DATA621-Homework3-HoddeFarrisBurmood

Rob Hodde, Matt Farris, JeffreyBurmood

3/28/2017

DATA621 Homework #3

Team Members: Rob Hodde, Matt Farris, Jeffrey Burmood

Problem Description

Explore, analyze and model a data set containing information on crime for various neighborhoods of a major city. Using the data set build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. Provide classifications and probabilities for the evaluation data set using the developed binary logistic regression model.

Data Exploration

```
# Load required libraries
library(ggplot2)
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(RCurl)
```

```
## Loading required package: bitops
```

```
# Read in the dataset from github
crime <- read.csv(text=getURL("https://raw.githubusercontent.com/jeffreyburmood/data621/master/Homework3
crime_eval <- read.csv(text=getURL("https://raw.githubusercontent.com/jeffreyburmood/data621/master/Hom
# First, get a general look at the data
head(crime)
```

```
##   zn indus chas   nox    rm   age    dis rad tax ptratio  black lstat medv
## 1  0 19.58    0 0.605 7.929  96.2 2.0459   5 403    14.7 369.30  3.70 50.0
## 2  0 19.58    1 0.871 5.403 100.0 1.3216   5 403    14.7 396.90 26.82 13.4
## 3  0 18.10    0 0.740 6.485 100.0 1.9784  24 666    20.2 386.73 18.85 15.4
## 4 30  4.93    0 0.428 6.393   7.8 7.0355   6 300    16.6 374.71  5.19 23.7
```

```
## 5  0  2.46    0 0.488 7.155  92.2 2.7006   3 193    17.8 394.12  4.82 37.9
## 6  0  8.56    0 0.520 6.781  71.3 2.8561   5 384    20.9 395.58  7.67 26.5
##   target
## 1      1
## 2      1
## 3      1
## 4      0
## 5      0
## 6      0
```

```
# Let's start by exploring the type of each variable
types <- sapply(1:length(crime),function(x) typeof(crime[,x]))
types.df <- data.frame(VAR=names(crime),TYPE=types)
types.df
```

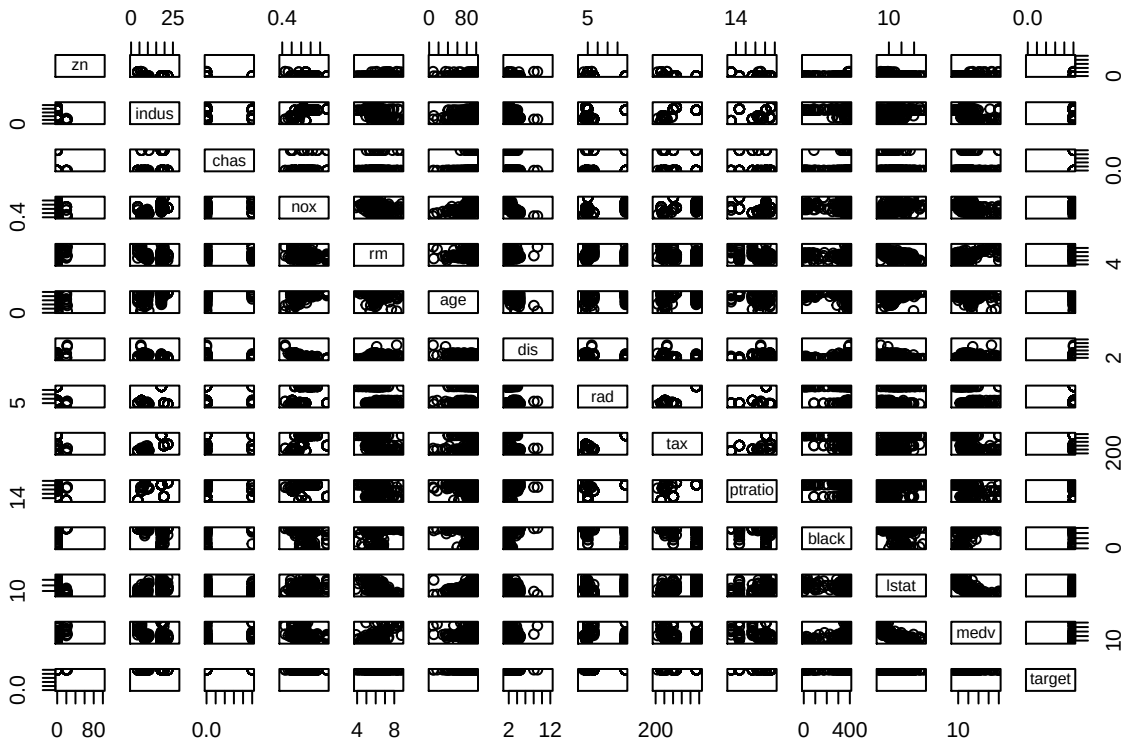
```
##        VAR    TYPE
## 1       zn  double
## 2    indus  double
## 3     chas integer
## 4      nox  double
## 5       rm  double
## 6      age  double
## 7      dis  double
## 8      rad integer
## 9      tax integer
## 10 ptratio  double
## 11   black  double
## 12   lstat  double
## 13    medv  double
## 14  target integer
```

```
# Now generate some summary statistics
print(summary(crime))
```

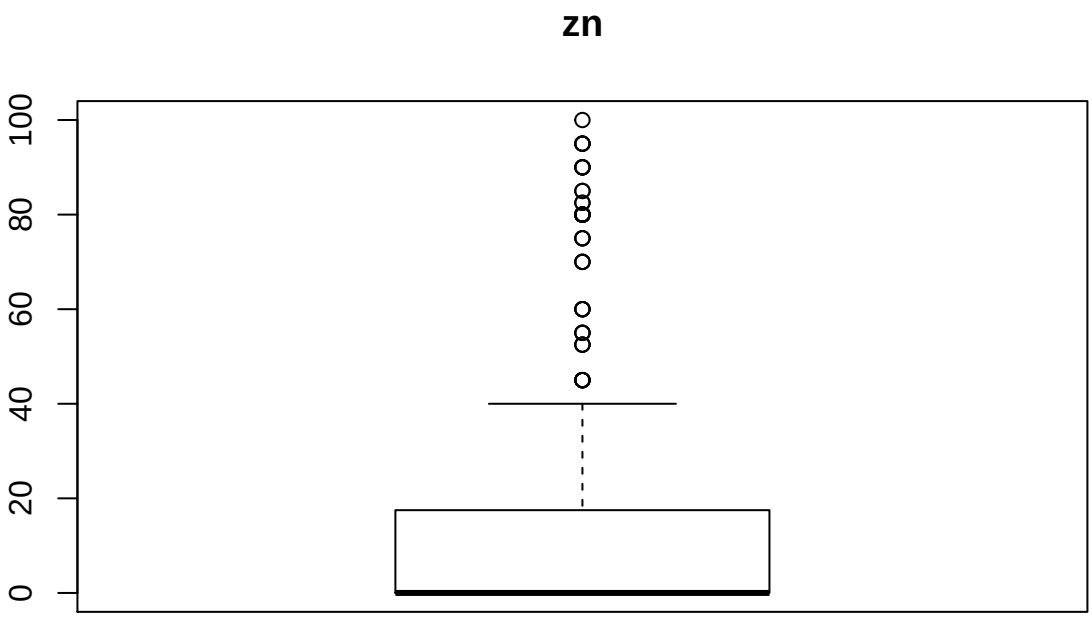
```
##        zn              indus             chas              nox
##  Min.   :  0.00   Min.   : 0.460   Min.   :0.00000   Min.   :0.3890
##  1st Qu.:  0.00   1st Qu.: 5.145   1st Qu.:0.00000   1st Qu.:0.4480
##  Median :  0.00   Median : 9.690   Median :0.00000   Median :0.5380
##  Mean   : 11.58   Mean   :11.105   Mean   :0.07082   Mean   :0.5543
##  3rd Qu.: 16.25   3rd Qu.:18.100   3rd Qu.:0.00000   3rd Qu.:0.6240
##  Max.   :100.00   Max.   :27.740   Max.   :1.00000   Max.   :0.8710
##        rm              age              dis              rad
##  Min.   :3.863   Min.   :  2.90   Min.   : 1.130   Min.   : 1.00
##  1st Qu.:5.887   1st Qu.: 43.88   1st Qu.: 2.101   1st Qu.: 4.00
##  Median :6.210   Median : 77.15   Median : 3.191   Median : 5.00
##  Mean   :6.291   Mean   : 68.37   Mean   : 3.796   Mean   : 9.53
##  3rd Qu.:6.630   3rd Qu.: 94.10   3rd Qu.: 5.215   3rd Qu.:24.00
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.00
##       tax           ptratio          black            lstat
##  Min.   :187.0   Min.   :12.6   Min.   :  0.32   Min.   : 1.730
##  1st Qu.:281.0   1st Qu.:16.9   1st Qu.:375.61   1st Qu.: 7.043
##  Median :334.5   Median :18.9   Median :391.34   Median :11.350
##  Mean   :409.5   Mean   :18.4   Mean   :357.12   Mean   :12.631
```

```
##  3rd Qu.:666.0   3rd Qu.:20.2   3rd Qu.:396.24   3rd Qu.:16.930
##  Max.   :711.0   Max.   :22.0   Max.   :396.90   Max.   :37.970
##       medv           target
##  Min.   : 5.00   Min.   :0.0000
##  1st Qu.:17.02   1st Qu.:0.0000
##  Median :21.20   Median :0.0000
##  Mean   :22.59   Mean   :0.4914
##  3rd Qu.:25.00   3rd Qu.:1.0000
##  Max.   :50.00   Max.   :1.0000
```

