REPORT ON ACCIDENT DATA 2020

# SHORT INTRODUCTION OF THE DATASET

The dataset includes detailed information on UK road traffic accidents for 2020, with four tables: accident, casualty, lsoa (Lower Layer Super Output Area), and vehicle. It covers various attributes such as location, weather, road conditions, vehicle details, and casualties. The goal is to analyze these factors to identify accident causes, assess casualty severity, and explore the relationship between accidents and some other factors to improve road safety.

# Task 1:

Accidents by Hour of Day

The Figure 1 shows a peak in accidents between the 9th and 17th hour, coinciding with rush hour and increased traffic congestion. A secondary, smaller peak occurs around the 8th hour, likely due to the morning commute. Nighttime hours have the lowest accident rates due to reduced traffic. The highest number of accidents occurs at 17th hour with a count of 7,813.

A graph of a number of blue bars

Description automatically generated with medium confidence

**Figure 1: Accidents by Hour of the Day**

Accidents by Day of the Week

Figure 2shows that Friday has the highest number of accidents, followed by Thursday, likely due to increased weekend traffic. Weekend days (Saturday and Sunday) see fewer accidents, possibly due to lower traffic volumes. The highest count is on Friday with 14,889 accidents.

A graph of blue bars

Description automatically generated

**Figure 2 : Accidents by Day of week**.

**Multi-Factor Analysis (Day of week)**

CASE 1: 10 accidents by Local Authority and weather conditions 1 (Fine without high winds).

Figure 3 that Birmingham has the highest number of accidents, followed by Leeds and Westminster. Friday is the day with the most accidents, followed by Tuesday and Thursday.

A graph with different colored bars

Description automatically generated

**Figure 3 : Accidents by Local Authority and weather condition 1.**

CASE 2: 10 accidents by Local Authority and Light conditions 1 (Daylight).

The Figure 4shows that Birmingham has the most accidents, followed by Leeds and Lambeth. The highest number of accidents occurs on Friday, with Tuesday and Thursday also seeing significant numbers.

A graph of a bar chart

Description automatically generated with medium confidence

**Figure 4 : Accidents by Local Authority and Light condition 1**.

CASE 3: 10 accidents by Local Authority and Road Type 6 (Single carriageway).

Figure 5 shows that Birmingham, Cornwall, and Leeds have the highest number of accidents. The days with the most accidents vary by local authority.

A graph with different colored bars

Description automatically generated

**Figure 5 : Accidents by Local Authority and Road Type 6.**

**Multi-Factor Analysis (Hour of the day)**

CASE 1: Maximum accidents by hour of the day and weather conditions 1 (Fine without high winds).

The Figure 6 shows a peak in accidents from the 9th to 15th hour of the day, likely due to rush hour, with a smaller peak around 8th hour. Accidents drop significantly during nighttime hours.

A graph of blue bars

Description automatically generated

**Figure 6 : Maximum Accidents by Hour of Day and Weather Condition 1**.

CASE 2: Maximum accidents by hour of the day and light conditions 1 (Daylight) and light condition 4 (Darkness: streetlights present and lit).

Figure 7 shows a clear peak in accidents between 9th to 15th hour of the day, primarily under light conditions 1. Accident numbers decrease significantly during nighttime hours.

A graph of a graph

Description automatically generated with medium confidence

**Figure 7 : Maximum Accidents by Hour of Day and light Condition 1 and 4**.

CASE 3: Maximum accidents by hour of the day and road type 6 (Single carriageway).

Figure 8 shows a clear peak in accidents between 9th to 15th hour of the day, primarily for road type 6. Noticeable decrease in accidents after the 17th hour.

A graph of a number of bars

Description automatically generated with medium confidence

**Figure 8 : Maximum Accidents by Hour of Day and Road type 6.**

# Task 2:

Accidents by Hour of Day

Figure 9 shows that motorcycle accidents peak at 18th hour for 125 cc and at 17th hour for motorcycles over 500 cc and those between 125 cc and 500 cc. Motorbikes under 125 cc have the highest number of accidents overall, with distinct patterns by engine size during peak hours.

