IBM Data Science Professional Certificate

Applied Data Science Capstone

Analysis of Severity of Accidents in UK

Submitted by:

Somya Bhushan

September 2020

INTRODUCTION

1.1 Background

There are many inventories in automobile industries to design and build safety measures for automobiles, but road accidents are unavoidable. Road accidents are a major world economic and social problem as shown by the report of loss of lives and properties in many countries around the world. Reporting indicated the number of fatalities from road accidents per year of about 1.3 million and 50 million injuries were recorded or an average of 3000 deaths per day and 30,000 injuries per day. Furthermore, its consequences have an impact on economic and social conditions in terms of health care costs of injuries and disabilities. On average, five people die every day on the road in Great Britain and countless more are seriously injured. Britain's road safety record has stagnated in recent years, with the number of road deaths remaining broadly constant for several years. In recent years, there is an increase in the researchers' attention to determine the significant factors that affect the severity of the injuries which is caused due to the road accidents. Accurate and comprehensive accident records are the basis of accident analysis.

1.2 Problem

The objective of this analysis is to know and understand more about the severity of accidents and to see whether different variables considered have any effect to the cause of the accident and what could be done with respect to this.

1.3 Target Audience

The purpose of accident analysis is to help decision-makers understand the nature, causes, and injury outcomes of crashes. The mental and emotional injuries after a car accident can include mental anguish, emotional distress, sleep disturbances, etc. This information provides context for the design of strategies and interventions that will reduce accidents and their consequences.

DATA ACQUISITION AND CLEANING

2.1 Data Sources

The secondary data used for this analysis has been taken from Kaggle consisting of road accidents in UK over a period of 10 years ranging from 2004 to 2014 and which is further divided into three datasets namely Casualties, Accidents and Vehicles.

2.2 Data Cleaning & Feature Selection

Data downloaded or scraped from the secondary source were put together and a thorough study of data was done. There were a lot of missing values for some features else rest was kept recorded.

There were some problems with the data and in order to make it more appealing data cleaning was done to address the missing values, in particular. The three datasets were on different files, so I merged those files and made a new dataset to further work upon. I merged the casualties and accident dataset first and then again merged it with vehicle dataset on the basis of accident index. To address the missing values in the new dataset, I dropped some features like latitude, longitude, accident location because there were quite a few missing values in those categories and the results would not have been valid.

METHODOLOGY

Exploratory Data Analysis

3.1 Target Variable

After examining the dataset and the features, following questions were considered for the analysis part that would give us the relationship between variables:

1. Is there any relationship between the timing of the accident and number of fatal accidents?
2. Does the age of driver have any effect on the number of accidents happened?
3. How does the weather impact the severity of accidents?
4. Are certain designs of vehicles safer than others?

These all answer the same where different dependent variables considered gave the severity of accidents and the number of fatal accidents which is our target variable for this study.

3.2 Relationship between Timing and Fatality of Accidents

It is accepted that to know whether the variables show any degree of dependence on each other, correlation matrix must be used. Here, I have made a correlation matrix between Hour, Day, Month of the accident and its severity where -1 shows that there exists a perfect negative relationship between the variables, 0 shows that there is no correlation between the variables and 1 shows a perfect positive relationship between the variables. The bandwidth at the right side of figure 1 depicts the degree of correlation between the variables and accordingly results are computed.

