A PROJECT REPORT ON 2019 ACCIDENT DATA IN UNITED KINGDOM.

Abstract

This is a project on 2019 accident data in the United Kingdom. It contains visualisations, insights and recommendations as well as procedures followed to achieved these outcomes. It also contains a trained model to make predictions for accident severity.

WALLACE TUDEME

Table of Contents

List of Figures

Introduction

About The Dataset

The data to be explored is the 2019 accident data of the United Kingdom.

Aims and Objectives

1. Provide insights with visualisations on trends found.
2. Advise the government on possible measures that could help reduce the risk of accidents or accident severity.
3. Build an artificial intelligent model that would predict the accident severity.

Initial Data Exploration

The available data for our project consist of three datasets, namely:

- Accidents data
- Vehicles data
- Casualties' data

The accidents data has 32 columns and 117,536 rows of data. For casualties' data, it has 16 columns and 153,158 rows of data. While vehicles data has 23 columns and 216,381 rows of data. Combining all the datasets we generated a total of 68 columns and 109,518 rows of data.

Carrying-out further initial data insights with the "info" method on the combined dataset, some discoveries were made and are listed below;

- Some of the columns had missing data, these columns and the number of entries they posses are listed below as against the 109,518 entries they ought to possess.
    i. Location Easting OSGR - 109,478 entries
    ii. Location Northing OSGR - 109,478 entries
    iii. Longitude - 109,478 entries
    iv. Latitude - 109,478 entries
    v. Time - 109,457 entries
    vi. LSOA of Accident Location - 101,887 entries
- The Date series had a wrong data type as it was an Object data type instead of the required DateTime data type.

Having discovered this, there was the need for data cleaning.

Data Cleaning

Cleaning the listed columns above, I had to write a function to generate a dictionary of police force as keys and the mode value as per the given police force and the column in question as the value and thereafter replacing the empty cell as appropriate.

Cleaning the Location Easting OSGR, Location Northing OSGR, Longitude and Latitude Data Series

For cleaning these data series, I had to write a function to create a dataframe of all the empty rows for each column and then used another function to compare the police force of each entry in individual columns with the already generated dictionary of police force for the column in question and thereafter replacing the empty cell as appropriate with value of the dictionary key.

Cleaning the Time Data Series

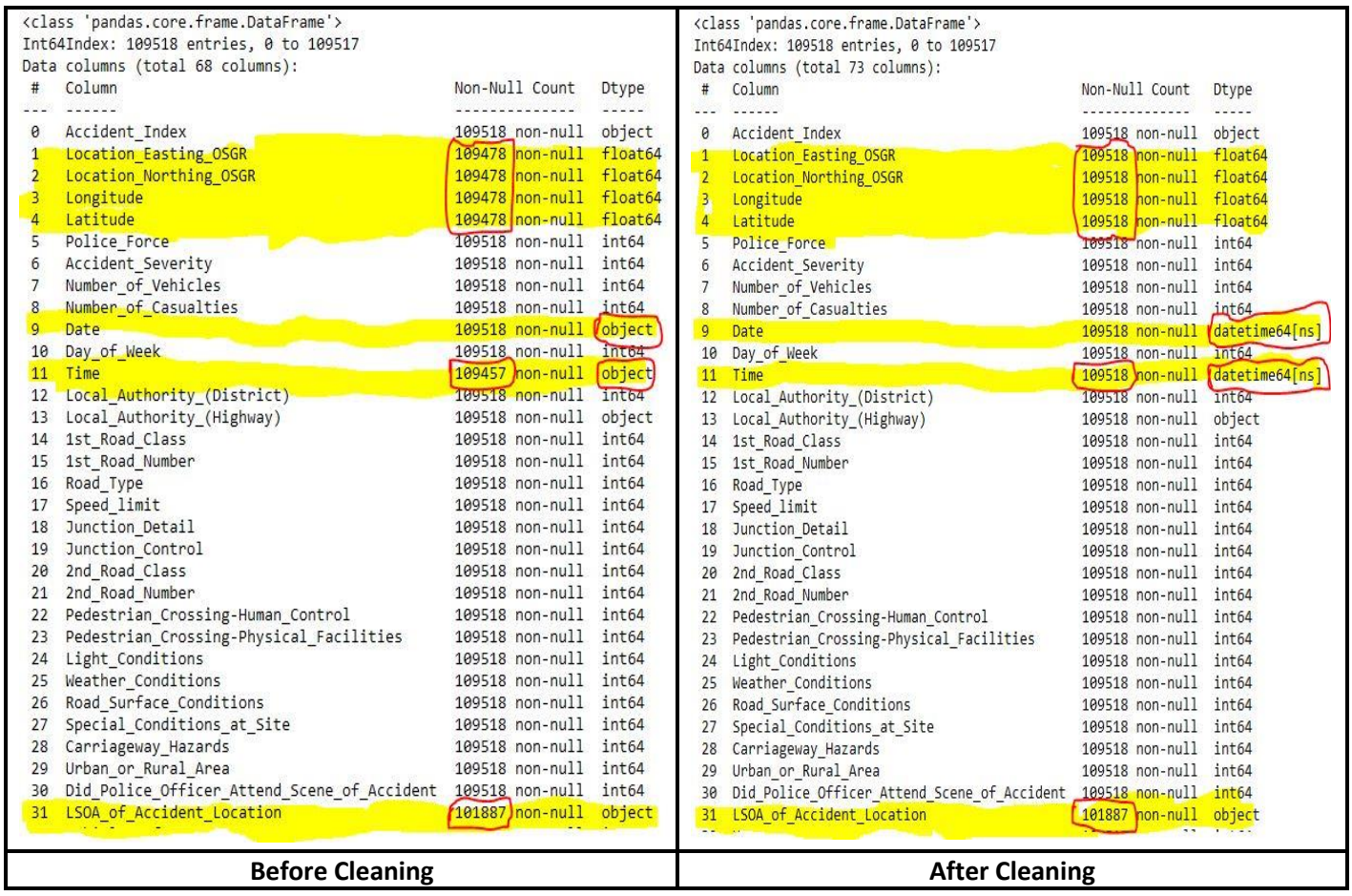
The same procedure was followed as in the cleaning procedure as the location easting osgr, location northing osgr, longitude and latitude data series. Thereafter the time series was converted from an Object data type to a DateTime data type.

Cleaning the LSOA of Accident Location Data Series

After exploring the LSOA of Accident Location data series, I discovered that the locations that had null entries were locations in Scotland. I went on to research why this was the case and discovered that Scotland does not use the LSOA as it is used by only England and Wales. Searching further to discover if there was any location in Scotland with an LSOA entry so I could replace them, I was unable find a suitable way to replace the data. Hence, I had to drop the column when I was to train my models (Reid et al., 2017).

Cleaning the Date Data Series

The date data series was converted to DateTime data type.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 109518 entries, 0 to 109517
Data columns (total 68 columns):
 #   Column                                      Non-Null Count   Dtype
---  ------                                      --------------   -----
 0   Accident_Index                              109518 non-null  object
 1   Location_Easting_OSGR                       109478 non-null  float64
 2   Location_Northing_OSGR                      109478 non-null  float64
 3   Longitude                                   109478 non-null  float64
 4   Latitude                                    109478 non-null  float64
 5   Police_Force                                109518 non-null  int64
 6   Accident_Severity                           109518 non-null  int64
 7   Number_of_Vehicles                          109518 non-null  int64
 8   Number_of_Casualties                        109518 non-null  int64
 9   Date                                        109518 non-null  object
 10  Day_of_Week                                 109518 non-null  int64
 11  Time                                        109457 non-null  object
 12  Local_Authority_(District)                  109518 non-null  int64
 13  Local_Authority_(Highway)                   109518 non-null  object
 14  1st_Road_Class                              109518 non-null  int64
 15  1st_Road_Number                             109518 non-null  int64
 16  Road_Type                                   109518 non-null  int64
 17  Speed_limit                                 109518 non-null  int64
 18  Junction_Detail                             109518 non-null  int64
 19  Junction_Control                            109518 non-null  int64
 20  2nd_Road_Class                              109518 non-null  int64
 21  2nd_Road_Number                             109518 non-null  int64
 22  Pedestrian_Crossing-Human_Control           109518 non-null  int64
 23  Pedestrian_Crossing-Physical_Facilities     109518 non-null  int64
 24  Light_Conditions                            109518 non-null  int64
 25  Weather_Conditions                          109518 non-null  int64
 26  Road_Surface_Conditions                     109518 non-null  int64
 27  Special_Conditions_at_Site                  109518 non-null  int64
 28  Carriageway_Hazards                         109518 non-null  int64
 29  Urban_or_Rural_Area                         109518 non-null  int64
 30  Did_Police_Officer_Attend_Scene_of_Accident 109518 non-null  int64
 31  LSOA_of_Accident_Location                   101887 non-null  object
```

Before Cleaning

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 109518 entries, 0 to 109517
Data columns (total 73 columns):
 #   Column                                      Non-Null Count   Dtype
---  ------                                      --------------   -----
 0   Accident_Index                              109518 non-null  object
 1   Location_Easting_OSGR                       109518 non-null  float64
 2   Location_Northing_OSGR                      109518 non-null  float64
 3   Longitude                                   109518 non-null  float64
 4   Latitude                                    109518 non-null  float64
 5   Police_Force                                109518 non-null  int64
 6   Accident_Severity                           109518 non-null  int64
 7   Number_of_Vehicles                          109518 non-null  int64
 8   Number_of_Casualties                        109518 non-null  int64
 9   Date                                        109518 non-null  datetime64[ns]
 10  Day_of_Week                                 109518 non-null  int64
 11  Time                                        109518 non-null  datetime64[ns]
 12  Local_Authority_(District)                  109518 non-null  int64
 13  Local_Authority_(Highway)                   109518 non-null  object
 14  1st_Road_Class                              109518 non-null  int64
 15  1st_Road_Number                             109518 non-null  int64
 16  Road_Type                                   109518 non-null  int64
 17  Speed_limit                                 109518 non-null  int64
 18  Junction_Detail                             109518 non-null  int64
 19  Junction_Control                            109518 non-null  int64
 20  2nd_Road_Class                              109518 non-null  int64
 21  2nd_Road_Number                             109518 non-null  int64
 22  Pedestrian_Crossing-Human_Control           109518 non-null  int64
 23  Pedestrian_Crossing-Physical_Facilities     109518 non-null  int64
 24  Light_Conditions                            109518 non-null  int64
 25  Weather_Conditions                          109518 non-null  int64
 26  Road_Surface_Conditions                     109518 non-null  int64
 27  Special_Conditions_at_Site                  109518 non-null  int64
 28  Carriageway_Hazards                         109518 non-null  int64
 29  Urban_or_Rural_Area                         109518 non-null  int64
 30  Did_Police_Officer_Attend_Scene_of_Accident 109518 non-null  int64
 31  LSOA_of_Accident_Location                   101887 non-null  object
```

After Cleaning

Figure 1 Summary snippet showing the combined dataset number of data entries before and after data cleaning.

