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

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

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

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

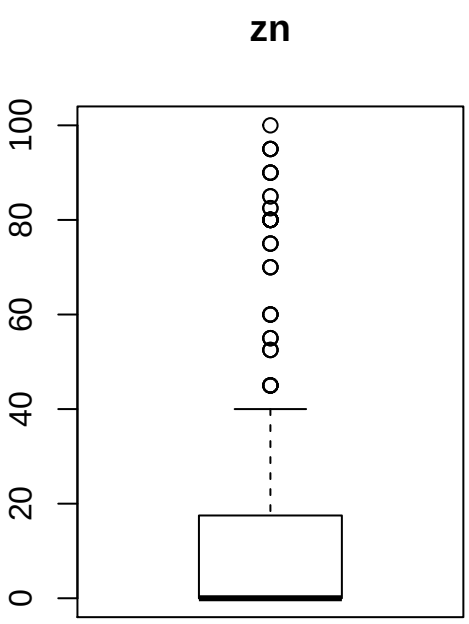
```
## Loading required package: bitops
```

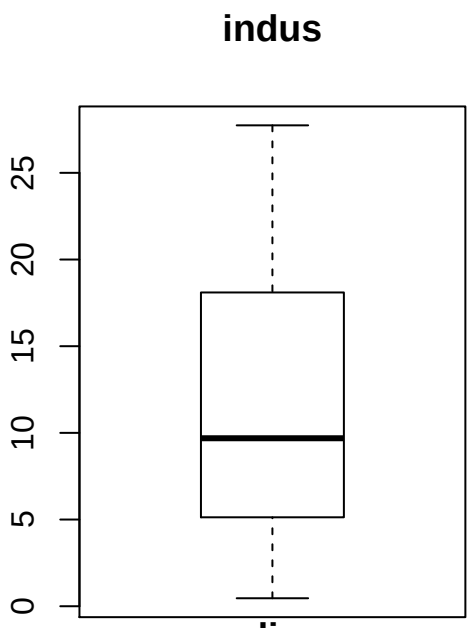
```
##   zn indus chas   nox    rm   age    dis rad tax ptratio  black lstat medv
## 1  0 19.58    0 0.605 7.929  96.2 2.0459   5 403    14.7 369.30  3.70 50.0
## 2  0 19.58    1 0.871 5.403 100.0 1.3216   5 403    14.7 396.90 26.82 13.4
## 3  0 18.10    0 0.740 6.485 100.0 1.9784  24 666    20.2 386.73 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
```

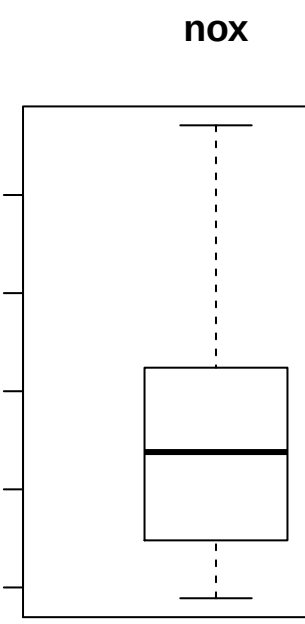
```
##        VAR    TYPE
## 1       zn  double
## 2    indus  double
## 3     chas integer
```

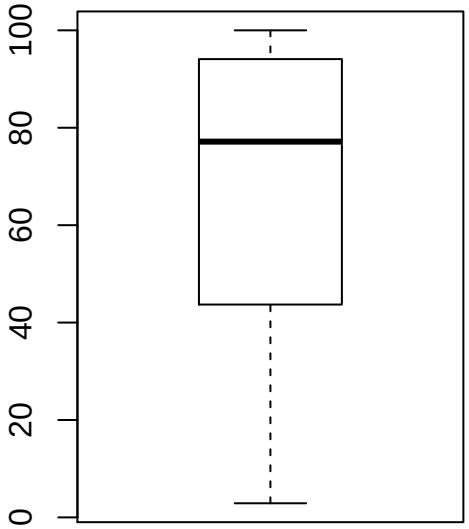
```
## 4      nox  double
## 5       rm  double
## 6      age  double
## 7      dis  double
## 8      rad integer
## 9      tax integer
## 10 ptratio  double
## 11   black  double
## 12   lstat  double
## 13    medv  double
## 14  target integer
```

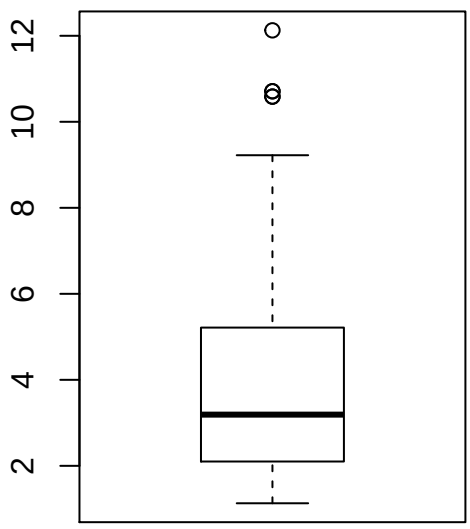
```
##        zn              indus             chas              nox        
##  Min.   :  0.00   Min.   : 0.460   Min.   :0.00000   Min.   :0.3890  
##  1st Qu.:  0.00   1st Qu.: 5.145   1st Qu.:0.00000   1st Qu.:0.4480  
##  Median :  0.00   Median : 9.690   Median :0.00000   Median :0.5380  
##  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
```

