Road Traffic Accident Report 2020

**Introduction**

The dataset is on reported road traffic accidents in Great Britain of varying degrees, released by the government every year but we have limited our analysis to just the year 2020. With the use of python and other relevant libraries, I have attempted to query the dataset to generate answers to pertinent questions as well as employed machine learning techniques in an attempt to predict whether an accident is fatal given the conditions it happened under.

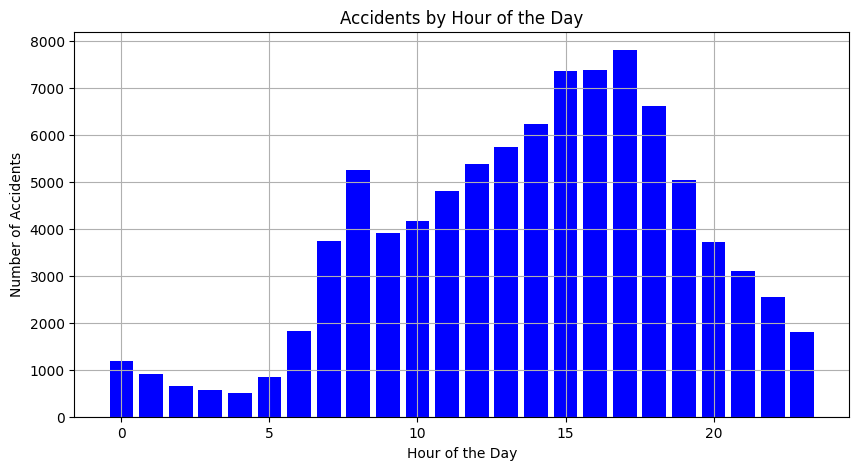
**Analysis**

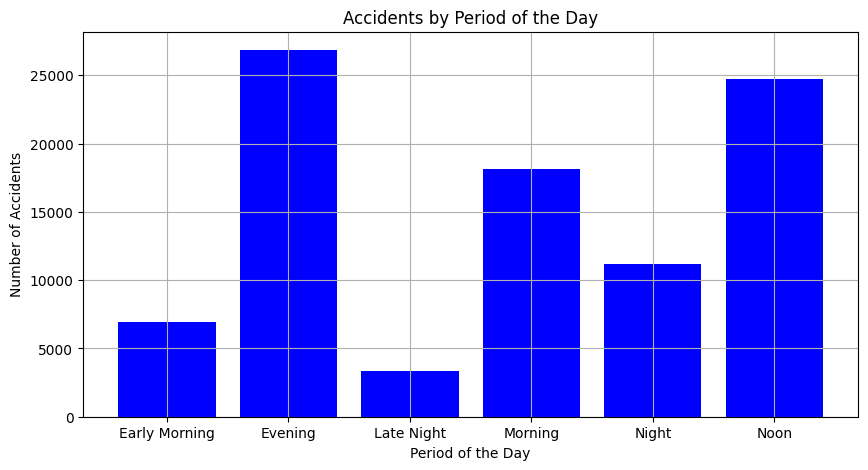
The first line of action was to clean the data to make sure the accuracy of our analysis and predictions were not in doubt. There are three datasets(Accidents\_df, vehicle\_df and casualty\_df) and they were cleaned individually and in a similar pattern as they all shared similar issues.

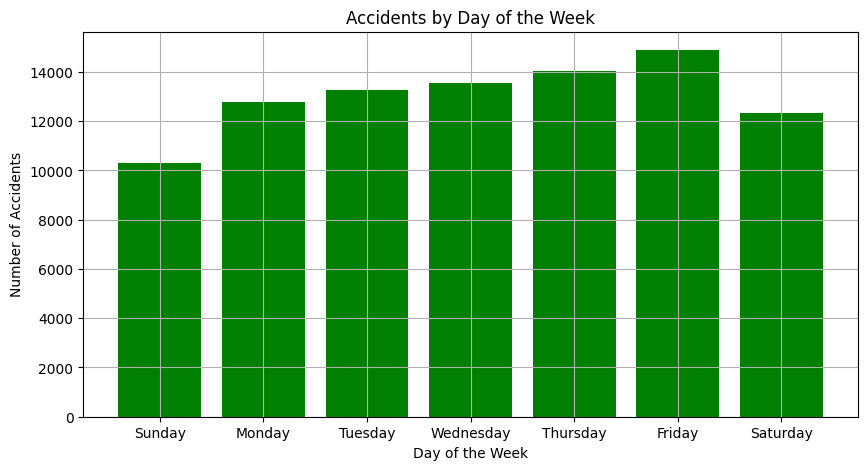
Most of the unknown values for variables were represented with -1 and so I had to replace all -1 entries with Nan to explicitly indicate that the value is missing. There was the possibility that there would be the case of an unknown entry for the longitude column and so i inspected the longitude column for values that are exactly -1, to adjust to a similar value of -1.0001, just to make that distinction. but apparently, the slice of our dataset from the year 2020, does not have this problem.

