BIG DATA AND DATA MINING

(REPORT)

BY

PASCAL UGONNA AKANO

202445626

TASK A

Introduction

This report investigates road traffic accidents in the UK during 2019 using official data provided by the Department for Transport (Department for Transport, 2020). The primary objective is to understand where, when, and under what conditions accidents tend to happen, and how these factors affect accident severity.

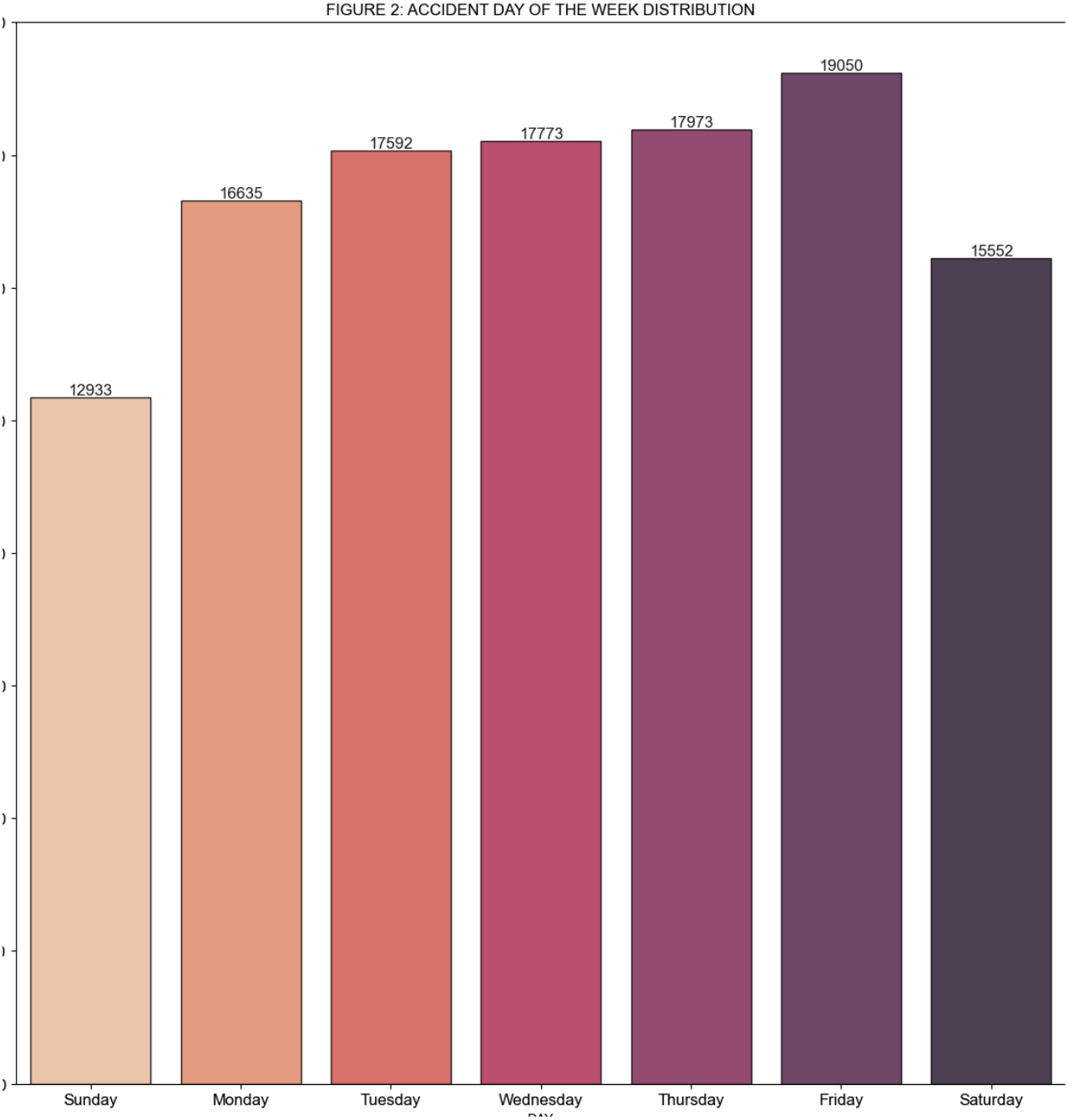
The dataset contains granular records on the accidents themselves, vehicles involved, and casualties, which were cleaned and integrated using a structured Python pipeline. By applying data mining techniques and predictive modelling, we identify accident patterns, develop forecasts, and propose policy recommendations. Exploratory analysis using clustering, association rule mining, and time series forecasting helps reveal the dynamics behind road risks, while also offering a data-driven foundation for safety planning (Hyndman and Athanasopoulos, 2018; Bhandari, 2022). The insights drawn from this project aim to assist policymakers and transportation authoriyes ties in reducing accident risks and improving road safety outcomes.

