771762 Big Data and Data Mining Project on Road Accident Prediction UK -2019

UNIVERSITY OF Hull

Uzair Ahmed Khan

Table of Contents

1. Introduction: 3
2. Methodology: 3
3. Exploratory Data Analysis: 4
   3.1. Accidents based on significant hours, days: 4
   3.2. Accidents based on Motorbikes: 5
   3.3. Pedestrian related accidents: 6
   3.4. Impact of Daylight Savings: 6
   3.5. Impact of Sunrise and Sunset: 7
   3.6. Vehicle Type involved in road accidents: 7
   3.7. Weather Conditions, Locations & Situational factors influencing accidents: 8
   3.8. Driver related variables affecting number of accidents: 9
4. Data Mining: 10
5. Predictions: 10
6. Comparison with government Models: 11
7. Recommendations: 12
8. References 13

1. Introduction:

The government agencies are analysing road accidents data to determine the best way to improve and develop road safety standards. All the accidents involving vehicles and causalities are logged in the United Kingdom. The government primarily identifies a report related with the probabilities of non-fatal accidents every year.

The purpose of the project is to make recommendations to government authorities on how to interpret the data and anticipate the number of accidents and injuries that caused in the year 2019. By building and estimating the model, the road safety can be established, and the upcoming circumstances of accidents can be avoided.

2. Methodology:

This section encompasses different process in providing better recommendations using the given dataset and the below flow diagram Figure 1 depicts the same. The major components as illustrated in the diagram below, are involved in the predictions of the outcome which is demonstrated by code.

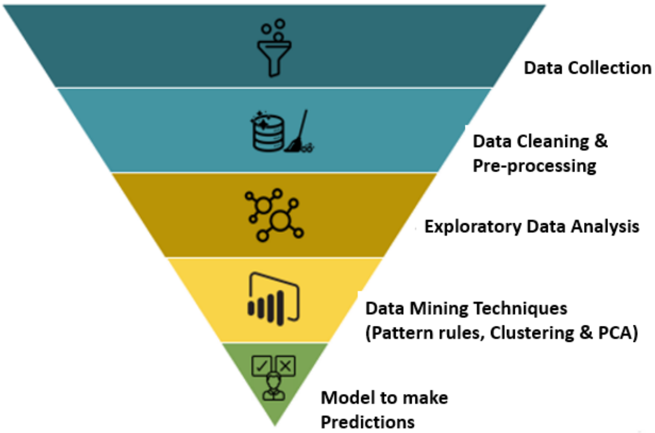
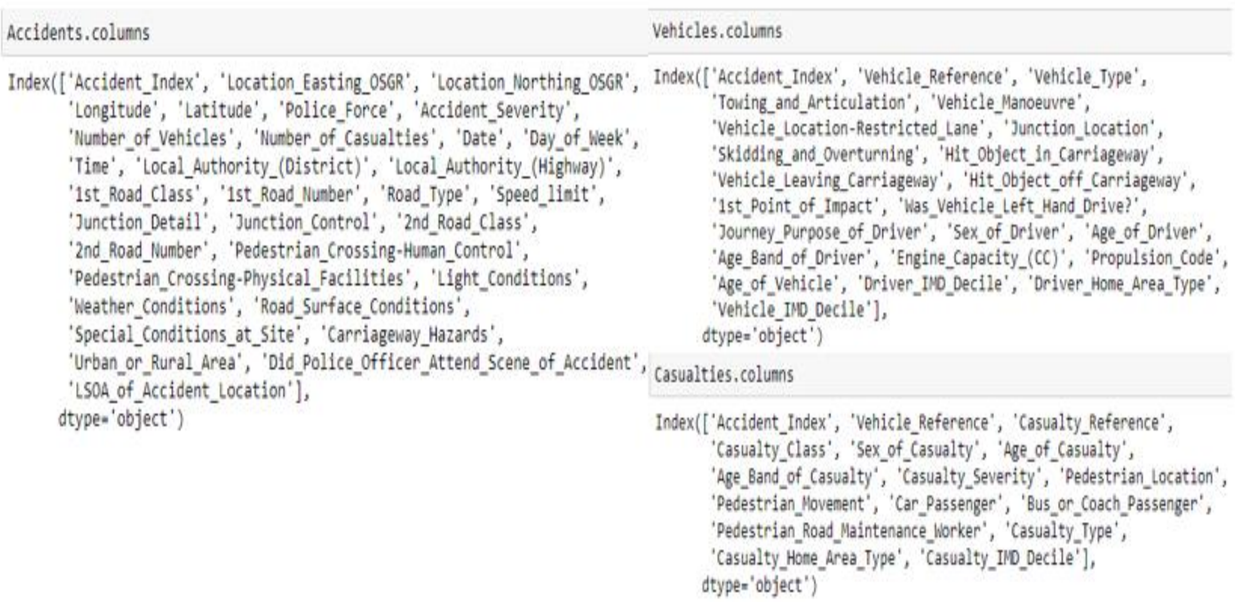

Data Collection
Data Cleaning & Pre-processing
Exploratory Data Analysis
Data Mining Techniques (Pattern rules, Clustering & PCA)
Model to make Predictions

Figure 1: Process Diagram Followed

A subset of accidents data along with the ancillary data from 2019 has been shared to perform necessary analysis over the data. The given dataset comprises of accidents, vehicles, and the casualty's data. The below Figure 2 represents the summary of the data shared. On observing the data, it contains inappropriate data. I aim to perform several data pre-processing techniques and methods to standardize the data.

