

## ASSIGNMENT2

### CHAPTER-2

#### Group members

Dhavului saimanonma-16321823

Sarika jakkidi – 16321394

8 a) use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

Source

Console Terminal x Jobs x

R 4.1.0 · ~/

> library(ISLR)  
> data("College")  
> College= read.csv("C:/Users/chakr/Documents/ISL/college.csv")  
> 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

b. Look at the data using the `fix()` function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

The screenshot shows the R console on the left and the Data Editor window on the right. The console displays the output of `summary(College)` and the commands to fix the row names.

```
> summary(College)
Private      Apps
No :212  Min.   : 81
Yes:565  1st Qu.: 776
        Median : 1558
        Mean   : 3002
        3rd Qu.: 3624
        Max.   :48094

Top25perc    F.Under
Min.   : 9.0  Min.   :
1st Qu.: 41.0 1st Qu.:
Median : 54.0 Median :
Mean   : 55.8 Mean   :
3rd Qu.: 69.0 3rd Qu.:
Max.   :100.0 Max.   :

Room.Board    Books
Min.   :1780  Min.   :
1st Qu.:3597 1st Qu.:
Median :4200 Median :
Mean   :4358 Mean   :
3rd Qu.:5050 3rd Qu.:
Max.   :8124 Max.   :

Terminal      S.F.Ra
Min.   : 24.0  Min.   :
1st Qu.: 71.0 1st Qu.:
Median : 82.0 Median :
Mean   : 79.7 Mean   :
3rd Qu.: 92.0 3rd Qu.:
Max.   :100.0 Max.   :

Grad.Rate
Min.   : 10.00
1st Qu.: 53.00
Median : 65.00
Mean   : 65.46
3rd Qu.: 78.00
Max.   :118.00

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

The Data Editor window shows a table with 19 rows and 4 columns. The first column, labeled 'row.names', contains the names of 19 colleges. The other three columns are 'Private', 'Apps', and 'Accept'.

row.names	Private	Apps	Accept
1 Abilene Christian University	Yes	1660	1232
2 Adelphi University	Yes	2186	1924
3 Adrian College	Yes	1428	1097
4 Agnes Scott College	Yes	417	349
5 Alaska Pacific University	Yes	193	146
6 Albertson College	Yes	587	479
7 Albertus Magnus College	Yes	353	340
8 Albion College	Yes	1899	1720
9 Albright College	Yes	1038	839
10 Alderson-Broadus College	Yes	582	498
11 Alfred University	Yes	1732	1425
12 Allegheny College	Yes	2652	1900
13 Allentown Coll. of St. Francis de Sales	Yes	1179	780
14 Alma College	Yes	1267	1080
15 Alverno College	Yes	494	313
16 American International College	Yes	1420	1093
17 Amherst College	Yes	4302	992
18 Anderson University	Yes	1216	908
19 Andrews University	Yes	1130	704

c.

i) Use the summary() function to produce a numerical summary of the variables in the data set.

---

```
> summary(College)
```

Private	Apps	Accept
Length:777	Min. : 81	Min. : 72
Class :character	1st Qu.: 776	1st Qu.: 604
Mode :character	Median : 1558	Median : 1110
	Mean : 3002	Mean : 2019
	3rd Qu.: 3624	3rd Qu.: 2424
	Max. :48094	Max. :26330

Enroll	Top10perc	Top25perc
Min. : 35	Min. : 1.00	Min. : 9.0
1st Qu.: 242	1st Qu.:15.00	1st Qu.: 41.0
Median : 434	Median :23.00	Median : 54.0
Mean : 780	Mean :27.56	Mean : 55.8
3rd Qu.: 902	3rd Qu.:35.00	3rd Qu.: 69.0
Max. :6392	Max. :96.00	Max. :100.0

F.Undergrad	P.Undergrad	Outstate
Min. : 139	Min. : 1.0	Min. : 2340
1st Qu.: 992	1st Qu.: 95.0	1st Qu.: 7320
Median : 1707	Median : 353.0	Median : 9990
Mean : 3700	Mean : 855.3	Mean :10441
3rd Qu.: 4005	3rd Qu.: 967.0	3rd Qu.:12925
Max. :31643	Max. :21836.0	Max. :21700

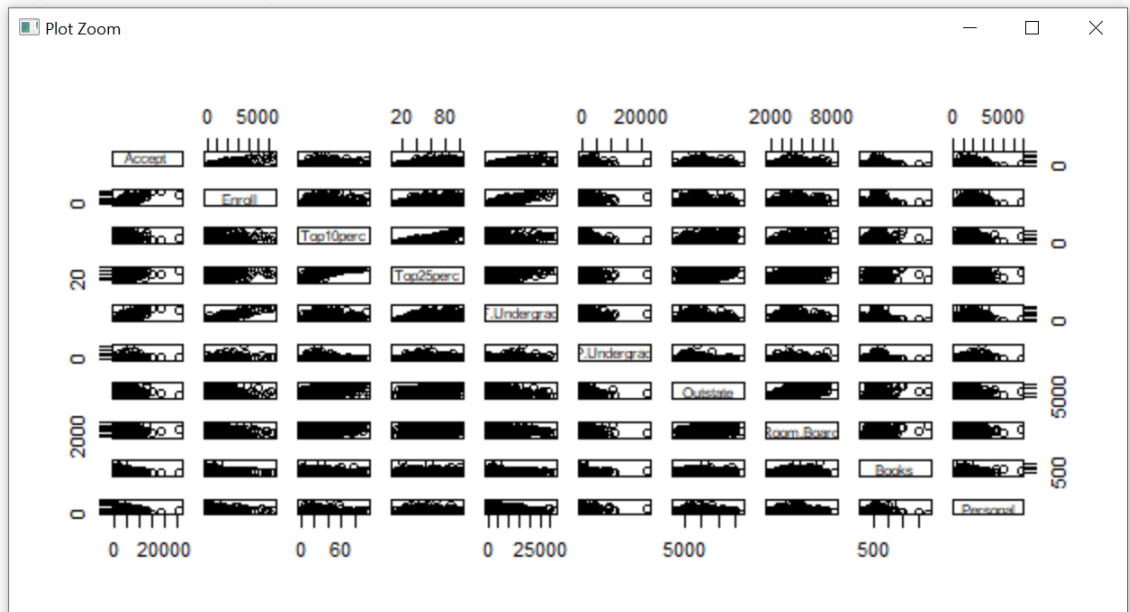
  

Room.Board	Books	Personal
------------	-------	----------

---

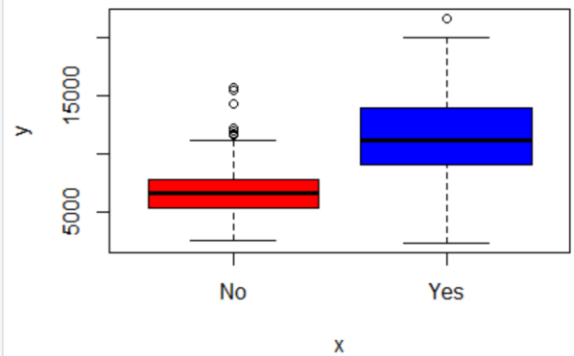
ii) Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

```
> pairs(College[,1:10])
```



iii) use the plot() function to produce side-by-side boxplots of Outstate versus Private.

