

14/Aug/2020

Noor Mohamed Abul Kalam Asath

1901711J

**Diploma in Big Data & Analytics**

**Data Science Essentials (CIA2C14)**

**AY2020/2021 April Semester**

**Assignment**

**Submitted by:** <1901711J > <Noor Mohamed Abul Kalam Asath>

**Date:** <2/08/2020 >

**“By submitting this work, I am declaring that I am the originator of this work and that all other original sources used in this work have been appropriately acknowledged.**

**I understand that plagiarism is the act of taking and using the whole or any part of another person’s work and presenting it as my own without proper acknowledgement.**

**I also understand that plagiarism is an academic offence and that disciplinary action will be taken for plagiarism.”**

**Name and Signature of student: <Noor Mohamed Abul Kalam Asath> <Asath>**

***Data Understanding using EDA***

**1. Visualization**

The first chart I’ve plotted using R studio is a bar chart. I used ggplot to plot it. The chart below shows number of peoples involved in an accident based on gender and age. The blue bar represents male drivers and the red bar represent female drivers.

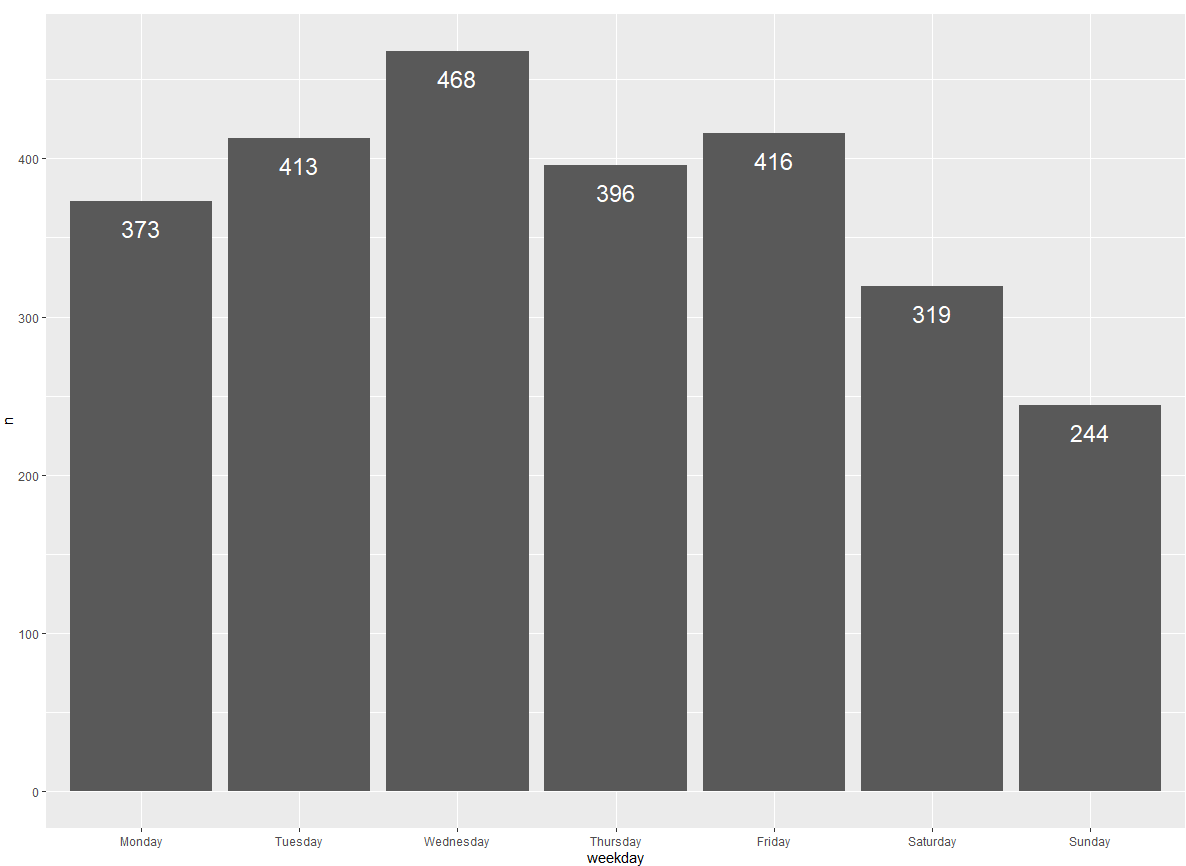
**Takeaway:** As we can observe from the chart, more male then female drivers tend to get into an accident. The age where the number of accidents tends to peak is between the age 25 – 40.



I used count in stat for geom\_bar and I used position dodge for better view of the bars. I used fill for Sex\_of\_Driver.

The chart above shows data of only Drivers who are involved in accident. Pedestrian and passenger class are removed.

