Predictive Analytics for Road Accidents-A Big Data Approach

Implementing predictive modeling techniques on large-scale traffic, weather, and historical accident datasets using big data frameworks

Contents

1. Introduction

The database used in this project, titled accident_data_v1.0.0_2023.db, serves as a comprehensive repository for analyzing road safety incidents. It is composed of four interconnected tables: Accident, Casualty, Vehicle, and LSOA. Each table is structured to capture specific dimensions of road accidents, with the Accident table being the central focus, containing 461,352 rows and 36 columns of detailed records about individual accidents. The Casualty table, with 600,332 rows and 19 columns, documents details about people involved in these accidents. The Vehicle table, housing 849,091 rows and 28 columns, records vehiclespecific details, while the LSOA table includes 34,378 rows and 7 columns detailing geographic areas.

The database's relational design is pivotal in facilitating complex queries and cross-referencing between datasets. Key relationships include a One-to-Many link between the Accident table and both the Casualty and Vehicle tables, reflecting that a single accident can involve multiple individuals and vehicles. Similarly, a Many-to-One relationship connects the Accident table with the LSOA table, indicating that multiple accidents can occur within the same local geographical area.

2. Dataset Cleaning

The local_authority_district column had invalid values (-1), replaced with code 37 for Suffolk and imputed using patterns in related data for others. Similarly, local_authority_highway values were corrected using local_authority_ons_district relationships. Missing speed_limit values were replaced with the modal value within each district to ensure contextually accurate imputation.

Light_conditions values (-1) were logically imputed based on the time of day, while -1 entries in weather_conditions and road_surface_conditions were set to 9 (unknown). For the lsoa_of_accident_location column, -1 was replaced with "unknown," reflecting the absence of reliable alternatives for categorical text.

3. Data Analysis

3.1.Significant Accident Times

Figure 1: Accidents by Hour of the day

The line graph showing accidents by hour of the day highlights distinct peaks during specific times. The most significant peak occurs during evening rush hours, around 16:00 to 18:00, likely due to heavy traffic as people return home from work or school. A smaller peak is observed in the morning, around 08:00, coinciding with the morning rush hour when people commute to work or school. Accidents are least frequent in the early morning hours, between 02:00 and 05:00, when road usage is minimal.

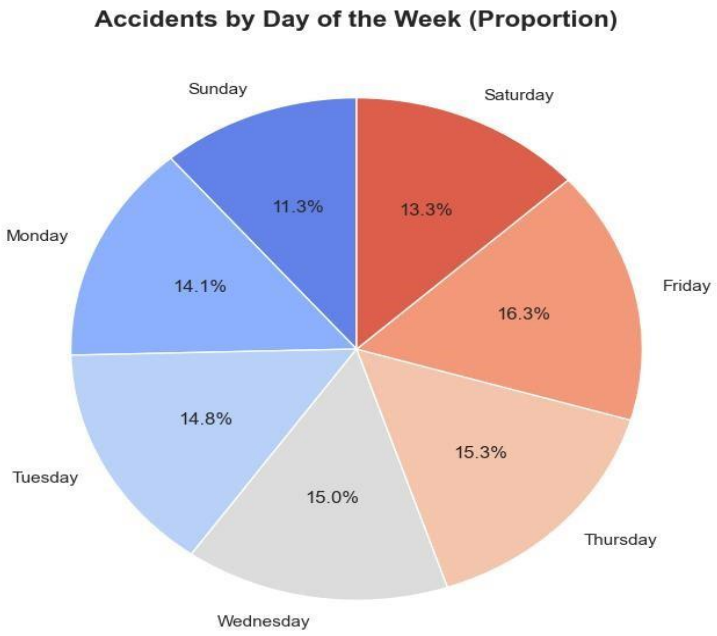

Figure 2: Accidents by Day of the Week

The pie chart shows the distribution of accidents by day, revealing Friday as the most accidentprone day, with 16.3% of the total accidents. This may result from increased traffic volume and fatigue as people wrap up their workweek. Saturday (13.3%) and Sunday (11.3%) have comparatively fewer accidents, potentially due to reduced commuter traffic.

3.2.Motorbike Accident Patterns

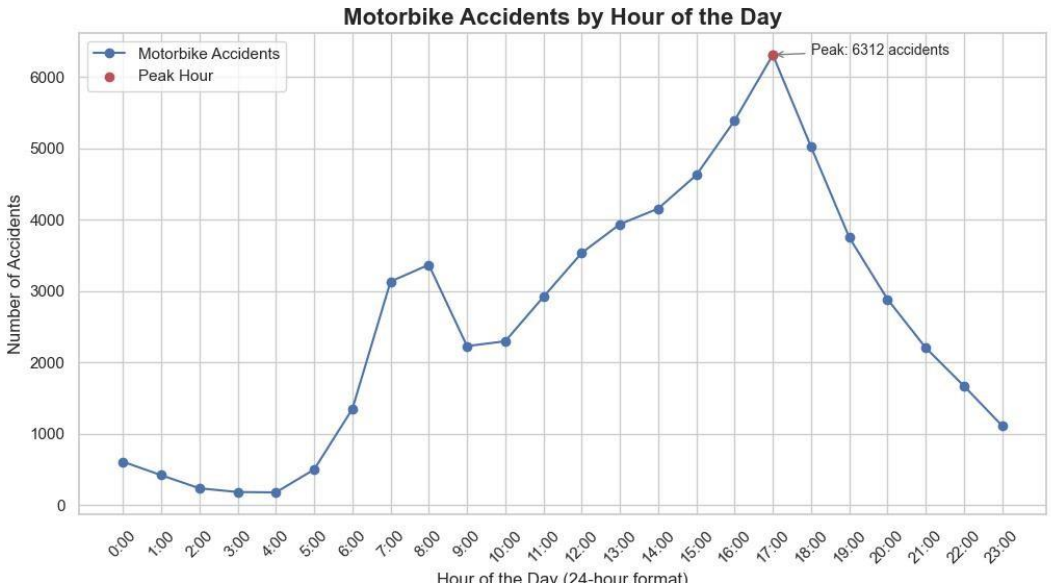

Figure 3: Motorbike Accidents by Hour of the Day

The most significant peak occurs during evening rush hours, around 17:00, likely due to heavy traffic as people return home from work or school. A smaller peak is observed in the morning, around 08:00, coinciding with the morning rush hour when people commute to work or school.

