STAT 702 - ASSIGNMENT-1

Sumukh Sagar Manjunath

February 9, 2016

```
> library(MASS)

> head(Boston)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7

> names(Boston)
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"
 [9] "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

1.a Size and Class of Dataset.

```
> size = c("rows"=nrow(Boston),"cols"=ncol(Boston))
> size
rows cols
 506   14
> sapply(Boston,class)
     crim        zn     indus      chas       nox        rm       age       dis
"numeric" "numeric" "numeric" "integer" "numeric" "numeric" "numeric" "numeric"
      rad       tax   ptratio     black     lstat      medv
"integer" "numeric" "numeric" "numeric" "numeric" "numeric"
```

1.b Relationship between the predictors in the data set and per-capita crime rate.

```
> cor(Boston[2:length(Boston)],Boston$crim,method="pearson")
               [,1]
zn      -0.20046922
indus    0.40658341
chas    -0.05589158
nox      0.42097171
rm      -0.21924670
age      0.35273425
dis     -0.37967009
rad      0.62550515
tax      0.58276431
ptratio  0.28994558
black   -0.38506394
lstat    0.45562148
medv    -0.38830461
```

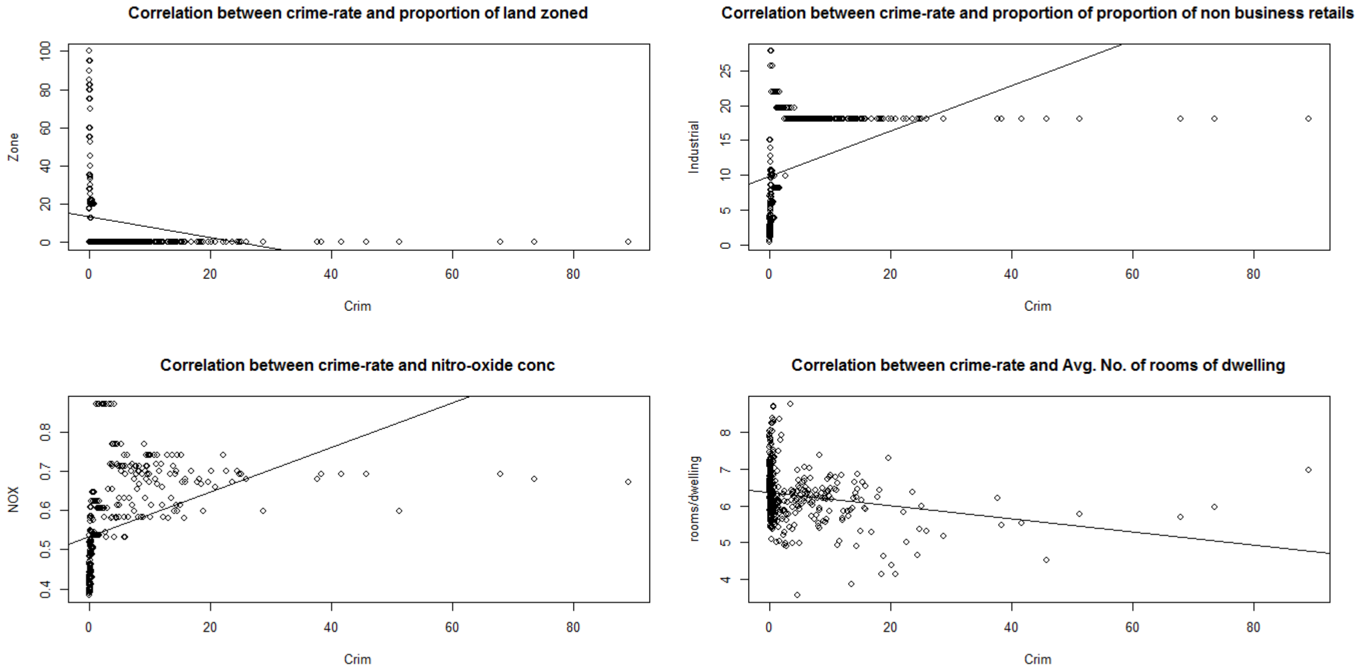
```
> # Function to retreive significant correlations
crim_cor = function()
{
  sig_cnt = 1
  res_corr = c()
  p_val = c()
  ind_var_name = c()
  assoc = c()
  ind_var = Boston[,names(Boston)!="crim"]
  for(i in 1:length(Boston))
  {
    if(cor.test(Boston[[i]],Boston$crim,method="pearson")$p.value < 0.05 && names(Boston[i])!="crim")
    {
      res_corr[sig_cnt] = cor(Boston[[i]],Boston$crim,method="pearson")
      ind_var_name[sig_cnt] = names(Boston[i])
      p_val[sig_cnt] = cor.test(Boston[[i]],Boston$crim,method="pearson")$p.value
      if(res_corr[[sig_cnt]]<0)
      {
        assoc = c(assoc,"neg")
      }
      else
      {
        assoc = c(assoc,"pos")
      }
      sig_cnt = sig_cnt + 1
    }
  }
  names(res_corr) = ind_var_name
  res_corr = rbind(res_corr,p_val)
  res_corr = rbind(res_corr,assoc)
  return (res_corr)
  # print(res_corr)
}

