MSc Artificial Intelligence and Data Science

Module 771762- Big Data & Data Mining

UK Road Traffic Accident Project Report

By

Student ID – 202403820 | Samuel Datubo Jaja

1.0 Introduction

Road traffic accidents are a significant concern for policymakers and the public alike, as they have great implications for safety, public health, and urban planning. Using the dataset provided by the UK Department of Transport, this study seeks to explore patterns in road accidents, identify critical factors contributing to accidents, and predict future accident trends. Specifically, we focus on answering key questions about when, where, and under what conditions accidents occur, and we forecast weekly accident counts for 2020 using historical data from 2017 to 2019. In addition, will use the Apriori algorithm to explore the impact of selected variables on accident severity.

2.0 Analysis

2.1 Data Cleaning

Data cleaning is always done on a copy of the original data (Skiena, 2017). The fuel of Data Science, Artificial Intelligence and Machine Learning is data and to achieve the most accurate insight from any data, data must be of high quality (Datascientest, 2023). A data cleaning class was implemented to handle invalid values, clean white spaces and drop duplicates.

2. 2 Data Analysis

2.2.1 Temporal Accident Analysis

After thorough cleaning and analysis, the heatmap in figure 1a shows that accident frequency is highest on Friday at 5 PM (746 accidents) in the United Kingdom and generally peaks during late afternoons and early evenings (4 PM to 7 PM) across all days, with lower occurrences during early morning hours.

The analysis of significant hours of the day, and days of the week, on which accidents occur revealed that accident hours occur during commuting periods, specifically between 7 AM to 9 AM and 4 PM to 7 PM, with the highest frequency recorded at 6 PM (4,275 accidents). In contrast, early morning hours between 2 AM to 5 AM see the lowest accident rates, with 4 AM having the minimum frequency of 222 accidents. Regarding days of the week, Saturday records the highest number of accidents at 8,065, followed by Friday with 7,614 accidents, likely due to increased travel and recreational activities. Monday has the lowest accident frequency at 5,307, reflecting a cautious start to the week. Accident frequencies gradually increase from Tuesday through Friday, with Wednesday recording 7,208 accidents and Thursday having 7,464 accidents as seen from figure 1b and 1c below.

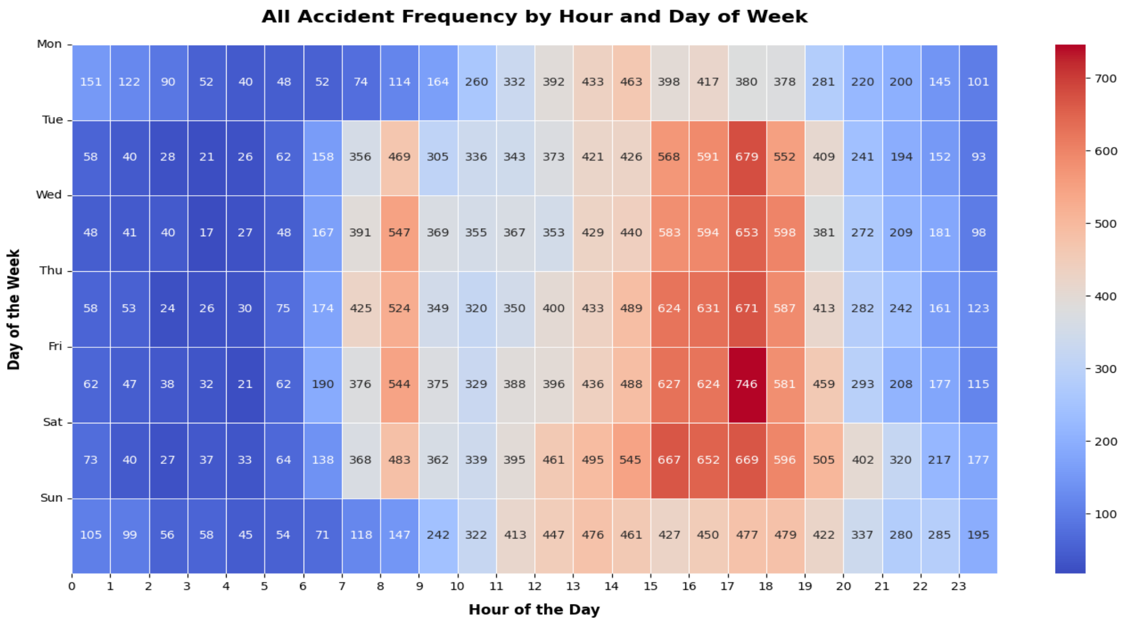


Figure 1a: Accident Frequency by Hour and Day of Week

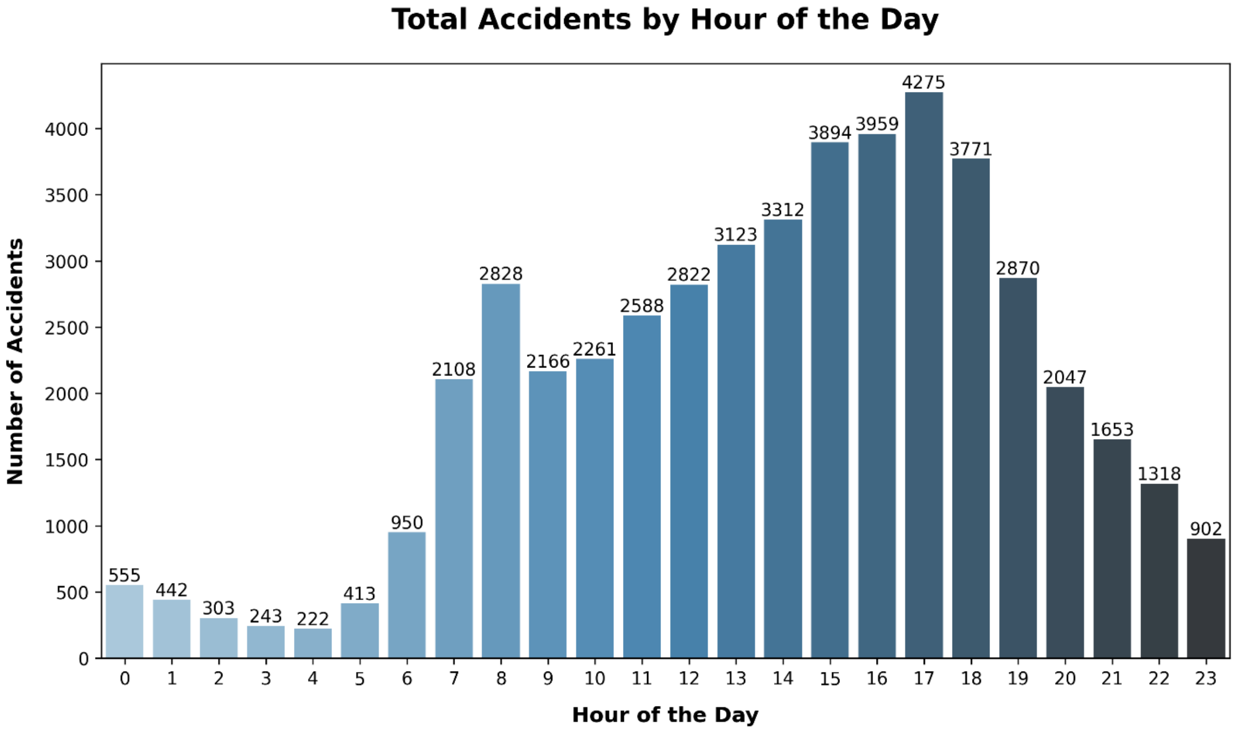


Figure 1b: Total Accidents by Hour of the Week

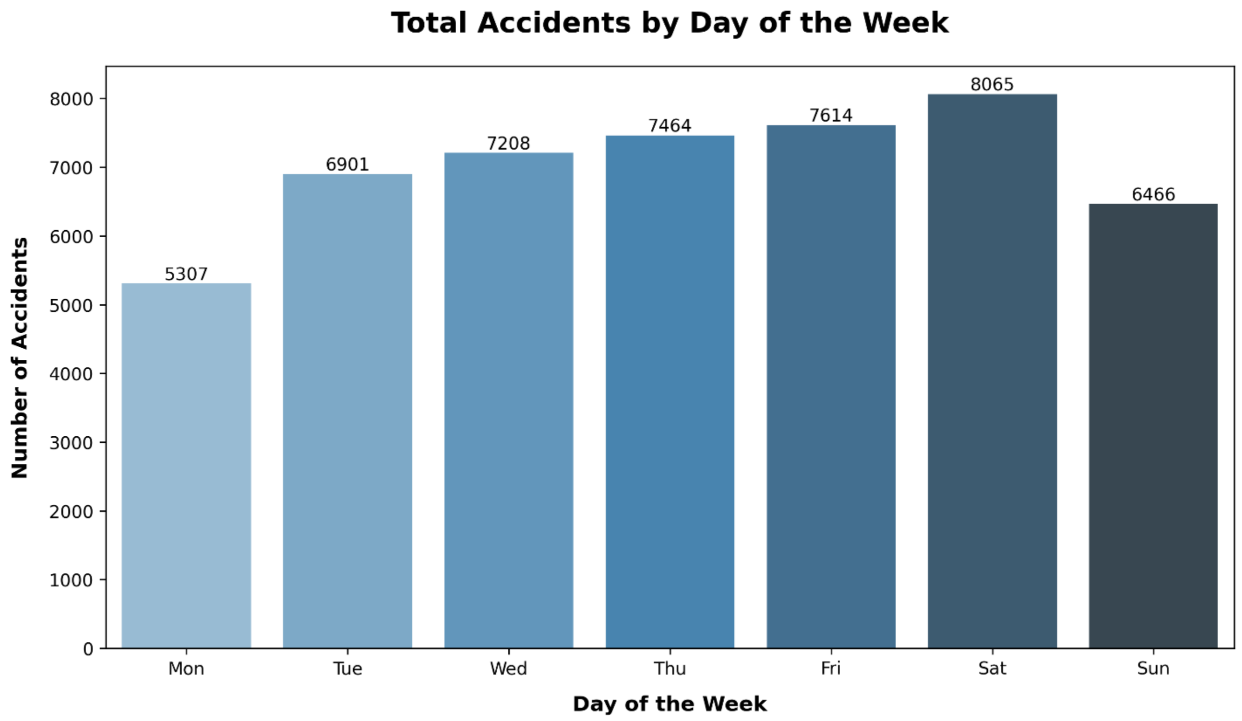


Figure 1c: Total Accidents by Day of the Week

From figure 2a, motorbike accident frequency peaks on Saturdays at 5 PM with 124 accidents, with high incidents also observed on Fridays and during afternoon hours (2 PM to 6 PM), particularly on weekends. Weekdays show relatively lower accident rates, especially in the early morning hours (midnight to 8 AM), likely reflecting cautious riding during utilitarian commutes. The weekend spikes suggest increased leisure riding and riskier behavior, aligning with favorable weather conditions.

