Big Data and Data Mining

INTRODUCTION

Road accidents are one of the world's most important concerns, as they result in a large number of casualties, injuries, and deaths each year, as well as considerable financial losses. There exist many factors that contribute to road accidents. It may be possible to take efforts to lessen the damages and their severity if these factors can be better understood and predicted. The purpose of the report is to use the UK road accidents 2019 dataset and understand the factors that cause accidents. The prediction based on the factors will undoubtedly reduce the road accidents and advise government agencies about how to improve road safety.

ANALYSIS

Multiple data sets are analysed and cleaned to effectively address the below questions

```
print(accidents19_df.isnull().sum())
```

```
Accident_Index                                0
Location_Easting_OSGR                         0
Location_Northing_OSGR                        0
Longitude                                     0
Latitude                                      0
Police_Force                                  0
Accident_Severity                             0
Number_of_Vehicles                            0
Number_of_Casualties                          0
Date                                          0
Day_of_Week                                   0
Time                                          0
Local_Authority_(District)                    0
Local_Authority_(Highway)                     0
1st_Road_Class                                0
1st_Road_Number                               0
Road_Type                                     0
Speed_limit                                   0
Junction_Detail                               0
Junction_Control                              0
2nd_Road_Class                                0
2nd_Road_Number                               0
Pedestrian_Crossing-Human_Control             0
Pedestrian_Crossing-Physical_Facilities       0
Light_Conditions                              0
Weather_Conditions                            0
Road_Surface_Conditions                       0
Special_Conditions_at_Site                    0
Carriageway_Hazards                           0
Urban_or_Rural_Area                           0
Did_Police_Officer_Attend_Scene_of_Accident   0
LSOA_of_Accident_Location                     0
dtype: int64
```

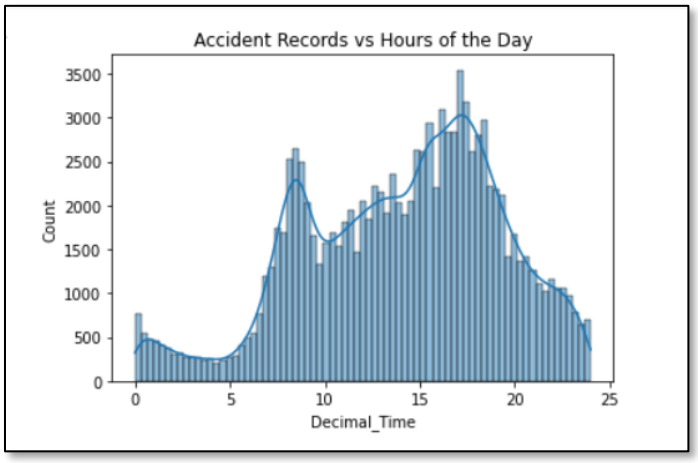
```
print(vehicles19_df.isnull().sum())
```

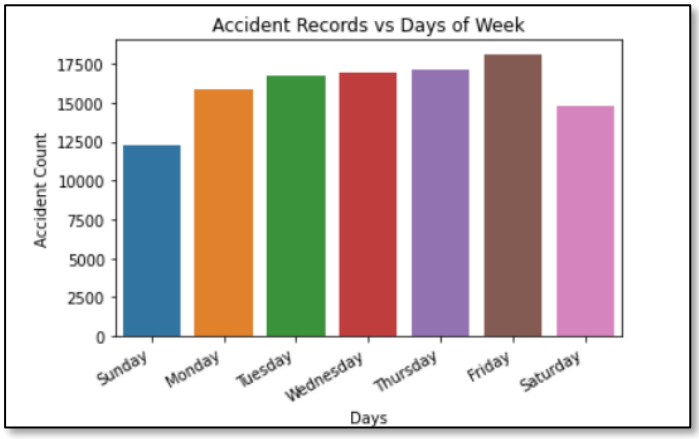
```
Accident_Index                    0
Vehicle_Reference                 0
Vehicle_Type                      0
Towing_and_Articulation           0
Vehicle_Manoeuvre                 0
Vehicle_Location-Restricted_Lane  0
Junction_Location                 0
Skidding_and_Overturning          0
Hit_Object_in_Carriageway         0
Vehicle_Leaving_Carriageway       0
Hit_Object_off_Carriageway        0
1st_Point_of_Impact               0
Was_Vehicle_Left_Hand_Drive?      0
Journey_Purpose_of_Driver         0
Sex_of_Driver                     0
Age_of_Driver                     0
Age_Band_of_Driver                0
Engine_Capacity_(CC)              0
Propulsion_Code                   0
Age_of_Vehicle                    0
Driver_IMD_Decile                 0
Driver_Home_Area_Type             0
Vehicle_IMD_Decile                0
dtype: int64
```

```
print(casualties19_df.isnull().sum())
```

```
Accident_Index                        0
Vehicle_Reference                     0
Casualty_Reference                    0
Casualty_Class                        0
Sex_of_Casualty                       0
Age_of_Casualty                       0
Age_Band_of_Casualty                  0
Casualty_Severity                     0
Pedestrian_Location                   0
Pedestrian_Movement                   0
Car_Passenger                         0
Bus_or_Coach_Passenger                0
Pedestrian_Road_Maintenance_Worker    0
Casualty_Type                         0
Casualty_Home_Area_Type               0
Casualty_IMD_Decile                   0
dtype: int64
```

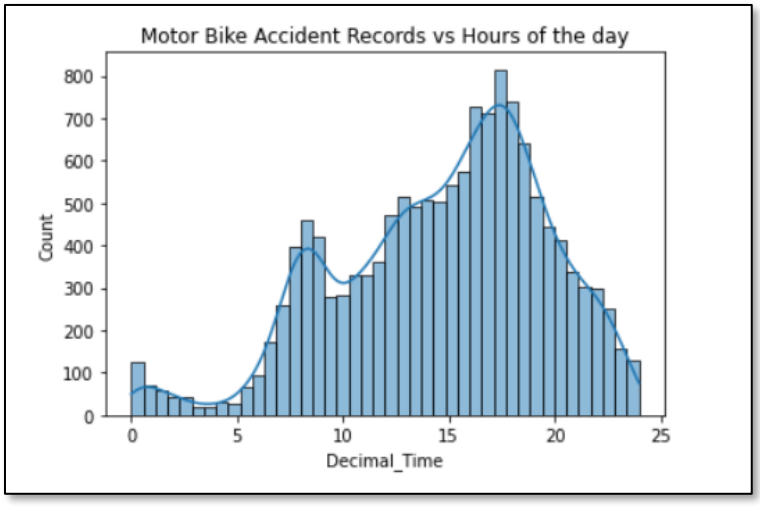
(a) Are there significant hours of the day, and days of the week, on which accidents occur?

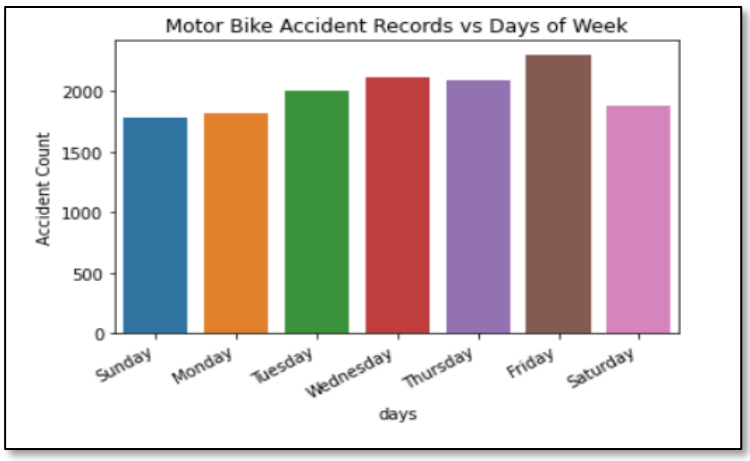




While considering the hours of the day, accidents peaked between **3 pm - 8pm**. For the Days, most of the accidents reported on **Friday & Thursday**. This time and day are considered as the rush hour, and on the final day of the working week. Also, it will be harder to understand speed and distance on evening as it becomes dark.

