Assignment 1

Mrunal Ghorpade
UIN: 677441117

1) Compare the classification performance of linear regression and k–nearest neighbor classification on the zipcode data. In particular, consider only the 2's and 3's, and k = 1, 3, 5, 7 and 15. Show both the training and test error for each choice. The zipcode data are available from the book website www-stat.stanford.edu/ElemStatLearn.

Linear Regression:

For linear regression R2 is 90.05% which means model can explain 90.05% variation in the training data. The RMSE error for training data is 0.1575174 and for test data is 0.3894424.

As linear regression gives output as continuous variable, for converting the predicted output in the form of classifier i.e. 2's or 3's, threshold value of 2.5 is selected and so if the predicted value is more than 2.5 then it is converted to 3 else 2.

Following is the confusion matrix;

Training data	Test data

`> table(X1pred,zipTrain$X1)`
 	
`X1pred  2  3`
`2 728  5`
`3  3 653`
	`> table(X1pred_test,zipTest$X1)`
	
	`X1pred_test  2  3`
	`2 191  8`
	`3  7 158`

Error Rate= (3+5)/(1389) =0.005759 Accuracy: 1- Error= 99.4241%	Error Rate = (7+8)/(364)= 0.0412 Accuracy = 95.879%

K nearest neighbor:

Following are the Error rates for different values of K;

	Training	Test

For K=1	*R output:* Error Rate=0
	`> table(knn_1_train,zipTrain$X1)`
	`knn_1_train  2  3`
	`2 731  0`
	`3  0 658`
	Calculation:
	Error Rate= 0 Accuracy= 100%
		R output: Error Rate= 0.02472527
		`> table(knn_1,zipTest$X1)`
		`knn_1  2  3`
		`2 192  3`
		`3  6 163`
		Calculation:
		Error Rate =(6+3)/(364)= 0.02472 Accuracy=0.9752

	Training	Test

For K=3	*R output:* Error Rate= 0.005039597
	`> table(knn_3_train,zipTrain$X1)`
	`knn_3_train  2  3`
	`2 728  4`
	`3  3 654`
	Calculation:
	Error Rate =(4+3)/(1389)= 0.005039597 Accuracy=0.9949
		R output: Error Rate= 0.03021978
		`> table(knn_3,zipTest$X1)`
		`knn_3  2  3`
		`2 191  4`
		`3  7 162`
		Calculation:
		Error Rate =(7+4)/(364)= 0.03021978 Accuracy=0.9697

For K=5	*R output:* Error Rate= 0.005759539
	`> table(knn_5_train,zipTrain$X1)`
	`knn_5_train  2  3`
	`2 726  3`
	`3  5 655`
	Calculation:
	Error Rate = (5+3)/(1389) = 0.005759539 Accuracy=0.9942
		R output: Error Rate= 0.03021978
		`> table(knn_5,zipTest$X1)`
		`knn_5  2  3`
		`2 191  4`
		`3  7 162`
		Calculation:
		Error Rate =(7+4)/(364)= 0.03021978 Accuracy=0.9697

For K=7	*R output:* Error Rate= 0.006479482
	`> table(knn_7_train,zipTrain$X1)`
	`knn_7_train  2  3`
	`2 725  3`
	`3  6 655`
	Calculation:
	Error Rate =(6+3)/(1389)= 0.006479482 Accuracy=0.9935
		R output: Error Rate= 0.03296703
		`> table(knn_7,zipTest$X1) #`
		`knn_7  2  3`
		`2 189  3`
		`3  9 163`
		Calculation:
		Error Rate =(9+3)/(364)= 0.03296703 Accuracy=0.96704

For K=15	*R output:* Error Rate= 0.009359251
	`> table(knn_15_train,zipTrain$X1)`
	`knn_15_train  2  3`
	`2 721  3`
	`3  10 655`
	Calculation:
	Error Rate =(10+3)/(1389)= 0.009447 Accuracy=0.99055
		R output: Error Rate= 0.03846154
		`> table(knn_15,zipTest$X1)`
		`knn_15  2  3`
		`2 187  3`
		`3  11 163`
		Calculation:
		Error Rate =(11+3)/(364)= 0.03846154 Accuracy=0.96153

By comparing linear Regression model and K-n-n with different values of K it can be seen that for k=1 classification performance has the accurate predicted output.

2) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

In less flexible approach we assume a model which gives a functional form of original function (f). Now we just estimate a set of parameters rather than estimating f. Disadvantage of this is that there is always a probability that the functional form used to estimate f is very different than true f, in which case the resulting estimate will be different from the data. This issue can be addressed by using more flexible approach where no assumptions regarding the functional form of f is made. Flexible approach considers all predictors from the data due to which they have more probability to accurately fit the original function. In general, fitting a more flexible model requires estimating a greater number of parameters. But as the model is too flexible, it can overfit the train data leading to large test error.
Less flexible approach can be used when we are interested in interference. When we want to understand how an individual predictor is associated with the response we use less flexible approach. However, when we are only interested in prediction and not the interpretability we use more flexible approach.

3) The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X1	X2	X3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

Ans: Euclidean distance between each observation in training data and the test point is as follows:

Y	Calculations	Distance
Red	sqrt((0-0)^2 + (0-3)^2 + (0-0)^2) = sqrt(9)	3
Red	sqrt((0-2)^2 + (0-0)^2 + (0-0)^2) = sqrt(4)	2
Red	sqrt((0-0)^2 + (0-1)^2 + (0-3)^2) = sqrt(10)	3.162278
Green	sqrt((0-0)^2 + (0-1)^2 + (0-2)^2) = sqrt(5)	2.236068
Green	sqrt((0+1)^2 + (0-0)^2 + (0-1)^2) = sqrt(2)	1.414214
Red	sqrt((0-1)^2 + (0-1)^2 + (0-1)^2) = sqrt(3)	1.732051

(Note: Also computed in R as shown below and validated the R code results by above calculations)

R output:

```
> print(euc_dis)
       R        R        R        G        G        R
3.000000 2.000000 3.162278 2.236068 1.414214 1.732051
```

(b) What is our prediction with $K = 1$? Why?

