Data 621: Assignment 3

Binary Logistic Regression

Ritesh Lohiya

June 30, 2018

Overview

In this homework assignment, you will explore, analyze and model a data set containing information on crime for various neighborhoods of a major city. Each record has a response variable indicating whether or not the crime rate is above the median crime rate (1) or not (0).

Your objective is to build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. You will provide classifications and probabilities for the evaluation data set using your binary logistic regression model. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

Data Exploration:

Our training data comprises 466 observations and 13 variables. Below is a brief description of the variables in our data set:

Variable Name	Description
zn	proportion of residential land zoned for large lots (over 25000 square feet)
indus	proportion of non-retail business acres per suburb
chas	a dummy var. for whether the suburb borders the Charles River (1) or not (0)

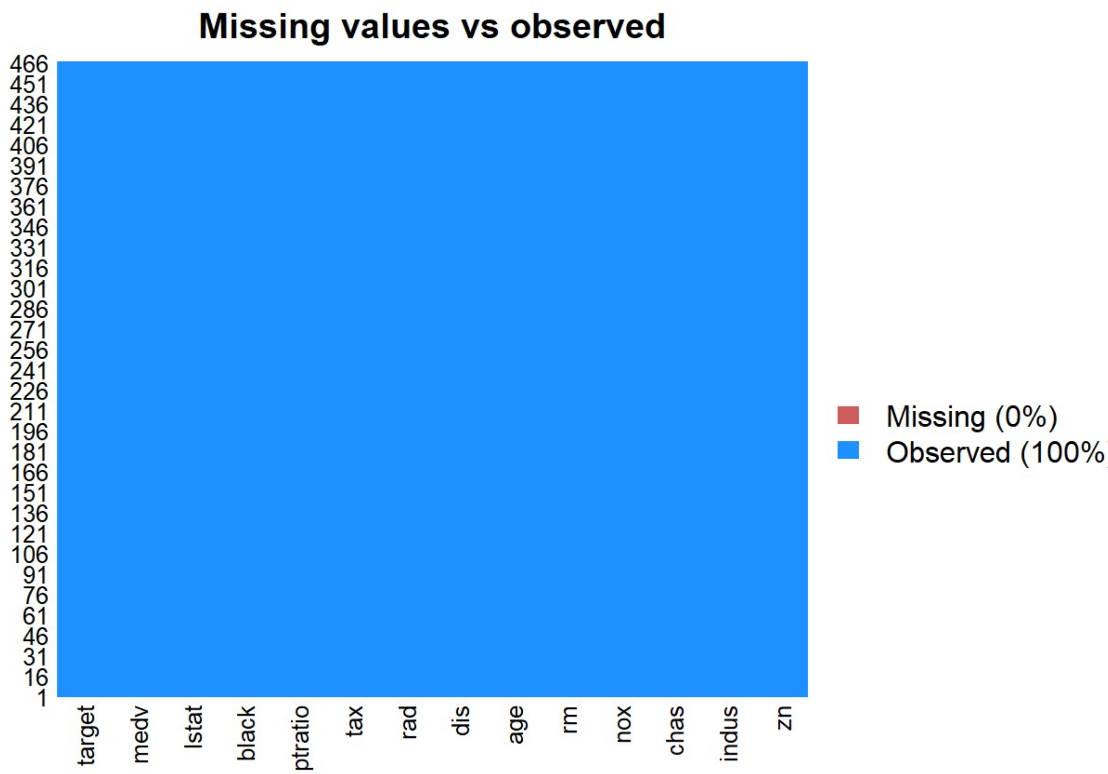
Variable Name	Description
nox	nitrogen oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted mean of distances to five Boston employment centers
rad	index of accessibility to radial highways
tax	full-value property-tax rate per $10,000
ptratio	pupil-teacher ratio by town
black	1000 (Bk−0.63)2(Bk−0.63)2 where Bk is the proportion of blacks by town
lstat	lower status of the population (percent)
medv	median value of owner-occupied homes in $1000s
target	whether the crime rate is above the median crime rate (1) or not (0) (response v

Target is our binary response variable. For this data exploration, we will be focusing on a binary logistic regression.

Visual Exploration:

Missing Values

We will see the missing values in the dataset. For this i have used Amelia package. According to the graph, the data shows no missing variables.

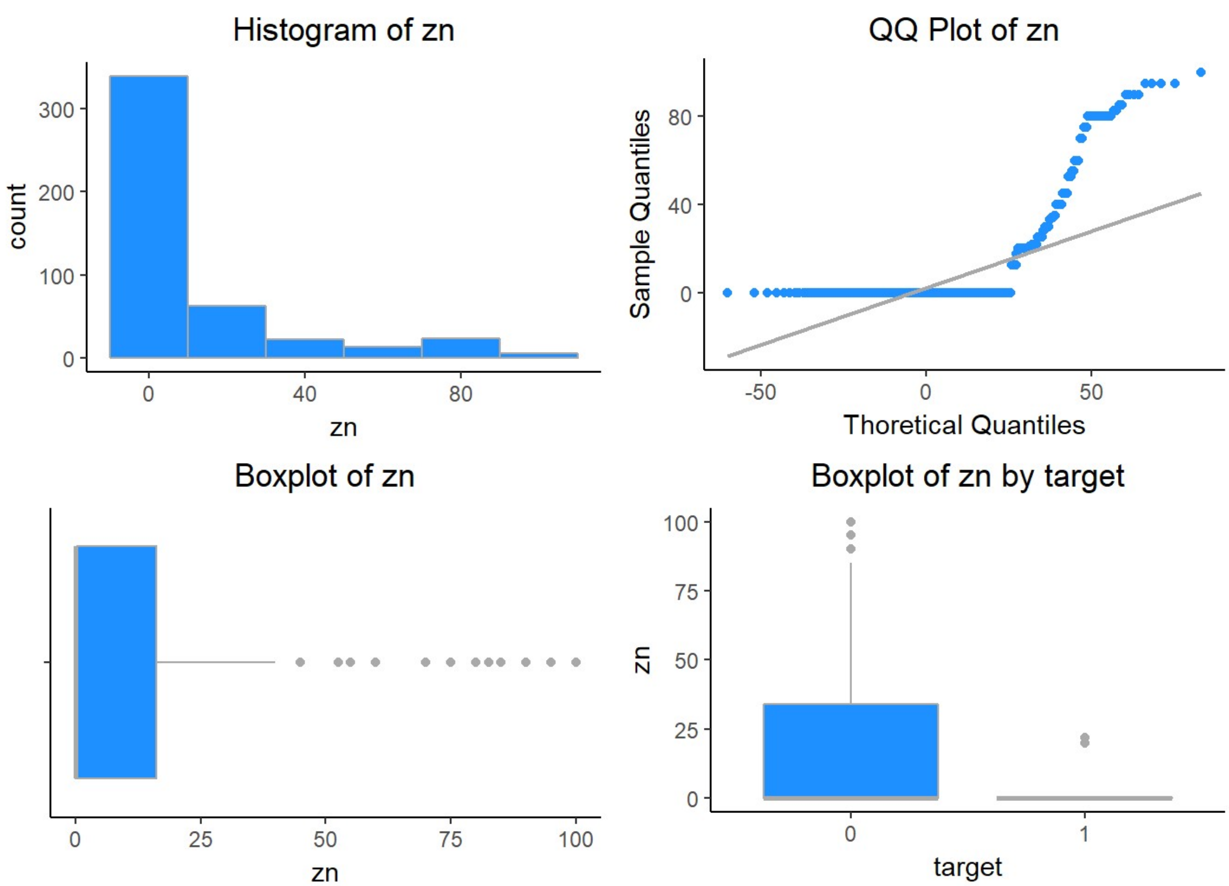


Lets now dig into available variables.

1. zn - proportion of residential land zoned for large lots (over 25000 square feet). We can see there are more zeros values for zn and also has positive skewness. Also there appears to be relationship between crime rates and zn.

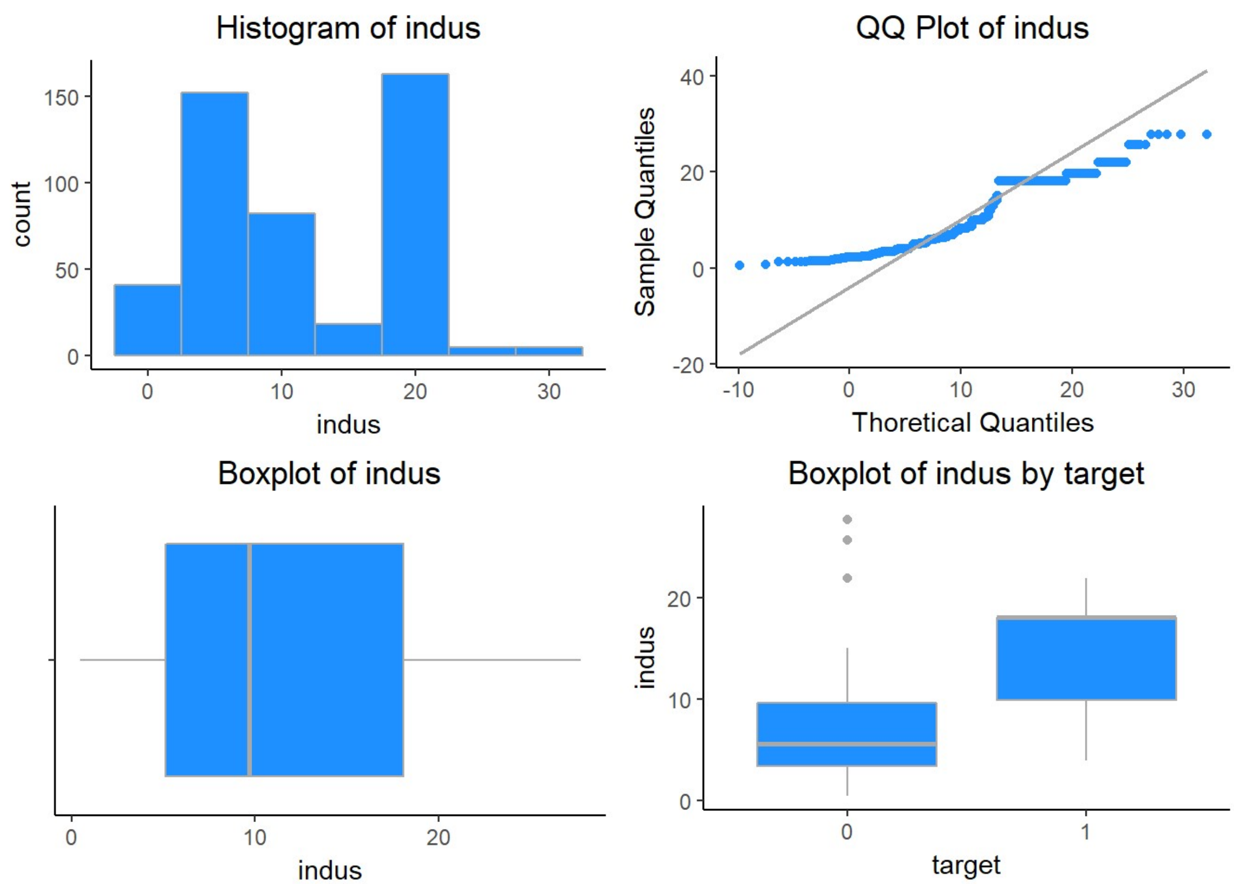
```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
##   0.000000   0.000000   0.000000  11.577253  16.250000 100.000000
##         SD       Skew       Kurt
```

```
## 23.364651   2.183841   6.842914
```