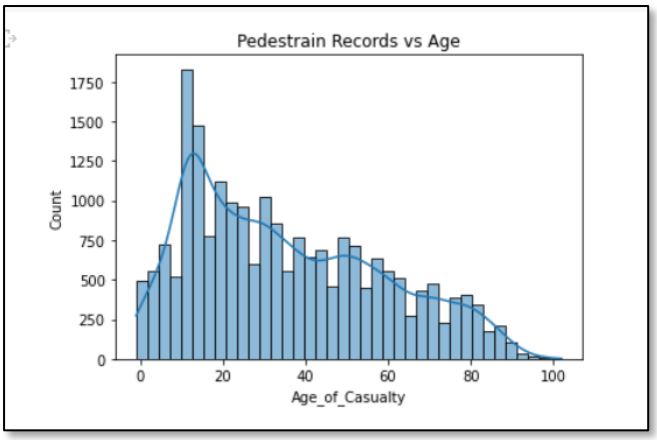
(b) For motorbikes, are there significant hours of the day, and days of the week, on which accidents occur?

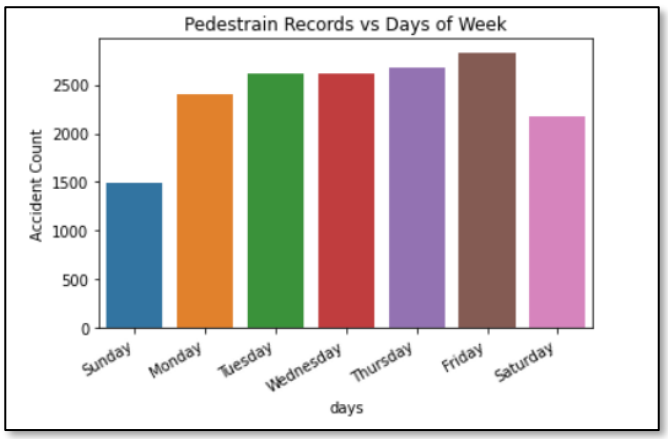


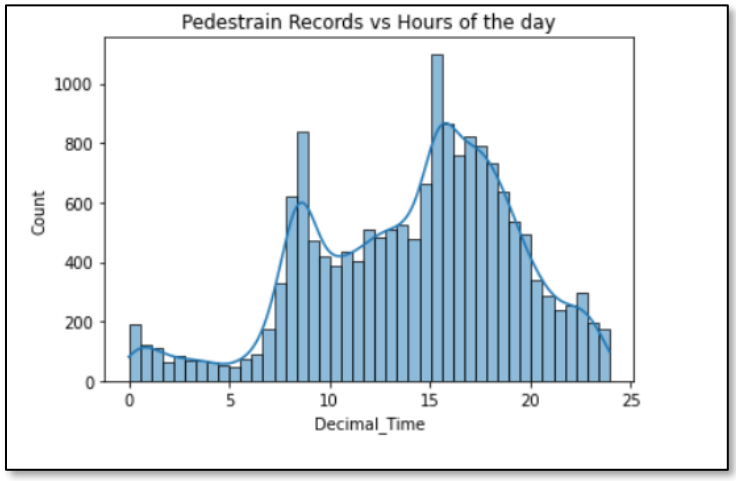


These reports show that about **50%** of all motorbike accidents occur between the hours of **3 - 8 pm.** Overall, it seems that most motorbike accidents occur on **Friday**. Since the people are trying their best to get home from work to enjoy their weekend. Also, accidents can occur due to failing to notice a motorcyclist on the road.

(c) 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?



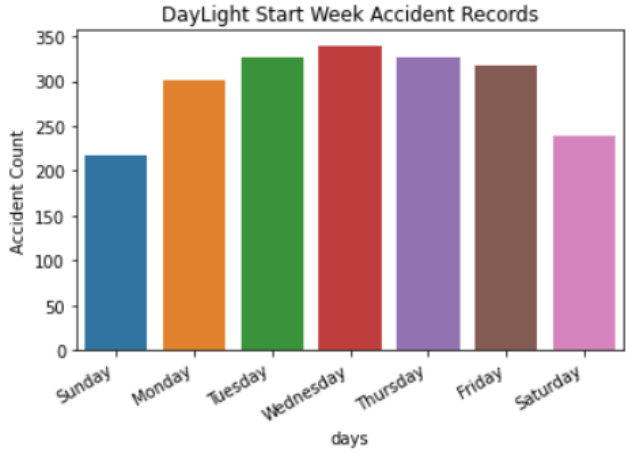




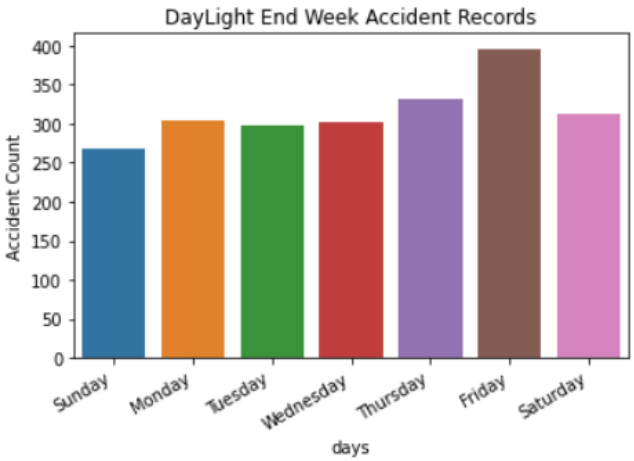
From the visualisations, it is clear that **Friday** has more likely to be involved in Pedestrian accidents Also, while analysing in Hours, the peak accident time ranges from **3pm – 8pm** and **8 am to 10 am** When comparing the age of Pedestrian, it ranges from **15-20** aged peoples. Mostly the youth fail to look properly and also fails to judge the vehicle's speed or path. Most accidents occur during the office hours. Since Pedestrians are non-protected, they are more at risk of being involved in accidents and sustaining injuries.

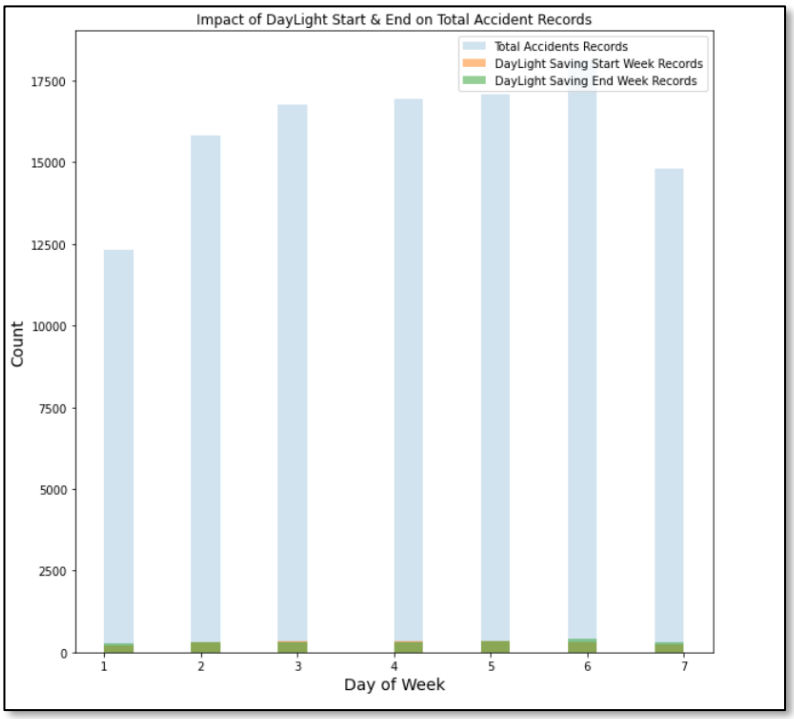
(d) What impact, if any, does daylight savings have on road traffic accidents in the week after it starts and stops?

```
Total Number of Accident Records Occured in a Week after Day light Starts:- 2067
[Text(0.5, 1.0, 'DayLight Start Week Accident Records')]
```



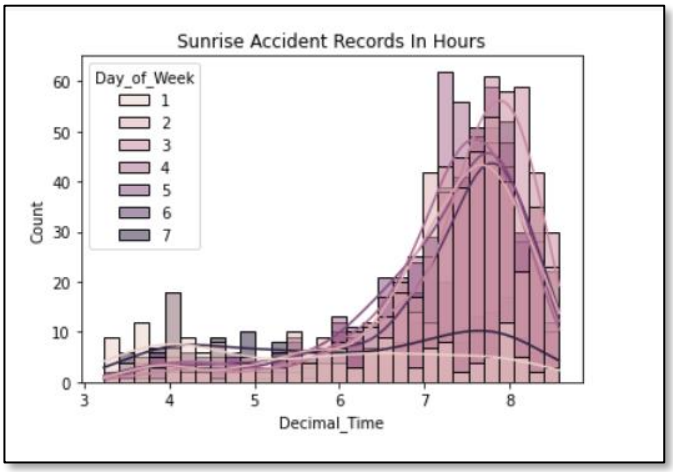

```
Total Number of Accident Records Occured in a Week after Day light Ends:- 2214
[Text(0.5, 1.0, 'DayLight End Week Accident Records')]
```

