INTRODUCTION

The report aims to conduct analysis of road traffic accident data from the year 2020 in Great Britain and develop a predictive model that can access fatal injuries resulting from these accidents. The goal is to provide insights into accident patterns, identify critical periods and locations, and recommend measures to improve road safety.

DATASET DESCRIPTION

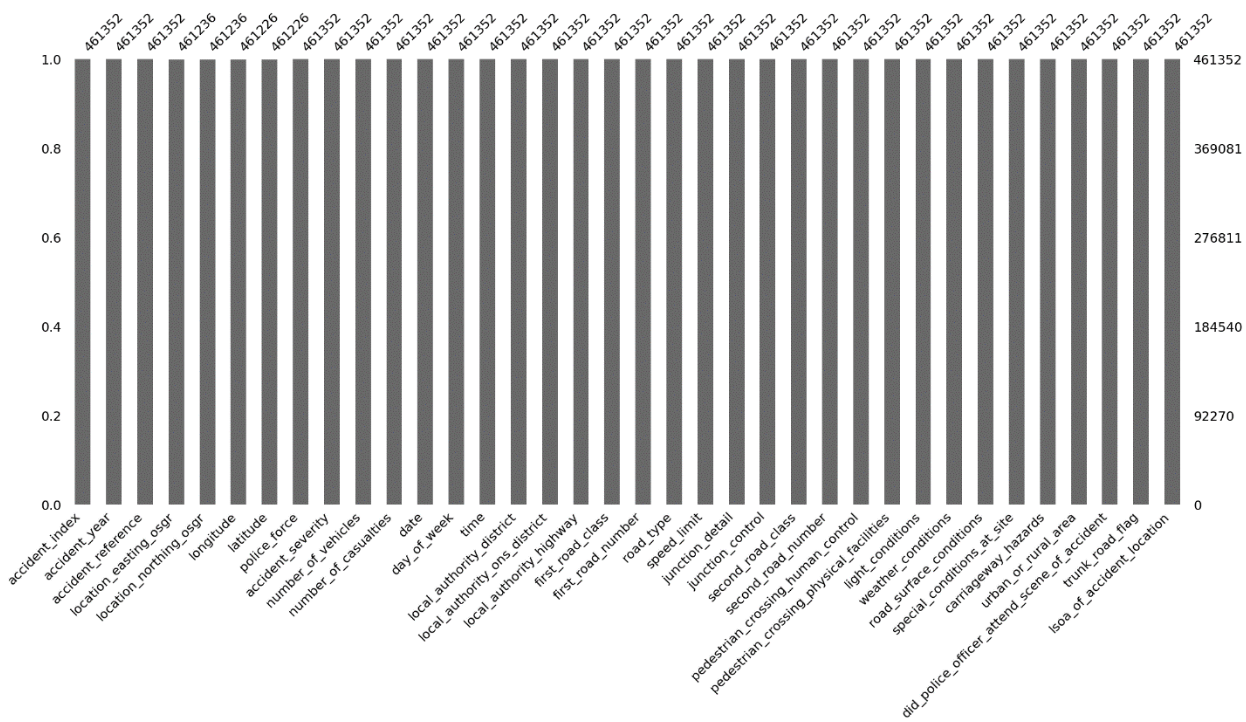
The dataset used for this analysis is a SQLite database named "accident_data_v10.0_2023.db" containing road traffic accident data from 2020 in Great Britain. The data has been collected and reported by government agencies and consists of information about accidents, casualties, vehicles involved, and various factors related to accident.

The main tables in the dataset includes:

1. **Accidents**: Contains information about each recorded accident, including the accident ID, weather conditions, accident severity, number of vehicles, number of casualties, speed limit, road type
2. **Vehicles**: The vehicle data consists of factors describing vehicles involved in accidents. These factors include driver data, junction information, journey purpose, and the condition and features of the vehicle.
3. **Casualty**: The casualties data includes variables describing the casualties involved in accidents, with each accident potentially having multiple casualties. These variables include the condition of casualties, specific casualty group data, casualty data, and information about pedestrians.
4. **LSOA**: The LSOA (Lower Layer Super Output Area) contains information related to the geographical areas in Great Britain. These variables include lsoa01cd, lsoa01nm, shape area, shape length, global id.

DATA CLEANING

HANDLING MISSING DATA



Upon further review of certain variables like 'Location Northing OSGR,' 'Location Easting OSGR,' 'Longitude' and 'Latitude,' it was observed that most rows contained different information. These columns had very few missing details, so the decision was made to remove the rows with missing information from these columns, as it did not significantly affect the analysis.

For columns with fewer missing details, the approach taken involved filling in the missing information by identifying the value that appeared most frequently in those columns. This method allowed the dataset to retain a substantial amount of information, enabling more comprehensive data analysis.