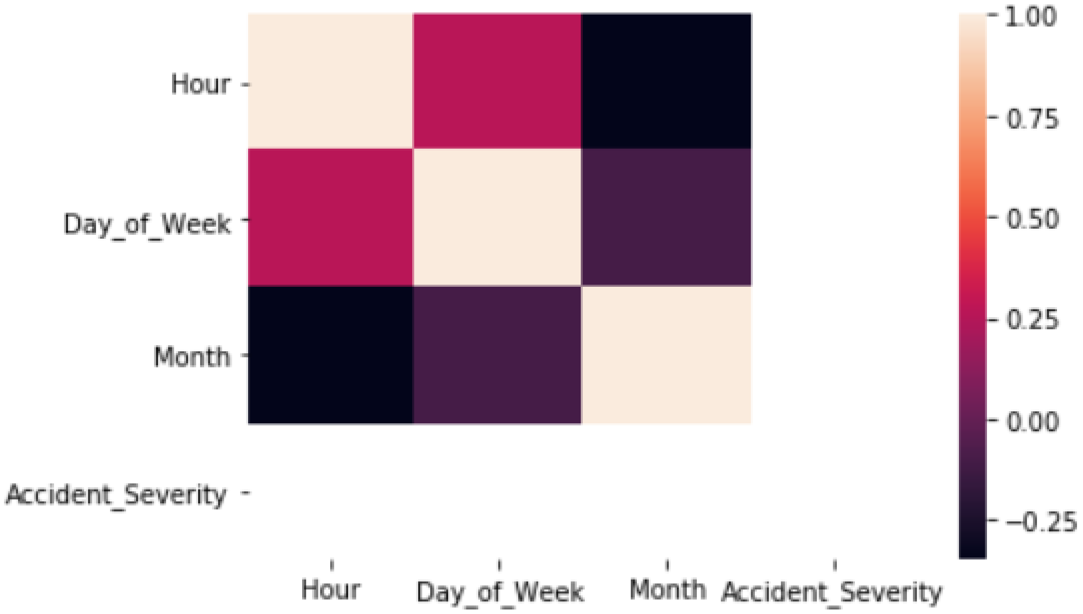


Figure 1. Correlation Matrix between Timing of the Accident and Accident Severity

This bar graph depicts the number of accidents and its timing of occurrence (whether it was evening, night, morning, afternoon, early morning or late night) and legend shows the accident severity, 1 being Fatal, 2 being serious and 3 being slight.

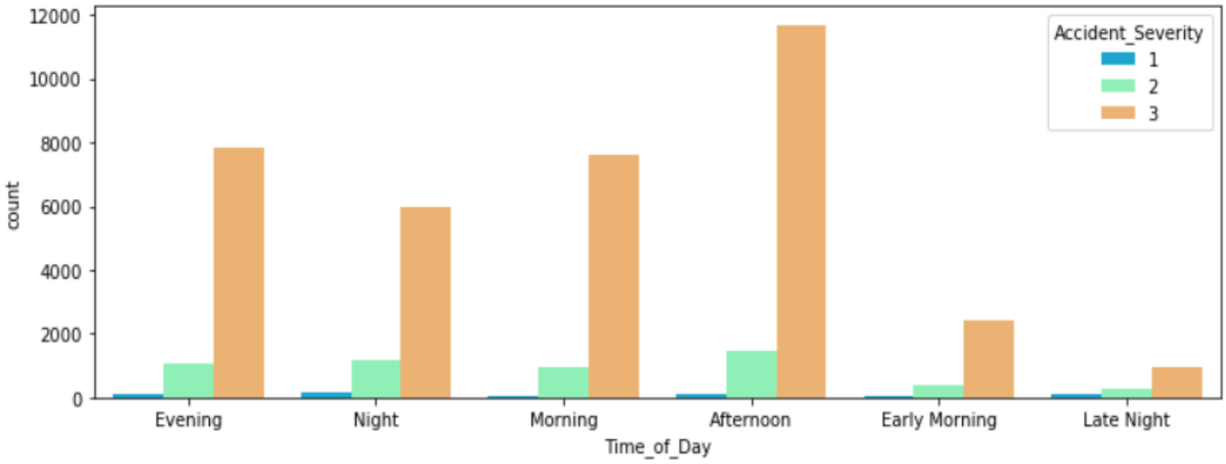


Figure 2. Bar Graph showing Timing of Day and the Number of Accidents

3.3 Relationship between Driver and Accidents

For this, I made a multiple bar chart to know whether the driver's age has any relation with the number of casualties happened in accidents and legend shows the sex of the driver, 1 being male, 2 being female and 3 being others.

