Assignment Part 1 (HD 88%)

Statistical Computing (Swinburne University of Technology)

Question 1

1. Discuss all the aspects of the dataset including the context and properties. [5 marks]

- Context:
    - Weight versus age of chicks on different diets
        - The data is on the growth curves of chicks. The growth curves differ possibly based on their diet.
    - Details
        - The body weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. There were four groups on chicks on different protein diets.
        - This dataset was originally part of package nlme, and that has methods (including for [, as.data.frame, plot andprint) for its grouped-data classes.

- An object of class **c("nfnGroupedData", "nfGroupedData", "groupedData", "data.frame")** containing the following columns:
    - weight
        - a numeric vector giving the body weight of the chick (gm).
    - Time
        - a numeric vector giving the number of days since birth when the measurement was made.
    - Chick
        - an ordered factor with levels 1 < ... < 50 giving a unique identifier for the chick. The ordering of the levels groups chicks on the same diet together and orders them according to their final weight (lightest to heaviest) within diet.
    - Diet
        - a factor with levels 1, ..., 4 indicating which experimental diet the chick received.

2. Look at carefully the variable and discuss any inconsistencies dataset has. Explain your reasoning and the steps you have taken. [5 marks]

- Time interval of weight recording per chick
    - Each of the 50 chicks should have their weight recorded once every 2 days from day 0 until day 20, with the last record on day 21.

- Using **tapply(ChickWeight$Time, ChickWeight$Chick, FUN=var)**
    - The variance of each chick is displayed.
    - If each chick has their weight recorded as per the time interval above, the variance for each chick should be 50.08333.

- Using **tapply(ChickWeight$Time, ChickWeight$Chick, FUN=function(x)diff(range(x)))**
    - The last day of the recording of the chick's weight is displayed.

  - If each chick has their weight recorded as per the time interval above, the last day of weight recording should be day 21.

- Using **summary(ChickWeight$Chicks)**
  - The total number of days each chick had their weight recorded.
  - If each chick has their weight recorded as per the time interval above, the total number of recordings would be 12.

- Observing the outputs of the 3 methods used, inconsistencies in dataset have been found for the following chicks:

	Chick 8	Chick 15	Chick 16	Chick 18	Chick 44
tapply(ChickWeight$Time, ChickWeight$Chick, FUN=var)	44.00000	24.00000	18.66667	2.00000	36.66667
tapply(ChickWeight$Time, ChickWeight$Chick, FUN=function(x)diff(range(x)))	20	14	12	2	18
summary(ChickWeight$Chicks)	11	8	7	2	10

Therefore,

- Chick 8 has no record of its weight on day 21.
- Chick 15 has no record of its weight on days 16, 18, 20, 21.
- Chick 16 has no record of its weight on days 14, 16, 18, 20, 21.
- Chick 18 has no record of its weight on days 4, 6, 8, 10, 12, 14 16, 18, 20, 21.
- Chick 44 has no record of its weight on days 20, 21.

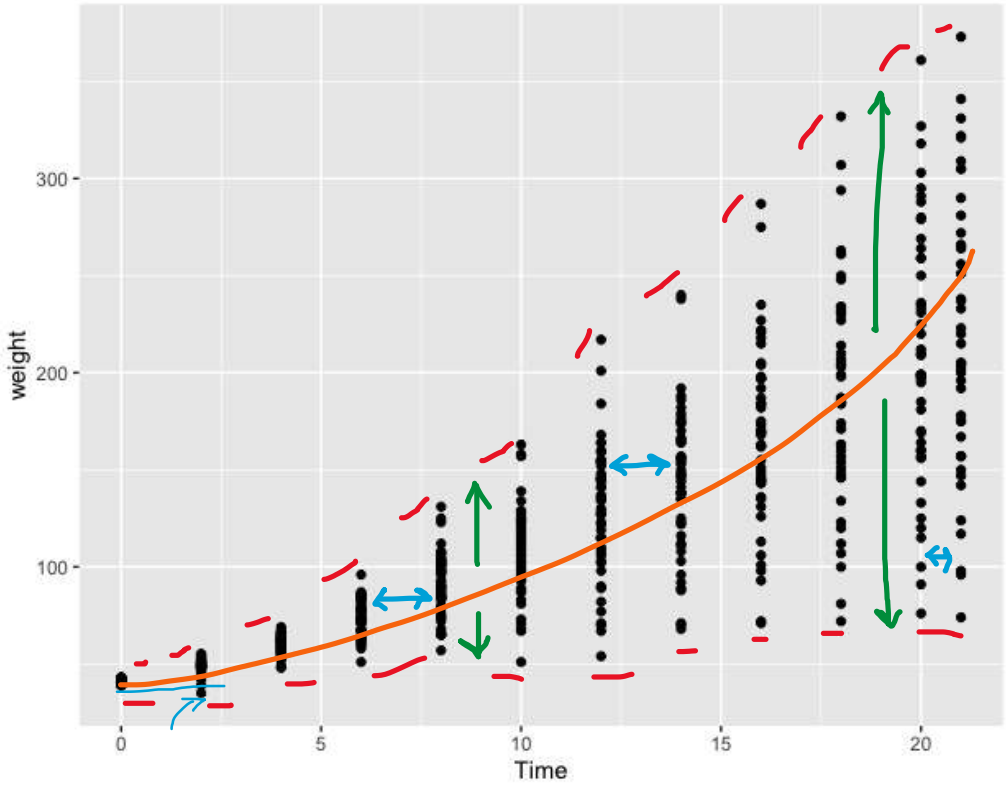
- Another inconsistency in the time interval is that the body weights of the chicks were measured at regular intervals from day 0 to 20, but not at the end from day 20 to 21.