Next was the problem of missing data, and so I opted to fill these missing values with the median value of their respective columns.

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

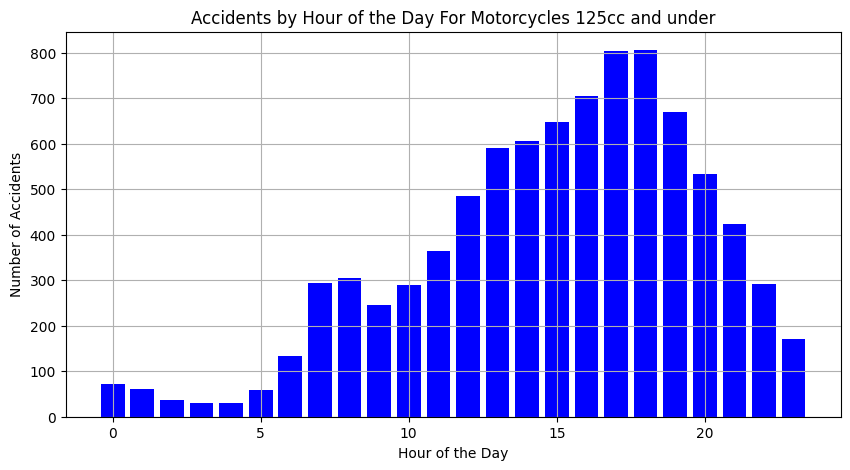
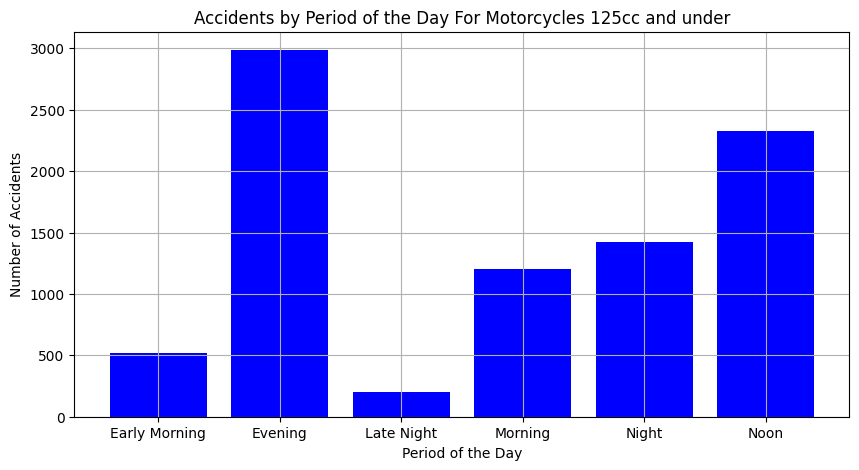


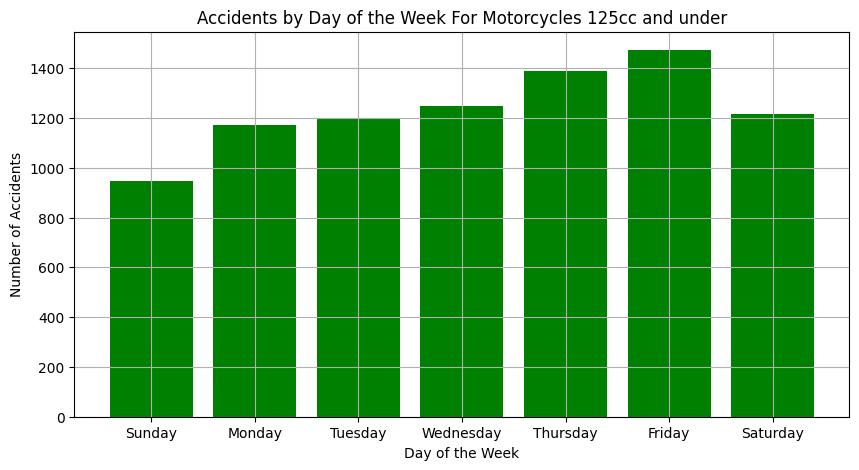




From the visuals above, it is evident that the majority of accidents in the dataset occur during the Evening (4:00 PM - 7:59 PM), Noon (12:00 PM - 3:59 PM) and Mornings (8:00 AM - 11:59 AM). This trend may be attributed to the fact that these hours represent peak activity times, encompassing commuting to and from work, school, and other work-related activities. It is worth noting that Friday exhibits the highest frequency of accidents, and there is a noticeable pattern of consistent increase from Sunday through Friday, followed by a decline on Saturday.