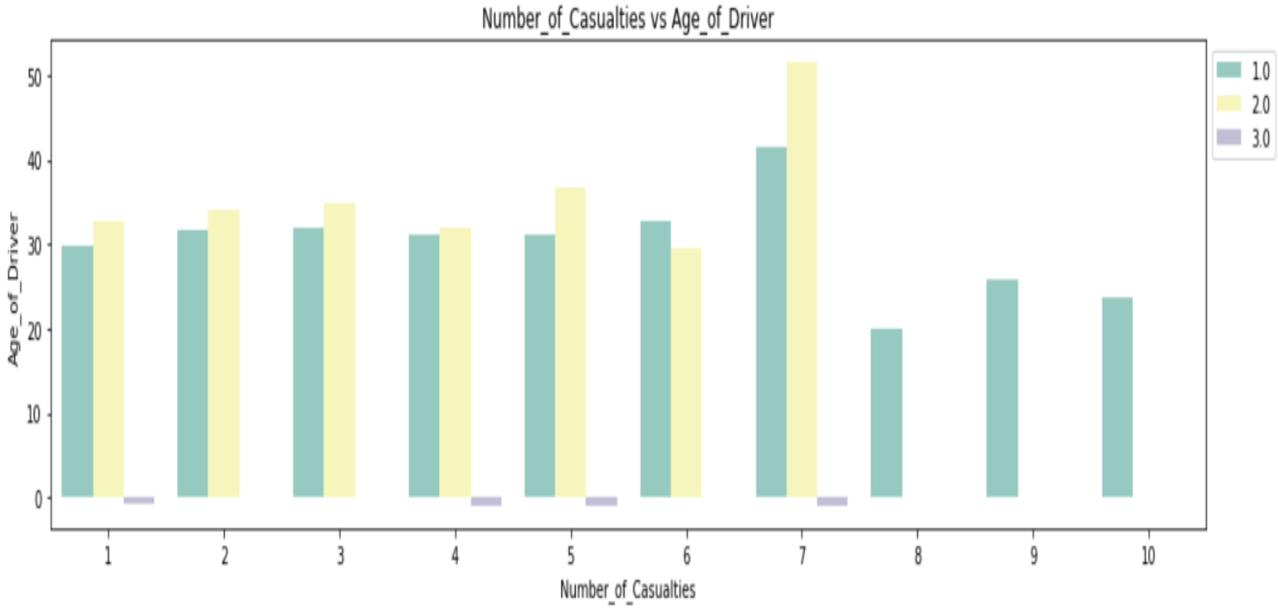


Figure 3. Bar graph showing relationship between Number of Casualties and Age of Driver

This box plot shows the shape of distribution of age of driver on the basis of their location of residence, that is, whether they come from urban area or some small town or rural area. This type of graph usually has its central value, and its variability. In a box and whisker plot, the ends of the box are the upper and lower quartiles, so the box spans the interquartile range. the median is marked by a vertical line inside the box and the outliers are shown that lies outside this region.

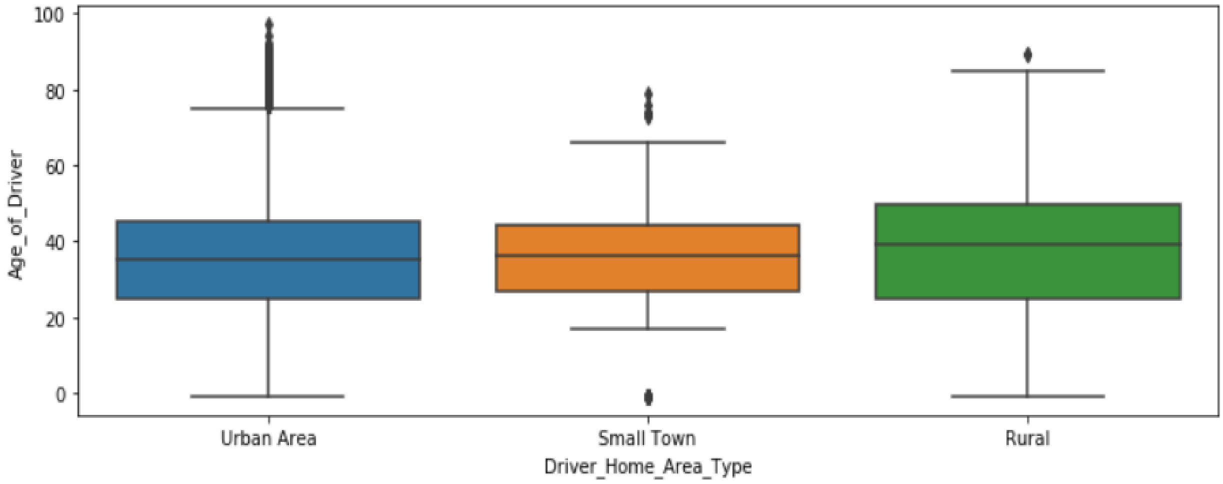


Figure 4. Box Pot showing the distribution of Age of Driver and their residence type