As a part of data transformation, cleaned dataset is sorted and classified using group by function. After they have been transformed, detailed analysis and trends are carried out using the data mining approach. Findings and Discussion on various trends are provided in order to make a choice on this project investigation.



```
Accidents.columns
```

```
Index(['Accident_Index', 'Location_Easting_OSGR', 'Location_Northing_OSGR',
       'Longitude', 'Latitude', 'Police_Force', 'Accident_Severity',
       'Number_of_Vehicles', 'Number_of_Casualties', 'Date', 'Day_of_Week',
       'Time', 'Local_Authority_(District)', 'Local_Authority_(Highway)',
       '1st_Road_Class', '1st_Road_Number', 'Road_Type', 'Speed_limit',
       'Junction_Detail', 'Junction_Control', '2nd_Road_Class',
       '2nd_Road_Number', 'Pedestrian_Crossing-Human_Control',
       'Pedestrian_Crossing-Physical_Facilities', 'Light_Conditions',
       'Weather_Conditions', 'Road_Surface_Conditions',
       'Special_Conditions_at_Site', 'Carriageway_Hazards',
       'Urban_or_Rural_Area', 'Did_Police_Officer_Attend_Scene_of_Accident',
       'LSOA_of_Accident_Location'],
      dtype='object')
```

```
Vehicles.columns
```

```
Index(['Accident_Index', 'Vehicle_Reference', 'Vehicle_Type',
       'Towing_and_Articulation', 'Vehicle_Manoeuvre',
       'Vehicle_Location-Restricted_Lane', 'Junction_Location',
       'Skidding_and_Overturning', 'Hit_Object_in_Carriageway',
       'Vehicle_Leaving_Carriageway', 'Hit_Object_off_Carriageway',
       '1st_Point_of_Impact', 'Was_Vehicle_Left_Hand_Drive?',
       'Journey_Purpose_of_Driver', 'Sex_of_Driver', 'Age_of_Driver',
       'Age_Band_of_Driver', 'Engine_Capacity_(CC)', 'Propulsion_Code',
       'Age_of_Vehicle', 'Driver_IMD_Decile', 'Driver_Home_Area_Type',
       'Vehicle_IMD_Decile'],
      dtype='object')
```

```
Casualties.columns
```

```
Index(['Accident_Index', 'Vehicle_Reference', 'Casualty_Reference',
       'Casualty_Class', 'Sex_of_Casualty', 'Age_of_Casualty',
       'Age_Band_of_Casualty', 'Casualty_Severity', 'Pedestrian_Location',
       'Pedestrian_Movement', 'Car_Passenger', 'Bus_or_Coach_Passenger',
       'Pedestrian_Road_Maintenance_Worker', 'Casualty_Type',
       'Casualty_Home_Area_Type', 'Casualty_IMD_Decile'],
      dtype='object')
```

Figure 2: Summary of Accidents Dataset

3. Exploratory Data Analysis:

On examining the dataset, there are null values in the columns such as Latitude, Longitude, and Time. Local authority district of the mentioned null values is identified, and the mean latitude and longitude of those districts are calculated. Later, the null values are replaced with the obtained mean value. The total number of accidents in 2019 was 117536 where each accidents involved number of vehicles and casualties. Accident Severity column contains information such as slight, serious, and fatal accidents. Accident Index is the primary key in accidents dataframe and its is the foreign key in vehicles and causality dataframes.

3.1. Accidents based on significant hours, days:

Most of the accidents occur during the peak hours of the day. To identify the peak hours the date column is used to identify the hours, days and weeks of the accidents occurred during 2019. Below graphs are generated to quantify (in terms of standard deviations) the likelihood of the accident that are more in the evening rush hours than the rest of the day. Based on Figure 3 it is evident that most of the accidents occurs in the morning 8:00 hrs and evening 17:00 hrs where people tend to travel to and from school and work.

Figure 4 shows the Sunday and Friday peaks with number of accidents in the year and it can be seen that the 2019-04 – 3rd Sunday has the highest number of accidents among the year which is the Easter festive where people travel during the holidays. Also, the accidents count increased during September where lot of people travel during school/college break and Christmas in December.

