DAB501 Project #2

Univariate and Bivariate Analysis

Alejandro Rodriguez (w0795089) and Jenson Jacob (w0794547)

2022-07-20

Contents

Academic Integrity policies statement.

We, Alejandro Rodriguez Orama and Jenson Jacob, hereby state that we have not communicated with or gained information in any way from any person or resource that would violate the College's academic integrity policies, and that all work presented is our own. In addition, we also agree not to share our work in any way, before or after submission, that would violate the College's academic integrity policies.

Academic Integrity at St. Clair College

R and RStudio environment.

- RStudio : 2022.02.3+492
- R : 4.2.0
- Libraries
    - tidyverse
    - knitr
    - plotly
    - ggpubr
    - ggridges

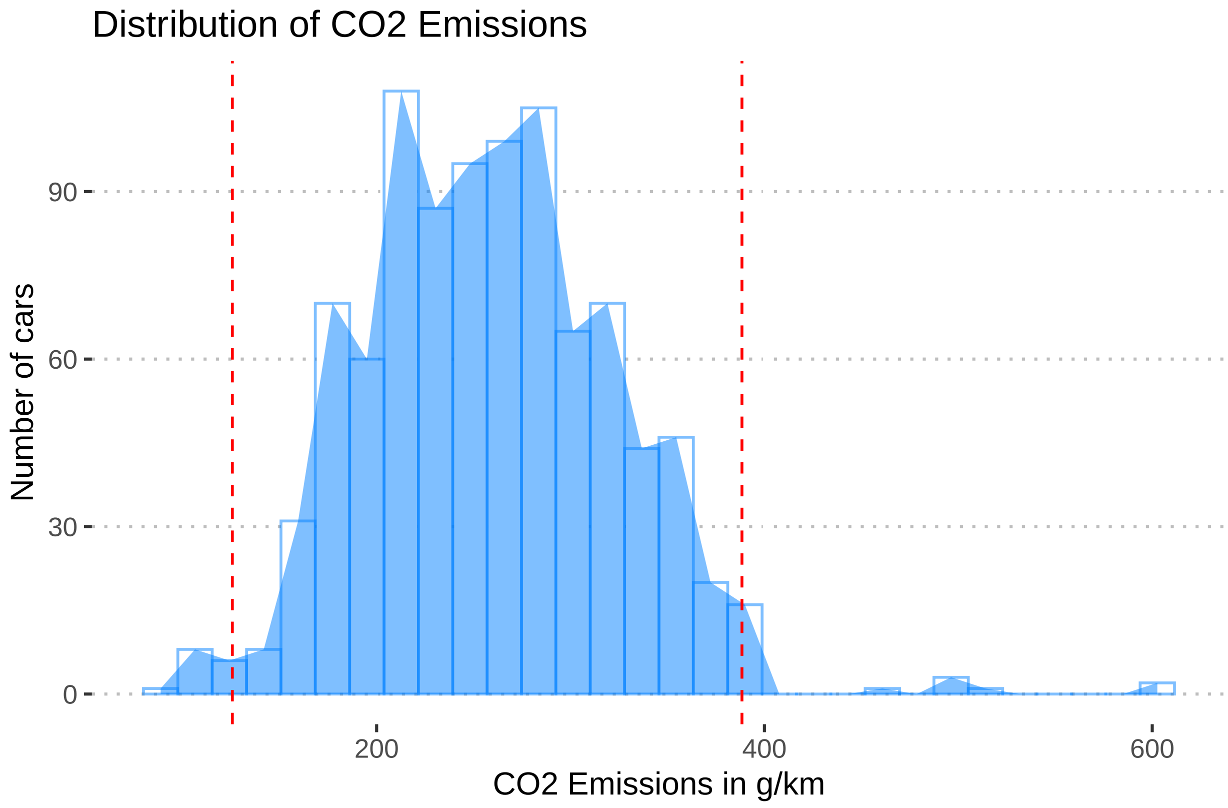
Univariate Analysis.

Analysis of Numeric Variables.

For each numerical variable selected we are going to discuss the following points:
1. Distribution of the variable.
2. Consider any outliers present in the data.
3. Describe the shape and skewness of the distribution.
4. Decide if it is appropriate to apply a transformation to your data.
5. Choose and calculate an appropriate measure of central tendency.
6. Explain why you chose this as your measure of central tendency. Provide supporting evidence for your choice.
7. Choose and calculate a measure of spread that is appropriate for your chosen measure of central tendency. Explain why you chose this as your measure of spread.

Case 1: CO2 Emissions.

```r
p1.1 <- ggplot(Automobile_data, mapping = aes(x = `CO2 Emissions(g/km)`)) +
          geom_histogram(fill = "white",
                         color = rgb(0, 0.5, 1, alpha = 0.5)) +
          geom_area(stat = "bin", fill = rgb(0, 0.5, 1, alpha = 0.5)) +
          labs(x = "CO2 Emissions in g/km",
               y = "Number of cars",
               title = 'Distribution of CO2 Emissions',
               caption = "Automobile Dataset found from Kaggle") +
          geom_vline(xintercept=125.6, linetype="dashed", color = "red") +
          geom_vline(xintercept=388.4, linetype="dashed", color = "red") +
          theme_pubclean()
p1.1
```



Automobile Dataset found from Kaggle

From the above visualization, we believe that this variable has a normal distribution.

