UNIVERSITÀ
DEGLI STUDI
DI MILANO

DATA SCIENCE FOR ECONOMICS

Statistical Learning

Road Safety Analysis

Daniele PIAZZA

Matriculation number n° 45740A

Academic year 2023/2024

Contents

1 Abstract

This project uses statistical learning techniques applied to road safety analysis in two main areas: unsupervised learning and supervised learning.

The Unsupervised Learning section examines a dataset that contains information about various countries and their road safety laws, along with the number of traffic deaths per 100,000 inhabitants. The goal is to identify clusters between countries using laws and to determine whether these groups have similar death rates.

In the part concerning Supervised Learning, I used a dataset containing data regarding individual road accidents. The goal was to develop predictive models in order to estimate the seriousness of an accident, severe or slight, based on its characteristics, such as weather conditions, driver's age, road type, time and other relevant factors.

Through these two approaches the project aims to provide insights into road safety.

2 Road safety by countries

2.1 Introduction

In this chapter I will explore unsupervised learning techniques to analyze road safety in different countries and the relationship with fatality rate. I have combined multiple datasets obtained from the World Health Organization (WHO)[1], which provide comprehensive road safety information across five main themes: national legislation, policy frameworks, post-crash response, road traffic mortality rates and institutional framework data.

Dataset Name	**Description**
Institutional Framework	Existence of a road safety lead agency Existence of a road safety strategy Availability of funding for road safety strategy
National Legislation	Existence of a drink-driving law Blood Alcohol Concentration (BAC) limit Attribution of deaths to alcohol (%) Existence of a seat-belt law Applicability of seat-belt law Seat-belt wearing rate (%) Existence of a child-restraint law Existence of speed limits Maximum speed limits Motorcycle helmet law to all occupants Law requires helmet to be fastened
Policy	Vehicle standards
Post-crash Response	Existence of a universal access number
Road Traffic Mortality	Estimated road traffic death rate Road traffic deaths by type of road user (%)

Table 1: Datasets Used in Unsupervised Learning Analysis

I used these datasets to analyze road safety and identify clusters between

countries with similar road laws. The goal was to determine if these clusters are related to road mortality rates.

To create a single dataset for analysis, I merged all individual datasets after examining them separately. This process involved:

- Selecting relevant variables.
- Converting categorical and ordinal variables into numerical ones.
- Transforming binary values, like ”yes” and ”no,” into 1 and 0, respectively.

After merging the datasets, I analyzed the columns and removed those with a large number of missing data, I did this to not have a limited number of countries.

The following is a list of variables, along with their description and possible values:

- **safety**: Indicates whether there is a road safety lead agency (0 = No, 1 = Yes).
- **strategy**: Represents the existence of a national road safety strategy (0 = No, 1 = Yes).
- **funding**: Refers to the availability of funding for the road safety strategy (0 = ”Not funded”, 1 = ”Partially funded”, 2 = ”Fully funded”).
- **seat_law**: Whether a national seat-belt law exists (0 = No, 1 = Yes).
- **restraint_law**: Whether a national child-restraint law exists (0 = No, 1 = Yes).
- **speed_limits**: Indicates whether national speed limits exist (0 = No, 1 = Yes).
- **speed_rural**: The maximum speed limits on rural roads (Integer).
- **speed_urban**: The maximum speed limits on urban roads (Integer).
- **helmet_child**: Applicability of motorcycle helmet laws to children (0 = No, 1 = Yes).
- **helmet_adult**: Applicability of motorcycle helmet laws to adults (0 = No, 1 = Yes).

- **helmet_driver**: Applicability of motorcycle helmet laws to drivers (0 = No, 1 = Yes).
- **helmet_fasten**: Whether the law requires helmets to be fastened (0 = No, 1 = Yes).
- **std_child_seats**: Whether there are standards for child seats (0 = No, 1 = Yes).
- **std_esc**: Refers to the presence of Electronic Stability Control (ESC) standards (0 = No, 1 = Yes).
- **std_front**: Indicates whether there are standards for frontal impact (0 = No, 1 = Yes).
- **std_moto**: Indicates whether there are standards for motorcycles (0 = No, 1 = Yes).
- **std_pedestrian**: Indicates whether there are standards for pedestrian protection (0 = No, 1 = Yes).
- **std_belt**: Whether there are standards for seat-belts (0 = No, 1 = Yes).
- **std_belt_anc**: Whether there are standards for seat-belt anchorages (0 = No, 1 = Yes).
- **std_side**: Indicates whether there are standards for side impact (0 = No, 1 = Yes).
- **emergency_number**: Indicates the existence of a universal access telephone number for pre-hospital care (0 = ”None”, 1 = ”Partial coverage”, 2 = ”National, single number”, 3 = ”National, multiple numbers”).
- **death**: Represents the number of road traffic deaths per 100,000 inhabitants (real number).

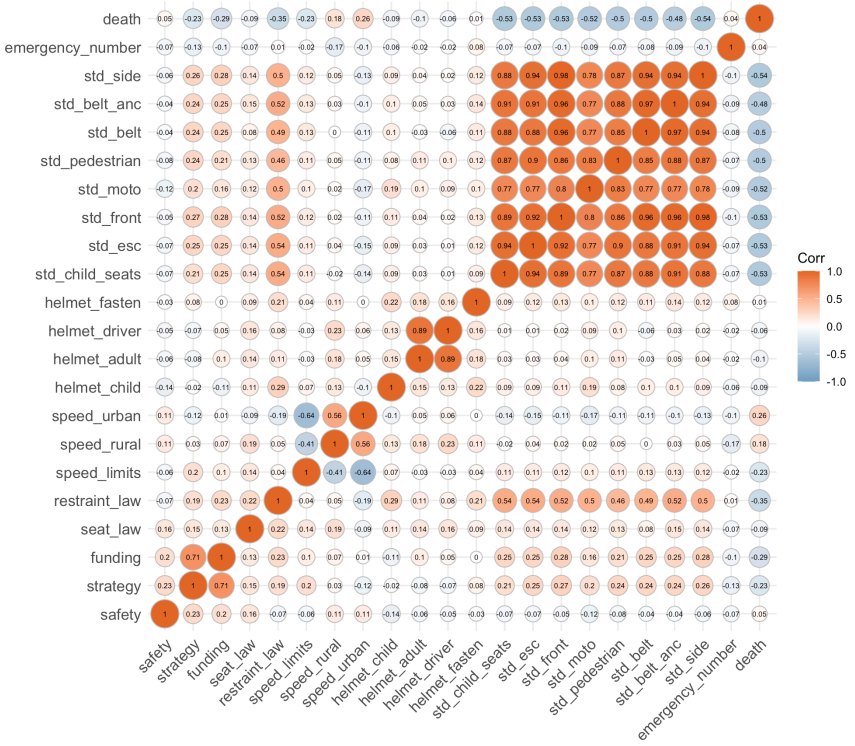

Figure 1: Correlation Matrix

Looking at the correlation matrix we see that different vehicle safety standards are highly correlated with each other, indicating that some safety standards are often implemented together.

To reduce redundancy and simplify the dataset, I decided to calculate the average, creating a new variable ”std_car”. The original columns were removed to avoid redundant information.

Below is the updated correlation matrix, which reflects this change and shows the new relationship between the remaining variables:

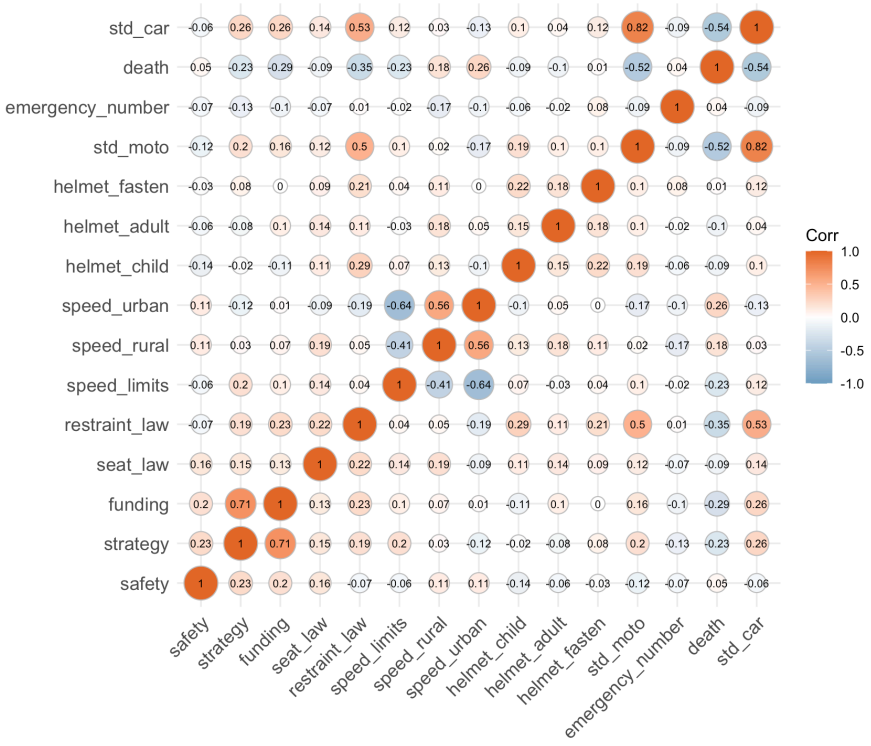

Figure 2: Correlation Matrix with std_car

The new correlation matrix reveals several key relationships. As expected, the number of traffic deaths is inversely related to safety standards, seat belt laws, and child restraint systems. This suggests that higher safety standards are associated with lower road traffic fatality rates.

