Kevin Gilmore

Professor Risley

BANA 6043

Statistical Computing Final

**Part 1: Describe the Data**

A.1.1 Describe the data set high-level with the R help file or Google.

* install.packages(‘MASS’)

library(‘MASS)

data(‘Boston’)

**?Boston** – This dataset is over the housing values in the suburbs of Boston, the description states that the Boston data frame has 506 rows and 14 columns

A.1.2 How many observations are present? How many variables are present?

* **Str(Boston)**
* There are 506 observations of 14 variables

A.1.3 What are the data types of the variables?

* **Str(Boston)**
* There is both numeric and integer variables

A.1.4 We want to understand what factors affect home values. What is the response (ie, dependent) variable?

* The home value itself would be the response. It changes through the independent variables which are given in the dataset such as crime or tax

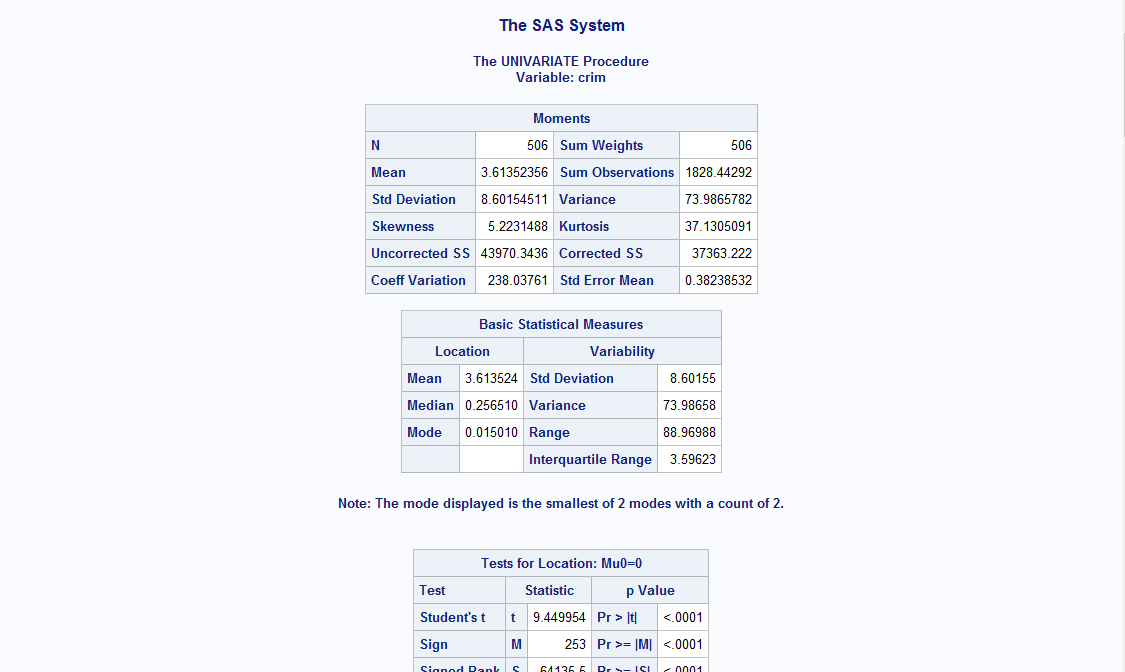
### Part 2: Univariate Summary

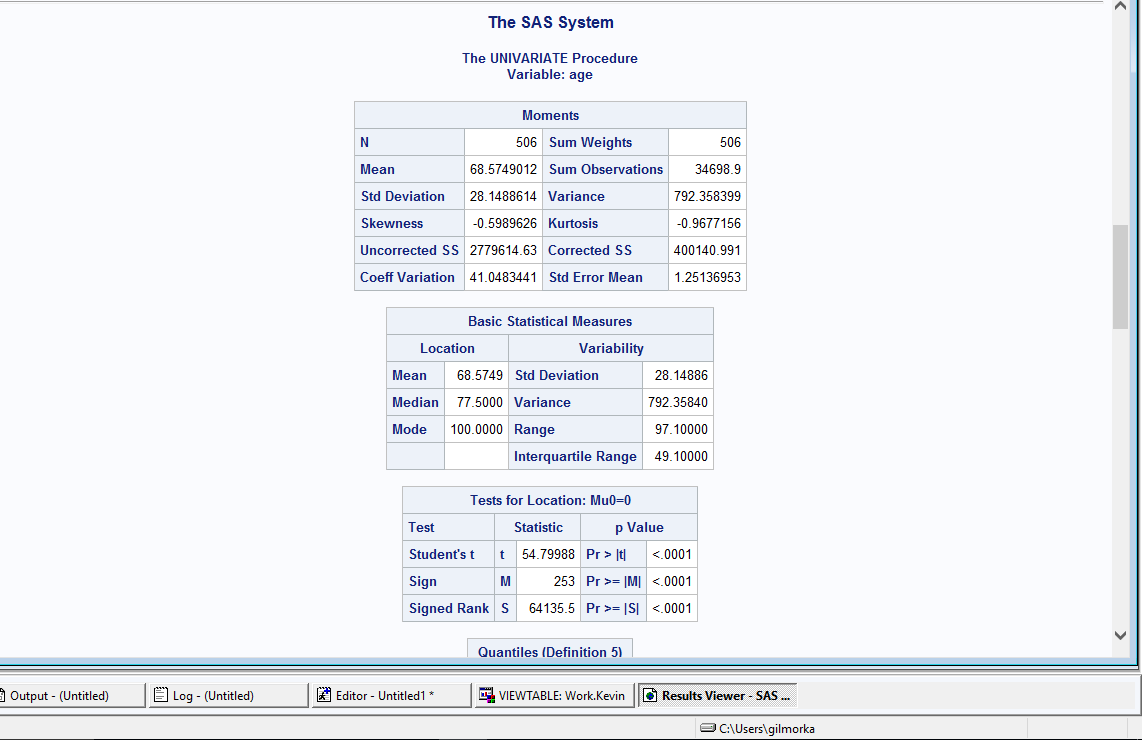
A.2.1 Produce summary statistics for three variables: crim, age, and medv. Screenshot the output.

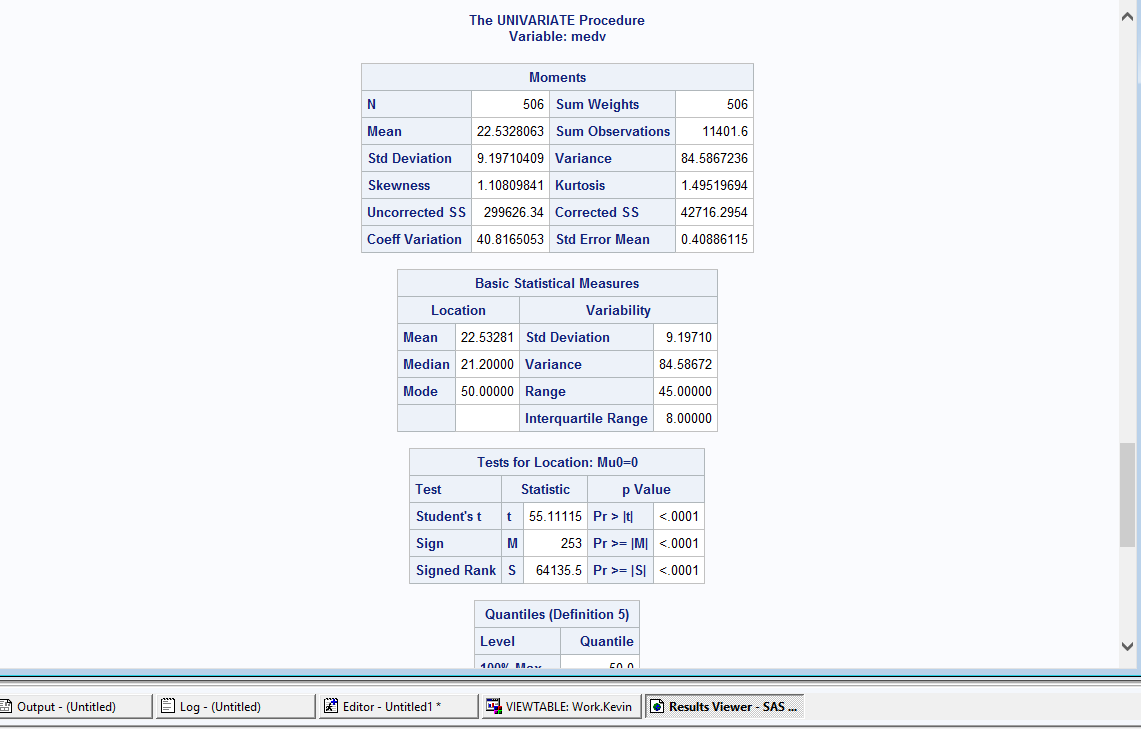
* **proc** **univariate** data=work.kevin;

