Analysis of CO2_Emissions_Vehicles in the USA

Aaniyah Garner

2024-09-10

Loading packages

```
library(readr)
library(tidyverse)
library(openintro)
library(infer)
library(GGally)
```

Loading/Exploring the dataset

```
CO2 <- read_csv("CO2_Emissions_Vehicles.csv")
glimpse(CO2)
```

```
## Rows: 7,385
## Columns: 12
## $ Make                                <chr> "ACURA", "ACURA", "ACURA", "ACURA",~
## $ Model                               <chr> "ILX", "ILX", "ILX HYBRID", "MDX 4W~
## $ `Vehicle Class`                     <chr> "COMPACT", "COMPACT", "COMPACT", "S~
## $ `Engine Size(L)`                    <dbl> 2.0, 2.4, 1.5, 3.5, 3.5, 3.5, 3.5, ~
## $ Cylinders                           <dbl> 4, 4, 4, 6, 6, 6, 6, 6, 6, 4, 4, 6,~
## $ Transmission                        <chr> "AS5", "M6", "AV7", "AS6", "AS6", "~
## $ `Fuel Type`                         <chr> "Z", "Z", "Z", "Z", "Z", "Z", "Z", ~
## $ `Fuel Consumption City (L/100 km)`  <dbl> 9.9, 11.2, 6.0, 12.7, 12.1, 11.9, 1~
## $ `Fuel Consumption Hwy (L/100 km)`   <dbl> 6.7, 7.7, 5.8, 9.1, 8.7, 7.7, 8.1, ~
## $ `Fuel Consumption Comb (L/100 km)`  <dbl> 8.5, 9.6, 5.9, 11.1, 10.6, 10.0, 10~
## $ `Fuel Consumption Comb (mpg)`       <dbl> 33, 29, 48, 25, 27, 28, 28, 25, 24,~
## $ `CO2 Emissions(g/km)`               <dbl> 196, 221, 136, 255, 244, 230, 232, ~
```

a. How many cases (instances/rows) are in your dataset? 7385 rows

b. How many variables (attributes/columns) are in your dataset? 12 columns

c. Does your dataset contain missing values? Which variables contain missing values? No missing Data

```
anyNA(CO2)
```

```
## [1] FALSE
```

Introduction

This data set captures the details of how CO2 emissions by a vehicle can vary with the different features. The data set has been taken from Canada Government official open data website. This is a compiled version. This contains data over a period of 7 years. There is total 7385 rows and 12 columns. There are few

abbreviations that has been used to describe the features. I am listing them out here. The same can be found in the Data Description sheet.

Description: Amount of CO2 emissions by a vehicle depending on their various features

Research Question and Hypotheses

Research Question

Identify at least two potential research questions that you plan to answer using your data set:

1) Does fuel type determine how much CO2 is emitted by driving?
2) Can the CO2 emitted by a car be predicted by make, model, class, engine size, cylinders, transmission, fuel type, and fuel consumption?

Hypotheses

We will test out the first research question using hypothesis testing with a 95% confidence level. Our hypothesis is that there is no difference in CO2 emissions for vehicles that use different fuel types.

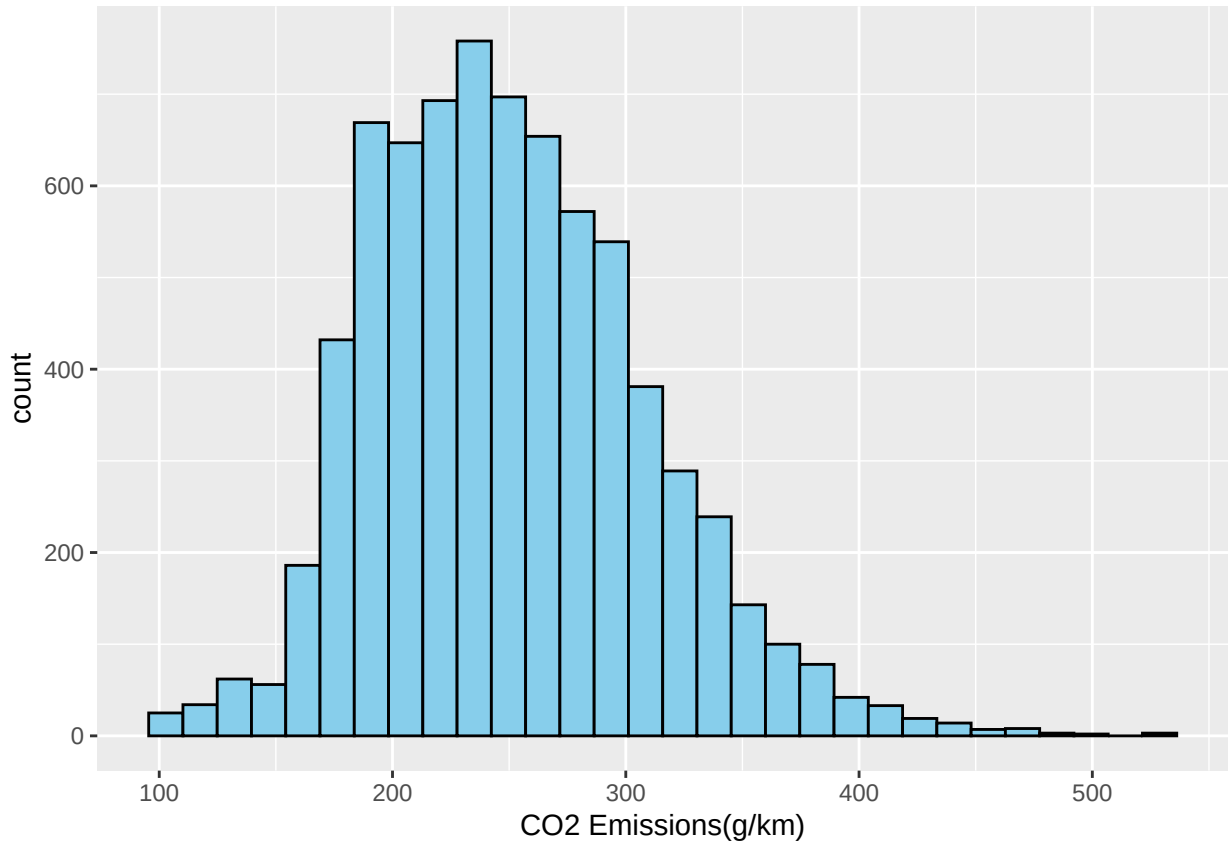
H0: CO2 Emissions = 0, Ha: CO2 Emissions != 0

