University of HULL

MSc. in Artificial Intelligence & Data Science

771762_C24_T1: Big Data and Data Mining

Accident & Facebook Data Analysis

Name:

Muhammad Huzaifa Owais

Student ID:

202412741

Word Count: 2038

Course Instructor:

Dr Julius Sechang Mboli

INTRODUCTION

Road traffic accidents pose significant public safety risks, often leading to injuries, fatalities, and economic losses. Using the **accident_data_v1.0.0_2023.db** dataset, containing four interconnected encoded tables; **accident, casualty, vehicle, and lsoa**, this analysis aims to uncover actionable insights to improve road safety.

Table 1: Database Description

Table	Rows	Columns
Accident	461352	36
Casualty	600332	19
Vehicle	849091	28
Isoa	34378	7

The tables are interconnected through specific relationships:

- **Accident-to-Casualty**: One-to-Many relationship, as each accident may involve multiple casualties.
- **Accident-to-Vehicle**: One-to-Many relationship, as each accident may involve multiple vehicles.
- **Accident-to-LSOA**: Many-to-One relationship, as multiple accidents can occur within the same local area.

As data scientists, our objective is to leverage this dataset to uncover actionable insights that can aid government agencies in improving road safety.

DATA CLEANING

The dataset contained several issues such as blank entries, -1 values, and redundancies. Cleaning efforts were focused on columns used in the analysis, employing logical imputation methods where possible. Below is a summary of the cleaning procedures:

Table 2: Data Cleaning Details

Column(s)	Problem	Procedure
location_easting_osgr, location_northing_osgr, longitude, latitude	Blank entries	Replaced blanks with the mode values of respective columns within the same local_authority_district, assigning the most frequent accident location in the district.

local_authority_district	-1 Values	Replaced -1 with 37 for Suffolk (East & West). For remaining, imputed valid values from similar rows based on matching other columns.
local_authority_highway	-1 Values	Replaced -1 with corresponding values from the local_authority_ons_district column.
speed_limit	-1 Values	Replaced -1 with the mode of speed_limit within the respective local_authority_district.
light_conditions	Few -1 entries	Imputed 1 (daylight) for early hours and 7 (darkness: street lighting unknown) for late hours based on the time of day.
weather_conditions	-1 Values	Replaced -1 with 9 (unknown), as weather conditions are difficult to predict.
road_surface_conditions	-1 Values	Replaced -1 with 9 (unknown), as road surface conditions are also difficult to predict.
lsoa_of_accident_location	-1 Values	Replaced -1 with "unknown," as this column is a string (varchar) and cannot be reliably predicted.

ANALYSIS

A. TEMPORAL ANALYSIS

Temporal analysis was conducted to identify significant patterns in road traffic accidents based on the time of day and the day of the week. This section highlights the findings for all accidents, motorcycle-specific incidents, and pedestrian-related accidents. Data was carefully filtered and grouped by hour and day to provide meaningful visualizations and insights.

1. Accidents by Time and Day

Day of the Week

The highest number of accidents occurred on **Friday**, likely due to increased traffic as people commute home at the end of the workweek. The buildup of traffic congestion from office closures and other activities contributes to this trend. The lowest number of accidents was observed on **Sunday**, typically a quieter day with reduced traffic.

Hour of the Day

During the first 12 hours, accidents peaked at **8:00 AM**, coinciding with the morning rush hour as offices and schools open. In the latter 12 hours, accidents peaked at **5:00 PM**, corresponding to the evening rush hour when workplaces close for the day.

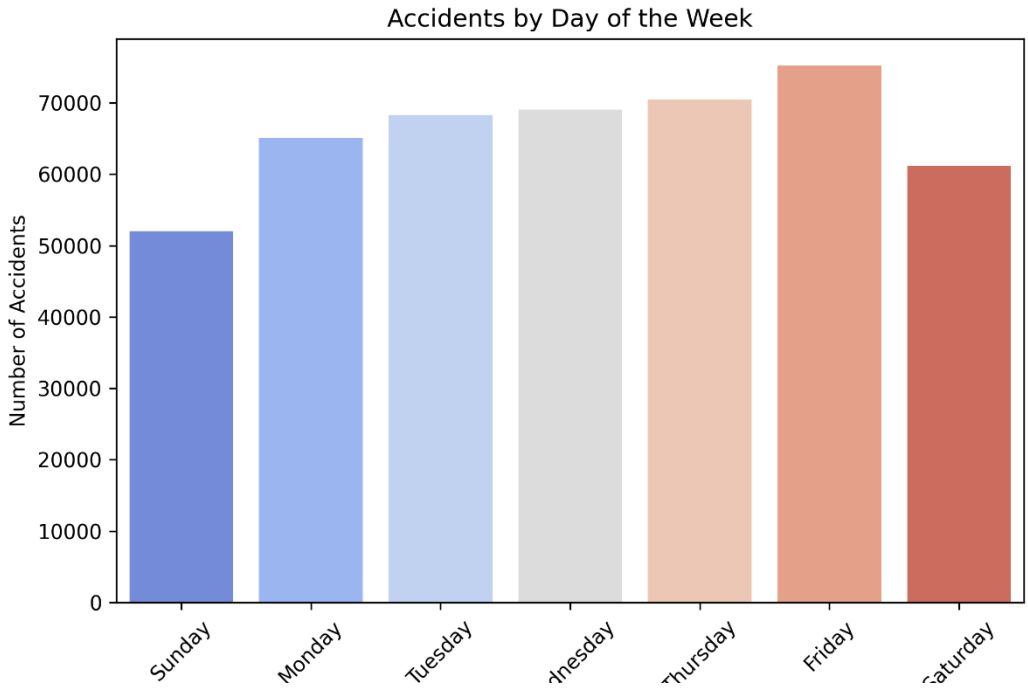

Figure 1: Accident count by Days

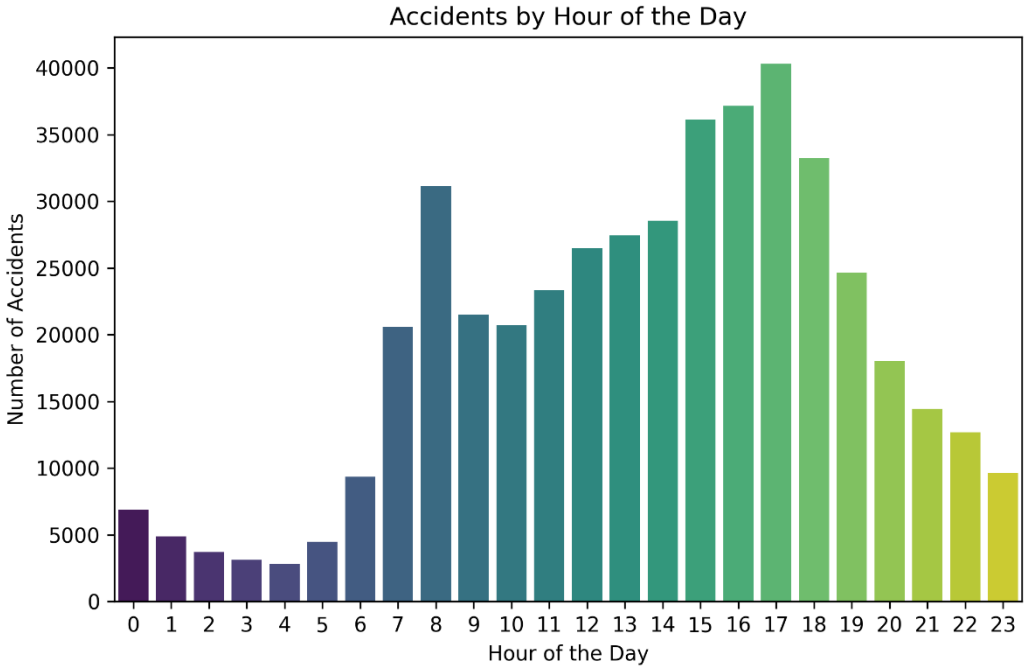

Figure 2: Accident count by Hour each Day

2. Motorcycle Accidents

Focusing on motorcycles of different engine capacities (125cc and under, 125cc–500cc, and over 500cc), similar patterns were observed:

Day of the Week

Friday recorded the highest number of motorcycle accidents, aligning with the trends seen for overall traffic accidents. **Sunday** recorded the lowest number, likely reflecting reduced road usage.