R output:

```
> knn_1<- knn(train = training[,1:3],test = testset, k=1, cl=as.factor(Y))
> print(knn_1)
[1] G
Levels: G R
```

The prediction is Green with K=1. For K=1, we pick 1 nearest neighbor for the data set from the training data. The closest one is Green i.e. X1 = -1, X2 = 0, X3 = 1 from the training data set at 1.414214 and therefore the prediction was Green.

(c) What is our prediction with $K = 3$? Why?

R output:

```
> # Prediction with knn=3
>
> knn_3<- knn(train = training[,1:3],test = testset, k=3, cl=as.factor(Y))
> print(knn_3)
[1] R
Levels: G R
>
```

The prediction is Red with K=3. For K=1 we pick 3 nearest neighbors for the data set from the training data predict on the bases of whichever color occurs most number of time. The 3 closest neighbors are as follows:

Y	X1	X2	X3	Distance
Green	-1	0	1	1.414214
Red	1	1	1	1.732051
Red	2	0	0	2

As there are 2 Red's and 1 Green the prediction was Red.

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the *best* value for *K* to be large or small? Why?

When K is small the decision boundary of K-n-n is more flexible but as K grows, the method becomes less flexible and produces a decision boundary that is close to linear. Now if the Bayes decision boundary is highly non-linear then we would need small K values. Whereas when the boundary becomes more rigid, we need large K values.

4) This exercise involves the Boston housing data set.

(a) How many rows are in this data set? How many columns? What do the rows and columns represent?

Ans:

Number of rows: 506

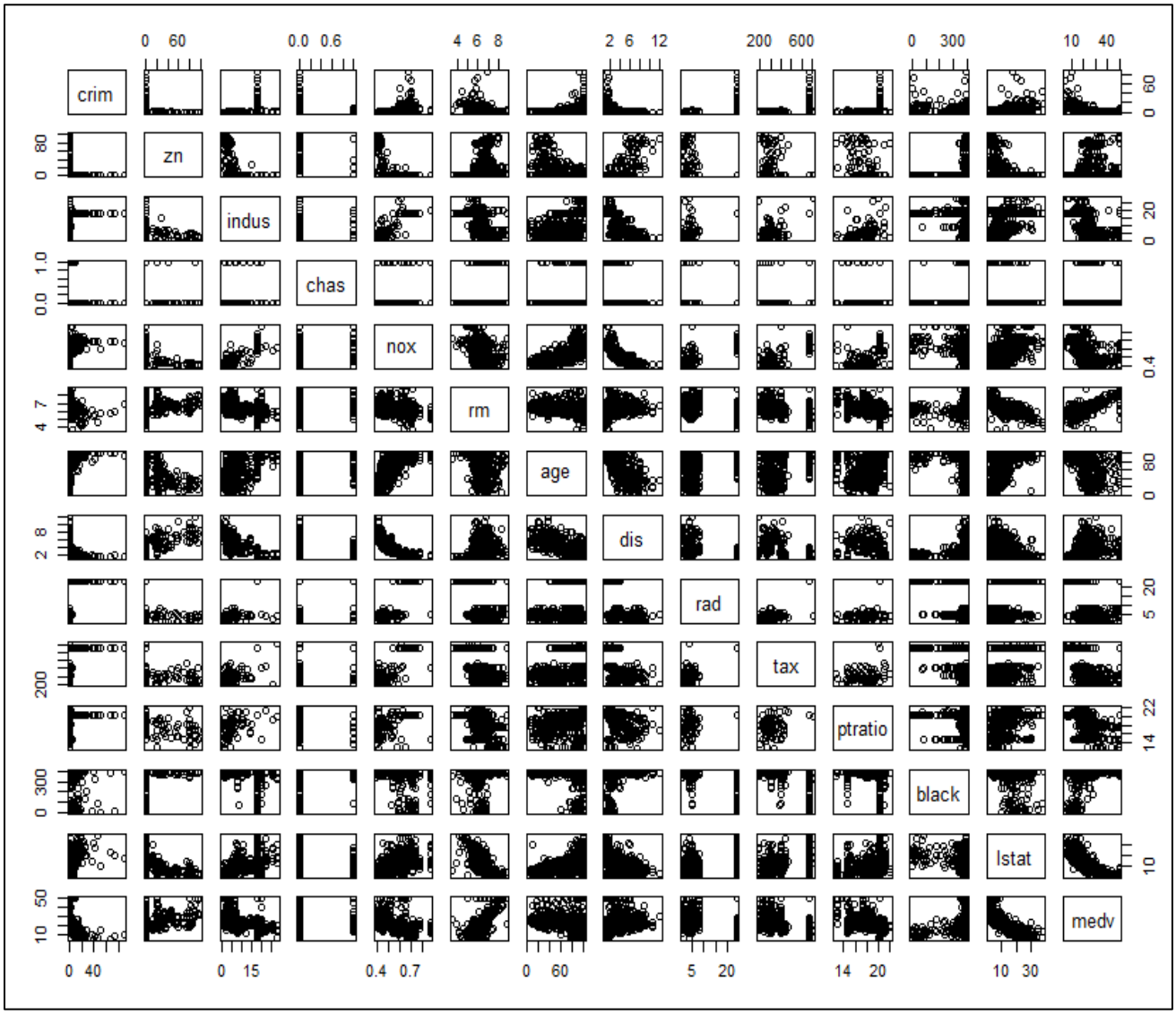
Number of columns: 14

The data represents number of variables for 505 different records of housing values in suburbs of Boston. The variables are as follows:

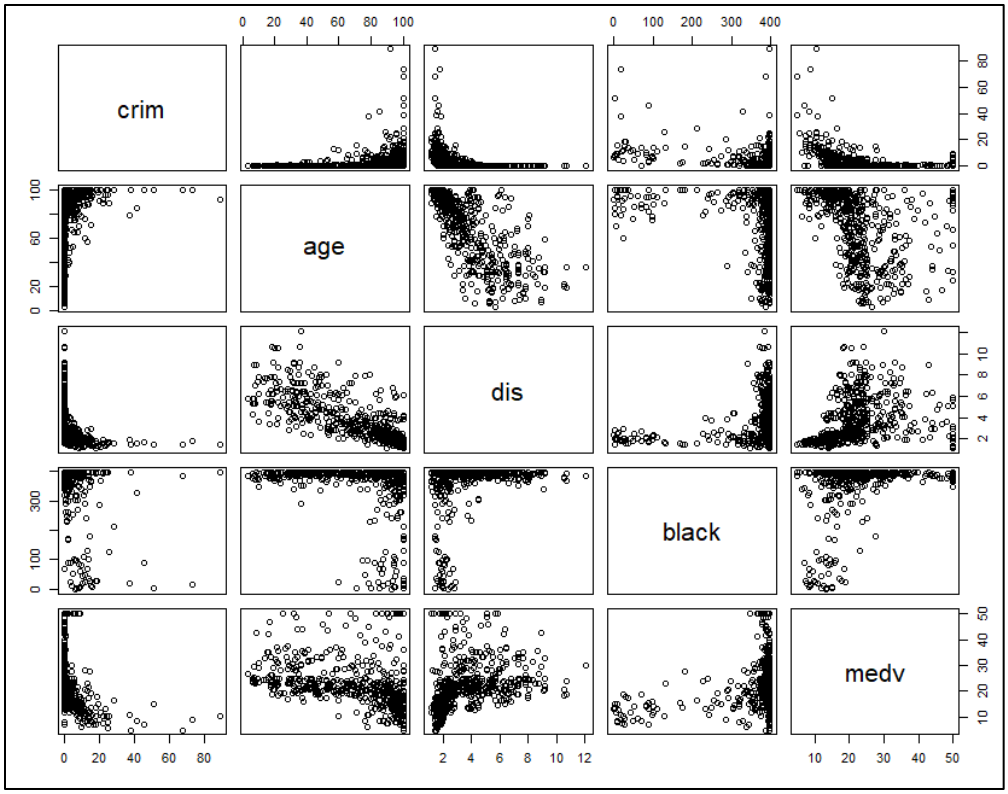
Column Names	Description
crim	per capita crime rate by town.
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town.
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
nox	nitrogen oxides concentration (parts 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 centers.
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) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings

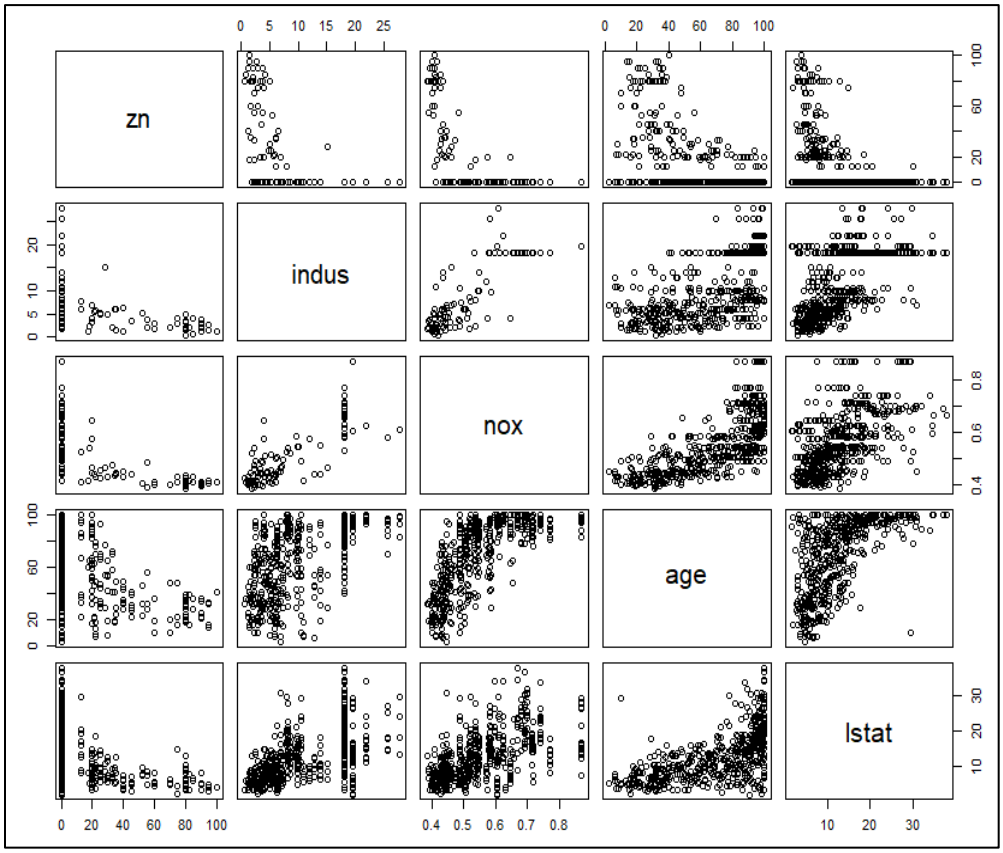
R output: Scatterplot for all the Predictor columns is as follows;



It is very difficult to understand the relation between all the columns when we plot all of them together, so visualizing them pairwise as follows;



The scatterplot shows that the crime is more near older homes. Near the employment centers the crime rate is more. As distance to the Employment centers increases, more newer houses are observed. Crime rate is less where the owner owns expensive houses.



