Introduction

It is important to analyse and predict traffic accidents in order to limit their occurrence. This project will focus on the analysis of UK road accident data given by the Department of Transport. The UK government publishes data studies on traffic accidents every year. As of 2019, there were 1,752 recorded road traffic fatalities and 153,158 reported casualties of all types, with 25,945 of them suffering significant injuries.

There are two objectives for this project; Conduct an exploratory analysis of 2019 road accident datasets to get insight into the current condition of road accidents in the UK; to investigate how well machine learning algorithms predict the occurrence and severity of traffic accidents in the United Kingdom to enhance road safety.

In this project, mining of road traffic data will be done using an association rule-based algorithm. The Apriori Algorithm is used to establish a relationship between features in the dataset, and then the rules are investigated with high lift and high support. K-nearest Neighbors and Naive Bayes classifiers are used in classifying the independent of attributes. To create clusters and evaluate them based on characteristics, K-means is utilized. The positive findings will surely help the government make better decisions and prevent traffic accidents.

Dataset

The dataset in this project is separated into four categories as collected from the UK government data portal.

- Accident Datasets
- Vehicle Datasets
- Casualty Datasets
- Adjustments Datasets
- Variable Lookup Dataset

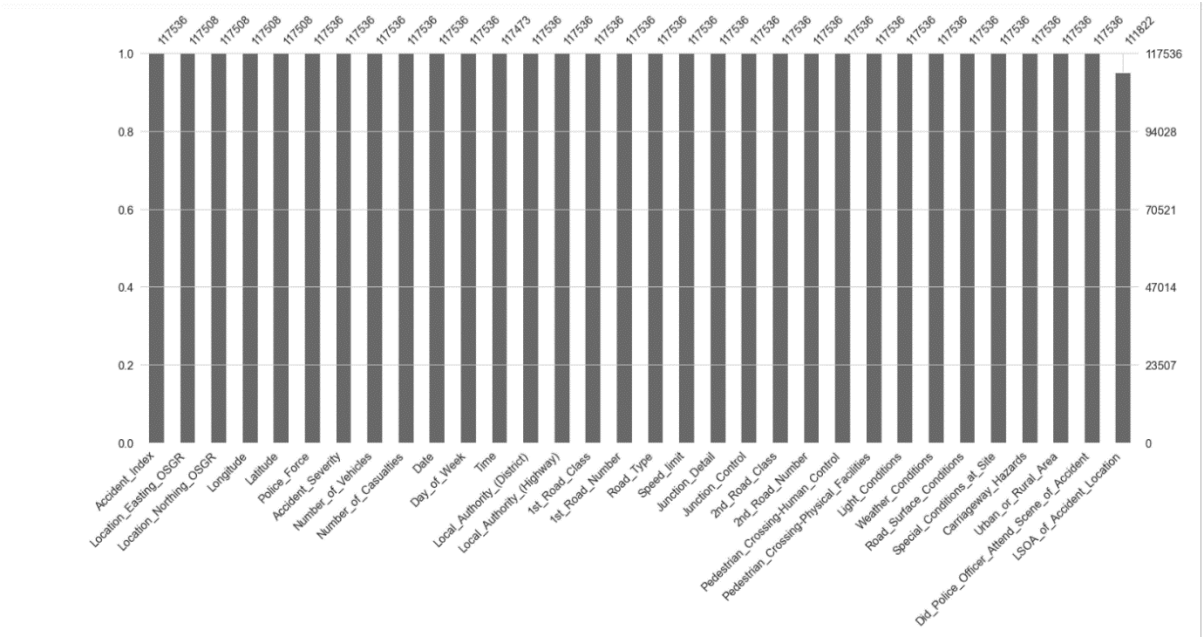
At most granular level, the Accidents dataset may be regarded as a superclass of the vehicle and casualty dataset linked via a shared 'accident index' identifier.

Data Pre-processing

The process of preparing data for analysis is known as data pre-processing. Cleaning the data, engineering new features through data transformation, and determining which portions of the data to keep and which to discard were all part of the process(Alkheder et al., 2017). This section explains how the dataset was prepared for analysis, modeling, and predictive analysis using pre-processing techniques.

Data Cleaning

Missing value



The source datasets' overall strategy was to assign unknown data items a value of (-1), which is a standard method in data reporting. However, there were times when data was completely missing. To comply with the basic strategy, the entire dataset was analysed to detect this (null,missing n/a, None). The initial thought was to replace all missing values with -1 to comply with the report's general approach, but a closer examination of some columns ('Location Northing OSGR', 'Location Easting OSGR','Longitude','Latitude', 'Time','LSOA of Accident Location') reveals that most of the data in them have unique values, and the percentage of missing values are insignificant.Missing values in these columns were dropped entirely while that of other columns were replaced with -1.

After standardising the value for all missing data, colunms with less than 5% missing values (-1) were replaced with the most frequent (modal) value and the rows of the ones with more missing values (especially in the vehicle and casualty dataset) were dropped from the record. This was an appropriate strategy to utilize and not have a detrimental influence on the model's performance since a greater percentage of the dataset was available for analysis.

Feature Selection

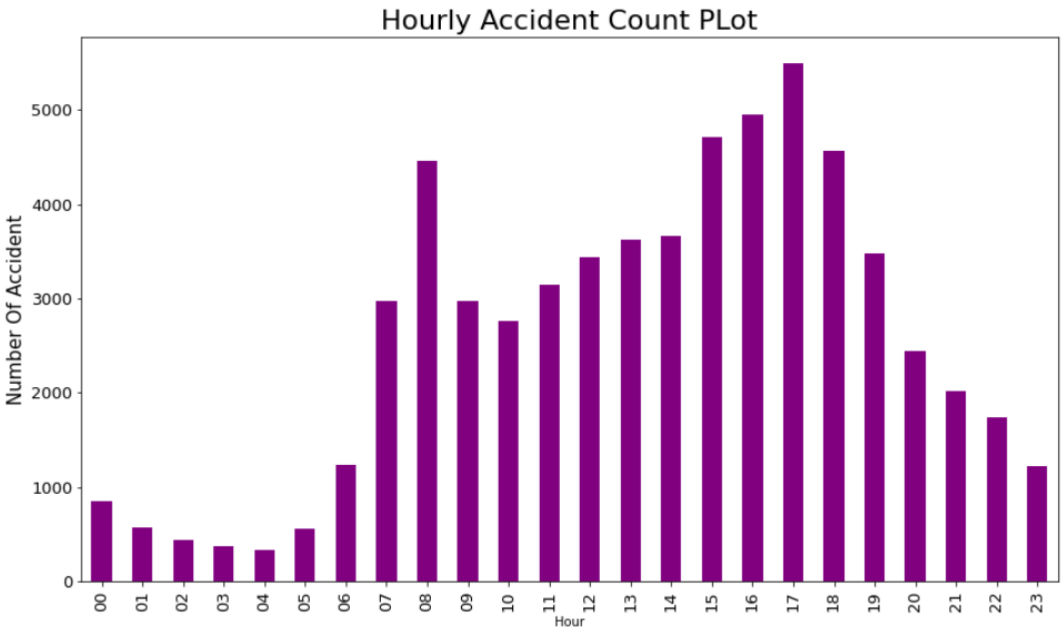
This project used an iterative feature selection method as different analyses and models needed different data transformations and feature engineering with the goal of enhancing the analysis quality.An initial examination of each of the three datasets was conducted to ensure that the structure of the datasets was consistent

Exploratory Data Analysis

An exploratory analysis of the dataset was done to better grasp the information and acquire a better knowledge of the attributes in the data. Focus points were considered, and certain EDA are made to guide this study.

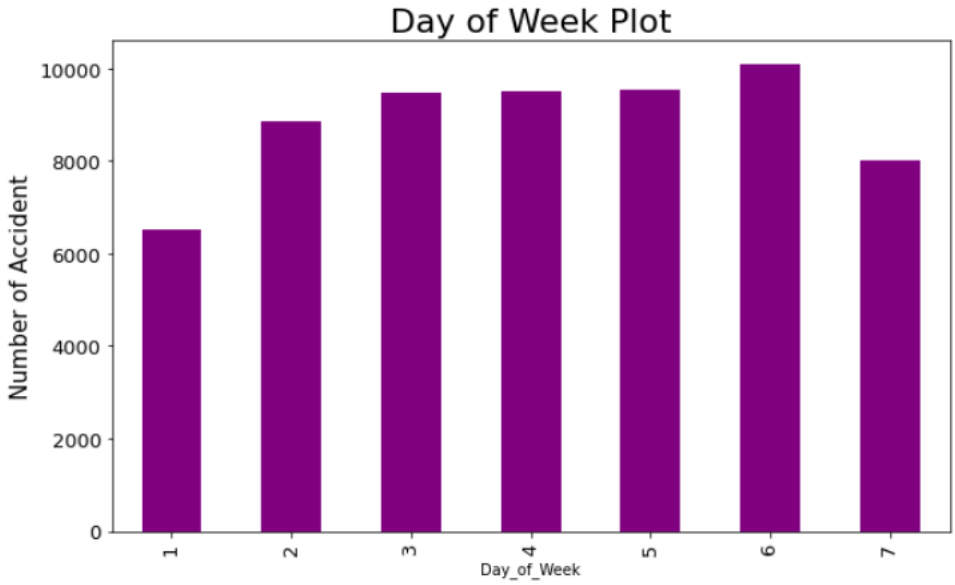
A. Are there significant hours of the day, and days of the week, on which accidents occur?