On the contrary, the number of deaths from road accidents shows a direct correlation with maximum speed limits. So higher speed limits could contribute to the increase of the risk of fatal accidents.

These correlations highlight the importance of road safety laws in reducing road fatalities.

2.2 Exploratory Data Analysis

After preparing the dataset, I conducted an exploratory data analysis (EDA). I have created a series of maps illustrating various aspects of road safety, looking at various key factors such as fatal road accidents, speed limits and safety standards.

Thanks to the following maps we want to have a better understanding about how the various road safety factors are applied in different countries

and to note a difference between continents.

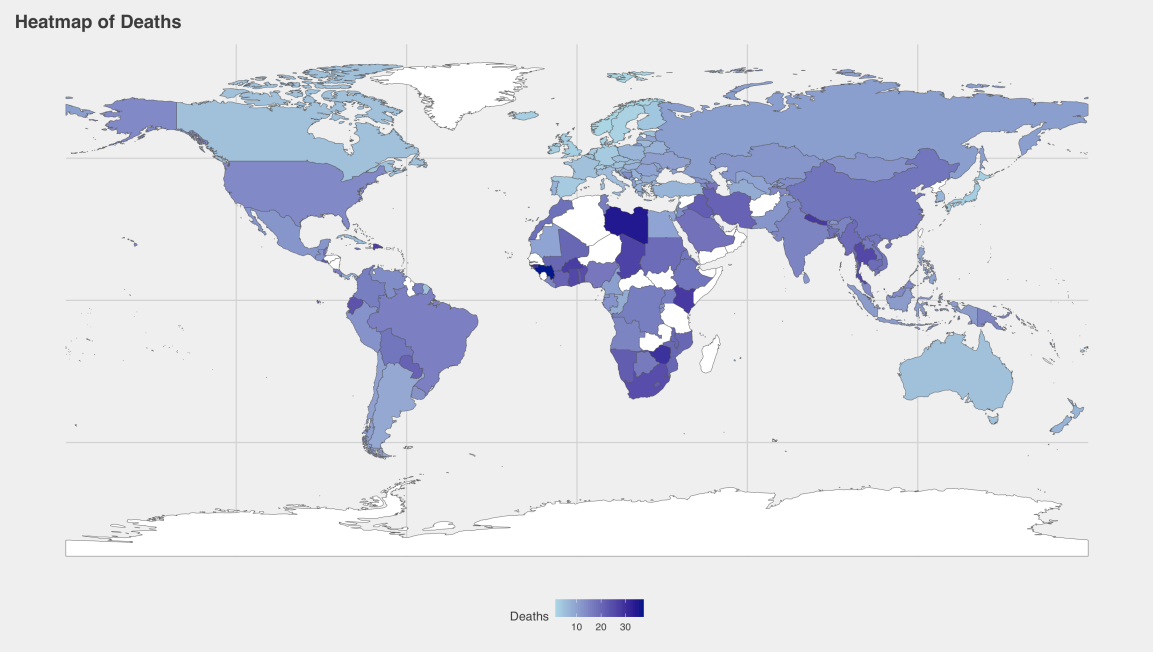

Figure 3: Heatmap of road traffic deaths per 100,000 inhabitants

This heat map shows the number of road accident deaths per 100,000 inhabitants in different countries. The countries with higher rates are African, Asian and South American. The ones with lower rates are to be found in Europe, Oceania and North America. This suggests that there may be factors contributing to these disparities.

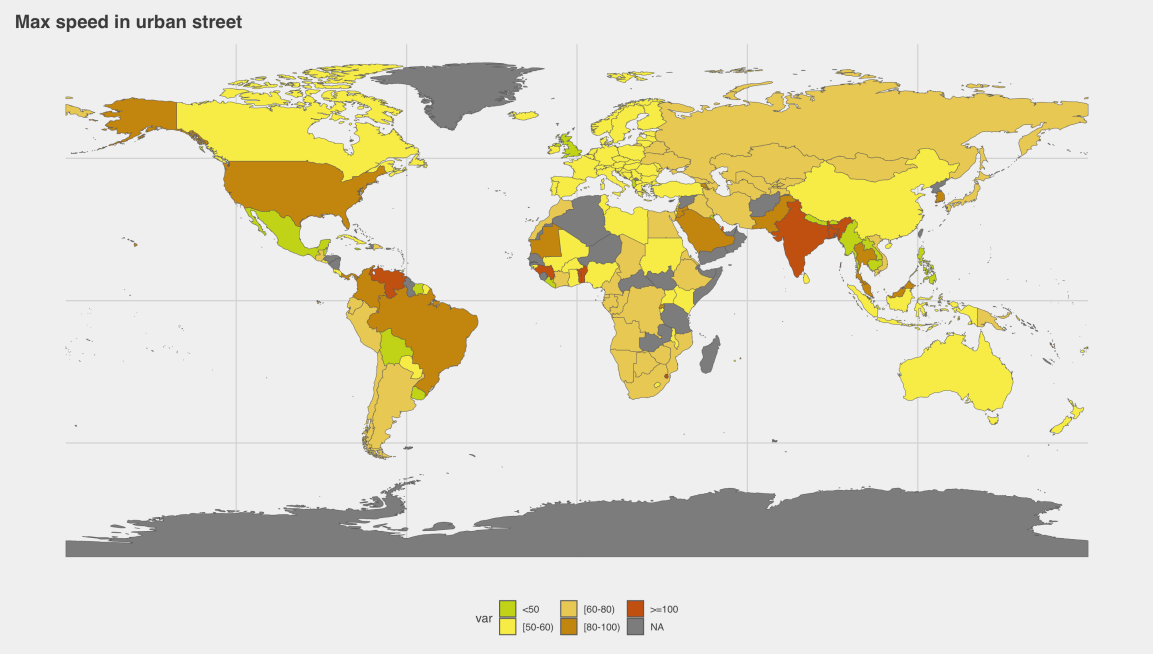

Figure 4: Map of Maximum Speed Limits in Urban Roads

This map shows the maximum speed limits in urban areas of different countries. European and Oceanic countries generally have limits ranging from 50 to 60 km/h. In Asian countries, most limits are between 60 and 80 km/h, with some exceptions, such as India, where limits can reach 100 km/h or higher. African countries typically have figures between 60 and 80 km/h, with some between 50 and 60 km/h.

North and South America show greater variability.

The only European country with a speed limit lower than 50 km/h is England.

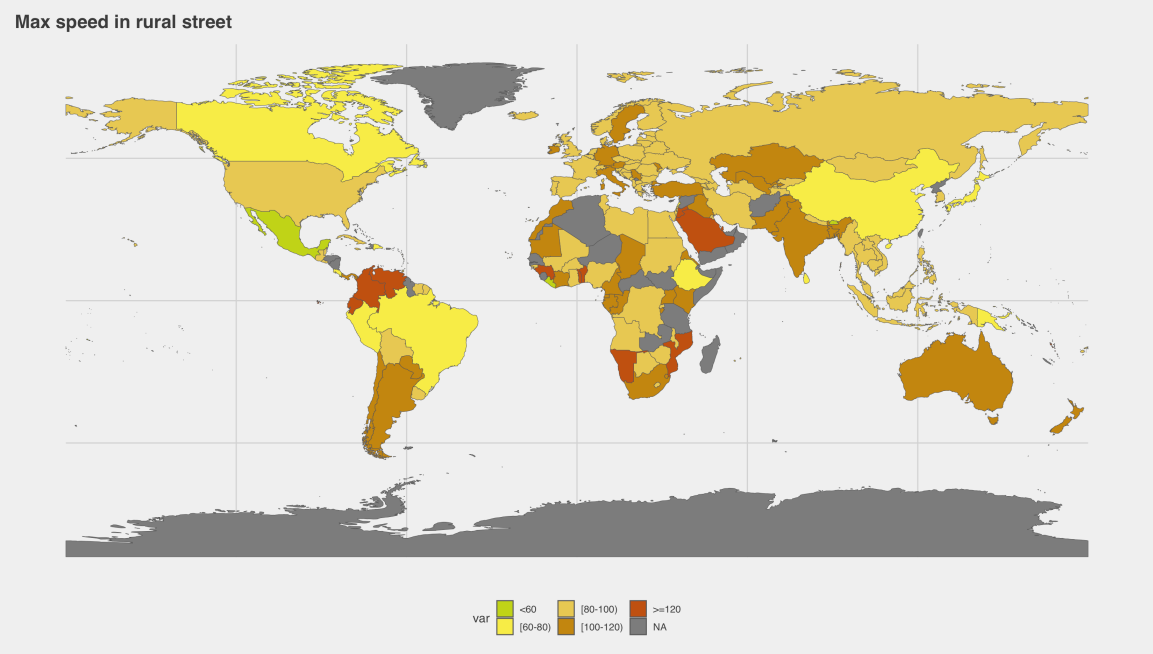

Figure 5: Map of Maximum Speed Limits on Rural Roads

This map shows maximum speed limits on rural roads in various countries. In most countries they are between 80 and 100 km/h or between 100 and 120 km/h. However, there are exceptions such as: China, Canada and Mexico which have speed limits lower than 80 km/h. In contrast, some South American and African countries, such as Mozambique and Venezuala, have limits greater or equal to 120 km/h, suggesting a potentially higher risk.

