# FIT5147 Data Exploration Project

April 2019

**Roopesh Kumar Ramesh**
Student ID: 30344565
Email: rram0019.student.monash.edu
Master of Data Science
Monash University

# Introduction

The UK government collects and publishes (usually on an annual basis) detailed information about traffic accidents across the country. This information includes, but is not limited to, geographical locations, weather conditions, type of vehicles, number of casualties and vehicle manoeuvres, making this a very interesting and comprehensive dataset for analysis and research.

# Data Wrangling

Vehicle_data and Accident_data was taken from the following URL: https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles.

Car features and MSRP was taken from the following URL: https://www.kaggle.com/CooperUnion/cardataset/version/1#

There are two datasets:

- Accident Information
- Vehicle Information

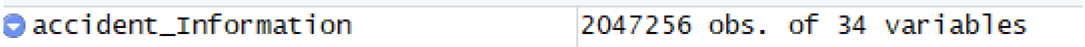
Accident Information contains the information regarding the Date, Latitude, Number of Causalities, Number of Vehicles, Severity of the accident etc

Vehicle Information contains information about the type of vehicle, Age band of driver, vehicle make, vehicle model etc.

Car dataset contains information about the make, model MSRP etc.

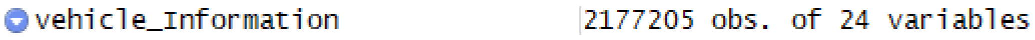
R programming language has been used for all Data Wrangling involved.

I have merged the two datasets based on a common column called Accident_Index.



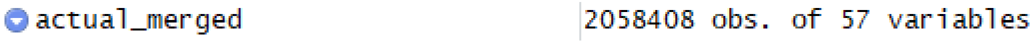
```
accident_Information          2047256 obs. of 34 variables
```

*Figure 1: Accident Information*



```
vehicle_Information           2177205 obs. of 24 variables
```

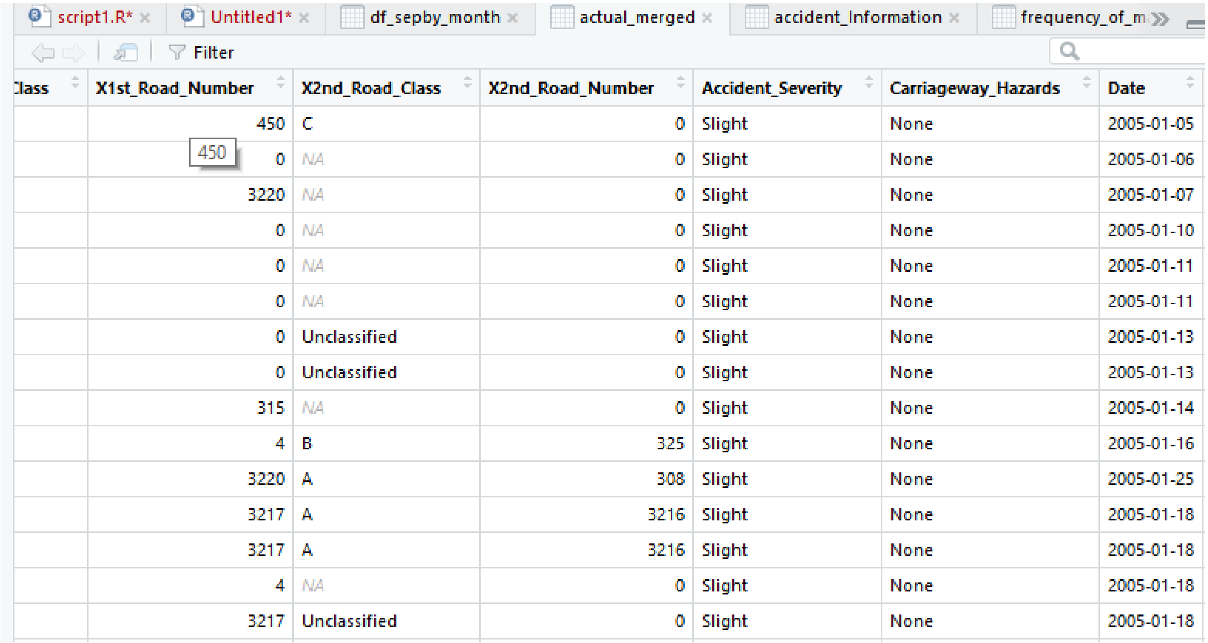
*Figure 2: Vehicle Information*



```
actual_merged                 2058408 obs. of 57 variables
```