Exploratory Data Analysis

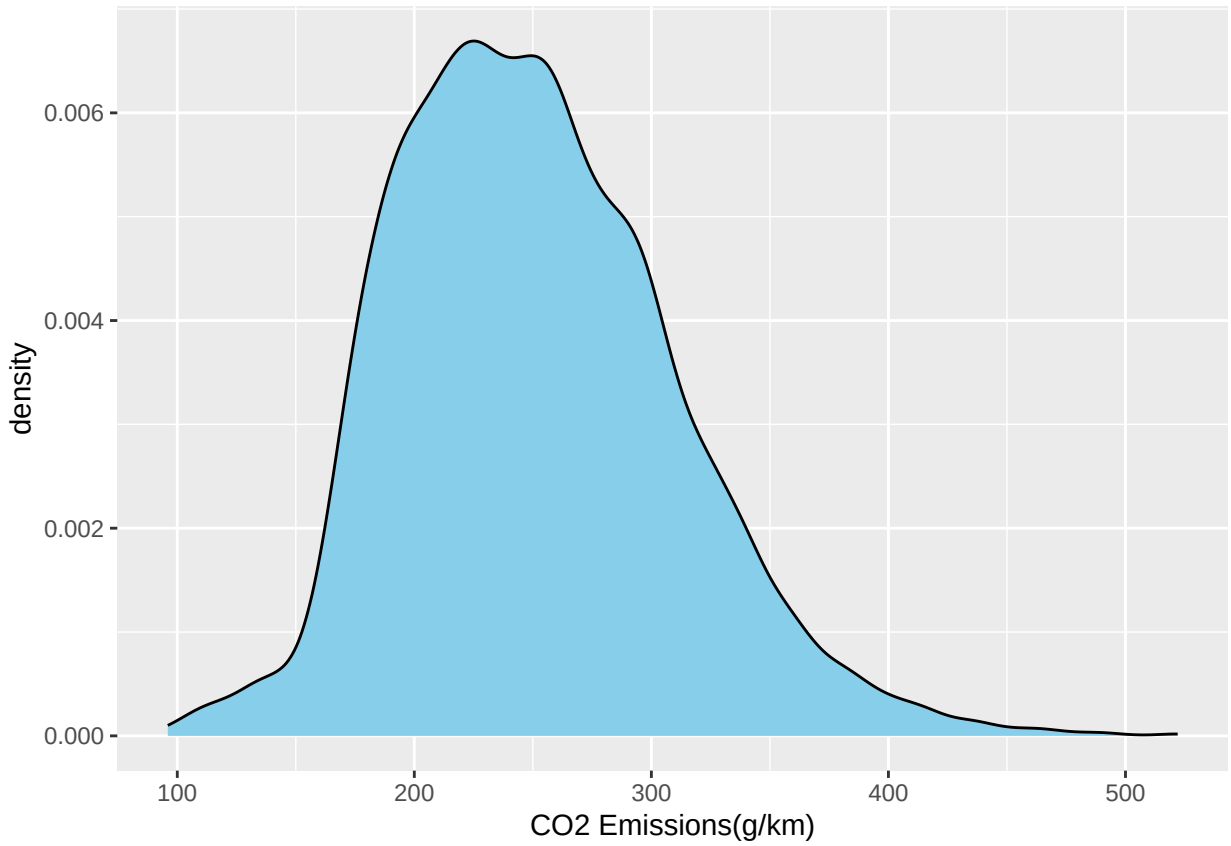
Checking the distribution for my response variable

```
CO2 %>%
  ggplot(aes(x=`CO2 Emissions(g/km)`)) + geom_histogram(color="black", fill="skyblue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
CO2 %>%
  ggplot(aes(x=`CO2 Emissions(g/km)`)) + geom_density(color="black", fill="skyblue")
```



#Comment

The Co2 Emissions seem to be skewed to the right. Most outliers of the data and the tail are to the right.

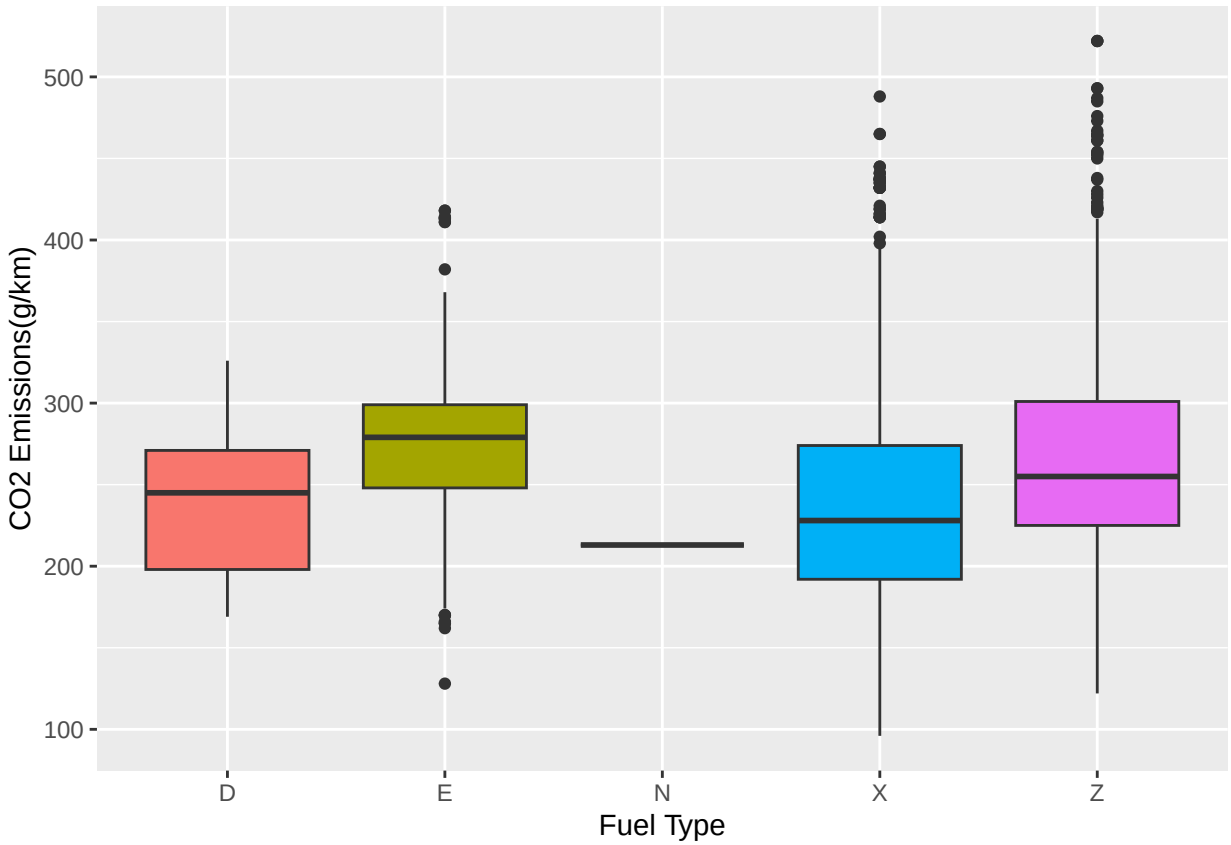
```
##Compute the summary statistics of your response variable
CO2 %>%
  summarize(Mean=mean(`CO2 Emissions(g/km)`),
            Std = sd(`CO2 Emissions(g/km)`),
            Min = min(`CO2 Emissions(g/km)`),
            Q1 = quantile(`CO2 Emissions(g/km)`, .25),
            Median = median(`CO2 Emissions(g/km)`),
            Q3 = quantile(`CO2 Emissions(g/km)`, .75),
            Max = max(`CO2 Emissions(g/km)`))
```

```
## # A tibble: 1 x 7
##    Mean   Std   Min    Q1 Median    Q3   Max
##   <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  251.  58.5    96   208    246   288   522
```