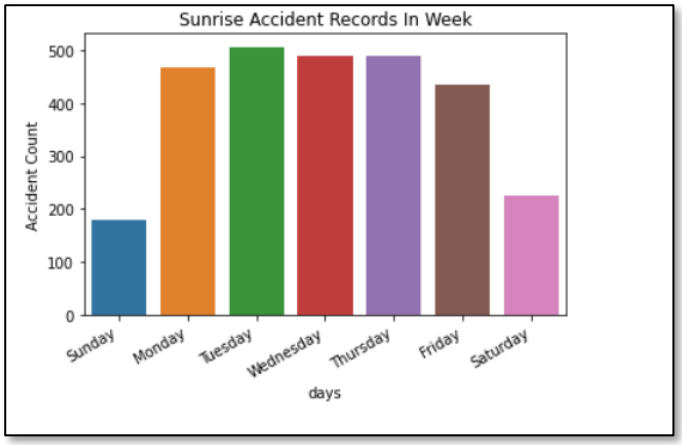


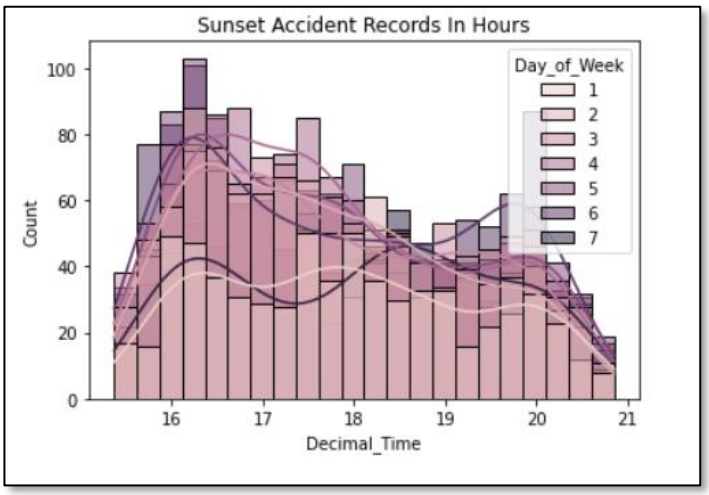


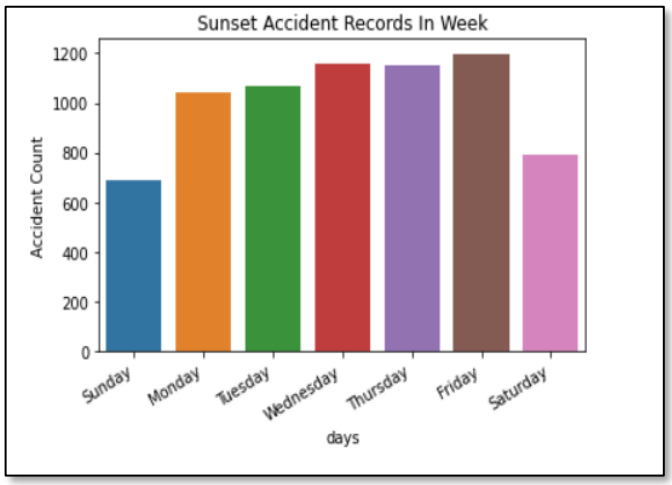
The daylight saving has a minor impact on the accidents rate. The total number of accidents occurred after the day light starts are **2067 in which Wednesday has the higher rate.** Also, the total number of accidents occurred after the daylight ends are **2214 in which Friday has the higher rate.** So more than **4000** accidents are occurred in the daylight start and end week time which cannot be neglected. The annual time change causes more sleep loss, which causes jet lag-like symptoms that make people drowsy behind the wheel. Thus, Daylight causes interrupted sleep schedules which can be factor for the accidents occurred in 2019.

(e) What impact, if any, does sunrise and sunset times have on road traffic accidents?





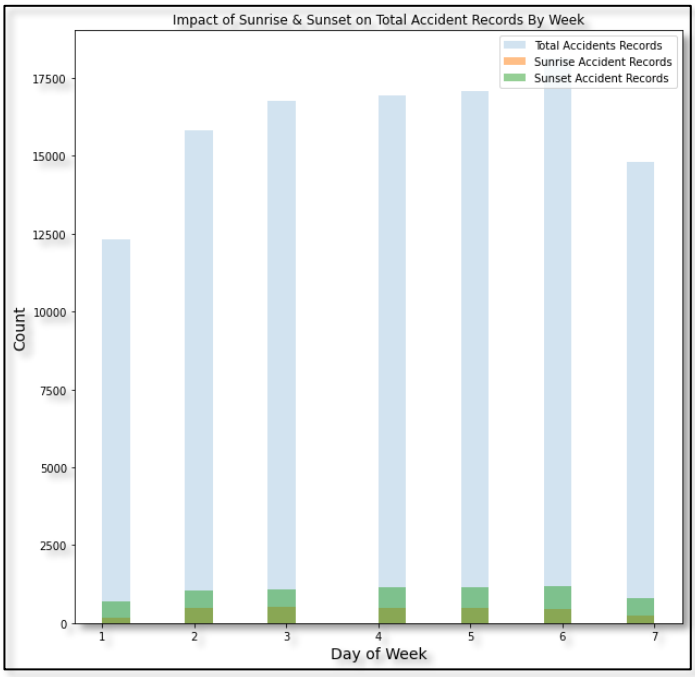


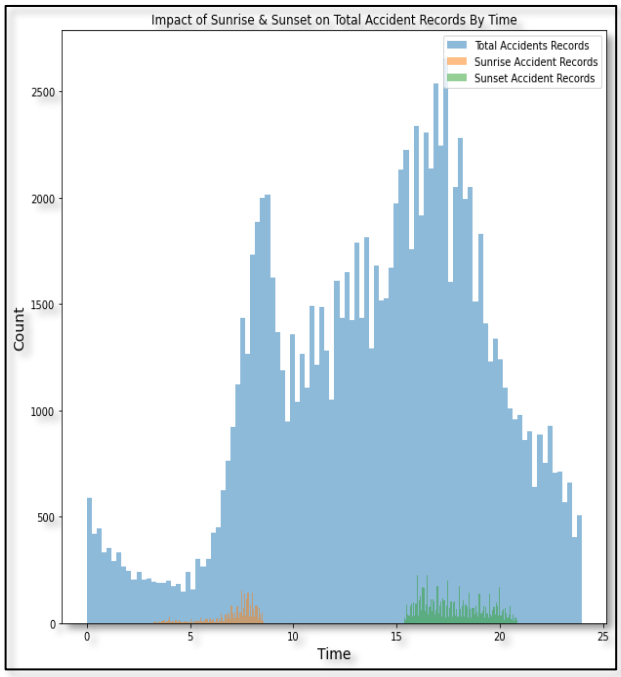



```
Total Number of Accidents recorded in Sunrise 2789
Total Number of Accidents recorded in Sunset 7096
```

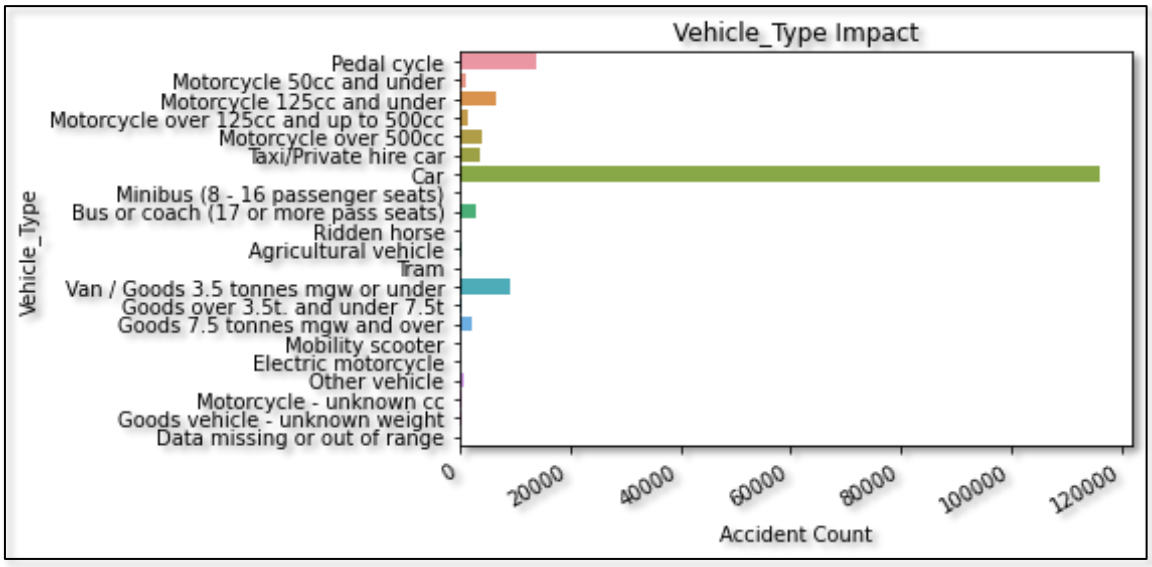
	Count	Avg per day Accidents
Total accidents occurred in sunrise	2789	398
Total accidents occurred in sunset	7096	1014
Total Accidents	117536	322

On comparing the Sunrise accident records, the peak time is **7 am to 8am on Week days.** Similarly on comparing the Sunset accident records, the peak time is **4pm to 6pm on Week days.** The sun can shine straight into drivers' eyes to the shifting level of brightness. During sunset, even if the sky is still light, the road will be darker, with deeper shadows and less colour contrast.

