INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Road Accident Analysis in Lisbon

Ana Catarina Melanda Oliveira

Master Degree in Integrated Decision Support Systems,

Supervisor:
PhD João Carlos Amaro Ferreira, Assistant Professor with Habilitation,
ISCTE-IUL

Co-Supervisor:
PhD José Miguel de Oliveira Monteiro Sales Dias, Associate Professor with Habilitation,
ISCTE-IUL

October, 2021

Department of Information Science and Technology

Road Accident Analysis in Lisbon

Ana Catarina Melanda Oliveira

Master Degree in Integrated Decision Support Systems,

Supervisor:
PhD João Carlos Amaro Ferreira, Assistant Professor with Habilitation,
ISCTE-IUL

Co-Supervisor:
PhD José Miguel de Oliveira Monteiro Sales Dias, Associate Professor with Habilitation,
ISCTE-IUL

October, 2021

Aknowledgements

Achieving this goal wouldn’t be possible without the knowledge acquired over the past years neither without the people whose paths have crossed mine.

To my advisor and co-supervisor, Professor João Carlos Ferreira and Professor José Miguel Sales Dias, for all the guidance and availability. A special thanks to Vitória Albuquerque for dispending her time helping and motivating me, and for pushing me to do better and never give up.

To my family, especially to my mum and dad, for all the support, patience, motivation. Without them, this could not be possible. Thanks for being always by my side, and for always believing in my skills and capacities.

To my friends for being so supportive and for all the laughs, coffees to discuss ideas, and also for understanding the time that I didn’t spend with them.

Last but not the least, I would like to thank my workmates who helped me to improve my professional and personal skills.

My sincere "Thank you" to all that I met along the way.

Resumo

A mobilidade urbana nas grandes cidades europeias tem sido cada vez mais estudada devido ao elevado volume de dados e de interesse que existe sobre a mesma. Como tal, as autoridades competentes sentem a necessidade de desenhar soluções inteligentes que auxiliem na mitigação de problemas de mobilidade. O presente trabalho de investigação foi desenvolvido com os dados de mobilidade da Câmara Municipal de Lisboa, nomeadamente dos acidentes rodoviários ocorridos no ano de 2019 nesta cidade, através da abordagem CRISP-DM em Python. Os dados foram previamente integrados e limpos para posteriormente serem submetidos a métodos de visualização, de forma a identificar padrões de ocorrência de acidentes rodoviários na cidade de Lisboa.

Palavras-Chave: Acidentes rodoviários; Análise de dados; Fatores externos; Mobilidade; Visualização de dados.

Abstract

Studies about urban mobility in big European cities have been increasing due to the high volume of data and interest that exists about this topic. As such, competent authorities feel the need to design intelligent solutions that help to mitigate mobility problems. This research work was developed using mobility data from the Câmara Municipal de Lisboa, namely road accidents that occurred in 2019 in this city, using the CRISP-DM approach in Python. The data were previously integrated and cleaned to later be submitted to visualization methods, to identify patterns of occurrence of road accidents in the city of Lisbon.

Keywords: Data analysis; Data visualization; External Factors; Mobility; Road accidents

Index

Tables Index

Figures Index

List of abbreviations

ANSR - Autoridade Nacional de Segurança Rodoviária

CML - Câmara Municipal de Lisboa

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

GNR - Guarda Nacional Republicana

PGIL - Plataforma de Gestão Inteligente

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PSP - Polícia de Segurança Pública

RQ - Research Question

SLR - Systematic Literature Review

Chapter 1 – Introduction

1.1. Topic context

The population inhabiting metropolitan areas around the world is increasing at an alarming rate. In 2008, 50% of the world's population lived in urban areas [1] and was growing exponentially. By 2050 [1] it is expected that 70% of the world's population will live in metropolitan regions. Due to the rapid population growth, cities will face new challenges [2], such as waste, pollution, traffic congestion, and road accidents increase.

The World Health Organization (WHO) estimates road accidents are the ninth leading cause of death globally across all age groups and are the main cause of death among people between 15 and 29. Furthermore, road accidents are a cause of life losses and bring health and socioeconomic costs [3].

Worldwide, countries are adopting measures to decrease road fatalities. Speed management, infrastructure design and improvement, enforcement of traffic laws, leadership on road safety, vehicle safety standards, and post-crash survival are some of the currently ongoing initiatives that aim to mitigate this hazard [3]. A large number of road accidents occur due to various factors that directly or indirectly affect conditions on the road for drivers, passengers, and pedestrians [4]. Factors such as gender and age, or environmental factors such as low brightness (dawn or dusk) and adverse weather conditions [5], influence the world's global number of road accidents.

In Portugal, road accidents are considered one of the most serious problems in nowadays Portuguese society and a public health issue. According to Autoridade Nacional de Segurança Rodoviária (ANSR) [6], the number of accidents with victims increased by 4% in 2019 compared to 2018, with a reduction of 9% in the number of fatalities. Despite this fall in fatalities, serious injuries increased by 9%. The geographic distribution of road traffic victims in 2019 shows that Lisbon and Porto have approximately 40% of the country's total number of victims.

This study aims to present an overview of road accidents in Lisbon with 2019 data, provided by Câmara Municipal de Lisboa (CML) and Autoridade Nacional de Segurança Rodoviária (ANSR). The data was collected by law enforcement agents, such as Polícia de Segurança Pública (PSP) and Guarda Nacional Repúblicana (GNR). The data is collected when an accident occurs, either by PSP or GNR. They are called to collect and fill ANSR form with information on the driver's age, road, weather, and much more, which is later added to the ANSR system.

Road traffic accidents happen everywhere at any time, as such, it is important to know which pattern conditions when most accidents occur. This study pretends to present a big picture of road accidents in Lisbon by analyzing data of 2019 and sharing knowledge with authorities on locations and patterns where accidents are more likely to occur and mitigate it.

1.2. Motivation and topic relevance

Road accident analysis is essential to help authorities to take early measures to reduce the occurrence of road accident fatalities.

Road accidents have an economic cost and a social one, with impacts on the National Gross Domestic Product (GPD) [3].

Factors such as the conditions of the driver, the vehicle, and the infrastructure are the most determinant in a road accident.

The analysis of road accidents is very important to help the authorities, such as the city council, ANSR, among others, to identify the roads and areas with the highest accident occurrence. Therefore, to take early action with preventive measures and reduce the road accident occurrence through actions, namely at the signage level (traffic lights, vertical signage, etc.), or placement of devices to promote drivers' speed reduction. Also, it helps authorities to define guidelines and to organize awareness campaigns on risk behaviors and their consequences, as well as the importance of using restraint devices, targeting age groups that are most involved in road accidents.

Moreover, the thesis motivation and relevance regard the fact that more vehicles circulating and more people with a driving license promote more traffic, hence more road accidents. Therefore, the need to tackle road accidents in Lisbon is the motivation to develop this analysis.

On the other hand, with the increase of road accident data, there is a need to analyze it and correlate it with other data, such as traffic, weather data. Furthermore, with the increase of the volume of data generated in cities, public entities such as Câmara Municipal de Lisboa (CML) can take advantage of their value, to improve decision-making. The analysis of this data through analytics and data visualization solutions aims to improve operational and emergency management in the city of Lisbon, contributing to the sustainable improvement of the resilience and quality of life of those who live, or visit it [7].

1.3. Research questions

CML and Lisboa Inteligente [7] launched a challenge on road accidents in Lisbon and provided data to the academic community. The research is on the scope of this challenge The main goal is to analyze and characterize road accidents patterns in Lisbon. This dissertation aims to answer the following research questions (RQs):

1. How can road accident patterns in Lisbon be characterized?
2. What are the external factors that contribute the most to this phenomenon?

1.4. Objectives

The objective of the research is to investigate road accidents in Lisbon by analyzing data of 2019. The main objective is to identify Lisbon road accident patterns and external factors by type and characteristics. This research also aims to provide insights on road accidents in Lisbon in the framework of a traffic management as well as a visualization component of this phenomenon. Moreover, pretends to help policymakers and stakeholders to make better-informed decisions on traffic management in order to mitigate road accidents in Lisbon.

1.5. Methodologic approach

The Cross-Industry Standard Process for Data Mining (CRISP-DM) was applied in the research analysis. This methodology has a standard approach to data-mining projects to reduce costs and increase reliability, repeatability, and manageability, making the data-mining process more efficient [8].

CRISP-DM [9] has six phases, although for this research, was adapted to 5 phases (Figure 1-1), for specific features and requirements aimed by CML regarding data characterization and visualization of road accident patterns in Lisbon. The first and second phases, business understanding and data understanding, were merged in one. In this phase, the strategy to address the research questions was defined and road accident datasets were explored described, and categorized in features. In the third phase, data preparation, data provided was cleaned, and data quality was validated and normalized, followed by data integration. In the fourth phase, data fusion, the data sources, such as accidents and city infrastructure, were selected. In the fifth phase, data visualization, visualization templates were defined to automatically visualize spatial data and achieve needs and defined goals by CML. Finally, the last step was decision and visualization of information.

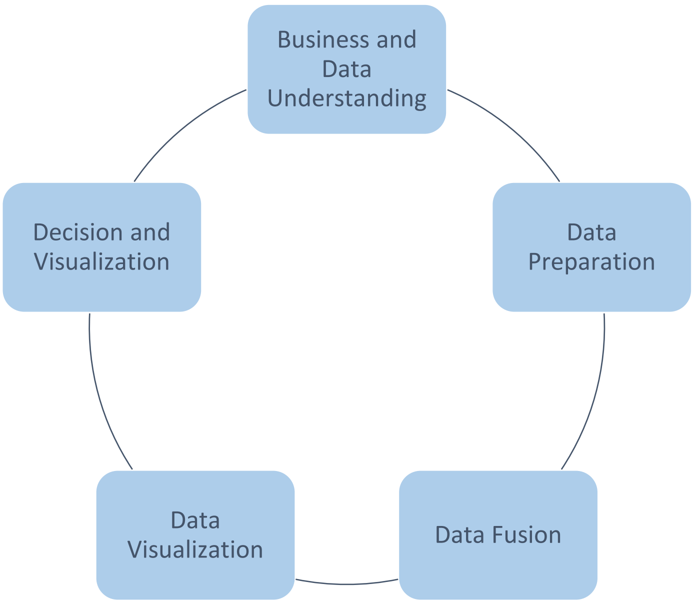

Figure 1-1 - CRISP-DM

1.6. Structure and organization of dissertation

The dissertation is organized into four chapters:

The first chapter introduces the subject of the investigation, its RQs, and objectives, as well as a brief description of the work structure.

The second chapter is dedicated to the Literature review.

The third chapter is dedicated to the dissertation framework used in collecting and processing data, the methods of analysis used, and analysis of the results obtained, according to the proposed methodology.

The fourth chapter presents the discussion and conclusions of this study as well as the recommendations, limitations, and future work.

Chapter 2– Literature review

2.1. Road accidents context

Studies [3], [10]–[12] on road accidents in Spain, India, and the United States of America (Washington) showed different approaches and results. Palazón-Bru [5] studied Spain's road accident data from 2015 and considered variables associated with the accident, the vehicle, and individuals. This information can be used by both police authorities and health services to predict and determine where to undertake possible interventions to reduce death risk. Factors, such as not using a seatbelt, unfavorable lighting, interurban roads with higher speed, and small vehicles driving alongside buses or trucks (which, in the event of a small vehicle rollover, could cause its occupants' death), were the more prevalent ones. The author [13] concluded that individuals younger than 60 years had lower mortality rates.

Road accidents in India were analyzed, and the author [4] concluded that between 3 p.m. and 6 p.m. was when most occurred, with a clear correlation to the peak traffic hours, although it deferred slightly between states. Furthermore, the study [4] showed that two-wheeled vehicles were involved in more accidents, and over 80% were the driver's fault. The authors also concluded that accident severity was growing due to the increase in the number of vehicles. Also, 61% of accidents occurred on the weekend (Friday and Saturday), which correlated with the days of highest alcohol consumption.

Most analyzed road accident papers present analysis and prediction methodologies combined. As such, it will be discussed how the scientific community has addressed road accident prediction.

Regarding road accidents, Norros [14] found that factors such as weather, traffic, and location could be included as variables when creating models to predict road accidents. For example, driver characteristics do not depend on the time of day, traffic, or weather, but weather and other external conditions affect traffic intensity and constitution.

Tang [12] used the Washington Incident Tracking System and applied a combined analysis using eight methods: four statistical methods (accelerated failure time (AFT), finite mixture (FM), random parameters hazard-based duration (RPHD), and quantile regression (QR); and four machine-learning methods (K-nearest neighbor (KNN), support vector machine (SVM), BPNN, and random forest (RF)) used in the traffic-incident clearance-time analysis. All showed that temporal factors like day of the week, time of day, and month of year influence accidents.

Calvert [15] introduced the concept of Anticipation Reliance (AR), which acts as a demand lowering compensative effect for the driving task by relying more on anticipation. Hence, human factors were included in mathematical models such as AR which introduces a ground-breaking concept that explains and models the mechanisms that allow drivers to compensate and avoid accidents in many circumstances. Driver's anticipation can be influenced by internal and external factors, especially when multiple driving tasks are considered. AR allows drivers to perform multiple driving tasks without becoming cognitively oversaturated through their tasks.

Adanu [16] understands and analysis human-centered crashes, with data from Alabama, that contains regional differences between cities and countries. The not use of seatbelts and driving under the influence of alcohol vary across states. With a framework that helps to understand a broader spectrum of factors contributing to crashes, strategies to reduce crashes and their negative outcomes can be taken.

Mehdizadeh [17] focuses to understand and quantify crash risk based on different driving conditions and also on minimizing crash risk through route selection and rest-break scheduling. Data collection, data exploration, and predictive modeling were the three phases taken to follow the standard data analytics framework. Risk factors (sleep and fatigue, distracted driving, weather, traffic conditions, and road geometry) that affect crash risk and statistical, machine learning models were studied, and authors concluded that there are a lot of accessible data, descriptive analytics tools are widely used in the pre-processing of driving-related data and there is an opportunity for statistical analysis of a larger scale.

