FINAL REPORT ON ACCIDENT DATA

INTRODUCTION

Road accidents pose a significant threat to public safety and require thorough analysis to understand the underlying patterns and risk factors. The goal of this report is to analyze road accident data in the UK for the year 2020 to gain insights into when, where, and under what conditions road traffic accidents occur in the UK. The report will also develop a model to predict whether an accident is fatal given the conditions it happened under. Finally, the report will make recommendations to government agencies based on the data and analysis to improve road safety.

DATA PREPROCESSING

To ensure data integrity, the dataset was thoroughly checked for preprocessing. The data contained four tables in the database schema namely, Accident, Casualty, Vehicle and LSOA (Lower Layer Super Output Area) and each of these tables has various columns.

Each table was checked for missing (NaN) values and all tables were clean with no missing values in exception of the "Accident" table which had 14 missing values each in 4 different columns namely**, location_easting_osgr, location_northing_osgr, longitude and latitude.** The missing values were appropriately imputed using **Mean** as the measure of central tendency for probability distribution.

Outliers were kept since analysis revealed they were proportionally insignificant and would not pose a distortion in the analysis.

Columns were selected based on their relevance to the objectives of the report. 316 rows with -1 as value under **road_surface_conditions** column were replaced with "**Unknown**" since there is enough representation of the class of conditions needed for analysis.

EXPLORATORY DATA ANALYSIS

Analysis was carried out by examining the data to identify significant accident related patterns. Visualizations, such as histograms, heatmaps, and clusters, provided a clear representation of spatial and temporal patterns of accidents. These findings provide valuable insights for implementing targeted road safety measures and optimizing traffic management during high-risk hours and days.

KEY FINDINGS

In Fig.1 below, the highest number of accidents in 2020 occurred on Fridays with a total of 14,889 followed by Thursday with 14,056. Fridays could record higher numbers since a lot of individuals during this period move about in preparation for the weekend. Saturdays and Sundays which are weekends tend to record lower numbers because a lot of people (workers and students) spend their weekends at home. Averagely, a significant number of accidents occurred throughout the week ranging between 10,000 and 15,000 with Friday recording slightly higher number.

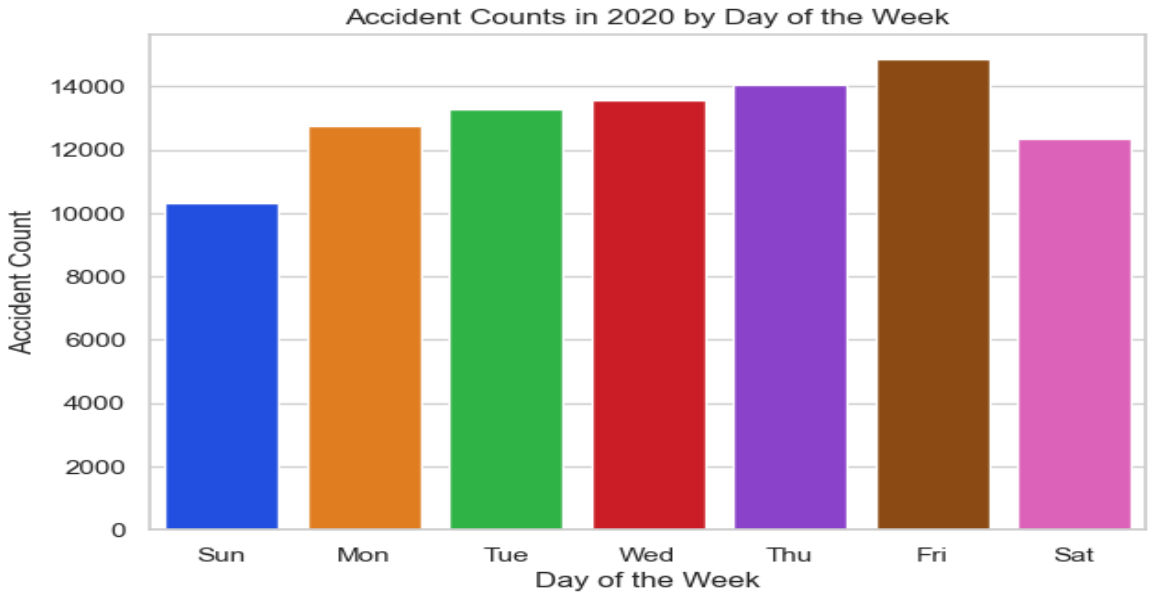


Fig. 1: Number of Accidents by each day of the week

Figure 2 below is a heatmap representing the hours of the day juxtaposed against days of the week on which significant number of accidents occur. Accidents occurred at about 8:00 am, 15:00 and 18:00 hours from Mondays to Fridays. This is a period of rush hour where a lot of commuters (workers or students) are either going to or returning from work or school.

