HW1

Anish Mohan

January 25, 2016

1. Q1

- 1a.
  - Scenario: Is school rating more important than Real-Estate-Factors for the value of house value
    - Response: Sale value of a house
    - Predictors: # of bedrooms, bathrooms, size, location, rating of school, football program in school etc.
    - Goal: Inference because goal is to find if the rating of the school has more impact than other predictors
  - Scenario: Weather of a particular place in January-2017
    - Response: Temperature of a particular place
    - Predictors: Lat/Long of the place, time of day, day of the year, temperature previous year etc.
    - Goal: Prediction because we are trying to use the predictors to predict the weather in future.
  - Scenario: Sales and impact of Internet Advertising.
    - Response: Volume sales
    - Predictors: Advertising in various media channels: radio, newspaper, tv and internet.
    - Goal: Inference, as are trying to find if advertising on interet has more siginificant impact than other mediums.
- 1b.
  - Given certain measurements of fish samples, categorizing them into different species/classes. This is an unsupervised learning problem as no labels are given.
  - Given a bunch of pictures of different pose of different animals, catergorize the pictures. Since no labels are given, this would be an unsupervised learning.
  - Given writing samples of words from various languages, clustering the words that belong to same script together. Again, since the input is just a collection of written scripts without labels, this is an unsupervised learning problem.

- 1c.
  - Sound sample of words from various languages are given; classifying new sounds of words into their respective languages. Response= List of language, Predictors: Characterics of sound (tone, timber, pitch, frequency) of different words. The goal is prediction.
  - Certain fish samples are given with lables; Classifying a new fish into different species/classes based on sample measurements. Response= category of fish (e.g sea bass, tuna). Predictors= Length, breadth, height, weight and color of the fish. The goal of the application is prediction.
  - Pictures of various animals are given with labels. Classify new pictures into categories of animals. Response= Cateogory of animal. Predictors: Average color in picture, ratio of length/height, ratio of lenght of front legs/hind legs etc. The goal is prediction.

2. Q2

- 2a.
  - Unsupervised learning
  - n=42000
  - p=6 (age, high school GPA, SAT reading score, SAT Math Score, SAT writing score, Domestic/International)
- 2b.
  - classification
  - n=200
  - p=3 (age, zip code gender)
- 2c.
  - Regression problem
  - n=500
  - p=6 (population, state, avg income, crime rate, high school passed students, unemployment level)
- 2d.
  - Supervised learning problem as training sample and labels are provided.
  - Prediction because the neuroscientist in interested in predicting the cell types based on measurements.
  - n=48
  - p=3 (# of branch points, # of active processes, avg process length)

3. Q3

- 3a.
  - Advantages of less flexible
  - Works well if the true function is very simple (e.g linear)
  - Generally does not have the problem of overfitting.
  - Fewer number of parameters to learn hence does not need a large number of learning samples
  - Less Flexible method will better than flexible models when there are large number of predictors and few sample points as it will not overfit for fewere sample points
  - Less flexible methods are better when the data has lot of variance as they are less susceptible to noisy data.
  - Disadvantages of less flexible
  - Does not work well if the true function is not simple (e.g non-linear).
  - Cannot effectively use a large dataset to model complex functions.
- 3b. Non-Flexible approach is better than a flexible method
  - Problem with large number of predictors and few sample points
  - An inflexible model will generally find a better fit to combine the predictors to produce the results close to the few samples. With large number of predictors and few number of sample points, the flexible model will combine the predictors but will be constrainted to the small number of existing data points, thus overfitting the limited data.
- Given data has high variance and noise.
  - Given the high variance in noise, flexible models will tend to fit the error data and give poor results. Inflexible models will do a better job of ignoring the noise and finding a reasonable fit
- 3c. Flexible approach is better than a non-flexible method
  - Given dataset with the underlying function being highly non-linear and sufficient amount of data.
  - Inflexible models cannot generalize for non linear functions. Flexible models will have better ability to fit to non-linear models, hence they will generally perform better
  - Given dataset with small number of predictors and large number of sample.
  - Inflexible methods can only fit to specific combination (e.g linear combination) of the small number of predictors and cannot utlize the large number of samples to find a good fitting model. However, if the underlying function is linear, then the inflexible method will do well. Flexible learning methods will be able to utlize the large number of samples to find a reasonable fit for the true function with p predictors. However, there is a risk of over fitting the large number of input training points.