Comprehensive Data Insights

Upon completion of the data cleaning process, the datasets were explored and charts and plots derived are outlined with implications explained as seen below;

Exploring the Time Series

Visualising the available data, there seem to be a trend of slight increase in the number of accidents between 8:00am and 9:00am and 4:00pm and 6:00pm daily. The time of the day with most recorded accidents is 5:00pm.

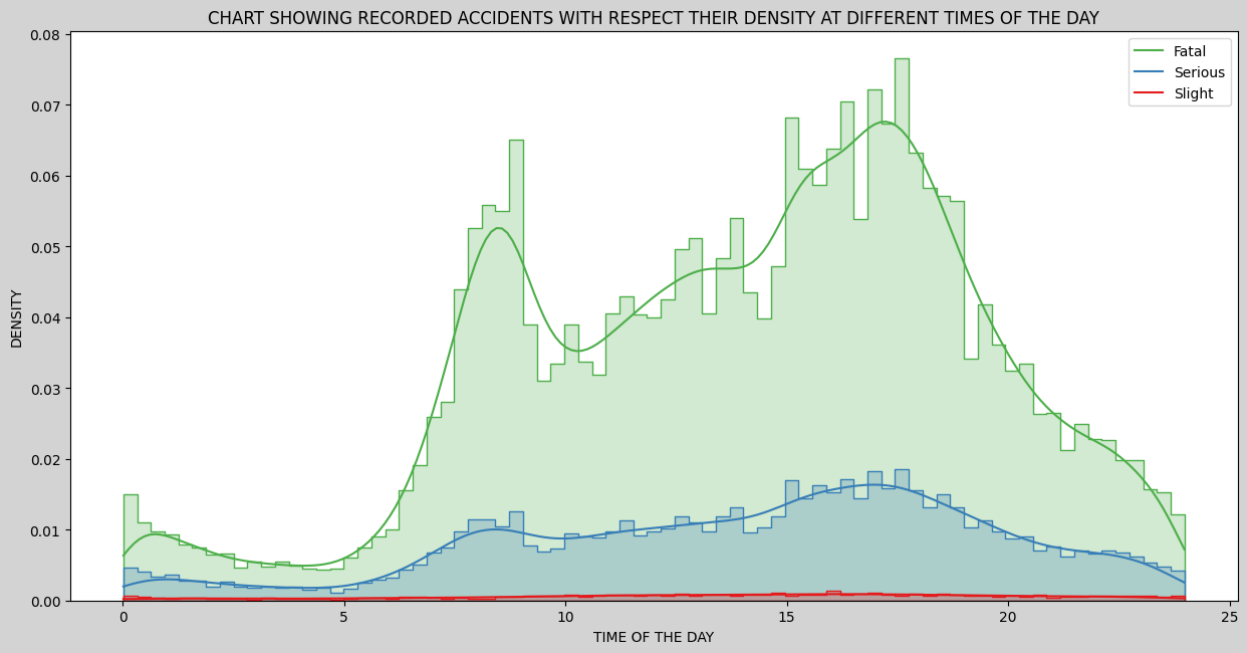


Figure 2 Density of accidents with regards to time of the day.

Also, a bulk of the accidents recorded seem to happen between the hours of 10:00am and 6:00pm daily as shown by the box plot below and has mean time of 2:04pm with a standard deviation 5.20 hours.

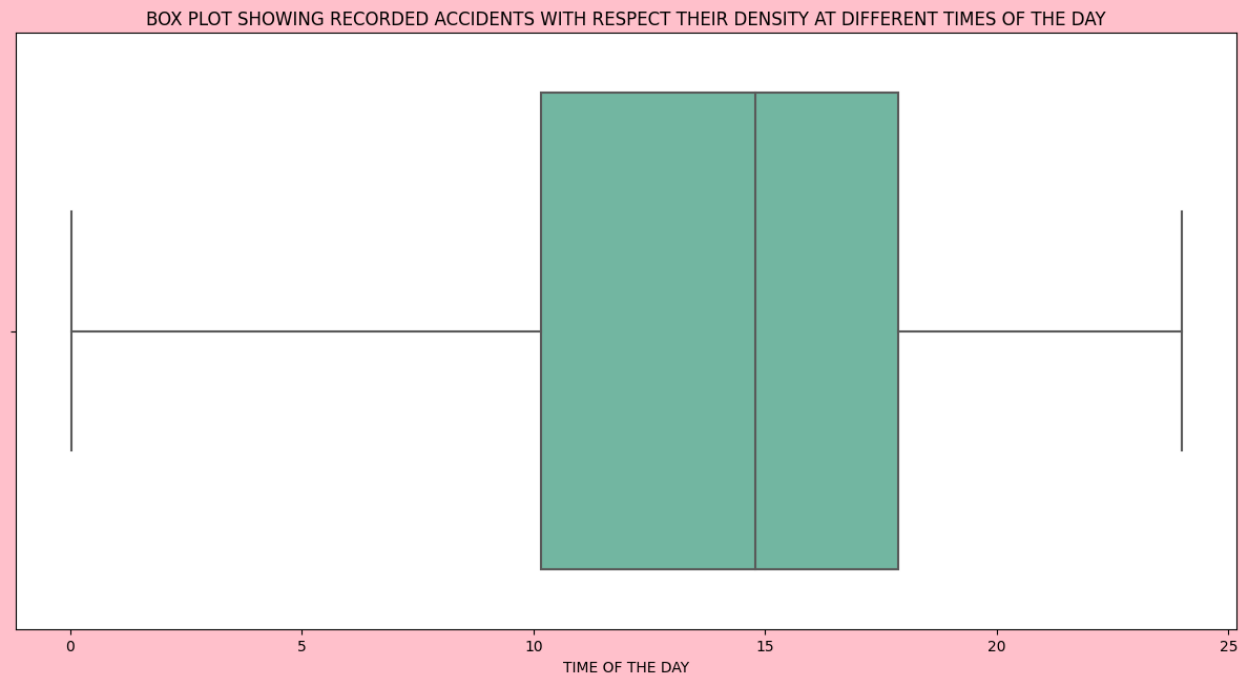


Figure 3 Box plot showing concentration of recorded accidents with regards to time of the day.

```
Most of the accidents happen at about 17.00 hours which is mode of the time data series
The  time of the day with the average number of accidents is 14.04 hours
The standard deviation of the time data series is 5.20 hours
```

Figure 4 Image showing result of time series exploration of mean, mode and standard deviation.

Exploring the Day of Week Series

There seem to be a trend of slow but steady rise in the number of accidents from the start of a business week down through to the end of the business week and the trend descends as we move to into the weekend (Saturday and Sunday).

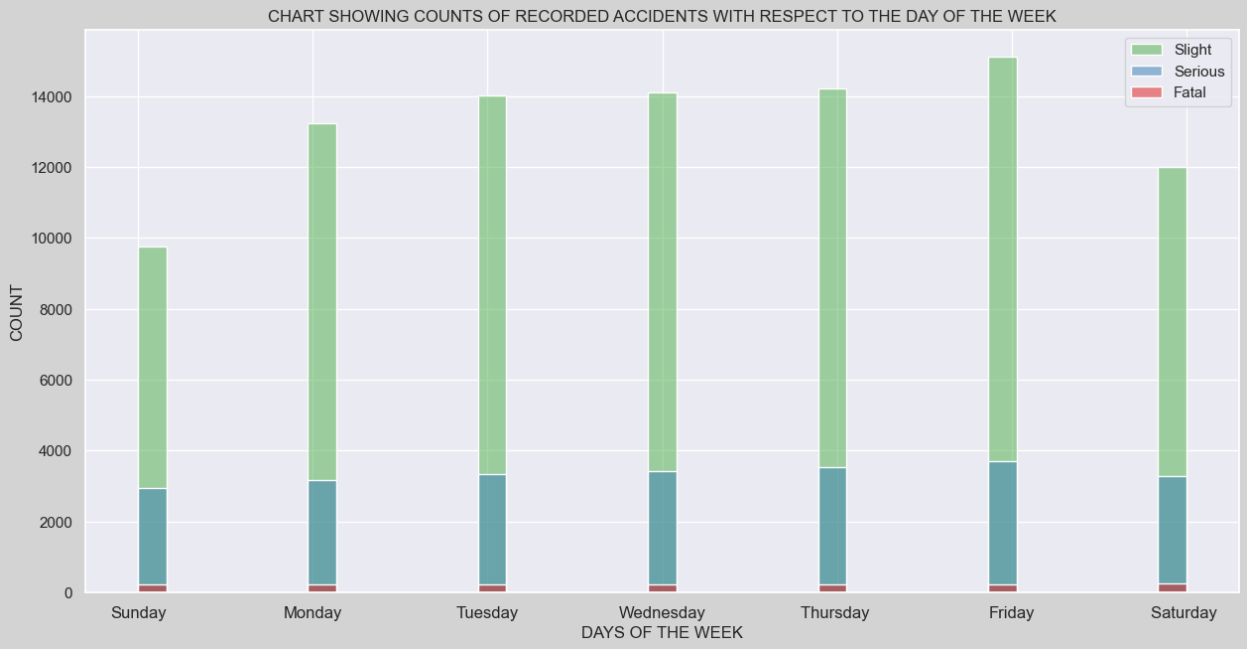


Figure 5 Plot showing accident count for each day of the week

Exploring the Transport Vehicles Data

Exploring Motorcycles

For motorcycles, the peak of the accidents appears to be at 5:00pm. Also there seem to be slightly higher percentage of fatal injuries on Sundays with 0.50 percent and Saturdays with 0.45 percent of the total motorcycle accidents. For time of the day, it follows the general time trend.