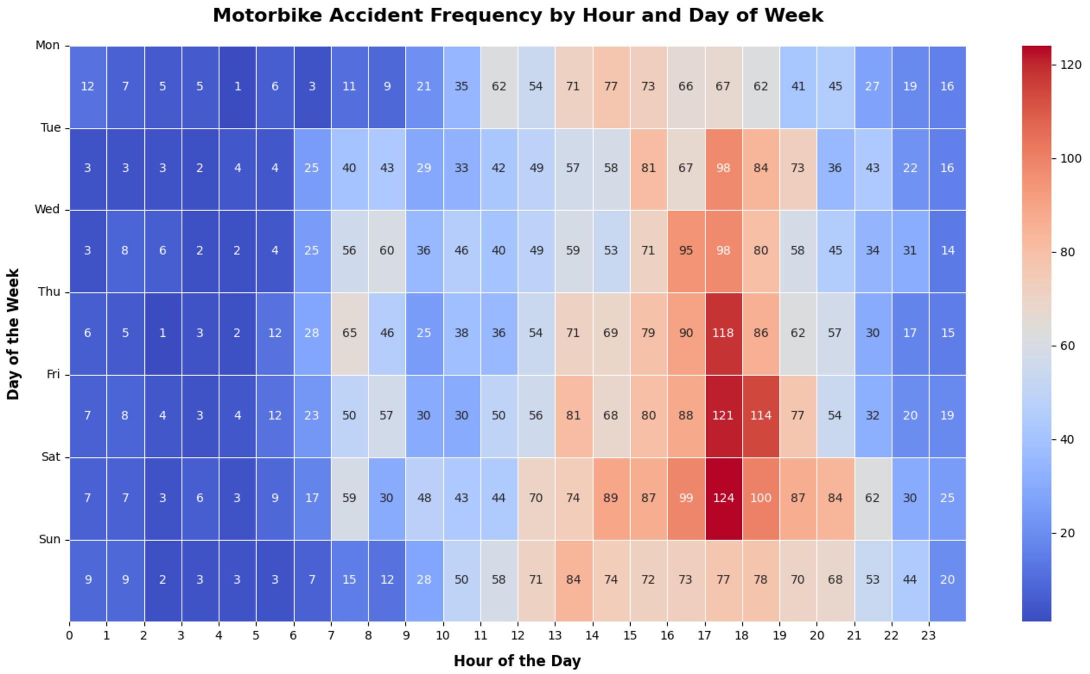


Figure 2a: Motorbike Accident Frequency by Hour and Day of Week

From figure 2b, motorbike accidents are most prevalent on Saturdays, with a total of 1,207 incidents, followed by Fridays with 1,088 accidents, while Mondays recorded the least accidents at 795 incidents.

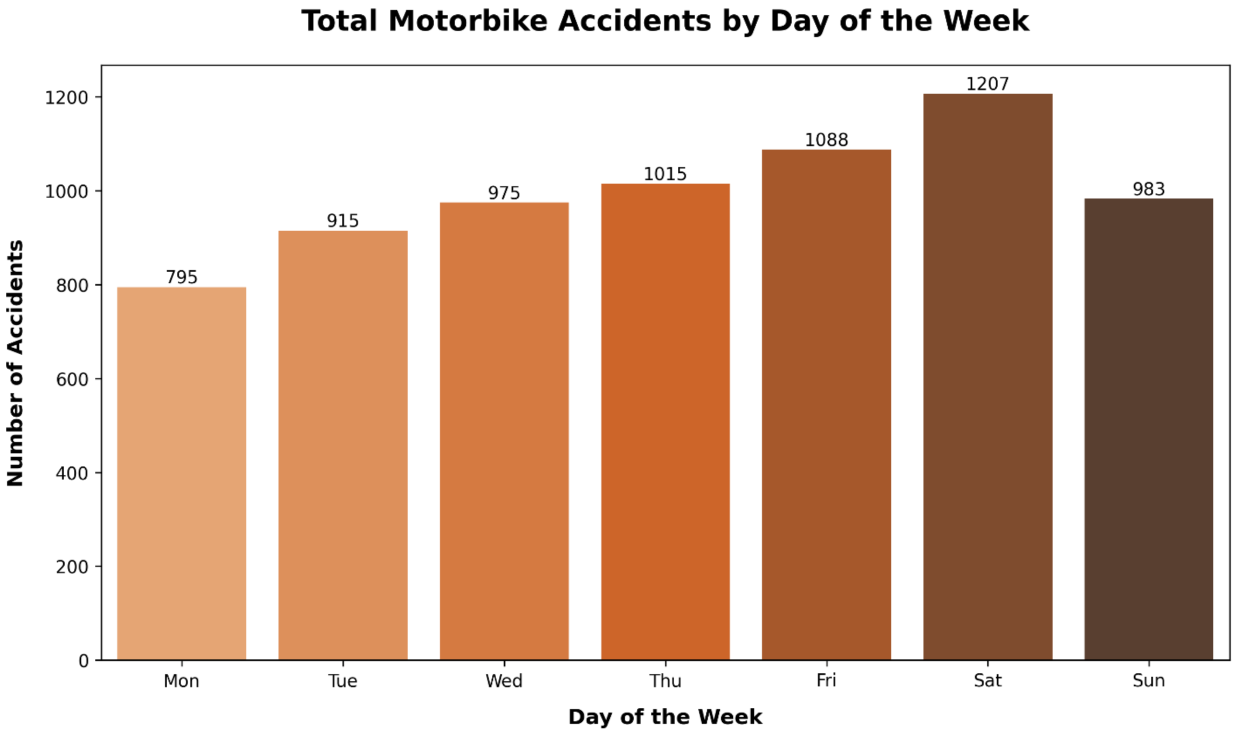


Figure 2b: Total Motorbike Accident Frequency by Day of Week

From figure 2c, the peak hour for motorbike accidents is 5 PM, with 703 accidents, indicating that the evening rush hour is a high-risk period. Other significant times include 6 PM (604 accidents) and 4 PM (578 accidents), suggesting increased risks during late afternoon hours. In contrast, the early morning hours (12 AM to 5 AM) saw the lowest occurrences, with accident counts ranging from 19 to 50.

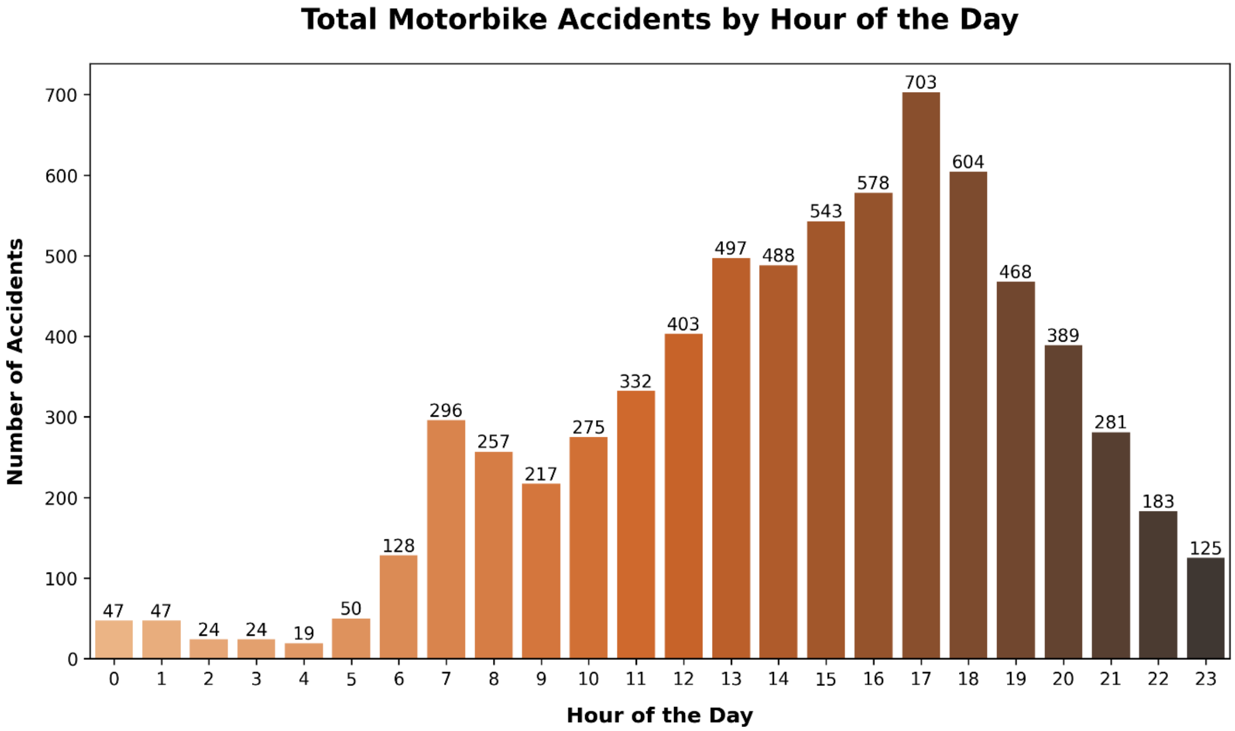
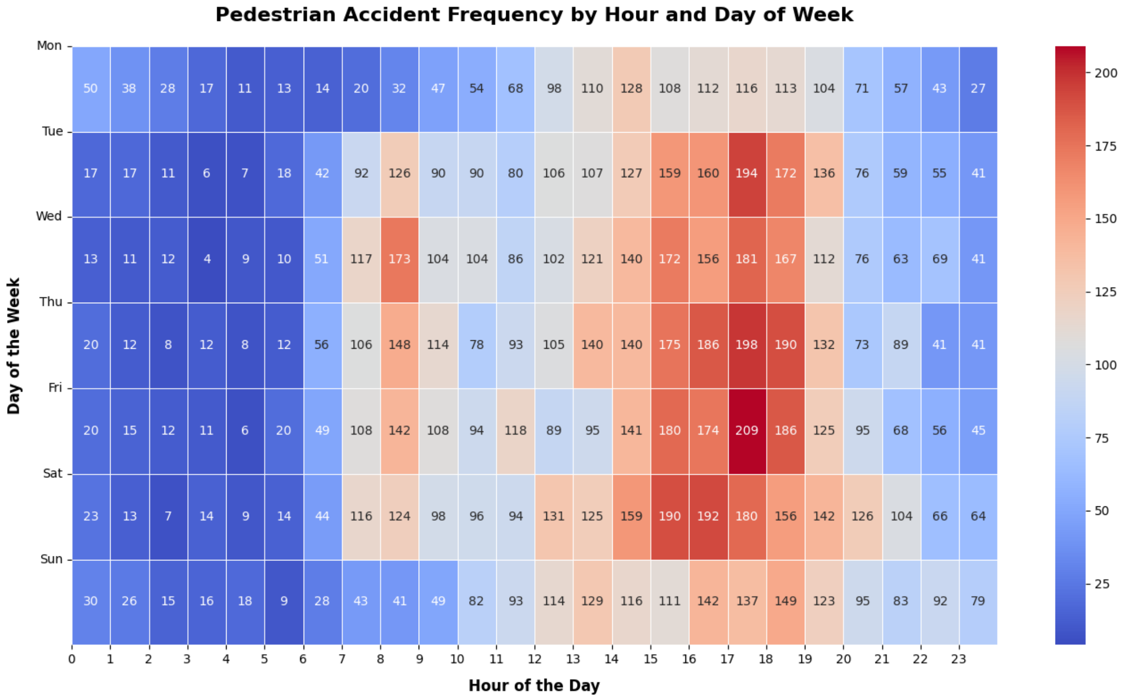


