## Introduction

The United Kingdom Road traffic accident dataset is a dataset that records events and occurrences of accidents in the UK. It consists of 4 tables (accident, casualty, vehicle and lsoa) with over 400,000 records. My goal is to analyze the dataset to:

1. Understand the causes or contributing factors of road accidents, their time and frequency of occurrence.
2. Assess the severity of road accidents based and the responsible factors.
3. Develop a classification model that accurately predicts fatal injuries sustained in road traffic accidents.
4. Provide recommendations to government to aid accident prevention and traffic management.

## Data Analysis

The goal of this analysis is to analyze and provide insights from road traffic accidents in the year 2020. SQL Queries was performed to extract the 3 relevant tables namely: Accident, Casualty and Vehicle tables. Analysis into the dataset is outlined under the below subsections.

- **Data Cleaning**:

Data cleaning was done in two phases. The first phase was to identify and clean the missing values. 14 missing values were identified in 4 columns in the accident table namely: location_easting_osgr, location_northing_osgr, longitude and latitude. The data cleaning involved converting the longitude and latitude columns from string type to float to ensure columns are captured as numeric columns. A function was created to fill all the missing values with the mean as these columns may be useful to my future analysis. The second phase of the cleaning was performed on the selected columns after the feature selection process. The stats 19 and 20 forms aided the discovery of invalid entries (-1, 99 ,9) in the following columns: speed_limit, junction_control second_road_class, age_of_casualty, age_band_of_casualty, vehicle_location_restricted_lane, junction_location hit_object_in_carriageway. The continuous variables or columns (age_of_casualty, age_band_of_casualty, speed_limit) were cleaned using a function by replacing the invalid entries with their median value. The choice between using mean and median depends on the distribution of the data and the presence of outliers (Statology, 2021). Outliers will be discussed extensively in its subsection. Furthermore, a function was created to clean the invalid entries in the categorical columns. This was achieved by dividing the all the records for the invalid entries equally and assigning them equally to the valid categories to maintain balance in the dataset.

To aid the development of my classification model, 3 tables (accident, casualty, and vehicle) were merged into one. All duplicated records that emerged from this process were dropped.

- **Accident Occurrence, Day and Time:**

The below plots (fig 1, fig 2) show the significant hours of the days and days of the week in which accidents occur in the UK in the year 2020.

