The United Kingdom Road Safety Analysis

Group Name: Royal
Members: Yining Wang, Tian Gu, Shenglan Zheng, Zihao Wang

I. Introduction

A. Overview:

According to the World Health Organization, more than 1.25 million people die each year as a result of road traffic crashes. We take UK road safety data from 2005 to 2016 as our research object to analyze the car accidents conditions and vehicle information, and decide to go deep in the following 4 parts:

1. the accidents geographic distribution
2. the accidents distribution across the season and how the weather and road surface conditions will affect the accidents
3. the accidents distribution across day during weekday and day during weekend, to know which time period incur the most accidents
4. the factors affect accidents severity

We hope that our analysis would potentially aid the UK Department for Transport to implement accident prevention in certain high incidence of accidents area and time period. In addition to that, people can also get some advice from us in purchasing vehicles, and get to know some matters need attention.

B. Data Resource:

Our original data is from Kaggle, and the original author downloaded it from an open data website of the UK government.
https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles
https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

Our data includes two datasets:
Accident_Information.csv: every line in the file represents a unique traffic accident (identified by the Accident_Index column), featuring various properties related to the accident as columns. Date range: 2005-2016
Vehicle_Information.csv: every line in the file represents the involvement of a unique vehicle in a unique traffic accident, featuring various vehicle and passenger properties as columns. Date range: 2004-2016.

II. Data Cleaning

1. We filter the `Accident` dataset to only include accidents whose number of vehicles equals 1. Since accidents that have two or more vehicles include both trouble cars and innocent cars, and those innocent cars were not influenced by factors that we want to explore.
2. We joined the `Accident` and `Vehicle` by a common column called `Accident_Index` into a new dataset called `Aggregate`
3. We clean `Aggregate` by selecting variables that are useful for our analysis. Which includes : `Accident_Index,Longitude,Latitude, Accident_Severity, Date, Day_of_Week, Age_Band_of_Driver, Age_of_Vehicle, Sex_of_Driver,Engine_Capacity_.CC.`

Here is a partial look of the dataset Aggregate.

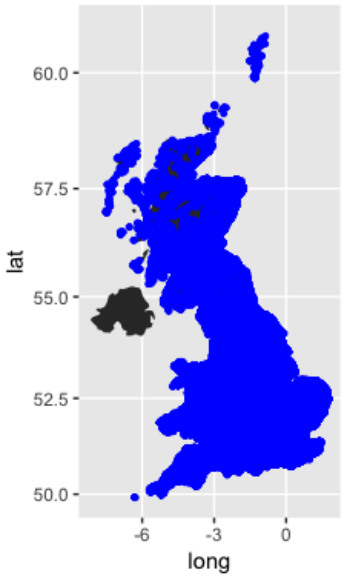
	Accident_Index	Longitude	Latitude	Accident_Severity	Date	Day_of_Week
1	200501BS00001	-0.191170	51.48910	Serious	2005-01-04	Tuesday
2	200501BS00002	-0.211708	51.52007	Slight	2005-01-05	Wednesday
3	200501BS00004	-0.173862	51.48244	Slight	2005-01-07	Friday

Age_Band_of_Driver	Age_of_Vehicle	Sex_of_Driver	Engine_Capacity_.CC.
NA	*NA*	*NA*	*NA*
36 - 45	3	Male	8268
46 - 55	4	Female	1769

