1. INTRODUCTION

Road traffic accidents are a major public health concern in any economy. This research uses 2020 United Kingdom traffic accident data to demonstrate an extensive analysis of the trends, accident patterns, and possible risk factors. Specific information in the dataset includes the casualty types, vehicle type, location, timing, etc. The study aims to use statistical analysis and predictive modelling to identify when, where, and under what conditions an accident occurs to support government road safety measures, and social network analysis to explore broader connectivity patterns.

2. Analysis

2.1 Focusing on “when” accidents happen.

This section answers the question below:

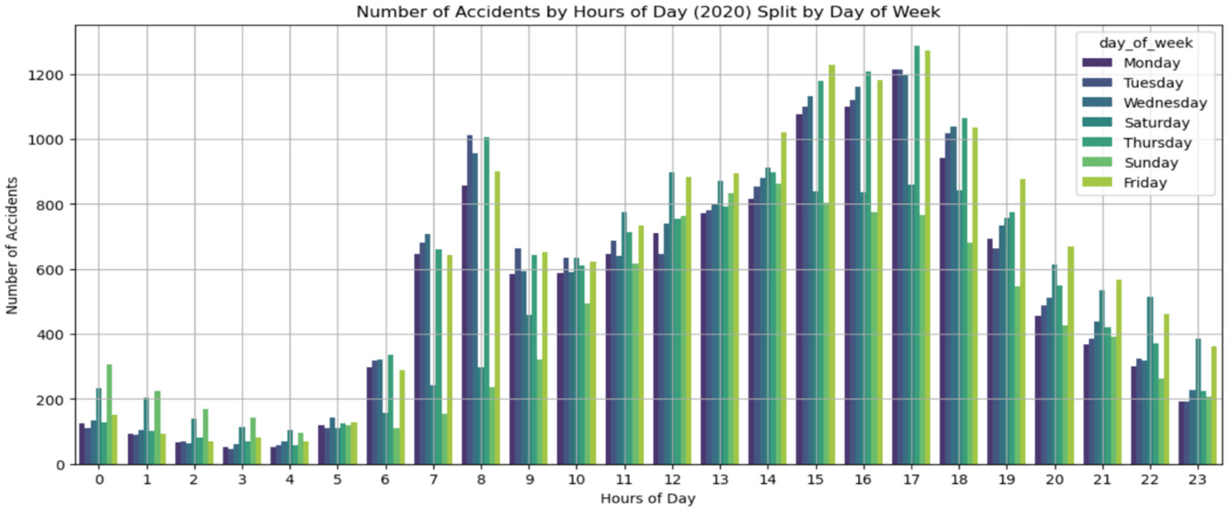
Significant hours of the day and days of the week, on which accidents occur

Data Preparation

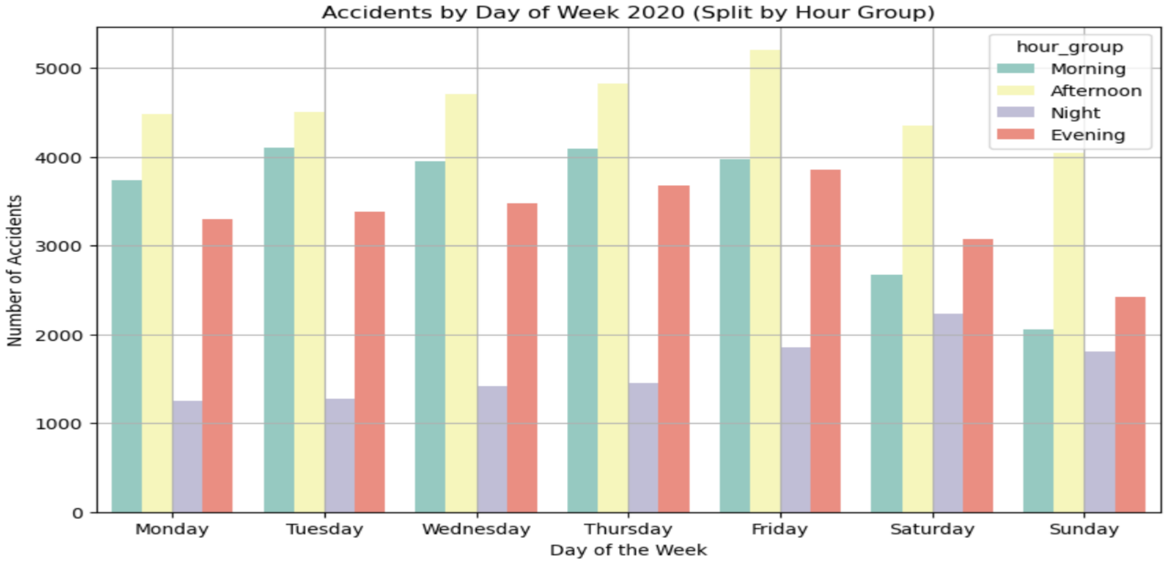
The year 2020 incidents were extracted from the accident table, focusing on the date, time, and day of the week columns. Time values were changed to a 24-hour format to support accident counts by hour, and day of the week codes were changed to weekday names for better interpretation.

Findings

Number of accidents by hour of the day



Morning and evening accident rates show around 8 AM and 5-6 PM, which confirms normal commuting hours. The accidents decrease during the late-night hours and early morning hours (1 AM – 5 AM). The accident rate starts to increase from 6 AM and spikes during the rush hour period.



The day_of_week bar plot shows that accidents are more persistent on weekdays, mostly on Fridays, indicating it may be the continuous build-up of tiredness over time. Sunday recorded the lowest number of accidents, suggesting lower weekday traffic congestion. The heatmap combining hour vs. weekday disclosed more information that weekday mornings to evening rush hours can be considered high-risk periods, and Friday evening stood out as the single riskiest time block across all combinations, likely due to an increase in after-work and weekend travel.

Interpretation

This analysis shows that road accidents are remarkably related to human activity patterns, especially commuting behaviour.

Analysing motorcycle accident patterns by hours of the day and day of the week.