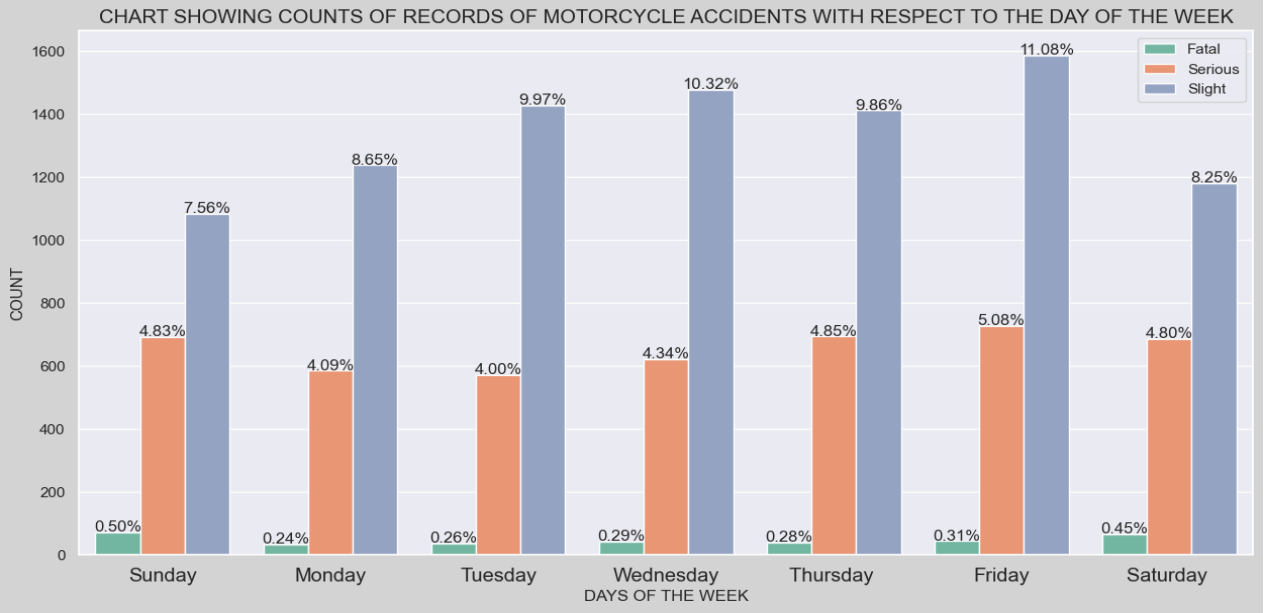


Figure 6 Motorcycle accidents day of the week plot

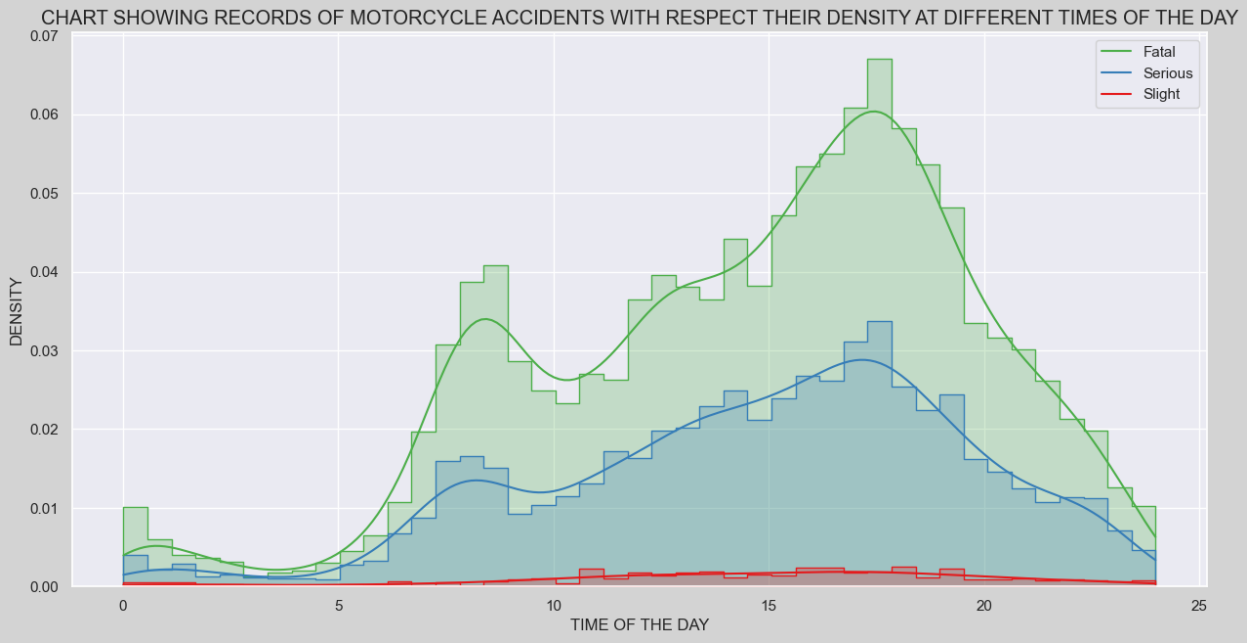


Figure 7 Motorcycle accidents time of the day plot.

Another observation on motorcycle accident records shows that motorcycles with over 500cc seem to have more fatal accidents than all other types of motorcycle with 1.66 percent of the total motorcycle accidents.

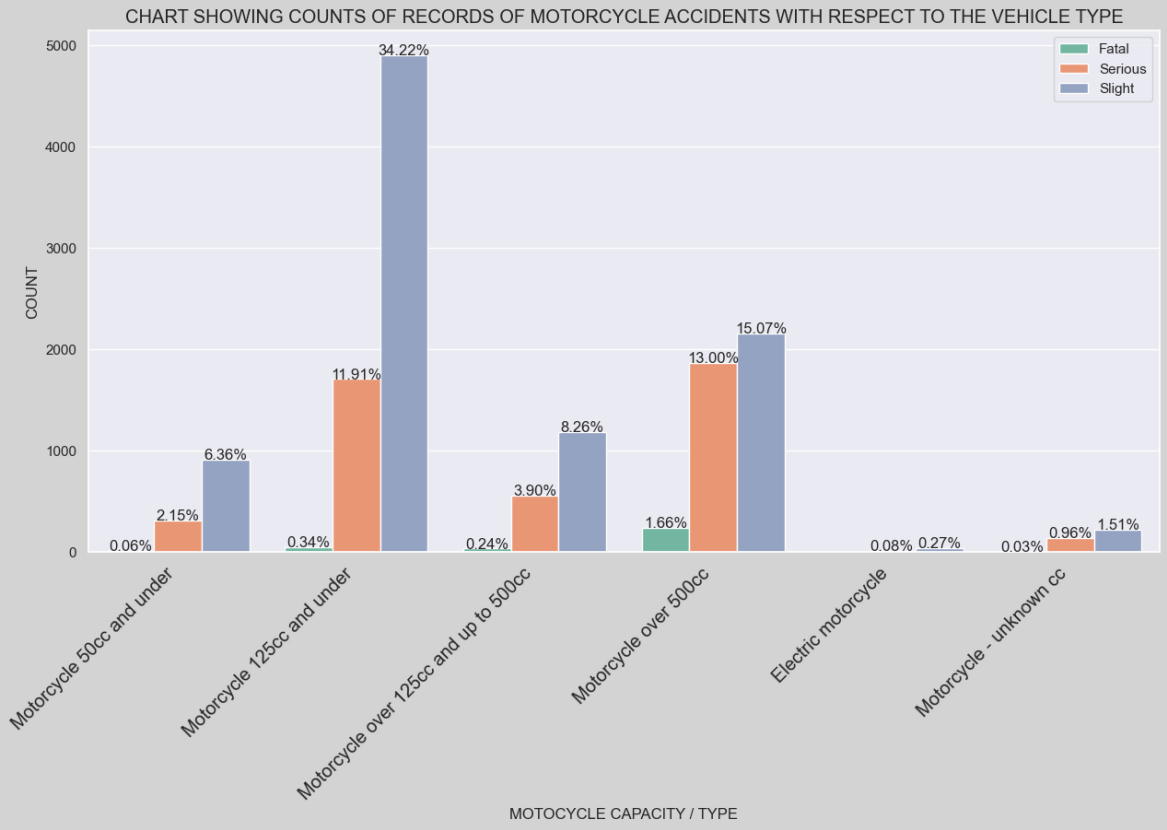


Figure 8 Motorcycle accident plot with respect to the vehicle type.

Exploring Pedestrians

Most of the accidents seem to occur around 3pm with increase in fatal accident severity from Thursday through Saturday.

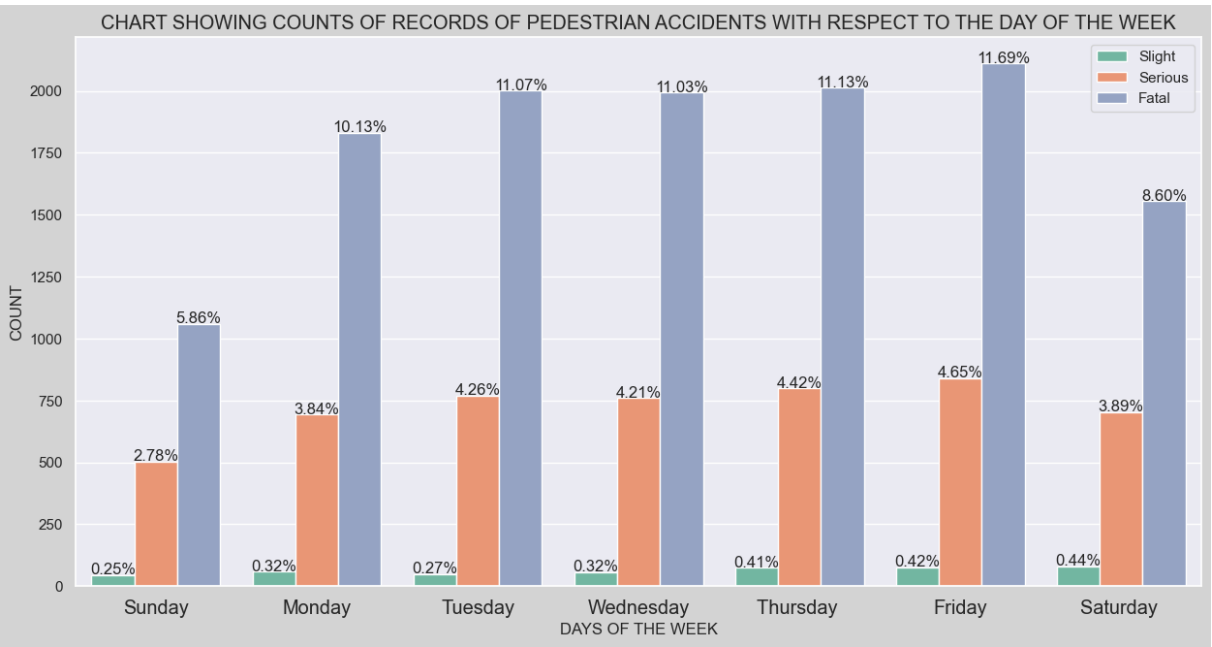


Figure 9 Pedestrian accidents day of the week plot.