3. Create appropriate summary statistics for each of the four variables. [5 marks]

```
> summary(df)
     weight           Time           Chick     Diet
 Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
 1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
 Median :103.0   Median :10.00   20     : 12   3:120
 Mean   :121.8   Mean   :10.72   10     : 12   4:118
 3rd Qu.:163.8   3rd Qu.:16.00   17     : 12
 Max.   :373.0   Max.   :21.00   19     : 12
                                 (Other):506
```

4. Create appropriate plot for each of the four variables and create plots to see any association between variables and discuss. [5 marks]

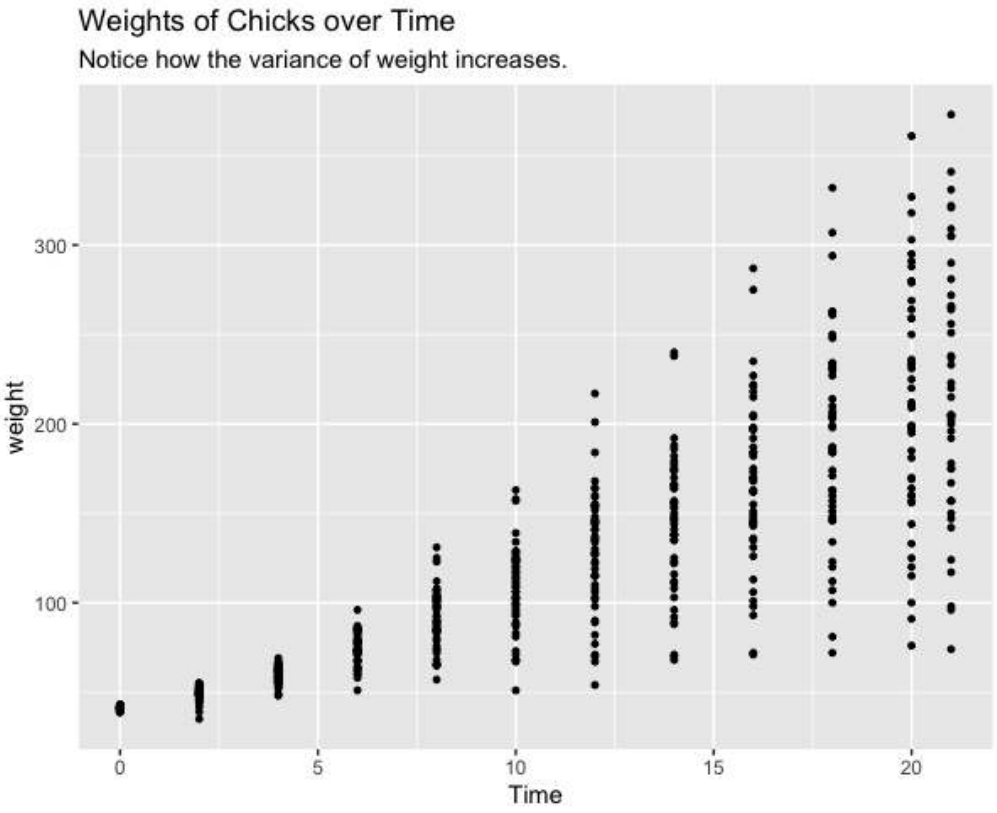
ggplot() + geom_point(data = ChickWeight, aes(x=Time, y=weight))



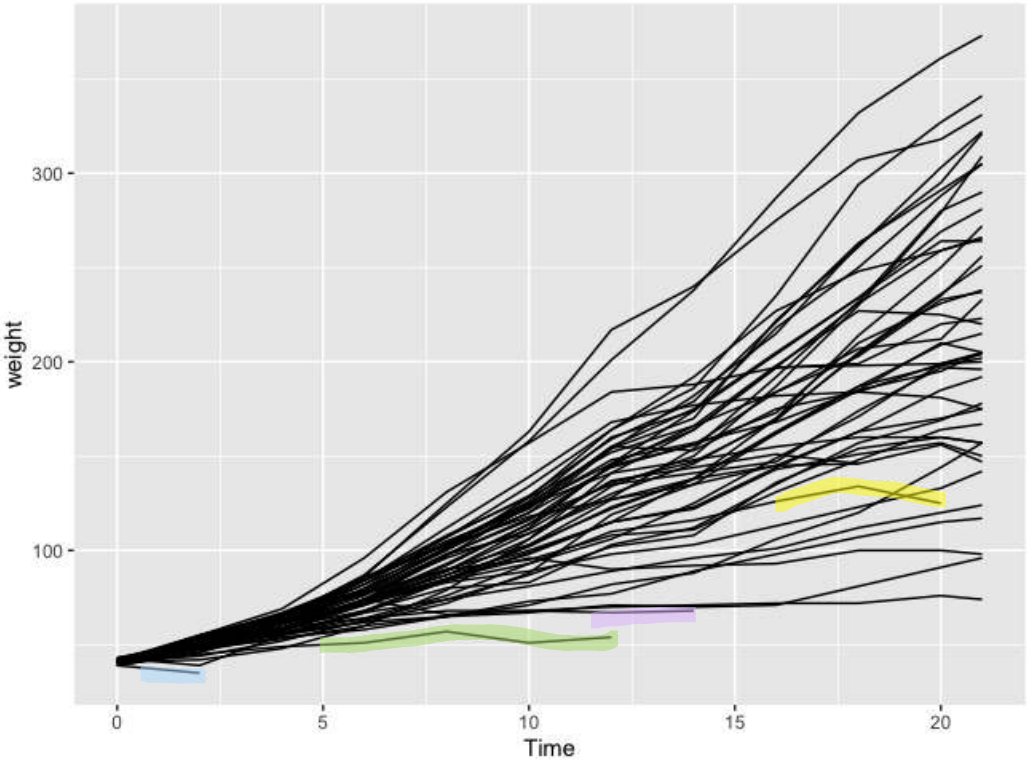
Observations:

- There is a regular 2-day interval from day 0-20 but a 1-day interval from day 20-21.
- On average, the chicks do seem to be increasing in weight overtime (i.e., there is an increasing trend).
- The spread is increasing over time.
- At the end of the timeframe, there is a significant change in variance than in the beginning.
- There seems to be weight loss in one of the chicks.

ggplot() + geom_point(data = ChickWeight, aes(x=Time, y=weight), size=1) + ggtitle("Weights of Chicks over Time", subtitle = "Notice how the variance of weight increases.")