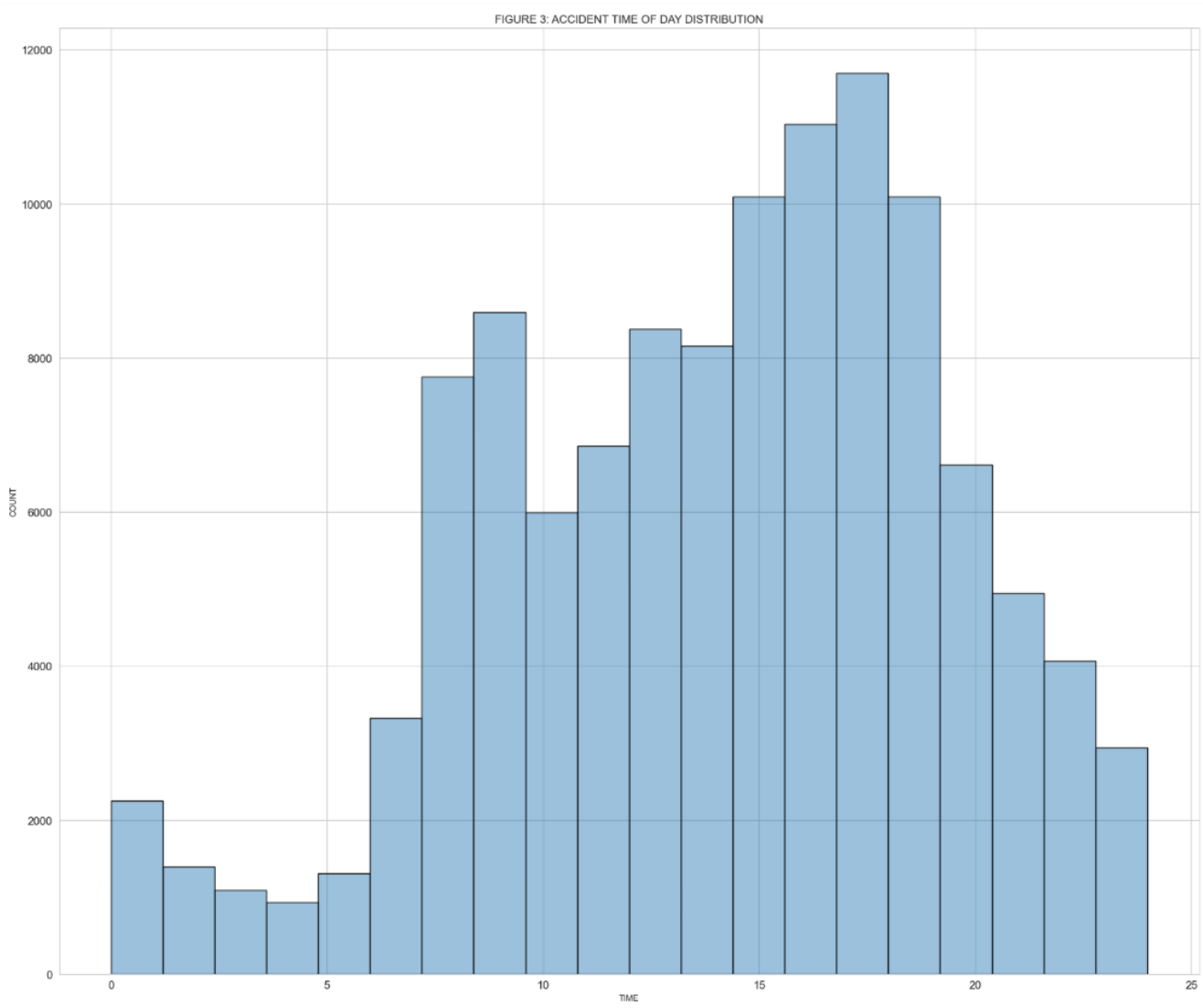
Analysis and Results

Task 1: When Do Accidents Occur?

Accident data showed that Fridays and Thursdays recorded the highest number of incidents, likely due to increased travel at the end of the workweek. Sundays had the fewest, reflecting reduced weekday commuting. In terms of time, accidents peaked between 8–9 AM and 3–6 PM, which corresponds to typical rush hour periods. These trends point to congestion and commuting as key factors influencing accident rates (RoSPA, 2001).

FIGURE 2: ACCIDENT DAY OF THE WEEK DISTRIBUTION





- *Figure 1: Weekly Accident Distribution by Day*
- *Figure 2: Accident Frequency by Time of Day*

Task 2: Motorcycle Accident Patterns

Motorcycle accidents were analysed across engine sizes:

- **Under 125cc**: Accidents peaked on weekdays, especially Fridays, indicating commuter use.

- **125cc–500cc**: Most incidents occurred on Thursdays and Fridays, suggesting a mix of commuting and weekend travel.
- **Over 500cc**: Incidents increased from midweek and peaked on Saturdays, linked to recreational riding.

