104_Project_1

Krish Methi, Krithik Jatavallabhula

1/18/2024

Load in the data from AER Package

```
library(mlbench)
data("BostonHousing")
df <- BostonHousing[-c(4,12)]
```

Step 1: Descriptive analysis of each variable

Dataset as a whole

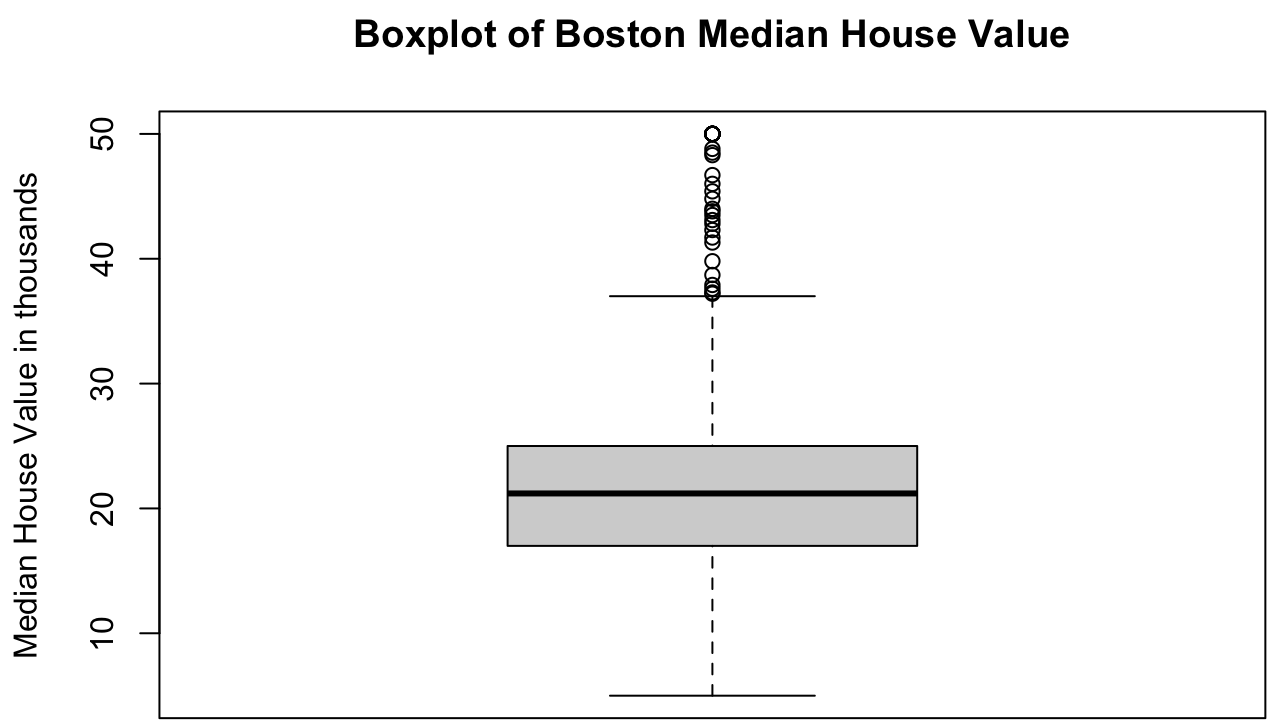
The Boston Housing dataset provides information on the details of various factors that may affect the median value of owner-occupied homes in different neighborhoods in boston, with the response variable being the median house price in thousands of dollars. In this dataset, there is originally 506 observations of 14 variables, however we have eliminated two columns so the dataset we are analyzing has 506 observations of 12 variables. These variables include things such as property tax, rooms, crime rates and so on, factors that may impact the value of these homes.

MedV

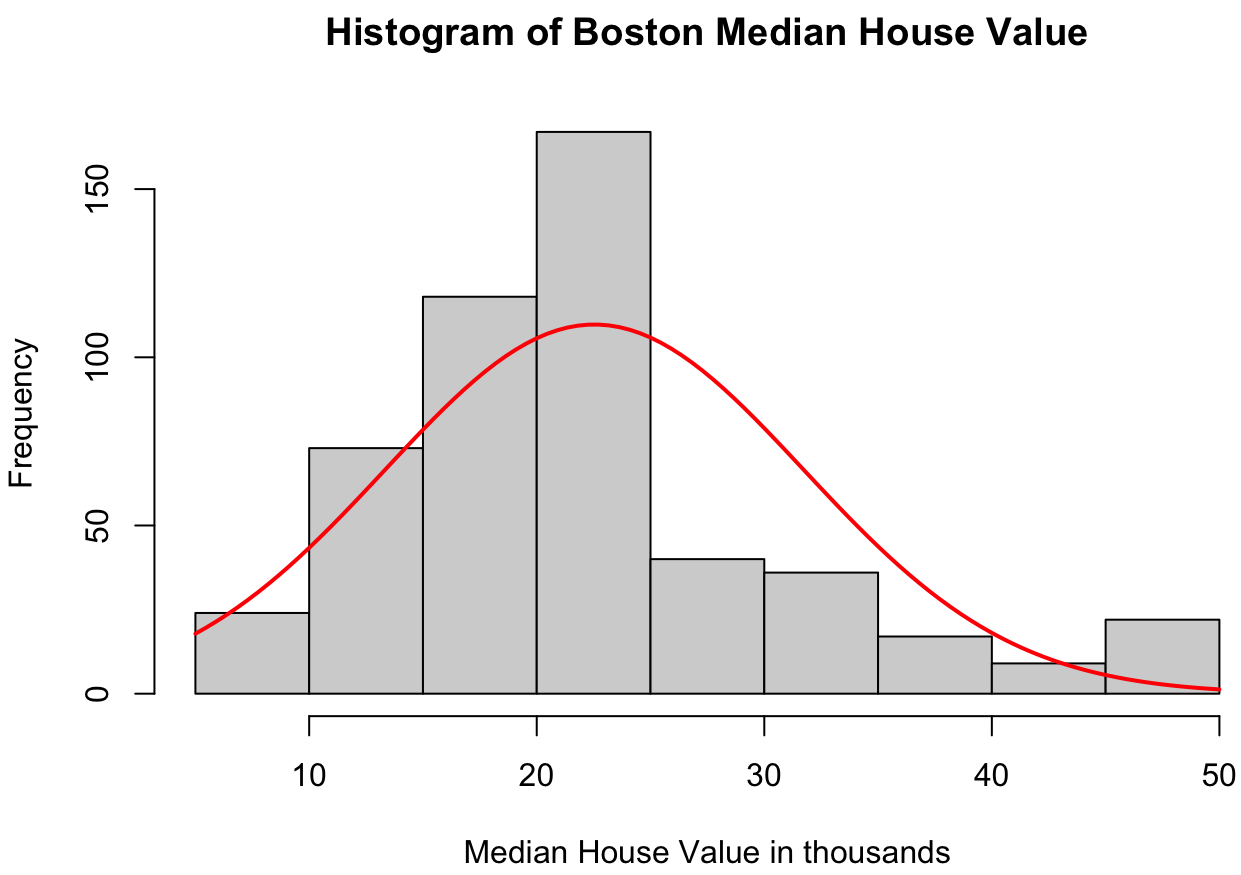
```
summary(df$medv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   17.02   21.20   22.53   25.00   50.00
```

```
boxplot(BostonHousing$medv, main = "Boxplot of Boston Median House Value", ylab = "Media
n House Value in thousands")
```




```
medv_hist <- hist(df$medv, main = "Histogram of Boston Median House Value", xlab = "Medi
an House Value in thousands", ylab = "Frequency")
xfit <- seq(min(df$medv), max(df$medv), length = 100)
yfit <- dnorm(xfit, mean = mean(df$medv), sd = sd(df$medv))
yfit <- yfit * diff(medv_hist$mids[1:2]) * length(df$medv)
lines(xfit, yfit, col = "red", lwd = 2)
```



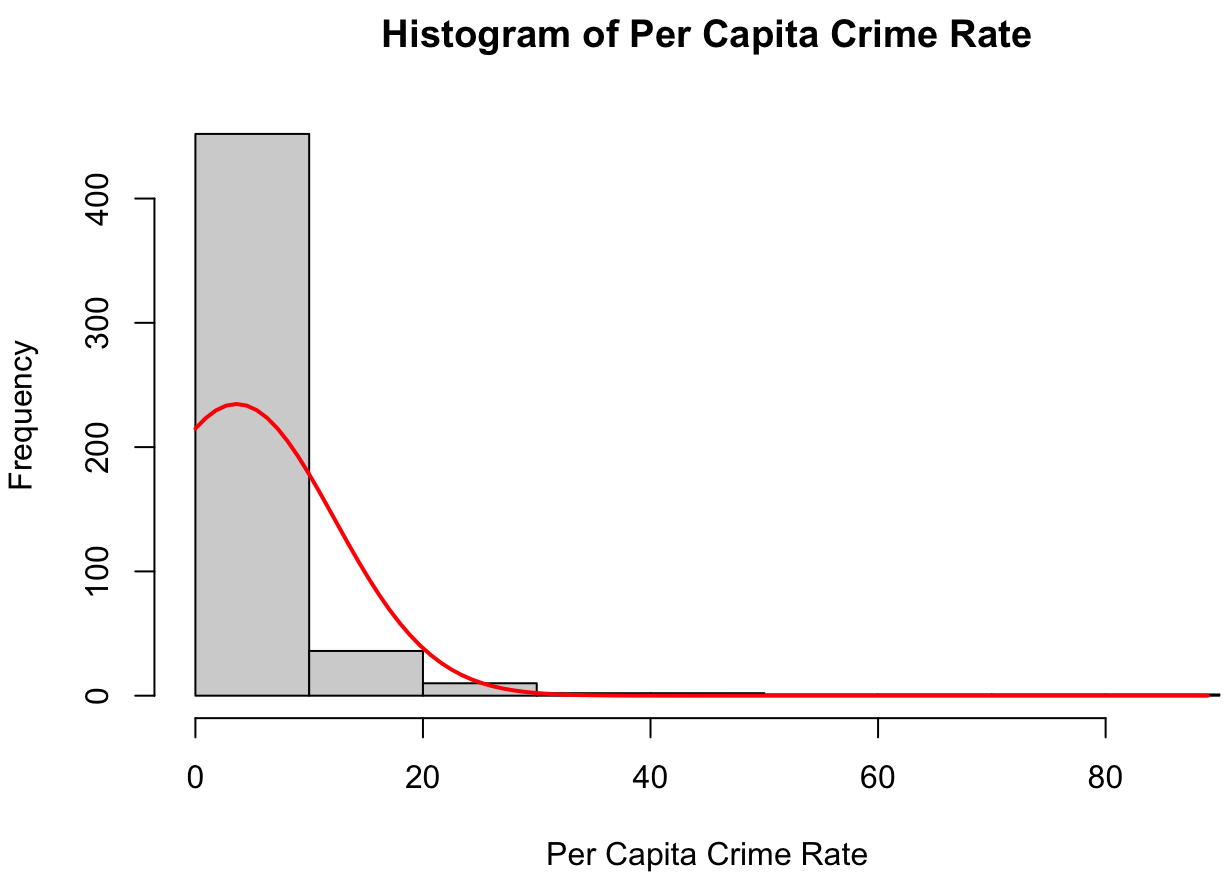
Comments: Medv refers to the median owner Boston house value in thousands, which is our response variable in this dataset. From the histogram, we can see that the data is roughly skewed to the right, with one main peak around in between 20-25 thousand. There doesn't seem to be a big outlier present from the histogram. Looking at the fitted red line normal curve, we can see the distribution somewhat resembles a normal distribution, but not fully. From the boxplot, we can see that the median is pretty much right in between the interquartile range, which suggests there might not be a great deal of skewness in the data. However we can see that from the box plot, there are quite a bit of outliers on the high side of the data. From the 5 number summary, we can identify the IQR of 17.02 thousand to 25 thousand dollars, and the median and mean of 21.20 and 22.53 thousand respectively, with a min of 5,000 and a max of 50,000, a range of 45,000 dollars

Crim

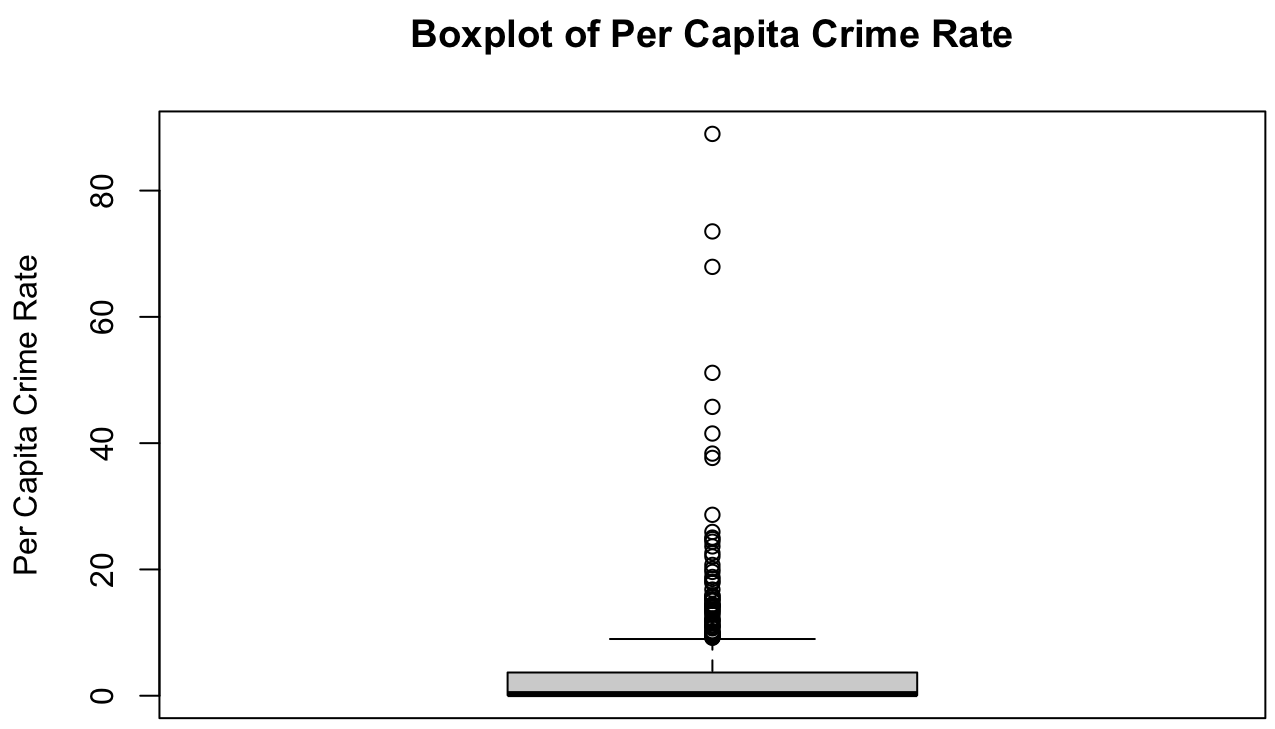
```
summary(df$crim)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

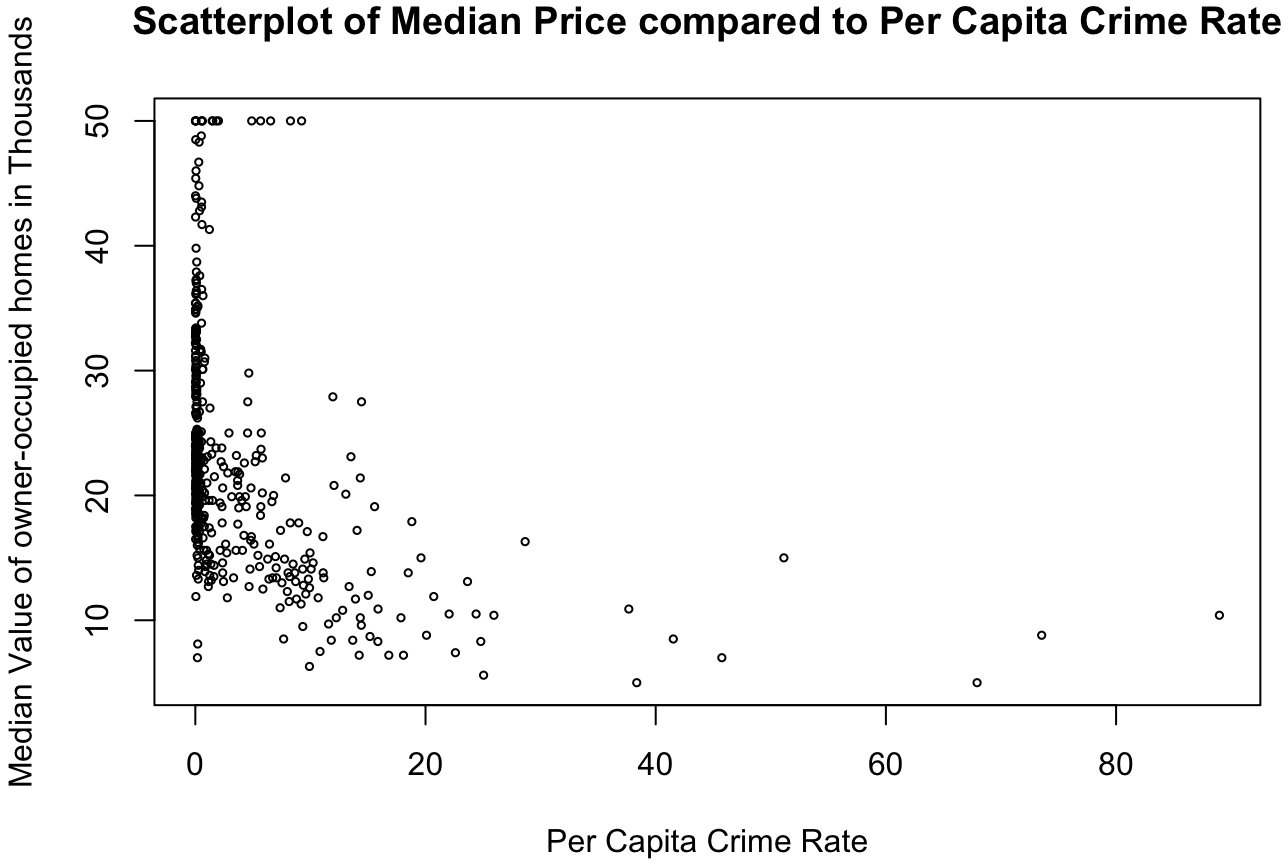
```
crim_hist <- hist(df$crim, main = "Histogram of Per Capita Crime Rate", xlab = "Per Capi
ta Crime Rate")
xfit <- seq(min(df$crim), max(df$crim), length = 100)
yfit <- dnorm(xfit, mean = mean(df$crim), sd = sd(df$crim))
yfit <- yfit * diff(crim_hist$mids[1:2]) * length(df$crim)
lines(xfit, yfit, col = "red", lwd = 2)
```




```
boxplot(df$crim, main = "Boxplot of Per Capita Crime Rate", ylab = "Per Capita Crime Rat
e")
```