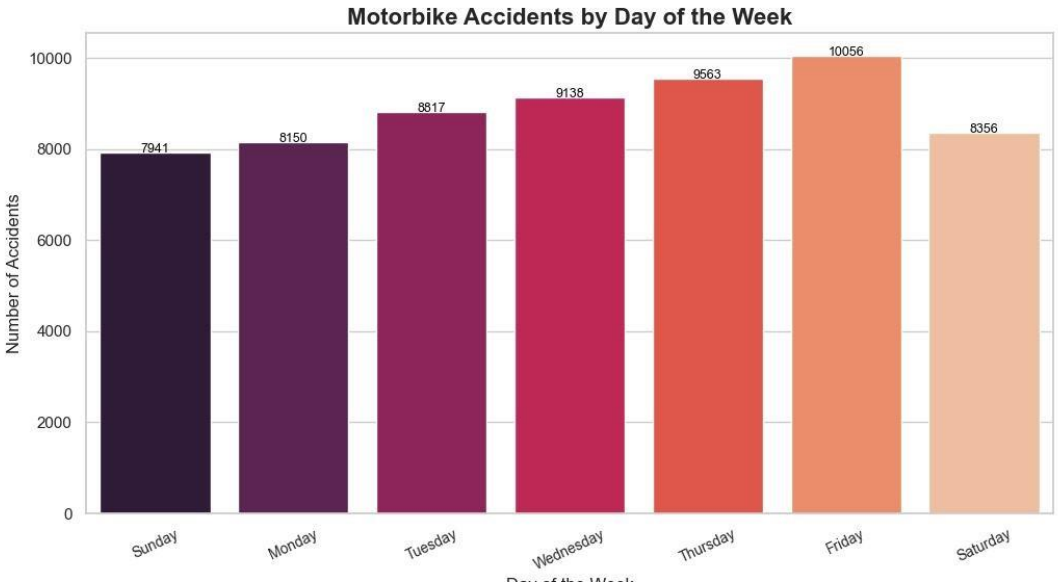

Figure 4: Motorbike Accidents by Day of the Week

The bar chart shows the distribution of accidents by day, revealing Friday as the most accidentprone day, with upwards of 10,000 accidents. This may result from increased traffic volume and fatigue as people wrap up their workweek. Saturday (8300) and Sunday (7900) have comparatively fewer accidents, potentially due to reduced commuter traffic.

3.3.Pedestrian Accident Trends

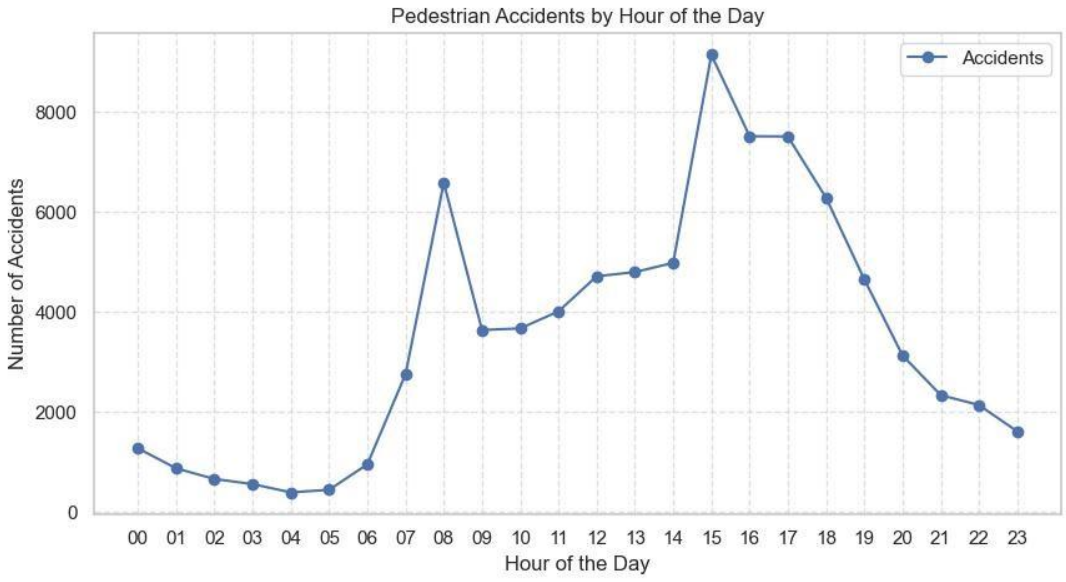

Figure 5: Pedestrian Accidents by Hour of the Day

The line graph showing motorbike accidents by hour of the day highlights distinct peaks during specific times. The most significant peak occurs during evening rush hours, around 15:00, likely due to heavy traffic as people return home from work or school. A smaller peak is observed in the morning, around 08:00, coinciding with the morning rush hour when people commute to work or school.

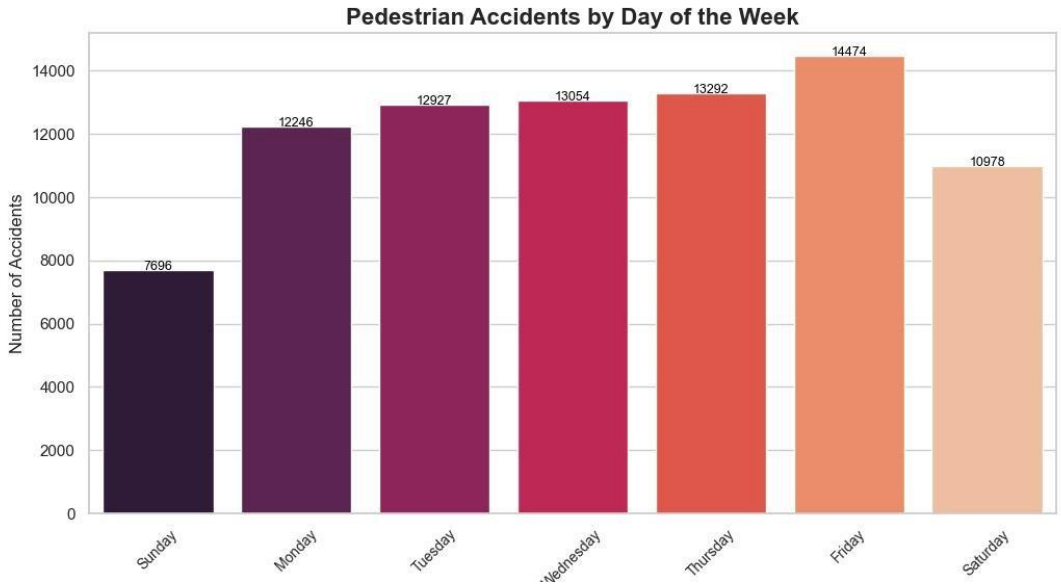

Figure 6: Pedestrian Accident by Day of the Week

The bar chart shows the distribution of accidents by day, revealing Friday as the most accidentprone day, with upwards of 14,400 accidents. This may result from increased traffic volume and fatigue as people wrap up their workweek. Saturday (10900) and Sunday (7600) have comparatively fewer accidents, potentially due to reduced commuter traffic.

