Econ 104L: Group Project

Project #1

Omer Abdelrahim

Contents

1 Part 1

1.1 Step 1: Descriptive Analysis of Variables

Relevant Information:

Concerns housing values in suburbs of Boston.

Number of Instances: 506

Number of Attributes: 13 continuous attributes (including "class" attribute "MEDV"), 1 binary-valued attribute.

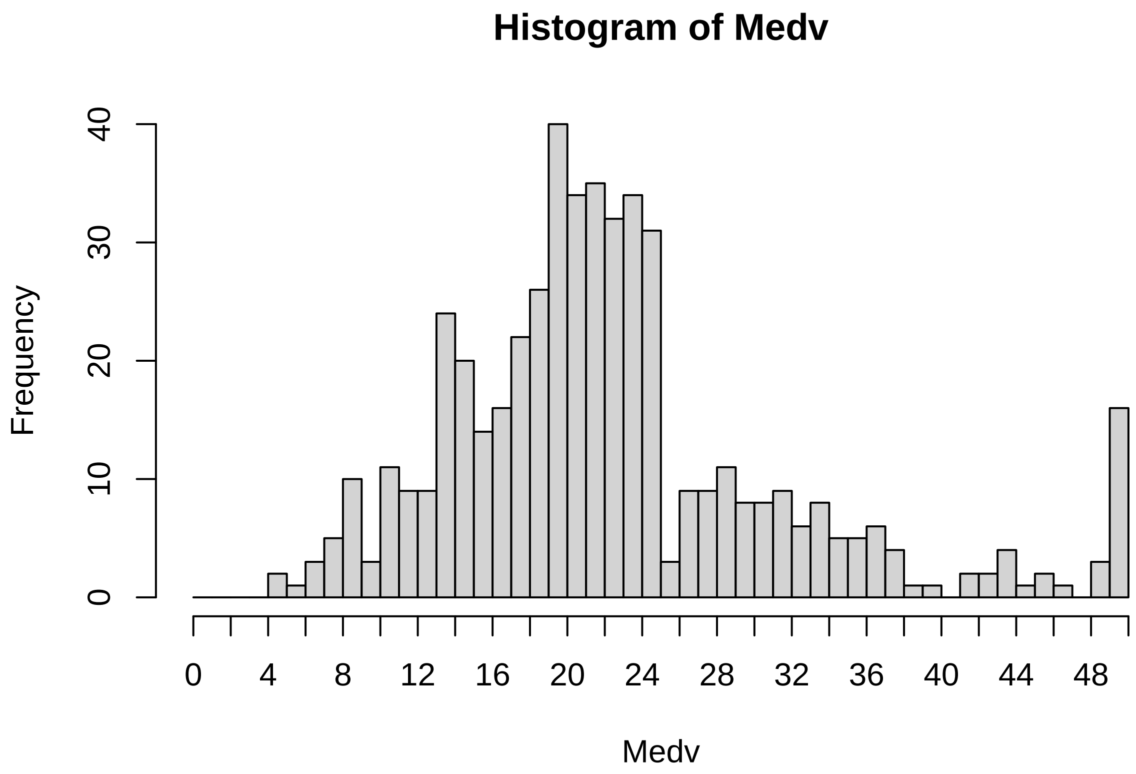
Attribute Information:

```
1. Crm       per capita crime rate by town
2. Zn        proportion of residential land zoned for lots over
             25,000 sq.ft.
3. Indus     proportion of non-retail business acres per town
4. Chas      Charles River dummy variable (= 1 if tract bounds
             river; 0 otherwise)
5. Nox       nitric oxides concentration (parts per 10 million)
```

