Assignment 2
Introduction to statistical learning
Abhijith Chowdary Eanuga
16290740

Chapter 2:

8)

a) Code :

```
getwd()
#Place the College.csv file in the working directory
college<-read.csv("College.csv")
```

b) Code:

```
rownames (college )=college [,1]
fix(college)
college =college [,-1] > fix(college)
fix(college)
glimpse(college)
```

C)

i) Code :

```
summary(college)
```

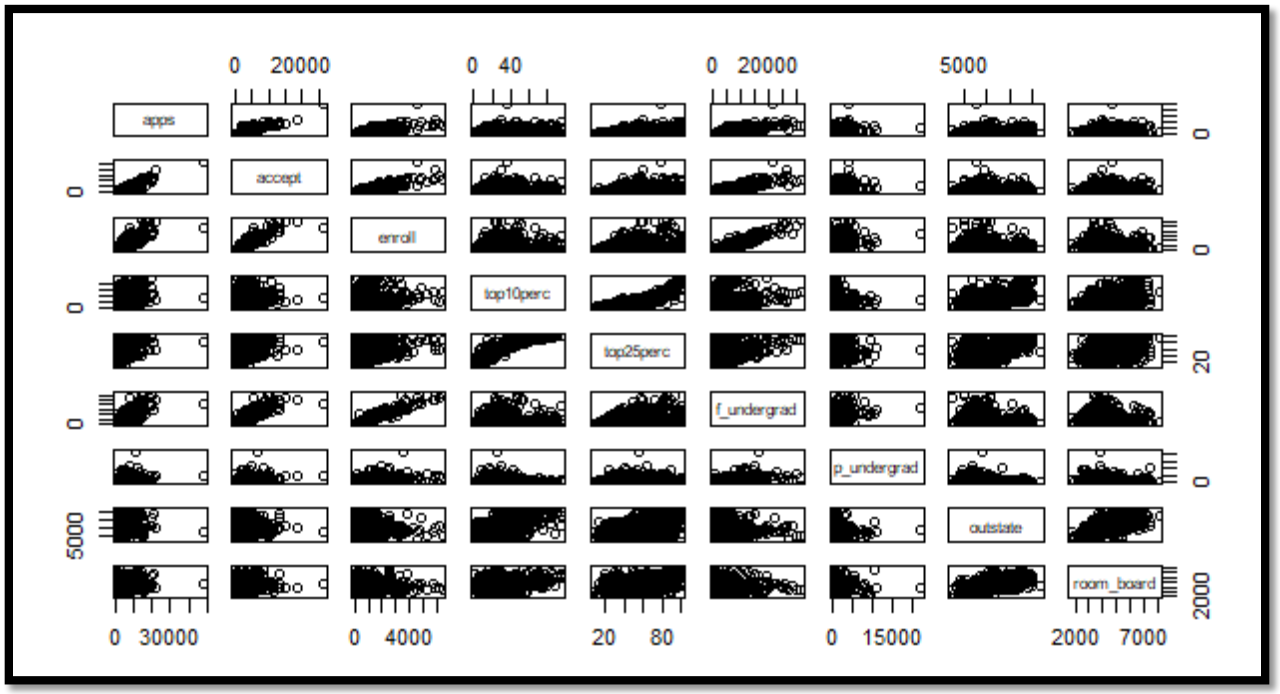
Output:

```
> summary(college)
   private               apps           accept          enroll       top10perc
 Length:777         Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
 Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
 Mode  :character   Median : 1558   Median : 1110   Median : 434   Median :23.00
                    Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
                    3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
                    Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
   top25perc      f_undergrad     p_undergrad         outstate       room_board
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
 1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
 Median : 54.0   Median : 1707   Median :  353.0   Median : 9990   Median :4200
 Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
 3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
 Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
     books           personal         phd            terminal       s_f_ratio
 Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50
 1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50
 Median : 500.0   Median :1200   Median : 75.00   Median : 82.0   Median :13.60
 Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7   Mean   :14.09
 3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50
 Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0   Max.   :39.80
  perc_alumni        expend        grad_rate      Elite
 Min.   : 0.00   Min.   : 3186   Min.   : 10.00   Yes: 78
 1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00   No  :699
 Median :21.00   Median : 8377   Median : 65.00
 Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :64.00   Max.   :56233   Max.   :118.00
```

ii) Code :

```
pairs(college[ ,1:10])
```

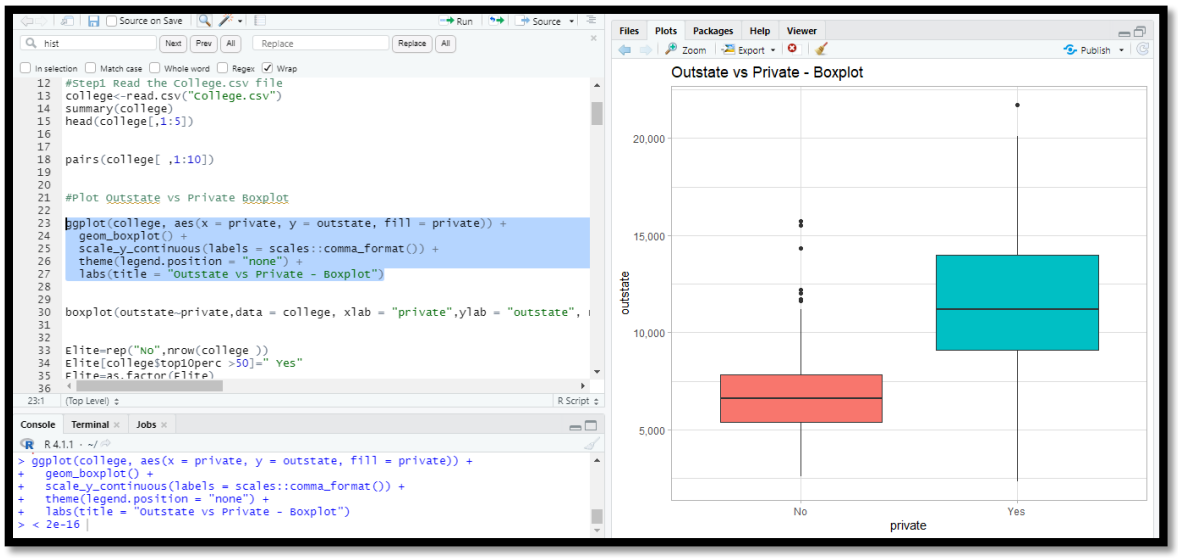
Output :



iii)Code :

```
ggplot(college, aes(x = private, y = outstate, fill = private)) +
  geom_boxplot() +
  scale_y_continuous(labels = scales::comma_format()) +
  theme(legend.position = "none") +
  labs(title = "Outstate vs Private - Boxplot")
```

Output:



iv) Code :

```
Elite=rep("No",nrow(college ))
Elite[college$top10perc >50]=" Yes"
Elite=as.factor(Elite)
college=data.frame(college , Elite)

summary(college$Elite)
```

Outtput :

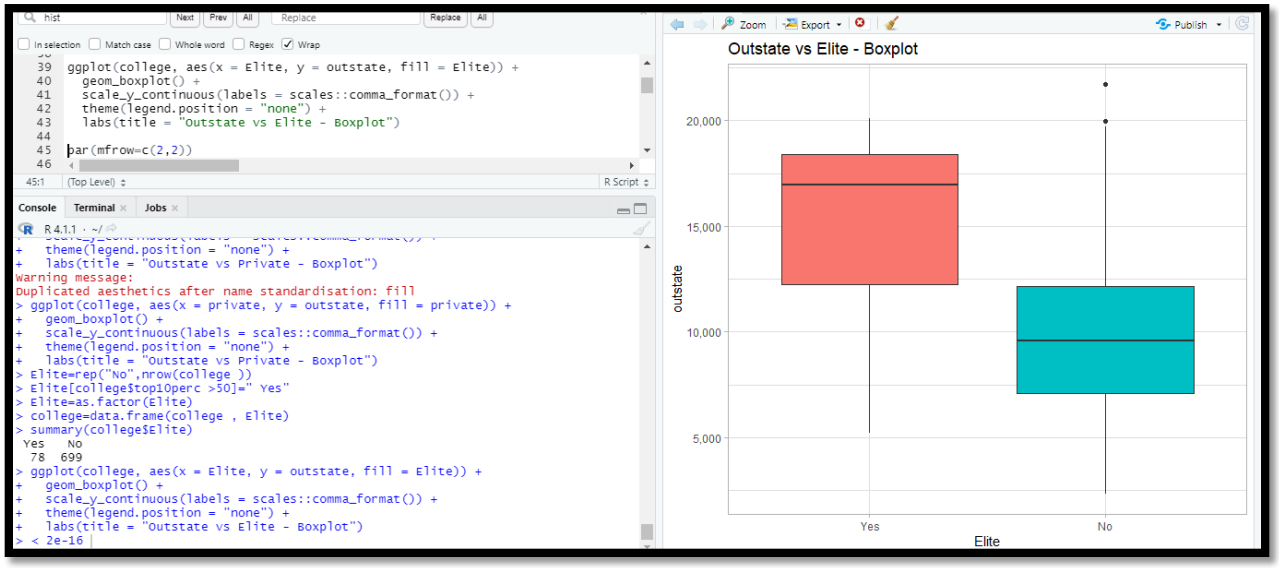
```
> summary(college$Elite)
 Yes    No
  78   699
```

Now we will plot Outstate vs Elite

Code :

```
ggplot(college, aes(x = Elite, y = outstate, fill = Elite)) +
  geom_boxplot() +
  scale_y_continuous(labels = scales::comma_format()) +
  theme(legend.position = "none") +
  labs(title = "Outstate vs Elite - Boxplot")
```

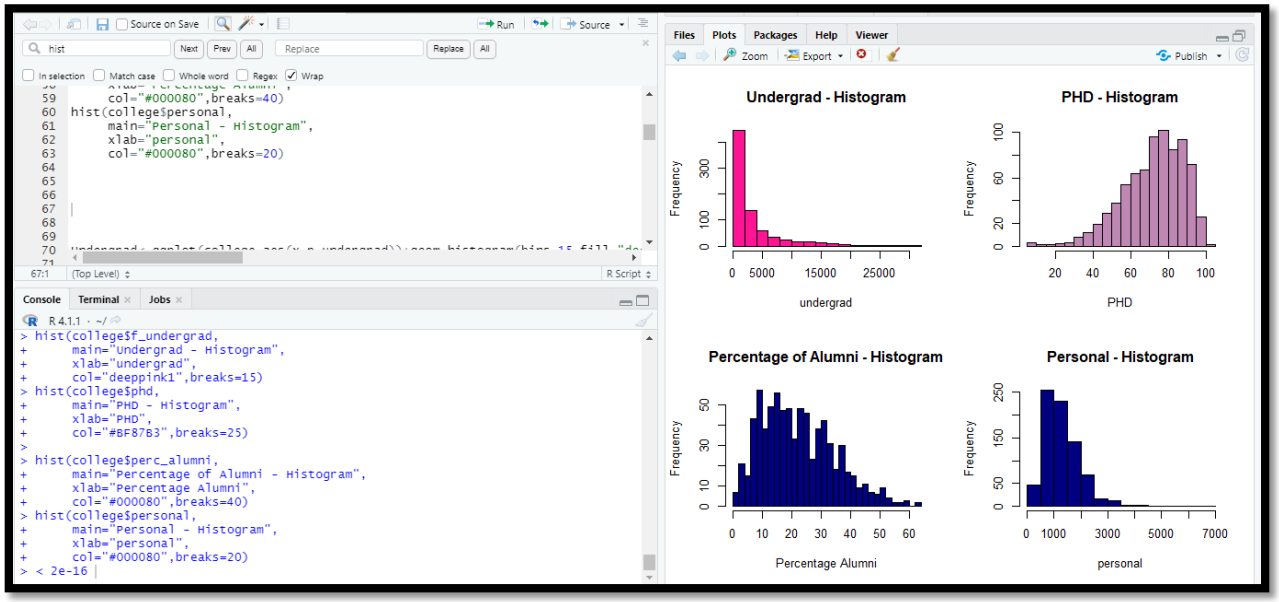
Output:



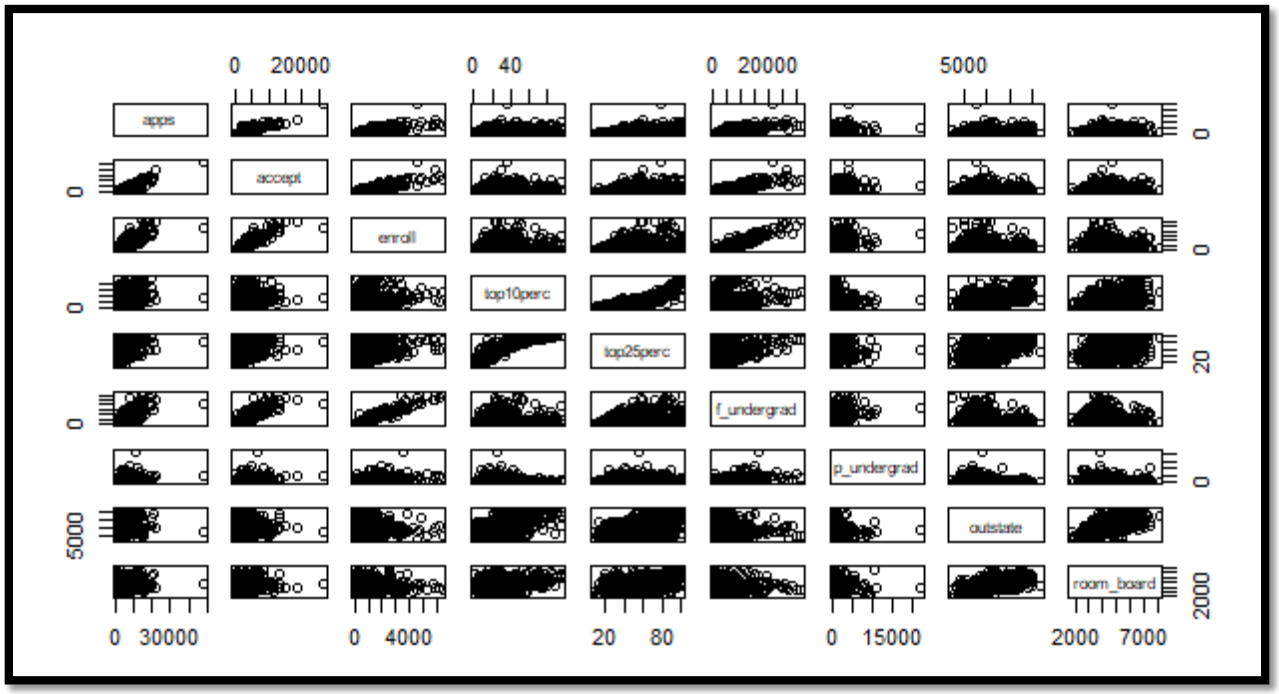
V) I will plot histograms for the following predictors – undergrad,phd,perc_alumni,personal
Code :

```
# To plot 4 histograms on the plot screen
par(mfrow=c(2,2))
hist(college$f_undergrad,
    main="Undergrad - Histogram",
    xlab="undergrad",
    col="deeppink1",breaks=15)
hist(college$phd,
    main="PHD - Histogram",
    xlab="PHD",
    col="#BF87B3",breaks=25)

hist(college$perc_alumni,
    main="Percentage of Alumni - Histogram",
    xlab="Percentage Alumni",
    col="#000080",breaks=40)
hist(college$personal,
    main="Personal - Histogram",
    xlab="personal",
    col="#000080",breaks=20)
```