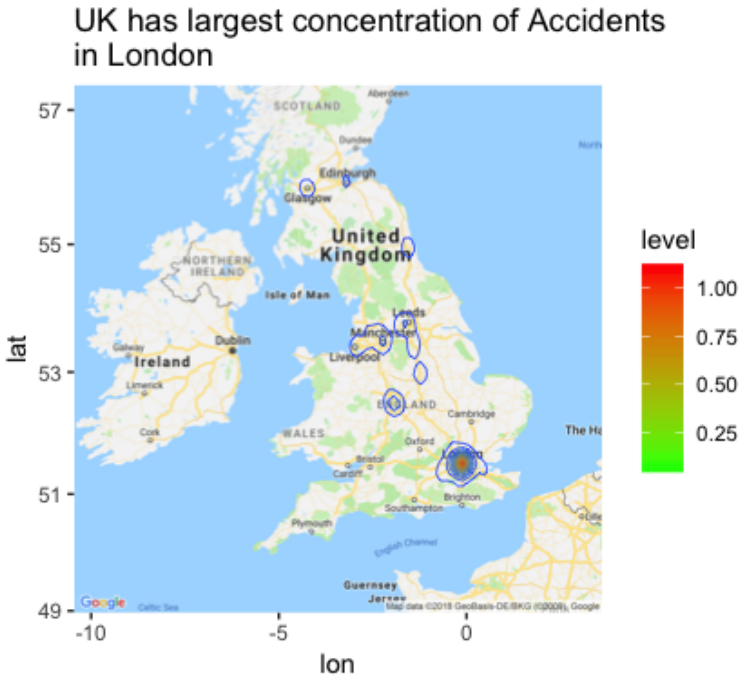
III. Factor Analysis

A. Factor 1: Location

Since the data is very huge, the first thing we do is to visualize the accidents on a map. By doing so, we could see how accidents distributed, and which locations happened most frequently, so that the government and citizens could be cautioned. We firstly dot plot the accidents into a UK map called `dotmap` by their Latitude and Longitude, and dots are colored in blue.

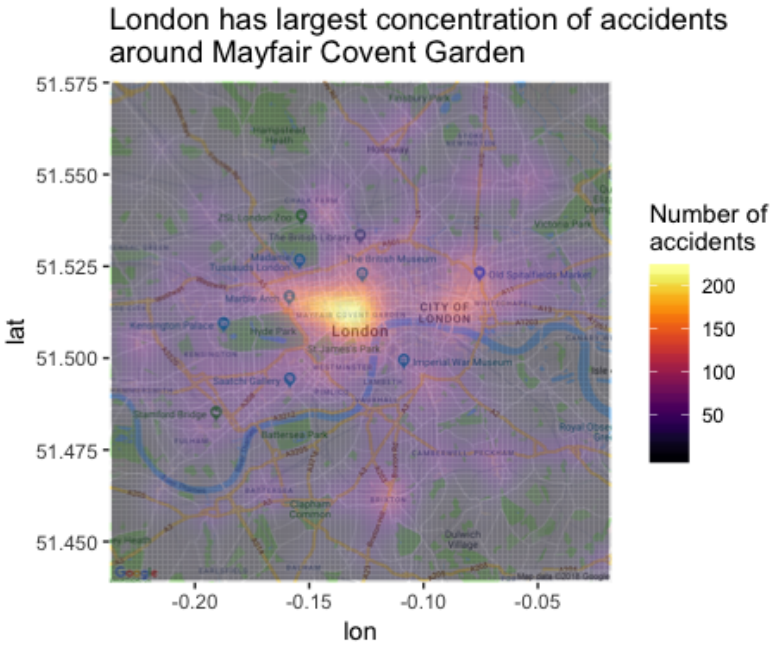


However, since our data are too large, we cannot see the intensity of accidents, even we change the dot size into the smallest size. So we then visualize the result by a density map to see the degree of accident concentration, and we use google map as a background because it helps us directly see the name of the location.



Through this density map, we could accidents are concentrated in London, Birmingham, Liverpool, Manchester, Leeds, and Glasgow. Unsurprisingly, those cities are ranked in the top ten of the list of major cities in England.