```
plot(df$crim, df$medv, main = "Scatterplot of Median Price compared to Per Capita Crime
Rate", xlab = "Per Capita Crime Rate", ylab = "Median Value of owner-occupied homes in T
housands", cex=0.5)
```



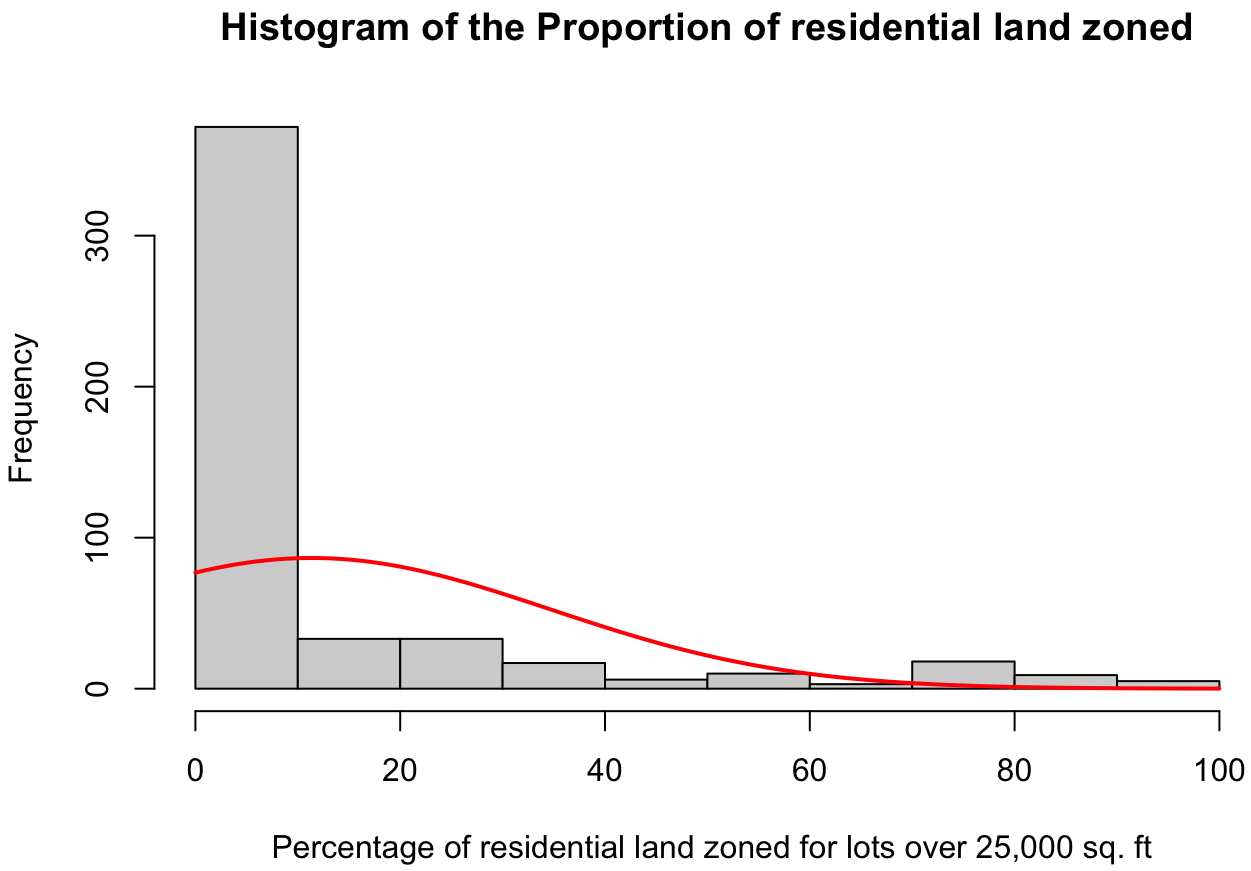
Comments: The Crim variable refers to the per capita crime rate by town, one of the predictors in the dataset. From the histogram, we can clearly see that the data is majorly right skewed, with one major peak between 0-10 per capita crime rate. There seem to be many outliers towards the right side of the histogram. With the fitted red line normal curve, that doesn't really give us too much information with the massive skewness. Looking at the box plot, we can see the Interquartile range box, with the median well at the front, also suggesting the major right skewness of the data, and we can see the large number of outliers in the data. Looking at the 5 number summary, we have the median and mean at 0.25651 and 3.61352, and we have the IQR from 0.08204 to 3.67708. And we can see that the range is very big. Looking at the scatterplot, we can start to see a trend of as the crime rate increases, the median house value decreases. We can see here that the range is quite large, around 88 crimes per capita, however most observations have a rate under 0.1.

ZN

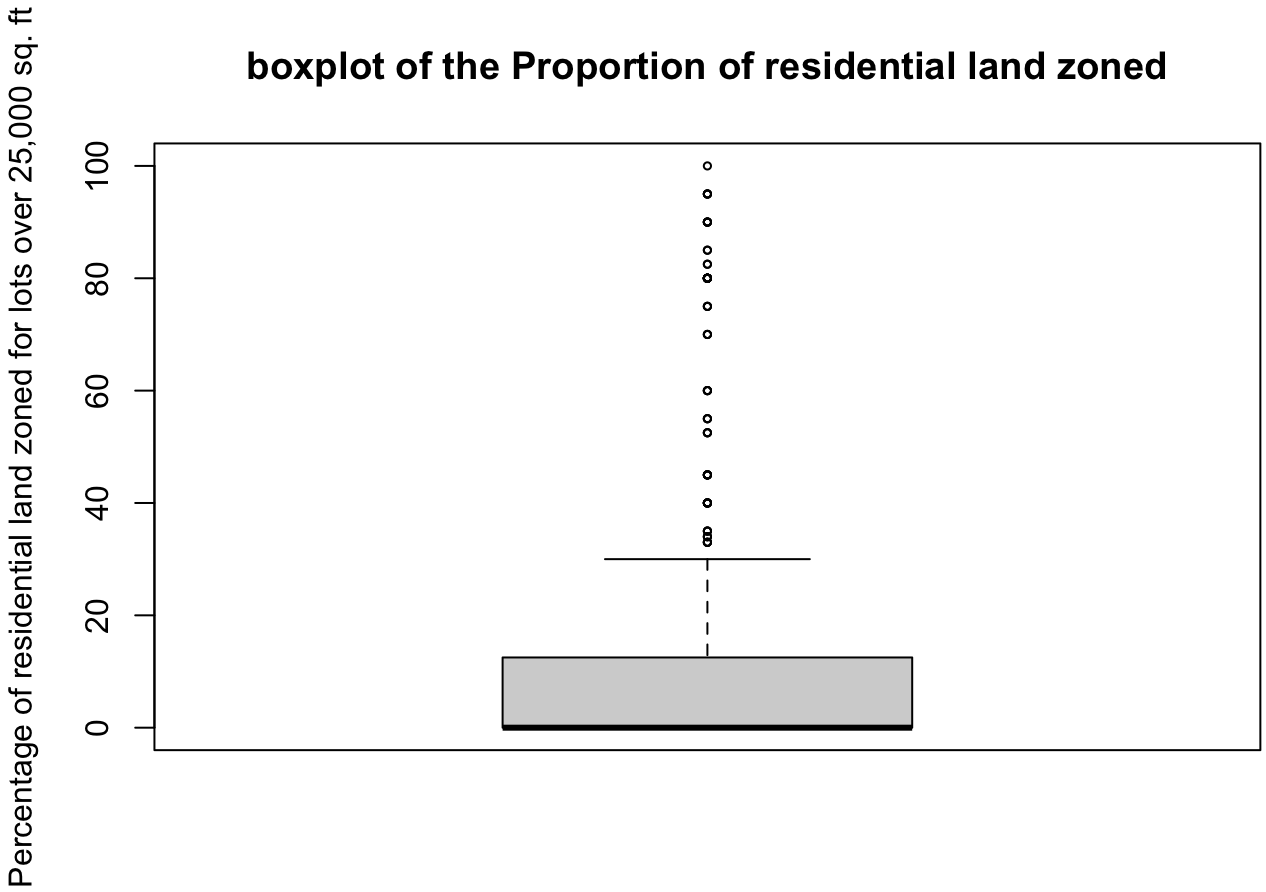
```
summary(df$zn)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   11.36   12.50  100.00
```

```
zn_hist <- hist(df$zn, main = "Histogram of the Proportion of residential land zoned", x
lab = " Percentage of residential land zoned for lots over 25,000 sq. ft")
xfit <- seq(min(df$zn), max(df$zn), length = 100)
yfit <- dnorm(xfit, mean = mean(df$zn), sd = sd(df$zn))
yfit <- yfit * diff(zn_hist$mids[1:2]) * length(df$zn)
lines(xfit, yfit, col = "red", lwd = 2)
```




```
boxplot(df$zn, main = "boxplot of the Proportion of residential land zoned", ylab = "Per
centage of residential land zoned for lots over 25,000 sq. ft", cex=0.5)
```




```
plot(df$zn, df$medv, main = "Scatterplot of Median Price compared to the Proportion of r
esidential land zoned", xlab = "Percentage of residential land zoned", ylab = "Median Va
lue of owner-occupied homes in Thousands", cex=0.5,
cex.main=0.90)
```

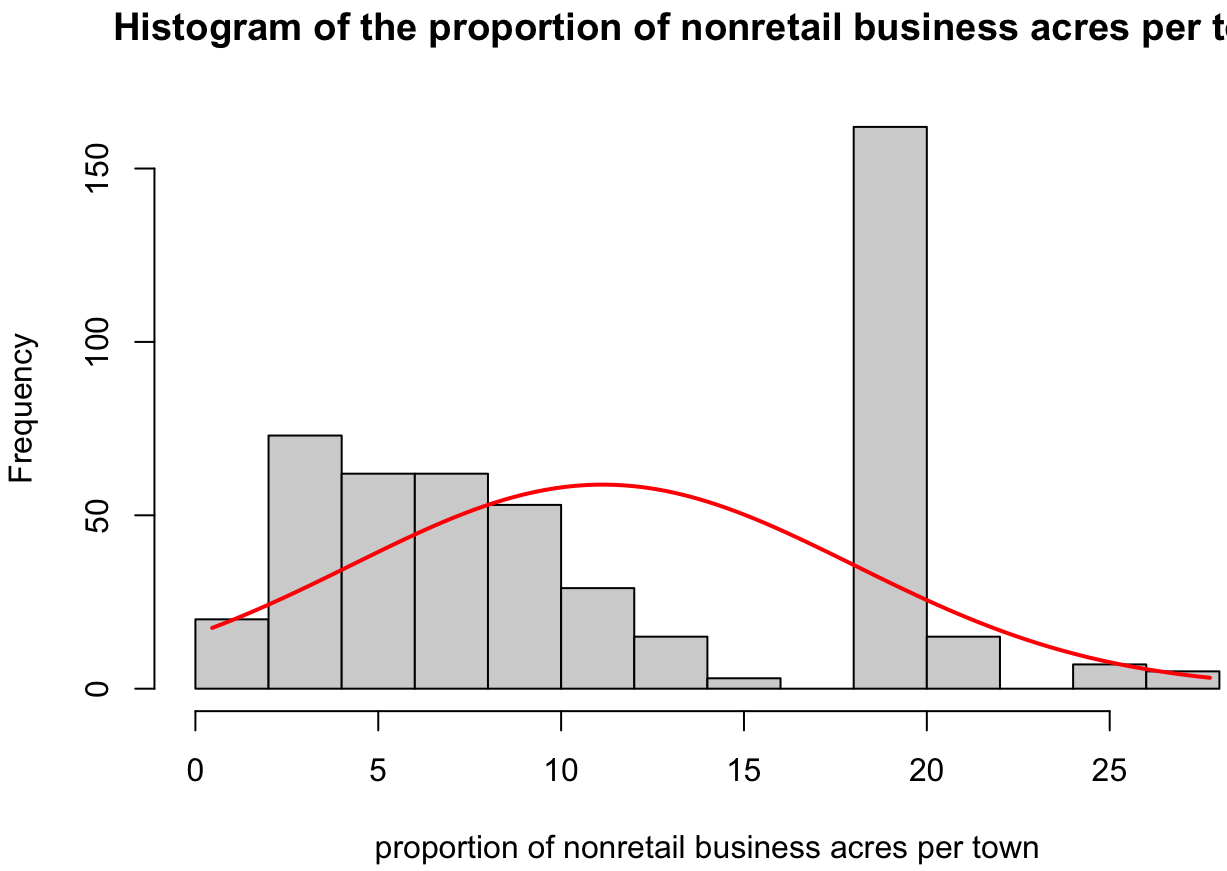
Comments: The Zm variable refers to the proportion of residential land zoned for lots over 25,000 sq.ft, one of the predictors in the dataset. From the histogram, we can clearly see that the data is majorly right skewed, with one major peak between 0-10 percent. There seem to be many outliers towards the right side of the histogram. With the fitted normal red curve, we can see that the data doesn't resemble a normal distribution at all, with heavy right skewness. Looking at the box plot, we can see the Interquartile range box, with the median well at the front, also suggesting the major right skewness of the data, and we can see the large number of outliers in the data. Looking at the 5 number summary, we have the median and mean at 0 and 11.36%, and we have the IQR from 0 to 12.50 percent And we can see that the range is very big, encompassing areas that have none and all the land home to houses with an area above 25,000 square feet. Looking at the scatterplot, there isn't too much of a trend, however there is a slight positive association that as the proportion of residential land for houses over 25,000 sq ft increase, the median house value increases. Seeing how the median is at 0 is also indication that there are a lot of houses in areas where there are no houses above 25,000 sq ft.

INDUS

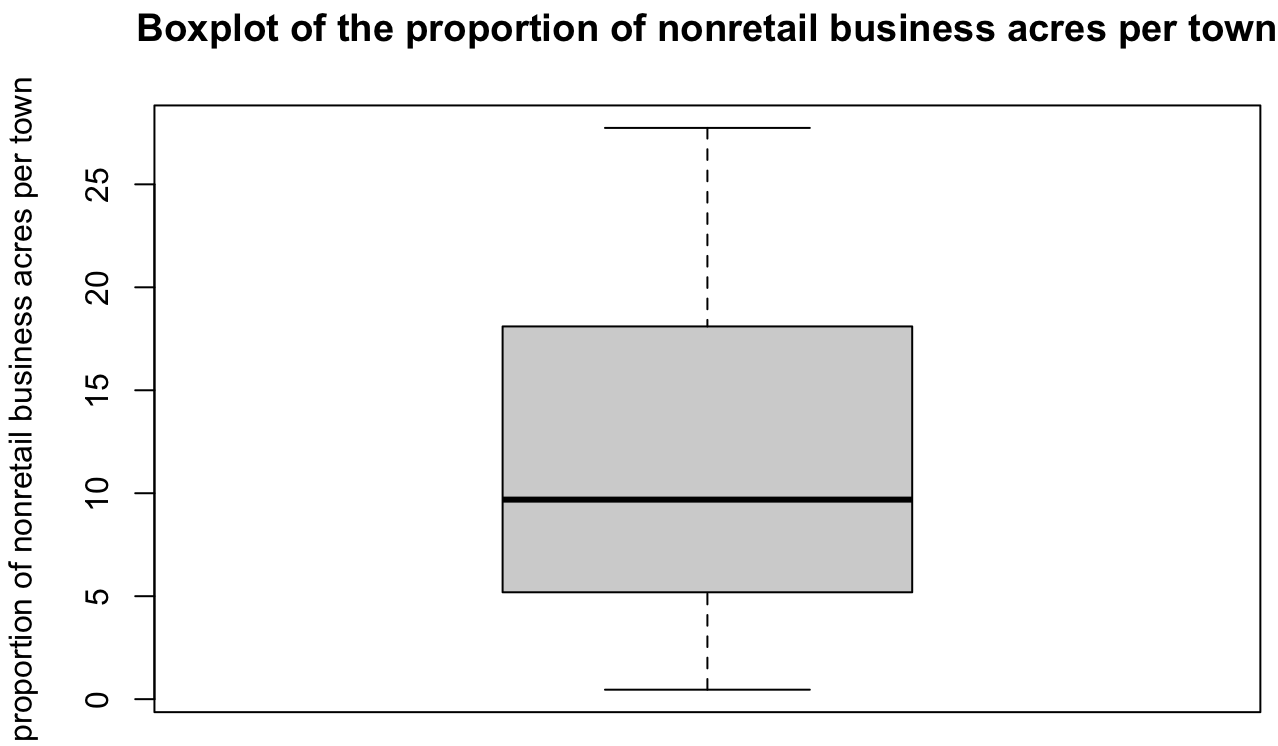
```
summary(df$indus)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.46    5.19    9.69   11.14   18.10   27.74
```

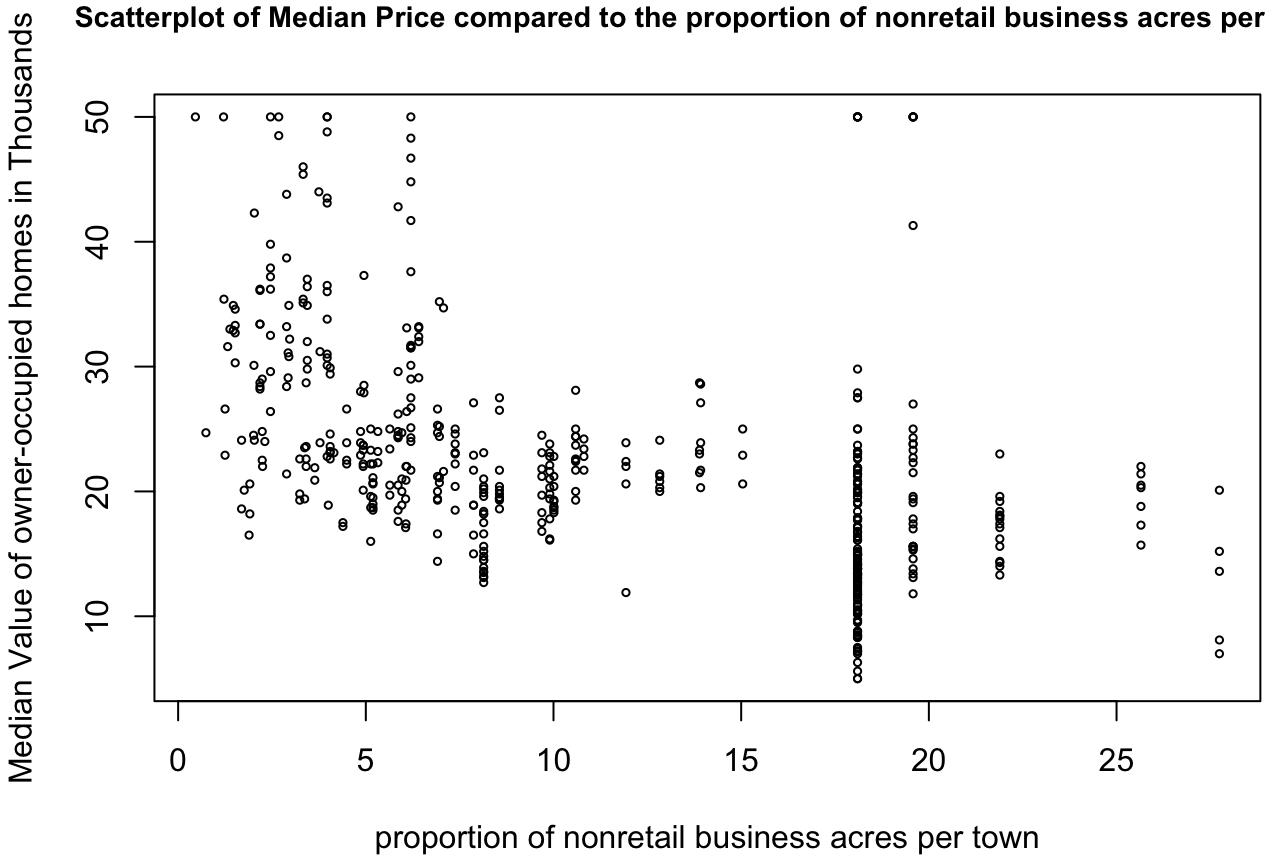
```
indus_hist <- hist(df$indus, main = "Histogram of the proportion of nonretail business a
cres per town", xlab = "proportion of nonretail business acres per town")
xfit <- seq(min(df$indus), max(df$indus), length = 100)
yfit <- dnorm(xfit, mean = mean(df$indus), sd = sd(df$indus))
yfit <- yfit * diff(indus_hist$mids[1:2]) * length(df$indus)
lines(xfit, yfit, col = "red", lwd = 2)
```




```
boxplot(df$indus, main = "Boxplot of the proportion of nonretail business acres per tow
n", ylab = "proportion of nonretail business acres per town", cex=0.5)
```




```
plot(df$indus, df$medv, main = "Scatterplot of Median Price compared to the proportion o
f nonretail business acres per town", xlab = "proportion of nonretail business acres per
town", ylab = "Median Value of owner-occupied homes in Thousands", cex=0.5, cex.main=0.9
0)
```



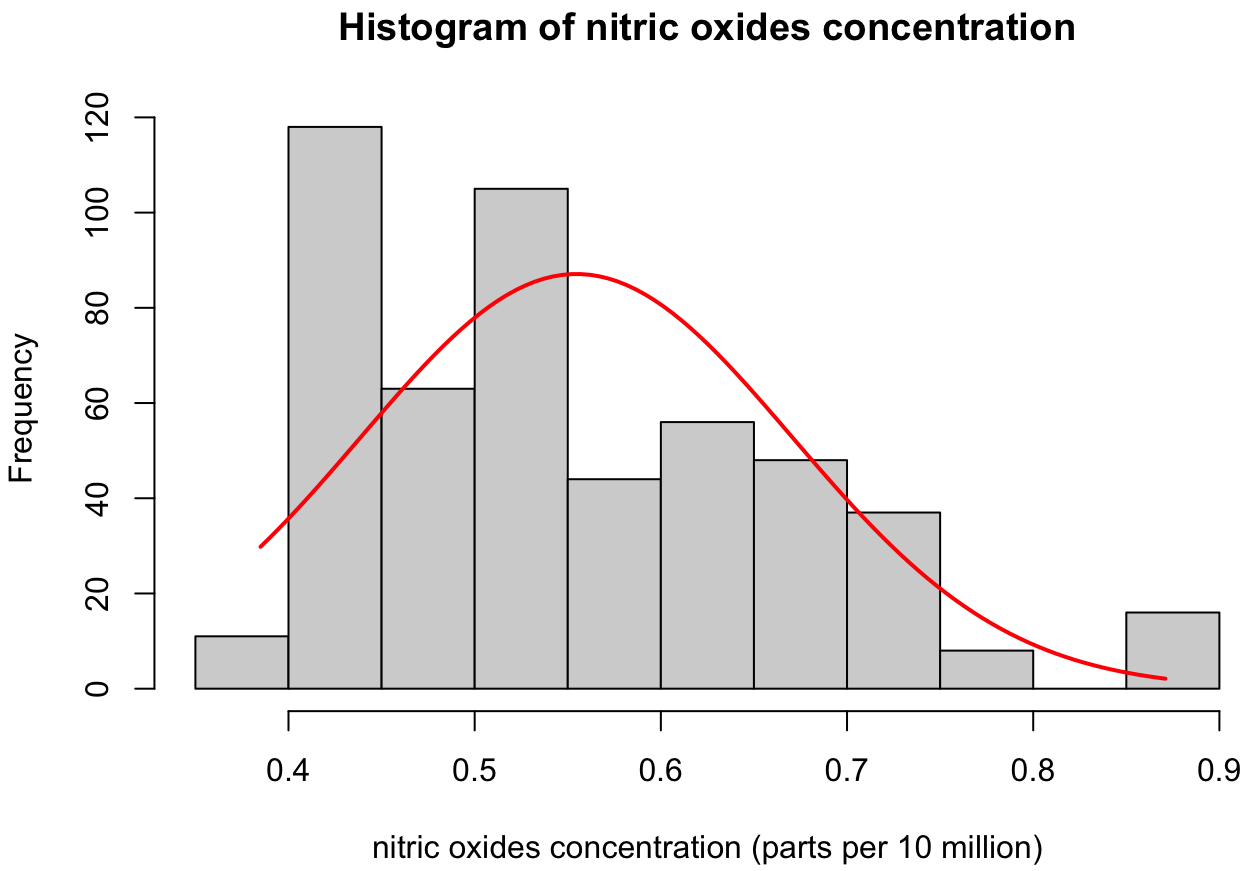
Comments: The indus variable refers to the proportion of non retail business acres per town, in percentage, one of the predictors in the dataset. From the histogram, the data looks slightly rightly skewed, with one major peak of between 15-20 percent. There doesn't seem to be too many outliers at first glance from the histogram. With the fitted normal red curve, we can see the data doesn't really resemble a normal distribution, however it seems to be more symmetric than the previous variables. Looking at the box plot, we can see the Interquartile range box, with the median not completely in the middle, also suggesting the major right skewness of the data, and we can see that there are no outliers in the data. Looking at the 5 number summary, we have the median and mean at 9.69 and 11.14 percent, and we have the IQR from 5.19 to 18.10. Looking at the scatterplot, we can start to see a slight trend of as the percentage of nonretail business acres per town increase, the median house value decreases.

NOX

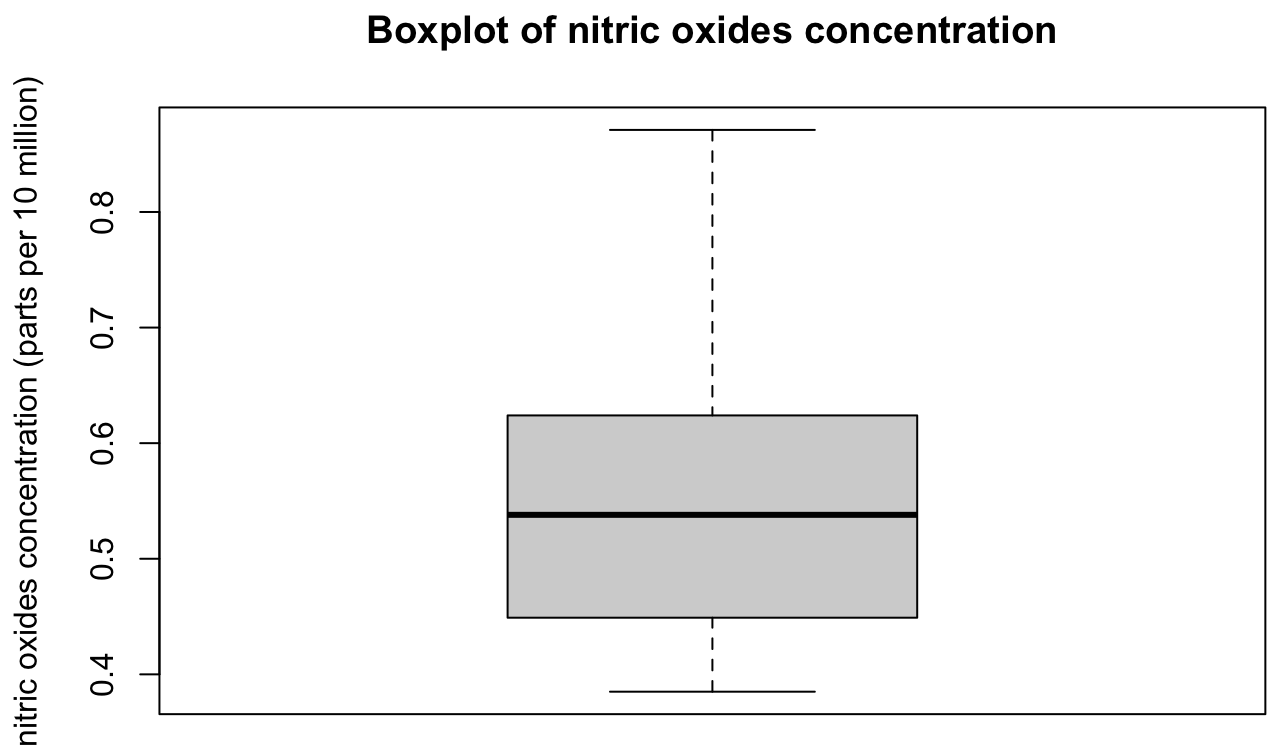
```
summary(df$nox)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3850  0.4490  0.5380  0.5547  0.6240  0.8710
```

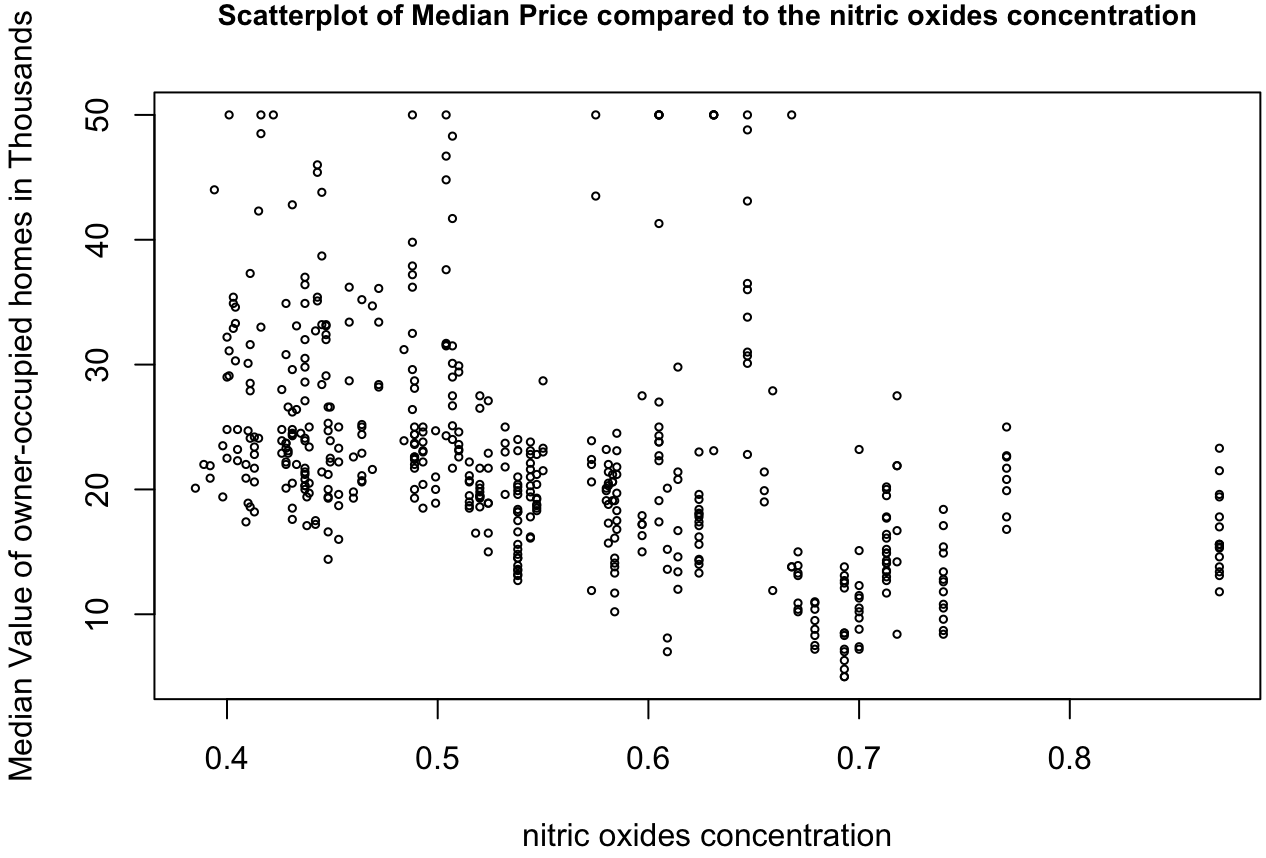
```
nox_hist <- hist(df$nox, main = "Histogram of nitric oxides concentration", xlab = "nitr
ic oxides concentration (parts per 10 million)")
xfit <- seq(min(df$nox), max(df$nox), length = 100)
yfit <- dnorm(xfit, mean = mean(df$nox), sd = sd(df$nox))
yfit <- yfit * diff(nox_hist$mids[1:2]) * length(df$nox)
lines(xfit, yfit, col = "red", lwd = 2)
```