Once plotted the numeric variable of interest we are going to consider as a rule of thumb to determine outliers the $1.5 * IQR$ rule. Here is a brief summary of our variable of interest which we obtained using the function "summary".

```
summary(Automobile_data$`CO2 Emissions(g/km)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    94.0   213.2   257.0   259.2   300.8   608.0
```

With these stats, we can calculate the IQR and arrive at the conclusion that all values greater than 388.4 $(257 + 1.5 * (300.8 - 213.2))$ or lower than 125.6 $(257 - 1.5 * (300.8 - 213.2))$ are considered outliers.

After obtaining the boundaries for determine the outliers we can check how many cars in our data have a CO2 Emission that is considered an outlier with the following code:

```
table(Automobile_data$`CO2 Emissions(g/km)` > 388.4 | Automobile_data$`CO2 Emissions(g/km)` < 125.6)
```

```
## 
## FALSE  TRUE 
##   922    24
```

As we can see we have only 24(2.5%) cars considered as outliers by their CO2 emissions, in addition, we want to spot out the fact that the difference between the Mean(259.2) and the Median(257.0) is not a huge

difference so it's mean that the outliers are not affecting too much the mean.
Even though, we are not experts on this topic we can't say with total security that 2 unit points of CO2 Emissions are not a big deal in terms of constituting a bigger hazard to the environment.

Regarding the shape and skewness of the distribution, we can say that even though it is not a perfect Gauss's Bell distribution, it's close to it and it's what usually we going to face in real problems.

We considered that as it is a distribution that visually looks like a normal distribution is not necessary to make any transformation to the data in order to obtain a better bell's shape looks, in addition, it will only increase the complexity of any further analysis.

All that said, we consider that for this particular variable, the Mean is the best option when we must take a measure of central tendency, we have a distribution that looks like a normal distribution and the outliers do not have a big impact on pulling the mean from the median.

The main reason on what we base our decision of selecting the Mean as the best measure of central tendency is the fact that the difference between the Mean and the Median is only +2 (slightly skewed to the right) unit approximately.

We think that for this particular case the best spread measure that we can obtain and get the level of spread of our variable is the standard deviation which has a value of 64.4.

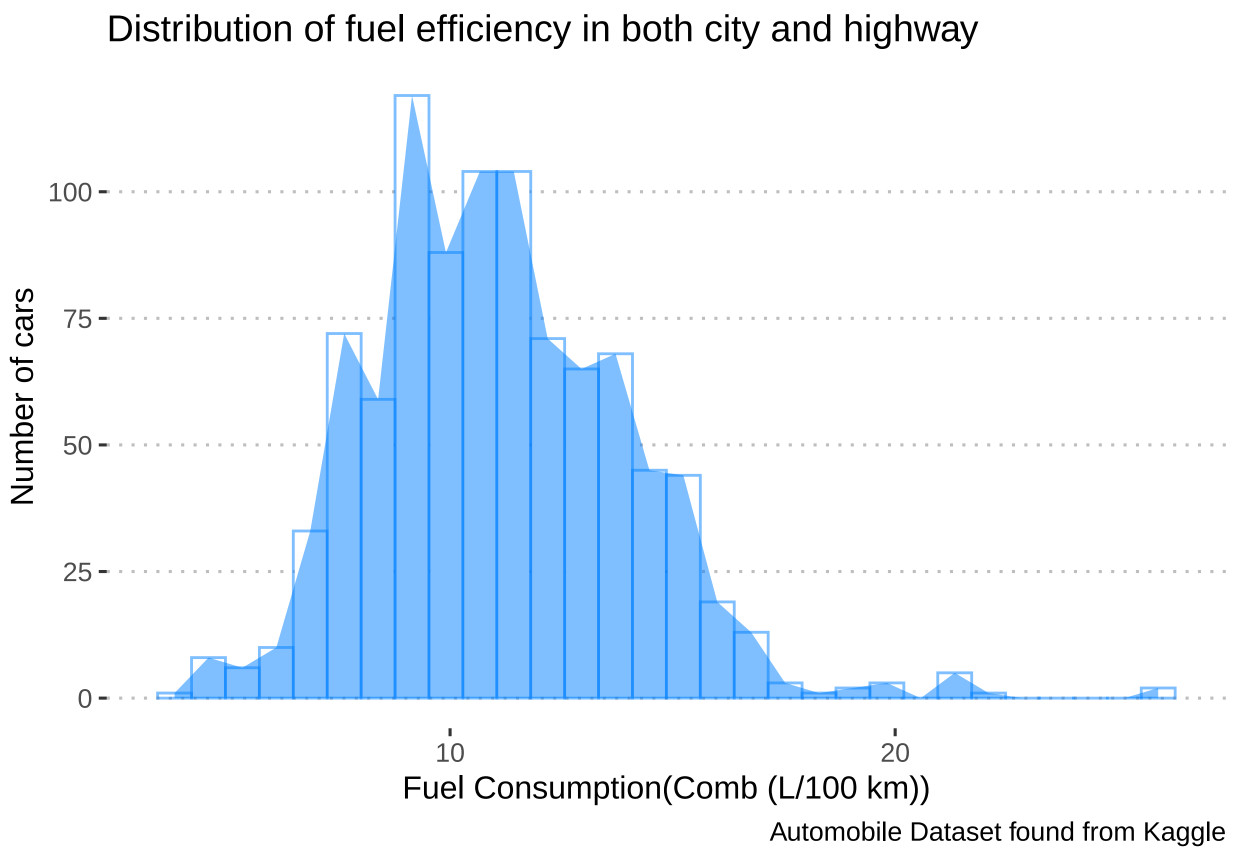
To sum up, we have a distribution that looks like a Normal distribution and a Mean that is not affected by outliers to a high degree. All this combined with the last measure of spread, the standard deviation, helps us to say that 95% of all our data will be in the range 257 ± 128.8 ($Mean \pm 2sd$).

```r
sd(Automobile_data$`CO2 Emissions(g/km)`)
```

```
## [1] 64.44315
```

