# Homework 1

Naga Kartheek Peddisetty, 50538422

9/11/2023

# Question 1

Consider the “Smarket” data in the ISLR2 package data(Smarket)

## a) What is the dimension of this data?

```
library(ISLR2)
data(Smarket)
dim(Smarket)
```

```
## [1] 1250    9
```

```
## OR

nrow(Smarket)
```

```
## [1] 1250
```

```
ncol(Smarket)
```

```
## [1] 9
```

1250(rows) Observations of 9(columns) variables.

## b) What are the average for: Lag1 – Lag5. Calculate this in two different ways.

```
#### Type - 1
colMeans(Smarket[,2:6])
```

```
##      Lag1      Lag2      Lag3      Lag4      Lag5
## 0.0038344 0.0039192 0.0017160 0.0016360 0.0056096
```

OR

```
#### Type - 2
apply(Smarket[,2:6],2,mean)
```

```
##      Lag1      Lag2      Lag3      Lag4      Lag5
## 0.0038344 0.0039192 0.0017160 0.0016360 0.0056096
```

## c) What type of variable is Direction.Using the table function, comment on the frequency of Up and Down.

```
class(Smarket$Direction)
```

```
## [1] "factor"
```

```
typeof(Smarket$Direction)
```

```
## [1] "integer"
```

```
table(Smarket$Direction)
```

```
##
## Down   Up
##  602  648
```

Frequency of Down is 602 and Frequency of Up is 648.

### d) Create a list object with data.frames for 2001, 2002, 2003, 2004 and 2005.

```
Year_2001 <- Smarket[Smarket$Year =="2001", ]
Year_2002 <- Smarket[Smarket$Year =="2002", ]
Year_2003 <- Smarket[Smarket$Year =="2003", ]
Year_2004 <- Smarket[Smarket$Year =="2004", ]
Year_2005 <- Smarket[Smarket$Year =="2005", ]

Yearly_data <- list(Year_2001,Year_2002,Year_2003,Year_2004,Year_2005)
# Yearly_data[1] for 2001.
```

As shown above i created a list object with yearly data frames. We can access them by using Yearly_data[1] for 2001, Yearly_data[2] for 2002, ….,Yearly_data[5] for 2005.

# Question 2

### Consider the “Hitters” dataset in the ISLR2 package. Suppose that you are getting this data ready to build a predictive model for salary.Pre-process/clean the data, investigate the data using exploratory data analysis such as scatterplots, and other tools we have discussed. Describe your process and justify any changes you have made to the dataset. Submit the cleaned dataset as an *.RData file to BrightSpace.

```
library(ISLR2)
data(Hitters)
```

```
str(Hitters)
```

```
## 'data.frame':    322 obs. of  20 variables:
##  $ AtBat    : int  293 315 479 496 321 594 185 298 323 401 ...
##  $ Hits     : int  66 81 130 141 87 169 37 73 81 92 ...
##  $ HmRun    : int  1 7 18 20 10 4 1 0 6 17 ...
##  $ Runs     : int  30 24 66 65 39 74 23 24 26 49 ...
##  $ RBI      : int  29 38 72 78 42 51 8 24 32 66 ...
##  $ Walks    : int  14 39 76 37 30 35 21 7 8 65 ...
##  $ Years    : int  1 14 3 11 2 11 2 3 2 13 ...
##  $ CAtBat   : int  293 3449 1624 5628 396 4408 214 509 341 5206 ...
##  $ CHits    : int  66 835 457 1575 101 1133 42 108 86 1332 ...
##  $ CHmRun   : int  1 69 63 225 12 19 1 0 6 253 ...
##  $ CRuns    : int  30 321 224 828 48 501 30 41 32 784 ...
##  $ CRBI     : int  29 414 266 838 46 336 9 37 34 890 ...
##  $ CWalks   : int  14 375 263 354 33 194 24 12 8 866 ...
##  $ League   : Factor w/ 2 levels "A","N": 1 2 1 2 2 1 2 1 2 1 ...
##  $ Division : Factor w/ 2 levels "E","W": 1 2 2 1 1 2 1 2 2 1 ...
##  $ PutOuts  : int  446 632 880 200 805 282 76 121 143 0 ...
##  $ Assists  : int  33 43 82 11 40 421 127 283 290 0 ...
##  $ Errors   : int  20 10 14 3 4 25 7 9 19 0 ...
##  $ Salary   : num  NA 475 480 500 91.5 750 70 100 75 1100 ...
##  $ NewLeague: Factor w/ 2 levels "A","N": 1 2 1 2 2 1 1 1 2 1 ...
```

```
head(Hitters)
```

```
##                   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun
## -Andy Allanson      293   66     1   30  29    14     1    293    66      1
## -Alan Ashby         315   81     7   24  38    39    14   3449   835     69
## -Alvin Davis        479  130    18   66  72    76     3   1624   457     63
## -Andre Dawson       496  141    20   65  78    37    11   5628  1575    225
## -Andres Galarraga   321   87    10   39  42    30     2    396   101     12
## -Alfredo Griffin    594  169     4   74  51    35    11   4408  1133     19
##                   CRuns CRBI CWalks League Division PutOuts Assists Errors
## -Andy Allanson       30   29     14      A        E     446      33     20
## -Alan Ashby         321  414    375      N        W     632      43     10
## -Alvin Davis        224  266    263      A        W     880      82     14
## -Andre Dawson       828  838    354      N        E     200      11      3
## -Andres Galarraga    48   46     33      N        E     805      40      4
## -Alfredo Griffin    501  336    194      A        W     282     421     25
##                   Salary NewLeague
## -Andy Allanson        NA         A
## -Alan Ashby        475.0         N
## -Alvin Davis       480.0         A
## -Andre Dawson      500.0         N
## -Andres Galarraga   91.5         N
## -Alfredo Griffin   750.0         A
```

```
summary(Hitters$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    67.5   190.0   425.0   535.9   750.0  2460.0      59
```

Salary contains 59 Null(NA) values.

```
dim(Hitters)
```

```
## [1] 322  20
```

322(rows) Observations of 20(columns) variables.

```
## Finding mean of the salary column by excluding null values.

mean(Hitters$Salary, na.rm = TRUE)
```

```
## [1] 535.9259
```

```
Hitters_dats <- Hitters

## Imputing Salary with mean.
## replacing null values in Salary column with it's mean.

Hitters_dats$Salary <- replace(Hitters_dats$Salary,is.na(Hitters_dats$Salary),535.9259)
summary(Hitters_dats$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    67.5   226.2   535.9   535.9   700.0  2460.0
```

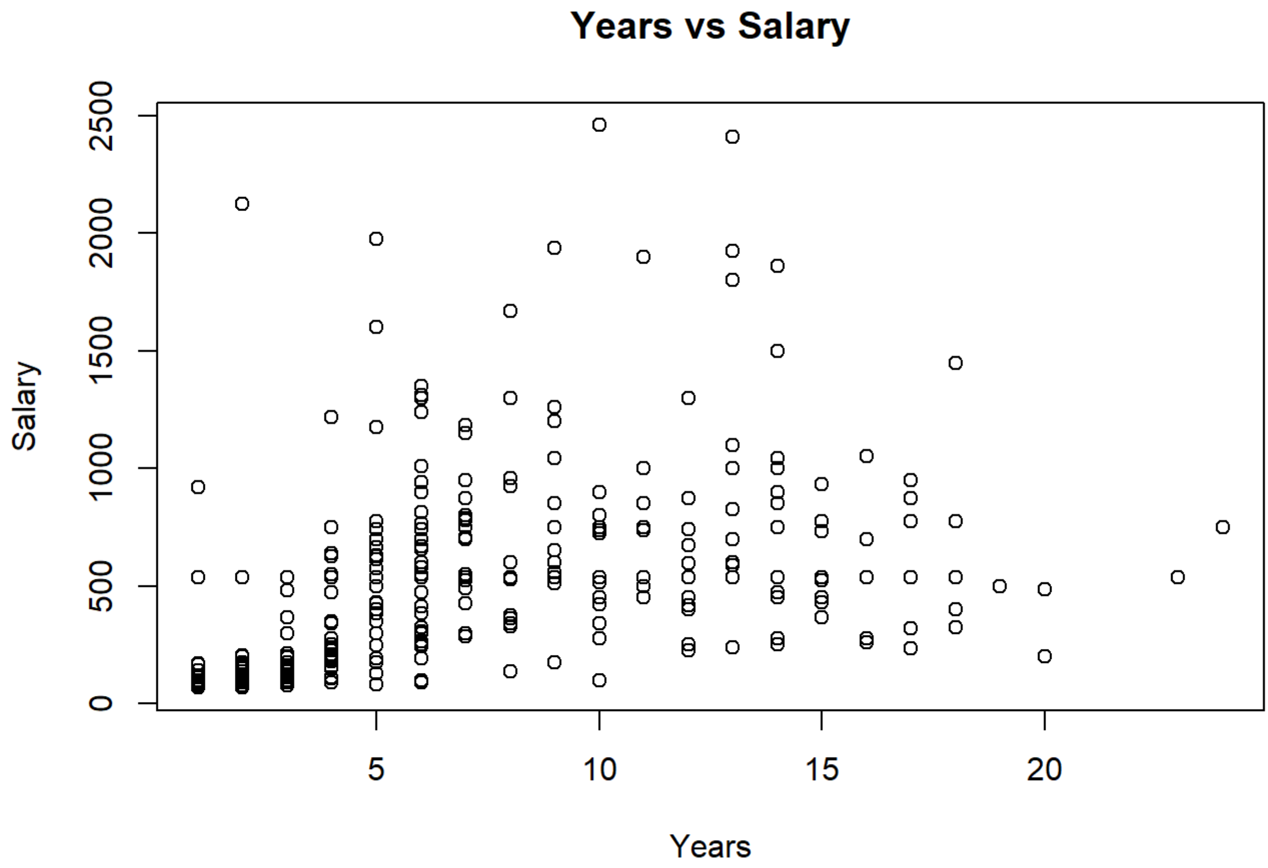
No null(NA) values in Salary column.

```
head(Hitters_dats$Salary)
```

```
## [1] 535.9259 475.0000 480.0000 500.0000  91.5000 750.0000
```