4. Q4

- 4a.

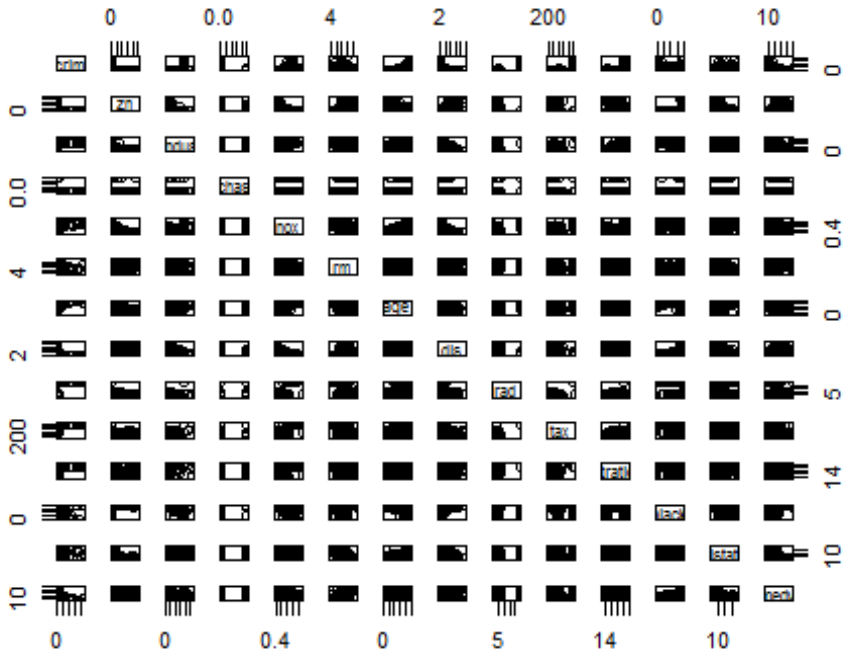
```
library(MASS)
dim(Boston)
```

```
## [1] 506  14
```

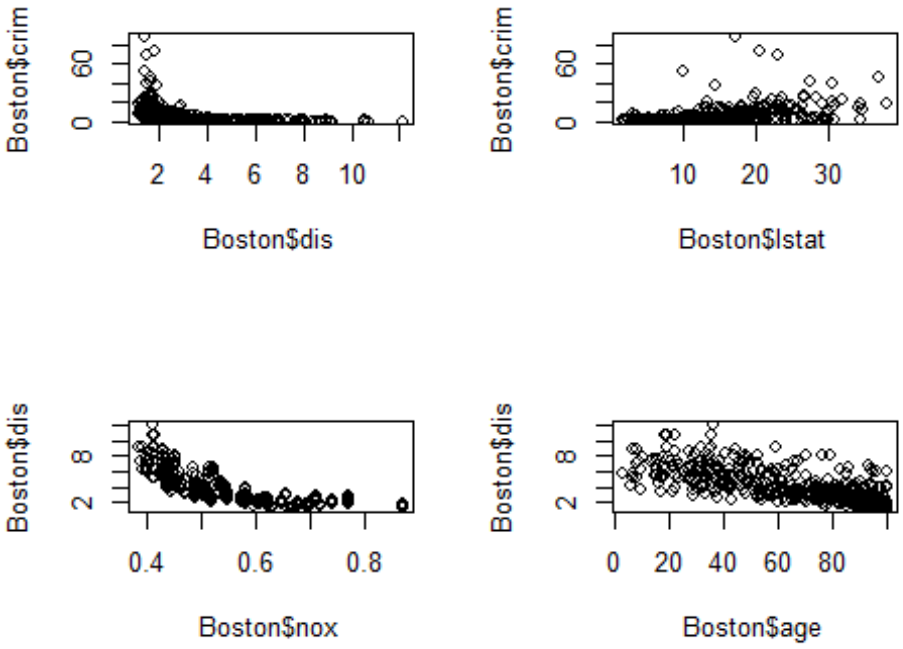
```
+ There are 506 rows and 14 columns. The rows represent the suburbs of Boston and the columns are various parameters like Tax rate, pupil-teacher ration in each town.
```