Figure 2c: Total Motorbike Accident Frequency by Hour of Day

Based on figure 3a, pedestrian accident frequency is highest on Fridays at 5 PM (209 accidents), with notable peaks during weekday evenings (4 PM to 7 PM) and moderate activity on Saturdays, particularly in the late afternoon. Early mornings (12 AM to 7 AM) and Sundays generally see fewer accidents. The weekday evening peaks align with rush hours, reflecting high pedestrian and vehicle interaction.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Mon	50	38	28	17	11	13	14	20	32	47	54	68	98	110	128	108	112	116	113	104	71	57	43	27
Tue	17	17	11	6	7	18	42	92	126	90	90	80	106	107	127	159	160	194	172	136	76	59	55	41
Wed	13	11	12	4	9	10	51	117	173	104	104	86	102	121	140	172	156	181	167	112	76	63	69	41
Thu	20	12	8	12	8	12	56	106	148	114	78	93	105	140	140	175	186	198	190	132	73	89	41	41
Fri	20	15	12	11	6	20	49	108	142	108	94	118	89	95	141	180	174	209	186	125	95	68	56	45
Sat	23	13	7	14	9	14	44	116	124	98	96	94	131	125	159	190	192	180	156	142	126	104	66	64
Sun	30	26	15	16	18	9	28	43	41	49	82	93	114	129	116	111	142	137	149	123	95	83	92	79

Figure 3a: Heatmap for Pedestrian Accident Frequency by Hour and Day of Week

From figure 3b and 3c, pedestrian accidents are most frequent on Saturdays (2,287 accidents), followed by Wednesdays (2,177) and Fridays (2,166), with the fewest occurring on Mondays (1,479). The most dangerous hour is 5PM, with 1,215 accidents, followed by 6 PM (1,133 accidents) and 4 PM (1,122 accidents), highlighting late afternoon to early evening (4 PM to 7 PM) as critical periods for intervention.

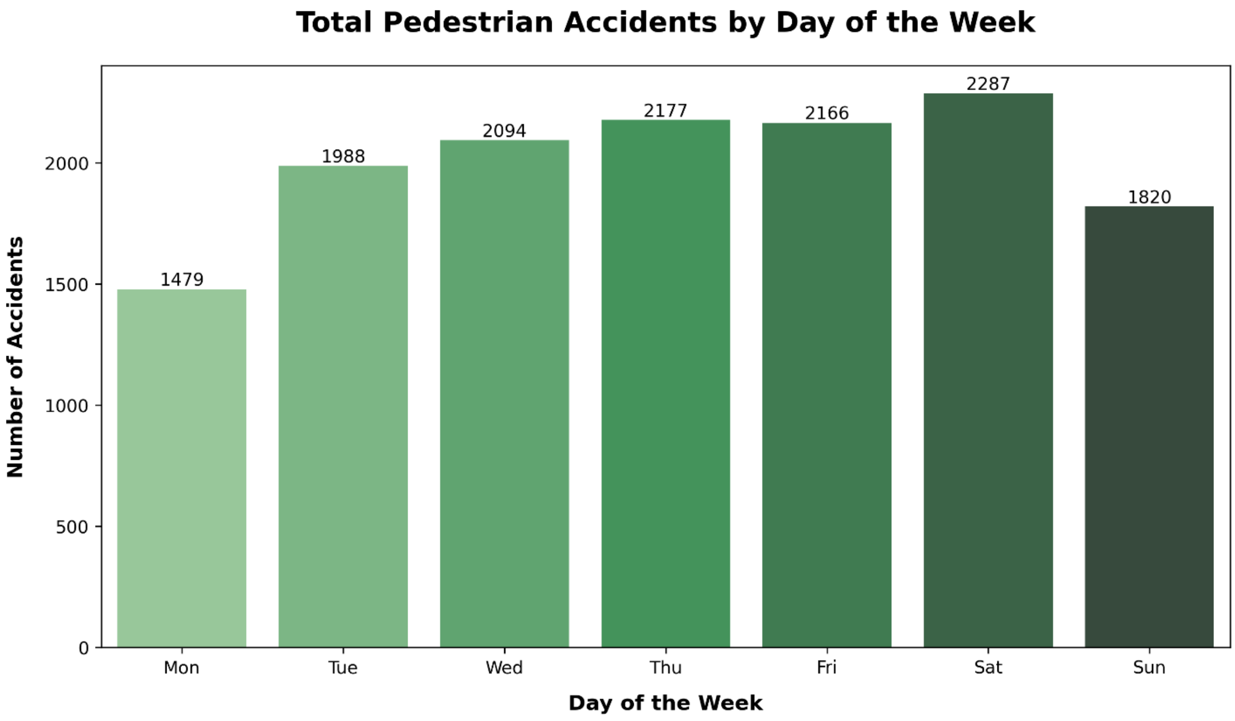


Figure 3b: Total Pedestrian Accident by Day of Week

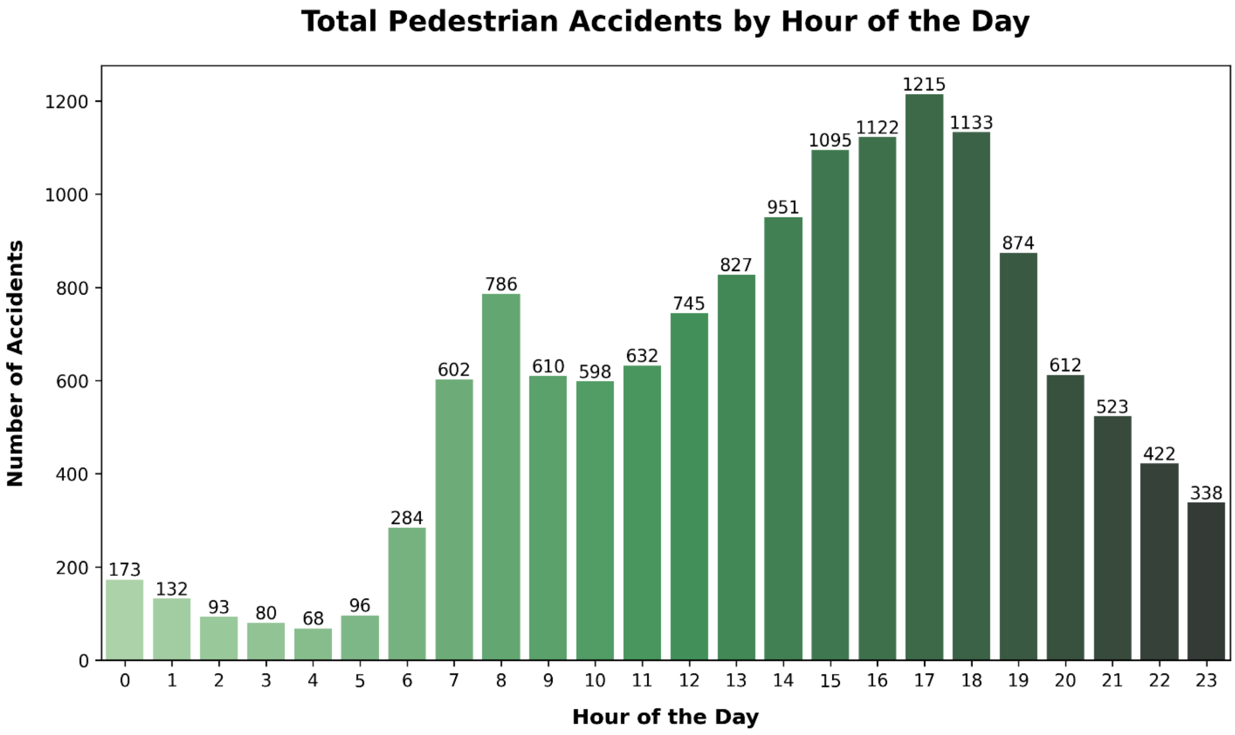
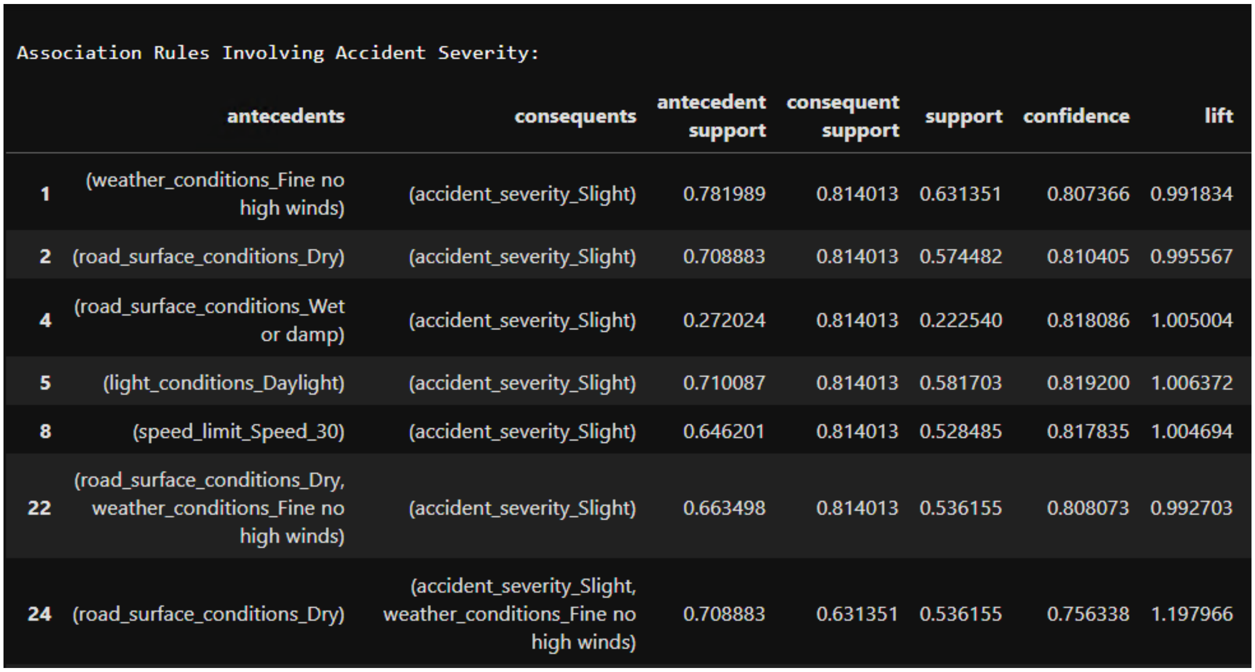


Figure 3c: Total Pedestrian Accidents by Hour of the Day

2.2.2 Analysis of Variables Impacting Accident Severity Using Apriori Algorithm

According to Perez (2023) association rule mining using the Apriori Algorithm produces significant patterns and insights that help identify the causes of road accidents. The analysis of features from the data impacting accident severity was conducted using the Apriori algorithm, with a minimum support threshold of 0.2 and confidence threshold of 0.5 to explore relationships between environmental conditions and accident severity. The findings reveal that slight severity accidents dominate the dataset with 81.4% support, while fatal and serious accidents show notably lower occurrence rates. Strong associations emerged from the analysis, particularly in favorable conditions: fine weather with no high winds showed 78.2% support and 80.7% confidence, dry road surfaces demonstrated 70.9% support and 81.0% confidence, daylight conditions exhibited 71.0% support and 81.9% confidence, and 30mph speed limits indicated 64.6% support with 81.8% confidence. A notable pattern emerged showing that most accidents occur under favorable conditions, with high confidence values exceeding 0.80 indicating strong relationships across these factors. The lift values consistently hovering near 1.0 suggest reliable associations between these conditions and accident outcomes, highlighting that favorable environmental conditions do not necessarily prevent accidents from occurring. These can be confirmed from figures 4a and 4b below and full details are in the Jupyter notebook associated with this report.