```
boxplot(df$nox, main = "Boxplot of nitric oxides concentration", ylab = "nitric oxides c
oncentration (parts per 10 million)")
```




```
plot(df$nox, df$medv, main = "Scatterplot of Median Price compared to the nitric oxides
concentration", xlab = "nitric oxides concentration", ylab = "Median Value of owner-occu
pied homes in Thousands", cex=0.5, cex.main=0.90)
```



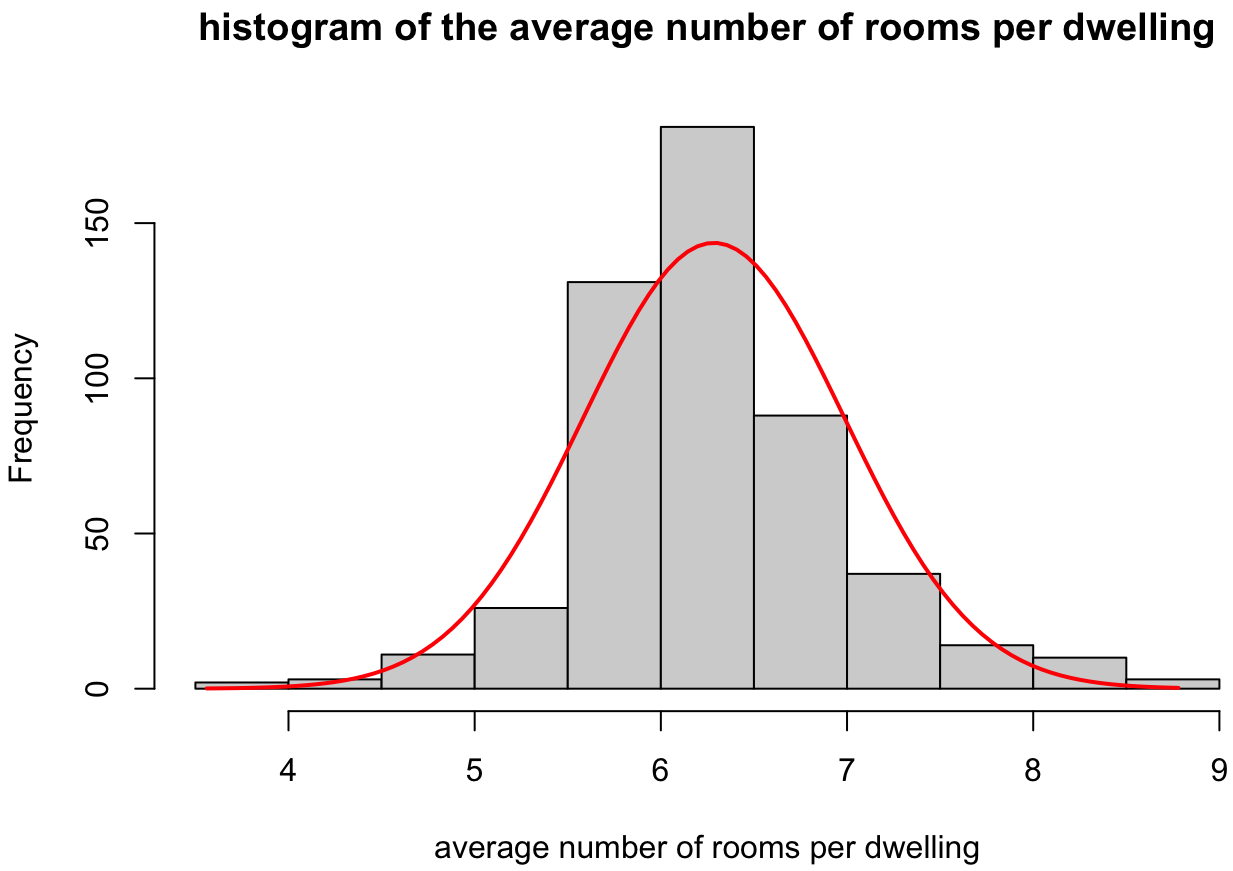
Comments: The nox variable refers to the nitric oxides concentration in parts per 10 million, one of the predictors in the dataset. From the histogram, the data looks relatively symmetrical, with some potential right skewness, with a couple of peaks from 0.4-0.45 and 0.5 to 0.55 . There doesn't seem to be too many outliers at first glance from the histogram except one observation towards the right side. With the fitted normal distribution curve, we can see that the curve aligns somewhat well with the data, but not completely, suggesting some symmetric nature. Looking at the box plot, we can see the Interquartile range box, with the median essentially right in the middle, also suggesting symmetrical data, and we can see that there are no outliers in the data. Looking at the 5 number summary, we have the median and mean at 0.538 and 0.5547 parts per 10 million, and we have the IQR from 0.449 to 0.624 Looking at the scatterplot, we can start to see a trend of as the nitric oxides concentration increases, the median house value decreases.

RM

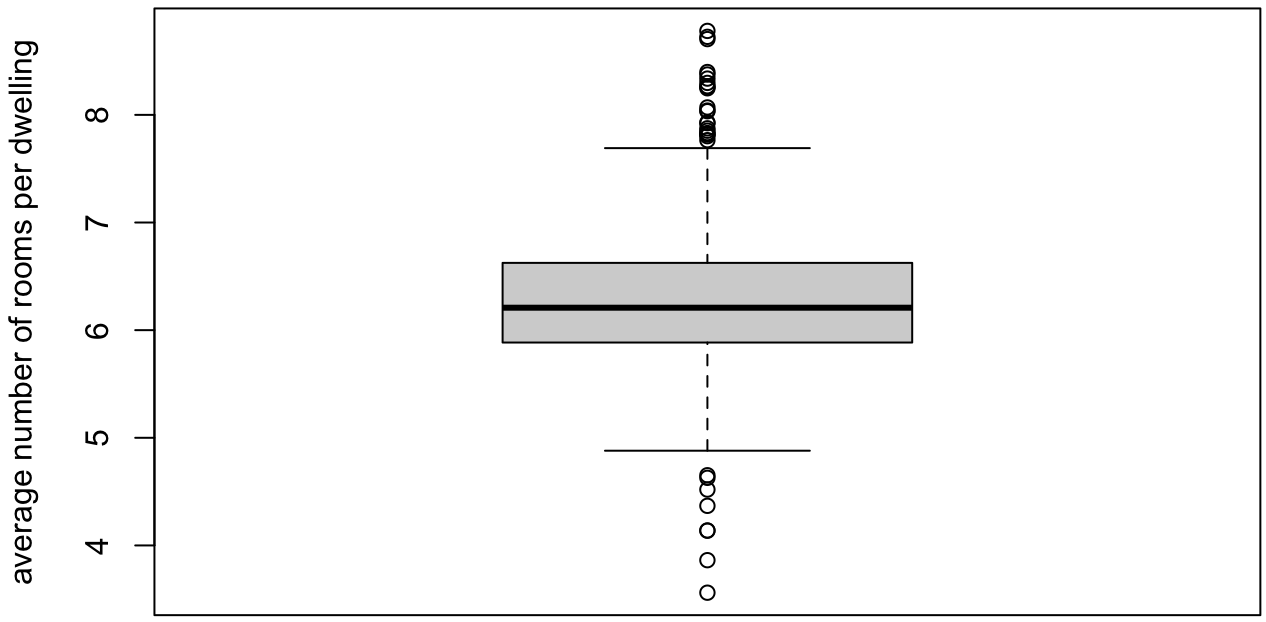
```
summary(df$rm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.561   5.886   6.208   6.285   6.623   8.780
```

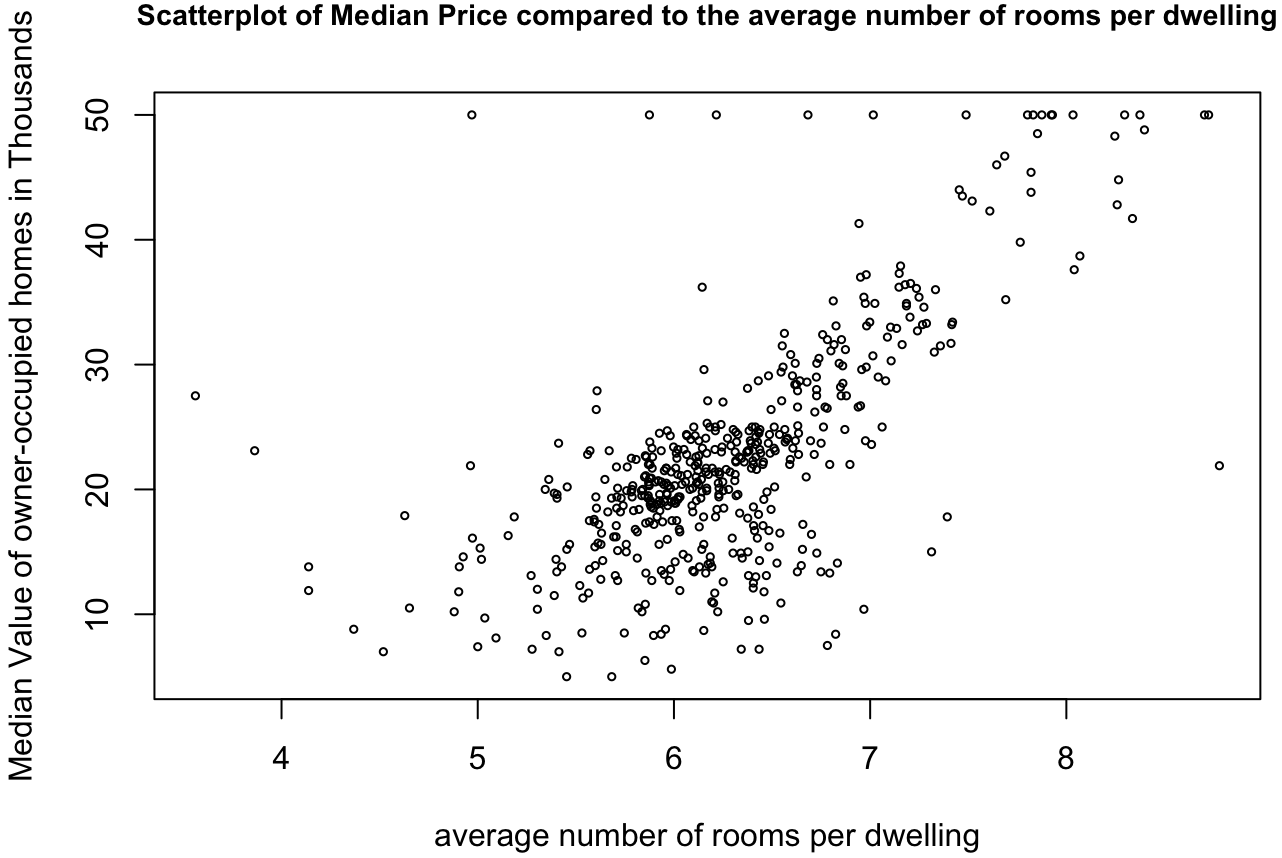
```
rm_hist <- hist(df$rm, main = "histogram of the average number of rooms per dwelling", x
lab = "average number of rooms per dwelling")
xfit <- seq(min(df$rm), max(df$rm), length = 100)
yfit <- dnorm(xfit, mean = mean(df$rm), sd = sd(df$rm))
yfit <- yfit * diff(rm_hist$mids[1:2]) * length(df$rm)
lines(xfit, yfit, col = "red", lwd = 2)
```




```
boxplot(df$rm, main = "boxplot of the average number of rooms per dwelling", ylab = "ave
rage number of rooms per dwelling")
```




```
plot(df$rm, df$medv, main = "Scatterplot of Median Price compared to the average number
of rooms per dwelling", xlab = "average number of rooms per dwelling", ylab = "Median Va
lue of owner-occupied homes in Thousands", cex=0.5, cex.main=0.90)
```



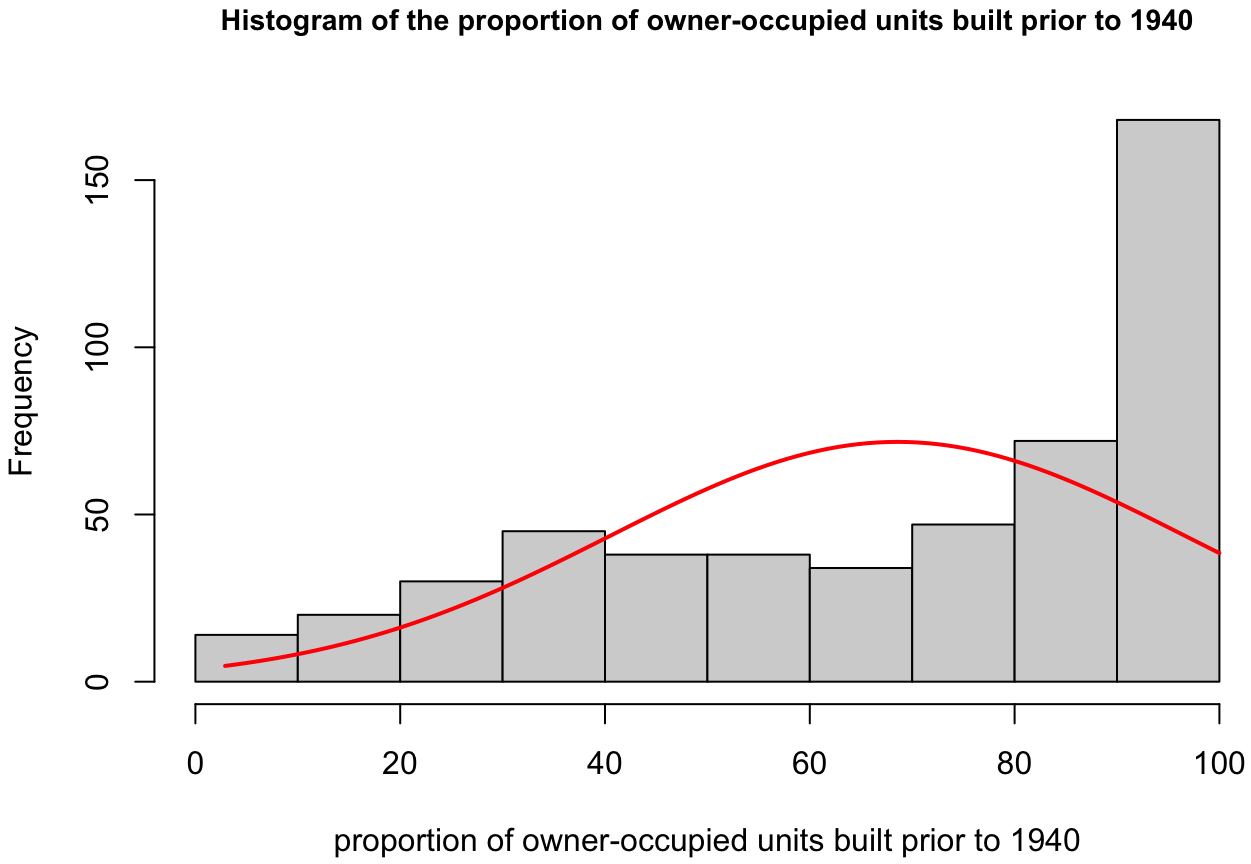
Comments: The rm variable refers to the average number of rooms per dwelling, one of the predictors in the dataset. From the histogram, the data looks relatively symmetrical with a hint of right skewness, with one main peak in between 6-7 rooms. However, there seems to be a number outliers at first glance from the histogram as the distribution is quite narrow towards both sides. With the fitted normal distribution curve, we can see that the curve actually aligns really well with the data, potentially suggesting a symmetric distribution. Looking at the box plot, we can see the Interquartile range box, with the median slighly towards the bottom side, suggesting some potential right skewness, and we can see that there are quite a number of outliers on both sides. Looking at the 5 number summary, we have the median and mean at 6.208 and 6.285 rooms, and we have the IQR from 5.886 to 6.623. Looking at the scatterplot, we can see a pretty clear trend, as the average number of rooms per dwelling increases, the median house value increases. This logically and economically makes sense.

AGE

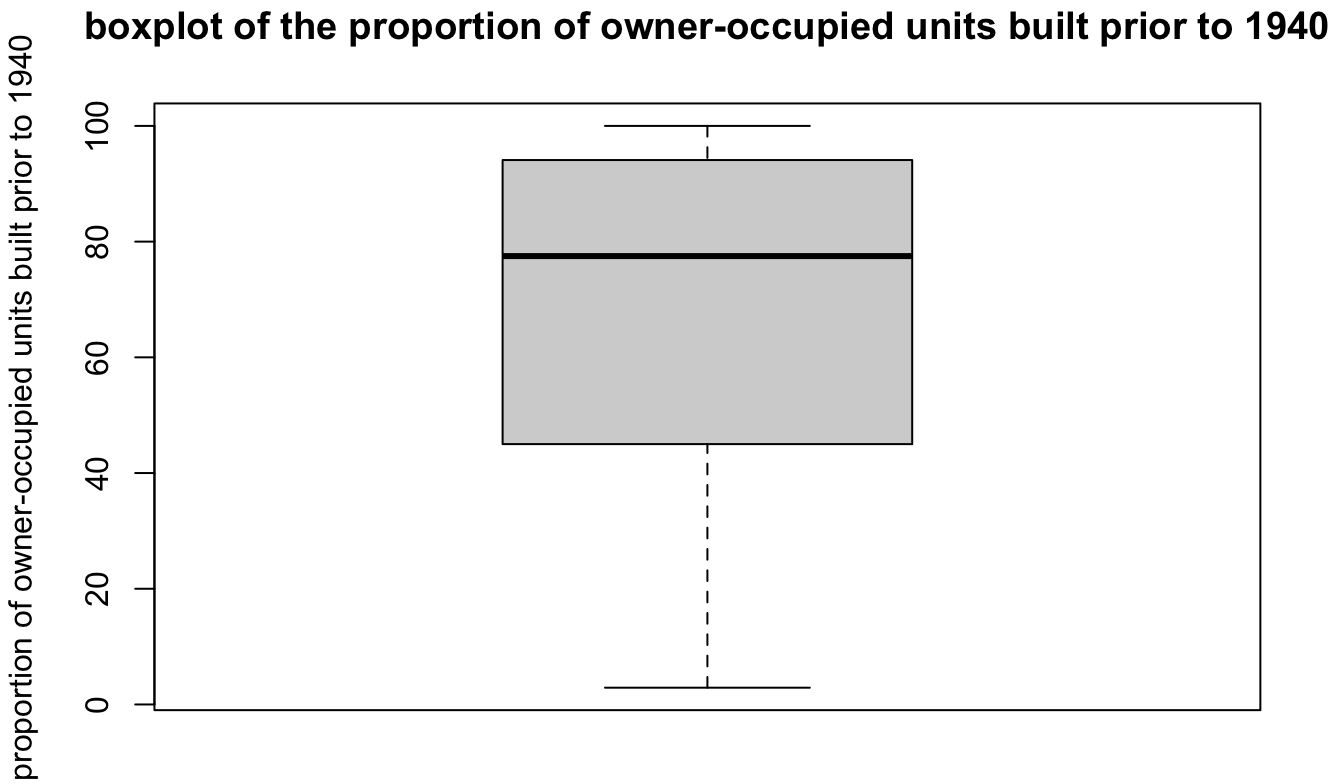
```
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.90   45.02   77.50   68.57   94.08  100.00
```

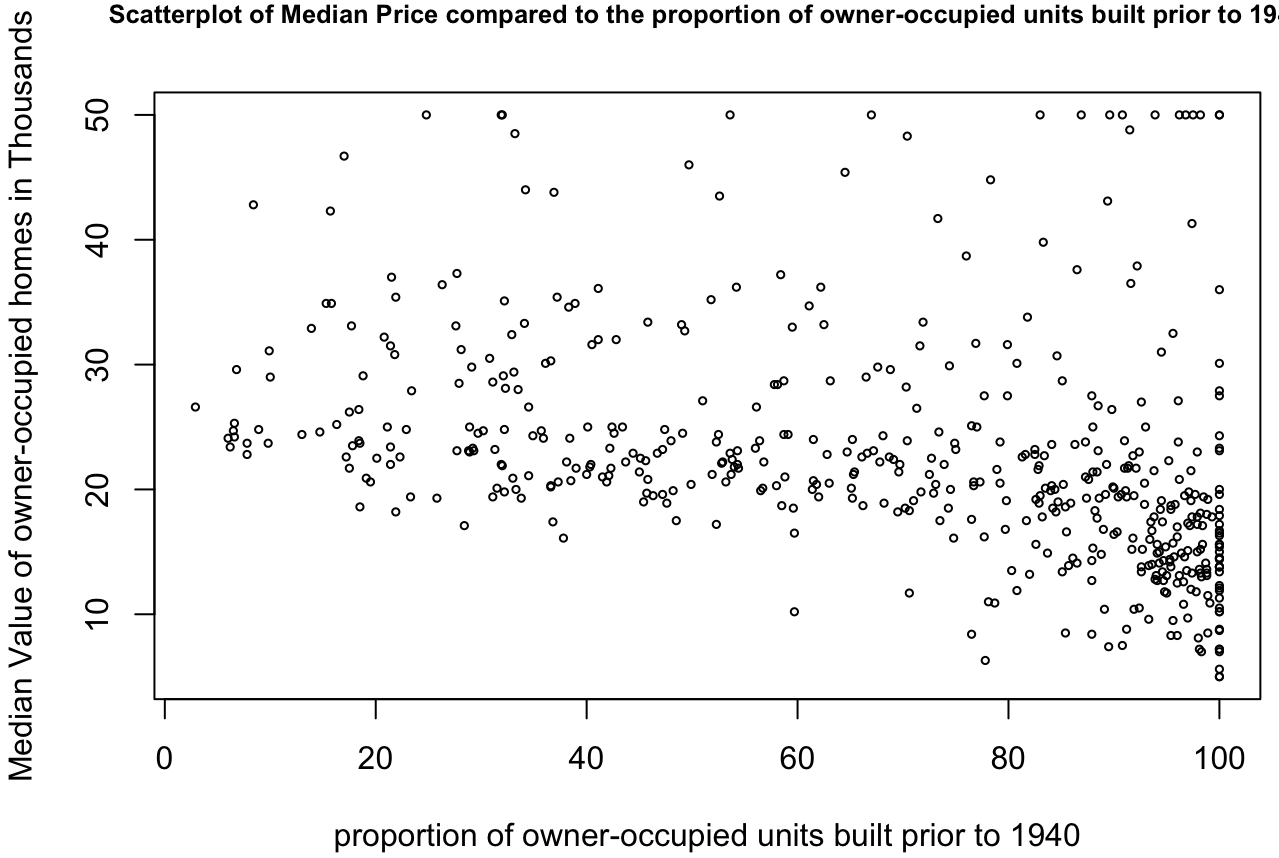
```
age_hist <- hist(df$age, main = "Histogram of the proportion of owner-occupied units bui
lt prior to 1940", xlab = "proportion of owner-occupied units built prior to 1940", cex.
main=0.9)
xfit <- seq(min(df$age), max(df$age), length = 100)
yfit <- dnorm(xfit, mean = mean(df$age), sd = sd(df$age))
yfit <- yfit * diff(age_hist$mids[1:2]) * length(df$age)
lines(xfit, yfit, col = "red", lwd = 2)
```




