

# Home Work Assignment - 03

*Critical Thinking Group 5*

## Contents

<b>Overview</b>	<b>2</b>
<b>1 Data Exploration Analysis</b>	<b>2</b>
1.1 Variable identification . . . . .	2
1.2 Data Summary Analysis . . . . .	8
1.3 Outliers and Missing Values Identification . . . . .	8
<b>2. Data Preparation</b>	<b>10</b>
2.1 Outliers treatment and transformation . . . . .	10
<b>3 Build Models</b>	<b>13</b>
3.1.1 Model One by using all given variable . . . . .	13
3.1.3 Model three- model with transformed variables . . . . .	17
<b>4 Model Selection</b>	<b>22</b>
4.1 Model selection strategy: . . . . .	22
4.1.1 Model1 Evaluation . . . . .	22
4.1.2 Model2 Evaluation . . . . .	23
4.1.3 Model3 Evaluation . . . . .	24
4.1.4 Model4 Evaluation . . . . .	25
4.1.5 Model5 Evaluation . . . . .	26
4.1.6 Model6 Evaluation . . . . .	27
4.2 Final Model Seletion . . . . .	28

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

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.

###Following is the summary of the train data set

```
##          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
```

###Following is the analysis on missing values in data set

Table: 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

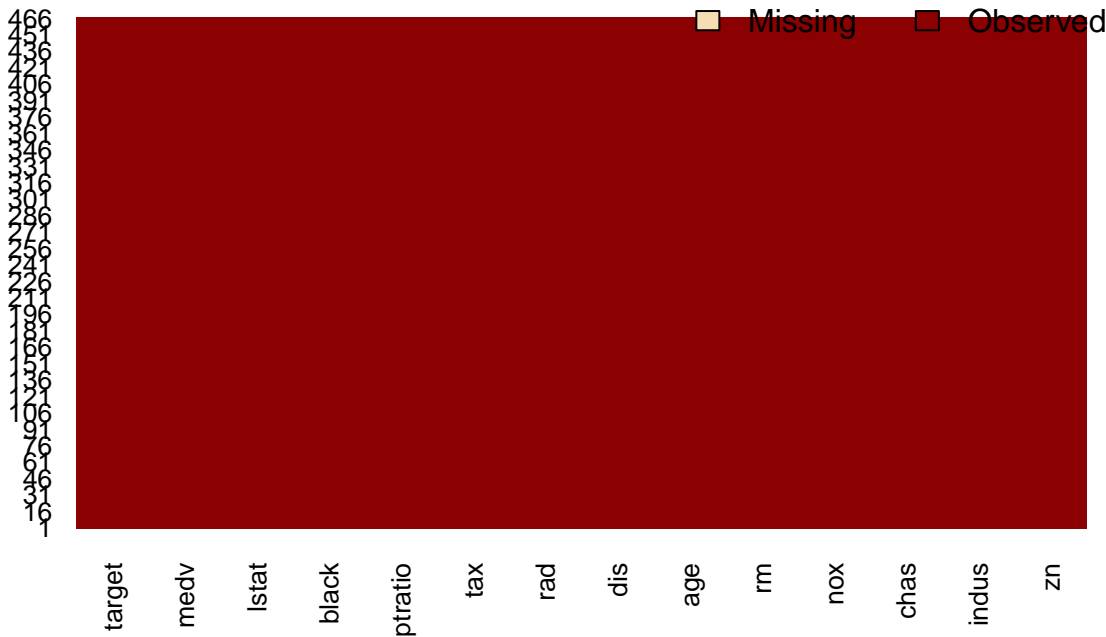


Table 3: Unique Values

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

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.

