iscte

INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Machine Learning Insights: Analyzing Factors Influencing Happiness Score

José Nascimento da Silva

M.Sc. in Computer Science and Business Management

Supervisor:
PhD Sancho Moura Oliveira, Associate Professor,
ISCTE-IUL

June 2024

"Happiness can change, and does change, according to the quality of the society in which people live."

John F. Helliwell is one of the chief editors of the World Happiness Report.

Direitos de cópia ou Copyright
©Copyright: José Nascimento da Silva.

O Iscte - Instituto Universitário de Lisboa tem o direito, perpétuo e sem limites geográficos, de arquivar e publicitar este trabalho através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, de o divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

Agradecimentos

Em primeiro lugar, gostaria de expressar a minha mais sincera gratidão, apoio, confiança, dedicação e disponibilidade aos meus orientadores do Prof. Dr. Sancho Moura Oliveira.

De seguida, deixo um agradecimento especial à minha família, em especial aos meus pais, irmã e à Maria, pelo exemplo de resiliência, entrega e por serem o meu pilar.

Por último, agradeço aos meus companheiros desta viagem, os meus amigos. Obrigada pela motivação, pela partilha e pela companhia em todos os momentos.

Acknowledgements

First and foremost, I would like to express my sincerest gratitude, support, trust, dedication, and availability to my supervisors of Prof. Dr. Sancho Moura Oliveira.

Next, I would like to thank my family, especially my parents, sister, and Maria, for their example of resilience, dedication and for being my pillar.

Finally, I would like to thank to my partners in this adventure, my friends. Thank you for the motivation and support.

Resumo

Este caso de estudo visa examinar o conjunto de dados do Relatório Mundial da Felicidade, focando na identificação de fatores-chave que influenciam significativamente as pontuações da felicidade na vida. A felicidade serve como um objetivo vital tanto para governos quanto para indivíduos e atua como um indicador confiável do desenvolvimento social. Utilizando técnicas de machine learning, especificamente modelos de regressão e classificação, este estudo classifica e seleciona características essenciais. Os resultados, derivados de uma análise de dados abrangente, destacam que o PIB per capita como o principal determinante da felicidade na vida, seguido pela expectativa de vida. Os resultados do estudo são substanciados através de várias métricas de desempenho, assegurando a validade dos dados obtidos.

Palavras-Chave: “Machine Learning” e “Pontuação de Felicidade” e “Visualização de Dados”

Abstract

This case study aims to examine the World Happiness Report dataset, focusing on identifying key factors that significantly influence life happiness scores. It posits that happiness serves as a vital goal for both governments and individuals and acts as a reliable indicator of societal development. Utilizing supervised machine learning techniques, specifically regression and classifications models, this study classifies and selects essential features. The findings, derived from comprehensive data analysis, highlight GDP per capita as the foremost determinant of life happiness, followed by health life expectancy. The study's outcomes are substantiated through various performance metrics, ensuring the validity of the obtained data.

keywords: “Machine Learning” AND “Happiness Score” AND “Data Visualization”

General Index

Table Index

Figure Index

Equation Index

Glossary of Abbreviations and Acronyms

WHR - World Happiness Report

HS - Happiness Score

ML - Machine Learning

KPI - Key Performance Indicator

RQs - Research Questions

1. Chapter 1 – Introduction

This chapter presents the background of the problem area, the purpose of the thesis, research questions, limitations, and system requirements.

1.1. Background

In recent decades, the pursuit of understanding human happiness has transcended beyond philosophical discourse into the realm of data-driven science. The emergence of machine learning (ML) as a potent tool in data analysis has significantly contributed to this transformation. Machine learning, a subset of artificial intelligence, leverages algorithms and statistical models to enable computers to perform tasks without explicit instructions, relying on patterns and inference instead [1]. This technological advancement has opened new avenues in psychological research, particularly in studying and quantifying the abstract concept of happiness.

Happiness, often referred to as subjective well-being, is a complex, multi-dimensional phenomenon that encompasses various aspects of human life. It is a subject of interest not only to psychologists but also to sociologists, economists, and policymakers. The quest to understand what makes individuals and societies happy has led to the development of numerous models and indices, such as the World Happiness Report, which attempt to quantify and rank happiness across different populations and cultures [2]. However, the subjective nature of happiness, interwoven with a myriad of socio-economic, psychological, and environmental factors, makes it a challenging concept to measure and analyze. This is where machine learning comes into play, offering sophisticated tools to dissect, interpret, and predict elements influencing happiness.

The applications of machine learning in understanding happiness are diverse and extensive. In educational settings, ML algorithms are being utilized to monitor and enhance the emotional well-being of students, recognizing the importance of a supportive learning environment for academic and personal growth [3]. In the corporate world, machine learning techniques are employed to analyze employee satisfaction and retention, acknowledging the link between employee well-being and organizational productivity[4]. The advent of emotion AI and real-time emotion detection systems further exemplifies the potential of machine learning in capturing and responding to human emotions in various contexts.

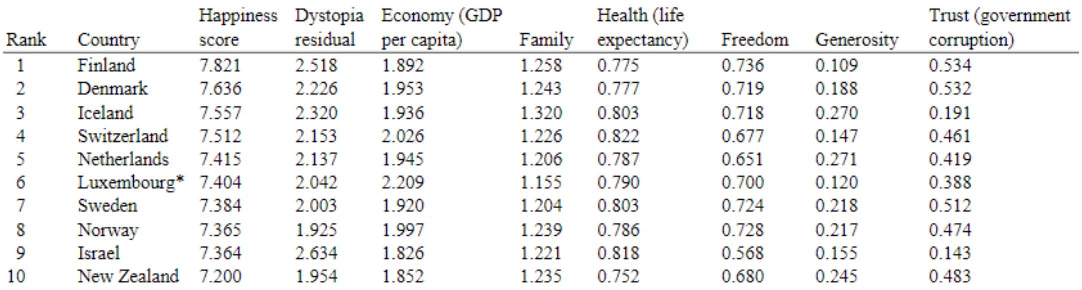
In this era where understanding human happiness transcends philosophical debate to embrace data-driven methodologies, the introduction of the World Happiness Report on April 1, 2012, has been a pivotal development. This report, updated annually, is rooted in the insights derived from the Cantril ladder survey, offering an invaluable global perspective on well-being [5]. This research scrutinizes the data presented in these reports from 2015 to 2022, covering critical parameters such as GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption to formulate a comprehensive happiness index for each nation.

By augmenting the existing dataset with a thorough exploration of selected variables from multiple perspectives through exploratory data analysis and leveraging machine learning techniques, this study significantly deepens our understanding of the factors that contribute to national happiness levels, aiming to shed light on the complex interplay of elements that influence global well-being.

1.2. Research Purpose

Much research has been done on happiness score, focusing on happiness to understand individual(s) as well as nation's life satisfaction and happiness rating. Following a review of various studies, the research gap is to determine the degree of importance of variables that lead to happiness and have a high impact on it, as which machine learning approaches can derive the highest accuracy in happiness prediction[6]. The Kaggle World Happiness Report dataset is a landmark survey of the state of global happiness, ranking 155 countries in terms of happiness levels. Happiness is rated on a scale of zero to ten based on survey respondent's perceptions of a ladder with the best possible life for them. As critical and supporting features, the dataset provides several insights from variables affecting global happiness scores. The data set features rely on getting responses from respondents based on their day-to-day life experiences, considering the most persuasive life and the most extremely bad life being.

Table 1: Top 10 countries' specific data (world happiness report 2021)



Rank	Country	Happiness score	Dystopia residual	Economy (GDP per capita)	Family	Health (life expectancy)	Freedom	Generosity	Trust (government corruption)
1	Finland	7.821	2.518	1.892	1.258	0.775	0.736	0.109	0.534
2	Denmark	7.636	2.226	1.953	1.243	0.777	0.719	0.188	0.532
3	Iceland	7.557	2.320	1.936	1.320	0.803	0.718	0.270	0.191
4	Switzerland	7.512	2.153	2.026	1.226	0.822	0.677	0.147	0.461
5	Netherlands	7.415	2.137	1.945	1.206	0.787	0.651	0.271	0.419
6	Luxembourg*	7.404	2.042	2.209	1.155	0.790	0.700	0.120	0.388
7	Sweden	7.384	2.003	1.920	1.204	0.803	0.724	0.218	0.512
8	Norway	7.365	1.925	1.997	1.239	0.786	0.728	0.217	0.474
9	Israel	7.364	2.634	1.826	1.221	0.818	0.568	0.155	0.143
10	New Zealand	7.200	1.954	1.852	1.235	0.752	0.680	0.245	0.483

Equation 1: Happiness Score

$$\text{HS} = \text{Dystopia Residual} + \text{Economy (GDP per Capita)} + \text{Family} + \text{Health (Life Expectancy)} + \text{Freedom} + \text{Generosity} + \text{Trust (Government Corruption)}$$

The factors outlined in the equation (Economy, Family, Health, Freedom, Trust, Generosity, and Dystopia Residual) are fundamental variables utilized in happiness reports such as the World Happiness Report [6]. These elements collectively aid in the assessment and measurement of a nation's overall happiness.

This case study targets the significant interest in measuring and predicting happiness in its many dimensions, emphasizing the importance of analyzing these metrics. The analysis is crucial for policymakers and global organizations, providing insights into the principal factors that influence happiness scores. Understanding these factors enriches academic research and is vital for identifying the nuances and variations in happiness scores across various contexts.

By applying machine learning techniques, this research seeks to uncover and interpret the critical determinants of happiness, offering valuable perspectives for those looking to understand and improve well-being on a global scale.

This approach not only advances the academic discourse on happiness but also equips decision-makers with the necessary insights to formulate policies that enhance overall happiness.

1.3. Research Questions

The objective of this thesis is to leverage machine learning models for the categorization of data from the World Happiness Report. By applying these models, insights capable of answering specific research questions through the analytical capabilities of machine learning are aimed to be generated:

RQ 1 - Can the region be accurately predicted based on the features extracted from the World Happiness Report data?

RQ 2 - To what extent can the Happiness Score be forecasted utilizing the features from the World Happiness Report, and how influential is the Economy in affecting this score?

RQ 3 - Is it feasible to predict an indicator (score) by leveraging other variables present within the dataset?

RQ 4 - How effectively can corruption/Trust levels between countries be predicted, considering the interrelation with other indicators within the dataset?

1.4. Limitations

L1: The dataset excludes 2020-2022, possibly differing in format due to altered data collection or reporting methods.

L2: The machine learning models will only be trained using the English language.

1.5. Document Outline

This research report consists five chapters in total, except for Chapter 1, the following sections can be identified:

- Chapter 2 – Literature Review: provides a review of research related to and important in terms of the research question and experiment. It will describe the concept of subjective happiness, including its definition and affecting factors identified to data.

- Chapter 3 – Implementation and Results: presents detailed analysis and evaluation of experimental results. It critically examines the strengths and weaknesses of the experiment.

- Chapter 4 – Conclusions and Recommendations: this chapter summarizes the research, encompassing the problem definition and the evaluation of results. It also discusses study limitations, potential improvements, and avenues for future work.

2. Chapter 2 – Literature Review and Related Work

2.1. Planning Review

Using a systematic literature review (SLR), which is the recommended ‘‘Evidence-based Software Engineering” (EBSE) method for aggregating evidence, it was possible to select the articles to answer the questions referred to previously.

SLR methodology used is based on Kitchenham, 2007: Planning, Conducting, and Reporting. The resume of the different steps is presented on Table 2.

Table 2: Systematic Literature Review Methodology

Outlining systematic literature review		Conducting Systematic Literature Review		Reporting the review
• Identification of the need for a review. • How to measure happiness.		• Applying filters and get final articles. ○ 22 articles		• Report the findings: ○ Machine Learning. ○ Factors that influence happiness. ○ ML models and analysis. ○ Challenges and future directions.
• The objective of the review. ○ Which are the features that influence happiness?	➡	• Perform data extraction and analysis of the sample. ○ Information about predict quality of life in different ways and from various perspectives. ○ Identification of the factors that affect happiness.	➡	
• Review protocol. ○ The search string, filters, repositories, and inclusion and exclusion criteria.				

2.2. Identification of the Need for a Review

We are facing many challenges as the economy grows, population growth and sustainability issues are arising, happiness is a desirable goal for governments and individuals both, and it may be used to assess social development success [7]. The purpose of this study is to categorize the most important variables influencing the life happiness score using the World Happiness Report dataset.

2.3. The Objective of the Review

The primary objective of this literature review is to critically examine and synthesize existing research on the application of machine learning techniques in analyzing factors influencing happiness scores. This review aims to provide a comprehensive understanding of how various machine learning methods have been utilized to interpret and predict aspects of human happiness.

2.4. Review Protocol

First, I identified the main keywords and databases. Second, I defined the filters to be applied to obtain the final number of articles to review.

Table 3: List of Keywords

ID	Keywords
1	“Machine Learning” AND “Happiness Score” AND “Data Visualization”

To obtain the articles I have selected 3 databases as displayed in the following table.

Table 4: List of databases

ID	Databases
1	Scopus
2	IEEExplore
3	Science Direct

To select recognizable journals and scientific articles I have applied 8 different filters.

Table 5: Quantitative filters

Inclusion	Exclusion
Full Text	Not in Full Text
Abstract	Not in Abstract
Title	Not in Title
Articles in English	Not in English
Articles and Conference Proceedings	Not in Articles and Conference Proceedings
Date (from 2015 to 2022)	2023 and 2024
Computer Science and Social Sciences	Not Computer Science and Social Sciences
No Duplicates	Duplicates

2.5. Applying Filters and Get Articles

After applying the filters, I have obtained 22 different articles as stated in the following table.

Table 6: Articles

Database	**Keywords**		**Filters**								
			Quantitative								
			Inclusion/Exclusion								
			1	2	3	4	5	6	7	8	9
	ID	**Word**	**Full Text**	**Abstract**	**Title**	**English**	**Articles/ Conference**	**Date** (2015-2022)	**Area** (Computer Science and Social Sciences)	**Duplicates**	**Manual Filter**
Scopus	1	“Machine Learning” AND “Happiness” Score" AND “Data Visualization"	54	42	34	34	28	14	6	1	22
IEEE Xplore	1	“Machine Learning” AND “Happiness” Score" AND “Data Visualization"	113	84	64	64	54	13	10	3	
Science Direct	1	“Machine Learning” AND “Happiness” Score" AND “Data Visualization"	717	547	86	84	80	76	8	2	
TOTAL									**34**	**28**	**22**

2.6. Perform Data Extraction

Statistical analysis was made to better understand the quality of the articles. Most articles obtained were related to journals, and it was possible to obtain their qualification through the H-Index, from which it was possible to obtain an average value of 70 points.

Discussion

Based on the findings, this section includes overview, the results and discussions based on the happiness analysis.

2.7. Happiness Overview

Aristotle's teachings highlighted the universal pursuit of happiness among humans, which embodies a state of well-being, joy, satisfaction, or contentment [8]. This sense of happiness is often associated with feelings of success, safety, or fortune [9]. The global community is keen on ensuring future generations can experience sustained happiness. The World Happiness Report ranks countries based on their happiness levels, using various indicators to measure the average citizen's sense of well-being. This report reveals how happiness evolves over time, influenced by technology, environment, culture, conflicts, and government policies. Many nations, including New Zealand, which allocates nearly NZ$2billion to health services, aim to create happier societies. Finland, Norway, and Denmark were recognized as the top three happiest countries in 2018, aligning with the United Nations' Sustainable Development Goals to foster a sustainable and joyful world [10].

Some research suggests happiness stems from an individual's satisfaction with life and their ability to find meaning in their daily activities and interactions [7]. It is proposed that happiness is less about external events and more about personal interpretation and internal control [11]. People who manage their inner experiences well tend to lead fulfilling lives, characterized by continuous learning, strong connections, and a positive engagement with their environment [12]. Various scales, such as the Satisfaction with Life Scale, the Positive and Negative Affect Schedule (PANAS), and the Subjective Happiness Scale, have been developed to quantitatively measure happiness. The Oxford Happiness Questionnaire, an advancement of the Oxford Happiness Inventory, uses a six-point Likert scale to assess well-being.

Studies are increasingly focusing on predicting happiness through behavioral adaptations and social network dynamics. The role of communication channels in influencing happiness, particularly through electronic commerce and social media, highlights the significance of technology in our daily lives and its impact on well-being. This exploration of happiness extends to psychological studies [13], emphasizing its importance as a consistent aspect of human experience that influences perceptions and reactions to life events [14].

The dataset from the World Happiness Report, is available on Kaggle. This dataset facilitates exploratory analysis and model building within the data science community, offering insights into the factors contributing to national happiness levels.

2.8. The Role of Environmental and Social Factors

Modern research has increased understanding of how individual components affect subjective well-being, including studies conducted by psychologists, sociologists, and economists [15]. Economic circumstances, social relationships health habits are the main groups of factors. These are discussed in greater depth below:

- **Economic**
  The most frequently analyzed economic factors are employment and income. The positive correlation between income and life satisfaction exists only until the presence of "frustrated achievement," where the increase in income is associated with a decrease in life satisfaction, due to a decrease in areas such as health and social relationship quality [5]. The other factor of concern is unemployment, that provide evidence of a consistent strong negative correlation with happiness.

- **Social Relationships**
  Another group of factors that have been shown to have a strong influence on subjective well-being is social relationships [16]. Family relationships influence on the reported level of happiness. In contrast to family and social support, which had a strong positive relationship with happiness change, there was a small, positive impact on happiness as also positive correlation between successful marriage and overall feeling of happiness. In fact, the quantity and quality of social relationships [17] (friendships, marriage, religious organization membership) have both short-term and long-term effects on life happiness.

- **Health**
  According to studies, if your health problems interfere with your daily life, you will struggle the most to find happiness. There is a strong relationship between, physical and mental health, with the stronger correlation being present for the mental health than physical health [18][7]. A healthy lifestyle is more enjoyable regardless of its additional physical health benefits, such as regular exercise, which makes life more satisfying in general [19]. However, it is also possible that healthy living is not particularly pleasurable, and that health educators frequently try to force us to do things we do not want to do, which probably will bring us to mental health. These factors can determine how healthy we are.

2.9. Machine Learning for Happiness Purpose

Machine Learning techniques have become invaluable in analyzing happiness levels among individuals and populations. These techniques allow for the extraction of patterns and insights from large-scale data sets, enabling researchers to gain a deeper understanding of factors that contribute to happiness and identify potential interventions or policy changes that can improve overall well-being[15].

Additionally, ML techniques can help in predicting happiness levels based on various inputs such as demographic information, social media activity, and survey responses. By training algorithms on labelled data that includes indicators of Happiness Score, such as health, ML techniques can learn to accurately classify and predict a country happiness level. Moreover, ML techniques can also uncover complex relationships and interactions between different variables, shedding light on the intricate interplay of factors that influence happiness. These advancements in ML techniques allow for a more nuanced and comprehensive analysis of happiness, leading to evidence-based strategies that can enhance individual.

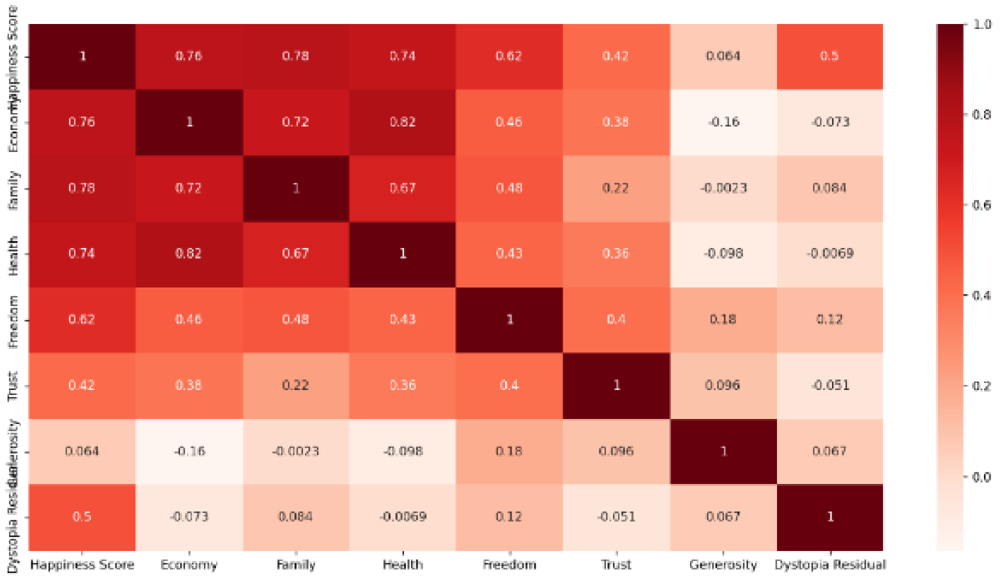
Furthermore, machine learning techniques can unearth complex relationships and interactions between different variables, offering a more comprehensive understanding of the multifaceted nature of happiness [20]. Overall, the integration of machine learning in happiness analysis stands to revolutionize the field.

2.10. Happiness Score Features Correlations

In this article [1], the author embarks on an investigation into the measurement of happiness, particularly focusing on the challenges posed by the COVID-19 pandemic's impact on 2021-2022 happiness scores [21]. Inspired by Bhutan's innovative Gross Happiness National (GHN) metric, the author seeks to explore how happiness is quantified globally, with an emphasis on the United Nations' World Happiness Report methodology and its reliance on six key variables for calculating a country's Happiness Score.

