**Boston Housing Dataset**

**EDA and Regression – Assignment 2**

**Chris Steege and Shuki Saito**



**Exploratory Data Analysis:**

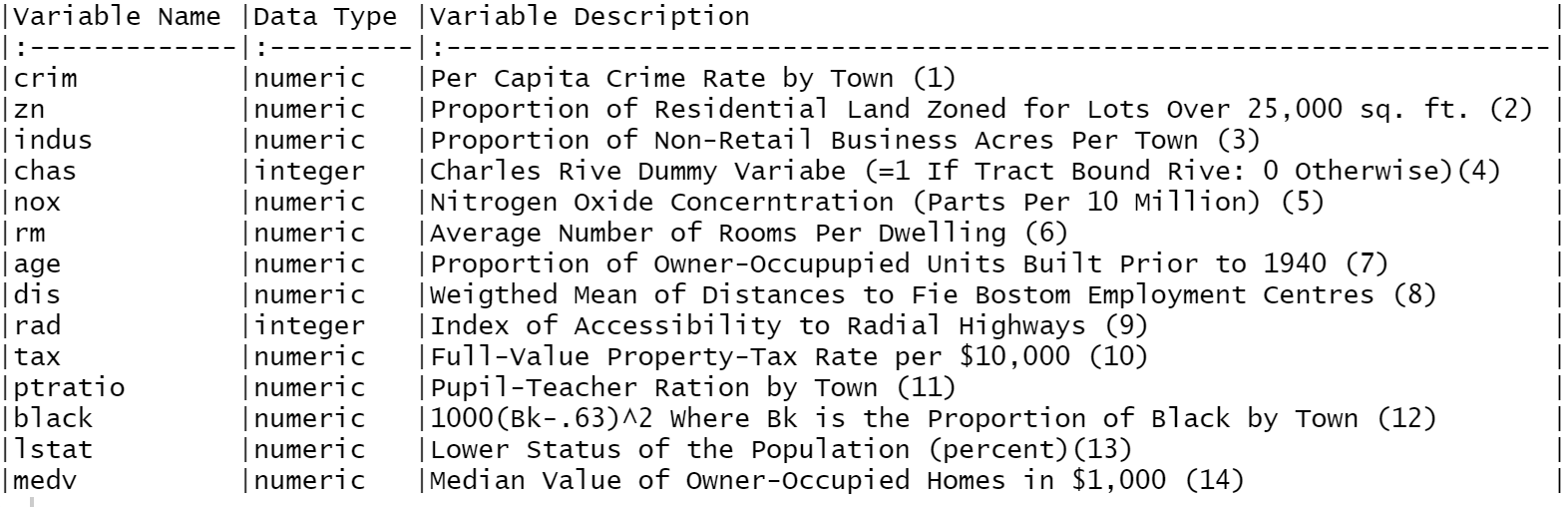
1. **Background**

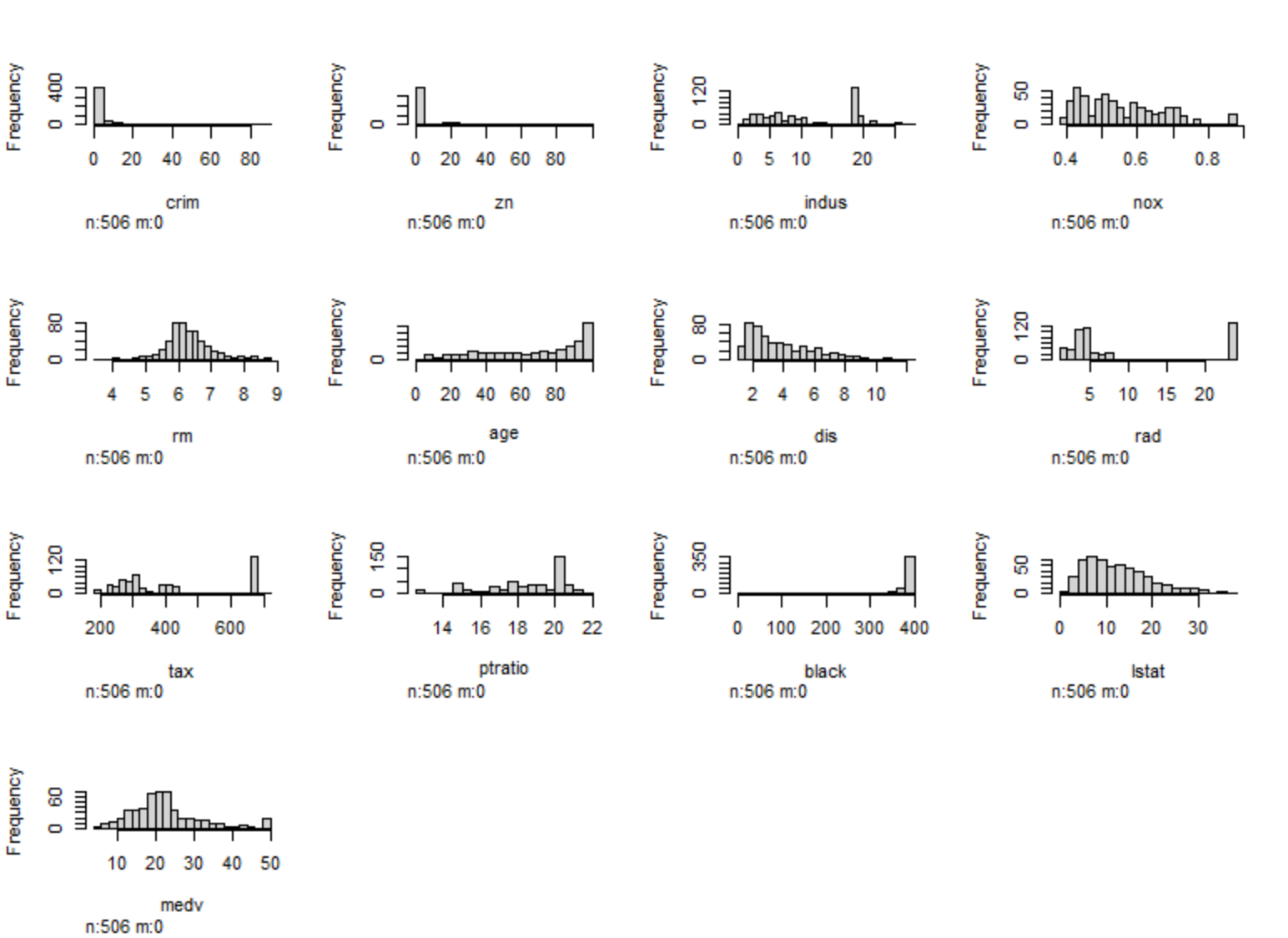
The Housing Data set was used in the 1976 publication *Hedonic Housing Prices and Demand for Clean Air.* The purpose of the data collection and analysis is to discover the nuances of housing market value with respect to air quality. The authors of this paper wanted to create an equation to model housing 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.