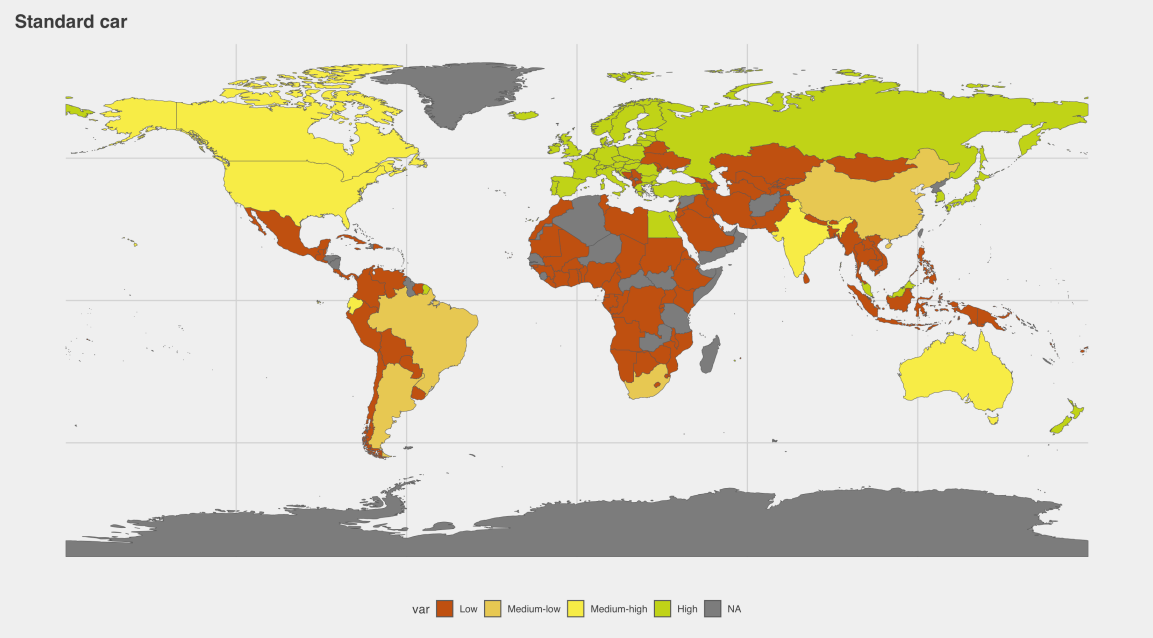

Figure 6: Map of Vehicle Safety Standards by Region

This map illustrates the level of car safety standards across the world. European countries generally have high to complete safety standards. In contrast, African countries tend to have low or no safety standards, suggesting a potential area for improvement in road safety regulations.

Asian countries exhibit a mixed approach, with some countries having high safety standards while others falling below average. North America tends toward medium to high safety standards. On the other hand, South American countries typically have low to medium safety standards.

2.3 Unsupervised Learning

In this chapter, I will explore some unsupervised learning algorithms to analyze road safety data. My goal is to find clusters based on countries' road laws and see if death rates are distributed among those.

I used a combination of Principal Component Analysis (PCA), K-Means clustering, and hierarchical clustering. PCA helps to reduce the dimensionality of the dataset allowing for better visualization and interpretation. K-Means and hierarchical clustering were employed to group countries into clusters

with similar characteristics.

To prepare the dataset for these analyses, I scaled the numerical variables in order to avoid that some variables could dominate due to differing scales.

Moreover, I removed the ”death” variable, this step was necessary to avoid biasing the analysis with a direct outcome-related metric. By excluding this variable, the clustering process focuses on the underlying factors contributing to road safety, without being deviated by the actual fatality rates.

After these preparations, I proceeded to apply unsupervised learning algorithms to explore the dataset and identify meaningful clusters among countries.

2.3.1 Principal Component Analysis (PCA)

The goal of Principal Component Analysis (PCA) is to identify a smaller set of components that explain the majority of the variance in the dataset, which can then be used for further analysis and clustering.

For this analysis, I calculated the PCA and selected the first three principal components, which together explain 64% of the total variance. This reduction in dimensionality helps to represent the data.

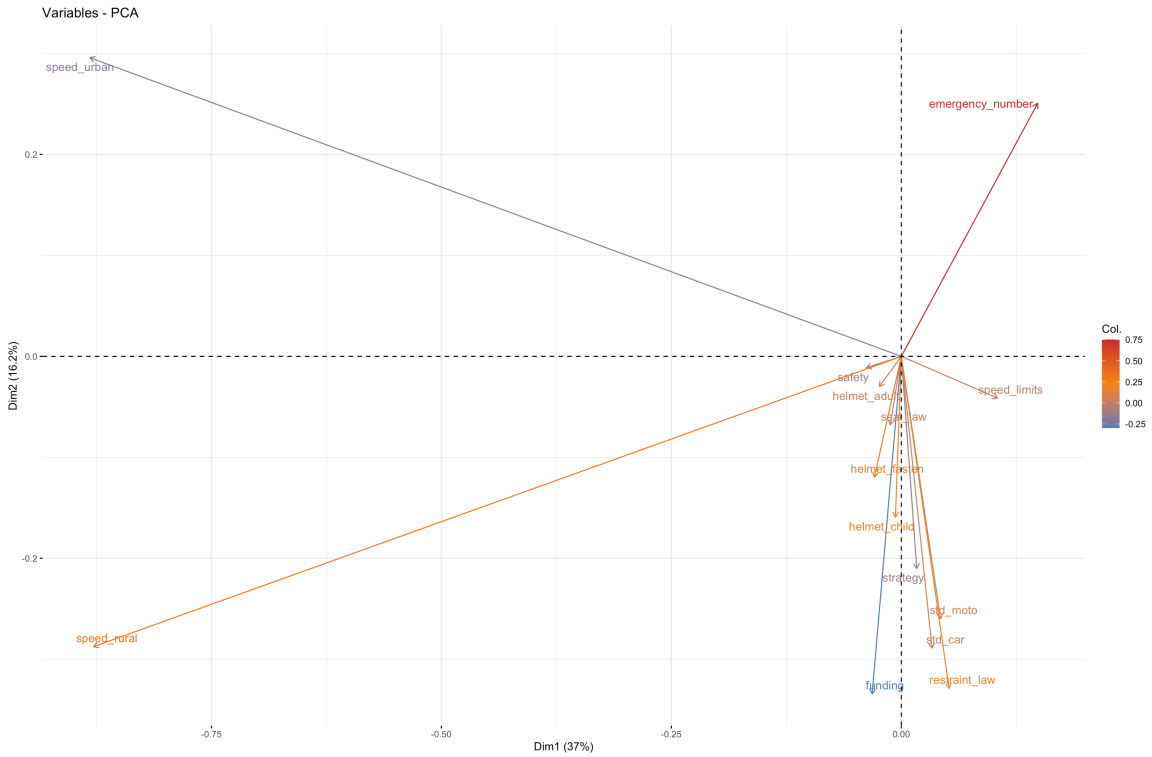

Figure 7: PCA loadings rapresentation. The color represents the third component.

Observing the PCA loadings, we can see that the safety standards for cars and motorcycles are mostly represented by the second component in the negative direction. The urban speed limits are represented in a negative direction in the first component and in a positive direction in the second component. The rural speed limits are negative in both of them.

The presence of emergency numbers is strongly represented by the second and third components.

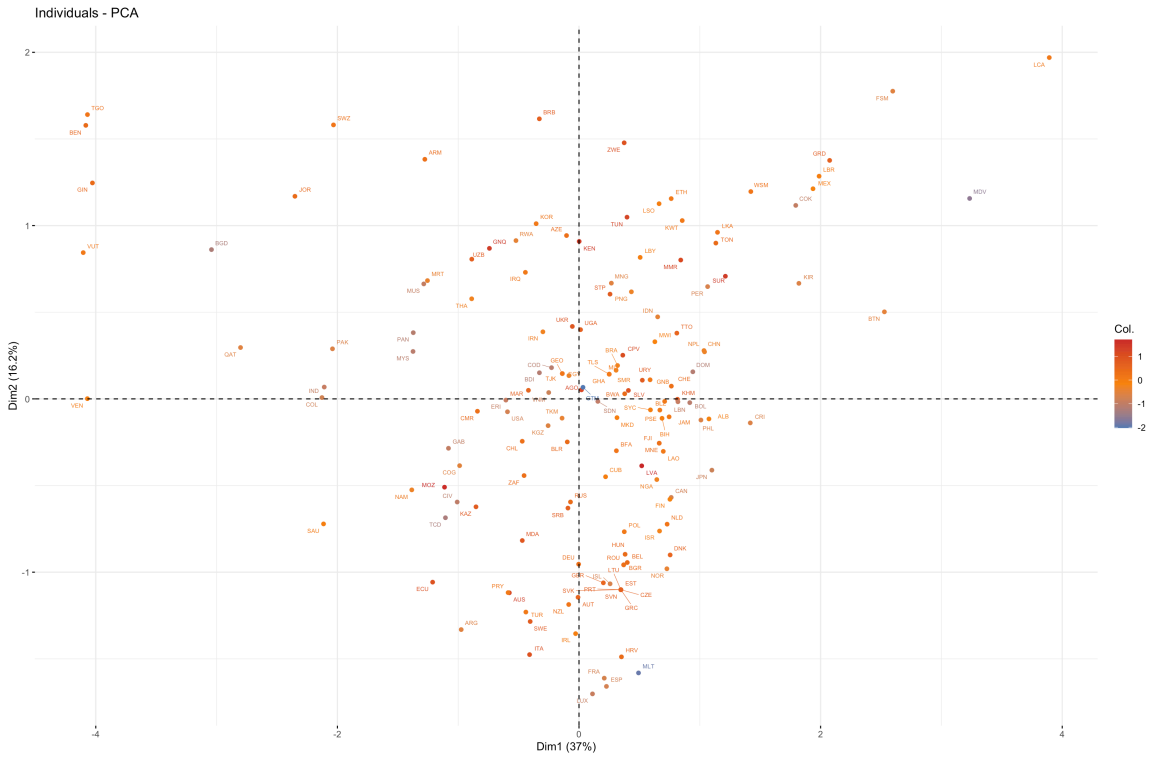

Figure 8: Countries represented in the PCA 2-D space. The color represents the third component.