- 4b.

```
pairs(Boston)
```

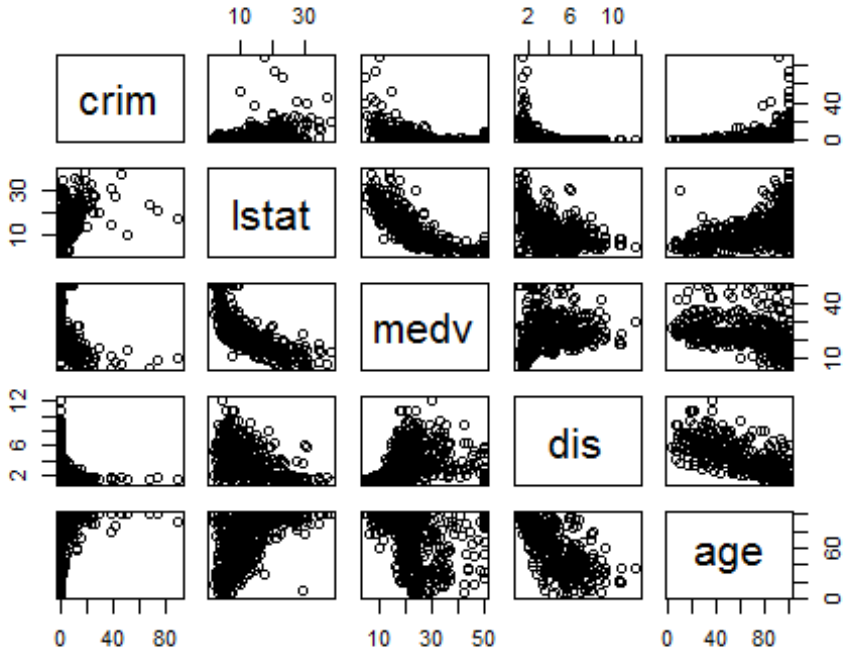



```
par(mfrow=c(2,2))
plot(Boston$dis, Boston$crim,)
plot(Boston$lstat,Boston$crim)
plot(Boston$nox,Boston$dis)
plot(Boston$age, Boston$dis)
```



- From the data we can see that crime-rate is influenced by many factors like average age of population in the suburbs/towns, the distance from the industrial employment center, tax rate, pupil-teacher ratio, median value of owner occupied homes etc. Here are some other examples
  - Crime rate is high in towns that have low median value of the owner occupied homes. Crime rate is very low in towns that have high median value of owner occupied homes
  - Proportion of zoned lots over 25K sqft are fairly high near the industrial employment centers
  - Nitrogen oxide concentration levels drop significantly as we move away from the Bostons 5 employment centers
- 4c.

```
pairs(~crim+lstat+medv+dis+age,Boston)
```



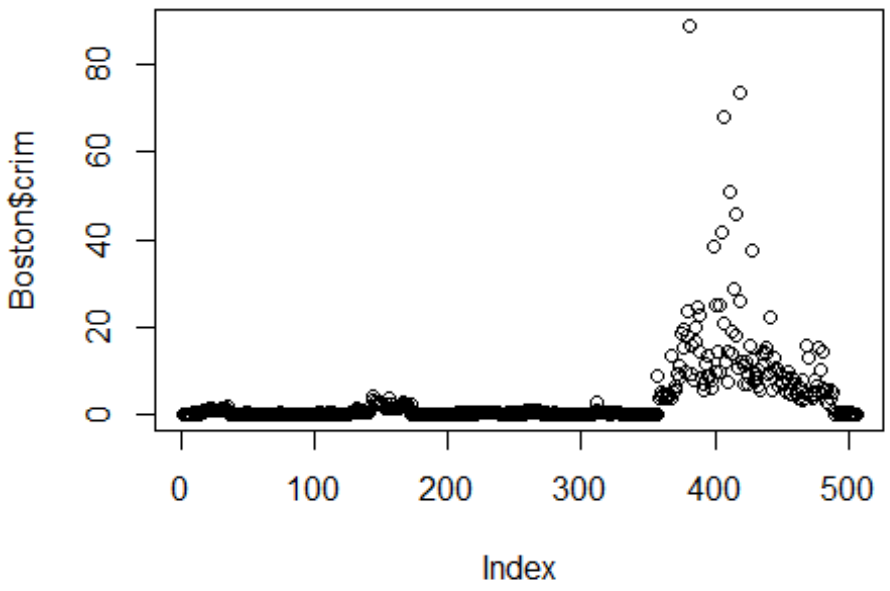

```
+ Some of predictors of the per capita crime rates are medv(median house valu
e), lstat(% of lower stat poputlation), dis (% disance away from employment c
enter).

+ Crime rate is very high in the areas with lower median value and decreases
exponentially in the areas where the median house values are larger. Inverse
relationship between crime rate and median value of house.

+ Crime rate is very high in the areas that are nearer to the employment cent
ers. Crime rate decreases as we move away from the employment areas. Again an
inverse relationship between crime rate and weighted distance to employment c
enter.

+ Crime rate is lower in areas where there smaller percentage of lower status
population. There is postive correlation between the percent of lower status
population and the crime rate.
```