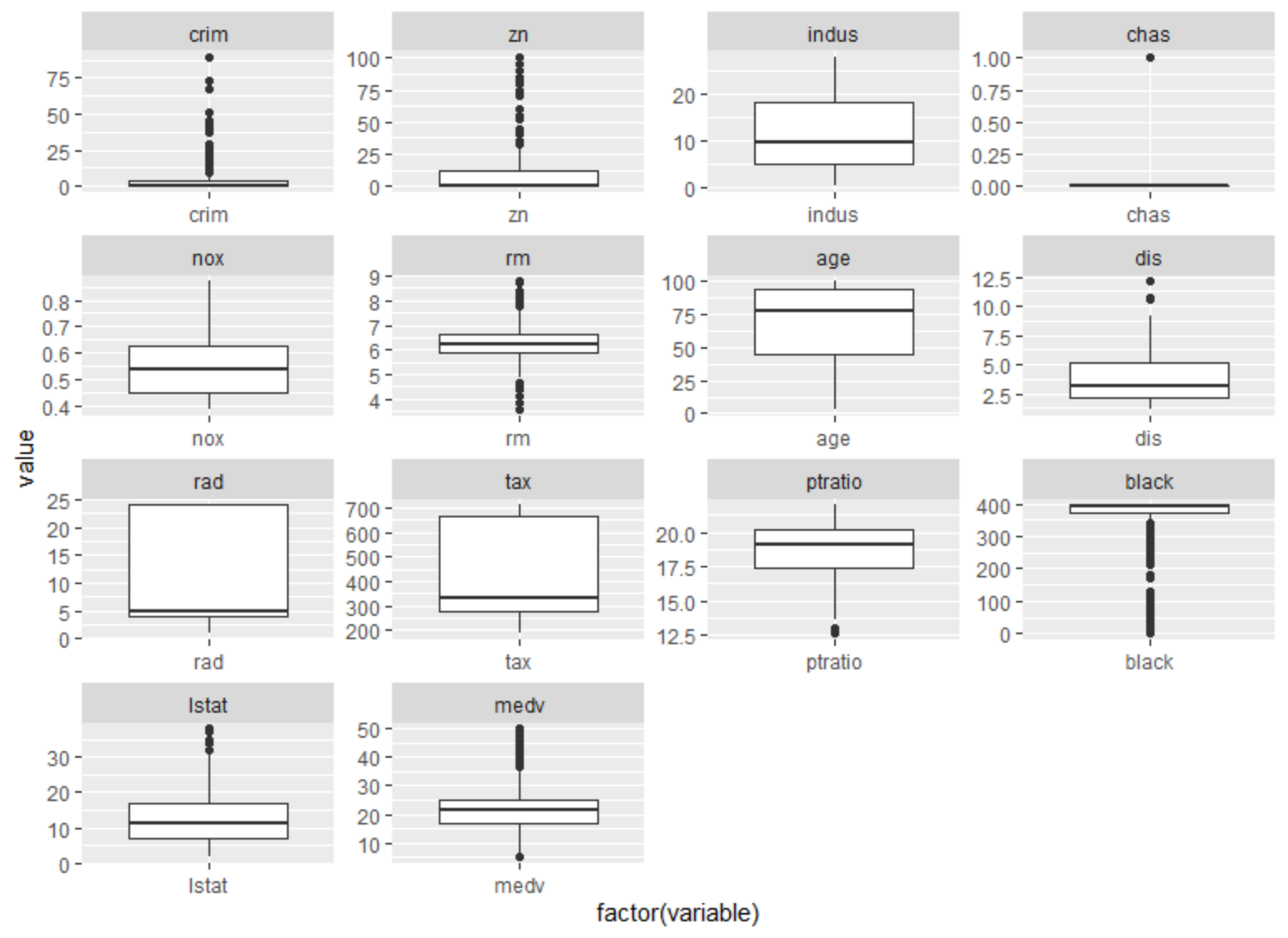
1. **Summary and Statistics**

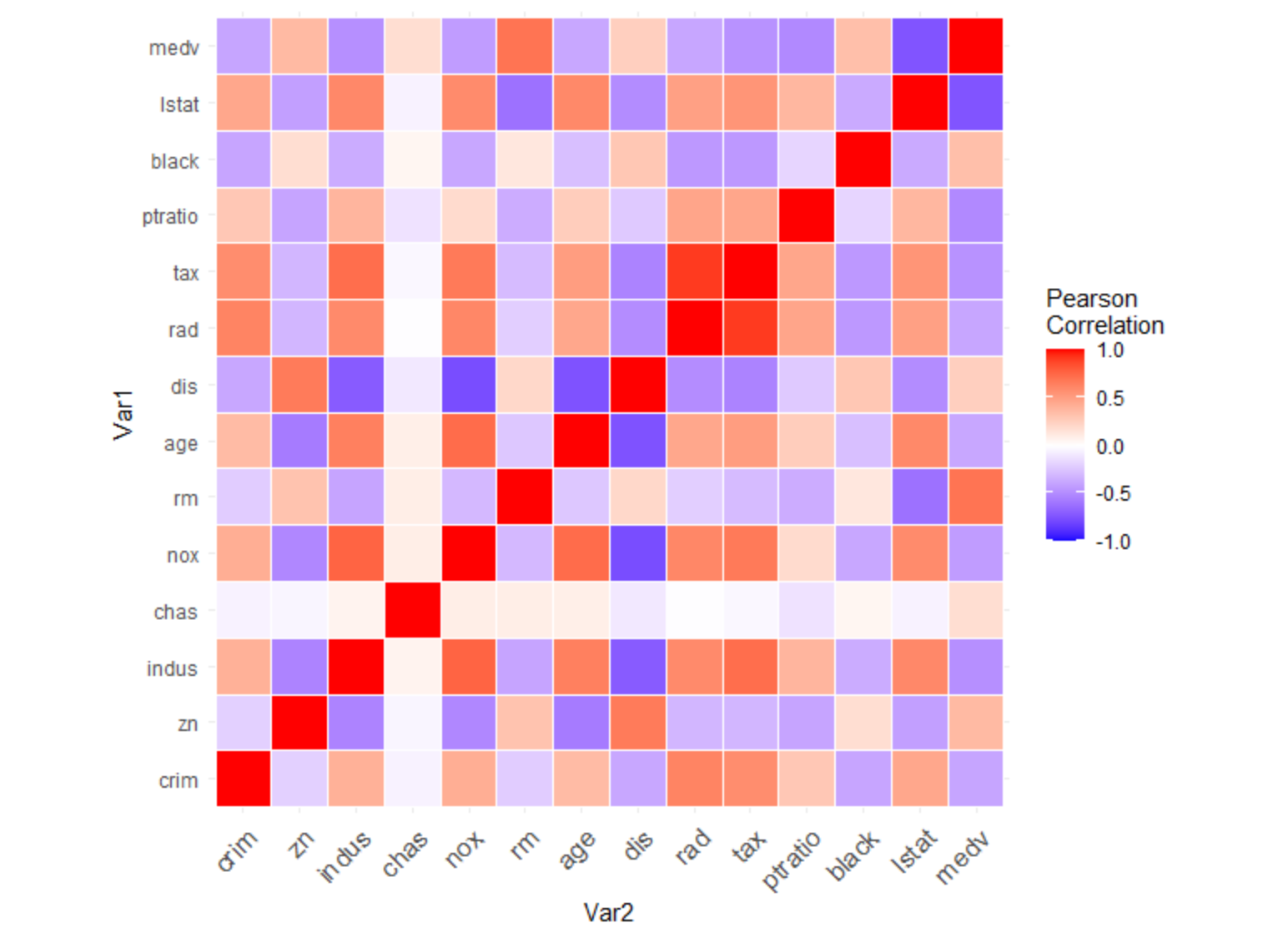
In order to understand the data we are dealing with, we created a table with variable descriptions and the count of attributes and observations. We are working with 506 neighborhoods with 14 attributes associated with each. We have 12 numeric continuous variables and 2 integer variables – one binary 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**