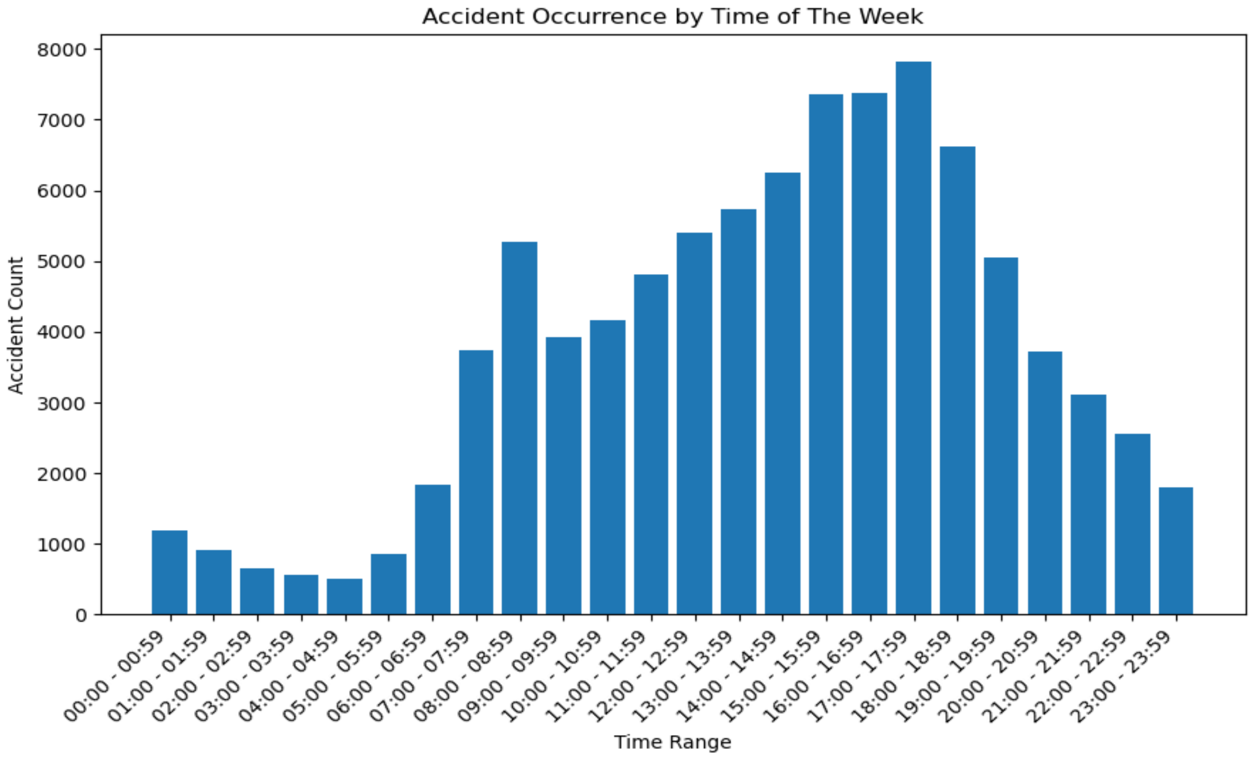


Fig 1

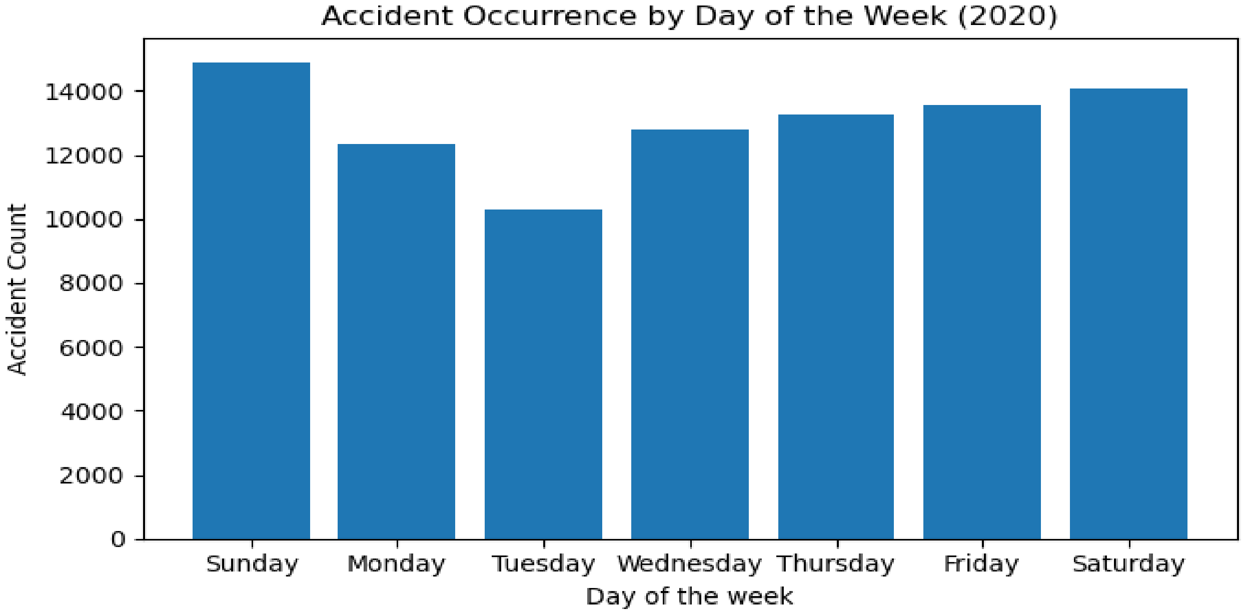


Fig 2

In the year 2020, the most significant number of accidents was recorded on day 1 (14,889), day 7 (14,056) and day 6 (13,564) representing Sunday, Saturday, and Friday respectively. These occurrences fall within the time range of 15:00 – 19:00 hours. These days are typical weekend period that witnesses increased level of recreational activities such as partying, drinking, travelling, reckless driving etc. The Cambridge Dictionary (n.d.) defines the term "weekend" as the period comprising Friday evening, Saturday, and Sunday. During this period, many individuals are not engaged in work activities.

Motorcycle accidents in the year 2020 exhibit a comparable pattern. The following visualizations depict the days and specific time periods during which accidents occur for the four motorcycle categories.