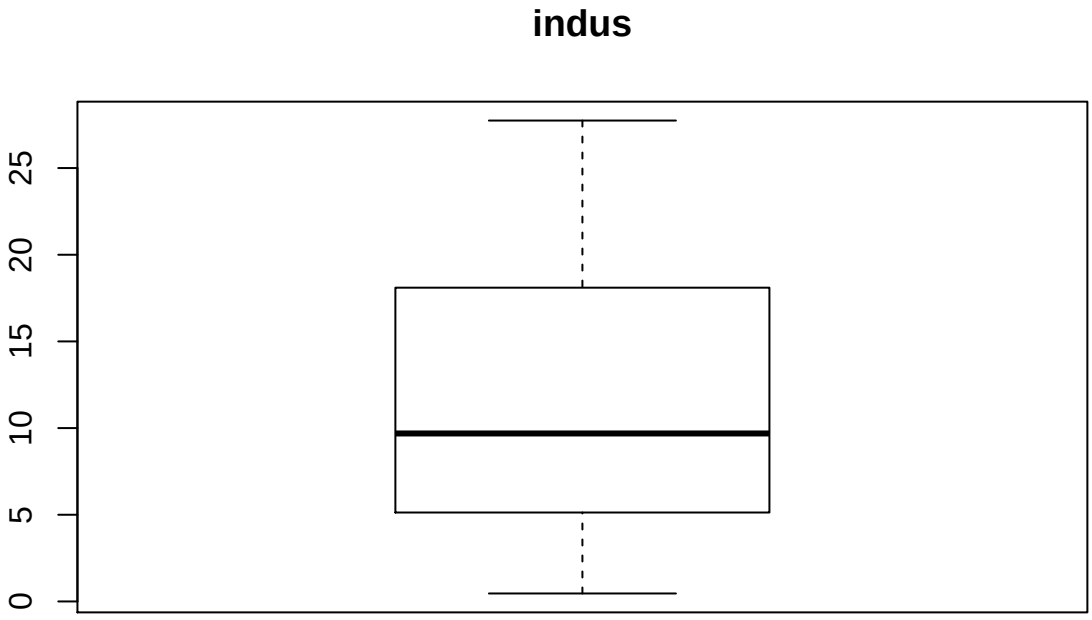
```
# Visual check for obvious correlations
pairs(crime,col=crime$target)
```



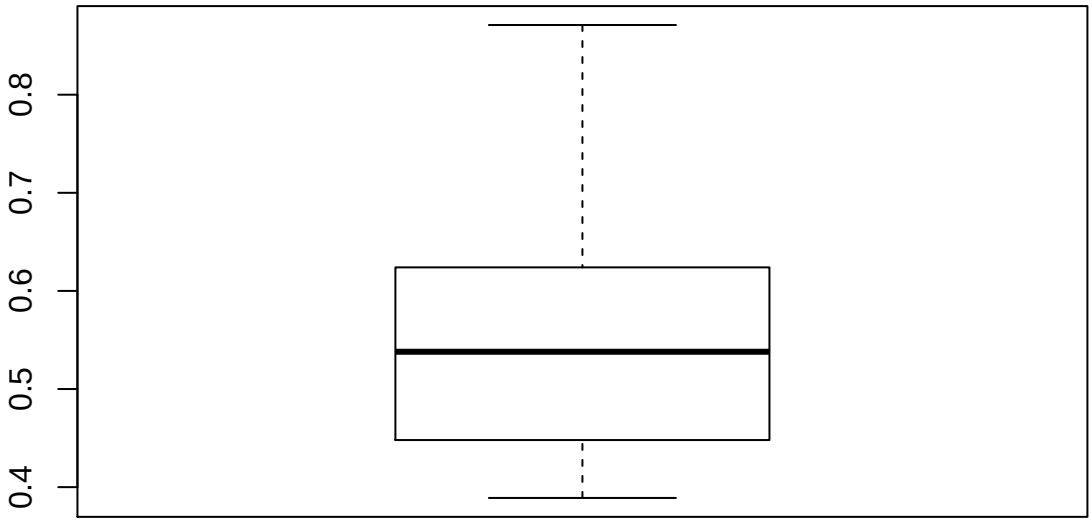

```
#
# no NAs found so no missing values to remove or fix?
#
# Look over the variables checking for outliers/influencial points, correlation between variables, etc.
#
boxplot(crime$zn, main="zn")
```



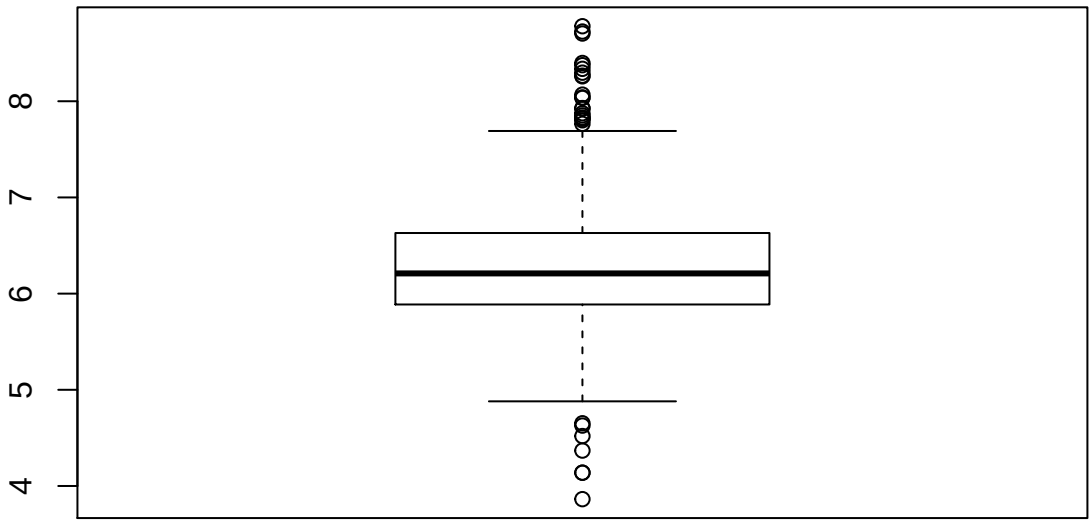

```
boxplot(crime$indus, main="indus")
```