```
> plot(College$Private, College$Outstate, col=c("red", "blue"))
>
```



iv) Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classes exceeds 50 %.

x and y lengths differ

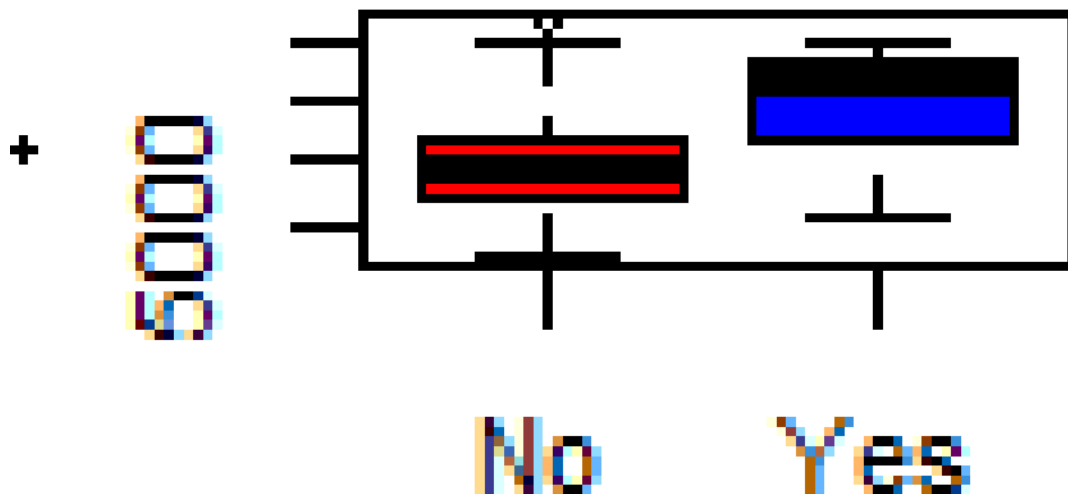
```
> Elite=rep("No",nrow(College))
> Elite[College$Top10perc > 50]="Yes"
> Elite=as.factor(Elite)
> College=data.frame(College,Elite)
> summary(College$Elite)
```

```
No Yes
699  78
```

```
> |
```

Error in plot.new() : figure margins too large

```
> plot(College$Elite,College$Outstate,col=c("red","blue"))
> |
```

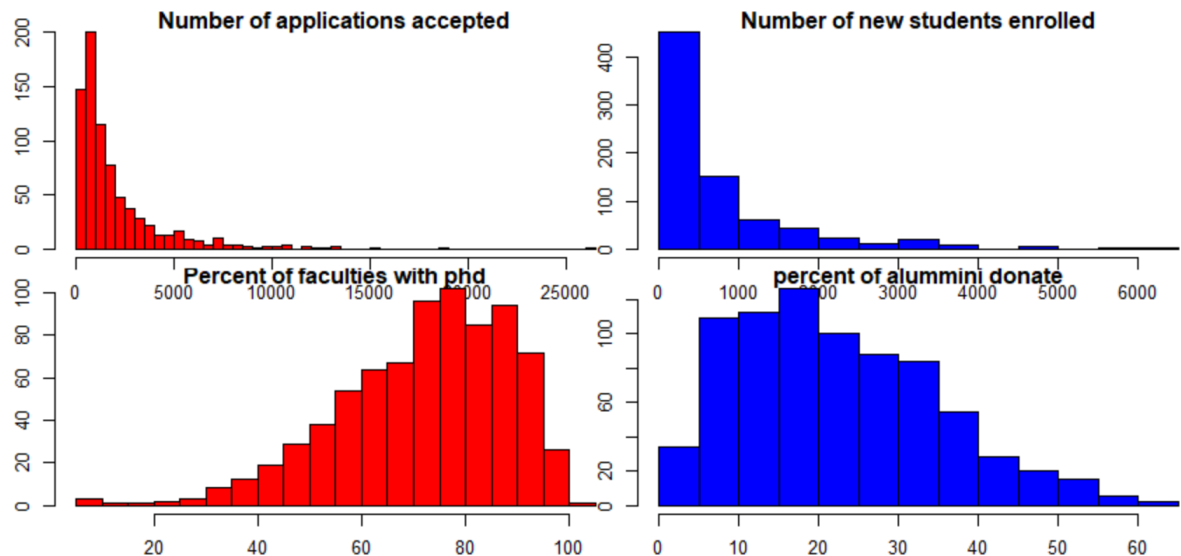


v) use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command `par(mfrow=c(2,2))` useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

```

> par(mfrow=c(2,2))
Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
> hist(College$Accept,main = "Number of applications accepted", col = "red",breaks = 50)
> hist(College$Enroll,main = "Number of new students enrolled",col = "blue",bin=100)
Warning messages:
1: In plot.window(xlim, ylim, "", ...) :
  "bin" is not a graphical parameter
2: In title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...) :
  "bin" is not a graphical parameter
3: In axis(1, ...) : "bin" is not a graphical parameter
4: In axis(2, ...) : "bin" is not a graphical parameter
> hist(College$PhD,main = "Percent of faculties with phd",col = "red",breaks = 20)
Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  "bin" is not a graphical parameter
2: In doTryCatch(return(expr), name, parentenv, handler) :
  "bin" is not a graphical parameter
8: In doTryCatch(return(expr), name, parentenv, handler) :
  "bin" is not a graphical parameter
> hist(College$perc.alumni,main = "percent of alummini donate",col = "blue")
Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  "bin" is not a graphical parameter
2: In doTryCatch(return(expr), name, parentenv, handler) :
  "bin" is not a graphical parameter
3: In doTryCatch(return(expr), name, parentenv, handler) :
  "bin" is not a graphical parameter

```



vi) Continue exploring the data, and provide a brief summary of what you discover.

```
> summary(College$PhD)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.00  62.00   75.00   72.66  85.00  103.00

> summary(College$Enroll)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   35   242    434    780   902   6392

> |
```

9)

a) Which of the predictors are quantitative, and which are qualitative?