### % break up of target variable in train data set

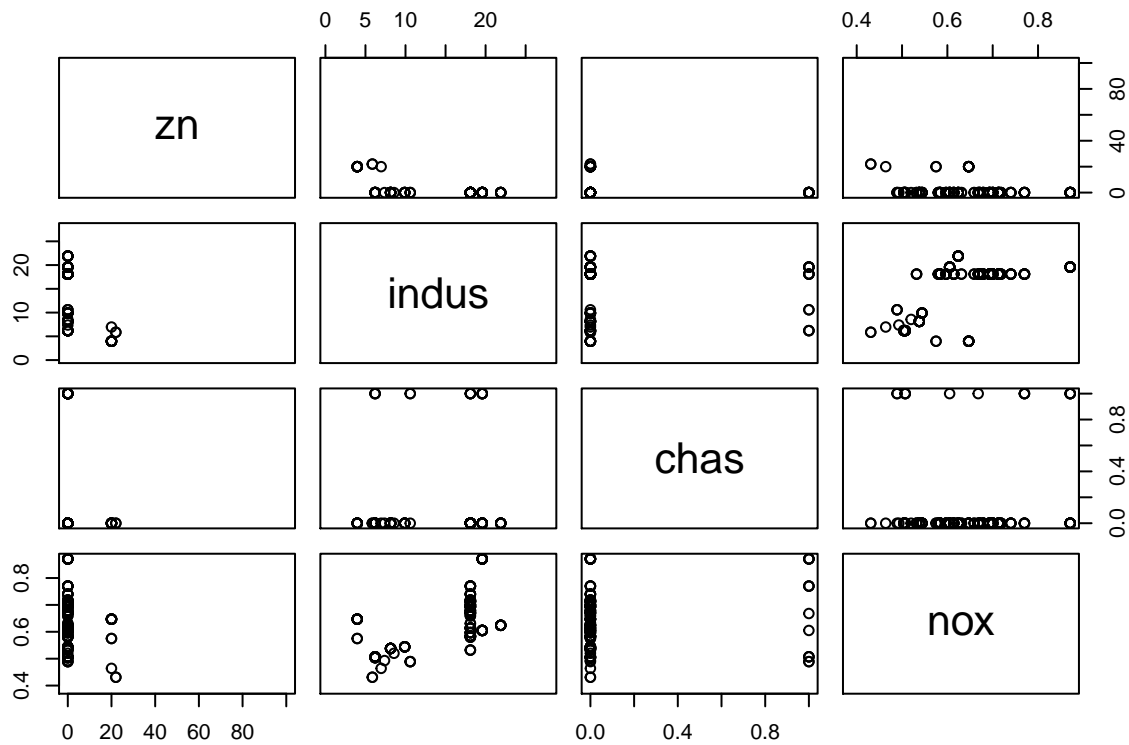
```
##
##      0      1
```

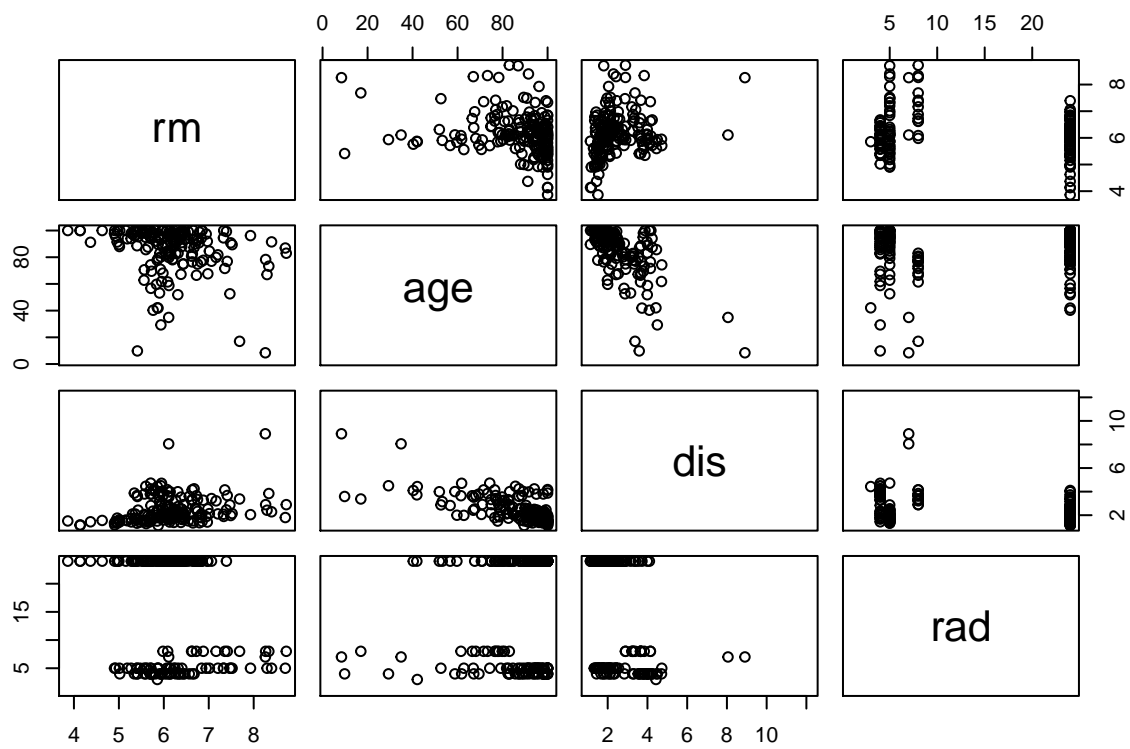
## 0.5085837 0.4914163

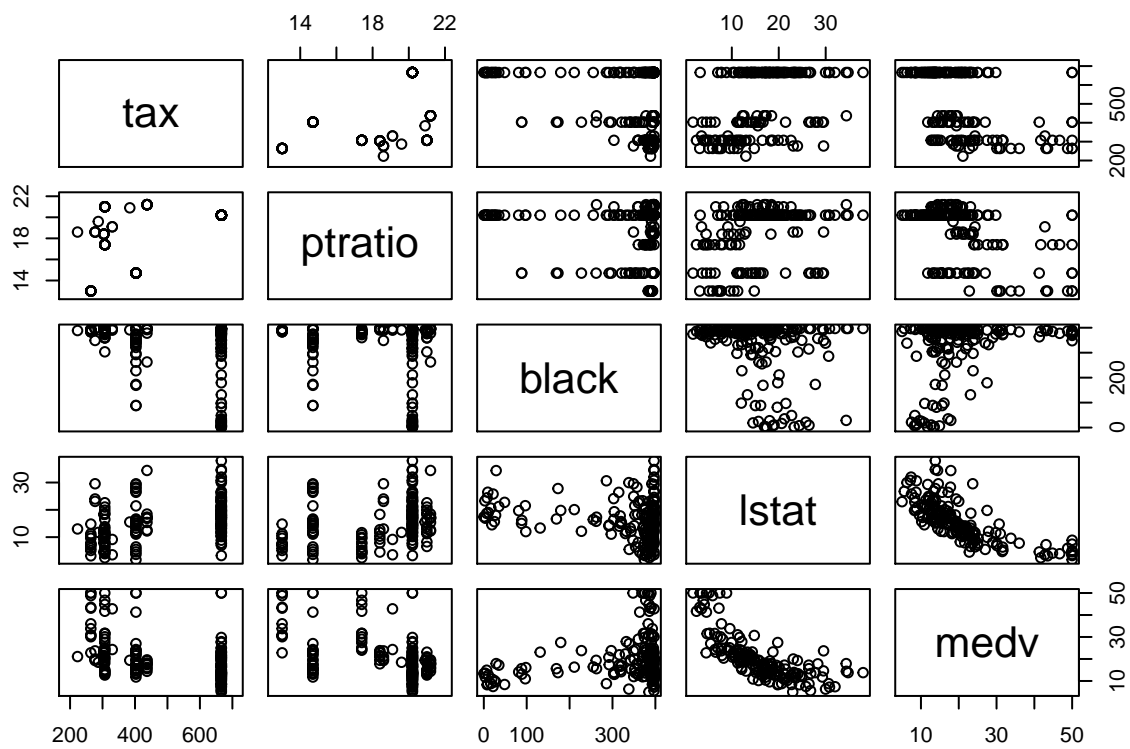
% split of target variable is given above table which shows data is almost evenly split between binary outcome 0 and 1.

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.

##	vars	n	mean	sd	median	trimmed	mad	min	max	range
## zn	1	372	12.36	24.06	0.00	6.04	0.00	0.00	100.00	100.00
## indus	2	372	10.90	6.90	8.56	10.66	7.90	0.46	27.74	27.28
## chas	3	372	0.06	0.25	0.00	0.00	0.00	0.00	1.00	1.00
## nox	4	372	0.55	0.12	0.52	0.54	0.12	0.39	0.87	0.48
## rm	5	372	6.30	0.70	6.21	6.27	0.53	3.86	8.72	4.86
## age	6	372	67.41	28.69	76.50	69.83	30.91	2.90	100.00	97.10
## dis	7	372	3.84	2.13	3.32	3.60	2.05	1.13	12.13	11.00
## rad	8	372	9.20	8.54	5.00	8.28	1.48	1.00	24.00	23.00
## tax	9	372	403.69	167.05	330.00	394.00	108.23	187.00	711.00	524.00
## ptratio	10	372	18.23	2.22	18.60	18.41	2.37	12.60	22.00	9.40
## black	11	372	359.63	88.60	391.96	384.77	7.33	0.32	396.90	396.58
## lstat	12	372	12.40	7.03	10.93	11.62	6.77	1.73	37.97	36.24
## medv	13	372	22.85	9.07	21.60	21.98	6.97	5.00	50.00	45.00
## target	14	372	0.47	0.50	0.00	0.47	0.00	0.00	1.00	1.00
##	se									
## zn	1.25									
## indus	0.36									
## chas	0.01									
## nox	0.01									
## rm	0.04									
## age	1.49									
## dis	0.11									
## rad	0.44									
## tax	8.66									
## ptratio	0.12									
## black	4.59									
## lstat	0.36									
## medv	0.47									
## target	0.03									