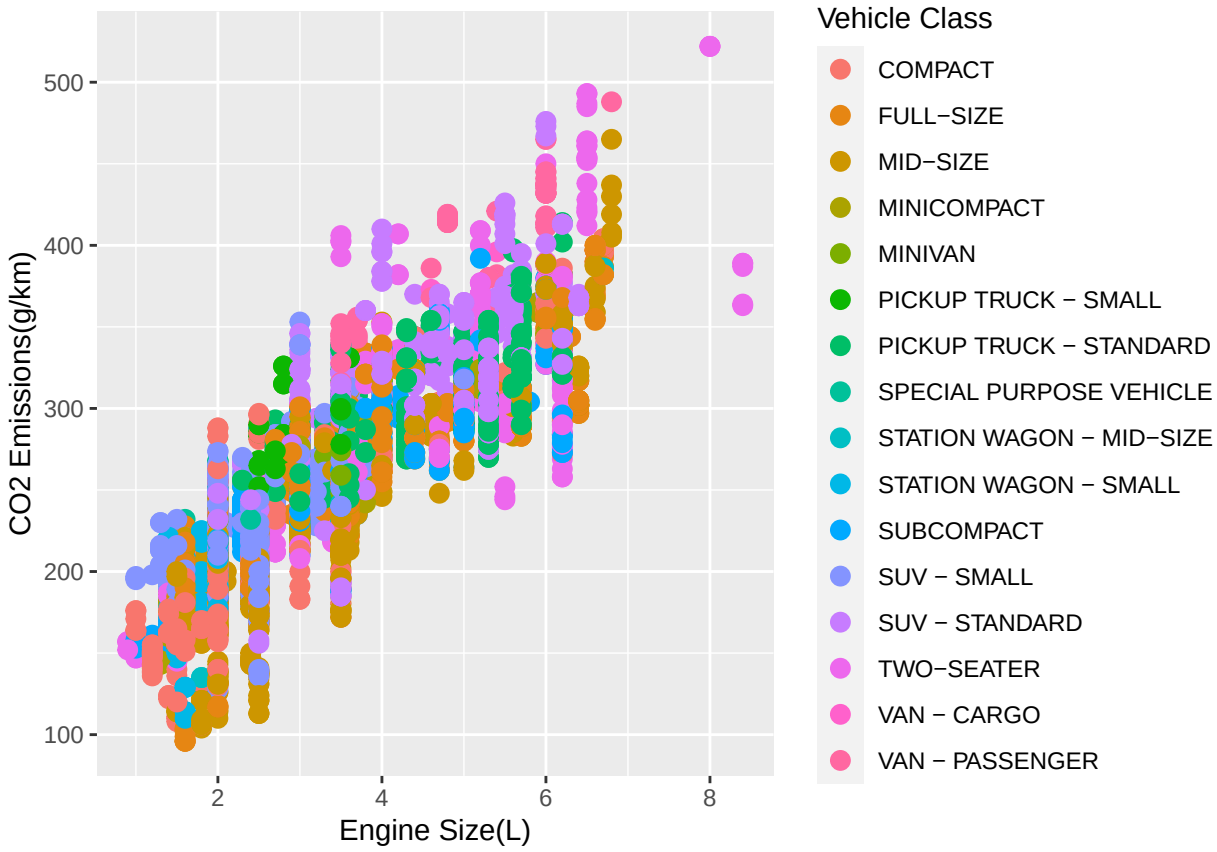
Report

The right skewness is supported by mean being higher than the median. The median CO2 Emissions is 246(g/km). Minimum CO2 Emissions is 96(g/km). Maximum is 522(g/km).

```
CO2%>%
  ggplot(aes(x=`Fuel Type`, y=`CO2 Emissions(g/km)`, fill=`Fuel Type`)) + geom_boxplot(show.legend = FA
```



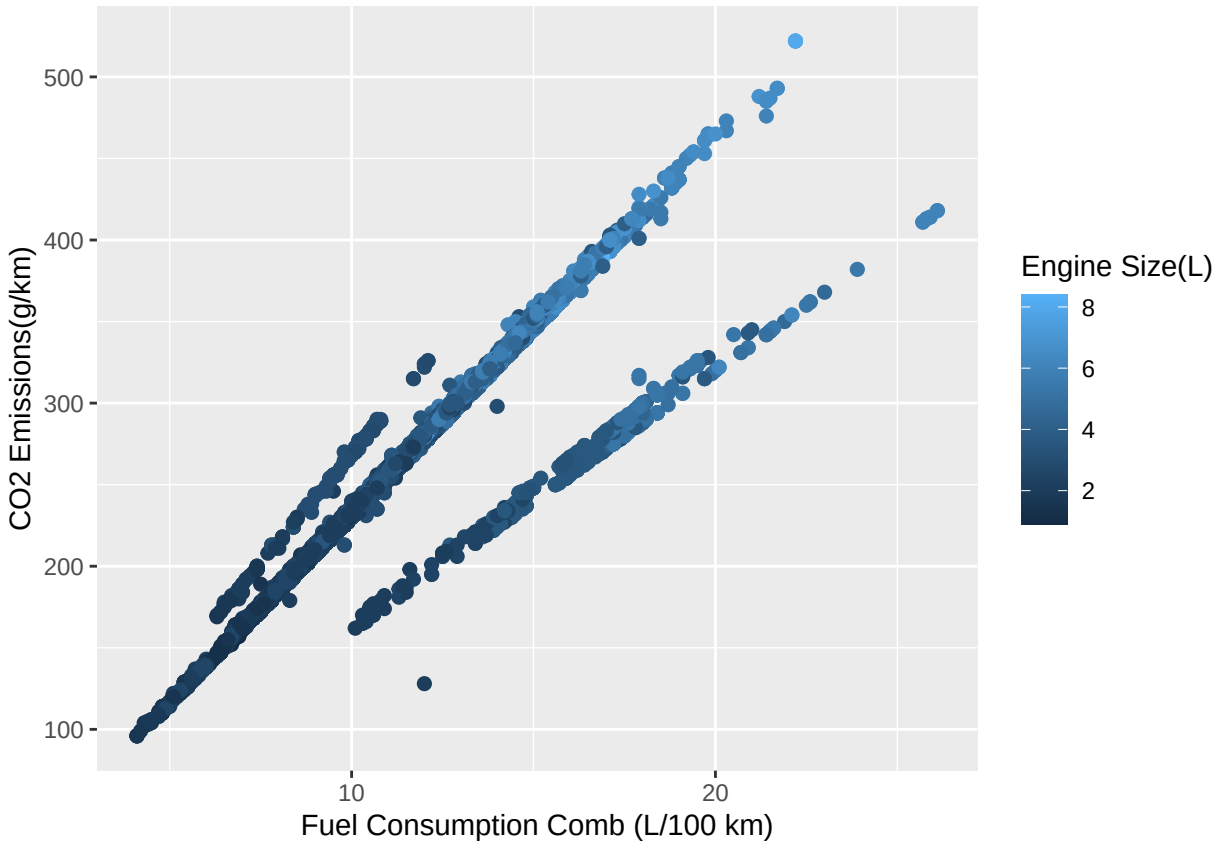

```
CO2%>%
  ggplot(aes(x=`Engine Size(L)`, y=`CO2 Emissions(g/km)`, col=`Vehicle Class`)) + geom_point(size = 3)
```