> a = crim_cor()
> a
         zn                    indus               nox                 rm                    age                   dis
res_corr "-0.200469219662547"  "0.406583411406259" "0.420971711392456" "-0.219246702862514"  "0.352734250901364"   "-0.379670086951024"
p_val    "5.50647210767929e-06" "0"                "0"                 "6.34670298468773e-07" "4.44089209850063e-16" "8.51994876692635e-19"
assoc    "neg"                 "pos"               "pos"               "neg"                 "pos"                 "neg"
         rad                 tax                 ptratio               black                 lstat               medv
res_corr "0.625505145262602" "0.582764312032585" "0.28994557927952"    "-0.385063941994224"  "0.455621479447946" "-0.388304608586812"
p_val    "0"                 "0"                 "2.94293478475538e-11" "2.4872739737737e-19" "0"                 "1.17398708219436e-19"
assoc    "pos"               "pos"               "pos"                 "neg"                 "pos"               "neg"
```

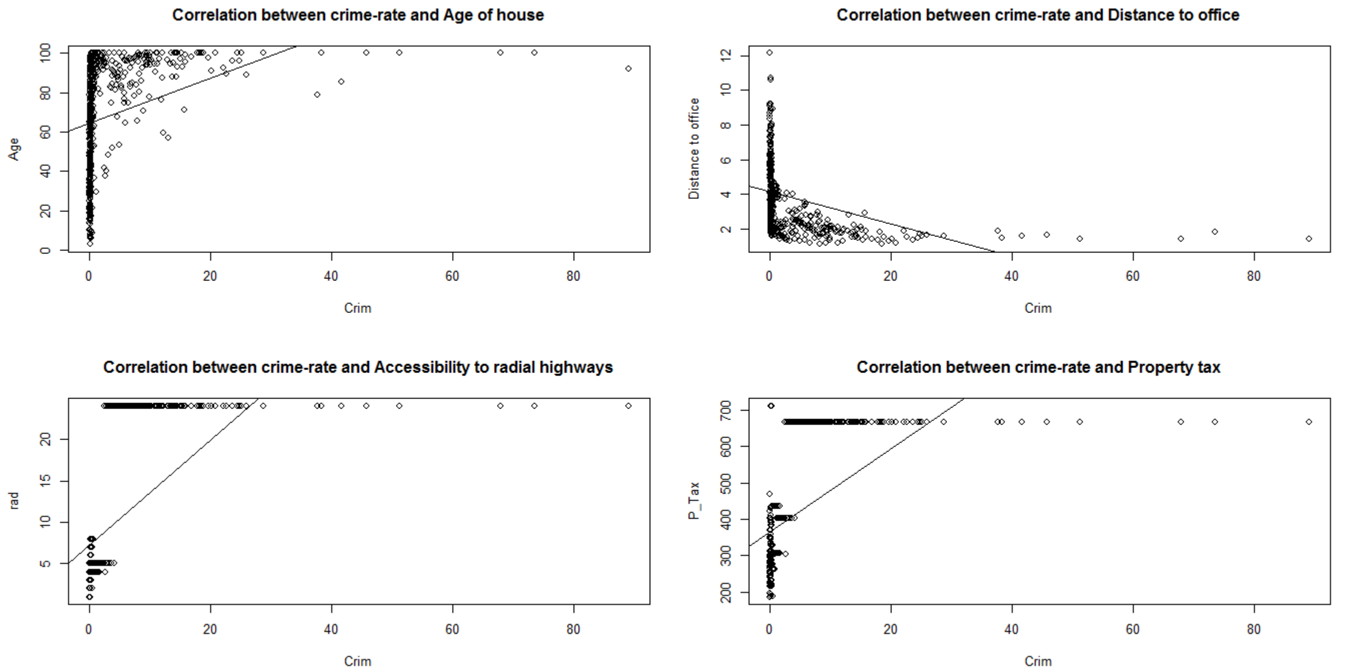
Here the Variable 'a' stores the the predictor's r-value, p-value and their associations row wise.

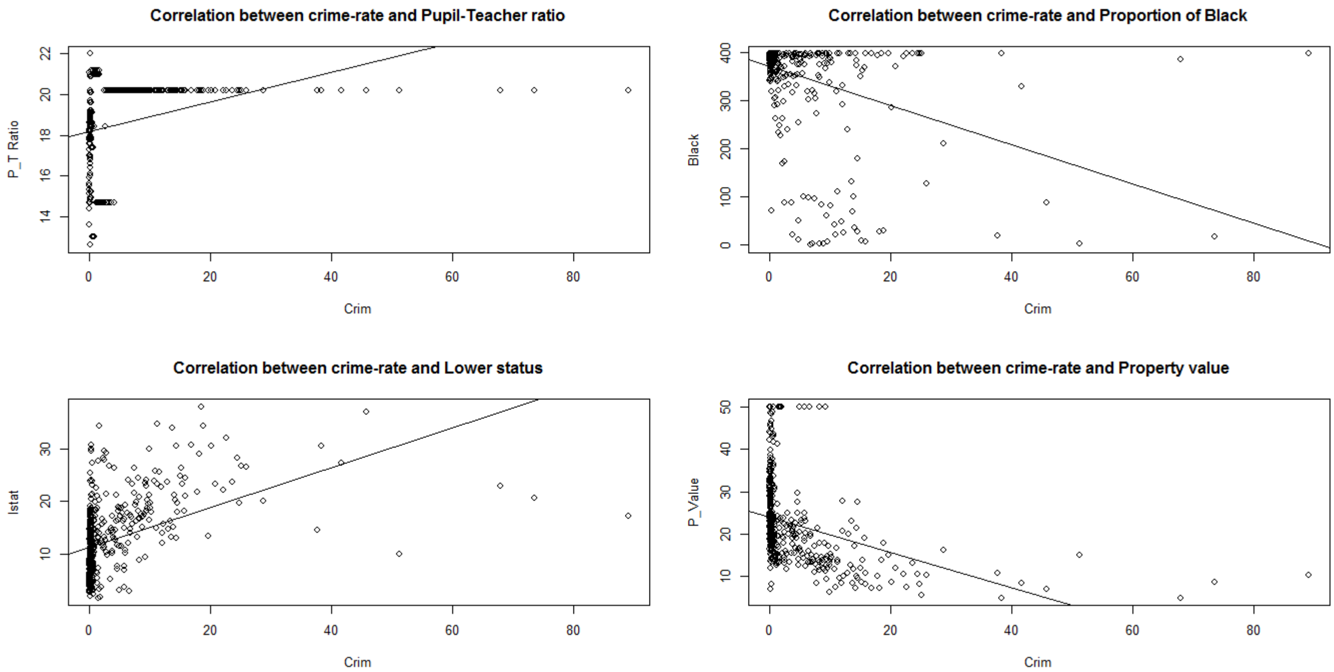
1.c X-Y Plots for the significant correlations.

```
par(mfrow=c(2,2))
plot(Boston$crim,Boston$zn,main="Correlation between crime-rate and proportion of land zoned",xlab="Crim", ylab="Zone")
abline(lm(Boston$zn ~ Boston$crim))
plot(Boston$crim,Boston$indus,main="Correlation between crime-rate and proportion of non business retails",xlab="Crim", ylab="Industrial")
abline(lm(Boston$indus ~ Boston$crim))
plot(Boston$crim,Boston$nox,main="Correlation between crime-rate and nitro-oxide conc",xlab="Crim", ylab="NOX")
abline(lm(Boston$nox ~ Boston$crim))
plot(Boston$crim,Boston$rm,main="Correlation between crime-rate and Avg. No. of rooms of dwelling",xlab="Crim", ylab="rooms/dwelling")
abline(lm(Boston$rm ~ Boston$crim))

par(mfrow=c(2,2))
plot(Boston$crim,Boston$age,main="Correlation between crime-rate and Age of house",xlab="Crim", ylab="Age")
abline(lm(Boston$age ~ Boston$crim))
plot(Boston$crim,Boston$dis,main="Correlation between crime-rate and Distance to office",xlab="Crim", ylab="Distance to office")
abline(lm(Boston$dis ~ Boston$crim))
plot(Boston$crim,Boston$rad,main="Correlation between crime-rate and Accessibility to radial highways",xlab="Crim", ylab="rad")
abline(lm(Boston$rad ~ Boston$crim))
plot(Boston$crim,Boston$tax,main="Correlation between crime-rate and Property tax",xlab="Crim", ylab="P_Tax")
abline(lm(Boston$tax ~ Boston$crim))
```

```
par(mfrow=c(2,2))
plot(Boston$crim,Boston$ptratio,main="Correlation between crime-rate and Pupil-Teacher ratio",xlab="Crim", ylab="P_T Ratio")
abline(lm(Boston$ptratio ~ Boston$crim))
plot(Boston$crim,Boston$black,main="Correlation between crime-rate and Proportion of Black",xlab="Crim", ylab="Black")
abline(lm(Boston$black ~ Boston$crim))
plot(Boston$crim,Boston$lstat,main="Correlation between crime-rate and Lower status",xlab="Crim", ylab="lstat")
abline(lm(Boston$lstat ~ Boston$crim))
plot(Boston$crim,Boston$medv,main="Correlation between crime-rate and Property value",xlab="Crim", ylab="P_Value")
abline(lm(Boston$medv ~ Boston$crim))
```







.

1.d Number of suburbs in this data set which bound the Charles River.