```
> data("Auto")
> summary(Auto)
```

mpg	cylinders	displacement	horsepower
Min. : 9.00	Min. : 3.000	Min. : 68.0	Min. : 46.0
1st Qu.: 17.00	1st Qu.: 4.000	1st Qu.: 105.0	1st Qu.: 75.0
Median : 22.75	Median : 4.000	Median : 151.0	Median : 93.5
Mean : 23.45	Mean : 5.472	Mean : 194.4	Mean : 104.5
3rd Qu.: 29.00	3rd Qu.: 8.000	3rd Qu.: 275.8	3rd Qu.: 126.0
Max. : 46.60	Max. : 8.000	Max. : 455.0	Max. : 230.0

weight	acceleration	year	origin
Min. : 1613	Min. : 8.00	Min. : 70.00	Min. : 1.000
1st Qu.: 2225	1st Qu.: 13.78	1st Qu.: 73.00	1st Qu.: 1.000
Median : 2804	Median : 15.50	Median : 76.00	Median : 1.000
Mean : 2978	Mean : 15.54	Mean : 75.98	Mean : 1.577
3rd Qu.: 3615	3rd Qu.: 17.02	3rd Qu.: 79.00	3rd Qu.: 2.000
Max. : 5140	Max. : 24.80	Max. : 82.00	Max. : 3.000

name
amc matador : 5
ford pinto : 5
toyota corolla : 5

Quantitative variables: mpg, cylinders, displacement, horsepower, weight, acceleration

Qualitative variables: Year, origin, name.

b) What is the range of each quantitative predictor? You can answer this using the range() function.



```

> sapply(Auto[,c(1:6),],range)
      mpg cylinders displacement horsepower
[1,]   9.0         3          68         46
[2,]  46.6         8         455        230
      weight acceleration
[1,]  1613           8.0
[2,]  5140          24.8
> |

```

- c) What is the mean and standard deviation of each quantitative predictor?

```

R 4.1.0 · ~/
> sapply(Auto[,c(1:6),],mean)
      mpg      cylinders displacement
23.445918  5.471939    194.411990
horsepower      weight acceleration
104.469388 2977.584184    15.541327
> sapply(Auto[,c(1:6),],sd)
      mpg      cylinders displacement
7.805007  1.705783    104.644004
horsepower      weight acceleration
38.491160  849.402560     2.758864
> |

```

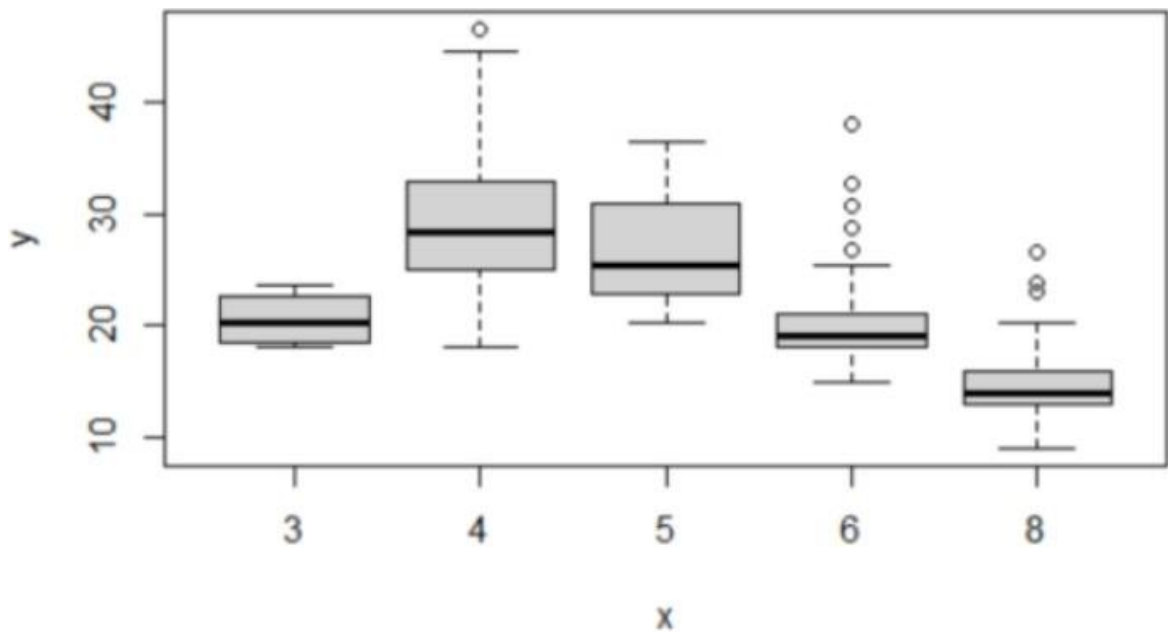
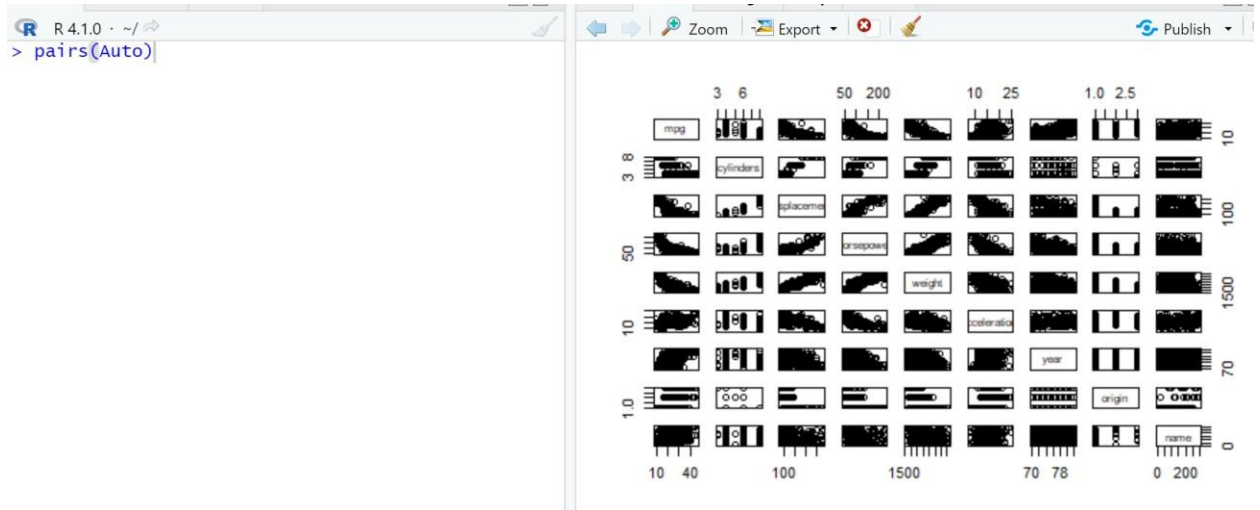
- d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?