This review is the continuation of Mehdizadeh [17] and Hu [18] work on the optimization, and prescriptive analytic models that focus on minimizing crash risk, especially hazardous materials (hazmat). Four popular predictive models (logistic regression, Poisson regression, neural networks, and XGBoost) were used to predict the probability of a crash or the number of crashes as a function of the aforementioned predictors, and then k-shortest path algorithm to identify the shortest routes ranked by the distance between two nodes. The logistic regression, Poisson regression, and XGBoost models indicate that the shorter the route, the less likely one is involved in a crash; the neural network shows an inverse relationship where maybe some “safety” benefits from selecting longer routes.

Nidhi [19] used data from USA states on Naïve Nayes (presumption of independence between each pair of variables) and K-means clustering algorithm. Naïve Nayes showed that the environmental factors like roadway surface, weather, and light condition do not strongly affect the fatal rate, while the human factors like being drunk or not, and the collision type, have a stronger effect on the fatal rate. K-means resulted in 3 clusters A, B, and C, and showed that some states/regions have higher fatal rates, while some others lower.

Kumar [20] with data from Dehradun (India) between 2009 and 2014 proposed a framework on K-modes clustering and association rule mining algorithm. Six clusters were created by k modes clustering based on attributes accident type, road type, lightning on road, and road feature.

Gutierrez-Osorio [21], due to the impact on public health and socio-economic, made a review about the prediction of road accidents through algorithms. Data were obtained from different sources, and clustering, decision trees, and classifiers, and natural language processing were used as analytic methods for road accident analysis. made prediction of road accidents through machine learning algorithms

Sodikov [22] analyzed data from a road traffic accident in R programming environment to assist reduce the number of accidents. The author analyzed data of traffic accidents, collisions, and collisions with pedestrians, in Tashkent city, from 2005 to 2012. Regarding pedestrian collision, Friday was the day of the week with more occurrences, especially March and September. About collisions in general, August was the month with more occurrences, and more likely to occur on Saturdays.

Wegman [23] developed a new approach of evidence-based and data-driven road safety management: ex-post and ex-ante evaluation of both individual interventions and intervention packages in road safety strategies, and transferability (external validity) of the research results. Western Australia, the Netherlands, Sweden, and Switzerland were the jurisdictions where the approach was applied to understand how exactly the implementation of (a multitude of) road safety interventions has influenced the positive road safety developments in many countries. Ex-post and ex-ante evaluations deal with the relation between policy output of interventions (safety measures and programs) and outcomes in terms of the number of people killed or seriously injured and the associated social costs. They concluded that fully evidence-based and data-driven road safety management is too complicated at the moment and not a realistic option, and measuring safety performance indicators could be expensive. Although, all the jurisdictions have made progress over the years.

As road accidents, disasters cause loss of lives and immeasurable economic losses. For Li [24], through data-driven disaster manager, it is possible to reduce the losses caused by disasters. With data from different partners and entities it is valuable to interconnect all data which allows users to find valuable information. Besides being a challenge the integration of data from distinct sources into one offers real-time services and enables efficient information digestion and quick response for users in disaster situations.

Road accidents number are increasing due to the growth of cars on road. Prashant [4] analyzed and contemplate the solution to accidents and accident response in urban areas. Authors used algorithms such as decision trees to classify accidents as more or less severe and k-means clustering to cluster regions in which accidents have happened to find ‘hotspot’ zones of accidents. Authors concluded that around 61% of accidents involving alcohol happen on the weekend, the accident severity index (deaths per 100 accidents) has shown an increasing trend from 2012-2016 and the percentage of accidents involving heavy-duty vehicles is a substantial number compared to the number of those vehicles on the road.

Oviedo-Trespalacios [25] made a systematic review about driver inattention and distraction because both are recognized as two of the most critical factors for road safety worldwide. Task-Capability Interface (TCI) Model, a seminal theoretical framework that explains determinants of driving behavior and crashes risk, was used to explain the potential safety impact of roadside advertising. Besides the research, it was not possible to conclude that there is a direct relationship between the driving behavior changes that can be attributed to roadside advertising and subsequent road crashes.

Due to a large number of loss of life in road traffic crashes, Kabli [26] proposed high-resolution crash severity models based on driver injury severity reported using the Abbreviated Injury Scale (AIS) by body region. Authors used a dataset with information about driver characteristics, vehicle characteristics, crash characteristics, roadway characteristics, and environmental characteristics from 2005 to 2015 and concluded that more factors can influence model estimation results, and more occupants besides drivers will increase the order of the dependent variables at the rate of six per additional occupant.

Nour [27] used data from the UK government from 2005 to 2019 to use predictive modeling techniques to identify risk and key factors that contribute to accident severity. After data collection and storage and data pre-processing, authors applied classification methods like logistic regression models, deep neural networks, support vector machines, decision trees, extreme gradient boosting to analyze 63 attributes and their relation with accident severity and concluded that tree-based techniques such as XGBoost outperform regression-based ones, such as ANN.

2.2. Literature review methodology

Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) methodology was used to perform our Systematic Literature Review (SLR). By following the PRISMA methodology, the PRISMA checklist phases were adapted, which help as a guide to the researcher [28]. PRISMA flow diagram is part of the process and depicts the flow of information through the different phases of an SLR. It maps out the number of records identified, included, and excluded, and the reasons for exclusions [29]. After the full-text screening phase, articles out of scope were removed, resulting in the final included articles of SLR.

The SLR of this research was performed using the Institute of Electrical and Electronics Engineers (IEEE), Science Direct, and Web of Science repositories. On Science Direct, the articles were filtered by subject area Decision Science, to get the best-fitted ones in the research scope.

The articles for the SLR focused on the latest state of the art, published between 2016 and 2021 and in English only.

The keywords identification and search in the repositories was performed with the query: (road accident) AND (external factors) AND (car) AND (analysis) NOT (autonomous) NOT (smart cities) NOT (bus).

2.3. Literature review results

2.3.1. PRISMA flow diagram

The PRISMA flow diagram (Figure 2-1) was performed to achieve SLR analysis, starting with the identification phase, based on the articles found on the selected repositories and queries. Duplicated records were removed in the identification phase, followed by an abstract screening phase where records unrelated to road accidents or data analysis were excluded.

The result was 74 records, and 19 were added through other sources as literature of reference in the field, resulting in a total of 93 eligible articles. Of these 93 eligible articles, 38 were excluded because were not related to the subject. In the next phase, 55 full-text articles were assessed for eligibility, 35 were excluded as they were off-topic, mainly for their approach to other transport modes and reference to cognitive aspects.

After the full-text screening, 20 studies were included for the SLR.

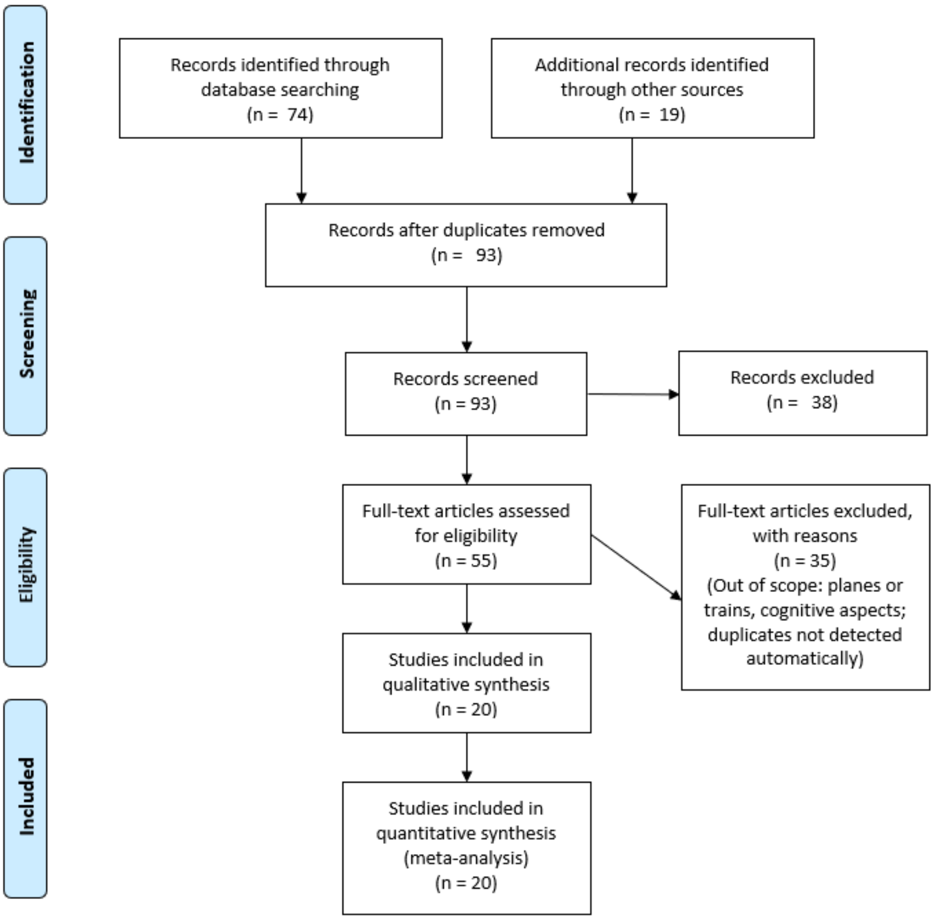

Figure 2-1 - PRISMA Flow Diagram

2.3.2. Identification of publication

In this study, a total of 20 papers were analyzed, including 6 papers published on Analytic Methods in Accident Research, 1 paper on ACM Computing Surveys, 1 paper on IATSS Research, 1 paper on International Journal of Advanced Computer Science and Applications, 1 paper on Journal of Traffic and Transportation Engineering, 2 papers on Sensors, 1 on International Journal of Civil, Structural, Environmental and Infrastructure Engineering Research and Development, 1 International Journal of Environmental Research and Public Health, 1 paper on Journal of Big Data, 1 paper on Proceedings 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, 1 paper on International Journal of Engineering and Technology (UAE), 1 paper on Journal of Big Data, 1 paper on Transportation Research Interdisciplinary Perspectives, 1 paper on Transportation Research Part B: Methodological.

As shown in Table 2-1, all of the top 5 most cited papers are Q1-quartile-ranked.

Overall, 10 papers are Q1-quartile-ranked, 3 are Q2-quartile-ranked and 1 are Q3-quartile-ranked.

Table 2-1 - Author, title, year, publication, and quartile ranked by citation

N. of citations	Authors	Title	Year	Publication	Quartile rank
46	Fountas, Grigorios; Anastasopoulos, Panagiotis Ch.	Analysis of accident injury-severity outcomes: The zero-inflated hierarchical ordered prohibit model with correlated disturbances	2018	Analytic Methods in Accident Research	Q1
35	Li, Tao; Xie, Ning; Zeng, Chunqiu; Zhou, Wubai; Zheng, Li; Jiang, Yexi; Yang, Yimin; Ha, Hsin Yu; Xue, Wei; Huang, Yue; Chen, Shu; Ching Navlakha, Jainendra; Iyengar, S. S.	Data-Driven Techniques in Disaster Information Management	2017	ACM Computing Surveys	Q1
24	Oviedo-Trespalacios, Oscar;	The impact of road advertising signs on driver behavior and	2020	Analytic Methods in Accident Research	Q1

	Truelove, Verity; Watson, Barry; Hinton, Jane A.	implications for road safety: A critical systematic review			
26	Tang, Jinjun; Zheng, Lanlan; Han, Chunyang; Yin, Weiqi; Zhang, Yue; Zou, Yajie; Huang, Helai.	Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review	2020	Analytic Methods in Accident Research	Q1
19	Wegman, Fred; Berg, Hans; Yngve Cameron, Iain; Thompson, Claire; Siegrist, Stefan; Weijermars, Wendy.	Evidence-based and data-driven road safety management	2015	IATSS Research	Q1
10	Norros, Ilkka; Kuusela, Pirkko; Innamaa, Satu; Pilli-Sihvola, Eetu; Rajamäki, Riikka.	The Palm distribution of traffic conditions and its application to accident risk assessment	2016	Analytic Methods in Accident Research	Q1
7	Gutierrez-Osorio, Camilo; Pedraza, César.	Modern data sources and techniques for analysis and forecast of road accidents: A review	2020	Journal of Traffic and Transportation Engineering (English Edition)	Q2
6	Kabli, Ahmed; Bhowmik, Tanmoy; Eluru, Naveen.	A multivariate approach for modeling driver injury severity by body region	2020	Analytic Methods in Accident Research	Q1
4	Mehdizadeh, Amir; Cai, Miao; Hu, Qiong; Yazdi, Mohammad Ali Alamdar; Mohabbati-Kalejahi, Nasrin; Vinel, Alexander; Rigdon, Steven E.; Davis, Karen C.; Megahed, Fadel M.	A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling	2020	Sensors (Switzerland)	
2	Sodikov, Jamshid	Road Traffic Accident Data Analysis and Visualization in R	2018	International Journal of Civil, Structural, Environmental and Infrastructure Engineering Research and Development	
2	Calvert, Simeon C; Schakel, Wouter J; van Lint, J. W.C.	A generic multi-scale framework for microscopic traffic simulation part II – Anticipation Reliance as a compensation mechanism for potential task overload	2020	Transportation Research, Series B: Methodological	Q1
1	Palazón-Bru, Antonio;	Development, and internal, and external	2020	International Journal of	Q2

	Prieto-Castelló, María José; De la Rosa, David Manuel Folgado; Macanás-Martínez, Ana; Mares-García, Emma; Carbonell-Torregrosa, María de los Ángeles; Gil-Guillén, Vicente Francisco; Cardona-Llorens, Antonio; Marhuenda-Amorós, Dolores.	validation of a scoring system to predict 30-day mortality after having a traffic accident traveling by private car or van: An analysis of 164,790 subjects and 79,664 accidents		Environmental Research and Public Health	
1	Prashant Krishnan, V; Chandra Sheel, Vivek; Viswanadh, M. V.S.; Shetty, Chetan; Seema, S.	Data Analysis of Road Traffic Accidents to Minimize the rate of Accidents	2018	Proceedings 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2018	
	Nidhi, R.; Kanchana, V.	Analysis of road accidents using Data mining techniques	2018	International Journal of Engineering and Technology(UAE)	
	Hu, Qiong; Cai, Miao; Mohabbati-Kalejahi, Nasrin; Mehdizadeh, Amir; Yazdi, Mohammad Ali Alamdar; Vinel, Alexander; Rigdon, Steven E.; Davis, Karen C.; Megahed, Fadel M.	A review of data analytic applications in road traffic safety. Part 2: Prescriptive modeling	2020	Sensors (Switzerland)	
	Kumar, Sachin; Toshniwal, Durga.	A data mining framework to analyze road accident data	2015	Journal of Big Data	Q1
	Adanu, Emmanuel Kofi; Penmetsa, Praveena; Wood, Dustin; Jones, Steven L.	Incorporating systems thinking approach in a multilevel framework for human-centered crash analysis	2019	Transportation Research Interdisciplinary Perspectives	Q2
	Feng, Mingjie; Wang, Xuesong; Quddus, Mohammed.	Developing multivariate time series models to examine the interrelations between police enforcement, traffic violations, and traffic crashes	2020	Analytic Methods in Accident Research	Q1

	Nour, Mohamed K; Naseer, Atif; Alkazemi, Basem; Jamil, Muhammad Abid.	Road Traffic Accidents Injury Data Analytics	2020	International Journal of Advanced Computer Science and Applications	Q3

2.3.3. Method and application

The SLR analysis (Table 2-2) shows that most methods applied in road accidents studies are statistical analysis and machine learning techniques.