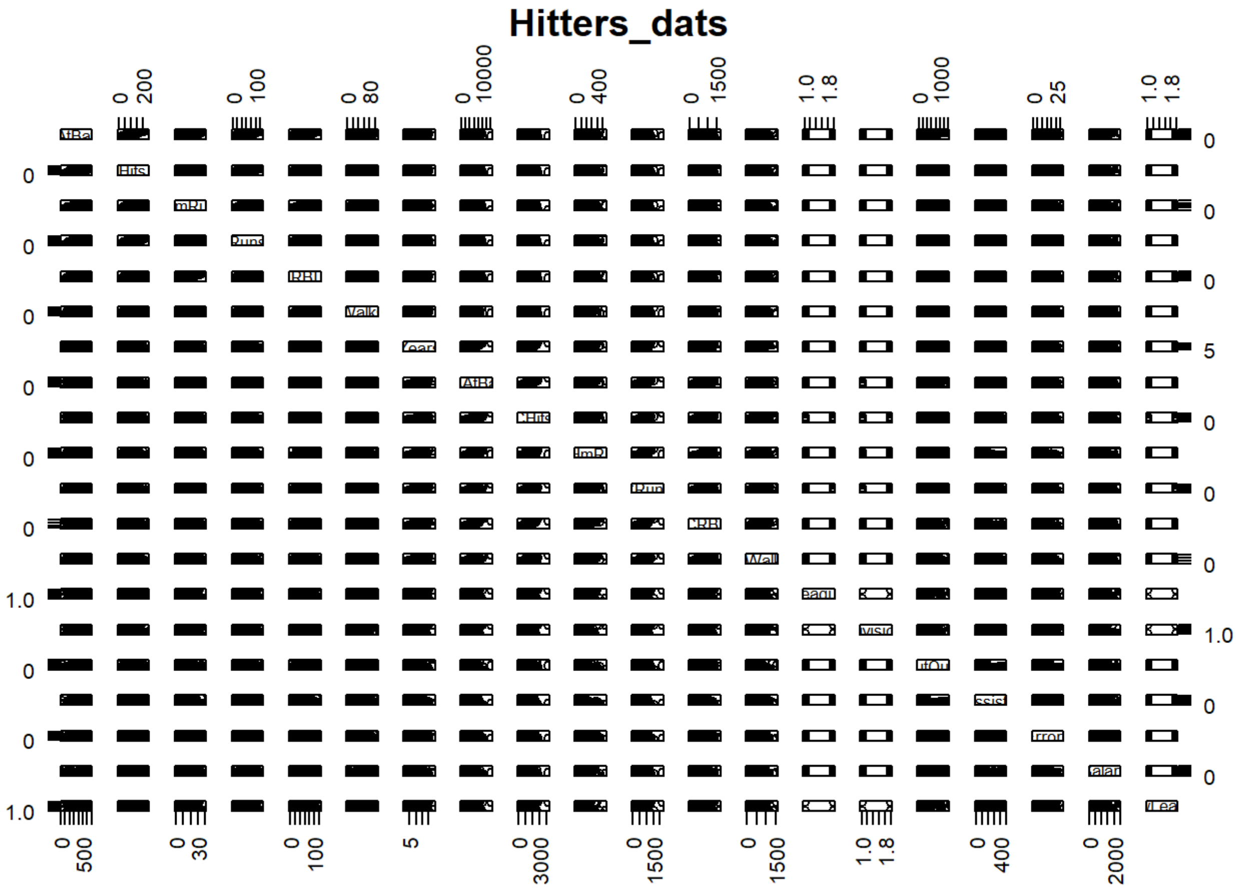
## Exploratory Data analysis

```
plot(Hitters_dats$Years,Hitters_dats$Salary, xlab = "Years", ylab = "Salary", main = "Years v
s Salary")
```

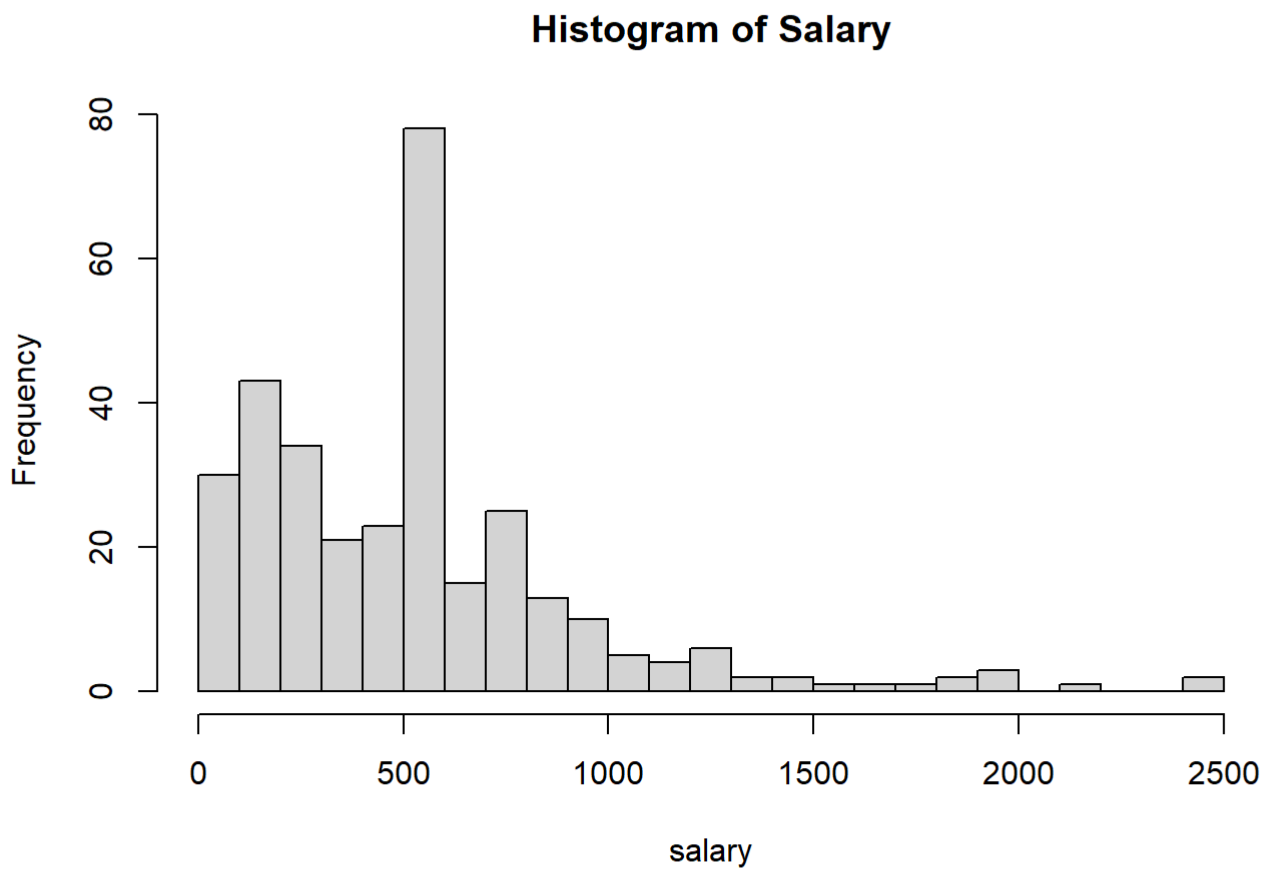



Number of years in the major leagues is more, then salary will increase upto 15 years after it’s decreasing.

```
pairs(Hitters_dats, las = 2, main = "Hitters_dats")
```

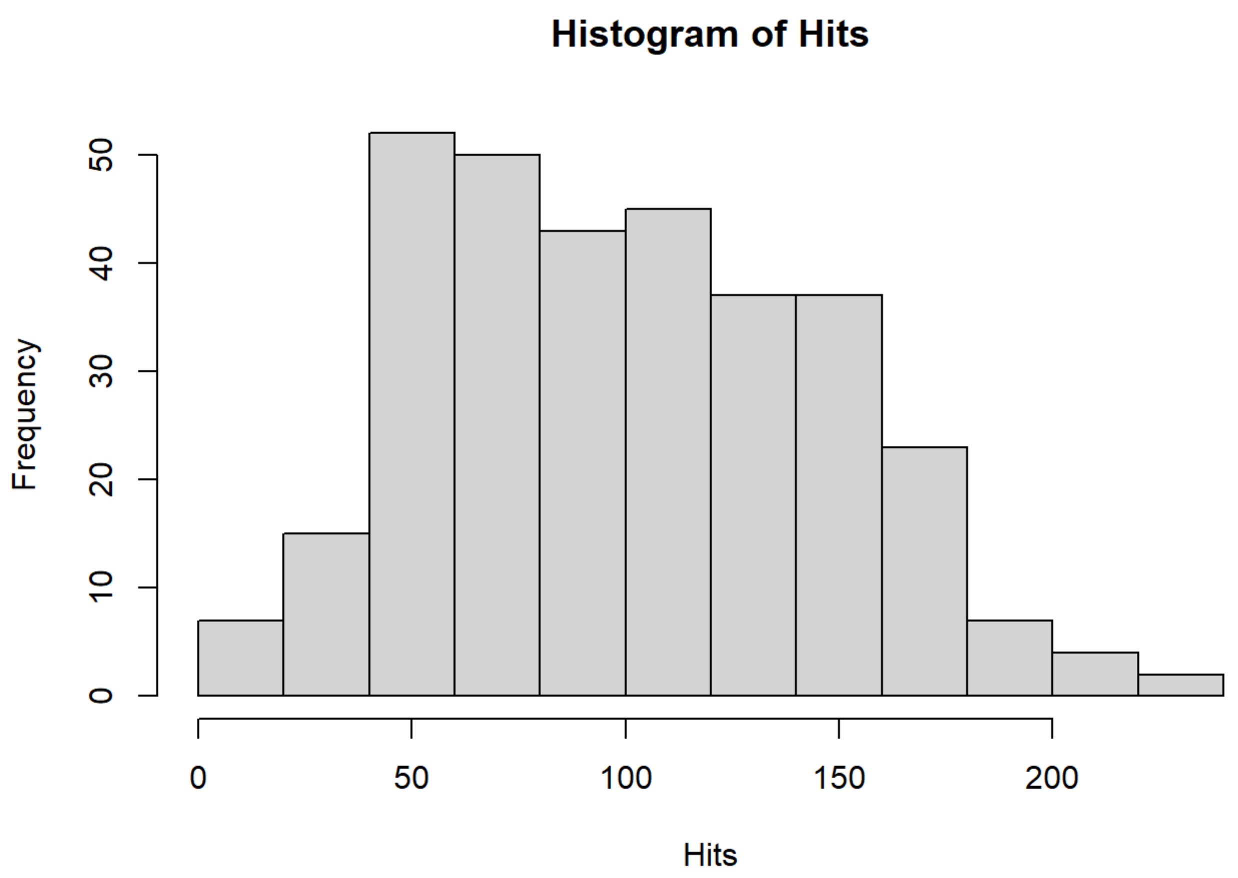



```
hist(Hitters_dats$Salary, breaks = 25, main = "Histogram of Salary", xlab="salary")
```