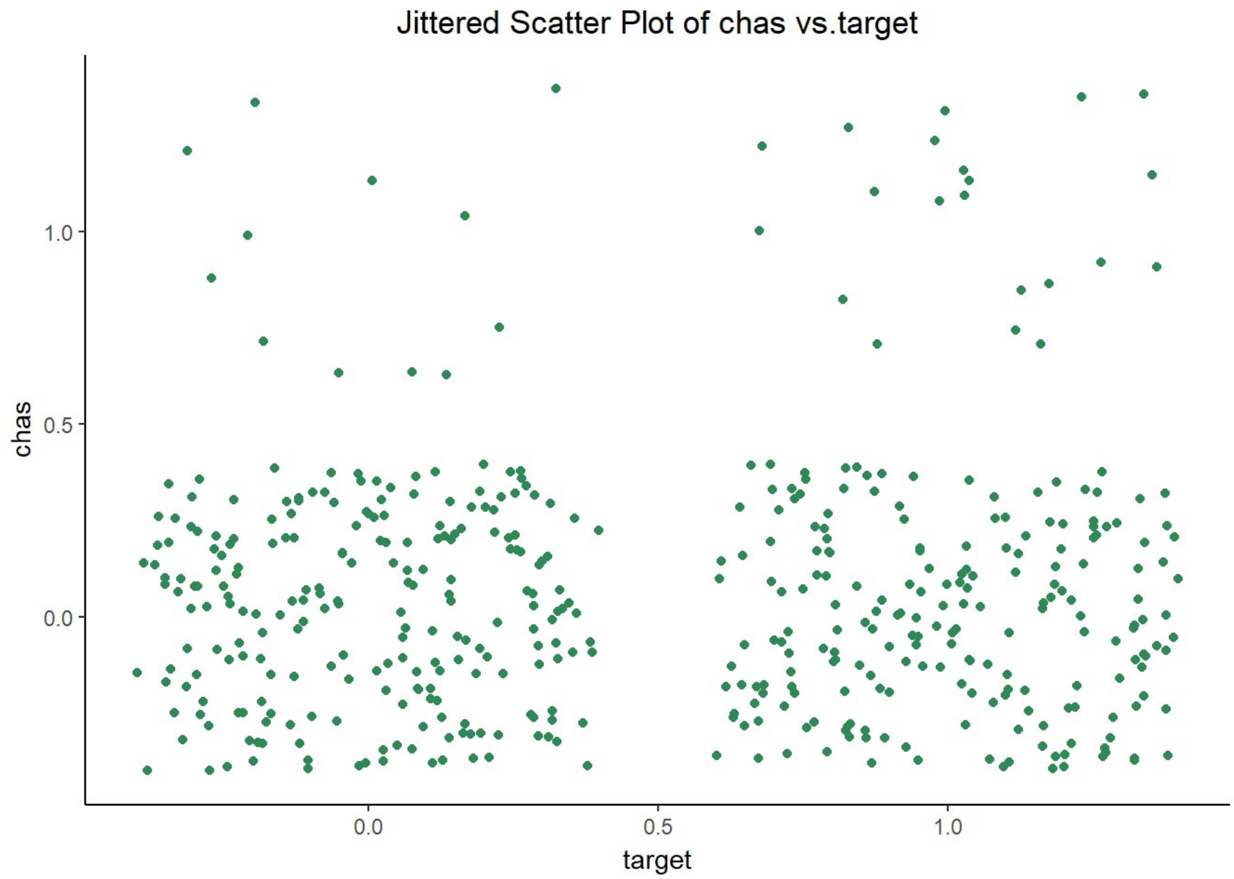
2. indus - proportion of non-retail business acres per suburb. The histogram below indicates a bi-modal quality to the variable's distribution, with many values clustering in two ranges.

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##   0.4600000   5.1450000   9.6900000  11.1050215  18.1000000  27.7400000
##          SD        Skew        Kurt
##   6.8458549   0.2894763   1.7643510
```



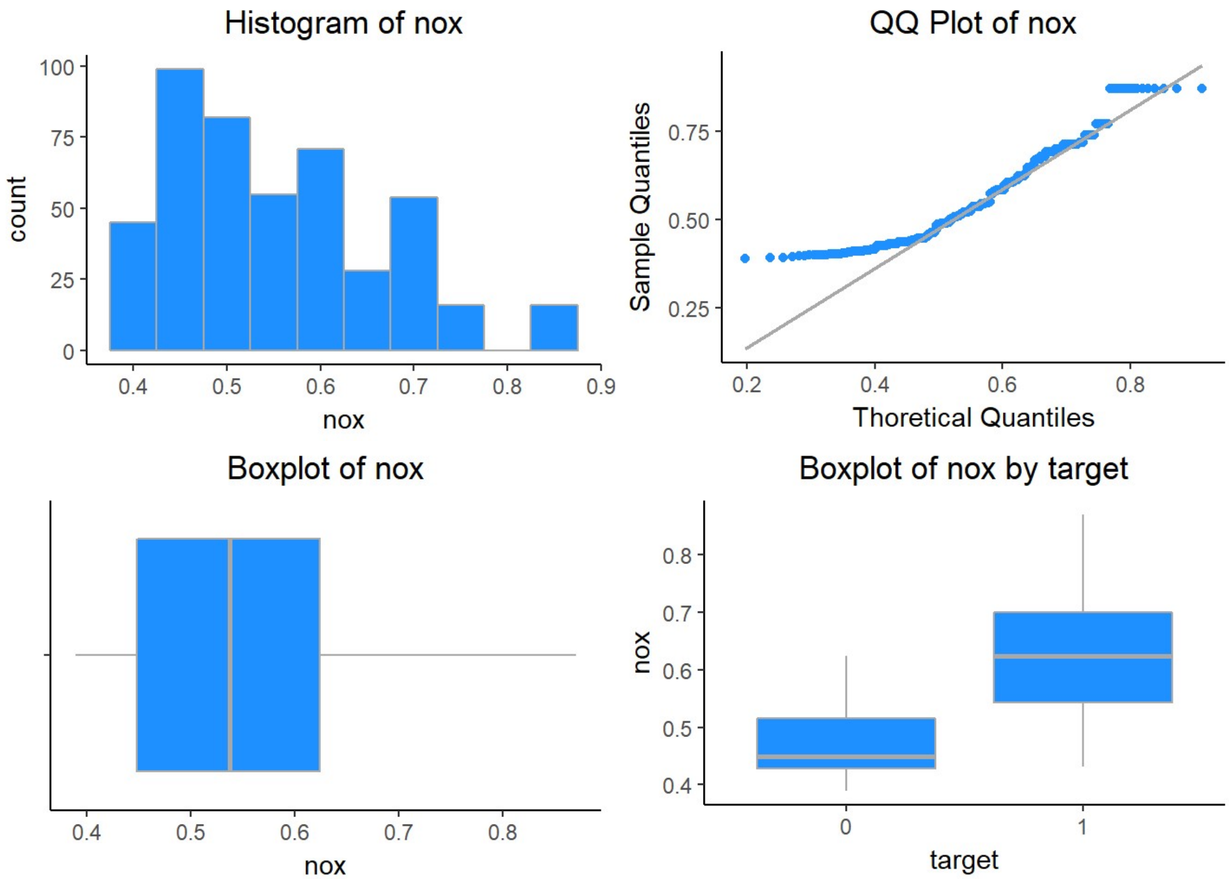
3. chas - a dummy var. for whether the suburb borders the Charles River (1) or not (0). This variable tells us if the neighborhood borders the Charles River (1) or not (0). Close to 7% of the neighborhood borders the Charles River. Of the areas bordering the Charles river 21 are in high crime areas.

```
##
##    0    1  Sum
##  433   33  466
```



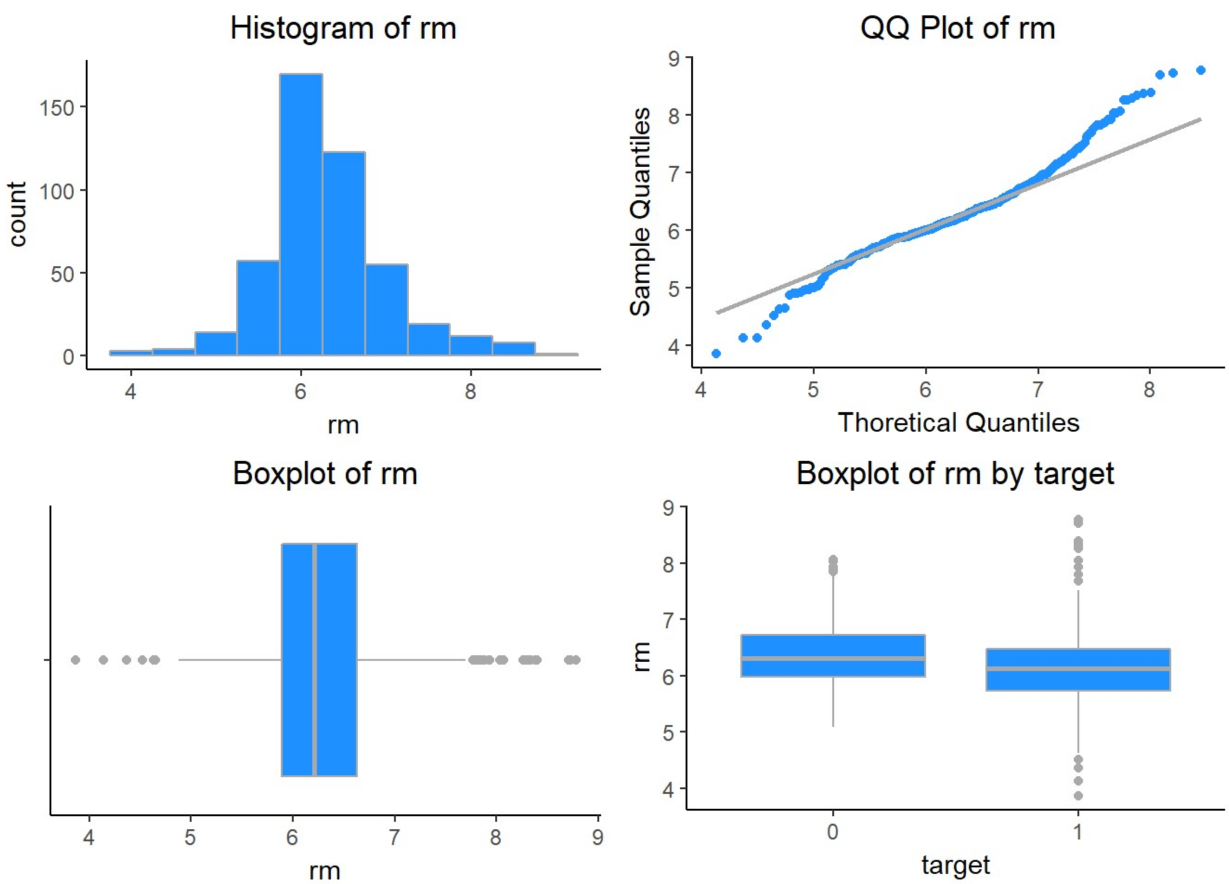
4. nox - nitrogen oxides concentration (parts per 10 million). The variable nox represents the concentration of nitrogen oxide in each Boston area. There is also positive skewness. We also see moderately higher nox variance in high crime areas.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.        SD
## 0.3890000 0.4480000 0.5380000 0.5543105 0.6240000 0.8710000 0.1166667
##      Skew      Kurt
## 0.7487369 2.9769895
```