A graph of a number of vehicles

Description automatically generated

**Figure 9 : Accident by hour of day for Motorbikes**.

Accidents by Day of the Week

Motorcycle accidents are most frequent on Thursdays and Fridays across all engine sizes, while Sundays generally see the fewest accidents. Motorcycles 125cc and under are involved in more accidents compared to other engine sizes as shown in Figure 10.

A graph of a motorcycle accident

Description automatically generated

**Figure 10 : Accident by day of the week for Motorbikes.**

**Multi-Factor Analysis: By Hour of the Day**

CASE 1: By weather conditions for maximum hour of the day

Figure 11 reveals the most accidents occur in weather condition 1, followed by condition 2, with other conditions seeing significantly fewer accidents at 18th hour. Figure 12 and Figure 13 at 17th hour also show that weather condition 1 has the highest accident count, while conditions 2 and 6 are notably higher than others in Figure 12. Figure 13 indicates that conditions 2 and 7 have the second-highest accident counts.

A graph of a motorcycle

Description automatically generated

**Figure 11 : Accident for Motorcycle 125cc and under**.

A graph with blue squares

Description automatically generated

**Figure 12 : Accident for Motorcycle over 500cc**.

A graph of a motorcycle

Description automatically generated

**Figure 13 : Accident for Motorcycle 125cc and up to 500cc.**

**Multi-Factor Analysis: By Day of Week**

CASE 1: By weather conditions for maximum day of the week

Figure 14 and Figure 15 show that most accidents for Motorcycle 125cc and under and Motorcycle 125cc and up to 500 occur in weather condition 1 on Fridays, followed by weather condition 2. Figure 16 shows more accidents for motorcycles over 500cc occur on Sundays under weather condition 1.

A graph of a motorcycle

Description automatically generated

**Figure 14: Accident on Friday for Motorcycle 125cc and under.**

A graph with blue squares

Description automatically generated

**Figure 15: Accident on Friday for Motorcycle 125cc and up to 500**.

A graph of accident on a white background

Description automatically generated

**Figure 16: Accident on Friday for Motorcycle over 500cc**.

**Task 3:**

Accidents by Hour of Day

Figure 17 shows a significant increase in pedestrian accidents between 9th to 15th hour, with a smaller peak around 8th hour, likely due to morning commutes. Accidents drop significantly during nighttime hours (midnight to 5th hour).

A graph of a number of people

Description automatically generated with medium confidence

**Figure 17 : Accidents in hour of day for pedestrian**.

Accidents by Day of the Week

Pedestrian accident numbers are relatively consistent from Monday to Thursday, with a significant peak on Friday. There is a noticeable decrease in accidents on both Saturday and Sunday as shown in Figure18.

A graph of accident count

Description automatically generated

**Figure 18 : Accidents in hour of day for pedestrian.**

**Multi-Factor Analysis: By 15th hour of the day.**

CASE 1: By weather conditions for maximum hour of the day

Figure 19 shows most pedestrian accidents occur in weather condition 1, but there's a notable increase under conditions 2 and 9, indicating these weather types may increase accident risk.

**A graph with blue squares

Description automatically generated**

**Figure 19 : Accident count by weather conditions for pedestrian**.

CASE 2: By Road type for maximum hour of the day

Road type 6 experiences the highest number of pedestrian accidents at 15th hour. In contrast, road types 1 and 7, have significantly lower pedestrian accident rates as shown in Figure 20.

**A graph of accident count

Description automatically generated**

**Figure 20: Accident count by road type for pedestrian.**

**Multi-Factor Analysis: By day of the week (Friday)**

CASE 1: By weather conditions for maximum day of the week

Figure 21 shows that most pedestrian accidents on Fridays occur in weather condition 1, with conditions 2 and 9 having slightly higher accident counts compared to other conditions.

A graph with blue squares

Description automatically generated

**Figure 21: Accident count on Friday by weather conditions**.

CASE 2: By road type for maximum day of the week

Figure 22 shows that road type 6 accounts for a significantly higher number of pedestrian accidents compared to all other road types.

