SCHOOL OF ENGINEERING AND TECHNOLOGY

ASSESSMENT FOR THE MASTER OF DATA SCIENCE

SUBJECT CODE AND TITLE:		MDS5033 Statistical Methods for Data Science
ASSESSMENT DUE DATE:		**12/12/2023**
NO.	**STUDENT ID**	**STUDENT NAME**
1	**14086334**	**KAN JUN FAI**

IMPORTANT

The University requires students to adhere to submission deadlines for any form of assessment. Penalties are applied in relation to unauthorised late submission of work. Coursework submitted after the deadline will be subjected to the prevailing academic regulations. Please check your respective programme handbook.

Academic Honesty Acknowledgement

"I KAN JUN FAI (name) verify that this paper contains entirely my own work. I have not consulted with any outside person or materials other than what was specified (an interviewee, for example) in the assignment or the syllabus requirements. Further, I have not copied or inadvertently copied ideas, sentences, or paragraphs from another student. I realise the penalties *(refer to the student handbook and undergraduate programme handbook)* for any kind of copying or collaboration."

..... (Student' Signature / Initial)

Project Title: Evaluating Environmental Impact of Automotive Vehicles.

Dataset of Case Study

A dataset entitled "Carbon Dioxide Emission by Vehicles" was obtained from Kaggle . Below are all the features available in this dataset.

Feature	Description
Year	The model year of the vehicle (all entries are from 2022).
Make	The manufacturer of the vehicle (all entries are Acura).
Model	The specific model of the vehicle (e.g., ILX, MDX SH-AWD).
Vehicle Class	The category of the vehicle (e.g., Compact, SUV: Small).
Engine Size (L)	The volume of the engine in liters.
Cylinders	The number of cylinders in the vehicle's engine.
Transmission	Type of transmission (e.g., AM8 for 8-speed automated manual, AS10 for 10-speed automatic).
Fuel Type	The type of fuel the vehicle uses (e.g., Z).
Fuel Consumption [City (L/100 km)]	Fuel consumption in liters per 100 kilometers in city driving.
Fuel Consumption [Hwy (L/100 km)]	Fuel consumption on the highway.
Fuel Consumption [Comb (L/100 km)]	Combined city and highway fuel consumption.
Fuel Consumption [Comb (mpg)]	Combined fuel consumption in miles per gallon.
CO2 Emissions (g/km)	Carbon dioxide emissions in grams per kilometer.
Motor (kW)	The power of the electric motor in kilowatts, if applicable.
Fuel Type 1	Another category for fuel type, if applicable.
Fuel Consumption Combined Le/100 km	Combined fuel consumption in liters equivalent per 100 kilometers, for electric or hybrid vehicles.
Range 1 (km)	The driving range in kilometers under certain conditions.
Recharge Time (h)	Time required to recharge the vehicle, for electric or hybrid vehicles.
Fuel Type 2	Secondary fuel type, if the vehicle is a hybrid.
Range 2 (km)	Additional range information for a secondary fuel type.
Fuel Consumption [City (kWh/100 km)]	City fuel consumption for electric vehicles, measured in kWh per 100 kilometers.
Fuel Consumption [Comb (kWh/100 km)]	Combined fuel consumption for electric vehicles.
Consumption [City (Le/100 km)]	City consumption in liters equivalent.
Consumption [Hwy (Le/100 km)]	Highway consumption in liters equivalent.
Consumption [Comb (Le/100 km)]	Combined consumption in liters equivalent.
Range (km)	The total range of the vehicle in kilometers.

Case Study Aim and Objective

The aim of this case study is to perform comparative analysis of driving conditions on fuel efficiency and CO2 emissions. The objective is to investigate the relationship between driving conditions (city and highway) and the environmental impact of automotive vehicles, focusing on carbon dioxide (CO2) emissions and fuel consumption.

Relevant Variables / Attributes for Objectives

Based on the dataset above, the first motivation of this case study is to explore and analyse the variability in CO2 emissions and fuel consumption under different driving conditions to identify if there are any trends and patterns. The relevant variables used are CO2 Emissions (g/km), Fuel Consumption [City (L/100 km)] (will now be named FC-CT), Fuel Consumption [Hwy (L/100 km)] (will now be named FC-HW), and Fuel Consumption [Comb (L/100 km)] (will now be named FC-COMB).

The second motivation of this study is to investigate and analyse if the vehicle type driven – SUV, Full-size, Mid-size, Two-seaters, and Pickup trucks – influence the CO2 emissions based on the driving conditions. The relevant variables used are CO2 emissions, FC-CT, FC-HW, FC-COMB, and Vehicle Class (will now be named VClass).

The table below summarizes the short-formed variables names mentioned above for future reference in this report, and in RStudio application.

Original Variable Name	**Short Name**
CO2 Emissions (g/km)	CO2
Fuel Consumption [City (L/100 km)]	FC-CT
Fuel Consumption [Highway (L/100 km)]	FC-HW
Fuel Consumption [Comb (L/100 km)]	FC-COMB
Vehicle Class (e.g., SUV, Full-size, Mid-size, Two-seaters, Pickup Trucks)	VClass

Descriptive Statistics Using R

Using RStudio program as indicated in the image, the dimensions of the dataset named "CO2_Emission" was tabulated to be 27549 rows and 26 columns. The summary was generated and the mentioned data features above is indicated as in red boxes in the image below. The feature "Vehicle.Class" contains only strings as its values, while "FC-CT" , "FC-HW" , "FC-COMB" , and "CO2" is a continuous numerical variable, with its minimum, maximum, median, mean, 1st quartile and 3rd quartile shown in the image.

```
dim(CO2_emission)

# Number of Rows:        27549
# Number of Columns:     26

summary(CO2_emission)
```

```
Console   Terminal ×   Background Jobs ×
R 4.3.2 · C:/Users/Admin/Downloads/
> dim(CO2_emission)
[1] 27549    26
> summary(CO2_emission)
      Year          Make              Model          Vehicle.Class      Engine.Size..L.   Cylinders      Transmission        Fuel.Type
 Min.   :1995   Length:27549       Length:27549       Length:27549       Min.   :0.600   Min.   : 2.00   Length:27549       Length:27549
 1st Qu.:2004   Class :character   Class :character   Class :character   1st Qu.:2.200   1st Qu.: 4.00   Class :character   Class :character
 Median :2010   Mode  :character   Mode  :character   Mode  :character   Median :3.000   Median : 6.00   Mode  :character   Mode  :character
 Mean   :2010                                                           Mean   :3.343   Mean   : 5.83
 3rd Qu.:2016                                                           3rd Qu.:4.200   3rd Qu.: 8.00
 Max.   :2022                                                           Max.   :8.400   Max.   :16.00
                                                                        NA's   :323     NA's   :323
 Fuel.Consumption..City..L.100.km.. Fuel.Consumption..Hwy..L.100.km.. Fuel.Consumption..Comb..L.100.km.. Fuel.Consumption..Comb..mpg..
 Min.   : 4.00                      Min.   : 3.90                     Min.   : 4.00                     Min.   :10.00
 1st Qu.:11.30                      1st Qu.: 8.30                     1st Qu.: 9.90                     1st Qu.:20.00
 Median :13.40                      Median : 9.60                     Median :11.60                     Median :24.00
 Mean   :13.84                      Mean   :10.15                     Mean   :12.12                     Mean   :24.86
 3rd Qu.:15.90                      3rd Qu.:11.50                     3rd Qu.:13.90                     3rd Qu.:28.00
 Max.   :33.30                      Max.   :35.00                     Max.   :27.50                     Max.   :71.00
 NA's   :323                                                          NA's   :323                       NA's   :548
 CO2.Emissions..g.km.   Motor..kW.      Fuel.Type.1        Fuel.Consumption.Combined.Le.100.km  Range.1..km.     Recharge.Time..h. Fuel.Type.2
 Min.   :  0.0        Min.   : 35.00   Length:27549       Length:27549                         Min.   : 18.00   Min.   : 1.300    Length:27549
 1st Qu.:228.0        1st Qu.: 80.75   Class :character   Class :character                     1st Qu.: 27.00   1st Qu.: 3.000    Class :character
 Median :267.0        Median :135.00   Mode  :character   Mode  :character                     Median : 34.00   Median : 7.000    Mode  :character
 Mean   :271.4        Mean   :209.91                                                           Mean   : 45.13   Mean   : 6.892
 3rd Qu.:313.0        3rd Qu.:313.50                                                           3rd Qu.: 50.00   3rd Qu.:10.100
 Max.   :633.0        Max.   :829.00                                                           Max.   :203.00   Max.   :15.000
                      NA's   :27001                                                            NA's   :27324    NA's   :27001
  Range.2..km.   Fuel.Consumption..City..kWh.100.km.. Fuel.Consumption..Comb..kWh.100.km.. Consumption..City..Le.100.km.. Consumption..Hwy..Le.100.km..
 Min.   :116.0   Min.   :13.70                        Min.   :14.80                        Min.   :1.600                  Min.   :1.800
 1st Qu.:512.0   1st Qu.:16.90                        1st Qu.:18.50                        1st Qu.:1.900                  1st Qu.:2.200
 Median :666.0   Median :19.00                        Median :20.20                        Median :2.100                  Median :2.400
 Mean   :654.2   Mean   :20.22                        Mean   :21.02                        Mean   :2.273                  Mean   :2.468
 3rd Qu.:798.0   3rd Qu.:23.00                        3rd Qu.:22.70                        3rd Qu.:2.600                  3rd Qu.:2.600
 Max.   :995.0   Max.   :32.90                        Max.   :32.40                        Max.   :3.700                  Max.   :3.900
 NA's   :27324   NA's   :27226                        NA's   :27226                        NA's   :27226                  NA's   :27226
 Consumption..Comb..Le.100.km..   Range..km.
 Min.   :1.700                    Min.   : 92.0
 1st Qu.:2.100                    1st Qu.:286.0
 Median :2.300                    Median :385.0
 Mean   :2.359                    Mean   :374.3
 3rd Qu.:2.600                    3rd Qu.:478.0
 Max.   :3.600                    Max.   :837.0
 NA's   :27226                    NA's   :27226
```