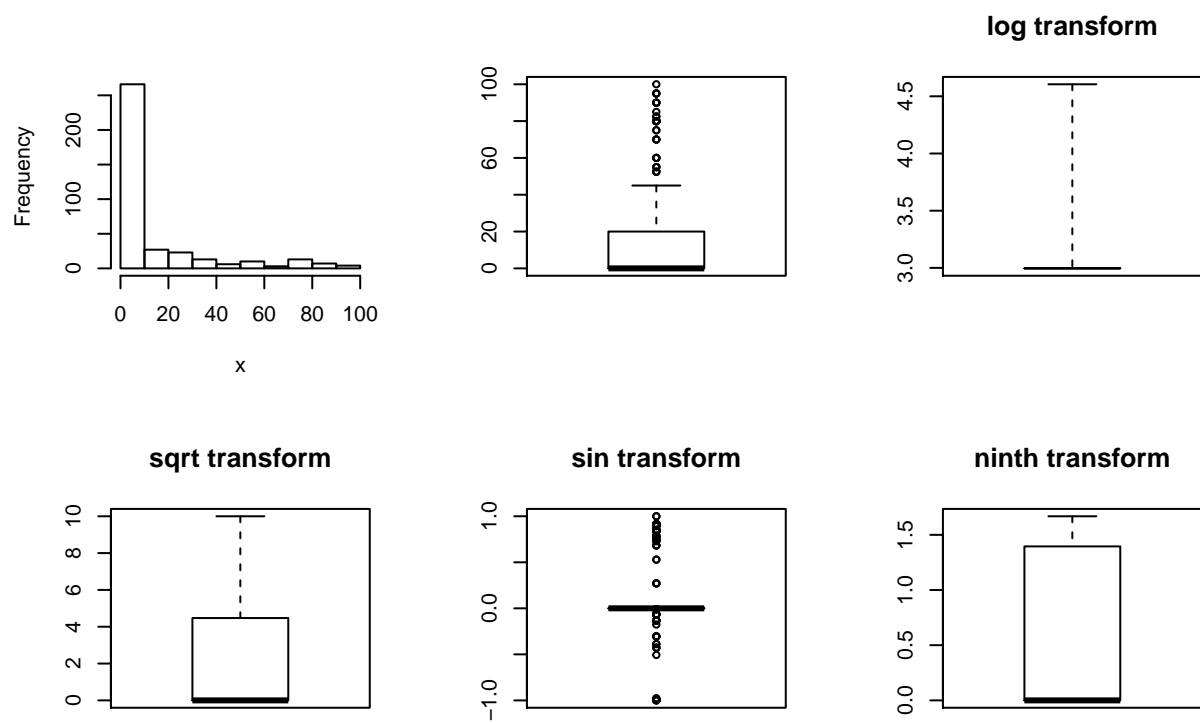
It is clear from the table that most of the variables are having strong correlation with the target variable.

## 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.

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 like 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

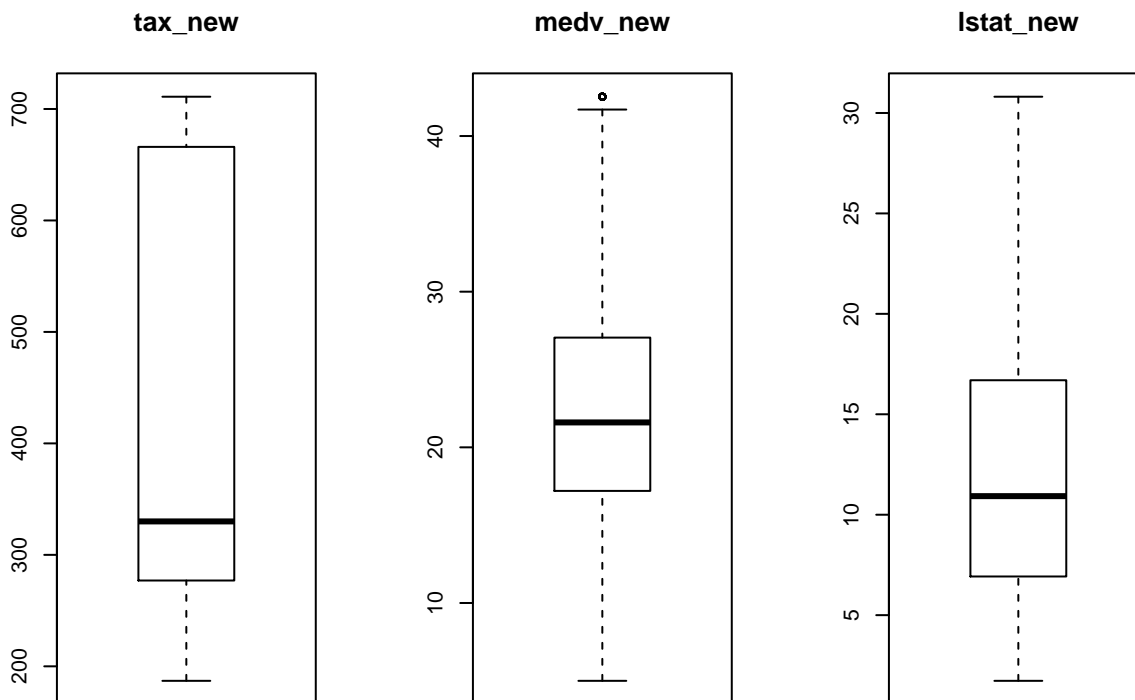
For outliers, we will create 2 sets of variables.