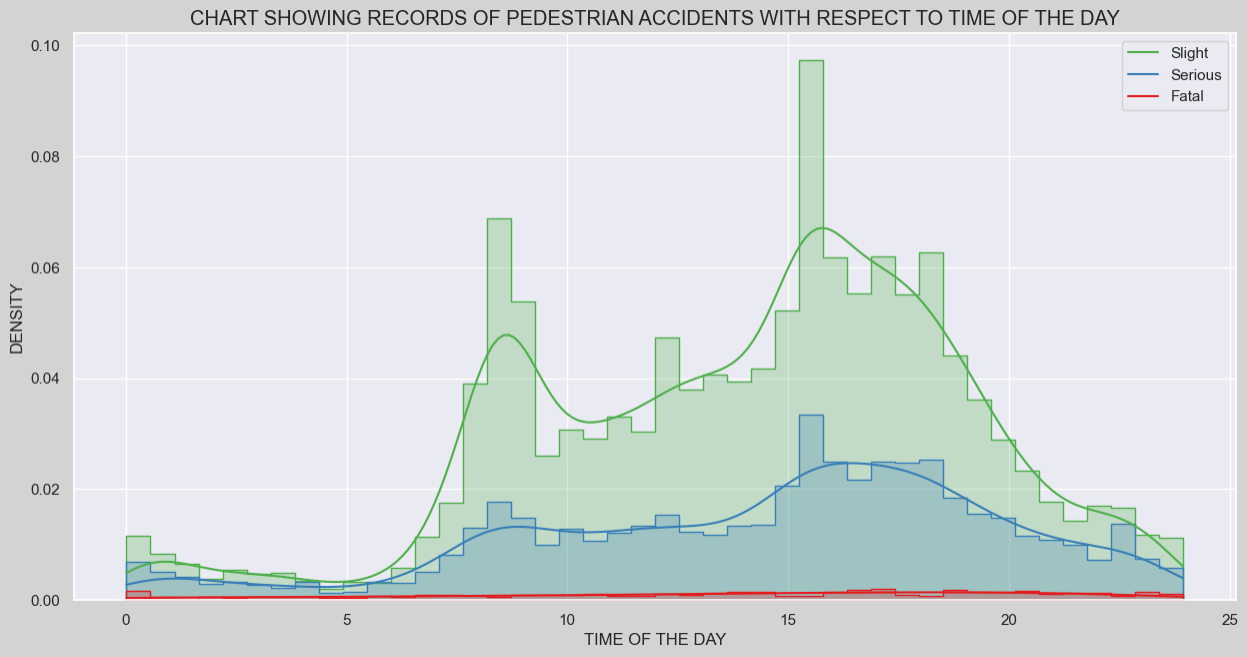


Figure 10 Pedestrian accidents time of the day plot.

Exploring Cycles

Most of the cycle accidents occur at 5:00pm and it appears to be similar to the general data trends with regards to time and day of the week.

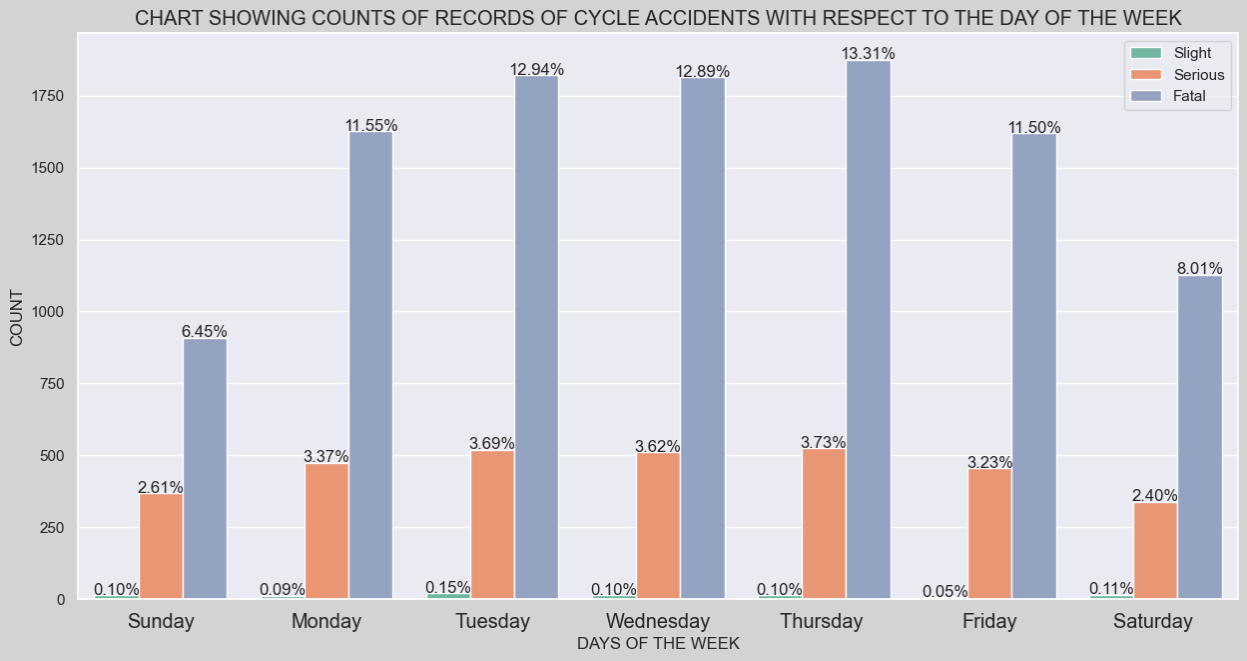


Figure 11 Cycle accidents day of the week plot.

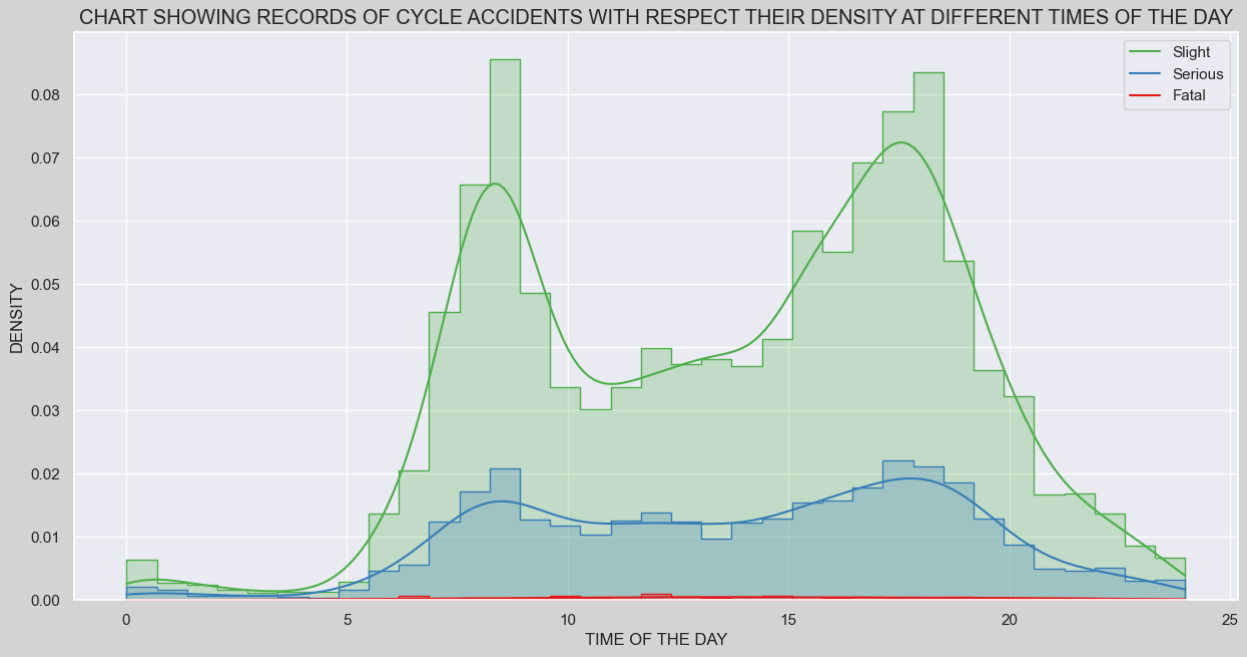


Figure 12 Cycle accidents time of the day plot.

Exploring Cars

There appears to be quite a small ratio of accidents for cars at the start of a business day compared to the end of a business day as there seem to be a high volume of car accidents in ratio between the hours of 4:00pm and 6:00pm and the highest number of car accidents recorded is at about 5:00pm.

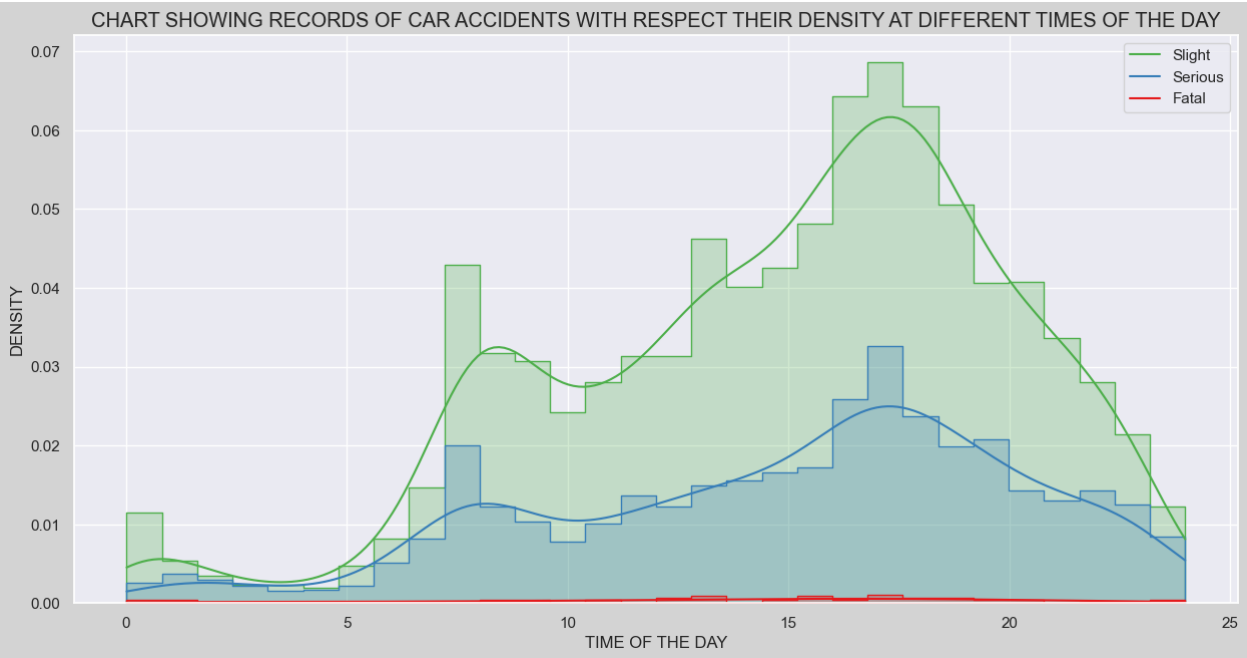


Figure 13 Car's accidents time of the day plot

While for day of the week it seems to follow the normal trend for the general day of the week with slight increase in fatalities on Friday and Wednesday compared to other days of the week.