This research reveals a gap in understanding the specific variables, like GDP and health, influencing happiness scores, leading the author to conduct a study analyzing the 2021-2022 World Happiness Report data. By employing linear regression analysis in Python, the author aims to predict happiness scores based on GDP, offering new insights into the economic determinants of happiness.

Figure 1: Source - Y. Zhang, "Analyze and Predict the 2022 World Happiness Report Based on the Past Year's Dataset," The heat map of each parameter´s correlations



The author's analysis focuses on the six key variables identified by the World Happiness Report - economic production (GDP), social support, life expectancy, freedom, absence of corruption, and generosity—to determine their impact on happiness scores. The initial findings, illustrated in Figure 1, reveal that GDP per capita and social support are the most significant predictors of happiness, with correlation coefficients of 0.78 and 0.76, respectively, indicating strong relationships with happiness scores. In contrast, generosity is identified as the least influential variable, with a correlation coefficient of just 0.18 [1].

2.11. Relationship Between Happiness Scores and GDP per Capita

The author presents an analysis based on six years of World Happiness Report data (2016-2021) sourced from Kaggle. This dataset, comprising 1084 entries, is segmented into training (75%), testing (15%), and comparison (10%) sets to develop a predictive model. The model, visualized in Figures 2 and 3, employs linear regression to establish a relationship between happiness scores and GDP per capita, yielding a target equation with coefficients indicating GDP as a significant predictor of happiness [1].

Figure 2: Source - Y. Zhang, "Analyze and Predict the 2022 World Happiness Report Based on the Past Year's Dataset," Distribution of the training model

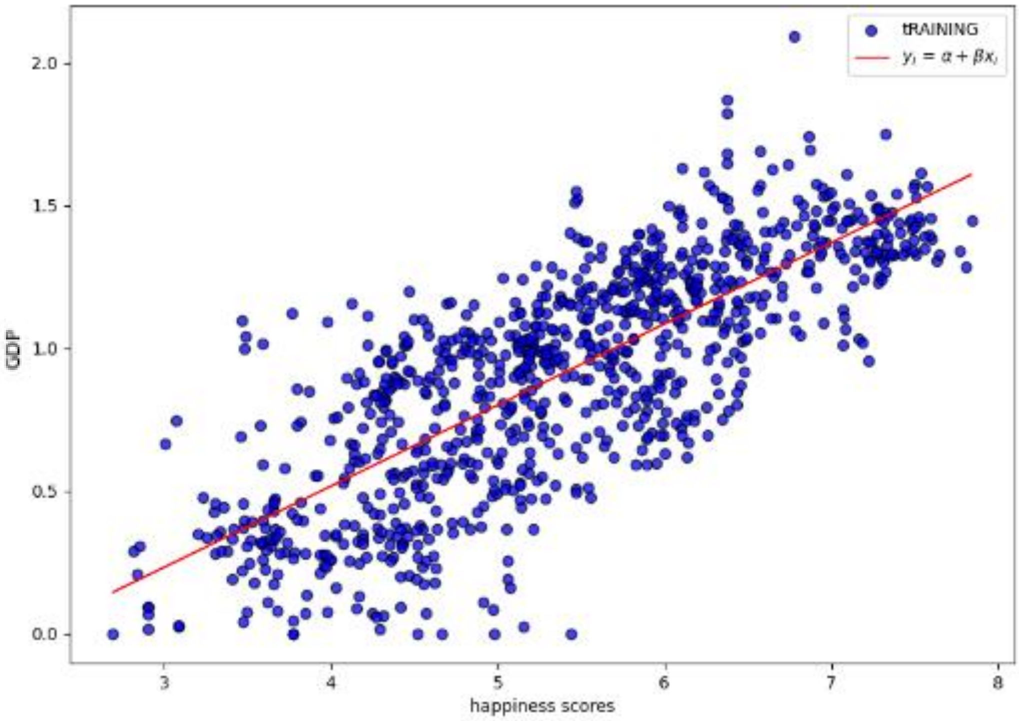
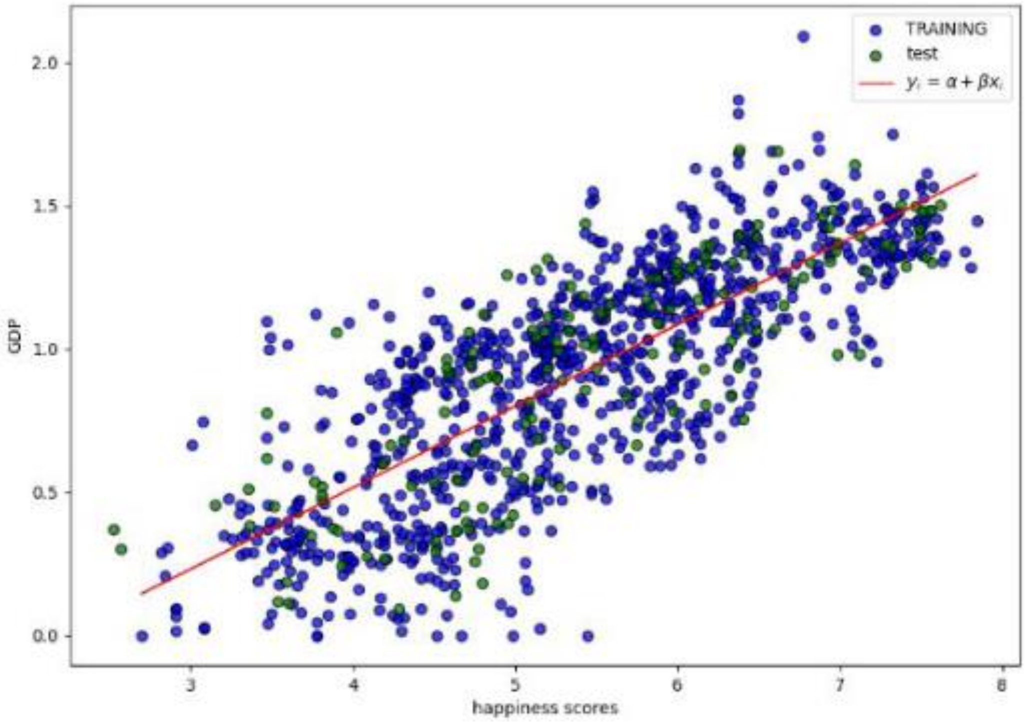


Figure 3: Source - Y. Zhang, "Analyze and Predict the 2022 World Happiness Report Based on the Past Year's Dataset," The training and test data distribution



The author's findings, corroborated by testing and validation phases (Figures 2 and 3), demonstrate the linear regression model's efficacy in fitting the data, with performance metrics (MAE, MSE, and RMSE) indicating satisfactory prediction accuracy. However, when comparing the predictive power of GDP against the family variable, despite the latter's high correlation with happiness scores, GDP emerges as the superior predictor due to lower error metrics in the GDP-based model.

The discussion acknowledges the limitations of linear regression, especially when considering the multifaceted nature of happiness influenced by various factors beyond GDP. The author suggests that while GDP is a strong predictor, it may not always accurately reflect a country's happiness score, citing the example of Singapore and Hong Kong's discrepancy between GDP rank and happiness rank.

The examination of the relationship between happiness scores and GDP per capita within this study underscores the multifaceted nature of happiness. Contrary to the simplistic view of happiness as a singularly attainable objective, our findings elucidate that happiness emerges from a complex interplay of factors, including but not limited to freedom, corruption levels, generosity, social support, economic output, and life expectancy within a nation [10]. Specifically, the correlation between GDPs per capita and happiness scores highlights the economic dimension as a significant, albeit not exclusive, determinant of societal well-being.

This insight aligns with the broader understanding that while economic prosperity can contribute to enhancing quality of life, the essence of happiness is also intricately linked to other elements of human experience and governance.

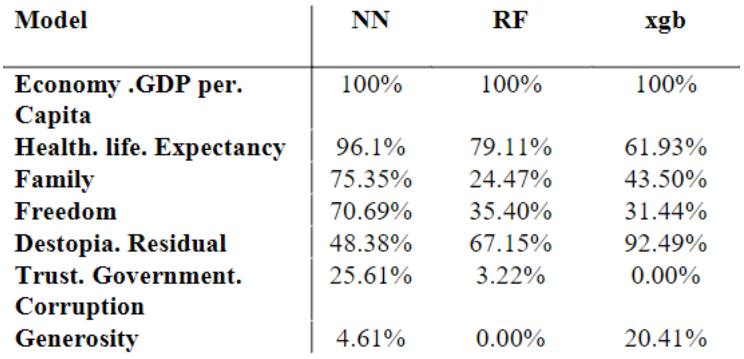
2.12. Machine Learning Approaches

In this article [20], the analysis of the World Happiness Report dataset underscores the pivotal role of GDP per capita as a determinant of the world happiness score. Initial results, leveraging a neural network-trained model, highlighted GDP per capita's significance, a finding corroborated by both Random Forest (RF) and Extreme Gradient Boosting (XGBoost) models, as depicted in figures demonstrating variable importance. These models collectively affirmed GDP per capita's critical influence on happiness scores, substantiated through ROC curve analysis and variable importance rankings, thereby addressing the first research question by codifying GDP per capita as a fundamental element in determining life happiness scores.

Subsequently, the investigation revealed health life expectancy as a paramount attribute, prioritized in logistic regression and information gain models, spotlighting its direct correlation with higher happiness scores. This insight was further validated by OneR rules, which classified health life expectancy as the foremost rule in determining life happiness scores across different classes. The consistency of this finding was also evident in the decision tree approach, where health life expectancy emerged as the primary rule, answering the second research question regarding its critical role in enhancing life happiness scores.

A comparative summary of insights gained from applying neural network, RF, and XGBoost models showcased the varying degrees of influence of different variables on happiness scores, with GDP per capita and health life expectancy ranking as the most impactful. This analysis provided a hierarchical understanding of variables affecting happiness, where economic and health factors stood out as primary contributors, followed by other variables such as family support, freedom, dystopia residual, government trust, and generosity to varying extents.

Table 7:Source - M. A. Khder, M. A. Sayfi, and S. W. Fujo, "Analysis of World Happiness Report Dataset Using Machine Learning Approaches," Gained insights after applying models: nn, rf and xgb



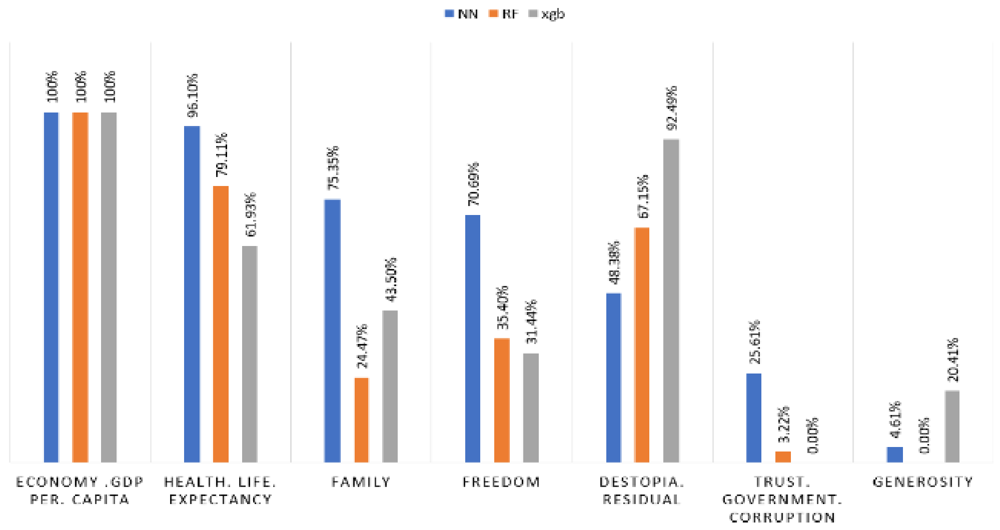
Model	NN	RF	xgb
Economy .GDP per. Capita	100%	100%	100%
Health. life. Expectancy	96.1%	79.11%	61.93%
Family	75.35%	24.47%	43.50%
Freedom	70.69%	35.40%	31.44%
Destopia. Residual	48.38%	67.15%	92.49%
Trust. Government. Corruption	25.61%	3.22%	0.00%
Generosity	4.61%	0.00%	20.41%

This comprehensive evaluation of machine learning methods and the derived insights significantly contribute to understanding the critical determinants of happiness, offering a nuanced perspective on how different factors interplay to shape overall life satisfaction and happiness scores.

The findings not only elucidate the paramount importance of economic and health conditions in predicting happiness but also highlight the relevance of social and governance factors, thus enriching the literature on happiness research and its predictors.

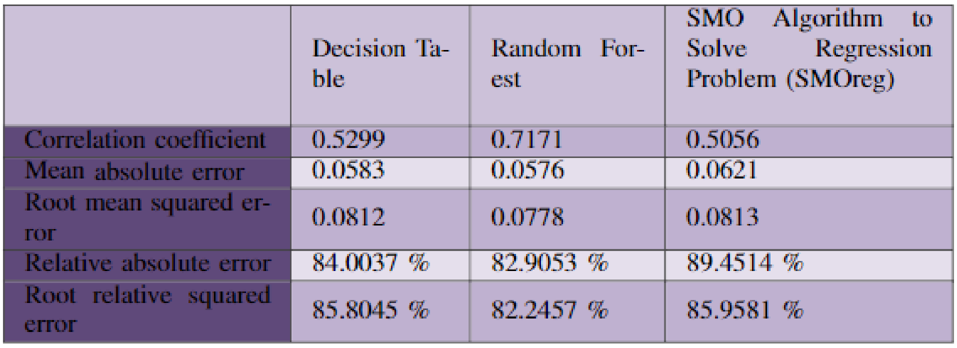
To complement the analysis, Figure 4 provides a conclusive summary of the importance of happiness variables, vividly illustrating that the most critical variables influencing happiness scores are the economy (GDP per capita) and health life expectancy [20]. This visual representation underscores the predominant impact of these variables, reinforcing the findings from the neural network, RF, and XGBoost models. The emphasis on economic stability and health as foundational to happiness not only aligns with previous research insights but also highlights the integral role these factors play in shaping the well-being of populations globally. Through this detailed exploration, the study contributes significantly to the literature, offering a clearer understanding of the key drivers behind happiness scores and underlining the importance of fostering economic and health advancements to enhance societal happiness.

Figure 4: Source - M. A. Khder, M. A. Sayfi, and S. W. Fujo, "Analysis of World Happiness Report Dataset Using Machine Learning Approaches," Summary for importance of the happiness variables



In other study, a series of machine learning algorithms were deployed, utilizing Weka software to assess the World Happiness Report data. The algorithms applied, including decision tables, random forest, and SMOreg, culminated in a comprehensive table of results [22]. A pivotal aspect of this analysis was the feature ranking, which aimed to delineate the most significant attributes contributing to happiness as per the dataset.

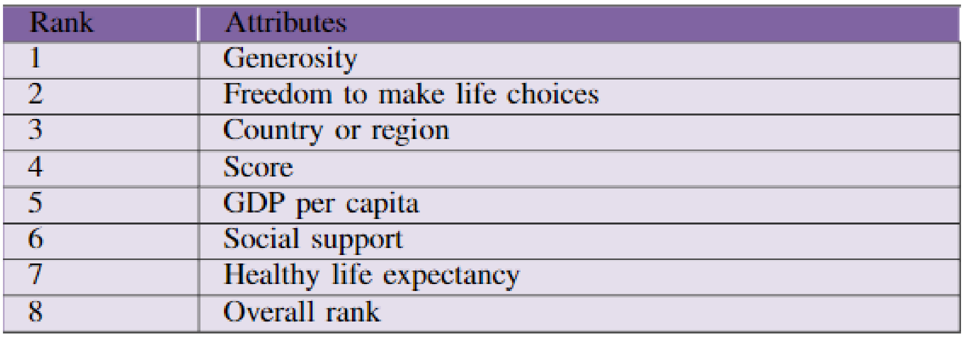
Table 8: Source -F. Ibnat, J. Gyalmo, Z. Alom, M. A. Awal, and M. A. Azim, "Understanding World Happiness using Machine Learning Techniques," Results based on classifiers



	Decision Table	Random Forest	SMO Algorithm to Solve Regression Problem (SMOreg)
Correlation coefficient	0.5299	0.7171	0.5056
Mean absolute error	0.0583	0.0576	0.0621
Root mean squared error	0.0812	0.0778	0.0813
Relative absolute error	84.0037 %	82.9053 %	89.4514 %
Root relative squared error	85.8045 %	82.2457 %	85.9581 %

Table 8 from the study presents a synthesis of the results from these classifiers, offering insights into the predictive accuracy of the models. It notes the correlation coefficients, mean absolute error (MAE), root mean squared error (RMSE), and relative errors, providing a quantitative measure of each classifier's performance. Notably, the Random Forest algorithm exhibited a superior correlation coefficient of 0.7171, suggesting a strong predictive relationship.

Table 9: Source -F. Ibnat, J. Gyalmo, Z. Alom, M. A. Awal, and M. A. Azim, "Understanding World Happiness using Machine Learning Techniques,"Ranked attributes



Rank	Attributes
1	Generosity
2	Freedom to make life choices
3	Country or region
4	Score
5	GDP per capita
6	Social support
7	Healthy life expectancy
8	Overall rank

Table 9 focuses on the ranked attributes, highlighting the relative importance of various factors in determining happiness scores. It is particularly noteworthy that 'Generosity' and 'Freedom to make life choices' are ranked as the top two attributes, preceding even 'GDP per capita', which is traditionally considered a significant predictor of national happiness levels.

This prioritization of attributes, with 'Generosity' at the helm, followed by 'Freedom to make life choices', and then 'GDP per capita', may indicate a paradigm shift in understanding what drives happiness across nations [22]. Moreover, the ranking posits 'Healthy life expectancy' lower than might be anticipated, suggesting that while health is important, other social and economic factors play a more pronounced role in shaping happiness.

2.13. Challenges and Future Directions

These research paper has illuminated critical aspects of happiness determinants, leveraging data science and machine learning methodologies to scrutinize the variables within the dataset crucial to understanding life happiness scores. Predominantly, the analysis classified GDP per Capita and health life expectancy as the most significant factors influencing happiness, with GDP recognized as a primary indicator through a Neural Network approach. The robustness of these findings was further enhanced by employing various analytical approaches, thereby augmenting the precision and reliability of the results.

Moreover, the study unearthed that high life expectancy is a pivotal determinant of life happiness, a conclusion drawn using the OneR classification method. This insight was substantiated by evaluating the outcome through diverse performance metrics, which solidified the argument's credibility. The intention for future research endeavors includes the application of a broader spectrum of machine learning techniques to an expanded dataset covering a more extended period. Additionally, there is a planned exploration into deep learning methodologies to unearth deeper insights into the drivers of happiness.

Happiness research, a complex yet vital area of societal development, benefits significantly from the advancements in data science tools for modeling and analyzing happiness predictions. These methodologies offer substantial potential in addressing the intricacies and challenges inherent in happiness studies. However, the limitations of current machine learning approaches, including the omission of significant events that could impact specific countries' happiness scores, underscore the necessity for a nuanced and comprehensive analytical framework.

This article integrates these discussions and findings to highlight the ongoing challenges and prospective directions in happiness research. By conducting a comparative analysis of World Happiness Reports between 2021-2022, which revealed minimal changes in happiness scores, this study emphasizes the potential of linear regression models for predicting happiness and underscores the importance of expanding analytical approaches. The combined conclusions of those research not only contribute to the academic discourse on happiness but also chart a path forward for utilizing advanced data science techniques to uncover the multifaceted nature of happiness determinants, thereby aiding in the development of more effective societal well-being strategies.

Future research, as suggested by the study, will expand the machine learning models employed to include linear regression, logistic regression, SVM, and others. This expansion is crucial, as the world happiness dataset from 2019 is not the most current, and newer reports may yield divergent results and attribute significance.

Upcoming analyses will be instrumental in discerning changes in attribute distribution and how these shifts manifest over time, offering a dynamic view of the factors that influence global happiness. This ongoing research will undoubtedly provide valuable contributions to the literature, enriching our understanding of happiness and its complex interplay with socioeconomic factors.

3. Chapter 3 – Implementation and Results

3.1 Business Understanding

As previously stated, the purpose of the research is to use machine learning models by using data from the World Happiness Report. We will employ the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology (Chapman, P. et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide), a widely accepted framework for data mining and machine learning projects. CRISP-DM consists of six phases:

1. Business Understanding,
2. Data Understanding,
3. Data Preparation,
4. Modeling,
5. Evaluation,
6. Deployment.

This structured approach ensures a thorough and systematic analysis, enhancing the reliability and validity of our findings.

I aim to demonstrate that different algorithms can be utilized to answer research questions through an exploratory analysis of the World Happiness Report dataset.

The outlined stages of chapter 3 include:

1) Data preparation, including missing values handling, feature selection, data resampling and normalization.

2) Exploratory analysis of the World Happiness Report dataset.

3) Modelling, including models selection, training, and testing machine learning models.

4) Evaluation of results.