Association Rules Involving Accident Severity:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
1	(weather_conditions_Fine no high winds)	(accident_severity_Slight)	0.781989	0.814013	0.631351	0.807366	0.991834
2	(road_surface_conditions_Dry)	(accident_severity_Slight)	0.708883	0.814013	0.574482	0.810405	0.995567
4	(road_surface_conditions_Wet or damp)	(accident_severity_Slight)	0.272024	0.814013	0.222540	0.818086	1.005004
5	(light_conditions_Daylight)	(accident_severity_Slight)	0.710087	0.814013	0.581703	0.819200	1.006372
8	(speed_limit_Speed_30)	(accident_severity_Slight)	0.646201	0.814013	0.528485	0.817835	1.004694
22	(road_surface_conditions_Dry, weather_conditions_Fine no high winds)	(accident_severity_Slight)	0.663498	0.814013	0.536155	0.808073	0.992703
24	(road_surface_conditions_Dry)	(accident_severity_Slight, weather_conditions_Fine no high winds)	0.708883	0.631351	0.536155	0.756338	1.197966

Figure 4a: Association Rules Involving Accident Severity

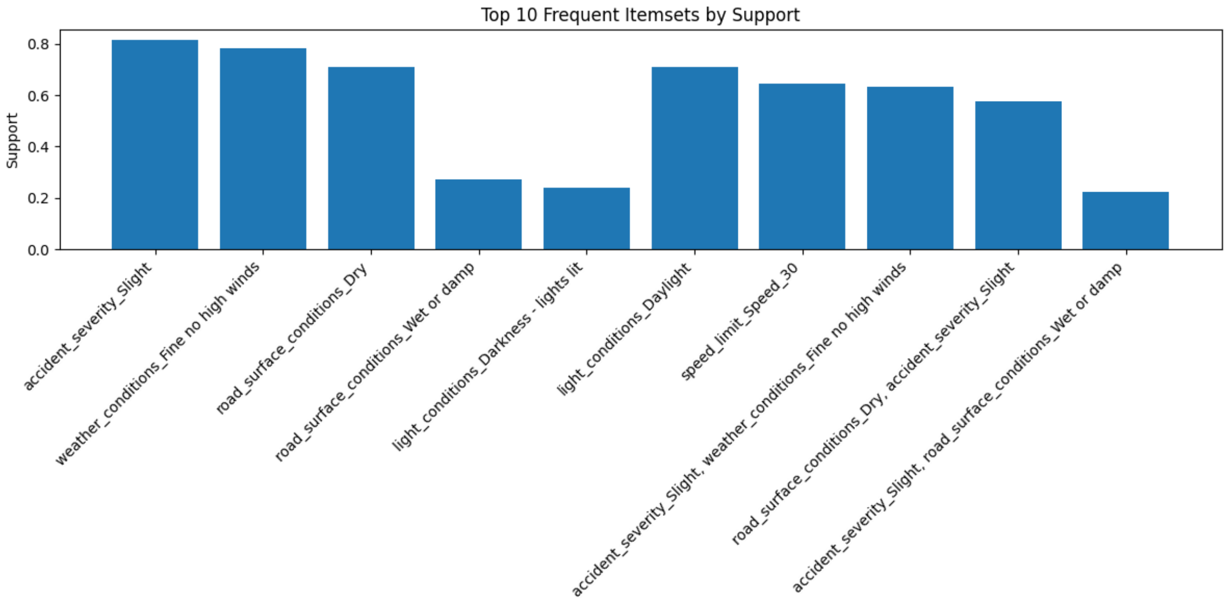


Figure 4b: Top 10 Frequent Itemset by Support

2.2.3 Analysis on Geographical Clustering

From figure 5 the cluster analysis of accidents in the Hull, Humberside, and East Riding of Yorkshire region reveals distinct patterns of accident distribution across the area. The most significant concentration is observed in Kingston upon Hull with 358 accidents, followed by notable clusters in Grimsby Northeast Lincolnshire (163 accidents), the western region (83 accidents, Scunthorpe), and areas east (42 accidents) and north (25 accidents) of Hull. The visualization demonstrates a clear correlation between urban density and accident frequency, with major clusters centered around urban areas and along key transportation corridors. Smaller clusters, represented by single-digit numbers in green, are distributed throughout rural areas and smaller settlements, particularly along main transportation routes and in the northern parts of the East Riding. These clusters when clicked also tell the severity of the accident. This pattern strongly suggests that accident concentrations align with population density and major transportation infrastructure, with the highest rates occurring in urban centers, at major road junctions, and along key connecting routes between settlements. These colors on the map are for visual distinction only and do not indicate accident severity or any other attribute, full details are in the Jupyter notebook.

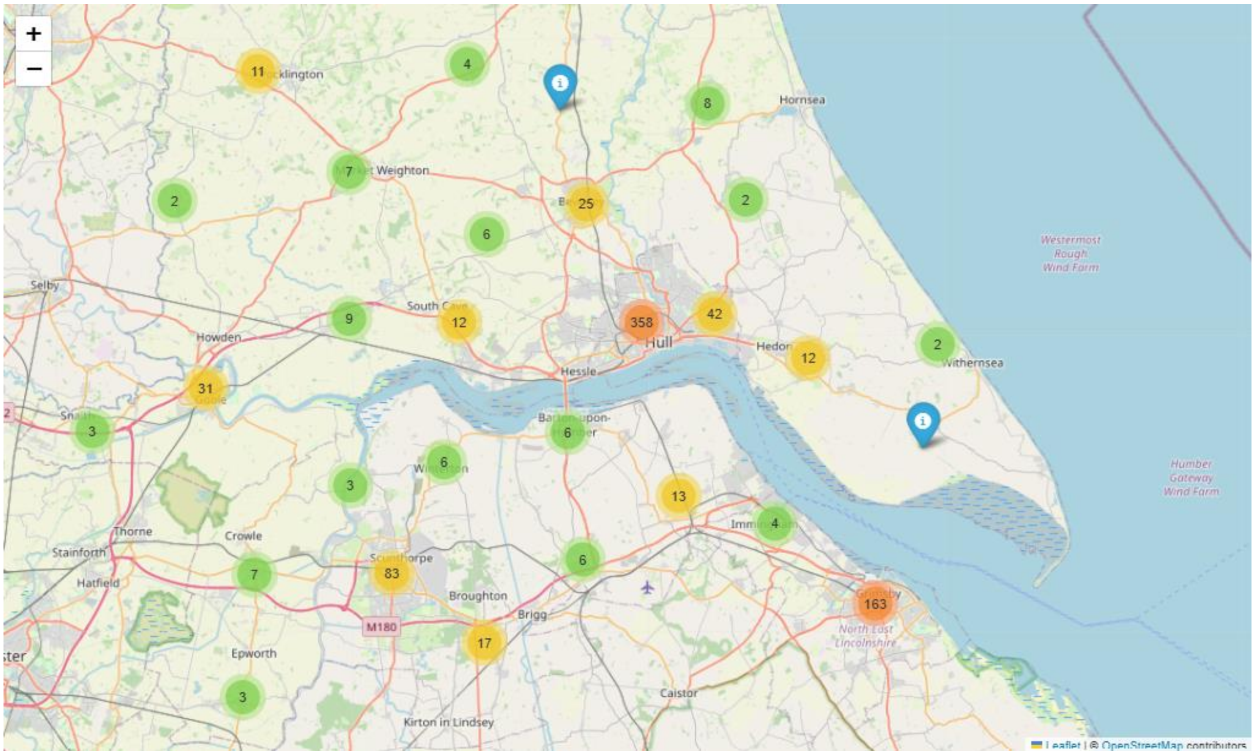


Figure 5: Geographical Accidents Clustering

2.2.4 Analysis on Time Series Modelling

The time series analysis reveals a two-tier pattern of predictability across geographic scales:

Regional Weekly Analysis with Key Performance Metrics:

- North Yorkshire: MSE 38.03 (Optimal accuracy)
- South Yorkshire: MSE 77.05
- West Yorkshire: MSE 87.05 (Highest variance)

The analysis reveals an inverse relationship between urbanization and prediction accuracy, with rural areas demonstrating more predictable patterns than complex urban environments as evident from figure 6a, 6b and 6c.