Output :



vi)
I will start with referring to the scatterplot matrix to get an overview of the relationships between various predictors,



9)
a) The following predictors are **Quantitative** :

- mpg – Miles per gallon
- Cylinders – Number of cylinders
- Displacement – Measured in Cubic inches
- Horsepower – Engine power
- Weight – Weight of the vehicle
- Acceleration – Measures as number of miles travelled per hour
- Year –Model year

The following predictors are **Qualitative**:

- Origin – Describes origin of the car
- Name – Vehicle name

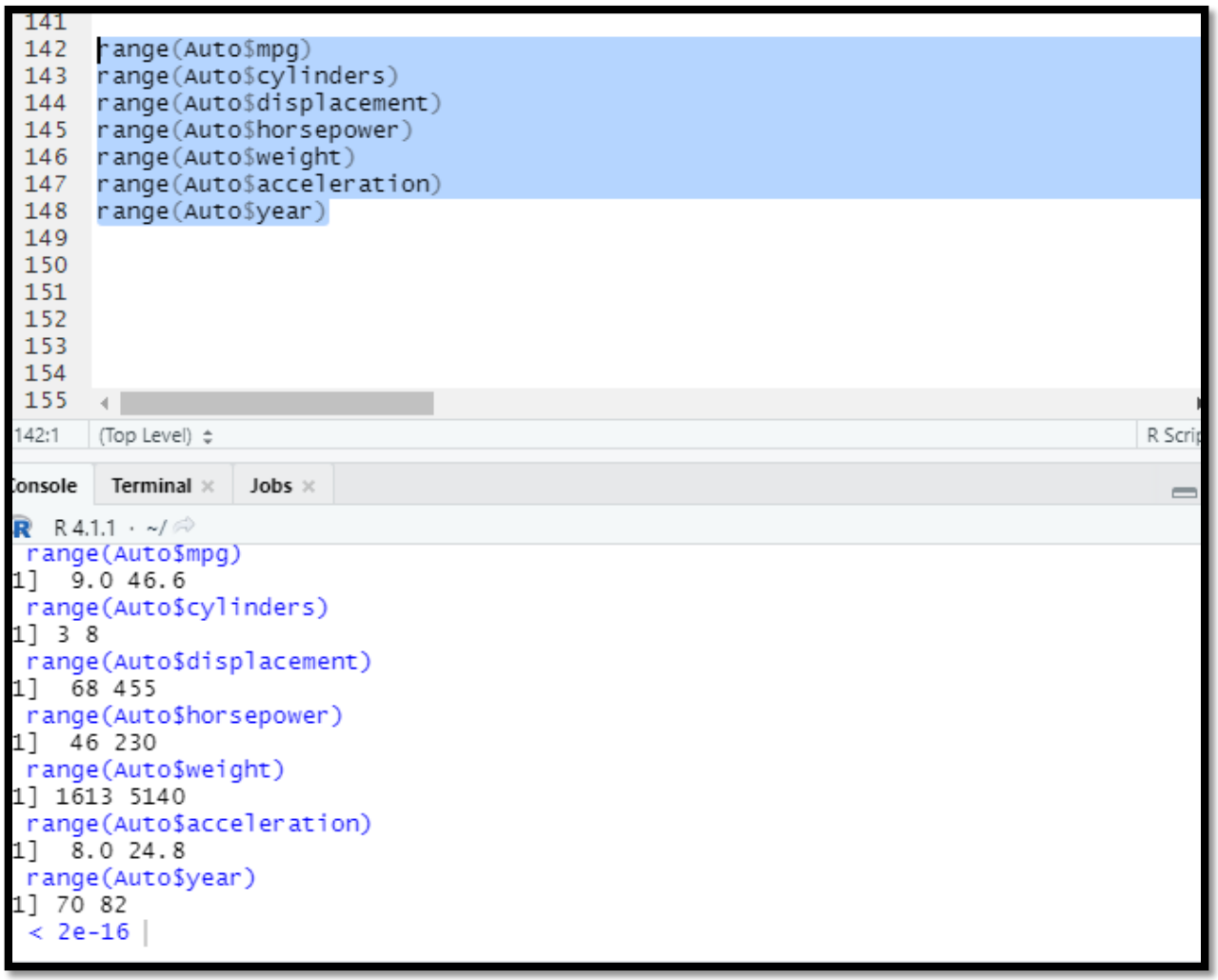
b) Range of Quantitative predictors:

Code :

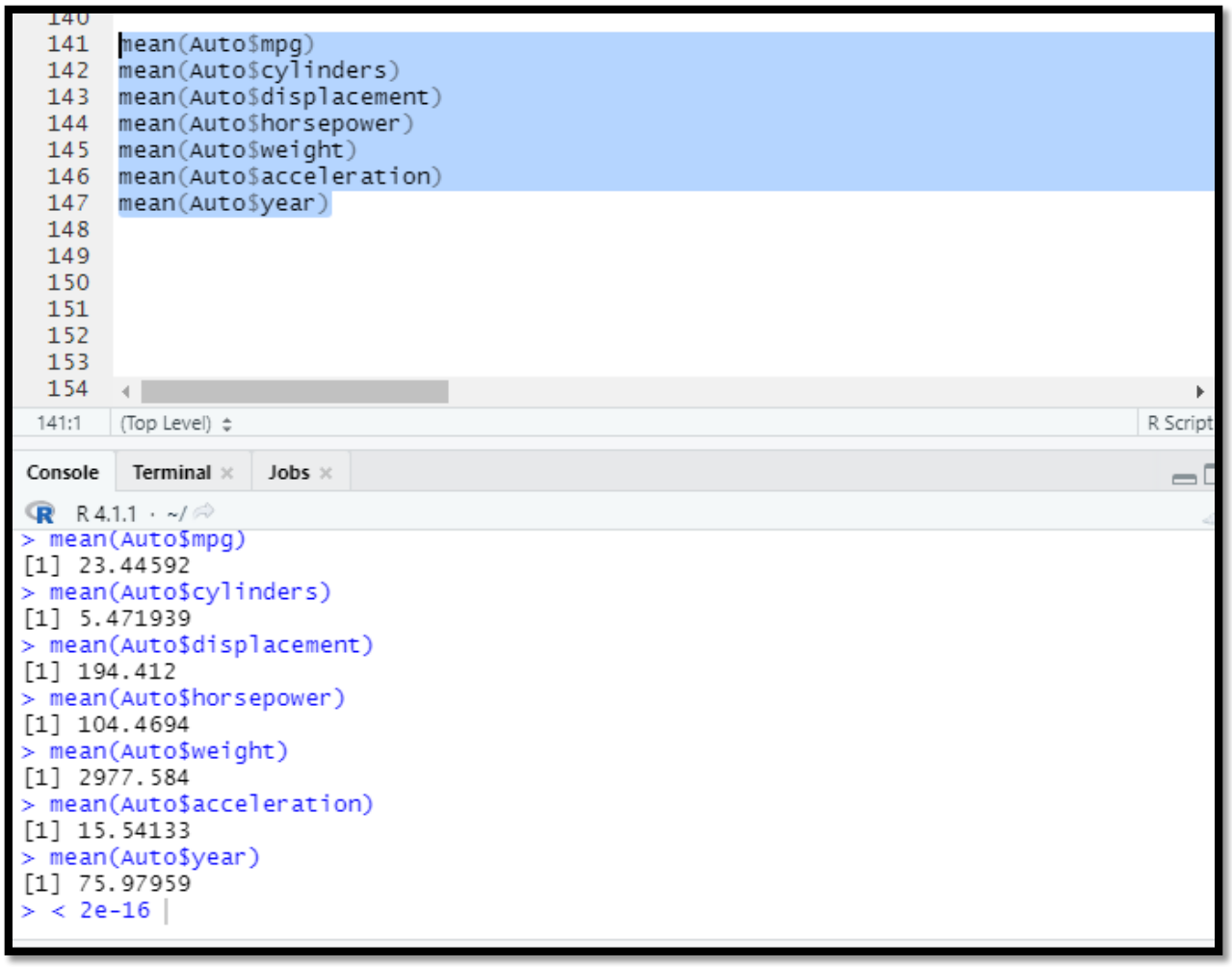
```
range(Auto$mpg)
range(Auto$cylinders)
range(Auto$displacement)
```

range(Auto$horsepower)
range(Auto$weight)
range(Auto$acceleration)
range(Auto$year)

Output:



c)
Mean :
Code :
mean(Auto$mpg)
mean(Auto$cylinders)
mean(Auto$displacement)
mean(Auto$horsepower)
mean(Auto$weight)
mean(Auto$acceleration)
mean(Auto$year)
Output for Mean :

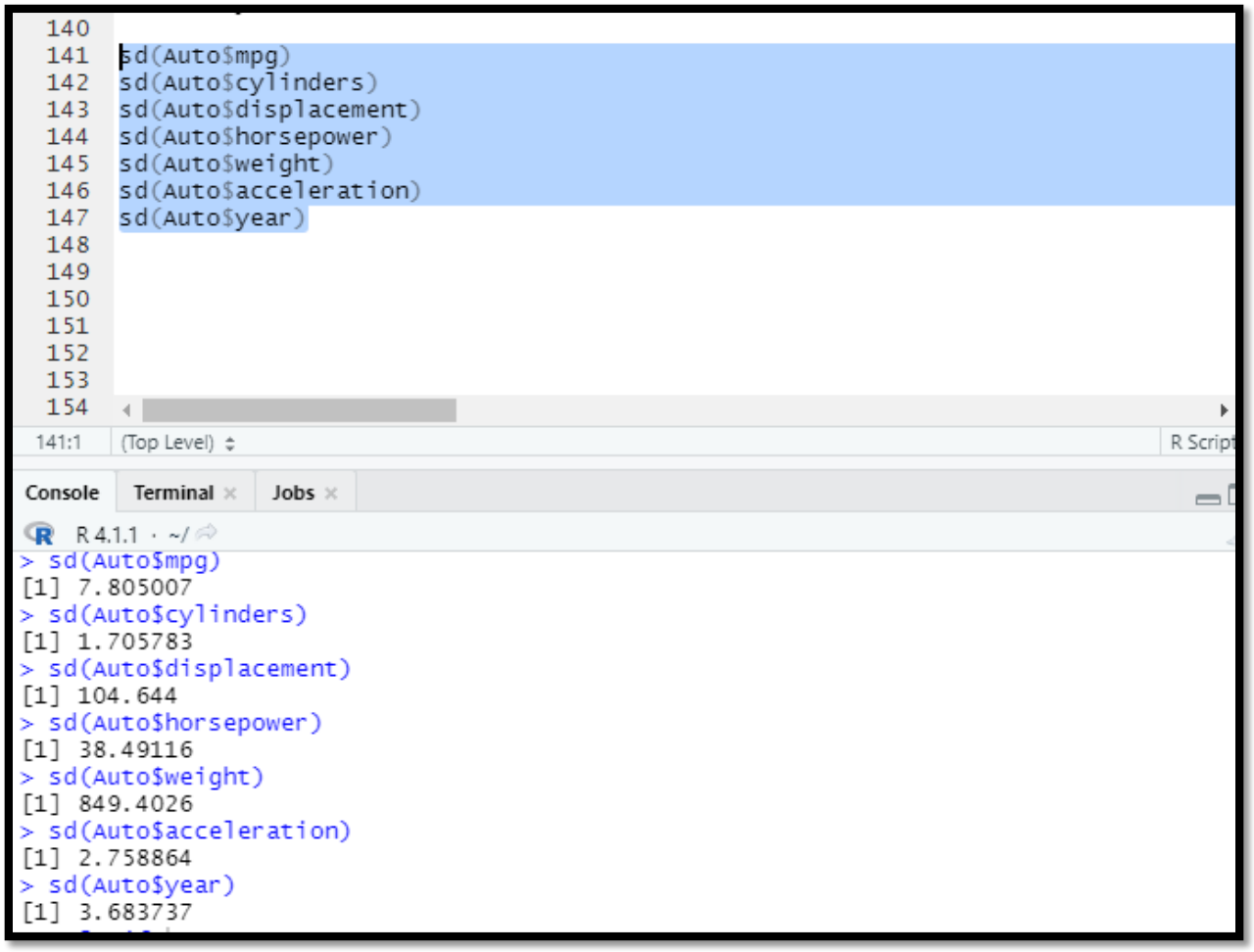


Standard Deviation :

Code:

```
sd(Auto$mpg)
sd(Auto$cylinders)
sd(Auto$displacement)
sd(Auto$horsepower)
sd(Auto$weight)
sd(Auto$acceleration)
sd(Auto$year)
```

Output for SD:



d)

Code to remove the 10th and 85th observations:

Auto1<-Auto[-c(10:85),]

Code to calculate range of all predictors for remaining data :

Output:

range(Auto1$mpg)

range(Auto1$cylinders)

range(Auto1$displacement)

range(Auto1$horsepower)

range(Auto1$weight)