```
> new.auto=subset(Auto[-c(10:85),])
> sapply(new.auto[, -c(9)], range)
      mpg cylinders displacement horsepower
[1,]  11.0         3           68         46
[2,]  46.6         8          455        230
      weight acceleration year origin
[1,]   1649           8.5    70     1
[2,]   4997          24.8    82     3
> |
```

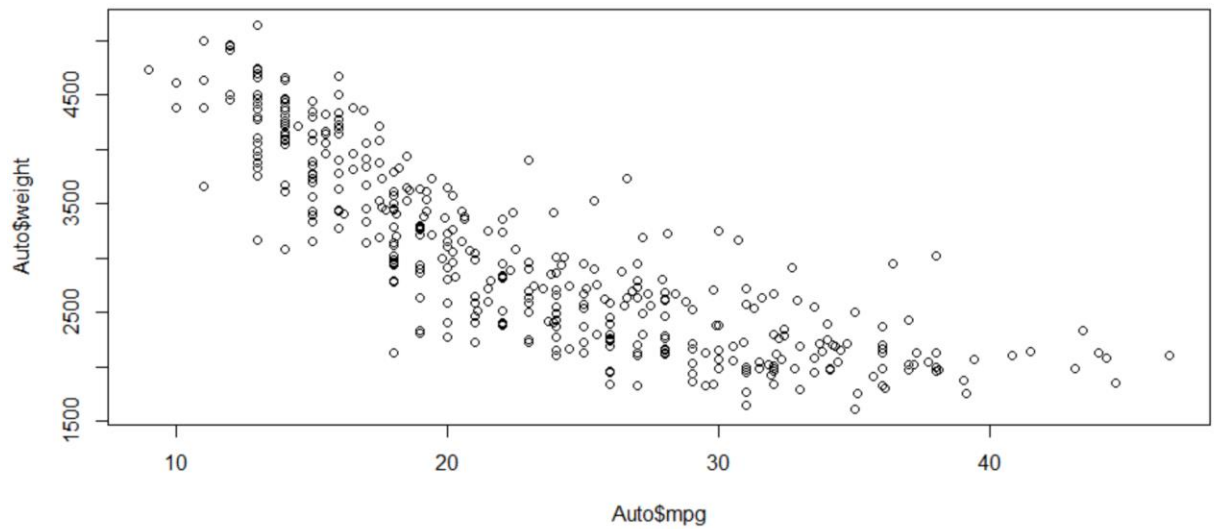
```
> sapply(new.auto[, -c(9)], mean)
      mpg      cylinders displacement
24.404430    5.373418    187.240506
horsepower      weight acceleration
100.721519 2935.971519    15.726899
      year      origin
77.145570    1.601266
> sapply(new.auto[, -c(9)], sd)
      mpg      cylinders displacement
7.867283    1.654179    99.678367
horsepower      weight acceleration
35.708853  811.300208    2.693721
      year      origin
3.106217    0.819910
> |
```

- e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings

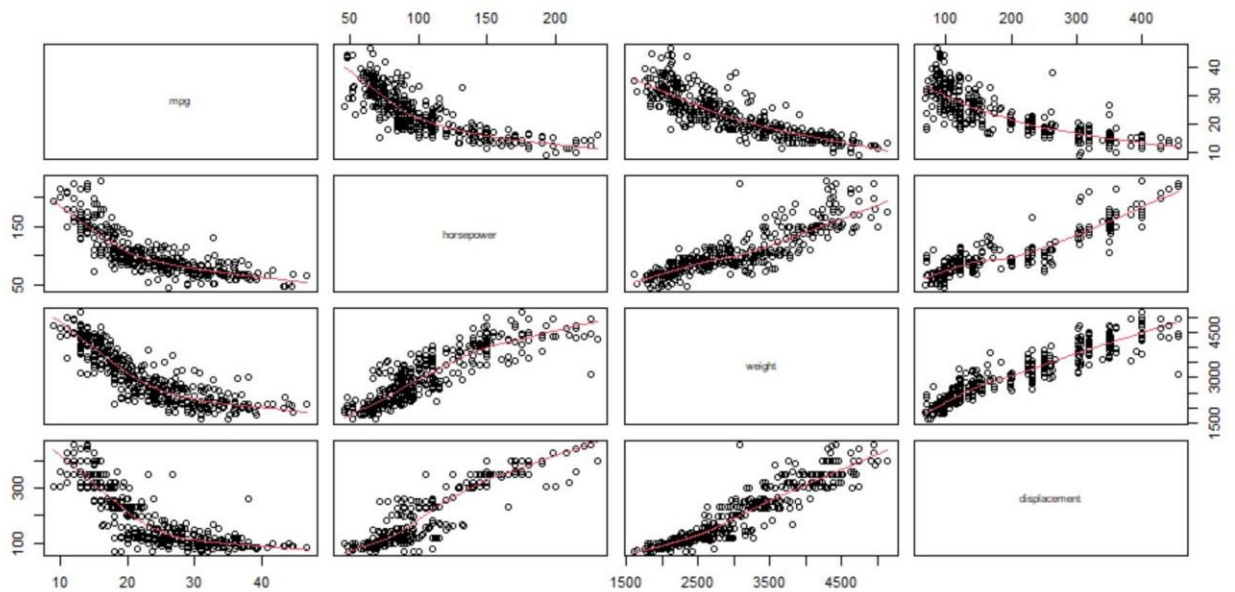


In this we can see that mph is increasing till number of cylinders is equal to 4 and then drops. So larger no of cylinders doesn't mean fuel efficiency.

```
> plot(Auto$mpg, Auto$weight)
> |
```



```
> pairs(~mpg+ horsepower + weight + displacement, data = Auto, panel=panel.smooth)
```

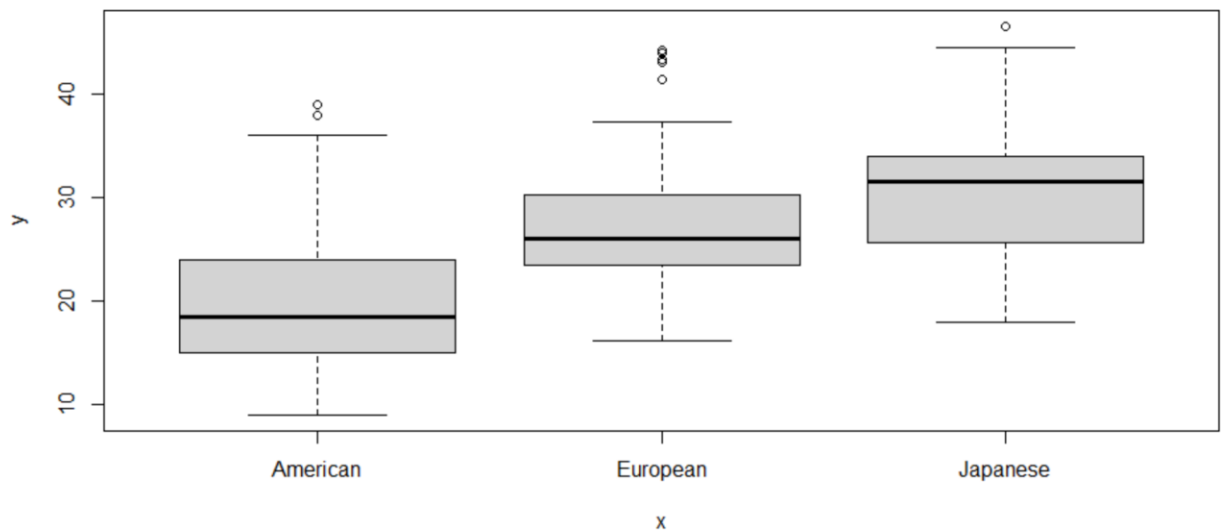


In this mpg is inversely proportional to weight, horsepower, and displacement.

Weight, horsepower and displacement are directly proportional to each other.