- 4d

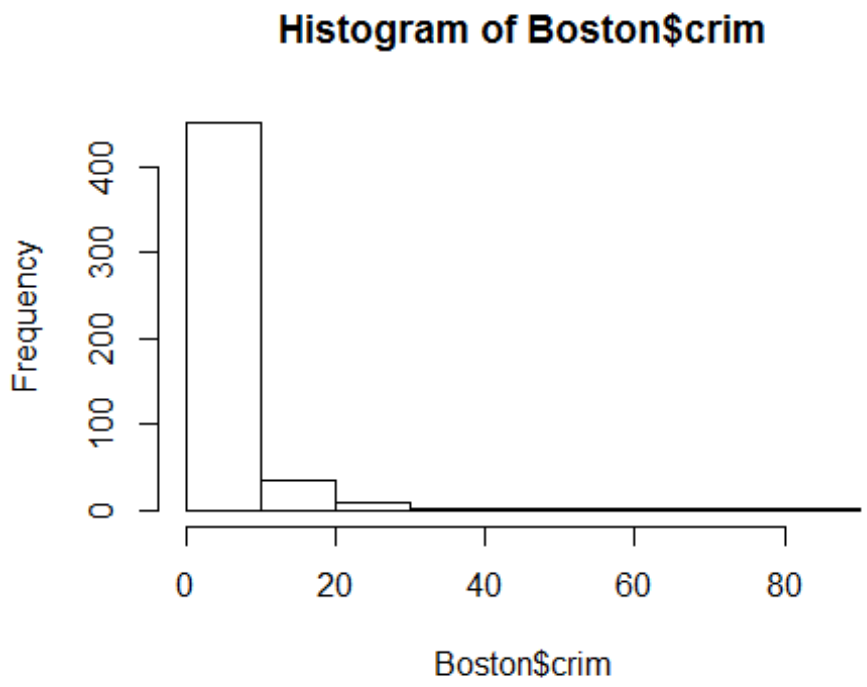
```
max(Boston$crim)

## [1] 88.9762

which.max(Boston$crim)

## [1] 381

summary(Boston$crim)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00632   0.08204   0.25650   3.61400   3.67700 88.98000
```

```
plot(Boston$crim)
```




```
hist(Boston$crim)
```




```
length(which(Boston$crim>10))
```

```
## [1] 54
```

```
+ Per capita crime rate is the highest at ~89 in suburb with index #381.
+ Per capita crime rate varies from 0.006 to 88.98, with median being 0.256 a
nd mean being 3.614
+ 54 suburbs have crime rate >10. Most of the suburbs have low per capita cri
me rate.
```

```
max(Boston$tax)
```

```
## [1] 711
```

