Introduction

According to the Road Casualties, Great Britain: 2019 Annual Report[1], over 500,000 injury collisions have been estimated between 2015-2019 in the UK[2]. The data on these collisions have been recorded and published annually by the Department for Transport since 1975, allowing for use for further investigations and analysis of Road safety and accidents in the UK.[3]

The report provides interesting insights into the trends in accidents and fatality rates. The report also breaks down the data by looking at road user types e.g., Car drivers, cyclists, pedestrians, etc.[4] However, there is little information presented in the report regarding regional trends or patterns over the course of the year or day. There is a need for further investigation in these areas, as it could aid in guiding future road and infrastructure developments, thereby enhancing road safety in the UK.

In this report, I will analyse the following datasets (Road Safety Data - Accidents 2019.csv
Road Safety Data - Casualties 2019.csv, Road Safety Data- Vehicles 2019.csv, and accompanying documents)[5] to address when, where and under what conditions accidents happen.

Analysis

Data loading and processing:

Before proceeding into deeper analysis, I first load the data files (Accidents, Vehicles, and casualties). I later uploaded the variable lookup excel file which contains a large number of variables associated with each dataset. I follow this with the evaluation of any required pre-processing, cleaning, or transformation.[6]

Data Cleaning: There are a lot of missing entries in the datasets – particularly in the Accidents dataset (features: Location_Eastin g_OSGR, Location_Northing_OSGR, Longitude, Latitude, Time, and LSOA_of_Accident_Location). There were a total of 5, 889 mi ssing entries on the Accidents dataset. The Vehicles and Casualties datasets did not present any null or missing entries. Due to th e number of missing entries, I decided to forward fill in the values to ensure the consistency of the data. I did not remove values such as “Unknow” from any of the datasets and I didn’t deem it necessary to remove these.

Data preparation and Pre-processing: A lot of the features are coded in numerical format, so I used the variable lookup mapping sheets provided to convert them to their textual strings’ equivalents.

Feature Transformation: I created more features (such as decimal time, Hour, Month, and Day of the week) by using the datetime module. I created new features as needed. Furthermore, I also linked some features from the Vehicles dataset to the main Accidents dataset. All of this will be useful in building a predictive model later on.

When:

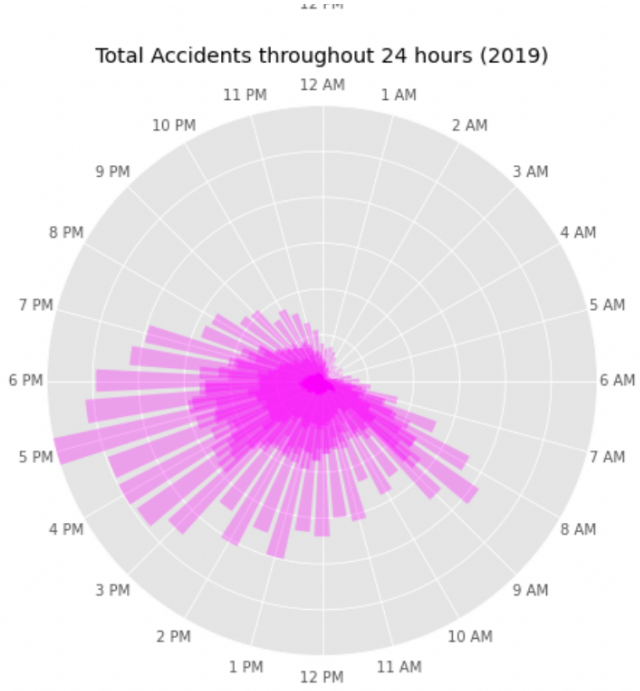
In this section I looked at when accidents happen:

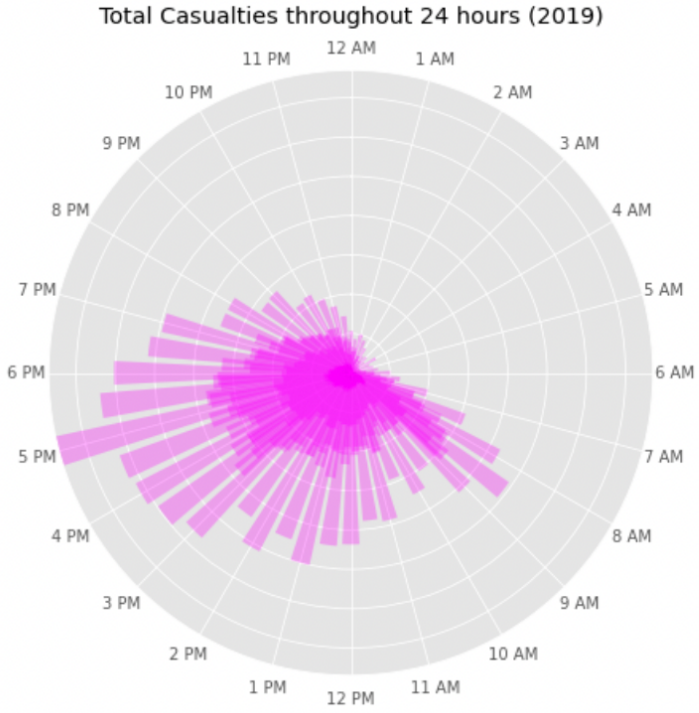
Are there significant hours of the day, and days of the week, on which accidents occur?
To get more details about when accidents happen, I analysed the hours, days of the week and even months, to discover when accidents happen the most. By grouping time and number of accidents, then getting the radian of this group. The 24-hour graphs were plotted with polar projection.[7]

a) **Hours of the Day:**

---

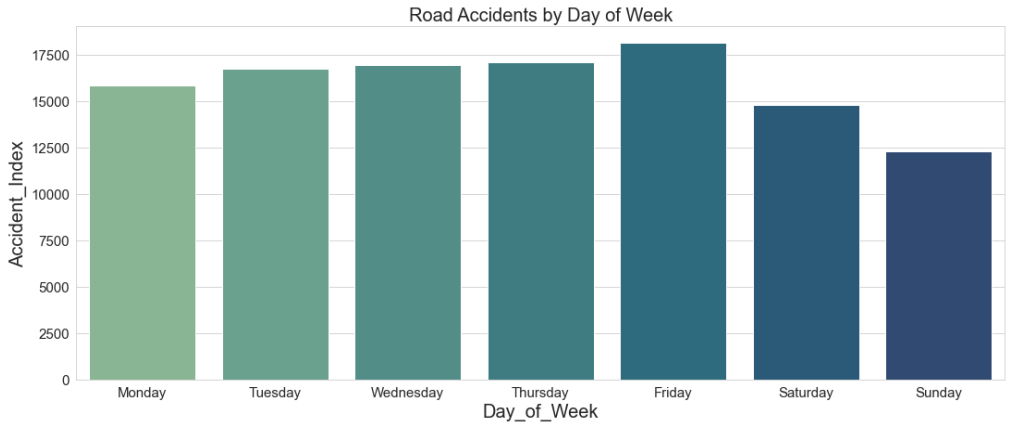
[1] *Reported Road Casualties Great Britain, Annual Report: 2019*
[2] *Reported Road Casualties Great Britain, Annual Report: 2019*
[3] Staines, T. (2018) *Car Crashes and the Weather: An Exploratory Analysis of Environmental Conditions’ Impact on Traffic….* Medium. Available online: https://towardsdatascience.com/car-crashes-and-the-weather-an-exploratory-analysis-of-environmental-conditions-impact-on-traffic-12bcb7f9afed [Accessed 17/5/2022].
[4] *Reported Road Casualties Great Britain, Annual Report: 2019*
[5] *Road Safety Data - data.gov.uk. (2022) Data.gov.uk. Available online: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data [Accessed 17/5/2022].*
[6] Almohimeed, R. (2019) *U.K. Traffic Accidents — Data Analysis (10+years)*. Medium. Available online: https://medium.com/@rawanme/u-k-traffic-accidents-data-analysis-10-years-c81293180ee5 [Accessed 17/5/2022].
[7] *Advanced EDA of UK's Road Safety Data using Python* (-02-19T03:00:39+00:00) Available online: https://omdena.com/blog/advanced-eda/ [Accessed Aug 10,2022].

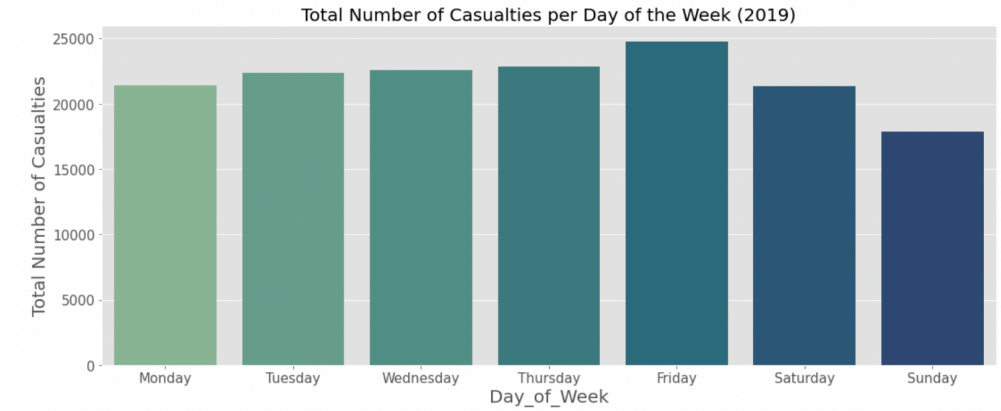




In the graphs above, there are a lot of accidents during morning peak period: between 8am and 9 am (with around 8.30am being the hour of highest accident occurrences, which is when most accidents occur. There are also afternoon-evening rush hours, in which a high number of accidents occur starting at around 15:00 to 19:00, with the highest peak in accidents at 17:00pm. These are the rush hours which means that the accident could be a result of people going and returning from their daily activities, such as work, and school runs. It is crucial that safety measures are enhanced during these periods. Consequently, the number of casualties, is just as high as it follows the number of accidents within the 24-hour period.