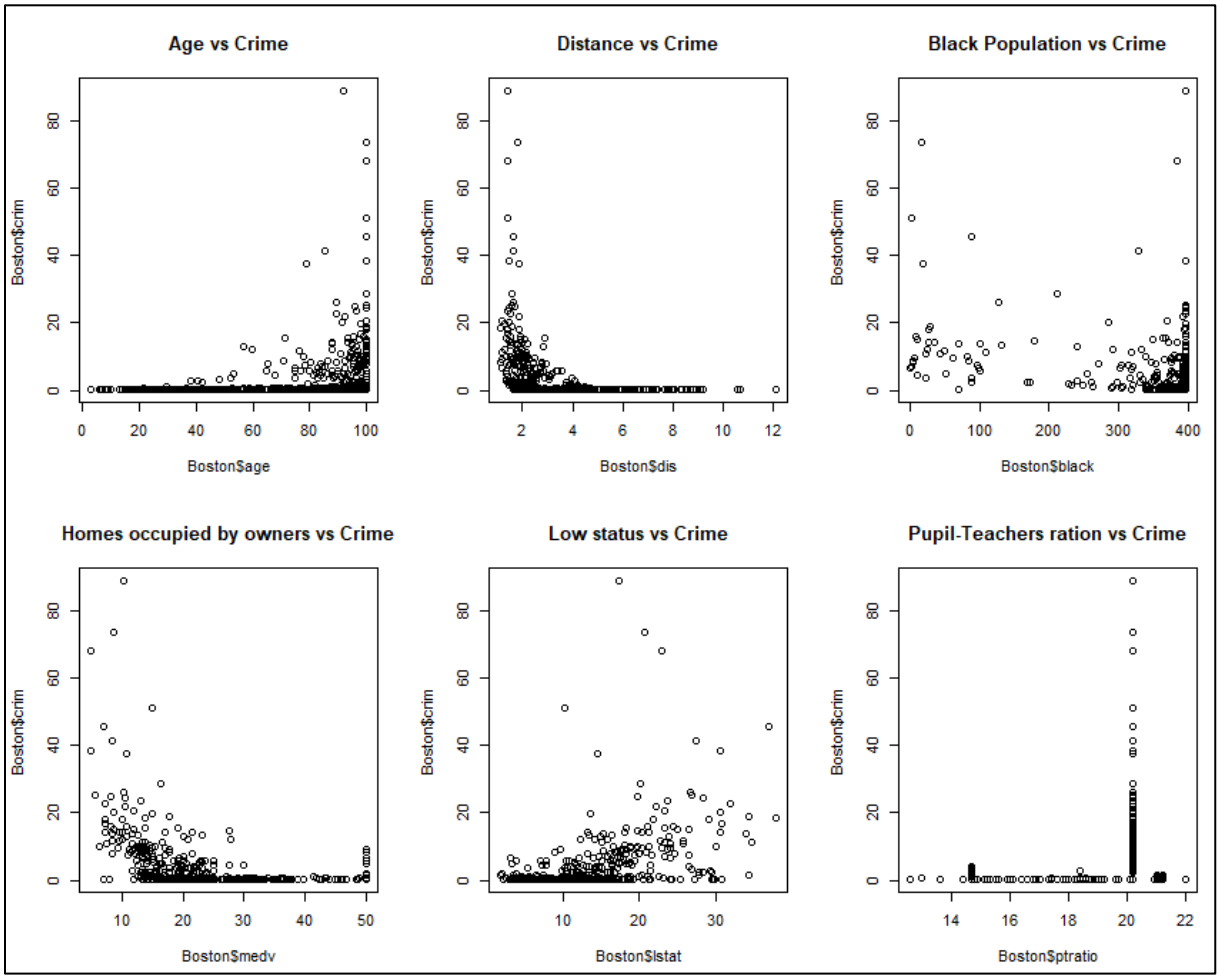
The proportion of residential land decreases as the proportional of non-residential business acres increase. Nitrogen oxide concentration is lower where proportion of residential land more. For the older houses Nitrogen oxide concentration is high. Percentage of lower status population staying in residential area is less.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

Ans: For the suburbs were the crime rate higher than 4, tax rates are high (ranging from 403 – 666), 72.72% of the black population stays in those suburbs, lower status population is more (range: 2.96-37.97) and pupil-teacher ratio is high (range: 14.7 – 20.2). Also, the plots show that crime rate is more near many older buildings. Near the employment centers the crime is more. Crimes are less where the owner owned expensive houses.

R output:

```
> Sub_Boston<-(Boston[Boston$crim>4,]) #selecting suburbs who have crime rate > 4
> dim(Sub_Boston)
[1] 121  14
> nrow(Sub_Boston[Sub_Boston$black>200,])/nrow(Sub_Boston) # percentage of black population where crime rate > 4
[1] 0.7272727
> range(Boston[Boston$crim>4,]$tax)
[1] 403 666
> range(Boston[Boston$crim>4,]$ptratio)
[1] 14.7 20.2
> range(Boston[Boston$crim>4,]$lstat)
[1]  2.96 37.97
```