Motorcyclists are often a vulnerable group in traffic, due to their exposure compared to cars.

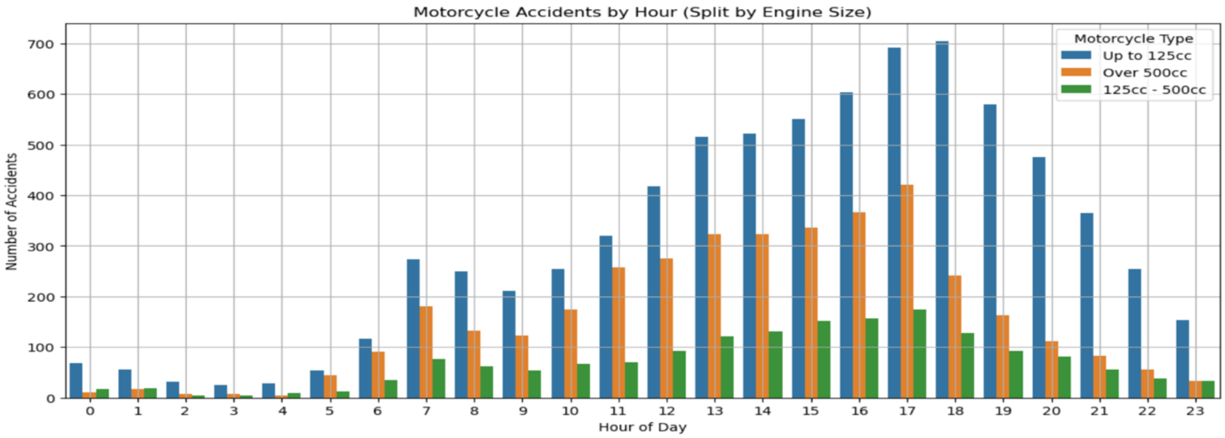
Data Preparation and Interpretation.

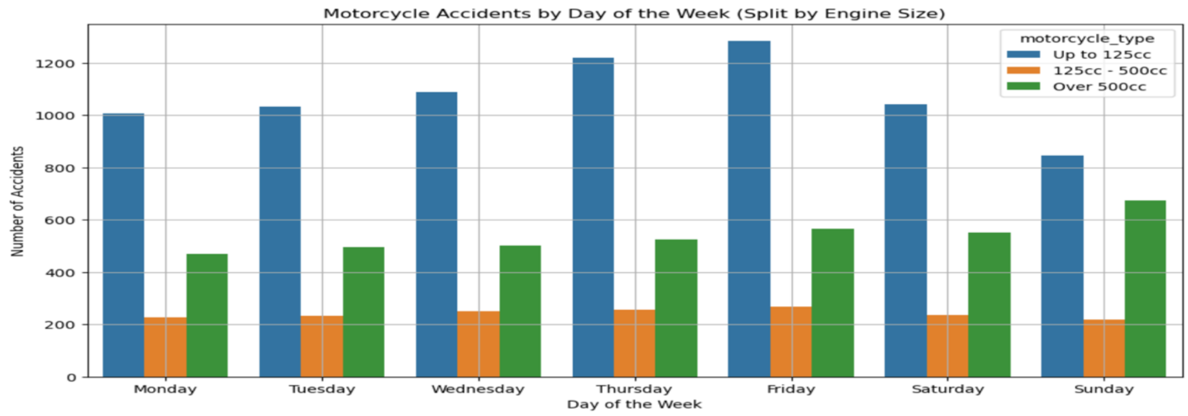
Motorcycle accident records were filtered by engine size and joined with accident date, time, and day data. Analysis showed that 125cc and under motorcycles had accident peaks during morning (7 AM) and late afternoon (3–6 PM) weekday rush hours, especially on Wednesdays and Fridays, with fewer accidents at night.

The result shows that motorcycles over 125cc and up to 500cc had a steady distribution across the week, accidents were high around 5 PM and weekends, implying various leisure uses. The bar plot also reveals that motorcycles over 500cc mostly had accidents on weekends at noon time (1-5 PM), which means heavy recreational use.

Findings

- 125cc and under motorcycles are mainly for commuting
- Over 125cc and up to 500cc motorbikes are for both utility and leisure
- Over 500cc are mainly leisure motorbikes.