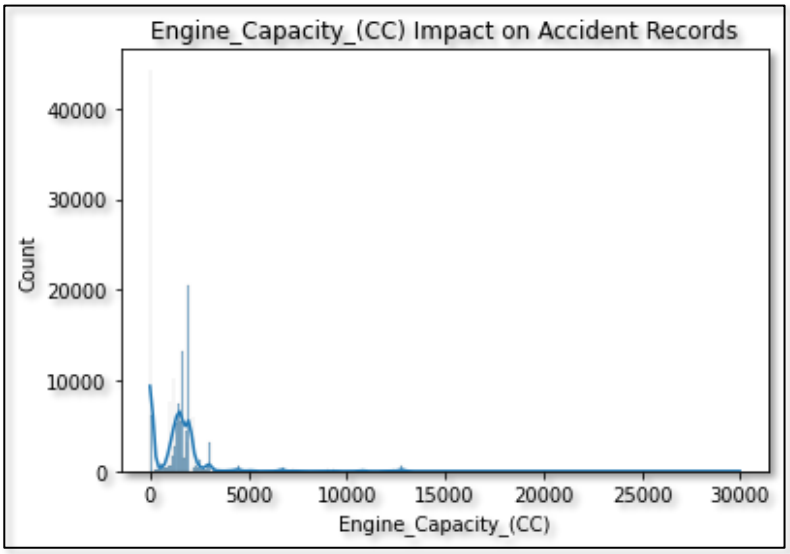




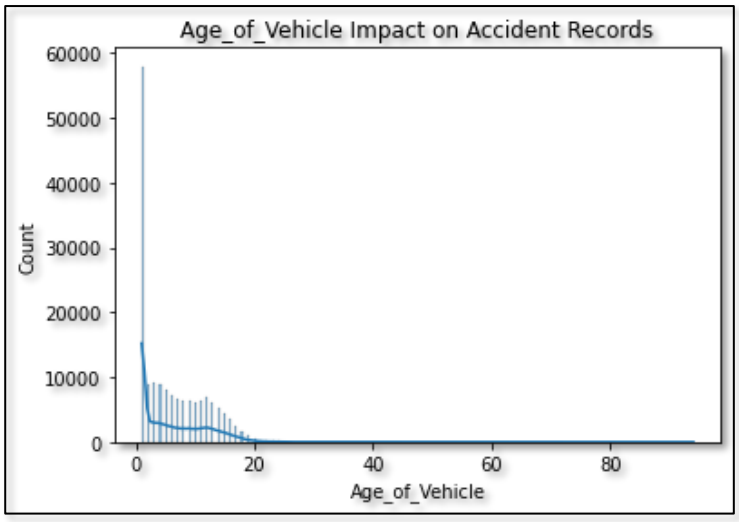
(f) Are there particular types of vehicles (engine capacity, age of vehicle, etc.) that are more frequently involved in road traffic accidents?



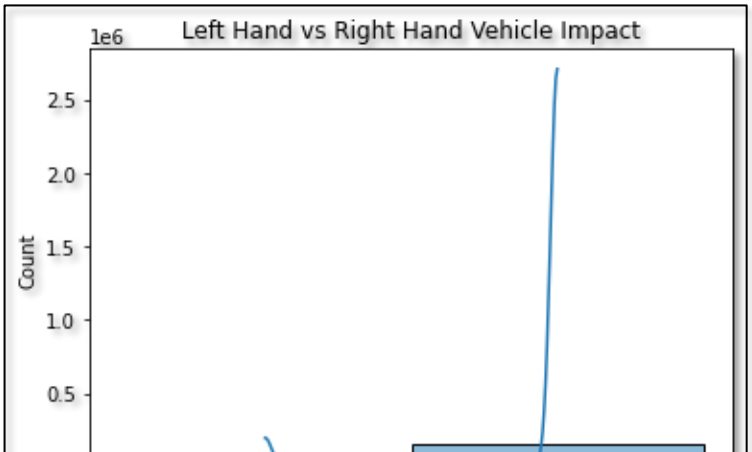
From the above visualisation, it is clearly understandable that **Car** is most likely to be in accident compared to any other vehicle type. Distraction during driving can be a reason for the accident. The other factors that contribute the peak in car accidents such as Driver fatigue, smoking, alcohol, over speeding and aggressive driving.



The vehicles that are below 5000cc are frequently occuring in the accidents.



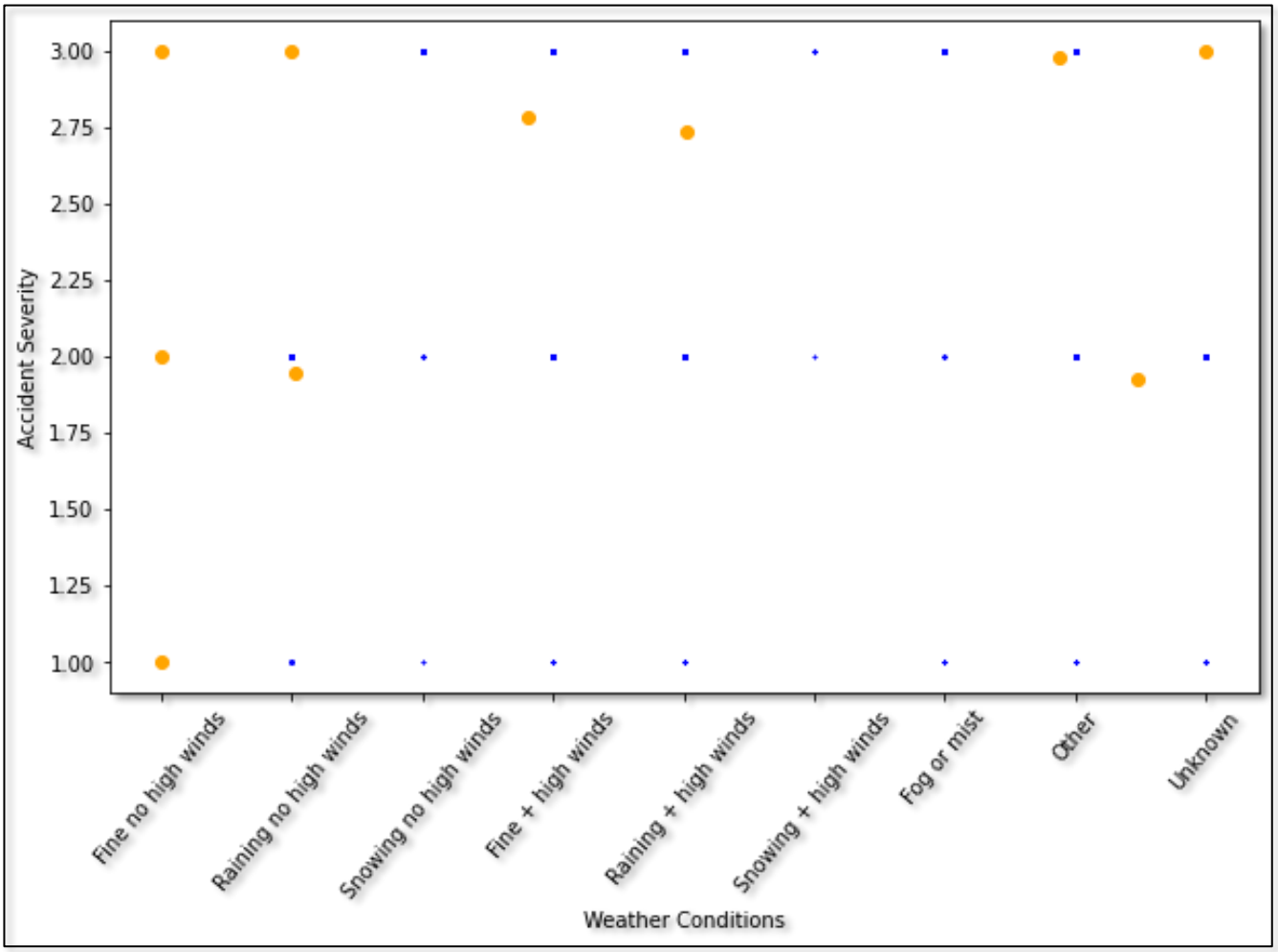
The age of the vehicle that are frequently occurred in road accidents ranges from 0 to 20 years old. Since most of the vehicles comes under this range, it can be concluded that vehicles that are under 20 years old has highest number of accidents reported