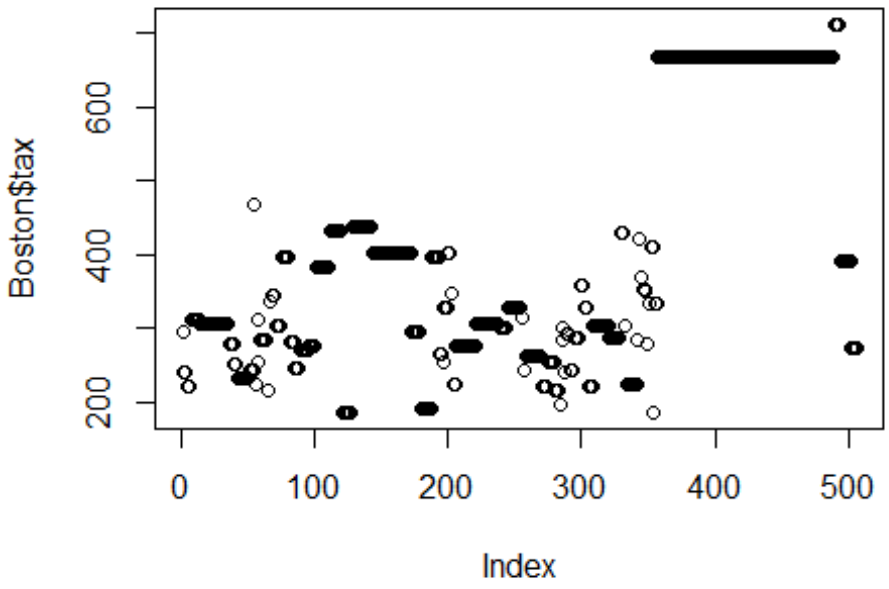
```
which.max(Boston$tax)
```

```
## [1] 489
```

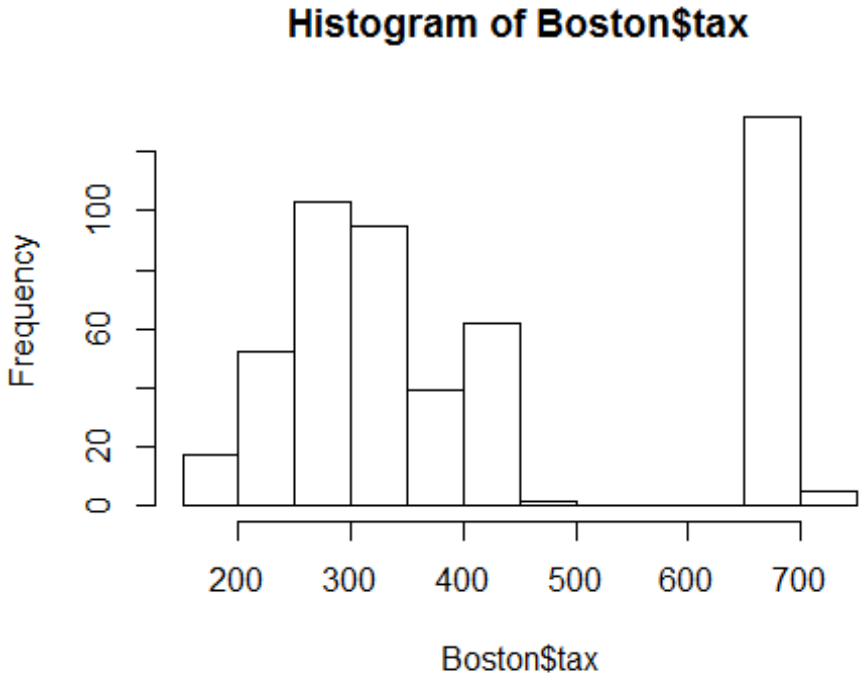
```
summary(Boston$tax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   279.0   330.0   408.2   666.0   711.0
```

```
plot(Boston$tax)
```




```r
hist(Boston$tax)
```




```r
length(which(Boston$tax>500))
```

```
## [1] 137
```

```
+ Max tax rate (per $10000) is $711 in the boston suburb with index 489
+ Tax rate varies from $187 to $711 (for per $10000)
+ 137 suburbs have a tax rate > $500 (per $10000)
```

```
max(Boston$ptratio)
```

```
## [1] 22
```