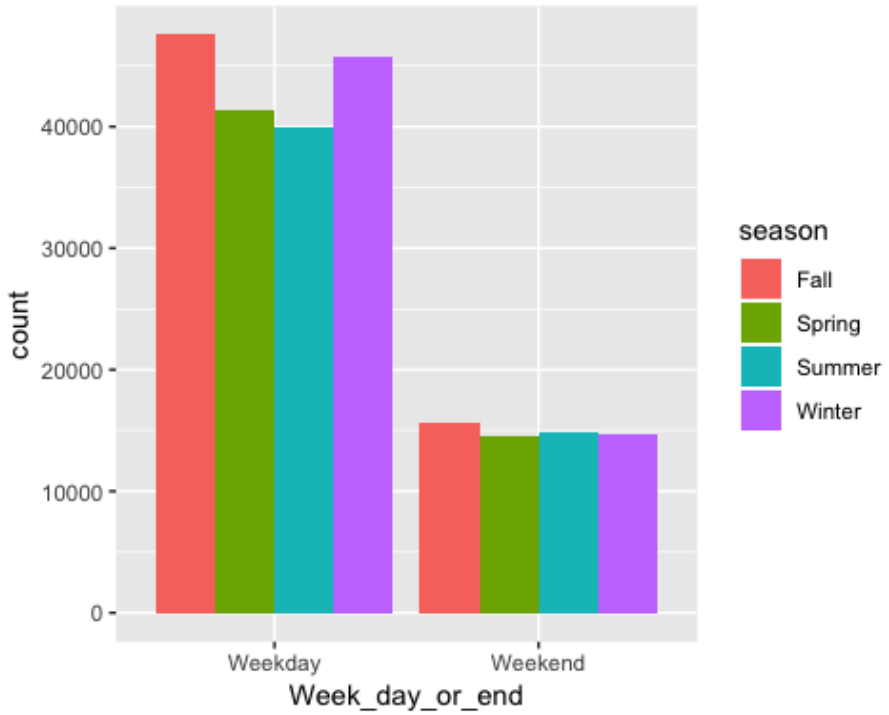
In this graph above, London has the most complex (highest intensity and most types to intensity) traffic accident events over the United Kingdom. So we decided to zoom the map and take a further look at the accidents distribution in London, and we draw a heatmap, to see the distribution of the number of accidents.



In this graph, locations that are more accidents concentrated is colored in yellow. We could see that only one area dominated the highest number of accidents over 2005 to 2026, and it is the area around Covent Garden. According to our research, Covent Garden is London's main theatre and entertainment area for both shopping and sightseeing. There are many narrow streets around this area, filled with street artists, visitors, and shops. So we think the London government should measure to improve in this area by controlling the visitors' flow rate, improve road construction and etc.

B. Factor 2: Season Change and Weather Conditions

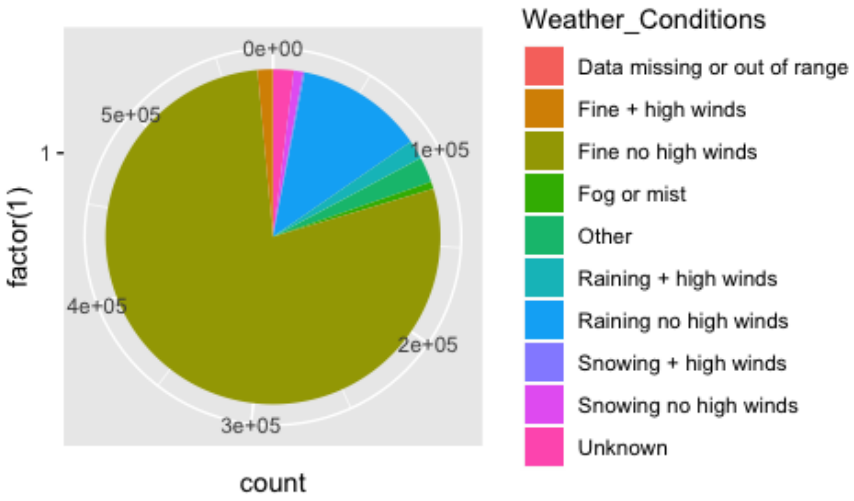
In order to know how the season will affect the number of accidents happened in terms of weekday or weekend, we make the following graph,

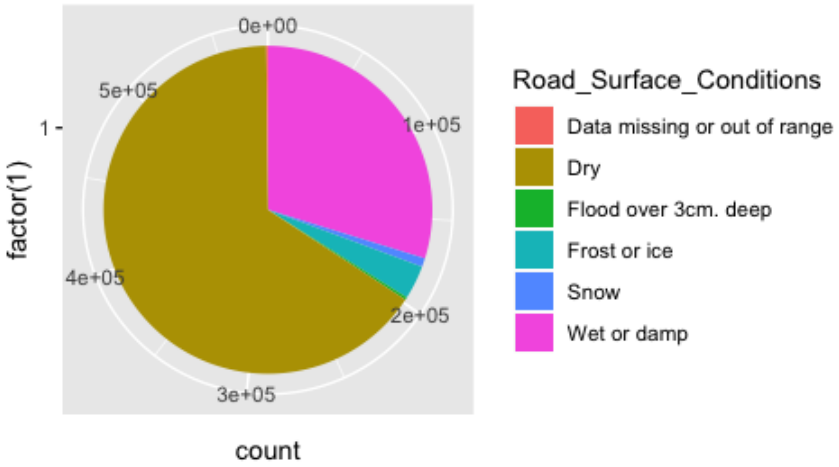


According to the graph above, we find that during the weekday, accidents happened in fall and winter is much higher than accidents happened during spring and summer. For the weekend, seasonal differences don't have much impact on it.