HOURS OF THE WEEK



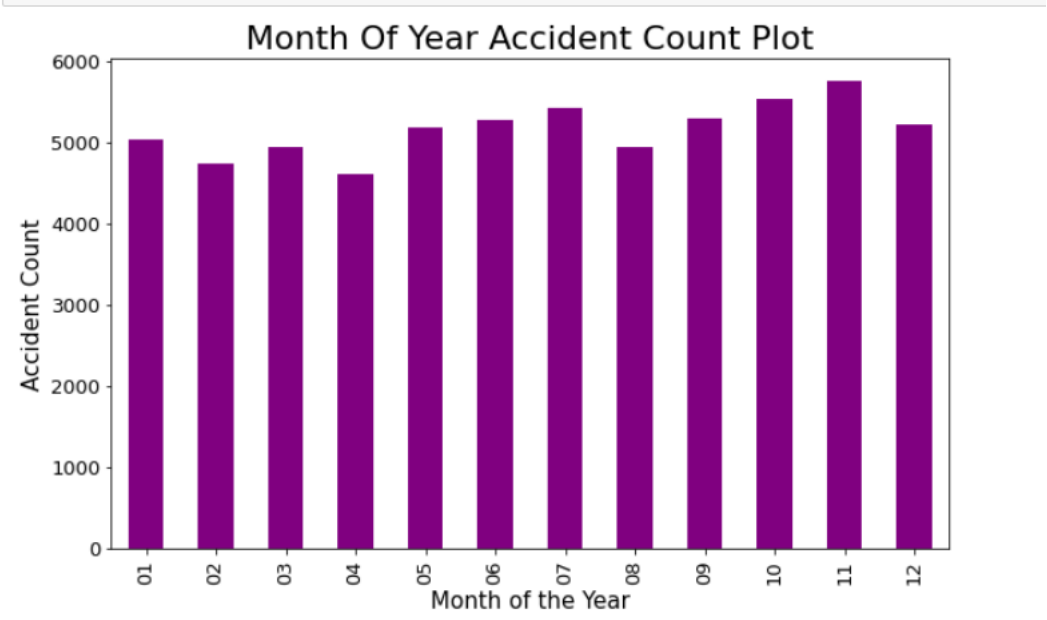
Accidents are most prevalent in the evening, as one might expect because people are going home after a hard day at work, and the plot reveals that more accidents occur at this time, with 5pm having the largest number of accidents. From 3 to 6 p.m., the number of accidents tends to increase. By 8 a.m., when people are hurrying to work, the graph reveals a substantial number of accidents during this time. The safest time to be on the road is between 12 to 5 a.m., when the roads are normally clear.

DAYS OF THE WEEK



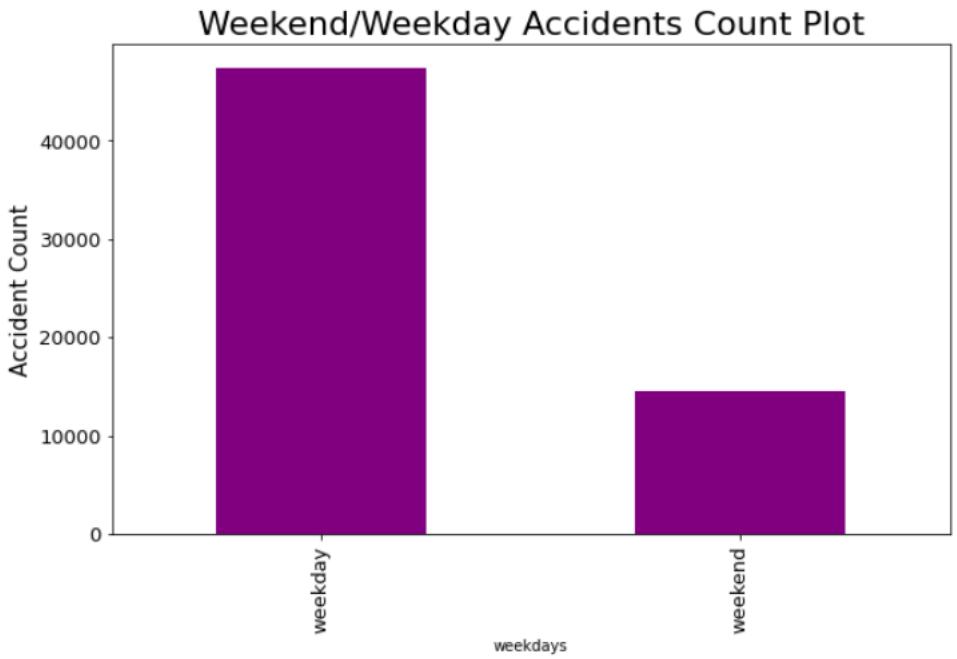
The graph shows Monday having the least accidents while Saturday has the highest. The high number of accidents on Saturday might be due to a multitude of causes.There are usually a lot more vehicles on the road on Saturday. People want to spend their days off work by going out which entails driving to numerous places. Second, drivers may be less alert or aware on Saturdays than they are during the week, which may hinder them from being as vigilant or aware as they are during the week

MONTH OF THE YEAR



The month of November has the highest number of traffic accidents. While one may anticipate the winter to have the most number of accidents because of the poor weather and long nights, the summer and spring months constitute 3 of the top 6 in terms of the highest number of accidents. This might be because there are more vehicles on the road as people enjoy the sunlight.

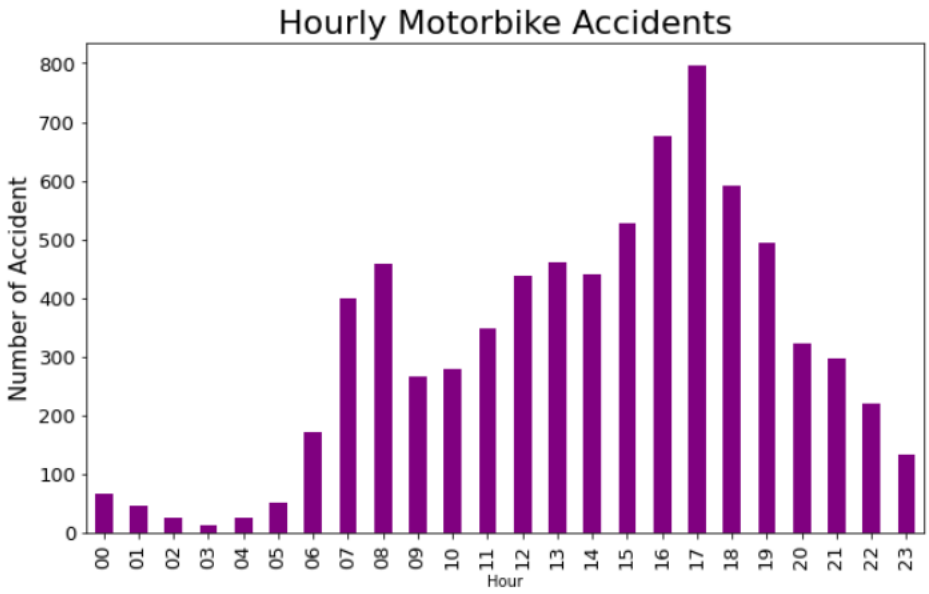
ACCIDENTS ON WEEKDAY OR WEEKEND



The previous graph reveals Saturday has more accidents than any day of the week.

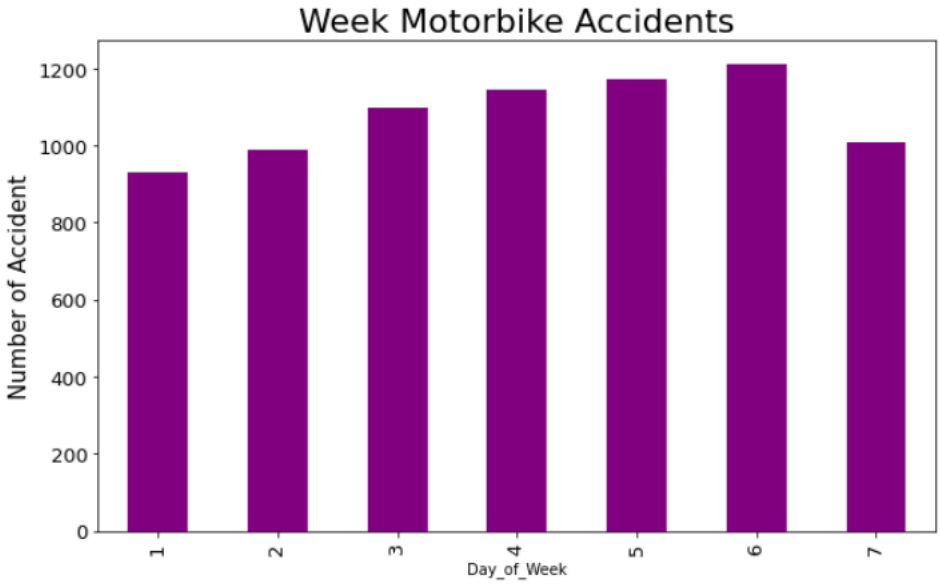
B. For motorbikes, are there significant hours of the day, and days of the week, on which accidents occur?

MOTORBIKES – HOURS OF THE DAY



According to the data, most motorbike accidents occur in the evening, between 4:00pm and 6:00 p.m., with a peak around 5 p.m. Between the hours of 12:00 a.m. and 5:00 a.m., there are the fewest motorbike accidents.

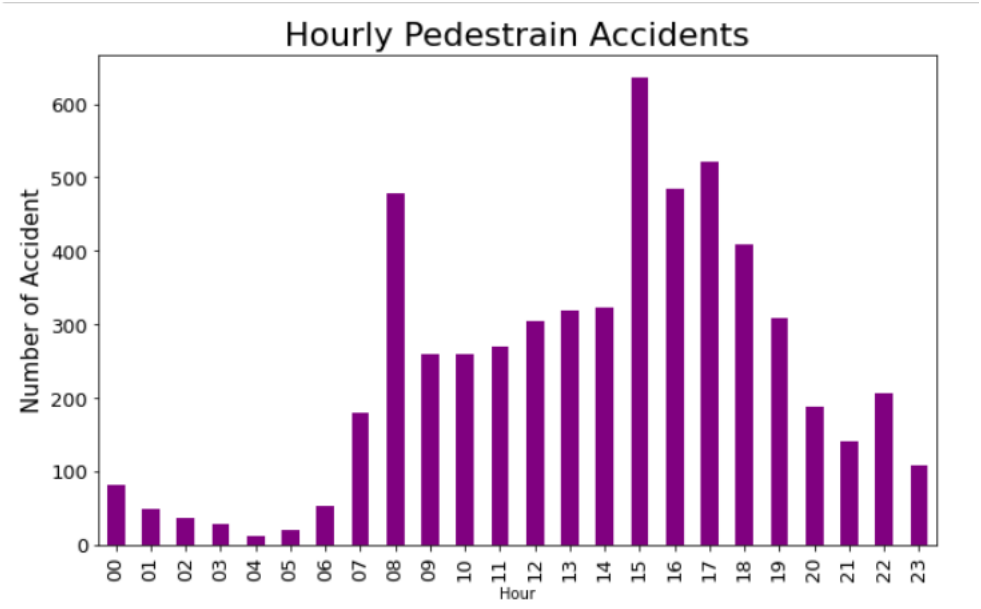
MOTORBIKES – DAYS OF THE WEEK



Just like the general accident, motorbike accidents tend to be more on Saturdays and the least on Monday.