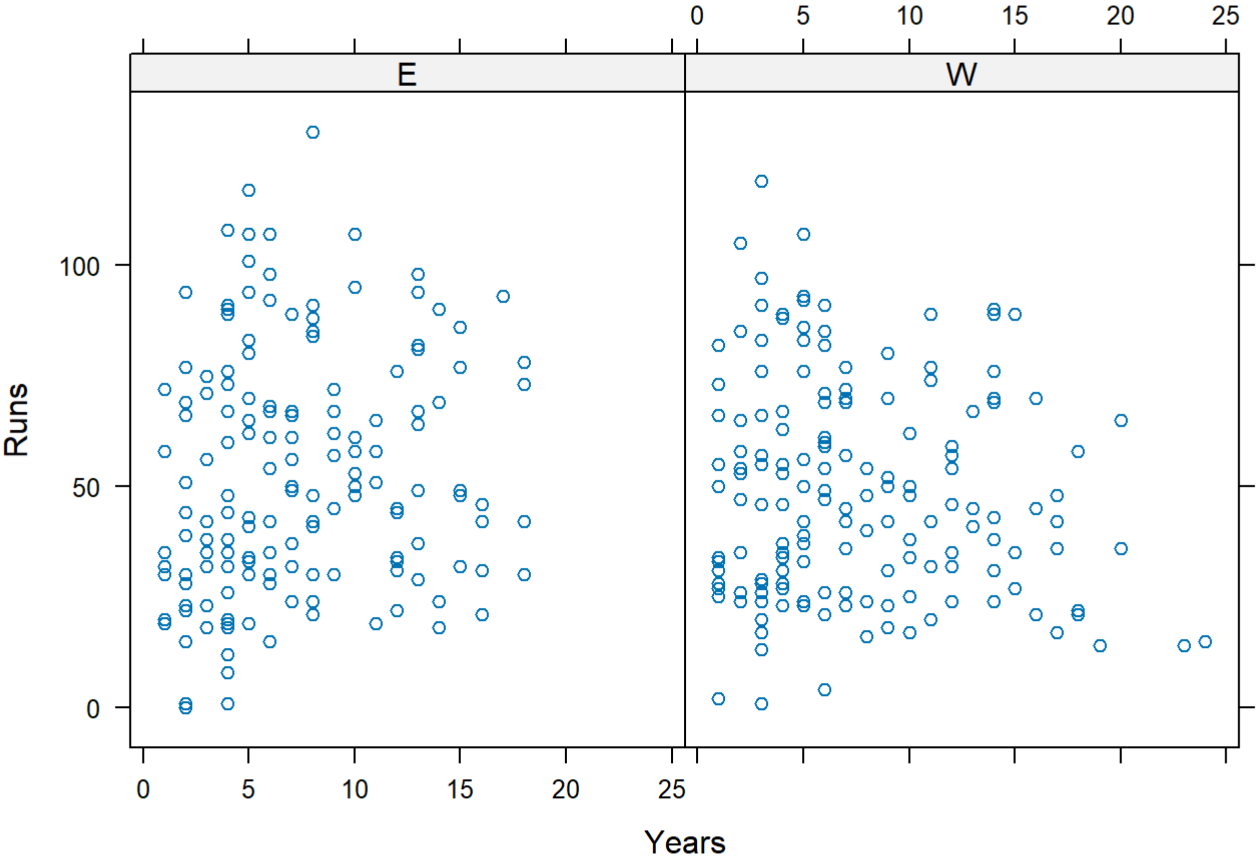
Frequency of Salary is peak at 535.92 because we impute the mean of the salary.

```
hist(Hitters_dats$Hits, xlab = "Hits", main = "Histogram of Hits")
```



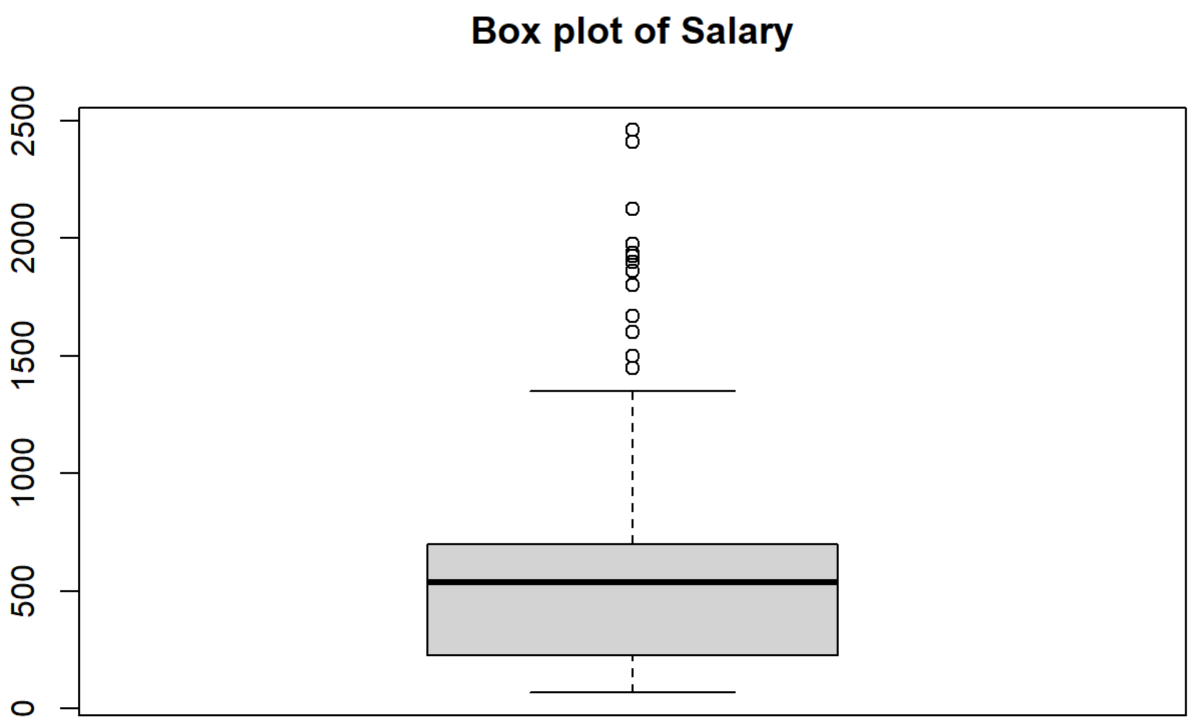

Lattice plot

```
## Plot of two continuous variables runs and years vs categorical variable division.

library(lattice)
xyplot(Runs~Years | Division, data = Hitters_dats)
```




### Box plot

```
boxplot(Hitters_dats$Salary,main = "Box plot of Salary" )
```




```
summary(Hitters_dats$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    67.5   226.2   535.9   535.9   700.0  2460.0
```

# Eliminating Outliers :

IQR(Inter quartile range) = Q3 - Q1 = 700 - 226.2 = 473.8

Outliers are greater than Q3 + (1.5 * IQR) or less than Q1 - (1.5 * IQR)

Q3 + (1.5 * IQR) = 700 + (1.5 * 473.8) = 1410.7

```
##indexes of data where Salary values are greater than 1410.7, storing in elim.

elim <- which(Hitters_dats$Salary > 1410.7)
elim
```

```
##  [1]  83  85  97 101 111 113 164 180 218 230 249 279 314
```

```
## data of salary greater than 1410.7

Hitters_dats[Hitters_dats$Salary > 1410.7, ]
```

```
##                    AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun
## -Don Mattingly       677  238    31  117 113    53     5   2223   737     93
## -Dale Murphy         614  163    29   89  83    75    11   5017  1388    266
## -Dave Winfield       565  148    24   90 104    77    14   7287  2083    305
## -Eddie Murray        495  151    17   61  84    78    10   5624  1679    275
## -George Brett        441  128    16   70  73    80    14   6675  2095    209
## -Gary Carter         490  125    24   81 105    62    13   6063  1646    271
## -Jim Rice            618  200    20   98 110    62    13   7127  2163    351
## -Keith Hernandez     551  171    13   94  83    94    13   6090  1840    128
## -Mike Schmidt         20    1     0    0   0     0     2     41     9      2
## -Ozzie Smith         514  144     0   67  54    79     9   4739  1169     13
## -Rickey Henderson    608  160    28  130  74    89     8   4071  1182    103
## -Steve Garvey        557  142    21   58  81    23    18   8759  2583    271
## -Wade Boggs          580  207     8  107  71   105     5   2778   978     32
##                    CRuns CRBI CWalks League Division PutOuts Assists Errors
## -Don Mattingly       349  401    171      A        E    1377     100      6
## -Dale Murphy         813  822    617      N        W     303       6      6
## -Dave Winfield      1135 1234    791      A        E     292       9      5
## -Eddie Murray        884 1015    709      A        E    1045      88     13
## -George Brett       1072 1050    695      A        W      97     218     16
## -Gary Carter         847  999    680      N        E     869      62      8
## -Jim Rice           1104 1289    564      A        E     330      16      8
## -Keith Hernandez     969  900    917      N        E    1199     149      5
## -Mike Schmidt          6    7      4      N        E      78     220      6
## -Ozzie Smith         583  374    528      N        E     229     453     15
## -Rickey Henderson    862  417    708      A        E     426       4      6
## -Steve Garvey       1138 1299    478      N        W    1160      53      7
## -Wade Boggs          474  322    417      A        E     121     267     19
##                      Salary NewLeague
## -Don Mattingly     1975.000         A
## -Dale Murphy       1900.000         N
## -Dave Winfield     1861.460         A
## -Eddie Murray      2460.000         A
## -George Brett      1500.000         A
## -Gary Carter       1925.571         N
## -Jim Rice          2412.500         A
## -Keith Hernandez   1800.000         N
## -Mike Schmidt      2127.333         N
## -Ozzie Smith       1940.000         N
## -Rickey Henderson 1670.000         A
## -Steve Garvey      1450.000         N
## -Wade Boggs        1600.000         A
```

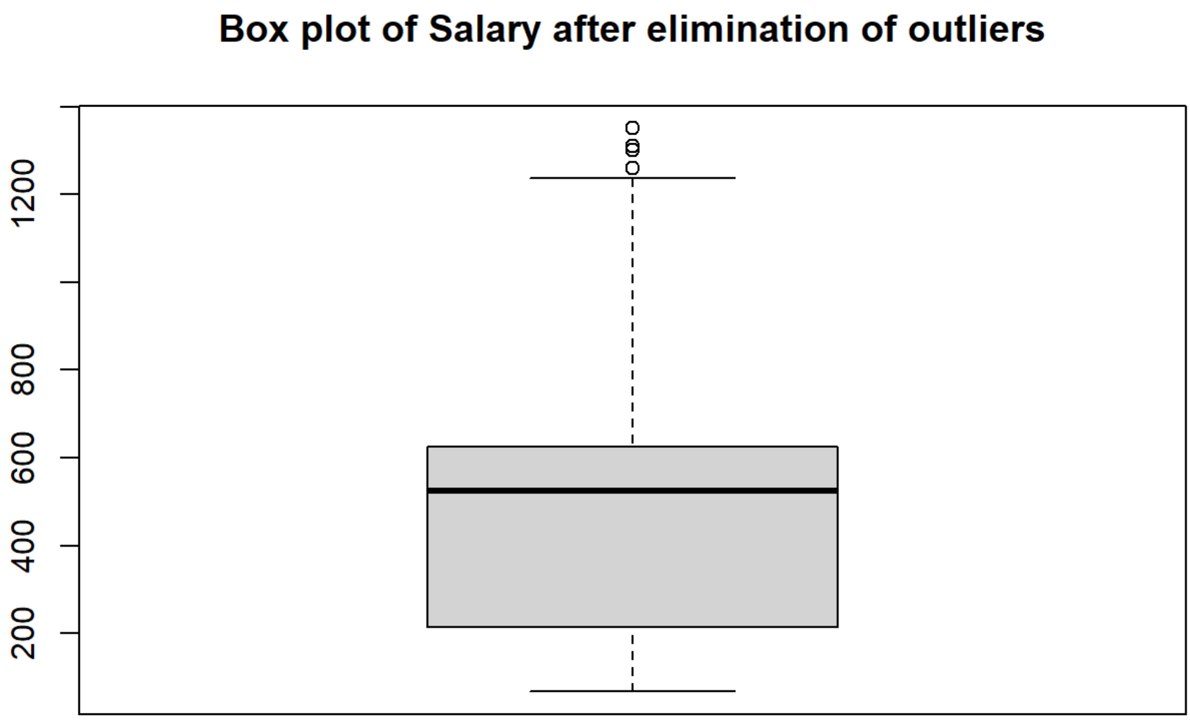
```
dim(Hitters_dats)
```

```
## [1] 322  20
```

```
## Eliminating (outliers from box plot) 13 rows from Hitters_Dats and saving in Hitters_dats_
elim.

Hitters_dats_elim <- Hitters_dats[-elim, ]
dim(Hitters_dats_elim)
```

```
## [1] 309  20
```

```
boxplot(Hitters_dats_elim$Salary, main = "Box plot of Salary after elimination of outliers" )
```




```
## saving the modified or cleaned dataset as RData file.

save(Hitters_dats_elim, file = "Hitters_dats_elim.RData")
```