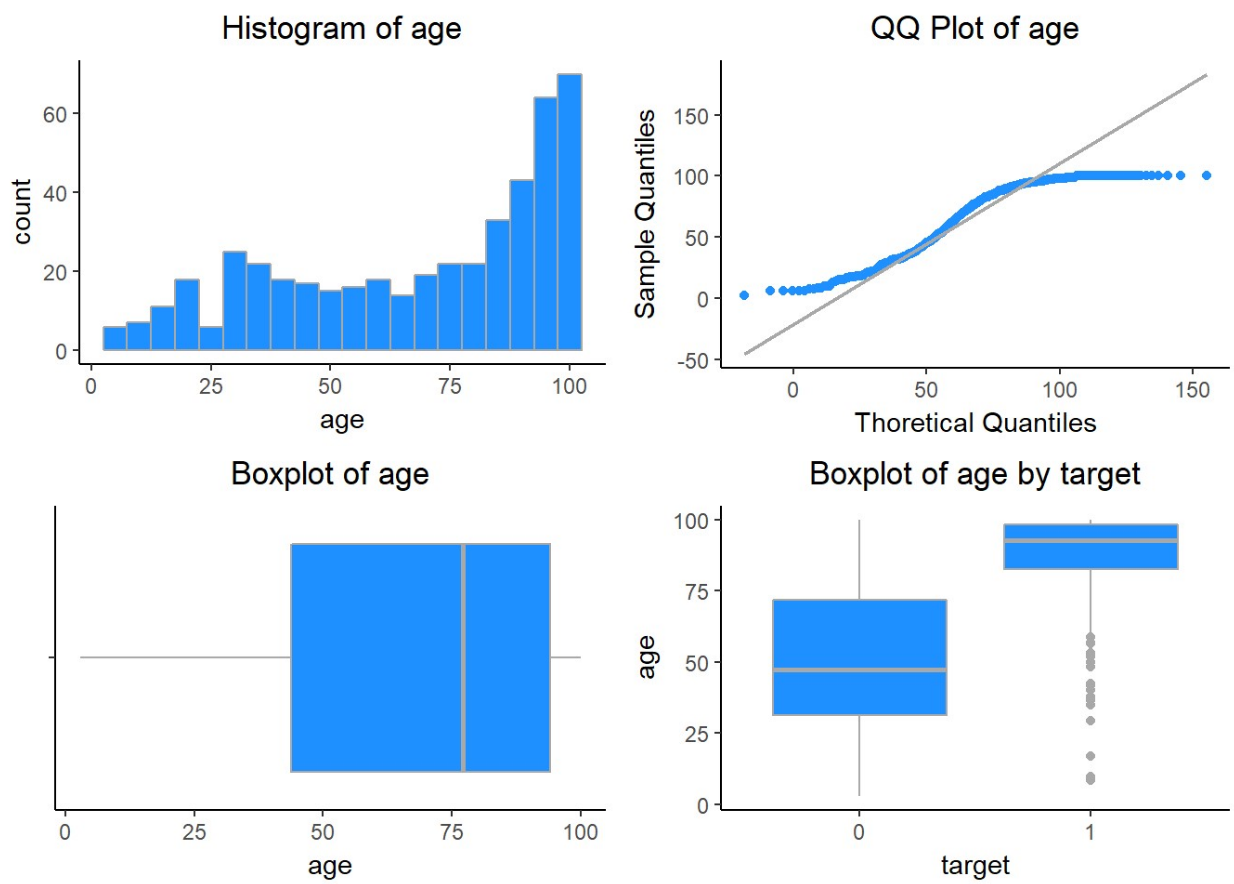
5. rm - average number of rooms per dwelling. The predictor rm is count measure describing the average number of rooms per dwelling. The distridution has heavy tail and has bell curve.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.        SD
## 3.8630000 5.8872500 6.2100000 6.2906738 6.6297500 8.7800000 0.7048513
##      Skew      Kurt
## 0.4808673 4.5619962
```



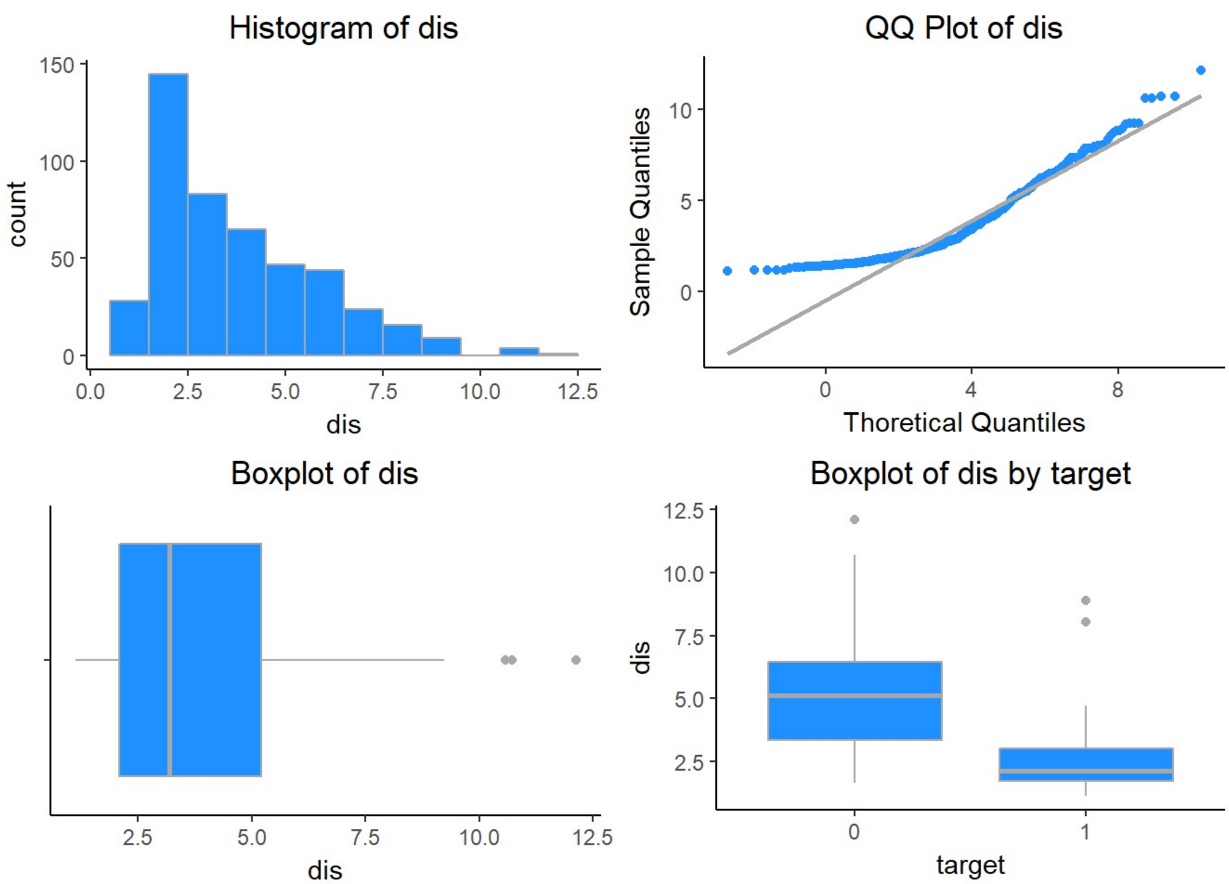
6. age - proportion of owner-occupied units built prior to 1940. The variable age indicates the proportion of owner occupied units built prior to 1940. This variable has high left skewness. Also there is significantly higher mean percentage of older homes in high crime areas.

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##    2.9000000   43.8750000   77.1500000   68.3675966   94.1000000  100.0000000
##           SD         Skew         Kurt
##   28.3213784   -0.5795721    1.9986874
```



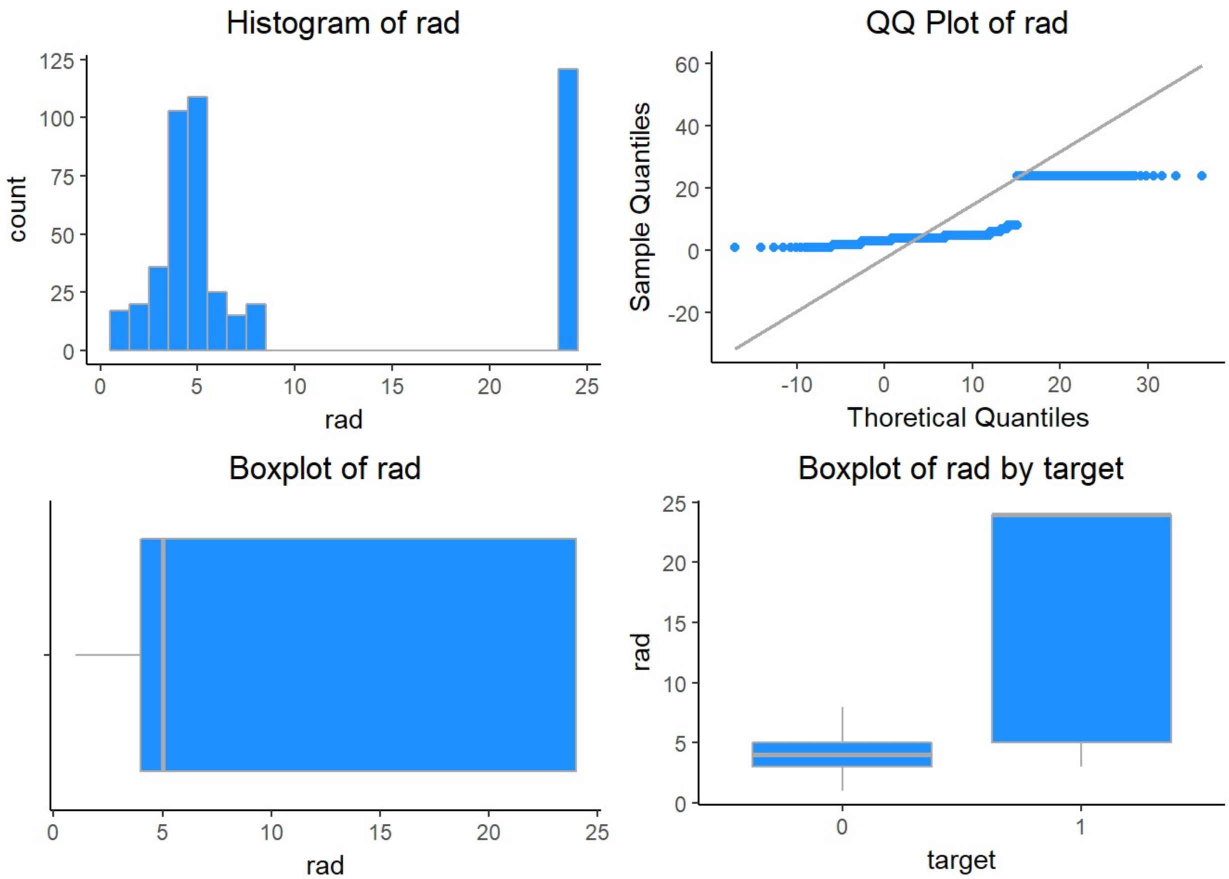
7. dis - weighted mean of distances to five Boston employment centers. The predictor dist describes the average distance to Boston employment centers. The variable is moderately right skewed. Also we can see that low crime areas are associated with higher average distances to employment centers.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.        SD
##  1.129600  2.101425  3.190950  3.795693  5.214600 12.126500  2.106950
##      Skew      Kurt
##  1.002117  3.486917
```



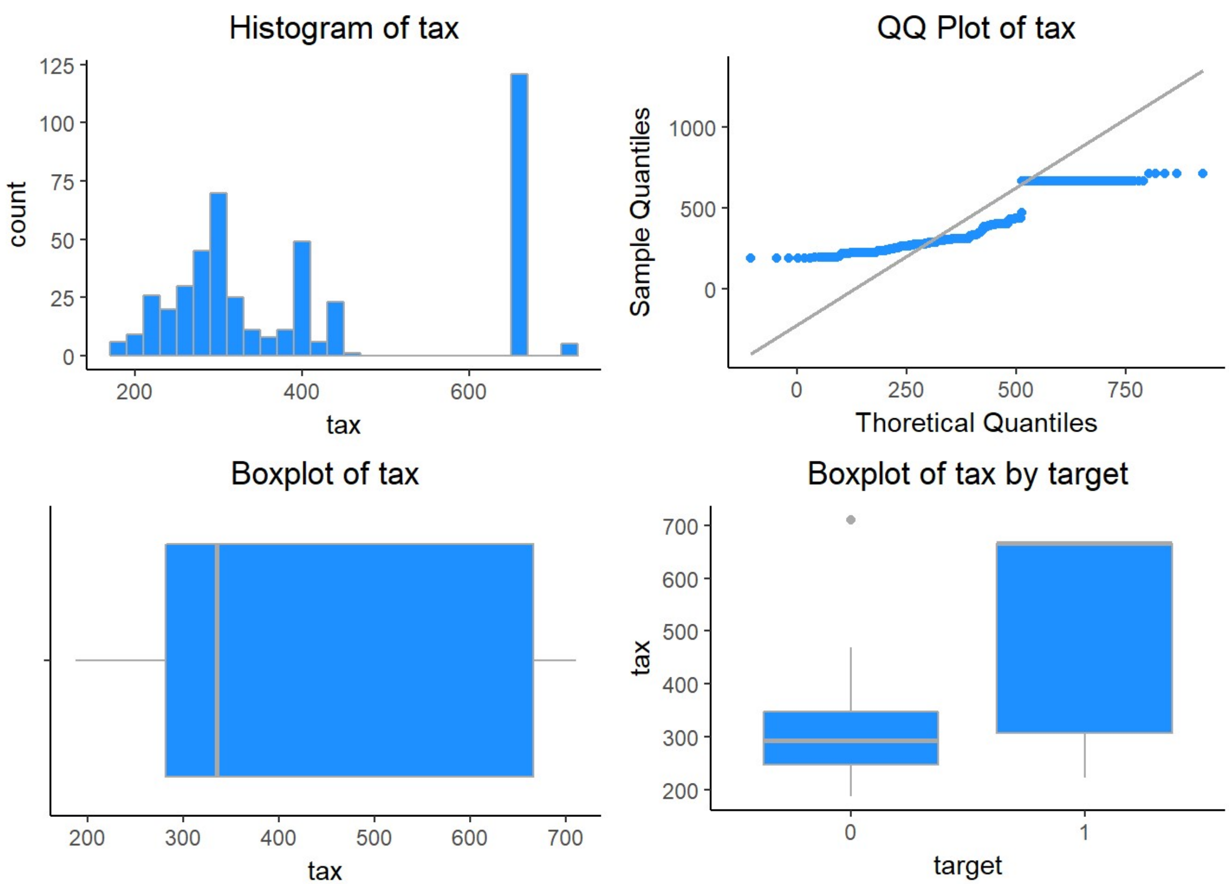
8. rad - index of accessibility to radial highways. The rad variable is an integer-valued index measure indicating an area's accessibility to radial highways. In the boxplots below, there appears to be a significant positive association between high crime rates and rad value.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.        SD
##  1.000000  4.000000  5.000000  9.530043 24.000000 24.000000  8.685927
##      Skew      Kurt
##  1.013539  2.147295
```



