INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Conhecer os clientes para melhor vender: Caso de estudo de uma empresa de transfer de turistas a operar na região do Algarve

Cláudio Manuel Neves Rocha

Mestrado em Sistemas Integrados de Apoio à Decisão

Orientadora:
Doutora Ana Maria Carvalho de Almeida, Professora Associada,
Iscte - Instituto Universitário de Lisboa

Coorientador:
Doutor Nuno Miguel da Conceição António, Professor Auxiliar Convidado,
Nova IMS - Universidade Nova de Lisboa

Outubro, 2020

TECNOLOGIAS
E ARQUITETURA

Departamento de Ciências e Tecnologias da Informação

Conhecer os clientes para melhor vender: Caso de estudo de uma empresa de transfer de turistas a operar na região do Algarve

Cláudio Manuel Neves Rocha

Mestrado em Sistemas Integrados de Apoio à Decisão

Orientadora:
Doutora Ana Maria Carvalho de Almeida, Professora Associada,
Iscte - Instituto Universitário de Lisboa

Coorientador:
Doutor Nuno Miguel da Conceição António, Professor Auxiliar Convidado,
Nova IMS - Universidade Nova de Lisboa

Outubro, 2020

Agradecimentos

Em primeiro lugar gostaria de expressar o meu profundo agradecimento aos meus orientadores, Professora Doutora Ana de Almeida e Professor Doutor Nuno António, pela sua dedicação e profissionalismo. Foram muito importantes em todo este longo processo, já que estiveram sempre presentes, em todos os momentos, e foram determinantes na minha motivação para a conclusão deste trabalho.

Gostaria ainda de agradecer a todos os professores do Mestrado em Sistemas Integrados de Apoio à Decisão do ISCTE-IUL, pelos ensinamentos transmitidos ao longo do curso. Esta dissertação reflete o conhecimento que me foi passado, aplicado a um caso prático.

Também não poderia deixar de referir o papel fulcral da empresa YellowFish Transfers, na realização deste trabalho, ao facultarem os dados utilizados nesta dissertação. O meu especial agradecimento vai para a equipa que acedeu em receber-me, nas instalações da empresa, e que contribuiu ativamente no esclarecimento das questões levantadas, no âmbito da compreensão dos dados e do negócio.

Por último gostaria de agradecer à minha mulher Marta e ao meu filho Frederico pela motivação e por acreditarem neste projeto. A sua compreensão e companhia tornaram os dias em que estive a trabalhar na dissertação mais divertidos e fáceis. Aos meus pais, João e Irene que apesar da distância sempre me apoiaram em todos os meus projetos.

A todos os que enumerei o meu sincero “Obrigado”.

Resumo

O presente trabalho apresenta o caso prático de uma empresa de transporte de passageiros, a operar no setor do turismo internacional, na região do Algarve. Propõe-se a utilização de técnicas de Extração de Informação e *Text Mining* para encontrar padrões nos dados que permitam conhecer os clientes e ainda estudar o impacto do marketing digital na procura dos serviços da empresa.

Foram utilizadas técnicas de *Text Mining* para extrair padrões dos comentários dos clientes de forma a condensar em tópicos e sumarizar o que estes pensam sobre o serviço. Recorreu-se ao histórico de transações, tendo sido aplicado algoritmos de aprendizagem não supervisionada para descobrir padrões nos dados que configuram segmentos de clientes. Os padrões revelados poderão ser utilizados em diferentes processos de tomada de decisão como por exemplo na criação de campanhas de marketing direcionadas para a criação de produtos específicos para cada segmento.

Na produção de previsões de impacto do marketing digital foi utilizada uma combinação entre modelos de regressão múltipla e técnicas de análise de séries temporais, de forma a compreender os fatores que explicam a procura dos serviços da empresa e consequentemente a receita da empresa. Investigaram-se técnicas mais recentes de Aprendizagem Automática de forma a estabelecer uma comparação entre os métodos estatísticos tradicionais de análise de series temporais e os algoritmos de Aprendizagem Automática. Os resultados de previsão de procura de serviços foram satisfatórios, tendo sido identificado a sazonalidade como o fator que mais afeta a procura dos serviços.

Palavras-Chave: Sistemas de Apoio à Decisão; Extração de Informação; Aprendizagem Automática; *Marketing Digital*; Procura Turística; Transfers; Segmentação Clientes.

Abstract

This work presents the case study of a passenger transport company, operating in the international tourism sector, in the Algarve region. It is proposed to use Data Mining and Text Mining techniques that allow to know the customers and also to find patterns in the data to study digital marketing's impact on demand for its services.

Text Mining techniques were used to extract customer comments patterns to condense into topics and summarize what they think about the service. We used transaction history and applied unsupervised learning algorithms to discover patterns in the data that configure customer segments. The revealed patterns can be used in different decision-making processes, such as creating targeted marketing campaigns to create specific products for each segment.

In producing digital marketing impact forecasts, a combination of multiple regression models and time-series analysis techniques were used to understand the factors that explain the demand for the company's services and, consequently, the company's revenue. More recent Machine Learning techniques were investigated to compare traditional statistical methods of time series analysis and Machine Learning algorithms. The service demand forecasting results were satisfactory, with seasonality having been identified as the factor that most affects the demand for services.

Keywords: Decision Support Systems; Data Mining; Machine Learning; Digital Marketing; Tourist Demand; Transfers; Clients Segmentation.

Índice

Índice de Quadros

Índice de Figuras

Lista de Abreviaturas e Siglas

AIC - Akaike's information criterion

ACP - Análise de Componentes Principais

APA - Agência Portuguesa do Ambiente

ARIMA - Auto-Regressive Integrated Moving Average

BI - Business Intelligence

BIC - Bayesian information criterion

CM - Contextual match

CPA - Cost per action

CPC - Cost per click

CPI - Cost per impression

CPM - Cost per mille

CTR - Click through rates

CRISP-DM - Cross-Industry Standard Process for Data Mining

CRM - Customer Relationship Management

DM - Data Mining

DSS - Decision Suport Systems

LDA - Latent Semantic Analysis

MAE - Mean absolute error

MAPE - Mean absolute percentage error

ML - Machine Learning

NLTK - Natural Language Toolkit

KDD - Knowledge Discovery and Data Mining

OLTP - Online Transaction Processing

PPC - Pay per click

RDBMS - Relational Database Management System

RFM - Recency, frequency e monetary

RGPD - Regime Geral de Proteção de Dados

RMSE - Root mean squared error

ROI - Return on Investment

SQL - Structured Query Language

SS - Sponsored search

SW - Shopping website

TIC - Tecnologias de Informação e Comunicação

KPI - Key performance indicators

Capítulo 1 – Introdução

1.1. Enquadramento e relevância do tema

Atualmente os modos de vida e padrões de consumo alteram-se a um ritmo cada vez mais acelerado, por via, nomeadamente da rápida evolução tecnológica. É crescente a procura da interação digital na escolha de produtos e serviços. Os utilizadores do digital deixam um rasto de informação pessoal que é agregada num conjunto de dados, formando o designado por *Big Data*. Neste contexto, dá-se o desenvolvimento das tecnologias de aprendizagem automática, capazes de fazer inferências sobre categorias, transformando-se em ferramentas poderosas para a análise de dados [1].

Impulsionado pela grande quantidade de dados e pelas técnicas de análise de dados disponíveis surge uma crescente sofisticação das operações de marketing digital. Estas passam a assumir um com grande valor, tanto para as empresas publicadoras de conteúdos digitais, como para as empresas que tentam promover os seus produtos ou serviços.

É neste contexto que a empresa *YellowFish Transfers* (*www.yellowfishtransfers.com*), a operar no ramo do turismo internacional no Algarve se enquadra. A empresa efetua transfers do aeroporto de Faro para toda a região do Algarve e vice-versa, tendo aderido a plataforma *Google Ads,* onde se encontram anúncios dos seus serviços. A empresa *YellowFish Transfers*, que se encontra em expansão e detém um elevado manancial de dados que pretende ver analisados e interpretados.

Assim, o objetivo do presente trabalho é o de efetuar um estudo que permita compreender que fatores contribuem positivamente para o aumento do volume de negócios e, consequentemente, das receitas da empresa. Pretende-se ainda entender que tipos de clientes procuram os seus serviços e como avaliam o serviço prestado, de modo a criar melhores "serviços", avaliar uma melhor estratégia de *pricing*, bem como definir uma melhor estratégia de marketing. Para atingir este objetivo, a empresa facultou os seus dados e disponibiliza-se para o acompanhamento e esclarecimentos necessários para o desenvolvimento do estudo.

1.2. Questões e objetivos de investigação

O projeto visa analisar os dados de reservas que a empresa de transfer de passageiros *YellowFish Transfers* (*YFT*) tem no seu histórico de operação e construir um sistema de avaliação de impacto do Marketing Digital, bem como de segmentação de clientes recorrendo às técnicas da área de conhecimento de *Extração de Conhecimento* e *Text Mining*.

Foram definidos os seguintes objetivos específicos de investigação:

- Conhecer e caraterizar os clientes da empresa, procurando identificar a sua segmentação.
- Identificação das características que os clientes mais valorizam no serviço prestado.
- Compreender o impacto de fatores como as campanhas do *Google Ads* no número de reservas efetuadas e, consequentemente, nas receitas da empresa, através da extração de padrões.

Os objetivos propostos permitem a empresa identificar o perfil dos seus clientes e conhecer os fatores que mais contribuem para o aumento da receita. Na prossecução dos objetivos propostos, foram formuladas as seguintes questões de investigação:

Q1: Que características do serviço são mais valorizados pelos clientes?

Q2: Que tipo de clientes procuram os serviços da empresa?

Q3: Que fatores contribuem positivamente para o aumento das receitas da empresa?

1.3. Metodologia

1.3.1 Abordagem seguida para a revisão da literatura

A área de estudo dos Sistemas de Informação é interdisciplinar. Essa interdisciplinaridade introduz complexidade e torna desafiante a construção de uma revisão de literatura, onde é necessário construir uma teoria tendo como ponto de partida diferentes áreas do saber.

Algumas publicações de revisão de literatura seguem uma abordagem mais tradicional centrada nos autores, onde é apresentado um sumário dos artigos mais relevantes. Alguns autores têm alertado para o fato de este método apresentar falhas, não sintetizando os conceitos abordados. Em oposição a esta abordagem centrada nos autores, Webster e Watson [2] argumentam que uma revisão de literatura com elevado padrão de qualidade deve seguir uma abordagem centrada em conceitos, sendo que a própria organização do trabalho de revisão de literatura deverá configurar uma *framework* orientada em conceitos.

Neste estudo foi seguida uma abordagem sistemática de revisão da literatura, centrada em conceitos, conforme proposto por autores como Webster e Watson [2], tendo sido efetuada em três etapas: (1) com recurso a repositórios de artigos científicos para a pesquisa de literatura, tais como: b-on; *Research Gate* e *Google Scholar*. Foram pesquisadas palavras-chave na língua inglesa relacionadas com o tema em estudo, como por exemplo: "*adwords*", "*Digital Marketing*", "online *advertising*", "*Data Mining and Marketing*", "*Pattern Mining*", "*Client Segmentation*". Webster e Watson [2] argumentam que, pelo fato do campo dos Sistemas de Informação ser interdisciplinar, durante uma revisão de literatura, para desenvolver uma teoria é necessário pesquisar fora da esfera dos Sistemas de informação. No presente estudo foi ainda necessário pesquisar termos fora da área dos Sistemas de Informação como por exemplo, "*Travel and tourism*", "*Airport Transfers*", "*Airport Ground access*", "*Market Segmentation*"; (2) foram revistas as citações dos artigos identificados no passo anterior, de forma a identificar artigos anteriores que pudessem ser considerados relevantes; (3) foi efetuada uma nova pesquisa, utilizando os repositórios de dados escolhidos, para encontrar artigos recentes que referenciam os artigos selecionados nos pontos anteriores.

Considerou-se imperiosa a validação da relevância científica dos conceitos abordados no conjunto inicial de artigos que serviram de ponto de partida para a investigação. Para

tal, seguiu-se uma aproximação convencional de validação cruzada do conjunto inicial de artigos, onde foi analisado o título, o resumo e o conteúdo tendo em consideração cinco critérios: (1) a pesquisa empírica, que procura compreender as formas de marketing digital e a importância da segmentação dos clientes na definição de uma estratégia de marketing; (2) a identificação das métricas de marketing digital e a utilização das tecnologias de aprendizagem automática para avaliar a sua efetividade; (3) o *Marketing Digital* aplicado ao turismo e aos transportes, com ênfase nos modos de acesso aos aeroportos, tentando compreender o papel dos serviços de transfers de passageiros entre os aeroportos e um local; (5) a utilização de técnicas de Data Mining para a construção de sistemas de suporte à decisão na implementação de uma estratégia de marketing na área do turismo.

1.3.2 Metodologia de desenvolvimento do projeto

No desenvolvimento do projeto procurou-se seguir as melhores práticas, de forma a maximizar as possibilidades de sucesso do projeto. Num projeto de Extração de Informação, existem diferentes metodologias de implementação. No presente projeto foi utilizado uma das mais populares metodologias, proposta nos anos 90, por um consórcio de empresas Europeias como uma metodologia standard não proprietária, o *CRISP-DM*. A Figura 1, ilustra a metodologia proposta, constituída por 6 grandes fases, iniciando com a compreensão do negócio e terminando com o *deployment* de uma solução que pretende dar resposta às necessidades específicas de negócio.

Na metodologia *CRISP-DM*, formular o problema e encontrar uma solução é um processo interativo de descoberta, representado como ciclos dentro de um ciclo, em vez de um simples processo linear. Como as etapas são construídas tendo por base os resultados das etapas anteriores, foi necessário conferir especial atenção às etapas iniciais de forma a não comprometer o resultado da investigação.

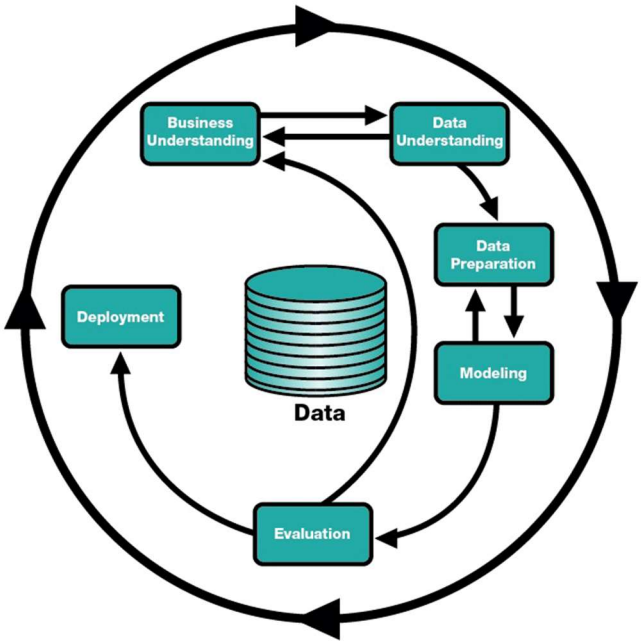


Figura 1 Fases da metodologia CRISP-DM. Fonte: crisp-dm.eu

1.4. Estrutura e organização da dissertação

O presente estudo está organizado em 5 capítulos que pretendem refletir as diferentes fases do processo de pesquisa até a conclusão da dissertação.

O atual capítulo introduz o tema da investigação os objetivos e questões de investigação da mesma, bem como uma breve descrição da estrutura do trabalho.

O segundo capítulo designado por Revisão da literatura, reflete o enquadramento teórico.

O terceiro e quarto capítulos são dedicados à Metodologia utilizada para o desenvolvimento do projeto, onde é descrito o processo de recolha e tratamento de dados, bem como os métodos de análise utilizados. Estes capítulos apresentam ainda a análise dos resultados obtidos, de acordo com o *CRISP-DM*.

No quinto e último capítulo apresentam-se as conclusões deste estudo assim como as recomendações, limitações da pesquisa e proposta para trabalhos futuros.

Capítulo 2 – Revisão da literatura

O presente capítulo inicia com uma revisão de literatura dos transportes e turismo apresentando estudos que realçam o papel da liberalização do espaço aéreo internacional e o surgimento das companhias aéreas de baixo custo no aumento da procura de viagens e consequente aumento do tráfego aéreo e terrestre. No que diz respeito a utilização dos modos de acesso aos aeroportos, foram escassos os estudos encontrados que exploram a utilização dos transfers. Na secção 2.1.4 são apresentados alguns motivos para a necessidade da utilização desse tipo de transportes, apesar de ecologicamente ineficientes.

A secção 2.2 procura fazer uma revisão de literatura do Marketing Digital aplicado aos transportes e turismo, apresentando trabalhos que abordam os diferentes formatos de Marketing Digital e as métricas e algoritmos utilizados.

Na secção 2.3 é efetuada uma breve introdução aos sistemas de apoio à decisão e a sua relação com as tecnologias de informação orientadas para os dados e as questões do ponto de vista sociológico, que devem ser acauteladas, para que seja garantida a proteção do cidadão.

Por último são apresentadas as técnicas utilizadas para a análise preditiva de series temporais e apresentados estudos que utilização os algoritmos de aprendizagem automática em alternativa aos métodos estatísticos de análise de séries temporais.

2.1. Transportes e turismo

2.1.1. A liberalização do espaço aéreo

A liberalização do espaço aéreo teve início nos Estados Unidos da América em 1978 e, desde então, tem vindo a ser implementada a nível mundial. Os países que inicialmente permitiram a concorrência entre companhias aéreas foram o Canadá, a Austrália e o Reino Unido, seguidos da União Europeia, no início dos anos 90. Desde o ano 2000 que a China, a Índia, a Tailândia e outras nações asiáticas aderiram à liberalização do espaço aéreo [3]. Anteriormente, em 1997, o processo de liberalização do espaço aéreo europeu comunitário conduziu a profundas mudanças na forma como as pessoas viajam. As novas regras comunitárias permitiram o surgimento das companhias aéreas de baixo custo, com um modelo de negócio distinto das tradicionais companhias aéreas regulares e *charters*,

permitindo a criação de novas rotas aéreas e levando a abertura de novos aeroportos e ao surgimento de novos destinos turísticos.

O crescimento das companhias aéreas de baixo custo e o desenvolvimento das tecnologias de informação associadas a utilização massiva da internet permitiram ao turista ter acesso direto aos fornecedores de serviços turísticos de forma a poderem customizar as suas viagens. Essas novas tendências do turismo conduziram a um aumento da mobilidade turística no país ou região de acolhimento. Alguns estudos demonstram que as novas tendências do turismo introduzidas pelo surgimento das companhias aéreas de baixo custo, levaram a um aumento na procura de viagens, provocando o aumento do tráfego terrestre [3].

2.1.2. O setor do turismo internacional no Algarve

A liberalização do espaço aéreo europeu provocou mudanças consideráveis na estrutura do tráfego aéreo do aeroporto de Faro, passando a ter, em 2010, cerca de 75% dos passageiros transportados por companhias aéreas de baixo custo [4]. Segundo os dados publicados pelo aeroporto de Faro, atualmente a maioria das ligações é assegurada por companhias aéreas baixo custo como a *Easyjet*, *Jet2.com* ou a *Ryanair*. Mais de 90% do tráfego aéreo do aeroporto é internacional, sendo que 80% dos voos são oriundos do Reino Unido, seguidos da Alemanha, Holanda e República da Irlanda, constituindo os principais mercados que o setor do turismo no Algarve explora.

Apesar da introdução das companhias áreas de baixo custo, o aeroporto de Faro continua a depender excessivamente de alguns mercados, como o Reino Unido, Alemanha e Holanda. No contexto atual de intensa competição entre destinos turísticos, esta dependência pode influenciar negativamente a procura a médio prazo [5]. A eminente saída do Reino Unido da União Europeia (Brexit), envolta em incerteza, representa uma séria ameaça para o setor do turismo dos países do mediterrâneo devido à dependência do mercado Britânico. Considera-se que deverão ser realizados estudos de modo a compreender de que forma o Brexit afeta os diferentes destinos turísticos da União Europeia e extra União Europeia [6].

2.1.3. O crescimento do turismo e possíveis impactos na mobilidade

O aumento do poder de compra a nível mundial permite aos indivíduos e famílias viajarem para destinos cada vez mais distantes [7]. O turismo constituiu o setor da atividade económica que mais tem crescido nas últimas décadas, enfrentado diferentes desafios devido a globalização. O rápido crescimento do turismo permitiu o surgimento de uma indústria de viagens e turismo fortemente competitivo, onde os países competem para o controlo dos destinos turísticos. A competitividade pelos destinos turísticos tem crescido principalmente nos países que dependem em grande medida da indústria das viagens e turismo. Navickas e Malakauskaite [7] destacam ainda a importância do setor do turismo em economias de países cuja atividade principal se centra em indústrias não relacionadas com o setor do turismo. Os autores defendem que fatores como o desenvolvimento tecnológico e das infraestruturas, como é o caso da rede de transportes terrestres e linhas aéreas, estão fortemente correlacionados com a competitividade dos destinos turísticos globais.

A relação entre o turismo internacional e os sistemas de transporte é amplamente debatida na literatura académica, centrando-se essencialmente no contributo dos transportes aéreos para o aumento da procura de destinos turísticos.

Autores como Khan et al. [8] apontam para a necessidade de considerar outros meios de transporte relacionados com o turismo, como por exemplo, o transporte ferroviário e outros serviços de viagens e transporte. Nas suas pesquisas analisaram o impacto dos transportes no turismo internacional de entrada (*inbound*) e saída (*outbound*), em 19 destinos turísticos, entre 1990 e 2014. Concluíram que a concentração de diferentes meios de transporte ajuda a promover o turismo internacional, destacando o impacto positivo dos serviços de viagem e transporte no turismo internacional nos próximos 10 anos.

Aguiló et al. [9], conscientes do impacto do trânsito automóvel e dos níveis de poluição atmosférica na procura do turismo e na qualidade de vida da população local, investigaram a escolha dos principais meios de transporte nas ilhas baleares, de acordo com os perfis dos turistas. Para estes autores, o aumento do uso das companhias aéreas de baixo custo, o abandono dos pacotes turísticos e dos resorts tradicionais, aliados ao aumento do poder de compra, conduziram para níveis mais elevados da utilização do automóvel privado nos destinos turísticos, com impacto no tráfego local e consequente aumento na emissão de gases poluentes. Nesse sentido, propuseram diferentes alternativas para reduzir o impacto da utilização do transporte privado na procura do

turismo, incluindo a melhoria das infraestruturas, para evitar congestionamento no trânsito, ou o recurso a subsídios para incentivar o desenvolvimento do setor dos transportes coletivos e impostos para dissuadir a utilização do transporte privado pelos turistas.

Para reduzir os impactos ambientais e económicos, o incentivo para a utilização dos transportes públicos tem sido prioridade em quase todos aeroportos mundiais. No entanto, poucos aeroportos conseguiram atrair os passageiros para os sistemas de transporte público [10].

No Aeroporto Internacional de Hamad no Qatar, algumas características da viagem como, o número de viajantes, a quantidade de bagagem transportada, o objetivo da viagem, o custo e tempo de viagem, os custos de estacionamento são fatores determinantes na escolha do meio de transporte para as deslocações para o aeroporto. Além disso, as características socioeconómicas, como a nacionalidade, o rendimento mensal do agregado familiar, o tipo de emprego, a posse de veículos e a idade, foram consideradas determinantes na escolha de automóveis particulares nas deslocações para o aeroporto [11].

A análise dos dados recolhidos de 282 turistas na cidade de Macau, revelou que a qualidade dos serviços de transporte de turistas (shuttles) condiciona a intenção de revisita do turista. Os resultados sugerem que os esforços devem ser direcionados no sentido de manter um serviço de transporte turístico de alta qualidade, especialmente nas áreas de atendimento, eficiência, pontualidade e segurança [12]. Estudos anteriores indicam ainda que os transportes passaram a fazer parte da experiência do turista [13], afetando o padrão das viagens e condicionando a escolha do destino [14]. Assim, a qualidade dos transportes pode influenciar a perceção da qualidade do destino [15]. Alguns estudos, [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], defendem que o segmento a que os passageiros pertencem condicionam a escolha do modo de acesso aos aeroportos. Um passageiro que viaja em negócio tem diferentes prioridades na escolha do modo de acesso ao aeroporto quando comparado com um passageiro que viaja em lazer ou ainda com passageiros de voos domésticos ou internacionais. Klähna e Hallb [29], consideram as características do destino como um conceito chave no turismo, influenciando na decisão da escolha do meio de transporte.

A revisão da literatura é consistente nas variáveis utilizadas para explicar o perfil dos passageiros e a escolha do modo de acesso aos aeroportos. Essas variáveis podem ser

categorizadas em dois grupos. Um primeiro grupo compreende as características da viagem, onde se destaca a duração, o custo da viagem, o tipo de destino e um segundo grupo que incluiu as características socioeconómicas e demográficos dos viajantes.

2.1.4. Escolha de transfer privado como modo de acesso aos aeroportos

O serviço de transfer consiste no transporte pré-combinado de um passageiro entre o aeroporto e um local, podendo ser um hotel ou outro local de interesse. Normalmente trata-se de um serviço pré-pago, através de um sistema de reservas, uma agência de viagem, um operador turístico, ou até serviços gratuitos de shuttle disponibilizados pelos hotéis. Não são considerados serviços de transfer, os táxis, *rideshare*, como por exemplo, o uber ou os transportes públicos.

Um dos benefícios da utilização do transfer privado reside na conveniência, eliminado o stress de ter de se chamar um táxi ou de transportar bagagens nos transportes públicos, principalmente quando se trata de famílias.

Embora na literatura sejam escassas as pesquisas que exploram o tópico da utilização dos transfers como modo de acesso aos aeroportos, alguns autores como, Mandle et al. [30] e Neufville [3], apontam algumas razões para a necessidade dos aeroportos recorrerem a este tipo de transportes. Segundo Neufville [3], o modelo de negócio das companhias aéreas de baixo custo, conduziu a uma dispersão do tráfego aéreo por múltiplos aeroportos a servir a mesma área metropolitana com o intuito de baixar os custos aeroportuários. Este modelo de negócio introduz a necessidade dos aeroportos investirem em modos de acesso mais flexíveis, capazes de expandir e contrair de acordo com as necessidades e fornecendo um serviço de transporte porta-a-porta. O autor aponta como exemplo, os serviços privados de transporte de passageiros (transfers ou *shuttles*). Já Mandle et al. [30] sugere que o utilizador típico dos transportes públicos que servem os aeroportos, são viajantes familiarizados com estes sistemas de transporte, que transportam pouca ou nenhuma bagagem, viajam sós e são capazes de caminhar da paragem ou estação até ao destino final. Por esse motivo, o autor conclui que os operadores aeroportuários devem encorajar a utilização de modos de acesso mais eficientes, como o serviço de transfers.

Apesar da necessidade dos serviços de transporte porta-a-porta, deve ser referido que atualmente os aeroportos estão sob uma crescente pressão no sentido de seguirem uma

política de baixa emissão de carbono. Uma das principais fontes de emissão de carbono relacionadas com os aeroportos são as viagens de e para os aeroportos, nesse sentido as questões relativas às opções de transporte de trabalhadores e passageiros devem ser geridas e acompanhadas para que seja minimizada o impacto negativo ao nível da qualidade do ar.

A afirmação anterior é corroborada pelo estudo recente de impacto ambiental, TUA [31], para a construção do novo aeroporto do Montijo efetuado pela APA. O estudo aponta como uma das principais fontes de emissão associadas a exploração do futuro aeroporto, o transporte rodoviário e os consumos energéticos inerentes ao funcionamento do Terminal. O mesmo estudo inclui pareceres de entidades que apontam alguns aspetos a considerar, a título de exemplo, a Metro – Transportes do Sul, é da opinião de que as apostas nos transportes coletivos de passageiros para reduzir a emissão de carbono não são suficientemente ambiciosas face aos desafios do futuro. Na realidade, a opção de se apostar exclusivamente no reforço do transporte público rodoviário e fluvial poderá vir a tornar-se um desincentivo à utilização de um transporte coletivo de passageiros, incentivando a que os passageiros optem naturalmente por veículos ligeiros de passageiros, quer através de um serviço de táxi/Uber/Cabify quer através do serviço de transfer privado. São ainda referidos como exemplo os dados relativos ao Aeroporto Humberto Delgado que indicam uma fraca utilização da rede de transportes públicos de Lisboa, a Carris, tem uma utilização de apenas 2%, sendo que metade dos passageiros utiliza o serviço *shuttle*. O estudo recomenda que seja potenciado como instrumento estratégico o papel dos serviços de transportes no desenvolvimento do sector do turismo nacional e promover a melhoria das condições de acolhimento e da qualidade do serviço bem como da informação ao público sobre os transportes, contribuindo para o grau de satisfação dos passageiros, em particular dos turistas.

2.2. Marketing Digital nos transportes e turismo

2.2.1. Visão geral do Marketing Digital aplicado ao turismo e transportes

O desenvolvimento das tecnologias de informação e comunicação (*TIC*), especialmente a internet, conduziu a profundas mudanças na aplicação do marketing no turismo. Atualmente os turistas são cada vez mais exigentes e apresentam necessidades específicas de consumo, principalmente devido à grande quantidade de informação

disponível na internet que lhes permitem encontrar facilmente os produtos que melhor se adaptam ao seu orçamento e necessidades.

Do ponto de vista das empresas de marketing, a internet trouxe uma nova era para o marketing, permitindo ter acesso a potenciais clientes a nível internacional a um baixo custo. Vários fatores contribuíram para a adoção massiva da internet como plataforma de marketing, tais como o acesso rápido e contínuo a uma grande variedade de informação, a capacidade de comunicação com qualquer pessoa a qualquer momento, a competência de personalização das atividades de marketing e a possibilidade de medir o desempenho das estratégias comerciais, [32]. A utilização dos meios de comunicação tradicionais tem vindo a perder terreno como plataformas para campanhas de marketing direcionados ao turismo. Todos os anos, desde 2010, a taxa de crescimento de publicidade na internet tem ultrapassado todos os outros canais de comunicação, com taxas de crescimento de dois dígitos, conforme ilustrado na Figura 2.

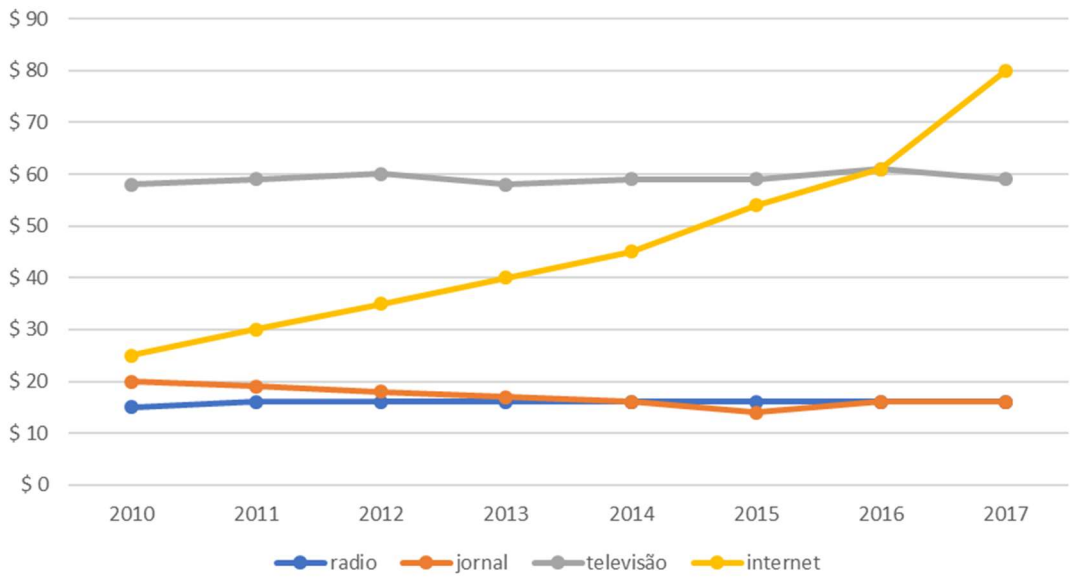


Figura 2 Lucro (mil milhões de dólares) em publicidade (Fonte: Adaptado de IAB, The Interactive Advertising Bureau)

As *TICs* revolucionaram o turismo e criaram ferramentas que permitem interagir com o contexto dos utilizadores, dando origem a oportunidades para novos níveis de serviços e interação [33]. A utilização das redes sociais, baseada em informação de contexto e o uso dos dispositivos móveis deu origem a um novo conceito que permitiu criar produtos e serviços de viagens e turismo de forma dinâmica, fornecendo recomendações personalizadas e ajustadas às alterações que ocorrem no contexto de uma viagem [33].

No turismo, o termo *contexto,* é definido como sendo "qualquer informação relevante que caracteriza o turista" e influencia o seu padrão de comportamento [34]. A informação de contexto pode ser considerada como: (i) interna ou fator humano, quando inclui objetivos pessoais, eventos sociais ou o estado emocional; (ii) externa, quando relacionada com as características do meio ambiente, como, por exemplo, a proximidade de um objeto ou as condições climatéricas [35]. A captura da informação de contexto do turista através das redes sociais e dispositivos móveis permitiu fornecer serviços de guia turístico [36], serviços de alertas associados a localização geográfica [37] e a criação de sistemas de recomendação baseados no contexto [38], [39], [40], [41], [42].

2.2.2. Os formatos de Marketing Digital

Inicialmente o *Marketing* digital procurou reproduzir os formatos de publicidade *off-line* onde os anúncios impressos, mais gráficos e menos orientados ao texto, tendem a causar maior impacto com maior probabilidade de virem a ser recordados [43].

Existem diferentes formatos de marketing digital: *Opt-In de emails, newsletters*, *instant messaging* e apresentação de anúncios em páginas *web*. Os *Opt-In de emails* e *newsletters* são realizados através do envio de anúncios por e-mail, para uma lista de interessados em receber informação relacionados com um determinado tema [44].

Foram identificadas quatro formas de apresentação de anúncios em páginas web [45]: *pop-up* e *popunder*, anúncios flutuantes (*floating ads)*, intersticial, e *banners*. Os anúncios que recorrem a *Pop-ups,* abrem uma pequena janela sobre a página principal quando o utilizador entra num determinado *site*. Nos anúncios *pop-under* a janela é aberta na parte inferior da página principal. Os anúncios flutuantes (*floating ads*) são apresentados quando o utilizador entra na página e flutuam por alguns segundos antes de ocuparem uma determinada posição na página. Os anúncios intersticiais são apresentados na transição entre páginas. Quando o utilizador pretende navegar de uma página A para B é apresentado uma página C com o anúncio e com um *link* para a página B. Muitos utilizadores consideram este tipo de anúncios intrusivos provocando demoras no tempo de carregamento das páginas [45]. O conceito dos *banners* consiste em apresenta uma caixa retangular de anúncio, numa área da página de um fornecedor de serviço publicitário, que ao ser clicado, redireciona o utilizador para a página do anunciante.

Existem três abordagens principais para o *marketing digital*: a abordagem não segmentada, a segmentada ou filtrada e a personalizada [46]. A abordagem não segmentada foi utilizada inicialmente por empresas pequenas, com o objetivo de apresentar anúncios estáticos nas suas páginas por um determinado período. Apesar deste tipo de anúncios ser bastante simples de implementar e de não apresentar constrangimentos que ponham em causa a privacidade dos utilizadores, na sua maioria não captam o interesse dos utilizadores, não satisfazendo o anunciante pela baixa taxa de *clicks* que apresenta [44]. Nos anúncios filtrados ou segmentados, o anunciante pode definir alguns parâmetros como a localização geográfica, o tipo de sistema operativo e a hora. Um sistema de anúncios filtrados possui um mecanismo de seleção que analisa o pedido do utilizador e apresenta apenas os anúncios que satisfazem determinados critérios [44]. Os sistemas de anúncios personalizados utilizam o *web mining* e os algoritmos de aprendizagem automática para apresentar anúncios personalizados, a um determinado utilizador tendo por base um conjunto de variáveis que podem estar relacionados com o perfil de navegação do utilizador, ou com os dados demográficos [46].

Os anúncios personalizados podem ser classificados em três categorias, de acordo com a sua aplicação: anúncios através de motores de pesquisas (*sponsored search* ou *SS*), anúncios com base em informação de contexto (*contextual match* ou *CM*), e os anúncios através de lojas online (*shopping website* ou *SW*). As *SS* apresentam os anúncios nos resultados das pesquisas efetuadas pelos utilizadores nos motores de pesquisa [46]. Os anúncios de contexto ou *CM* são apresentados dentro de uma página *web* de acordo com a sua relevância para conteúdo da página [47], [48]. Os anúncios em lojas online, *SW*, são apresentados na página principal das lojas online, de acordo com critérios como a última compra ou os dados demográficos dos clientes [49].

Atualmente a *Interactive Advertising Bureau* [50] identificou as seguintes formas de Marketing Digital: *Banner advertising Sponsorship, Email, Search, Lead generation, Classifieds and auctions, Rich media, Digital audio, Digital video advertising, Mobile advertising, Social media advertising, Impression-based , Performance-based, Hybrid*.

Na Figura 3 é apresentada a comparação, entre dispositivos móveis e computadores, da distribuição do rendimento pelos principais formatos de *marketing* digital, no último semestre de 2018 e no primeiro semestre 2019.

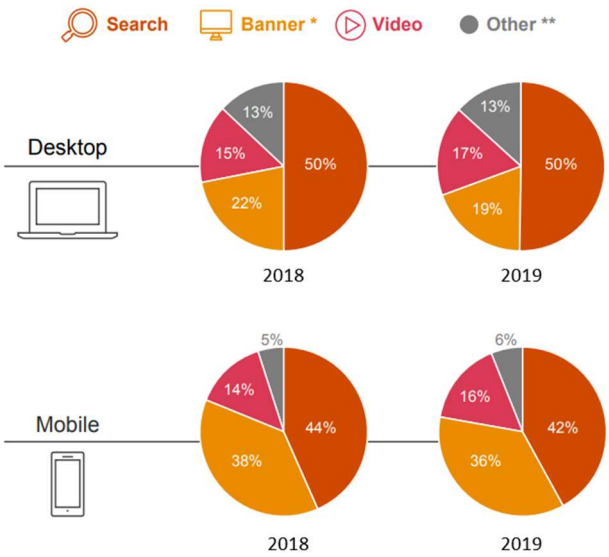


Figura 3 Distribuição dos lucros da publicidade por formato. (Fonte: IAB, The Interactive Advertising Bureau, relatório de 2019

2.2.3. Métricas e modelo financeiro do Marketing Digital

Todas as formas de marketing digital e, em particular, o *Marketing* de contexto (*CM*) apresentam vantagens quando comparadas com as formas de marketing tradicional. Uma dessas vantagens reside na relativa facilidade em medir a sua efetividade [47].

Os anúncios através de motores de pesquisas (*SS*) e os baseados em informação de contexto (*CM*) dependem, em grande medida, da capacidade de criar modelos de aprendizagem capazes de predizerem com precisão, rapidez e com elevado nível de confiança a probabilidade de um anúncio ser clicado, caso este seja apresentado (*click through rates ou CTR*) [51]. Através do histórico dos cliques, os motores de pesquisa podem fazer previsões acerca do *CTR* esperado [52].

O processo de submissão de anúncios online, consubstancia-se em encontrar o espaço na internet para disponibilizar o anúncio certo no momento certo [49]. Para atingir esse objetivo, estão envolvidas três perspetivas: a perspetiva do publicitário, que coloca o anúncio, a perspetiva do publicador que vende o espaço publicitário online e a perspetiva do consumidor que interage com os anúncios na forma de cliques. Esta interação com o consumidor é o que a empresa de publicidade paga [53]. Quando o utilizador clica no *link* do anúncio é redirecionado para a página do anunciante. Nesse caso, o modelo financeiro

predominante consiste no anunciante pagar um montante por cada clique efetuado no anúncio (*cost per click*, *CPC*) [47].

Desde o início que as empresas publicitárias se preocupam com as formas de medir a efetividade dos seus anúncios. Por outro lado, as empresas publicadoras de anúncios preocupam-se com as formas de atribuir um preço aos seus serviços [54]. Algumas empresas publicadoras cobram uma taxa fixa para apresentar os anúncios numa página web, enquanto outras, como a *Netscape* e a *Infoseek,* em 2005, utilizaram o custo por mil apresentações (*CPM*) – o custo de apresentar um anúncio mil vezes [54]. Existem outras variantes como, o custo por impressão, (*cost-per-impression CPI*) onde o publicitário paga pelo número de vezes que o anúncio é mostrado e o custo por ação (*cost-per-action CPA*) onde a empresa que submeteu o anúncio paga apenas se houve uma transação completa [49]. Por exemplo, um *site* de viagens sabe que um determinado passageiro que comprou um bilhete de avião chegou ao *site,* porque pesquisou pelo termo "viagem" num determinado motor de pesquisa [54]. Dependendo das circunstâncias, o anunciante poderá traçar um objetivo e medir a sua execução através das métricas *CPA*, *CPC* ou *CPM* [54]. O rácio entre o *CPC* e *CPA* é conhecido com a taxa de conversão. A taxa de conversão e a percentagem de cliques (*CTR*) são probabilidades condicionais: de uma ação dado um clique e de um clique dado uma apresentação [54].

A introdução das métricas do marketing digital como o *CTR* e *CPA,* por companhias como a Google, permitiram que os gestores das campanhas de marketing pudessem gerir de forma mais eficaz o investimento feito em publicidade online em comparação com o investimento feito noutros meios de comunicação como a televisão [55].

Foram delineadas várias estratégias para aumentar a efetividade do marketing digital. Uma dessas estratégias consistia em aumentar o investimento na publicidade que recorre aos motores de pesquisa (*SS*). Foi identificado que os utilizadores da *internet* têm a tendência de encontrar os produtos pretendidos, através de pesquisas de palavras-chave relevantes [56].

Contudo, utilizar as métricas identificadas anteriormente para aferir sobre a efetividade do marketing digital apresenta alguns desafios, dado que as mesmas não avaliam o impacto da interação entre os diferentes formatos de marketing digital nas vendas [55]. A título de exemplo, a publicidade através de *banners* influencia as pesquisas futuras do consumidor, o que produz impacto nas vendas. Nesse sentido foi proposta uma análise das métricas, considerando a dinâmica existente entre as formas de marketing e concluiu-

se que ao considerar a dinâmica existente entre as diferentes formas de marketing, o *CPA* dos motores de pesquisa foi 48% inferior quando comparado com o *CPA* estático, enquanto o retorno do investimento (*ROI*) das pesquisas foi 38% superior [55]. Essa nova forma de avaliação das métricas conduziu a novas formas de alocação de recursos para o marketing digital.

O dinamismo existente entre as diferentes formas de marketing online foi confirmado pelo estudo da *Google* [57] que concluiu que na banca tradicional os clientes são influenciados pelas pesquisas que efetuaram, entre um a dois meses antes da conversão. O estudo concluiu ainda que para produtos relacionados com seguros e automóveis a conversão é afetada por pesquisas efetuadas entre um a três meses antes da conversão, mas para produtos tecnológicos, o período é de apenas entre 1 a 4 semanas antes da compra. É esperado que o dinamismo existente entre as formas de marketing online perdure e que o seu efeito seja ainda maior para produtos e serviços que exigem maior ponderação por parte dos consumidores, como é o caso dos produtos financeiros, a compra de casa, de automóveis e de viagens. Esse dinamismo verifica-se com menor intensidade nos bens de consumo uma vez que o consumidor tende a ponderar a compra desse tipo de bens por um curto período.

