

Home Work Assignment - 03

Critical Thinking Group 5

Contents

Overview	2
1 Data Exploration Analysis	2
1.1 Variable identification	2
1.2 Variable Relationships	3
1.3 Data summary analysis	3
1.2 Data Summary Analysis	8
1.3 Outliers and Missing Values Identification	9
2. Data Preparation	12
2.1 Outliers treatment and transformation	12
3 Build Models	15
3.1.1 Model One by using all given variable	18
3.1.3 Model three- model with transformed variables	22
4 Model Selection	25
4.1 Model selection strategy:	25
4.1.1 Model1 Evaluation	26
4.1.2 Model2 Evaluation	26
4.1.3 Model3 Evaluation	27
4.1.4 Model4 Evaluation	28
4.1.5 Model5 Evaluation	29
4.1.6 Model6 Evaluation	29

Overview

The data set contains approximately 466 records and 14 variables. Each record has 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).

The 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. In addition, we will provide classifications and probabilities for the evaluation data set using the binary logistic regression model.

1 Data Exploration Analysis

In section we will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:

- Variable identification
- Variable Relationships
- Data summary analysis
- Outliers and Missing Values Identification

1.1 Variable identification

First let's display and examine the data dictionary or the data columns as shown in table 1

Table 1: Variable Description

Variable	Description	Datatype	Role
zn	proportion of residential land zoned for large lots (over 25000 square feet)	numeric	predictor
indus	proportion of non-retail business acres per suburb	numeric	predictor
chas	a dummy var. for whether the suburb borders the Charles River (1) or not (0)	binary	predictor
nox	nitrogen oxides concentration (parts per 10 million)	numeric	predictor
rm	average number of rooms per dwelling	numeric	predictor
age	proportion of owner-occupied units built prior to 1940	numeric	predictor
dis	weighted mean of distances to five Boston employment centers	numeric	predictor
rad	index of accessibility to radial highways	integer	predictor
tax	full-value property-tax rate per \$10,000	integer	predictor
ptratio	pupil-teacher ratio by town	numeric	predictor
black	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town	numeric	predictor
lstat	lower status of the population (percent)	numeric	predictor
medv	median value of owner-occupied homes in \$1000s	numeric	predictor
target	whether the crime rate is above the median crime rate (1) or not (0)	binary	response

We notice that all variables are numeric except for two variables: the response variable “target” which is binary and the predictor variable “chas” which is a dummy binary variable indicating whether the suburb borders the Charles River (1) or not (0).

Based on the original dataset, our predictor input is made of 13 variables. And our response variable is one variable called target.

1.2 Variable Relationships

The variables seem to not have any arithmetic relations. In other words, there are no symmetricity or transitivity relationships between any two variable in the independent variable set.

In addition, since this is Logistic Regression, we will be making the below assumptions on the variables:

-The dependent variable need not to be normally distributed

-Errors need to be independent but not normally distributed.

- We will be using GLM and GLM does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.

- Also does not use OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE)

1.3 Data summary analysis

```
summary(city_crime_train_full)
```

```
##           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
```

Table 2: Missing Values

zn	0
indus	0
chas	0
nox	0
rm	0
age	0
dis	0
rad	0
tax	0
ptratio	0
black	0
lstat	0
medv	0
target	0

Missing values vs observed

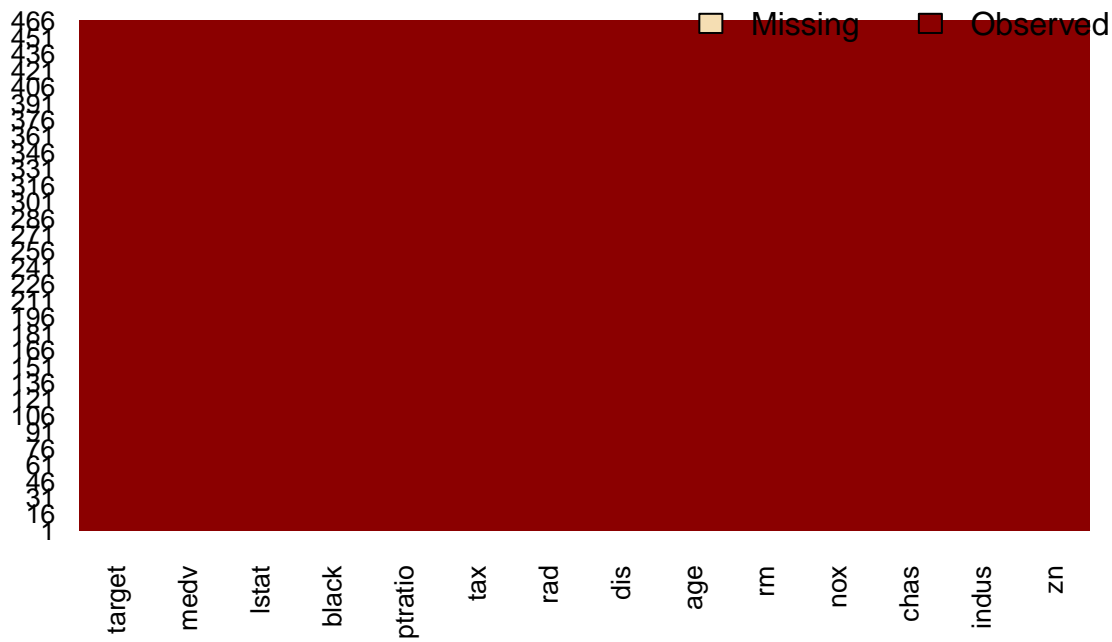


Table 3: Unique Values

zn	26
indus	73
chas	2
nox	79
rm	419
age	333
dis	380
rad	9

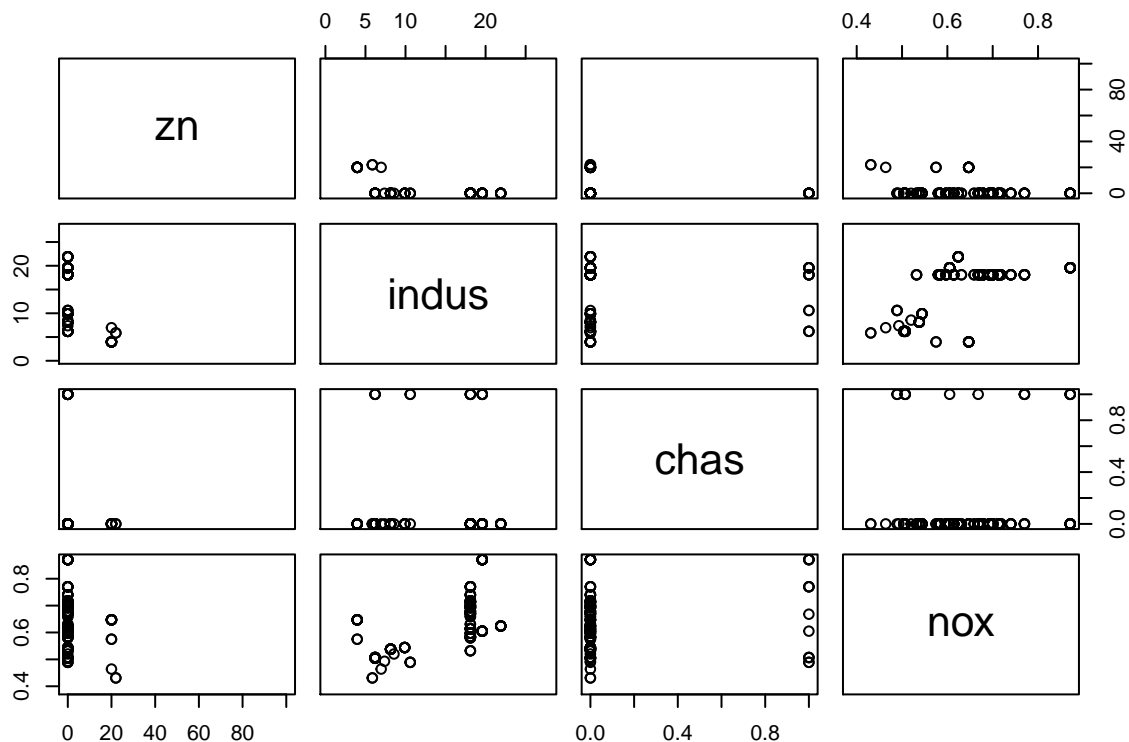
tax	63
ptratio	46
black	331
lstat	424
medv	218
target	2