Then we search the UK weather and find out that, during autumn and winter, the weather is wet and windy with average monthly rainfall of 80mm, and in sometimes UK residences suffer very wintry conditions that come with frost and snowfall. However, during spring and summer, the weather is dry and hot with average rainfall of 60mm.

Based on the fact, we assume that rainy or windy day with a wet road surface will increase the probability of accidents happen. Therefore, we make the following graph:



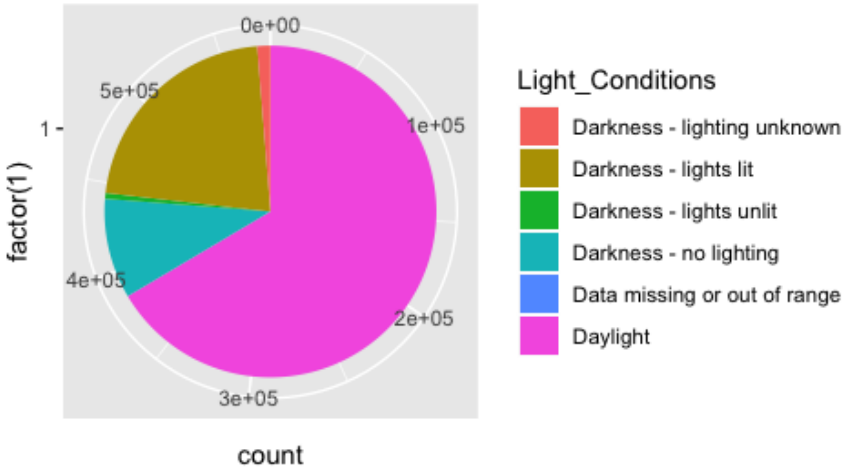


However, the truth is unexpected. A major number of accidents were happened in fine with no high winds weather and in dry road surface condition. Accidents happened in raining with no high winds, and in wet or damp road surface condition comes into second place.

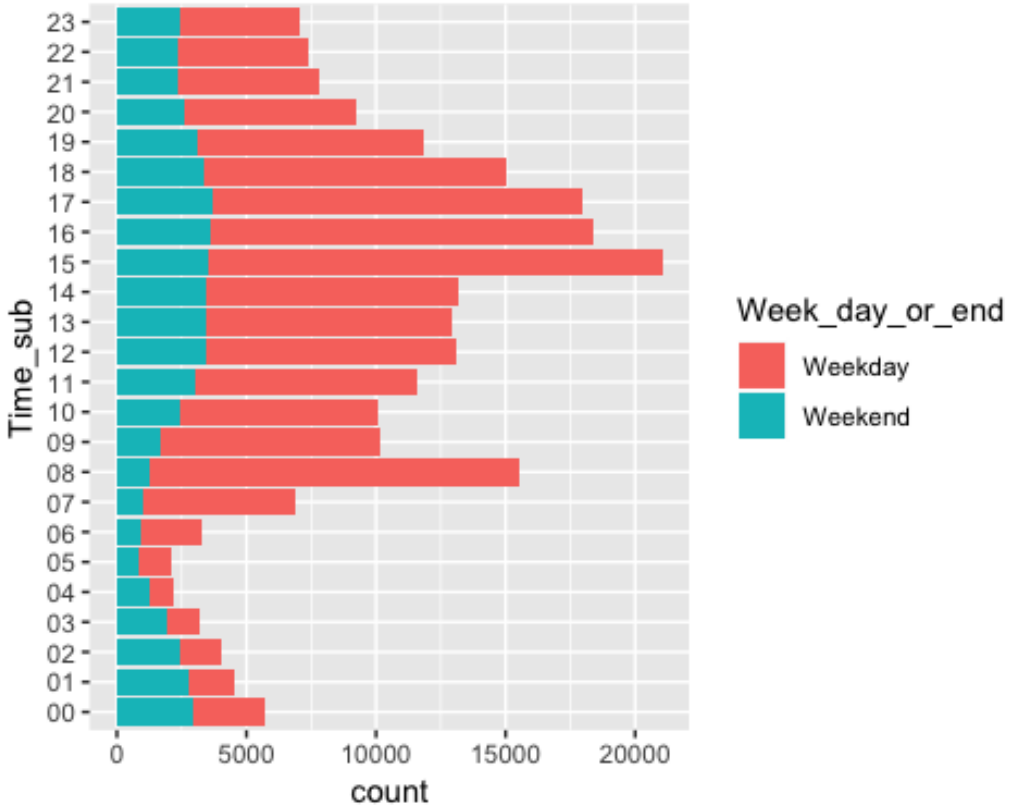
In other words, the weather is not a decisive factor in car accidents, but it is a very important factor. Therefore, choosing strong grip tires to remain traction on a slippery surface will help.