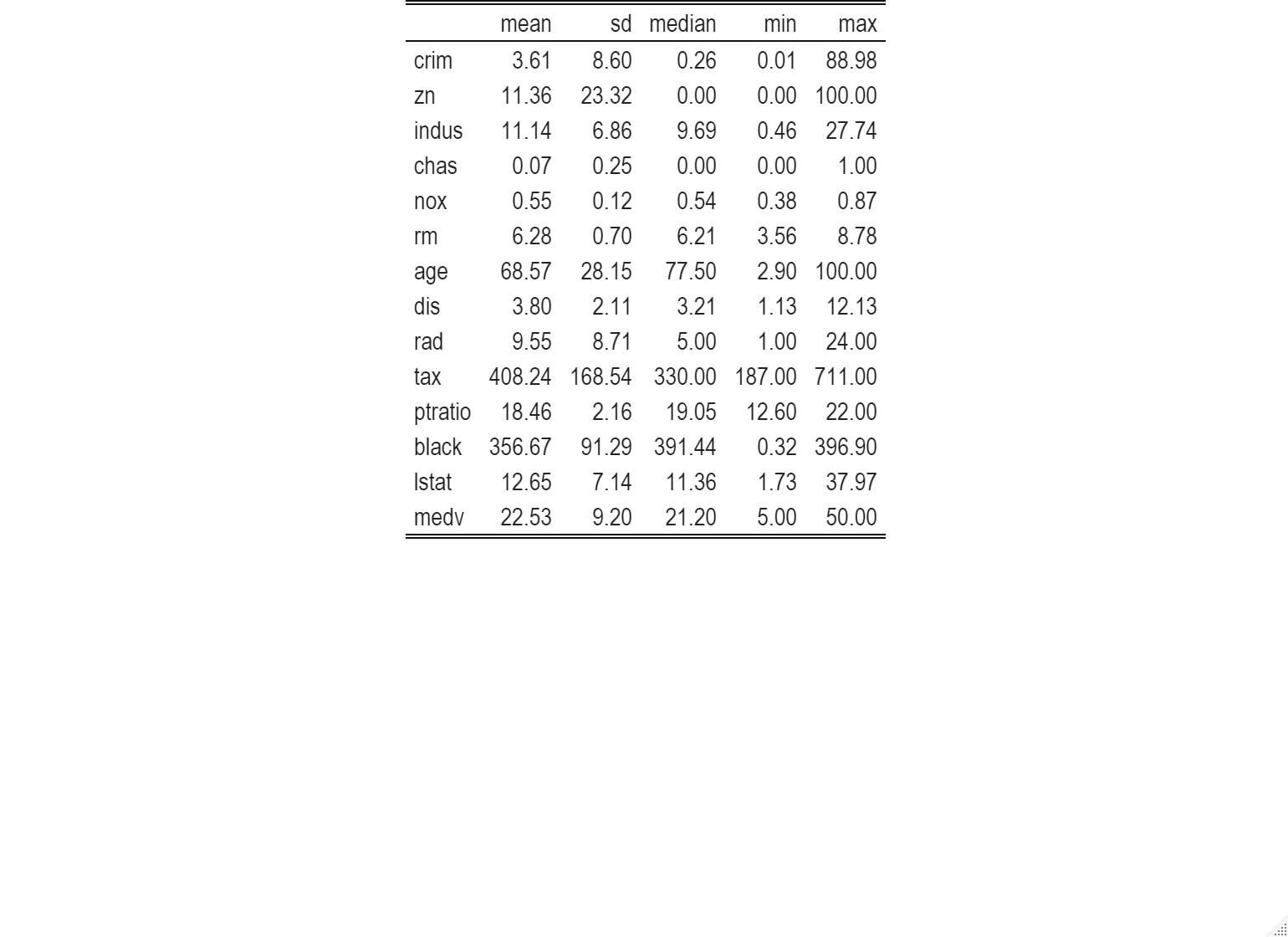


**Plot 1: Histogram of Housing Attributes**

**Plot 2: Boxplot of Housing Attributes**

**Plot 3: Correlation Heat Map**

**Table 2: Summary Statistics**



**Report:**

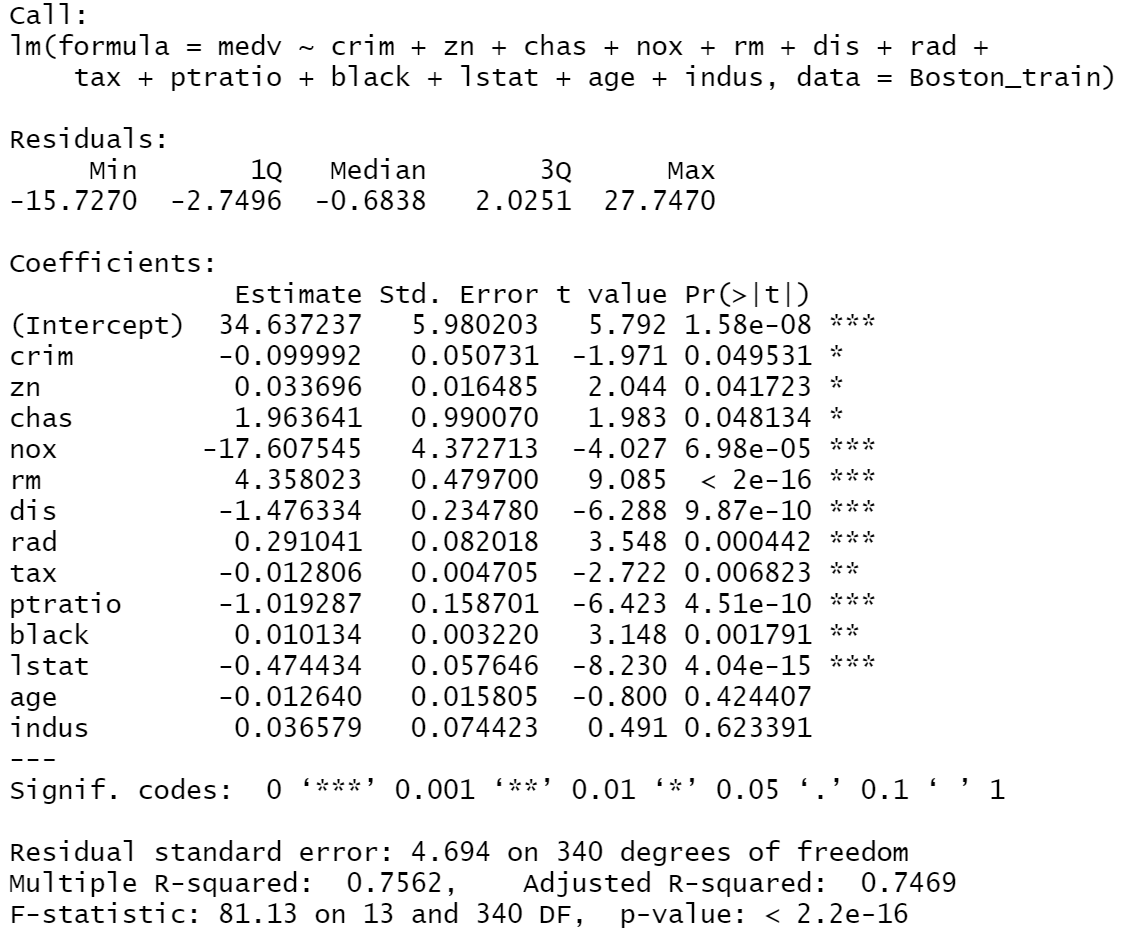
In looking at plot 1, it is clear the much of the data is highly skewed with irregular patterns across the spectrum of continuous data points. Rm and medv are the variables which are closest to a normal distribution. Dis, nox, and lstat appear to follow something like a power log normal distribution. 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.