Case 2: Fuel Consumption in both city and highway.

```r
p1.2 <- ggplot(Automobile_data, mapping = aes(x = `Fuel Consumption(Comb (L/100 km))`)) +
          geom_histogram(fill = "white",
                         color = rgb(0, 0.5, 1, alpha = 0.5)) +
          geom_area(stat = "bin", fill = rgb(0, 0.5, 1, alpha = 0.5)) +
          labs(x = "Fuel Consumption(Comb (L/100 km))",
               y = "Number of cars",
               title = 'Distribution of fuel efficiency in both city and highway',
               caption = "Automobile Dataset found from Kaggle") +
          theme_pubclean()
p1.2
```


Distribution of fuel efficiency in both city and highway

Automobile Dataset found from Kaggle

Following are the responses to the questions on the basis of the above visualization:

Normal distribution of Fuel consumption in both highway and city.

Fuel consumption after the count of 20 are outliers which are majorly found for approximately range of 10 cars count. right 16.5 left 5.1.

The above depicted graph shows right skewness where all the majority outliers are clustered around right tail.

No transformation required as the shape and skewness closely looks like normal distribution.

Mean is the appropriate measure of central tendency for this data.

The reason to choose mean is because both median and mean has 0.29 difference and whenever mean and median is having less difference then its preferable to go with the mean.

Standard deviation is the measure of spread used in this data and the standard deviation is 2.87. Data is pretty close to normal distribution, we can say that 95.45% of the data will be in the range mean - (2* S.D), mean +(2*S.D). This range will give me a perfect idea of the spread of my data.

```
sd(Automobile_data$`Fuel Consumption(Comb (L/100 km))`)
```

```
## [1] 2.876276
```

```
summary(Automobile_data$`Fuel Consumption(Comb (L/100 km))`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    4.00    9.10   10.80   11.09   12.90   26.10
```

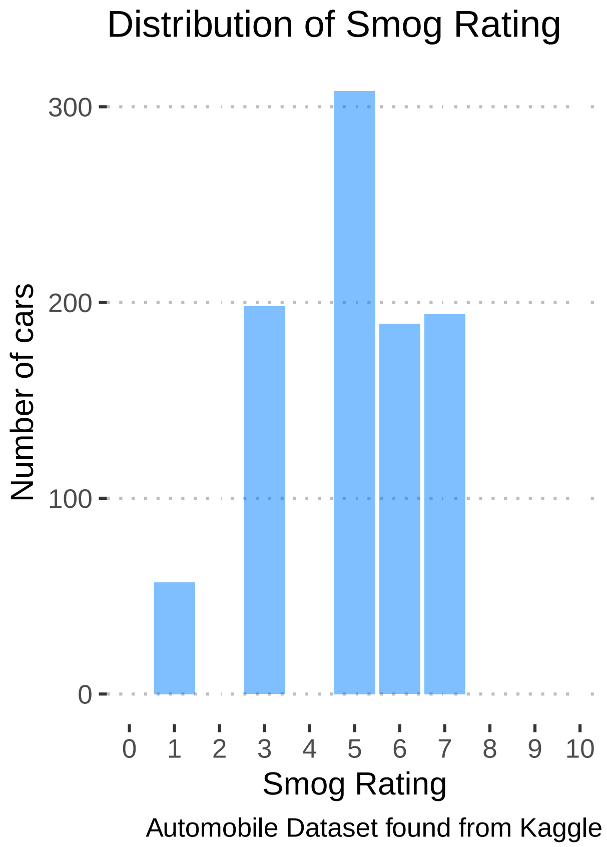
Analysis of Categorical variables.

Similar as we did with the numerical variable, we are going to discuss the next point for each categorical variable that we selected:
1. Plot to visualize the distribution of counts for this variable.
2. Plot to visualize the distribution of proportions for this variable.
3. Discuss any unusual observations for this variable?
4. Discuss if there are too few/too many unique values?

Case 1: Smog Rating.

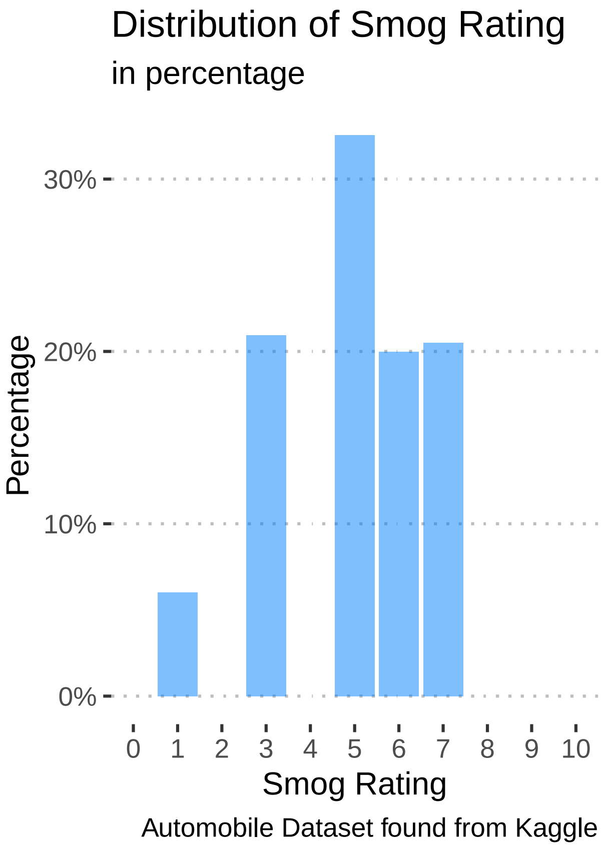
```
p2.1 <- ggplot(Automobile_data, mapping = aes(x = `Smog Rating`, y = )) +
          geom_bar(fill = rgb(0, 0.5, 1, alpha = 0.5)) +
          labs(x = "Smog Rating",
               y = "Number of cars",
               title = 'Distribution of Smog Rating',
               caption = "Automobile Dataset found from Kaggle") +
          scale_x_continuous(breaks=seq(0,10,by=1), limits = c(0, 10)) +
          theme_pubclean()
```

```
p2.1.1 <- ggplot(Automobile_data, mapping = aes(x = `Smog Rating`, y = stat(prop), group = 1 )) +
          geom_bar(fill = rgb(0, 0.5, 1, alpha = 0.5)) +
          labs(x = "Smog Rating",
               y = "Percentage",
               title = 'Distribution of Smog Rating',
               subtitle = "in percentage",
               caption = "Automobile Dataset found from Kaggle") +
          scale_x_continuous(breaks=seq(0,10,by=1), limits = c(0, 10)) +
          scale_y_continuous(labels = scales::percent) +
          theme_pubclean()
grid.arrange(p2.1, p2.1.1, ncol=2)
```