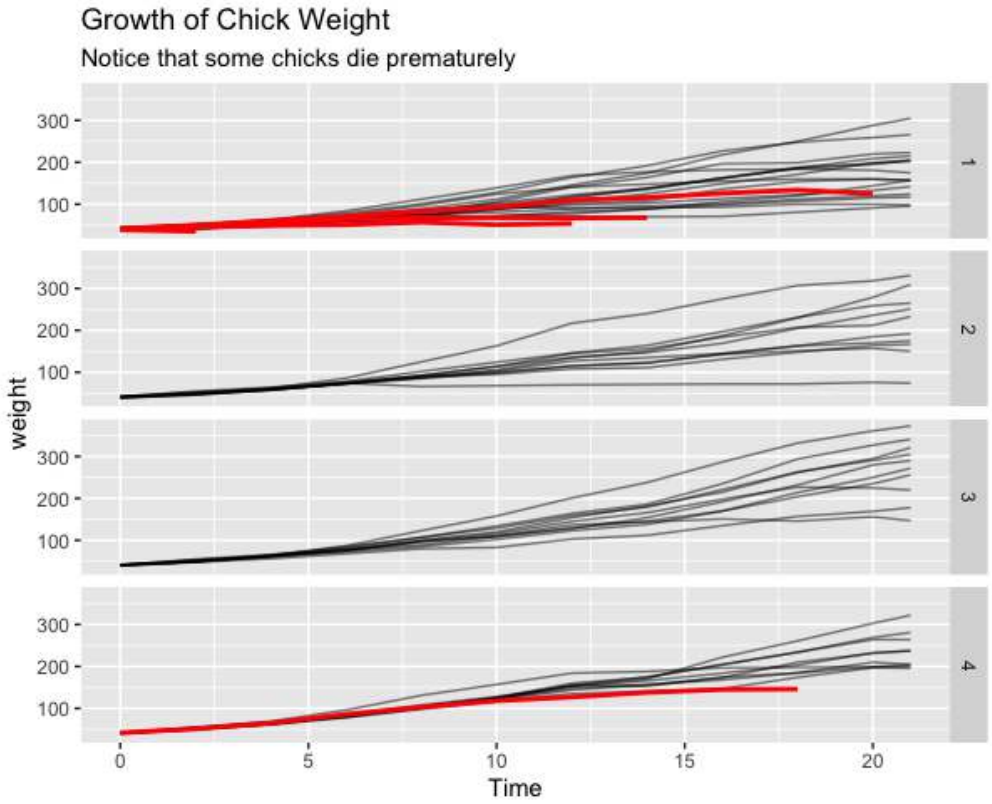
ggplot() + geom_line(data = ChickWeight, aes(x=Time, y=weight, group = Chick))



Observations:

- 4 chicks whose weights were not fully recorded for 12 days.

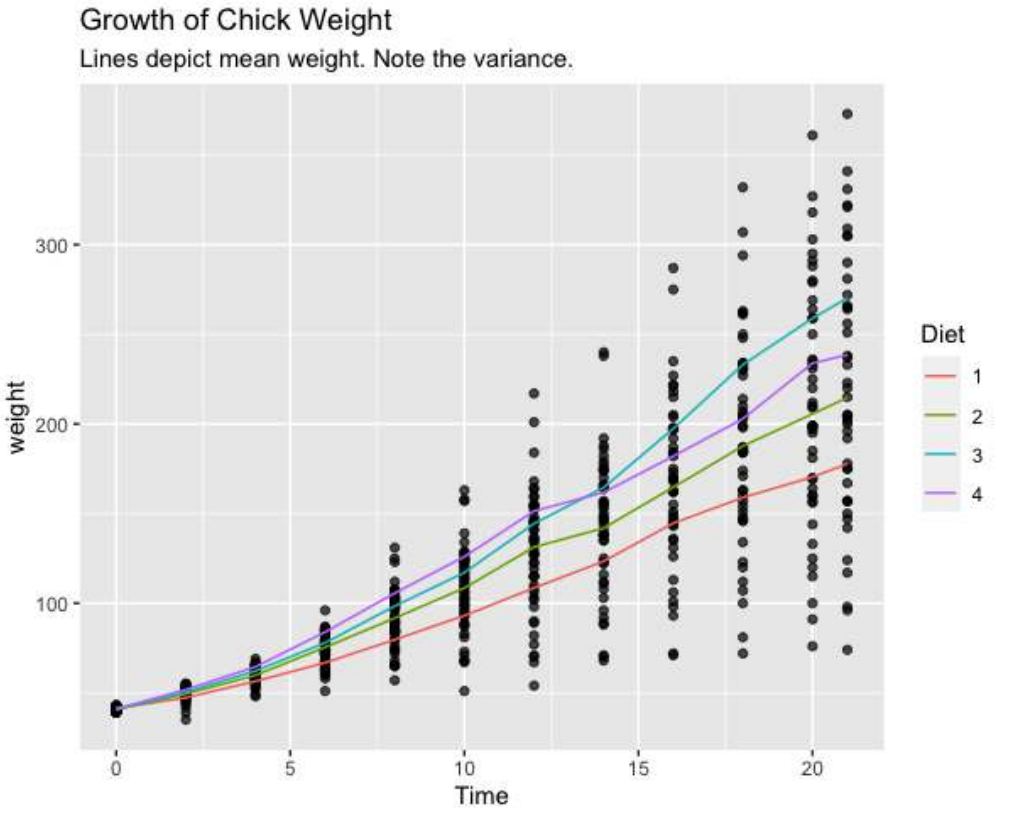
```
  # Create a subset data frame
  sub_df <- ChickWeight %>%
  group_by(Chick) %>%
  mutate(max_time = max(Time)) %>%
  ungroup() %>%
  filter(max_time < max(Time))

# Make a chart that highlights the subset
ggplot() +
  geom_line(data=ChickWeight, aes(x=Time, y=weight, group=Chick), alpha=0.5) +
  geom_line(data=sub_df, aes(x=Time, y=weight, group=Chick), color="red", size=1) +
  facet_grid(Diet ~ .) +
  ggtitle("Growth of Chick Weight",
          subtitle = "Notice that some chicks die prematurely")
```



Observations:

- There are 5 chicks that die prematurely from consuming diet 1 or 4.
- Diet 1 has caused the most premature deaths.