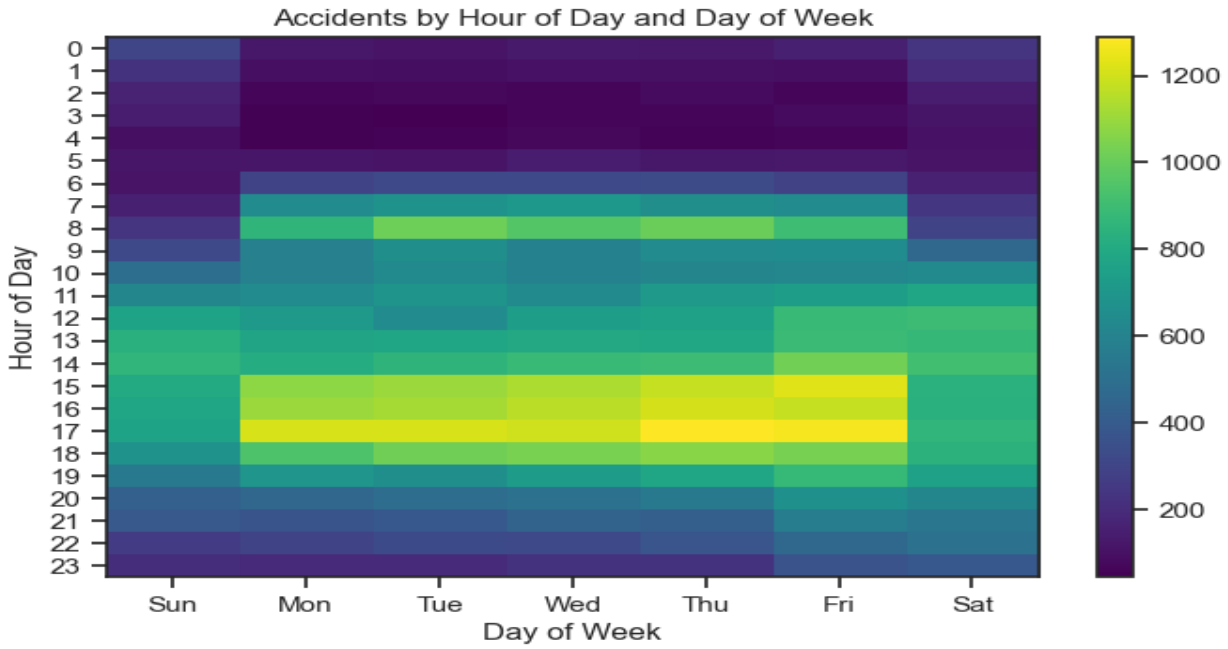
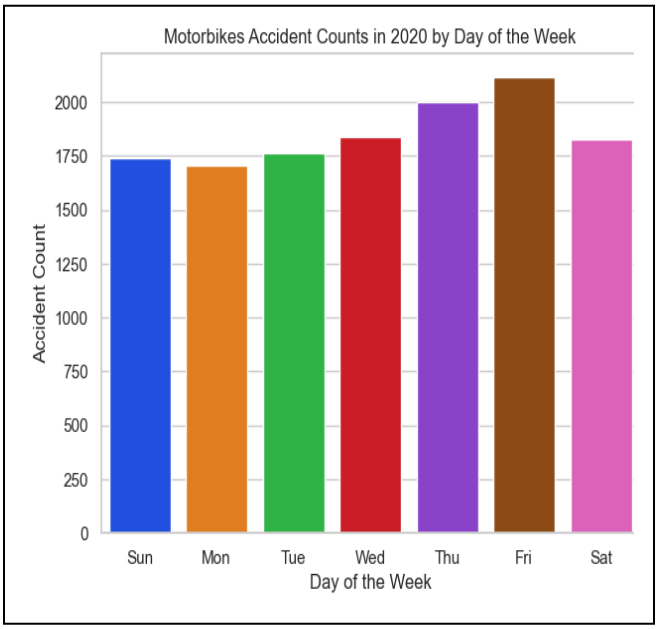


Fig. 2: Heatmap showing Hours of Day and Days of Week

Motorbike Related Accidents

For this analysis, motorbikes considered were Motorcycle 125cc and under, Motorcycle over 125cc and up to 500cc, and Motorcycle over 500cc. By analyzing significant hours and days of the week when motorbike accidents are more likely to occur, valuable insights revealed are discussed below.

Fig. 3 below shows averagely, a significant number of accidents occurred throughout the week in the year 2020 with slightly higher numbers recorded for Thursdays and Fridays as the highest but not too significant. From Table 1, Friday (day 6) recorded slightly higher percentage of 2.32% whiles Sunday (day 1) recorded the lowest 1.91%. It is therefore evident a significant motorbike related accidents were recorded on each day even though Friday is slightly higher.



	day_of_week	accident_count	percentage_motorcycle
0	1	1738	1.905723
1	2	1703	1.867345
2	3	1765	1.935328
3	4	1841	2.018662
4	5	2002	2.195200
5	6	2119	2.323490
6	7	1830	2.006601

Fig. 3: Motorbike accidents by Day of the Week

Table 1. Percentage of Motorbike accidents by Day of Week

Further Analysis on Motorbike related Accidents.