Comment

It seems that fuel type E produces more CO2 emissions than the other fuel types. This is not enough evidence to see the frequency of each fuel type or the actual amount of CO2 being emitted by each.

```
CO2 %>%
  ggplot(aes(x= `Fuel Consumption Comb (L/100 km)`, y= `CO2 Emissions(g/km)`, color = `Engine Size(L)`)
```



Comment

In the plot above we plotted Fuel consumption on cities and highways (L/100km) against CO2 Emissions. There seems to be higher CO2 Emissions and fuel consumption with bigger engine sizes. Fuel consumption and CO2 emissions have a positive relationship.

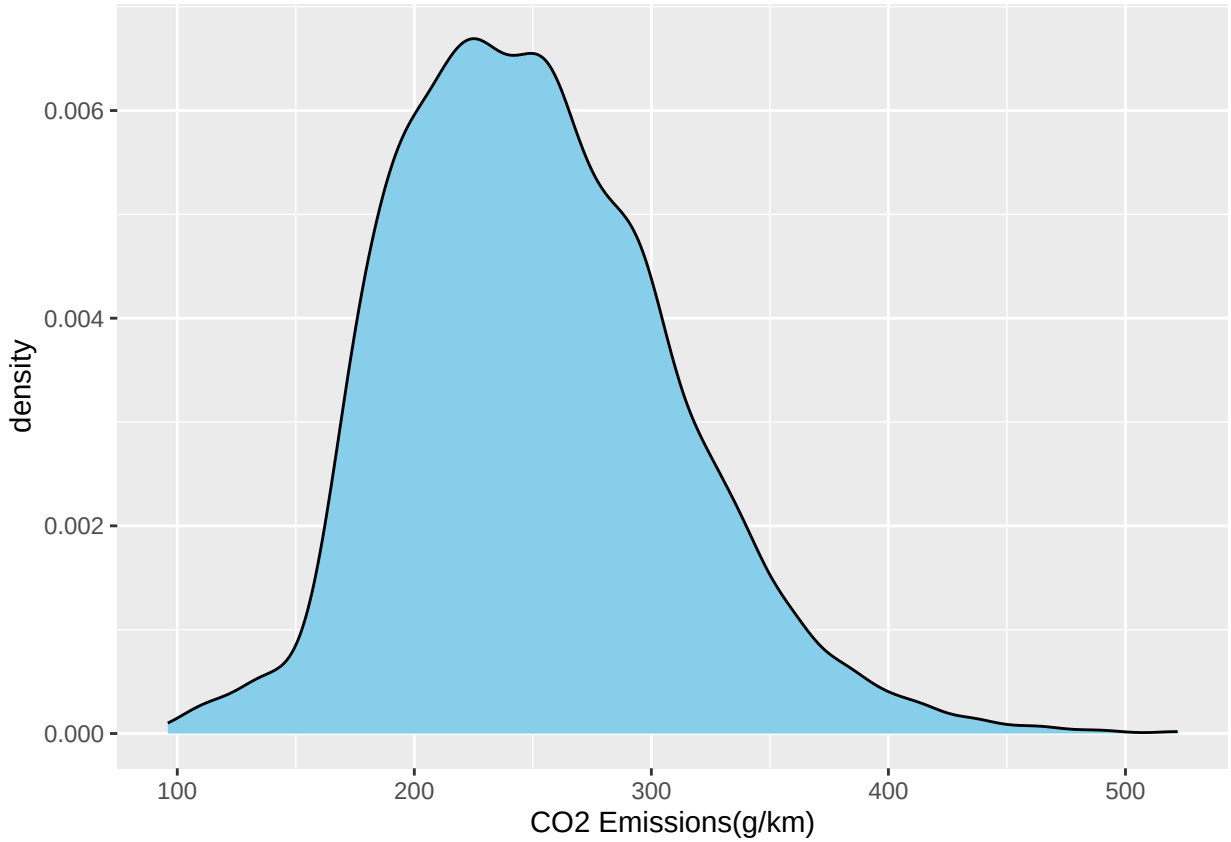
PART II

```
##Compute the summary statistics of your response variable
CO2 %>%
  summarize(Mean=mean(`CO2 Emissions(g/km)`),
            Std = sd(`CO2 Emissions(g/km)`),
            Min = min(`CO2 Emissions(g/km)`),
            Q1 = quantile(`CO2 Emissions(g/km)`, .25),
            Median = median(`CO2 Emissions(g/km)`),
            Q3 = quantile(`CO2 Emissions(g/km)`, .75),
            Max = max(`CO2 Emissions(g/km)`))
```

```
## # A tibble: 1 x 7
##    Mean   Std   Min    Q1 Median    Q3   Max
##   <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  251.  58.5    96   208    246   288   522
```