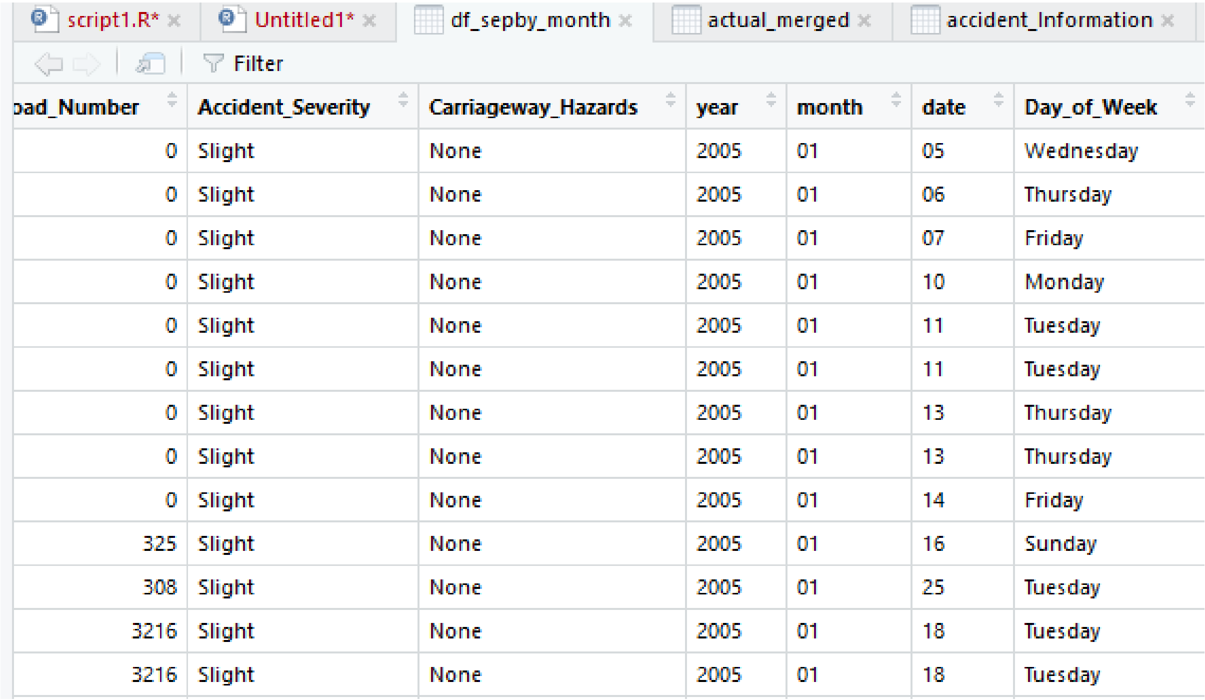
*Figure 3: Merged DataFrame*

Further, I have restructured the Date column to remove the Date, Month and Year independently and convert them to Appropriate *Date* datatypes.



| lass | X1st_Road_Number | X2nd_Road_Class | X2nd_Road_Number | Accident_Severity | Carriageway_Hazards | Date |
|---|---|---|---|---|---|---|
| | 450 | C | 0 | Slight | None | 2005-01-05 |
| | 0 | NA | 0 | Slight | None | 2005-01-06 |
| | 3220 | NA | 0 | Slight | None | 2005-01-07 |
| | 0 | NA | 0 | Slight | None | 2005-01-10 |
| | 0 | NA | 0 | Slight | None | 2005-01-11 |
| | 0 | NA | 0 | Slight | None | 2005-01-11 |
| | 0 | Unclassified | 0 | Slight | None | 2005-01-13 |
| | 0 | Unclassified | 0 | Slight | None | 2005-01-13 |
| | 315 | NA | 0 | Slight | None | 2005-01-14 |
| | 4 | B | 325 | Slight | None | 2005-01-16 |
| | 3220 | A | 308 | Slight | None | 2005-01-25 |
| | 3217 | A | 3216 | Slight | None | 2005-01-18 |
| | 3217 | A | 3216 | Slight | None | 2005-01-18 |
| | 4 | NA | 0 | Slight | None | 2005-01-18 |
| | 3217 | Unclassified | 0 | Slight | None | 2005-01-18 |