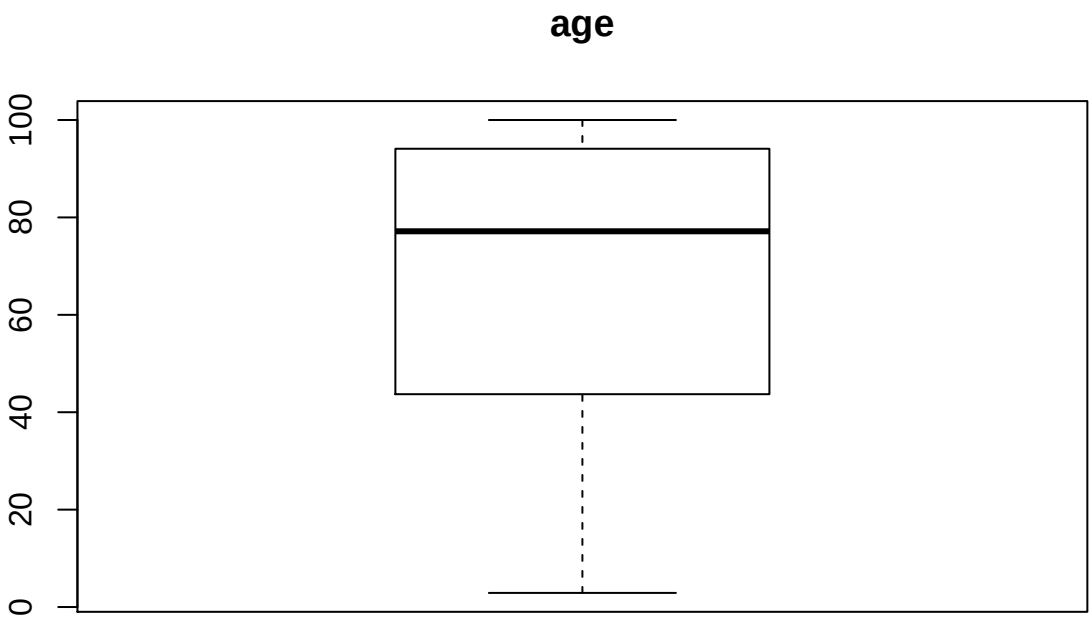

```
boxplot(crime$nox, main="nox")
```



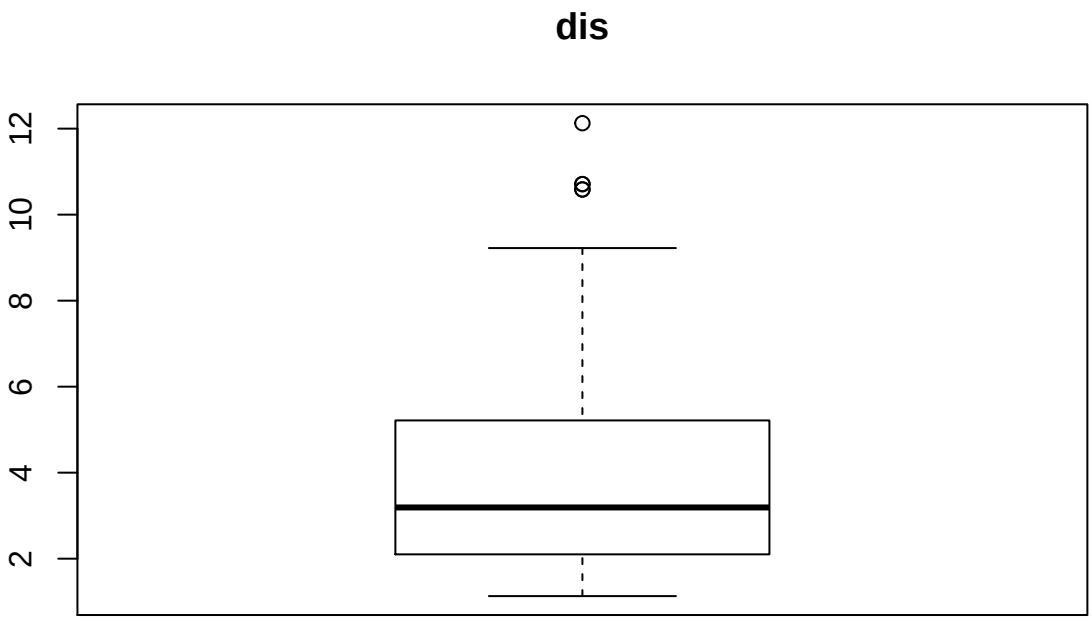

```
boxplot(crime$rm, main="rm")
```



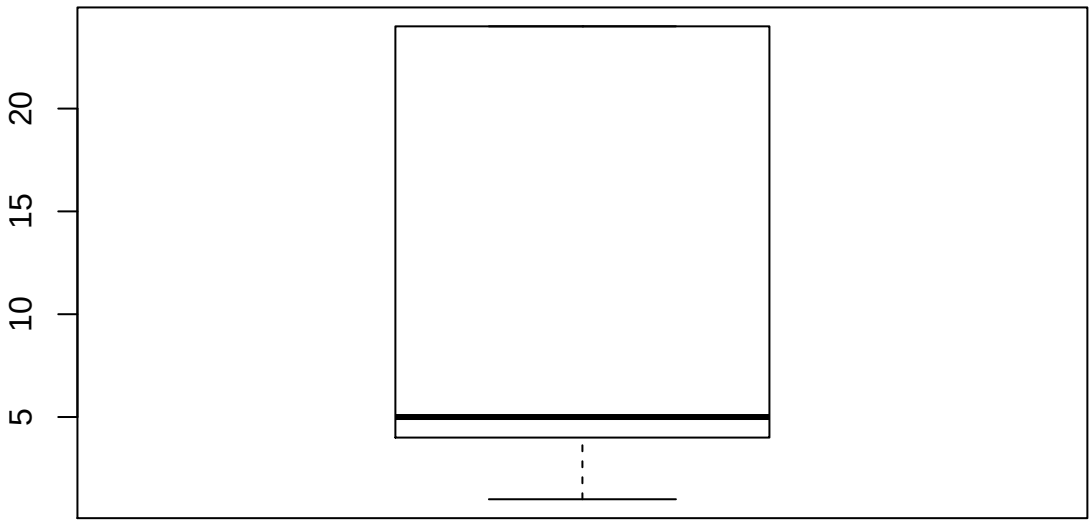

```
boxplot(crime$age, main="age")
```



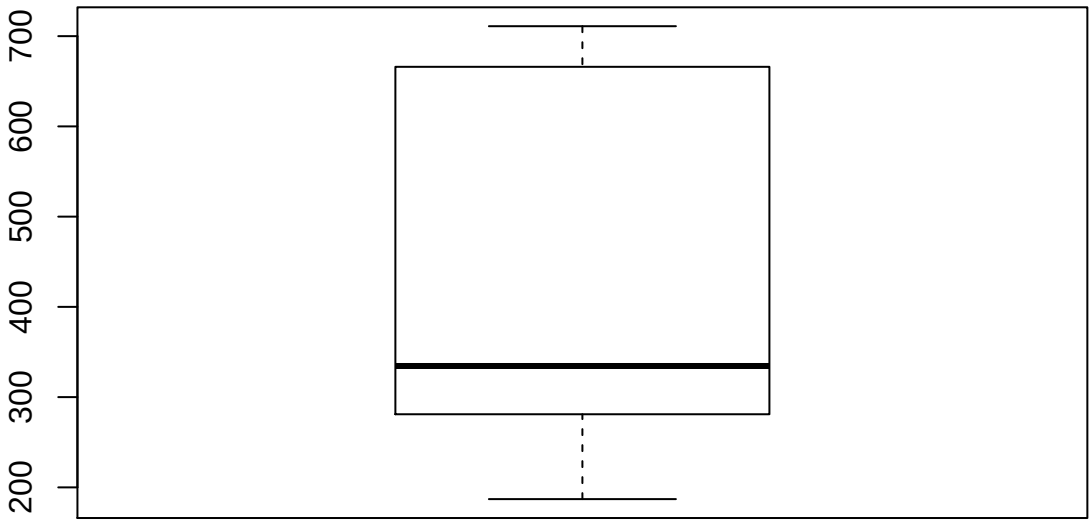

```
boxplot(crime$dis, main="dis")
```



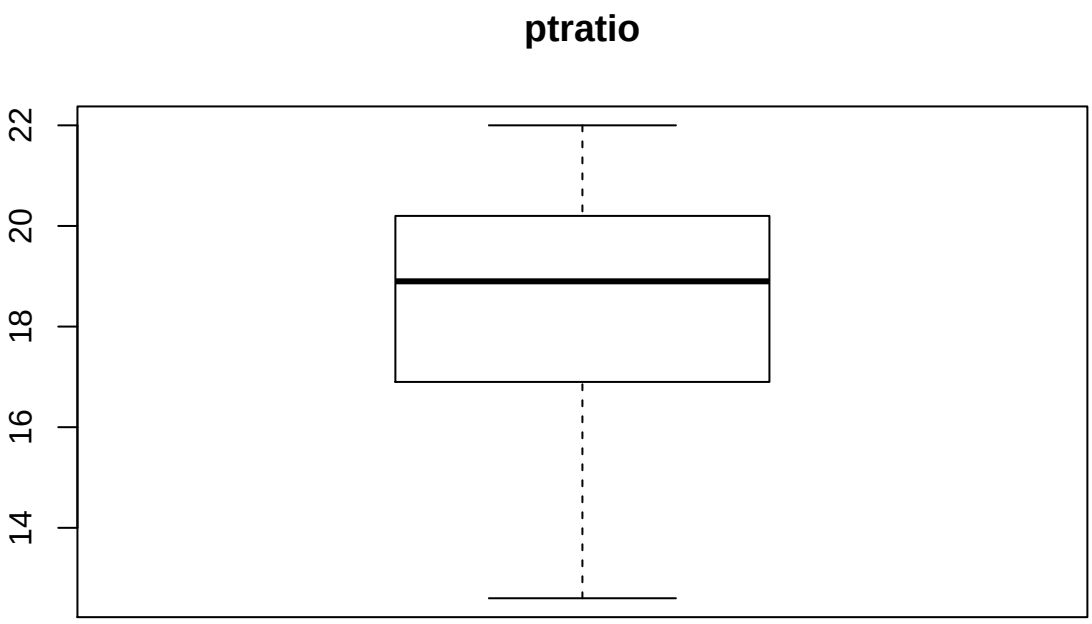

```
boxplot(crime$rad, main="rad")
```



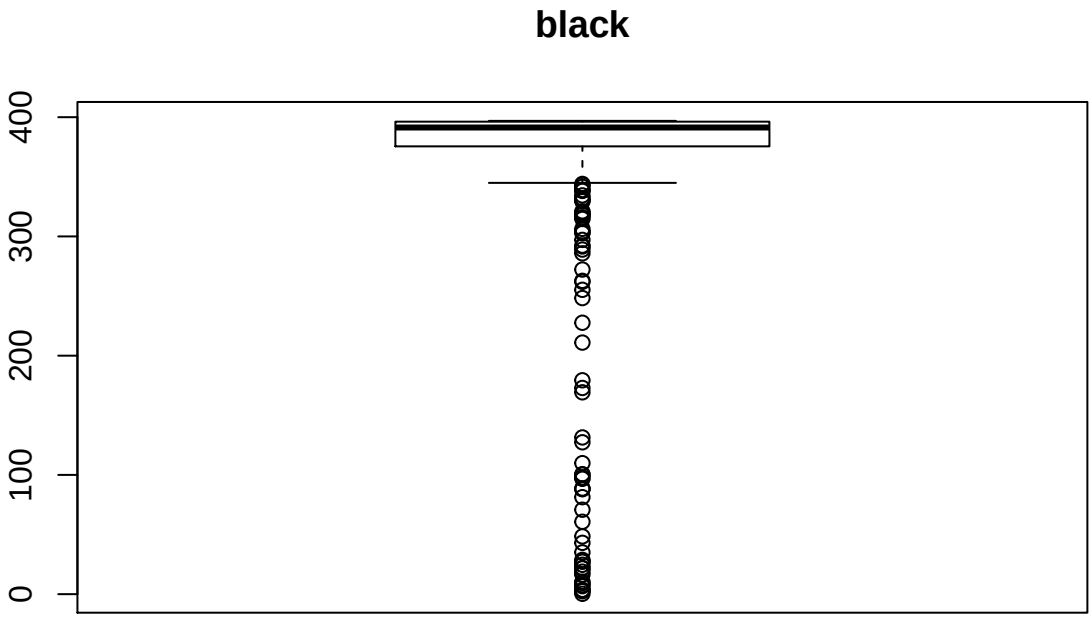

```
boxplot(crime$tax, main="tax")
```



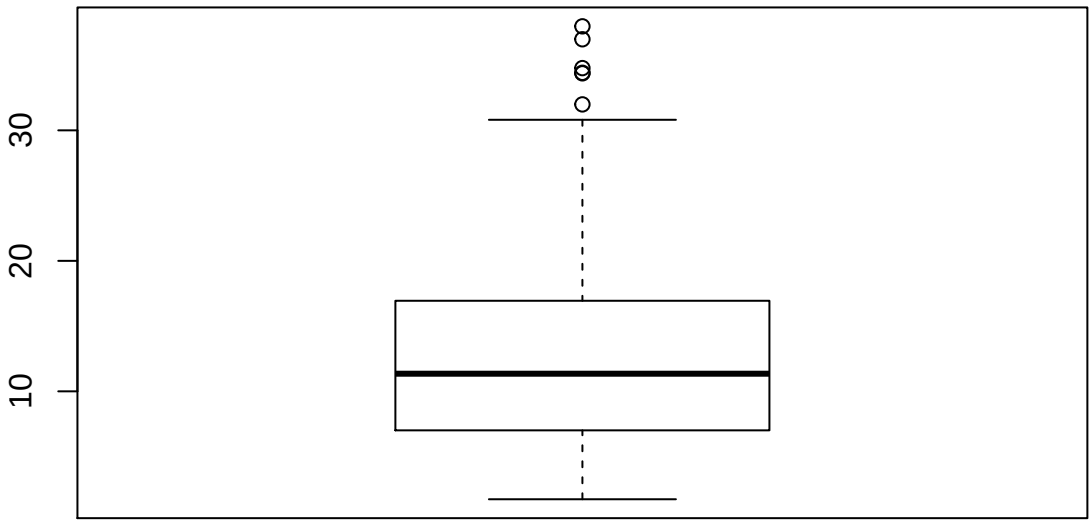

```
boxplot(crime$ptratio, main="ptratio")
```