```
> plot(factor(Auto$origin), Auto$mpg, names=c("American", "European", "Japanese"))
```

```
> |
```



Japanese vehicles have more mpg than American and European vehicles.

f. All of the predictors show correlation with mpg. The name predictor has to little observations per name though, so using this as a predictor is likely to result in overfitting the data and will not generalize well.

10)

a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

```
> library(MASS)
> data("Boston")
> summary(Boston)
```

crim	zn	indus	chas	nox	rm
Min. : 0.00632	Min. : 0.00	Min. : 0.46	Min. : 0.00000	Min. : 0.3850	Min. : 3.561
1st Qu.: 0.08205	1st Qu.: 0.00	1st Qu.: 5.19	1st Qu.: 0.00000	1st Qu.: 0.4490	1st Qu.: 5.886
Median : 0.25651	Median : 0.00	Median : 9.69	Median : 0.00000	Median : 0.5380	Median : 6.208
Mean : 3.61352	Mean : 11.36	Mean : 11.14	Mean : 0.06917	Mean : 0.5547	Mean : 6.285
3rd Qu.: 3.67708	3rd Qu.: 12.50	3rd Qu.: 18.10	3rd Qu.: 0.00000	3rd Qu.: 0.6240	3rd Qu.: 6.623
Max. : 88.97620	Max. : 100.00	Max. : 27.74	Max. : 1.00000	Max. : 0.8710	Max. : 8.780

age	dis	rad	tax	ptratio	black
Min. : 2.90	Min. : 1.130	Min. : 1.000	Min. : 187.0	Min. : 12.60	Min. : 0.32
1st Qu.: 45.02	1st Qu.: 2.100	1st Qu.: 4.000	1st Qu.: 279.0	1st Qu.: 17.40	1st Qu.: 375.38
Median : 77.50	Median : 3.207	Median : 5.000	Median : 330.0	Median : 19.05	Median : 391.44
Mean : 68.57	Mean : 3.795	Mean : 9.549	Mean : 408.2	Mean : 18.46	Mean : 356.67
3rd Qu.: 94.08	3rd Qu.: 5.188	3rd Qu.: 24.000	3rd Qu.: 666.0	3rd Qu.: 20.20	3rd Qu.: 396.23
Max. : 100.00	Max. : 12.127	Max. : 24.000	Max. : 711.0	Max. : 22.00	Max. : 396.90

lstat	medv
Min. : 1.73	Min. : 5.00
1st Qu.: 6.95	1st Qu.: 17.02
Median : 11.36	Median : 21.20
Mean : 12.65	Mean : 22.53
3rd Qu.: 16.95	3rd Qu.: 25.00

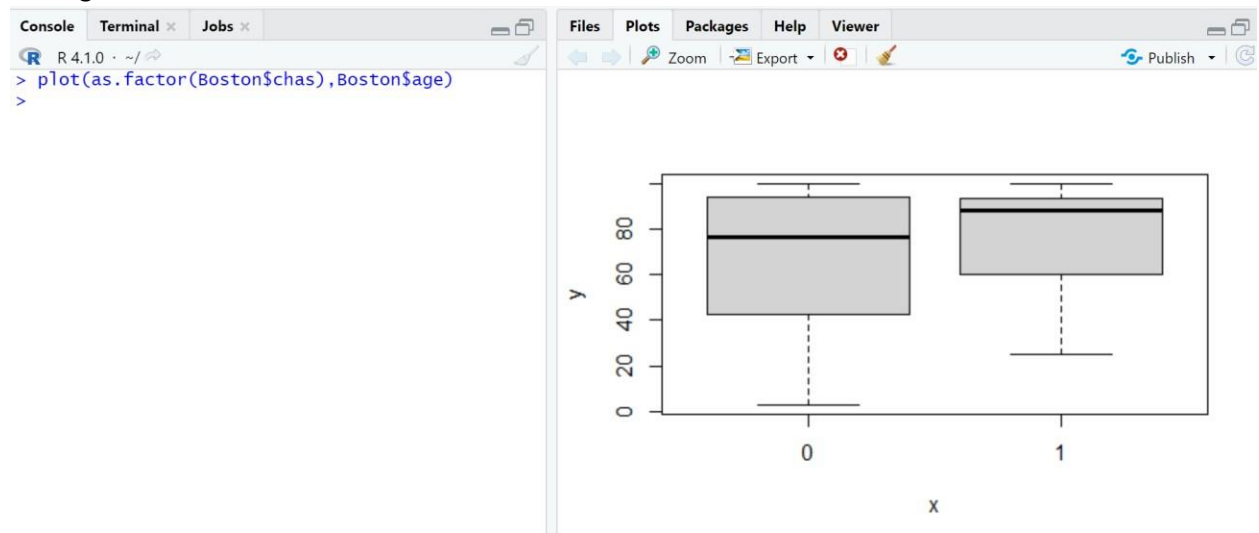
```

> str(Boston)
'data.frame':  506 obs. of  14 variables:
 $ crim    : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn      : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus   : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas    : int   0 0 0 0 0 0 0 0 0 0 ...
 $ nox     : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm      : num  6.58 6.42 7.18 7 7.15 ...
 $ age     : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis     : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad     : int   1 2 2 3 3 3 5 5 5 5 ...
 $ tax     : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black   : num  397 397 393 395 397 ...
 $ lstat   : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv    : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
>