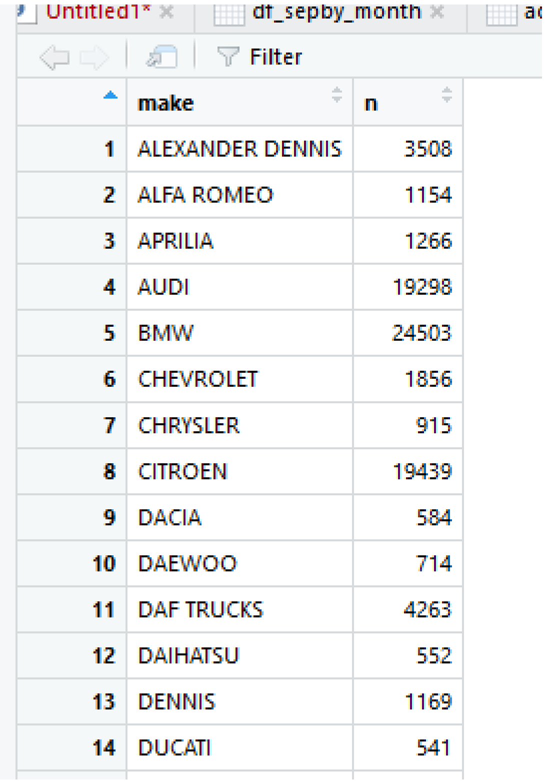
*Figure 4: Date in single column*



| oad_Number | Accident_Severity | Carriageway_Hazards | year | month | date | Day_of_Week |
|---|---|---|---|---|---|---|
| 0 | Slight | None | 2005 | 01 | 05 | Wednesday |
| 0 | Slight | None | 2005 | 01 | 06 | Thursday |
| 0 | Slight | None | 2005 | 01 | 07 | Friday |
| 0 | Slight | None | 2005 | 01 | 10 | Monday |
| 0 | Slight | None | 2005 | 01 | 11 | Tuesday |
| 0 | Slight | None | 2005 | 01 | 11 | Tuesday |
| 0 | Slight | None | 2005 | 01 | 13 | Thursday |
| 0 | Slight | None | 2005 | 01 | 13 | Thursday |
| 0 | Slight | None | 2005 | 01 | 14 | Friday |
| 325 | Slight | None | 2005 | 01 | 16 | Sunday |
| 308 | Slight | None | 2005 | 01 | 25 | Tuesday |
| 3216 | Slight | None | 2005 | 01 | 18 | Tuesday |
| 3216 | Slight | None | 2005 | 01 | 18 | Tuesday |

*Figure 5: Date split into appropriate datatypes*

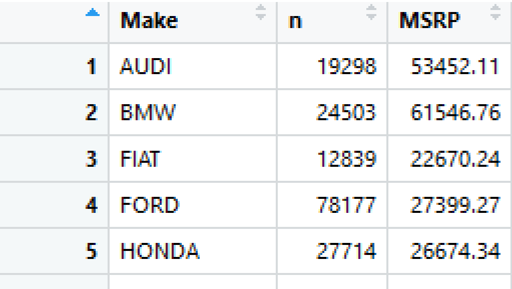
I have transformed dataset to obtain frequency of various occurrences. For example, to get the frequency of the make of vehicle, I had transformed the dataset appropriately.



| | make | n |
|---|---|---|
| 1 | ALEXANDER DENNIS | 3508 |
| 2 | ALFA ROMEO | 1154 |
| 3 | APRILIA | 1266 |
| 4 | AUDI | 19298 |
| 5 | BMW | 24503 |
| 6 | CHEVROLET | 1856 |
| 7 | CHRYSLER | 915 |
| 8 | CITROEN | 19439 |
| 9 | DACIA | 584 |
| 10 | DAEWOO | 714 |
| 11 | DAF TRUCKS | 4263 |
| 12 | DAIHATSU | 552 |
| 13 | DENNIS | 1169 |
| 14 | DUCATI | 541 |

*Figure 6: Dataset transformed to get frequency of make*

Also, the dataset I had chosen did not have the price of the cars, so I had to look for a separate dataset to answer my questions. I merged the dataset I found with the frequency_of_model dataset by *Make* to get the price included.



| | Make | n | MSRP |
|---|---|---|---|
| 1 | AUDI | 19298 | 53452.11 |
| 2 | BMW | 24503 | 61546.76 |
| 3 | FIAT | 12839 | 22670.24 |
| 4 | FORD | 78177 | 27399.27 |
| 5 | HONDA | 27714 | 26674.34 |

*Figure 7: Added MSRP to frequency_car_make*

# Data Cleaning and Checking

I have used R programming language for data checking and cleaning purposes.

I have plotted a graph of the missing values.