The chosen features above were then consolidated into one data frame named "Emissions" for better view. Based on the image below, there are 323 null values indicated as "NA's" in FC-CT, and FC-COMB.

```
> selected_features <- c("Vehicle.Class", "CO2.Emissions..g.km.", "Fuel.Consumption..City..L.100.km..", "Fuel.Consumption..Hwy..L.100.km..", "Fuel.Consumption..
Comb..L.100.km..")
> Emissions <- CO2_emission[selected_features]
> summary(Emissions)
 Vehicle.Class      CO2.Emissions..g.km. Fuel.Consumption..City..L.100.km.. Fuel.Consumption..Hwy..L.100.km.. Fuel.Consumption..Comb..L.100.km..
 Length:27549       Min.   :  0.0        Min.   : 4.00                      Min.   : 3.90                      Min.   : 4.00
 Class :character   1st Qu.:228.0        1st Qu.:11.30                      1st Qu.: 8.30                      1st Qu.: 9.90
 Mode  :character   Median :267.0        Median :13.40                      Median : 9.60                      Median :11.60
                    Mean   :271.4        Mean   :13.84                      Mean   :10.15                      Mean   :12.12
                    3rd Qu.:313.0        3rd Qu.:15.90                      3rd Qu.:11.50                      3rd Qu.:13.90
                    Max.   :633.0        Max.   :33.30                      Max.   :35.00                      Max.   :27.50
                                         NA's   :323                                                           NA's   :323
```

The null values in FC-CT were imputed with values from the FC-HW, under the assumption that city and highway consumption are equivalent for those cases. This approach preserves more data than removing that row of data entirely.

```
> summary(Emissions)    #Summary after adjusting NULL values.
 Vehicle.Class      CO2.Emissions..g.km. Fuel.Consumption..City..L.100.km.. Fuel.Consumption..Hwy..L.100.km.. Fuel.Consumption..Comb..L.100.km..
 Length:27549       Min.   :  0.0        Min.   : 4.00                      Min.   : 3.90                     Min.   : 4.00
 Class :character   1st Qu.:228.0        1st Qu.:11.30                      1st Qu.: 8.30                     1st Qu.:10.00
 Mode  :character   Median :267.0        Median :13.40                      Median : 9.60                     Median :11.70
                    Mean   :271.4        Mean   :13.94                      Mean   :10.15                     Mean   :12.23
                    3rd Qu.:313.0        3rd Qu.:16.00                      3rd Qu.:11.50                     3rd Qu.:14.00
                    Max.   :633.0        Max.   :35.00                      Max.   :35.00                     Max.   :35.00
>
```

The minimum value of “CO2” is 0. The 0 values are removed entirely, with the reason that cars cannot emit 0 carbon dioxide while still producing movement of the vehicle.

```
> summary(VClass)  # Summary after extracting only "SUV", "FULL-SIZE", "MID-SIZE", "TWO-SEATER", "PICKUP TRUCK"
 Vehicle.Class      CO2.Emissions..g.km. Fuel.Consumption..City..L.100.km.. Fuel.Consumption..Hwy..L.100.km.. Fuel.Consumption..Comb..L.100.km..
 Length:17004       Min.   :  0.0        Min.   : 4.0                       Min.   : 3.90                     Min.   : 4.00
 Class :character   1st Qu.:242.0        1st Qu.:11.9                       1st Qu.: 8.70                     1st Qu.:10.60
 Mode  :character   Median :279.0        Median :14.2                       Median :10.10                     Median :12.30
                    Mean   :281.6        Mean   :14.5                       Mean   :10.61                     Mean   :12.75
                    3rd Qu.:327.0        3rd Qu.:16.7                       3rd Qu.:12.20                     3rd Qu.:14.70
                    Max.   :633.0        Max.   :35.0                       Max.   :35.00                     Max.   :35.00
>
```

After removing the 0 values from the dataset, the new minimum value is 36.

```
> summary(VClass) # Removing "0" values from CO2.Emissions
 Vehicle.Class      CO2.Emissions..g.km. Fuel.Consumption..City..L.100.km.. Fuel.Consumption..Hwy..L.100.km.. Fuel.Consumption..Comb..L.100.km..
 Length:16770       Min.   : 36.0        Min.   : 4.0                       Min.   : 3.90                     Min.   : 4.00
 Class :character   1st Qu.:244.0        1st Qu.:11.9                       1st Qu.: 8.70                     1st Qu.:10.50
 Mode  :character   Median :281.0        Median :14.2                       Median :10.00                     Median :12.30
                    Mean   :285.6        Mean   :14.4                       Mean   :10.45                     Mean   :12.62
                    3rd Qu.:327.0        3rd Qu.:16.6                       3rd Qu.:12.10                     3rd Qu.:14.57
                    Max.   :633.0        Max.   :33.3                       Max.   :22.10                     Max.   :27.50
>
```