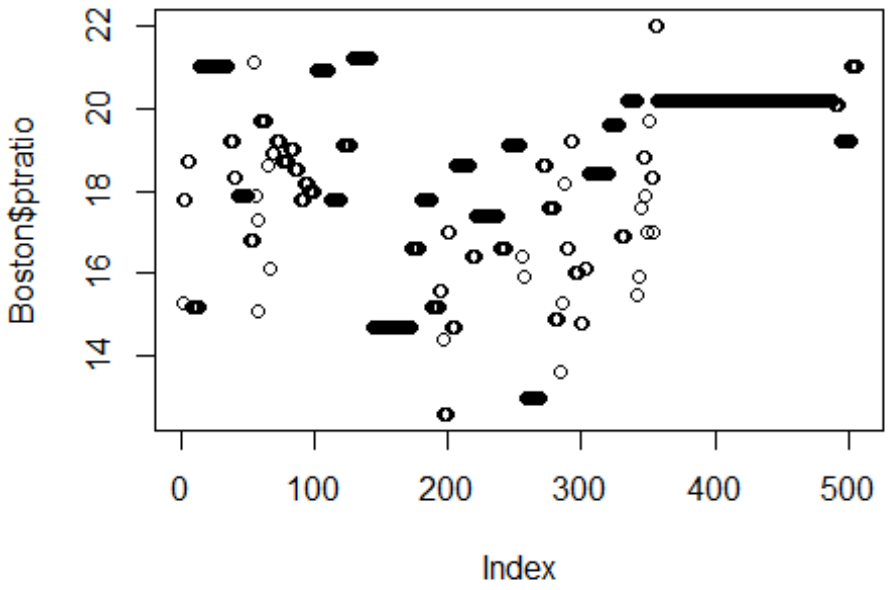
```
which.max(Boston$ptratio)
```

```
## [1] 355
```

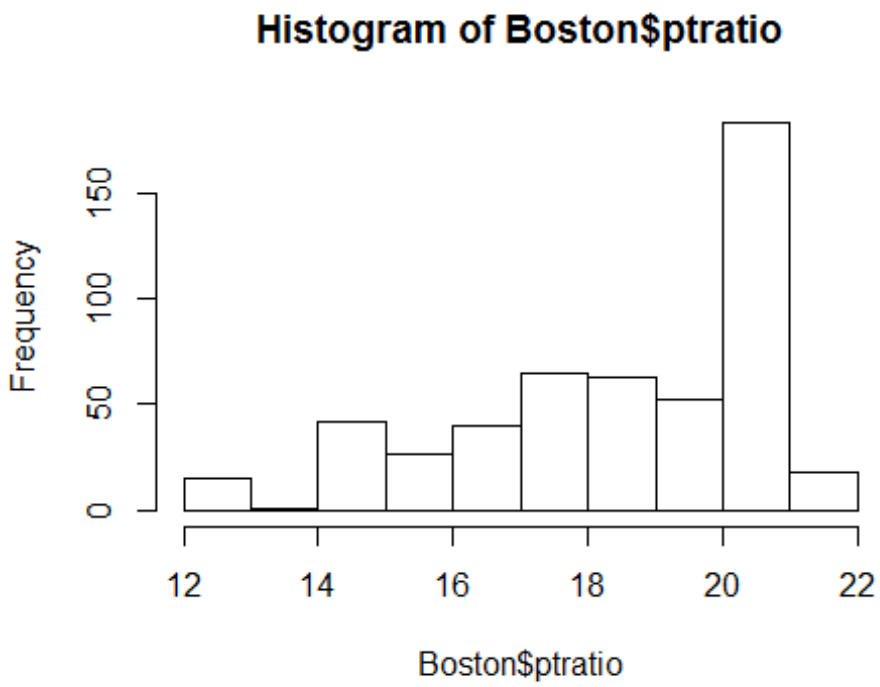
```
summary(Boston$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```

```
plot(Boston$ptratio)
```




```
hist(Boston$ptratio)
```




```
length(which(Boston$ptratio>20))
```

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

```
+ Maximum pupil/teacher ratio is 22 in suburb with index 355.
+ pupil teacher ratio varies from 12.60 to 22.00.
+ 201 suburbs have the pupil/teacher ration >20.
```

- 4e.

```
length(which(Boston$chas==1))
```

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

```
 +  There are 35 suburbs that are bound to Charles river
```

- 4f.