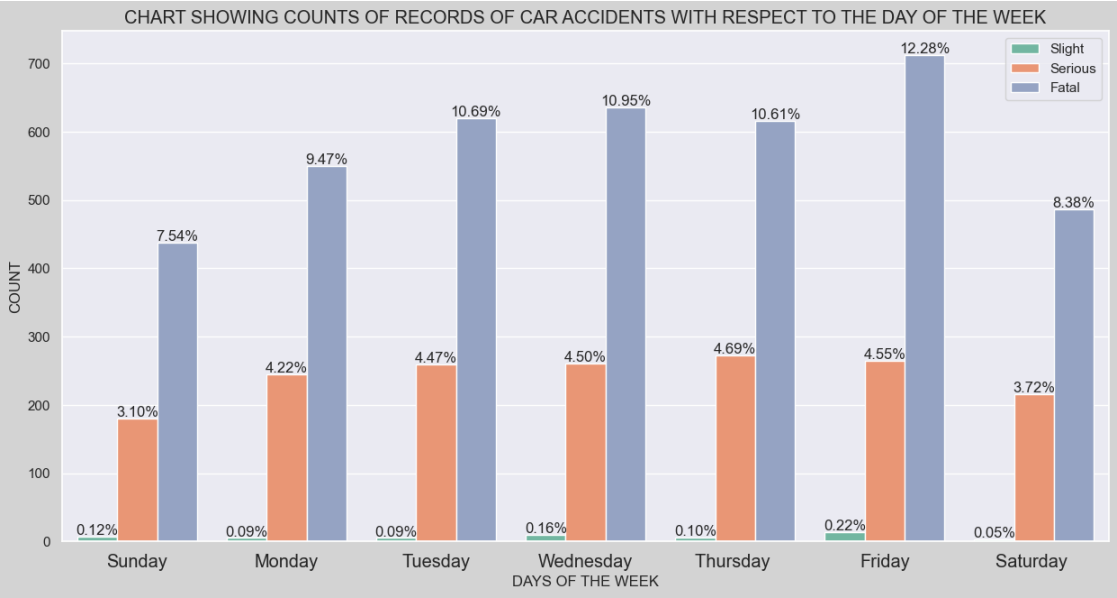


Figure 14 Car's accidents day of the week plot.

Exploring Day Light Saving Data

Day light saving in the United Kingdom for the year 2019 occurred between 1:00am 31st March, 2019 and 2:00am 27th November, 2019 (Time and Date, 2022). The start week for day light savings happen to be the 13th week indicated by a green bar and the end week the 43rd week indicated by an orange bar of the year.

In the trends according to visualisations, the two preceding weeks after week 13 seem to increase in the number of accidents recorded. Also, the week preceding week 43 also increases.

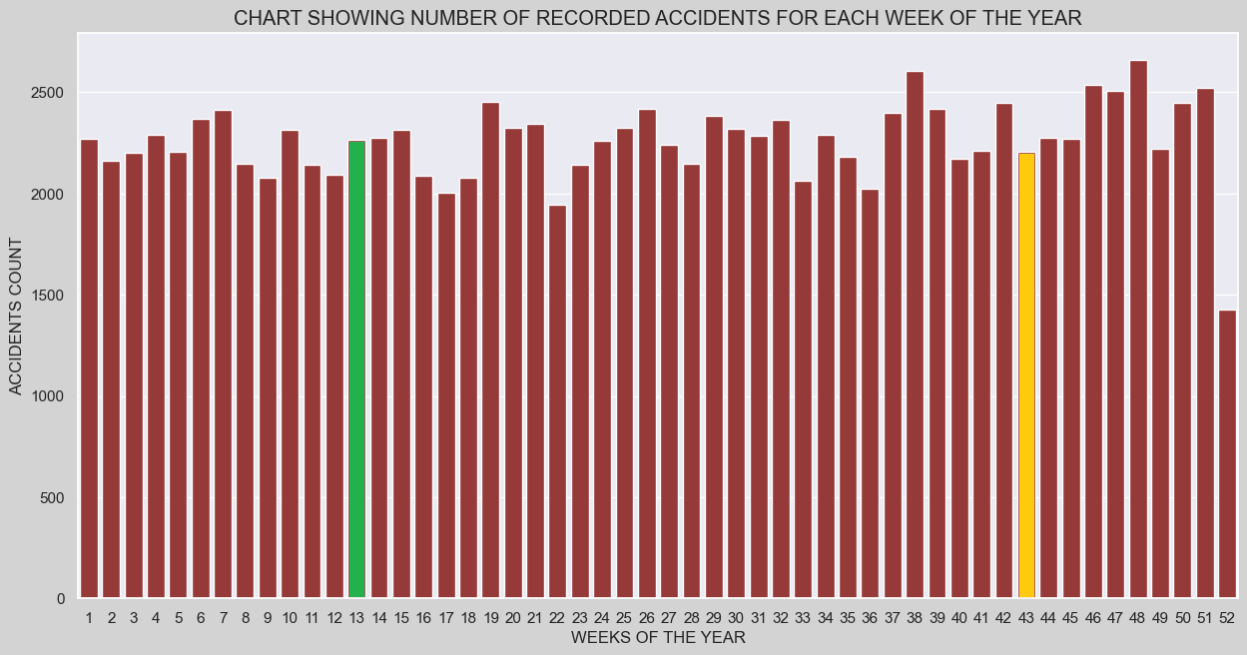


Figure 15 Plot showing number of accidents recorded for each week of the year

November appears to be the month with the highest number of accidents recorded in 2019 with 8.96 percent while February is the month with the least record of accidents with 7.75 percent of all accidents recorded in 2019. The month with the most percentage of slight accidents is November with 7.25 percent, while for serious and fatal injuries, it is July with 1.79 percent and 0.14 percent respectively.

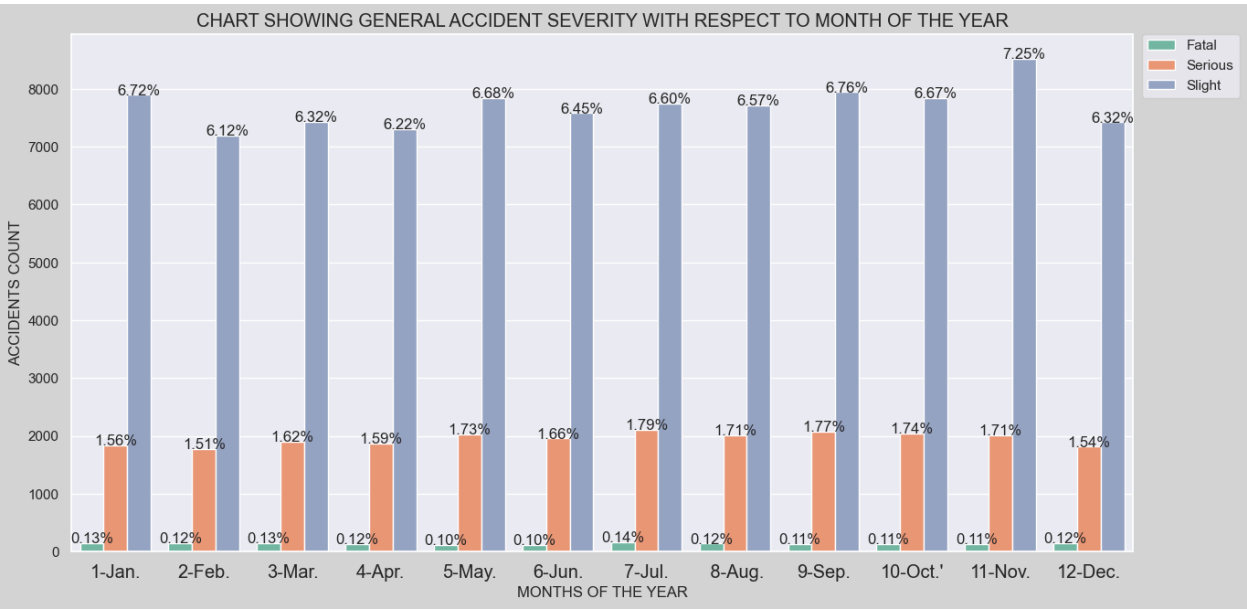


Figure 16 Plot showing accident severity with respect to months of the year

Another observation from analysing day light savings and standard time data is that the ratio of accidents that occur per day is the same. A hypothesis test was carried out to confirm if the accidents that occur during sunrise is similar to accidents that occur at sunset and a pvalue of 0.82 was achieved making a

hypothesis valid. Another hypothesis test was conducted to confirm if the accidents preceding the start week of day light savings and the end week of daylight savings are similar and a pvalue of 0.99 was achieved confirming the hypothesis.

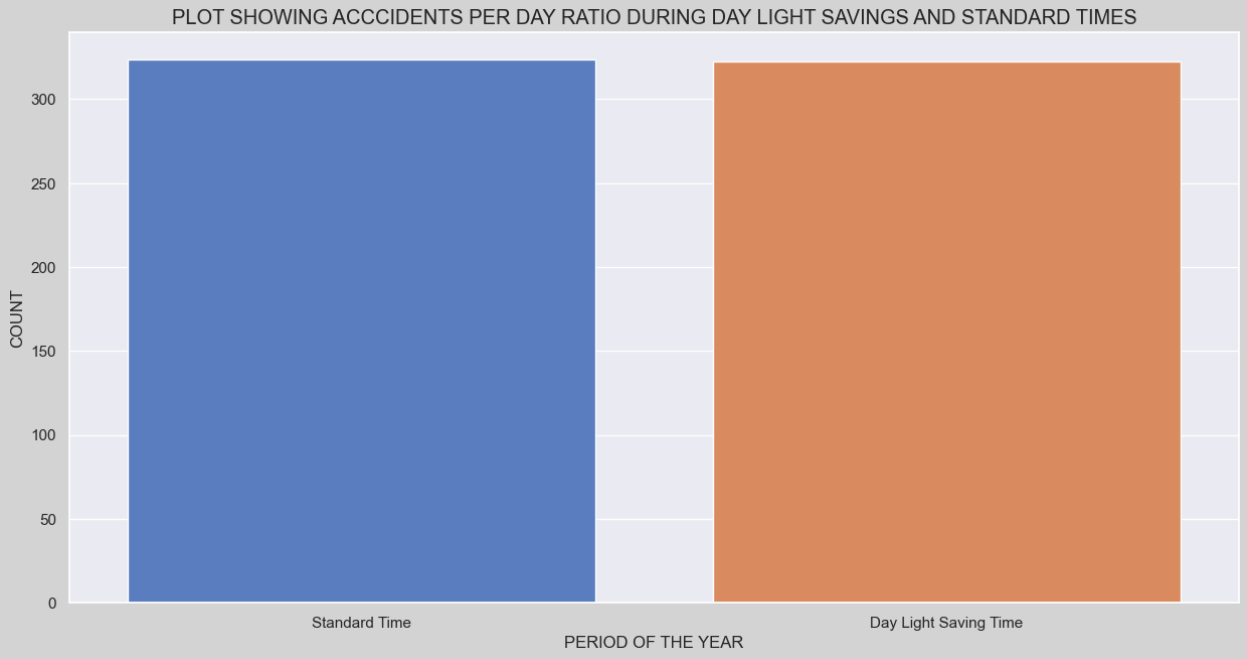


Figure 17 Plot showing number of accidents recorded per day between daylight savings time and standard time

Below is the hourly plot for accidents that occur during daylight savings and standard time