```
> with(Boston, subset(Boston, Boston$chas==1 ))
       crim zn indus chas    nox    rm   age    dis rad tax ptratio  black lstat medv
143 3.32105  0 19.58    1 0.8710 5.403 100.0 1.3216   5 403    14.7 396.90 26.82 13.4
153 1.12658  0 19.58    1 0.8710 5.012  88.0 1.6102   5 403    14.7 343.28 12.12 15.3
155 1.41385  0 19.58    1 0.8710 6.129  96.0 1.7494   5 403    14.7 321.02 15.12 17.0
156 3.53501  0 19.58    1 0.8710 6.152  82.6 1.7455   5 403    14.7  88.01 15.02 15.6
161 1.27346  0 19.58    1 0.6050 6.250  92.6 1.7984   5 403    14.7 338.92  5.50 27.0
163 1.83377  0 19.58    1 0.6050 7.802  98.2 2.0407   5 403    14.7 389.61  1.92 50.0
164 1.51902  0 19.58    1 0.6050 8.375  93.9 2.1620   5 403    14.7 388.45  3.32 50.0
209 0.13587  0 10.59    1 0.4890 6.064  59.1 4.2392   4 277    18.6 381.32 14.66 24.4
210 0.43571  0 10.59    1 0.4890 5.344 100.0 3.8750   4 277    18.6 396.90 23.09 20.0
211 0.17446  0 10.59    1 0.4890 5.960  92.1 3.8771   4 277    18.6 393.25 17.27 21.7
212 0.37578  0 10.59    1 0.4890 5.404  88.6 3.6650   4 277    18.6 395.24 23.98 19.3
213 0.21719  0 10.59    1 0.4890 5.807  53.8 3.6526   4 277    18.6 390.94 16.03 22.4
217 0.04560  0 13.89    1 0.5500 5.888  56.0 3.1121   5 276    16.4 392.80 13.51 23.3
219 0.11069  0 13.89    1 0.5500 5.951  93.8 2.8893   5 276    16.4 396.90 17.92 21.5
220 0.11425  0 13.89    1 0.5500 6.373  92.4 3.3633   5 276    16.4 393.74 10.50 23.0
221 0.35809  0  6.20    1 0.5070 6.951  88.5 2.8617   8 307    17.4 391.70  9.71 26.7
222 0.40771  0  6.20    1 0.5070 6.164  91.3 3.0480   8 307    17.4 395.24 21.46 21.7
223 0.62356  0  6.20    1 0.5070 6.879  77.7 3.2721   8 307    17.4 390.39  9.93 27.5
235 0.44791  0  6.20    1 0.5070 6.726  66.5 3.6519   8 307    17.4 360.20  8.05 29.0
237 0.52058  0  6.20    1 0.5070 6.631  76.5 4.1480   8 307    17.4 388.45  9.54 25.1
270 0.09065 20  6.96    1 0.4640 5.920  61.5 3.9175   3 223    18.6 391.34 13.65 20.7
274 0.22188 20  6.96    1 0.4640 7.691  51.8 4.3665   3 223    18.6 390.77  6.58 35.2
275 0.05644 40  6.41    1 0.4470 6.758  32.9 4.0776   4 254    17.6 396.90  3.53 32.4
277 0.10469 40  6.41    1 0.4470 7.267  49.0 4.7872   4 254    17.6 389.25  6.05 33.2
278 0.06127 40  6.41    1 0.4470 6.826  27.6 4.8628   4 254    17.6 393.45  4.16 33.1
283 0.06129 20  3.33    1 0.4429 7.645  49.7 5.2119   5 216    14.9 377.07  3.01 46.0
284 0.01501 90  1.21    1 0.4010 7.923  24.8 5.8850   1 198    13.6 395.52  3.16 50.0
357 8.98296  0 18.10    1 0.7700 6.212  97.4 2.1222  24 666    20.2 377.73 17.60 17.8
358 3.84970  0 18.10    1 0.7700 6.395  91.0 2.5052  24 666    20.2 391.34 13.27 21.7
359 5.20177  0 18.10    1 0.7700 6.127  83.4 2.7227  24 666    20.2 395.43 11.48 22.7
364 4.22239  0 18.10    1 0.7700 5.803  89.0 1.9047  24 666    20.2 353.04 14.64 16.8
365 3.47428  0 18.10    1 0.7180 8.780  82.9 1.9047  24 666    20.2 354.55  5.29 21.9
370 5.66998  0 18.10    1 0.6310 6.683  96.8 1.3567  24 666    20.2 375.33  3.73 50.0
371 6.53876  0 18.10    1 0.6310 7.016  97.5 1.2024  24 666    20.2 392.05  2.96 50.0
373 8.26725  0 18.10    1 0.6680 5.875  89.6 1.1296  24 666    20.2 347.88  8.88 50.0
> nrow(with(Boston, subset(Boston, Boston$chas==1 )))
[1] 35
```

.

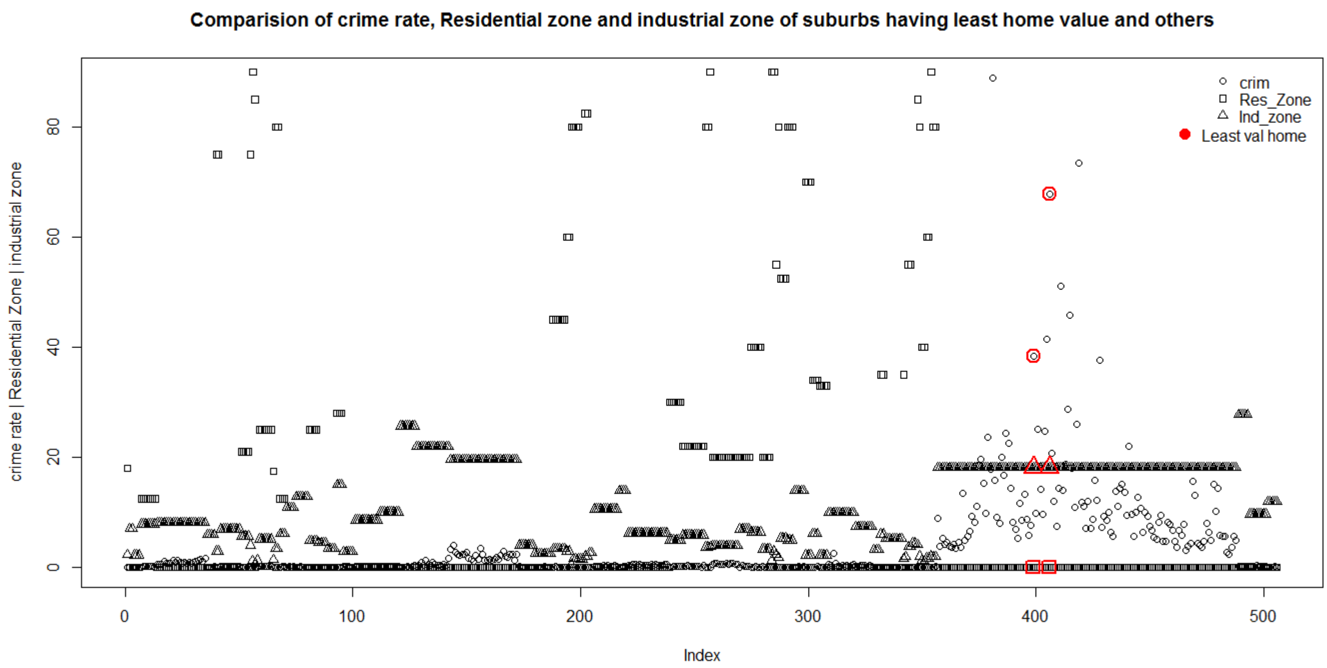
1.e Median pupil-teacher ratio among the towns in this data set.

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

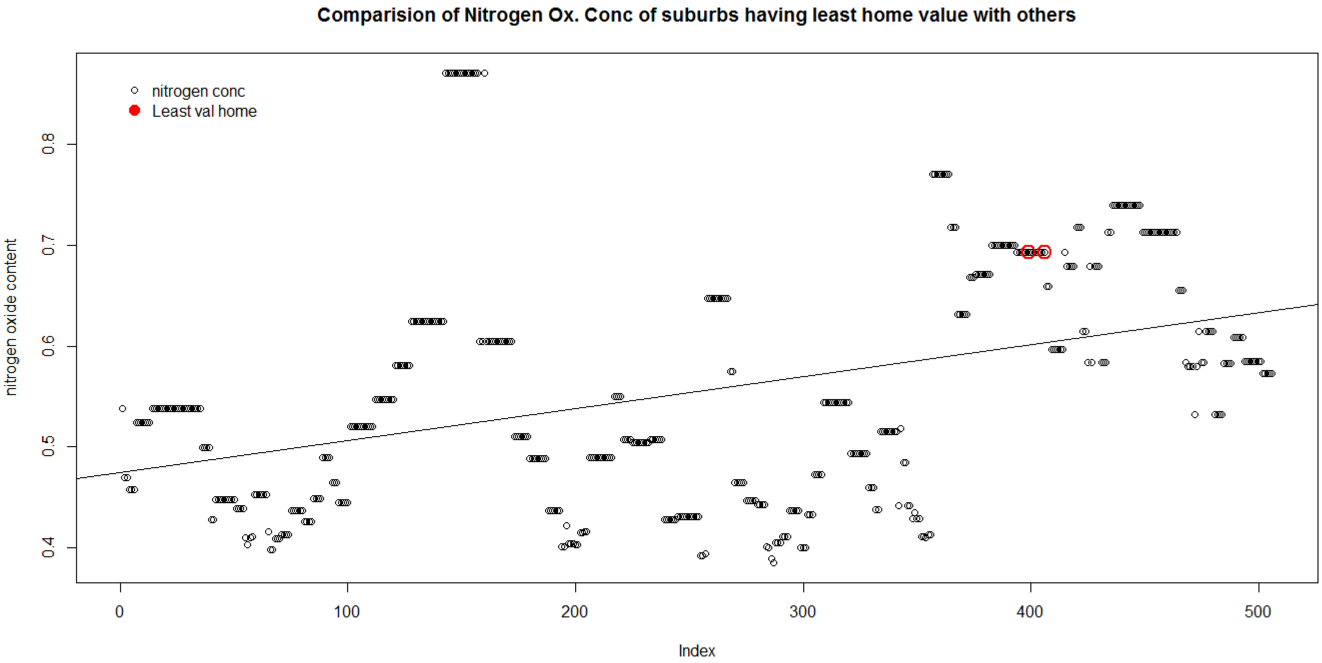
.

1.f Lowest median value of owner occupied homes and comparison of other predictors from this group with rest of the sample.

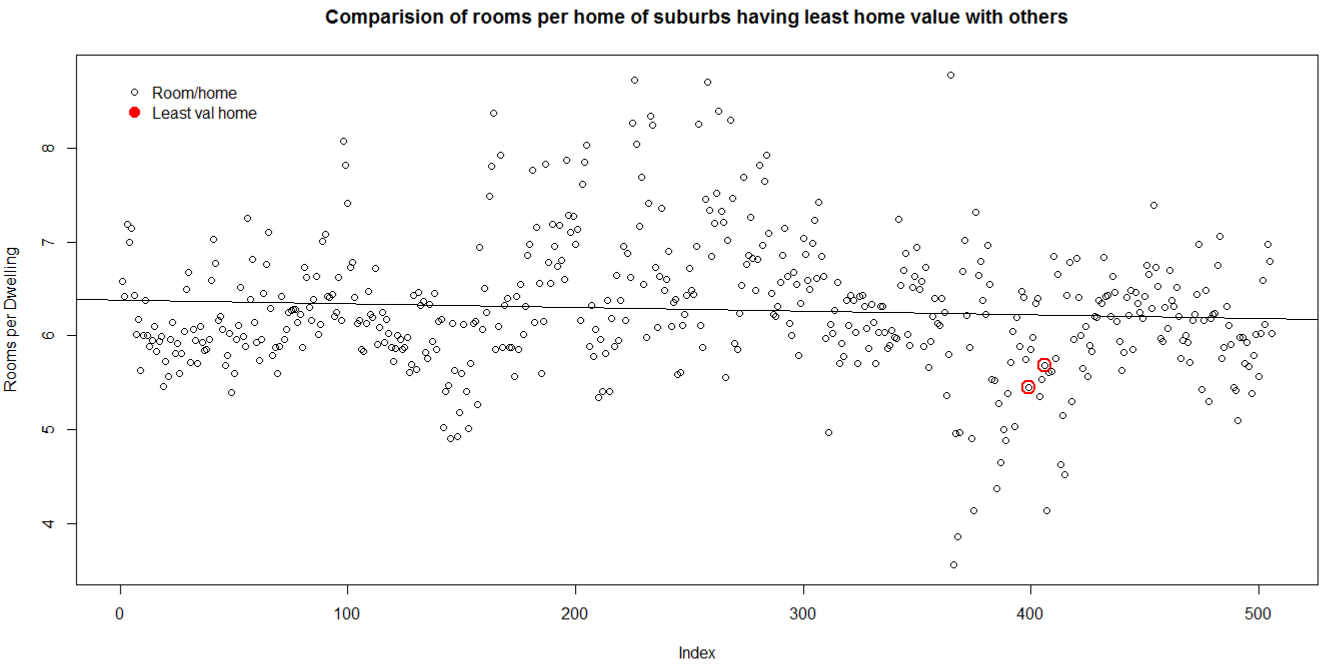
```
> Boston[Boston$medv == min(as.numeric(Boston$medv)),]
       crim zn indus chas   nox    rm age    dis rad tax ptratio  black lstat medv
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90 30.59    5
406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98    5