3.2 Data Understanding

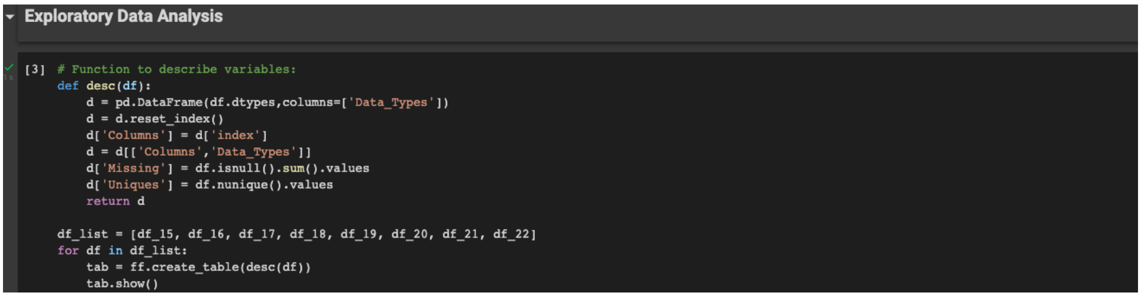
The dataset for the World Happiness Report is a flat file with several features that change from year to year. The goal of the initial data quality investigation is to identify any potential issues by analyzing descriptive statistics, data trends, data distribution and missing values.

3.3 Data Preparation and Exploratory Analysis of the WHR dataset

Initially, I imported the necessary libraries and machine learning algorithms. These algorithms are essential for training the models in subsequent steps. For the purpose of evaluation, I utilized the mean_squared_error metric. This metric is crucial for assessing the performance of both classification and regression models post-training. A lower mean_squared_error, closer to zero, indicates a more accurate model. Conversely, a mean_squared_error closer to 1 suggests the model may not perform well on this particular dataset.

At this stage, after describing the variables for all the years from 2015 to 2022, I detected that the features name of each column sometimes has different names, such as, for example, the column “Happiness Rank” in other years appears as “Happiness.Rank” or “Overall rank”, another example the column “Country” sometimes appears as “Country or region” in other columns from other years. In addition, the number of columns increases with the years, the specificity of criteria increases with the years.

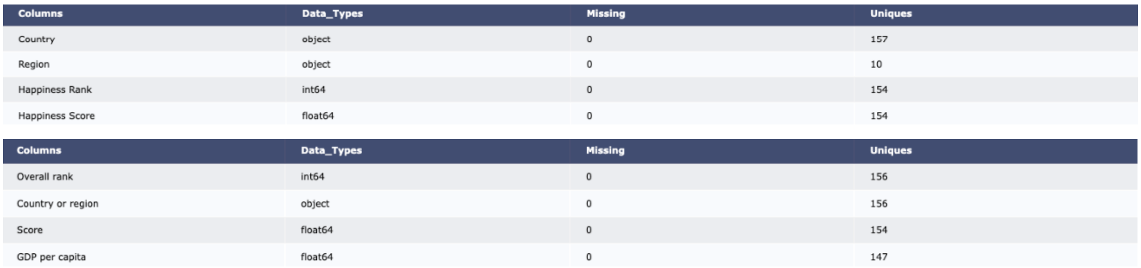
Figure 5:Function to describe feature.



Exploratory Data Analysis

```
[3]  # Function to describe variables:
     def desc(df):
         d = pd.DataFrame(df.dtypes,columns=['Data_Types'])
         d = d.reset_index()
         d['Columns'] = d['index']
         d = d[['Columns','Data_Types']]
         d['Missing'] = df.isnull().sum().values
         d['Uniques'] = df.nunique().values
         return d

     df_list = [df_15, df_16, df_17, df_18, df_19, df_20, df_21, df_22]
     for df in df_list:
         tab = ff.create_table(desc(df))
         tab.show()
```

Table 10: Column name example



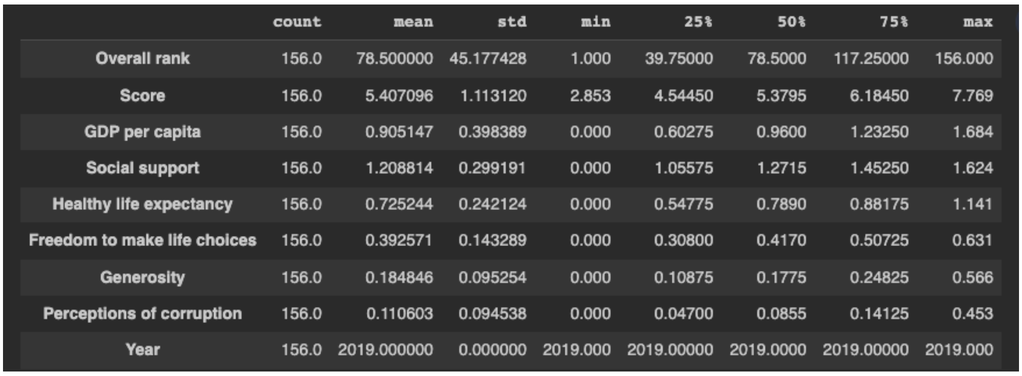
Columns	Data_Types	Missing	Uniques
Country	object	0	157
Region	object	0	10
Happiness Rank	int64	0	154
Happiness Score	float64	0	154

Columns	Data_Types	Missing	Uniques
Overall rank	int64	0	156
Country or region	object	0	156
Score	float64	0	154
GDP per capita	float64	0	147

3.4.1 Statistical Analysis

In order to have a deeper analysis of the data, I described the year 2019 (the same can be done for any year) with the aim of detailing the information that exists in, as we can see in table 11.

Table 11: WHR feautures



	count	mean	std	min	25%	50%	75%	max
Overall rank	156.0	78.500000	45.177428	1.000	39.75000	78.5000	117.25000	156.000
Score	156.0	5.407096	1.113120	2.853	4.54450	5.3795	6.18450	7.769
GDP per capita	156.0	0.905147	0.398389	0.000	0.60275	0.9600	1.23250	1.684
Social support	156.0	1.208814	0.299191	0.000	1.05575	1.2715	1.45250	1.624
Healthy life expectancy	156.0	0.725244	0.242124	0.000	0.54775	0.7890	0.88175	1.141
Freedom to make life choices	156.0	0.392571	0.143289	0.000	0.30800	0.4170	0.50725	0.631
Generosity	156.0	0.184846	0.095254	0.000	0.10875	0.1775	0.24825	0.566
Perceptions of corruption	156.0	0.110603	0.094538	0.000	0.04700	0.0855	0.14125	0.453
Year	156.0	2019.000000	0.000000	2019.000	2019.00000	2019.0000	2019.00000	2019.000

For all variables it is now possible to see the total entries, the mean, standard deviation, minimum and maximum.

3.4.2 Data ta Manipulations

In conducting a comprehensive analysis of the dataset spanning from 2015 to 2019, it was observed that these years maintained a consistent structure, sharing the same number of columns, which is crucial for the integrity and comparability of our analysis. This structural consistency allowed for a cohesive analytical framework, thereby supporting the decision to exclude data from 2020 to 2022, which deviated from this uniformity.

The decision to not utilize data from 2020 to 2022 was not made lightly. Upon closer examination, these datasets introduced significant structural changes and potentially different variables, which could introduce complexity and inconsistency into the analysis. The aim was to ensure a robust and coherent evaluation of trends over time without the confounding effects of changing data schemas.

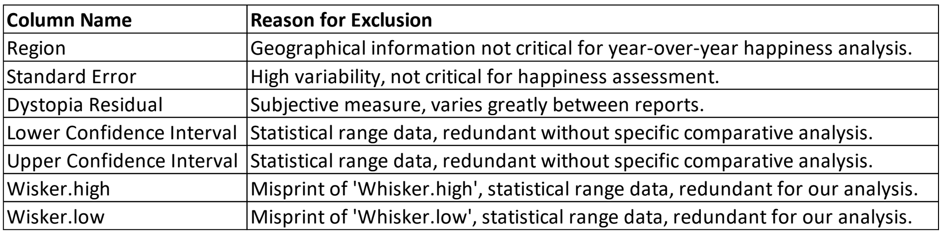
Moreover, while the datasets from 2015 to 2019 did not adhere to a uniform naming convention, efforts were made to standardize this aspect by renaming columns for clarity and ease of integration. For the years 2015 and 2016, where the data conversion by features was identical, adjustments were made to align the column order with those of 2017, 2018, and 2019 to facilitate a seamless aggregation of the datasets.

The decision to exclude later datasets stems from a rigorous assessment of the potential impact of integrating dissimilar data structures on the study's outcomes. It is recognized that with additional resources and time, a more complex data harmonization process could potentially integrate the 2020 to 2022 data. However, given the scope and objectives of the current analysis, which prioritizes maintaining a consistent and comparable data framework over time, the inclusion of these datasets was deemed not to contribute meaningfully to the analysis' objectives, possibly detracting from the reliability and clarity of insights derived.

This approach underscores a methodological choice aimed at preserving analytical precision and relevance, acknowledging that while data uniformity efforts are feasible, they must be balanced against the potential for introducing variability that could obscure the analysis' foundational trends and insights.

In the process of combining the datasets from 2015 to 2019 using the pandas.concat function, I strategically omitted certain columns that were not pertinent to our analysis objectives. This decision was made to streamline the dataset and focus on the most relevant factors. Below is a table outlining the columns excluded, along with the rationale for each:

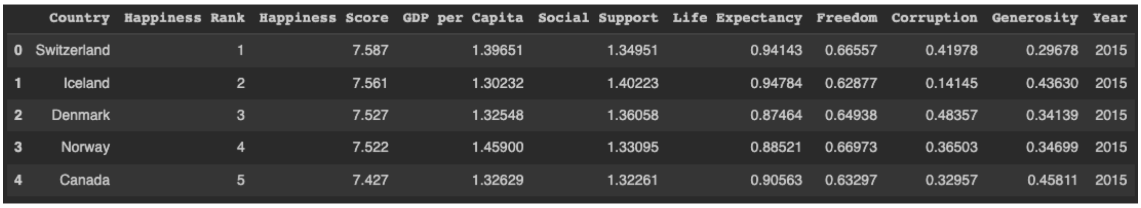
Table 12: Dropping columns



Column Name	Reason for Exclusion
Region	Geographical information not critical for year-over-year happiness analysis.
Standard Error	High variability, not critical for happiness assessment.
Dystopia Residual	Subjective measure, varies greatly between reports.
Lower Confidence Interval	Statistical range data, redundant without specific comparative analysis.
Upper Confidence Interval	Statistical range data, redundant without specific comparative analysis.
Wisker.high	Misprint of 'Whisker.high', statistical range data, redundant for our analysis.
Wisker.low	Misprint of 'Whisker.low', statistical range data, redundant for our analysis.

The filtration process streamlined our dataset to focus on crucial variables from 2015 to 2019, enhancing data consistency and relevance. The resulting data frame includes key metrics: Happiness Rank, Happiness Score, GDP per Capita, Social Support, Life Expectancy, Freedom, Corruption (inversely measured), Generosity, and Year. This selection underpins a focused analysis of the World Happiness Report, aiming for a deeper insight into global happiness determinants.

Table 13: Data frame features

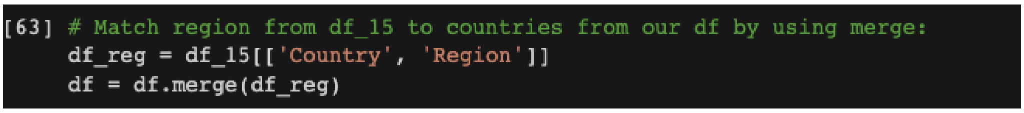


	Country	Happiness Rank	Happiness Score	GDP per Capita	Social Support	Life Expectancy	Freedom	Corruption	Generosity	Year
0	Switzerland	1	7.587	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678	2015
1	Iceland	2	7.561	1.30232	1.40223	0.94784	0.62877	0.14145	0.43630	2015
2	Denmark	3	7.527	1.32548	1.36058	0.87464	0.64938	0.48357	0.34139	2015
3	Norway	4	7.522	1.45900	1.33095	0.88521	0.66973	0.36503	0.34699	2015
4	Canada	5	7.427	1.32629	1.32261	0.90563	0.63297	0.32957	0.45811	2015

3.4.3 Region and Continent

In this section I added Region and Continent columns and use Regions from df_15 and merge function on my df data frame with the purpose of assigning to Continent. For there is the problem where in some cases the region is provided, but the country is not. To solve this problem this, I used pycountry_convert library.

Figure 6: Match Regions from 2015 dataset

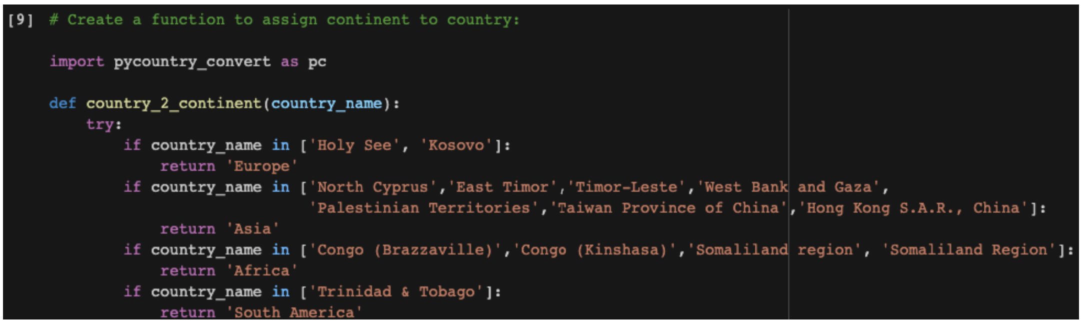


```
[63] # Match region from df_15 to countries from our df by using merge:
     df_reg = df_15[['Country', 'Region']]
     df = df.merge(df_reg)
```

Upon consolidating the datasets spanning 2015 to 2019 into a single dataframe, it became apparent that there are missing values, particularly in the 'Region' and 'Continent' columns. To address this, I leveraged the 2015 dataset—which contains comprehensive regional information—and utilized the merge function to enrich our consolidated dataframe with the necessary geographic details. Additionally, the pycountry_convert library was employed to facilitate the accurate mapping of countries to their respective continents and regions, ensuring a more robust dataset for analysis.

This step is crucial for enhancing the dataset's completeness and reliability, thereby supporting more nuanced geographical insights in this study.

Figure 7: Function to assign continent to country

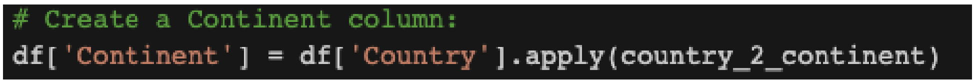


```
[9] # Create a function to assign continent to country:

    import pycountry_convert as pc

    def country_2_continent(country_name):
        try:
            if country_name in ['Holy See', 'Kosovo']:
                return 'Europe'
            if country_name in ['North Cyprus','East Timor','Timor-Leste','West Bank and Gaza',
                                'Palestinian Territories','Taiwan Province of China','Hong Kong S.A.R., China']:
                return 'Asia'
            if country_name in ['Congo (Brazzaville)','Congo (Kinshasa)','Somaliland region', 'Somaliland Region']:
                return 'Africa'
            if country_name in ['Trinidad & Tobago']:
                return 'South America'
```

Following the integration of the pycountry_convert library, I proceeded to map each country within our dataset to its respective continent. For instance, when identifying East Timor, the library categorizes it under Asia. This mapping applies exclusively to the countries present in our newly consolidated dataset from 2015 to 2019. The primary objective of this step is to facilitate a more structured and insightful graphical analysis by providing a clear geographical context. This enhancement allows for a deeper exploration of regional patterns and trends within the World Happiness Report data, ultimately contributing to a richer and more comprehensive analysis.

Finally, I applied this function to all countries in the data frame, in order to create a continent column.

Figure 8: Create a continent column



```
# Create a Continent column:
df['Continent'] = df['Country'].apply(country_2_continent)
```

3.4.4 Happiness Score

Now with the data worked on, and considering the statistical analysis, I can start to visualize the different features. For this, I selected the following features of the data set to create the following map from 2019:

- Happiness Score
- GDP per Capita
- Social Support
- Life Expectancy
- Freedom
- Corruption
- Generosity

Figure 9:Happiness Score

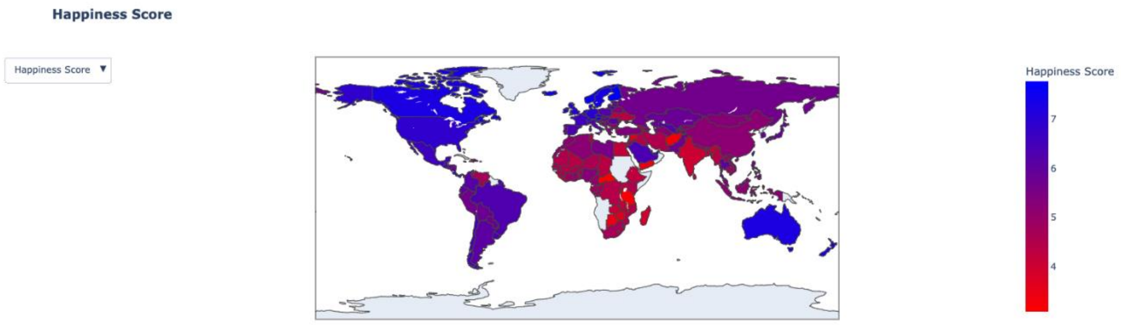


Figure 9 shows an overview about which countries have highest Happiness Score, continents like North America, Europe and Oceania have the highest score.

Figure 10: GPD per Capita Score

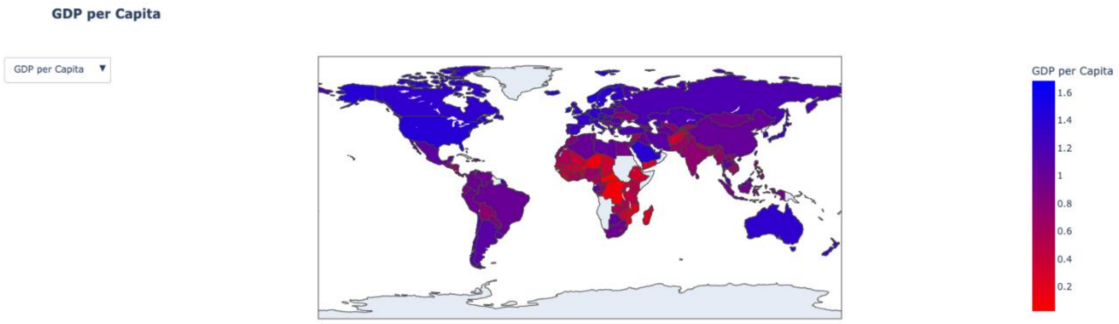


Figure 10 shows an overview about what countries have highest GPD per capita score, continents like North America, Europe, Oceania, and Saudi Arabia have the highest score.

Figure 11: Corruption Score

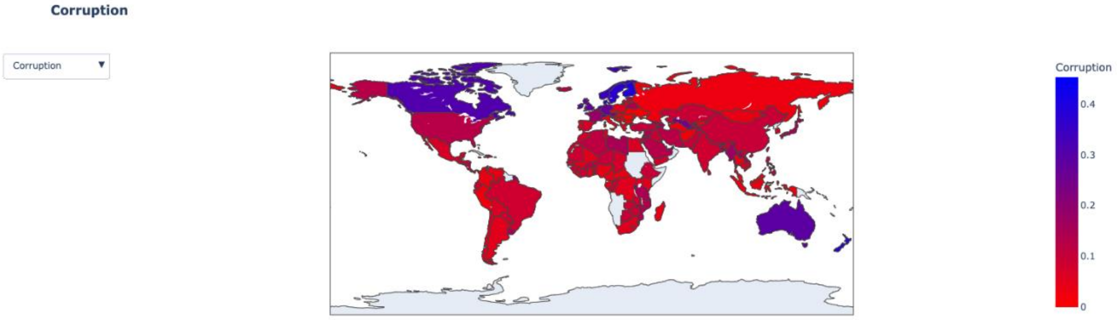
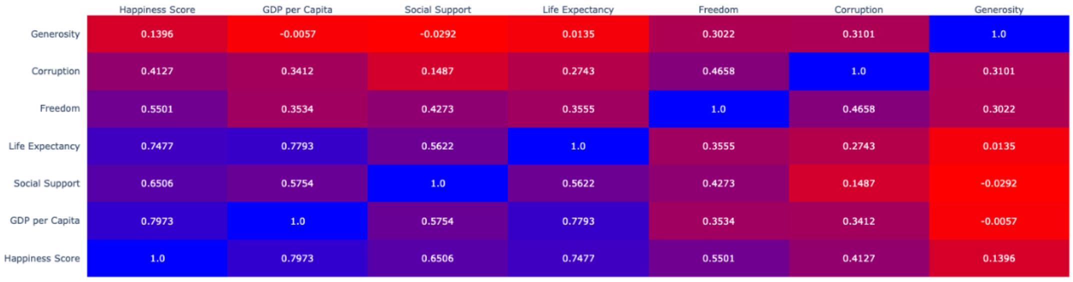


Figure 11 shows an overview about what countries have lowest corruption score, the closer the score is to 1, the less corruption there is countries. Countries like Canada, Norway, Sweden, Finland, and New Zealand have lowest corruption score.

3.4.5 Correlation Analysis

In order to find which features are related to each other, I create a data frame with the variables. The goal is to know the coalition between the features. To explain the following matrix, the bluer the link, it means that the variables are dependent on each other, and the redder they are, it means that there is less dependence between them.