Do ponto de vista dos sistemas comerciais, existem programas informáticos que selecionam os anúncios a apresentar, recorrendo a sistemas de leilões em tempo real, onde as empresas que publicam os seus anúncios licitam sobre uma palavra-chave [44]. Neste tipo de sistemas o *CPC* de um anúncio corresponde a uma licitação num sistema de leilão [52]. Juntamente com o sistema de leilões, são utilizados algoritmos de aprendizagem automática para prever o *CTR* e determinar se o anúncio deve ser apresentado, em que posição deve ser apresentado na página de resultados e o preço que a empresa que publicou o anúncio deve pagar, caso seja clicado [51].

O Google Adsense [58] foi um programa pioneiro na apresentação de anúncios de contexto *CM*, permitindo que as empresas publicadoras de anúncios fornecessem os serviços de pesquisa do *Google* e apresentassem anúncios do *Google* nos seus próprios *sites*. O *Google Adwords* [59] é um serviço fornecido às empresas de publicidade que lhes permite criar anúncios, selecionar e licitar sobre palavras-chave. Quando um utilizador pesquisa a palavra-chave no *Google*, o anúncio é apresentado em conjunto com o resultado da pesquisa, de acordo com um mecanismo de classificação e o valor licitado. O *Google Ads* permite ainda dirigir os anúncios a utilizadores da internet de uma

determinada região geográfica e utiliza o modelo de pagamento *PPC* [44]. Outras empresas como a *Yahoo*! e a *MSN* adotaram a mesma tecnologia, sendo as principais fornecedoras de publicidade online conforme listadas no site Contextual Advertising [60].

2.2.4. Marketing Digital através de pesquisas pagas (*sponsored search* ou SS)

Na secção 2.2.2 deste documento foram apresentados os diferentes formatos de marketing digital existentes, contudo, na literatura analisada foi atribuído maior destaque ao marketing, através das pesquisas pagas (SS). Os autores Saura et al. [61] recorreram ao marketing através de SS para estabelecerem os principais *key performance indicators* (KPI) e métricas de desempenho do marketing digital, alegando que os motores de pesquisa constituem o principal canal de contato entre as empresas e clientes. Este facto é evidente nas estatísticas apresentadas nos relatórios da IAB ao longo dos anos (vide Figura 3).

Antes de a publicidade estar inserida nos resultados das pesquisas *web (SS),* os motores de pesquisa baseavam-se nos mecanismos de *banners* para apresentar os anúncios. Por esse motivo tiveram de lidar com o dilema de reter o utilizador o maior tempo possível na página de pesquisa, de forma a apresentar o maior número de *banners* possível, ou direcionar o utilizador para os sites apresentados no resultado da pesquisa [54]. Em 1996 esse dilema foi ultrapassado quando o motor de pesquisa *Open Text* apresentou o modelo de listagens preferenciais, onde as páginas *web* pagam para serem incluídas nos resultados de pesquisas para uma terminada palavra-chave [54].

As perspetivas introduzidas por este modelo foram tão atrativas que nos meados dos anos noventa, muitas empresas estavam dispostas a investir grandes quantidades de dinheiro na publicidade online*,* sem qualquer preocupação com o retorno do investimento [62]. Contudo, com a crise das *dot-com,* no primeiro trimestre do ano de 2001, esta situação inverteu-se provocando uma redução considerável do investimento, nas tecnologias da internet e consequentemente na publicidade online [62], [63]. No entanto, esta tendência negativa viria a inverter-se nos finais de 2002, devido a adoção crescente de um certo formato de publicidade online, a publicidade através de motores de pesquisas [45].

A tarefa principal dos sistemas de anúncios que utilizam os motores de pesquisa (SS) consiste em determinar qual o anúncio a apresentar, em que ordem deve ser apresentado

para cada pesquisa efetuada e o preço a cobrar por cada clique no anúncio [52]. Para maximizar o lucro e a satisfação dos clientes, os sistemas devem ser capazes de predizer o comportamento dos utilizadores para cada anúncio apresentado, e estimar se este irá clicar no anúncio.

Os sistemas de pesquisa recorrem a modelos preditivos, que utilizam a informação histórica dos cliques dos anúncios para prever o comportamento dos utilizadores relativamente aos anúncios [52]. Por exemplo, se um anúncio for apresentado 100 vezes no passado, e obteve 5 cliques, então o sistema pode estimar o *CTR* de 0,05. No entanto, devido à alta variância que este tipo de estimativa pode apresentar, só é razoável ser aplicado para anúncios com várias apresentações. Para os anúncios inseridos recentemente no sistema, sem informação histórica, esta estimativa não se aplica, não sendo possível atribuir um valor de *CTR*. Alguns autores estudaram esta problemática, e propuseram utilizar o valor do *CTR* de anúncios com termos similares, como estimativa do *CTR* de anúncios sem histórico de cliques [52], [64]. Prever o *CTR* foi considerado um problema de aprendizagem em larga escala. É necessário efetuar vários milhões de previsões ao longo do dia e atualizar os modelos de aprendizagem, sempre que são observados novos cliques ou quando são inseridos novos anúncios, sem qualquer clique [51]. O problema de estimar a probabilidade do utilizador clicar no anúncio (*CTR*) foi largamente debatido na literatura. A Tabela 1 apresenta algumas abordagens de implementação e lista as variáveis e os algoritmos de aprendizagem utilizados.

Tabela 1 Abordagens para prever o CTR

Autor	**Variáveis utilizadas**	**Algoritmo/Ferramenta**
Richardson et al. [52]	**Aparência:** Quantas palavras estão no título e no corpo do anúncio? O texto do anúncio tem boa formatação? Contém demasiadas exclamações , sinais monetários (ex. $, €), ou outra pontuação? Utiliza palavras curtas ou palavras longas? **Captura de atenção:** O título contém palavras que denotam ação, tais como "comprar", "aderir", "subscrever", etc.? O anúncio apresenta valores numéricos (tais como descontos específicos, preços, etc.)? **Reputação:** O URL termina com .com, .net, .org, .edu? Qual é o seu comprimento? Existem muitos segmentos no URL (por exemplo books.com é geralmente melhor do que books.something.com)? Contém travessões ou números?	Regressão Logística

Autor	Variáveis utilizadas	Algoritmo/Ferramenta
	Qualidade da página de destino: A página contém código em flash? Qual a percentagem da página que está é coberta com imagens? A página é compatível com o W3C? Usa um guia de estilos? Apresenta muitos anúncios? **Relevância**: O termo licitado é igual ao título ou apenas parte corresponde a parte do título? Se o termo é apresentado no corpo da página, em que fração é apresentado?	
Broder et al. [65]	Palavras utilizadas na pesquisa. Resultado da Pesquisa	
McMahan et al. [51]	- Termo utilizado na pesquisa. - Texto do anúncio. - Metadados do anúncio	- Classificador baseado nos centróides dos documentos. - Refinamento com Altavista Prisma
Shi e Li [66]	- CTR nos últimos 7 dias - CPC Máximo. - Posição médio do resultado da pesquisa. - Posição medio dos últimos 7 dias.	- Regressão Linear - Árvores de Decisão - *Gradient Boost*

2.2.5. Segmentação de clientes

Smith [67] foi um dos primeiros autores a introduzir o conceito de segmentação de clientes/mercado na definição de uma estratégia de Marketing, ao defender a ideia de que o sucesso das ações de Marketing depende da combinação entre a diferenciação dos produtos e a segmentação do mercado. O autor defende que a segmentação consiste em dividir o mercado heterogéneo em pequenos grupos homogéneos.

Os termos segmentação de mercado e segmentação de clientes estão estritamente relacionadas e, muitas vezes, são utilizados sem distinção, [68]. A segmentação de clientes é uma ferramenta utilizada pelos gestores de Marketing, no processo de tomada de decisão, para a seleção de um mercado alvo de um determinado produto [69]. A segmentação consiste num processo de divisão do mercado em grupos de clientes ou consumidores com necessidades similares. Quanto maior a similaridade das necessidades dos clientes menor são os grupos formados. Em contrapartida aumenta a probabilidade dos clientes do mesmo grupo (segmento de mercado) virem a adquirir o mesmo produto por este ir de encontro às suas necessidades [70].

Nos estudos levados a cabo sobre a segmentação de clientes, surgem 4 modelos base de segmentação [71]: (i) segmentação geográfica onde o mercado é dividido por região geográfica, densidade populacional e clima; (ii) segmentação demográfica em que a segmentação dos clientes é feita pela idade, género e número de elementos do agregado familiar, etc; (iii) segmentação psicográfica, onde a segmentação é efetuada recorrendo a variáveis relacionadas com o estilo de vida; (iv) segmentação comportamental que considera o comportamento do cliente em relação ao produto, e tem em conta fatores como, a lealdade a marca, compras por impulso ou ocasionais.

Na literatura predominam estudos que consideram os fatores demográficos e psicográficos para a segmentação de mercado, [72], [73], [74]. Goyat [75] é da opinião que diferentes formas de segmentação têm diferentes níveis de importância de acordo com a área de negócio, no entanto, existem outras variáveis de grande importância como o preço, as condições e tendências do mercado que devem ser consideradas.

Estudos empíricos demonstram que diferentes fatores podem influenciar a segmentação de mercado, de acordo com a área de negócio. Amandeep Singh [76] revela na sua investigação, recorrendo a dados de 500 clientes que os fatores demográficos como a idade e as habilitações literárias são indiferentes na escolha de produtos de higiene pessoal, mas que o género pode afetar na escolha destes produtos. Wells et al. [77], verificaram que apenas alguns fatores demográficos como, o tamanho da habitação, o nível de habilitações literárias e o rendimento influenciam em grande medida a preferência na escolha de uma determinada marca de um produto. Higgs e Ringer [78] sugerem no seu estudo sobre formas de segmentação de clientes, a necessidade de traçar o perfil do cliente de forma progressiva recorrendo à captura incremental de dados em diferentes pontos de interação online, como em sites e dispositivos eletrónicos, recolhendo informação do comportamento online como, a visita aos sites, envolvimento com os conteúdos e exposição à publicidade. Ansari e Riasi [79], defendem que para melhorar a qualidade dos produtos e serviços e aumentar a competitividade, as empresas devem procurar identificar as necessidades dos seus clientes alvo e desenvolver uma estratégia de marketing orientada à satisfação das suas necessidades. Assim, os serviços não devem ser fornecidos de igual modo a todos os clientes. A segmentação e a gestão da relação com os clientes (*customer relationship management ou CRM*) devem ser considerados os principais determinantes da viabilidade do negócio.

Na área do turismo, Donlicar [80] é da opinião que qualquer indústria a operar no ramo do turismo pode ser segmenta, e aponta como exemplos a hotelaria, os serviços de viagens e turismo, a restauração e os serviços de atração turísticos. Para a autora, definir uma estratégia de segmentação na indústria do turismo significa satisfazer necessidades específicas para um determinado tipo de turistas (segmento de mercado) em vez de tentar servir todo o mercado turístico.

Existe uma grande variedade de técnicas de segmentação onde a abordagem pode ser conceptual, também denominado de *a priori segmentation,* onde os turistas são segmentados tendo por base características pré-definidas, [81], [82]. A abordagem também pode ser multidimensional, orientada aos dados, onde um conjunto de características são obtidos a partir de dados empíricos [83], [84], [85]. Independentemente da abordagem de segmentação adotada, [80] sugere que deve ser solicitado aos gestores de negócio que avaliem não só a técnica utilizada bem como os segmentos propostos. A abordagem de segmentação orientada aos dados foi introduzida no marketing por Haley [86]. O autor critica a abordagem conceptual por ser meramente descritiva e propõe utilizar técnicas estatísticas que agrupam os objetos em *clusters* de acordo com critérios de similaridade. Após a publicação de Harley [86], alguns investigadores na área do turismo adotaram a mesma abordagem, como pode ser evidenciado nos estudos [87], [88], [89], [90].

Existem muitas publicações sobre a segmentação recorrendo a uma abordagem orientada aos dados, o que conduziu à publicação de algumas revisões de literatura sobre o tema. Donlicar [91] efetuou uma revisão de literatura analisando as metodologias aplicadas na abordagem orientada aos dados para a segmentação de clientes no turismo e concluiu que apenas uma pequena parte dos algoritmos são utilizados e que grande parte dos investigadores, na área do turismo optam pelos algoritmos de *cluster* hierárquicos e pelo algoritmo iterativo *k-means.*

2.2.6. Análise de clusters para a segmentação de clientes

Desde a publicação do artigo de Sminth [67], a segmentação de clientes passou a constituir uma importante ferramenta, tanto para a investigação académica, como para a aplicação prática empresarial.

A primeira aplicação de análise de *cluster* em Marketing foi utilizada para a segmentação de clientes [92]. Nos anos 60 e 70 surgem as primeiras publicações onde a análise de *cluster*, ou simplesmente *Clustering*, foi utilizada para resolver problemas de Marketing, [93], [94]. Estes autores utilizaram os algoritmos *K-Means* e Cluster Hierárquico para a segmentação de clientes. Para Punj e Stewart [92] a análise de *cluster* consiste num método estatístico de classificação. Ao contrário de outros métodos estatísticos de classificação como é o caso da análise discriminante, não define quaisquer pressupostos relativamente a diferenças na população em estudo.

Clustering é utilizado para a Extração de Conhecimento de dados multidimensionais, na literatura muitas vezes o termo é referido como *pattern recognition* ou aprendizagem não supervisionada [95]. De acordo com Han [96], a análise de *cluster*, consiste no processo de dividir um conjunto de objetos ou observações de objetos em subconjuntos. Cada subconjunto forma um *cluster* em que cada objeto partilha características semelhantes. Neste contexto não se identifica um método estatístico particular, ou um modelo de *clustering*, mas sim um conjunto de métodos que permitem classificar casos em grupos. Sendo este método um processo exploratório de dados confere a vantagem de poder revelar grupos de objetos nos dados que anteriormente não eram conhecidos.

De referir ainda que a aplicação do algoritmo não necessita da validação de pressupostos relativamente à distribuição dos dados. A escolha de um método depende sobretudo do número de casos em análise e do tipo de variáveis que serão usadas para formar os *clusters*. Diferentes métodos de *clustering* podem gerar *clusters* diferentes, quando aplicados sobre o mesmo conjunto de dados.

Han [96], considera que o *core* dos algoritmos de *cluster* consiste em medir a distância entre os *clusters*.

Aldenderfer e Blashfield [97] consideram que o conceito de similaridade e a sua medição são importantes para compreender o desempenho de qualquer procedimento de *clustering*. Para estes autores, estimar quantitativamente a similaridade de objetos recorrendo ao conceito de métrica é determinante. Essa abordagem de similaridade assume que o grau de dissimilaridade entre casos pode ser representado como uma distância entre pontos projetados no espaço multivariado. A escolha da medida de distância é de extrema importância, influenciando o resultado da análise. É usual ser utilizada a distância euclidiana, mas dependendo dos dados em análise e das questões de investigação podem ser utilizadas outras medidas de dissimilaridade como os coeficientes

de correlação, descritos como medidas de forma por serem insensíveis às diferenças de magnitude das variáveis envolvidas no cálculo dos coeficientes, [97], [95]. Por exemplo, muitas vezes são utilizadas as correlações de *Pearson* mais sensível a *outliers* ou de *Spearman* quando se pretende mitigar a presença de *outliers*. Dois objetos são considerados similares se as suas variáveis estiverem altamente correlacionadas mesmo estando distantes em termos de distância euclidiana. A dissimilaridade baseada na correlação dos objetos é utilizada quando se pretende formar *clusters* de objetos com o mesmo perfil independentemente da sua magnitude. Em marketing pode ser aplicada para agrupar clientes com o mesmo perfil em termos de preferência de artigos comprados, independentemente do volume de compras [95].

A medição da distância entre os objetos está intimamente relacionada com a escala em que as medições são efetuadas. Por esse motivo, muito frequentemente, as variáveis são standardizadas antes de medir a dissimilaridade dos objetos. A standardização é particularmente recomendada quando as variáveis que representam um objeto estão em diferentes escalas métricas, caso contrário a similaridade é severamente afetada. O objetivo consiste em tornar as variáveis comparáveis aplicando uma transformação na escala, standardizando as variáveis para variância unitária e media zero, [95]. Donlicar [91] questiona a aplicação de transformações nos dados, como a standardização ou a redução de dimensionalidade, justificando que essa prática pode eliminar parte da estrutura de dados que a análise de *cluster* procura espelhar.

Existe uma grande variedade de técnicas que podem ser utilizadas para dividir os dados em grupos homogéneos. Aldenderfer e Blashfield [97] dividem os algoritmos de *cluster* em 5 principais famílias: (1) hierárquicos aglomerativos, (2) hierárquicos divisivos, (3) não hierárquicos ou de particionamento iterativo, (4) baseados na densidade e (5) variantes da análise fatorial como, por exemplo, a análise de componentes principais. Os algoritmos hierárquicos e iterativos são os mais utilizados com um grande número de publicações académicos. Nos algoritmos hierárquicos aglomerativos são utilizados diferentes métodos de *linkage* (*single linkage*, *complete linkage*, *average linkage*, e o método de *Ward*) para agrupar os objetos de acordo com a sua (di)similaridade. A ideia base consiste em agrupar os casos passo-a-passo iniciando com cada objeto constituindo-se individualmente como um grupo. Progressivamente os objetos são agrupados até existir um único grupo denominado, o grupo de topo da hierarquia. Nos algoritmos hierárquicos divisivos, o processo inicia-se com todos os objetos agrupados num único

grupo, em cada iteração um *cluster* é divido num grupo menor e o processo termina quando cada *cluster* for constituído por um único objeto. O resultado da aplicação dos métodos hierárquicos é representado graficamente numa estrutura hierárquica, denominada de dendrograma que representa a sequências de passos do algoritmo e os níveis de agrupamento dos *clusters*. Os algoritmos de *cluster* não hierárquicos ou de particionamento iterativo operam diretamente sobre os dados *raw* (em bruto) e tem a vantagem de poderem lidar com um maior volume de dados quando comparados com os métodos hierárquicos [97].

O algoritmo *K-means* é considerado o mais popular dos algoritmos de aprendizagem não supervisionada para dividir conjuntos de dados em grupos (*clusters*), onde *k* representa o número de grupos predefinido pelo analista. Os grupos são classificados de forma a que os objetos pertencentes a cada grupo sejam o mais similar possível (isto é, alta similaridade dos objetos do mesmo grupo), e que os objetos de *clusters* diferentes são o mais dissimilares possíveis (baixa similaridade dos objetos de grupos diferentes). Cada *cluster* é representado pelo seu centro (ou *centroid*) que corresponde à média dos pontos atribuídos ao *cluster* [98], [99].

A decisão do número de *clusters* que melhor representa uma solução é crucial para o resultado da análise. Embora a discussão de como determinar o número de *cluster* remonte o trabalho de Thorndike [100], nenhuma solução satisfatória para o problema foi encontrada até hoje, apenas são sugeridos um conjunto de heurísticas e índices [91] que são analisados e comparados [101], [102]. Uma solução muita utilizada para determinar o número de *clusters* consiste em utilizar o dendrograma dos *clusters* hierárquicos para identificar graficamente o número de clusters que melhor representa a estrutura. Esta solução também continua a ser subjetiva [95].

Com a crescente utilização e sofisticação das técnicas de *Data Mining*, no setor das viagens e turismo, têm surgido novas abordagens de segmentação de clientes como é o caso do modelo *RFM* (*Recency*, *Frequency* e *Monetary*) [103], [104] e do algoritmo *biclustering* [105]. Hughes [106], propõem a utilização do modelo analítico *RFM* para agrupar os clientes utilizando três variáveis, a compra mais recente (*Recency*), a frequência das compras (*Frequency*) e o valor monetário gasto pelo cliente (*Monetary*). O modelo *RFM* permite extrair às caraterísticas do cliente utilizando um número reduzido de critérios, apenas com três dimensões como atributos para formar os *clusters*, reduzindo a complexidade na construção das análises [107]. Do ponto de vista do comportamento

dos clientes, Schijns e Schroder [108] referem a importância do modelo *RFM* na gestão da relação com os clientes, permitindo analisar o valor do cliente. O custo de reter os clientes é muito superior ao custo de desenvolver ações de Marketing para a aquisição de novos clientes, assim o modelo *RFM* pode ser utilizado nos dados das empresas para encontrar os clientes que criam maior valor para empresa [109], [110]. Alford [111] critica o modelo *RFM*, por não permitir que clientes com diferentes valores pertencem ao mesmo segmento. Apesar de raras as evidências da aplicação do modelo RFM no setor das viagens e turismo, Wong et al. [103] aplicou o modelo para agrupar os viajantes de acordo com o seu valor tendo em vista identificar dados demográficos e padrões nas escolhas dos destinos. Já Shelly et al. [104] utilizaram o modelo *RFM* para compreender o que leva os clientes a tornarem-se sócios de um clube de viagens *all-inclusive*. Os resultados do estudo apontam que a frequência (*frequency*) é o principal preditor dos clientes para estes se virem a tornar sócios do clube quando comparado com as variáveis *recency* e *monetary*.

Donlicar et al. [105] introduziram um novo algoritmo de *clustering* para a segmentação do mercado de viagens e turismo, o algoritmo *biclustering,* anteriormente utilizado por Madeira e Oliveira [112] na análise de dados genéticos. Os autores anunciam que este algoritmo supera as limitações dos algoritmos tradicionais de *clustering*, especificamente no que toca a dimensionalidade dos dados. Este algoritmo consegue lidar com poucas amostras nos dados e um elevado número de variáveis, efetuando em simultâneo a seleção de variáveis e a constituição dos grupos. Desta forma permite identificar nichos de mercado e possibilita reproduzir os resultados obtidos, superando os algoritmos tradicionais paramétricos e de particionamento iterativo que podem conduzir a resultados diferentes em cada execução. Esse facto está relacionado com o carater aleatório desses algoritmos na inicialização dos pontos para a constituição dos grupos. A desvantagem do algoritmo *biclustering*, no contexto da segmentação de mercado, reside no facto dos segmentos serem definidos de forma restritiva, em que todos os elementos do segmento partilham as mesmas atividades e estas caracterizam o segmento. Como consequência, os segmentos resultantes do *biclustering* são muito distintos e de pequena dimensão. Contudo, esta desvantagem pode ser ultrapassada enfraquecendo o nível de restrição dos elementos de um segmento de forma a permitir alguma discordância entre os seus elementos [105].

2.3. Os Sistemas de Suporte à Decisão

Os Sistemas de Suporte à Decisão ou *DSS* estão diretamente relacionados com o conceito de *Business Intelligence* (*BI*). Apesar de *DSS* e *BI* não serem sinónimos, ambos podem ser visualizados como um conceito abrangente, usualmente referidos como um termo "*umbrella*" que inclui tecnologias de informação orientadas para os dados (*data-driven*), como *data warehouses* e *data mining* (*DM*) [113]. Este tipo de *DSS* permite a extração de padrões e fazer previsões, a partir dos dados históricos das organizações.

O modelo proposto neste trabalho pretende implementar uma *DSS* para assistir os gestores de negócio na definição de uma estratégia de negócio e de investimento em marketing digital.

2.3.1. Extração de Informação e Aprendizagem Automática

A área de Extração de Informação (ou KDD: *Knowledge Discovery and Data Mining*) está estritamente relacionada com a Aprendizagem Automática (ou *Machine Learning*). Ambas tentam encontrar padrões e extrair conhecimento recorrendo a algoritmos e técnicas de análise de dados. Apesar da partilha de algoritmos e técnicas, existem diferenças entre as duas áreas que devem ser realçadas.

Resumidamente, a Aprendizagem Automática inclui subáreas relacionadas com algoritmos de procura, algoritmos evolutivos e com a programação de agentes inteligentes, através da aprendizagem por reforço. As subáreas mencionadas têm aplicação em domínios como a robótica, sistemas de conversação e jogos de computador, que não fazem parte do *KDD*. A KDD tem uma aplicação mais comercial e preocupa-se com a resolução de problemas de negócio, através de um rigoroso processo de análise de dados que compreende etapas que vão desde a preparação dos dados até à construção e avaliação de modelos estatísticos de aprendizagem, com foco especial na explicação causal do histórico.

Tanto em Aprendizagem Automática como em Extração de Informação, o processo de aprendizagem divide-se em duas categorias: supervisionada e não supervisionada. A categoria de aprendizagem não supervisionada é conhecida como tal porque não é possível supervisionar a análise, dado que não existe uma variável de resposta. A aprendizagem supervisionada é conhecida por aprender com base em exemplos, onde uma função é aprendida a partir de exemplos. No presente trabalho foram abordadas as

duas categorias de aprendizagem. A não supervisionada é utilizada para a identificação de *cluster* de clientes e para a extração de informação de texto, com foco na modelação de tópicos relevantes nos comentários dos clientes. A aprendizagem supervisionada é aplicada para a análise preditiva de séries temporais, com o objetivo de compreender o impacto das campanhas *adwords* na procura de serviços. Quando a resposta que queremos prever é quantitativa denomina-se a tarefa a resolver como sendo um problema de regressão. Para respostas qualitativas o problema é considerado como um problema de classificação. Contudo, essa distinção não é clara, existem algoritmos como é o caso da regressão logística que são utilizados para obter respostas qualitativas mas, como o resultado é uma probabilidade, pode ser encarado tanto como um problema de regressão como de classificação. Outros métodos como *K-nearest neighbors* e o *boosting* podem ser utilizados tanto para respostas quantitativas como qualitativas [114].

2.3.2. Extração de Informação e *Big Data*

O termo *Big Data* surge originalmente para descrever o aumento no volume de dados, circunstância que tornou incompatível capturar, armazenar ou analisar nas bases de dados tradicionais.

Ao longo dos anos, a definição tem vindo a adaptar-se aos avanços tecnológicos. Em 2001, *Meta Group*, agora designado de *Gartner*, definiu *Big Data* como sendo algo tridimensional para representar o aumento do volume de dados, a velocidade no seu processamento e a variedade de dados e fontes existentes. Mais tarde, em 2012, *Gartner* atualizou a definição para grande-volume, velocidade e variedade [115], [116]. Estudos mais recentes consideram que a definição dos 3Vs (volume, velocidade e variedade) não é suficiente para explicar o conceito de *Big Data* [117].

De todos os Vs apresentados nas definições de *Big Data*, talvez o Valor seja aquele que mais contribuiu para atrair a atenção de empresas de vários setores de atividade para a demanda dos dados. O valor está na análise que é efetuada aos dados e como estes são transformados em informação e em conhecimento.

Segundo Cordeiro [118], o conceito de *Big Data* não assenta apenas na existência de um conjunto de dados de grande dimensão, fala-se de *Big Data* para referir as tecnologias e processos implicados na recolha, armazenamento, tratamento e análise de grande

volume de dados, envolvendo processos de Extração de Informação dos dados (*Data Mining*), com vista à criação de nova informação. O autor refere ainda a importância dos processos de tratamento do *Big Data*, realizados através de modelos de análise de dados, construídos por algoritmos, normalmente assentes em técnicas de Aprendizagem Automática e que servem para extrair um resultado que servirá o propósito de auxiliar nos processos de tomada de decisões.

Tsai [117] revela que estudos realizados sobre as plataformas tradicionais de análise de dados indicam que estes se centram no desenho e desenvolvimento de formas eficientes de extrair conhecimento dos dados. Ao entrarmos na era do *Big Data*, a maior parte destes sistemas tradicionais não estão preparados para lidar com o grande volume de dados, pelo que o seu desenho e implementação são importantes fatores a ter em conta no processo de análise de dados. Os mesmos autores apontam ainda várias soluções comerciais que foram apresentadas para a análise de dados em ambientes *Big Data*. Estas podem ser divididas, de acordo com as seguintes características:

Processamento: Hadoop, Nvidia CUDA

Armazenamento: HDFS (Hadoop), Titan

Analítico: Mahout, MLPACK

Mais recentemente, a *HortonWorks* e a *Cloudera*, duas das principais empresas fornecedoras de soluções para *Big Data*, fundiram-se e apresentaram a sua plataforma de código aberto, assente na infraestrutura de sistemas distribuídos, o *HDFS* (vide Figura 4). Com a chegada destas novas plataformas, a utilização de algoritmos de Extração de Informação para *Big Data* passou a ser uma realidade, podendo ser encontrados em atividades do nosso quotidiano.

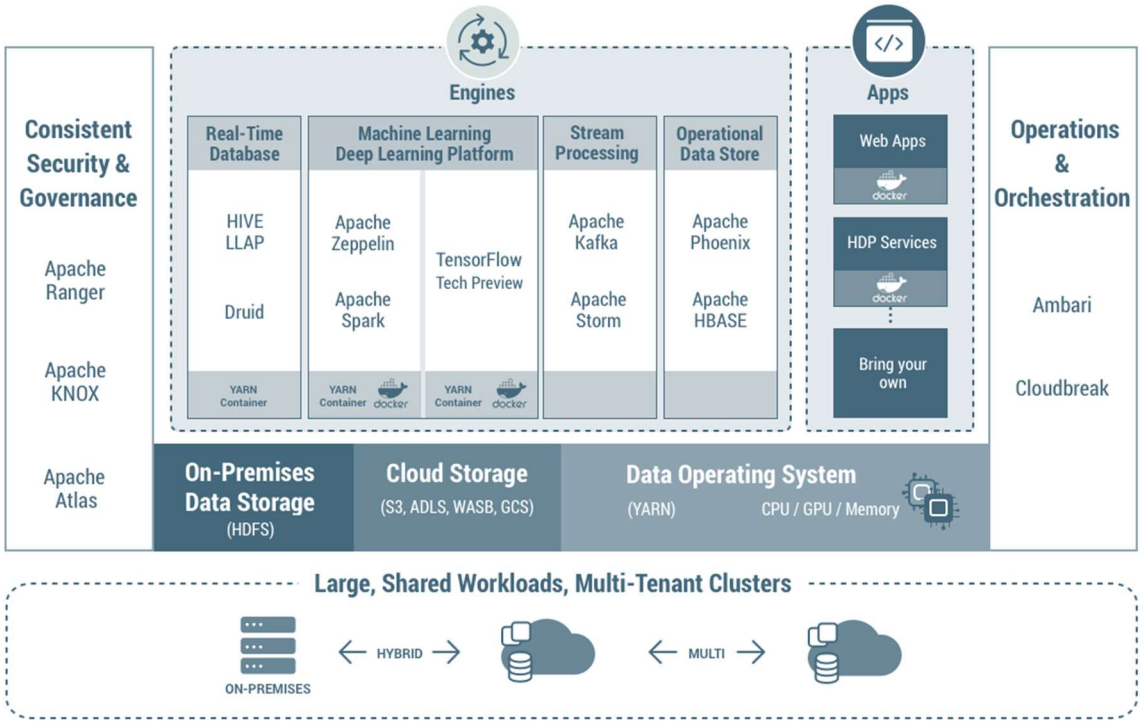


Figura 4 Plataforma de dados da Hortonworks/Cloudera. Fonte: cloudera.com/products/hdp.html

2.3.3. Proteção e privacidade dos dados

O desenvolvimento das técnicas de Extração de Informação, capazes de fazer inferências sobre categorias, passaram a constituir ferramentas poderosas para a análise de dados. Assim, não será de estranhar os constantes debates em diferentes círculos sobre a privacidade e a proteção contra a intromissão nos dados pessoais. A forma como as tecnologias estão a ser aplicadas e os dados são recolhidos levantam questões do ponto de vista sociológico, que devem ser acauteladas, para que seja garantida a proteção do cidadão [1].

A privacidade diferencia-se da segurança, os desafios que as empresas enfrentam relativamente à segurança dos dados relacionam-se com o armazenamento dos dados de forma segura, com a segurança na sua transferência e com a forma de prevenir como os dados pessoais podem ser acedidos por terceiros. A privacidade representa um papel diferente e comporta outras preocupações a ter em conta, por estar relacionada com a capacidade dos sistemas recuperarem ou inferirem informação que anteriormente era anónima, através da aplicação de algoritmos de Extração de Informação.

Com o aumento da utilização das técnicas de Extração de Informação, a informação privada passou a estar exposta após a aplicação do processo de análise de dados. Por este motivo, os dados devem ser utilizados e protegidos de forma criteriosa. O fator crítico reside em saber como utilizar, o que utilizar e quando utilizar, os dados recolhidos.

É importante referir o regulamento de proteção de dados adotado por todos os Estados Membros da União Europeia, em maio de 2018. O regulamento dá enfase à transparência, introduzindo novos elementos em que o titular dos dados deve ser informado do processamento dos seus dados, podendo solicitar os mesmos em formato eletrónico para fornecer a uma terceira entidade. O direito de eliminação dos dados também foi reforçado com a introdução do direito-de-ser-esquecido, o que obriga a um cuidado suplementar na forma como as empresas transmitem informação pessoal a terceiros. Em particular, este novo direito obriga a uma empresa que tornou público os dados pessoais de um utilizador, a comunicar a terceiros que o titular dos dados requereu a eliminação de qualquer *link* ou forma de replicação dos seus dados pessoais.

Cordeiro [118], refere o objetivo Europeu relativo à proteção dos dados pessoais como um "*global gold standard* ", assente na conciliação entre as liberdades e direitos fundamentais de pessoas singulares e na livre circulação de dados. Este autor entende ser necessário assegurar que as regras impostas sejam ajustadas à rápida evolução tecnológica e às novas formas de partilha, recolha, tratamento e transmissão de dados pessoais.

2.4. Séries Temporais

Uma série temporal consiste numa observação sequencial ao longo do tempo, podendo apresentar uma variedade de padrões, possíveis de serem decompostos em diferentes categorias de padrões subjacentes com origem temporal.

Se assumirmos uma decomposição aditiva de uma série temporal, podemos escrever $Yt = St + Tt + Rt$, onde Yt representam os dados, St representa a componente sazonal, Tt a tendência e Rt o resíduo, ou seja, representa o componente restante após retirar todos os outros componentes e t representa o período da observação. Em alternativa pode-se considerar uma decomposição multiplicativa, $Yt = St\,.\,Tt\,.\,Rt$. A decomposição multiplicativa de uma série temporal é indicada quando existem variações na componente sazonal e na tendência da série. Opcionalmente, pode ser aplicada uma transformação

nos dados de forma a estabilizar as variações de sazonalidade e tendência e de seguida pode ser aplicada a decomposição aditiva.

No que diz respeito à utilização de séries temporais, na área de Marketing, existe um crescente consenso do valor empírico da aplicação de técnicas orientadas para os dados, onde as observações dos padrões existentes nos dados históricos são consideradas na especificação dos modelos. Dekimpe et al. [119] fazem uma análise detalhada desse tema, referindo que no passado a relutância de seguir uma abordagem orientada para os dados deveu-se à escassez de pessoal com conhecimento adequado, à inexistência de *software* específico e a poucas fontes de dados disponíveis. Como constatado na revisão de literatura de Dekimpe et al. [119] a utilização de séries temporais em Marketing começou gradualmente a surgir na literatura, para: (1) fazer previsões (*forcasting*); (2) determinar a ordem temporal de variáveis através do teste de casualidade de *Granger* [1] e (3) para determinar o impacto das variáveis de marketing ao longo do tempo.

2.4.1. Análise preditiva de Séries Temporais

Na análise preditiva de séries temporais o objetivo é estimar a continuação futura de uma sequência de observações. O método mais simples de análise de séries temporais utiliza apenas a informação da variável a predizer, sem considerar os fatores que possam afetar o comportamento dessa variável. Apesar desse método explorar a sazonalidade e a tendência, ignora a restante informação, como por exemplo, iniciativas de marketing, alteração das condições económicas, as atividades dos competidores, entre outra informação relevante [120].

A especificação do modelo preditivo de séries temporais depende em grande medida dos dados disponíveis e da forma como os modelos são utilizados [120].

O modelo *Exponential smoothing* foi proposto nos finais dos anos 50 [121], [122], [123]. Este método utiliza o peso médio das observações passadas, sendo que nas observações mais antigas, o peso diminui de forma exponencial. Este modelo produz previsões confiáveis para uma grande variedade de séries temporais, o que constitui uma vantagem de grande importância para a aplicação na indústria. Holt [122] e Winters [123]

[1] O teste de causalidade proposto por Granger visa superar as limitações do uso de simples correlações entre variáveis. Essa distinção é de fundamental importância porque correlação não implica por si só causalidade (relação de causa e efeito).

estenderam o método anterior de forma a incorporar a sazonalidade e a tendência dos dados, dando origem ao método sazonal Holt-Winters**.**

O modelo *ARIMA* **(**Auto-Regressive Integrated Moving Average**)** introduziu uma nova abordagem na análise de series temporais, passando a ser um dos modelos mais utilizados, complementando os modelos anteriores. Enquanto o modelo exponencial descreve a tendência e sazonalidade nos dados, o modelo *ARIMA* pretende descrever a autocorrelação existente nos dados [120].

Os modelos apresentados anteriormente permitem a inclusão da informação histórica das séries temporais, mas não permitem a inclusão de variáveis externas que possam explicar a variação dos dados históricos e assim conduzir a previsões mais precisas. Por outro lado, os modelos clássicos de regressão permitem a inclusão de variáveis preditoras relevantes, mas não permitem capturar o dinamismo das séries temporais como é o caso da captura da autocorrelação existente nos dados que o modelo *ARIMA* introduziu. Por vezes, os impactos das variáveis preditoras nos modelos de regressão não se verificam de imediato. Por exemplo, o efeito de uma campanha publicitária pode ter impacto nas vendas apenas após o seu término. As vendas de um mês podem depender das campanhas publicitárias dos meses anteriores.

2.4.2. Modelos de regressão dinâmicos

Um modelo de regressão dinâmico configura uma equação onde uma variável é explicada em função da sua observação em períodos temporais passados e presente e/ou da observação de períodos temporais passados de outras variáveis com ela relacionadas. Desta forma, o efeito da alteração numa variável explicativa não é instantâneo, podendo demorar algum tempo até que se verifique o seu efeito na variável de resposta.

Uma aplicação de modelos dinâmicos na análise de séries temporais em marketing consiste nos modelos VAR [124]. Na literatura, muitos autores apresentam estudos que se baseiam em diferentes modelos VAR (VAR-based persistence models), para aferirem os efeitos ao longo do tempo dos instrumentos de marketing, como por exemplo nas promoções e no lançamento de novos produtos [125], na publicidade [126], e na distribuição [127].

Mais recentemente [128] apontam duas formas de formular um modelo de regressão dinâmico: (1) como um modelo de regressão múltiplo, definido de forma usual, incluindo

diferentes períodos temporais das variáveis preditoras e da variável de resposta, este modelo é designado por modelo *Autoregressive Distributed Lags* (ADL), ou (2) como uma estrutura de polinómios designado por modelo de Transfer Function (TF).

Se considerarmos um exemplo simples de um modelo ADL, que inclui um único período temporal passado da variável dependente, $Yt - 1$ e uma variável explicativa Xt com o correspondente período temporal passado, $Xt - 1$, podemos escrever a seguinte equação ADL:

$$Yt = c + a1.Yt - 1 + b0.Xt - 1 + t$$

Hyndman e Athanasopoulos [120] introduzem o conceito de modelos de regressão dinâmicos para permitir que outras informações sejam incluídas no modelo *ARIMA*. Desta forma é possivel incluir nos modelos *ARIMA* informação de outras variáveis externas que podem explicar algumas das variações históricas e conduzir a previsões mais precisas.

2.4.3. Aprendizagem Automática e análise preditiva de Series Temporais

Na literatura são vários os estudos que propõem a utilização dos algoritmos de aprendizagem automática para a análise preditiva de series temporais, em alternativa aos métodos estatísticos. Por exemplo, a *Amazon* utilizou redes neuronais para prever a venda dos seus produtos [129] alegando bons resultados quando se trata de series temporais longas e homogéneas.

Ahmed et al. [130] concebe a comparação de uma multiplicidade de algoritmos de aprendizagem automática utilizando as séries temporais disponibilizadas na competição M3[2]. Os resultados obtidos estabelecem um ranking de resultados que coloca em primeiro plano as redes neuronais multicamada (*multilayer perceptron*) seguidas de *Bayesian neural network* (*BNN*) e *support vector regression (SVR)* com resultados idênticos, de seguida surge as *Generalized Regression Neural Networks (GRNN)* e *K-nearest neighbor* (*KNN*), por último as árvores de decisão *CART e radial basis functions (RBF)*. Os resultados apresentados mantiveram-se estáveis quando aplicados em diferentes categorias de variáveis, das centenas de series temporais utilizadas no estudo.

[2] A competição M3, organizada por Makridakis e Hibon [176], consiste numa sequela de competições com o objetivo de estabelecer um ranking de modelos de análise preditiva de series temporais

Foi considerado ainda que a aplicação de transformações, na fase de pré processamento nos dados, tem grande importância nos resultados obtidos.

Apesar dos estudos que recorrem aos algoritmos de aprendizagem automática para a análise de séries temporais apresentarem bons resultados, existem poucas evidências empíricas da sua superioridade, quando comparados com os métodos estatísticos tradicionais.

Os estudos existentes apenas comparam as redes neuronais com as técnicas tradicionais de análise de séries temporais. Por exemplo, Sharda e Patil [131], comparam as redes neuronais com o modelo *ARIMA*. Alon et al. [132] comparam as redes neuronais com o *Exponential Smoothing*, *Box–Jenkins* e *ARIMA*, utilizando dados de vendas a retalho. O resultado de todos os estudos indicados são unanimes em considerarem que as redes neuronais apresentam melhores resultados, quando comparados com os métodos tradicionais. Contudo, os estudos apresentados incidem apenas nas redes neuronais, revelando uma lacuna na literatura, no que diz respeito à comparação com os restantes algoritmos de aprendizagem automática. Esta lacuna na literatura foi confirmada recentemente pelos estudos de Makridakis et al. [133] e Papacharalampous [134] que concluem que a maioria dos trabalhos publicados apresentam resultados satisfatórios dos algoritmos de aprendizagem automática sem estabelecerem uma comparação com os métodos estatísticos tradicionais.

No trabalho de Makridakis et al. [133], o autor apresenta provas de que os métodos estatísticos tradicionais apresentam melhores resultados quando comparados com os algoritmos de aprendizagem automática. No recente trabalho de Cerqueira et al. [135], os autores contrariam os resultados de Makridakis et al. [133] alegando que o tamanho das amostras tem implicação na capacidade de generalização dos algoritmos de aprendizagem automática, sendo esse o motivo dos resultados inferiores relativamente aos métodos estátisticos.

Capítulo 3 – Segmentação de Clientes da empresa *YellowFish Transfers*

Ao longo do presente capítulo foi conduzido um processo de Extração de Informação dos dados da empresa *Yellowfish Transfers* onde procurou-se seguir as melhores práticas recorrendo a metodologia standard de desenvolvimento de projetos de Extração de Informação, o CRISP-DM.

O processo descrito ao longo das secções seguintes procura dar resposta as duas primeiras questões de investigação que consistem em encontrar padrões nos dados que permitam identificar que características do serviço são mais valorizados pelos clientes e ainda perceber que tipo de clientes procuram os serviços da empresa.

3.1. Compreensão do negócio

A compreensão do negócio (*Business understanting*) constitui o estágio inicial da metodologia *CRISP-DM*. Nesta fase, o foco reside em compreender os requisitos de negócio da empresa, de forma a formular um problema de Extração de Informação e definir um plano para atingir os objetivos propostos.

Tendo em vista uma melhor compreensão do negócio da empresa em estudo foi realizada uma apresentação aos órgãos decisores da empresa, recorrendo a ferramentas de visualização de dados, no sentido de efetuar uma exploração inicial dos dados fornecidos. A apresentação serviu como forma de comunicar os objetivos gerais do projeto, mas principalmente, para compreender a perspetiva dos decisores da empresa, relativamente às regras de negócio que deram origem aos dados em análise.

3.1.1. Sistema de reservas online da *YellowFish Tranfers*

A *Yellowfish Transfers* é uma empresa do grupo *Yellowfish* fundada em 2010, inicialmente como uma agência de viagens que rapidamente cresceu passando a fornecer diferentes serviços de viagens e turismo na região do Algarve.

