1.0 INTRODUCTION

Road traffic accidents are a serious yet often underestimated public health issue. Globally, they cause around Every year, road accidents cause around 50 million injuries and 1.2 million deaths worldwide. (Pawłowski et al., 2019). In the UK alone, the Department for Transport (2024) reported 1,607 road deaths and over 29,000 serious injuries between 2023 and 2024 contributing to a total of nearly 129,000 casualties. Road traffic injuries are projected to become the seventh most common cause of death worldwide by 2030 (Ahmed et al., 2023).

This report provides analysis on road accident data from 2020 was analysed using advanced methods to uncover patterns, forecast future risks, and offer practical recommendations to help improve road safety.

2.0 ANALYSIS, VISUALIZATION AND PREDICTION

The dataset is from the UK Government Accident database and includes records across four main tables: Accident, Vehicle, Casualty, and LSOA.

To get the dataset ready for analysis, the following preprocessing steps were taken:

- Filtered the data to only include records from 2020, as required for the project.
- Converted the date column to the UK's standard datetime format.
- Mapped categorical variables to numerical values where needed, making it easier to apply the Apriori Algorithm.
- Removed any rows with missing or invalid values such as -1 for the Apriori algorithm to clean up the data for analysis.
- Ensured the data was consistent and accurate to get the best results from the analysis.

- SIGNIFICANT HOURS OF THE DAY, AND DAYS OF THE WEEK WHICH ACCIDENTS OCCUR

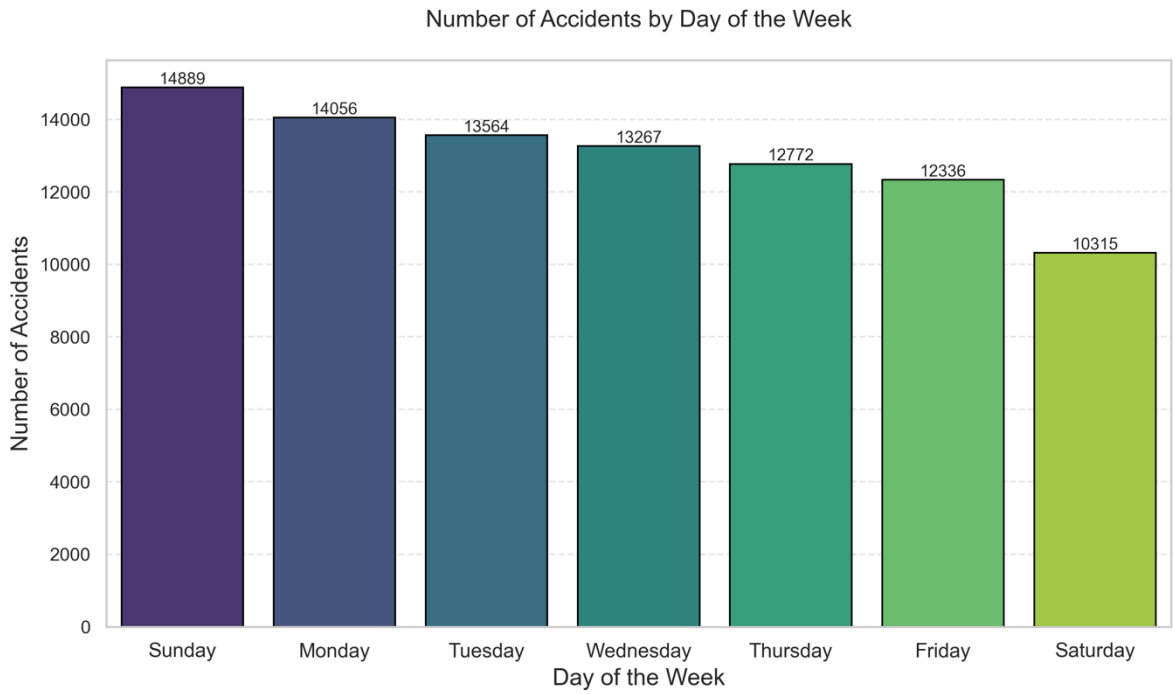

Figure 1: Number of Accidents by Day of the Week

Figure 1 shows that most accidents happen on Fridays, likely because of heavier traffic as people travel at the end of the week. Wednesday and Thursday have relatively high accident counts, while Sunday sees the fewest, which makes sense given the lower traffic volume on weekends when fewer people are commuting.

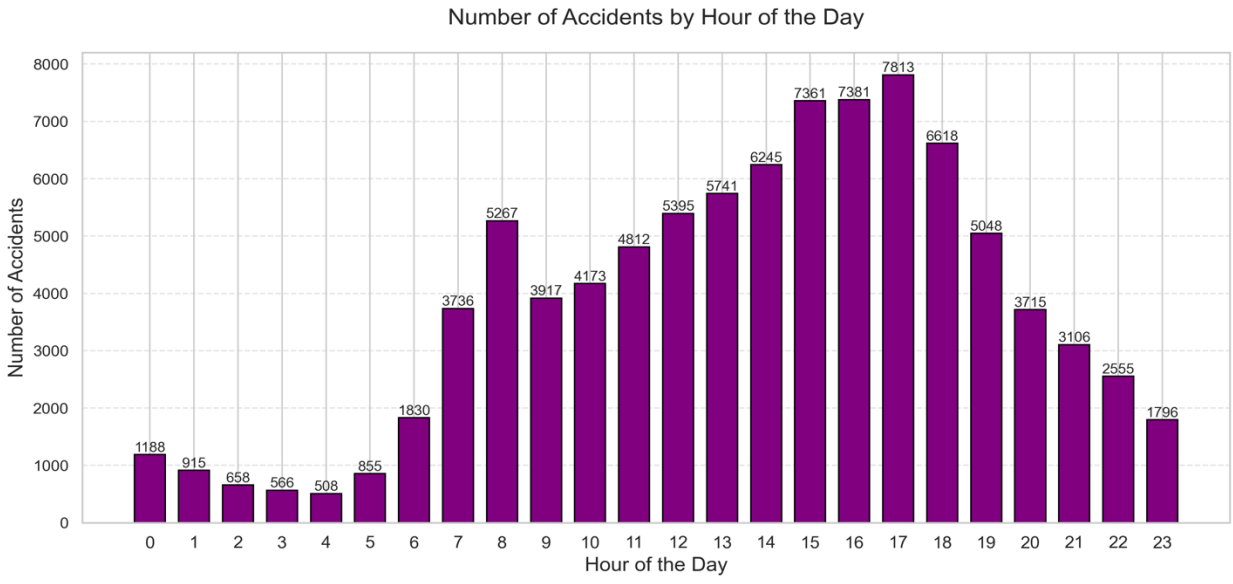

Figure 2: Number of Accidents by Hour of the Day

Figure 2 above shows that the highest accident counts 7813 happens at 17:00 (5PM) reflecting the time people retire from the day's activities.

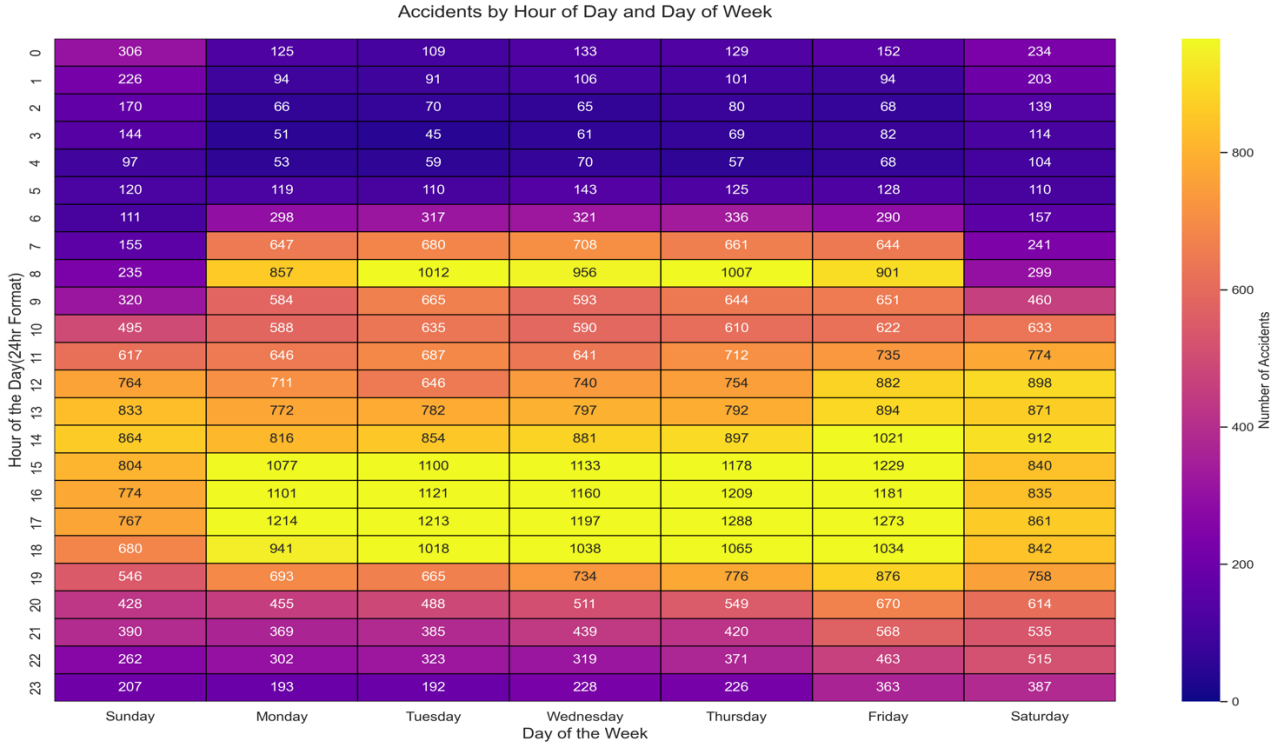

Figure 3: Heatmap of Accidents by Hour of the Day and Day of the Week

From the heatmap above the hours with the highest accidents marked in yellow, we can deduce that the hours of 8am and 3pm-6pm have the highest accidents. With 8am signifying people starting the day's activities such as going to work and hours between 3pm and 6pm signifying people returning from the day's activities. Reduced accident counts between the hours of 12AM - 6AM can also be observed.