3.4 Relationship between Weather and Severity of Accidents

Here I made a correlation matrix between weather conditions, hour of the accident and severity of the accident to know whether there is any dependence or not between the variables. To know whether the weather conditions have an effect on the severity of the accident.

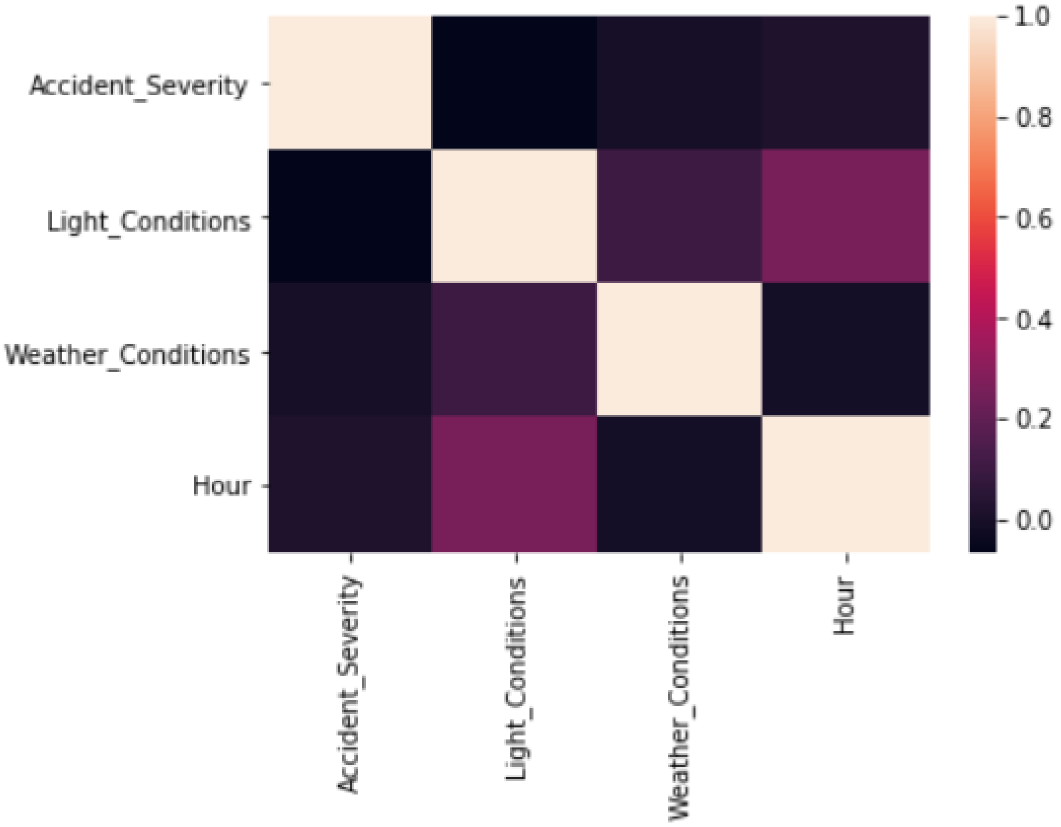


Figure 5. Correlation Matrix between Weather Conditions and Accident Severity

This multiple bar chart depicts the relationship between different weather conditions (1 being Fine with no high winds, 2 being Rain with no high winds, 3 being Snow with no high winds, 4 being Fine with high winds, 5 being Rain with high winds, 6 being Snow with high winds, 7 being Fog or mist, 8 being other and 9 being Unknown conditions) and Hour of the accident. The legend shows the severity of accident where 1 is Fatal, 2 is Serious and 3 is Slight consequences.

