**Exploratory Data Analysis:**

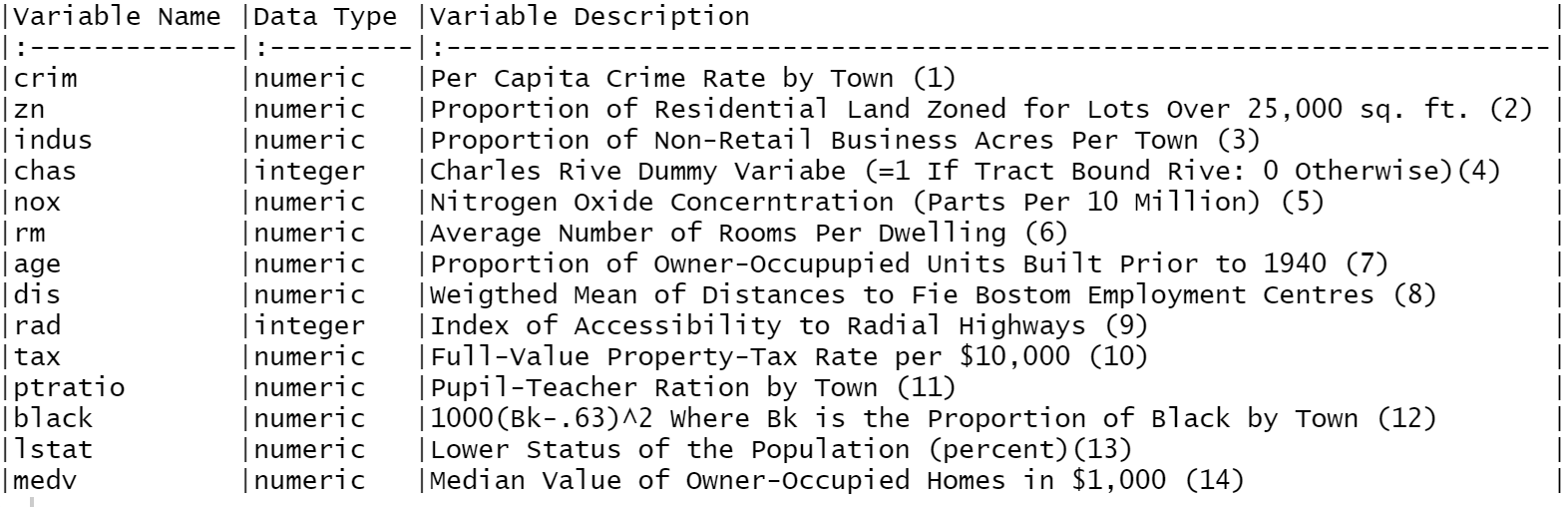
1. **Background**

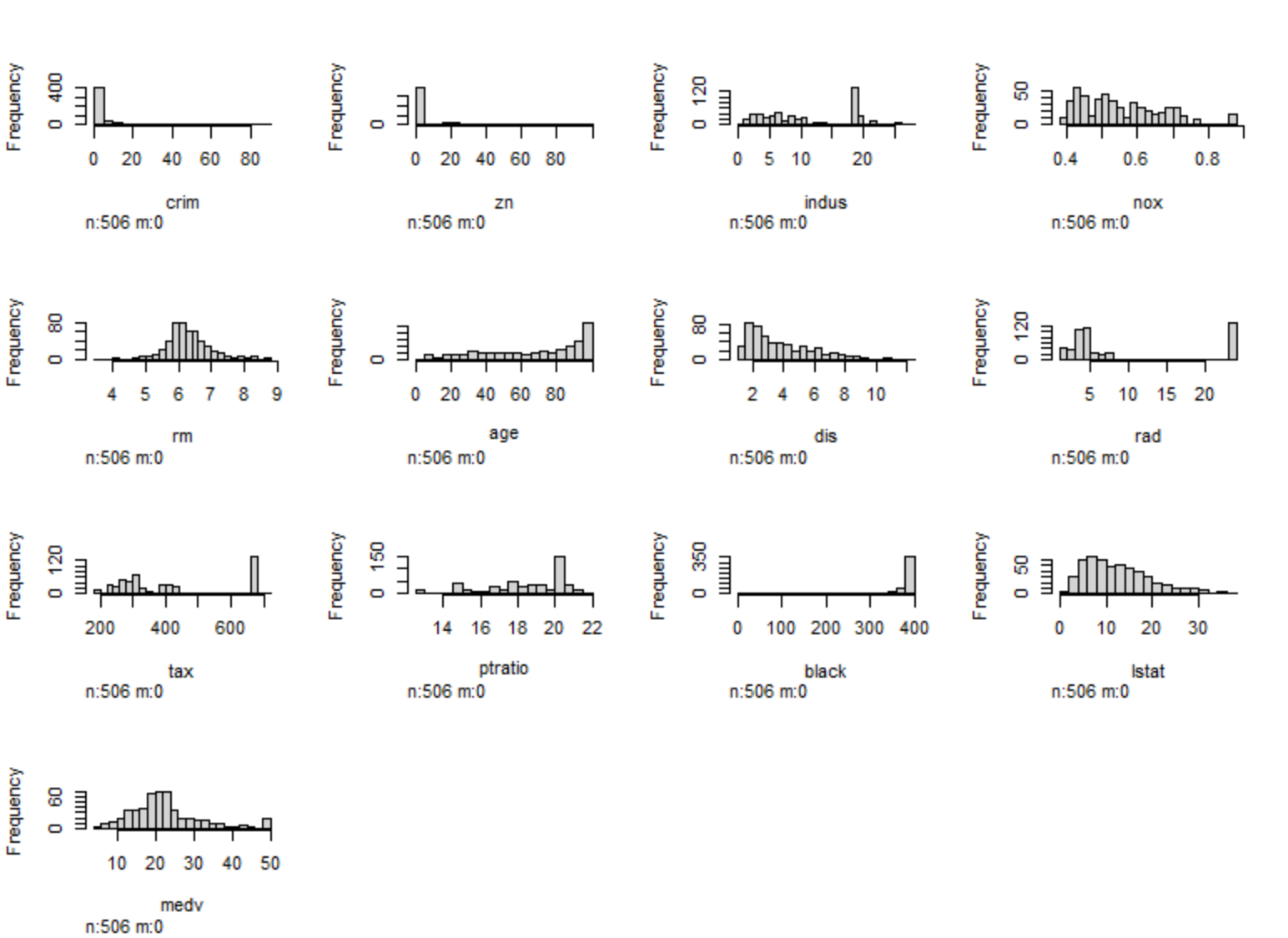
The Housing Data set was used in the 1976 publication *Hedonic Housing Prices and Demand for Clean Air.* The purpose of the data collection and analysis is to discover the nuances of housing market value with respect to air quality. The authors of this paper wanted to create an equation to model housing value with respect to air quality, estimate the value of housing to homeowners and buyers using this equation, and estimate the dollar benefits of pollution control per household.