Distribution of Smog Rating

Automobile Dataset found from Kaggle



Distribution of Smog Rating
in percentage

Automobile Dataset found from Kaggle

Regarding the previous charts, we would like to talk about the fact that we were expecting a more distributed variable along the 10 categories of rating. We didn't expect that approximately 95% of the cars got ratings from 3 to 7, which means that there are no cars that have too good smog ratings or too bad smog ratings, at least in this sample dataset.

In terms of unique values, we think that this variable has a midterm of unique values, not a few but they are not a lot as well.

Case 2: CO2 Rating.

We have used categorical variables in the following visualization to explore the questions asked:

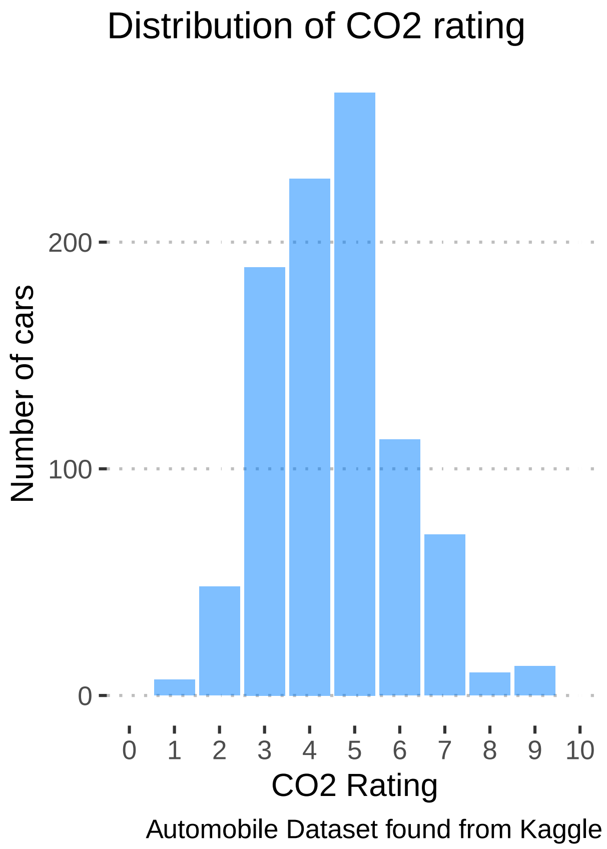
1. Plot to visualize the distribution of counts for this variable

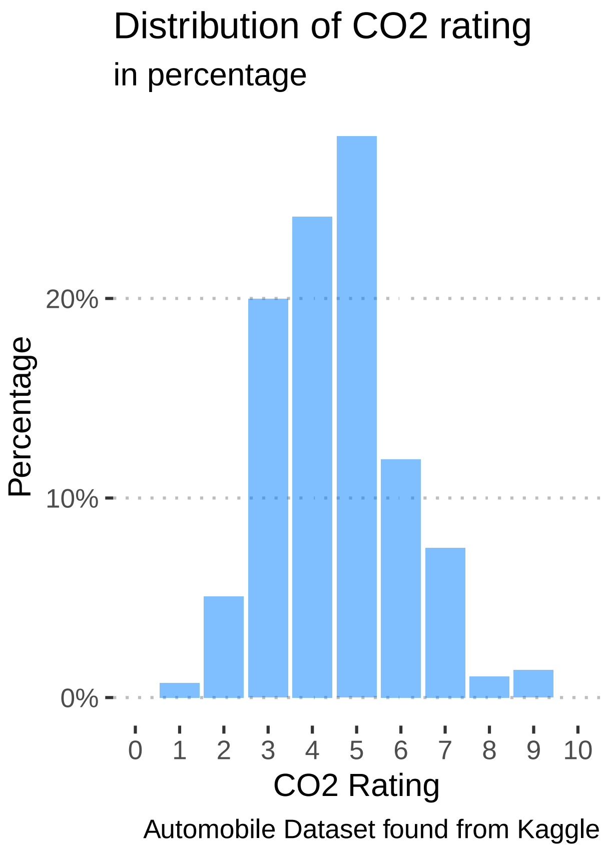
```
p2.2 <- ggplot(Automobile_data, mapping = aes(x = `CO2 Rating`)) +
          geom_bar(fill = rgb(0, 0.5, 1, alpha = 0.5))+
          labs(x = "CO2 Rating",
               y = "Number of cars",
               title = 'Distribution of CO2 rating',
               caption = "Automobile Dataset found from Kaggle") +
  scale_x_continuous(breaks=seq(0,10,by=1), limits = c(0, 10)) +
          theme_pubclean()
```

