**Andrew Hillard**

**Data Analysis 3**

**Homework 1**

**2.5)**

When inference is the goal, less flexible models are preferred because they are easier to interpret. Less flexible models are also preferred when you have data with a lot of noise because less flexible models have low model variance and will be more robust to senseless variation. Another benefit of less flexible models is that they estimate fewer parameters and, as a result, need fewer observations. However, less flexible models can yield poor predictions, especially if the true shape of the data does not fit the model’s assumptions or is unknown (i.e. linear regression is applied to non-linear data). In contrast, when the goal is prediction, flexible models can often provide much better results. Flexible models are also highly useful when the true shape of the data is unknown, or there is a lot of variation that needs to be explained by the model. However, compared to less flexible models, flexible models can often be very challenging to interpret, require more observations, can suffer from the curse of dimensionality, and can overfit the data, leading to high model variance.

**2.6)**

Parametric methods make an assumption about the form of the data’s function, *f.* While simplifying the problem of estimating *f,* parametric methods run the risk of assuming a functional form that does not match reality, leading to poor estimates and prediction. This problem can be addressed by flexible parametric methods that fit multiple functional forms. However, flexible parametric methods run the risk of overfitting the data. Unlike parametric methods, non-parametric methods make no explicit assumptions about the form of the data’s function, *f*. This means non-parametric methods can fit a much wider range of possible shapes for *f.* However, since non-parametric methods make no assumptions about *f,* the problem of estimating *f* becomes much more complex and requires a much larger sample size.

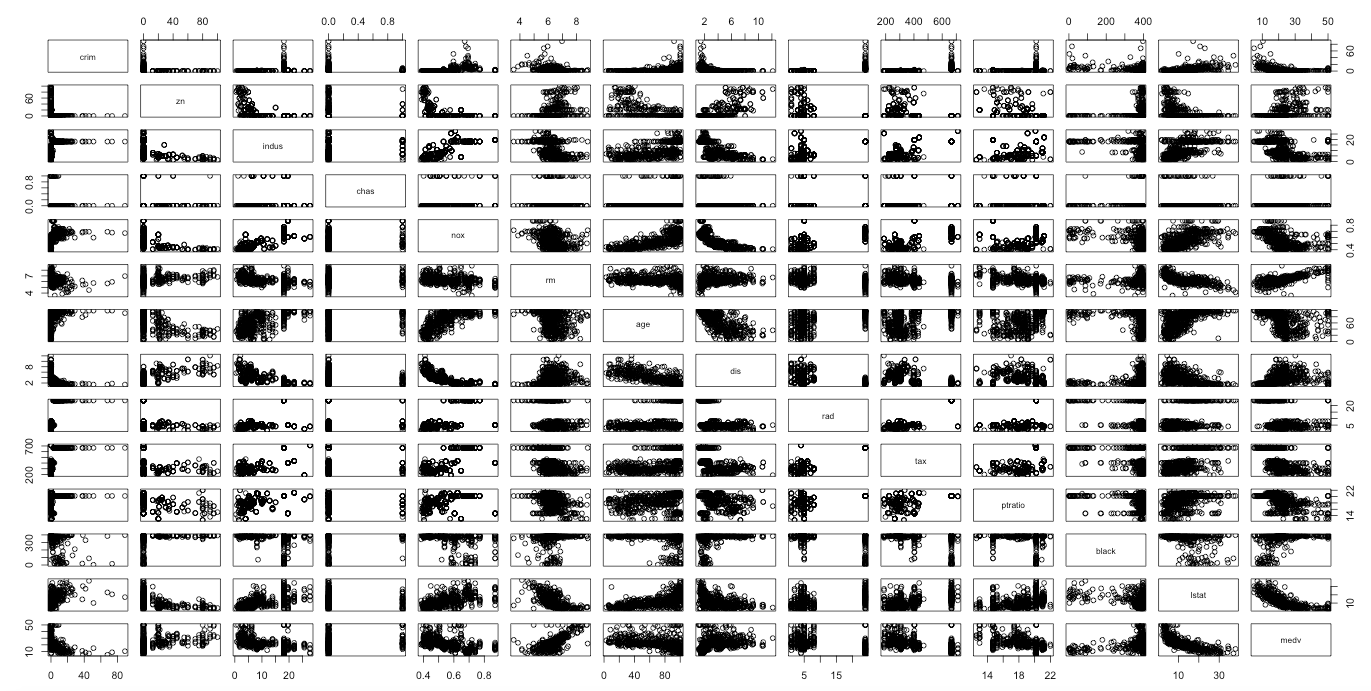
**2.10)**

a. The Boston data frame has 506 rows and 14 columns. Each row represents an observation for a Boston suburb. The column names and descriptions are as follows.