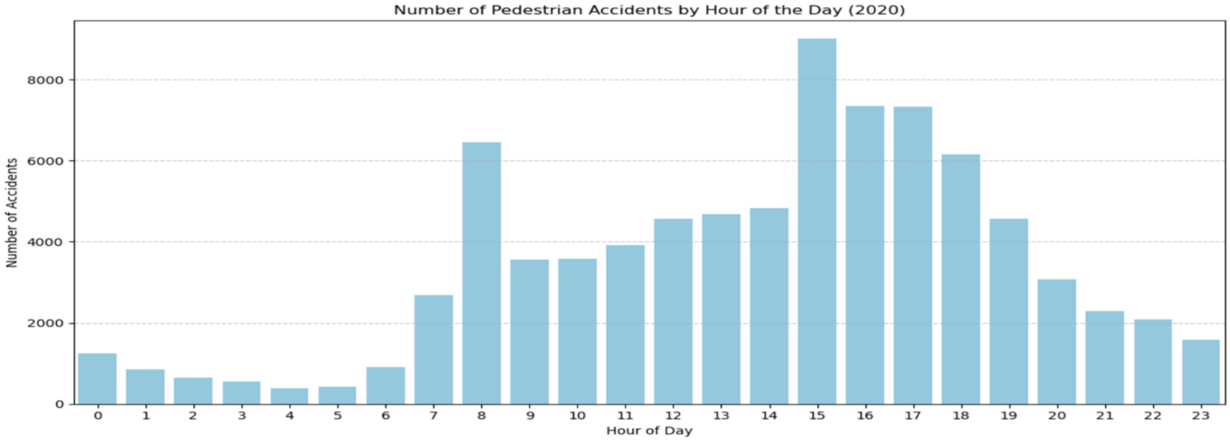


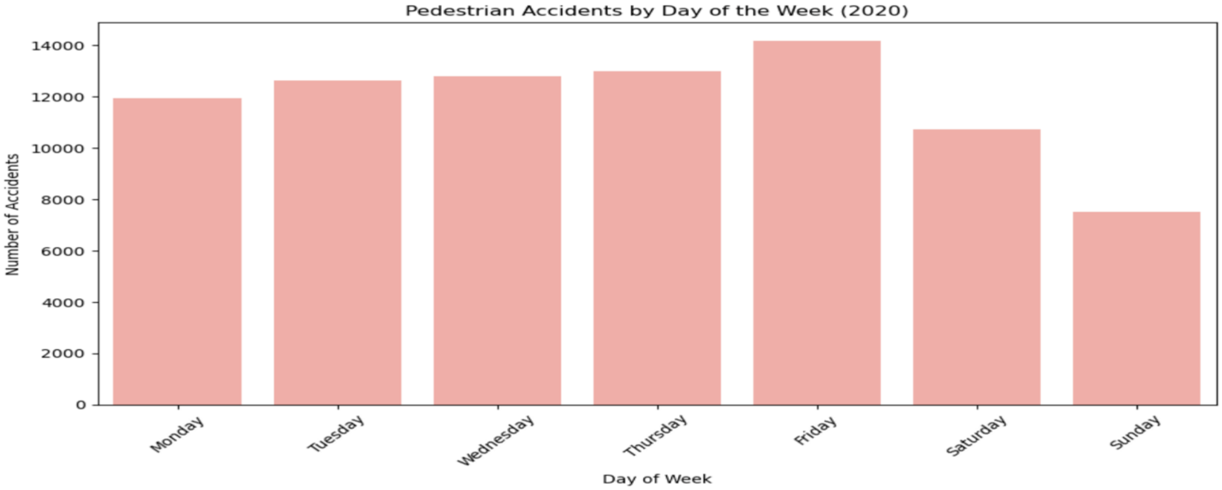
Analysis for pedestrian accident patterns by hours of the day and days of the week.

Data Preparation

The casualty table was filtered to select only casualties where the casualty_type was coded as 0 (Pedestrian). The results were then joined with the accident table using the accident_index, which allowed the gain of important fields like date, time, and day of week. Hour was extracted from the time field, and weekday names from the day_of_week column.

Findings





Hours of the day pose two major peaks: (1) Morning peak between 7 AM and 9 AM, (2) Evening peak between 3 PM and 6 PM. Accidents decrease across the late night and early morning hours, 11 PM to 5 AM.

For the day of week, weekdays (Monday–Friday) had a glaringly vast number of pedestrian accidents than weekends. Friday noted the highest number of pedestrian accidents in total, and Sunday showed the lowest pedestrian accident figures.

Interpretation

This analysis reveals that morning and afternoon peaks tally closely with school commutes, work travel, and shopping trip activities that increase pedestrian movement on streets. High accident rates on Fridays may be related to increased traffic. The lower accident rates on weekends could be due to decreased pedestrian exposure.

2.2 Focusing on “where” accidents happen

This section answers the following questions:

Clustering for Kingston upon Hull and surrounding areas.

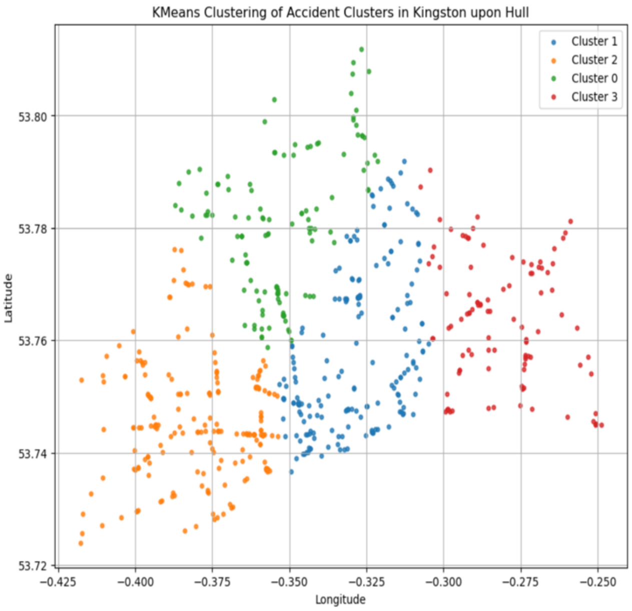
Data extraction

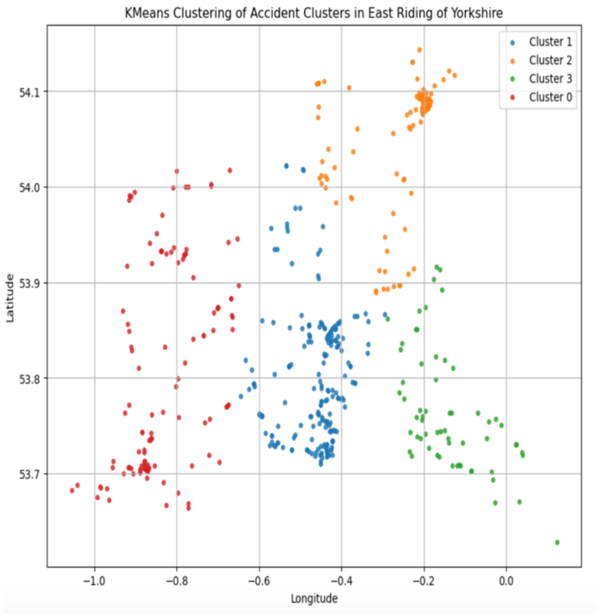
The accident dataset using LSOA codes and local authority information was filtered, and entries were matched specifically where the Lsoa_of_accident_location or local_authority_district codes indicated the target regions.