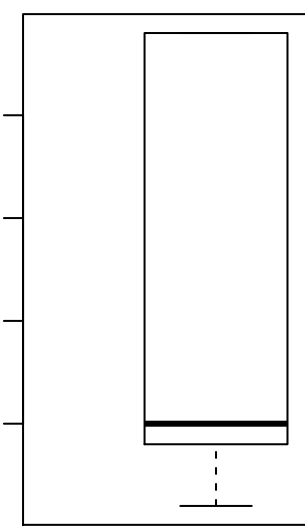

zn
0
20
40
60
80
100

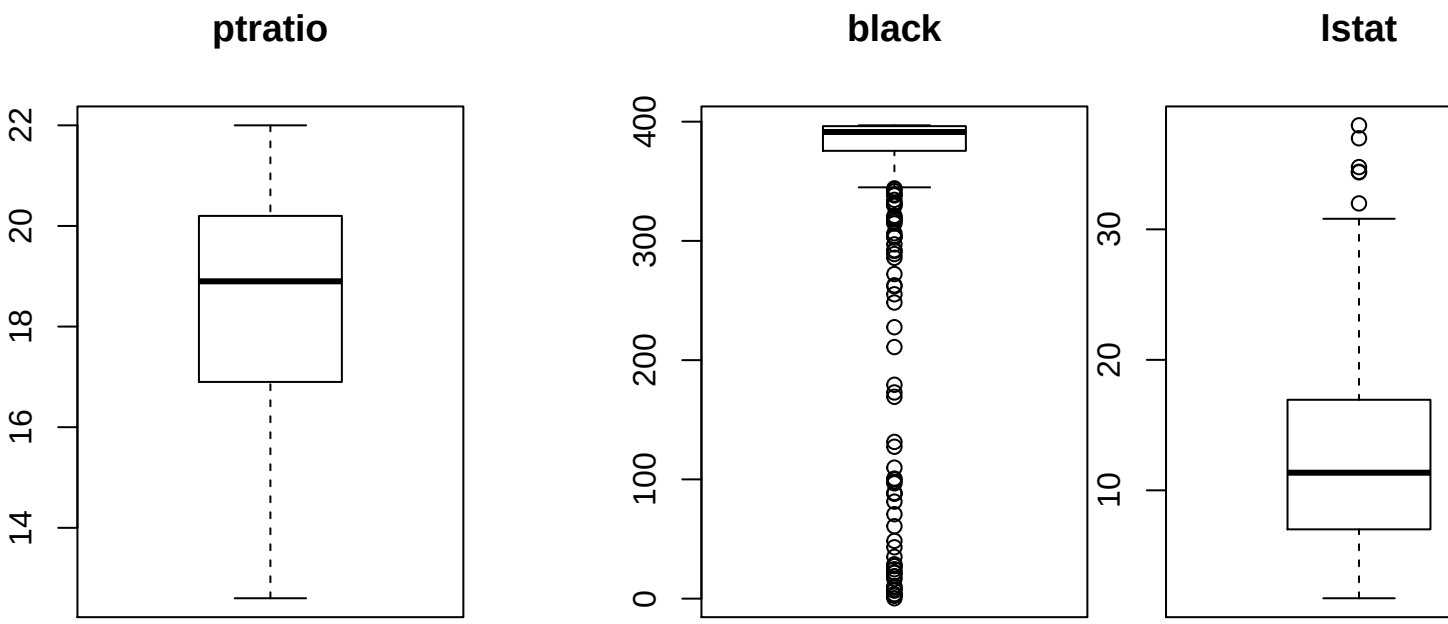

indus
0
5
10
15
20
25


nox
0.4
0.5
0.6
0.7
0.8


age
0
20
40
60
80
100

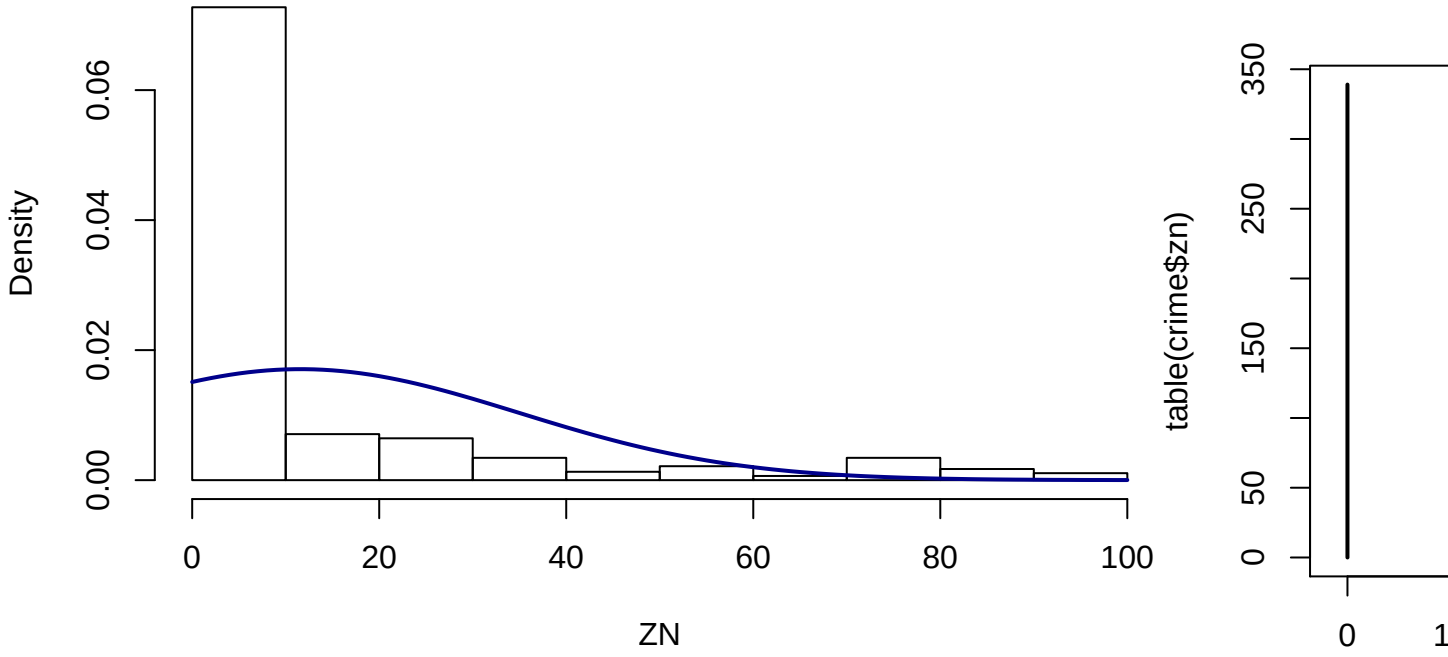

dis
2
4
6
8
10
12