Hour of the Day

Accidents involving motorcycles showed a prominent peak at **5:00 PM**, coinciding with heavy evening traffic.

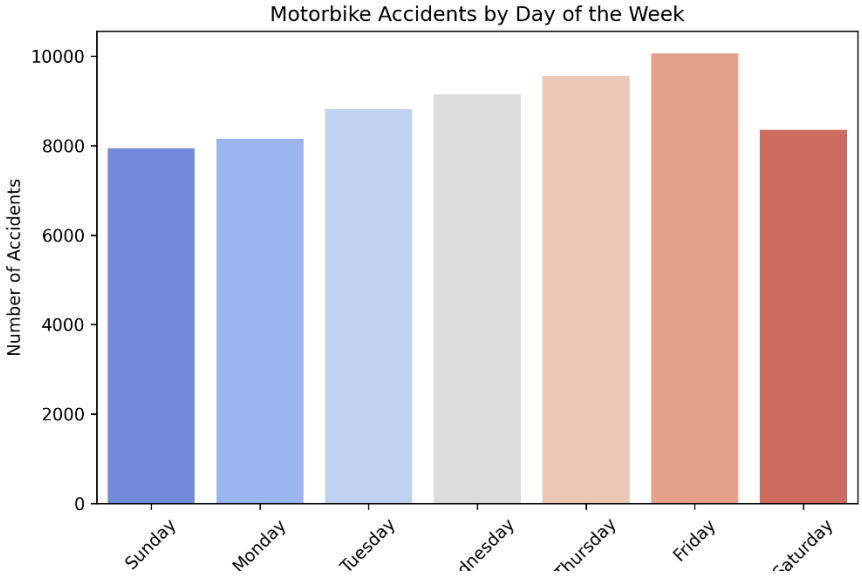

Figure 3: Motorbike Accident count by Days

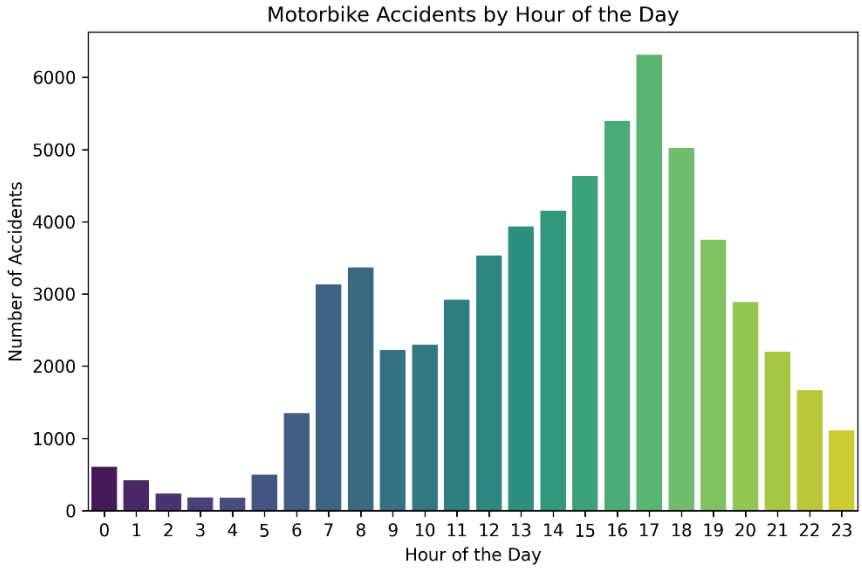

Figure 4: Motorbike Accident count by Hour each Day

3. Pedestrian-Involved Accidents

Day of the Week

Most pedestrian accidents occurred on **Friday**, mirroring the trend for overall accidents. The least pedestrian accidents occurred on **Sunday**, when pedestrian and vehicular traffic is reduced.

Hour of the Day

A unique trend was observed, with pedestrian accidents peaking at **3:00 PM**, possibly reflecting increased pedestrian activity during school hours and mid-afternoon errands.

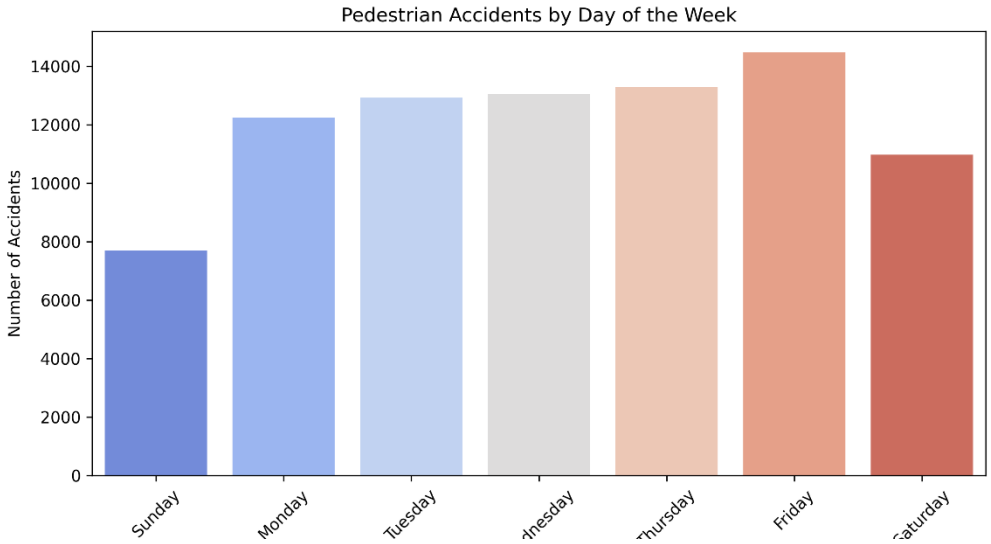

Figure 5: Pedestrian Accident count by Day

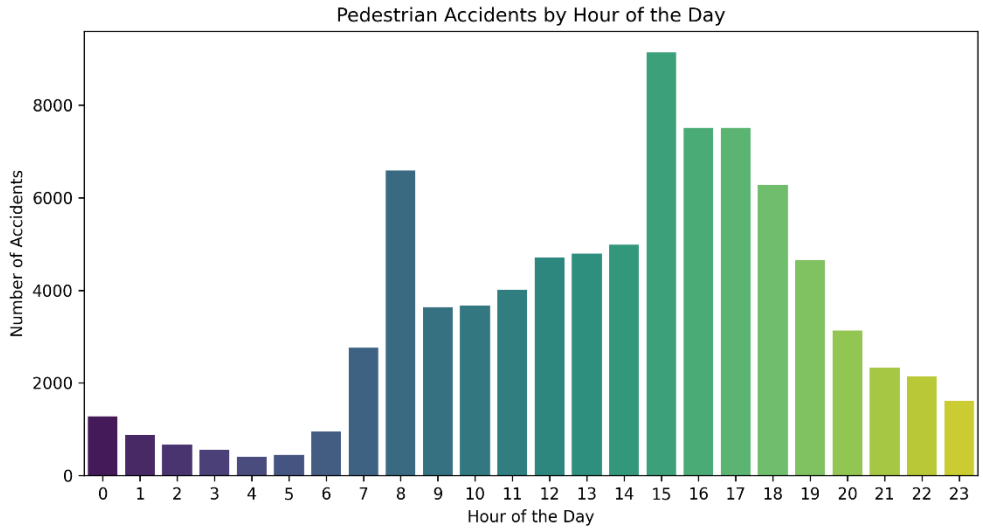

Figure 6: Pedestrian Accident count by Hour each Day

B. ACCIDENT SEVERITY AND INFLUENCING VARIABLES

To analyze the impact of various factors on accident severity, key variables including speed_limit, weather_conditions, road_surface_conditions & light_conditions were examined using the Apriori algorithm. This method identified frequent patterns and associations, with lift metrics used to quantify the strength of these associations.