# Question 3

Divide your data from Q2 into training and test. Use k-nearest neighbors to predict salary. Justify your choice of "k". Provide a profile of the test and training error as a function of the number of neighbors "k".

```
# dividing the data from Question 2 into training and test.

## taking the length of row in data set. Defining total number of data points.

N = length(Hitters_dats_elim[,1])
N
```

```
## [1] 309
```

```
Hitters_dats_elim1 <- Hitters_dats_elim

#### Converting league, New league & division into numeric values.

Hitters_dats_elim1$League <- as.numeric(Hitters_dats_elim1$League)

Hitters_dats_elim1$NewLeague <- as.numeric(Hitters_dats_elim1$NewLeague)

Hitters_dats_elim1$Division <- as.numeric(Hitters_dats_elim1$Division)

set.seed(219)

#### Generating a 2/3 of sample data for training with out replacing.

train_Hitters_dats_elim <- sample(1:N, size = (2/3)*N, replace = FALSE)

#### Dividing data into training and testing,

training_data <- Hitters_dats_elim1[train_Hitters_dats_elim, ]
testing_data <- Hitters_dats_elim1[-train_Hitters_dats_elim, ]
dim(training_data)
```

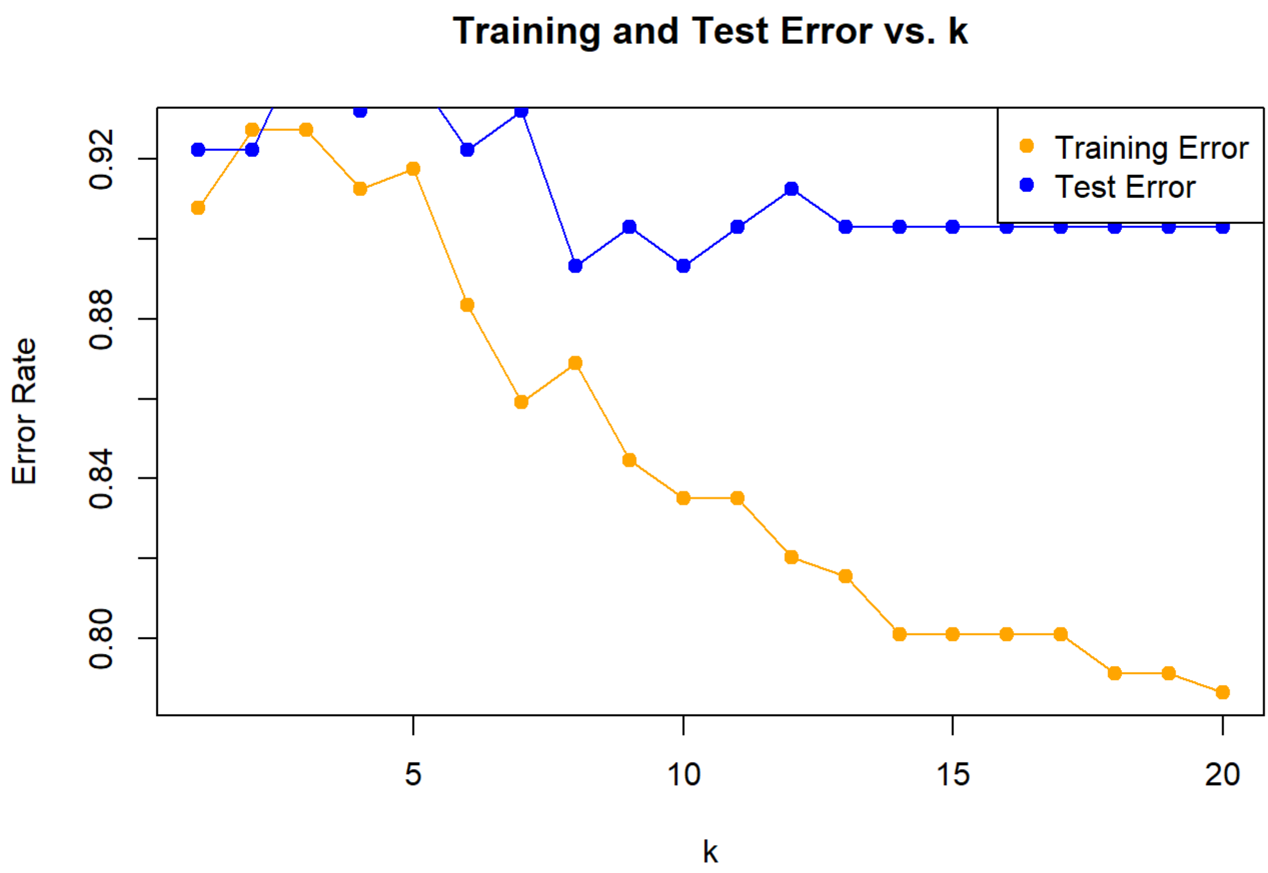
```
## [1] 206  20
```

```
dim(testing_data)
```

```
## [1] 103  20
```

```
### training data except Salary column

X_train = training_data[, -19]

### training data of salary column

Y_train = training_data[, 19]

### testing data except Salary column

X_test =  testing_data[, -19]

### training data of salary column

Y_test = testing_data[, 19]
################################################

library(class)

### Taking k values from 1 to 20.

k_values <- 1:20
train_err <- numeric(length(k_values))
test_err  <- numeric(length(k_values))

##### using for loop to calculate knn, train and test error values.

for (k in k_values) {

  knn_model <- knn(X_train,X_test,Y_train,k=k)
  train_err[k] <- mean(knn_model != Y_train)
  test_err[k] <- mean(knn_model != Y_test)

}

# Plotting of training and test errors as a function of k

plot(k_values, train_err, type = "o", col = "orange", pch = 19, xlab = "k", ylab = "Error Rat
e", main = "Training and Test Error vs. k")
points(k_values, test_err, type = "o", col = "blue", pch = 19)
legend("topright", legend = c("Training Error", "Test Error"), col = c("orange", "blue"), pch
= 19)
```

(-1.5 marks) performed KNN classification instead of knn regression use knn.reg function




```
#### We can choose the best value for k at which the minimum test error occurs.
best_value_k <- k_values[which.min(test_err)]
best_value_k
```

```
## [1] 8
```

The best value for k = 8. Because, the test error is minimum at k=8 as shown in the above plot.

# Question 4

To begin,load in the Boston data set. The Boston data set is part of the ISLR2 library.

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

```
library(ISLR2)
data(Boston)
?Boston
```

```
## starting httpd help server ... done
```

```
dim(Boston)
```

```
## [1] 506  13
```

```
## OR

nrow(Boston)
```

```
## [1] 506
```

```
ncol(Boston)
```

```
## [1] 13
```

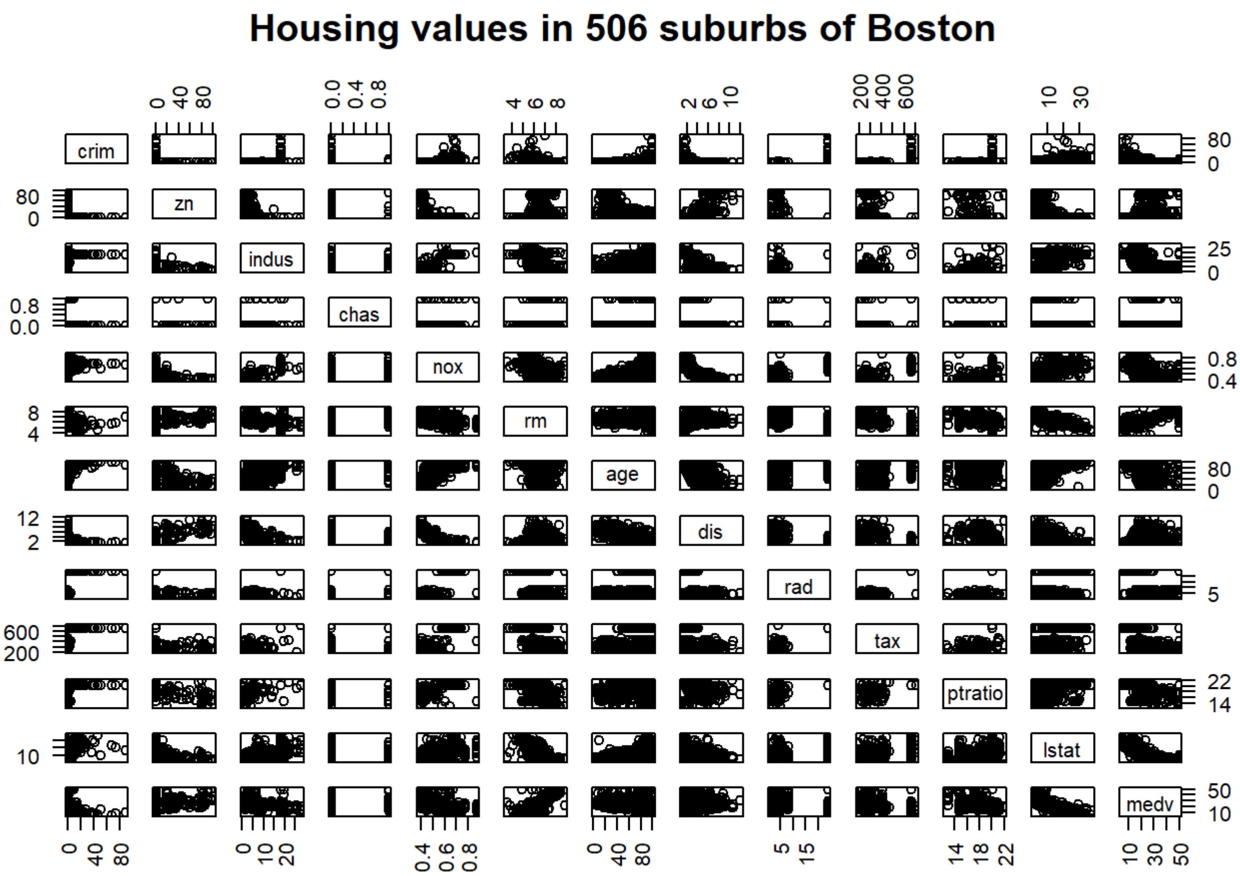
506(rows) observations of 13(columns) variables. And data(rows and columns) represents the housing values in 506 suburbs of Boston.

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

```
str(Boston)
```

```
## 'data.frame':    506 obs. of  13 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 ...
##  $ 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 ...
```

```
Boston_dats <- Boston

## Converting integer to numeric for chas and rad.

Boston_dats$chas <- as.numeric(Boston_dats$chas)
Boston_dats$rad <- as.numeric(Boston_dats$rad)
pairs(Boston_dats, las = 2, main = "Housing values in 506 suburbs of Boston")
```




-> ‘crim’ is positively correlated with indus,nox,age,rad,tax and lstat. Negatively correlated with dis and medv.

-> ‘zn’ is positively correlated with dis and medv. Negatively correlated with indus,nox,age,ptratio and lstat.

-> ‘indus’ is positively correlated with nox,age,rad,tax,lstat and ptratio. Negatively correlated with dis,medv and rm.

-> ‘nox’ is positively correlated with age,rad,tax and lstat. Negatively correlated with dis and medv.

-> ‘rm’ is positively correlated with medv. Negatively correlated with lstat and ptratio.

-> ‘age’ is positively correlated with lstat,tax and rad. Negatively correlated with dis and medv.