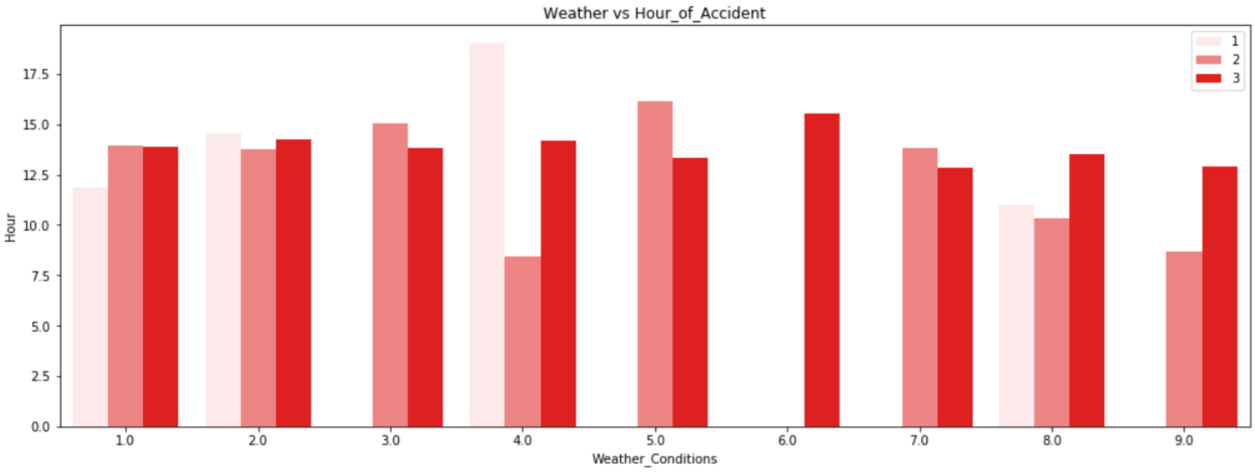


Figure 6. Bar Graph showing relationship between Weather Conditions and Hour of Accident

3.5 Severity of Accidents

This graph shows the severity of accidents and depicts which kind occurred the most and which gender was a casualty. The legend shows the gender with 1 being male and 2 being female.

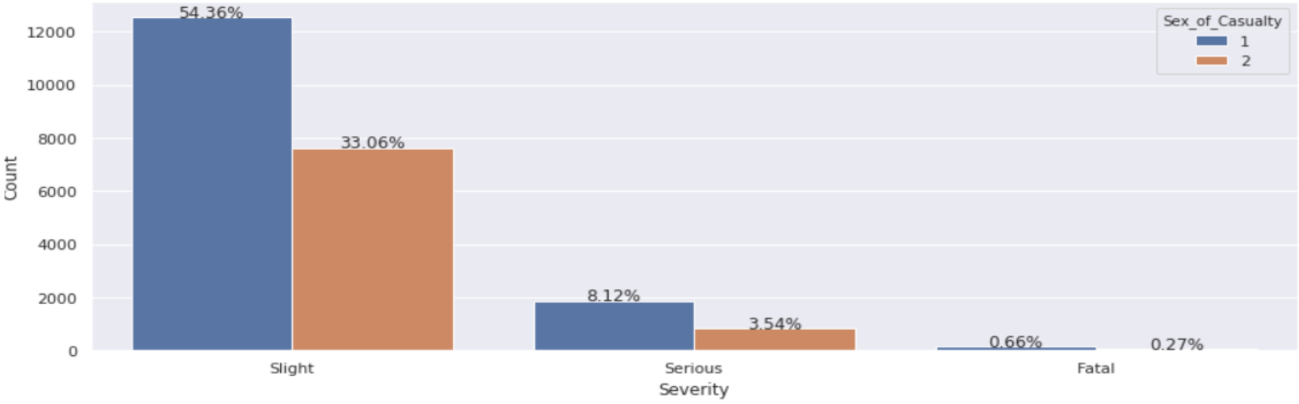


Figure 7. Bar Graph showing Severity of Accidents on the basis of Gender

3.6 Relationship between Vehicle and Accidents

This graph shows the type of vehicle that have been involved in an accident the greatest number of times.