Fig 4 shows motorbike accidents are low from 00.00 to 05:00 hours where individuals are deemed to be asleep. The number rises from 06:00 and 07:00 hours when individuals begin to commute to work or school. The significant number of accidents occurred within the day from 10:00 hours with a steady rise till 16:00 and then sees a sharp rise at 17:00 hours which is considered a rush hour where individuals have either closed from work or school and are returning home.

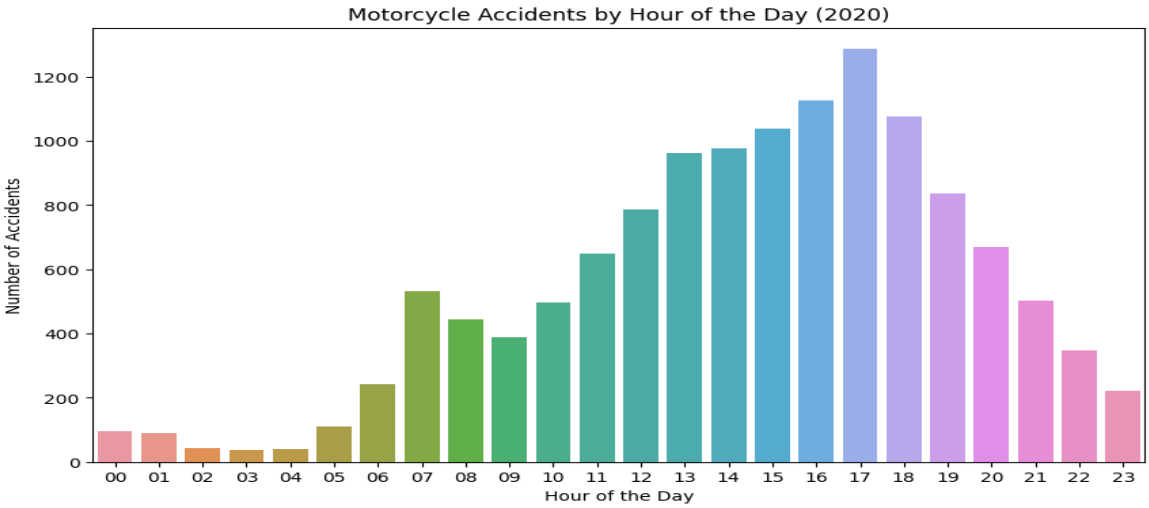


Fig. 4: Motorcycle Accidents by Hour of the day

Pedestrian related accidents

Further analysis into pedestrian accidents identified significant hours and days of the week when pedestrians are more vulnerable to accidents. Understanding peak times for pedestrian accidents is essential for implementing measures to improve pedestrian safety and reduce accidents involving pedestrians.

In Fig. 5 below, pedestrian related accidents are relatively low between 00:00 and 06:00 hours. The number rises at 07:00 hours and jumps to higher levels at 08:00 hours. The numbers are significantly high at 15:00 hours but begins to steadily fall from 16:00 hours after it peaks at 15:00. The higher numbers recorded for 08:00 and 15:00 hours is due to rush hours for commuters going to work or school or returning to their homes within those hours respectively.

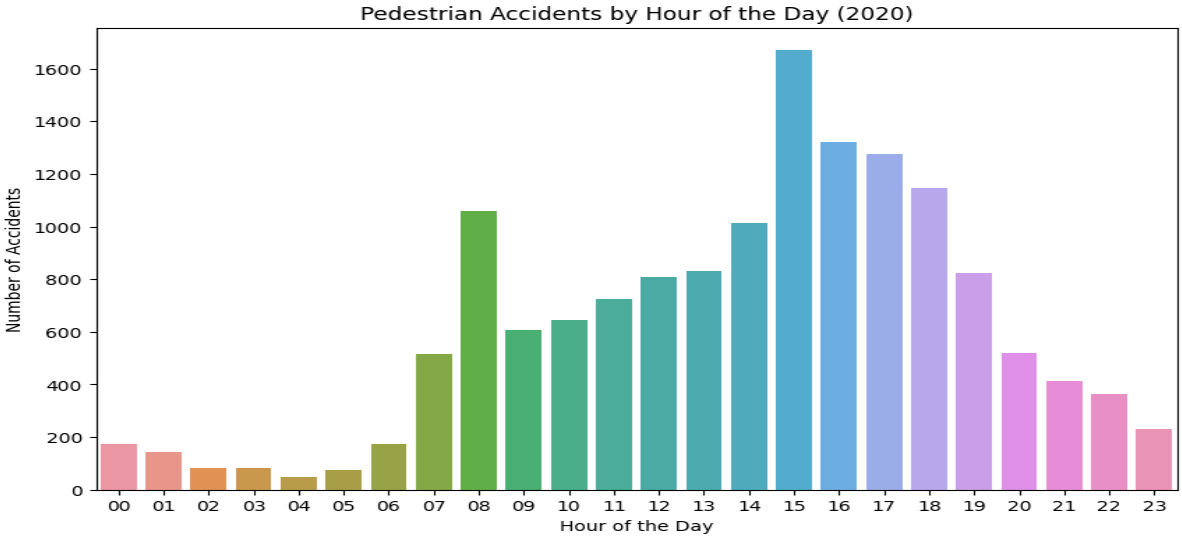


Fig. 5: Pedestrian related accidents by Hour of Day

As shown in Fig. 6 and Table 2 below, on the average, a significant number of pedestrian related accidents occurred throughout the week. Friday recorded the highest number of 2543 representing 17.24% as shown in Table 2. The lowest numbers, Sunday 1242 and Saturday 1878 with 8.42% and 12.73% respectively are obviously weekends where most commuters are at home.