**2. Visualization**

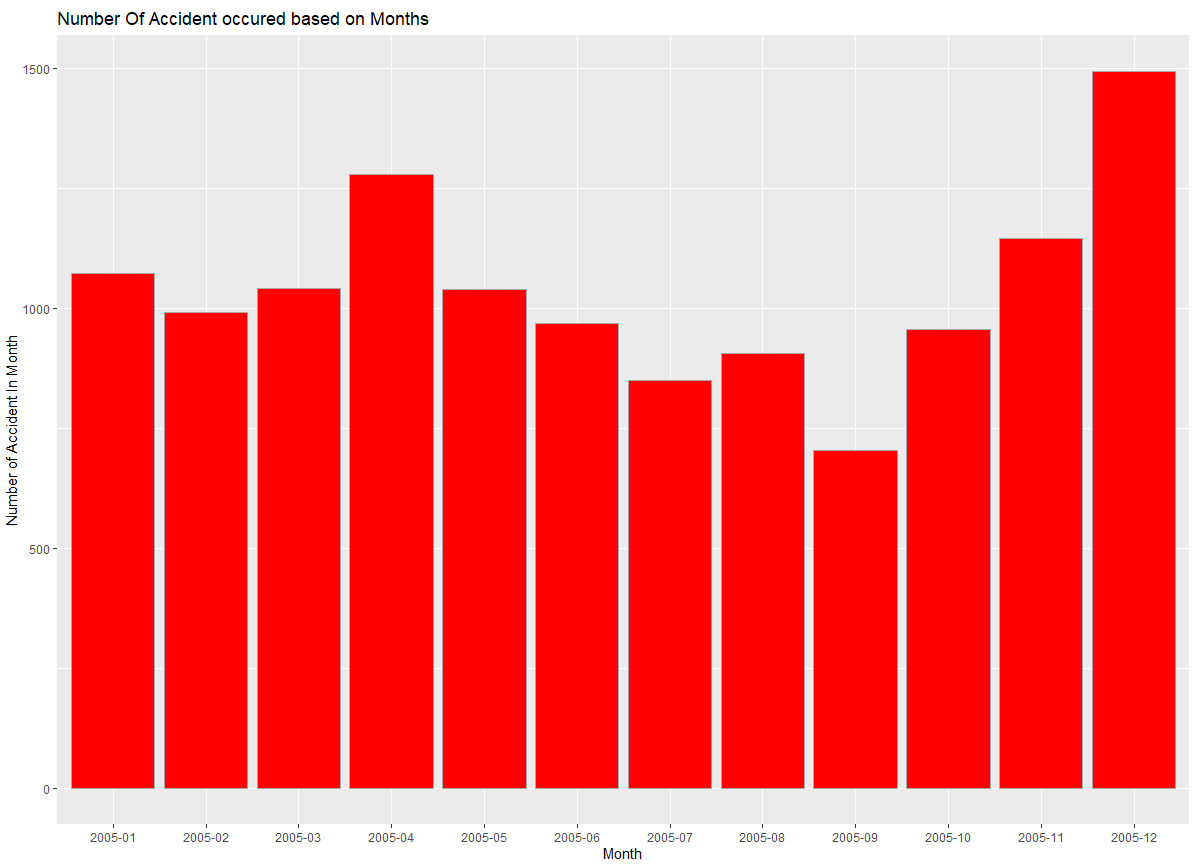


This is my second bar chart it shows the number of accidents that occurred on each day of the week.

**Takeaway:** As we can see from the bar chart, we have more accident on Wednesday then any other day of the week.

From the data, Sunday tend to have the lowest amount of accidents. This can be due to the fact that Sunday is a holiday and peoples tend to spend their time at home.

**3.Visualization**



The above chart represents a histogram which show the number of accidents that occurred on each month. I first created a data frame first called accident\_in\_Day which group\_by Date column to summarise the number of accidents that occurred in a day. Then I created another frame called Accident\_in\_Month which mutates Date format to %Y-%m so I could display my chart based on month.

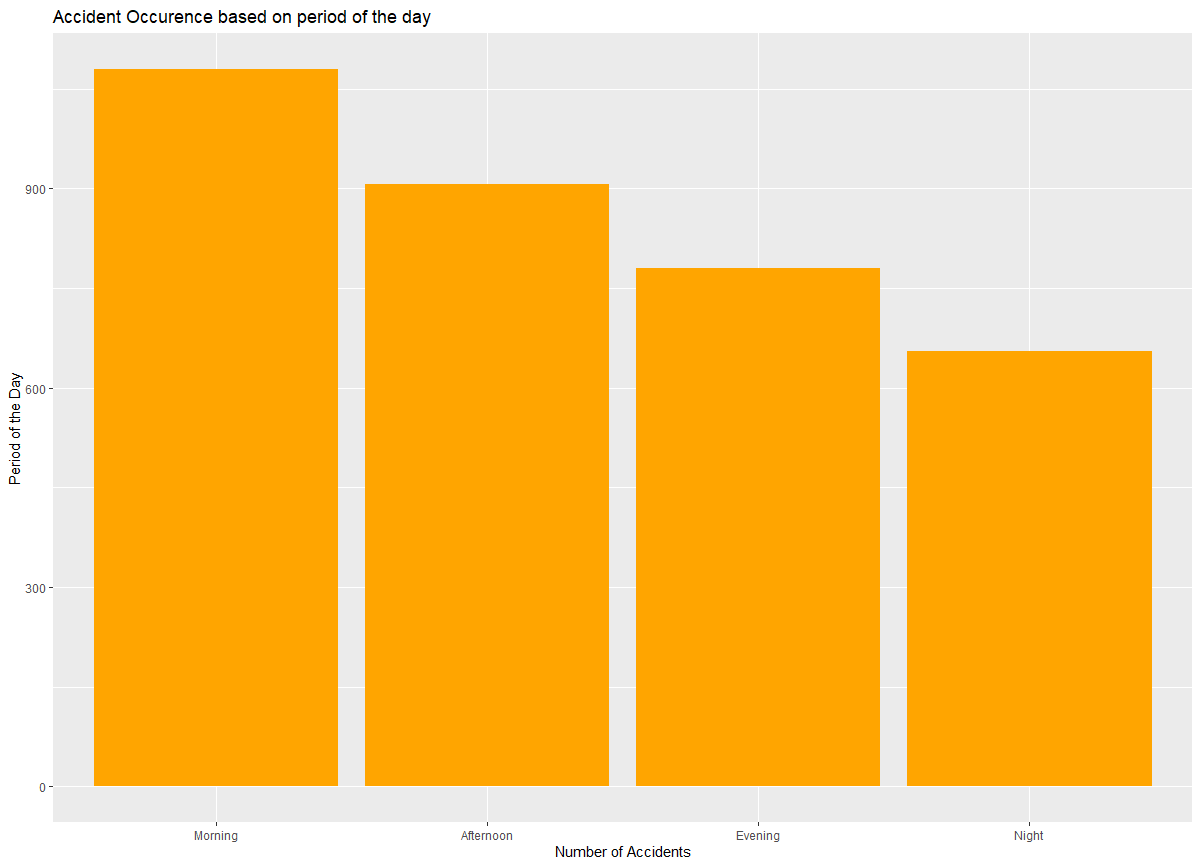
Then I created another data frame to store the number of accidents based on month. I used this data frame to create a histogram. I used identity as stat as each row only represent the total.