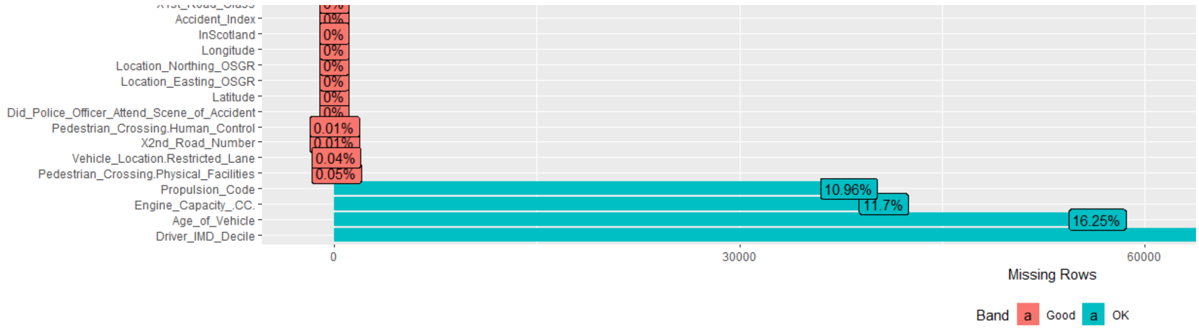


*Figure 8: Missing Values*

From the graph, quite clearly, Propulsion_Code, Engine_Capacity _.CC, Age_of_Driver, Driver_IMD_Decile have a lot of missing values. Subsequently, I have dropped these columns from the dataset.

I have removed major outliers from some of the columns. I have plotted frequency of the variable to get the outliers.

Number of Vehicles involved in accident:

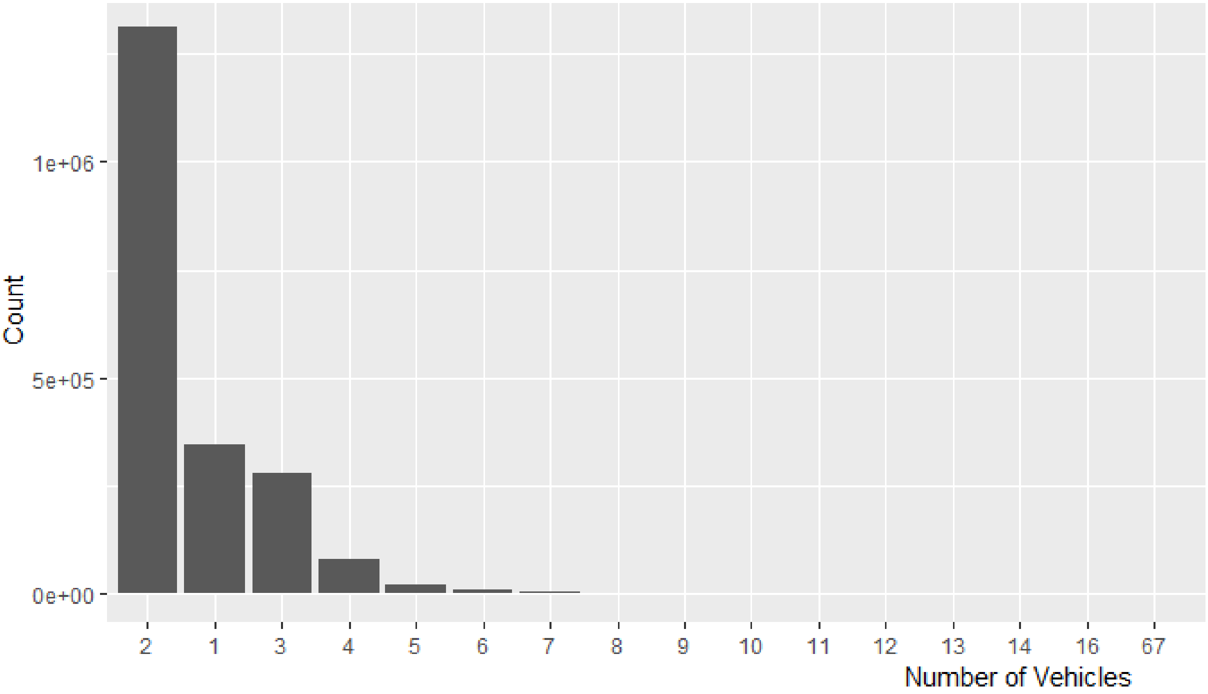


*Figure 9: Frequency of number of vehicles involved*

Clearly, any number of vehicles above 5 is redundant in the dataset and while they may not exactly be outliers, they are so exactly helpful in the analysis and therefor have been removed.

Number of Casualties involved in accident:

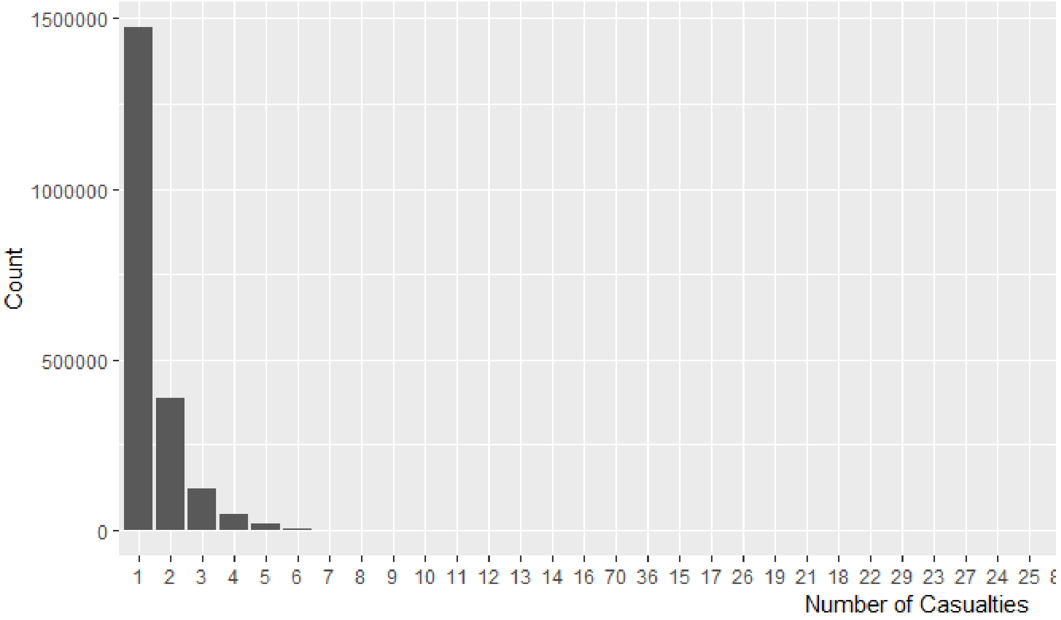
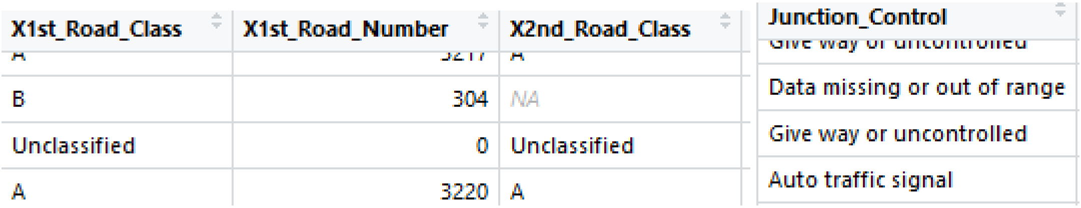


*Figure 10: Frequency of number of Casualties*

Again, any number of casualties above 5 is negligible.

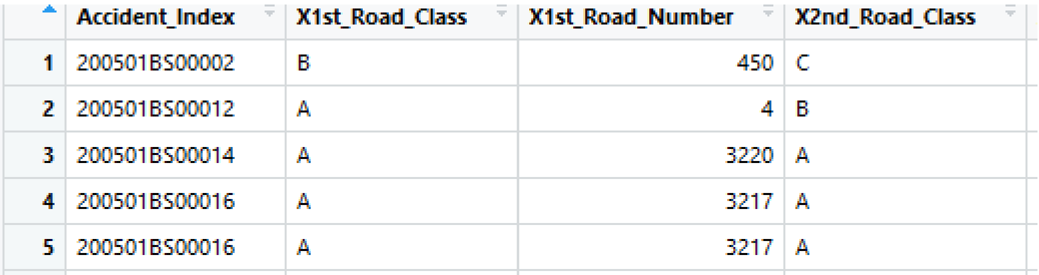
I have removed all the *NA*, *NULL* and *NOT APPLICABLE* values wherever possible.



| X1st_Road_Class | X1st_Road_Number | X2nd_Road_Class |
|---|---|---|
| A | 3217 | A |
| B | 304 | NA |
| Unclassified | 0 | Unclassified |
| A | 3220 | A |

| Junction_Control |
|---|
| Give way or uncontrolled |
| Data missing or out of range |
| Give way or uncontrolled |
| Auto traffic signal |

*Figure 11: Dataframe with redundant values*



| | Accident_Index | X1st_Road_Class | X1st_Road_Number | X2nd_Road_Class |
|---|---|---|---|---|
| 1 | 200501BS00002 | B | 450 | C |
| 2 | 200501BS00012 | A | 4 | B |
| 3 | 200501BS00014 | A | 3220 | A |
| 4 | 200501BS00016 | A | 3217 | A |
| 5 | 200501BS00016 | A | 3217 | A |

*Figure 12: Redundant values cleaned*

# Data Exploration

I have used R programming language for all Data Exploration and Visualizations.