Application of statistical analysis in studies [4], [6], [13] - [16], [22] and [23] were used to assess the influence of various road, weather, and traffic conditions on traffic accidents and machine learning techniques [13], [17] - [21], [24] and [27] to perform prediction models for road accidents.

Study [5] used a logistic regression model to determine the risk of mortality following a traffic accident.

Another study of Gutierrez-Osorio [21] used clustering algorithms, decision trees and classifiers, association rules, and natural language processing algorithms to analyze and predict traffic accidents and determine the most relevant elements that contribute to road accidents.

Tang [12] used four statistical models, Accelerated Failure Time (AFT) model, Quantile Regression (QR) model, Finite Mixture (FM) model, and Random Parameters Hazard-Based Duration (RPHD) model to examine road accidents. And four machine learning models: K- Nearest Neighbor (KNN) model, Support Vector Machine (SVM) model, Back Propagation Neural Network (BPNN) model, and Random Forest (RF) model to perform road accident prediction.

Table 2-2 - Author, title, method, and application ranked by citation

N. of citations	Author	Title	Method	Application
46	Fountas, Grigorios Anastasopoulos, Panagiotis	Analysis of accident injury-severity outcomes: The zero-inflated hierarchical ordered probit	Statistical Analysis	Using injury-severity data from single-vehicle accidents that occurred in the State of Washington, from 2011 to 2013, the implementation

		model with correlated disturbances		potential of the proposed approach is demonstrated
35	Li, Tao; Xie, Ning; Zeng, Chunqiu; Zhou, Wubai; Zheng, Li; Jiang, Yexi; Yang, Yimin; Ha, Hsin Yu; Xue, Wei; Huang, Yue; Chen, Shu; Ching Navlakha, Jainendra; Iyengar, S. S.	Data-driven techniques in disaster information management	Advanced data management and analysis techniques	Provide a systematic treatment of the recent developments in data-driven disaster management
26	Tang, Jinjun Zheng, Lanlan Han, Chunyang Yin, Weiqi Zhang, Yue Zou, Yajie Huang, Helai	Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review	Four statistical models: Accelerated Failure Time (AFT) model, Quantile Regression (QR) model, Finite Mixture (FM) model, and Random Parameters Hazard-Based Duration (RPHD) model; and four machine learning models: K- Nearest Neighbor (KNN) model, Support Vector Machine (SVM) model, Back Propagation Neural Network (BPNN) model, and Random Forest (RF) model as candidates	Examining performance in incident clearance time prediction, especially, when omitted variables present significant impacts on selected variables
24	Oviedo-Trespalacios, Oscar; Truelove, Verity; Watson, Barry; Hinton, Jane A.	The impact of road advertising signs on driver behavior and implications for road safety: A critical systematic review	Revised literature using a systematic approach informed by the Task-Capability Interface (TCI) model (a seminal theoretical framework that explains determinants of driving behavior and crash risk)	The impact of roadside advertising signs on driver behavior and road safety
19	Wegman, Fred Berg, Hans Yngve Cameron, Iain Thompson, Claire Siegrist, Stefan Weijermars, Wendy	Evidence-based and data-driven road safety management	Statistical Analysis	Evidence-based and data-driven road safety management: ex-post and ex-ante evaluation of both individual interventions and intervention packages in road safety strategies, and

				transferability (external validity)
10	Norros, Ilkka Kuusela, Pirkko Innamaa, Satu Pilli-Sihvola, Eetu Rajamäki, Riikka	The Palm distribution of traffic conditions and its application to accident risk assessment	Statistical Analysis	Assessing the influence of various road, weather and traffic conditions on traffic accidents
7	Gutierrez-Osorio, Camilo Pedraza, César	Modern data sources and techniques for analysis and forecast of road accidents: A review	Clustering algorithms, decision trees and classifiers, association rules, and natural language processing algorithms	Analyze and predict traffic accidents and determine the most relevant elements that contribute to road accidents
6	Kabli, Ahmed; Bhowmik, Tanmoy; Eluru, Naveen.	A multivariate approach for modeling driver injury severity by body region	Empirical analysis	Development of high-resolution crash severity models based on driver severity reported by a medical professional using Abbreviated Injury Scale (AIS) by body region
4	Mehdizadeh, Amir Cai, Miao Hu, Qiong Yazdi, Mohammad Ali Alamdar Mohabbati-Kalejahi, Nasrin Vinel, Alexander Rigdon, Steven E. Davis, Karen C. Megahed, Fadel M.	A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling	Prediction and optimization	Understand and quantify crash risk based on different driving conditions; minimizing crash risk through route/path selection and rest-break scheduling
2	Jamshid Sodikov, Jamshid Sodikov	Road Traffic Accident Data Analysis and Visualization in R	Statistical analysis	Road traffic accidents data analysis and visualization
2	Calvert, Simeon C. Schakel, Wouter J. van Lint, J. W.C.	A generic multi-scale framework for microscopic traffic simulation part II – Anticipation Reliance as a compensation mechanism for potential task overload	Statistical analysis	Multi-level modeling and simulation framework to describe the role of anticipation in human driving.
1	Palazón-Bru, Antonio Prieto-Castelló, María José De la Rosa, David Manuel Folgado Macanás-Martínez, Ana	Development, and internal, and external validation of a scoring system to predict 30-day mortality after having a traffic	Statistical analysis: logistic regression model	Determine the risk of mortality following a traffic accident

	Mares-García, Emma Carbonell-Torregrosa, María de los Ángeles Gil-Guillén, Vicente Francisco Cardona-Llorens, Antonio Marhuenda-Amorós, Dolores	accident traveling by private car or van: An analysis of 164,790 subjects and 79,664 accidents		
1	Prashant Krishnan, V. Chandra Sheel, Vivek Viswanadh, M. V.S. Shetty, Chetan Seema, S.	Data Analysis of Road Traffic Accidents to Minimize the rate of Accidents	Statistical analysis: k-means and decision trees	Identification of patterns and trends to measure occurrences of accidents on roads
	Nidhi, R. Kanchana, V.	Analysis of road accidents using Data mining techniques	Statistics analysis and data mining algorithms: classification model (Naïve Bayes) and clustering algorithm (K-means)	Roadway traffic data analysis to find variables that are related to fatal accidents
	Hu, Qiong Cai, Miao Mohabbati-Kalejahi, Nasrin Mehdizadeh, Amir Yazdi, Mohammad Ali Alamdar Vinel, Alexander Rigdon, Steven E. Davis, Karen C. Megahed, Fadel M.	A review of data analytic applications in road traffic safety. Part 2: Prescriptive modeling	K-shortest path algorithm with four risk indicators: logistic regression, Poisson regression, XGBoost, and neural network	Crash risk prediction
	Kumar, Sachin Toshniwal, Durga	A data mining framework to analyze road accident data	K-modes clustering	Identify the main factors associated with a road and traffic accident
	Adanu, Emmanuel Kofi Penmetsa, Praveena Wood, Dustin Jones, Steven L.	Incorporating systems thinking approach in a multilevel framework for human-centered crash analysis	Statistical analysis	Understand how various risk factors contribute to crash occurrence: road user behavior (driver characteristics)
	Feng, Mingjie; Wang, Xuesong; Quddus, Mohammed.	Developing multivariate time series models to examine the interrelations between police enforcement, traffic violations, and traffic crashes	Vector autoregressive models	Reveal the dynamic interactions and contemporaneous relation- ships between police enforcement, traffic violations, and traffic crashes

	Nour, Mohamed K; Naseer, Atif; Alkazemi, Basem; Jamil, Muhammad Abid.	Road Traffic Accidents Injury Data Analytics	Advanced data analytics methods: tree-based techniques (XGBoost), outperform regression-based (ANN)	Identify risk and key factors that contribute to accident severity

2.3.4. Keywords occurrence analysis

For this analysis, VOSviewer, a software tool for constructing and visualizing bibliometric networks [30], was used to create and visualize bibliometric networks.

The keyword co-occurrence network is created by treating each keyword as a node and each co-occurrence of a pair of words as a link between those two words. The number of times that a pair of words co-occurs constitutes the weight of the link connecting these two keywords. The network constructed in this manner represents a weighted network [31].

Figure 2-2 shows the 5 keywords with more weight in the 20 papers analyzed in SLR. All the keywords are related to the presented study because most of the papers were related to machine learning, statistical methods, or clearance time prediction, all about influence factors that lead to road incidents.

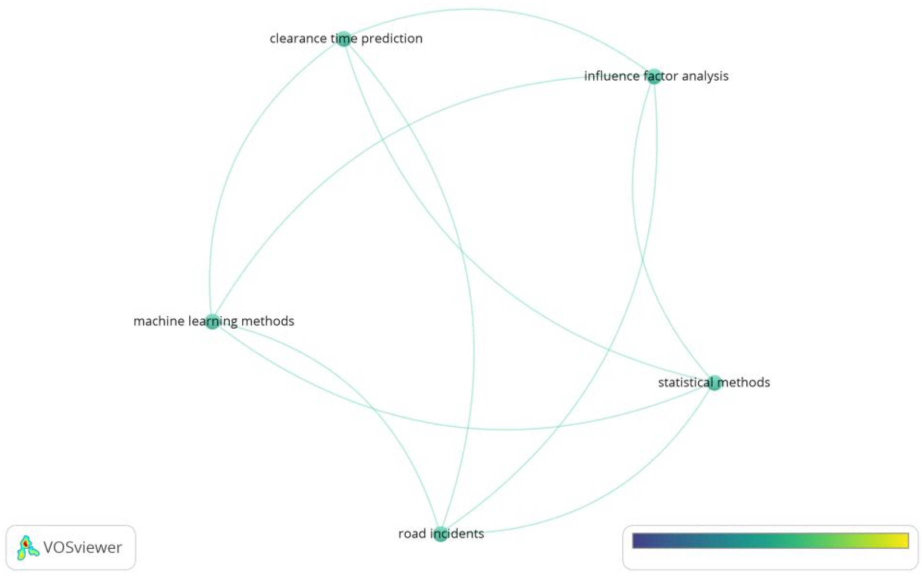

Figure 2-2 - VOSviewer keyword occurrence analysis

2.3.5. Authors and co-authorship occurrence analysis

Figure 2-3 presents authors' co-authorship and it is visible two clusters which means that authors are relatively strongly related to each other. The cluster present on the right side shows authors who wrote 3 of the 20 papers of the SLR, and those papers were related to machine learning and data-driven. The cluster at the left presents authors who wrote 2 related papers.

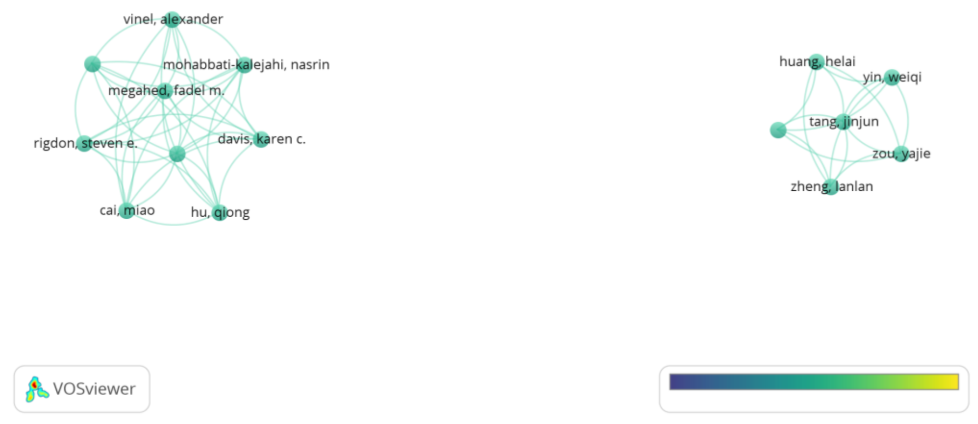

Figure 2-3 - VOSViewer authors co-authorship analysis

Chapter 3 – Road accident analytics process

3.1. Business and data understanding

This analysis addresses challenge 51 of “Identification of road accidents patterns and correlation with external factors” [32], launched by Lisboa Inteligente—LxDataLab [7] for academia. This analysis objective is to investigate and identify accident patterns in Lisbon’s metropolitan area, define when and where accidents road occurs, and how external factors such as weather and pandemics effect such phenomena.

The data-understanding phase aimed to collect, describe, explore, and verify the data’s quality. Therefore, this step was structured in three subphases: describe, explore, and verify data quality.

Road accident data sets were provided and collected from Lisboa Inteligente—LxDataLab [7] on the scope of the challenges launched for academia, as mentioned previously.

In the scope of the Lisboa Inteligente—LxDataLab [7] challenge 51, the provided data set was a single Excel file with aggregated data from ANSR, PSP, GNR, with a list of road accidents (RSB) events in Lisbon involving vehicles and motorcycles from January to December 2019. The Excel file included four sheets with the following structure: accidents (Table 3-1) with 37 columns and 2768 rows; vehicles involved and their driver (Table 3-2), with 18 columns and 4834 rows; passengers (Table 3-3), with eight columns and 631 rows; and pedestrians (Table 3-4), with seven columns and 700 rows.