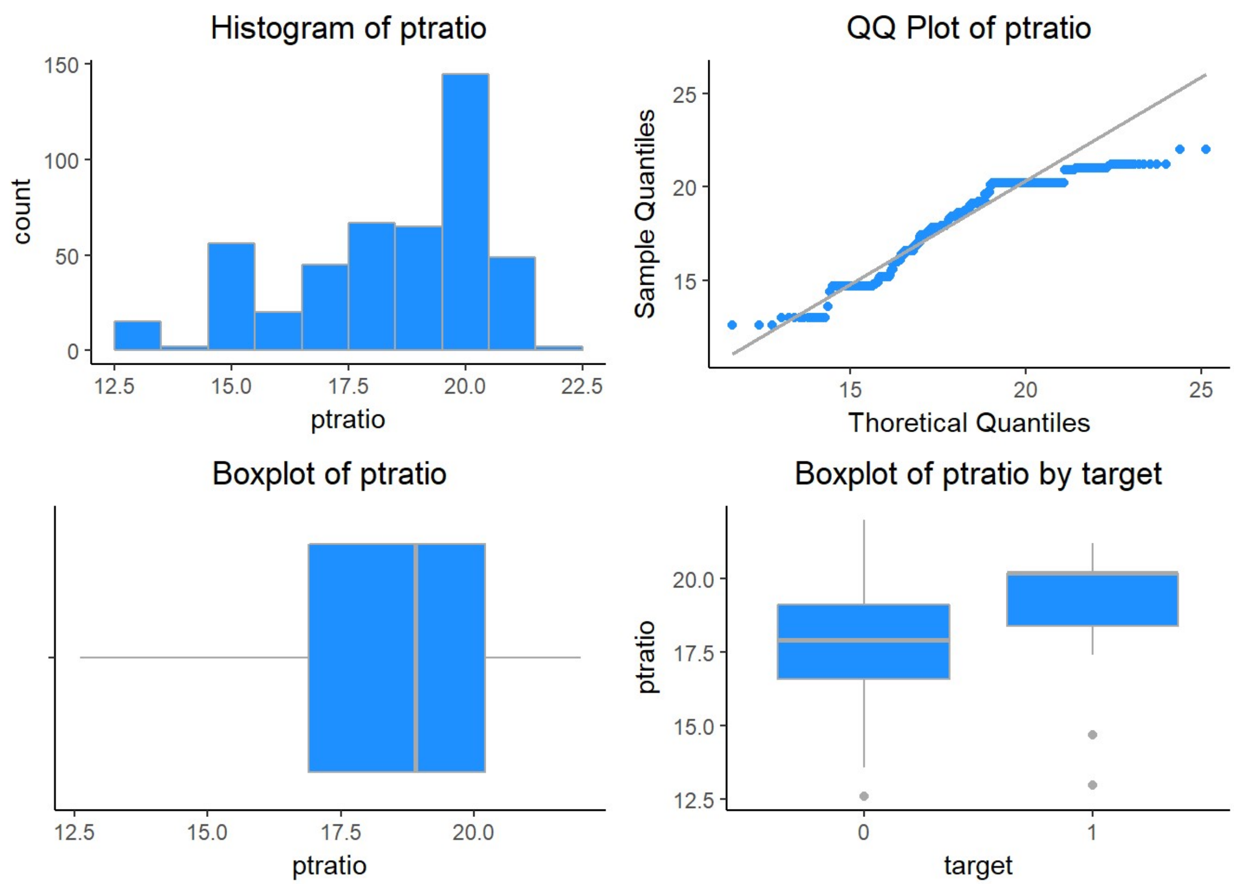
9. tax - full-value property-tax rate per $10,000. The tax variable refers to the the tax rate per $10k of property value. High crime areas also appear to have a strong, positive association with the tax value. This variable is densely distributed around two of the following approximate values: 300 and 700.

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## 187.0000000 281.0000000 334.5000000 409.5021459 666.0000000 711.0000000
##          SD        Skew        Kurt
## 167.9000887   0.6614416   1.8599284
```



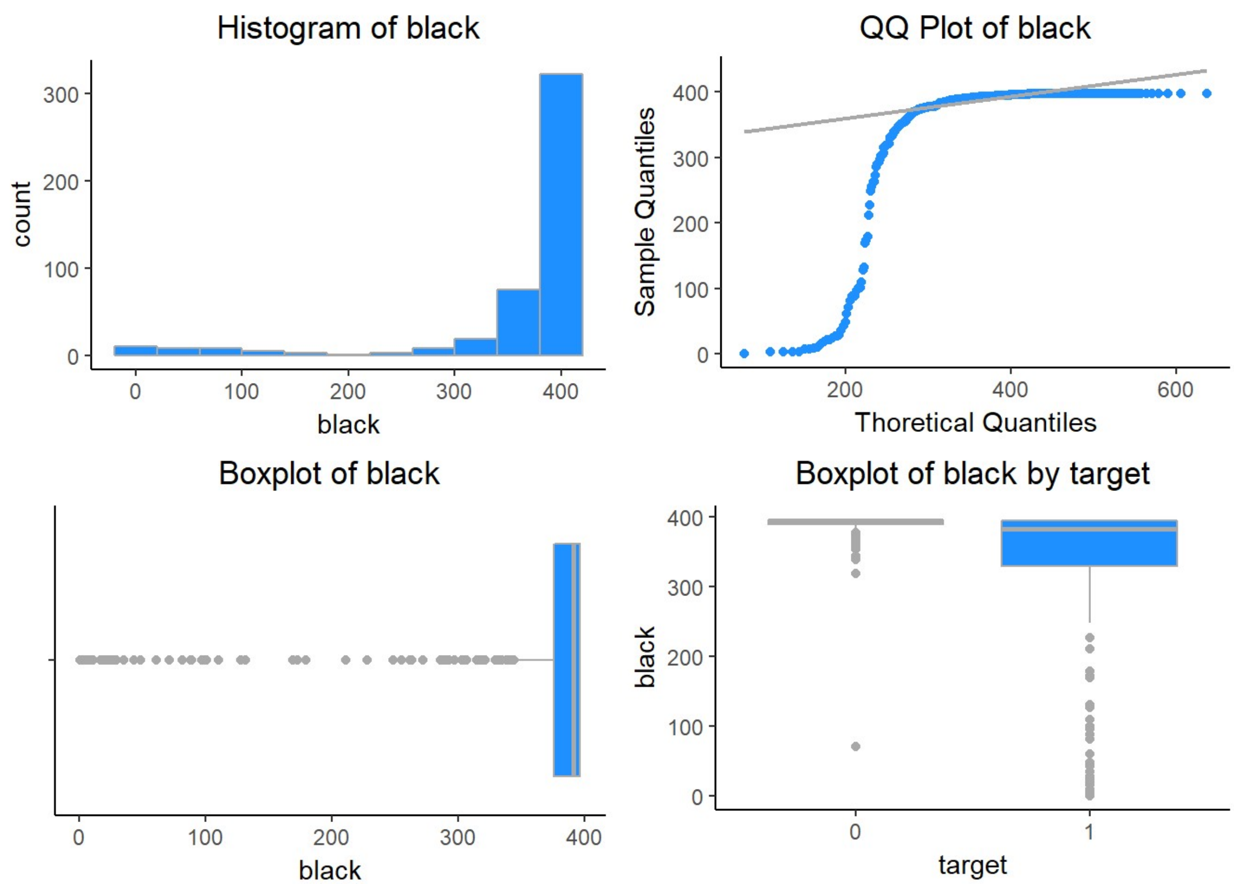
10. ptratio - ptratio: pupil-teacher ratio by town. The predictor ptratio indicates the average school, pupil-to-student ratio, and has a left skewed distribution. We can see a positive relationship between ptratio and high crime.

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## 12.6000000 16.9000000 18.9000000 18.3984979 20.2000000 22.0000000
##         SD       Skew       Kurt
##  2.1968447 -0.7567025  2.6108306
```