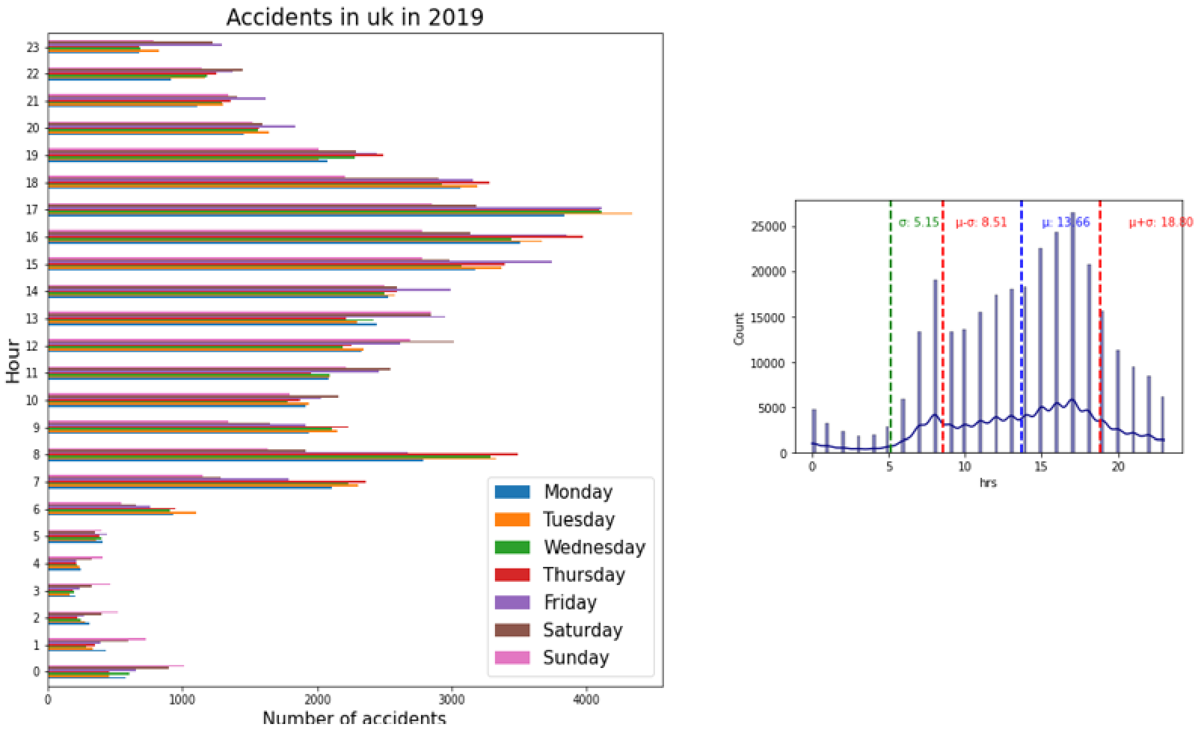


Figure 3: Number of accidents per hour in a day

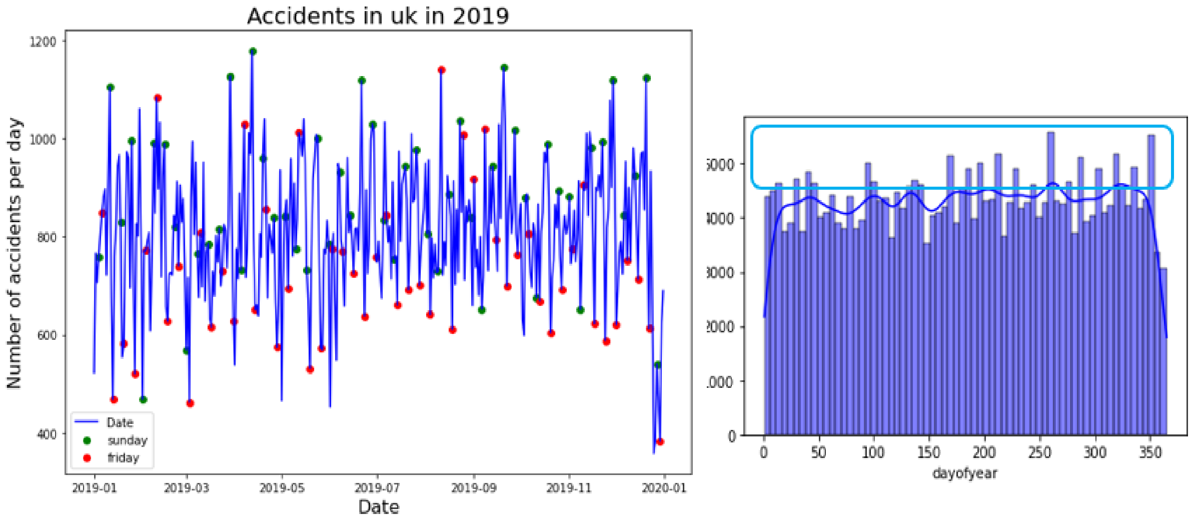


Figure 4: Number of accidents per day in a year

3.2. Accidents based on Motorbikes:

There are various motorbikes with varied specs, but older 125cc and 500c bikes are frequently engaged in accidents. As previously noted, motobike accidents also occur at peak hours and weekends such as morning 8:00 a.m. and evening 17:00 p.m – Fridays and Saturdays as shown in Figure 5. The majority of persons who own a motorcycle also own a car, and the most of them are men. During the summer, motorcycles are usually used during the months of May and June (UK, 2016).