Table 3-1 - Accident data schema

Characteristics	**Description**
IdAcidente	Accident ID
Datahora	Date and hour
Dia da semana	Day of the week
Sentidos	Upward and downward
Latitude GPS	Latitude

Longitude GPS	Longitude
Via Trânsito	Left, right, or central transitway
Localizações	Inside or outside localities
Freguesia	Parish
Pov. Proxima	Nearby village
Tipo natureza	Run over, collision, or screen
Natureza	Type of run over, collision, or screen
Traçado 1	Straight or curved
Traçado 2	With slope, in level or bump
Traçado 3	With or without roadside
Traçado 4	Place where the accident occurred
Estado de conservação	State of road
Características Tecnicas	Highway or other
Reg Circulação	One or both ways
Marca Via	Marks on the road
Obstáculos	Obstacles
Sinais	Signals
Sinais Luminosos	Light signals
Tipo Piso	Pavement type
Intersecção Vias	Road intersection
Factores Atmosféricos	Weather conditions
Luminosidade	Luminosity
Cond Aderência	Adhesion conditions
VM	No description
FG	No description
FL	No description
Tipos Vias	Type of road
Via	Lane

Num arruamento	Street number
Km	No description
Nome arruamento	Street name
Localização 2	GPS signal

Table 3-2 - Vehicles involved data schema

Characteristics	**Description**
IdAcidente	Accident ID
Datahora	Date and hour
Id. Veiculo	Vehicle ID
Categoria Veículos	Vehicle category
Idade	Age
Sexo	Sex
Lesões a 30 dias	Type of injury
Acessórios Condutores	Driver accessories
Acções Condutores	Driver actions
Inf. Comp. a Acções e Manobras	No description
Licença Condução	Driver license
Tempo Condução Continuada	Driving time
Teste Alcool	Alcohol test
Carga Lotação	Freight
Certificado Adr	No description
Inspecção Periódica	Periodic inspection
Seguros	Insurance

Table 3-3 - Passengers data schema

Characteristics	**Description**
IdAcidente	Accident ID
Datahora	Date and hour
Id. Veículo	Vehicle ID
Id. Passageiro	Pedestrian ID
Idade	Passenger ID
Sexo	Sex
Lesões a 30 dias	Type of injury
Acessórios Passageiro	Passenger accessories

Table 3-4 - Pedestrians data schema

Characteristics	**Description**
IdAcidente	Accident ID
Datahora	Date and hour
Id. Peao	Pedestrian ID
Idade	Age
Sexo	Sex
Lesões a 30 dias	Type of injury
Acções Peão	Pedestrian actions

3.2. Data Preparation and Fusion

This phase of the CRISP-DM methodology was subdivided into four subphases: data selection, data cleaning, feature selection, and data integration.

Data cleaning and pre-processing were performed in Python, using the Spyder platform [33] and Python libraries, such as Numpy [34], Pandas [35], Matplotlib [36], and Seaborn [37].

From the initial data set comprising 4 Excel sheets, four corresponding new datasets were generated: ‘acidentes’, referring to accidents; ‘veíc-cond’, referring to vehicles involved and their drivers; ‘passageiros’, corresponding to passengers; and ‘peões’, meaning pedestrians.

The next step, data cleaning, replaced some strings with null values since they did not add value. For instance, ‘NÃO DEFINIDO’ was replaced by ‘nan’. Likewise, some columns were deleted due to the high number of null values (acidentes: Sentidos, Pov. Proxima, Km; veíc-cond: certificado_adr; passageiros: IdAcidente, id_veiculo, id_passageiro; peões: IdAcidente, id_peao), or because they did not add value for this analysis (acidentes: IdAcidente; veíc-cond: IdAcidente, id_veiculo; passageiros: IdAcidente, id_veiculo, id_passageiro; peões: IdAcidente, id_peao), while others were renamed as they had spaces in the name and were then converted to the string type.

Additionally, due to the occurrence of a large number of ‘nan’ in the categoria_veiculo column, regarding the veíc-cond data frame, ‘nan’ values were replaced, taking into account the values in the tipo_veiculo column. The dataset was divided by type of accident: mislead (despiste), collision (colisão), and runover (atropelamento). With the division of the dataset, it is pretended to explore if there are boroughs in Lisbon with more probability of mislead, collision, run over, or even both, to occur. The Datahora column was divided in date and hour to show what day of the week and what hour of the day more accidents occurred and in what season of the year (spring, summer, autumn, or winter). From the Datahora, hora (hour) and mes (month) were generated, as well as new columns: nome_estacao (season) and momentoDia (moment of the day), to analyze the moment of the day and the moment of the year when accidents were more likely to happen. In the veíc-cond, passageiros, and peões datasets, age was aggregated to create an average column. A new dataset was created with latitude and longitude columns to georeference the road accidents by type and visualize the Lisbon boroughs with the highest number of accidents.

Still, during data preparation, outliers were found in latitude and longitude because some points appeared outside Lisbon. Due to this, a border polygon was implemented to remove outliers and get only the points located in Lisbon.

After a meeting with the engineers of the CML, two additional datasets were created. One with accidents with drivers less than 18 years (legal age in Portugal to drive a car) and another with more than 65 years drivers. The analysis of road accidents involving people with more than 65 years was suggested because they are increasing mainly due to the existence of more drivers with this age on the roads and the analysis with less than 18 was suggested due to the possible occurrence of accidents where transport was a bicycle or a scooter.

A correlation map was created using heatmap to understand what variables are more likely to influence others. The heatmap was created using the Seaborn [37] library that is built on top of Matplotlib [38]. It represented the data in a 2-dimensional form by a colored visual summary of information. The heatmap function uses the corr() to return the correlation matrix [39]. With the correlation matrix, it is possible to visualize simply the correlation between variables, to what degree, and in which direction [40]. As the heatmap uses a colorful palette, it is easier to see that stronger correlation ends of the spectrum pop out in darker, weaker correlation in lighter shades. Variables such as latitude and longitude have a strong correlation (0.64).

Also, with Folium [39], interactive leaflet maps were generated according to the latitude and longitude of each road accident to see which zones of Lisbon had more accidents in 2019.

K-means, a clustering algorithm, was performed to visualize accident clusters in the city of Lisbon. The Elbow method was used to determine the optimal number of clusters in which the data may be clustered. To this dataset, 4 clusters were used so similarity data could be aggregated and presented in a map.

3.3. Data visualization

Modeling, which involved analysis and visualization, was performed in Python with the Spyder platform. Python libraries such as Numpy [35], Pandas [36], Matplotlib [37], and Seaborn [38] were used for statistical analysis, and Folium [39] and Geopandas [40] were used for spatial analysis visualization.

In the entire universe of accidents, it is a gauge that 56.8% of the accidents were collisions, 24.3% were runovers, and 18.8% were misleads. Accidents that occurred during the day represented 66%, mainly between 3 p.m. and 6 p.m. It was also identified that 52% of accidents occurred in autumn and spring [41].

Figure 3-1, Figure 3-2 and Figure 3-3 shows the road accident distribution by type, the hour of the day, and the season of the year.

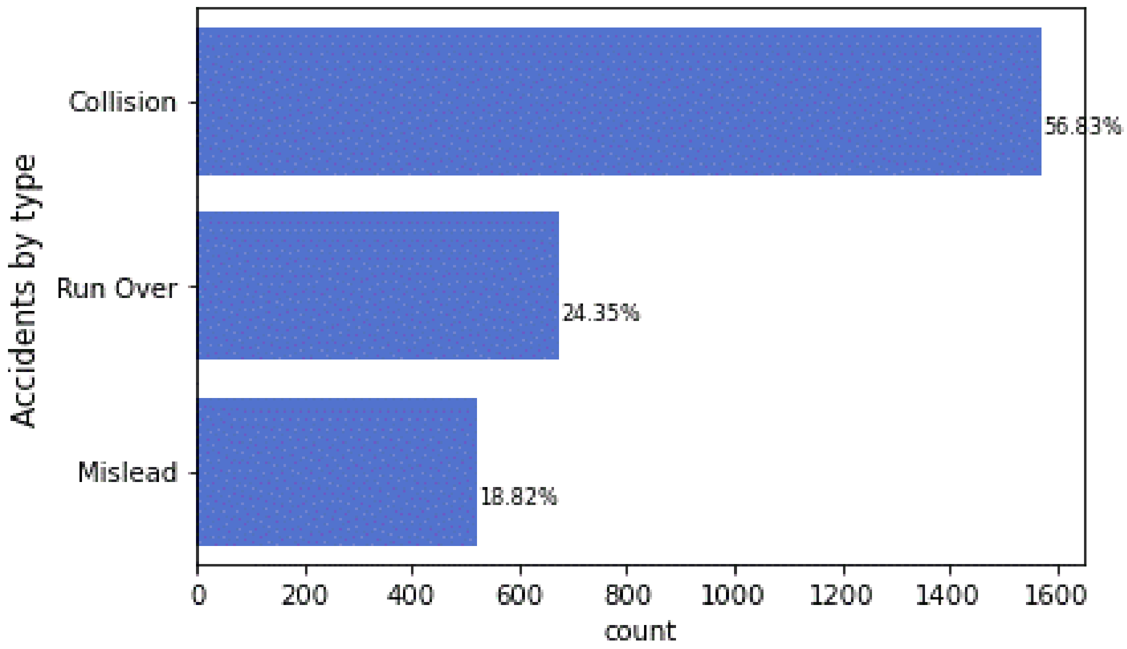

Figure 3-1 - Distribution of road accident by type - collisions, runovers, and misleads

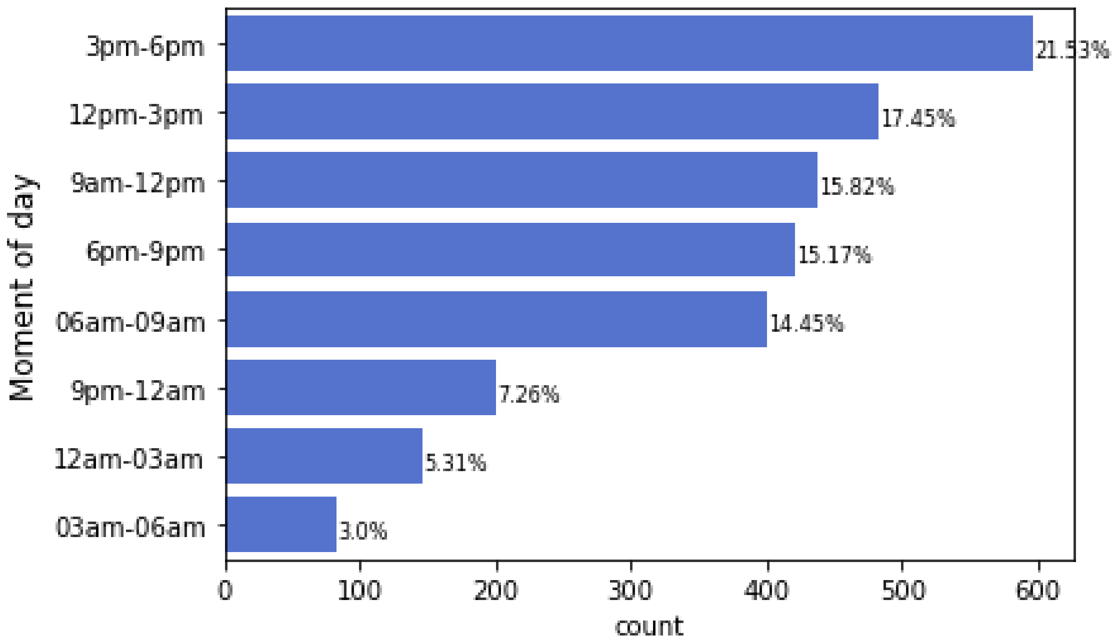

Figure 3-2 - Road accident count by the time

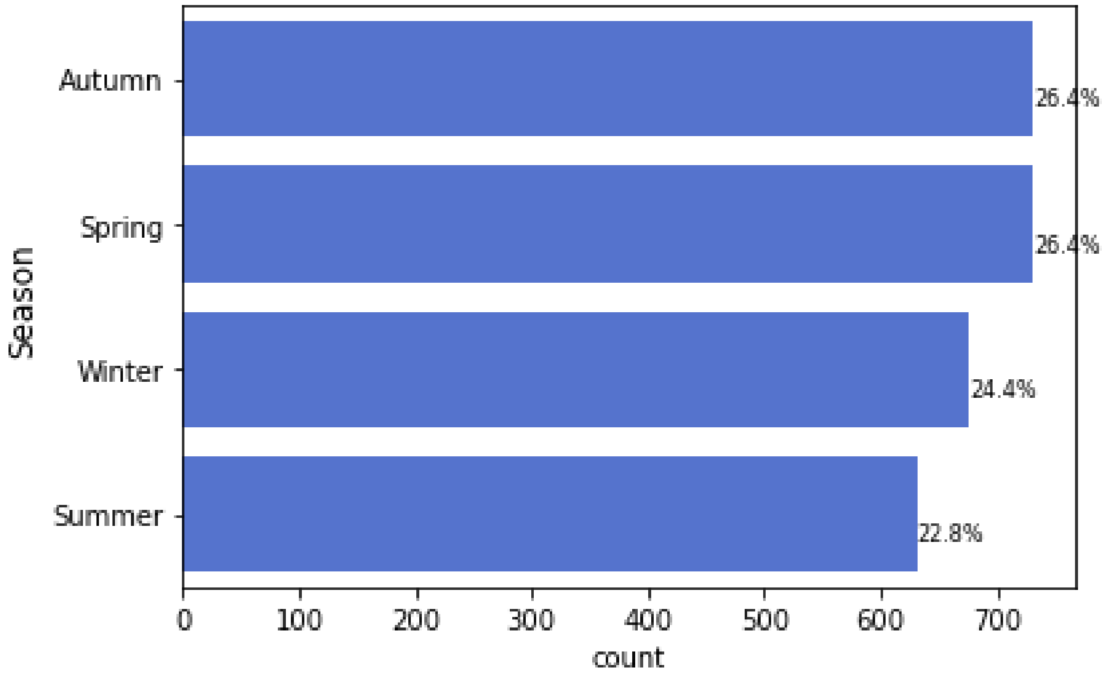

Figure 3-3 - Road accident count by season

October and November were the months where 30% of misleads happened; of these, 54% were in broad daylight, and 39% were at night but with illumination. On Thursday and Saturday, there was a higher prevalence of misleads; this was lesser on Sunday.