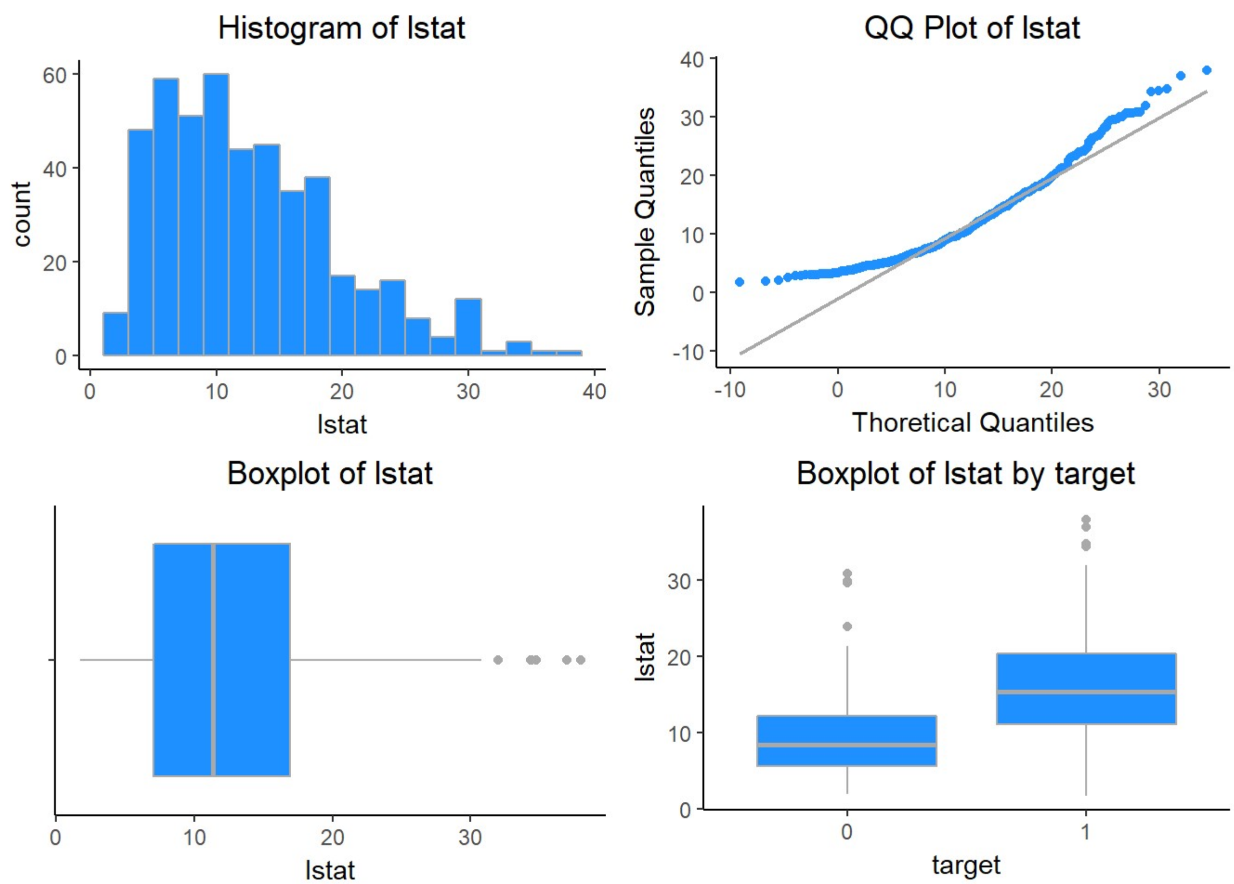
11. black - 1000 (Bk−0.63)2(Bk−0.63)2 where Bk is the proportion of blacks by town. This variable is heavily left skewed.

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
##   0.320000 375.607500 391.340000 357.120150 396.237500 396.900000
##         SD       Skew       Kurt
##  91.321130  -2.925723  10.386460
```



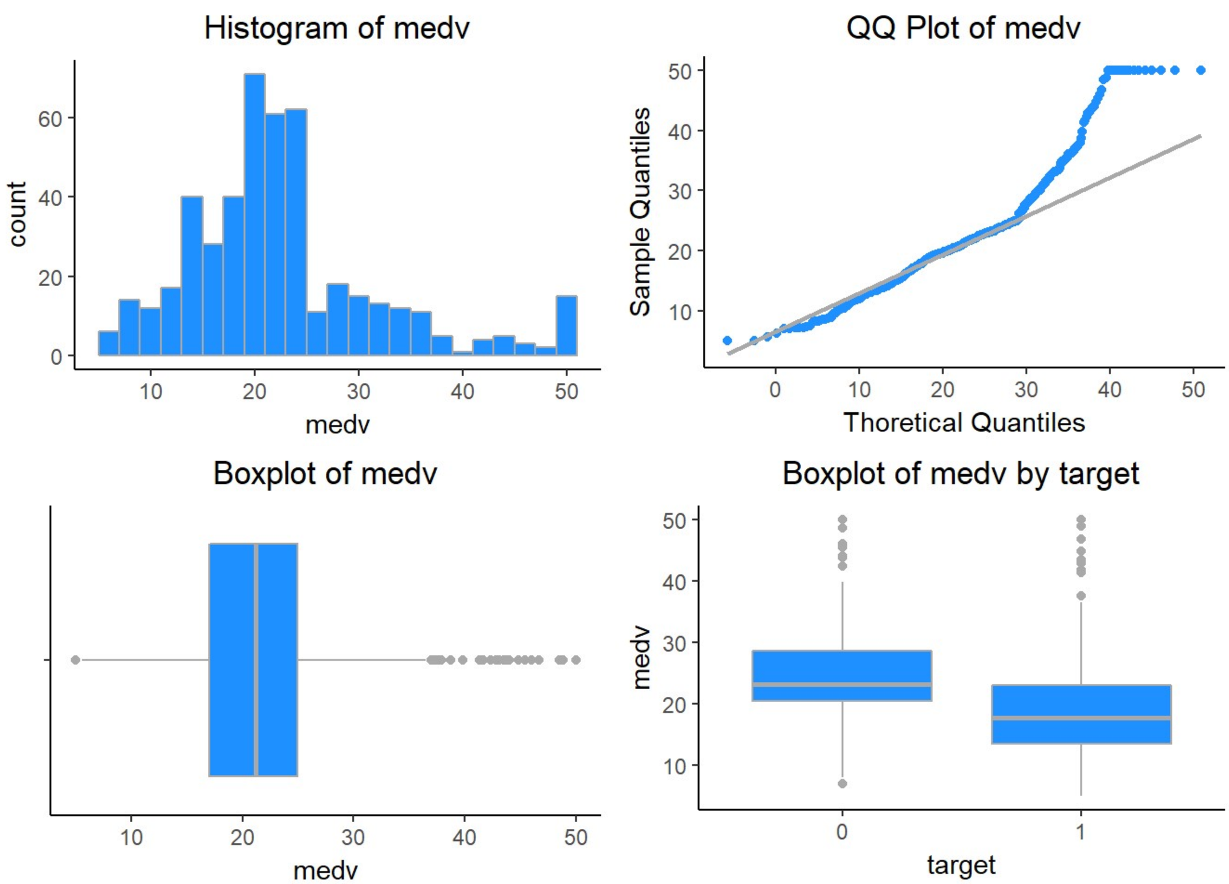
12. lstat - lower status of the population (percent). The variable lstat indicates the proportion of the population deemed to be of lower status. lstat is right skewed. High crime areas tend to have be associated with larger lstat values.

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
##  1.7300000  7.0425000 11.3500000 12.6314592 16.9300000 37.9700000
##         SD       Skew       Kurt
##  7.1018907  0.9085092  3.5184532
```



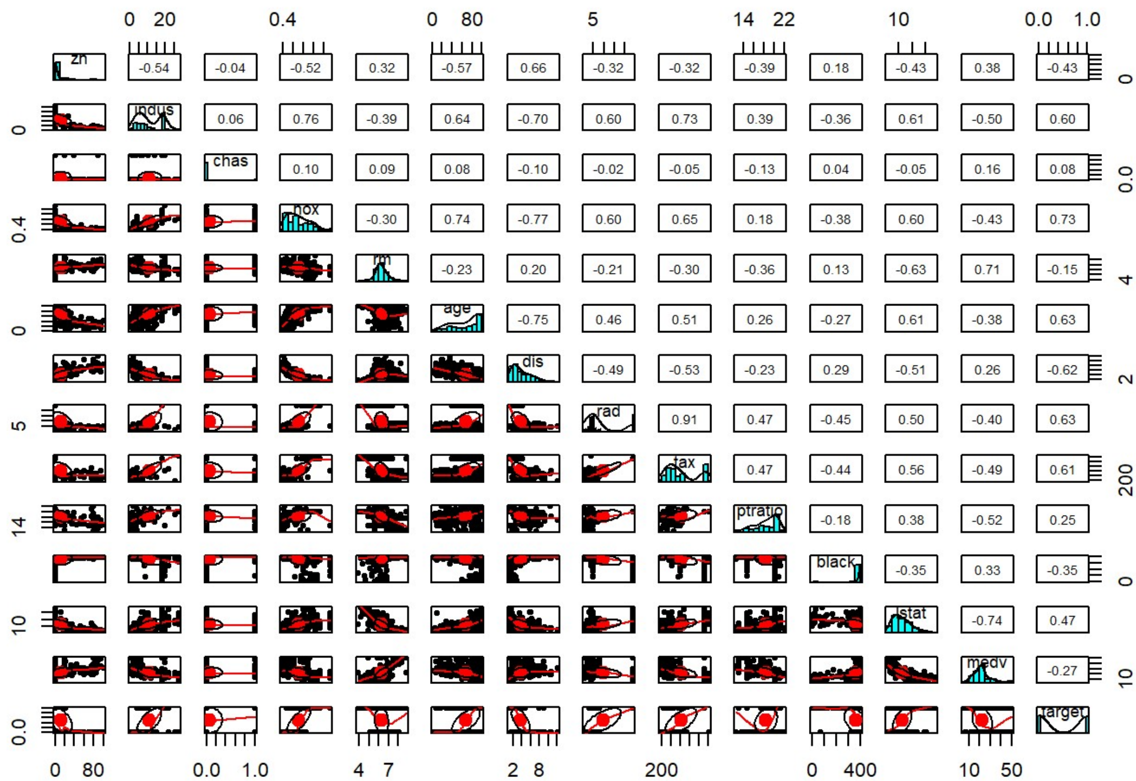
13. medv - median value of owner-occupied homes in $1000s. The median value of residential homes in a given area. The variable is slightly right skewed, and high values of medv appear to be associated with lower crime rates.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.        SD
##  5.000000 17.025000 21.200000 22.589270 25.000000 50.000000  9.239681
##      Skew      Kurt
##  1.080167  4.392615
```



Correlation

The correlation plot below shows how variables in the dataset are related to each other. Looking at the plot, we can see that certain variables are more related than others.



Data Preparation

Data preparation or the preprocessing is the most important part in model development. We need to remove the noise in the data so as to build a good model. We may use the transformation such as log, power transformation etc

a. Missing Values - there are no missing values, so we will not do any missing value treatment.

b. outliers: I think we don't have any outliers that we should be removing at this stage.

c. Transformation –

age and lstat are both skewed, so let's see boxcox transformation suggestions.

```
## Fitted parameters:
##       lambda         beta      sigmasq
##     1.317655   205.697942 10492.780979
##
## Convergence code returned by optim: 0
```

```
## Fitted parameters:
##    lambda      beta   sigmasq
## 0.2328042 3.2351269 1.0549878
##
## Convergence code returned by optim: 0
```

So for age the boxcox fit suggested power transformation of 1.3 and for lstat boxcox fit suggested power transformation of 0.23. Let's apply the same.