In the “Vehicle.Type” feature, the dataset contains segregated subsets with different names.

```
R 4.3.2 · C:/Users/Admin/Downloads/
                  1                    2                    1                    1                    1
        Subcompact          SUBCOMPACT          SUBCOMPACT?          SUBCOMPACT<          Subcompact>
                496                 2025                    1                    1                    1
        Subcompatc          SUBCOMPATC                  SUV          SUV - S%MALL     SUV - S?M>AL>L
                 23                   61                 3015                    1                    1
  SUV - S<TANDARD     SUV - SM%A?LL     SUV - SM>AL!L     SUV - SM>ALL     SUV - SMA!LL
                  1                    1                    1                    1                    1
    SUV - SMA>LL          SUV - SMALL     SUV - ST>ANDARD     SUV - STAND<AR%D     SUV - STANDADR
                  1                  831                    1                    1                   14
 SUV - STANDADR>     SUV - STANDARD     SUV ?- ?STANDAR!D     SUV ?- S>MALL     SUV >- S!TANDARD
                  1                  515                    1                    1                    1
             SUV!     SUV! - STANDARD       SUV!: Smal?l     SUV!: Smal!l                  SUV%
                  1                    1                    1                    1                    1
    SUV% - SMALL     SUV: ?S>tandard       SUV: <Small     SUV: >Small     SUV: Sm%all
                  1                    1                    1                    1                    1
   SUV: Sma!l>l         SUV: Sma!ll       SUV: Sma>ll     SUV: Smal<l     SUV: Smal<l?
                  1                    1                    1                    1                    1
     SUV: Small         SUV: Small>     SUV: St>and!ard     SUV: Sta!nda%rd     SUV: Sta%ndard
               1019                    1                    1                    1                    1
 SUV: Sta?ndard     SUV: Stan!dard     SUV: Stan%dard     SUV: Standadr     SUV: Standard
                  1                    1                    1                   22                  691
   SUV:! Smal>l         SUV:! Small     SUV:? Sma?ll!     SUV:< Standar>d     SUV:> Sm%a>l>l
                  1                    1                    1                    1                    1
    SUV:> Small     SUV:> Standa!rd                 SUV?     SUV? - S%MALL     SUV?: Small
                  1                    1                    2                    1                    1
   SUV< - ?SMALL     SUV< - STANDARD     SUV< !- SMALL                  SUV>     SUV> - <SMALL%
                  1                    2                    1                    4                    1
 SUV> - STAN%DARD         SUV>: Small                  SVU     T<WO-S%EATER     T>W>O-SEATER
                  1                    2                   95                    1                    1
   TW!O>-<SEATER     TW?O-SEA>TER     TW<O-S?EATER     Tw>o-seater     TW>O-SEATER
                  1                    1                    1                    1                    1
    TWO-!S%EATER     TWO-%SEATER     TWO-<SEATER     TWO->SEATER     TWO-S!EATER
                  1                    1                    1                    1                    1
    TWO-S>EAT<ER     TWO-SE%ATER     TWO-SE?ATER     TWO-SEA?TER<     TWO-SEA>TER
                  1                    1                    1                    2                    1
     TWO-SEATE>R      Two-seater     TWO-SEATER     TWO-SEATER<      Two-seatre
                  1                  330                 1080                    1                   13
     TWO-SEATRE     TWO%-S!EATER     TWO<-SEATER     TWO>-SEATE?R     TWO>-SEATER?
                 30                    1                    1                    1                    1
                 UL     V%AN - PASSENGER     V%AN -? PASSE?NG?ER?     V%AN% - P%ASS?ENG?ER     V?AN> - CARGO
                  1                    1                    1                    1                    1
    VA%N - CARGO      VAN - !C<ARGO     VAN - ?C>ARGO     VAN - CAR%GO     VAN - CAR>GO
                  1                    1                    2                    2                    1
     VAN - CARGO      VAN - CAROG     VAN - PASSENGER     VAN - PASSENGRE     VAN !- CARGO
                457                   13                  394                    7                    1
 VAN !- PASSENGER     VAN ?- <PASS>ENGER     VAN ?- PASSENGER     Van!: Passe%nger     Van: Passenger
                  1                    1                    1                    1                   10
   Van: Passengre      VAN? - CA?R<GO
                  1                    1
```

The action taken was to combine all “SUV” type into just one feature name “SUV”. The same was done for “FULL-SIZE”, “MID-SIZE”, “TWO-SEATER”, and “PICKUP TRUCK”. The new values after combining each respective type are as follows:

New SUV Value: 6166

New FULL-SIZE Value: 1824

New MID-SIZE Value: 4120

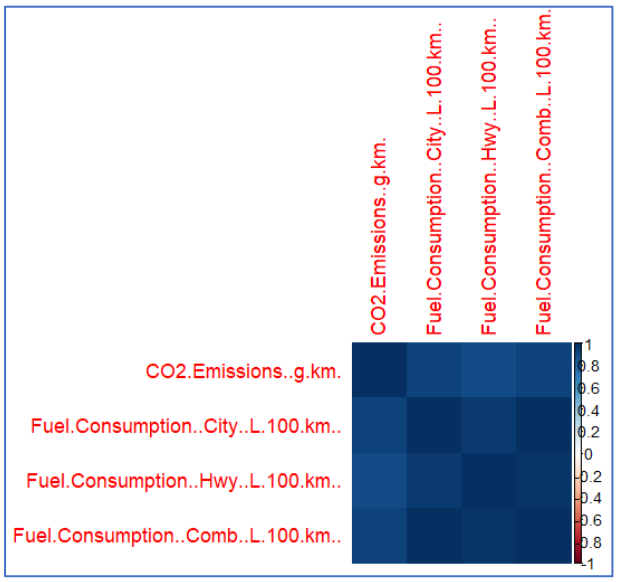
New TWO-SEATER Value: 1470

New PICKUP TRUCK Value: 3424

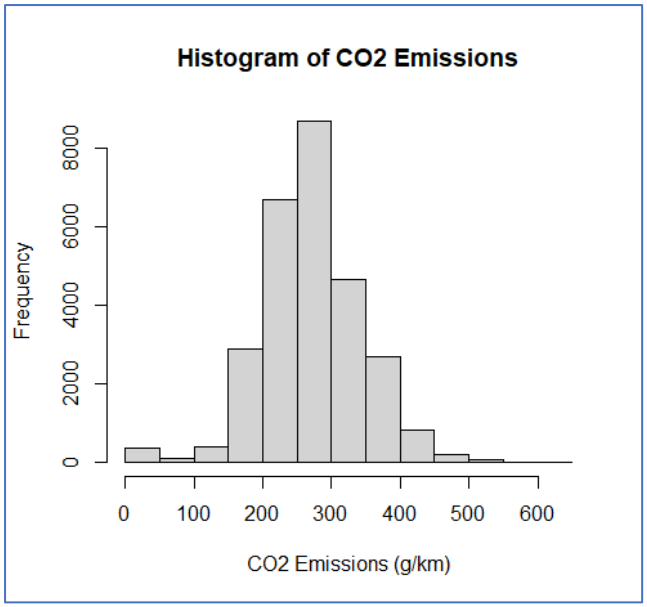
The summary for each Vehicle Type are shown as the below image.