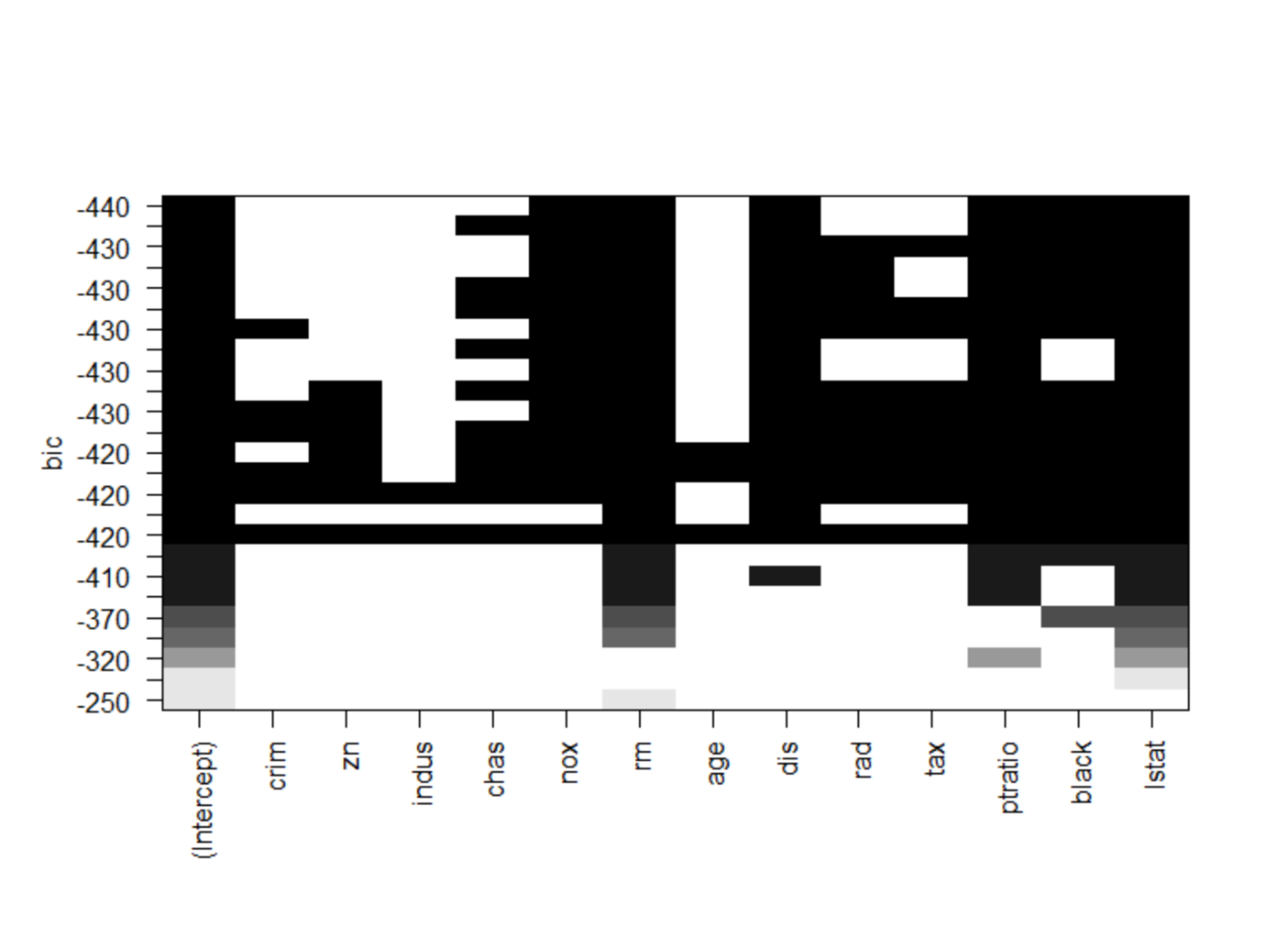
1. **Linear Regression (No Variable Transformation)**

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

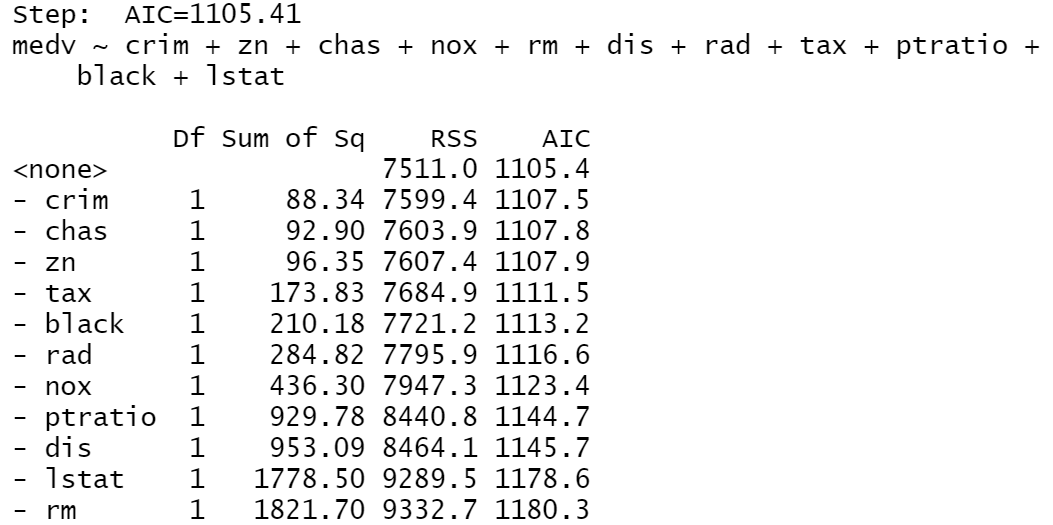


1. **Linear Regression (subset, stepwise, LASSO)**

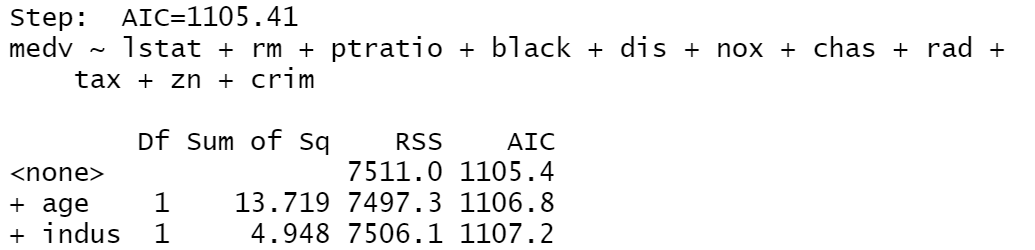
**Plot 4: Best Subset Regression Variable Selection**



