Road Traffic Accidents Comprehensive Report

INTRODUCTION

Road traffic accidents (RTAs) are a significant public health issue in the United Kingdom, leading to considerable social and economic impacts (Department for Transport, 2021). In 2020 alone, there were approximately 115,584 reported road casualties of all severities, a decrease of 25% compared to 2019, largely due to the impact of the COVID-19 pandemic and associated restrictions (Department for Transport, 2021). Despite this decrease, RTAs remain a leading cause of death and serious injury, particularly among young adults and vulnerable road users such as pedestrians and cyclists. The objective of this report is to gain some understanding into the contributing factors for road accidents by analyzing the data we have, through which we aim to glean insights that can inform the development and implementation of effective road safety measures.

DATA DESCRIPTION

The dataset used in this project is sourced from the UK government's annual release of road traffic accident data. This comprehensive dataset logs all road traffic accidents involving casualties in Great Britain, along with a majority of other non-fatal road traffic accidents. For the purpose of this study, we focus on the data from the year 2020.

The data is stored in a SQLite database named accident_data_v1.0.0_2023.db. This database contains detailed information about each accident event, including the involved vehicles and casualties, as well as geographic information at the Lower Layer Super Output Area (LSOA) level.

METHODOLOGY

The methodology for this project involves several stages of data extraction and preprocessing to ensure the data is suitable for analysis.

Data Extraction

The accident data for the year 2020 is extracted from the SQLite database. This involves querying the database to retrieve the relevant data, which includes details on accident events, involved vehicles and casualties, as well as geographic information at the Lower Layer Super Output Area (LSOA) level.

Merging and Filtering Dataframes

The merging process involves combining dataframes based on common identifiers, while the filtering process involves selecting relevant records based on specific criteria.

Data Cleaning and Preliminary Processing

Once the data is extracted and consolidated, we carried out a series of cleaning and preprocessing steps. Including:

- **Validating Data Types in the DataFrame**: Ensuring that each column in the dataframe has the correct data type (e.g., numerical, categorical) is crucial for subsequent analysis and modeling tasks.
- **Addressing Duplicate Entries**: Duplicate entries in the dataset are identified and removed to prevent skewing the analysis results.
- **Addressing Null Values**: Null values in the dataset are identified and handled appropriately, either by filling them with appropriate values (e.g., mean, median, mode) or by removing the records containing them, depending on the nature and extent of the missing data.
- **Correcting Anomalies in the Numerical Features**: Anomalies or outliers in the numerical features are identified and corrected. This could involve applying transformations to the data or replacing the anomalous values with more representative ones.

ANALYTICAL TECHNIQUES

This project employs a range of analytical techniques to uncover patterns and risk factors influencing road accident severity. These techniques include association rule learning, clustering, outlier detection and classification modeling.

Association Rule Learning (Apriori Algorithm)

The Apriori algorithm operates on the principle that an itemset cannot be frequent unless all its subsets are frequent (Toivonen, 2023). In the context of this project, the Apriori algorithm is used to uncover relationships between different variables in the accident data.

Clustering

By grouping accidents based on their geographic coordinates and other relevant features, we can identify areas with high accident rates or particular types of accidents. K-means was used for this.

Outlier Detection

Outlier detection was performed on the accident data to identify unusual entries. This was achieved using the Isolation Forest algorithm, which is particularly effective for detecting outliers in high-dimensional datasets.

Classification Modeling

We developed and compared classification models to predict the severity of injuries sustained in road traffic accidents between fatal and non-fatal accidents.

RESULTS AND DISCUSSION

Temporal Patterns

The analysis of the accident data reveals several temporal patterns related to the frequency and timing of accidents. Some general trends include:

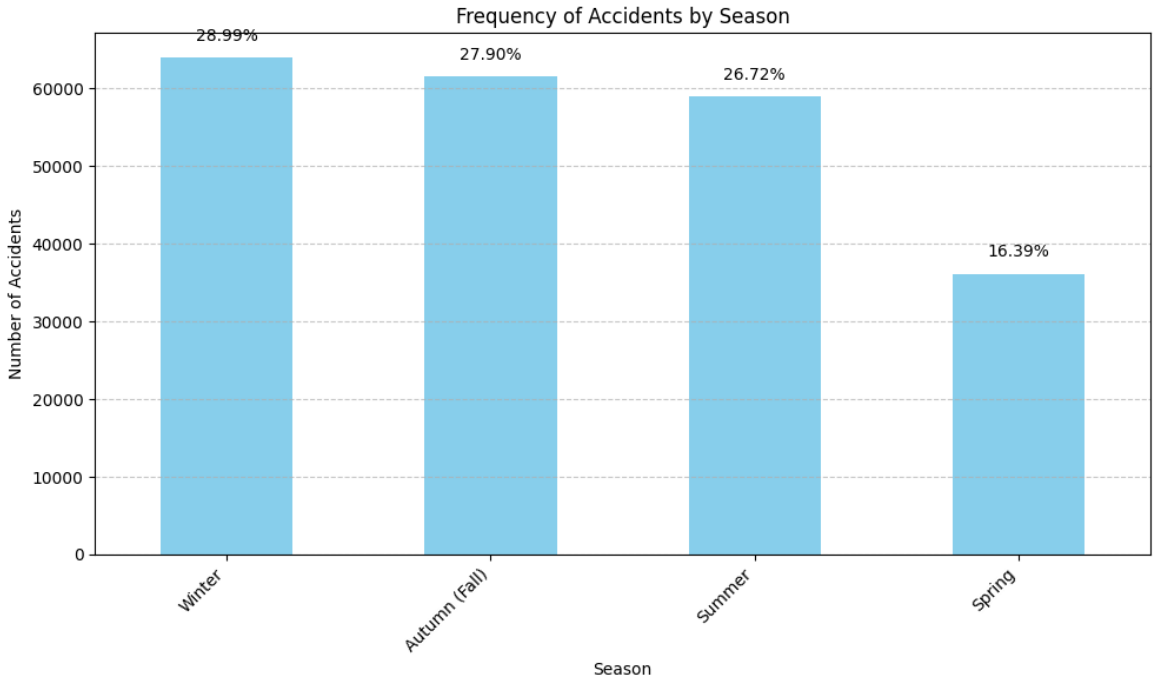

Figure 1 Plot of Frequency of Accidents by Season

Accidents are most prevalent during the winter season, accounting for 28.99% of all accidents, closely followed by the autumn season at 27.90%. This could be attributed to the challenging driving conditions during these seasons, such as reduced daylight hours, wet or icy roads, and poor visibility (World Health Organization, 2013).

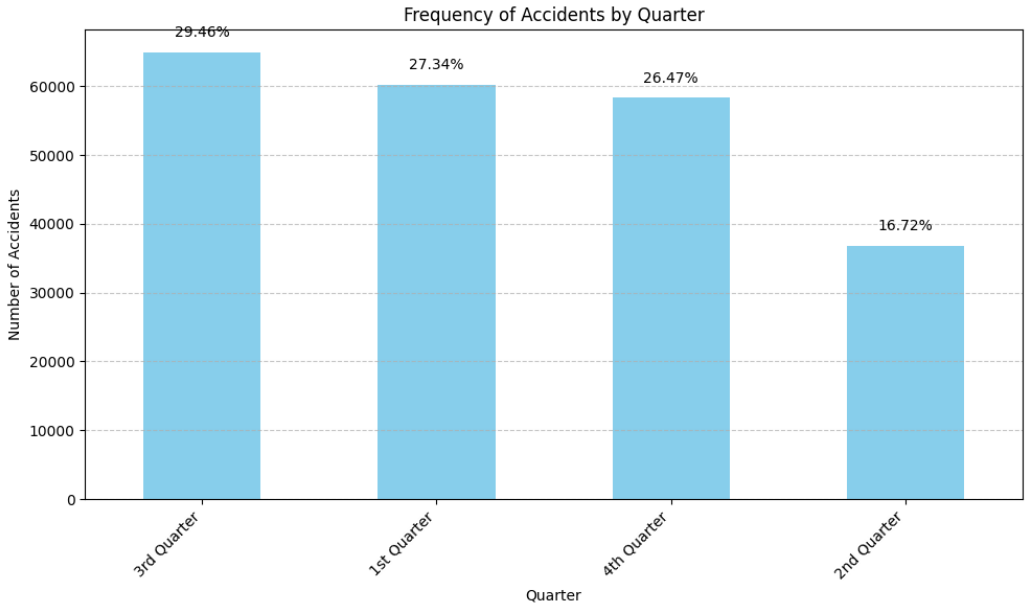

Figure 2: Frequency of Accidents by Quarter

In terms of quarters, the third quarter of the year stands out as the period with the highest accident frequency, representing 29.46% of all accidents. This could be due to increased travel during the summer holidays (Ritchie, 2018). The first quarter follows closely behind at 27.34%.

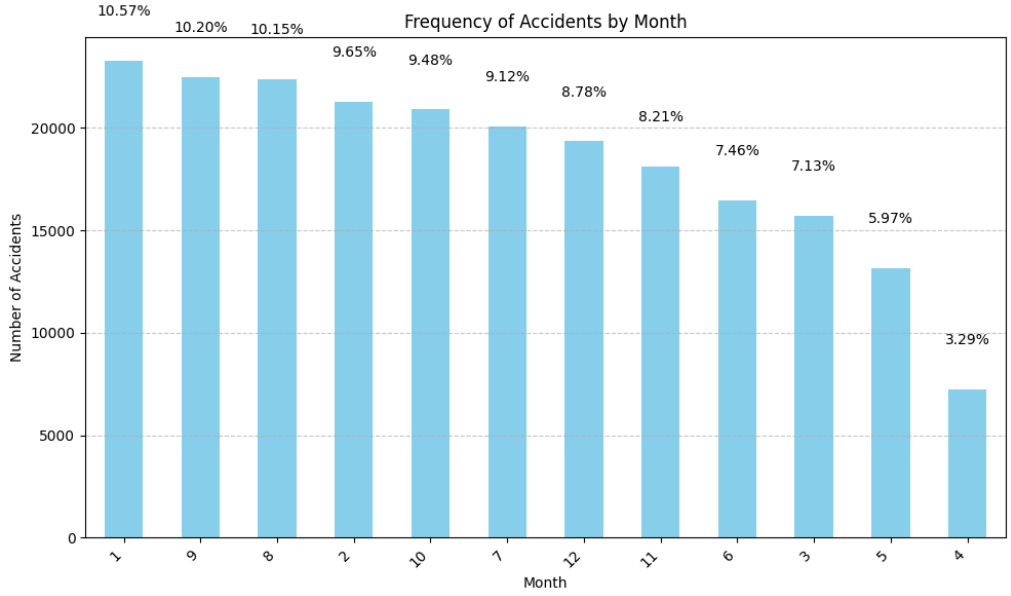

Figure 3 Frequency of Accidents by Month

The analysis also shows that January has the highest accident frequency at 10.57%, with September ranking second at 10.20%. Accidents are most likely to occur during the third week of the month (23.34%), followed by the second week (23.21%).