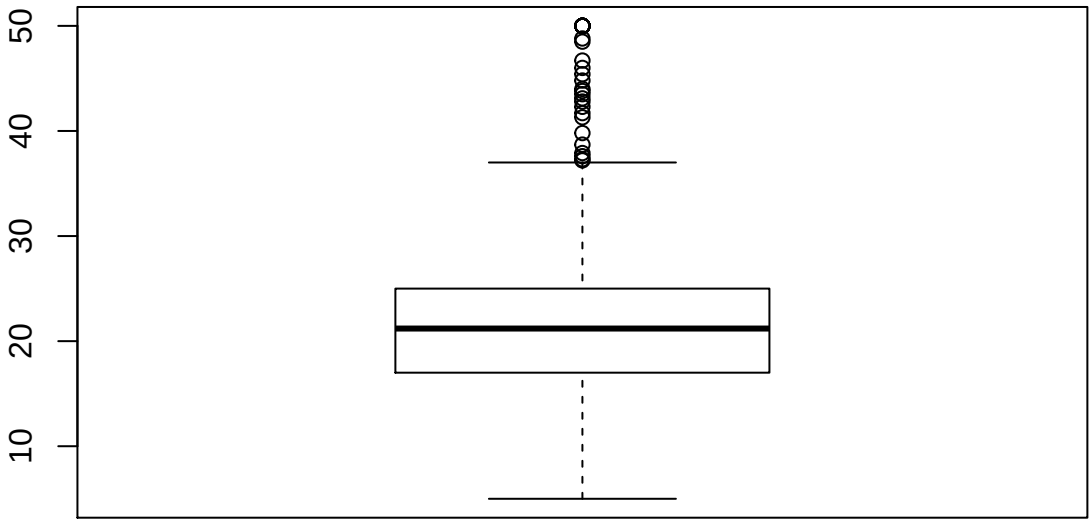

```
boxplot(crime$black, main="black")
```



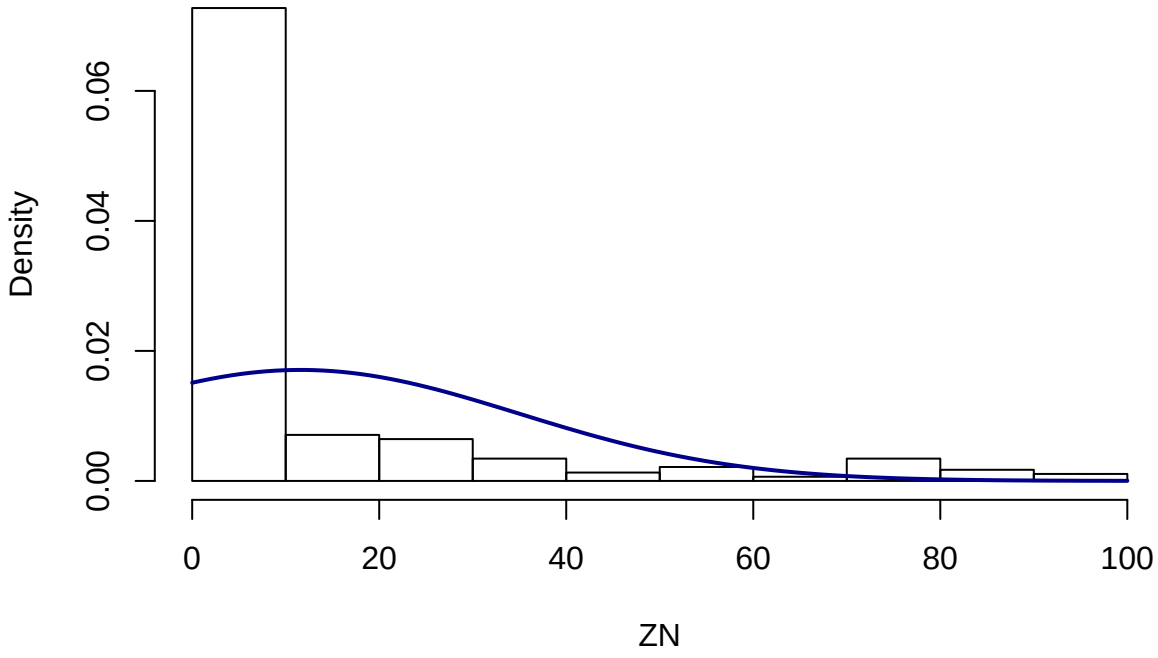

```
boxplot(crime$lstat, main="lstat")
```



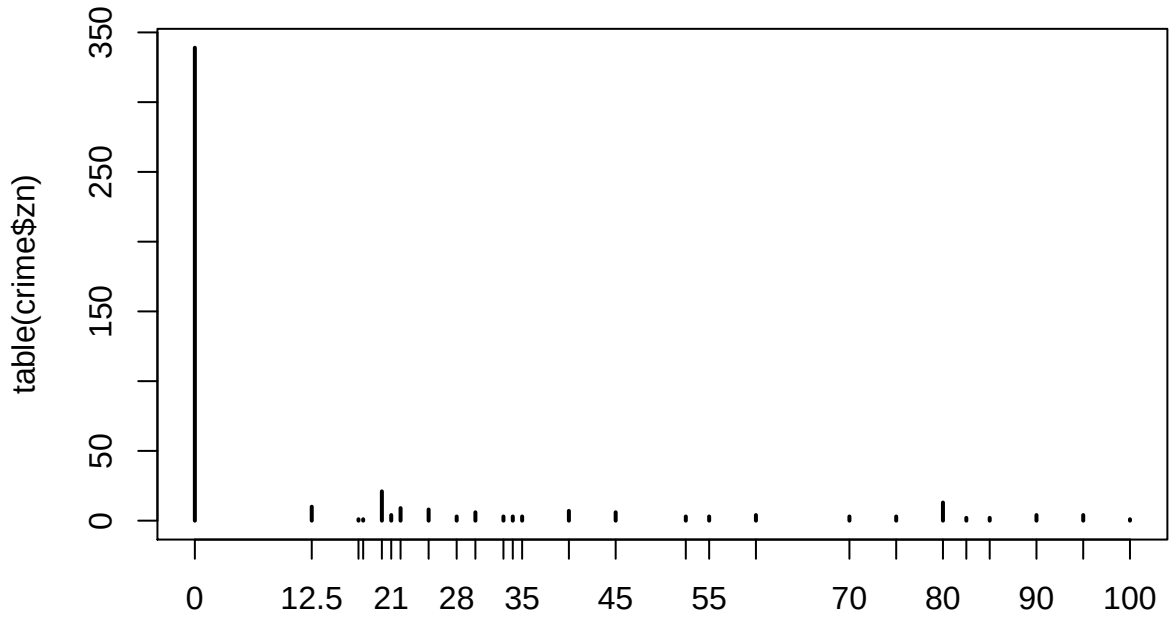

```
boxplot(crime$medv, main="mdev")
```



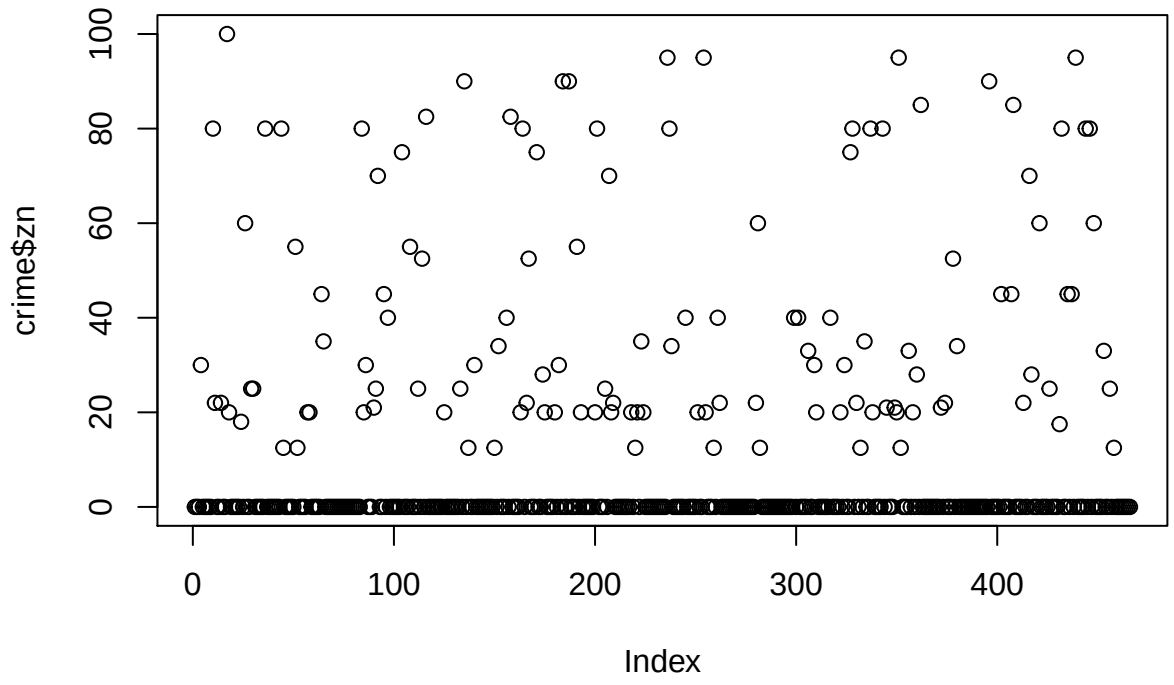

```
#
# the following variables look like they have some outliers,
# zn,rm, dis, black, lstat, medv so let's look at their historgrams

# zn
#zn.plot <- ggplot(crime, aes(x=zn,color=target)) + geom_histogram(position="dodge",binwidth=5)
#print(zn.plot)
m <- mean(crime$zn)
s <- sd(crime$zn)
hist(crime$zn,prob=TRUE,xlab="ZN",main='')
curve(dnorm(x,mean=m,sd=s),col="darkblue",lwd=2,add=TRUE)
```