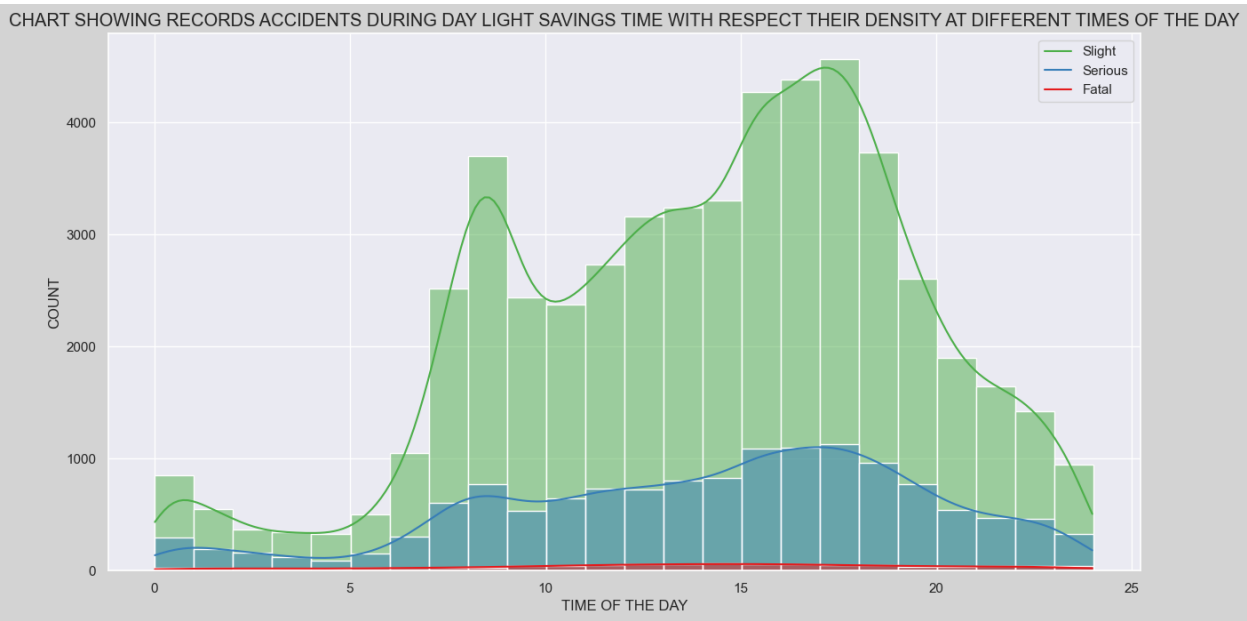


Figure 18 Daylight savings hourly accident plot

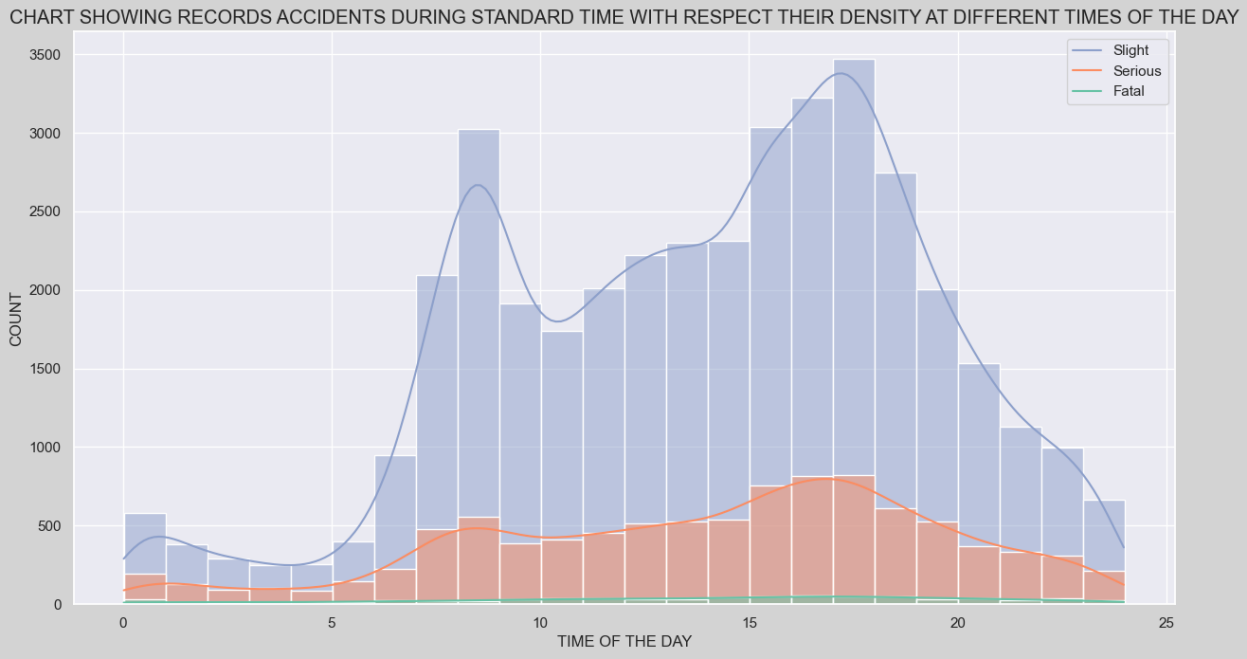


Figure 19 Standard time hourly accident plot

Conditions that may have Influenced Accident Severity

Vehicle Related Conditions

As shown in the charts below, it can be seen that the vehicles majority of accidents occur with two vehicles involved. Also, vehicles going ahead seem to have the greatest number of accidents and most fatal severity. Also as shown by the apriori test, it can be seen that;

i. slight accident severity is popular among skidding and overturning.
ii. when two vehicles are involved is more likely to involve skidding and over turning than any other number of vehicles involved.
iii. Cars are more likely to be involved in skidding and over turning than any other type of vehicle.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Skidding_and_Overturning_0)	(Accidents_severity_3)	0.811684	0.766276	0.631960	0.778579	1.016056	0.009986	1.055564
1	(Accidents_severity_3)	(Skidding_and_Overturning_0)	0.766276	0.811684	0.631960	0.824716	1.016056	0.009986	1.074349
2	(Accidents_severity_3)	(Vehicle_Type_9)	0.766276	0.673040	0.531584	0.693724	1.030732	0.015850	1.067533
3	(Vehicle_Type_9)	(Accidents_severity_3)	0.673040	0.766276	0.531584	0.789825	1.030732	0.015850	1.112045
4	(Skidding_and_Overturning_0)	(Number_of_Vehicles_2)	0.811684	0.619259	0.518271	0.638513	1.031093	0.015628	1.053264
5	(Number_of_Vehicles_2)	(Skidding_and_Overturning_0)	0.619259	0.811684	0.518271	0.836921	1.031093	0.015628	1.154756
6	(Skidding_and_Overturning_0)	(Vehicle_Type_9)	0.811684	0.673040	0.556292	0.685356	1.018298	0.009996	1.039141
7	(Vehicle_Type_9)	(Skidding_and_Overturning_0)	0.673040	0.811684	0.556292	0.826536	1.018298	0.009996	1.085623

Figure 20 Apriori test result for vehicle data.

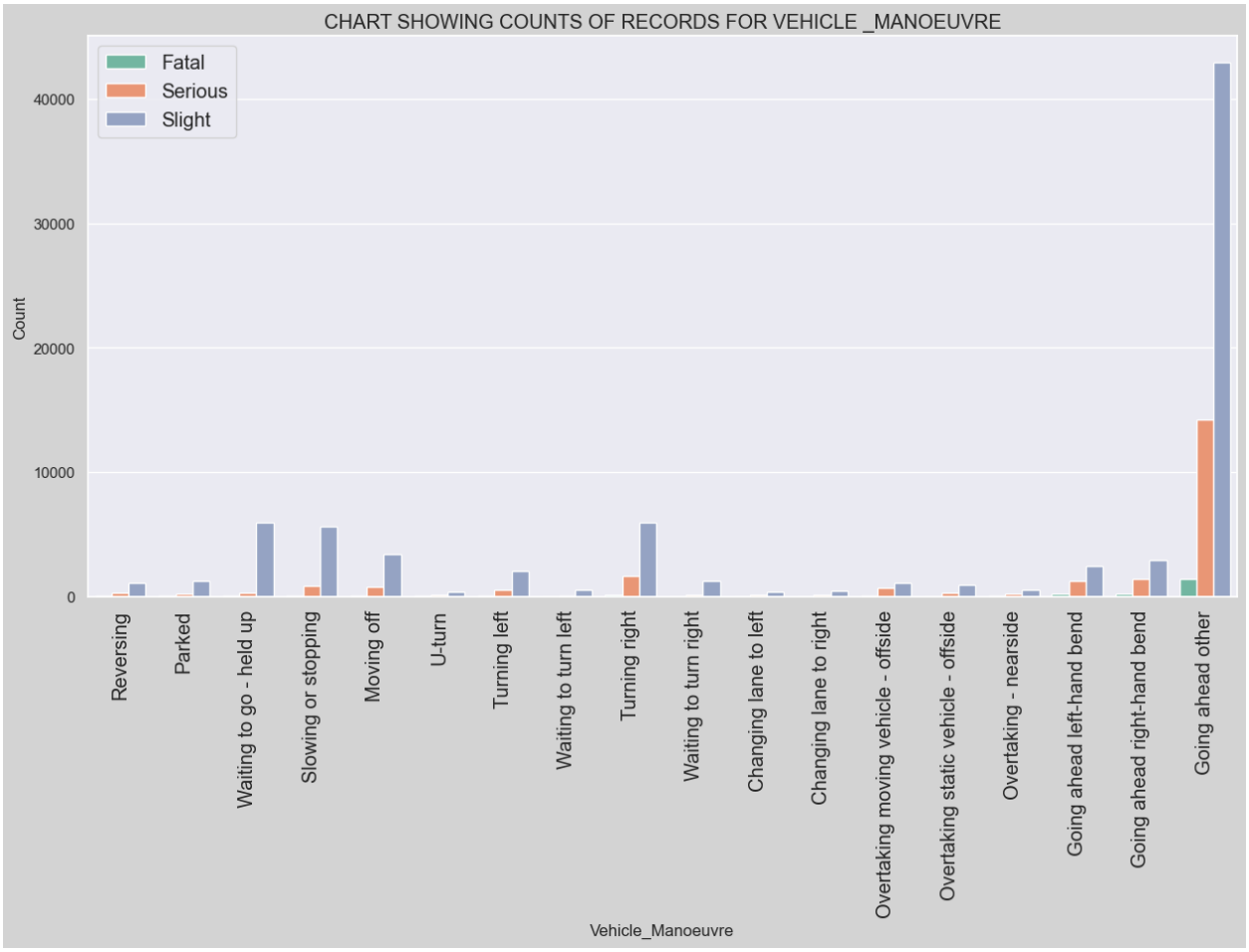


Figure 21 Chart showing explored data for vehicle manoeuvre.