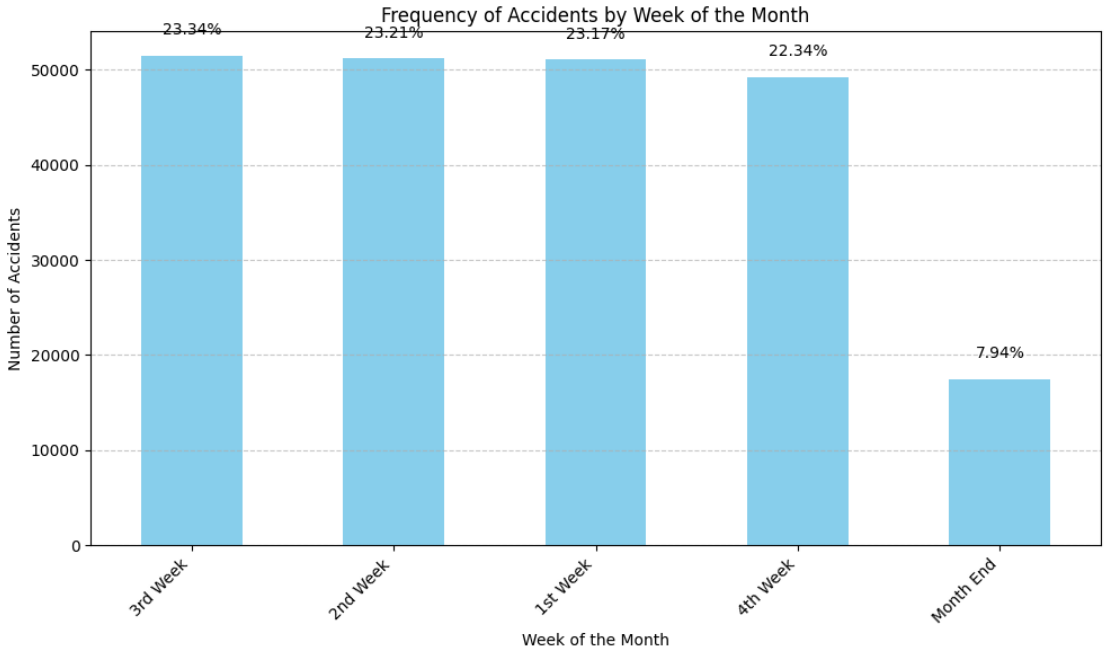

Figure 4 Frequency of Accidents by Week of the Month

The data indicates that Saturdays exhibit the highest accident frequency among weekdays, with 16.30% of accidents occurring on this day, while Fridays follow closely behind at 15.17%. This could be due to increased travel during the weekend (Ritchie, 2018).

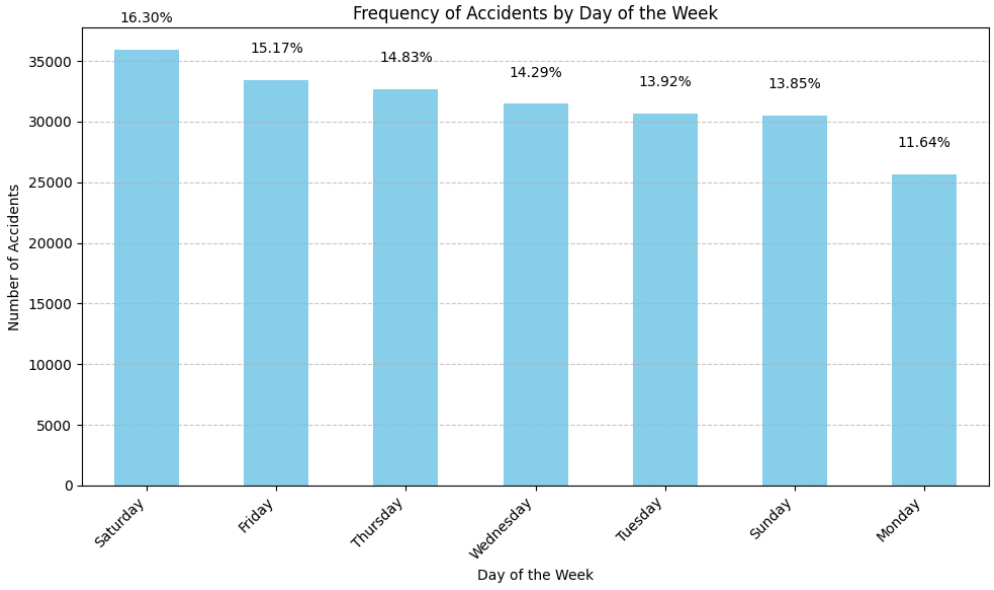

Figure 5 Frequency of Accidents by Day of the Week

During the day, the afternoon session accounts for the highest proportion of accidents at 27.86%, followed by the evening rush period at 24.43%. Finally, accidents peak at 17:00, representing 8.70% of all accidents, with 16:00 closely following at 8.43%.