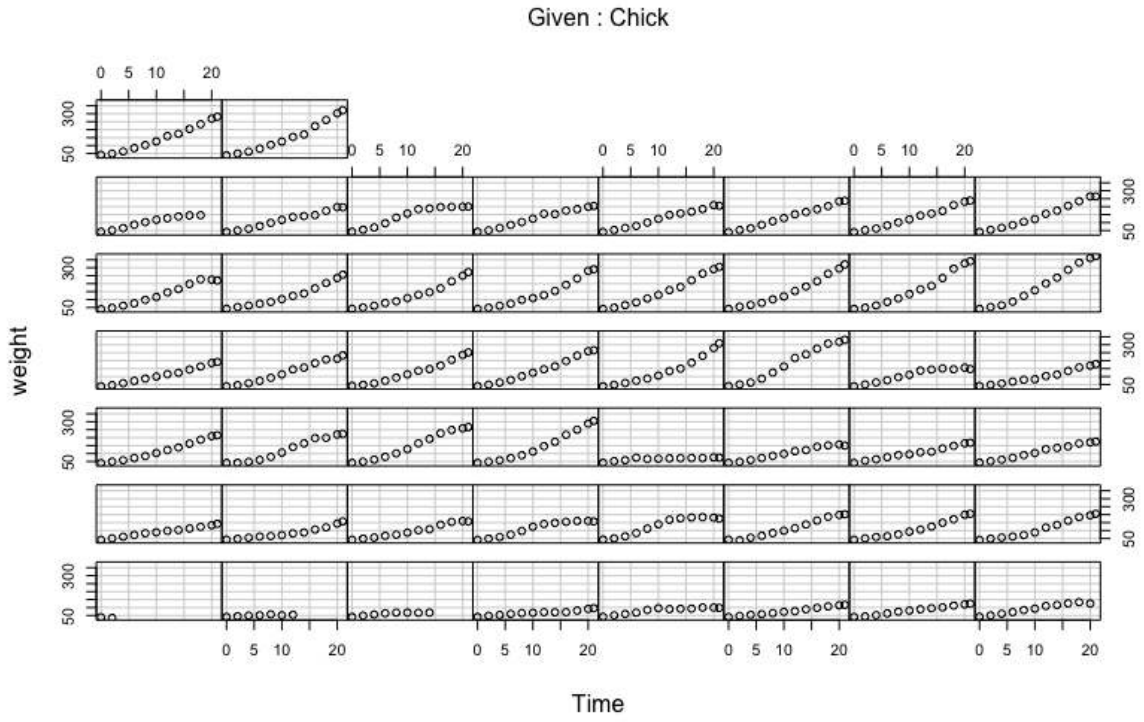
```r
#Final plot
agg_df <- ChickWeight %>%
  group_by(Diet, Time) %>%
  summarise(m = mean(weight))

ggplot() +
  geom_point(data=ChickWeight, aes(x=Time, y=weight), alpha=0.7) +
  geom_line(data=agg_df, aes(x=Time, y=m, color=Diet)) +
  ggtitle("Growth of Chick Weight",
          subtitle="Lines depict mean weight. Note the variance.")
```



Observations:

- On average, diet 3 increases the chick weights the most.
- There still is a significant variance in weights.

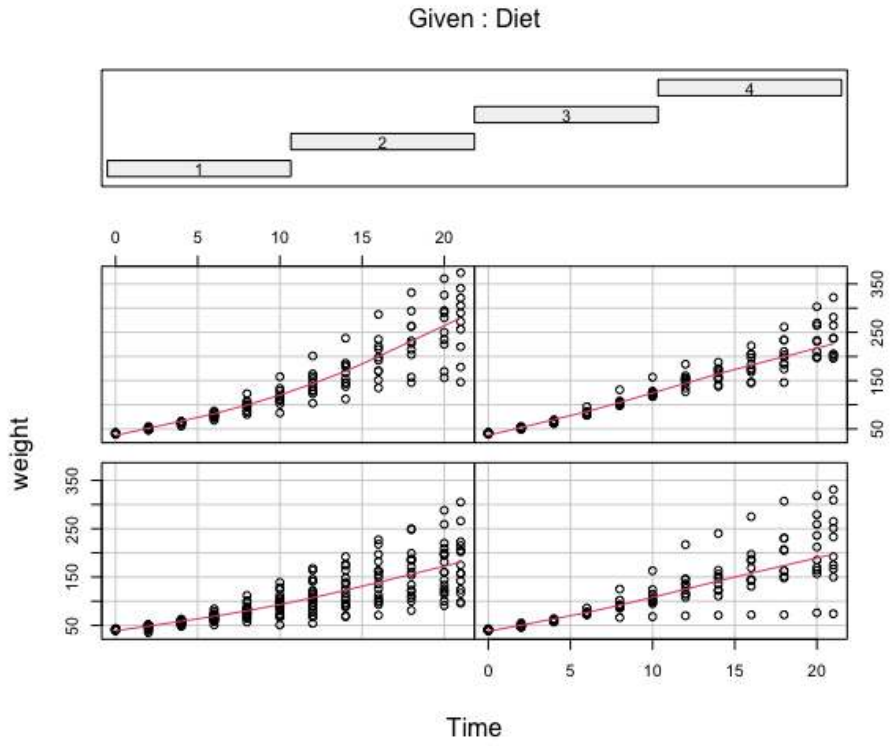
coplot(weight ~ Time | Chick, data = ChickWeight, type = "b", show.given = FALSE)



Observations:

- Majority of chicks increase in weight as they grow overtime.
- Some chicks die prematurely.

coplot(weight ~ Time | Diet, data = ChickWeight, panel = panel.smooth)



Observations:

- Diet 3 results in the largest weight growth of chicks.
- Diet 1 results in the smallest weight growth of chicks.

Question 2

1. Read the dataset in R, obtain the structure of the dataset, and discuss. [3 marks]

```
> str(data)
'data.frame':   101 obs. of  16 variables:
 $ ID     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ AGE    : int  18 23 39 24 27 26 26 26 28 24 ...
 $ GENDER : chr  "Male" "Female" "Male" "Male" ...
 $ YRSTUDY: int  3 7 10 6 16 16 10 8 9 6 ...
 $ Q1     : int  8 15 9 10 10 10 14 12 9 10 ...
 $ Q2     : int  14 21 12 15 20 16 18 17 15 13 ...
 $ Q3     : int  16 20 14 15 21 16 16 11 21 23 ...
 $ Q4     : chr  "19" "23" "12" "18" ...
 $ Q5     : chr  "18" "19" "22" "17" ...
 $ Q6     : int  14 16 20 15 23 18 11 10 23 19 ...
 $ Q7     : int  21 20 16 20 29 20 15 18 21 24 ...
 $ Q8     : int  27 22 22 29 26 27 27 21 23 23 ...
 $ Q9     : int  25 24 26 30 27 24 25 27 20 26 ...
 $ Q10    : int  26 26 26 25 25 25 25 25 25 25 ...
 $ Q11    : int  22 18 23 27 21 25 24 24 26 19 ...
 $ Q12    : int  14 17 15 13 24 14 13 10 22 18 ...
```

- The effects on each student's scaled scores for each question varies by their year of study, age and possibly gender.
- There are 101 students (rows) and 16 variables (columns).
- There are 13 numeric variables. They are ID, AGE, YRSTUDY, Q1-3, 6-12.
- There are 3 nominal variables. They are GENDER, Q4 and 5.