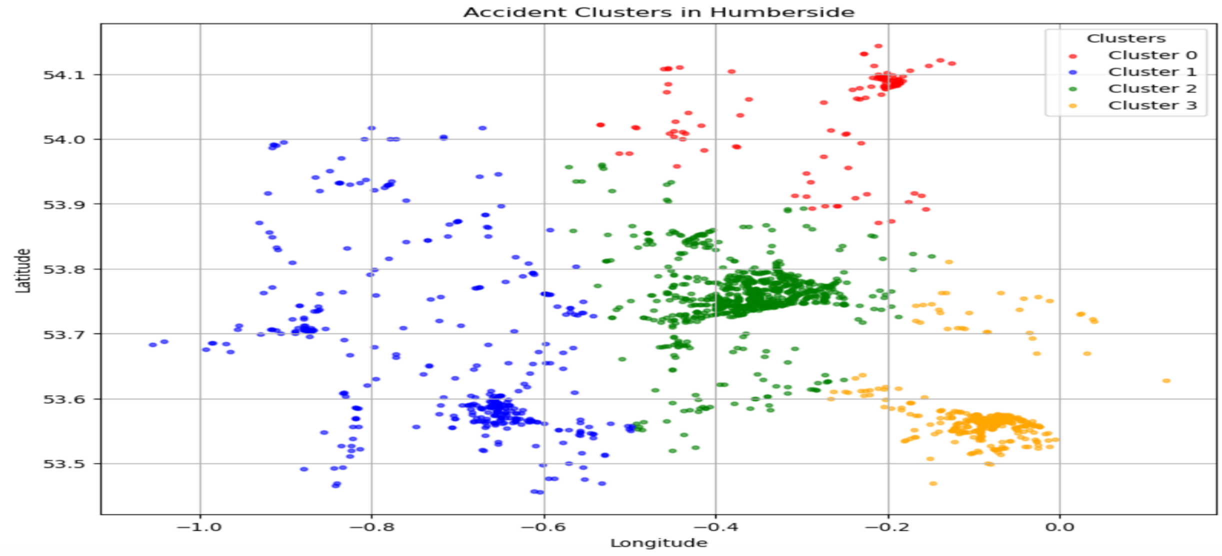
Clustering Method

The geographic coordinates (longitude and latitude) of each accident were selected, and K-means clustering was applied. The optimal k value, k = 4, was chosen where the line starts to flatten out. Standard scaling was applied to latitude and longitude before clustering to prevent scale bias.

Results and Interpretation







Four clusters were distributed across different regions as shown in the scatter plots above. High-density clusters suggest busy areas in the regions, indicating that accidents center more around key urban areas, highways, and major transport junctions.

Justification for selecting k = 4

The application of the elbow method to identify the optimal k; points where the line starts to flatten out, approximately at point 4 in the elbow curve. Justifying the reason for choosing the selected number of clusters, which offers a good balance between model efficiency and capturing key accident hotspots, avoiding underfitting and overfitting.

Time series predicting weekly accident counts for three police forces.

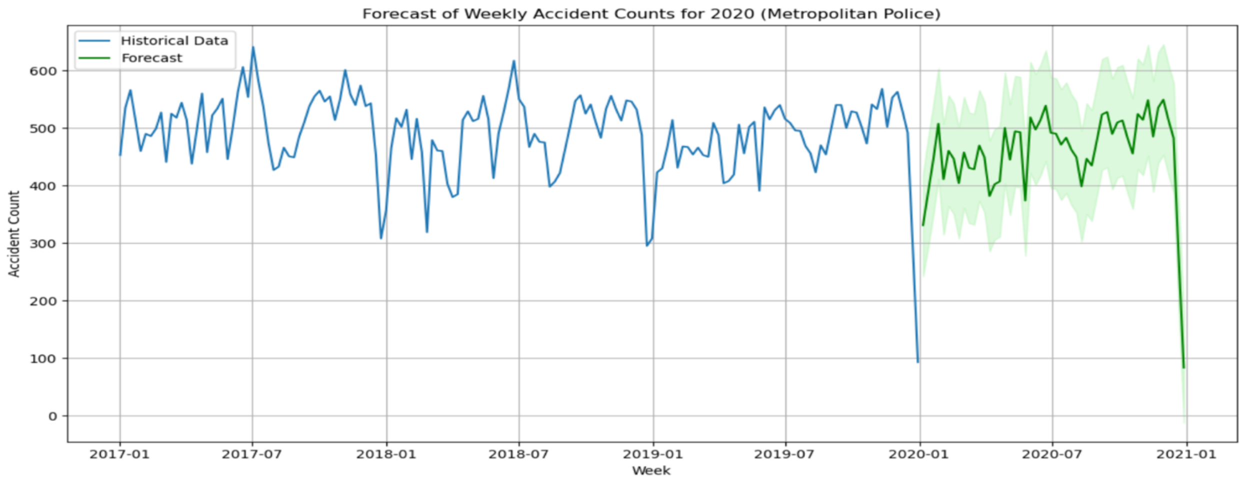
Data Extraction

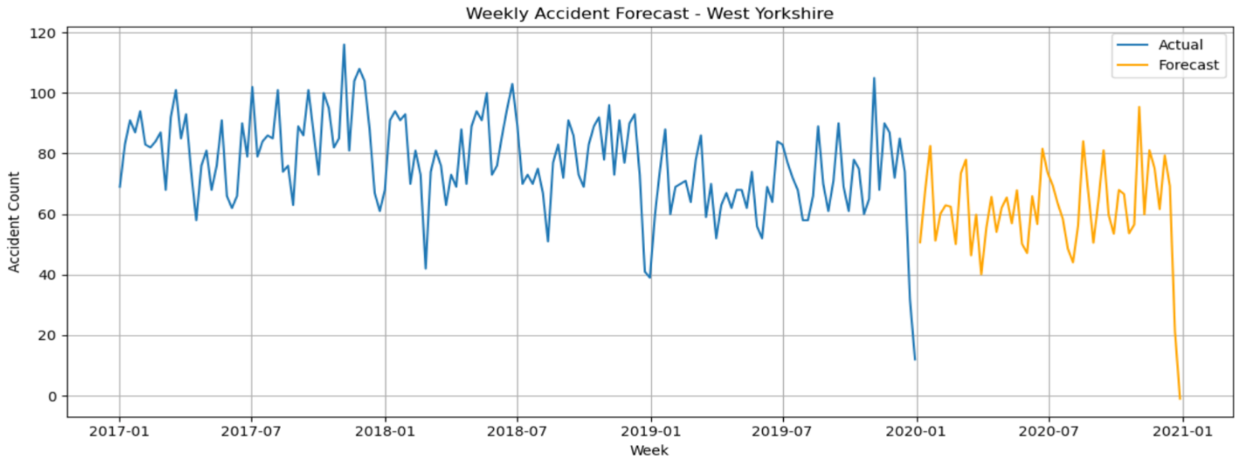
To build the time series models, selected accident records were filtered by three police forces: Metropolitan police (code 1), West Yorkshire police (code 13), Humberside Police (code 16). The data was restricted to accidents occurring between 2017 and 2019 to form the historical training set. The accident data was used to aggregate the data into weekly counts for each police force separately.

