STATISTICAL MODELLING

Assignment by Muhammad Shayan Anwar [27027]

Q-5: Consider the data file 'auto.csv' which contains data of mpg (miles per gallon) and related variables for 392 different cars (after deleting missing rows. The other variables are cylinders (number of cylinders), displacement(cubic inches), horsepower, weight (kg), acceleration(m/s^2), and year (year of manufacturing).

Access the file and prepare the data frame with missing data omitted using the na. Omit command. In this exercise, first, use your sample.

Solution:

access file
using read.csv command.

```
auto = read.csv(file.choose())
attach(auto)
head(auto)
```

#removing rows with missing values

```
clean_auto <- na.omit(auto)
```

```
> head(auto)
  mpg cylinders displacement horsepower weight acceleration year
1  18         8          307        130   3504         12.0   70
2  15         8          350        165   3693         11.5   70
3  18         8          318        150   3436         11.0   70
4  16         8          304        150   3433         12.0   70
5  17         8          302        140   3449         10.5   70
6  15         8          429        198   4341         10.0   70
>
> #removing rows with missing values
> clean_auto <- na.omit(auto)
```

part(a):

Descriptive statistics

```
mpg_stat <- summary(clean_auto$mpg)
mpg_mean <- mean(clean_auto$mpg)
mpg_sd <- sd(clean_auto$mpg)
mpg_quartiles <- quantile(clean_auto$mpg)
```

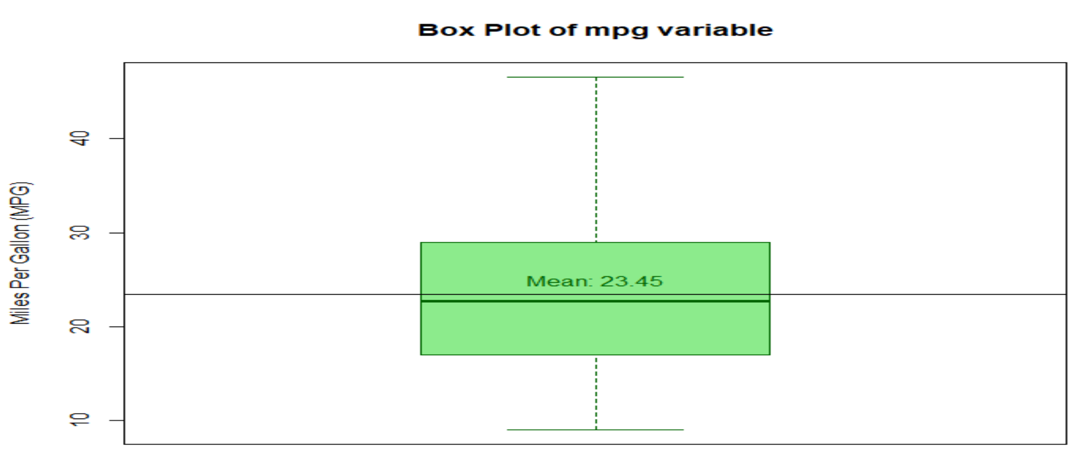
```
mpg_stat
mpg_mean
mpg_sd
Mpg_quartiles
```

```
> mpg_stat <- summary(clean_auto$mpg)
> mpg_mean <- mean(clean_auto$mpg)
> mpg_sd <- sd(clean_auto$mpg)
> mpg_quartiles <- quantile(clean_auto$mpg)
> mpg_stat
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   9.00   17.00   22.75   23.45   29.00   46.60 
> mpg_mean
[1] 23.44592
> mpg_sd
[1] 7.805007
> mpg_quartiles
   0%   25%   50%   75%  100% 
 9.00 17.00 22.75 29.00 46.60 
> 
```

```
# Box plot
boxplot(clean_auto$mpg,
    main = "Box Plot of mpg variable",
    ylab = "Miles Per Gallon (MPG)",
    col = "lightgreen",
    border = "darkgreen")

# Add a mean line to the box plot
abline(h = mpg_mean, col = "black")

# Add label to mean line
text(x = 1,  y = mpg_mean,  labels = paste("Mean:", round(mpg_mean, 2)), pos = 3, col = "darkgreen")
```



Comment:

The dataset is positively skewed as the mean is greater than the median and the longer part of the box to the right side of the median.

The range of data is about (46.6-9)=37.6 miles per gallon, which depicts the overall spread of data.

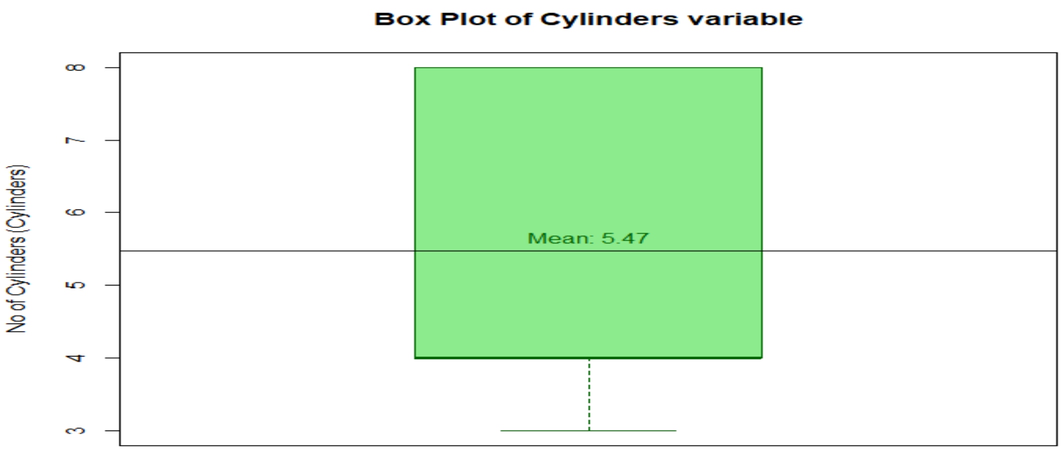
The interquartile range is 12, which relates to the length of that green box in the data. In this case, the IQ is not too large, which certainly depicts less variation in the mid-50 % data.

Furthermore, the data has a standard deviation of 7.80, which also suggests that there is high variability of data from the mean.

Additionally, there are no outliers in MPG data.

part(b)

```
> # Descriptive statistics of cylinders
> cylinders_stat <- summary(clean_auto$cylinders)
> cylinders_sd <- sd(clean_auto$cylinders)
> cylinders_stat
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   4.000   5.472   8.000   8.000
> cylinders_sd
[1] 1.705783
# Box plot
> boxplot(clean_auto$cylinders,
+        main = "Box Plot of Cylinders variable",
+        ylab = "No of Cylinders (Cylinders)",
+        col = "lightgreen",
+        border = "darkgreen")
> # Add mean line to the box plot
> abline(h = mean(clean_auto$cylinders), col = "black")
> # Add label to mean line
> text(x = 1, y = mean(clean_auto$cylinders), labels = paste("Mean:",
round(mean(clean_auto$cylinders), 2)), pos = 3, col = "darkgreen")
```



Comment:

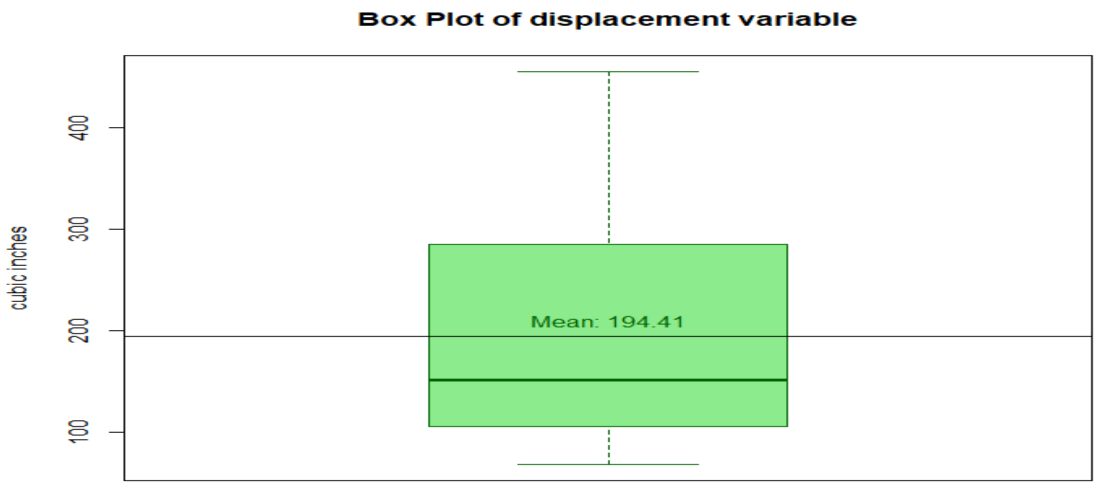
This plot seems bizarre because the max point in the data is the upper quartile, which is 8, and the lower quartile is also the median, which is 4.