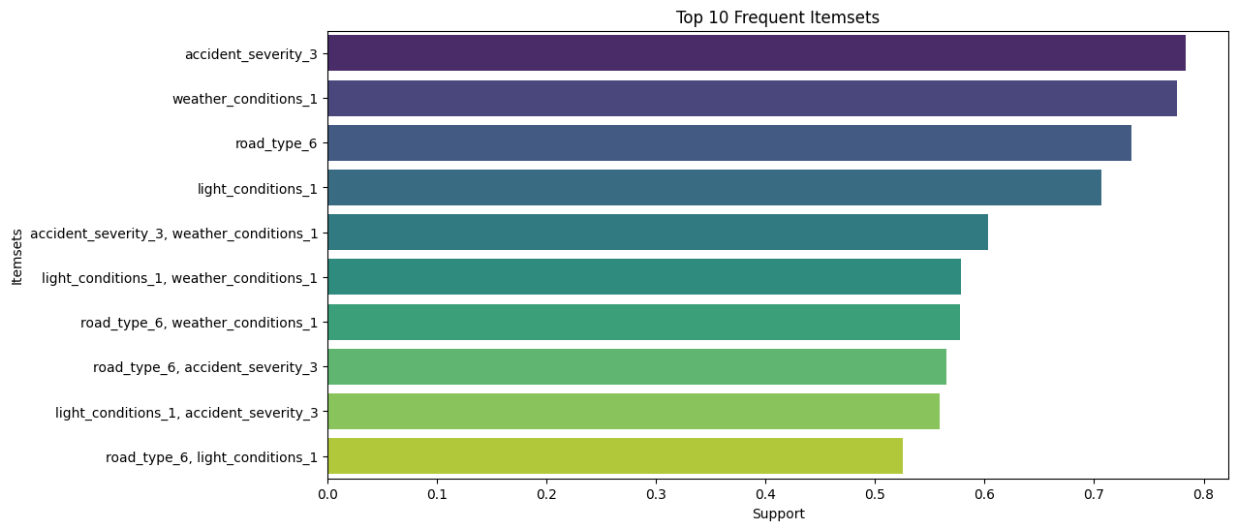
A graph of a road type

Description automatically generated

**Figure 22: Accident count on Friday by road type.**

**Task 4:**

The Apriori algorithm is used in data mining to find frequent itemset and generate association rules. Applied to selected columns of the accidents\_2020 Data Frame (like accident severity, road type, and weather conditions), it identifies frequent itemset with a minimum support of 0.01, as shown in Figure 23.



**Figure 23 : Top 10 Itemset with their support.**

The algorithm then generates and filters association rules based on accident severity using a minimum lift threshold of 0.2, displaying key metrics like support, confidence, and lift, as shown in Figure 24, Figure 25 and Figure 26.

A screenshot of a computer

Description automatically generated

**Figure 24 : Top 5 Association rules for Accident severity 3.**

A screenshot of a computer

Description automatically generated

**Figure 25 : Top 5 Association rules for Accident severity 2.**

A screenshot of a computer

Description automatically generated

**Figure 26 : Association rules for Accident severity 1.**

**Task 5:**

To cluster accident data for Humberside, the Elbow method with KMeans identifies 5 as the optimal number of clusters, shown in Figure 27. Clustering with k=5 yields a Silhouette Score of 0.65 and a Davies-Bouldin Score of 0.57, indicating reasonably well-defined clusters.

A graph of a curve

Description automatically generated

**Figure 27 : Elbow Method.**

Figure 28 displays the spatial distribution, with Cluster 1 being the largest and most concentrated, while Clusters 0 and 2 are sizable but more spread out. Clusters 3 and 4 are smaller, indicating less populated or specific accident locations.

A diagram of a number of objects

Description automatically generated with medium confidence

**Figure 28 : Clusters of Accidents.**

**Task 6:**

The Table 1 compares the performance of ARIMA and SARIMA models for three police forces, measured by Mean Squared Error (MSE). SARIMA models consistently outperform ARIMA models, with the lowest MSE achieved by Police Force 46 using SARIMA.

|  |  |  |  |
| --- | --- | --- | --- |
| Model/Police Force | Police Force 1 | Police Force 20 | Police Force 46 |
| Arima (MSE) | 8269.23 | 548.87 | 325.72 |
| Sarima (MSE) | 698.12 | 56.76 | 54.55 |

**Table 1 : MSE value of ARIMA and SARIMA Model.**

The Figure 29 show weekly accident count forecasts for two police forces from 2017 to 2021. Police Force 1 has a higher MSE, suggesting a less accurate forecast. Forecasted drops in accidents might indicate specific events or changes. Police Force 46 has a lower MSE, indicating better model fit as shown in Figure 30