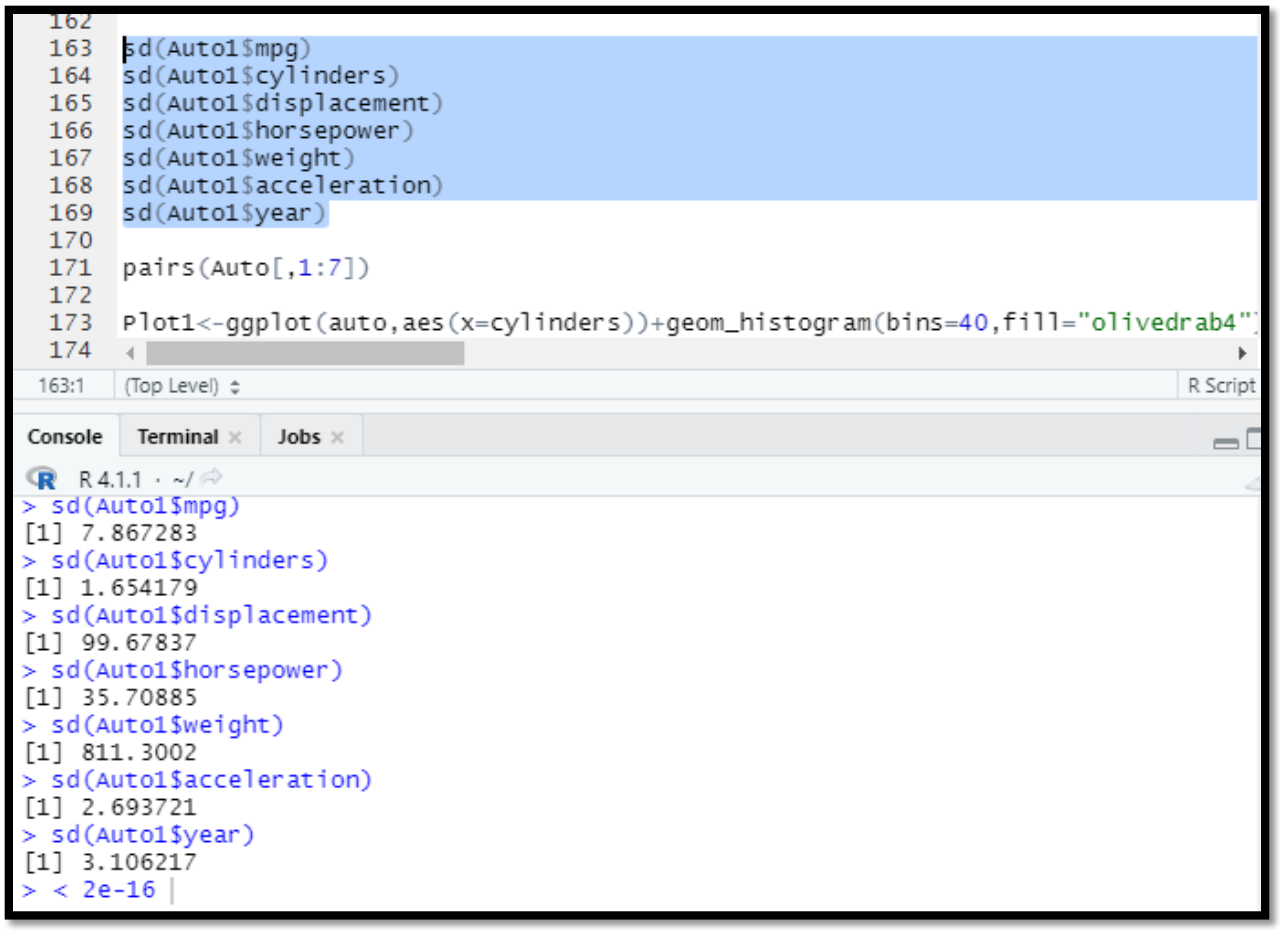
range(Auto1$acceleration)

range(Auto1$year)

Code to calculate the standard deviation for the remaining data :

```
sd(Auto1$mpg)
sd(Auto1$cylinders)
sd(Auto1$displacement)
sd(Auto1$horsepower)
sd(Auto1$weight)
sd(Auto1$acceleration)
sd(Auto1$year)
```

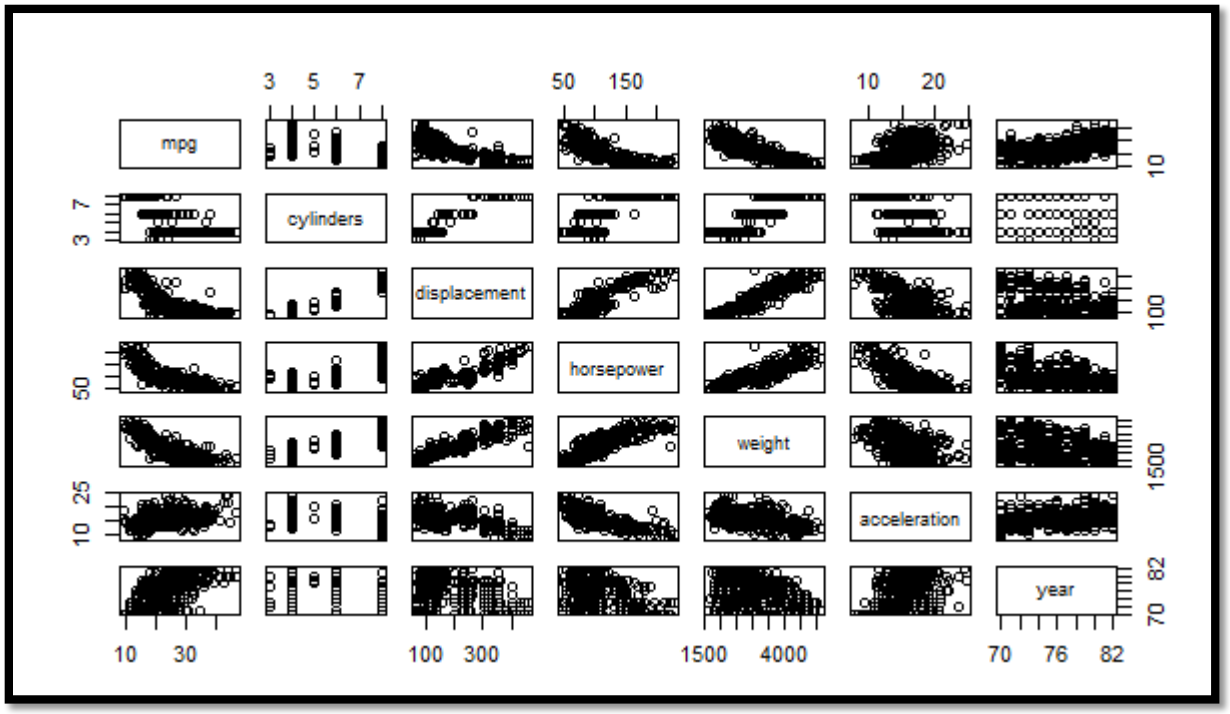
Output:



Code to calculate the mean for the remaining data :

e) I would like to determine the relationship between various predictors. I will plot a scatterplot matrix to get an overview of the relationships between various predictors,
The scatter matrix below provides an overview of the relations between various predictors,
Code:
pairs(Auto[,1:7])

Output:

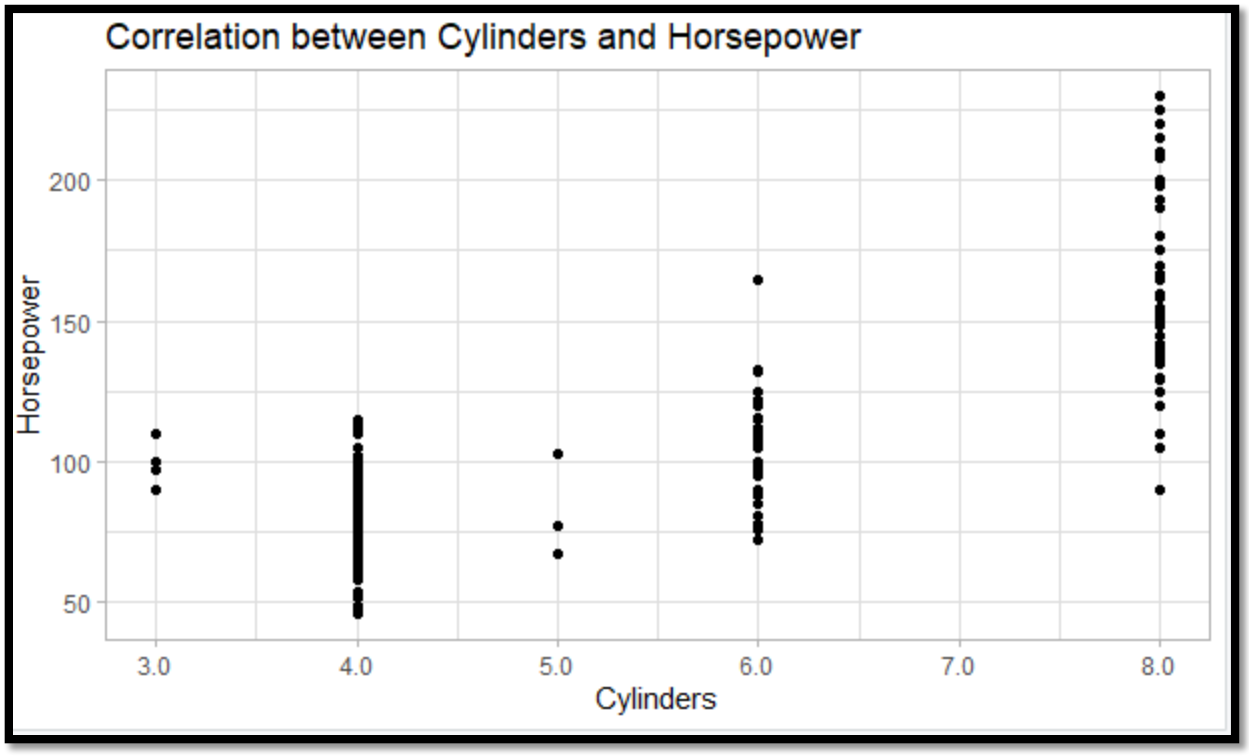


The scatterplot below depicts the Correlation between **Number of Cylinders and horsepower**. It can be observed that the predictors are positively related.

Code:

```
chrel<-ggplot(Auto, aes(x =cylinders , y = horsepower)) +
  geom_point() +
  scale_x_continuous(labels = scales::comma_format()) +
  labs(x = "Cylinders",
       y = "Horsepower",
       title = "Correlation between Cylinders and Horsepower")
chrel
```

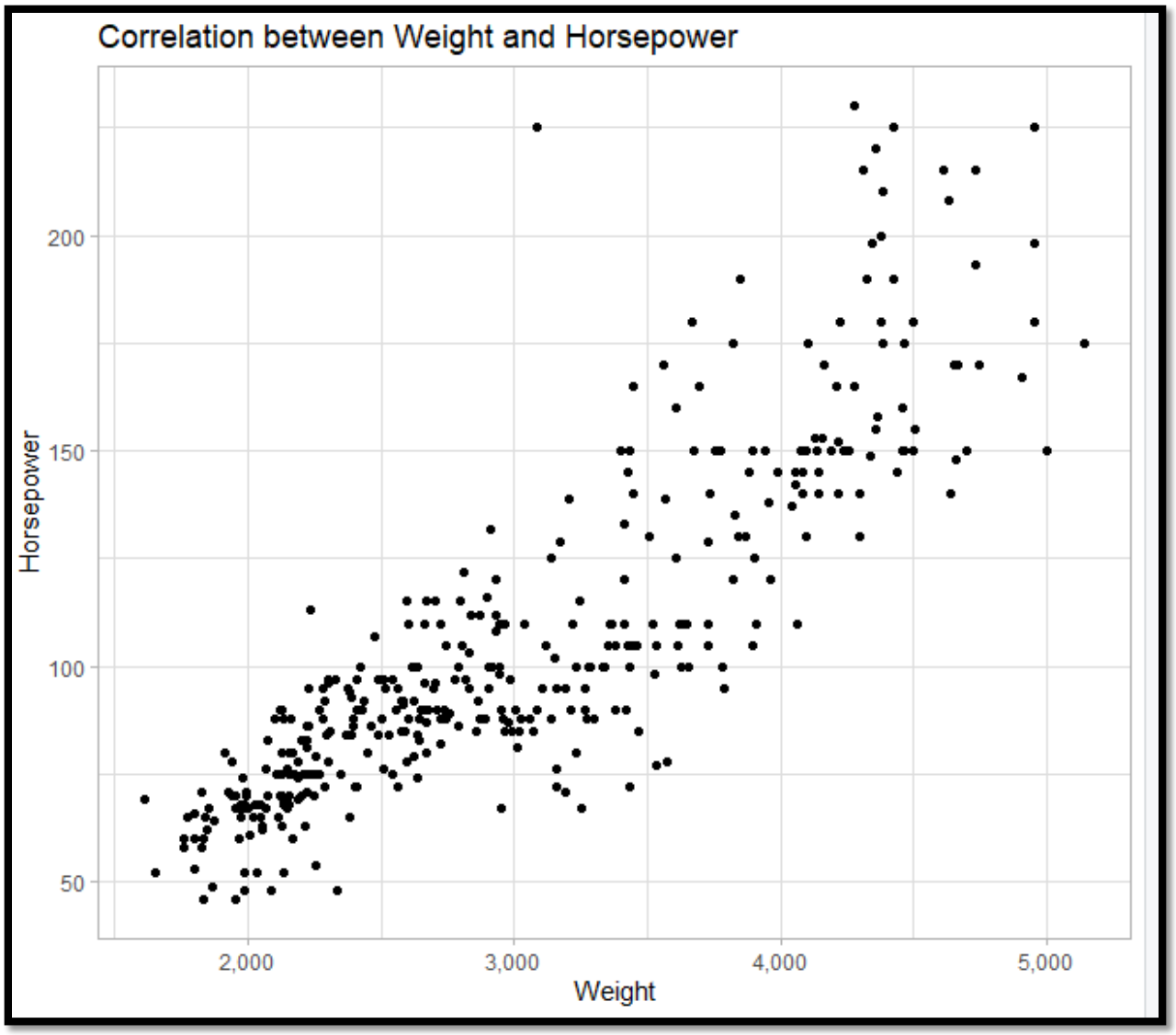
Output:



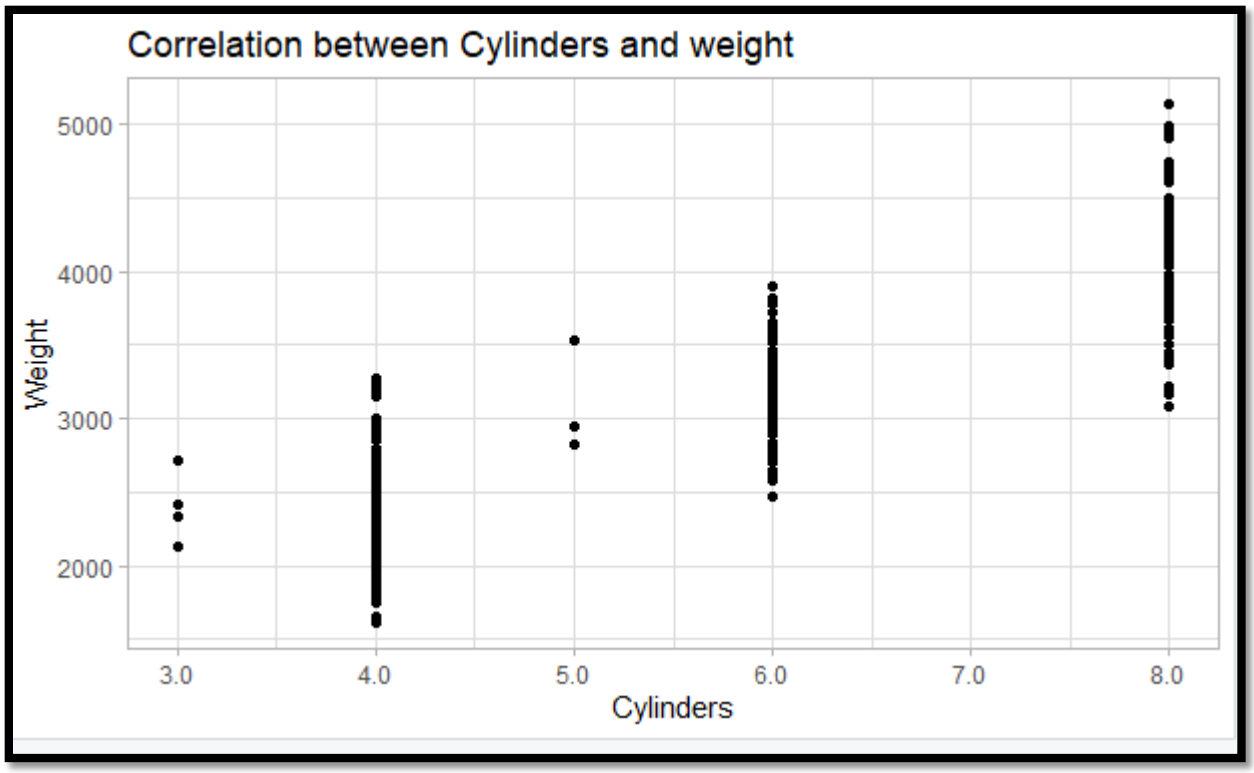
The scatterplot below depicts correlation between **horsepower and Weight**.It can be observed that the 2 predictors are positively related.

Code:

```
wH<-ggplot(Auto, aes(x =weight , y = horsepower)) +
  geom_point() +
  scale_x_continuous(labels = scales::comma_format()) +
  labs(x = "Weight",
       y = "Horsepower",
       title = "Correlation between Weight and Horsepower")
wH
```



The scatterplot below depicts correlation between Number of Cylinders and weight. It can be inferred that these predictors have positive relationship.



I will now use a scatterplot to determine the correlation between weight and acceleration

Code:

```
wa<-ggplot(Auto, aes(x = weight, y = acceleration)) +
  geom_point() +
  theme(legend.position = "none") +
  labs(x = "Weight",
       y = "Acceleration",
       title = "Correlation between weight and acceleration")
wa
```

Output:

The data points in the graph are too scattered, therefore the correlation is weak. I will try to examine the correlation based on origin

Code :

```
Auto$origin <- factor(Auto$origin, labels = c("American", "Australian", "Indian"))
ggplot(Auto, aes(x = weight, y = acceleration, col = origin)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~ origin) +
  theme(legend.position = "none") +
  labs(x = "Weight",
       y = "Acceleration",
       title = "Correlation between weight and acceleration, by origin")
```

Output:

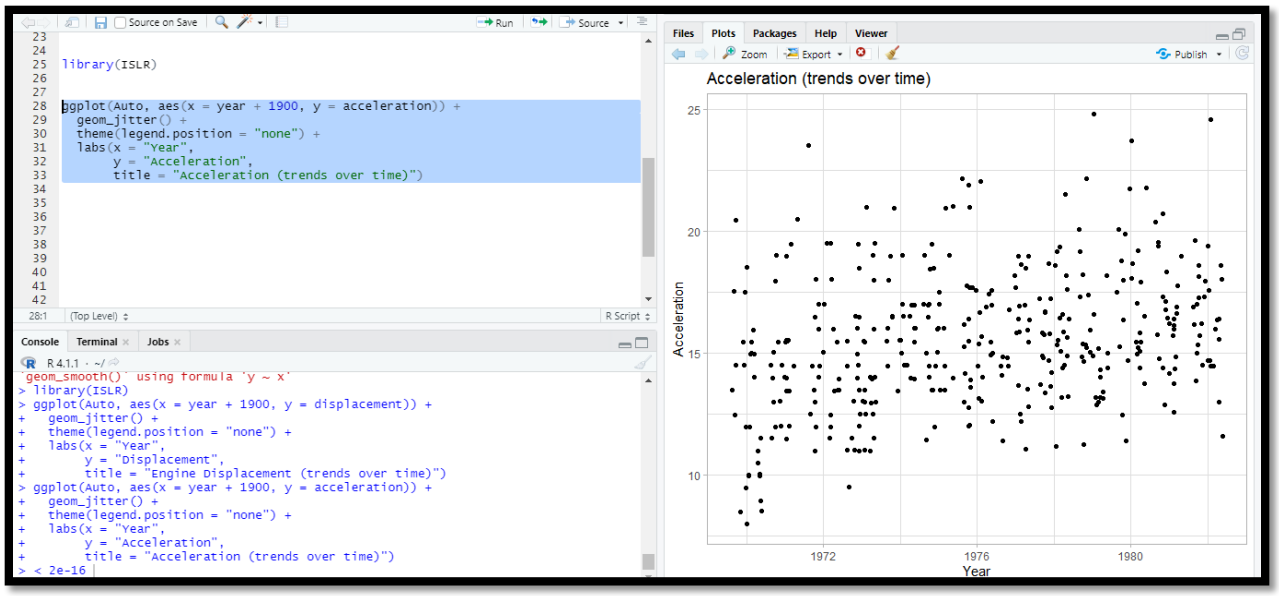
It can be observed that the weight of cars is low in Australia and India when compared to America.The weight of cars did not exceed 4000 in Australia and India. Whereas in America, cars have exceeded 4000 in weight.

I will now determine how acceleration of cars has changed over time,

Code :

```
ggplot(Auto, aes(x = year + 1900, y = acceleration)) +
  geom_jitter() +
  theme(legend.position = "none") +
  labs(x = "Year",
       y = "Acceleration",
       title = "Acceleration (trends over time)")
```

Output:

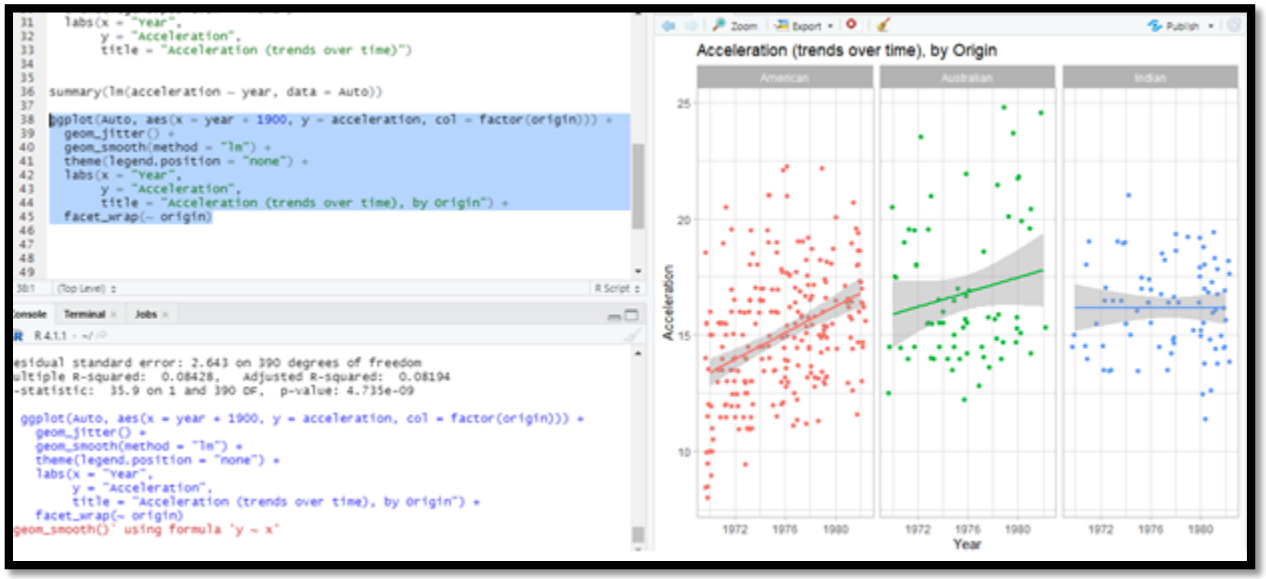


The data points are too scattered, so I will now plot this trend across various origins to analyze further

Code:

```
ggplot(Auto, aes(x = year + 1900, y = acceleration, col = factor(origin))) +
  geom_jitter() +
  geom_smooth(method = "lm") +
  theme(legend.position = "none") +
  labs(x = "Year",
       y = "Acceleration",
       title = "Acceleration (trends over time), by Origin") +
  facet_wrap(~ origin)
```

Output:



It can be observed that the of acceleration of American and Australian cars has increased over time. Whereas, acceleration of Indian cars has remained almost constant.

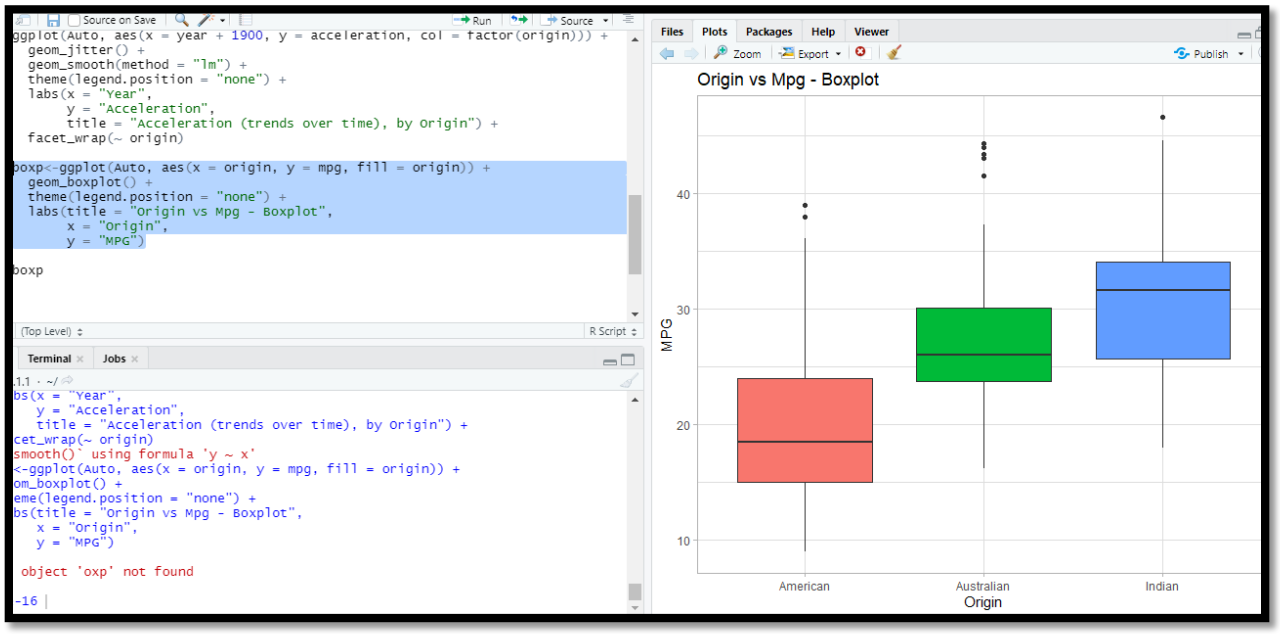
f) By looking at the scatter matrix we plotted earlier,it can be observed that mpg holds a relationship with other predictors.

I would like to determine how mpg is varying across various origins.I will plot a boxplot to determine the same,

Code:

```
boxp<-ggplot(Auto, aes(x = origin, y = mpg, fill = origin)) +
  geom_boxplot() +
  theme(legend.position = "none") +
  labs(title = "Origin vs Mpg - Boxplot",
       x = "Origin",
       y = "MPG")
```

Output:



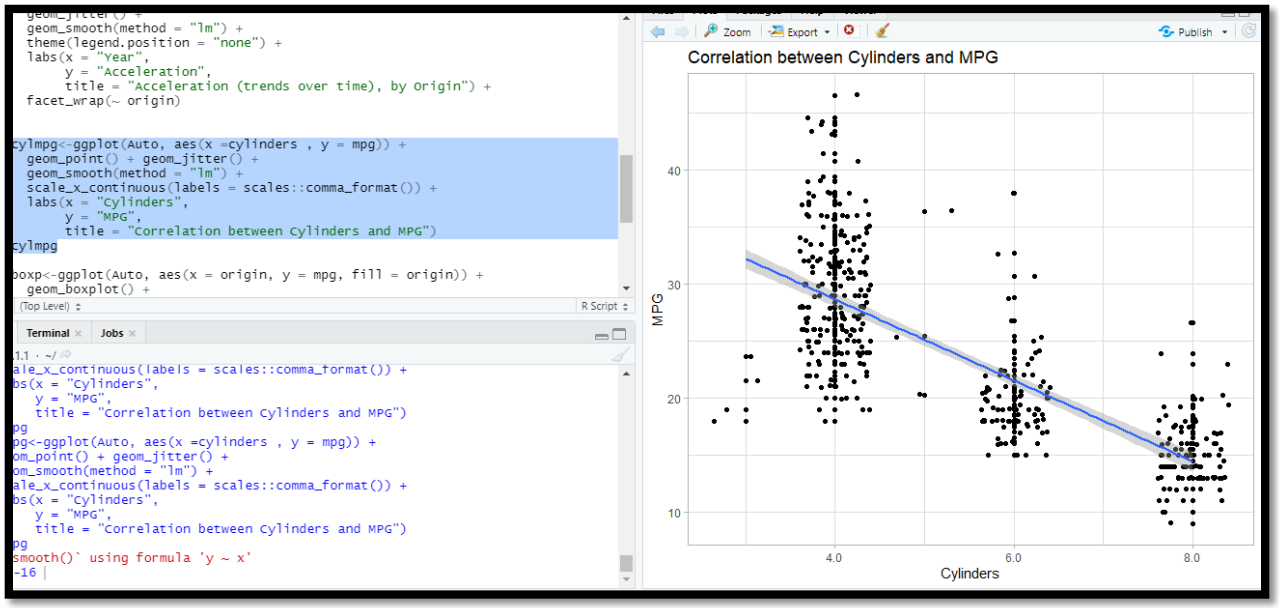
It can be inferred that the MPG is greater in Australian and Indian cars in comparison with American cars.