- **SIGNIFICANT HOURS OF THE DAY, AND DAYS OF THE WEEK WHICH ACCIDENTS OCCUR FOR MOTORBIKES**

Motorcyclists have a much poorer safety record compared to other road users. In the UK, they are about twice as likely to be killed or seriously injured per million vehicle kilometres as pedal cyclists and over 16 times more likely than passengers and car drivers (Clarke et al., 2007). An SQL query was written to filter motorcycle types based on engine size. The accident statistics form did not explicitly have "Motorcycle 125cc and under", The available related categories were used to approximate it**.** "Motorcycle under 50cc" as a starting point. **"**Motorcycle over 50cc and up to 125cc" is the closest available category to cover the rest of the "125cc and under" range, the code for "under 50cc" was replaced with the one for "over 50cc and up to 125cc"**.** Motorcycle over 500cc and Motorcycle over 125cc and up to 500cc and were filtered too.

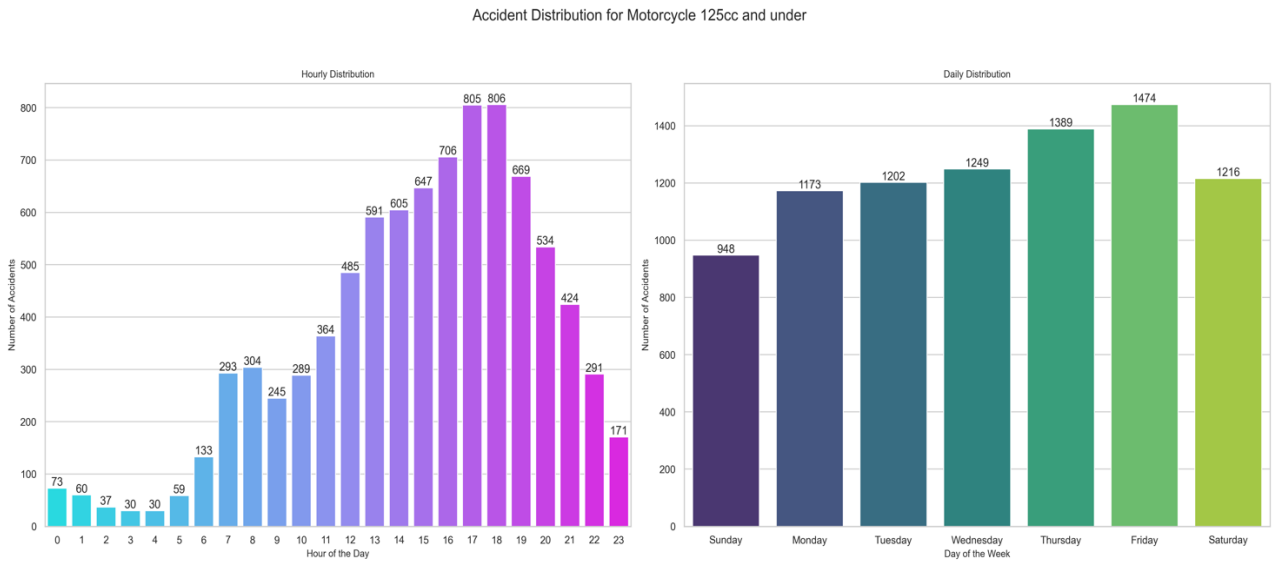

Figure 4: Hourly and daily accidents for Motorcycle 125cc and under

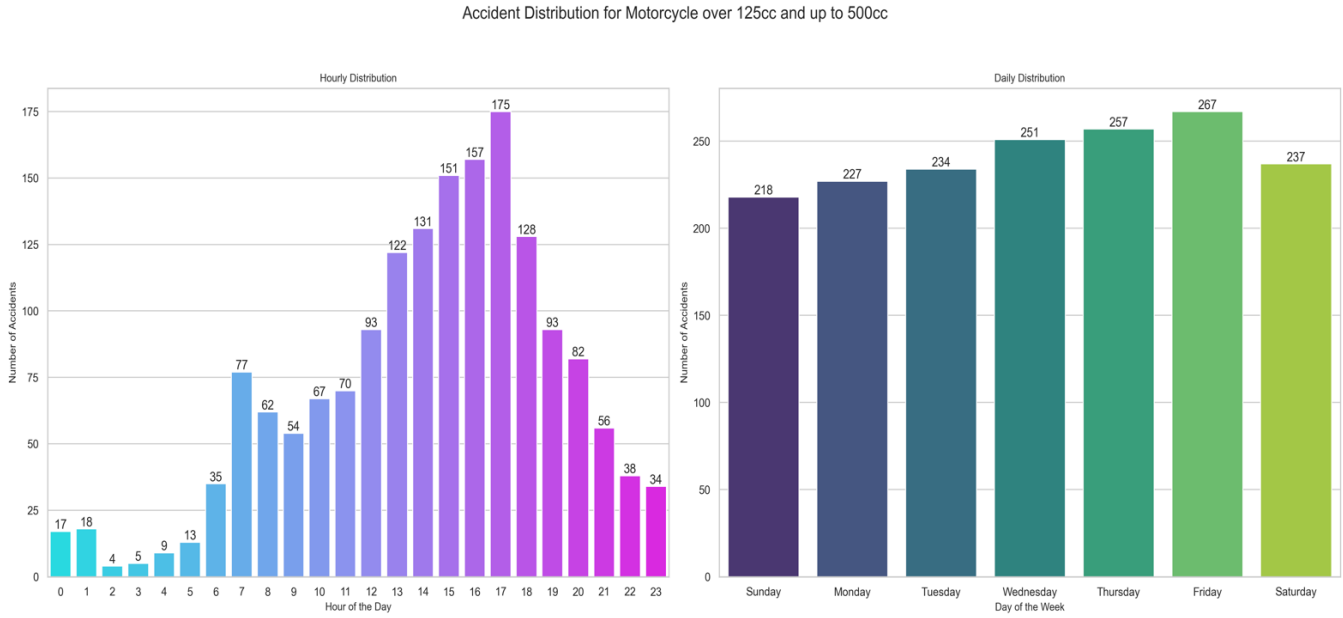

Figure 5: Hourly and daily accidents for Motorcycle over 125cc and up to 500cc

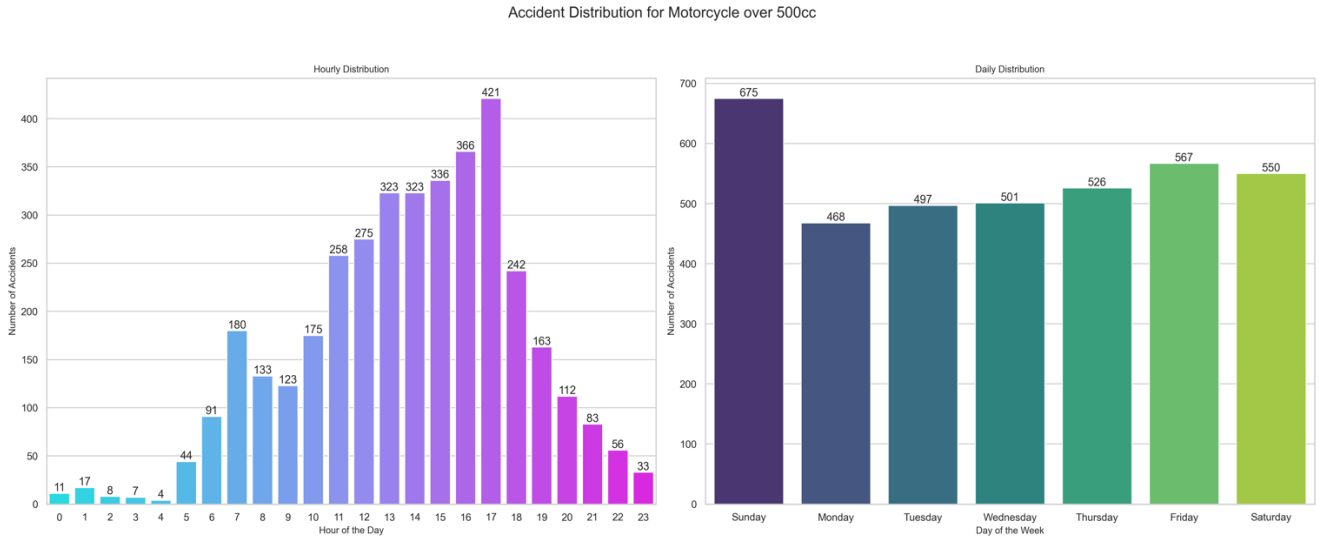

Figure 6: Hourly and daily accidents for Motorcycle over 500cc

Motorcycle Engine Types	3 Peak Accident Days	3 Peak Accident Hours
Motorcycle 125cc and under	Fridays- 1474 accidents Thursdays- 1389 accidents Wednesdays- 1249 accidents	18:00(6pm)- 806 accidents 17:00(5pm)- 805 accidents 16:00(4pm)- 706 accidents
Motorcycle over 125cc and up to 500cc	Fridays-267 accidents Thursdays-257 accidents Wednesdays-251 accidents	17:00(5pm)-175 accidents 16:00(4pm)- 157 accidents 15:00(3pm)- 151 accidents
Motorcycle over 500cc	Sundays- 675 accidents Fridays- 567 accidents Thursdays- 526 accidents	17:00(5pm)- 421 accidents 16:00(4pm)- 366 accidents 15:00(3pm)- 336 accidents

Table 1: Summary of Peak accident hours and day for motorbikes

The summary in Table 1 above shows Motorcycles under 125cc have a peak accident time at 6 PM (806 accidents) and a peak day on Fridays (1,474 accidents). Motorcycles over 125cc and up to 500cc peak at 5 PM (175 accidents) and on Fridays (267 accidents). Motorcycles over 500cc have their peak at 5 PM (421 accidents), but their highest accident day is Sunday (675 accidents).