2. Discuss and report any missing values and unusual characters in the dataset. [5 marks]

```
#specify column
i <- c(8, 9)

# Specify own function within apply
data[ , i] <- apply(data[ , i], 2,
                    function(x) as.integer(x))
df = data.frame(data)
str(data)
View(df)
summary(df)
```

```
> summary(df)
       ID              AGE            GENDER             YRSTUDY            Q1              Q2              Q3       
 Min.   :  1   Min.   :18.00   Length:101         Min.   : 0.000   Min.   : 7.00   Min.   : 9.00   Min.   :10.00  
 1st Qu.: 26   1st Qu.:19.00   Class :character   1st Qu.: 3.000   1st Qu.: 9.00   1st Qu.:13.00   1st Qu.:15.00  
 Median : 51   Median :23.00   Mode  :character   Median : 6.000   Median :10.00   Median :15.00   Median :18.00  
 Mean   : 51   Mean   :23.35                      Mean   : 6.634   Mean   :10.94   Mean   :15.61   Mean   :18.02  
 3rd Qu.: 76   3rd Qu.:26.00                      3rd Qu.: 9.000   3rd Qu.:12.00   3rd Qu.:19.00   3rd Qu.:21.00  
 Max.   :101   Max.   :39.00                      Max.   :20.000   Max.   :19.00   Max.   :28.00   Max.   :30.00  
                                                                                   NA's   :1                      
       Q4              Q5              Q6              Q7              Q8              Q9              Q10       
 Min.   : 9.00   Min.   : 10.00   Min.   : 9.00   Min.   :13.00   Min.   :19.00   Min.   :16.00   Min.   :22.00  
 1st Qu.:15.75   1st Qu.: 15.00   1st Qu.:13.00   1st Qu.:16.00   1st Qu.:24.00   1st Qu.:24.00   1st Qu.:23.00  
 Median :20.00   Median : 18.00   Median :16.00   Median :20.00   Median :26.00   Median :26.00   Median :24.00  
 Mean   :19.11   Mean   : 19.23   Mean   :16.19   Mean   :20.29   Mean   :25.45   Mean   :25.59   Mean   :23.73  
 3rd Qu.:23.00   3rd Qu.: 21.00   3rd Qu.:19.00   3rd Qu.:22.00   3rd Qu.:28.00   3rd Qu.:27.00   3rd Qu.:25.00  
 Max.   :30.00   Max.   :120.00   Max.   :26.00   Max.   :52.00   Max.   :32.00   Max.   :33.00   Max.   :26.00  
 NA's   :1       NA's   :1                                                                       NA's   :1      
      Q11             Q12       
 Min.   :13.00   Min.   :10.00  
 1st Qu.:21.00   1st Qu.:15.00  
 Median :23.00   Median :17.50  
 Mean   :23.01   Mean   :17.61  
 3rd Qu.:25.00   3rd Qu.:20.00  
 Max.   :52.00   Max.   :26.00  
 NA's   :1       NA's   :1      
```

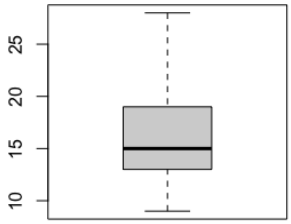
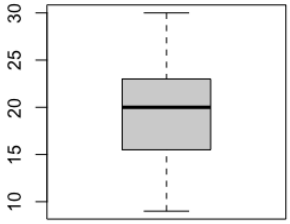
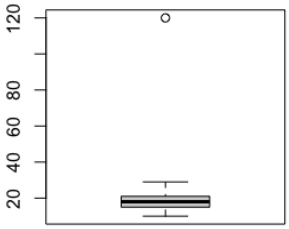
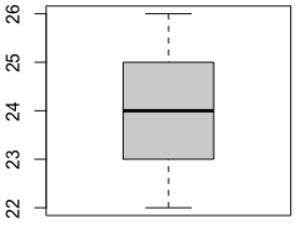
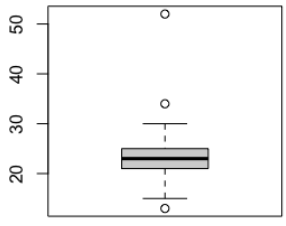
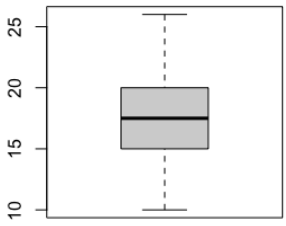
- There are 6 missing values indicate as NA, one each in Q2, 4, 5, 10, 11 and 12.
- As the variables Q1-12 are measured in 0-40 scales, it is unusual that:
    - The maximum values are 120.00 and 52.00 for Q5 and Q7 and 11 respectively.
    - The initial values for Q4 and 5 were characters instead of integers.

3. Replace unusual values and missing values if exists, in the dataset with NA. [5 marks]

```
> # Return the column names containing missing observations
> list_na <- colnames(df)[ apply(df, 2, anyNA) ]
> list_na
[1] "Q2"  "Q4"  "Q5"  "Q10" "Q11" "Q12"
# Replace missing/unusual values with NA
list_na[is.na(list_na)] = NA
```

4. Impute missing values with appropriate statistic (mean or median) for each variable and explain why you have chosen that particular statistic. [7 marks]

```
# Create a new variable with the mean and median
df_replace <- df %>%
  mutate(replace_mean_Q2  = ifelse(is.na(Q2), average_missing[1], Q2),
         replace_mean_Q4  = ifelse(is.na(Q4), average_missing[2], Q4),
         replace_mean_Q5  = ifelse(is.na(Q5), average_missing[3], Q5),
         replace_mean_Q10 = ifelse(is.na(Q10), average_missing[4], Q10),
         replace_mean_Q11 = ifelse(is.na(Q11), average_missing[5], Q11),
         replace_mean_Q12 = ifelse(is.na(Q12), average_missing[6], Q12))


head(df_replace)

# Create box plot of mean to check for outliers
boxplot(df$Q2)
boxplot(df$Q4)
boxplot(df$Q5)
boxplot(df$Q10)
boxplot(df$Q11)
boxplot(df$Q12)
```