Figure 12: Pending matrix between features



	Happiness Score	GDP per Capita	Social Support	Life Expectancy	Freedom	Corruption	Generosity
Generosity	0.1396	-0.0057	-0.0292	0.0135	0.3022	0.3101	1.0
Corruption	0.4127	0.3412	0.1487	0.2743	0.4658	1.0	0.3101
Freedom	0.5501	0.3534	0.4273	0.3555	1.0	0.4658	0.3022
Life Expectancy	0.7477	0.7793	0.5622	1.0	0.3555	0.2743	0.0135
Social Support	0.6506	0.5754	1.0	0.5622	0.4273	0.1487	-0.0292
GDP per Capita	0.7973	1.0	0.5754	0.7793	0.3534	0.3412	-0.0057
Happiness Score	1.0	0.7973	0.6506	0.7477	0.5501	0.4127	0.1396

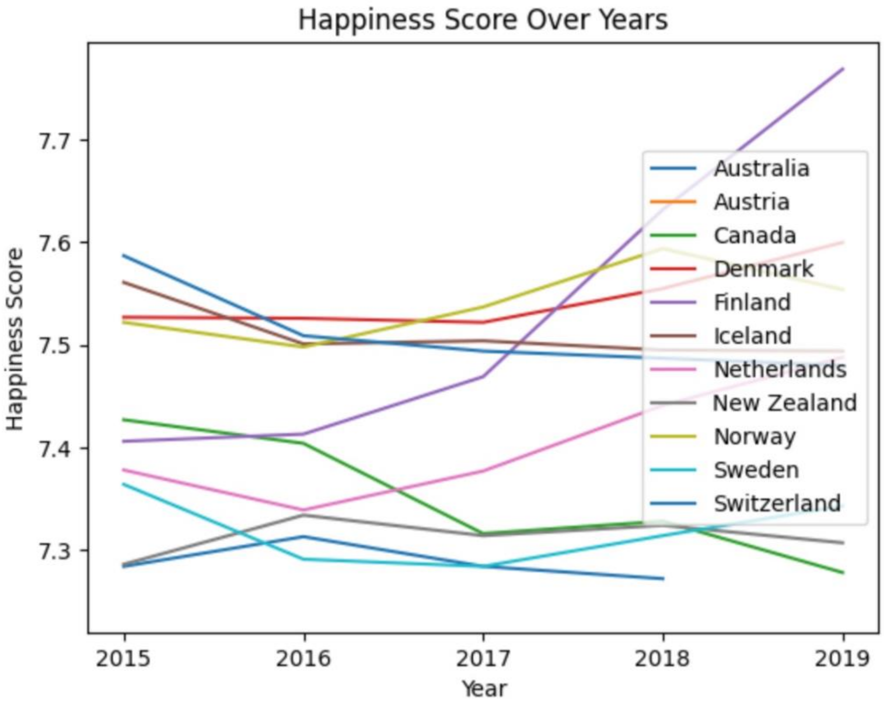
For example, in this case we can see that there is a high link between Happiness Score and GPD per Capita (0.7977, close to 1), if GPD per Capita increases, it is very likely that the Happiness Score will also increase. The opposite is true for the variable GPD per Capita, it is not dependent on Generosity, so this Correlation is red (0.0057), as I can see in the matrix above.

Happiness Score has strong positive correlation with GDP per Capita (0.7973), Life Expectancy (0.7477), Social Support (0.6506) followed by Freedom (0.5501). This means that the population on rich countries with higher life expectancy, social support and freedom tends to be happier.

3.4.6 Top 10 Happiness Countries 2015 to 2019

For this step, only the countries with the Happiness Rank feature equal or less than 10 were selected. As mentioned above, I´m working with data from 2015 to 2019, and the figure 13 graphs depict the variation over these same years:

Figure 13:Top 10 Happiness Countries from 2015 to 2019



Throughout this five-year period, it's evident that most of the top-ranked countries are members of the European Union, suggesting that residing in Europe significantly correlates with a higher quality of life. Particularly in the last three years, an observable trend is that Northern European countries have consistently ranked among the top three happiest nations. This observation highlights the region's substantial contribution to well-being and life satisfaction.

3.4.7 Happiness vs. GDP per capita

Through the use of a simple scatter plot, it was possible to create the link between the Happiness Score and the GPD per capita. This way I can analyze the countries and continents where an increase in GPD per capita reflects a high Happiness score. The same analysis was performed for the years between 2015 and 2019.

Figure 14: Happiness vs. GDP 2015

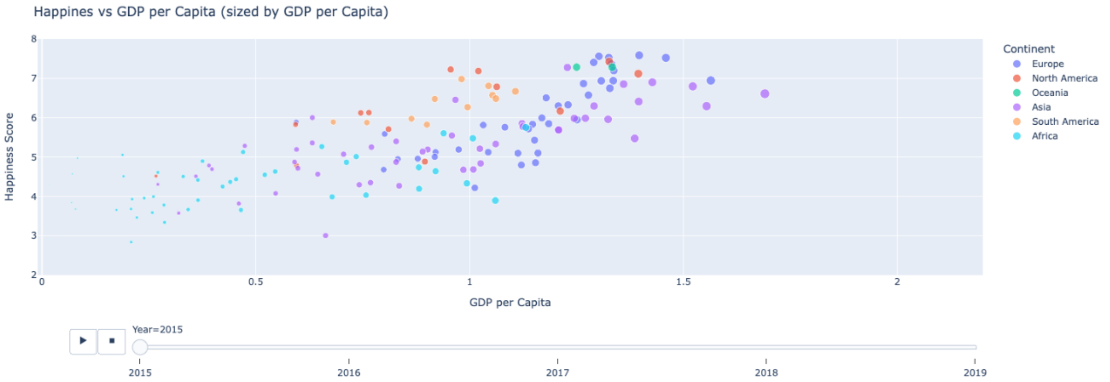


Figure 15: Happiness vs. GDP 2016

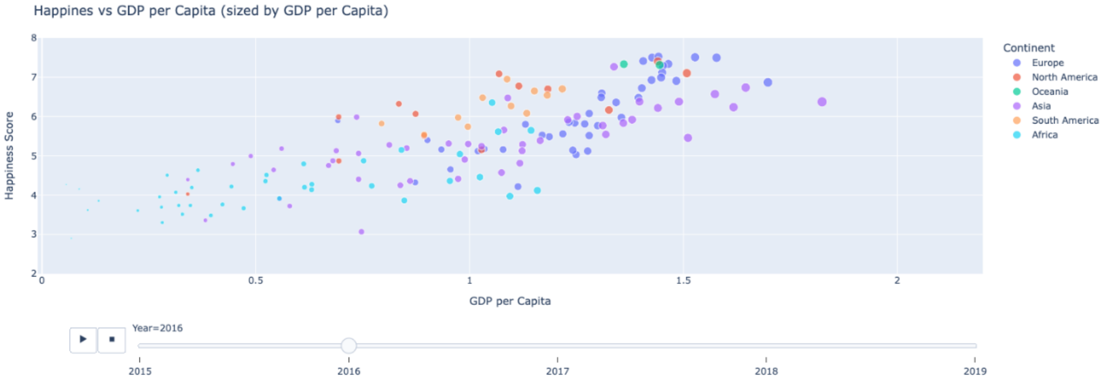


Figure 16: Happiness vs. GDP 2017

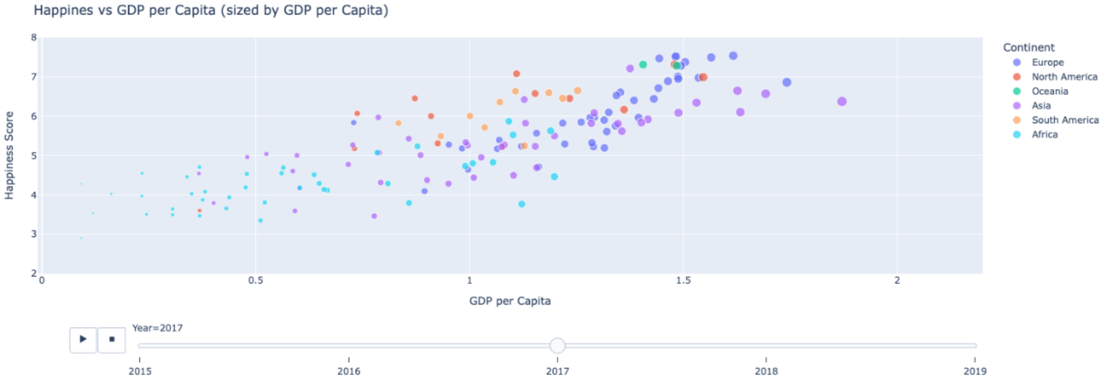


Figure 17: Happiness vs. GDP 2018

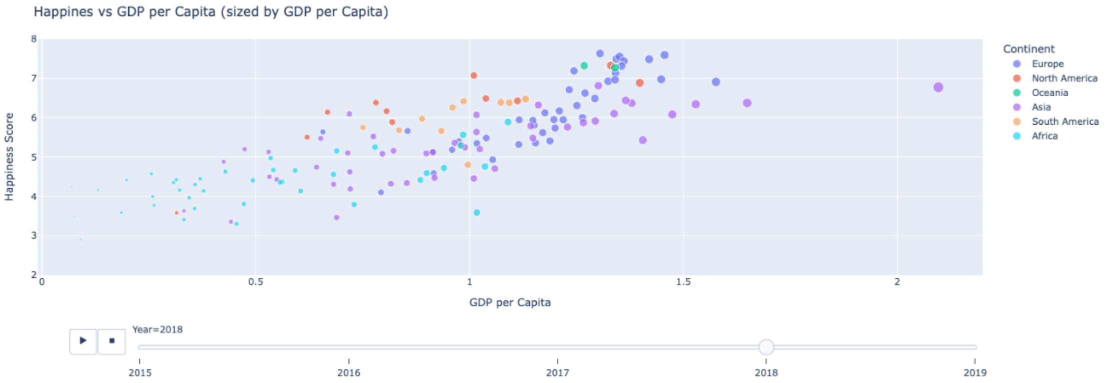
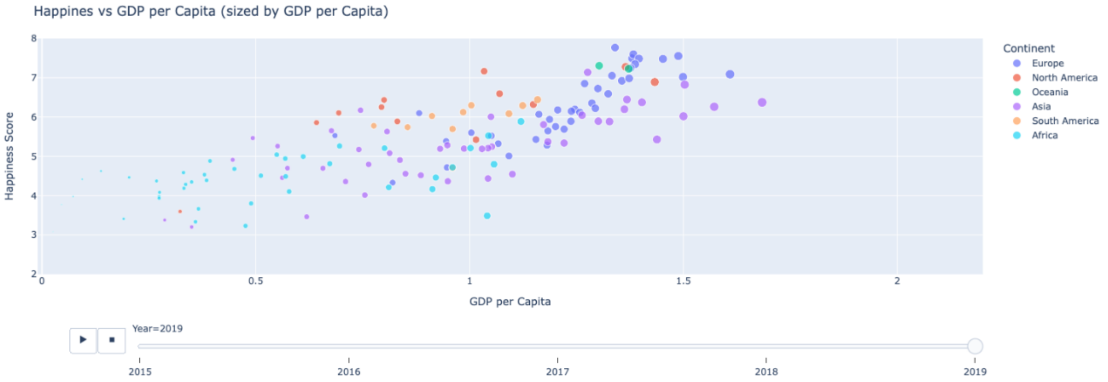


Figure 18: Happiness vs. GDP 2019



Upon conducting an initial analysis of the data spanning 2015 to 2019, it was observed that countries from the European continent consistently appear in the upper right quadrant of the cluster analysis. Specifically, Switzerland led in 2015, followed by Denmark in 2016, Norway in 2017, and Finland in 2018. Notably, in 2019, this trend persisted, with Finland maintaining its position as the country with the highest score. This pattern underscores a consistent dominance of European nations in terms of their rankings within the dataset.

3.4.8 Happiness vs. Life Expectancy

Building on the previously mentioned insights, I employed a scatter plot to further explore the relationship, this time between happiness score and life expectancy, while also considering GDP per capita as a size factor for each data point.

Figure 19: Happiness vs. Life Expectancy 2015

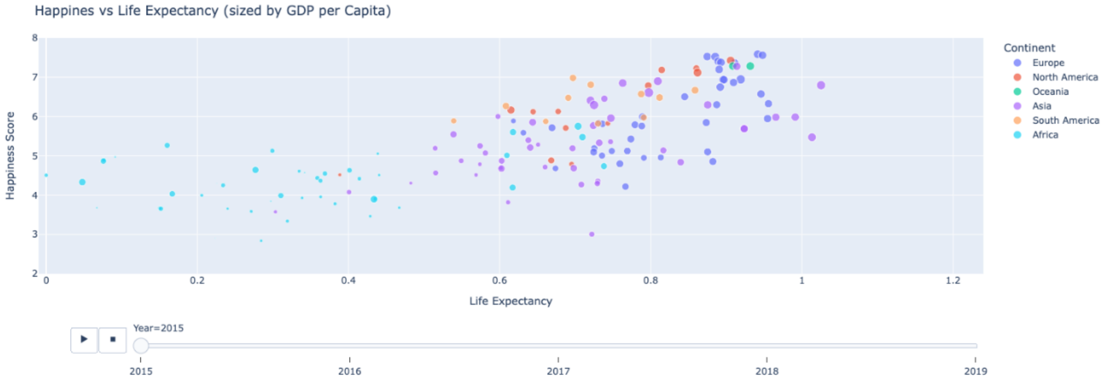


Figure 20: Happiness vs. Life Expectancy 2016

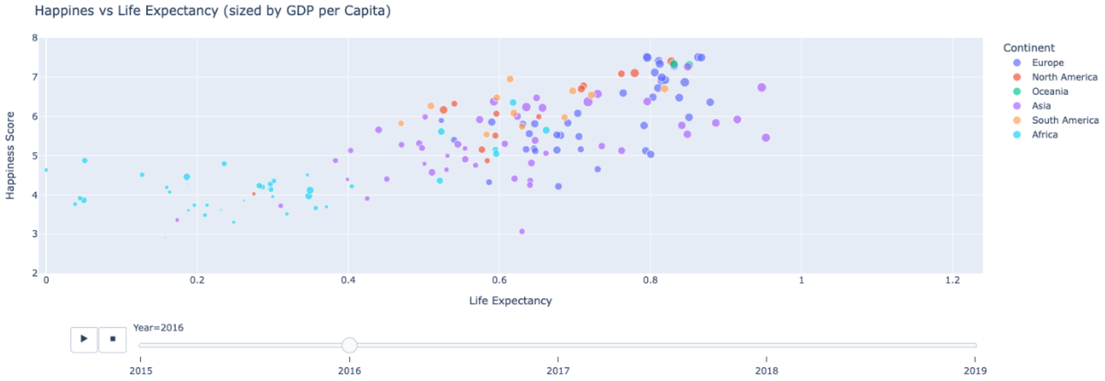


Figure 21: Happiness vs. Life Expectancy 2017

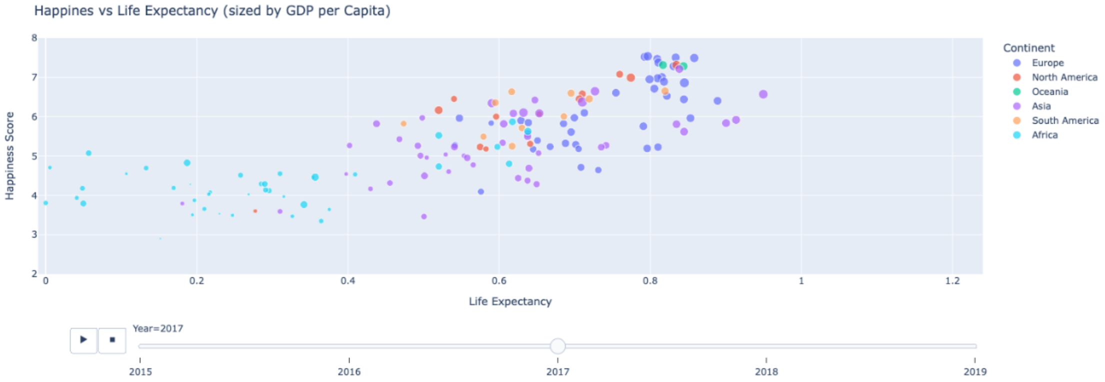


Figure 22: Happiness vs. Life Expectancy 2018

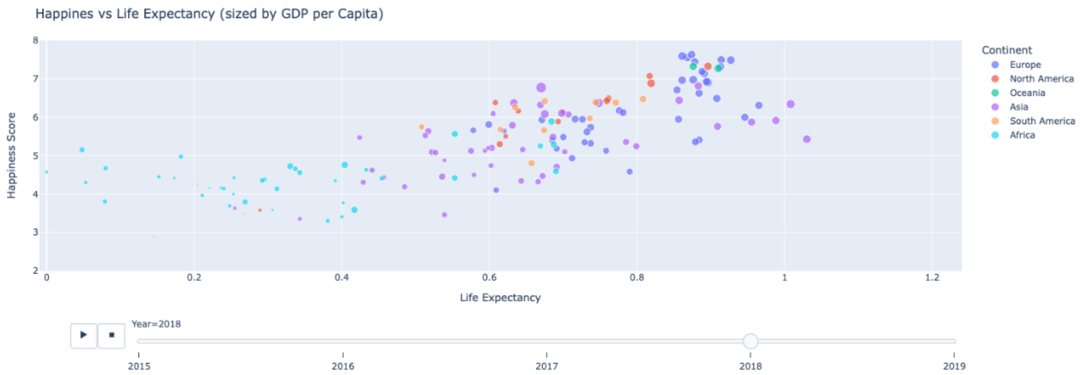
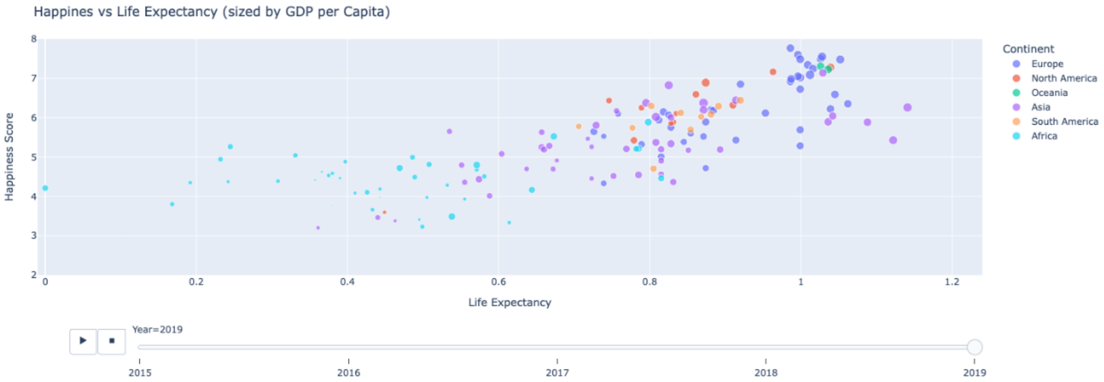


Figure 23: Happiness vs. Life Expectancy 2019



The analysis reveals a consistent trend where happiness scores correlate closely with life expectancy, similar to their correlation with GDP per capita. Notably, European countries, alongside Canada, New Zealand, and Australia, register the highest scores. This pattern suggests that an increase in GDP per capita significantly contributes to improvements in life expectancy.