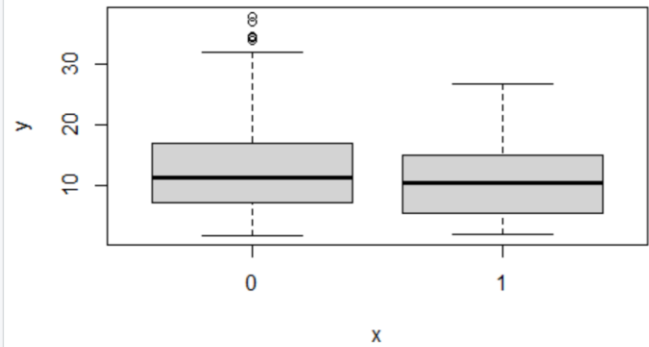
```

Each row represents the set of predictor observations for a given neighborhood in Boston. Each column represent each predictor variable for which an observation was made in 506 neighborhoods of Boston.

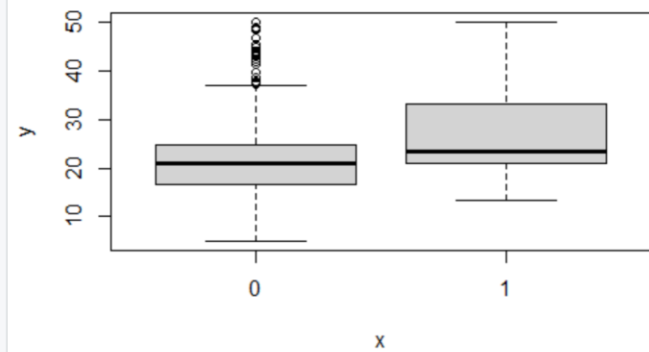
b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.



```
R 4.1.0 ~/
> plot(as.factor(Boston$chas), Boston$lstat)
>
```

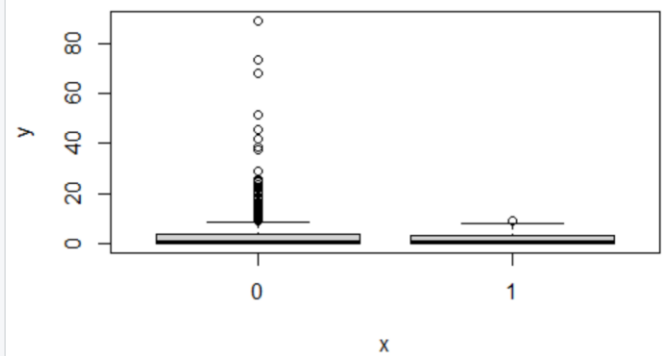


```
> plot(as.factor(Boston$chas), Boston$medv)
>
```

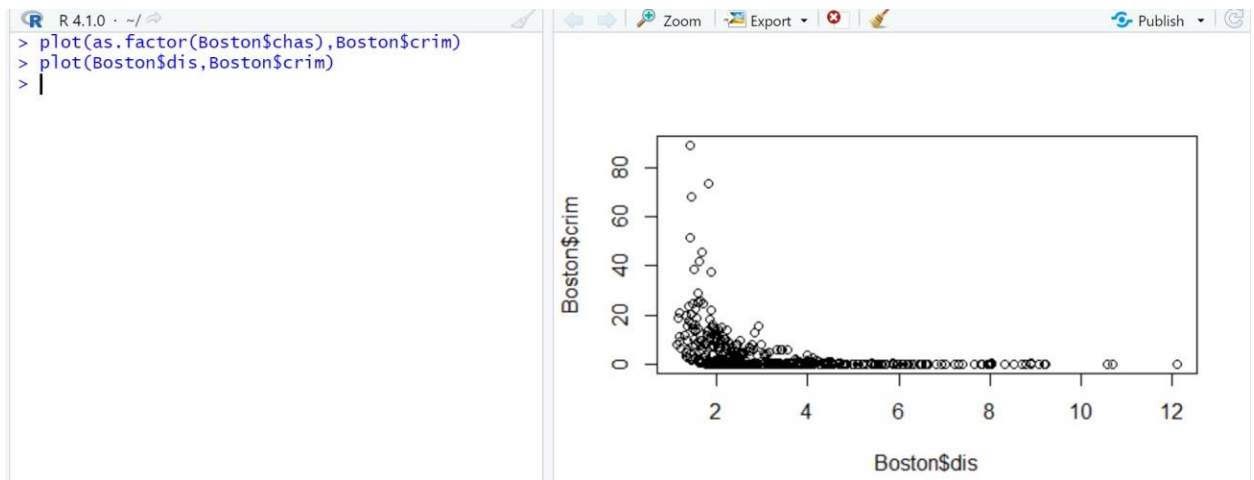


c. Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

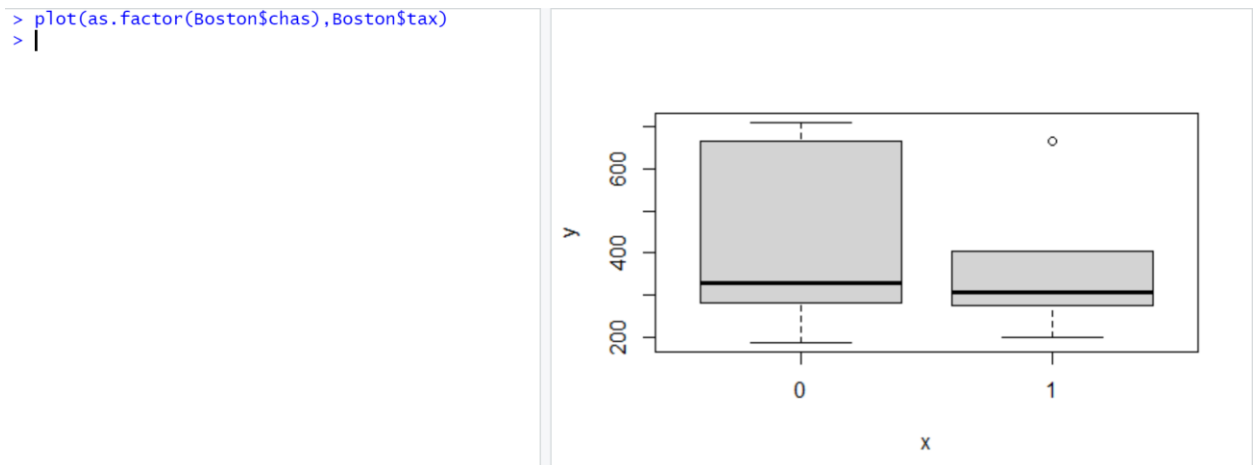
```
> plot(as.factor(Boston$chas), Boston$crim)
>
```







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.

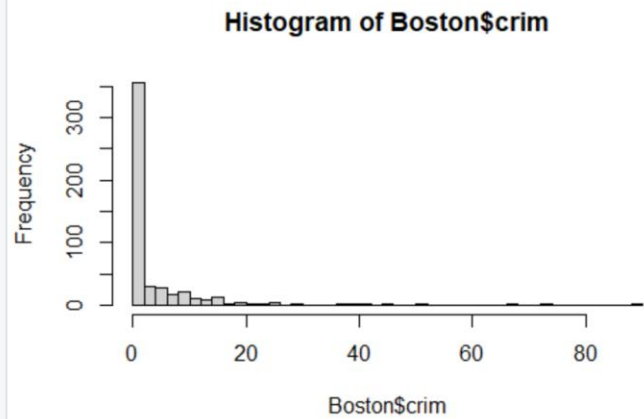


```
> range(Boston$crim)
[1] 0.00632 88.97620
> |
```

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



```
> range(Boston$crim)
[1] 0.00632 88.97620
> hist(Boston$tax)
> hist(Boston$crim,breaks = 50)
> |
```



e. How many of the suburbs in this data set bound the Charles river?