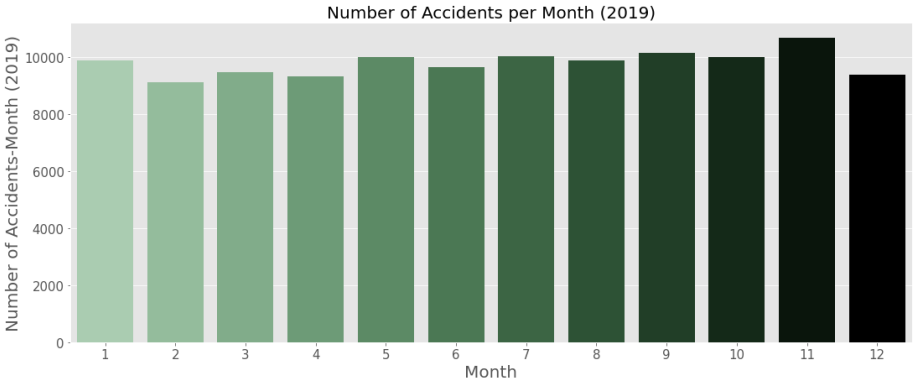
b) Accidents Occur by Days of Week:



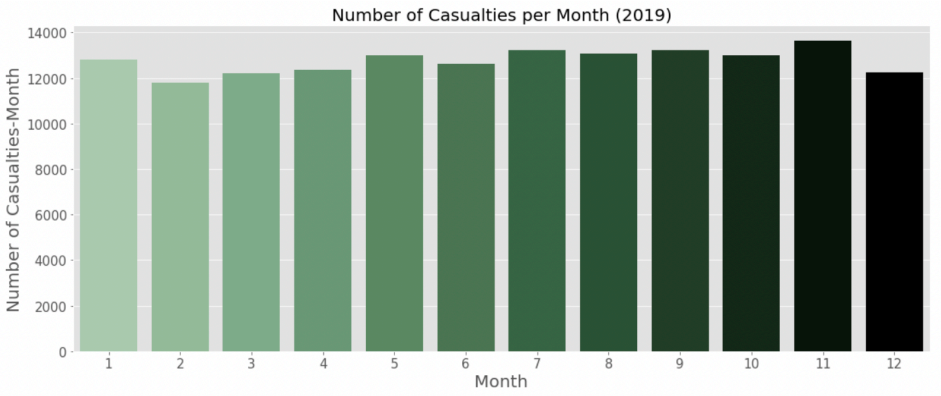


It can be seen that the day of the week on which the most accidents happen is Friday. This could be because it is the end of the workweek and the beginning of the weekend, therefore, there may be more people on the road. The day with the least number of accidents is Sunday. This may be because Sunday is viewed as a rest day, so there isn't much need to be on the road. Checking the total number of casualties per accident per day of the week, Friday also has the highest number total of casualties per accident.

c) Accidents Occur by Months:



It is noticed that month 11 (November) is the month in which the highest number of accidents occur. December (though being a month of festivities) has lesser accidents - this could be due to the weather conditions, making road users take better precautions when on the road. Also, a lesser number of accidents seem to occur in Springtime (February, March, and April), - the month with the least number of accidents in February.

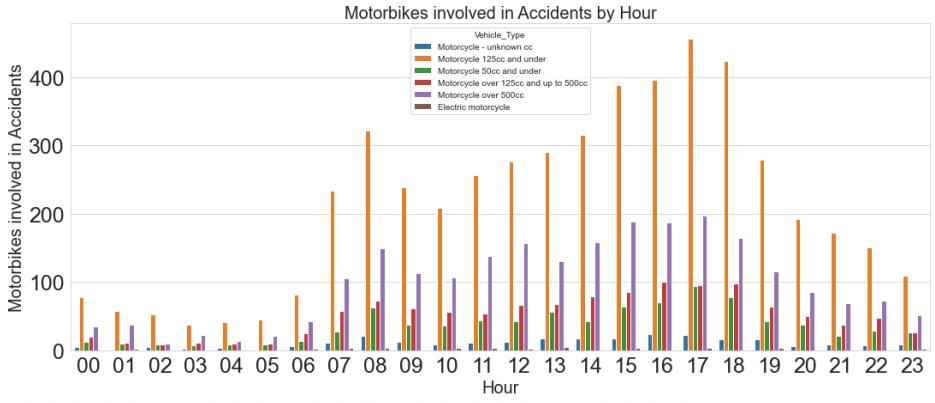


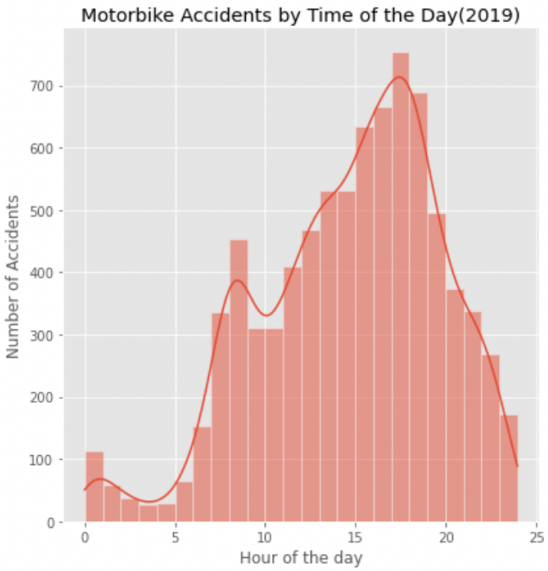
The month of November is also the month in which there are more casualties per accident compared to the other months. Safety measures during these periods may need to be enhanced.

Motorbikes:

To analyse the hours and days of the week in which motorbike accidents occur, I created a dataset with all motorbike types from the Vehicle dataset. I then grouped this with the "hour" feature and afterwards with the "Day of the Week" feature of the Vehicle dataset.

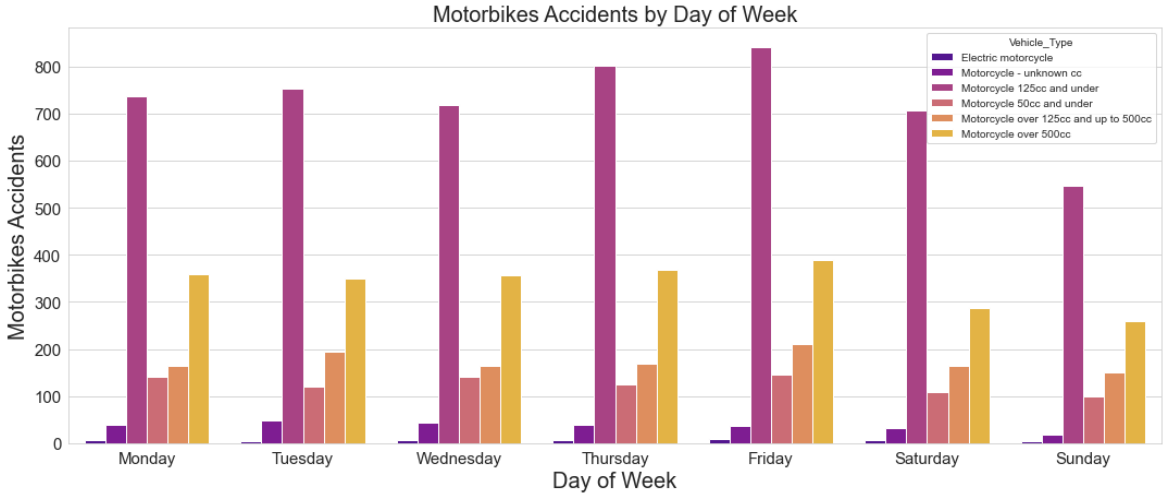
a) Hours of the day:





Most motorbikes accidents occur at around 17:00 in the afternoon. This agrees with the conclusion reached above that most accidents occur around the same hour.
Motorcycles 125cc and under are the type of motorbikes mostly involved in the accidents.

b) Day of Week:

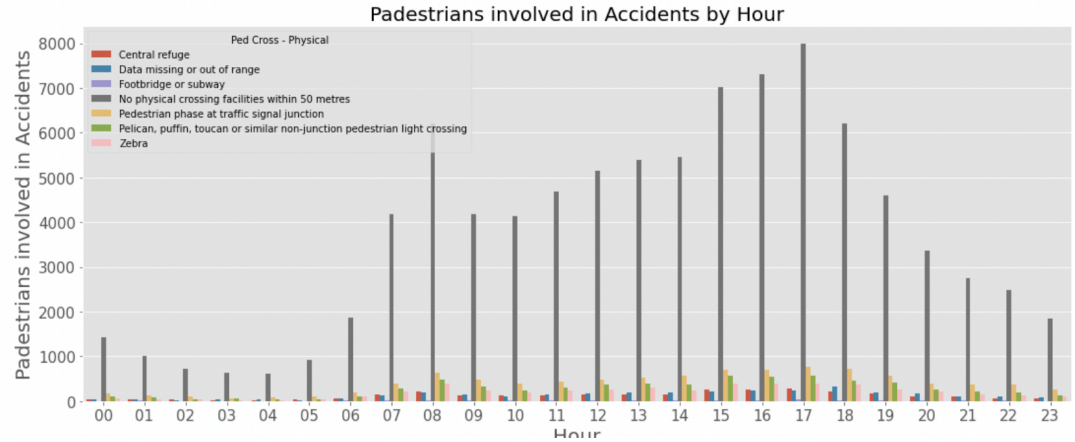


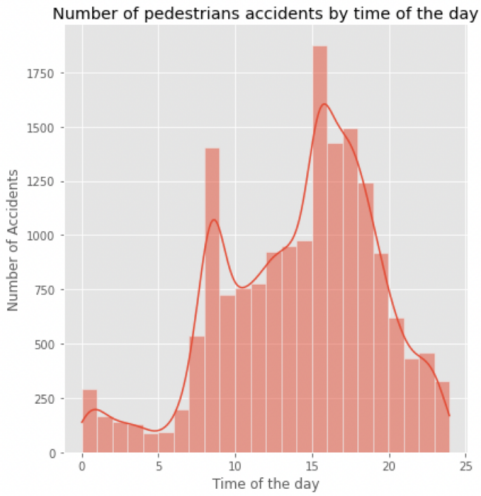
It looks from this graph that the day on which most motorbike accidents occur is Fridays.
Given that the type of motorbikes involved the most in these accidents are Motorcycles 125cc, an in-depth review of the motorcycle may be required to determine if the accidents are caused by a fault of the driver of the motorcycle or as a fault of the vehicle.