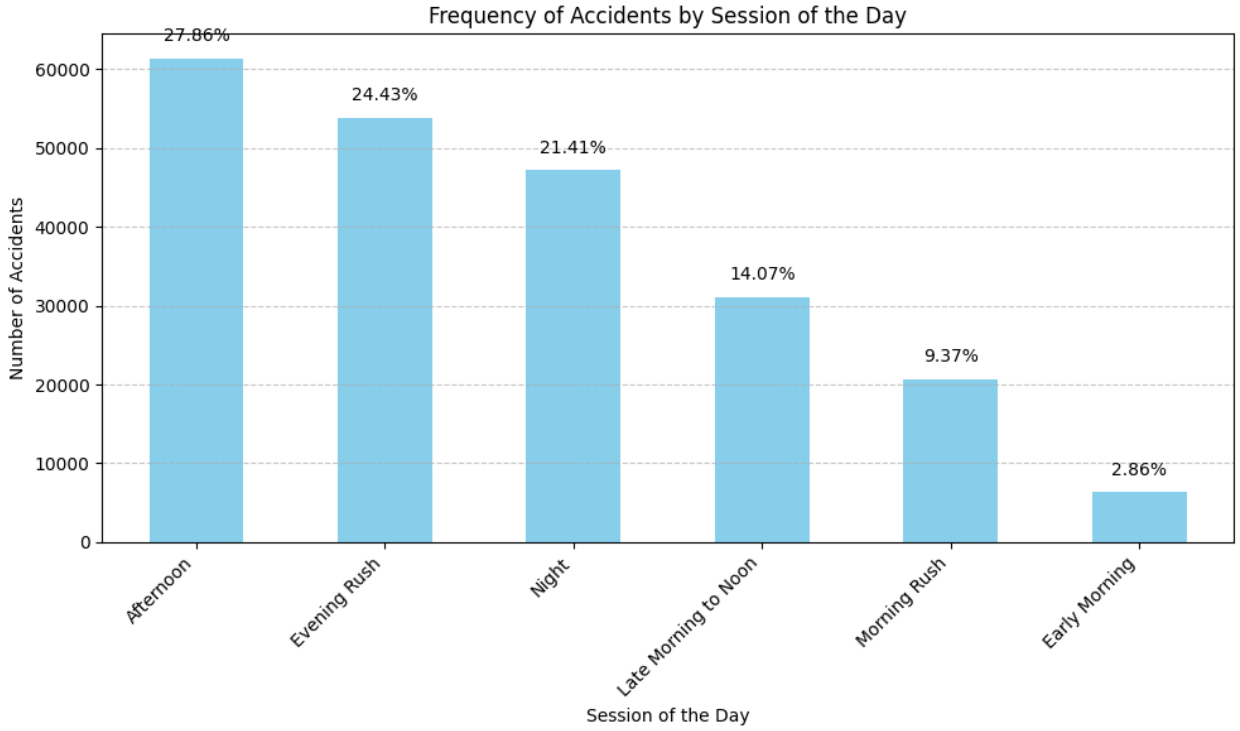

Figure 6 Frequency of Accidents by Session of the Day

Specific Vehicle Types and Road Users

When analyzing specific vehicle types and road users, some interesting patterns emerged.

For motorbikes, accidents are most prevalent during the summer season, contributing to 29.01% of all incidents, closely followed by autumn at 27.92%. This could be due to increased motorcycle use during warmer weather (Clarke et al., 2004). Motorcycle accidents peak during the third quarter of the year, comprising 30.51% of all incidents. The second week of the month sees the highest proportion of motorcycle accidents at 26.40%, and Saturdays top the list as the day with the highest proportion of motorcycle accidents, representing 16.21%. Afternoons witness the highest proportion of motorcycle accidents, constituting 29.09% of all incidents, and motorcycle accidents peak at 17:00, accounting for 10.10% of all incidents.