```
crime_train$age_mod <- crime_train$age^1.3
crime_train$lstat_mod <- crime_train$lstat^0.23
```

The predictor dis, rm and medv has a moderate positive skew. Let's transform using the box-cox transformation

```
## Fitted parameters:
##     lambda       beta    sigmasq
## -0.1467279  1.0719066  0.2051234
##
## Convergence code returned by optim: 0
```

```
## Fitted parameters:
##     lambda       beta    sigmasq
## 0.20380031 2.22393623 0.02626356
##
## Convergence code returned by optim: 0
```

```
## Fitted parameters:
##    lambda      beta   sigmasq
## 0.2348612 4.4693904 0.6926584
##
## Convergence code returned by optim: 0
```

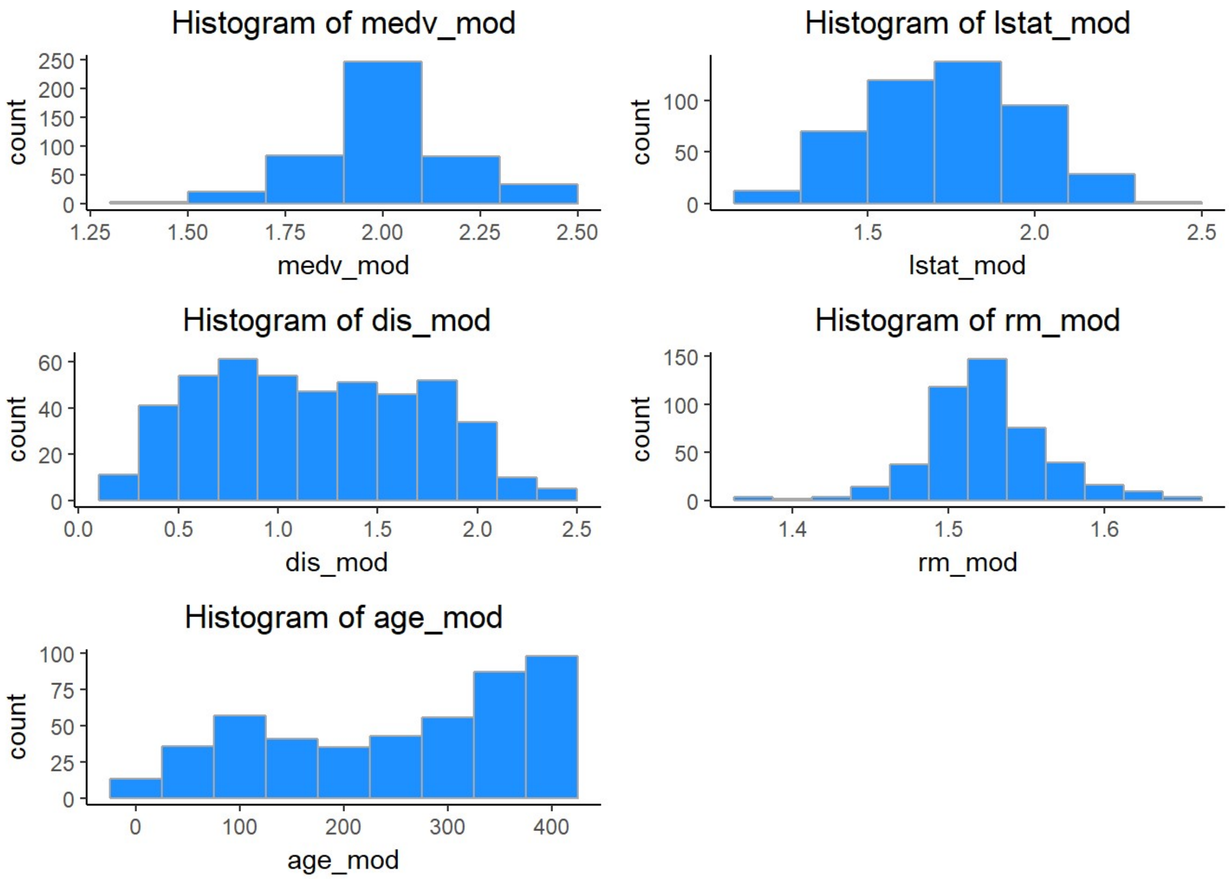
For medv and rm the boxcox fit suggested power transformation of .23. Let's apply the same.

```
crime_train$medv_mod <- crime_train$medv^0.23
crime_train$rm_mod <- crime_train$rm^0.23
```

The lamda for the boxcoxfit for is dis is alose to 0, so we can apply log transformation.

```
crime_train$dis_mod <- log(crime_train$dis)
```

Let's plot to see the status of the variables after transformation:



We can see that the skewness of the transformed variables improved.

Build Models

We will build 4 models and see which one is a good fit model.

Model 1 - All original variables model

In this model we will use all the variables. This can be our base model and this model will not include any transformations. We can see which variables are significant. This will help us in looking at the P-Values and removing the non-significant variables.

```r
model1 <- glm(target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
ptratio + black + lstat + medv , family="binomial", data=crime_train)

summary(model1)
```

```
##
## Call:
## glm(formula = target ~ zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat + medv, family = "binomial",
##     data = crime_train)
##
## 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
```

Model 2: - All significant original variables model.

I came up with this models after analyzing the output of model1. I removed all the variables that are not significant after seeing their P-Value.

```
model2 <- glm(target ~ nox + age + dis + rad + tax + ptratio + black +  medv
, family="binomial", data=crime_train)
summary(model2)
```

```
##
## Call:
## glm(formula = target ~ nox + age + dis + rad + tax + ptratio +
##     black + medv, family = "binomial", data = crime_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.42422  -0.19292  -0.01400   0.00279   3.06740
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -32.301655   6.382694  -5.061 4.17e-07 ***
## nox          42.160350   6.674149   6.317 2.67e-10 ***
## age           0.031017   0.010681   2.904 0.003684 **
## dis           0.437803   0.172533   2.538 0.011165 *
```

```
## rad           0.703446   0.140296   5.014 5.33e-07 ***
## tax          -0.008744   0.002611  -3.348 0.000813 ***
## ptratio       0.395580   0.112482   3.517 0.000437 ***
## black        -0.012490   0.006760  -1.848 0.064662 .
## medv          0.101177   0.034116   2.966 0.003020 **
## ---
## 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: 198.28  on 457  degrees of freedom
## AIC: 216.28
##
## Number of Fisher Scoring iterations: 9
```

Model 3: - All variables with transformations(will keep variables that were not transformed)

Model 3 includes original variables, plus the transformed variables from the transformations like power transformation and log transformations. This transformation should help in reducing the skewness in the data or help them to become more normalized. This will help us in looking at the P-Values and removing the non-significant variables.

```
model3 <- glm(target ~ zn + indus + chas + nox + rm_mod + age_mod + dis_mod +
rad + tax + ptratio + black + lstat_mod + medv_mod , family="binomial", data=
crime_train)

summary(model3)
```

```
##
## Call:
## glm(formula = target ~ zn + indus + chas + nox + rm_mod + age_mod +
##     dis_mod + rad + tax + ptratio + black + lstat_mod + medv_mod,
##     family = "binomial", data = crime_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4018  -0.1416  -0.0029   0.0032   3.4233
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -42.515655  17.813038  -2.387 0.016997 *
## zn           -0.037515   0.029842  -1.257 0.208703
## indus        -0.051749   0.049379  -1.048 0.294636
## chas          0.970813   0.768970   1.262 0.206774
## nox          54.149495   8.472349   6.391 1.64e-10 ***
## rm_mod      -15.802136  12.885763  -1.226 0.220076
## age_mod       0.010277   0.003204   3.208 0.001336 **
## dis_mod       3.824093   0.986732   3.876 0.000106 ***
## rad           0.634929   0.164849   3.852 0.000117 ***
## tax          -0.004892   0.003173  -1.542 0.123132
## ptratio       0.500107   0.141497   3.534 0.000409 ***
## black        -0.013934   0.007189  -1.938 0.052588 .
## lstat_mod     0.363908   1.824782   0.199 0.841930
## medv_mod     11.900134   4.008860   2.968 0.002993 **
## ---
## 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: 182.76  on 452  degrees of freedom
## AIC: 210.76
```

```
##
## Number of Fisher Scoring iterations: 9
```

Model 4: - Only the significant variables from model3 are used in this model.

I removed all the variables that are not significant after seeing their P-Value.

```
model4 <- glm(target ~ nox + age_mod + dis_mod + rad + ptratio + medv_mod , f
amily="binomial", data=crime_train)

summary(model4)
```

```
##
## Call:
## glm(formula = target ~ nox + age_mod + dis_mod + rad + ptratio +
##     medv_mod, family = "binomial", data = crime_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8866  -0.2127  -0.0217   0.0064   3.2168
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -57.955389   9.233814  -6.276 3.46e-10 ***
## nox          46.172648   7.022160   6.575 4.86e-11 ***
## age_mod       0.009192   0.002487   3.695 0.000220 ***
## dis_mod       3.488834   0.807859   4.319 1.57e-05 ***
## rad           0.529064   0.123587   4.281 1.86e-05 ***
## ptratio       0.398295   0.110069   3.619 0.000296 ***
## medv_mod      7.928413   1.927376   4.114 3.90e-05 ***
## ---
```

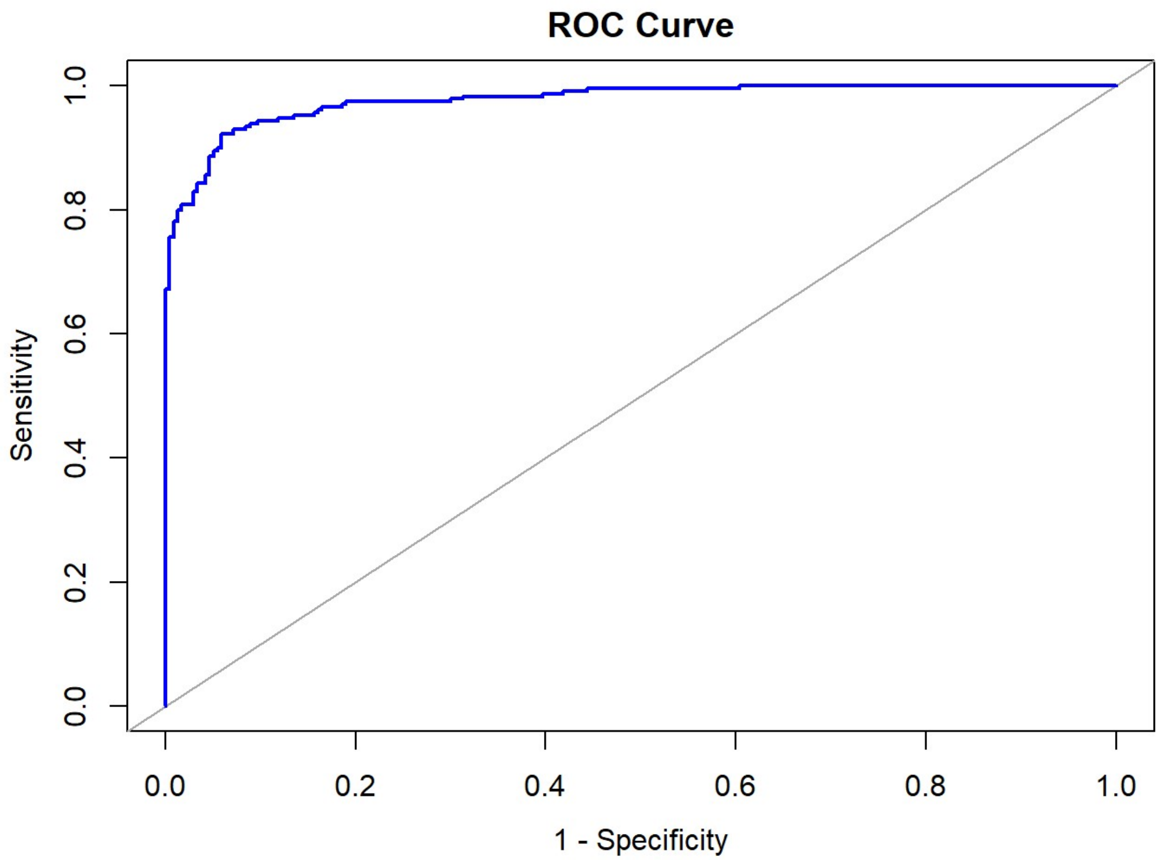
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 203.43  on 459  degrees of freedom
## AIC: 217.43
##
## Number of Fisher Scoring iterations: 9
```

Model Selection.

I would like to select Model3. The AIC and residual deviance for this model seemed to give the best values that would be suited for the prediction. Below is the ROC curve for model3 and to me it looks good. So i would like to proceed with model3.