-> ‘dis’ is negatively correlated with tax,rad and lstat.

-> ‘rad’ is positively correlated with tax,lstat and ptratio. Negatively correlated with medv.

-> ‘tax’ is positively correlated with lstat and ptratio. Negatively correlated with medv.

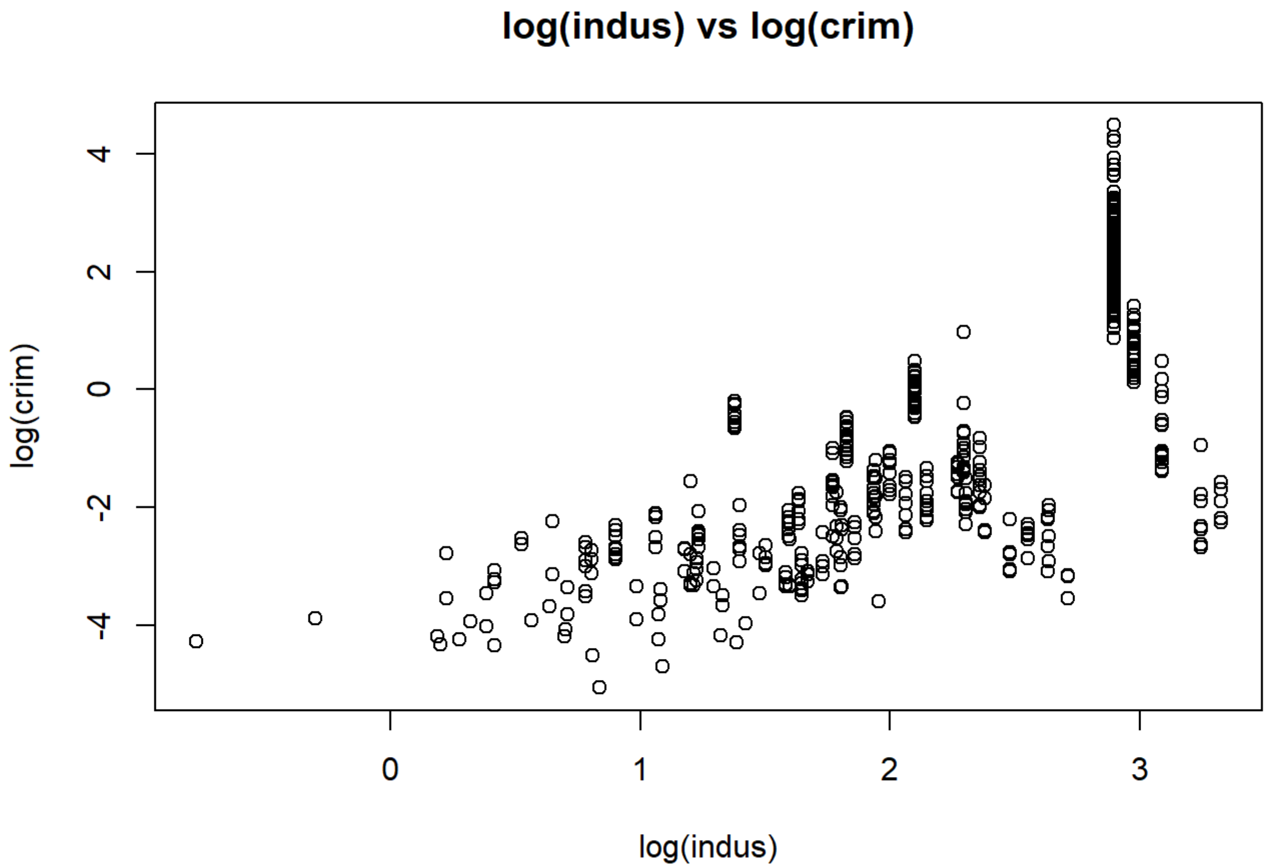
-> ‘ptratio’ is positively correlated with lstat. Negatively correlated with medv.

-> ‘lstat’ is negatively correlated with medv.

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

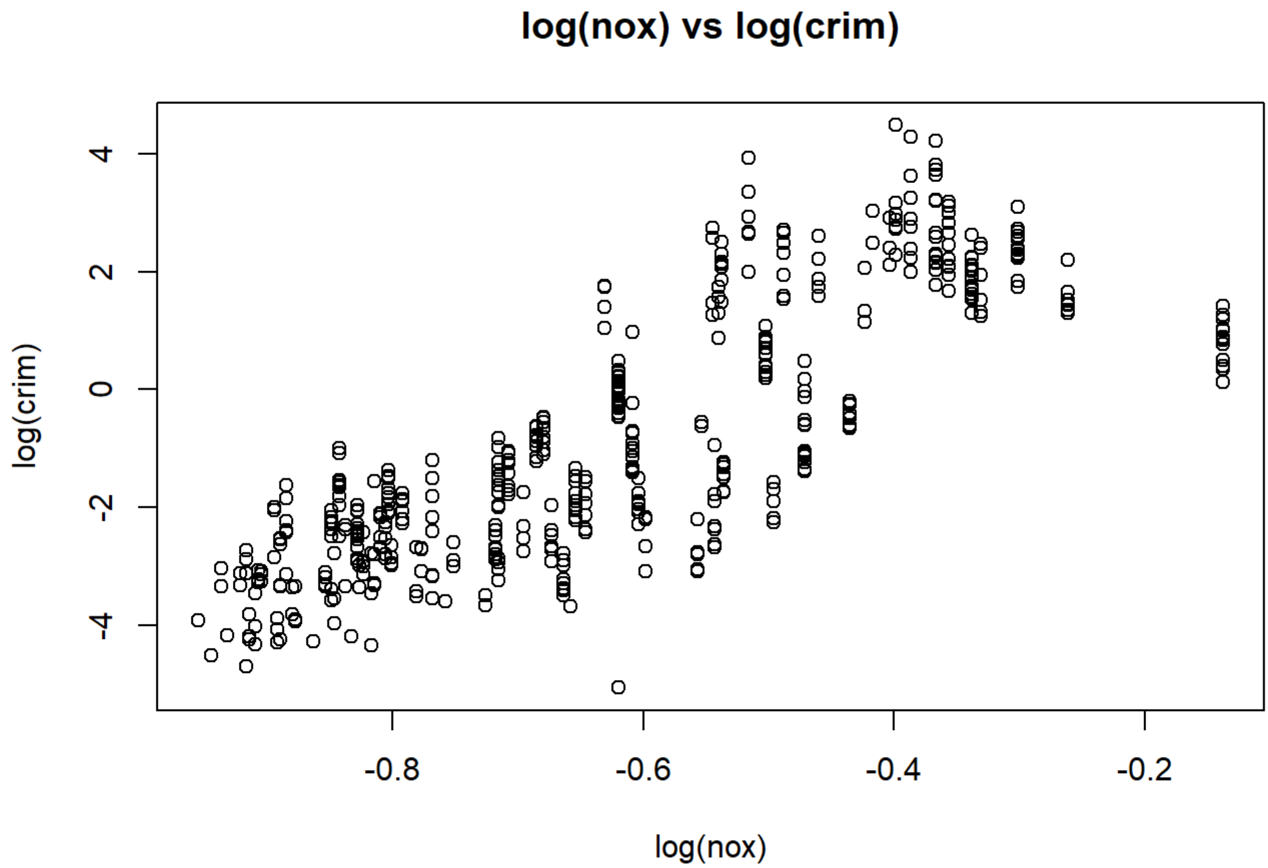
As mentioned above in question b ‘crim’ is positively correlated with indus,nox,age,rad,tax and lstat. Negatively correlated with dis and medv.

```
## For better understanding of the plot i am using transformations.

## Transforming indus and crim values into logarithm.

log_indus <- log(Boston$indus)
log_crim  <- log(Boston$crim)
plot(log_indus,log_crim,xlab = 'log(indus)',ylab = 'log(crim)',main = "log(indus) vs log(cri
m)")
```



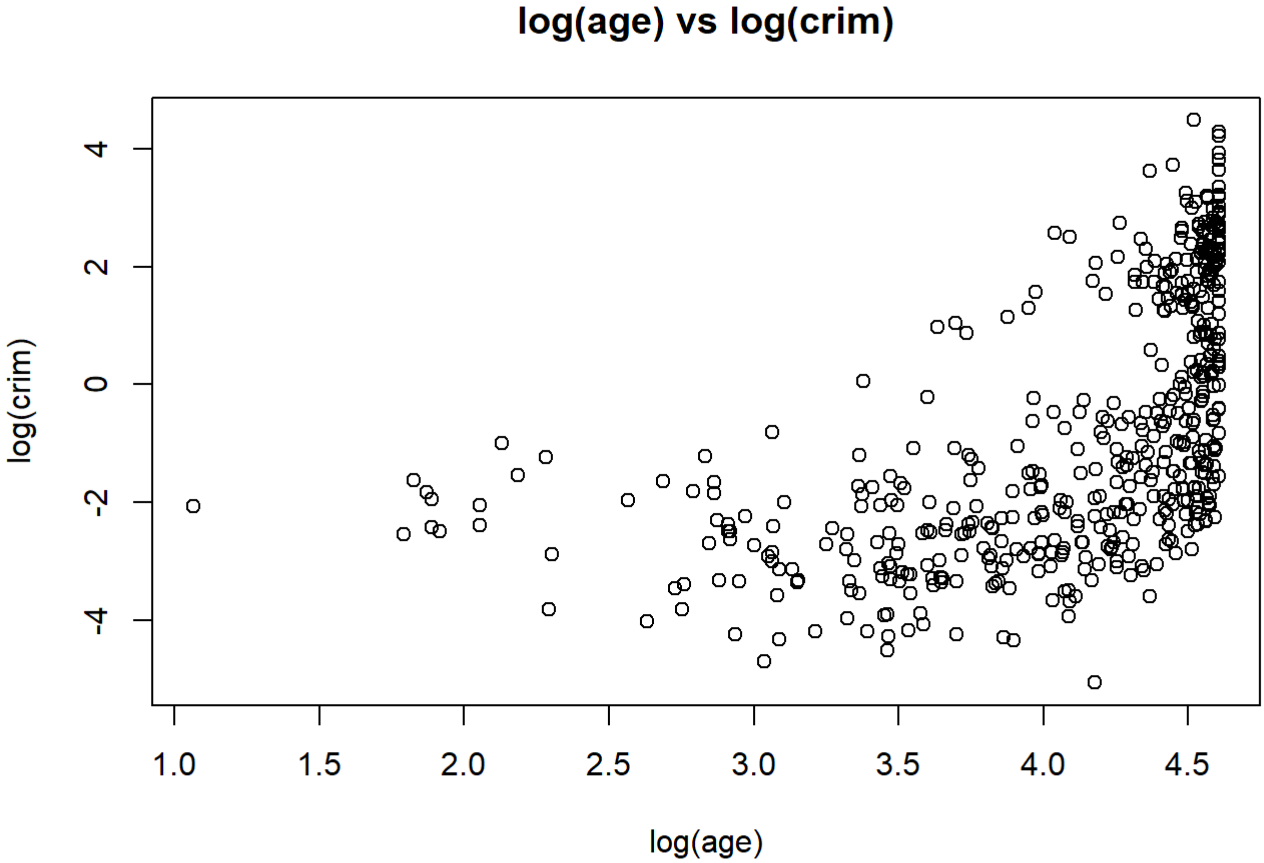

As shown in the above plot crim is positively correlated with indus. Because if crime rate is increasing, indus is also increasing.

```
## Transforming nox values into logarithm.

log_nox <- log(Boston$nox)
plot(log_nox,log_crim,xlab = 'log(nox)',ylab = 'log(crim)',main = "log(nox) vs log(crim)")
```