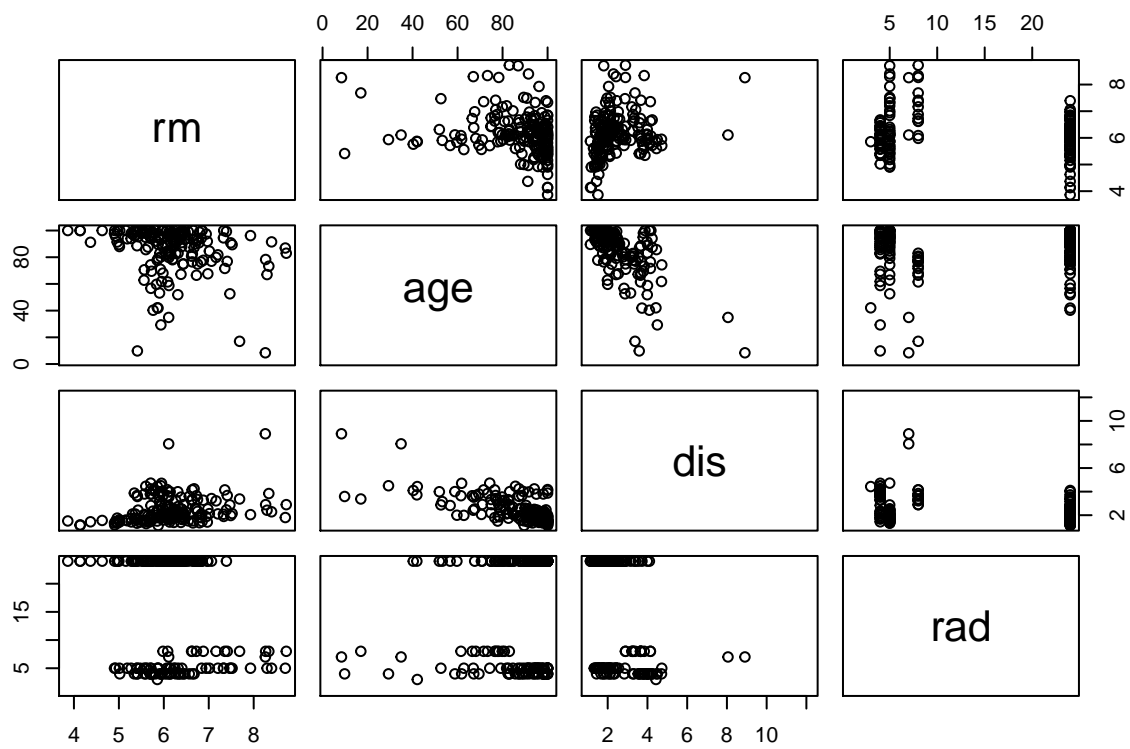
```
##
##          0          1
## 0.5085837 0.4914163
```

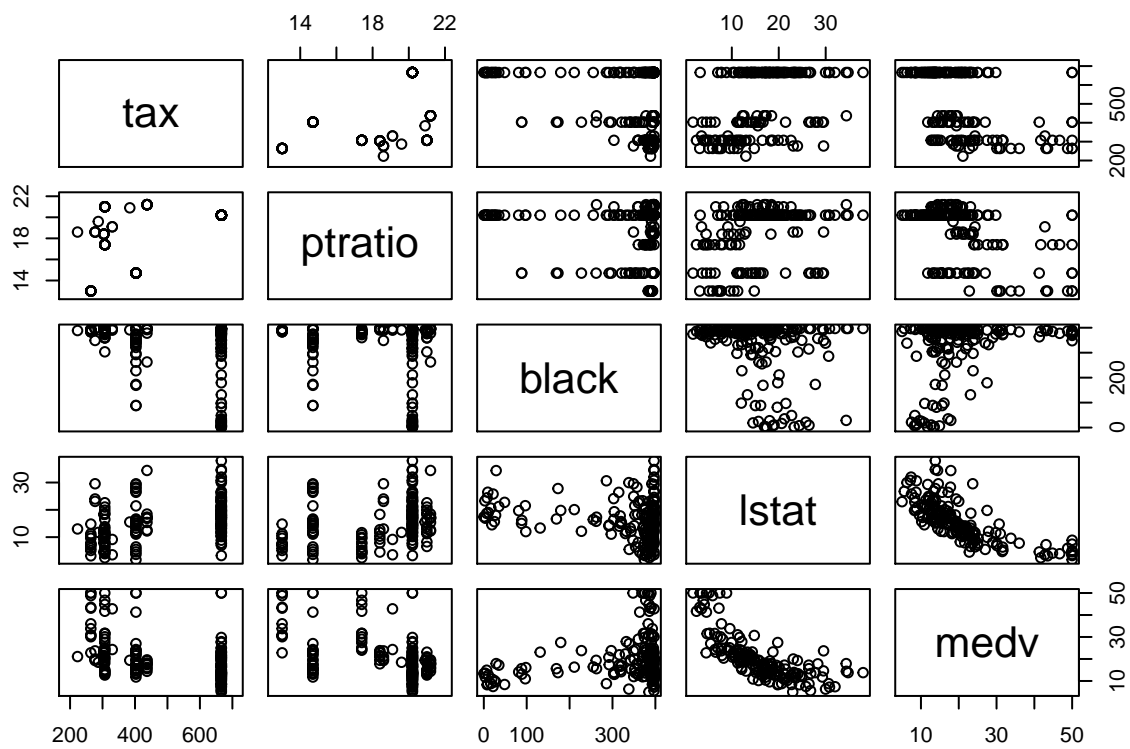
Based on the analysis above it can be seen that there is no missing value in the data set. Also count of unique values for each variable is shown above. Also % split of target variable is given above table which shows data is almost evenly split between binary outcome 0 and 1.

Train data set will be Split into train data(80% of train set) and validation set (20% of train set)to evaluate the performnce of the models on the validation set. Train subset will be used to build the models.

Two data set has been created city_crime_train (80% of train data), and train_test (20% of train data). In next step below relationship between the target variable and dependent variables is shown in three charts.







1.2 Data Summary Analysis

In this section, we will create summary data to better understand the initial relationship variables have with our dependent variable using correlation, central tendency, and dispersion As shown in table 2.

Table 4: Data Summary

	vars	n	mean	sd	median	trimmed	mad
zn	1	372	12.3615591	24.0566280	0.0000	6.0436242	0.0000000
indus	2	372	10.8997043	6.9049747	8.5600	10.6588591	7.9022580
chas	3	372	0.0645161	0.2460010	0.0000	0.0000000	0.0000000
nox	4	372	0.5512384	0.1192316	0.5220	0.5394315	0.1230558
rm	5	372	6.2950027	0.6994228	6.2055	6.2664161	0.5315121
age	6	372	67.4131720	28.6905417	76.5000	69.8332215	30.9122100
dis	7	372	3.8438124	2.1293108	3.3246	3.5968611	2.0548095
rad	8	372	9.2043011	8.5398184	5.0000	8.2818792	1.4826000
tax	9	372	403.6854839	167.0523120	330.0000	394.0033557	108.2298000
ptratio	10	372	18.2325269	2.2232031	18.6000	18.4144295	2.3721600
black	11	372	359.6269355	88.5960400	391.9550	384.7739262	7.3314570
lstat	12	372	12.3974731	7.0278483	10.9250	11.6246309	6.7680690
medv	13	372	22.8473118	9.0745857	21.6000	21.9842282	6.9682200
target	14	372	0.4731183	0.4999493	0.0000	0.4664430	0.0000000

Table 5: Data Summary (Cont)

	min	max	range	skew	kurtosis	se
zn	0.0000	100.0000	100.0000	2.0480221	3.1952491	1.2472781
indus	0.4600	27.7400	27.2800	0.3403513	-1.2138927	0.3580063
chas	0.0000	1.0000	1.0000	3.5309878	10.4961121	0.0127546
nox	0.3890	0.8710	0.4820	0.8365808	0.0921839	0.0061819
rm	3.8630	8.7250	4.8620	0.3906669	1.4757590	0.0362634
age	2.9000	100.0000	97.1000	-0.5301804	-1.0930455	1.4875353
dis	1.1296	12.1265	10.9969	0.9561753	0.3805285	0.1103996
rad	1.0000	24.0000	23.0000	1.0969598	-0.6717893	0.4427690
tax	187.0000	711.0000	524.0000	0.7166865	-1.0538103	8.6612589
ptratio	12.6000	22.0000	9.4000	-0.6734659	-0.5183076	0.1152677
black	0.3200	396.9000	396.5800	-3.0998189	8.5467305	4.5934907
lstat	1.7300	37.9700	36.2400	0.9457306	0.6003832	0.3643770
medv	5.0000	50.0000	45.0000	0.9704021	1.1065501	0.4704954
target	0.0000	1.0000	1.0000	0.1072487	-1.9938358	0.0259212

Now we will produce the correlation table between the independent variables and the dependent variable

```
Correlation <- sort(Correlation, decreasing = TRUE)
kable(Correlation, caption = "Variable Correlation")
```

Table 6: Variable Correlation

target	1.0000000
nox	0.7290920
rad	0.6307187
age	0.6275762
indus	0.6034795
tax	0.6021403
lstat	0.4808888
ptratio	0.2198922
chas	0.0579716
rm	-0.1605913
medv	-0.2724789
black	-0.3463425
zn	-0.4239382
dis	-0.6167264

*** Curious It is clear from the table that most of the variables are having strong correlation with the target variable.

Correlation analysis suggests that there are strong positive and negative between the independent variables and the dependent variable. For instance, we notice that there is a strong correlation of .73 between the concentration of nitrogen oxides and crime rate being above average. We will need to perform more investigations about this correlation as it is not obvious the concentration of nitrogen oxides would results in high crime rate; perhaps it impacts the crime rate indirectly by impacting other independent variables that we may or may not have in our data set.