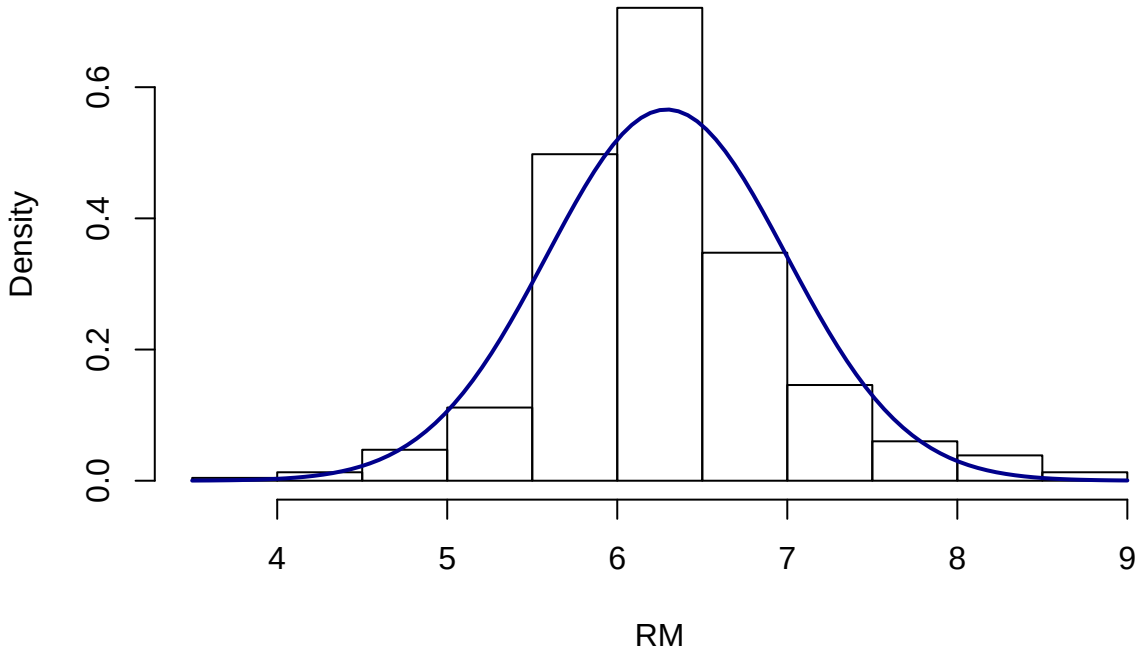

```
# zn is so skewed, let's look at a frequency count
plot(table(crime$zn))
```



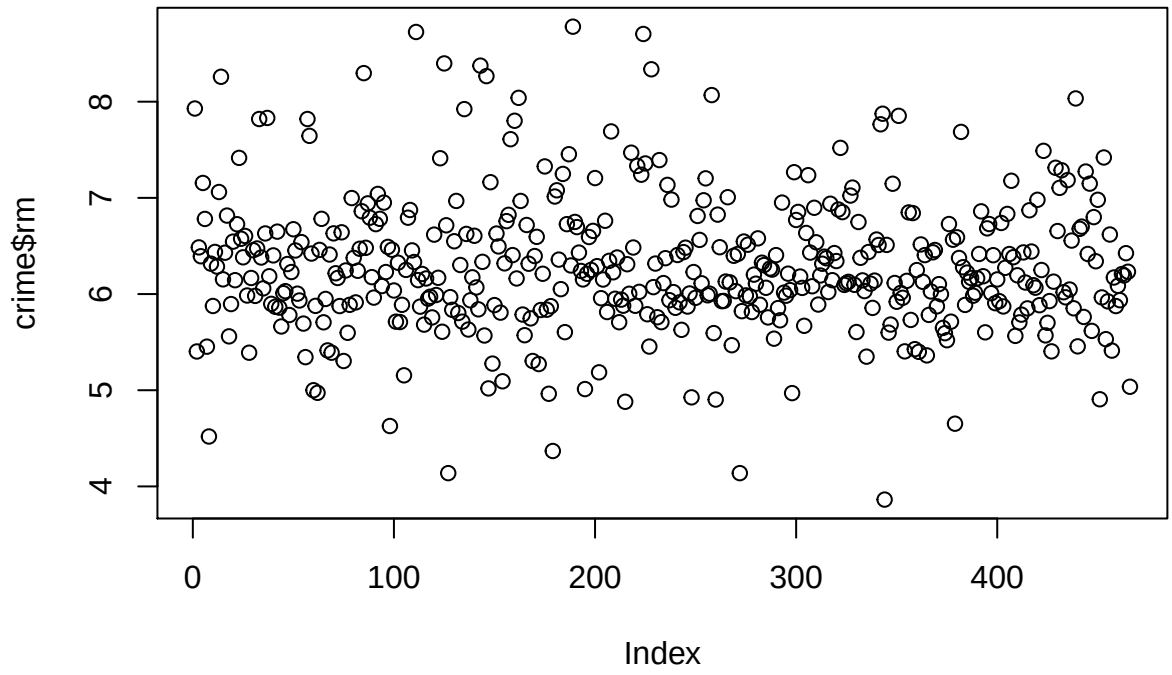

```
# let's look at a plot of the values
plot(crime$zn)
```



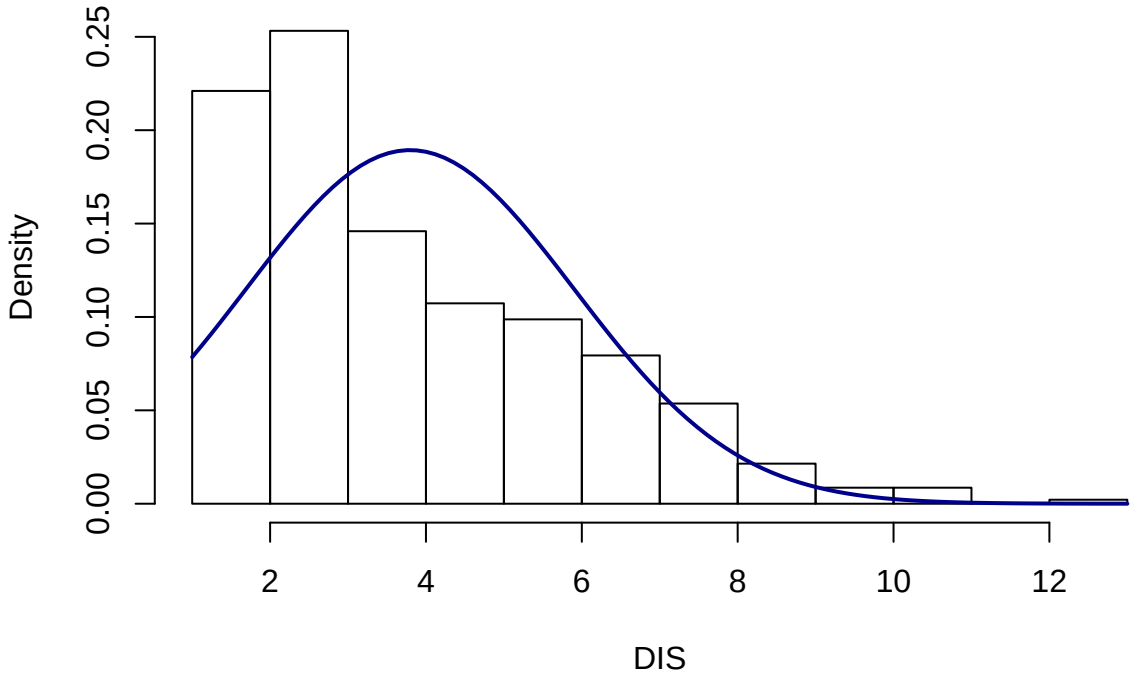

```
# rm
m <- mean(crime$rm)
s <- sd(crime$rm)
hist(crime$rm,prob=TRUE,xlab="RM",main='')
curve(dnorm(x,mean=m,sd=s),col="darkblue",lwd=2,add=TRUE)
```



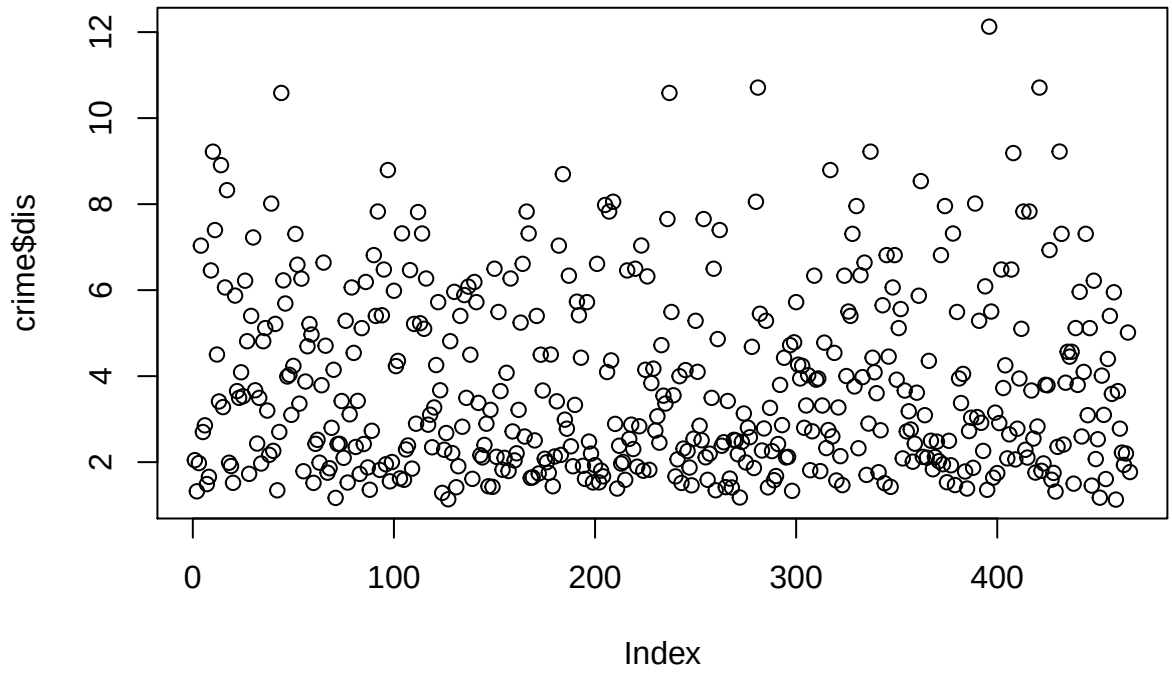

```
# let's look at a plot of the values
plot(crime$rm)
```



