621 Assignment3

Raghu

Oct 20, 2018

Contents

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:

Explanatory Variables:

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)

1. DATA EXPLORATION

Describe the size and the variables in the crime training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job.

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

Data View:

Lets have a quick view of crime data.

zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv	target
0	19.58	0	0.605	7.929	96.2	2.0459	5	403	14.7	369.30	3.70	50.0	1
0	19.58	1	0.871	5.403	100.0	1.3216	5	403	14.7	396.90	26.82	13.4	1
0	18.10	0	0.740	6.485	100.0	1.9784	24	666	20.2	386.73	18.85	15.4	1
30	4.93	0	0.428	6.393	7.8	7.0355	6	300	16.6	374.71	5.19	23.7	0
0	2.46	0	0.488	7.155	92.2	2.7006	3	193	17.8	394.12	4.82	37.9	0
0	8.56	0	0.520	6.781	71.3	2.8561	5	384	20.9	395.58	7.67	26.5	0

Basic Stats

There are 466 observations and 14 variables. * 10 variables of type dble. * 4 variables of type int.

```
## Observations: 466
## Variables: 14
## $ zn      <dbl> 0, 0, 0, 30, 0, 0, 0, 0, 0, 80, 22, 0, 0, 22, 0, 0, 10...
## $ indus   <dbl> 19.58, 19.58, 18.10, 4.93, 2.46, 8.56, 18.10, 18.10, 5...
## $ chas    <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ nox     <dbl> 0.605, 0.871, 0.740, 0.428, 0.488, 0.520, 0.693, 0.693...
## $ rm      <dbl> 7.929, 5.403, 6.485, 6.393, 7.155, 6.781, 5.453, 4.519...
## $ age     <dbl> 96.2, 100.0, 100.0, 7.8, 92.2, 71.3, 100.0, 100.0, 38....
## $ dis     <dbl> 2.0459, 1.3216, 1.9784, 7.0355, 2.7006, 2.8561, 1.4896...
## $ rad     <int> 5, 5, 24, 6, 3, 5, 24, 24, 5, 1, 7, 5, 24, 7, 3, 3, 5,...
## $ tax     <int> 403, 403, 666, 300, 193, 384, 666, 666, 224, 315, 330,...
## $ ptratio <dbl> 14.7, 14.7, 20.2, 16.6, 17.8, 20.9, 20.2, 20.2, 20.2, ...
## $ black   <dbl> 369.30, 396.90, 386.73, 374.71, 394.12, 395.58, 396.90...
## $ lstat   <dbl> 3.70, 26.82, 18.85, 5.19, 4.82, 7.67, 30.59, 36.98, 5....
## $ medv    <dbl> 50.0, 13.4, 15.4, 23.7, 37.9, 26.5, 5.0, 7.0, 22.2, 20...
## $ target  <int> 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, ...
```

- Summary Statistics shows none of the variables have missing values
- The mean of target is below 0.5 which means there are more observations where the crime rate is below the median.

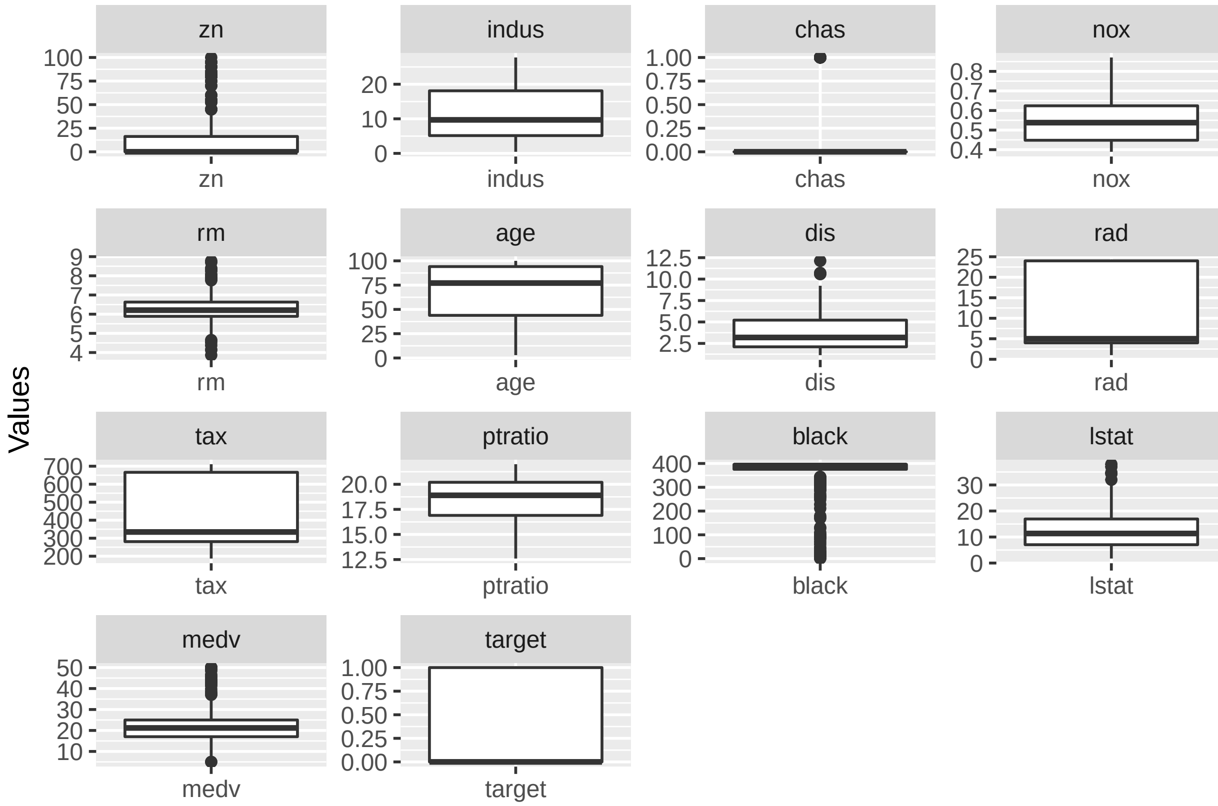
	vars	n	mean	sd	median	trimmed	mad	min	max	ra
zn	1	466	11.5772532	23.3646511	0.00000	5.3542781	0.0000000	0.0000	100.0000	100.0
indus	2	466	11.1050215	6.8458549	9.69000	10.9082353	9.3403800	0.4600	27.7400	27.2
chas	3	466	0.0708155	0.2567920	0.00000	0.0000000	0.0000000	0.0000	1.0000	1.0
nox	4	466	0.5543105	0.1166667	0.53800	0.5442684	0.1334340	0.3890	0.8710	0.4
rm	5	466	6.2906738	0.7048513	6.21000	6.2570615	0.5166861	3.8630	8.7800	4.9
age	6	466	68.3675966	28.3213784	77.15000	70.9553476	30.0226500	2.9000	100.0000	97.1
dis	7	466	3.7956929	2.1069496	3.19095	3.5443647	1.9144814	1.1296	12.1265	10.9
rad	8	466	9.5300429	8.6859272	5.00000	8.6978610	1.4826000	1.0000	24.0000	23.0
tax	9	466	409.5021459	167.9000887	334.50000	401.5080214	104.5233000	187.0000	711.0000	524.0
ptratio	10	466	18.3984979	2.1968447	18.90000	18.5970588	1.9273800	12.6000	22.0000	9.4
black	11	466	357.1201502	91.3211298	391.34000	383.5064439	8.2432560	0.3200	396.9000	396.5
lstat	12	466	12.6314592	7.1018907	11.35000	11.8809626	7.0720020	1.7300	37.9700	36.2
medv	13	466	22.5892704	9.2396814	21.20000	21.6304813	6.0045300	5.0000	50.0000	45.0
target	14	466	0.4914163	0.5004636	0.00000	0.4893048	0.0000000	0.0000	1.0000	1.0

Data Visualization

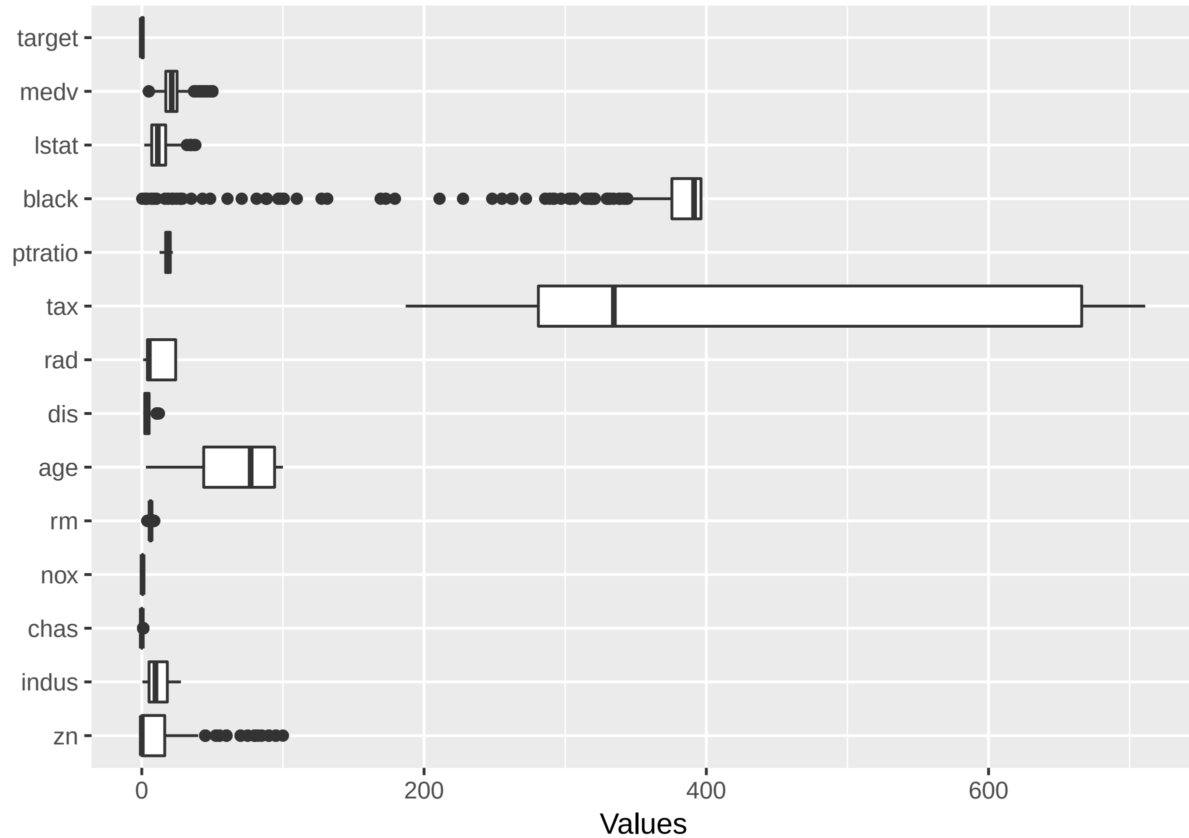
BoxPlot Distribution

Boxplot demonstrating the mean, median and quartiles of the independent variables. rad, tax and black has high variances.