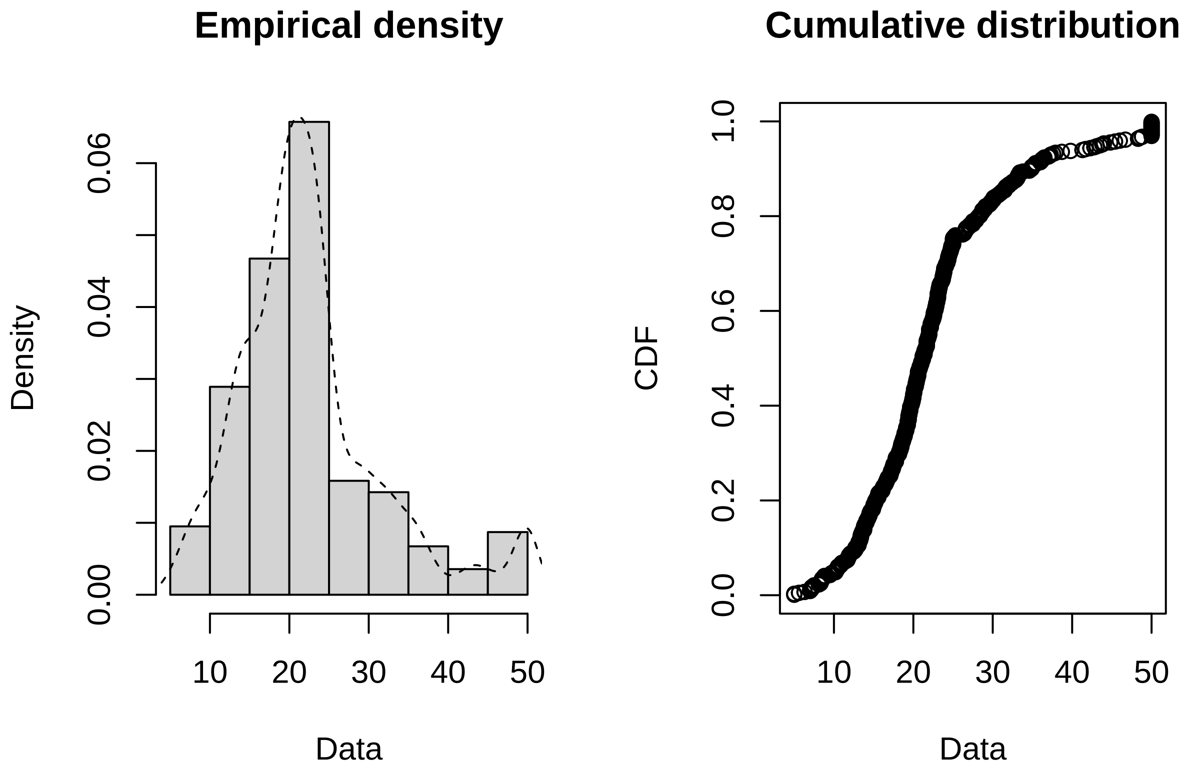
```
6. RM         average number of rooms per dwelling
7. Age        proportion of owner-occupied units built prior to 1940
8. Dis        weighted distances to five Boston employment centres
9. Rad        index of accessibility to radial highways
10. Tax       full-value property-tax rate per $10,000
11. Ptratio   pupil-teacher ratio by town
12. Lstat     % lower status of the population
13. Medv      Median value of owner-occupied homes in $1000's
```

Dependent Value (y): Median Housing Values in the Suburb (Medv) Predictors: Crm, Zn, Indus, Chas, Nox, Rm, Age, Dis, Rad, Tax, Ptratio, Lstat

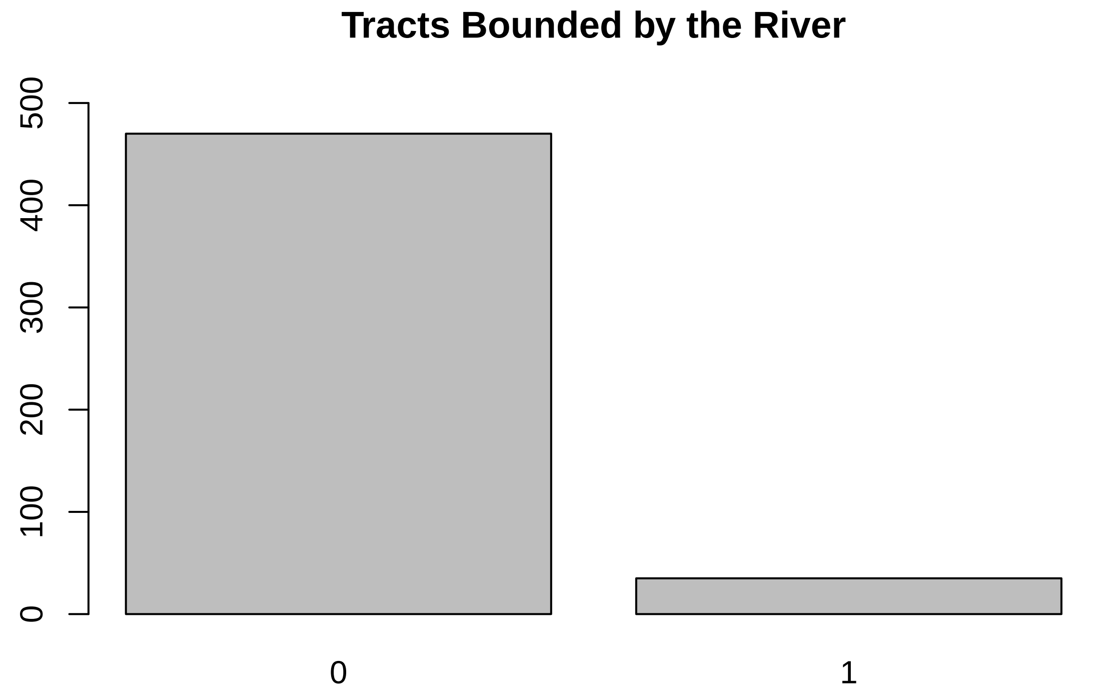
```
attach(Bhousing)
hist(Medv, breaks = seq(0,50,1), xaxp=c(0,50,25))
```



```
plotdist(Medv, histo = TRUE, demp = TRUE)
```