* crim: per capita crime rate by town.
* zn: proportion of residential land zoned for lots over 25,000 sq.ft.
* indus: proportion of non-retail business acres per town.
* chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
* nox: nitrogen oxides concentration (parts per 10 million).
* rm: average number of rooms per dwelling.
* age: proportion of owner-occupied units built prior to 1940.
* dis: weighted mean of distances to five Boston employment centres.
* rad: index of accessibility to radial highways.
* tax: full-value property-tax rate per \$10,000.
* ptratio: pupil-teacher ratio by town.
* black: *1000(Bk - 0.63)^2* where *Bk* is the proportion of blacks by town.
* lstat: lower status of the population (percent).
* medv: median value of owner-occupied homes in \$1000s.

b. Based on the pairwise plots pasted below, there are a number of interesting relationships that could be analyzed further.

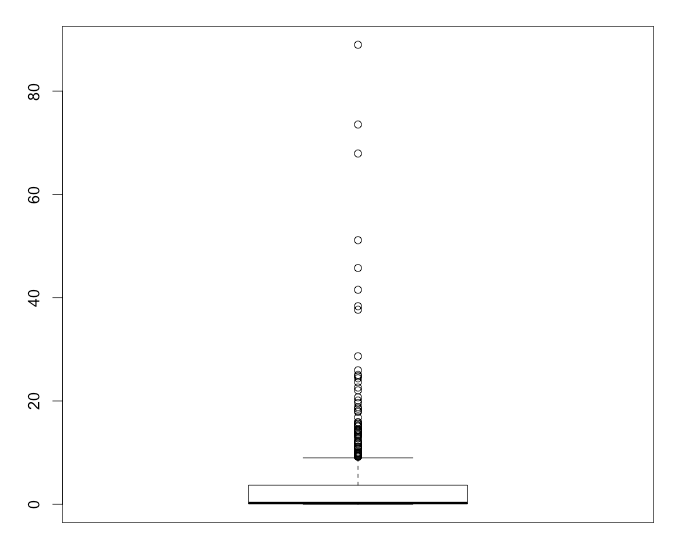
* Crim seems to have a slight positive linear relationship with nox, lstat, and age. There are interesting spikes for crim for certain values of zn, indus, chas, rad, tax, ptratio.
* Zn seems to have a negative linear relationship with indus, nox, and age, and a positive linear relationship with rm, dis, and medv.
* While the same sort of analysis as above can be done on each column, the remaining findings will only talk about the most striking visual relationships.
* There appears to be a positive linear relationship between rm and medv.
* There appears to be a negative linear relationship between rm and lstat, age and dis, and lstat and medv.
* There appears to be a decreasing non-linear relationship between nox and dis. There appears to be an increasing non-linear relationship between nox and age.



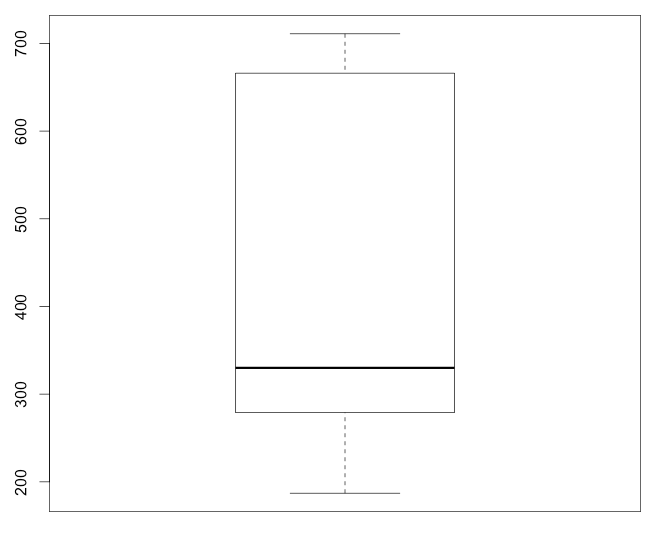
c. Based on visual analysis, crime appears to be associated with a number of predictors (see answer in part b). The correlation matrix also shows that crime has a strong positive correlation with tax and rad (correlation greater than .50), and a medium positive correlation with indus, nox, age, ptratio, and lstat (correlation between .20 and .50). In addition, crime has negative correlation with medv, black, dis, rm, and zn (correlation between -.20 and -0.40)