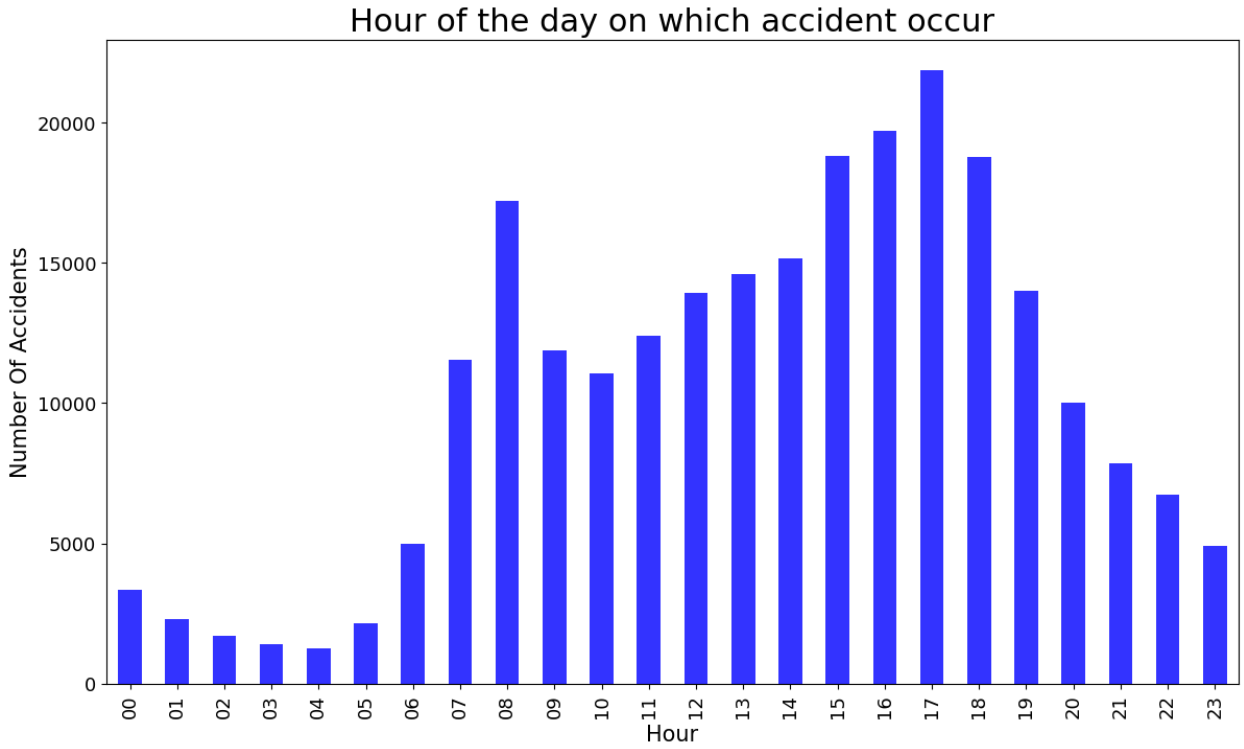
However, in columns where a substantial amount of information was missing, particularly evident in the vehicle and casualty data, the choice was made to exclude rows with such missing details. This step was taken to maintain the integrity of the analysis by working with a more complete dataset, ensuring that the results obtained were reliable and less influenced by the missing information.

DATA ANALYSIS

We conducted an exploratory analysis of the dataset to gain deeper insights into the information and achieve a thorough understanding of the attributes within the data. We pinpointed specific areas of interest and employed various techniques for Exploratory Data Analysis (EDA) to extract insights.

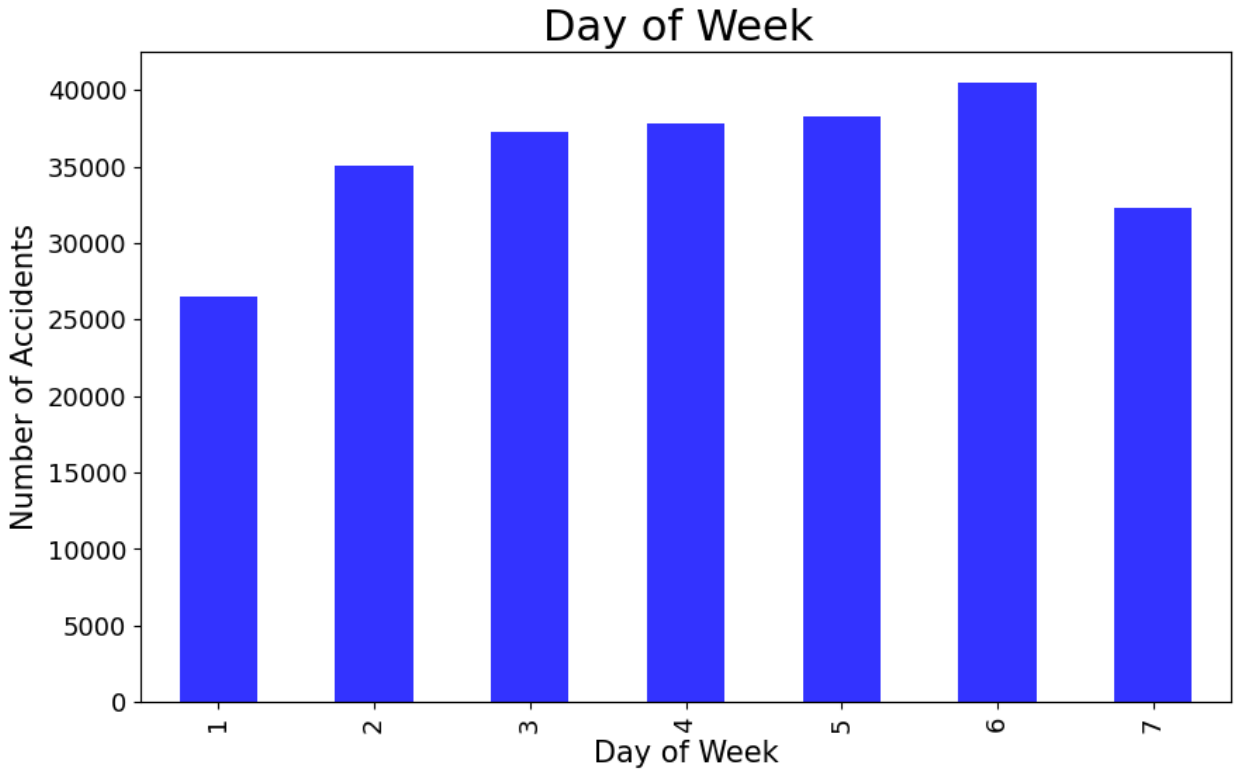
QUESTION 1: Are there significant hours of the day, and days of the week, on which accident occur?