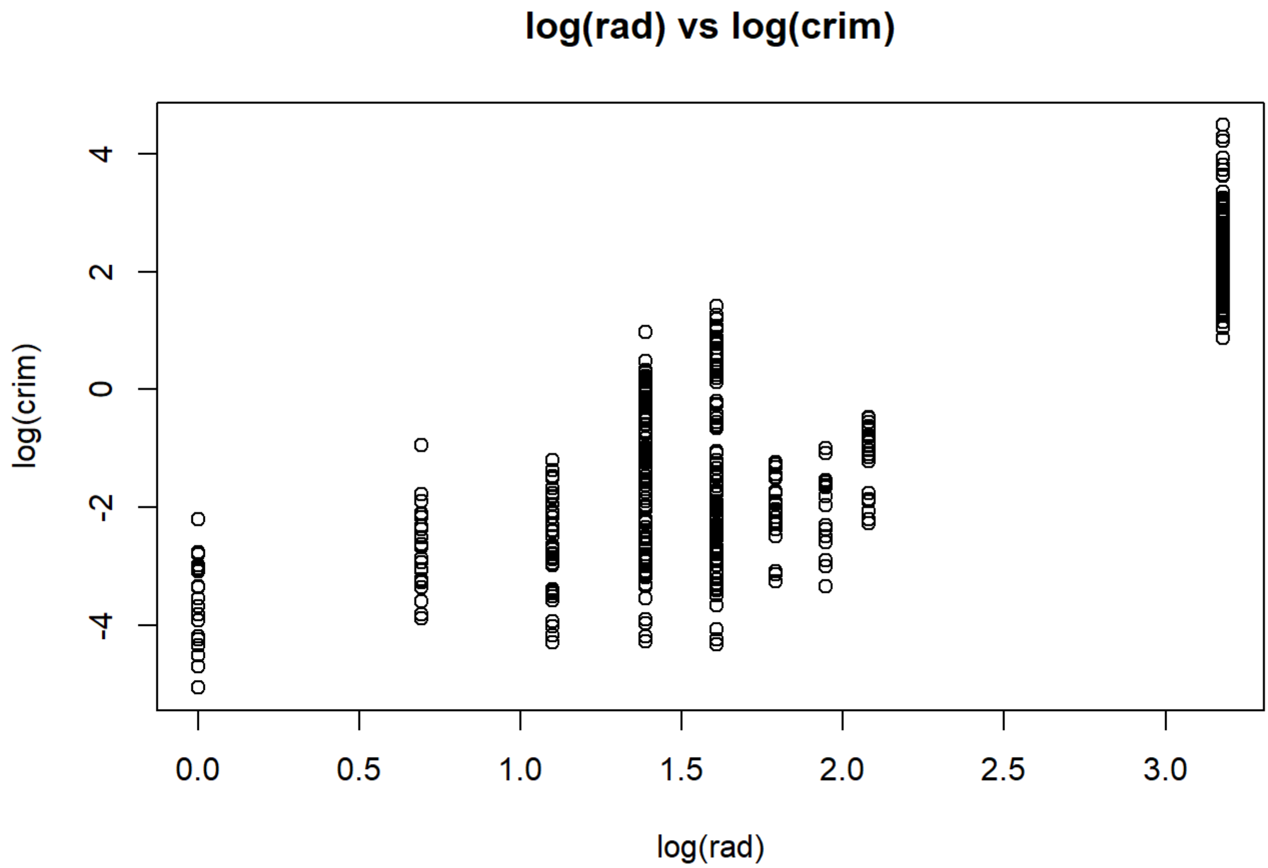

As shown in the above plot crim is positively correlated with nox. Because if crime rate is increasing, nox is also increasing.

```
## Transforming age values into logarithm.

log_age <- log(Boston$age)
plot(log_age,log_crim,xlab = 'log(age)',ylab = 'log(crim)',main = "log(age) vs log(crim)")
```



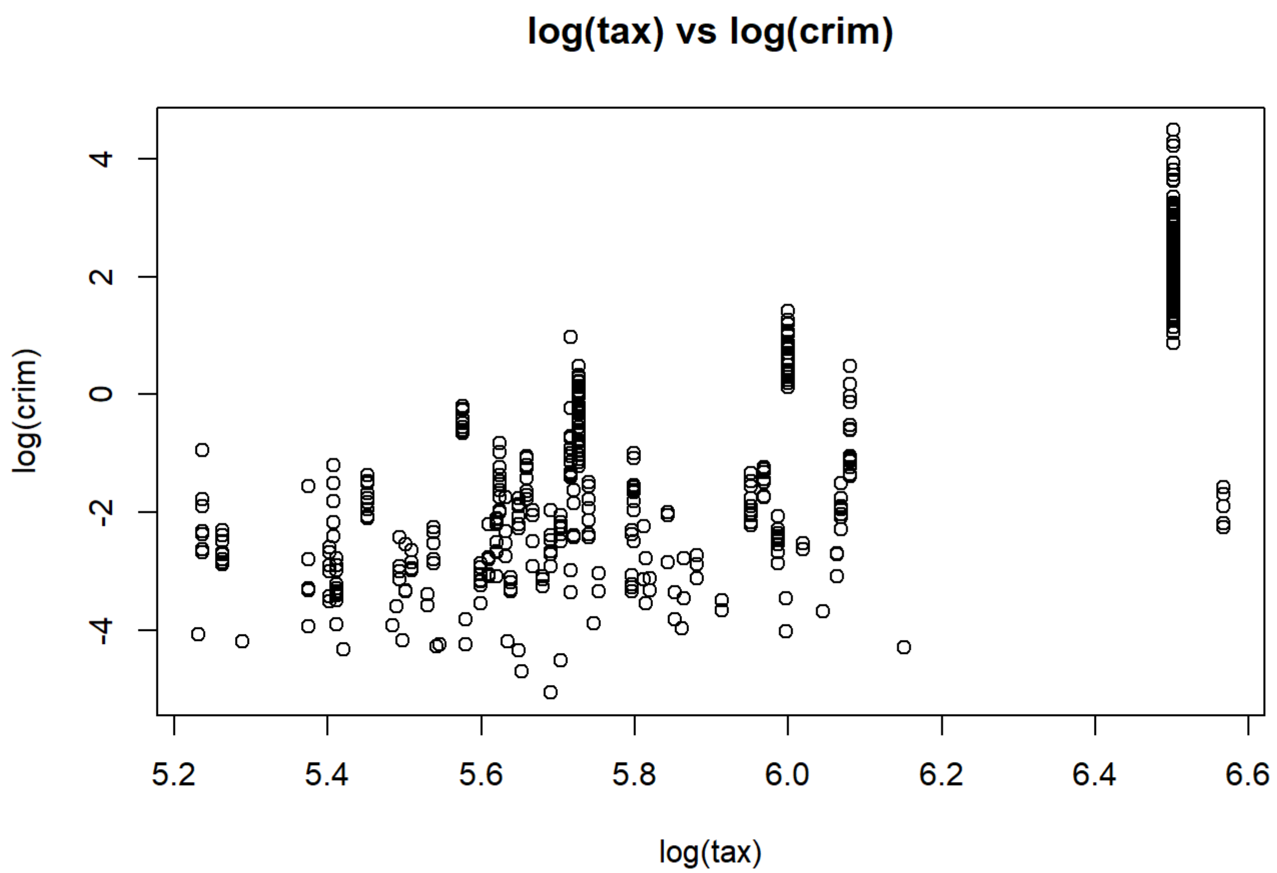

As shown in the above plot crim is positively correlated with age. Because if crime rate is increasing, age is also increasing.

```
## Transforming rad values into logarithm.

log_rad <- log(Boston$rad)
plot(log_rad,log_crim,xlab = 'log(rad)',ylab = 'log(crim)',main = "log(rad) vs log(crim)")
```



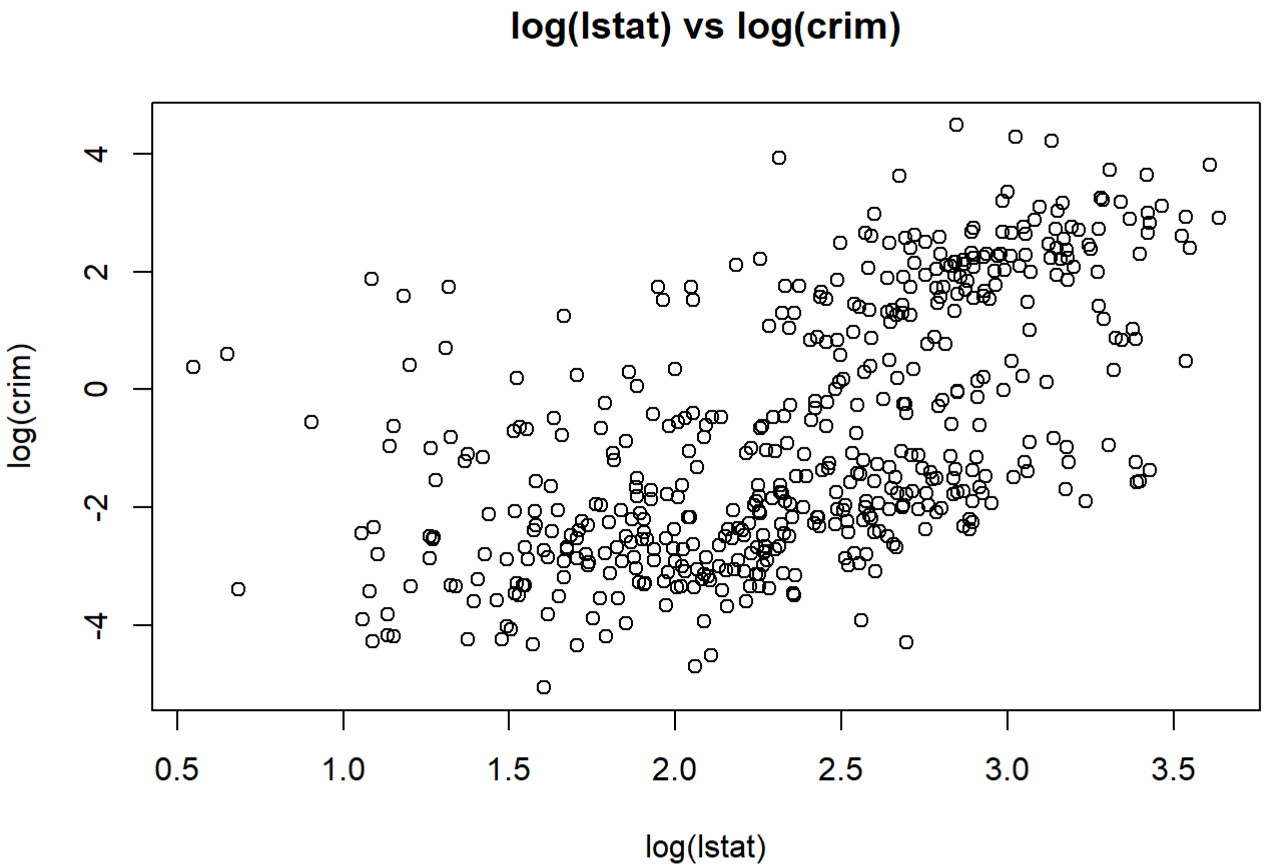

As shown in the above plot crim is positively correlated with rad. Because if crime rate is increasing, rad is also increasing.

```
## Transforming tax values into logarithm.

log_tax <- log(Boston$tax)
plot(log_tax,log_crim,xlab = 'log(tax)',ylab = 'log(crim)',main = "log(tax) vs log(crim)")
```