```
boxplot(df$age, main = "boxplot of the proportion of owner-occupied units built prior to
1940", ylab = "proportion of owner-occupied units built prior to 1940", cex=0.5)
```




```
plot(df$age, df$medv, main = "Scatterplot of Median Price compared to the proportion of
owner-occupied units built prior to 1940", xlab = "proportion of owner-occupied units bu
ilt prior to 1940", ylab = "Median Value of owner-occupied homes in Thousands", cex=0.5,
cex.main=0.80)
```



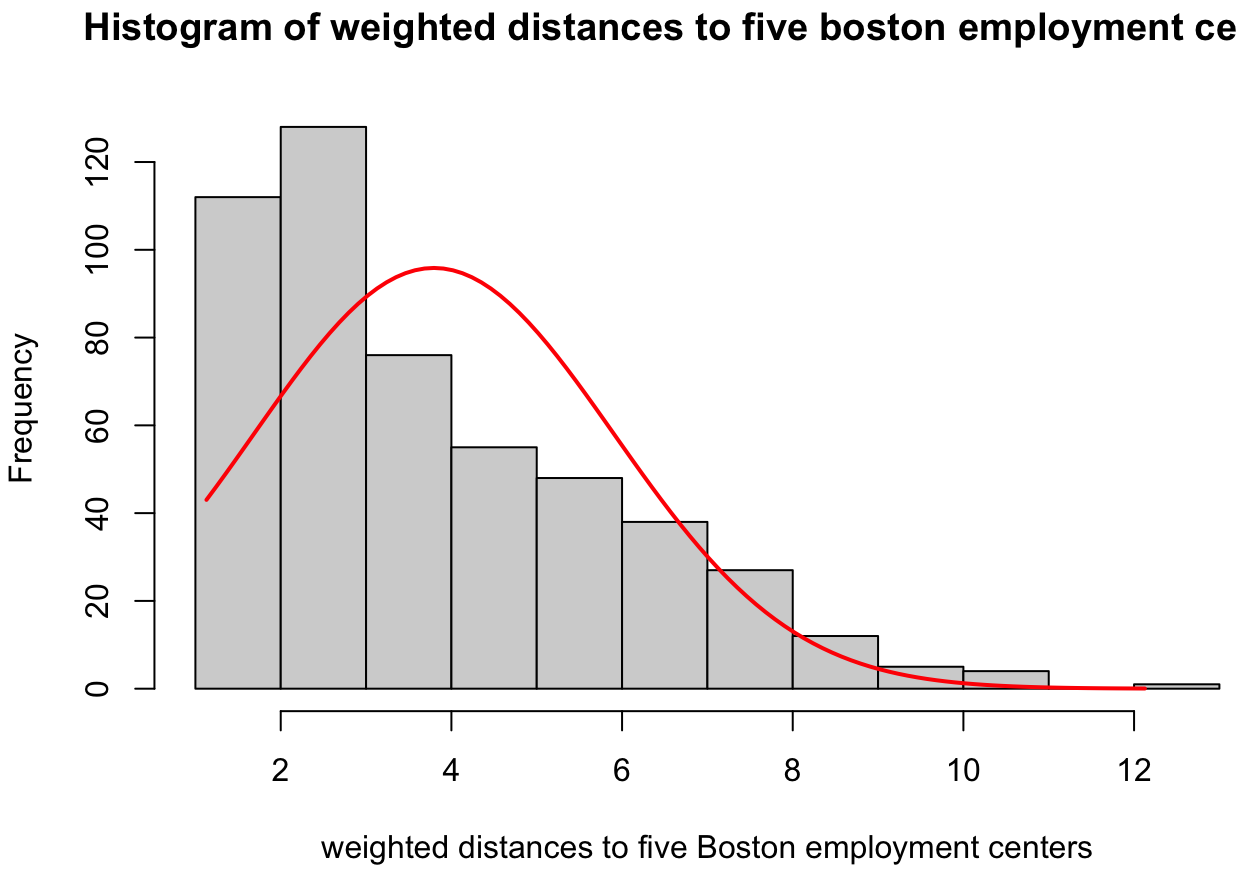
Comments: The age variable refers to the proportion of owner-occupied units built prior to 1940, one of the predictors in the dataset. From the histogram, the data looks heavily left skewed, with the main peak all the way on the right side between 80-100 percent. There doesn't seem to be too many outliers at first glance from the histogram. With the fitted normal distribution curve, we can seee that the curve does not really align well with the data, with the massive skewness. Looking at the box plot, we can see the Interquartile range box, with the median towards the top side, suggesting some left skewness, and we can see that there are no outliers in the data from the boxplot. Looking at the 5 number summary, we have the median and mean at 77.5 and 68.57 percent, and we have the IQR from 45.02 to 94.08 percent, quite a large range. Looking at the scatterplot, there isn't too much of a trend evident, but the slight trend seems to say that as the proportion of owner-occupied units built prior to 1940 increases, the median house value decreases.

DIS

```
summary(df$dis)
```

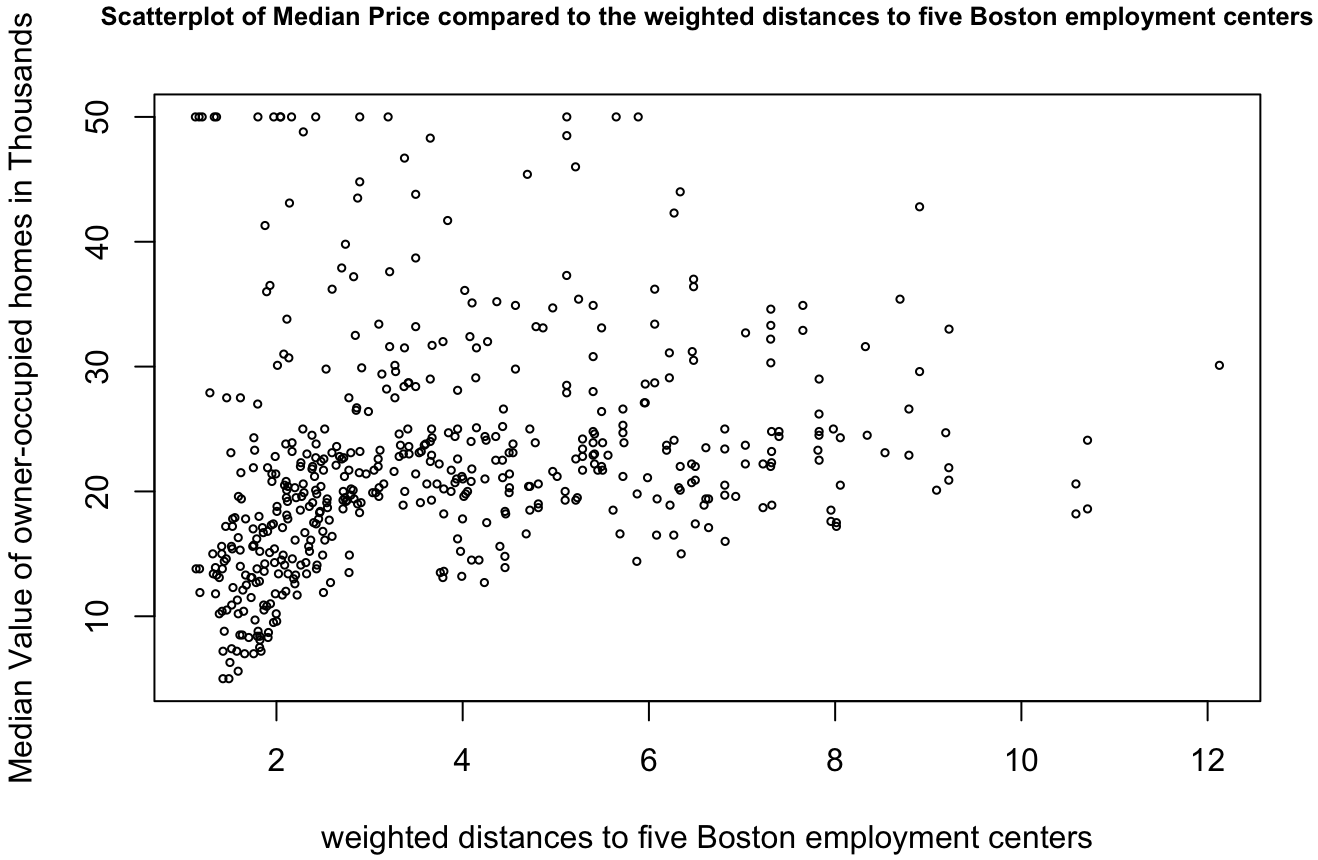
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.130   2.100   3.207   3.795   5.188  12.127
```

```
dis_hist <- hist(df$dis, main = "Histogram of weighted distances to five boston employme
nt centers", xlab = "weighted distances to five Boston employment centers")
xfit <- seq(min(df$dis), max(df$dis), length = 100)
yfit <- dnorm(xfit, mean = mean(df$dis), sd = sd(df$dis))
yfit <- yfit * diff(dis_hist$mids[1:2]) * length(df$dis)
lines(xfit, yfit, col = "red", lwd = 2)
```




```
boxplot(df$dis, main = "boxplot of weighted distances to five boston employment center
s", ylab = "weighted distances to five Boston employment centers", cex=0.75)
```


```
plot(df$dis, df$medv, main = "Scatterplot of Median Price compared to the weighted dista
nces to five Boston employment centers", xlab = "weighted distances to five Boston emplo
yment centers", ylab = "Median Value of owner-occupied homes in Thousands", cex=0.5, ce
x.main=0.80)
```



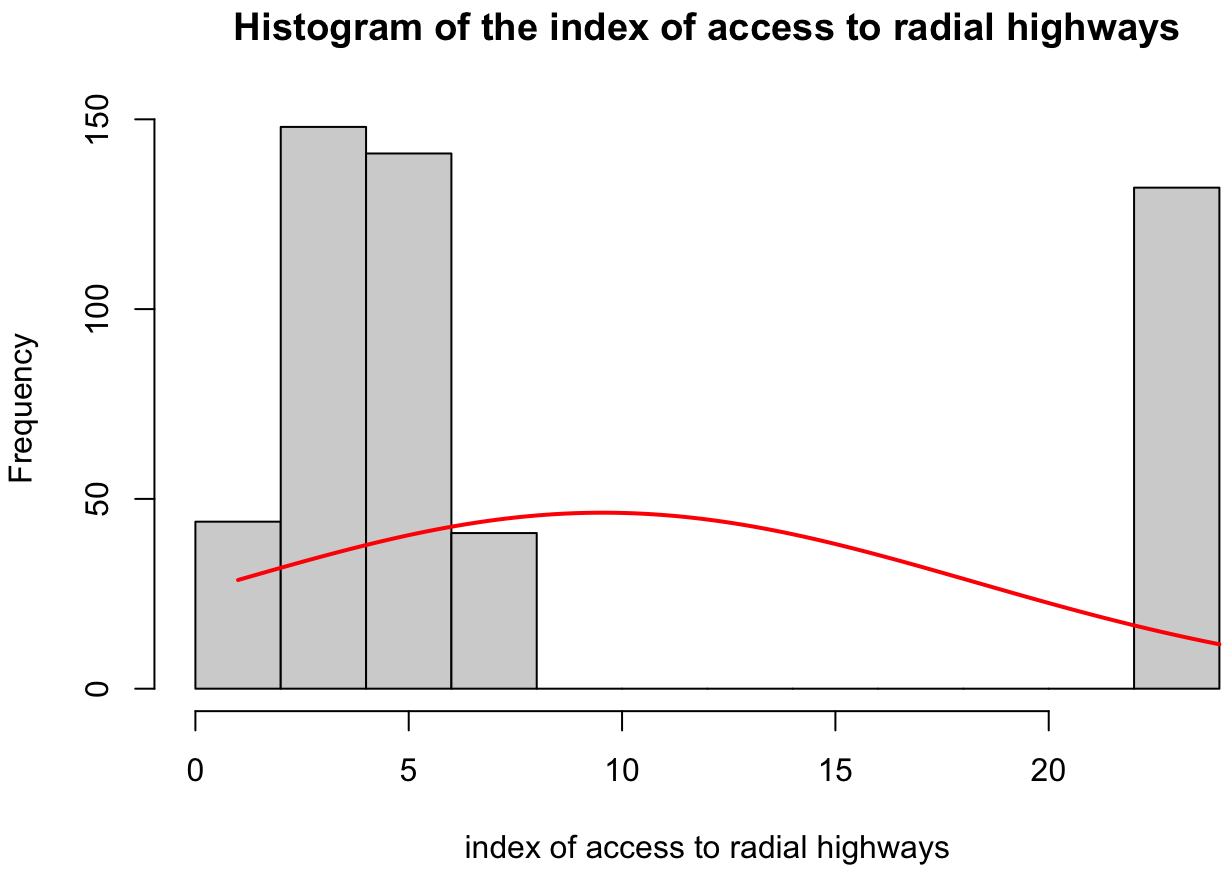
Comments: The dis variable refers to the weighted distances to five Boston emplyment centers one of the predictors in the dataset. From the histogram, the data looks right skewed, with the main peaks all the way on the left side between 0-3 miles. There seems to be outliers towards the right side of the histogram. With the fitted normal distribution curve, we can see that the data slightly fits the curve, however not really. Looking at the box plot, we can see the Interquartile range box, with the median towards the bottom side, suggesting some right skewness as well, and we can see that there are a couple outliers in the data on the higher side from the boxplot. Looking at the 5 number summary, we have the median and mean at 3.207 and 3.795 miles, and we have the IQR from 2.1 to 5.188 miles. Looking at the scatterplot, there isn't too much of a trend evident, but the slight trend seems to say that as the weighted distances to five boston employment centers increases, the median house value increases.

RAD

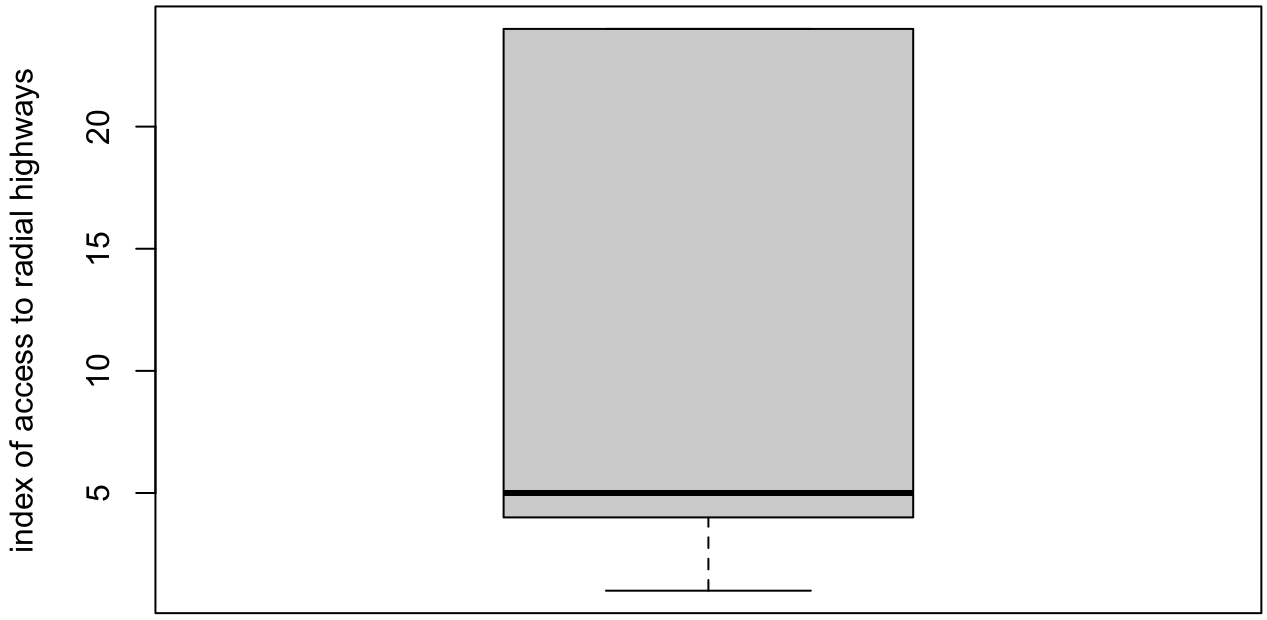
```
summary(df$rad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   5.000   9.549  24.000  24.000
```