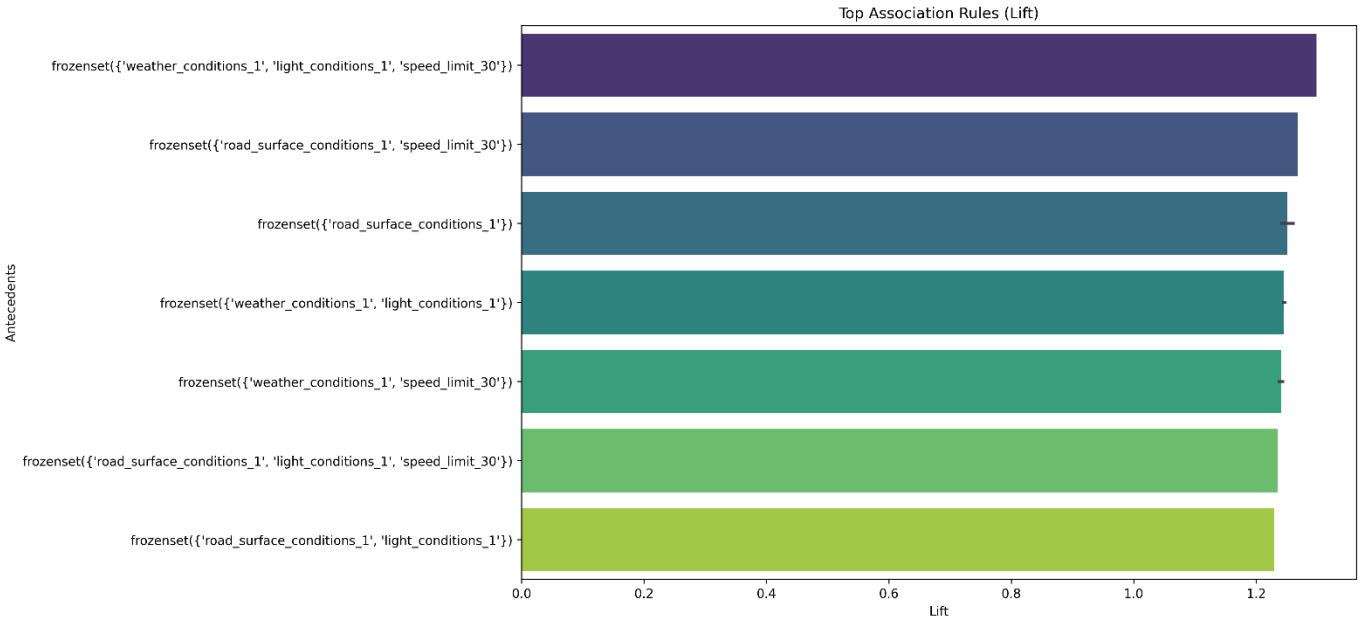

Figure 7: Top Association Rules by LIFT

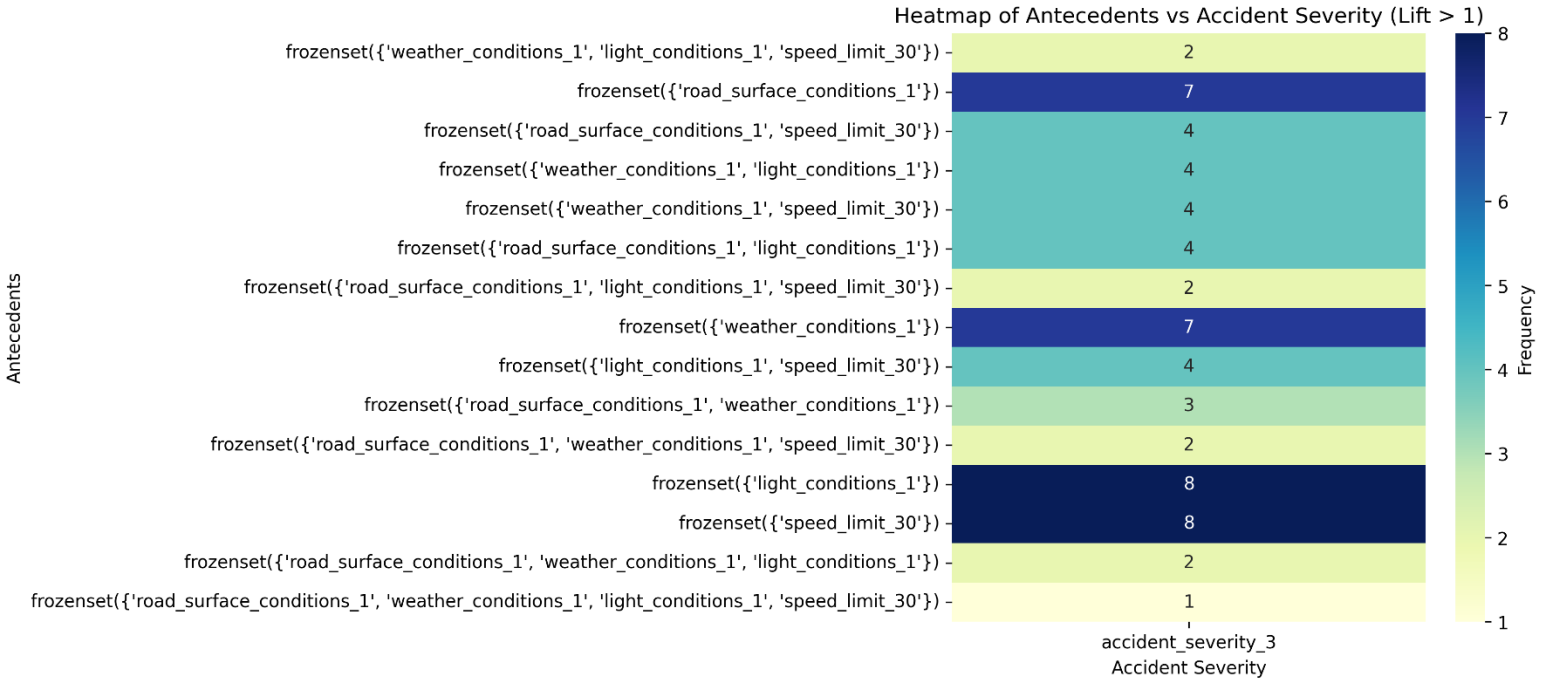

Figure 8: Heatmap of Antecedents that lead to Accident by LIFT

Table 3: Variables & their encoded details

Variables	Encoding	Encoding interpretation
speed_limit	30	30 mph
weather_conditions	1	Fine no high winds
road_surface_conditions	1	Dry
light_conditions	1	Daylight
accident_severity	3	Slight

The Apriori algorithm identified above conditions as being strongly associated with accidents of **slight** severity. This suggests that accidents under these conditions are relatively frequent, though not necessarily severe. It is likely that in these circumstances, accidents are caused by human factors such as inattention, distraction, or minor driver error, as the environmental conditions are generally favorable. While the Apriori algorithm identifies associations based on data frequency, from a practical standpoint, these are typical conditions on a normal driving day, and such accidents may occur randomly rather than being the result of specific dangerous circumstances.