- <u>SIGNIFICANT HOURS OF THE DAY, AND DAYS OF THE WEEK WHICH PEDESTRIANS HAVE ACCIDENTS</u>

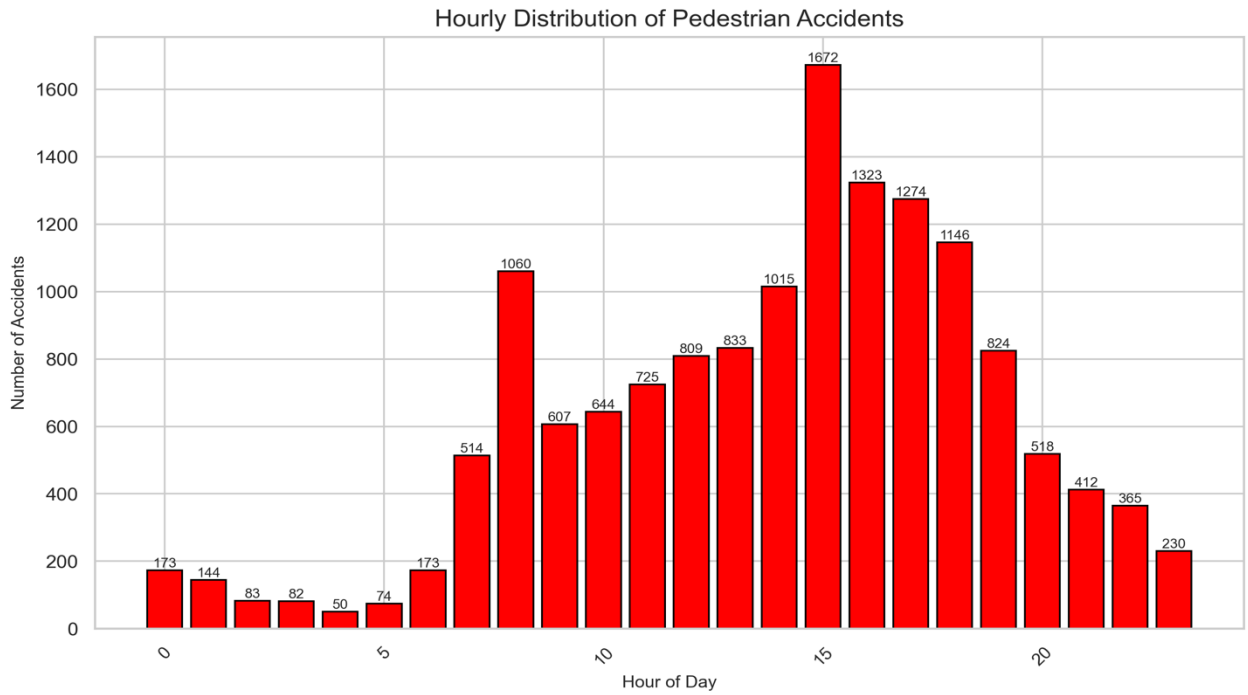

Figure 7: Significant hours of pedestrian accidents

From Figure 7 above pedestrian accidents peak at 15:00(3PM)

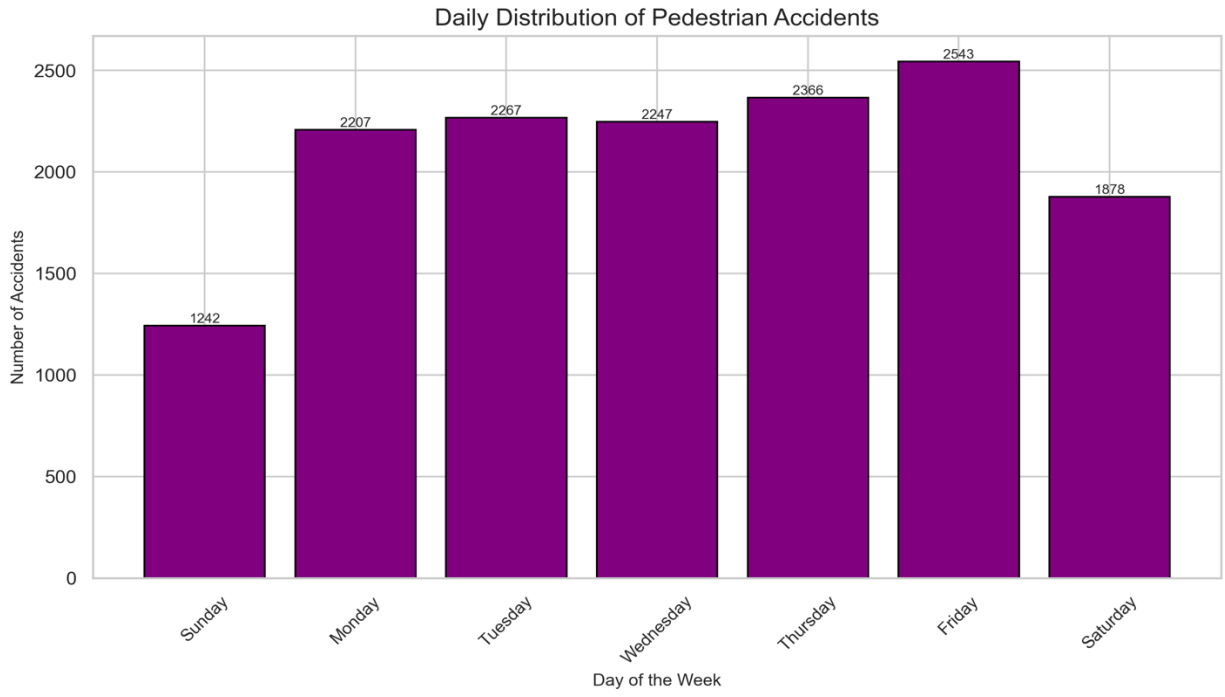

Figure 8: Significant Days of Pedestrian Accidents

From Figure 8 above pedestrian accidents peak on Fridays.

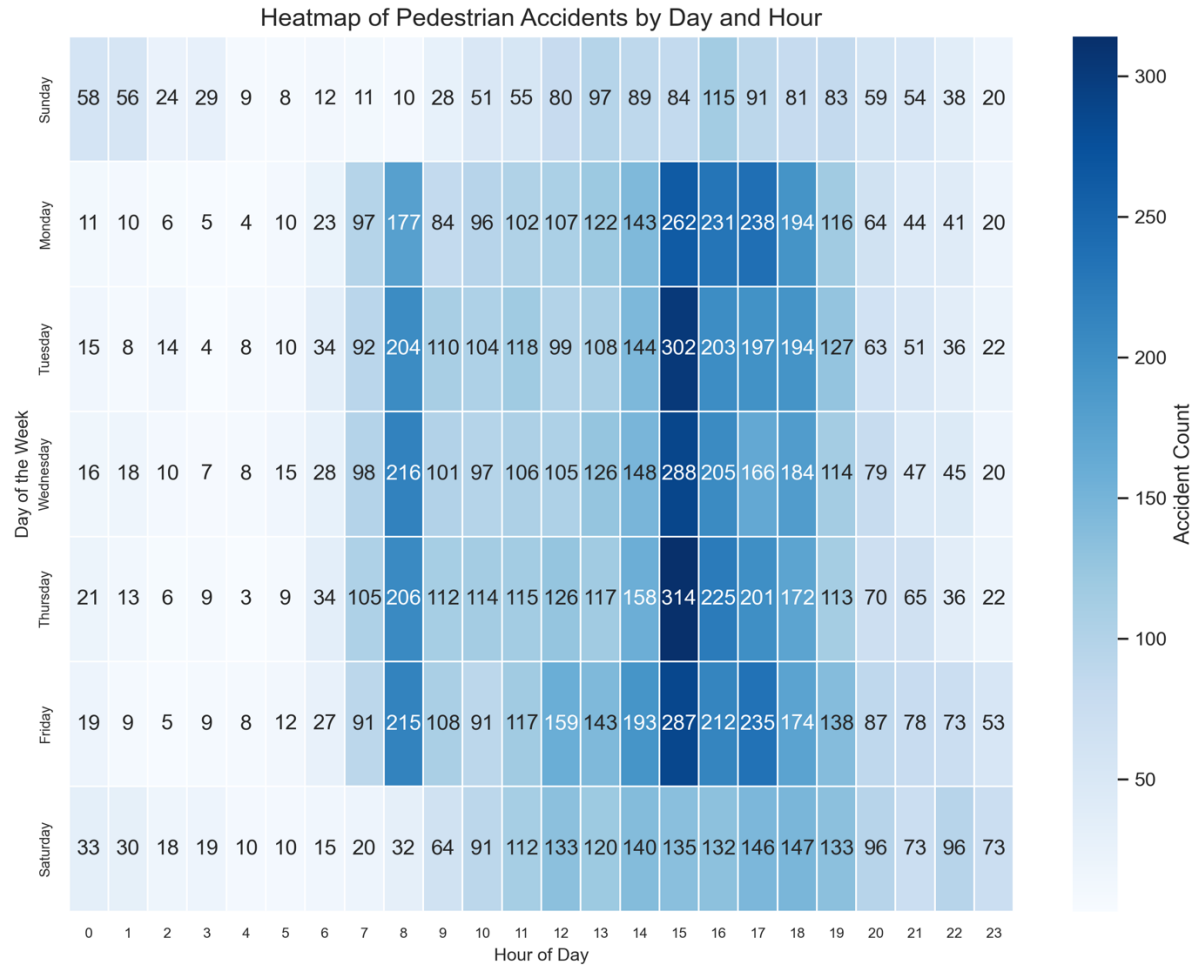

Figure 9: Heat map of Pedestrian Accidents by Hour of the Day and Day of the Week

From Figure 9 we can observe that Pedestrian accidents are most frequent between 3 PM and 4PM (15:00-16:00) and are highest on Fridays closely followed by Wednesdays and Thursdays.