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



BoxPlot of all variables

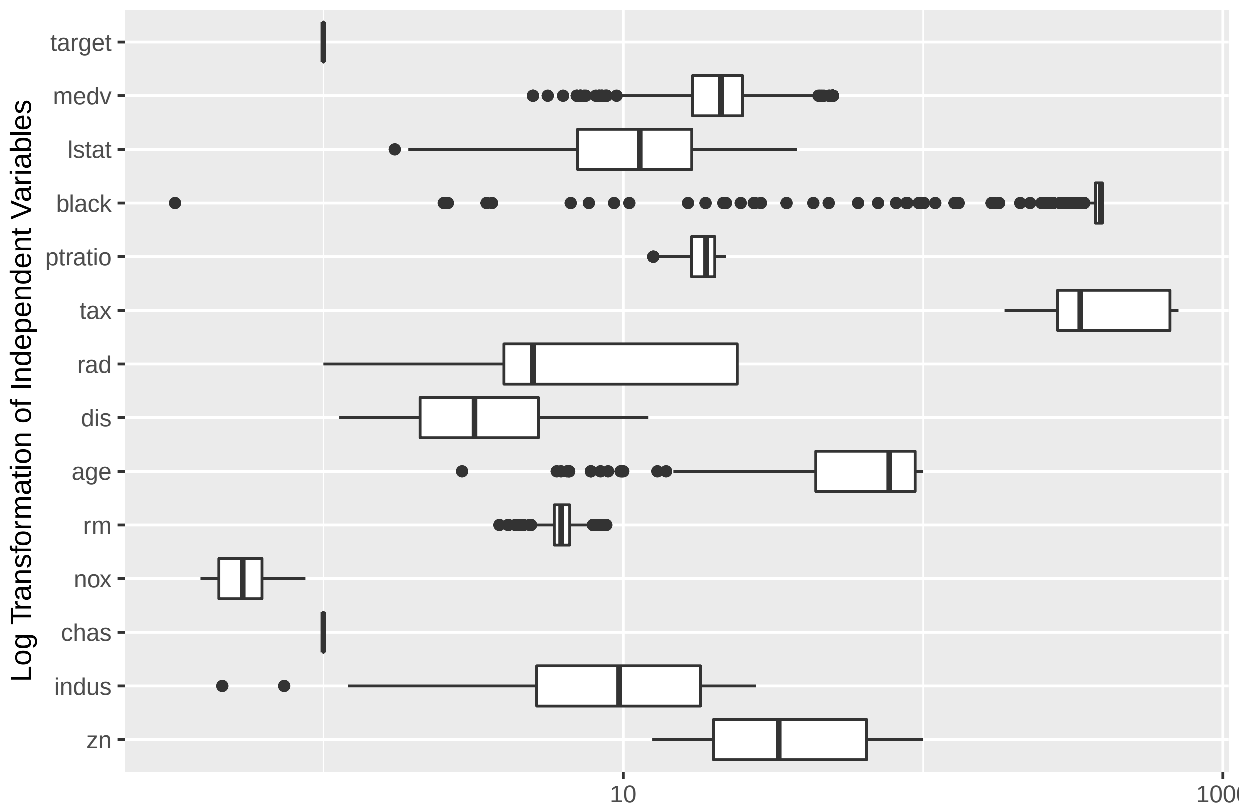


The box plots displays that many of the variables have low variances.

Lets look at the log scale of the independent variables.

BoxPlot with Log Scale

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 1009 rows containing non-finite values (stat_boxplot).
```



Density Plots

Lets look at the Density Plots for skewness.

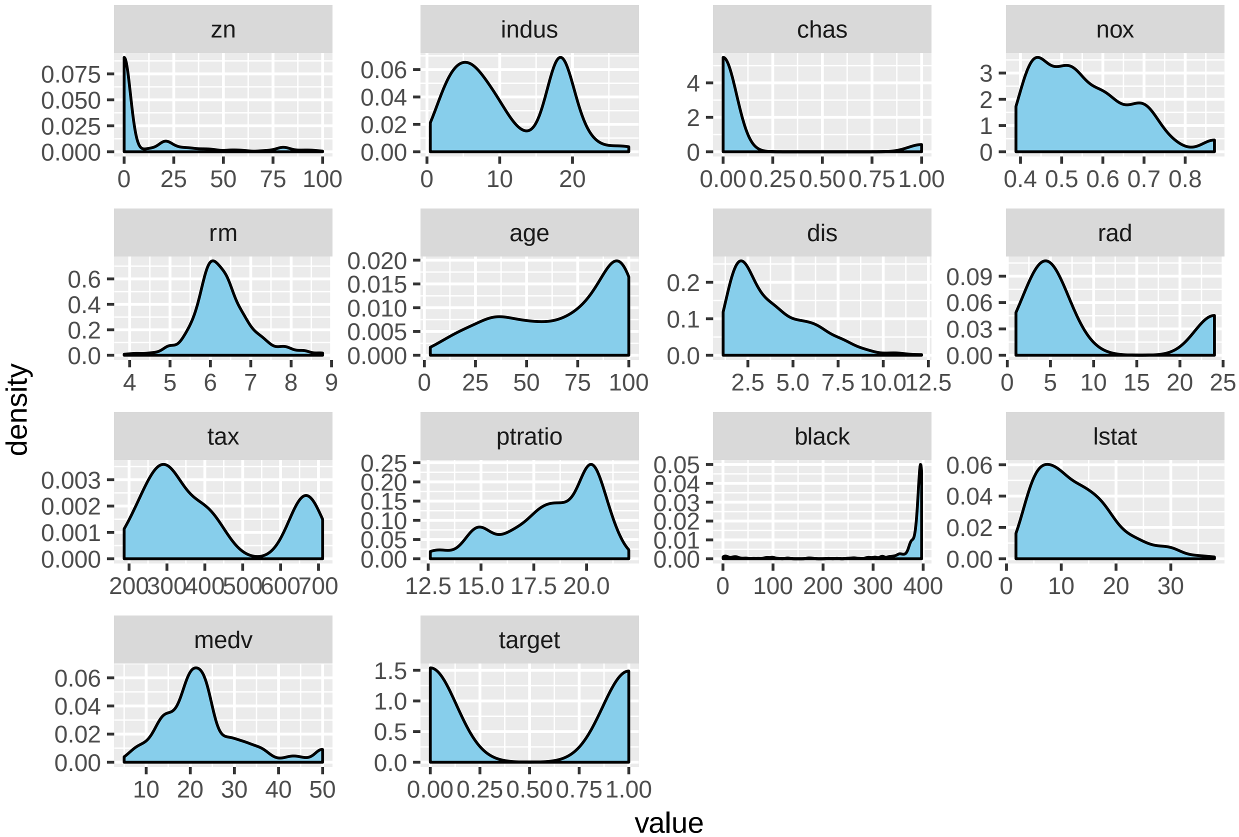
–rm is the only variable that closely mirrors a normal distribution.

–zn, chas, and dis are heavily skewed right.

–nox, lstat, and medv are are also skewed right.

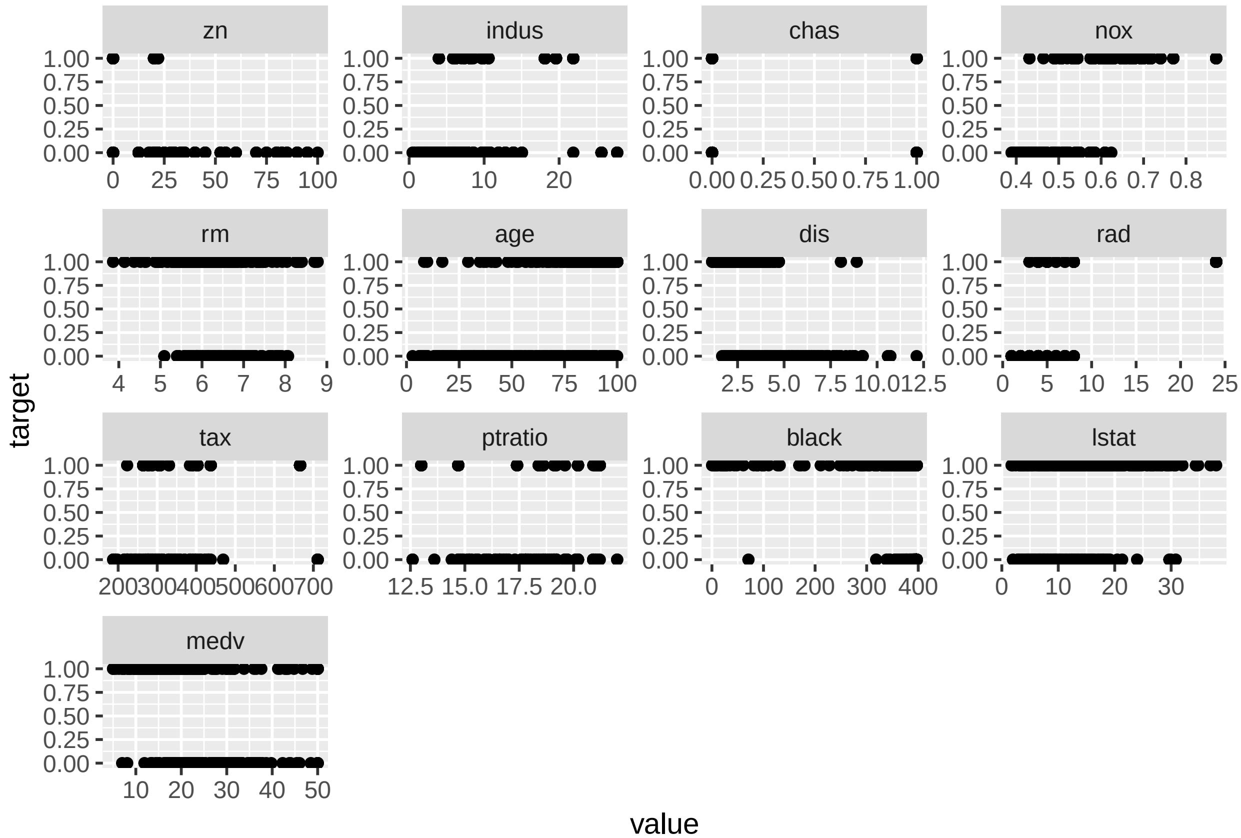
–indus, rad, tax, and target are multi-modal.

The density plots reveal that most is the data is not normal.



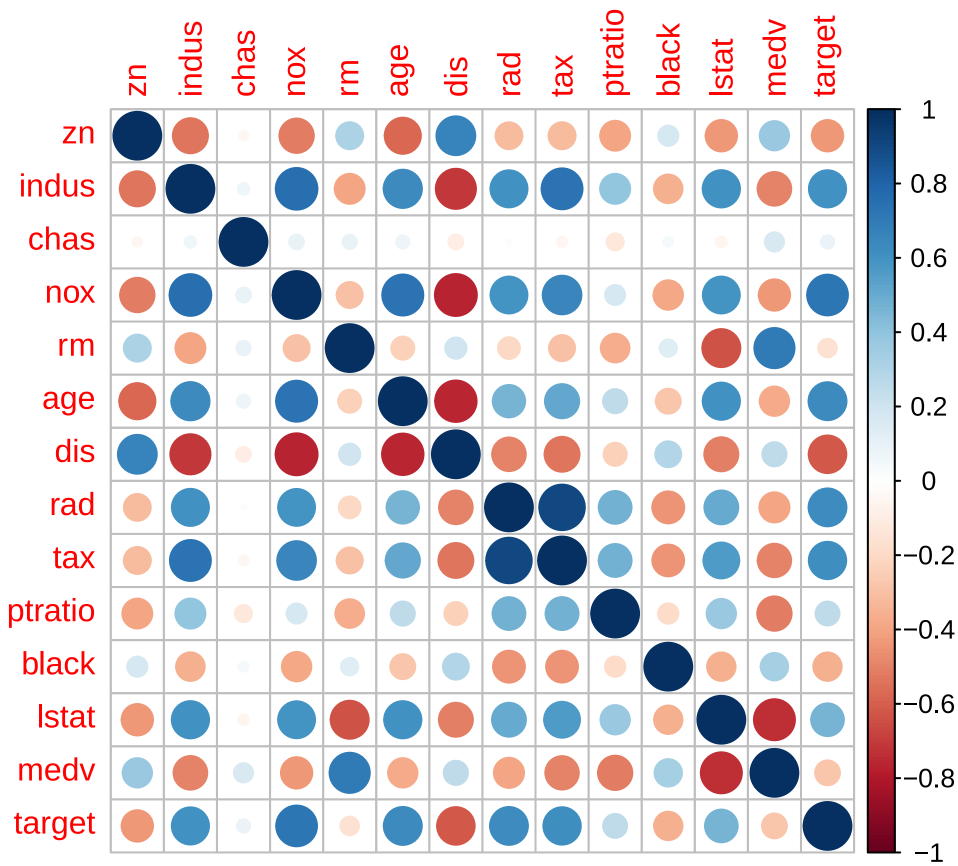
Scatterplot

Interpreting a binomial reseponse variable may not be the most best way to visualize the data using scatterplot.



Correlations

“nox”" has the highest positive correlation and “dis”" has the highest negative correlation. “tax”" and “rad” are correlated to each other.