The correlation between mpg and cylinders seems interesting.I will plot a jitterplot to analyze further,

Code :

```
cylmpg<-ggplot(Auto, aes(x =cylinders , y = mpg)) +
  geom_point() + geom_jitter() +
  geom_smooth(method = "lm") +
  scale_x_continuous(labels = scales::comma_format()) +
  labs(x = "Cylinders",
       y = "MPG",
       title = "Correlation between Cylinders and MPG")
cylmpg
```

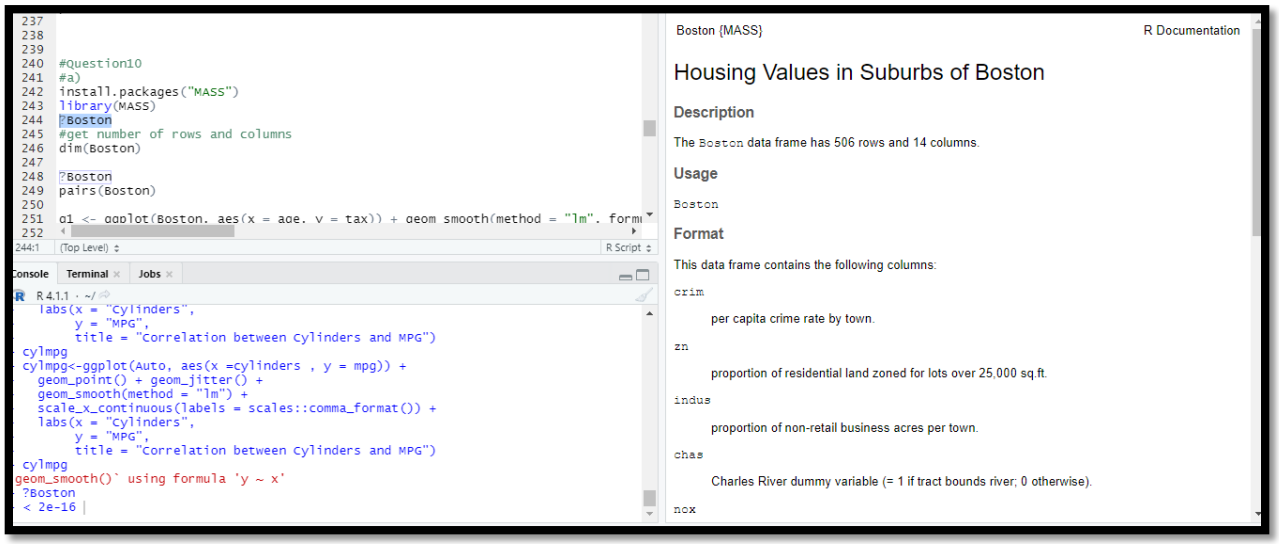
Output:



It can be inferred that the relationship between Cylinders and mpg is negative.

10)

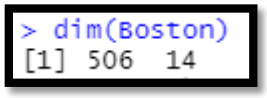
a)Code :



Code to determine number of rows and columns in the data set:

dim(Boston)

Output:

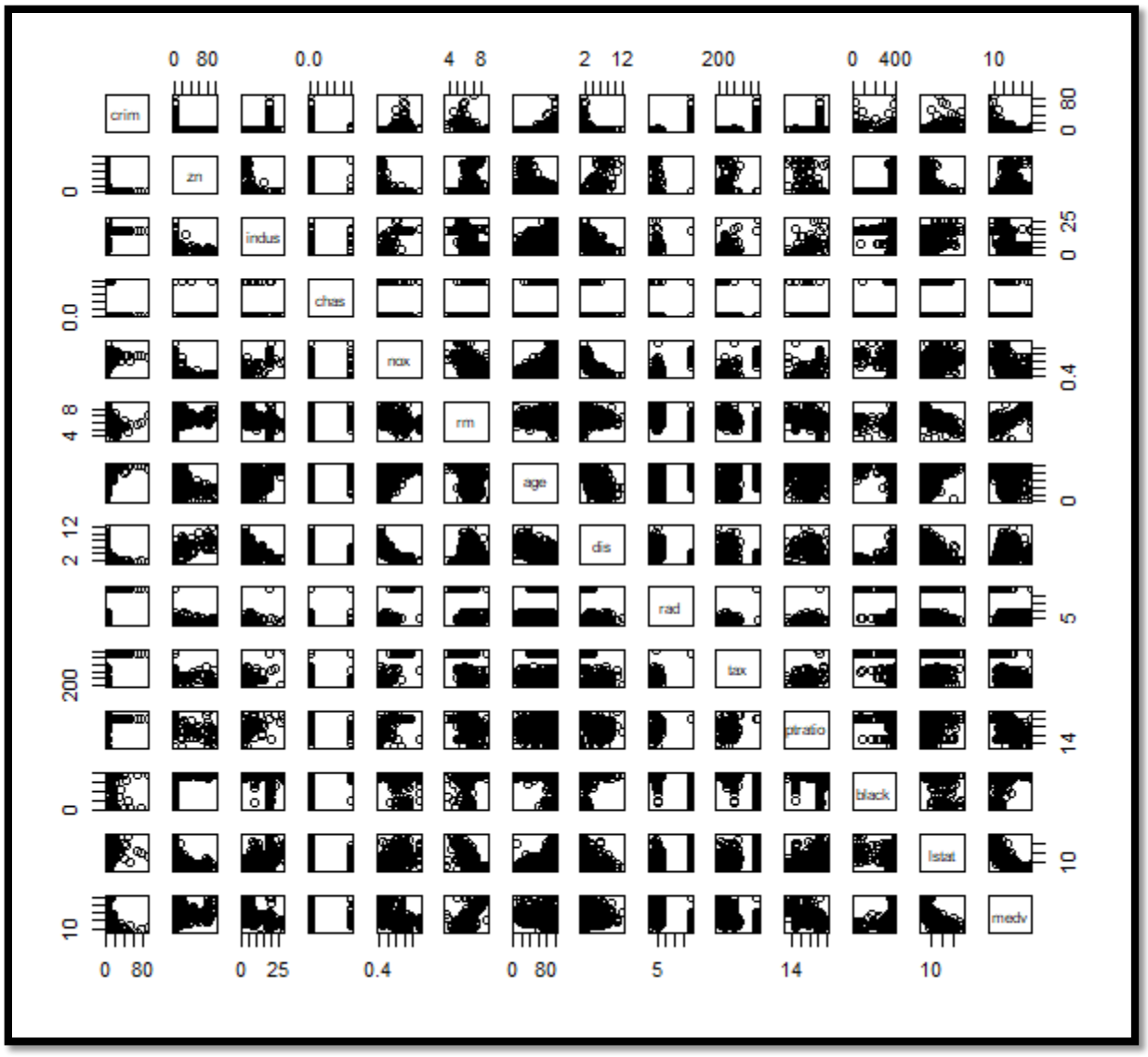


There are 506 rows and 14 columns in the data. Rows represent the suburbs of Boston.

Columns represent as below:

- *crim* - Per capita crime rate by town
- *zn* - Proportion of residential land zoned for lots over 25,000 sq.ft.
- *indus* - Proportion of non-retail business acres per town
- *chas* - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- *nox* - Nitrogen oxides concentration (parts per 10 million)
- *rm* - Average number of rooms per dwelling
- *age* - Proportion of owner-occupied units built prior to 1940
- *dis* - Weighted mean of distances to five Boston employment centres
- *rad* - Index of accessibility to radial highways
- *tax* - Full-value property-tax rate per $10,000
- *ptratio* - Pupil-teacher ratio by town
- *black* - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- *lstat* - Lower status of the population (percent)
- *medv* - Median value of owner-occupied homes in $1000s

b)Below is the pairs scatterplot:



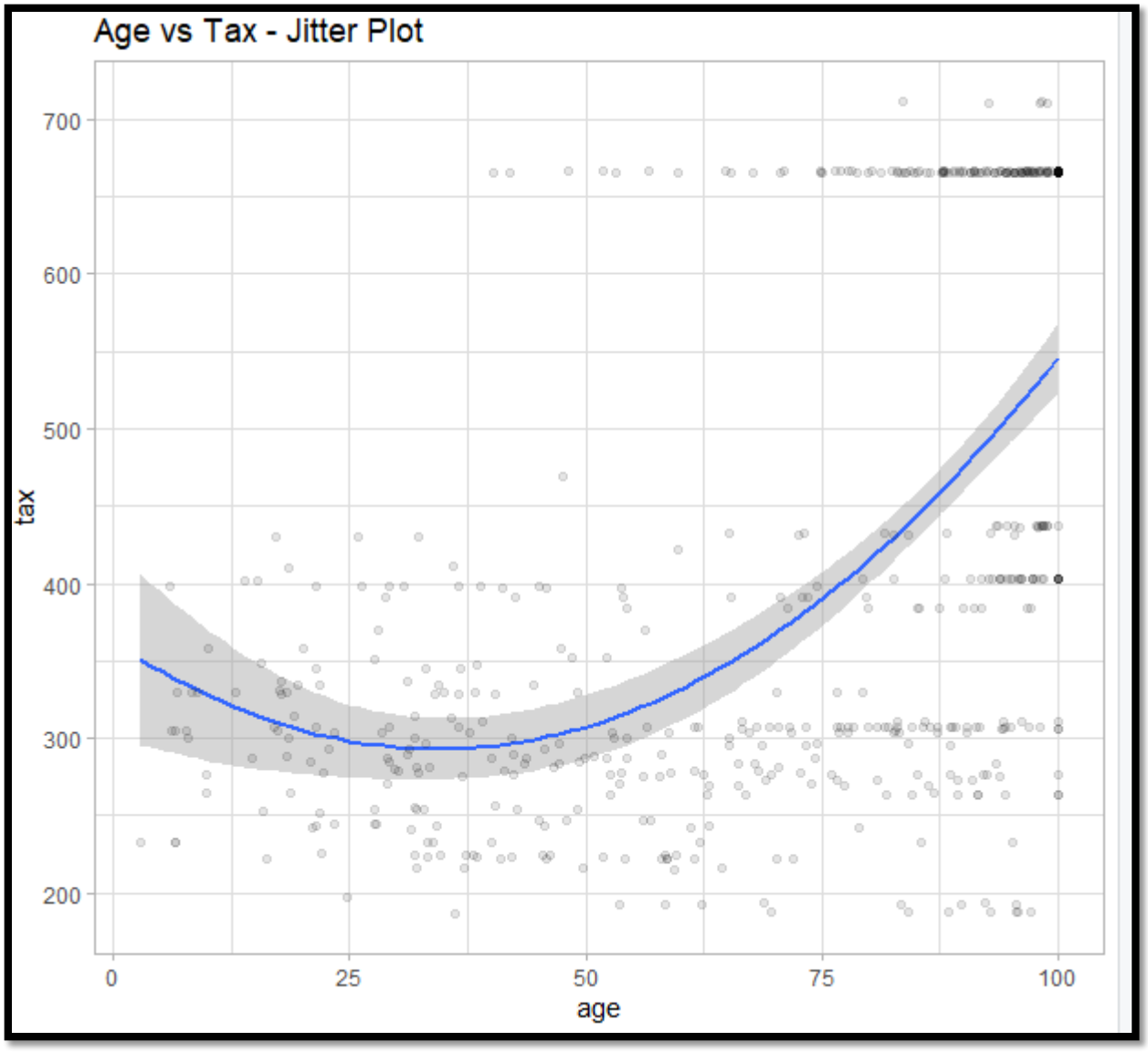
Below jitterplot depicts the relationship between age and tax,

Code:

```
g1 <- ggplot(Boston, aes(x = age, y = tax)) + geom_smooth(method = "lm", formula = "y ~ x + I(x^2)") +

  scale_x_continuous(labels = scales::comma_format()) +

  geom_jitter(alpha = 0.1) + labs(title = "Age vs Tax - Jitter Plot",

       x = "age",

       y = "tax")
```

g1

Output:



It can be observed from the scatterplot that after a specific age(Proportion of owner-occupied units built prior to 1940) tax is positively related to age.