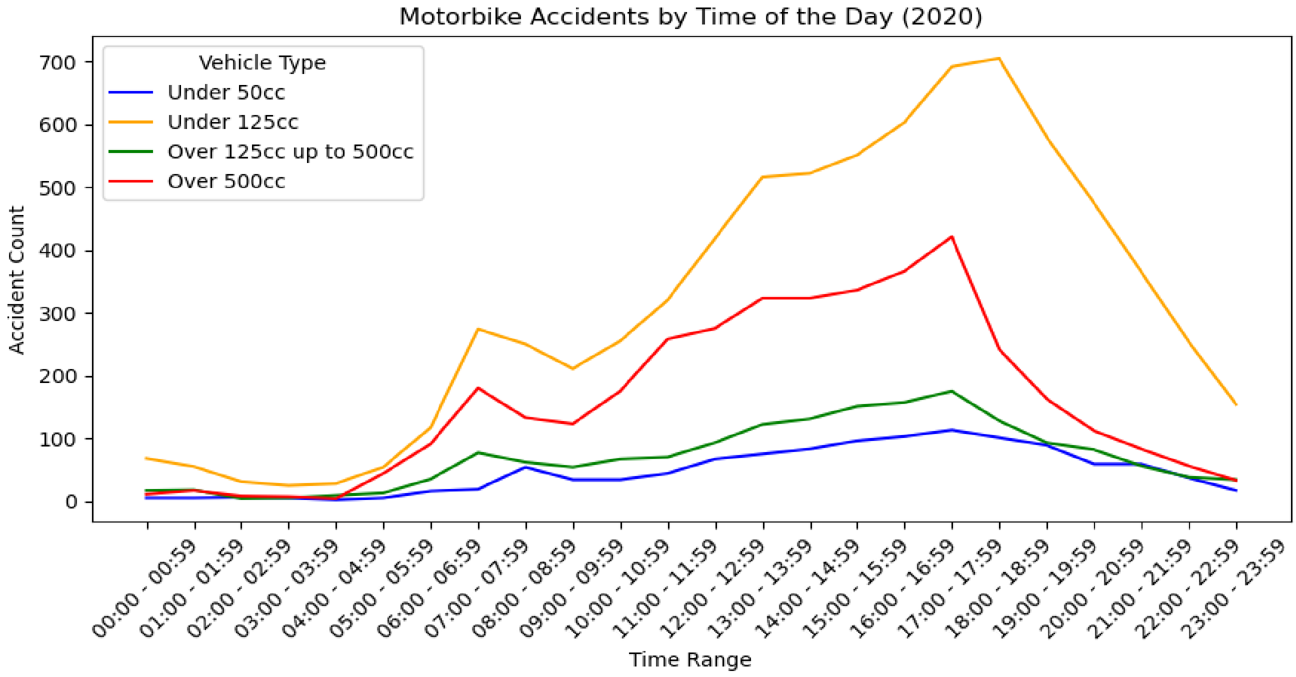


Fig 3

The plot illustrates a notable decrease in accident occurrences across all four motorcycle categories during the early morning hours until around 05:00 AM, after which accidents start to rise as human activities for the day commence, and then gradually decline as the day comes to an end. The highest incidence or rate of occurrence of these incidents are recorded within the time range of 15:00 – 19:00 hours. These are rush hour periods where people are leaving their workplaces and heading home, they are tired for the day, distracted and tend to lose mental alertness. This could also be attributable to reckless driving and increased leisure activities during the weekends. Furthermore, the chart illustrates a notable trend wherein motorcycles falling within the "over 50cc and up to 125cc" (under 125cc) category exhibit the highest frequency of accidents, reaching a peak count of 705 incidents. Following closely is the "over 500cc" motorcycle type, registering 421 accidents. On the use of the two motorcycle types, immediate action is required by the government on the former while the latter should be closely monitored.