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$tax_new = city_crime_train$tax
city_crime_train$medv_new = city_crime_train$medv
city_crime_train$lstat_new = city_crime_train$lstat
```

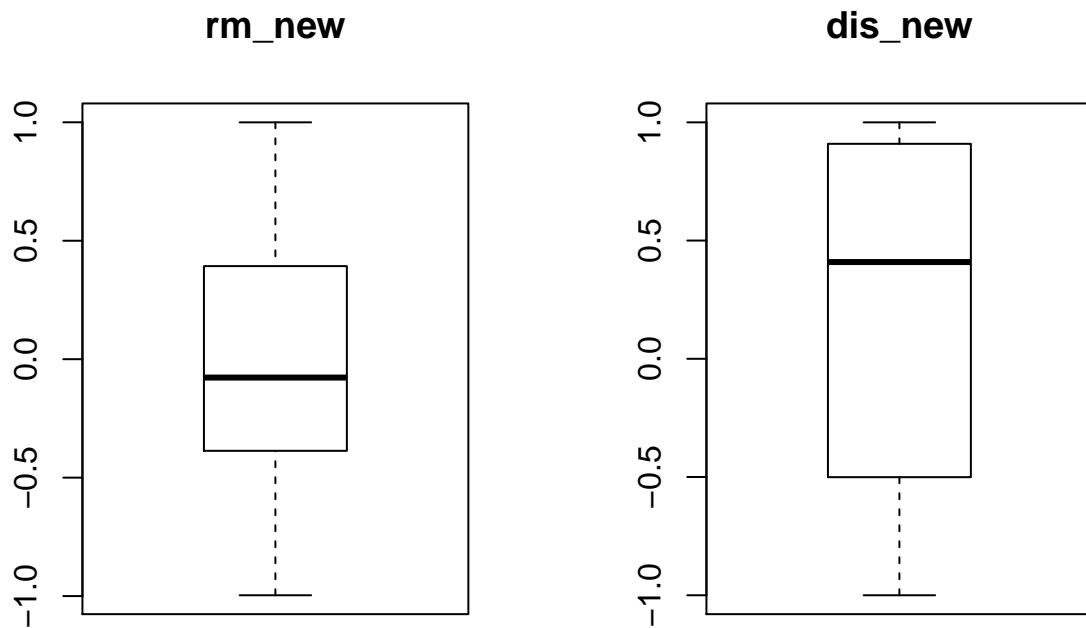
Below boxplots shows distribution of variables after outliers treatment.



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

```
city_crime_train_mod$rm_new = city_crime_train$rm
city_crime_train_mod$dis_new = city_crime_train$dis
```

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

Following strategy has been adopted to build models for this scenario:

- (i) Building model 1 using given variables by using logit function.
- (ii) Using step function to enhance model 1 and create model 2.
- (iii) Building model 3 using tranformed variables also by using logit function.
- (iv) Using step function to enhance model 3 and create model 4.
- (v) Using Linear discrement analysis model create model 5 with given variables.
- (vi) Using Linear discrement analysis model create model 6 with transformed variables.

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

### 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
```

### 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 model1 =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 incremases by 53.41.
  - b. For a one unit increase in dis, the log odds of crime rate above median value incremases by 0.80.
  - c. For a one unit increase in rad, the log odds of crime rate above median value incremases by 0.72.
  - d. For a one unit increase in tax, the log odds of crime rate above median value incremases by -0.007.
  - e. For a one unit increase in ptratio, the log odds of crime rate above median value incremases by 0.44.
  - f. For a one unit increase in medv , the log odds of crime rate above median value incremases by 0.23.

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

```
## 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> 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> 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> 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

##
## 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 incremases by 46.61.

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



- c. For a one unit increase in rad, the log odds of crime rate above median value incremases by 0.77.
- d. For a one unit increase in tax, the log odds of crime rate above median value incremases by -0.009.
- e. For a one unit increase in ptratio, the log odds of crime rate above median value incremases by 0.35.
- f. For a one unit increase in medv , the log odds of crime rate above median value incremases by 0.18

(iv) there were 9 ierations 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
```

- (i) From this model it can be seen following variables are relevant for this model-`nox`, `dis`, `rad`, `ptratio`, `tax_new`, `medv_new`, `lstat_new`
- (ii) number of integration is 9 and AIC value =169.71.

**Notes:** Similar explanation of model coefficient will be applicable here as described for model 1 & 2 and is not repeated here.

### 3.1.4 Model with transformed variable and with backward step function

```
## 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
##
##
## Call:
## glm(formula = target ~ nox + age + dis + rad + ptratio + black +
##      tax_new + medv_new + rm_new + dis_new, family = "binomial",
##      data = city_crime_train_mod)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0158  -0.1472  -0.0031   0.0005   3.1030
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -52.779764   9.739144  -5.419 5.98e-08 ***
## nox          56.509319   9.188179   6.150 7.74e-10 ***
## age           0.051467   0.016215   3.174 0.001503 **
## dis           0.564992   0.255943   2.207 0.027280 *
## rad           0.849127   0.212643   3.993 6.52e-05 ***
## ptratio       0.533319   0.159365   3.347 0.000818 ***
## black        -0.010960   0.005943  -1.844 0.065147 .
## tax_new       -0.004534   0.003144  -1.442 0.149355
## medv_new       0.342778   0.095427   3.592 0.000328 ***
## rm_new        -2.358513   1.028472  -2.293 0.021835 *
## dis_new       -1.865533   0.488896  -3.816 0.000136 ***
## ---
## 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: 126.80  on 361  degrees of freedom
## AIC: 148.8
##
## Number of Fisher Scoring iterations: 9

```

- (i) From this model it can be seen following variables are relevant for this model-nox, dis, rad, ptratio , tax\_new, medv\_new, lstat\_new  
(ii) number of integration is 9 and AIC value =165.8.

### 3.1,5 Model three with Linear discrement analysis

```
## Call:
## lda(target ~ ., data = city_crime_train)
##
## Prior probabilities of groups:
##      0      1
## 0.5268817 0.4731183
##
## Group means:
##      zn      indus      chas      nox      rm      age      dis
## 0 22.012755  6.956327 0.05102041 0.4689730 6.401296 50.37398 5.086538
## 1  1.613636 15.291193 0.07954545 0.6428523 6.176631 86.38864 2.459868
##      rad      tax ptratio      black      lstat      medv
## 0  4.107143 308.4949 17.76990 388.6647  9.199235 25.18724
## 1 14.880682 509.6932 18.74773 327.2894 15.959148 20.24148
##
## Coefficients of linear discriminants:
##      LD1
## zn      -0.0047914631
## indus    0.0281044279
## chas     -0.0556293189
## nox       7.9109306913
## rm        0.1658180998
## age       0.0131973114
## dis       0.0840623852
## rad       0.1027832012
## tax      -0.0019152605
## ptratio   0.0090391049
## black    -0.0009160458
## lstat     0.0248449648
## medv      0.0425514709
```