```
summary(Boston$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```

- Median ptratio in Boston data set is 19.05
- 4g.

```
minmedv=which.min(Boston$medv)
Boston[minmedv,]
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio black
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9
##     lstat medv
## 399 30.59    5
```

- Suburb with index 399 has the lowes median value of owner occupied homes of 5. Values of other predictors are shown above.
  - For this suburb, it is not bound to Charles river has and following charactersitics:
  - Following predictors are on the higher side(compared to median of Boston): Crime rate (above 3rd Quartile(Qu)), proportion of non-retail business acres per town (3rd Qu), nitrogen oxide concentration(above 3rd Qu), accebilitty to radial highways(3rd Qu), tax rate (3rd Qu), pupil-parent ratio (3rd Qu), % of black population (Maximum), % of lower status population (Above 3 u) and additionally all the owner occupied units were built before 1940 (maximum)
  - Following predictors are on the lower side (compared to median of Boston): No residential lots zone for over 25K (minmum=0), distance to boston's employment center (below 1st Qu), # of rooms per dwelling (below 1st Qu)
  - Overall it seems to be reasonable place to live with older building and easy access to job center.

- 4h.

```
length(which(Boston$rm>7))
```

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

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

```
##  [1]   3   5  41  56  65  89  90  98  99 100 162 163 164 167 181 183 187
## [18] 190 193 196 197 198 199 201 203 204 205 225 226 227 228 229 232 233
## [35] 234 238 254 257 258 259 261 262 263 264 265 267 268 269 274 277 281
## [52] 283 284 285 292 300 305 307 342 365 371 376 454 483
```

- 64 suburbs ahve more than 7 rooms per dwelling

```
length(which(Boston$rm>8))
```

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

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

```
##  [1]  98 164 205 225 226 227 233 234 254 258 263 268 365
```

- 13 suburbs have more than 8 rooms per dwelling.

```
rm8=which(Boston$rm>8)
total=Boston[98,]

for (i in rm8[2:13]){
```

```
  total=total+Boston[i,]
}

avg=total/13
```

- For the 13 suburbs that have on an avg more than 8 rooms per dwelling:
  - Greater proportion of lots zoned over 25K sqft (3 Qu)
  - Lower proportion of non-retail business acres (below 1st Qu)
  - Most of the tracts are away from Charles river(median)
  - Concentration of nitrogen oxide is comparable to other Boston suburbs (Median)
  - Crime rate is just above the median crime rate for Boston.
  - Proportion of owner occupied units built before 1940 is near the median
  - Distance to employment center are around the median.
  - Accesibility to high ways is bit above the median.
  - Tax rates are below the median
  - pupil/teacher is better and on the lower side (1 Qu)
  - Black population is near the median.
  - % of lower status folks are very low (1 Qu)
  - Median values of houses are on the higher side (3 Qu)
- Overall, quality of life factors seems to be better in teh 8 suburbs that have an on an average more than 8 rooms per dwelling.

5. Q5

- 5a.

```
# sampling the data to split into test and training samples
  set.seed(1)
  train=sample(506,253)
  test=-train
  train_sample=Boston[train,]
  test_sample=Boston[test,]
```

- 5b.

```
  # Running linear model on training samples
  lm.fit=lm(crim~.,data=train_sample)
  summary(lm.fit)

##
## Call:
## lm(formula = crim ~ ., data = train_sample)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.126 -2.416 -0.388  1.170 72.833
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.534874  11.096678   1.400   0.1628
## zn            0.049855   0.029678   1.680   0.0943 .
## indus        -0.022048   0.127855  -0.172   0.8632
## chas         -0.810878   1.750798  -0.463   0.6437
## nox         -15.894987   7.926449  -2.005   0.0461 *
## rm            1.470057   0.891093   1.650   0.1003
## age           0.006866   0.027854   0.246   0.8055
## dis          -1.307743   0.436109  -2.999   0.0030 **
## rad           0.650409   0.134286   4.843 2.29e-06 ***
## tax          -0.006347   0.007774  -0.816   0.4151
## ptratio      -0.398470   0.295614  -1.348   0.1790
## black        -0.002436   0.005403  -0.451   0.6524
## lstat         0.174022   0.122562   1.420   0.1569
## medv         -0.260635   0.095805  -2.720   0.0070 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.728 on 239 degrees of freedom
## Multiple R-squared:  0.4548, Adjusted R-squared:  0.4251
## F-statistic: 15.33 on 13 and 239 DF,  p-value: < 2.2e-16

  # calculating root mean square error for training data
  train_RMSE=sqrt(mean(residuals(lm.fit)^2))
  print(paste0("Training RMSE= ",train_RMSE))
```

```
## [1] "Training RMSE= 6.53968192857284"
```

```
  # calculating root mean square error for test data
  test_predict=predict(lm.fit,newdata = test_sample)
  test_error=test_predict-test_sample$crim
  test_RMSE=sqrt(mean(test_error^2))
  print(paste0("Test RMSE= ",test_RMSE))
```

```
## [1] "Test RMSE= 6.26705041014832"
```