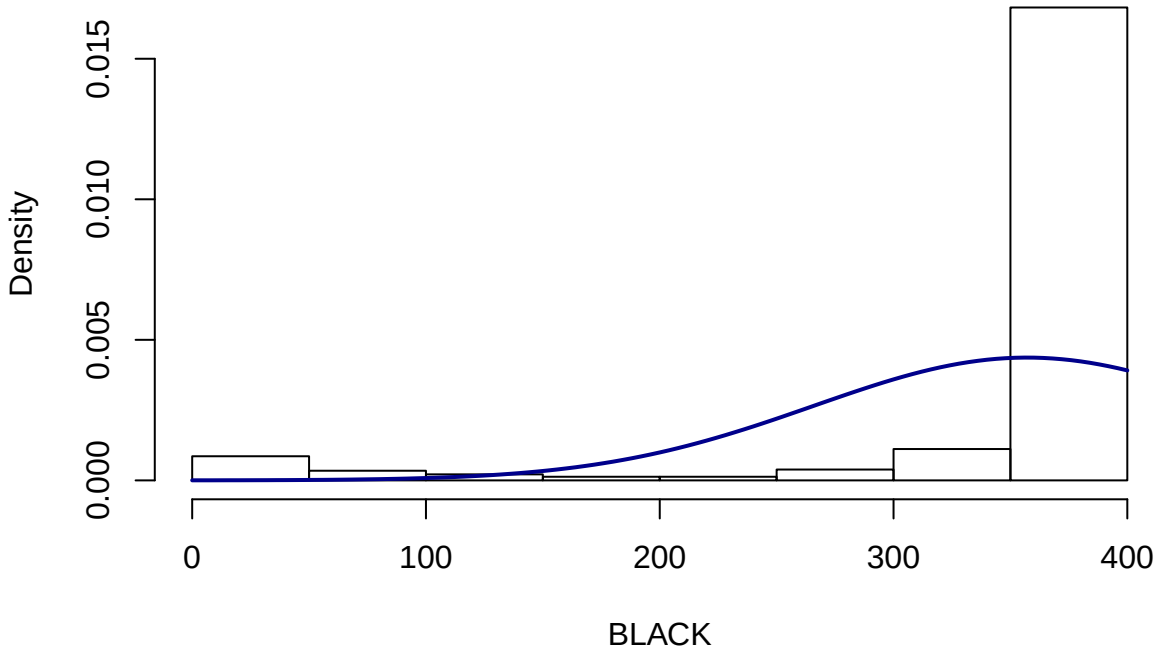

```
# dis
m <- mean(crime$dis)
s <- sd(crime$dis)
hist(crime$dis,prob=TRUE,xlab="DIS",main='')
curve(dnorm(x,mean=m,sd=s),col="darkblue",lwd=2,add=TRUE)
```



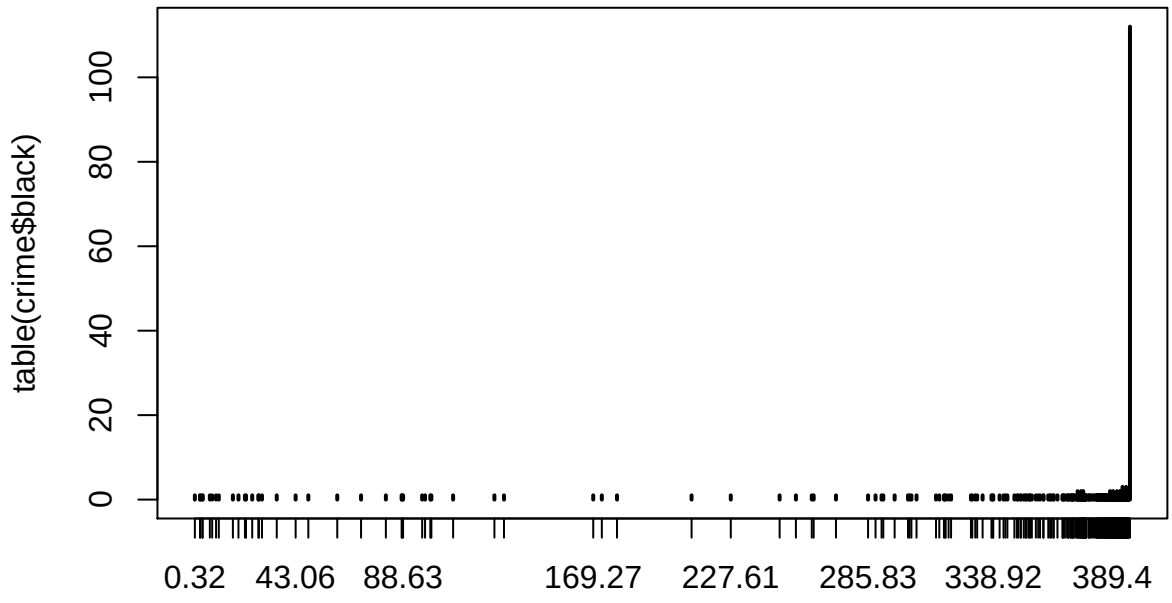

```
# let's look at a plot of the values
plot(crime$dis)
```



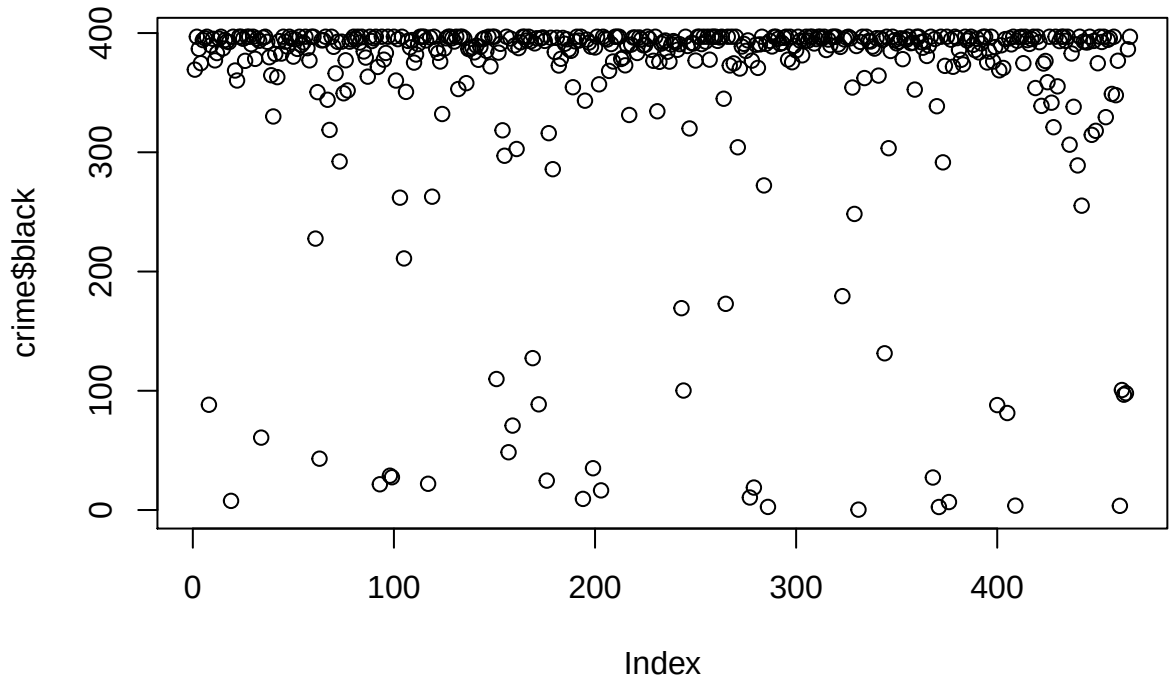

```
# black
m <- mean(crime$black)
s <- sd(crime$black)
hist(crime$black,prob=TRUE,xlab="BLACK",main='')
curve(dnorm(x,mean=m,sd=s),col="darkblue",lwd=2,add=TRUE)
```