var crim age medv;

**run**;







A.2.2 With the appropriate units of measurement, interpret the median of medv.

* **proc** **means** data=work.kevin median;

var medv;

**run**;

* **Median of medv is 21.2**

A.2.3 With the appropriate units of measurement, interpret the maximum of age.

* **proc** **means** data=work.kevin maximum;

var age;

**run**;

* Maximum of age is 100

A.2.4 With the appropriate units of measurement, interpret the 1st Quartile of crim. “per capita” means “per person”.

* **proc** **univariate** data=work.kevin;

var crim;

r**un**;

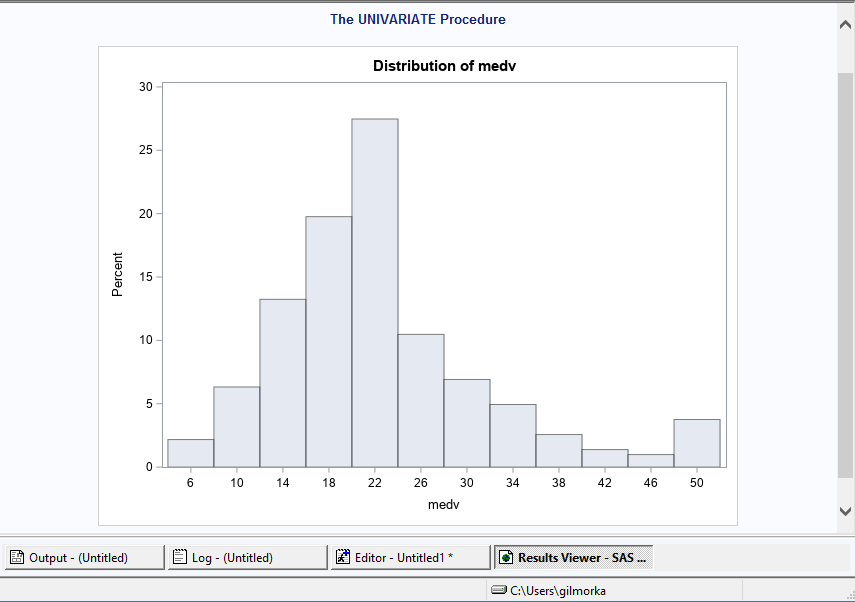
* The 1st Quartile of crim per capita is 0.08199

A.2.5 Plot a histogram for the variable medv. Screenshot the output.

* **proc** **univariate** data=work.kevin noprint;

histogram medv / midpoints=uniform;

**run**;



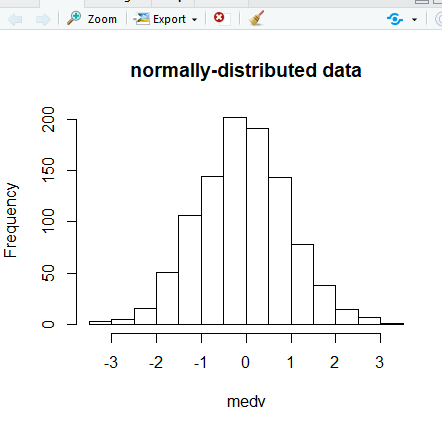
A.2.6 According to the definition below, do the data appear **normally distributed**?

* No, it is not symmetrically distributed around the average value

A.2.7 Multiply medv by 1,000 in order to convert it to original units. Create a histogram for this transformed variable. Screenshot the output.