**Takeaway:** As we can observe from the chart more accidents happened on December month then any other months. From my personal belief, I think the accident rate is very high as many people went for shopping as it is the end of the year or because since it snows on December and the road is slippery many accident happened on that month.

Between the period of Jun to September (2005-06 to 2005-09) the accident rate decreases. This could be main because it is the summer season and since during summer season the road is dry and less accidents happens.

We can also see a trend were after September from (2005-10 to 2005-12) the rate of accident increase. As I mentioned before it could be because there is a transition from autumn to winter season were the roads are more likely to be wet and slippery.

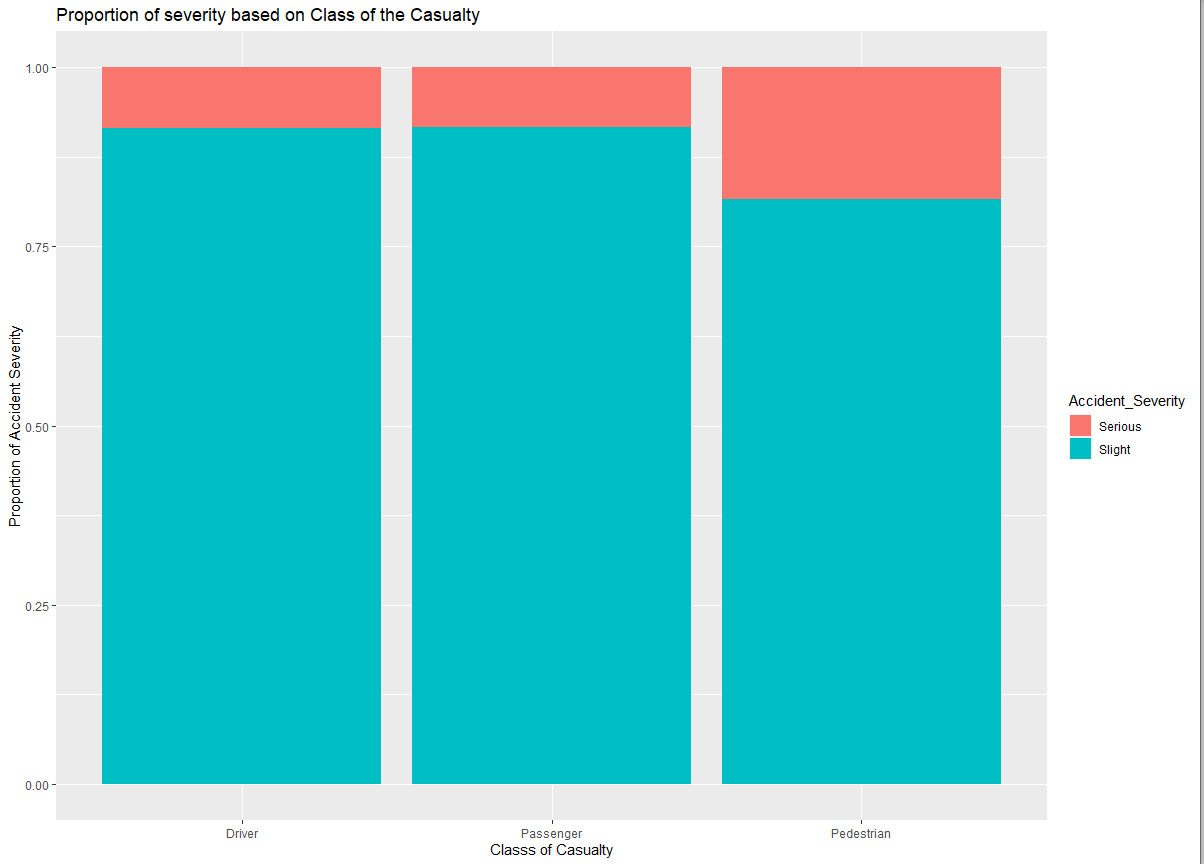
**4. Visualization**

****

This chart shows the occurrence of Accident based on the period of the day. For this chart I first changed the Time attributes type from factor to POSIXct. I then created a column in my data frame called time2 to only store hours as I wanted to do a calculation to divide time based on period of the day. Using time2 column I created another column in my data frame called time\_category which allocates the time based on period of the day. time\_category column has 4 levels Morning, Afternoon, Evening and Night.