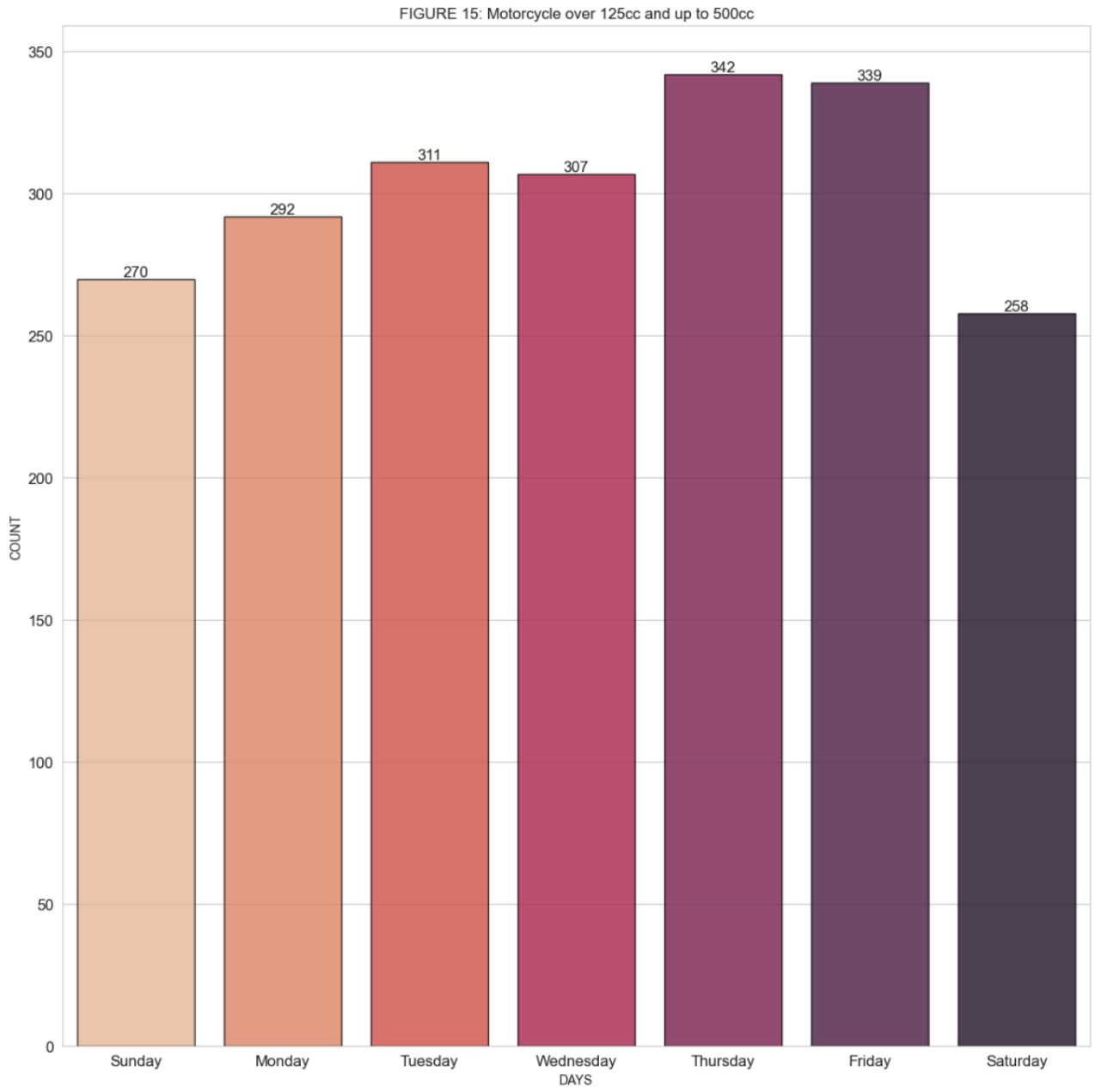
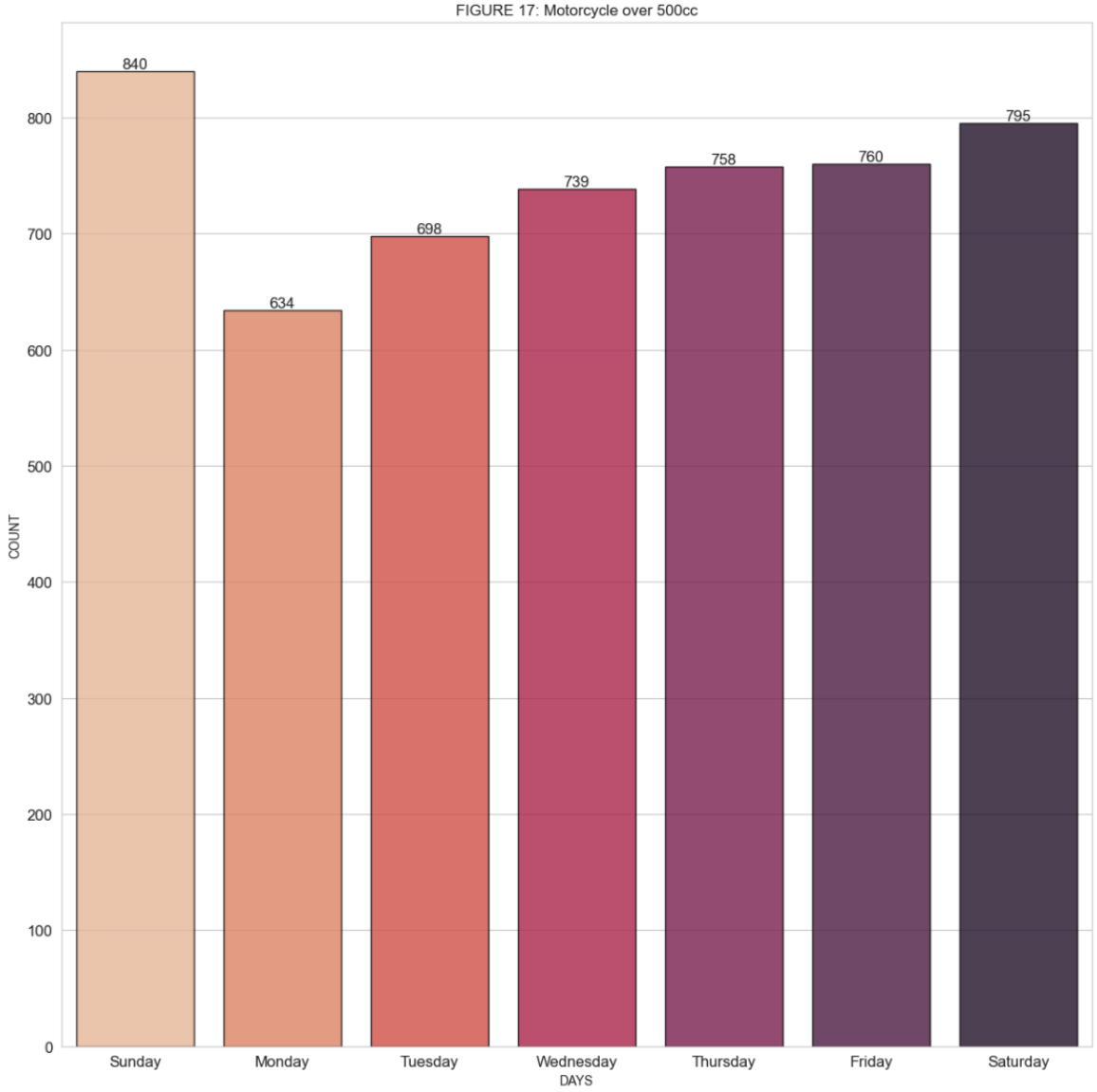
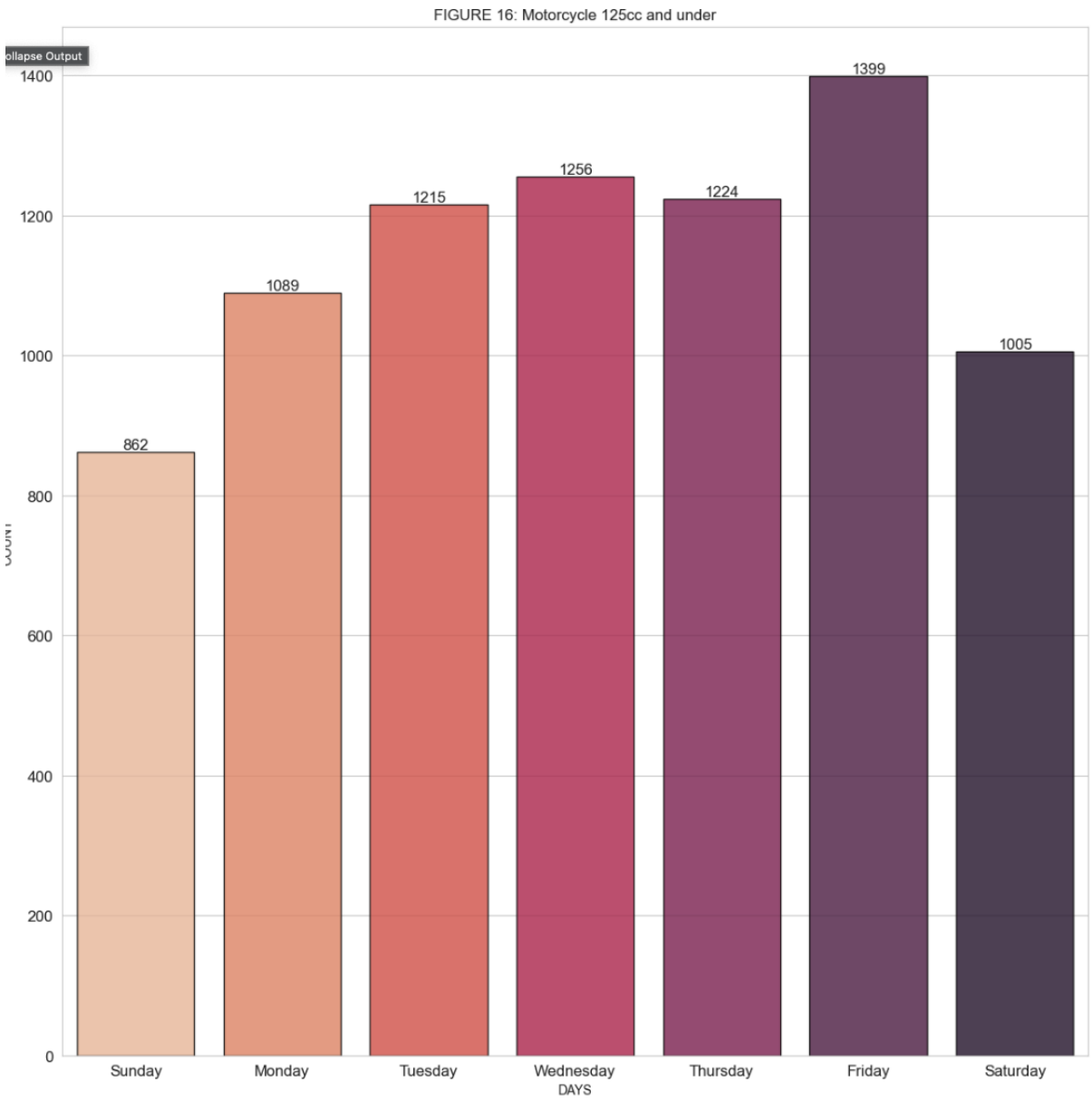


FIGURE 15: Motorcycle over 125cc and up to 500cc
COUNT
DAYS
0
50
100
150
200
250
300
350
270
292
311
307
342
339
258
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday

FIGURE 17: Motorcycle over 500cc



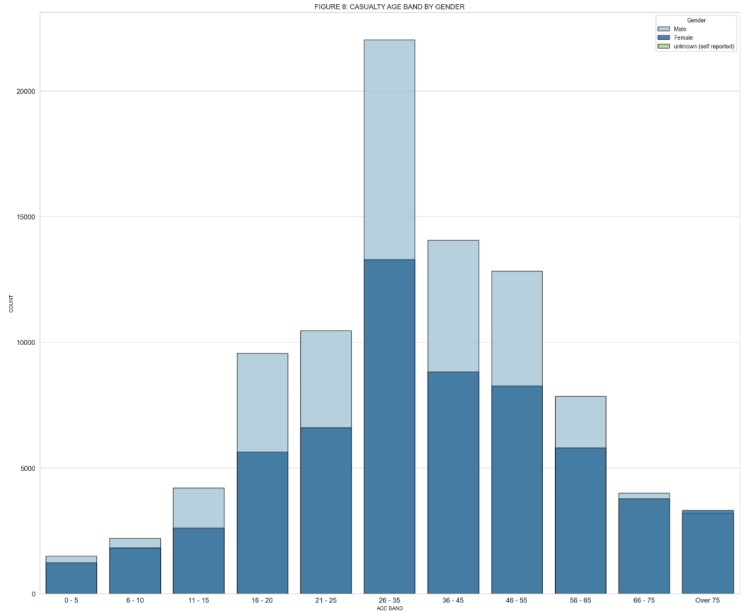


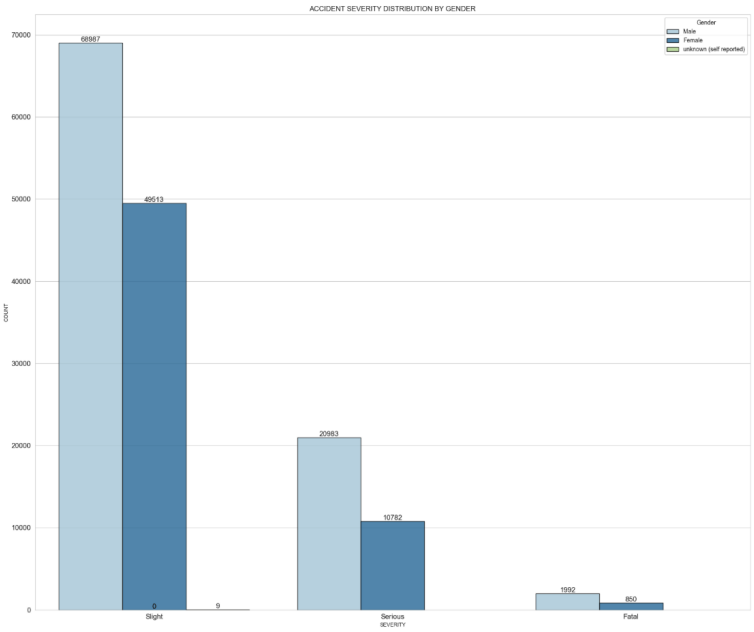
Figures 3-5: Motorcycle Accidents by Day and Engine Size

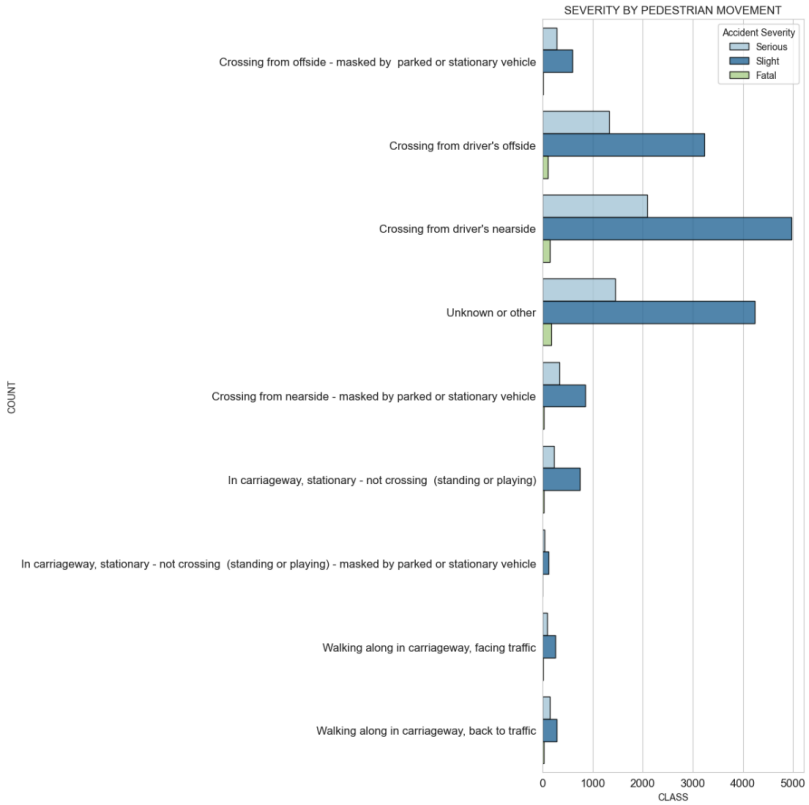
Task 3: Pedestrian Accident Trends

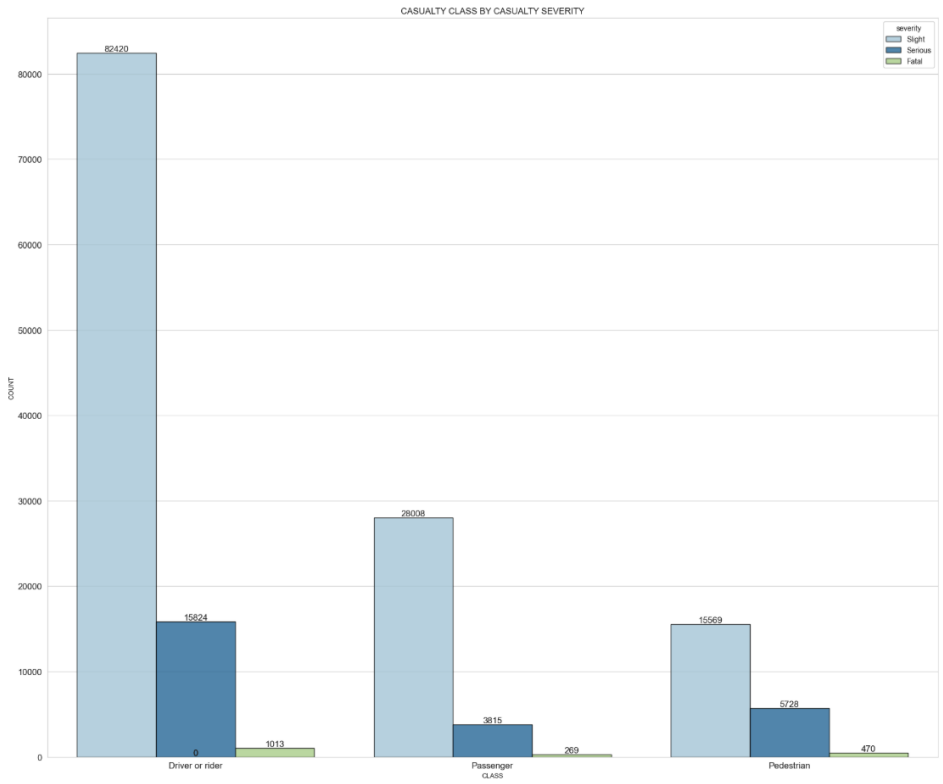
Pedestrian accidents showed a higher frequency among males, particularly those aged 16 to 35, with serious and fatal injuries being more common in this group (Figure 6: Casualty Age Band by Gender, Figure 7: Casualty Severity by Gender). Many incidents occurred while individuals crossed from the driver's nearside or offside, often outside designated crossings or from behind parked vehicles (Figure 8: Severity by Pedestrian Movement). This highlights visibility issues and the absence of safe crossing infrastructure. Though pedestrians made up a smaller portion of total casualties, their injuries were more severe compared to drivers or passengers (Figure 9:

Casualty Class by Severity). These findings underline the need for targeted pedestrian safety measures, such as better crossings, traffic calming zones, and visibility enhancements in urban areas.