Para além do serviço de transfer a empresa também fornece serviços de guia turístico e criou a empresa *Yellowfish Adventures* que oferece experiências em veículos fora de estrada, utilizando moto quatro e *buggys*.

A empresa está sediada em Albufeira e possui um *website* (Figura 5) onde são efetuadas as reservas dos serviços de transfer de passageiros, entre uma origem e um destino e ainda um serviço de transfer exclusivo para campos de golf.

O objetivo principal é fornecer um serviço privado de transporte porta-a-porta de alta qualidade, tendo em vista satisfazer as necessidades individuais dos turistas e viajantes em negócio, não familiarizados com os sistemas de transporte locais. Para atingir esse objetivo o perfil do motorista desempenha um papel importante devendo este ser fluente na língua inglesa, ter um bom conhecimento da região e ser proativo na prestação de informações sobre a mesma. A maior parte dos serviços são de transfers de turistas do aeroporto de Faro para um hotel na região do Algarve, existindo alguns serviços ocasionais em Lisboa, Alentejo e no sul de Espanha.

A atividade da empresa depende em grande medida da sazonalidade do turismo na região do Algarve que pode ter diferentes causas, como as estações do ano, período de férias, a conjuntura económica ou a periodicidade de alguns voos no aeroporto de Faro.

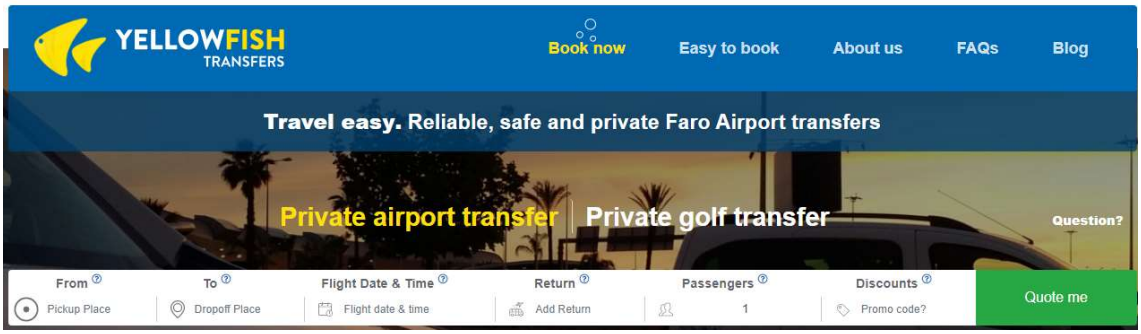


Figura 5 web site de reservas da empresa www.yellowfishtransfers.com

Existe um conjunto de empresas afiliadas que vendem os serviços da *YFT*, onde através de *cookies* identificam a origem dos clientes que chegam ao *website* e pagam uma comissão, por cada reserva que teve origem numa empresa afiliada.

As reservas também podem ser efetuadas por agentes de viagens que acedem diretamente ao sistema de *backoffice* para efetuar as reservas, neste caso, o cliente final "pertence" ao agente e não à *YFT*. Quando a *YTF* não tem capacidade de satisfazer todas as reservas, contrata serviços de fornecedores terceiros.

A uma reserva pode estar associado os seguintes tipos de serviços: (i) serviço de chegada, em que o cliente é transportado do aeroporto de Faro para o hotel de destino; (ii) serviço de partida, onde o transporte é feito no sentido inverso, do hotel para o aeroporto de Faro; (iii) serviço de *golf* que inclui o serviço de ida e volta.

Se numa reserva o cliente incluir o serviço de chegada e partida, esta fica registada com um dos possíveis códigos: (i) *RAL* ou *Return Airport Local* em que são efetuados dois serviços, um do aeroporto para um local e um segundo serviço no sentido inverso; (ii) *RLA* ou *Return Local Airport* em que o primeiro serviço parte de um local e tem como destino o aeroporto e o segundo serviço inicia no aeroporto para um local (iii) *RLL* ou *Return Local Local* em que o serviço de ida e volta é efetuado entre dois locais; (iv) *GOF* que identifica o serviço de ida e volta de golf.

Quando uma reserva inclui apenas o serviço num sentido, fica identificada com os códigos *OAL* ou *One-Way Airport Local, OLA* ou *One-Way Local Airport* e *OLL ou One-Way Local*.

3.2. Compreensão dos dados

A compreensão dos dados (*data understanding*) constitui uma fase critica da metodologia *CRISP-DM*, determinante para o sucesso e para antecipar o surgimento de problemas inesperados nos passos subsequentes. A compreensão dos dados envolve uma recolha inicial dos dados, a sua exploração recorrendo a técnicas de visualização e análise, permitindo aferir acerca da sua qualidade e a identificação de grupos de atributos relevantes para responder às questões de investigação. Muitas vezes os dados que as empresas mantêm nas bases de dados não estão relacionados com o problema de negócio que se pretende resolver. Autores como Provost e Fawcett [136], acrescentam que a exploração dos dados é de extrema relevância, já que permite não só avaliar se existe correspondência com o problema de negócio que se pretende solucionar, bem como encontrar erros nos dados.

A empresa YFT, disponibilizou um conjunto de ficheiros de dados da sua operação diária, extraídos da base de dados do sistema de reservas online. De forma a garantir a proteção dos dados dos seus clientes e motoristas, não foi disponibilizada qualquer informação que permitisse identificá-los de forma direta ou indireta, recorrendo a aplicação das técnicas de Extração de Informação durante o processo de análise de dados.

3.2.1 Recolha inicial de dados

O acesso direto ao sistema de registo das transações operacionais (*OLTP*) não foi disponibilizado, portanto, os dados foram analisados fora do sistema de origem. Os dados que a empresa *YellowFish Transfer* disponibilizou para análise, foram extraídos do sistema *OLTP* de reservas online para ficheiros de texto no formato csv (*Comma-separated values*). Foram fornecidos ficheiros de dados referente aos veículos utilizados nos transfers *(veiculos.csv* e *trasportes.csv)*, os motoristas (*motoristas.csv*), o pais de origem (*paises.csv*), os meios de pagamento utilizados (*payments.csv*), os locais (*local.csv*) onde são efetuados os serviços de transporte, o questionário de satisfação dos clientes (*feedback.csv*), o tipo e quantidade de bagagem (*luggage.csv*) transportado, a identificação das reservas canceladas (*cancelbook.csv*) e do motivo do seu cancelamento e o detalhe dos serviços associados a uma reserva (*manifesto.csv*).

Para melhor compreender as relações de integridade dos dados fornecidos, estes foram carregados num sistema de base de dados relacional, *RDBMS.* Este tipo de sistemas foi proposto por Codd [137] através de um modelo relacional onde a informação é armazenada em estruturas tabulares relacionadas entre si através dos seus atributos e recorrendo a álgebra relacional como linguagem formal para a consulta dos dados. Foi utilizada a base de dados relacional *Microsoft SQL Server* como repositório de dados e recorreu-se a linguagem *Structured Query Language* (*SQL*) para a consulta e exploração inicial dos dados.

Na Figura 6 encontra-se representado o modelo relacional onde cada ficheiro de dados representa uma entidade e as suas relações. A entidade *manifesto* constitui a fonte principal de dados e contém a informação das reservas (*book_id*) e o detalhe dos serviços associados a uma reserva. Para identificar univocamente uma reserva houve a necessidade de criar a entidade *Booking* com informação extraída do ficheiro *manifesto.csv.* A entidade *Booking* permite relacionar uma reserva com os dados recebidos nos ficheiros *cancelbook.csv*, *feedback.csv* e *luggage.csv*.

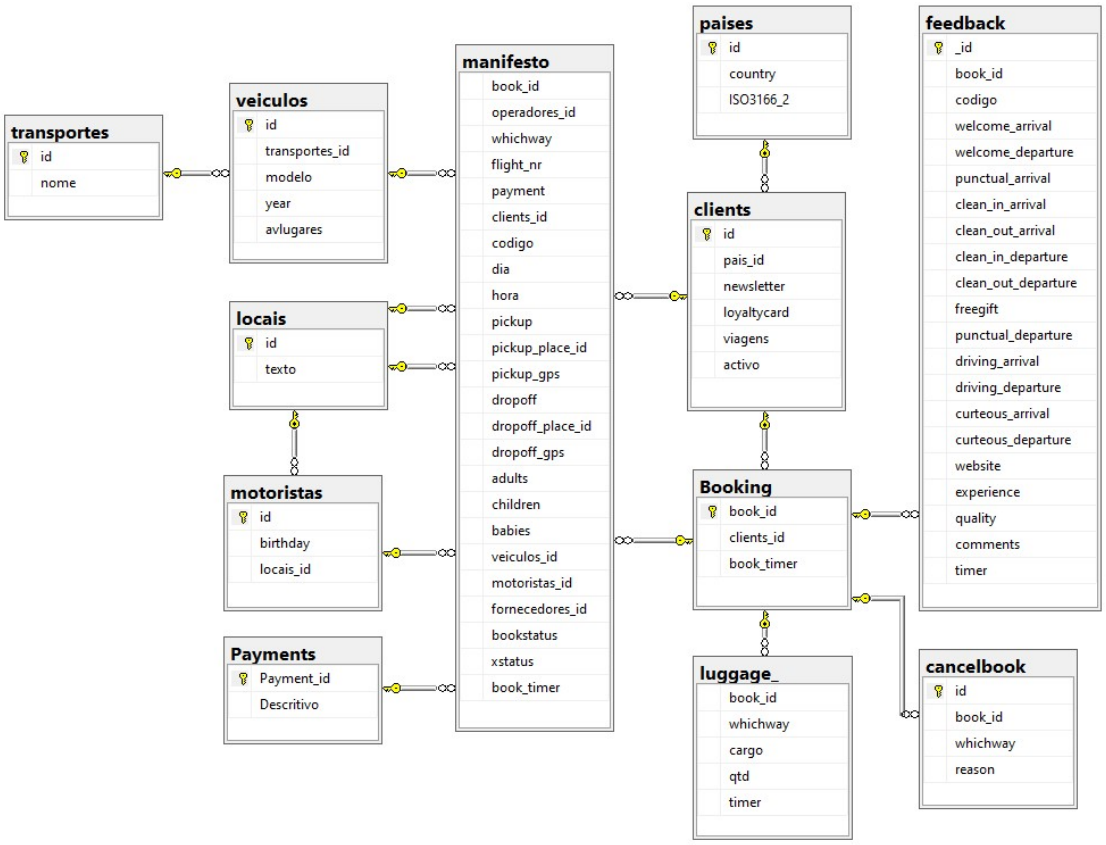


Figura 6 Modelo relacional criado a partir dos dados extraídos do sistema de reservas online

3.2.2 Análise descritiva dos dados

A análise descritiva dos dados constitui uma das etapas da compreensão dos dados identificada na metodologia *CRISP-DM*. onde o objetivo é detetar falhas relacionadas com a qualidade e a relevância dos atributos. Esta constitui uma etapa determinante para avaliar se os dados recolhidos reúnem toda a informação necessária, tendo em vista conferir uma resposta ao problema em análise e prosseguir com o projeto de Extração de Informação.

Nesta etapa foi identificado o método utilizado para a captura dos dados, foram analisados os formatos e a dimensão dos dados (número de observações e atributos), e efetuada uma análise preliminar dos mesmos através das estatísticas descritivas de forma a compreender a sua distribuição.

Conforme descrito no ponto anterior, os dados utilizados no projeto foram disponibilizados através de um conjunto de ficheiros com o histórico das reservas,

extraídos do Sistema de reservas online. A conjugação dos dados dos diferentes ficheiros foi efetuada através de atributos chave representados por códigos idênticos como o *book_id, client_id* ou *o pais_id*.

O processo de carregamento dos ficheiros em formato *csv* para uma base de dados relacional permitiu identificar o formato dos atributos e validar a integridade relacional dos dados. Foram necessárias algumas interações com o gestor da empresa no sentido de avaliarem situações relacionadas com a integridade dos dados.

Um caso que pode ser exemplificado prende-se com a existência de reservas com o código de veículo 0 (*veiculo_id*) sem correspondência no ficheiro de veículos. A explicação para essa ocorrência nos dados tem significado funcional e deve ser interpretado como uma situação em que o serviço de transporte foi prestado por um fornecedor logo os atributos *veiculo_id* e *motorista_id* são valorizados a zero e o *fornecedor_id* tem um valor diferente de zero. Outro caso que pode ser exemplificado são os clientes que solicitaram a anonimização dos dados, no âmbito do regime geral de proteção dos dados (*RGPD*) que ficam associados a um identificador de país (*pais_id*) com valor 0, sem correspondência no ficheiro de países.

Para efeitos de validação da informação recolhida foi efetuado o estudo preliminar das variáveis observadas em cada uma das fontes de dados, recorrendo às estatísticas descritivas. As estatísticas descritivas demonstram que de uma forma geral os dados apresentam uma boa qualidade para prosseguir com a análise, como evidenciado na Tabela 14 do Apêndice A onde são apresentadas as estatísticas descritivos dos dados do ficheiro *feedback.csv*.

3.2.3 Exploração dos dados

Muitas vezes os dados que as empresas mantêm nas bases de dados não estão relacionados com o problema de negócio atual. A exploração dos dados permite avaliar se existe correspondência com o problema que se pretende solucionar [136].

A exploração dos dados integra a fase da compreensão dos dados na metodologia *CRISP-DM* e tem como objetivo a utilização de técnicas de Extração de Informação para a pesquisas e visualização de dados. Pretende-se nesta fase do estudo identificar padrões iniciais e características que permitam identificar subconjuntos de atributos com possível interesse para a análise.

As técnicas de visualização de dados constituem uma poderosa ferramenta para a exploração dos dados, permitindo a identificação de atributos relevantes e a deteção de outliers. Fornecem ainda pistas para a escolha de algoritmos de aprendizagem, mais adequados e ainda permitem uma melhor compreensão dos dados [138].

Foi realizada uma análise exploratória dos dados históricos, recorrendo a ferramenta *Oracle Data Visualization Desktop* (*DVD*), no sentido de investigar se nos dados das reservas, dos serviços e nos questionários de satisfação dos clientes, existem atributos que poderão vir a ser utilizados para extrair conhecimento relativamente ao perfil dos clientes e o seu nível de satisfação com o serviço prestado.

A Figura 7 apresenta uma análise anual das reservas dos clientes, onde foi possível identificar um padrão que vai de encontro com a sazonalidade do turismo na região do Algarve. É notável o aumento do número de reservas desde o início da atividade da empresa em 2012, demonstrando um crescimento notável no volume de negócio da empresa. Os dados demonstram que 2017 foi o ano com o maior número de reservas. Esse facto pode estar relacionado com a divulgação internacional da região do Algarve. De acordo com a informação publicada na página do turismo do Algarve [139], desde 2016 que personalidades de todo o mundo e profissionais de turismo e viagens vinham elegendo o Algarve como o melhor destino turístico de Portugal, da Europa e até do globo, em diversas categorias. Em 2017 a região do Algarve venceu vários prémios turísticos, entre os quais o de melhor destino de praia da Europa nos *World Travel Awards*.

Para confirmar que a procura dos serviços é influenciada pela sazonalidade, foi efetuada a visualização mensal das reservas na Figura 8, onde é evidenciada a sazonalidade da atividade. Na realidade, a procura dos serviços da empresa começa a crescer no mês de fevereiro, atingindo o seu pico nos meses de verão e decrescendo nos meses outono e inverno.

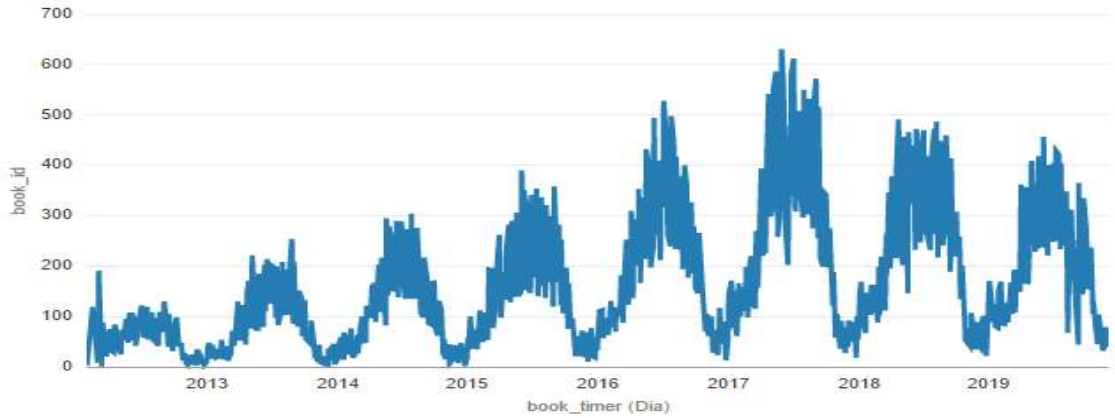


Figura 7 Reservas anuais

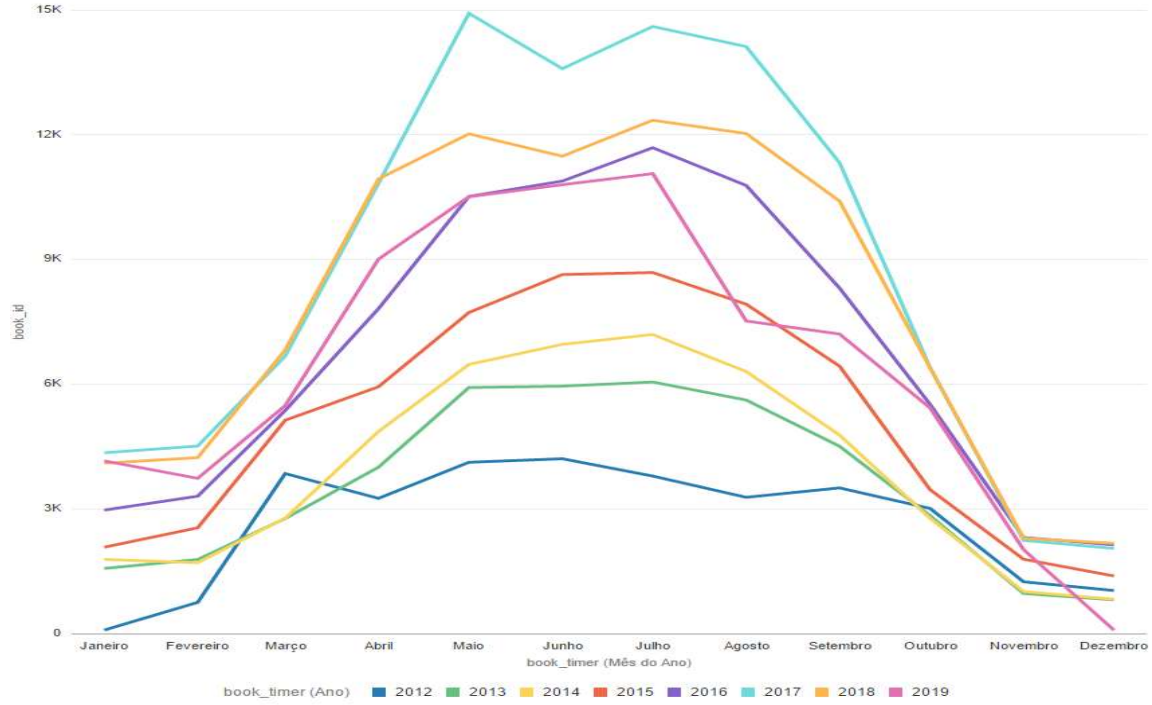


Figura 8 Reservas Mensais

A Figura 9. demonstra que as reservas diárias dos serviços de chegada e partida estão igualmente distribuídas. Como referido anteriormente um cliente pode contratar numa reserva mais do que um tipo serviço. Na Figura 10 é possível identificar que a maior parte das reservas estão associadas a serviços *RAL, cerca de 75,45%, seguidos dos serviços OAL e OLA* representando 11,38% e 11%, o serviço de *GOLF* apenas representam *0,11%*

das reservas. A visualização da distribuição das percentagens pode ser consultada na Figura 48 do Apêndice B.

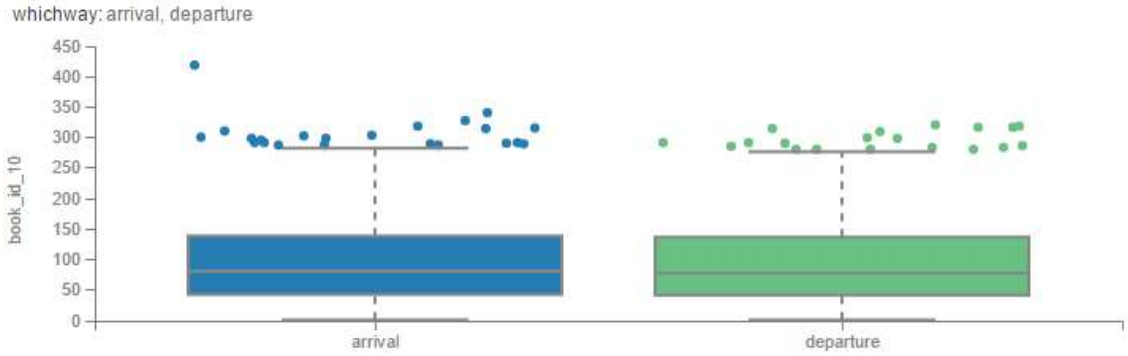


Figura 9 Reservas diárias por Tipos de Serviços

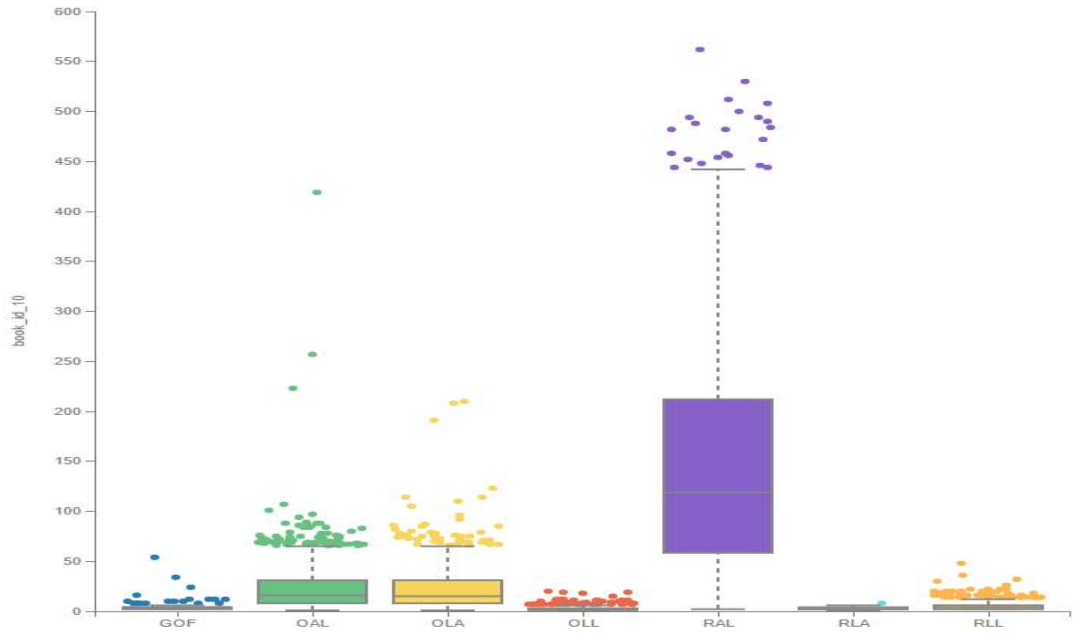


Figura 10 Reservas diárias por Código de Serviços

Ao visualizar os locais de origem e destino do serviço *GOF,* na Figura 11, constata-se que todos os serviços deste tipo têm início nas localidades de Carvoeiro, Vilamoura e Albufeira, sendo as duas últimas localidades as que concentram o maior número de serviços.

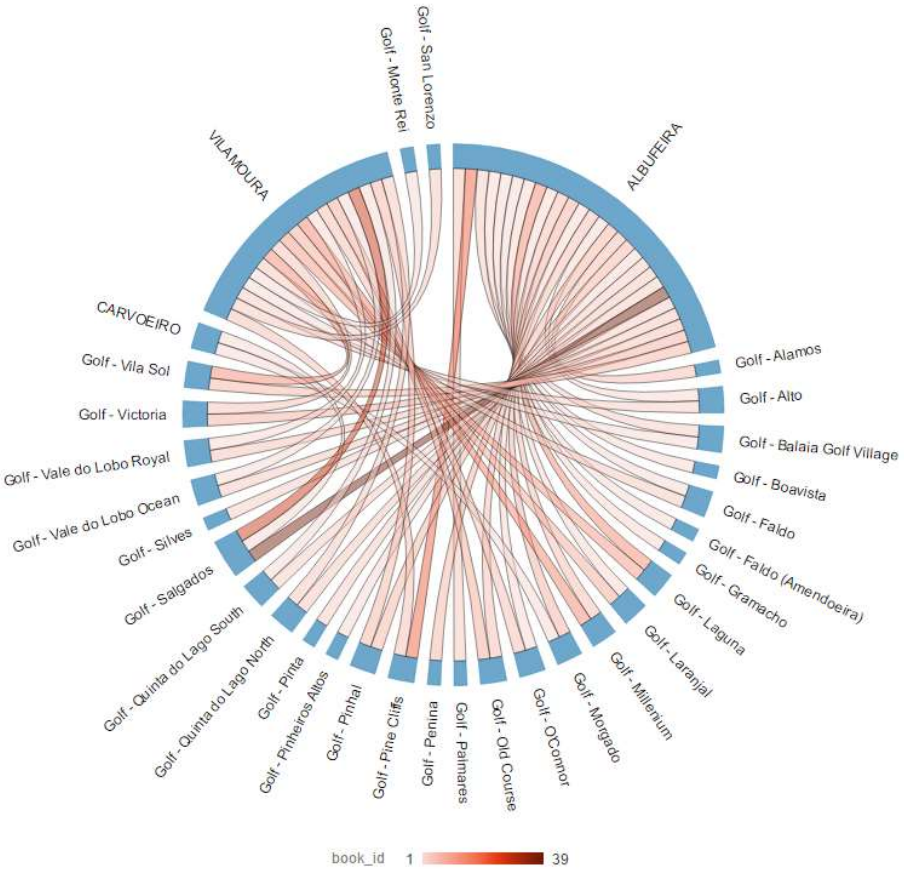


Figura 11 Locais do serviço de Golf

Baseando-se na contagem do número de reservas (*book_id*) por pais de origem dos clientes, foram criadas as visualizações apresentas nas Figura 12 e Figura 13 para compreender a origem dos clientes. A Figura 12 representa o mapa onde os países com a cor mais intensa representam aqueles com maior quantidade de reservas.

A Figura 13 apresenta a percentagem de reservas por pais de origem e mostra que 65,5% das reservas são efetuadas por clientes do Reino Unido seguido da Irlanda com 23,98%. Portugal representa o terceiro país com 2,71% das reservas seguido dos clientes que optaram por não fornecer informações sobre a sua origem, identificados com o código *unknown*.

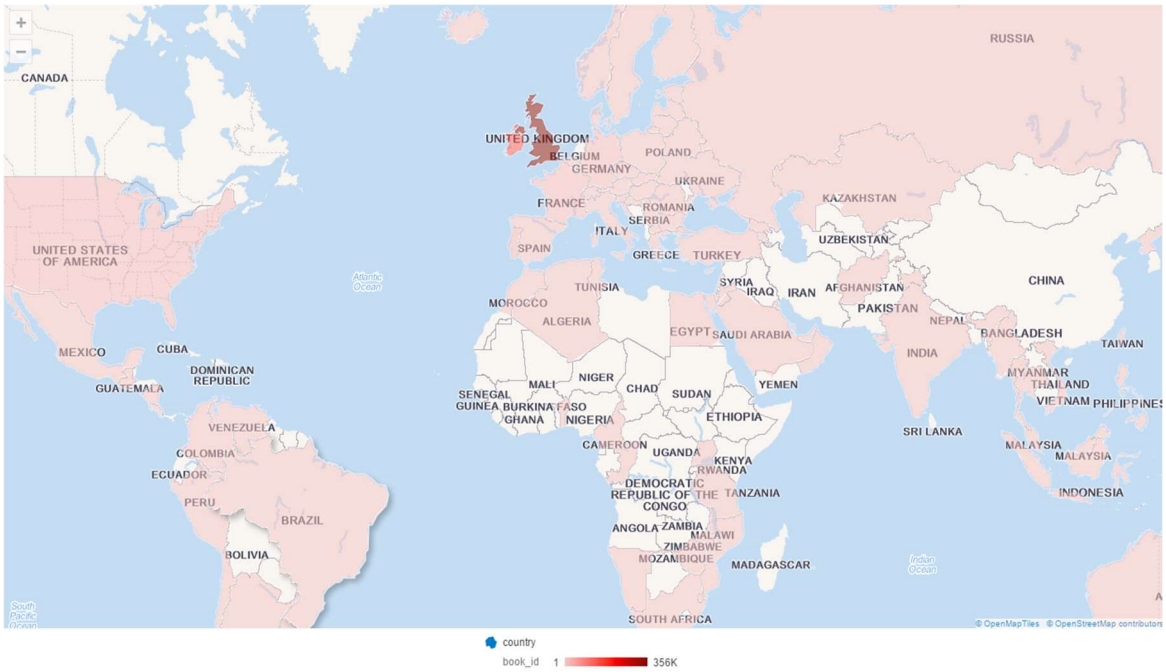


Figura 12 Países de residência dos clientes da empresa

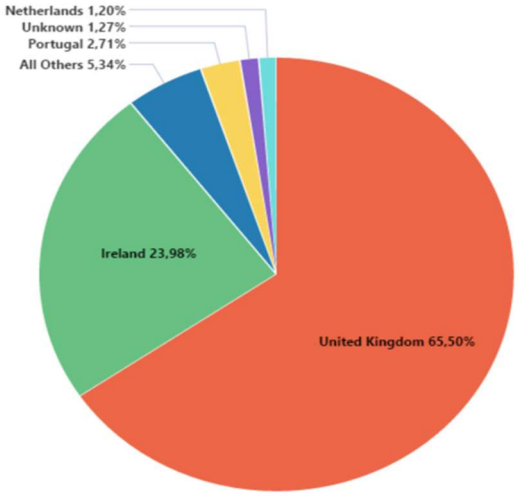


Figura 13 Percentagem de reservas por país de origem

Analisou-se também os cancelamentos das reservas, Figura 14, registados com códigos numéricos, com o seguinte significado: (1) o serviço foi efetuado; (-1) a reserva foi cancelada; e (0) o cliente não cancelou a reserva, mas não apareceu no dia do serviço. Os dados revelam que 9,95% dos clientes cancelam as reservas e que 1,86% de clientes com reservas confirmadas não aparecem no dia do serviço.

O cancelamento de reservas tem grande impacto na gestão da receita e previsão da procura em empresas que prestam serviços relacionados com a indústria do turismo,

[140]. Para mitigar o efeito do cancelamento das reservas as empresas adotam políticas de *overbooking,* com efeitos negativos na sua reputação social.

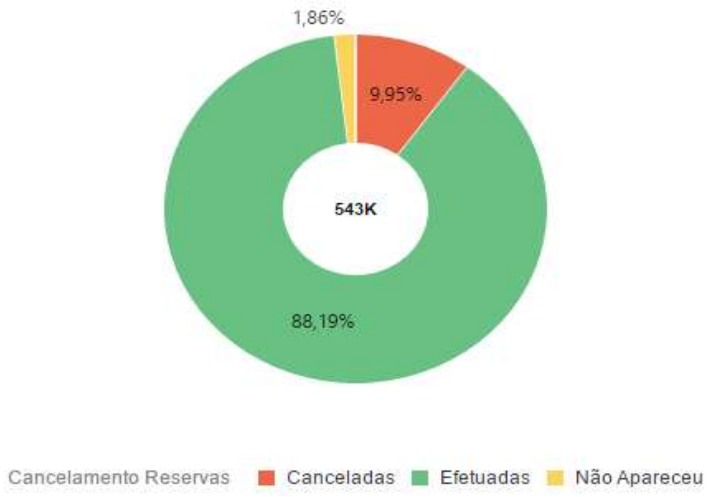


Figura 14 Cancelamento de reservas

No que diz respeito aos meios de pagamento utilizados, Figura 15, destaca-se a existência de pagamentos efetuados por operadores turísticos (*Operator Pays*), que contratam os serviços de transporte da empresa. Nesse caso o cliente da empresa é um operador turístico e não o turista que é transportado, estes casos representam 9,63% dos meios de pagamento. Existem situações em que o serviço é prestado a um operador turístico e o pagamento é efetuado mais tarde no tempo, identificado como *YTF Collects, representando* 8,41% dos casos. Dos restantes meios de pagamento, destaca-se a preferência pelo pagamento em dinheiro em que 55.13% dos serviços são pagos desta forma. A preferência por pagar em dinheiro pode estar relacionada com as características inerentes dos serviços de transporte, em que normalmente o pagamento é efetuado no dia da prestação do serviço.

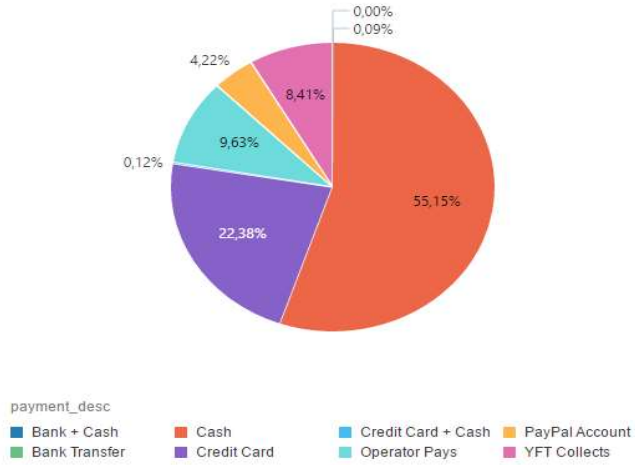


Figura 15 Meios de pagamentos utilizados

Na Figura 16 foi analisado o número de adultos, crianças e bebés transportados. A partir dos meses da primavera verifica-se um aumento do número de crianças e bebés transportados, atingindo o pico nos meses de verão e diminuindo nos meses de inverno. A identificação deste padrão nos dados pode ser um indicador de que as viagens em família podem estar relacionadas com o período das férias escolares.

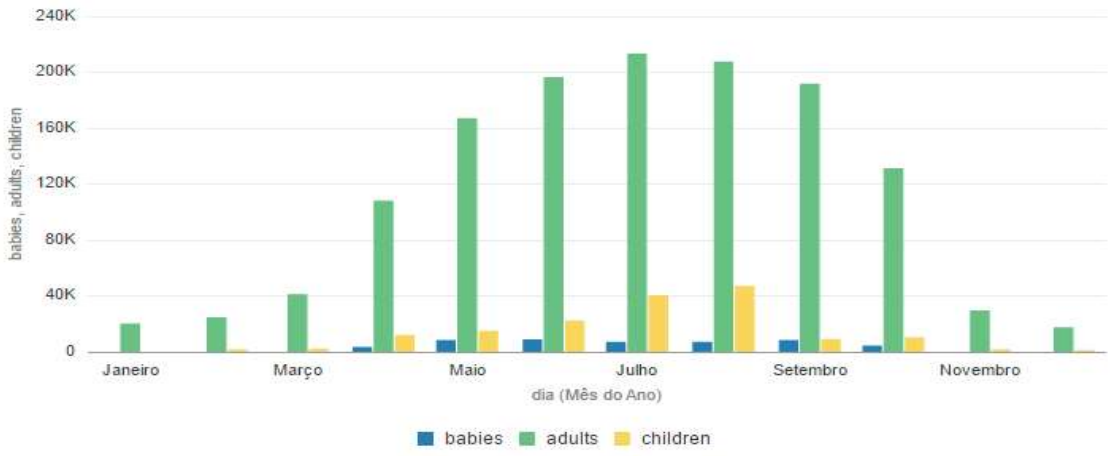


Figura 16 Pessoas Transportadas

O ficheiro de respostas do questionário de satisfação dos clientes, *feedback.csv*, possui ainda um atributo de texto livre onde os clientes podem comentar a sua experiência relativamente ao serviço prestado. A análise do texto dos comentários pode revelar pontos interessantes que ainda não se tenham revelado. Para investigar a existência de padrões relativamente à satisfação dos clientes, recorreu-se a técnicas de *Text Mining* para a visualização de texto.

Na Figura 17., é apresentado o *Word Cloud* dos comentários dos clientes. O *Word Cloud* define-se como um modelo de representação de palavras de fácil interpretação, em que as palavras mais frequentes encontram-se destacadas através da sua dimensão e por diferentes cores.

Para o pré processamento do texto, recorreu-se à biblioteca de processamento de linguagem natural, *Natural Language Toolkit* ou *NLTK,* disponível na linguagem *Python*. Nesta foi efetuado a limpeza e pré-processamento simples do texto, através normalização das palavras, em que as letras maiúsculas foram convertidas em minúsculas, foram removidas as palavras não discriminantes (*stop words*), foi eliminada a pontuação e foram criados *tokens* de forma a permitir identificar as palavras mais frequentes dos comentários.

Apesar de existirem alguns comentários em português, a maioria encontra-se escrito na língua inglesa. Foi evidente pela visualização das palavras mais frequentes, na Figura 17, que estas sugerem um grande nível de satisfação dos clientes.

Figura 17 Word Cloud dos comentários dos clientes

3.3. Preparação dos dados

A fase de preparação dos dados do *CRISP-DM* constitui uma das fases mais importantes num projeto de Extração de Informação e onde geralmente se consome a maior parte do tempo.

É estimado que a preparação dos dados pode consumir entre 50-70% do tempo do projeto. Investir tempo nas fases iniciais de compreensão do negócio e dos dados, pode

minimizar o tempo de preparação dos dados. Dependendo do tipo de organização e dos objetivos a atingir, a preparação dos dados engloba as seguintes tarefas: (1) conjugar *data sets* e registos;(2) selecionar um subconjunto de dados;(3) construir novos atributos;(4) ordenação dos dados para a construção de modelos;(5) remoção ou substituição de valores omissos;(6) dividir os dados em conjunto de treino e testes.

Partindo do conjunto de dados iniciais, é ainda nesta fase do *CRISP-DM* onde é efetuada a seleção dos dados considerados relevantes para cumprir os objetivos do projeto. Existem duas formas de selecionar os dados: (1) seleção de amostras (linhas), que envolve decidir sobre o que considerar na amostra a analisar e (2) seleção de atributos ou características (colunas) que envolve decidir sobre a utilização de determinadas caraterísticas como, por exemplo, um determinado tipo de serviço ou clientes com uma determinada caraterística.

A próxima etapa desta fase consiste na limpeza dos dados selecionados para a análise, onde é necessário investigar possíveis problemas que necessitam de ser corrigidos. Existem diferentes tipos de problemas e formas de os mitigar, algumas delas apresentadas na Tabela seguinte.

Tabela 2 Limpeza dos dados. Adaptado de CRISP-DM (1999), pag. 22

Problema nos dados	**Possível solução**
Dados em falta	Eliminar as linhas ou preencher os dados em falta com um valor estimado.
Erros nos dados	Utilizar a lógica para detetar os erros manualmente e substituir ou excluir os dados errados.
Inconsistências de codificação	Decidir sobre um único esquema de codificação, depois converter e substituir valores.
Falta de metadados ou maus metadados	Examinar manualmente os campos suspeitos e identificar o significado correto.

Nas secções seguintes foi criada uma secção especifica para a preparação dos dados de cada modelo construído.

3.4. Extração de padrões dos comentários dos clientes

Os clientes da empresa preenchem um questionário, onde partilham a sua opinião sobre o serviço na forma de *reviews.* Neste sentido procura-se analisar os comentários dos clientes, esperando identificar tópicos que permitem identificar os aspetos que os

clientes mais comentam e em última instância permitir a empresa compreender que decisões estratégicas estão na origem dos comentários.

Os algoritmos de *Text Mining* surgem como importantes ferramentas para a extração de informações com foco na modelação de tópicos relevantes nos *reviews*. A análise de tópicos é utilizada para descobrir os principais temas abordados numa coleção de documentos não estruturados, sem recorrer a qualquer tipo de anotação nos documentos analisados, constituindo assim uma abordagem de aprendizagem não supervisionada.

Para introduzir o método de análise de tópicos utilizado, *Latent Dirichlet Allocation* (*LDA*), recorre-se ao conceito de *word space*, apresentado por Sahlgren [141]. Este autor define o *word space*, como um modelo computacional para obter o significado das palavras. Baseia-se na distribuição das palavras nos documentos para representar a sua semelhança semântica em termos de proximidade espacial, num vetor de contexto.

De acordo com Alghamdi e Alfalqi [142], *LDA* é um algoritmo de text mining baseado em modelos de tópicos que utilizam a estatística bayesiana. São considerados modelos generativos que tentam reproduzir o processo de escrita, tentado reproduzir o documento que deu origem ao tópico. A ideia base é modelar cada documento como uma mistura de tópicos em que cada tópico constitui uma distribuição discreta de probabilidades de uma palavra pertencer a um tópico.

3.4.1 Preparação dos dados

Da análise inicial dos comentários dos clientes, verificou-se que a maioria estava escrita em inglês, totalizando 50 937 comentários escritos em inglês, registados entre abril de 2012 e dezembro de 2019.

O pré-processamento dos dados trata-se de uma etapa relevante, uma vez que proporciona uma primeira fase de estruturação dos mesmos. Foram utilizadas várias técnicas de pré-processamento de texto recorrendo a bibliotecas da linguagem *Python*, foi utilizada a biblioteca *Gensim* [143] para a modelação de tópicos com *LDA*, para a representação e pré-processamento dos documentos recorreu-se ao *NLTK* [144]. Para a visualização dos dados foram utilizadas as bibliotecas *WordCloud* e *pyLDAvis* [145]. O pré-processamento do texto iniciou por (i) remover a pontuação e as palavras não discriminantes (*stop words*), (ii) normalização das palavras, em que as letras maiúsculas foram convertidas em minúsculas e (iii) aplicou-se a lematização com intuito de reduzir

as palavras que se encontram em formas derivadas para a sua forma base (radical). (iv) Recorreu-se a bigramas e trigramas com o objetivo de identificar as palavras únicas. (v) Para captar as classes gramaticais abertas, aplicou-se filtros, recorrendo a técnica de *part-ofspeech tagging* com o intuito de identificar a categoria gramatical dos nomes. (vi) Por último foram eliminados os extremos que consistem nas palavras pouco ou muito frequentes, que ocorrem em menos do que 20 documentos ou mais do que 80% dos documentos. Ainda nesta fase, optou-se por reservar 50 documentos para realizar inferência de tópicos.

Na Figura 18 é apresentada o *Word Cloud* onde foram aplicadas técnicas de pré-processamento utilizadas em *Text Mining* para a construção do dicionário de termos a utilizar na análise de tópicos.

Após o pré processamento do texto continuam a estar destacadas palavras que indicam a satisfação dos clientes. São frequentes as palavras em português, "muito" e "obrigado", o que indica que apesar da língua dos comentários ser a inglesa, muitos *reviews* são terminados com o agradecimento na língua portuguesa. Surgem ainda nomes de motoristas como palavras frequentes, o que corrobora com o papel importante que o perfil dos motoristas desempenha para o sucesso da empresa.