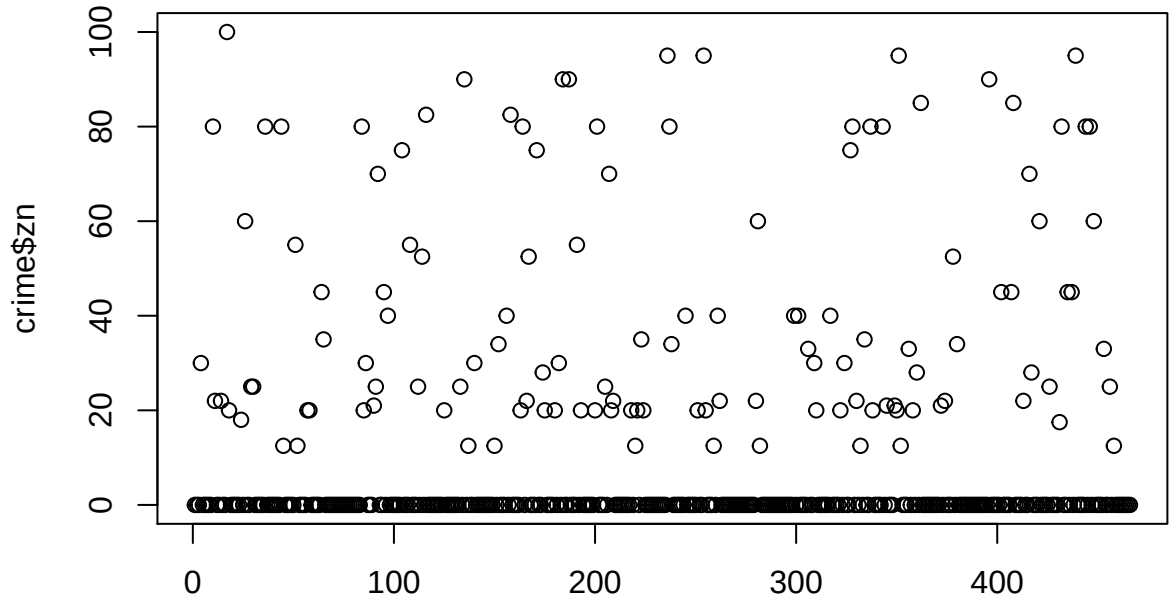

rad
5
10
15
20

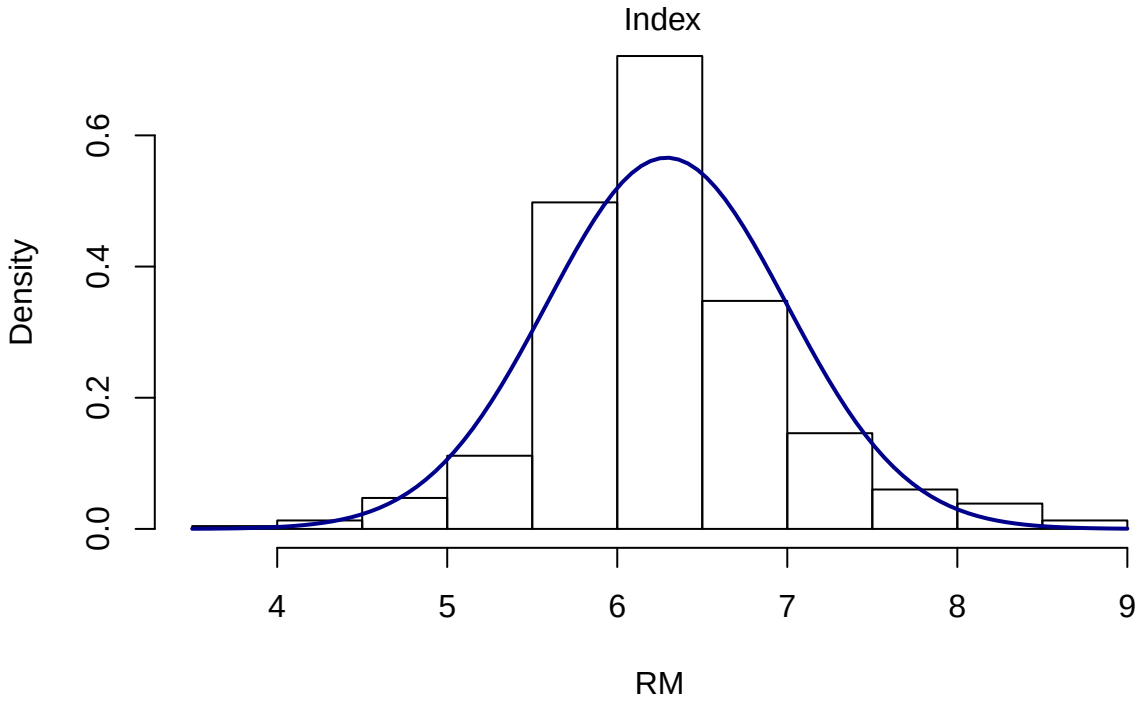


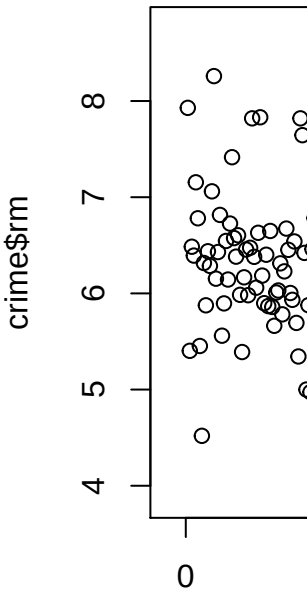
Based on an analysis of the box plots, the following variables have some outliers that may, or may not, exert influence on the regression results.