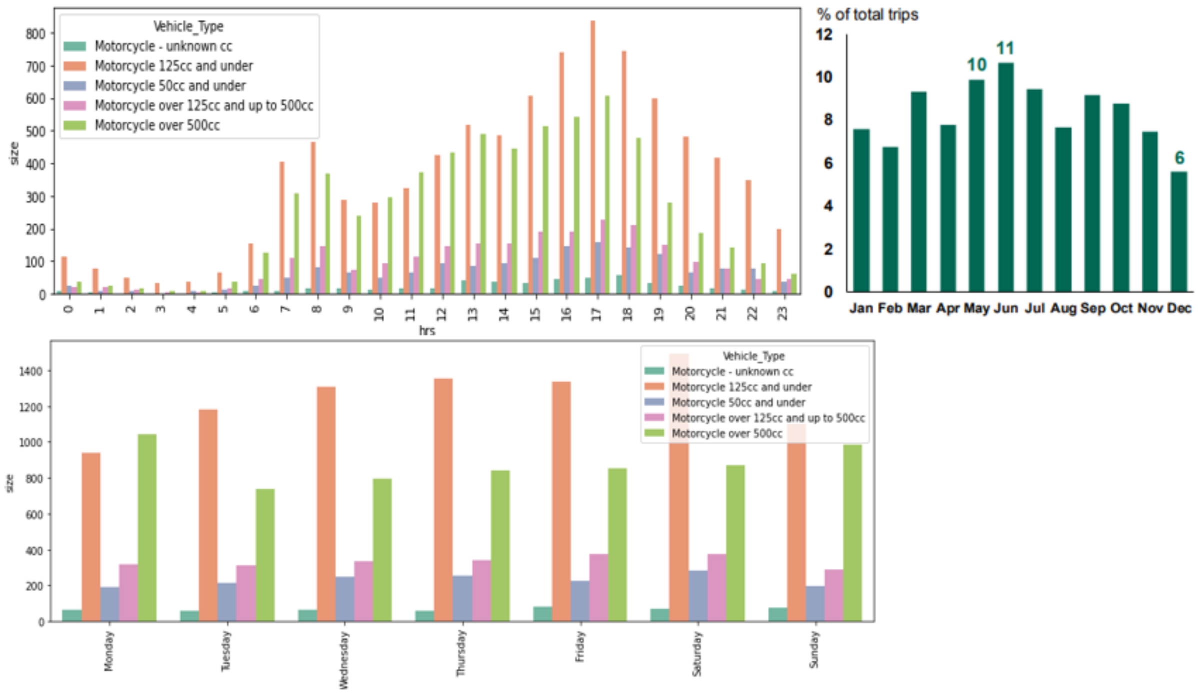


Figure 5: Significance of hours & days on Motorbikes

3.3. Pedestrian related accidents:

Pedestrians are one of the unprotected causality groups while crossing roads. As Figure 6 depicts most pedestrian accidents happened during the week, with a higher number on Saturday around 8 a.m. and 3 p.m. Young people walking to and from school and college would make up the majority of the casualties during this time.

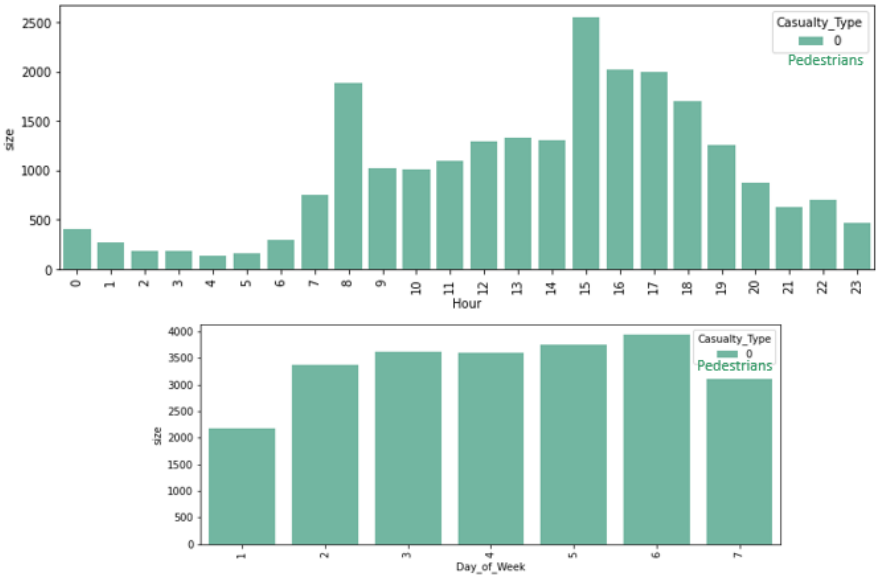


Figure 6:Pedestrians related to accidents

3.4. Impact of Daylight Savings:

The Impact of changes in the clock timings results in some hours of light being reduced at the evening when daylight savings ends (timeanddate.com, n.d.). The entire population have to align their activity with the early timing which results accidents in roads (Garnsey, 2009). in

Based on the Figure 7 it is observed that March and October month contribute to the one of the maximum accidents due to the changes in the daylights. Weeks 13 and 42 are the start and end of the daylight savings where accidents are most likely to occur.

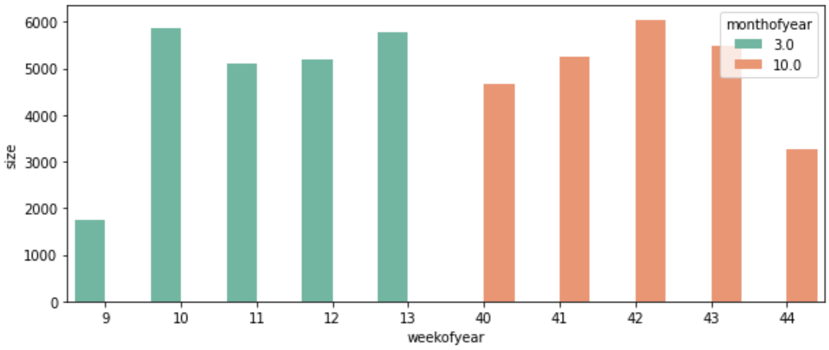


Figure 7: Number of accidents on the start and end weeks of daylight savings

3.5. Impact of Sunrise and Sunset:

Sunrise and sunset time vary per month. World data for sunrise and sunset in UK (WorldData, n.d.) was used to extract sunrise and sunset times into dataframe column for each month by creating a function, which were then added to the dataset to produce the visualisations as shown in Figure 8. Its is observed that maximum accidents occurs duing the daylight and sunset. There are more accidents at sunset than sunrise where daylight ends at different times causing chaos in traffic. (road-observatory, n.d.)