Pedestrians:

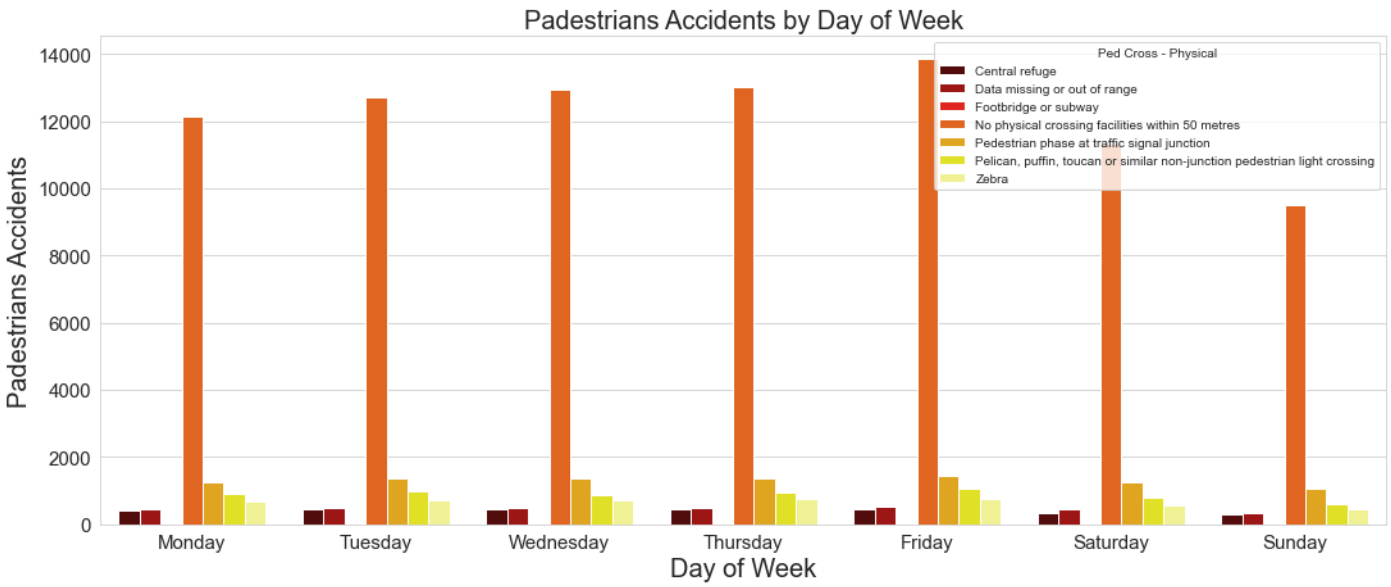
I also looked at the pedestrians involved in the accidents. Following the current theme, I analysed the hours of the day and days of the week in which pedestrians are likely to be involved in accidents

a) Hour:





As in the above graphs, the hour in which pedestrians are mostly involved in accidents is around 17:00. There are also a high number of pedestrian accidents at hours: 8:00, 15:00, 16:00 and 18:00, - so during the morning rush hour and afternoon-evening rush hours.

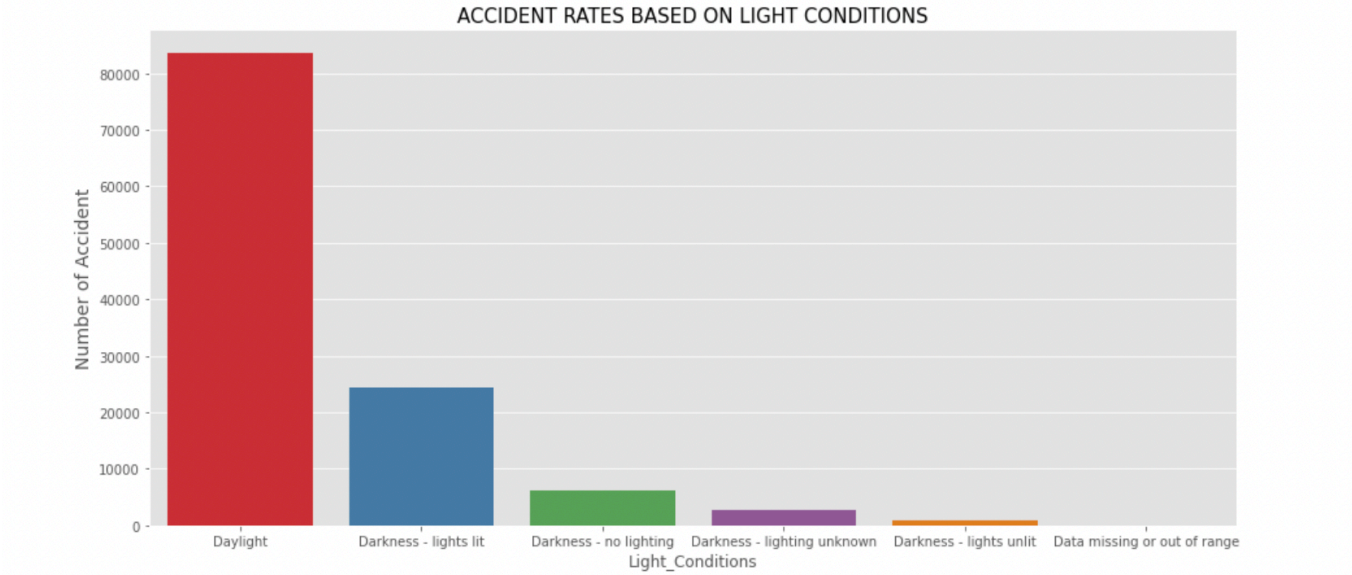


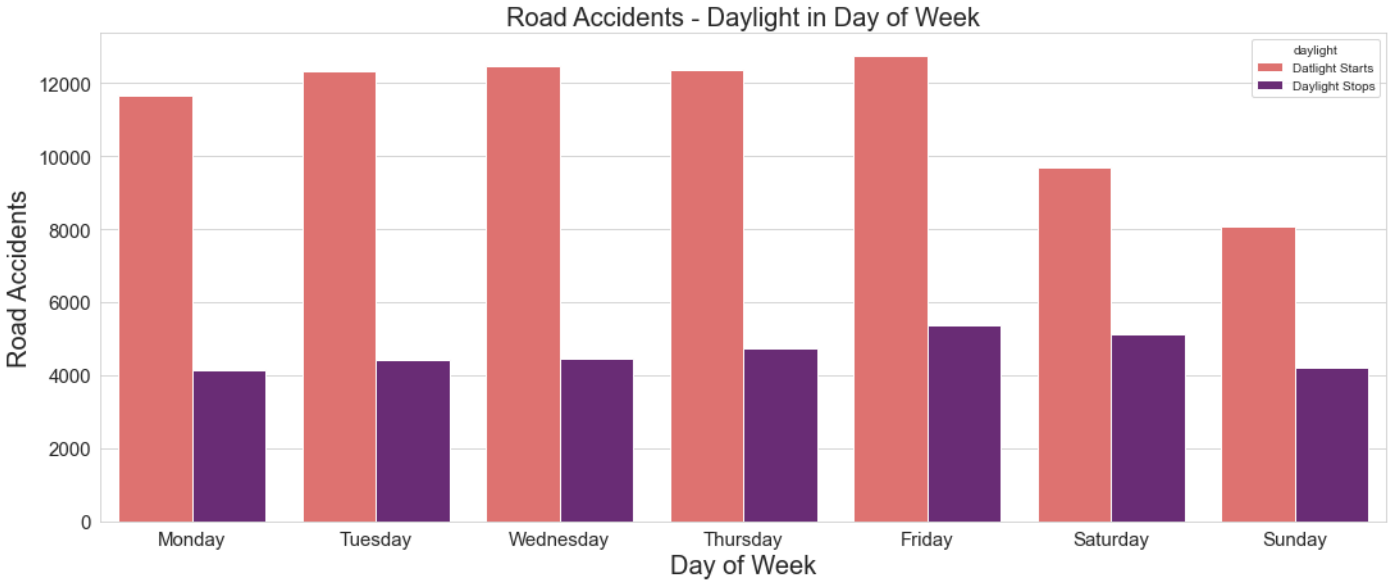
b) Day of the Week:

The day with the highest pedestrian accidents is Fridays. From this last graph, it can be noted that the areas with the most pedestrian accident occurrence are areas in which there are No physical crossing facilities within 50 metres. This means that pedestrians were accessing or crossing the roads which they shouldn't have due to a lack of crossing facilities. There is a possibility for extensive security measures to avoid pedestrians from accessing these areas or this could be seen as a need for the creation of crossing facilities for the safety of road users.