```
> table(Boston$chas)
```

```

  0      1
471    35
> |
```

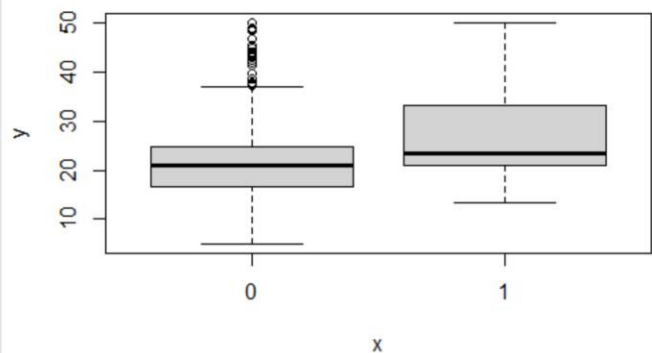
There are 35 Near Charles river

f. What is the median pupil-teacher ratio among the towns in this data set

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

- f. Suburb of Boston has lowest median value of owneroccupied 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.

```
> plot(as.factor(Boston$chas), Boston$medv)
> which.min(Boston$medv)
[1] 399
> Boston[which.min(Boston$medv),]
      crim zn indus chas
399 38.3518  0 18.1    0
      nox  rm age  dis
399 0.693 5.453 100 1.4896
      rad tax ptratio black
399 24 666 20.2 396.9
      lstat medv
399 30.59 5
> summary(Boston$crim)
      Min. 1st Qu.  Median
0.00632  0.08204  0.25651
      Mean 3rd Qu.    Max.
3.61352  3.67708 88.97620
> |
```



- 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.

```

> table(Boston$rm >7)

FALSE  TRUE
  442    64
> table(Boston$rm>8)

FALSE  TRUE
  493    13
> rooms8=Boston[Boston$rm>8, ]
> summary(rooms8)
      crim
Min.   :0.02009
1st Qu.:0.33147
Median :0.52014
Mean   :0.71879
3rd Qu.:0.57834
Max.   :3.47428
      zn
Min.   : 0.00
1st Qu.: 0.00
Median : 0.00
Mean   :13.62
3rd Qu.:20.00
> table(rooms8$chas)

 0  1
11  2
> summary(rooms8$black)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
354.6  384.5   386.9   385.2   389.7   396.9
> summary(Boston$black)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.32  375.38   391.44   356.67   396.23   396.90
>

```

There are 64 houses with more than 7 rooms and 13 houses with more than 8 rooms. Crime is less in the areas with 8 rooms. All the 8 room houses are having blacks population.

```

> library(ISLR)
> data("Auto")
> lm.fit=lm(mpg~horsepower, data = Auto)
> summary(lm.fit)

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

(Intercept) ***
horsepower  ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

g. Is there a relationship between the predictor and the response?

The P-values for the regression coefficients are nearly zero. This implies statistical significance, which in turn mean that there is a relationship.

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

The  $R^2$  value indicates that about 61% of the variation in the response variable(mpg) is due to the predictor variable(horsepower).

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

The regression coefficient for 'horsepower' is negative. Hence, relationship is negative.

iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?

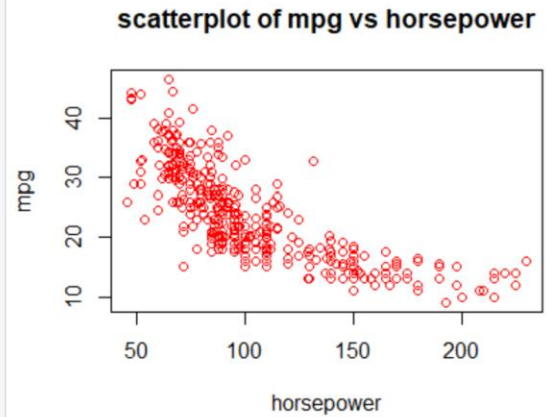
```

> predict(lm.fit,data.frame(horsepower=c(98)),interval = "prediction")
      fit      lwr      upr
1 24.46708 14.8094 34.12476
> predict(lm.fit,data.frame(horsepower=c(98)),interval = "confidence")
      fit      lwr      upr
1 24.46708 23.97308 24.96108
>

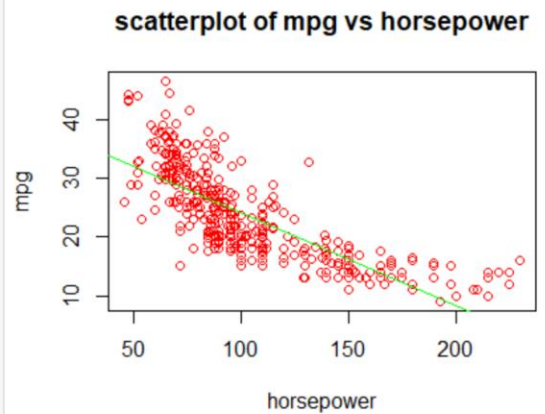
```

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

```
> plot(Auto$horsepower,Auto$mpg,main = "scatterplot of mpg v
s horsepower", xlab = "horsepower",ylab = "mpg",col="red")
> |
```

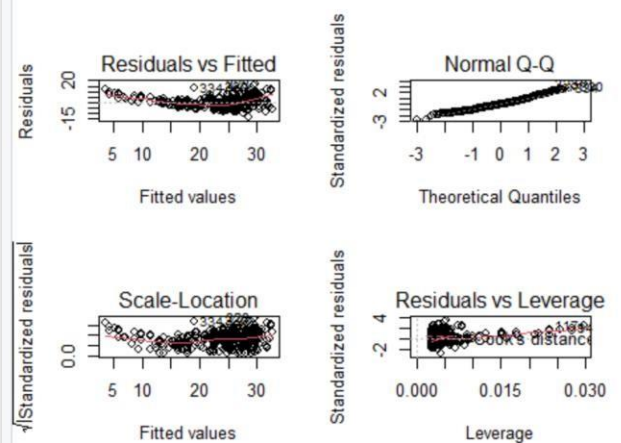


```
> plot(Auto$horsepower,Auto$mpg,main = "scatterplot of mpg v
s horsepower", xlab = "horsepower",ylab = "mpg",col="red")
> abline(lm.fit,col="green")
> |
```