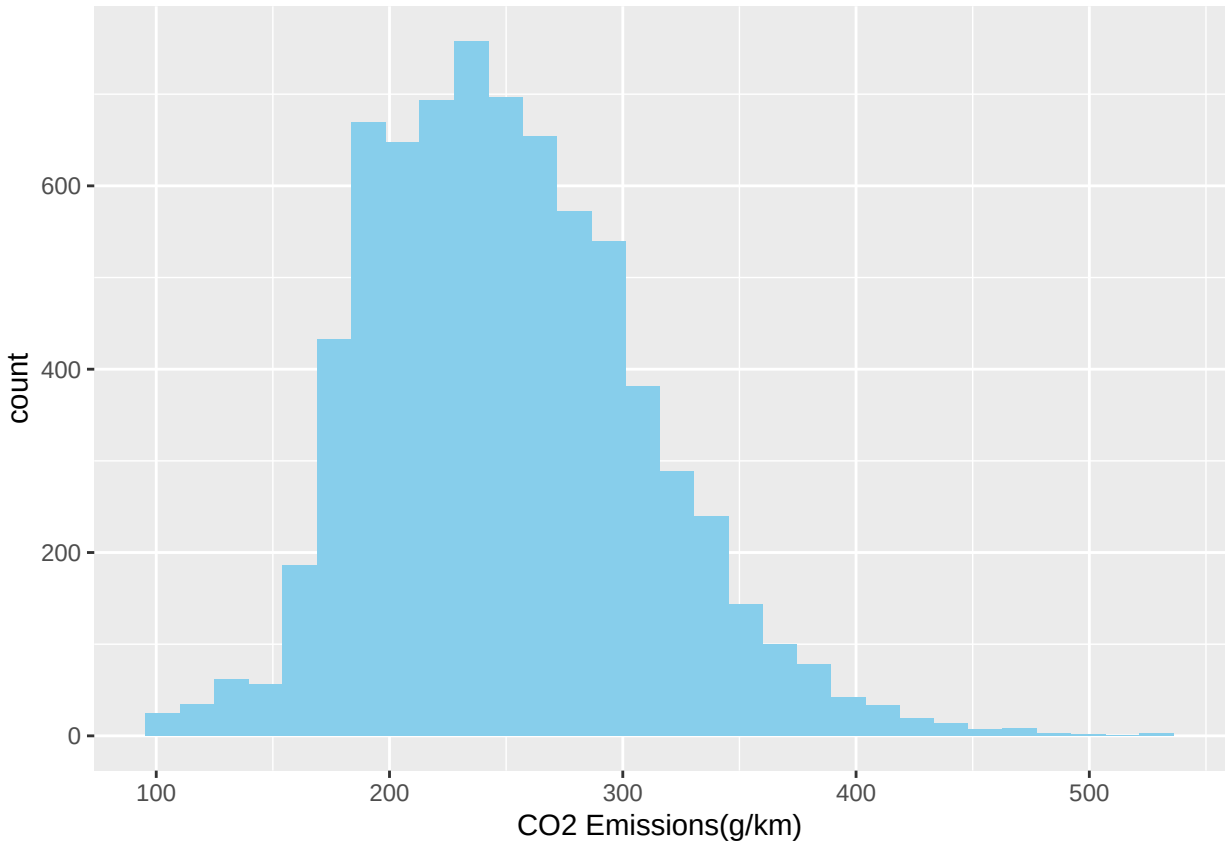
```
CO2 %>%
    filter(!is.na(`CO2 Emissions(g/km)`)) %>%
    ggplot(aes(x = `CO2 Emissions(g/km)`)) +
```

```
    geom_density(fill = "skyblue")
```




```
CO2 %>%
    filter(!is.na(`CO2 Emissions(g/km)`)) %>%
    ggplot(aes(x = `CO2 Emissions(g/km)`)) +
    geom_histogram(fill = "skyblue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



My datas et is skewed to the right so we can use the median as the appropriate statistic. The sample size is large enough for the application of the central limit theorem normal approximation to hold.

```
nrow(CO2)
```

```
## [1] 7385
```

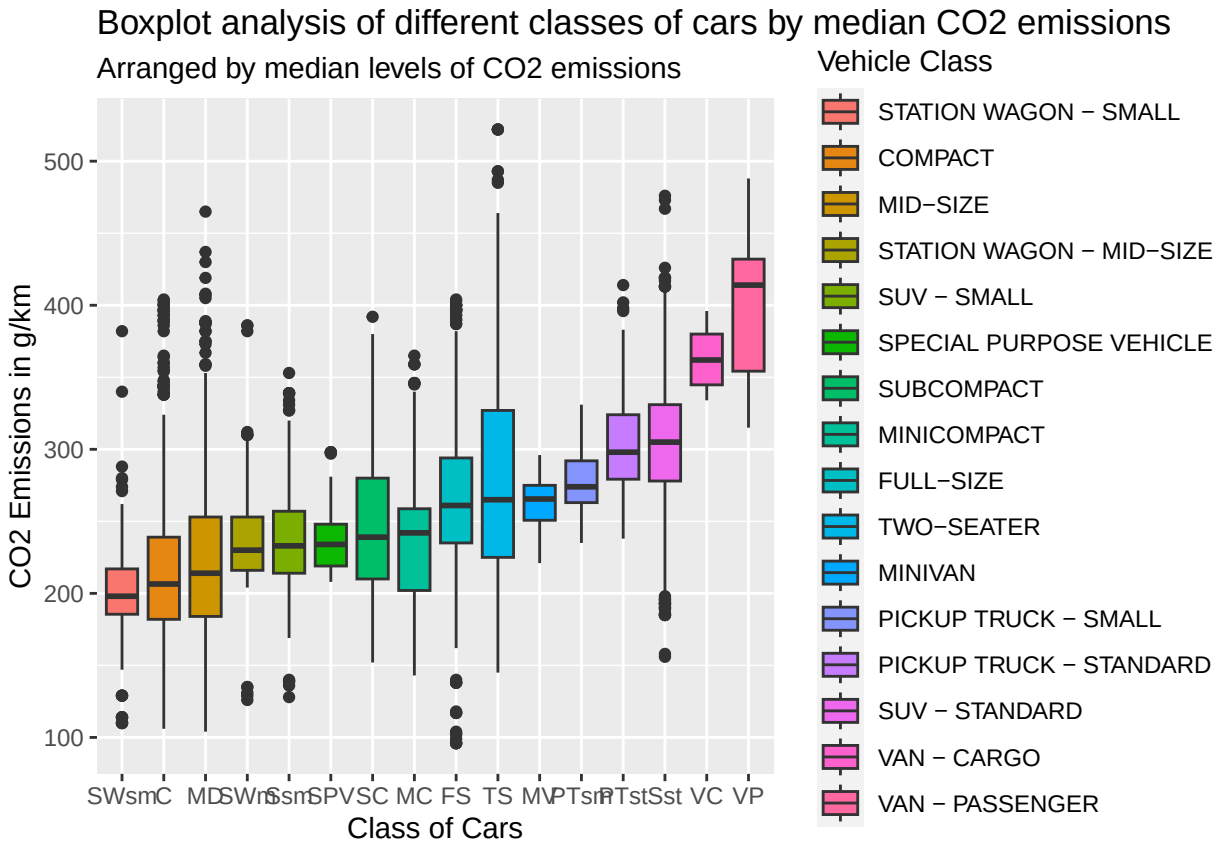
$H_0 : \mu = 246$ *vs* $H_a : \mu \neq 246$ (2 tail test) $H_0 : \mu = 246$ *vs* $H_a : \mu \geq 246$ (upper tail test) $H_0 : \mu = 246$ *vs* $H_a : \mu \leq 246$ (lower tail test)