d. Crime rates have a very large range, .00632 to 88.98. The range indicates that some suburbs are much higher than others since the average is 3.614. This is supported by the crime boxplot (pasted below), which shows a number of suburbs that are outliers for crime rate. Tax rates have a range of 187 to 711, and a mean of 408.2. Since the mean is in the center of the range, there are likely fewer suburbs that have extreme tax rate values. This is supported by the tax boxplot given below, which shows no outliers. The ptratio has a range of 12.60 to 22. The range appears to be reasonable, though the boxplot for ptratio shows that two Boston suburbs are outliers and have very low ptratio compared to the other suburbs.

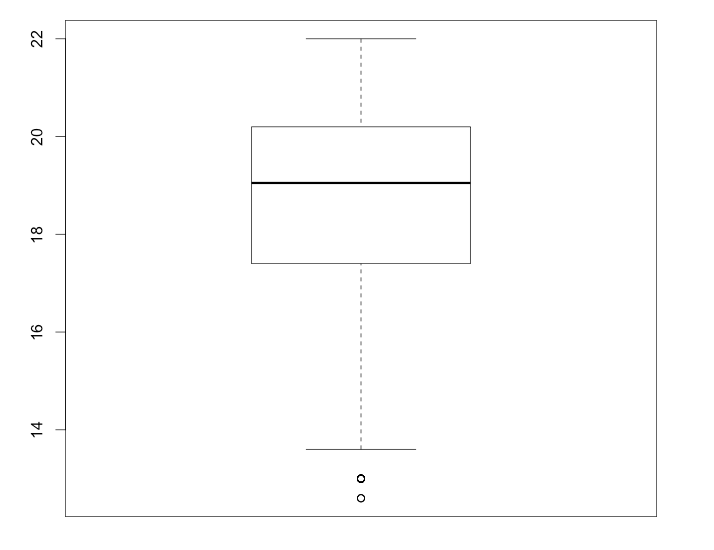
Boxplot of Crime Rates in Boston Suburbs



Boxplot of Tax in Boston Suburbs



Boxplot of ptratio in Boston Suburbs



e. 35 Boston suburbs bound the Charles River.

f. 19.05

g. Suburb 399 and 406 have the lowest median value of owner-occupied homes, 5. The crime rate for both these suburbs is very high, 38.35 and 67.92 respectively. Below gives a summary of how the predictors for these suburbs compare to other suburbs. Overall, the predictors are either very high or very low compared to a typical suburb.

* Zn is low at 0, but not unusual since 25% of the observations have a value of 0.
* Indus is a little higher than most and has the same value as the 3rd Quartile of all observations.
* Chas is normal at 0. The median of the observations is 0.
* Nox is a little higher than most and falls above the 3rd quartile observation.
* Rm lower than most suburbs and falls below the 1st quartile observation.
* Age is very high and shares the same value as the max observation.
* Dis is lower than most suburbs and falls below the 1st quartile observation.
* Rad is very high and shares the same value as the max observation.
* Tax is a little higher than most and falls at the 3rd quartile observation.
* Ptratio is a little higher than most and falls at the 3rd quartile observation.
* Black is very high and falls at or near the maximum observation.
* Lstat is higher than most and falls above the 3rd quartile.
* Medv is vey low and falls at the value of the minimum observation.

h. 64 suburbs average more than 7 rooms per dwelling. 13 suburbs average more than 8 rooms per dwelling. The suburbs that average more than 8 rooms per dwelling appear to have below average crime rates, lower than average indus, lower than average lstat, and higher than average medv. This makes sense because suburbs with more rooms per dwelling would likely be more affluent and have less crime, lower poverty rates, less industry, and higher median income.