**Takeaway:** From the chart above we can see that a lot of accidents tend to happen in the morning, this could be due to the fact that peoples tend to rush to work or places early in the morning. They rush to not get late or miss their appointment.

**5. Visualization**

****

Using x as casualty\_class and accident\_severity as fill I was able to create the bar chart above. I used fill in position to view the proportion of Accident Severity.

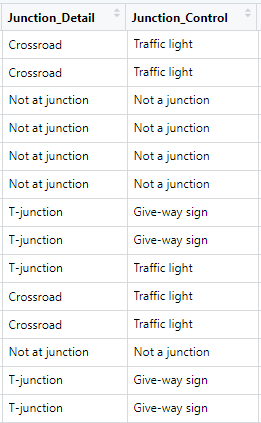
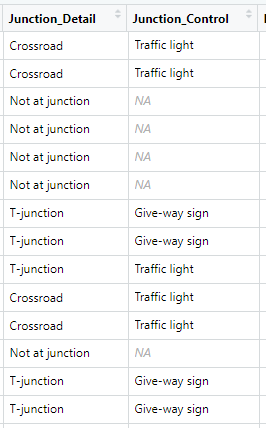
**Takeaway:** We can observe from the chart above that pedestrian tends to have more serious accident compared with driver and passengers.

***Data Preparation***

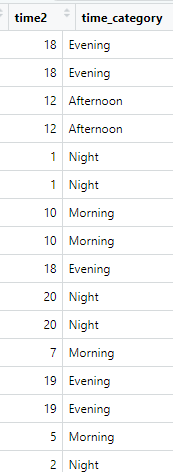
**Data Transformations**

After

Before

******

I changed na value in Junction control to Not at junction as I noticed that it was NA because it was not at junction similar to junction detail.



I created two table in my data frame, one is time2 and another is time\_category time\_category will denote the period of the day based of a calculated condition.

**Correlation Analysis**

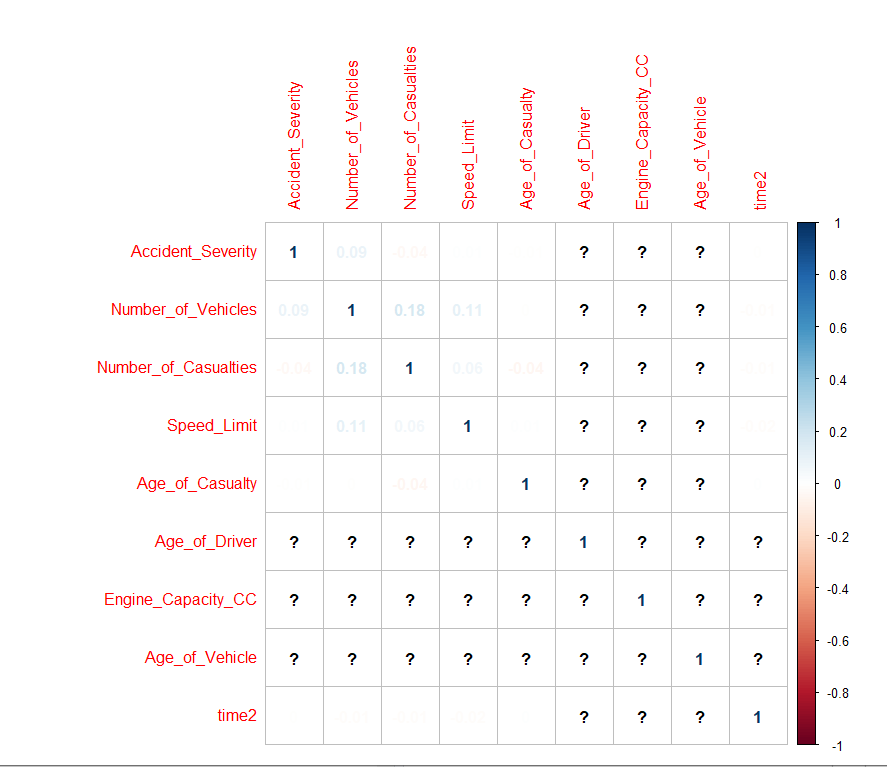
For correlation analysis, you need data with either integer or numeric type. Since most of the data column are nominal and ordinal, I selected columns with int and num with the help of str() function.