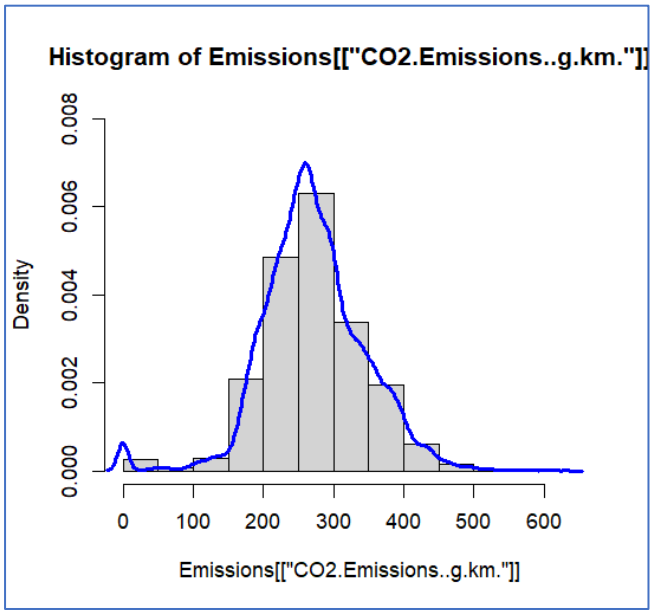
```
> summary(V.SUV)
 Vehicle.Class      CO2.Emissions..g.km. Fuel.Consumption..City..L.100.km.. Fuel.Consumption..Hwy..L.100.km.. Fuel.Consumption..Comb..L.100.km..
 Length:6166        Min.   :  0.0        Min.   : 5.40                      Min.   : 6.00                      Min.   : 5.8
 Class :character   1st Qu.:239.0        1st Qu.:11.60                      1st Qu.: 8.90                      1st Qu.:10.4
 Mode  :character   Median :278.0        Median :13.90                      Median :10.30                      Median :12.3
                    Mean   :282.3        Mean   :14.38                      Mean   :10.88                      Mean   :12.8
                    3rd Qu.:329.0        3rd Qu.:16.70                      3rd Qu.:12.40                      3rd Qu.:14.7
                    Max.   :476.0        Max.   :32.10                      Max.   :32.10                      Max.   :32.1
> summary(V.FULL.SIZE)
 Vehicle.Class      CO2.Emissions..g.km. Fuel.Consumption..City..L.100.km.. Fuel.Consumption..Hwy..L.100.km.. Fuel.Consumption..Comb..L.100.km..
 Length:1824        Min.   :  0.0        Min.   : 4.00                      Min.   : 3.90                      Min.   : 4.00
 Class :character   1st Qu.:244.0        1st Qu.:12.90                      1st Qu.: 8.70                      1st Qu.:11.00
 Mode  :character   Median :278.0        Median :14.70                      Median : 9.80                      Median :12.50
                    Mean   :268.6        Mean   :14.74                      Mean   :10.18                      Mean   :12.69
                    3rd Qu.:304.0        3rd Qu.:16.10                      3rd Qu.:10.80                      3rd Qu.:13.80
                    Max.   :504.0        Max.   :25.30                      Max.   :24.00                      Max.   :24.00
> summary(V.MID.SIZE)
 Vehicle.Class      CO2.Emissions..g.km. Fuel.Consumption..City..L.100.km.. Fuel.Consumption..Hwy..L.100.km.. Fuel.Consumption..Comb..L.100.km..
 Length:4120        Min.   :  0.0        Min.   : 4.30                      Min.   : 4.200                     Min.   : 4.30
 Class :character   1st Qu.:209.8        1st Qu.:10.50                      1st Qu.: 7.500                     1st Qu.: 9.10
 Mode  :character   Median :253.0        Median :12.90                      Median : 8.800                     Median :11.10
                    Mean   :243.3        Mean   :12.64                      Mean   : 8.935                     Mean   :10.97
                    3rd Qu.:278.0        3rd Qu.:14.40                      3rd Qu.: 9.800                     3rd Qu.:12.30
                    Max.   :497.0        Max.   :29.20                      Max.   :29.200                     Max.   :29.20
> summary(V.TWO.SEATER)
 Vehicle.Class      CO2.Emissions..g.km. Fuel.Consumption..City..L.100.km.. Fuel.Consumption..Hwy..L.100.km.. Fuel.Consumption..Comb..L.100.km..
 Length:1470        Min.   :  0.0        Min.   : 4.90                      Min.   : 4.00                      Min.   : 4.50
 Class :character   1st Qu.:242.0        1st Qu.:11.90                      1st Qu.: 8.70                      1st Qu.:10.50
 Mode  :character   Median :278.0        Median :13.85                      Median : 9.70                      Median :12.10
                    Mean   :293.8        Mean   :14.98                      Mean   :10.35                      Mean   :12.89
                    3rd Qu.:339.0        3rd Qu.:17.60                      3rd Qu.:11.80                      3rd Qu.:14.90
                    Max.   :633.0        Max.   :33.30                      Max.   :23.10                      Max.   :27.50
> summary(V.PICKUP.TRUCK)
 Vehicle.Class      CO2.Emissions..g.km. Fuel.Consumption..City..L.100.km.. Fuel.Consumption..Hwy..L.100.km.. Fuel.Consumption..Comb..L.100.km..
 Length:3424        Min.   :  0.0        Min.   : 9.50                      Min.   : 7.10                      Min.   : 8.80
 Class :character   1st Qu.:294.0        1st Qu.:14.70                      1st Qu.:10.90                      1st Qu.:13.00
 Mode  :character   Median :326.0        Median :16.30                      Median :12.30                      Median :14.50
                    Mean   :328.2        Mean   :16.63                      Mean   :12.47                      Mean   :14.76
                    3rd Qu.:359.0        3rd Qu.:18.30                      3rd Qu.:13.70                      3rd Qu.:16.20
                    Max.   :573.0        Max.   :35.00                      Max.   :35.00                      Max.   :35.00
>
```

Based on the correlation matrix generated among the four numerical variables, all features have positive correlation of at least 0.6 or higher towards one another. This needs further verification in later tests.

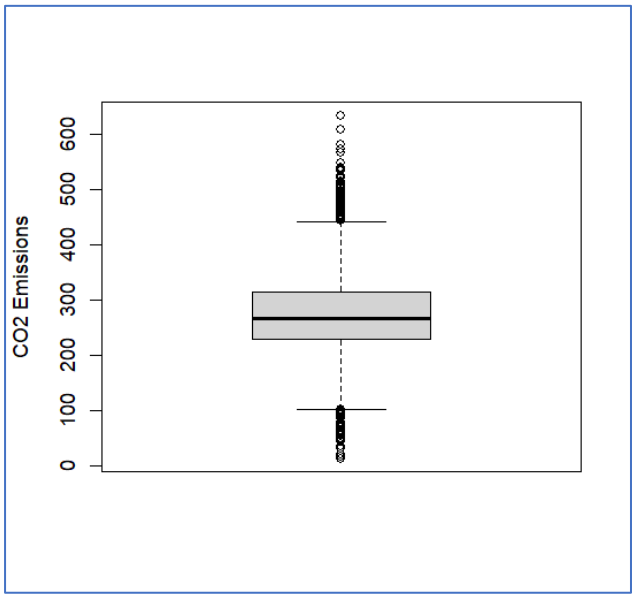


A histogram and density plot was plotted, showing that majority of the data for CO2 are concentrated between 200 – 300 g/km . The density plot further reinforces the fact that the peak distribution is concentrated at the said range.

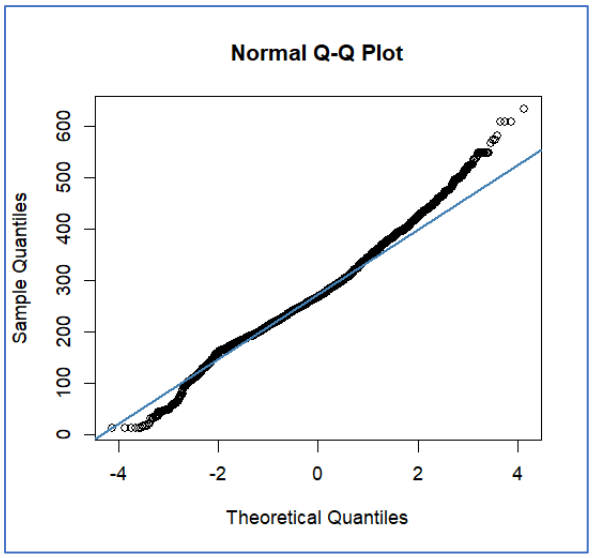




From the box plot generated, we can see that CO2 is skewed to the right, with outliers on both the lower and higher ends of the data.



The Normal Q-Q plot shows that there are some separations from the normal line at both the lower and upper tail ends respectively.



One Sample T-Test – "CO2 Emissions"

```
ttest_result1_LESS <- t.test(VClass$CO2.Emissions..g.km., mu = 0, alternative = "less")
print(t_test_result_LESS)

ttest_result1_GREATER <- t.test(VClass$CO2.Emissions..g.km., mu = 0, alternative = "greater")
print(ttest_result1_GREATER)
```

```
> ttest_result1_LESS <- t.test(VClass$CO2.Emissions..g.km., mu = 0, alternative = "less")
> print(t_test_result_LESS)

        One Sample t-test

data:  VClass$CO2.Emissions..g.km.
t = 551.6, df = 16769, p-value = 1
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
     -Inf 286.4169
sample estimates:
mean of x
 285.5654

> ttest_result1_GREATER <- t.test(VClass$CO2.Emissions..g.km., mu = 0, alternative = "greater")
> print(ttest_result1_GREATER)

        One Sample t-test

data:  VClass$CO2.Emissions..g.km.
t = 551.6, df = 16769, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 284.7138      Inf
sample estimates:
mean of x
 285.5654
```

When performing a left-tailed T-Test (t_test_result_LESS) where null hypothesis is the true mean of CO2 emissions is equal to 0, and the alternative hypothesis is that the true mean is less than 0, the p-value obtained is 1. This suggests that there is no evidence to reject the null hypothesis, thus not enough evidence to conclude that the true mean of CO2 emissions is less than 0.