In addition, we noticed that accessibility to radial highways also has a strong correlation with the crime rate being average average. Again we will investigate such correlation. We also noticed that unit or house age, property tax, and non-retail businesses having a positive impact on the crime rate being above average.

It is also worth noting that that distances to five Boston employment centers, large residential lots, the proportion of blacks by town, median value of owner-occupied homes, and the average number of rooms per dwelling, all have negative correlation to the crime rate being above crime rate average. In other words, the closer people are to the five Boston employment centers, the more likely the crime rate will be below the crime average.

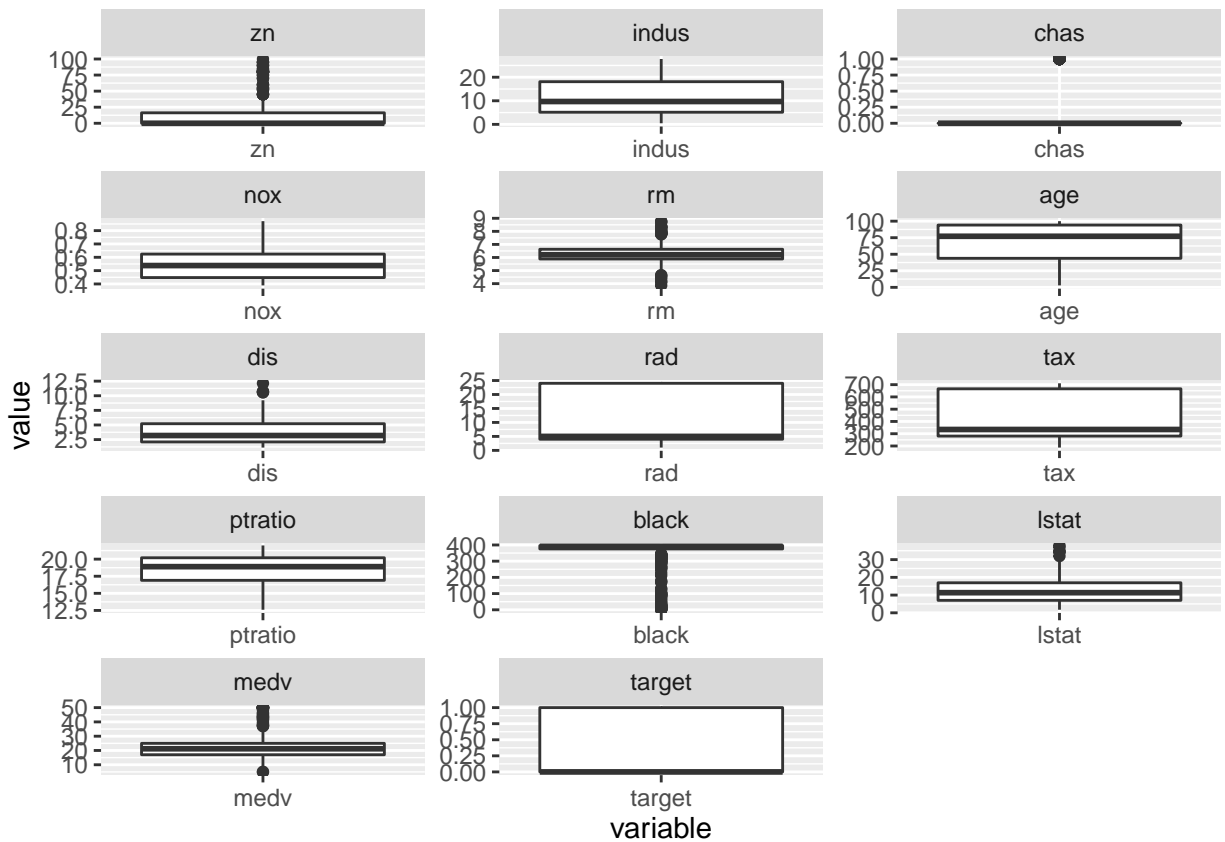
1.3 Outliers and Missing Values Identification

In this section univariate analysis is being carried out and boxplots diagrams are being used to determine the outliers in variables and decide on whether to act on the outliers. Along with boxplot, Histogram, Sin, Log,Sqrt,nth transformation diagrams are used to evaluate best transformation to handle outliers.

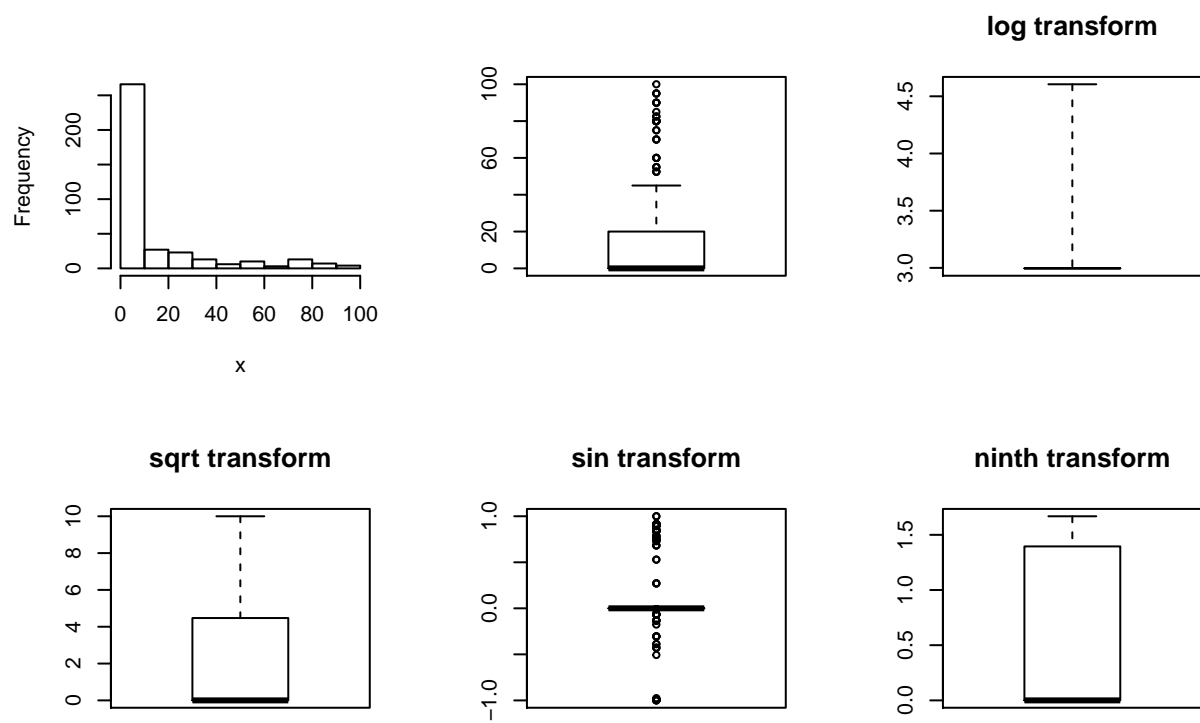
```
library(ggplot2)
library(reshape2)
#create a new data frame with two columns only (variable, value) for all three predictors
mdata <- melt(city_crime_train_full)
```

```
## No id variables; using all as measure variables
```

```
# Output the boxplot
p <- ggplot(data = mdata, aes(x=variable, y=value)) +
  geom_boxplot()
p + facet_wrap( ~ variable, scales="free", ncol=3)
```



Analysis of variable zn:proportion of residential land zoned for large lots



For zn , we can see that there are large number of values with 0. ninth transformation seem better for this variable..(1)

***Please note that we have created similar figures to figure 1 above for each remaining variable. However, we hid the remaining figures for ease of streamlining the report as they have similar shapes. However, we have drawn the below observations from each remaining figure.

For indus, we can see that there is a spike toward right side of the distribution. Looking at the sqrt transformation it appears that distribution is close to normal and having two peaks after transformation.

For nox, there is a long right tail.

For rm, there are some outliers as we can see from box plot. This variable will need some transformation to handle the outliers.

age of the building variable is skewed heavily towards right side. We will need some transformation for this variable and looks sin transformation is best option for this case

For this variable dis, there are some outliers which needs transformation to handle those outliers. log transformation looks best suited for this scenario.

For rad variable distribution is not uniform as seen from the chart and will need transformation.

For tax variable is not uniformly distributed but there is no outlier for this variable.

For pratio has right aligned peak but no outliers are there in data set.