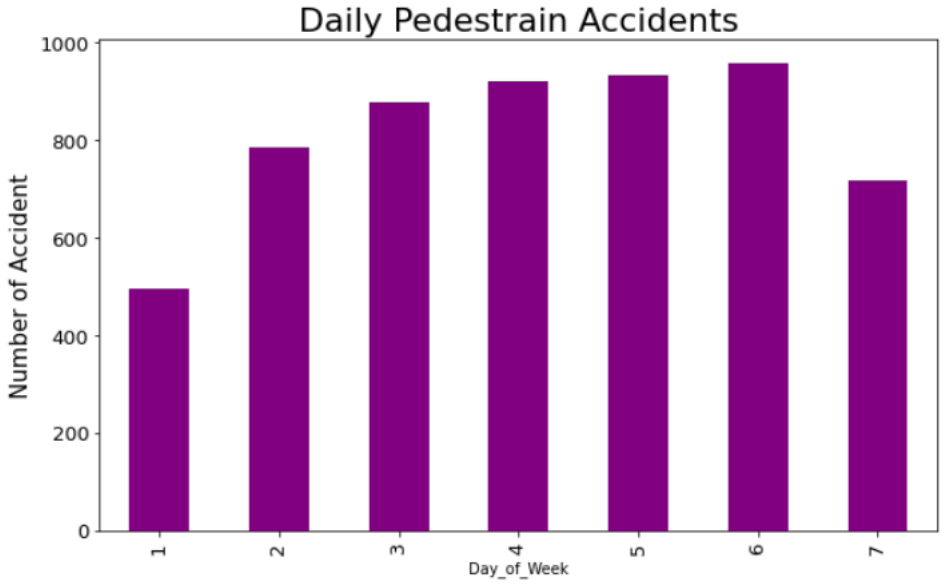
C. For pedestrians involved in accidents, are there significant hours of the day, and days of the week, on which they are more likely to be involved?

PEDESTRIANS – HOURS OF THE DAY



According to the graph, walking by 8 a.m. and in the evening hours is particularly perilous, since more pedestrian accidents occur at these times, possibly due to increased traffic as people attempt to get home.

PEDESTRIANS – HOURS OF THE DAY



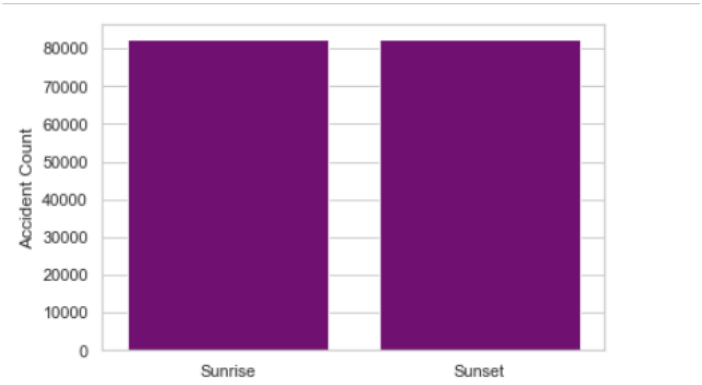
Just like the general accident, pedestrian accidents tend to be more on Saturdays and the least on Monday.

D. What impact, if any, does daylight savings have on road traffic accidents in the week after it starts and stops?

TIME	NUMBER OF ACCIDENTS	DIFFERENCE
Week before Start of DST	1916	
Week after DST	1623	293
Week before End of DST	1584	
Week after After end of DST	1559	25

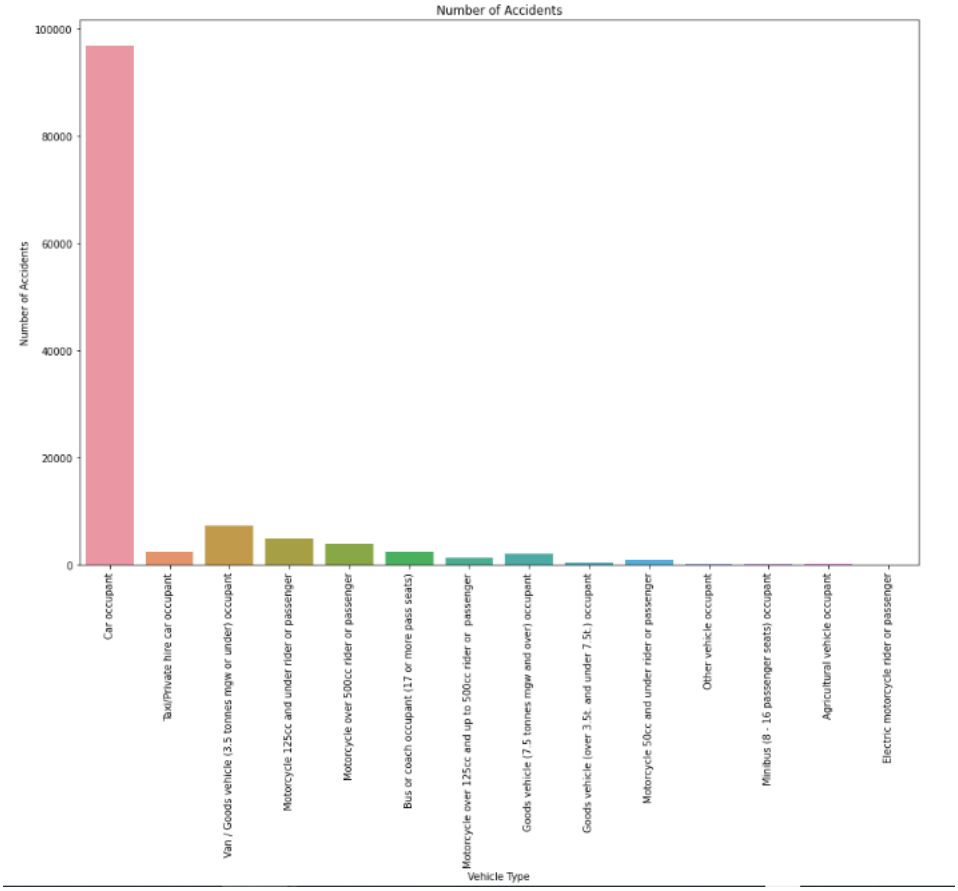
Changes to and from DST may have an impact on the frequency of the days after the shift because of sleep loss(Lahti et al., 2010). In 2019, we looked at the number of road accidents a week before and after the switch. Changes in DST did not increase rather it decreased the number of road accidents, according to our findings.

E. What impact, if any, do sunrise and sunset times have on road traffic accidents?



The light sparkles straight into the eyes of drivers at the sunrise, causing glare drivers. At sunset, drivers' eyes battle to acclimate to the shifting degree of brightness, making it harder to identify potential hazards(Gomes-Franco et al., 2020). The report does not show any impact of sunset and sunrise on road accidents in 2019.

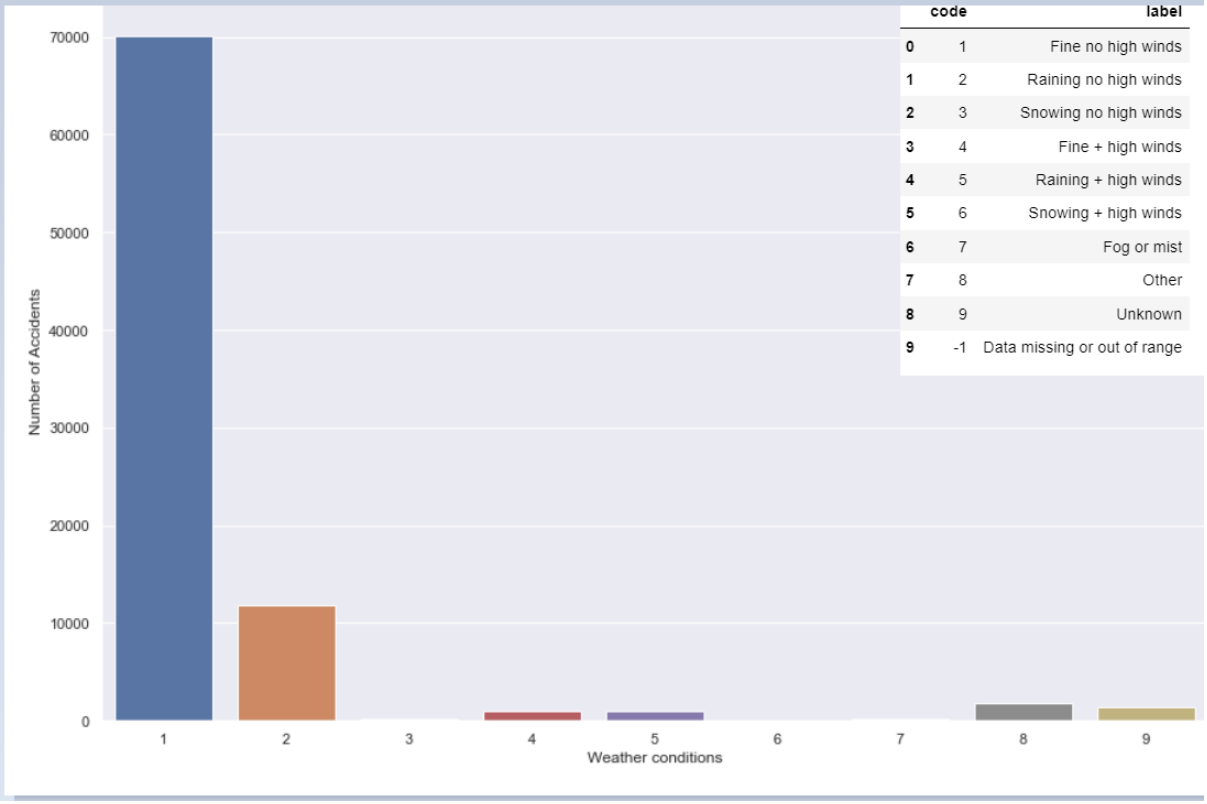
F. Are there particular types of vehicles (engine capacity, age of vehicle, etc.) that are more frequently involved in road traffic accidents?