The standard deviation is about 1.7 cylinders approximately, and this suggests appropriate variability from the mean.

The graph has no outliers but is rightly skewed(large difference between mean and median) and is uneven in nature to normal whiskey plots.

The graph has high variability in central data, and it, according to my opinion, reflects 75% by that rectangular block.

```
> # Descriptive statistics of displacement
> displacement_stat <- summary(clean_auto$displacement)
> displacement_sd <- sd(clean_auto$displacement)
> displacement_stat
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   68.0   105.0   151.0   194.4   275.8   455.0 
> displacement_sd
[1] 104.644
# Box plot
> boxplot(clean_auto$displacement,
+        main = "Box Plot of displacement variable",
+        ylab = "cubic inches",
+        col = "lightgreen",
+        border = "darkgreen")
> # Add mean line to the box plot
> abline(h = mean(clean_auto$displacement), col = "black")
>
> # Add label to mean line
> text(x = 1, y = mean(clean_auto$displacement), labels = paste("Mean:", round(mean(clean_auto$displacement), 2)), pos = 3, col = "darkgreen")
```



Comment:

The graph has a range of 387 cubic inches, and we can see that it doesn't have a very large variation within 50% of central data(just about 170.8 cubic inches).

The graph is rightly skewed as, again mean is greater than the median.

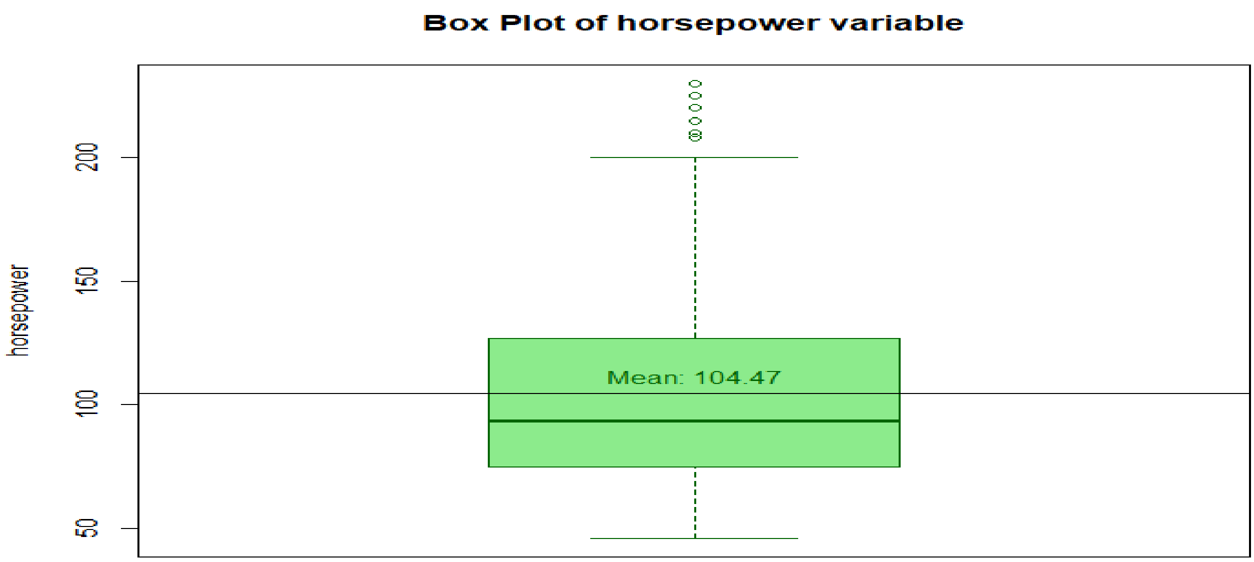
The standard deviation of 104.644 cubic inches shows a very high deviation in displacement from the mean.

Also, there's a large difference between the max value and the 75th percentile,

Additionally, there are no outliers in the data.

```
> # Descriptive statistics of horsepower
>
> horsepower_stat <- summary(clean_auto$horsepower)
> horsepower_sd <- sd(clean_auto$horsepower)
> horsepower_stat
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   46.0    75.0    93.5   104.5   126.0   230.0
> horsepower_sd
[1] 38.49116
> # Box plot
> boxplot(clean_auto$horsepower,
+        main = "Box Plot of horsepower variable",
+        ylab = "horsepower",
+        col = "lightgreen",
+        border = "darkgreen")
>
> # Add mean line to the box plot
> abline(h = mean(clean_auto$horsepower), col = "black")
> # Add label to mean line
> text(x = 1, y = mean(clean_auto$horsepower), labels = paste("Mean:",
round(mean(clean_auto$horsepower), 2)), pos = 3, col = "darkgreen")
```

Comments:



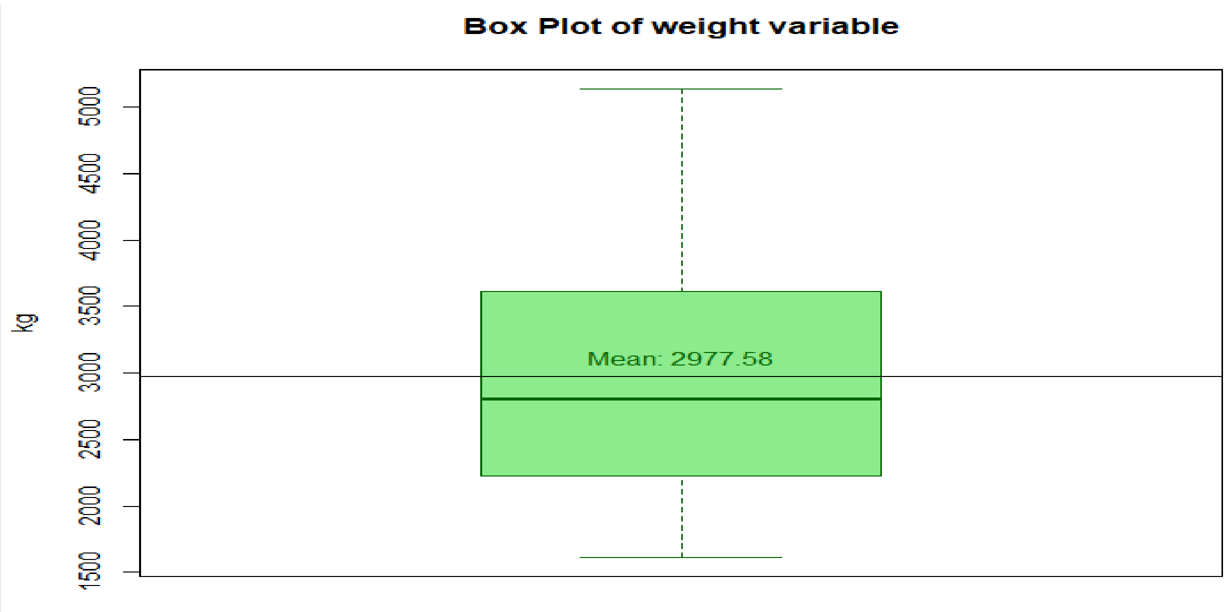
The horsepower data varies from 46 to 230. The interquartile of 51 horsepower and the box plot of considerably smaller length show low variability within its central data.

The standard deviation of 38.49 horsepower explains the variability of horsepower from the mean, which is 104.
47.

The diagram plot(off-centered median position) and the stats(mean>median) also show positive skewness.