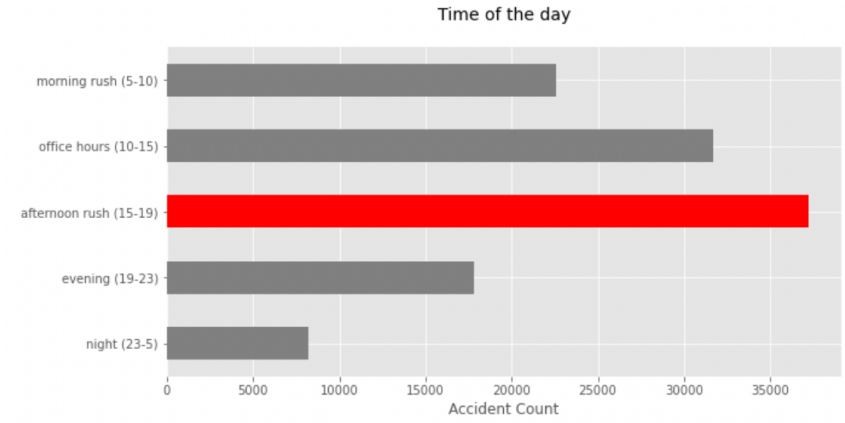
Accidents – Daylight:

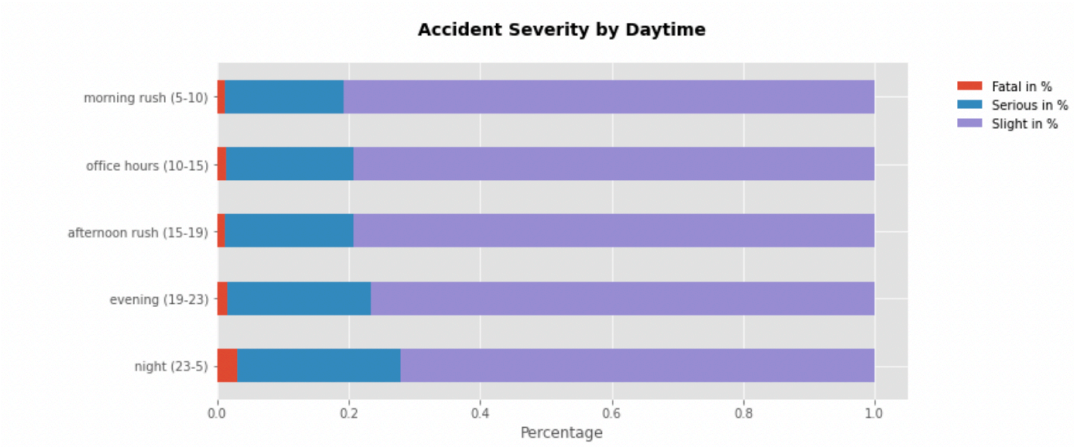
- Looking at the below graph, the greatest number of accidents occur during daytime or when there is daylight, with the highest number of accidents occurring on Fridays during the Daylight-saving season (when it starts and when it ends). This is closely followed by more accidents Wednesday and Thursday when Daylight savings start.





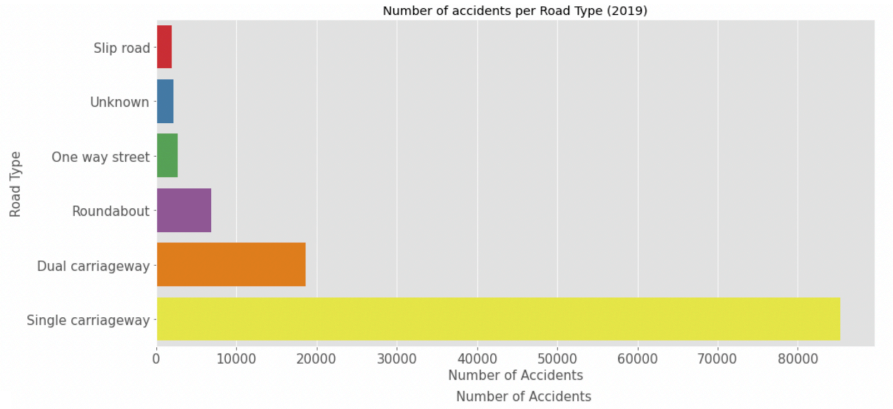
The following graph shows more accidents during the sunset hours during the daylight saving period.
Most accident happen during the afternoon rush hour but most fatal accidents happen during the night time between 23:00 and 5:00

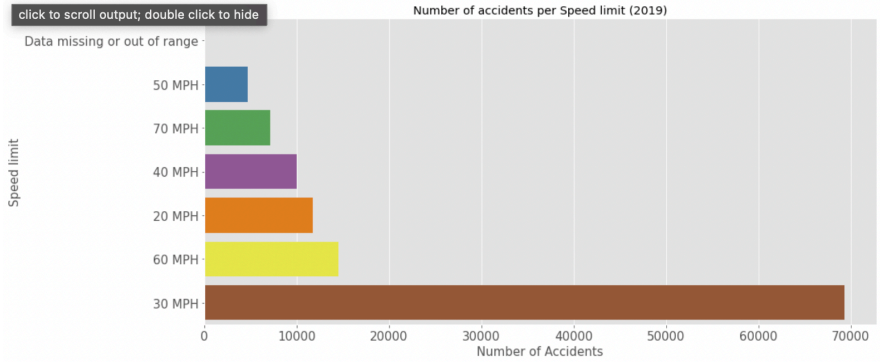


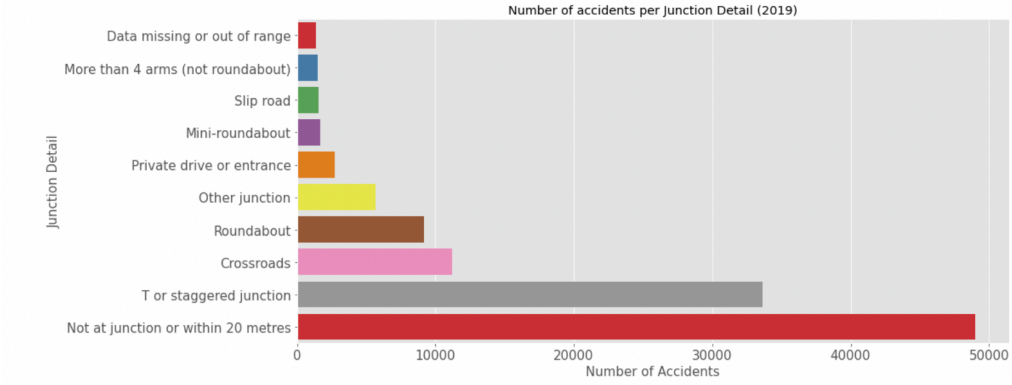


Where:

1. The graphs below show that most accidents happen in urban areas, on single carriageway “A” roads. A-roads are major roads connecting cities and regional towns; they are known as ‘trunk’ roads or principal roads. Most accidents seem to happen in areas there isn’t a junction or are within 20 metres of a junction.[8] Roads with (a 30-speed limit) account for 60%of accidents,[9] This is understandable since most accidents are categorized as (Slight) so it is more likely to happen on roads with a low-speed limit. These graphs indicate the absence of separation between traffic moving in different direction makes it a critical situation from a safety perspective, and therefore must be addressed.