The second component is extremely useful since it splits the countries in those with restrictive roads' safety laws and the ones with less harsh ones. In fact, European countries tend to be closer in the lower part of the graph, indicating that they have similar road safety standards. The tight grouping could be a result of similar regulatory frameworks and traffic laws across Europe.

2.3.2 Clustering

K-Means clustering is an unsupervised learning technique used to group similar data points into clusters. To determine the optimal number of clusters, I used the elbow method, which plots the within-cluster sum of squares (WSS) against the number of clusters.

The following sections discuss the outcomes of the K-Means clustering, including the visualization of the clusters and insights into what distinguishes each cluster.

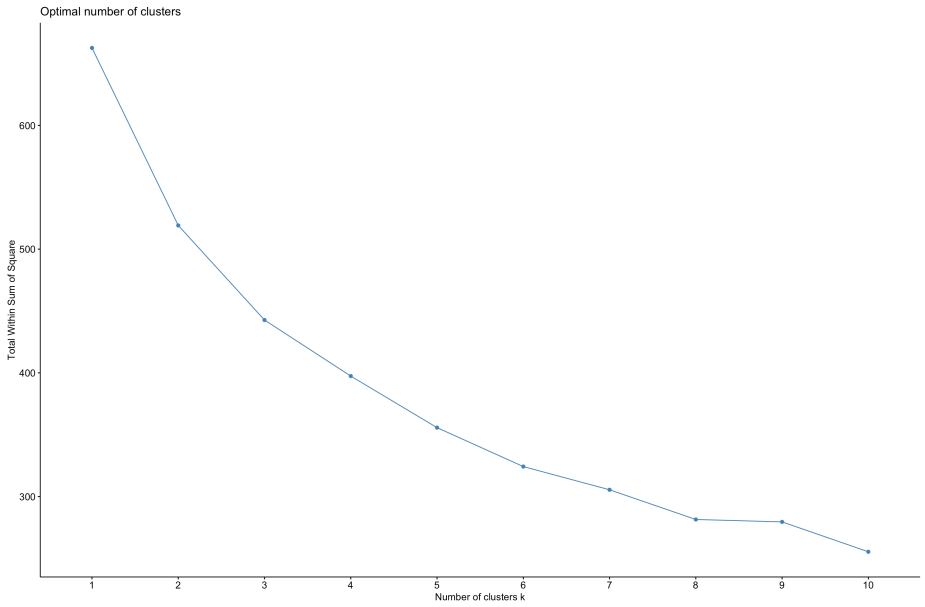

Figure 9: Elbow plot (WSS)

Based on the elbow plot, I selected five clusters.

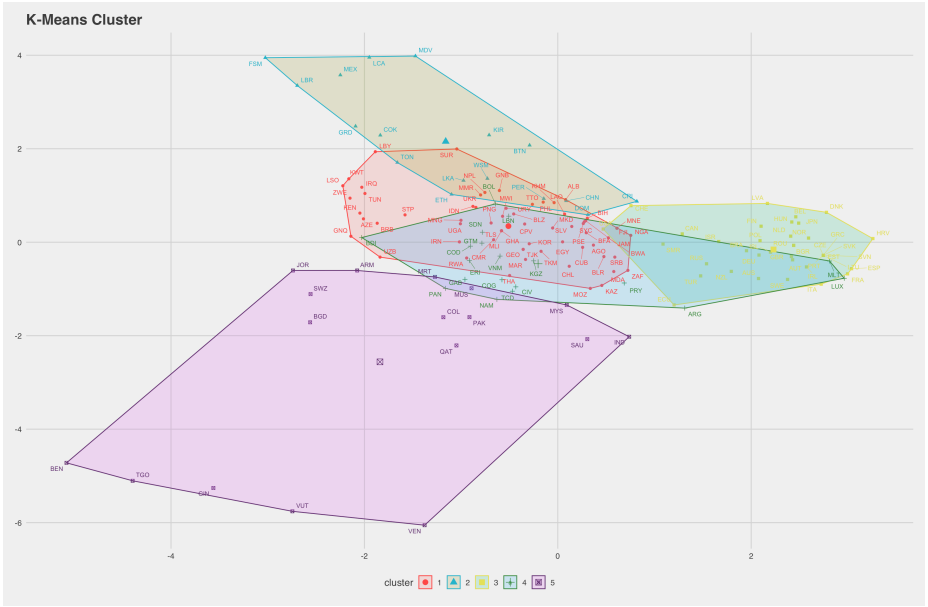

Figure 10: Clusters formed by K-Means

The plot above visualizes the clusters formed by K-Means clustering, with data points grouped in the PCA-reduced feature space. This visualization helps to identify distinct clusters based on shared road safety characteristics.

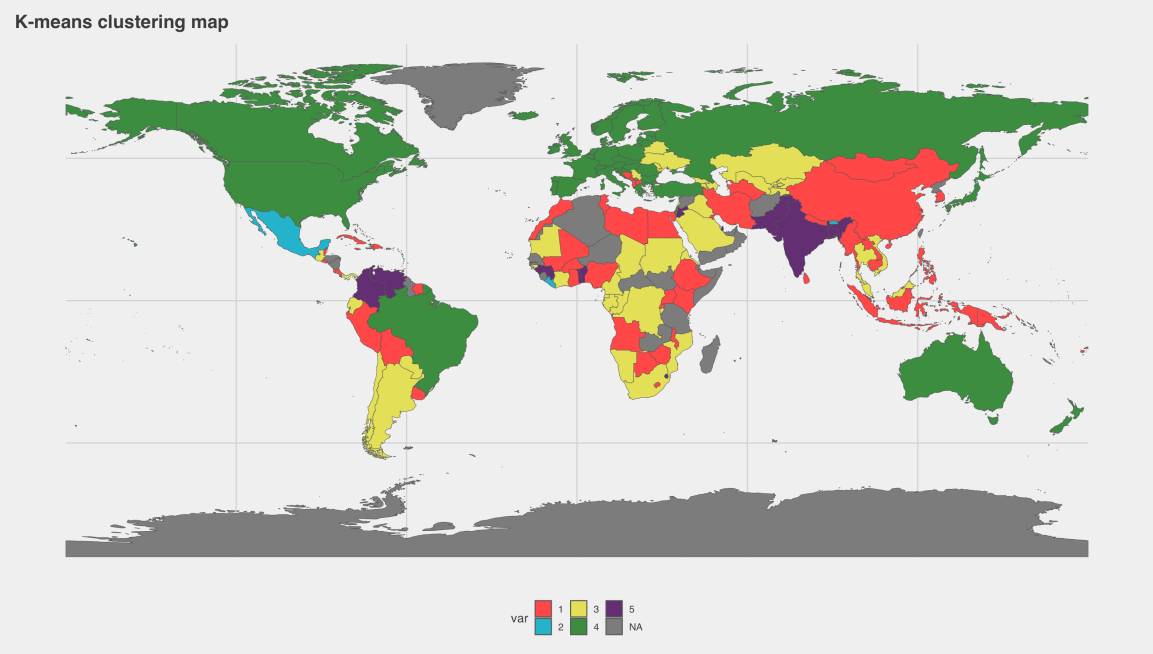

Figure 11: K-Means Clusters by Country

Visualizing the K-Means clustering on the map, it's easy to observe a division between more developed countries with stricter road safety laws and less developed or developing countries.

Here is a detailed breakdown of the clusters and their notable characteristics:

- **Cluster 1 - Red**: This cluster consists of several East Asian countries, some Oceanic islands and a few African states.
- **Cluster 2 - Light Blue**: This cluster includes smaller islands, with Mexico being a key member. The geographical distribution and smaller size of these regions could explain the grouping, as they might face unique road safety challenges.
- **Cluster 3 - Yellow**: Primarily composed of African countries, suggesting common road safety characteristics.
- **Cluster 4 - Green**: Including European, North American countries and Australia, this cluster generally represents more developed regions with stricter safety standards and advanced road safety regulations.

- **Cluster 5 - Purple**: This cluster includes India and its neighboring countries, as well as some South American countries.

These clusters provide valuable insights into global road safety trends and highlight the potential influence of development status and road safety laws. Understanding these groupings can guide further analysis and support the development of targeted road safety policies.

I've also used hierarchical clustering using the Ward method to evaluate different clusters.