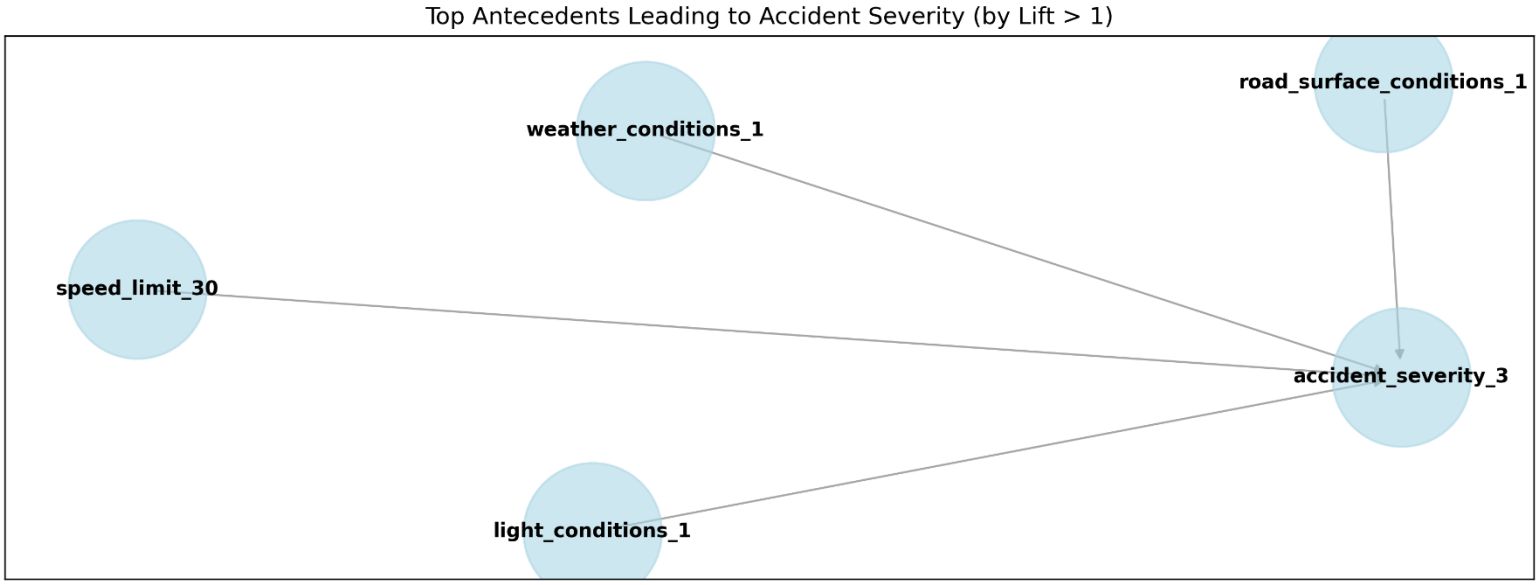

Figure 9: Network of Top 4 Antecedents that lead to Accident

C. REGIONAL ANALYSIS

The regional analysis focused on accident data from the **Humberside region**, encompassing **Kingston upon Hull**, **East Riding of Yorkshire**, **North Lincolnshire**, and **North East Lincolnshire**. Among these areas, **Kingston upon Hull** recorded the highest number of accidents, underscoring it as a significant focal point for road safety interventions.

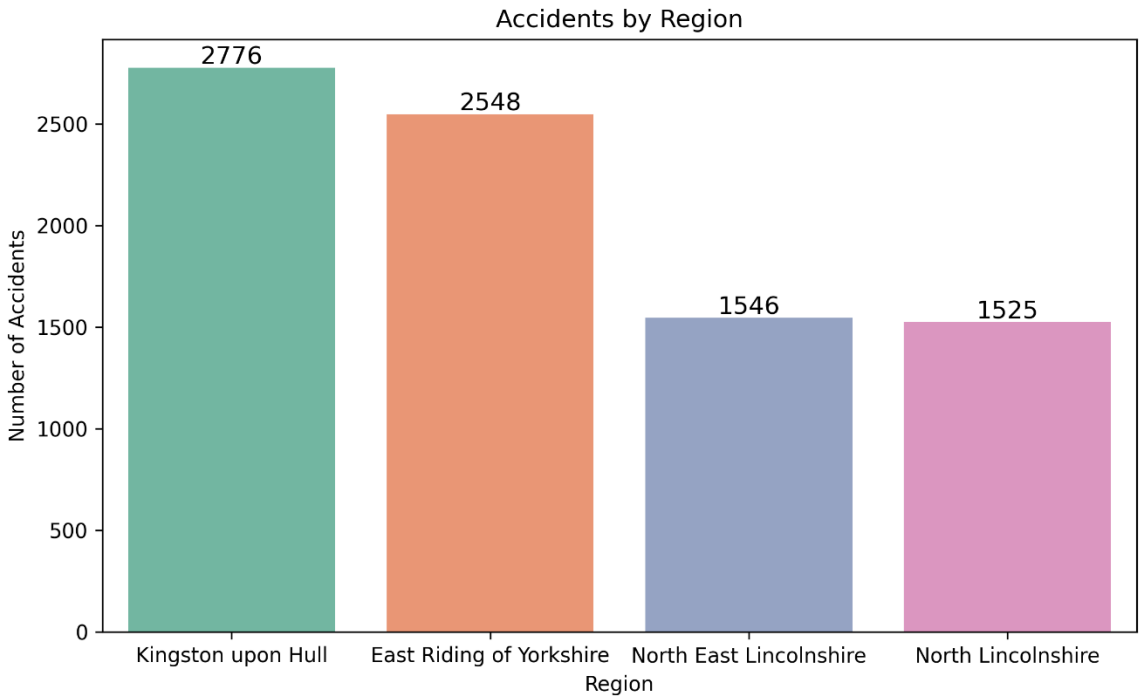

Figure 10: Accident count of each region

1. Geographical Clustering

To identify accident hotspots across the entire region, **geographical clustering** was applied to the dataset. The optimal number of clusters was determined using the **Elbow Method** and further validated by the **Silhouette Score**. This analysis revealed several high-incident locations, with **Kingston upon Hull** serving as an example of such clusters. Specific streets in Kingston upon Hull identified as hotspots include:

- **Faversham Avenue connecting to Anlaby Road**: Coordinates [-0.40953776, 53.74625224].
- **De Grey Street connecting to Beverley Road**: Coordinates [-0.35352298, 53.76227102].
- **Portobello Street connecting to Holderness Road**: Coordinates [-0.28293334, 53.76309072].

These examples illustrate how clustering identifies areas with high accident concentrations, aiding in targeted interventions.