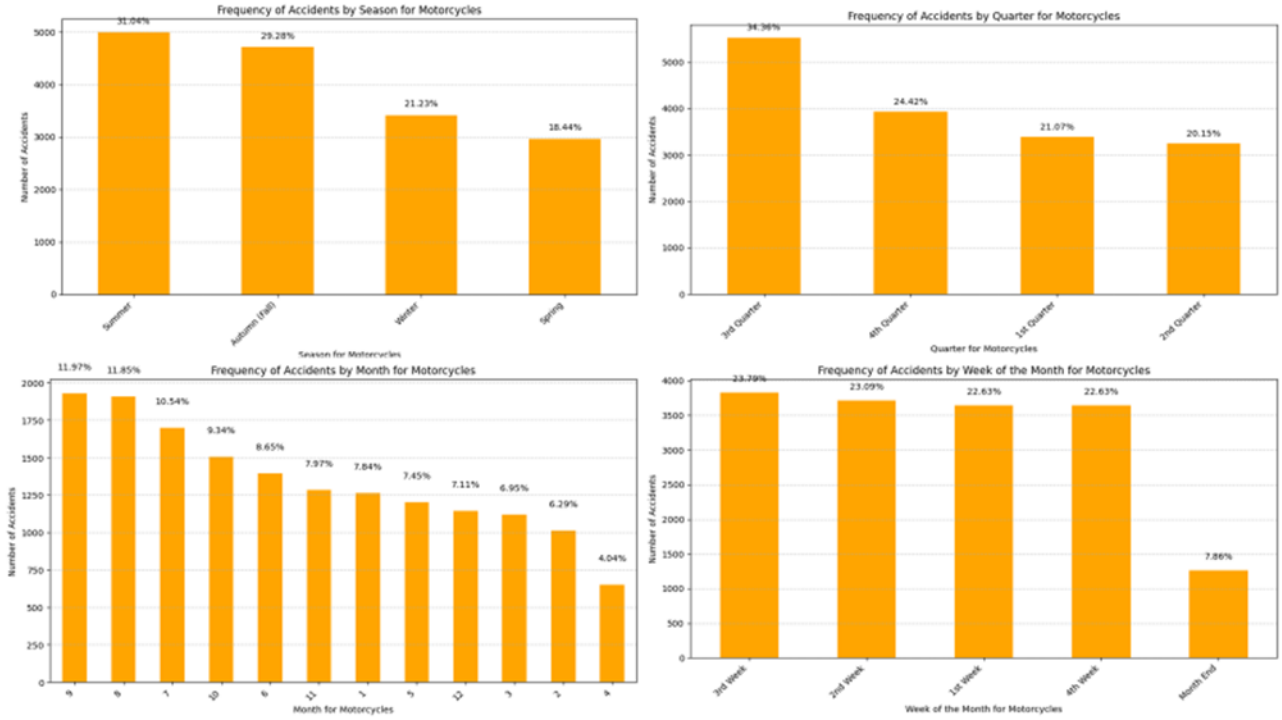

Figure 7 Accident Frequencies across Time for Motorcycles

For pedestrians, the data reveals that accidents follow similar trends to the general patterns, with the highest occurrence in winter (30.94%), the first quarter of the year (29.33%), and on Saturdays (17.26%). However, the highest hourly frequency for pedestrian accidents is at 3:00 PM (11.14%), which could be due to school finishing times (Davies et al., 1996).

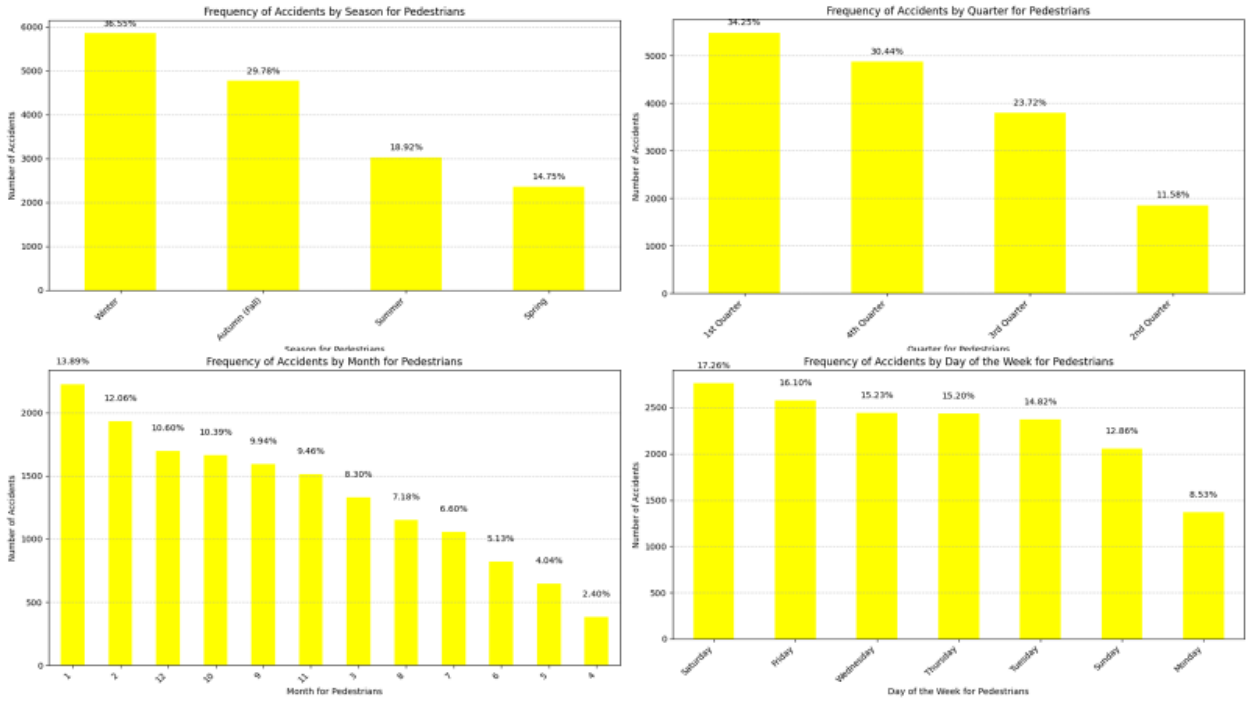

Figure 8 Accident Frequencies across Time for Pedestrians

Driver Characteristics and Accident Severity

Driver characteristics such as age, gender, and experience play a significant role in road traffic accidents. Younger drivers, particularly those in their teens and early twenties, are often overrepresented in traffic accidents as shown in the plot below.