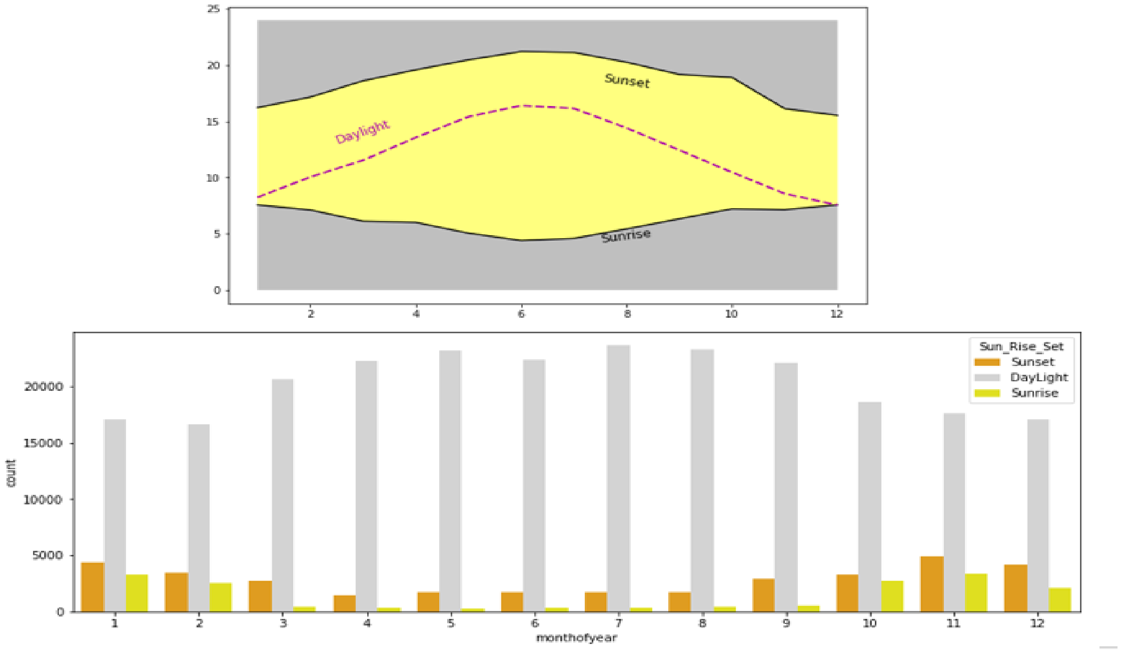


Figure 8: Number of accidents on sunrise and sunset for each month

3.6. Vehicle Type involved in road accidents:

Most of the vehicles involved in the accidents were cars as they are the most common means of transport by public to travel everyday. The majority of accident-causing vehicle types include Pedal Cycles and Goods/Vans. It's possible that there's a lack of awareness among pedal riders, resulting in accidents. Cars and Vans with engine capacity 1500 – 2000cc are facing more accidents. According to age and engine capacity, older taxis, private rentals, and long-distance motorcycles are involved in the majority of incidents, as illustrated in Figure 9.

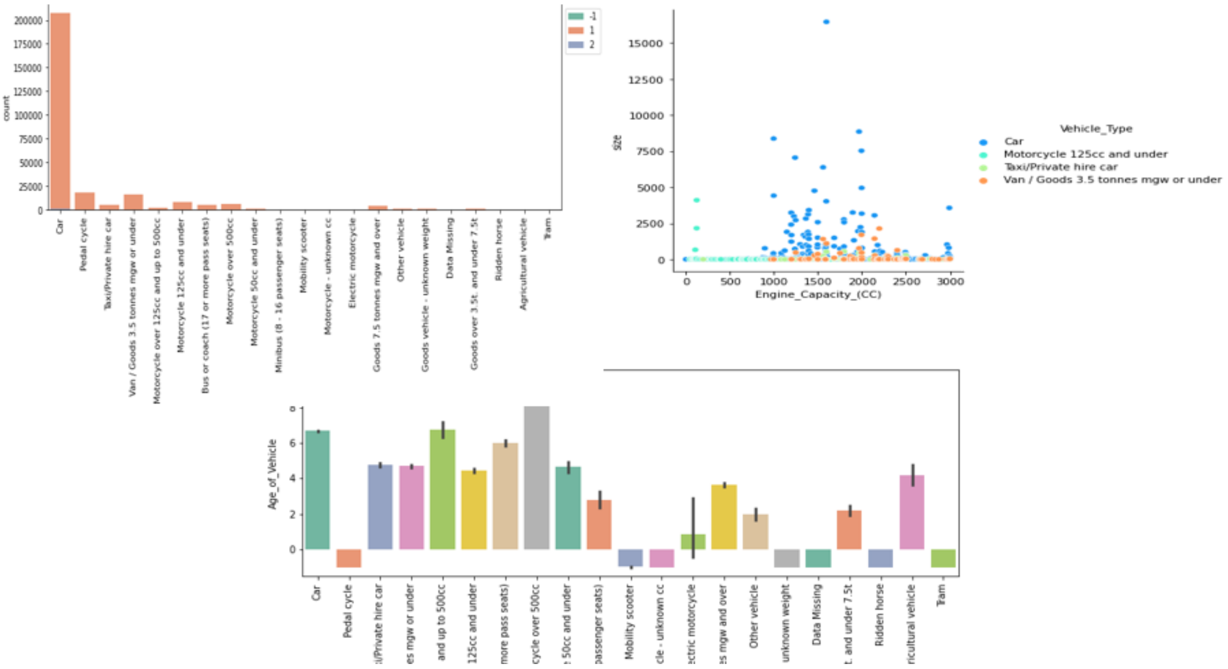


Figure 9: Vehicle type, age, and engine capacity

3.7. Weather Conditions, Locations & Situational factors influencing accidents:

Number of accidents are more during fine weather and rainy days involving slightly injured casualities than seriously injured and fatal. Use of clustering helps us identify accidents at various speed limit and weather conditions, from Figure 10 accidents occurs for all speed limits at fine weather and speed limit 60-70 at Foggy weather is seen. Number of clusters 5 is identified using elbow method where cluster inertia is fades out at that particular point. Using 3 features weather conditions, light conditions and accident severity in Kmeans 3d view of the clusters is plotted. Its evident that snow,fog weather conditions and no light conditions does not impact much with the accidents.