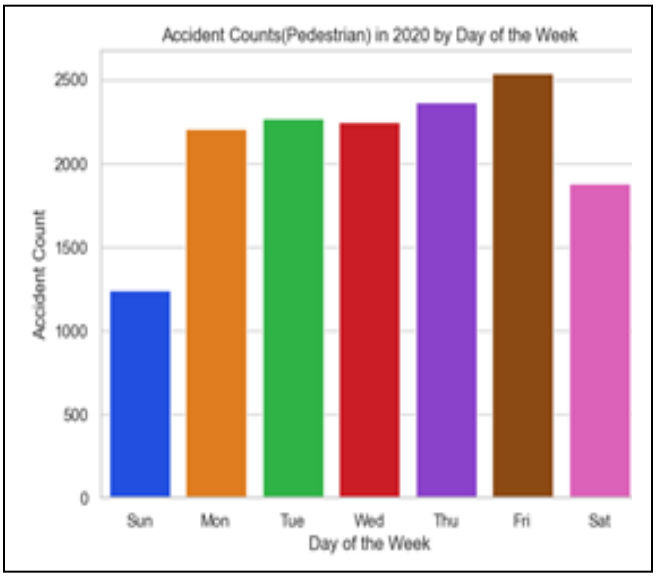


Figure 6: Pedestrian accidents by Day of Week

	day_of_week	pedestrian_count	percentage_pedestrians
0	1	1242	8.420339
1	2	2207	14.962712
2	3	2267	15.369492
3	4	2247	15.233898
4	5	2366	16.040678
5	6	2543	17.240678
6	7	1878	12.732203

Table 2. Percentage (Pedestrian) accidents by Day of Week

IMPACT OF SELECTED VARIABLES ON ACCIDENT SEVERITY

Leveraging the Apriori algorithm data mining technique to discover patterns and associations that significantly influence accident severity. Variables chosen for this analysis were Weather conditions, Accident severity and Road surface conditions.

Some Rules discussed

Rule 1 in Table 3 below indicates that, if the road is dry (road_1), then there is a high probability of slight accident severity (severity_3). The support of this rule is 0.687486, which means that it appears in 68.74% of the data set. The confidence of 0.783039, indicates if the road is dry (road_1), then there is a 78.30% probability that severity is going to be less serious.

Rule 11 also indicates that, if the road is dry (road_1) and the weather is fine without high winds (weather_1), then there is a high probability the accident would be less sever (severity_3). The support of this rule is 0.503399, which means that it appears in 50.34% of the data set. The confidence of 0.780155, indicates if the road is dry (road_1), and the weather is fine without high winds (weather_1), then there is a 78.02% probability that severity is going to less serious.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(road_1)	(weather_1)	0.687493	0.775554	0.645255	0.938563	1.210183	0.112067	3.653244	0.555760
1	(weather_1)	(road_1)	0.775554	0.687493	0.645255	0.831993	1.210183	0.112067	1.860077	0.773812
2	(road_1)	(severity_3)	0.687493	0.783493	0.538334	0.783039	0.999421	-0.000312	0.997909	-0.001851
3	(weather_2)	(road_2)	0.127009	0.287726	0.122382	0.963567	3.348910	0.085838	19.550413	0.803440
4	(road_2)	(severity_3)	0.287726	0.783493	0.224336	0.779687	0.995143	-0.001095	0.982727	-0.006806
5	(severity_2)	(weather_1)	0.201265	0.775554	0.160004	0.794988	1.025058	0.003911	1.094792	0.030605
6	(weather_1)	(severity_3)	0.775554	0.783493	0.603193	0.777757	0.992679	-0.004448	0.974192	-0.031812
7	(severity_3)	(weather_1)	0.783493	0.775554	0.603193	0.769877	0.992679	-0.004448	0.975328	-0.032940
8	(weather_2)	(severity_3)	0.127009	0.783493	0.101724	0.800915	1.022236	0.002213	1.087511	0.024917
9	(road_1, severity_2)	(weather_1)	0.139027	0.775554	0.132086	0.950075	1.225027	0.024263	4.495651	0.213353
10	(severity_2, weather_1)	(road_1)	0.160004	0.687493	0.132086	0.825521	1.200770	0.022085	1.791084	0.199049
11	(road_1, weather_1)	(severity_3)	0.645255	0.783493	0.503399	0.780155	0.995739	-0.002154	0.984816	-0.011918
12	(road_1, severity_3)	(weather_1)	0.538334	0.775554	0.503399	0.935105	1.205725	0.085892	3.458620	0.369583
13	(weather_1, severity_3)	(road_1)	0.603193	0.687493	0.503399	0.834557	1.213914	0.088708	1.888914	0.444091
14	(road_1)	(weather_1, severity_3)	0.687493	0.603193	0.503399	0.732224	1.213914	0.088708	1.481863	0.563886