## Which car Make is likely to cause most of the accidents?

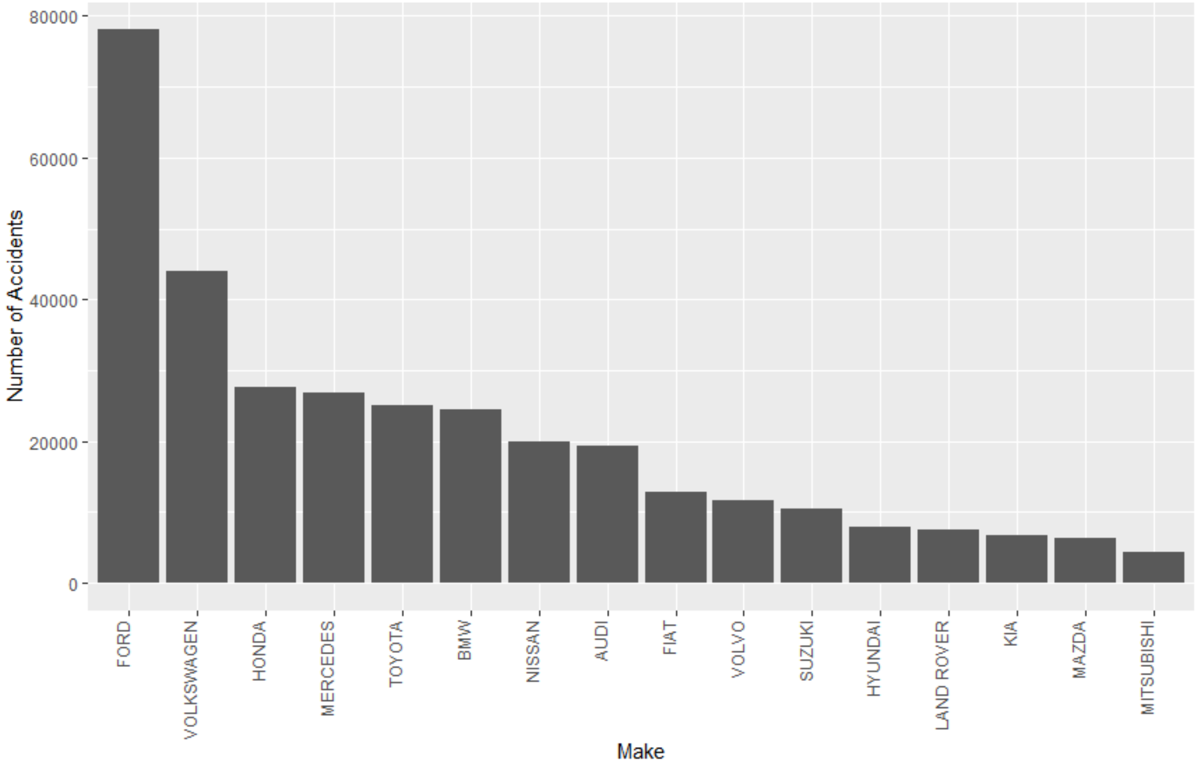


*Figure 13: Make vs Number of Accidents*

From the above figure it is clear that Ford is involved in the greatest number of accidents.  Followed, after quite a margin by Volkswagen.

The high number of Ford cars involved in accidents can be taken as an indication of safety features Ford cars are equipped with.

I have only shown car models involved in at least 3000 accidents from 2005-2016.

## Economic Background of Drivers that cause most accidents

This was a question could not be fully answered as the only financial data given in the dataset was the price of the cars.

As such, I have estimated the financial background of the driver solely based on the average price of the car he drives.