**A graph of a graph of a person

Description automatically generated with medium confidence**

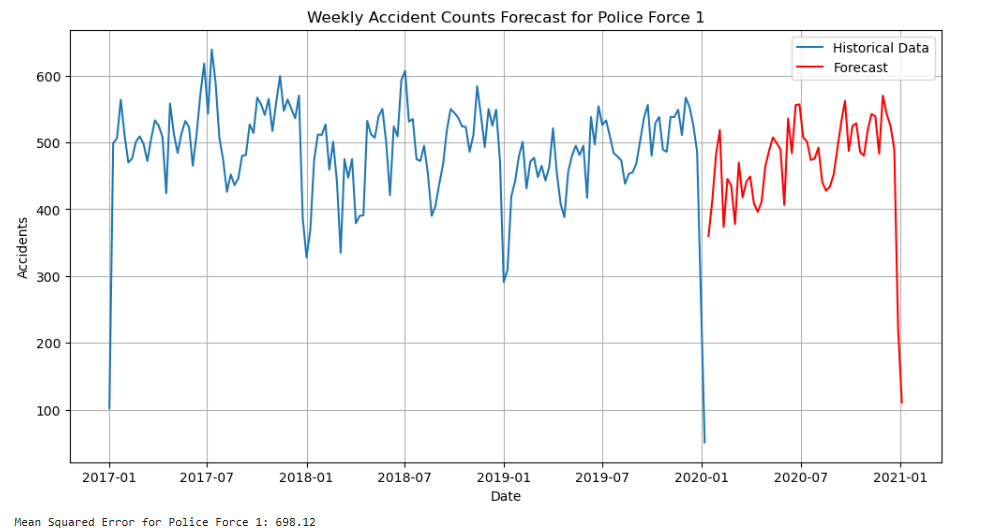
**Figure 29 : Weekly Accident Forecast (ARIMA) for police force 1 and 20.**