The variable lstat has long right tail and left skewed

2. Data Preparation

Now that we have completed the preliminary analysis, we will be cleaning and consolidating data into one dataset for use in analysis and modeling. We will be purging the below steps as guidelines:

- Outliers treatment
- Missing values treatment
- Data transformation

2.1 Outliers treatment and transformation

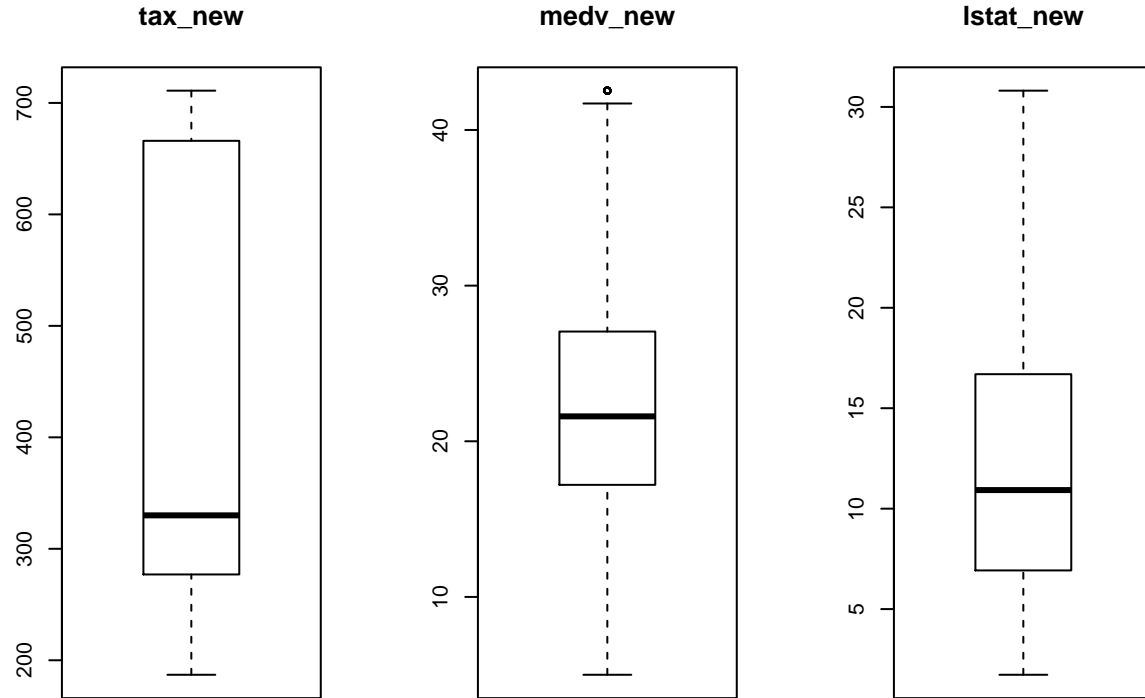
For outliers, we will create 2 sets of variables.

The first set uses the capping method. In this method, we will replace all outliers that lie outside the 1.5 times of IQR limits. We will cap it by replacing those observations less than the lower limit with the value of 5th %ile and those that lie above the upper limit with the value of 95th %ile.

Accordingly we create the following new variables while retaining the original variables.

```
city_crime_train$tax city_crime_train$medv  
city_crime_train$lstat
```

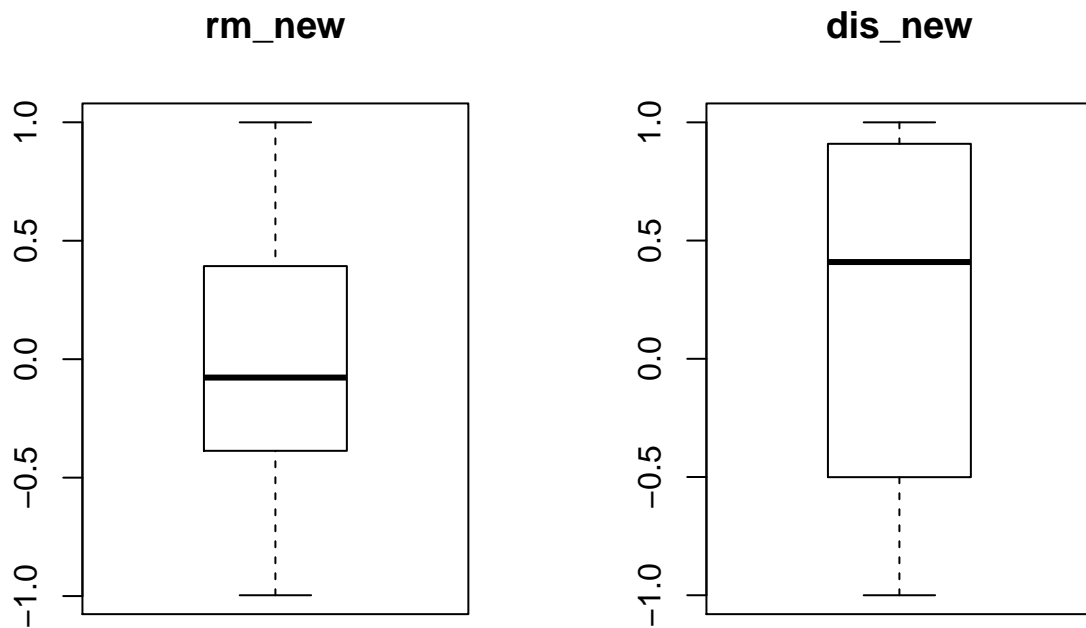
Below boxplots shows distribution of variables after outliers treatment.



In the second set, we will use the sin transformation and create the following variables:

`city_crime_train_modrmnew` `citycrimetrainmoddis_new`

Below is the boxplot after sin transformation of above variable.



Additional transformation was performed on following variables

1. using bucket for zn, with set of values 0 and 1
2. Converting chas to a factor variable of 0 and 1
3. Converting target to a factor variable of 0 and 1

below we evaluate correlation of target with new variables

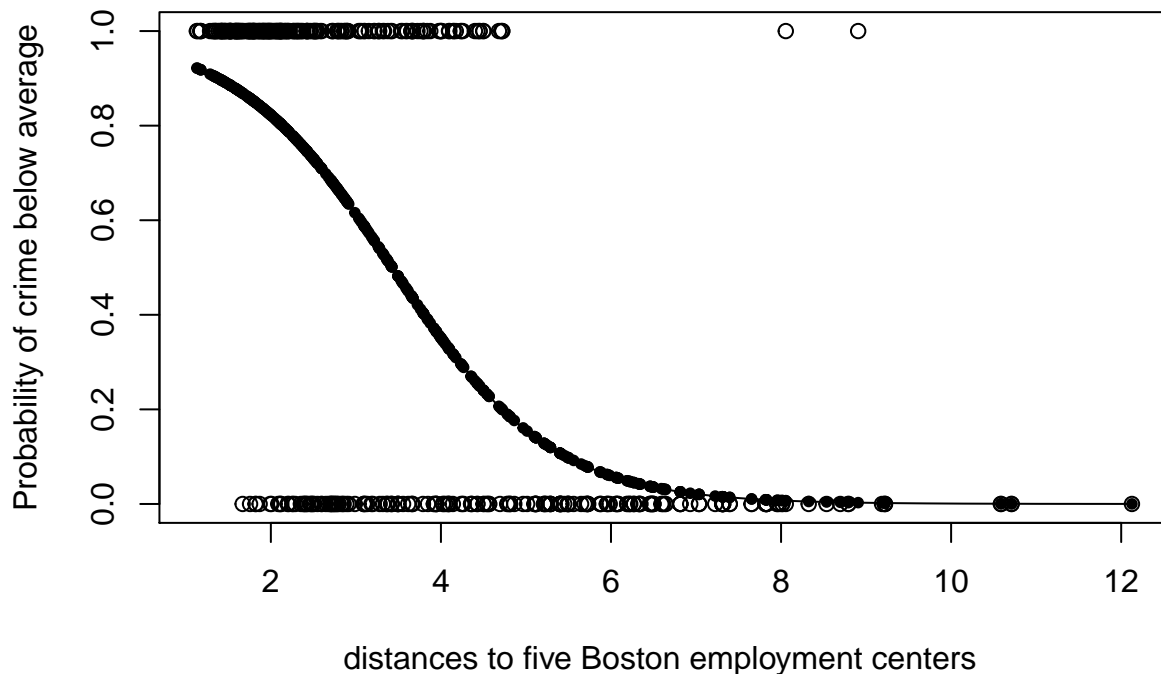
All new variables seem to have a positive correlation with target. However, some of them do not seem to have a strong correlation. Lets see how they perform while modeling.