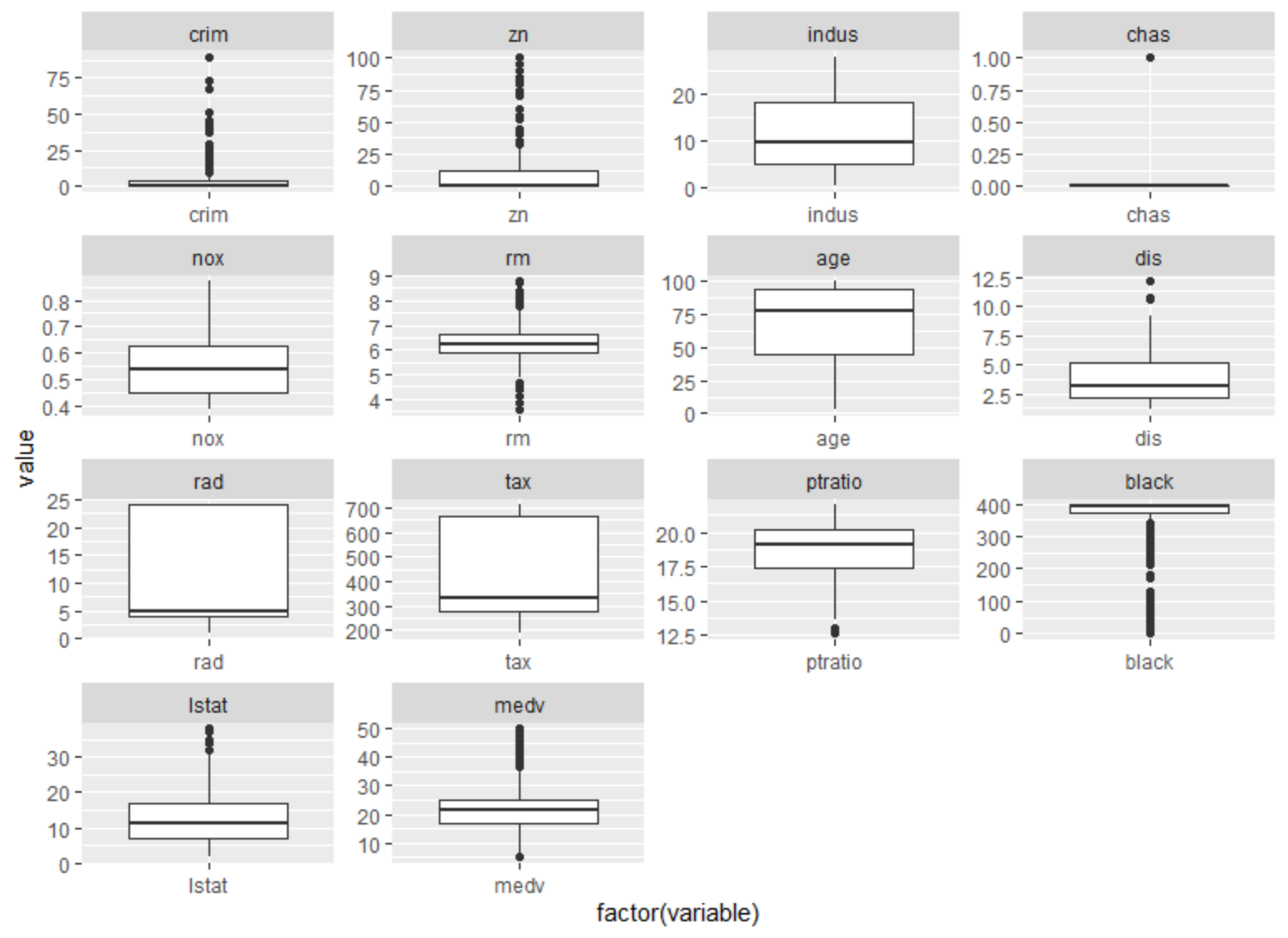
1. **Summary and Statistics**

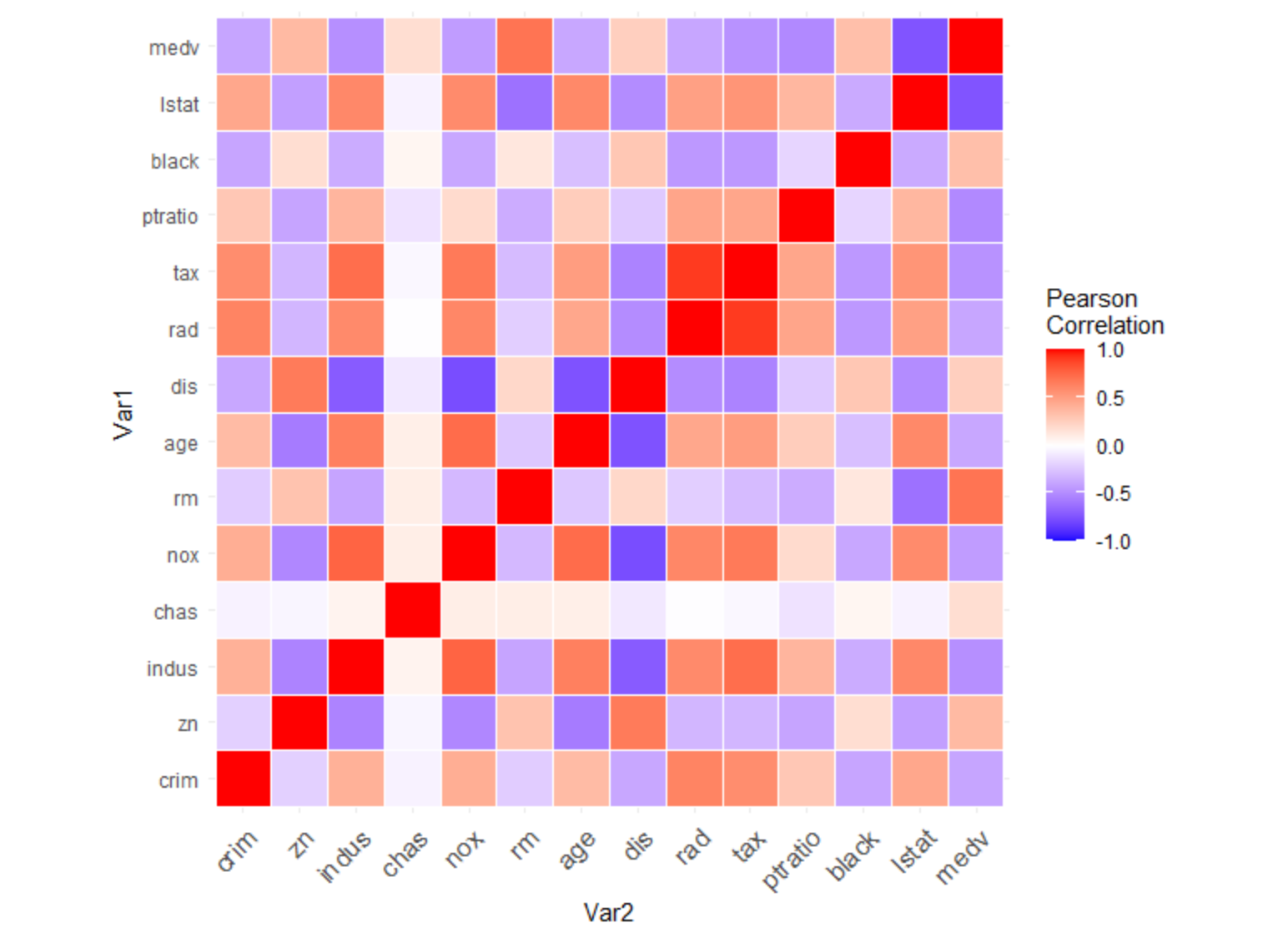
In order to understand the data we are dealing with, we created a table with variable descriptions and the count of attributes and observations. We are working with 506 neighborhoods with 14 attributes associated with each. We have 12 numeric continuous variables and 2 integer variables – one binary and one ordinal. I will be using the abbreviated variable names throughout the presentation. The table below can be used as a reference.