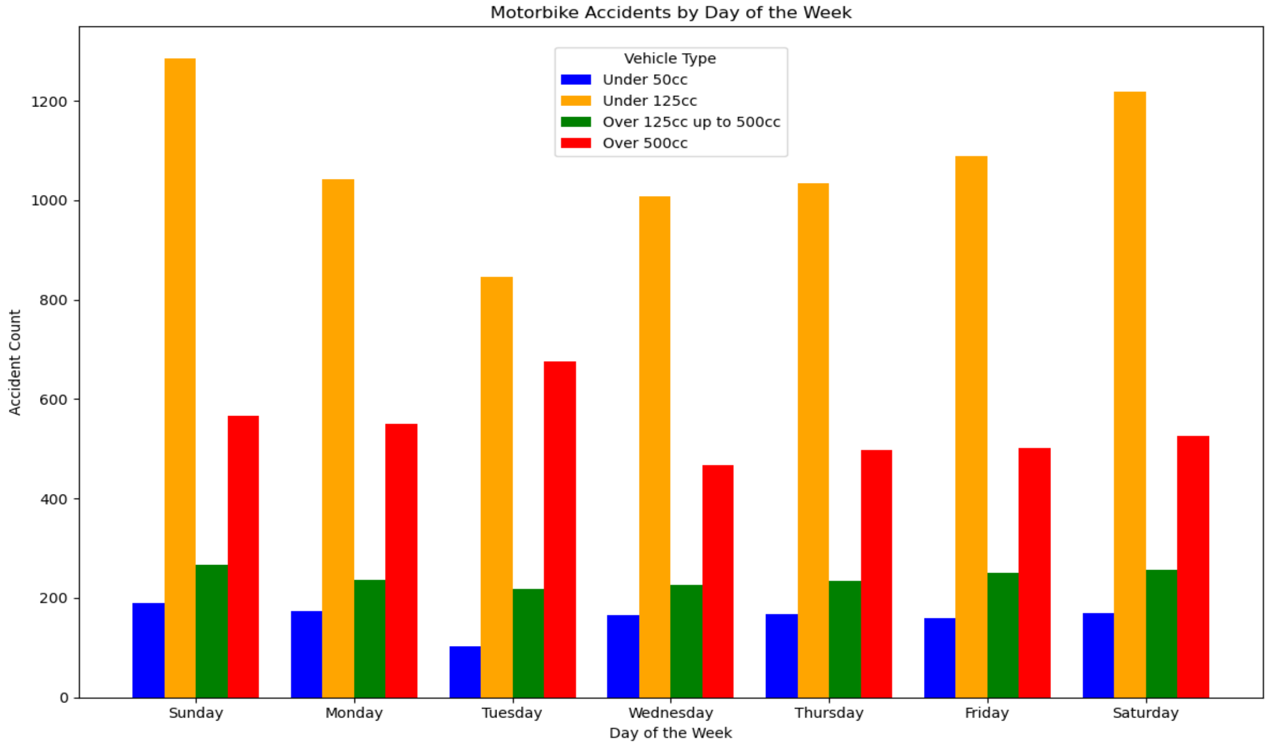


Fig 4

The multi-bar plots in fig 4 illustrate that Sunday (day 1), Saturday (day 7), and Friday (day 6) exhibit the most notable frequency of motorcycle accidents. This observation aligns with the high incidence or trend observed in the two aforementioned motorcycle categories. These patterns can also be attributed to increased recreational activities during the weekends.

Pedestrian accidents also maintain similar trends. The below plots show their time and incidence of occurrence. Fig 5 and fig 6 plots below show pedestrians are involved in more accidents during the weekends (Sunday, Friday, Saturday) from the time range of 15:00 – 19:00 hours. The possible factors contributing to these include increased relaxation activities, higher foot traffic in shopping areas and entertainment centers, display of recklessness and carefree attitude of some road users (Car, Motorcycle) during these periods.