3.4.Apriori on Accident Severity

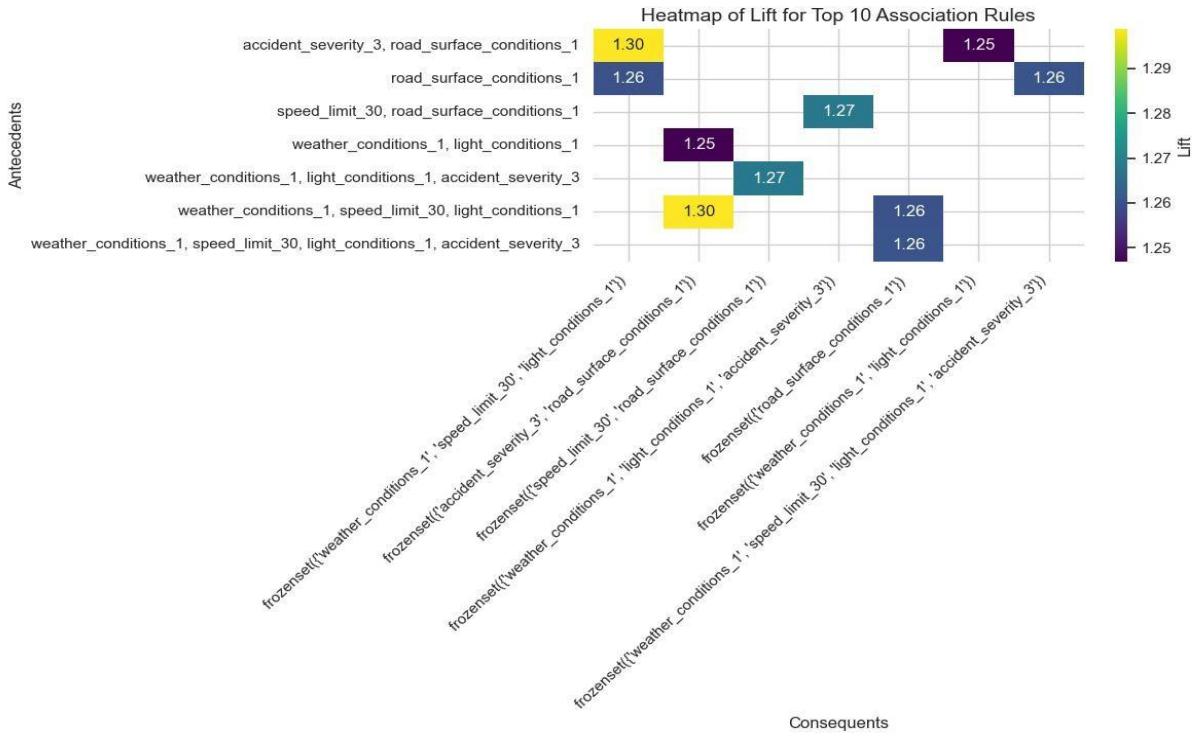

Figure 7: Heatmap of Lift for Top 10 Association Rules

The heatmap illustrates the lift values for the top 10 association rules derived from a dataset. The rows represent the antecedents (e.g., weather conditions, road surface conditions), while the columns indicate the consequents. Each cell's color corresponds to the lift value, with higher values (e.g., 1.30) highlighted in yellow, indicating stronger associations. For instance, the combination of "weather_conditions_1, speed_limit_30, and light_conditions_1" is strongly associated with "accident_severity_3" (lift = 1.30). The heatmap helps identify relationships between factors influencing accidents, emphasizing how combinations of environmental and traffic conditions correlate with accident severity or other outcomes.