From the graph, cars are the major culprits of road traffic accidents consisting of more than 90% of all road traffic accidents. The modal age of vehicles involved in an accident is 3 years while the cars with Engine capacity(cc) 1598 tend to be more prone to accidents.

G. Are there particular conditions (weather, geographic location, situations) that generate more road traffic accidents?

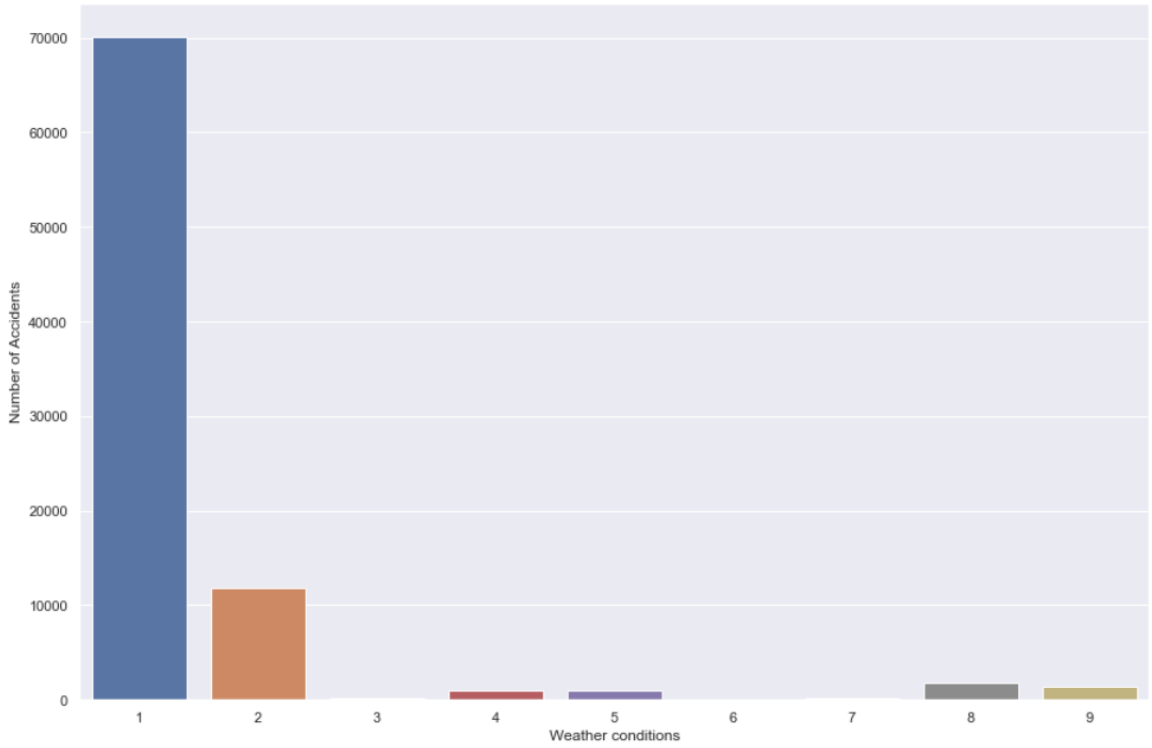
Impact of Weather on Road Traffic Accidents



	code	label
0	1	Fine no high winds
1	2	Raining no high winds
2	3	Snowing no high winds
3	4	Fine + high winds
4	5	Raining + high winds
5	6	Snowing + high winds
6	7	Fog or mist
7	8	Other
8	9	Unknown
9	-1	Data missing or out of range

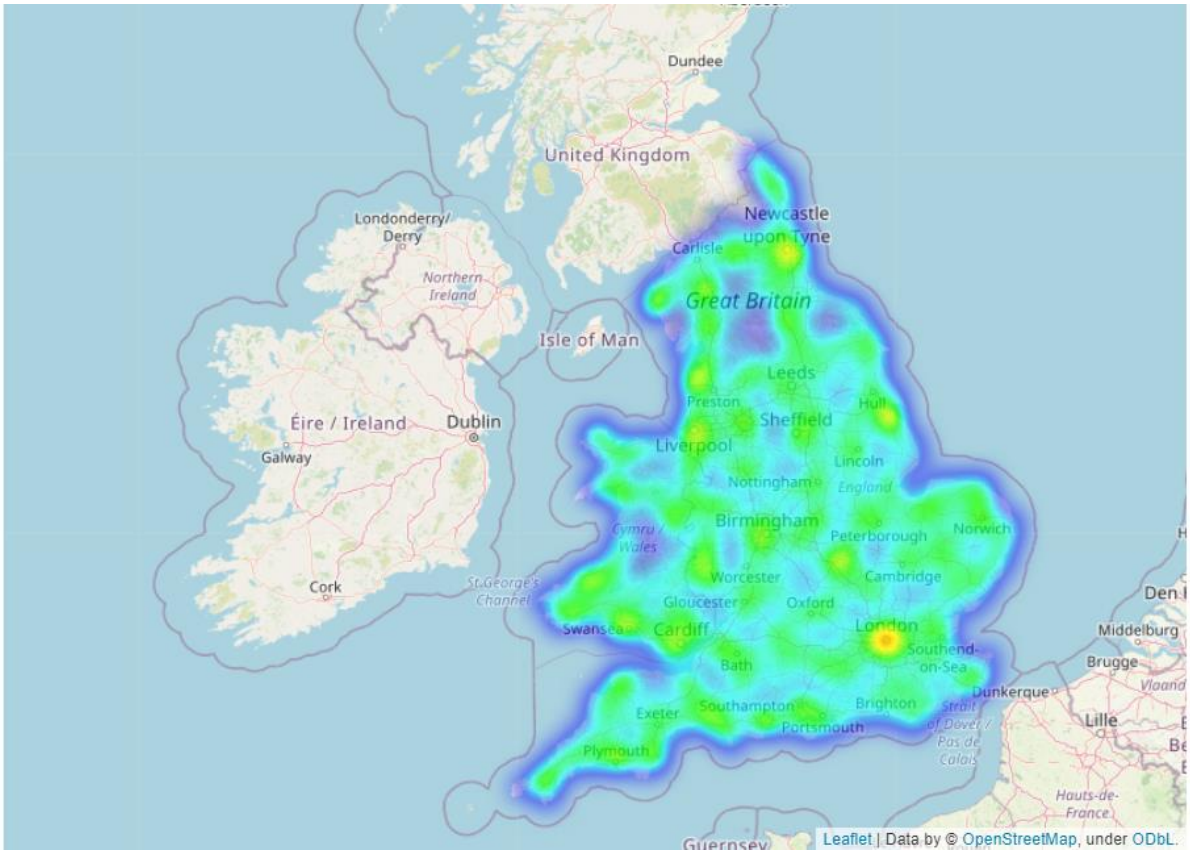
Most accidents happened while the weather was fine. During the rain, a significant number of accidents occurred. There were very few incidents while it was snowing, fog and mist. It is worthy of note that more cars tend to be on the road when the weather is fine than during any other weather conditions.

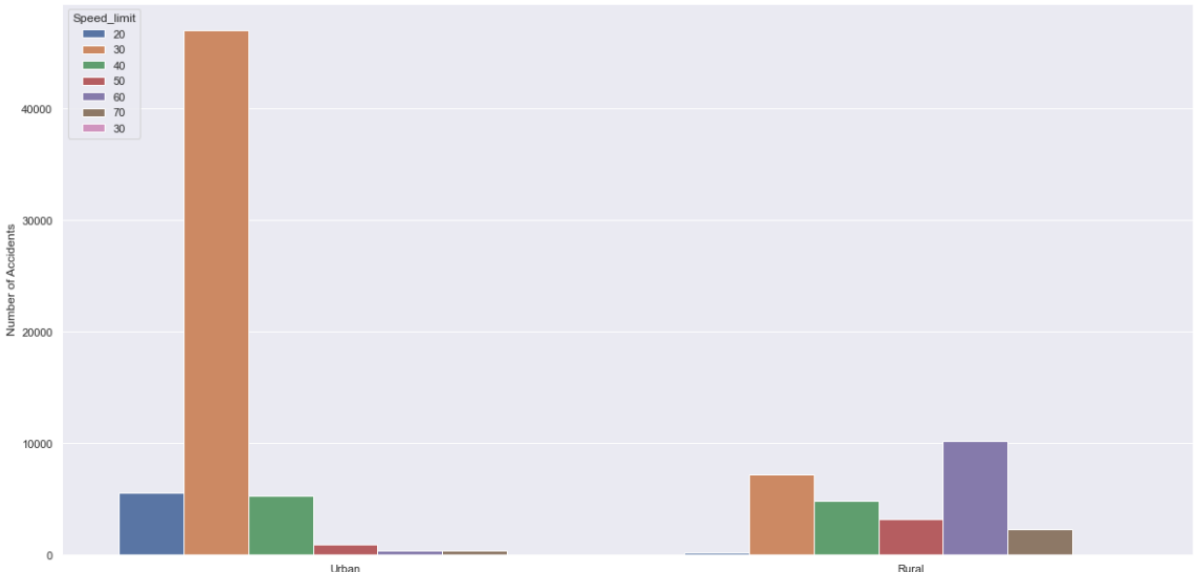
Impact of Light Conditions on Road Traffic Accidents