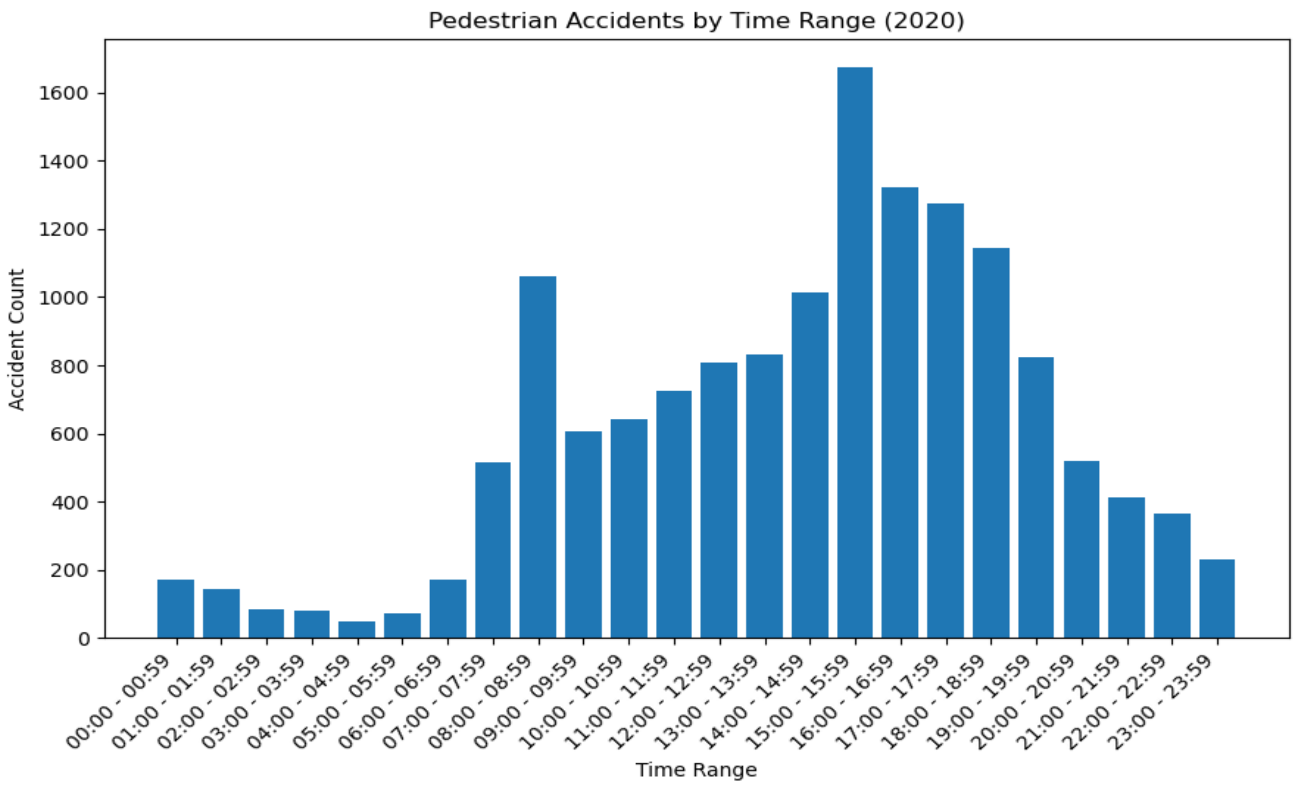

Pedestrian Accidents by Time Range (2020)
Accident Count
0
200
400
600
800
1000
1200
1400
1600
00:00 - 00:59
01:00 - 01:59
02:00 - 02:59
03:00 - 03:59
04:00 - 04:59
05:00 - 05:59
06:00 - 06:59
07:00 - 07:59
08:00 - 08:59
09:00 - 09:59
10:00 - 10:59
11:00 - 11:59
12:00 - 12:59
13:00 - 13:59
14:00 - 14:59
15:00 - 15:59
16:00 - 16:59
17:00 - 17:59
18:00 - 18:59
19:00 - 19:59
20:00 - 20:59
21:00 - 21:59
22:00 - 22:59
23:00 - 23:59
Time Range

Fig 5

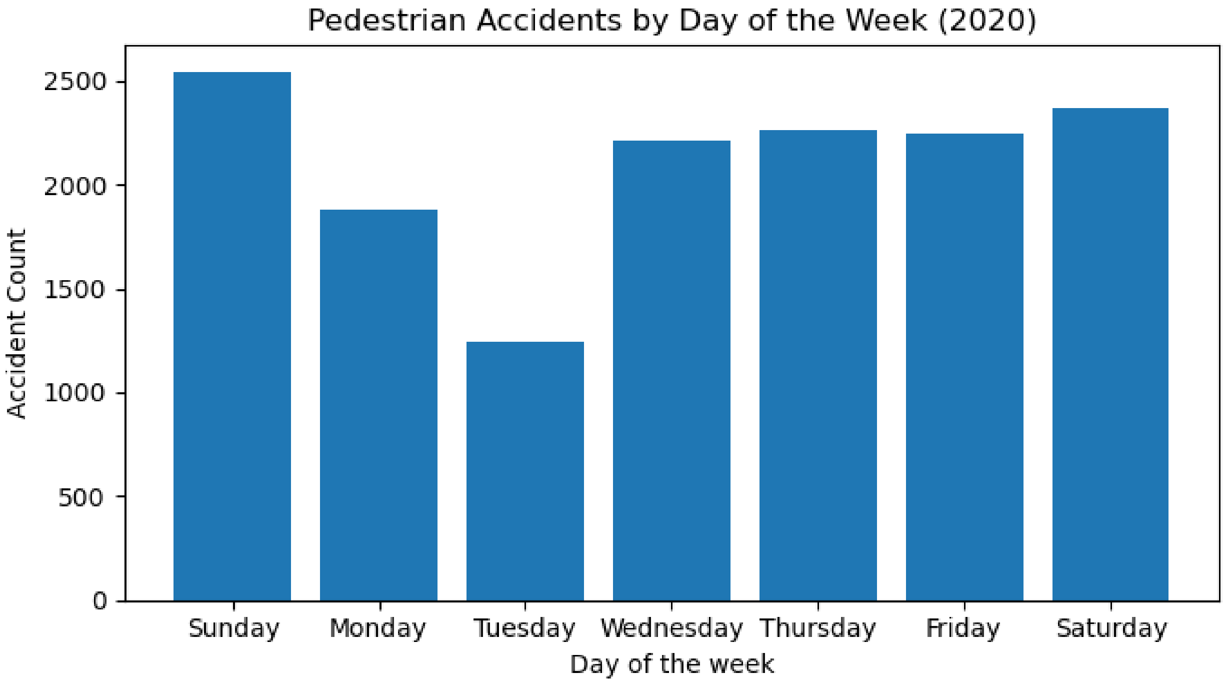

Pedestrian Accidents by Day of the Week (2020)
Accident Count
0
500
1000
1500
2000
2500
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of the week

Fig 6

Summarily, the UK road traffic accidents data of 2020 has shown high prevalence of accidents involving motorcycles, pedestrians and other vehicle types during specific days and times (evenings and weekends). Government should strengthen law enforcement efforts to ensure compliance with traffic rules and regulations, especially during peak accident hours.

## Apriori Algorithm

The apriori algorithm helps identify common patterns or associations among different items in a dataset. The below table shows the pattern of association between the selected variables: accident_severity, speed limit and weather conditions for the UK road traffic accidents for the year 2020. The support value represents the percentage of times the itemset appears in the dataset, indicating its frequency of occurrence.

| SN | Support | Itemset |
|---|---|---|
| 0 | 0.201263 | (severity_2) |
| 1 | 0.783484 | (severity_3) |
| 2 | 0.573033 | (Speedlimit_30) |
| 3 | 0.775546 | (weathercondition_1) |
| 4 | 0.459983 | (Speedlimit_30, severity_3) |
| 5 | 0.603186 | (weathercondition_1, severity_3) |
| 6 | 0.450137 | (weathercondition_1, Speedlimit_30) |
| 7 | 0.359697 | (weathercondition_1, Speedlimit_30, severity_3) |

According to the table, around 20.13% of accidents in the year 2020 were classified as serious (severity_2), while a significant 78.35% were categorized as slight (severity_3). Moreover, approximately 57.30% of accidents involved a speed limit of 30 mph (Speedlimit_30), and about 77.5% of the accidents occurred under fine weather conditions without high winds (weathercondition_1).