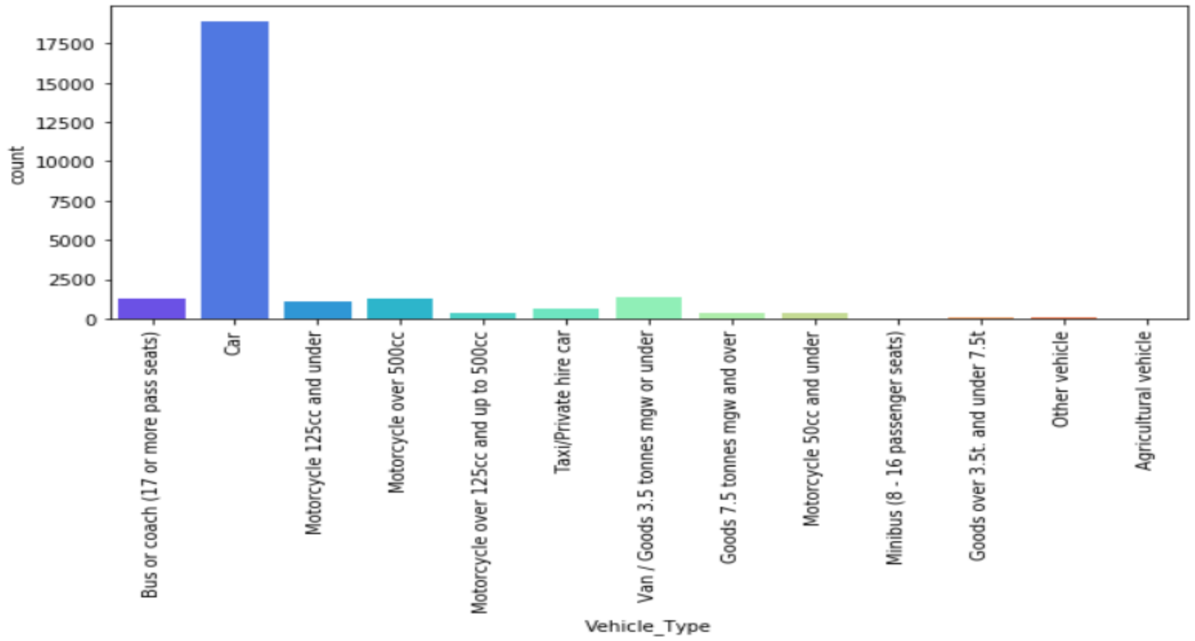


Figure 8. Bar Graph showing the Vehicle Types that have been met with an Accident

This graph shows the different vehicle types and their engine capacity in terms of Heavy Oil, Petrol, Petrol gas (LPG), Hybrid Electric, Gas or Bi-Fuel and Gas.