And the plot also shows that outliers exist in data.

```
> # Descriptive statistics of weight
> weight_stat <- summary(clean_auto$weight)
> weight_sd <- sd(clean_auto$weight)
> weight_stat
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1613    2225    2804    2978    3615    5140
> weight_sd
[1] 849.4026
> # Box plot
> boxplot(clean_auto$weight,
+        main = "Box Plot of weight variable",
+        ylab = "kg",
+        col = "lightgreen",
+        border = "darkgreen")
>
> # Add mean line to the box plot
> abline(h = mean(clean_auto$weight), col = "black")
> # Add label to mean line
> text(x = 1, y = mean(clean_auto$weight), labels = paste("Mean:",
round(mean(clean_auto$weight), 2)), pos = 3, col = "darkgreen")
```



Comment:

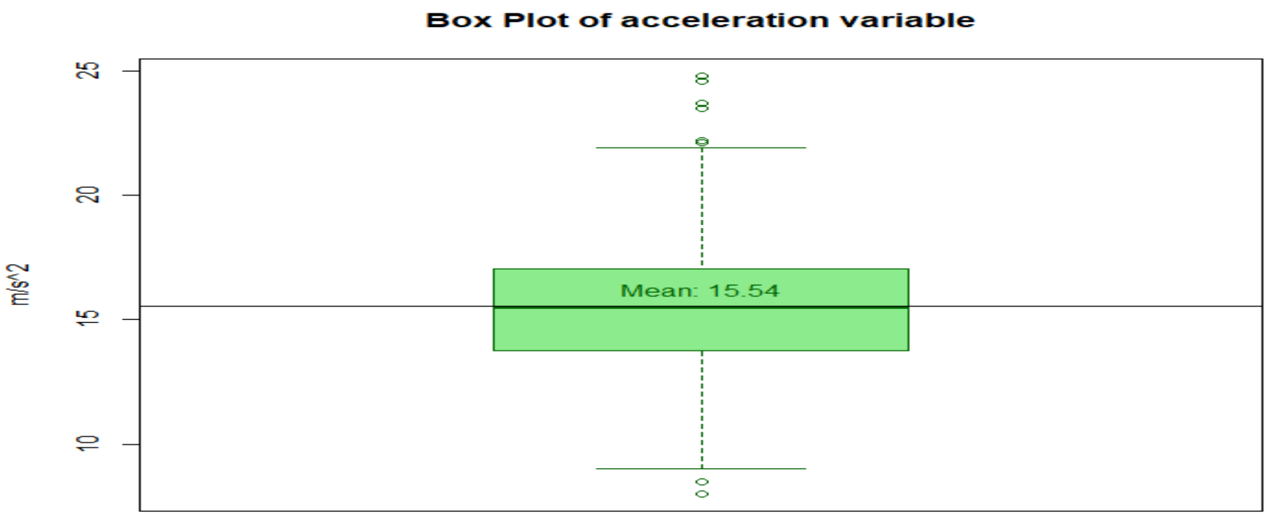
The weight data varies from 1613 to 5140 kgs. The central tendency is explained by the mean and median, which shows the mean (2977.58) to be greater than median (2804), highlighting right/positive skewness in data.

There seems to be good variation within 50% of central data, which lies between the lower and upper quartiles, i.e., 2225 kg and 3615 kg, respectively.

The standard deviation of 849 kg depicts a significant variation in weights from it's mean.

```
> # Descriptive statistics of acceleration

> acceleration_stat <- summary(clean_auto$acceleration)
> acceleration_sd <- sd(clean_auto$acceleration)
> acceleration_stat
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   8.00   13.78   15.50   15.54   17.02   24.80 
> acceleration_sd
[1] 2.758864
> 
> # Box plot
> boxplot(clean_auto$acceleration,
+        main = "Box Plot of acceleration variable",
+        ylab = "m/s^2",
+        col = "lightgreen",
+        border = "darkgreen")
> 
> # Add mean line to the box plot
> abline(h = mean(clean_auto$acceleration), col = "black")
> 
> # Add label to mean line
> text(x = 1, y = mean(clean_auto$acceleration), labels = paste("Mean:",
round(mean(clean_auto$acceleration), 2)), pos = 3, col = "darkgreen")
```



Comment

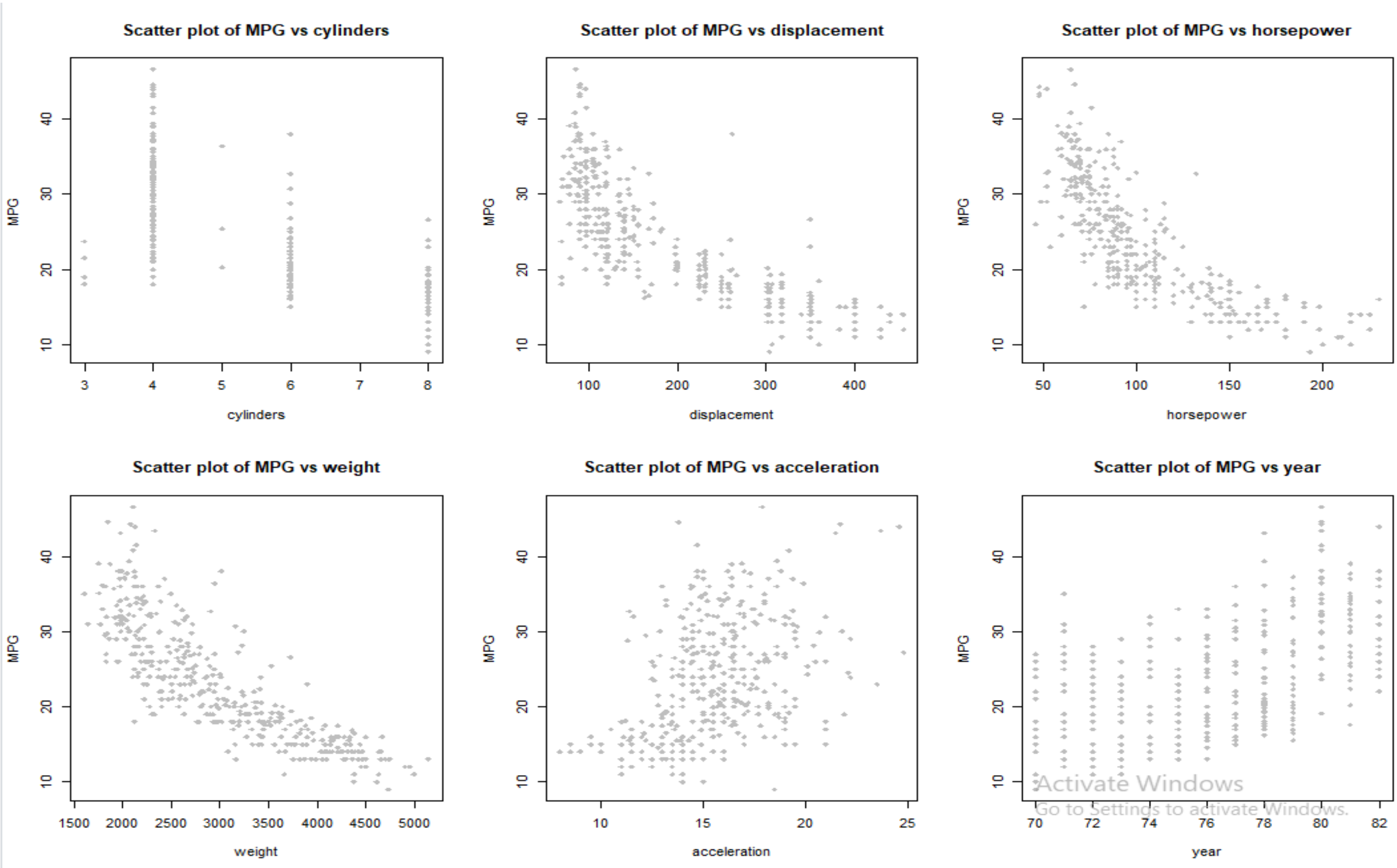
The acceleration data lie between 8 and 24.8 m/s2. It has a lower quartile of 13.78 and an upper quartile of 17.02, which showcases a very low variation of acceleration within the central 50% data. This is also vividly illustrated in the boxplot.