```
CO2 %>%
  filter(!is.na(`CO2 Emissions(g/km)`)) %>%
  t_test(
    response = `CO2 Emissions(g/km)`,
    conf_int = TRUE,
    conf_level = .95,
    mu = 246,
    alternative = "two-sided"
  )
```

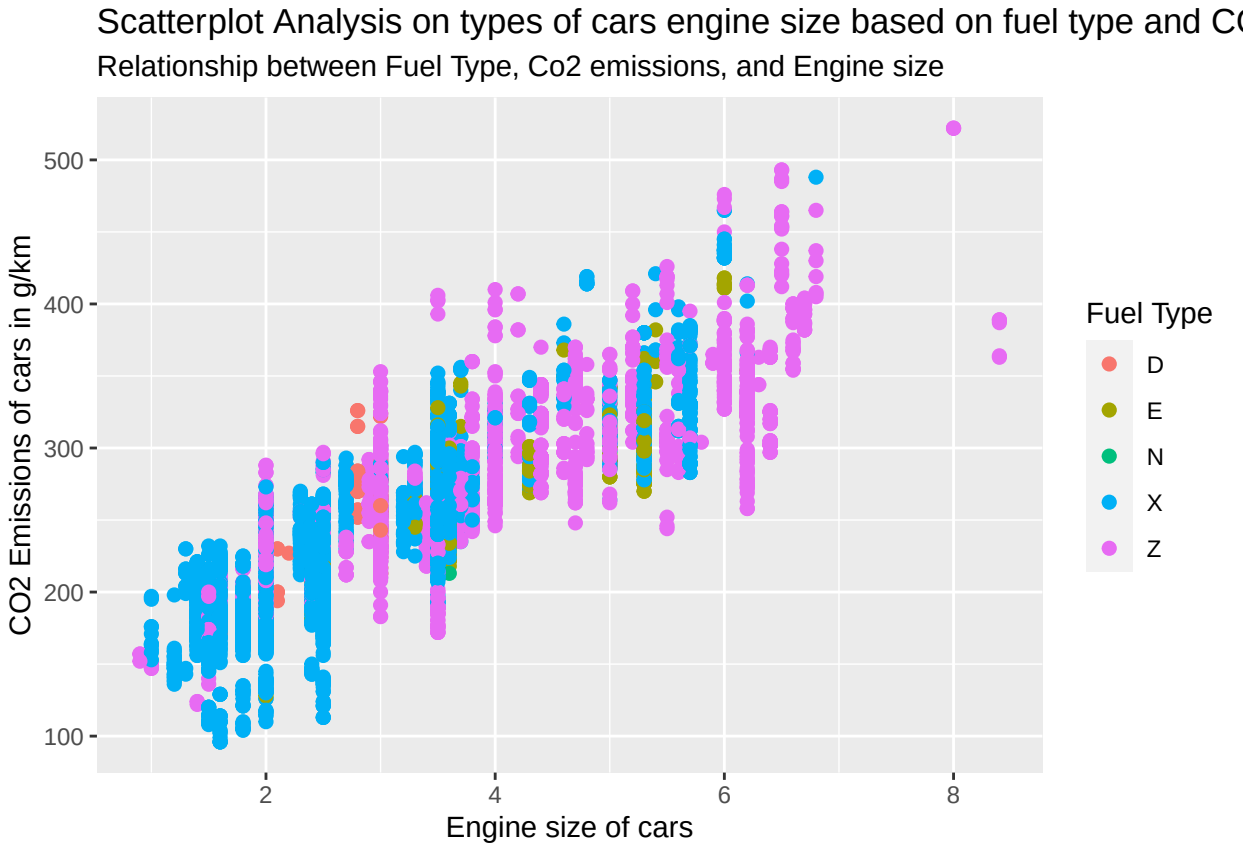
```
## # A tibble: 1 x 7
##   statistic  t_df  p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>    <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      6.73  7384 1.78e-11 two.sided       251.     249.     252.
```

We are 95% confident that the true mean does not lie between the upper and lower confidence interval(249.25, 251.92). Therefore, we can conclude that the distribution of the response variable does not have a median statistic as the appropriate statistic. So, we reject the null hypothesis.

```
CO2 %>%
  filter(!is.na(`CO2 Emissions(g/km)`), !is.na(`Vehicle Class`)) %>%
  mutate(`Vehicle Class` = reorder(`Vehicle Class`, `CO2 Emissions(g/km)`, FUN =median)) %>%
  ggplot(aes(x = `Vehicle Class`, y = `CO2 Emissions(g/km)`, fill = `Vehicle Class`)) +
  geom_boxplot() +
  labs(x = "Class of Cars", y = "CO2 Emissions in g/km", title = "Boxplot analysis of different classes
       subtitle = "Arranged by median levels of CO2 emissions") +
  scale_x_discrete(labels = c("SWsm", "C", "MD", "SWm", "Ssm", "SPV", "SC", "MC", "FS", "TS", "MV",
                              "PTsm", "PTst", "Sst", "VC", "VP"))
```




```
CO2%>%
  ggplot(aes(x=`Engine Size(L)`, y=`CO2 Emissions(g/km)`, col= `Fuel Type`)) + geom_point(size = 2) +
  labs(x = "Engine size of cars", y = "CO2 Emissions of cars in g/km", title = "Scatterplot Analysis on
```




```
CO2 %>%
  filter(!is.na(`CO2 Emissions(g/km)`), !is.na(`Vehicle Class`)) %>%
  group_by(`Vehicle Class`) %>%
  summarize(Mean=mean(`CO2 Emissions(g/km)`),
            Std = sd(`CO2 Emissions(g/km)`),
            Min = min(`CO2 Emissions(g/km)`),
            Q1 = quantile(`CO2 Emissions(g/km)`, .25),
            Median = median(`CO2 Emissions(g/km)`),
            Q3 = quantile(`CO2 Emissions(g/km)`, .75),
            Max = max(`CO2 Emissions(g/km)`))
```

```
## # A tibble: 16 x 8
##    `Vehicle Class`           Mean   Std   Min    Q1 Median    Q3   Max
##    <chr>                    <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
##  1 COMPACT                   217.  50.4   106  182    206.  239    404
##  2 FULL-SIZE                 263.  55.1    96  235    261   294    404
##  3 MID-SIZE                  222.  55.6   104  184    214   253    465
##  4 MINICOMPACT               237.  41.0   143  202    242   259.   365
##  5 MINIVAN                   262.  17.7   221  251.   266.  275    296
##  6 PICKUP TRUCK - SMALL      279.  22.9   235  263    274   292    331
##  7 PICKUP TRUCK - STANDARD   302.  30.5   238  279.   298   324    414
##  8 SPECIAL PURPOSE VEHICLE   238.  22.0   208  219    234   248    298
##  9 STATION WAGON - MID-SIZE  239.  56.4   126  216    230   253    386
## 10 STATION WAGON - SMALL     200.  33.4   110  186.   198   217    382
## 11 SUBCOMPACT                246.  49.8   152  210    239   280    392
## 12 SUV - SMALL               236.  31.2   128  214    233   257    353
## 13 SUV - STANDARD            305.  44.4   156  278    305   331    476
```