HOUR OF THE DAY ON WHICH ACCIDENT OCCUR



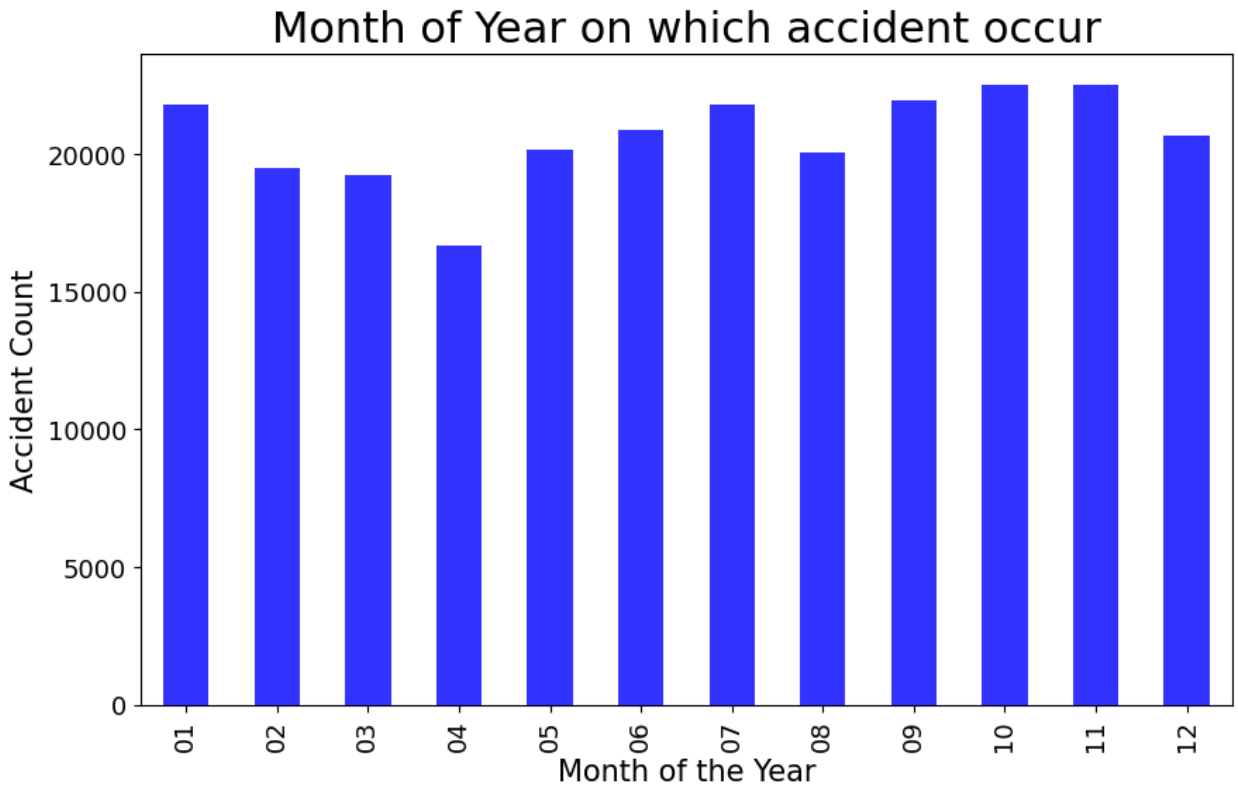
The period of least risk for being on the road is from midnight to 5 in the morning, as indicated by the plotted data. During this period, the roads are generally clear, resulting in a lower frequency of accidents. However, as the day progresses and people begin their daily commutes, the number of accidents tends to increase. The evening hours, particularly from 3 p.m. to 6 p.m., show a significant rise in accidents, likely due to rush hour traffic and higher road activity. The peak number of accidents is observed at 5 p.m., likely when many individuals are heading home after a busy day at work. Additionally, around 8 a.m., when people are hurrying to work, there is another substantial increase in accidents.