When performing a right-tailed T-Test (t_test_result_GREATER) where null hypothesis is the true mean of CO2 emissions is equal to 0, and the alternative hypothesis is that the true mean is greater than 0, the p-value obtained is $2.2e^{-16}$. Since the p-value is less than 0.05, it has strong evidence to reject the null hypothesis, as the data suggests that the true mean is greater than 0.

In summary, for the "LESS" alternative hypothesis, there is no evidence to conclude that the true mean of CO2 emissions is less than 0. For the "GREATER" alternative hypothesis, there is strong evidence to conclude that true mean of CO2 emissions is greater than 0.

Two Sample T-Test – "CO2 Emissions" and "FC-Comb"

```
# TWO SAMPLE T-TEST - CO2 Emissions & FC-COMB

# Splitting Fuel Consumption according to two categorical groups High and Low
VClass$Category <- ifelse(VClass$Fuel.Consumption..Comb..L.100.km.. > median(VClass$Fuel.Consumption..Comb..L.100.km..),
                          "HighFuelConsumption", "LowFuelConsumption")

# Perform a t-test comparing CO2 emissions between the two groups
t_test_result <- t.test(CO2.Emissions..g.km. ~ Category, data = VClass)

print(t_test_result)
```

```
> # Splitting Fuel Consumption according to two categorical groups High and Low
> VClass$Category <- ifelse(VClass$Fuel.Consumption..Comb..L.100.km.. > median(VClass$Fuel.Consumption..Comb..L.100.km..),
+                           "HighFuelConsumption", "LowFuelConsumption")
> # Perform a t-test comparing CO2 emissions between the two groups
> t_test_result <- t.test(CO2.Emissions..g.km. ~ Category, data = VClass)
> print(t_test_result)

        Welch Two Sample t-test

data:  CO2.Emissions..g.km. by Category
t = 146.85, df = 15769, p-value < 2.2e-16
alternative hypothesis: true difference in means between group HighFuelConsumption and group LowFuelConsumption is not equal to 0
95 percent confidence interval:
  99.48434 102.17600
sample estimates:
mean in group HighFuelConsumption  mean in group LowFuelConsumption
                         337.2912                          236.4610
```

When testing for the difference in means between groups "High Fuel Consumption" and "Low Fuel Consumption" in terms of their respective CO2 emissions, the t-value is 146.85, degree of freedom (df) is 15769, and p-value is $2.2e^{-16}$. The 95% confidence interval of the true difference in means (99.48434 , 102.17600). The mean in "High Fuel Consumption" group is 337.2912 , and for "Low Fuel Consumption" is 236.4610 . The alternative hypothesis is that the true difference in means between the two groups is not equal to 0.

Based on the data above, it can be concluded that the very low p-value indicates that there is strong evidence to reject the null hypothesis, suggesting that there is a significant difference in means between the two groups. The 95% confidence interval also does not include 0, which further supports that there is a significant difference in means. The mean of CO2 Emissions in "High Fuel Consumption" group (337.2912) is significantly higher than in "Low Fuel Consumption" group (236.4610). These information supports the project hypothesis that driving conditions based on vehicles with higher fuel consumption also have significantly higher CO2 emissions, as compared to vehicles with lower fuel consumption, thus having the greater environmental impact.

ANOVA – Vehicle Types ("SUV", "Full-size", "Mid-Size", "two-seater", and "pickup truck") and its respective CO2 Emissions

```
anova_result <- aov(CO2.Emissions..g.km. ~ Vehicle.Class, data = VClass)
summary(anova_result)
```

```
> anova_result <- aov(CO2.Emissions..g.km. ~ Vehicle.Class, data = VClass)
> summary(anova_result)
                Df   Sum Sq Mean Sq F value Pr(>F)
Vehicle.Class    4 12490457 3122614   832.6 <2e-16 ***
Residuals    16765 62879586    3751
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When testing for the variability in CO2 Emissions across its respective 5 different vehicle types, the degree of freedom (DF) , Sum of Squares, Means Squared, F-statistic, and p-value associated with F-statistic (Pr(>F)) were calculated. The ANOVA yielded a high F-statistic and a p-value < 0.001, indicating a statistically significant difference in mean CO2 emissions across vehicle classes. This supports the project hypothesis that different vehicle class have different CO2 emissions, which contribute to varying environmental impact.

ANOVA – Vehicle driving conditions (City & Highway) and its respective CO2 Emissions

```
# ANOVA on "FC-COMB" and its CO2 Emissions
anova_result2 <- aov(VClass$CO2.Emissions..g.km. ~ VClass$Fuel.Consumption..Comb..L.100.km.., data = VClass)
summary(anova_result2)
```

```
> # ANOVA on "FC-COMB" and its CO2 Emissions
> anova_result2 <- aov(VClass$CO2.Emissions..g.km. ~ VClass$Fuel.Consumption..Comb..L.100.km.., data = VClass)
> summary(anova_result2)
                                      Df   Sum Sq  Mean Sq F value Pr(>F)
VClass$Fuel.Consumption..Comb..L.100.km..     1 62926365 62926365   84794 <2e-16 ***
Residuals                             16768 12443678      742
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

The conclusion of the interpretation of the above results are that the Pr(>F) value is less than 0.001, which indicates strong evidence that the null hypothesis should be rejected. The results also show that there are significant difference in the fuel consumption across different vehicle classes.

Linear Regression – Vehicle Class and Its CO2 Emissions

```
# LinearRegression on "Vehicle Class" and its CO2 Emissions
LinearRegression <- lm(VClass$CO2.Emissions..g.km. ~ Vehicle.Class, data = VClass)
summary(LinearRegression)
```

```
> # LinearRegression on "Vehicle Class" and its CO2 Emissions
> LinearRegression <- lm(VClass$CO2.Emissions..g.km. ~ Vehicle.Class, data = VClass)
> summary(LinearRegression)

Call:
lm(formula = VClass$CO2.Emissions..g.km. ~ Vehicle.Class, data = VClass)

Residuals:
    Min      1Q  Median      3Q     Max 
-242.01  -40.64   -1.64   34.99  336.79 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                279.480      1.463 191.068  < 2e-16 ***
Vehicle.ClassMID-SIZE      -32.114      1.751 -18.344  < 2e-16 ***
Vehicle.ClassPICKUP TRUCK   49.164      1.799  27.328  < 2e-16 ***
Vehicle.ClassSUV             6.527      1.660   3.932 8.47e-05 ***
Vehicle.ClassTWO-SEATER     16.732      2.171   7.708 1.35e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.24 on 16765 degrees of freedom
Multiple R-squared:  0.1657,	Adjusted R-squared:  0.1655 
F-statistic: 832.6 on 4 and 16765 DF,  p-value: < 2.2e-16
```

In the above Linear Regression model results for Vehicle Class and its CO2 Emissions, the “FULL-SIZE” vehicle class represents the “intercept”. The variability of residuals range from -242.01 to 336.79 grams, suggesting a huge variation. In terms of CO2 emissions based on each vehicle class, “MID-SIZE” vehicles produce 32.114 grams lesser compared to baseline emissions of a “FULL-SIZE” vehicle. Whereas for PICKUP TRUCK, SUV, and TWO-SEATER is 49.164 grams, 6.527 grams, and 16.732 grams higher as compared to baseline emissions of “FULL-SIZE” respectively. The Pr(>|t|) values are all lesser than 0.001, which indicates that the null hypothesis should be rejected, and there is evidence that there are significant differences in CO2 emissions among all vehicle classes. However, due to the very low R-squared value of 0.1655 , this indicates that there are other factors not included that can explain the CO2 emissions variability. Further analysis of other features need to be considered. The next variable that will be examined is the Fuel Consumption – City & Highway Combined (FC-COMB) with its respective CO2 emissions.

Linear Regression - Fuel Consumption – City & Highway Combined (FC-COMB) with its respective CO2 emissions.