> plot(Boston$crim,ylab="crime rate | Residential Zone | industrial zone")
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$crim[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2)
> points(Boston$zn,lwd=1,cex=1,pch=22)
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$zn[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2,pch=22)
> points(Boston$indus,lwd=1,cex=1,pch=24)
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$indus[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2,pch=24)
> legend(x=630,y=95,c("crim","Res_Zone","Ind_zone"),pch=c(21,22,24),cex=1,pt.cex = 1,xjust=1,bty="n",x.intersp=0.3,y.intersp=0.3)
> legend(x=440,y=89,c("Least val home"),pch=19,col="red",cex=1,bty='n',x.intersp=0.3,pt.cex = 1.5)
> title("Comparision of crime rate, Residential zone and industrial zone of suburbs having least home value with others")
```



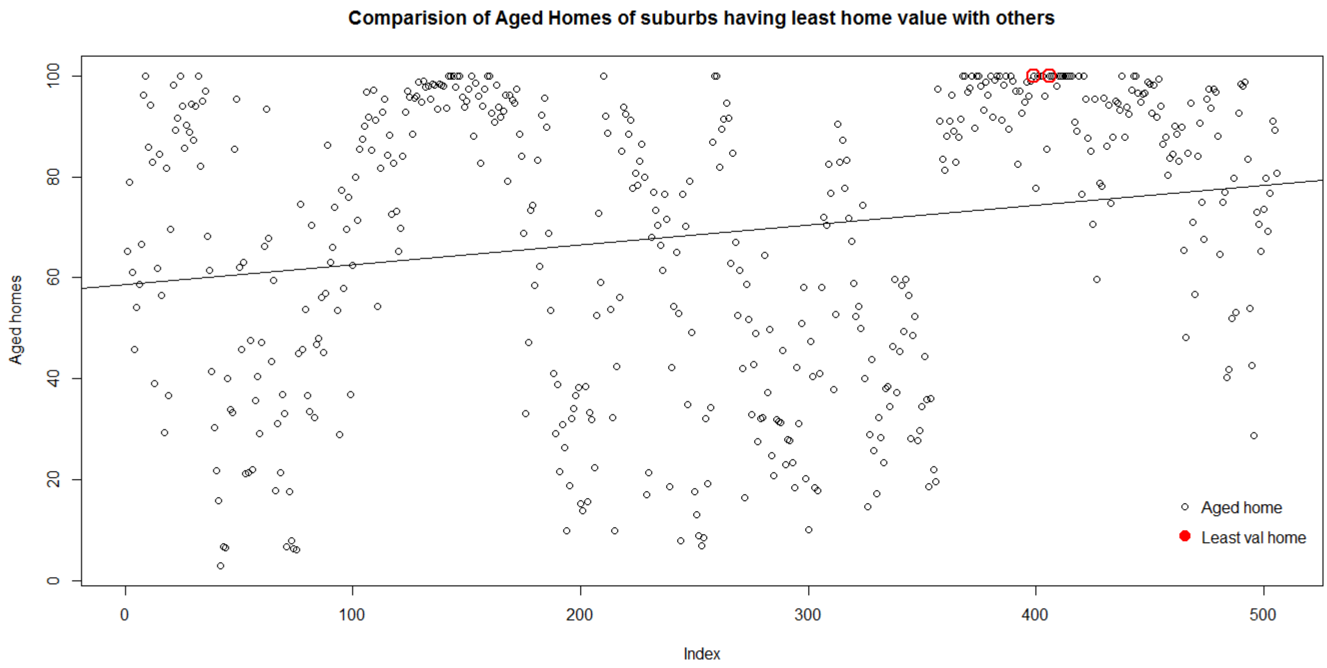