DAY OF THE WEEK ON WHICH ACCIDENT OCCUR



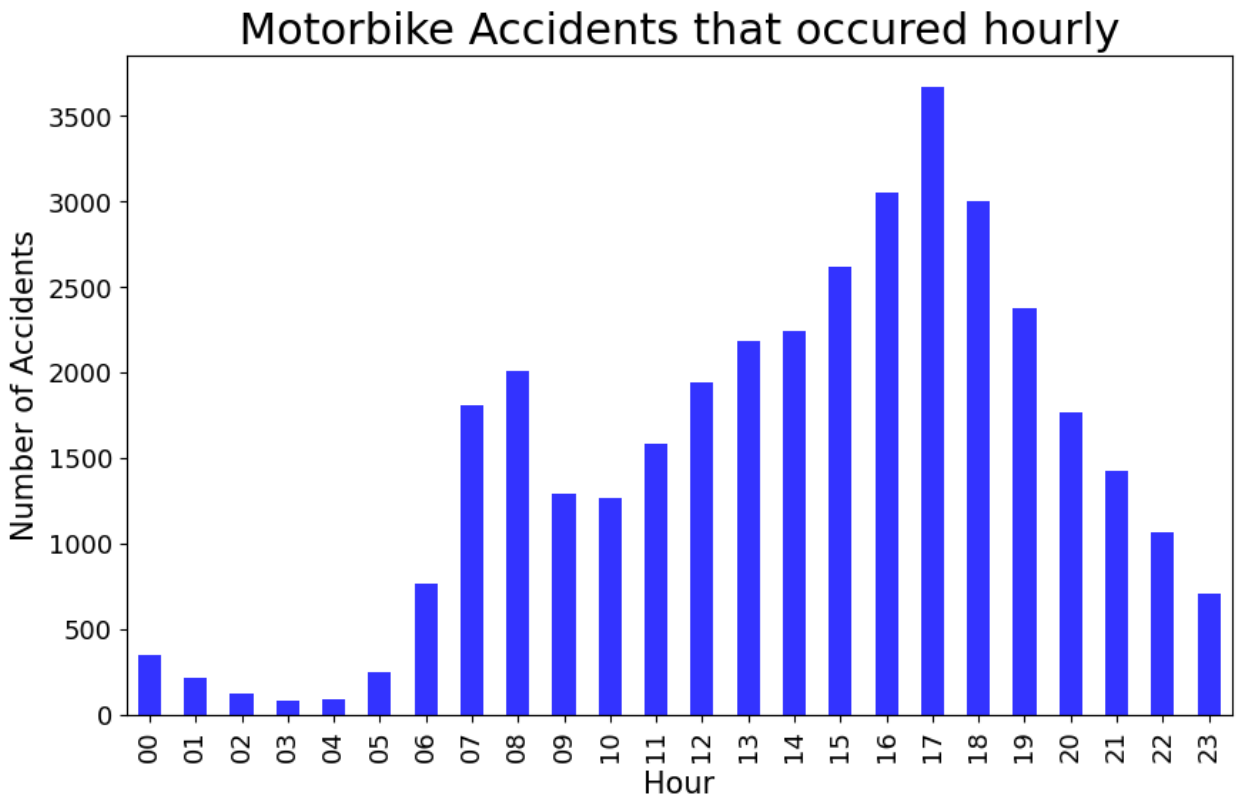
Accidents are most prevalent on Saturdays, likely due to higher traffic volume as people go out on their days off. Additionally, drivers may be less alert or aware on Saturdays, contributing to the increased accident rate. Mondays have the least accidents, likely because of reduced traffic and increased driver alertness as people return to work. Promoting road safety awareness is crucial, especially during high-traffic periods.

MONTH OF THE YEAR ON WHICH ACCIDENT OCCUR

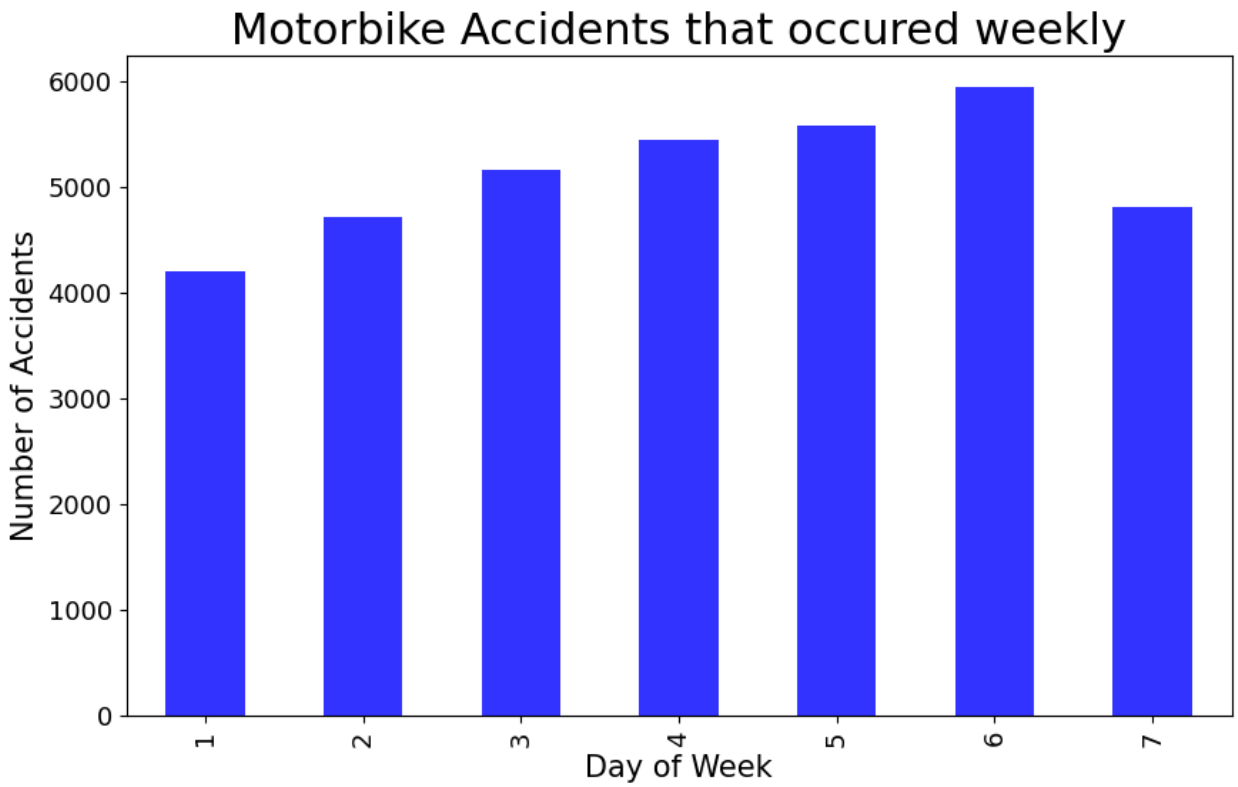


October and November have more accidents because of tricky weather, shorter daylight, and holidays, while April has fewer accidents due to better weather, cautious driving, and less traffic.