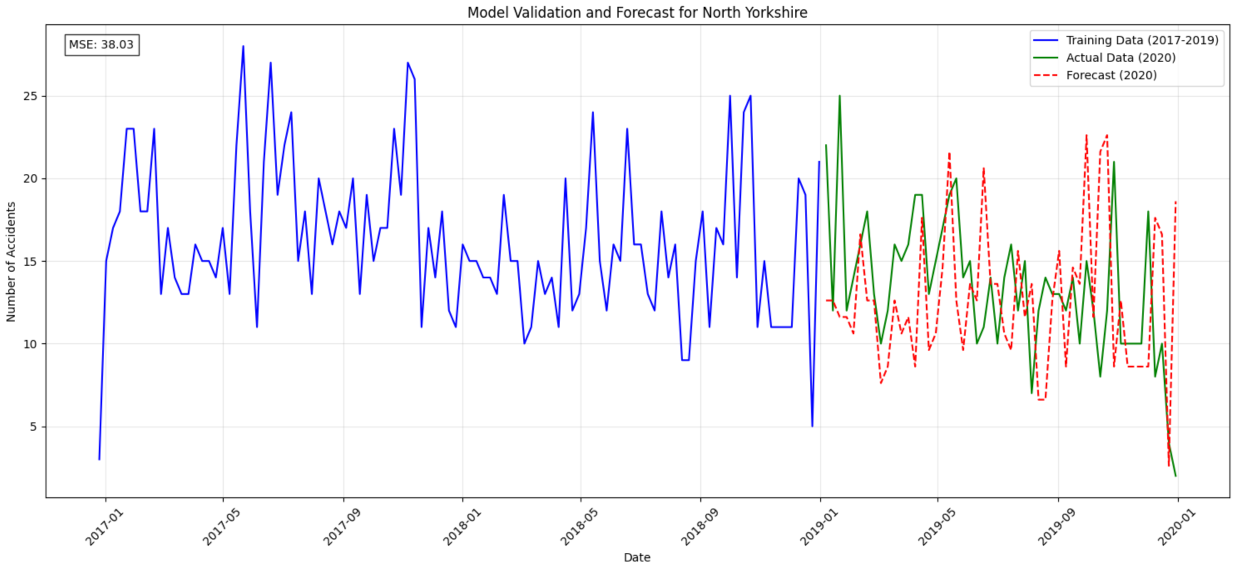


Figure 6a: Model Validation and Forecast for North Yorkshire

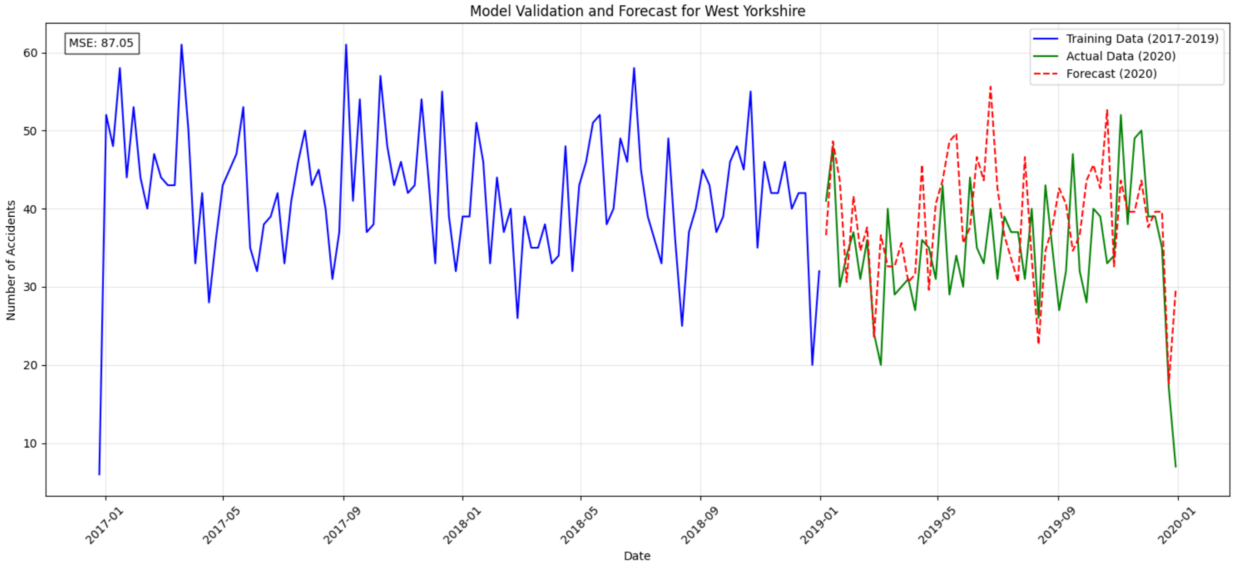


Figure 6b: Model Validation and Forecast for West Yorkshire

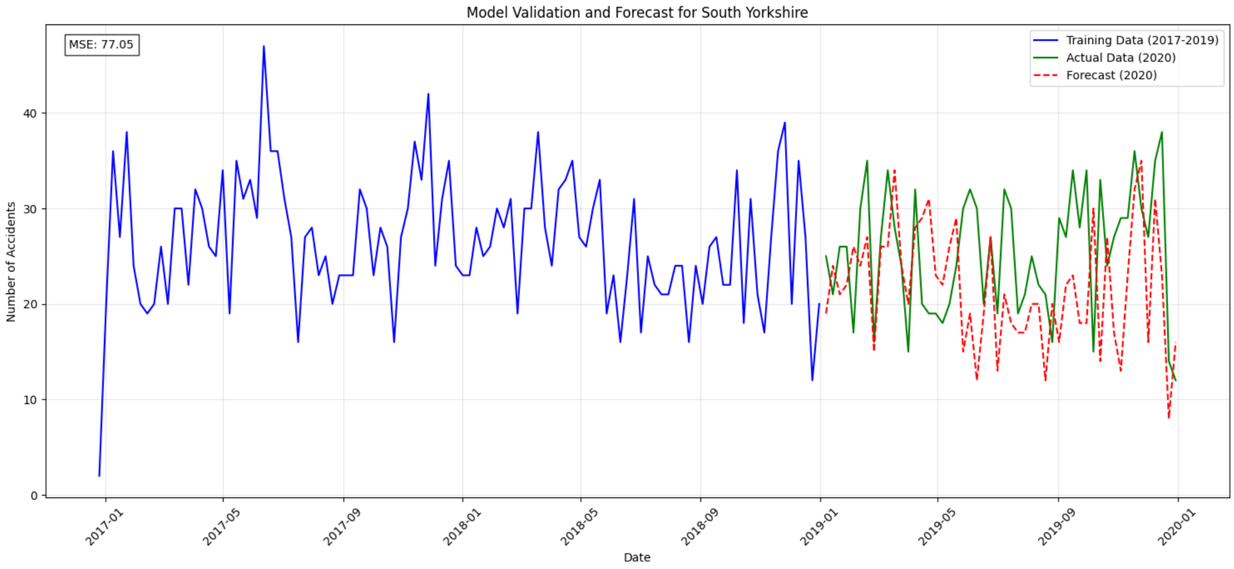


Figure 6c: Model Validation and Forecast for South Yorkshire

Granular Daily Analysis (Hull LSOAs):

- Hull 016D: MSE 0.06 - Sporadic single-accident spikes
- Hull 020B: MSE 0.03 - Extreme data sparsity
- Hull 024B: MSE 0.00 - Near-zero continuous pattern

From figure 7a, 7b and 7c, the analysis shows that it's easier to predict accidents for larger regions over weeks than for small neighborhoods over days. While weekly regional models maintain reasonable accuracy (MSE 38-87), daily LSOA-level predictions face fundamental challenges with extreme sparsity (MSE 0.00-0.06), suggesting that accident prediction requires different methodological approaches based on both geographic and temporal scales. This dual-scale understanding demonstrates why a one-size-fits-all prediction approach is inadequate for comprehensive accident forecasting.