- 5c.
  - R^2 value is 0.4548, which implies that regression with predictors here is only able to explain about 45% variability in the TSS. This indicates that the are other non-linear terms (e.g interaction terms) that needs to be considered to get a better fit for the model.
  - The most important predictors for crime rate based on low p values are distance to employment centers(dis), accessibility to highways(rad) and median values of home values (medv). These predictors have lower than 0.01 p value.

6. Q6

```r
# splitting data into training and test sample
library(MASS)
attach(Boston)
set.seed(1)
train=sample(506,253)
test=-train

train_sample=Boston[train,]
test_sample=Boston[test,]

#setting up response variable to be binary for training data
crime_train=rep(0,253)
crime_train[train_sample$crim>median(Boston$crim)]=1
train_sample$crim=crime_train

#setting up response variable to be binary for test data
crime_test=rep(0,253)
crime_test[test_sample$crim>median(Boston$crim)]=1
test_sample$crim=crime_test

#Running logistic regression on training data
glm.fit=glm(crim~.,data=train_sample, family=binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = crim ~ ., family = binomial, data = train_sample)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5914  -0.1747  -0.0028   0.0037   3.4997
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -44.492631  10.370769  -4.290 1.79e-05 ***
## zn           -0.044964   0.040574  -1.108   0.2678
## indus        -0.076522   0.068962  -1.110   0.2672
## chas          0.013430   1.047477   0.013   0.9898
## nox          52.722192  11.982655   4.400 1.08e-05 ***
## rm            0.339584   1.048227   0.324   0.7460
## age           0.047307   0.020520   2.305   0.0211 *
## dis           0.797836   0.336235   2.373   0.0177 *
## rad           0.522050   0.224384   2.327   0.0200 *
## tax          -0.007807   0.004375  -1.784   0.0744 .
## ptratio       0.324977   0.184932   1.757   0.0789 .
## black        -0.004292   0.005490  -0.782   0.4343
## lstat         0.068336   0.079535   0.859   0.3902
## medv          0.129651   0.098128   1.321   0.1864
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 350.729  on 252  degrees of freedom
## Residual deviance:  94.604  on 239  degrees of freedom
## AIC: 122.6
##
## Number of Fisher Scoring iterations: 9
```

```
      #Using the model produced by logistic regression to predict values for
training data
      glm.probs=predict(glm.fit,type="response")
      glm.pred=rep(0,253)
      glm.pred[glm.probs>0.5]=1

      # Measuring the number of errors in the training data prediction
      table(glm.pred,train_sample$crim)
```

```
##
## glm.pred   0   1
##        0 120  10
##        1   7 116
```

```
      mean(glm.pred==train_sample$crim)
```

```
## [1] 0.9328063
```

```
      #predicting the response for test data set
      glm.probs.test=predict(glm.fit,test_sample,type="response")
      glm.pred.test=rep(0,253)
      glm.pred.test[glm.probs.test>0.5]=1

      # Measuring the number of errors in the test data prediction
      table(glm.pred,test_sample$crim)
```

```
##
## glm.pred  0  1
##        0 62 68
##        1 64 59
```

```
      mean(glm.pred==test_sample$crim)
```

```
## [1] 0.4782609
```

- Important predictors: Most important predictors with low p value (<0.05) nitrogen oxide levels(nox), % of building built before 1940(age), distance to the employment center(dis) and accessibility to highways(rad)
- Training Error: Logistic regression does really well in classifying the training data. The classification rate is very high with ~93% of training data being classified correctly.

- Test Error: Logistic regression model does not fit really well for test data with the misclassification rate being high ~52% i.e only 48% of test data points were classified correctly.
- The logistic regression model does very well with the training data but does poorly with the test data. This implies the logistic regression model was over fit to the training data and hence performed poorly to the training data. The linear regression model has a consistent error rate in the test and the training data, thus proving to be a consistent fit for training and test data.