```
## [1] "Correlation to Target"
```

	target
zn	-0.4316818
indus	0.6048507
chas	0.0800419
nox	0.7261062
rm	-0.1525533
age	0.6301062
dis	-0.6186731
rad	0.6281049
tax	0.6111133
ptratio	0.2508489
black	-0.3529568
lstat	0.4691270
medv	-0.2705507
target	1.0000000

```
## [1] "Positive Correlative Factors:"

## [1] "indus"   "nox"     "age"     "rad"     "tax"     "ptratio" "lstat"
## [8] "target"
```

```
## [1] "Negative Correlative Factors:"
## [1] "zn"    "dis"   "black" "medv"
## [1] "Neutral Correlative Factors"
## [1] "chas" "rm"
## [1] "Highly Positive Correlated Variables"
```

	target
nox	0.7261062

```
## [1] "Highly Negative Correlated Variables"
```

	target
dis	-0.6186731

2. DATA PREPARATION

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

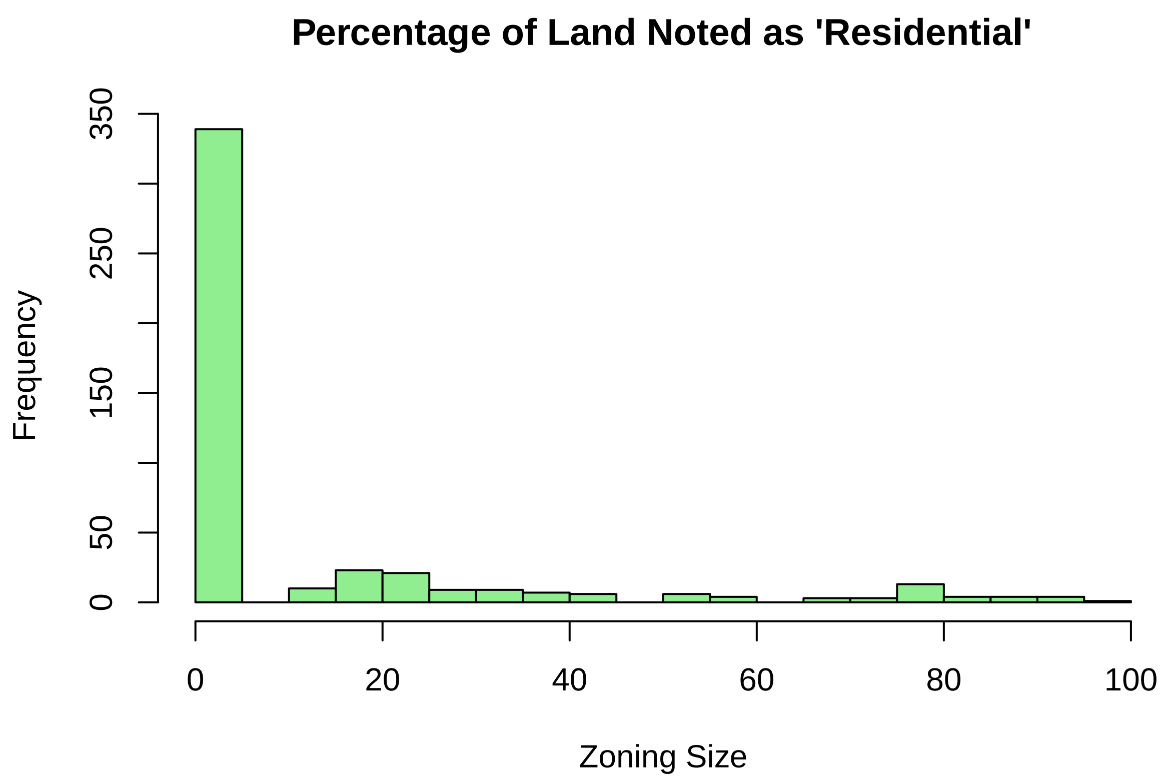
a. Fix missing values (maybe with a Mean or Median value)
b. Create flags to suggest if a variable was missing
c. Transform data by putting it into buckets
d. Mathematical transforms such as log or square root (or use Box-Cox)
e. Combine variables (such as ratios or adding or multiplying) to create new variables

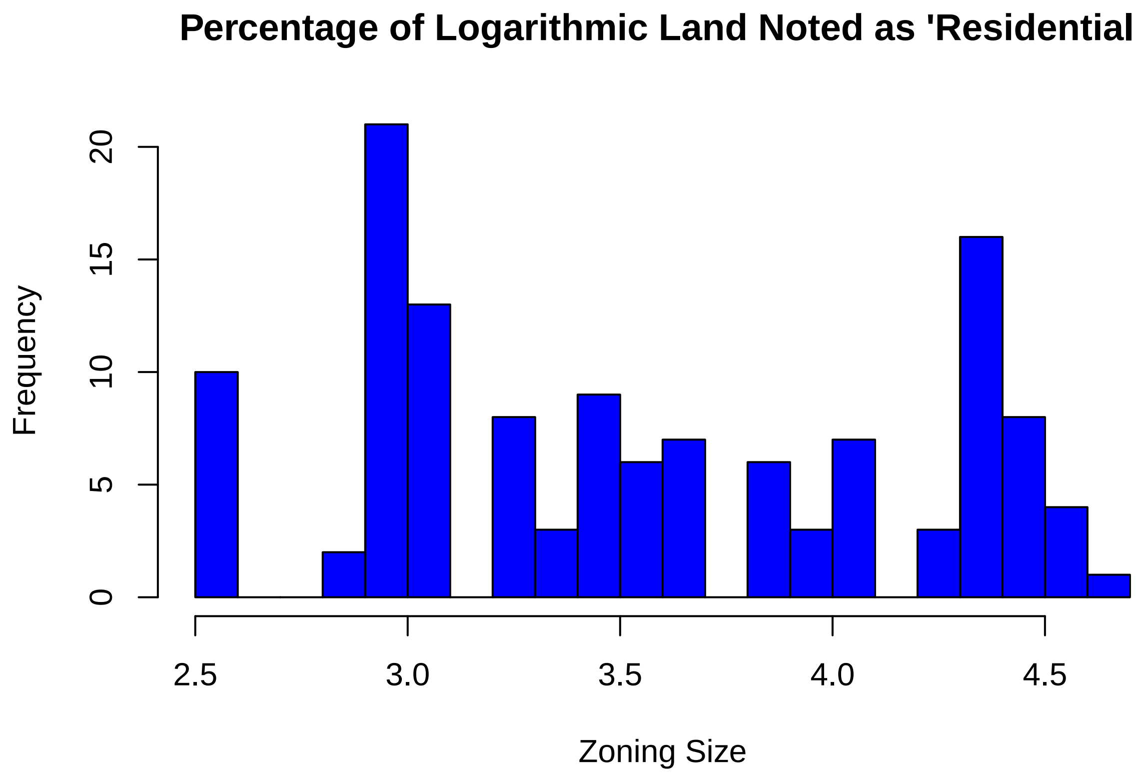
Transformation:

To reduce the effect of skewness on the model, lets do log transformations on all the variables except the variables that are binary(Zn,chas).