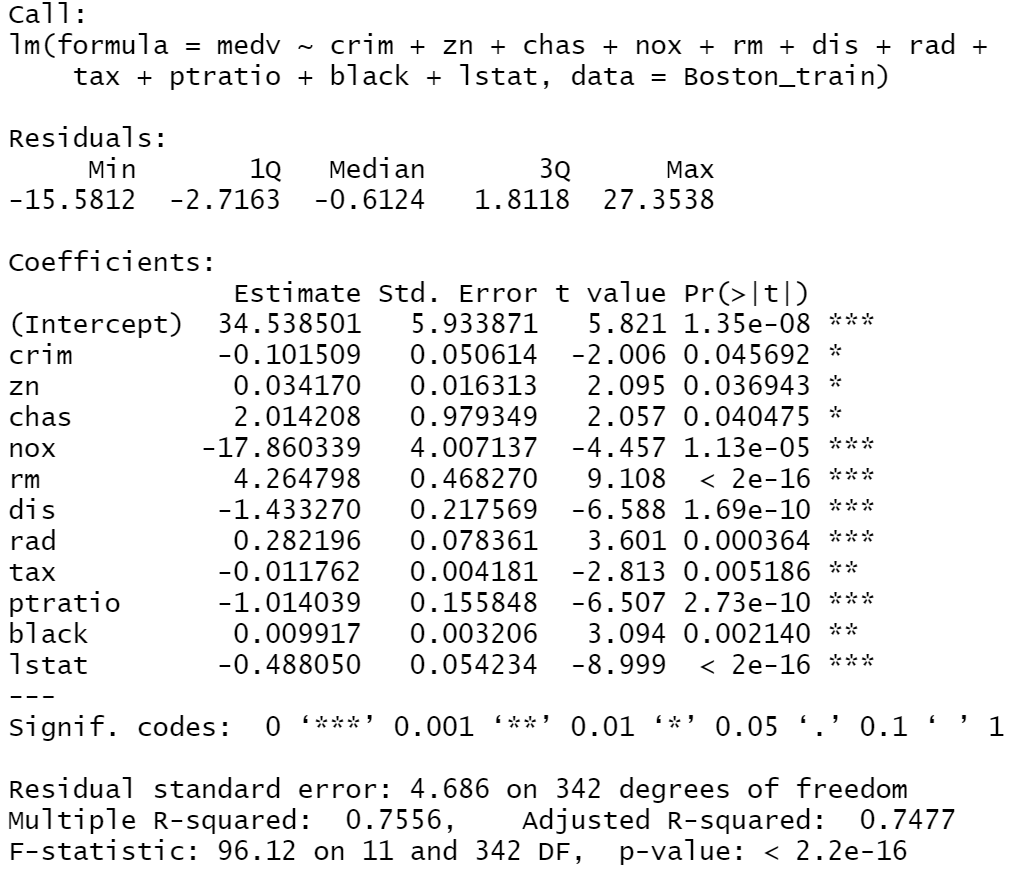
**Figure 2: Stepwise Backward Selection**