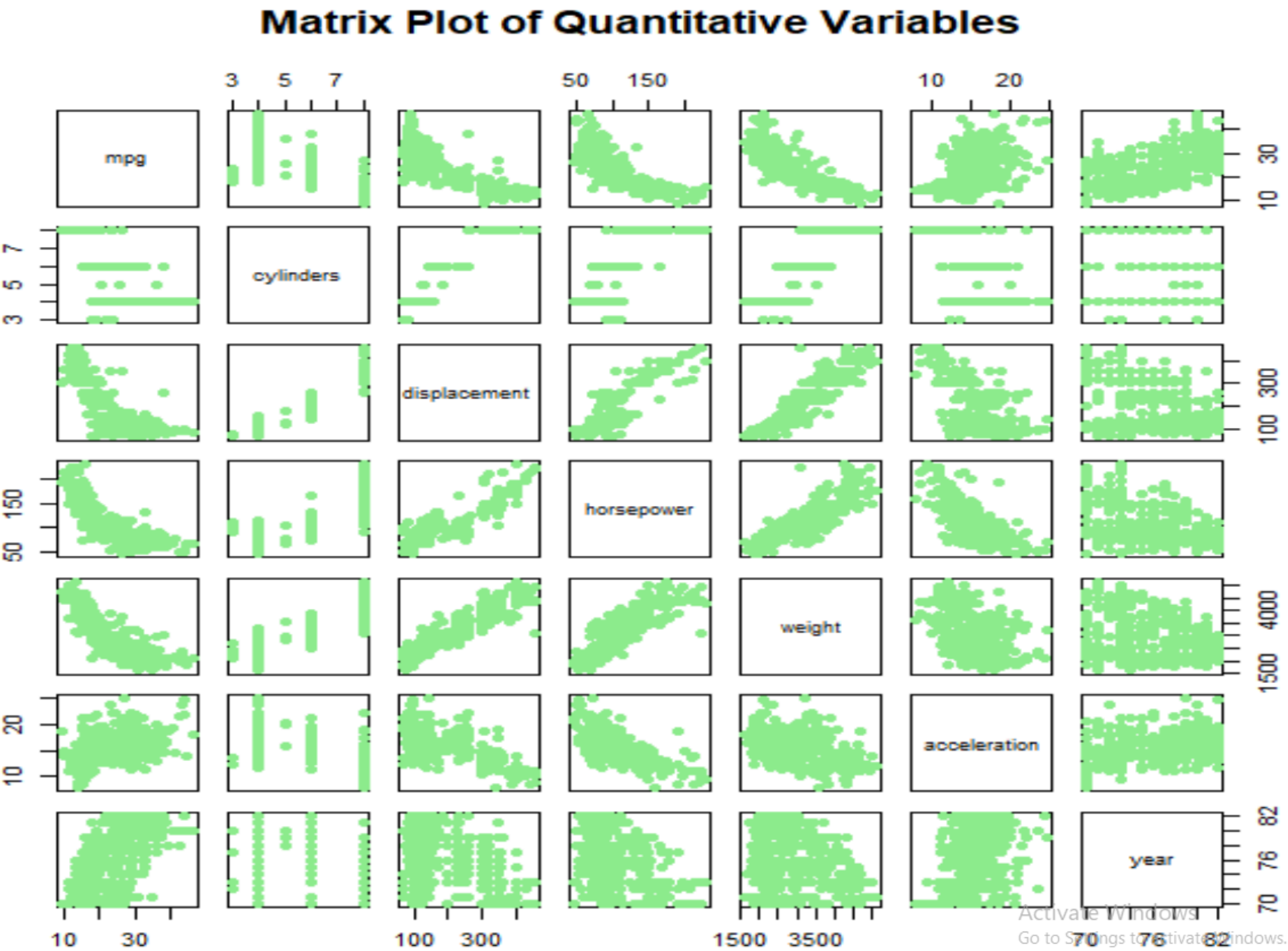
The median and mean are approximately 15.5 and are nearly equal. On the box plot, we can also see that both the lines of mean and median are very close to each other, depicting no or negligible skewness in data.

There also exist a few outliers in this data,

part(c):

```
> # Extracting quantitative variables
> quantitative_vars <-
c("cylinders","displacement","horsepower","weight","acceleration","year")
>
> # Set up layout for multiple plots
> par(mfrow = c(2, 3)) # Adjust dimensions based on the number of variables
>
> # Construct scatter plots
> for (var in quantitative_vars) {
+  plot(clean_auto[[var]], clean_auto$mpg,
+       xlab = var, ylab = "MPG",
+       main = paste("Scatter plot of MPG vs", var),
+       col = "gray", pch = 18)
+ }
> # Matrix plot
> pairs(clean_auto[, quantitative_vars],
+      main = "Matrix Plot of Quantitative Variables",
+      pch = 19,col="lightgreen")
```





Comments:

It is evident through scatterplots and matrix plots that the mpg varies linearly with no of cylinders.

On the other hand, mpg has decreasing and exponential variation with horsepower, displacement, and weight.

And, mpg has increasing exponential variation with year and acceleration.

part(d)

```
> model1=lm(mpg ~
cylinders+displacement+horsepower+weight+acceleration+year, data =
clean_auto)
```

```
> summary(model1)

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year, data = clean_auto)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6927 -2.3864 -0.0801  2.0291 14.3607

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.454e+01  4.764e+00  -3.051  0.00244 **
cylinders    -3.299e-01  3.321e-01  -0.993  0.32122
displacement  7.678e-03  7.358e-03   1.044  0.29733
horsepower   -3.914e-04  1.384e-02  -0.028  0.97745
weight       -6.795e-03  6.700e-04 -10.141  < 2e-16 ***
acceleration  8.527e-02  1.020e-01   0.836  0.40383
year          7.534e-01  5.262e-02  14.318  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.435 on 385 degrees of freedom
Multiple R-squared:  0.8093,  Adjusted R-squared:  0.8063
F-statistic: 272.2 on 6 and 385 DF,  p-value: < 2.2e-16
```

Comments:

Cylinders, horsepower, and weight are negatively related to mpg.
Displacement, acceleration, and year are positively related.

According to expectations(shown by scatterplot), displacement should also negatively impact mpg, but it does not.

Only weight and year are statistically significant.

part(e):

Interpretations:

- **Intercept:** When the number of cylinders, displacement, horsepower, weight, acceleration, and years is taken to be zero, there exist -14.54 miles per gallon. This is impractical.
- **Cylinders:** When the number of cylinders increases by one, the miles per gallon decreases by 0.3299, keeping all other predictors constant.
- **Displacement:** When the displacement increases by cubic inches, the miles per gallon increases by 0.007678, keeping all other predictors constant.
- **Horsepower:** When horsepower increases by 1 unit, the miles per gallon decrease by 0.0003914, keeping all other predictors constant.
- **Weight:** When weight increases by 1kg, the miles per gallon decrease by 0.006795, keeping all other predictors constant.

- **Acceleration:** When acceleration increases by 1m/s^2, the miles per gallon increase by 0.08527, keeping all other predictors constant.
- **Year:** When one year passes, the miles per gallon increase by 0.7534, keeping all other predictors constant.