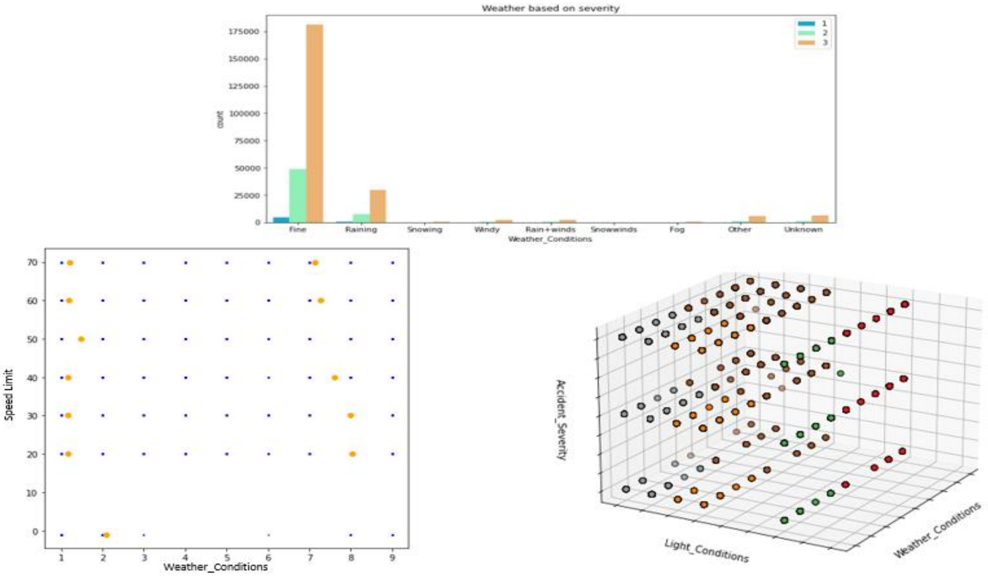


Figure 10: External Factors grouped using K-means clustering

As the number of columns in the dataset is more our model predictions might be poor. Princpal componen analysis (PCA) is one of the way to reduce the complexity of the features and find the feature having more variablity. As seen below the majority of the accidents are with Accident severity 3 (red) and Accident severity 2 (lime).

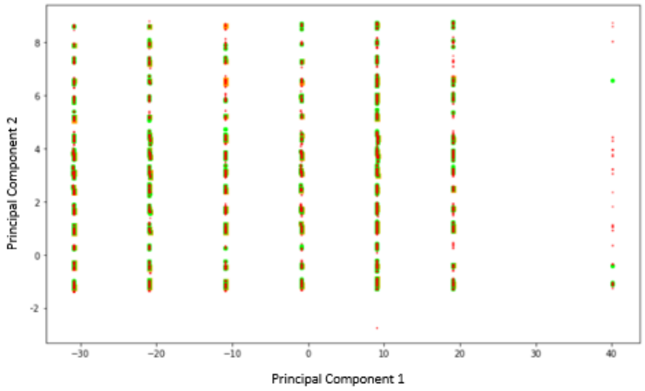


Figure 11: Principal Component Analysis for the features

On applying clustering techniques to the latitude and longitude of the dataset we were abled to identify the most accidents occuring at the denser part of the below shown map. The denser part of the map are areas around London at june being the highest accidents count, Birmingham with all months at higher accidents count. Major events such as cricket world cup 2019 held on June-July 2019 at London (wikipedia, n.d.) and local football match & FIA World Rally Championship made thounsands of fans go crazy at Birmingham which is also the second populated city in england (wrc, n.d.) where more precautions should be taken in road safety during these occasions.

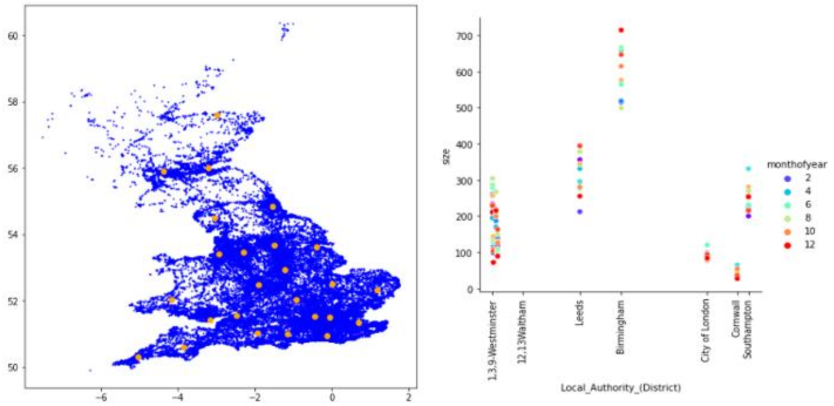


Figure 12: UK accidents based on locations using K-Means

3.8. Driver related variables affecting number of accidents:

Based on the Figure 13, we can see that the majority of the accidents are caused by drivers aged 30-50 who are travelling to work located in cities. There have been a few incidents involving school-aged pupils and persons responsible for dropping them to school.