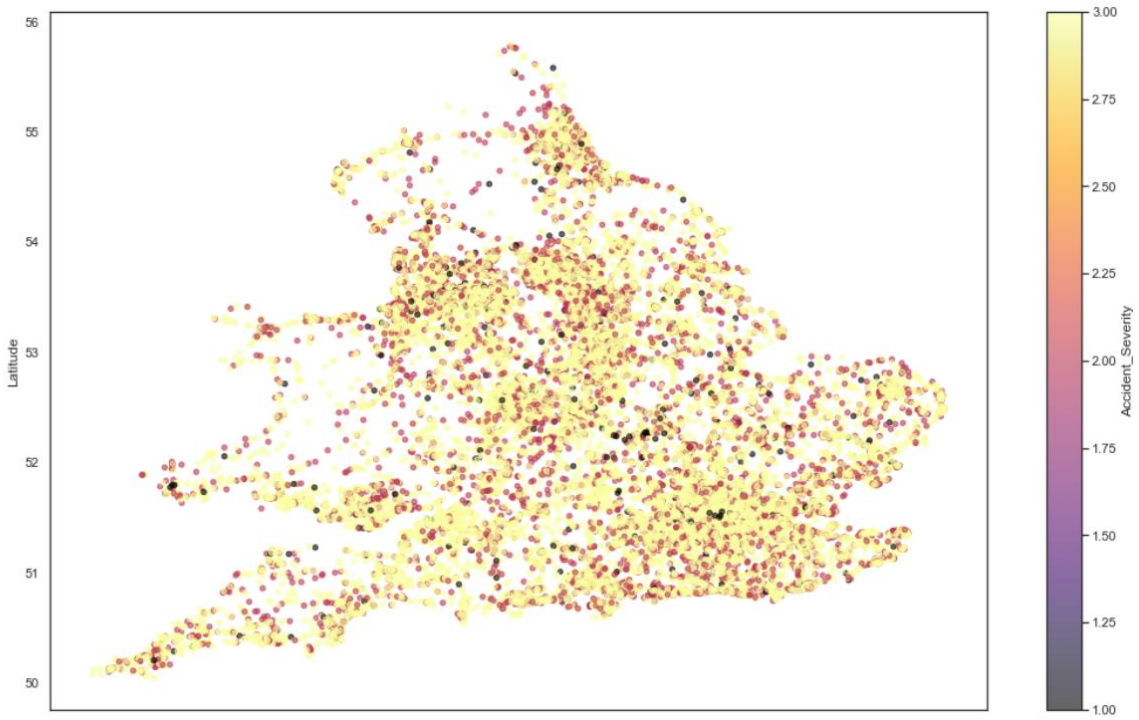


More accidents tend to happen during daylight and a significant amount happened during darkness. Just because there are more accidents during the day does not indicate that driving at night is safer. According to the graph, more than 35% of all accidents occurred in the dark. This is a startling result given that there is 60percent lesser traffic. That implies your chances of being in an accident are substantially higher at night.

IMPACT OF SPEED LIMIT AND GEOGRAPHIC LOCATIONS ON ACCIDENT


Dundee
United Kingdom
Newcastle upon Tyne
Carlisle
Londonderry/ Derry
Northern Ireland
Great Britain
Isle of Man
Leeds
Preston
Hull
Sheffield
Éire / Ireland
Dublin
Galway
Liverpool
Lincoln
Nottingham
England
Cymru / Wales
Birmingham
Peterborough
Norwich
Cork
St.George's Channel
Worcester
Cambridge
Gloucester
Oxford
Swansea
Cardiff
London
Southend-on-Sea
Bath
Middelburg
Brugge
Dunkerque
Exeter
Southampton
Brighton
Portsmouth
Lille
Strait of Dover / Pas de Calais
Plymouth
Hauts-de-France
Guernsey
Leaflet | Data by © OpenStreetMap, under ODbL.

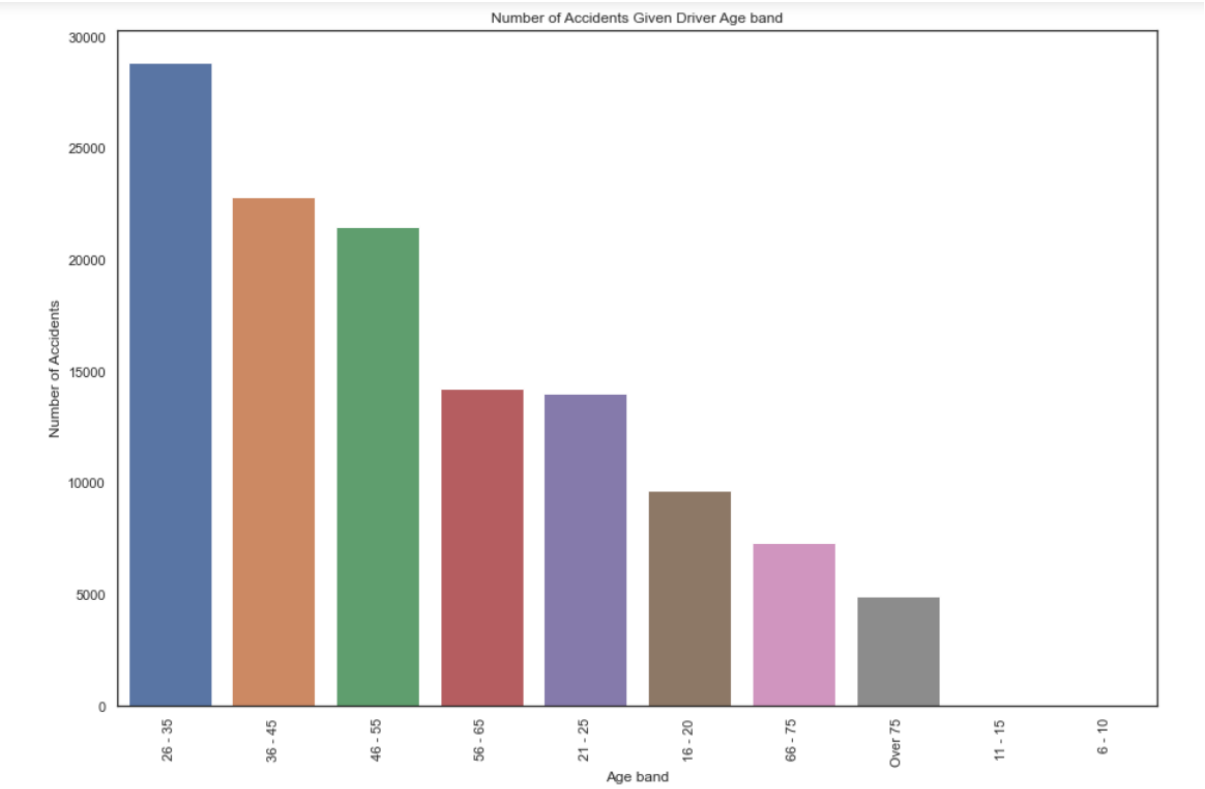

Speed_limit
20
30
40
50
60
70
30
40000
30000
20000
10000
0
Number of Accidents
Urban
Rural

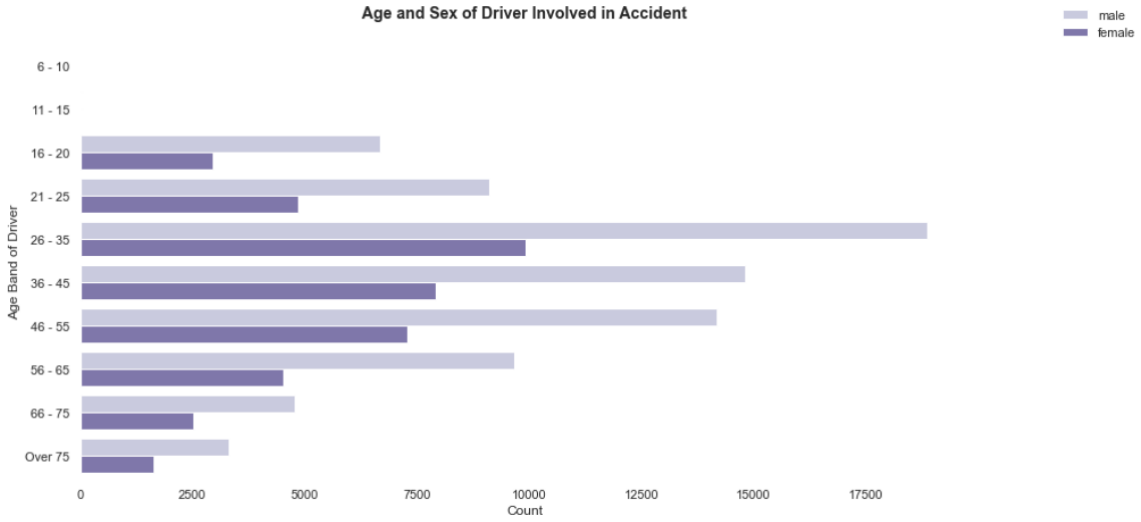


The graph and maps shows that more accidents happened in the urban cities than in rural areas but interestingly, it reveals that the severity of most of these accidents is slight as the speed is just 30mph whereas most accidents in the rural areas are fatal with a speed limit above 30mph.

H. How does driver related variables affect the outcome (e.g., age of the driver, and the purpose of the journey)?

AGE OF THE DRIVER





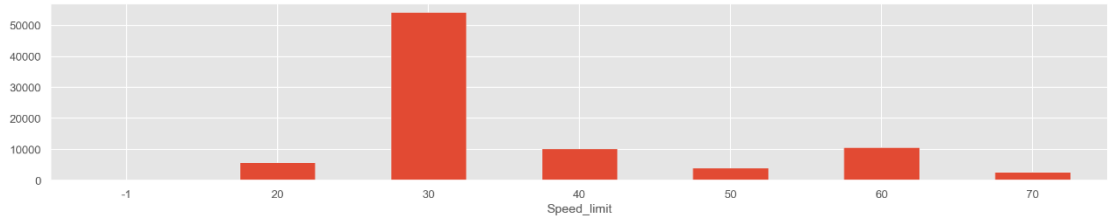
For a long time, it has been known that younger and older drivers are more likely to be involved in an accident(Gomes-Franco et al., 2020). In the report, the age band of 26-35 are prone to accidents while ages over 75 are least involved in an accident. Also, more males of different age bands are involved.

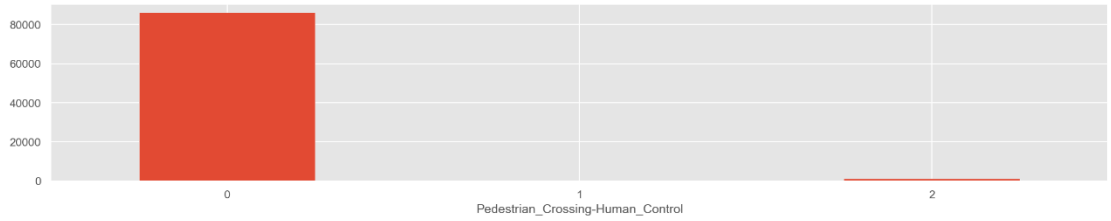
PURPOSE OF JOURNEY

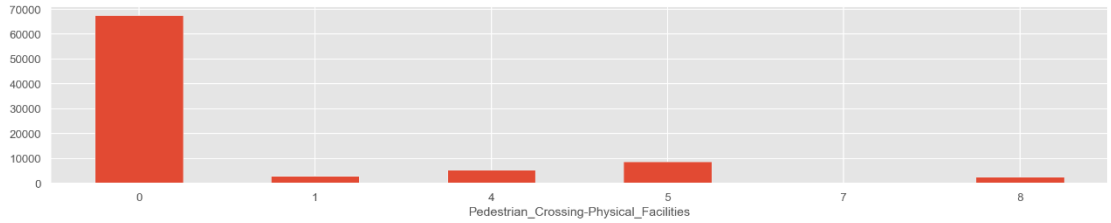
According to the data, the purpose of the drivers' journey in most of the incidents is unknown, but for all known purposes, the two most involved in an accident are 'commuting to/from work' and 'journey as part of work.'