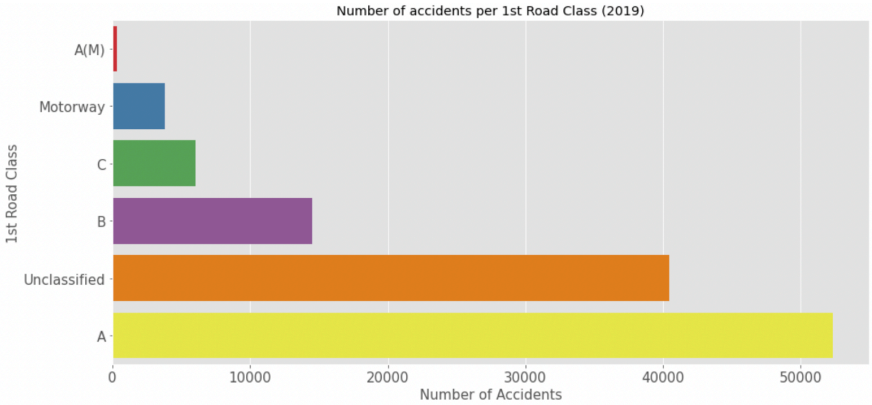


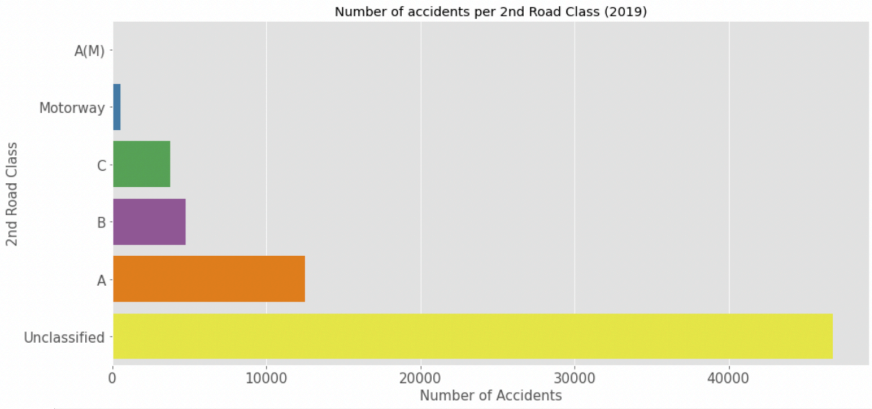


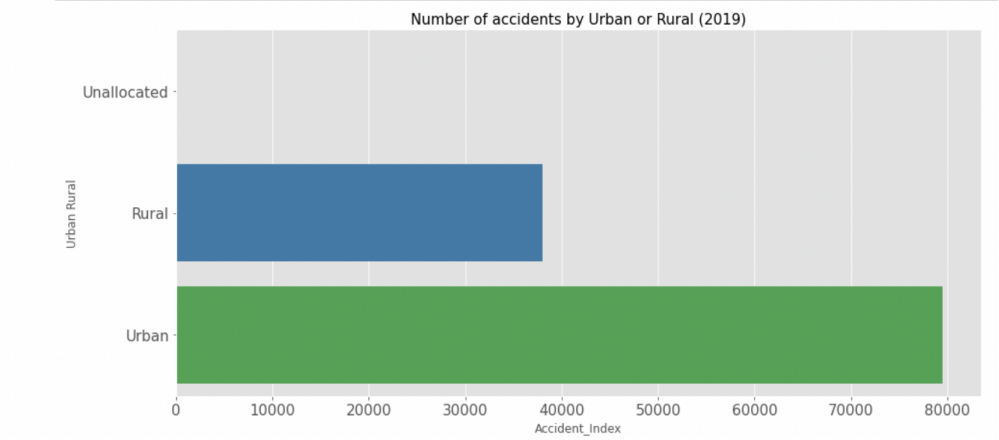


[8] *Great Britain Road Numbering Scheme*, 2022b). https://en.wikipedia.org/w/index.php?title=Great_Britain_road_numbering_scheme&oldid=1087647981. [Accessed 10/8/2022].

[9] Almohimeed, R. (2019) *U.K. Traffic Accidents — Data Analysis (10+years)*. Medium. Available online: https://medium.com/@rawanme/u-k-traffic-accidents-data-analysis-10-years-c81293180ee5 [Accessed 17/5/2022].



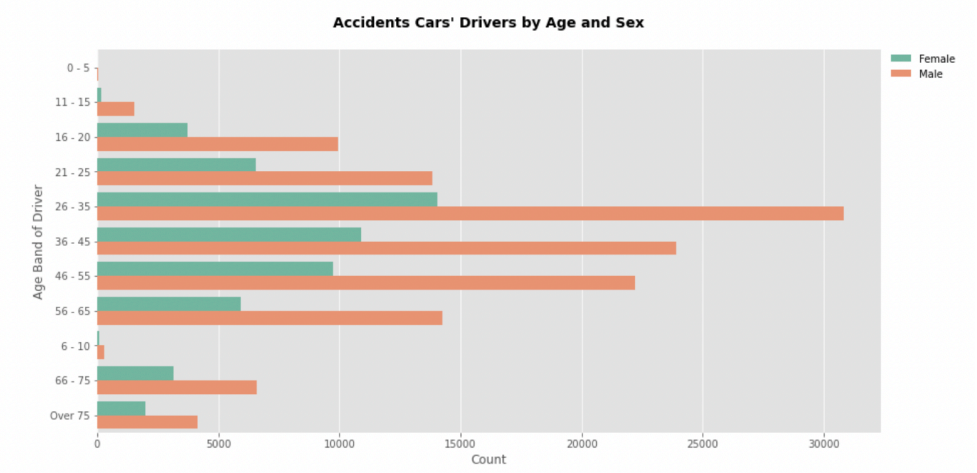




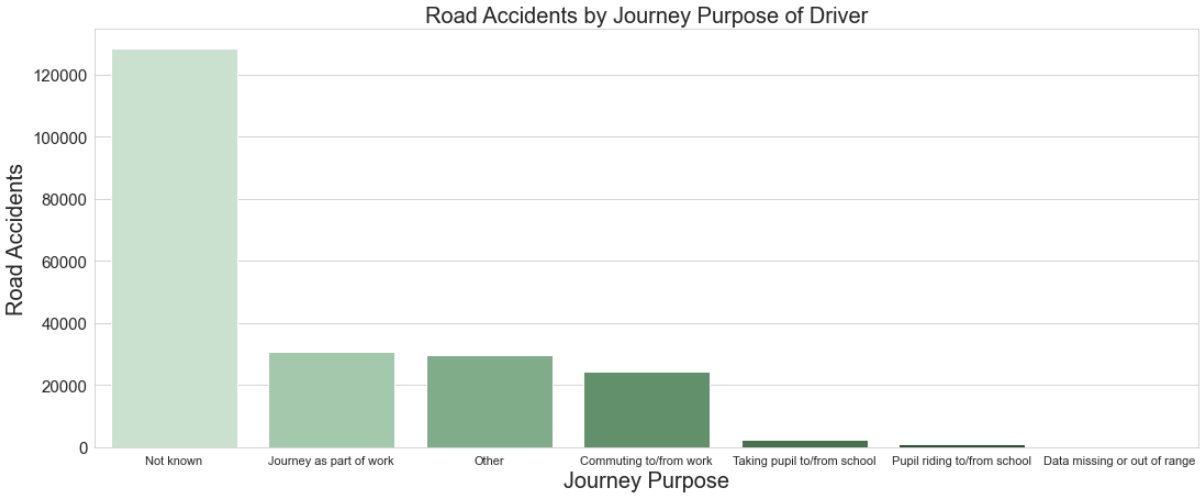
Under what conditions do accidents happen:

Age of Driver:

It is important to understand the target audience of road safety campaigns should need be. The highest 'Age of Driver' for most accident seems to be around age 26-35, this means that drivers that may cause the most accidents are around age 30. This is followed by drivers in the age range: 36-45.



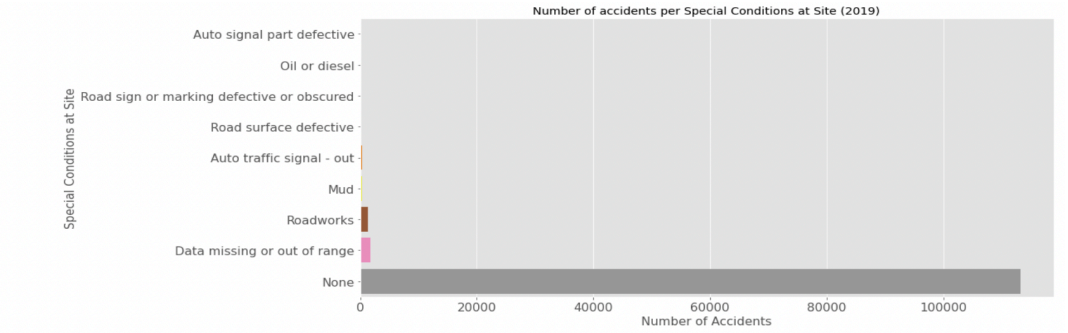
For most of the accidents, the purpose of the journey is not known (or not reported at the time of the accident). It can also be seen that people travelling as part of work or commuting to and from work have a high amount have accidents:

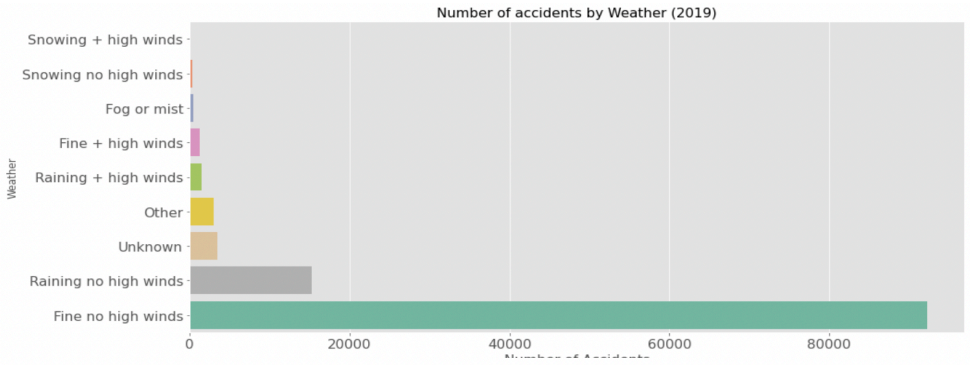


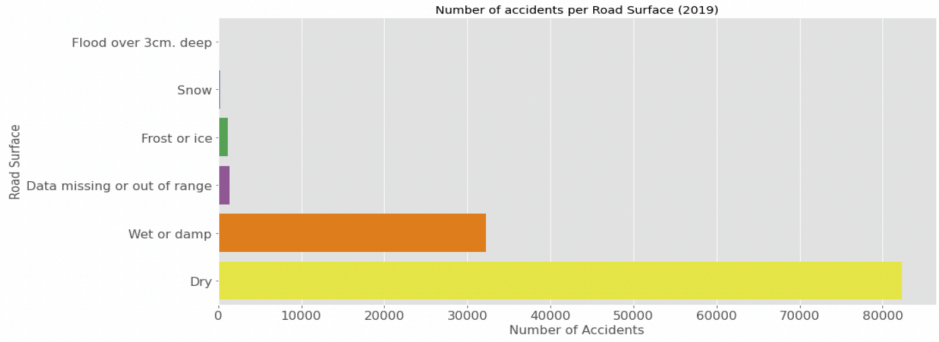
Conditions at the time of accidents:

The conditions are not surprising results, for example, accidents happen more on Dry Surfaces, in Daylight, and in Clear weather.