**3.5)**

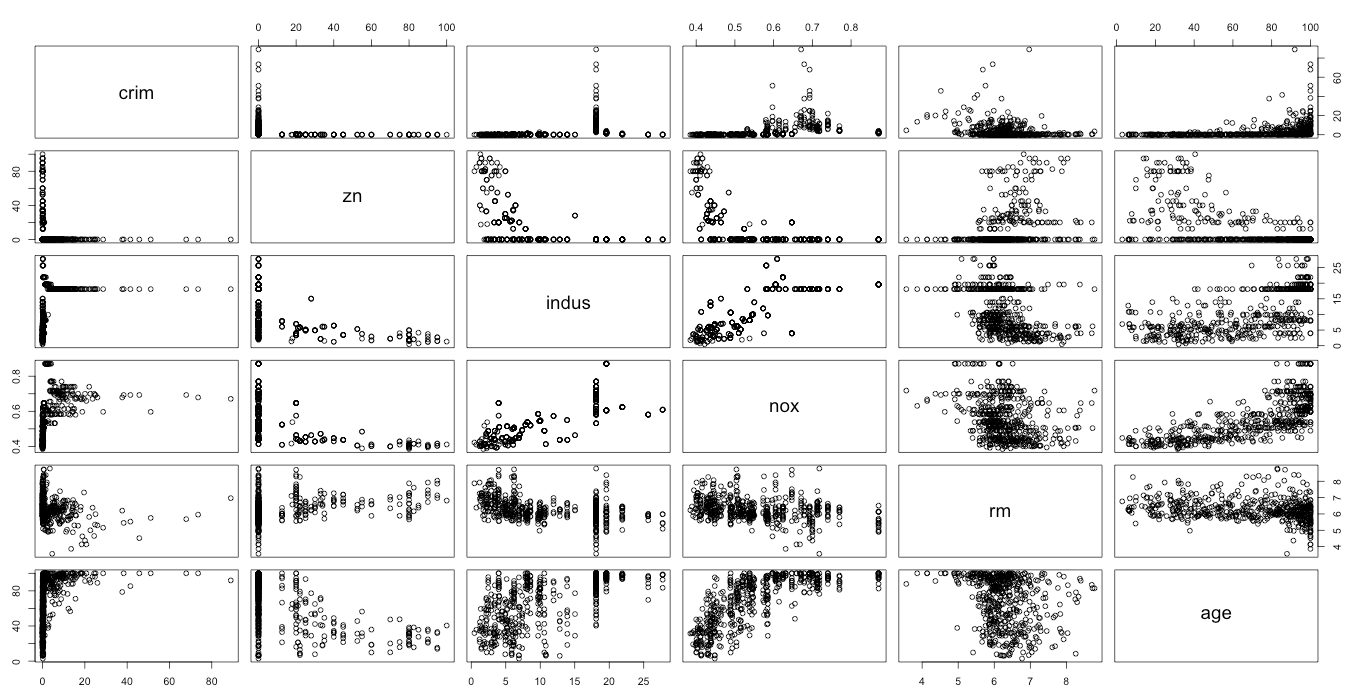
Makes it a lot easier if you change index letters to distinguish the observation from the summation.