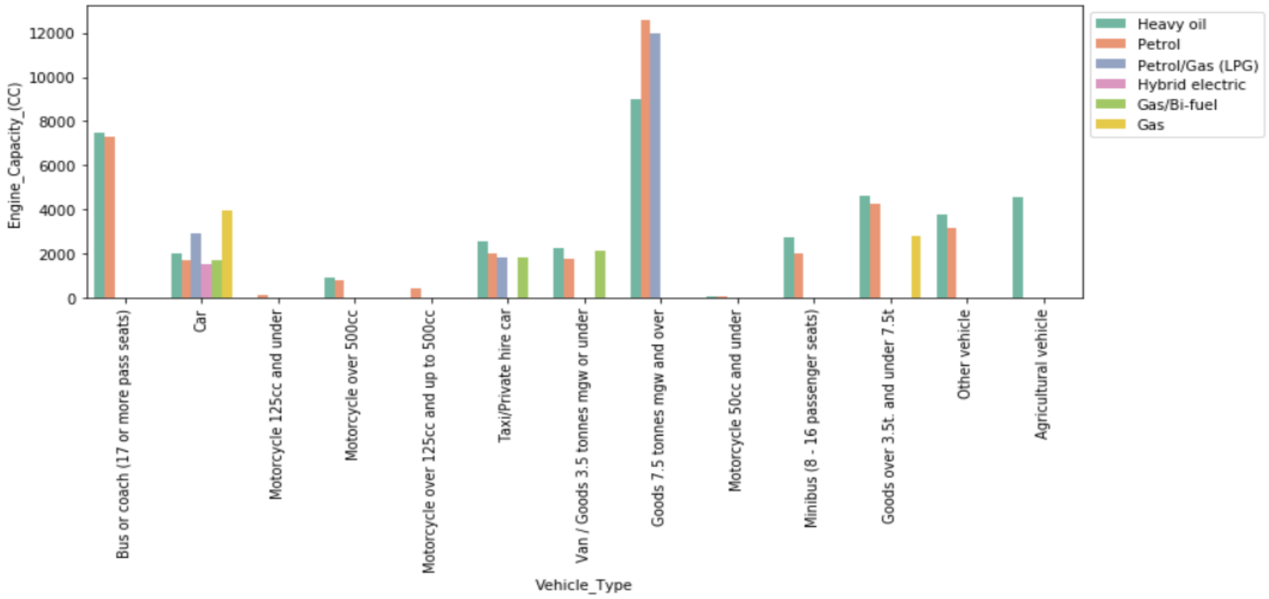


Figure 9. Bar Graph showing the Vehicle type and their Engine Capacity

3.6 Impact of Accidents

This graph shows the first point of impact on the vehicle during an accident.