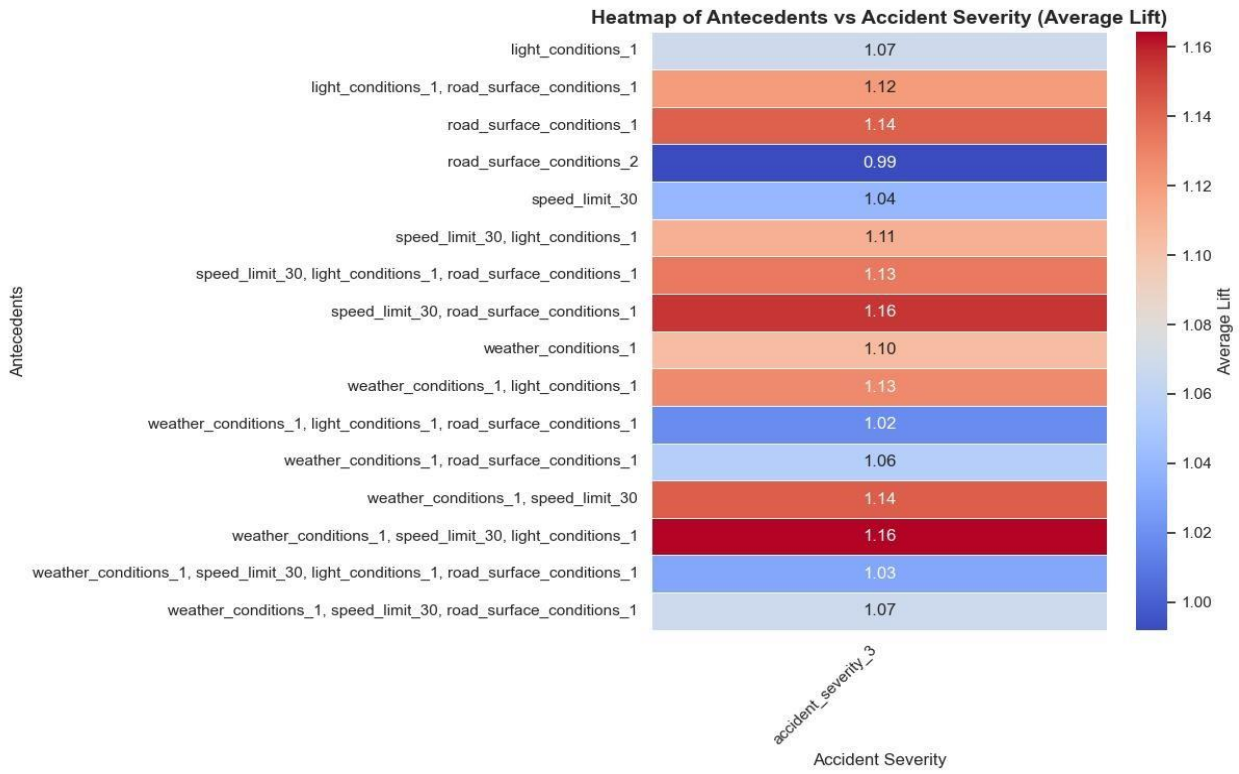

Figure 8: Heatmap of Antecedents vs Accident Severity

The heatmap illustrates the average lift values for various combinations of antecedents (e.g., weather conditions, road surface conditions, light conditions, and speed limits) in relation to accident severity (specifically, accident_severity_3). The rows represent distinct combinations of antecedents, while the colors indicate the strength of association, with darker red representing higher lift values and blue indicating weaker associations. For example, "speed_limit_30, road_surface_conditions_1" shows a strong lift of 1.16, highlighting a significant relationship with severe accidents. Conversely, "road_surface_conditions_2" has a lower lift (0.99), suggesting a weaker link. This detailed analysis identifies key environmental and situational factors contributing to severe accidents, supporting targeted interventions.

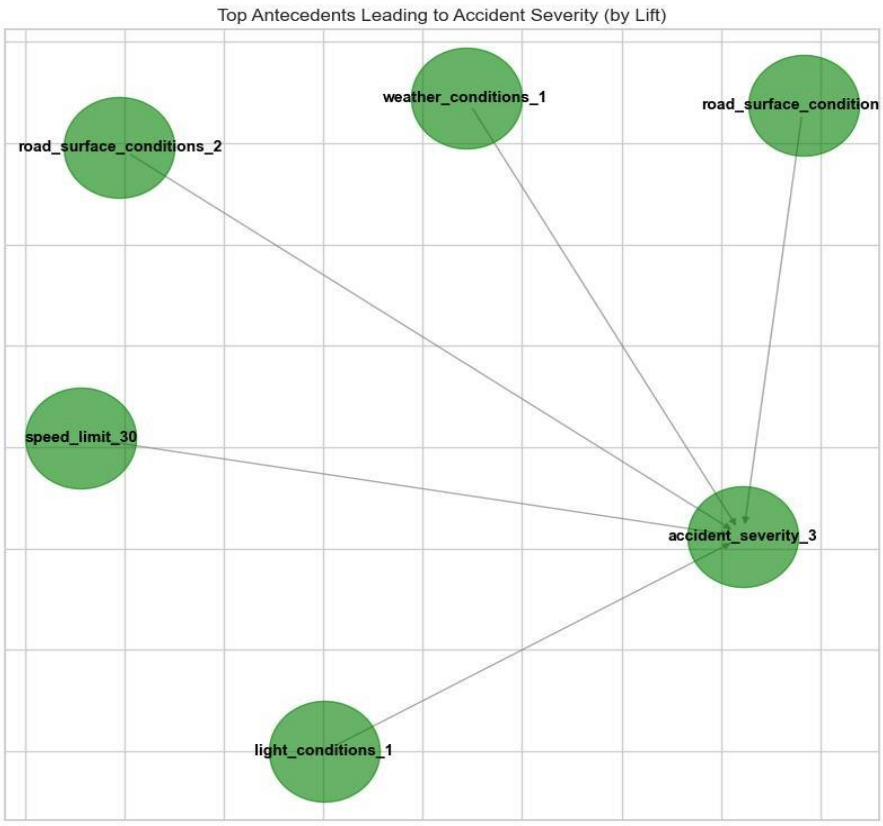

Figure 9: Top Antecedents leading to accident severity (by lift)

This graph represents a network diagram highlighting the associations between antecedent factors (e.g., road surface conditions, light conditions, weather conditions, etc.) and the consequent, accident_severity_3.

Larger nodes, such as road_surface_conditions_2, indicate higher significance in terms of contributing to accident severity.

Edges between nodes show relationships, and the thickness suggests the strength of association.

Factors like weather_conditions_1 and speed_limit_30 are connected to accident_severity_3, indicating they may play an important role.

3.5.Regional Accident Clustering

Using data specific to regions such as Kingston upon Hull, Humberside, and the East Riding of Yorkshire, clustering was applied to identify accident hotspots.

The Elbow Method and Silhouette Score helped determine the optimal number of clusters, ensuring meaningful grouping of accident data points. The resulting clusters reveal that highincident locations, such as Kingston upon Hull, are particularly significant accident-prone areas. The density of points around these cluster centers indicates concentrated accident occurrences, which could be attributed to factors like high traffic volume, complex intersections, or specific environmental conditions.