3.4.9 Happiness vs. Generosity

In order to apply the same ideology presented above, through the use of scatter plot I recreate the link, this time taking into account two other variables, in this case between happiness score and generosity between the same years, but still sized by GPD per capita. The results obtained were as follows:

Figure 24: Happiness vs. Generosity 2015

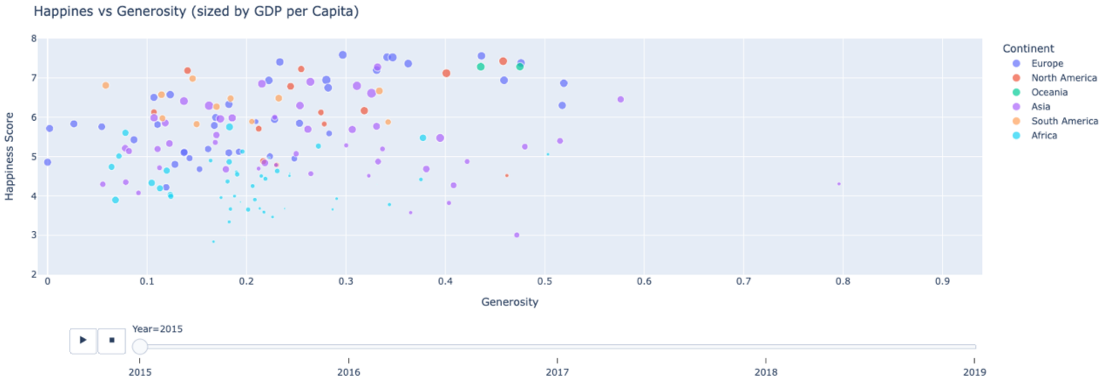


Figure 25: Happiness vs. Generosity 2016

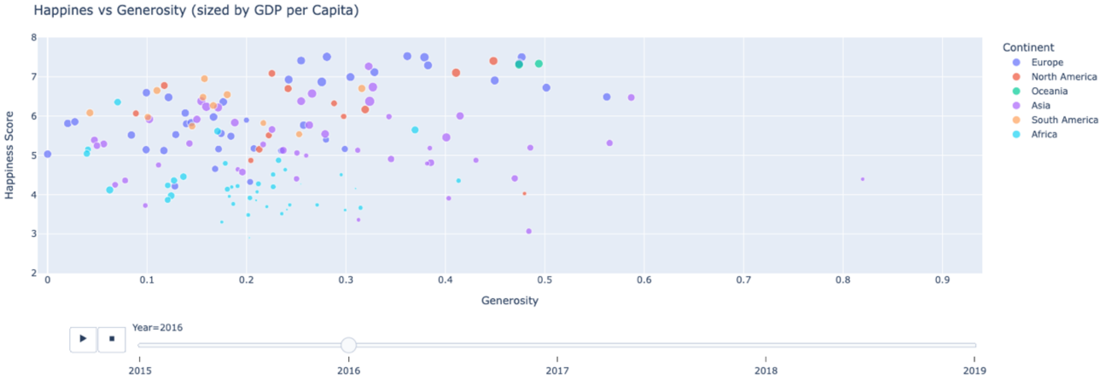


Figure 26: Happiness vs. Generosity 2017

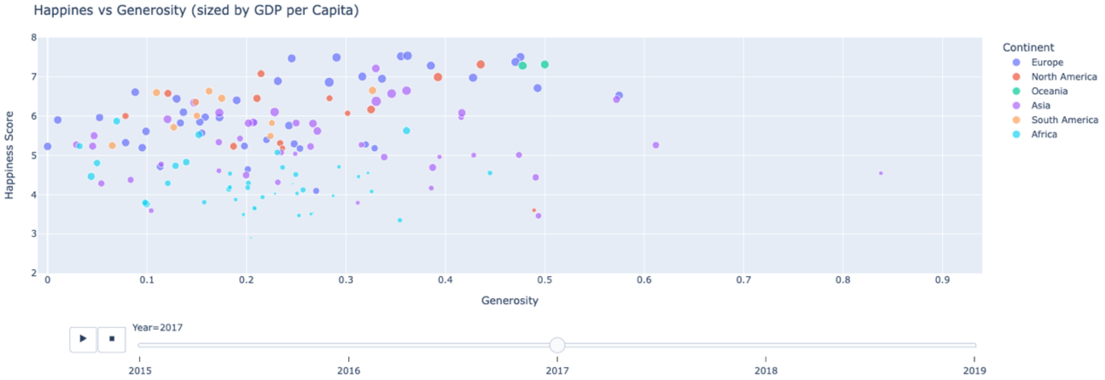


Figure 27: Happiness vs. Generosity 2018

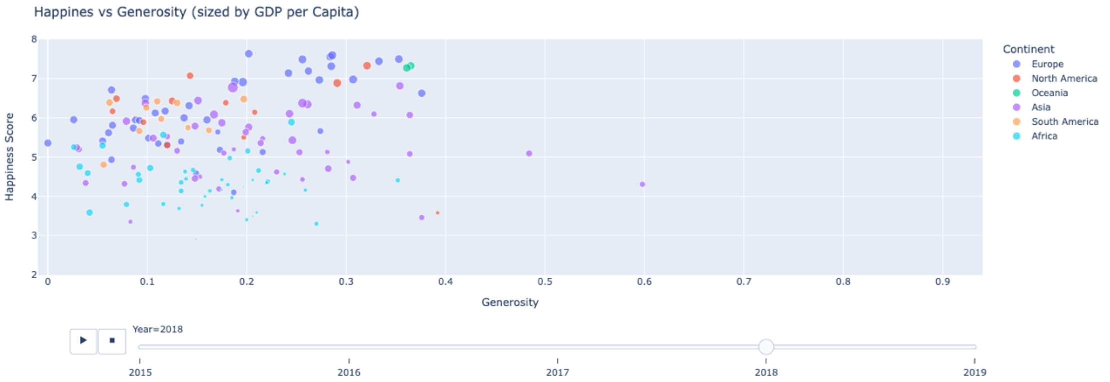
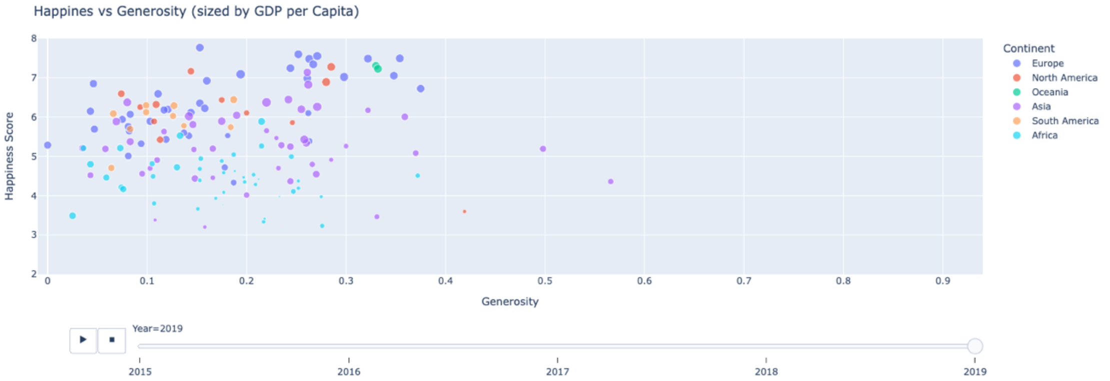


Figure 28: Happiness vs. Generosity 2019



Reflecting on the analyses conducted, it becomes clear that a notable trend emerges, as happiness scores and GDP per capita rise, so does the level of generosity observed within populations. This pattern suggests that higher economic prosperity and greater individual well-being are associated with increased altruistic behavior. This relationship underscores the interconnectedness of economic factors, subjective well-being, and social values such as generosity, highlighting the complex dynamics that contribute to societal happiness and prosperity.

3.4.10 Happiness vs. Corruption

To complete this analysis of these four scenarios, using the scatter plot I recreated a link, this time considering two other variables, in this case between happiness score and corruption, but still sized to GPD per capita. So, the countries with less corruption score are happier as we know. To explain the following graphs, the less corrupt a country is score will be close to 1, and the least corrupt I got was 0.6, and in case the score is close to 0 (zero) it follows that there is a high level of corruption.

Figure 29: Happiness vs. Corruption 2015

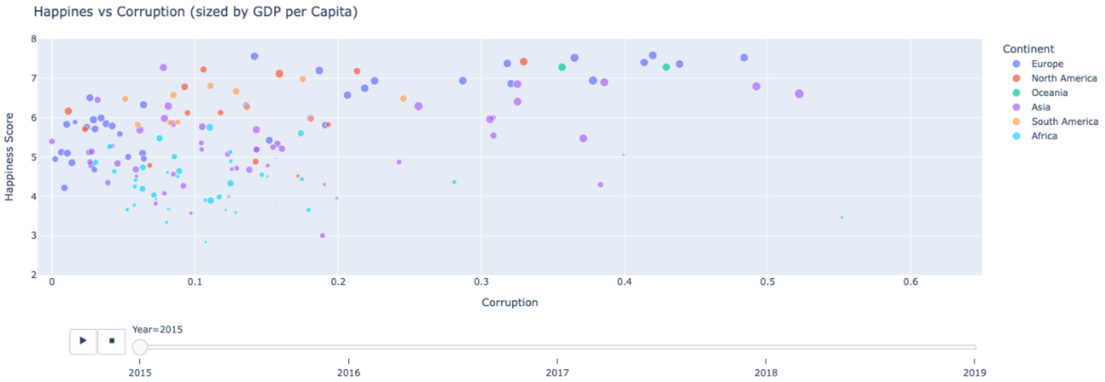


Figure 30: Happiness vs. Corruption 2016

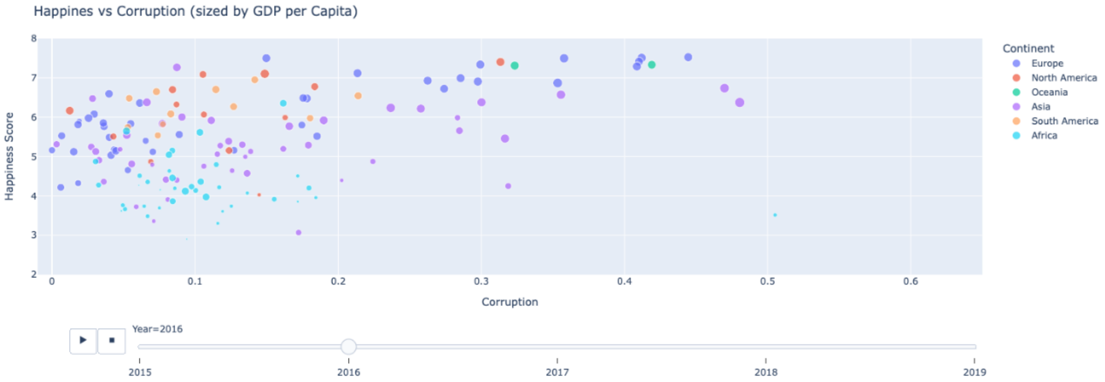


Figure 31: Happiness vs. Corruption 2017

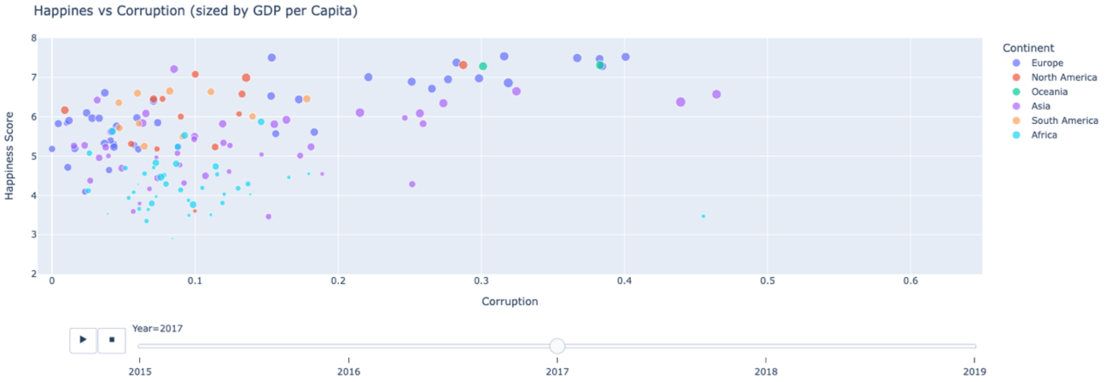


Figure 32: Happiness vs. Corruption 2018

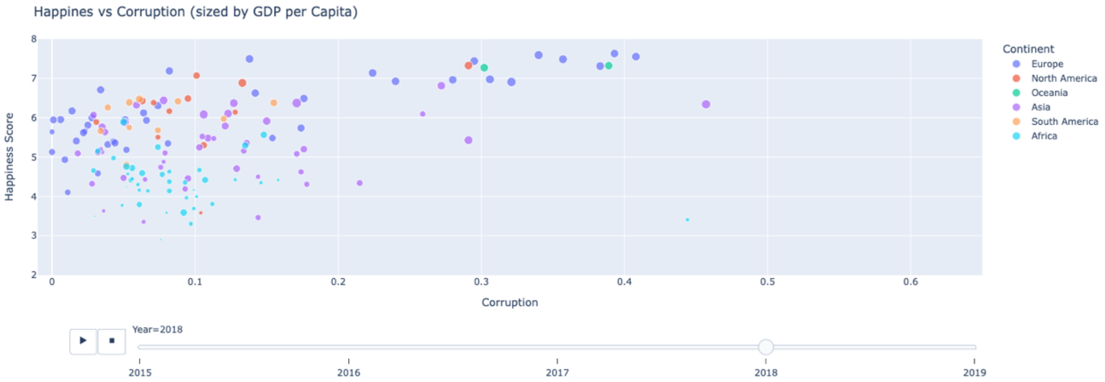
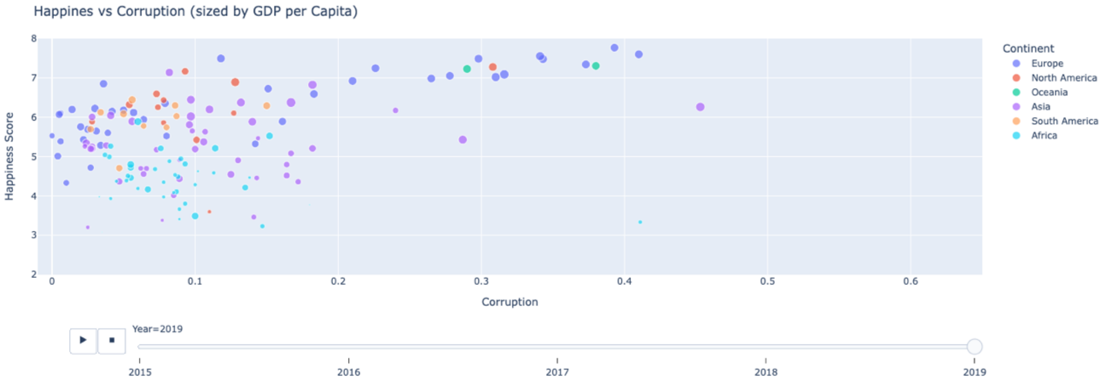


Figure 33: Happiness vs. Corruption 2019



Upon examining the data from 2015 to 2019, it becomes apparent that Eastern European nations exhibit higher levels of corruption yet maintain happiness scores above 4. In contrast, African and Asian countries also display elevated corruption levels but with significantly lower happiness scores when compared to their European counterparts.

Conversely, Northern European countries, including Denmark and Sweden, consistently show high happiness scores coupled with minimal corruption levels. This pattern is similarly observed in nations such as Singapore and Qatar, indicating a correlation between low corruption and high happiness scores across diverse geographical regions.

3.5. Machine Learning Application

At this stage, I will initiate the training of machine learning models, focusing on four key hypotheses. This phase aims to apply predictive analytics to uncover hidden patterns and insights, ensuring the models are well-selected, effectively trained, and accurately evaluated to support informed decision-making. The key objectives are:

1. Region can be predicted using the features from the world happiness report data;

2. Happiness Score can be predicted using the features from the world happiness report whereas Economy will affect it the most;

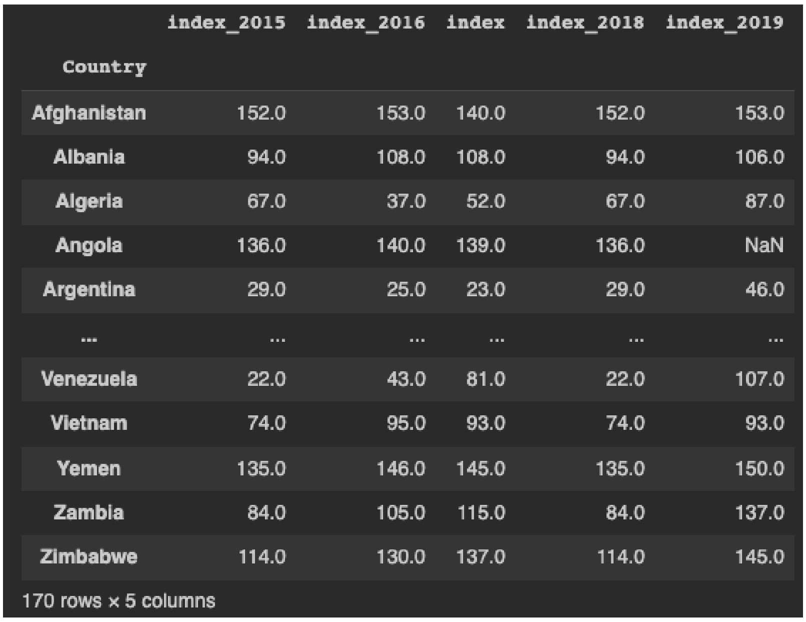
3. Predict an indicator (score) through others;

4. Predict Corruption/Trust between countries considering other indicators.

3.5.1 Machine Learning Data Cleaning

First, I sorted the data frame and then I set the country as the index so it's possible to have the data frame to train the machine learning models.

Figure 34: Setting the Country as Index

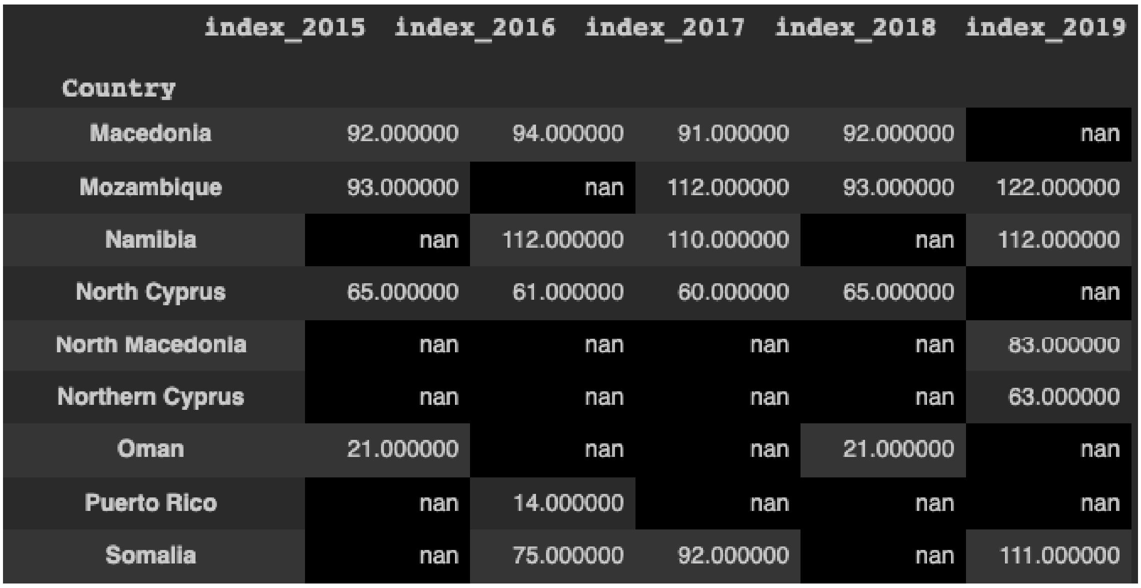


	index_2015	index_2016	index	index_2018	index_2019
Country					
Afghanistan	152.0	153.0	140.0	152.0	153.0
Albania	94.0	108.0	108.0	94.0	106.0
Algeria	67.0	37.0	52.0	67.0	87.0
Angola	136.0	140.0	139.0	136.0	NaN
Argentina	29.0	25.0	23.0	29.0	46.0
...	...	...	...	...	...
Venezuela	22.0	43.0	81.0	22.0	107.0
Vietnam	74.0	95.0	93.0	74.0	93.0
Yemen	135.0	146.0	145.0	135.0	150.0
Zambia	84.0	105.0	115.0	84.0	137.0
Zimbabwe	114.0	130.0	137.0	114.0	145.0

170 rows × 5 columns

There is a possibility that some countries exist in 2015 and do not exist in 2016, 2017, 2018 and 2019. To solve that problem, I highlighted the missing values in black, to identify them.

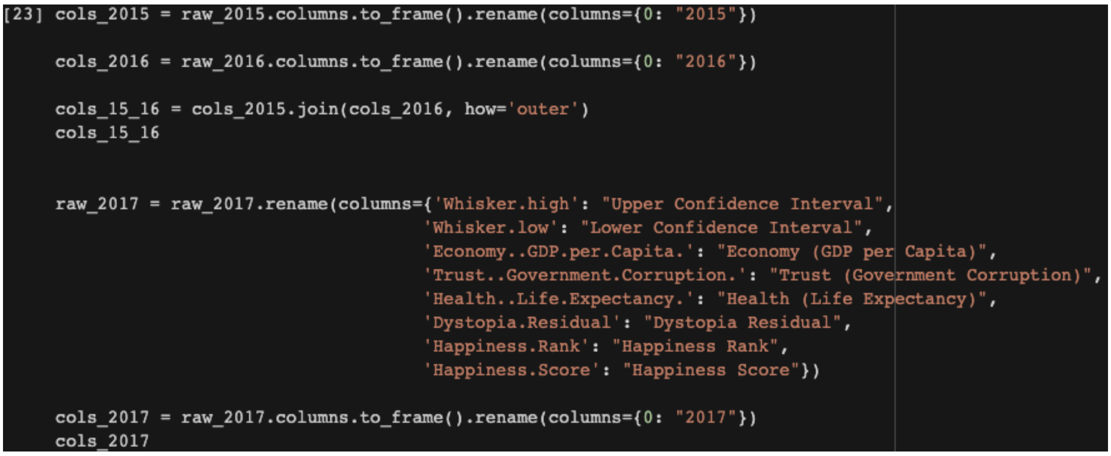
Table 13: Missing Countries throughout the years



Country	index_2015	index_2016	index_2017	index_2018	index_2019
Macedonia	92.000000	94.000000	91.000000	92.000000	nan
Mozambique	93.000000	nan	112.000000	93.000000	122.000000
Namibia	nan	112.000000	110.000000	nan	112.000000
North Cyprus	65.000000	61.000000	60.000000	65.000000	nan
North Macedonia	nan	nan	nan	nan	83.000000
Northern Cyprus	nan	nan	nan	nan	63.000000
Oman	21.000000	nan	nan	21.000000	nan
Puerto Rico	nan	14.000000	nan	nan	nan
Somalia	nan	75.000000	92.000000	nan	111.000000

Table 13, there are several missing countries in our time span between 2015 and 2019, for example, Puerto Rico only appears in the year 2016. To fill the missing values, I need to do a unification of columns and append all years, for this, I did something similar as was done in the visualization. For this I will convert the variable names so that the variable names are the same for the different years. The data from 2015 to 2016 follow the same sequence of variables, but as I saw earlier, the same does not happen for 2017, 2018 and 2019.