Table 3. Association rules (Impact on Accident severity)

Geographical Clustering

To gain insights into the distribution of accidents in specific regions, emphasis was placed on accidents in Kingston upon Hull, Humberside, and East Riding of Yorkshire. By employing clustering techniques, identified various clusters or hotspots of accidents in the region. This information can inform targeted safety interventions and resource allocation to reduce accidents in high-risk areas.

Choice of Number of Clusters

The choice of number of clusters was informed by the use of the Elbow method given a range of 1 to 25. Figure 7 below revealed 5 number of clusters at the elbow hence the choice of 5 to perform clustering on regions earlier stated.

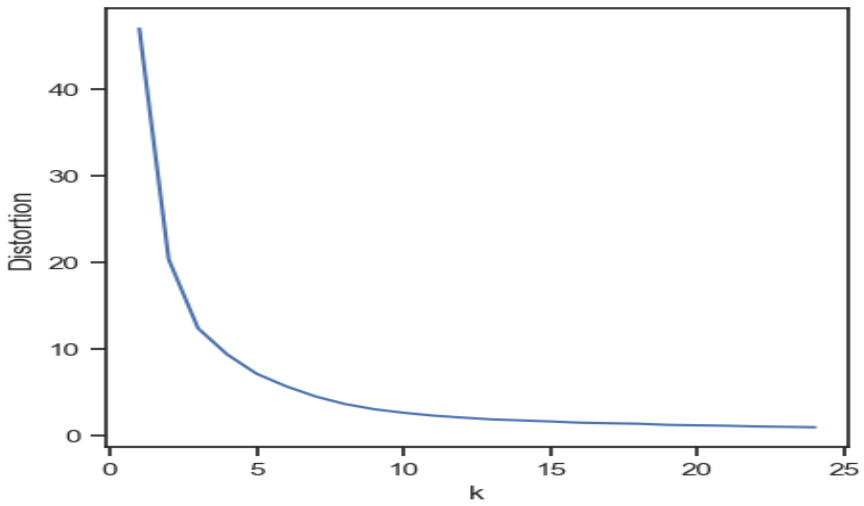


Fig. 7: Elbow method for Number of Clusters

Geographical Hotspots for accidents

The map below presents 3 specific regions, Kingston upon Hull, Humberside and East Riding Yorkshire. The map revealed 25 hotspots with Red indicating very high numbers, Yellow for medium and Green for relatively low numbers. However, there are 5 out of the 25 hotspots with higher number of accidents in the region. The city of Kingston upon Hull recorded the highest number of 627 accidents. Beverley and Bridlington recorded 60 accidents each whiles Goole and the area around Hedon recorded 51 number of accidents each.