Q2	Q4	Q5
		
Q10	**Q11**	**Q12**
		

- Q5 and 11 will be imputed with their median because otherwise the value of their mean would be dominated by the outliers rather than the typical values. Q2, 4, 10 and 12 will be imputed with their mean as they have no outliers.

```
> # Replace NA with mean
> df1$Q2[is.na(df1$Q2)] <- mean(df1$Q2, na.rm = TRUE)
> summary(df1$Q2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   13.00   15.00   15.61   19.00   28.00
> summary(df$Q2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   9.00   13.00   15.00   15.61   19.00   28.00       1
> df1$Q4[is.na(df1$Q4)] <- mean(df1$Q4, na.rm = TRUE)
> summary(df1$Q4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   16.00   20.00   19.11   23.00   30.00
> summary(df$Q4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   9.00   15.75   20.00   19.11   23.00   30.00       1
> df1$Q10[is.na(df1$Q10)] <- mean(df1$Q10, na.rm = TRUE)
> summary(df1$Q10)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.00   23.00   24.00   23.73   25.00   26.00
> summary(df$Q10)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  22.00   23.00   24.00   23.73   25.00   26.00       1
> df1$Q12[is.na(df1$Q12)] <- mean(df1$Q12, na.rm = TRUE)
> summary(df1$Q12)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   15.00   17.61   17.61   20.00   26.00
> summary(df$Q12)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  10.00   15.00   17.50   17.61   20.00   26.00       1
> # Replace NA with median
> df1$Q5[is.na(df1$Q5)] <- mean(df1$Q5, na.rm = TRUE)
> summary(df1$Q5)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   15.00   18.00   19.23   21.00  120.00
> summary(df$Q5)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  10.00   15.00   18.00   19.23   21.00  120.00       1
> df1$Q11[is.na(df1$Q11)] <- mean(df1$Q11, na.rm = TRUE)
> summary(df1$Q11)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.00   21.00   23.00   23.01   25.00   52.00
> summary(df$Q11)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  13.00   21.00   23.00   23.01   25.00   52.00       1
```

Question 3

1. Read the dataset "Fuel_Cons_2022.csv" in R, obtain the structure of the dataset and discuss it. [2 marks]

```
> str(data)
'data.frame':   967 obs. of  11 variables:
 $ Make        : chr  "Acura" "Acura" "Acura" "Acura" ...
 $ Model       : chr  "ILX" "MDX SH-AWD" "MDX SH-AWD A-SPEC" "MDX Hybrid AWD" ...
 $ Cylinders   : int  4 6 6 6 4 4 6 4 6 6 ...
 $ Transmission: chr  "AM8" "AS9" "AS9" "AM7" ...
 $ Fuel_type   : chr  "Z" "Z" "Z" "Z" ...
 $ City_Fuel   : num  9.9 12.3 12.2 9.1 11 11.3 8.4 10.2 11.4 12 ...
 $ Hwy_Fuel    : num  7 9.2 9.5 9 8.6 9.1 8.2 7.4 7.7 8.2 ...
 $ Comb_Fuel   : num  8.6 10.9 11 9 9.9 10.3 8.4 8.9 9.8 10.3 ...
 $ Emission_co2: int  199 254 258 210 232 241 196 209 228 240 ...
 $ Rating_Co2  : int  6 4 4 5 5 5 6 5 5 5 ...
 $ Smog_Rating : int  3 3 3 3 6 6 7 3 3 3 ...
```

- The fuel consumption for each vehicle varies by their Model, Fuel_type and where the fuel was bought (i.e., City_Fuel, Hwy_Fuel or Comb_Fuel) and size of "Cylinders".
- There are 7 numeric variables. They are Cylinders, City_Fuel, Hwy_Fuel, Comb_Fuel, Emission_co2, Rating_Co2 and Smog_Rating.
- There are 4 nominal variables. They are Make, Model, Transmission and Fuel_type.

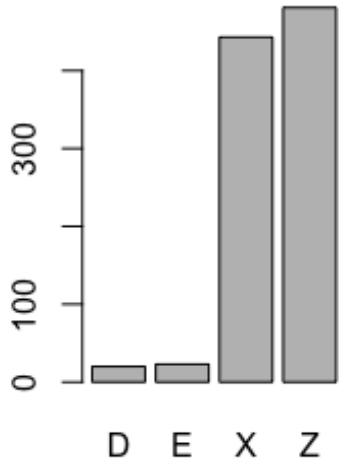
2. Produce a frequency table for the variable "Fuel_type" and discuss it. [2 marks]

```
> freq_table <- table(data$Fuel_type)
> freq_table

  D   E   X   Z
 20  23 443 481
```

- The frequency table has been ordered in ascending order, from left to right.
- Fuel D is the least used fuel type and fuel Z is the most used fuel type.
- Fuel D could be the most expensive, scarce, and least environmentally friendly fuel type, thus the least used.
- Fuel Z could be the least expensive, most abundant, and most environmentally friendly fuel type, thus most used.

3. Obtain a bar plot for the variable “Fuel_type”. [2 marks]

barplot(freq_table)



4. Obtain mean and standard deviation for the variable “City_Fuel” based on “Cylinders” and discuss. [4 marks]

```
> psych::describeBy(data$City_Fuel, data$Cylinders)

 Descriptive statistics by group
group: 3
   vars  n mean   sd median trimmed  mad min max range
X1    1 12 8.52 0.93    8.6    8.58 0.52 6.6 9.8   3.2
    skew kurtosis   se
X1 -0.61     -0.5 0.27
------------------------------------------
group: 4
   vars   n mean   sd median trimmed  mad min  max range
X1    1 425 9.89 1.78   10.1   10.03 1.63 4.2 14.3  10.1
    skew kurtosis   se
X1 -0.85     1.01 0.09
------------------------------------------
group: 5
   vars n mean sd median trimmed mad  min  max range skew
X1    1 2 12.1  0   12.1    12.1   0 12.1 12.1     0  NaN
   kurtosis se
X1      NaN  0
------------------------------------------
group: 6
   vars   n  mean  sd median trimmed  mad min  max range
X1    1 295 12.89 1.7   12.8    12.8 1.33 7.5 22.1  14.6
   skew kurtosis  se
X1 1.17      5.6 0.1
------------------------------------------
group: 8
   vars   n  mean   sd median trimmed  mad  min  max range
X1    1 202 16.25 2.14   15.8   15.96 1.63 12.8 24.5  11.7
   skew kurtosis   se
X1 1.29     1.66 0.15
------------------------------------------
```

```
group: 10
   vars n  mean   sd median trimmed mad  min max range
X1    1 6 17.83 0.26     18   17.83   0 17.5  18   0.5
    skew kurtosis   se
X1 -0.54    -1.96 0.11
------------------------------------------
group: 12
   vars  n  mean   sd median trimmed  mad  min  max range
X1    1 23 20.61 3.37     20   20.32 1.19 15.5 28.1  12.6
   skew kurtosis  se
X1 1.09     0.24 0.7
------------------------------------------
group: 16
   vars n mean   sd median trimmed mad  min  max range
X1    1 2   27 0.28     27      27 0.3 26.8 27.2   0.4
   skew kurtosis  se
X1    0    -2.75 0.2
```