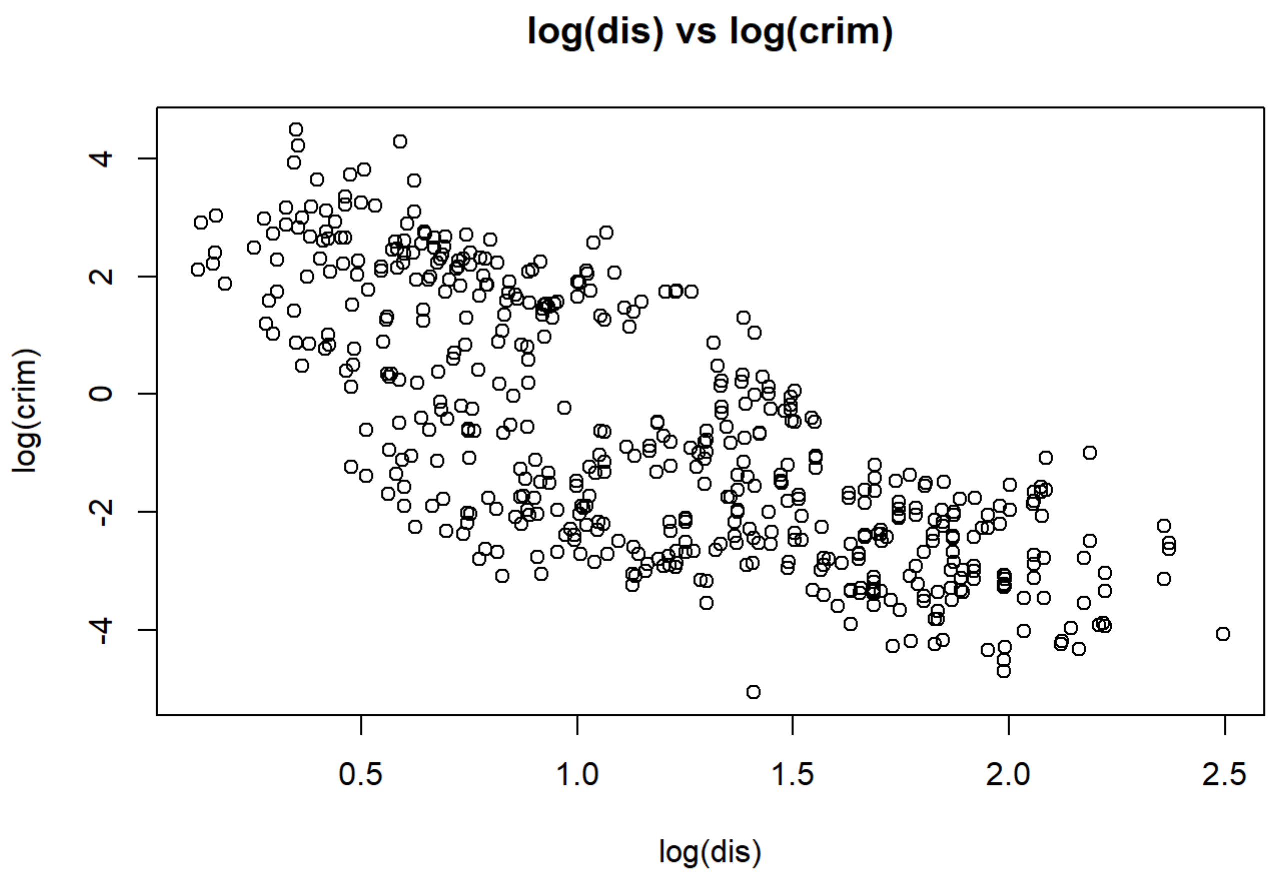

As shown in the above plot crim is positively correlated with tax. Because if crime rate is increasing, tax is also increasing.

```
## Transforming lstat values into logarithm.

log_lstat <- log(Boston$lstat)
plot(log_lstat,log_crim,xlab = 'log(lstat)',ylab = 'log(crim)',main = "log(lstat) vs log(cri
m)")
```



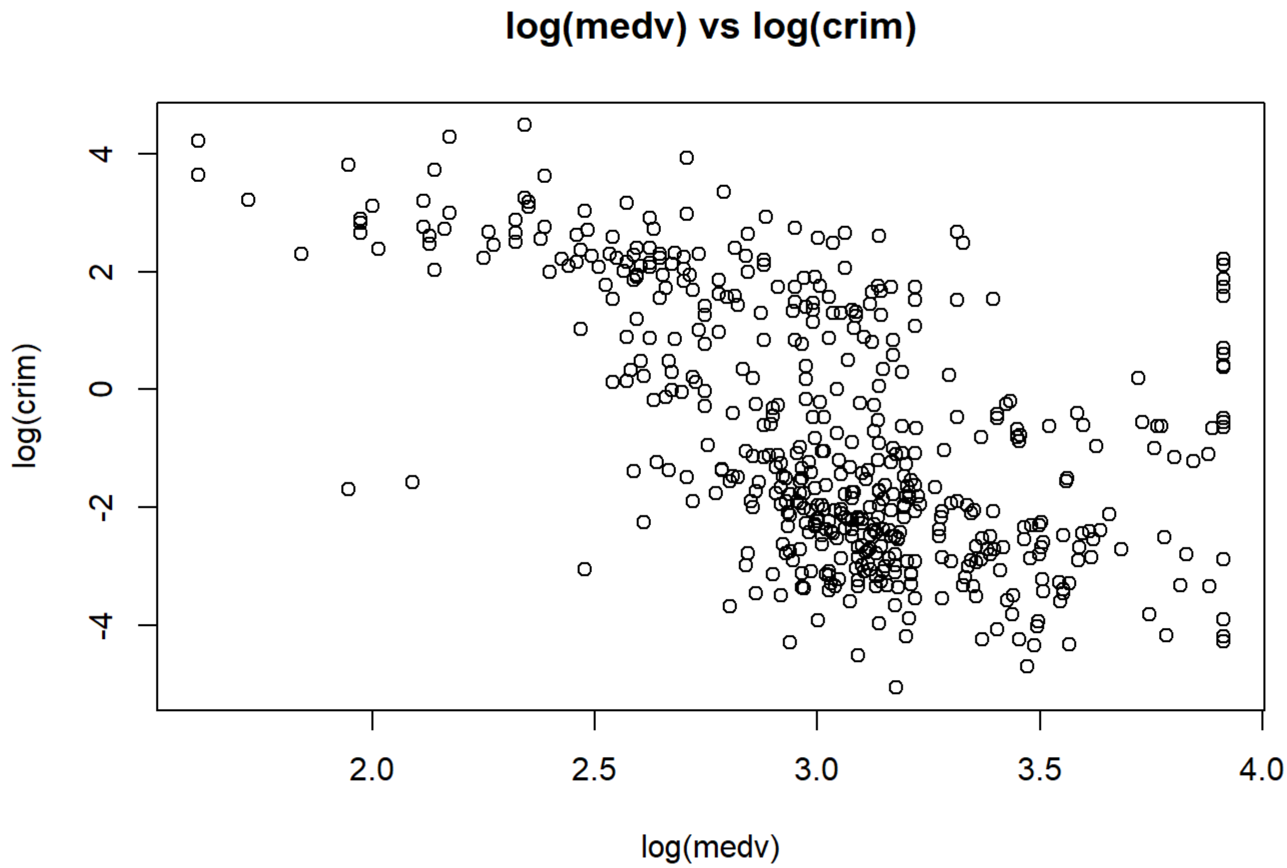

As shown in the above plot crim is positively correlated with lstat. Because if crime rate is increasing, lstat is also increasing.

```
## Transforming dis values into logarithm.

log_dis <- log(Boston$dis)
plot(log_dis,log_crim,xlab = 'log(dis)',ylab = 'log(crim)',main = "log(dis) vs log(crim)")
```




As shown in the above plot crim is negatively correlated with dis. Because if crime rate is increasing, dis is decreasing.

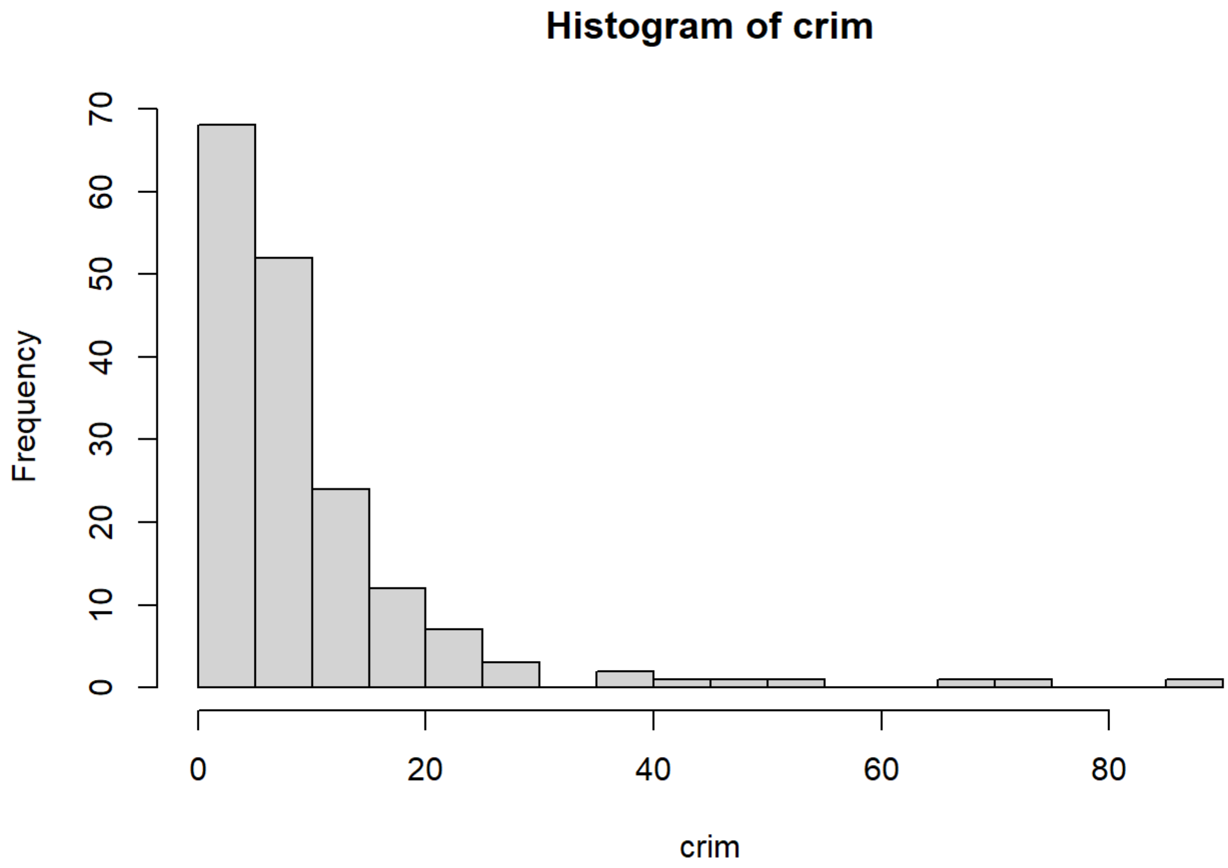
```
## Transforming medv values into logarithm.

log_medv <- log(Boston$medv)
plot(log_medv,log_crim,xlab = 'log(medv)',ylab = 'log(crim)',main = "log(medv) vs log(crim)")
```




As shown in the above plot crim is negatively correlated with medv. Because if crime rate is increasing, medv is decreasing.