Furthermore, when considering specific combinations of factors, it was observed that around 46% of accidents with a speed limit of 30 mph resulted in slight injuries (Speedlimit_30, severity_3). Additionally, approximately 60% of slight injury accidents happened under fine weather conditions without high winds (weathercondition_1, severity_3). Moreover, around 45% of accidents occurred under fine weather conditions without high winds and involved a speed limit of 30 mph (weathercondition_1, Speedlimit_30). Lastly, about 36% of accidents in 2020 were characterized by a combination of fine weather conditions, a speed limit of 30 mph, and slight injuries (weathercondition_1, Speedlimit_30, severity_3).

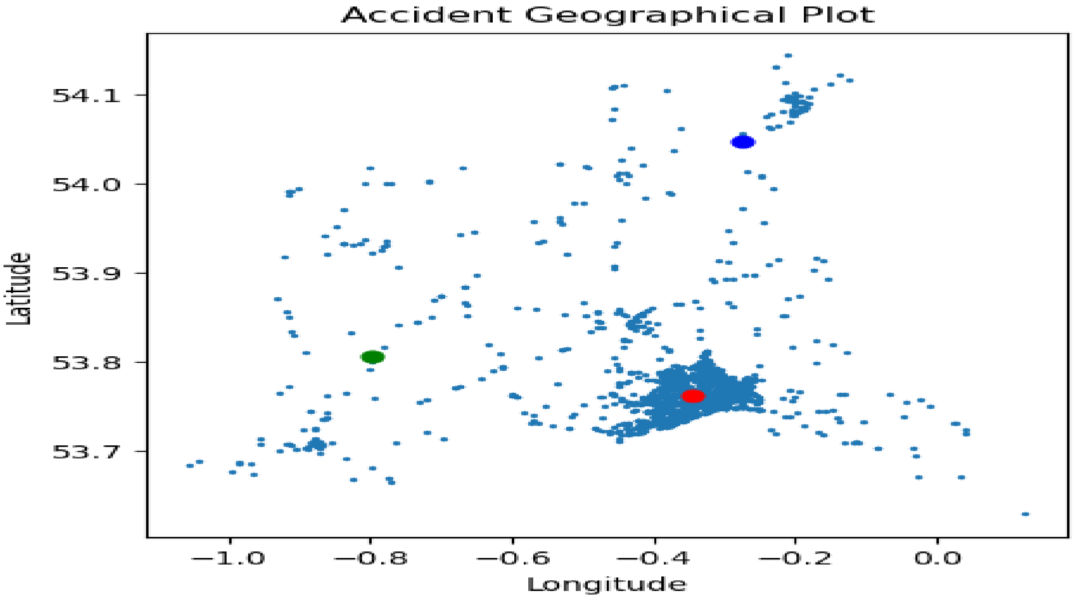
| SN | antecedents | consequents | Confidence | lift | Conviction |
|---|---|---|---|---|---|
| 1 | (weathercondition_1) | (Speedlimit_30, severity_3) | 0.463798 | 1.008294 | 1.007115 |
| 2 | (Speedlimit_30) | (weathercondition_1, severity_3) | 0.627708 | 1.040653 | 1.065865 |
| 3 | (severity_3) | (weathercondition_1, Speedlimit_30) | 0.459099 | 1.019911 | 1.016570 |

Table 1

The above table shows a few selected rules, their values and interpretations. From the above, the first rule shows that when accidents occurred under favorable weather conditions without high winds, there was a 46.38% confidence that these accidents were also associated with both a speed limit of 30 mph and slight injuries. The lift value of 1.008294 suggests a slight positive correlation between these factors. Secondly, accidents in areas with a 30 mph speed limit showed a 62.77% confidence of being linked to favorable weather conditions without high winds and resulted in slight injuries, indicating a positive correlation with a lift value of 1.040653. Lastly, when accidents resulted in slight injuries, there was a 45.91% confidence that they were related to favorable weather conditions without high winds and a 30 mph speed limit, indicating a slight positive correlation with a lift value of 1.019911.

### Clustering

The clustering plot below reveals that the K-means algorithm has grouped the data points into three clusters based on the 'longitude' and 'latitude' coordinates for Kingston Upon Hull. The coordinates of the centroids indicate the center points of each cluster. The labels assigned to each data point indicate which cluster the point belongs to.