QUESTION 2: FOR MOTOBIKES, ARE THERE SIGNIFICANT HOURS OF THE DAY, AND DAYS OF THE WEEK, ON WHICH ACCIDENTS OCCUR

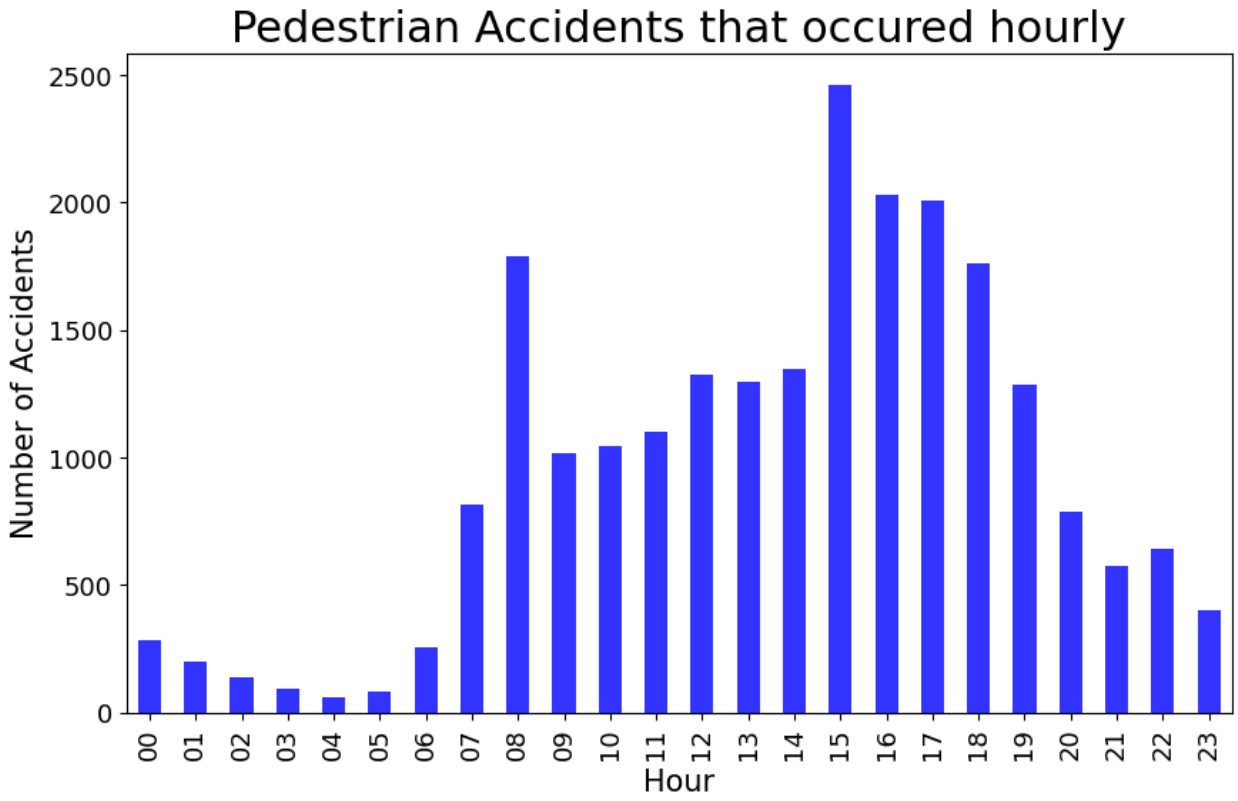


The majority of motorcycle accidents occur during the evening hours, with the highest concentration happening between 4:00 p.m. and 6:00 p.m., particularly peaking around 5:00 p.m. Conversely, the lowest number of accidents is observed between midnight and 5:00 a.m.