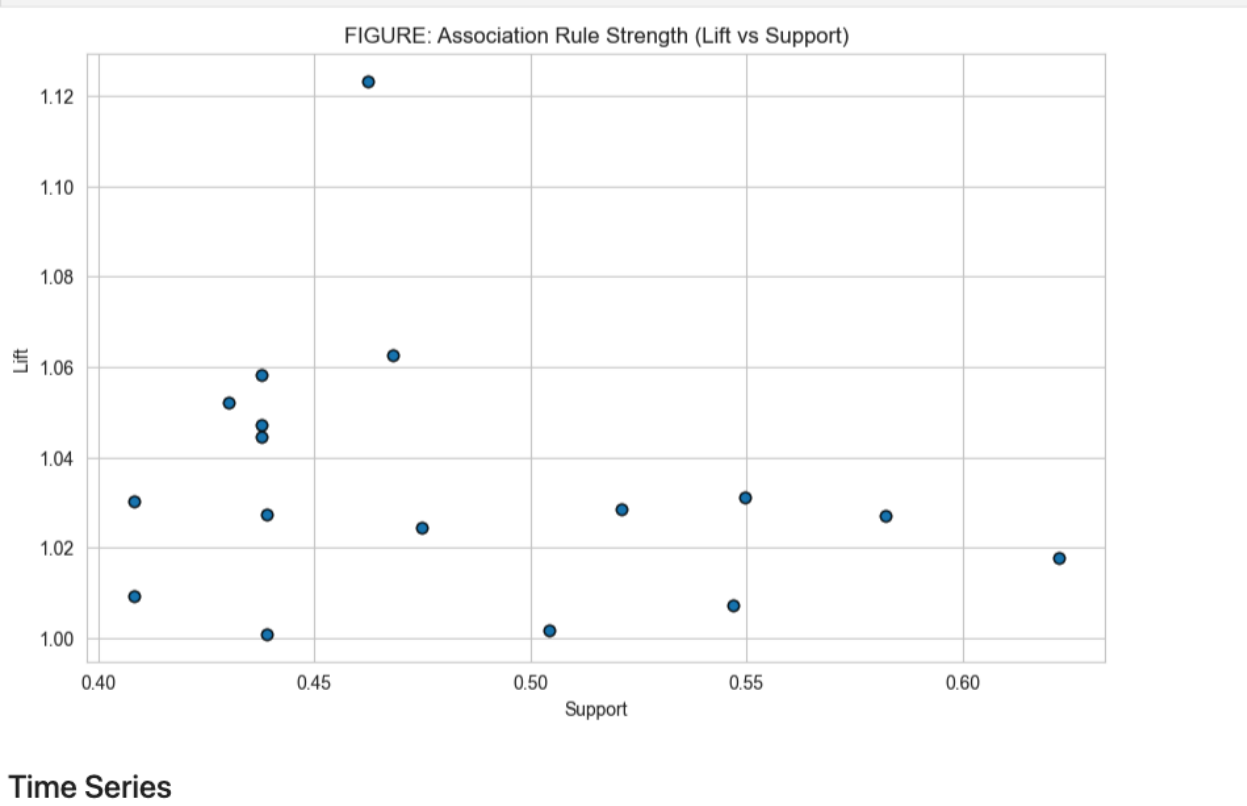


Figures 6-9: Pedestrian Accident Demographics and Severity

Task 4: Accident Severity via Association Rules

Using the Apriori algorithm on one-hot encoded features, several meaningful patterns were discovered. The strongest rules linked vehicle type 9 (cars) and road type 6 (urban roads) with accident severity 3 (slight). For example: (vehicle_type_9, road_type_6) → (accident_severity_3) had over 70% confidence and a lift above 1.0, indicating a reliable association.

Other frequent itemsets involved 30 mph speed limits and two vehicles involved, supporting prior findings that slight accidents are more common on urban roads with lower speeds (Brownlee, 2014; DataTechNotes, n.d.).



Time Series

Figure 10: Lift vs Support Plot

Task 5: Regional Accident Clustering

The analysis focused on accident locations within Humberside, especially Kingston upon Hull and nearby areas. Using K-Means clustering on geographic coordinates, several high-density

zones were identified, mostly around busy roads and intersections. To explore unusual patterns, the Local Outlier Factor (LOF) was used to highlight data points that stood out from the rest. These could reflect rare incidents, misreported data, or disruptions like roadworks or bad weather. Visual breakdowns by severity and urban versus rural settings showed that most severe accidents occurred in urban areas, which suggests that city centres may benefit most from targeted road safety improvements.

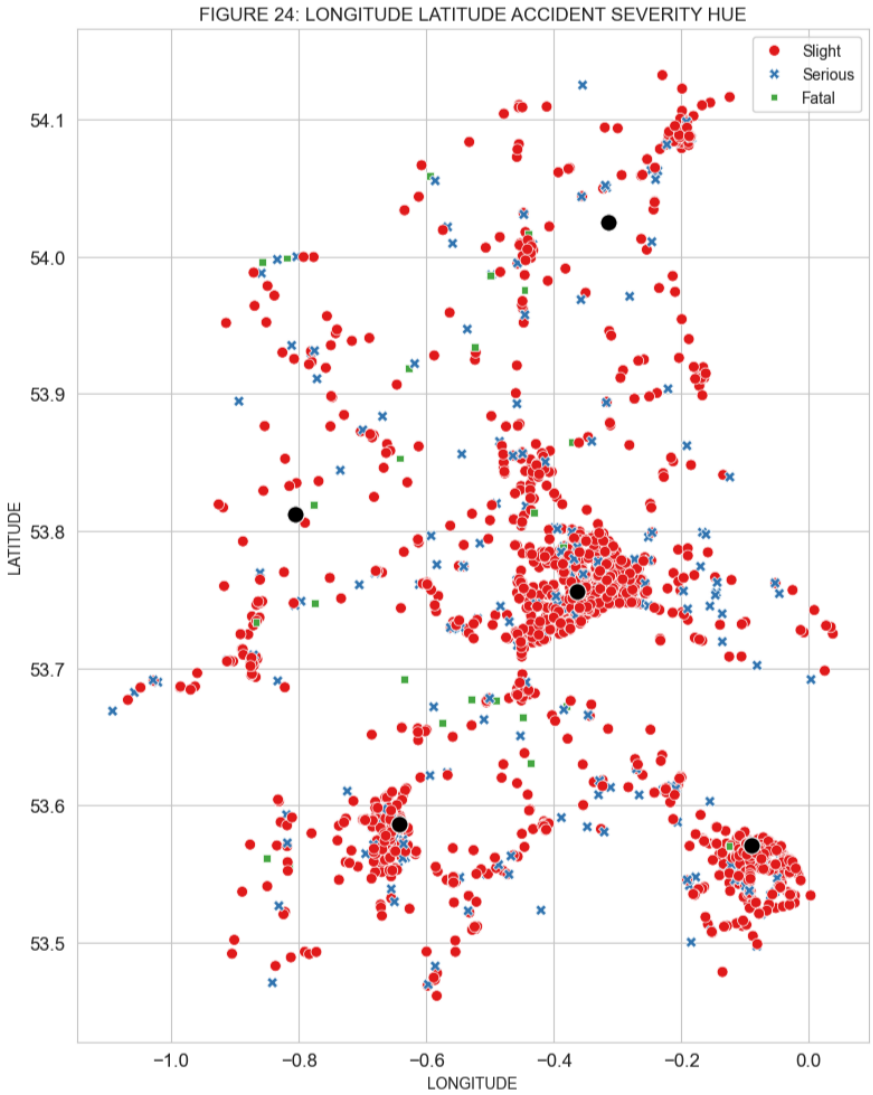


Figure 11: Severity Heatmap

Tasks 6 and 7: Forecasting Accident Trends

To build on the descriptive analysis, Holt-Winters exponential smoothing was used to forecast accident trends based on historical data (Hyndman and Athanasopoulos, 2018). In Task 6,