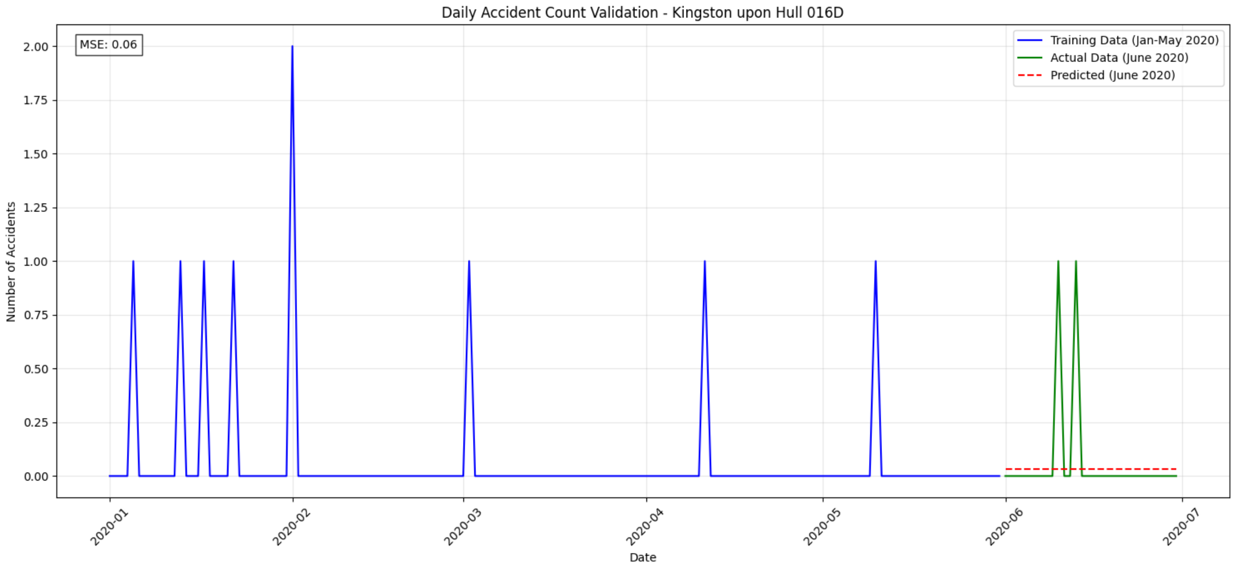


Figure 7a: Daily Accident Count Validation for Kingston upon Hull 016D

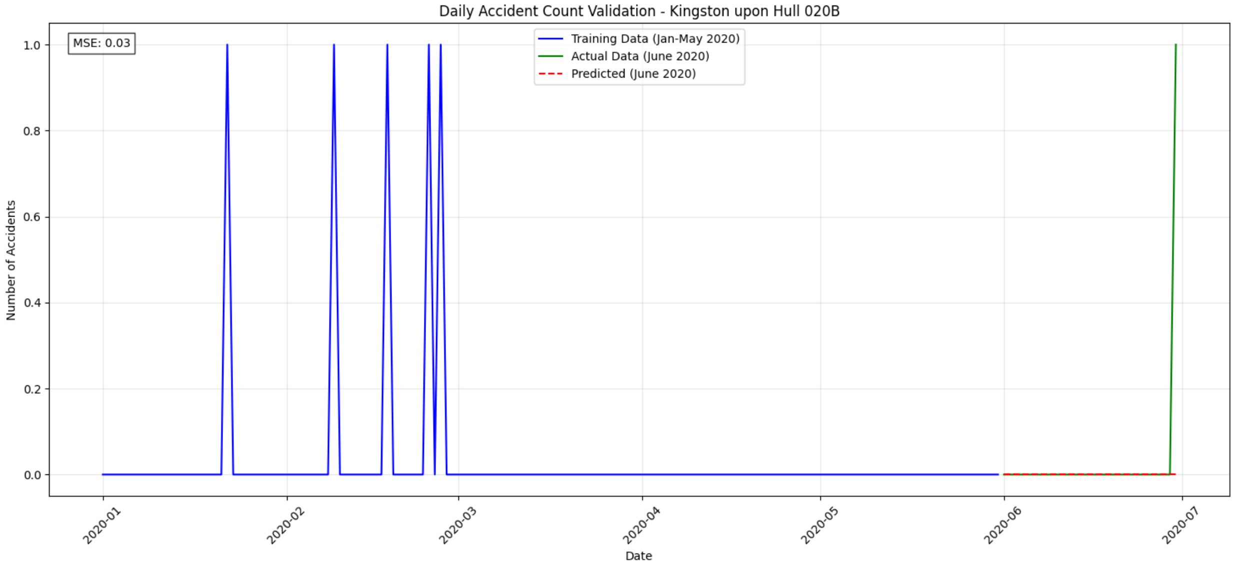


Figure 7b: Daily Accident Count Validation for Kingston upon Hull 020B

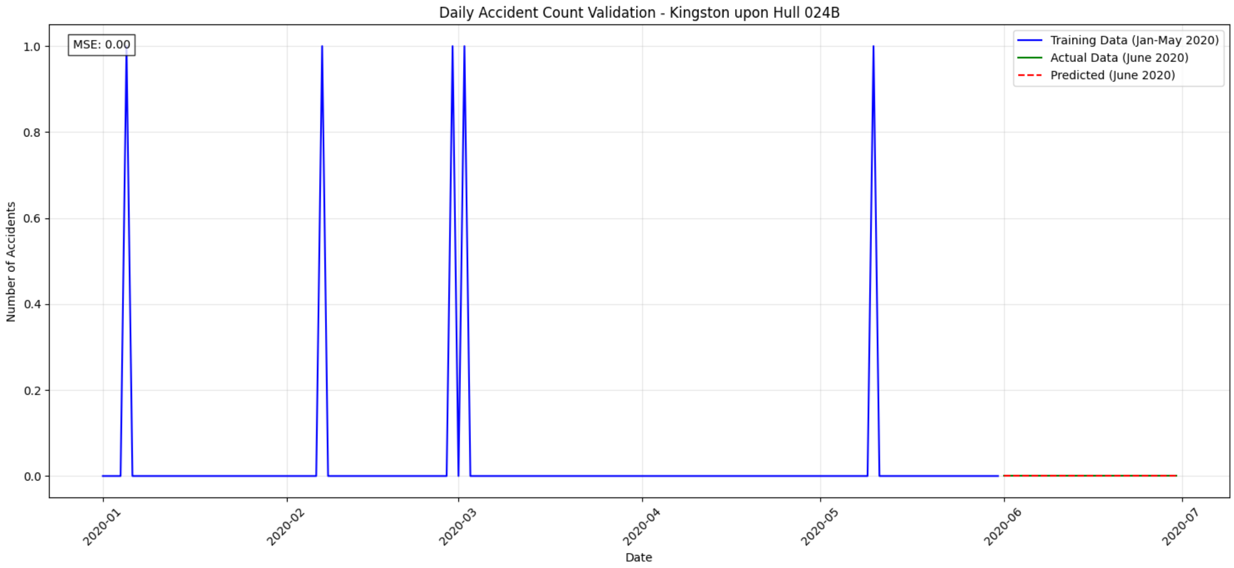


Figure 7c: Daily Accident Count Validation for Kingston upon Hull 024B

2.2.5 Social Network Analysis

The network visualization in figure 8 reveals a sophisticated structure of 4,039 nodes and 88,234 edges, exhibiting distinct characteristics of a real-world complex network. Key metrics demonstrate compelling network characteristics:

- Density: 0.0108 (indicating strategic sparsity)
- Average Degree: 43.69 connections per node
- Total Nodes: 4,039
- Total Edges: 88,234

The network's topology reveals a critical balance between efficiency and resilience, evidenced by its sparse density (1.08%) yet robust average connectivity (43.69 connections). This structure, visualized using NetworkX's spring layout algorithm, demonstrates scale-free characteristics - a pattern typically associated with evolved, optimized networks. The clear clustering patterns suggest naturally formed communities, with darker blue regions indicating high-density interaction zones.

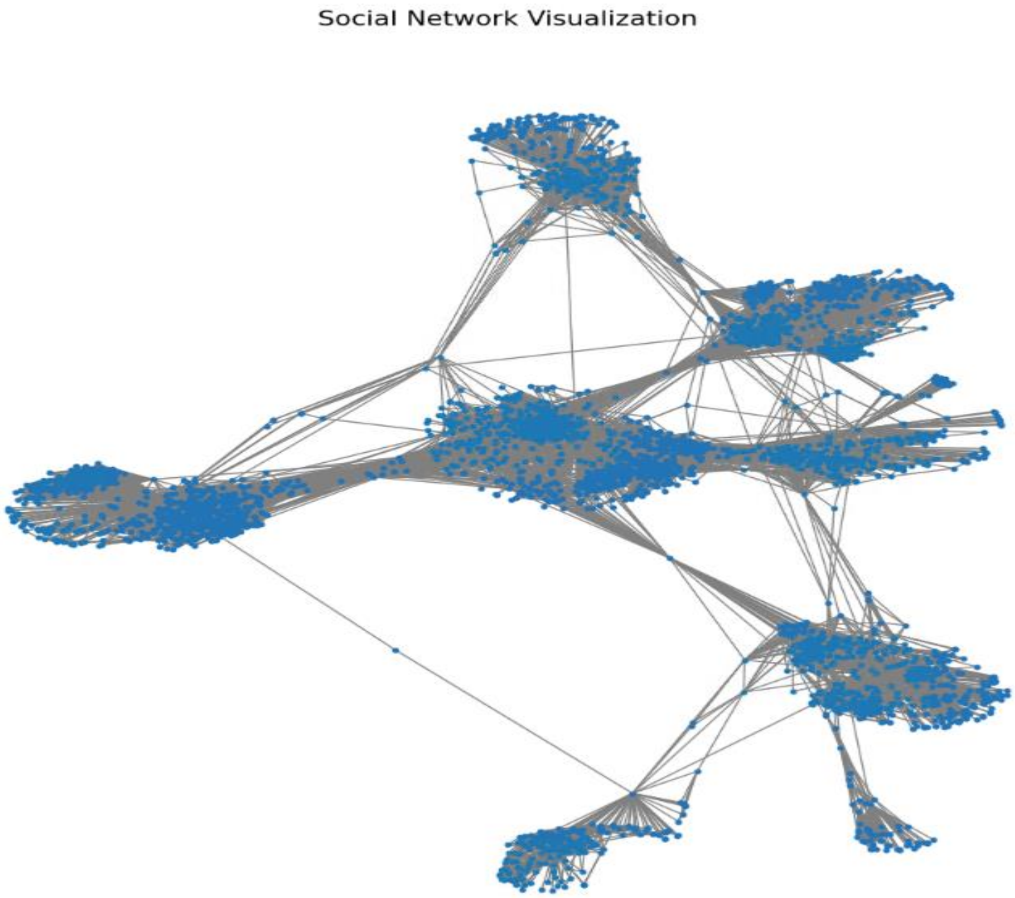


Figure 8: Social Network Visualization

Edge Betweenness Centrality

The distribution on figure 9 reveals a striking hierarchical structure in the network's connectivity patterns with key metrics:

- Peak Frequency: ~80,000 edges at near-zero centrality
- Maximum Centrality: 0.175
- Distribution: Heavily right skewed

The extreme concentration of low centrality values (≈ 0) with a long tail extending to 0.175 reveals a critical "bridge and cluster" network architecture. This highly skewed distribution demonstrates that while most connections serve local community functions, a select few edges function as vital inter-community bridges. This topology creates a robust yet efficient network structure where approximately 1% of edges facilitate critical cross-community information flow, suggesting an evolved, optimized network architecture typical of resilient social systems.

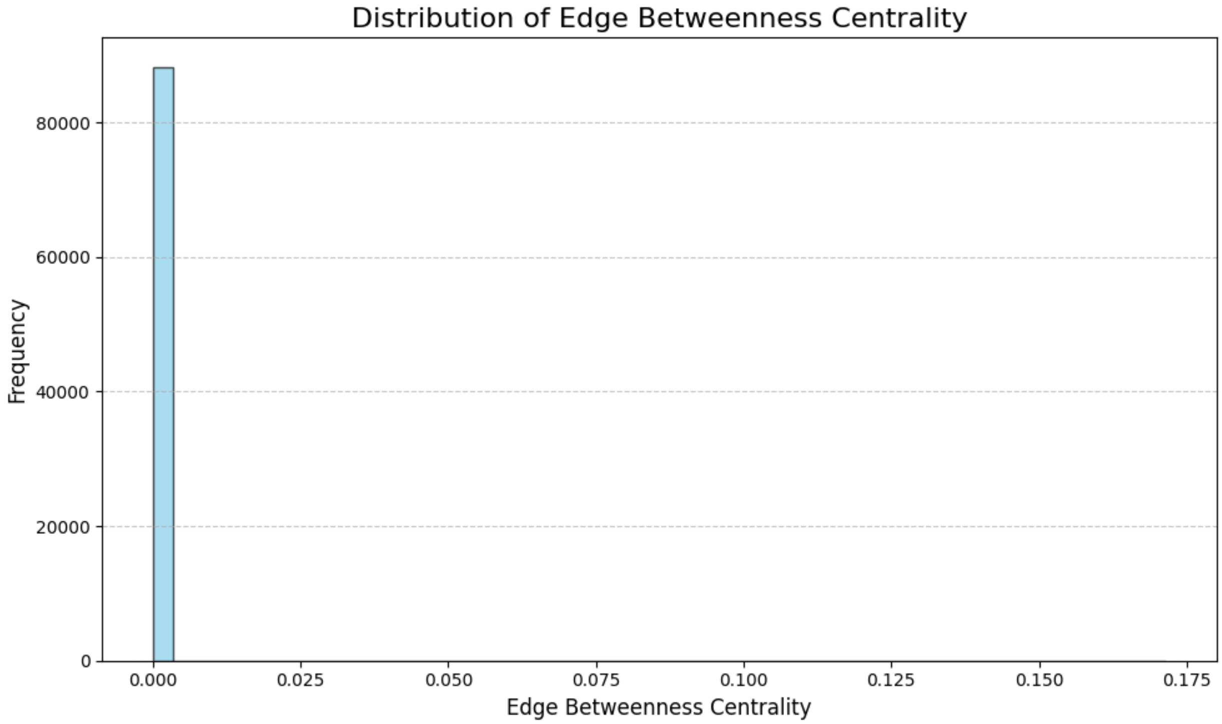


Figure 9: Distribution of Edge Betweenness Centrality

Community Detection Algorithm

The strategic application of both Girvan-Newman (GN) and Louvain community detection algorithms reveals compelling multi-scale network organization insights. GN identified a stark binary division (3,833 and 206 nodes), suggesting a core-periphery structure, while Louvain uncovered 16 distinct communities (ranging from 19 to 548 nodes), revealing granular substructures as seen from figure 10 and 11 respectively.

Girvan-Newman Justification:

- Excels in identifying hierarchical structures through edge betweenness
- Provides deterministic results ensuring reproducibility
- Specifically effective for transportation networks where flow patterns are crucial
- Better at detecting global community structures

Louvain Method Justification:

- Offers superior computational efficiency for large networks
- Optimizes modularity through multi-level optimization
- Particularly effective at detecting natural groupings in spatial networks
- Better suited for identifying local, tightly knit communities

The stark contrast in results (2 vs. 16 communities) validates this dual approach, revealing both macro-level divisions and micro-level clustering patterns.