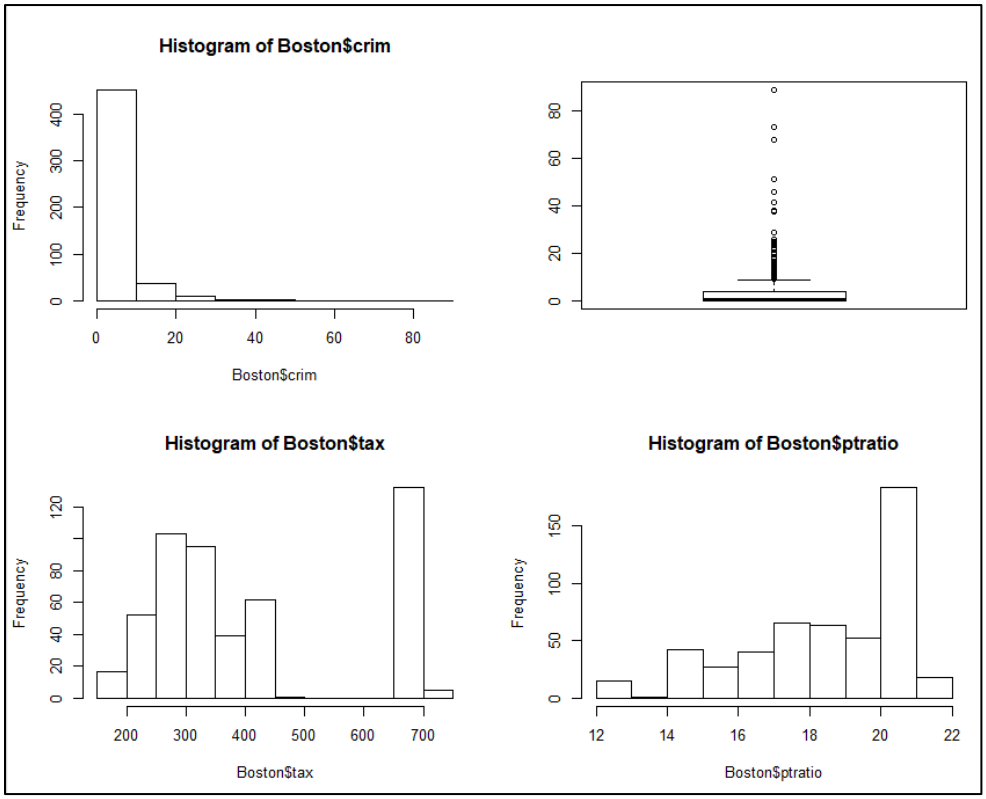
(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

Ans:

- The histogram of crime is right tailed which shows most cities have less crime rate, but 18 suburbs have crime rate more than 20 and only 18% of suburbs have crime rate more than 6
- Tax rates are lower in most of the suburbs, but have a peak near 700 (Tax Rates Range: 187-711) for some suburbs and for those suburbs crime rate is greater than 20 (tax rates are 666)
- Histogram for Pupil-teacher ratio is left skewed and most of the suburbs have higher pupil-teacher ratio. For Crime-rate greater than 20 the pupil teacher ratio is 20.2

R output:

```
> nrow(Boston[Boston$crim>20,])
[1] 18
> nrow(Boston[Boston$crim>6,])/nrow(Boston)
[1] 0.1837945
> range(Boston$tax)
[1] 187 711
> hist(Boston$tax)
> range(Boston$ptratio)
[1] 12.6 22.0
> hist(Boston$ptratio)
> nrow(Boston[Boston$crim>20,])
[1] 18
> nrow(Boston[Boston$crim>6,])/nrow(Boston)
[1] 0.1837945
> range(Boston$tax)
[1] 187 711
> hist(Boston$tax)
> range(Boston[Boston$crim>20,]$tax)
[1] 666 666
> range(Boston$ptratio)
[1] 12.6 22.0
> hist(Boston$ptratio)
> range(Boston[Boston$crim>20,]$ptratio)
[1] 20.2 20.2
```



(e) How many of the suburbs in this data set bound the Charles river?
Ans: 35 suburbs in data set bound to Charles river

R output:

```
> nrow(Boston[Boston$chas==1,]) # number of suburbs near charles river
[1] 35
```

(f) What is the median pupil-teacher ratio among the towns in this data set?
Ans: The median pupil-teacher ratio = 19.05

R output:

```
> median(Boston$ptratio)
[1] 19.05
```

(g) Which suburb of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

Ans: Suburbs that have following predictors have lowest value of owner occupied homes;

Columns	Range	Comments
crim	38.3518-67.9208	above 3rd quintile
zn	0	Least
indus	18.1	at 3rd quintile
chas	0	Not near the river
nox	0.693	above 3rd quintile
rm	5.453-5.683	below 1st quintile
age	100	Maximum
dis	1.4896-1.4254	below 1st quintile
rad	24	Maximum
tax	666	at 3rd quintile
ptratio	20.2	at 3rd quintile
black	396.9-384.97	above 1st quintile
lstat	30.59-22.98	above 3rd quintile
medv	5	Least

R output:

```
> summary(Boston[Boston$medv==min(Boston$medv),])
      crim                zn             indus            chas             nox               rm             age              dis        
 Min.   :38.35   Min.   :0   Min.   :18.1   Min.   :0   Min.   :0.693   Min.   :5.453   Min.   :100   Min.   :1.425  
 1st Qu.:45.74   1st Qu.:0   1st Qu.:18.1   1st Qu.:0   1st Qu.:0.693   1st Qu.:5.511   1st Qu.:100   1st Qu.:1.441  
 Median :53.14   Median :0   Median :18.1   Median :0   Median :0.693   Median :5.568   Median :100   Median :1.458  
 Mean   :53.14   Mean   :0   Mean   :18.1   Mean   :0   Mean   :0.693   Mean   :5.568   Mean   :100   Mean   :1.458  
 3rd Qu.:60.53   3rd Qu.:0   3rd Qu.:18.1   3rd Qu.:0   3rd Qu.:0.693   3rd Qu.:5.625   3rd Qu.:100   3rd Qu.:1.474  
 Max.   :67.92   Max.   :0   Max.   :18.1   Max.   :0   Max.   :0.693   Max.   :5.683   Max.   :100   Max.   :1.490  
      rad              tax             ptratio          black            lstat            medv      
 Min.   :24   Min.   :666   Min.   :20.2   Min.   :385.0   Min.   :22.98   Min.   :5  
 1st Qu.:24   1st Qu.:666   1st Qu.:20.2   1st Qu.:388.0   1st Qu.:24.88   1st Qu.:5  
 Median :24   Median :666   Median :20.2   Median :390.9   Median :26.79   Median :5  
 Mean   :24   Mean   :666   Mean   :20.2   Mean   :390.9   Mean   :26.79   Mean   :5  
 3rd Qu.:24   3rd Qu.:666   3rd Qu.:20.2   3rd Qu.:393.9   3rd Qu.:28.69   3rd Qu.:5  
 Max.   :24   Max.   :666   Max.   :20.2   Max.   :396.9   Max.   :30.59   Max.   :5  
```