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

```
hist(Boston$crim[Boston$crim>1], breaks=25, xlab = "crim", main = "Histogram of crim" )
```

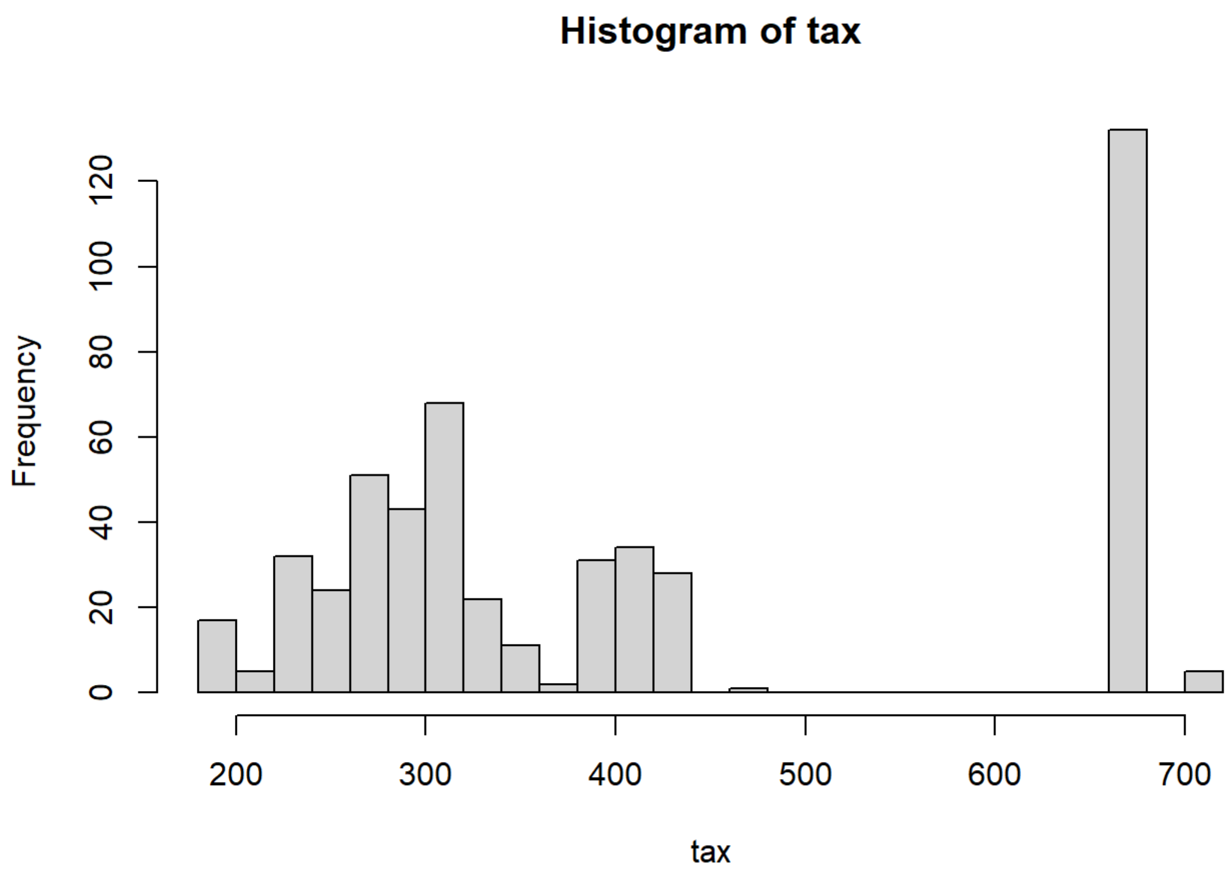



```
sum(Boston$crim>20)
```

```
## [1] 18
```

most suburbs have low crime rates, but there is 18 suburbs appear to have a crime rate greater than 20 reaching to above 80.

```
hist(Boston$tax, breaks=25,xlab = "tax", main = "Histogram of tax")
```

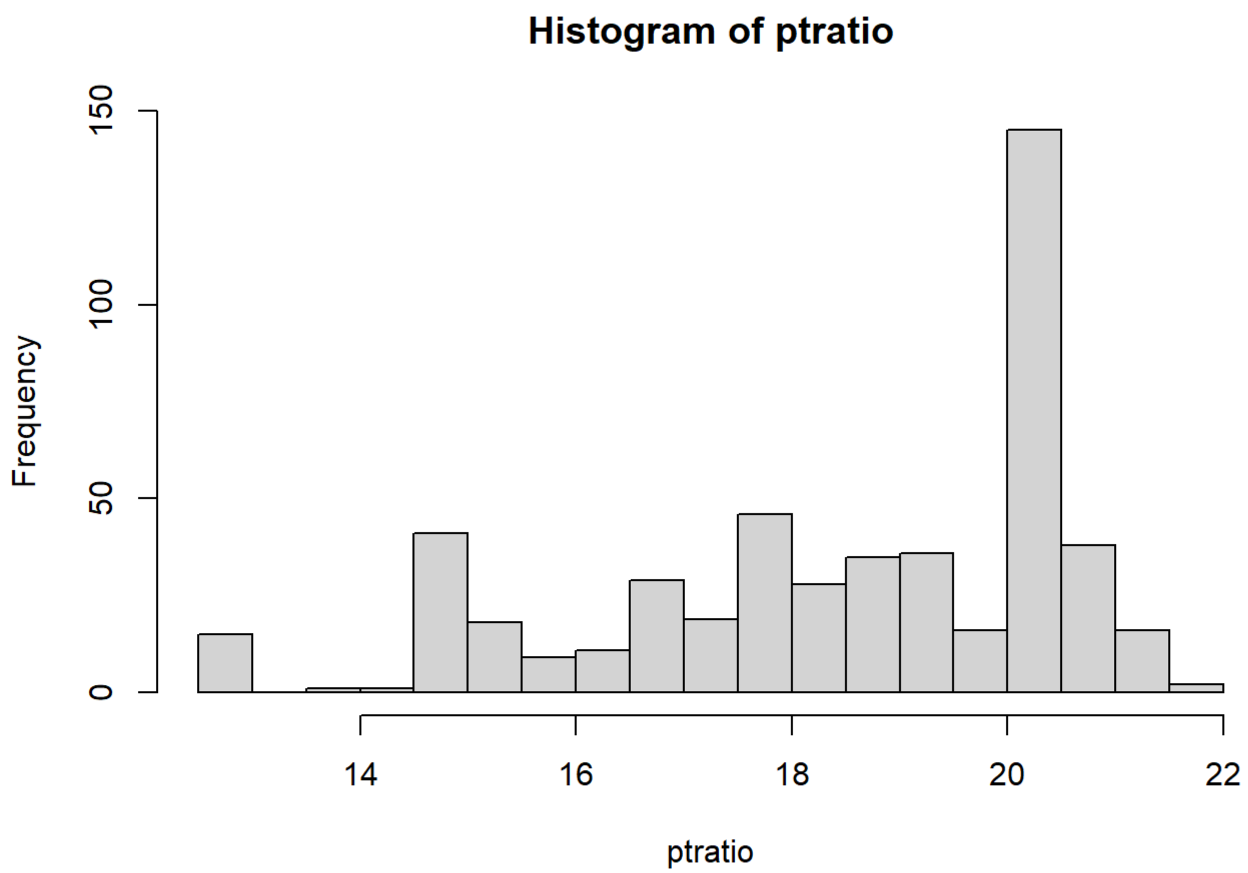



```
sum(Boston$tax > 660 & Boston$tax < 680)
```

```
## [1] 132
```

There is a large gap between suburbs with low tax rates and high tax rates, we can see a peak at 660 to 680(132 suburbs are in between 660 to 680).

```
hist(Boston$ptratio, breaks=25,xlab = "ptratio", main = "Histogram of ptratio")
```




```
sum(Boston$ptratio > 20)
```

```
## [1] 201
```

There is a peak towards high ratios and particularly greater than 20(201 suburbs have ptratio greater than 20).

## e) How many of the census tracts in this data set bound the Charles river?

```
sum(Boston$chas == 1)
```

```
## [1] 35
```

```
## OR

nrow(subset(Boston, chas == 1))
```

```
## [1] 35
```

35 census tracts in this data set bound the Charles river.

## f) What is the median pupil-teacher ratio among the towns in this data set?

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

19.05 is the median pupil-teacher ratio among the towns in this data set. Means 19 pupils for each teacher.

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

```
summary(Boston)
```

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          lstat      
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95  
##  Median : 5.000   Median :330.0   Median :19.05   Median :11.36  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97  
##       medv      
##  Min.   : 5.00  
##  1st Qu.:17.02  
##  Median :21.20  
##  Mean   :22.53  
##  3rd Qu.:25.00  
##  Max.   :50.00
```

```
t(subset(Boston[Boston$medv == min(Boston$medv), ]))
```

```
##              399      406
## crim     38.3518  67.9208
## zn        0.0000   0.0000
## indus    18.1000  18.1000
## chas      0.0000   0.0000
## nox       0.6930   0.6930
## rm        5.4530   5.6830
## age     100.0000 100.0000
## dis       1.4896   1.4254
## rad      24.0000  24.0000
## tax     666.0000 666.0000
## ptratio  20.2000  20.2000
## lstat    30.5900  22.9800
## medv      5.0000   5.0000
```

```
## OR

t(subset(Boston,medv == min(Boston$medv)))
```

```
##              399      406
## crim     38.3518  67.9208
## zn        0.0000   0.0000
## indus    18.1000  18.1000
## chas      0.0000   0.0000
## nox       0.6930   0.6930
## rm        5.4530   5.6830
## age     100.0000 100.0000
## dis       1.4896   1.4254
## rad      24.0000  24.0000
## tax     666.0000 666.0000
## ptratio  20.2000  20.2000
## lstat    30.5900  22.9800
## medv      5.0000   5.0000
```

Below are the comparision of predictors(columns) with overall range,

“crim” is above the 3rd quartile

“zn” is at minimum

“indus” is at 3rd quartile

“chas” not bounded by river

“nox” is above 3rd quartile

“rm” is below 1st quartile

“age” is at maximum

“dis” is below 1st quartile

“rad” is at maximum

“tax” is at 3rd quartile

“ptratio” is at 3rd quartile

“lstat” is above 3rd quartile

“medv” is at minimum

Because of the crime rate is above the 3rd quartile it is not the best place to live, but certainly not the worst.

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

```
sum(Boston$rm > 7)
```

```
## [1] 64
```

There are 64 census tracts that average more than seven rooms per dwelling.

```
sum(Boston$rm > 8)
```

```
## [1] 13
```

There are 13 census tracts that average more than eight rooms per dwelling.

```
summary(Boston)
```

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          lstat      
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95  
##  Median : 5.000   Median :330.0   Median :19.05   Median :11.36  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97  
##       medv      
##  Min.   : 5.00  
##  1st Qu.:17.02  
##  Median :21.20  
##  Mean   :22.53  
##  3rd Qu.:25.00  
##  Max.   :50.00
```

```
summary(subset(Boston[Boston$rm >8, ]))
```

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

While comparing with the overall range it has relatively lower “crime rate(crim)” and lower “lower status of the population percent(lstat)”.