### 3.1.6 Model with Linear discrement analysis with transformed data

```
## Call:
## lda(target ~ . - zn - rm - dis - tax - lstat - medv, data = city_crime_train_mod)
##
## Prior probabilities of groups:
##      0      1
## 0.5268817 0.4731183
##
## Group means:
##      indus      chas1      nox      age      rad ptratio      black
## 0  6.956327 0.05102041 0.4689730 50.37398  4.107143 17.76990 388.6647
## 1 15.291193 0.07954545 0.6428523 86.38864 14.880682 18.74773 327.2894
##      tax_new medv_new lstat_new      rm_new      dis_new      zn_new
## 0 308.4949 25.04528  9.199235  0.08333182 -0.0504096 0.46938776
## 1 509.6932 19.86151 15.724247 -0.11166891  0.5106930 0.07954545
```

```
##
## Coefficients of linear discriminants:
##          LD1
## indus      0.022452946
## chas1     -0.186416323
## nox        7.970446650
## age        0.015169354
## rad        0.100159450
## ptratio   -0.014404341
## black     -0.001159202
## tax_new   -0.001196341
## medv_new   0.047596449
## lstat_new  0.016840318
## rm_new    -0.008946209
## dis_new   -0.340985994
## zn_new    -0.001832533
```

## 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:

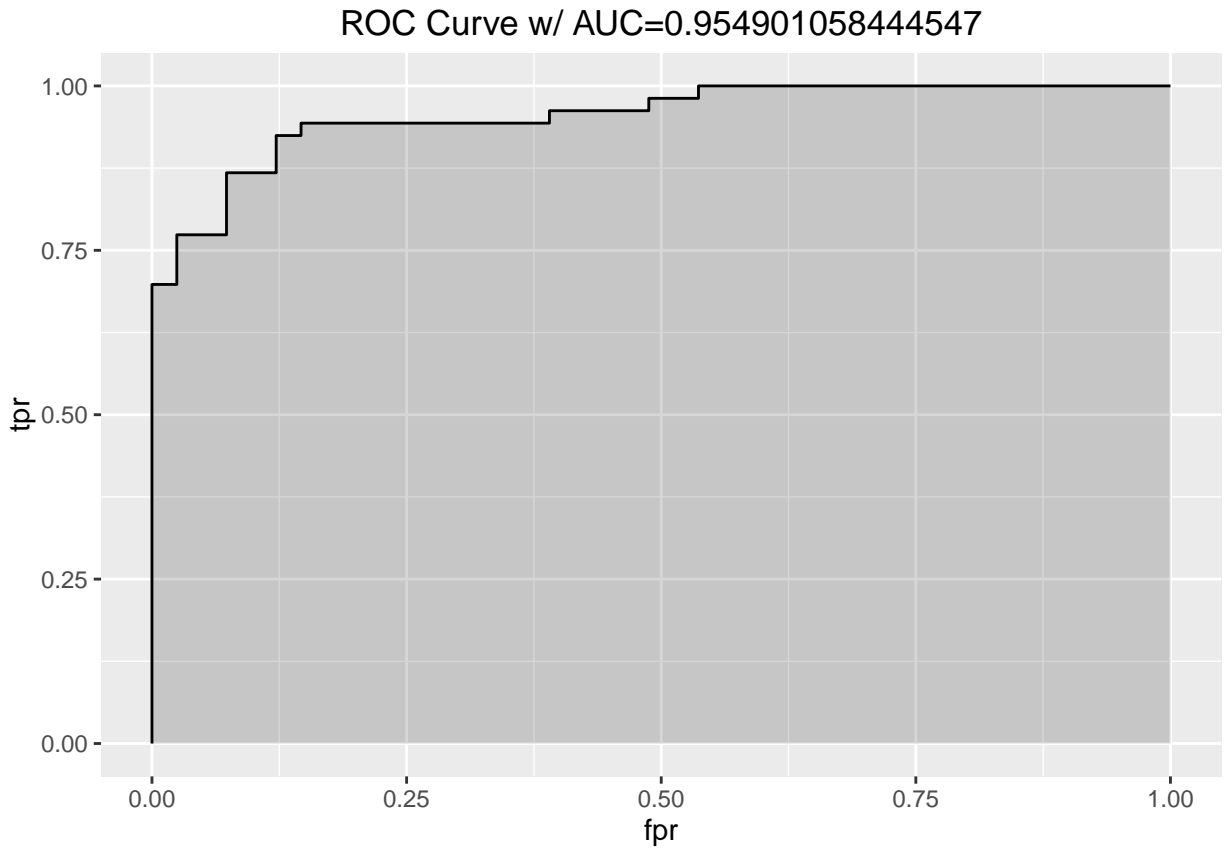
Following model selection strategy has been used for this assignment:

- (i) Compare accuracy of the models & confusion matrix
- (ii) Compare Precision,Sensitivity,Specificity,F1 score
- (iii) Compare AUC curve for the models

Following function Eval() will be used to calculate various metrics related to the model like Accuracy, Sensitivity, Precision , Specificity and F1 score

#### 4.1.1 Model1 Evaluation

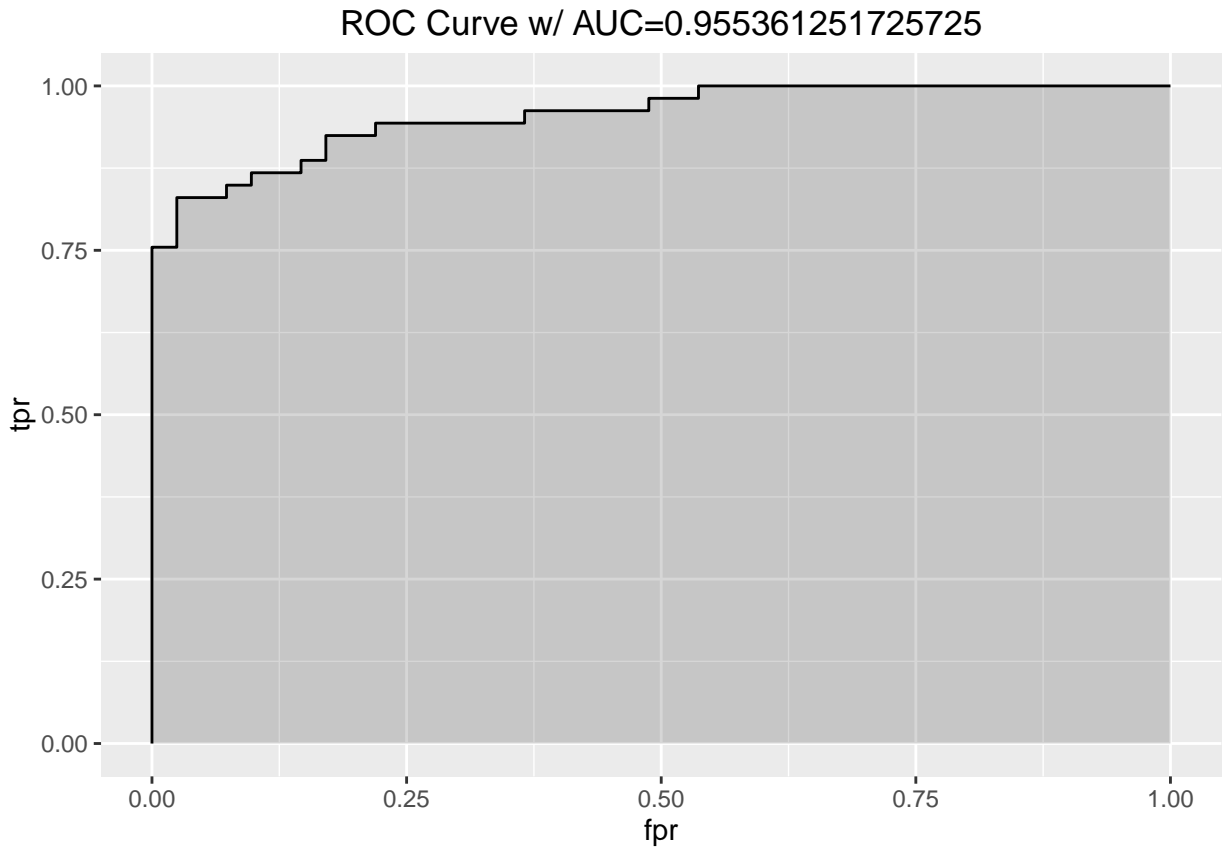
```
##      Accuracy Error_Rate Precision sensitivity specificity F1_Score
## 1 0.9042553 0.09574468 0.9245283  0.9074074          0.9 0.9283174
```