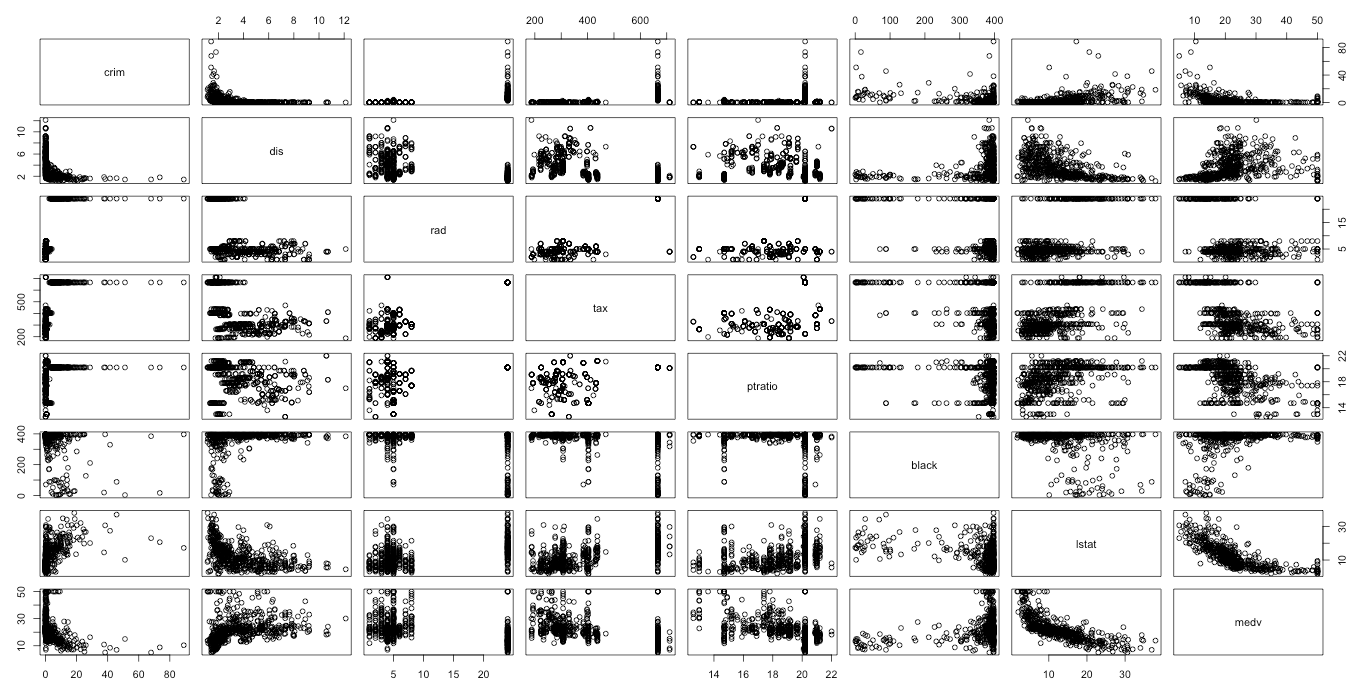
=

Therefore,

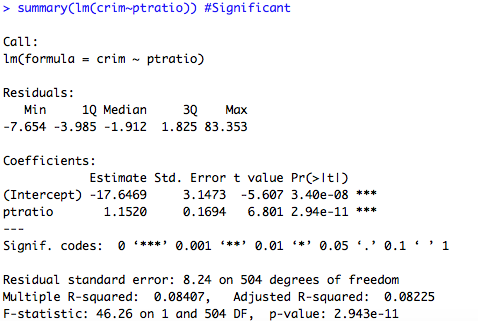
**3.15)**

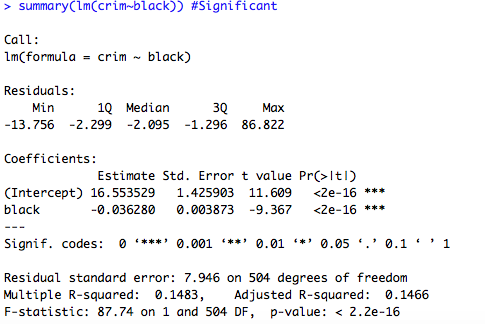
a. Every predictor except chas is statistically significant and has a p-value less than .05 when incorporated in the simple linear regression model with crime as the response. The below graph shows crim plotted against every significant predictor (first row of each graph). In each pairwise plot, crim and the predictors appear to have some sort of linear relationship.