- **<u>EXPLORING THE IMPACT OF SELECTED VARIABLES ON ACCIDENT SEVERITY USING APRIORI ALGORITHM</u>**

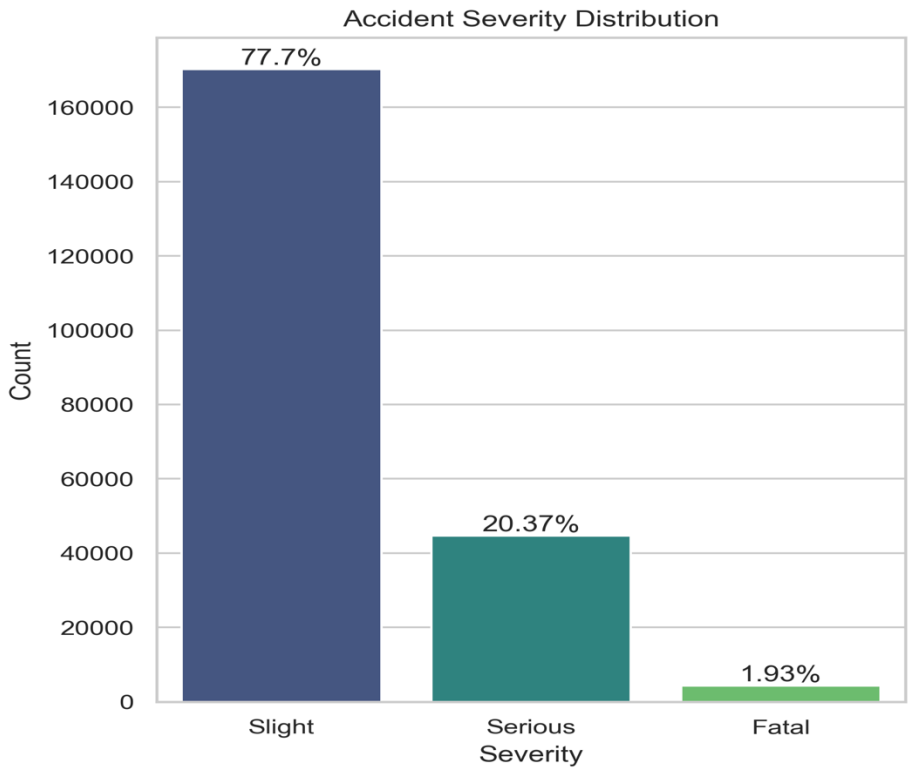

Figure 10: Accident Severity Distribution Plot

Antecedents	Consequents	Support	Confidence	Lift
casualty class- Driver or rider, light conditions- Daylight, vehicle type- Car, urban or rural area- Urban, weather conditions- Fine no high winds	accident severity- slight	0.117	0.703	1.715
casualty class- Driver or rider, speed limit- 30mph, vehicle type- Car, road surface conditions- Dry	accident severity- slight	0.132	0.748	1.553
Casualty class - Driver or rider, speed limit- 30mph, road surface conditions- Dry, light conditions- Daylight, vehicle type- Car	accident severity- slight	0.103	0.741	1.540

Table 2: Top 3 Rules for Accident Severity Using Apriori Algorithm

In all Top 3 cases, the casualty is a driver/rider, and the vehicle type is a car. Slight accidents are more likely in urban areas during daylight, especially when the weather is fine, and roads are dry.30 mph zones also appear frequently, suggesting that slight accidents tend to happen in lower-speed, everyday driving environments. The high confidence and lift values mean these conditions are strong indicators of slight accidents.

The Apriori algorithm is picking up only slight accidents because they make up most of the data (77.7%), while serious (20.37%) and especially fatal accidents (1.93%) are much rarer as seen in the plot in Figure 10.

- **IDENTIFYING ACCIDENTS IN HULL, CLUSTERING ON THIS DATA AND REVEALING THE DISTRIBUTION OF ACCIDENTS**

Accident data was narrowed down by region, concentrating on Kingston upon Hull, Humberside, and East Riding of Yorkshire. Regional data like latitude, LSOA and longitude were used to pinpoint accidents in these areas. The elbow method, shown in Figure 11, helped determine that 4 clusters was the best choice.

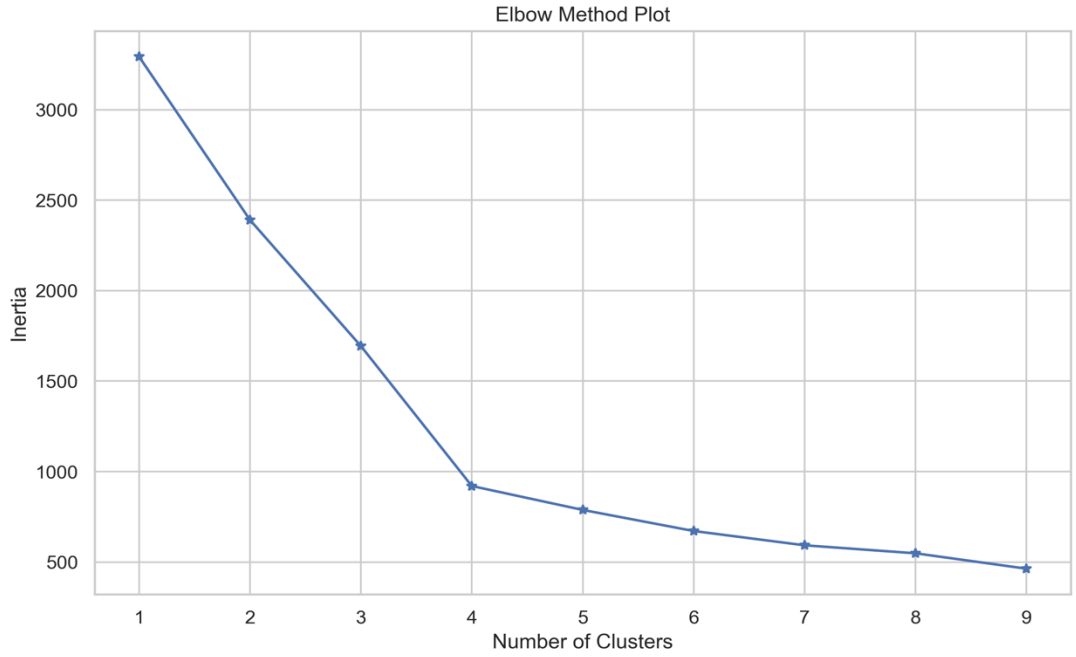

Figure 11: Determining Number of Clusters (Elbow Method)

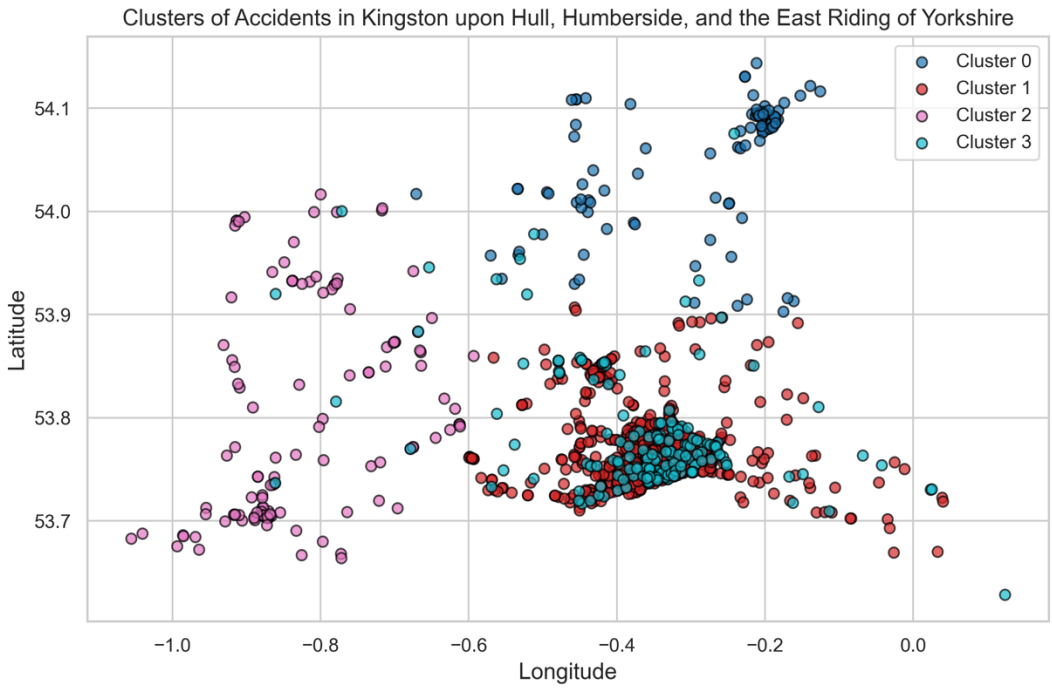

Figure 12: Clusters of Accidents in Kingston Upon Hull

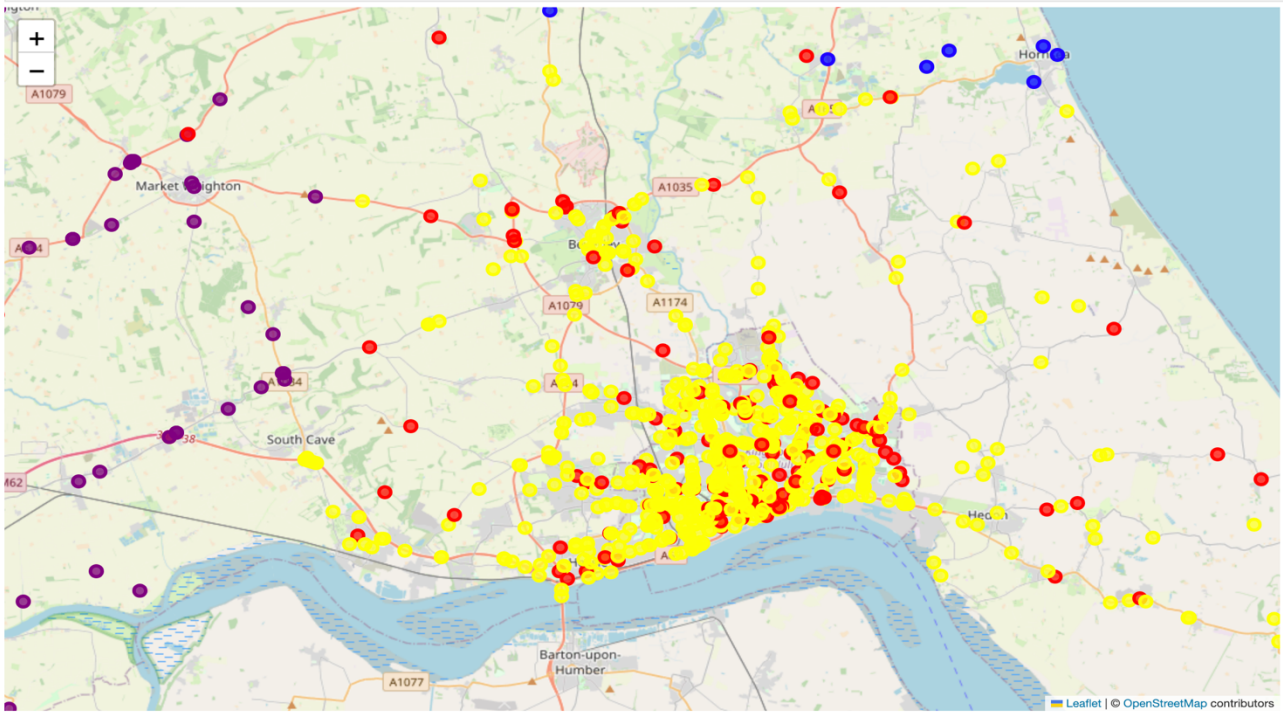

Figure 13: Map Showing Cluster Areas

Cluster	Type/ Area	Total Accidents	Slight Accidents	Serious Accidents	Fatal Accidents
0	Rural	106	88	18	0
1	City/Urban	691	691	0	0
2	Rural/Countryside	128	98	30	0
3	High risk areas	173	0	154	19

Table 3: Summary and Interpretation of Clusters

Cluster 1 has the highest number of accidents (691), but all are slight. **Cluster 0 (Rural)** and **Cluster 2 (Rural/Countryside)** show fewer total accidents but more serious ones, especially in Cluster 2, which has 30 serious accidents out of 128. **Cluster 3 (High-risk areas)** is the most concerning, with 173 total accidents, the majority being serious (154) and 19 fatal. Area types was deduced from the map in Figure 13.