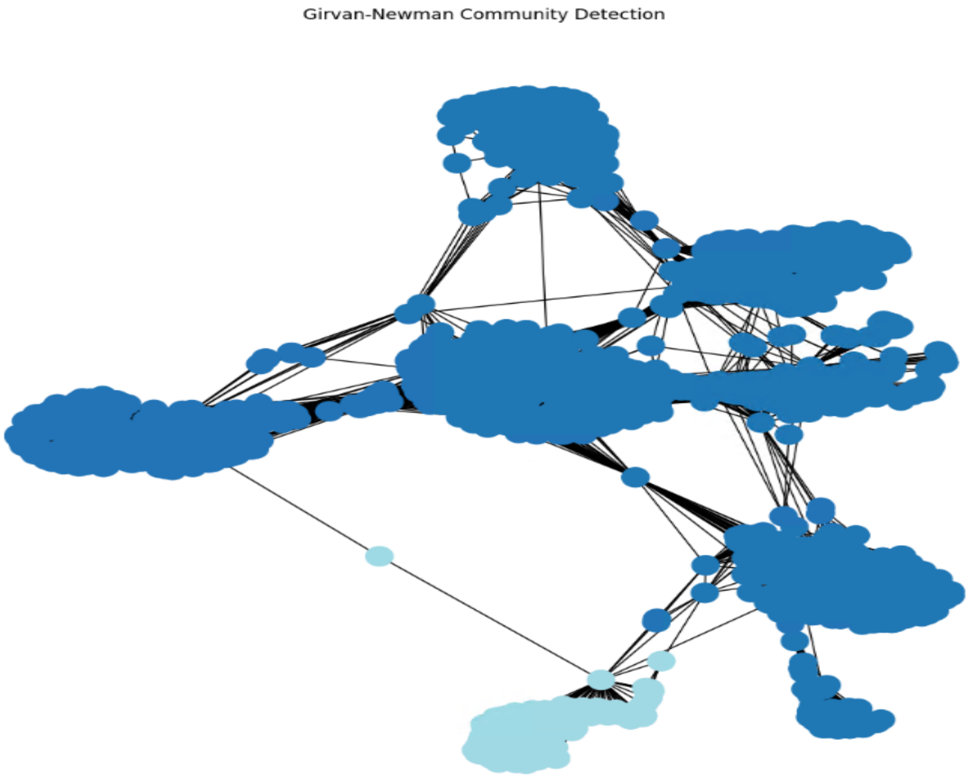


Figure 10: Girvan-Newman Community Detection

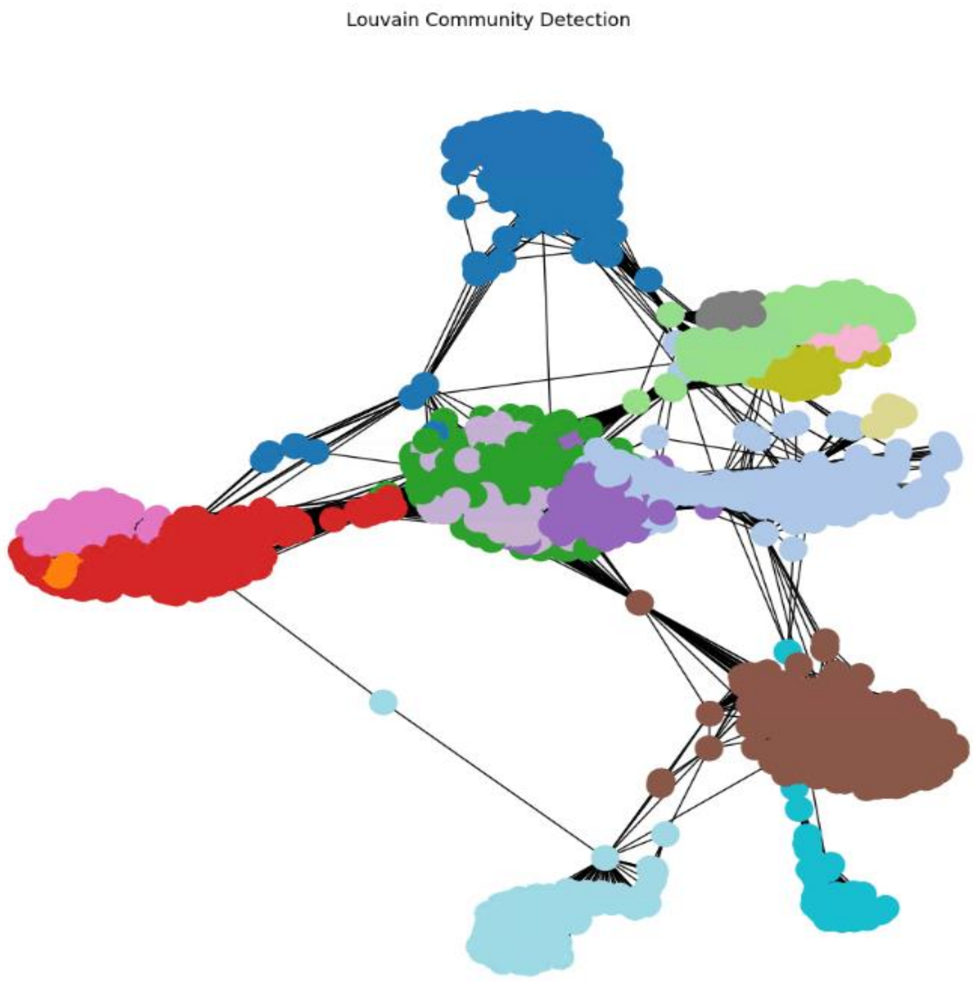


Figure 11: Louvain Community Detection

3.0 Predictions

As seen from figure 12a, 12b, and 12c, the forecasting results revealed distinct patterns across regions, with West Yorkshire showing the highest accident counts (30-55 per week), followed by South Yorkshire (15-35 per week), and North Yorkshire with the lowest counts (2-22 per week). While the models effectively captured seasonal patterns, their prediction accuracy varied with urban density, showing better performance in less populous areas. These insights can inform resource allocation and emergency response planning.

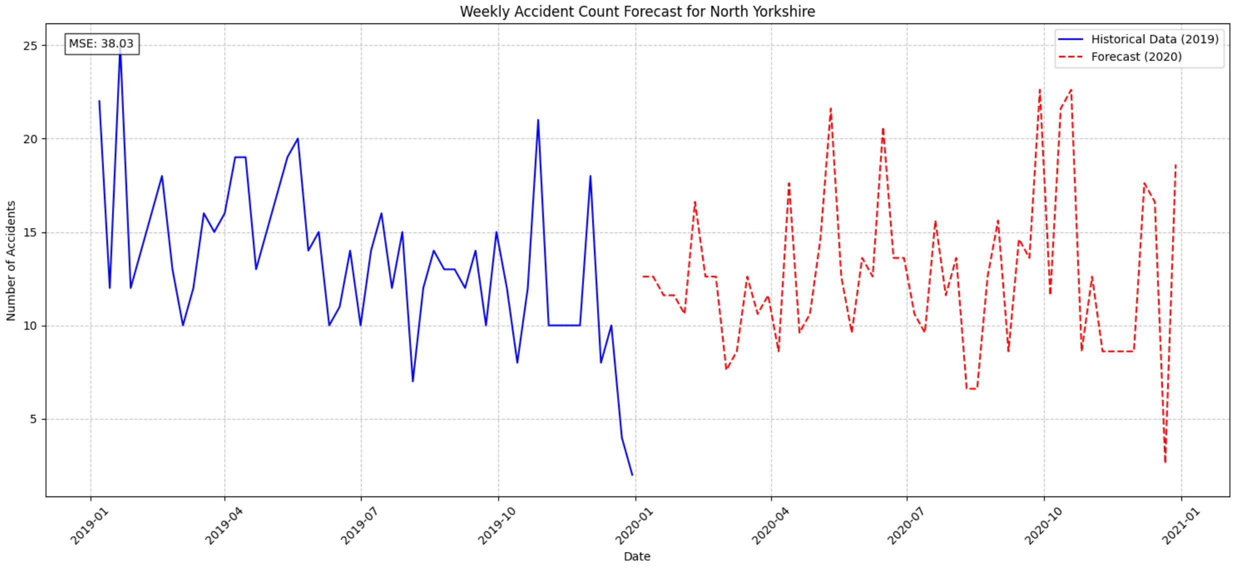


Figure 12a: Weekly Accident Count Forecast for North Yorkshire

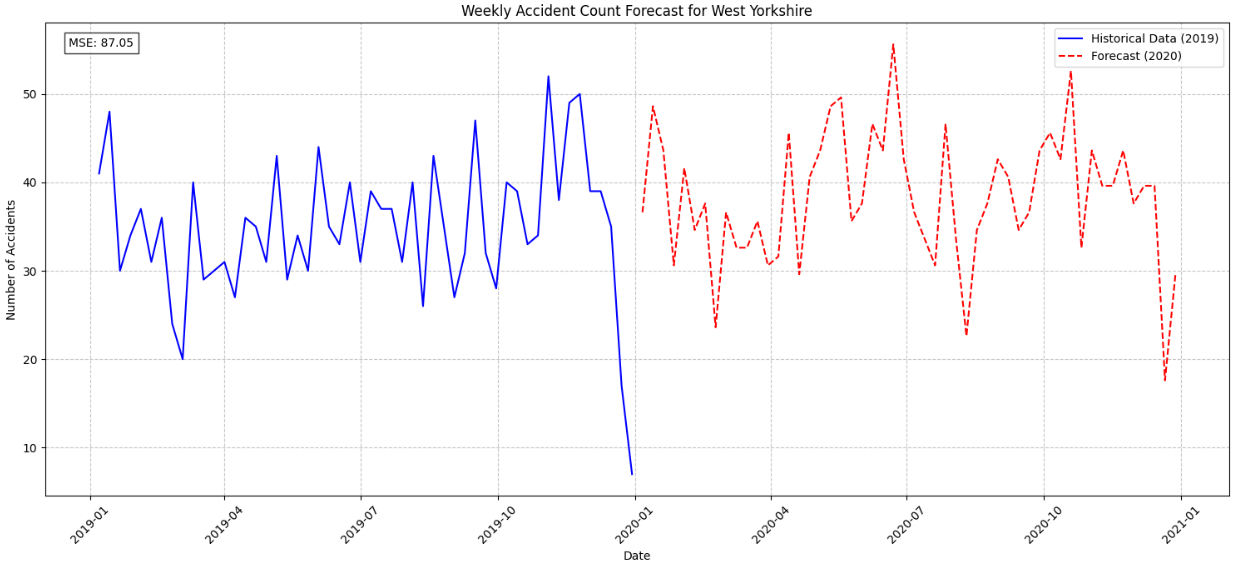


Figure 12b: Weekly Accident Count Forecast for West Yorkshire

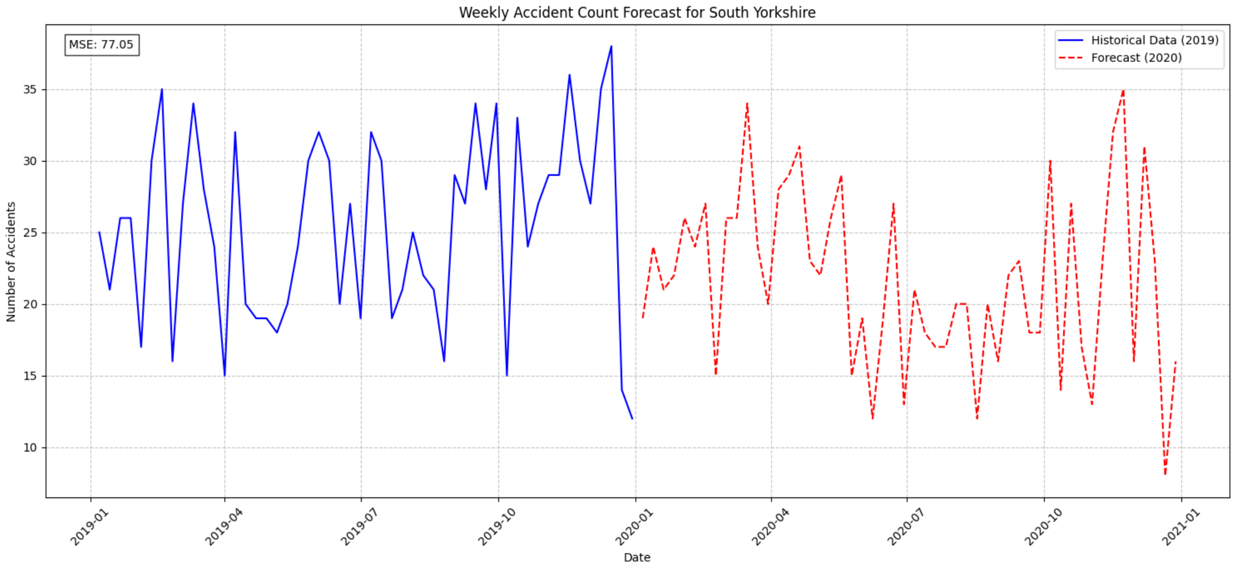


Figure 12c: Weekly Accident Count Forecast for South Yorkshire

Micro-Level (Hull Lower Super Output Areas)