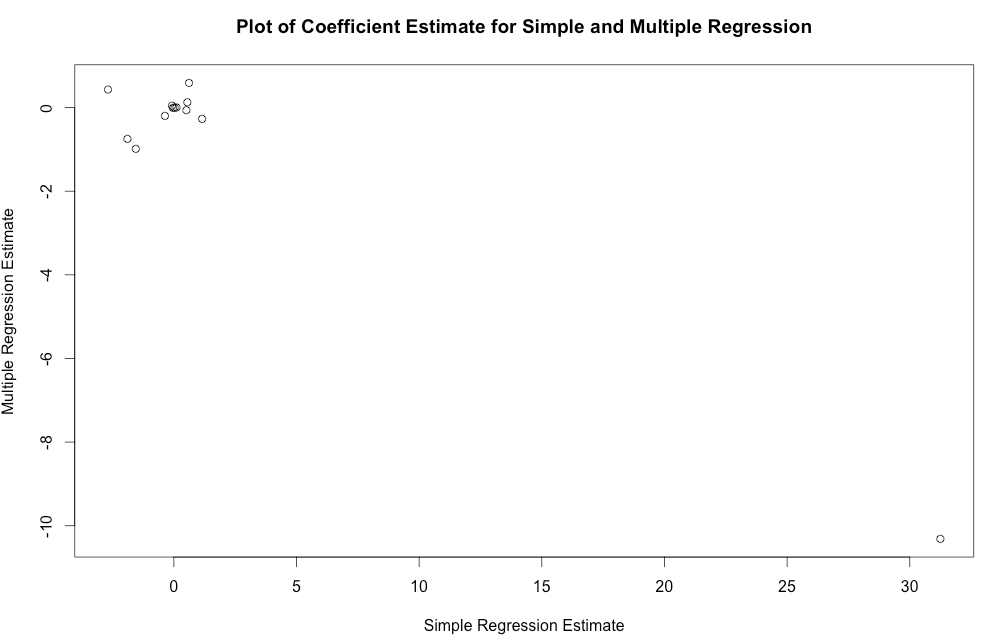
I’ve also pasted the summaries of the linear regression for two variables below. These summaries show that these predictors are statistically significant at the 5% level. The same logic applies to the other predictors that were found to be significant.





b. The model is statistically significant and has a p-value of <2.2e-16. The Adjusted R-squared is 0.4396. Of the predictors, only zn, dis, rad, black, and medv are significant and allow the null hypothesis, , to be rejected.

c. In part a, there were many predictors that were significant in the simple linear regression. In part b, only five of the thirteen predictors were significant. This is not surprising because, unlike the simple regression, the significance of a predictor in multiple regression takes into account the other predictors. If any of the predictors are correlated, which is likely, one predictor may not give any additional information about the response if it’s correlated with a predictor already in the model.



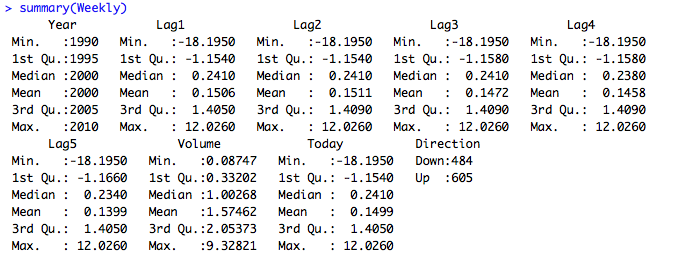
d. There is not sufficient evidence to support that black is non-linear. However, there is sufficient evidence to support that zn, rm, rad, tax, and lstat are quadratic, and indus, nox, age, dis, ptratio, and medv are cubic.

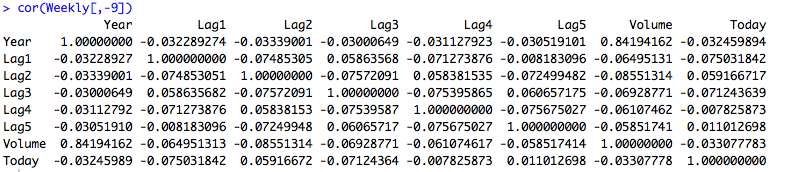
**4.3)**

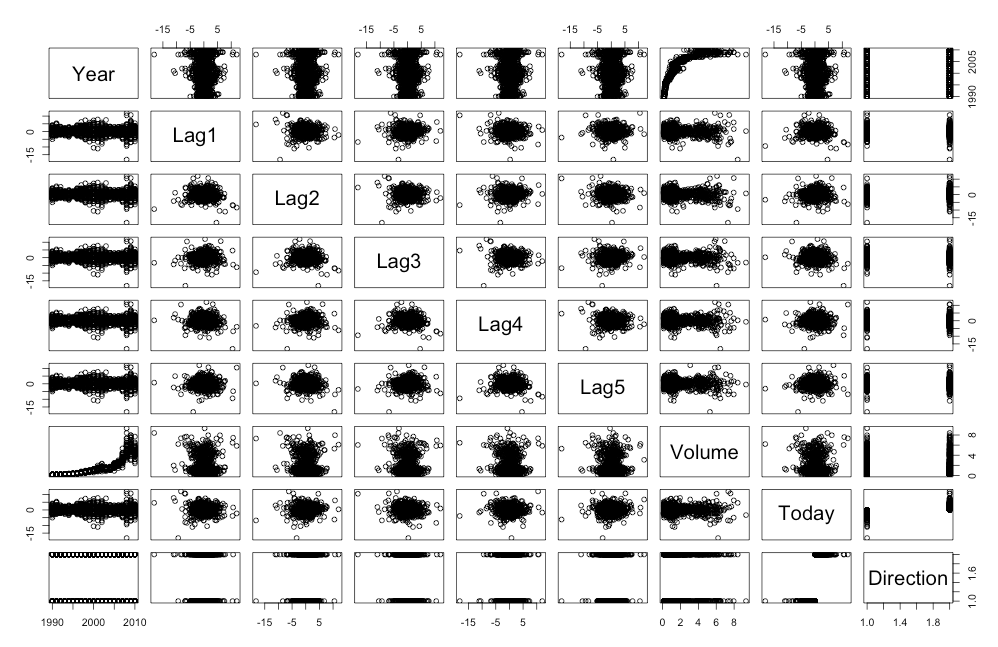
Please see notebook paper attached.