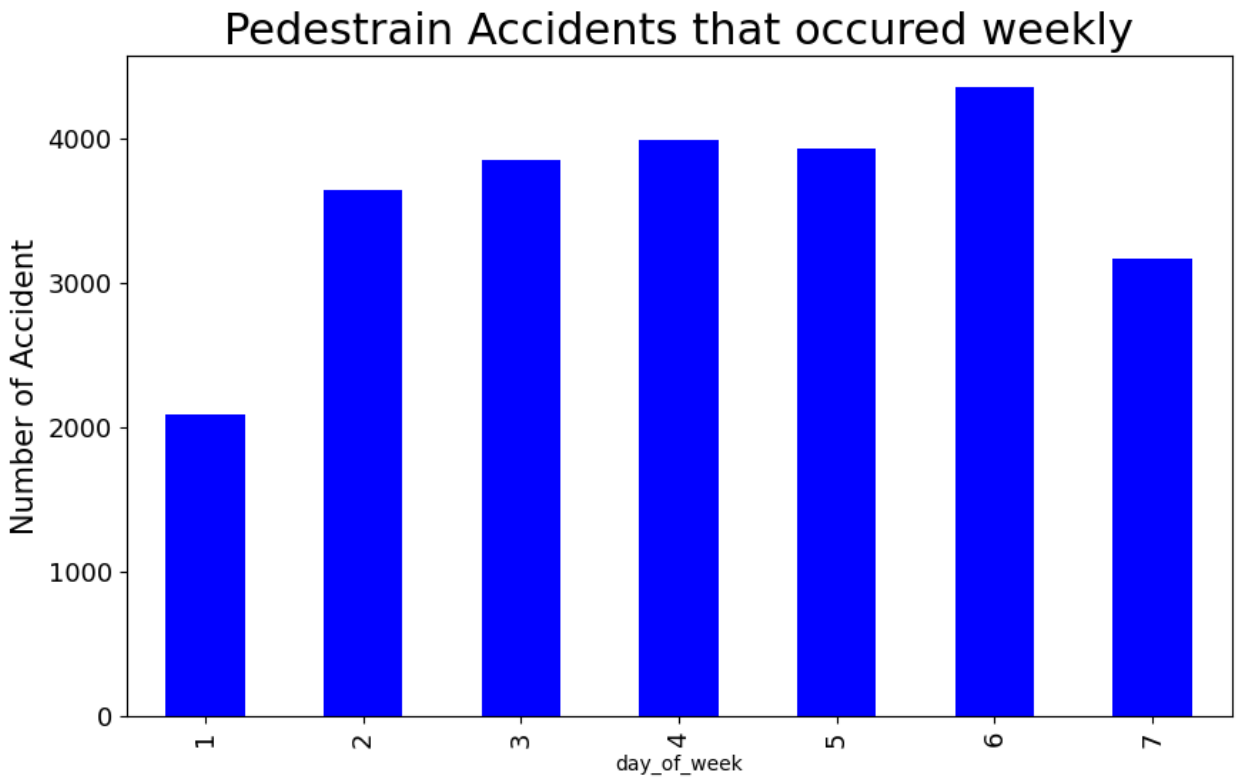


Motorbike accidents are most prevalent on Saturdays and least on Mondays, reflecting higher weekend activity and lower road traffic on weekdays.

QUESTION 3: FOR PEDESTRIANS INVOLVED IN ACCIDENTS, ARE THERE SIGNIFICANT HOURS OF THE DAY, AND DAYS OF THE WEEK, ON WHICH THEY ARE MORE LIKELY TO BE INVOLVED?



Pedestrian accidents peak around 3pm, while the fewest occur between midnight and 5am. Walking before 8am appears particularly hazardous, as indicated by the graph.



Similar to overall accidents, pedestrian accidents show a higher frequency on Saturdays and a lower occurrence on Mondays.

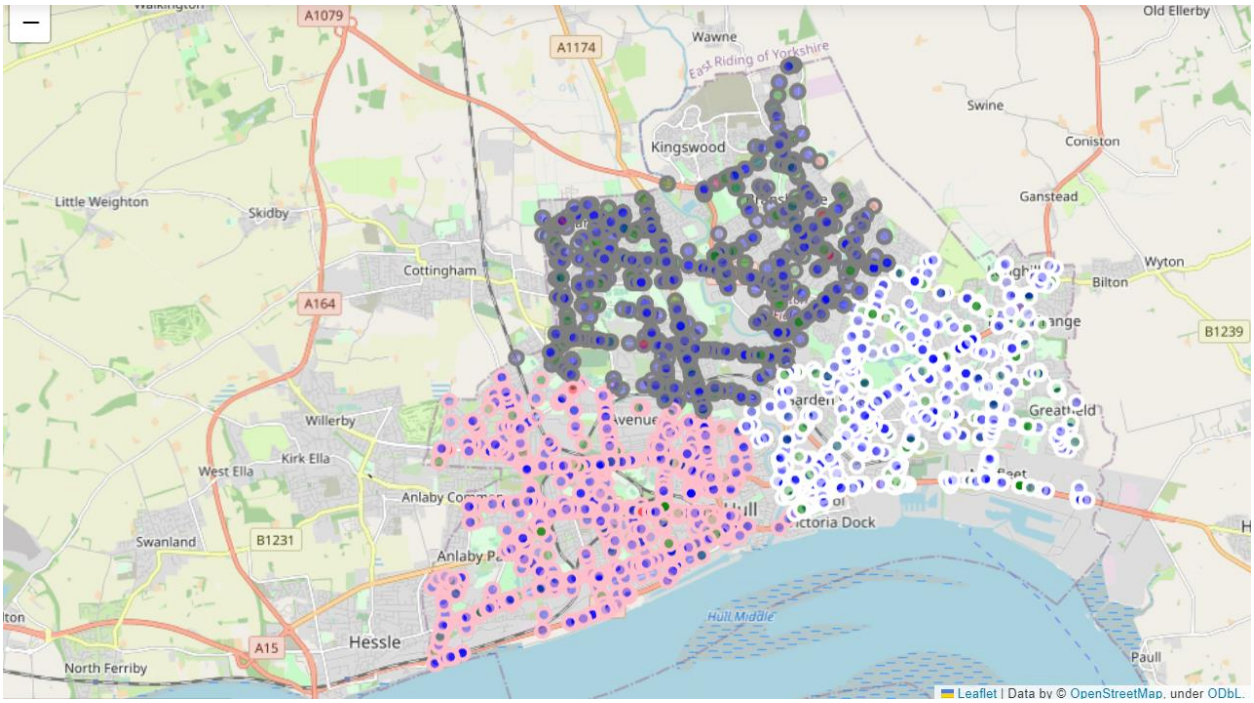
QUESTION 4: USING APRIORI ALGORITHM, EXPLORE THE IMPACT OF SELECTED VARIABLES ON ACCIDENT SEVERITY