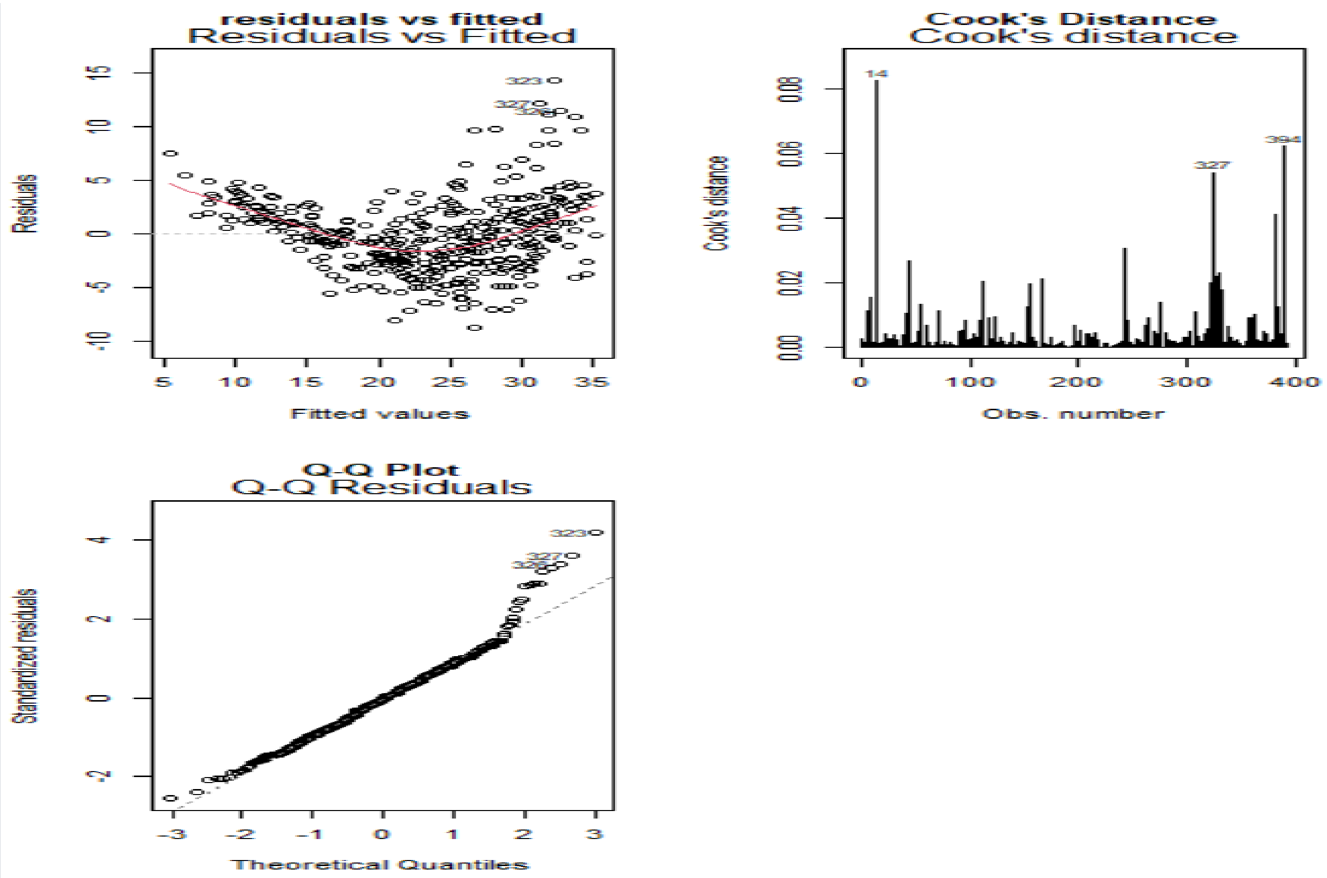
part(f):

Multiple R-squared: 0.8093, Adjusted R-squared: 0.8063
The R2 suggests that the model is 80.93% explainable by the predictors: cylinder, displacement, acceleration, year, horsepower, and weight.

Adjusted R2 is about 80.63% and is 0.2% lower than R2. This depicts a bit of redundancy and irrelevance in predictors when explaining mpg.

part(g):

```
> plot(model1, , which=2, main = "Q-Q Plot")
> plot(model1, , which=4, main = "Cook's Distance")
> plot(model1, which=1 ,main = "residuals vs fitted")
```



Comments:

- Non-linear graph as shown in residuals plot.
- Heteroscedastic due to nonconstant variance
- Normality is detected in the Q-Q plot but is rightly skewed
- Outliers are visible due to long bars on the cook's graph.

part(h):

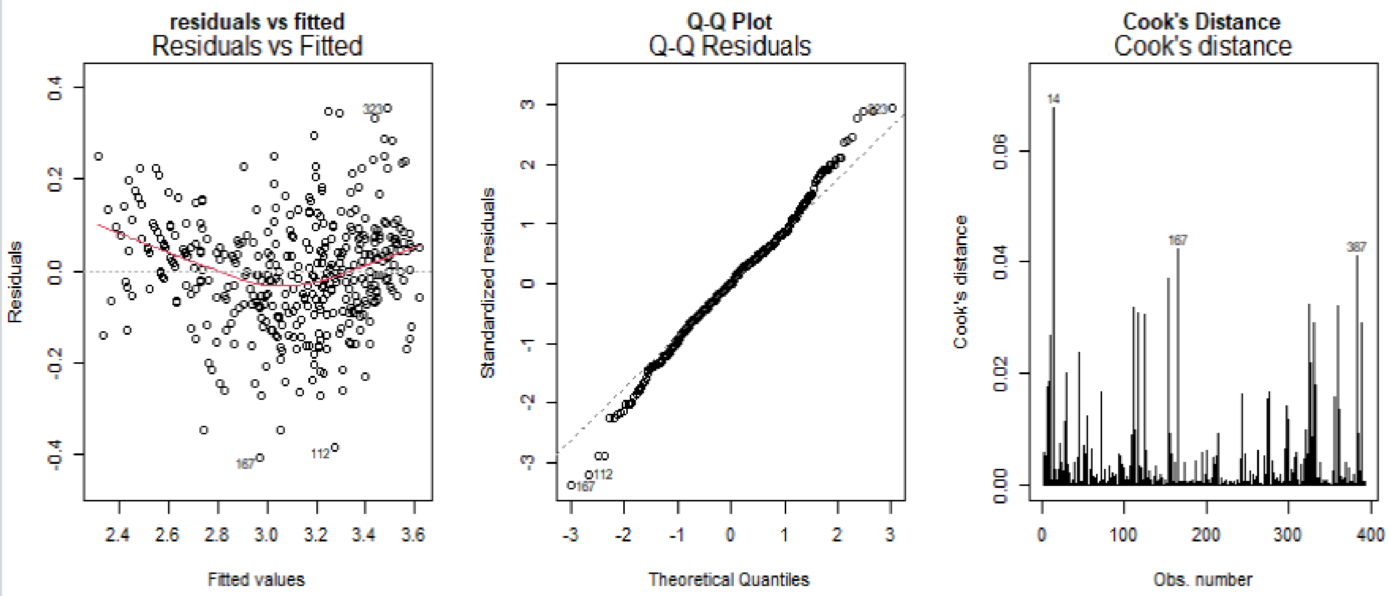
```
> data=data.frame(cylinders=48,displacement=350,horsepower=140 , weight=
3500, acceleration = 11.5, year= 1970)
>
> predicted <- predict(model1,newdata=data )
> predicted
       1
1433.597

> predict(model1, data, interval='confidence')
       fit      lwr      upr
1 1433.597 1235.303 1631.891
```

part(i):

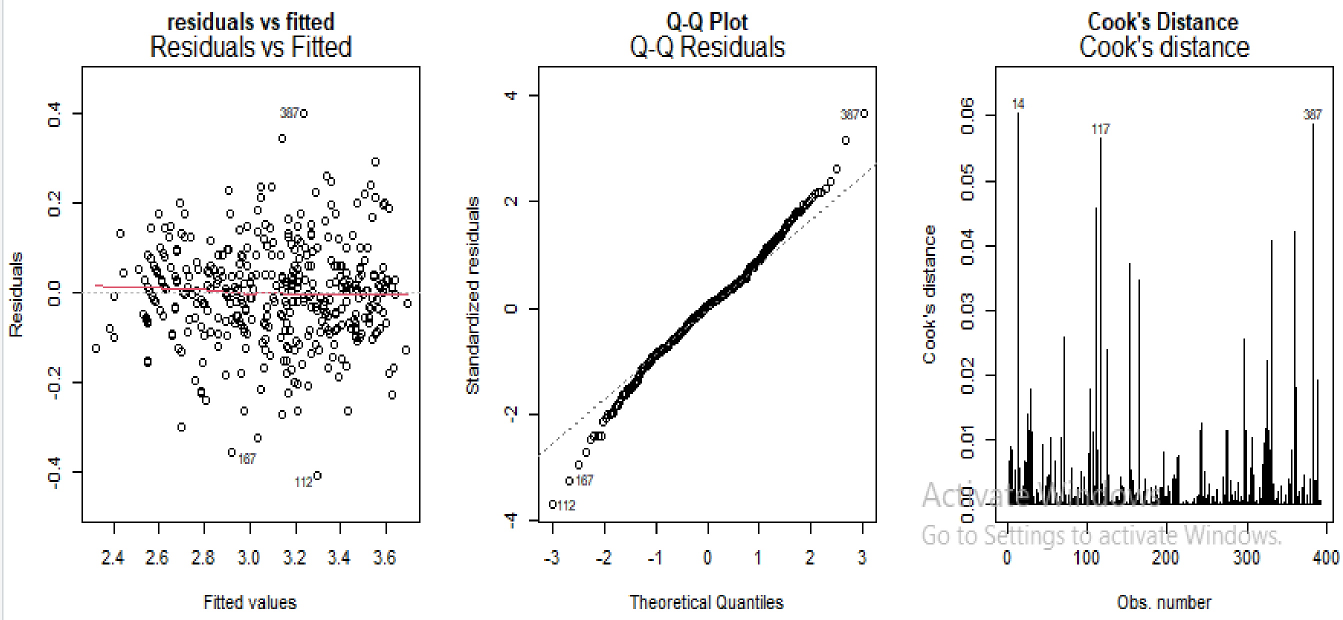
```
> model2= lm(log(mpg) ~
cylinders+displacement+horsepower+weight+acceleration+year, data =
clean_auto)
> model3= lm(log(mpg) ~
cylinders+displacement+I(displacement^2)+horsepower+I(horsepower^2)+weight+
acceleration+year, data = clean_auto)
> model4= lm(log(mpg) ~
cylinders+displacement+I(displacement^2)+horsepower+I(horsepower^2)+weight+
acceleration+year+(year*horsepower) + (weight*cylinders), data =
clean_auto)


> plot(model2, which=1 ,main = "residuals vs fitted")
> plot(model2, , which=2, main = "Q-Q Plot")
> plot(model2, , which=4, main = "Cook's Distance")
```