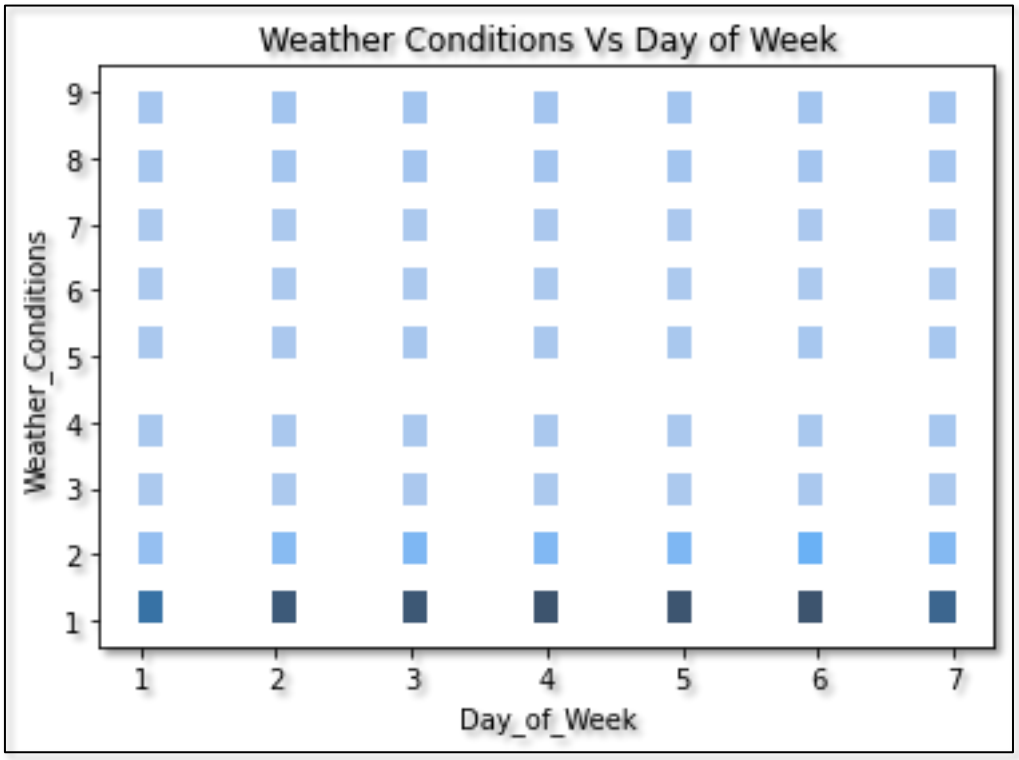


Since most of the vehicles in UK are Right Hand drive, the majority of the accidents that reported would be Right Hand Drive Vehicle.

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

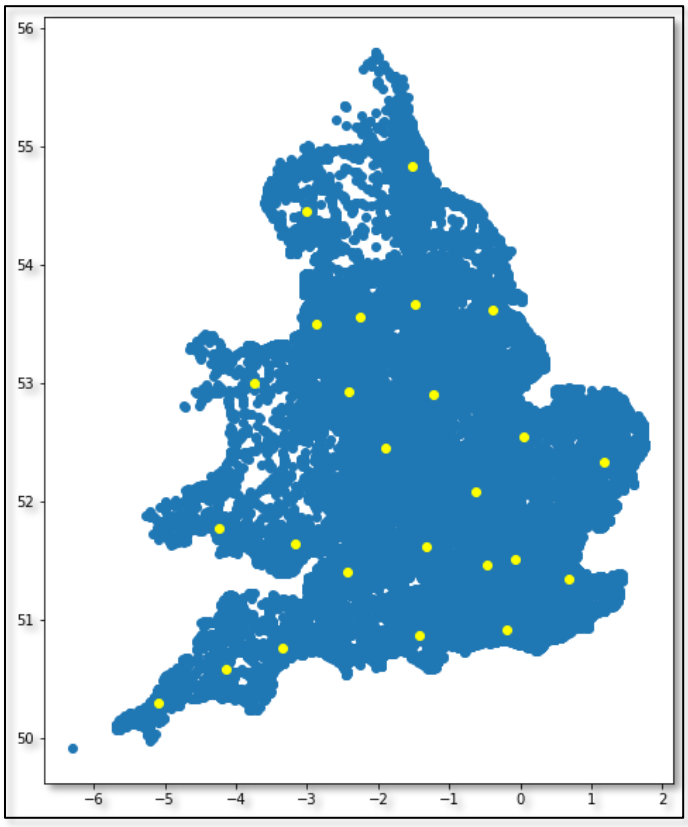
Weather Conditions: -





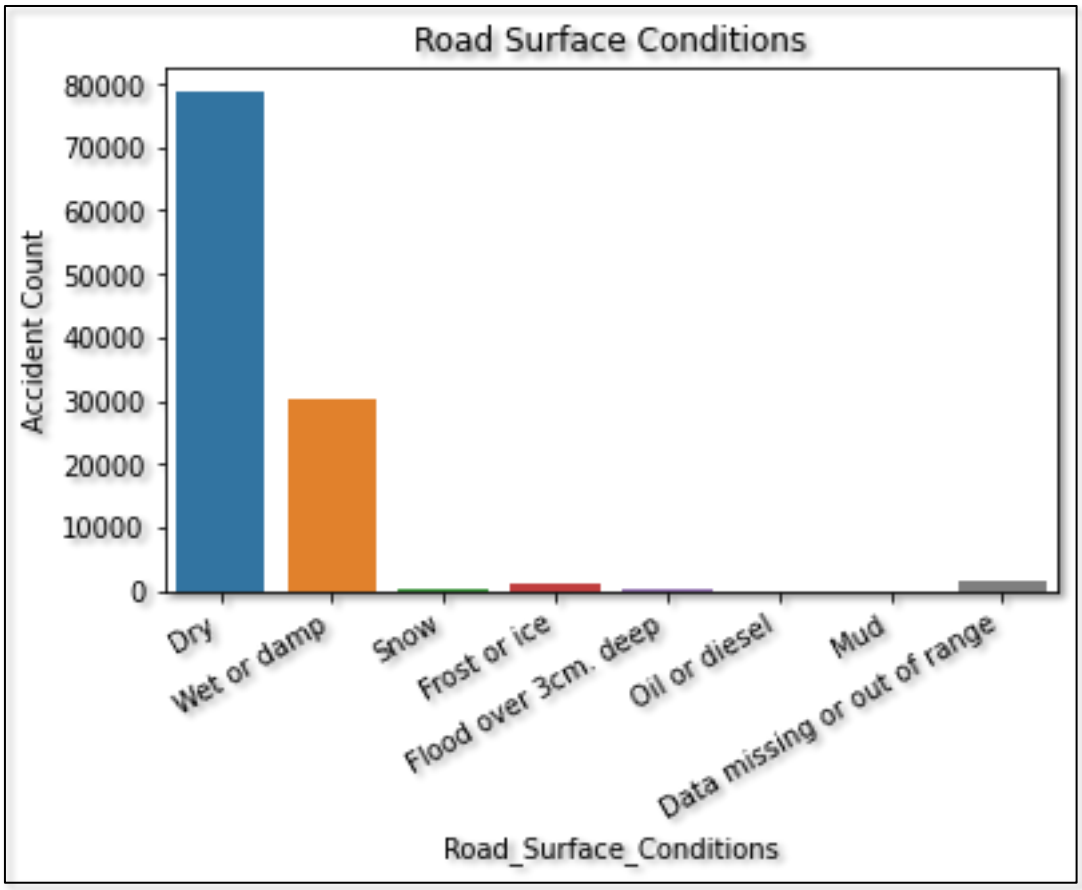
Regarding the Weather conditions, the most accidents occur on Fine and Raining day. When driving, rain is an inconvenient weather disturbance. Rain can cause vehicle to skid, drive erratically, and impair your vision, particularly at night. Fog, rain, snow, and ice can not only make sight difficult, but they can also influence the performance of your car. Lines on the road may be obscured by snow, and slippery patches might catch you off guard and cause your vehicle to spin out of control.