- zn, rm, dis, black, lstat, medv

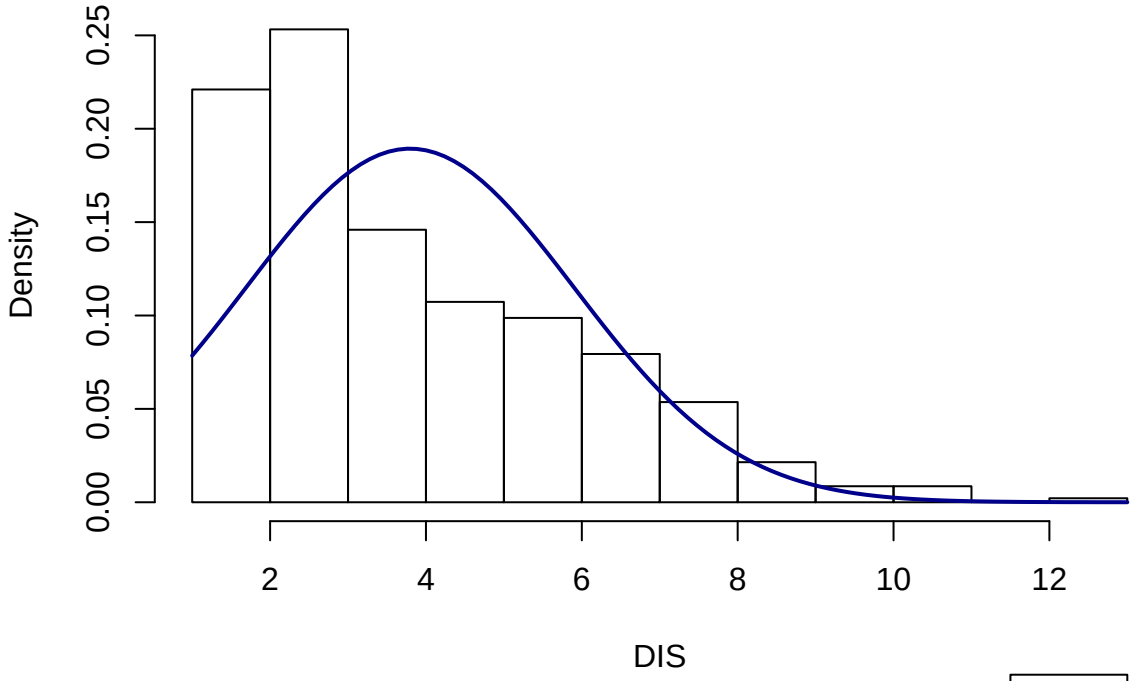
We'll next look at these variable mroe closely, starting with there histograms and frequency counts to better understand the nature of their distribution.