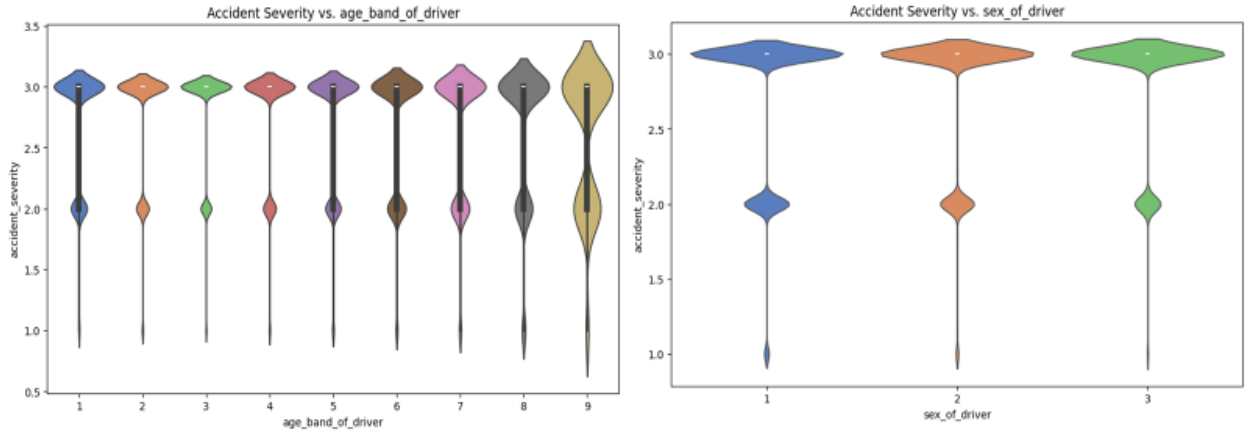

Figure 9 Accident Severity by Age Band of Driver and Sex of Driver

This can be attributed to a combination of factors, including lack of driving experience, risk-taking behavior, and susceptibility to distractions (Williams, 2003). On the other hand, older drivers may also be at a higher risk due to declining cognitive and physical abilities (Langford & Koppel, 2006).

Vehicle Characteristics and Accident Severity

Vehicle characteristics such as type, age, and size can significantly influence the likelihood and severity of road traffic accidents.

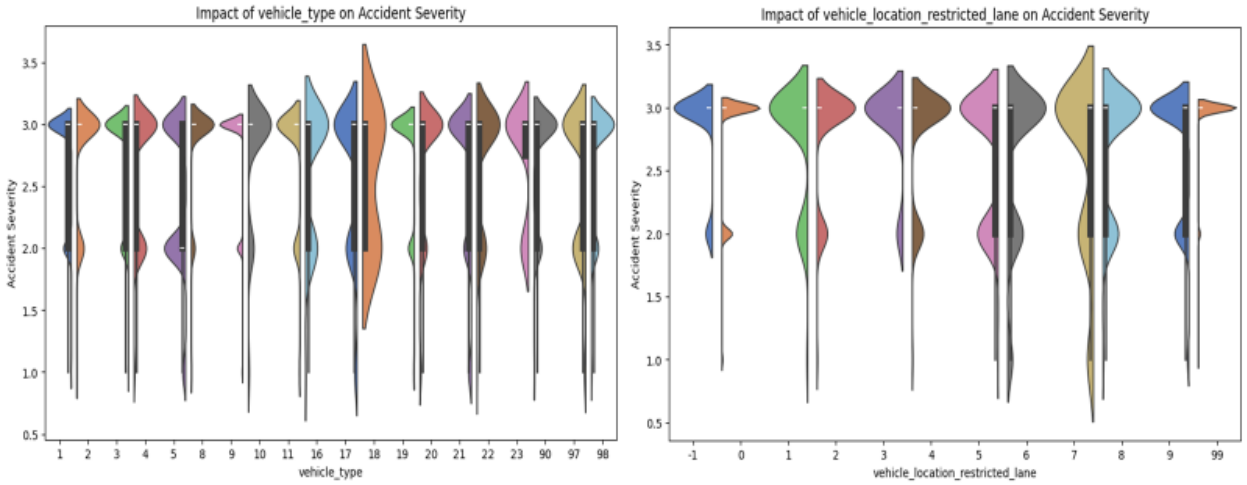

Figure 10 Impact of vehicle characteristics on Accident Severity

The violin plot in the provided image above illustrates the impact of different vehicle types on accident severity. This visualization provides insights into the distribution and density of accident severity for each type of vehicle. From the plot, it can be inferred that different types of vehicles have different distributions of accident severity.

Association Rules

1. **Association Rules for Fatal Accident Severity:**

   No association rules could be generated for fatal accident severity. This could be due to the low frequency of fatal accidents in the dataset, making it challenging for the Apriori algorithm to find

meaningful associations. However, considering more features could help fish out rules, though this would require a powerful processing unit.

2. **Association Rules for Serious Accident Severity:**

Several interesting patterns emerge from these rules:

a. *Vehicle maneuver of "Reversing" (vehicle_manoeuvre_18.0)* has a strong association with serious accidents, either alone or combined with other factors like urban areas.

b. *Poor weather conditions* (weather_conditions_1.0) and poor road surface conditions (road_surface_conditions_1.0) are associated with serious accidents.

c. *Driving during poor light conditions* (light_conditions_4.0) is linked to serious accidents, especially in urban areas.

d. *Male drivers* (sex_of_driver_1.0) are associated with serious accidents on certain road classes (first_road_class_3.0).