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(speed_30)	(severity_3)	0.672691	0.817349	0.554515	0.824324	1.008533	0.004691	1.039699	0.025849
1	(severity_3)	(speed_30)	0.817349	0.672691	0.554515	0.678431	1.008533	0.004691	1.017850	0.046321
2	(speed_30, weather_conditions_1)	(severity_3)	0.535783	0.817349	0.439569	0.820423	1.003761	0.001647	1.017116	0.008071
3	(weather_conditions_1, severity_3)	(speed_30)	0.648870	0.672691	0.439569	0.677437	1.007055	0.003080	1.014714	0.019953
4	(speed_30)	(weather_conditions_1, severity_3)	0.672691	0.648870	0.439569	0.653448	1.007055	0.003080	1.013210	0.021405
5	(severity_3)	(speed_30, weather_conditions_1)	0.817349	0.535783	0.439569	0.537798	1.003761	0.001647	1.004359	0.020512

These association rules provide statistical insights into the relationships between the speed limit, weather conditions, and accident severity. The support indicates the frequency of occurrence of the rule, and confidence measures the likelihood of the consequents given the antecedents. Lift values close to 1 indicate weak associations, while values greater than 1 suggest positive associations. Conviction values greater than 1 suggest non-randomness, and Zhang's metric

measures statistical significance. The row represents the association rule “If the accident severity is 3 which is (Slight), then the speed limit is 30 and the weather condition is 1 which is (fine without high winds). These associations are powerful rules that tell us the impact of speed limit and weather conditions on accident severity.

QUESTION 5: Identify accidents in our region: Kingston upon Hull,Humberside, and the east riding of yorkshire etc. You can do this by filtering on LSOA, or police region or another method if you can find one. Run clustering on this data. What do these clusters reveal about the distribution of accidents in our region?



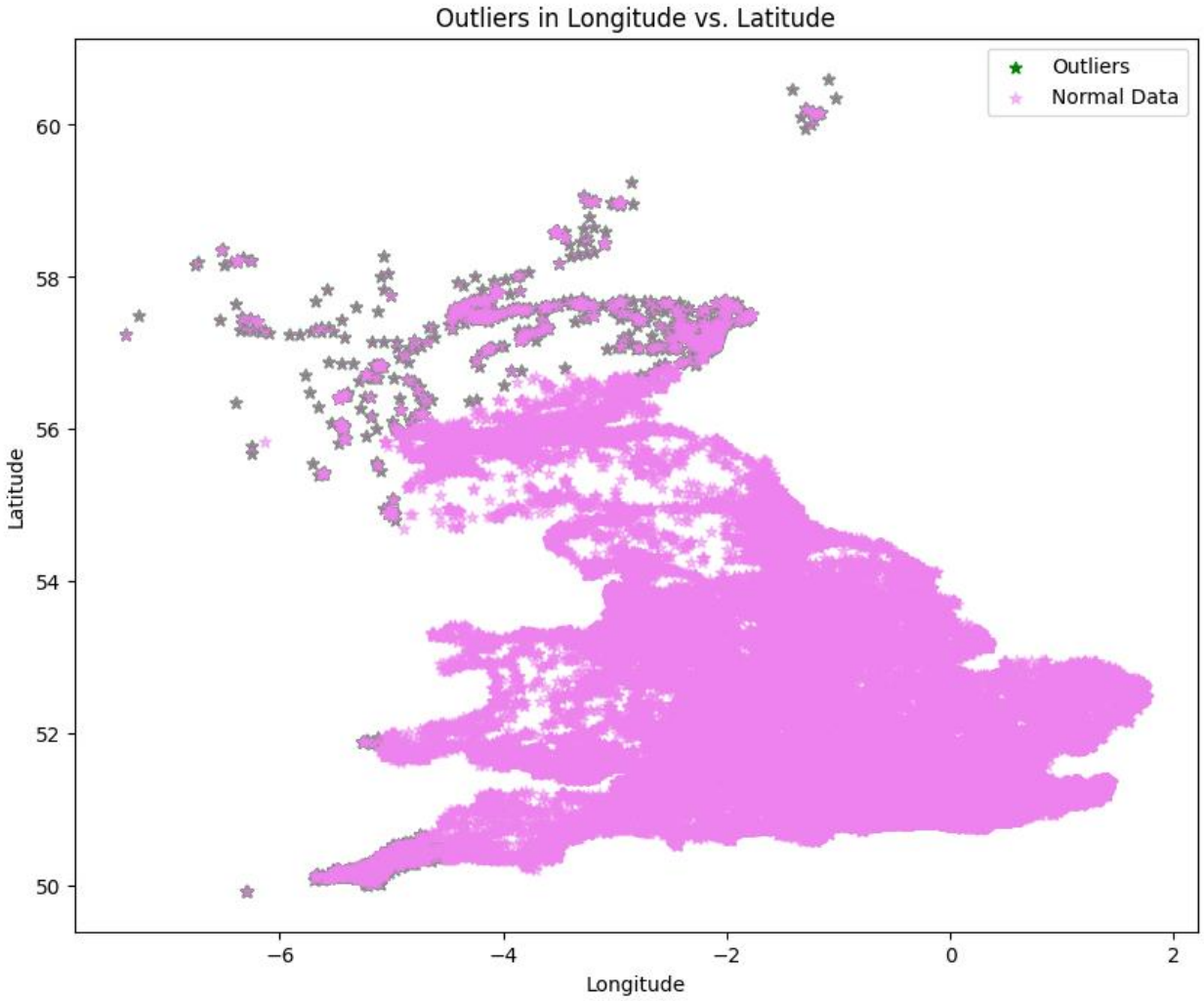
AN ANALYSIS OF ACCIDENT DATA IN HULL USING CLUSTER ANALYSIS AND ACCIDENT SEVERITY CLASSIFICATION