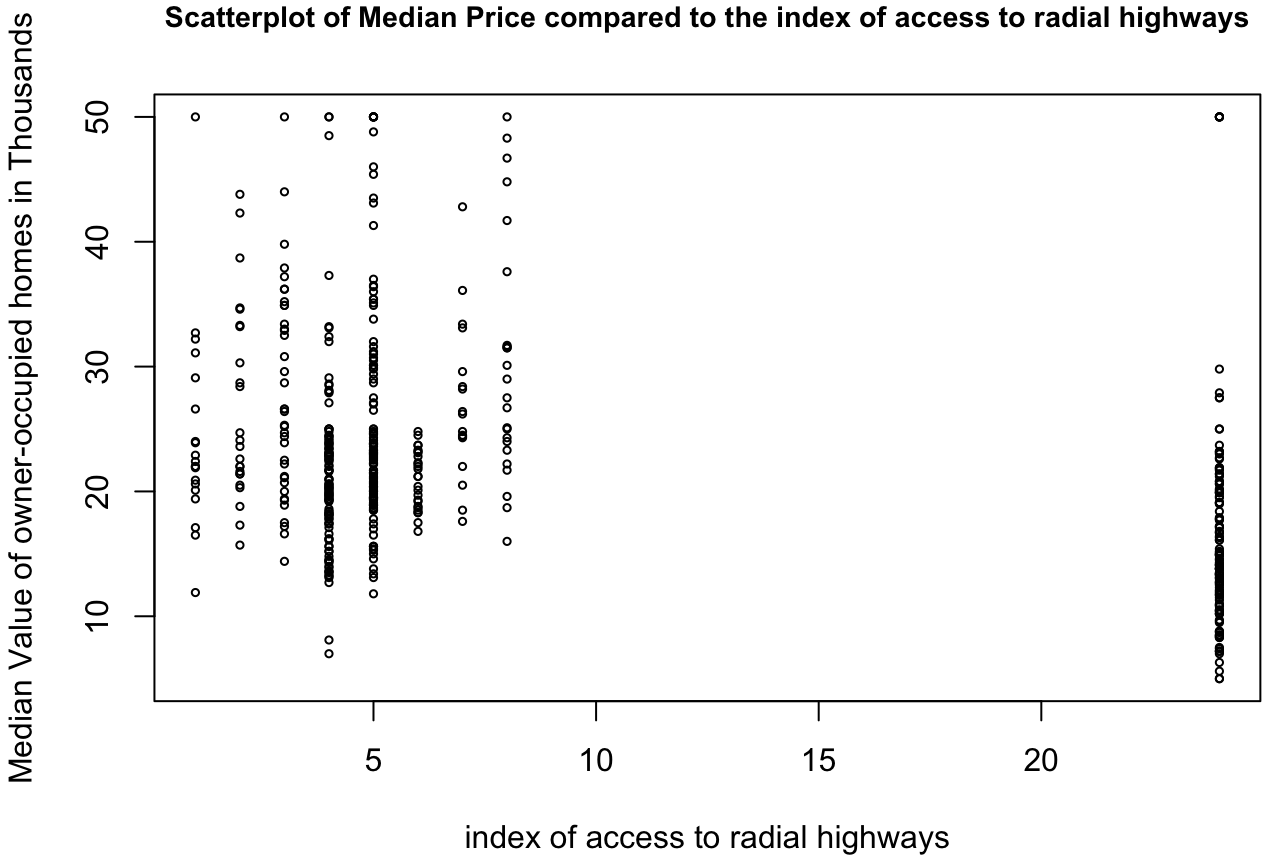
```
rad_hist <- hist(df$rad, main = "Histogram of the index of access to radial highways",
xlab = "index of access to radial highways")
xfit <- seq(min(df$rad), max(df$rad), length = 100)
yfit <- dnorm(xfit, mean = mean(df$rad), sd = sd(df$rad))
yfit <- yfit * diff(rad_hist$mids[1:2]) * length(df$rad)
lines(xfit, yfit, col = "red", lwd = 2)
```




```
boxplot(df$rad,  main = "Boxplot of the index of access to radial highways", ylab = "ind
ex of access to radial highways")
```




```
plot(df$rad, df$medv, main = "Scatterplot of Median Price compared to the index of acces
s to radial highways", xlab = "index of access to radial highways", ylab = "Median Value
of owner-occupied homes in Thousands", cex=0.5, cex.main=0.90)
```



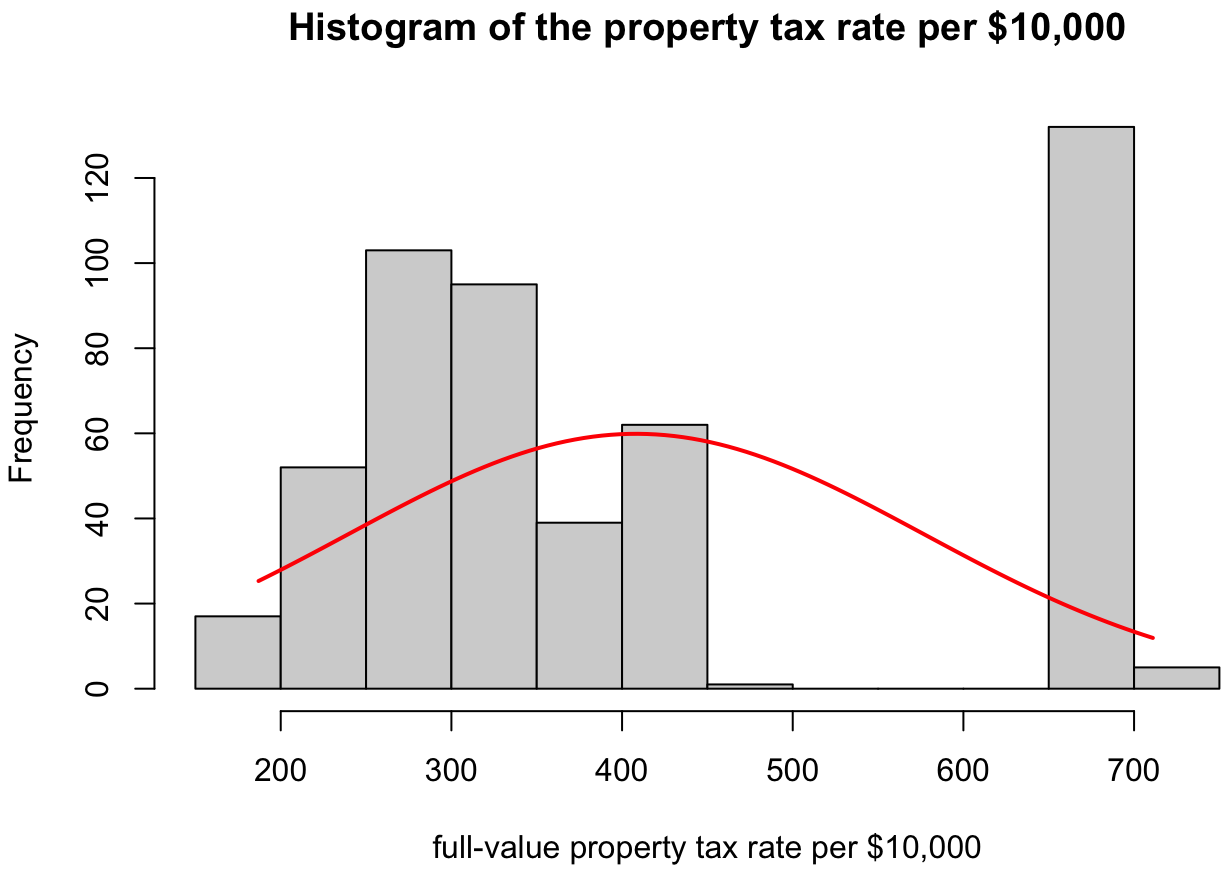
Comments: The rad variable refers to the index of access to radial highways, one of the predictors in the dataset. From the histogram, it's a little difficult to tell the skewness of the data, with such a massive gap and peaks beteen 2-7 and above 20. There seems to be outliers towards the right side of the histogram, however with the massive frequency, we are not certain. With the fitted normal distribution curve, we can see that the data does not really fit the normal curve. Looking at the box plot, we can see the Interquartile range box, with the median way towards the bottom side, suggesting a lot right skewness, and we can see that there are no outliers identified from the boxplot. Looking at the 5 number summary, we have the median and mean at 5 and 9.549, and we have the IQR from 4 to 24 miles. Looking at the scatterplot, the slight trend indicates that as the index of access to radial highways increases, the median house value decreases, however it is not clear with the massive gap and the volatility seeming to be constant.

TAX

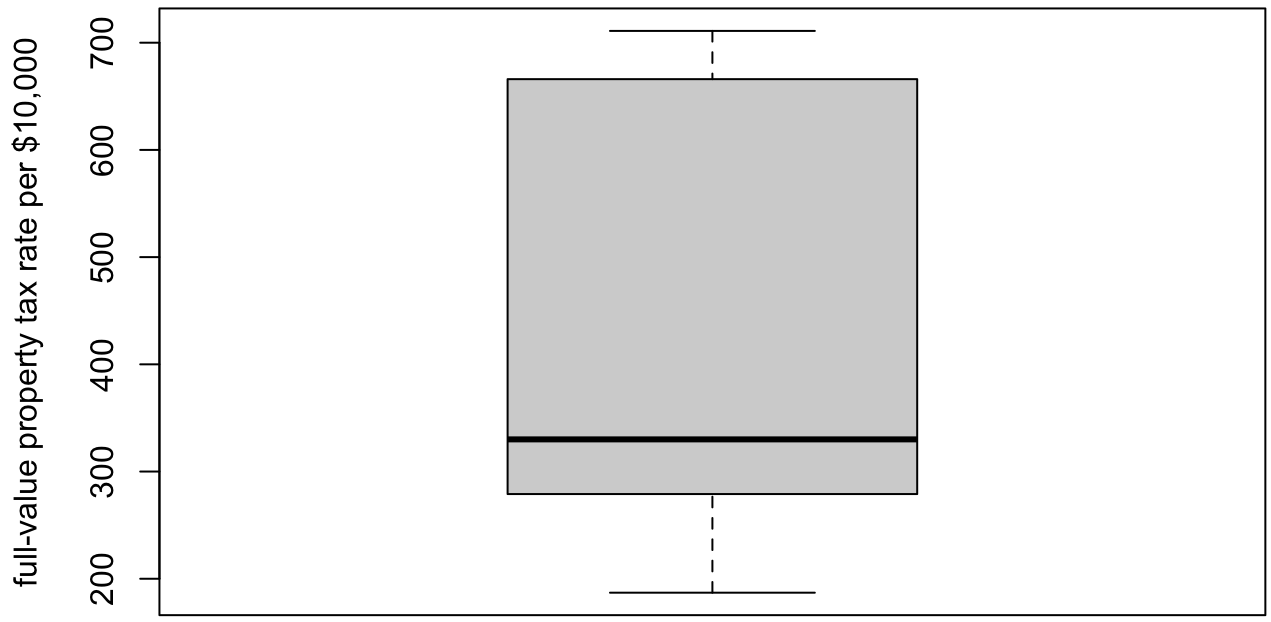
```
summary(df$tax)
```

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

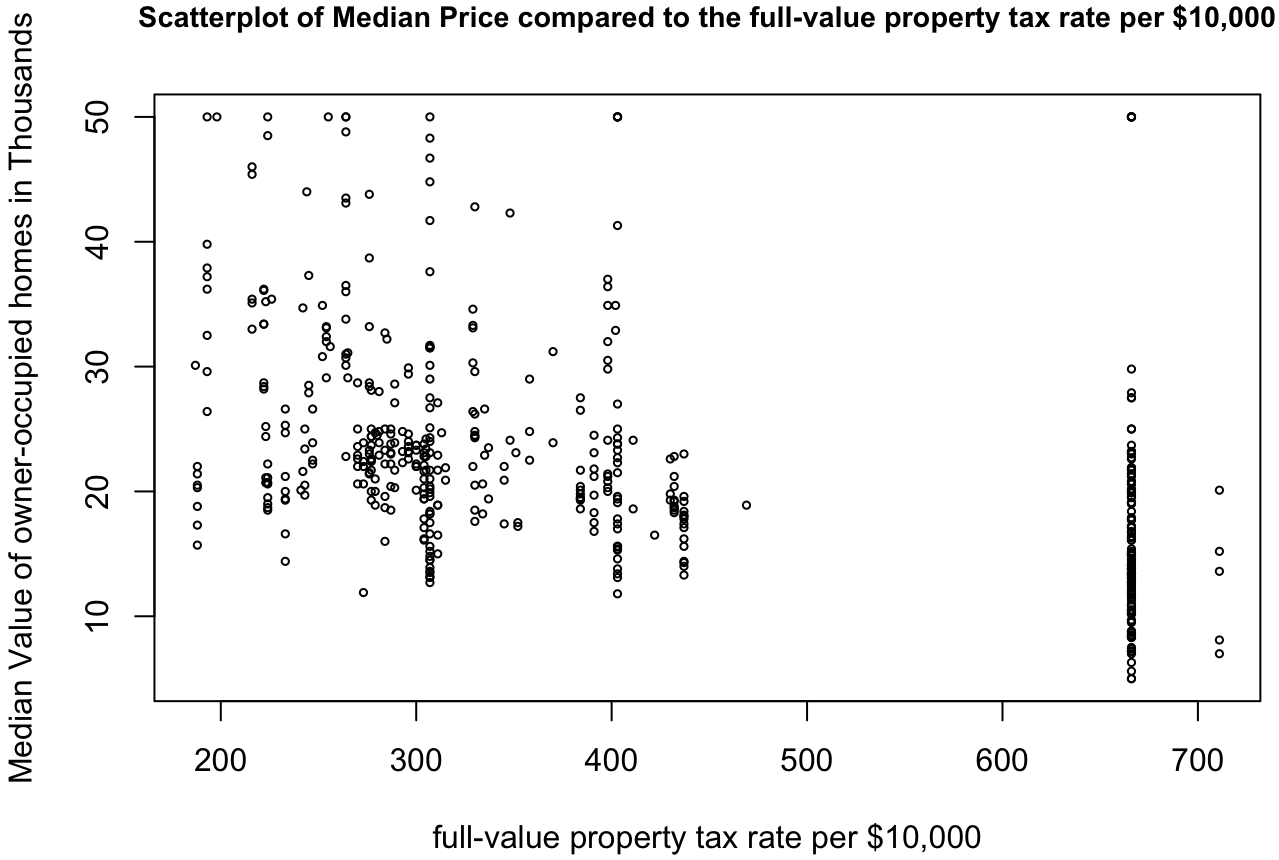
```
tax_hist <- hist(df$tax, main = "Histogram of the property tax rate per $10,000",  xlab
= "full-value property tax rate per $10,000")
xfit <- seq(min(df$tax), max(df$tax), length = 100)
yfit <- dnorm(xfit, mean = mean(df$tax), sd = sd(df$tax))
yfit <- yfit * diff(tax_hist$mids[1:2]) * length(df$tax)
lines(xfit, yfit, col = "red", lwd = 2)
```




```
boxplot(df$tax,  main = "Boxplot of the property tax rate per $10,000", ylab = "full-val
ue property tax rate per $10,000")
```




```
plot(df$tax, df$medv, main = "Scatterplot of Median Price compared to the full-value pro
perty tax rate per $10,000", xlab = "full-value property tax rate per $10,000", ylab =
"Median Value of owner-occupied homes in Thousands", cex=0.5, cex.main=0.90)
```



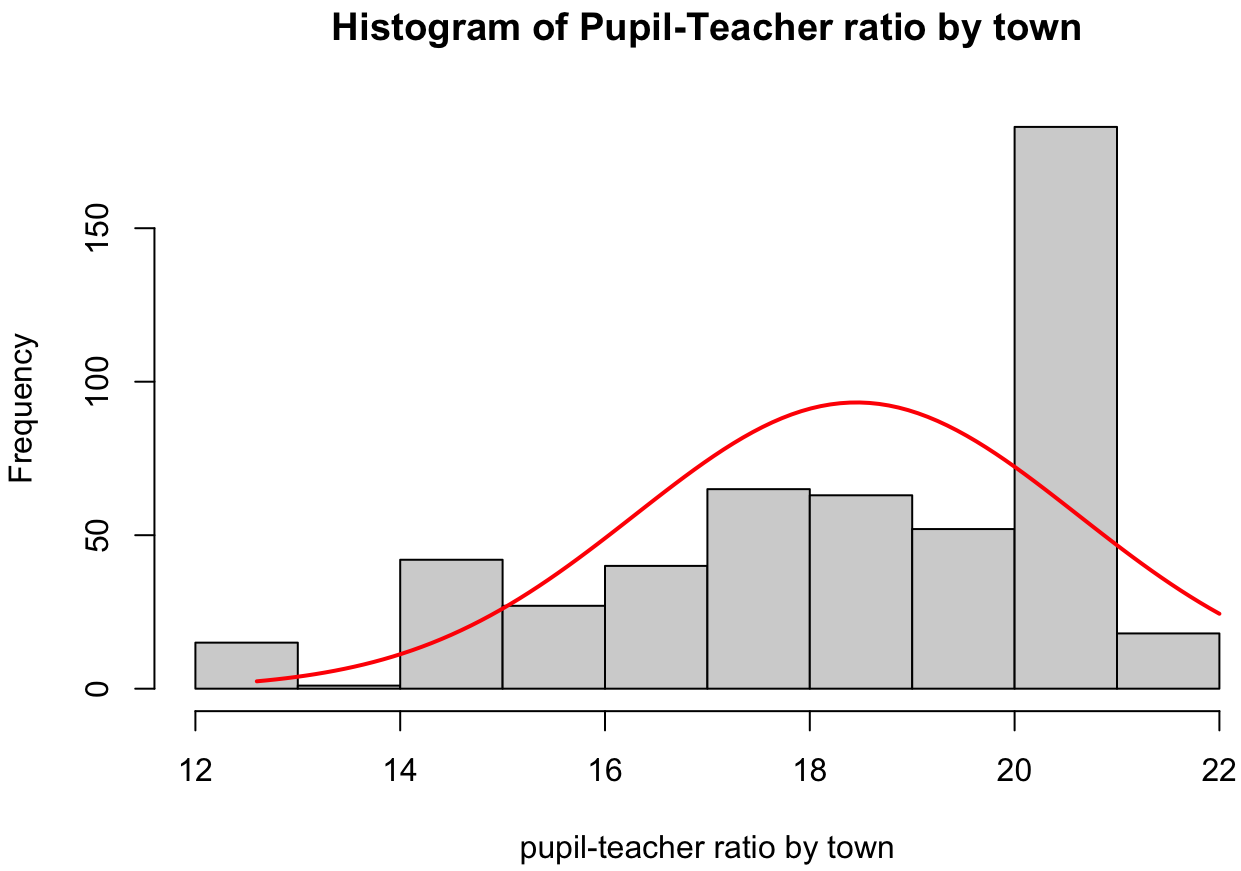
Comments: The tax variable refers to the full-value property tax rate per $10,000, one of the predictors in the dataset. From the histogram, its hard to tell the skewness in the data, as there is a gap in the data from 500-650 dollars roughly. Potentially the right side after the gap may be considered as outliers. There are some main peaks from 250-350 dollars and on the right side 650-700 dollars. With the fitted normal distribution curve, we can see that the curve does not really fit the data well in this case, with the gap in the data and the big outlier peak. Looking at the box plot, we can see the Interquartile range box, with the median towards the bottom side, suggesting heavy right skewness, and we can see that there are no outliers according to the boxplot. Looking at the 5 number summary, we have the median and mean at 330 and 408.2 dollars, and we have the IQR from 279 to 666 dollars Looking at the scatterplot, there isn't too much of a trend evident, other than the gap the data looks relatively random.

PTRATIO

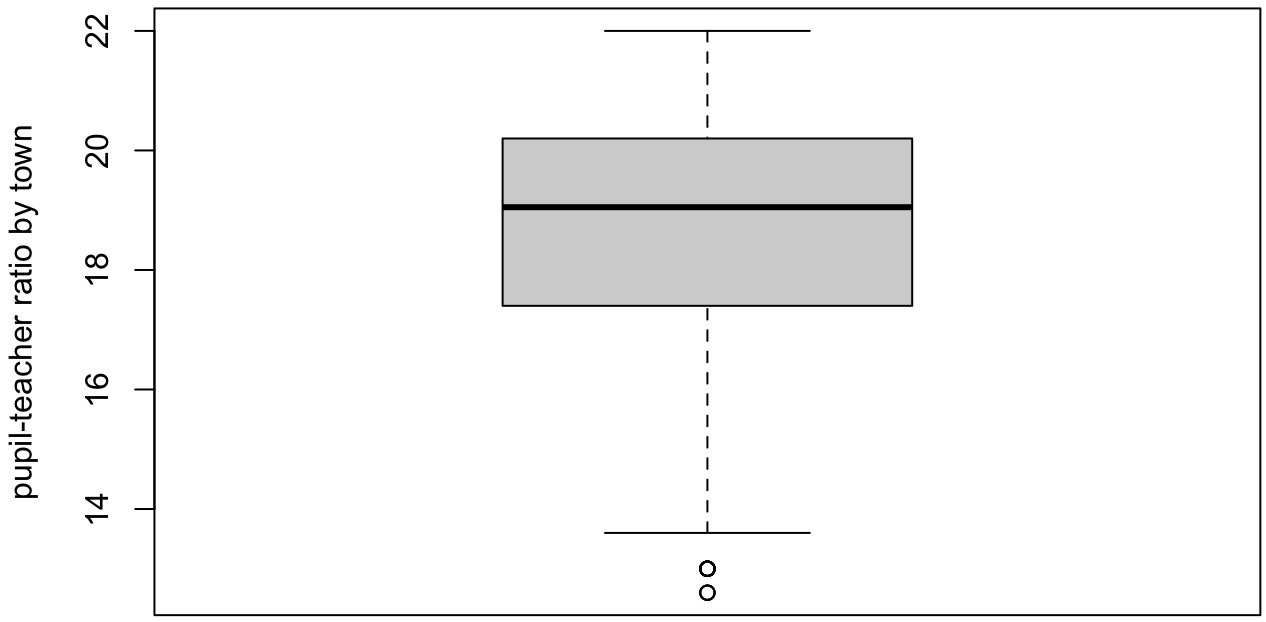
```
summary(df$ptratio)
```

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

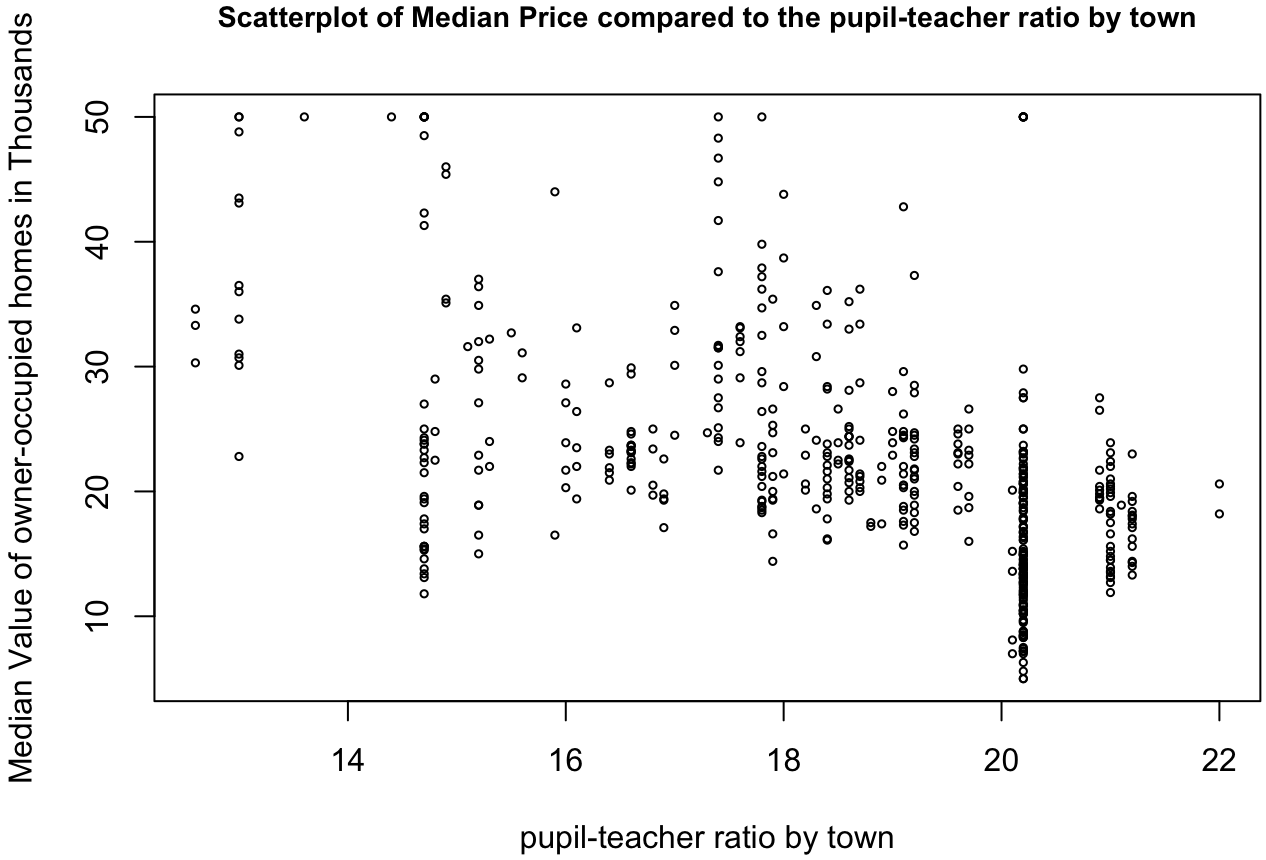
```
ptratio_hist <- hist(df$ptratio, main = "Histogram of Pupil-Teacher ratio by town", xlab
= "pupil-teacher ratio by town")
xfit <- seq(min(df$ptratio), max(df$ptratio), length = 100)
yfit <- dnorm(xfit, mean = mean(df$ptratio), sd = sd(df$ptratio))
yfit <- yfit * diff(ptratio_hist$mids[1:2]) * length(df$ptratio)
lines(xfit, yfit, col = "red", lwd = 2)
```




```
boxplot(df$ptratio,  main = "Boxplot of Pupil-Teacher ratio by town", ylab = "pupil-teac
her ratio by town")
```




```
plot(df$ptratio, df$medv, main = "Scatterplot of Median Price compared to the pupil-teac
her ratio by town", xlab = "pupil-teacher ratio by town", ylab = "Median Value of owner-
occupied homes in Thousands", cex=0.5, cex.main=0.90)
```