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

```
> par(mfrow=c(2,2))
> plot(lm.fit)
> |
```

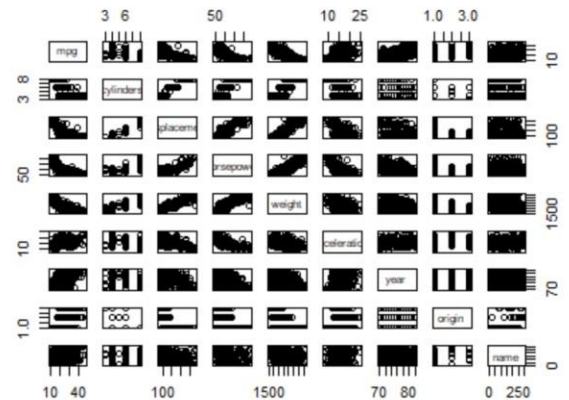


9.

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



```
> pairs(Auto)
> 
```



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

```
> names(Auto)
[1] "mpg"           "cylinders"      "displacement"
[4] "horsepower"    "weight"         "acceleration"
[7] "year"          "origin"         "name"

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

	mpg	cylinders	displacement
mpg	1.0000000	-0.7776175	-0.8051269
cylinders	-0.7776175	1.0000000	0.9508233
displacement	-0.8051269	0.9508233	1.0000000
horsepower	-0.7784268	0.8429834	0.8972570
weight	-0.8322442	0.8975273	0.9329944
acceleration	0.4233285	-0.5046834	-0.5438005
year	0.5805410	-0.3456474	-0.3698552
origin	0.5652088	-0.5689316	-0.6145351

	horsepower	weight	acceleration
mpg	-0.7784268	-0.8322442	0.4233285
cylinders	0.8429834	0.8975273	-0.5046834
displacement	0.8972570	0.9329944	-0.5438005
horsepower	1.0000000	0.8645377	-0.6891955
weight	0.8645377	1.0000000	-0.4168392
acceleration	-0.6891955	-0.4168392	1.0000000
year	-0.4163615	-0.3091199	0.2903161
origin	-0.4551715	-0.5850054	0.2127458

c.

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

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

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

(Intercept) \*\*\*  
cylinders

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

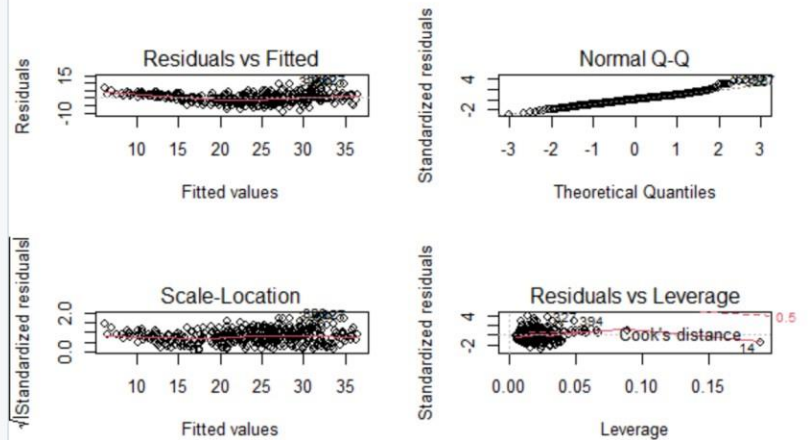
We can answer this by checking P-values associated with each predictors t-statistic. We may all included this by statistically except 'cylinders', 'horsepower' and 'acceleration'.

iii. What does the coefficient for the year variable suggest?

The coefficient of the 'year' variable suggests that the average effect of an increase of 1 year is an increase pf 0.7507727 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?

```
> par(mfrow=c(2,2))
> plot(lm.fit2)
> |
```



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

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

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

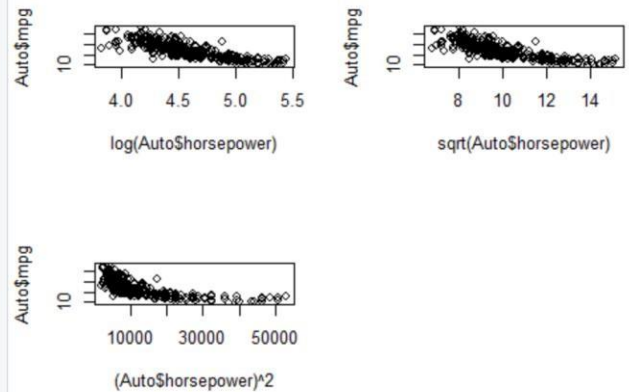
Residual standard error: 4.103 on 386 degrees of freedom

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

```

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

```



10.

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

```

> data("Carseats")
> lm.fit4=lm(Sales~Price+Urban+US, data = Carseats)
> summary(lm.fit4)

```

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
(Intercept)	13.043469	0.651012	20.036
Price	-0.054459	0.005242	-10.389
UrbanYes	-0.021916	0.271650	-0.081
USYes	1.200573	0.259042	4.635

	Pr(> t )
(Intercept)	< 2e-16 ***
Price	< 2e-16 ***
UrbanYes	0.936
USYes	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. Plot the response and the predictor. Use the `abline()` function to display the least squares. The coefficient of the 'price' variables may be interpreted by saying that the average effect of a price of 1 dollar is a decrease of 54.4588492 units in the sales all other predictors remaining fixed. The coefficient of the 'urban' variables may be interpreted by saying that the average sales in the us store are 1200.572 units more than in a no US store all other predictors.

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

The model may be written as

$$\text{Sales} = 13.0434 + (-0.0544) * \text{price} + (-0.02191) * \text{urban} + (1.2005727) * \text{US} + E$$

d. For which of the predictors can you reject the null hypothesis  $H_0 : \beta_j = 0$ ?

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.

```
> lm.fit5=lm(Sales~Price+US, data = Carseats)
> summary(lm.fit5)
```

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
(Intercept)	13.03079	0.63098	20.652
Price	-0.05448	0.00523	-10.416
USYes	1.19964	0.25846	4.641

Pr(>|t|)

(Intercept)	< 2e-16	***
Price	< 2e-16	***
USYes	4.71e-06	***

---  
Signif. codes:  
0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.'  
0.1 ' ' 1

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

The R square for the smaller model is marginally better than for the bigger model. Essentially about 23.92% of the variability is explained by the model.

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

```
> confint(lm.fit5)
```

	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)?