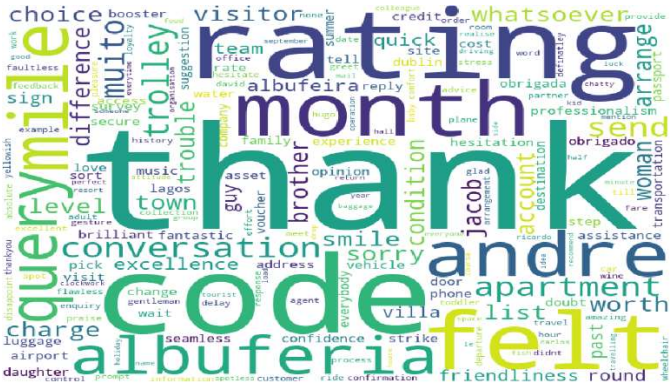


Figura 18 WordCloud após pré-processamento das reviews

3.4.2 Construção e avaliação do modelo

Na avaliação de um modelo de tópicos o objetivo principal consiste em avaliar se o modelo encontrado vai de encontro com a interpretação humana dos tópicos e se estes permitem diferenciar os temas tratados nos documentos analisados [146]. Trabalhos

anteriores sugerem que a qualidade de um tópico é determinada pela coerência das palavras que formam o tópico e pela sua importância para o problema relativamente aos restantes tópicos [147], [148], [144].

Têm sido aplicados métodos internos de avaliação, como a perplexidade, onde é avaliada a quantidade de informação representada num tópico [149]. Por outro lado, existem os métodos externos que utilizam os tópicos obtidos pelos modelos para representar um vocabulário que é comparado com a interpretação humana dos documentos [150]. A perplexidade é uma medida estatística de quão bem um modelo de probabilidade prevê uma amostra. No entanto, esta estatística não faz muito sentido, quando aplicada a um único método, ou seja, o benefício desta estatística é comparar a perplexidade entre diferentes modelos, sendo que o modelo com menor perplexidade é geralmente considerado o melhor.

Em relação a coerência, é uma das principais técnicas utilizadas para estimar o número de tópicos. A medida da coerência atribui uma pontuação ao tópico, de acordo com o grau de similaridade semântico entre as palavras com a pontuação mais elevada do tópico. Trata-se de uma métrica que auxilia na distinção dos tópicos que são semanticamente interpretáveis dos que foram obtidos por artefactos resultantes de inferência estatística [146].

O algoritmo *LDA* configura um método paramétrico onde é necessário indicar o número de tópicos. A abordagem para identificar o número de tópicos ótimo presentes nas *reviews* consistiu em calcular o valor da coerência de vários modelos *LDA* fazendo variar o número de tópicos. A Figura 49 do Apêndice C demonstra que a coerência dos tópicos é crescente e atinge o seu valor mais elevado quando o número de tópico é igual a 8. A experiência realizada com 8 tópicos, Figura 19, demonstrou que o modelo capturou tópicos que correspondem a subcategorias de reviews com sobreposição de termos entre os tópicos. Pela análise do gráfico é visível que para além da sobreposição dos tópicos existem tópicos que estão muito próximos formando três grupos. Por esse motivo a escolha de 3 tópicos demonstrou ser a opção mais adequada, por não apresentar sobreposição de termos entre os tópicos e por estes estarem claramente separados, Figura 20.

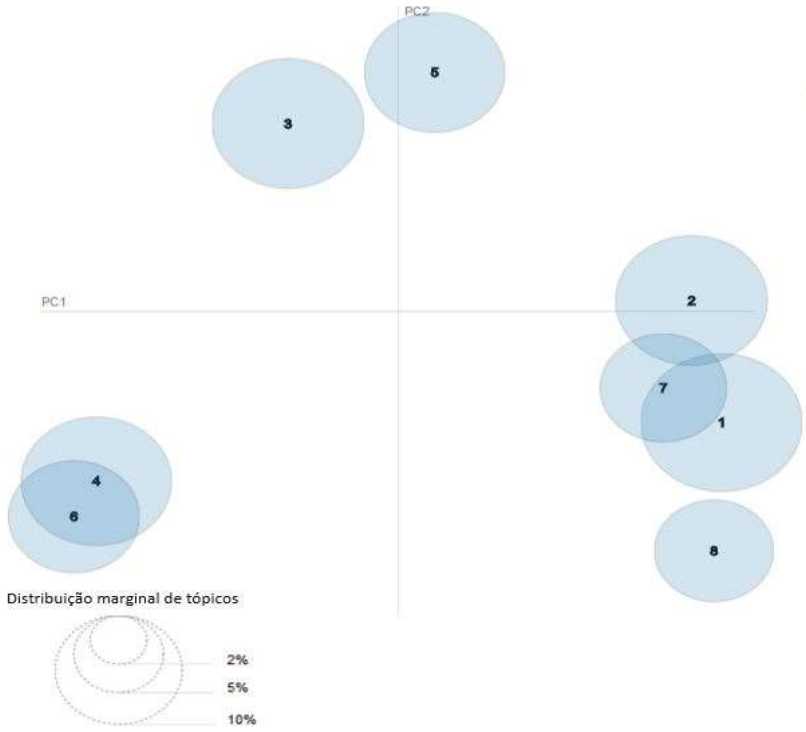


Figura 19 Distância intertópico do modelo LDA com 8 tópicos

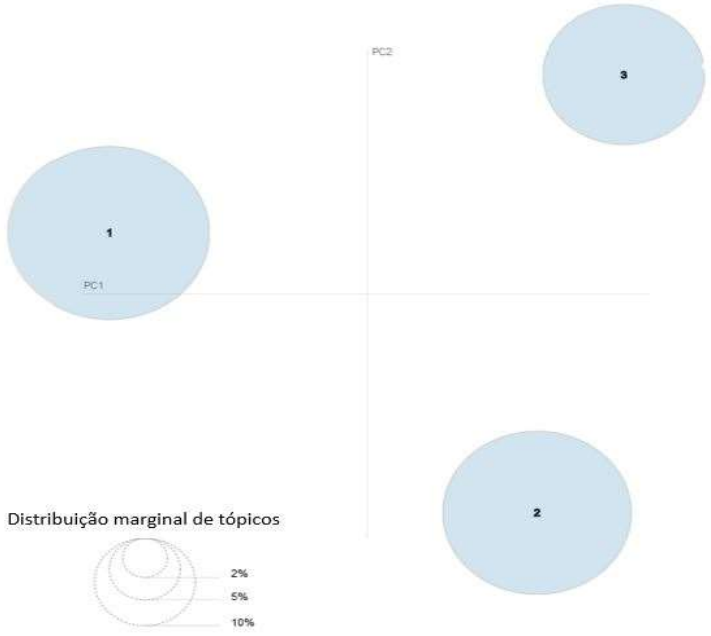


Figura 20 Distância intertópico do modelo LDA com 3 tópicos

A Figura 21 apresenta o *word cloud* dos termos mais frequentes identificados nos três tópicos obtidos no modelo *LDA*. O tópico 1 refere-se a comentários sobre o tipo de serviço

prestado, a relação serviço pela qualidade/preço e se recomendam o serviço a amigos ou familiares. O tópico 2 refere-se a comentários com ênfase nos condutores e da experiência do transporte de ida e volta entre o aeroporto e o hotel. O tópico 3 agrupa comentários sobre a empresa *YellowFish Transfer*, da experiência com o serviço de reservas e com os empregados.

Figura 21 Word Cloud dos 3 tópicos obtidos pelo algoritmo LDA

3.4.3 Discussão dos resultados

Devidas às características paramétricas do modelo *LDA* em que é necessário indicar o número de tópicos, torna-se complexo conseguir encontrar um valor ótimo para o número de tópicos quando não conhecemos as categorias dos documentos. Assim, as técnicas de pré-processamento e representação dos documentos constituem-se como um passo relevante na criação do modelo.

Deste modo, foram realizadas diversas experiências de representação de documentos e conclui-se que a simples eliminação da pontuação, a conversão das palavras para minúsculas, assim como a eliminação da "*stop words*" fazem realçar os termos que transmitem o conteúdo semântico dos tópicos, mas os modelos de tópicos produzidos são de difícil interpretação.

A aplicação de técnicas como a lematização, a filtragem de palavras com *part-of-speech* e a filtragem de extremos (palavras pouco ou muito frequentes), fez melhorar os resultados de forma notável.

A Tabela 3 apresenta o resultado da utilização do modelo *LDA* treinado, para fazer inferência, recorrendo a uma amostra dos documentos reservados na fase de pré-processamento dos dados. Os valores da tabela representam a probabilidade do comentário pertencer a cada um dos tópicos. Alguns tópicos são claramente identificados pelo modelo com probabilidade muito alta, mas existem situações em que o modelo atribuiu probabilidades muito próximos entre dois tópicos, pelo que nestes casos a interpretação humana é crucial para avaliar o desempenho do modelo.

Tabela 3 Utilização do modelo LDA para inferência

	Probabilidade de pertencer aos tópicos		
Comentário dos clientes	**Tópico 1**	**Tópico 2**	**Tópico 3**
We had a very pleasant experience with Yellowfish.It was our first time using this service. Both drivers were very pleasant and courteous. They drove very carefully and both were very informative and attentive. We will reccommend Yellowfish to others, just as it was reccommended by another family member.	0,429	0,042	0,530
Drivers are always polite and helpful	0,167	0,651	0,182
As always excellent service, used for many years and hopefully many more to come. Don't even think about anyone else.	0,824	0,083	0,093
We will definately use your company again for future trips and I have recommended you to friends too.	0,333	0,086	0,580
One of the best transfers I have ever had	0,167	0,167	0,667
I've used your service in the past and again I found it to be excellent. I would highly recommend your service. Both drivers were courteous and professional and are a credit to your organisation. Hope to see you again next year. Well done Yellowfish.	0,387	0,158	0,387
Service was extremely efficient and we would recommend Yellowfish to anyone. Thank you.	0,583	0,083	0,333
Thanks again, using yellow fish now for a few years, so reliable it's unbelievable. Clean cars, friendly, helpful drivers. The only way to travel, highly recommend yellow fish	0,302	0,068	0,630
The driver was late picking us up for the return to the airport. She had the incorrect pick up address. She arrived just as we were getting a taxi to take us to the airport. We did get to the airport in time for our flight. The drives was courteous and apologized for being late	0,028	0,943	0,028
All 6 of us ladies, taught the service with excellent, both drivers was extremely courteous, and we will definitely use again.	0,444	0,434	0,121

3.5. Padrões relativos à satisfação dos clientes

No capítulo anterior analisou-se os comentários dos clientes esperando identificar tópicos que permitissem identificar os aspetos que os clientes mais comentam sobre o serviço que lhes foi prestado. Em complemento à análise anterior, no sentido de avaliar as características do serviço que mais são valorizadas, recorreu-se ao questionário de satisfação dos clientes.

3.5.1 Preparação dos dados

Na Tabela 4 encontram-se listadas as perguntas do questionário de satisfação dos clientes utilizadas na análise. De notar que as opções de resposta para cada questão assumem valores entre 1 e 5, configurando um *rating* que o cliente atribui a cada questão. Apesar dos *ratings* serem valores em escalas de *Likert*, para o efeito da análise foram consideradas como variáveis qualitativas.

Cada serviço dá origem ao preenchimento de um questionário. Quando um cliente solicita um serviço de ida e volta é-lhe pedido para preencher dois questionários, um referente ao serviço de chegada e outro para o serviço de partida.

Como ponto de partida foram analisadas as estatísticas descritivas dos *ratings* (ver Tabela 14 do Apêndice A) e estas indicam que foram respondidos 84122 questionários, não apresentando valores omissos.

Antes de prosseguir com a análise foi avaliada a fiabilidade do questionário. Existem diferentes estatísticas para estimar a fiabilidade de um instrumento de medida.

Na área das ciências sociais, em que são efetuados questionários, é usual recorrer a testes estatísticos para estimar a consistência interna das respostas, ou seja, a proporção da variabilidade nas respostas que resulta de diferenças nos inquiridos. Para medir a consistência interna de um teste ou uma escala, Cronbach [151] desenvolveu o coeficiente alfa, que hoje é a estatística mais utilizada para medir a consistência de um questionário. Um questionário é considerado consistente quanto mais perto de 1 estiver o coeficiente alfa. Há muita discussão sobre os valores aceitáveis de alfa. De um modo geral, um instrumento ou teste é classificado como tendo fiabilidade apropriada quando o alfa é de pelo menos 0,70. A maneira prática de interpretar o valor de alfa é comparar o valor calculado com o valor preconizado em tabelas apresentadas na literatura [152].

Grande parte das variáveis contribui para o valor do alpha de Cronbach em 0,83, o que é considerado um valor bom, pois se alguma das variáveis fosse eliminada deste conjunto o valor de alpha diminuiria. O resultado do teste pode ser consultado na Tabela 15 e Tabela 16 no Apêndice D.

Tabela 4 Questões do questionário de satisfação dos clientes à chegada

Your Arrival
How was our welcome at the airport?
How punctual was your driver at the airport?
How clean was the interior of your Yellowfish vehicle?
How clean was the exterior of your Yellowfish vehicle?
How was the driving experience to your resort?
How friendly was your driver?
Was the vehicle boot big enough for your luggage?
Our booking process
How easy was it to book with Yellowfish?
In General
Would you recommend Yellowfish Transfers to anyone?
Anything else to add?
comments If there's something else you'd like to tell us, this is the place to do it

A existência de um número elevado de atributos nos dados pode tornar complexa a utilização dos algoritmos de extração de padrões, devido não só ao custo computacional como também pela dificuldade na interpretação dos resultados obtidos [153]. Como forma de mitigar este tipo problemas é usual a utilização de técnicas de redução de dimensionalidade, trazendo ganhos no desempenho dos algoritmos, na visualização dos dados e na compressão dos dados de grande dimensionalidade [154].

Nesta fase da análise, pretende-se diminuir o número de variáveis, por forma a poder utilizar essa informação na análise de Clusters. Para tal recorreu-se a Análise de Componentes Principais ou ACP como técnica de redução de dimensionalidade.

Na secção seguinte são descritas todas as etapas da ACP.

3.5.2 Análise de Componentes Principais

A ACP é uma técnica de análise exploratória de dados multivariados cujo objetivo é o de resumir a informação existente, num conjunto de variáveis correlacionadas num número menor de variáveis independentes que contenham a maior parte da informação presente nas variáveis originais [155].

Pretende-se que as primeiras componentes sejam suficientes para explicar grande parte da variabilidade das variáveis originais de forma a resumir os dados sem grande perda de informação. As componentes obtidas são independentes entre si, pelo que podem ser utilizadas não só na análise de *clusters,* mas também em análises onde a multicolinearidade (existência de correlação entre variáveis explicativas) pode constituir um problema, como é o caso de problemas de regressão linear múltipla [154].

A ACP deve respeitar os seguintes requisitos [156]: (1) as variáveis devem ser métricas (as escalas de Likert não são quantitativas mas podem ser consideradas como tal); (2) a dimensão da amostra deve ser adequada, devendo existir pelo menos cinco vezes mais casos do que o número de variáveis;(3) existir correlações entre as variáveis originais que podem ser confirmadas através da estatística de Kaiser- Meyer-Olkin (KMO) e do teste de Bartlett [157]; (4) as comunalidades devem ser altas, sendo a comunalidade a proporção de variância de cada variável que é explicada pelo conjunto das componentes retidas.

A primeira etapa da análise consistiu em analisar a matriz de correlações de forma a identificar grupos de variáveis correlacionadas e verificar a sua adequação para a ACP. Para tal, recorreu-se ao teste de esfericidade de Bartlett [157], para verificar se a matriz de correlações é significativamente diferente da matriz identidade e a estatística de *Kaiser-Meyer-Olkin (*KMO*)* para confirmar a adequabilidade das correlações.

As hipóteses do teste de *Bartlett* são: (*H0*) a matriz de correlações é a matriz identidade, ou seja, as correlações entre as variáveis são zero; vs. (*H1*) a matriz de correlações não é a matriz identidade. Para prosseguir com a análise, pretende-se rejeitar *H0*. Analisando a significância do teste de *Bartlett*, pode-se inferir que as variáveis estão correlacionadas entre si ($Sig = \ 2{,}2E^{-16} < 0{,}05$) logo rejeita-se *H0*.

Por outro lado, sendo o KMO de 0,84, a adequação da ACP aos dados é considerada boa [158]. Este valor não desaconselha a utilização de uma ACP para os dados em análise. O resultado do teste pode ser consultado na Tabela 17 do Apêndice D.

Os resultados obtidos podem ser confirmados através da visualização da matriz de correlações ordenada, na Figura 22, onde é possível identificar padrões de grupos de variáveis correlacionadas.

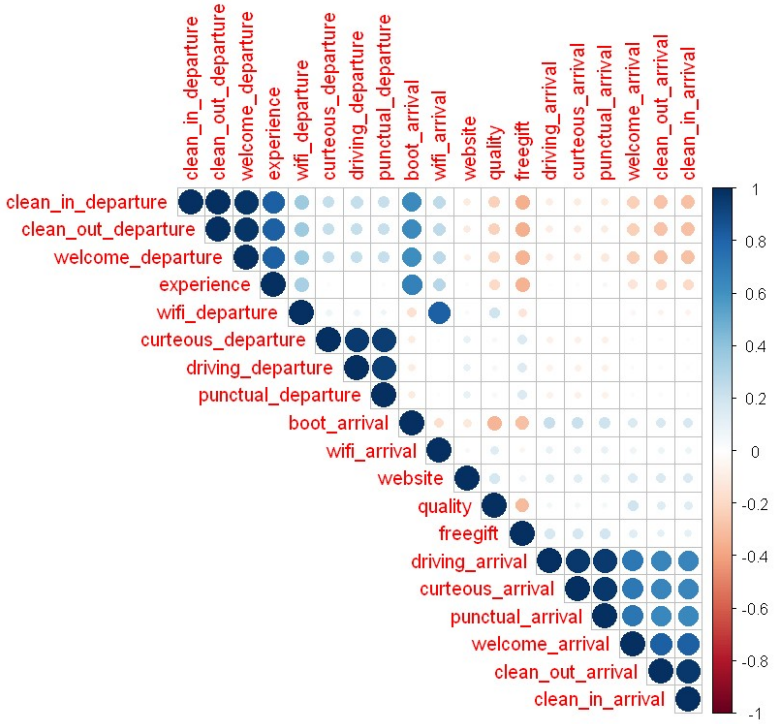


Figura 22 Matriz de correlações ordenada

A ACP implementada não defini a *priori* o número de componentes a reter, como resultado são apresentados tantos componentes quanto o número de variáveis analisadas.

O resultado da ACP encontra-se listado na Tabela 5 e da sua interpretação constata-se que:

- A CP1 é a que apresenta a maior variância (maior valor próprio) e explica 29.44% do total de informação contida em todas as variáveis.
- A percentagem de variância explicada pela CP2 é de 22,14 %, a CP3 explica 15,74%, e os valores vão diminuindo;
- As primeiras 4 componentes explicam por si 77,77% da variância total das variáveis em análise; se extrairmos mais uma componente, a CP5, a

variabilidade total é explicada em 84,46%, o que corresponde a um incremento de apenas 6,69%.

- Só quando se considera as 19 componentes (tantos componentes quantas as variáveis originais) é que se explica 100% da informação, ou seja, 100% da variância total das 19 variáveis em análise;

Tabela 5 Resultado da ACP

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Desvio Padrão	2,365	2,051	1,72	1,402	1,128	0,966	0,874	0,592	0,451	0,441
Proporção da Variância	0,294	0,221	0,157	0,105	0,067	0,049	0,040	0,018	0,011	0,010
Proporção acumulada	0,294	0,516	0,673	0,778	0,845	0,894	0,934	0,952	0,963	0,973
	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	
Desvio Padrão	0,401	0,294	0,253	0,211	0,205	0,199	0,180	0,171	0,069	
Proporção da Variância	0,008	0,005	0,003	0,002	0,002	0,002	0,002	0,002	0,001	
Proporção acumulada	0,982	0,986	0,990	0,992	0,994	0,997	0,998	0,999	1,000	

A decisão do número de componentes principais a reter depende da informação que se pode desprezar. No entanto, quantas mais componentes forem retidas, menos úteis se tornam cada uma delas, porque vai diminuindo o seu valor próprio. De acordo com o critério do Scree Plot [159] apresentado na Figura 23, a escolha deve recair sobre uma solução de 5 componentes, uma vez que, no gráfico, ainda se regista um declive acentuado do 4º para o 5º. Repare-se que, para a retenção de 4 componentes, a percentagem de variância explicada seria 77,77% e, para 5 componentes, seria de 84,46%, resultando num incremento de apenas 6,69%. Como o ganho em termos de valor próprio é pequeno, optou-se por uma solução de 4 componentes.

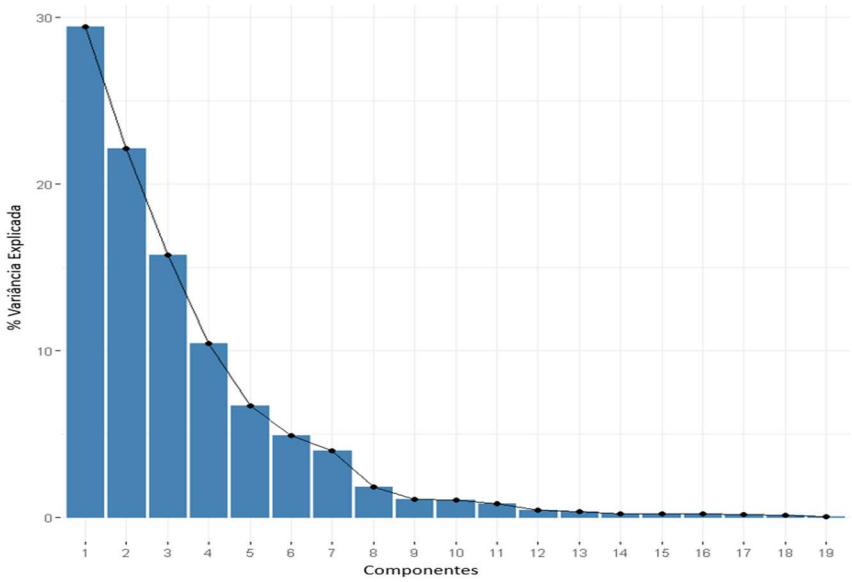


Figura 23 Proporção de variância explicada de cada componente

Assim, a ACP permitiu a obtenção de quatro componentes. Na Figura 50, Figura 51, Figura 52 e Figura 53 do Apêndice D é apresentada a contribuição de cada variável nos diferentes componentes onde é possível apurar as seguintes características:

- Na CP1(Figura 50, Apêndice D) as variáveis com maior peso dizem respeito a aspetos de limpeza interior e exterior do veículo e do bom acolhimento, tanto na chegada como na partida bem como a pontualidade na chegada.
- Na CP2 (Figura 51, Apêndice D), as variáveis com maior peso dizem respeito a aspetos de espaço para transporte de bagagens à chegada.
- Na CP3 (Figura 52, Apêndice D) a importância recai sobre aspetos relacionados com a cortesia do condutor, tipo de condução e pontualidade na partida.
- Na CP4 (Figura 53, Apêndice D) o maior peso está nos aspetos de conforto a bordo, como o acesso a *wifi*.

Recorrendo a análise anterior, foi atribuído um nome a cada variável, de acordo com as contribuições mais importantes das variáveis em cada componente.

Tabela 6 Atribuição de um nome a cada componente

Componente	Variável	Aspetos mais valorizados
CP1	cleaness_punctual_friendly_arrival	Limpeza interior e exterior do veículo, bom acolhimento na chegada e partida e pontualidade na chegada.
CP2	lugage_space	Espaço para transporte de bagagens à chegada.
CP3	curteous_driving_puntual_departure	Cortesia do condutor, tipo de condução e pontualidade na partida
CP4	Comfort_wifi	Conforto a bordo, como o acesso a wifi.

Os 4 componentes obtidos foram utilizados na análise de *clusters* apresentada nas secções seguintes.

3.5.3 Construção e avaliação do modelo

Os recursos computacionais utilizados na análise, consistiram num computador com um processador Intel Core i5-9400F a 2.9Ghz, com 16Gb de RAM e sistema operativo Windows 10.

Inicialmente recorreu-se ao algoritmo *k-means,* mas devido a limitações de recursos computacionais, não foi possível executar o algoritmo para todas as 84122 observações, apresentando erros de memória insuficiente. Para ultrapassar as limitações encontradas com o *K-means*, optou-se pelo algoritmo *CLARA* (*Clustering Large Applications*) [160]. Este algoritmo é apresentado pelos seus autores como sendo uma extensão do algoritmo *PAM*(*k-Medoids*), que recorre a uma abordagem de *sampling* para reduzir o tempo de execução e problemas relacionados com recursos de memória *RAM*. Os mesmos autores ainda referem que o algoritmo *k-Medoids* é mais robusto que *o k-means*, menos sensível a *outliers*, por utilizar *medoids* em vez da média como ponto central.

Uma das fragilidades dos algoritmos de *clustering* advém do facto de devolverem sempre grupos, mesmo tratando-se apenas de grupos aleatórios. Por esse motivo, antes de aplicar o algoritmo de *cluster* é importante efetuar uma análise preliminar para avaliar se os padrões existentes nos dados não se tratam de estruturas aleatórias. Para avaliar a tendência de *clusters* nos dados, foi utilizada a abordagem *VAT* (*Visual Assessment of Cluster Tendency*) [161]. A aplicação do *VAT* consistiu no cálculo da matriz de

dissimilaridade ordenada, onde o grau de dissimilaridade entre casos é representado como uma distância entre pontos projetados no espaço multivariado [161].

A matriz de dissimilaridade é apresentada com um padrão de cores interpretadas da seguinte forma: vermelho significa alta similaridade (i.e baixa dissimilaridade) e azul baixa similaridade. A análise do gráfico da Figura 54 do Apêndice D, indica que existem grupos que representam padrões não aleatórios nos dados.

Na secção 2.2.6 foi referida a importância da escolha da medida de distância a utilizar e da sua influência no resultado da análise. Uma vez que pretendemos agrupar os objetos de acordo com a magnitude da pontuação das respostas do questionário, foram avaliadas as distâncias Euclidianas e de Manhattan. Foi escolhida a distância Euclidiana por ter apresentado melhores resultados em termos de coeficiente de silhueta.

Dado que o *CLARA* configura um algoritmo de particionamento, onde o número de *clusters* K é um parâmetro obrigatório é necessário determinar o valor ótimo para k. Tendo presente a subjetividade dos métodos utilizados para determinar um possível valor de k, optou-se por seguir a abordagem apresentada por Kassambara [95], aplicando métodos diretos e estatísticos.

Os métodos diretos consistem em critérios de otimização como, por exemplo, o coeficiente da silhueta ou o método do cotovelo (*elbow*). Os métodos de testes estatísticos consistem em comparar evidências com a hipótese nula como, por exemplo, o método de estatística *GAP*.

O método do cotovelo (*Elbow Metthod*) consiste na visualização gráfica de uma medida da qualidade da partição, obtida em função do número de *clusters*.

O coeficiente de silhueta é uma técnica utilizada para avaliar o grau de afastamento dos objetos de um *cluster*. O valor do coeficiente de silhueta encontra-se entre -1 e 1 e reflete o grau de afastamento do objeto em relação aos objetos do *cluster* onde este pertence e de outros *clusters*. Quando o valor se aproxima de 1 significa que o cluster que contém o objeto está compacto e que o objeto está distante de outros *clusters*, sendo esta a situação ideal [160].

Tabela 7 Interpretação do valor de silhueta. Fonte: Adatpada de Kaufman e Rousseew (1990)

Coeficiente	Sugestão de interpretação
0,71 − 1,00	Foram encontrados grupos com uma estrutura muito forte
0,51 − 0,70	Os Grupos possuem uma estrutura razoável
0,26 − 0,50	A estrutura encontrada é fraca e pode ser artificial; recomenda-se a aplicação de outros métodos no conjunto de dados.
≤ 0.25	Não foi encontrada qualquer estrutura no conjunto de dados

Os resultados encontrados para o valor ótimo de K, pela aplicação dos diferentes métodos enunciados anteriormente, apresentam valores compreendidos entre 5 e 8 *clusters* conforme apresentado na Figura 59, Figura 60 e Figura 61 do Apêndice D.

Adicionalmente foi utilizada a biblioteca *NbClust* que recorre a diferentes métodos para calcular simultaneamente o número de *cluster* utilizando 30 índices diferentes [95]. De acordo com este método, recorrendo a regra da maioria dos índices, o número ótimo de *clusters* é de 6. O resultado pode ser confirmado na Figura 62 do Apêndice D.

O algoritmo *CLARA* foi executado fazendo variar o parâmetro K entre 4 e 8 *clusters*. Foi selecionado como solução final o valor de K=6, por ter apresentado o maior valor de coeficiente de silhueta, 0,8. De acordo com a Tabela 7, o coeficiente de silhueta de 0,8 corresponde a uma solução onde foram encontrados grupos com uma estrutura muito forte.

Na Figura 24 é apresentada a visualização gráfica das observações de cada *cluster* e na Figura 25 o coeficiente de silhueta. No gráfico de silhueta, apresentado, o eixo horizontal representa as observações de cada grupo e o eixo vertical o valor da silhueta. O gráfico da Figura 24 recorre a biblioteca R, *factoextra* [95] para representar graficamente os *clusters*. A biblioteca transforma as 4 variáveis iniciais num novo conjunto de variáveis através de componentes principais e seleciona as duas primeiras componentes, *Dim1* e *Dim2 que* representam 28,5% e 37,9%, respetivamente, de variação (ou seja, informação) contida no conjunto de dados original. As duas dimensões ou componentes foram utilizadas para representar o conjunto de dados original e projetar graficamente os *clusters*.

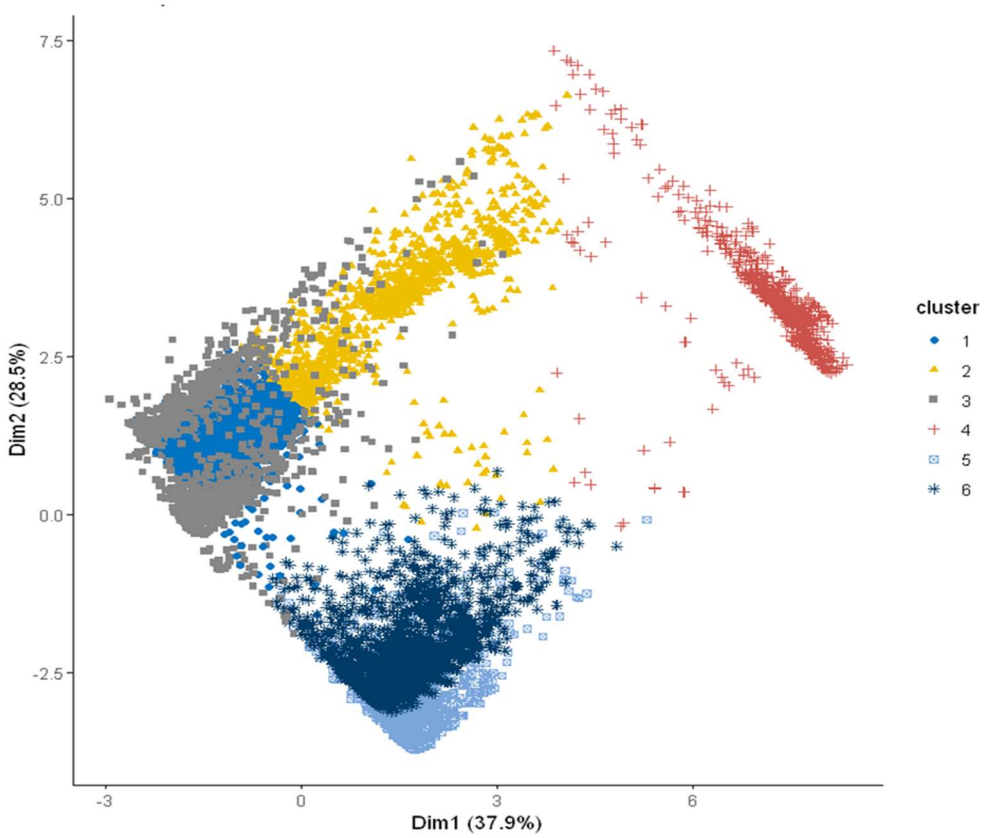


Figura 24 Visualização dos clusters para K=6

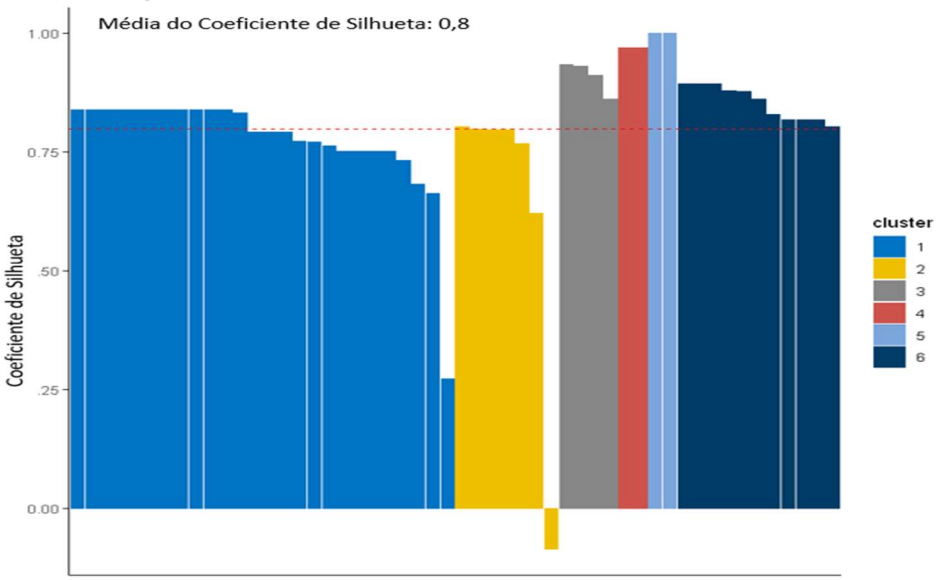


Figura 25 Coeficiente de silhueta para K=6

3.5.4 Discussão dos resultados

Para uma melhor compreensão dos resultados obtidos pela solução de 6 clusters, cada cluster foi analisado em termos dos centroides. Pela interpretação da Figura 26, podemos concluir que os grupos possuem algumas características a destacar, tais como:

- O Grupo 1 com o maior número de casos, 42576, é o que mais valoriza aspetos da CP3, relacionados com a cortesia do condutor, tipo de condução e pontualidade na partida.

- O Grupo 2 com 6077 casos, a par do Grupo 1 valoriza aspetos da CP3, relacionados com a cortesia do condutor, tipo de condução e pontualidade na partida, ainda este grupo destaca-se por ser o grupo que menos valoriza aspetos da CP2, relacionados com o espaço para transporte de bagagem.

- O Grupo 3 com 8156 casos é o que menos valoriza aspetos relacionados com a cortesia do condutor, tipo de condução e pontualidade na partida, CP3.

- O Grupo 4 com 4829 casos é o que mais valoriza aspetos da CP1, relacionados com a limpeza interior e exterior do veículo, simpatia do condutor e pontualidade na chegada.

- O Grupo 5 com 4645 é o que menos valoriza aspetos da CP4, relacionados com o conforto a bordo, como o acesso a wifi e é o que mais valoriza aspetos da CP2, relacionados com o espaço para transporte de bagagem.

- O Grupo 6 é o segundo maior, com 17839 casos, é o que mais valoriza aspetos da CP4, relacionados com o conforto a bordo, como o acesso a wifi.

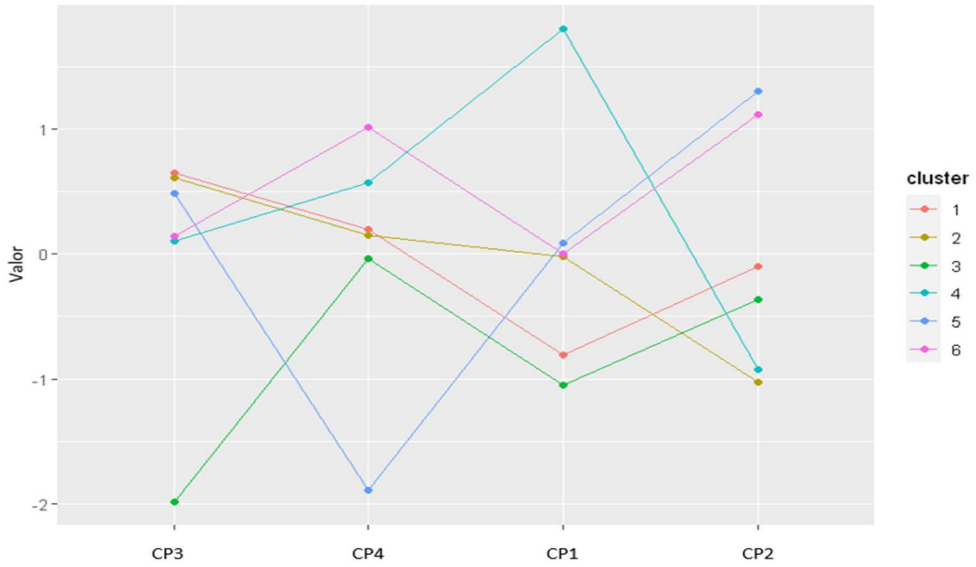


Figura 26 Visualização dos Centroides de cada cluster

Adicionalmente, com o intuito da solução poder ser aplicada no apoio à tomada de decisão em diferentes áreas de ação como, por exemplo, para decidir qual o perfil de condutor ou qual o veículo é o mais adequado para determinado grupo de clientes, foi efetuada uma análise exploratória e analisadas características como, o tipo de serviço contratado, o mês do serviço, o local de *dropoff* e país de origem.

Relativamente ao tipo de serviço contratado pelos clientes, pela análise da Figura 27 conclui-se que os *clusters* 1, 5 e 6 agrupam apenas clientes que contratam serviços de ida e volta para o aeroporto (*RAL*) e os *clusters* 2 e 4 incluem os dois tipos de serviços num único sentido, *OAL* e *OLA,* sendo o serviço que tem como destino final o aeroporto o que mais se destaca. A maior parte das observações do *cluster* 3 diz respeito a serviços num único sentido do aeroporto para um local (*OAL*), relembro que este é o grupo que menos valorizava aspetos relacionados com a cortesia do condutor, tipo de condução e pontualidade à partida.

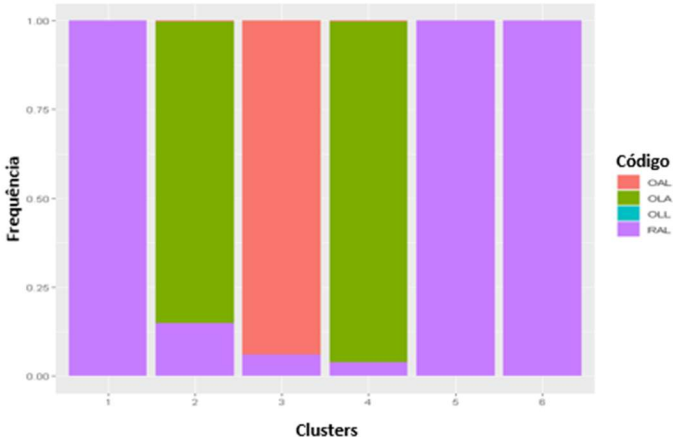


Figura 27 Tipo de serviço contratado em cada cluster

Analisando o mês do serviço de cada *cluster*, pela interpretação da Figura 28 é possível identificar que o *cluster* 5, identificado como sendo o grupo que menos valoriza aspetos relacionados com o conforto a bordo e que mais valoriza o espaço para transporte de bagagem, apresenta um padrão que o distingue dos restantes *clusters*. Neste *cluster* os serviços efetuados estão concentrados nos meses de primavera em oposição dos restantes *clusters* onde os serviços estão concentrados nos meses de verão.

Na Figura 29 é apresentado o top 10 locais de *dropoff* de cada *cluster.* É possível identificar que Vilamoura e Albufeira são os destinos preferidos dos clientes agrupados nos *clusters* 1,3,5 e 6. Os *clusters* 2 e 4 incluem os clientes que contratam serviços num único sentido do tipo *OLA,* em que o destino final é o aeroporto, este facto é visível no gráfico.

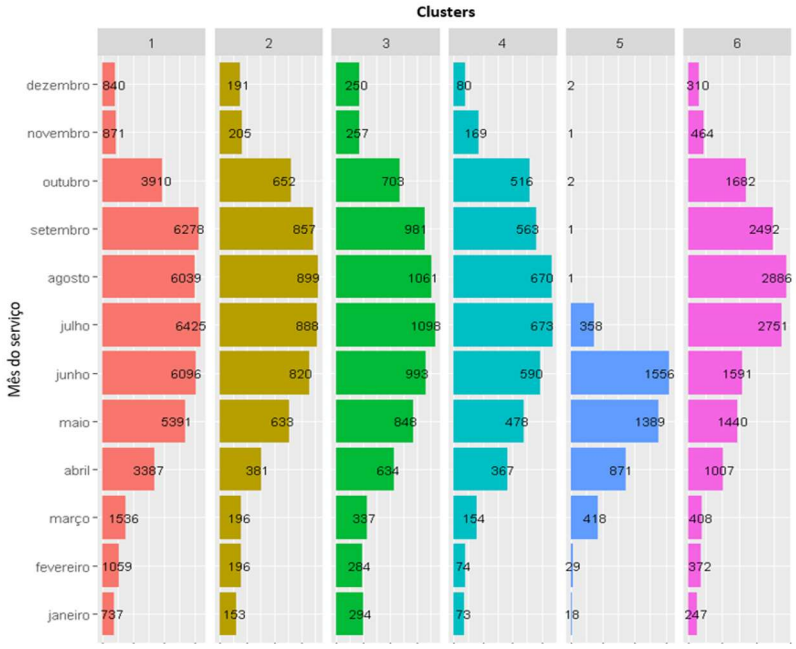


Figura 28 Mês de prestação do serviço em cada cluster a chegada

Figura 29 Local de dropoff dos clientes de cada cluster a chegada

Relativamente ao aeroporto de origem dos clientes, a Figura 30 representa o top 30 dos aeroportos por *cluster*. Em todos os *clusters* a maior parte dos clientes viajam a partir dos aeroportos de Dublin e London Gatwick. O *cluster* 3 diferencia-se dos demais, por incluir clientes com origem nos aeroportos de Estocolmo, Eindhoven e Amsterdam. Relembro que este foi o *cluster* que foi identificado como o que menos valoriza aspetos relacionados com a cortesia do condutor, tipo de condução e pontualidade na partida.

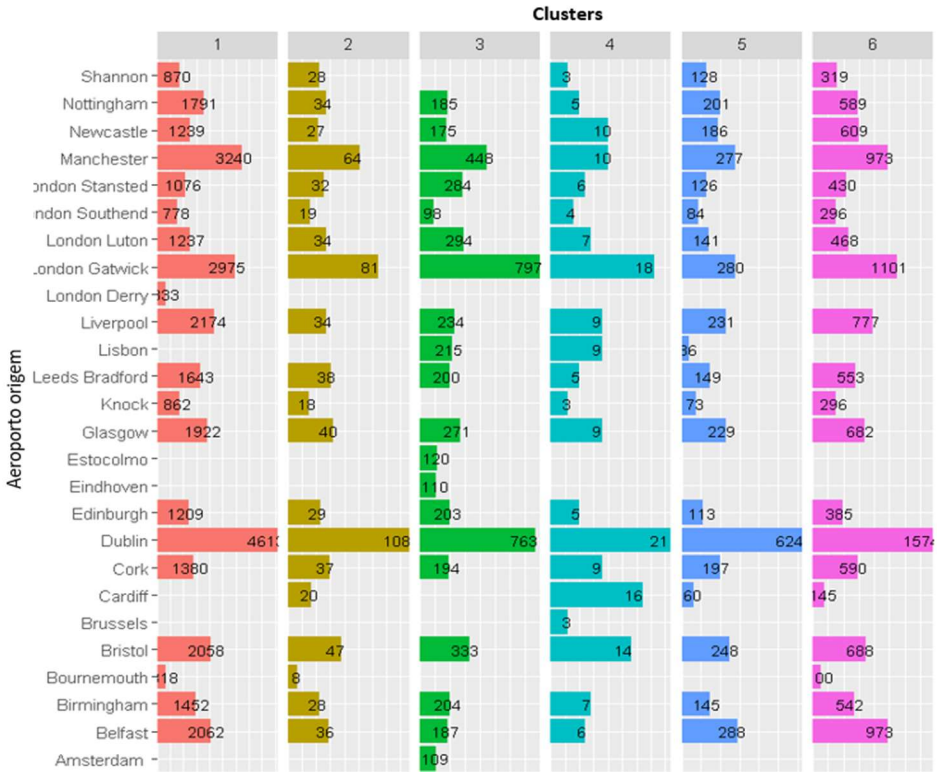


Figura 30 Aeroporto de origem do cliente

A Figura 31 representa os top 10 países de origem dos clientes de cada cluster. Em todos os *clusters* a maioria dos clientes provém do Reino Unido e da Irlanda.