Due to high correlation between two indepdent variables that is between tax and rad, We can build an interactive term for this when we build our models as they are possibly likely very dependent on each other with one term affecting the other.

Lets look at Zn:proportion of residential land zoned for large lots

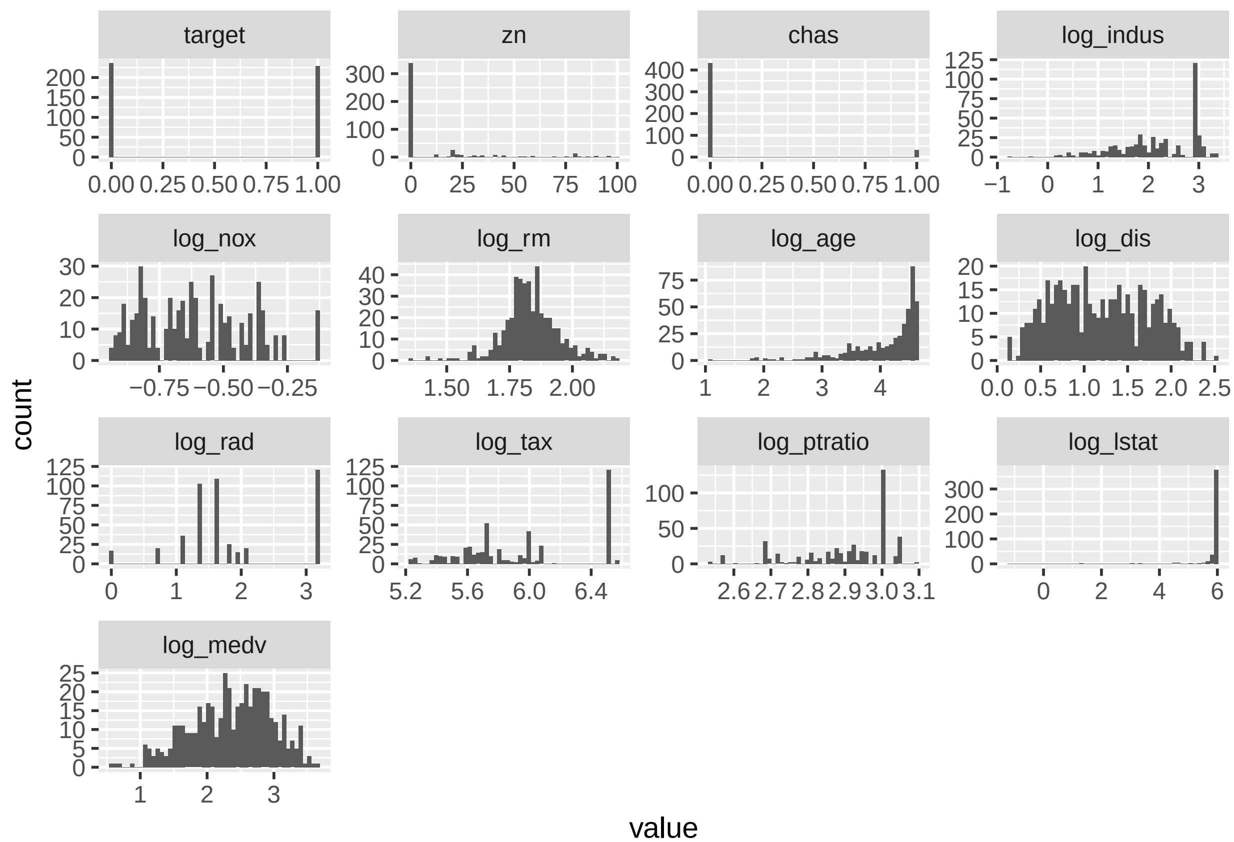



Percentage of Logarithmic Land Noted as 'Residential'
Frequency
0
5
10
15
20
2.5
3.0
3.5
4.0
4.5
Zoning Size

Let's create a scatterplot with these new variables for the logarithmic transformations.

ScatterPlot of log transformations

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



3. BUILD MODELS

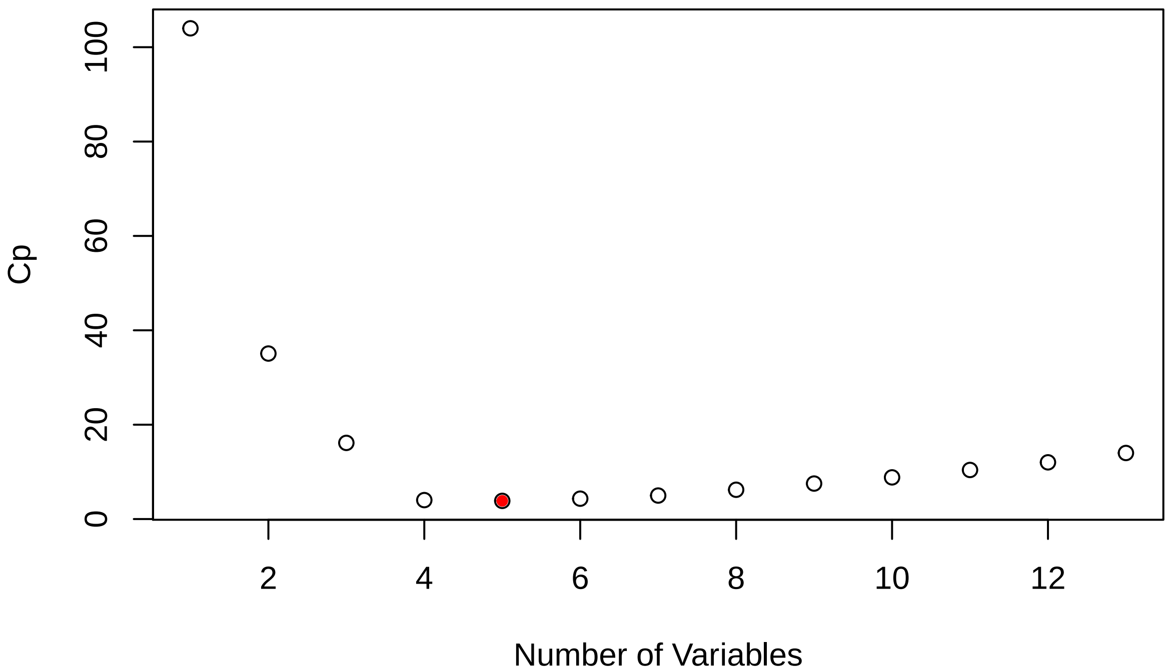
Using the training data, build at least three different binary logistic regression models, using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Be sure to explain how you can make inferences from the model, as well as discuss other relevant model output. Discuss the coefficients in the models, do they make sense? Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

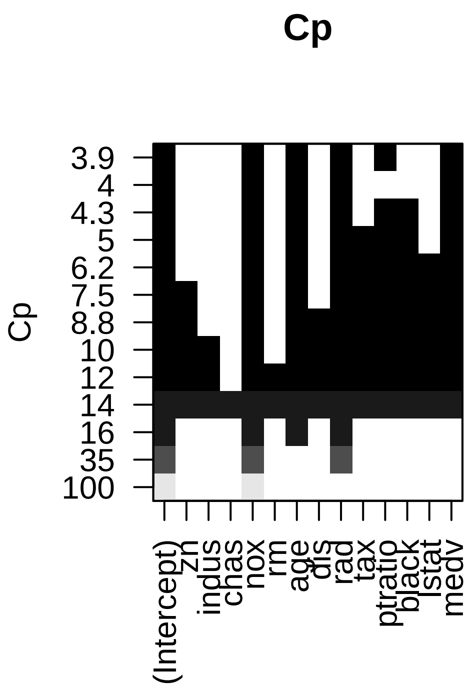
Leaps Subsetting of Untransformed Data

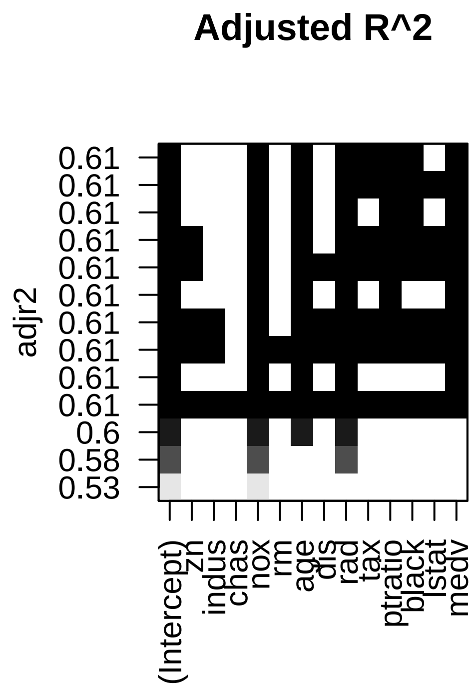
The Leaps package is an "regression subset selection" tool. The package automatically generates all possible models. The tool is basically used to find the "best" model.