Further to the above, I also checked if all the rows in Rain conditions are dry.



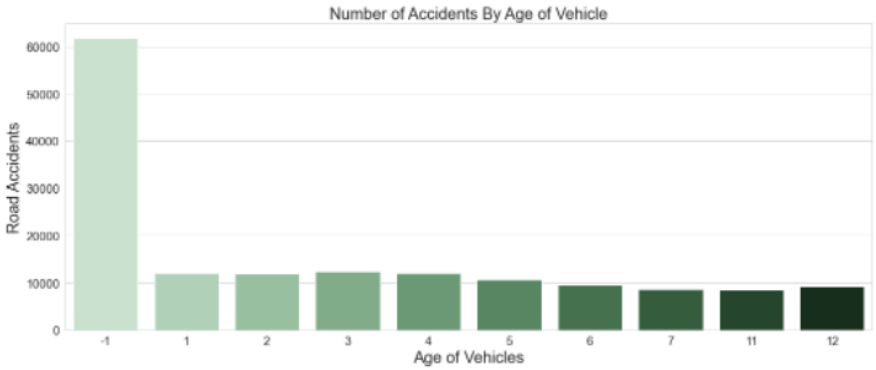




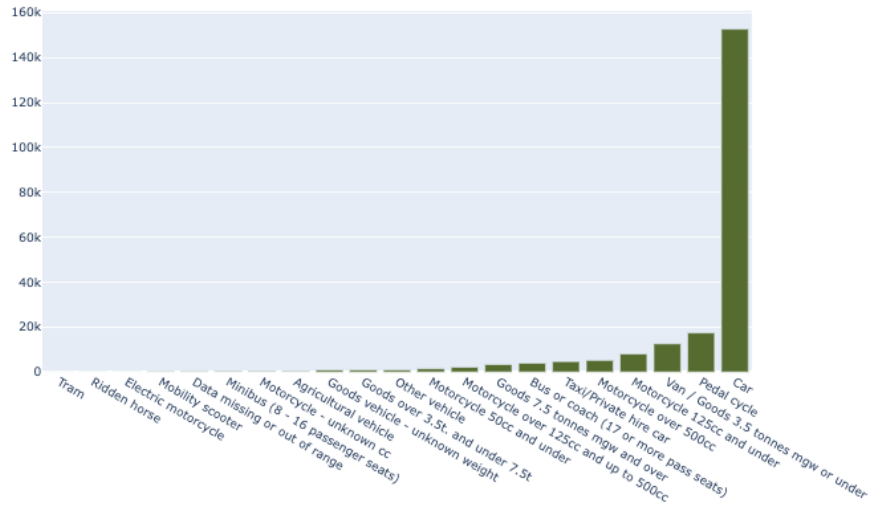
	Weather	Road Surface	Accident_Index
0	Fine + high winds	Data missing or out of range	4
1	Fine + high winds	Dry	711
2	Fine + high winds	Flood over 3cm. deep	1
3	Fine + high winds	Frost or ice	11
4	Fine + high winds	Wet or damp	425
5	Fine no high winds	Data missing or out of range	282
6	Fine no high winds	Dry	74324
7	Fine no high winds	Flood over 3cm. deep	18
8	Fine no high winds	Frost or ice	638
9	Fine no high winds	Snow	29
10	Fine no high winds	Wet or damp	12632
14	Fog or mist	Wet or damp	246
13	Fog or mist	Frost or ice	38
11	Fog or mist	Data missing or out of range	1
12	Fog or mist	Dry	87
15	Other	Data missing or out of range	44
16	Other	Dry	1305
17	Other	Flood over 3cm. deep	5
18	Other	Frost or ice	252
19	Other	Snow	13
20	Other	Wet or damp	1321
26	Raining + high winds	Wet or damp	1263
25	Raining + high winds	Snow	1
24	Raining + high winds	Frost or ice	3
22	Raining + high winds	Dry	13
21	Raining + high winds	Data missing or out of range	7
23	Raining + high winds	Flood over 3cm. deep	28
27	Raining no high winds	Data missing or out of range	58
28	Raining no high winds	Dry	286
29	Raining no high winds	Flood over 3cm. deep	101
30	Raining no high winds	Frost or ice	32
31	Raining no high winds	Snow	12
32	Raining no high winds	Wet or damp	13892
36	Snowing + high winds	Snow	20
37	Snowing + high winds	Wet or damp	6
34	Snowing + high winds	Dry	3
33	Snowing + high winds	Data missing or out of range	1
35	Snowing + high winds	Frost or ice	5
38	Snowing no high winds	Data missing or out of range	2
39	Snowing no high winds	Dry	14
40	Snowing no high winds	Frost or ice	46
41	Snowing no high winds	Snow	139
42	Snowing no high winds	Wet or damp	116
46	Unknown	Frost or ice	17
45	Unknown	Flood over 3cm. deep	5
47	Unknown	Snow	2
43	Unknown	Data missing or out of range	944
44	Unknown	Dry	1945
48	Unknown	Wet or damp	412

There are 4 circumstances in which Snowing and Wet/Rainy weather conditions have Dry road surfaces which is strange.

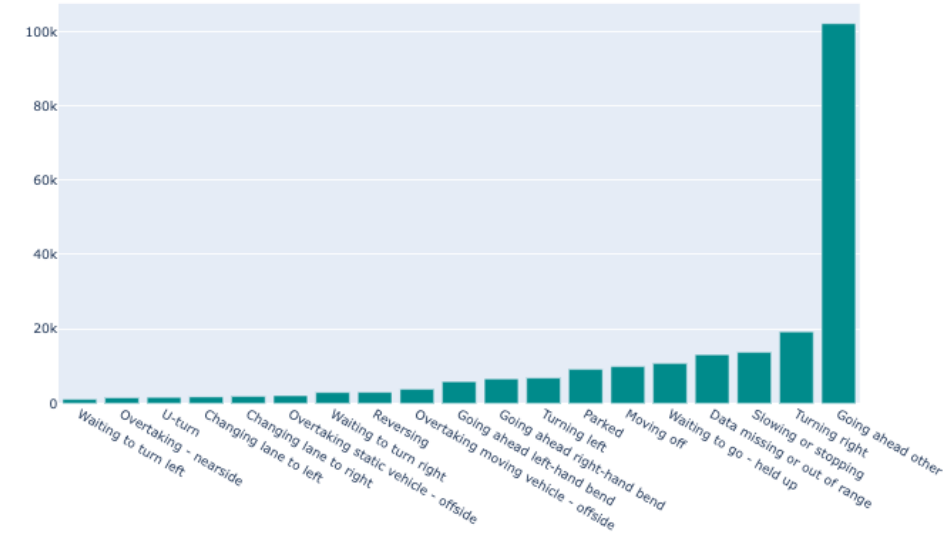
Vehicle Conditions:



Vehicle Type



Vehicle Manoeuvre



From the above graph I observed the following:

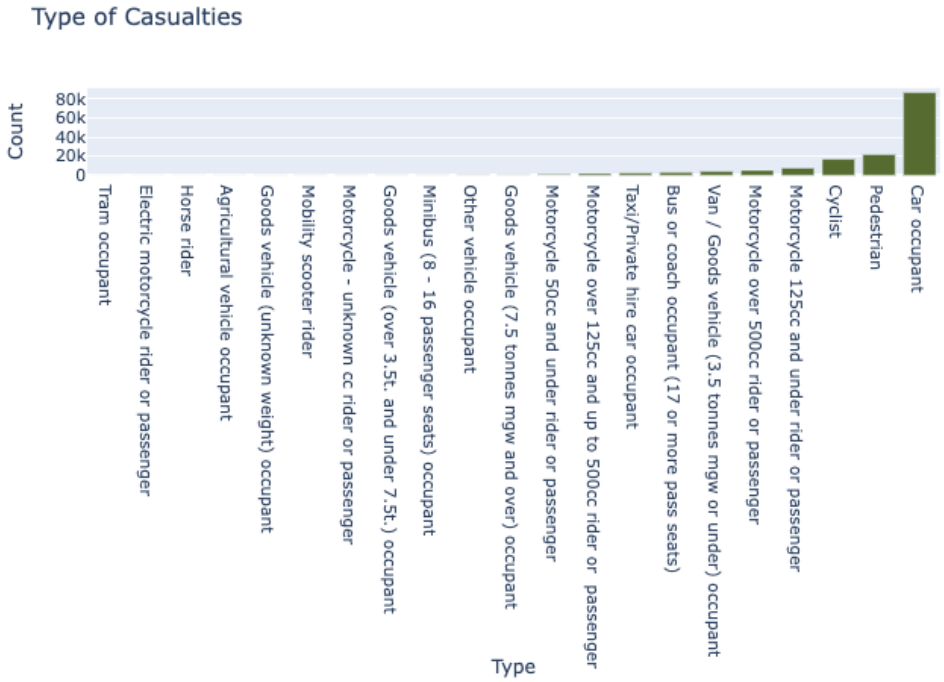
1. Cars are the top vehicles involved in car accidents
2. Electric cars, Environment-friendly cars, were the lowest among other types of cars as compared to Petrol fuelled cars, which probably is because not a lot of people in the population use them as compared to Petrol fuelled cars.
3. New and relatively new Cars scored the highest when looking at age of vehicles, this is an important factor to consider for insurance companies and policymakers.
4. Vehicle Manoeuvre, where Cars are (going ahead of others) was the highest condition, from this information policymakers may establish new penalties to limit the risk of accidents.