Runovers and collisions (Figure 3-5 and Figure 3-6) occurred more at Avenidas Novas and misleads (Figure 3-4) at São Domingos de Benfica. On the other hand, runovers were more likely to occur in Lisbon's more touristic areas, and collisions mostly in the Lisbon city center.

Data visualizations of the incidence of accidents in Lisbon are shown as heatmaps in Figure 3-4, Figure 3-5 and Figure 3-6. Areas represented in orange and yellow were the most active, representing streets where more accidents occurred; green and blue represent areas with lower occurrences of accidents.

Misleads occurred more in São Domingos de Benfica, Benfica, Carnide, and Lumiar (Figure 3-4), corresponding to entrances and exits to Lisbon's outskirts. The central city axis from Campo Grande, Avenida da República, passing by Saldanha, Marquês do Pombal, to Avenida Infante Santo had a strong incidence of mislead occurrences [41].

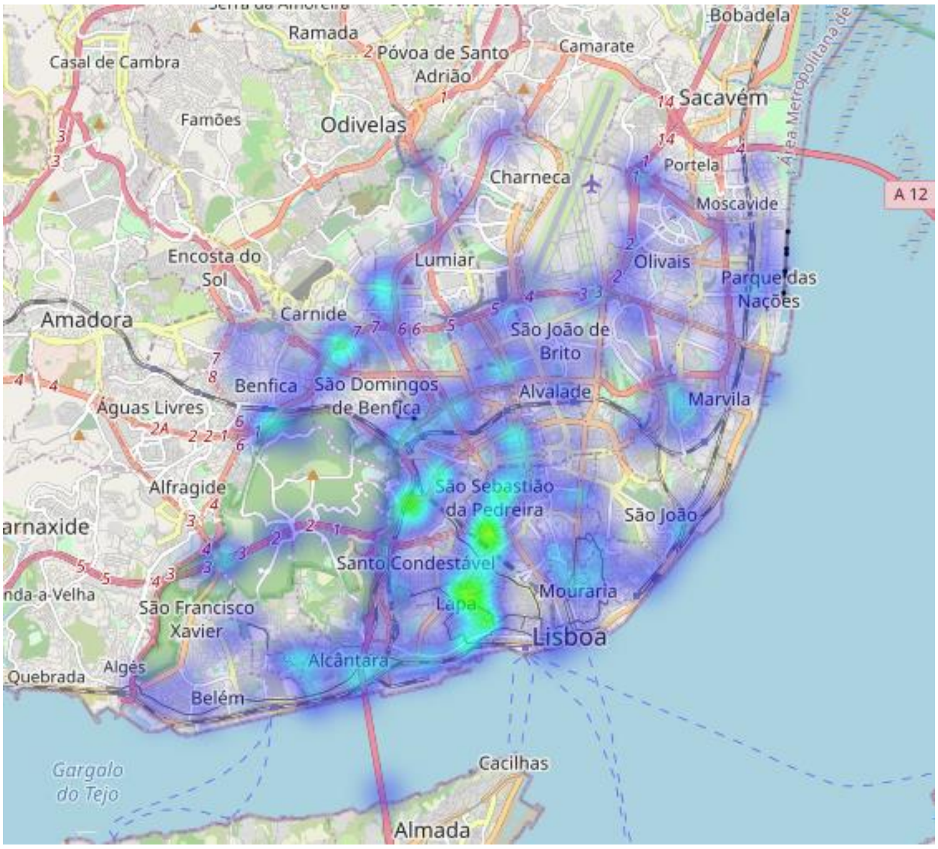

Figure 3-4 - Mislead distribution in Lisbon, during 2019. Orange and yellow represent streets where more misleads occur; green and blue represent less common places for misleads occurrence.

Approximately 95% of runovers involved pedestrians in broad daylight, and most happened between 3 p.m. and 6 p.m. on Tuesday and Friday. Runover incidents were scattered throughout the city (Figure 3-5). There was a substantial incidence in the downtown area, Terreiro do Paço, that could be associated with distracted tourists strolling. Other areas include Parque das Nações, São Domingos de Benfica, and Alcântara, corresponding to entrances to and exits from Lisbon. From Campo Grande to Avenida da República, passing by Saldanha, Marquês do Pombal, Avenida Infante Santo to Avenida 24 de Julho are the streets where most runovers took place. Moreover, the same pattern was also observed in the analysis and visualization of misleads (Figure 3-4).

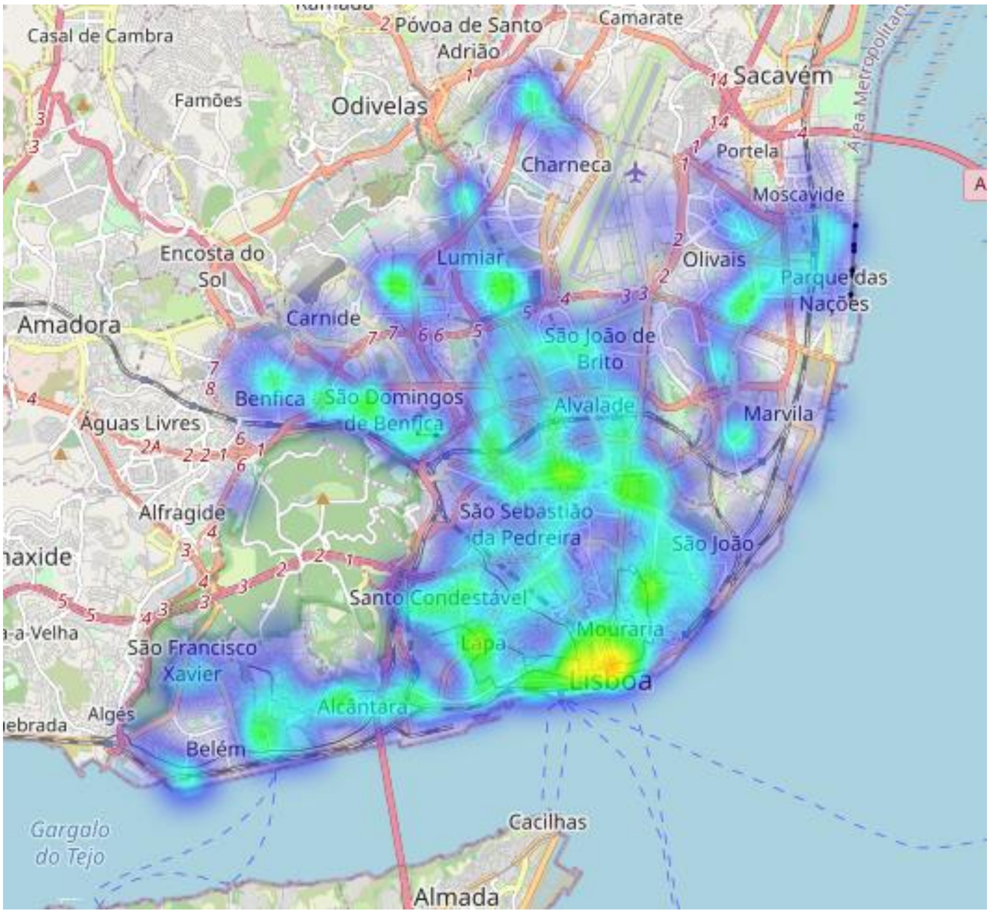

Figure 3-5 - Run-over distribution in Lisbon during 2019. Orange and yellow represent streets where more runovers occur; green and blue represent less common places for the run-over occurrence.

Finally, regarding collisions, 65% of them occurred during daylight, and 21% between 3 p.m. and 6 p.m. on Thursdays and Fridays. Side collisions with another moving vehicle represented 35% of the total, while the more expressed others were rear collisions with another moving vehicle and collisions with other situations involved.

Collision occurrence was scattered throughout the city, with focuses in Marvila, Alcântara (entrance and exit to Ponte 25 de Abril) and downtown, Terreiro do Paço, and stronger occurrences in Alvalade, Areeiro, Avenidas Novas, Campo de Ourique, and Estrela boroughs.

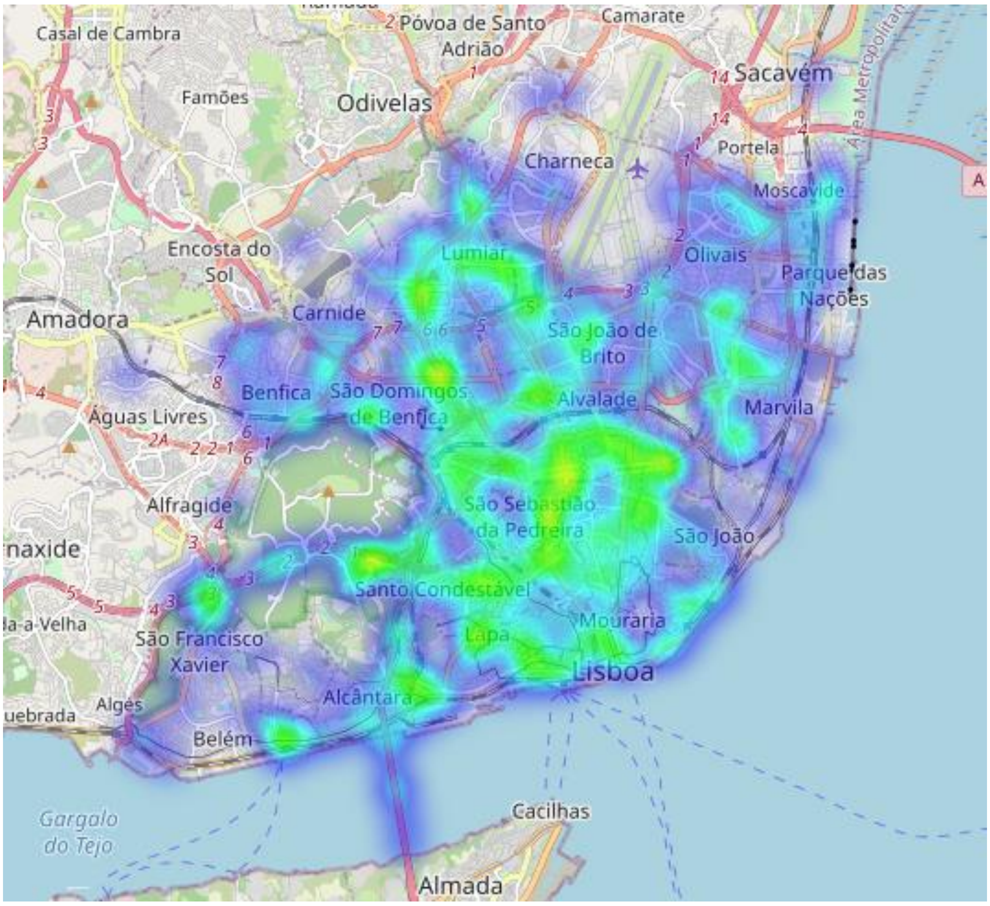

Figure 3-6 - Collision distribution in Lisbon during 2019. Yellow represents streets where more collisions occur; green and blue represent less common places for collisions occurrence.

Overall, vehicle analysis showed that 89% of accidents corresponded to a passenger vehicle with a driver 18–29 years of age (28%). In addition, 84% of accidents occurred on a dry and clean road, and 15% on a wet road. It was possible to see that of all the passengers, 64.8% were females, and 33.8% were between 18–29 years of age. Of these, 99% suffered minor injuries and were wearing a seat belt or a helmet.

Pedestrians involved in accidents were mainly females (55.9%), and 94% suffered minor injuries, of which 41% were crossing roads on a signalized zebra crossing. Pedestrians aged 18 to 29 years and 70+ were the main citizens involved.

Moreover, misleads and runovers had a high incidence at the entrances and exits of Lisbon, and along the central axis of the city from Campo Grande to Avenida Infante Santo. However, these were phenomena that occurred all over the city.

The four datasets related to accidents, drivers and vehicles, passengers and pedestrians were merged into one by the column IdAcidente. It resulted in an overall visualization of accidents and context. This was made to analyze the black spots in the city of Lisbon, to visualize the accidents with severe injuries or death, using the column lesoes_30dias. Alvalade (Figure 3-7) is the borough where more accidents occurred with victims in 2019, especially at Avenida Marechal Craveiro Lopes (Figure 3-8), where the victims were mainly men with 39 years. Figure 3-9 presents a heatmap for better visualization of road accidents by street.