2. Plot to visualize the distribution of proportions for this variable

```
p2.2.1 <- ggplot(Automobile_data, mapping = aes(x = `CO2 Rating`, y = stat(prop), group = 1)) +
          geom_bar(fill = rgb(0, 0.5, 1, alpha = 0.5))+
          labs(x = "CO2 Rating",
               y = "Percentage",
               title = 'Distribution of CO2 rating',
               subtitle = "in percentage",
               caption = "Automobile Dataset found from Kaggle") +
          scale_x_continuous(breaks=seq(0,10,by=1), limits = c(0, 10)) +
          scale_y_continuous(labels = scales::percent) +
          theme_pubclean()

grid.arrange(p2.2, p2.2.1, ncol=2)
```





Very few cars are having good CO2 ratings and also most of the cars fall under 3-7 rating category. We have moderate data so we don't have too few/too many unique values.

Bivariate Analysis.

Analysis of pair of numeric variables.

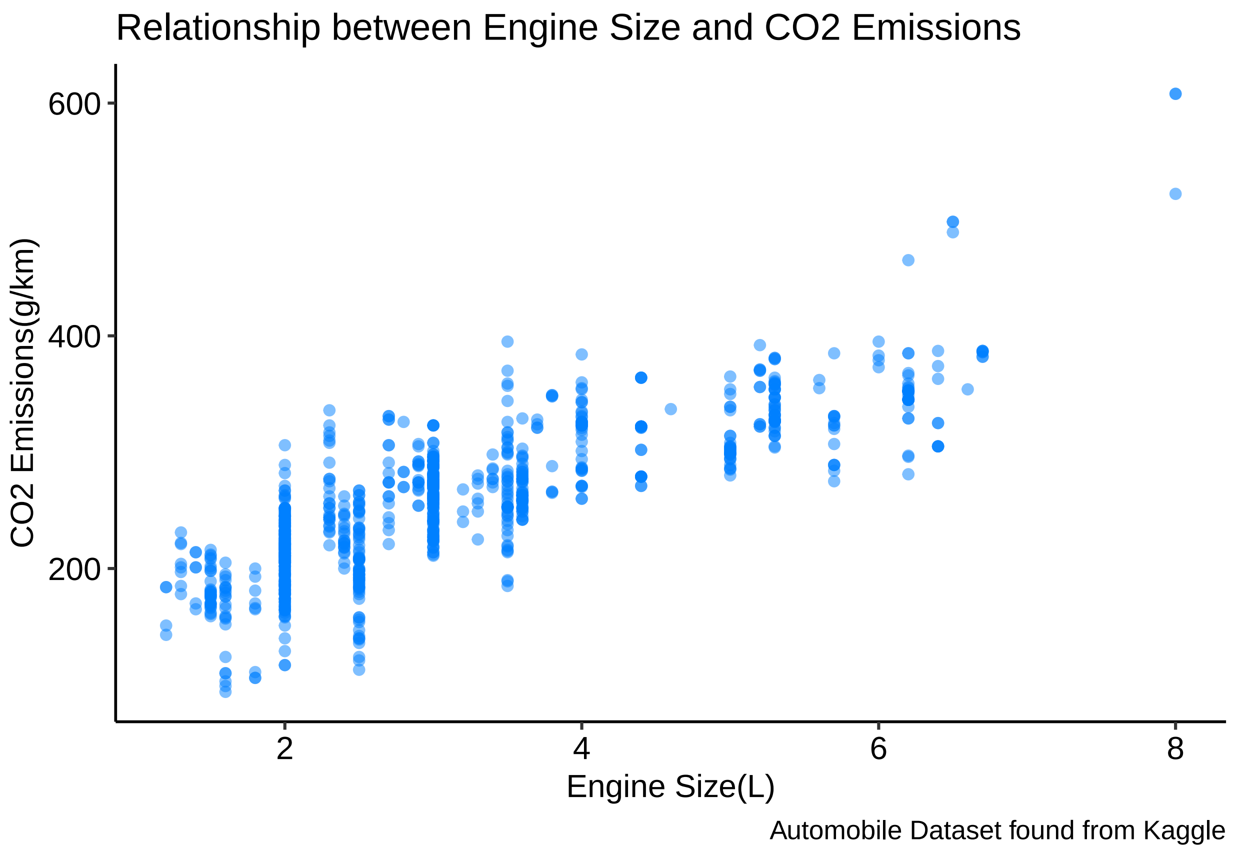
For the next analysis we are going to select one pair of variables, where both are numeric and developing the following points:
1. Plot to visualize the relationship between the two variables.
2. Describe the form, direction, and strength of the observed relationship. Include both qualitative and quantitative measures, as appropriate.
3. Explain what this relationship means in the context of the data.
4. Describe the variability that you observe in the plot and how that corresponds to the strength you calculated in #2 above.

Case 1: Engine Size and CO2 Emissions.