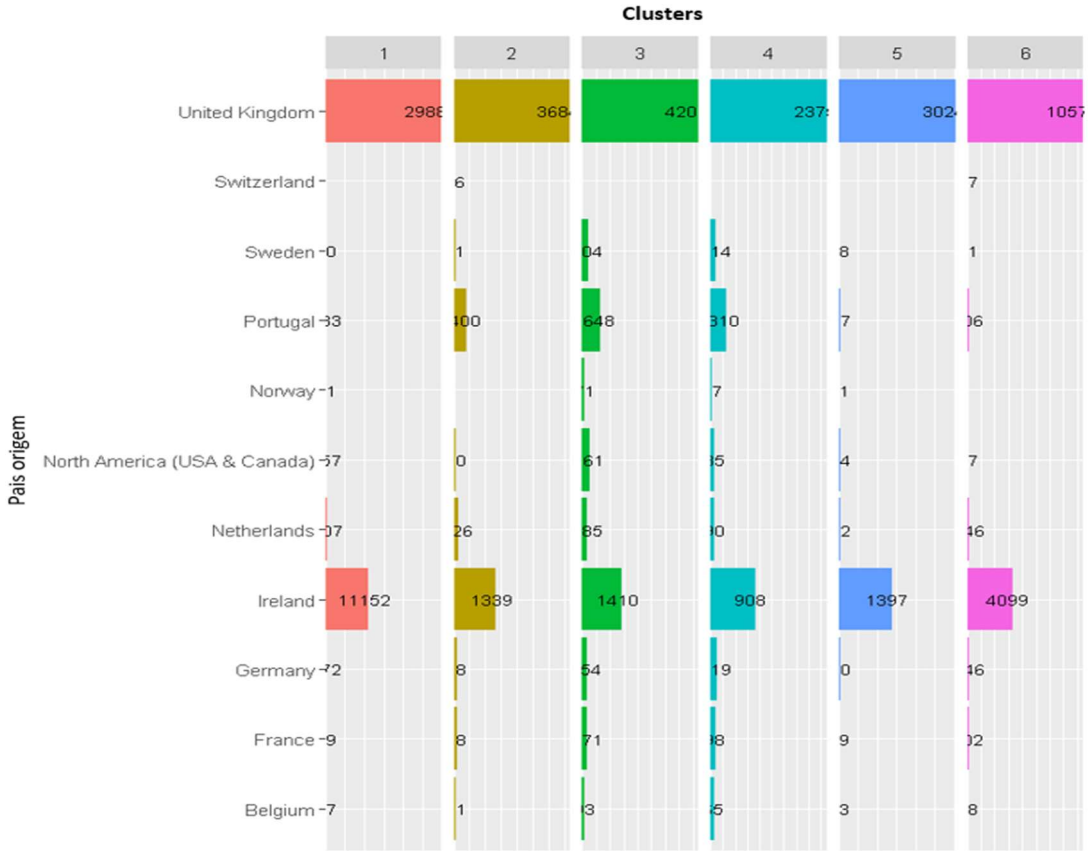


Figura 31 Pais de origem dos clientes de cada cluster

3.6.Padrões relativos aos tipos de clientes

Encontrar padrões relativos aos tipos de clientes remete para o conceito de segmentação de mercado/cliente, introduzido na secção 2.3.3 deste documento.

É sabido que na indústria do turismo a segmentação de mercado cria vantagem competitiva e permite a redução de custos na gestão da relação com os clientes [80]. Na gestão da relação com o cliente recorre-se a ferramentas tecnológicas como o *CRM,* onde a segmentação dos clientes é primordial para direcionar campanhas de marketing de forma eficaz para um segmento-alvo [80], [162]. Ainda, na secção 2.3.3 foram apresentados estudos sobre diferentes metodologias aplicadas para a segmentação de clientes onde grande parte dos investigadores, na área do turismo, optam pela metodologia orientada aos dados, recorrendo a algoritmos de cluster.

3.6.1 Preparação dos dados

Para encontrar padrões relativos aos tipos de clientes que procuram os serviços da empresa, recorreu-se a uma abordagem de segmentação orientada aos dados aplicando algoritmos de *cluster*.

A seleção de variáveis no contexto de aprendizagem não supervisionada constitui um desafio quando comparado com os métodos utilizados em aprendizagem supervisionada. Em aprendizagem supervisionada encontrar as variáveis relevantes consiste em determinar aquelas que melhor distinguem as classes identificadas por etiquetas. No contexto de aprendizagem não supervisionada, encontrar um subconjunto de variáveis continua a ser um passo importante, porque contribui para melhorar diferentes aspetos do processo de aprendizagem permitindo diminuir a complexidade computacional, reduzindo problemas de *performance,* criando modelos mais fáceis de interpretar e que tendem a fornecer melhor desempenho preditivo [163], [164].

Foram selecionadas e calculadas novas variáveis a partir dos dados existentes de forma a obter um conjunto de variáveis que mais se adequam ao problema. Ainda, baseando-se na literatura, foram calculadas as três variáveis do modelo *RFM, recency, frequency e monetary,* muito utilizadas para a segmentação de clientes em diferentes áreas de negócio [106], [109], [110], [108], [103], [104].

Tabela 8 Seleção de variáveis para a segmentação de clientes

Variável	**Tipo**	**Descritivo**	**Fonte**	**Método de Cálculo**
pickup	Categórica	Local de partida	*Manifesto*	
dropoff	Categórica	Destino	*Manifesto*	
country	Categórica	País de Origem	*Paises*	
code	categórica	Tipo do serviço: OAL; OLA; OLL; RAL; RLA; RLL e GOF	*Manifesto*	
arrival_month_name	categórica	Mês do Serviço	Manifesto	Calculado a partir da data do serviço (Manifesto.dia).
arrival_week_name	categórica	Dia da semana do serviço	Calculado	Calculado a partir da data do serviço (Manifesto.dia).
days_to_service	Numérico	Dias decorridos desde a reserva até ao dia do serviço	Calculado	Diferença entre a data do serviço e a data da reserva. *Manifesto.dia - Manifesto.booktimer*
adults	Numérico	Nº de adultos (>13 anos)	*Manifesto*	
children	Numérico	Nº de crianças (de 3 a 12 anos)	*Manifesto*	

Variável	Tipo	Descritivo	Fonte	Método de Cálculo
babies	Numérico	Nº de bebés (<3 anos)	*Manifesto*	
luggage	Numérico	Nº de bagagens de cabine e de porão transportado	Luggage	Soma de *Luggage.qtd* onde *cargo* é igual a *cabin_luggage* e *checked_luggage*
Child_buggy	Numérico	Nº de carrinhos de bebés transportados	Luggage	Soma de *Luggage.qtd* onde cargo é igual a Child_buggy
Golf_bags	Numérico	Nº de sacos de golf transportados	Luggage	Soma de *Luggage.qtd* onde cargo é igual a Golf_bags
frequency	Numérico	Frequência de serviços	Calculado	Contagem do nº de serviços de efetuados por um cliente até uma determinada dada
recency	Numérico	Nº de dias decorridos desde o último serviço	Calculado	Diferença entre a data atual e a data do serviço. Indica a antiguidade do serviço.
monetary	Numérico	Montante pago até a data	Calculado	Montantes pagos pelo cliente até uma determinada nada

Na Tabela 8 estão descritas as variáveis que foram consideradas para a análise de *clusters*, onde podem ser evidenciadas variáveis categóricas e numéricas. Os valores e variáveis foram obtidas recorrendo à linguagem *SQL* para juntar as tabelas *manifesto*, *luggage* e *países.*

Uma das características de um projeto de Extração de Informação reside no facto de que os dados podem ser de grande dimensão com diferentes tipos de variáveis. Essa caraterística pode ser evidenciada na Tabela 8, onde foram selecionadas variáveis categóricas e numéricas de um conjunto de dados com 132 745 observações. As estatísticas sumárias do conjunto de dados indicam a existência de 3 observações com valores omissos para a variável c*ountry* que foram removidos do conjunto de dados, passando a existir 132 742 observações.

Para o estudo das variáveis numéricas, o gráfico da Figura 32 apresenta na diagonal superior as correlações de *Spearman,* menos sensível a *outliers* quando comparado com o método de *Pearson* [95]. Na diagonal inferior é apresentada a dispersão das variáveis e na diagonal a densidade da distribuição normal.

O gráfico de dispersão revela a existência de outliers nos dados e as correlações entre as variáveis são baixas, exceto para *monetary* e *frequency* com uma correlação de 0,76 o que indica a existência de uma relação linear entre estas variáveis. A presença de *outliers* ou anomalias nos dados podem afetar a estrutura dos *clusters* [160], [95]. Para o tratamento de *outliers* foi utilizada a função *tsoutliers* da biblioteca *Forecast* [120], desenhado para detetar e sugerir valores para a substituição das anomalias detetadas.

Quando as variáveis numéricas estão em diferentes medidas, o cálculo da dissimilaridade entre os objetos pode ser afetada. Por esse motivo, com o objetivo de tornar as variáveis comparáveis é usual aplicar uma transformação na escala, standardizando as variáveis numéricas para variância unitária e media zero [95].

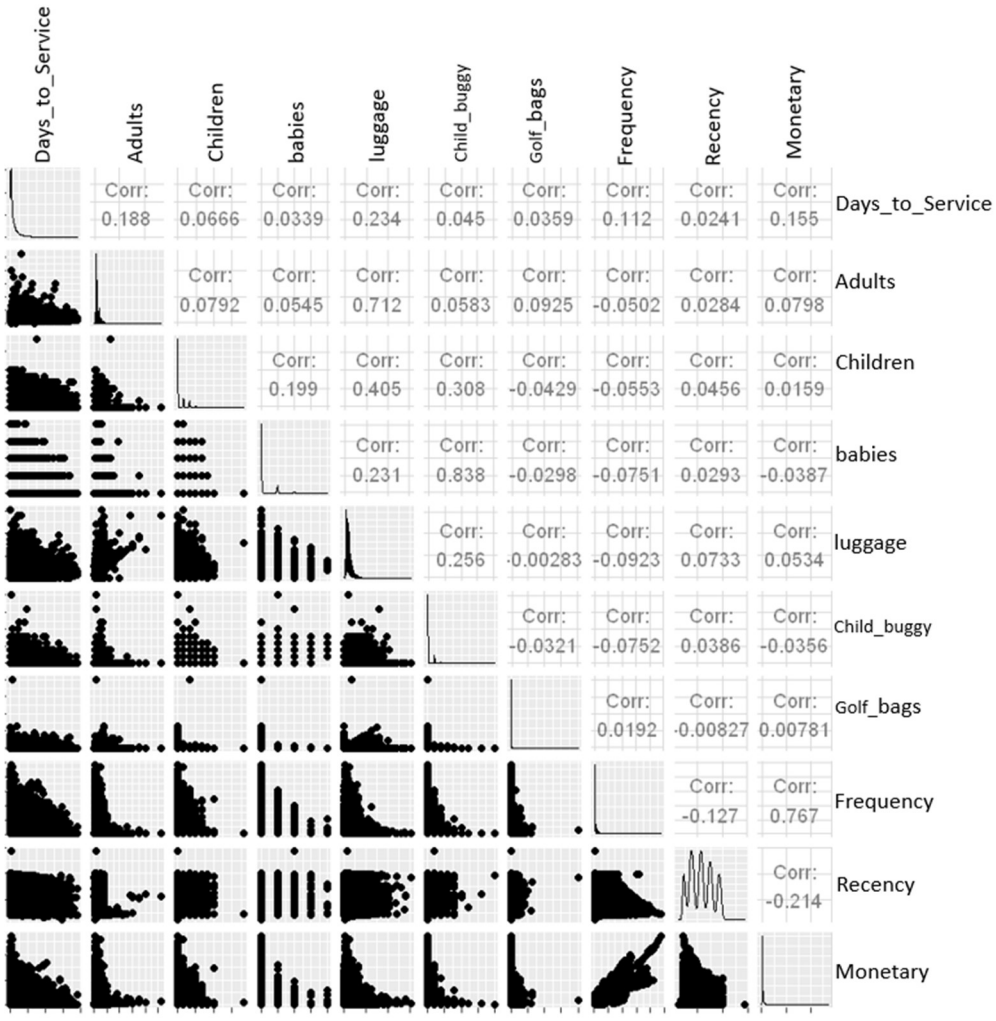


Figura 32 Estudo das variáveis numéricas

3.6.2 Construção e avaliação do modelo

A maior parte dos algoritmos de *clustering* existentes são capazes de tratar variáveis categóricas e numéricas, mas não são eficientes quando estamos perante uma grande quantidade de observações, ou, quando permitem processar grandes quantidades de dados estão limitados apenas às variáveis numéricas [165].

Para lidar com conjuntos de dados com diferentes tipos de variáveis, Ralambondrainy [166] apresentou uma abordagem em que as variáveis categóricas são codificadas como variáveis binárias, onde 1 indica que uma categoria está presente e 0 ausente. Após a codificação das variáveis categóricas em binárias, estas passam a ser consideradas como numéricas no algoritmo de *cluster*.

Hunag [165], critica a abordagem de Ralambondrainy [166] indicando como desvantagem a necessidade de ter tantas variáveis binárias quanto o número de categorias existente na variável original, conduzindo a um aumento do custo computacional e de espaço de memória. Outra desvantagem apontada pelo autor consiste no cálculo do valor médio dos *clusters* utilizando os valores 0 e 1 não representam as características dos *clusters*.

Para o conjunto de dados em análise, apresentado na Tabela 8, foi utilizado o algoritmo de cluster *k-prototypes* [167]. O algoritmo *k-prototypes* funciona como uma extensão do algoritmo *k-means* com o algoritmo *k-modes* [168] para incluir o processamento de variáveis categóricas. O algoritmo proposto por Huang [167], considera o quadrado da distância Euclidiana, s^r como medida de dissimilaridade para as variáveis numéricas e o número de categorias incompatíveis entre dois objetos, s^c, como medida de dissimilaridade para as variáveis categóricas. A medida de dissimilaridade entre dois objetos é obtida pela soma $s^r + \gamma s^c$ onde γ é o peso utilizado para equilibrar as partes de forma a evitar favorecer um dos tipos de variáveis. Huang [168] sugere a média do desvio padrão σ das variáveis numéricas como o valor a atribuir a γ.

Para determinar o número ótimo de *clusters* foi utilizado o *método do Cotovelo*, tendo como função objetivo a soma das distâncias dos objetos de cada *cluster,* e o método do *Coeficiente de Silhueta* que determina o grau de afastamento dos objetos de um *cluster.* Para a aplicação dos dois métodos, o algoritmo *k-prototypes* foi executado fazendo variar o número de *clusters* entre 1 e 10 e mantendo o valor de γ constante. Na Figura 33, o gráfico da esquerda utiliza a função objetivo e revela uma zona de pico (cotovelo) para 2

clusters. O gráfico da direita utiliza o coeficiente de silhueta e o valor mais elevado da silhueta indica que o valor ótimo de *K* para este método é 2. Assim, o valor escolhido para o parâmetro *K* foi 2.

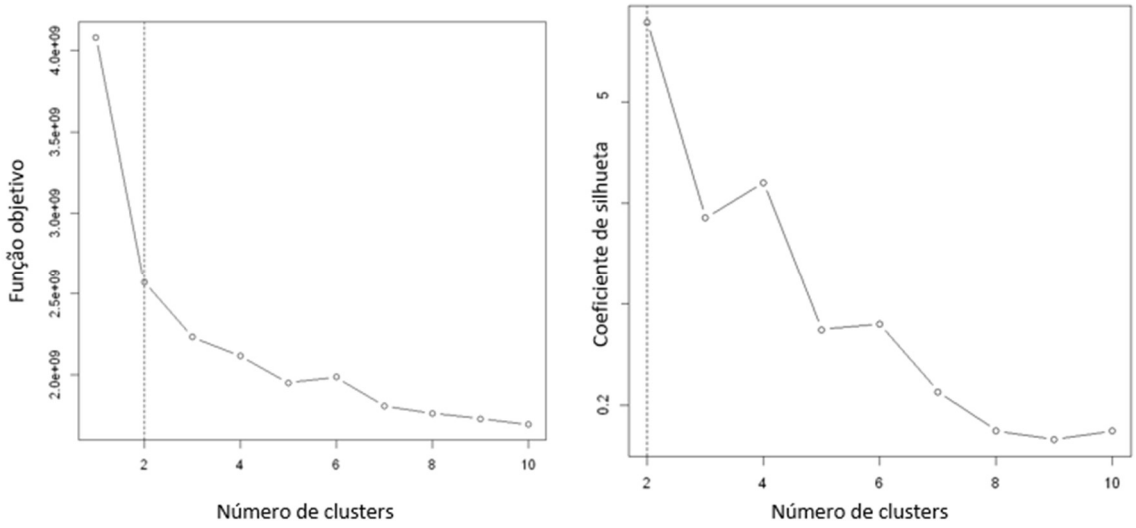


Figura 33 Número ótimo de cluster

3.6.3 Discussão dos resultados

O algoritmo *k-prototypes* foi executado considerando diferentes abordagens de tratamento de dados com o valor de k igual a 2 e mantendo o valor de γ constante. O valor de γ foi calculado recorrendo a função *lambdaest* que integra o pacote R *clustMixType* que implementa o algoritmo *k-prorotypes* [165]. A função *lambdaest* utiliza como heurística a média do desvio padrão de todas as variáveis numéricas.

A primeira abordagem, representada na Figura 63 do Apêndice E, consistiu na execução do algoritmo sem qualquer tratamento dos dados, apresentando um coeficiente de silhueta de 0,58. A segunda abordagem (Figura 64 do Apêndice E) consistiu na remoção de outliers e apresentou um coeficiente de silhueta de 0,63. Por último (Figura 65 do Apêndice E), foi aplicada a remoção de *outliers* e as variáveis numéricas foram standardizadas para variância unitária e media zero, fazendo cair o valor do coeficiente de silhueta para 0,47.

A segunda solução onde foi efetuada a remoção de *outliers*, foi selecionada como a solução final, por ter sido aquela que apresentou o maior valor de coeficiente de silhueta. Pela interpretação da Figura 34 que representa os centroides dos *clusters* foi possível

concluir que os clusters apresentam os mesmos resultados, em termos das variáveis categóricas, *pickup*, *dropoff*, *country* e *código*. As variáveis numéricas permitiram diferenciar os *clusters* nos seguintes aspetos:

- O segmento 1 (*cluster* 1) foi formado por 63 4884 observações e representa os clientes que efetuaram as reservas com pouca antecedência (*Days_to_Service*) e incluem o menor número de adultos, crianças e bebés. Este segmento ainda é caracterizado por incluir os clientes frequentes (*Frequency*) logo com maior valor monetário para a empresa. Este segmento ainda caracteriza-se por incluir os clientes que transportam sacos de golf e preferem viajar na primavera.
- O segmento 2 (*cluster* 2) é composto por 69 258 observações. Representa os clientes menos frequentes e com menor valor monetário para a empresa e que efetuaram as reservas com alguma antecedência, quando comparados com o segmento 1. Têm referência em viajar no verão e representam grupos maiores tanto em termos de adultos, crianças e bebés transportados.

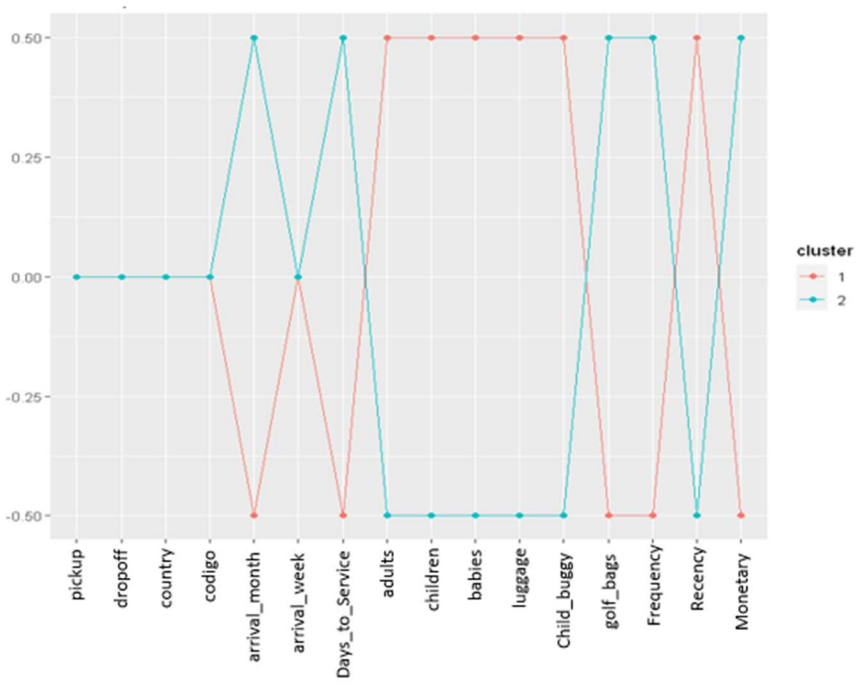


Figura 34 Visualização dos centroides de cada cluster obtidos pelo algoritmo K-prototype

A Figura 35 detalha a distribuição do tipo de serviço (variável *código*) que os clientes escolhem em cada um dos segmentos. Em ambos os segmentos, os clientes utilizam com

maior frequência o tipo de serviço *RAL* seguido do *OAL*, sendo que o serviço *RLA* apenas é escolhido pelos clientes do segmento 2.

A existência de um maior número de serviços *RAL* e *OAL* implica que o aeroporto será o local de *pickup* da maior parte dos serviços dos dois segmentos, conforme pode ser confirmado na Figura 36 que apresenta os top 10 locais de *pickuip* e *dropoff*. A distribuição do local de *dropoff* é idêntica nos dois segmentos, sendo os locais de Albufeira e Vilamoura os mais frequentes.

Figura 35 Visualização do código do serviço de cada cluster

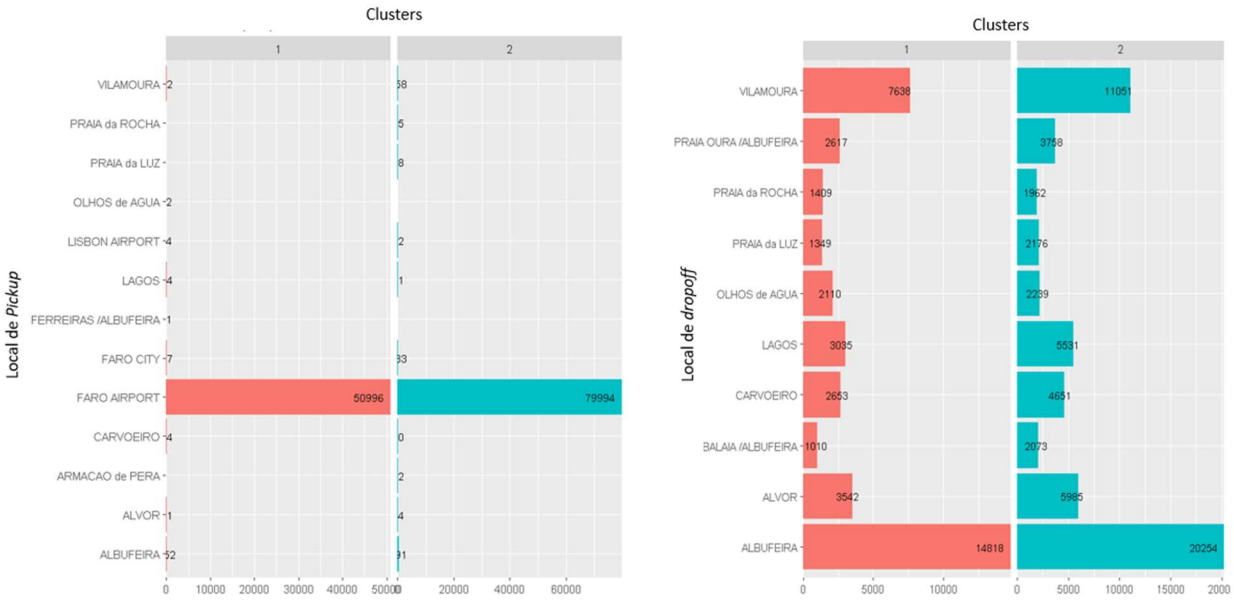


Figura 36 Visualização do local de pickup e dropoff à chegada de cada cluster

A Figura 37 apresenta o top 10 dos países por segmento. A distribuição é idêntica nos dois segmentos, com o maior número de clientes com proveniência no Reino Unido e na Irlanda. Em ambos os segmentos os serviços estão distribuídos ao longo da semana, mas na Figura 38 constata-se que no segmento 1 existe um ligeiro aumento da quantidade de serviços na quinta-feira, no sábado e no domingo e no segmento 2 verifica-se um pico no sábado.

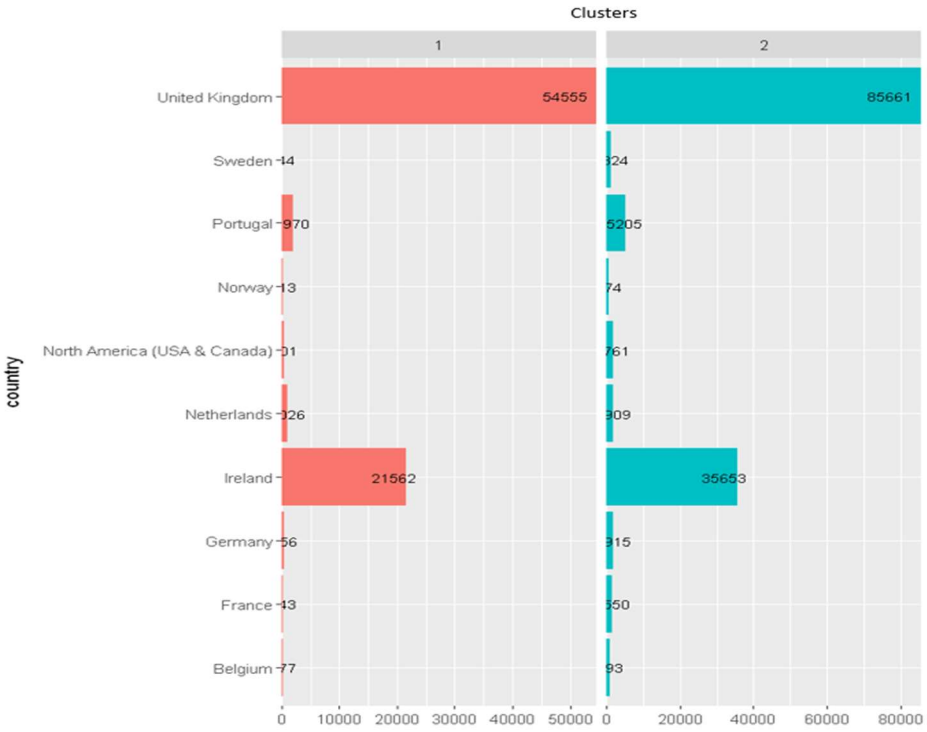


Figura 37 Visualização do país de origem dos clientes de cada cluster

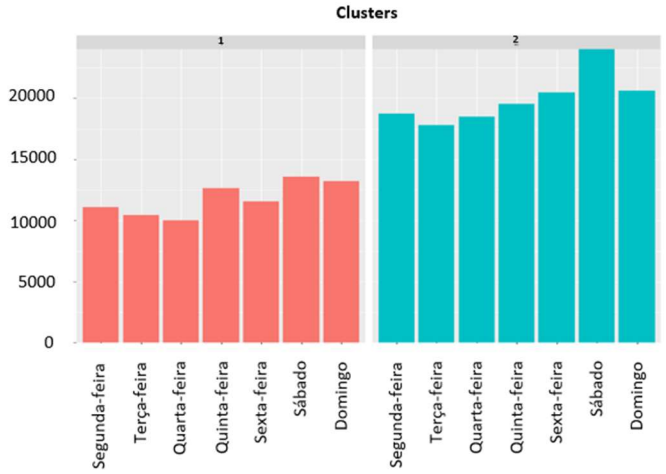


Figura 38 Visualização do dia da semana do serviço de cada cluster

A Figura 39 representa o mês escolhido pelos clientes de cada *cluster* para viajar. O segmento 1 identifica o grupo de clientes que preferem viajar nos meses de verão e o segmento 2, os clientes que viajam mais nos meses da primavera e outono.

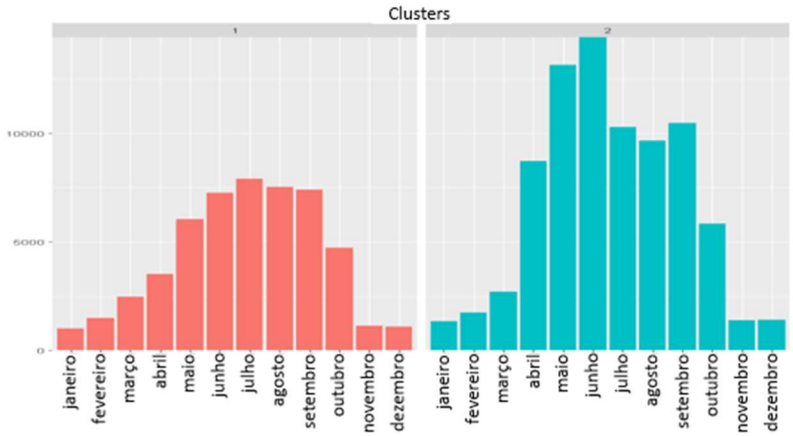


Figura 39 Visualização do mês do serviço de cada cluster

A análise efetuada anteriormente permitiu, pela aplicação de algoritmos cluster, identificar dois segmentos de clientes, dando resposta a questão de investigação que procurava compreender os tipos de clientes que procuram os serviços da empresa. Ainda, a análise exploratória das características de cada segmento, revelou padrões no comportamento de cada grupo que se julga de grande importância para os decisores de negócio. A interpretação das características de cada segmento pelos decisores de negócio pode servir de suporte à criação de campanhas de marketing de forma mais eficaz e permitindo uma melhor gestão da relação com o cliente, trazendo assim vantagem competitiva para a empresa.

Capítulo 4 – Avaliação de possíveis fatores que influenciam a procura

No capítulo anterior foi efetuado o processo de Extração de Informação dos dados existentes no histórico de transações da empresa, usando a metodologia CRISP-DM, para responder as duas primeiras questões de investigação.

O presente capítulo procura dar resposta a terceira questão de investigação cujo objetivo é o de compreender que fatores podem afetar as reservas efetuadas e, consequentemente, as receitas da empresa. Para atingir esse objetivo, foi seguida a mesma metodologia para a Extração de Informação do capítulo anterior. Nesta análise, os dados do histórico de transações foram conjugados com informação de campanhas digitais da empresa e com dados obtidos de fontes externas.

4.1 Preparação dos dados

Os dados existentes no histórico de transações da empresa podem captar alguns dos fatores que influenciam a procura como, a data do serviço o que pode indicar a existência de um padrão de sazonalidade, ou o número de cliques em resultados de pesquisas que redirecionam para o site de reservas da empresa, consequência de campanhas de marketing digital do *Google Ads*. Todavia, podem ainda existir fatores provenientes de fontes de dados externos com influência na procura tais como, a informação dos feriados nacionais no país de origem e de destino que podem indicar que os clientes aproveitam os fins-de-semana longos para viajar. Importa salientar que os feriados nacionais foram extraídos do website de acesso gratuito *TimeAndDate.com* aplicando técnicas de web scraping. Foram ainda selecionadas um conjunto de variáveis relacionadas com as campanhas do *Google Ads*. A decisão de quais as variáveis das campanhas a incluir foi baseada na literatura existente (vide Tabela 1). A informação das campanhas *Google Ads* foi conjugada com a informação das reservas (*Manifesto*) formando um novo conjunto de dados, apresentado na Tabela 9.

Os dados foram agregados semanalmente, considerando que a previsão semanal da procura é de grande importância para o apoio do planeamento e alocação dos condutores e dos veículos de forma a adequar os recursos a procura, com a devida antecedência.

Tabela 9 Possíveis variáveis que influenciam a procura dos Serviços

Variável	Descritivo	Fonte	Método de Cálculo
hollidays_origem	Feriados no país de origem	*TimeAndDate.com*	*web scraping*
holidays_pt	Feriados em Portugal	*TimeAndDate.com*	*web scraping*
Campanhas_ativas	Número de campanhas ativas	Google Analytics	Número médio de campanhas ativas por semana
cliques	Número de cliques numa campanha	Google Analytics	Número médio de cliques por semana
dias_campanha_activo	Número de dias em que a campanha esteve ativa	Google Analytics	Número médio de dias que a campanha esteve ativa
palavras_chave	Número de palavras chave de uma campanha	Google Analytics	Número médio de palavras chave nas campanhas por semana
Impressoes	Número de vezes que as palavras chave das campanhas são apresentadas em resultados de pesquisas	Google Analytics	Número Máximo de impressões da semana
Custo	Montante gasto em campanhas	Google Analytics	Montante gasto em campanhas por semana
CPC	Custo por clique no resultado de uma pesquisa por uma palavra chave de uma campanha.	Google Analytics	CPC Máximo da semana
CTR	Rácio entre os cliques e as impressões.	Google Analytics	CTR médio da semana
Totalservicos	Total de serviços	Manifesto	Total de serviços da semana

No capítulo anterior, foi possível verificar que a procura apresenta um padrão que vai de encontro à sazonalidade do turismo na região do Algarve e onde se pôde concluir que a sazonalidade se constitui como um dos principais fatores que influenciam a procura.

O objetivo dos modelos apresentados nesta secção é prever a variável quantitativa, *totalServicos*, que agrega a procura semanal dos serviços, recorrendo a um conjunto de fatores que possam afetar o comportamento da variável que se pretende prever. O objetivo proposto configura um problema de aprendizagem supervisionada, onde podem ser aplicados algoritmos de regressão múltipla ou métodos de análise multivariada de series temporais.

A variável objetivo, *totalServicos,* foi analisada na perspetiva de uma serie temporal. Os dados das séries temporais podem ser divididos em diferentes componentes. Cada

componente representa um padrão subjacente que pode ser utilizado como um preditor. Quando a série é suficientemente longa, com mais de um ano de informação, então poderá ser necessário efetuar uma decomposição para analisar a sazonalidade anual, e semanal [120].

Para decompor a série temporal foi utilizado o método *STL (Seasonal and Trend decomposition using Loess)*, *Loess* é um método para estimar relações não lineares [169]. O *STL* é considerado robusto, por conseguir lidar com qualquer tipo de sazonalidade e pela presença de *outliers* não afetarem a obtenção das componentes sazonais e da tendência.

Na Figura 40 a série temporal foi decomposta em quatro componentes: (1) uma componente que indica a tendência; (2) uma componente sazonal semanal; (3) uma componente sazonal anual; e (4) uma componente remanescente (contendo outro tipo de informação que não se enquadra na série temporal).

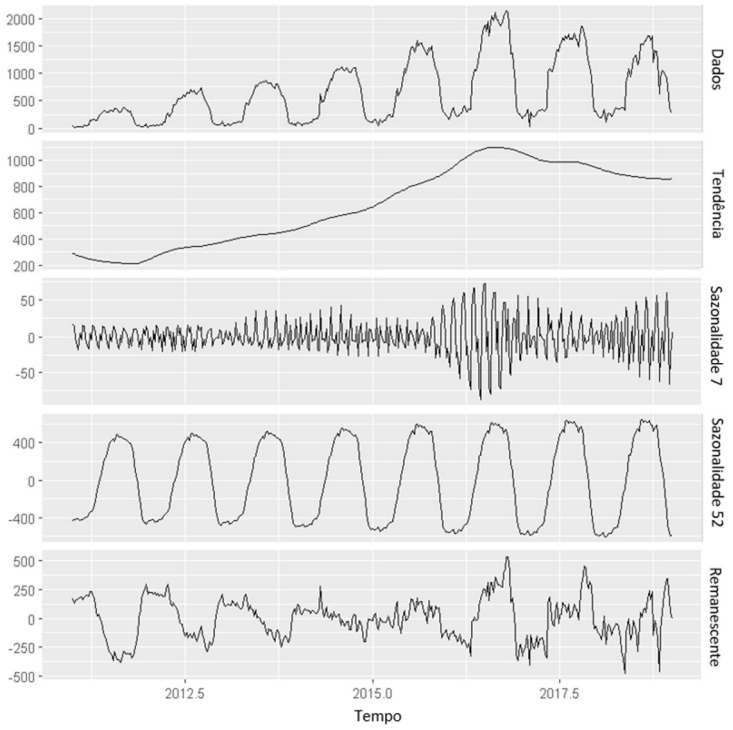


Figura 40 Decomposição da procura semanal e exploração de multi-sazonalidade

Para além da sazonalidade anual, a decomposição da série temporal apresenta um padrão que sugere a existência de sazonalidade semanal. A sazonalidade semanal pode

ser explicada pelos voos semanais, provenientes do Reino Unido e da Irlanda que representam 89% dos clientes da empresa, serem suscetíveis de afetar a procura.

Se a série temporal apresentar múltiplos períodos sazonais, uma alternativa pode ser a utilização dos termos de *Fourier* como preditor. Jean-Baptiste Fourier foi um matemático francês, nascido em 1700 que demonstrou que uma série de termos seno e cosseno pode aproximar-se de qualquer função periódica e podem ser utilizados para obter padrões sazonais.

A Figura 41 apresenta a análise das correlações entre as variáveis selecionadas. Os *outliers* identificados foram removidos recorrendo-se à função *tsoutliers* da biblioteca *Forecast* [120].

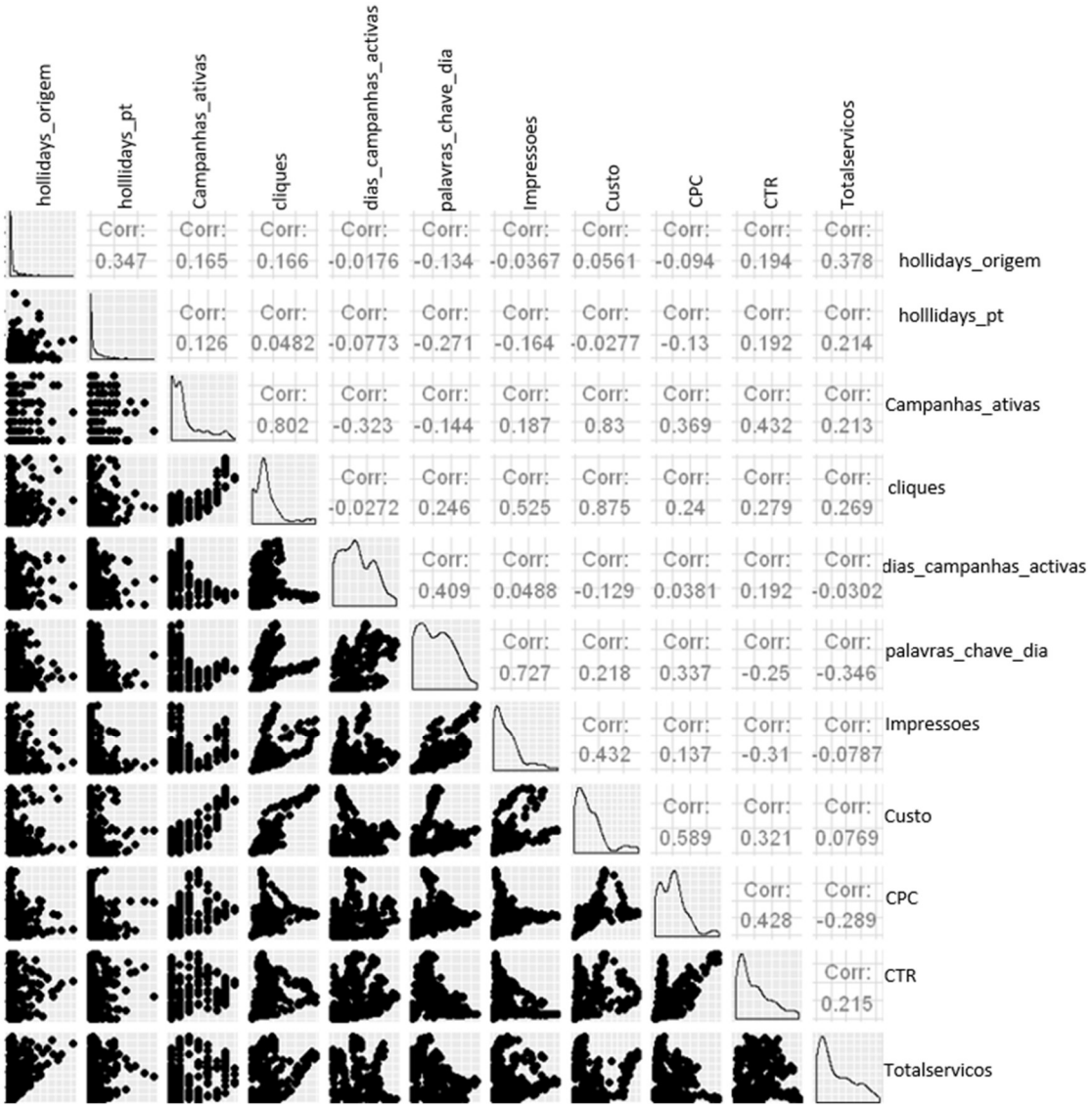


Figura 41 Estudo das correlações

Tal como a correlação mede a relação linear entre duas variáveis, a autocorrelação, ou ACF, mede a relação linear entre os valores verificados em diferentes momentos de uma série temporal. O ACF mostra as autocorrelações entre Y_t e Y_{t-k} para diferentes valores de k. Se Y_t e Y_{t-1} estão correlacionados, então Y_{t-1} e Y_{t-2} também devem estar correlacionados. No entanto, Y_t e Y_{t-2} podem estar correlacionados, simplesmente, porque ambos estão relacionados com Y_{t-1} e não por existir nova informação em Y_{t-2} que possa ser utilizada na previsão yt. Para ultrapassar este problema, podem ser utilizadas as autocorrelações parciais ou *PACF*.

O *PACF* mede a relação entre Y_t e Y_{t-k}, após a eliminação dos efeitos dos valores intermédios. Assim, a primeira autocorrelação parcial é idêntica à primeira autocorrelação, porque não há nada entre elas a remover.

As séries temporais que não mostram autocorrelação são denominadas de ruído branco (*white noise*) ou séries estacionárias [120]. Uma série temporal estacionária é aquela cujas propriedades não dependem do momento em que a série é observada. As séries temporais que apresentam padrões de tendências ou sazonalidade não são estacionárias, porque a tendência ou a sazonalidade afetarão o valor das séries temporais em momentos diferentes. Por outro lado, uma série de ruído branco é estacionária, porque deve ter o mesmo especto em qualquer altura que for observado [120].

Para a construção de modelos de previsão com séries temporais, um dos pressupostos é que estas devem de ser estacionárias, caso contrário é necessário efetuar transformações nos dados para estabilizar a variância. Box e Jenkins [170] sugeriram a utilização da transformação *BoxCox* para validar não só a hipótese da variância constante, bem como todos os pressupostos subjacentes aos modelos *ARIMA*.