Validating the model:

I would like to validate the model using some techniques such as ROC curve, confusion Matrix as see the Accuracy, CER, Precision, Sensitivity, Specificity and F1 Score.




```
Area under the curve: 0.9766
```

Now let's do the confusion matrix:

```
crime_train$predict_target <- ifelse(crime_train$predict >=0.5, 1, 0)
crime_train$predict_target <- as.integer(crime_train$predict_target)

myvars <- c("target", "predict_target")
crime_train_cm <- crime_train[myvars]

cm <- table(crime_train_cm$predict_target,crime_train_cm$target)
knitr:: kable(cm)
```

```
##             PredictedValue
## ActualValue FALSE TRUE
##           0   221   18
##           1    16  211
```

Accuracy : 0.9270386

Classification Error Rate: 0.07296137

Precision: 0.9213974

Sensitivity: 0.9295154

Specificity: 0.9246862

F1 Score: 0.9254386

Testing the evaluation data with mode 3

In this final step we will be testing the evaluation data using model3. We need to first pre-preprocess the data in the exact similar way as we did for train data. The Predicted Evaluation data is present at https://github.com/Riteshlohiya/Data621-Week3-Assignment3/blob/master/Evaluation_Data.csv

```
crime_eval <- read.csv("https://raw.githubusercontent.com/Riteshlohiya/Data621-Week3-Assignment3/master/crime-evaluation-data.csv")


crime_eval$age_mod <- crime_eval$age^1.3
crime_eval$lstat_mod <- crime_eval$lstat^0.23
crime_eval$dis_mod <- log(crime_eval$dis)
crime_eval$medv_mod <- crime_eval$medv^0.23
crime_eval$rm_mod <- crime_eval$rm^0.23


crime_eval$predict_prob <- predict(model3, crime_eval, type='response')
crime_eval$predict_target <- ifelse(crime_eval$predict_prob >= 0.50, 1,0)


write.csv(crime_eval,"Evaluation_Data.csv", row.names=FALSE)
```

Appendix

---
title: "Data621 Assignment3"
author: "Ritesh Lohiya"
date: "June 30, 2018"
output: html_document

#Overview
objective is to build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. Below is a short description of the variables in the dataset.

zn: proportion of residential land zoned for large lots (over 25000 square feet)

indus: proportion of non-retail business acres per suburb

chas: a dummy var. for whether the suburb borders the Charles River (1) or not (0)

nox: nitrogen oxides concentration (parts per 10 million)

rm: average number of rooms per dwelling

age: proportion of owner-occupied units built prior to 1940

dis: weighted mean of distances to five Boston employment centers

rad: index of accessibility to radial highways

tax: full-value property-tax rate per $10,000

ptratio: pupil-teacher ratio by town

black: 1000 $(B_k - 0.63)^2$ where Bk is the proportion of blacks by town lstat: lower status of the population (percent)

medv: median value of owner-occupied homes in $1000s

target: whether the crime rate is above the median crime rate (1) or not (0) (response variable)

```{r}
library(readr)
library(kableExtra)
library(tidyverse)
library(knitr)
library(psych)
library(gridExtra)
library(usdm)
library(mice)
library(ggiraph)
library(cowplot)
library(reshape2)
library(corrgram)
library(caTools)
library(caret)
library(ROCR)
library(pROC)
library(reshape2)
library(Amelia)
library(qqplotr)
library(moments)
library(car)
library(MASS)
library(geoR)
```

#DATA EXPLORATION:

```{r}
```

```
crime_train <-
read.csv("https://raw.githubusercontent.com/Riteshlohiya/Data621-Week3-Assignment3/master/crime-training-data.csv")
crime_eval <-
read.csv("https://raw.githubusercontent.com/Riteshlohiya/Data621-Week3-Assignment3/master/crime-evaluation-data.csv")
summary(crime_train)
```

Visual Exploration:

Now we will see the missing values in the dataset. For this i have used Amelia package

```{r}
missmap(crime_train, main = "Missing values vs observed",
color='dodgerblue')
```

There are no missing values in the dataset.

Lets now dig into available variables.

1. zn - proportion of residential land zoned for large lots (over 25000 square feet). We can see there are more zeros values for zn and also has positive skewness. Also there appears to be relationship between crime rates and zn.

```{r}
with(crime_train, c(summary(zn), SD=sd(zn), Skew=skewness(zn),
Kurt=kurtosis(zn)))

hist <- ggplot(crime_train, aes(zn)) + geom_histogram(fill = 'dodgerblue',
binwidth = 20, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of zn') + theme(plot.title =
element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=zn)) +
stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
```

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of zn") + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", zn)) + geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +
  labs(title = 'Boxplot of zn', x="") + theme(plot.title = element_text(hjust = 0.5)) + coord_flip()

box_target <- ggplot(crime_train, aes(x=factor(target), zn)) + geom_boxplot(fill='dodgerblue', color='darkgrey') +
  labs(x='target', title = 'Boxplot of zn by target') + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

2. indus - proportion of non-retail business acres per suburb. The histogram below indicates a bi-modal quality to the variable’s distribution, with many values clustering in two ranges.

```{r}
with(crime_train, c(summary(indus), SD=sd(indus), Skew=skewness(indus), Kurt=kurtosis(indus)))

hist <- ggplot(crime_train, aes(indus)) + geom_histogram(fill = 'dodgerblue', binwidth = 5, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of indus') + theme(plot.title = element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=indus)) + stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
  labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of indus") + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", indus)) + geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +

  labs(title = 'Boxplot of indus', x="") + theme(plot.title = element_text(hjust = 0.5)) + coord_flip()

box_target <- ggplot(crime_train, aes(x=factor(target), indus)) + geom_boxplot(fill='dodgerblue', color='darkgrey') +
  labs(x='target', title = 'Boxplot of indus by target') + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

3. chas - a dummy var. for whether the suburb borders the Charles River (1) or not (0). This variable tells us if the neighborhood borders the Charles River (1) or not (0). Close to 7% of the neighborhood borders the Charles River. Of the areas bordering the Charles river 21 are in high crime areas.

```{r}
addmargins(table(crime_train$chas))
addmargins(table(crime_train$chas, crime_train$target))
```

```{r}
ggplot(crime_train, aes(x=target, y=chas)) + geom_jitter(color='seagreen4') + theme_classic() +
  labs(title ='Jittered Scatter Plot of chas vs.target') + theme(plot.title = element_text(hjust = 0.5))
```

4. nox - nitrogen oxides concentration (parts per 10 million). The variable nox represents the concentration of nitrogen oxide in each Boston area. There is also positive skewness. We also see moderately higher nox variance in high crime areas.

```{r}
with(crime_train, c(summary(nox), SD=sd(nox), Skew=skewness(nox), Kurt=kurtosis(nox)))

hist <- ggplot(crime_train, aes(nox)) + geom_histogram(fill = 'dodgerblue', binwidth = .05, color = 'darkgray' ) +

theme_classic() + labs(title = 'Histogram of nox') + theme(plot.title = element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=nox)) +
stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
 labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of nox")
+ theme_classic() +
 theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", nox)) +
geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +
 labs(title = 'Boxplot of nox', x="") + theme(plot.title = element_text(hjust = 0.5)) + coord_flip()

box_target <- ggplot(crime_train, aes(x=factor(target), nox)) +
geom_boxplot(fill='dodgerblue', color='darkgrey') +
 labs(x='target', title = 'Boxplot of nox by target') + theme_classic() +
 theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

5. rm - average number of rooms per dwelling. The predictor rm is count measure describing the average number of rooms per dwelling. The distridution has heavy tail and has bell curve.

```{r}
with(crime_train, c(summary(rm), SD=sd(rm), Skew=skewness(rm), Kurt=kurtosis(rm)))

hist <- ggplot(crime_train, aes(rm)) + geom_histogram(fill = 'dodgerblue', binwidth = 0.5, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of rm') + theme(plot.title = element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=rm)) +
stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
 labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of rm")
+ theme_classic() +

```
  theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", rm)) +
geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +
  labs(title = 'Boxplot of rm', x="") + theme(plot.title = element_text(hjust =
0.5)) + coord_flip()

box_target <- ggplot(crime_train, aes(x=factor(target), rm)) +
geom_boxplot(fill='dodgerblue', color='darkgrey') +
  labs(x='target', title = 'Boxplot of rm by target') + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

6. age - proportion of owner-occupied units built prior to 1940. The variable age indicates the proportion of owner occupied units built prior to 1940. This variable has high left skewness. Also there is significantly higher mean percentage of older homes in high crime areas.

```{r}
with(crime_train, c(summary(age), SD=sd(age), Skew=skewness(age),
Kurt=kurtosis(age)))

hist <- ggplot(crime_train, aes(age)) + geom_histogram(fill = 'dodgerblue',
binwidth = 5, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of age') + theme(plot.title =
element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=age)) +
stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
  labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of age")
+ theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", age)) +
geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +
  labs(title = 'Boxplot of age', x="") + theme(plot.title = element_text(hjust =
0.5)) + coord_flip()
```

```
box_target <- ggplot(crime_train, aes(x=factor(target), age)) +
geom_boxplot(fill='dodgerblue', color='darkgrey') +
  labs(x='target', title = 'Boxplot of age by target') + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

7. dis - weighted mean of distances to five Boston employment centers. The predictor dist describes the average distance to Boston employment centers. The variable is moderately right skewed. Also we can see that low crime areas are associated with higher average distances to employment centers.

```{r}
with(crime_train, c(summary(dis), SD=sd(dis), Skew=skewness(dis),
Kurt=kurtosis(dis)))

hist <- ggplot(crime_train, aes(dis)) + geom_histogram(fill = 'dodgerblue',
binwidth = 1, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of dis') + theme(plot.title =
element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=dis)) +
stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
  labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of dis")
+ theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", dis)) +
geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +
  labs(title = 'Boxplot of dis', x="") + theme(plot.title = element_text(hjust =
0.5)) + coord_flip()

box_target <- ggplot(crime_train, aes(x=factor(target), dis)) +
geom_boxplot(fill='dodgerblue', color='darkgrey') +
  labs(x='target', title = 'Boxplot of dis by target') + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))
```

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

8. rad - index of accessibility to radial highways. The rad variable is an integer-valued index measure indicating an area’s accessibility to radial highways. In the boxplots below, there appears to be a significant positive association between high crime rates and rad value.

```{r}
with(crime_train, c(summary(rad), SD=sd(rad), Skew=skewness(rad), Kurt=kurtosis(rad)))

hist <- ggplot(crime_train, aes(rad)) + geom_histogram(fill = 'dodgerblue', binwidth = 1, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of rad') + theme(plot.title = element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=rad)) + stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
  labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of rad") + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", rad)) + geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +
  labs(title = 'Boxplot of rad', x="") + theme(plot.title = element_text(hjust = 0.5)) + coord_flip()

box_target <- ggplot(crime_train, aes(x=factor(target), rad)) + geom_boxplot(fill='dodgerblue', color='darkgrey') +
  labs(x='target', title = 'Boxplot of rad by target') + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

9. tax - full-value property-tax rate per $10,000. The tax variable refers to the the tax rate per $10k of property value. High crime areas also appear to have
```

a strong, positive association with the tax value. This variable is densely distributed around two of the following approximate values: 300 and 700.

```{r}
with(crime_train, c(summary(tax), SD=sd(tax), Skew=skewness(tax),
Kurt=kurtosis(tax)))

hist <- ggplot(crime_train, aes(tax)) + geom_histogram(fill = 'dodgerblue',
binwidth = 20, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of tax') + theme(plot.title =
element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=tax)) +
stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
  labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of tax")
+ theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", tax)) +
geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +
  labs(title = 'Boxplot of tax', x="") + theme(plot.title = element_text(hjust =
0.5)) + coord_flip()

box_target <- ggplot(crime_train, aes(x=factor(target), tax)) +
geom_boxplot(fill='dodgerblue', color='darkgrey') +
  labs(x='target', title = 'Boxplot of tax by target') + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

10. ptratio - ptratio: pupil-teacher ratio by town. The predictor ptratio indicates the average school, pupil-to-student ratio, and has a left skewed distribution. We can see a positive relationship between ptratio and high crime.

```{r}
with(crime_train, c(summary(ptratio), SD=sd(ptratio),
Skew=skewness(ptratio), Kurt=kurtosis(ptratio)))
```

```
hist <- ggplot(crime_train, aes(ptratio)) + geom_histogram(fill = 'dodgerblue',
binwidth = 1, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of ptratio') + theme(plot.title =
element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=ptratio)) +
stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
  labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of
ptratio") + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", ptratio)) +
geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +
  labs(title = 'Boxplot of ptratio', x="") + theme(plot.title = element_text(hjust
= 0.5)) + coord_flip()

box_target <- ggplot(crime_train, aes(x=factor(target), ptratio)) +
geom_boxplot(fill='dodgerblue', color='darkgrey') +
  labs(x='target', title = 'Boxplot of ptratio by target') + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

11. black - 1000 (Bk−0.63)2(Bk−0.63)2 where Bk is the proportion of blacks by town. This variable is heavily left skewed.