* medv <- rnorm(1000)

hist(medv, main="normally-distributed data")



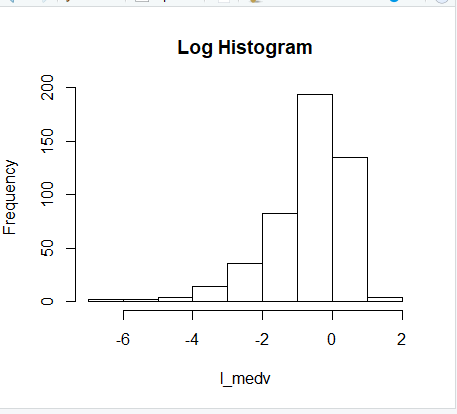
A.2.8 Did the shape of the histogram change from A.2.5?

* Yes, it is now normally-distributed

A.2.9 Take the natural log of medv. Create a histogram for this transformed variable. Screenshot the output.

* l\_medv<-log(medv)

hist(l\_medv, main='Log Histogram')



A.2.10 Do the data under the log transformation appear to be more normally distributed than the original variable?

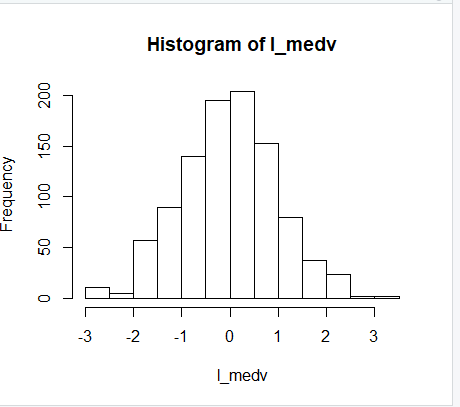
* The data is the new histogram is going down from 0 to -6. The data in the original variable does have the high point at the average and then goes down on both sides, however not symmetrical. The original data went down until 46 and then it went back up, while the new log historgram is going down on each number.

A.2.11 Finally, take the natural log of medv **after multiplying it by 1,000**. Create a histogram for this transformed variable. Screenshot the output.

* l\_medv<-log(medv)

l\_medv <- rnorm(1000)

hist(l\_medv)



A.2.12 Did the shape of the histogram change from A.2.9? What can you conclude about the use of the log transformation on already-transformed values?

* Yes, the shape of the histogram changed. The use of the log transformation on already-transformed values ensures that the data is distributed normally if it is not already.

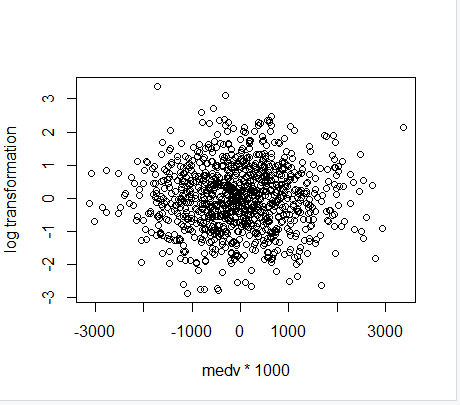
A.2.13 Of the 4 transformations analyzed, which would you recommend to use to begin a multivariate analysis on medv?

* (Second option) I would recommend to multiply medv by 1,000 to convert it to its original units.
* medv <-rnorm(1000)

### Part 3: Bivariate Summary

A.3.1 Plot medv\*1000 on the x-axis and its log transformation on the y-axis. Screenshot the output. Briefly explain how a change in x relates to a change in y.

* plot(medv\*1000,l\_medv,ylab = 'log transformation')
* X and y can have correlation, a positive relationship has two variables increasing or decreasing together, a negative relationship has one variable increasing while the other decreases, and there is no relationship when a change in one variable does not have a change in the other



A.3.2 Produce Spearman’s Rank-Order Correlation coefficient for the variables in A.3.1. Interpret the coefficient. What can we infer about the nature of the log transformation?

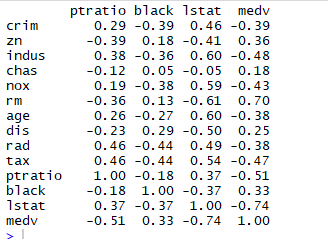
* cor(medv\*1000,l\_medv,method='spearman')
* The coefficient is 0.03369701
* We can infer that there is no linear relationship because -0.3 < r < 0.3 indicates no linear relationship.