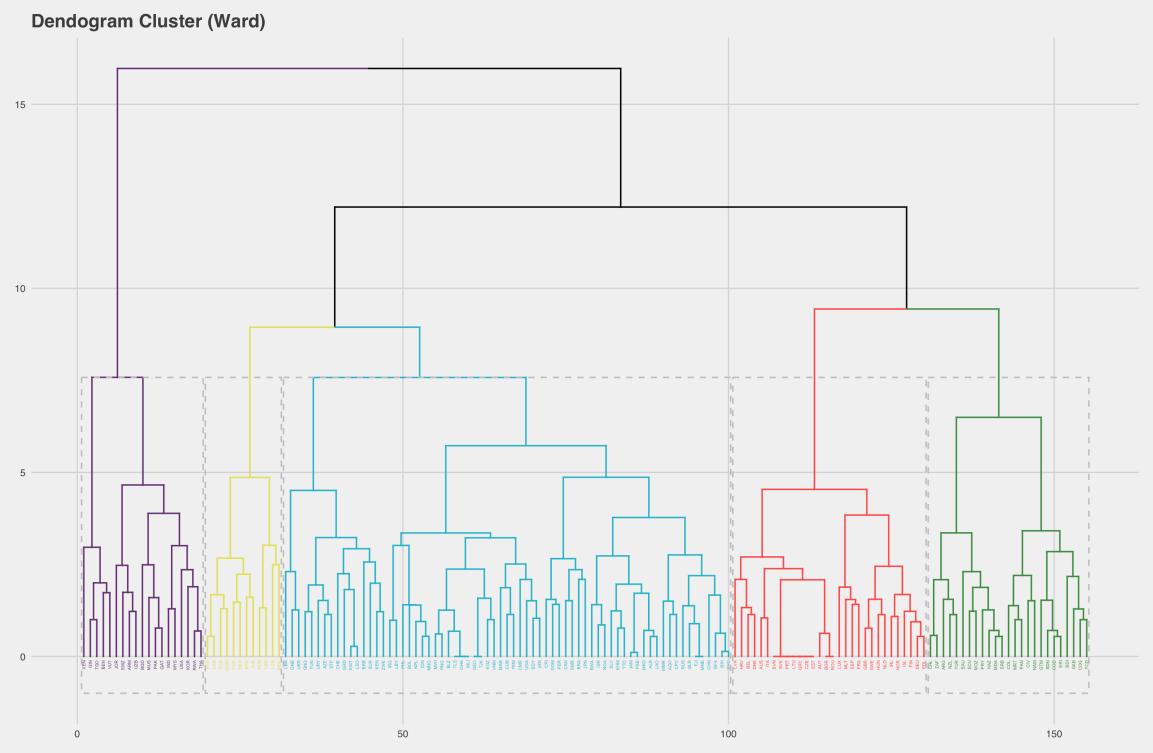

Figure 12: Dendrogram of Clusters with the Ward Method

This dendrogram shows the relationships among different clusters and indicates how countries are grouped based on shared characteristics.

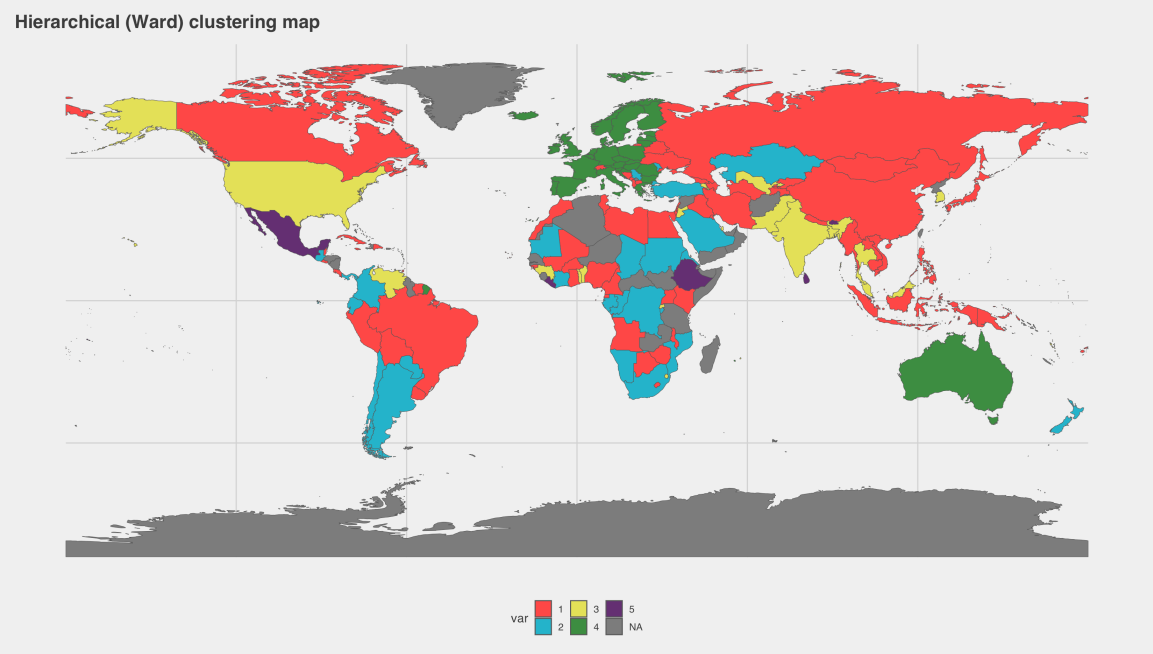

Figure 13: Hierarchical Clusters by Country

The hierarchical clustering map, compared to K-Means, reveals a slightly different pattern.

Here's an explanation of the clusters formed by hierarchical clustering, along with notable characteristics for each:

- **Cluster 1 - Red**: This cluster consists primarily of Asian countries, but also includes regions from South America, Africa, and Oceanic islands. This broad composition might suggest common characteristics shared by diverse regions.
- **Cluster 2 - Light Blue**: Composed of African and South American countries not included in Cluster 1. These countries could have similar road safety issues, such as lower safety standards.
- **Cluster 3 - Yellow**: This cluster features the United States and India, along with several smaller islands.
- **Cluster 4 - Green**: Includes European countries and Australia, this cluster represents regions with generally high road safety standards.

- **Cluster 5 - Purple**: Mainly consisting of islands, indicating a different approach to road safety, possibly influenced by geographical factors or smaller population sizes.

These clusters highlight the different approaches to road safety around the world and suggest that geographical, cultural, and economic factors could play a significant role in how countries manage road safety. Understanding these differences is critical in order to identify potential areas for improvement and could be very useful in order to reduce road traffic mortality.

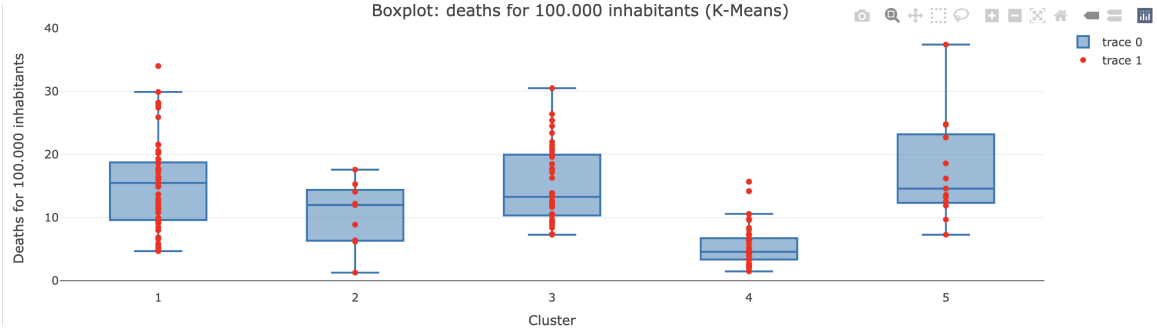

Figure 14: Boxplot for K-Means Clusters

The boxplot derived from K-Means clustering shows the range of road traffic deaths across different clusters. Here's the order from the cluster with ascending number of deaths:

- **Cluster 4**: This cluster has the lowest number of deaths. Most countries in this cluster, including several European nations, have relatively low fatality rates. However, outliers like the United States and Brazil show higher rates.
- **Cluster 2**: Composed of smaller islands and Mexico, this cluster shows a range of road traffic deaths. The unique geography of these smaller regions might play a role in the results.
- **Cluster 3**: This cluster represents a moderate number of road traffic deaths. It's observable that there are two groups of countries based on the death rates. The ones with the highest number are the African, the other ones are from South America and Eastern Europe.
- **Cluster 5**: This cluster shows only a few countries: on the bottom there is a group with fewer deaths and on the upper whisker there are some countries that increase the average.

- **Cluster 1**: This cluster has the highest number of road traffic deaths, with a mix of East Asian countries, Oceanian islands, and several African states. This indicates that these regions may require stronger road safety regulations to reduce traffic fatalities.

These results from the K-Means clustering boxplot highlight distinct patterns in road traffic deaths, emphasizing where road safety measures might be lacking. Understanding these trends can help guide further analysis and inform policy decisions to improve road safety in regions with higher fatality rates.