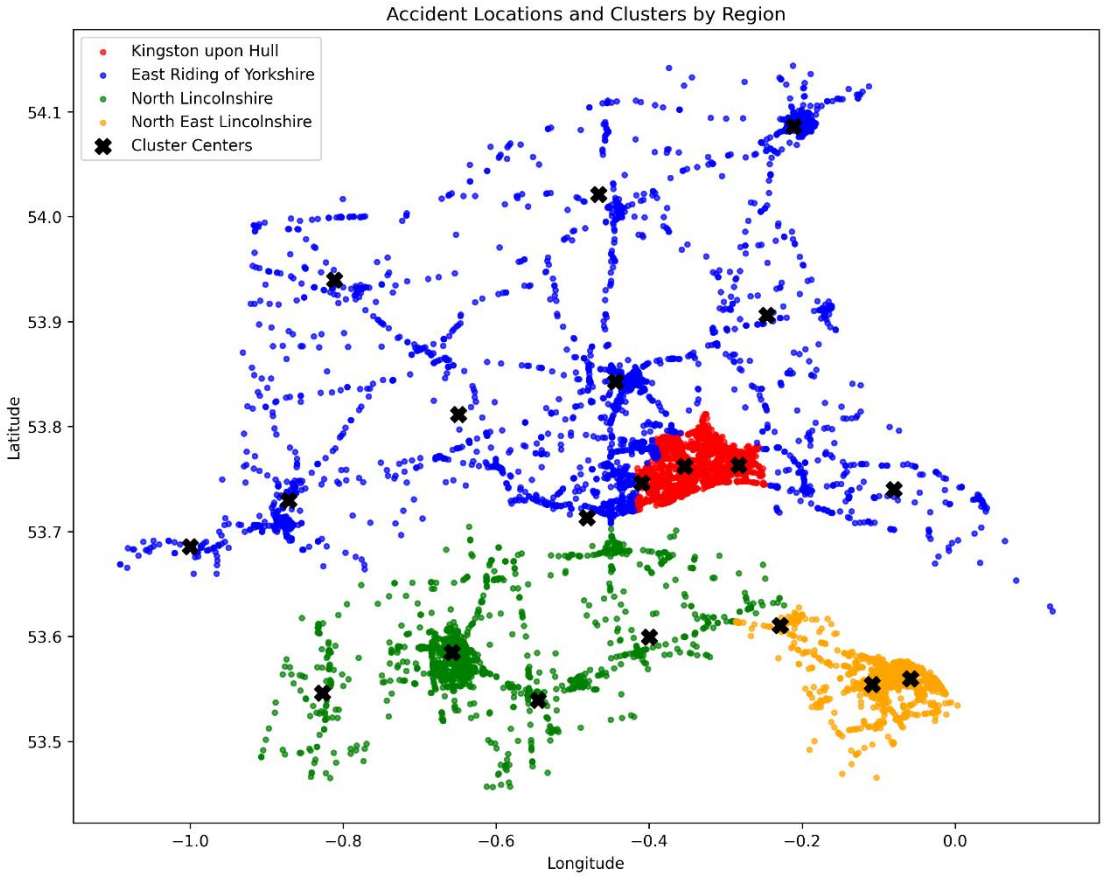

Figure 11: Clustering on Accident Locations

2. Weather and Road Surface Clustering

Further clustering was conducted to explore the relationship between **weather 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 (Dry)**, and the **weather condition** was either **1 (Fine with no high winds)**, **4 (Fine with high winds)**, or **9 (Condition unknown)**.

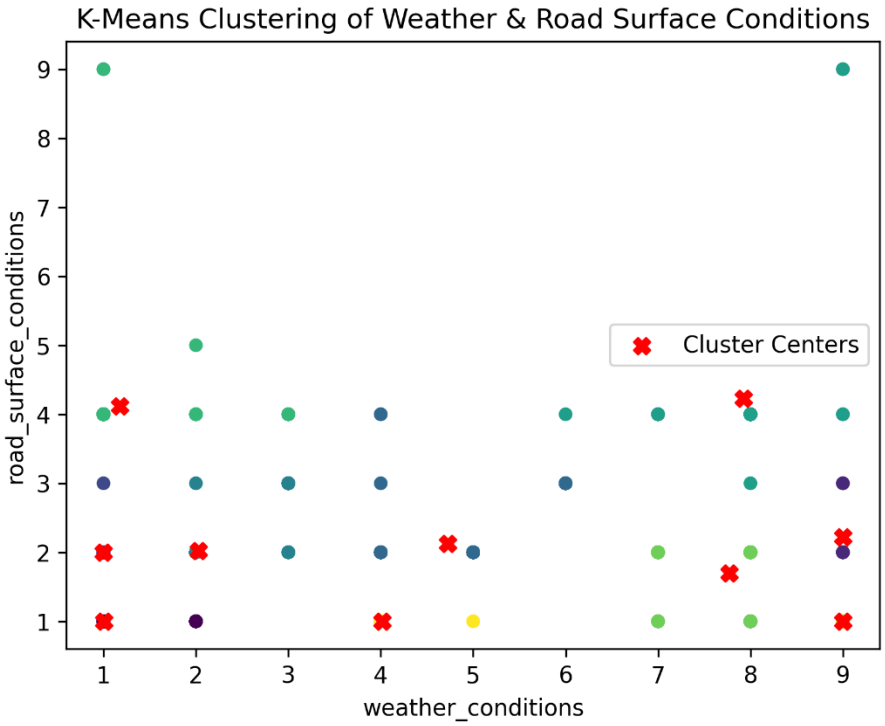

Figure 12: Clustering on Weather & Road Surface Conditions

Among these, the combination of **dry roads (1)** and **fine weather with no high winds (1)** showed the highest likelihood of accidents, as indicated by the greater number of cases in that cluster **(1,1)**, as visualized in the bar graph below.

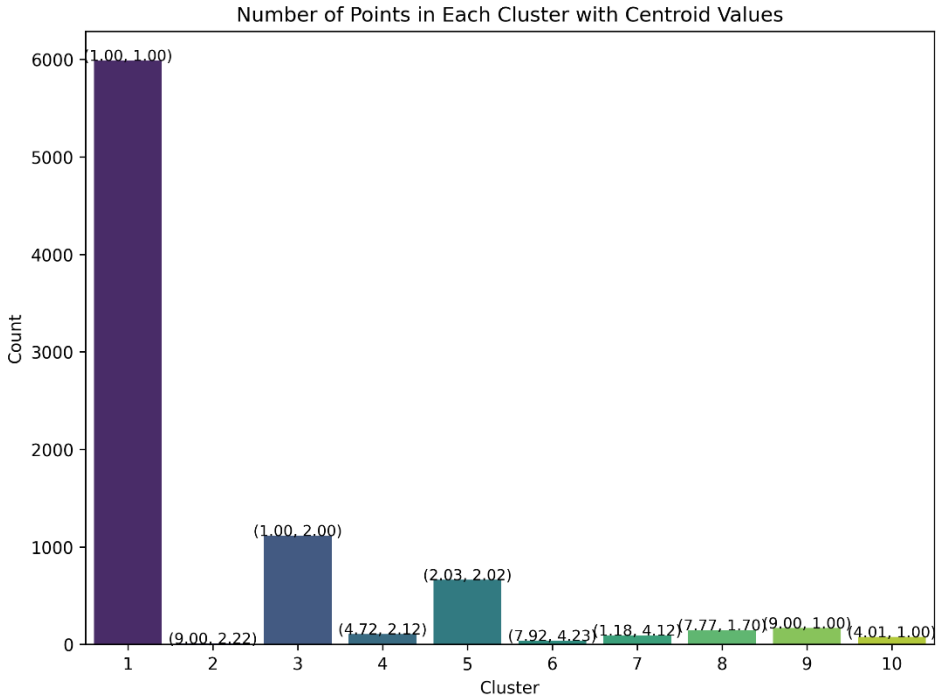

Figure 13: Count of Points in each cluster

FUTURE ACCIDENT PREDICTIONS

This study aims to forecast future accident counts at both weekly and daily intervals using robust time series modeling techniques. To achieve this, the dataset was carefully filtered based on specific regions and analytical requirements. A comprehensive process was followed to prepare, analyze, and model the time series data effectively:

- **Visualization and Decomposition:**

The time series data was visualized and decomposed into its key components—trend, seasonality, and residuals. This decomposition provided insights into underlying patterns, enabling a better understanding of the factors driving accident occurrences over time.

- **Stationarity Testing:**

To ensure that the data was suitable for time series modeling, the Augmented Dickey-Fuller (ADF) test was conducted. This test evaluates whether a time series is stationary by comparing the p-value to a significance threshold. If the null hypothesis (non-stationarity) was not rejected, the series was made stationary using differencing techniques.

- **Determination of Model Parameters (p and q):**

The orders of autoregressive (p) and moving average (q) terms were determined by analyzing the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.

  - ✓ To refine parameter selection, a grid search approach was employed to identify the combination of p and q that minimized the Akaike Information Criterion (AIC), ensuring the model's optimal balance between complexity and performance.

- **Model Validation:**
  The selected parameters were further validated using:

  - ✓ Ljung-Box Test: To confirm the adequacy of the model by checking for randomness in residuals.
  - ✓ Q-Q Plot Analysis: To ensure residuals followed a normal distribution.

This rigorous process ensured the development of accurate and reliable predictive models. The details and results for both weekly and daily accident count predictions are discussed below.

A. WEEKLY ACCIDENT PREDICTIONS

The weekly accident counts for three policing areas—**City of London (48), Greater Manchester (6), and Humberside (16)**—were analyzed. Data from **2017 to 2019** was used to predict accident counts for the year **2020**.