- **PREDICTING WEEKLY ACCIDENT COUNTS FOR THE UPCOMING YEAR BASED ON HISTORICAL DATA FROM 2017 TO 2019 USING TIME SERIES MODELS**

The weekly accidents in Greater Manchester, Bedfordshire, and Durham were forecasted using three time series models.

ARIMA: Known for identifying patterns and forecasting in single-variable time series (Mondal, Shit and Goswami, 2014).

SARIMA: This model deals with the seasonal behaviour/trend in the data and used to forecast time-series data (Adineh, Narimani and Satapathy, 2020)

XGBOOST: Captures complex, non-linear patterns with strong predictive power (Fang et al., 2022).

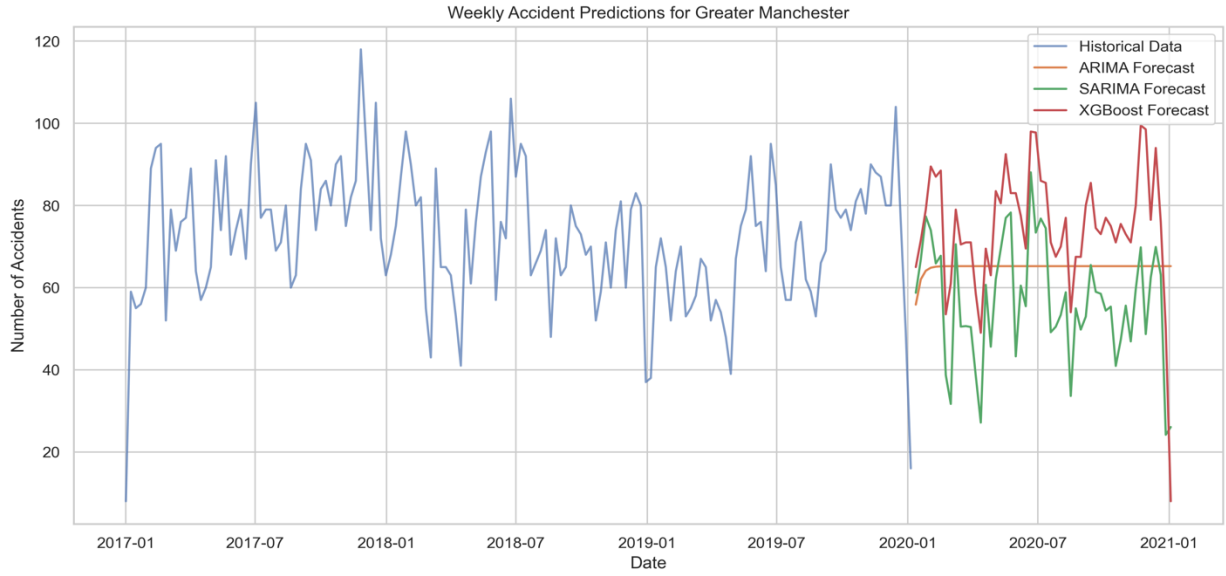

Figure 14: Weekly Accident Predictions for Greater Manchester

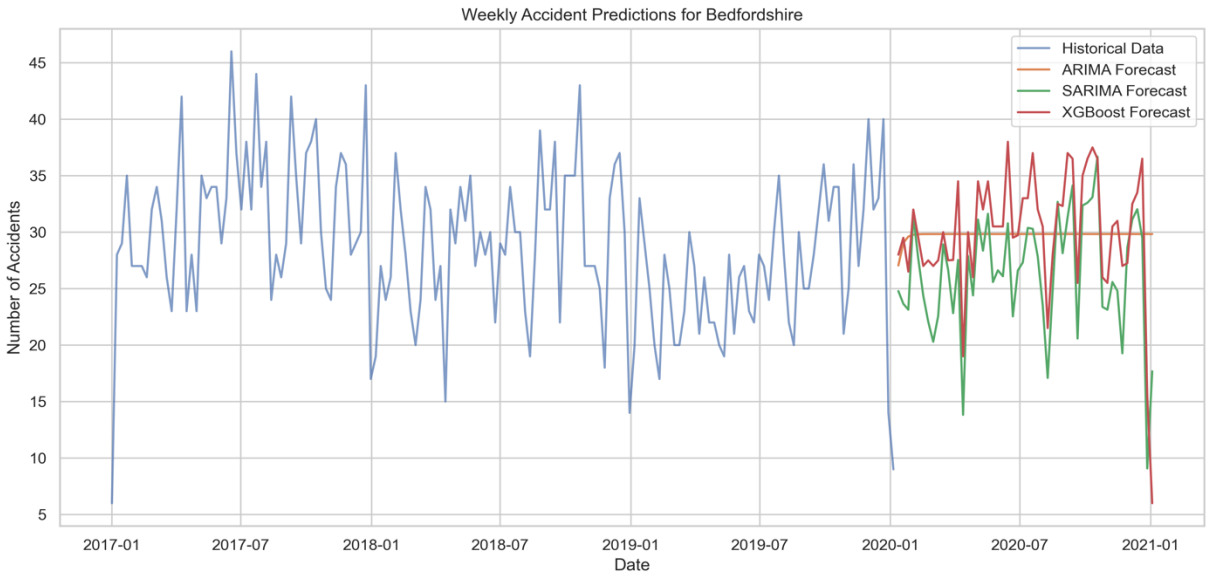

Figure 15: Weekly Accident Predictions for Bedfordshire

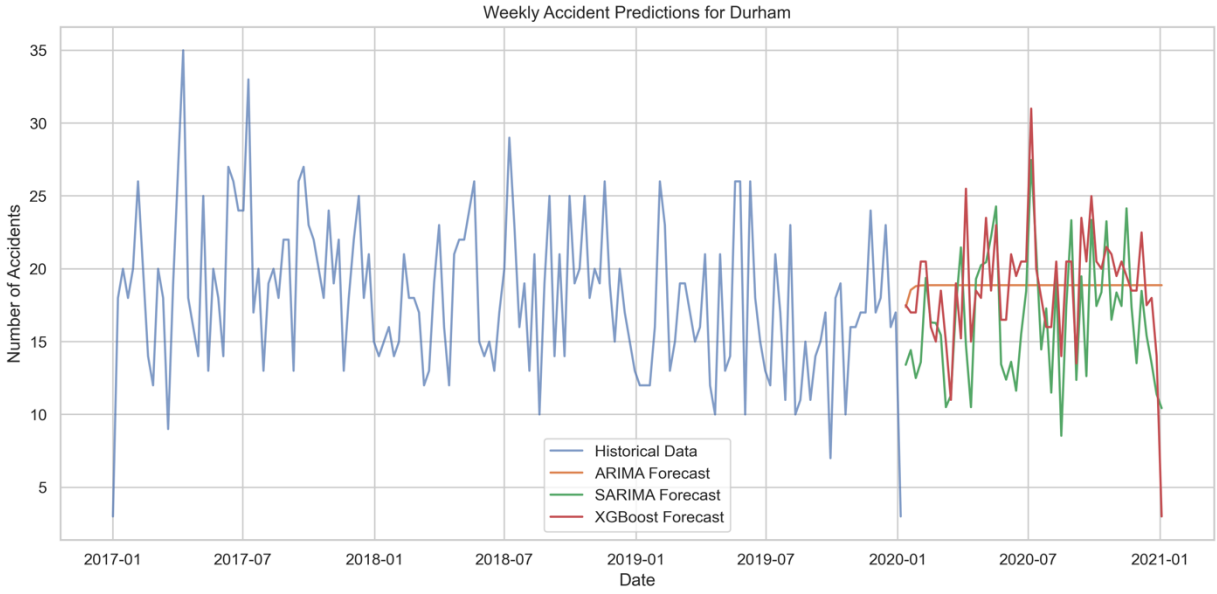

Figure 16: Weekly Accident Predictions for Durham

Augmented Dickey Fuller (ADF) Test showed that the time series data for all three cities was stationary (p-values < 0.05). The data was splitted into training and test sets, and model performance was evaluated using RMSE and MAE. Each model generated weekly forecasts for the next 52 weeks, visualized against historical trends (Figures 14–16).

Region	Date	ARIMA	SARIMA	XGBoost
Greater Manchester				
	2020-01-12	55.86	58.74	65.00
	2020-01-19	62.00	66.79	71.50
	2020-01-26	64.12	77.33	79.00
	2020-02-02	64.84	73.97	89.49
	2020-02-09	65.09	65.89	87.01
Average		**64.95**	**56.62**	**74.53**
Bedfordshire				
	2020-01-12	27.05	24.78	28.00
	2020-01-19	29.04	23.65	29.50
	2020-01-26	29.60	23.13	26.50
	2020-02-02	29.76	31.42	32.00
	2020-02-09	29.81	27.64	29.50
Average		**29.75**	**26.31**	**29.89**
Durham				
	2020-01-12	17.38	13.41	17.5 0
	2020-01-19	18.55	14.42	17.00
	2020-01-26	18.80	12.50	17.00
	2020-02-02	18.86	13.59	20.50
	2020-02-09	18.87	19.37	20.50
Average		**18.84**	**16.60**	**18.70**

Table 4: Time Series Models Weekly Accidents Predictions for each city

The predictions suggest accident rates will remain like past patterns. Manchester is expected to see weekly accidents ranging between 56 and 74. Bedfordshire shows lower rates at around 26–30, while Durham forecasts a steadier count of about 19 accidents per week.