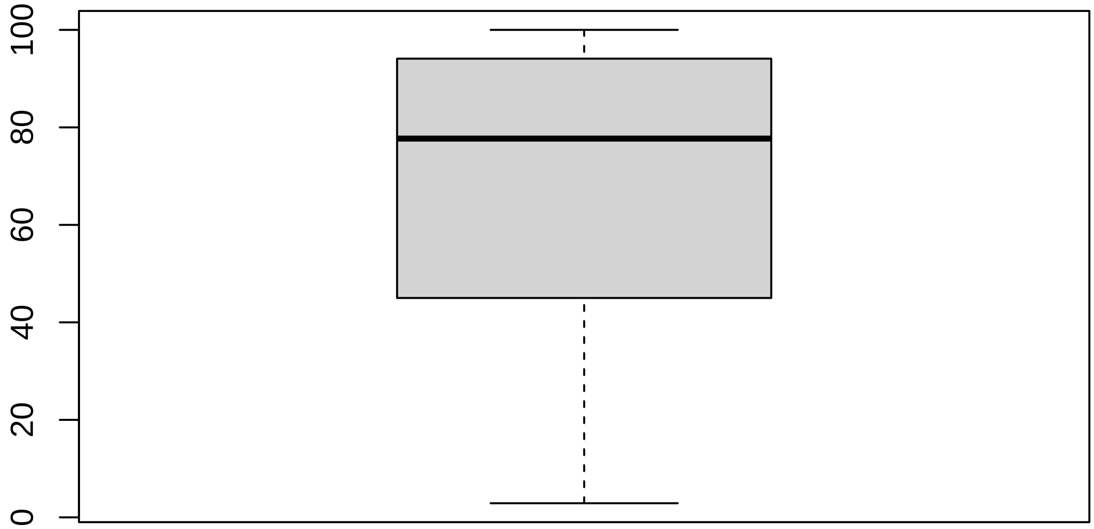


```r
River <- table(Chas)
barplot(River, main = "Tracts Bounded by the River", ylim =c(0,500))
```



Majority of the tracts are not bound by the Charles River, about a 10:1 ratio.

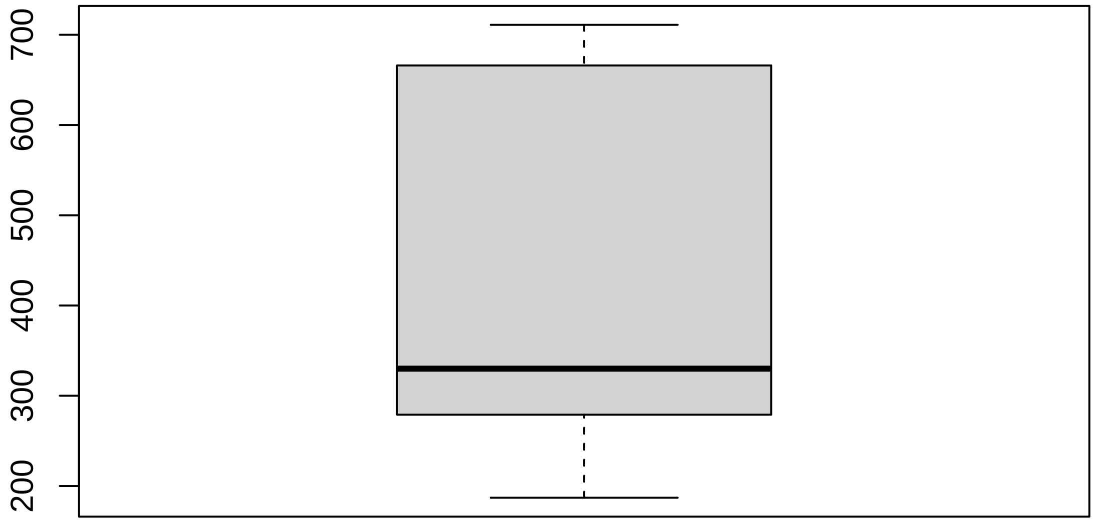
```
boxplot(Age)
```



```
fivenum(Age)
```

```
## [1]   2.9  45.0  77.7  94.1 100.0
```