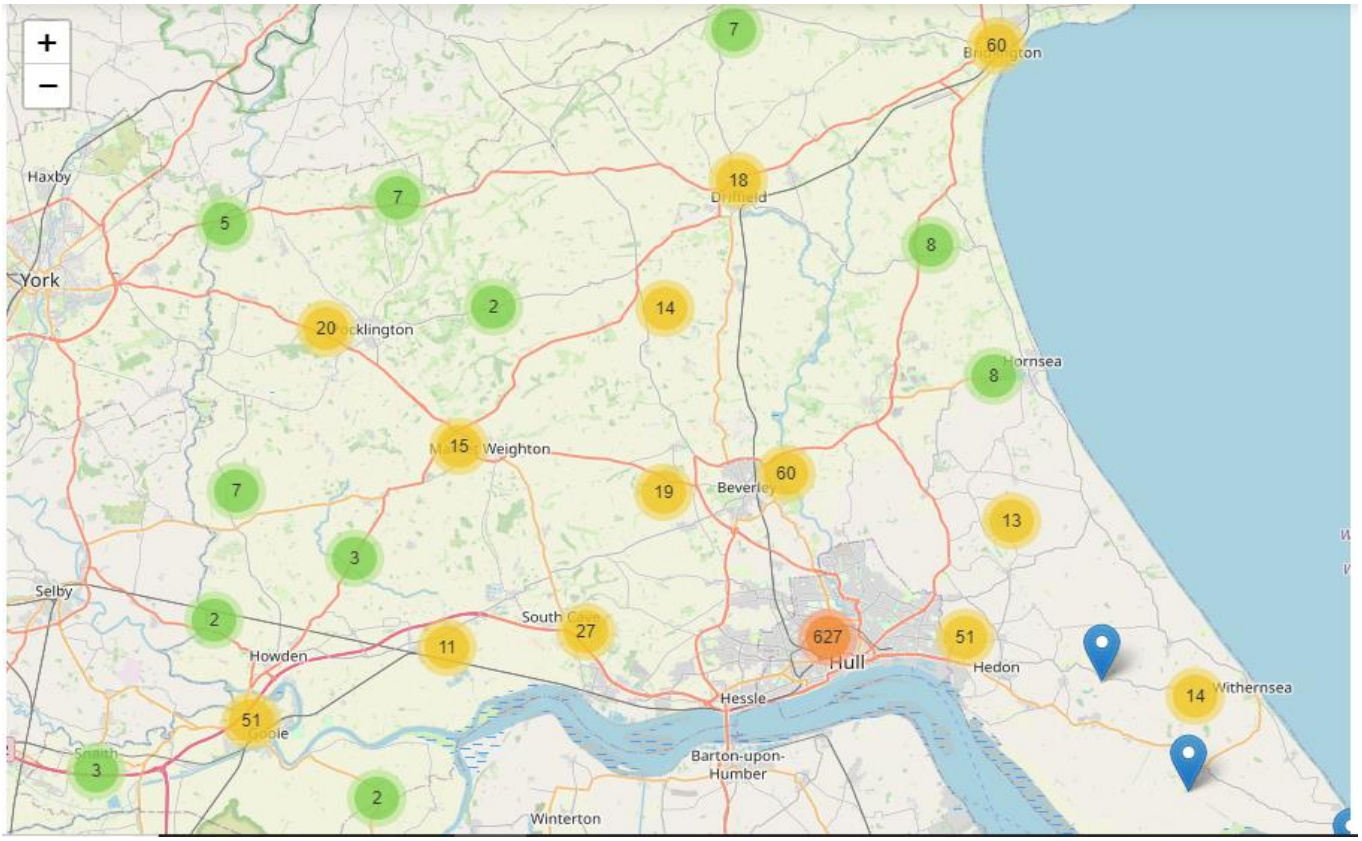


Fig. 8: Map of 3 Regions

Fig. 9 and 10 below show different hotspots with 5 number of clusters. They confirm the insights revealed in the regional map above and a clear indication of the various hotspots identified in the regions. Figure 9 identified various points and a clumsy spot for these points is an indication of a hotspot. Subsequently, in Figure 10, the 5 points represent the clusters or the hotspots identified in the map.

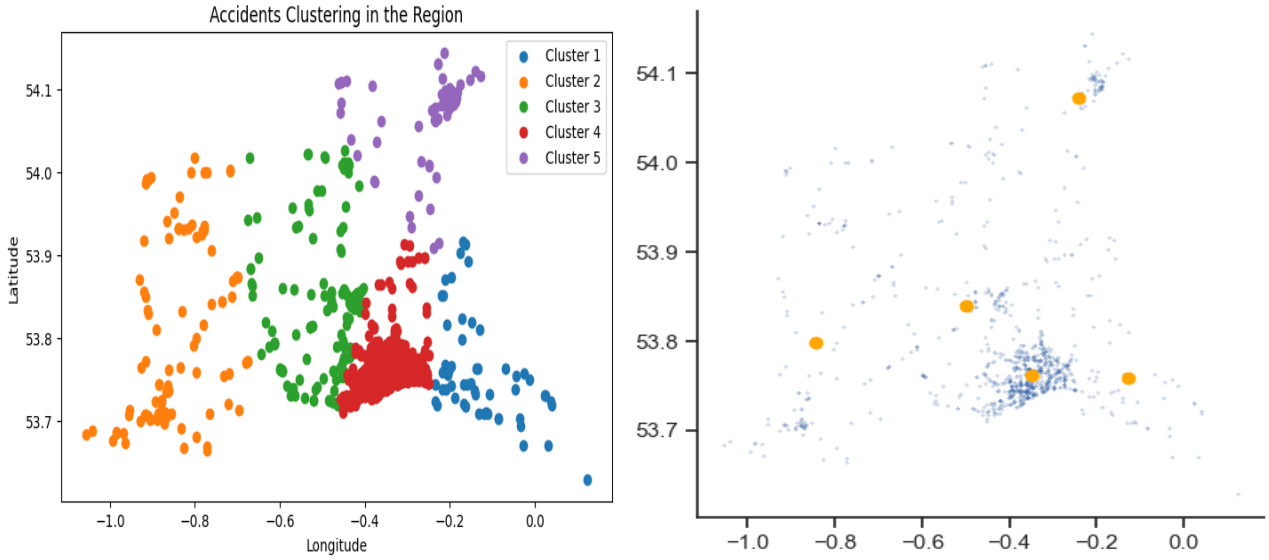


Fig 9 and 10: Geographical clustering in 3 regions

Outlier Analysis in Dataset

Using Isolation Forests Method and labeling normal data points as 1 as well as unusual data points as -1, the following were discovered.