Geographical Locations:

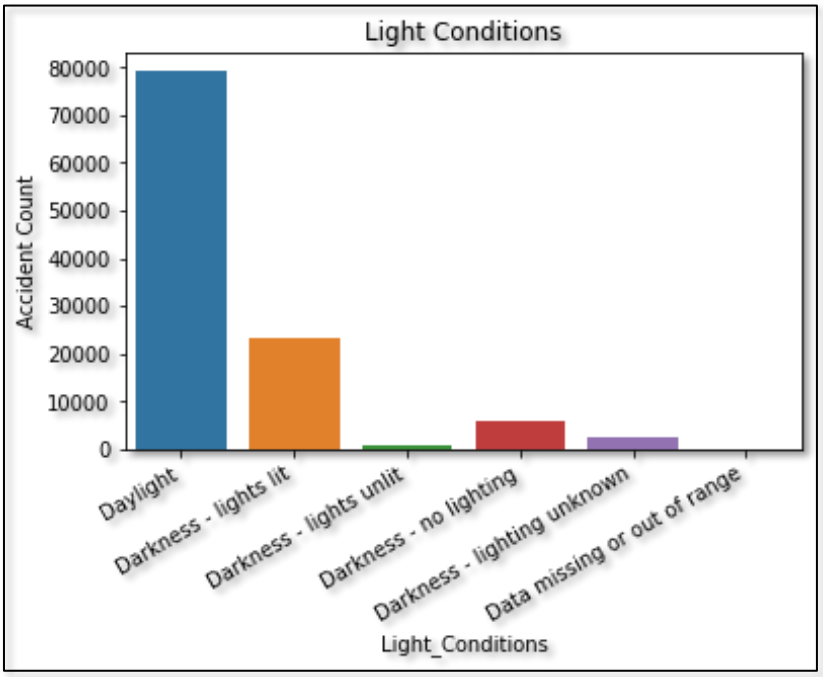


From above visualisation, it clearly understandable that accidents occur almost all over the UK. The highlighted points are the centroids cluster points. These areas have high accident rate while compared to the rest of the area. Most of the points cover crowded cities and high traffic areas.

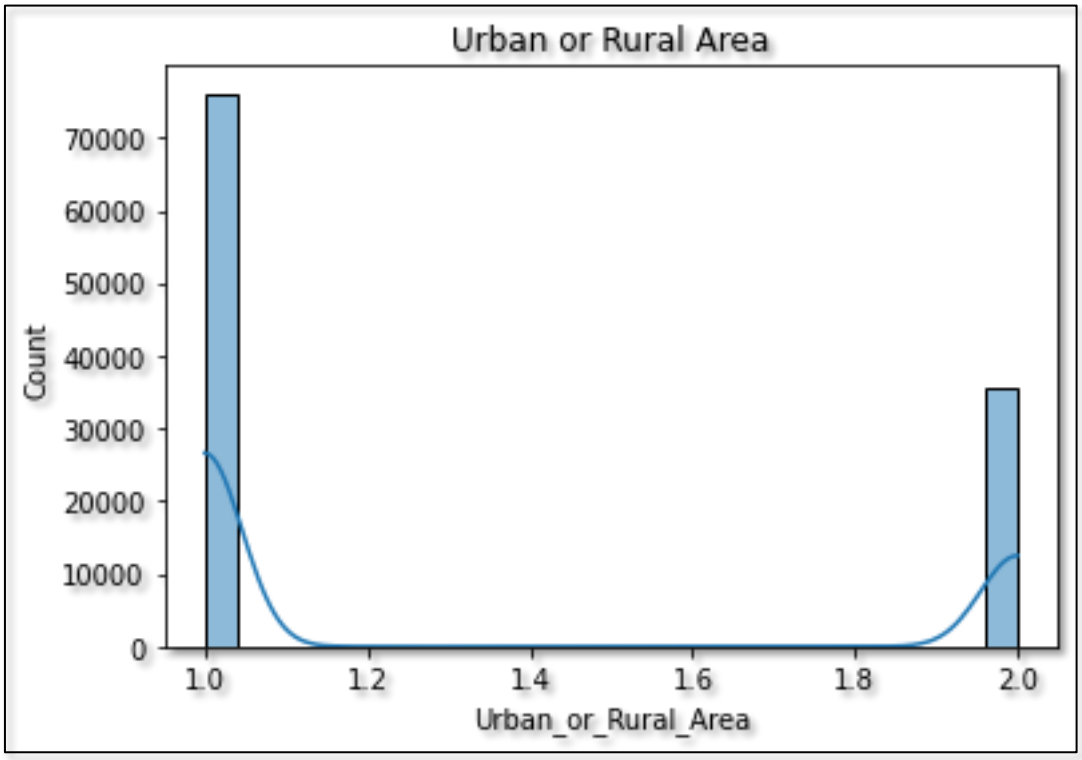
Conditions:



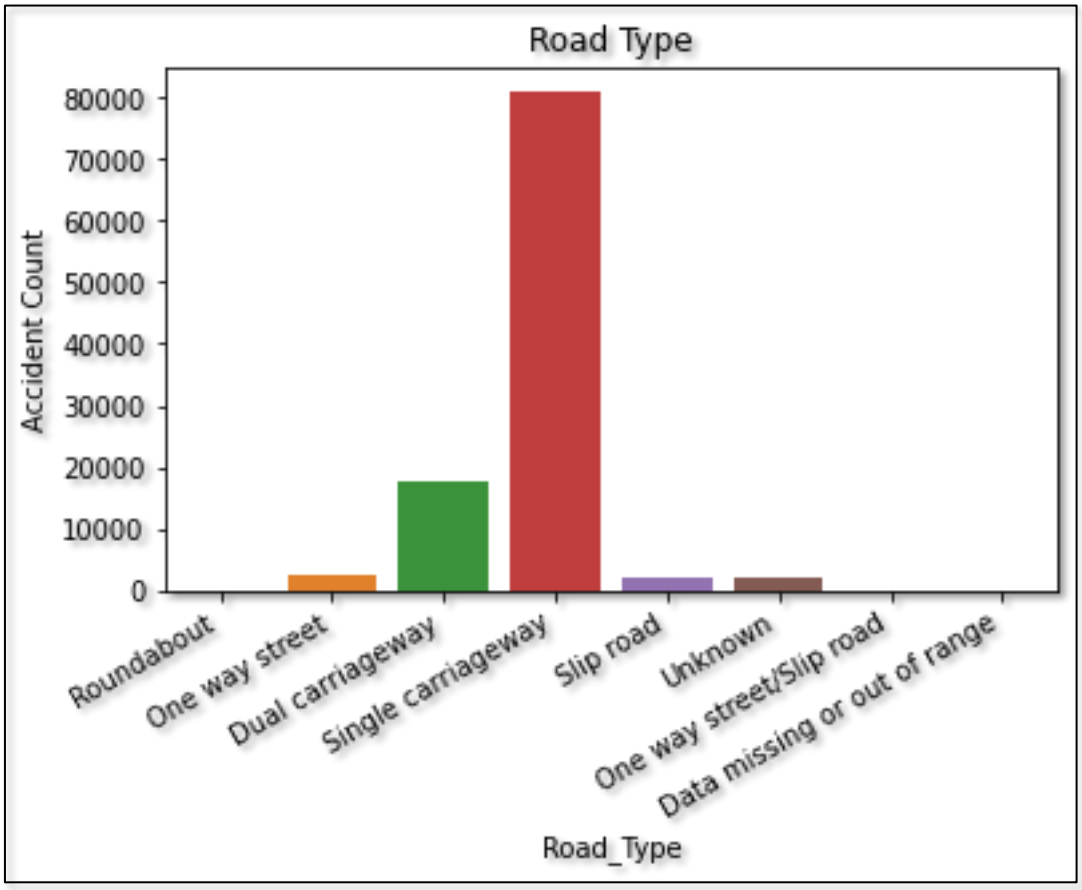
The most of the accidents reported are on Dry Road Surface conditions.



Above visualisation shows most of the accidents are reported in **Daylight time and Darkness time.** During daytime, as it is rush hour, traffic loads are higher, and drivers are more agitated on the road.