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

Ans: 64 suburbs have average more than seven rooms per dwelling. 13 suburbs have average more than 8 rooms per dwelling. The Suburbs with average more than 8 rooms per dwelling have following predictors;

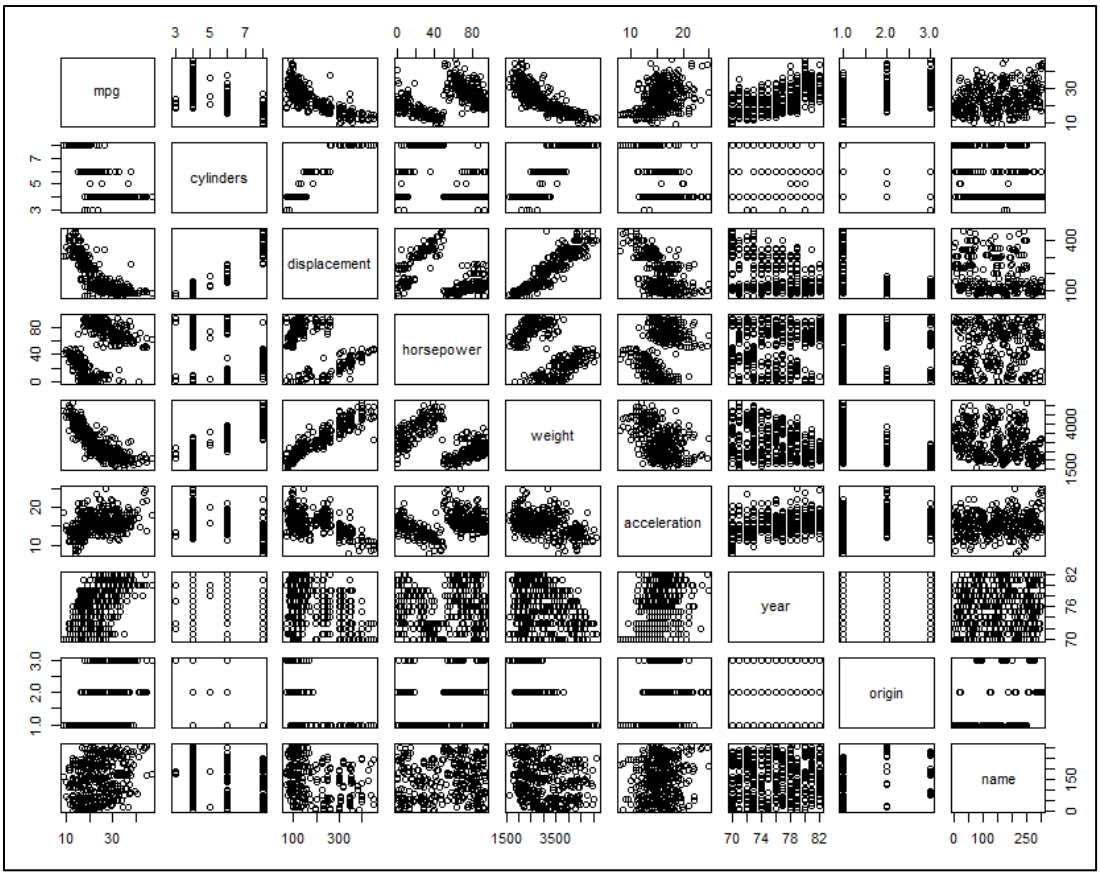
Columns	Range	Comments
crim	0.02-3.47	below 3rd quintile
zn	0-95	No effect
indus	2-19.58	below 3rd quintile
chas	0-1	not bounded by river
nox	0.41-0.71	above 1st quintile
rm	8.03-8.78	Maximum
age	8.40-93.9	Minimum
dis	1.8 - 8.9	No effect
rad	2.0-24.0	No effect
tax	224-666	above 1st quintile
ptratio	13-20.2	below 3rd quintile
black	354.6-396.9	above 1st quintile
lstat	2.47-7.44	below 1st quintile
medv	21.9-50	above 3rd quintile

R output:

```
> nrow(Boston[Boston$rm>7,])
[1] 64
> nrow(Boston[Boston$rm>8,])
[1] 13
> summary(Boston[Boston$rm>8,])
      crim                zn             indus            chas              nox              rm             age       
 Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000   Min.   :0.4161   Min.   :8.034   Min.   : 8.40  
 1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000   1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40  
 Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000   Median :0.5070   Median :8.297   Median :78.30  
 Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538   Mean   :0.5392   Mean   :8.349   Mean   :71.54  
 3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000   3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50  
 Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000   Max.   :0.7180   Max.   :8.780   Max.   :93.90  
      dis              rad              tax            ptratio          black            lstat            medv      
 Min.   :1.801   Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6   Min.   :2.47   Min.   :21.9  
 1st Qu.:2.288   1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5   1st Qu.:3.32   1st Qu.:41.7  
 Median :2.894   Median : 7.000   Median :307.0   Median :17.40   Median :386.9   Median :4.14   Median :48.3  
 Mean   :3.430   Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2   Mean   :4.31   Mean   :44.2  
 3rd Qu.:3.652   3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7   3rd Qu.:5.12   3rd Qu.:50.0  
 Max.   :8.907   Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9   Max.   :7.44   Max.   :50.0  
```

9) This question involves the use of multiple linear regression on the Auto data set.