The time series analysis conducted for the three highest-accident LSOAs in Hull revealed important insights about forecasting rare events at such a granular geographic level, this is evident from figure 13a, 13b and 13c. This sparsity violates fundamental SARIMA assumptions of continuous data and normal distribution, resulting in mostly zero-value predictions with very low confidence. The forecasting challenge is evidenced by the validation plots showing large gaps between sparse actual events and the model's conservative predictions near zero. While regional models effectively capture accident trends across Yorkshire (as evidenced in Question 6), neighborhood-level prediction face a mathematical barrier where increased geographic precision reduces predictive power, challenging conventional approaches to road safety forecasting.

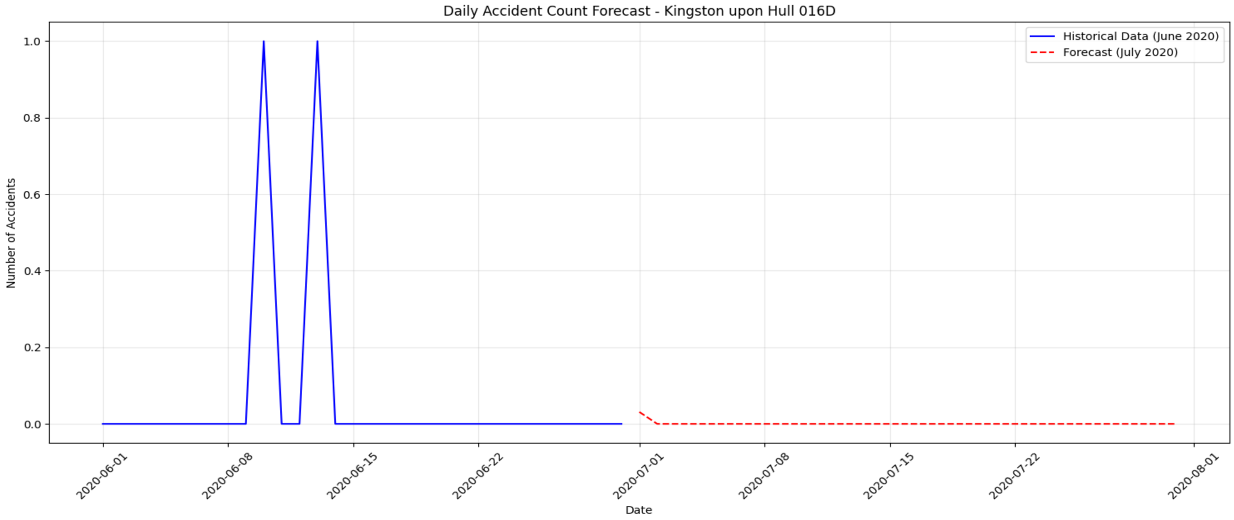


Figure 13a: Daily Accident Count Forecast for Kingston upon Hull 016D

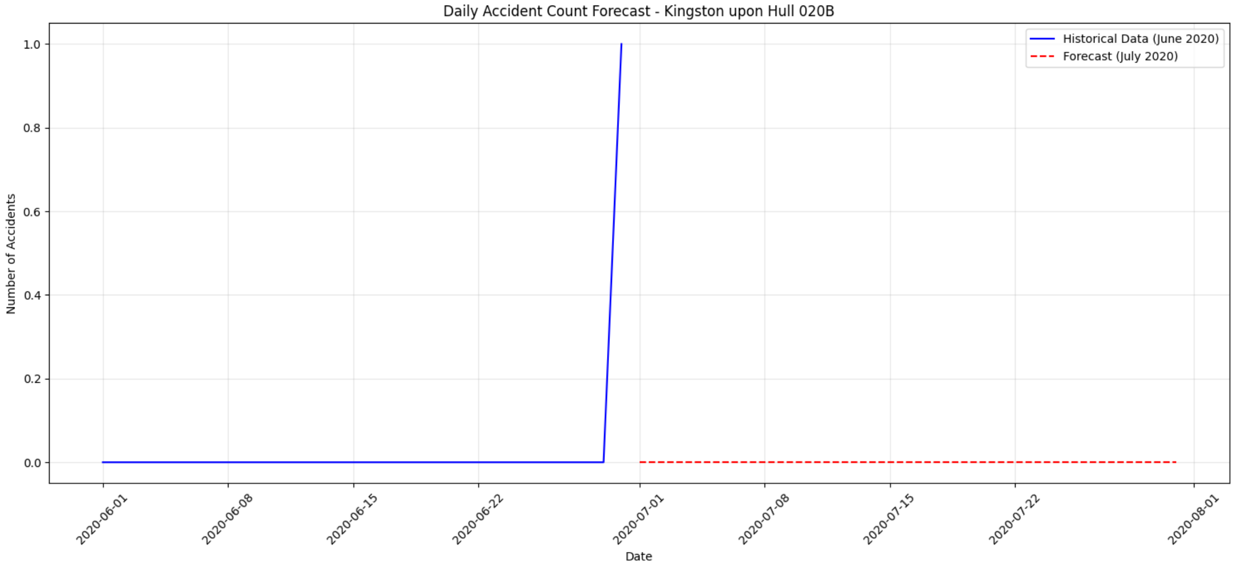


Figure 13b: Daily Accident Count Forecast for Kingston upon Hull 020B

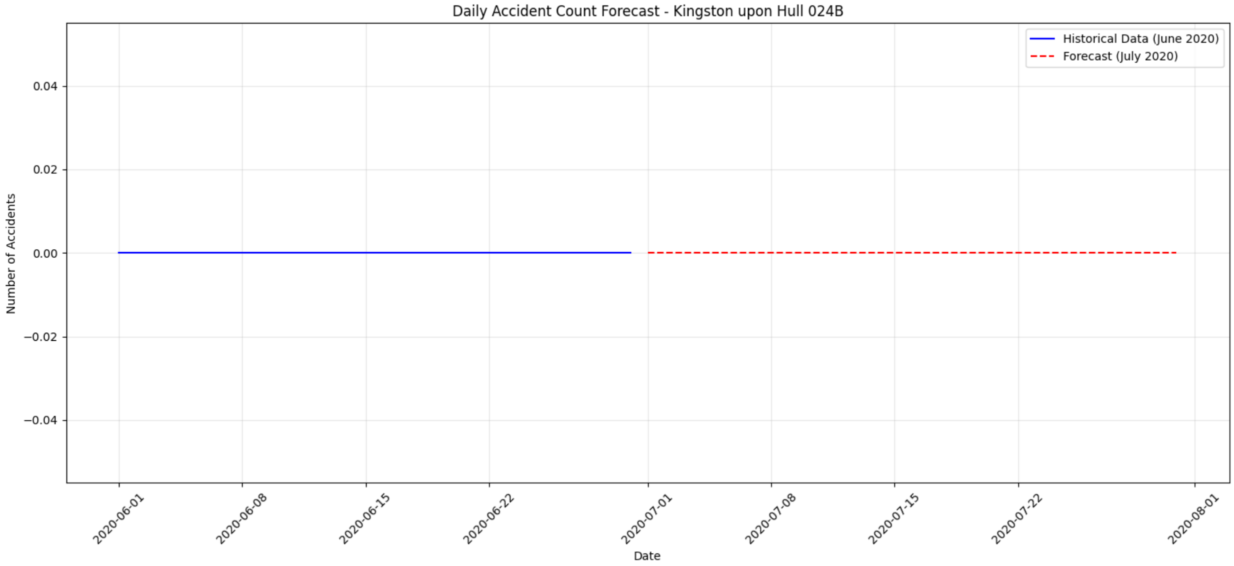


Figure 13c: Daily Accident Count Forecast for Kingston upon Hull 020B

4.0 Recommendations

Based on the analysis and predictions from the data, the following recommendations are proposed to government agencies for improving road safety:

- Focus on driver behavior in favorable conditions and enhance monitoring in 30mph zones, as the data shows most accidents (81.4%) occur in good weather, dry roads, and daylight conditions, particularly in urban areas with 30mph limits.
- Implement targeted traffic management during peak hours and high-risk days (particularly Fridays and Saturdays), based on clear temporal patterns in the accident data showing higher incident rates during these periods.
- Prioritize infrastructure improvements in identified accident hotspots, particularly in Hull City and Northeast Lincolnshire, where data shows consistently higher accident frequencies.
- Launch focused public safety campaigns for high-risk periods (late afternoon and early evening), supported by enhanced law enforcement presence, as temporal analysis shows these times have elevated accident rates.
- Develop specific safety programs around schools and workplaces during morning rush hours (7-9 AM), as data indicates significant accident clusters during these commuting periods.

5.0 References

Datascientest (2023) Data cleaning: Definition, methods and relevance in Data Science Available online: https://datascientest.com/en/data-cleaning-definition-methods-and-relevance-in-data-science [Accessed 03/12/2024].

Skiena, S.S. (2017). The Data Science Design Manual. Springer International Publishing https://link.springer.com/book/10.1007/978-3-319-55444-0

Perez, R.C.L. (2023) 'Data Mining technique: Application of Apriori algorithm for road accident analysis', HCMCOUJS-Engineering and Technology, 13(2), pp. 60-68. https://doi.org/10.46223/HCMCOUJS.tech.en.13.2.2831.2023

Reported road casualties in Great Britain: notes, definitions, symbols and conventions:Links to an external site. government guidance on the data set.

https://www.gov.uk/government/publications/road-accidents-and-safety-statistics-notes-and-definitions/reported-road-casualties-in-great-britain-notes-definitions-symbols-and-conventions

stats20-2011.pdf: Download stats20-2011.pdf:Detailed guidance on how to complete accident reporting forms.

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995423/stats20-2011.pdf

Road Traffic Accidents Statistics Form: Links to an external site. the form used to report road traffic accidents.

https://assets.publishing.service.gov.uk/media/60d0cc548fa8f57ce4615110/stats19.pdf

dft-road-casualty-statistics-road-safety-open-dataset-data-guide-2023-1.xlsx : Download dft-road-casualty-statistics-road-safety-open-dataset-data-guide-2023-1.xlsx :This form states the meaning of the numerical values in each column.

https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-accidents-safety-data