A graph of a crash

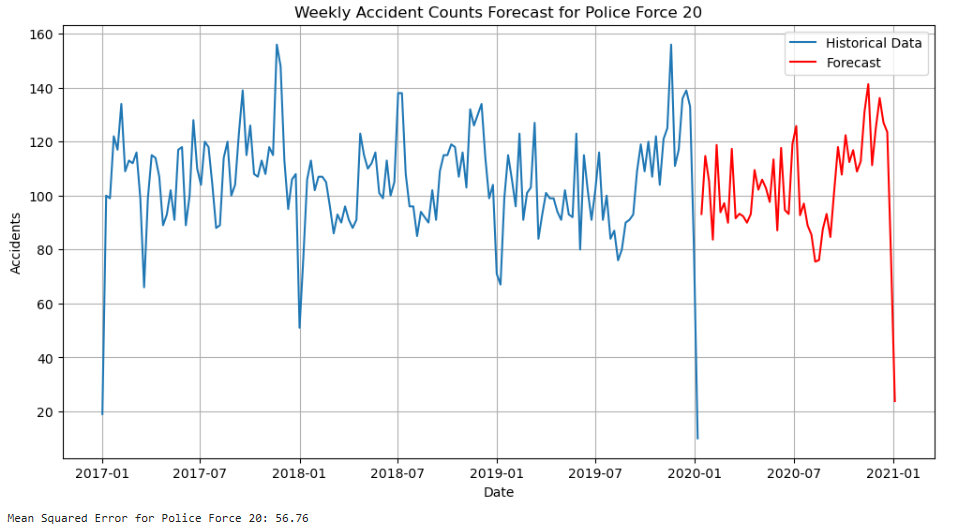
Description automatically generated

**Figure 30 : Weekly Accidents Forecast (ARIMA) for police force 46.**

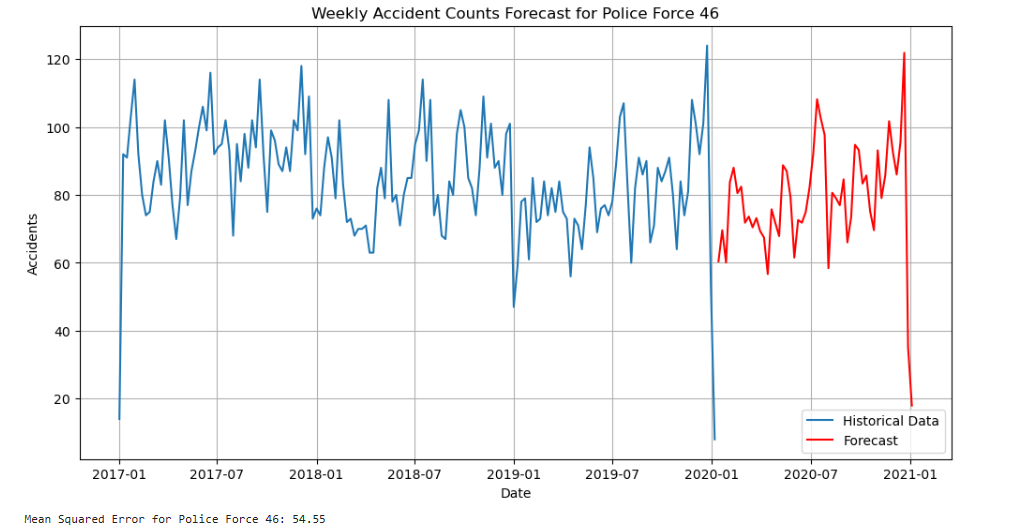
The forecasts for Police Forces 20 and 46 are more accurate as shown in Figure 32 and Figure 33 than the forecast for Police Force 1 as shown in Figure 31. Police Force 1's forecast shows a general upward trend but with a downward slope towards the end, and its high MSE indicates a less accurate model. In contrast, Police Forces 20 and 46 both demonstrate a close alignment with historical data, suggesting effective model fits and low MSE values, indicative of more accurate predictions.



**Figure 31 : Weekly Accidents Forecast (SARIMA) for police force 1.**



**Figure 32 : Weekly Accidents Forecast (SARIMA) for police force 20.**



**Figure 33 : Weekly Accidents Forecast (SARIMA) for police force 46.**

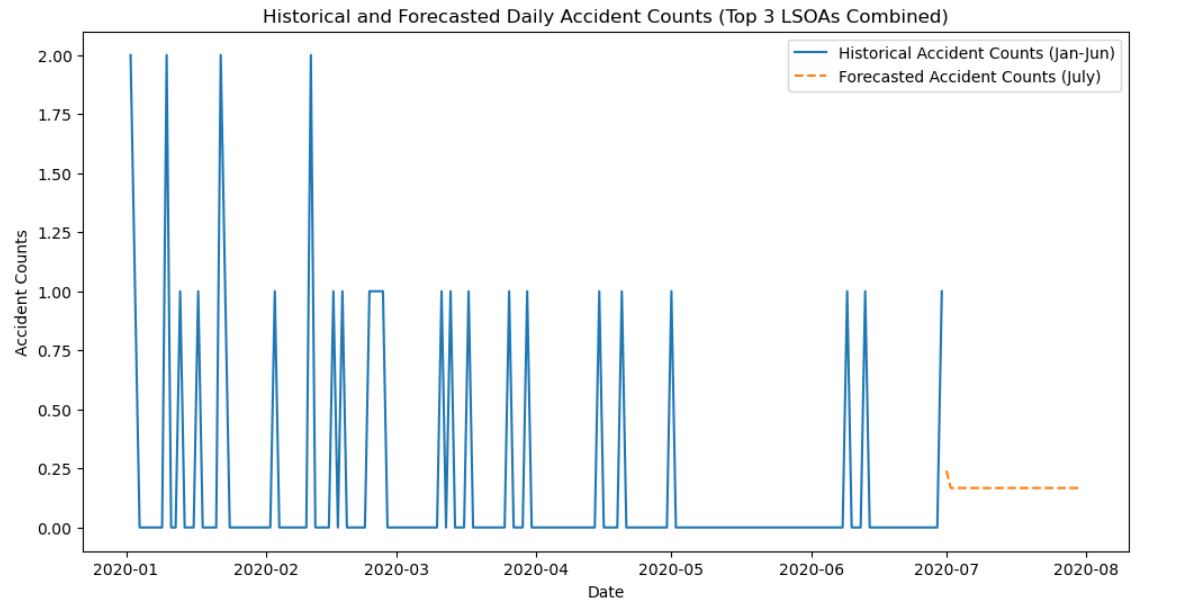
**Task 7:**

The table shows the Mean Squared Error (MSE) values for ARIMA and SARIMA models applied to historical and daily accident counts in the top 3 LSOAs. ARIMA models consistently outperform SARIMA models, with lower MSE values indicating better predictive accuracy.