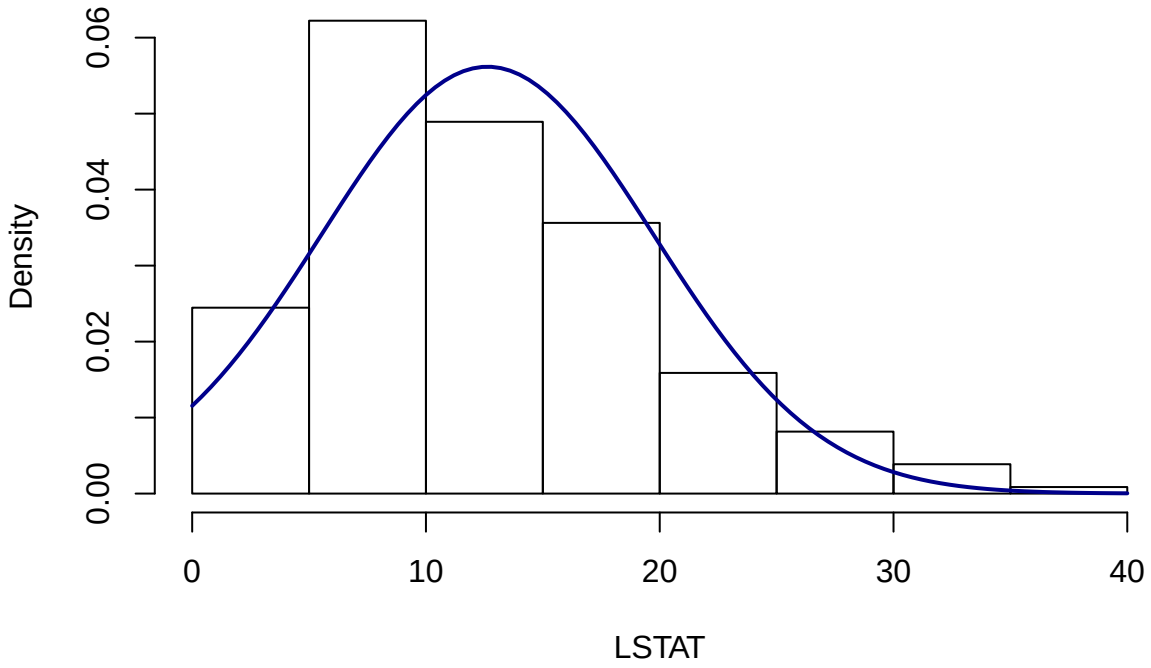

```
# black is so skewed, let's look at a frequency count
plot(table(crime$black))
```



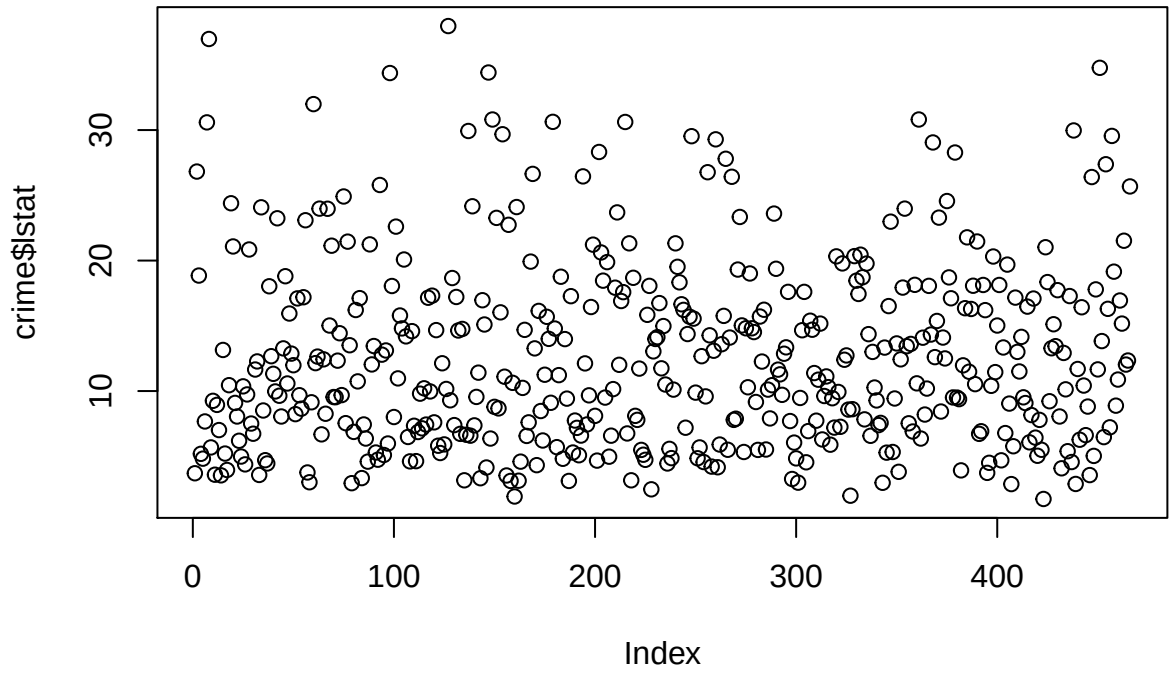

```
# let's look at a plot of the values
plot(crime$black)
```



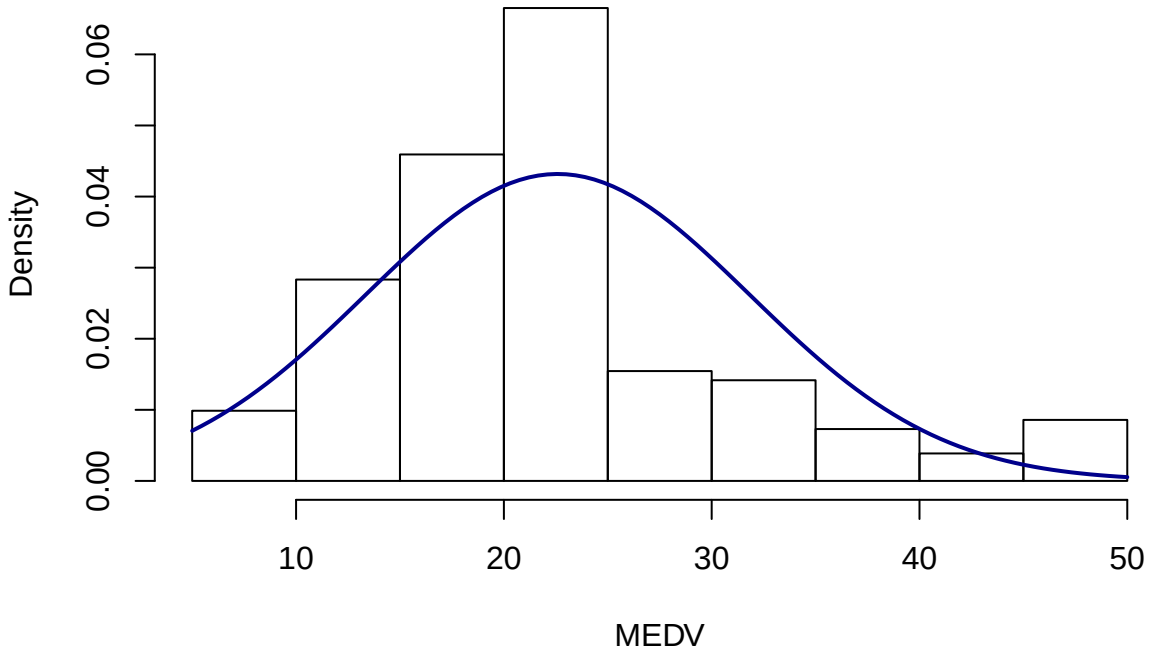

```
# lstat
m <- mean(crime$lstat)
s <- sd(crime$lstat)
hist(crime$lstat,prob=TRUE,xlab="LSTAT",main='')
curve(dnorm(x,mean=m,sd=s),col="darkblue",lwd=2,add=TRUE)
```



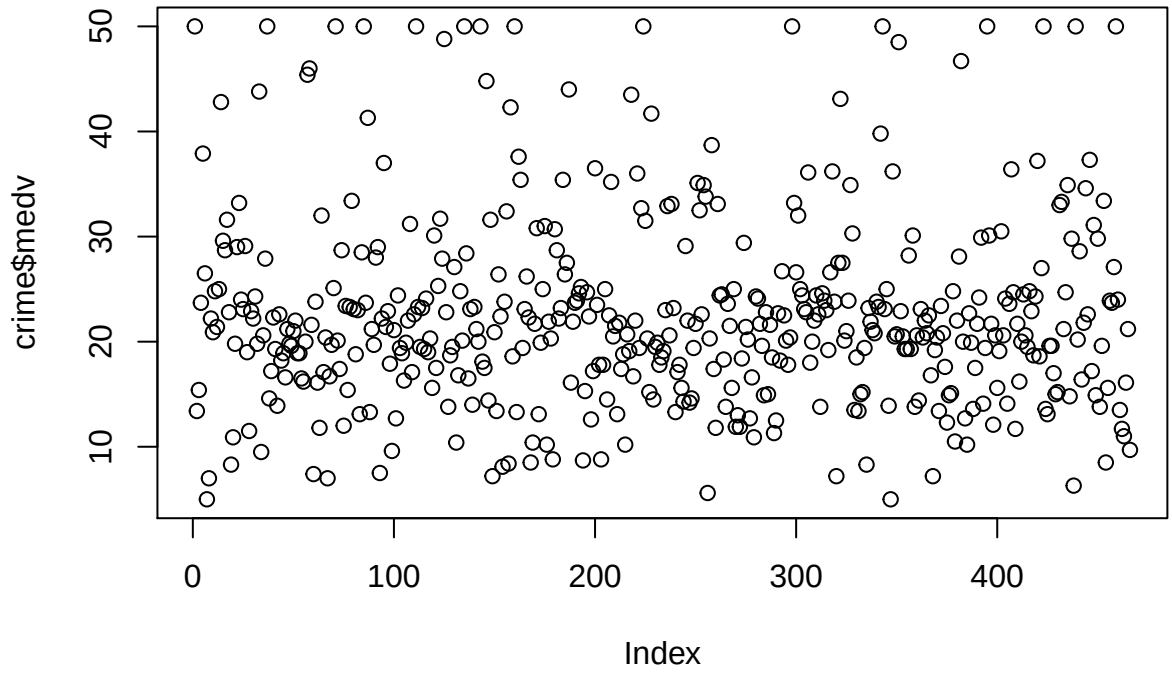

```
# let's look at a plot of the values
plot(crime$lstat)
```




```
# medv
m <- mean(crime$medv)
s <- sd(crime$medv)
hist(crime$medv,prob=TRUE,xlab="MEDV",main='')
curve(dnorm(x,mean=m,sd=s),col="darkblue",lwd=2,add=TRUE)
```




```
# let's look at a plot of the values
plot(crime$medv)
```




```
# quick look at model with all variables
crime.model <- glm(target ~ .,family=binomial(link='logit'),data=crime)
print(summary(crime.model))
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
```

```
##     data = crime)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2854  -0.1372  -0.0017   0.0020   3.4721
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -36.839521   7.028726  -5.241 1.59e-07 ***
## zn           -0.061720   0.034410  -1.794 0.072868 .
## indus        -0.072580   0.048546  -1.495 0.134894
## chas          1.032352   0.759627   1.359 0.174139
## nox          50.159513   8.049503   6.231 4.62e-10 ***
## rm           -0.692145   0.741431  -0.934 0.350548
## age           0.034522   0.013883   2.487 0.012895 *
## dis           0.765795   0.234407   3.267 0.001087 **
## rad           0.663015   0.165135   4.015 5.94e-05 ***
## tax          -0.006593   0.003064  -2.152 0.031422 *
## ptratio       0.442217   0.132234   3.344 0.000825 ***
## black        -0.013094   0.006680  -1.960 0.049974 *
## lstat         0.047571   0.054508   0.873 0.382802
## medv          0.199734   0.071022   2.812 0.004919 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 186.15  on 452  degrees of freedom
## AIC: 214.15
##
## Number of Fisher Scoring iterations: 9
```

According to the description, the variables zn, indus, and age are area, or land, proportions. According to the statistical summary, the values for these variables are all within the range [1,100] as you would expect.