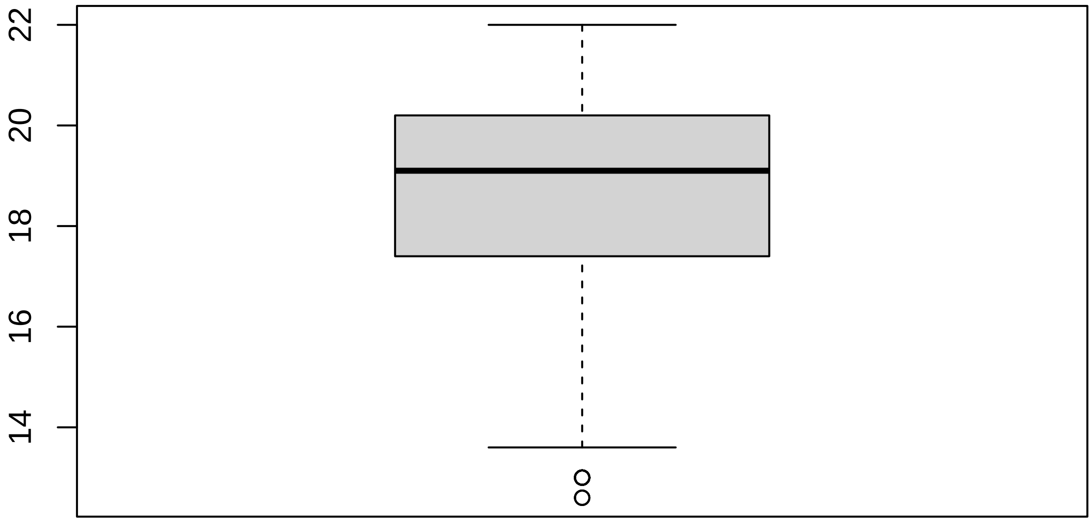
```
boxplot(Tax)
```



```
fivenum(Tax)
```

```
## [1] 187 279 330 666 711
```