(a) Produce a scatterplot matrix which includes all of the variables in the data set.

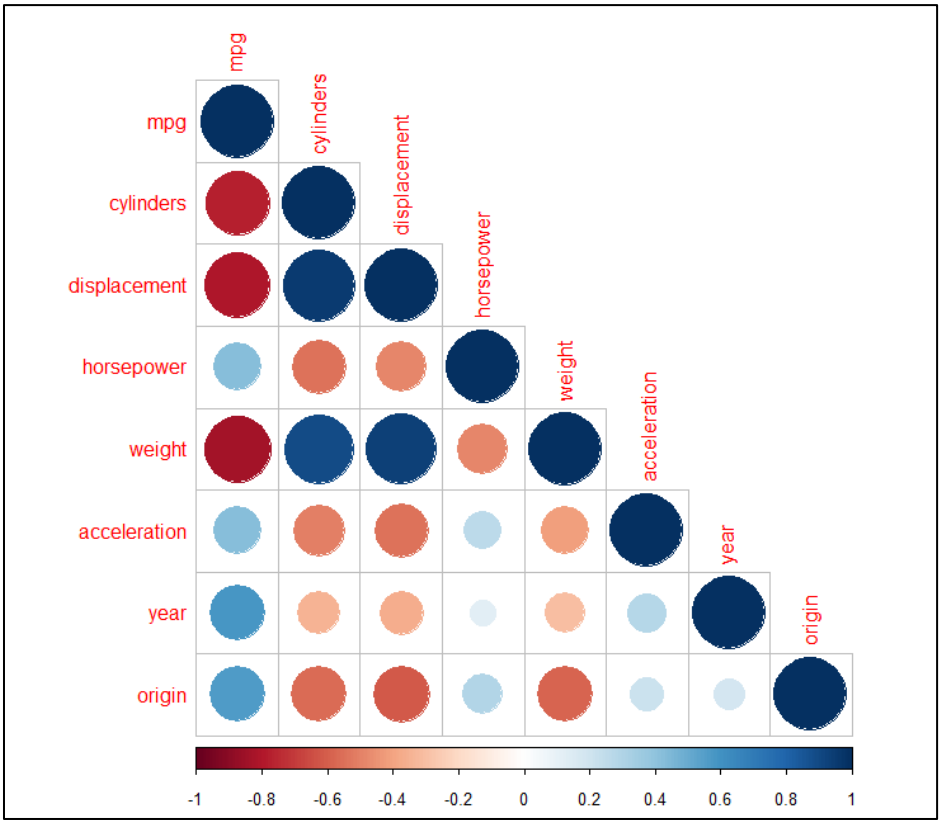


(b) Compute the matrix of correlations between the variables using the function cor (). You will need to exclude the name variable, which is qualitative.

R output:

```
> cor(Auto[,-9])
                     mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7762599   -0.8044430  0.4228227 -0.8317389    0.4222974  0.5814695  0.5636979
cylinders    -0.7762599  1.0000000    0.9509199 -0.5466585  0.8970169   -0.5040606 -0.3467172 -0.5649716
displacement -0.8044430  0.9509199    1.0000000 -0.4820705  0.9331044   -0.5441618 -0.3698041 -0.6106643
horsepower    0.4228227 -0.5466585   -0.4820705  1.0000000 -0.4821507    0.2662877  0.1274167  0.2973734
weight       -0.8317389  0.8970169    0.9331044 -0.4821507  1.0000000   -0.4195023 -0.3079004 -0.5812652
acceleration  0.4222974 -0.5040606   -0.5441618  0.2662877 -0.4195023    1.0000000  0.2829009  0.2100836
year          0.5814695 -0.3467172   -0.3698041  0.1274167 -0.3079004    0.2829009  1.0000000  0.1843141
origin        0.5636979 -0.5649716   -0.6106643  0.2973734 -0.5812652    0.2100836  0.1843141  1.0000000
```

Plot Correlation:



(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.

R output:

```
Call:
lm(formula = mpg ~ ., data = Auto[, -9])

Residuals:
    Min      1Q  Median      3Q     Max
-9.629 -2.034 -0.046   1.801 13.010

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.128e+01  4.259e+00  -4.998 8.78e-07 ***
cylinders    -2.927e-01  3.382e-01  -0.865   0.3874
displacement  1.603e-02  7.284e-03   2.201   0.0283 *
horsepower    7.942e-03  6.809e-03   1.166   0.2442
weight       -6.870e-03  5.799e-04 -11.846  < 2e-16 ***
acceleration  1.539e-01  7.750e-02   1.986   0.0477 *
year          7.734e-01  4.939e-02  15.661  < 2e-16 ***
origin        1.346e+00  2.691e-01   5.004 8.52e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.331 on 389 degrees of freedom
Multiple R-squared:  0.822,     Adjusted R-squared:  0.8188
F-statistic: 256.7 on 7 and 389 DF,  p-value: < 2.2e-16
```