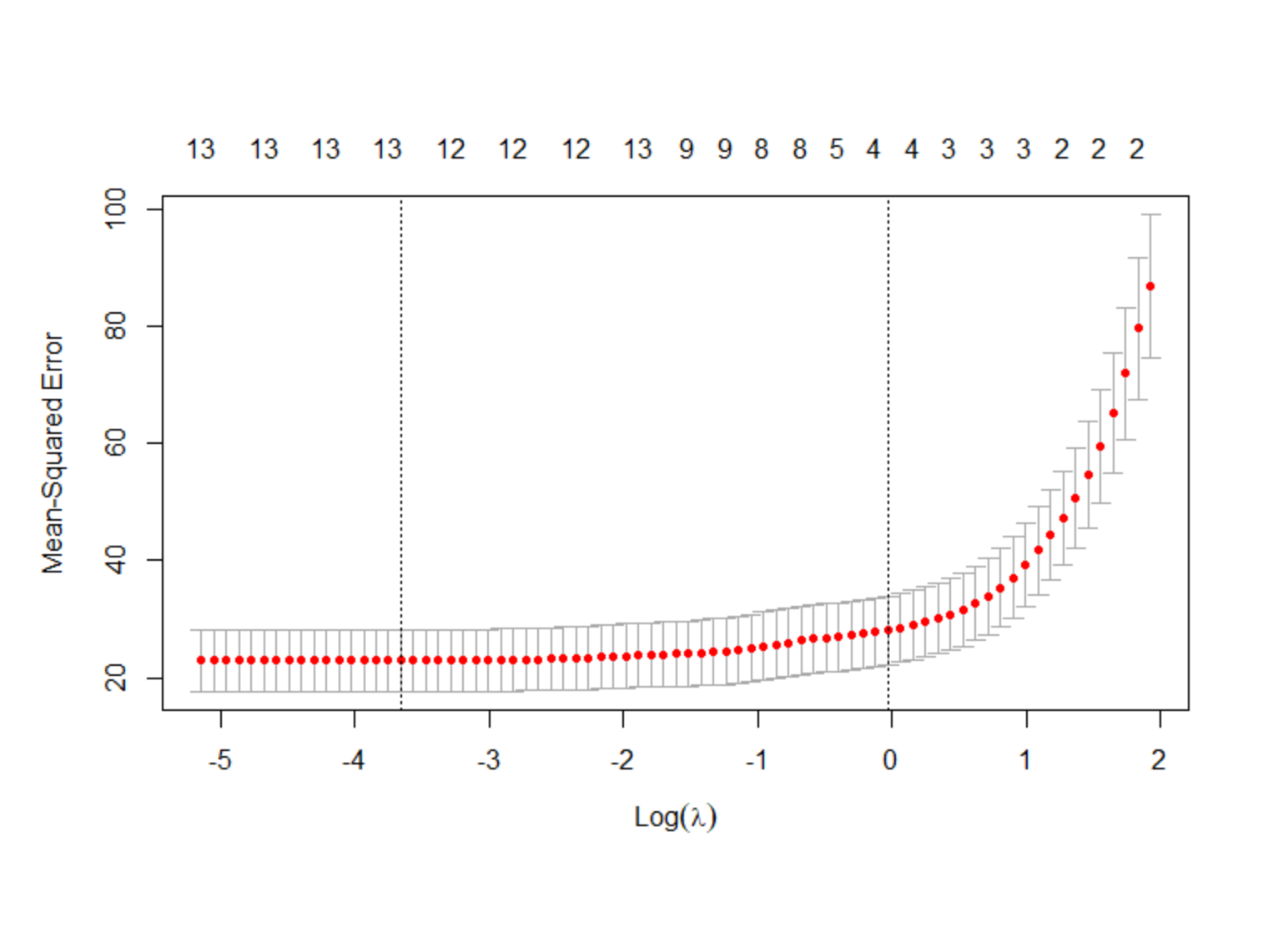


**Figure 3: Stepwise Forward Selection**

****

**Figure 4: Linear Regression Without Age and Indus**



**Plot 4: Optimal Lambda Search for Lasso**

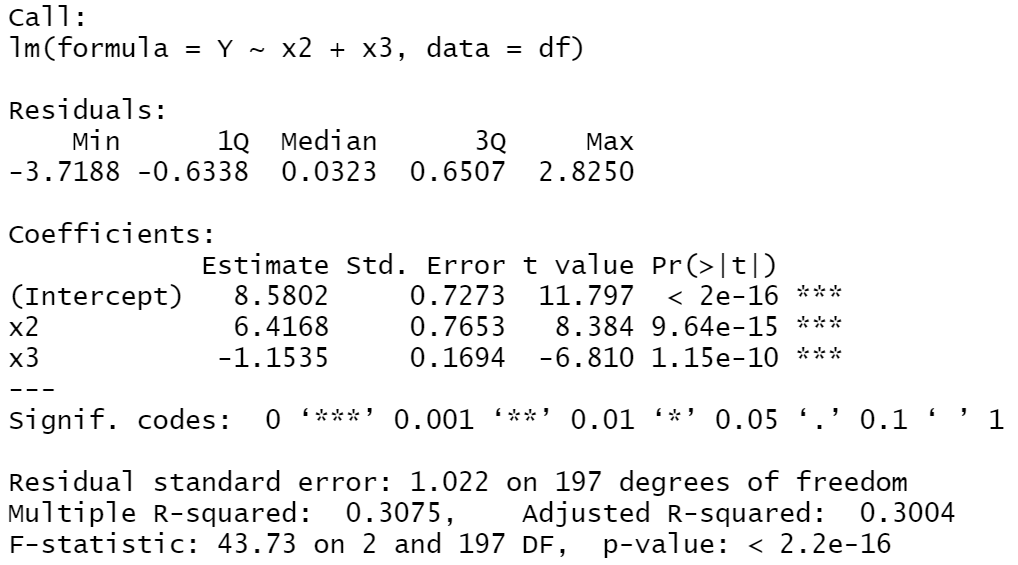
**Report:**

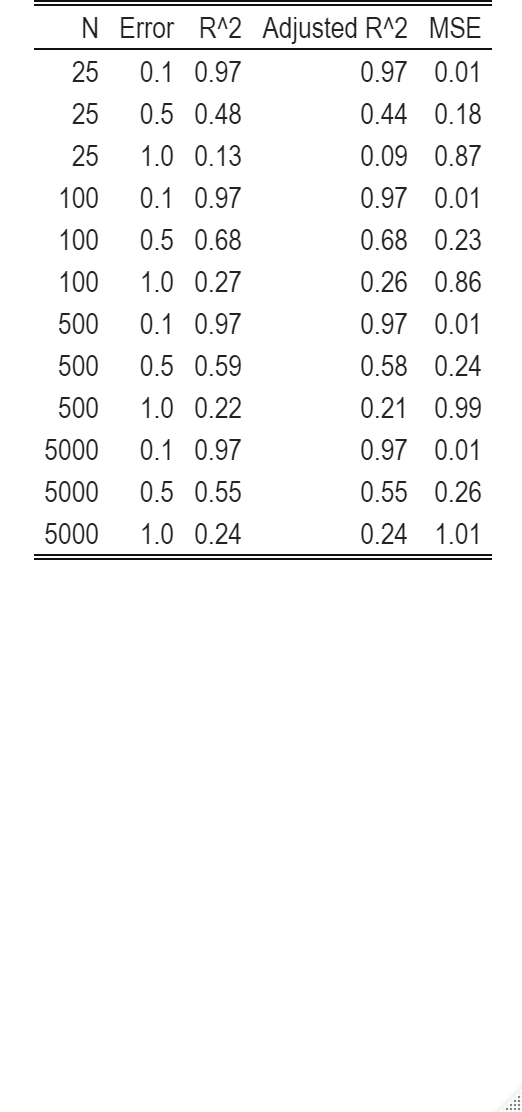
The worst performing model for MSE of the among the no variable transformation, stepwise, and LASSO was the LASSO variable selection model. The stepwise variable selection models and the model with no variable transformation both performed equally well in regards to MSE. LASSO had an MSE of 24.4 using lambda.min, no transformation had an MSE of 24.37, and stepwise had an MSE of 24.07. The model I would choose to use are the stepwise models. As indicated from figure 2 and 3, they both ended up selecting the same set of predictor variables by simply leaving out age and indus. The MSE again was 24.07, the R^2 was 0.76, and the adjusted R^2 was 7.5.

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 more effective as a sparse matrix using LASSO regression. The forward and backward stepwise regression, however, proved to have a lower squared error when comparing the predictions with the actual test data set values. Notably the models did not appear to have a significant difference in MSE or adjusted R^2. As indicated by plot 4 for our best subset regression model, we can expect a better model with more variables. For optimal performance, it seems that having nearly every variable is best, but taking out the age and indus variables optimized it a little further with the stepwise regression.