```
## 14 TWO-SEATER                277.  73.6   145  225    265   327    522
## 15 VAN - CARGO               362.  17.9   334  345.   362   380    396
## 16 VAN - PASSENGER           397.  42.3   315  354.   414   432    488
```

```
CO2 %>%
  filter(!is.na(`CO2 Emissions(g/km)`), !is.na(`Vehicle Class`)) %>%
  t_test(
    response = `CO2 Emissions(g/km)`,
    explanatory = `Vehicle Class`,
    order = c("COMPACT", "FULL-SIZE"),
    conf_int = TRUE,
    conf_level = .90,
    alternative = "two-sided"
  )
```

```
## # A tibble: 1 x 7
##   statistic  t_df  p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>    <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1     -17.3 1266. 1.33e-60 two.sided      -46.6    -51.1    -42.2
```

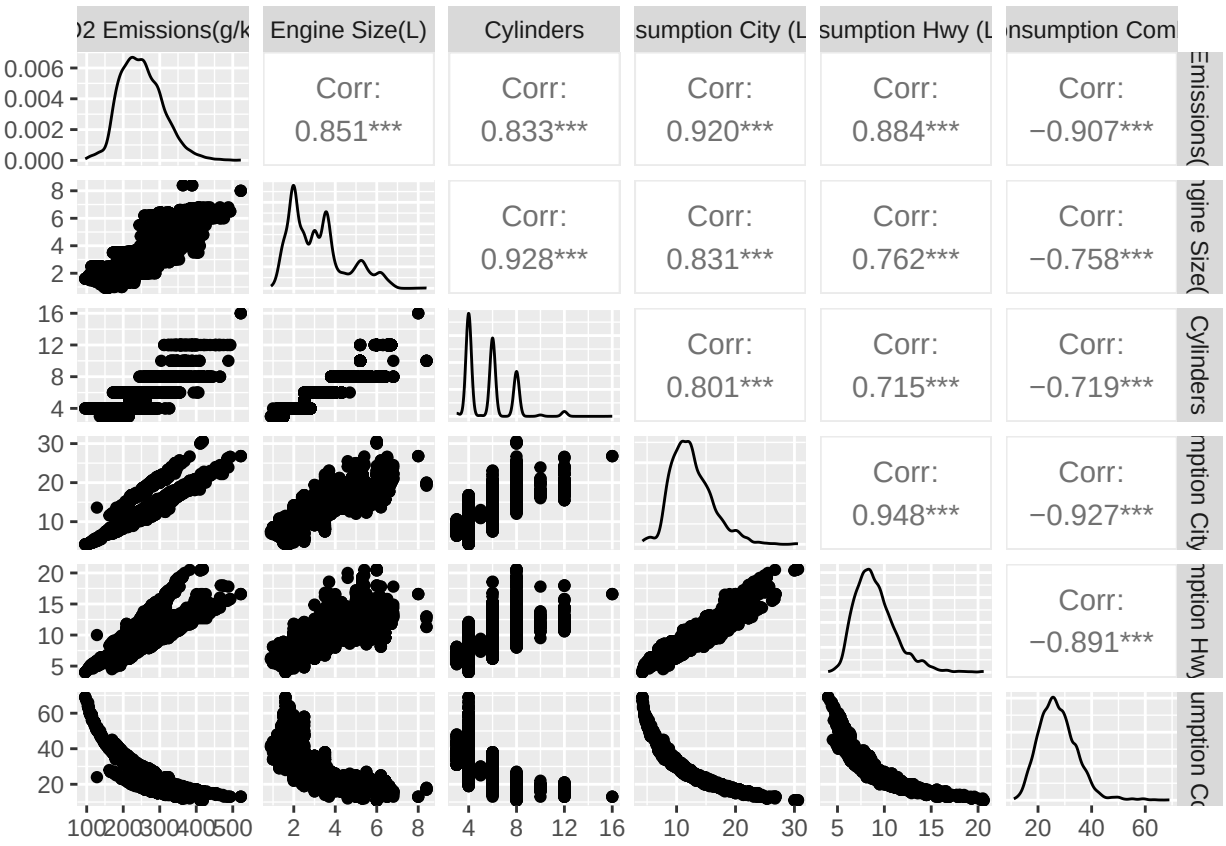
We are 90% confident that the difference in the Mean of Compact and Full-size cars lie within the upper and lower confidence interval.

Part III

Correlation

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. Correlation gives us a quantitative sense of how strong the relationship between two variables is. Correlation ranges from -1 to 1 inclusive. A positive correlation means there is a positive relationship (slope) between the two variables, that is, as one variable increases, the second variable tends to increase as well. A negative correlation means there is a negative slope between the two variables. The close the correlation is to -1 or 1, the closer it is for the relationship between the variables to be in a straight line with no variation. From the price vs horsepower plot above, we can see that the points aren't really close to each other but there is positive slope, then we can guess that the correlation is positive but isn't very close to 1. We can find the correlation between the response and the explanatory variables using the ggpairs() function from the GGally package. ggpairs(), not only calculates the correlation but it also plots the relationship for us to have a visual check for the corresponding correlation. The code is shown below:

```
CO2 %>%
  dplyr::select(`CO2 Emissions(g/km)`, `Engine Size(L)`, Cylinders, `Fuel Consumption City (L/100 km)`,
                `Fuel Consumption Hwy (L/100 km)`, `Fuel Consumption Comb (mpg)`) %>%
  ggpairs()
```



Correlation number and interpretation

We can see that CO2 Emissions(g/km) and Fuel Consumption City (L/100 km) has the highest positive correlation of 0.920. This is pretty close to 1 as we suspected. CO2 Emissions(g/km) and Fuel Consumption Comb (mpg) has a negative correlation of -0.907 and we can clearly see a negative relation in the corresponding plot. The asterisks beside the correlation number dictates the significance of the correlation. If there are no asterisks then the correlation isn't strong at all. As we can see, each correlation is above .600 and has three asterisks.

Predictive Modeling

Methodology

$\hat{Y} = \beta_0 + \beta_1 X1 + \beta_2 X2 + ... + \beta_p X_p$

$\hat{CO2Emissions(g/km)} = \beta_0 + \beta_1 \times EngineSize(L) + \beta_2 \times Cylinders + ... + \beta_p \times FuelConsumptionComb(mpg)$