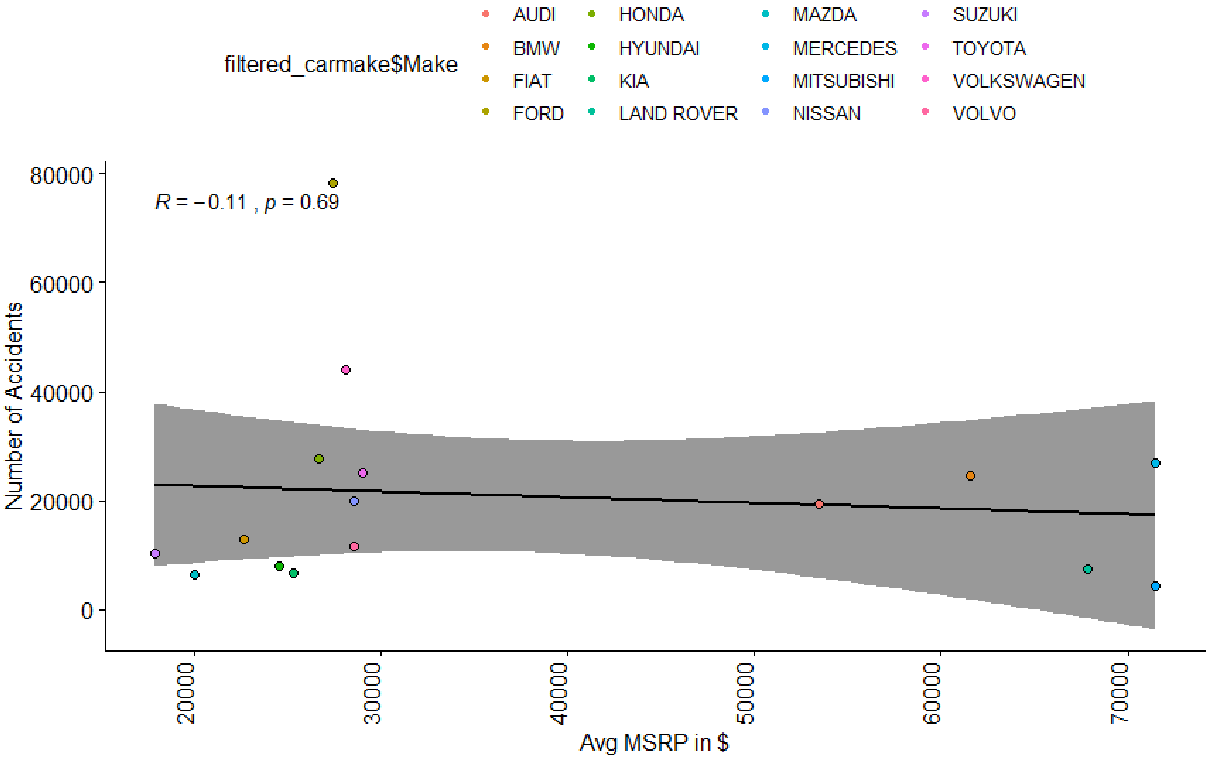


*Figure 14: Correlation between Number of Accidents and average price of Car Make*

The above figure depicts the correlation between the average price of a Car Make and the number of accidents it is involved in.

Abnormally high values for Ford and Volkswagen can also be attributed to their large numbers on the streets,

Fitting a linear regression curve, we find that there is a weak downhill linear relationship between the average price of cars and the number of accidents they are involved in i.e., as the price of a car increases, it is just slightly less probable to be involved in a crash. This is denoted by the Pearson’s correlation coefficient r, which in this case is -0.11.

Granted, the P value is quite high to take this as a strong evidenced analysis, that might be due to fact that we only have the number of vehicles that are involved in the crash but not the total number of vehicles that are present on the road.

This analysis shows that the involvement of the car in an accident is irrespective of the price of the car, for the most part.