weekly accident counts were forecasted for three police regions: Metropolitan Police (1), Greater Manchester (6), and Cleveland (17). The model performed best in Cleveland, where patterns were more stable. It had the lowest mean absolute error (MAE: 4.26), followed by Greater Manchester (MAE: 11.10). The Metropolitan area had higher variability and the highest error (MAE: 38.73), likely due to complex traffic conditions.

In Task 7, daily forecasts were generated for the top 30 high-risk LSOAs in Hull, using data from January to June 2019 to predict accident patterns in July. The model captured weekday trends, particularly peaks on Mondays and Fridays, offering useful insight for planning local interventions.

Overall, the Holt-Winters method proved effective for short-term forecasting, especially in areas with regular traffic behaviour. These forecasts can help local authorities plan road safety initiatives and allocate emergency response more efficiently.

Predictions and Discussion

Task 6: Forecasting Weekly Accidents by Police Region

To evaluate the potential for short-term forecasting, weekly accident counts from 2017 and 2018 were used to predict 2019 trends across three selected police jurisdictions: Metropolitan (1), Greater Manchester (6), and Cleveland (17). The Holt-Winters Exponential Smoothing method was applied, as it effectively handles both seasonality and trend in time series data (Hyndman and Athanasopoulos, 2018).

The forecasting model performed with varying levels of accuracy depending on the region. Cleveland exhibited the best results, with a mean absolute error (MAE) of 4.26 and root mean square error (RMSE) of 5.34, reflecting its relatively stable traffic flow. Greater Manchester followed with a moderate MAE of 11.10 and RMSE of 13.10, while the Metropolitan Police area had the highest error rates (MAE = 38.73, RMSE = 54.39) due to higher variability and urban complexity. These results suggest that Holt-Winters is more reliable in regions with less fluctuation in weekly accident patterns. In highly urbanised areas like Metropolitan, additional factors such as roadworks, events, and dense traffic may reduce the model's effectiveness, highlighting a need for more complex or adaptive methods (Hyndman, 2010).

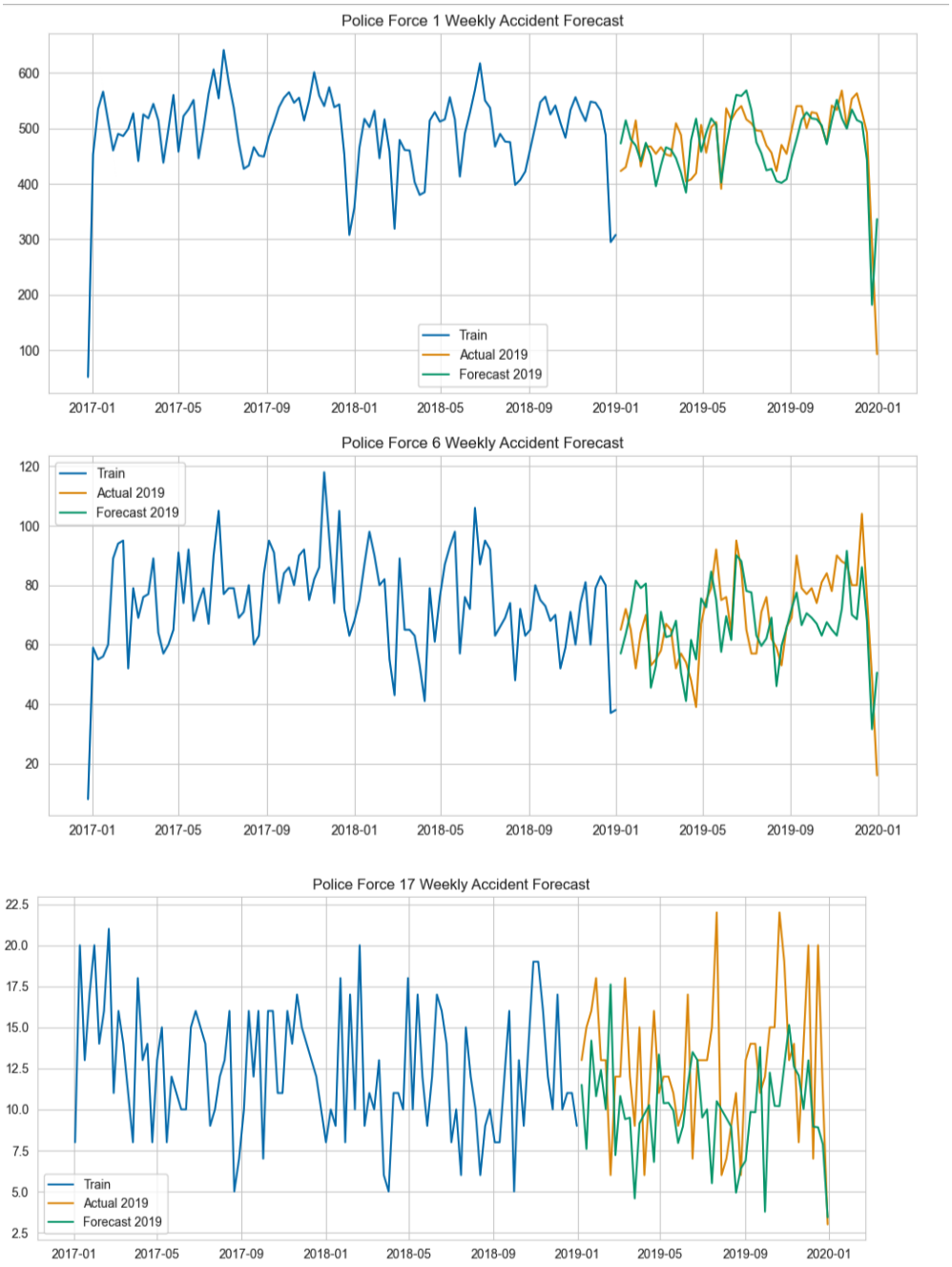


Figure 12: Forecasted vs Actual Weekly Accidents

Task 7: Daily Forecasting for High-Incident LSOAs

For the City of Hull, the top 30 Local Super Output Areas (LSOAs) with the highest number of accidents between January and March 2019 were identified. Their daily accident trends from January to June were used to forecast incident rates for July. Despite the inherent volatility of daily data, the Holt-Winters model captured clear patterns. The forecast aligned well with actual

data, especially highlighting peak incidents on Mondays and Fridays a trend consistent with earlier descriptive analyses. This reinforces the idea that even with smaller data windows, simple seasonal models can provide actionable insights for short-term planning.