C. Factor 3: Light Conditions & Accidents Happen Across A Day

Furthermore, we predict that lighting conditions may also be a significant factor that will drive road casualties. In common sense, the amount of accidents at night would be larger since low visibility conditions tend to trigger car accidents and drivers are more likely to be tired at night. Therefore, we decide to draw a pie chart to see the proportions of the number of accidents under different light conditions.



According to the above graph , it is clear that most accidents happened in daylight conditions which differ from our prediction. This surprising result might be affected by the assumption that the sample size of driving in daylight is much larger than the sample size of driving in darkness. But we can still conclude that light conditions do not have a significant influence on the amount of accidents. Since light conditions are partially controlled by the sun, we draw a histogram plot to observe the time distribution of accidents during a day.

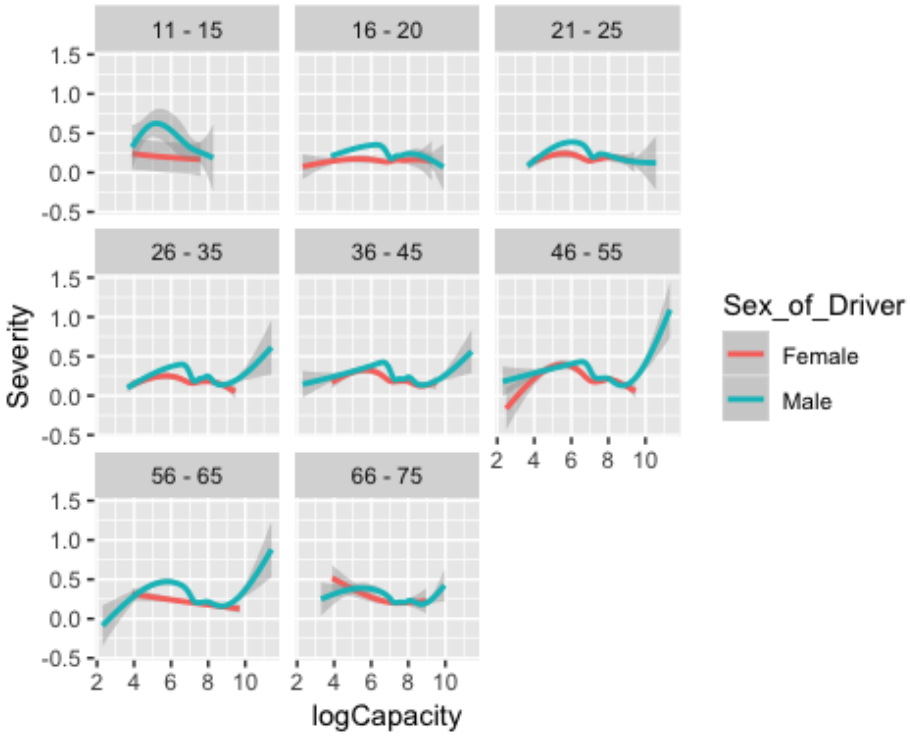


From this plot, the number of accidents is relatively low between 1 am and 6 am when the light condition is insufficient. This fact supports our previous conclusion that light condition has no strong relationship to the number of accidents. We also notice that the number of accidents on weekday is much larger than the number of accidents on weekend. And the distribution of accidents on weekend is much flatter. This information implies that there exist other factors that contribute to road casualties. Heavy traffics could be one of them. In the graph, we find out that the number of accidents happened around 8 am and 3 pm are relatively large. UK drivers may want to avoid these time periods or at pay extra attention when they drive, especially on weekday.