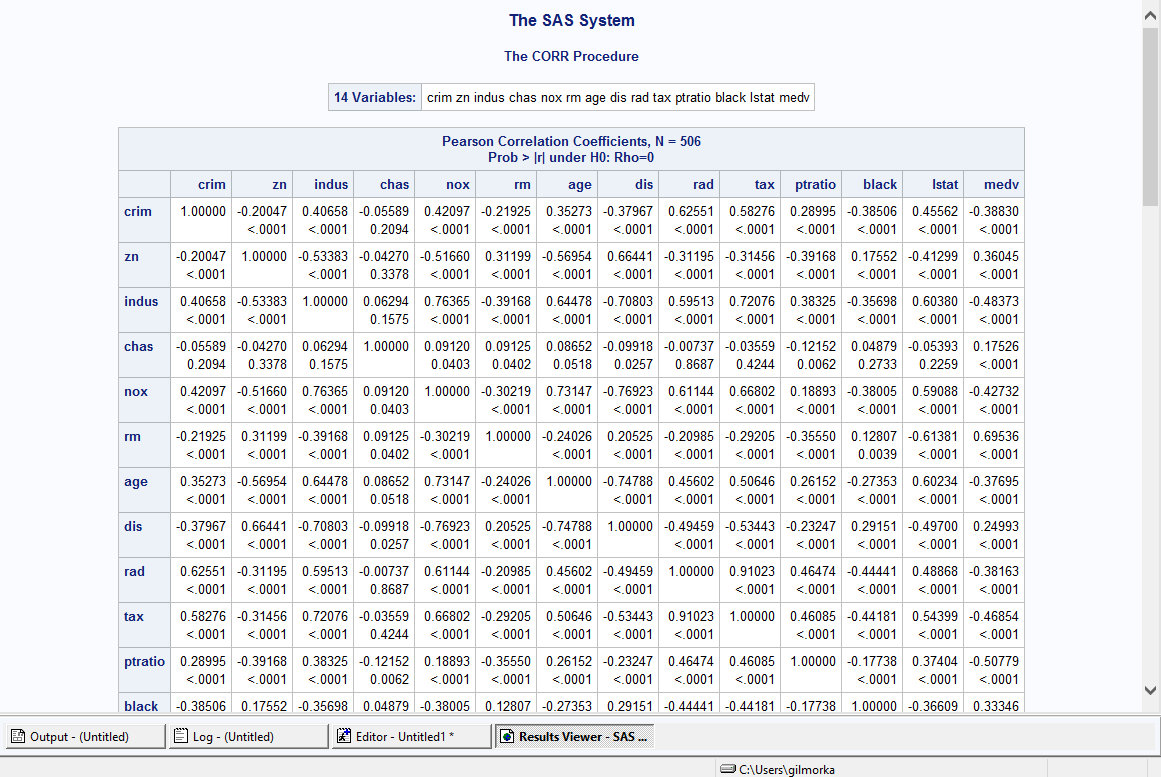
A.3.3 Produce a correlation matrix with Pearson’s r with all variables in the data set. For the variable medv, use the transformation you recommend in A.2.13. Screenshot the matrix. If you use R for this process, read the notes below. Otherwise, proceed to A.3.4.

* library(MASS)

res <- cor(crim zn indus chas nox rm age dis rad tax ptratio black lstat medv\*1000)

round(res, 2)





A.3.4 Which two variables have the strongest linear relationship with your chosen transformation of medv?

* Rm and lstat

A.3.5 Produce two scatter plots for each of these variables on the x-axis with the chosen transformation of medv on the y-axis.