```
## Subset selection object
## Call: regsubsets.formula(target ~ ., data = crime_train, method = "exhaustive",
##     nvmax = NULL, nbest = 1)
## 13 Variables  (and intercept)
##         Forced in Forced out
## zn          FALSE      FALSE
## indus       FALSE      FALSE
## chas        FALSE      FALSE
## nox         FALSE      FALSE
## rm          FALSE      FALSE
## age         FALSE      FALSE
## dis         FALSE      FALSE
## rad         FALSE      FALSE
## tax         FALSE      FALSE
## ptratio     FALSE      FALSE
## black       FALSE      FALSE
## lstat       FALSE      FALSE
## medv        FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: exhaustive
##           zn  indus chas nox rm  age dis rad tax ptratio black lstat medv
## 1  ( 1 )  " " " "   " "  "*" " " " " " " " " " " " "     " "   " "   " "
## 2  ( 1 )  " " " "   " "  "*" " " " " " " "*" " " " "     " "   " "   " "
## 3  ( 1 )  " " " "   " "  "*" " " "*" " " "*" " " " "     " "   " "   " "
## 4  ( 1 )  " " " "   " "  "*" " " "*" " " "*" " " " "     " "   " "   "*"
## 5  ( 1 )  " " " "   " "  "*" " " "*" " " "*" " " "*"     " "   " "   "*"
## 6  ( 1 )  " " " "   " "  "*" " " "*" " " "*" " " "*"     "*"   " "   "*"
## 7  ( 1 )  " " " "   " "  "*" " " "*" " " "*" "*" "*"     "*"   " "   "*"
## 8  ( 1 )  " " " "   " "  "*" " " "*" " " "*" "*" "*"     "*"   "*"   "*"
## 9  ( 1 )  "*" " "   " "  "*" " " "*" " " "*" "*" "*"     "*"   "*"   "*"
## 10  ( 1 ) "*" " "   " "  "*" " " "*" "*" "*" "*" "*"     "*"   "*"   "*"
## 11  ( 1 ) "*" "*"   " "  "*" " " "*" "*" "*" "*" "*"     "*"   "*"   "*"
## 12  ( 1 ) "*" "*"   " "  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"   "*"
## 13  ( 1 ) "*" "*"   "*"  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"   "*"
```



Cp
0
20
40
60
80
100
2
4
6
8
10
12
Number of Variables





Based on Cp, a model that includes nox, age, rad, ptratio, and medv would be the best predictor.

Based on Adjusted R^2, a model that includes nox, age, rad, tax, ptratio, black, and medv would be the best predictor.

Both metrics share the nox, age, rad, ptratio, and medv variables.

Model 1: All Variables

The glmulti package is an "automated model selection and model averaging" tool. The package automatically generates all possible models "with the specified response and explanatory variables".

All of the variables will be tested to determine the base model they provided. This will allow us to see which variables are significant in our dataset, and allow us to make other models based on that. This model will be based off of the original data - before transformed (log) variables have been added to account for potential issues in the data.

"nox", "rad","ptratio" are highly statistically significant and "dis" and "medv" are somewhat significant. "nox" has high impact on target. tax has minimum impact and also negative.

Positive coefficients: chas, nox, age, dis, rad, ptratio, lstat, medv

Negative coefficients: zn, indus, rm, tax, black

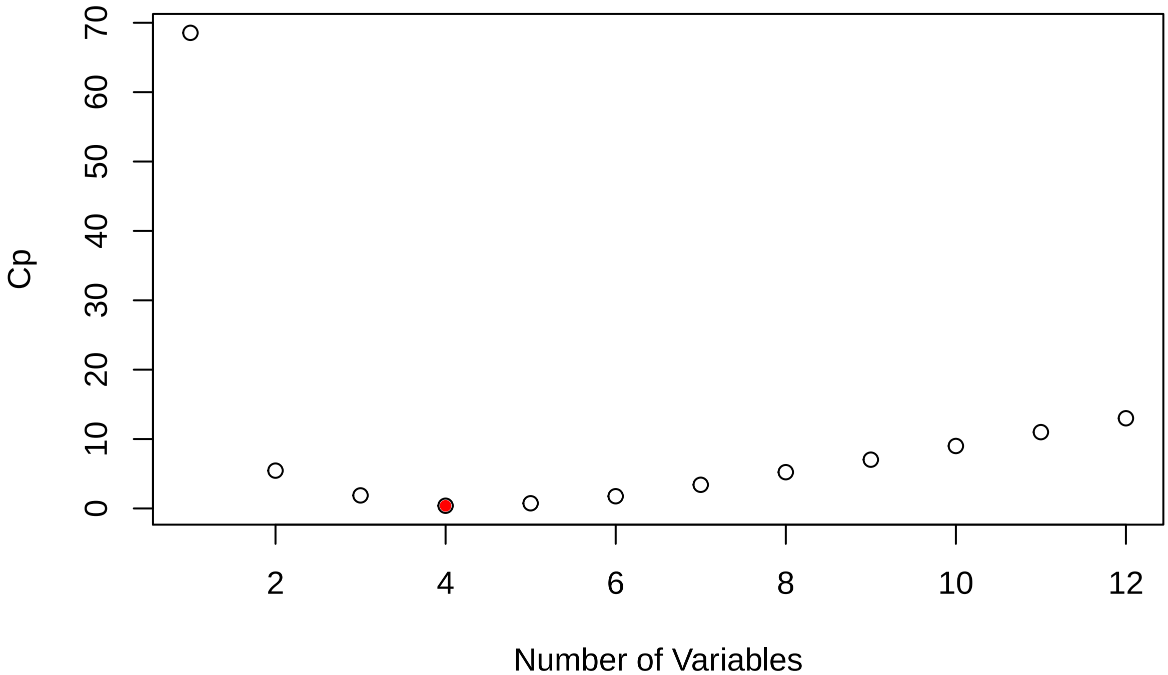
lets see how the other models reports on the deviance and AIC for comparison.

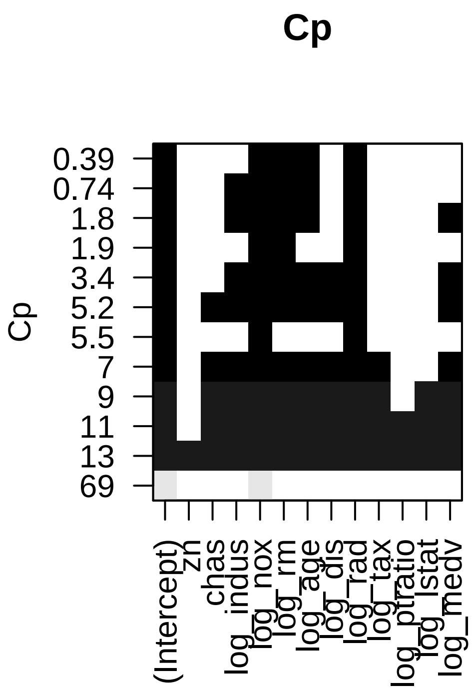
```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = crime_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2854  -0.1372  -0.0017   0.0020   3.4721
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (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
```

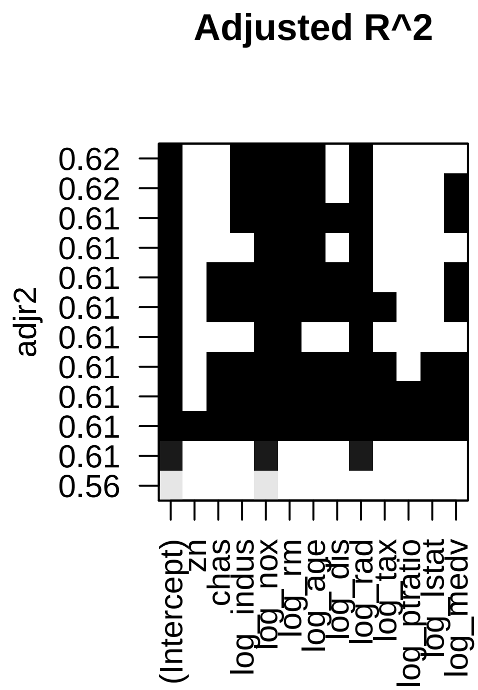
Transformed Data Analysis.

Lets look at the transformed Data.

```
## Subset selection object
## Call: regsubsets.formula(target ~ ., data = crime_train_log, method = "exhaustive",
##     nvmax = NULL, nbest = 1)
## 12 Variables  (and intercept)
##             Forced in Forced out
## zn              FALSE      FALSE
## chas            FALSE      FALSE
## log_indus       FALSE      FALSE
## log_nox         FALSE      FALSE
## log_rm          FALSE      FALSE
## log_age         FALSE      FALSE
## log_dis         FALSE      FALSE
## log_rad         FALSE      FALSE
## log_tax         FALSE      FALSE
## log_ptratio     FALSE      FALSE
## log_lstat       FALSE      FALSE
## log_medv        FALSE      FALSE
## 1 subsets of each size up to 12
## Selection Algorithm: exhaustive
##           zn  chas log_indus log_nox log_rm log_age log_dis log_rad
## 1  ( 1 )  " " " "  " "       "*"     " "    " "     " "     " "
## 2  ( 1 )  " " " "  " "       "*"     " "    " "     " "     "*"
## 3  ( 1 )  " " " "  " "       "*"     "*"    " "     " "     "*"
## 4  ( 1 )  " " " "  " "       "*"     "*"    "*"     " "     "*"
## 5  ( 1 )  " " " "  "*"       "*"     "*"    "*"     " "     "*"
## 6  ( 1 )  " " " "  "*"       "*"     "*"    "*"     " "     "*"
## 7  ( 1 )  " " " "  "*"       "*"     "*"    "*"     "*"     "*"
## 8  ( 1 )  " " "*"  "*"       "*"     "*"    "*"     "*"     "*"
## 9  ( 1 )  " " "*"  "*"       "*"     "*"    "*"     "*"     "*"
## 10  ( 1 ) " " "*"  "*"       "*"     "*"    "*"     "*"     "*"
## 11  ( 1 ) " " "*"  "*"       "*"     "*"    "*"     "*"     "*"
## 12  ( 1 ) "*" "*"  "*"       "*"     "*"    "*"     "*"     "*"
##           log_tax log_ptratio log_lstat log_medv
## 1  ( 1 )  " "     " "         " "       " "
## 2  ( 1 )  " "     " "         " "       " "
## 3  ( 1 )  " "     " "         " "       " "
## 4  ( 1 )  " "     " "         " "       " "
## 5  ( 1 )  " "     " "         " "       " "
## 6  ( 1 )  " "     " "         " "       "*"
## 7  ( 1 )  " "     " "         " "       "*"
## 8  ( 1 )  " "     " "         " "       "*"
## 9  ( 1 )  "*"     " "         " "       "*"
## 10  ( 1 ) "*"     " "         "*"       "*"
## 11  ( 1 ) "*"     "*"         "*"       "*"
## 12  ( 1 ) "*"     "*"         "*"       "*"
```


Cp
0
10
20
30
40
50
60
70
2
4
6
8
10
12
Number of Variables





CP has reduced to 4 from 5 after transformation.

Both CP and Rsquare indicates “nox”, “rm” , age“,”rad" are best predictors.

Model 2: Transformed Variables

Model2 is the log transformation of all the variables and the interacetive term is included.

The log variables should help negate the large amount of skew in the data - or help them to become more normalized.