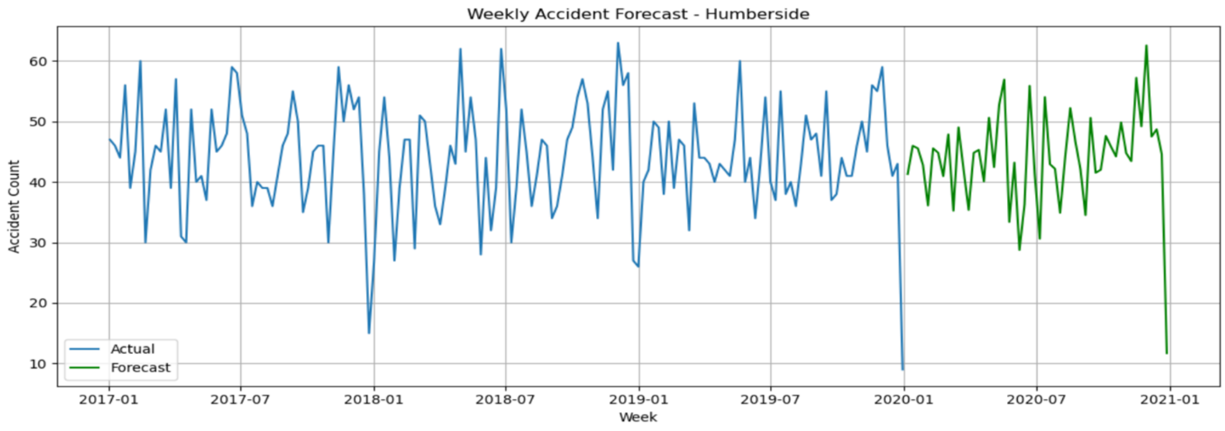
Model Choice and Findings

SARIMA model was chosen for forecasting because it captures both seasonality and trends, it is robust for univariate time series with repeated seasonal patterns. The SARIMA models fit reasonably well for each region. Seasonal patterns were observed in the data, especially for the Metropolitan Police, showing peaks during summer months and dips in winter. Which suggests high road users, high speed use during the summer months.

Larger regions, such as the Metropolitan Police, had better prediction accuracy because of their robust dataset, creating more opportunities to identify accident patterns and trends. Smaller police forces (Humberside) showed more noise and irregularities in their time series, making predictions slightly less stable.







Key Challenges:

A few key challenges to point out.

1. Limited historical data (2017-2019), restricted long-term seasonality learning.

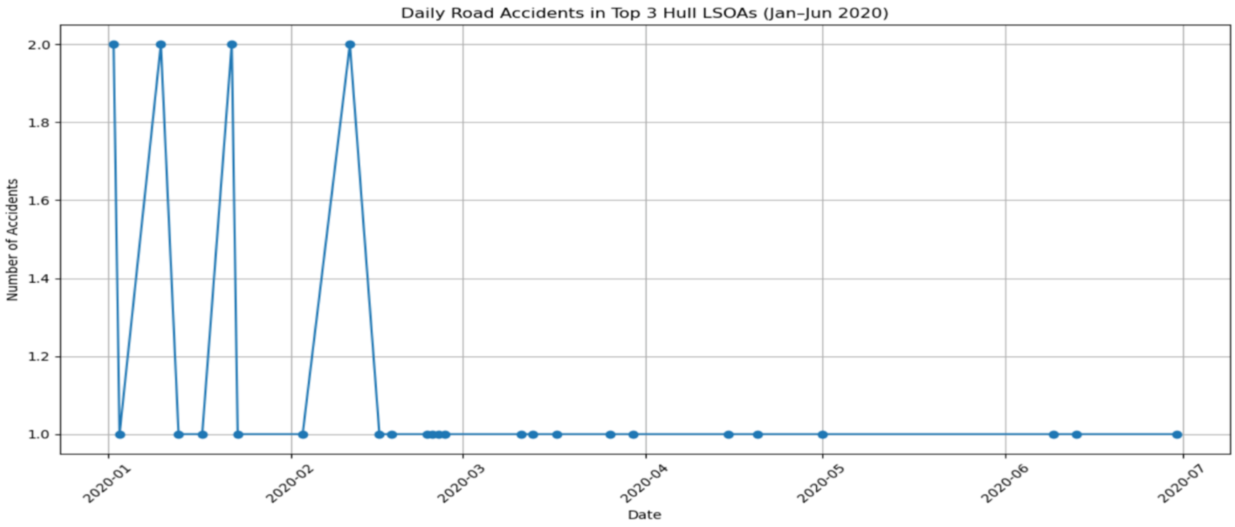
2. The COVID-19 pandemic affected accident patterns in 2020, causing anomalies in the trends.