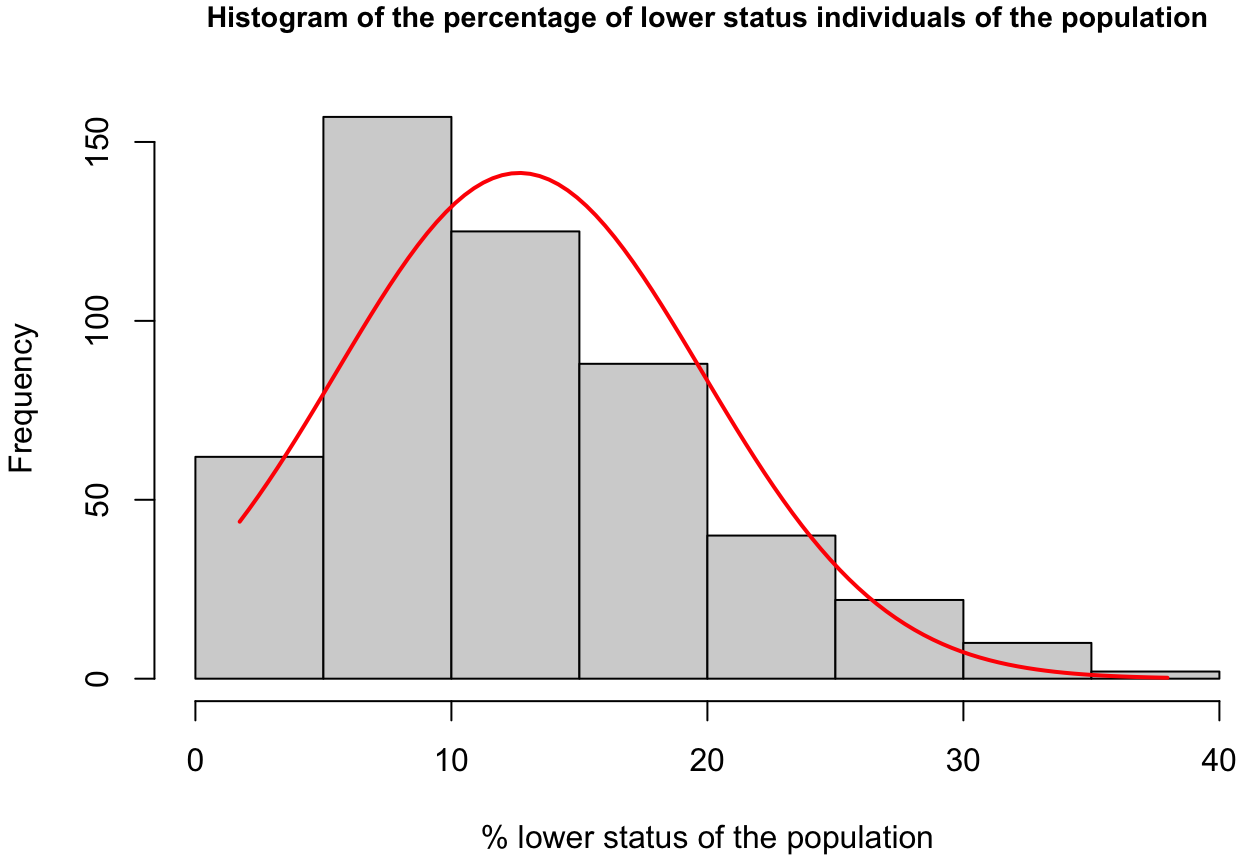
Comments: The ptratio variable refers to the Pupil-Teacher ratio by town, one of the predictors in the dataset. From the histogram, the data seems left skewed, and there is a tiny gap in the data from 13-14. Potentially the left side observations may be considered as outliers. There one main peak on the right side at 20-21. With the fitted normal distribution curve, we can see that other than the big peak at 20-21, the curve actually fits the data decently well. Looking at the box plot, we can see the Interquartile range box, with the median slightly towards the top side, suggesting left skewness, and we can see that there are a couple outliers towards the lower end according to the boxplot. Looking at the 5 number summary, we have the median and mean at 19.05 and 18.46 dollars, and we have the IQR from 17.4 to 20.2 dollars Looking at the scatterplot, we can see a slight trend that as the pupil teacher ratio increases, the median value of the owner-occupied homes decreases.

LSTAT

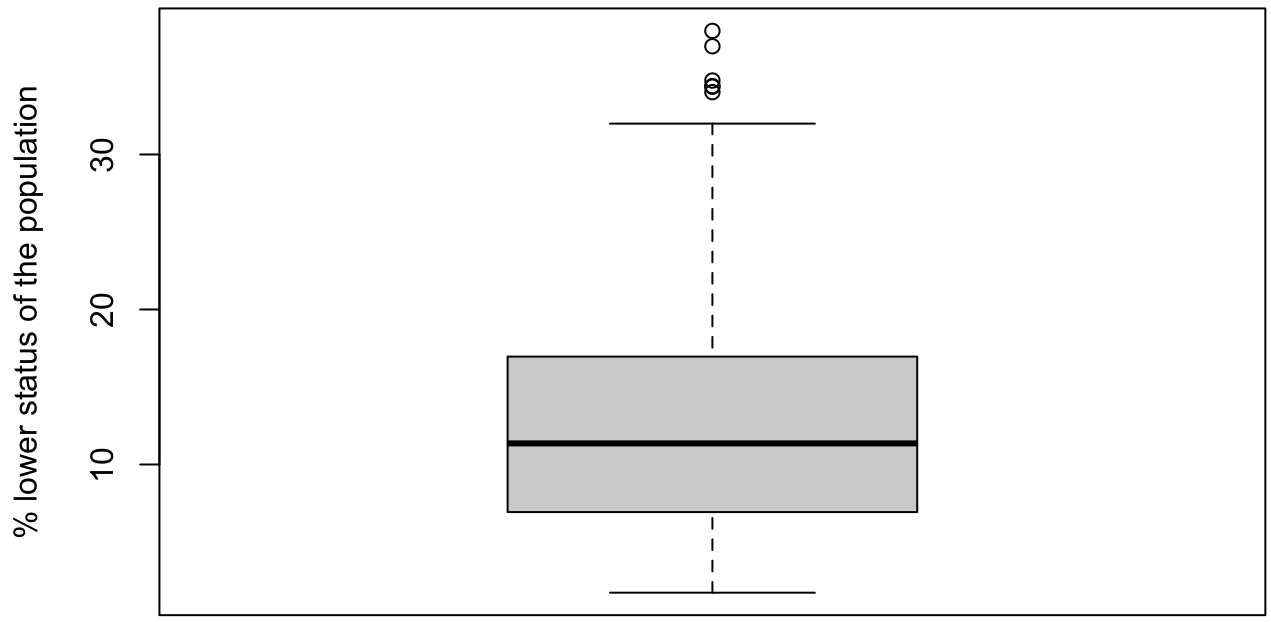
```
summary(df$lstat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.73    6.95   11.36   12.65   16.95   37.97
```

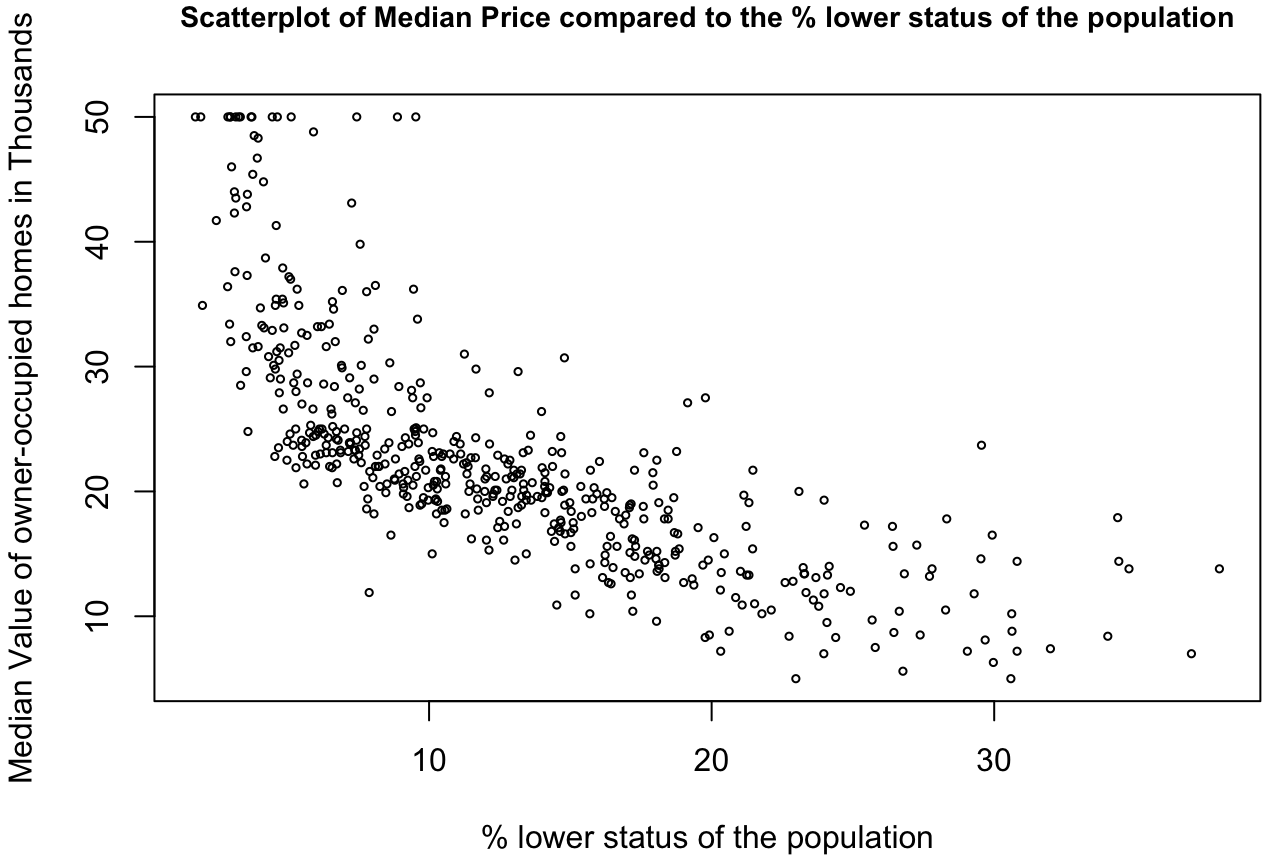
```
lstat_hist <- hist(df$lstat, main = "Histogram of the percentage of lower status individ
uals of the population", xlab = "% lower status of the population", cex.main=0.9)
xfit <- seq(min(df$lstat), max(df$lstat), length = 100)
yfit <- dnorm(xfit, mean = mean(df$lstat), sd = sd(df$lstat))
yfit <- yfit * diff(lstat_hist$mids[1:2]) * length(df$lstat)
lines(xfit, yfit, col = "red", lwd = 2)
```




```
boxplot(df$lstat, main = "Boxplot of the percentage of lower status individuals of the p
opulation", ylab = "% lower status of the population", cex.main=0.90)
```




```
plot(df$lstat, df$medv, main = "Scatterplot of Median Price compared to the % lower stat
us of the population", xlab = "% lower status of the population", ylab = "Median Value o
f owner-occupied homes in Thousands", cex=0.5, cex.main=0.90)
```



Comments: The Lstat variable refers to the percentage of lower status individuals in the population, one of the predictors in the dataset. From the histogram, the data seems right skewed. Potentially there may be some outliers on the right side. There are a couple main peaks at 5-15 percent. With the fitted normal distribution curve, we can see that the curve actually fits the data pretty well, suggesting a pretty symmetric distribution. Looking at the box plot, we can see the Interquartile range box, with the median slightly towards the bottom side, suggesting some right skewness, and we can see that there are a couple outliers towards the top end according to the boxplot. Looking at the 5 number summary, we have the median and mean at 11.36 and 12.65 percent, and we have the IQR from 6.95 to 16.95 perfent Looking at the scatterplot, we can see a pretty clear trend that as the percentage of lower status people in the population increase, the median value of the owner-occupied homes decreases. Again, this makes economic sense as well.

Step 2: multiple linear regression model with the main effects

```
reg.mod <- lm(df$medv ~., data = df)
summary(reg.mod)
```

```
##
## Call:
## lm(formula = df$medv ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7785  -2.8017  -0.6185   2.0915  26.5619
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.327911   4.979197   8.501 2.24e-16 ***
## crim         -0.127477   0.033268  -3.832 0.000144 ***
## zn            0.047660   0.014012   3.401 0.000725 ***
## indus         0.034056   0.062425   0.546 0.585627
## nox         -18.382995   3.887082  -4.729 2.94e-06 ***
## rm            3.693754   0.424190   8.708  < 2e-16 ***
## age           0.005970   0.013439   0.444 0.657070
## dis          -1.501890   0.203555  -7.378 6.83e-13 ***
## rad           0.312184   0.067190   4.646 4.34e-06 ***
## tax          -0.014205   0.003809  -3.729 0.000214 ***
## ptratio      -0.977286   0.132925  -7.352 8.15e-13 ***
## lstat        -0.563314   0.051032 -11.038  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.845 on 494 degrees of freedom
## Multiple R-squared:  0.7286, Adjusted R-squared:  0.7225
## F-statistic: 120.5 on 11 and 494 DF,  p-value: < 2.2e-16
```

In terms of the statistical and economic significance, we can see that the P value is very small which shows that this model is significant. However, we can see that not all predictors are statistically significant. The age and indus predictors are not statistically significant which shows that they may not be impactful for the model.

Here are the interpretations of the estimates: Crim: a 1 % increase in the per capita crime rate is predicted to be associated with a -0.1275 thousand dollar change in the median value of owner occupied homes.

zn: a 1 % increase in the proportion of residential land zoned for lots over 25,000 sq.ft is predicted to be associated with a 0.04766 thousand dollar change in the median value of owner occupied homes.

indus: a 1 % increase in the proportion of non retail business acres per town is predicted to be associated with a 0.034 thousand dollar change in the median value of owner occupied homes.

nox: a 1 unit increase in the nitric oxides concentration is predicted to be associated with a -18.383 thousand dollar change in the median value of owner occupied homes

rm: a 1 unit increase in the average number of rooms per dwelling is predicted to be associated with a 3.69 thousand dollar change in the median value of owner occupied homes

age: a 1% increase in the proportion of owner occupied units built prior to 1940 is predicted to be associated with a 0.00597 thousand dollar change in the median value of owner occupied homes

dis: a 1 unit increase in the weighted distance to five boston employment centers is predicted to be associated with a -1.502 thousand dollar change in the median value of owner occupied homes

rad: a 1 unit increase in the index of accessibility to radial highways is predicted to be associated with a 0.3122 thousand dollar increase in the median value of owner occupied homes

tax: a $1 increase in the full value property tax rate per $10,000 is predicted to be associated with a -0.0142 thousand dollar change in the median value of owner occupied homes

ptratio: a 1 unit increase in the pupil-teacher ratio by town is predicted to be associated with a -0.9773 thousand dollar change in the median value of owner occupied homes

LSTAT: a 1% increase in the proportion of lower status individuals in the population is predicted to be associated with a -0.633 thousand dollar change in the median value of owner occupied homes

Step 3: identifying any outliers, high levarage and or influential observations worth removing and re-estimating model:

Medv: the boxplot had identified a number of potential outliers, however looking at the plots there don't seem to be any massive gaps and significant outliers so we will not identify any outliers in this variable

Crim: From the histogram and box plot of this variable, there definitely seems to be high influential outliers in this variable. The boxplot identified a number of them, however it looks like there are 3 main massive outliers so we will remove those 3 highest observations.

Zn: From the histogram and box plot of this variable, There do seem to be a number of potential outliers on the high side far away from the peak towards the left side, however they all seem relatively grouped together so we will leave them in the dataset.

indus: From the histogram and boxplot, there are no outliers visible, and no outliers identified as well, so we will not remove any observations.

nox: From the histogram there seems to be one or two potential outliers, however the boxplot did not identify any so we will not remove any observations.

rm: The boxplot identified a number of outliers on both sides, however from the histogram the data seems relatively symmetrical so we will not remove any observations.

age: Neither the histogram nor boxplot showed signs of any big outliers so we will not remove any observations.

dis: From the histogram, there seems to be a couple potential outliers on the high side, and the boxplot also identified a couple outliers. Since a couple of the outliers are relatively close to the rest of the data and close to each other, we will only remove the largest outlier in the variable.

rad: Looking at the histogram, there seems to be a main distribution to the left, and then a major gap and then a number of observations that could be potential outliers to the right. From the boxplot, there weren't any outliers identified. Although there is a massive gap, due to the frequency of observations after the gap, we will not remove any observations.

tax: Looking at the histogram, there again is a massive gap in the data, and therefore potential outliers to the right, however the frequency is high. According to the boxplot, there are no outliers identified. Therefore, we will not remove any observations even with the massive gap in the data.

ptratio: Looking at the histogram, it seems as though there may be potential outliers on bottom side. The boxplot also identified a couple of outliers on the bottom side. As a result, we will remove the bottom 3 observations from this variable(2 observations are the same value).

Lstat: Looking at the histogram, there doesn't seem to be any clear outliers. According to the boxplot, there were some outliers identified on the top side, however they are all grouped decently together therefore we will not remove any observations from the variable.