**4.10)**

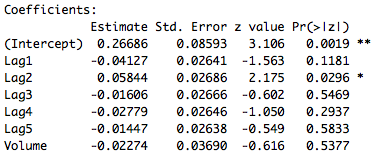
a. There is a definite relationship, potentially non-linear, between year and volume. The correlation between year and volume is very high, .842. Other correlations between predictors are very low and appear to have no visual association with one another (see plot below).







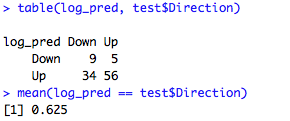
b. The only predictor that is significant in the logistic regression model is lag2.



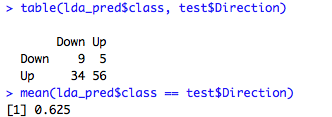
c. The confusion matrix is telling us that when the logistic model predicts down, it is wrong 47.05%, and when it predicts up, it is wrong 43.56% of the time. The error rate is 43.89% and the correct prediction rate is 56.11%.



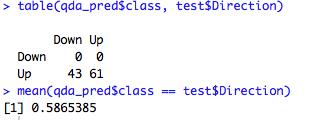
d. The confusion matrix is given below. The logistic model correctly predicted 62.5% of the time.



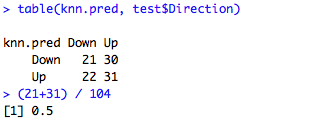
e. The confusion matrix is given below. The LDA model correctly predicted 62.5% of the time.



f. The confusion matrix is given below. The QDA model correctly predicted 58.7% of the time.



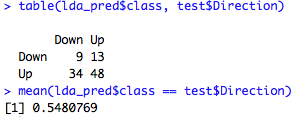
g. The confusion matrix is given below. The KNN model correctly predicted 50% of the time.



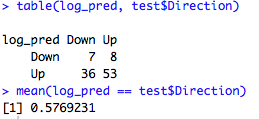
h. It appears that the logistic regression and LDA have the minimum error rates and are the best models, followed by QDA and KNN.

i.

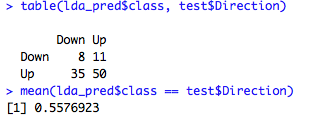
The LDA model with all predictors results in the confusion matrix and correct prediction rate below.



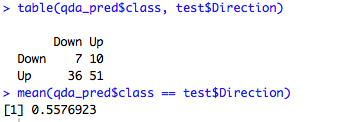
When all the predictors are entered into a logistic regression model, only Lag 1 and Lag 2 were significant. The Logistic Regression with predictors Lag 1 and Lag 2 was therefore run and resulted in the following confusion matrix and correct prediction rate. The same confusion matrix and correct prediction rate results if I include the interaction between Lag 1 and Lag 2.



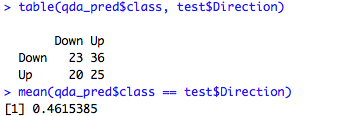
The LDA model is run for Lag1, Lag2, and Lag2 squared. The below confusion matrix and correct prediction rate results.



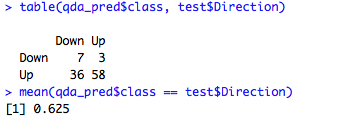
The QDA model with Lag 1 and Lag 2 as the predictors results in the following confusion matrix and correct prediction rate.