Outliers in the accident table data are 912 while normal data points are 90286. Outliers in the vehicle data are 1673 and normal data points are 165702. Outliers from the casualty table are 1156 and normal data points are 114428. It is evident the outliers generally represent just 1 percent of the data hence the decision to keep the entries since they would not pose distortion to the analysis. These are all presented in Outlier distribution figures 11, 12, 13 and 14 below

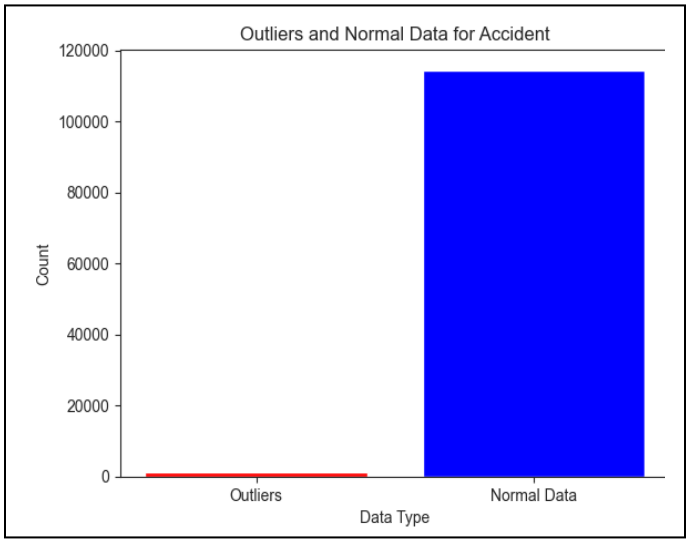


Fig. 11: Distribution of outliers in Accident data

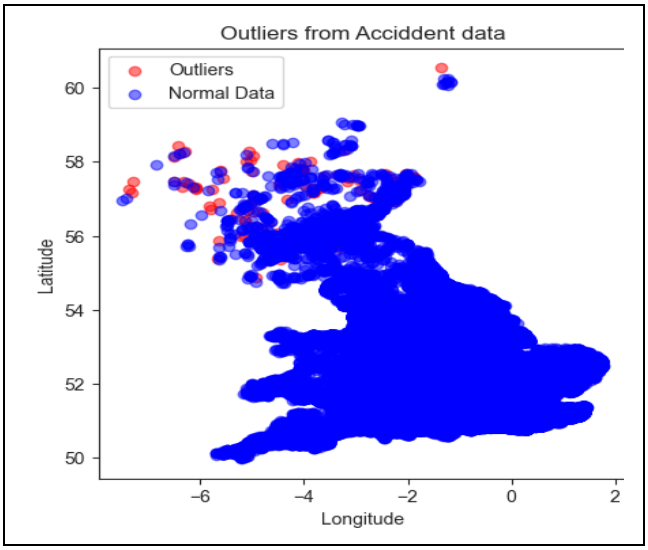


Fig 12. Geographical outliers in Accident data

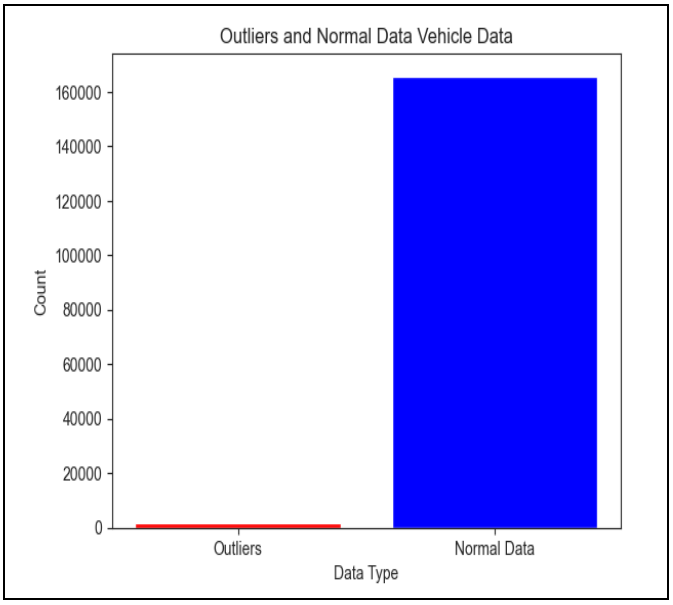


Fig. 13: Distribution of outliers in Vehicle data

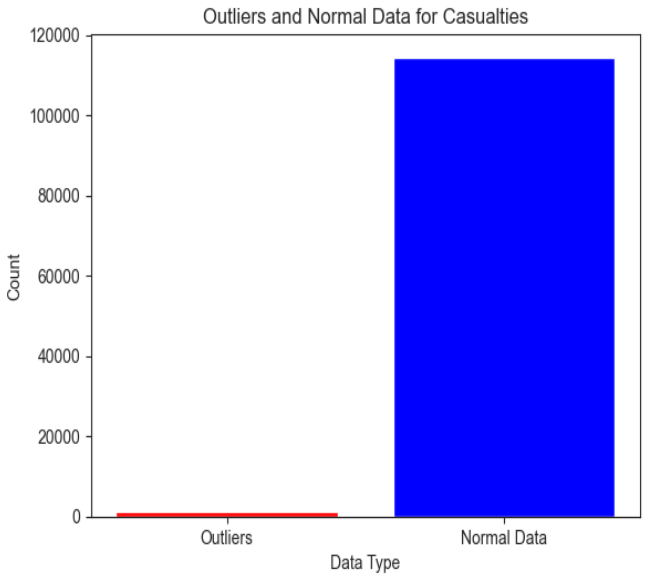


Fig. 14: Distribution of outliers in Casualty data

PREDICTION

A classification model was developed using Decision Tree Classifier machine learning algorithms to predict whether an accident is likely to result in a fatal injury. SelectKBest method was used to select 15 best features for training the model. The model's performance was evaluated using metrics such as accuracy, precision, f1 score, and recall and confusion matrix, providing valuable insights into its predictive capabilities.