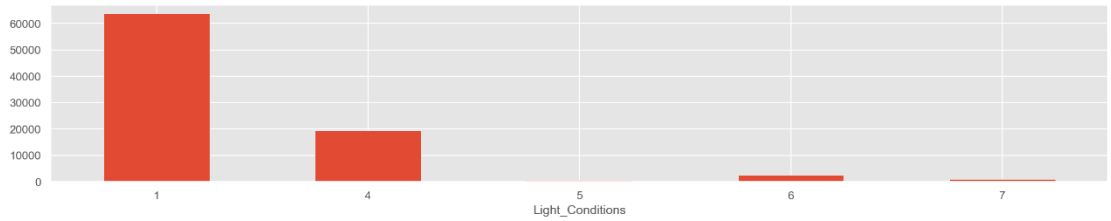
I. Can we make predictions about when and where accidents will occur, and the severity of the injuries sustained from the data supplied to improve road safety? How well do our models compare to government models?

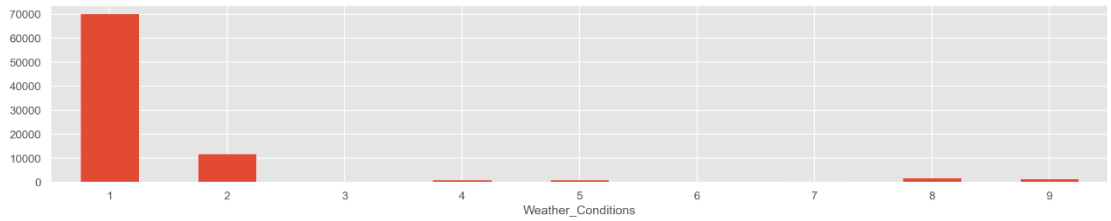
From analysis and as seen in the graphs below, most accidents happen when speed limit = 30, Ped Cross - Human = None within 50 metres, Ped Cross - Physical = No physical crossing facilities within 50 metres, light condition = daylight, weather condition = Fine no high winds, road surface condition = dry . The Heatmap below shows some correlations between attributes.

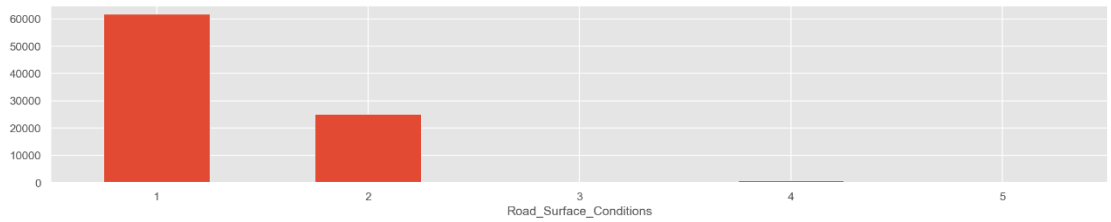

50000
40000
30000
20000
10000
0
-1
20
30
40
50
60
70
Speed_limit

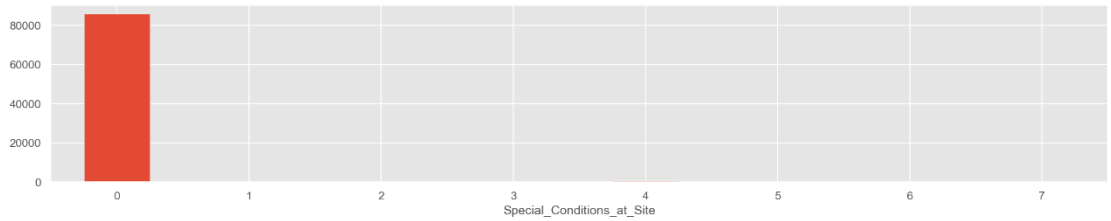

80000
60000
40000
20000
0
0
1
2
Pedestrian_Crossing-Human_Control

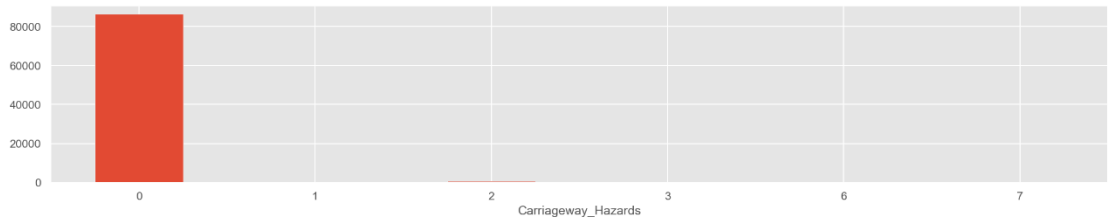

70000
60000
50000
40000
30000
20000
10000
0
0
1
4
5
7
8
Pedestrian_Crossing-Physical_Facilities