Code to determine correlation between age and tax:

summary(lm(age~tax,data=Boston))

Output:

```
> summary(lm(age~tax,data=Boston))

Call:
lm(formula = age ~ tax, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-61.709 -16.030   4.518  18.396  47.054

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.043105   2.832797   12.02   <2e-16 ***
tax          0.084588   0.006415   13.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value of tax vs age is significant i.e. these predictors are strong related.

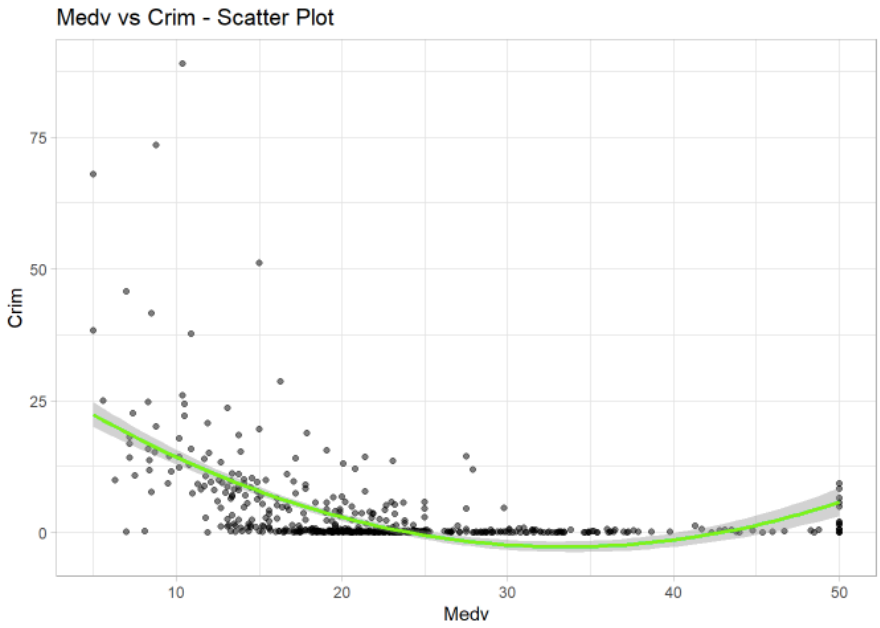
c)

best_predictor(Boston, "crim")

Code:

```
ggplot(Boston, aes(x = medv, y = crim)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = "y ~ x + I(x^2)", col = "green") +
  labs(title = "Medv vs Crim - Scatter Plot",
       x = "Medv",
       y = "Crim")
```

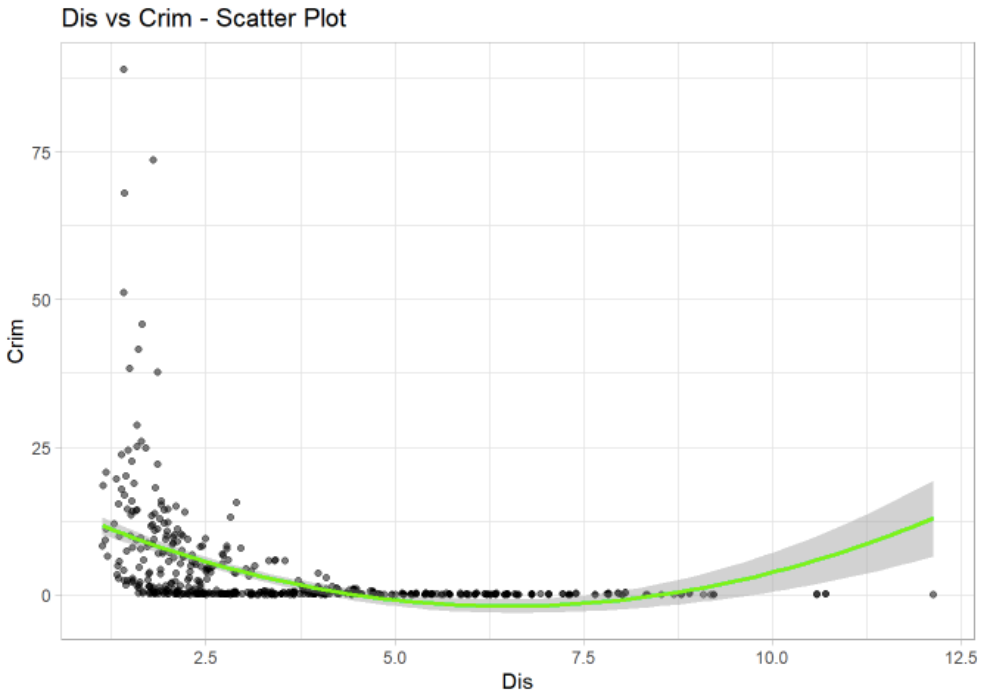
Output:



code:

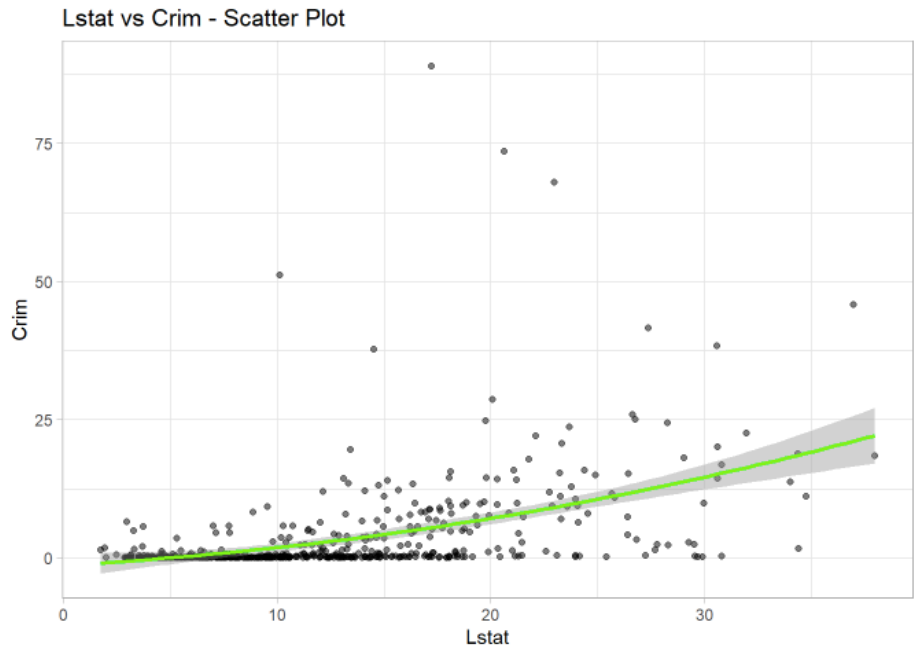
```
ggplot(Boston, aes(x = dis, y = crim)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = "y ~ x + I(x^2)", col = "green") +
  labs(title = "Dis vs Crim - Scatter Plot",
       x = "Dis",
       y = "Crim")
```

output:




```
ggplot(Boston, aes(x = lstat, y = crim)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = "y ~ x + I(x^2)", col = "green") +
  labs(title = "Lstat vs Crim - Scatter Plot",
       x = "Lstat",
       y = "Crim")
```

Output:



d)

Crime Rates

Range of crim = 0.00632 to 88.97620

Code:

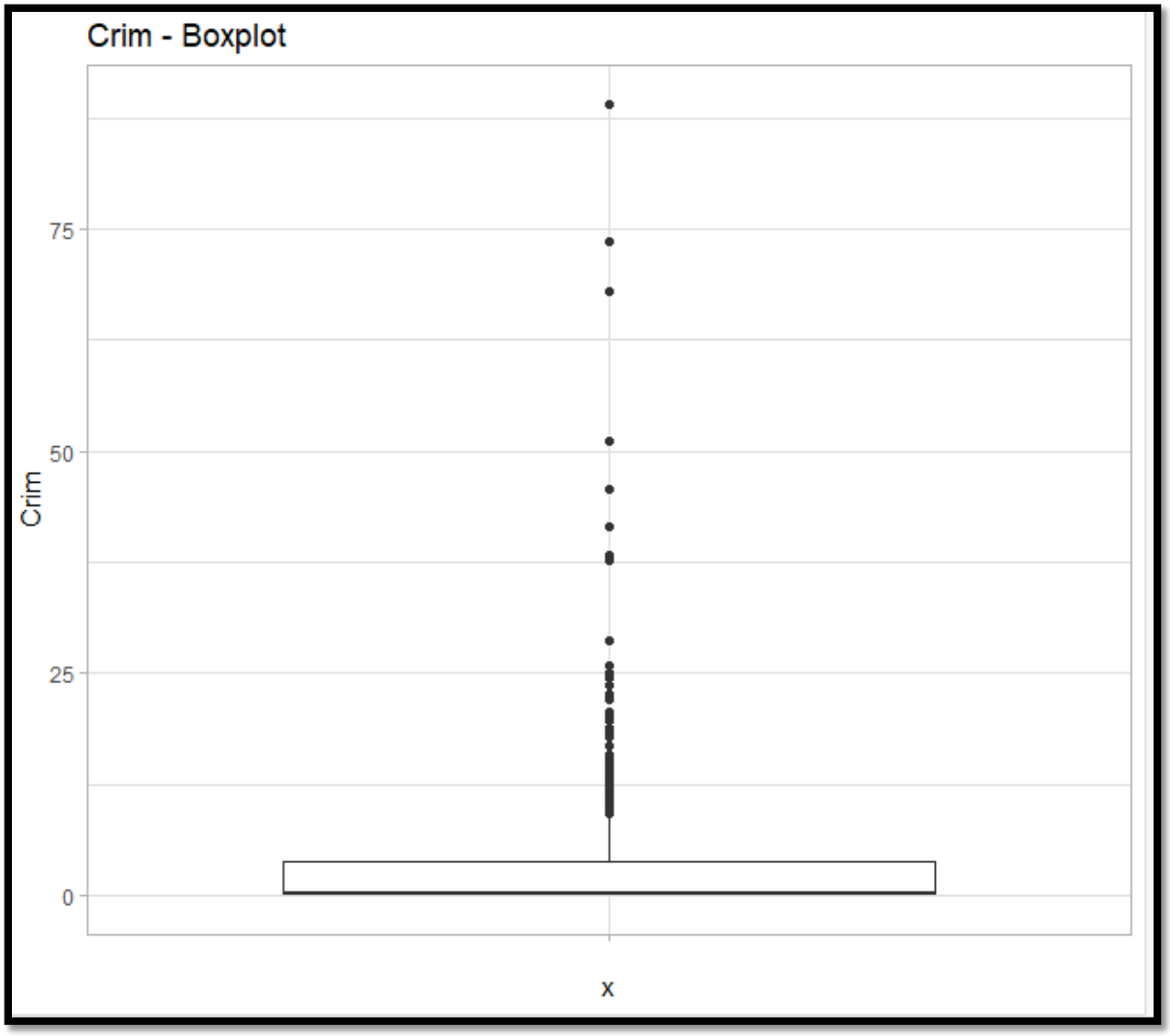
```
> range(Boston$crim)
[1]  0.00632 88.97620
```

This indicates a very wide range.I will use a boxplot to understand how Per capita crime rate is varying across various suburbs.

Code:

```
cb<-ggplot(Boston, aes(x = "", y = crim)) +
  geom_boxplot() + labs(y = "Crim",
       title = "Crim - Boxplot")
cb
```

Output:



It can be observed from the above boxplot that these is a huge variation in the Per capita crime rate across various suburbs .Though most areas show less crime rate, there are few areas in the suburbs where crime rate is very high.

Tax Rates

Range of tax- 187 to 711

Code:

```
> range(Boston$tax)
[1] 187 711
```

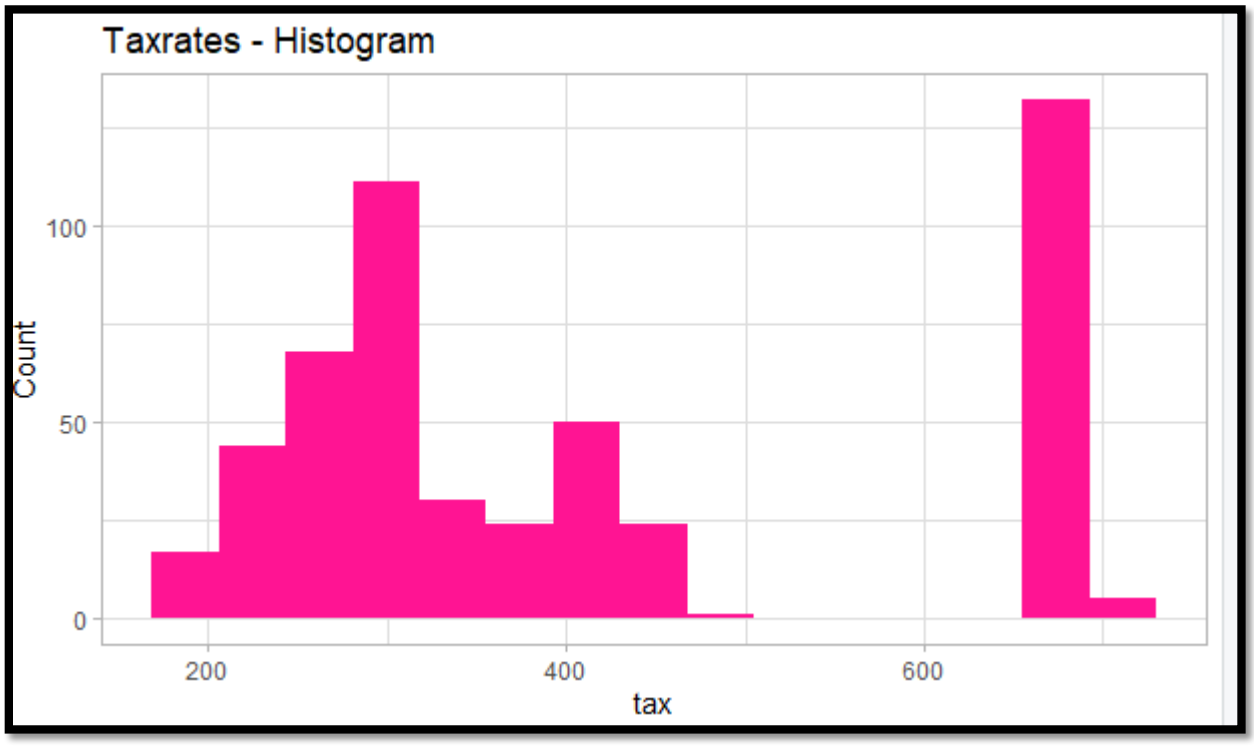
This range is smaller when compared to Crime rates.I will plot a histogram to understand how taxes vary across different areas,

Code:

taxrates<-ggplot(Boston,aes(x=tax))+geom_histogram(bins=15,fill="deeppink1")+labs(title = "Taxrates - Histogram",y = "Count")

taxrates

Output:



It can be observed that most areas have similar tax rate. group_by function cann be used to analyze further,

Code:

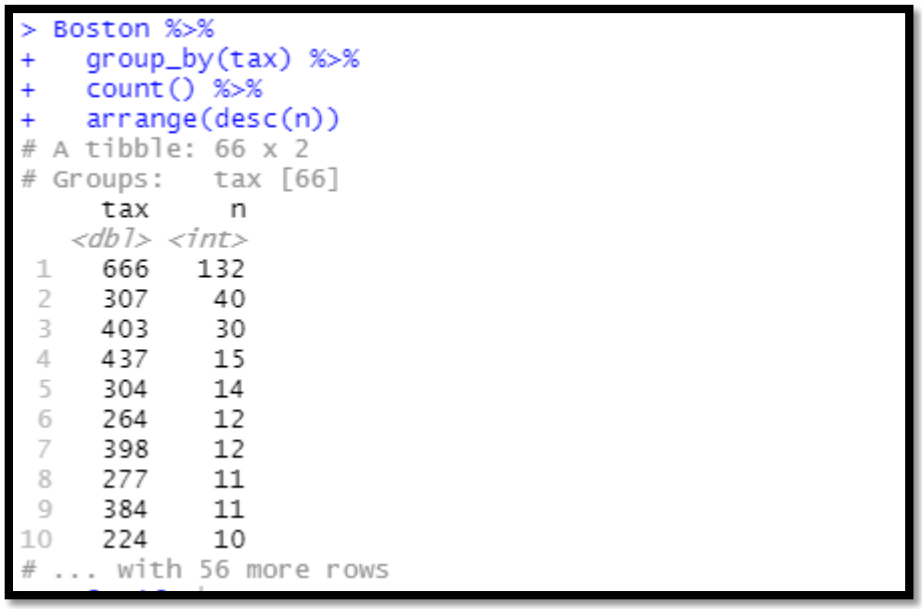
Boston %>%

group_by(tax) %>%

count() %>%

arrange(desc(n))

Output:



```
> Boston %>%
+   group_by(tax) %>%
+   count() %>%
+   arrange(desc(n))
# A tibble: 66 x 2
# Groups:   tax [66]
     tax     n
   <dbl> <int>
 1   666   132
 2   307    40
 3   403    30
 4   437    15
 5   304    14
 6   264    12
 7   398    12
 8   277    11
 9   384    11
10   224    10
# ... with 56 more rows
```

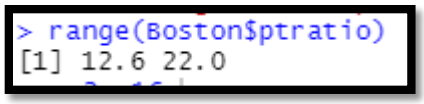
Pupil-teacher ratios

Range of ptratio : 12.6 22.0 – This is a very small range.

Code:

Range(Boston$ptratio)

Output:



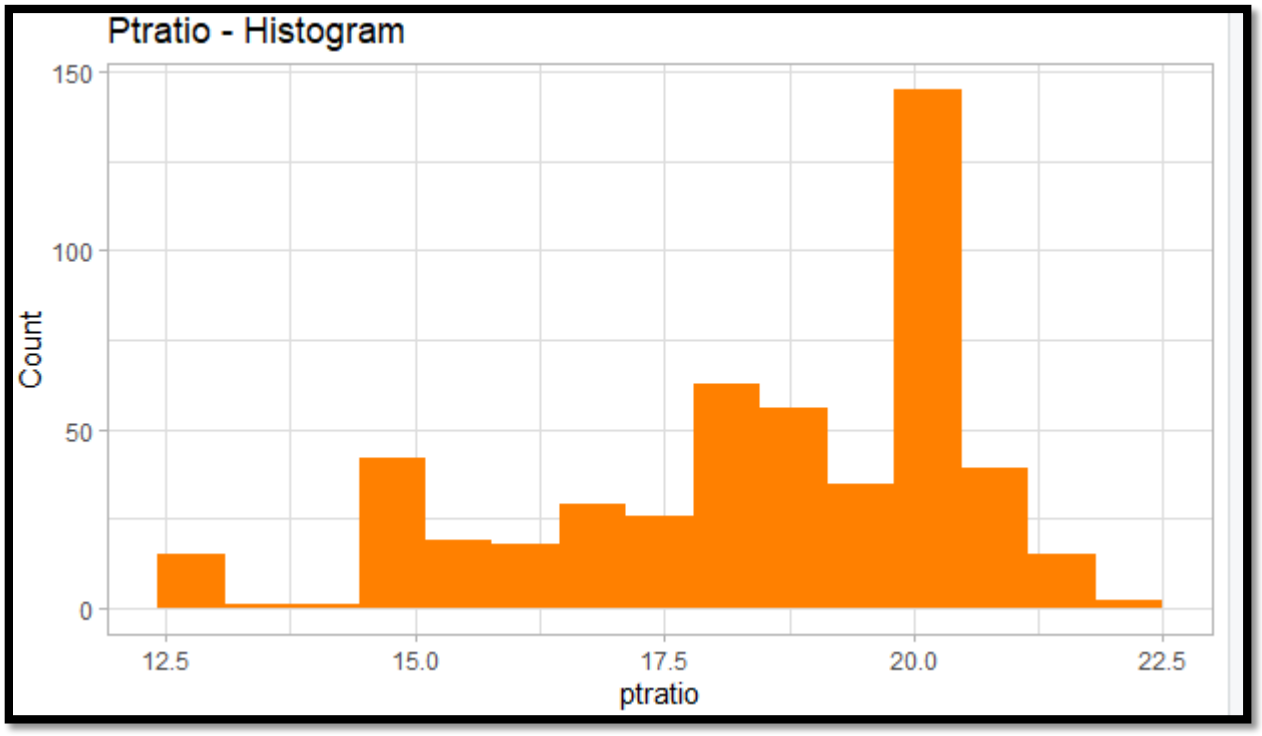
```
> range(Boston$ptratio)
[1] 12.6 22.0
```

Histogram below depicts the values of ptratio

Code:

```
ptratio<-ggplot(Boston,aes(x=ptratio))+geom_histogram(bins=15,fill="#FF8000FF")+labs(title = "Ptratio - Histogram",y = "Count")

ptratio
```

Output:



It can be observed that most suburbs have a similar Pupil-teacher ratio.group_by function can be used to analyze this further,

Code:

```
Boston %>%

 group_by(ptratio) %>%

 count() %>%

 arrange(desc(n))
```

Output:

```
> Boston %>%
+   group_by(ptratio) %>%
+   count() %>%
+   arrange(desc(n))
# A tibble: 46 x 2
# Groups:   ptratio [46]
   ptratio     n
     <dbl> <int>
 1    20.2   140
 2    14.7    34
 3    21      27
 4    17.8    23
 5    19.2    19
 6    17.4    18
 7    18.6    17
 8    19.1    17
 9    16.6    16
10    18.4    16
# ... with 36 more rows
```

Most observations have same values.

e) 35 suburbs are bound to the Charles river.I used the below line of code to determine the same

table(Boston$chas)

Output :

0 1

471 35

f) 19.05 is the median of Pupil-teacher ratio of various suburbs.For determining median at town level- we would require more information on the factors for grouping suburbs to towns.

median(Boston$ptratio)

g) There are 2 suburbs with the lowest median: 399 and 406

Code :

Boston[Boston$medv == min(Boston$medv),]

Output:

```
        crim zn indus chas   nox    rm age    dis rad tax ptratio  black lstat medv
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90 30.59    5
406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98    5
```

I will compute the percentile rank for each of the observations.Dataframe –HDP will be used to calculate the percentile rank of all the predictors.

Code :

```
HDP <- sapply(Boston[ ,-4], function(x) rank(x)/length(x)) %>%

  as.data.frame()

HDP[c(399, 406),]
```

Output :

```
> HDP[c(399, 406),]
         crim        zn     indus       nox        rm      age        dis
399 0.9881423 0.3685771 0.7579051 0.8448617 0.0770751 0.958498 0.05731225
406 0.9960474 0.3685771 0.7579051 0.8448617 0.1363636 0.958498 0.04150198
          rad       tax   ptratio     black     lstat        medv
399 0.8705534 0.8606719 0.7519763 0.8814229 0.9782609 0.002964427
406 0.8705534 0.8606719 0.7519763 0.3498024 0.8992095 0.002964427
```

Observations about the 2 suburbs:

- Both 399 and 406 take almost similar values for most of the predictors except for rm ,dis, black and lstat. There is a wide difference in values for rm and black redictors.
- >=90th percentile –crim,age,lstat
  >50th percentile – indus,nox,rad,tax,pratio
- Medv and zn have very low percentiles
- As chas=0,both suburbs are not near the Chas river

h)

```
sum(Boston$rm > 7)

 [1] 64
```

```
Boston_gt_8rooms <- Boston[Boston$rm > 8, ]

nrow(Boston_gt_8rooms)
```

Output:

```
[1] 13

prop.table(table(Boston_gt_8rooms$chas))
```

Output:

```
##

##         0         1

## 0.8461538 0.1538462

Boston_gt_8rooms_perc <- Boston_percentiles[as.numeric(rownames(Boston_gt_8rooms)), ]

glimpse(Boston_gt_8rooms_perc)
```

Output:

```
## Rows: 13

## Columns: 13

## $ crim    <dbl> 0.34387352, 0.69367589, 0.03557312, 0.52766798, 0.58893281,...

## $ zn      <dbl> 0.3685771, 0.3685771, 0.9950593, 0.3685771, 0.3685771, 0.36...

## $ indus   <dbl> 0.09683794, 0.91798419, 0.08992095, 0.34090909, 0.34090909,...

## $ nox     <dbl> 0.21541502, 0.68873518, 0.08893281, 0.38833992, 0.38833992,...

## $ rm      <dbl> 0.9802372, 0.9920949, 0.9762846, 0.9861660, 0.9980237, 0.97...

## $ age     <dbl> 0.48418972, 0.74604743, 0.14130435, 0.50988142, 0.55830040,...

## $ dis     <dbl> 0.5454545, 0.2766798, 0.7480237, 0.4634387, 0.4634387, 0.50...

## $ rad     <dbl> 0.06422925, 0.49407115, 0.27173913, 0.71640316, 0.71640316,...

## $ tax     <dbl> 0.21541502, 0.63932806, 0.07806324, 0.41600791, 0.41600791,...
```

$ ptratio <dbl> 0.37154150, 0.06818182, 0.06818182, 0.26778656, 0.26778656,...

$ black <dbl> 0.8814229, 0.4100791, 0.4624506, 0.3537549, 0.3201581, 0.39...

$ lstat <dbl> 0.075098814, 0.035573123, 0.011857708, 0.071146245, 0.09881...

$ medv <dbl> 0.9367589, 0.9851779, 0.9851779, 0.9565217, 0.9851779, 0.93...

sapply(Boston_gt_8rooms_perc, mean)

Output:

crim zn indus nox rm age dis

0.53815749 0.54651870 0.33612040 0.47514442 0.98814229 0.48008513 0.47202797

rad tax ptratio black lstat medv

0.57236242 0.37473396 0.24407115 0.43987534 0.09007297 0.93227425

Q8. This question involves the use of simple linear regression on the "Auto" data set.

a) Use the lm() function to perform a simple linear regression with "mpg" as the response and "horsepower" as the predictor. Use the summary() function to print the results. Comment on the output. For example:

i) Is there a relationship between the predictor and the response ?

Code:

```
library(ISLR)
data(Auto)
fit <- lm(mpg ~ horsepower, data = Auto)
summary(fit)
```

Output:

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
```

```
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,  Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

We can answer this question by testing the hypothesis $H_0: \beta_i = 0 \ \forall i$ H0:βi=0 ∀i. *The p-value corresponding to the F-statistic is 7.03198910^{-81}, this indicates a clear evidence of a relationship between “mpg” and “horsepower”.*

ii) How strong is the relationship between the predictor and the response ?

To calculate the residual error relative to the response we use the mean of the response and the RSE. The mean of mpg is 23.4459184. The RSE of the lm.fit was 4.9057569 which indicates a percentage error of 20.9237141%. We may also note that as the R^2 R2 *is equal to 0.6059483, almost 60.5948258% of the variability in “mpg” can be explained using “horsepower”.*

iii) Is the relationship between the predictor and the response positive or negative ?

As the coeficient of “horsepower” is negative, the relationship is also negative. The more horsepower an automobile has the linear regression indicates the less mpg fuel efficiency the automobile will have.

iv) What is the predicted mpgmpg associated with a “horsepower” of 98 ? What are the associated 95% confidence and prediction intervals ?

Code:

```
predict(fit, data.frame(horsepower = 98), interval = "confidence")
```

Output:

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

Code:

```
predict(fit, data.frame(horsepower = 98), interval = "prediction")
```

Output:

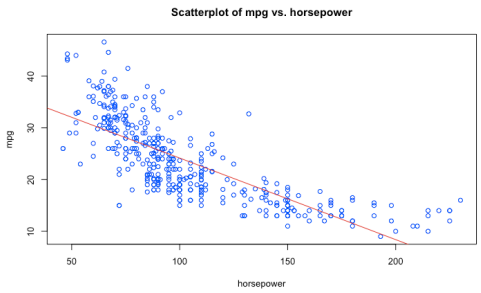
```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

b)Plot the response and the predictor. Use the abline() function to display the least squares regression line.

Code:

```
plot(Auto$horsepower, Auto$mpg, main = "Scatterplot of mpg vs. horsepower", xlab =
"horsepower", ylab = "mpg", col = "blue")
abline(fit, col = "red")
```

Output:

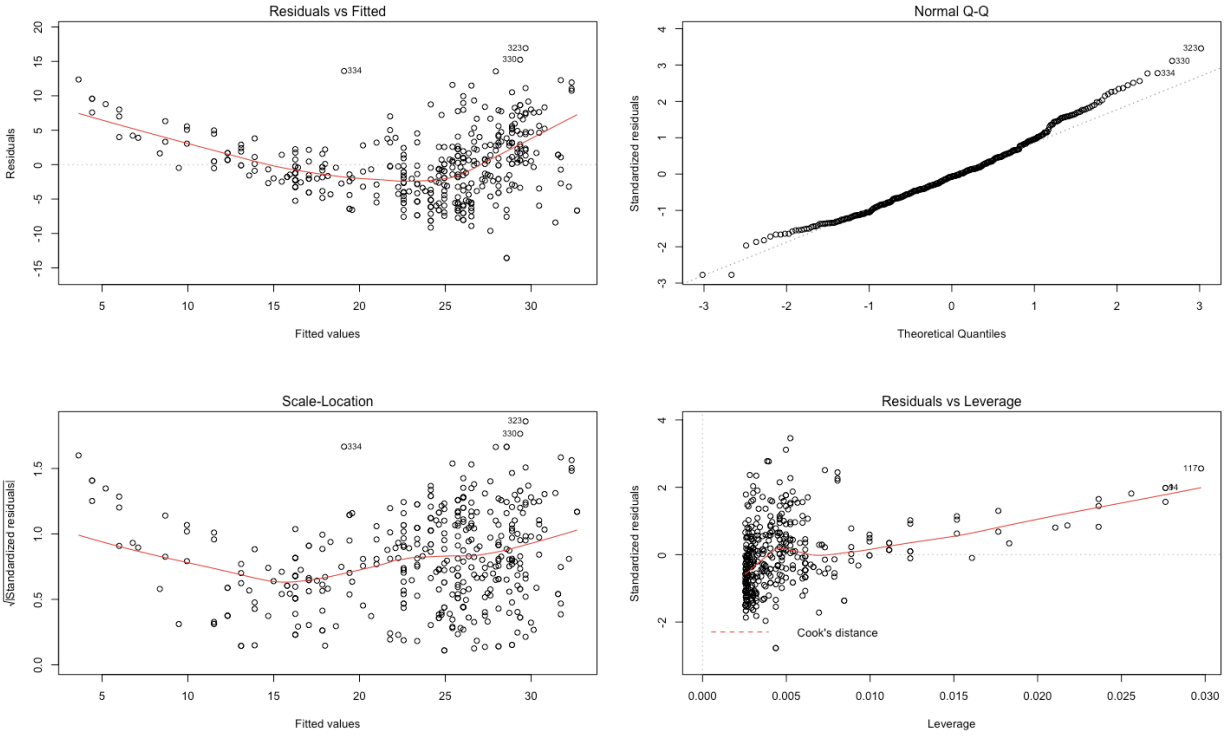


c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

Code:

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

Output:



The plot of residuals versus fitted values indicates the presence of non linearity in the data. The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and a few high leverage points.