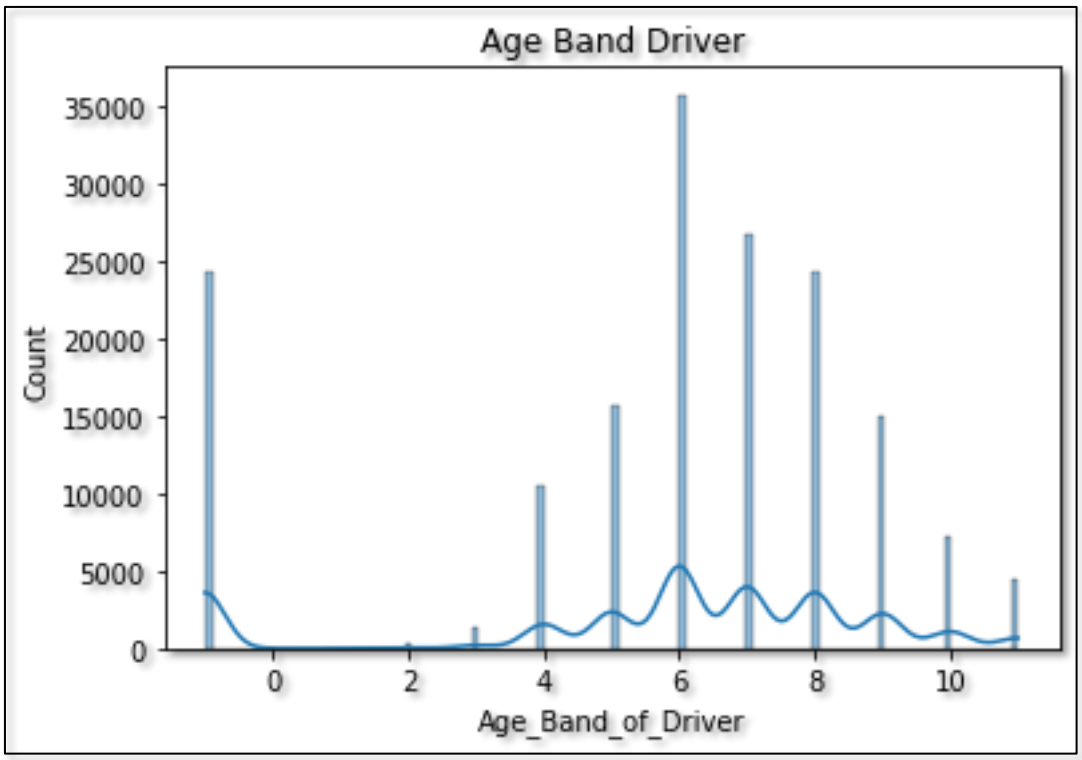


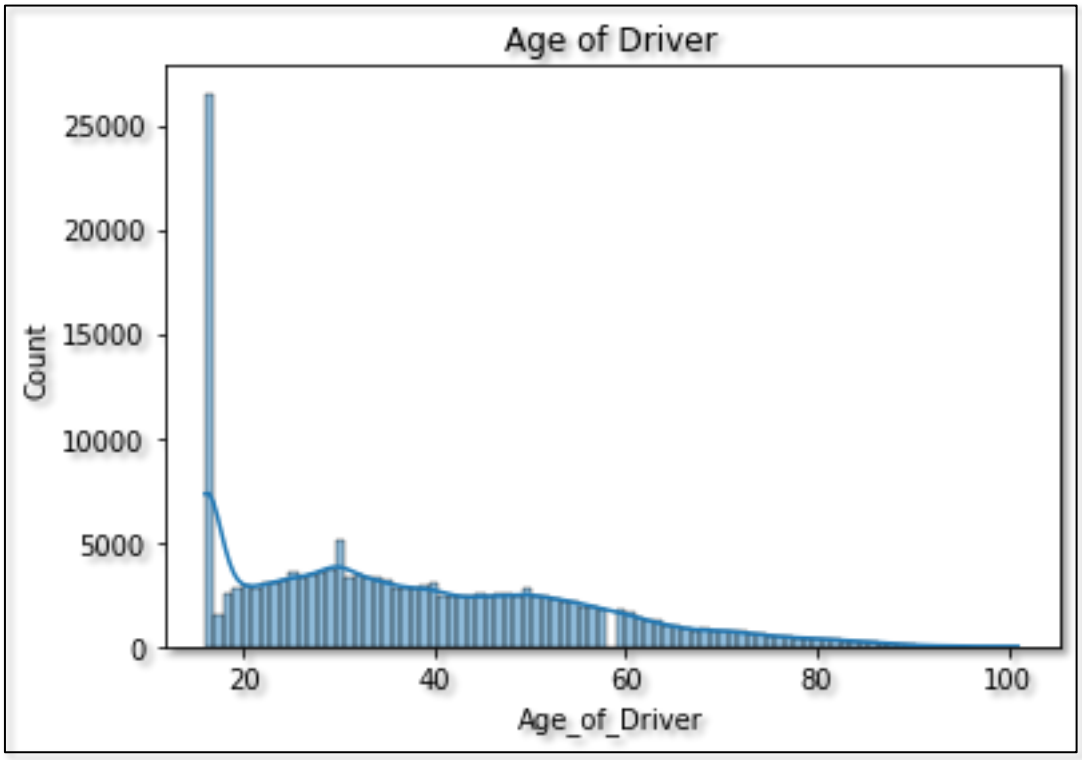
Most of the accidents are reported on Urban areas than the Rural areas. Since the population would be higher in the Urban areas when compared to Rural areas.



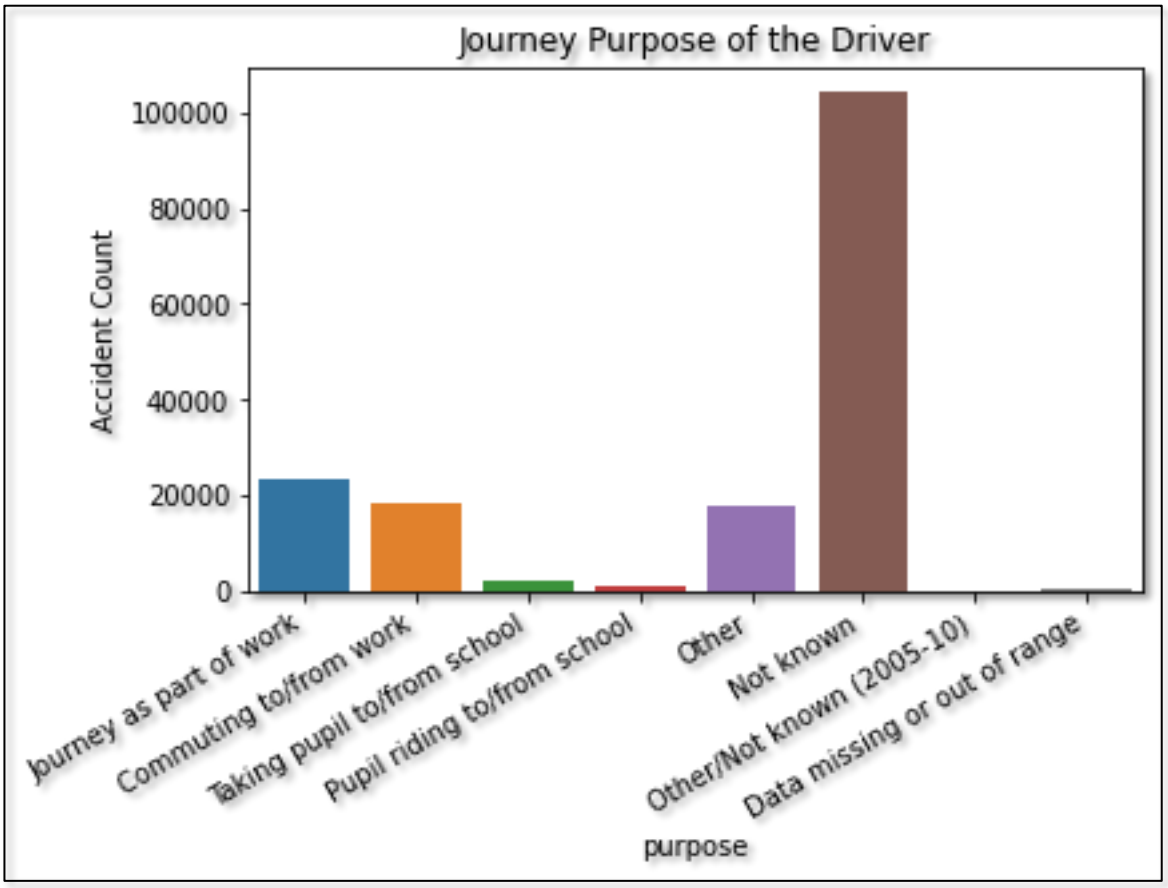
The most of the accidents are reported on Single carriageway and Dual Carraigeway.