```
> plot(Boston$nox,ylab="nitrogen oxide content")
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$nox[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2)
> legend('topleft',c("nitrogen conc"),pch=21,bty="n",x.intersp=0.3,y.intersp=0.3)
> legend('topleft',c("Least val home"),pch=19,col="red",cex=1,bty='n',x.intersp=0.3,pt.cex = 1.5)
> abline(lm(Boston$nox ~ as.numeric(rownames(Boston)) ))
> title("Comparision of Nitrogen Ox. Conc of suburbs having least home value with others")
```



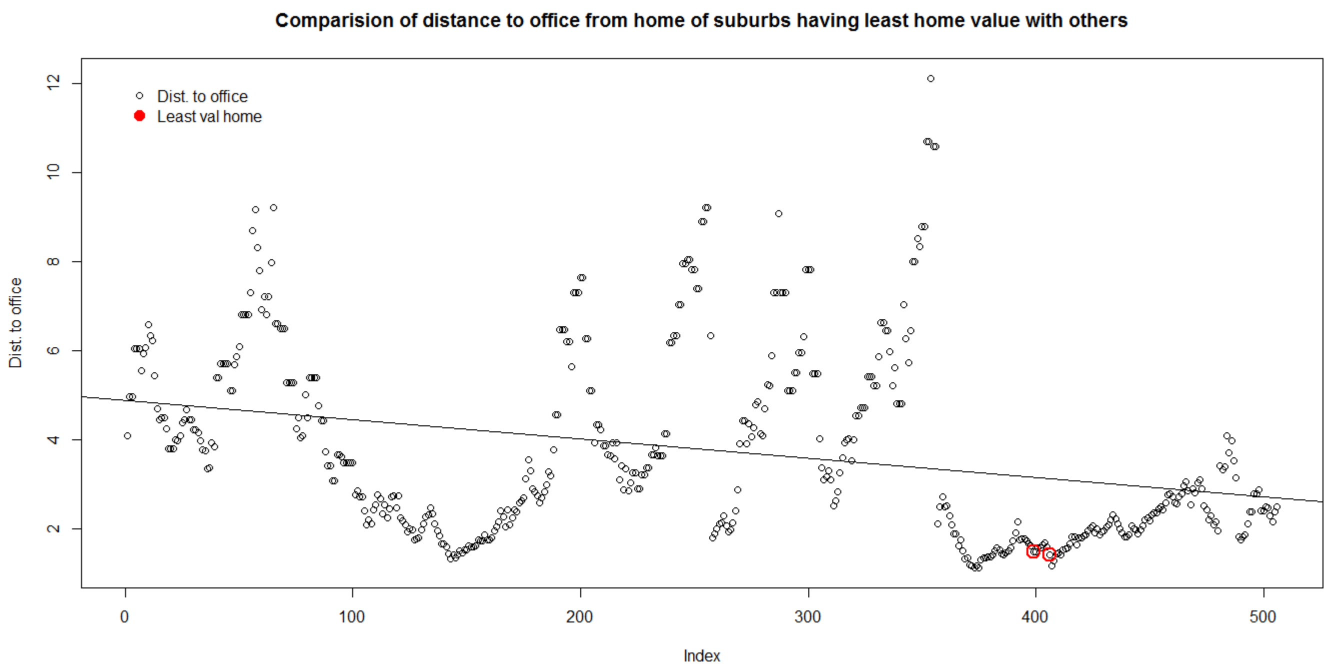

```
> plot(Boston$rm,ylab="Rooms per Dwelling")
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$rm[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2)
> legend('topleft',c("Room/home"),pch=21,bty="n",x.intersp=0.3,y.intersp=0.3)
> legend('topleft',c("Least val home"),pch=19,col="red",cex=1,bty='n',x.intersp=0.3,pt.cex = 1.5)
> abline(lm(Boston$rm ~ as.numeric(rownames(Boston)) ))
> title("Comparision of rooms per home of suburbs having least home value with others")
```



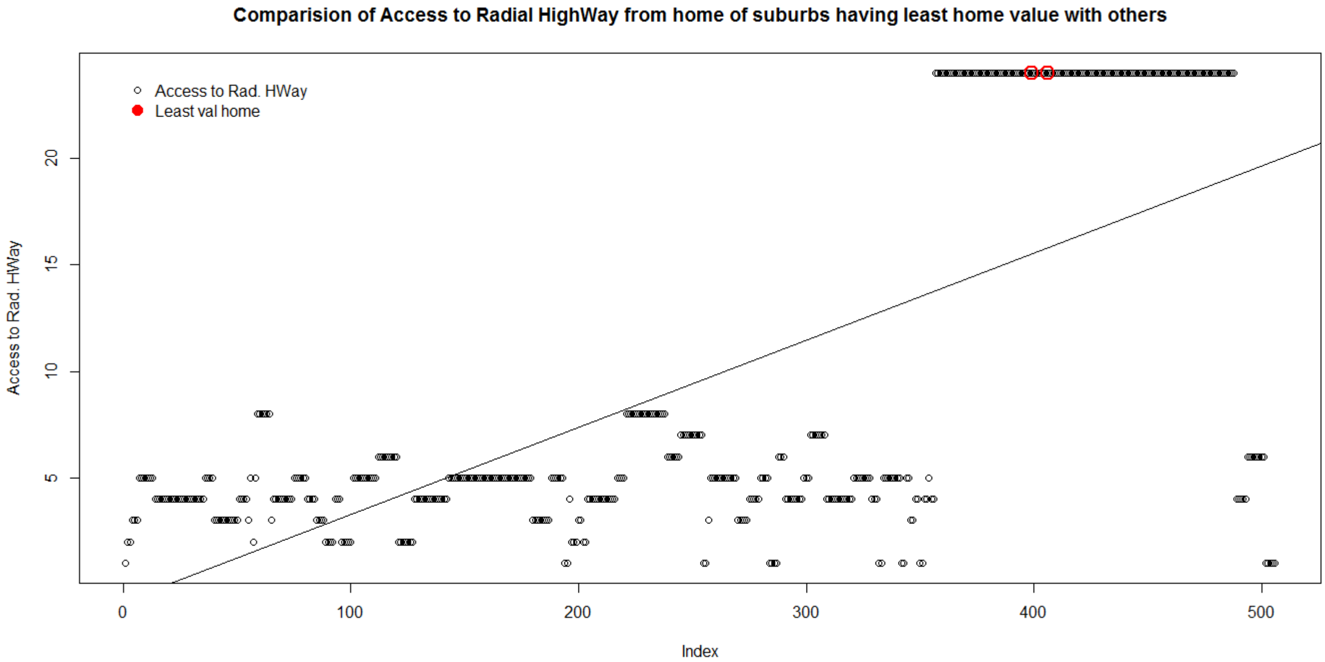