EVALUATING THE BEST PERFORMING MODEL

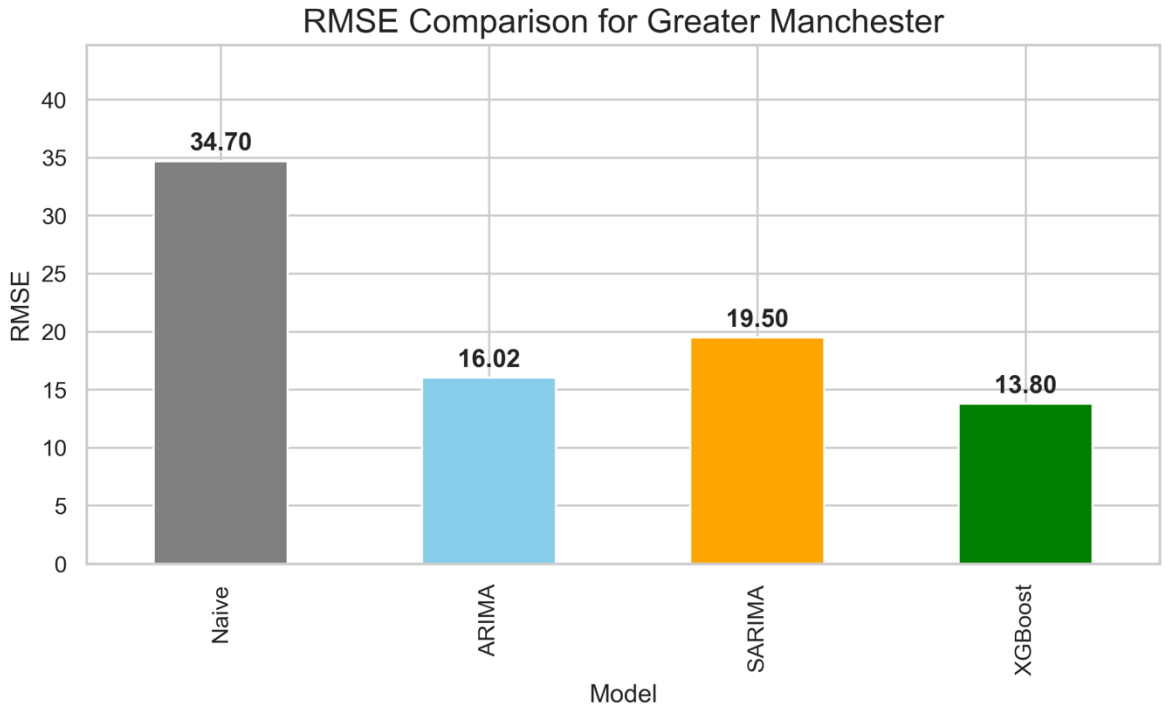

Figure 17: Models RMSE Comparison for Greater Manchester

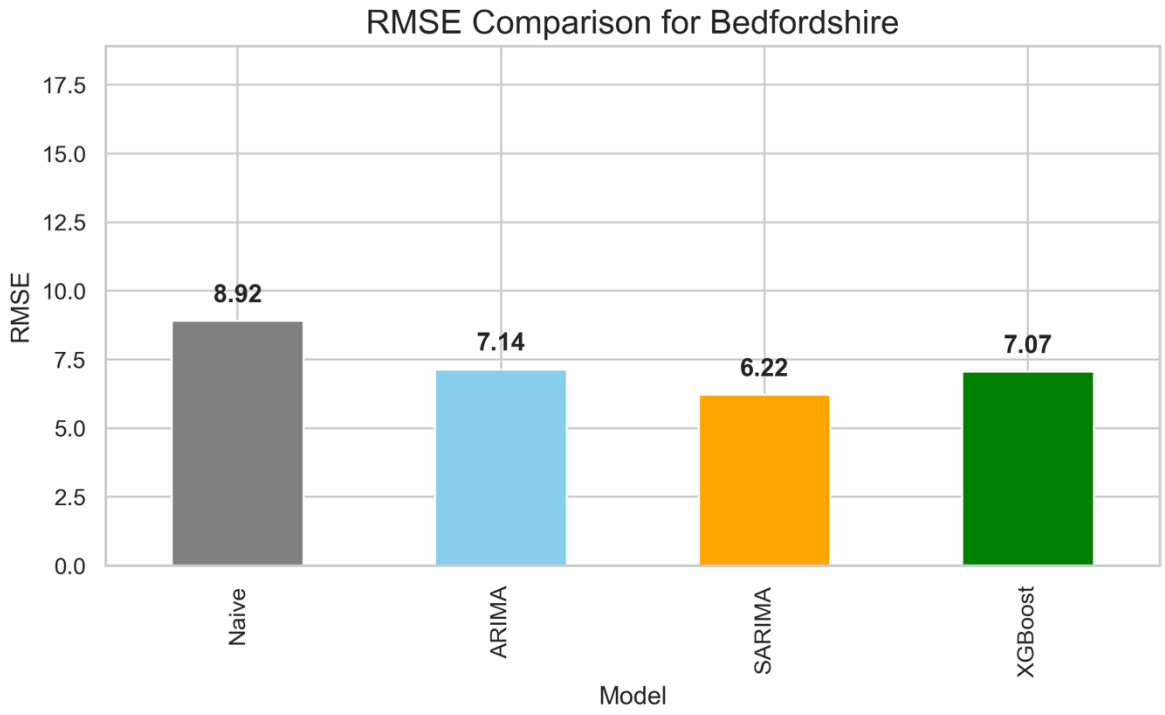

Figure 18: Models RMSE Comparison for Bedfordshire

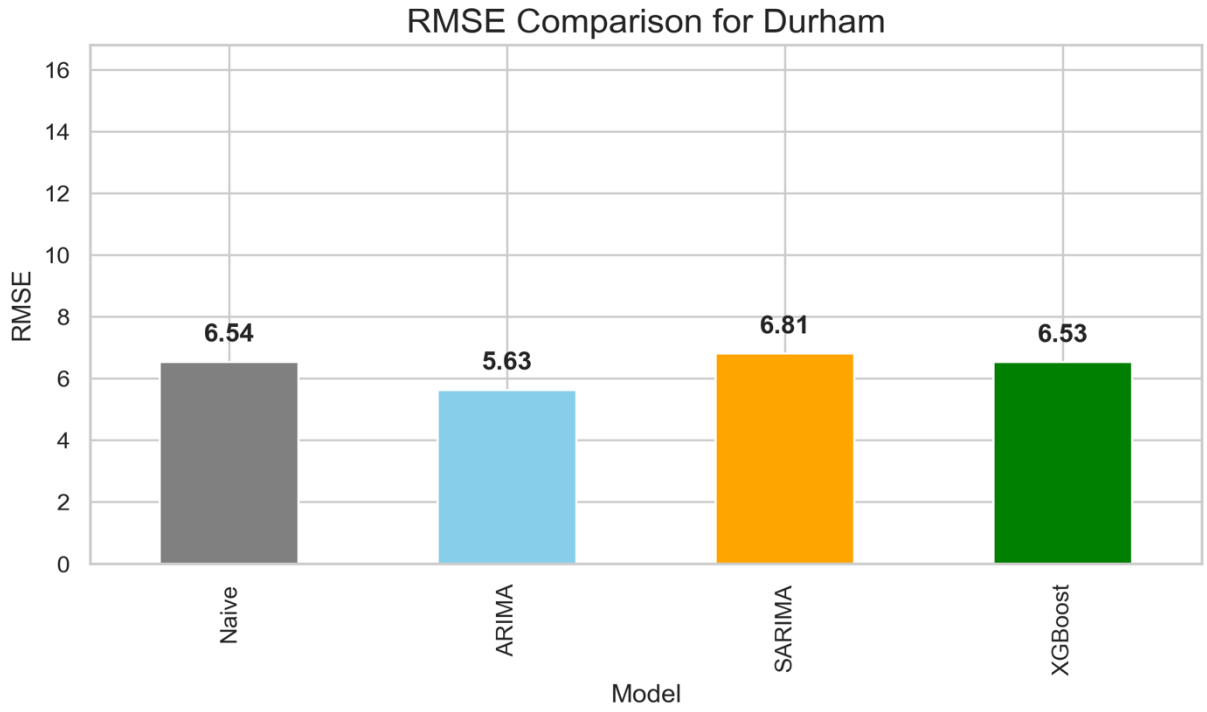

Figure 19: Models RMSE Comparison for Durham

Cities	Naïve (Baseline Model)	ARIMA	SARIMA	XGBOOST
Greater Manchester	34.7	16.02	19.50	13.80
Bedfordshire	8.92	7.14	6.26	7.07
Durham	6.54	5.63	6.81	6.53

Table 5: RMSE of all Models

From table 5 above Greater Manchester: XGBoost had the lowest RMSE (13.80), showing better prediction accuracy. Bedfordshire: SARIMA performed better than ARIMA, XGBoost and Naive with the lowest RMSE (6.26) and Durham: ARIMA outperformed other models with the lowest RMSE (5.63)

- **EMPLOYING A TIME SERIES MODEL TO FORECAST DAILY ACCIDENTS IN HIGH-INCIDENT LSOAs OF HULL USING DATA FROM JANUARY TO JUNE 2020**

An SQL query was used to extract LSOA codes for the selected areas and accident data from January to June 2020, for the historical dataset. The results were loaded into a Data Frame and converted into a datetime series. Figure 20 below highlights the highest number of accidents by the three LSOAs during the first quarter of 2020.

```
Top 3 LSOAs of Hull with Highest Accident Count (January–March 2020)
================================================
LSOA Name                      LSOA Code        Accident Count
------------------------------------------------
Kingston upon Hull 016D        E01012817                  10
Kingston upon Hull 020B        E01012848                   7
Kingston upon Hull 030B        E01012889                   7
------------------------------------------------
Top 3 LSOAs Total Accidents are: 0
================================================
```