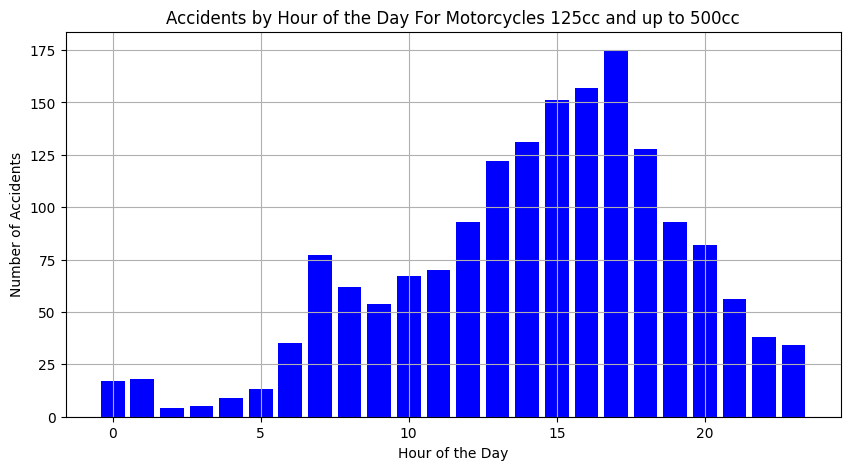
* **For Motorbikes, are there significant hours of the day, days of the week, on which accidents occur?**
* **For motorcycles 125cc and under**

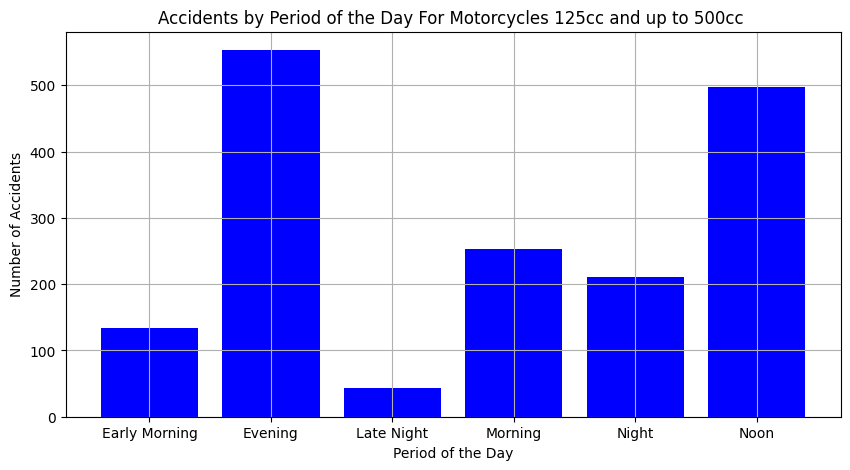


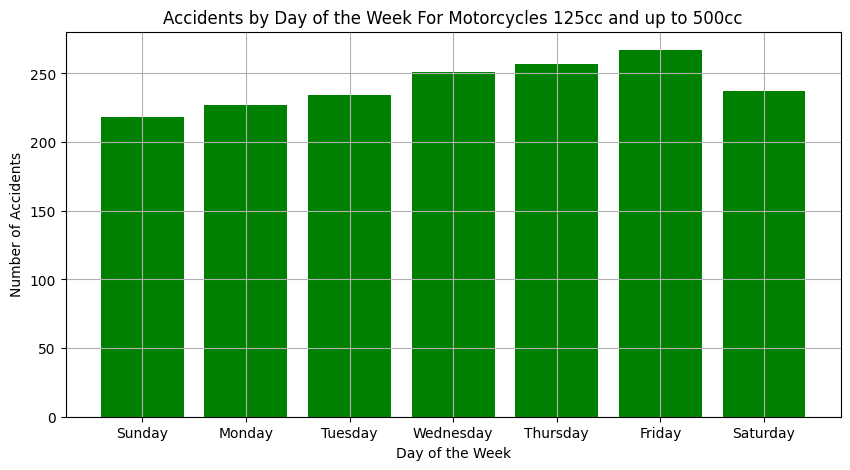


Thursdays and Fridays show the highest occurrence of accidents involving motorcycles of 125cc and under. Additionally, evenings and noons are the times when these accidents are most frequent.