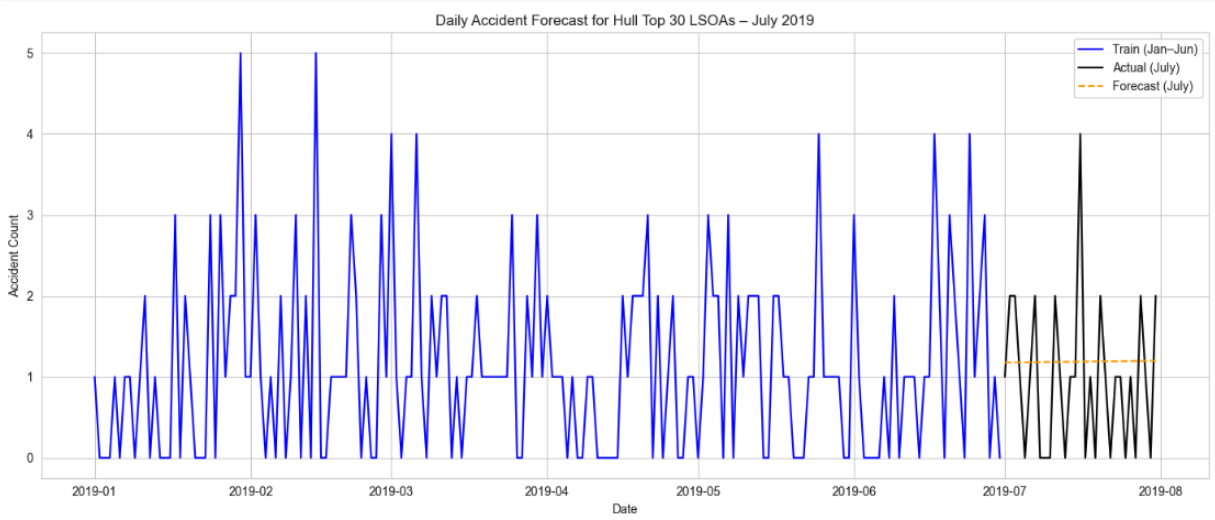


Figure 13: Daily Accident Forecast in High-Risk Hull LSOAs (July 2019)

Together, these forecasting tasks demonstrate the practical value of time series models in accident analysis. They show that while Holt-Winters may not capture every irregularity in high-density zones, it remains a useful tool for informing weekly and daily safety planning particularly in more consistent or localised environments.

Recommendations

Based on the analysis and predictions carried out in this project, the following recommendations are proposed:

- Prioritise High-Risk Time Periods : Accidents peak around 8 AM and 5 PM, especially on Thursdays and Fridays. These time slots should be targeted through road safety campaigns, traffic monitoring, and public alerts.
- Focus Urban Interventions on Hotspots Clustering analysis pinpointed dense accident zones in central Hull. Urban planners should prioritise traffic calming measures, clear signage, and better lighting in these areas.
- Enhance Motorcycle Safety by Type
    - Under 125cc bikes: High weekday usage suggests commuter focus.

    - Over 500cc bikes: Higher weekend accident rates point to leisure riders. Campaigns should be engine-size specific and address rider behaviors on different days.
- Improve Pedestrian Infrastructure Pedestrians aged 26–45, particularly males, were most frequently injured. Many incidents occurred while crossing roads without traffic controls. Authorities should install zebra crossings, barriers, enhance school zone safety and urban footpaths.
- Forecast-Informed Emergency Response Holt-Winters forecasting showed strong alignment in Cleveland, moderate accuracy in Greater Manchester, and higher error in Metropolitan due to variability. Forecast models should guide emergency resource planning in areas where patterns are stable, while more adaptive methods should be explored for complex urban zones.
- Investigate Outliers and Data Anomalies Outlier detection revealed unusual accident clusters. These should be flagged for manual review, as they may indicate roadwork impacts, event-related congestion, or data reporting issues.
- Promote Public Transparency and Education Share road accident heatmaps and trends with the public. Visualised data can help raise awareness and encourage safer behaviours in known danger zones.

These recommendations stem from data-driven insights and can support national and local agencies in forming adaptive, targeted, and cost-effective strategies to reduce traffic-related injuries and fatalities.

Reference

Bhandari, P. (2022) *How to Find Outliers | Meaning, Formula & Examples*. Scribbr. Available at: https://www.scribbr.co.uk/stats/statistical-outliers

Brownlee, J. (2014) *Feature selection to improve accuracy and decrease training time*. Machine Learning Mastery. Available at: https://machinelearningmastery.com/feature-selection-to-improve-accuracy-and-decrease-training-time/

DataTechNotes (n.d.) *SelectKBest feature selection example in Python*. Available at: https://www.datatechnotes.com/2021/02/seleckbest-feature-selection-example-in-python.html

Department for Transport (DfT) (2020) *Reported road casualties in Great Britain: Annual Report 2019*. GOV.UK. Available at: https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2019

Hyndman, R.J. and Athanasopoulos, G. (2018) *Forecasting: Principles and Practice*. 2nd edn. OTexts. Available at: https://otexts.com/fpp2/

Hyndman, R.J. (2010) *Initializing the Holt-Winters method*. Hyndsight. Available at: https://robjhyndman.com/hyndsight/hw-initialization/

RoSPA (2001) *Driver Fatigue and Road Accidents: A Literature Review and Position Paper.* Royal Society for the Prevention of Accidents. Available at: https://www.rospa.com/siteassets/images/road-safety/road-safety-projects/road-safety-observatory/drivers-driver-fatigue.pdf

TASK B

Introduction

This section explores the structure and behaviour of a real-world social network derived from the facebook_combined.txt dataset (Leskovec & Mcauley, 2012). The data contains anonymised edge pairs representing friendships among Facebook users. Using this edge list, a social network was constructed with NetworkX, and several key structural properties, centrality measures, and community detection techniques were applied to uncover insights about user connectivity and subgroups within the network.

Analysis and Results

1. Network Construction and Basic Properties

Using the provided edge list, we constructed an undirected graph where nodes represent users and edges represent friendship ties. The graph consists of 4,039 nodes and 88,234 edges. The average degree was approximately 43.69, indicating a moderately connected user base. The network density was 0.0108, typical for large sparse social networks, and the network formed a single connected component. The diameter of the largest connected component was 8, suggesting the classic "small-world" phenomenon in online social graphs (Newman, 2010).

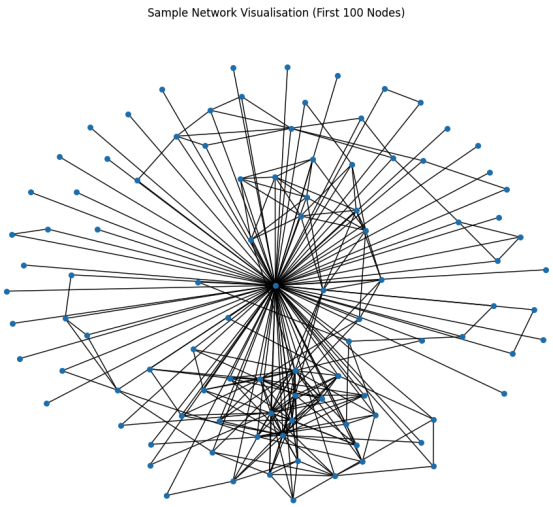


Figure 1: Sample Visualisation of the First 100 Nodes

2. Node and Edge Centrality Analysis

To assess influence and connectivity within the network, we analysed degree centrality and edge betweenness centrality.