Forecasting Daily Accident Counts for the Top 3 Hull LSOAs

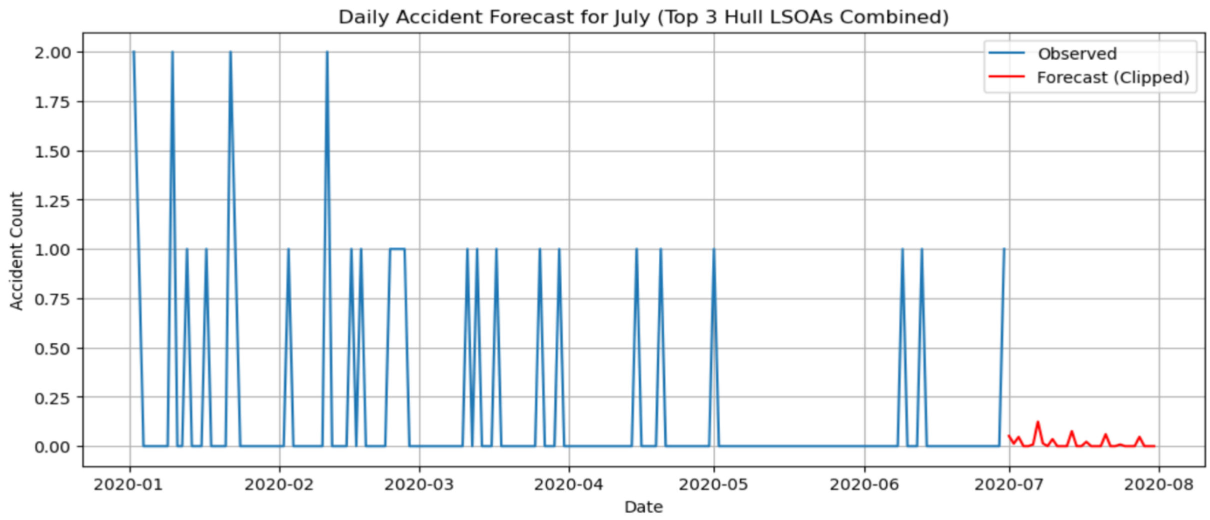
The top 3 Hull LSOAs with the highest accident counts during January–March 2020 were selected, and a SARIMA model was built to forecast the daily counts.

Findings

The model predicted low daily accident counts (typically between 0-2 accidents/day per LSOA). Accidents were relatively rare on a day-to-day basis, making prediction performance harder to evaluate purely on RMSE (Root Mean Square Error). Forecasted trends remained stable, with occasional small spikes aligning with observed historical surges (weekends).



Daily Road Accident Count,



Forecast of Daily Road Accidents.

Key Challenge

Sparse data (showing very few accidents on many days) made modelling more sensitive to random variations.

2.3 Focusing on "under what conditions" accidents happen.

This section will answer the following question:

Evaluating factors impacting accident severity using the Apriori Algorithm.

Data preparation

The accident severity subset variables important for the analysis were filtered. Weather conditions, road surface conditions, and light conditions. Accident severity was encoded into categories (slight, serious, fatal) based on the existing labels. The dataset was transformed from numeric codes to categorical string values, given that Apriori requires categorical data. Each selected feature was one-hot-encoded, giving itemset Light_1 (code 1 representing 'Day_light in the field name; light_conditions, according to the accident spreadsheet) we had items like Surface_Dry, Light_Daylight. Etc, accident severity was also incorporated. The Apriori algorithm was employed with a minimum support threshold of 0.01, denoting that the rule must apply to at least 1% of accidents. Confidence threshold of 0.6, stating how often the rule is correct. Lift values were considered to filter out non-informative rules; lift > 1 indicates strong association. Association rules were then generated linking combinations of conditions to different severities.

Results

It shows fine weather and dry road surfaces were frequently associated with slight accidents, suggesting that environmental factors alone do not guarantee safety. Wet or icy roads combined with poor lighting (e.g., nighttime without street lighting) were associated with serious accidents.

Interpretation reveals that the severity of accidents is multi-factorial, often resulting from combinations of environmental conditions and vehicle types. Motorcyclists are very exposed, particularly under poor visibility and adverse surface conditions. Justification for the Apriori algorithm was best because it highlights frequent itemsets and discovers hidden patterns, it does not assume any causality, only association, and handles categorical and binary data very well.

3. Network Analysis.

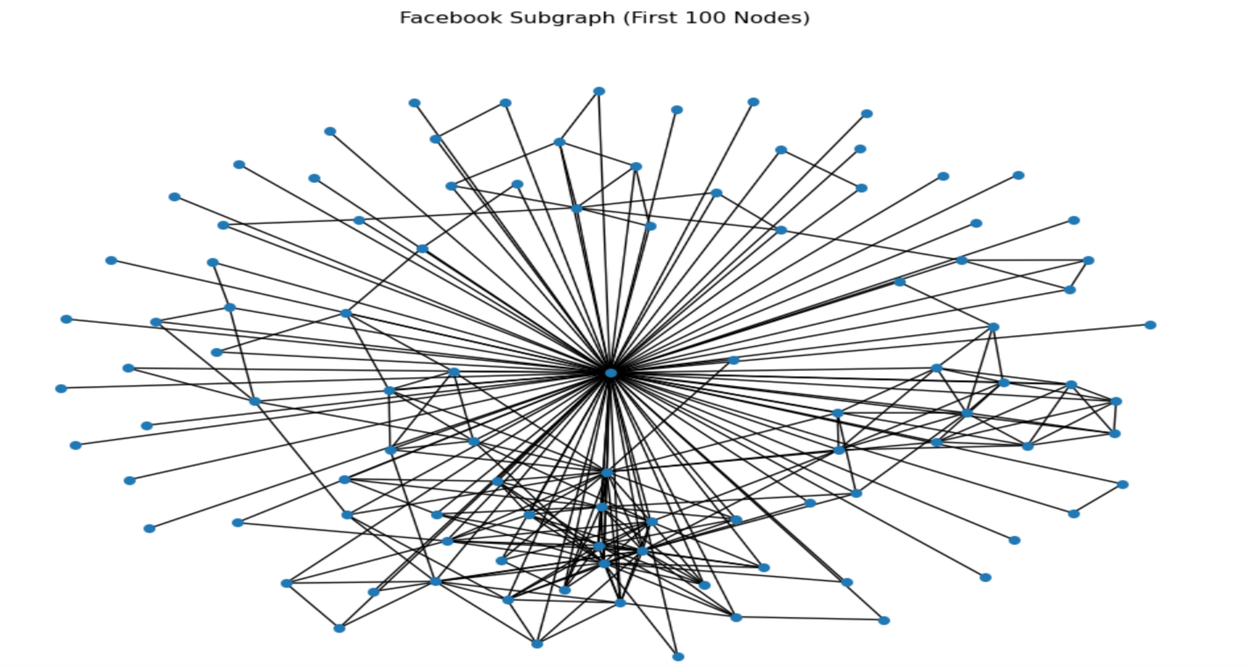
Social Network Construction and Basic Network Analysis.