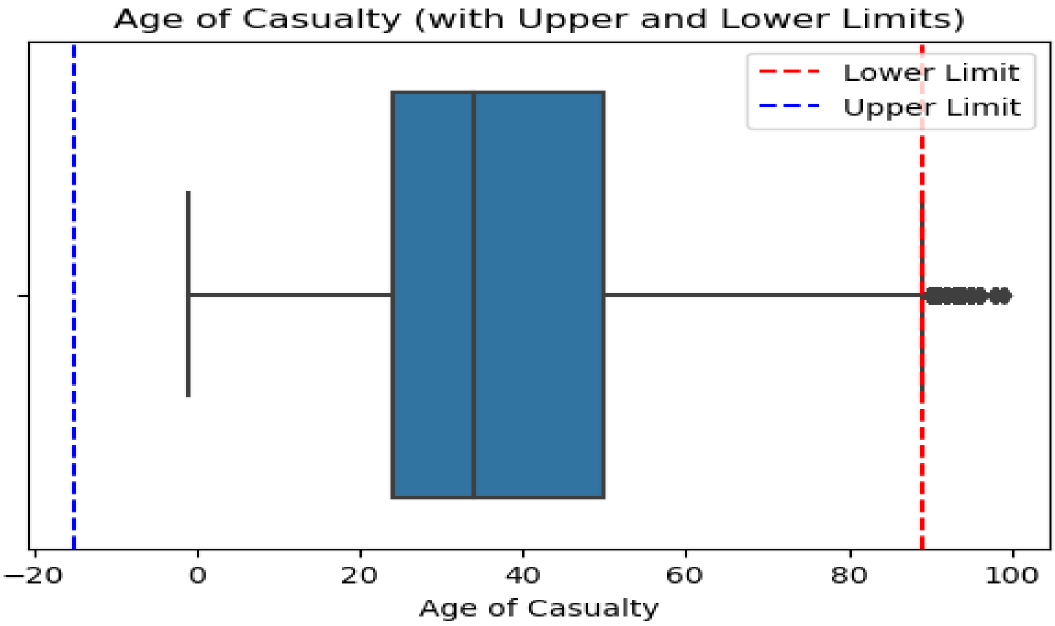


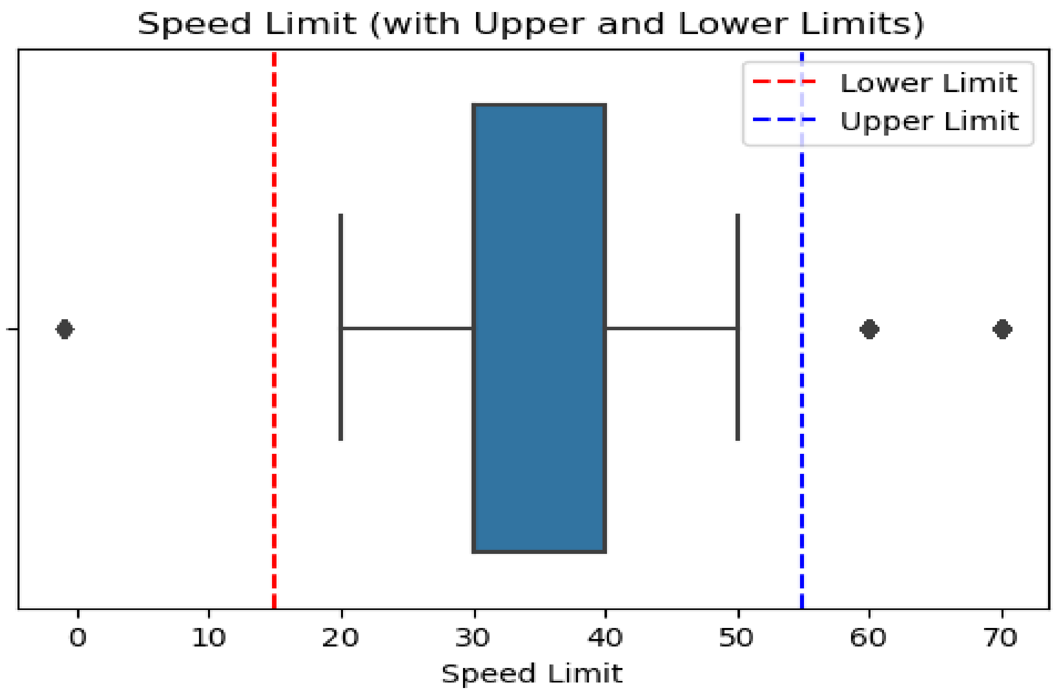
-0.34557787 longitude and 53.76132643 latitude have been identified as the mid-point for accidents in Kingston Upon Hull. It shows a high saturation of data points which indicates a higher frequency of accidents in these areas compared to other areas of Kingston Upon Hull. It is a hotspot for accidents in the region.