```
model <- lm(`CO2 Emissions(g/km)` ~ `Engine Size(L)` + Cylinders + `Fuel Consumption City (L/100 km)` +
              `Fuel Consumption Hwy (L/100 km)` + `Fuel Consumption Comb (mpg)`, data = CO2)

summary(model)
```

```
##
## Call:
## lm(formula = `CO2 Emissions(g/km)` ~ `Engine Size(L)` + Cylinders +
##     `Fuel Consumption City (L/100 km)` + `Fuel Consumption Hwy (L/100 km)` +
##     `Fuel Consumption Comb (mpg)`, data = CO2)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -122.497   -5.690   -0.400    7.668   93.311
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       227.94950    4.19676  54.316  < 2e-16 ***
## `Engine Size(L)`                    4.99613    0.45547  10.969  < 2e-16 ***
## Cylinders                           7.53756    0.31863  23.656  < 2e-16 ***
## `Fuel Consumption City (L/100 km)`  0.89357    0.27048   3.304 0.000959 ***
## `Fuel Consumption Hwy (L/100 km)`   5.24424    0.30831  17.010  < 2e-16 ***
## `Fuel Consumption Comb (mpg)`      -3.42460    0.07855 -43.600  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.14 on 7379 degrees of freedom
## Multiple R-squared:  0.9039, Adjusted R-squared:  0.9038
## F-statistic: 1.388e+04 on 5 and 7379 DF,  p-value: < 2.2e-16
```

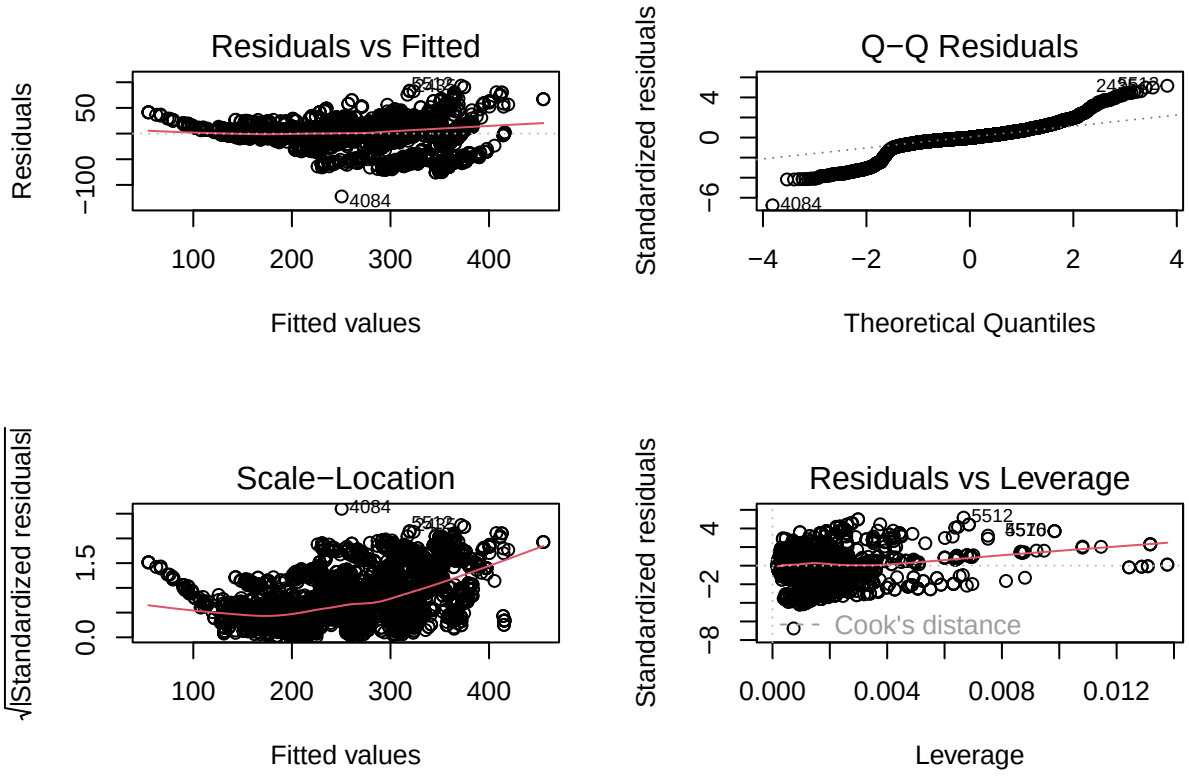
$\widehat{CO2Emissions}(g/km) = 227.94950 + 4.99613 \times EngineSize(L) + 7.53756 \times Cylinders + 0.89357 \times FuelConsumptionCity(L/100km) + 5.24424 \times FuelConsumptionHwy(L/100km) - 3.42460 \times FuelConsumptionComb(mpg)$

From the summary table above, Engine Size(L), Cylinders, Fuel Consumption City (L/100 km), and Fuel Consumption Hwy (L/100 km) have a positive estimate of 4.99, 7.53, 0.89, and 5.24 respectively. For example, this means that a unit increase in Engine Size(L) correlates to an increase of 4.99 units in CO2 Emissions(g/km) on average. We can make similar interpretations for the the other positive and negative estimate variables.From the results, we can conclude that 90.38% of total variation in the outcome (response) variable is explained by the explanatory variables collectively. The whole regression model is statistically significant as all of the variables have a p-value much lower than 0.05.

Multiple R-squared: 0.9039; 90.39% of the total variation of the response {CO2 Emissions(g/km)} variable can be explained by all the predictors (Engine Size(L), Cylinders, ..., Fuel Consumption Comb (mpg))

p-value: $<$ 2.2e-16: Since the p-value $<$ 0.05 we can conclude that the whole regression model is statistically significant

```
par(mfrow = c(2,2))
plot(model)
```



Comment

From the plot above, the normality assumption of linear regression is valid since most of the data is on the line foo each of the plots. The data is relatively close to the lines nest fit

Statistical Inference

$H_0 : \mu = 246$ *vs* $H_a : \mu \neq 246$ (2 tail test) $H_0 : \mu = 246$ *vs* $H_a : \mu \geq 246$ (upper tail test) $H_0 : \mu = 246$ *vs* $H_a : \mu \leq 246$ (lower tail test)

```
CO2 %>%
  filter(!is.na(`CO2 Emissions(g/km)`)) %>%
  t_test(
    response = `CO2 Emissions(g/km)`,
    conf_int = TRUE,
    conf_level = .95,
    mu = 246,
    alternative = "two-sided"
  )
```

```
## # A tibble: 1 x 7
##   statistic  t_df  p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>    <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      6.73  7384 1.78e-11 two.sided       251.     249.     252.
```

Results

Engine Size and fuel type are significant predicting variables for CO2 Emissions(g/km)

Interpretation

The analysis indicates a positive correlation between engine displacement and CO2 emissions, suggesting that vehicles equipped with larger engines tend to produce higher levels of carbon dioxide during operation. This relationship is likely attributable to the increased fuel consumption typically associated with larger engine sizes, resulting in greater CO2 output per unit of distance traveled.

Conclusion

In conclusion, Bigger vehicles that use more fuel will emit more CO2 into the air.

Discussion

No improvements great project.