Data Source

The facebook_combined.txt dataset was provided; it contains an edge list format where each line represents a connection (friendship) between two users.

Network Construction, Result, and Interpretation

Networkx Python library was used to read the edge list, create an undirected graph, because friendships are mutual.



Undirected Facebook subgraph.

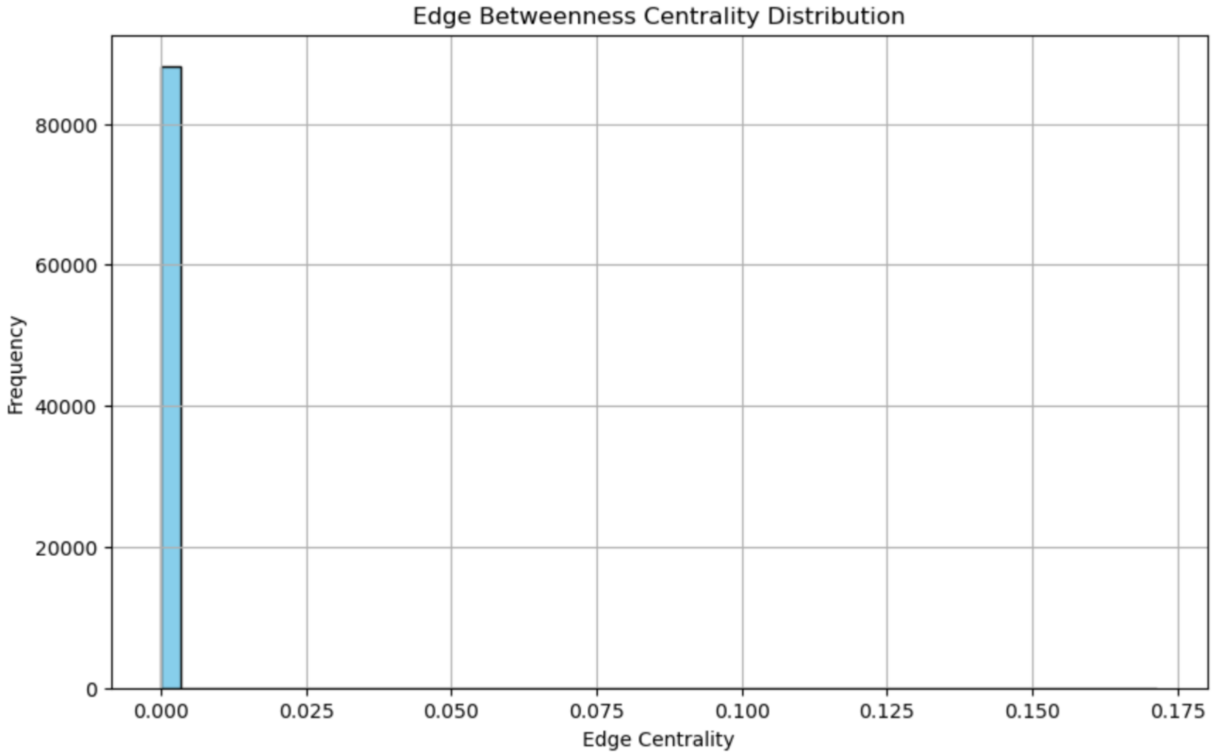
Basic Network Characteristics

Characteristic	Result
Number of nodes	4,039
Number edges	88, 234
Network Density	0.0108
Average Degree	43.68

Nodes records 4039 users in this Facebook dataset, edges records 88,234 friendship connections exist between users, the density is only about 1.08% of all possible connections realised, and average degree records each user has about 44 friends on average. Key observations reported that the network is highly clustered, and there are many “friendship groups” or local communities.

Edge Centrality Analysis

Edge centrality describes the number of shortest path between other pairs of nodes. It determines the key connections that act as bridges in the network. Edge betweenness centrality was computed for every edge using NetworkX's edge_betweenness_centrality() function.



Distribution of Edge Centrality Values.

Findings from the Distribution Plot and Interpretation

Most edges have a very low centrality because they are not important for connecting major parts of the network. The distribution is highly skewed (large peak near zero).

The network structure relies heavily on a small number of key friendships to maintain global connectivity. This behaviour is typical of real-world social networks, where tight groups are strongly connected internally.

Community detection in the Social Network

Two commonly used community detection algorithms (Girvan-Newman algorithm and Louvain algorithm) were used to identify clusters in the Facebook social network dataset provided, which gave the following analysis.