Looking at the key metrics this can be concluded this model has high accuracy 0.9042553 and low error rate 0.09574468. AUC curve for this model is 0.9549 which is very good.

#### 4.1.2 Model2 Evaluation

##	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score
## 1	0.8723404	0.1276596	0.9056604	0.8727273	0.8717949	0.9061444

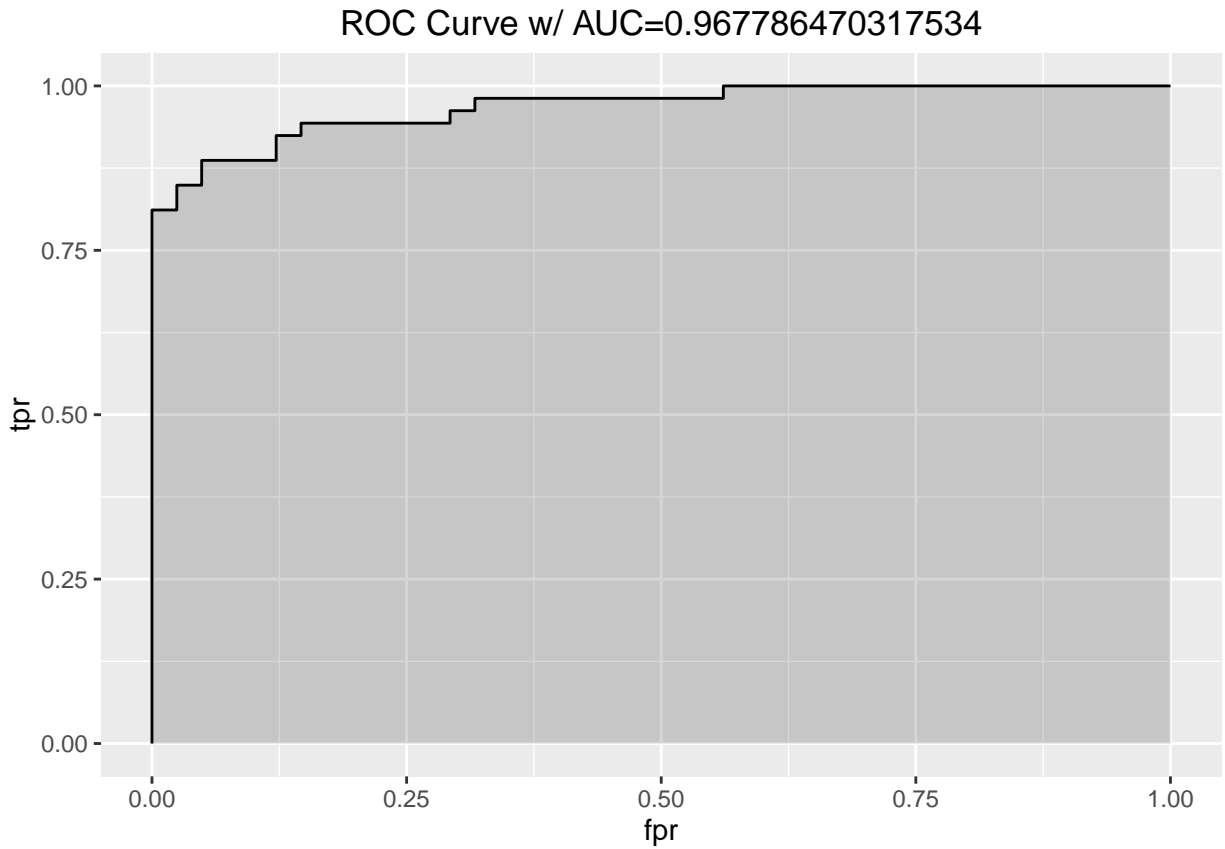


Looking at the key metrics this can be concluded this model has high accuracy 0.8723404 and low error rate 0.12765957. AUC curve for this model is 0.9553 which is very good.

#### 4.1.3 Model3 Evaluation

```
## Accuracy Error_Rate Precision sensitivity specificity F1_Score
## 1 0.893617 0.106383 0.9245283 0.8909091 0.8974359 0.9211541
```