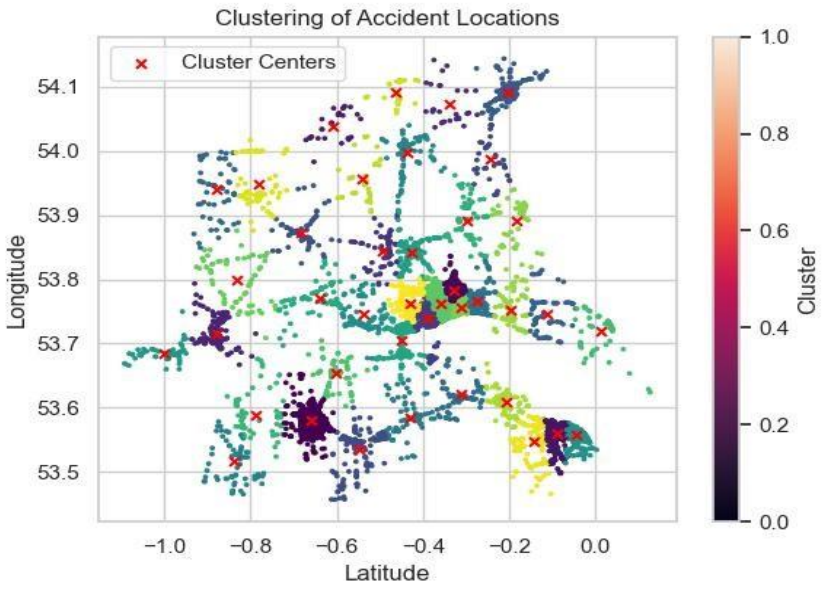

Figure 10: Clustering of Accident Locations

3.5.1. Clustering of Light & Road Surface Conditions

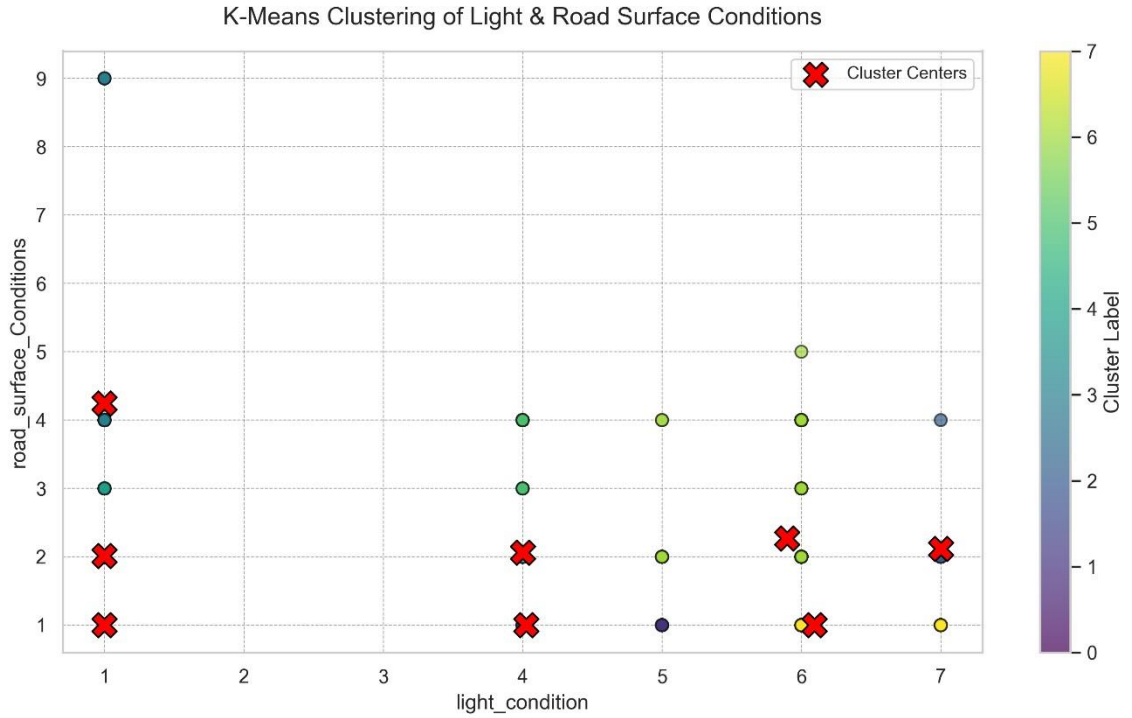

Figure 11: Clustering of Light and Road Surface Conditions

Further clustering was conducted to explore the relationship between light conditions and road surface conditions in accidents. The analysis revealed specific combinations of conditions that were associated with accidents. For instance, accidents frequently occurred when the road surface condition was classified as 1, and the light condition was either 1, 4, or 6.

3.6.Weekly Accident Forecasting

The time series data was analyzed through visualization and decomposition to extract key components—trend, seasonality, and residuals—providing insights into accident patterns. Stationarity was assessed using the Augmented Dickey-Fuller (ADF) test, with differencing applied as needed. Model parameters (p and q) were identified using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, refined via a grid search to minimize Akaike Information Criterion (AIC). Validation included the Ljung-Box test for residual randomness and Q-Q plots to confirm normality.

Three policing areas chosen were: City of London = 48, Derbyshire = 30, Humberside = 16. Data from 2017 to 2019 was used to predict accident counts for the year 2020.

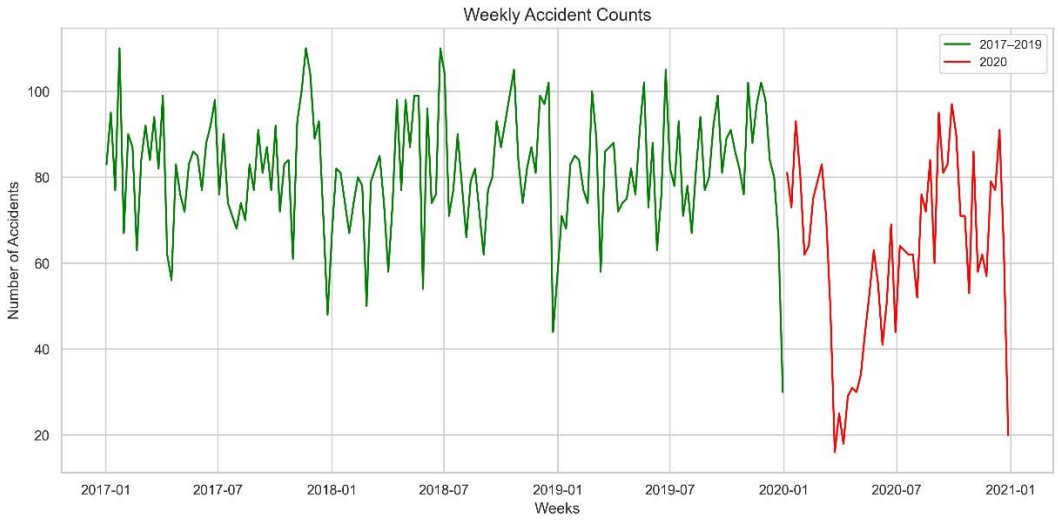

Figure 12: Weekly Accidents Counts

The original time series data was stationary as such a grid search determined that the ARMA (6,0,7) model provided the best fit for the dataset. The model's predictions are visualized in the graph below:

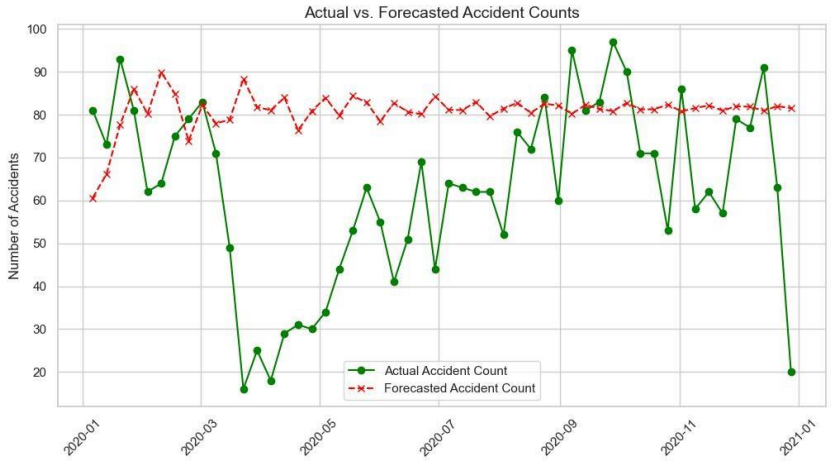

Figure 13: Actual vs Forecasted Accident Counts

The RMSE was calculated to be 28.85, which represents approximately 46.27% of the mean accident count. While the RMSE indicates a better fit compared to higher error margins, it still suggests the need for further refinement to enhance accuracy.

3.6.1. Hull High-Incident LSOAs Forecast

The analysis concentrated on the Kingston upon Hull region, specifically examining three key locations: E01012817, E01012848, and E01012889, which recorded the highest number of road accidents during the first quarter of 2020. These locations were then chosen to predict accident counts for July 2020, utilizing data from the first half of the year (January to June 2020).

An analysis of the time series data revealed that the series for E01012817 and E01012889 were non-stationary, requiring differencing to achieve stationarity. Through a grid search, the optimal models were determined for each location: an ARMA (5,6) model for E01012848, and ARIMA (15,14) and ARIMA (2,3) models for E01012817 and E01012889, respectively. Predictions made using these models, as depicted in the graph, resulted in an RMSE of 0.0 for all three locations, indicating a perfect alignment between the predicted and actual values.