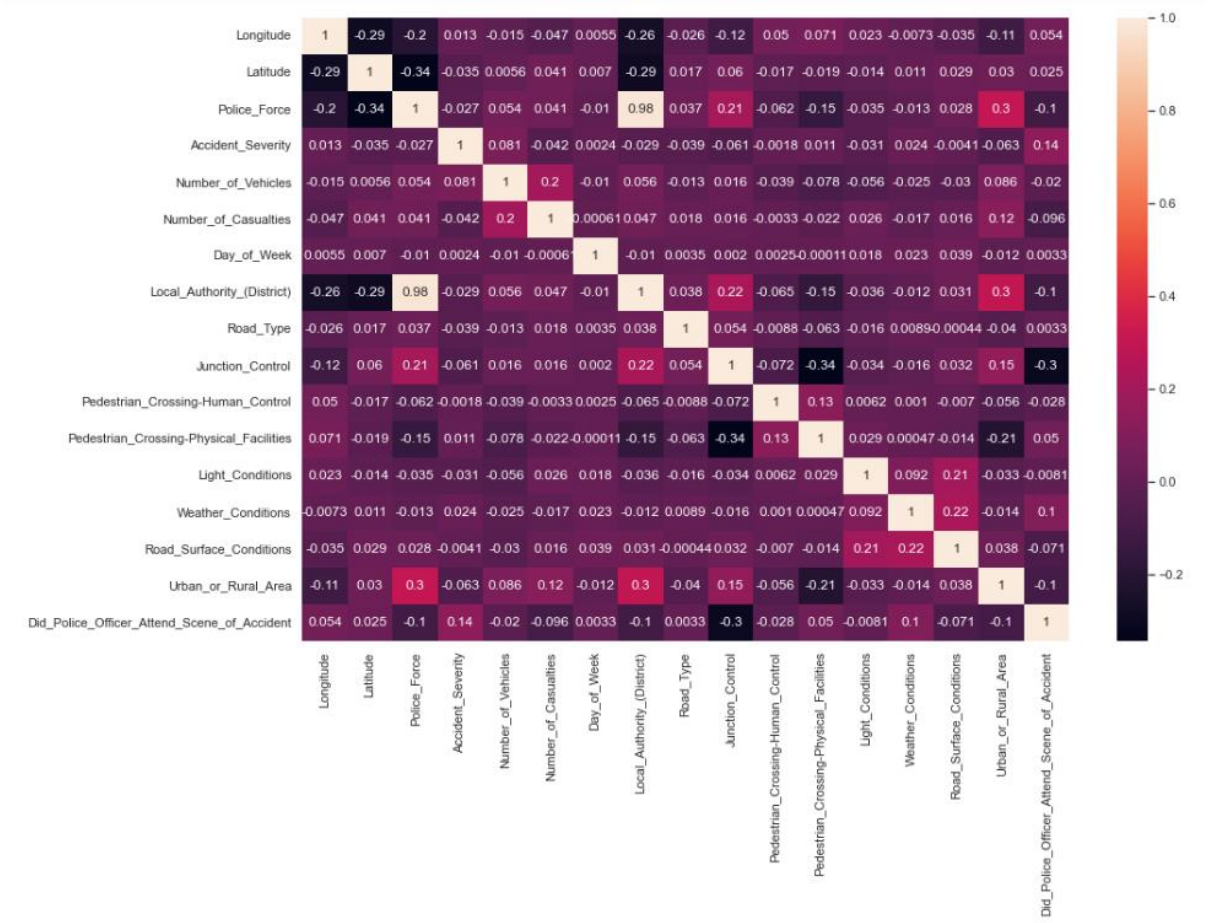

60000
50000
40000
30000
20000
10000
0
1
4
5
6
7
Light_Conditions

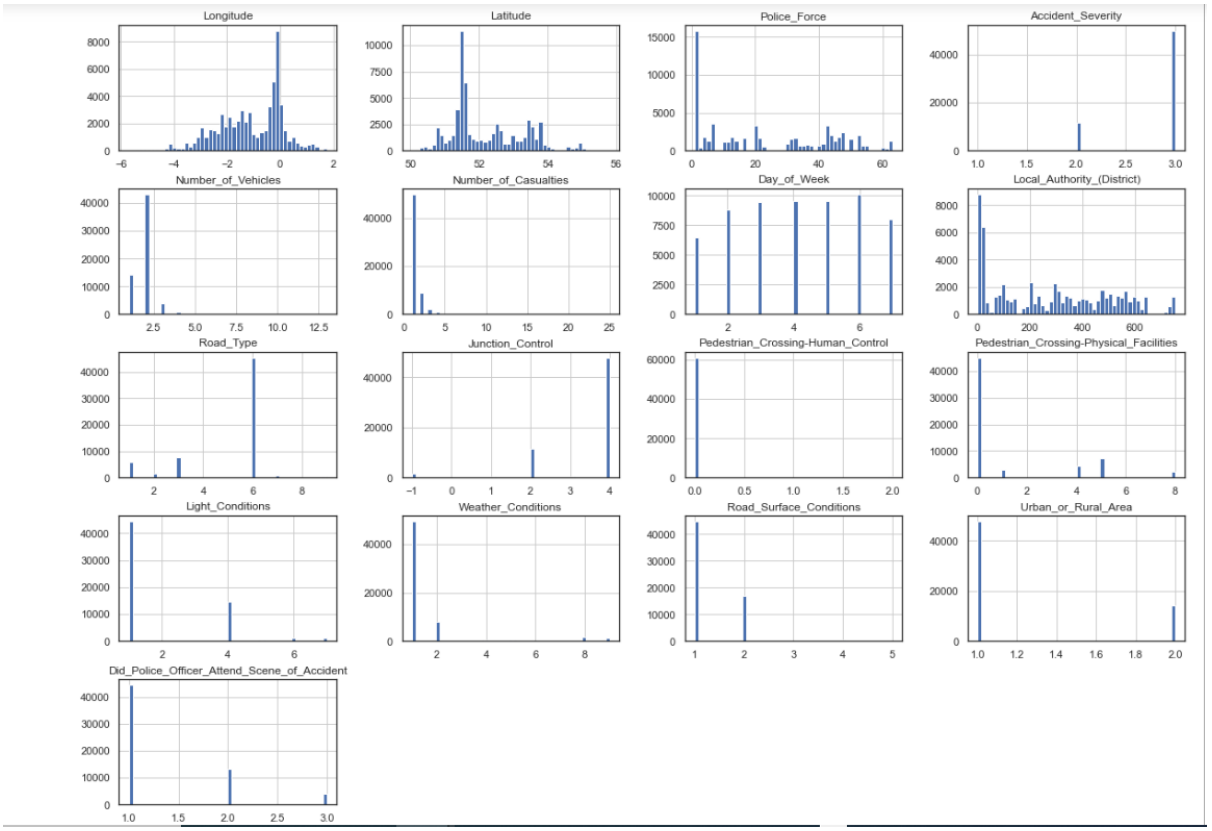

70000
60000
50000
40000
30000
20000
10000
0
1
2
3
4
5
6
7
8
9
Weather_Conditions


60000
50000
40000
30000
20000
10000
0
1
2
3
4
5
Road_Surface_Conditions


80000
60000
40000
20000
0
0
1
2
3
4
5
6
7
Special_Conditions_at_Site


80000
60000
40000
20000
0
0
1
2
3
6
7
Carriageway_Hazards


Longitude 1 -0.29 -0.2 0.013 -0.015 -0.047 0.0055 -0.26 -0.026 -0.12 0.05 0.071 0.023 -0.0073 -0.035 -0.11 0.054
Latitude -0.29 1 -0.34 -0.035 0.0056 0.041 0.007 -0.29 0.017 0.06 -0.017 -0.019 -0.014 0.011 0.029 0.03 0.025
Police_Force -0.2 -0.34 1 -0.027 0.054 0.041 -0.01 0.98 0.037 0.21 -0.062 -0.15 -0.035 -0.013 0.028 0.3 -0.1
Accident_Severity 0.013 -0.035 -0.027 1 0.081 -0.042 0.0024 -0.029 -0.039 -0.061 -0.0018 0.011 -0.031 0.024 -0.0041 -0.063 0.14
Number_of_Vehicles -0.015 0.0056 0.054 0.081 1 0.2 -0.01 0.056 -0.013 0.016 -0.039 -0.078 -0.056 -0.025 -0.03 0.086 -0.02
Number_of_Casualties -0.047 0.041 0.041 -0.042 0.2 1 0.00061 0.047 0.018 0.016 -0.0033 -0.022 0.026 -0.017 0.016 0.12 -0.096
Day_of_Week 0.0055 0.007 -0.01 0.0024 -0.01 -0.0006 1 -0.01 0.0035 0.002 0.0025 -0.00011 0.018 0.023 0.039 -0.012 0.0033
Local_Authority_(District) -0.26 -0.29 0.98 -0.029 0.056 0.047 -0.01 1 0.038 0.22 -0.065 -0.15 -0.036 -0.012 0.031 0.3 -0.1
Road_Type -0.026 0.017 0.037 -0.039 -0.013 0.018 0.0035 0.038 1 0.054 -0.0088 -0.063 -0.016 0.0089 -0.00044 -0.04 0.0033
Junction_Control -0.12 0.06 0.21 -0.061 0.016 0.016 0.002 0.22 0.054 1 -0.072 -0.34 -0.034 -0.016 0.032 0.15 -0.3
Pedestrian_Crossing-Human_Control 0.05 -0.017 -0.062 -0.0018 -0.039 -0.0033 0.0025 -0.065 -0.0088 -0.072 1 0.13 0.0062 0.001 -0.007 -0.056 -0.028
Pedestrian_Crossing-Physical_Facilities 0.071 -0.019 -0.15 0.011 -0.078 -0.022 -0.00011 -0.15 -0.063 -0.34 0.13 1 0.029 0.00047 -0.014 -0.21 0.05
Light_Conditions 0.023 -0.014 -0.035 -0.031 -0.056 0.026 0.018 -0.036 -0.016 -0.034 0.0062 0.029 1 0.092 0.21 -0.033 -0.0081
Weather_Conditions -0.0073 0.011 -0.013 0.024 -0.025 -0.017 0.023 -0.012 0.0089 -0.016 0.001 0.00047 0.092 1 0.22 -0.014 0.1
Road_Surface_Conditions -0.035 0.029 0.028 -0.0041 -0.03 0.016 0.039 0.031 -0.00044 0.032 -0.007 -0.014 0.21 0.22 1 0.038 -0.071
Urban_or_Rural_Area -0.11 0.03 0.3 -0.063 0.086 0.12 -0.012 0.3 -0.04 0.15 -0.056 -0.21 -0.033 -0.014 0.038 1 -0.1
Did_Police_Officer_Attend_Scene_of_Accident 0.054 0.025 -0.1 0.14 -0.02 -0.096 0.0033 -0.1 0.0033 -0.3 -0.028 0.05 -0.0081 0.1 -0.071 -0.1 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Longitude
Latitude
Police_Force
Accident_Severity
Number_of_Vehicles
Number_of_Casualties
Day_of_Week
Local_Authority_(District)
Road_Type
Junction_Control
Pedestrian_Crossing-Human_Control
Pedestrian_Crossing-Physical_Facilities
Light_Conditions
Weather_Conditions
Road_Surface_Conditions
Urban_or_Rural_Area
Did_Police_Officer_Attend_Scene_of_Accident

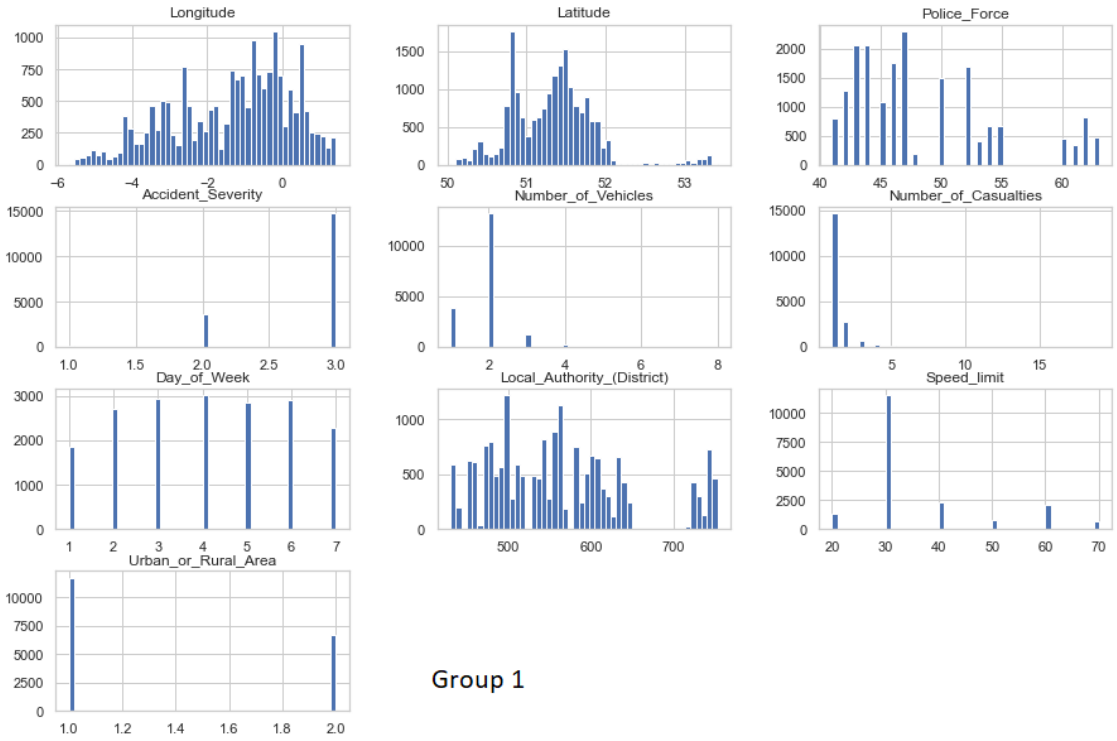

Longitude
Latitude
Police_Force
Accident_Severity
Number_of_Vehicles
Number_of_Casualties
Day_of_Week
Local_Authority_(District)
Road_Type
Junction_Control
Pedestrian_Crossing-Human_Control
Pedestrian_Crossing-Physical_Facilities
Light_Conditions
Weather_Conditions
Road_Surface_Conditions
Urban_or_Rural_Area
Did_Police_Officer_Attend_Scene_of_Accident