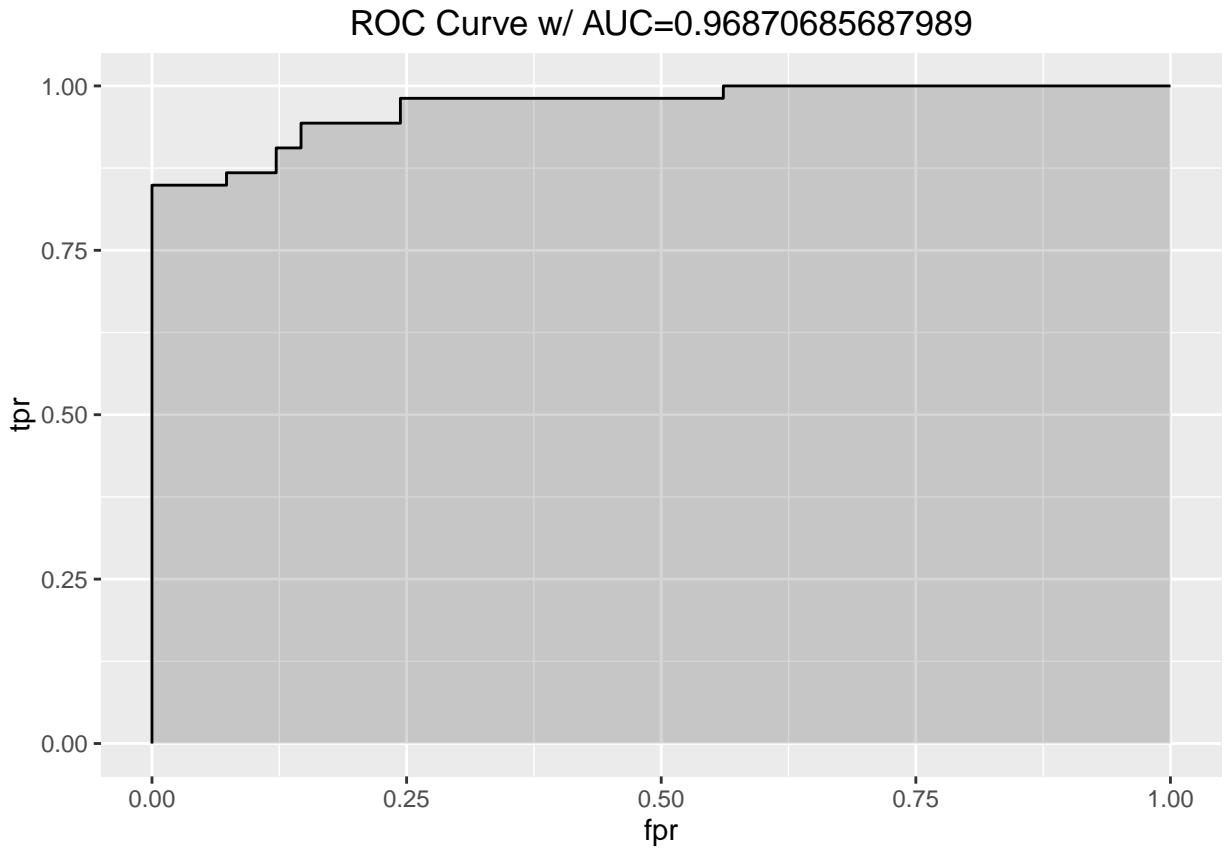




Looking at the key metrics this can be concluded this model has high accuracy 0.8936170 and low error rate 0.10638298. AUC curve for this model is 0.9558 which is very good.

#### 4.1.4 Model4 Evaluation

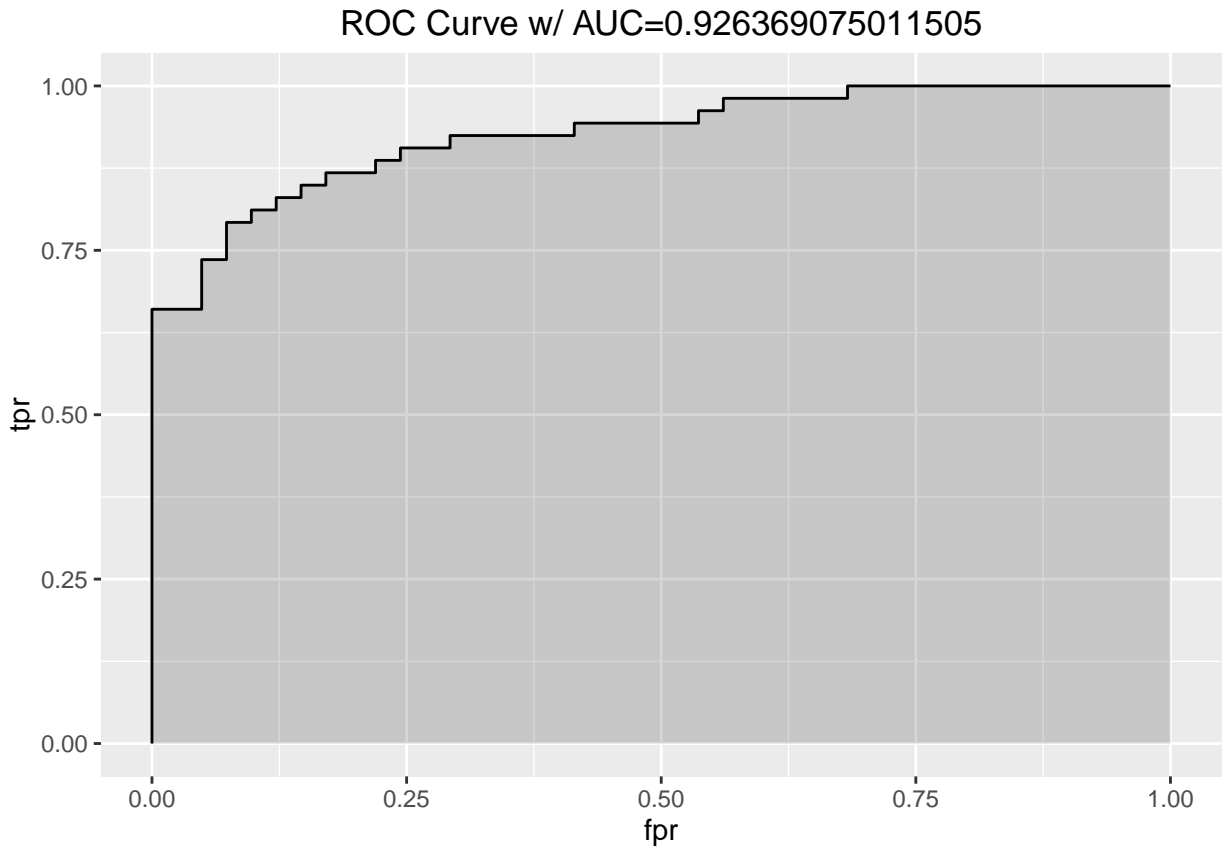
##	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score
## 1	0.8829787	0.1170213	0.9056604	0.8888889	0.875	0.9127916



Looking at the key metrics this can be concluded this model has high accuracy 0.8829787 and low error rate 0.11702128. AUC curve for this model is 0.9549 which is very good.