Figure 20: Three LSOAs with Highest Accident Count in Hull (January-March 2020)

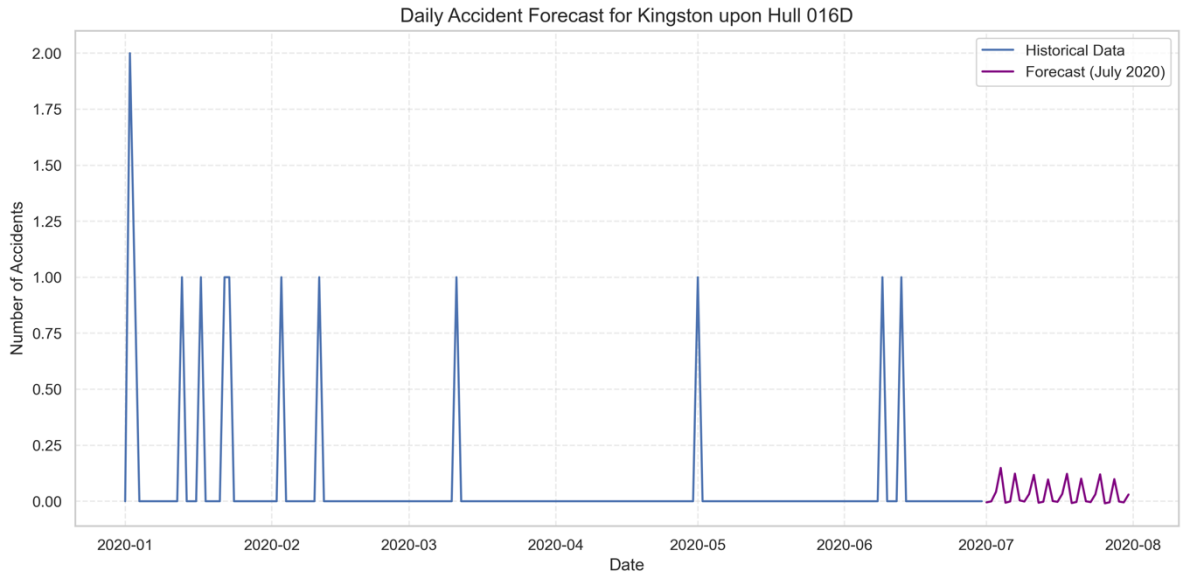

Figure 21: Daily Accident Prediction for Kingston Upon Hull 016D

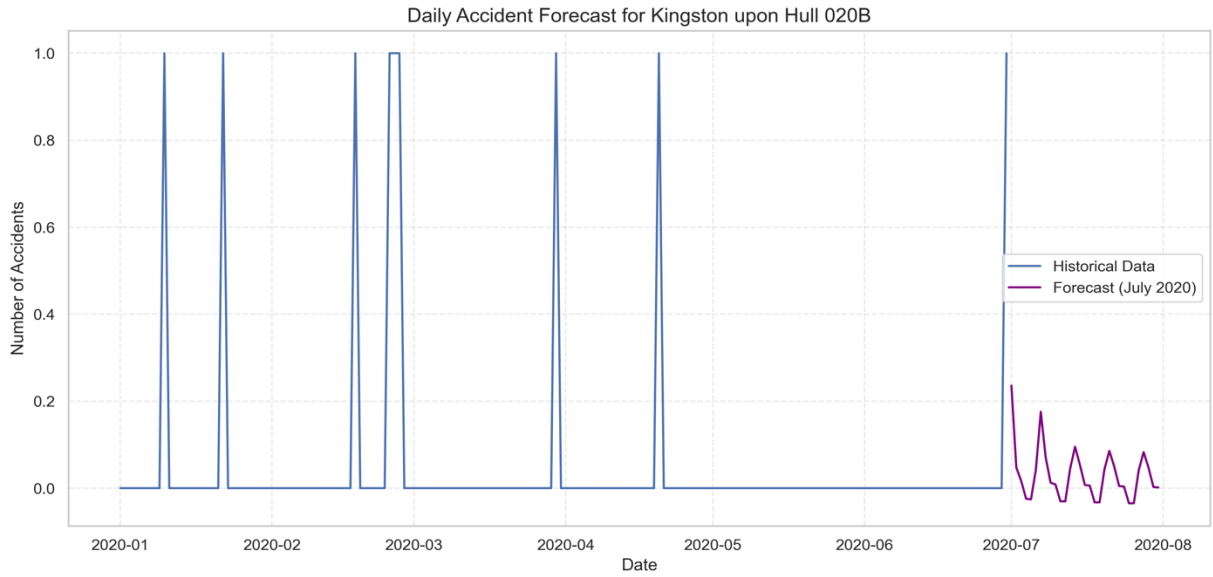

Figure 22: Daily Accident Prediction for Kingston Upon Hull 020B

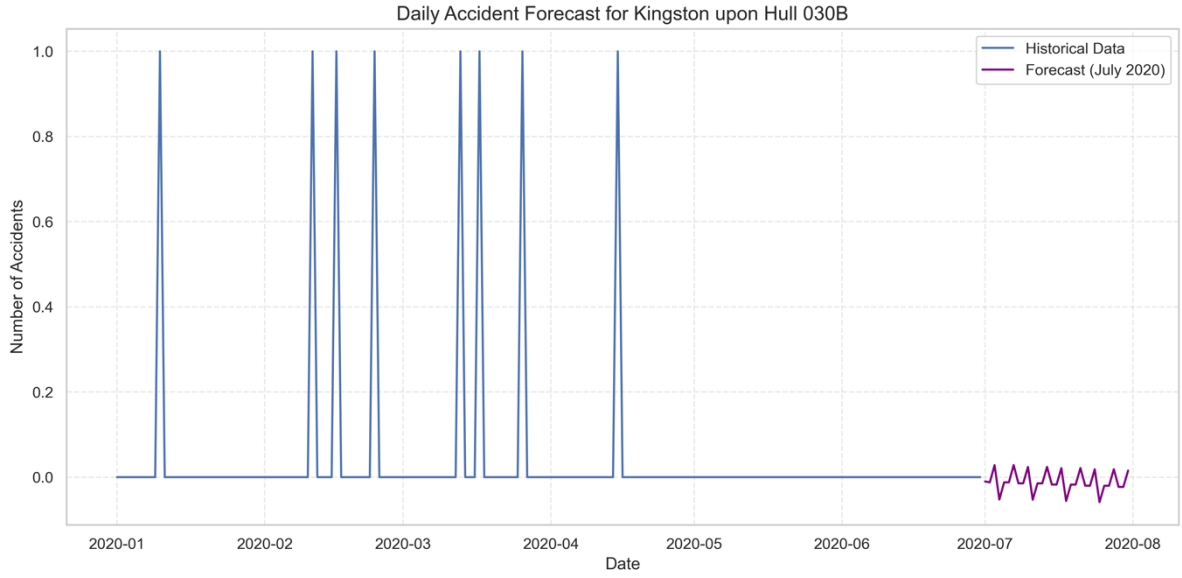

Figure 23: Daily Accident Prediction for Kingston Upon Hull 030B

LSOA	CATEGORY	ANALYSIS
Kingston Upon Hull 016D	Past Data	Accidents didn't happen every day, and the busiest days varied without a clear pattern.
	Predictions	The predicted daily accident counts for July 2020 show a steady trend with consistently low numbers.
Kingston Upon Hull 020B	Past Data	This LSOA also experienced few accidents, with no

		consistent pattern in peak days.
	Predictions	The July 2020 forecast shows low daily accident counts, closely resembling historical trends with slight fluctuations.
Kingston Upon Hull 030B	Past Data	Accident patterns were not consistent, with most days having no incidents and occasional sudden spikes.
	Predictions	Daily accident counts for July 2020 are expected to remain low.

Table 6: Top 3 LSOAs Daily Accident Prediction for July 2020

Date	Accidents
2020-07-01	**0.0**
2020-07-02	**0.0**
2020-07-03	**0.0**
2020-07-04	**0.0**
2020-07-05	**0.0**
2020-07-06	**0.0**

The three Hull LSOAs with the highest accident counts in early 2020 are forecasted to maintain low daily accidents(0), consistent with past trends. Accident occurrences remain irregular, with most days showing no incidents and occasional, unpredictable spikes. This highlights the difficulty of modelling sparse, uneven data. Still, SARIMA managed to capture the general trend effectively, showing that even with limitations, time series models can offer valuable insights for planning and safety interventions.

- **CONSTRUCTING AND VISUALISING A SOCIAL NETWORK USING THE DATA, SHOWING THE BASIC NETWORK CHARACTERISTICS (NETWORK DENSITY, AVERAGE DEGREE, NUMBERS OF NODES AND EDGES**

Networkx provides a wide range of centrality measures to help detect the most influential member in the network (Richard, Faadhilah and Qomariyah, 2022).

```
--- Social Network Summary ---
Nodes: 4039
Edges: 88234
Density: 0.0108
Average Degree: 43.69
```

Figure 24: Characteristics of the Basic Network

The network has 88,234 connections and 4,039 nodes**,** with a density of 0.0108 signifying it's relatively sparse. On average, each node is connected to 44 others, showing moderate connectivity as seen in figure 24.