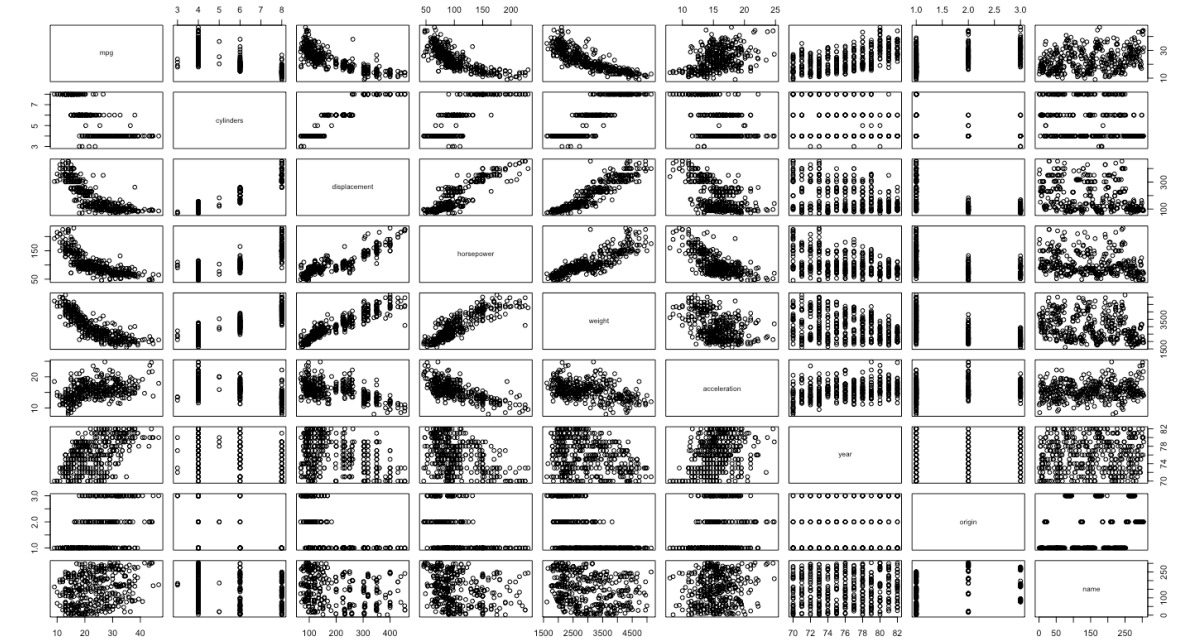
Q9. This question involves the use of multiple linear regression on the "Auto" data set.

a) Produce a scatterplot matrix which include all the variables in the data set.

Code:

```
pairs(Auto)
```

Output:



b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the "name" variable, which is qualitative.

Code:

```
names(auto)
```

output:

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"
## [5] "weight"       "acceleration" "year"         "origin"
## [9] "name"
```

Code:

```
cor(Auto[1:8])
```

Output:

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
```

```
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

c) Use the lm() function to perform a multiple linear regression with “mpg” as the response and all other variables except “name” as the predictors. Use the summary() function to print the results. Comment on the output. For instance:

i) Is there a relationship between the predictors and the response?

Code:

```
fit2 <- lm(mpg ~ . - name, data = Auto)
summary(fit2)
```

Output:

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

We can answer this question by again testing the hypothesis $H_0: \beta_i = 0 \ \forall i$ $H0: \beta i = 0 \ \forall i$. *The p-value corresponding to the F-statistic is 2.037105910^{-139}, this indicates a clear evidence of a relationship between “mpg” and the other predictors.*

ii) Which predictors appear to have a statistically significant relationship to the response ?

We can answer this question by checking the p-values associated with each predictor's t-statistic. We may conclude that all predictors are statistically significant except "cylinders", "horsepower" and "acceleration".

iii) What does the coefficient for the "year" variable suggest ?

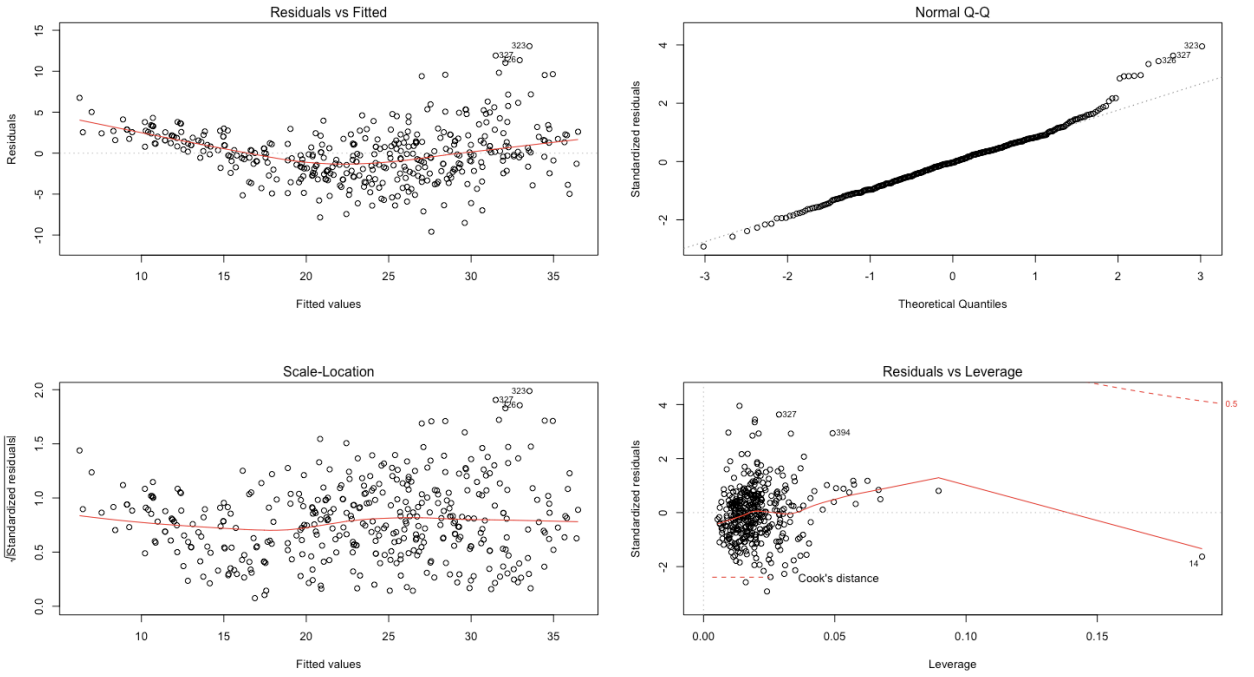
The coefficient ot the "year" variable suggests that the average effect of an increase of 1 year is an increase of 0.7507727 in "mpg" (all other predictors remaining constant). In other words, cars become more fuel efficient every year by almost 1 mpg / year.

d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers ? Does the leverage plots identify any observations with unusually high leverages ?

Code:

```
par(mfrow = c(2, 2))
plot(fit2)
```

Output:



As before, the plot of residuals versus fitted values indicates the presence of mild non linearity in the data. The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and one high leverage point (point 14).

e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant ?

From the correlation matrix, we obtained the two highest correlated pairs and used them in picking interaction effects.

Code:

```
fit3 <- lm(mpg ~ cylinders * displacement+displacement * weight, data = Auto[, 1:8])
summary(fit3)
```

Output:

```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight, data = Auto[, 1:8])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2934  -2.5184  -0.3476   1.8399  17.7723
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.262e+01  2.237e+00  23.519  < 2e-16 ***
## cylinders                7.606e-01  7.669e-01   0.992    0.322
## displacement            -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
## weight                  -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
## cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
## displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.103 on 386 degrees of freedom
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
## F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

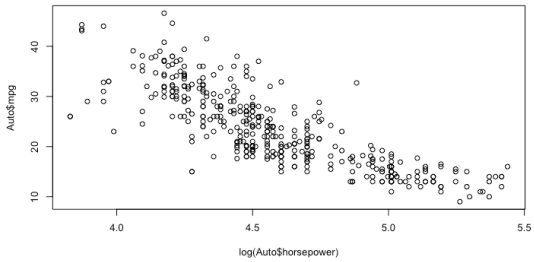
From the p-values, we can see that the interaction between displacement and weight is statistically signifcant, while the interactiion between cylinders and displacement is not.

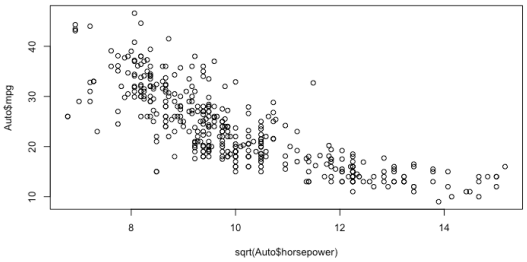
f) Try a few different transformations of the variables, such as $\log X$log⁡X, $\sqrt{X}$X, X^2X2. Comment on your findings.

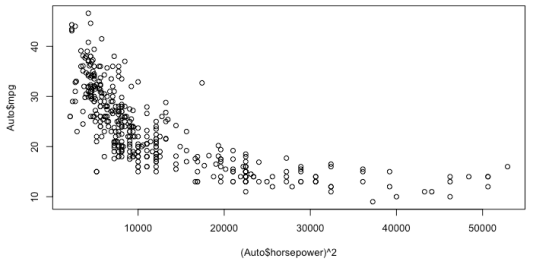
Code:

```
par(mfrow = c(2, 2))
plot(log(Auto$horsepower), Auto$mpg)
plot(sqrt(Auto$horsepower), Auto$mpg)
plot((Auto$horsepower)^2, Auto$mpg)
```

Output:







We limit ourselves to examining "horsepower" as sole predictor. It seems that the log transformation gives the most linear looking plot.

Q10. This question should be answered using the "Carseats" data set.

a) Fit a multiple regression model to predict "Sales" using "Price", "Urban" and "US".

Code:

```
data(Carseats)
fit3 <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(fit3)
```

Output:

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b) Provide an interpretation of each coefficient in the model. Be careful - some of the variables in the model are qualitative !

The coefficient of the "Price" variable may be interpreted by saying that the average effect of a price increase of 1 dollar is a decrease of 54.4588492 units in sales all other predictors remaining fixed. The coefficient of the "Urban" variable may be interpreted by saying that on average the unit sales in urban location are 21.9161508 units less than in rural location all other predictors remaining fixed. The coefficient of the "US" variable may be interpreted by saying that on average the unit sales in a US store are 1200.5726978 units more than in a non US store all other predictors remaining fixed.

c) Write out the model in equation form, being careful to handle the qualitative variables properly.

The model may be written as

$\text{Sales}=13.0434689+(-0.0544588)\times\text{Price}+(-0.0219162)\times\text{Urban}+(1.2005727)\times\text{US}+\varepsilon\text{Sales}=13.0434689+(-0.0544588)\times\text{Price}+(-0.0219162)\times\text{Urban}+(1.2005727)\times\text{US}+\varepsilon$

with $\text{Urban}=1\text{Urban}=1$ *if the store is in an urban location and* 00 *if not, and* $\text{US}=1\text{US}=1$ *if the store is in the US and* 00 *if not.*

d)For which of the predictors can you reject the null hypothesis $H0:\beta j=0H0:\beta j=0$?

We can reject the null hypothesis for the "Price" and "US" variables.

e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

Code:

```
fit4 <- lm(Sales ~ Price + US, data = Carseats)
summary(fit4)
```

Output:

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

f) How well do the models in (a) and (e) fit the data ?

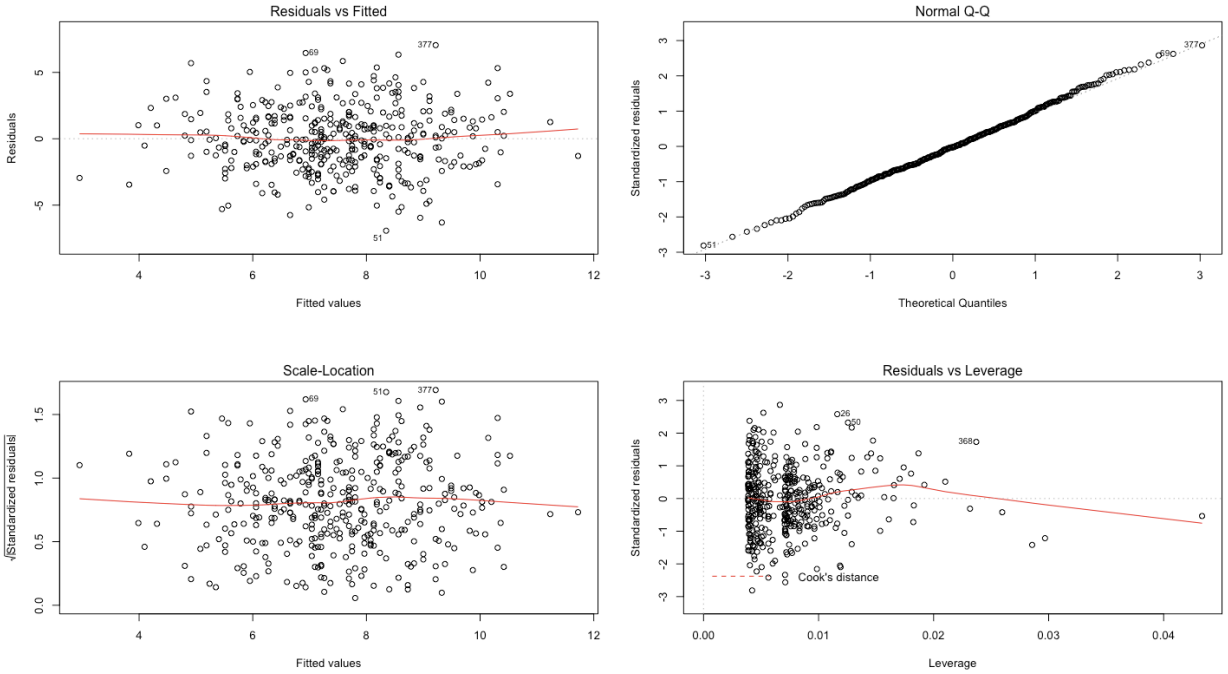
The $R2R2$ *for the smaller model is marginally better than for the bigger model. Essentially about 23.9262888% of the variability is explained by the model.*

g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

Code:
confint(fit4)
Output:

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

h) Is there evidence of outliers or high leverage observations in the model from (e) ?
Code:
par(mfrow = c(2, 2))
plot(fit4)
Output:



The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and some leverage points as some points exceed (p+1)/n(p+1)/n *(0.01).*