With the help of select function in dplyr I run my first model by selecting all the integer and number column. The column I selected for first correlation model are Accident\_Severity,Number\_of\_Vehicle, Number\_of\_Casualties, Speed\_Limit (Changed it to numbers), Age\_of\_Casualty, Age\_of\_Driver,

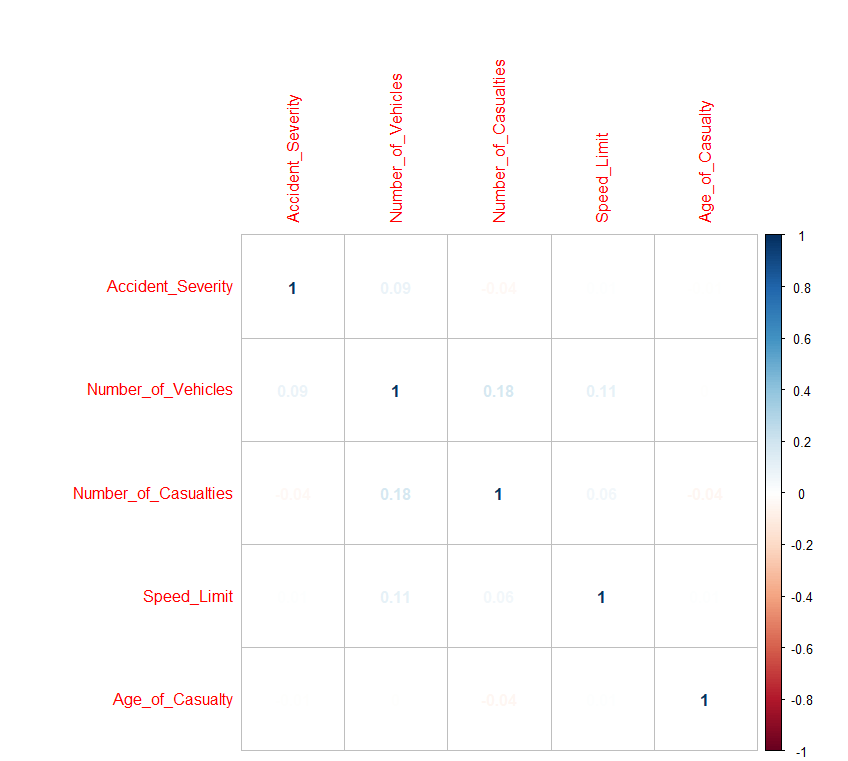
Engine\_Capacity\_CC, Age-of\_Vehicle and finally the time2 column I created.

I installed corrplot package in r and called the library.

Using the selected data frame, I created a correlation coefficient matrix and named it correlation. The I used corrplot package to plot the matrix



From the corrplot we can see that there are lot of ? markes due to missing values in Age\_of\_Driver, Engine\_Capacity\_CC, and Age\_of\_Vehicel variable. Using sum is.na function I was able to see that there are 1040 na values in Age\_of\_Driver variable, there are 2078 na values in Engine\_Capacity\_CC variable and 2181 missing values in Age\_of\_Vehicle variable. I cannot omit all the na values in the dataframe as doing so will delete valuable rows. Hence, I removed these three variables and rerun the correlation coefficient.

After removing the three variable I sum all the na in my data frame and get 0 as value. 

From the corrplot we can observe that there is no correlation between any variable. We cannot fully conclude our data frame based only on corrplot so we must do regression with the variables to confirm if there is any correlation.

***Data Modelling***

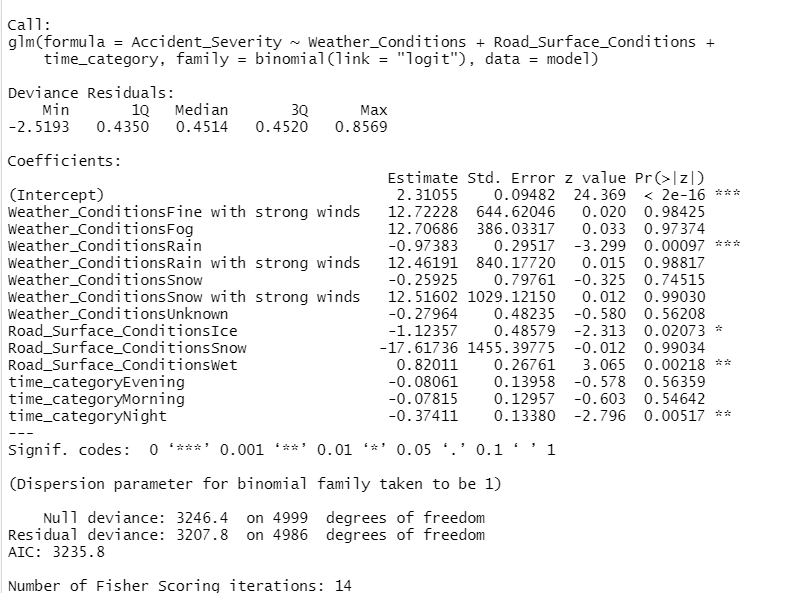
Regression Analysis 1

For my regression model I want to predict if weather condition, road surface condition and time of the day has any effect on accident severity.

The data attributed I choose for data modelling are accident\_severity as dependent variable, Weather\_Condition, Road\_Surface\_Condition, Casualty\_Class and time\_category as independent variable.

I used the Generalized Linear Model function, glm(), to apply the logistic regression model to the attributes. I used binomial family to make the glm() function do logistic regression.