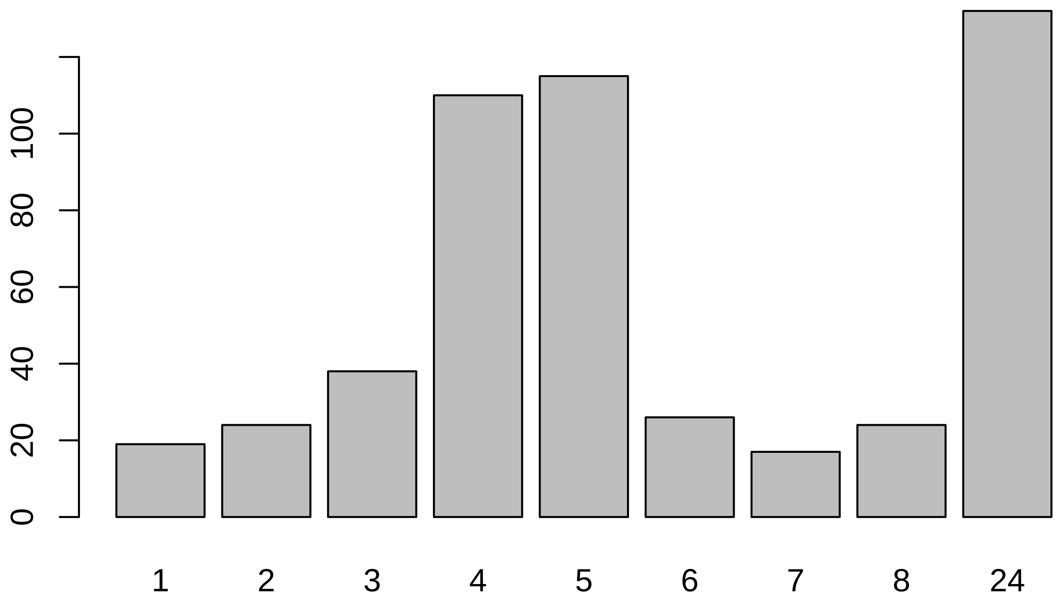
```
boxplot(Ptratio)
```



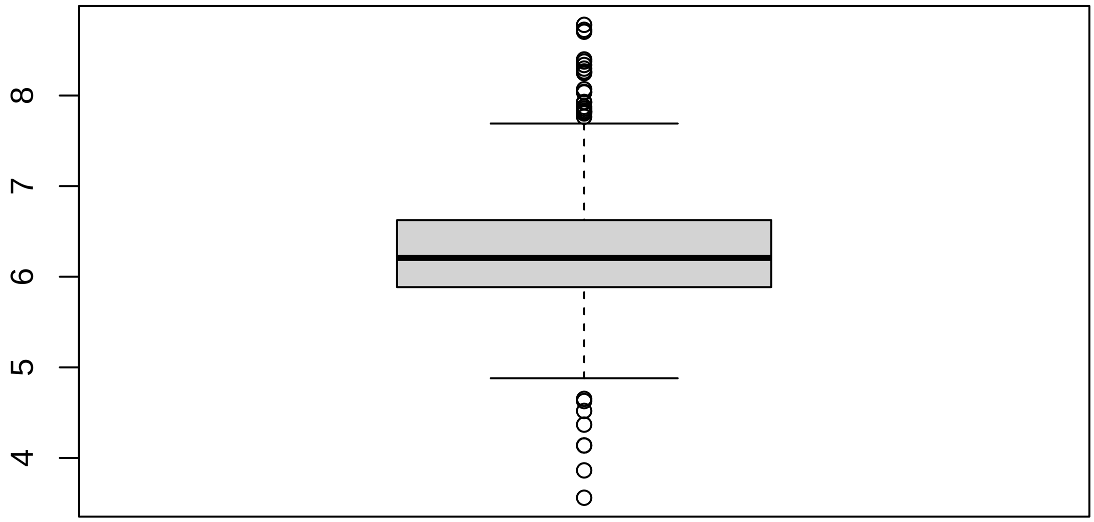
```
fivenum(Ptratio)
```

```
## [1] 12.6 17.4 19.1 20.2 22.0
```

```
Access <-table(Rad)
barplot(Access)
```



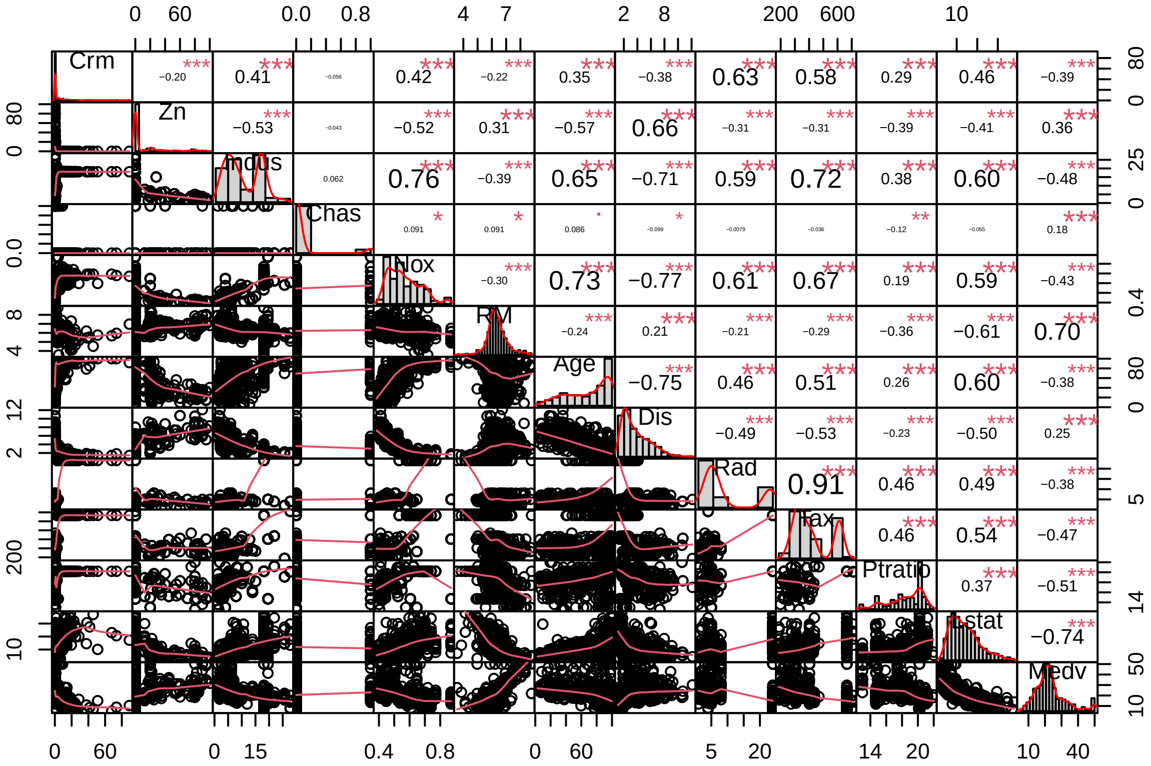
```
boxplot(RM)
```



```
fivenum(RM)
```

```
## [1] 3.561 5.885 6.208 6.625 8.780
```

```
chart.Correlation(Bhousing, histogram = TRUE)
```