3 Build Models

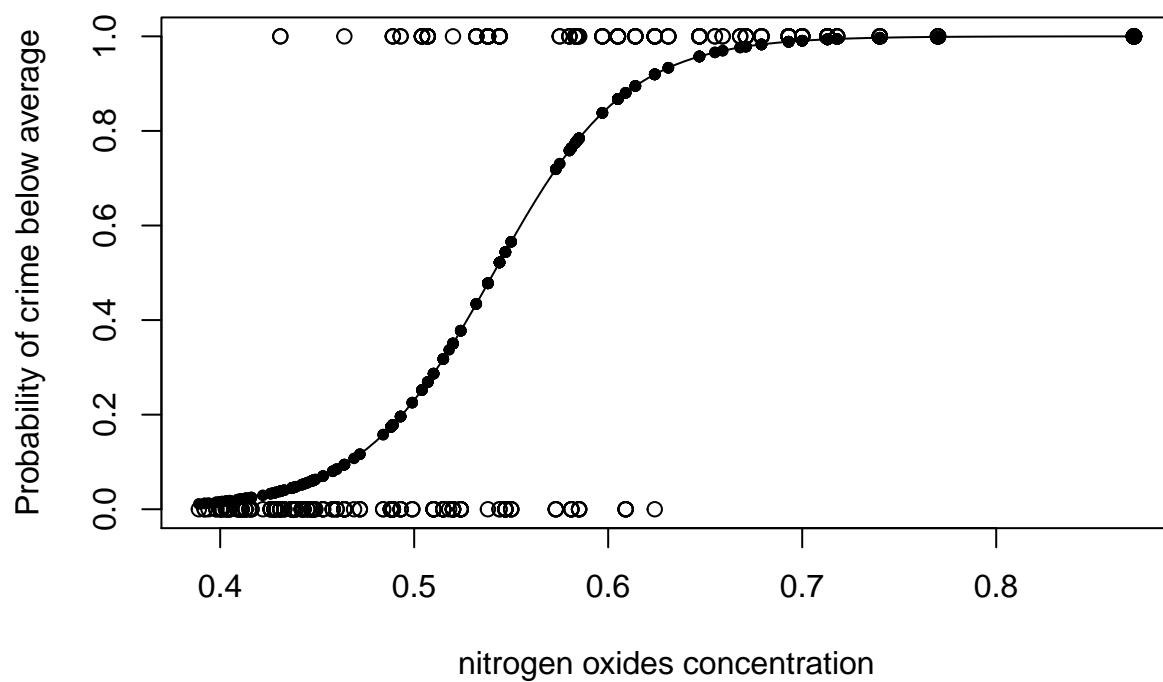
Below is a summary table showing models and their respective variables.

*****For every variable the the model model

```
x = city_crime_train_full
plot(x$dis,x$target,xlab="distances to five Boston employment centers ",ylab="Probability of crime below average")
g=glm(target~dis,family=binomial,x) # run a logistic regression model (in this case, generalized linear model)
curve(predict(g,data.frame(dis=x),type="resp"),add=TRUE) # draws a curve based on prediction from logistic regression
points(x$dis,fitted(g),pch=20)
```



```
x = city_crime_train_full
plot(x$nox,x$target,xlab="nitrogen oxides concentration",ylab="Probability of crime below average") #
g=glm(target~nox,family=binomial,x) # run a logistic regression model (in this case, generalized linear model)
curve(predict(g,data.frame(nox=x),type="resp"),add=TRUE) # draws a curve based on prediction from logistic regression
points(x$nox,fitted(g),pch=20)
```

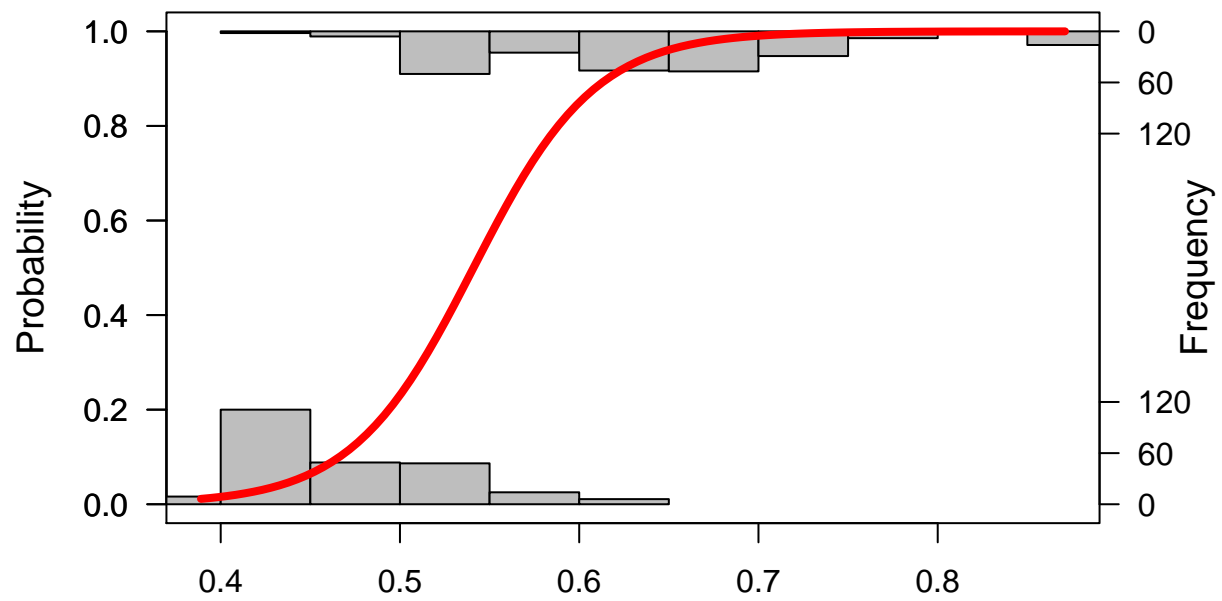


```
library(popbio)
```

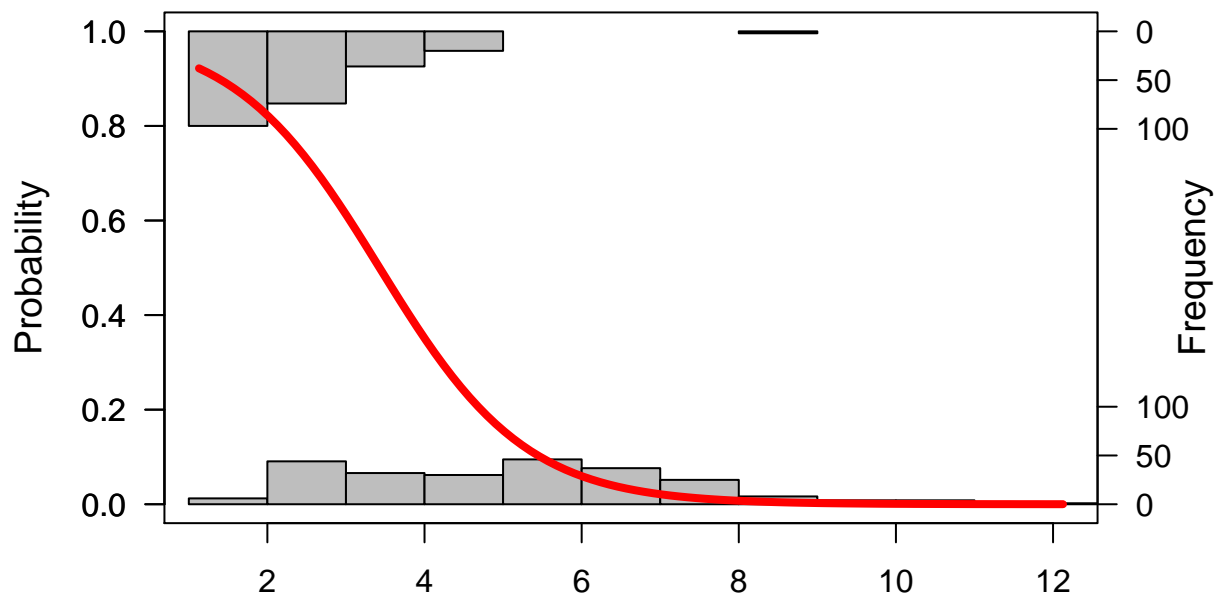
```
##
## Attaching package: 'popbio'

## The following object is masked from 'package:AUC':
##
##     sensitivity
```

```
logi.hist.plot(x$nox,x$target,boxp=FALSE,type="hist",col="gray")
```

```
library(popbio)
logi.hist.plot(x$dis, x$target,boxp=FALSE,type="hist",col="gray")
```



3.1.1 Model One by using all given variable

In this model, we will be using the original variables. We will create model and we will highlight the variables that being recommended using the AIC value.

First we will produce the summary model as per below:

```
##
## Call:
## glm(formula = target ~ ., family = "binomial", data = city_crime_train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.8791  -0.1299  -0.0025   0.0011   3.4785
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -41.462153   8.250799  -5.025 5.03e-07 ***
## zn          -0.060580   0.039153  -1.547 0.121799
## indus       -0.063885   0.059335  -1.077 0.281618
## chas         0.789391   0.865818   0.912 0.361912
## nox         53.413503  10.013666   5.334 9.60e-08 ***
## rm          -0.647942   0.904430  -0.716 0.473739
## age         0.028835   0.015680   1.839 0.065915 .
## dis         0.800917   0.268877   2.979 0.002894 **
## rad         0.721751   0.195662   3.689 0.000225 ***
```

```

## tax          -0.007065    0.003490   -2.024 0.042948 *
## ptratio      0.440768    0.159366    2.766 0.005679 **
## black        -0.009591    0.006025   -1.592 0.111412
## lstat        0.096941    0.062429    1.553 0.120469
## medv         0.236940    0.091276    2.596 0.009436 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 514.63  on 371  degrees of freedom
## Residual deviance: 140.71  on 358  degrees of freedom
## AIC: 168.71
##
## Number of Fisher Scoring iterations: 9

## (Intercept)          zn          indus          chas          nox
## 9.844998e-19 9.412183e-01 9.381125e-01 2.202054e+00 1.574670e+23
##          rm          age          dis          rad          tax
## 5.231212e-01 1.029255e+00 2.227583e+00 2.058033e+00 9.929600e-01
##          ptratio          black          lstat          medv
## 1.553900e+00 9.904547e-01 1.101795e+00 1.267365e+00

```

model interpretation for model 1 Below we analyze and the fitting and interpret what the model is telling us.