```
> plot(Boston$age,ylab="Aged homes")
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$age[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2)
> legend(x=440,y=22,c("Aged home"),pch=21,bty="n",x.intersp=0.3,y.intersp=0.3)
> legend(x=440,y=20,c("Least val home"),pch=19,col="red",cex=1,bty='n',x.intersp=0.3,pt.cex = 1.5)
> abline(lm(Boston$age ~ as.numeric(rownames(Boston)) ))
> title("Comparision of Aged Homes of suburbs having least home value with others")
```



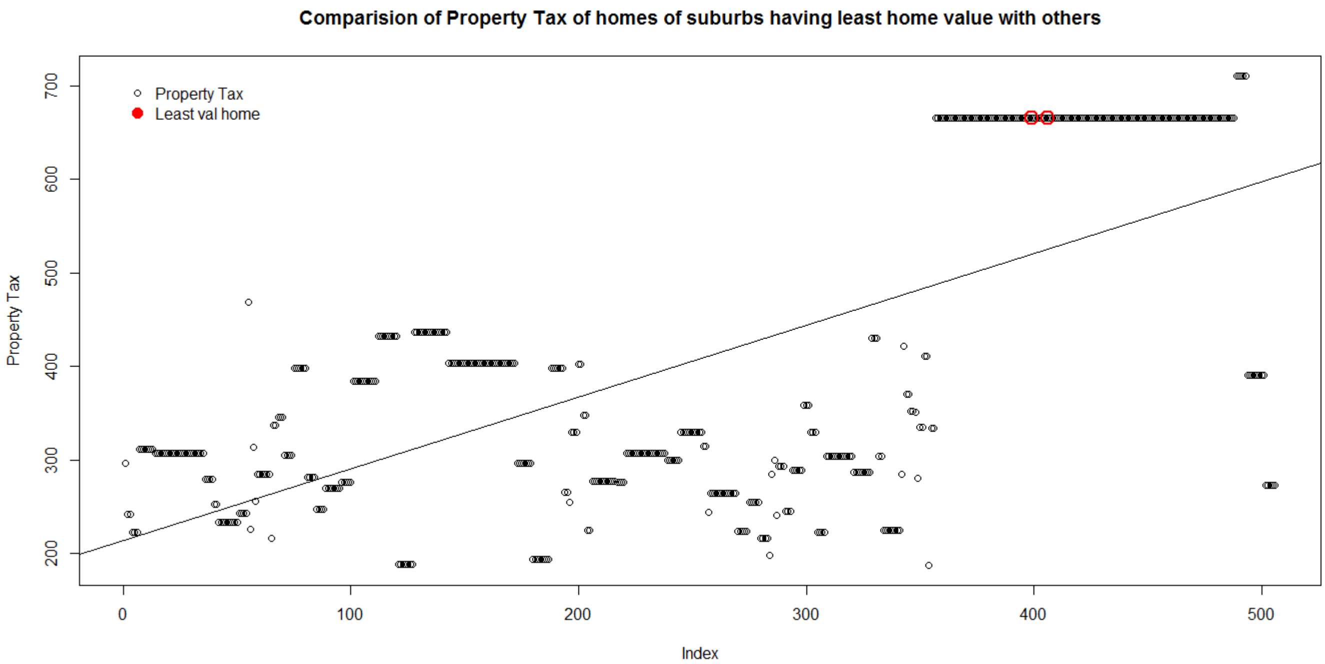

```
> plot(Boston$dis,ylab="Dist. to office")
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$dis[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2)
> legend('topleft',c("Dist. to office"),pch=21,bty="n",x.intersp=0.3,y.intersp=0.3)
> legend('topleft',c("Least val home"),pch=19,col="red",cex=1,bty='n',x.intersp=0.3,pt.cex = 1.5)
> abline(lm(Boston$dis ~ as.numeric(rownames(Boston)) ))
> title("Comparision of distance to office from home of suburbs having least home value with others")
```




```
> plot(Boston$rad,ylab="Access to Rad. HWay")
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$rad[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2)
> legend('topleft',c("Access to Rad. HWay"),pch=21,bty="n",x.intersp=0.3,y.intersp=0.3)
> legend('topleft',c("Least val home"),pch=19,col="red",cex=1,bty='n',x.intersp=0.3,pt.cex = 1.5)
> abline(lm(Boston$rad ~ as.numeric(rownames(Boston)) ))
> title("Comparision of Access to Radial HighWay from home of suburbs having least home value with others")
```

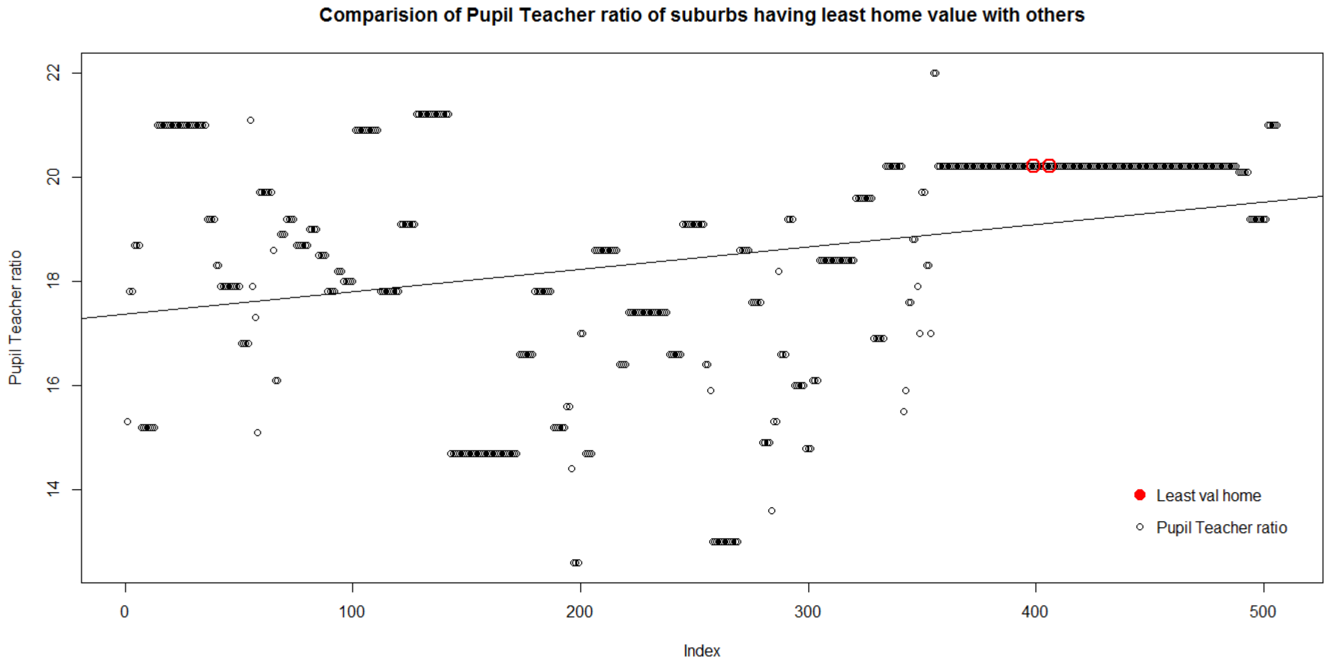