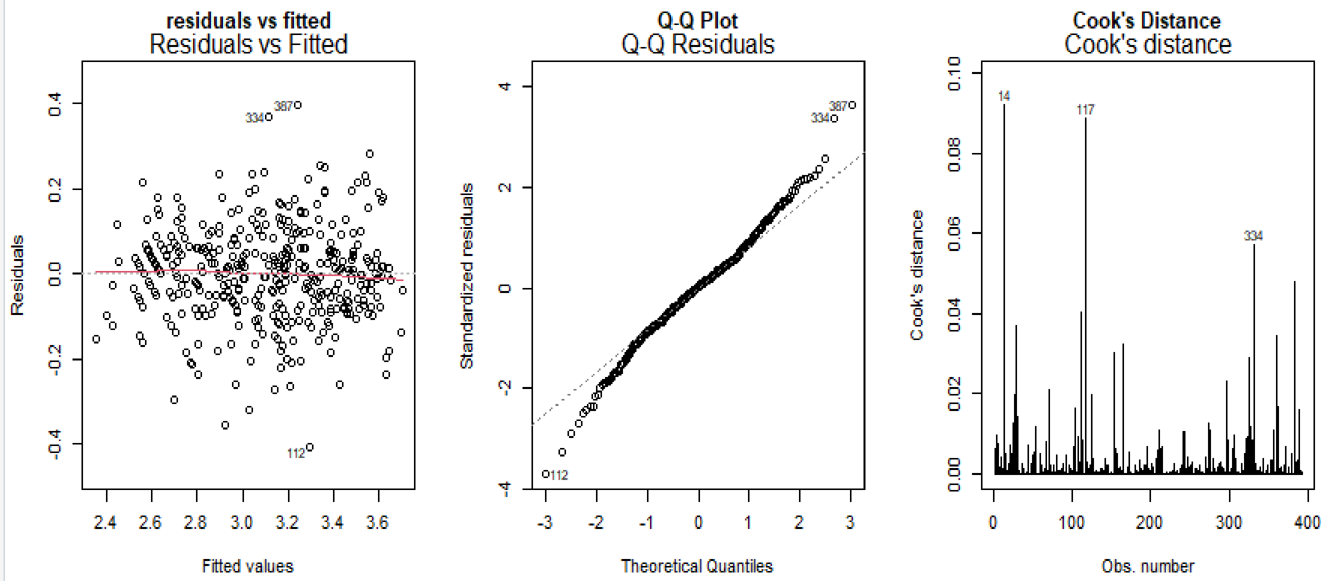
-> Nonlinear, normal, heteroscedastic outliers are present.

```
> plot(model3, which=1 ,main = "residuals vs fitted")
> plot(model3, , which=2, main = "Q-Q Plot")
> plot(model3, , which=4, main = "Cook's Distance")
```



> nonlinear, normal, heteroscedastic outliers present..

```
> plot(model4, which=1 ,main = "residuals vs fitted")
> plot(model4, , which=2, main = "Q-Q Plot")
> plot(model4, , which=4, main = "Cook's Distance")
```



-> nonlinear, normal, heteroscedastic outliers present

part(j):

Best Subset Regression:

```
> best_subset <- regsubsets(log(mpg) ~ cylinders + displacement +
I(displacement^2) + horsepower + I(horsepower^2) + weight + acceleration +
year + (year * horsepower) + (weight * cylinders) , data = clean_auto)
> # Print out the summary
> summary(best_subset)
```

```
                  Forced in Forced out
cylinders             FALSE      FALSE
displacement          FALSE      FALSE
I(displacement^2)     FALSE      FALSE
horsepower            FALSE      FALSE
I(horsepower^2)       FALSE      FALSE
weight                FALSE      FALSE
acceleration          FALSE      FALSE
year                  FALSE      FALSE
horsepower:year       FALSE      FALSE
cylinders:weight      FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         cylinders displacement I(displacement^2) horsepower I(horsepower^2) weight acceleration year
1  ( 1 ) " "       " "          " "               " "        " "             "*"    " "          " "
2  ( 1 ) " "       " "          " "               " "        " "             "*"    " "          "*"
3  ( 1 ) " "       " "          " "               " "        " "             "*"    " "          "*"
4  ( 1 ) " "       " "          " "               " "        "*"             "*"    " "          "*"
5  ( 1 ) " "       "*"          "*"               " "        " "             "*"    " "          "*"
6  ( 1 ) " "       "*"          "*"               "*"        " "             "*"    " "          "*"
7  ( 1 ) " "       "*"          "*"               " "        "*"             "*"    "*"          "*"
8  ( 1 ) "*"       "*"          "*"               " "        "*"             "*"    "*"          "*"
         horsepower:year cylinders:weight
1  ( 1 ) " "             " "
2  ( 1 ) " "             " "
3  ( 1 ) "*"             " "
4  ( 1 ) "*"             " "
5  ( 1 ) "*"             " "
6  ( 1 ) "*"             " "
7  ( 1 ) "*"             " "
8  ( 1 ) "*"             " "
>
```

```
> model4 <- lm(log(mpg) ~ cylinders + displacement + I(displacement^2) +
horsepower + I(horsepower^2) + weight + acceleration + year + (year *
horsepower) + (weight * cylinders), data = clean_auto)
>
> model11=lm(log(mpg) ~ weight, data = auto)
> model12=lm(log(mpg) ~ weight + year, data = auto)
> model13=lm(log(mpg) ~ weight + year + (year * horsepower) , data =
clean_auto)
> model14=lm(log(mpg) ~ weight + year + (year * horsepower) +
I(horsepower^2), data = clean_auto)
> model15=lm(log(mpg) ~ displacement + I(displacement^2) + weight + year +
(year * horsepower), data = clean_auto)
> model16=lm(log(mpg) ~ displacement + I(displacement^2) + weight + year +
(year * horsepower) +horsepower, data = clean_auto)
```

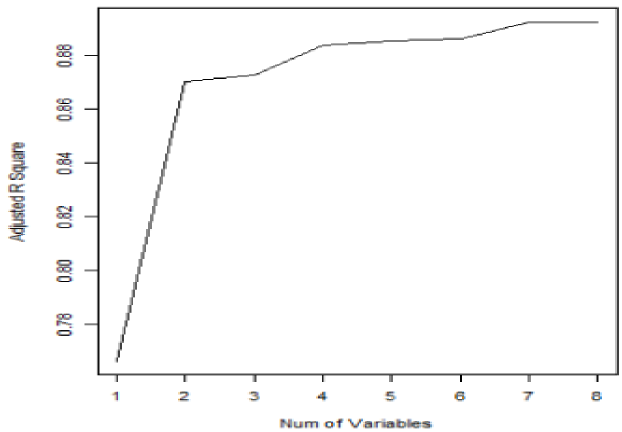
```
> model17=lm(log(mpg) ~ displacement + I(displacement^2) + I(horsepower^2)
+ weight + acceleration + year + (year * horsepower), data = clean_auto)
> model18=lm(log(mpg) ~ cylinders+displacement + I(displacement^2) +
I(horsepower^2) + weight + acceleration + year + (year * horsepower), data
= clean_auto)
>
> PRESS=function(model4){
+  i=residuals(model4)/(1-lm.influence(model4)$hat)
+  sum(i^2)
+
+ }
>
>
PRESS=c(PRESS(model11),PRESS(model12),PRESS(model13),PRESS(model14),PRESS(m
odel15),PRESS(model16),PRESS(model17),PRESS(model18))
> data.frame(Adj.R2=measures$adjr2,CP=measures$cp, BIC=measures$bic ,
PRESS)
     Adj.R2         CP       BIC     PRESS
1 0.7661793 455.382521 -558.7159 10.783688
2 0.8700418  81.552288 -783.9891  5.966912
3 0.8727762  72.538280 -787.3626  5.569715
4 0.8835764  34.705220 -817.1784  5.400759
5 0.8911633   8.543523 -838.6369  5.039772
6 0.8917787   7.345449 -835.9055  5.039772
7 0.8922760   6.578458 -832.7592  5.070274
8 0.8922683   7.609410 -827.7821  5.083555
```