|  |  |
| --- | --- |
| Model | MSE value |
| Arima | 0.18 |
| Sarima | 0.34 |

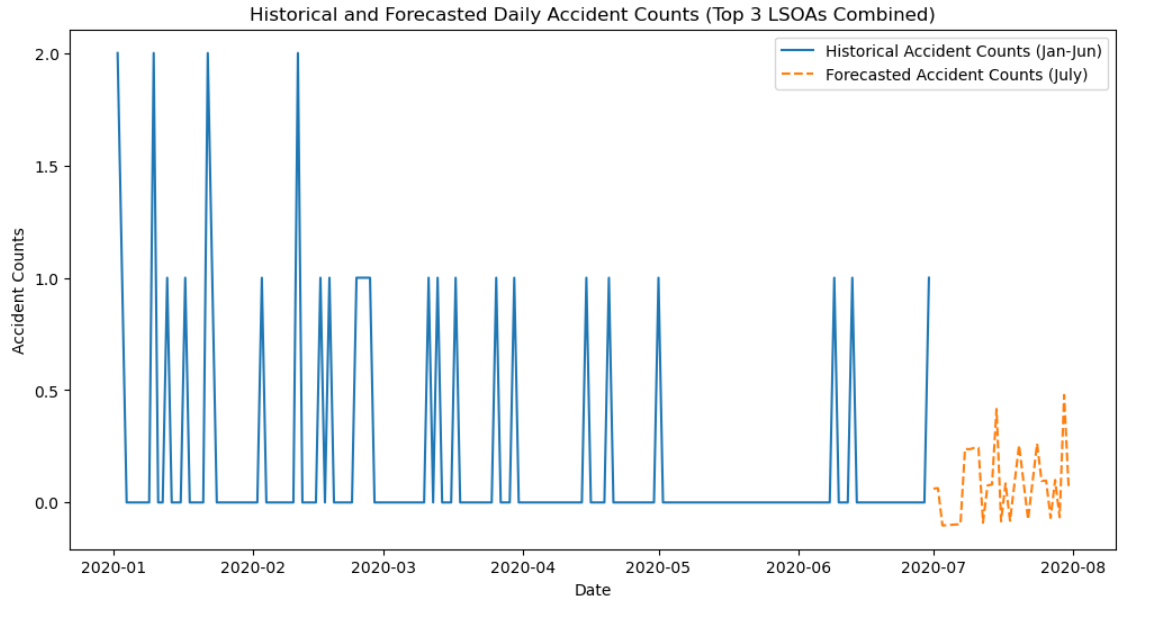
**Table 2 : MSE value of ARIMA and SARIMA model.**

The Figure 34 shows a decrease in forecasted accident counts for July 2020 compared to historical data. The historical data shows a consistent pattern with some fluctuations, while the forecast suggests a more significant drop in accidents for July.



**Figure 34 : Accident count of top 3 LSOAs (ARIMA model).**

The Figure 35 shows a significant decrease in forecasted accident counts for July 2020 compared to the historical average. The historical data shows a consistent pattern with some fluctuations, while the forecast suggests a more significant drop in accidents for July. This could be attributed to various factors such as changes in traffic patterns, weather conditions, or public health measures.



**Figure 35 : Accident count of top 3 LSOAs (SARIMA model)**

**Task 8:**

The analysis of a social network graph, likely from Facebook, involves 4,039 users and 88,234 friendships, resulting in a sparse network with a density of 0.0108. Each user has an average of 44 friends, indicating a large but moderately connected network.

Figure 36 visualizes a subgraph of 100 users from a larger social network. It features a central hub with many connections, surrounded by nodes with varying connectivity. Some nodes link directly to the hub, while others form smaller, tightly knit clusters, creating a mix of star-like and clustered patterns.