D. Factor 4: Accident Severity

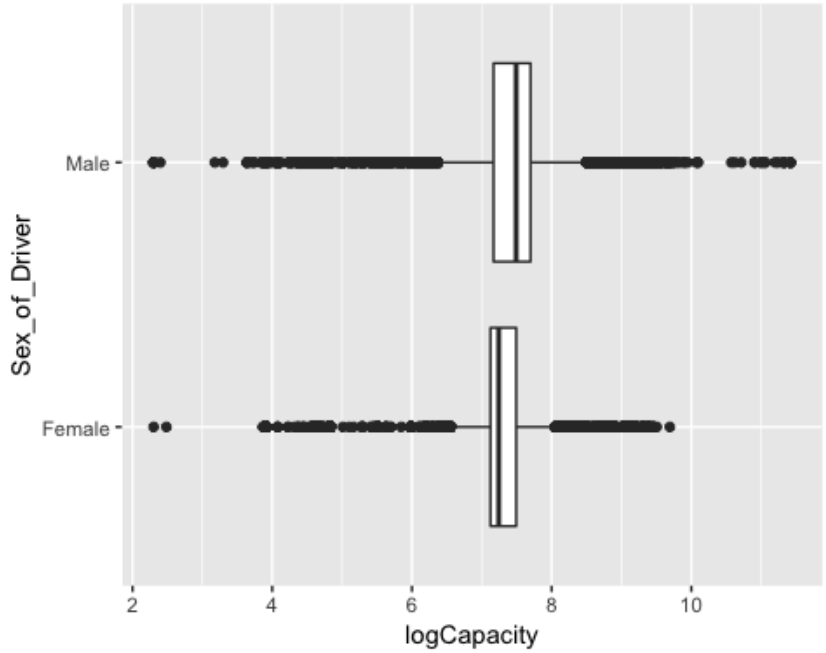
In this part, we studied the factors that may be related to the accident severity and based on that to offer a suggestion on car purchasing. We thought variables `Accident_Severity,Engine_Capacity_.CC.,Age_of_Vehicle,Age_Band_of_Driver,Sex_of_Driver` may be correlated to it. Before any analyzation, we predict that `Engine_Capacity_.CC.` might be positively correlated to `Accident_Severity`, since if we increase the engine capacity, it could increase the max running speed of the car, which could lead to larger damage to the surrounding and the car. Also, the probability of serious accidents might increase if the drive is a female. To verify if our prediction is correct, we start to analyze.

To optimize our result, we did some transformation to the data. Since the magnitude of engine capacity is quite large, we used a log transformation and generate the variable `logCapacity.` Also, In this case, we set `Accident_Severity` as our response which is a categorical data, and we transfer it into numeric data `Severity` and assign 1 to `Serious`and 0 to `Slight`.