#### 4.1.5 Model5 Evaluation

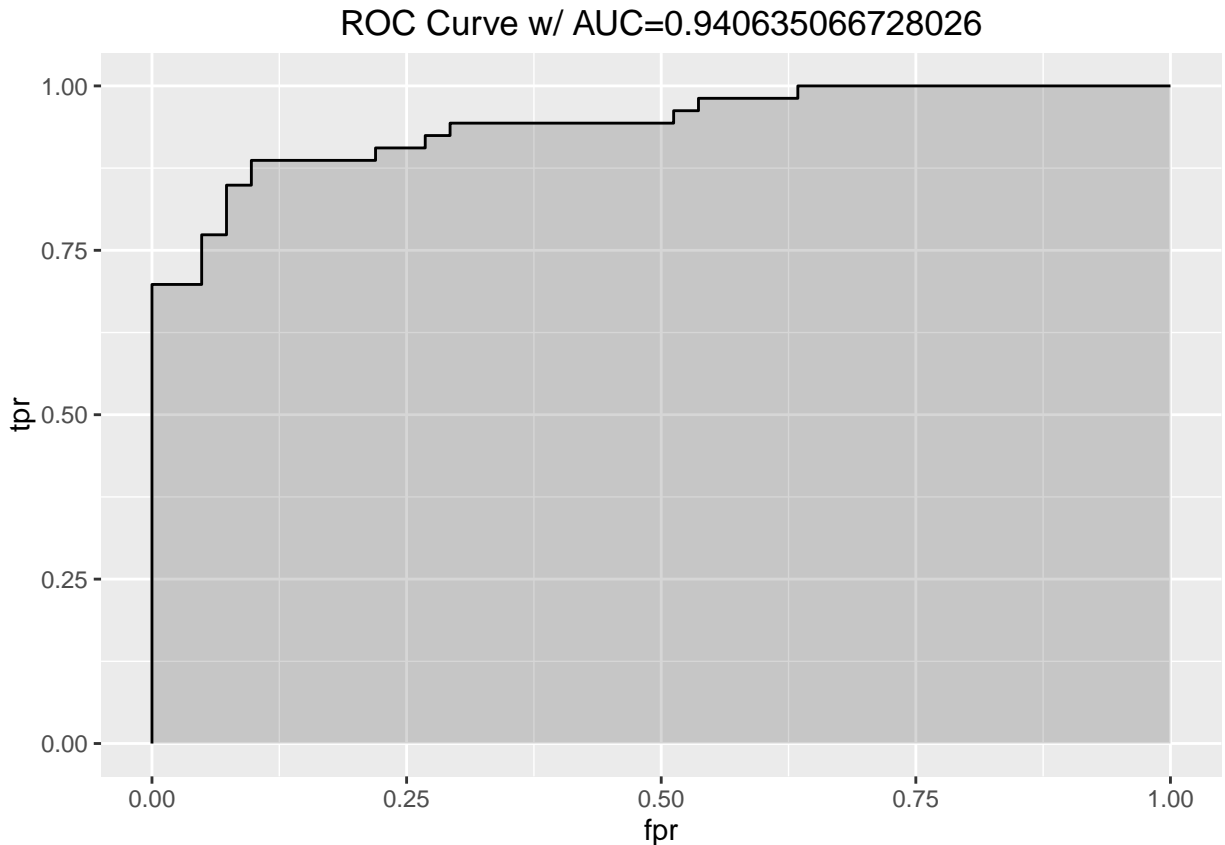
##	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score
## 1	0.8297872	0.1702128	0.7358491	0.9512195	0.7358491	0.8297872



Looking at the key metrics this can be concluded this model has relatively low accuracy 0.8297872 and higher error rate 0.1702127 compared to other models. AUC curve for this model is 0.9263.

#### 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



Looking at the key metrics this can be concluded this model has relatively low accuracy 0.8297872 and higher error rate 0.17021277 compared to other models. AUC curve for this model is 0.930.

## 4.2 Final Model Seletion

Following is the comparison of various metrics for above 6 models

##	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score
## 1	0.9042553	0.09574468	0.9245283	0.9074074	0.9000000	0.9283174
## 2	0.8723404	0.12765957	0.9056604	0.8727273	0.8717949	0.9061444
## 3	0.8936170	0.10638298	0.9245283	0.8909091	0.8974359	0.9211541
## 4	0.8829787	0.11702128	0.9056604	0.8888889	0.8750000	0.9127916
## 5	0.8297872	0.17021277	0.7358491	0.9512195	0.7358491	0.8297872
## 6	0.8297872	0.17021277	0.7358491	0.9512195	0.7358491	0.8297872

From the comparison table it can be concluded model 1 is the best model with very high accuracy rate of 90.42%. Further analysis has been carried out on this model below-

(i) Estimate confidence interval for coefficient (ii) wald test to understand effect of variable in the model (iii) odds ratios and 95% CI

##		2.5 %	97.5 %
## (Intercept)		-57.633422451	-2.529088e+01
## zn		-0.137318897	1.615847e-02
## indus		-0.180179675	5.240896e-02

```
## chas      -0.907581097  2.486362e+00
## nox       33.787078022  7.303993e+01
## rm        -2.420592622  1.124709e+00
## age       -0.001896677  5.956762e-02
## dis       0.273928524  1.327906e+00
## rad       0.338261333  1.105240e+00
## tax       -0.013905482 -2.242516e-04
## ptratio   0.128416628  7.531198e-01
## black     -0.021400089  2.217745e-03
## lstat     -0.025418331  2.193000e-01
## medv      0.058041222  4.158385e-01
```

```
## Wald test:
```

```
## -----
```

```
##
```

```
## Chi-squared test:
```

```
## X2 = 13.6, df = 1, P(> X2) = 0.00023
```

```
##              OR          2.5 %      97.5 %
## (Intercept) 9.844998e-19 9.335179e-26 1.038266e-11
## zn          9.412183e-01 8.716922e-01 1.016290e+00
## indus       9.381125e-01 8.351201e-01 1.053807e+00
## chas        2.202054e+00 4.034991e-01 1.201748e+01
## nox         1.574670e+23 4.715650e+14 5.258208e+31
## rm          5.231212e-01 8.886894e-02 3.079319e+00
## age         1.029255e+00 9.981051e-01 1.061378e+00
## dis         2.227583e+00 1.315121e+00 3.773133e+00
## rad         2.058033e+00 1.402507e+00 3.019950e+00
## tax         9.929600e-01 9.861908e-01 9.997758e-01
## ptratio     1.553900e+00 1.137027e+00 2.123615e+00
## black       9.904547e-01 9.788273e-01 1.002220e+00
## lstat       1.101795e+00 9.749020e-01 1.245205e+00
## medv        1.267365e+00 1.059759e+00 1.515641e+00
```

The chi-squared test statistic of 13.6 , with one degrees of freedom is associated with a p-value of 0.00023 indicating that the overall effect of tax is statistically significant.

odds ratios and 95% CI