Remove the outlier observations from the variables and re-estimate the model:

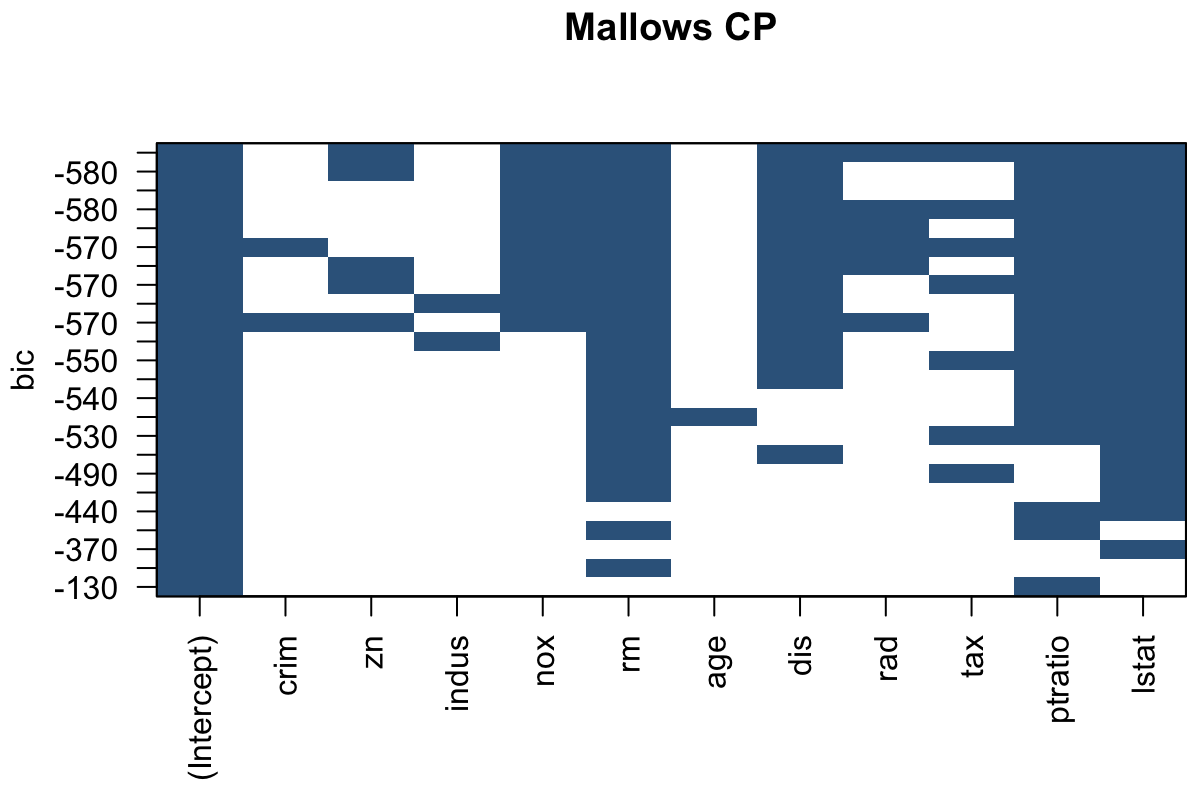
```
crim_outliers <- df$crim > 60
dis_outliers <- df$dis > 12
ptratio_outliers <- df$ptratio < 12.8
outliers_rows <- crim_outliers | dis_outliers | ptratio_outliers
df_cleaned <- df[!outliers_rows, ]

cleaned_reg.mod <- lm(df_cleaned$medv ~., data = df_cleaned)
summary(cleaned_reg.mod)
```

```
##
## Call:
## lm(formula = df_cleaned$medv ~ ., data = df_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7452  -2.7979  -0.5416   2.0765  26.5782
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.645860   5.043249   8.456 3.24e-16 ***
## crim         -0.112995   0.052795  -2.140 0.032828 *
## zn            0.047216   0.014253   3.313 0.000992 ***
## indus         0.029149   0.063054   0.462 0.644084
## nox         -18.724068   3.930347  -4.764 2.51e-06 ***
## rm            3.723614   0.428926   8.681  < 2e-16 ***
## age           0.005070   0.013590   0.373 0.709248
## dis          -1.529392   0.209661  -7.295 1.22e-12 ***
## rad           0.296552   0.070502   4.206 3.09e-05 ***
## tax          -0.013493   0.003876  -3.481 0.000544 ***
## ptratio      -0.992386   0.135595  -7.319 1.04e-12 ***
## lstat        -0.562763   0.052100 -10.802  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.87 on 487 degrees of freedom
## Multiple R-squared:  0.723,  Adjusted R-squared:  0.7167
## F-statistic: 115.5 on 11 and 487 DF,  p-value: < 2.2e-16
```

Step 4: Use Mallows CP for identifying the terms you will keep in the model, and use the Boruta algorithm for variable selection

Mallows_CP

```
library(leaps)
ss=regsubsets(df_cleaned$medv ~., method =c("exhaustive"), nbest = 3, data=df_cleaned)
plot(ss, statistic="cp", legend=F, main = "Mallows CP", col = "steelblue4", ylim = c(0,5
0)) #originally subsets(), gave me an error
```



Boruta Algorithm:

```
library(Boruta)
Bor.res <- Boruta(df_cleaned$medv ~., data = df_cleaned, doTrace = 2)
```

```
##  1. run of importance source...
```

```
##  2. run of importance source...
```

```
##  3. run of importance source...
```

```
##  4. run of importance source...
```

```
##  5. run of importance source...
```

```
##  6. run of importance source...
```

```
##  7. run of importance source...
```

```
##  8. run of importance source...
```

```
##  9. run of importance source...
```

```
##  10. run of importance source...
```

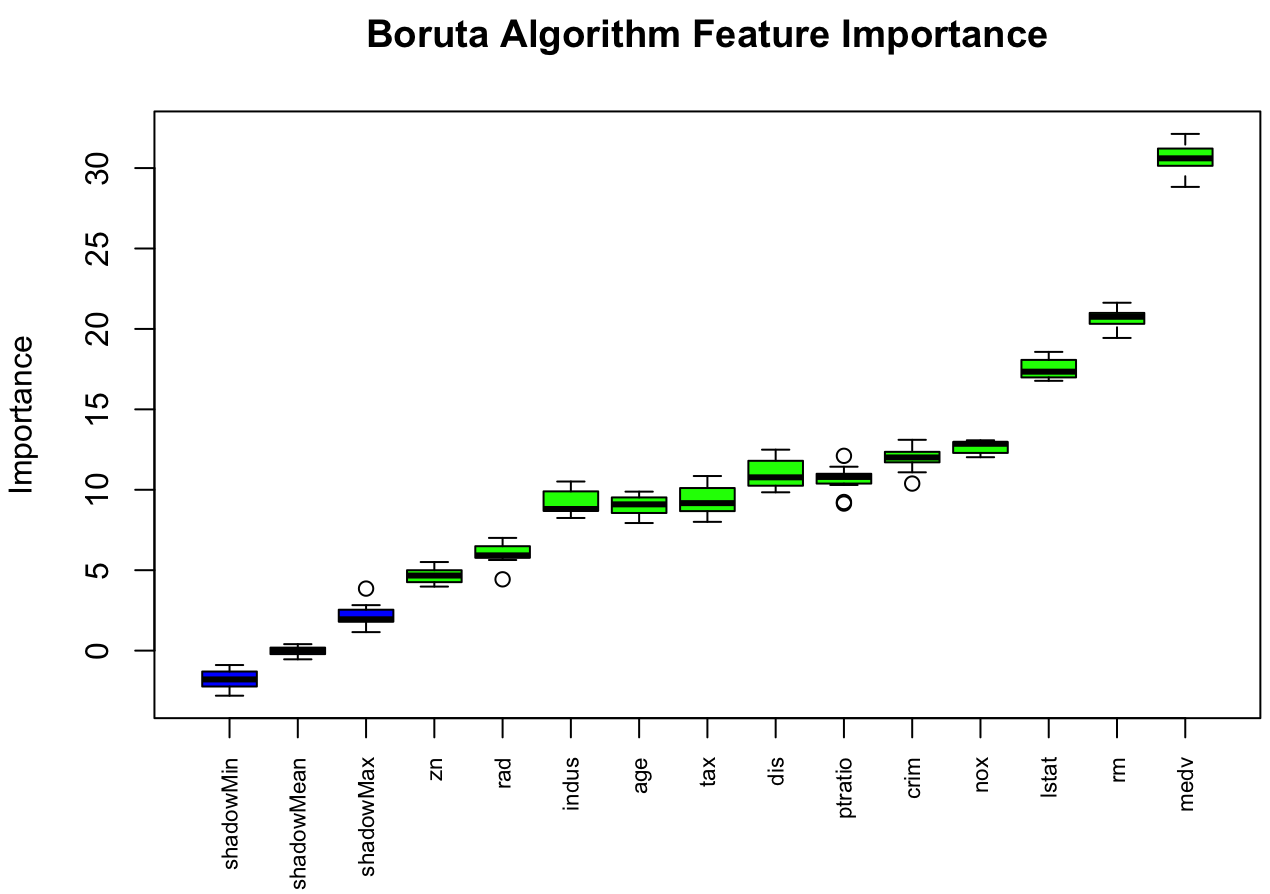
```
##  11. run of importance source...
```

```
## After 11 iterations, +1.1 secs:
```

```
##  confirmed 12 attributes: age, crim, dis, indus, lstat and 7 more;
```

```
##  no more attributes left.
```

```
#plot(Bor.res,sort=TRUE)
plot(Bor.res, xlab = "", xaxt = "n", main="Boruta Algorithm Feature Importance")
lz<-lapply(1:ncol(Bor.res$ImpHistory),function(i)
Bor.res$ImpHistory[is.finite(Bor.res$ImpHistory[,i]),i])
names(lz) <- colnames(Bor.res$ImpHistory)
Labels <- sort(sapply(lz,median))
axis(side = 1,las=2,labels = names(Labels),
at = 1:ncol(Bor.res$ImpHistory), cex.axis = 0.7)
```



Interpretation: Overall, from the Mallows CP and Boruta algorithm, we are getting different results. According to the Mallows CP, the best model removes the crim, indus and age variables. On the other hand, according to the Boruta algorithm, all the predictors are confirmed. In this case, because indus and age were ommited using mallows CP, and are the least important in the boruta algorithm, we will choose to keep all the variables except for the indus and age variables. We will keep the crim variable as in the regression, it was significant(albeit not 3 stars, and it is more important than the other 2 variables)

```
summary(cleaned_reg.mod)
```

```
##
## Call:
## lm(formula = df_cleaned$medv ~ ., data = df_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7452  -2.7979  -0.5416   2.0765  26.5782
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.645860   5.043249   8.456 3.24e-16 ***
## crim         -0.112995   0.052795  -2.140 0.032828 *
## zn            0.047216   0.014253   3.313 0.000992 ***
## indus         0.029149   0.063054   0.462 0.644084
## nox         -18.724068   3.930347  -4.764 2.51e-06 ***
## rm            3.723614   0.428926   8.681  < 2e-16 ***
## age           0.005070   0.013590   0.373 0.709248
## dis          -1.529392   0.209661  -7.295 1.22e-12 ***
## rad           0.296552   0.070502   4.206 3.09e-05 ***
## tax          -0.013493   0.003876  -3.481 0.000544 ***
## ptratio      -0.992386   0.135595  -7.319 1.04e-12 ***
## lstat        -0.562763   0.052100 -10.802  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.87 on 487 degrees of freedom
## Multiple R-squared:  0.723,  Adjusted R-squared:  0.7167
## F-statistic: 115.5 on 11 and 487 DF,  p-value: < 2.2e-16
```

```
var_selected_reg.mod <- lm(medv~.-age-indus, data=df_cleaned)
summary(var_selected_reg.mod)
```

```
##
## Call:
## lm(formula = medv ~ . - age - indus, data = df_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7650  -2.7684  -0.5747   2.0572  26.6185
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.40617    5.01806   8.451 3.34e-16 ***
## crim         -0.11466    0.05262  -2.179  0.02980 *
## zn            0.04584    0.01403   3.267  0.00116 **
## nox         -17.87894    3.65387  -4.893 1.35e-06 ***
## rm            3.73799    0.41777   8.947  < 2e-16 ***
## dis          -1.57528    0.19349  -8.141 3.27e-15 ***
## rad           0.28628    0.06785   4.219 2.92e-05 ***
## tax          -0.01267    0.00348  -3.642  0.00030 ***
## ptratio      -0.98136    0.13408  -7.319 1.03e-12 ***
## lstat        -0.55434    0.04889 -11.338  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.862 on 489 degrees of freedom
## Multiple R-squared:  0.7228, Adjusted R-squared:  0.7177
## F-statistic: 141.7 on 9 and 489 DF,  p-value: < 2.2e-16
```

When comparing the original model and the variable reduced model, we see that the model with fewer predictors has a higher adjusted R squared, so we will use that one for part 5.

Step 5 Test for multicollinearity using VIF on the model from 4, and based on the test, remove any appropriate variables and estimate a new regression model

```
library(car)
```

```
## Loading required package: carData
```

```
vif(var_selected_reg.mod)
```

```
##     crim       zn      nox       rm      dis      rad      tax  ptratio
## 2.450054 2.113825 3.752527 1.812688 3.353315 7.296899 7.210985 1.713137
##    lstat
## 2.565627
```

Interpretation: In this case, we can see that the VIF values for most of the variables are relatively low, however we can see for the rad and tax variables, the VIF is quite high. Therefore, we will estimate the new model by removing those two variables

```
new_reg_mod <- lm(medv ~.-tax-rad-age-indus, data = df_cleaned)
summary(new_reg_mod)
```

```
##
## Call:
## lm(formula = medv ~ . - tax - rad - age - indus, data = df_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8745  -3.0022  -0.5951   1.9221  27.3074
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.73328    4.68296   7.844 2.74e-14 ***
## crim         -0.04076    0.04385  -0.929  0.35315
## zn            0.03862    0.01396   2.766  0.00589 **
## nox         -17.96203    3.29341  -5.454 7.82e-08 ***
## rm            4.09249    0.41495   9.863  < 2e-16 ***
## dis          -1.50169    0.19589  -7.666 9.58e-14 ***
## ptratio      -0.94081    0.12177  -7.726 6.28e-14 ***
## lstat        -0.56481    0.04964 -11.379  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.942 on 491 degrees of freedom
## Multiple R-squared:  0.7124, Adjusted R-squared:  0.7083
## F-statistic: 173.7 on 7 and 491 DF,  p-value: < 2.2e-16
```

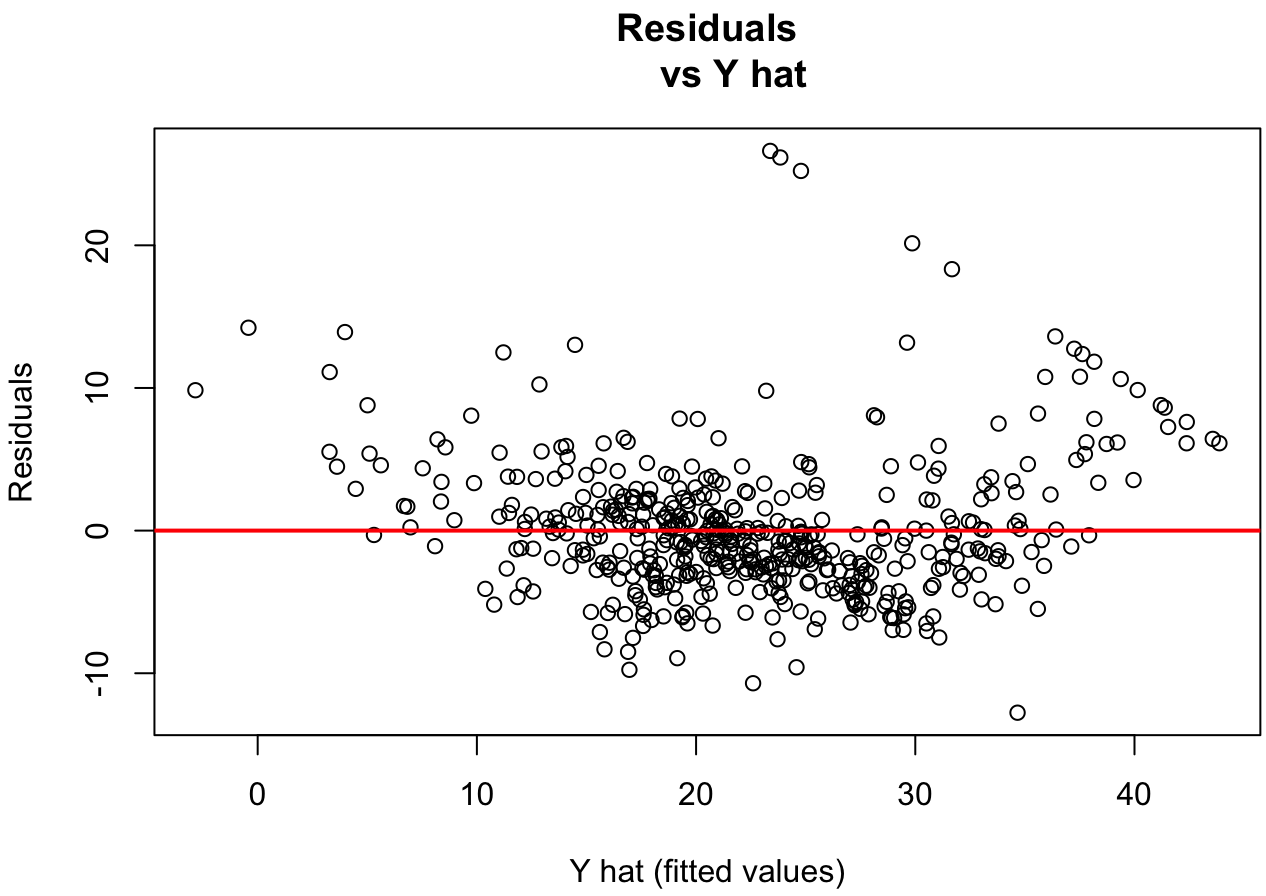
```
summary(var_selected_reg.mod)
```

```
##
## Call:
## lm(formula = medv ~ . - age - indus, data = df_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7650  -2.7684  -0.5747   2.0572  26.6185
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.40617    5.01806   8.451 3.34e-16 ***
## crim         -0.11466    0.05262  -2.179  0.02980 *
## zn            0.04584    0.01403   3.267  0.00116 **
## nox         -17.87894    3.65387  -4.893 1.35e-06 ***
## rm            3.73799    0.41777   8.947  < 2e-16 ***
## dis          -1.57528    0.19349  -8.141 3.27e-15 ***
## rad           0.28628    0.06785   4.219 2.92e-05 ***
## tax          -0.01267    0.00348  -3.642  0.00030 ***
## ptratio      -0.98136    0.13408  -7.319 1.03e-12 ***
## lstat        -0.55434    0.04889 -11.338  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.862 on 489 degrees of freedom
## Multiple R-squared:  0.7228, Adjusted R-squared:  0.7177
## F-statistic: 141.7 on 9 and 489 DF,  p-value: < 2.2e-16
```

In this new model, we see that except crim is significant. This has an adjusted R squared of 70.83%, while the original one that is just missing age and indus (both are cleaned versions without outliers) also has a 71.77% adjusted R squared, implying that it is better to use tax and rad. Also, in the var_selected model, we see that all predictors are significant, so this is the better one to use.

Step 6 For your model in part (5) plot the respective residuals vs y_hate and comment on your findings.

```
plot(var_selected_reg.mod$fitted.values, var_selected_reg.mod$residuals,
     xlab = "Y hat (fitted values)", ylab= "Residuals", main = "Residuals
     vs Y hat")
abline(h=0, col="red", lwd=2)
```



There are some values that have very high residuals, both it seems that most of the residuals are centered around 0, meaning that the fitted values did a good job of predicting the actual y values.

Step 7 perform a RESET test and comment on your findings

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
resettest(var_selected_reg.mod, power=2)
```

```
##
##  RESET test
##
## data:  var_selected_reg.mod
## RESET = 200.97, df1 = 1, df2 = 488, p-value < 2.2e-16
```

We reject the null hypothesis (null states that the model is fine as is and we don't need to add more terms to it). We did delete two variables earlier, so this could imply that we should bring them back. However, in this case, the reset tests contradicts the Mallows Cp and Boruta test and the summaries of the original model and the reduced model show that the reduced model is probably the better option.

Step 8 For the model in part (5) test for heteroskedasticity and comment on your findings.

```
#There are three tests: we'll start with BP test
#null hypothesis is that variance is constant

library(lmtest)
bptest(var_selected_reg.mod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  var_selected_reg.mod
## BP = 52.745, df = 9, p-value = 3.267e-08
```

We can reject the null hypothesis, meaning that heteroskedasticity is present. However, we will need to run the other two tests to confirm this.

```
gqtest(var_selected_reg.mod)
```

```
##
##  Goldfeld-Quandt test
##
## data:  var_selected_reg.mod
## GQ = 2.4391, df1 = 240, df2 = 239, p-value = 5.97e-12
## alternative hypothesis: variance increases from segment 1 to 2
```

We can reject the null hypothesis, meaning heteroskedasticity is present. Since 2/3 tests suggest heteroskedasticity, we can say that heteroskedasticity is present.

Step 9 Estimate a model based on all your findings that also includes interaction terms (if appropriate) and if needed, any higher power terms. Comment

on the performance of this model compared to your other models. Make sure to use AIC and BIC for model comparison

There are multiple ways of doing this, and we'll start with Robust Standard Errors (adjusts Standard Errors).

```
robust <- hccm(var_selected_reg.mod, type="hc1")
coeftest(var_selected_reg.mod, vcov.=robust)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  42.4061655   7.2961603  5.8121 1.114e-08 ***
## crim         -0.1146634   0.0565957 -2.0260 0.0433064 *
## zn            0.0458348   0.0138394  3.3119 0.0009952 ***
## nox         -17.8789382   3.2621542 -5.4807 6.796e-08 ***
## rm            3.7379864   0.7729512  4.8360 1.778e-06 ***
## dis          -1.5752808   0.2319184 -6.7924 3.212e-11 ***
## rad           0.2862776   0.0680649  4.2059 3.093e-05 ***
## tax          -0.0126720   0.0028483 -4.4489 1.069e-05 ***
## ptratio      -0.9813587   0.1149008 -8.5409 < 2.2e-16 ***
## lstat        -0.5543363   0.0869733 -6.3736 4.272e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(var_selected_reg.mod)
```

```
##
## Call:
## lm(formula = medv ~ . - age - indus, data = df_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7650  -2.7684  -0.5747   2.0572  26.6185
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.40617    5.01806   8.451 3.34e-16 ***
## crim         -0.11466    0.05262  -2.179  0.02980 *
## zn            0.04584    0.01403   3.267  0.00116 **
## nox         -17.87894    3.65387  -4.893 1.35e-06 ***
## rm            3.73799    0.41777   8.947  < 2e-16 ***
## dis          -1.57528    0.19349  -8.141 3.27e-15 ***
## rad           0.28628    0.06785   4.219 2.92e-05 ***
## tax          -0.01267    0.00348  -3.642  0.00030 ***
## ptratio      -0.98136    0.13408  -7.319 1.03e-12 ***
## lstat        -0.55434    0.04889 -11.338  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.862 on 489 degrees of freedom
## Multiple R-squared:  0.7228, Adjusted R-squared:  0.7177
## F-statistic: 141.7 on 9 and 489 DF,  p-value: < 2.2e-16
```

We see that when we use robust standard errors, there actually isn't really that much of a difference in standard errors, and for some predictors, the standard error actually goes up.

The next method is GLS Known Form of Variance.

```
library(broom)
weights = 1/df_cleaned$nox #nox has the highest coefficient so we chose this #variable f
or weights
weighted_model <- lm(medv~.-age-indus, weights=weights, data=df_cleaned)
gls <- coeftest(weighted_model, vcov.=robust)
tidy(gls)
```

```
## # A tibble: 10 × 5
##    term        estimate std.error statistic  p.value
##    <chr>          <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)  33.6      7.30         4.61 5.24e- 6
##  2 crim         -0.103    0.0566      -1.81 7.06e- 2
##  3 zn            0.0415   0.0138       3.00 2.88e- 3
##  4 nox         -15.6      3.26        -4.78 2.34e- 6
##  5 rm            4.54     0.773        5.87 7.89e- 9
##  6 dis          -1.40     0.232       -6.05 2.90e- 9
##  7 rad           0.277    0.0681       4.06 5.64e- 5
##  8 tax          -0.0131   0.00285     -4.59 5.67e- 6
##  9 ptratio      -0.897    0.115       -7.81 3.51e-14
## 10 lstat        -0.511    0.0870      -5.87 8.02e- 9
```

Once again, there doesn't really seem to be much of a difference between this version and the original model, so it is probably better to just stick to the original model.

The final method is GLS Unknown Form of Variance

```
log_df <- log(df_cleaned)

library(stats)
ehatsq <- resid(var_selected_reg.mod) ^ 2
log_df["ehatsq_log"] <- log(ehatsq)


sighatsq.ols <- lm(ehatsq_log~.-age-indus-medv-zn, data=log_df)
vari <- exp(fitted(sighatsq.ols))
vari_model <- lm(medv ~ .-age-indus, weights=1/vari, data=df_cleaned)
tidy(vari_model)
```

```
## # A tibble: 10 × 5
##    term        estimate std.error statistic  p.value
##    <chr>          <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)  24.8      3.92         6.31 6.19e-10
##  2 crim         -0.126    0.0550      -2.29 2.27e- 2
##  3 zn            0.0320   0.00973      3.28 1.10e- 3
##  4 nox         -14.6      2.78        -5.26 2.21e- 7
##  5 rm            4.93     0.376       13.1  5.86e-34
##  6 dis          -0.893    0.126       -7.10 4.48e-12
##  7 rad           0.218    0.0547       3.99 7.73e- 5
##  8 tax          -0.0132   0.00260     -5.07 5.62e- 7
##  9 ptratio      -0.817    0.0884      -9.24 7.35e-19
## 10 lstat        -0.305    0.0337      -9.06 3.18e-18
```

```
summary(vari_model)
```

```
##
## Call:
## lm(formula = medv ~ . - age - indus, data = df_cleaned, weights = 1/vari)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1323 -1.1779 -0.1306  1.1008  8.2434
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.750084   3.921150   6.312 6.19e-10 ***
## crim         -0.125780   0.055028  -2.286   0.0227 *
## zn            0.031960   0.009733   3.284   0.0011 **
## nox         -14.628891   2.783764  -5.255 2.21e-07 ***
## rm            4.933354   0.375653  13.133  < 2e-16 ***
## dis          -0.893075   0.125822  -7.098 4.48e-12 ***
## rad           0.218204   0.054734   3.987 7.73e-05 ***
## tax          -0.013165   0.002596  -5.071 5.62e-07 ***
## ptratio      -0.816931   0.088379  -9.244  < 2e-16 ***
## lstat        -0.305002   0.033672  -9.058  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.944 on 489 degrees of freedom
## Multiple R-squared:  0.7374, Adjusted R-squared:  0.7325
## F-statistic: 152.5 on 9 and 489 DF,  p-value: < 2.2e-16
```

```
bptest(vari_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  vari_model
## BP = 2.38, df = 9, p-value = 0.9839
```

```
gqtest(vari_model)
```

```
##
##  Goldfeld-Quandt test
##
## data:  vari_model
## GQ = 2.4391, df1 = 240, df2 = 239, p-value = 5.97e-12
## alternative hypothesis: variance increases from segment 1 to 2
```

```
ncvTest(vari_model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.887775e-06, Df = 1, p = 0.9989
```

Here we see that standard errors are much lower than in the original model (for each predictor, the standard error in the new model is lower with the exception of crim, whose standard error increased by .003). On top of that, the new adjusted R squared is 73.25%, which is a lot better than the original model, that had an adjusted R squared of 71.77%, so we will use this as our final model unless the AIC and BIC suggests we should remove variables. Also, when it comes to heteroskedasticity in the new model, we see that 2/3 tests suggest constant variance, so we won't have to worry about that issue anymore.