```
##
## Call:
## glm(formula = target ~ . + log_rad:log_tax, family = binomial(link = "logit"),
##     data = crime_train_log)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.00318  -0.17093  -0.00164   0.10619   3.13261
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.46216   14.45018  -0.170  0.86470
## zn              -0.03409    0.02674  -1.275  0.20227
## chas             0.94719    0.77212   1.227  0.21992
## log_indus        0.34000    0.56795   0.599  0.54941
## log_nox         22.81455    3.76435   6.061 1.36e-09 ***
## log_rm           5.30404    2.99089   1.773  0.07616 .
## log_age          0.48567    0.56896   0.854  0.39332
## log_dis          2.13517    0.78734   2.712  0.00669 **
## log_rad          6.89690    7.26768   0.949  0.34263
## log_tax         -1.33460    1.69875  -0.786  0.43208
## log_ptratio      3.97363    1.82156   2.181  0.02915 *
## log_lstat       -1.17666    1.13585  -1.036  0.30023
## log_medv        -0.16766    0.61577  -0.272  0.78541
## log_rad:log_tax -0.63883    1.18935  -0.537  0.59118
## ---
## 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: 207.08  on 452  degrees of freedom
## AIC: 235.08
##
## Number of Fisher Scoring iterations: 8
```

Analysis:

nox has the greatest impact on target.

nox, rad are highly statistically significant.

AIC has increased compared to model1.

Null deviance is the same. Residual deviance has increased.

interactive term did not add much value.

So, this model may not be the best choice. let me try stepwise for Model1 and Model2.

Model 3: (Logarithmic Model) with stepwise.

Let me try both forward and backward elimination stepwise algorithm here.

```
## Start:  AIC=235.08
## target ~ zn + chas + log_indus + log_nox + log_rm + log_age +
##     log_dis + log_rad + log_tax + log_ptratio + log_lstat + log_medv +
##     log_rad:log_tax
##
##                   Df Deviance    AIC
## - log_medv         1   207.15 233.15
## - log_rad:log_tax  1   207.35 233.35
## - log_indus        1   207.44 233.44
## - log_age          1   207.84 233.84
## - log_lstat        1   208.53 234.53
## - chas             1   208.61 234.61
## - zn               1   208.99 234.99
## <none>                 207.08 235.08
## - log_rm           1   210.17 236.17
## - log_ptratio      1   211.92 237.92
## - log_dis          1   214.86 240.86
## - log_nox          1   267.00 293.00
##
## Step:  AIC=233.15
## target ~ zn + chas + log_indus + log_nox + log_rm + log_age +
##     log_dis + log_rad + log_tax + log_ptratio + log_lstat + log_rad:log_tax
##
##                   Df Deviance    AIC
## - log_rad:log_tax  1   207.42 231.42
## - log_indus        1   207.50 231.50
## - log_age          1   207.85 231.85
## - log_lstat        1   208.61 232.61
## - chas             1   208.63 232.63
## <none>                 207.15 233.15
## - zn               1   209.24 233.24
## + log_medv         1   207.08 235.08
## - log_ptratio      1   211.93 235.93
## - log_rm           1   214.62 238.62
## - log_dis          1   214.88 238.88
## - log_nox          1   267.25 291.25
##
## Step:  AIC=231.42
## target ~ zn + chas + log_indus + log_nox + log_rm + log_age +
##     log_dis + log_rad + log_tax + log_ptratio + log_lstat
##
##                   Df Deviance    AIC
## - log_indus        1   207.62 229.62
## - log_age          1   208.11 230.11
## - log_lstat        1   208.75 230.75
## - chas             1   209.24 231.24
## <none>                 207.42 231.42
## - zn               1   209.57 231.57
## + log_rad:log_tax  1   207.15 233.15
## - log_tax          1   211.31 233.31
```

```
## + log_medv         1   207.35 233.35
## - log_ptratio      1   212.07 234.07
## - log_rm           1   215.38 237.38
## - log_dis          1   216.15 238.15
## - log_rad          1   248.83 270.83
## - log_nox          1   268.82 290.82
##
## Step:  AIC=229.62
## target ~ zn + chas + log_nox + log_rm + log_age + log_dis + log_rad +
##     log_tax + log_ptratio + log_lstat
##
##                  Df Deviance    AIC
## - log_age         1   208.28 228.28
## - log_lstat       1   208.99 228.99
## <none>                207.62 229.62
## - chas            1   209.88 229.88
## - zn              1   210.12 230.12
## + log_indus       1   207.42 231.42
## + log_rad:log_tax  1   207.50 231.50
## + log_medv        1   207.56 231.56
## - log_tax         1   211.64 231.64
## - log_ptratio     1   212.53 232.53
## - log_rm          1   215.41 235.41
## - log_dis         1   216.40 236.40
## - log_rad         1   249.58 269.58
## - log_nox         1   278.93 298.93
##
## Step:  AIC=228.28
## target ~ zn + chas + log_nox + log_rm + log_dis + log_rad + log_tax +
##     log_ptratio + log_lstat
##
##                  Df Deviance    AIC
## - log_lstat       1   209.76 227.76
## <none>                208.28 228.28
## - chas            1   210.87 228.87
## - zn              1   211.06 229.06
## + log_age         1   207.62 229.62
## + log_indus       1   208.11 230.11
## - log_tax         1   212.12 230.12
## + log_rad:log_tax  1   208.17 230.17
## + log_medv        1   208.27 230.27
## - log_ptratio     1   213.08 231.08
## - log_rm          1   216.16 234.16
## - log_dis         1   216.50 234.50
## - log_rad         1   249.61 267.61
## - log_nox         1   287.61 305.61
##
## Step:  AIC=227.76
## target ~ zn + chas + log_nox + log_rm + log_dis + log_rad + log_tax +
##     log_ptratio
##
##                  Df Deviance    AIC
## <none>                209.76 227.76
## - chas            1   212.20 228.20
```

```
## + log_lstat        1   208.28 228.28
## - zn               1   212.64 228.64
## + log_age          1   208.99 228.99
## - log_tax          1   213.37 229.37
## + log_indus        1   209.55 229.55
## + log_rad:log_tax  1   209.73 229.73
## + log_medv         1   209.74 229.74
## - log_ptratio      1   214.75 230.75
## - log_rm           1   217.34 233.34
## - log_dis          1   217.64 233.64
## - log_rad          1   255.17 271.17
## - log_nox          1   290.84 306.84
```

```
##
## Call:
## glm(formula = target ~ zn + chas + log_nox + log_rm + log_dis +
##     log_rad + log_tax + log_ptratio, family = binomial(link = "logit"),
##     data = crime_train_log)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.93386  -0.18898  -0.00247   0.09744   3.11267
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.89208    8.50307  -0.458  0.64715
## zn          -0.04000    0.02603  -1.537  0.12439
## chas         1.10355    0.72056   1.532  0.12565
## log_nox     23.90164    3.52542   6.780 1.20e-11 ***
## log_rm       5.68049    2.10479   2.699  0.00696 **
## log_dis      2.05974    0.75548   2.726  0.00640 **
## log_rad      2.97131    0.60525   4.909 9.14e-07 ***
## log_tax     -1.75087    0.93090  -1.881  0.05999 .
## log_ptratio  3.86798    1.74937   2.211  0.02703 *
## ---
## 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: 209.76  on 457  degrees of freedom
## AIC: 227.76
##
## Number of Fisher Scoring iterations: 8
```

Analysis:

AIC has decreased compared to model-2 but still above model-1.

No difference on the Null deviance and Residual deviance.

"nox", "rad" are highly significant and "rm", "dis" are less significant.

Model 4: (Logarithmic Model) with Principal Components.

want to check how many variables are selecgted in this model.

```
## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to
## do classification? If so, use a 2 level factor as your outcome column.
```

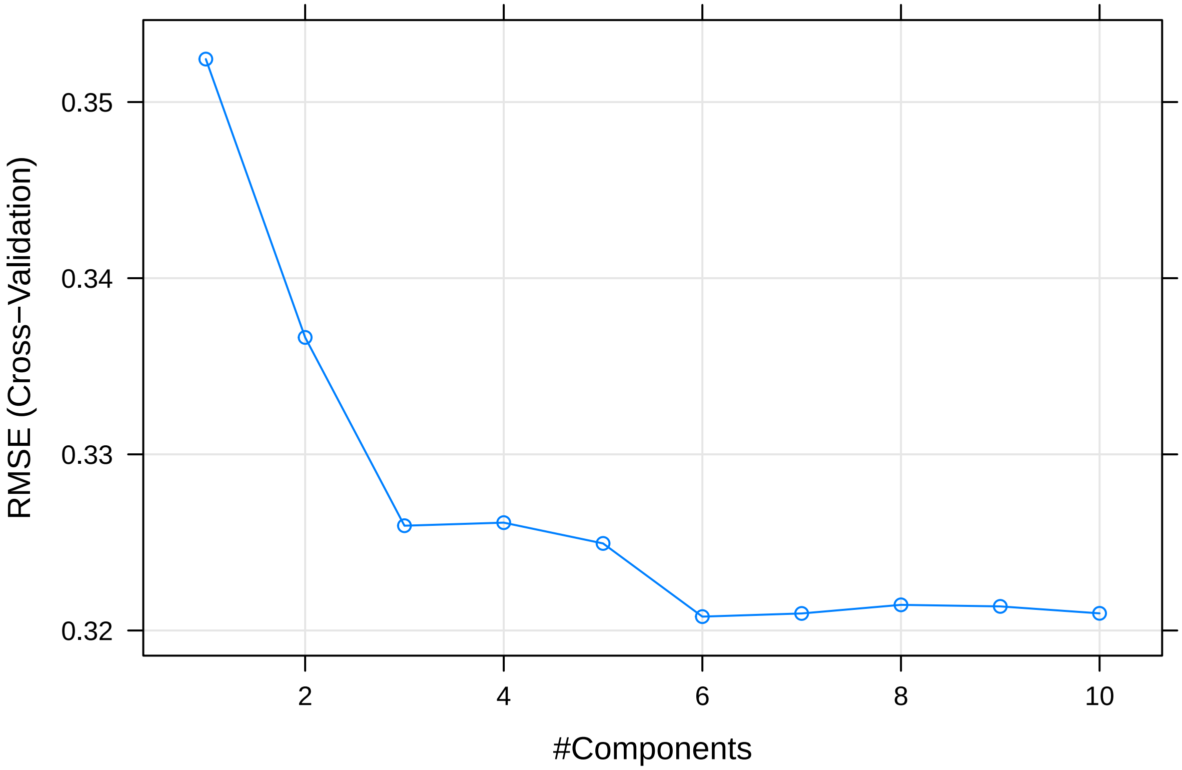
Model summary