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

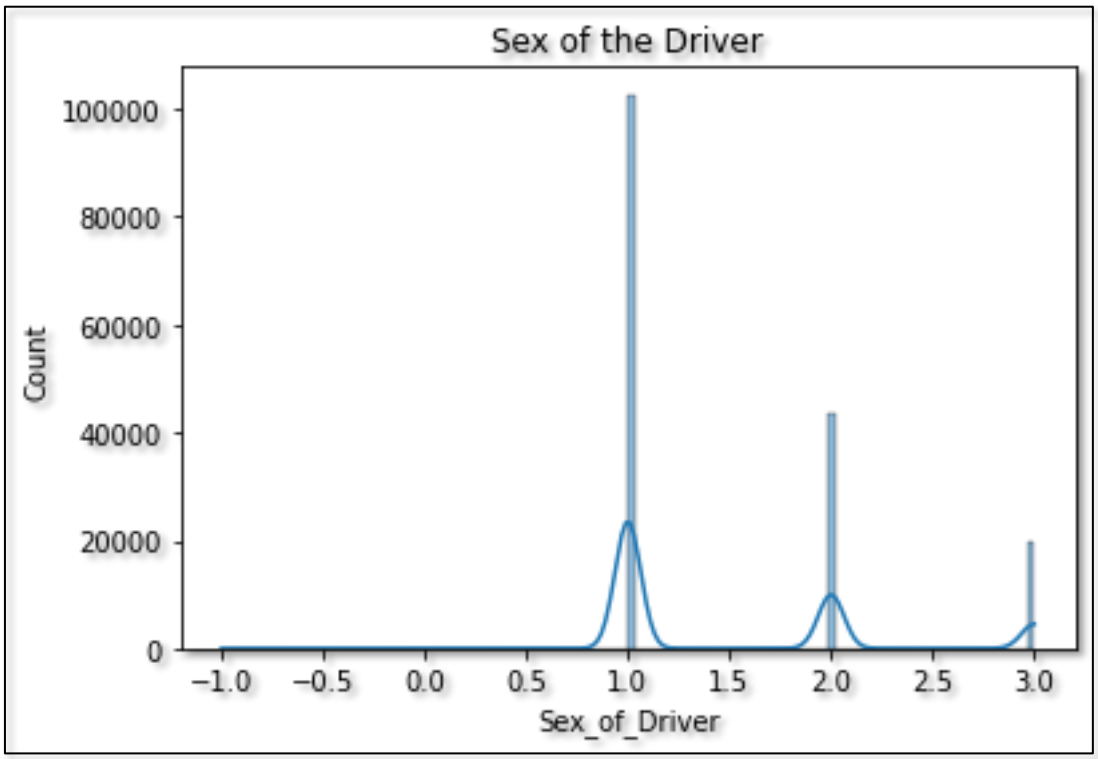




Young vehicle drivers, ages 17 to 25, are significantly overrepresented in recorded road accidents when compared to elder car drivers, ages 30 and up. These age groups are new to the Road and its traffic rules and hence it increases chance of accidents due to carelessness while driving. Also, Old age people has poor vision compared to younger ones. This can also contribute to the accidents rate.



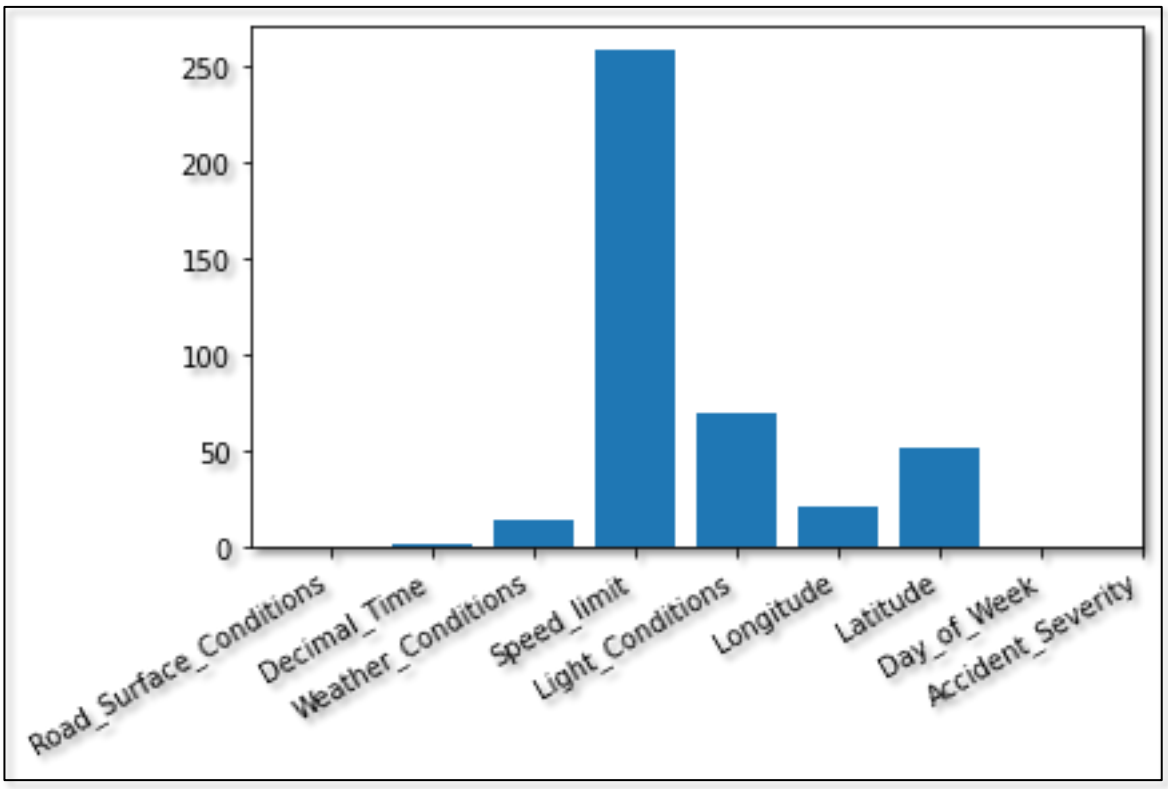
The most of the accidents are occurred during Journey as a part of work. Since most of the time will be involved in work and its commute.



The Female drivers are more in accident rate when compared to the Male drivers.

(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?

Feature Selection:



Feature Selection helps to remove unwanted columns from model Thus resulting in accuracy of the model.

Accident predictions accuracy based on the stacked model after feature selection is done. The stacked models include different algorithms to predict the accident occurrence based on the combination of features.

```
>Random Forest 0.911  (0.002)
>knn 0.789  (0.004)
>Logistic Regression 0.806  (0.000)
>Naive Bayes 0.803  (0.001)
```

Here The trained model, Random Forest gives the high accuracy compared to other models.

Then the government model accuracy is calculated to understand which model performs the prediction of accidents well using the dataset.

Accuracy Score of Government Model

```
[ ]  y_true = accident_merge_gov['Accident_Severity']
     y_prod = accident_merge_gov['Govt_Pred_Accident_Severity']
     accuracy_score(y_true, y_prod)
```

```
0.9443050482935005
```

The accuracy score is 94 % for the government model.

```
>Random Forest 0.842  (0.001)
>knn 0.765  (0.003)
>Logistic Regression 0.779  (0.000)
>Naive Bayes 0.774  (0.002)
```

Our trained model gives maximum accuracy of 84 % .

	Random Forest	knn	Logistic Regression	Naive Bayes
0	0.840915	0.763505	0.779013	0.771905
1	0.839493	0.765572	0.779013	0.774102
2	0.841432	0.757302	0.779013	0.773326
3	0.841820	0.768674	0.779013	0.772163
4	0.840915	0.762988	0.779013	0.775911
5	0.840398	0.763892	0.779013	0.776816
6	0.842078	0.764539	0.779013	0.775523
7	0.841561	0.766865	0.778883	0.775265
8	0.843241	0.767511	0.778883	0.773456
9	0.845160	0.765284	0.778984	0.774977

On comparison with the government model, our model has the lowest accuracy score for predicting the accidents based upon of the features. Thus government model can be used for predicting the accidents more accurately.