The following is the output:



From the binomial output we can see that Weather-Condition Rain has p value of less then 0.05, this suggest that rain has an influence on the severity of the accident.

Surface-road condition Ice and wet also has p value less then 0.05 which suggestion that if the road surface is Wet or Ice the severity of the accident gets worse. We can see that p-value of time\_category Night also has small p-value. This suggest that if the driver ride on night, the accident tend to be more serious.

From the p-value we can conclude that these 3 attribute are the correct attributes for predicting severity of the accident.

**Model Equation:**

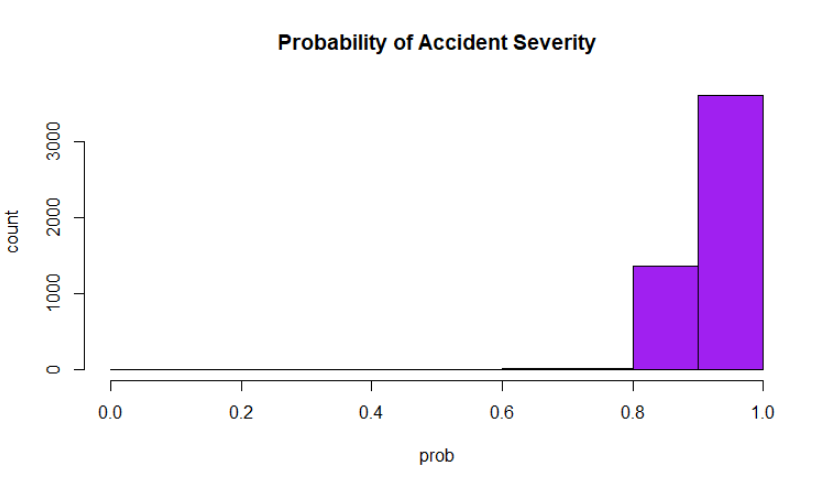
Accident\_Severity = -0.37411\*time\_category Night +0.82011\*Road\_Surface\_Condition wet -0.97383 \*Weather\_ConditionsRain -1.12357 \*Road\_Surface\_ConditionsIce +0.02073

Nighttime accident severity serious

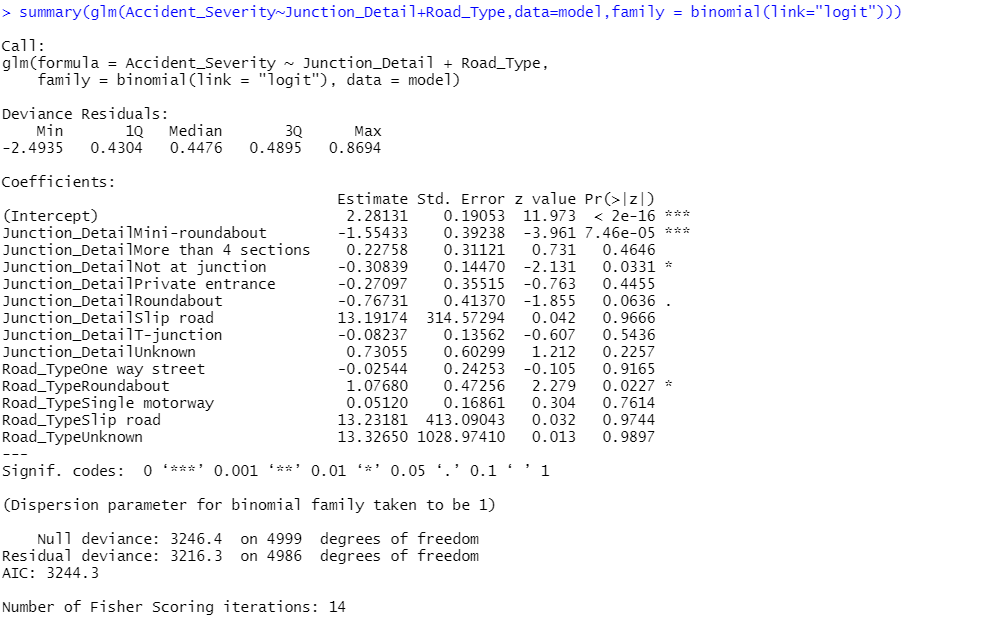
Road surface is wet or Ice accident severity serious

Weather condition is Rain accident severity serious

**This is my fitted result.**



Regression Analysis 2



For the binomial regression above I used Junction\_detail variable and Road\_type variable. we can observe that the severity of the accident drastically increases in mini\_roundabout from considering its P-value which very

low. Road\_Type roundabout also has p-value less then 0.05.