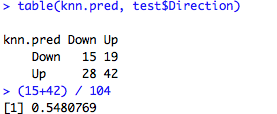
The QDA model with Lag 1, Lag 2, and the interaction between Lag 1 and Lag 2 results in the following confusion matrix and correct prediction rate.



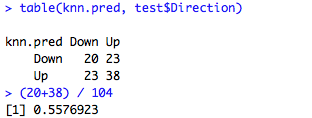
The QDA model with Lag2, and Lag2 squared results in the confusion matrix and correct prediction rate below.



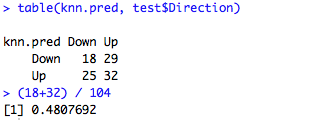
The KNN model with Lag 2 and 3 neighbors results in the confusion matrix and correct prediction rate below.



The KNN model with Lag 2 and 50 neighbors results in the confusion matrix and correct prediction rate below.



The KNN model with Lag 1, Lag 2, and 50 neighbors results in the confusion matrix and correct prediction rate below.

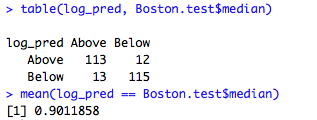


Despite running different models with different predictors, no model out performed the correct prediction rate of the logistic and LDA model with Lag 2 as a predictor. Interestingly, the QDA model with the predictors Lag 2, and Lag 2 squared, performed just as well as this original model. All the other model variants performed worse in terms of correct prediction rate.

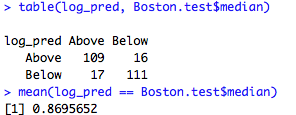
**4.13)**

The data was first randomly divided into two sets of data for training and testing. A column was added to the data to classify whether the data had a crime rate above or below the median.

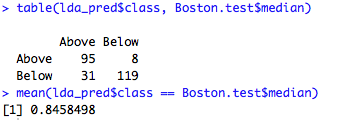
The logistic regression using all predictors results in the below confusion matrix and correct prediction rate.



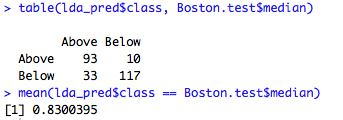
When only the significant predictors from the first logistic regression (nox, dis, rad, ptratio, and medv) are used, the logistic regression results in the following confusion matrix and correct prediction rate.



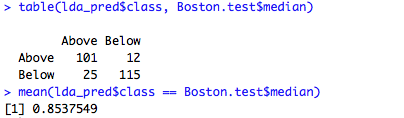
The LDA model with all predictors results in the following confusion matrix and correct prediction rate.



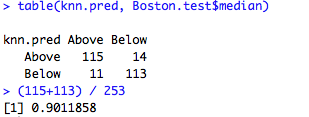
The LDA with only nox, dis, rad, ptratio, and medv as the predictors results in the following confusion matrix and correct prediction rate.



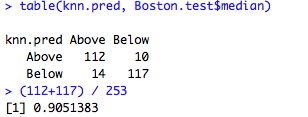
The LDA with nox, nox squared, dis, rad, ptratio, and medv as the predictors results in the following confusion matrix and correct prediction rate.



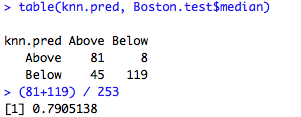
The KNN with all predictors and 1 neighbor results in the following confusion matrix and correct prediction rate.



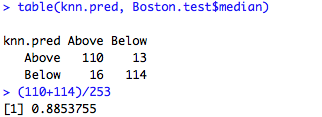
The KNN with all predictors and 3 neighbors results in the following confusion matrix and correct prediction rate.



The KNN with all predictors and 50 neighbors results in the following confusion matrix and correct prediction rate.



The KNN with only nox, nox squared, dis, rad, ptratio, and medv as predictors and 3 neighbors results in the following confusion matrix and correct prediction rate.



In conclusion, the best model discovered in this exercise to predict whether a suburb is above or below the median crime rate is the KNN model with all predictors and 3 neighbors, which has a correct prediction rate of 90.5% and a test error rate of 9.5%. The next best model was the logistic regression and KNN with 1 neighbor using all predictors; both models had a correct prediction rate of 90.1% and a test error rate of 9.9%. The models with the worst prediction rates were the LDA models.