2 Part 2

2.1 Multiple Regression Predicting Median House value in the Boston Suburbs

```
reg.Bfull <-lm(Medv~Crm+Zn+Tax+Nox+Ptratio+Rad+Dis+RM+Age+Lstat)
summary(reg.Bfull)
```

```
##
## Call:
## lm(formula = Medv ~ Crm + Zn + Tax + Nox + Ptratio + Rad + Dis +
##     RM + Age + Lstat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7205  -2.8185  -0.6101   2.1375  26.5382
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.367430   4.963780   8.535  < 2e-16 ***
## Crm          -0.127368   0.033197  -3.837 0.000141 ***
## Zn            0.046688   0.013913   3.356 0.000853 ***
## Tax          -0.013177   0.003431  -3.841 0.000139 ***
## Nox         -17.781536   3.734971  -4.761 2.54e-06 ***
## Ptratio      -0.976625   0.131889  -7.405 5.71e-13 ***
## Rad           0.299347   0.064651   4.630 4.68e-06 ***
## Dis          -1.524610   0.198564  -7.678 8.71e-14 ***
## RM            3.660946   0.422050   8.674  < 2e-16 ***
## Age           0.006391   0.013422   0.476 0.634142    
## Lstat        -0.564578   0.050869 -11.099  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.837 on 494 degrees of freedom
## Multiple R-squared:  0.7294, Adjusted R-squared:  0.7239
## F-statistic: 133.1 on 10 and 494 DF,  p-value: < 2.2e-16
```

Nox has an outsized negative effect on median value, removing it from the model will probably result in an increased accuracy for the model, and may help to improve accuracy of the Age statistic. This may be doubtful though, as in a city such as Boston, many of the houses are post 1940, and should have no real effect on the price, unless age is indicative of a lack of amenities among other things.

RM and Dis also seem like prime candidates to remove from the regression as they have outsized affects in comparison to peer statistics, but using a bit of of real world knowledge, location and the number of rooms do in fact have signficant effects in terms of property evaluation in the real world. As a result both of these predictors will stay.

The residuals look good, and the R value is quite high for a financial regression.

3 Part 3

3.1 Re-evaluation of the multiple regression with the removal of the predictor Nox

```
reg.BfullA <-lm(Medv~Crm+Zn+Tax+Ptratio+Rad+Dis+RM+Age+Lstat)
summary(reg.BfullA)
```

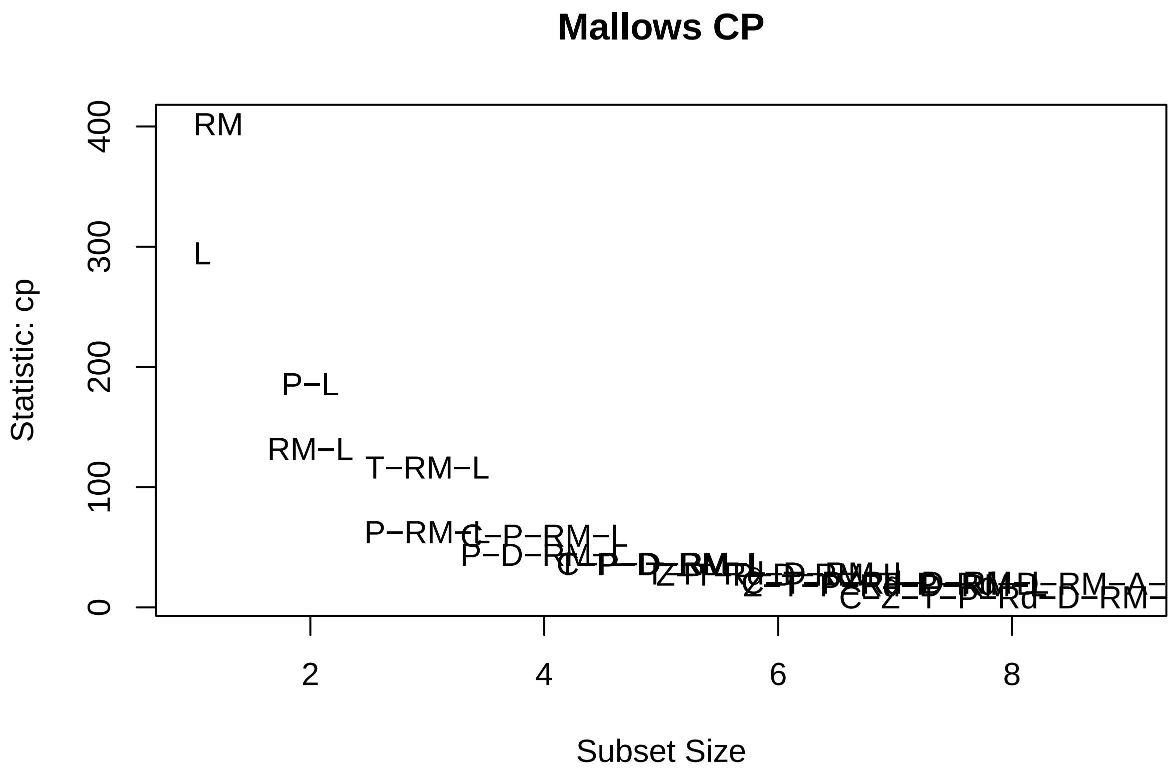
```
##
## Call:
## lm(formula = Medv ~ Crm + Zn + Tax + Ptratio + Rad + Dis + RM +
##     Age + Lstat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4726  -2.9292  -0.7583   1.6302  26.9302
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.237841   4.216510   6.934 1.28e-11 ***
## Crm         -0.116861   0.033841  -3.453 0.000601 ***
## Zn           0.051190   0.014181   3.610 0.000338 ***
## Tax         -0.016774   0.003419  -4.906 1.26e-06 ***
## Ptratio     -0.776669   0.127729  -6.081 2.40e-09 ***
## Rad          0.273005   0.065808   4.149 3.94e-05 ***
## Dis         -1.182893   0.189146  -6.254 8.66e-10 ***
## RM           3.889621   0.428385   9.080  < 2e-16 ***
## Age         -0.011193   0.013183  -0.849 0.396278
## Lstat       -0.589957   0.051684 -11.415  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.942 on 495 degrees of freedom
## Multiple R-squared:  0.717,  Adjusted R-squared:  0.7118
## F-statistic: 139.3 on 9 and 495 DF,  p-value: < 2.2e-16
```