* **For motorcycles 125cc and up to 500cc**

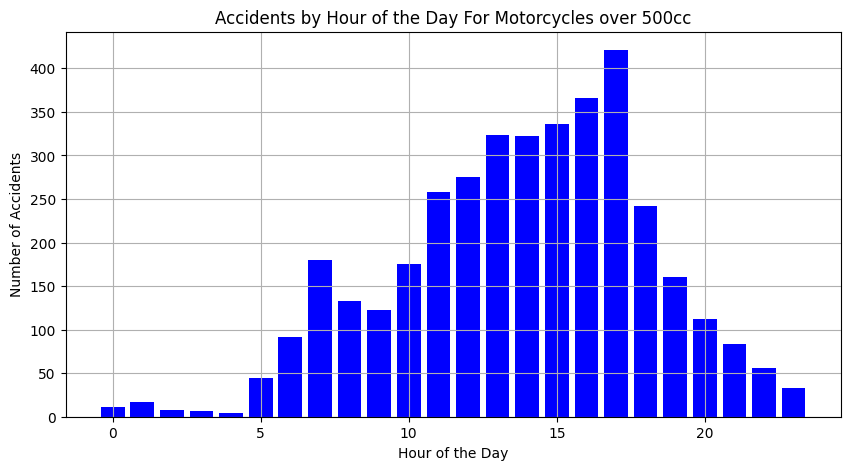


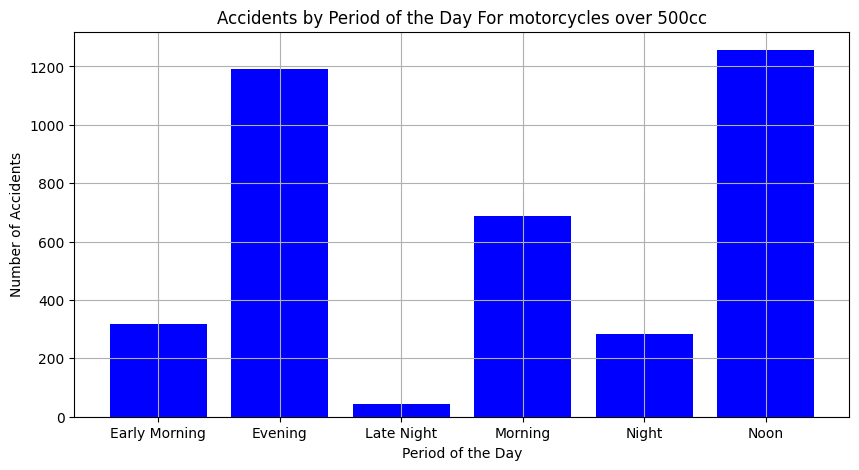


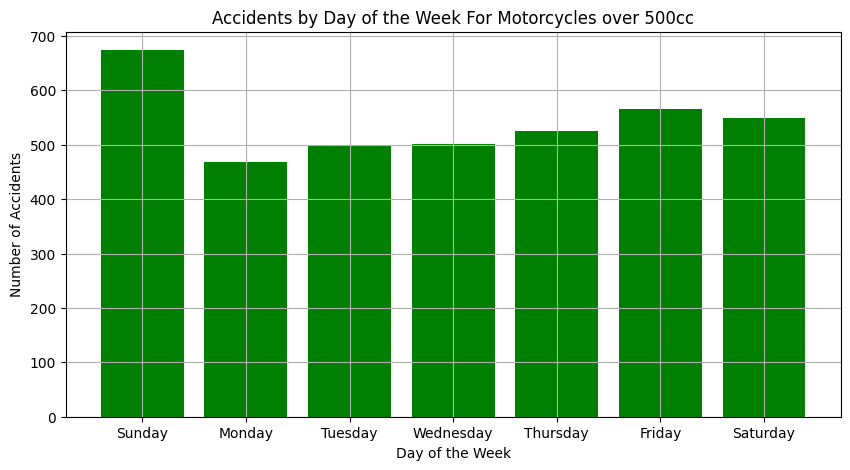


There are significantly fewer accidents involving motorcycles with engine capacities between 125cc and 500cc (1691) compared to motorcycles with engine capacities of 125cc and under (8650). The pattern observed in the overall dataset is also evident in this subset, with Friday having the highest occurrence of accidents, and Noon and Evening time also having more occurrences.