Best Forward Selection:

```
> frwd <- regsubsets(log(mpg) ~ cylinders + displacement +
I(displacement^2) + horsepower + I(horsepower^2) + weight + acceleration +
year + (year * horsepower) + (weight * cylinders) , data = clean_auto,
method ="forward")
>
> # Print out the summary
> summary(frwd)
```

```
Selection Algorithm: forward
         cylinders displacement I(displacement^2) horsepower I(horsepower^2) weight acceleration year horsepower:year cylinders:weight
1  ( 1 ) " "       " "          " "               " "        " "             "*"    " "          " "  " "             " "
2  ( 1 ) " "       " "          " "               " "        " "             "*"    " "          "*"  " "             " "
3  ( 1 ) " "       " "          " "               " "        " "             "*"    " "          "*"  "*"             " "
4  ( 1 ) " "       " "          " "               " "        "*"             "*"    " "          "*"  "*"             " "
5  ( 1 ) " "       " "          " "               " "        "*"             "*"    "*"          "*"  "*"             " "
6  ( 1 ) " "       "*"          " "               " "        "*"             "*"    "*"          "*"  "*"             " "
7  ( 1 ) " "       "*"          "*"               " "        "*"             "*"    "*"          "*"  "*"             " "
8  ( 1 ) "*"       "*"          "*"               " "        "*"             "*"    "*"          "*"  "*"             " "
```

```
> plot(b$adjr2,xlab="Num of Variables", ylab="Adjusted R Square", type="l")
> coef(frwd,6)
    (Intercept)    displacement I(horsepower^2)          weight
acceleration             year horsepower:year
   1.459011e+00   -3.632565e-04    2.178431e-05   -1.817692e-04
-1.111763e-02    3.934857e-02   -1.060705e-04
> which.max(b$adjr2)
[1] 7
```

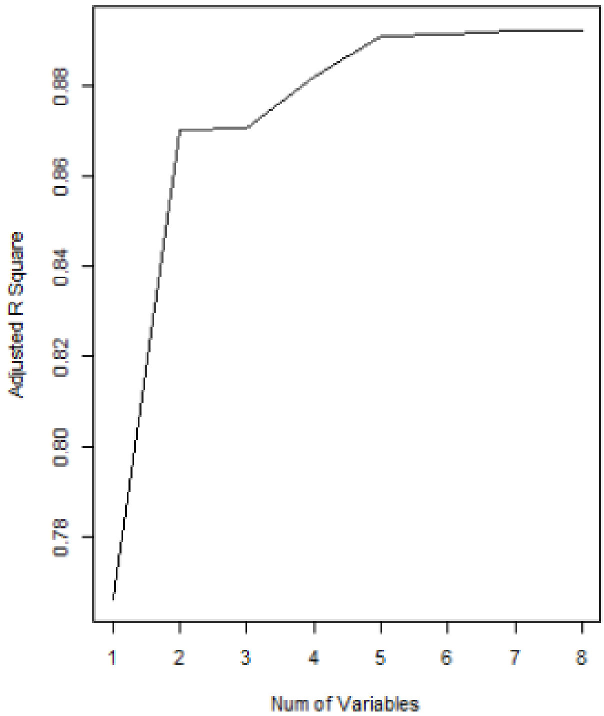


Best Backward Selection:

```
> # Best Backward selection
> bwd <- regsubsets(log(mpg) ~ cylinders + displacement + I(displacement^2)
+ horsepower + I(horsepower^2) + weight + acceleration + year + (year *
horsepower) + (weight * cylinders) , data = clean_auto, method ="backward")
>
> # Print out the summary
> summary(bwd)
```

```
Selection Algorithm: backward
         cylinders displacement I(displacement^2) horsepower I(horsepower^2) weight acceleration year horsepower:year cylinders:weight
1  ( 1 ) " "       " "          " "               " "        " "             "*"    " "          " "  " "             " "
2  ( 1 ) " "       " "          " "               " "        " "             "*"    " "          "*"  " "             " "
3  ( 1 ) " "       "*"          " "               " "        " "             "*"    " "          "*"  " "             " "
4  ( 1 ) " "       "*"          "*"               " "        " "             "*"    " "          "*"  " "             " "
5  ( 1 ) " "       "*"          "*"               " "        " "             "*"    " "          "*"  "*"             " "
6  ( 1 ) " "       "*"          "*"               " "        " "             "*"    "*"          "*"  "*"             " "
7  ( 1 ) " "       "*"          "*"               " "        "*"             "*"    "*"          "*"  "*"             " "
8  ( 1 ) " "       "*"          "*"               " "        "*"             "*"    "*"          "*"  "*"             "*"
>
```

```
> plot(a$adjr2,xlab="Num of Variables", ylab="Adjusted R Square", type="l")
> coef(bwd,6)
      (Intercept)      displacement I(displacement^2)            weight
acceleration              year
     1.763954e+00     -2.841869e-03      5.819127e-06     -1.867638e-04
-5.187498e-03      3.327449e-02
  horsepower:year
    -3.656752e-05
> which.max(a$adjr2)
[1] 7
```

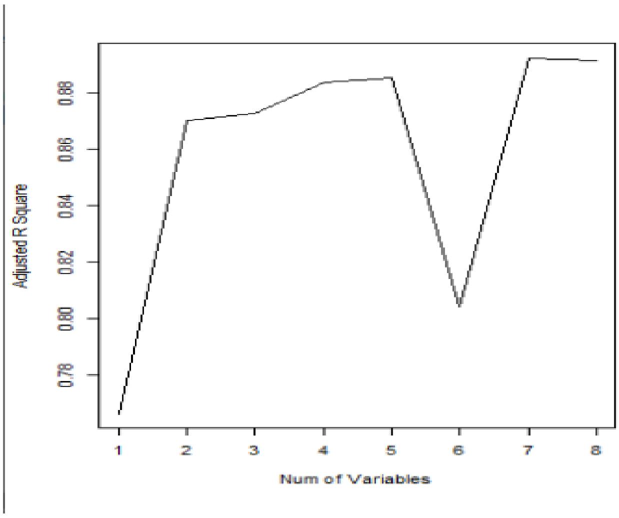


Stepwise Selection:

```
# Perform stepwise selection
> stepwise_selection <- regsubsets(log(mpg) ~ cylinders + displacement +
I(displacement^2) + horsepower + I(horsepower^2) + weight + acceleration +
year + (year * horsepower) + (weight * cylinders) , data = clean_auto,
method ="seqrep")
>
> # Print out the summary
> summary(stepwise_selection)
```

```
         cylinders displacement I(displacement^2) horsepower I(horsepower^2) weight acceleration year horsepower:year
1  ( 1 ) " "       " "          " "               " "        " "            "*"    " "          " "  " "
2  ( 1 ) " "       " "          " "               " "        " "            "*"    " "          "*"  " "
3  ( 1 ) " "       " "          " "               " "        " "            "*"    " "          "*"  "*"
4  ( 1 ) " "       " "          " "               " "        "*"            "*"    " "          "*"  "*"
5  ( 1 ) " "       " "          " "               " "        "*"            "*"    "*"          "*"  "*"
6  ( 1 ) "*"       "*"          "*"               "*"        "*"            "*"    " "          " "  " "
7  ( 1 ) " "       "*"          "*"               " "        "*"            "*"    "*"          "*"  "*"
8  ( 1 ) "*"       "*"          "*"               "*"        "*"            "*"    "*"          "*"  " "
         cylinders:weight
1  ( 1 ) " "
2  ( 1 ) " "
3  ( 1 ) " "
4  ( 1 ) " "
5  ( 1 ) " "
6  ( 1 ) " "
7  ( 1 ) " "
8  ( 1 ) " "
>
```

```
plot(stepwise_selection_sum$adjr2,xlab="Num of Variables", ylab="Adjusted R
Square", type="l")
> coef(stepwise_selection,6)
      (Intercept)         cylinders      displacement I(displacement^2)
horsepower   I(horsepower^2)            weight
     4.285483e+00      5.854548e-03     -2.701500e-03      4.680986e-06
-5.222995e-03      7.124613e-06     -1.560643e-04
> which.max(stepwise_selection_sum$adjr2)
[1]
```



Interpretation:

- Subset Procedure:

  Model 7 is the best model based on adjR2 and cp mallow.