These associations highlight the impact of various factors, such as driver behavior, environmental conditions, and road characteristics, on the likelihood of serious accidents.

3. **Association Rules for Slight Accident Severity**:

Interestingly, the top 10 association rules for slight accident severity are identical to those for serious accident severity. This suggests that the same factors may contribute to both slight and serious accidents, but the degree of severity may depend on other variables not captured in this analysis.

Clustering Analysis

Based on the clustering analysis performed on the Humberside region data, the following observations can be made:

The optimal number of clusters for this dataset was determined to be 6, based on the Silhouette score method. From the clusters, we see that accidents the region can be grouped into six distinct areas.

a. *Royal Dock,* ***Grimsby***: This cluster represents a concentration of accidents around the Royal Dock area near the port city of Grimsby. This could be attributed to factors such as high traffic volumes, congestion, or potential issues with road infrastructure in this industrial and maritime area.

b. ***Scunthorpe***: The second cluster is centered around the town of Scunthorpe, likely indicating a higher accident rate in and around this urban area.

c. ***Old Goole***: The third cluster is located in the Old Goole area, suggesting a higher incidence of accidents in this part of the region. This could be related to factors such as road conditions, traffic management, or other local factors unique to this area.

d. *Princess Avenue,* ***Hull***: The fourth cluster is centered around Princess Avenue in Hull, indicating a higher concentration of accidents in this specific location within the city. This could be influenced by factors such as road layout, traffic volumes, or the presence of intersections or other road features in this area.

e. *Between* ***Shiptonthorpe*** *and* ***Londesborough***: The fifth cluster is situated between the villages of Shiptonthorpe and Londesborough, potentially highlighting a higher accident rate on rural roads or highways connecting these areas. Factors such as speed limits, road conditions, or driver behavior on these routes may contribute to this cluster.

f. Carnaby by **Bridlington**: The sixth cluster is located near the village of Carnaby, close to the coastal town of Bridlington. This could indicate a higher accident rate on roads leading to or from this tourist destination, potentially influenced by factors such as increased traffic during peak seasons or road conditions in this area.

Outlier Detection

A total of 9,986 observations were identified as outliers based on their longitude and latitude coordinates. These outliers are displayed on a scatter plot, where they are represented by larger red dots, contrasting with the normal data points shown as smaller blue dots.

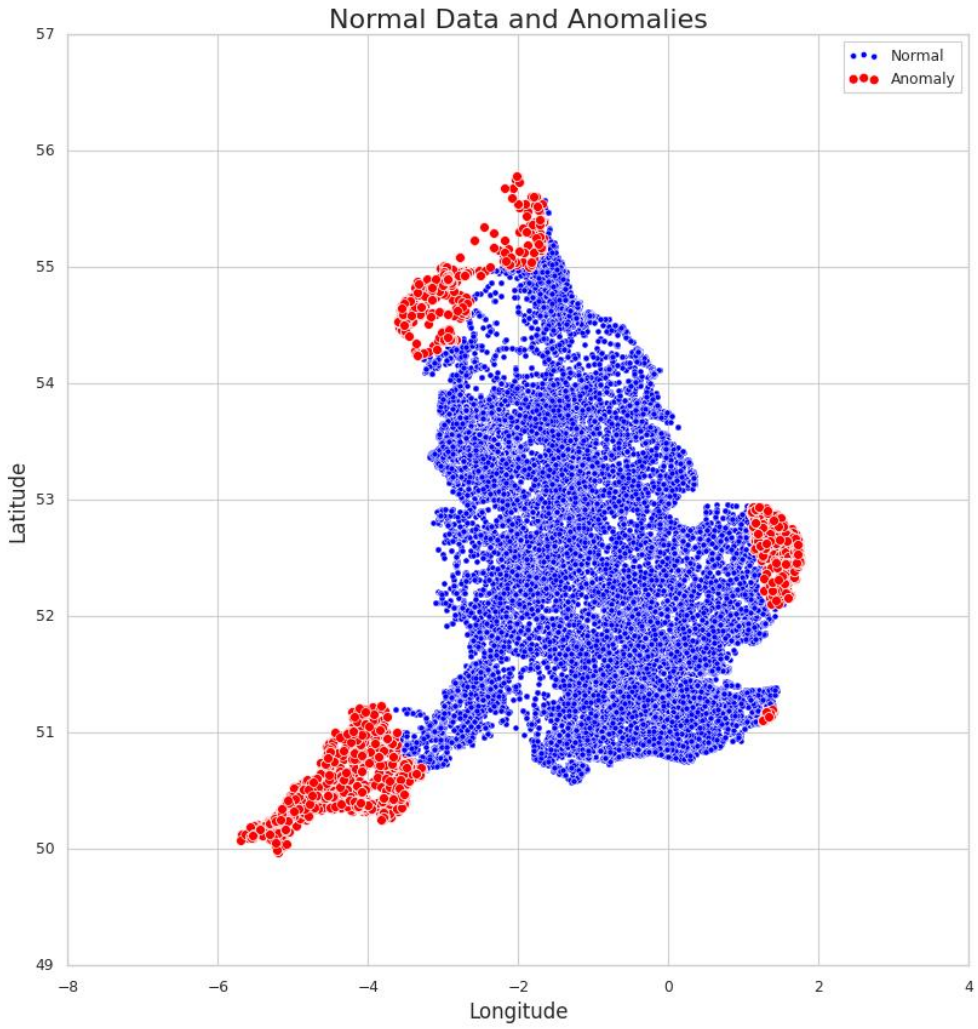

Figure 11 Outliers Detected (Using Isolation Forest) shown by red dots on the plot

The scatter plot reveals a nice spatial pattern. The majority of the normal data points form a dense cluster resembling the shape of England, suggesting that the dataset primarily covers accidents within a specific geographical region. However, the outliers are scattered around the periphery of this main cluster and across a wider geographical area, potentially representing data points from different regions or countries. The presence of these outliers could be attributed to various factors, such as data entry errors, exceptional cases, or missing or incomplete data, which is logical given the different imputation strategies we had to adopt for this analysis.