**Table 1: Data Description**

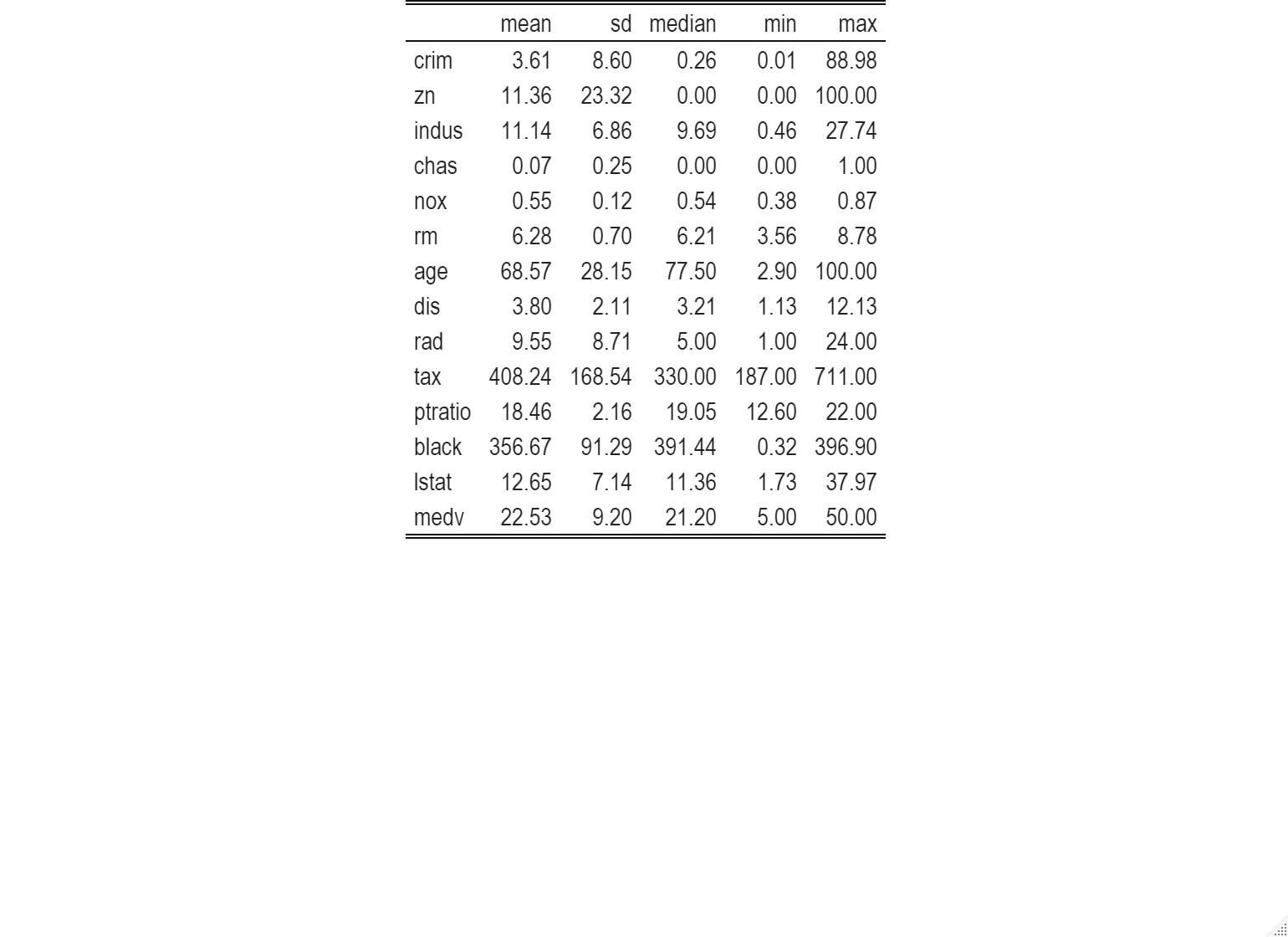


**Plot 1: Histogram of Housing Attributes**

**Plot 2: Boxplot of Housing Attributes**

**Plot 3: Correlation Heat Map**

**Table 2: Summary Statistics**



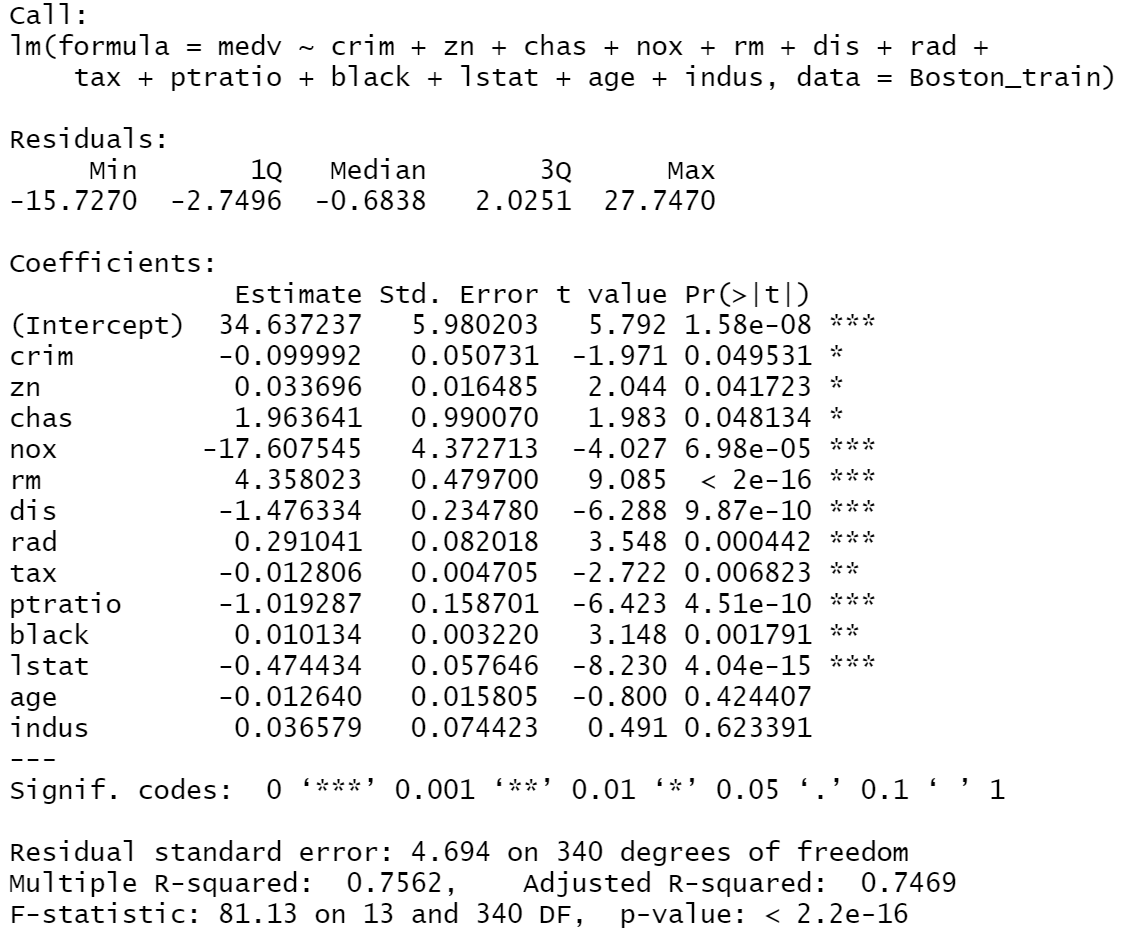
**Report:**

In looking at plot 1, it is clear the much of the data is highly skewed with irregular patterns across the spectrum of continuous data points. Rm and medv are the variables which are closest to a normal distribution. Dis, nox, and lstat appear to follow something like a power log normal distribution. Age appears to follow a power law distribution. Most of the other variables’ distributions are difficult to interpret, because they follow irregular distrbutions.

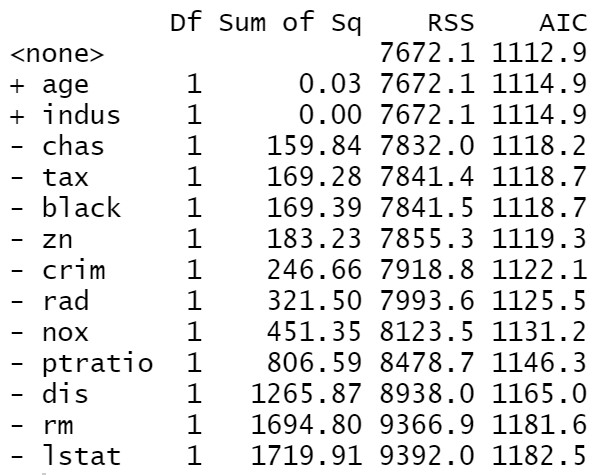
Upon observing our boxplots in plot 2 and the wide range and irregular distributions presented in our histograms, it is apparent that our data is riddled with outliers. Variables such as crim, zn, black, medv, lstat, and ptratio display many data point that lie far outside the upper quartile range. Our summary statistics reveal a large difference between the medians and means of these variables further highlighting our observations that the data is skewed and irregular. For example, crim has a mean of 3.61, a median of .26, a standard deviation of 8.6, a min of nearly 0, and a max of 88.98. The mean and median are nearly a half standard deviation in this case and the maximum value is 10 standard deviations away from the mean. Because of this kind of skewness and irregularity in the data, it makes me slightly concerned for the performance of the linear models soon to be created.