```
AIC(vari_model)
```

```
## [1] 2786.167
```

```
AIC(lm(medv~.-age-indus-crim, weights=1/vari, data=df_cleaned))
```

```
## [1] 2789.47
```

```
AIC(lm(medv~.-age-indus-crim-zn, weights=1/vari, data=df_cleaned))
```

```
## [1] 2796.562
```

```
AIC(lm(medv~.-age-indus-crim-zn-rad, weights=1/vari, data=df_cleaned))
```

```
## [1] 2802.863
```

```
AIC(lm(medv~.-age-indus-crim-zn-rad-nox, weights=1/vari, data=df_cleaned))
```

```
## [1] 2827.852
```

```
AIC(lm(medv~.-age-indus-crim-zn-rad-nox-dis, weights=1/vari, data=df_cleaned))
```

```
## [1] 2837.512
```

AIC test suggests that vari_model (all relevant predictors) is the best option.

```
BIC(vari_model)
```

```
## [1] 2832.505
```

```
BIC(lm(medv~.-age-indus-crim, weights=1/vari, data=df_cleaned))
```

```
## [1] 2831.596
```

```
BIC(lm(medv~.-age-indus-crim-zn, weights=1/vari, data=df_cleaned))
```

```
## [1] 2834.476
```

```
BIC(lm(medv~.-age-indus-crim-zn-rad, weights=1/vari, data=df_cleaned))
```

```
## [1] 2836.564
```

```
BIC(lm(medv~.-age-indus-crim-zn-rad-nox, weights=1/vari, data=df_cleaned))
```

```
## [1] 2857.34
```

```
BIC(lm(medv~.-age-indus-crim-zn-rad-nox-dis, weights=1/vari, data=df_cleaned))
```

```
## [1] 2862.788
```

BIC test also suggests to just keep all the relevant predictors, so it is in our best interest to just stick to vari_model.

Step 10 valuate your model performance (from 9) using cross-validation, and also by dividing your data into the traditional 2/3 training and 1/3 testing samples, to evaluate your out-of-sample performance. Comment on your results

```
set.seed(123)
row.number <- sample(1:nrow(df_cleaned), 0.67*nrow(df_cleaned))
train <- df_cleaned[row.number, ]
test <- df_cleaned[-row.number, ]
dim(train)
```

```
## [1] 334  12
```

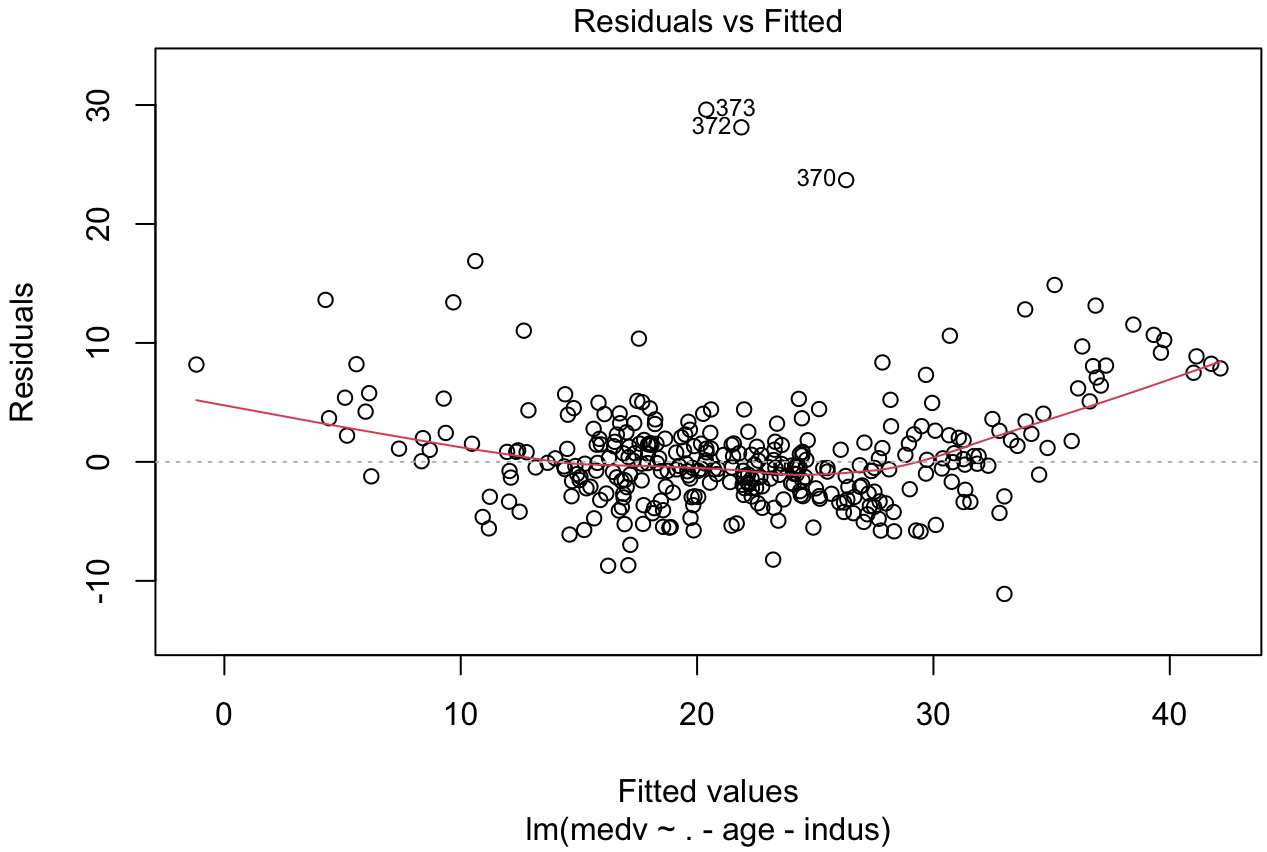
```
dim(test)
```

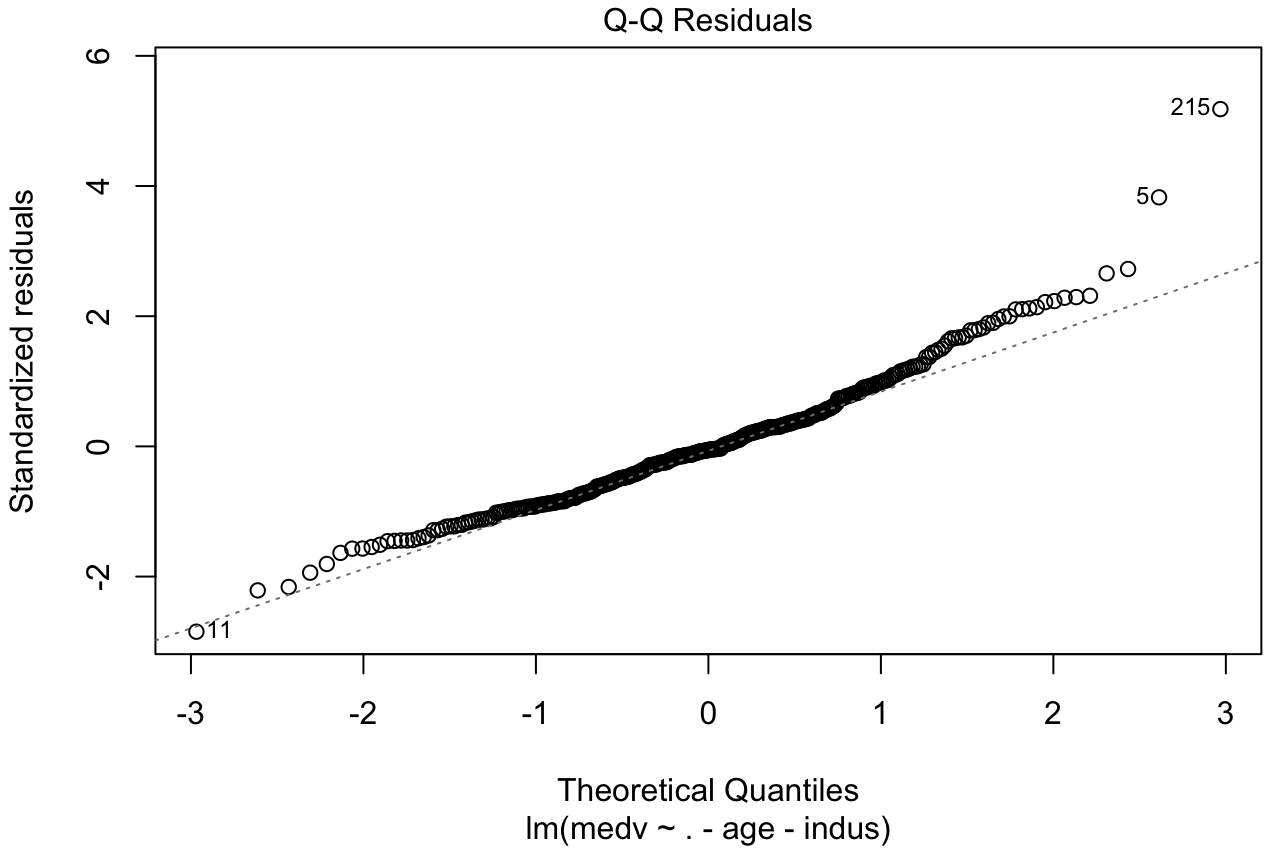
```
## [1] 165  12
```

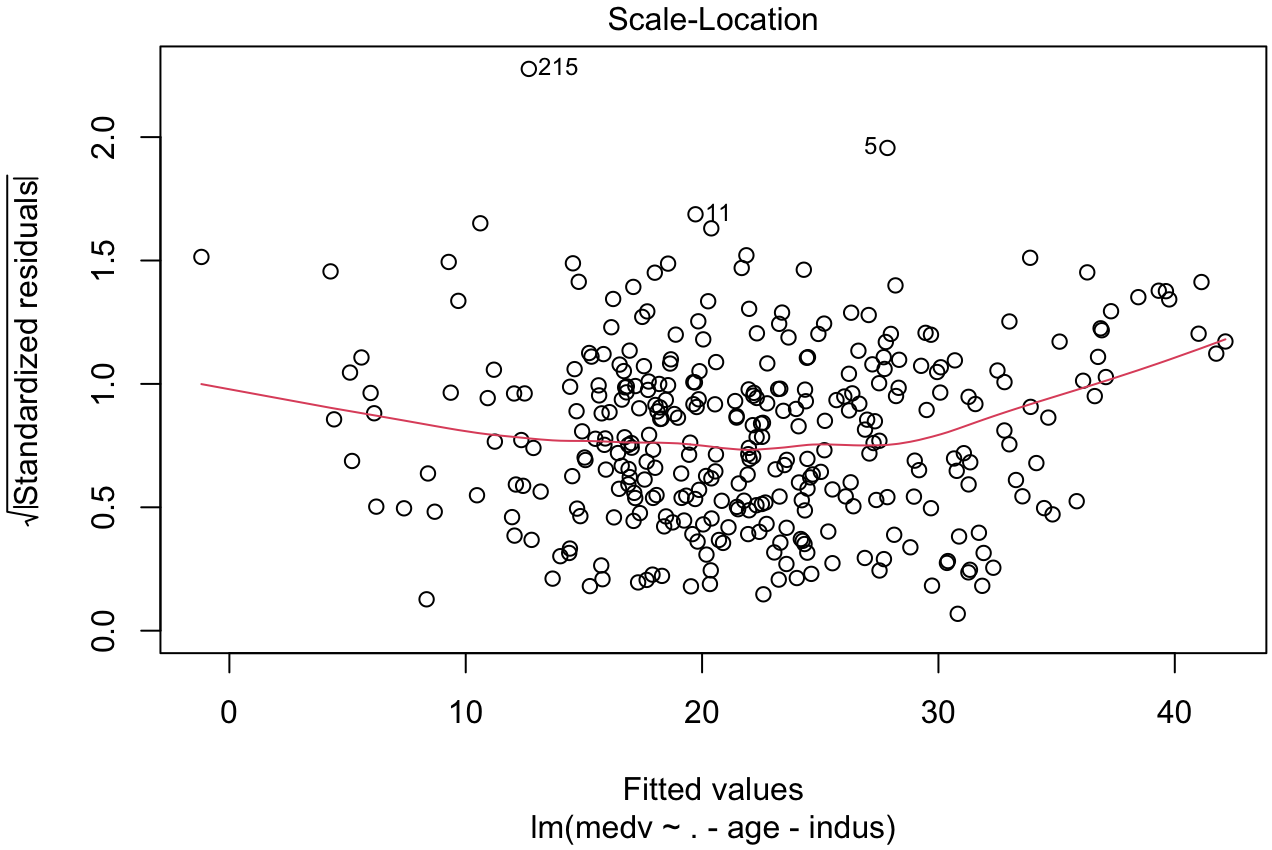
```
#Need to redo weights before doing anything
ehatsq <- resid(var_selected_reg.mod) ^ 2
log_df["ehatsq_log"] <- log(ehatsq)


sighatsq.ols_train <- lm(ehatsq_log~.-age-indus-medv-zn, data=log_df[row.number,])

vari_train <- exp(fitted(sighatsq.ols_train))

train_model <- lm(medv~.-age-indus, weights=1/vari_train, data=train)

sqrt(mean((train$medv - train_model$fitted.values) ^ 2))
```

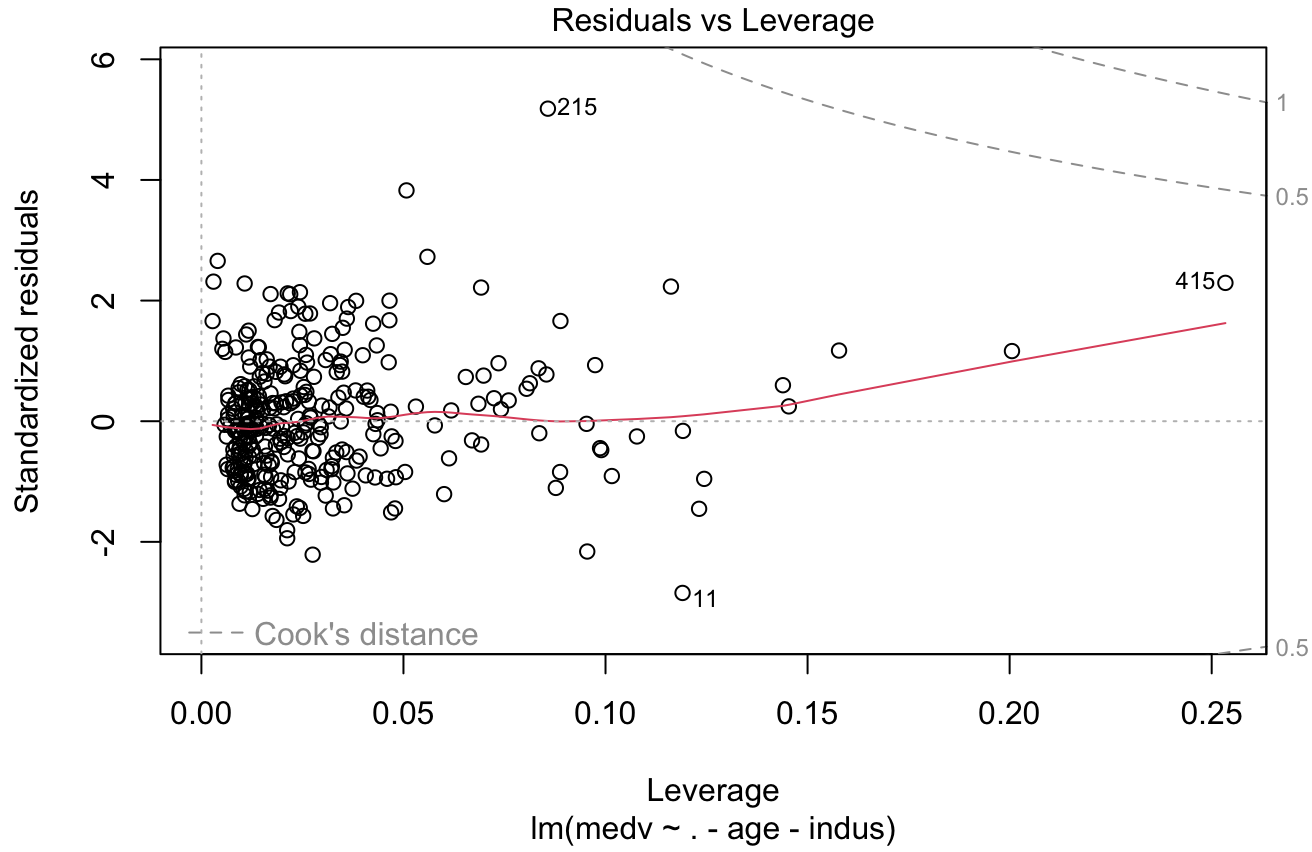
```
## [1] 4.81027
```

```
plot(train_model)
```


Residuals vs Fitted
373
372
370
Residuals
30
20
10
0
-10
0
10
20
30
40
Fitted values
lm(medv ~ . - age - indus)


Q-Q Residuals
215
5
11
Standardized residuals
6
4
2
0
-2
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(medv ~ . - age - indus)


Scale-Location
215
5
11
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
Fitted values
lm(medv ~ . - age - indus)


Residuals vs Leverage
215
415
11
Cook's distance
1
0.5
0.5
Standardized residuals
-2
0
2
4
6
0.00
0.05
0.10
0.15
0.20
0.25
Leverage
lm(medv ~ . - age - indus)

```
summary(train_model)
```

```
##
## Call:
## lm(formula = medv ~ . - age - indus, data = train, weights = 1/vari_train)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9794 -1.2572 -0.1029  1.0030  9.2327
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.339574   4.469960   6.340 7.71e-10 ***
## crim         -0.098675   0.060431  -1.633 0.103472
## zn            0.052286   0.010644   4.912 1.43e-06 ***
## nox         -12.341275   3.059921  -4.033 6.87e-05 ***
## rm            4.189700   0.431065   9.719  < 2e-16 ***
## dis          -1.106000   0.142753  -7.748 1.21e-13 ***
## rad           0.224960   0.064416   3.492 0.000545 ***
## tax          -0.014412   0.003155  -4.568 7.02e-06 ***
## ptratio      -0.717106   0.096408  -7.438 9.23e-13 ***
## lstat        -0.402143   0.039449 -10.194  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.863 on 324 degrees of freedom
## Multiple R-squared:  0.7738, Adjusted R-squared:  0.7675
## F-statistic: 123.2 on 9 and 324 DF,  p-value: < 2.2e-16
```

The RMSE is 4.81027 for the training set. Base don the graph, apart from some inconsistencies, the training model seems to be pretty solid, and the adjusted R squared is 76.75%.

```
sqrt(mean((test$medv - predict(train_model, test)) ^ 2))
```

```
## [1] 5.337472
```

The RMSE here is 5.337472. It does make sense that the testing error is higher than the training error (smaller sample size), but this is still pretty good regardless.

K-fold Cross Validation (5 folds)

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
ctrl <- trainControl(method="cv", number = 5)
cv_model <- train(medv~.-age-indus, data=df_cleaned, weights=1/vari, method="lm", trCont
rol=ctrl)

print(cv_model)
```

```
## Linear Regression
##
## 499 samples
##  11 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 399, 399, 400, 399, 399
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   5.107857  0.6931803  3.344809
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
summary(cv_model)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat, weights = wts)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1323 -1.1779 -0.1306  1.1008  8.2434
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.750084   3.921150   6.312 6.19e-10 ***
## crim         -0.125780   0.055028  -2.286   0.0227 *
## zn            0.031960   0.009733   3.284   0.0011 **
## nox         -14.628891   2.783764  -5.255 2.21e-07 ***
## rm            4.933354   0.375653  13.133  < 2e-16 ***
## dis          -0.893075   0.125822  -7.098 4.48e-12 ***
## rad           0.218204   0.054734   3.987 7.73e-05 ***
## tax          -0.013165   0.002596  -5.071 5.62e-07 ***
## ptratio      -0.816931   0.088379  -9.244  < 2e-16 ***
## lstat        -0.305002   0.033672  -9.058  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.944 on 489 degrees of freedom
## Multiple R-squared:  0.7374, Adjusted R-squared:  0.7325
## F-statistic: 152.5 on 9 and 489 DF,  p-value: < 2.2e-16
```

The 5-fold cross validation didn't change the adjusted R squared, but it still does a good job of fine tuning the model by switching training and testing as opposed to splitting the data for training and testing since you're only doing it once while k-fold does it 5 times.

Step 11 Provide a short (1 paragraph) summary of your overall conclusions/findings.

From this project and dataset, we were able to learn how to take a dataset and use different methods to create a solid model. The first step was to deal with outliers, then we used multiple variable selection methods to figure out which variables we didn't need. After experimenting with different heteroskedasticity methods, we found one that improved our model and from there, it was just fine tuning our model. At the beginning, our model had 2 insignificant predictors and a 72.25% adjusted R squared and at the end, we got a 73.25% adjusted R squared and 0 insignificant predictors.