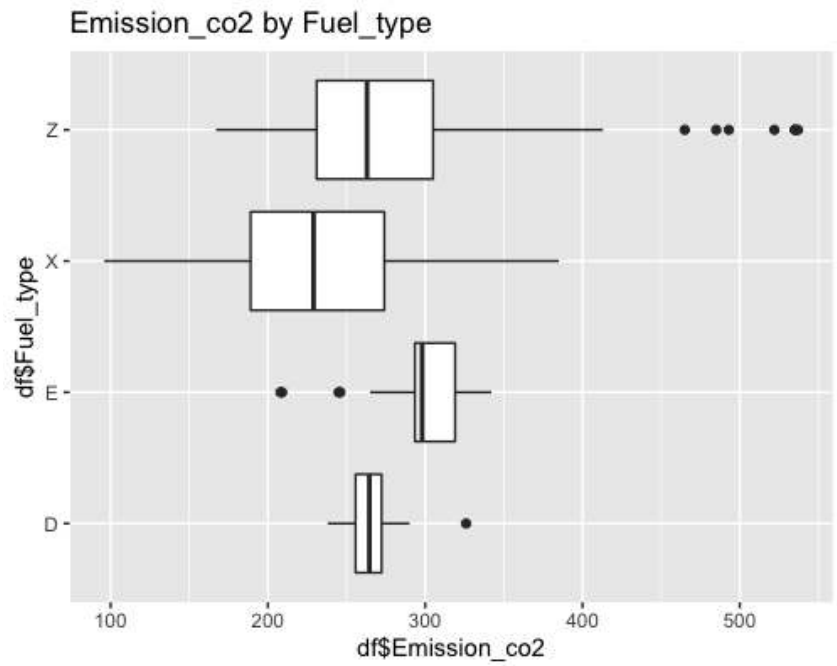
- The larger the value of the “Cylinders” variable, the larger the fuel consumption of the vehicle. This can be seen through the increase in mean values of fuel consumption for the vehicles running on “City_Fuel”.
- The standard deviation is used to quantify the amount of variation or dispersion of a set of data values from the mean. It can be seen that using “Cylinders” 5, there is a high consistency in the fuel consumption of “City_Fuel” with data only collected from 2 vehicles, whereas there is a lower consistency in the fuel consumption of “City_Fuel” using “Cylinders” 12 with data collected from a larger dataset of 12 vehicles.

5. List the records of the vehicles where Smog_Rating= 7, Transmission= “A6” and Fuel_type= “X”. [2 marks]

```
> list <- subset(df, df$Smog_Rating == 7 & df$Transmission == "A6" & df$Fuel_type == "X")
> list
        Make       Model Cylinders Transmission Fuel_type
208 Chevrolet     Equinox         4           A6         X
209 Chevrolet Equinox AWD         4           A6         X
    City_Fuel Hwy_Fuel Comb_Fuel Emission_co2 Rating_Co2
208       8.9      7.7       8.4          196          6
209       9.3      8.0       8.7          204          6
    Smog_Rating compare
208           7    4900
209           7    5100
```

6. Obtain a parallel boxplot for the variable “Emission_co2” by “Fuel_type” variable and discuss. [2 marks]

```
library(ggplot2)
ggplot() +
  geom_boxplot(aes(x = df$Emission_co2, y = df$Fuel_type))
  ggtitle("Emission_co2 by Fuel_type")
```



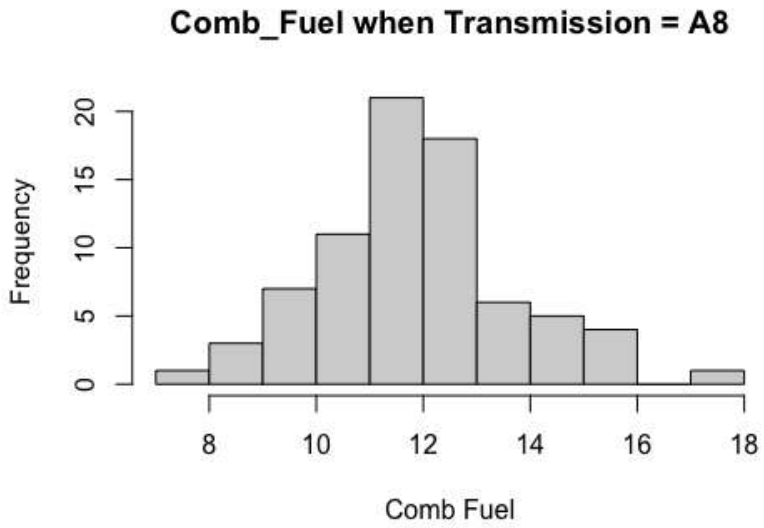

```
> library(psych)
> describeBy(df$Emission_co2, group = df$Fuel_type)

 Descriptive statistics by group
group: D
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis  se
X1    1 20 266.8 19.66  264.5  264.94 13.34 238 326    88 1.16     1.72 4.4
------------------------------------------------------------------------------
group: E
   vars  n   mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 23 293.87 36.13    298  298.53 31.13 208 342   134 -1.11     0.29 7.53
------------------------------------------------------------------------------
group: X
   vars   n   mean   sd median trimmed   mad min max range skew kurtosis   se
X1    1 443 232.68 57.3    229  231.56 62.27  96 385   289 0.15    -0.49 2.72
------------------------------------------------------------------------------
group: Z
   vars   n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 481 272.8 59.47    263  267.08 53.37 167 537   370  1.2     2.47 2.71
```

- The boxplot of Fuel X has no outliers and its approximately symmetrical, with a slight skew of 0.15 to the right. It ranks second in variability of emissions with a a range of 289.
- Comparing the median values of all 4 boxplots, Fuel X can be seen as the most environmentally friendly fuel type with 50% of its emissions below 229, and a minimum value of 96.