- i. First of all, we can see that indus, chas, rm, age, black, and lstat are not statistically significant.
- ii. As for the statistically significant variables, nox has the lowest p-value suggesting a strong association of the nox of the target variable. Other important variables are dis, rad, tax, ptratio, medv. AIC value for the model = 168.71.
- iii. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variables.
 - a. For every one unit change in nox, the log odds of crime rate above median value increases by 53.41.
 - b. For a one unit increase in dis, the log odds of crime rate above median value increases by 0.80.
 - c. For a one unit increase in rad, the log odds of crime rate above median value increases by 0.72.
 - d. For a one unit increase in tax, the log odds of crime rate above median value increases by -0.007.
 - e. For a one unit increase in ptratio, the log odds of crime rate above median value increases by 0.44.
 - f. For a one unit increase in medv, the log odds of crime rate above median value increases by 0.23.

3.1.2 Model two- with backward step function with all given variables

```
stepmodel1<- step(model1, direction="backward")
```

```
## Start: AIC=168.71
## target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
##   ptratio + black + lstat + medv
##
##           Df Deviance    AIC
## - rm      1   141.22 167.22
## - chas    1   141.55 167.55
## - indus   1   141.93 167.93
## <none>    1   140.71 168.71
## - lstat   1   143.06 169.06
## - black   1   143.68 169.68
## - zn      1   143.99 169.99
## - age     1   144.45 170.45
## - tax     1   144.93 170.93
## - medv    1   148.67 174.67
## - ptratio 1   149.29 175.29
## - dis     1   150.97 176.97
## - rad     1   171.94 197.94
## - nox     1   195.65 221.65
##
## Step: AIC=167.22
## target ~ zn + indus + chas + nox + age + dis + rad + tax + ptratio +
##   black + lstat + medv
##
##           Df Deviance    AIC
## - chas    1   142.10 166.10
## - indus   1   142.37 166.37
## <none>    1   141.22 167.22
## - black   1   144.02 168.02
## - age     1   144.48 168.48
## - zn      1   144.74 168.74
## - lstat   1   145.13 169.13
## - tax     1   145.97 169.97
## - ptratio 1   149.78 173.78
## - dis     1   150.97 174.97
## - medv    1   156.73 180.73
## - rad     1   172.26 196.26
## - nox     1   196.29 220.29
##
## Step: AIC=166.1
## target ~ zn + indus + nox + age + dis + rad + tax + ptratio +
##   black + lstat + medv
##
##           Df Deviance    AIC
## - indus   1   142.85 164.85
## <none>    1   142.10 166.10
## - black   1   144.69 166.69
## - age     1   145.65 167.65
## - zn      1   146.09 168.09
## - lstat   1   146.43 168.43
## - tax     1   148.34 170.34
```

```
## - ptratio 1 149.90 171.90
## - dis 1 151.42 173.42
## - medv 1 157.16 179.16
## - rad 1 177.68 199.68
## - nox 1 196.44 218.44
##
## Step: AIC=164.85
## target ~ zn + nox + age + dis + rad + tax + ptratio + black +
## lstat + medv
##
## Df Deviance AIC
## <none> 142.85 164.85
## - black 1 145.21 165.21
## - age 1 146.69 166.69
## - lstat 1 146.75 166.75
## - zn 1 146.89 166.89
## - ptratio 1 150.46 170.46
## - dis 1 151.87 171.87
## - tax 1 154.08 174.08
## - medv 1 157.59 177.59
## - rad 1 184.71 204.71
## - nox 1 203.12 223.12
```

```
summary(stepmodel1)
```

```
##
## Call:
## glm(formula = target ~ zn + nox + age + dis + rad + tax + ptratio +
## black + lstat + medv, family = "binomial", data = city_crime_train)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -1.9258 -0.1459 -0.0024 0.0013 3.3934
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -39.282116 7.705519 -5.098 3.43e-07 ***
## zn -0.064656 0.037414 -1.728 0.083964 .
## nox 46.617168 8.074920 5.773 7.78e-09 ***
## age 0.025273 0.013545 1.866 0.062065 .
## dis 0.710480 0.249767 2.845 0.004447 **
## rad 0.775881 0.182072 4.261 2.03e-05 ***
## tax -0.009144 0.003082 -2.967 0.003011 **
## ptratio 0.359297 0.135081 2.660 0.007817 **
## black -0.008384 0.005737 -1.462 0.143871
## lstat 0.110624 0.055650 1.988 0.046829 *
## medv 0.181460 0.053572 3.387 0.000706 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 514.63 on 371 degrees of freedom
## Residual deviance: 142.85 on 361 degrees of freedom
```

```
## AIC: 164.85
##
## Number of Fisher Scoring iterations: 9
```

model interpretation for model 2 Below we analyze and the fitting and interpret what the model is telling us.

i. First of all, we can see that zn, age, black are not statistically significant.

ii. As for the statistically significant variables, nox has the lowest p-value suggesting a strong association of the nox of the target variable. Other important variables are dis, rad, tax, ptratio, medv, lstat. AIC value for the model1 = 164.85

iii. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variables.

a. For every one unit change in nox, the log odds of crime rate above median value increases by 46.61.

b. For a one unit increase in dis, the log odds of crime rate above median value increases by 0.71.

c. For a one unit increase in rad, the log odds of crime rate above median value increases by 0.77.

d. For a one unit increase in tax, the log odds of crime rate above median value increases by -0.009.

e. For a one unit increase in ptratio, the log odds of crime rate above median value increases by 0.35.

f. For a one unit increase in medv, the log odds of crime rate above median value increases by 0.18

iv. there were 9 iterations in backward steps before final model was selected

3.1.3 Model three- model with transformed variables

In this model, we will be using the some transformed variables.

First we will produce the summary model as per below:

```
##
## Call:
## glm(formula = target ~ . - zn - tax - lstat - medv, family = "binomial",
##      data = city_crime_train_mod)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7883  -0.1410  -0.0026   0.0005   3.3645
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -68.319369  16.418997  -4.161 3.17e-05 ***
## indus       -0.001867   0.067017  -0.028 0.977778
## chas1         0.366993   0.849076   0.432 0.665577
## nox          56.080643  10.147964   5.526 3.27e-08 ***
## rm           2.995884   2.385419   1.256 0.209147
```

```
## age          0.043435    0.018166    2.391 0.016805 *
## dis          0.472036    0.331312    1.425 0.154231
## rad          0.838409    0.237364    3.532 0.000412 ***
## ptratio      0.468316    0.176293    2.656 0.007896 **
## black        -0.010739    0.005922   -1.813 0.069782 .
## tax_new      -0.005285    0.003663   -1.443 0.149151
## medv_new      0.283102    0.106228    2.665 0.007698 **
## lstat_new     0.050027    0.074958    0.667 0.504515
## rm_new       -5.052053    2.830695   -1.785 0.074304 .
## dis_new      -1.886385    0.552223   -3.416 0.000636 ***
## zn_new       -0.363834    1.036508   -0.351 0.725574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 514.63  on 371  degrees of freedom
## Residual deviance: 124.11  on 356  degrees of freedom
## AIC: 156.11
##
## Number of Fisher Scoring iterations: 9
```