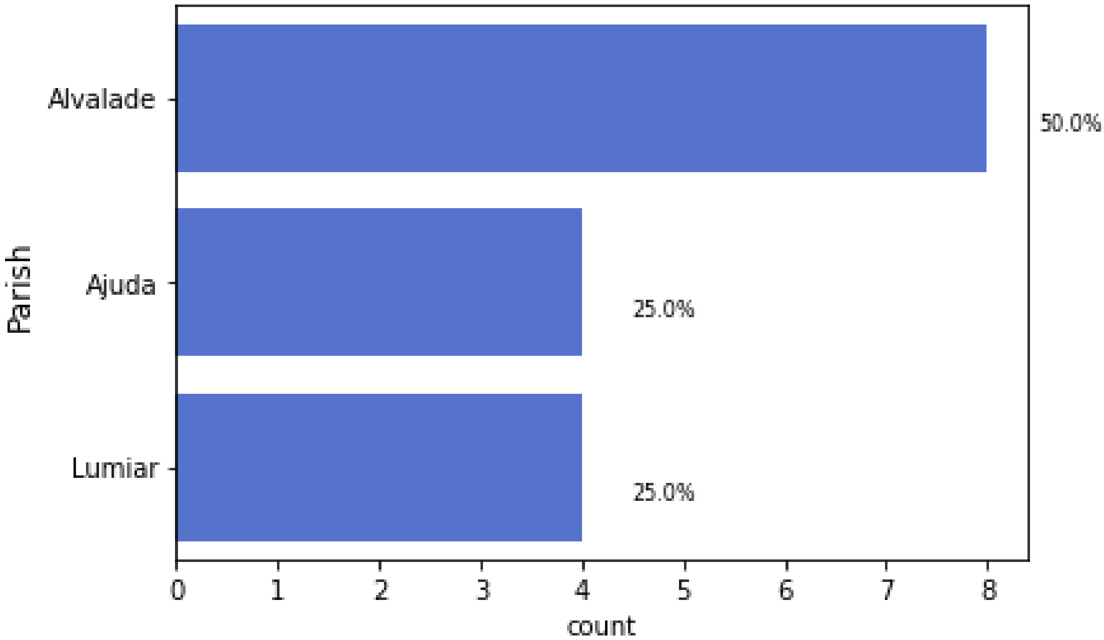

Figure 3-7 - Road accidents by parish

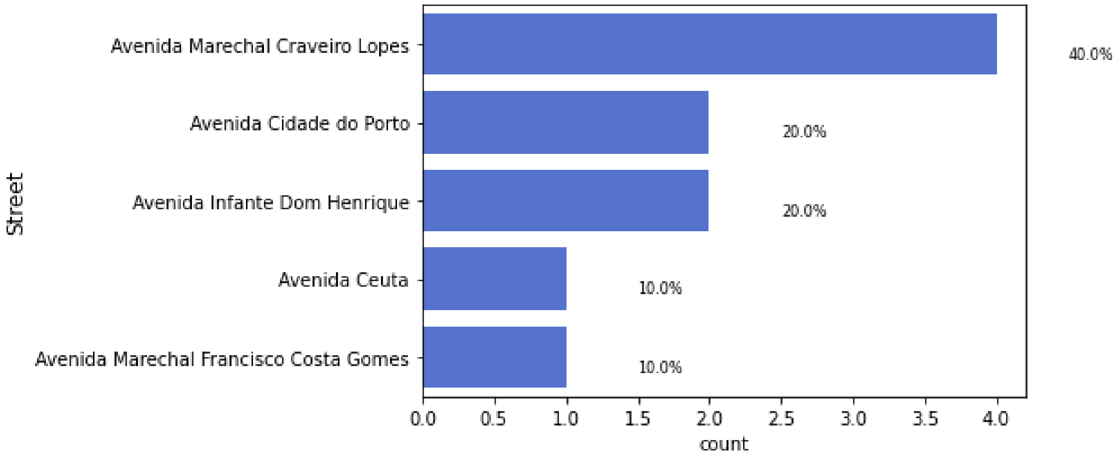

Figure 3-8 - Road accidents by street

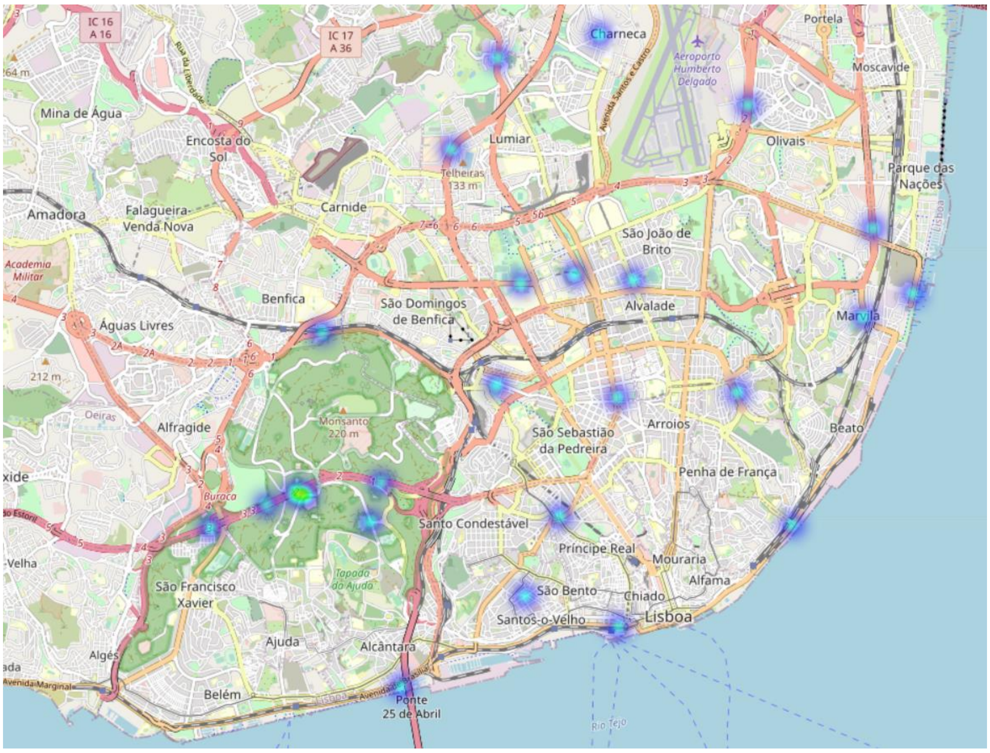

Figure 3-9 - Road accidents by street in a heatmap

The application of K-means algorithm, with 4 clusters, generated the map presented in Figure 3-10. It is possible to see 4 different colors that represent the clusters mentioned before. Table 3-5 shows the number of cluster and his color.

Table 3-5 - Cluster's color

Cluster	**Color**
1	Red
2	Purple
3	Cyan
4	Orange

The 4 clusters are noticeable: one in the north, the other at the east, the other at the west, and another in the center. Cluster 3, in the center, is denser which means that most accidents occur in this zone.

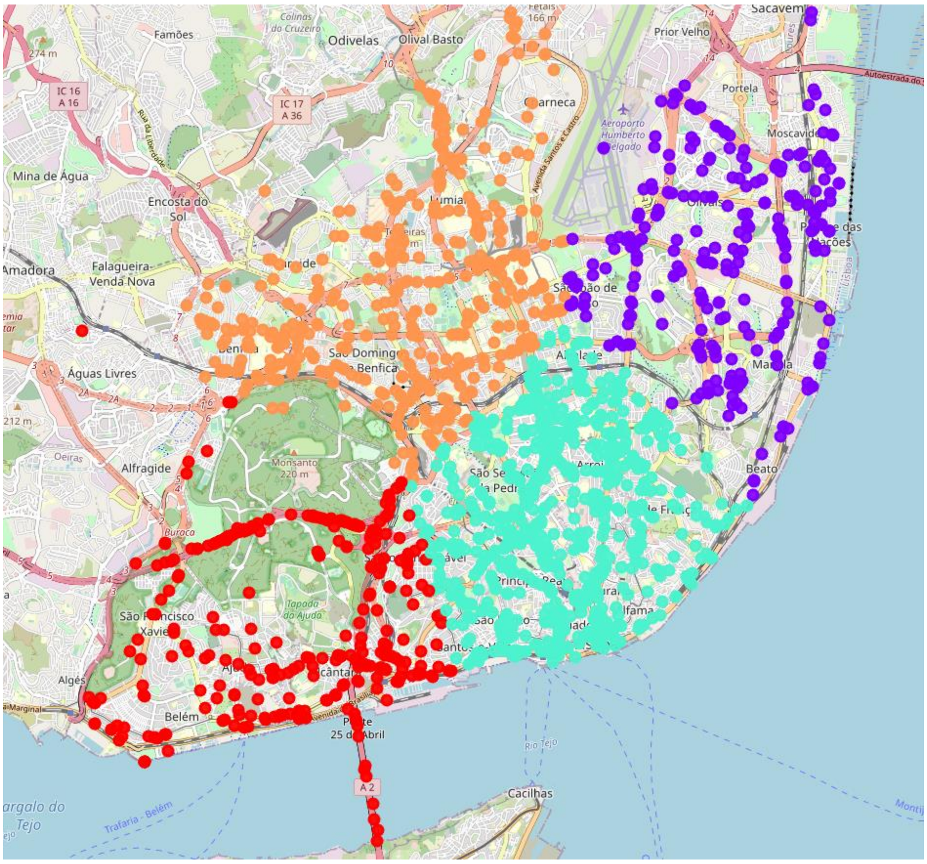

Figure 3-10 - Map of K-means distribution in Lisbon

Analyzing the data it is perceived that some accidents occurred in the footway, which is not common. Figure 3-11 shows that parishes like Santa Maria Maior and Alvalade had the strongest prevalence of accidents on the footway. Note that Santa Maria Maior is one of the most tourist parishes in Lisbon.

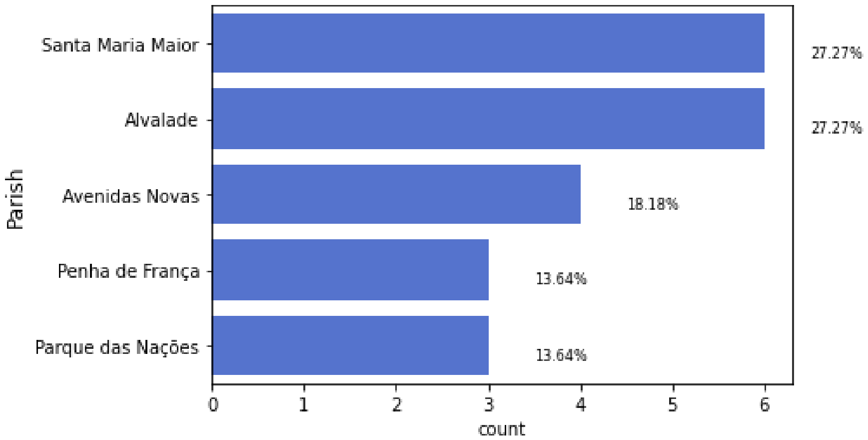

Figure 3-11 - Accidents in footway by parish

Still, about accidents in footways, Avenida da República is the street with the highest incidence as shown in Figure 3-12.

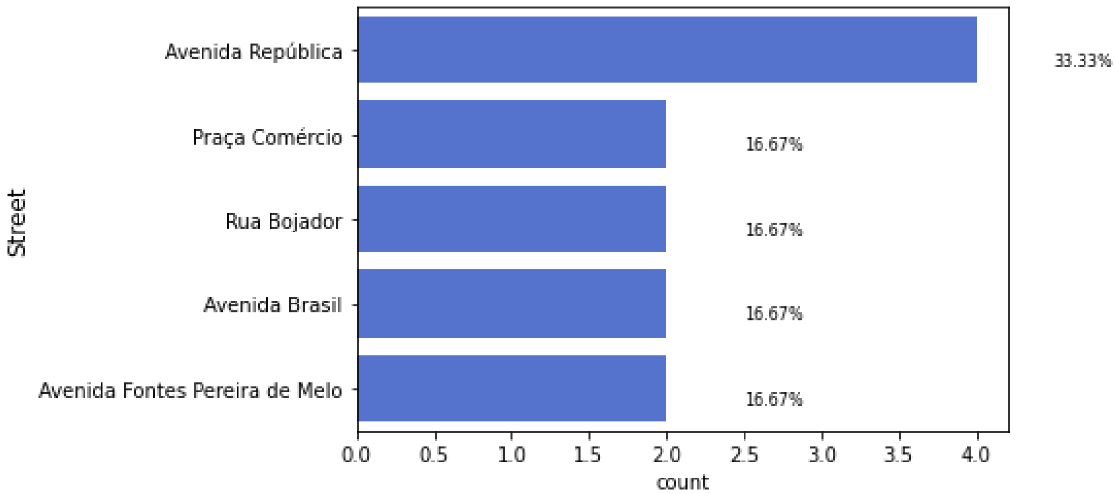

Figure 3-12 - Accidents in footway by street

Despite being subject to a more rigorous driving license revalidation after 60 years old, people older than 65 are involved in road accidents as seen in data analyzed.

As seen in Figure 3-13, the period of the day between 12 p.m. and 6 p.m. is when most accidents take place.

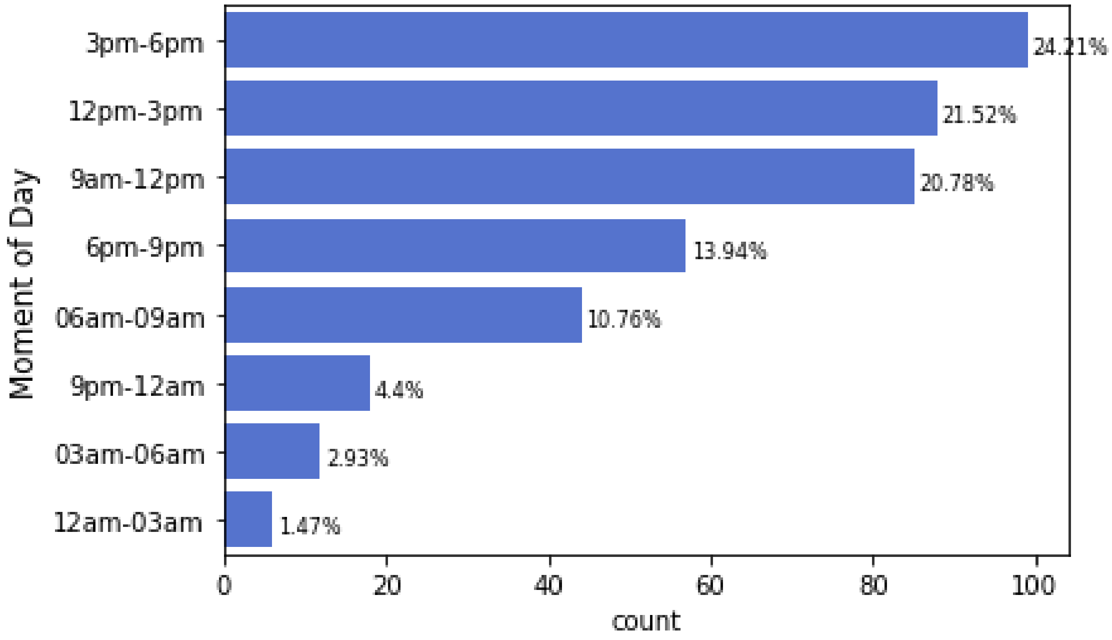

Figure 3-13 - Road accidents by the moment of the day (65+ years)

Although the age, in 75.79% of the accidents people leave unscathed (Figure 3-14), and 78.59% were in normal march (Figure 3-15), and most were not committing an infraction.