The removing of Nox as a predictor heavily affects the Intercept and as a result it was probably in the best interest of accuracy to remove it.

4 Part 4

4.1 Part 1: Mallows Cp

```
MCPBH=regsubsets(Medv~Crm+Zn+Tax+Ptratio+Rad+Dis+RM+Age+Lstat,method=c("exhaustive") ,nbest = 2, data =
subsets(MCPBH,statistic="cp",legend=F,main="Mallows CP")
```



```
##          Abbreviation
## Crm                C
## Zn                 Z
## Tax                T
## Ptratio            P
## Rad               Rd
## Dis                D
## RM                RM
## Age                A
## Lstat              L
```

```
model1<-lm(Medv~Crm)
model2<-lm(Medv~Crm+Zn)
model3<-lm(Medv~Crm+Zn+Tax)
model4<-lm(Medv~Crm+Zn+Tax+Ptratio)
model5<-lm(Medv~Crm+Zn+Tax+Ptratio+Rad)
model6<-lm(Medv~Crm+Zn+Tax+Ptratio+Rad+Dis)
model7<-lm(Medv~Crm+Zn+Tax+Ptratio+Rad+Dis+RM)
model8<-lm(Medv~Crm+Zn+Tax+Ptratio+Rad+Dis+RM+Age)
model9<-lm(Medv~Crm+Zn+Tax+Ptratio+Rad+Dis+RM+Age+Lstat)
model10<-lm(Medv~Crm+Zn+Tax+Ptratio+Rad+Dis+RM+Lstat)

ols_mallows_cp(model1, reg.BfullA)
```

```
## [1] 984.2534
```

```
ols_mallows_cp(model2, reg.BfullA)
```

```
## [1] 840.7123
```

```
ols_mallows_cp(model3, reg.BfullA)
```

```
## [1] 747.0342
```

```
ols_mallows_cp(model4, reg.BfullA)
```

```
## [1] 618.7718
```

```
ols_mallows_cp(model5, reg.BfullA)
```

```
## [1] 552.2234
```

```
ols_mallows_cp(model6, reg.BfullA)
```

```
## [1] 524.5789
```

```
ols_mallows_cp(model7, reg.BfullA)
```

```
## [1] 167.901
```

```
ols_mallows_cp(model8, reg.BfullA)
```

```
## [1] 138.2951
```

```
ols_mallows_cp(model9, reg.BfullA)
```