Na Figura 42 é apresentado o *ACF* e o *PACF* após a aplicação da transformação de *BoxCox* para obter uma série estacionária. As linhas azuis tracejadas indicam se as correlações são significativamente diferentes de zero. Quando os dados apresentam uma tendência, as autocorrelações tendem a ser grandes e positivas para as observações próximas no tempo. Assim, as ACF das séries temporais que apresentam uma tendência, tendem a ter valores positivos que diminuem lentamente à medida que as observações se vão distanciando.

Quando os dados são sazonais, as autocorrelações tendem a ser maiores, se os dados apresentam ao mesmo tempo uma tendência e sazonalidade é possível ver uma combinação destes efeitos.

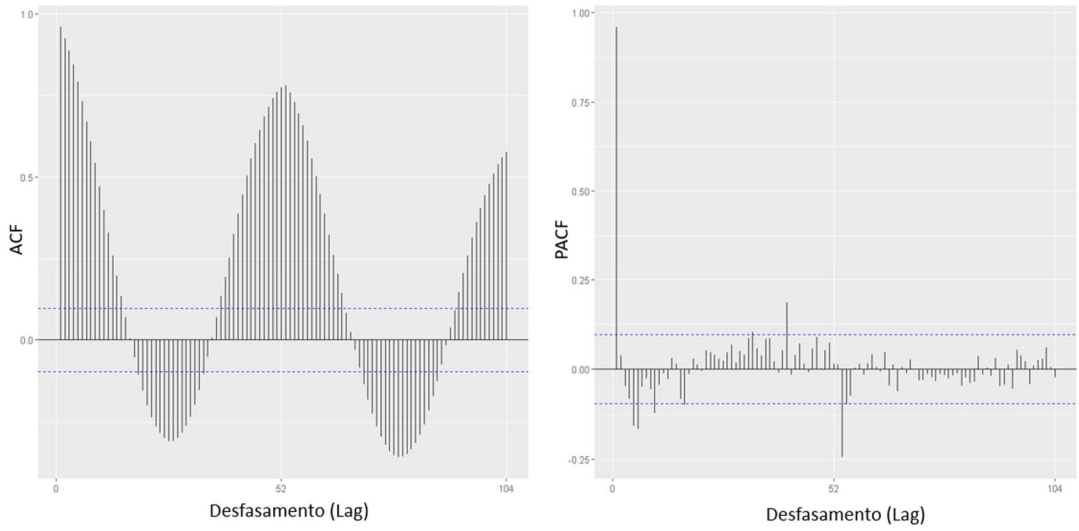


Figura 42 ACF e PACF da procura semanal dos serviços, com a transformação de Box Cox

4.2 Construção e avaliação dos modelos

4.2.1. Modelos de regressão de Series Temporais

Quando se utiliza um modelo de regressão linear, são assumidos alguns pressupostos sobre as variáveis da equação. Assume-se que a relação entre a variável dependente e as variáveis preditoras satisfaz uma equação linear. Não deve existir multicolinearidade, ou seja, uma relação linear entre duas ou mais variáveis independentes. Os *outliers* devem ser pouco significativos, sob pena de influenciar negativamente a qualidade do modelo, reduzindo o significado estatístico dos resultados. São assumidos os seguintes pressupostos sobre resíduos (erros):

- seguem, aproximadamente, uma distribuição normal, com média de 0. Ou seja, as diferenças entre o modelo estimado e os dados observados são muito reduzidas, caso contrário, as previsões seriam sistematicamente enviesadas.
- não estão autocorrelacionadas; caso contrário, as previsões seriam ineficientes, pois há mais informação nos dados que podem ser explorados.

- não estão relacionadas com as variáveis preditoras; caso contrário, haveria mais informação que deveria ser incluída na modelo.

Um dos principais requisitos na construção de modelos computacionais é que estes tenham uma elevada capacidade preditiva e de generalização. A fraca capacidade de generalização pode ser caracterizada como *over-fitting*, em que o modelo memoriza os dados de treino e não é capaz de produzir resultados corretos para dados nunca vistos. Uma técnica para evitar o *over- fitting* consiste na validação cruzada (*cross-validation*), que requer a divisão dos dados numa componente de testes e de treino [171]. A validação cruzada aplica-se quando os dados são independentes e distribuídos de forma idêntica. Na literatura foram propostas algumas variantes de validação cruzada para dados dependentes, como é o caso das séries temporais. No entanto, surgem questões relacionados com a dependência entre as observações que podem conduzir a modelos com uma fraca capacidade de generalização para novas observações [172].

Para lidar com a dependência entre observações, o modelo é testado com observações futuras utilizando a técnica *Out-of-sample* (OOS). Esta técnica divide os dados em duas partes: um período inicial onde o modelo é treinado, e um período de testes reservado para validar o modelo [135]. Esta abordagem é retratada na Figura 43, onde foi reservado o período de 2012 a 2017 para treino e 2018 e 2019 para testes.

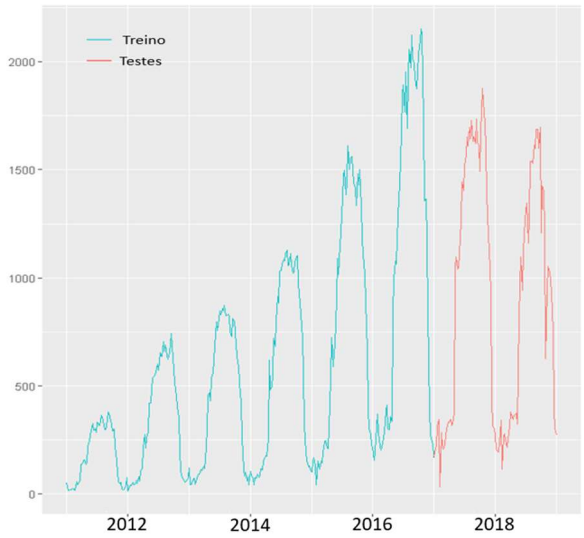


Figura 43 Divisão (Out-of-sample) dos dados de treino e testes

Para a construção do modelo de regressão foi necessário definir uma estratégia para selecionar os preditores a incluir no modelo. Uma abordagem usual consiste em utilizar a matriz de correlações para identificar as variáveis que estão correlacionadas com a variável dependente. Se não existir uma relação percetível, a variável não deve ser incluída no modelo [120]. Outra abordagem seria incluir todos os preditores no modelo e remover as variáveis com o valor estatístico *p-value* superior a 0,05. Constata-se que o significado estatístico nem sempre indica o valor preditivo, o valor do *p-value* não tem em conta, por exemplo, a multicoliniaridade [120].

A estratégia utilizada, que também recorre a matriz de correlações, consiste em identificar as relações de multicoliniaridade entre as variáveis preditoras. No caso de estudo, as variáveis "*custo*", *campanhas_activas* e "*cliques*", na Figura 41, apresentam correlações elevadas, o que é um indicador de existência de multicoliniaridade. Por esse motivo, não foram incluídas em conjunto no mesmo modelo. A existência de correlações altas entre os preditores é um indicador de redundância no sentido de que não acrescentam informação ao modelo. Foram ainda utilizadas medidas de precisão preditiva dos modelos com diferentes combinações de preditores. As medidas de precisão preditiva utilizadas foram o R^2 ajustado, o *AIC* (ou Akaike's Information Criterion) e o *BIC* (Schwarz's Bayesian Information Criterion). Para as medidas *AIC*, e *BIC*, foi selecionado o modelo com o valor mais baixo; para o R^2ajustado, foi selecionado o modelo com o valor mais elevado. Estas medidas e a sua aplicação na seleção de preditores em séries temporais foram descritas por Hyndman e Athanasopoulos [120].

A Tabela 10 apresenta diferentes modelos onde foi adotada a estratégia *forward stepwise* para a seleção dos preditores, em que as variáveis preditoras vão sendo adicionadas aos modelos e são mantidas aquelas que introduzem melhorias ao modelo, caso contrário a variável não é considerada. Esta tabela foi ordenada e apresenta o melhor conjunto de preditores na última linha.

Para além das variáveis indicadas na Tabela 9, foram ainda incluídas as variáveis tendências e sazonalidade, obtidas recorrendo ao método *STL*. Para capturar múltiplos períodos sazonais foram utilizados termos de *Fourier*, passando a existir dois pares de termos seno e cosseno, com o objetivo de representar a sazonalidade semanal, *S1-7, C1-7, S2-7, C2-7* e quatro pares para representar a sazonalidade anual *S1-52, C1-52, S2-52, C2-52, S3-52, C3-52, S4-52, C4-52*.

As variáveis "*Holliday na origem*" e "*Holliday no destino*" não introduziram melhorias ao modelo, pelo que não foram consideradas. As variáveis "*Tendência*" e "*Sazonalidade anual e semanal*" foram obtidas pela decomposição da variável "*totalServicos*", conforme descrito no início desta secção. Como constatado anteriormente, as variáveis "*custo*", "*campanhas_activas*" e "*cliques*" podem apresentar problemas de multicoliniaridade, pelo que apenas foi considerada a variável custo no modelo final, *M13*.

Tabela 10 Seleção de Preditores

	Preditores												Medidas de Precisão		
Modelos	Tendência	Sazonalidade anual e semanal (S1-7 a C4-52)	Holliday na origem	Holliday no destino (pt)	campanhas ativas	Dias campanhas ativas	Palavras chave dia	Cliques	Impressoes	Custo	CPC	CTR	AIC	BIC	R2 Ajustado
M1	x												207,268	218,507	0,300
M2	x	x											-440,261	-384,068	0,915
M3	x	x	x										-440,245	-380,305	0,915
M4	x	x		x									-438,263	-378,324	0,914
M5	x	x			x								-446,676	-386,737	0,917
M6	x	x			x	x							-476,046	-412,360	0,924
M7	x	x			x	x	x						-488,092	-420,660	0,928
M8	x	x			x	x	x	x					-489,695	-418,517	0,928
M9	x	x			x	x	x	x	x				-495,369	-420,445	0,930
M10	x	x			x	x	x	x	x	x			-533,917	-455,247	0,938
M11	x	x			x	x	x	x	x	x	x		-532,084	-449,668	0,938
M12	x	x			x	x	x	x	x	x	x	x	-546,881	-460,718	0,941
M13	x	x				x	x		x	x	x	x	-549,215	-470,545	0,941

Na análise de séries temporais é provável que o valor de uma variável, observada no período atual, seja semelhante ao seu valor no período anterior, sendo comum existir autocorrelação nos resíduos. Neste caso, o modelo estimado viola o pressuposto de não haver autocorrelação nos erros, resultando em previsões pouco eficientes porque existe

informação que não está a ser considerada no modelo, para serem obtidas melhores previsões.

Para avaliar os pressupostos assumidos sobre os resíduos, a Figura 44 apresenta diferentes gráficos que contêm a informação sobre os resíduos do modelo M13. O primeiro gráfico da figura mostra que os resíduos não apresentam um padrão visível, podendo concluir que não é significativamente diferente de um ruído branco. O histograma dos resíduos, apresentado no gráfico da direita, aproxima-se de uma distribuição normal. Da mesma forma, da observação do *Q-Q Plot*, na Figura 66 do Apêndice F, verifica-se o pressuposto de normalidade pois os resíduos estão aproximadamente em linha reta. No entanto, o gráfico da esquerda aponta para a existência de autocorrelação nos resíduos, indicando a existência de informação que o modelo não identificou.

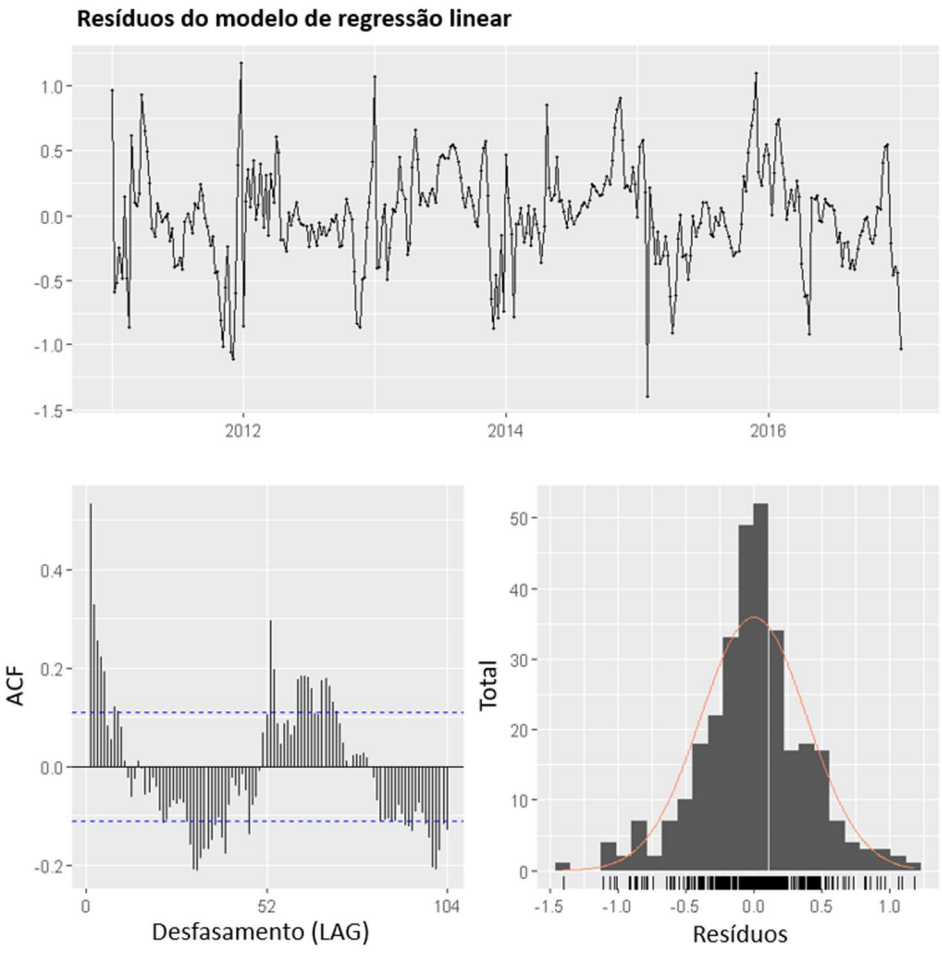


Figura 44 Avaliação do modelo de regressão (interpretação dos resíduos)

A análise dos resíduos não constitui um indicador fiável dos erros reais do modelo preditivo. É de salientar que os erros reais do modelo diferem dos resíduos, porque estes são calculados sobre o conjunto de dados, enquanto que os erros são calculados sobre o conjunto de teste. Neste sentido, a precisão das previsões é determinada avaliando o desempenho do modelo nos dados de teste que nunca foram "vistos" pelo modelo.

A Figura 45 apresenta o resultado da utilização do modelo para efetuar previsões, recorrendo aos dados guardados para teste, onde se pode verificar que o modelo conseguiu capturar o padrão de sazonalidade e a tendência, produzindo bons resultados de previsão dos dados de teste.

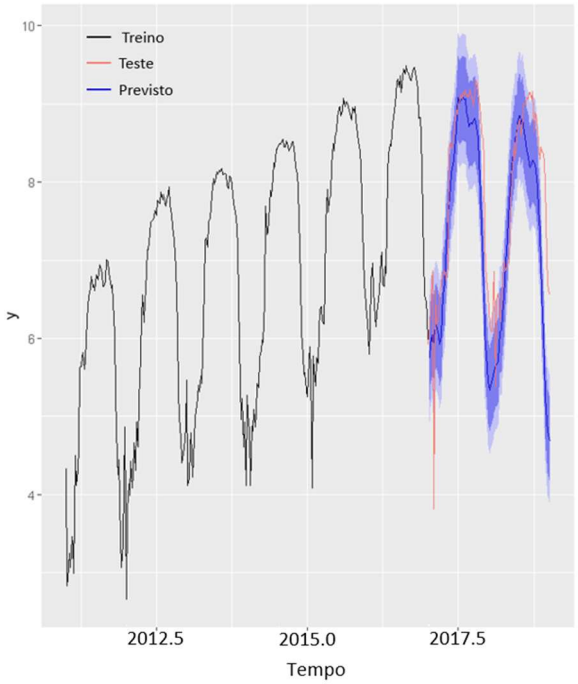


Figura 45 Previsões utilizando os dados de testes (intervalo de confiança de 80 a 95 %)

Na Tabela 11 apresenta os parâmetros do modelo. De salientar que se fosse seguida a abordagem de incluir todos os preditores no modelo e remover as variáveis com o valor estatístico *p-value* superior a 0,05, os parâmetros referentes a sazonalidade semanal e a variável "*palavras chave dia*" e "*custo"* seriam removidos do modelo.

Tabela 11 Parâmetros do modelo de regressão

Coeficientes	Estimado	Std. Error	Estatística-t	Pr (>|t|)
(Intercept)	4,278216	0,1634636	26,172	< 2e-16 ***
tendência	0,0032783	0,0001628	20,133	< 2e-16 ***
seasonFourierS1-7	0,0056416	0,0322678	0,175	0,86133
seasonFourierC1-7	0,0278843	0,0322086	0,866	0,38734
seasonFourierS2-7	-0,013557	0,0321516	-0,422	0,67359
seasonFourierC2-7	-0,024348	0,032164	-0,757	0,44966
seasonFourierS1-52	-0,864629	0,0356039	-24,285	< 2e-16 ***
seasonFourierC1-52	-1,583684	0,0422457	-37,487	< 2e-16 ***
seasonFourierS2-52	-0,253048	0,0324432	-7,8	1,10e-13 ***
seasonFourierC2-52	-0,314474	0,0341232	-9,216	< 2e-16 ***
seasonFourierS3-52	0,1412667	0,0329088	4,293	2,40e-05 ***
seasonFourierC3-52	-0,17849	0,0323966	-5,51	7,88e-08 ***
seasonFourierS4-52	0,1351961	0,0324488	4,166	4,08e-05 ***
seasonFourierC4-52	-0,060171	0,0323613	-1,859	0,06398 .
dias_campanha_activo	0,0006244	0,0001457	4,286	2,47e-05 ***
palavras_chave_dia	-0,00911	0,0095762	-0,951	0,34222
Impressoes	0,0005769	0,0001845	3,127	0,00195 **
Custo	0,0052253	0,0028596	1,827	0,06868 .
CPC	1,5018337	0,2202459	6,819	5,24e-11 ***
CTR	-1,490906	0,2978128	-5,006	9,60e-07 ***
Códigos de significância(p-value): 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

4.2.2. Modelos de regressão dinâmicos (modelo *ARIMA* com preditores)

Os modelos de regressão da secção anterior permitem a inclusão de informação relevante das variáveis preditoras, mas não capturam a dinâmica que os modelos *ARIMA* podem obter das séries temporais.

Nesta secção do documento foi considerada a utilização dos modelos *ARIMA,* de modo a permitir a inclusão de outras informações nos modelos, como, a autocorrelação existente nos dados, para capturar o dinamismo das séries temporais. Foi utilizada a função *auto.arima()* da linguagem R que consiste na implementação de uma variação do algoritmo Hyndman e Khandakar [173] cujo objetivo é o de encontrar um modelo de regressão com erros *ARIMA,* quando são incluídos regressores exógenos no modelo. Os argumentos do modelo são denotados como *ARIMA (p, d, q)* em que o parâmetro *p* é a ordem do modelo auto-regressivo, *d* é o grau de diferenciação e *q* a ordem do modelo de média móvel. A função *auto.arima()* calcula os argumentos *p*,*d*,*q* do modelo *ARIMA*

minimizando o *AIC* depois de diferenciar os dados *d* vezes. Em vez de considerar todas as combinações possíveis de *p* e *q*, o algoritmo utiliza uma abordagem *stepwise* para percorrer todo o espaço do modelo.

Utilizando os mesmos preditores, identificados no modelo de regressão, M13 foi possível verificar pela interpretação do gráfico *ACF,* apresentado na Figura 46, que a autocorrelação dos resíduos diminuiu consideravelmente. Como se pode verificar, apenas são apresentados quatro períodos que apresentam autocorrelação.

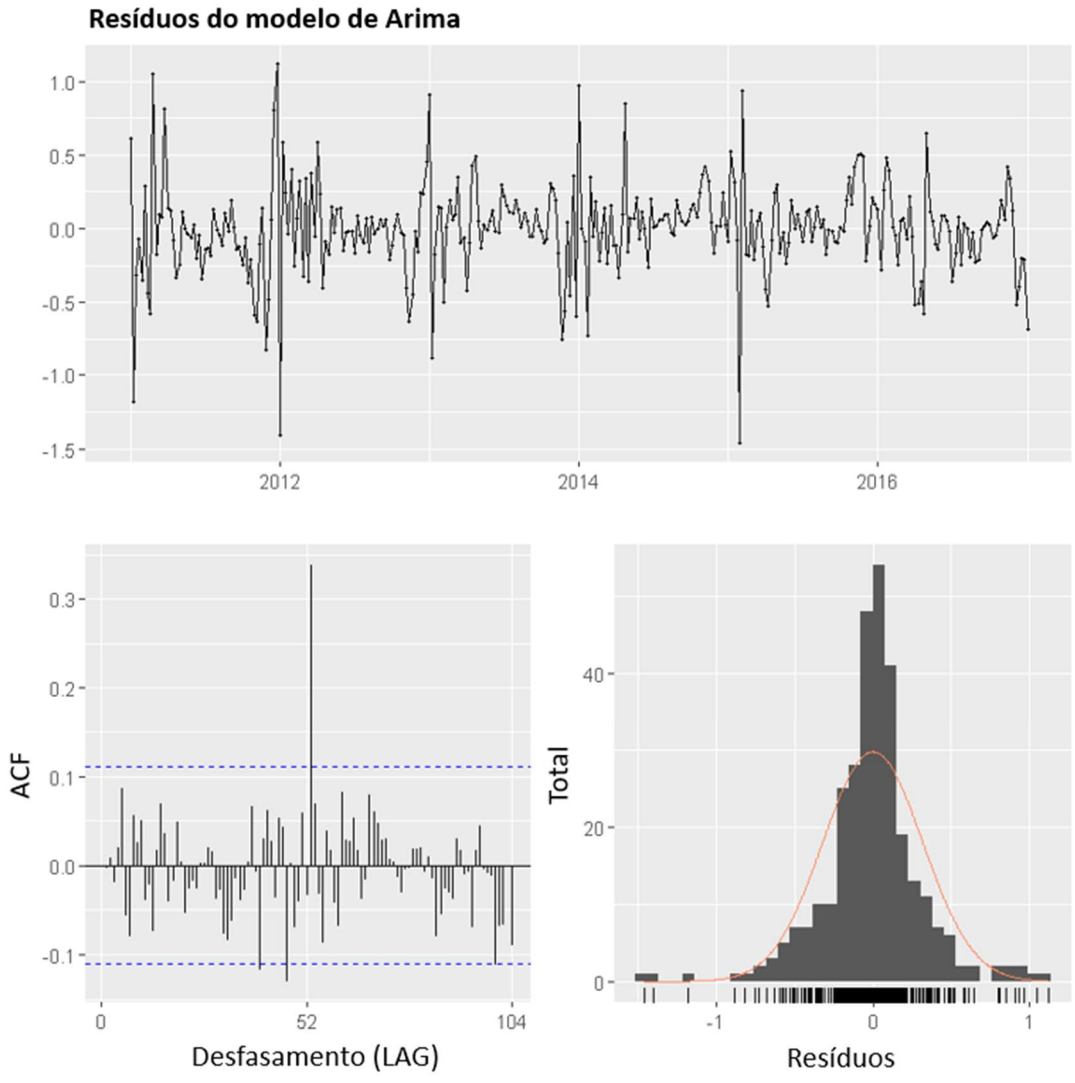


Figura 46 Avaliação do modelo ARIMA (interpretação dos resíduos)

Para comparar a precisão de modelos de previsão distintos, recorre-se à medição dos erros através de diferentes métricas. Uma das métricas consiste em calcular a percentagem

dos erros. Este método tem a vantagem de não depender da escala em que estão os dados, pelo que são frequentemente utilizados para comparar modelos que são treinados em diferentes conjuntos de dados. A métrica de percentagem mais utilizada é o *MAPE* (*Mean absolute percentage error*) [120]. Quando os valores estimados e os erros dos diferentes modelos estão na mesma unidade, as métricas *MAE* (*Mean absolute error*) e *RMSE* (*Root mean squared error*) são muito utilizadas, porque dependem da escala dos dados e dos valores estimados [120].

Para efetuar a validação cruzada dos resultados, o cálculo das métricas foi efetuado a partir dos valores estimados e dos observados do conjunto de dados de teste e foram utilizadas as métricas *MAE* e *RMSE,* por se tratar do mesmo conjunto de dados onde foram aplicadas as mesmas transformações.

O modelo *ARIMA* foi o que apresentou melhores resultados, conforme apresentado na Tabela 12.

Tabela 12 Medidas de precisão dos modelos

	RMSE	**MAE**
Modelo *ARIMA* - Regressão Dinâmico	0,319	0,214
Modelo de Regressão	0,389	0,287

4.2.3. Modelos de Aprendizagem Automática

Os algoritmos de aprendizagem automática podem ser utilizados como métodos para selecionar de forma automática, as variáveis que mais contribuem para o desempenho de um modelo preditivo.

Segundo Guyon e Elisseeff [163], existem três métodos de seleção de variáveis: (i) *Filter Methods*, que utilizam medidas estatísticas para atribuir uma pontuação as variáveis; (ii) *Wrapper Methods*, onde um modelo preditivo é utilizado para avaliar diferentes combinações de variáveis; e (iii) *Embedded Methods* que avaliam, na fase de construção do modelo, qual a variável que melhor contribui para o seu desempenho.

Para compreender que variáveis mais contribuíram para a procura dos serviços, foi efetuada uma análise exploratória da importância de cada variável, aplicando diferentes algoritmos de aprendizagem automática, listado no Apêndice G. Pela interpretação dos

gráficos (Apêndice G) da importância das variáveis apresentada pelos diferentes modelos, verificou-se que a tendência (*trend*) e a sazonalidade (termos *Fourier*) são consideradas as mais importantes. Para a análise exploratória, foram utilizadas as mesmas variáveis dos modelos de regressão e *ARIMA*, conforme apresentadas na Tabela 9 da secção 4.2.1.

Foi ainda aplicado o método *wrapper* de eliminação recursiva de variáveis, onde inicialmente foram consideradas todas as variáveis e progressivamente foram sendo eliminadas de acordo com o seu grau de importância. Os métodos *wrapper*, por utilizarem os algoritmos de aprendizagem automática que funcionam como uma caixa negra, são considerados universais e de utilização simples [163]. Na Figura 47 está representada a aplicação do método recursivo (*RFE*) de eliminação de variáveis, aplicando o algoritmo de "*Random Forest*". Este método foi aplicado para encontrar diferentes subconjuntos de variáveis e recorrendo à validação cruzada foi encontrado o subconjunto com melhor pontuação. Tendo como medida o *RMSE*, o subconjunto com melhor pontuação indica que o número ótimo de variáveis com maior contribuição para o modelo é de 16 e consistem nas seguintes variáveis: *seasonality.S1.52, seasonality.C1.52, Trend, CPC, dias_campanha_activo, custo, cliques, palavras_chave_dia, holidays_origem, seasonality.S2.52, CTR, Impressoes, campanhasativas, seasonality.C2.52, seasonality.C3.52, seasonality.S3.52*.

Ao comparamos as variáveis selecionadas por este método com as variáveis selecionados no modelo de regressão na Tabela 9, verificou-se que a variável *holidays_origem* passou a ser considerada como um preditor com alguma importância. Nenhum dos termos *Fourier* que representam a sazonalidade semanal (*S1-7, C1-7, S2-7, C2-7*) foram incluídos como preditores importantes. Dos quatro pares de termos utilizados para representar a sazonalidade anual (*S1-52, C1-52, S2-52, C2-52, S3-52, C3-52, S4-52, C4-52)* o último par de termos (*S4-52, C4-52)* não foi incluído no modelo.

Optou-se por executar os algoritmos de aprendizagem automática, considerando as 16 variáveis obtidas pelo método recursivo de seleção de variáveis.

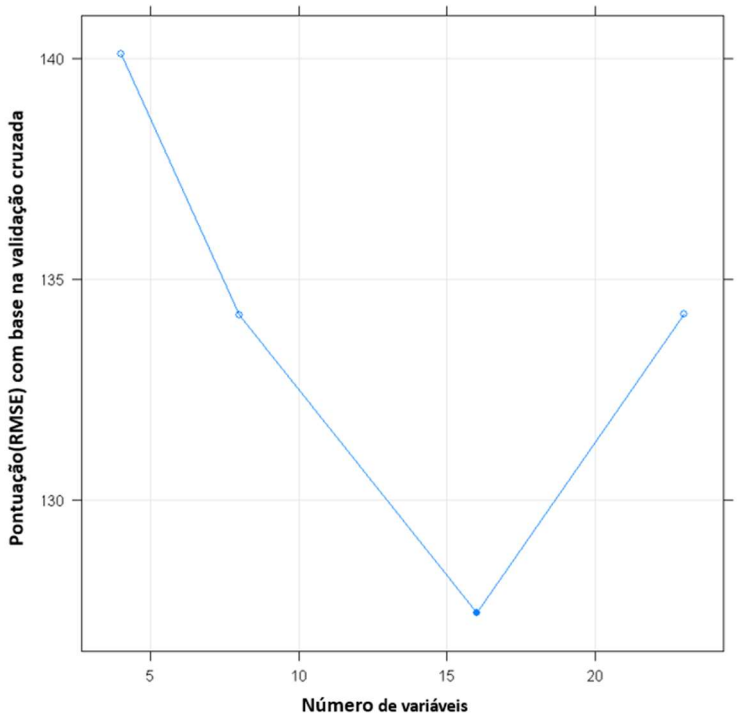


Figura 47 Eliminação recursiva de variáveis com validação cruzada

4.3 Discussão dos resultados

Quando se utiliza algoritmos de aprendizagem automática, uma abordagem típica de validação cruzada consiste no método *K-fold* [135]. Este método divide aleatoriamente os dados em K blocos do mesmo tamanho onde cada bloco é um subconjunto dos dados que compreende t=K observações atribuídas aleatoriamente, e t é o número de observações. Após a divisão dos dados em K blocos, cada bloco é escolhido iterativamente para testes. Um modelo é treinado no bloco K-1 e os erros são estimados no bloco que ficou de fora.

Apesar da divisão aleatória dos dados ser uma prática comum em aprendizagem automática, quando se trata de séries temporais, surgem questões relacionadas com a dependência entre as observações, conforme referido na secção 4.2.1 deste documento. Para lidar com dependência entre as observações, Snijders [174] apresenta o método *Blocked Cross-Validation*. O procedimento é semelhante ao método *K-fold* descrito acima. A diferença reside no facto de não ser efetuada a divisão aleatória das observações, mantendo os K blocos de observações contíguos e preservando a ordem natural das observações dentro de cada bloco.

Para o treino dos modelos de aprendizagem automática, recorreu-se ao método *Blocked Cross-Validation* com 10 blocos contíguos.

A Tabela 13 apresenta a comparação das métricas *RMSE* e *MAE* dos diferentes modelos de aprendizagem. O modelo *ARIMA* e o modelo *Stochastic Gradient Boosting* foram os que apresentaram melhores resultados de performance. O resultado obtido pelo modelo *ARIMA* pode estar relacionado com a capacidade do modelo incluir informações como a autocorrelação existente nos dados, de forma a capturar o dinamismo das séries temporais. Os resultados inferiores dos algoritmos de aprendizagem automática, relativamente ao modelo *ARIMA,* podem ser consequência do número de amostras ser reduzido, 313, o que tem impacto na capacidade de generalização dos algoritmos [135].

Tabela 13 Performance dos modelos de aprendizagem

	RMSE	**MAE**
ARIMA	0,319	0,214
Stochastic Gradient Boosting	0,378	0,281
Linear Regression	0,389	0,287
Random Forest	0,390	0,275
Suport Vector Machine	0,407	0,290
eXtreme Gradient Boosting	0,410	0,296
Multi Layer Perceptron	2,217	1,810
Neural Network	7,120	6,855

Capítulo 5 – Conclusões e recomendações

5.1 Principais conclusões

A presente dissertação visa conduzir um processo de extração de conhecimento dos dados de uma empresa de transfer de passageiros a operar na região do Algarve, com o objetivo de conhecer os seus clientes e construir um sistema de segmentação bem como avaliar possíveis fatores que influenciam a procura dos serviços, recorrendo às técnicas da área de conhecimento de Extração de Informação e *Text Mining*.

Os objetivos propostos permitiram formular três questões de investigação que foram respondidos recorrendo a metodologia CRISP-DM para formular o problema e encontrar uma solução.

As etapas inicias de desenvolvimento do projeto que consistiram na integração e análise exploratória dos dados, permitiram adquirir conhecimento sobre os turistas que procuram a região do Algarve e ter uma visão geral da evolução do negócio e das principais características da empresa. Por outro lado, foi possível nesta fase perceber se o conjunto de dados disponibilizado seria suficiente para responder as questões de investigação propostas.

Seguidamente é apresentado um resumo de como foram alcançadas as respostas às questões de investigação.

Q1: Que características do serviço são mais valorizados pelos clientes?

Para dar resposta a esta questão, na secção 3.4, foi efetuada uma análise da opinião dos clientes sobre o serviço na forma de *reviews*, esperando identificar tópicos de aspetos mais comentados. O modelo de tópicos construído permitiu identificar três tópicos com os aspetos que os clientes mais comentam, constituído um importante indicador para a empresa avaliar que decisões estão na origem dos comentários.

Em complemento a análise de tópicos, na secção 3.5, recorreu-se ao questionário de satisfação dos clientes onde foi realizado uma análise de componentes principais para resumir a informação dos questionários num número menor de variáveis que contenham a maior parte da informação presente nas variáveis originais. O resultado das componentes principais obtidos foram utilizados numa análise de clusters que resultou em seis grupos de clientes. A solução de clusters apresentado pode ser aplicada, por

exemplo, para decidir qual o perfil de condutor é o mais adequado para determinado grupo de clientes.

Q2: Que tipo de clientes procuram os serviços da empresa?

A resposta a esta questão encontra-se detalhada na secção 3.6, onde recorreu-se a análise de *clusters* para a segmentação de clientes.

Foram selecionadas e calculadas novas variáveis a partir dos dados existentes e baseando-se na literatura. Foram ainda calculadas as três variáveis do modelo *RFM*, *recency*, *frequency* e *monetary*, muito utilizadas para a segmentação de clientes em diferentes áreas de negócio. A análise de cluster permitiu identificar dois segmentos de clientes. Ainda, com o intuito da solução de clusters poder vir a ser aplicada no apoio à tomada de decisão, foi efetuada uma análise exploratória onde foram analisadas características como o tipo de serviço contratado, o mês do serviço, o local de *dropoff* e país de origem dos clientes de cada *cluster*.

Q3: Que fatores contribuem positivamente para o aumento das receitas da empresa?

No capítulo 4 foi efetuado o estudo dos possíveis fatores que influenciam a procura dos serviços e consequentemente a receita da empresa.

A informação dos possíveis fatores que influenciam a procura como a data do serviço ou as métricas extraídas das campanhas de marketing digital do *Google Ads*, foram obtidos dos dados operacionais da empresa. Foram ainda explorados fatores provenientes de fontes de dados externos que se julga ter influência na procura, tais como, a informação dos feriados nacionais no país de origem e de destino que podem indicar se os clientes aproveitam os fins-de-semana longos para viajar.

Concluiu-se que a procura dos serviços da empresa configura uma série temporal com padrões de sazonalidade. Para além da inclusão das métricas de campanhas de marketing digital, foram ainda incluídos fatores como a tendência e a sazonalidade, extraídos da decomposição das séries temporais. Foram utilizados métodos estatísticos e algoritmos de Aprendizagem Automática como técnica para selecionar as variáveis que mais contribuem para o desempenho de um modelo preditivo. Foi possível concluir que a sazonalidade e a tendência foram os fatores que mais contribuiriam para o desempenho dos modelos.

5.2 Contributos para a comunidade científica e empresarial

5.2.1 Implicações ao nível académico

A literatura académica tem apresentado os algoritmos de aprendizagem automática como uma alternativa aos métodos estatísticos para a previsão de séries temporais. No entanto, grande parte dos estudos publicados não comparam a precisão das previsões obtidas com os métodos estatísticos tradicionais de previsão de séries temporais. Até a data, do conhecimento do autor, apenas foram publicados dois estudos que estabelecem essa comparação:

No primeiro, Makridakis et al. [133], recomenda que os profissionais e investigadores que utilizam os algoritmos de aprendizagem automática necessitam ainda de encontrar uma solução para melhorar o desempenho dos modelos. Por outro lado, Cerqueira et al. [135] afirmam que para obter melhores resultados é necessário recolher o maior número de observações possível e a inclusão de ambos os métodos de previsão de forma a enriquecer os estudos. Surge assim uma lacuna na literatura na apresentação de provas empíricas que comparam os dois métodos de previsão de séries temporais.

O presente trabalho contribui estabelecendo uma comparação entre o desempenho de um dos modelos estatísticos mais populares de previsão de séries temporais, o modelo *ARIMA*, com algoritmos populares de aprendizagem automática. Com os dados agregados semanalmente, foram obtidas 313 observações. O modelo *ARIMA* apresentou melhores resultados relativamente aos algoritmos de Aprendizagem Automática. Os resultados obtidos corroboram com os trabalhos anteriores que afirmam que os modelos estatísticos apresentam melhores resultados quando as amostras são reduzidas.

5.2.2 Implicações ao nível empresarial

O estudo apresentado realça a importância do valor dos dados nos processos de tomada de decisão nas empresas.

Através de um caso de estudo, foi conduzido um processo de extração de conhecimento e reconhecimento de padrões nos dados de forma a compreender que tipo de clientes procuram os serviços de uma empresa e que fatores podem afetar a procura semanal dos serviços. Por exemplo, recorrendo a dados do questionário de satisfação dos clientes, foi possível agrupar os clientes pelos aspetos que mais valorizam. Ainda,

utilizando dados do histórico das transações e das campanhas de marketing digital foi possível obter previsões satisfatórios da procura semanal dos serviços e identificar a sazonalidade como o fator que mais afeta a procura dos serviços.

Assim, é importante que as empresas compreendam o valor dos dados que possuem e das vantagens competitivas que podem ser obtidas quando os processos de tomada de decisão são suportados pela análise de dados.

5.3 Limitações e proposta de investigação futura

Apesar do âmbito deste trabalho colocar o enfoque na construção dos modelos de previsão e na apresentação de resultados em formato de protótipo, considera-se de elevada importância a implementação dos modelos num ambiente real, no sentido de avaliar a sua aplicabilidade nos processos de tomada de decisão.

Uma limitação encontrada na construção dos modelos relaciona-se com a informação mais detalhada dos clientes e motoristas que não foi disponibilizada pela empresa, de modo a garantir a proteção dos dados. Concretamente, não foram disponibilizados dados demográficos como, por exemplo, a idade e o género, o que iria permitir uma comprensão mais profunda dos clientes e a criação de segmentos mais personalizados.

Como trabalho futuro, considera-se que a implementação dos modelos num ambiente de produção seria da maior importância, tendo em vista avaliar o impacto da solução nas operações comerciais e nos processos de tomada de decisão, desenvolvidos pelos decisores de negócio.

Bibliografia

[1] H. Coelho, "Privacidade na era digital," 2016. [Online]. Available: https://ciencias.ulisboa.pt/pt/noticia/16-05-2016/privacidade-na-era-digital. [Acedido em 07 07 2020].

[2] J. Webster e R. T. Watson, "Analyzing the past to prepare for the future: writing a literature review," *MIS Quarterly,* 2002.

[3] R. de Neufville, "Planning Airport Access in an Era of Low-Cost Airlines," *Journal of the American Planning Association,* vol. 72, nº 3, pp. 347-356, 2006.

[4] C. Ribeiro de Almeida, "Low cost airlines, airport and tourism.The case of Faro airport.," em *51st ERSA 2011. Annual Conference*, Barcelona, Espanha, 2011.

[5] D. P. Ramos e J. I. Izquierdo Misiego, "Flying from Europe to the Algarve: The Geographical Impacts of the Growth of Low-cost Carriers (1996-2013).," *Journal of Spatial and Organizational Dynamics, CIEO-Research Centre for Spatial and Organizational Dynamics, University of Algarve,* vol. 3, nº 4, pp. 275-295, 2015.

[6] J. F. P. Ribes, A. B. R. Rodríguez e A. O. Padilla, "Brexit Announcement: Immediate Impact on British Tourism in Spain," *Cornell Hospitality Quarterly,* vol. 60, nº 2, pp. 97-103, 2018.

[7] V. Navickas e A. Malakauskaite, "The Possibilities for the Identification and Evaluation of Tourism Sector Competitiveness Factors," *Engineering Economics,* vol. 61, nº 1, 2009.

[8] S. A. Khan, D. Qianli, W. SongBo, K. Zaman e Y. Zhang, "Travel and tourism competitiveness index: The impact of air transportation, railways transportation, travel and transport services on international inbound and outbound tourism," *Journal of Air Transport Management,* vol. vol. 58, nº issue C, pp. 125-134, 2017.

[9] E. Aguiló, T. Palmer e J. Rosselló, "Road Transport for Tourism: Evaluating Policy Measures from Consumer Profiles," *Tourism Economics,* vol. 18, nº 2, pp. 281-293, 2012.

[10] M. Pasha e M. Hickman, "Airport Ground Accessibility: Review and Assessment," em *38th Australasian Transport Research Forum (ATRF 2016)*, Melboure, Australia, 2016.

[11] E. Zaidana e A. Abulibdehb, "Modeling ground access mode choice behavior for Hamad International Airport in the 2022 FIFA World Cup city," *Journal of Air Transport Management,* vol. 73, pp. 32-45, 2018.

[12] L. T. I. Loi, A. S. I. So, I. S. Lo e L. H. N. Fong, "Does the quality of tourist shuttles influence revisit intention through destination image and satisfaction? The case of Macao.," *Journal of Hospitality and Tourism Management,* vol. 32, pp. 115-123, 2017.

[13] D. T. Duval, "Tourism and Transport: Modes, Networks and Flows," *Clevedon, United Kingdom: Channel View Publications,* 2007.

[14] B. Prideaux, "The role of the transport system in destination development.," *Tourism Management,* vol. 21, nº 1, pp. 53-63, 2000.

[15] C. M. Law, Urban Tourism: The Visitor Economy and the Growth of Large Cities (2nd ed.), London: Continuum, 2002.

[16] G. Harvey, "Study of airport access mode choice," *Journal of transportation Engineering,* vol. 112, pp. 525-545, 1986.

[17] M. Furuichi e F. S. Koppelman, "An analysis of air travelers departure airport and destination choice behavior.," *Transportation Research Part A: Policy and Practice,* nº 28, pp. 187-195, 1994.

[18] V. Psaraki e C. Abacoumkin, "Access mode choice for relocated airports: the new Athens International Airport," *Journal of Air Transport Management,* vol. 8, nº 2, pp. 89-98, 2002.

[19] H. Chebli e H. S. Mahmassani, "Air travelers stated preferences towards new airport landside access mode services," *Annual Meeting of Transportation Research Board, Washington DC,* 2003.

[20] S. Gupta, P. Vovsha e R. Donnelly, "Air Passenger Preferences for Choice of Airport and Ground Access Mode in the New York City Metropolitan Region.," *Transportation Research Record: Journal of the Transportation Research Board, 2042,* pp. 3-11, 2008.

[21] R. C. Jou, D. A. Hensher e T. L. Hsu, "Airport ground access mode choice behavior after the introduction of a new mode: A case study of Taoyuan International Airport in Taiwan," *Transportation Research Part E: Logistics and Transportation Review,* vol. 47, nº 3, pp. 371-381, 2011.