The correlation head map in plot 3 reveals that we have many notable correlations present in the data. Some of the most notable negative correlations being nox and dis, chas and dis, age and dis, and medv and lstat. The most notable positive correlations are tax and rad, tax and indus, nox and age, and medv and rm. Correlation can be a hinderance to the predictive power of a linear model do to multiple collinearity. There is a potential to give too much weight to a single cluster of colinear variables leading to inaccuracies in the predictions. There are different methods of variables selection we can use to include or exclude some of the correlated variables in our search for the most precise model. For instance, Elastic Net variable selection will tend to select one or more predictor variables from a group of correlated variables, while Lasso variable selection will usually only select one variable from a group of correlated variables.

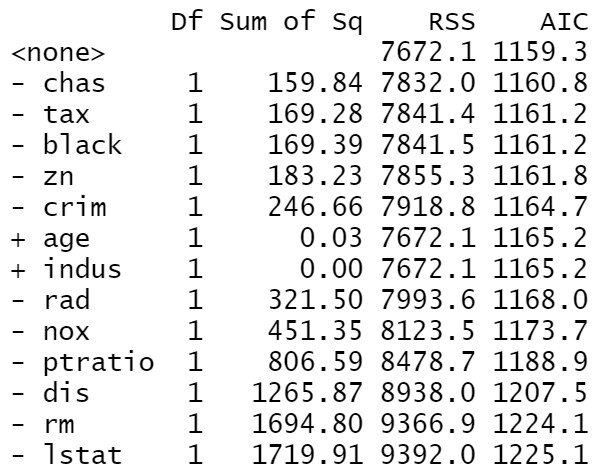
1. **Linear Regression**

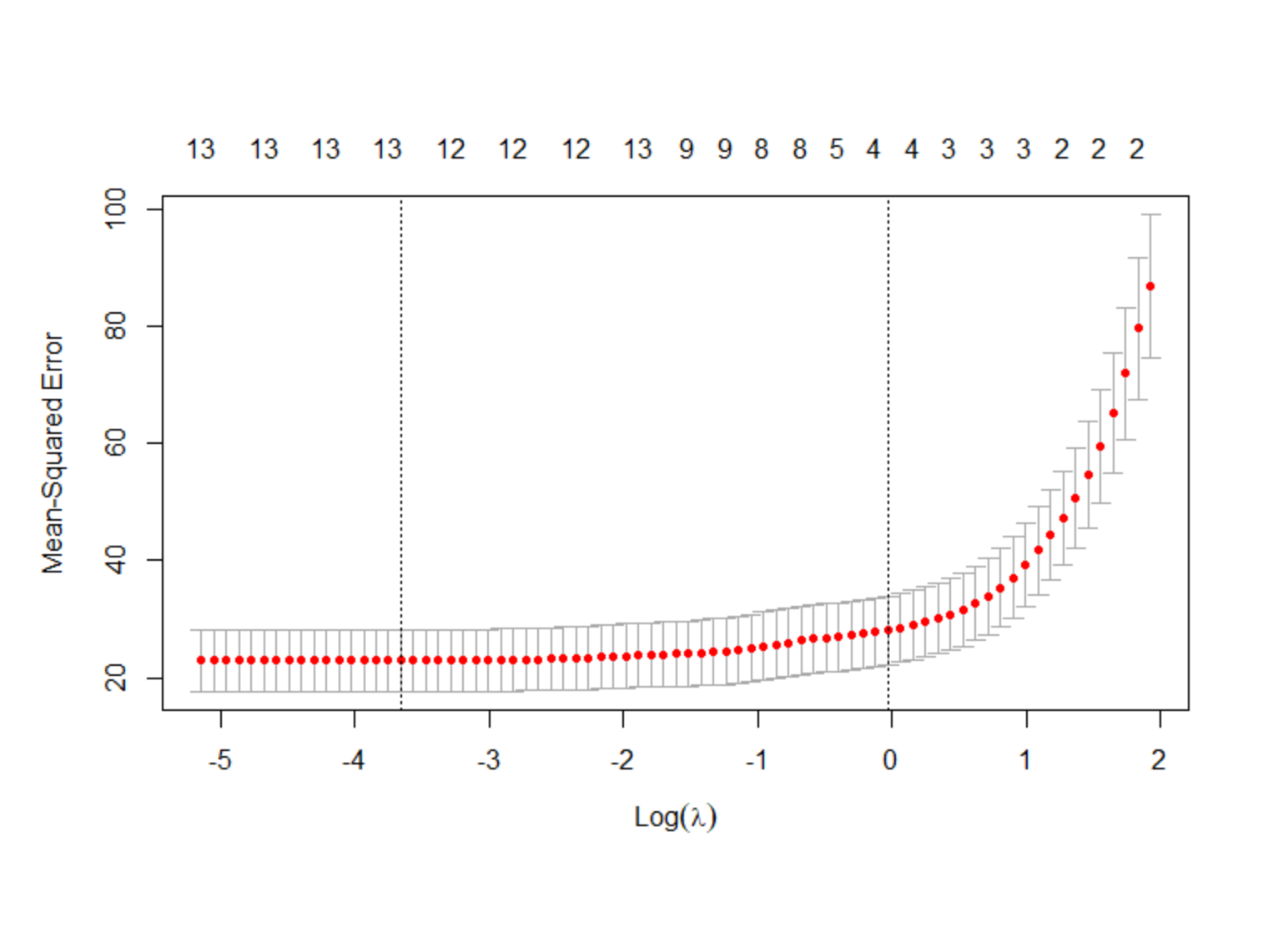
**Figure 1: Linear Regression with no Transformation**

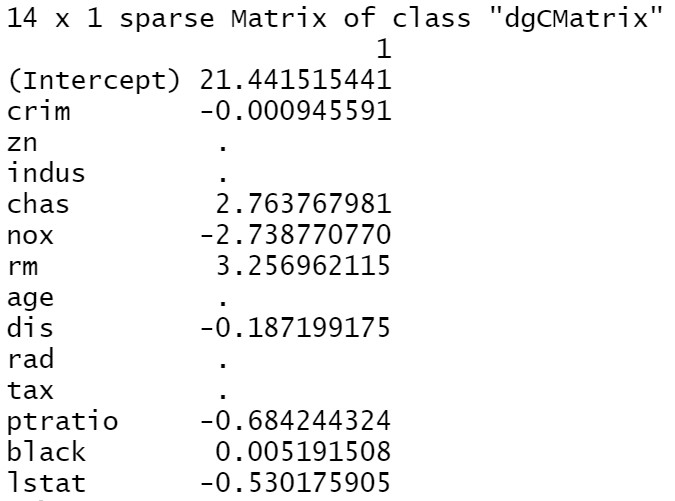
**Figure 2: Linear Regression (AIC – Stepwise Variable Selection)**



**Figure 3: Linear Regression (BIC – Stepwise Variable Selection)**



**Plot 1: Optimal Lambda Search for Lasso**

**Figure 4: LASSO using lambda = lambda.min**

**Report on 3 methods of variable selection:**

The 3 models creating using AIC and BIC as the criteria in stepwise regression and LASSO all preformed nearly equally in-regards-to MSE.

**MSE:**

*AIC – 22.115*

*BIC – 22.115*

*LASSO – 22.101*

**Out-of-Sample MSE:**

*AIC – 23.144*

*BIC – 23.114*

*LASSO – 23.091*

The variables I will select for my final model are the LASSO coefficients. I will choose these variables because it has a slightly lower out of sample MSE. This is important, because it is indicative of better predication on data outside of the sample.

The result that I got were interesting and a little surprising to me. I thought that since the correlation heat map in plot 3 indicated a lot of multicollinearity that a linear regression might be much more effective as a sparse matrix using LASSO regression. As expected, the variables removed by using LASSO did tend to be variables that were highly correlated as indicated by plot 3. The AIC and BIC stepwise regression, however, proved to have nearly equal mean squared errors when comparing the predictions with the actual test data set values.

Chosen Model: **medv = crim + chas + nox + rm + dis + ptratio + black + lstat**

When creating a model using only these variables, we get a slightly worse MSE plugging these into a linear model on their own.