Classification Models

Table 1: Table of Accident Fatality Classification Models

Model	**Mean Accuracy**	**Standard Deviation**
Random Forest	0.96	0.022
XGBoost	0.96	0.026
Gradient Boosting	0.95	0.025
KNN	0.70	0.061
SVM	0.51	0.005
Stacked Model	0.96	0.025

From the model performance values obtained, the Random Forest, XGBoost, and Gradient Boosting models achieved the highest mean accuracy scores, all hovering around 0.96, indicating a high level of predictive performance. The KNN model achieved a lower mean accuracy of 0.70, while the SVM model exhibited the lowest mean accuracy of 0.51. A stacked model was also explored, which combined the predictions of multiple models to generate a final prediction. This stacked model achieved a mean accuracy of 0.96, matching the performance of the best individual models.

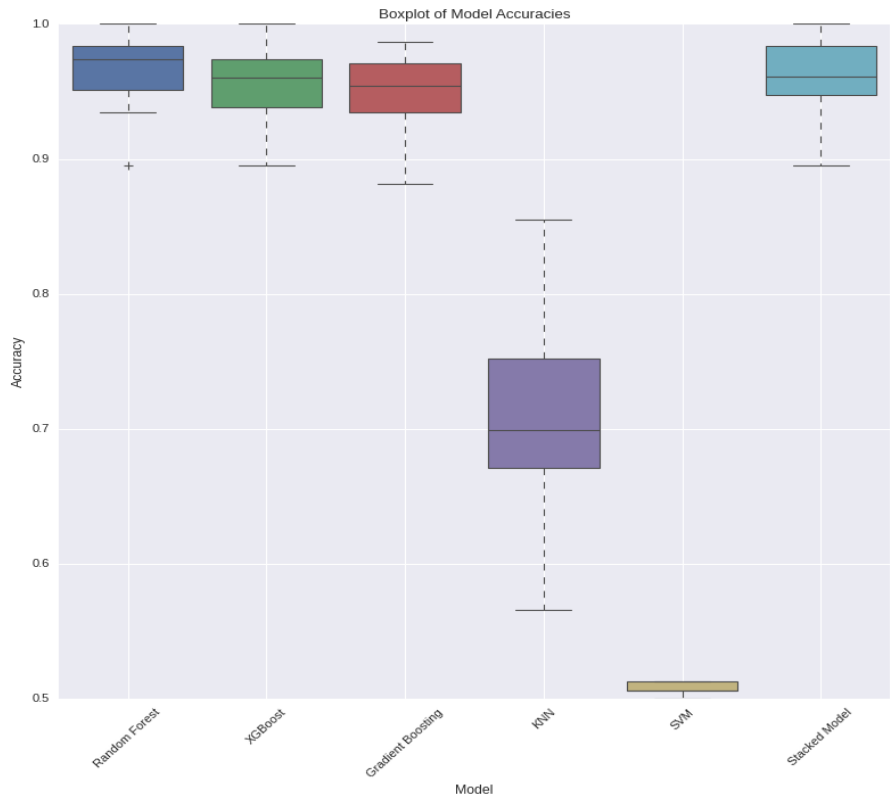

Figure 12 Boxplot of Model Accuracies

Conclusion

The findings suggest that machine learning models, particularly ensemble models like Random Forest, XGBoost, and Gradient Boosting, and the stacked model can effectively predict the severity of injuries sustained in road traffic accidents. The observations here can inform decision-making key in reducing road accidents.

References

1. Clarke, D.D., Ward, P., Bartle, C. and Truman, W., 2004. In-depth accident causation study of young drivers. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 218(8), pp.909-919

2. Davies, D.G., Halliday, M.E., Mayes, M. and Pocock, R.L., 1996. Collection and analysis of accident data from police, hospital and school sources. *Child: Care, Health and Development*, 22(1), pp.21-34

3. Department for Transport, 2021. Reported road casualties Great Britain: 2020 annual report

4. Department for Transport, 2021. Road accidents and safety statistics

5. Evans, L., 2004. *Traffic Safety*. Bloomfield Hills, MI: Science Serving Society

6. Ritchie, H., 2018. Road safety. *Our World in Data*

7. Rumar, K., 2000. Relative merits of the U.S. and E.U. approaches to improve car occupant protection. *Journal of Crash Prevention and Injury Control*, 2(2), pp.79-90

8. Toivonen, H., 2023. Apriori Algorithm. In: Phung, D., Webb, G.I., Sammut, C. (eds) *Encyclopedia of Machine Learning and Data Science*. Springer, New York, NY

9. World Health Organization, 2013. *Global status report on road safety 2013: Supporting a decade of action*. World Health Organization