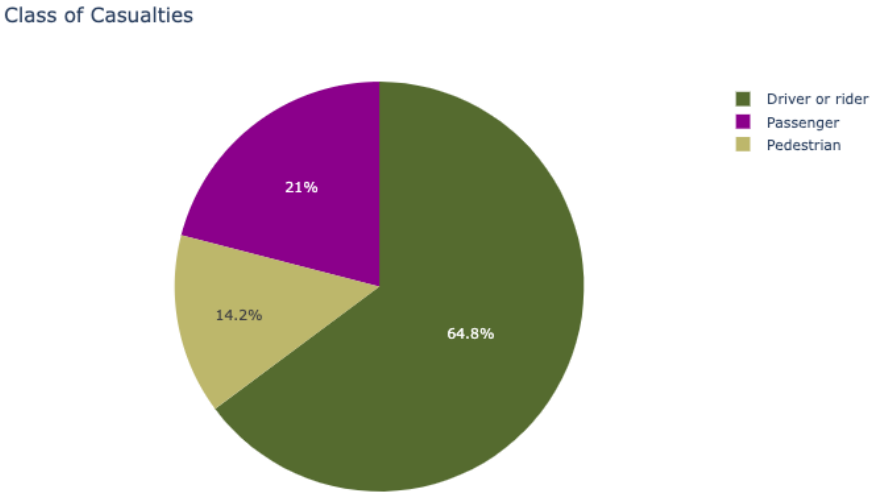
Casualties:

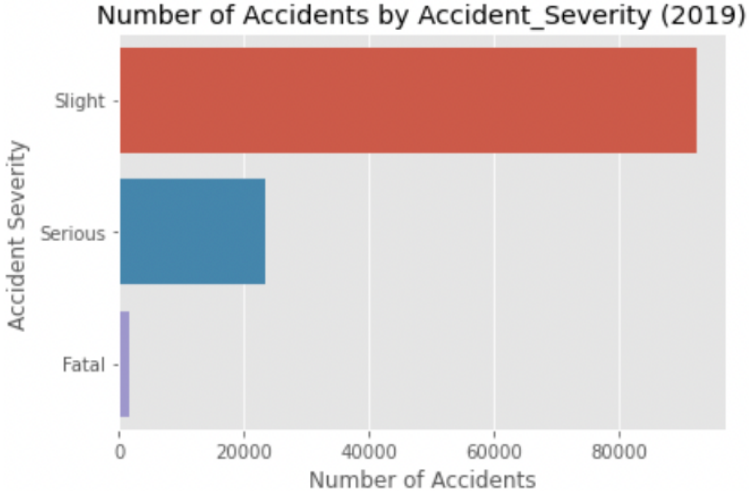
To gain a deeper insight into casualty rates, I explored more features related to casualties impacted by accidents:

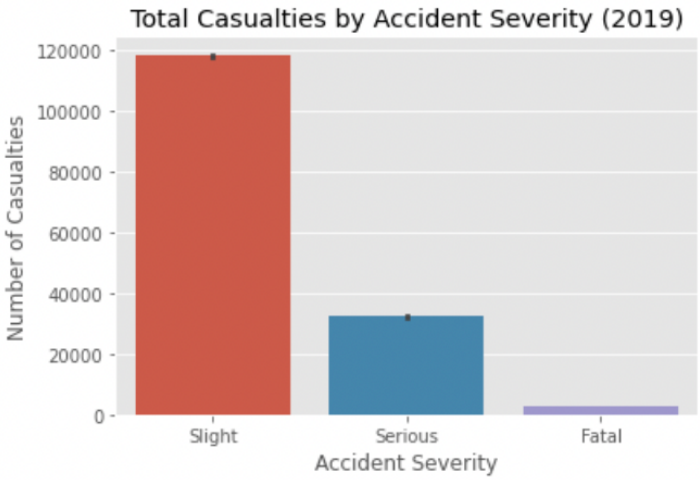
a) Type and class of casualties:

Car occupants account for the majority of casualties: Drives of the vehicles - 64.8% of casualties, followed by passengers with 21% of casualties and lastly Pedestrians with 14% of casualties in the accidents that occurred





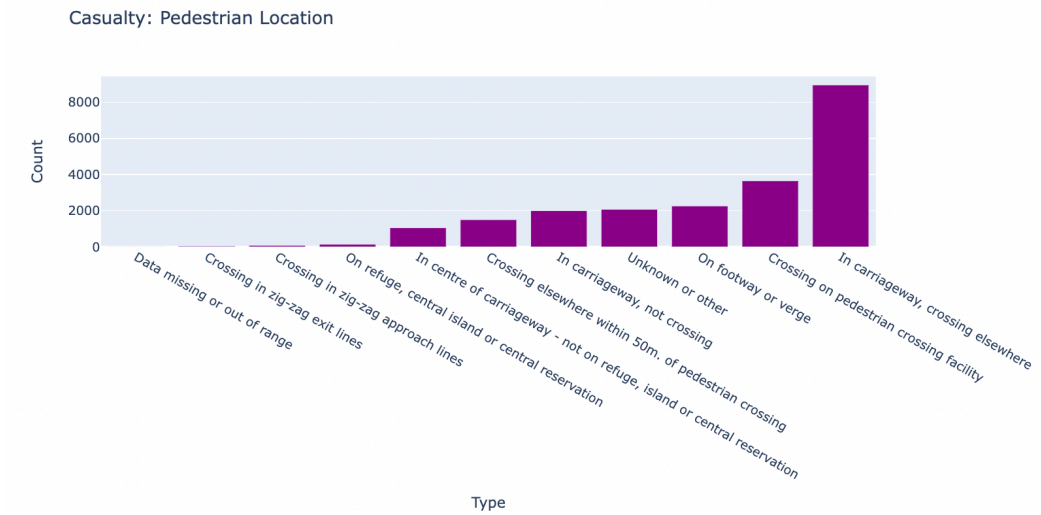




Looking at the graphs above, most of the accidents are labelled “Slight”, in which there is also the highest number of casualties.

b) Location:

There are more casualties in accidents on the carriageway – more than 8000 casualties:



From the above graph, there is a large number of casualties in the accidents that occurred. Measures must be taken to limit this from happening.

When developing road safety, these insights can be extremely useful, for example, when determining pedestrian crossing locations, increasing junction regulations, and installing surveillance and security measures in congested areas.

Predictions

In this section, I use the data to build a predictive model to predict when and where accidents may occur, as well as the severity of the injuries.

I carried out further pre-processing and cleaning of the data so that I can use it for the Machine Learning algorithm. I used Supervised Learning to obtain the severity of an accident.[10] As it is an unbalanced dataset, I focused on achieving good F1 results rather than accuracy, this will aid in detecting the likelihood of an accident occurring.[11]

F1-score is the harmonic mean of Precision and Recall. It provides a better measure of the unbalanced cases than the Accuracy Metric.[12]

---

[10] Almohimeed, R. (2019) *U.K. Traffic Accidents — Data Analysis (10+years)*. Medium. Available online: https://medium.com/@rawanme/u-k-traffic-accidents-data-analysis-10-years-c81293180ee5 [Accessed 17/5/2022].

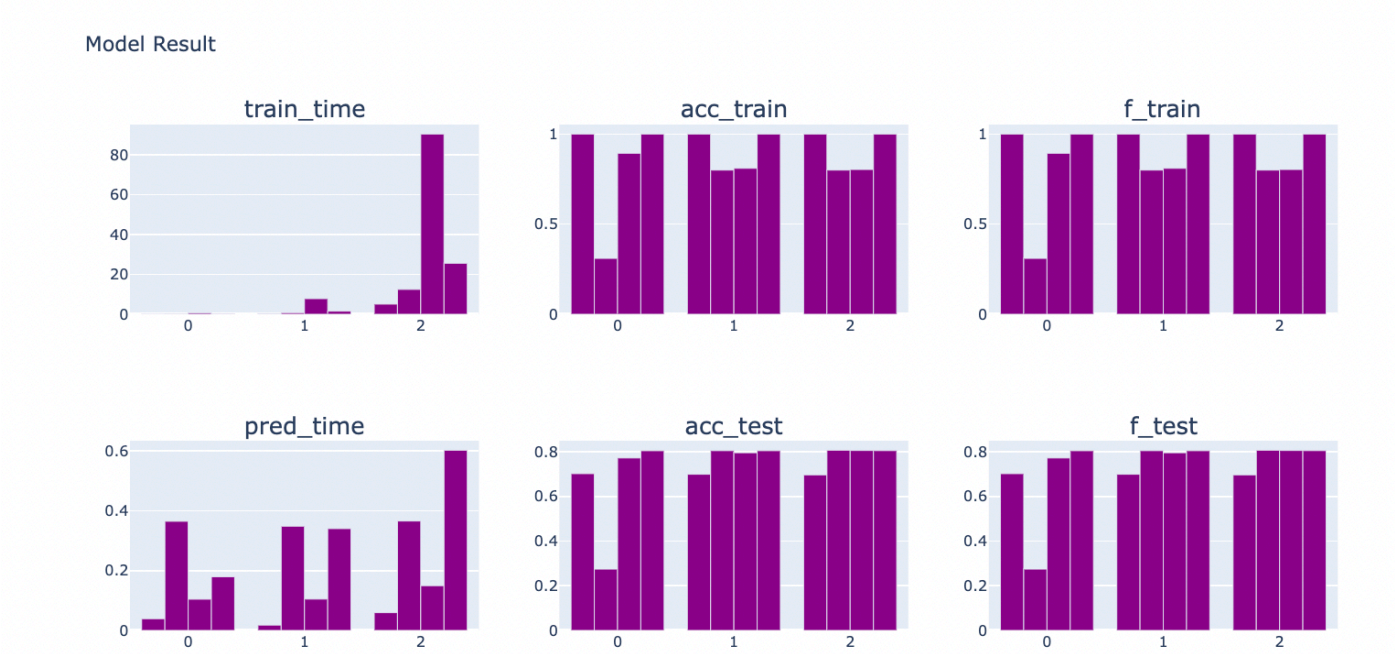
[11] Almohimeed, R. (2019) *U.K. Traffic Accidents — Data Analysis (10+years)*. Medium. Available online: https://medium.com/@rawanme/u-k-traffic-accidents-data-analysis-10-years-c81293180ee5 [Accessed 17/5/2022].