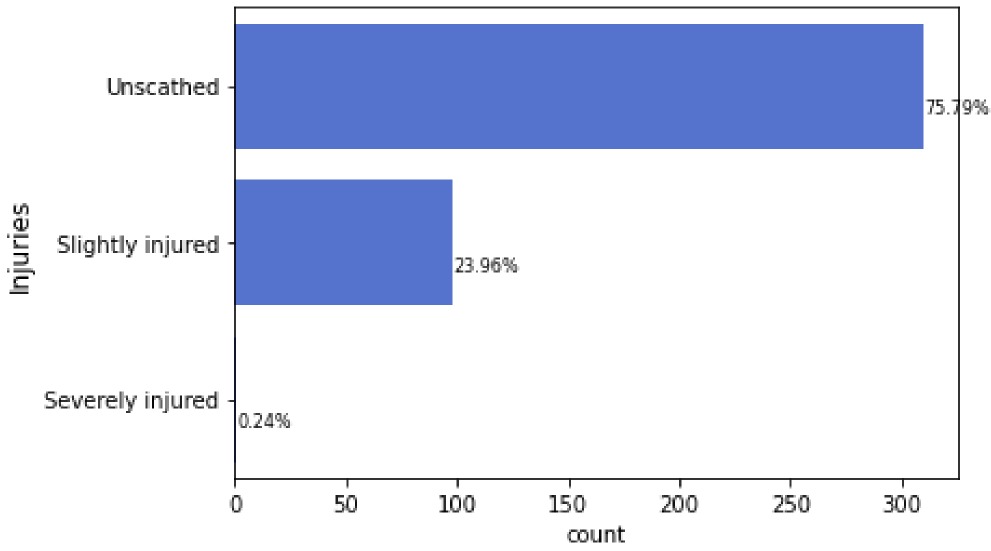

Figure 3-14 - Injuries (65+ years)

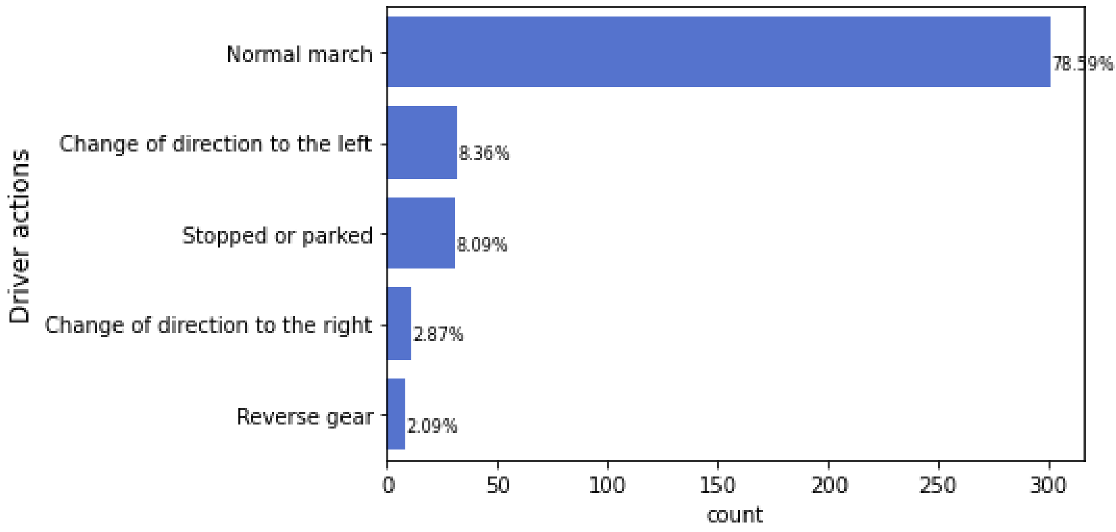

Figure 3-15 - Driver's actions (65+ years)

In Portugal, only people more than 18 years old can have a driving license to drive cars but, in the data provided, it is possible to see accidents involving drivers less than 18 years. In Figure 3-16, 2.38% of young people were committing an infraction driving a car, and the rest were legal because after 16 years old it is possible to get a driver's license to drive 2 wheeled motors. Younger less than 16 years involved in accidents were driving a bicycle or an electric scooter.

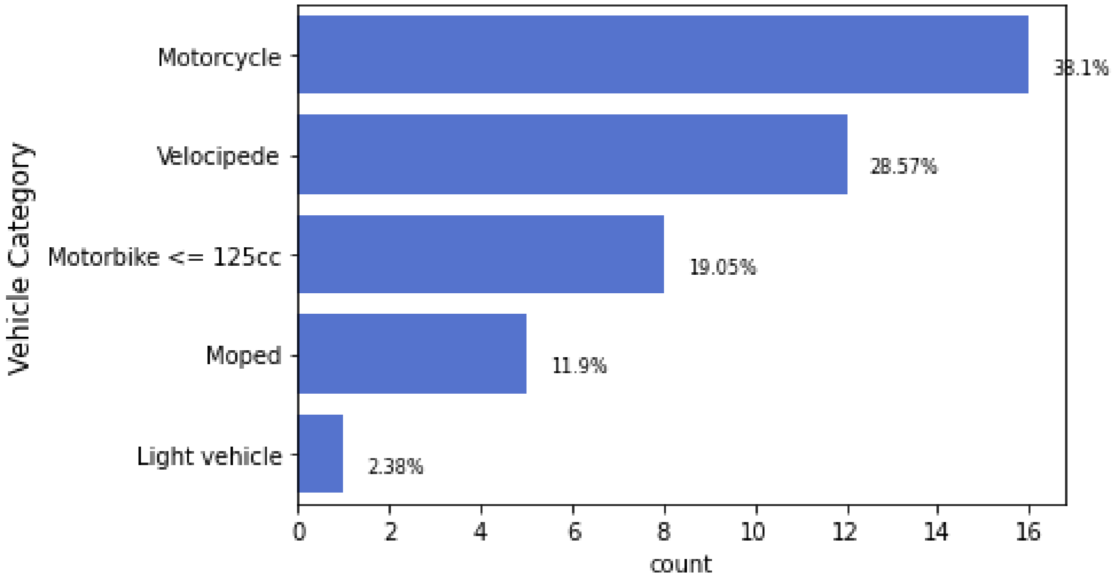

Figure 3-16 - Road accidents by vehicle category (<18 years)

Also, 87.8% of young people were driving their vehicles without taking risky actions, as seen in Figure 3-17, while others were against hand, change of lane of transit, to brake sharply.

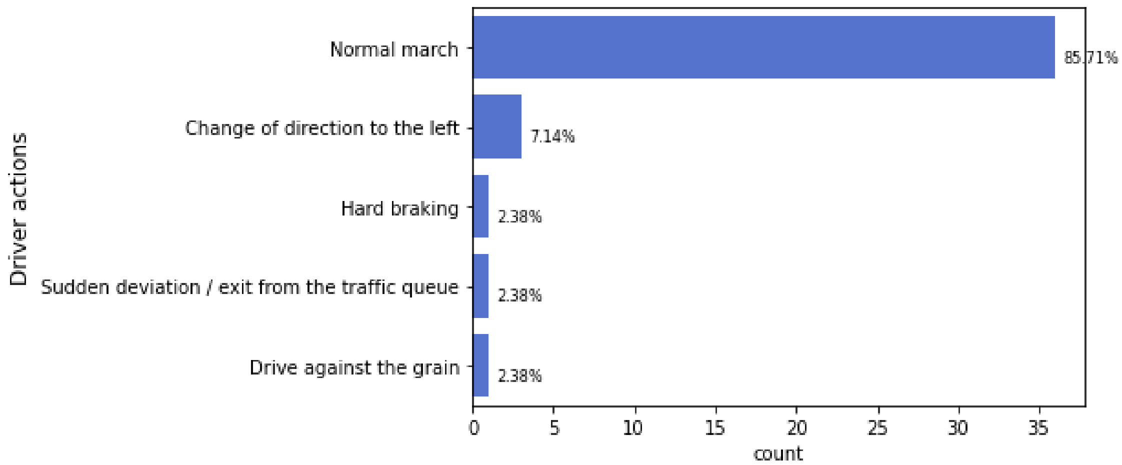

Figure 3-17 - Driver's actions (<18 years)

3.4. Decision and visualization

Results were presented at a meeting with CML, and one of the engineers gave her opinion and discussed new ideas about the subject. The concept of soft mobility was also introduced, which consists of removing automobile traffic from cities and improving the quality of life of communities by reducing the occurrence of road traffic accidents and moving to more environmentally friendly transport modes. One of the points raised in the discussion was accidents that occur on sidewalks and bicycle lanes, but there is not data yet in this regards. With the increasing of bicycles in the cities, accidents involving bicycles as well as scooters and pedestrians are likely to grow, and are of major concern for safety reasons.

Although the CML already has dashboards [42] with different city data on various parameters (air quality, noise, environmental index) as well as reports about urban mobility, the data has as well to be visualized in graphs, geolocated maps, or even tables, in order to make decisions based on data. For example, through the visualization of geographic data on a map, the places where more and fewer road accidents occur are highlighted, as well as the black spots. Having pertinent data can lead to decision-making that leads to better policies such as signaling improvement, speed limits shifting, and other measures that can help to reduce the number of road accidents.

Road accident visualization provided knowledge and insights on how citizens move in Lisbon enabling a better understanding of road accident scenarios and related events. Also providing data-driven guidelines and knowledge about road accidents in Lisbon to the city authorities and policymakers in the framework of a traffic management and visualization tool to help them mitigate such phenomena.

Chapter 4 – Discussion and conclusions

4.1. Discussion

Lisbon road accidents are an ongoing issue in Lisbon's urban mobility, carbon emissions reduction, and road safety. In a city where 370,000 vehicles enter every day, adding to the 200,000 vehicles that already circulate in the city, road accidents are key challenges that need to be tackled to improve citizens' quality of life.

This study accomplished the RQs in characterizing road accidents in Lisbon as well as characterizing road accident patterns in Lisbon and the external factors that contribute the most to this phenomenon.

To this aim and aligned with the literature review findings, most methods applied in this study on road accidents in Lisbon were statistical analysis, visualization and machine learning techniques.

The use of statistical analysis in studies [4], [6], [13] - [16], [22] and [23] were applied in road accident characterization and patterns identification, as well as on external factors as such as road type, weather, and traffic conditions of road accidents.About RQ1, road accident patterns in Lisbon strongly correlate with traffic congestion [41] and external factors but vary according to the type of occurrence – mislead, run-over and collision. This study shows that most accidents occur in bustling traffic areas, during the daytime, between 3 pm to 6 pm, corresponding to the afternoon traffic peak where traffic congestion peaks. Friday is the weekday with more road accidents prevalence, mostly when commuters travel back home and rush for weekend leave [41].Seasonality also influences road accidents, namely autumn and spring, when most accidents occur.

Hence, in RQ2 about external factors that contribute to road accidents, it was possible to conclude that external factors such as weather, location, luminosity, and time of the day are crucial to understanding the Lisbon road accident phenomenon [41]. Age also influences accidents, population between 18 and 29 years are the most involved in road accidents. Similarities with the literature [4], [5], [43] are observed, especially regarding the study's external factors.

Road accidents in Lisbon are scattered in the city by type of occurrence. Mislead and run-over have a higher incidence in Lisbon's entrances and exits and along the city's central axis from Campo Grande to Avenida Infante Santo. These occurrences correspond to traffic congestion patterns observed in primary streets and freeways [41]. São Domingos de Benfica, Carnide, Lumiar and Alcântara are zones where most mislead and run overs occur, corresponding to entrances and exits of freeways - IC 17, A5, and Eixo Norte-Sul - with high congestion levels. Collisions, on the other hand, are a phenomenon that spread all over the city.

The application of K-means algorithm for road accidents cluster analysis, identified 4 clusters, that were visualized in a georreferenced Lisbon map. With this cluster analysis of road accidents, it was possible to validate that the occurrence of accidents of all type (collisions, misleads or runover) mostly took place in the center of Lisbon which is the most touristic and business related areas.

In this analysis it was also perceived that most accidents occurred in Santa Maria Maior and Alvalade parishes had the strongest prevalence of accidents on footway. Furthermore, Santa Maria Maior is one of the most touristic parishes in Lisbon with many tourists walking around. This might influence the strongest prevalence of accidents on footway. Although, with the statistical analysis of road accidents, were found that the street with the highest footway incidence was located in another parish, Avenidas Novas, in Avenida da República, one of the busiest Avenues in the city.

A statistical analysis was performed to understand road accidents patterns according to the vehicles' drivers age group, namely, over 65 years old and under 18 years old. The main focus of this analysis was to address concerns expressed by CML engineers in the meetings along this study, especially regarding road accidents with under 18 years old.

The drivers age statistical analysis of people involved in road accident resulted in very interesting findings for this study. Although rigorous driving license revalidation policy was applied to people over 60 years old, most road accidents involve people older than 65, in normal march, not committing infraction and leaving people unscathed. Most of these road accidents occurred at noon till the end of the afternoon, between 12 p.m. and 6 p.m.

In the drivers age statistical analysis, it was also found that there was a prevalence of road accidents involving drivers under 18 years old. A very interesting factor for policy makers and authorities, as in Portugal, people under 18 years old are not allowed to have driving license for cars. Hence road accidents with people under 18 years old are considered infractions by their age. It was also found that road accidents with under 18 years old occurred as well while driving 2 wheeled motors. In this case, two wheeled and electric scooter drivers are allowed to have drivers' license as it is legal. Moreover, it was also found in this analysis that people under 16 years were involved in road accidents riding bicycles and driving electric scooters. Almost 90% of the road accidents with under 18 years old were while driving the vehicles without taking any risky actions but other road accidents occurred in against hand, change of lane of transit, to brake sharply scenarios.

Road accident data characterization, as well as pattern identification with external factors, and road accident data visualization provided a better understanding of 2019 Lisbon road accident scenarios and related events, leading to better knowledge of this phenomenon and enabling policies improvements such as signaling improvement, speed limits shifting, and other measures that can help to reduce the number of road accidents.

4.2. Conclusions

This study achieved the proposed objectives by characterizing road accidents in Lisbon as well as road accident patterns in Lisbon and the external factors that contribute the most to this phenomenon.

Lisbon road accidents are an ongoing issue in Lisbon's urban mobility and a key challenge to improve public health and citizens' quality of life.

In this study was developed a data analysis and visualization of road accidents in Lisbon in 2019, assessed on road accident data characteristics based on a multi-variable analysis on road accidents patterns and external factors understanding.

It was concluded, responding to the research questions, that Lisbon road accident patterns strongly correlate with external factors but vary by the road accident type of occurrence. Traffic congestion areas also strong correlate to where most road accidents occur, corresponding mainly to the afternoon traffic congestion peak. Friday is definitely the weekday with more road accidents in Lisbon, especially related to commuters traveling back home as well as the weekend leave rush hour [41]. Seasonality also influences road accidents, namely autumn and spring, when most accidents occur.