```{r}
with(crime_train, c(summary(black), SD=sd(black), Skew=skewness(black),
Kurt=kurtosis(black)))

hist <- ggplot(crime_train, aes(black)) + geom_histogram(fill = 'dodgerblue',
binwidth = 40, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of black') + theme(plot.title =
element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=black)) +
stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
```

```
 labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of
black") + theme_classic() +
 theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", black)) +
geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +
 labs(title = 'Boxplot of black', x="") + theme(plot.title = element_text(hjust =
0.5)) + coord_flip()

box_target <- ggplot(crime_train, aes(x=factor(target), black)) +
geom_boxplot(fill='dodgerblue', color='darkgrey') +
 labs(x='target', title = 'Boxplot of black by target') + theme_classic() +
 theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

12. lstat - lower status of the population (percent). The variable lstat indicates the proportion of the population deemed to be of lower status. lstat is right skewed. High crime areas tend to have be associated with larger lstat values.

```{r}
with(crime_train, c(summary(lstat), SD=sd(lstat), Skew=skewness(lstat),
Kurt=kurtosis(lstat)))

hist <- ggplot(crime_train, aes(lstat)) + geom_histogram(fill = 'dodgerblue',
binwidth = 2, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of lstat') + theme(plot.title =
element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=lstat)) +
stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
 labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of
lstat") + theme_classic() +
 theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", lstat)) +
geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +
```

  labs(title = 'Boxplot of lstat', x="") + theme(plot.title = element_text(hjust = 0.5)) + coord_flip()

box_target <- ggplot(crime_train, aes(x=factor(target), lstat)) + geom_boxplot(fill='dodgerblue', color='darkgrey') +
  labs(x='target', title = 'Boxplot of lstat by target') + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

13. medv - median value of owner-occupied homes in $1000s. The median value of residential homes in a given area. The variable is slightly right skewed, and high values of medv appear to be associated with lower crime rates.

```{r}
with(crime_train, c(summary(medv), SD=sd(medv), Skew=skewness(medv), Kurt=kurtosis(medv)))

hist <- ggplot(crime_train, aes(medv)) + geom_histogram(fill = 'dodgerblue', binwidth = 2, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of medv') + theme(plot.title = element_text(hjust = 0.5))

qq_plot <- ggplot(crime_train, aes(sample=medv)) + stat_qq_point(color='dodgerblue') + stat_qq_line(color='darkgray') +
  labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of medv") + theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

box_plot <- ggplot(crime_train, aes(x="", medv)) + geom_boxplot(fill='dodgerblue', color='darkgray')+ theme_classic() +
  labs(title = 'Boxplot of medv', x="") + theme(plot.title = element_text(hjust = 0.5)) + coord_flip()

box_target <- ggplot(crime_train, aes(x=factor(target), medv)) + geom_boxplot(fill='dodgerblue', color='darkgrey') +
  labs(x='target', title = 'Boxplot of medv by target') + theme_classic() +

  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist, qq_plot, box_plot, box_target, ncol=2)
```

Finding correlations: The correlation plot below shows how variables in the dataset are related to each other. Looking at the plot, we can see that certain variables are more related than others.

```{r}
names(crime_train)
cor(drop_na(crime_train))
```

```{r}
pairs.panels(crime_train[1:14])
```

#DATA PREPARATION:

a. Missing Values - there are no missing values, so we will not do any missing value treatment.

b. outliers: I think we dont have any outliers that we should be removing at this stage.

c. Transformation -

age and lstat are both skewed, so lets see boxcox transformation suggestions.

```{r}
boxcoxfit(crime_train$age)
boxcoxfit(crime_train$lstat)
```

so for age the boxcox fit suggested power transformation of 1.3 and for lstat boxcox fit suggested power transformation of 0.23. Lets apply the same.

```{r}

crime_train$age_mod <- crime_train$age^1.3
crime_train$lstat_mod <- crime_train$lstat^0.23
```

The predictor dis, rm and medv has a moderate positive skew. Let's transform using the box-cox transformation

```{r}
boxcoxfit(crime_train$dis)
boxcoxfit(crime_train$rm)
boxcoxfit(crime_train$medv)
```

so for medv and rm the boxcox fit suggested power transformation of .23. Lets apply the same

```{r}
crime_train$medv_mod <- crime_train$medv^0.23
crime_train$rm_mod <- crime_train$rm^0.23
```

The lamda for the boxcoxfit for is dis is alose to 0, so we can apply log transformation.

```{r}
crime_train$dis_mod <- log(crime_train$dis)
```

Lets plot to see the status of the variables after transformation:

```{r}
hist_medv <- ggplot(crime_train, aes(medv_mod)) + geom_histogram(fill = 'dodgerblue', binwidth = 0.2, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of medv_mod') + theme(plot.title = element_text(hjust = 0.5))

hist_lstat <- ggplot(crime_train, aes(lstat_mod)) + geom_histogram(fill = 'dodgerblue', binwidth = 0.2, color = 'darkgray' ) +

theme_classic() + labs(title = 'Histogram of lstat_mod') + theme(plot.title = element_text(hjust = 0.5))

hist_dis <- ggplot(crime_train, aes(dis_mod)) + geom_histogram(fill = 'dodgerblue', binwidth = 0.2, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of dis_mod') + theme(plot.title = element_text(hjust = 0.5))

hist_rm <- ggplot(crime_train, aes(rm_mod)) + geom_histogram(fill = 'dodgerblue', binwidth = 0.025, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of rm_mod') + theme(plot.title = element_text(hjust = 0.5))

hist_age <- ggplot(crime_train, aes(age_mod)) + geom_histogram(fill = 'dodgerblue', binwidth = 50, color = 'darkgray' ) +
 theme_classic() + labs(title = 'Histogram of age_mod') + theme(plot.title = element_text(hjust = 0.5))

grid.arrange(hist_medv, hist_lstat, hist_dis, hist_rm , hist_age, ncol=2)
```

We can see that the skewness of the transformed variables improved.

#BUILD MODELS:

Model 1 - : All original variables model . In this model we will use all the variables. This can be our base model and this model will not include any transformations. We can see which variables are significant. This will help us in looking at the P-Values and removing the non significant variables.

```{r}
model1 <- glm(target ~ zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + black + lstat + medv , family="binomial", data=crime_train)
summary(model1)
```

Model 2: - All significant original variables model. I came up with this models after analyzing the output of model1. I removed all the variables that are not significant after seeing their P-Value.
```

```{r}
model2 <- glm(target ~ nox + age + dis + rad + tax + ptratio + black +  medv ,
family="binomial", data=crime_train)
summary(model2)
```

Model 3: - All variables with transformations(will keep variables that were not transformed)

Model 3 includes original variables, plus the transformed variables from the transformations like power transformation and log transformations. This transfornation should help in reducing the skewness in the data or help them to become more normalized. This will help us in looking at the P-Values and removing the non significant variables.

```{r}
model3 <- glm(target ~ zn + indus + chas + nox + rm_mod + age_mod +
dis_mod + rad + tax + ptratio + black + lstat_mod + medv_mod ,
family="binomial", data=crime_train)
summary(model3)
```

Model 4: - Only the significant variables from model3 are used in this model. I removed all the variables that are not significant after seeing their P-Value.

```{r}
model4 <- glm(target ~ nox + age_mod + dis_mod + rad + ptratio + medv_mod
, family="binomial", data=crime_train)
summary(model4)
```

#MODEL SELECTION:

I would like to select Model3. The AIC and residual deviance for this model seemed to give the best values that would be suited for the prediction. Below is the ROC curve for model3 and to me it looks good. So i would like to proceed with model3

```{r}
crime_train$predict <- predict(model3, crime_train, type='response')

roc_model3 <- roc(crime_train$target, crime_train$predict, plot=T, asp=NA,
            legacy.axes=T, main = "ROC Curve", col="blue")

roc_model3["auc"]
```

Now lets do the confusion matrix:

```{r}
crime_train$predict_target <- ifelse(crime_train$predict >=0.5, 1, 0)
crime_train$predict_target <- as.integer(crime_train$predict_target)
myvars <- c("target", "predict_target")
crime_train_cm <- crime_train[myvars]
cm <- table(crime_train_cm$predict_target,crime_train_cm$target)
knitr:: kable(cm)
```

```{r}
Accuracy <- function(data) {
tb <- table(crime_train_cm$predict_target,crime_train_cm$target)
TN=tb[1,1]
TP=tb[2,2]
FN=tb[2,1]
FP=tb[1,2]
return((TP+TN)/(TP+FP+TN+FN))
}
Accuracy(data)
```

```{r}
CER <- function(data) {
tb <- table(crime_train_cm$predict_target,crime_train_cm$target)
TN=tb[1,1]
TP=tb[2,2]
FN=tb[2,1]
FP=tb[1,2]
```

````
return((FP+FN)/(TP+FP+TN+FN))
}
CER(data)
```

```{r}
Precision <- function(data) {
tb <- table(crime_train_cm$predict_target,crime_train_cm$target)
TP=tb[2,2]
FP=tb[1,2]
return((TP)/(TP+FP))
}
Precision(data)
```

```{r}
Sensitivity <- function(data) {
tb <- table(crime_train_cm$predict_target,crime_train_cm$target)
TP=tb[2,2]
FN=tb[2,1]
return((TP)/(TP+FN))
}
Sensitivity(data)
```

```{r}
Specificity <- function(data) {
tb <- table(crime_train_cm$predict_target,crime_train_cm$target)
TN=tb[1,1]
TP=tb[2,2]
FN=tb[2,1]
FP=tb[1,2]
return((TN)/(TN+FP))
}
Specificity(data)
```

```{r}
F1_score <- function(data) {
````

```
tb <- table(crime_train_cm$predict_target,crime_train_cm$target)
TN=tb[1,1]
TP=tb[2,2]
FN=tb[2,1]
FP=tb[1,2]
Precision = (TP)/(TP+FP)
Sensitivity = (TP)/(TP+FN)
Precision =(TP)/(TP+FP)
return((2*Precision*Sensitivity)/(Precision+Sensitivity))
}
F1_score(data)
```

#TEST DATA PREPARATION AND TESTING THE MODEL ON EVALUATION DATA:

In the final step we will test our model by using the test data.

```{r}
crime_eval <-
read.csv("https://raw.githubusercontent.com/Riteshlohiya/Data621-Week3-Assignment3/master/crime-evaluation-data.csv")
crime_eval$age_mod <- crime_eval$age^1.3
crime_eval$lstat_mod <- crime_eval$lstat^0.23
crime_eval$dis_mod <- log(crime_eval$dis)
crime_eval$medv_mod <- crime_eval$medv^0.23
crime_eval$rm_mod <- crime_eval$rm^0.23
crime_eval$predict_prob <- predict(model3, crime_eval, type='response')
crime_eval$predict_target <- ifelse(crime_eval$predict_prob >= 0.50, 1,0)
write.csv(crime_eval,"Evaluation_Data.csv", row.names=FALSE)
```

The Predicted Evaluation data is present at https://github.com/Riteshlohiya/Data621-Week3-Assignment3/blob/master/Evaluation_Data.csv