PREDICTION

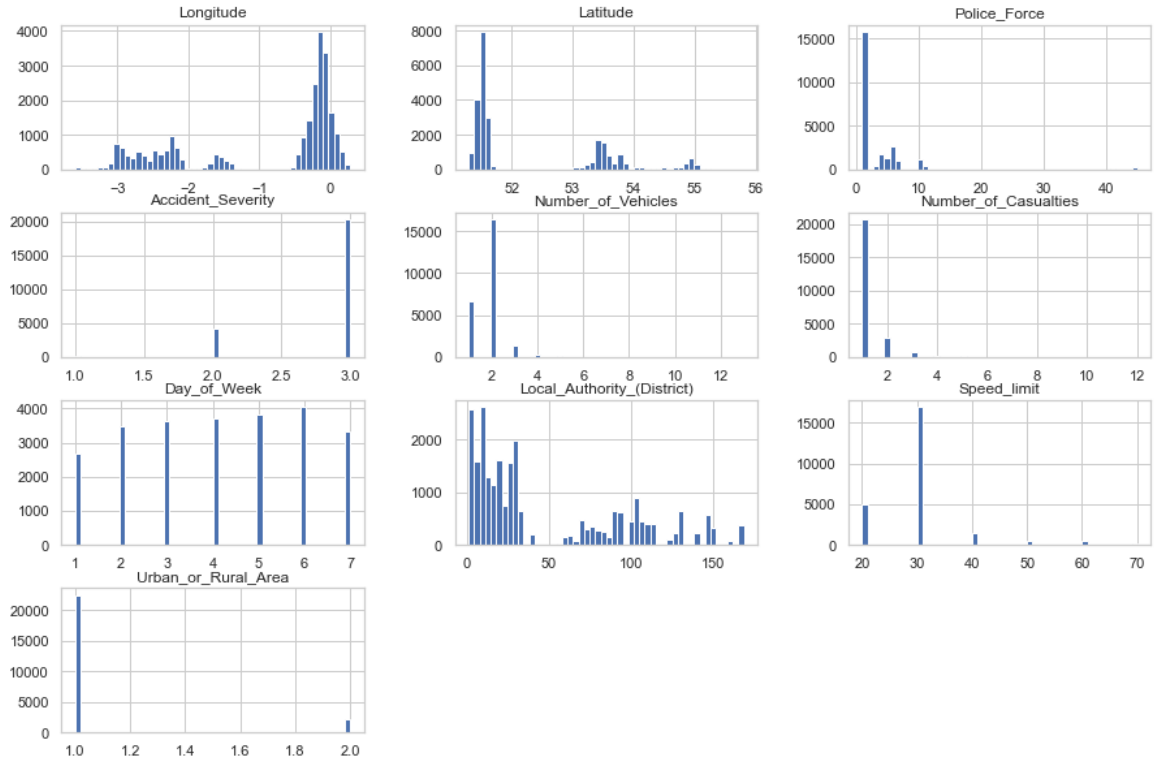
A prediction algorithm is a type of ML approach that analyses historical and current data to predict future behaviour.

K-Means clustering

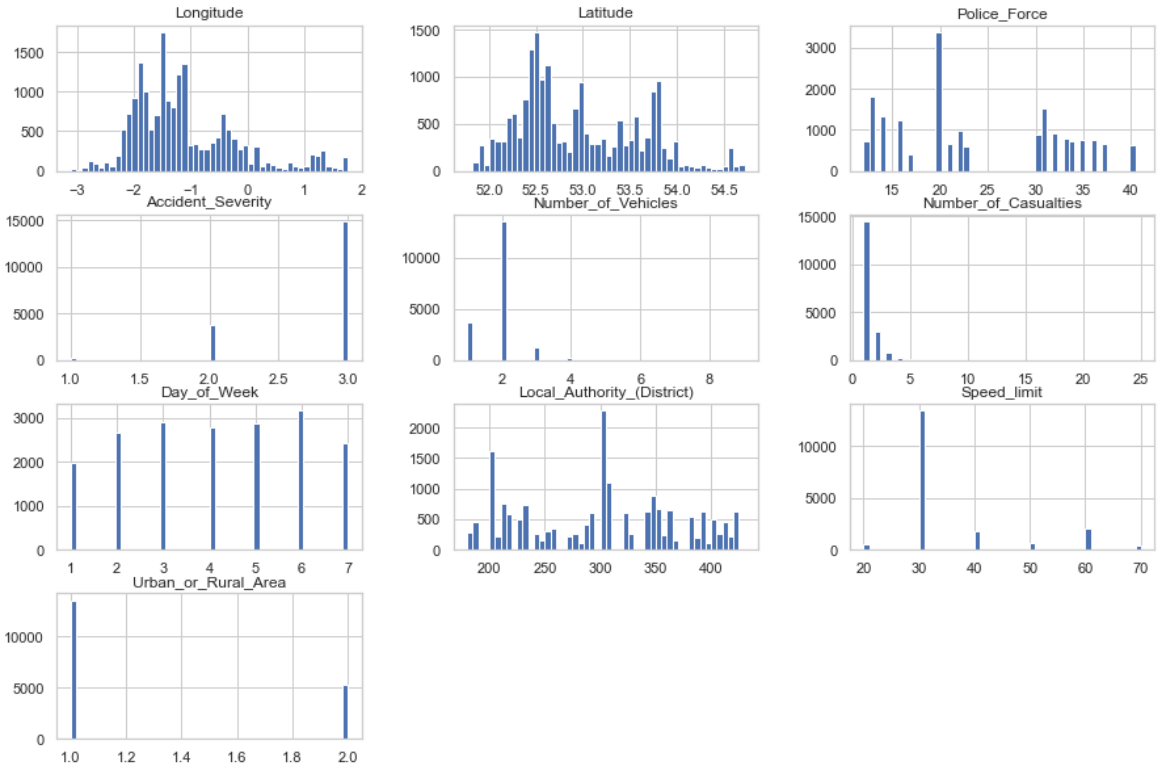
Heterogeneity in the data might lead to incorrect modelling and prediction(Kyriakopoulou & Kalamboukis, 2008). We are utilizing clustering to help us solve this challenge. In the database, we used cluster analysis to check the closest data to each other, allowing us to detect a pattern in the causes of the recorded accidents and perform trend analysis.



Group 1



GROUP 2



GROUP 3

Apriori Algorithm

We used the Apriori algorithm to generate useful rules from the data. For usage in the algorithm, the attributes in the dataset are inserted into an array. The amount of rules created is controlled by three parameters: support, confidence, and lift. The minimum support and confidence are set to 0.1 and 0.5, respectively, while the minimum lift and maximum lift are set to 2. We can examine how the values of various attributes are related and the rule for the occurrence of an accident in a specific circumstance. As a result, it becomes a complement to the K-mean. The table below summarizes some of the rules.

Rule: Saturday -> weekend	
-> Urban	
	Support: 0.10001935546307945
	Confidence: 0.7731920199501248
	Lift: 4.330294385986254
Rule: Saturday, 0 -> High	
-> weekend	
	Support: 0.10148714474660472
	Confidence: 0.7959519291587602
	Lift: 4.301929012639248
Rule: High, weekend -> Saturday	
-> 0	
	Support: 0.10148714474660472
	Confidence: 0.5485136430999913
	Lift: 4.301929012639248
Rule: Saturday, 0 -> High	
-> 1	
-> weekend	
	Support: 0.100212910093874
	Confidence: 0.7859582542694497
	Lift: 4.298883092033289
Rule: High, 1, weekend -> Saturday	
-> 0	
	Support: 0.100212910093874
	Confidence: 0.5481252756947508
	Lift: 4.298883092033289
Rule: Saturday -> High	
-> 0	
-> weekend	
	Support: 0.10148714474660472
	Confidence: 0.7845386533665836

	Confidence: 0.5557812914053528
	Lift: 4.296425000567216
Rule: Saturday -> High	
-> weekend	
	Support: 0.1028097680570341
	Confidence: 0.7947630922693267
	Lift: 4.295503634775845
Rule: High, weekend -> Saturday	
	Support: 0.1028097680570341
	Confidence: 0.5556621044372766
	Lift: 4.295503634775845

Association rules with high support

End of document

Preventive measures can be adopted, guided by the rules, to make faster choices and limit the likelihood of accidents as against the government model. For example, the rules (weather condition = Raining no high winds, light Condition = Darkness - lights lit, weekday = Saturday) Accident Severity = High suggest that fatal accident is more likely to occur on rainy night when the road lights are turned on. These associations are powerful rules that tell us where accidents will occur and how severe they will be.

KNN and Naïve Bayes Classifiers

Finally, we use predictive models (K-Nearest Neighbors and Naive Bayes) to train our database to forecast if a particular condition would result in a serious accident. The Bayes theorem is used to create a probabilistic classifier called Naive Bayes(Budiawan et al., 2019). It presupposes that variables are unrelated to one another. In KNN new data point are assigned a value based on how close it fits the features in the training set(Lv et al., 2009). We can see that the Naive Bayes method did better in categorization, and we can also infer that it is obviously possible to forecast the severity of an accident based on the scenario even with 79.41 percent accuracy.

Recommendation

It is vital for governments throughout the world to understand the key causes of road traffic accidents and the circumstances in which they occur to take action to reduce the number of fatalities because of these occurrences. The following guidelines can reduce the likelihood and severity of road accidents.

Control speed limit to keep it at a safe level.

Implement existing Laws and Regulations

Increase public awareness and education on road safety.

Accelerate emergency response and care

Design smarter Roads

References

Alkheder, S., Taamneh, M. & Taamneh, S. (2017) Severity prediction of traffic accident using an artificial neural network. *Journal of Forecasting,* 36 (1), 100-108.

Budiawan, W., Saptadi, S., Tjioe, C. & Phommachak, T. (2019) Traffic accident severity prediction using naive bayes algorithm-a case study of semarang toll road. *IOP Conference Series: Materials Science and Engineering.* IOP Publishing.

Gomes-Franco, K., Rivera-Izquierdo, M., Martín-delosReyes, L. M., Jiménez-Mejías, E. & Martínez-Ruiz, V. (2020) Explaining the association between driver's age and the risk of causing a road crash through mediation analysis. *International Journal of Environmental Research and Public Health,* 17 (23), 9041.

Kyriakopoulou, A. & Kalamboukis, T. (2008) Combining clustering with classification for spam detection in social bookmarking systems. *ECML PKDD.*

Lahti, T., Nysten, E., Haukka, J., Sulander, P. & Partonen, T. (2010) Daylight saving time transitions and road traffic accidents. *Journal of Environmental and Public Health,* 2010 .

Lv, Y., Tang, S. & Zhao, H. (2009) Real-time highway traffic accident prediction based on the k-nearest neighbor method. *2009 international conference on measuring technology and mechatronics automation.* IEEE.