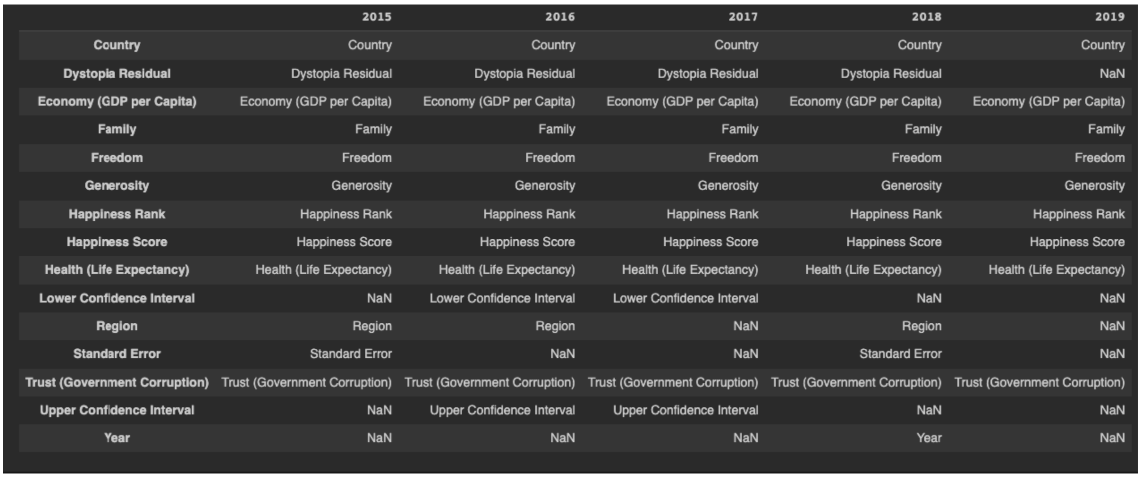
Figure 35: Rename the features for data from 2017 to 2019



```
[23] cols_2015 = raw_2015.columns.to_frame().rename(columns={0: "2015"})

     cols_2016 = raw_2016.columns.to_frame().rename(columns={0: "2016"})

     cols_15_16 = cols_2015.join(cols_2016, how='outer')
     cols_15_16


     raw_2017 = raw_2017.rename(columns={'Whisker.high': "Upper Confidence Interval",
                                   'Whisker.low': "Lower Confidence Interval",
                                   'Economy..GDP.per.Capita.': "Economy (GDP per Capita)",
                                   'Trust..Government.Corruption.': "Trust (Government Corruption)",
                                   'Health..Life.Expectancy.': "Health (Life Expectancy)",
                                   'Dystopia.Residual': "Dystopia Residual",
                                   'Happiness.Rank': "Happiness Rank",
                                   'Happiness.Score': "Happiness Score"})

     cols_2017 = raw_2017.columns.to_frame().rename(columns={0: "2017"})
     cols_2017
```

Upon completing the unification process, I now have a comprehensive overview of the columns that are missing across the years 2015 to 2019.

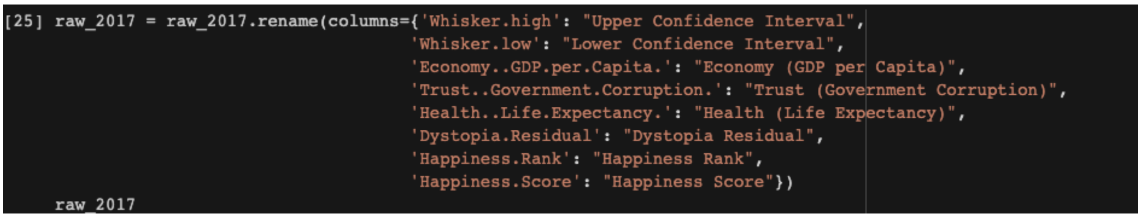
Figure 36: Missing columns



	2015	2016	2017	2018	2019
Country	Country	Country	Country	Country	Country
Dystopia Residual	Dystopia Residual	Dystopia Residual	Dystopia Residual	Dystopia Residual	NaN
Economy (GDP per Capita)	Economy (GDP per Capita)	Economy (GDP per Capita)	Economy (GDP per Capita)	Economy (GDP per Capita)	Economy (GDP per Capita)
Family	Family	Family	Family	Family	Family
Freedom	Freedom	Freedom	Freedom	Freedom	Freedom
Generosity	Generosity	Generosity	Generosity	Generosity	Generosity
Happiness Rank	Happiness Rank	Happiness Rank	Happiness Rank	Happiness Rank	Happiness Rank
Happiness Score	Happiness Score	Happiness Score	Happiness Score	Happiness Score	Happiness Score
Health (Life Expectancy)	Health (Life Expectancy)	Health (Life Expectancy)	Health (Life Expectancy)	Health (Life Expectancy)	Health (Life Expectancy)
Lower Confidence Interval	NaN	Lower Confidence Interval	Lower Confidence Interval	NaN	NaN
Region	Region	Region	NaN	Region	NaN
Standard Error	Standard Error	NaN	NaN	Standard Error	NaN
Trust (Government Corruption)	Trust (Government Corruption)	Trust (Government Corruption)	Trust (Government Corruption)	Trust (Government Corruption)	Trust (Government Corruption)
Upper Confidence Interval	NaN	Upper Confidence Interval	Upper Confidence Interval	NaN	NaN
Year	NaN	NaN	NaN	Year	NaN

The features may vary depending on the years, as I realized, there are still missing columns for a few years, so I will proceed with the same process which consist in the renaming the variable names, as was done for the year 2017.

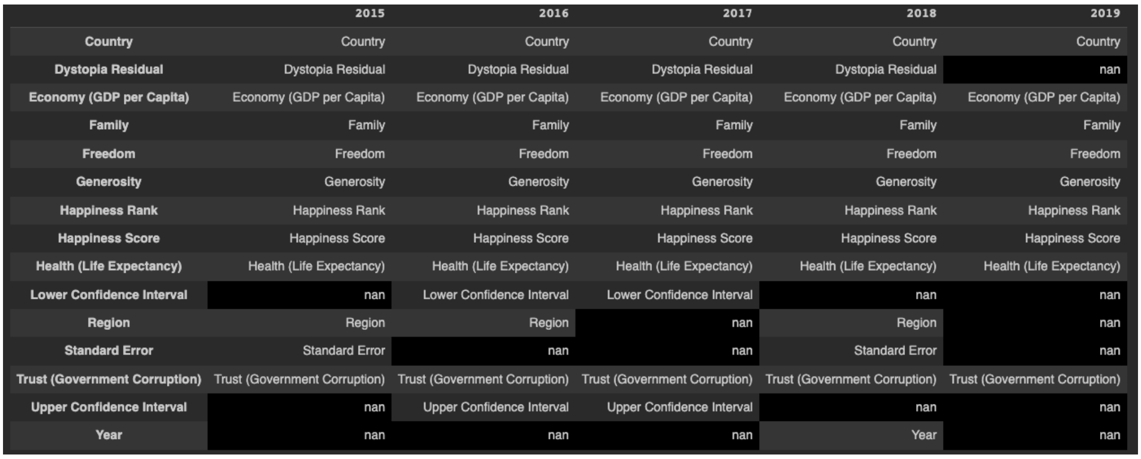
Figure 37: Rename the features columns



```
[25] raw_2017 = raw_2017.rename(columns={'Whisker.high': "Upper Confidence Interval",
                                        'Whisker.low': "Lower Confidence Interval",
                                        'Economy..GDP.per.Capita.': "Economy (GDP per Capita)",
                                        'Trust..Government.Corruption.': "Trust (Government Corruption)",
                                        'Health..Life.Expectancy.': "Health (Life Expectancy)",
                                        'Dystopia.Residual': "Dystopia Residual",
                                        'Happiness.Rank': "Happiness Rank",
                                        'Happiness.Score': "Happiness Score"})
     raw_2017
```

With the labels now standardized, the remaining issue is the identification of missing columns, which are currently highlighted in black for clarity.

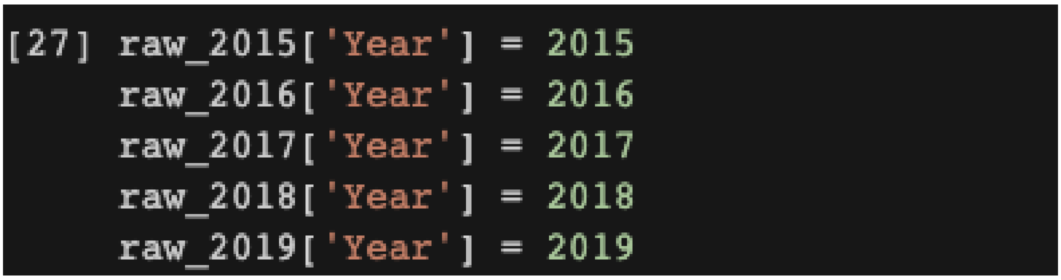
Figure 38: Result of the rename features column



	2015	2016	2017	2018	2019
Country	Country	Country	Country	Country	Country
Dystopia Residual	Dystopia Residual	Dystopia Residual	Dystopia Residual	Dystopia Residual	nan
Economy (GDP per Capita)	Economy (GDP per Capita)	Economy (GDP per Capita)	Economy (GDP per Capita)	Economy (GDP per Capita)	Economy (GDP per Capita)
Family	Family	Family	Family	Family	Family
Freedom	Freedom	Freedom	Freedom	Freedom	Freedom
Generosity	Generosity	Generosity	Generosity	Generosity	Generosity
Happiness Rank	Happiness Rank	Happiness Rank	Happiness Rank	Happiness Rank	Happiness Rank
Happiness Score	Happiness Score	Happiness Score	Happiness Score	Happiness Score	Happiness Score
Health (Life Expectancy)	Health (Life Expectancy)	Health (Life Expectancy)	Health (Life Expectancy)	Health (Life Expectancy)	Health (Life Expectancy)
Lower Confidence Interval	nan	Lower Confidence Interval	Lower Confidence Interval	nan	nan
Region	Region	Region	nan	Region	nan
Standard Error	Standard Error	nan	nan	Standard Error	nan
Trust (Government Corruption)	Trust (Government Corruption)	Trust (Government Corruption)	Trust (Government Corruption)	Trust (Government Corruption)	Trust (Government Corruption)
Upper Confidence Interval	nan	Upper Confidence Interval	Upper Confidence Interval	nan	nan
Year	nan	nan	nan	Year	nan

Following the information in table 15, the “Year” feature is missing, so it will have to be inserted.

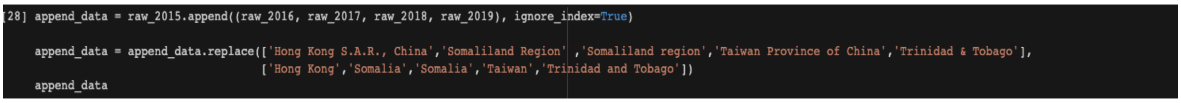
Figure 39: Insert year column in all data frames



```
[27] raw_2015['Year'] = 2015
     raw_2016['Year'] = 2016
     raw_2017['Year'] = 2017
     raw_2018['Year'] = 2018
     raw_2019['Year'] = 2019
```

Another step to consider before starting to train the models is the unification of the names of the countries, as some of the countries appear with different names, for example, Taiwan Province of China, let's rewrite the only Taiwan. The purpose of this action will correct the missing values in the country’s column.

Figure 40: Correction of countries missing values

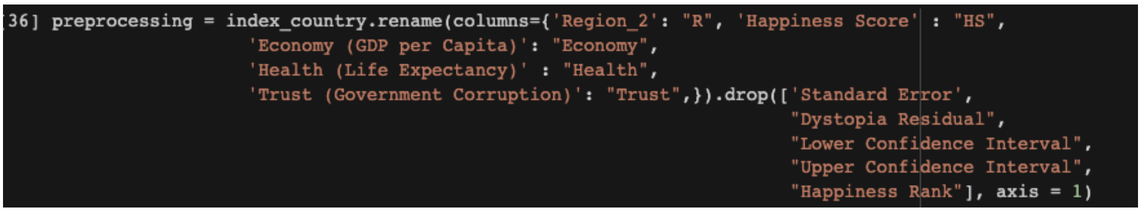


```
[28] append_data = raw_2015.append((raw_2016, raw_2017, raw_2018, raw_2019), ignore_index=True)

     append_data = append_data.replace(['Hong Kong S.A.R., China','Somaliland Region' ,'Somaliland region','Taiwan Province of China','Trinidad & Tobago'],
                                       ['Hong Kong','Somalia','Somalia','Taiwan','Trinidad and Tobago'])
     append_data
```

The last action prior to initiating the training of machine learning models is the renaming of certain features that will be utilized. The changes implemented are detailed below:

- Happiness Score = HS
- Economy (GPD per Capita = Economy
- Health (Life Expectancy) = Health
- Trust (Government Corruption) = Trust

Figure 41: Rename the features



```
36] preprocessing = index_country.rename(columns={'Region_2': "R", 'Happiness Score' : "HS",
                       'Economy (GDP per Capita)': "Economy",
                       'Health (Life Expectancy)' : "Health",
                       'Trust (Government Corruption)': "Trust",}).drop(['Standard Error',
                                                                         "Dystopia Residual",
                                                                         "Lower Confidence Interval",
                                                                         "Upper Confidence Interval",
                                                                         "Happiness Rank"], axis = 1)
```

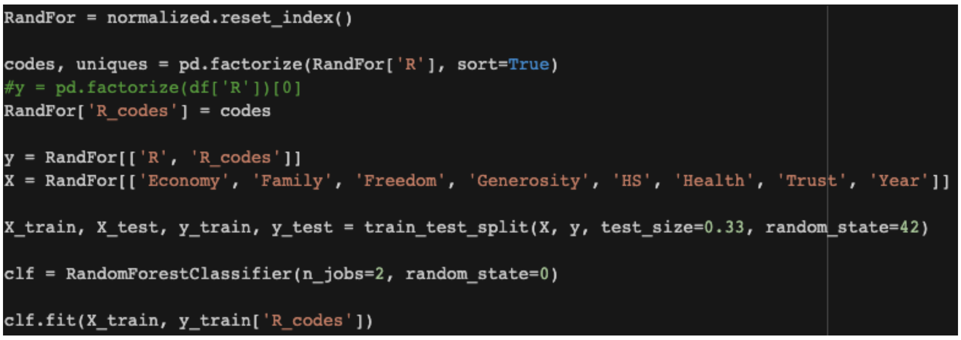
As highlighted previously, not all variables are consistent across each year from 2015 to 2019, leading to the introduction of selection criteria for analysis. This necessitated the exclusion of certain variables for the training and testing of machine learning models, as depicted in Figure 49.

To address the four machine learning objectives, a new dataframe data_df was created by duplicating the preprocessing dataset. The dataset was then divided into training and testing subsets. Regional data was assigned to variable "R", while the selected features for model training were stored in variable "x". Additionally, the MinMaxScaler was employed to standardize the dataset, ensuring that our models are trained on data with uniform scaling.

3.5.2 Hypothesis 1 - Random Forest Classifier

To be able to train and test this model, I will first store my regions in “y” variable, and the features of the data set economy, family, freedom, generosity, hs, health, trust, and year in I’ll store in “x” variable. The second step is split the data set, the test size is equal to 0.33 which means that I’m storing 33% of the data to the “y_test” and “x_test”, reaming approximately 67% of the data goes to “y_train” and “x_train”. The random state of 0.42 will make sure that split remains constant every time I run the notebook. This split of date will be used to test the four machine learning models. There are 3 steps to implement a machine learning model, model initiation, train the model and predict the model.

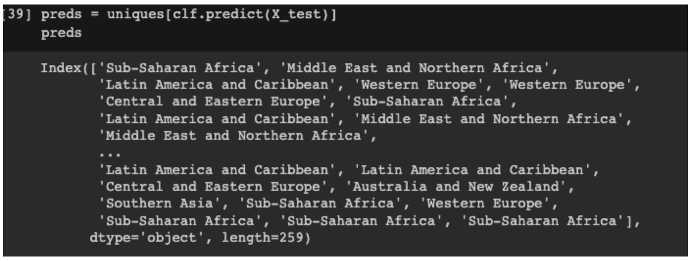
Figure 42: Implementation of random forest model



```
RandFor = normalized.reset_index()

codes, uniques = pd.factorize(RandFor['R'], sort=True)
#y = pd.factorize(df['R'])[0]
RandFor['R_codes'] = codes

y = RandFor[['R', 'R_codes']]
X = RandFor[['Economy', 'Family', 'Freedom', 'Generosity', 'HS', 'Health', 'Trust', 'Year']]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)

clf = RandomForestClassifier(n_jobs=2, random_state=0)

clf.fit(X_train, y_train['R_codes'])
```

To get predictions, I´m using clf.predict on my "x_test", in order to predict the following regions:

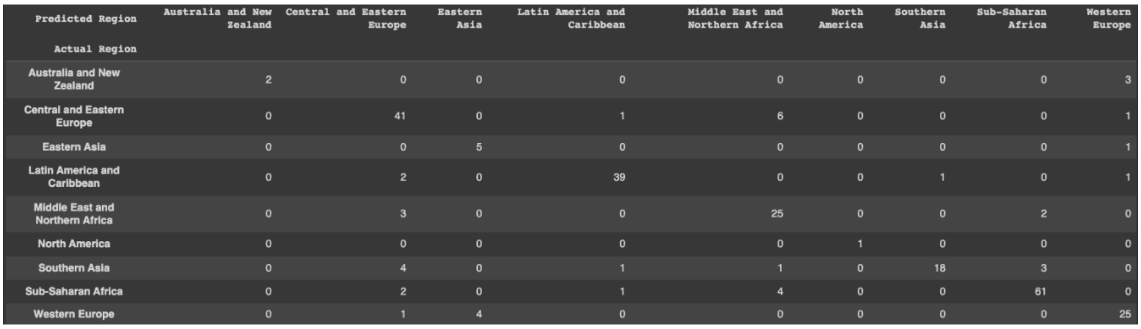
Figure 43: Clf.predict on x_test



```
[39] preds = uniques[clf.predict(X_test)]
     preds

     Index(['Sub-Saharan Africa', 'Middle East and Northern Africa',
            'Latin America and Caribbean', 'Western Europe', 'Western Europe',
            'Central and Eastern Europe', 'Sub-Saharan Africa',
            'Latin America and Caribbean', 'Middle East and Northern Africa',
            'Middle East and Northern Africa',
            ...
            'Latin America and Caribbean', 'Latin America and Caribbean',
            'Central and Eastern Europe', 'Australia and New Zealand',
            'Southern Asia', 'Sub-Saharan Africa', 'Western Europe',
            'Sub-Saharan Africa', 'Sub-Saharan Africa', 'Sub-Saharan Africa'],
           dtype='object', length=259)
```

On "x_test" I have the predictions, but originally, I had the "x_test" against "y_test", now it's possible to compare the "y_test" with the predictions I made and see how many correct predictions I have and wrong predictions.

Figure 44: Confusion matrix



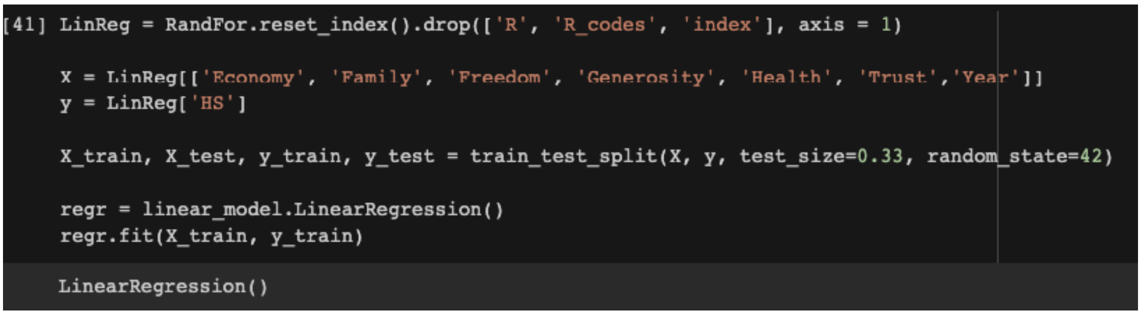
Predicted Region	Australia and New Zealand	Central and Eastern Europe	Eastern Asia	Latin America and Caribbean	Middle East and Northern Africa	North America	Southern Asia	Sub-Saharan Africa	Western Europe
Actual Region									
Australia and New Zealand	2	0	0	0	0	0	0	0	3
Central and Eastern Europe	0	41	0	1	6	0	0	0	1
Eastern Asia	0	0	5	0	0	0	0	0	1
Latin America and Caribbean	0	2	0	39	0	0	1	0	1
Middle East and Northern Africa	0	3	0	0	25	0	0	2	0
North America	0	0	0	0	0	1	0	0	0
Southern Asia	0	4	0	1	1	0	18	3	0
Sub-Saharan Africa	0	2	0	1	4	0	0	61	0
Western Europe	0	1	4	0	0	0	0	0	25

As it can be seen from above confusion matrix, the accuracy of the predictions are great and actual regions have a good prediction with the random forest, for example, out of 49 times this model correctly predicted 41times in central and eastern Europe, which proves the hypothesis 1, that regions can be predicted using this data from world happiness report.