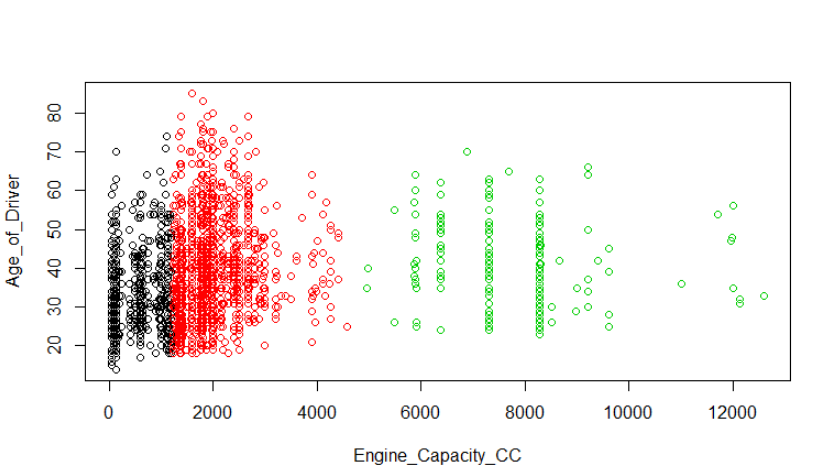
This regression model suggest that roundabouts are dangerous and accident

in roundabout are more likely to be serious.

**Model Equation:**

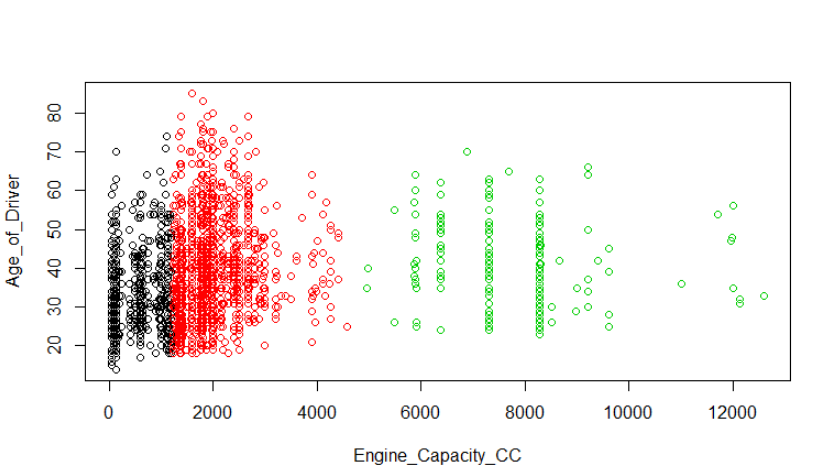
Accident Severity = 1.07\*Road\_TypeRoundabout-1.55433\*Juntion\_DetailMini\_Roundabout+2..28131

**K-mean Cluster**



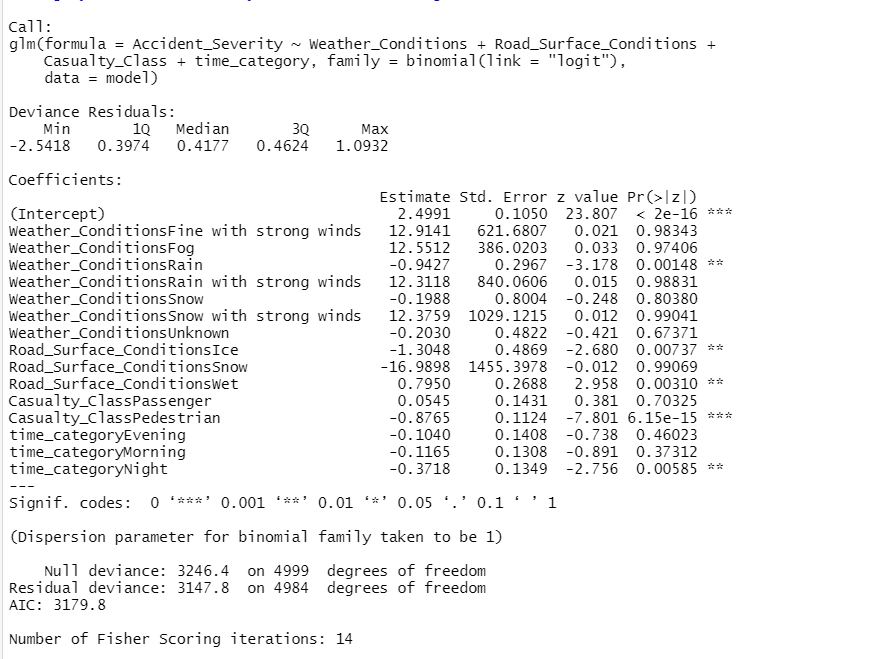
The above k-mean cluster show the cluster based on Engine capacity and Age of driver. We can observe that there are less people who use cars with engine capacity more then 6000 cc. Most of the people in different age group preferred cars with Engine capacity less then 4000CC.

I used 3 groups for K-mean. 2 group falls between 0 to 5000 and the last group covers from 4500 – 12000.