- The boxplot of Fuel Z has the greatest variability in emissions, with the largest range of 370, and has the greatest number of outliers, as seen on the boxplot. It is also the least environmentally friendly fuel type, with a maximum value of 537.
- Fuels D, X and Z are positively skewed while fuel E is negatively skewed.
- The boxplot of Fuel E although negatively skewed by 1.11 and has 2 outliers below the lower fence, it has a median of 298, highest amongst these 4 fuel types.
- The boxplot of Fuel D is slightly skewed to the right by 1.16, with an outlier above the upper fence. It is ranked third in terms of environmental friendliness with a median of 264.5.
- The median of emissions of all 4 boxplots are seen to lie between 200 and 300.

7. Obtain a histogram for variable “Comb_Fuel” when Transmission = “A8” and discuss. [2 marks]

```
subset <- subset(df$Comb_Fuel, df$Transmission=="A8")
subset
hist(subset,
     main = "Comb_Fuel when Transmission = A8",
     xlab = "Comb Fuel",
     ylab = "Frequency")
```




```
> summary(subset)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.90   11.00   11.80   11.95   13.00   17.70
```

```
> library(psych)
> describeBy(df$Comb_Fuel, df$Transmission=="A8")

 Descriptive statistics by group
group: FALSE
   vars   n  mean   sd median trimmed  mad min max range skew kurtosis  se
X1    1 890 10.89 2.87   10.5   10.68 2.52 4.1  23  18.9 0.87     1.56 0.1
------------------------------------------------------------------------
group: TRUE
   vars  n  mean  sd median trimmed  mad min  max range skew kurtosis   se
X1    1 77 11.95 1.8   11.8    11.9 1.48 7.9 17.7   9.8 0.38     0.52 0.21
```

- The data of Comb_Fuel is positively skewed, with the mode ranging between 11 and 12.
- The minimum and maximum of Comb_Fuel is 7.90 and 17.70 respectively.
- 25% of values are below the 1st Quartile of 11.
- 75% of values are below the 3rd Quartile of 13.
- The median and mean of 11.80 and 11.95 respectively are similar, which means that the histogram is approximately symmetrical with a slight skew by 0.38 to the right.
- IQR = 2, Upper fence = 13+1.5*2 = 16, Lower fence = 11-1.5*2 = 8. Having calculated the upper and lower fences, values below 8 and above 16 would be outliers.

8. Create a new variable Compare= Emission_co2/Cylinders*100, attach it to the dataset "Fuel_Cons_2022.csv". Using the "Compare" variable, discuss which "Model", and "Make" is more efficient in reducing Co2 emission. [4 marks]

df$compare <-(df$Emission_co2/df$Cylinders*100)
View(df)

Make	Model	Cylinders	Transmission	Fuel_type	City_Fuel	Hwy_Fuel	Comb_Fuel	Emission_co2	Rating_Co2	Smog_Rating	compare
Hyundai	IONIQ Blue	4	AM6	X	4.2	4.0	4.1	96	10	7	2400.00
Hyundai	IONIQ	4	AM6	X	4.2	4.2	4.2	99	10	7	2475.00
Aston Martin	DB11 AMR	12	A8	Z	15.5	10.6	13.3	312	3	3	2600.00
Toyota	Corolla Hybrid	4	AV	X	4.4	4.5	4.5	106	10	7	2650.00
Toyota	Prius	4	AV	X	4.4	4.7	4.5	106	10	7	2650.00
Aston Martin	DBS Superleggera	12	A8	Z	16.4	10.7	13.8	324	3	3	2700.00
Toyota	Prius AWD	4	AV	X	4.5	4.9	4.7	109	10	7	2725.00
Kia	Niro FE	4	AM6	X	4.5	4.8	4.7	110	10	7	2750.00
Toyota	Camry Hybrid LE	4	AV	X	4.9	4.8	4.9	113	10	7	2825.00
Kia	Niro	4	AM6	X	4.6	5.1	4.8	114	10	7	2850.00

	Make	Model	Cylinders	Transmission	Fuel_type	City_Fuel	Hwy_Fuel	Comb_Fuel	Emission_co2	Rating_Co2	Smog_Rating	compare
	All	All	All	am6	x	All	All	All	All	All	All	All
560	Kia	Optima Hybrid	4	AM6	X	5.9	5.2	5.6	132	9	7	3300
558	Kia	Niro Touring	4	AM6	X	5.1	5.8	5.4	129	9	7	3225
476	Hyundai	Sonata Hybrid	4	AM6	X	5.3	4.6	5.0	117	10	7	2925
556	Kia	Niro	4	AM6	X	4.6	5.1	4.8	114	10	7	2850
557	Kia	Niro FE	4	AM6	X	4.5	4.8	4.7	110	10	7	2750
464	Hyundai	IONIQ	4	AM6	X	4.2	4.2	4.2	99	10	7	2475
465	Hyundai	IONIQ Blue	4	AM6	X	4.2	4.0	4.1	96	10	7	2400

- Hyundai's IONIQ Blue is 1st in efficiency in reducing Co2 emissions with the lowest compare rate of 2400. Using fuel type X, it emits the least carbon dioxide where Emission_Co2 = 96, Rating_Co2 = 10 and Smog_Rating = 7.

- The 2nd best vehicle is Hyundai's IONIQ. Using fuel type X, it emits the only slightly more carbon dioxide than the IONIQ Blue model, with a compare value of 2475, Emission_Co2 = 99, Rating_Co2 = 10 and Smog_Rating = 7.

- However, the 3rd best vehicle is Aston Martin's DB11 AMR, despite its higher emissions, use of a larger cylinder and a low rating of 3 for Rating_Co2 and Smog_Rating. This is because the "compare" percentage is smaller when a large emission is divided by a large cylinder value. Thus, the "compare" percentage can only be used as a guiding value.

- Filtering the dataset, Hyundai and Kia "Make" have similar characteristics, where Transmission = "AM6", Fuel_type = "X", and "Cylinders" = 4, relatively emit less carbon dioxide as compared to other vehicles and have lower "compare" percentages as well.