```
> plot(Boston$tax,ylab="Property Tax")
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$tax[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2)
> legend('topleft',c("Property Tax"),pch=21,bty="n",x.intersp=0.3,y.intersp=0.3)
> legend('topleft',c("Least val home"),pch=19,col="red",cex=1,bty='n',x.intersp=0.3,pt.cex = 1.5)
> abline(lm(Boston$tax ~ as.numeric(rownames(Boston)) ))
> title("Comparision of Property Tax of homes of suburbs having least home value with others")
```

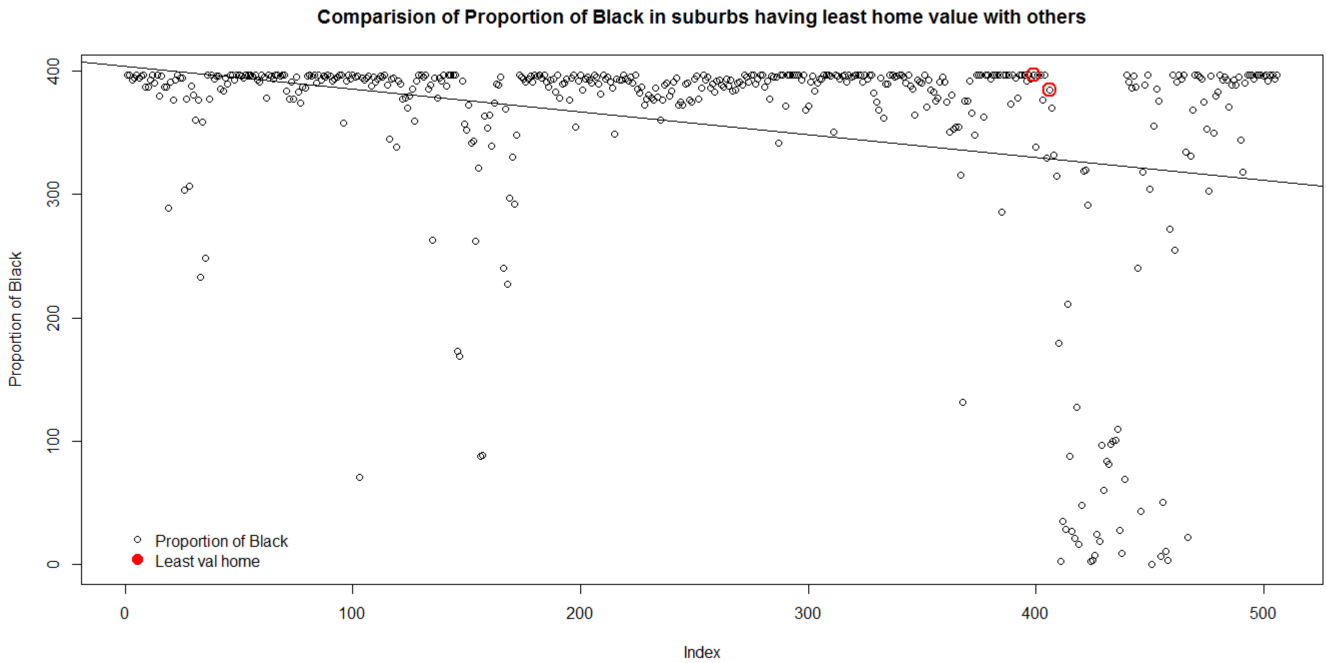



```
> plot(Boston$ptratio,ylab="Pupil Teacher ratio")
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$ptratio[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2)
> legend(x=420,y=14,c("Pupil Teacher ratio"),pch=21,bty="n",x.intersp=0.3,y.intersp=0.3)
> legend(x=420,y=15,c("Least val home"),pch=19,col="red",cex=1,bty='n',x.intersp=0.3,pt.cex = 1.5)
> abline(lm(Boston$ptratio ~ as.numeric(rownames(Boston)) ))
> title("Comparision of Pupil Teacher ratio of suburbs having least home value with others")
```

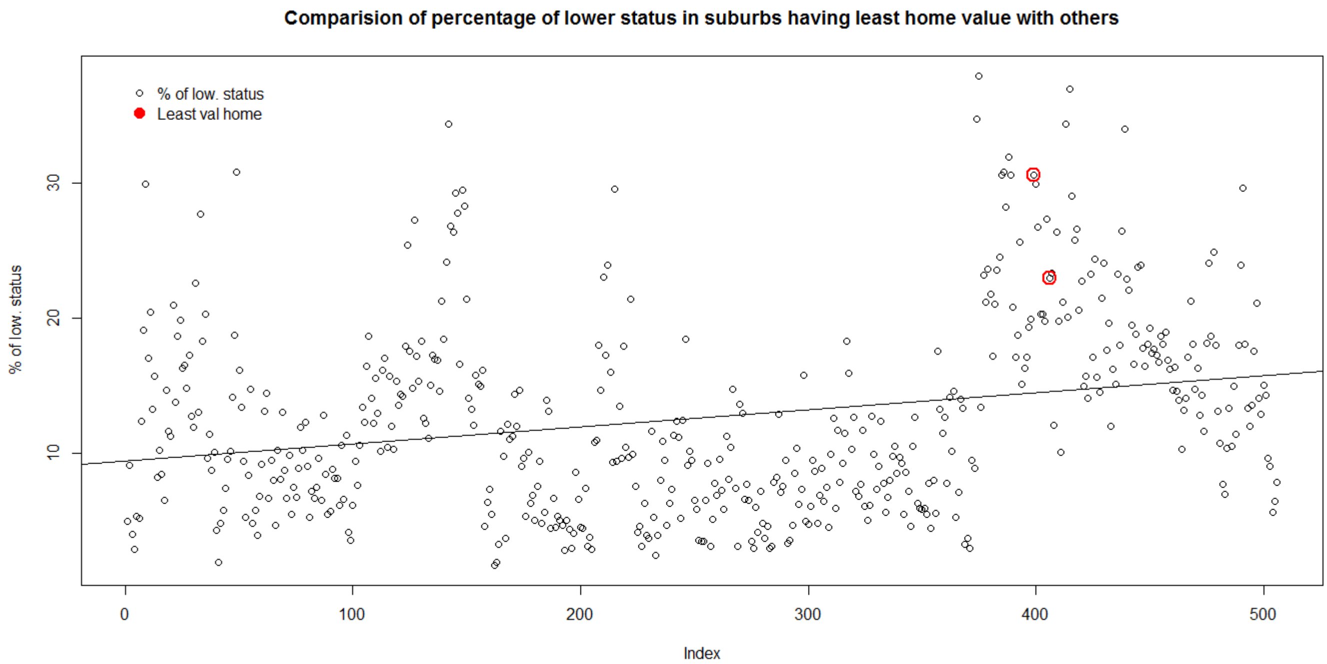
Comparision of Pupil Teacher ratio of suburbs having least home value with others




```
> plot(Boston$black,ylab="Proportion of Black")
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$black[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2)
> legend(x=-20,y=50,c("Proportion of Black"),pch=21,bty="n",x.intersp=0.3,y.intersp=0.3)
> legend(x=-20,y=50,c("Least val home"),pch=19,col="red",cex=1,bty='n',x.intersp=0.3,pt.cex = 1.5)
> abline(lm(Boston$black ~ as.numeric(rownames(Boston)) ))
> title("Comparision of Proportion of Black in suburbs having least home value with others")
```