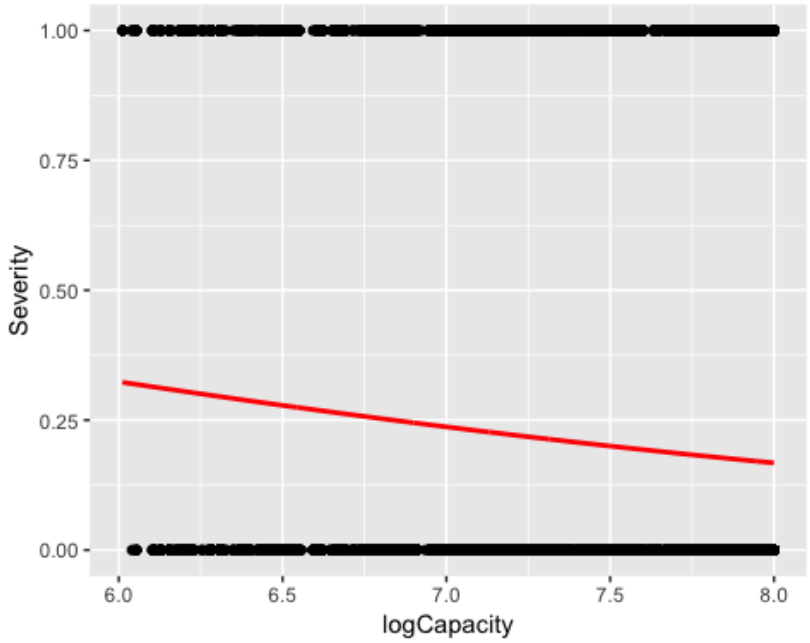
After visualizing our data based on `logCapacity` and `Severity` and divided by the sex of drivers and the age band of them, in this graph, we notice that generally, male is more likely to be related with a serious accident, especially for the male at the age band of 56-65 and when `logCapacity` is between 6 to 8, `logCapacity` and `Severity` seems to be negative correlated. Additionally, we noticed that when `logCapacity` is over 8, the probability of serious accident

increases sharply and most of them are male drivers. To further explore our data, we made a boxplot.



From the boxplot, we discovered that most `logCapacity` is in between (6,8) and we could also observe that for the accident already happened male are more likely to choose a car with higher engine capacity than women. By checking the engine capacity, we notice that when `logCapacity` >=8 it is more likely to be vans, thus, we could say that the number of accidents that male vans drivers involved in have higher than female vans drivers. However, in real life, female van drivers are indeed less than males, it is not accurate enough for us to conclude that female vans drivers are better than males.

Since we are going to give suggestion on car purchasing instead of vans, we then filter out `logCapacity` that is out of (6,8), which is more efficient for us to give advice to drivers. To verify our inference, we built a model. Since our response `Severity` is binary, we need to use the logistic regression. We assume that `Severity` follow the Bernoulli distribution(π_i), $\pi_i=P(y_i=1)$. By using the log-odds ratio, the logit of is $\log(\pi_{i/(}1-\pi_{i)})=\beta_0+\beta_1 X_i$.



From the predicted model, we observed that `Severity` and `logCapacity` is negative correlated, which is opposite to our prediction. As the magnitude of engine capacity increased, the probability of it is a serious accident is decreased.

By checking the P-value, we found out that both intercept and coefficient are significant. Thus, our model is significant.

To optimize our model, we want to analyze how `Severity` effects by the car's engine capacity, `Age_of_Vehicle`,`Age_of_Driver`,`Sex_of_Driver`. We only have the band of the age, thus we generate a new variable `Age_of_Driver` and assign the median of each band to it.

From the result, we found out that `Age_of_Vehicle`,`Age_of_Driver`,`Sex_of_DriverMale` are positive correlated to the our response and `logCapacity` have negative relationship with the `Severity` of the accident. It means by increasing the car engine, we can decrease the severity of the accident and as the car and the drivers get older, the severity of the accident will increase and males are more likely to be involved into serious accidents. However, the age of drivers and vehicles have weak effect on the `Severity`. Also, by considering the P value of each coefficient, we know that all of our variables are significant.

To reduce the damage in an accident, we could choose a car with higher engine capacity. We also observe that males are more likely to be involved in serious accidents than females and as the age of driver and vehicle increased, the probability of involvement in a serious accident also increased.

IV. Conclusion

1. Specific locations where accidents happen frequently should be pay more attention by both UK citizens and the government.
2. Even though the weather condition in raining with no high winds, and wet or damp road surface condition only play the second role in accidents. We still suggest that people choose strong grip tires to remain traction on a slippery surface and slow down in driving.
3. The lighting condition seems do not have a significant influence on the number of accidents. And UK drivers should be careful when they are driving around 8 am and 3 pm.
4. To reduce the damage in an accident, we could choose a car with higher engine capacity.
5. Males are more likely to be involved in serious accidents than females and as the age of driver and vehicle increased, the probability of involvement in a serious accident also increased.

V. Reference:

"UK Weather: Guide to the Seasons."http://www.foreignstudents.com/guide-to-britain/british-culture/weather/seasons

Worldlistmania Contributor. “List of Major Cities in England.” http://www.worldlistmania.com/list-biggest-cities-england/