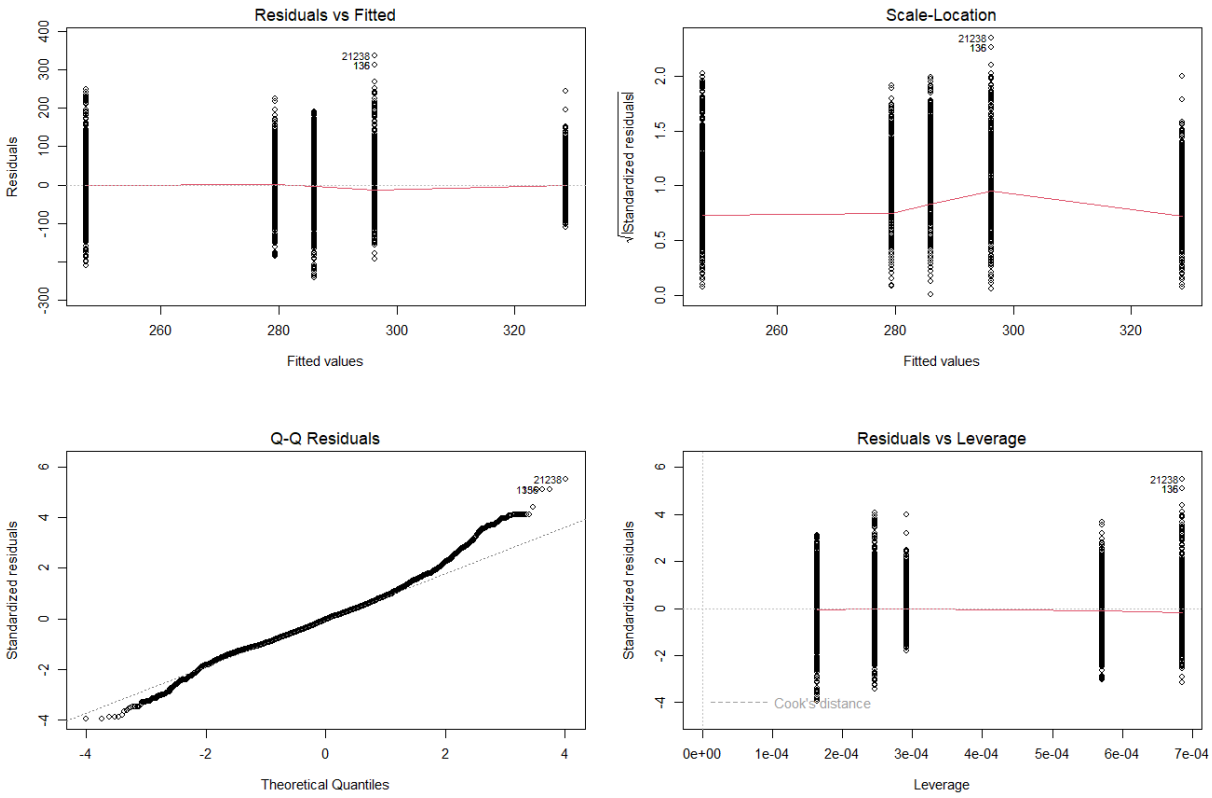
```
# Linear Regression on "FC-COMB" and its CO2 Emissions
LinearRegression2 <- lm(VClass$CO2.Emissions..g.km. ~ VClass$Fuel.Consumption..Comb..L.100.km.., data = VClass)
summary(LinearRegression2)


model1.full <- lm(VClass$CO2.Emissions..g.km. ~
                    VClass$Fuel.Consumption..City..L.100.km.. +
                    VClass$Fuel.Consumption..Hwy..L.100.km.. +
                    VClass$Fuel.Consumption..Comb..L.100.km.,
                  data = VClass)
```

```
> # Linear Regression on "FC-COMB" and its CO2 Emissions
> LinearRegression2 <- lm(VClass$CO2.Emissions..g.km. ~ VClass$Fuel.Consumption..Comb..L.100.km.., data = VClass)
> summary(LinearRegression2)

Call:
lm(formula = VClass$CO2.Emissions..g.km. ~ VClass$Fuel.Consumption..Comb..L.100.km..,
    data = VClass)

Residuals:
     Min       1Q   Median       3Q      Max
-153.780   -2.282    3.883   12.057   85.066

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             40.19609    0.86849   46.28   <2e-16 ***
VClass$Fuel.Consumption..Comb..L.100.km.. 19.44205    0.06677  291.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.24 on 16768 degrees of freedom
Multiple R-squared:  0.8349,	Adjusted R-squared:  0.8349
F-statistic: 8.479e+04 on 1 and 16768 DF,  p-value: < 2.2e-16
```

Based on the linear regression on FC-COMB with its respective CO2 emissions (represents the intercept), the variability of residuals range from -153.780 to 85.066 grams, suggesting a huge variation. In terms of CO2 emissions based on the fuel consumption of combined driving conditions, the estimated coefficient for fuel consumption is 19.44205, which indicates that for one-unit fuel consumption increase, there is 19.44205 grams change in CO2 emission. The Pr(>|t|) values are all lesser than 0.001, which indicates that the null hypothesis should be rejected, and there is evidence that there are significant differences in CO2 emissions with its driving conditions. This indicates that 83.49% of the variance in CO2 emissions ($R^2 = 0.8349$), suggesting a strong linear relationship with combined fuel consumption.

Diagnostic Plots – Linearity, Normality, Homogeneity of Variance

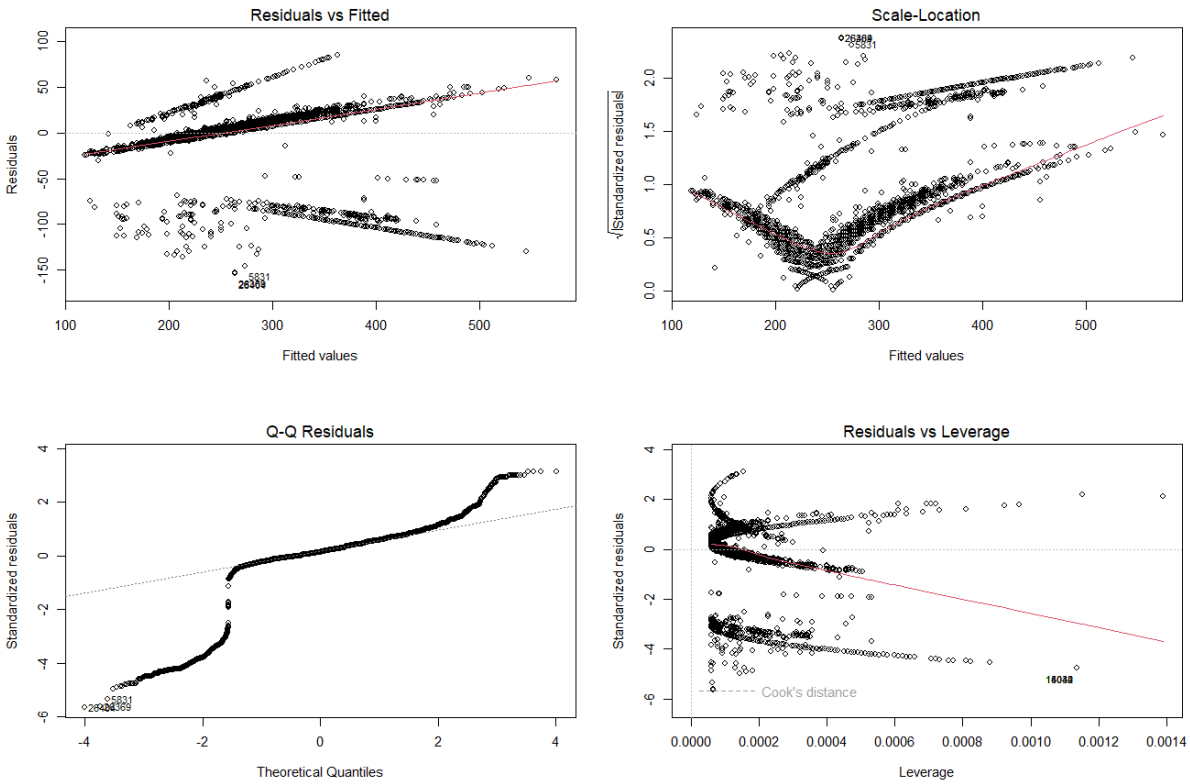
Linear Regression # 1 – Vehicle Class and its CO2 Emissions