- Girvan-Newman Algorithm

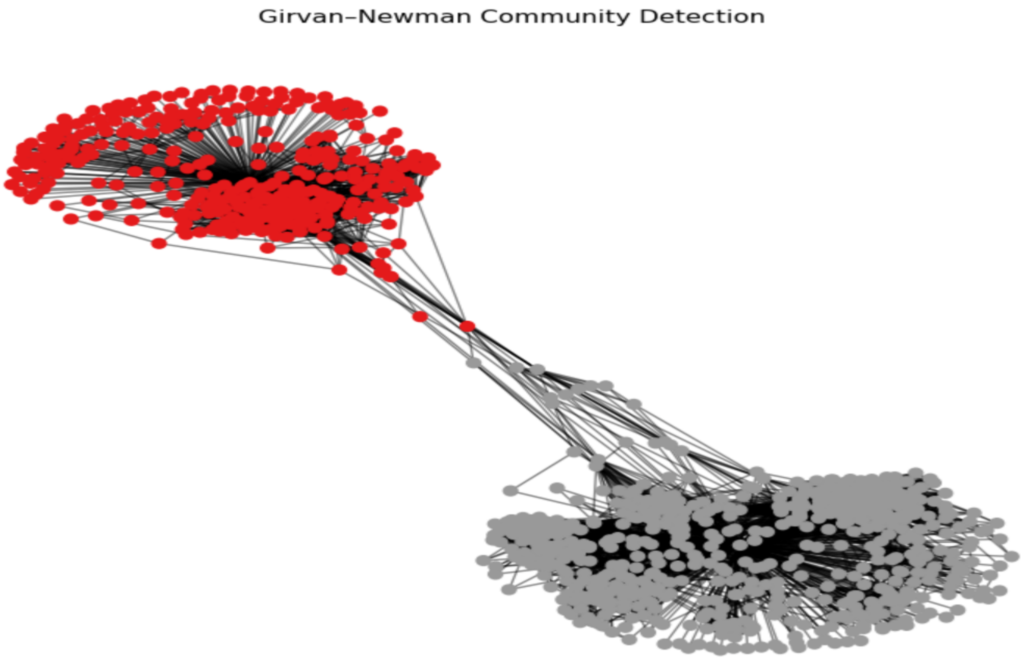
The algorithm iteratively removes the most valuable edges (edges with the highest betweenness), leading to the identification of other large, unique communities.

Results: After the Girvan algorithm was applied and the most valuable edges had been removed, 2 communities were detected in the sampled network of 1000 nodes: community 1: 344 nodes, community 2: 656 nodes.

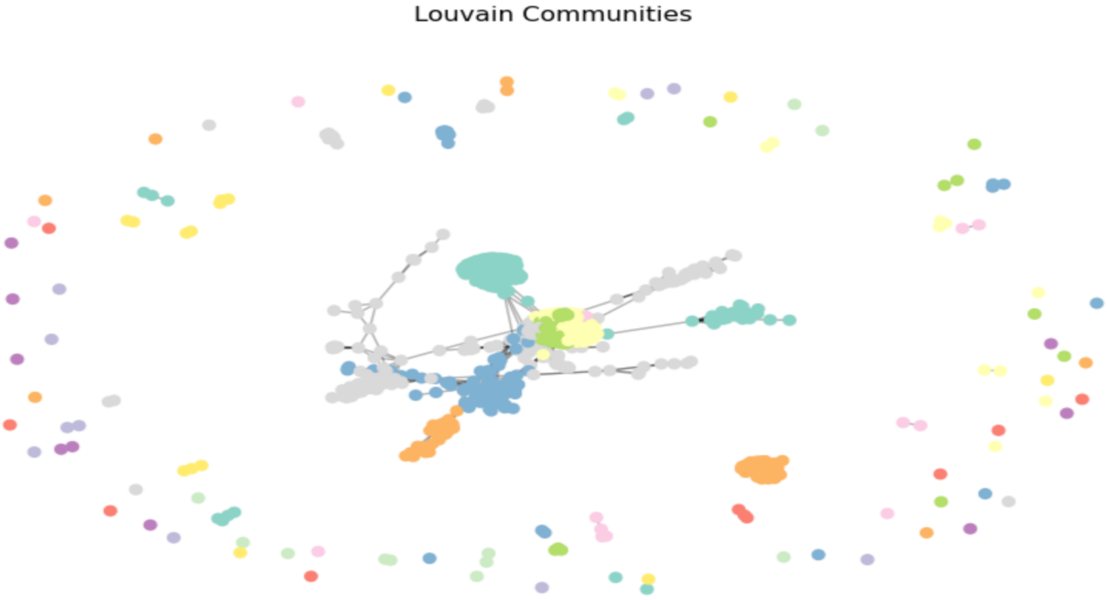
- Louvain Algorithm

The Louvain method optimises modularity to find dense communities within the network.

Results discovered were 103 communities of uneven sizes, varying from large clusters (e.g., 141, 119, 111 nodes) to many small and isolated communities.



Girvan-Newman 2 sampled community detection.



Louvain 103 community detection.

Comparison

- Girvan-Newman gives a broader division, ideal for high-level community structure.

- Louvain reveals finer-grained communities and substructures.
- Louvain is also more scalable and efficient on large graphs.

Both results show that the network is hugely modular, with many closely connected groups, a typical characteristic of social networks.

4. Recommendations

According to the knowledge drawn from the 2020 road traffic accident dataset analysis, the following recommendations are made.

- Increase Road Safety Policies During Peak Accident Times.

Late afternoons and evenings (15:00 – 18:00) and especially on weekends are the high point of accidents. The enforcement of traffic law should be intensified by government agencies, and carry out awareness campaigns during these peak hours and days to reduce the chance of accidents.

- Focus Motorcycle Safety Campaigns on Key Periods

There is a high accident rate during weekend afternoons, especially for motorcycles over 500cc, as shown by the distribution of incidents in the plots. Safety programs should be targeted at motorcyclists during weekends, including promoting protective gear and safe riding practices, should be encouraged.

- Implement Pedestrian Protection Strategies

More pedestrian crossings should be established, improve street lighting, and with pedestrian safety awareness campaigns in accident hotspots to reduce pedestrian casualties.

- Monitoring Emerging Trends with Ongoing Data Analysis.

Using methods like SARIMA can help establish a real-time data monitoring system and seasonal prediction models to assist in proactively responding to changes in accident trends.

Conclusion

Accidents peaked during rush hours and in key urban areas, which is mostly caused by poor conditions. The SARIMA model linked the 2020 accident drop to COVID-19, and community detection showed the network is highly modular.

Reference

Traag, V.A., Waltman, L. and van Eck, N.J., 2019. From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports, 9, p.5233. Available at: https://doi.org/10.1038/s41598-019-41695-z