* **For motorcycles over 500cc**

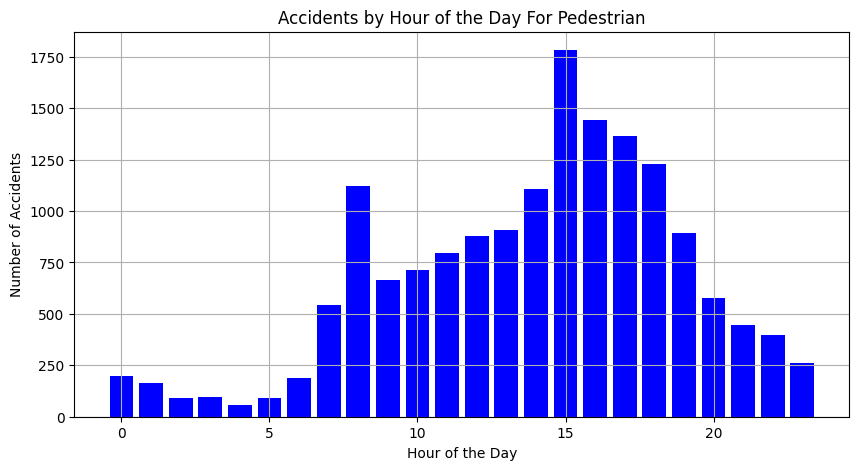


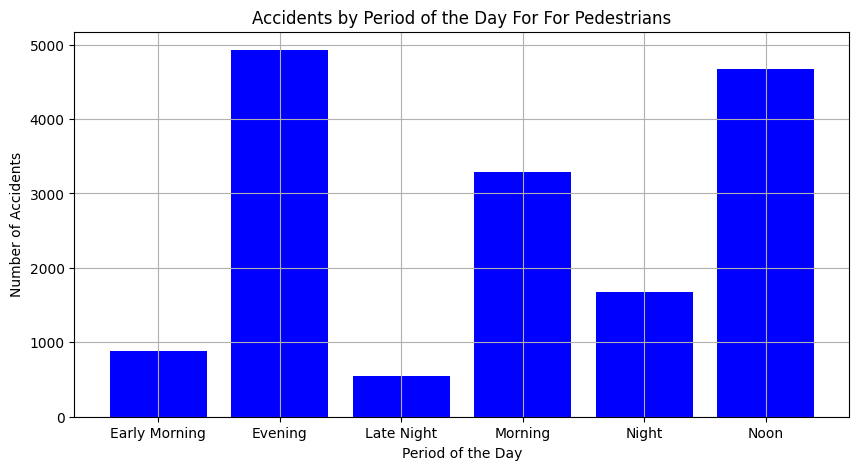


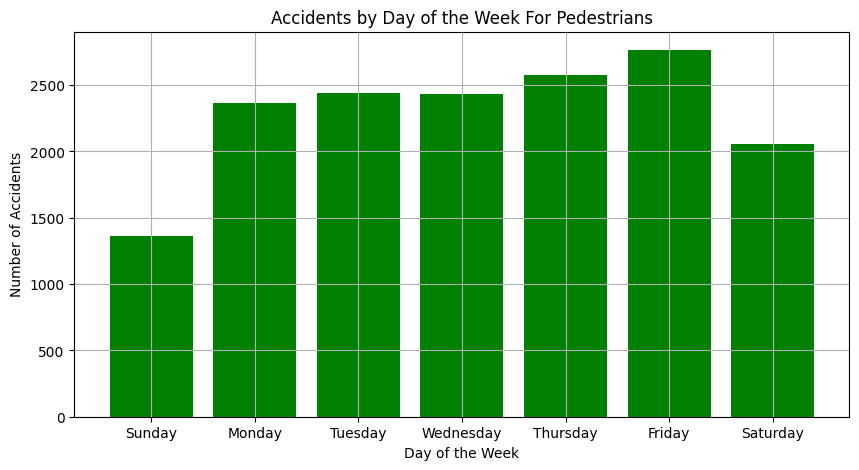


It is evident that Sundays have the highest occurrence of accidents involving motorcycles with engine capacities over 500cc, which differs from the pattern observed in other cases. However, the higher occurrence during Noon and Evening hours remains consistent, as observed with other motorcycle types.