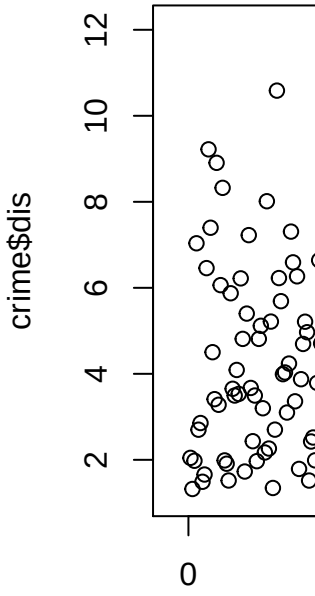


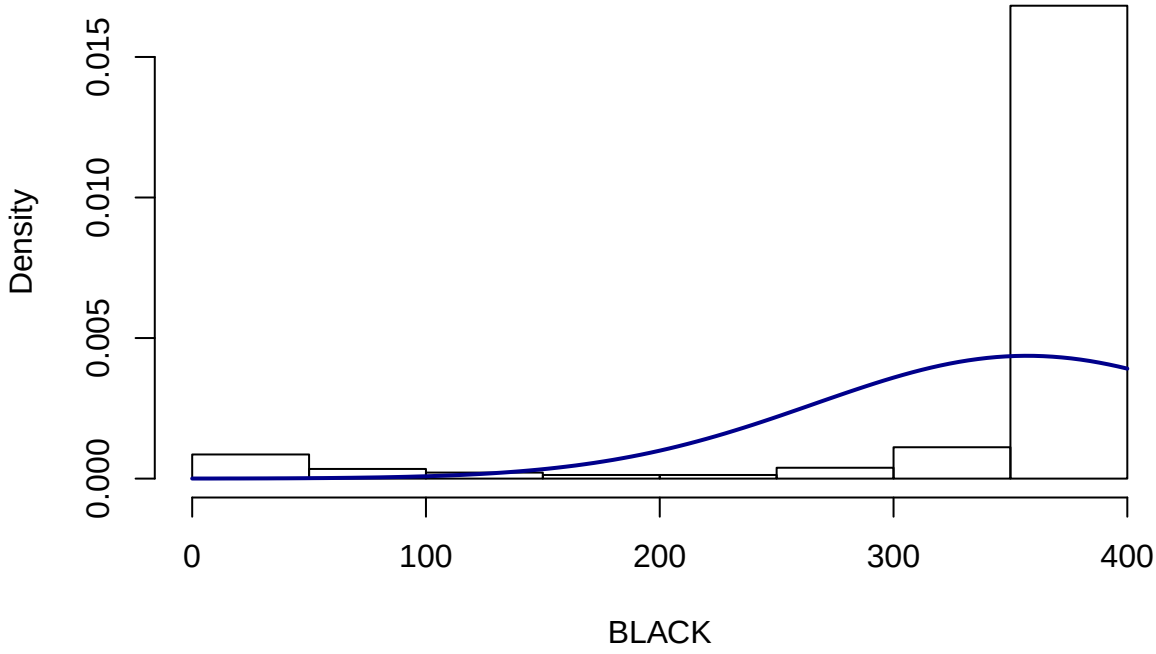

crime$zn
0
20
40
60
80
100
0
100
200
300
400
Index

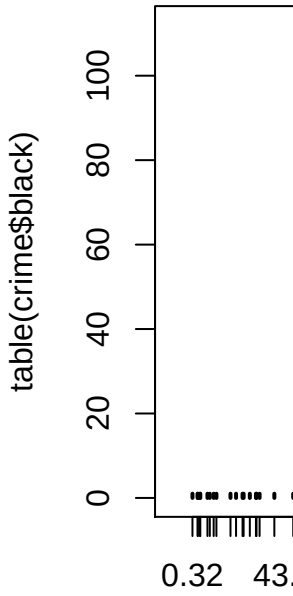

Density
0.0
0.2
0.4
0.6
4
5
6
7
8
9
RM

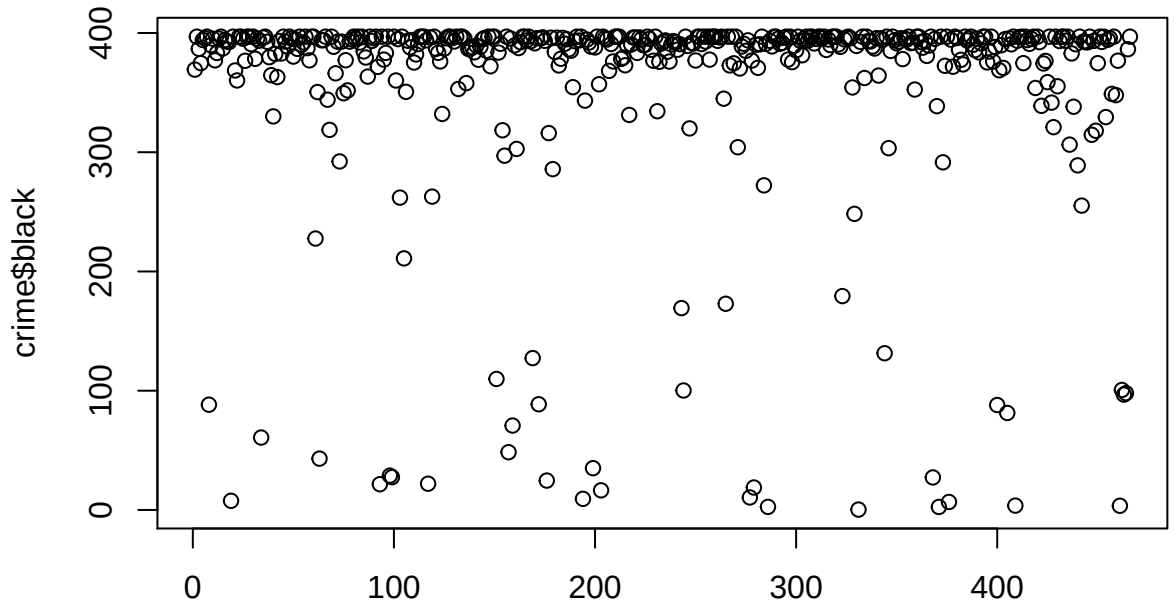

crime$rm
4
5
6
7
8
0

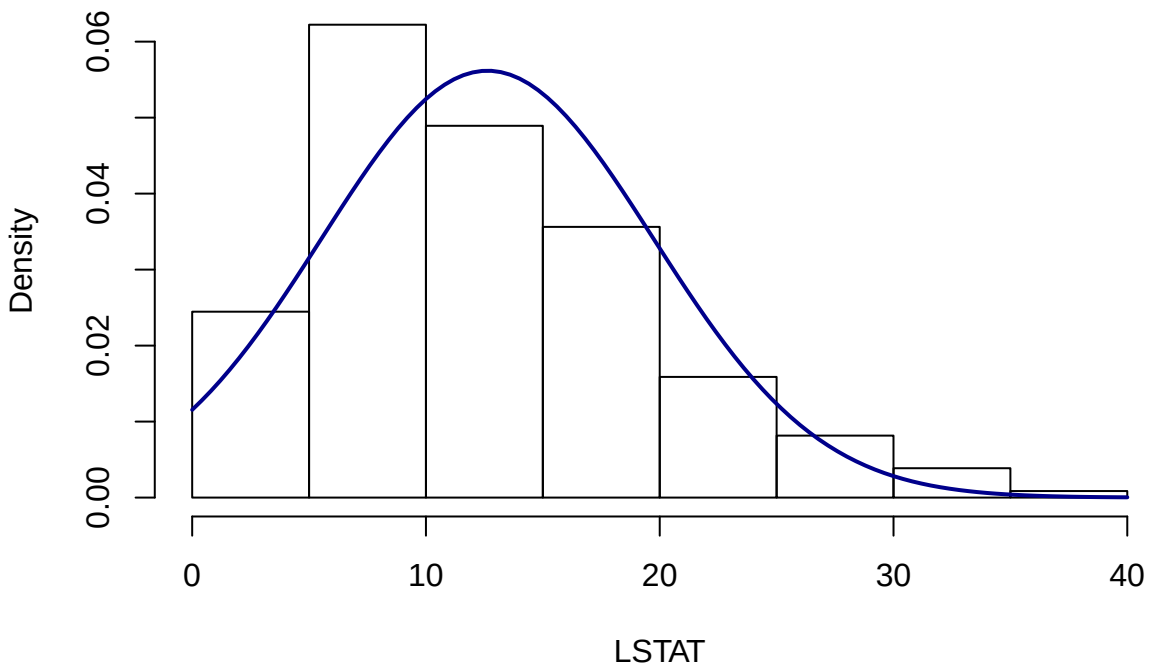

Density
0.00
0.05
0.10
0.15
0.20
0.25
2
4
6
8
10
12
DIS

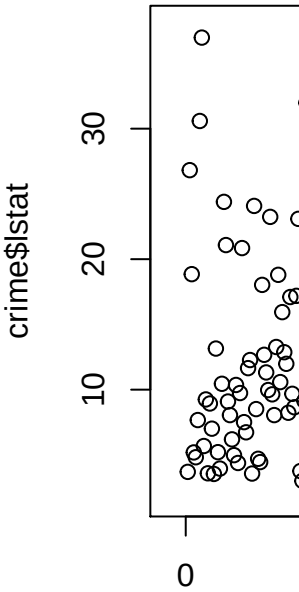

crime$dis
2
4
6
8
10
12
0

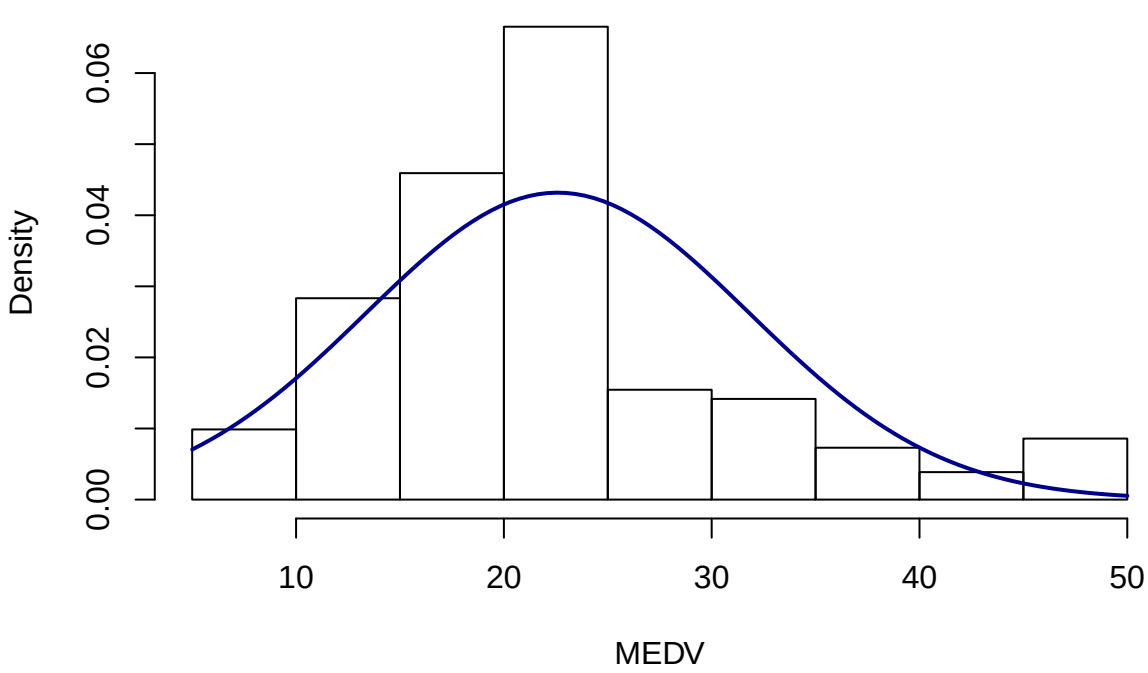

Density
0.000
0.005
0.010
0.015
0
100
200
300
400
BLACK

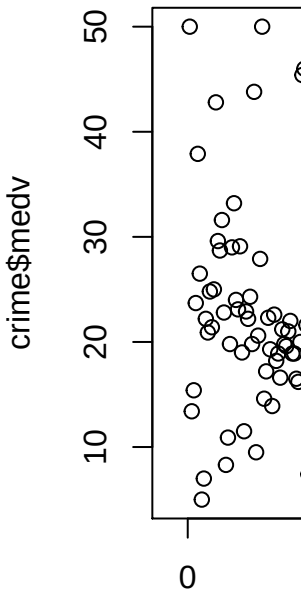

table(crime$black)
0
20
40
60
80
100
0.32
43


crime$black
0
100
200
300
400
0
100
200
300
400
Index


Density
0.00
0.02
0.04
0.06
0
10
20
30
40
LSTAT


crime$lstat
10
20
30
0






```
##
## 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.

Based on our detailed review of the variables that contained outliers, the following variables could be problematic:

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.

Data Preparation

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

	zn	indus	nox	rm	age	dis	rad	tax	ptra
zn	1.0000000	-0.5382664	-0.5170452	0.3198141	-0.5725805	0.6601243	-0.3154812	-0.3192841	-0.39103
indus	-0.5382664	1.0000000	0.7596301	-0.3927118	0.6395818	-0.7036189	0.6006284	0.7322292	0.39468
nox	-0.5170452	0.7596301	1.0000000	-0.2954897	0.7351278	-0.7688840	0.5958298	0.6538780	0.17626
rm	0.3198141	-0.3927118	-0.2954897	1.0000000	-0.2328125	0.1990158	-0.2084457	-0.2969343	-0.36034
age	-0.5725805	0.6395818	0.7351278	-0.2328125	1.0000000	-0.7508976	0.4603143	0.5121245	0.25544
dis	0.6601243	-0.7036189	-0.7688840	0.1990158	-0.7508976	1.0000000	-0.4949919	-0.5342546	-0.23333
rad	-0.3154812	0.6006284	0.5958298	-0.2084457	0.4603143	-0.4949919	1.0000000	0.9064632	0.47145
tax	-0.3192841	0.7322292	0.6538780	-0.2969343	0.5121245	-0.5342546	0.9064632	1.0000000	0.47442
ptratio	-0.3910357	0.3946898	0.1762687	-0.3603471	0.2554479	-0.2333394	0.4714516	0.4744223	1.00000
black	0.1794150	-0.3581356	-0.3801549	0.1326676	-0.2734677	0.2938441	-0.4463750	-0.4425059	-0.18163
lstat	-0.4329925	0.6071102	0.5962426	-0.6320245	0.6056200	-0.5075280	0.5031013	0.5641886	0.37735
medv	0.3767171	-0.4961743	-0.4301227	0.7053368	-0.3781560	0.2566948	-0.3976683	-0.4900329	-0.51591

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.

Build Models

We will attempt to create the simplest model possible by using only one variable - the one that provides us the highest overall AUC (performance) all by itself. We can plug in each variable separately and then select the highest result. The best variable is nox - the presence of nitrogen oxides (an industrial pollutant) on the property.

```
## [1] 0.9356955
```

By combining nos with all the remaining variables and selecting the highest resulting AUC result, we conclude that nox plus rad (access to radial highways) is the strongest combinaton of two variables.