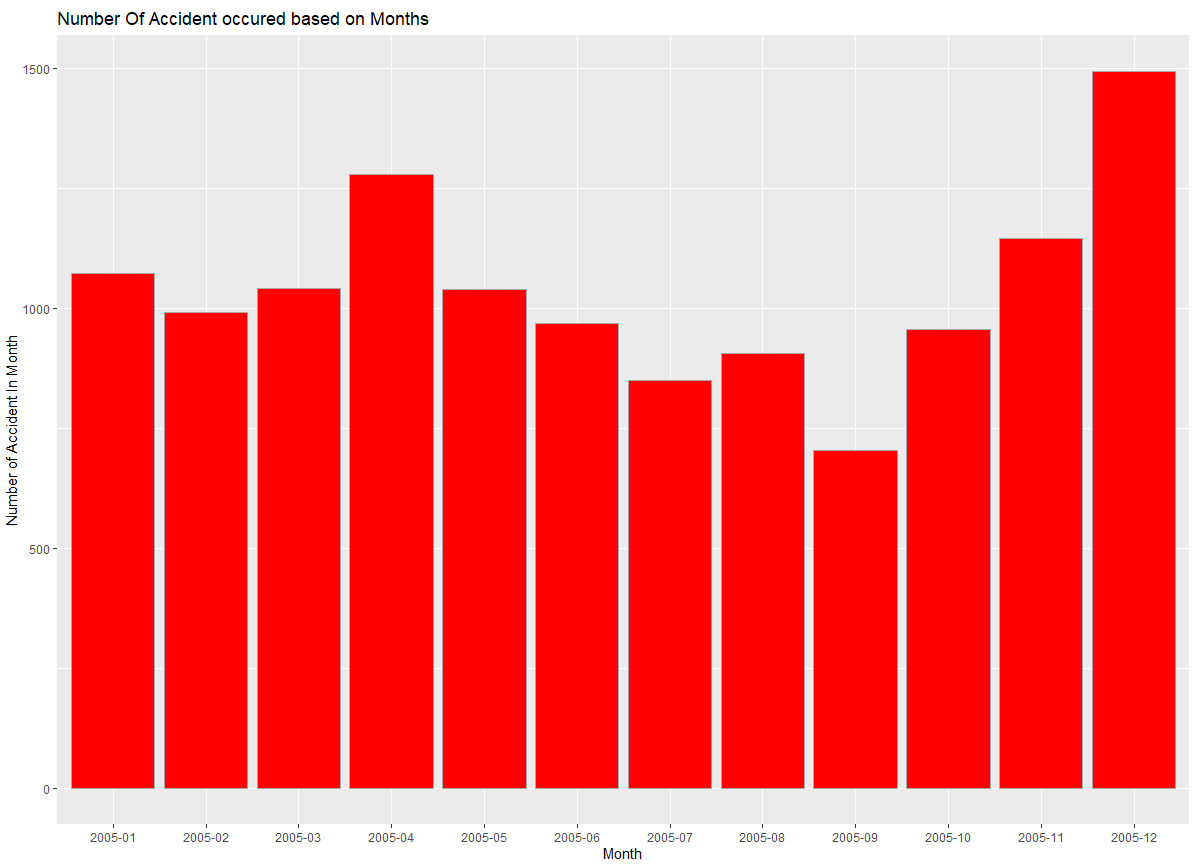
****

**Recommendations:**

1. From my regression model I was able to see that if you drive during night and if the road surface condition is wet, the Accident severity is high. Drivers should be precautious and drive slow during these conditions.



Traffic police department in country x should take measure to prevent accident during this road circumstances.



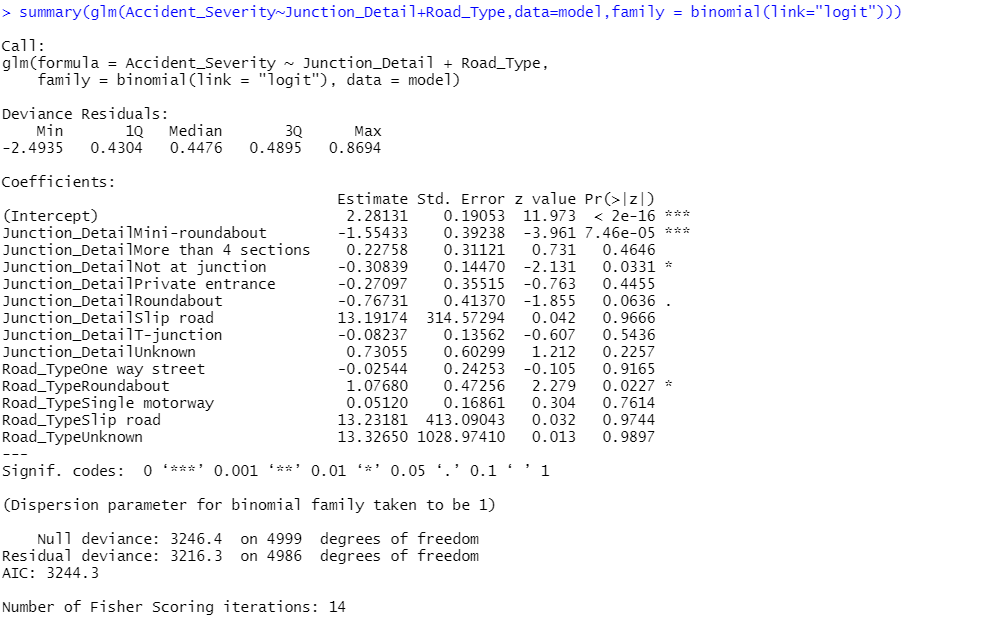
The number of accidents tend to increase in autumn and winter season. This could be due to snow and rain during this period.

2.

We can see from the chart below that young peoples cause most of the accident and it is male driver who cost most accident. Awareness programs should be implemented to warn young peoples to drive safe.



3.



In roundabouts accidents bound to be very serious. More focus should be put on roundabout and mini\_roundabout. Traffic police in country x should consider improving the infrastructure in this roundabout.