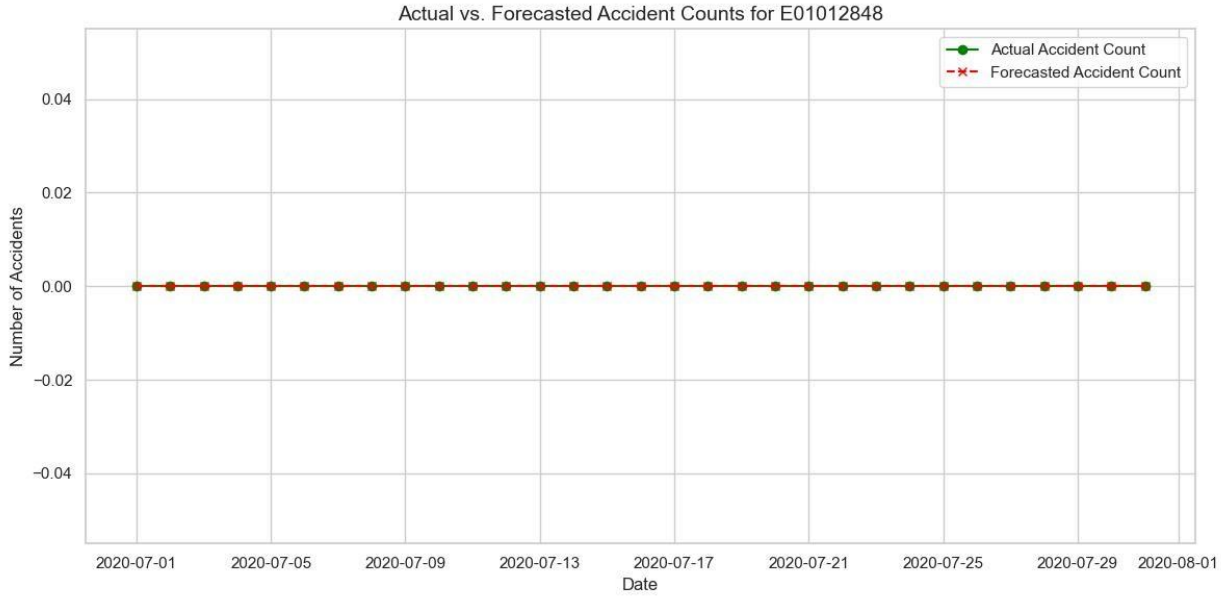

Figure 14: Counts for E01012848

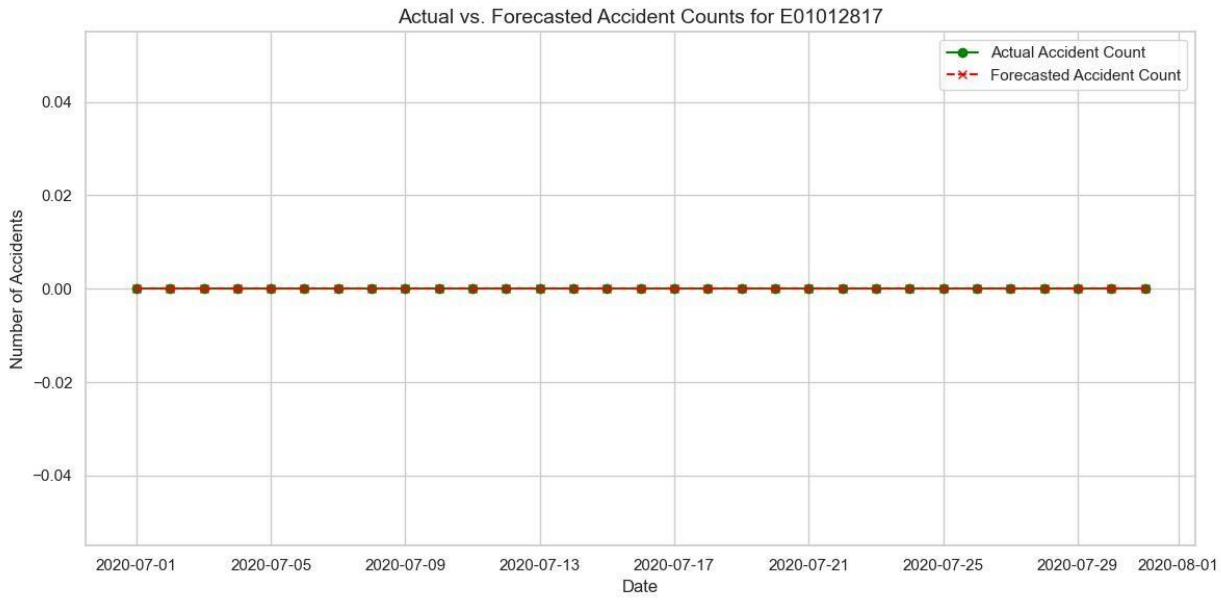

Figure 15: Count for E01012817

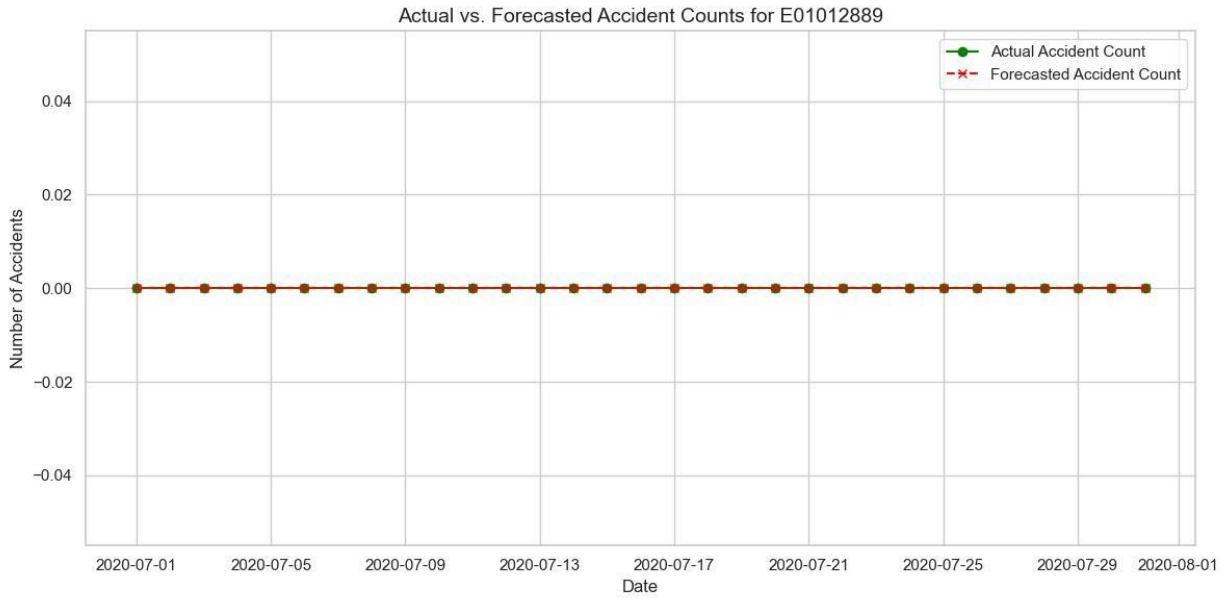

Figure 16: Count for E01012889

3.7.Social Network Characteristics

A social network was constructed from the dataset, with individuals represented as nodes and their connections depicted as edges. The spring layout algorithm was used to visualize the network, positioning the nodes to illustrate their structural relationships. The resulting visualization revealed the intricate and highly interconnected structure of the social network, as shown below:

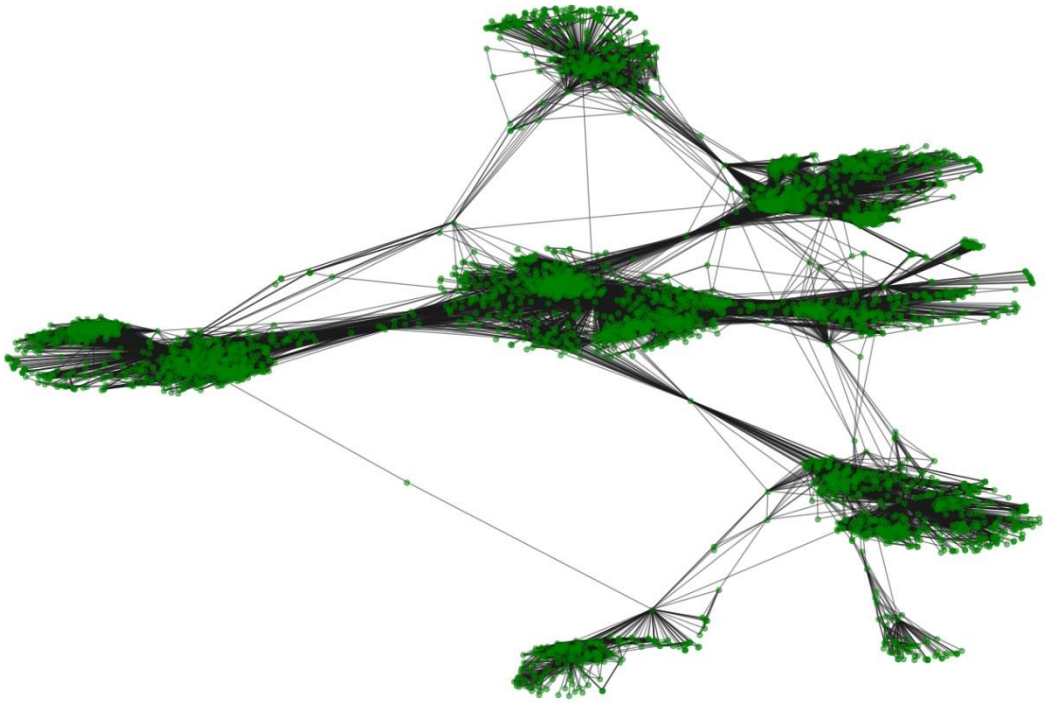

Figure 17: Social Network

The social network consists of 4,039 nodes and 88,234 edges, representing the individuals and their connections, respectively. The network density is 0.0108, indicating that only about 1% of all possible connections between nodes are present. This is typical for real-world social networks, which tend to be sparse. Additionally, the average degree of the network is 43.69, meaning that, on average, each individual (node) is connected to approximately 44 others. This highlights a relatively high level of interconnectivity within the network.