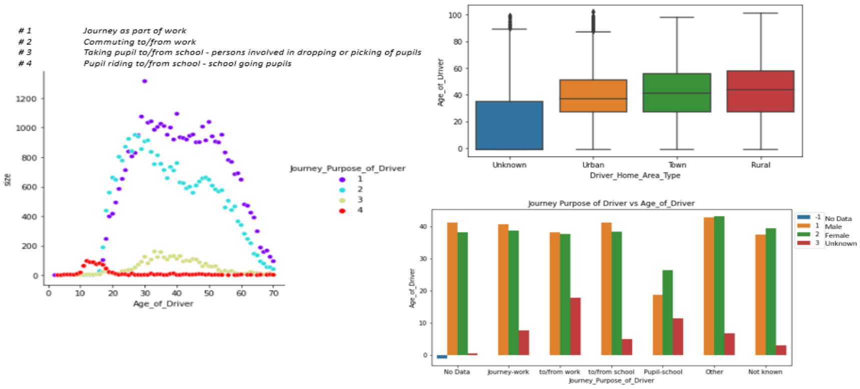
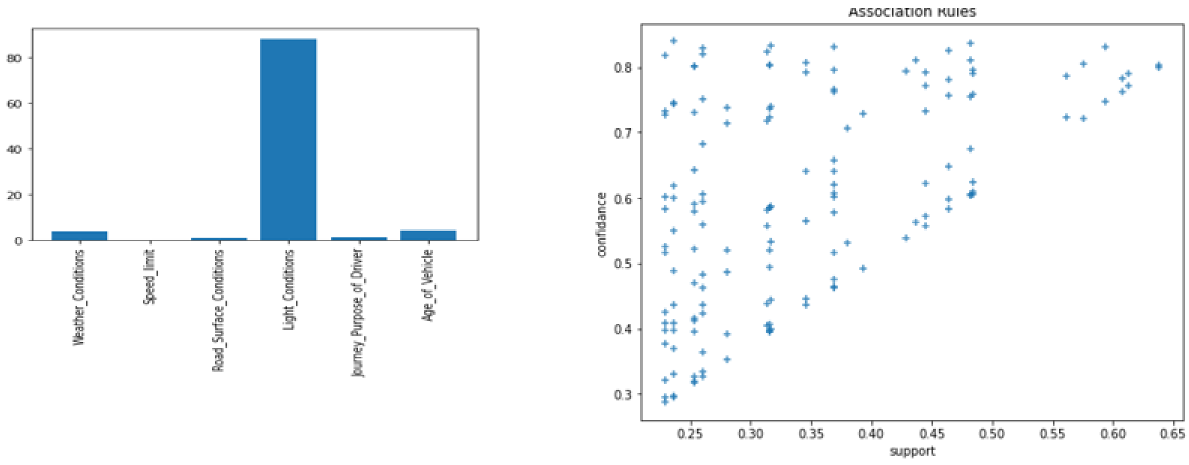
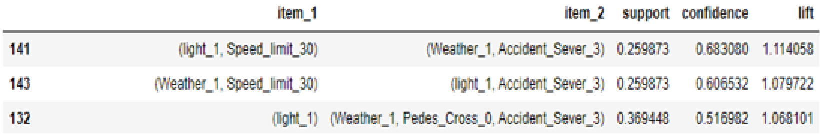


Figure 13: Driver related variables

4. Data Mining:

In this study, data mining techniques are used to investigate the relationship between various patterns found in accident data. The K best feature extraction method is first used to extract the most significant features from a complex dataset which are used in further analysis. One of the pattern discovery strategies used in data mining is apriori. Where the support and confidence of each association pattern are identified using association rules. The most common occurrence is indicated by sorting the lift value >1 and confidence >0.5. The resulting pattern identified are the accident severity-3 associated with light condition-1 (daylight) with fine weather conditions and speed limit-30.



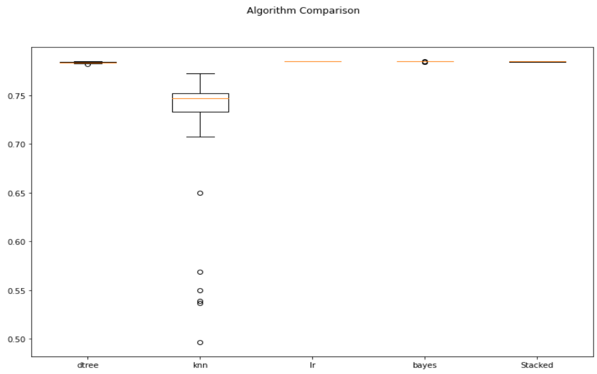


	item_1	item_2	support	confidence	lift
141	(light_1, Speed_limit_30)	(Weather_1, Accident_Sever_3)	0.259873	0.683080	1.114058
143	(Weather_1, Speed_limit_30)	(light_1, Accident_Sever_3)	0.259873	0.606532	1.079722
132	(light_1)	(Weather_1, Pedes_Cross_0, Accident_Sever_3)	0.369448	0.516982	1.068101

Figure 14: Data & Pattern Mining

5. Predictions:

In this part, Machine learning algorithms are applied to the obtained features to provide predictions. The performance of various models is evaluated using accuracy metrics as stated in the table below to select the optimal model. Because the dataset was unbalanced, there were some anomalies, therefore I tried scaling the dataset and applied the best model.

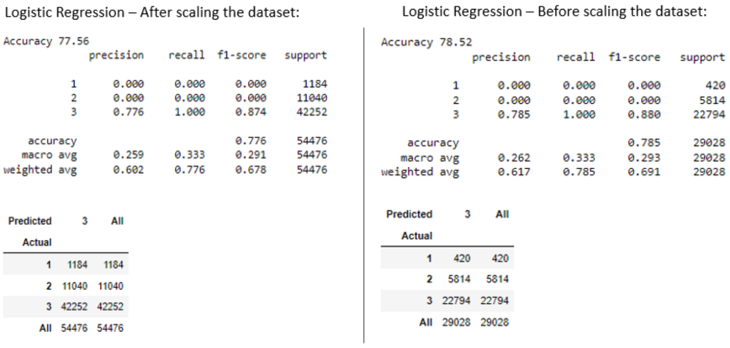


Box plot for applied algorithms

Model	Accuracy
Decision Tree Classifier	78%
K-Neighbors Classifier	71%
Logistic Regression	78.5%
Gaussian NB	78.5%
Stacked- (all mentioned models)	78.5%

Table 1: Model Performance

Considering Logistic Regression as the best model and evaluating it using other metrics such as precision with 77%, recall with 100% which overfits the model and f1 score with 87%. Also, the model predicted the accident severity as ‘3’ based on the conditions passed. We couldn't notice much of a difference in the dataset's accuracy after using scaling methods (Standard Scalar).