[12] Huilgol *Accuracy Vs. F1-Score.*

$$\text{F1-score} = \left(\frac{\text{Recall}^{-1} + \text{Precision}^{-1}}{2}\right)^{-1} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

[13]

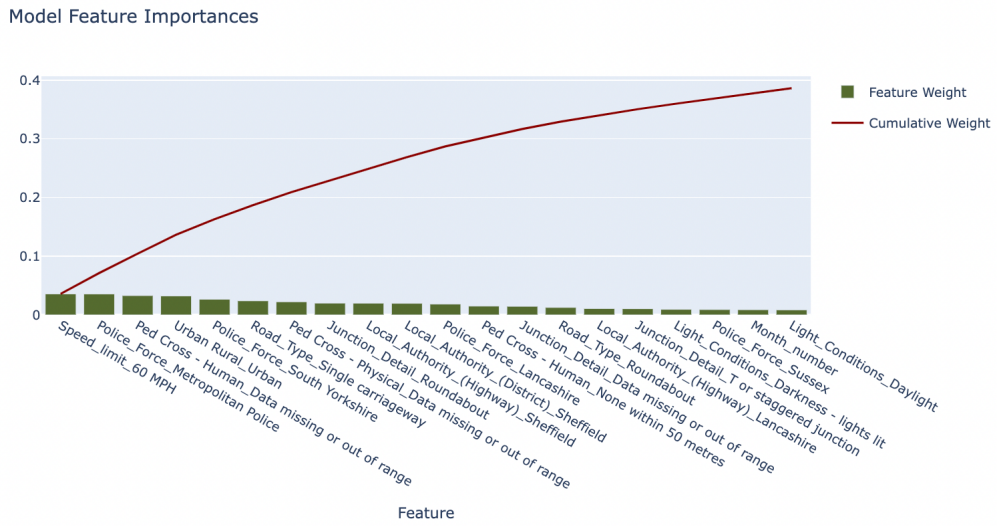
I tested with various classification algorithms - Gradient Boosting, AdaBoost, Decision Tree, and Random Forest. AdaBoost, and Random Forest; both scored very similarly, however considering the training and prediction time, Random Forest was chosen.

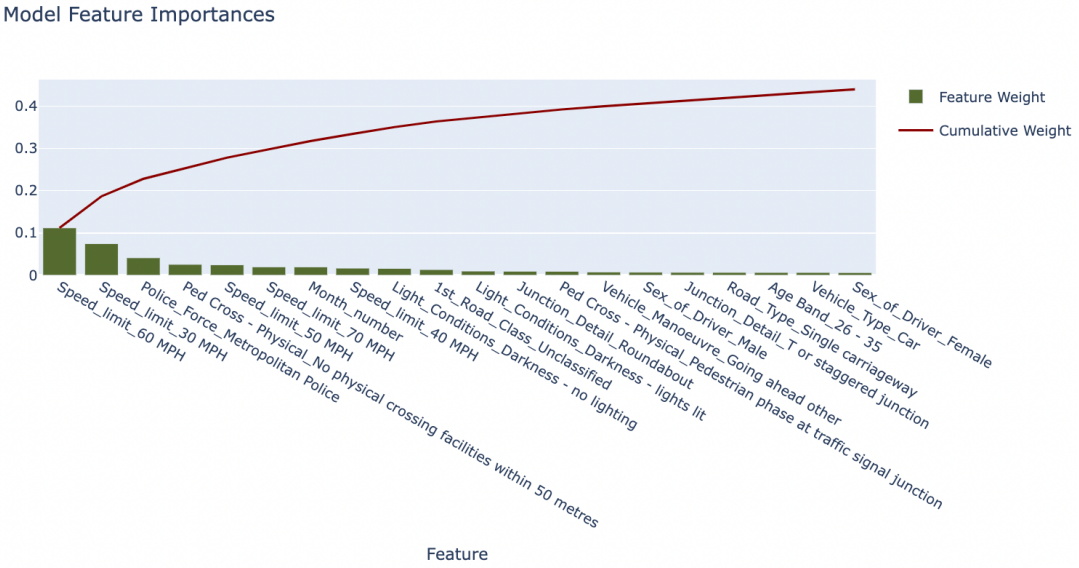


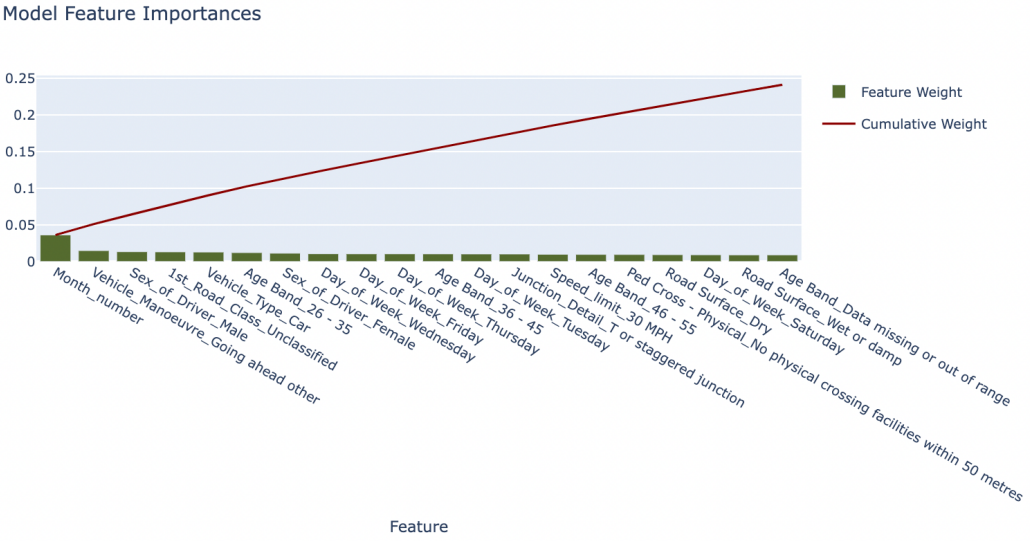

```
RandomForestClassifier(max_depth=15, max_features='sqrt', min_samples_split=100,
                       n_estimators=50, random_state=9)
Unoptimized model
------
Accuracy score on testing data: 0.8055
F-score on testing data: 0.8055

Optimized Model
------
Final accuracy score on the testing data: 0.8071
Final F-score on the testing data: 0.8071
```

[13] Huilgol *Accuracy Vs. F1-Score.*







Looking at the graphs above, the cumulative weight of the features gives the model a good result. Overall, some important outtakes are:

1. The top feature is the months and the speed limit.
2. I can recognize dominating feature: The speed limit of 60 MPH – this means that more accidents could occur on a road with a 60MPH speed limit. This is followed by areas of speed limit of 30 MPH, which as indicated in the analysis, are areas where a lot of accidents occur. These show that there needs to be an increase safety measures in these areas and during specific months.

Recommendations

After analysing the data, I can recommend the following:

- An increase in response time during rush hour periods.
- The construction of other roads to divert traffic from congested areas, thus limiting accidents in those specific areas.
- Enhanced safety and security measures in Urban areas, as these are the areas with a high number of accidents and casualties.
- Clear signs indicating adverse weather or special road conditions to limit and possibly avoid a high number of accidents.
- Easily accessible road safety regulations for the public as well as possibly increased penalties for infringement of the regulations.

I believe that these insights will be useful for policymakers to understand how accidents happen and provide much-needed solutions.

Bibliography/References:

Reported Road Casualties Great Britain, Annual Report: 2019 Available online: https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2019 [Accessed May 17,2022].

Great Britain Road Numbering Scheme, 2022b). https://en.wikipedia.org/w/index.php?title=Great_Britain_road_numbering_scheme&oldid=1087647981. [Accessed 10/8/2022].

Staines, T. (2018) Car Crashes and the Weather: An Exploratory Analysis of Environmental Conditions' Impact on Traffic…. Medium. Available online: https://towardsdatascience.com/car-crashes-and-the-weather-an-exploratory-analysis-of-environmental-conditions-impact-on-traffic-12bcb7f9afed [Accessed 17/5/2022].

Advanced EDA of UK's Road Safety Data using Python Available online: https://omdena.com/blog/advanced-eda/ [Accessed Aug 10,/ 2022].

Road Safety Data - data.gov.uk. (2022) Data.gov.uk. Available online: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data [Accessed 17/5/2022].

Almohimeed, R. (2019) U.K. Traffic Accidents — Data Analysis (10+years). Medium. Available online: https://medium.com/@rawanme/u-k-traffic-accidents-data-analysis-10-years-c81293180ee5 [Accessed 17/5/2022].

P. Huilgol (-08-24) *Accuracy Vs. F1-Score.* Available online: https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2 [Accessed May 17,2022].

Calculate and Plot a Correlation Matrix in Python and Pandas • datagy. (n.d.) datagy. Available online: https://datagy.io/python-correlation-matrix/ [Accessed 17/5/2022].