**Out-of-sample MSE:**

*Chosen Model – 24.615*

**Cross Validation using chosen model variables:**

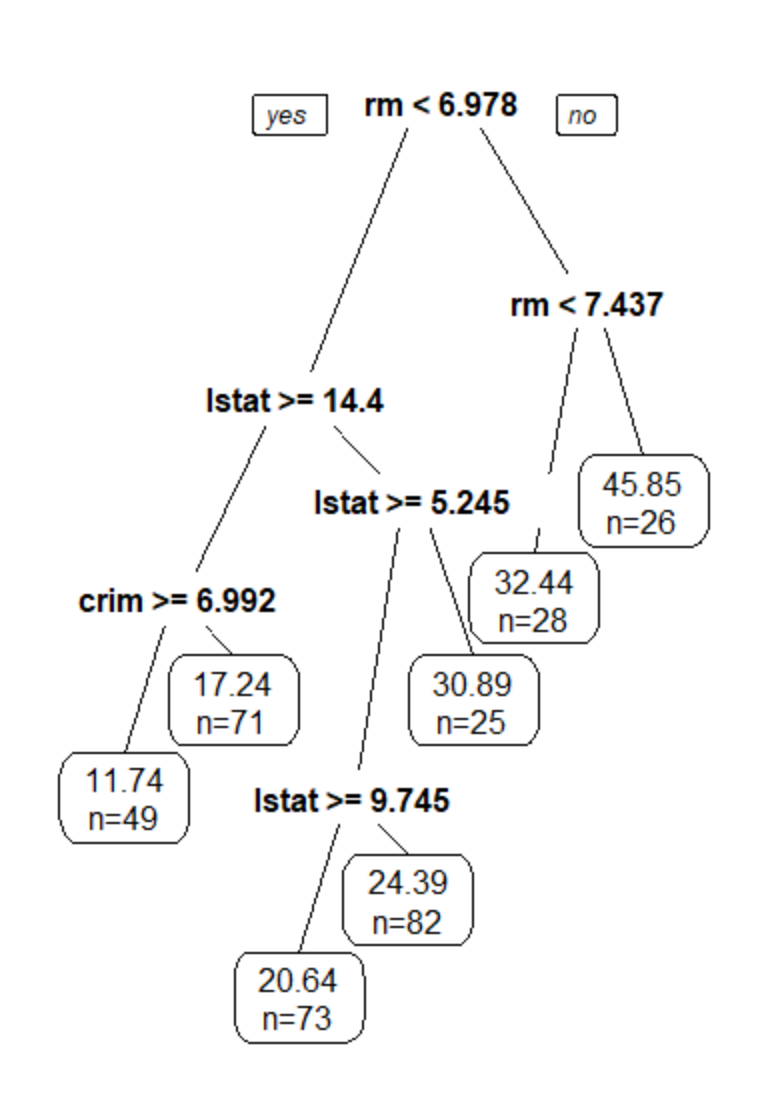
Here, I want to use 3 fold cross validation on our LASSO-selected variables with the full data set and compare the out-of-sample MSE from the chosen model without cross validation using just the training set.

**MSE:**

*CV LASSO – 24.282*

The first thing I notice when looking our the MSE for the cross validation model is lower. This makes sense because more data is being used to train the model which helps to increase its predictive power. The results are very similar, though.

**Regression Tree using chosen variables:**

**Figure 5: Regression Tree using chosen variables.**

**Regression Tree Out-of-Sample MSE:**

*Chosen Model – 23.67*

In fitting out model to a regression tree instead of a simple linear model as done in figure 2 and 3 and in homework, there is a slight improvement in the out-of-sample MSE. This indicates the regression tree does a slightly better job of creating a model for this data. This could be due to the data not being best defined by a strictly linear model rather than something nonlinear or that makes axis parallel splits such as decision trees.

**Starting over with a new training data set:**

I am going to resample the training data and repeat the previous steps to calculate the out-of-sample MSEs again.

**Out-of-Sample MSE:**

*AIC – 33.162*

*BIC – 34.313*

*LASSO – 33.171*

*Chosen Model – 34.216*

*Chosen Model (CV) – 24.579*

*Regression Tree – 32.308*

There are some similarities between our prior results and the new results, but there are also some significant differences. One difference is that the values for out-of-sample MSE are larger for each model except the cross-validated model. This is because the training data is randomly sampled and can be different between sampling procedures which causes different models to be built. This is not the case though with the cross-validation model, because the whole data set is used regardless.

Another difference between our model validations is that our stepwise variable section procedures chose to remove an additional variable from the model in the BIC criteria stepwise variable selection but not in the AIC with the second sampling. This is interesting because before they had chosen the exact same variables. This of course leads to the differing out-of-sample MSE values between the two.

The most significant consistency between our results from the different samplings is that the regression tree performed better than all the other models barring the cross validated model. It makes sense that this is the case, because it is likely a generally better algorithm for modeling this data set regardless of the sampling that occurs. In all likelihood, other forms of non-linear models will probably perform better because of the skewedness of many of our predictor variables.

**Conclusion:**

Two important things learned from this analysis are that the performance can be highly contingent upon the sampling that occurs when you are not working with a very large data set, and that the regression tree performs very well on the data compared to a simple linear model and the LASSO model. The latter was true in both cases of sampling. The former is interesting because it should indicate some hesitancy in assessing the integrity of MSE or other performance metrics that are not cross-validated, because it can fluctuate significantly depending on the training set that is sampled.

Also, we are able to decipher from the variable selection procedures which variables are potentially disposable in determining the true model of the data set. The stepwise selection procedures both removed age and indus while zn, rad, and tax were also deemed disposable by the LASSO regression. While other types of models may not find these variables to be disposable in predicting the median value of homes, it is probably the case that they are not necessary for linear models. In all likelihood, other forms of non-linear models will probably perform better because of the skewedness of many of our predictor variables.

**Bankruptcy Data:**

The bankruptcy data set contains 5436 observations and has a total of 13 attributes for each observation. The purpose of this data set is to use the R1-R10 variables to predict the likelihood of bankruptcy. Below is a list of the variables with descriptions and data types:

DLRSN = Credit; (Binary)

R1=Working Capital/Total Asset; (Numeric)

R2=Retained Earning/Total Asset; (Numeric)

R3=Earning Before Interest & Tax/Total Asset; (Numeric)

R4=Market Capital / Total Liability; (Numeric)

R5=SALE/Total Asset; (Numeric)

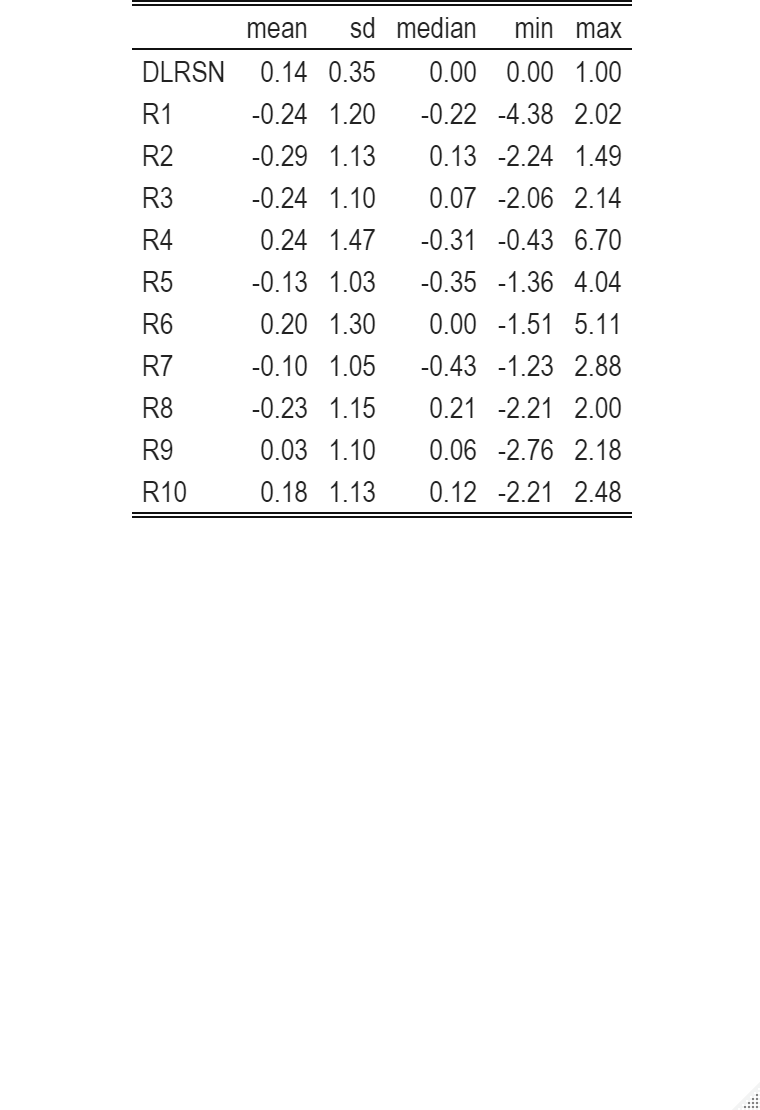
R6=Total Liability/Total Asset; (Numeric)