```
## [1] 0.9574743
```

```
## [1] 0.962486
```

By combining three variables - nox, rad and zn - that is, the concentration of nitrogen oxides, access to radial highways and the proportion of land zoned for large lots, we can predict with 95.8% accuracy whether the crime rate at this property is above or below average. Since this is very close to the performance of the model using all variables (96%), we can be confident in using these three variables for our decision support process, and disregarding the others.

Select Models

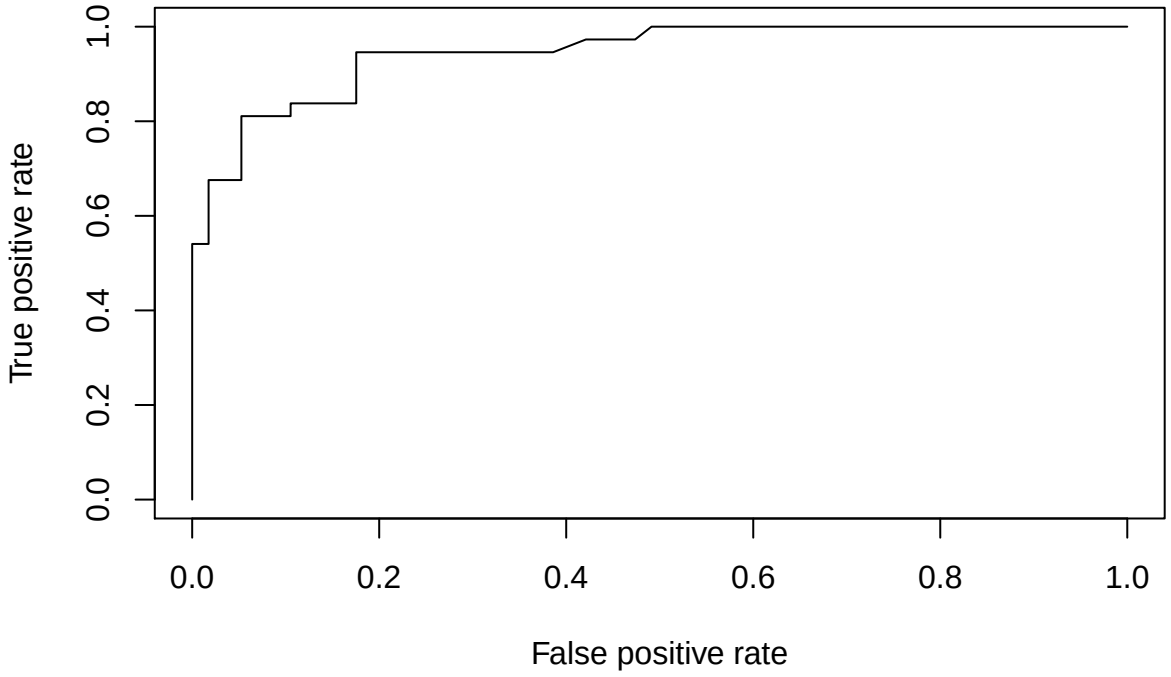
One way to test what variables to includes running univariate regression tests and analyse corresponding p-values and relative AIC values. Furthermore, we will investigate the AUC as well to see how accurate our univarate models are:

```
##        var        p_val      aic       auc
## 1       zn 2.287795e-10 413.2878 0.7076814
## 2    indus 3.980629e-26 345.8163 0.8091513
## 3    chas1 3.188437e-01 518.3011 0.5452821
## 4      nox 1.959853e-21 212.6269 0.8710289
## 5       rm 1.062379e-03 507.8644 0.5737316
## 6      age 7.459056e-25 317.3847 0.7937411
## 7      dis 1.086762e-22 307.0926 0.7970602
## 8      rad 1.468211e-06 330.3616 0.8440019
## 9      tax 1.304288e-19 353.7222 0.8319109
## 10 ptratio 1.079209e-06 493.3566 0.6600284
## 11   black 1.849434e-06 435.2948 0.7484590
## 12   lstat 1.003217e-16 416.8908 0.7015173
```

From the above table, we can see 1 variable that has no significance and under a univariate regression model, and have high relative AIC, and accuracy that is barely higher than a random variable. The Chas variable is a viable candidate to remove from our modelling.

```
## Start:  AIC=243.19
## target ~ nox + rad + zn
##
##        Df Deviance    AIC
## <none>      235.19 243.19
## - zn    1   239.51 245.51
## - rad   1   288.29 294.29
## - nox   1   344.89 350.89

##
## Call:
## glm(formula = target ~ nox + rad + zn, family = binomial(link = "logit"),
##     data = crime)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.87748  -0.32492  -0.02648   0.00566   2.73698
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -16.00865    1.99493  -8.025 1.02e-15 ***
## nox          24.60111    3.32256   7.404 1.32e-13 ***
## rad           0.53186    0.11272   4.719 2.37e-06 ***
## zn           -0.03617    0.01931  -1.873    0.061 .
## ---
## 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: 235.19  on 462  degrees of freedom
## AIC: 243.19
##
## Number of Fisher Scoring iterations: 8
```




```
## [1] 0.9447606
```

All Done!