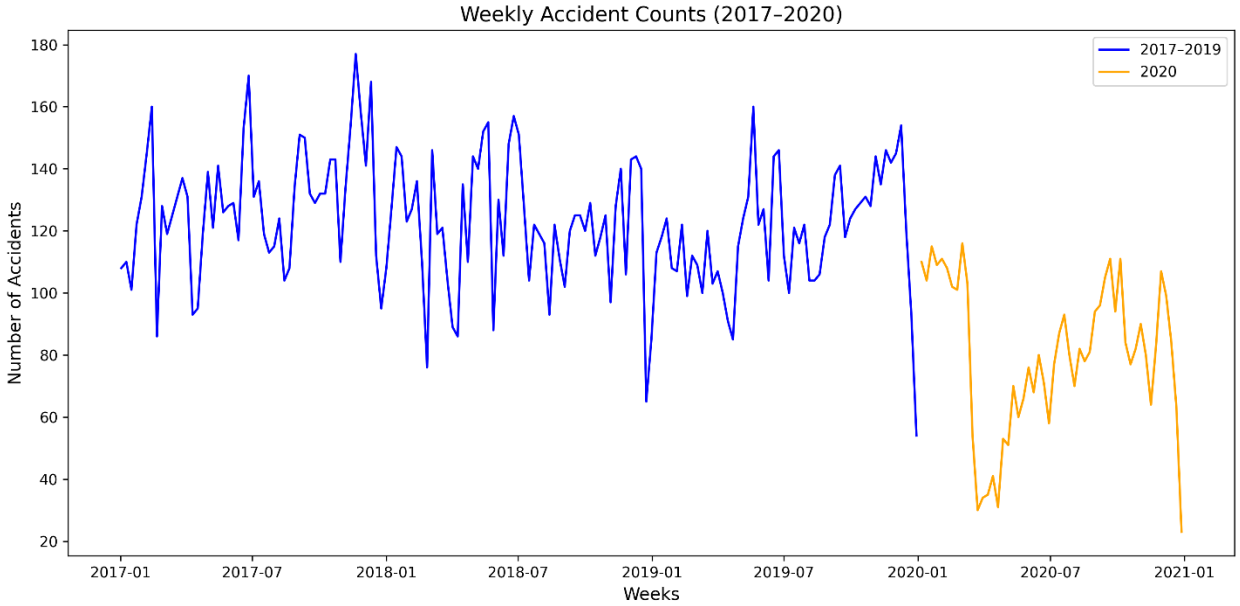

Figure 14: Weekly Accidents (2017-2020)

Since the original time series data was non-stationary, first differencing was applied to achieve stationarity. A grid search determined that the **ARIMA (1,1,1)** model provided the best fit for the dataset. The model's predictions are visualized in the graph below

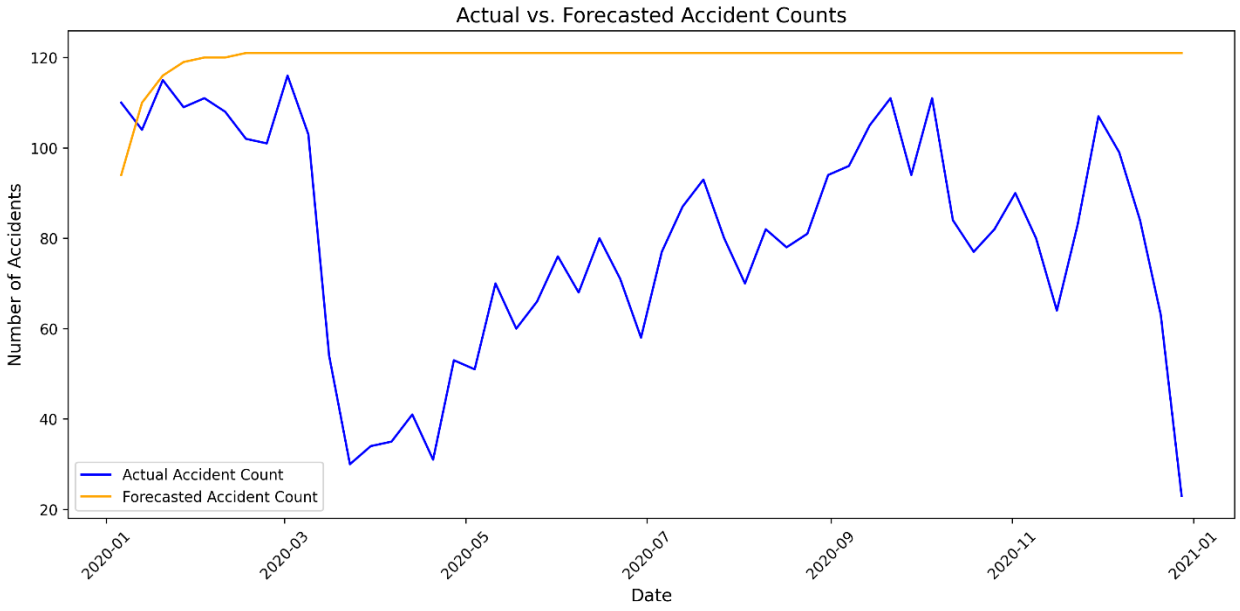

Figure 15: Actual vs. Forecasted Accident Counts

Further accuracy was assessed using the **Root Mean Square Error (RMSE)** and the **RMSE as a percentage of the mean**:

- **RMSE**: 47.77
- **RMSE as a percentage of the mean**: 59.83%

Given that actual accident counts range between **23 and 111**, an RMSE of 47.77 represents approximately **59.83%** of the typical range, which is considerable and indicates a **suboptimal predictive performance**.

REMARKS

- ✓ The **ARIMA (1,1,1)** model successfully captured the general trend and most of the variance in the accident data from 2017 to 2019. However, the model's predictive performance was significantly impacted by unforeseen external factors, notably the onset of the **COVID-19** pandemic towards the end of 2019.
- ✓ The forecasted accident count stabilizes at **121** after a few weeks, failing to reflect significant fluctuations in the actual data, particularly the sharp decline observed during **March–April 2020**.
- ✓ The dramatic shift in accident patterns due to COVID-19, which was not present in the training data (2017-2019), contributed to a mismatch between actual and predicted values.
- ✓ The high RMSE, relative to the range of accident counts, highlights the model's inability to adapt to dynamic patterns in the data.

B. DAILY ACCIDENT PREDICTIONS

The data analysis focused on Kingston upon Hull region, with particular emphasis on three locations: **E01012817, E01012848, and E01012889**, recording the highest number of road accidents during the first three months of 2020. These locations were subsequently selected to forecast the accident count for July 2020, using data from the first half of the year (January to June 2020).