```
> plot(Boston$lstat,ylab="% of low. status")
> points(rownames(Boston[Boston$medv == min(as.numeric(Boston$medv)),]),Boston$lstat[Boston$medv == min(as.numeric(Boston$medv))],col="#FF0000",lwd=2,cex=2)
> legend('topleft',c("% of low. status"),pch=21,bty="n",x.intersp=0.3,y.intersp=0.3)
> legend('topleft',c("Least val home"),pch=19,col="red",cex=1,bty='n',x.intersp=0.3,pt.cex = 1.5)
> abline(lm(Boston$lstat ~ as.numeric(rownames(Boston)) ))
> title("Comparision of percentage of lower status in suburbs having least home value with others")
```



.

Comparison/Characteristics of predictors :

- Higher Crime rates than other suburbs.
- No proportion of residential land zoned for lots over 25,000 sq.ft.
- Higher proportion of non-retail business acres per town.
- Very high proportion of owner-occupied units built prior to 1940 houses.
- High proportion of blacks by town.
- Very Low weighted mean of distances to five Boston employment centers.
- High percentage of lower status population.
- Above average level of nitrogen oxide concentration.
- Pupil-teacher ratio by town slightly higher than the mean.
- High index of accessibility to radial highways.
- Average number of rooms per dwelling is just below the average.
- High full-value property-tax rate per $10,000.

.

1.e Suburbs having average more than 7 and 8 rooms per dwelling.

```
> nrow(Boston[Boston$rm > 7,])
[1] 64
> nrow(Boston[Boston$rm > 8,])
[1] 13
```

Observations of characteristic of dwelling with more than 8 rooms:

- For Majority of the dwellings, the tract does not bound the Charles river.
- Percentage of the lowers status population is lower than the mean of sample population.
- Generally low crime rates.

- Mean of 'age' variable is slightly higher than that of the population. However median is very close to the sample population median. (Possible outlier with 'age'=8.4 whereas mean for the category being 71.53).
- Mean and median for 'medv' is more than twice for corresponding sample population values.

.

2.a Code for calculating Mahalanobi's distance.

```
> x= matrix(c(2,10,3,3,7,2),nrow=3)
> x
     [,1] [,2]
[1,]    2    3
[2,]   10    7
[3,]    3    2
> y=matrix(c(1,3,5,15,8,16,4,3,7,2,2,4,33,7),ncol=2)
> y
     [,1] [,2]
[1,]    1    3
[2,]    3    7
[3,]    5    2
[4,]   15    2
[5,]    8    4
[6,]   16   33
[7,]    4    7
> mean_x = matrix(c(mean(x[,1]),mean(x[,2])),nrow=2)
> mean_x
     [,1]
[1,]    5
[2,]    4
> mean_y = matrix(c(mean(y[,1]),mean(y[,2])),nrow=2)
> mean_y
         [,1]
[1,] 7.428571
[2,] 8.285714
> x1=matrix(c(x[,1]-mean(x[,1]),c(x[,2]-mean(x[,2]))),ncol=2)
> x1
     [,1] [,2]
[1,]   -3   -1
[2,]    5    3
[3,]   -2   -2
> y1=matrix(c(y[,1]-mean(y[,1]),c(y[,2]-mean(y[,2]))),ncol=2)
> y1
           [,1]      [,2]
[1,] -6.4285714 -5.285714
[2,] -4.4285714 -1.285714
[3,] -2.4285714 -6.285714
[4,]  7.5714286 -6.285714
[5,]  0.5714286 -4.285714
[6,]  8.5714286 24.714286
[7,] -3.4285714 -1.285714
> c1 = (1/nrow(x1))*(t(x1)%*%x1)
> c1
          [,1]     [,2]
[1,] 12.666667 7.333333
[2,]  7.333333 4.666667
> c2 = (1/nrow(y1))*(t(y1)%*%y1)
> c2
         [,1]      [,2]
[1,] 29.95918  31.59184
[2,] 31.59184 105.63265
> s = (nrow(x1)/(nrow(x1)+nrow(y1)))*c1 + (nrow(y1)/(nrow(x1)+nrow(y1)))*c2
> s
         [,1]     [,2]
[1,] 24.77143 24.31429
[2,] 24.31429 75.34286
> solve(s)
            [,1]        [,2]
[1,]  0.05908476 -0.01906755
[2,] -0.01906755  0.01942605
> Mahalanobis_dist = sqrt(t(mean_x - mean_y) %*% solve(s) %*% (mean_x - mean_y))
> Mahalanobis_dist
         [,1]
[1,] 0.555309
```

2.b Function that calculates Mahalanobi's distance.

```
> Mahalanobis_dist = function(x,y)
+ {
+     if(ncol(x)!=ncol(y))
+     {
+         print("ERROR! Number of columns in two matrices must be same!")
+         return()
+     }
+     mean_x = matrix(c(mean(x[,1]),mean(x[,2])),nrow=2)
+     mean_y = matrix(c(mean(y[,1]),mean(y[,2])),nrow=2)
+     x1=matrix(c(x[,1]-mean(x[,1]),c(x[,2]-mean(x[,2]))),ncol=2)
+     y1=matrix(c(y[,1]-mean(y[,1]),c(y[,2]-mean(y[,2]))),ncol=2)
+     c1 = (1/nrow(x1))*(t(x1)%*%x1)
+     c2 = (1/nrow(y1))*(t(y1)%*%y1)
+     s = (nrow(x1)/(nrow(x1)+nrow(y1)))*c1 + (nrow(y1)/(nrow(x1)+nrow(y1)))*c2
+     dist = sqrt(t(mean_x - mean_y) %*% solve(s) %*% (mean_x - mean_y))
+     return(dist)
+ }
> x= matrix(c(2,10,3,3,7,2),nrow=3)
> x
     [,1] [,2]
[1,]    2    3
[2,]   10    7
[3,]    3    2
> y=matrix(c(1,3,5,15,8,16,4,3,7,2,2,4,33,7),ncol=2)
> y
     [,1] [,2]
[1,]    1    3
[2,]    3    7
[3,]    5    2
[4,]   15    2
[5,]    8    4
[6,]   16   33
[7,]    4    7
> Mahalanobis_dist(x,y)
         [,1]
[1,] 0.555309
```

.

3 Difference between Artificial Intelligence, Machine Learning, Statistics, and Data Mining.

.

Though all of the four terms in discussion has many grounds overlapping each other, there are some significant differences among them. All are in way either related to each other or one form of the other.

Starting from Artificial Intelligence(AI), the goal of AI is simply to induce intelligence to machines, so that they are enabled to make independent decisions without human intervention. This is a very broad area and has given rise to many disciplines within itself like Natural Language Processing(NLP), Robotics, Computer vision, Reasoning etc. Machine Learning is one such discipline that has grown out of the need for pure AI.

Machine Learning tries to enable machines to make decisions on their own by feeding them with training data and using some generic algorithms. These algorithms can be employed to solve a variety of problems and are most of the time, if not always, are directly derived or inspired by classical statistics. One of the application of such algorithms are in Data Mining.

Data Mining uses algorithms or techniques that are mostly coined by Machine learning, and apply it to a specific domain/area of interest. Data mining has a very clear goal unlike Machine learning and tries to solve/understand a particular problem. Data Mining has been commonly used to leverage the data in hand and make predictions, draw inferences, reason associations, finding patterns etc. It has been branched out from the field of exploratory Statistics.

Lastly, Statistics is a branch of mathematics that concentrates on collection, analysis, interpretation, presentation and organization of the data. This is the oldest of the other three fields in discussion. Statistics can also be seen as a way to transform data into information or insights.