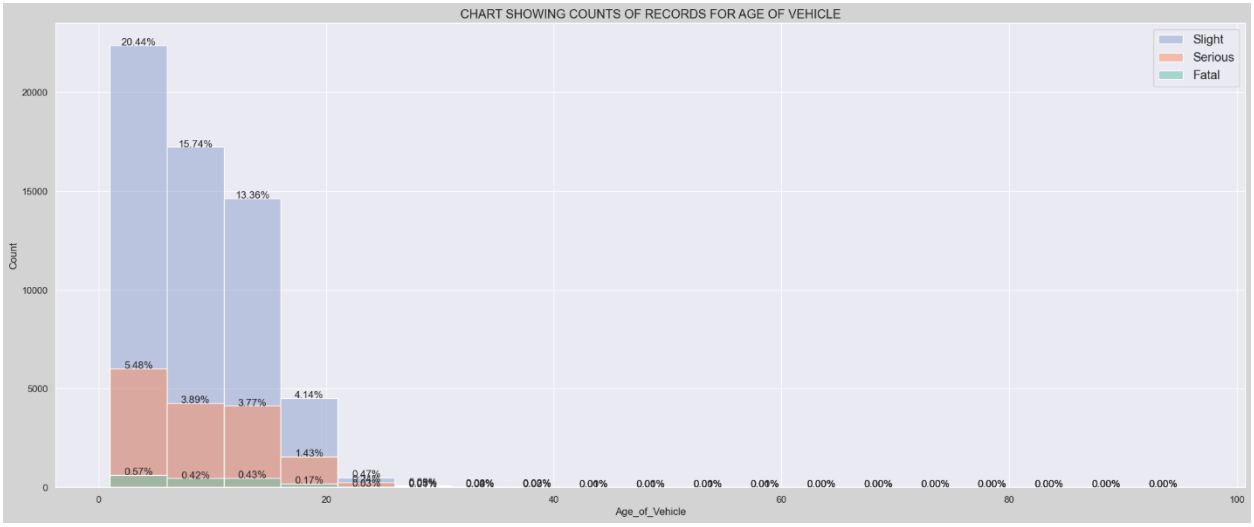


Figure 22 Plot showing age of vehicles

Junction Control

The junctions that are uncontrolled seem to have a large number of accidents as shown in the chart below.

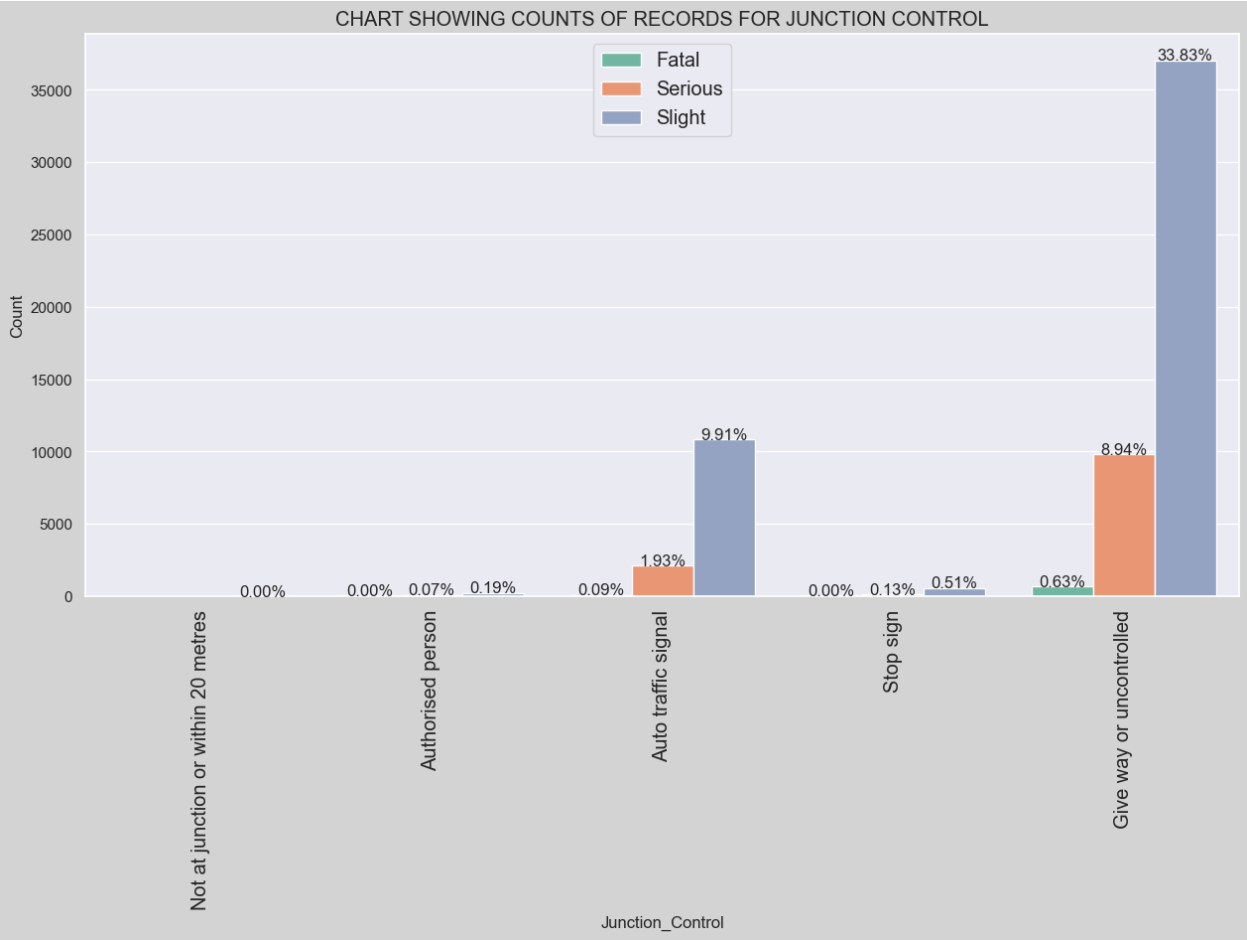


Figure 23 Chart showing junction control accident data

Weather Related Conditions

As seen in the charts below, most of the accidents seem to happen under daylight condition with fine weather.

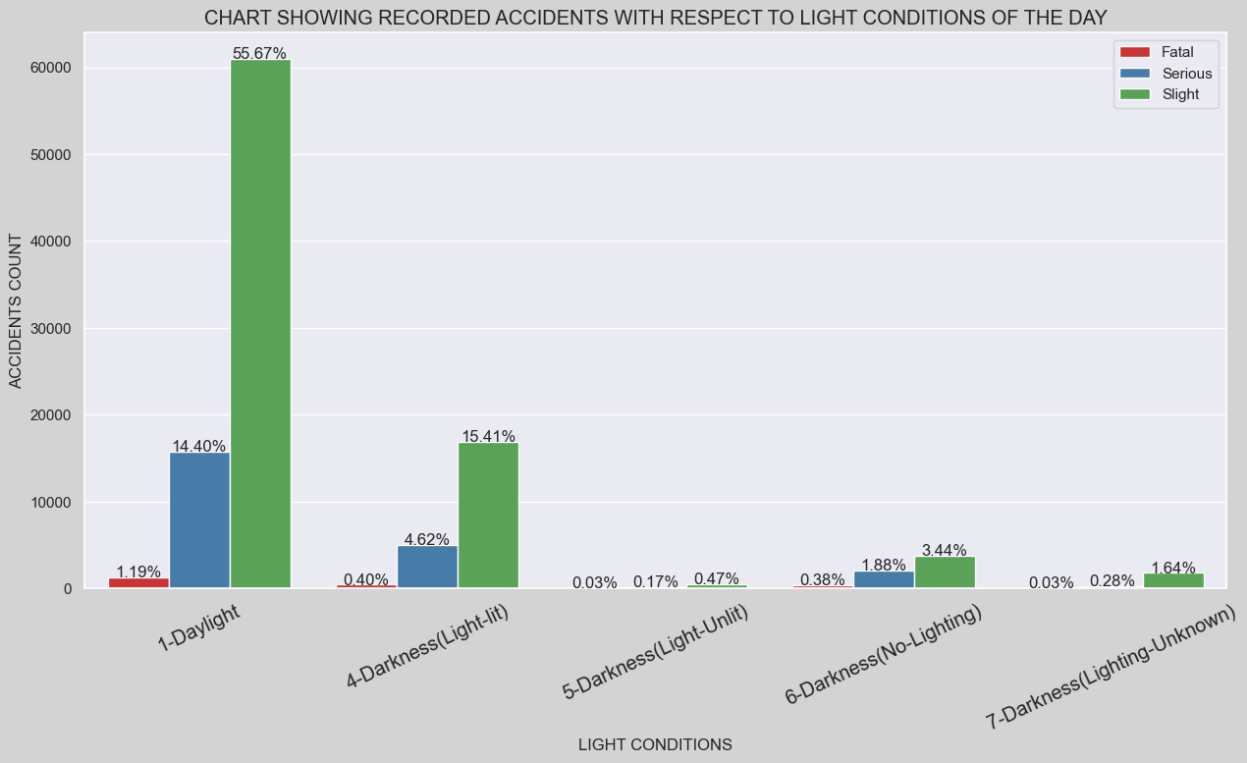


Figure 24 Plot showing different light conditions with respect to accident severity

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Accidents_severity_3)	(Light_Conditions_1)	0.766276	0.712531	0.556676	0.726469	1.019561	0.010680	1.050955
1	(Light_Conditions_1)	(Accidents_severity_3)	0.712531	0.766276	0.556676	0.781265	1.019561	0.010680	1.068526
2	(Accidents_severity_3)	(Weather_Conditions_1)	0.766276	0.796326	0.607170	0.792364	0.995025	-0.003036	0.980921
3	(Weather_Conditions_1)	(Accidents_severity_3)	0.796326	0.766276	0.607170	0.762464	0.995025	-0.003036	0.983952
4	(Light_Conditions_1)	(Weather_Conditions_1)	0.712531	0.796326	0.593236	0.832575	1.045521	0.025829	1.216511
5	(Weather_Conditions_1)	(Light_Conditions_1)	0.796326	0.712531	0.593236	0.744966	1.045521	0.025829	1.127179

Figure 25 Apriori test result for weather conditions

In the apriori test above it is important to note these facts:

i. Slight accidents occur more during daylight
ii. Slight accidents occur more under fine weather

Geography Related Conditions

The geographic location with the metropolitan police has a huge chuck of the accidents recorded in 22.88 percent of all recorded accidents in the United Kingdom as shown in the chart below.