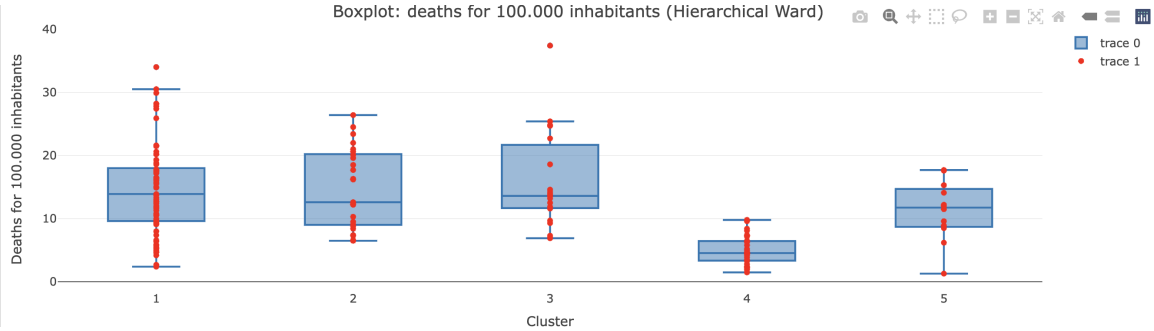

Figure 15: Boxplot for Hierarchical (Ward) Clusters

The boxplot derived from hierarchical clustering using the Ward method illustrates the distribution of road traffic deaths across different clusters. Here's a list of the clusters ordered by the number of road traffic deaths, from lowest to highest:

- **Cluster 4**: This cluster has a consistently low number of road traffic deaths, with most countries showing similar rates. There are no significant outliers, suggesting stable road safety across these regions. It could be like that because Europe possibly has some standard laws.
- **Cluster 5**: This cluster primarily consists of smaller islands, which exhibit a varied range of road traffic death rates. The geographical diversity could account for the differences observed in this cluster.
- **Cluster 2**: This cluster is a mix of African and South American countries, with a bimodal distribution in road traffic deaths. Some countries in this cluster have higher fatality rates, particularly in parts of Africa, indicating possible safety issues.

- **Cluster 3**: This cluster has a moderately high number of deaths but that's because of some outliers that increase the mean. The biggest countries, such as USA and India, are below the average of the cluster.
- **Cluster 1**: This cluster includes countries with a wide range of road traffic death rates, with several African countries that have significantly higher rates.

These results from the hierarchical clustering boxplot reveal patterns in road traffic deaths, pointing out where road safety measures may need to be improved.

2.4 Conclusions

The boxplot analysis reveals the variability in road traffic deaths across different clusters, as defined by both K-Means and hierarchical clustering methods. The data suggest that road safety outcomes are influenced by various factors, including safety standards, geographical distribution and development status. The clusters with higher mean death rates are the ones that include developing countries. On the other hand, the most advanced nations tend to have high security standards, therefore less road deaths.

However, outliers and variability within clusters indicate that other factors might play a role, for example:

- Road maintenance (paved versus unpaved).
- Type and age of vehicles.
- Road infrastructure relative to country size.
- Economic factors affecting safety regulations.
- Number of vehicle relative to the inhabitants.
- Mix of road users (pedestrians, cyclists, motorcyclists, drivers, ...).
- Number of Speed Cameras.

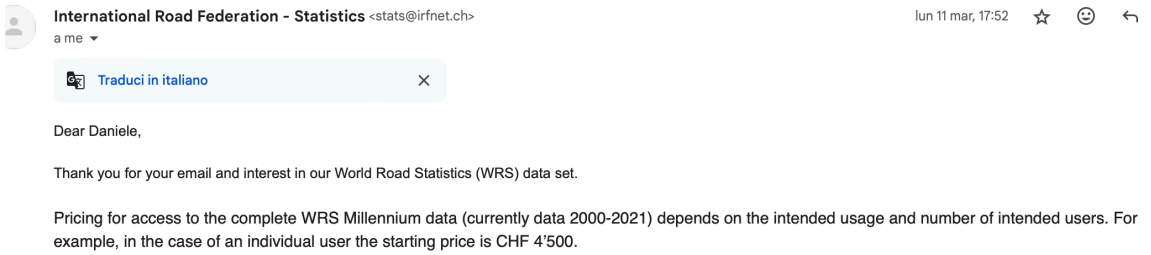

Figure 16: Mail with IRF

Unfortunately, obtaining data on these factors from sources like the International Road Federation proved to be challenging due to high costs, with a price of 4,500 CHF. As a result, the analysis was limited to publicly available data, focusing only on road safety variables.

3 Accident Analysis

3.1 Introduction

The dataset used for this supervised learning analysis is sourced from UK.gov[2] and contains detailed information about individual traffic accidents. The data includes a wide range of attributes, such as:

- The severity of the accident (fatal, serious, or slight).
- The type of road where the accident occurred (roundabout, one-way street, etc.).
- Weather conditions at the time of the accident.
- Light conditions, road surface, and pedestrian crossing facilities.
- Details about the vehicles involved, such as vehicle type and whether the vehicle was left-hand drive.
- Information about the drivers, including age and gender.

Given the raw state of the data, significant preprocessing was required to clean and transform the dataset into a suitable form for supervised learning. Here are the key steps involved in the data cleaning and transformation process:

1. **Data Merging**: The dataset was created by merging two separate datasets: ‘collision‘, and ‘vehicle‘. This allowed for a comprehensive view of each accident.
2. **Data Cleaning**: Columns with high null values or those not relevant to the analysis were removed. Rows with missing or unknown values in critical fields were also filtered out.
3. **Feature Engineering**: Several transformations were applied to convert categorical variables into numerical ones, typically by creating dummy columns. For example:
    - **Road Types**: Converted into separate dummy columns, with specific types grouped or removed to focus on essential road structures.

- **Weather Conditions**: Mapped to key categories and converted into dummy columns.
- **Light Conditions**: Simplified to a smaller set of categories.
- **Vehicle Types**: Grouped into broader categories and then converted into dummy columns.
- **Driver Information**: Gender was converted into a binary variable, while other driver attributes were retained or transformed as needed.
- **Road Surface and Other Road variables**: Consolidated into binary variables where applicable, such as whether the road was dry or not.

4. **Data Reduction**: Some features were dropped because they provided redundant information.

After these preprocessing steps, the final dataset contained a set of variables that were most relevant for predicting accident severity. These included:

- **Accident severity**: fatal/serious or slight.
- **Road conditions**: type, speed limit, weather, and lighting.
- **Driver attributes**: age and gender.
- **Vehicle information**: type, number of vehicles involved, skidding/overturning, etc.
- **Presence of pedestrian crossing and carriageway hazards**

The objective of this supervised learning analysis is to predict whether an accident with certain characteristics is likely to be severe or not. To achieve this, I used various supervised learning algorithms, including logistic regression, random forest, and XGBoosting. Additionally, I employed data balancing techniques to address potential class imbalance in the target variable.

3.2 Exploratory Data Analysis

The exploratory data analysis (EDA) examines various factors that may influence accident severity. The analysis focuses on several key points: the distribution of accidents by month, day of the week, time of day, accident severity, and speed limit. The following sections discuss the insights gained from these analyses.

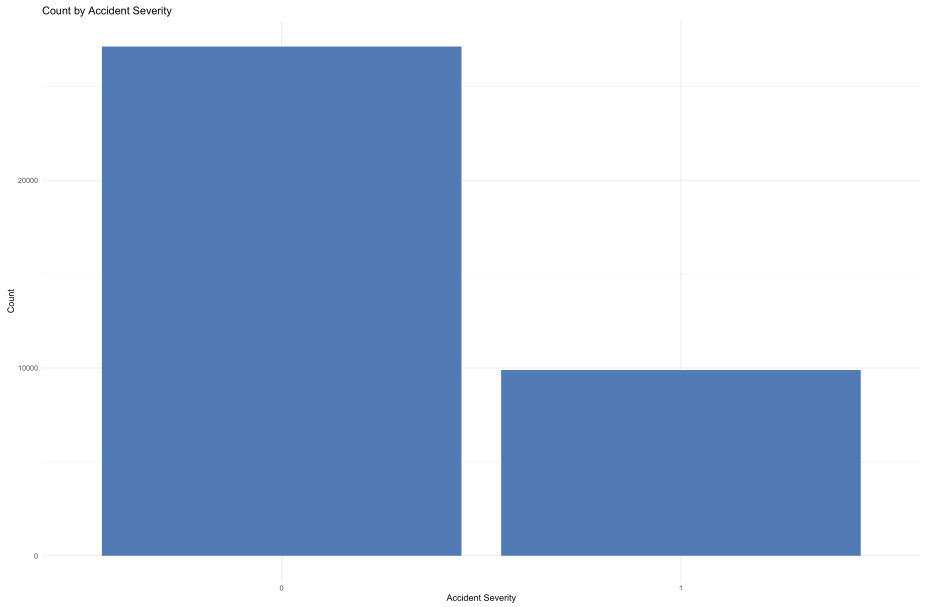

Figure 17: Count by Accident Severity

The distribution of accident severity shows that most accidents are classified as ”slight,” with fewer classified as ”severe.” This difference in the data indicates that safety measures might be reducing the number of severe accidents. This class imbalance is critical to address when developing supervised learning models. To make sure that the models are effective and unbiased, I used techniques to handle class imbalance. This will help to bring better learning and prediction outcomes.