Logistic Regression – After scaling the dataset:

```
Accuracy 77.56
              precision    recall  f1-score   support

           1      0.000     0.000     0.000      1184
           2      0.000     0.000     0.000     11040
           3      0.776     1.000     0.874     42252

    accuracy                          0.776     54476
   macro avg      0.259     0.333     0.291     54476
weighted avg      0.602     0.776     0.678     54476
```

Predicted	3	All
Actual		
1	1184	1184
2	11040	11040
3	42252	42252
All	54476	54476

Logistic Regression – Before scaling the dataset:

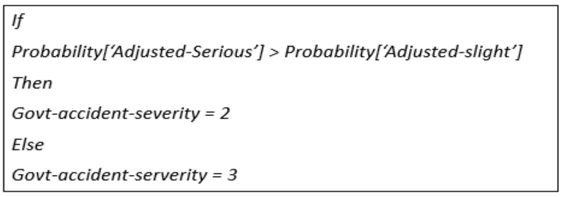
```
Accuracy 78.52
              precision    recall  f1-score   support

           1      0.000     0.000     0.000       420
           2      0.000     0.000     0.000      5814
           3      0.785     1.000     0.880     22794

    accuracy                          0.785     29028
   macro avg      0.262     0.333     0.293     29028
weighted avg      0.617     0.785     0.691     29028
```

Predicted	3	All
Actual		
1	420	420
2	5814	5814
3	22794	22794
All	29028	29028

Figure 15: Model Performance Comparison

6. Comparison with government Models:

The given cas adjustment dataset contains probabilities of different injuries occurring for each accident over past years. New column govt_accident_severity is added to cas dataset by values 2 & 3 Equation 1. Fatal accidents are excluded from the cas data and the given dataset which are then merged using accident index. On analysing the data, there is a significant decrease in the number of incidents that occurred each year, as shown in Figure 16. The government model's accuracy is calculated, and it turns out to be higher than the predicted accuracy of our model. As a result, the government model appears to outperform ours.



```
If
Probability['Adjusted-Serious'] > Probability['Adjusted-slight']
Then
Govt-accident-severity = 2
Else
Govt-accident-serverity = 3
```

Equation 1: Probability calculation for govt accident severity

Model	**Accuracy**
Our Model	78.5%
Government Model	88.3%

Table 2 : Model Comparison

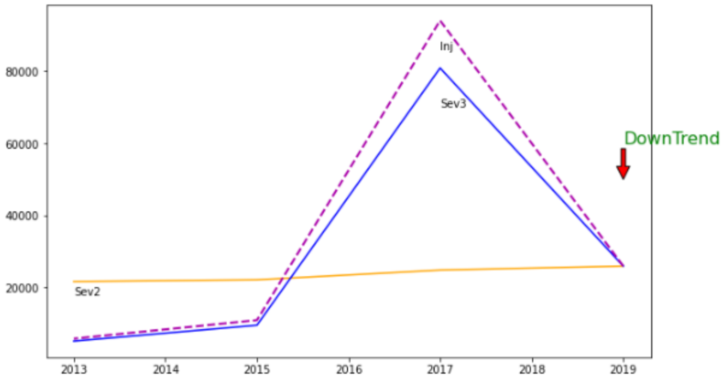


Figure 16: Number of accidents each year based on government data

7. Recommendations:

The United Kingdom has one of the best road safety management systems in the world. There is a downward trend in the number of injuries based on the figures from the previous year's accidents. The following is a list of suggestions that can be followed.

- Sampling techniques can be applied on the provided dataset to improve accuracy.
- The government can raise cyclist awareness and develop policies to reduce accidents involving cyclists and pedestrians.
- CrashMap is one of the applications by the government of UK that gives us access to maps that show us nearby accidents and traffic so that we may avoid traffic and additional accidents (Govt, n.d.).
- Create awareness of technologies and application used in road safety and traffic.
- Regular vehicle health check to be done by individuals as age of vehicle act as one of the criteria causing accidents.
- School/College safety zone to be formed at those areas as they are more likely to be affected as per the analysis done.
- Speed limit checks at motorways and highways to be monitored.

8. References

Garnsey, B. C. a. E., 2009. *Daylight Saving in GB:Is there evidence in favour of clock time on GMT,* University of Cambridge: s.n.

Govt, U., s.d. *crashmap uk.* [Online]
Available at: https://www.crashmap.co.uk/

road-observatory, n.d. *www.rospa.com.* [Online]
Available at: https://www.rospa.com/media/documents/road-safety/road-observatory/Other-Daylight-hours.pdf
[Accessed 2 4 2022].

timeanddate.com, s.d. *Daylight Savings 2019.* [Online]
Available at: Daylight Saving Time 2019 in the United Kingdom (timeanddate.com)

UK, N. T. S., 2016. *National Travel Survey UK.* [Online]
Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/694965/motorcycle-use-in-england.pdf
[Accessed 13 4 2022].

wikipedia, s.d. *WorldCup2019.* [Online]
Available at: https://en.wikipedia.org/wiki/2019_Cricket_World_Cup

WorldData, s.d. *WorldData.* [Online]
Available at: https://www.worlddata.info/europe/united-kingdom/sunset.php

wrc, s.d. [Online]
Available at: https://www.wrc.com/en/news/news-archive/wrc/birmingham-blast-off-for-2019-wrc/

Softwaretestinghelp.com. (2019). Apriori Algorithm in Data Mining: Implementation With Examples. [online] Available at: https://www.softwaretestinghelp.com/apriori-algorithm/.

Ng, A. (2016). Association Rules and the Apriori Algorithm: A Tutorial. [online] Kdnuggets.com. Available at: https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html.

Analytics Vidhya. (2020). Feature Scaling | Standardization Vs Normalization. [online] Available at: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/.