3.5.3 Hypothesis 2 - Extreme Boost Regressor

In the upcoming chapter, I'll explore the second hypothesis, in order to address this, I'll utilize the dataset prepared for the Random Forest model. This involves removing the regional columns to concentrate solely on the features, aiming to understand the influence of economic factors on happiness scores.

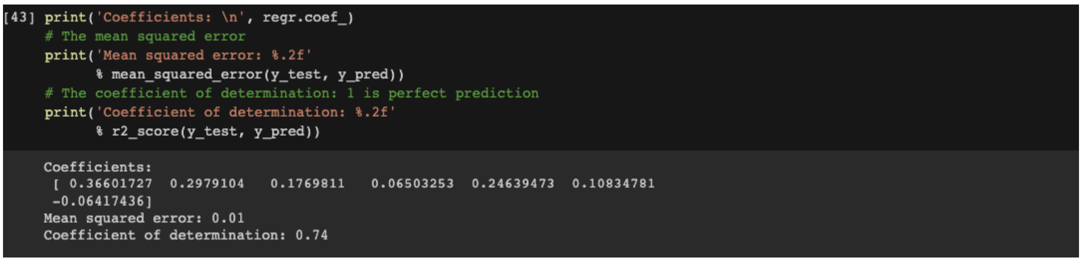
Figure 45: Training linear regression



```
[41] LinReg = RandFor.reset_index().drop(['R', 'R_codes', 'index'], axis = 1)

     X = LinReg[['Economy', 'Family', 'Freedom', 'Generosity', 'Health', 'Trust','Year']]
     y = LinReg['HS']

     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)

     regr = linear_model.LinearRegression()
     regr.fit(X_train, y_train)

     LinearRegression()
```

The following step involves generating predictions on the training dataset to obtain the coefficients of the linear regression model. Unlike in classification models where a confusion matrix might be used, as was the case with the Random Forest model, regression models require different metrics for evaluation. Here, I calculated both the mean squared error (MSE) and the coefficient of determination (R² score) to assess model performance. A lower MSE indicates a model that better fits the data, while a higher R² score suggests a higher proportion of variance explained by the model, both of which are indicative of a more effective machine learning model.

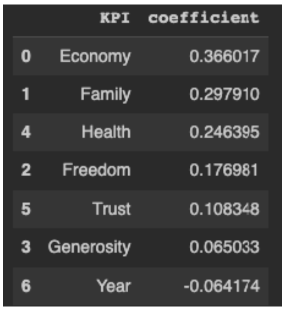
Figure 46: Linear regression evaluation



```
[43] print('Coefficients: \n', regr.coef_)
     # The mean squared error
     print('Mean squared error: %.2f'
           % mean_squared_error(y_test, y_pred))
     # The coefficient of determination: 1 is perfect prediction
     print('Coefficient of determination: %.2f'
           % r2_score(y_test, y_pred))

     Coefficients:
      [ 0.36601727  0.2979104   0.1769811   0.06503253  0.24639473  0.10834781
      -0.06417436]
     Mean squared error: 0.01
     Coefficient of determination: 0.74
```

This evaluation shows a mean square error close to zero and coefficient of determination with 0.74, which means, this is a good model to answer the hypothesis 2. Finally, the future importance, I grouped the features with the coefficient from the evaluation carried out above, making it possible to observe in order the variables that most affect the happiness score.

Table 14: KPI coefficient 1



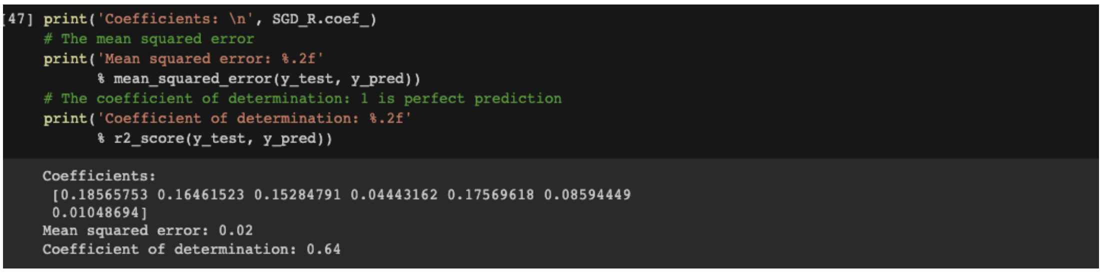
	KPI	coefficient
0	Economy	0.366017
1	Family	0.297910
4	Health	0.246395
2	Freedom	0.176981
5	Trust	0.108348
3	Generosity	0.065033
6	Year	-0.064174

From above linear regression model, happiness score can indeed be predicted using the available features and as in this hypothesis 2, table 14 shows from the above KPI that economy does affect the happiness score the most which proves the hypothesis.

3.5.4 Hypothesis 2 - Stochastic Gradient Regressor

Following the same steps performed in the evaluation of the previous regression model, where the model was trained, predicted, and performed, the same steps will be applied to this and the remaining two regression models. For this reason, I'm going to show the evaluation and feature importance, to be able to compare the machine learning models.

Figure 47: Stochastic gradient regressor evaluation

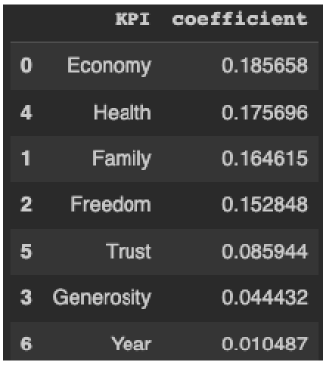


```
[47] print('Coefficients: \n', SGD_R.coef_)
     # The mean squared error
     print('Mean squared error: %.2f'
           % mean_squared_error(y_test, y_pred))
     # The coefficient of determination: 1 is perfect prediction
     print('Coefficient of determination: %.2f'
           % r2_score(y_test, y_pred))
```

```
Coefficients: 
 [0.18565753 0.16461523 0.15284791 0.04443162 0.17569618 0.08594449
 0.01048694]
Mean squared error: 0.02
Coefficient of determination: 0.64
```

This evaluation shows a mean square error close to zero and coefficient of determination with 0.64, which means, this is still a good model to answer the hypothesis 2.

Table 15: KPI coefficient 2

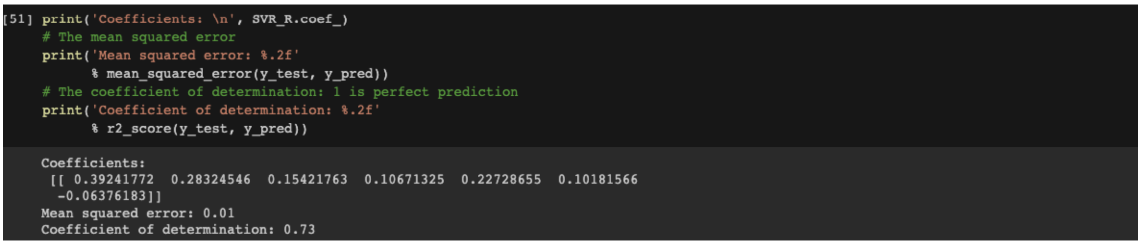


	KPI	coefficient
0	Economy	0.185658
4	Health	0.175696
1	Family	0.164615
2	Freedom	0.152848
5	Trust	0.085944
3	Generosity	0.044432
6	Year	0.010487

This time the result takes coefficients with lower values than the previous model, the KPI economy remains as one of the factors that most influences the happiness score, which proves hypothesis 2.

3.5.5 Hypothesis 2 - Support Vector Regressor

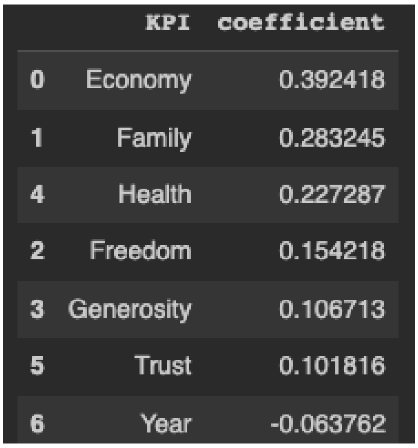
Figure 48: Support vector regressor evaluation



```
[51] print('Coefficients: \n', SVR_R.coef_)
     # The mean squared error
     print('Mean squared error: %.2f'
           % mean_squared_error(y_test, y_pred))
     # The coefficient of determination: 1 is perfect prediction
     print('Coefficient of determination: %.2f'
           % r2_score(y_test, y_pred))

     Coefficients:
      [[ 0.39241772  0.28324546  0.15421763  0.10671325  0.22728655  0.10181566
       -0.06376183]]
     Mean squared error: 0.01
     Coefficient of determination: 0.73
```

Evaluation in this model shows a mean square closer to zero and coefficient of determination with 0.73 which means, this is a good model to answer the hypothesis 2.

Table 16: KPI coefficient 3

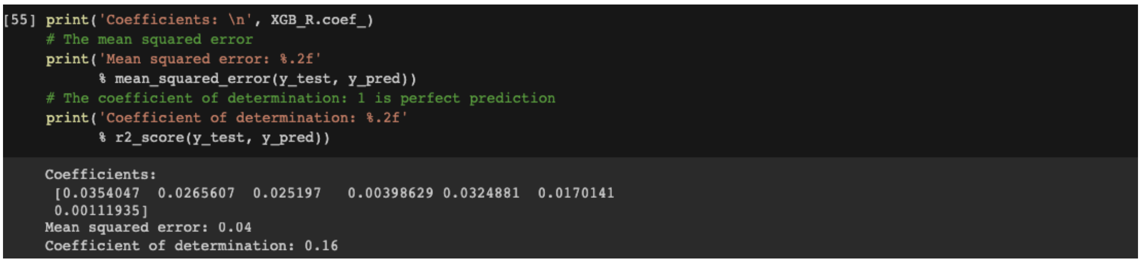


	KPI	coefficient
0	Economy	0.392418
1	Family	0.283245
4	Health	0.227287
2	Freedom	0.154218
3	Generosity	0.106713
5	Trust	0.101816
6	Year	-0.063762

This model shows the highest coefficient regarding the economy, the KPI economy remains one of the factors that most influences the happiness score, which proves hypothesis 2.

3.5.6 Hypothesis 2 - Extreme Gradient Boost Regressor

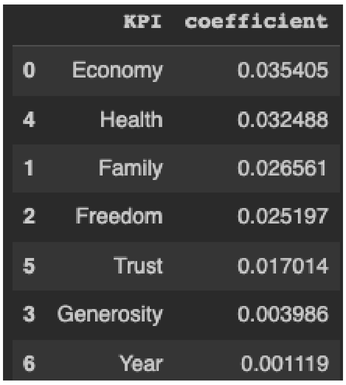
Figure 49: Extreme gradient boost regressor evaluation



```
[55] print('Coefficients: \n', XGB_R.coef_)
     # The mean squared error
     print('Mean squared error: %.2f'
           % mean_squared_error(y_test, y_pred))
     # The coefficient of determination: 1 is perfect prediction
     print('Coefficient of determination: %.2f'
           % r2_score(y_test, y_pred))

     Coefficients:
      [0.0354047  0.0265607  0.025197   0.00398629 0.0324881  0.0170141
      0.00111935]
     Mean squared error: 0.04
     Coefficient of determination: 0.16
```

For this regressor mode, evaluation shows the mean square error closer to 0.5 and coefficient of determination with 0.16, which means, this is not good model to answer the hypothesis 2. Compared to the other 3 models, this one has the lowest accuracy.

Table 17: KPI coefficient 4



	KPI	coefficient
0	Economy	0.035405
4	Health	0.032488
1	Family	0.026561
2	Freedom	0.025197
5	Trust	0.017014
3	Generosity	0.003986
6	Year	0.001119

The results show that the KPI economy continues to be the factor that most influences the happiness score, even using a worse model than the other three. Even so, it is noted that the health and family variables are the only ones that fluctuate throughout the training of these four models, while the others show almost a constant. Hypothesis 2 was confirmed in the application of these four regression models.

3.5.7 Hypothesis 3 - Ridge Regression and Lasso Regression

The aim is to predict health outcomes using a set of indicators through three regression techniques: Ridge Regression and Lasso Regression. These methods will help in understanding complex data relationships, managing multicollinearity, and identifying key health indicators. This approach is designed to yield accurate health outcome predictions and provide insights into the factors influencing these outcomes.

Figure 50: Feature Importance using Ridge Regressor

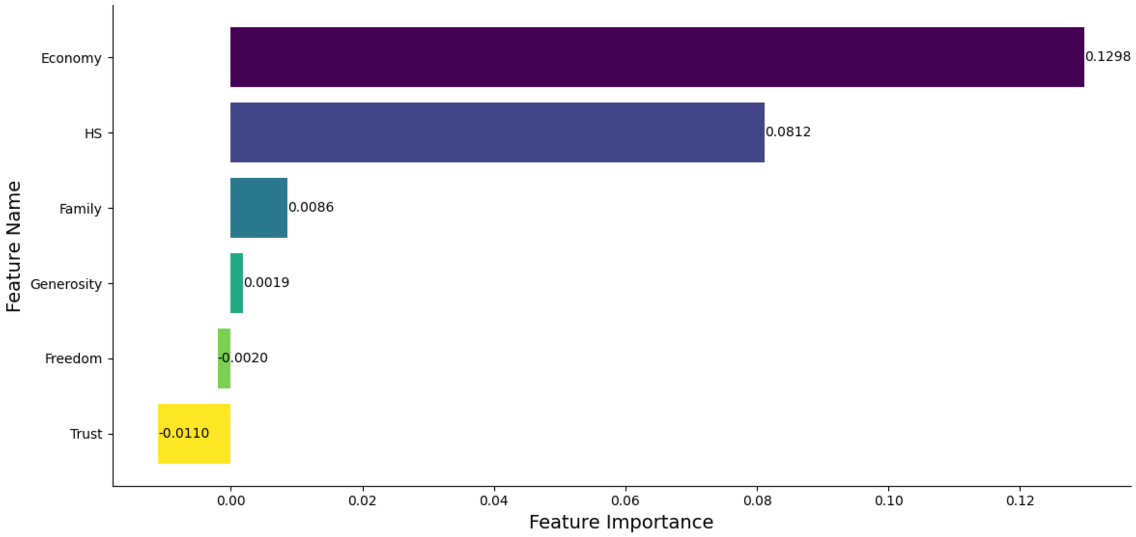
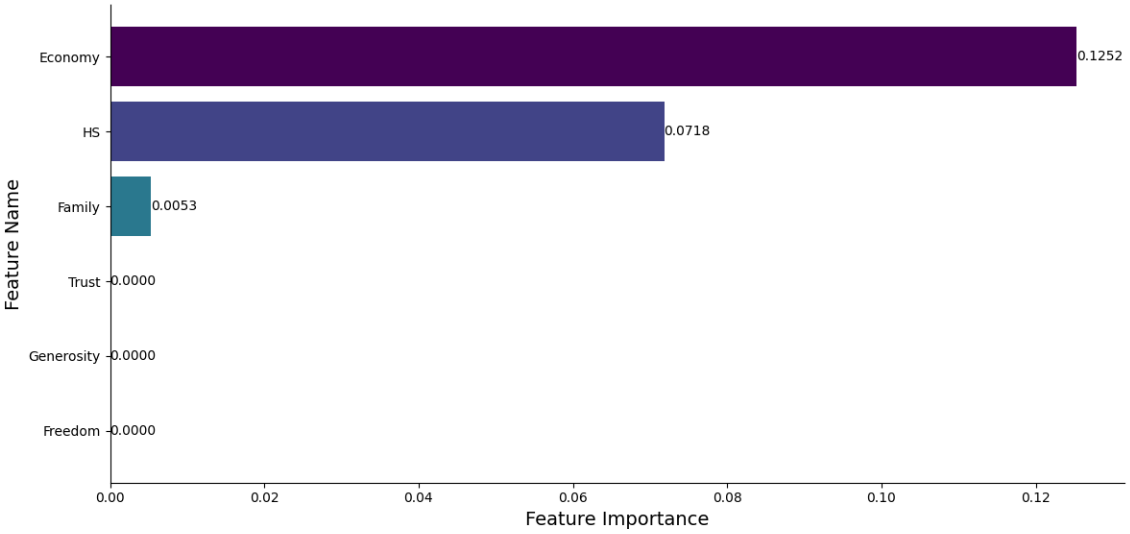


Figure 51: Feature Importance using Lasso Regression



Hypothesis 3 delves into the potential of predicting health outcomes through the strategic application of various indicators. This investigation primarily centers around the evaluation of model performance, aimed at forging a path towards insightful and accurate health outcome predictions.

The culmination of rigorous training and meticulous evaluation of regression models has yielded promising results, as quantified by the Mean Squared Error (MSE) metric:

- Ridge Regression: MSE = 0.0193
- Lasso Regression: MSE = 0.0196

The essence of these findings reveals a slight edge in performance for Ridge Regression model over Lasso Regression in the context of predicting health outcomes. This nuanced analysis provides a gateway into a deeper exploration of the results, acknowledging the complex interdependencies between indicators and the "Happiness score" as the focal point of this study.

The results not only showcase the nuanced capabilities of each regression model but also illuminate the path forward in harnessing the power of regression algorithms to predict health outcomes effectively. This journey through data analysis emphasizes the critical nature of model selection, tailored to the intricacies of the dataset at hand and aligned with the overarching analytical goals.

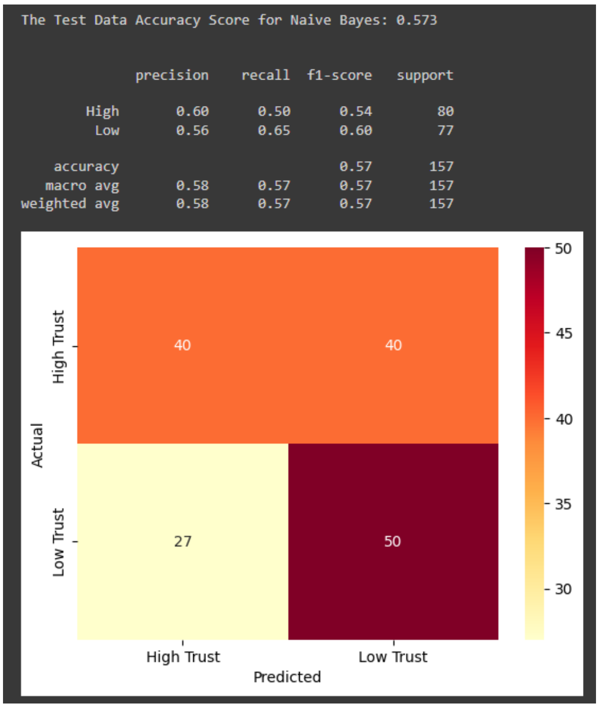
3.5.8 Hypothesis 4 - Naive Bayes Classifier, Logistic Regression and Decision Tree Classifier

Hypothesis 4 explores the prediction of countries' trust levels, specifically their perceived corruption, by transforming it into a binary classification task. By categorizing countries into "High" or "Low" trust categories based on the median value of the "Trust" column, this hypothesis aims to employ classification techniques to navigate the complexities of corruption perception.

To operationalize this, the "Explained by: Perceptions of corruption" metric will be dichotomized into binary or categorical terms. This simplification allows for the utilization of classification algorithms—Naive Bayes Classifier, Logistic Regression, and Decision Tree Classifier—to predict the newly defined corruption categories using various indicators.

The effectiveness of each classifier will be meticulously evaluated to determine the most accurate model in predicting trust levels among countries. This approach not only enhances our understanding of corruption perceptions across nations but also tests the efficacy of different classification models in handling binary outcomes based on a range of indicators.