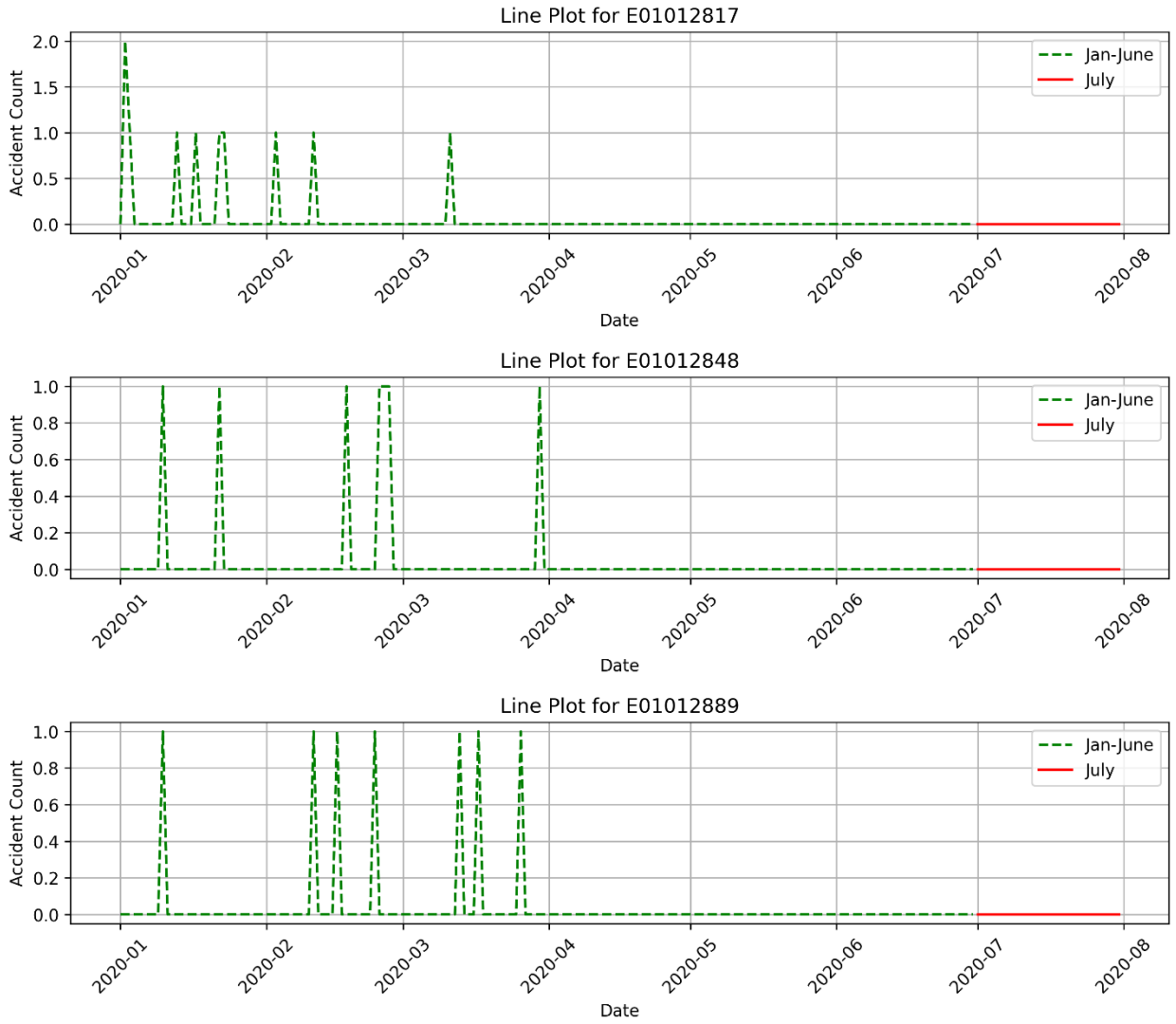

Figure 16: Regions Accident Count (Jan 2020-July 2020)

Upon reviewing the time series data, it was found that the series for **E01012817** and **E01012889** were **non-stationary**, necessitating the application of differencing to achieve stationarity. Following a grid search, the best-fit models for the different locations were identified: an **ARMA (5,6)** model for **E01012848**, and **ARIMA (15,14)** and **ARIMA (2,3)** models for **E01012817** and **E01012889**, respectively. The prediction results from these models, as shown in the graph,

yielded an **RMSE** value of **0.0** for all three locations, suggesting a perfect match between predicted and actual values.

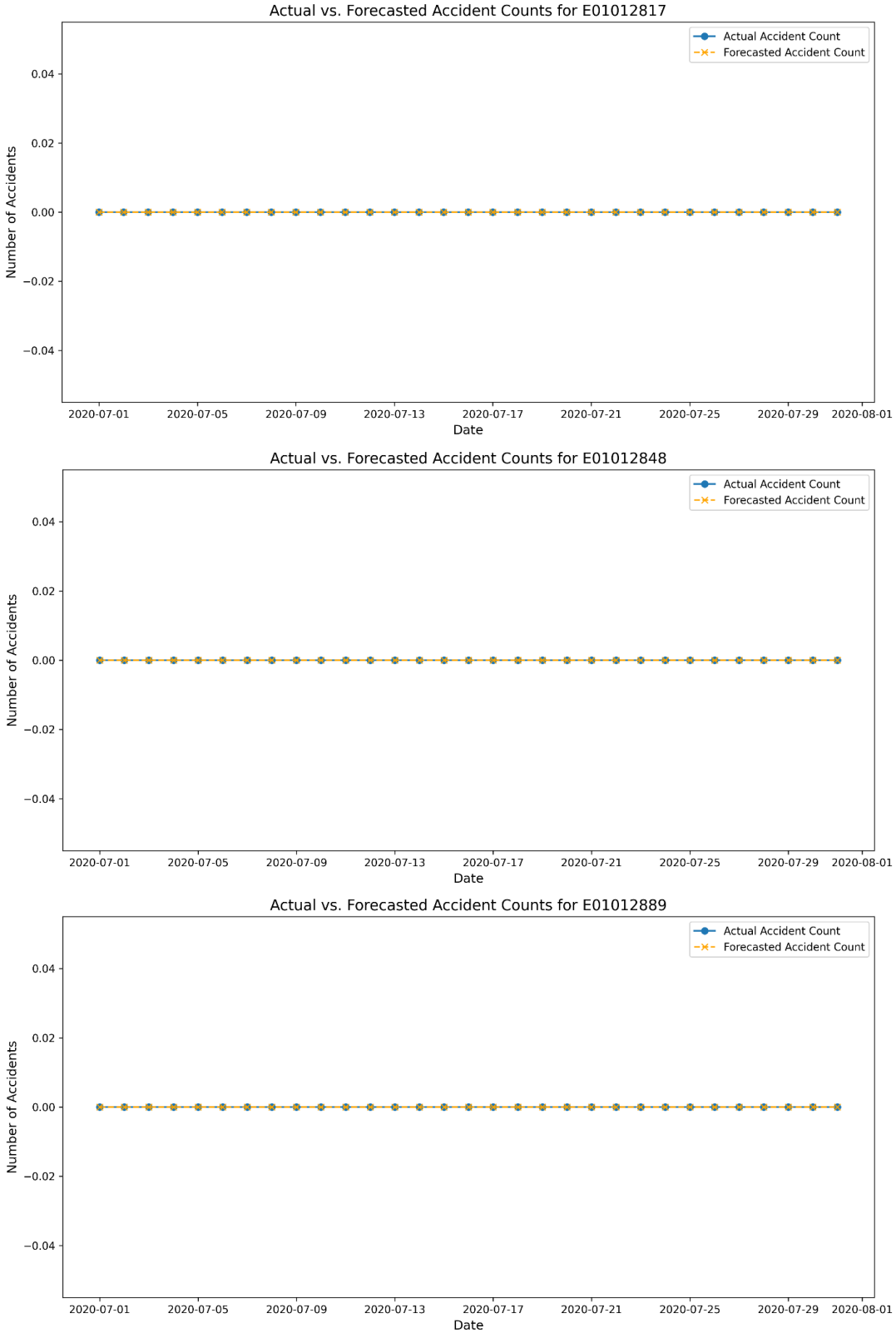

Figure 17: Regions Accident Prediction for month of July 2020

REMARKS

- ✓ However, this seemingly flawless prediction does not truly reflect the robustness of the models, primarily due to the nature of the data.
- ✓ The dataset, consisting of only 181 entries, was highly skewed, with a significant portion of the locations recording no accident counts, and only a small number of entries reporting the accident counts.
- ✓ This imbalance in the data might resulted in an overfitted model that failed to capture the variability and real-world complexity of accident occurrences.

RECOMMENDATIONS TO GOVERNMENT AGENCIES

- ✓ Implement targeted safety measures in high-risk areas such as Kingston upon Hull, particularly at identified streets (e.g., Faversham Avenue, De Grey Street). Include speed cameras & enhanced lighting.
- ✓ Deploy traffic police and enforce speed limits during peak accident hours (8:00 AM and 5:00 PM) and Fridays, which show the highest accident rates.
- ✓ Install additional crossings and introduce stricter pedestrian zones near schools to address the 3:00 PM accident peak for pedestrian-related incidents.
- ✓ Given that accidents are frequent under dry road conditions and fair weather, focus on enhancing road infrastructure in these conditions. For instance, ensure that road markings are clear and that there is adequate signage in areas prone to accidents

SOCIAL NETWORK & ITS ANALYSIS

Using the provided dataset, a social network was constructed by representing individuals as nodes and their connections as edges. The network was visualized using the **spring layout algorithm**, which positions nodes to reflect their structural relationships. The visualization highlighted the complex and interconnected nature of the social network as shown below:

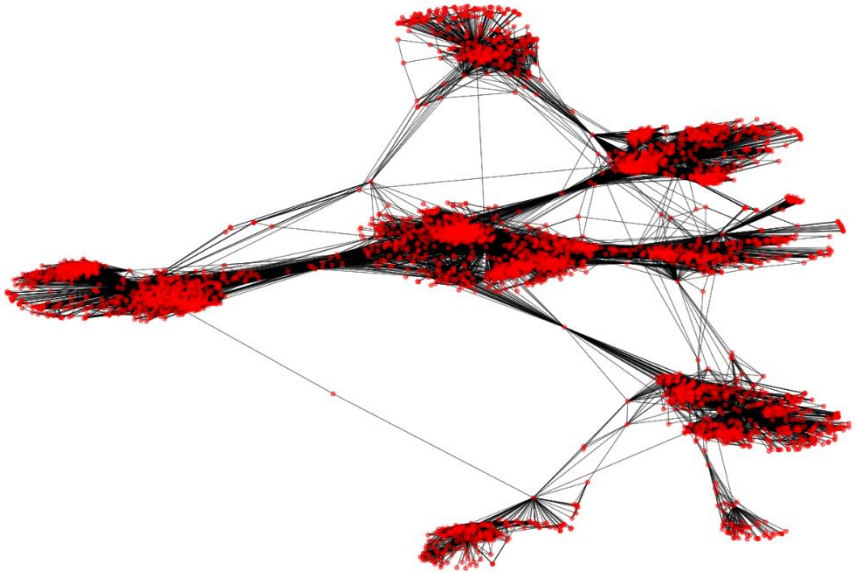

Figure 18: Social Network Visualization

Table 4: Network Characteristics

BASIC Network Characteristics		
Number of Nodes	4039	
Number of Edges	88234	
Network Density	0.0108	The **density** of 0.0108 indicates that only about 1% of all possible connections between nodes exist, which is typical for real-world social networks that tend to be sparse.
Average Degree	43.69	The **average degree** of 43.69 suggests that, on average, each individual (node) is connected to approximately 44 others. This reflects a high level of interconnectivity within the network.

EDGE CENTRALITY

Edge Betweenness Centrality was computed to identify the importance of edges in connecting different parts of the network. Edge centrality measures how often an edge lies on the shortest path between pairs of nodes.

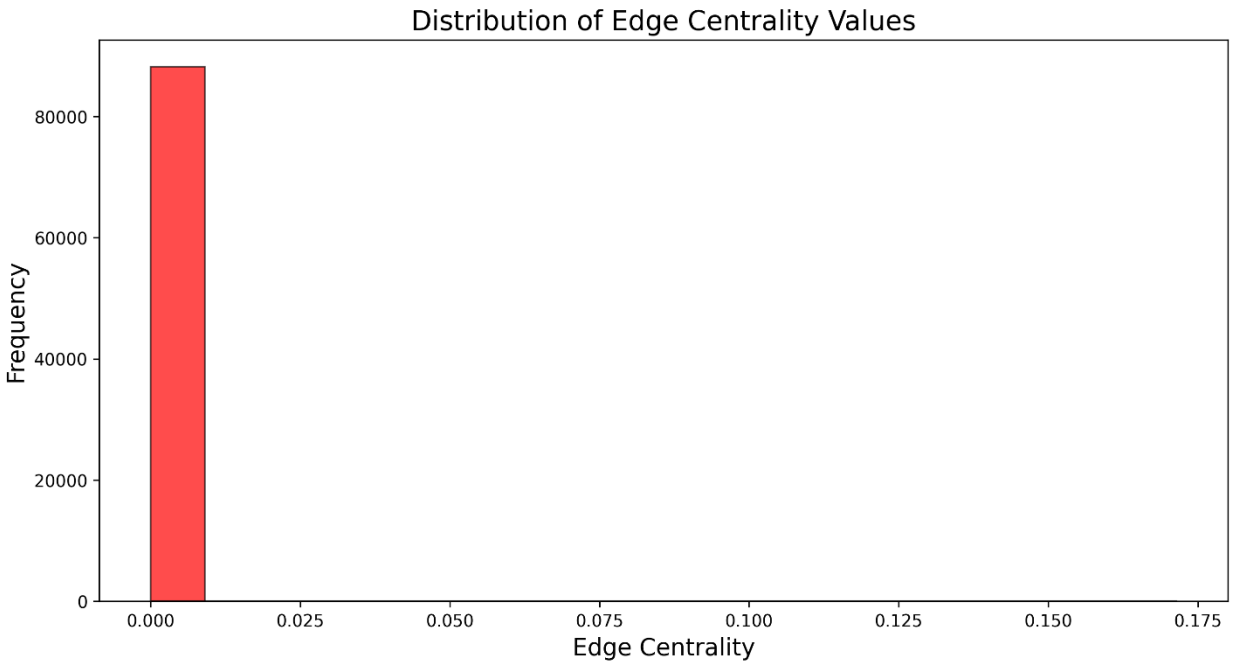

Figure 19: Edge Centrality Distribution

Table 5: Edge Centrality Characteristics

Range of Edge Centrality		
Minimum Edge Centrality	1.23×10−7	Very small value, indicating that most edges have a negligible impact on the overall network's connectivity
Maximum Edge Centrality	0.1715	Relatively larger value, showing that a small number of edges play a crucial role in connecting different parts of the network

COMMUNITY DETECTION AND COMPARISON

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

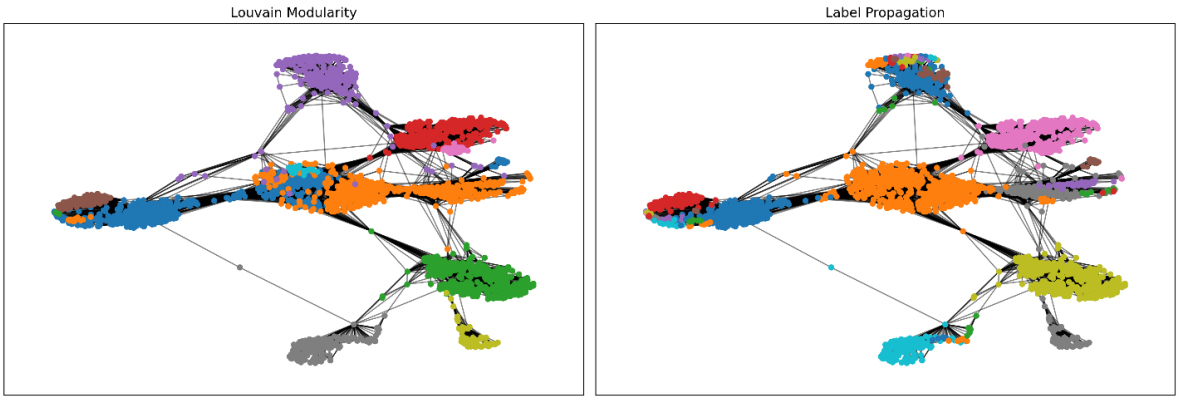

Figure 20: Social Network Visualization (Louvain vs. Label)

Table 6: Algorithms output & Performance

	LOUVAIN MODULARITY	LABEL 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	91.80
Modularity Score	0.7774	0.7368

COMPARISON

- ✓ Louvain Modularity achieved a higher modularity score (0.7774) compared to Label Propagation (0.7368), indicating better-defined community structures.
- ✓ Louvain's fewer, larger communities suggest more cohesive groups, while Label Propagation's numerous smaller communities may reflect finer-grained divisions within the network.
- ✓ Louvain's communities were relatively balanced in size, whereas Label Propagation's results exhibited greater variability, with some very small clusters (as small as 2-3 nodes).

JUSTIFICATION OF ALGORITHMS

- ✓ Louvain Modularity is ideal for uncovering broader, clearly defined group structures, making it effective for strategic decision-making, such as targeting major clusters in marketing.
- ✓ Label Propagation, being computationally efficient, is suited for dynamic, large-scale networks where identifying smaller, fine-grained communities is essential, such as studying niche social interactions.
- ✓ Both algorithms complement each other: **Louvain Modularity** provides insight into the overarching structure, while **Label Propagation** captures more granular community dynamics. Together, they ensure a comprehensive analysis of the network.