[22] M. L. Tam, W. H. Lam e H. P. Lo, "Incorporating passenger perceived service quality in airport ground access mode choice model," *Transportmetrica,* vol. 6, nº 1, pp. 3-17, 2010.

[23] A. Mamdoohi, M. Saffarzade, A. Taherpour e M. Y. Panah, "Modeling Air Passengers Ground Access Mode Choicea Case Study of IKIA," *International Journal of Modeling and Optimization,* vol. 2, nº 2, pp. 147-152, 2012.

[24] S. Choo, S. You e H. Lee, "Exploring characteristics of airport access mode choice: a case study of Korea," *Transportation Planning and Technology,* vol. 36, nº 3, pp. 335-351, 2013.

[25] H.-J. Roh, "Mode Choice Behavior of Various Airport User Groups for Ground Airport Access," *The Open Transportation Journal,* vol. 7, pp. 43-55, 2013.

[26] G. Akar, "Ground access to airports, case study: Port Columbus International Airport," *Journal of Air Transport Management, Elsevier,* vol. 30, pp. 25-31, 2013.

[27] S. Hess, T. Ryley, L. Davison e T. Adler, "Improving the quality of demand forecasts through cross nested logit: A stated choice case study of airport, airline and access mode choice.," *Transportmetrica A: Transport Science,* vol. 9, pp. 358-384, 2013.

[28] D. Bao, T. Guo e S. Hua, "Analysis of airport passenger's access mode choice based on SP/RP combined data," *Journal of Wuhan University of Technology (Transportation Science and Engineering),* vol. 39, pp. 763-767, 2015.

[29] D.-T. Le-Klähna e C. M. Hallb, "Tourist use of public transport atdestinations – a review," *Current Issues in Tourism,* vol. 18, nº 8, pp. 37-41, 2014.

[30] P. B. Mandle, D. M. Mansel e M. A. Coogan, "Use of public transportation by airport passengers.," *Transportation Research Record: Journal of the Transportation Research Board,* pp. 83-89, 2000.

[31] TUA - Titulo Único Ambiental, "Agência Portuguesa para o Ambiente," 21 01 2020. [Online]. Available: http://siaia.apambiente.pt/AIADOC/AIA3280/aia3280_dia_tua202012217290.pdf. [Acedido em 24 01 2020].

[32] S. Zlatanov e J. Popesku, "The Link Between Digital Media and Making Travel Choices.," *Marketing: časopis za marketing teoriju i praksu,* vol. 48, nº 2, pp. 75-85, 2017.

[33] D. Buhalis e M. Foerste, "SoCoMo marketing for travel and tourism: Empowering co-creation of value," *Journal of Destination Marketing & Management,* vol. 4, 2015.

[34] C. Lamsfus, C. Grün, A. Sorzabal e W. H. Aurkene, "Context-based matchmaking to enhance tourists experience," 2010.

[35] P. Prekop e M. Burnett, "Activities, context and ubiquitous computing," *Computer Communications ,* vol. 26, nº 11, pp. 1168-1176, 2003.

[36] A. Emrich, A. Chapko e D. Werth, "Context-Aware Recommendations on Mobile Services: The m:Ciudad Approach," *Lecture Notes in Computer Science book series,* vol. 5741, 2009.

[37] A. LaMarca, Y. Chawathe, S. Consolvo, J. Hightower, I. Smith, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, J. Tabert, P. Powledge, G. Borriello e B. Schilit, "Place Lab: Device Positioning Using Radio Beacons in the Wild," *Lecture Notes in Computer Science book series,* vol. 3468, 2005.

[38] W. Woerndl, C. Schueller e R. Wojtech., "A hybrid recommender system for context-aware recommendations of mobile applications.," em *IEEE 23rd International Conference on Data Engineering Workshop*, 2007.

[39] W. Woerndl, J. Huebner, R. Bader e D. Gallego-Vico., "A model for proactivity in mobile, context-aware recommender systems," em *Proceedings of the fifth ACM conference on Recommender systems*, 2011.

[40] X. Liu e K. Aberer., "SoCo: a social network aided context-aware recommender system," em *Proceedings of the 22nd international conference on World Wide Web*, 2013.

[41] K. J. Kim, H. Ahn e S. Jeong, "Context-aware recommender systems using data mining techniques.," em *Proceedings of world academy of science, engineering and technology*, 2010.

[42] A. Karatzoglou, X. Amatriain, L. Baltrunas e N. Oliver, "Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering," em *Proceedings of the fourth ACM conference on Recommender systems*, 2010.

[43] E. Du Plessis, The advertised mind: Ground-breaking insights into how our brains respond to advertising., Kogan Page, 2005.

[44] M. Shatnawi e N. Mohamed, "Statistical techniques for online personalized advertising: a survey," em *27th Annual ACM Symposium on Applied Computing*, 2012.

[45] N. Chowdhury, "A Survey of Search Advertising, A report .," 2007.

[46] M. Langheinrich, A. Nakamura, N. Abe, T. Kamba e Y. Koseki, "Unintrusive customization techniques for Web advertising," *NEC Corporation, C&C Media Research Laboratories,* pp. 216-8555, 1999.

[47] A. Broder, M. Fontoura, V. Josifovski e L. Riedel, "A semantic approach to contextual advertising.," pp. 559-566, 2007.

[48] D. F. M. Buhalis, "SoCoMo marketing for travel and tourism: Empowering co-creation of value," *Journal of Destination Marketing & Management,* vol. 4, pp. 151-161, 2015.

[49] R. Chatwin, "An overview of computational challenges in online advertising," em *Proceedings of the American Control Conference*, Washington, DC, USA, 2013.

[50] IAB, "The Interactive Advertising Bureau," 2018. [Online]. Available: www.iab.com/wp-content/uploads/2018/11/REPORT-IAB-Internet-Advertising-Revenue-Report-HY-2018.pdf. [Acedido em 01 02 2019].

[51] H. B. McMahan, G. Holt, D. Sculley, M. Young, E. Dietmar, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos e J. Kubica, "Ad click prediction: a view from the trenches," *The 19th ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 1222-1230, 2013.

[52] M. Richardson, E. Dominowska e R. Ragno, "Predicting clicks: estimating the click-through rate for new ads.," *In Proceedings of the 16th international conference on World Wide Web,* pp. 521-530, 2007.

[53] Y. Bart, T. Stephen e M. Sarvary, "Which products are best suited to mobile advertising? A field study of mobile display advertising effects on consumer attitudes and intentions.," *Journal of Marketing Research,* vol. 51, nº 3, p. 270–285, 2014.

[54] D. Fain e J. Pedersen, "Sponsored Search: a Brief History," em *Proceedings of the Second Workshop on Sponsored Search Auctions.*, 2006.

[55] P. Kireyev, K. Pauwels e S. Gupta, "Do display ads influence search? Attribution and dynamics in online advertising.," *International Journal of Research in Marketing,* vol. 33, p. 475–490, 2016.

[56] C. Kim, k. Kwon e W. Chang, "How to measure the effectiveness of online advertising in online marketplaces.," *Expert Systems with Applications,* vol. 38, p. 4234–4243, 2011.

[57] J. Lecinski, "ZMOT Winning the zero momento of truth," *Google,* 2011.

[58] "Google Adsense," [Online]. Available: https://www.google.com/adsense/. [Acedido em 01 09 2020].

[59] "Google Adwords," [Online]. Available: https://ads.google.com/. [Acedido em 01 09 2020].

[60] "Contextual Advertising," [Online]. Available: http://www.contextual-advertising.org. [Acedido em 01 12 2019].

[61] J. R. Saura, P. R. P. Sanchez e L. M. C. Suárez, "Understanding the Digital Marketing Environment with KPIs and Web Analytics," *Future Internet,* vol. 9, nº 4, 2017.

[62] M. Weideman, "Ethical issues on content distribution to digital consumers via paid placement as opposed to website visibility in search engine results.," em *Seventh ETHICOMP International Conference on the Social and Ethical Impacts of Information and Communication Technologies*, 2004.

[63] M. Weideman e H.-S. T., "An investigation into search engines as a form of targeted advert delivery.," *In Proceedings of the 2002 annual research*

conference of the South African institute of computer scientists and information technologists on Enablement through technology (SAICSIT'02), p. 258–258, 2002.

[64] M. F. D. Regelson, "Predicting click-through rate using keyword clusters.," *In Proceedings of the Second Workshop on Sponsored Search Auctions.,* 2006.

[65] A. Z. Broder, P. Ciccolo, M. F. M. C. D. Fontoura, E. Gabrilovich, J. Vanja e R. Lance, "Search Advertising using Web Relevance Feedback," *Proceedings of the 17th ACM conference on Information and knowledge management,* pp. 1013-1022, 2008.

[66] L. Shi e B. LI, "Predict the click-through rate and average cost per click for keywords using machine learning methodologies," em *Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management*, Michigan, USA, 2016.

[67] W. Smith, "Product Differentiation and Market Segmentation as Alternative Marketing Strategies," *Journal of Marketing,* vol. 21, pp. 3-8, 1956.

[68] T. E. Shreya, A. Bhardwaj e Poovammal, "Approaches to Clustering in Customer Segmentation," *International Journal of Engineering & Technology,* vol. 7, nº 3.12, pp. 802-807, 2018.

[69] A. C. Tynan e J. Drayton, "Market segmentation," *Journal of Marketing Management,* vol. 2, nº 3, pp. 301-335, 1987.

[70] J. Blythe, Marketing strategy, London: McGraw-Hill Education, 2003.

[71] P. Kotler, G. Armstrong, V. Wong e J. Saunders, Principles of Marketing, 3rd European Edition, London: Prentice-Hall, 2002.

[72] G. Barrie e A. Furnham, Consumer Profiles: An Introduction to Psychographics (Consumer Research and Policy), London: Routledge, 1992.

[73] A. Weinstein, Market segmentation using demographics, psychographics and other, Chicago: Probus publishing company, 1994.

[74] R. Straughan e J. Roberts, "Environmental segmentation alternatives: a look at green consumer behavior in the new millennium," *Journal of Consumer Marketing,* vol. 16, nº 6, pp. 558-575, 1999.

[75] S. Goyat, "The basis of market segmentation: a critical review of," *European Journal of Business and Management,* vol. 3, nº 9, 2011.

[76] A. Singh, "Impact of demographical factors on the purchasing behaviour of the customers with special reference to FMCG: An empirical study," *International journal of research in commerce and management,* vol. 2, nº 3, 2011.

[77] V. K. Wells, S. W. Chang, O. J. Castro e J. Pallister, "Market Segmentation from a Behavioral Perspective," *Journal of Organizational Behavior Management,* vol. 30, nº 2, pp. 176-198, 2010.

[78] B. Higgs e A. Ringer , "Trends in Consumer Segmentation," em *Australian and New Zealand Marketing Academy Conference*, University of Otago, Dunedin, NZ, 2007.

[79] A. Ansari e A. Riasi, "Modelling and evaluating customer loyalty using neural networks: Evidence from startup insurance companies," *Future Business Journal,* vol. 2, nº 1, pp. 15-30, 2016.

[80] S. Dolnicar, “Market Segmentation in Tourism,” *Tourism Management: Analysis, Behaviour and Strategy, CAB International, Cambridge,* pp. 29-150, 2008.

[81] S. Dolnicar, “Beyond Commonsense Segmentation – a Systematics of Segmentation Approaches in Tourism,” *Journal of Travel Research,* vol. 42, nº 3, pp. 244-250, 2004.

[82] J. H. Myers e E. Tauber, “Market structure analysis,” *American Marketing Association: Chicago,* 1977.

[83] K. D. Bailey, “Typologies and Taxonomies: An Introduction to Classification Techniques,” *Sage University Paper series on Quantitative Applications in the Social Sciences. Thousand Oaks: Sage,* vol. 102, pp. 1-96, 1994.

[84] J. David J. Ketchen e C. L. Shook, “The application of cluster analysis in strategic management research: an analysis and critique.,” *Strategic Management Journal,* vol. 17, nº 16, pp. 441-458, 1996.

[85] R. Baumann, “Marktsegmentierung in den Sozial- und Wirtschaftswissenschaften: eine Metaanalyse der Zielsetzungen und Zugänge,” *Diploma thesis at Vienna University of Economics and Management Science,* 2000.

[86] R. Haley, “Benefit Segmentation: A Decision-Oriented Research Tool,” *Journal of Marketing,* nº 32, pp. 30-35, 1968.

[87] R. Calantone, C. Schewe e C. Allen, “Targeting Specific Advertising Messages at Consumer Segment,” *In D.E. Hawkins, E.L. Shafer, and J.M, Tourism Marketing and Management,* pp. 133-147, 1980.

[88] J. Goodrich, “Benefit Segmentation of US International Travelers: An Empirical Study with American Express.,” *J.M., (Eds.), Tourism Marketing and Management,* pp. 133-147, 1980.

[89] M. Crask, “Segmenting the Vacationer Market: Identifying the Vacation Preferences, Demographics, and Magazine Readership of Each Group,” *Journal of Travel Research,* nº 20, pp. 20-34, 1981.

[90] J. Mazanec, “How to detect Travel Market Segments: A Clustering Approach,” *Journal of Travel Research,* vol. 23, nº 1, pp. 17-21, 1984.

[91] S. Dolnicar, "A review of data-driven market segmentation in tourism," *http://ro.uow.edu.au/commpapers/41,* 2002.

[92] G. Punj e D. W. Stewart, “Cluster analysis in marketing research: Review and suggestions for application,” *Journal of marketing research,* vol. 20, nº 2, pp. 134-148, 1983.

[93] R. J. Calantone e A. G. Sawyer, “The Stability of Benefit Segments,” *Journal of Marketing Research,* vol. 15, nº 3, 1978.

[94] P. E. Green e F. J. Carmone, “The Performance Structure of the Computer Market: A Multivariate Approach,” *Economic and Business Bulletin,* vol. 20, pp. 1-11, 1968.

[95] A. Kassambara, Practical Guide To Cluster Analysis in R, 1 ed., sthda.com , 2017.

[96] J. Han, J. Pei e M. Kamber, Data mining concepts and techniques third edition, The Morgan Kaufmann Series in Data Management Systems, 2012.

[97] M. S. Aldenderfer e R. K. Blashfield, “The Methods and Problems of Cluster Analysis. Sage Series on quantitative applications in the social sciences,” *Beverly Hills: Sage Publications,* 1988.

[98] J. MacQueen, “Some methods for classification and analysis of multivariate,” *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* pp. 281-297, 1967.

[99] L. Gary e A. Rangaswamy, “Marketing Engineering:Computer-Assisted Marketing Analysis and Planning.Mass,” *Addison-Wesley.,* 1998.

[100] R. Thorndike, “Who belongs in the family?,” *Psychometrika,* vol. 18, pp. 267-276, 1953.

[101] G. Milligan, “A monte carlo study of thirty internal criterion measures for cluster analysis,” *Psychometrika,* vol. 46, pp. 187-199, 1981.

[102] G. Milligan e M. Cooper, “An examination of procedures for determining the number of clusters in data sets,” *Psychometrika,* nº 50, pp. 159-179, 1985.

[103] J. Y. Wong, H. J. Chen, P. H. Chung e N.-C. Kao, “Identifying Valuable Travelers and Their Next Foreign Destination by the Application of Data Mining Techniques,” *Asia Pacific Journal of Tourism Research,* vol. 11, nº 4, p. 355 – 373, 2006.

[104] S. A. Lumsden, S. Beldona e A. M. Morrison, “Customer Value in an All-Inclusive Travel Vacation Club: An Application of the RFM Framework,,” *Journal of Hospitality & Leisure Marketing,* vol. 16, nº 3, pp. 270-285, 2008.

[105] S. Dolnicar, S. Kaiser e K. Lazarevski , “Overcoming Data Dimensionality Problems in Market Segmentation,” *Journal of Travel Research,* 2012.

[106] A. M. Hughes, “Strategic database marketing.,” *Chicago: Probus PublishingCompany.,* 1994.

[107] U. Kaymak, “Fuzzy target selection using RFM variables,” *In IFSA World congress and 20th NAFIPS international conference,* vol. 2, pp. 1038-1043, 2001.

[108] J. M. C. Schijns e G. Schroder, “Segment selection by relationship strength.,” *Journal of Direct Marketing,* vol. 10, pp. 69-79, 1996.

[109] P. Kotler, “Marketing management: Analysis, planning, implementation, and control.,” *New Jersey: Prentice-Hall.,* 1994.

[110] D. Peppers e M. Rogers, The one to one future: Building relationships one customer at a time, New York: Doubleday, 1996.

[111] P. Alford, “Database Marketing In Travel And Tourism.,” *Travel & Tourism Analyst,* vol. 1, pp. 87-104, 1999.

[112] S. C. Madeira e A. L. Oliveira, “Biclustering Algorithms for Biological Data Analysis: A Survey,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 1, nº 1, pp. 24-45, 2004.

[113] E. Turban, R. Sharda e D. Delen, “Decision Support and Business Intelligence Systems,” *9th edition Pearson,* 2011.

[114] J. Gareth, D. Witten, T. Hastie e R. Tibshirani, An Introduction to Statistical Learning with Applications in R, New York Heidelberg Dordrecht London: Springer Texts in Statistics, 2013.

[115] M. Van Rijmenam, “Why the 3v’s are not sufficient to describe big data, BigData Startups, Tech. Rep.,” 2013.

[116] D. Laney, "3D data management: controlling data volume, velocity, and variety, META Group, Tech. Rep.," 2001.

[117] C. W. Tsai, C. F. Lai, H. C. Chao e A. Vasilakos, "Big data analytics: A survey.," *Journal of Big Data,* vol. 2, nº 21, 2015.

[118] A. Cordeiro, A. Oliveira e D. Duarte, "Fintech Desafios da Tecnologia Financeira," *edições Almeida S.A.,* 2017.

[119] M. G. Dekimpe e D. M. Hanssen, "Time-series models in marketing: Past, present and future," *Intern. J. of Research in Marketing,* vol. 17, p. 183–193, 2000.

[120] R. J. Hyndman e G. Athanasopoulos, Forecasting: principles and practice 2nd Edition, Monash University, Australia, 2018.

[121] R. G. Brown, Statistical forecasting for inventory control, McGraw/Hill, 1959.

[122] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted averages," *O.N.R. Memorandum,* nº 52, 1957.

[123] P. R. Winters, "Forecasting sales by exponentially weighted moving averages," *Management Science,* pp. 324-342, 1960.

[124] K. Pauwels, I. Currim, M. Dekimpe, D. Hanssens, N. Mizik, G. Eric e N. Prasad, "Modeling Marketing Dynamics by Time SeriesEconometrics," *Marketing Letters,* vol. 15, pp. 167-183, 2004.

[125] K. Pauwels, D. M. Hanssens e S. Siddart, "The Long-Term Effects of Price Promotions on Category Incidence, Brand Choice, and Purchase Quantity," *Journal of Marketing Research,* vol. 39, nº 4, pp. 421-439, 2002.

[126] M. Dekimpe e D. M. Hanssens , "The Persistence of Marketing Effects on Sales," *Marketing Science,* vol. 14, nº 1, pp. 1-21, 1995.

[127] B. J. Bronnenberg, "The Emergence of Market Structure in New Repeat-Purchase Categories: The Interplay of Market Share and Retailer Distribution," *Journal of Marketing Research,* vol. 37, nº 1, pp. 16-31, 2000.

[128] A. Espasa e A. P. Espartero, "FORECASTING WITH DYNAMIC REGRESSION MODELS," Madrid, 2008.

[129] H. Böse, V. Flunkert, J. Gasthaus, T. Januschowski, D. Lange, D. Salinas, S. Schelter, M. Seeger e Y. Wang, "Probabilistic demand forecasting at scale," *Proceedings of the VLDB Endowment,* vol. 10, nº 12, pp. 1694-1705, 2017.

[130] N. K. Ahmed, A. F. Atiya, N. E. Gayar e H. El-Shishiny, "An Empirical Comparison of Machine Learning Models for Time Series Forecasting," 2010.

[131] R. Sharda e R. B. Patil, "Connectionist approach to time series prediction: An empirical test," *Journal of Intelligent Manufacturing,* vol. 3, nº 5, pp. 317-323, 1992.

[132] I. Alon, M. Qi e R. J. Sadowski, "Forecasting aggregate retail sales: A comparison of artifcial neural networks and traditional methods," *Journal of Retailing and Consumer Services,* vol. 8, nº 3, pp. 147-156, 2001.

[133] S. Makridakis, E. Spiliotis e V. Assimakopoulos, "Statistical and Machine Learning forecasting methods:Concerns and ways forward," *PLoS ONE,* vol. 13, nº 3, 2018.

[134] G. Papacharalampous, H. Tyralis e D. Koutsoyiannis, "Univariate Time Series Forecasting of Temperature and Precipitation with a Focus on Machine

Learning Algorithms: a Multiple-Case Study from Greece," *Water Resources Management,* vol. 32, pp. 5207-5239, 2018.

[135] V. Cerqueira, L. Torgo e C. Soares, "Machine Learning vs Statistical Methods for Time Series Forecasting: Size Matters," *ArXiv abs/1909.13316,* 2019.

[136] T. F. Foster Provost, Data Science for Business: What you need to know about data mining and data-analytic thinking, O'REILLY, 2013.

[137] E. F. Codd, "A Relational Model of Data for Large Shared Data Banks," *IBM Research Laboratory, San Jose, California,* 1970.

[138] I. H. Witten e E. Frank, "Data mining: Practical machine learning tools and techniques with java implementations, Morgan Kauffmann.," 2000.

[139] T. Algarve, "2020 Turismo do Algarve.," [Online]. Available: https://www.turismodoalgarve.pt/pt/5614/premios.aspx. [Acedido em 27 6 2020].

[140] N. Antonio, A. Almeida e L. Nunes, "Predicting hotel booking cancellation to decrease uncertainty and increase revenue," *Tour. Manag. Stud,* vol. 13, nº 2, pp. 25-39, 2017.

[141] M. Sahlgren, "The Word-Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high dimensional vector spaces.," *Ph.D. thesis, Stockholm University.,* 2006.

[142] R. Alghamdi e K. Alfalqi, "A Survey of Topic Modeling in Text Mining," *International Journal of Advanced Computer Science and Applications,* vol. 2, nº 1, 2015.

[143] P. S. Radim Rehurek, "Software Framework for Topic Modelling with Large Corpora," *Natural Language Processing LaboratoryMasaryk University, Faculty of Informatics,* 2010.

[144] S. Bird, E. Klein e E. Loper, Natural Language Processing with Python, O'Reilly Media, 2009.

[145] C. Sievert, S. Kenneth e E. Shirley, "LDAvis: A method for visualizing and interpreting topics," *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces (Association for Computational Linguistics, Stroudsburg, PA),* 2014.

[146] K. Stevens, P. Kegelmeyer, D. Andrzejewski e D. Buttler, "Exploring Topic Coherence over Many Models and Many Topics," *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* pp. 952-961, 2012.

[147] L. Alsumait, D. Barbar, J. Gentle e C. Domeniconi., "Topic significance ranking of LDA generative models," *Lecture Notes in Computer Science,* vol. 5781, pp. 67-82, 2009.

[148] Z. W. Lin e C. Yung, "Towards finding valuable topics," *ICDM,* pp. 720-731, 2010.

[149] H. Wallach, I. Murray e R. Salakhutdinov, "Evaluation methods for topic models," em *Proceedings of the 26th International Conference on Machine Learning (ICML). Omnipress*, 2009.

[150] D. Jurgens e K. Stevens, "The s-space package: an open source package for word space models," *Proceedings of the ACL 2010 System Demonstrations. Association for Computational Linguistics.,* pp. 30-35, 2010.

[151] L. J. Cronbach, "Coefficient Alpha and the internal structure of tests," *Psychometrika,* vol. 16, p. 297–334, 1951.

[152] J. Maroco e T. Garcia Marques, "Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas?," *Laboratório de Psicologia, I.S.P.A.,* vol. 4, nº 1, pp. 65-90, 2006.

[153] K. Fukunaga, Introduction to Statistical Pattern Recognition, Second Edition, WERNER RHEINBOLDT, 1990.

[154] L. Maaten, E. Postma e J. Herik, "Dimensionality Reduction: A Comparative Review," *J Mach Learn Res,* vol. 10, pp. 66-71, 2009.

[155] J. Maroco, Análise estatística com utilização do SPSS, Edições Sílabo ISBN: 972-618-331-6, 2003.

[156] S. Sharma, Applied Multivariate Techniques., New York: John William & Sons., 1996.

[157] M. S. Bartlett, The effect of standardization on a Chi-square approximation in factor analysis, Biometrika, 1951.

[158] M. H. Pestana e J. N. Gageiro, Análise de Dados para Ciências Sociais: A Complementaridade do SPSS, Lisboa: Edições Sílabo, 2005.

[159] R. B. Cattell, "The Scree Test For The Number Of Factors," *Multivariate Behavioral Research,* 1966.

[160] L. Kaufman e P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis., New York: Wiley, 1990.

[161] J. C. a. H. R. Bezdek, "VAT: A Tool for Visual Assessment of (Cluster) Tendency," *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint,* 2002.

[162] P. Dhandayudam e I. Krishnamurthi, "An Improved Clustering Algorithm for Customer Segmentation," *International Journal of Engineering Science and Technology (IJEST),* vol. 4, nº 2, 2012.

[163] I. Guyon e A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research,* pp. 1157-1182, 2003.

[164] J. Tang, S. Alelyani e H. Liu, "Feature selection for classification: A review.," *In Data Classification: Algorithms and Applications ,* pp. 37-64, 2014.

[165] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Mining and Knowledge Discovery,* vol. 2, pp. 283-304, 1998.

[166] H. Ralambondrainy, "A conceptual version of the K-means algorithm,," *Pattern Recognition Letters,* vol. 16, nº 11, pp. 1147-1157, 1995.

[167] Z. Huang, "Clustering large data sets with mixed numeric and categorical values.," *Proceedings of the FirstPacific Asia Knowledge Discovery and Data Mining Conference, Singapore: World Scientific, pp. 21–34.,* 1997a.

[168] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining.," *Proceedingsof the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Dept. of ComputerScience, The University of British Columbia, Canada, pp. 1–8.,* 1997b.

[169] R. B. Cleveland, W. S. Cleveland e I. Terpenning, "STL: A seasonal-trend decomposition procedure based on loess," *Journal of Official Statistics,* vol. 6, nº 1, pp. 3-33, 1990.

[170] G. E. P. Box e G. M. Jenkins, Time series analysis:Forecasting and control, San Francisco: Holden-Day, 1976.

[171] Z. Reitermanová, "Data splitting," *in WDS'10 Proceeding of Contributing Papers, Praha,* pp. 31-36, 2010.

[172] S. Arlot e A. Celisse, "A survey of cross-validation procedures for model," *Statistics surveys,* vol. 4, pp. 40-79, 2010.

[173] R. J. Hyndman e Y. Khandakar, "Automatic time series forecasting: The forecast package for R," *Journal of Statistical Software,* vol. 27, nº 1, pp. 1-22, 2008.

[174] T. Snijders, "On cross-validation for predictor evaluation in time series," *Model Uncertainty and its Statistical Implications,* pp. 56-69, 1988.

[175] I. CRISP-DM, "IBM SPSS Modeler CRISP-DM Guide," 1999. [Online]. Available: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.2.1/en/ModelerCRISPDM.pdf. [Acedido em 27 08 2020].

[176] S. Makridakis e M. Hibon, "The M3-Competition: Results, Conclusions and Implications," *International Journal of Forecasting,* vol. 16, nº 4, pp. 451-476, 2000.

[177] V. Cerqueira, L. Torgo e I. Mozetic, "Evaluating time series forecasting models: An empirical study," *arXiv:1905.11744,* 2019.

[178] M. Dekimpe, P. Franses, D. Hanssens e P. Naik, "Time-Series Models in Marketing," *ERIM Report Series Research in Management ERS-2006-049-MKT,* pp. 1-31, 2006.

[179] D. Trinh, L. Klahn e H. C. Michael, "Tourist use of public transport at destinations – a review," *Current Issues in Tourism,* vol. 18, nº 8, pp. 785-803, 2014.

Anexos e Apêndices

Apêndice A - Estatísticas sumárias dos dados

Nome: feedback.csv

Observações: 84 122

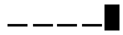
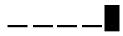
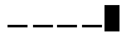
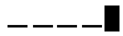
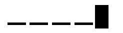
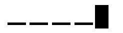
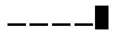
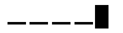
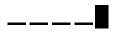
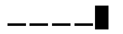
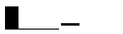
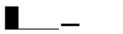
Atributos: 30

Tipos de atributos:

Fator :5

Numéricos:25

Tabela 14 Dados recolhidos do questionário de satisfação dos clientes(feedback.csv)

Variável	Omissos	Completos	Média	Desvio Padrão	Histograma
welcome_arrival	0	1	4.52	1.33	
welcome_departure	0	1	1.56	2.30	
punctual_arrival	0	1	4.33	1.64	
clean_in_arrival	0	1	4.60	1.21	
clean_out_arrival	0	1	4.59	1.21	
clean_in_departure	0	1	1.60	2.32	
clean_out_departure	0	1	1.60	2.32	
punctual_departure	0	1	4.42	1.52	
driving_arrival	0	1	4.28	1.64	
driving_departure	0	1	4.38	1.52	
curteous_arrival	0	1	4.31	1.64	
curteous_departure	0	1	4.42	1.51	
website	0	1	4.64	0.97	
experience	0	1	6.62	2.52	
quality	0	1	2.97	2.35	
freegift	0	1	1.01	1.99	
wifi_arrival	0	1	0.33	1.17	
wifi_departure	0	1	0.29	1.13	
boot_arrival	0	1	1.25	2.16	
boot_departure	0	1	1.27	2.17	

Apêndice B – Análise exploratória dos dados

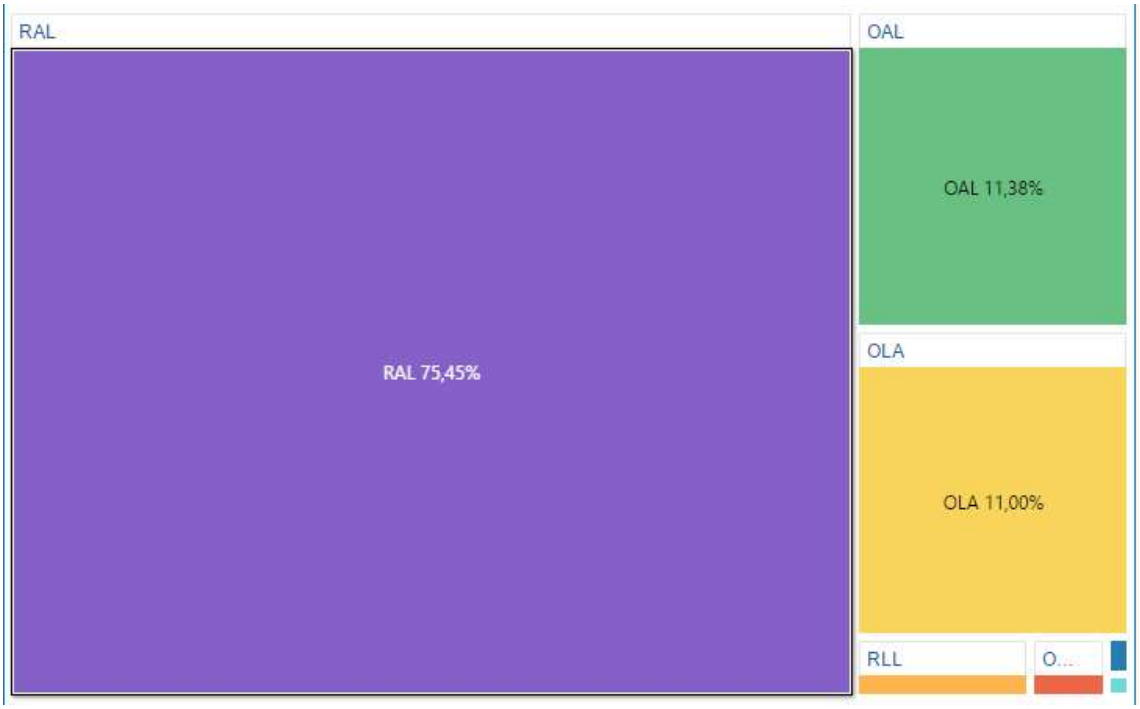


Figura 48 Distribuição diária das reservas por tipo de serviço

Apêndice C – Modelação de tópicos em *reviews*

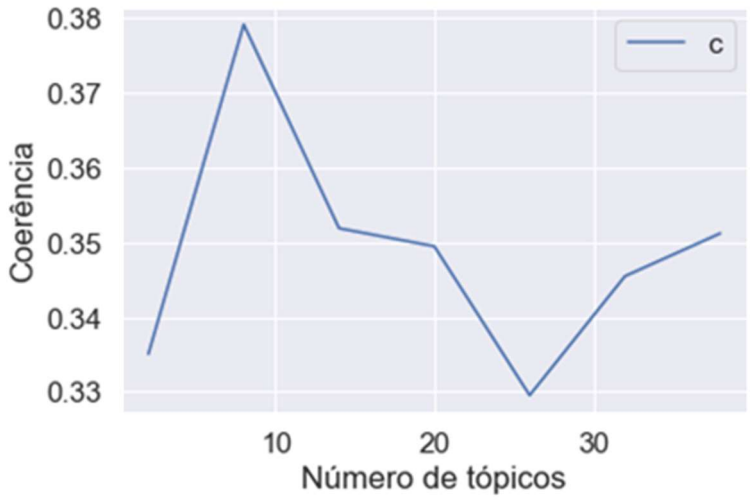


Figura 49 Coerência por número de tópicos do Algoritmo LDA

Apêndice D – Padrões relativo a satisfação dos clientes

Tabela 15 Estatística Alfa de Cronbach para estimar a confiabilidade do questionário.

Intervalo de confiança de 95%

raw_alpha	std.alpha	G6(smc)	average_r	S/N	ase	mean	sd	median_r
0,83	0,83	0,96	0,2	4,9	0,00086	-0,31	0,5	0,11

Tabela 16 Fiabilidade da estatística de Cronbach quando uma variável é eliminada

	raw_alpha	std.alpha	G6(smc)	average_r	S/N	ase	var,r
welcome_arrival	0,81	0,81	0,96	0,2	4,4	0,00094	0,092
welcome_departure-	0,81	0,81	0,95	0,19	4,1	0,00099	0,092
punctual_arrival	0,82	0,82	0,95	0,2	4,5	0,00092	0,09
clean_in_arrival	0,81	0,81	0,95	0,2	4,4	0,00094	0,092
clean_out_arrival	0,81	0,81	0,95	0,2	4,4	0,00094	0,092
clean_in_departure-	0,8	0,8	0,95	0,19	4,1	0,00099	0,092
clean_out_departure-	0,8	0,8	0,95	0,19	4,1	0,00099	0,092
punctual_departure-	0,83	0,83	0,96	0,21	4,9	0,00086	0,093
driving_arrival	0,82	0,82	0,95	0,2	4,5	0,00092	0,09
driving_departure-	0,83	0,83	0,96	0,21	4,9	0,00086	0,093
curteous_arrival	0,82	0,82	0,95	0,2	4,5	0,00092	0,09
curteous_departure-	0,83	0,83	0,96	0,21	4,9	0,00086	0,093
website	0,84	0,84	0,97	0,22	5,1	0,00084	0,101
experience-	0,82	0,82	0,95	0,2	4,4	0,00093	0,094
quality	0,84	0,84	0,96	0,22	5,1	0,00084	0,099
freegift	0,83	0,83	0,96	0,22	4,9	0,00086	0,099
wifi_arrival-	0,84	0,84	0,96	0,22	5,1	0,00083	0,096
wifi_departure-	0,83	0,83	0,96	0,21	4,9	0,00086	0,098
boot_arrival-	0,83	0,83	0,96	0,22	5	0,00084	0,09

Tabela 17 Estatística de Kaiser-Meyer-Olkin (KMO de adequabilidade dos dados.

KMO Global	
0,84	
KMO de Cada Variável	
welcome_arrival	0,96
welcome_departure-	0,97
punctual_arrival	0,89
clean_in_arrival	0,81
clean_out_arrival	0,81
clean_in_departure-	0,82
clean_out_departure-	0,82
punctual_departure-	0,85
driving_arrival	0,87
driving_departure-	0,8
curteous_arrival	0,86
curteous_departure-	0,78
website	0,72
experience-	0,91
quality	0,57
freegift	0,74
wifi_arrival-	0,64
wifi_departure-	0,73
boot_arrival-	0,79

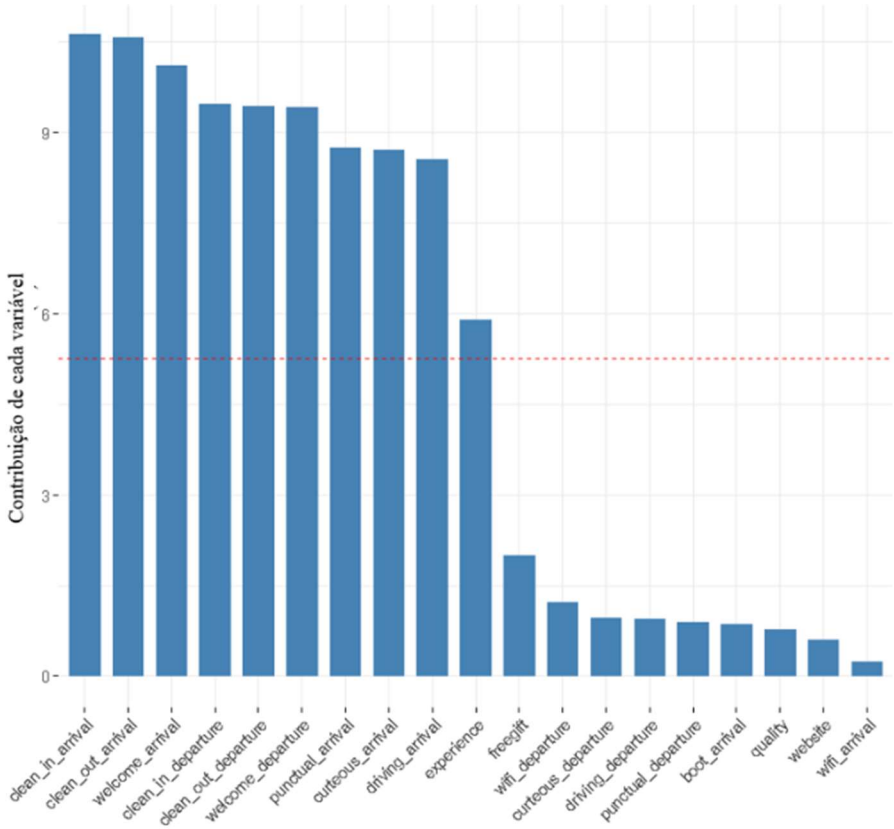


Figura 50 Contribuição das variáveis para a CP1

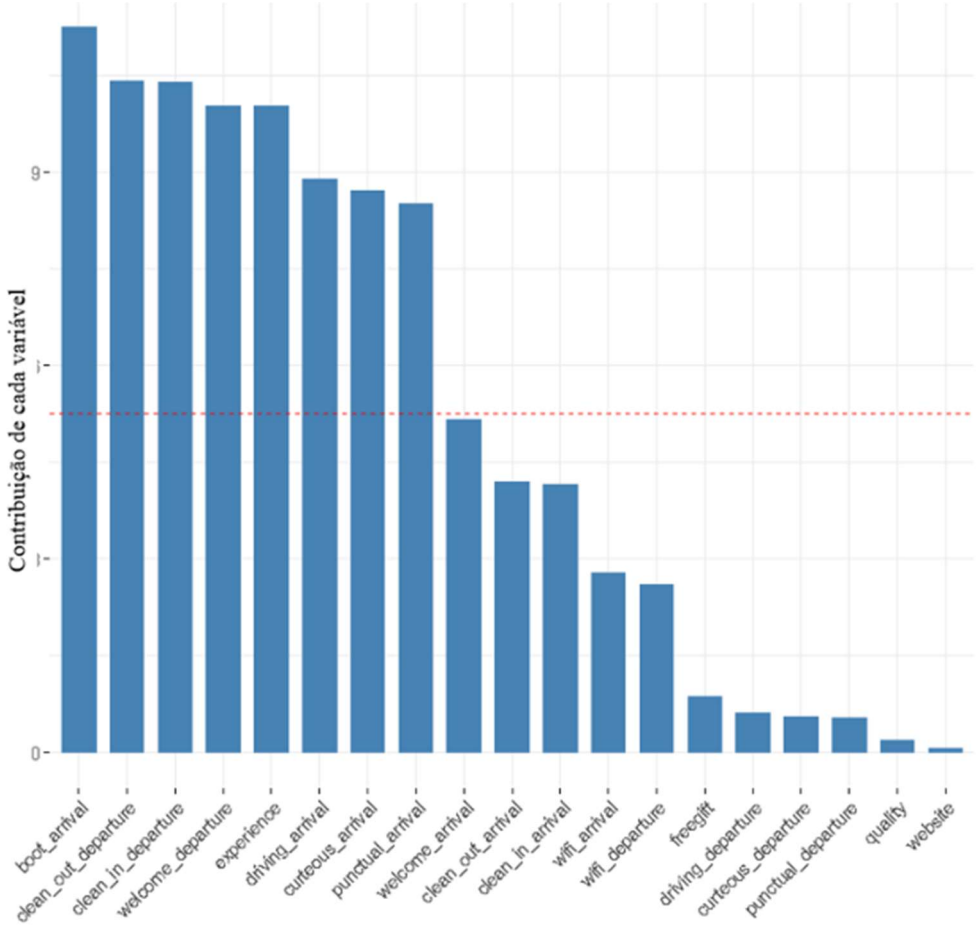


Figura 51 Contribuição das variáveis para a CP2

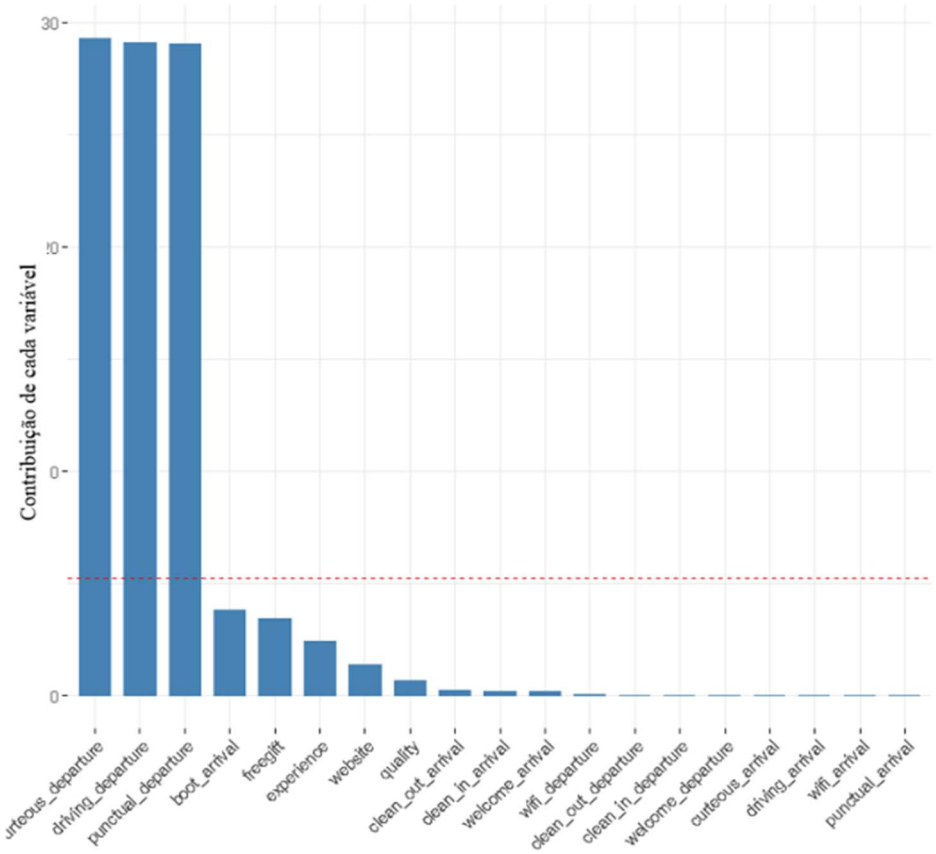


Figura 52 Contribuição das variáveis para a CP3

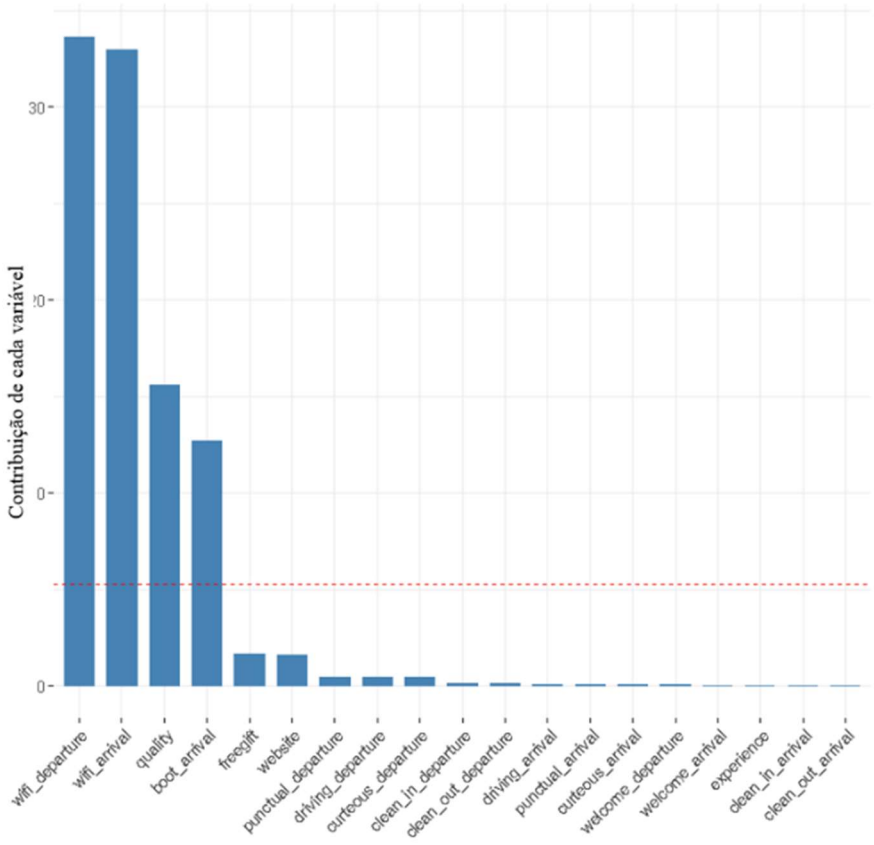


Figura 53 Contribuição das variáveis para a CP4

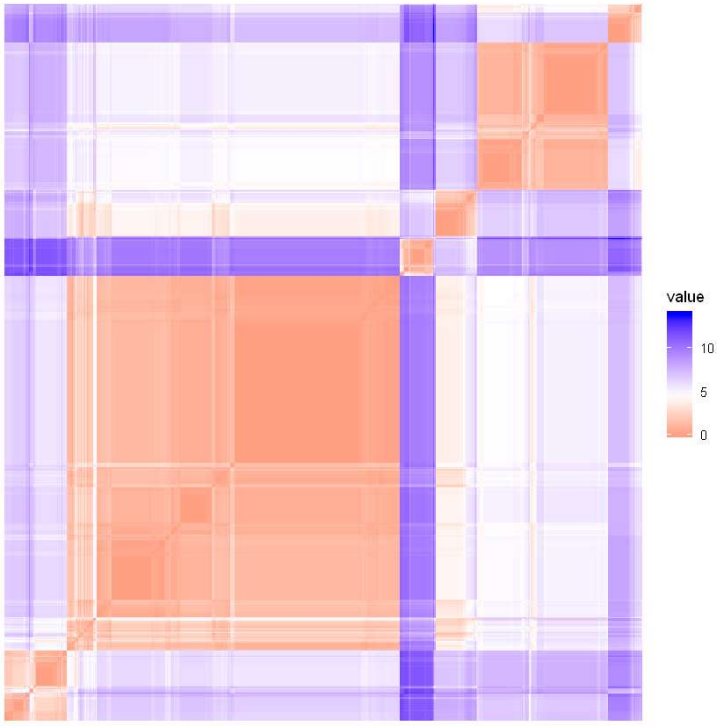


Figura 54 Método visual para avaliar a tendência de cluster nos dados. Vermelho, alta similaridade (i.e alta dissimilaridade); Azul, baixa similaridade.