```
## Data:    X dimension: 466 12
##  Y dimension: 466 1
## Fit method: svdpc
## Number of components considered: 6
## TRAINING: % variance explained
##           1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X           47.34    58.58    68.41    75.58    82.02    87.88
## .outcome    50.27    54.85    57.81    57.99    58.42    59.37
```

Model Results

ncomp	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
1	0.3524385	0.5178057	0.2812633	0.0417900	0.1121874	0.0387359
2	0.3366375	0.5565167	0.2663780	0.0415044	0.1106530	0.0362781
3	0.3259484	0.5781348	0.2554683	0.0345522	0.0888443	0.0321500
4	0.3261228	0.5776264	0.2545356	0.0325305	0.0852492	0.0312235
5	0.3249427	0.5793505	0.2542360	0.0330624	0.0855554	0.0316626
6	0.3207876	0.5900258	0.2403649	0.0333483	0.0860768	0.0331914
7	0.3209674	0.5893919	0.2400574	0.0335698	0.0867198	0.0330836
8	0.3214561	0.5886781	0.2418056	0.0329510	0.0855423	0.0320599
9	0.3213672	0.5895202	0.2414771	0.0325073	0.0829208	0.0312326
10	0.3209735	0.5898968	0.2425102	0.0318672	0.0812026	0.0299609

Model Plot



	ncomp
6	6

Analysis:

This model has selected upto 6 components.

4. SELECT MODELS

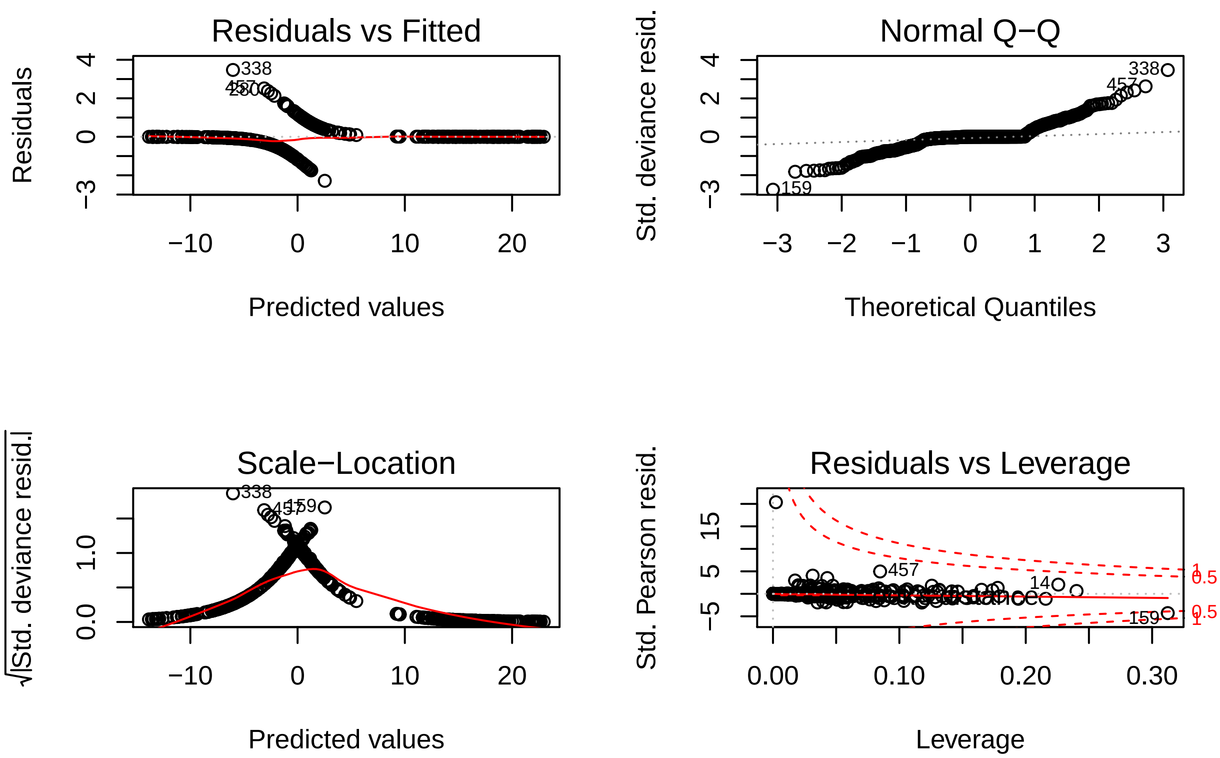
Decide on the criteria for selecting the best binary logistic regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models.

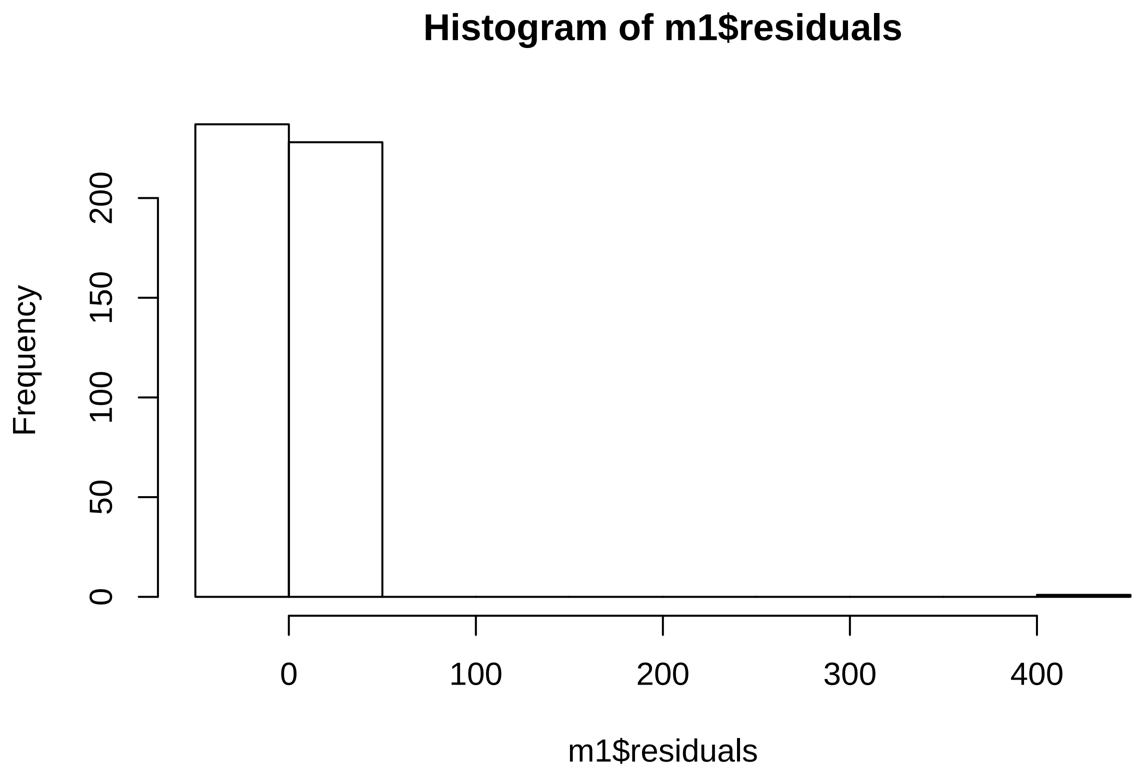
For the binary logistic regression model, will you use a metric such as log likelihood, AIC, ROC curve, etc.? Using the training data set, evaluate the binary logistic regression model based on (a) accuracy, (b) classification error rate, (c) precision, (d) sensitivity, (e) specificity, (f) F1 score, (g) AUC, and (h) confusion matrix. Make predictions using the evaluation data set.

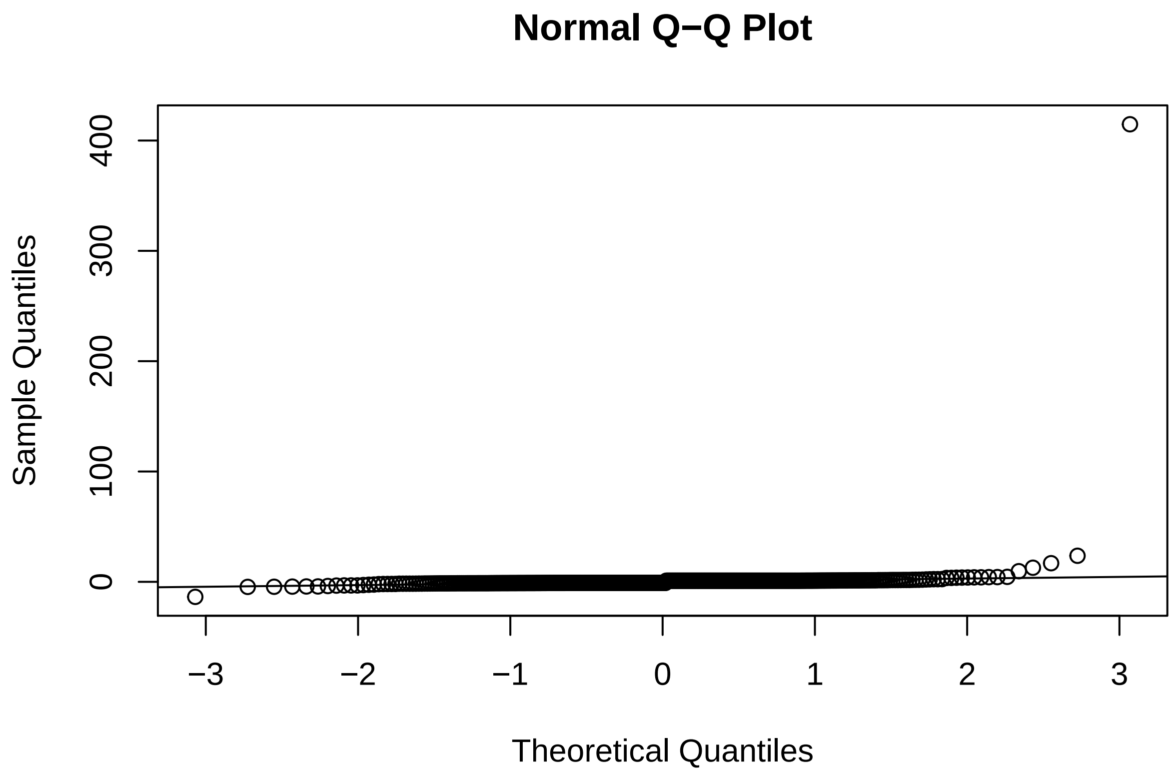
Analysis:

After looking at the residual deviance scores and AIC scores in the previous section, we'll evaluate the model1 here.

Let us evalute the Model Number 1 (baseline model). Next, we will develop a confusion matrix and create our evaluations there.




Histogram of m1$residuals
Frequency
0
50
100
150
200
0
100
200
300
400
m1$residuals



The histogram of the residuals do not show a normal distribution.

The qqplot shows a fairly linear relationship, except towards the tail end of the residuals.

The residual indicates that there is not constant variance throughout, as there is a noticable pattern around 0.

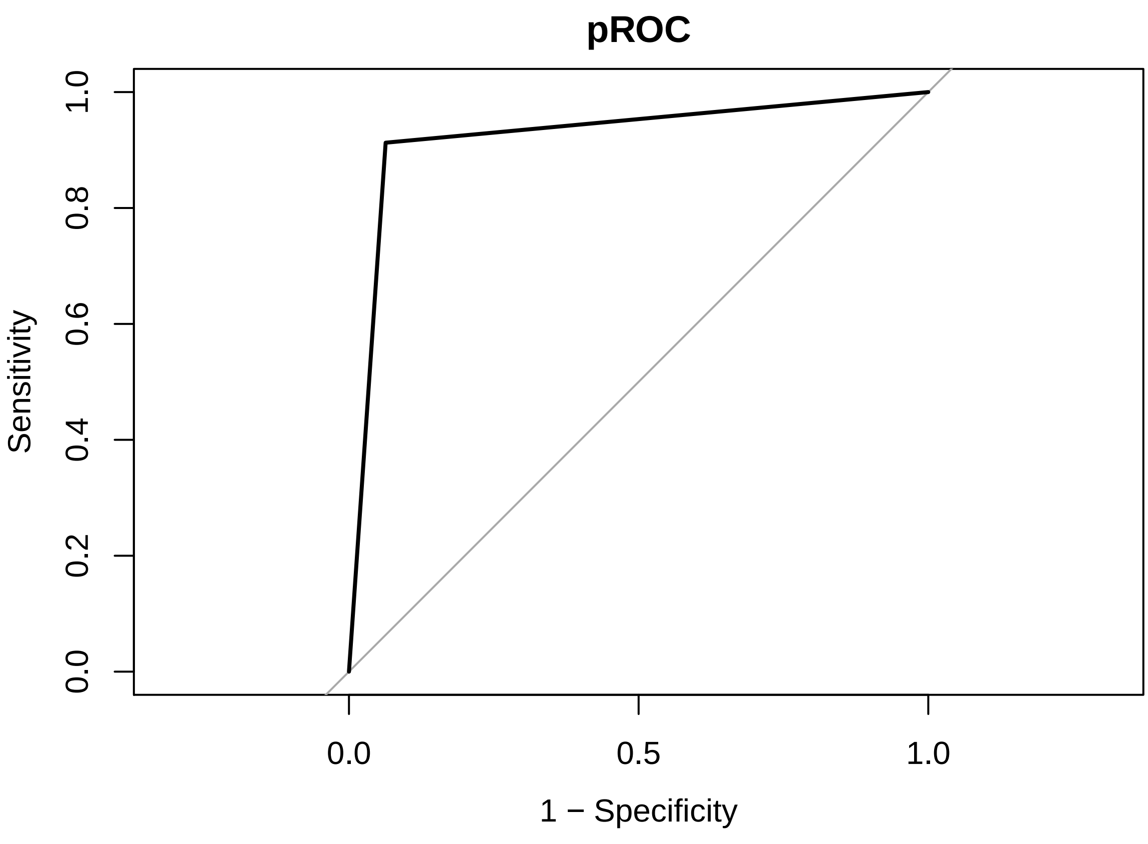
Test Model1

zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv	target	scored_targe
0	19.58	0	0.605	7.929	96.2	2.0459	5	403	14.7	369.30	3.70	50.0	1	
0	19.58	1	0.871	5.403	100.0	1.3216	5	403	14.7	396.90	26.82	13.4	1	
0	18.10	0	0.740	6.485	100.0	1.9784	24	666	20.2	386.73	18.85	15.4	1	
30	4.93	0	0.428	6.393	7.8	7.0355	6	300	16.6	374.71	5.19	23.7	0	
0	2.46	0	0.488	7.155	92.2	2.7006	3	193	17.8	394.12	4.82	37.9	0	

Performance

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 222  20
##          1  15 209
##
##                Accuracy : 0.9249
```

```
##                  95% CI : (0.8971, 0.9471)
##     No Information Rate : 0.5086
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8497
##  Mcnemar's Test P-Value : 0.499
##
##             Sensitivity : 0.9127
##             Specificity : 0.9367
##          Pos Pred Value : 0.9330
##          Neg Pred Value : 0.9174
##               Precision : 0.9330
##                  Recall : 0.9127
##                      F1 : 0.9227
##              Prevalence : 0.4914
##          Detection Rate : 0.4485
##    Detection Prevalence : 0.4807
##       Balanced Accuracy : 0.9247
##
##        'Positive' Class : 1
##
```