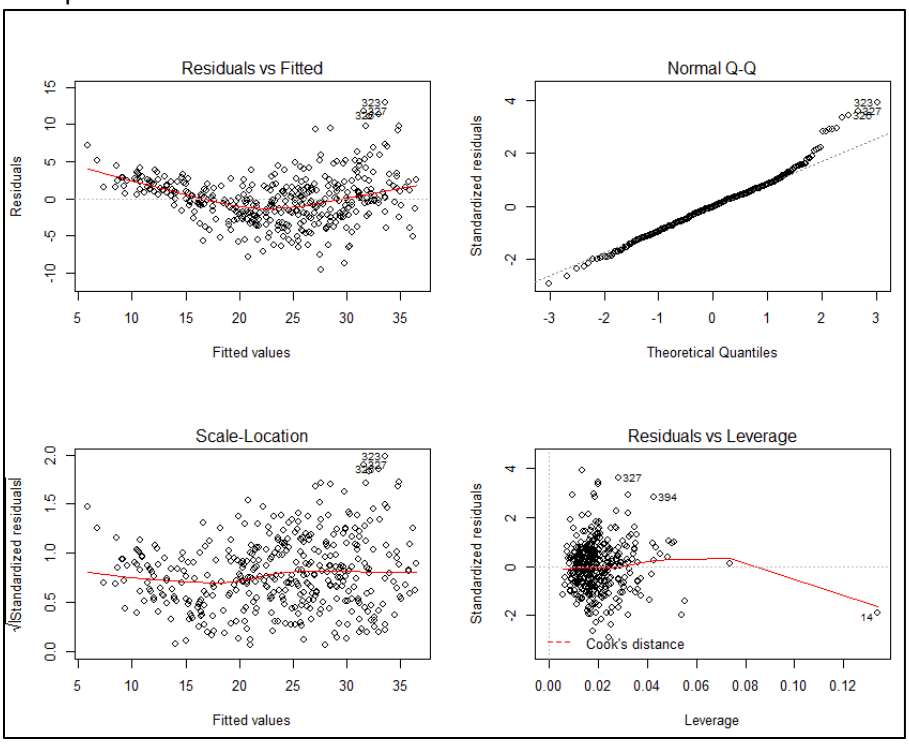
I. Is there a relationship between the predictors and the response?
Ans:
Yes, as R2= 82.2% i.e. Based on independent variables the linear regression model can predict 82.2% variability in dependent variable.

II. Which predictors appear to have a statistically significant relationship to the response?
Ans:
If alpha=5% then following predictors appear to be statistically significant:

Predictors	P-value
displacement	0.0283
weight	< 2e-16
acceleration	0.0477
year	< 2E-16
origin	8.52E-07

III. What does the coefficient for the year variable suggest?
Ans:
For every 1 Unit increase in year there is 0.7734 Unit increase in mpg.

(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 plot identify any observations with unusually high leverage?
R output:



Observations:

1) Residual vs Fitted: The plot shows that the residuals have non-linear patterns. There is a non-linear relationship between predictor variables and the response variable as the residuals are not equally spread around the horizontal line.
2) Normal Q-Q: Plot shows that the residuals are normally distributed, although there are some exceptions as observation 323, 326 and 327 which seem off.
3) Scale-Location: This plot shows that the residuals are not spread equally along the ranges of predictors i.e. the plot shows heteroscedasticity.
4) Residuals vs Leverage: It looks like there are no influential cases, although there are outliers (observations 394, 327 and 14) but all of them are inside the cook's distance line. That means they even though they have extreme values, they would not affect the regression line and the results wouldn't be much different if we either include or exclude them from analysis.

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

Ans:
Following interactions are statistically significant:

- Weight and Year
- Horsepower and acceleration
- Cylinder and displacement

R output:

```
Call:
lm(formula = mpg ~ . + weight * year + horsepower:acceleration +
    cylinders:displacement, data = Auto[, -9])

Residuals:
     Min       1Q   Median       3Q      Max
-10.0871  -1.6402  -0.0735   1.4532  12.5109

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.005e+02  1.373e+01  -7.325 1.41e-12 ***
cylinders               -1.438e+00  4.144e-01  -3.471 0.000576 ***
displacement            -6.117e-02  1.369e-02  -4.467 1.04e-05 ***
horsepower               9.616e-02  4.297e-02   2.238 0.025804 *
weight                   2.253e-02  4.537e-03   4.966 1.03e-06 ***
acceleration             6.402e-01  1.574e-01   4.068 5.75e-05 ***
year                     1.860e+00  1.700e-01  10.939  < 2e-16 ***
origin                   5.229e-01  2.596e-01   2.014 0.044717 *
weight:year             -3.867e-04  6.020e-05  -6.424 3.91e-10 ***
horsepower:acceleration -5.934e-03  2.663e-03  -2.228 0.026465 *
cylinders:displacement   9.445e-03  1.653e-03   5.715 2.20e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.969 on 386 degrees of freedom
Multiple R-squared:  0.8597,    Adjusted R-squared:  0.8561
F-statistic: 236.6 on 10 and 386 DF,  p-value: < 2.2e-16
```

(f) Try a few different transformations of the variables, such as log(*X*), *vX*, *X*2. Comment on your findings.

Ans:

R^2 has increased from 82.2% to 86.21% which mean model is able to explain higher amount of variation in data. Also, Residual vs Fitted plot shows that the non-linearity in the data is decreased and Residuals are equally spread around the horizontal line showing homoscedasticity.

R output:

```
Call:
lm(formula = mpg ~ . + log(weight) + I(acceleration^2), data = Auto[,
    -9])

Residuals:
    Min      1Q  Median      3Q     Max
-9.948 -1.688 -0.017   1.597 12.877

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      278.964075  30.202888   9.236  < 2e-16 ***
cylinders         -0.130159   0.301499  -0.432   0.6662
displacement      -0.002148   0.006932  -0.310   0.7569
horsepower         0.005981   0.006013   0.995   0.3205
weight             0.007626   0.001553   4.911 1.34e-06 ***
acceleration      -2.029122   0.486971  -4.167 3.81e-05 ***
year               0.837958   0.044008  19.041  < 2e-16 ***
origin             0.475596   0.251253   1.893   0.0591 .
log(weight)      -41.017628   4.304579  -9.529  < 2e-16 ***
I(acceleration^2)  0.064999   0.014444   4.500 8.99e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.94 on 387 degrees of freedom
Multiple R-squared:  0.8621,    Adjusted R-squared:  0.8589
F-statistic: 268.8 on 9 and 387 DF,  p-value: < 2.2e-16
```