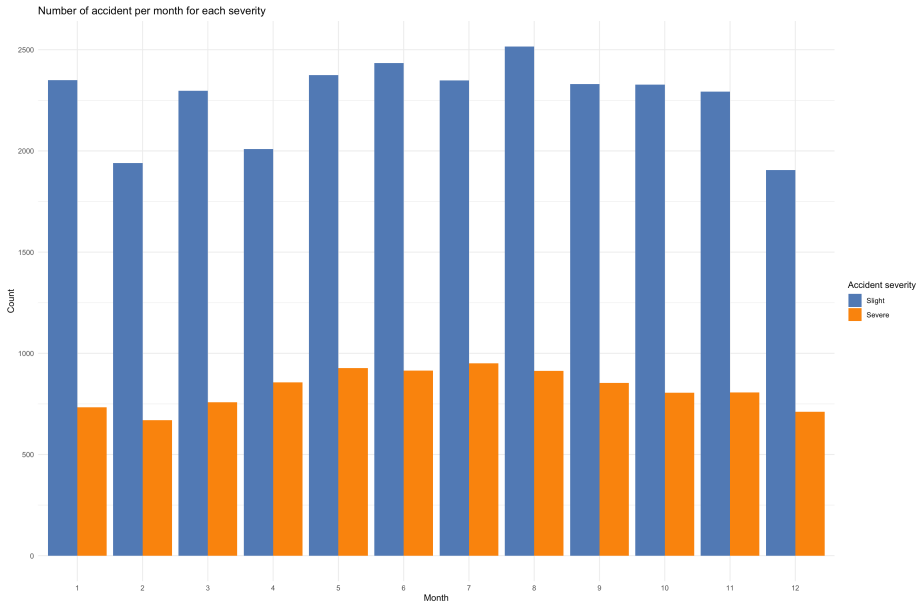

Figure 18: Accidents by Month for Each Severity

This bar chart shows the number of accidents per month, categorized by severity. The pattern indicates that severe accidents generally follow a similar distribution as slight accidents. The severity distribution across the year seems fairly consistent, indicating that no specific month has a significantly higher rate of severe accidents.

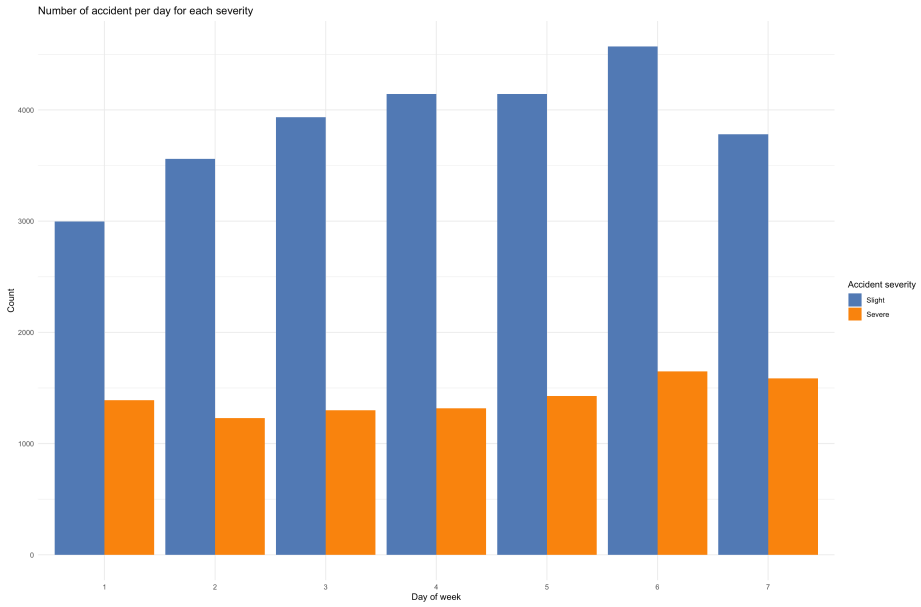

Figure 19: Accidents by Day of the Week for Each Severity

This chart shows the distribution of accidents by day of the week by sever-

ity. The pattern suggests that Saturday has the most accidents, with both slight and severe accidents peaking. This trend might be due to increased travel activities. Alcohol drinking on Saturday could also be influencing the number.

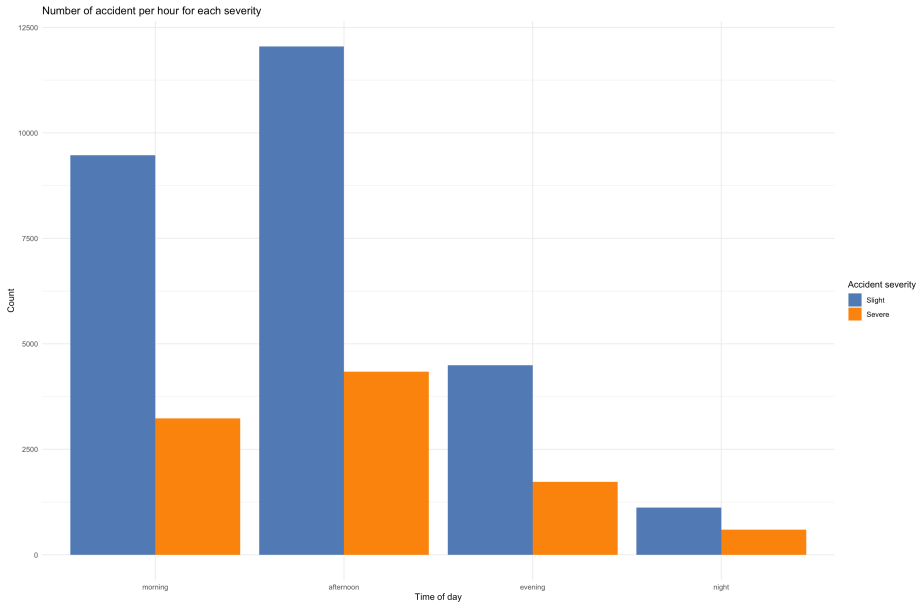

Figure 20: Accidents by Time of Day for Each Severity

Examining accidents by time of day, the data indicates that the afternoon has the highest number of accidents, with both slight and severe accidents peaking during this period. This could align with higher traffic volumes during afternoon rush hour and increased activity levels.

The morning also shows a significant count of accidents, suggesting that morning rush hour could contribute to the frequency of road incidents. Both slight and severe accidents follow a similar trend, peaking in the afternoon, then gradually declining through the evening and night.

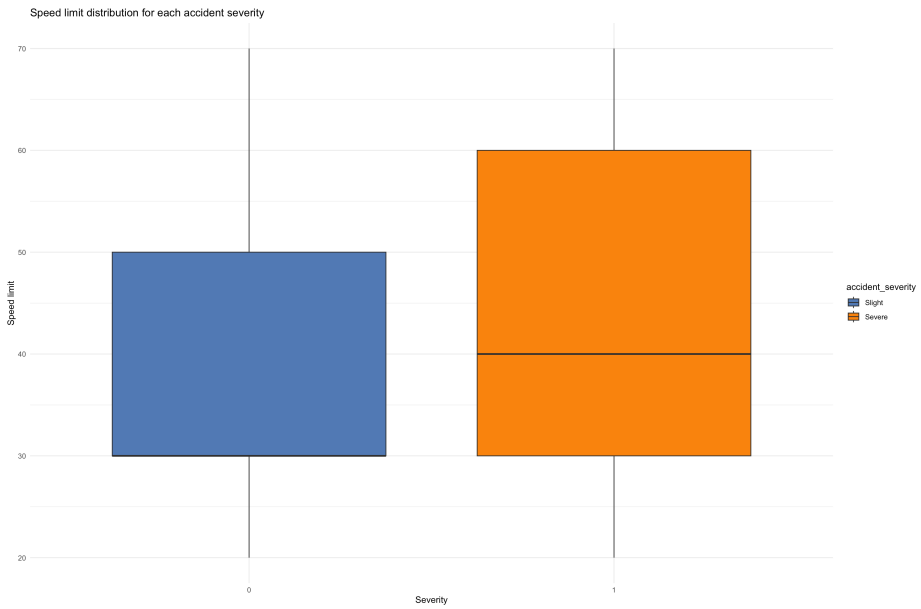

Figure 21: Speed Limit Distribution for Each Accident Severity

The boxplot showing speed limits for each accident severity reveals that most accidents occur in areas with speed limits of 30 or 60 Km/h. The data suggests that while slight accidents are more common at lower speeds, severe accidents might occur at higher speeds, possibly due to increased impact force or other risk factors.

These analyses provide a comprehensive view of the distribution and characteristics of traffic accidents, segmented by severity.