* **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?**

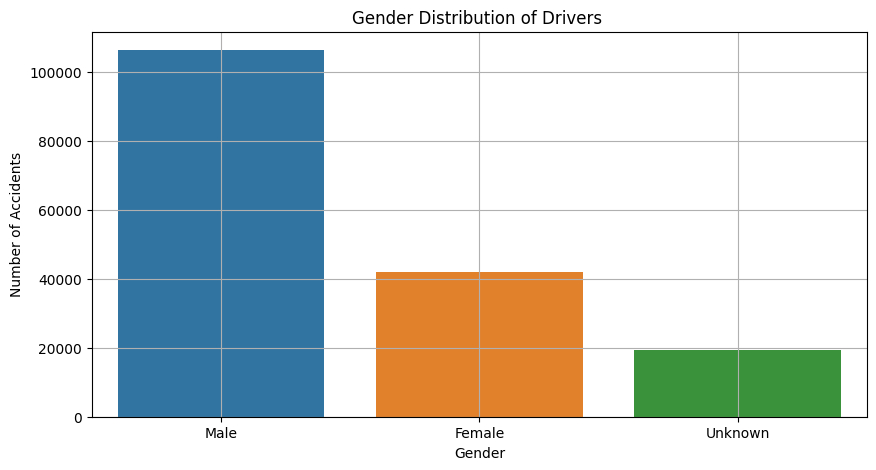


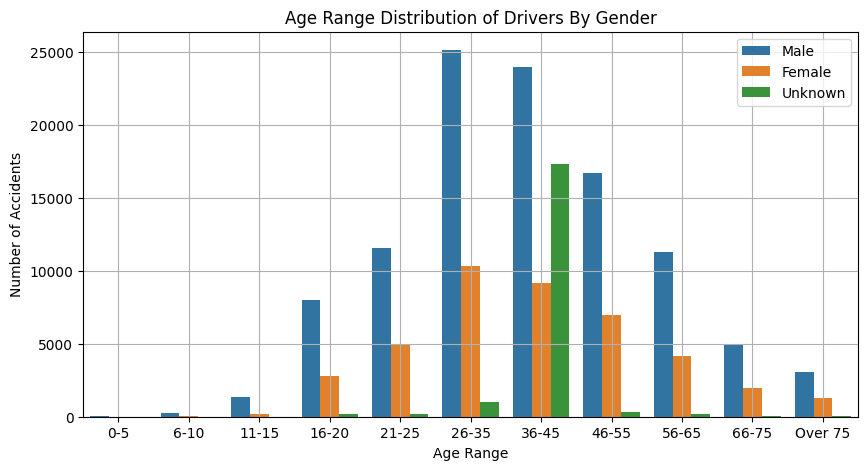




On Fridays, pedestrians are more likely to be involved in accidents, possibly due to the start of the weekend. Additionally, it is observed that evenings and noons are also more likely times for pedestrian accidents to occur.

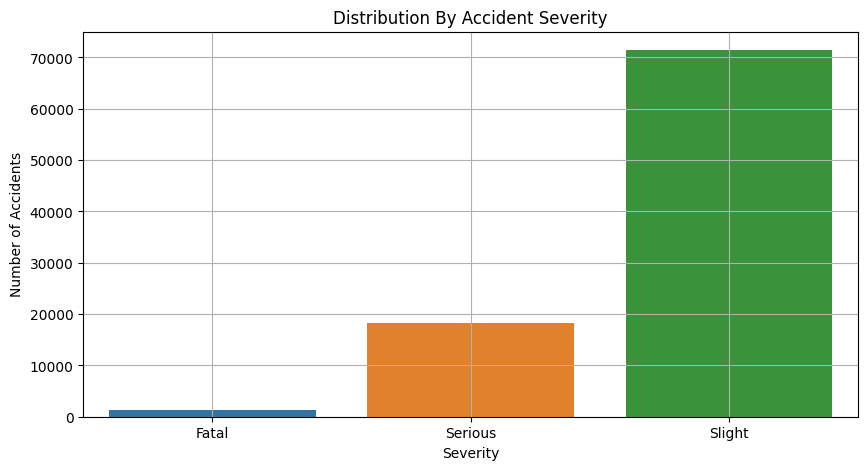
* **Age and Gender Distribution of Drivers**

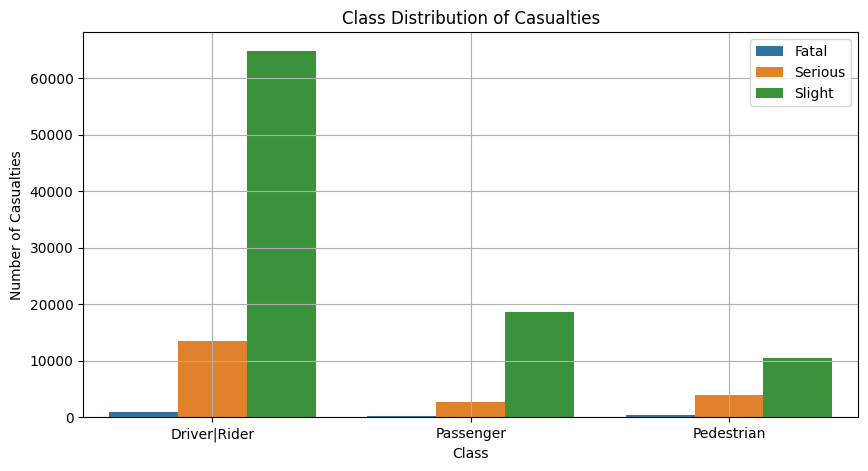
****

****

A large proportion of Accidents involved drivers within the age bracket of 26-35 and 36-45 (the youthful population). Also observed is that more than 60% of the drivers involved in accidents in the dataset were male. Targeted campaigns towards this age range by the government could serve to reduce the number of accidents experienced.

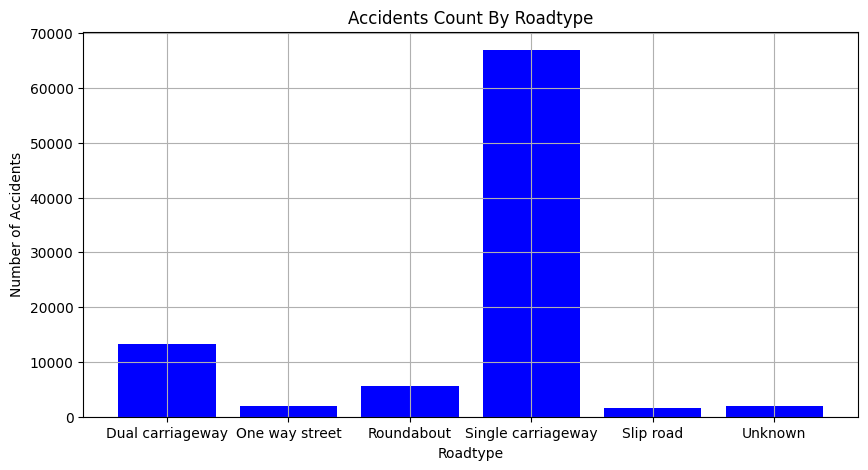
* **How Balanced is the Dataset in terms of the distribution of Accidents by severity?**

****

****

The dataset is highly imbalanced with the severity type of slight being so much more than the Serious and Fatal types combined. And we can see this filter down to the class distribution of Casualties with a large proportion (approximately 68%) of casualties involving Drivers|riders with some of the casualties for the other classes being caused by these drivers|riders.