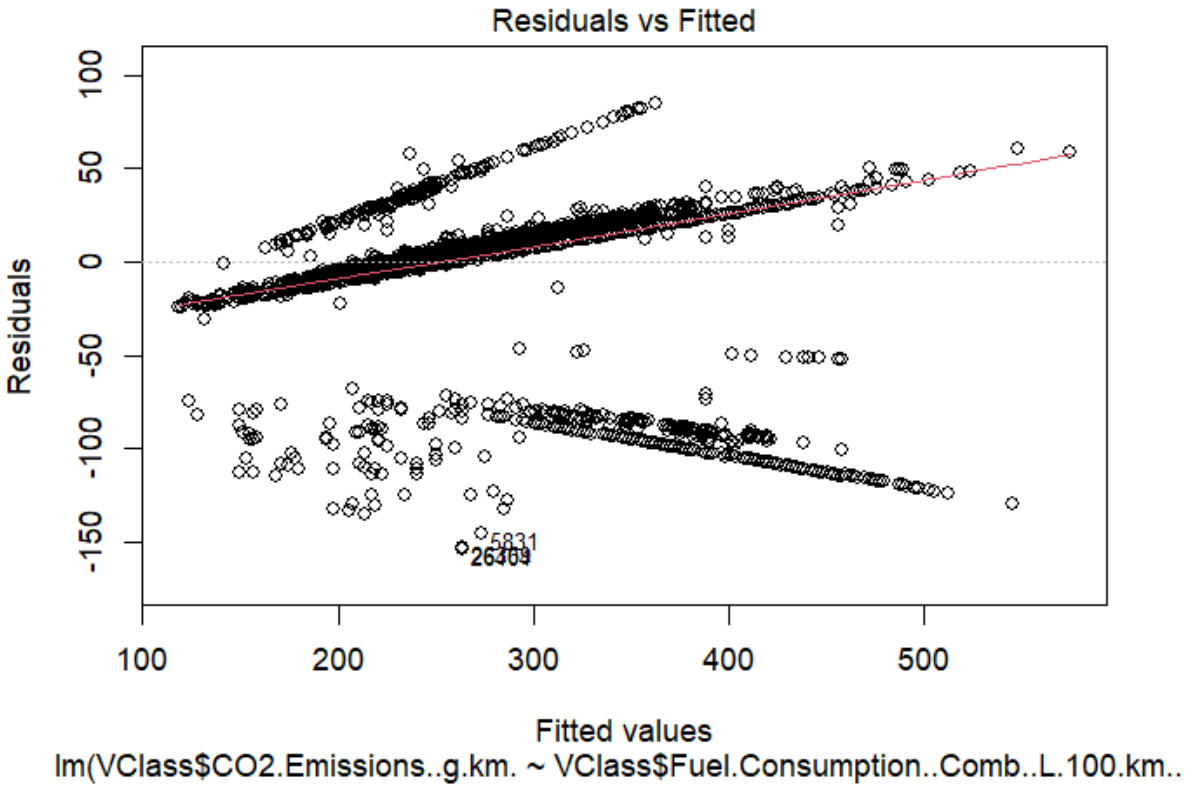
At a glance of the summary of results, there are vertical lines in Residuals vs Fitted, Residuals vs Leverage, and Scale-Location plots, which indicates possible linearity and homogeneity of variance issues with this linear regression model. The Q-Q Plot shows that this linear regression model has diversion in its head and tail ends, indicating normality issues with this model. The possible cause may be due to the presence of non-linear relationships, heteroscedasticity, multicollinearity and/or mis-specified model .

Therefore, this Vehicle Class and its CO2 Emissions variables may not be the accurate variables that describe the relationship between the features, as it does not meet the assumptions of linearity, normality, and homogeneity of variance.

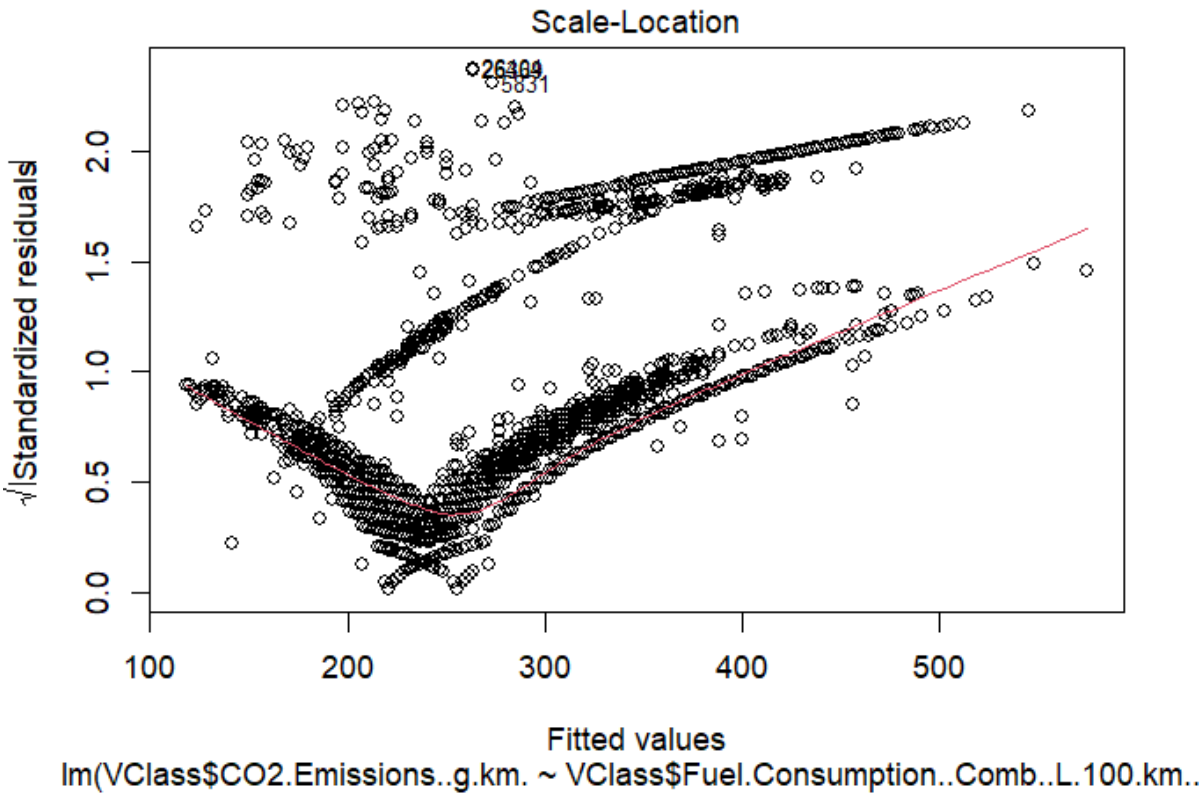
Linear Regression # 2 – Fuel Consumption – Combined (City & Highway) and its CO2 Emissions