- **Degree Centrality**: This identifies nodes with the highest number of connections. The distribution was heavily right-skewed, reflecting a small number of highly connected users (hubs) and many with few connections. This supports the scale-free nature of social networks (Barabási & Albert, 1999).

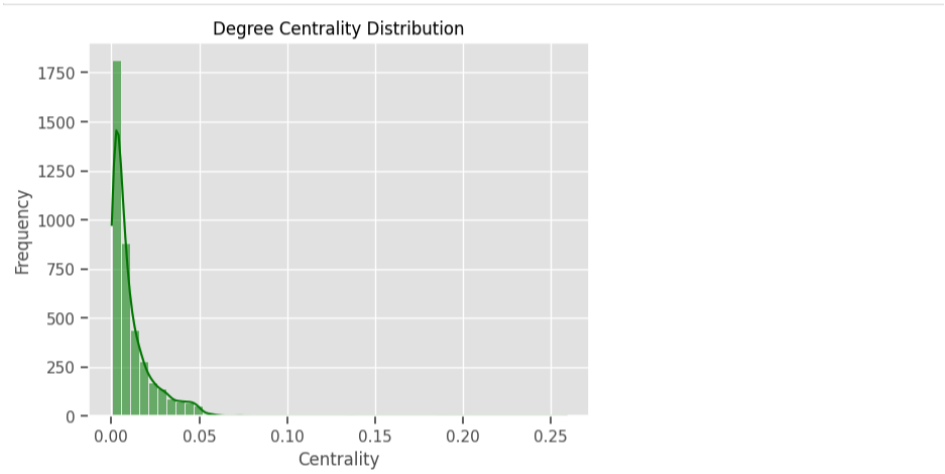


Figure 2: Degree Centrality Distribution

- **Edge Betweenness Centrality**: This metric highlights edges that serve as critical bridges between different parts of the network. The distribution showed that most edges had very low centrality, while a few carried significant inter-community influence. Such edges are crucial for understanding how information or influence may flow across the network (Girvan & Newman, 2002).

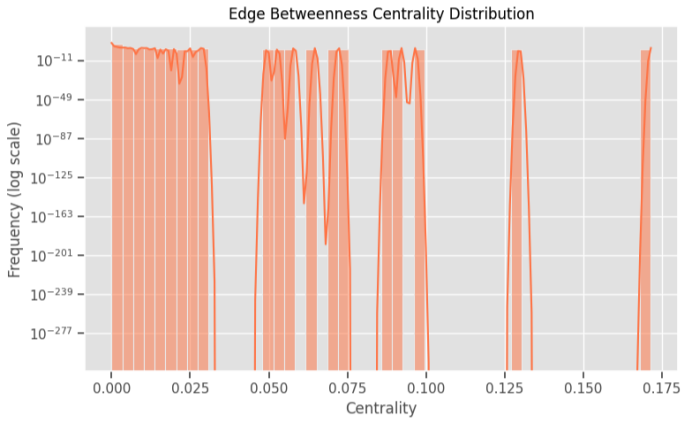


Figure 3: Edge Betweenness Centrality Distribution (Log Scale)

3. Community Detection

Two algorithms were applied to uncover community structures:

- **Louvain Method**: This modularity-based method found 16 communities within the graph. The largest five communities ranged from 423 to 548 nodes. It is widely used for its efficiency and accuracy in large networks (Blondel et al., 2008).

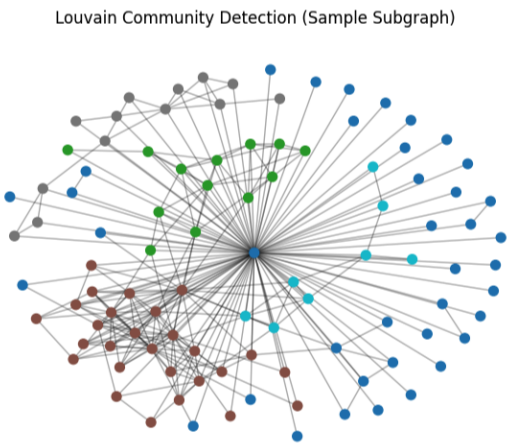


Figure 4: Louvain Community Detection on Subgraph (100 nodes)

- **Greedy Modularity**: Applied to a 1,000-node subgraph, this algorithm discovered 9 communities. Although not as granular as Louvain, it effectively grouped nodes based on modularity maximisation.

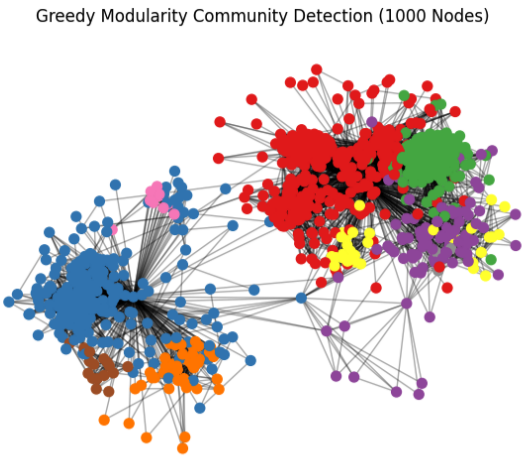


Figure 5: Greedy Modularity Community Detection (1,000 nodes)

These results confirm that the Facebook network exhibits strong community structure, a hallmark of real-world social networks.

Discussion

The analysis reveals key structural features of online social networks:

- **Connectivity**: The high average degree and short diameter reinforce the idea of a tightly-knit yet globally connected network. Users are likely to reach one another through only a few intermediaries.
- **Centrality Patterns**: The presence of hub nodes suggests influencers or central figures who may drive communication, trends, or influence. Meanwhile, edge centrality helps uncover critical links for inter-group connections.
- **Community Structure**: Both detection methods uncovered rich, modular substructures. This reflects real-world social groupings like friend circles, shared interests, or geographic proximity. Louvain was better at identifying more granular groupings, while greedy modularity performed well on medium-scale subgraphs.

Together, these findings provide a deeper understanding of the topological and social dynamics of online communities.

Conclusions

- The Facebook network forms a single, highly connected component with typical small-world characteristics.
- Degree centrality analysis highlighted the scale-free nature of the network, with a few central nodes.
- Edge centrality pinpointed key bridges between communities, essential for information flow.
- Louvain and greedy modularity algorithms successfully identified community structures, supporting modularity theory in social graphs.
- These insights have practical implications for understanding social influence, network resilience, and targeted outreach in digital platforms.

References

Barabási, A.-L. and Albert, R., 1999. Emergence of scaling in random networks. *Science*, 286(5439), pp.509–512. [online] Available at: https://doi.org/10.1126/science.286.5439.509

Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E., 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), p.P10008. [online] Available at: / https://doi.org/10.1088/1742-5468/2008/10/P10008

Girvan, M. and Newman, M.E.J., 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), pp.7821–7826. [online] Available at: https://doi.org/10.1073/pnas.122653799.

Leskovec, J. and McAuley, J.J., 2012. Learning to discover social circles in ego networks. *Advances in Neural Information Processing Systems (NeurIPS)*. [online] Available at: https://snap.stanford.edu/data/egonets-Facebook.html

Newman, M.E.J., 2010. *Networks: An Introduction*. Oxford: Oxford University Press. Available at: https://academic.oup.com/book/27303.