* **On what road type do accidents most frequently occur?**

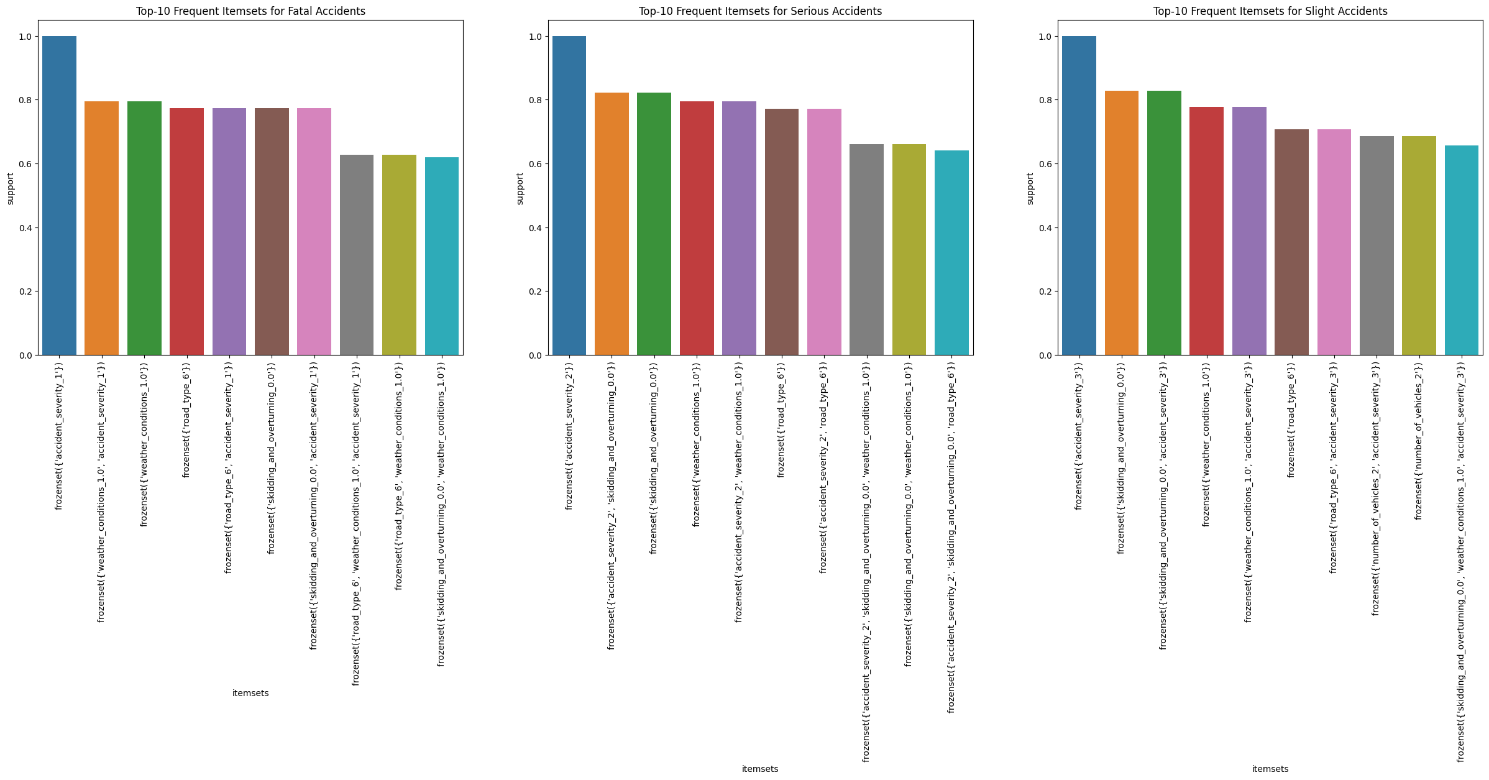
****

From the visual above it is clearly seen that more than 70% of accidents that occurred in 2020 were on single carriageways(A single carriageway, also known as an undivided highway, is a type of road with a single roadway for vehicles traveling in both directions. It means that all traffic, regardless of direction, shares the same road surface, often separated by a painted line or a center median.)

* **Using the apriori algorithm, explore the impact of selected variables on accident severity.**

Because of the imbalance in the dataset, a lot of the associations for accident severity that were fatal or serious are being neglected. And so, we will look at each of these subsets individually to try to draw out insightful relationships.