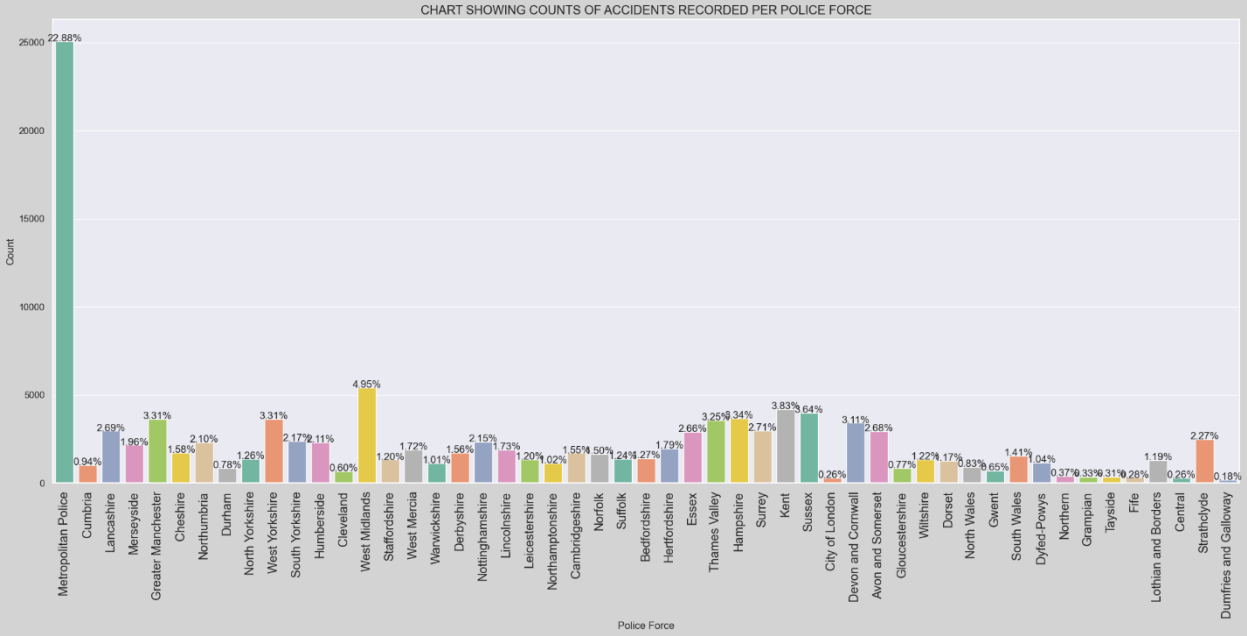


Figure 26 Plot showing number of accidents recorded per police force

Below is the elbow method graph using to predict the centroids and a scatterplot showing the accident concentrations and centroid points that can be used as points to site major emergency services for accident victims.

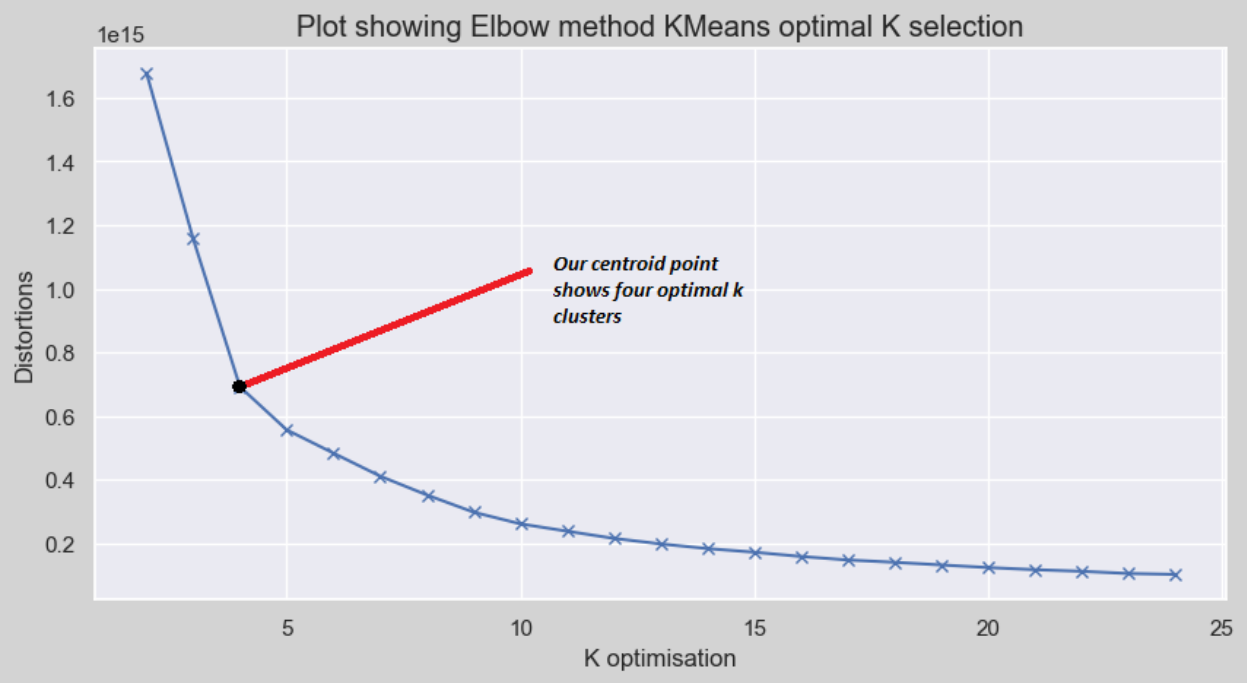


Figure 27 Elbow curve used to determine centroid of clusters

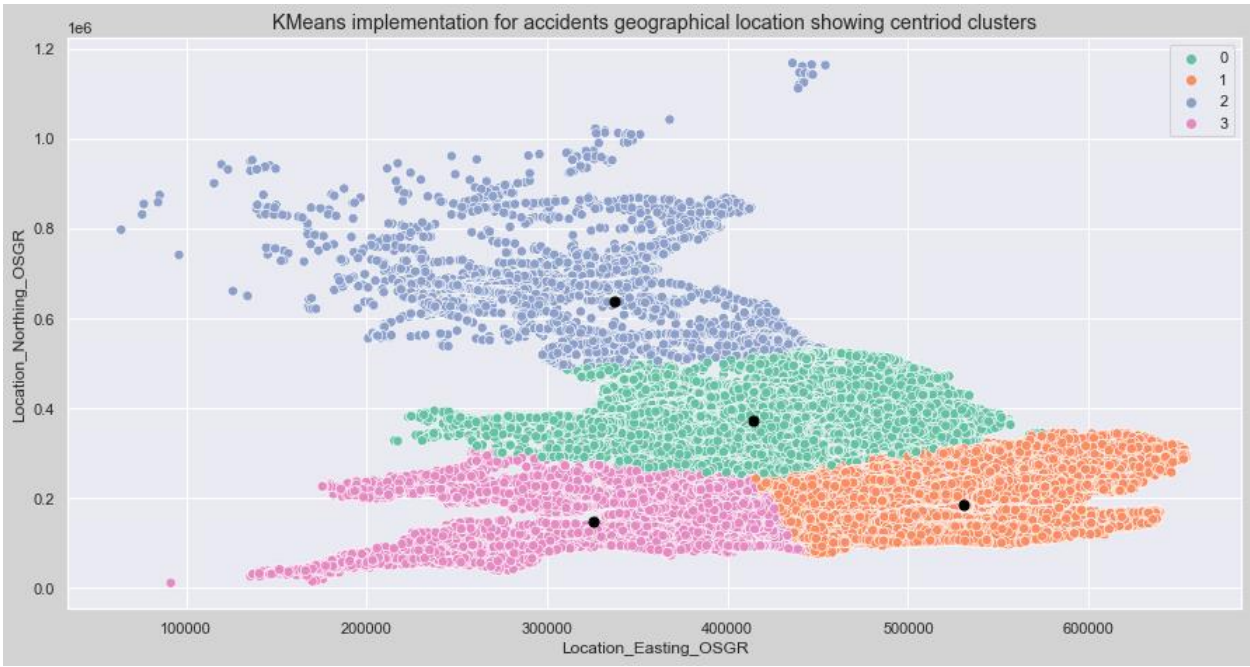


Figure 28 Scatter plot with centroid point of the clusters indicated with black points

Building an Artificial Intelligent Model to Predict Accident Severity

Model Methodology

To build my model I had to follow the following steps:

1. Split my training and test data from the main dataframe
2. Scale the data using standard scaler
3. Carry out Hyper-Parameter optimisation and feature selection using random forest classifier
4. Use PCA to reduce the dimensionality of the selected features
5. Train selected models with the best features and data already prepared

Model Summary and Test Result

The chart below show's the result after training my models.

	Accuracy
Logistic Regression	0.740000
K Nearest Neighbours	0.890000
Decision Tree	0.850000
Random Forest	0.950000
AdaBoost	0.530000
Bagging Classifier	0.950000
XGBoost	0.950000
GaussianNB	0.695397
Stacking Models	0.931217

Figure 29 Accuracy score produced by each classifier used to train our model

As seen above, the random forest classifier out performs a lot of its peers hence I had to select it as my preferred model.

Upon selection of the random forest model, I had to put in the hyperparameters earlier generated and there was a slight improvement in the performance of the model as shown below.

```
              precision    recall  f1-score   support

           1       1.00      1.00      1.00      7865
           2       0.88      0.89      0.89      7782
           3       0.89      0.88      0.88      7740

    accuracy                           0.92     23387
   macro avg       0.92      0.92      0.92     23387
weighted avg       0.92      0.92      0.92     23387

[[7865    0    0]
 [   0 6912  870]
 [   1  910 6829]]

The Training accuracy score is : 0.97
The Validation accuracy score is : 0.92
The number of correctly predicted points are : 21606
```

Figure 30 Random Forest improved model performance after implementing hyperparameter optimisation

When I compared my model prediction to the government model for the slight and serious labels prediction, it gave a 50 percent probability.

Conclusion and Recommendation

i. More intelligent traffic lights should be installed at junctions as this would have significant impact on the number and severity of accidents recorded (Qi et al., 2015).
ii. The speed limit for Motorcycles with over 500CC should be reduced and the use of protective gears strongly enforced as risk of fatal crash is higher among riders with powerful motorcycles (Mattsson & Summala, 2010).
iii. Movement of vehicles over 10 years should be strictly regulated as they tend to lose efficiency because of wear and tear and by this, the use of such vehicles should be discouraged.
iv. High pedestrian accidents might be as a result of primary and secondary school students commuting home (STANLEY PRIMARY SCHOOL, 2022). Therefore investing in teaching them more about road signs and measures required to commute safely as pedestrians will be help reduce the numbers.

References

Mattsson, M. & Summala, H. (2010) With power comes responsibility: Motorcycle engine power and power-to-weight ratio in relation to accident risk. *Traffic Injury Prevention,* 11 (1), 87-95.

Qi, L., Zhou, M. & Luan, W. (2015) Emergency traffic-light control system design for intersections subject to accidents. *IEEE Transactions on Intelligent Transportation Systems,* 17 (1), 170-183.

Reid, J., Crone, J. & Hayes, J. (2017) Geographic boundary data and their online access. *The Routledge Handbook of Census Resources, Methods and Applications: Unlocking the UK 2011 Census,* 90.

STANLEY PRIMARY SCHOOL (2022) *School-hours.* Available online: https://www.stanley.richmond.sch.uk/school-hours/ [Accessed 20/04/ 2022].

Time and Date (2022) *Time change 2019 in the united kingdom.* [Accessed 20/04/ 2022].