The predictor variable zn is highly right skewed, we can confirm this by comparing the median and mean where the median is 0.0, but the median is 11.58. The frequency count plot shows how poor the distribution is due to clustering of the data at one extreme.

The predictor variable black is highly left skewed. We can confirm this by comparing the median and mean where the median is 391.34 and the mean is 357.12. The frequency count plot shows how poor the distribution is due to clustering of the data at one extreme.

The predictor variable dis is slightly right skewed. We can confirm this by comparing the median and mean where the median is 3.191 and the mean is 3.796.

Fortunately, no missing data, or NAs, were found.

The following data corrections were identified in this section:

(1) The predictor variable "chas" and the response variable "target" are supposed to be categorical (binary), so we need to convert them to factors.

(2) Need to determine if there are other variables highly coorelated with the zn or black variable that don't have the severe skew and outliers. This would allow us to remove the zn or black variable from the model.

Why is the tax rate an integer and not a numeric???

Data Preparation

The variable changes we identified so far include converting the predictor variable "chas" and the response variable "target" to factors.

```
# Based on the data exploration results, identify any changes, transformations, and new or deleted vari
# Need to set variables to a factor as required
crime$target <- as.factor(crime$target)
crime$chas <- as.factor(crime$chas)
crime_eval$chas <- as.factor(crime_eval$chas)

# get a table of non-factor variables
crime.nofactor <- subset(crime,select=-c(chas,target))
# build a correlation table to study the variable relationships
cor.table <- cor(crime.nofactor) # build a table of inter-variable correlation values
(cor.table)
```

```
##                zn      indus        nox         rm        age        dis
## zn       1.0000000 -0.5382664 -0.5170452  0.3198141 -0.5725805  0.6601243
## indus   -0.5382664  1.0000000  0.7596301 -0.3927118  0.6395818 -0.7036189
## nox     -0.5170452  0.7596301  1.0000000 -0.2954897  0.7351278 -0.7688840
## rm       0.3198141 -0.3927118 -0.2954897  1.0000000 -0.2328125  0.1990158
## age     -0.5725805  0.6395818  0.7351278 -0.2328125  1.0000000 -0.7508976
## dis      0.6601243 -0.7036189 -0.7688840  0.1990158 -0.7508976  1.0000000
## rad     -0.3154812  0.6006284  0.5958298 -0.2084457  0.4603143 -0.4949919
## tax     -0.3192841  0.7322292  0.6538780 -0.2969343  0.5121245 -0.5342546
## ptratio -0.3910357  0.3946898  0.1762687 -0.3603471  0.2554479 -0.2333394
## black    0.1794150 -0.3581356 -0.3801549  0.1326676 -0.2734677  0.2938441
## lstat   -0.4329925  0.6071102  0.5962426 -0.6320245  0.6056200 -0.5075280
## medv     0.3767171 -0.4961743 -0.4301227  0.7053368 -0.3781560  0.2566948
##                rad        tax    ptratio      black      lstat       medv
## zn      -0.3154812 -0.3192841 -0.3910357  0.1794150 -0.4329925  0.3767171
## indus    0.6006284  0.7322292  0.3946898 -0.3581356  0.6071102 -0.4961743
## nox      0.5958298  0.6538780  0.1762687 -0.3801549  0.5962426 -0.4301227
## rm      -0.2084457 -0.2969343 -0.3603471  0.1326676 -0.6320245  0.7053368
## age      0.4603143  0.5121245  0.2554479 -0.2734677  0.6056200 -0.3781560
## dis     -0.4949919 -0.5342546 -0.2333394  0.2938441 -0.5075280  0.2566948
## rad      1.0000000  0.9064632  0.4714516 -0.4463750  0.5031013 -0.3976683
## tax      0.9064632  1.0000000  0.4744223 -0.4425059  0.5641886 -0.4900329
## ptratio  0.4714516  0.4744223  1.0000000 -0.1816395  0.3773560 -0.5159153
## black   -0.4463750 -0.4425059 -0.1816395  1.0000000 -0.3533659  0.3300286
## lstat    0.5031013  0.5641886  0.3773560 -0.3533659  1.0000000 -0.7358008
## medv    -0.3976683 -0.4900329 -0.5159153  0.3300286 -0.7358008  1.0000000
```

Based on the correlation table, the variable zn has a moderate correlation with the variable dis. The plot of the dis data shows a much better distribution of values. Consequently, one possibility is to remove the zn variable from the data set for modeling.

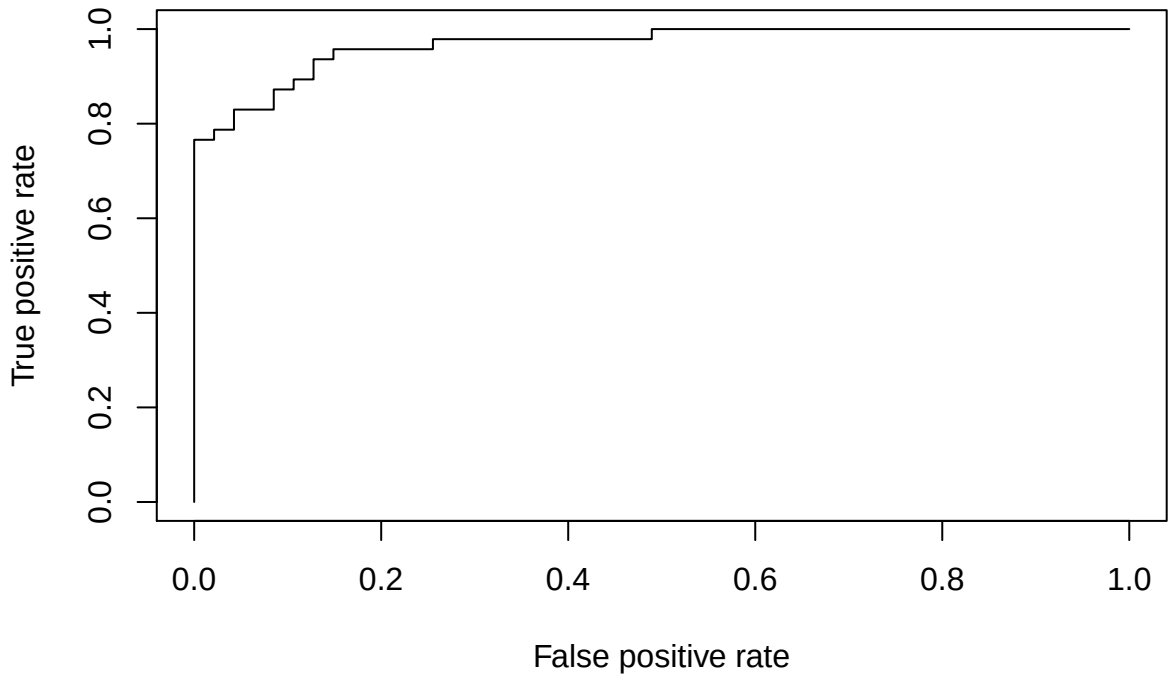
What to do about the black variable???

Build Models

```
## 75% of the sample size
smp_size <- floor(0.80 * nrow(crime))

## set the seed to make your partition reproductible
train_ind <- sample(seq_len(nrow(crime)), size = smp_size)

train <- crime[train_ind, ]
test <- crime[-train_ind, ]

# quick look at model with all variables
qm <- glm(target ~ .,family=binomial(link='logit'),data=train)
print(summary(qm))
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9629  -0.1324  -0.0018   0.0017   3.5793
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -42.297084   8.084705  -5.232 1.68e-07 ***
## zn           -0.055887   0.036129  -1.547 0.121896
## indus        -0.078901   0.057303  -1.377 0.168537
## chas1         0.892907   0.919841   0.971 0.331688
## nox          52.819058   9.285843   5.688 1.28e-08 ***
## rm           -0.870851   0.847559  -1.027 0.304194
## age           0.043100   0.016302   2.644 0.008198 **
## dis           0.911657   0.261870   3.481 0.000499 ***
## rad           0.708860   0.183314   3.867 0.000110 ***
## tax          -0.006169   0.003491  -1.767 0.077209 .
## ptratio       0.569289   0.173162   3.288 0.001010 **
## black        -0.010239   0.006061  -1.689 0.091189 .
## lstat         0.005510   0.066750   0.083 0.934214
## medv          0.230680   0.084445   2.732 0.006301 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.53  on 371  degrees of freedom
## Residual deviance: 146.27  on 358  degrees of freedom
## AIC: 174.27
##
## Number of Fisher Scoring iterations: 9
```

```
p <- predict(qm, newdata=subset(test,select=c(1,2,3,4,5,6,7,8,9,10,11,12,13)), type="response")
pr <- prediction(p, test$target)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```




```
#
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.9674061
```

Select Models

All Done!