Figure 52: Naive Bayes Confusion Matrix



```
The Test Data Accuracy Score for Naive Bayes: 0.573

              precision    recall  f1-score   support

        High       0.60      0.50      0.54        80
         Low       0.56      0.65      0.60        77

    accuracy                           0.57       157
   macro avg       0.58      0.57      0.57       157
weighted avg       0.58      0.57      0.57       157
```

Figure 53: Naive Bayes Feature Importances

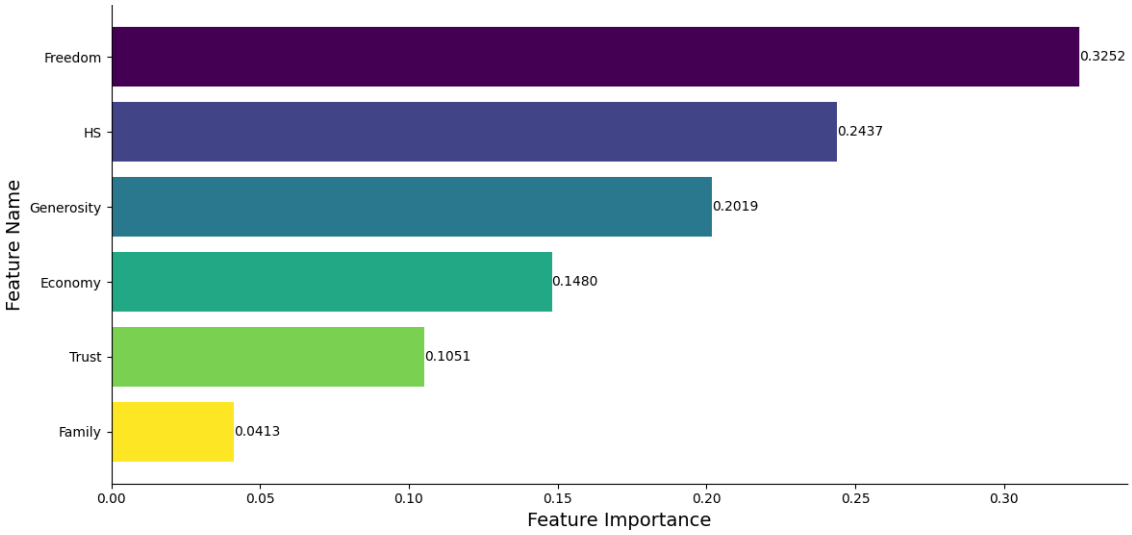
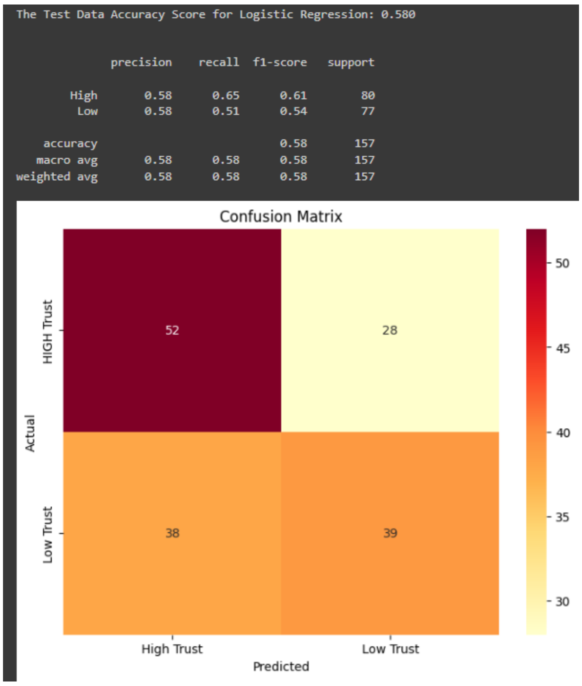


Figure 54: Logistic Regression Confusion Matrix



```
The Test Data Accuracy Score for Logistic Regression: 0.580

              precision    recall  f1-score   support

        High       0.58      0.65      0.61        80
         Low       0.58      0.51      0.54        77

    accuracy                           0.58       157
   macro avg       0.58      0.58      0.58       157
weighted avg       0.58      0.58      0.58       157
```

Figure 55: Logistic Regression Feature Importances

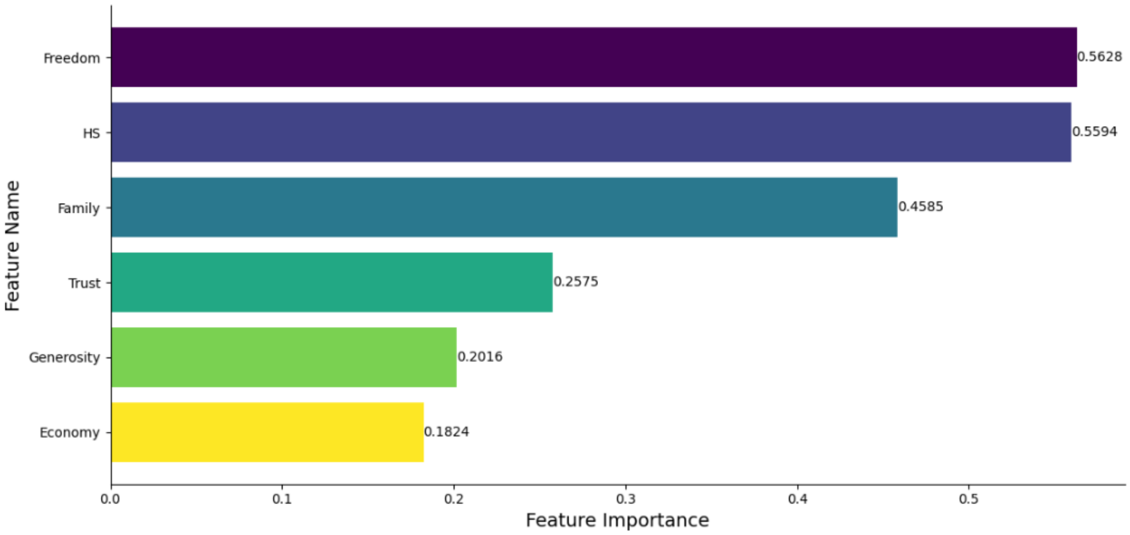
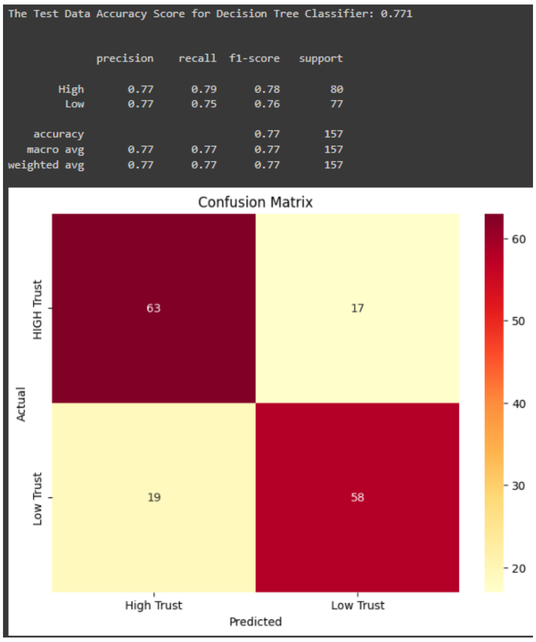
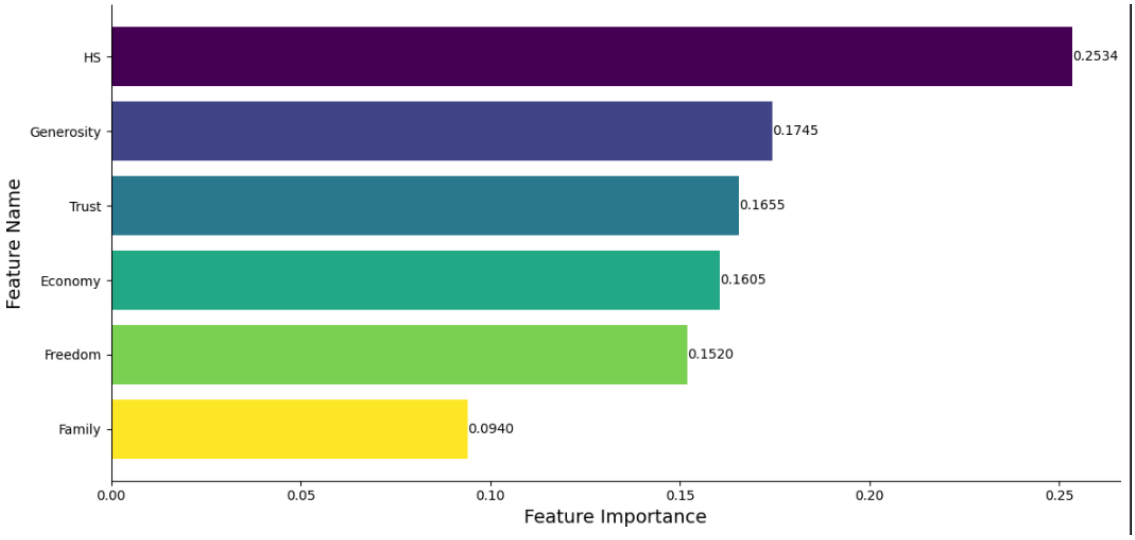


Figure 56: Decision Tree Classifier Confusion Matrix



```
The Test Data Accuracy Score for Decision Tree Classifier: 0.771

              precision    recall  f1-score   support

        High       0.77      0.79      0.78        80
         Low       0.77      0.75      0.76        77

    accuracy                           0.77       157
   macro avg       0.77      0.77      0.77       157
weighted avg       0.77      0.77      0.77       157
```

Figure 57: Decision Tree Feautures Importances



In the exploration of Hypothesis 4, the focus is on segmenting countries based on trust (or perceived corruption) into either "High" or "Low" categories, determined by the median of the "Trust" column. This segmentation transforms the analysis into a classification challenge, where the "Explained by: Perceptions of corruption" metric is recast into binary terms. To navigate this classification landscape, three algorithms Naive Bayes Classifier, Logistic Regression, and Decision Tree Classifier are deployed to forecast these newly defined corruption categories. The efficacy of each algorithm is critically assessed through performance metrics.

Classification Results:

- Naive Bayes Classifier delivered an accuracy of 57.3%, with precision, recall, and F1-scores reflecting varying degrees of effectiveness in distinguishing between high and low trust countries.

- Logistic Regression achieved a slightly higher accuracy of 58%, showcasing its capacity to moderately differentiate between the categories, as evidenced by its precision, recall, and F1-scores.

- Decision Tree Classifier significantly outpaced the other models with an accuracy of 78.3%, demonstrating a superior ability to identify and categorize based on trust levels.

The Decision Tree Classifier emerged as the leading model, surpassing the Naive Bayes Classifier and Logistic Regression in predictive accuracy.

This outcome underscores the Decision Tree's adeptness at unraveling complex feature interactions and its robustness against non-linear data patterns, enhanced by an inherent feature selection mechanism.

Decision Tree Classifier: The strength of this model lies in its capability to dissect intricate relationships within the data, accommodating non-linearities with ease and without the need for feature scaling adjustments.

Naive Bayes and Logistic Regression: These models encounter limitations due to their foundational assumptions—Naive Bayes' expectation of feature independence and Logistic Regression's presupposition of linear relationships—which may not hold in the face of the dataset's complexity.

While accuracy serves as a straightforward metric, it might not fully capture model performance, especially in datasets with imbalanced classes. Precision, recall, and the F1-score offer a more detailed understanding, crucial for evaluating models in scenarios with disparate class frequencies.

The selection of the Decision Tree Classifier is justified by its proficiency in handling the dataset's non-linear patterns and executing effective feature selection. This adaptability makes it particularly suited for the task at hand, highlighting the importance of model selection in alignment with the data's specific attributes and the analytical objectives.

4. Chapter 4 – Conclusions and Recommendations

4.1. Conclusions

In this case study, I embarked on a comprehensive journey to unravel the complexities of happiness through the lens of data science, employing machine learning algorithms to probe into the intricacies that define and influence happiness across nations. The World Happiness Report served as the foundation of our exploration, providing a rich dataset for analysis across multiple years. Through meticulous data preparation, exploratory analysis, and the application of various machine learning models, I sought to address four key research questions (RQs) that aimed to deepen our understanding of happiness and its determinants.

RQ 1 explored the feasibility of predicting a country's region based on features extracted from the World Happiness Report data. Employing classifiers like the Decision Tree Classifier, I demonstrated a high accuracy in region prediction, thereby validating the predictive power of happiness indicators in geographical categorization.

RQ 2 aimed to forecast the Happiness Score using the report's features, with a particular focus on the Economy's influence. The application of regression models confirmed the Economy's significant impact on the Happiness Score, illustrating its paramount importance among the predictors.

RQ 3 questioned the feasibility of predicting one indicator (score) by leveraging other variables within the dataset. Through various machine learning models, I affirmed this possibility, showcasing the predictive capabilities inherent in the interrelations among different happiness indicators.

RQ 4 addressed the prediction of corruption/Trust levels between countries, considering the interrelation with other indicators. The use of classification models such as the Naive Bayes Classifier, Logistic Regression, and Decision Tree Classifier allowed for effective prediction of Trust levels, highlighting the intricate relationship between corruption perceptions and other national characteristics.

In conclusion, this case study not only answered the posed research questions but also illuminated the profound and multifaceted nature of happiness. The successful application of machine learning models to predict which features from World Happiness Report have the most impact on happiness scores, understand the influence of economic and health factors, and classify countries based on trust levels demonstrates the potential of data science in providing valuable insights into the study of happiness. These findings contribute to a deeper understanding of the determinants of happiness and offer a framework for further research and policymaking aimed at enhancing well-being on a global scale.

4.2. Contributions and Impact

This case study makes significant contributions to both data science and happiness research by applying machine learning techniques to analyze the World Happiness Report. The key contributions and impacts include:

Contributions

- Innovative Methodology: Establishes a robust framework for analyzing happiness using data science, setting a precedent for future research.
- Determinants of Happiness: Offers insights into how economic, health, and social factors intertwine to influence national happiness levels.
- Predictive Modeling: Demonstrates the potential of machine learning to forecast happiness and related indicators, enhancing our quantitative understanding of well-being.
- Economic Impact Analysis: Provides empirical evidence on the economy's role in happiness, contributing to policy discussions.
- Corruption and Trust Insights: Advances our knowledge on the relationship between governance, trust, and societal well-being.

Impact

- Policy Guidance: Supplies policymakers with data-driven insights to craft strategies aimed at improving national well-being.
- Academic Enrichment: Fuels academic debate and future research on happiness, showcasing the integration of data science in social studies.
- Public Awareness: Raises awareness about the factors contributing to happiness, advocating for targeted improvements in societal conditions.
- Data Science Applications: Highlights the versatility of data science in addressing social phenomena, encouraging interdisciplinary approaches.

In essence, this case study bridges data science and happiness studies, providing valuable insights for enhancing well-being and fostering interdisciplinary research.

4.3. Study limitations

This study makes a significant stride in understanding happiness through data science, yet it's important to recognize its limitations for a comprehensive perspective. Firstly, the reliance on the World Happiness Report, while valuable, may not encapsulate the full spectrum of happiness or cultural variations, given its standardized data collection methods. Changes in variables over the years and data uniformity also present challenges in ensuring a seamless longitudinal analysis.

The use of machine learning models, though innovative, comes with its own set of constraints. These models operate under specific assumptions and may not capture the complex, multifaceted nature of happiness. The study's findings are inherently influenced by the limitations of these algorithms, affecting predictive accuracy and interpretation.

Moreover, the generalizability of results is a concern. The analysis, based on national-level data, might not accurately reflect the individual experiences of happiness, potentially overlooking significant within-country variations. Such aggregation could mask vital socio-economic and demographic differences that influence happiness.

Lastly, the study is bound by its temporal scope, focusing on a defined period that precludes the examination of long-term happiness trends or the effects of recent significant global events, such as the COVID-19 pandemic. These limitations underscore the need for cautious interpretation of the findings and suggest avenues for future research to explore, aiming to overcome these challenges and deepen our understanding of happiness.

4.4. Future Work and Recommendations

Future research in this field will aim to broaden the scope of analysis by incorporating a wider range of machine learning and deep learning techniques, applied to more extensive datasets over longer periods. This effort seeks to deepen our understanding of happiness and its evolving drivers.

A key recommendation for enhancing future studies is the standardization of formats and variables in the World Happiness Report, facilitating easier data analysis and comparison over time. Additionally, incorporating diverse data sources, including qualitative research and social media analysis, could provide a more comprehensive view of global happiness.

These steps forward will not only enrich academic insights into the factors influencing happiness but also offer valuable information for developing policies and interventions aimed at improving well-being across different populations and cultures.

References

[1] Y. Zhang, “Analyze and Predict the 2022 World Happiness Report Based on the Past Year’s Dataset,” *Journal of Computer Science*, vol. 19, no. 4, pp. 483–492, 2023, doi: 10.3844/jcssp.2023.483.492.

[2] R. Setiawan, Z. E. Rasjid, and A. Effendi, “The Effect of World Happiness Aspects with COVID-19 Infected using Machine Learning Multiple Regression Model,” in *International Conference on Electrical, Computer, and Energy Technologies, ICECET 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICECET52533.2021.9698628.

[3] M. Varsha, M. Ramya, C. C. Sobin, NP. Subheesh, and J. Ali, “Assessing Emotional Well-being of Students using Machine Learning Techniques,” Institute of Electrical and Electronics Engineers (IEEE), Mar. 2022, pp. 336–340. doi: 10.1109/ocit53463.2021.00073.

[4] S. Kino *et al.*, “A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects,” *SSM - Population Health*, vol. 15. Elsevier Ltd, Sep. 01, 2021. doi: 10.1016/j.ssmph.2021.100836.

[5] F. Marshall, R. Bond, and S. Zhang, “Measuring and Visualising Global Happiness,” BCS Learning & Development, 2018. doi: 10.14236/ewic/hci2018.210.

[6] A. Jannani, N. Sael, and F. Benabbou, “Predicting Quality of Life using Machine Learning: case of World Happiness Index,” in *2021 4th International Symposium on Advanced Electrical and Communication Technologies, ISAECT 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ISAECT53699.2021.9668429.

[7] G. IEEE Engineering in Medicine and Biology Society. Annual International Conference (41st : 2019 : Berlin, IEEE Engineering in Medicine and Biology Society, and Institute of Electrical and Electronics Engineers, *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) : Biomedical Engineering Ranging from Wellness to Intensive Care : 41st EMB Conference 2019 : July 23-27, Berlin.*

[8] J. A. Arias *et al.*, “The neuroscience of sadness: A multidisciplinary synthesis and collaborative review,” *Neuroscience and Biobehavioral Reviews*, vol. 111. Elsevier Ltd, pp. 199–228, Apr. 01, 2020. doi: 10.1016/j.neubiorev.2020.01.006.

[9] H. C. Cromwell *et al.*, “Mapping the interconnected neural systems underlying motivation and emotion: A key step toward understanding the human affectome,” *Neuroscience and Biobehavioral Reviews*, vol. 113. Elsevier Ltd, pp. 204–226, Jun. 01, 2020. doi: 10.1016/j.neubiorev.2020.02.032.

[10] M. Lubis, D. O. D. Handayani, and A. R. Lubis, “Prediction analysis of the happiness ranking of countries based on macro level factors,” *IAES International Journal of Artificial Intelligence*, vol. 11, no. 2, pp. 666–678, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp666-678.

[11] IEEE Staff, *2019 5th International Conference on Science in Information Technology (ICSITech)*. IEEE, 2019.

[12] H. Huang, X. Wang, S. Sengan, and T. Chandu, “Emotional intelligence for board capital on technological innovation performance of high-tech enterprises,” *Aggression and Violent Behavior*. Elsevier Ltd, 2021. doi: 10.1016/j.avb.2021.101633.

[13] D. De Ridder, D. Adhia, and S. Vanneste, “The anatomy of pain and suffering in the brain and its clinical implications,” *Neuroscience and Biobehavioral Reviews*, vol. 130. Elsevier Ltd, pp. 125–146, Nov. 01, 2021. doi: 10.1016/j.neubiorev.2021.08.013.

[14] A. A. Mocanu and A. Iftene, “How the Events in the Life of Painters Influence the Colors of their Paintings,” in *Proceedings - 2021 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 105–112. doi: 10.1109/SYNASC54541.2021.00028.

[15] D. Nieciecka, “Predicting Happiness-Comparison of Supervised Machine Predicting Happiness-Comparison of Supervised Machine Learning Techniques Performance on a Multiclass Classification Learning Techniques Performance on a Multiclass Classification Problem Problem.” [Online]. Available: https://arrow.tudublin.ie/scschcomdis

[16] M. M. N. Bieńkiewicz *et al.*, “Bridging the gap between emotion and joint action,” *Neuroscience and Biobehavioral Reviews*, vol. 131. Elsevier Ltd, pp. 806–833, Dec. 01, 2021. doi: 10.1016/j.neubiorev.2021.08.014.

[17] E. F. Pace-Schott *et al.*, “Physiological feelings,” *Neuroscience and Biobehavioral Reviews*, vol. 103. Elsevier Ltd, pp. 267–304, Aug. 01, 2019. doi: 10.1016/j.neubiorev.2019.05.002.

[18] K. Suzuki, P. Siriaraya, W. J. She, R. Tanaka, D. Li, and S. Nakajima, “HappyRec: Evaluation of a ‘happy Spot’ Recommendation System Aimed at Improving Mental Well-Being,” in *IEEE International Conference on Data Mining Workshops, ICDMW*, IEEE Computer Society, 2021, pp. 889–892. doi: 10.1109/ICDMW53433.2021.00116.

[19] C. Sweeney, E. Ennis, M. Mulvenna, R. Bond, and S. O’neill, “How Machine Learning Classification Accuracy Changes in a Happiness Dataset with Different Demographic Groups,” *Computers*, vol. 11, no. 5, May 2022, doi: 10.3390/computers11050083.

[20] M. A. Khder, M. A. Sayfi, and S. W. Fujo, “Analysis of World Happiness Report Dataset Using Machine Learning Approaches,” *International Journal of Advances in Soft Computing and its Applications*, vol. 14, no. 1, pp. 14–34, 2022, doi: 10.15849/IJASCA.220328.02.

[21] S. Li, Y. Wang, J. Xue, N. Zhao, and T. Zhu, “The impact of covid-19 epidemic declaration on psychological consequences: A study on active weibo users,” *Int J Environ Res Public Health*, vol. 17, no. 6, Mar. 2020, doi: 10.3390/ijerph17062032.

[22] F. Ibnat, J. Gyalmo, Z. Alom, M. A. Awal, and M. A. Azim, “Understanding World Happiness using Machine Learning Techniques,” in *6th International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering, IC4ME2 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/IC4ME253898.2021.9768407.