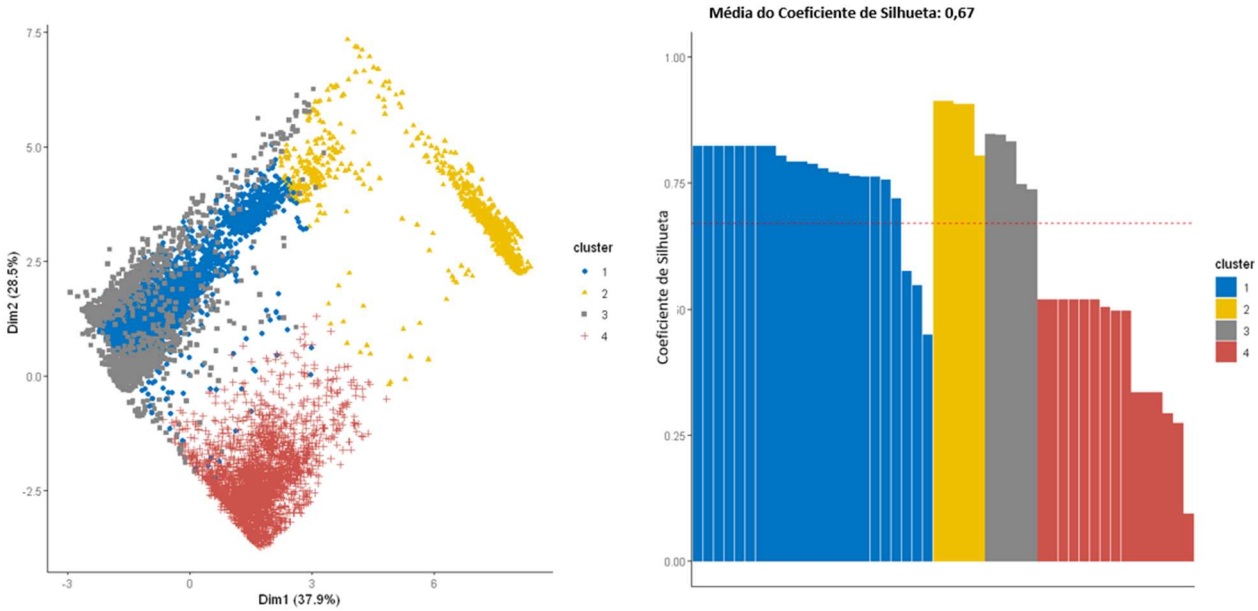


Figura 55 Visualização dos clusters e coeficiente de silhueta para K=4

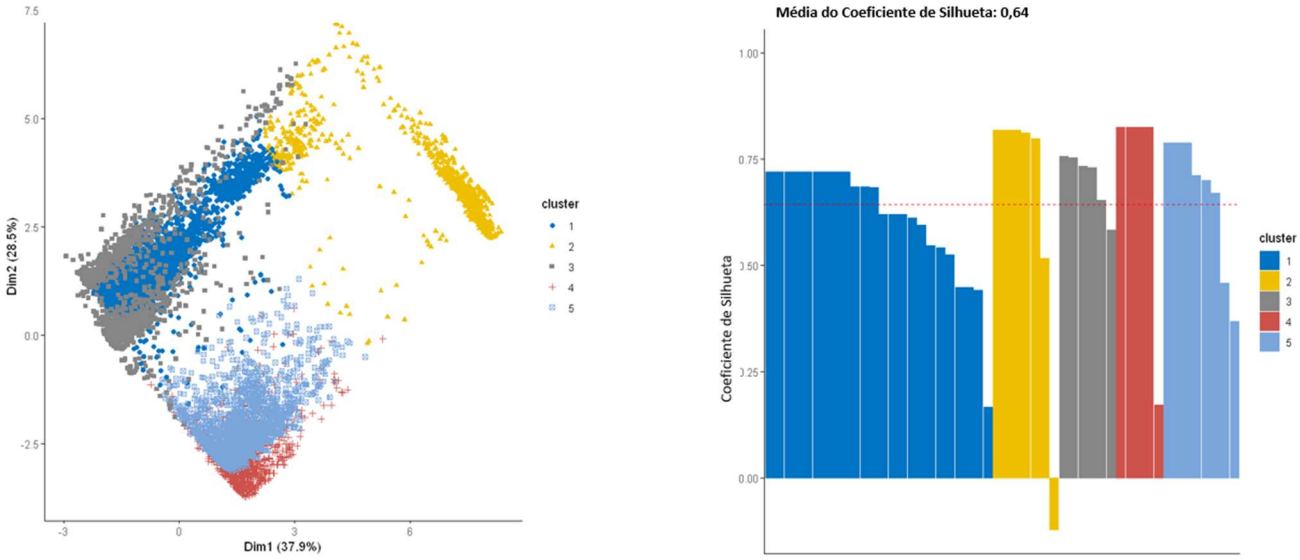


Figura 56 Visualização dos clusters e coeficiente de silhueta para K=5

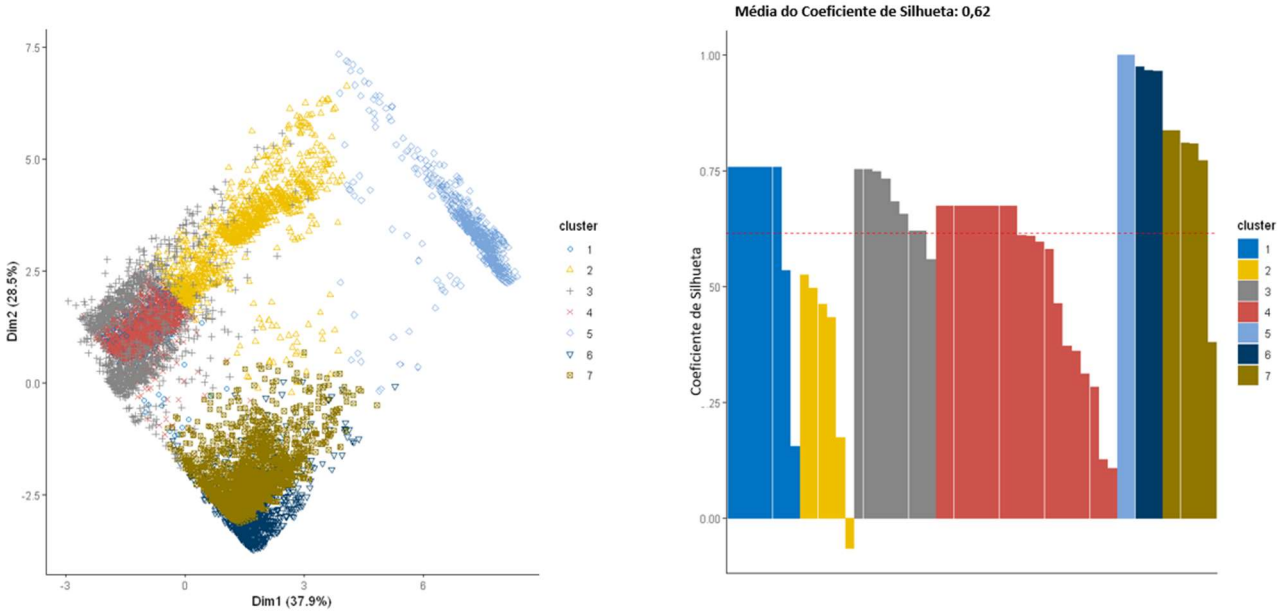


Figura 57 Visualização dos clusters e coeficiente de silhueta para K=7

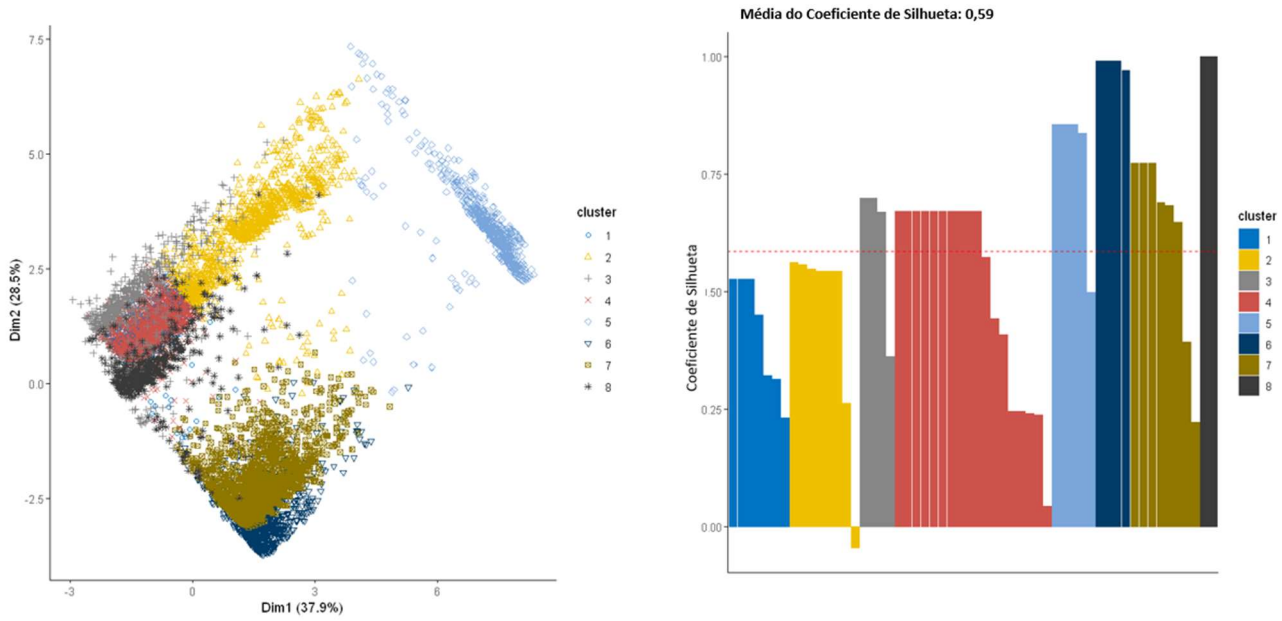


Figura 58 Visualização dos clusters e coeficiente de silhueta para K=8

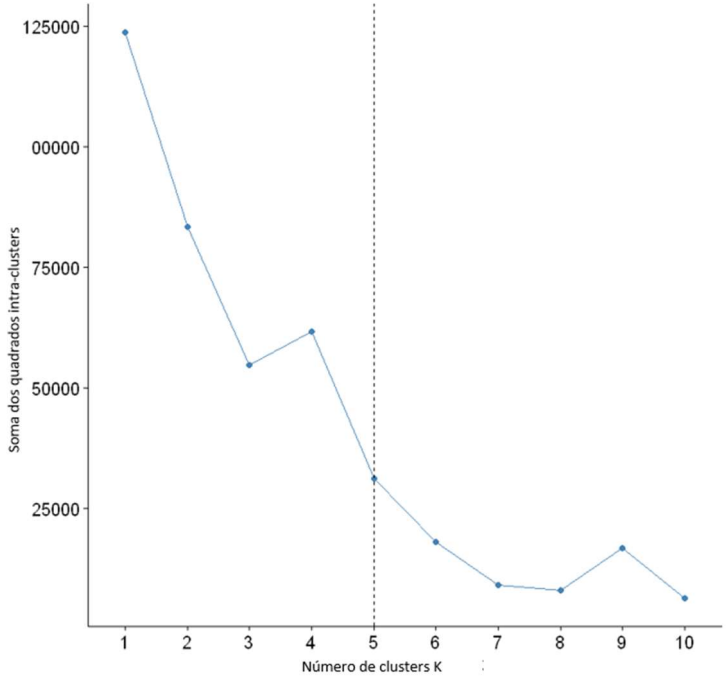


Figura 59 Número de clusters pelo método do cotovelo

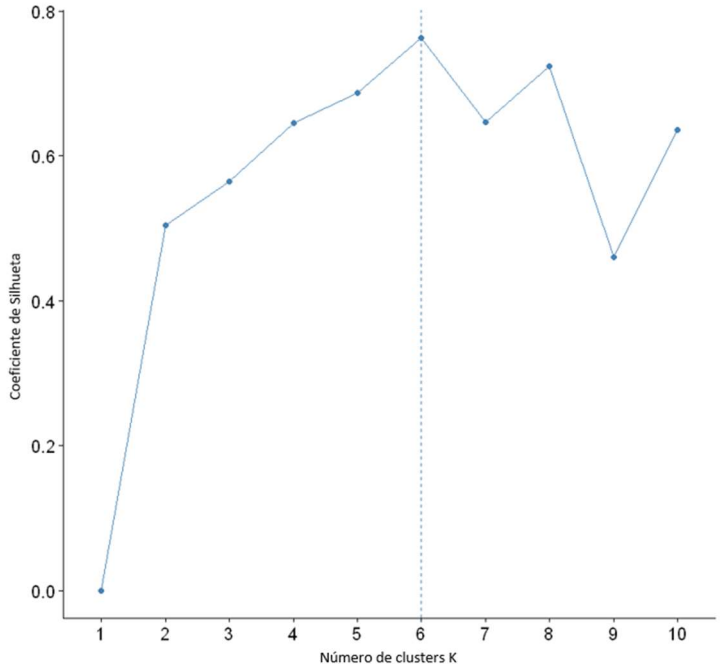


Figura 60 Número de clusters através do coeficiente de silhueta

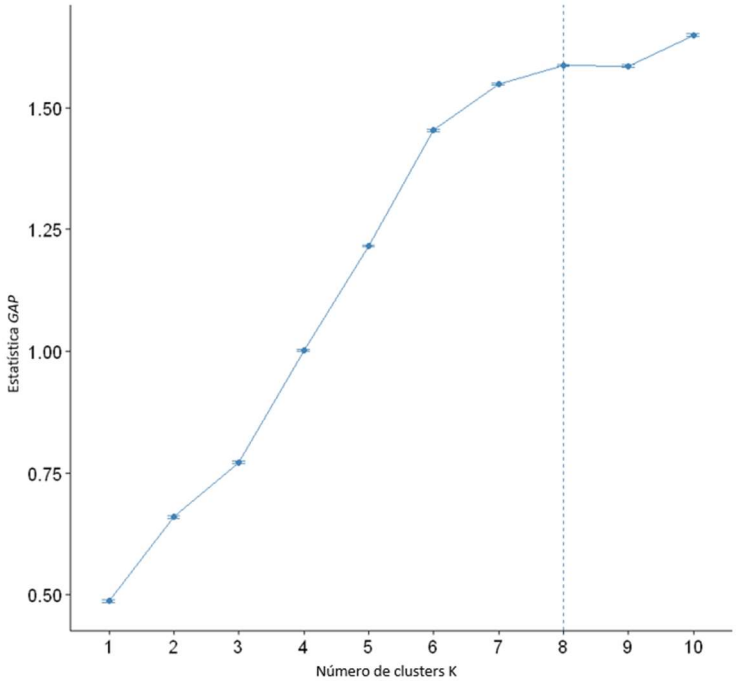


Figura 61 Número de clusters pelo método de estatística GAP

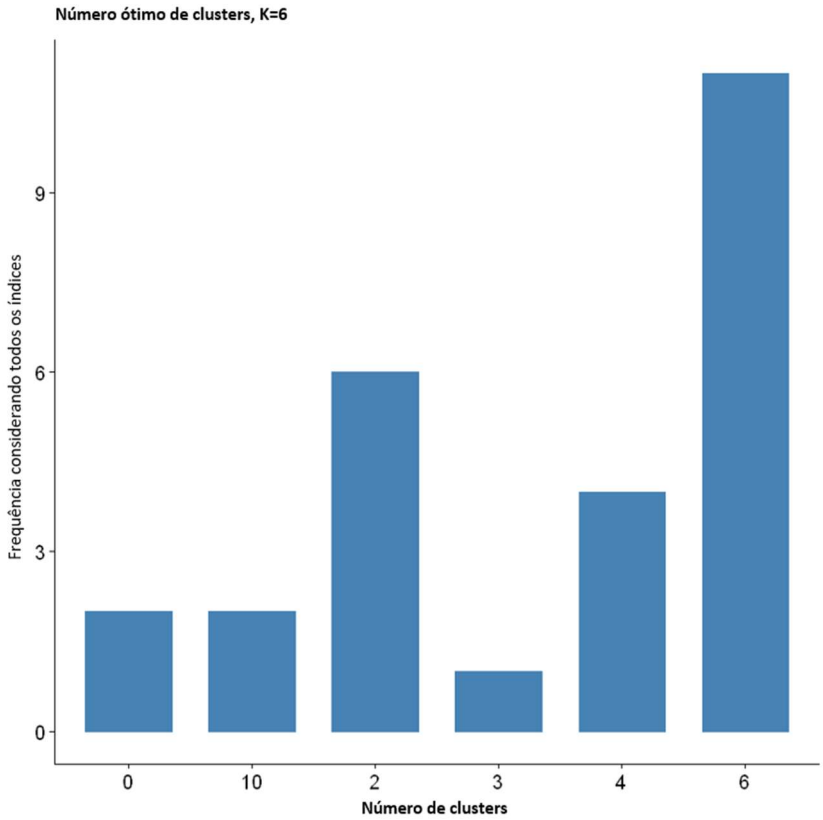


Figura 62 Número de clusters pelo critério da maioria de 30 índices

Apêndice E – Padrões relativo aos tipos de clientes

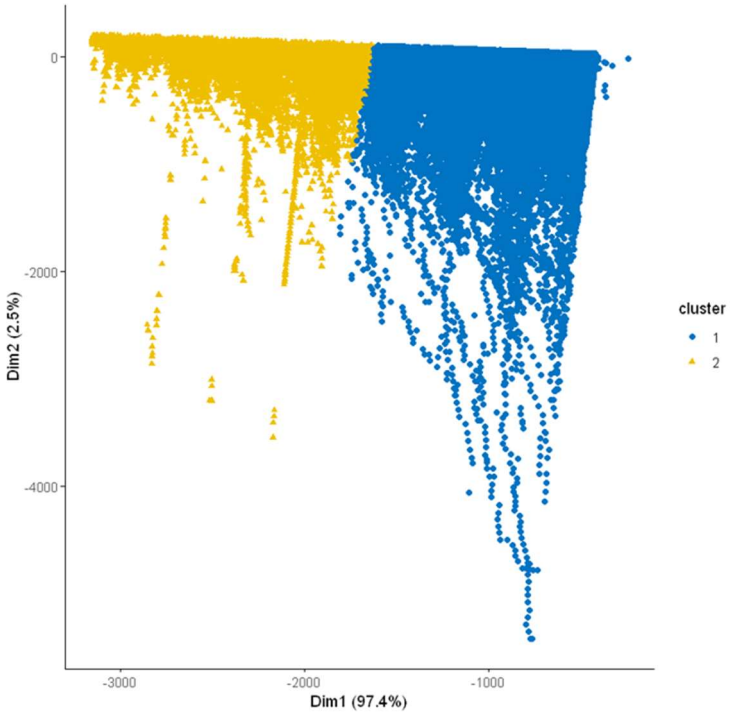


Figura 63 Visualização dos clusters com os dados sem tratamento

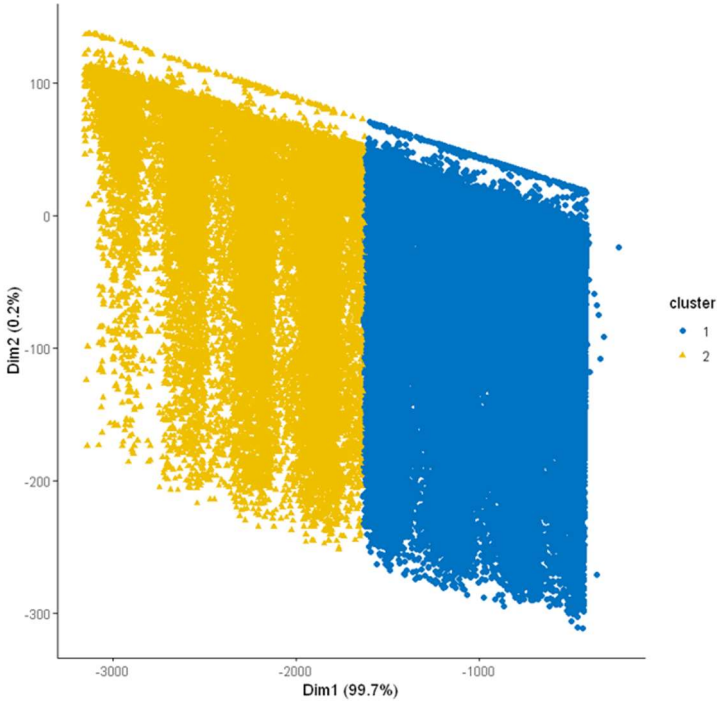


Figura 64 Visualização dos clusters com remoção de outliers

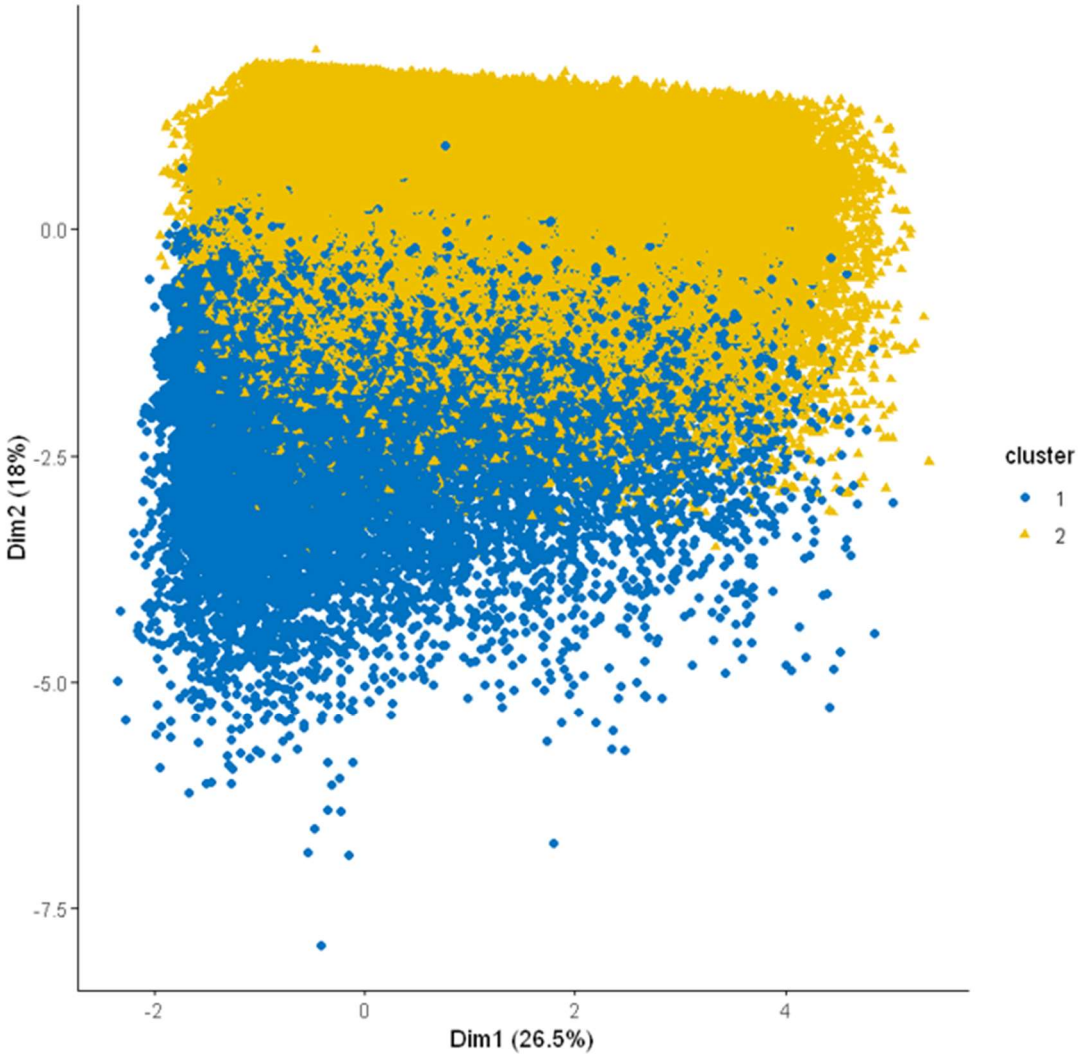


Figura 65 Visualização dos clusters com remoção de outliers e standardização dos dados

Apêndice F – Avaliação do modelo de regressão

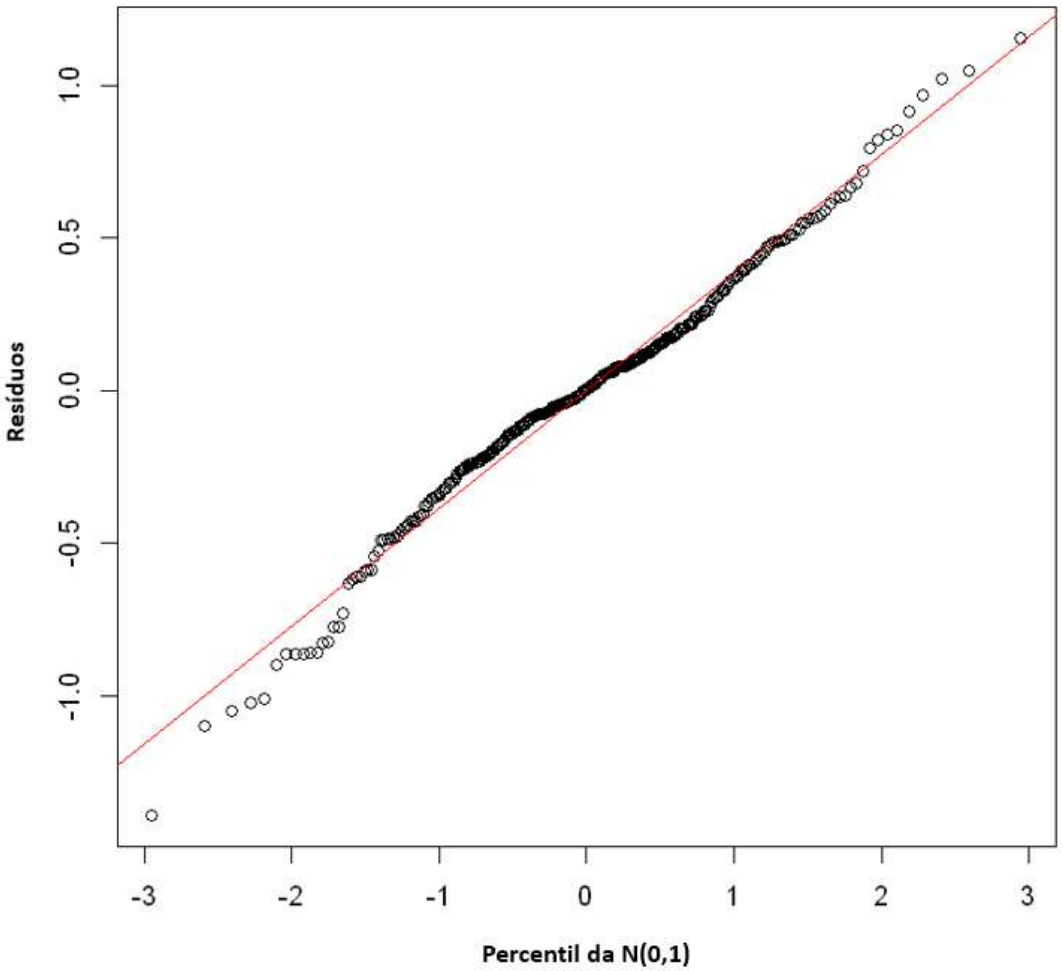


Figura 66 Q-Q Plot dos resíduos

Apêndice G – Análise exploratória da importância das variáveis

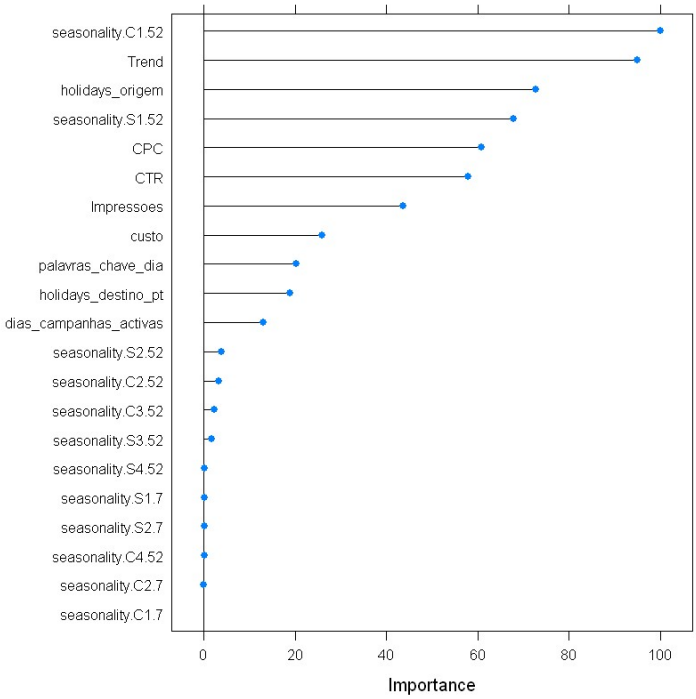


Figura 67 Importância das variáveis do algoritmo Multi Layer Perceptron

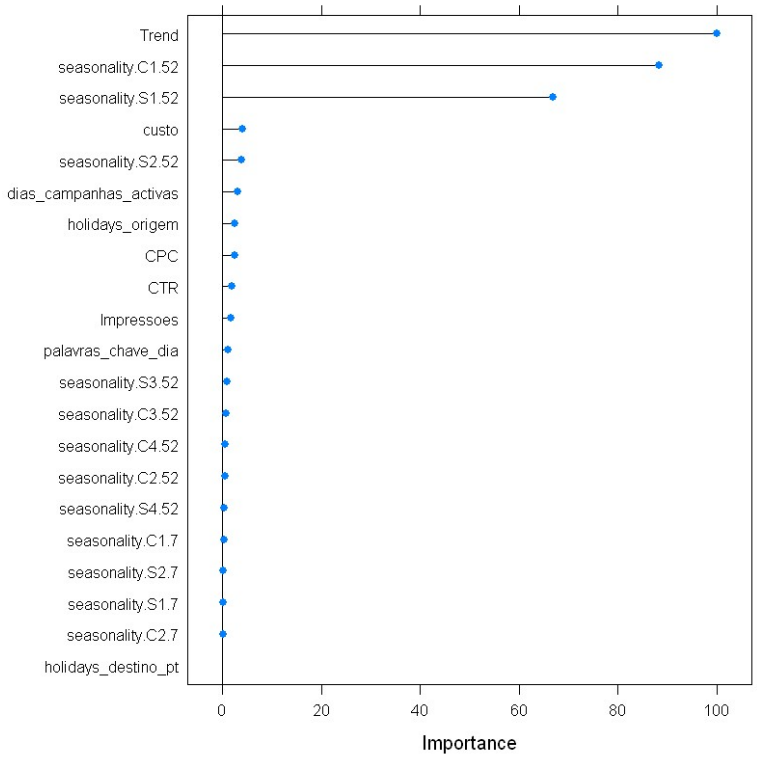


Figura 68 Importância das variáveis do algoritmo Random Forest

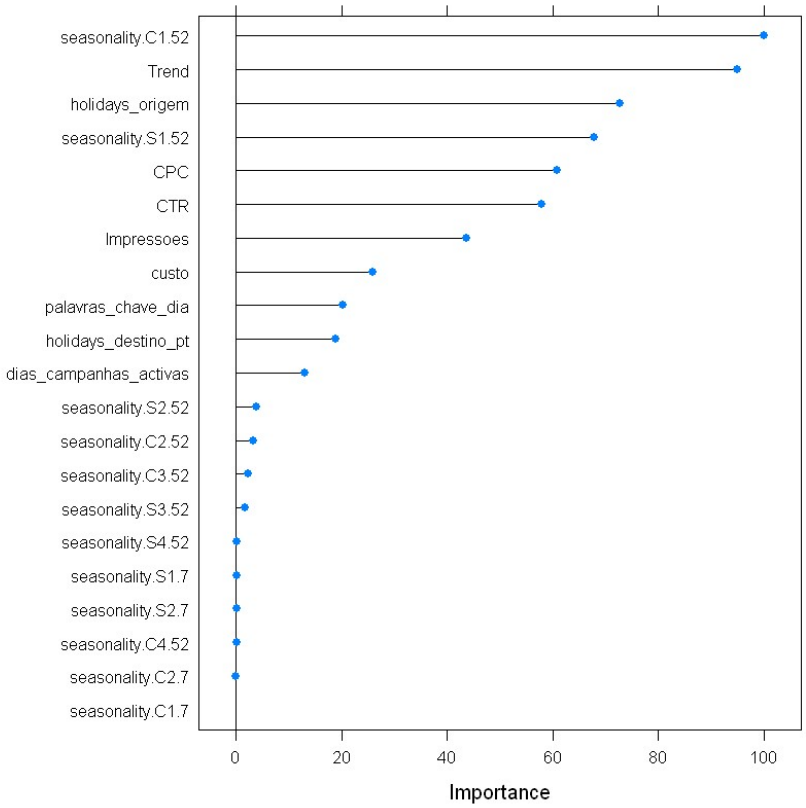


Figura 69 Importância das variáveis do algoritmo Suport Vector Machine

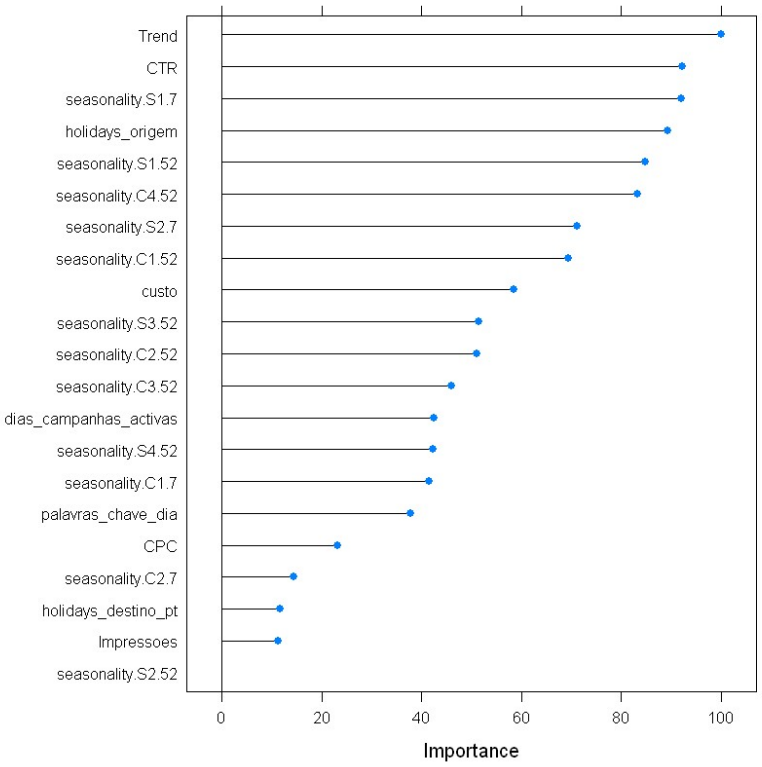


Figura 70 Importância das variáveis do algoritmo Nneural Net

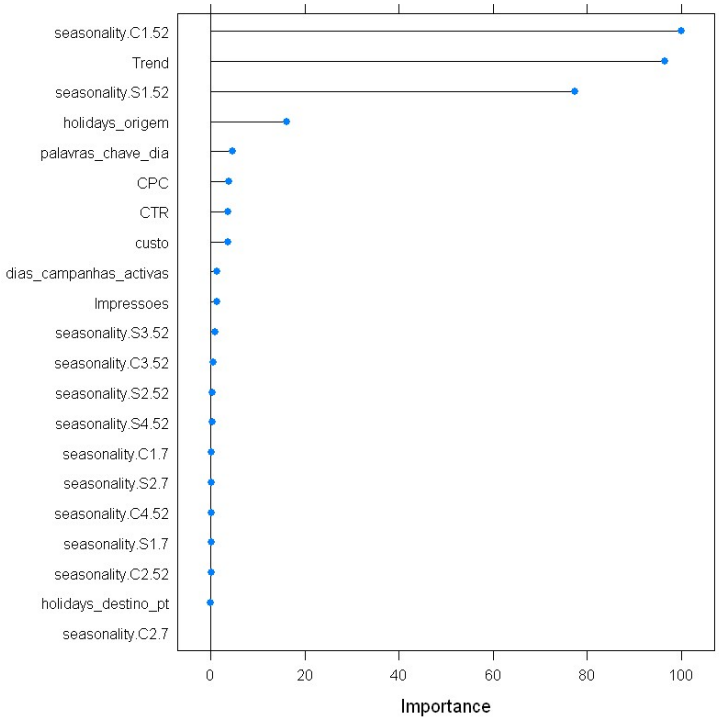


Figura 71 Importância das variáveis do algoritmo XGBOOST

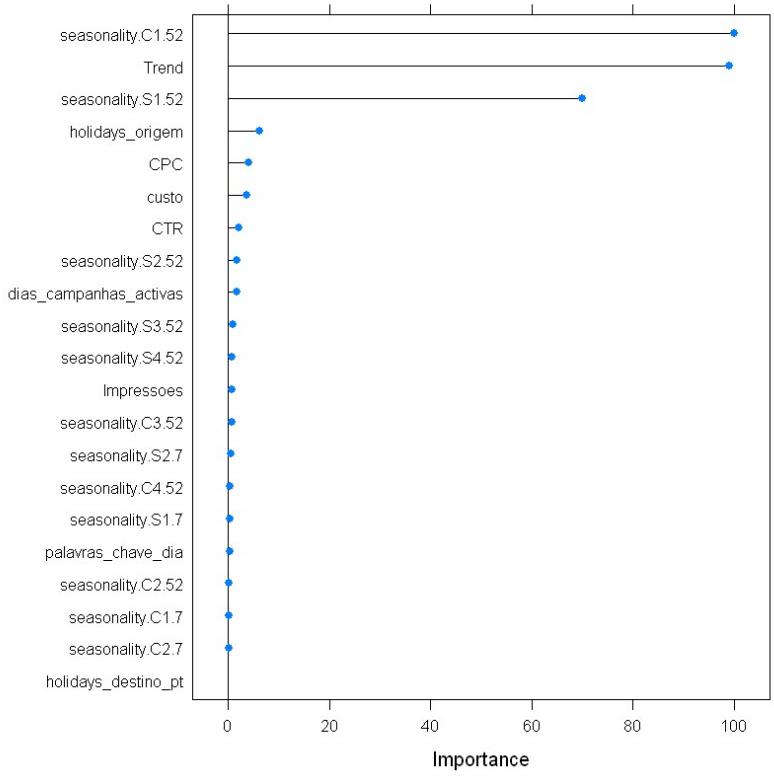


Figura 72 Importância das variáveis do algoritmo GMB