Insights

Two models developed, one on accident table data and the second model developed on merged data from the accident, vehicle and casualty tables. Model developed on merged data performed better with an accuracy score of 91% as shown in table 5 below while model built on only data from accident table had an overall accuracy score of 70% evident in table 4. However the decision tree from both models revealed some conditions under which an accident could be considered as fatal or non-fatal. The level of injury sustained by casualties (casualty_severity) is a good predictor, **Number of casualties** involved in the accident and **Speed limit** as a higher speed could result in result in a fatal accident resulting in fatal injuries. Others are lighting conditions under which accident happened as well as the number of vehicles involved.

	precision	recall	f1-score	support
Non-Fatal	0.72	0.66	0.69	11140
Fatal	0.68	0.74	0.71	11110
accuracy			0.70	22250
macro avg	0.70	0.70	0.70	22250
weighted avg	0.70	0.70	0.70	22250

Table 4. Model report for Accident Table data Table

	precision	recall	f1-score	support
Non-Fatal	0.91	0.91	0.91	17210
Fatal	0.91	0.92	0.91	17590
accuracy			0.91	34800
macro avg	0.91	0.91	0.91	34800
weighted avg	0.91	0.91	0.91	34800

5. Model Classification report for Merged Data

Confusion Matrix for the best performing model as shown in Fig. 15 below, the model predicted 15,602 instances as Non-Fatal and the actual label was Non-Fatal (True Negatives). It also predicted accurately 16,122 instances as fatal and the actual label was fatal (True Positives). The inaccurate predictions 1468 (False Negative) and 1608 (False Positives) are relatively low.

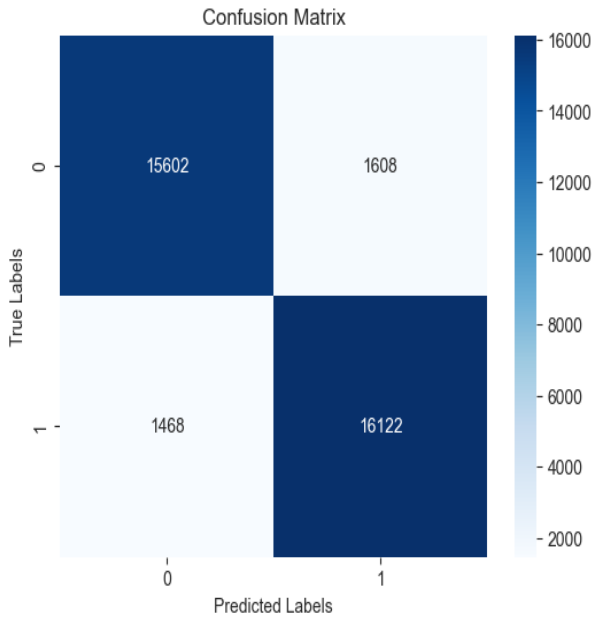


Fig. 15: Confusion Matrix for Model on merged data

RECOMMENDATIONS

Based on the analysis and predictions, the following recommendations can be adopted by government agencies to improve road safety:

1. Implement targeted road safety campaigns during peak hours and days identified in the analysis to enhance road safety awareness among drivers, motorbike riders, and pedestrians. Campaign should be intensified at identified hotspots.

2. Implement efficient traffic management strategies during identified high-risk periods to reduce the risk of accidents and traffic congestion.

3. Provide specialized training programs for motorbike riders, especially for those using larger engine-size motorcycles, to improve their handling skills and road awareness.

4. Invest in Road Infrastructure: Conduct thorough safety assessments of roads and intersections and invest in necessary improvements and maintenance to minimize accident-prone areas.

5. Strengthen Law Enforcement: Increase police presence during high-risk periods to enforce traffic regulations and deter reckless driving.

CONCLUSION

The comprehensive analysis of road accidents in the UK during 2020 revealed that, most of the accidents occur at about 8:00 and 17:00 hours from Mondays to Fridays and these are considered as rush hours for commuters. By following the suggested recommendations, government agencies can make significant progress in reducing the number of accidents and enhancing road safety for all road users. Overall, this study serves as a valuable resource for informing road safety policies and initiatives.

Reference

Department for Transport. (2022). *Reported Road Casualties in Great Britain - Notes, Definitions, Symbols and Conventions*.https://www.gov.uk/government/publications/road-accidents-and-safety-statistics-notes-and-definitions]. Accessed July 25, 2023.