A network with blue dots and lines

Description automatically generated

**Figure 36 : Social Network Subgraph with 100 Nodes.**

**Task 9:**

Figure 37 shows Edge Betweenness Centrality, where most edges have low centrality, indicating they're not critical for network connectivity. However, a few edges with high centrality serve as key bridges, maintaining network cohesion.

A graph with a number of bars

Description automatically generated with medium confidence

**Figure 37 : Edge Centrality of the network.**

The top 5 edges with the highest betweenness centrality are crucial connectors in the social network. Edges (107, 1684) and (107, 1085) are particularly influential, followed by (1085, 3437), (567, 3437), and (0, 107):

A number on a white background

Description automatically generated

**Task 10: Two community detection algorithms**

**Label Propagation Algorithm (LPA)** detects communities by iteratively assigning labels to nodes based on their neighbors, converging to stable communities. It's fast and scalable but can yield varying results due to random initialization.

**The Louvain Method** optimizes modularity by iteratively grouping nodes into local communities and merging them into larger ones until modularity no longer improves, effectively identifying high-quality community structures.

**Clusters/community within this social network -** Both Label Propagation Algorithm (LPA) and Louvain Method were applied to detect communities in the network. LPA identified 44 communities, with sizes ranging from 2 to 1030 nodes, while Louvain Method detected 16 larger and more cohesive communities.

**Comparison Chart**

The two graphs shown in the Figure 38 compare both methods.

|  |  |  |
| --- | --- | --- |
| Attributes | LPA Community Sizes | Louvain Community Sizes |
| Distribution | LPA creates skewed communities with one large and many small groups. | Louvain creates evenly sized communities |
| Range | Community sizes vary widely, with many small communities. | Community sizes are more balanced, with most communities being of mid-range size. |
| Interpretation | LPA finds few large and many small communities. | Louvain finds fewer, but more balanced communities., focusing on optimizing modularity. |

A green and blue graph

Description automatically generated with medium confidence

**Figure 38: LPA and Louvain Graph.**

**Recommendations**

The following recommendations can be made to government agencies to improve road safety:

1. Enhance rush hour management by optimizing signals and improving public transport (de Souza et al*.*, 2017 ; Chelugo, 2017).
2. Run targeted road and traffic safety campaigns and increase patrols on Fridays to address high accident rates (Schrock et al., 2017).
3. Prioritize infrastructure upgrades in high-accident areas, especially single carriageways, to reduce accidents. (Gitelman et al., 2014)
4. Develop safety measures for adverse weather, including wet and foggy conditions, as well as fine weather with no high winds (Mohammed et al*.*, 2020).
5. Implement safety measures in pedestrian zones, especially in peak hours, and in high-accident areas (Gitelman et al*.*, 2012).

**References**

Chelugo, N.C. (2017) *Effects of Road Improvement on Safety:A Case study of Nairobi – Thika Superhighway, Kenya*. Thesis. COETEC, JKUAT. Available at: http://localhost/xmlui/handle/123456789/3280 (Accessed: 14 August 2024).

Gitelman, V. *et al.* (2012) ‘Characterization of pedestrian accidents and an examination of infrastructure measures to improve pedestrian safety in Israel’, *Accident Analysis & Prevention*, 44(1), pp. 63–73. Available at: https://doi.org/10.1016/j.aap.2010.11.017.

Gitelman, V., Carmel, R. and Pesahov, F. (2014) ‘The evaluation of safety efficiency of non-urban infrastructure improvements; a case-study’, *European Transport Research Review*, 6(4), pp. 477–491. Available at: https://doi.org/10.1007/s12544-014-0145-1.

Mohammed, A.S. *et al.* (2020) ‘The Perception System of Intelligent Ground Vehicles in All Weather Conditions: A Systematic Literature Review’, *Sensors*, 20(22), p. 6532. Available at: https://doi.org/10.3390/s20226532.

Schrock, S.D. *et al.* (2017) *Improving the traffic safety culture in Kansas.* K-TRAN: KU-15-2. Available at: https://rosap.ntl.bts.gov/view/dot/31752 (Accessed: 14 August 2024).

de Souza, A.M. *et al.* (2017) ‘Traffic management systems: A classification, review, challenges, and future perspectives’, *International Journal of Distributed Sensor Networks*, 13(4), p. 1550147716683612. Available at: https://doi.org/10.1177/1550147716683612.