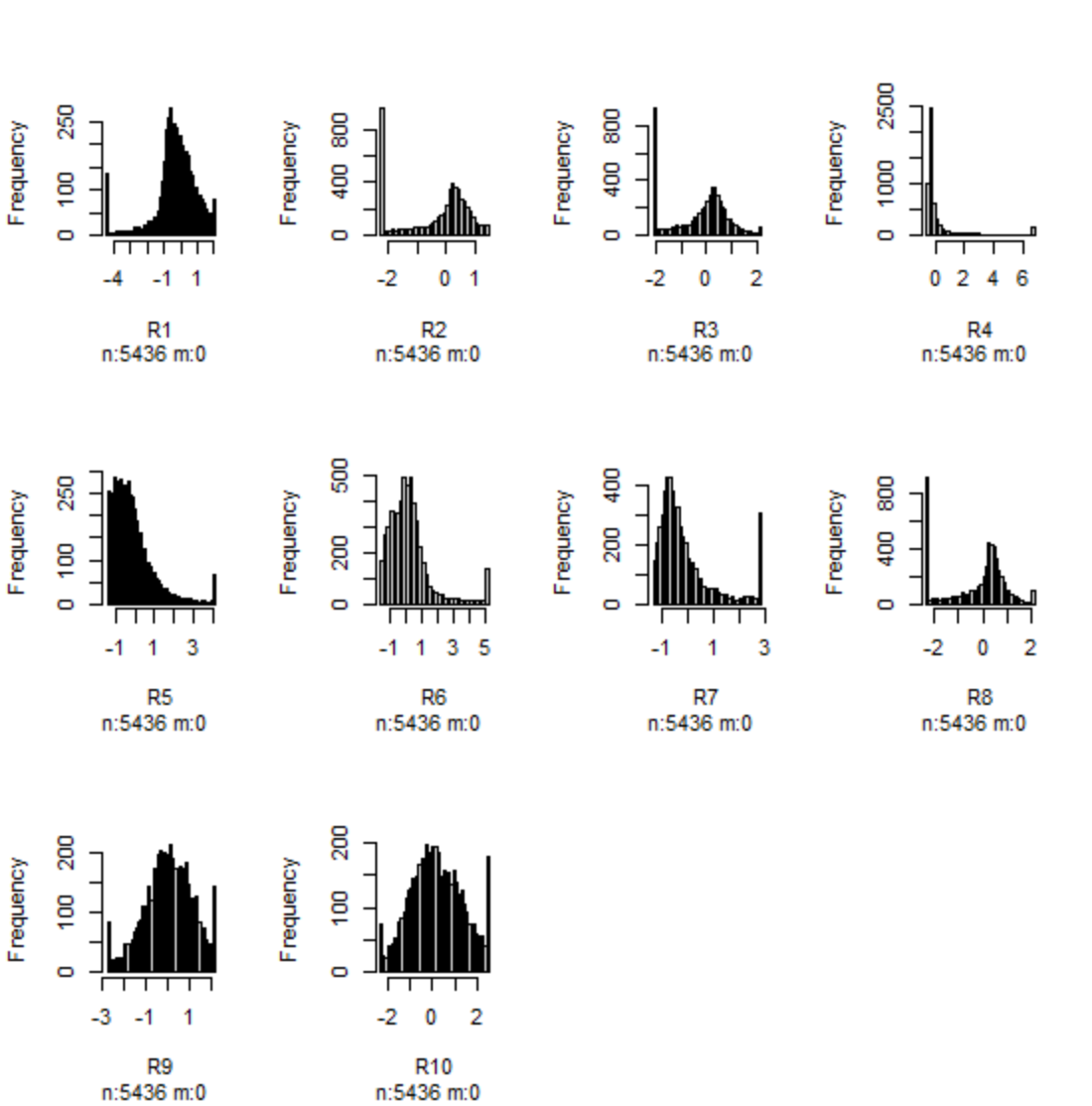
R7=Current Asset/Current Liability; (Numeric)

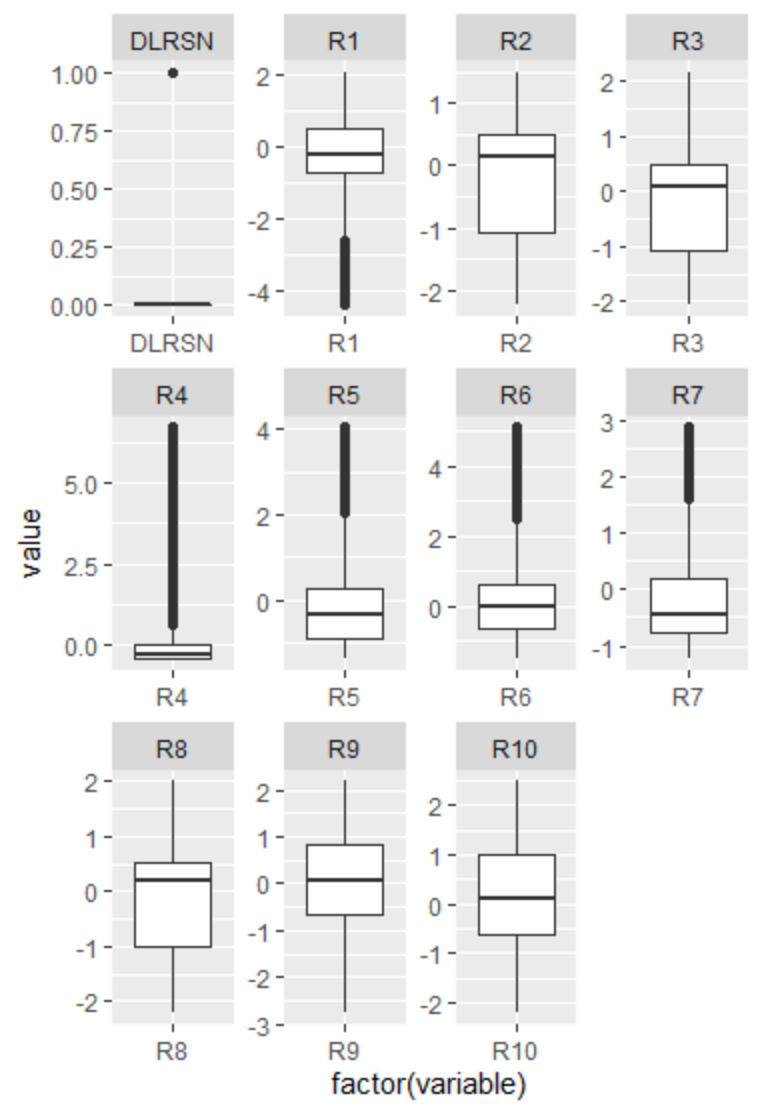
R8=Net Income/Total Asset; (Numeric)