## Number of Accidents according to Age Group

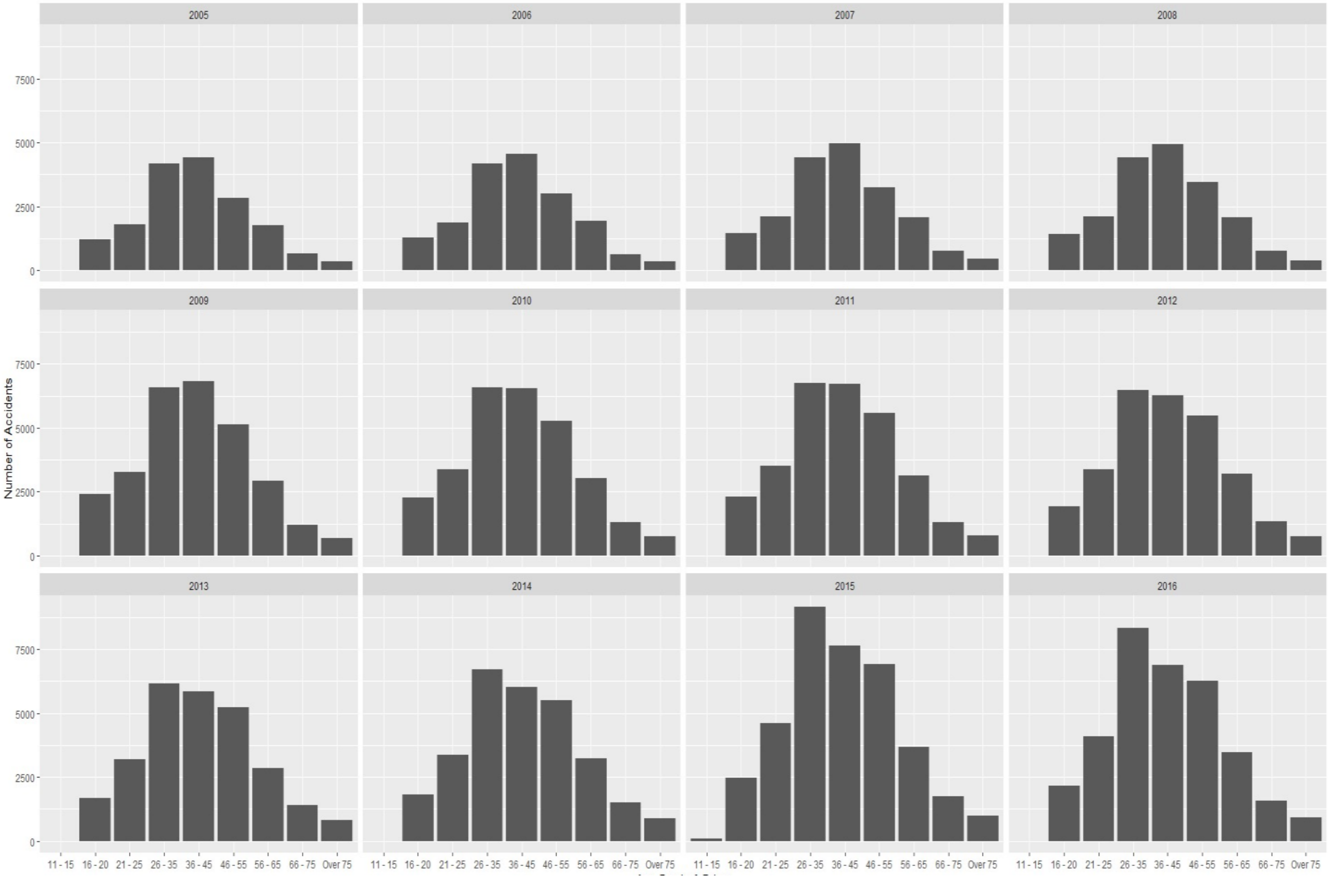


*Figure 15: Number of Accidents according to age Group over the course of the years*

As seen from the above visualization, in the past, maximum number of accidents were caused by Drivers within the age groups 36-45. But this has changed mid-way and as of 2016, the maximum accidents have been caused by drivers in the age group 26-35. This change can be credited to vehicles being made easily accessible to everyone.

Also, taking the overall effect of driver age group that cause the accidents, it is quite evident that Drivers in the age groups 26-35 cause the most accidents as shown in this visualization.

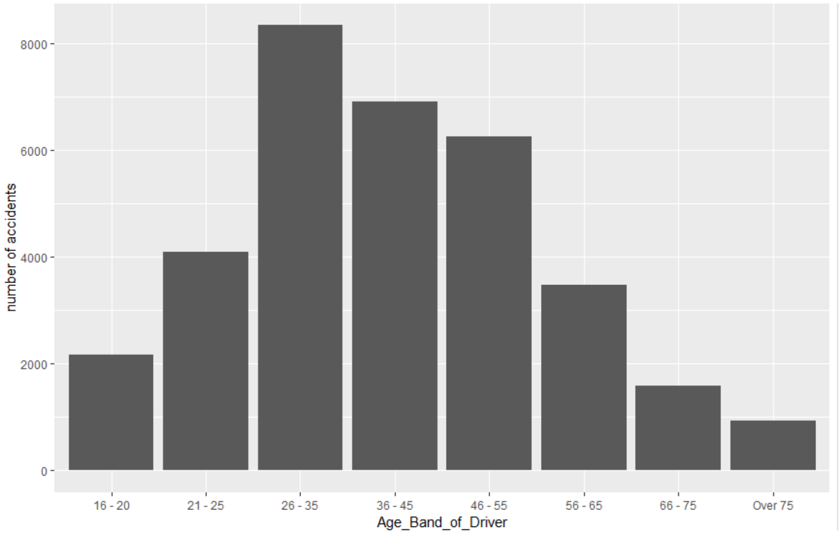


*Figure 16: Overall Age group of drivers involved in accidents*

# Conclusion

The data provides insights on the effects of Vehicle make, Age Band of Drivers, and many such factors on the number of accidents caused.

Think twice before buying a Ford vehicle. It seems to be more than twice as likely to be involved in an accident that the next car maker on the list.

The price of a car does not give us any indication of its likeliness to be involved in an accident. Expensive cars are just as likely to be involved in an accident as inexpensive ones, for the most part.

Looking at the trend in the number of accidents involving a particular age group, between 2005 there had been substantial increase in the percentage of drivers within the age group 18-25 over the percentage of drivers in the age group 25-35 up until 2014. After 2014, there doesn't seem to be any change as the percentage of drivers in the 18-25 age group reaches a stable state.

Most of the accidents occur during the week days and during rush hours indication that people getting to and from work to office and vice versa are more likely to be involved in accidents.

# Reflection

The dataset I had chosen was quite large. Around 2,000,000 rows and 57 columns in total. This project gave me an opportunity to wrangle, explore, and visualize such a huge dataset.

In most of my analysis, I have taken the total sum of accidents across the years. In hindsight, I think it would've been better to do year wise analysis and then compare all the analysis to figure out the trends.

# Bibliography

- Vehicle Data and Accident Data: https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles.
- Car features and MSRP: https://www.kaggle.com/CooperUnion/cardataset/version/1#