The following graph reflect the relation of how is the CO2 emission regarding the engine size of the car.

```r
p3.1 <- ggplot(Automobile_data, mapping = aes(x = `Engine Size(L)`,y =`CO2 Emissions(g/km)`)) +
          geom_point(color = rgb(0, 0.5, 1, alpha = 0.5)) +
          labs(x = "Engine Size(L)",            # adding the labels of the axis
               y = "CO2 Emissions(g/km)",
               title = 'Relationship between Engine Size and CO2 Emissions',
               caption = "Automobile Dataset found from Kaggle")

p3.1
```



At a simple glance, we can see that the relation among that variable is a positive linear relationship, given the fact that as the engine size of the car is bigger the trend is that the CO2 emission will increase too.

The function "cor()" allows us to check that effectively as we had thought with the help of the visualization, there is a strong relationship between both variable because we got a value of 0.82 which mean that exists a high level of correlation.

```
cor(Automobile_data$`Engine Size(L)`, Automobile_data$`CO2 Emissions(g/km)`)
```

```
## [1] 0.8241876
```

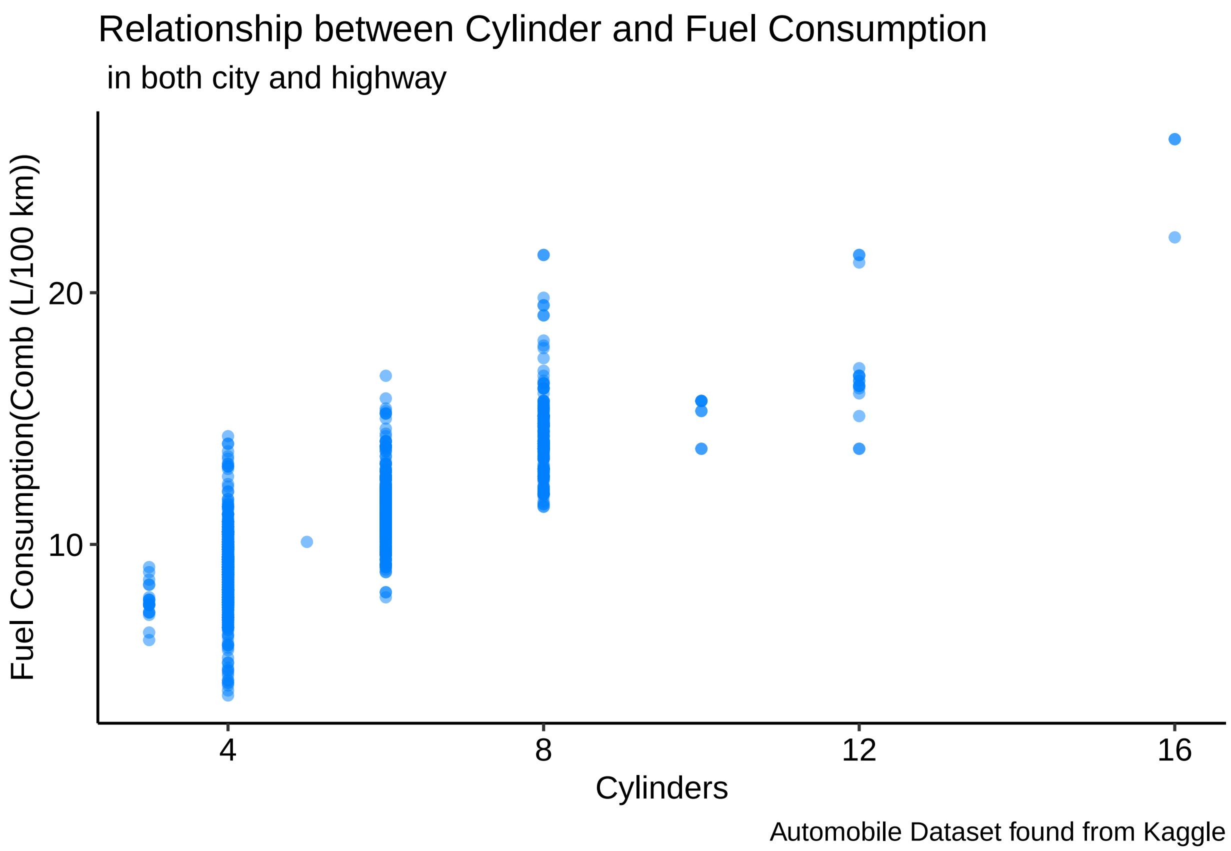
Another detail that we would like to bring out is the fact that the range of CO2 emission varies inside cars with the same engine size, for example, cars with a 2L of engine size can have a level of CO2 emission that ranges from 180g/km to 300g/km and the same trend is present along different engine sizes.

Case 2: Cylinders and Fuel Consumption.

Following graph depicts the pair of numeric variables

Q.1.plot to visualize the relationship between the two variables

```
p3.2 <- ggplot(Automobile_data, mapping = aes(x = Cylinders, y =`Fuel Consumption(Comb (L/100 km))`)) +
          geom_point(color = rgb(0, 0.5, 1, alpha = 0.5)) +
          labs(x = "Cylinders",                                   # adding the labels of the axis
               y = "Fuel Consumption(Comb (L/100 km))",
               title = 'Relationship between Cylinder and Fuel Consumption',
               subtitle = " in both city and highway",
               caption = "Automobile Dataset found from Kaggle")
p3.2
```



This is a linear form and in positive direction with a correlation value of 0.8217181.

```
cor(Automobile_data$Cylinders,Automobile_data$`Fuel Consumption(Comb (L/100 km))`)
```

```
## [1] 0.8217181
```

From the above visualization, we can understand that cars with more cylinders consume more fuel. One more notable thing is that similar to what we saw with the relation between Engine size and CO2 emission is happening in this relation too.
We can see that cars with 4 cylinders can be the ones with the best fuel consumption, but the range of fuel consumption among cars with 4 cylinders is so large that some cars with 8 cylinders can have a better performance in terms of fuel consumption. This situation is observable along the different number of cylinders and gives us an idea of the variability of fuel consumption inside each number of cylinders and in general.