3.8.Edge Centrality Distribution

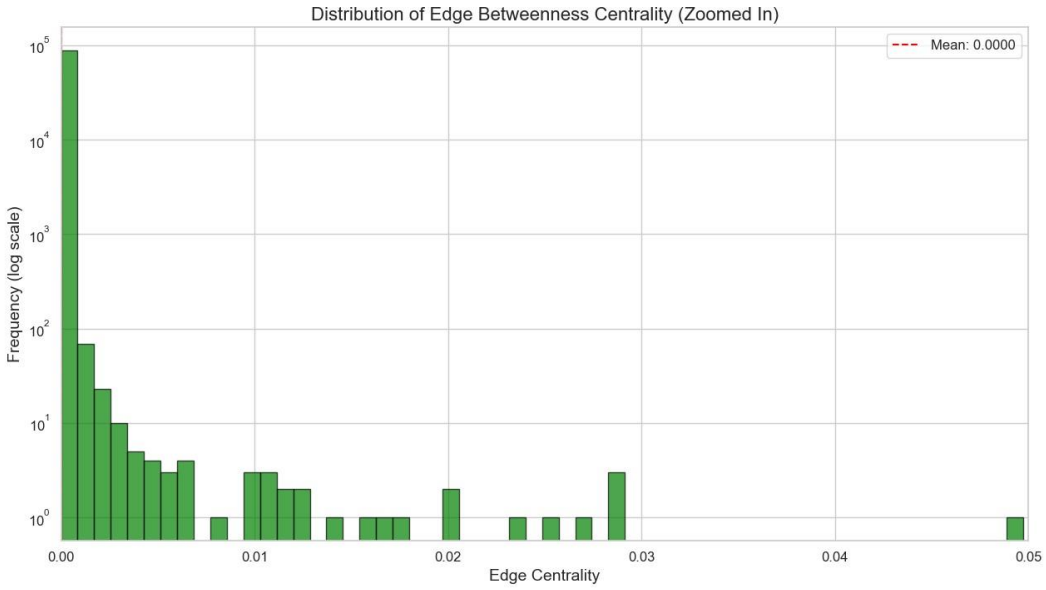

Figure 18: Distribution of Edge Between Centrality

The distribution is heavily skewed towards lower values, indicating that the majority of edges have very low centrality, meaning they are rarely part of the shortest paths connecting different node pairs. However, a few edges exhibit relatively higher centrality, suggesting they play a critical role in facilitating connectivity and information flow within the network. The mean value, marked as nearly zero, further emphasizes the dominance of edges with minimal centrality.

3.9.Community Detection Comparison

Two community detection algorithms were applied to identify clusters within the network: Louvain Modularity and Leiden Propagation. These methods aim to group nodes into communities based on the density of connections within groups compared to connections between groups.

Table 1: Comparison of community detection algorithm

	LOUVAIN MODULARITY	LEIDEN PROPAGATION
NUMBER OF COMMUNITIES	13	44

COMMUNITY SIZES	[983, 815, 548, 543, 372, 219, 208, 206, 59, 37, 25, 18, 6]	[198, 36, 10, 8, 8, 34, 2, 215, 16, 3, 3, 1030, 6, 7, 3, 3, 753, 10, 2, 2, 469, 13, 9, 3, 49, 25, 2, 60, 547, 179, 10, 9, 8, 226, 19, 4, 3, 8, 6, 14, 12, 7, 6, 2]
AVERAGE COMMUNITY SIZE	310.69	91.79
MODULARITY SCORE	0.7774	0.7368

Louvain Modularity identified 13 larger, balanced communities with a higher modularity score (0.7774), indicating well-defined structures, while Leiden Propagation detected 44 smaller, varied communities with a slightly lower modularity score (0.7368). Leiden's finer granularity reflects a more dispersed node distribution across communities. The diagrams are shown below:

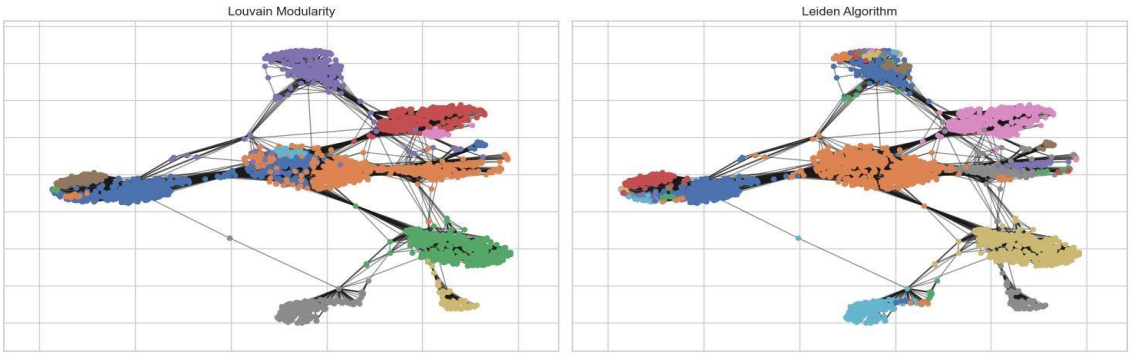

Figure 19: Community Detection Models

4. Recommendations

Implement Targeted Traffic Measures During Peak Hours:

- Increase patrols and enforce stricter speed limits during evening rush hours (4:00 PM to 6:00 PM) and morning rush hours (8:00 AM). These are the times when accidents, especially for motorcyclists and pedestrians, are most frequent.

Address High-Incident Days:

- Focus on Fridays, which account for the highest proportion of accidents. Measures could include public safety campaigns about fatigue and road awareness, as well as increased monitoring and enforcement.

Improve Safety in High-Risk Locations:

- Use clustering analysis to target accident hotspots like Kingston upon Hull and other high-density accident areas. Implement enhanced signage, better lighting, and infrastructure changes, such as roundabouts or pedestrian overpasses, in these areas.

Enforce Speed Limits:

- Strengthen compliance with speed limits, especially in areas where 30 mph limits are common, as these are often linked with severe accidents. Automated speed cameras could help ensure adherence.

Adapt to Weather Conditions:

- Develop predictive models to identify and mitigate risks associated with specific weather patterns. Use dynamic digital signage to alert drivers about adverse conditions and suggest alternative routes.

Pedestrian-Specific Safety Measures:

- Create safe crossing zones near schools, work areas, and public transport hubs. Pedestrian-focused campaigns and infrastructure improvements, like signalized crossings, can reduce accident frequency.