The map analysis reveals the presence of three distinct clusters within the accident data. The first cluster, denoted as Cluster 0 and represented by grey markers, encompasses accidents lacking a discernible pattern in terms of both location and severity. The second cluster, Cluster 1, is highlighted by white markers and encompasses accidents of a relatively moderate nature in regard to both their distribution and severity. The final cluster, labeled as Cluster 2 and depicted with pink markers, exhibits accidents that exhibit a pronounced concentration in specific geographic areas, potentially indicating higher severity levels.

In terms of accident severity categorization, three levels are identified: fatal severity marked by red indicators, serious severity indicated by green markers, and slight severity represented by blue markers. Notably, Cluster 2's pink markers exhibit a strong clustering in particular regions, such as Analby Road and Hull Paragon Interchange. A noteworthy observation is that red markers (indicative of fatal severity) are relatively dispersed across the map, whereas green markers (representing serious severity) demonstrate a more uniform distribution. Meanwhile, blue markers

(reflecting slight severity) are the most prevalent, encompassing a substantial portion of the map, particularly within various regions.

QUESTION 6: Using outlier detection methods, identify unusual entries in your data set. Should you keep these entries in your data?



From the map we can see that the green marker represents the outliers (unusual entries) while the pink marker represents the normal data. Certainly, keeping outliers in the road traffic accident data can be crucial, especially in extreme cases that hold significant meaning for the analysis. Here's is an emphasis on why keeping such outliers is important:

1. **Extreme Severity**: Outliers in the accident severity category may indicate incidents with severe consequences, such as fatalities or serious injuries. These outliers are critical for road safety analysis, as they represent incidents that have a higher impact on human lives and require special attention.

2**. Uncommon Locations**: Outliers in the geographical attributes (latitude and longitude) may represent accidents that occurred in remote or unusual locations

3. **Unusual Time Patterns**: Outliers in the time attributes might point to accidents that occurred during a typical time frames. For example, accidents that happen during late-night or early-morning hours may have different contributing factors compared to daytime accidents.

4. **Weather and Road Conditions**: Outliers in weather and road condition attributes could represent accidents that occurred under extreme weather circumstances, indicating the potential need for weather-related safety interventions.

In summary, outliers in road traffic accident data should not be automatically removed. They represent critical data points that can provide deep insights into the factors contributing to severe accidents and help shape effective road safety measures.

Question 7: Can you develop a classification model using the provided data that accurately predicts fatal injuries sustained in road traffic accidents, with the aim of informing and improving road safety measures?

Answer

This is the accuracy scores for all classifiers

Variable	Random Forest	KNN	Naïve Bayes
Accident Severity	0.81	0.79	0.73

Based on these findings, we employed classification models including Random Forest, KNN, and Naïve Bayes. Notably, the Random Forest Classifier demonstrated the highest performance in dataset classification, achieving an accuracy rate of 81 percent. Furthermore, the models effectively predict fatal injuries, as evidenced by their accuracy scores.

Recommendation

It's extremely important for governments to truly grasp the main reasons behind road traffic accidents and when and where they happen. This understanding is crucial for them to step in and work towards lowering the number of lives lost due to these unfortunate events. Here are some practical steps that can really make a difference in reducing the chances of accidents and how bad they get:

1. **Keep Speed in Check**: Make sure that speed limits are set at levels that ensure everyone's safety on the roads. It's about making sure people drive at speeds that are reasonable and safe for everyone sharing the road.

2. **Follow the Rules We Already Have**: Let's make sure the laws and rules we've got in place are followed and enforced. This includes cracking down on things like driving under the influence or not paying attention while driving.

3. **Spread the Word on Road Safety**: We need to make sure everyone knows the ins and outs of road safety. This means telling people about using seatbelts, avoiding distractions, and being careful pedestrians.

4. **Get Help Fast When Needed**: When accidents do happen, we want to be sure that help arrives quickly. This involves training and preparing our emergency response teams to get to the scene fast and provide the care that's needed.

5. **Build Smarter Roads**: Our roads should be designed with safety in mind. That means clear signs, well-lit intersections, and safe spots for people walking or biking. We can also use clever technology to make traffic flow better and lessen the chances of accidents.

These steps, when taken together, can go a long way in making our roads safer for everyone. It's about looking out for each other and doing what we can to prevent accidents and keep our roads a place where we all can travel without fear.

REFERENCES

Akour, M., Al Qasem, O. and Hanandeh, F., 2022. Traffic Accident Severity Prediction: A comparison Study. *International Journal of Transportation Systems*, *7*.

Kumar, A.P. and Santosh, D.T., 2022. Road Accident Severity Prediction Using Machine Learning Algorithms. *International Journal of Computer Engineering in Research Trends*, *9*(2), pp.175-183.

Mostafa, S.M., Salem, S.A. and Habashyis, S.M., 2022. Predictive model for accident severity. *IAENG Int. J. Comput. Sci*, *49*, pp.110-124.