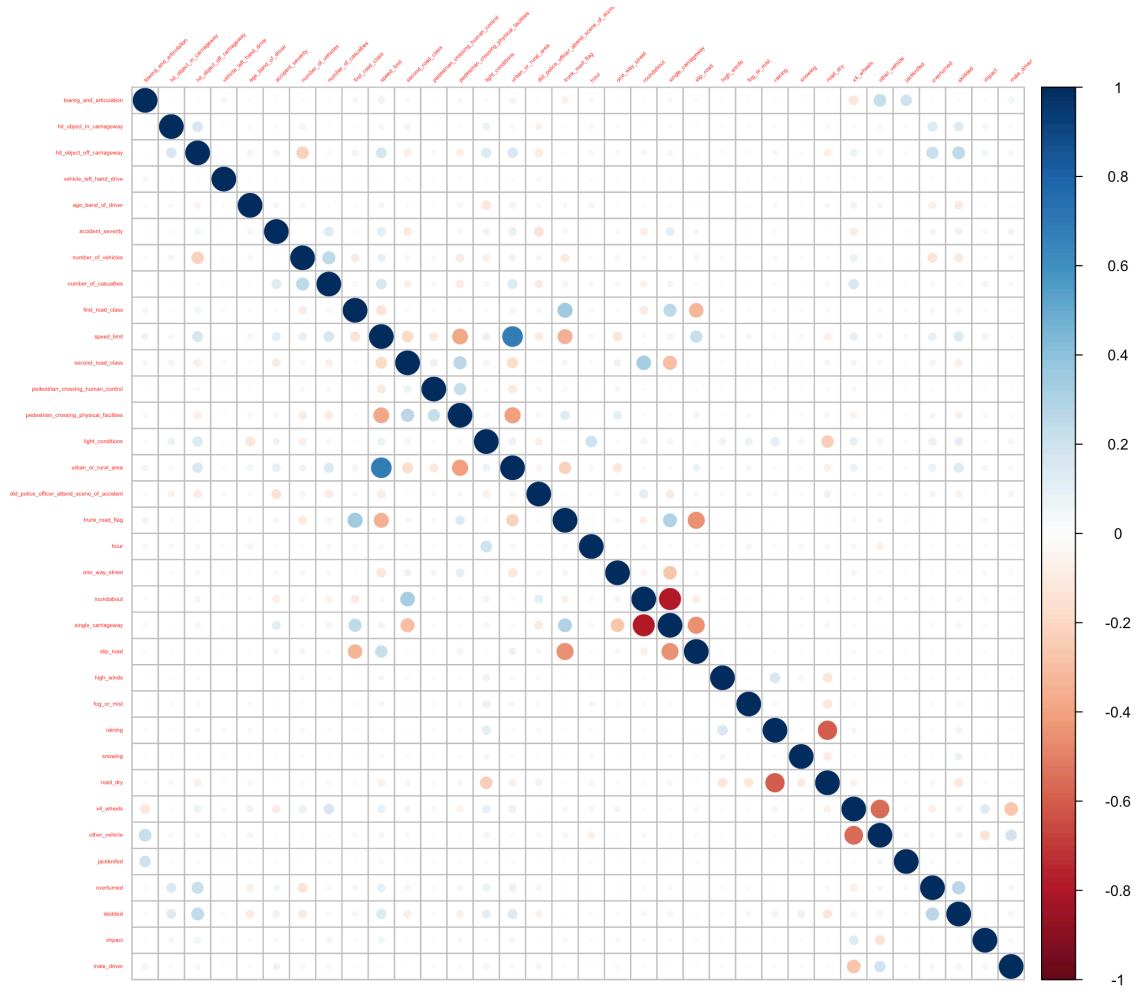

Figure 22: Correlation Matrix

Observing the correlation matrix, we can identify a few notable relationships between the variables and accident severity:

- **Speed Limit**: The correlation between speed limit and accident severity is moderately positive, suggesting that higher speed limits may contribute to more severe accidents as seen before.
- **Urban or Rural Area**: This variable shows a positive correlation with accident severity, indicating that accidents in rural areas tend to be more severe. This may be due to higher speed limits and longer ambulance' response time in rural locations.
- **Number of Casualties**: There's a positive correlation between the number of casualties and accident severity, indicating that more severe accidents tend to have more casualties.

- **Pedestrian Crossing Physical Facilities**: This variable has a moderate negative correlation with accident severity, suggesting that the presence may reduce the severity of accidents probably because of the reduction of the speed.
- **Light Conditions**: The correlation between light conditions and accident severity indicates that darker conditions may be associated with more severe accidents, emphasizing the importance of proper lighting for road safety.

These insights can guide further analysis and help develop strategies to reduce accident severity.

3.3 Supervised Learning

The dataset is imbalanced, with more accidents classified as ”slight” compared to those classified as ”severe.” This imbalance creates challenges for supervised learning algorithms, as they tend to be biased towards the majority class. This is particularly concerning in this context, where identifying severe accidents is crucial.

I decided to apply two specific techniques:

- **Oversampling**: This technique increases the number of samples in the minority class by duplicating existing samples. It is a simple approach to balance the data but can lead to overfitting.
- **SMOTE (Synthetic Minority Oversampling Technique)**: This method generates synthetic samples in the minority class by interpolating between existing data points. It's less prone to overfitting than simple oversampling.

These techniques were used in combination with different supervised learning algorithms to evaluate their effectiveness in predicting accident severity. The primary metrics analyzed were:

- **Accuracy** $= \frac{TP+TN}{TP+TN+FP+FN}$, measures the proportion of correct predictions out of the total predictions. Accuracy might be misleading in imbalanced datasets, as the majority class can dominate, making the accuracy appear high even if the minority class is poorly predicted.
- **Precision** $= \frac{TP}{TP+FP}$, indicates the proportion of true positives among all positive predictions. Precision, in imbalanced datasets, is crucial for understanding the proportion of correct positive predictions, but focusing only on precision may lead to missing many true positives.
- **Recall/Sensitivity** $= \frac{TP}{TP+FN}$, represents the proportion of true positives among all actual positives. Recall is crucial in imbalanced datasets, especially when it's essential to identify the minority class.
- **F1 Score** $= 2\frac{Precision \times Recall}{Precision+Recall}$, is the harmonic mean of precision and recall, providing a single metric that balances both. This score is valuable in imbalanced datasets because it considers both precision and recall, offering a more balanced view.

- **AUC (Area Under the ROC Curve)** reflects the ability of the model to distinguish between classes, with higher values indicating better separation. AUC is less affected by class imbalance, making it a reliable metric for evaluating model performance in imbalanced datasets.

Due to the large size of the dataset and the number of attributes, advanced tuning, such as altering the number of trees and the mtry parameter, was not performed to maintain computational efficiency.

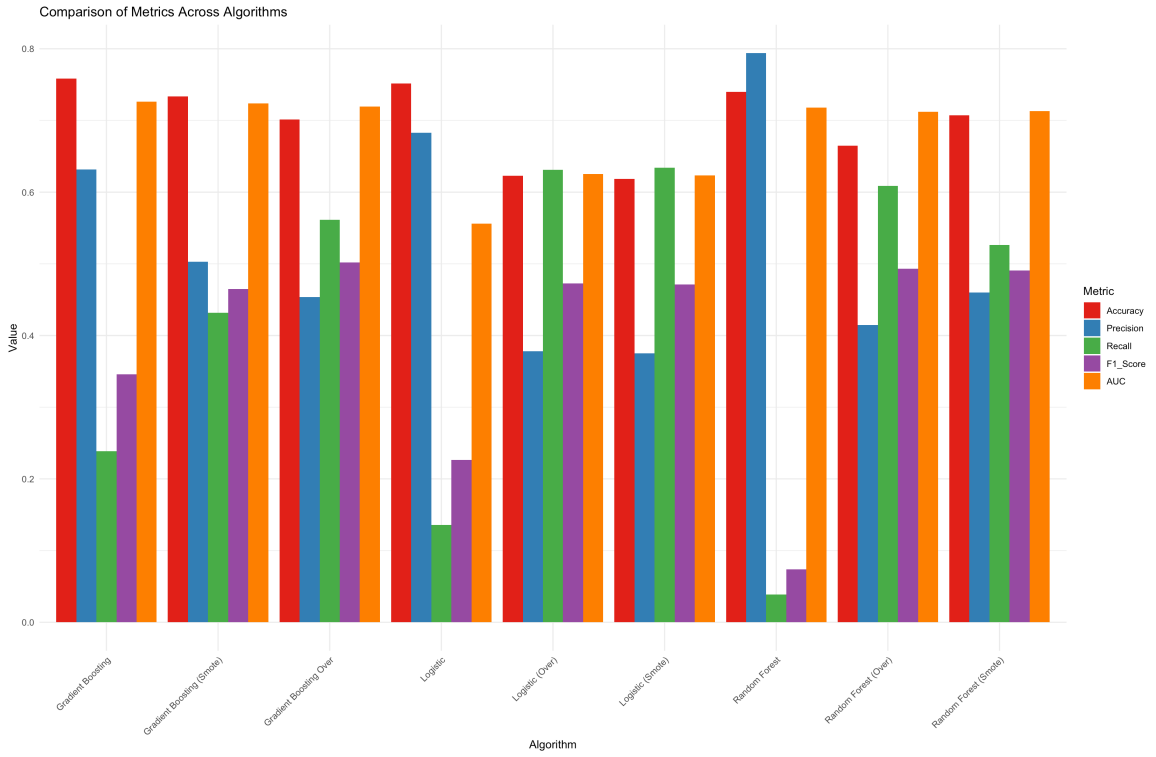

Figure 23: Results of Various Algorithms and Sampling Techniques

Examining the results, here are the key findings for each algorithm:

- **Logistic Regression**: The logistic model has high accuracy but low recall, suggesting that it struggles to identify severe accidents. Both oversampling and SMOTE improved recall but at the expense of accuracy.
- **Random Forest**: This model demonstrates high accuracy but low recall. The F1 Score increases when using oversampling or SMOTE, indicating a better balance between precision and recall.
- **Gradient Boosting**: This algorithm delivers the highest accuracy among the models. Although its recall is lower, it maintains a reasonable balance

between precision and recall, providing a higher F1 Score in the balanced training than other models.

3.3.1 Conclusion

The best overall result, in terms of a balanced approach to accuracy, recall, and F1 Score, came from Gradient Boosting. While data balancing techniques, such as oversampling and SMOTE, help improve recall and F1 Score, they can slightly reduce accuracy.

I am pleased with the improvement in results achieved by balancing the data, but I believe that better results could be achieved with more advanced parameter tuning.

4 Dataset

[1] World Healt Organization. *Road safety*. URL: https://www.who.int/data/gho/data/themes/road-safety.

[2] GOV.UK. *Road accident*. URL: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data.