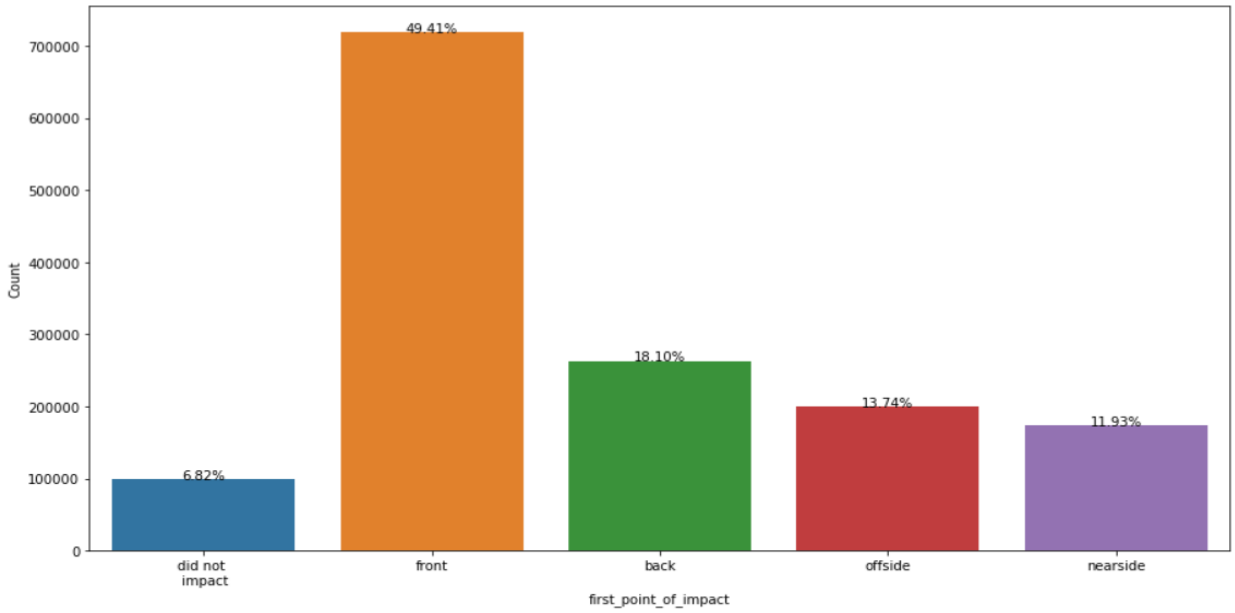


Figure 10. Bar Graph showing the point of First Impact of an Accident

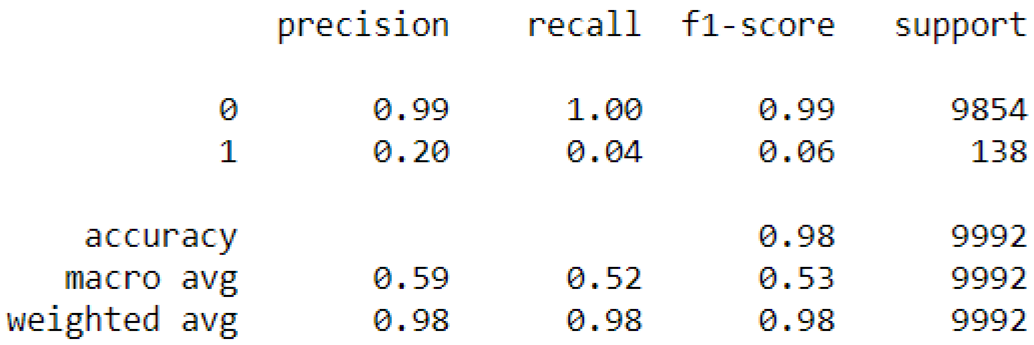
Predictive Modelling

3.7 Classification Model

In order to do this, we first have to find a function (model) that best describes the dependency between the variables in our dataset. This step is called training the model. The training dataset will be a subset of the entire dataset. I tried to create a model using decision tree to find the probability of the severity of accident using dummies for the features Accident Severity and Gender of the Driver where Accident_Severity_1 corresponds to fatal accident and Sex_of_Driver_1.0 corresponds to male driver.

In Classification report, *precision* is the ability of a classifier not to label an instance positive that is actually negative. For each class it is defined as the ratio of true positives to the sum of true and false positives. Said another way, “for all instances classified positive, what percent was correct?”, *recall* is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives. Said another way, “for all instances that were actually positive, what percent was classified correctly?”, the *F_1 score* is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. Generally speaking, F_1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F_1 should be used to compare classifier models, not global accuracy and *support* is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn’t change between models but instead diagnoses the evaluation process.

CLASSIFICATION REPORT



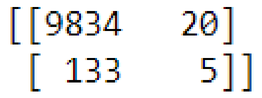
```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      9854
           1       0.20      0.04      0.06       138

    accuracy                           0.98      9992
   macro avg       0.59      0.52      0.53      9992
weighted avg       0.98      0.98      0.98      9992
```

The confusion matrix visualizer is a score visualizer that takes a fitted scikit-learn classifier and a set of test x and y values and returns a report showing how each of the test values predicted classes compare to their actual classes.

CONFUSION MATRIX



```
[[9834   20]
 [ 133    5]]
```

RESULT & CONCLUSION

In this study, the target variable taken is severity of accident and analysed the relationship between the target variable and different features like timing of the accident, timing of the day, driver's age, weather conditions, and different types of vehicle.

The study found out that the drivers who met with an accident were mostly in the age band of 30-40 years and were males. The accidents usually had occurred in the afternoon and the weather conditions didn't effectively contribute to the occurrence of accidents as they were mostly taken place when the weather was fine with no high winds. Mostly, the accidents resulted in slight severity with not much consequences. Only 1% of the casualties resulted with fatal injuries. Generally, males were injured in the accidents and these accidents occurred during the peak hours (commuting to work), that is, 8-9 and 16-17 hrs.

The study also found that number of accidents taking place is mostly with cars and other vehicle types are almost negligible as compared. The first point of impact during the crash is the front side of the vehicles (almost 50%) and can also be inferred that cars had low engine capacity with all types of fuel that could be one of possible reason for the greatest number of accidents.

The classification report is about a binary classification with linear SVM (Support Vector Machine). It can be seen that class 0 has a higher precision than class 1 and better recall as well. The support is the number of occurrences of the given class in the dataset, since it is 9854 of class 0 and 138 for class 1, it means that the dataset is imbalanced. It could be inferred that the model didn’t do well for this dataset even though the precision is good, the model had better predictions for class 0 as it neglected class 1.

The confusion matrix shows that this model predicted that 9854/9992 people did not suffer fatal injuries when there were actually 9967/9992 people not suffering fatal injuries. This model has an accuracy of 9839/9992 and the recall of this model is equal to 9834/9967, that is, 0.98%. This model is not fit for this imbalanced dataset and better data engineering is required.