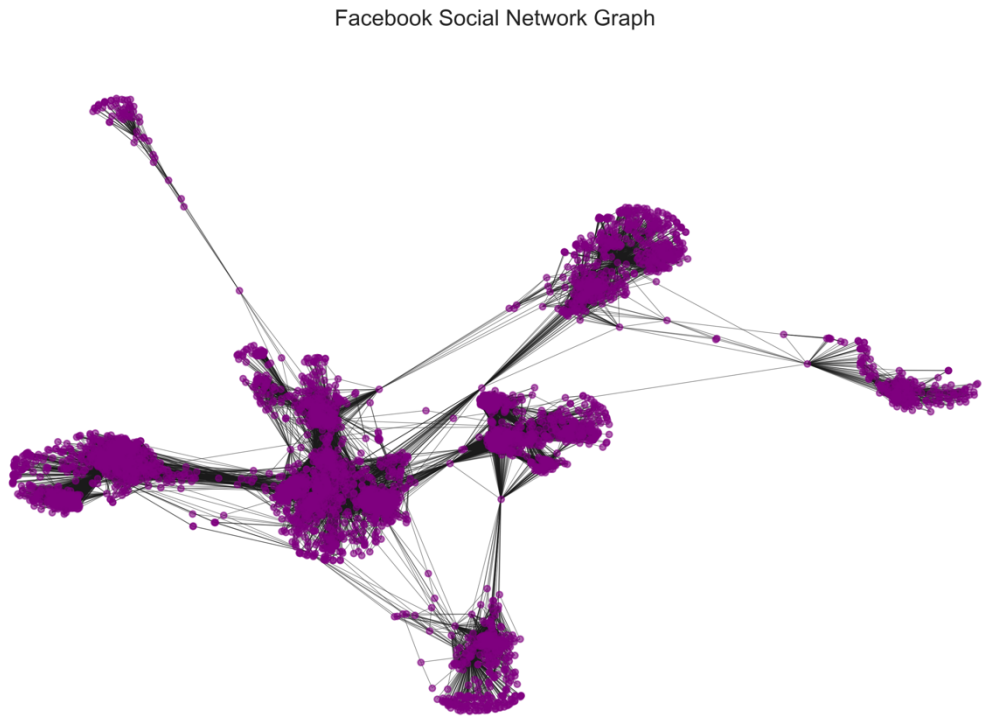

Figure 25: Facebook Social Network Graph

Figure 25 above is a social network graph showing groups of connected individuals. Each blue dot is a person, and lines show relationships. The clusters represent tight-knit groups, while a few key connections link different groups together, acting like bridges across the network.

- **CALCULATING THE EDGE CENTRALITY OF THE NETWORK**

Metric	**Value**
Edges Analysed	88,234
Minimum Centrality	0.0000
Maximum Centrality	0.1715
Average Centrality	0.0000

Table 7: Edge Betweenness Centrality

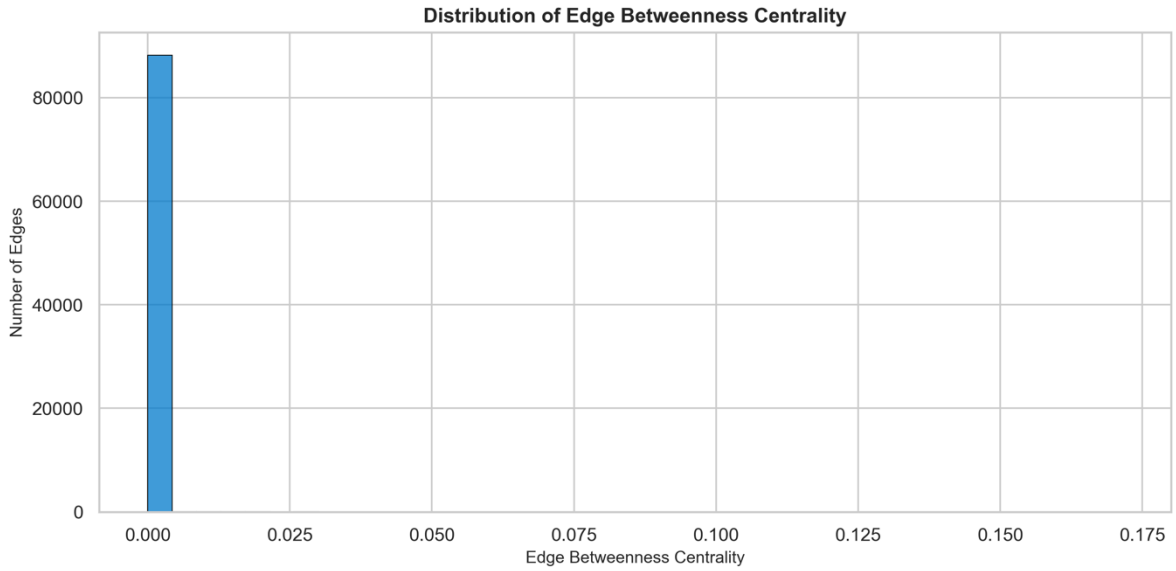

Figure 26: Distribution of Edge Betweenness Centrality

Most edges in the network have very low betweenness centrality as shown in Figure 26, meaning they're not often used in shortest paths between nodes. This suggests that only a few edges act as key connectors or bridges between different parts of the network.

- **DETECTING THE CLUSTERS/COMMUNITY WITHIN THE SOCIAL NETWORK USING TWO COMMUNITY DETECTION ALGORITHMS AND COMPARING THE DIFFERENCE OF RESULTS**

Louvain Modularity Algorithm captures the basic idea behind communities which are groups of nodes tightly connected to one another than the rest of the network (Yao *et al.*, 2023) and Label Propagation is easy to implement, clear and efficient (Raghavan, Albert and Kumara, 2007).

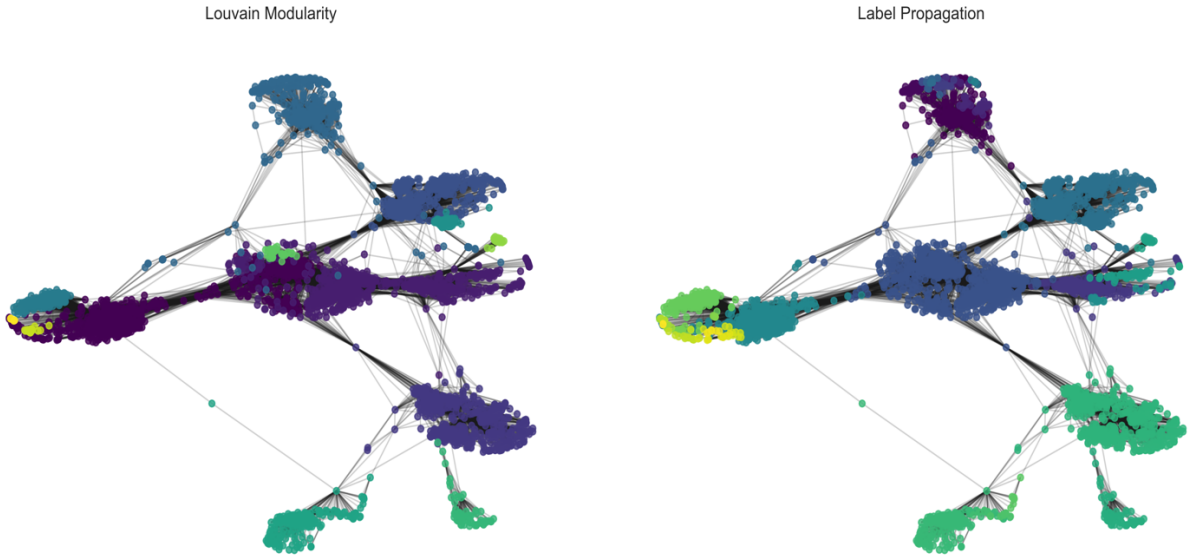

Figure 27: Plot of Louvain Modularity vs Label Propagation

Community Detection Algorithm	Louvain Modularity	Label Propagation
Modularity	0.7774	0.7368
Number of Communities	13	44
Average Community Size	310.69	91.80

Table 8: Results of Louvain Modularity vs Label Propagation

The Louvain algorithm found 13 well-defined communities with a high modularity (0.7774), meaning the groups are strongly connected internally. The average community (311) shows more structured and larger clusters.

Label Propagation identified 44 smaller communities with a lower modularity (0.7368) and an average size of 92**,** leading to more scattered and less cohesive groupings. Louvain provides a clearer and more meaningful community structure.

3.0 RECOMMENDATIONS FOR THE GOVERNMENT

- Focus awareness efforts on Fridays between 3–6 PM, when most accidents occur, especially for bikers and pedestrians.
- Upgrade signage, lighting, and speed controls in Cluster 3(High risk) areas of Hull and Kingston upon Hull 016D lsoa, where serious and fatal crashes are most common.
- Enforce stricter training and licensing for riders, especially those using bikes under 125cc, who are high-risk during rush hours.
- Apply models like XGBoost and SARIMA to forecast accident trends and guide timely police patrols.
- Review 30 mph zones and improve road design to reduce frequent low-severity crashes in daylight and good weather.

REFERENCES

Adineh, A.H., Narimani, Z. and Satapathy, S.C., 2020. Importance of data preprocessing in time series prediction using SARIMA: A case study. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, *24*(4), pp.331-342.

Ahmed, S., Mohammed, M., Abdulqadir, S., Abd Elkader, R., El-Shall, N., Chandran, D., Rehman, M. E. U. & Dhama, K. (2023) Road traffic accidental injuries and deaths: A neglected global health issue. *Health Science Reports,* 6 e1240. https://doi.org/10.1002/hsr2.1240

Clarke, D.D., Ward, P., Bartle, C. and Truman, W., 2007. The role of motorcyclist and other driver behaviour in two types of serious accident in the UK. *Accident Analysis & Prevention*, *39*(5), pp.974-981.

Department for Transport. (2024) 'Reported road casualties in Great Britain, provisional estimates: year ending June 2024', *GOV.UK*. Available at: https://www.gov.uk/government/statistics/reported-road-casualties-in-great-britain-provisional-estimates-year-ending-june-2024/reported-road-casualties-in-great-britain-provisional-estimates-year-ending-june-2024 (Accessed: 5 April 2025).

Fang, Z.G., Yang, S.Q., Lv, C.X., An, S.Y. and Wu, W., 2022. Application of a data-driven XGBoost model for the prediction of COVID-19 in the USA: a time-series study. *BMJ open*, *12*(7), p.e056685.

Mondal, P., Shit, L. and Goswami, S., 2014. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, *4*(2), p.13.

Pawłowski, W., Lasota, D., Goniewicz, M., Rzońca, P., Goniewicz, K. and Krajewski, P., 2019. The effect of ethyl alcohol upon pedestrian trauma sustained in traffic crashes. *International journal of environmental research and public health*, *16*(8), p.1471.

Raghavan, U.N., Albert, R. and Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, *76*(3), p.036106.

Richard, J., Faadhilah, R. and Qomariyah, N.N., 2022, August. Jaebot: Discord bot for network analysis with networkX. In *2022 International Conference on ICT for Smart Society (ICISS)* (pp. 1-6). IEEE.

Yao, B., Zhu, J., Ma, P., Gao, K. and Ren, X., 2023. A constrained louvain algorithm with a novel modularity. *Applied Sciences*, *13*(6), p.4045.