3.1.4 Model with transformed variable and with with backward step function

```
stepmodel2<- step(model2, direction="backward")
```

```
## Start:  AIC=156.11
## target ~ (zn + indus + chas + nox + rm + age + dis + rad + tax +
##          ptratio + black + lstat + medv + tax_new + medv_new + lstat_new +
##          rm_new + dis_new + zn_new) - zn - tax - lstat - medv
##
##              Df Deviance    AIC
## - indus      1   124.11 154.11
## - zn_new     1   124.24 154.24
## - chas       1   124.30 154.30
## - lstat_new  1   124.54 154.54
## - rm        1   125.88 155.88
## - dis       1   126.02 156.01
## <none>      124.11 156.11
## - tax_new   1   126.11 156.11
## - black     1   127.44 157.44
## - rm_new    1   127.97 157.97
## - age       1   130.93 160.93
## - ptratio   1   131.81 161.81
## - medv_new  1   132.41 162.41
## - dis_new   1   138.64 168.64
## - rad       1   149.17 179.17
## - nox       1   186.38 216.38
##
## Step:  AIC=154.11
## target ~ chas + nox + rm + age + dis + rad + ptratio + black +
##          tax_new + medv_new + lstat_new + rm_new + dis_new + zn_new
```

```

##
##           Df Deviance    AIC
## - zn_new      1   124.24 152.24
## - chas        1   124.31 152.31
## - lstat_new    1   124.55 152.55
## - rm          1   125.88 153.88
## - dis         1   126.04 154.04
## <none>         124.11 154.11
## - tax_new     1   127.03 155.03
## - black       1   127.45 155.45
## - rm_new      1   127.97 155.97
## - age         1   130.96 158.96
## - ptratio     1   131.82 159.82
## - medv_new    1   132.55 160.55
## - dis_new     1   140.43 168.43
## - rad         1   155.61 183.61
## - nox         1   196.97 224.97
##
## Step:  AIC=152.24
## target ~ chas + nox + rm + age + dis + rad + ptratio + black +
##          tax_new + medv_new + lstat_new + rm_new + dis_new
##
##           Df Deviance    AIC
## - chas        1   124.50 150.50
## - lstat_new    1   124.56 150.56
## - rm          1   125.97 151.97
## - dis         1   126.08 152.08
## <none>         124.24 152.24
## - tax_new     1   127.18 153.18
## - black       1   127.72 153.72
## - rm_new      1   128.22 154.22
## - age         1   131.29 157.29
## - medv_new    1   132.64 158.64
## - ptratio     1   134.36 160.36
## - dis_new     1   143.38 169.38
## - rad         1   157.08 183.08
## - nox         1   196.97 222.97
##
## Step:  AIC=150.5
## target ~ nox + rm + age + dis + rad + ptratio + black + tax_new +
##          medv_new + lstat_new + rm_new + dis_new
##
##           Df Deviance    AIC
## - lstat_new    1   124.91 148.91
## - rm          1   126.15 150.15
## - dis         1   126.19 150.19
## <none>         124.50 150.50
## - tax_new     1   127.58 151.58
## - black       1   127.91 151.91
## - rm_new      1   128.38 152.38
## - age         1   131.80 155.80
## - medv_new    1   133.04 157.04
## - ptratio     1   134.38 158.38
## - dis_new     1   144.36 168.36

```



```

## - rad          1    158.12 182.12
## - nox          1    196.98 220.98
##
## Step:  AIC=148.91
## target ~ nox + rm + age + dis + rad + ptratio + black + tax_new +
##      medv_new + rm_new + dis_new
##
##           Df Deviance    AIC
## - rm          1    126.80 148.80
## - dis          1    126.88 148.88
## <none>          124.91 148.91
## - tax_new      1    127.77 149.77
## - black        1    128.14 150.14
## - rm_new       1    130.21 152.21
## - medv_new     1    133.39 155.39
## - ptratio      1    135.25 157.25
## - age          1    135.57 157.57
## - dis_new      1    145.13 167.13
## - rad          1    159.22 181.22
## - nox          1    198.49 220.49
##
## Step:  AIC=148.8
## target ~ nox + age + dis + rad + ptratio + black + tax_new +
##      medv_new + rm_new + dis_new
##
##           Df Deviance    AIC
## <none>          126.80 148.80
## - tax_new      1    129.00 149.00
## - black        1    130.37 150.37
## - dis          1    130.87 150.87
## - rm_new       1    132.36 152.36
## - age          1    138.72 158.72
## - ptratio      1    139.68 159.68
## - medv_new     1    142.98 162.98
## - dis_new      1    146.97 166.97
## - rad          1    160.12 180.12
## - nox          1    203.79 223.79

```

3.1,5 Model three with Linear discrement analysis

3.1.6 Model with Linear discrement analysis with transformed data

4 Model Selection

In section we will further examine all six models. We will apply a model selection strategy defined below to compare the models.

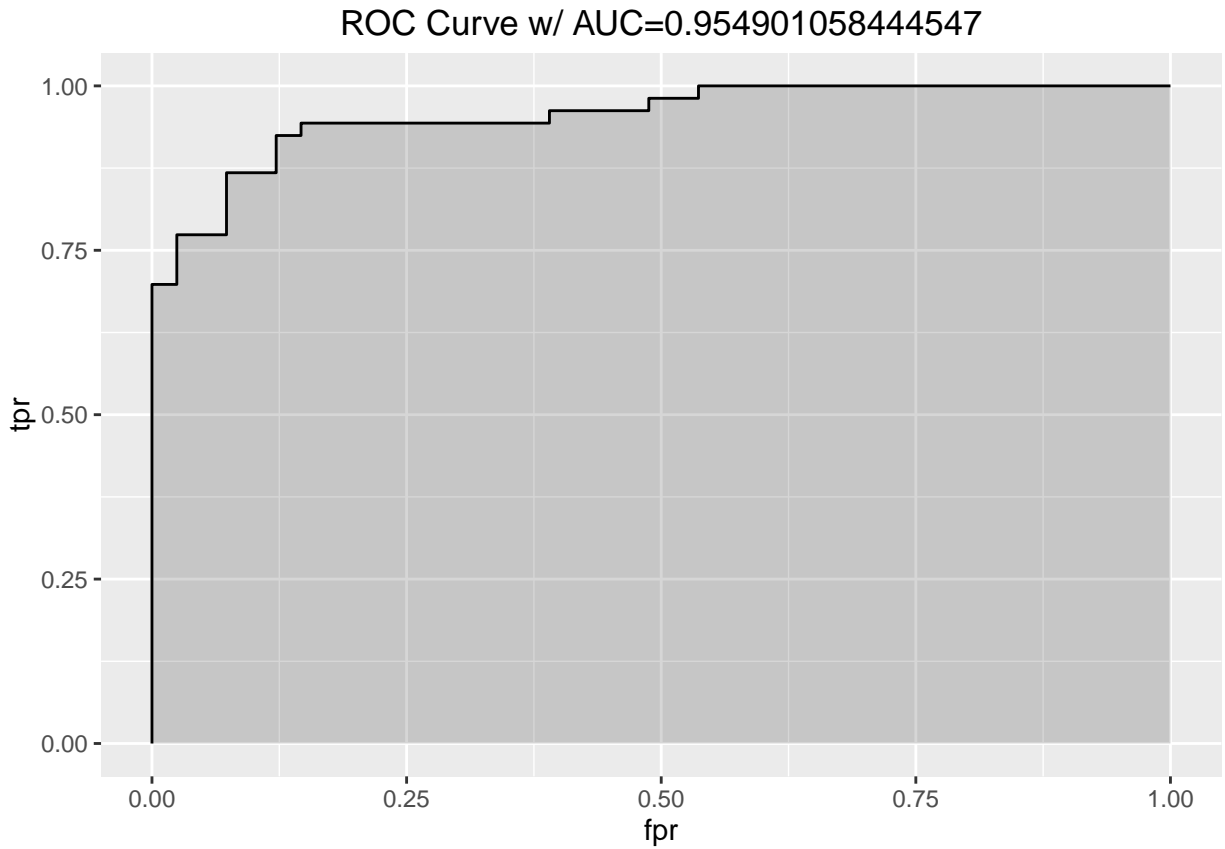
4.1 Model selection strategy:

Following model selection strategy has been used for this assignment:

- (1) Compare accuracy of the models & confusion matrix
- (2) Compare Precision,Sensitivity,Specificity,F1 score
- (3) Compare AUC curve for the models

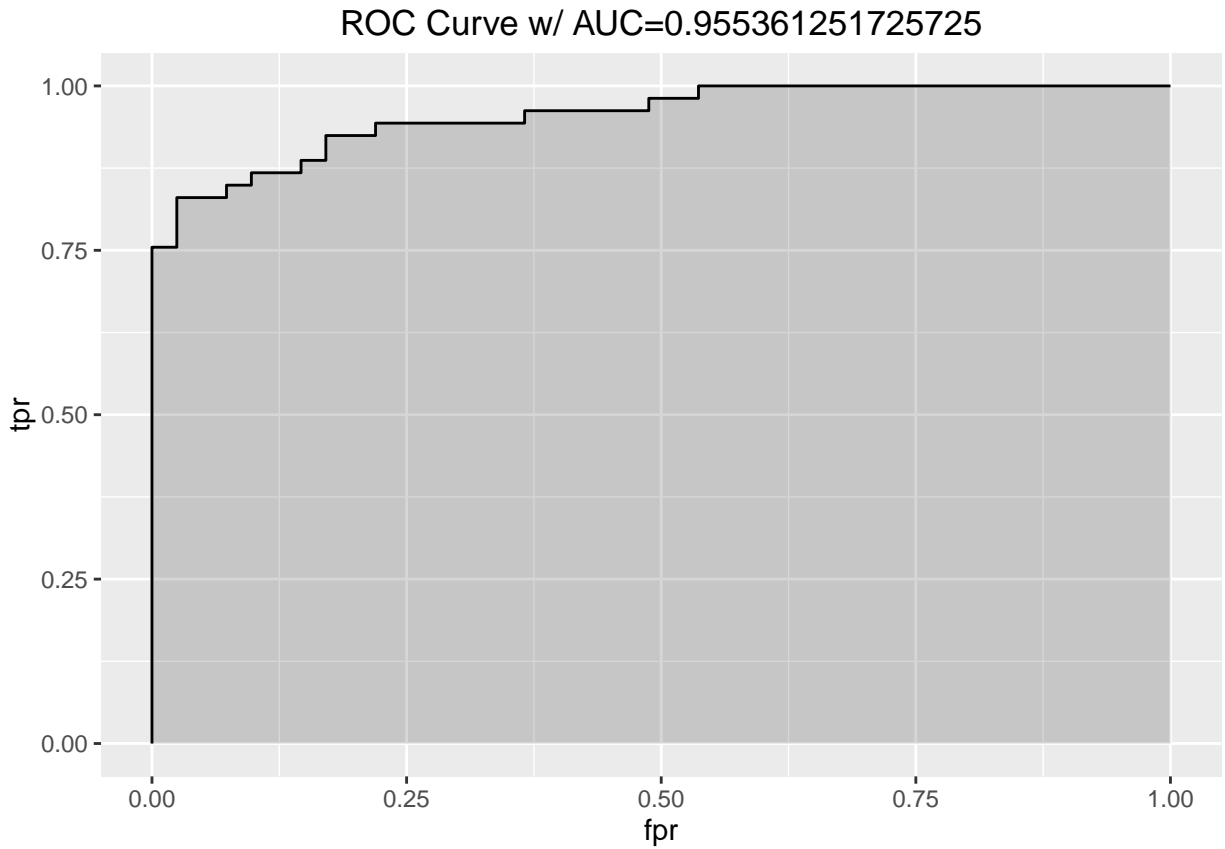
4.1.1 Model1 Evaluation

```
##      Accuracy Error_Rate Precision sensitivity specificity F1_Score
## 1 0.5744681  0.4255319 0.9245283   0.5764706   0.5555556 0.9416097
```