In the real world, many factors contribute to each distinct crash. Hence, it is likely to have several items in either the antecedent or consequent. The figure below visualizes the frequency of items. Overall, the top ten frequent items in each of the subsets are represented in the bar chart below.



After dividing the dataset into the three magnitudes of severity (slight, serious, and fatal), I highlighted some interesting patterns I noticed in generating rules using the apriori algorithm.

* **Fatal Accidents**

**Association**: Accidents with a speed limit of 60 km/h and fatal severity are more likely to occur on single-carriageways under fine weather conditions.

* **Serious Accidents**

**Association**: Accidents with a speed limit of 30 km/h resulting in serious accidents are more likely to occur on single-carriageways, and there is no skidding, jack-knifing, or overturning involved.

* **Slight Accidents**

**Association**: Accidents on single carriageways, with no skidding, jack-knifing, or overturning, resulting in slight accidents, and occurring in fine weather conditions with a speed limit of 30 km/h, are 1.19 times more likely compared to accidents in other conditions.

One common factor observed in these created or observed rules is the constant presence of the single-carriageway road type. Also observed is that the speed limit might play a crucial role in determining the severity of the accident seeing as speed limits of 60 are more associated with fatal accidents and speed limits of 30 are more associated with serious and slight accidents

* **Identify accidents in Kingston upon Hull, Humberside,and East Riding of Yorkshire. Run clustering on this Data. What do these clusters reveal about the distribution of accidents in these regions?**

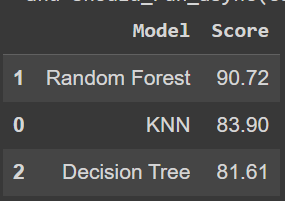
**Predictions**

Using a classification model, I attempted to predict the fatality of an accident given the conditions it happened under.

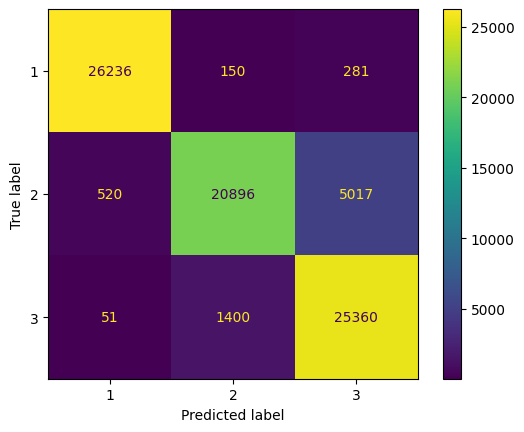
**Addressing Class Imbalance**

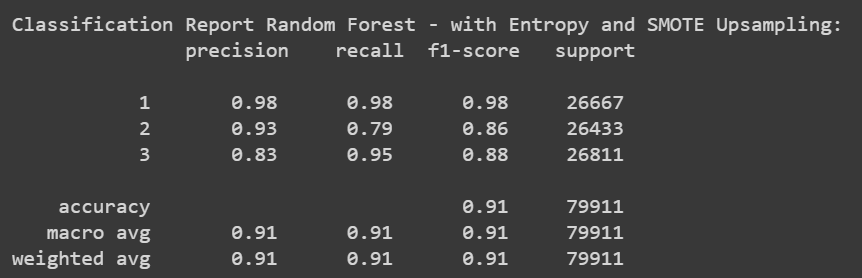
Upon conducting Exploratory Data Analysis (EDA), it became evident that the dataset exhibited a significant class imbalance. This phenomenon arises from the infrequent occurrence of Fatal accidents in comparison to the other two categories (Serious and Slight). Consequently, data are scarce for these classes. To mitigate this issue, we implemented the Synthetic Minority Over-sampling Technique (SMOTE), which is recognized as one of the most effective methods for handling imbalanced datasets. This approach involves generating synthetic samples to rebalance the class distribution.

After this came trying out different models to attempt to find the best-performing model. The accuracy scores for the three different models used were:



The best-performing model was the random forest classifier.





The precision, recall, F1 score and accuracy all are at high levels of 0.80+, meaning that the classification is successful and the accuracy of the model is more or less 90% when investigated on multiple metrics.

**Recommendations**

* The age of the driver was a very high contributing factor to the severity of the accident after running the random forest algorithm. With backing from our analysis, conducting targeted awareness campaigns focusing especially on male drivers within the age bracket of 26-45 (the youthful population) encouraging safe driving practices, by the government could serve to reduce the number of accidents experienced.
* Enhanced Signage and Warning Systems through the Installation of clear and prominent signs indicating the speed limit on single-carriageways could go a long way in reducing the frequency of these accidents. Additionally, consider using electronic message signs to alert drivers about the speed limit and road conditions.
* The Implementation of traffic calming measures such as speed bumps, chicanes, or roundabouts to naturally slow down vehicles and improve safety on single-carriageways should also be a consideration.