Analysis of Pair of numeric and categorical variable.

Finaly we are going to select one pair of variables, where one variable is categorical and the other is numeric and create and explain the same previous points:

Case 1: Engine Size given by Transmission type.

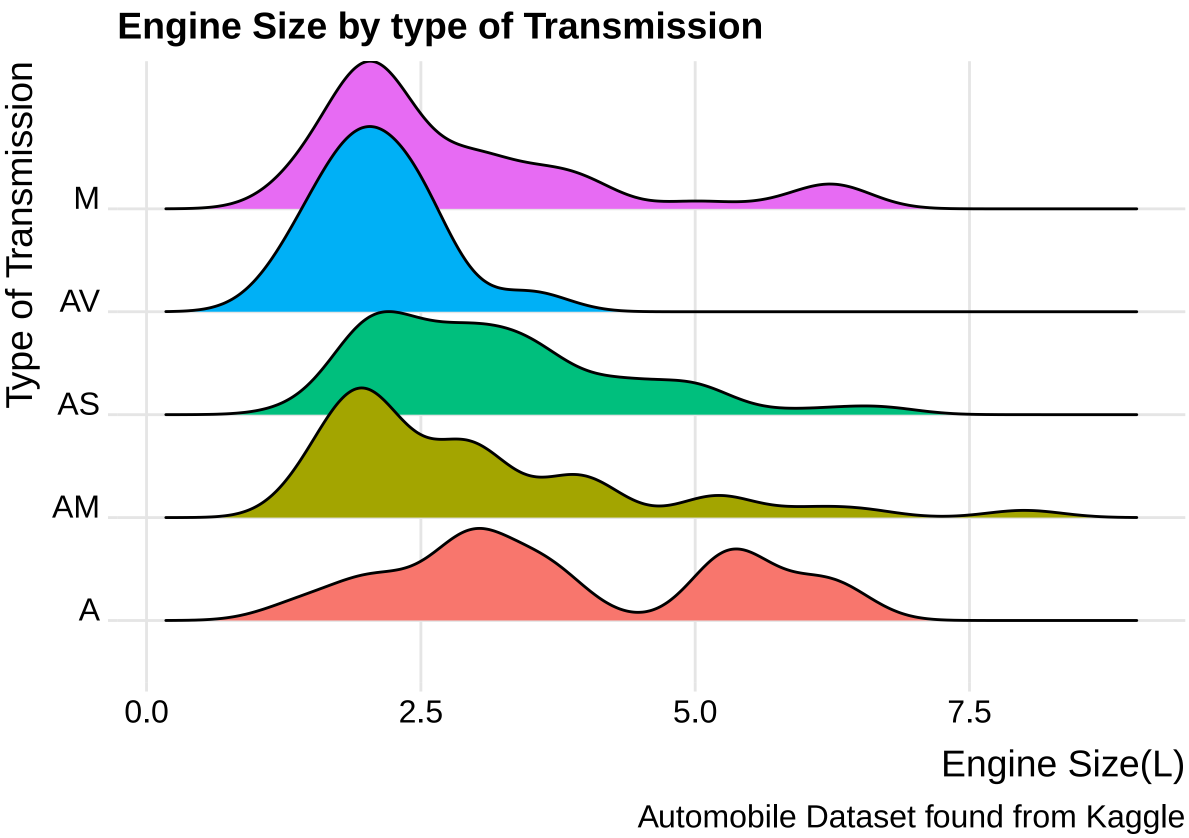
Before entering in full the analysis of the relation between the variables Engine Size and type of transmission we would like to explain that the original data came with different kinds of transmissions but also inside each kind of transmission there are different types given by the number of gears.

For this analysis we decided to group the type of transmission independently from the number of gears, so from a categorical variable of more than 15 values, we did our analysis with a new transmission variable with only 5 categories.
(Transmission: A = automatic; AM = automated manual; AS = automatic with select shift; AV = continuously variable; M = manual)

```
Automobile_data = Automobile_data %>%
  mutate(New.Transmission = case_when(
    Transmission %in% c("A10", "A6", "A7", "A8", "A9") ~ "A",
    Transmission %in% c("AM6", "AM7", "AM8") ~ "AM",
    Transmission %in% c("AS10", "AS5", "AS6", "AS7", "AS8", "AS9") ~ "AS",
    Transmission %in% c("M5", "M6", "M7") ~ "M",
    TRUE ~ "AV"
  ))
#table(Automobile_data$Transmission, Automobile_data$New.Transmission, exclude = NULL)


# Transmission: A = automatic; AM = automated manual; AS = automatic with select shift; AV = continuous
p4.1 <- ggplot(Automobile_data, mapping = aes(x = `Engine Size(L)`, y = New.Transmission,
                                            fill = New.Transmission)) +
       geom_density_ridges() +
       labs(x = "Engine Size(L)",
            y = "Type of Transmission",
            title = 'Engine Size by type of Transmission',
            caption = "Automobile Dataset found from Kaggle") +
       theme_ridges() +
       theme(legend.position = "none")

p4.1
```