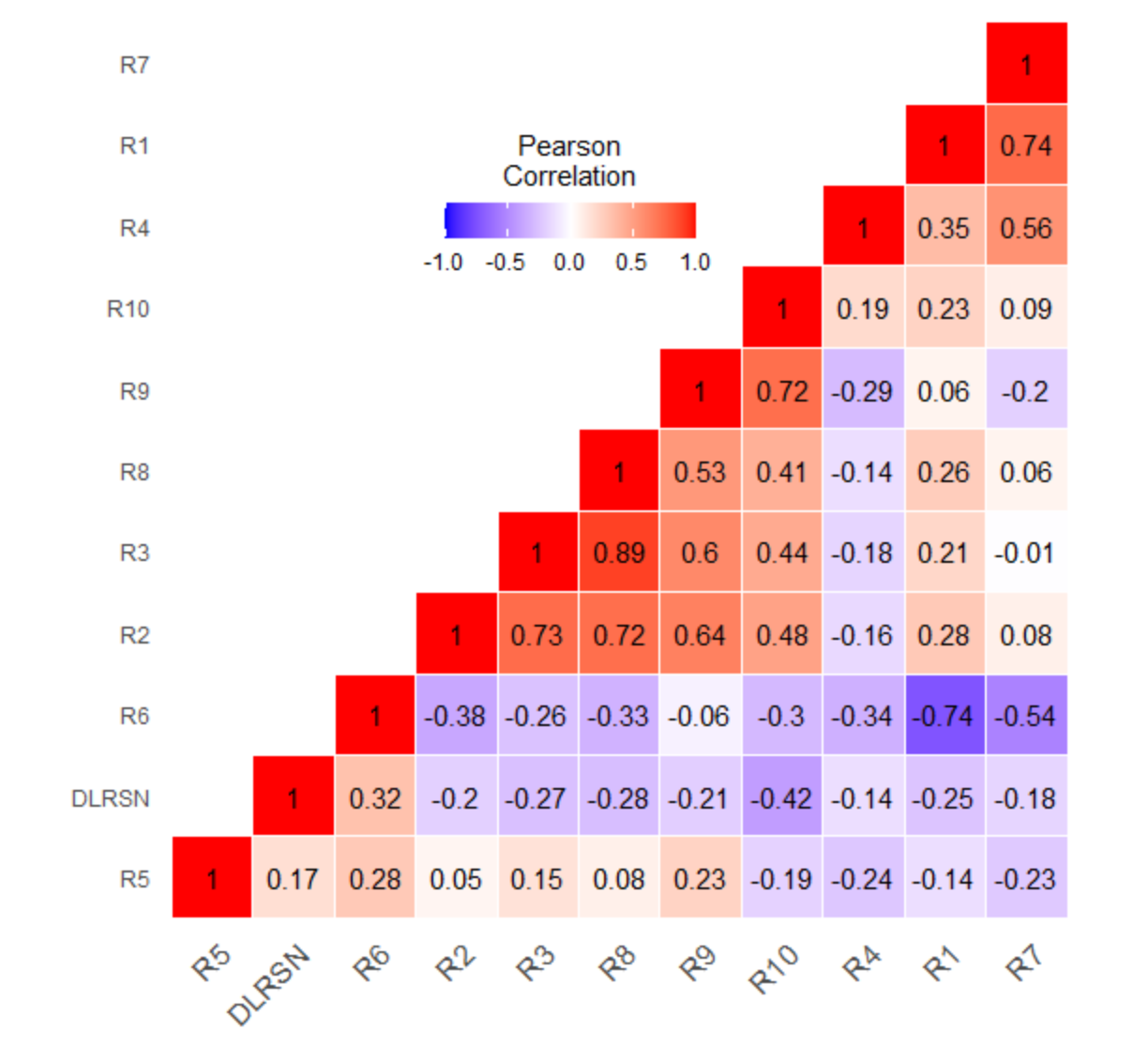
R9=LOG(SALE); (Numeric)

R10=LOG(Market Cap); (Numeric)

**Figure 6: Summary Statistics for Bankruptcy Data**

**Plot 4: Histograms of Bankruptcy Data Variables**

**Plot 5: Box Plots of Bankruptcy Data Variables**

**Plot 6: Correlation Heat Map for Bankruptcy Data**

**EDA Report:**

In looking at plot 4, we can see that most of our predictor variables are slightly skewed or near normal distributions. R1, R5, R6, and R7 are all quite skewed to the right, R4 seems to follow some sort of exponential distribution, and the others follow a near normal distribution. Judging by the histograms and summary statistics, it does seem like we have some outliers or spikes on the far end of the distributions. The boxplots (plot 5) shown also reveal skewedness and significant ouliers because of the many points that lie beyond the whiskers. For example, R8 has many observations near -2 and 2 which are on the far ends. Otherwise, it follows a normal distribution. Other variables follow a similar theme. R4 has some very significant outliers on the far-right end. The mean is 0.24 while the max value is all the way up at 6.40. Another common way of perceiving the irregularities in the data is looking at the differences between the mean and median. Many of the variables’ means and medians are quite far apart further reinforcing our prior observations about irregular distributions and skewedness.

The correlation head map in plot 6 reveals that we have many notable correlations present in the data. The most notable negative correlations being R1-R6, and R7-R6. The most notable positive correlations are R3-R8. R2-R3, R2-R8, R2-R9, R3-R9, R8-R9, and R9-R10. It makes perfect sense that we have many variables that are highly correlated because, as indicated by the data description. Many of our variables are built from the same data. For example, R1-2-3-5-6 and 8 all contain Total Asset in the denominator. This fact must be what leads to most of the correlation in the data. Also, generally in finance, many data points are going to be correlated because there is often some connection between different types of earnings and accumulations.

**Logit, Probit, Cloglog:**

**Significant Coefficients:**

*Logit – LIMIT\_BAL, SEX2, EDUCATION2,3,4, MARRIAGE2, AGE, PAY0 and 2, BILL\_AMT1, PAY\_AMT1 and 2*

*Probit – LIMIT\_BAL, SEX2, EDUCATION2,3,4, MARRIAGE2, AGE, PAY0 and 2, BILL\_AMT1, PAY\_AMT1 and 2*

*Cloglog - LIMIT\_BAL, SEX2, EDUCATION2,3,4, MARRIAGE2, PAY0, BILL\_AMT1, PAY\_AMT1 and 2*

The significant coefficients each model gave were the same except cloglog did not have AGE or PAY2 as significant variables. One thing I noticed is the coefficients and standard errors all tended to be slightly smaller for the probit model than logit and cloglog. Probit had the lowest AIC at 11099.

**Variable Selection For Best Model:**

AIC Variable Selection:

* LIMIT\_BAL + SEX + EDUCATION + MARRIAGE + PAY\_0 + PAY\_2 + PAY\_4 + PAY\_6 + BILL\_AMT1 + BILL\_AMT2 + BILL\_AMT4 + BILL\_AMT5 + PAY\_AMT1 + PAY\_AMT2 + PAY\_AMT4
* AIC = 7749
* Out-of-Sample MSE = 3.87

BIC Variable Selection:

* LIMIT\_BAL + PAY\_0 + PAY\_2 + BILL\_AMT1 + BILL\_AMT2 + PAY\_AMT1
* AIC = 7832
* Out-of-Sample MSE = 3.68

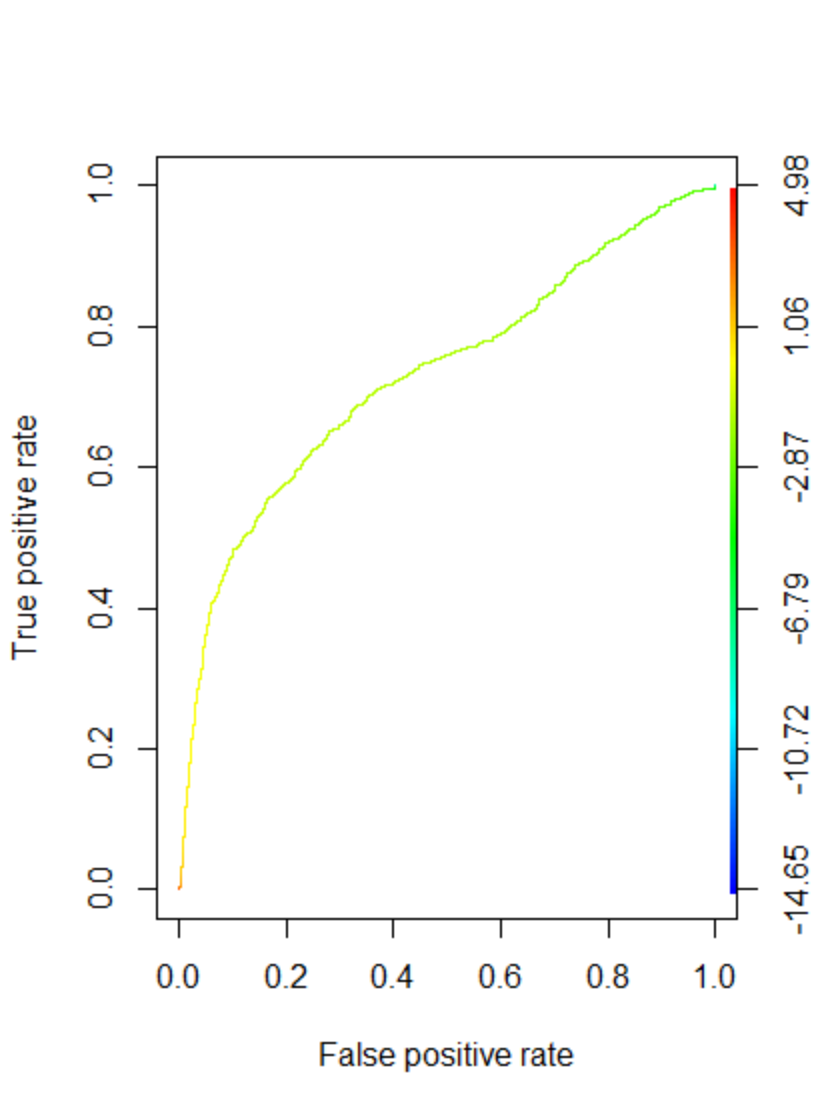
LASSO variable selection:

* LIMIT\_BAL+SEX+EDUCATION+MARRIAGE+AGE+PAY\_0+PAY\_2+PAY\_3+PAY\_4+PAY\_5+PAY\_6+ PAY\_AMT1+PAY\_AMT2+PAY\_AMT4+PAY\_AMT5
* AIC = 7782.5
* Out-of-Sample MSE = 3.83

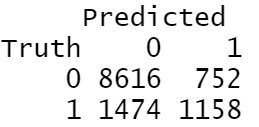
The model that will be chose here is the BIC Variable Selection model, because it has the lowest out-of-sample MSE. This is important, because it is indicative of better prediction power on observations not used to train the model.

* **Final Model: Default = LIMIT\_BAL + PAY\_0 + PAY\_2 + BILL\_AMT1 + BILL\_AMT2 + PAY\_AMT1**

**Plot7: ROC Curve for Final Model**

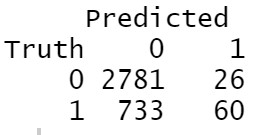
**Final Model AUC: 0.72**

**Figure 7: Confusion Matrix Using Final Model (0.5 cutoff)**



For our final model, the ROC curve looks pretty good. The TPR increases sharply in the beginning, but the returns start to diminish sharply is a the FPR increases past 1.5. At a cutoff of 0.5 for classification, the AMR = 0.19.

**Figure 8: Confusion Matrix – Out-of-Sample (0.5 cutoff)**



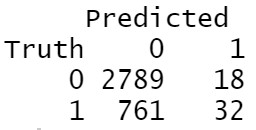
**AUC for out-of-sample performance: 0.73**

**AMR = .22 tat 0.5 cutoff**

**Optimal Cutoff Probability:**

Using a 5:1 asymmetric cost ratio for determining the optimal cutoff for default prediction of our final model on the credit data set, I found that choosing a cutoff of 0.92 would be optimal.

**Figure 9: Confusion Matrix – Out-of-Sample (.92 cutoff)**

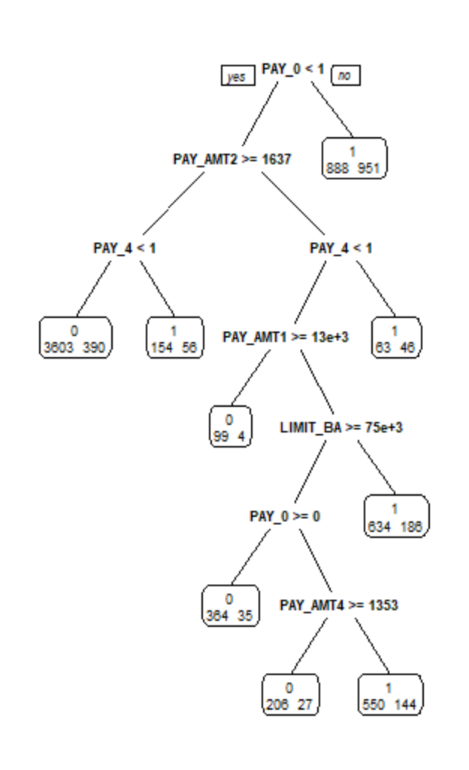


**Cross-Validation:**

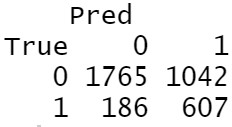
I fit the final model to a 3-fold cross validation to test the effectiveness of a cross-validated model. The result was that the AMR = 0.14. This is 8% better than the out of sample classification rate without cross validation. This is due to cross validation using all of the data which improves its predictions, and all the data is in-sample so it is predicting data used to train the model.

**Classification Tree:**

**Figure 10: Classification Tree – Final Model**



**Figure 11: Confusion Matrix – Classification Tree (Out-of-Sample)**



Our classification tree has an AMR of 0.34. This is worse than both the variable selected model without cross-validation and with cross-validation.

**Comparison to Logistic Regression from HW3:**

In comparing our result with Homework 3. I had similar results in that the AMR tended to be lower for out-of-sample predictions. This makes sense because the out-of-sample data points are not used to build the model. It is hard for me to do much more comparison, because I used a different data set here than in homework 3.

**Re-sampling Using 90% Of Data For Training:**

AIC Variable Selection:

* LIMIT\_BAL + SEX + EDUCATION + MARRIAGE + AGE + PAY\_0 + \_2 + PAY\_3 + PAY\_5 + BILL\_AMT1 + BILL\_AMT2 + BILL\_AMT4 + BILL\_AMT5 + PAY\_AMT1 + PAY\_AMT2 + PAY\_AMT4 + PAY\_AMT6
* AIC = 9990
* Out-of-Sample MSE = 3.91

BIC Variable Selection:

* LIMIT\_BAL + SEX + AGE + PAY\_0 + PAY\_2 + PAY\_5 + BILL\_AMT1 + BILL\_AMT2 + PAY\_AMT1 + PAY\_AMT2
* AIC = 10095
* Out-of-Sample MSE = 3.86

LASSO variable selection:

* LIMIT\_BAL+SEX+EDUCATION+MARRIAGE+AGE+PAY\_0+PAY\_2+PAY\_3+PAY\_4+PAY\_5+PAY\_6+ BILL\_AMT1 + PAY\_AMT1+PAY\_AMT2+PAY\_AMT4+PAY\_AMT5 +PAY\_AMT6
* Out-of-Sample MSE = 3.89

**AMR For 70:30 Split:**

* Full Model = 0.19
* Full-Model (Out-of-sample) = 0.22
* CV Model = 0.14
* Classification Tree = 0.34

**AMR For 90:10 Split:**

* Full Model = 0.21
* Full-Model (Out-of-sample) = 0.21
* CV Model = 0.14
* Classification Tree = 0.38

**Conclusion:**

There were a few significant differences between the samples. The 90:10 split tended to choose more variables in the variable selection procedure and the AIC values tended to be higher with this split as well. The variable selection procedures probably chose more variables because adding more data to training the models likely made the variables more significant or impactful. The AIC is higher simply because there is more data. The MSEs for the 90:10 split also tended to be just a little bit higher, though relative to each other, they performed the same. The AMR did not seem to change much between the two samples. What this implies to me is that the classification rate is likely not significantly impacted between the two samples, because there is already a lot of data. There is probably more variation in the classification rates due to the variance that occurs from random sampling than from choosing different sized samples.