**A Simulation Study:**

**Figure 5: Backward Variable Selection – n=200**



**Table 3: Backward Variable Selection – Adjusting Parameters**

**Report:**

My first observation when comparing the models with varying sizes and error is the model performance fluctuates significantly based on the size of the error term. As larger error term generally gives a smaller R^2 and adjusted R^2 while causing an increase in the mean squared error. This is persistent across all sample sizes.

It is difficult to determine any other trends across increasing number of observations and error, but I speculate that when the observation number is higher, adding additional observations would not lead to as significant of fluctuation in our performance metrics, because the larger sample size makes the model more robust.

The decreasing performance with a higher error makes sense, because it makes the model less likely to fit a linear trend with low residuals. As a result, our model will does not do as well in explaining the variance of the response variable. Larger residuals also will lead to a larger MSE, because data points are likely to be further away from the regression line in a geometric space.

**Monte Carlo Simulation Study (Linear Regression) :**

Linear Regression models are a critical way to understanding a pattern of a dataset, it can be used to predict a dependent response value using independent variables. However, for prediction, the value of each beta will vary in the real world.

In the model for y, y = 5 + 1.2\* x1 +3\* x2, each beta for x1 and x2 will vary in our 200 sample date set. Monte Carlo Simulation Study is one of the methods to find the range of each beta’s variance.

The result of Monte Carlo Simulation Study is the followings,

* Mean of Interception, beta(x1) and beta(x2):   
  Mean of Beta^0 = 5.1022   
  Mean of Beta^x1 = 1.1974   
  Mean of Beta^x2 = 3.0973

* Standard Deviation of Interception, beta(x1) and beta(x2):   
  Sd of Beta^0 = 0.7935   
  Sd of Beta^x1 = 0.0308   
  Sd of Beta^x2 = 0.6126

This is interpreted as x0 is normally distributed with mean of 5.1022 and Standard Deviation of 0.7935, so does x1 with mean of 1.1974 and Standard Deviation of 0.0308, and x2 with mean of 3.0973 and Standard Deviation of 0.6126.

The model for linear coefficients is:

y^=5.1022 + 1.1974\*x1^ + 3.0973\*x2^

The comparison table for y, y^, and the difference between y and y^ with 50 random values of each covariance is the below:

**Table 4: Monte Carlo Simulation – Actual vs. Predicted**

|  |  |  |
| --- | --- | --- |
| Y | Y^ | |Y^-Y| |
| 4.6653 | 4.6563 | 0.0090 |
| 3.3793 | 3.3576 | 0.0217 |
| 4.5183 | 4.5281 | 0.0098 |
| 4.9943 | 4.9958 | 0.0015 |
| 4.8237 | 4.8443 | 0.0206 |
| 3.9517 | 3.9622 | 0.0105 |
| 4.6661 | 4.6779 | 0.0119 |
| 4.3230 | 4.3340 | 0.0110 |
| 4.6323 | 4.6389 | 0.0065 |
| 5.1892 | 5.1829 | 0.0063 |
| 3.4826 | 3.4765 | 0.0061 |
| 4.8707 | 4.8711 | 0.0004 |
| 5.5512 | 5.5633 | 0.0121 |
| 4.3160 | 4.3168 | 0.0009 |
| 4.1287 | 4.1377 | 0.0090 |
| 4.6706 | 4.6643 | 0.0064 |
| 4.7724 | 4.7740 | 0.0016 |
| 4.3708 | 4.3681 | 0.0026 |
| 4.6461 | 4.6591 | 0.0130 |
| 4.2463 | 4.2434 | 0.0029 |
| 3.9290 | 3.9325 | 0.0036 |
| 4.3618 | 4.3576 | 0.0042 |
| 4.5783 | 4.5764 | 0.0019 |
| 4.5742 | 4.5624 | 0.0118 |
| 3.9674 | 3.9871 | 0.0196 |
| 4.5956 | 4.5987 | 0.0030 |
| 4.3582 | 4.3767 | 0.0185 |
| 4.4968 | 4.5133 | 0.0165 |
| 4.9759 | 4.9691 | 0.0067 |
| 4.8660 | 4.8779 | 0.0119 |
| 5.4221 | 5.4226 | 0.0005 |
| 4.4780 | 4.4718 | 0.0062 |
| 3.9860 | 3.9914 | 0.0054 |
| 4.4887 | 4.4854 | 0.0033 |
| 3.3984 | 3.3753 | 0.0231 |
| 4.1467 | 4.1483 | 0.0016 |
| 5.2376 | 5.2363 | 0.0014 |
| 3.9878 | 3.9797 | 0.0080 |
| 4.5883 | 4.5917 | 0.0034 |
| 3.5653 | 3.5507 | 0.0146 |
| 4.5561 | 4.5544 | 0.0017 |
| 4.2743 | 4.2785 | 0.0042 |
| 4.1951 | 4.1906 | 0.0045 |
| 4.1253 | 4.1072 | 0.0181 |
| 4.5314 | 4.5346 | 0.0031 |
| 4.5864 | 4.5935 | 0.0070 |
| 3.3661 | 3.3629 | 0.0032 |
| 4.9361 | 4.9426 | 0.0066 |
| 5.3759 | 5.3814 | 0.0055 |

As it is seen in the table, the value of y^ is very close to the value of y. Therefore, the estimated model seems to be very accurate.

In order to conduct more analysis for the accuracy of the estimated model, want to understand the estimation bias:

Bias(Beta^x1) = -3.8026

Bias(Beta^x2) = -1.9026

Thus, the average estimated MSE for each beta is:

MSE(x1) = 14.491

MSE(x2) = 4.2327

The average estimated model MSE is not close to 1 – our error factor standard deviation. It is slightly larger due to due to the standard deviation used to generation our variables x1 and x2.