Engine Size by type of Transmission

Automobile Dataset found from Kaggle

The chart showed us that there are differences in the engine size of a car given by the type of transmission. One of the conclusions that we arrived at and where the magic of the data science become a fact is that even though we are not expert in the cars industry we know that Manual transmission cars have fewer parts incorporated into the engine so the driver has the control of more features, and with this chart, we can observe that as we move from Automatic transmission to Manually transmission those distributions that look like mountains are moving slowly to the left, which means that the Mean and Median are decreasing so the engine size of the cars is smaller in general.
In the next table, we can see the Mean and Median for each type of car transmission and it can also help us to depict what we are explaining.

```
Automobile_data %>%
  group_by(New.Transmission) %>%
  summarise(Engine.Size_Mean = mean(`Engine Size(L)`),
            Engine.Size_Median = mean(`Engine Size(L)`))
```

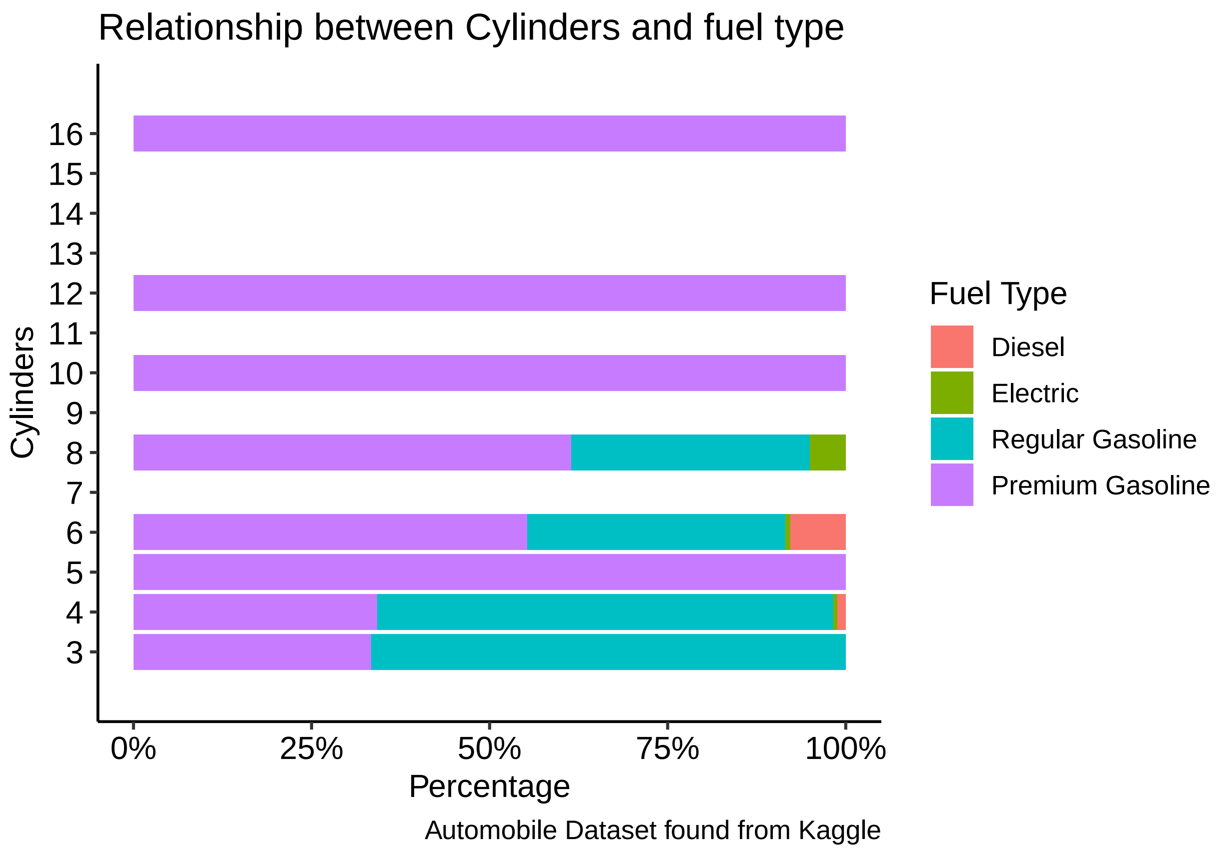
```
## # A tibble: 5 x 3
##   New.Transmission Engine.Size_Mean Engine.Size_Median
##   <chr>                       <dbl>              <dbl>
## 1 A                            3.85               3.85
## 2 AM                           3                  3
## 3 AS                           3.18               3.18
## 4 AV                           2.10               2.10
## 5 M                            2.82               2.82
```

Case 2: Number of Cylinders given by Fuel Type.

Following graph depicts the pair of numeric and categorical variables

Q.1. plot to visualize the relationship between the two variables

```
p4.2 <- ggplot(Automobile_data) +
          geom_bar(mapping = aes(x = Cylinders, fill = `Fuel Type`), position = "fill") +
          scale_y_continuous(labels = scales::percent) +
          labs(x = "Cylinders",
               y = "Percentage",
               title = 'Relationship between Cylinders and fuel type',
               caption = "Automobile Dataset found from Kaggle") +
          scale_fill_discrete(labels=c('Diesel', 'Electric', "Regular Gasoline", "Premium Gasoline" )) +
          scale_x_continuous(breaks=seq(3,16,by=1), limits = c(2, 17)) +
          theme(legend.position = "right") +
          coord_flip()
p4.2
```



Resulted interesting how the cars with more than 8 cylinders all work only with premium gasoline as fuel type, while cars with 8 or fewer cylinders look like the type of fuel has more variability.
From the visualized data we came to know that as the number of cylinders increases fuel selection is less, 4 cylinder engines have variety of fuel options whereas cylinders ranging from 10-16 has only premium gasoline option.

References

We referred the following websites for enhancing our knowledge in this project.

-R for Data Science
-ggplot2 - Tidyverse
-The R Graph Galery
-STHDA