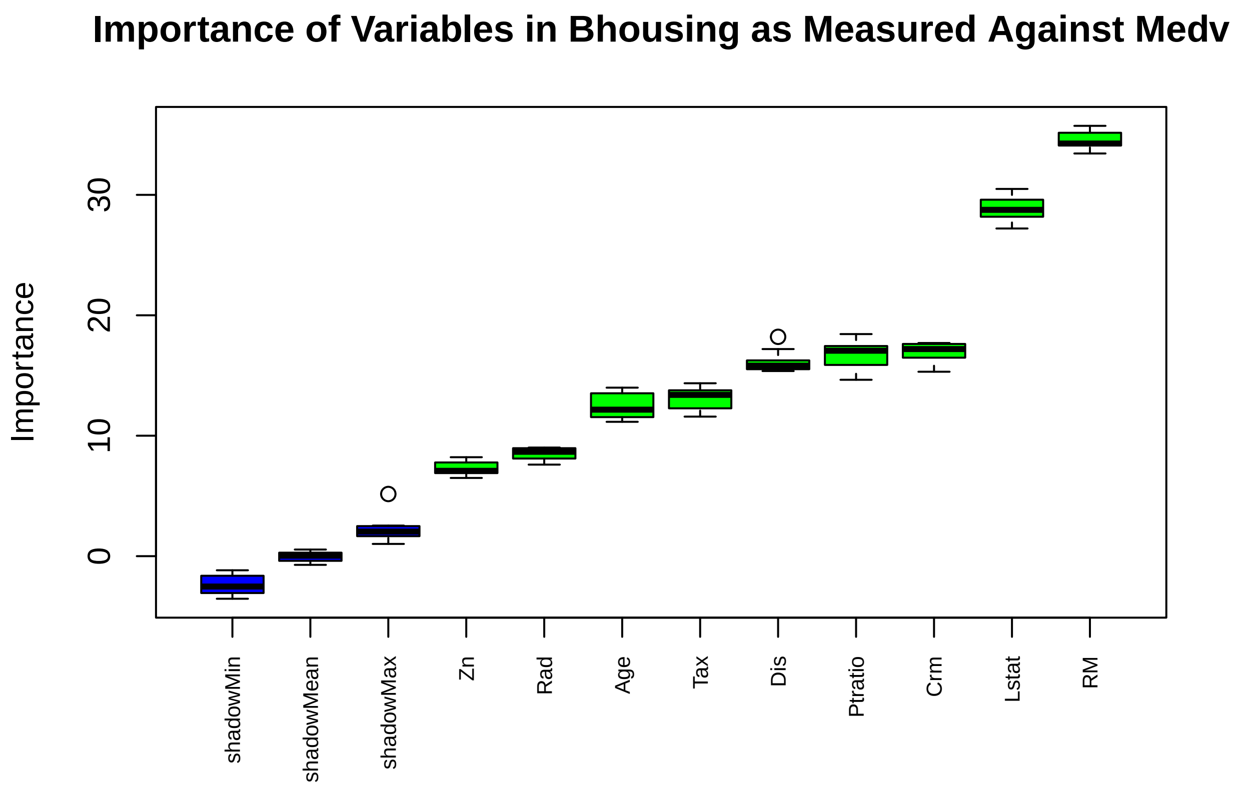
```
## [1] 10
```

```
ols_mallows_cp(model10, reg.BfullA)
```

```
## [1] 8.720842
```

4.2 Part 2: Boruta's Algorithm

```
Brt.res<-Boruta(Medv~Crm+Zn+Tax+Ptratio+Rad+Dis+RM+Age+Lstat, data=Bhousing)
plot(Brt.res,xlab = "", xaxt = "n",main="Importance of Variables in Bhousing as Measured Against Medv")
lz<-lapply(1:ncol(Brt.res$ImpHistory),function(i) Brt.res$ImpHistory[is.finite(Brt.res$ImpHistory[,i]),
names(lz) <- colnames(Brt.res$ImpHistory)
Labels <- sort(sapply(lz,median))
axis(side = 1,las=2,labels = names(Labels),
at = 1:ncol(Brt.res$ImpHistory), cex.axis = 0.7)
```



```
boruta_signif <- names(Brt.res$finalDecision[Brt.res$finalDecision %in% c("Confirmed")])
boruta_signif_Conf <- names(Brt.res$finalDecision[Brt.res$finalDecision %in% c("Confirmed")])

print(boruta_signif_Conf)
```

```
## [1] "Crm"     "Zn"      "Tax"     "Ptratio" "Rad"     "Dis"     "RM"
## [8] "Age"     "Lstat"
```

```
sorted_vars = attStats(Brt.res)[order(-attStats(Brt.res)$meanImp),]
print(sorted_vars)
```

```
##            meanImp medianImp    minImp    maxImp normHits  decision
## RM       34.460497 34.262175 33.435739 35.729980        1 Confirmed
## Lstat    28.821711 28.755652 27.207243 30.492231        1 Confirmed
## Crm      16.866059 17.194585 15.313624 17.700214        1 Confirmed
## Ptratio 16.829974 17.058137 14.643369 18.438551        1 Confirmed
## Dis      16.145792 15.815648 15.371269 18.211402        1 Confirmed
## Tax      13.048595 13.391982 11.586791 14.356370        1 Confirmed
## Age      12.448040 12.163436 11.154140 13.992209        1 Confirmed
## Rad       8.500602  8.665427  7.600352  9.018351        1 Confirmed
## Zn        7.254636  7.084966  6.492717  8.215643        1 Confirmed
```