**Outlier Analysis**

The UK accident dataset for the year 2020 was subjected to outlier analysis and treatment. Outliers are data points that significantly deviate from the entire dataset. For the purpose of conciseness, only the multiple IQR technique was used to identifier outliers on selected columns or variables relevant to my analysis and model development. Tukey (1977) introduced the concept of the interquartile range (IQR) and the use of multiples of IQR for outlier detection in his book "Exploratory Data Analysis”. This technique divides the distribution of the data into upper quartile and lower quartile and any datapoints that fall outside these thresholds are identified as potential outliers. Using this technique, outliers were identified in age_of_driver, age_of_casualty, and speed_limit columns. See Table below.

| Column Name | Lower limit | Upper limit |
|---|---|---|
| Age_of_driver | -16.0 | 88.0 |
| Age_of_casualty | -15.0 | 89.0 |
| Speed_limit | 15.0 | 55.0 |





The outliers identified in age_of_casualty column were cleaned by replacing with the median value. Furthermore, the negative speed entry of -1 (lower limit) was replaced with the median value. However, not all the outliers identified above the upper_limit (55.5mph) were deemed to be outliers. The table below outlines the national speed limits applicable to different categories of vehicles and road types (Gov.uk, nd).

| Car Type | Built up areas | Single Carriage | Dual Carriage | Roundabout | Slip road |
|---|---|---|---|---|---|
| Cars, motorcycles, car-derived vans, and dual-purpose vehicles | 30mph | 60mph | 70mph | 30mph | 50mph |
| Cars, motorcycles, car-derived vans and dual-purpose vehicles when towing caravans or trailers | 30mph | 50mph | 60mph | 30mph | 50mph |

An exploratory data analysis conducted on the balanced (numeric_data) dataframe revealed that only 21 accident records involved vehicles exceeding the prescribed speed limit while driving on roundabouts and slip roads. The rest of the vehicle categories were observed to be following the prescribed speed limit. Consequently, no adjustments were applied to the outlier values that exceeded the upper limit in the speed limit column.

### Model Predictions

In this section, the goal is here is to develop a classification model that can accurately predict road accidents in the UK for the year 2020.

- **Feature Selection**

The feature selection process involved the process of merging 3 tables into a single dataframe and removing all duplicated records. A total of 10 best numeric columns were selected as input variables using k-best algorithm to predict road accidents (accident_severity) for the classification model. The top performing columns selected include speed_limit, urban_or_rural_area, age_of_casualty, junction_control, junction_location. The data-frame was balanced to ensure equal representation of different classes to promote fair prediction and better generalization of the model. The selected columns were cleaned by removing all invalid entries and outliers' values were cleaned on a few selected columns.

- **Model Results**

| **Classification Algorithm** | **Accuracy Score** | **F1 Score** | **Precision** | **Recall** |
|---|---|---|---|---|
| Decision Tree | 0.79 | 0.78 | 0.82 | 0.74 |
| Decision Tree with Cross Validation | 0.94 | 0.93 | 0.97 | 0.90 |
| Naïve Bayes | 0.71 | 0.76 | 0.66 | 0.90 |
| Random Forest | 0.80 | 0.79 | 0.84 | 0.75 |

From the above, Decision Tree Cross Validation model achieved the highest accuracy rate of 94%, along with precision and recall values of 0.96 and 0.90, respectively. However, overfitting may be present, indicating a need for further evaluation and potential adjustments. The random forest model possessed accuracy of 80% and precision of 84%, outperforming other models. The model successfully predicted 84% of fatal accidents, making it a suitable model for predicting accidents in the United Kingdom.

## Recommendations

Sequel to the completion of analysis UK's 2020 accident data, the following actionable suggestions are proposed to the UK government.

1. Implement stricter traffic enforcement measures especially during the weekends and pick period during the day (3pm – 7pm). Focus on monitoring and enforcing speed limits in both urban and rural areas, review driving age and other traffic accident regulations and contributing factors.

2. Enforce stringent regulations on the utilization of "over 50cc and up to 125cc" and "over 500cc" motorcycle types. This includes the introduction of higher taxes (tariffs, excise duties) to discourage usage, thorough reassessment of licensing requirements, promotion of focused public awareness initiatives etc.

3. Deploy the use of machine learning and statistical techniques such outlier and clustering analysis to identify accident hotspots and unusual patterns and behaviors of road users.

4. -0.34557787 longitude, 53.76132643 latitude has been identified as accident hotspot in Kingston Upon Hull. The traffic authorities should implement targeted safety measures such as enhanced traffic signage, stricter speed limits, increased law enforcement presence and public awareness campaigns in this area.

# References

Cambridge Dictionary. (n.d.). *Weekend*. Retrieved from https://dictionary.cambridge.org/dictionary/english/weekend

Gov,uk(n.d).*National Speed limits*. Retrieved from https://www.gov.uk/speed-limits

Statology. (2021). *When to Use Mean vs. Median*. Retrieved from https://www.statology.org/when-to-use-mean-vs-median/

Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.