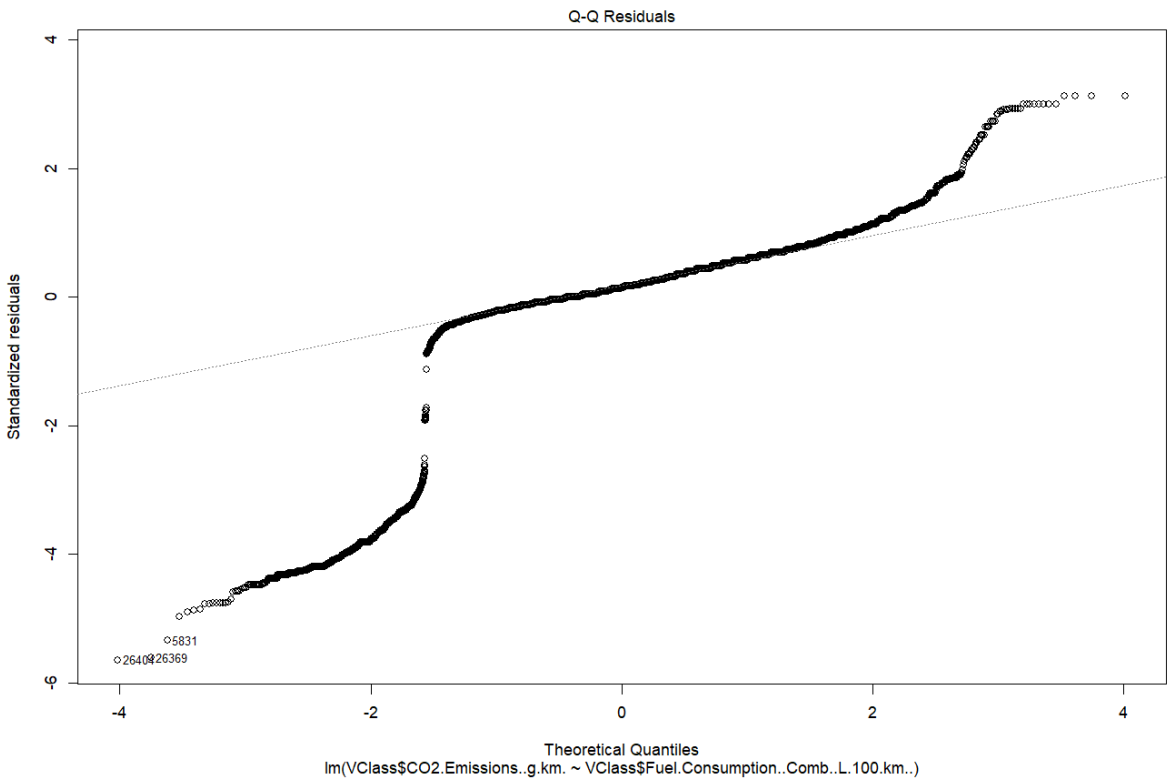
A quick glance at the general overview of plots indicate an improvement compared to the previous plot of Vehicle Class and its CO2 Emissions.



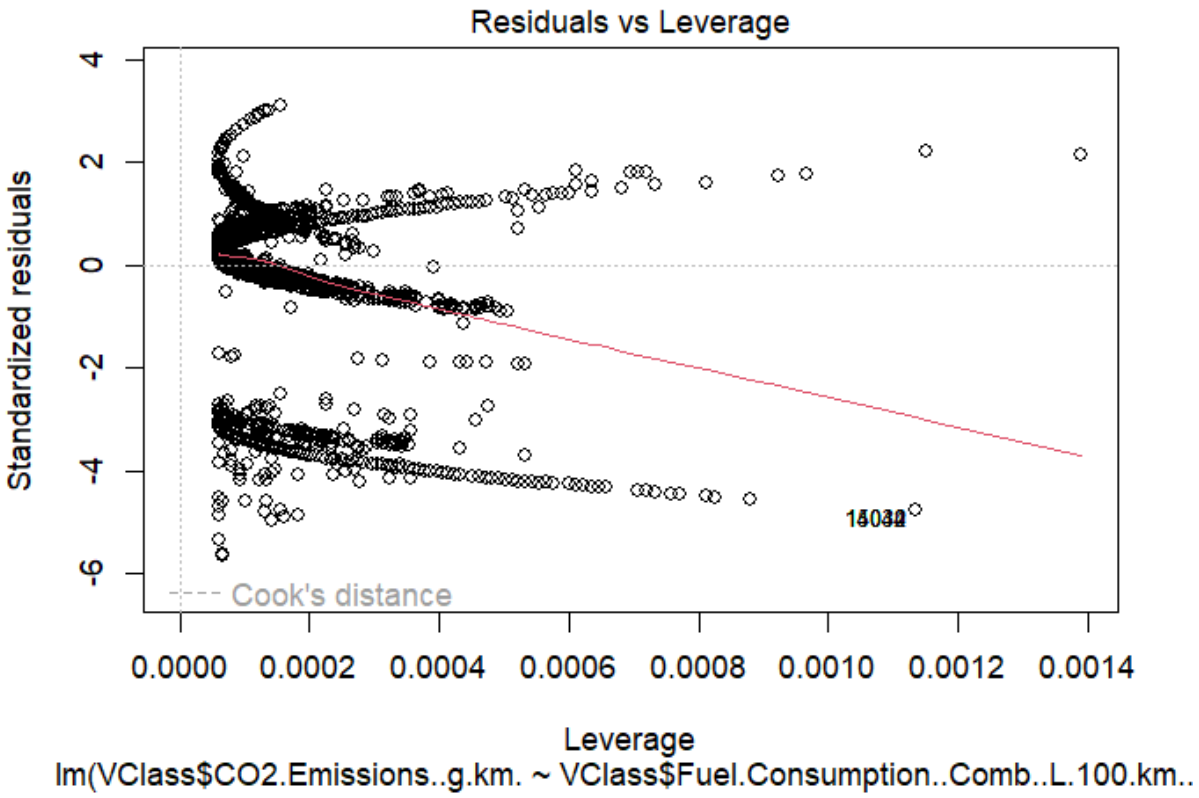
Based on the interpretation of this Residuals vs Fitted plot, there is a clear density of data residuals following the red line, indicating a potential trend and linearity. However, due to the presence of a downward-like trend of residuals in the lower-right quartile of the plot, this indicates that there are outliers, as well as a potential for identifying a trend through further analysis.



In the above Scale-Location plot, majority of the data points are situated close to the red line, which may potentially indicate constant variance of residuals. However, there are still some residual points that are far from the red line, which may potentially indicate heteroscedasticity or presence of outliers. Further analysis and data cleaning may help to improve this issue.



In the above Q-Q Residuals Plot, it indicates that there is a huge deviation from the line at both the head-end and the tail-end, which indicates that there are heavy numbers of outliers. At just a glance of the plot, it seems that the dataset is left-skewed. This Q-Q Residuals Plot indicates that this linear regression model may not be perfectly normally distributed. Possible solutions include further data transformation and cleaning before running the model again.



In this Residuals vs Leverage Plot, it seems that some points lie close to the red line, however, majority of the data residuals are away from the red line. This suggests that the model may not adequately capture the underlying relationship due to potential outliers or specification errors. This could be due to outliers or model misspecification. Potential solutions to this are re-examining data points whether there are presence of outliers and perform necessary data transformation before re-generating the linear regression model.

As a summary, this Fuel Consumption – Combined (City & Highway) and its CO2 Emissions diagnostic plots indicate some levels of linearity and homoscedasticity in the Residuals Vs Fitted, and Scale-Location Plots, although it may further be improved through further data examination and transformation. There are issues in normality and homogeneity of variance as indicated in the Q-Q and Residuals Vs Leverage Plots, which indicates that the model possibly do not fit the assumptions of normality and homogeneity of variance. Compared to the first linear regression model, this model is heavily preferred .

Conclusion

This study provides strong evidence of a significant relationship between driving conditions and CO2 emissions. Specifically, higher fuel consumption correlates with increased emissions, and vehicle class significantly affects environmental impact.

There is a significant relationship between fuel consumption of the vehicles and its CO2 emissions, whereby the greater the amount of fuel consumption, the higher the CO2 emissions. This relationship is also shown when examining different Vehicle Class and its CO2 emissions, whereby according to ranking from highest to lowest fuel consumption of vehicle classes, PICKUP TRUCKS emits the most CO2 per one unit of fuel consumed, followed by TWO-SEATERS, SUVs, and lastly MID-SIZE vehicles.

In terms of model performance, Linear Regression # 2 – Fuel Consumption – Combined (City & Highway) and its CO2 Emissions outperform Linear Regression # 1 – Vehicle Class and its CO2 Emissions. Linear Regression # 2 has greater R-squared value, and the assumption of Linearity, Normality, Homogeneity of Variance are significantly better than that of Linear Regression # 1.