- ## Forward Selection:
  Models 7 and 8 can be regarded as the best models in Forward regression as seen from the graph.

- ## Backward Elimination:
  Models 7 and 8 are the best models, as seen from the graph.

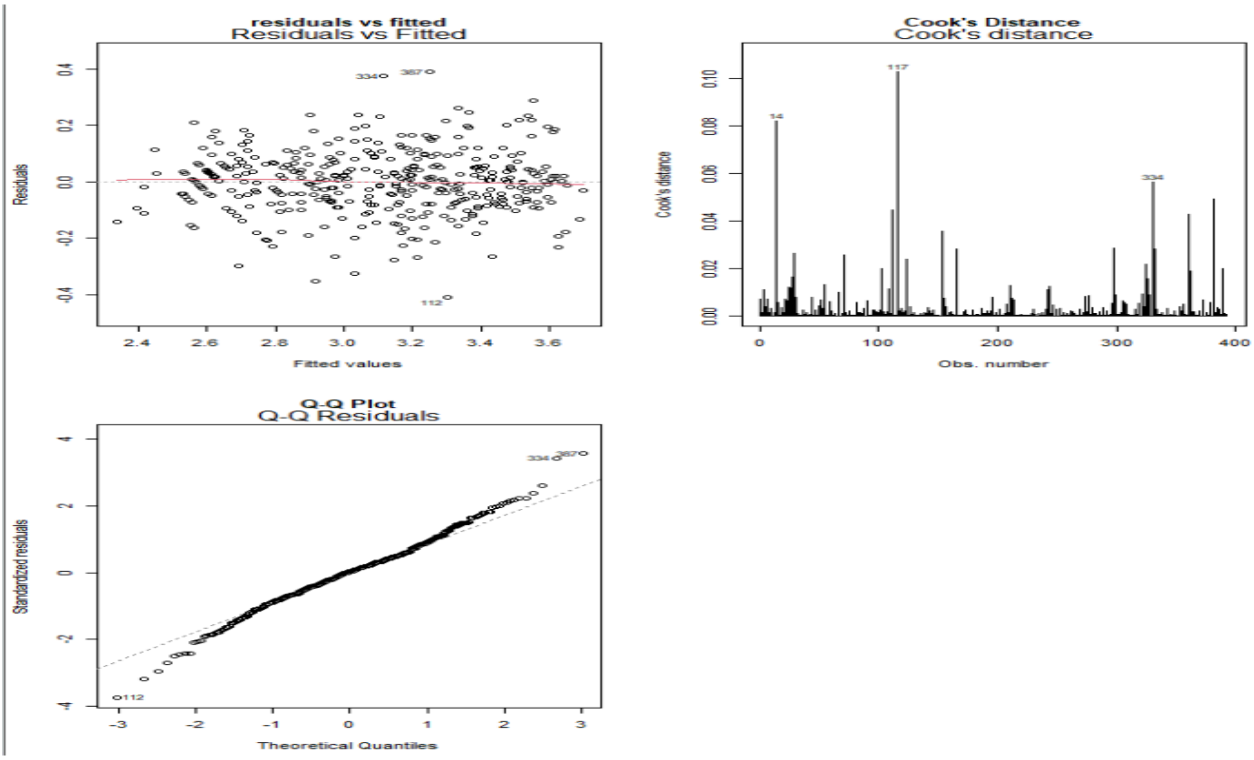
- ## Stepwise Selection:
  Model 7 is the best one and then it's model 8 that follows it.

Best 2 models chosen:

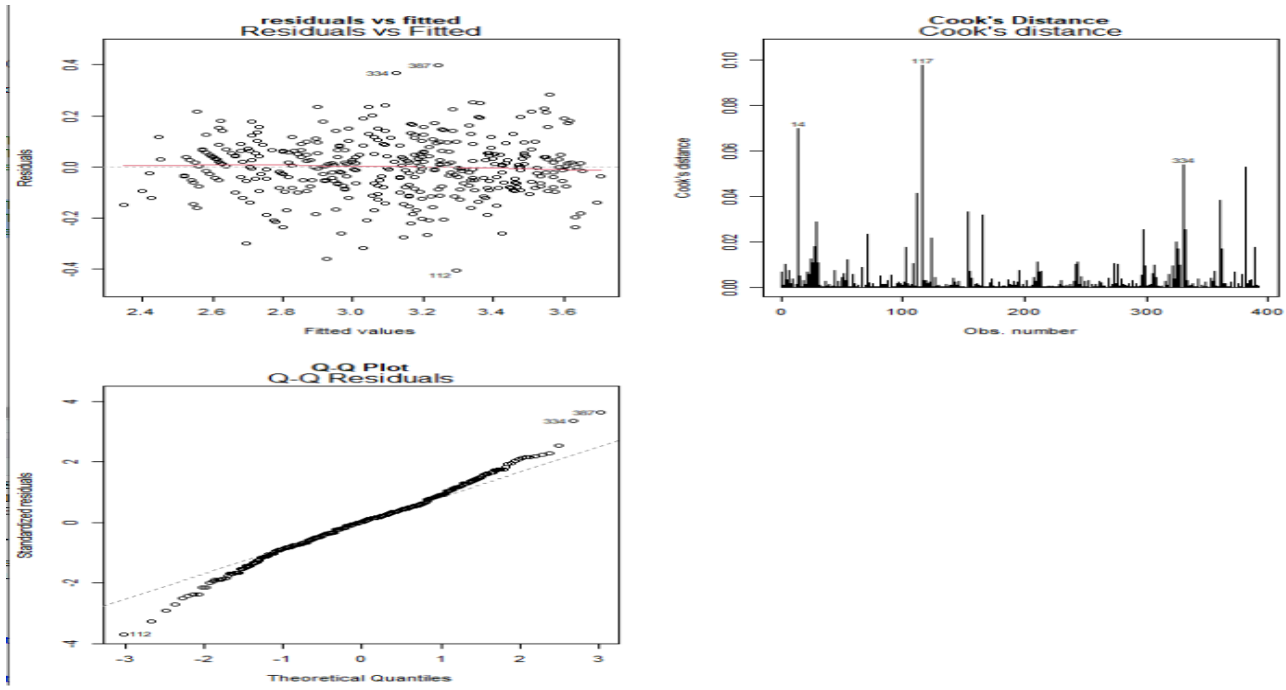
- [ ] log(mpg) ~ displacement + I(displacement^2) + I(horsepower^2) + weight + acceleration + year + (year * horsepower) -> Model 7
- [ ] log(mpg) ~ cylinders+displacement + I(displacement^2) + I(horsepower^2) + weight + acceleration + year + (year * horsepower) -> Model 8

part(k):

```
> plot(model17, which=1 ,main = "residuals vs fitted")
> plot(model17, , which=2, main = "Q-Q Plot")
> plot(model17, , which=4, main = "Cook's Distance")
```




```
> plot(model18, which=1 ,main = "residuals vs fitted")
> plot(model18, , which=2, main = "Q-Q Plot")
> plot(model18, , which=4, main = "Cook's Distance")
```



Model 8’s diagnostic plot shows a bit more non linearity than model 7 and in both graphs there's evidence of heteroscedasticity. Both graphs follow normality except for a slight deviation at the tails making tails as the majority of the points are close to the line in Q-Q plot. Outliers too exist in both the models.

CONCLUSION:
Model 7 seems to be the better fit overall through variable screening process due better adjr2, cp mallow, considerably low PRESS and BIC.

According to diagnostic plots, again model 7 seems a bit more linear and following normality better than model 8.