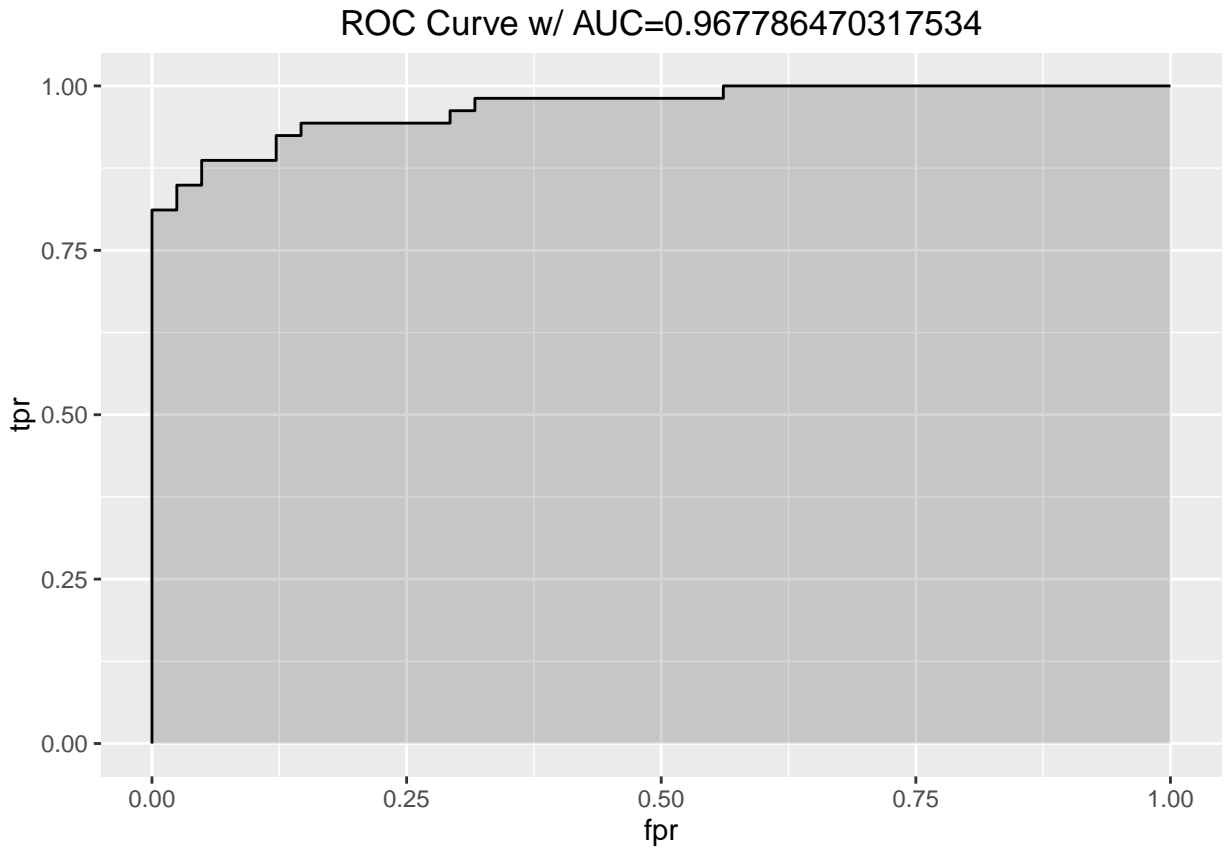
4.1.2 Model2 Evaluation

```
##      Accuracy Error_Rate Precision sensitivity specificity F1_Score
## 1 0.5851064  0.4148936 0.9056604   0.5853659   0.5833333 0.9072354
```



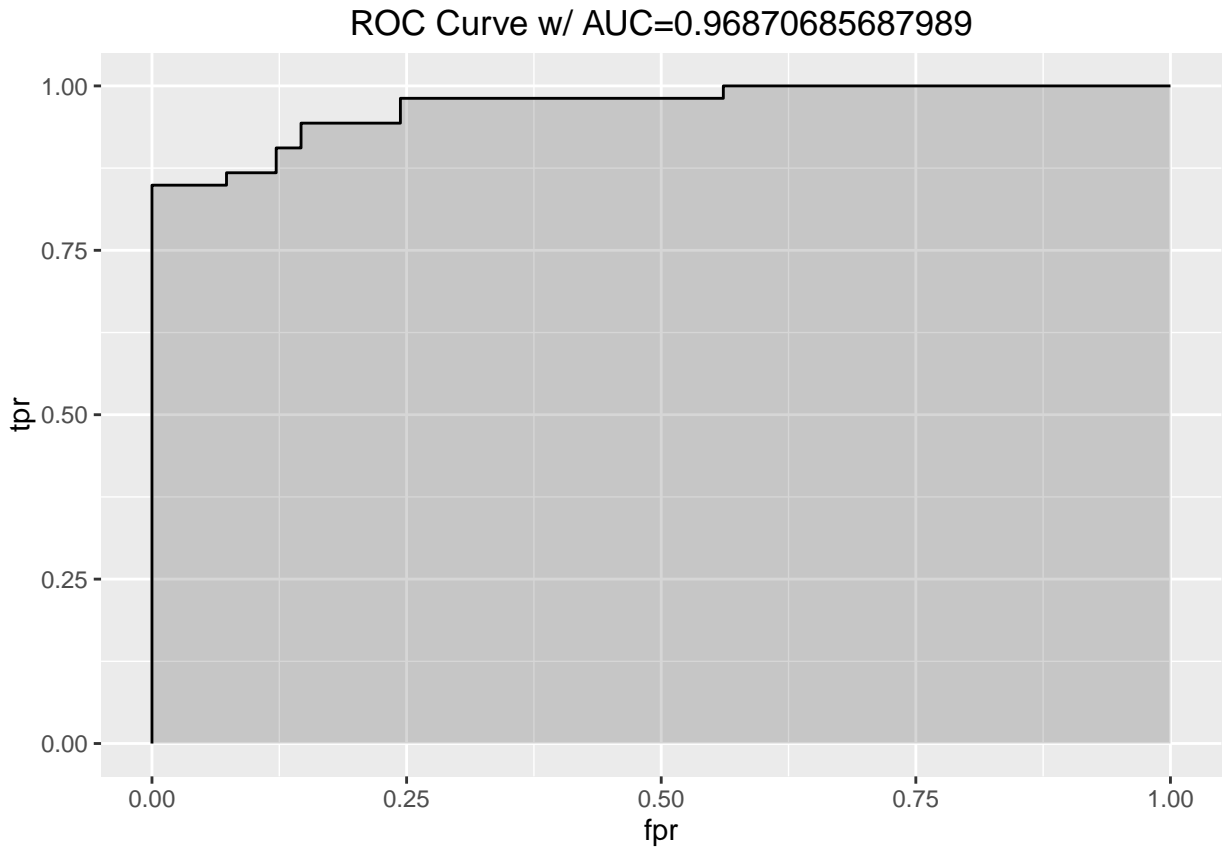
4.1.3 Model3 Evaluation

##	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score
## 1	0.5851064	0.4148936	0.9245283	0.5833333	0.6	0.9115068



4.1.4 Model4 Evaluation

##	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score
## 1	0.5744681	0.4255319	0.9056604	0.5783133	0.5454545	0.9321417



4.1.5 Model5 Evaluation

```
##      Accuracy Error_Rate Precision sensitivity specificity F1_Score
## 1 0.8297872  0.1702128 0.7358491   0.9512195   0.7358491 0.8297872
```

4.1.6 Model6 Evaluation

```
##      Accuracy Error_Rate Precision sensitivity specificity F1_Score
## 1 0.8297872  0.1702128 0.7358491   0.9512195   0.7358491 0.8297872
```