This study concluded as well that external factors contributed to road accidents namely, weather, location, luminosity, time of the day, day of the week, and drivers age, crucial to understanding Lisbon road accident phenomenon.

Lisbon road accidents are scattered in the city by type of occurrence. Mislead and run-over have a higher incidence in the city's entrances and exits and along the city's central axis from Campo Grande to Avenida Infante Santo. These occurrences correspond to traffic congestion patterns [41] and were visualized in georeferenced map. Collisions were spread all over the city. Furthermore, cluster analysis on Lisbon road accidents resulted in four clusters, located in the north, the east, the west, and in the center of the city. The center of the city cluster showed a denser concentration of road accidents occurrence that corresponds to the findings of the statistical analysis. This analysis was aligned with the georeferenced visualization of the road accident maps.

Lisbon road accidents analysis and visualization resulted in knowledge and insights on the more problematic Lisbon parishes and streets as well as correlation with type of road accident, vehicle category, age group, road and weather condition. This provided data-driven guidelines and knowledge on road accidents in Lisbon that help city authorities and policymakers in the improvement of the traffic management framework.

The present study on Lisbon road accidents provide results that can be cross referenced with other traffic management phenomena, such as traffic congestion [41] with key insights on how to tackle road accidents, new road traffic policies for a better public health and citizens' quality of life.

4.3. Research limitations

This study has several limitations, such as the time, only one year, which limits the diversity of the study as it is not possible to compare whether the number of accidents increased or decreased compared to other years, or the existence of several columns with null values.

After a meeting with CML, it was possible to verify that some of the coordinates associated with the accidents were not entered correctly, so some places do not correspond to the actual location of the accident. Also, it was alerted by one of the CML engineers that a specific text inserted in a column contained a spelling error.

4.4. Future work

Future work aims to understand better these results' implications, especially regarding external factors, such as traffic congestion, weather, air quality, events, crowd flow, and bike data. Moreover, pedestrian walkability, cycleway, bike station, and bike accident data are of interest to correlate with this study findings and understand the overall Lisbon urban mobility scenario.

To this aim, an integrated urban mobility dashboard with data analysis and visualization can contribute to providing city management authorities and policymakers with an overall picture of the city’s urban mobility, enabling smart solutions implementation towards a more resilient city.

This urban mobility tool would allow city management authorities and decision-makers to explore and better understand Lisbon metropolitan area commuters’ profiles through an interactive dashboard depicting georeferenced data with different features such as boroughs, geographical, demographic, economic, social, planning and environment, generating combined visualizations.

Moreover, this analytical and visualization tool would provide a complete monitoring and management resource of the entire urban ecosystem that could be replicated in other cities. It could also be integrated into the Lisbon Intelligent Management Platform - Plataforma de Gestão Inteligente de Lisboa (PGIL) [44], an existing data platform of the City Hall, further developing PGIL capacity to process and provide useful information for the operational and strategic management of the city to the various stakeholders.

A more ambitious project is the creation of an application that, with the integration of some variables and resources of artificial intelligence, can calculate the probability of an accident taking place in a certain zone.

References

[1] UN, "World Urbanization Prospects; The 2007 Revision Executive Summary," no. February, 2008, [Online]. Available: http://www.un.org/esa/population/publications/wup2007/2007WUP_ExecSum_web.pdf

[2] R. . Albino, V.; Berardi, U.; Dangelico, "Smart Cities: Definitions, Dimensions, Performance, and Initiatives," *J. Urban Technol.*, vol. 22, pp. 3–21, 2015.

[3] World Health Organization (WHO), *A Road Safety Technical Package*. 2017. [Online]. Available: http://iris.paho.org/xmlui/bitstream/handle/123456789/34980/9789275320013-por.pdf?sequence=1&isAllowed=y

[4] V. Prashant Krishnan, V. Chandra Sheel, M. V. S. Viswanadh, C. Shetty, and S. Seema, "Data Analysis of Road Traffic Accidents to Minimize the rate of Accidents," *Proc. 2018 3rd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2018*, pp. 247–253, 2018, doi: 10.1109/CSITSS.2018.8768736.

[5] A. Palazón-Bru, M. J. Prieto-Castelló, D. M. F. De la Rosa, A. Macanás-Martínez, E. Mares-García, M. Á. Carbonell-Torregrosa, V. F. Gil-Guillén, A. Cardona-Llorens, and D. Marhuenda-Amorós, "Development, and internal, and external validation of a scoring system to predict 30-day mortality after having a traffic accident traveling by private car or van: An analysis of 164,790 subjects and 79,664 accidents," *Int. J. Environ. Res. Public Health*, vol. 17, no. 24, pp. 1–13, 2020, doi: 10.3390/ijerph17249518.

[6] A. N. de S. Rodoviária, "Relatório Anual de Segurança Rodoviária," p. 154, 2019, [Online]. Available: http://www.ansr.pt/Estatisticas/RelatoriosDeSinistralidade/Documents/2019/RelatorioAnualSegurancaRodoviaria2019.pdf

[7] "LxDataLab—Laboratório de Dados Urbanos de Lisboa." https://lisboainteligente.cm-lisboa.pt/lxdatalab/ (accessed Jun. 29, 2021).

[8] R. Wirth and J. Hipp, "CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39," no. 24959, 2000,

[Online]. Available: https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining

[9] R. Chapman, Peter; Clinton, Janet; Kerber, Randy; Khabaza, Tom; Reinartz, Thomas; Russel H. Shearer, C.; Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," *Comput. Sci.*, 2000.

[10] P. Dimitrakopoulos, G.; Demestichas, "Intelligent Transportation Systems.," *IEEE Veh. Technol. Mag*, vol. 5, pp. 77–84, 2010.

[11] D. Nunes, "42 Minutos de Fila por Dia: Lisboa é a Cidade Ibérica com Mais Trânsito," *Diário de Notícias*. https://www.dn.pt/dinheiro/42-minutos-de-fila-por-dia-lisboa-e-a-cidade-iberica-com-mais-transito-10976480.html (accessed Jun. 29, 2021).

[12] J. Tang, L. Zheng, C. Han, W. Yin, Y. Zhang, Y. Zou, and H. Huang, "Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review," *Anal. Methods Accid. Res.*, vol. 27, p. 100123, 2020, doi: https://doi.org/10.1016/j.amar.2020.100123.

[13] G. Fountas and P. C. Anastasopoulos, "Analysis of accident injury-severity outcomes: The zero-inflated hierarchical ordered probit model with correlated disturbances," *Anal. Methods Accid. Res.*, vol. 20, pp. 30–45, 2018, doi: https://doi.org/10.1016/j.amar.2018.09.002.

[14] I. Norros, P. Kuusela, S. Innamaa, E. Pilli-Sihvola, and R. Rajamäki, "The Palm distribution of traffic conditions and its application to accident risk assessment," *Anal. Methods Accid. Res.*, vol. 12, pp. 48–65, 2016, doi: https://doi.org/10.1016/j.amar.2016.10.002.

[15] S. C. Calvert, W. J. Schakel, and J. W. C. van Lint, "A generic multi-scale framework for microscopic traffic simulation part II – Anticipation Reliance as compensation mechanism for potential task overload," *Transp. Res. Part B Methodol.*, vol. 140, pp. 42–63, 2020, doi: https://doi.org/10.1016/j.trb.2020.07.011.

[16] E. K. Adanu, P. Penmetsa, D. Wood, and S. L. Jones, "Incorporating systems

thinking approach in a multilevel framework for human-centered crash analysis," *Transp. Res. Interdiscip. Perspect.*, vol. 2, p. 100031, 2019, doi: https://doi.org/10.1016/j.trip.2019.100031.

[17] A. Mehdizadeh, M. Cai, Q. Hu, M. A. A. Yazdi, N. Mohabbati-Kalejahi, A. Vinel, S. E. Rigdon, K. C. Davis, and F. M. Megahed, "A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling," *Sensors (Switzerland)*, vol. 20, no. 4, pp. 1–24, 2020, doi: 10.3390/s20041107.

[18] Q. Hu, M. Cai, N. Mohabbati-Kalejahi, A. Mehdizadeh, M. A. A. Yazdi, A. Vinel, S. E. Rigdon, K. C. Davis, and F. M. Megahed, "A review of data analytic applications in road traffic safety. Part 2: Prescriptive modeling," *Sensors (Switzerland)*, vol. 20, no. 4, pp. 1–19, 2020, doi: 10.3390/s20041096.

[19] R. Nidhi and V. Kanchana, "Analysis of road accidents using Data mining techniques," *Int. J. Eng. Technol.*, vol. 7, no. 3, pp. 40–44, 2018, doi: 10.14419/ijet.v7i3.10.15626.

[20] S. Kumar and D. Toshniwal, "A data mining framework to analyze road accident data," *J. Big Data*, vol. 2, no. 1, 2015, doi: 10.1186/s40537-015-0035-y.

[21] C. Gutierrez-Osorio and C. Pedraza, "Modern data sources and techniques for analysis and forecast of road accidents: A review," *J. Traffic Transp. Eng. (English Ed.*, vol. 7, no. 4, pp. 432–446, 2020, doi: 10.1016/j.jtte.2020.05.002.

[22] J. Sodikov, "Road Traffic Accident Data Analysis and Visualization in R," *Int. J. Civil, Struct. Environ. Infrastruct. Eng. Res. Dev.*, vol. 8, no. 3, pp. 25–32, 2018, doi: 10.24247/ijcseierdjun20184.

[23] F. Wegman, H. Y. Berg, I. Cameron, C. Thompson, S. Siegrist, and W. Weijermars, "Evidence-based and data-driven road safety management," *IATSS Res.*, vol. 39, no. 1, pp. 19–25, 2015, doi: 10.1016/j.iatssr.2015.04.001.

[24] T. Li, N. Xie, C. Zeng, W. Zhou, L. Zheng, Y. Jiang, Y. Yang, H. Y. Ha, W. Xue, Y. Huang, S. C. Chen, J. Navlakha, and S. S. Iyengar, "Data-driven techniques in disaster information management," *ACM Comput. Surv.*, vol. 50, no. 1, 2017, doi: 10.1145/3017678.

[25] O. Oviedo-Trespalacios, V. Truelove, B. Watson, and J. A. Hinton, "The impact

of road advertising signs on driver behaviour and implications for road safety: A critical systematic review," *Transp. Res. Part A Policy Pract.*, vol. 122, pp. 85–98, 2019, doi: https://doi.org/10.1016/j.tra.2019.01.012.

[26] A. Kabli, T. Bhowmik, and N. Eluru, "A multivariate approach for modeling driver injury severity by body region," *Anal. Methods Accid. Res.*, vol. 28, p. 100129, 2020, doi: https://doi.org/10.1016/j.amar.2020.100129.

[27] M. K. Nour, A. Naseer, B. Alkazemi, and M. A. Jamil, "Road Traffic Accidents Injury Data Analytics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 12, pp. 762–770, 2020, doi: 10.14569/IJACSA.2020.0111287.

[28] Fabiano Filho, "PRISMA como metodologia para Revisão Sistemática." https://medium.com/dados-e-saude/prisma-como-metodologia-para-revisão-sistemática-b3f55b4ebc5c (accessed Apr. 25, 2021).

[29] "PRISMA Flow Diagram." http://www.prisma-statement.org/PRISMAStatement/FlowDiagram (accessed Apr. 25, 2021).

[30] "VOSviewer." https://www.vosviewer.com/ (accessed Aug. 29, 2021).

[31] S. Radhakrishnan, S. Erbis, J. A. Isaacs, and S. Kamarthi, "Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature," *PLoS One*, vol. 12, no. 9, pp. 1–16, 2017, doi: 10.1371/journal.pone.0185771.

[32] "Identificação de Pontos de Incidencia dos Acidentes Rodoviarios e da sua Correlação com Outros Fatores." https://lisboainteligente.cm-lisboa.pt/lxdatalab/desafios/identificacao-de-pontos-de-incidencia-dos-acidentes-rodoviariose-da-sua-correlacao-com-outros-fatores/ (accessed Jun. 29, 2021).

[33] "Spyder IDE." https://www.spyder-ide.org/ (accessed Jun. 29, 2021).

[34] "Nbsp; NumPy." https://numpy.org/ (accessed Jun. 29, 2021).

[35] "Nbsp; Pandas." https://pandas.pydata.org/ (accessed Jun. 29, 2021).

[36] "MatPlotLib." https://matplotlib.org/ (accessed Jun. 29, 2021).

[37] "Seaborn." https://seaborn.pydata.org/ (accessed Jun. 29, 2021).

[38] Mokhtar Ebrahim, “Seaborn heatmap tutorial (Python Data Visualization).” https://likegeeks.com/seaborn-heatmap-tutorial/ (accessed Jun. 25, 2021).

[39] “Mapa de Calor de Correlação.” https://craveindustry.com/mapa-calor-correlacao/ (accessed Jun. 25, 2021).

[40] Bibor Szabo, “How to Create a Seaborn Correlation Heatmap in Python?” https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e (accessed Jun. 25, 2021).

[41] V. Albuquerque, A. Oliveira, J. L. Barbosa, R. S. Rodrigues, F. Andrade, M. S. Dias, and J. C. Ferreira, “Smart cities: Data-driven solutions to understand disruptive problems in transportation—the lisbon case study,” *Energies*, vol. 14, no. 11, pp. 1–25, 2021, doi: 10.3390/en14113044.

[42] Câmara Municipal de Lisboa, “Monitorização de Parâmetros Ambientais,” 2021. https://dados.gov.pt/pt/reuses/monitorizacao-de-parametros-ambientais/ (accessed Sep. 12, 2021).

[43] J. Tang, L. Zheng, C. Han, W. Yin, Y. Zhang, Y. Zou, and H. Huang, “Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review,” *Anal. Methods Accid. Res.*, vol. 27, p. 100123, 2020, doi: 10.1016/j.amar.2020.100123.

[44] Câmara Municipal de Lisboa, “Plataforma de Gestão Inteligente de Lisboa.” https://lisboainteligente.cm-lisboa.pt/lxi-iniciativas/plataforma-de-gestao-inteligente-de-lisboa/ (accessed Jun. 29, 2021).