Analysis:

This model has 90% accuracy. Precision is 95%. Negative prediction rate is only 91%. Positive prediction rate is 93. Sensitivity is 91% Specificity is 93% F1 is 92% AUC is 92%

Prediction for Test Data

```
##    zn indus chas   nox    rm   age    dis rad tax ptratio  black lstat
## 1   0  7.07    0 0.469 7.185  61.1 4.9671   2 242    17.8 392.83  4.03
## 2   0  8.14    0 0.538 6.096  84.5 4.4619   4 307    21.0 380.02 10.26
## 3   0  8.14    0 0.538 6.495  94.4 4.4547   4 307    21.0 387.94 12.80
## 4   0  8.14    0 0.538 5.950  82.0 3.9900   4 307    21.0 232.60 27.71
## 5   0  5.96    0 0.499 5.850  41.5 3.9342   5 279    19.2 396.90  8.77
## 6  25  5.13    0 0.453 5.741  66.2 7.2254   8 284    19.7 395.11 13.15
## 7  25  5.13    0 0.453 5.966  93.4 6.8185   8 284    19.7 378.08 14.44
## 8   0  4.49    0 0.449 6.630  56.1 4.4377   3 247    18.5 392.30  6.53
## 9   0  4.49    0 0.449 6.121  56.8 3.7476   3 247    18.5 395.15  8.44
## 10  0  2.89    0 0.445 6.163  69.6 3.4952   2 276    18.0 391.83 11.34
## 11  0 25.65    0 0.581 5.856  97.0 1.9444   2 188    19.1 370.31 25.41
## 12  0 25.65    0 0.581 5.613  95.6 1.7572   2 188    19.1 359.29 27.26
## 13  0 21.89    0 0.624 5.637  94.7 1.9799   4 437    21.2 396.90 18.34
## 14  0 19.58    0 0.605 6.101  93.0 2.2834   5 403    14.7 240.16  9.81
## 15  0 19.58    0 0.605 5.880  97.3 2.3887   5 403    14.7 348.13 12.03
## 16  0 10.59    1 0.489 5.960  92.1 3.8771   4 277    18.6 393.25 17.27
## 17  0  6.20    0 0.504 6.552  21.4 3.3751   8 307    17.4 380.34  3.76
## 18  0  6.20    0 0.507 8.247  70.4 3.6519   8 307    17.4 378.95  3.95
## 19 22  5.86    0 0.431 6.957   6.8 8.9067   7 330    19.1 386.09  3.53
## 20 90  2.97    0 0.400 7.088  20.8 7.3073   1 285    15.3 394.72  7.85
## 21 80  1.76    0 0.385 6.230  31.5 9.0892   1 241    18.2 341.60 12.93
## 22 33  2.18    0 0.472 6.616  58.1 3.3700   7 222    18.4 393.36  8.93
## 23  0  9.90    0 0.544 6.122  52.8 2.6403   4 304    18.4 396.90  5.98
## 24  0  7.38    0 0.493 6.415  40.1 4.7211   5 287    19.6 396.90  6.12
## 25  0  7.38    0 0.493 6.312  28.9 5.4159   5 287    19.6 396.90  6.15
## 26  0  5.19    0 0.515 5.895  59.6 5.6150   5 224    20.2 394.81 10.56
## 27 80  2.01    0 0.435 6.635  29.7 8.3440   4 280    17.0 390.94  5.99
## 28  0 18.10    0 0.718 3.561  87.9 1.6132  24 666    20.2 354.70  7.12
## 29  0 18.10    1 0.631 7.016  97.5 1.2024  24 666    20.2 392.05  2.96
## 30  0 18.10    0 0.584 6.348  86.1 2.0527  24 666    20.2  83.45 17.64
## 31  0 18.10    0 0.740 5.935  87.9 1.8206  24 666    20.2  68.95 34.02
## 32  0 18.10    0 0.740 5.627  93.9 1.8172  24 666    20.2 396.90 22.88
## 33  0 18.10    0 0.740 5.818  92.4 1.8662  24 666    20.2 391.45 22.11
## 34  0 18.10    0 0.740 6.219 100.0 2.0048  24 666    20.2 395.69 16.59
## 35  0 18.10    0 0.740 5.854  96.6 1.8956  24 666    20.2 240.52 23.79
## 36  0 18.10    0 0.713 6.525  86.5 2.4358  24 666    20.2  50.92 18.13
## 37  0 18.10    0 0.713 6.376  88.4 2.5671  24 666    20.2 391.43 14.65
## 38  0 18.10    0 0.655 6.209  65.4 2.9634  24 666    20.2 396.90 13.22
## 39  0  9.69    0 0.585 5.794  70.6 2.8927   6 391    19.2 396.90 14.10
## 40  0 11.93    0 0.573 6.976  91.0 2.1675   1 273    21.0 396.90  5.64
##    medv scored_target
## 1  34.7             0
## 2  18.2             1
## 3  18.4             1
## 4  13.2             1
## 5  21.0             0
## 6  18.7             0
## 7  16.0             0
## 8  26.6             0
## 9  22.2             0
## 10 21.4             0
```

```
## 11 17.3                 0
## 12 15.7                 0
## 13 14.3                 1
## 14 25.0                 1
## 15 19.1                 1
## 16 21.7                 0
## 17 31.5                 0
## 18 48.3                 1
## 19 29.6                 0
## 20 32.2                 0
## 21 20.1                 0
## 22 28.4                 0
## 23 22.1                 0
## 24 25.0                 0
## 25 23.0                 0
## 26 18.5                 1
## 27 24.5                 0
## 28 27.5                 1
## 29 50.0                 1
## 30 14.5                 1
## 31  8.4                 1
## 32 12.8                 1
## 33 10.5                 1
## 34 18.4                 1
## 35 10.8                 1
## 36 14.1                 1
## 37 17.7                 1
## 38 21.4                 1
## 39 18.3                 1
## 40 23.9                 0
```

Appendix

For full code visit:

https://github.com/raghu74us/DATA-621/blob/master/Assignment3/621_Assignment3.Rmd