* ods graphics on;

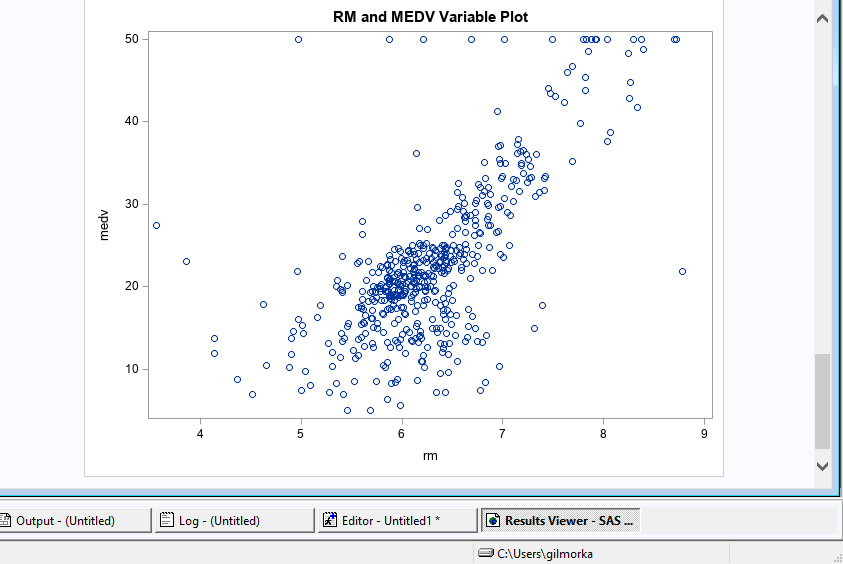
**proc** **sgplot** data=work.kevin;

scatter x=rm y=medv;

title 'RM and MEDV Variable Plot';

**run**;

ods graphics off;



* ods graphics on;

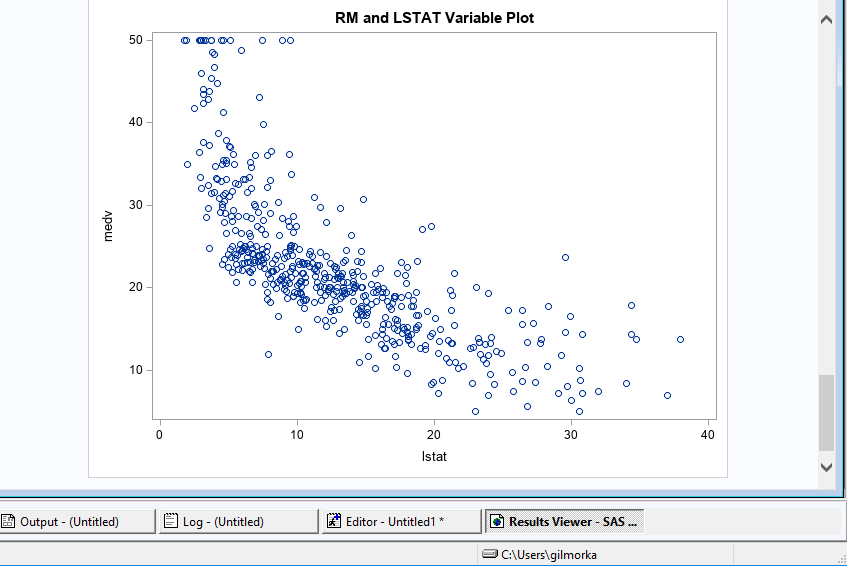
**proc** **sgplot** data=work.kevin;

scatter x=lstat y=medv;

title 'RM and LSTAT Variable Plot';

**run**;

ods graphics off;



A.3.6 Based on the plots in A.3.5, are the relationships linear? Why/why not?

* A linear relationship is one that follows a straight line. The first plot seems to have a linear relationship, however the second one is a bit more curved. It seems to somewhat follow a line, however even if it was considered linear it would not be that strong of a relationship.

A.3.7 Interpret the two relationships in everday language. Are the results intuitive? Why/why not?

* The first plot shows two variables rm and medv. The variable rm in this case stands for average number of rooms per dweeling and medv stands for the median value of owner-occupied homes in \$1000s. In the first graph as the rooms per dwelling go up, it also seems the case that median value of owner occupies homes in thousands does as well. The second plot has the variables of medv and lstat. Lstat is the lower status of the population. As the median value of owner occupied homes goes up, the LSAT variable does not. It has many points within a single range on the plot. I do believe these results are somewhat intuitive because typically if you had less rooms per dwelling then the home owner has less median value.

A.3.8 Run a regression with your chosen transformation of medv on the left-hand-side of the equation and with the two variables in A.3.4 on the right-hand-side. Produce a summary of the model and screenshot it.

* ods graphics on;

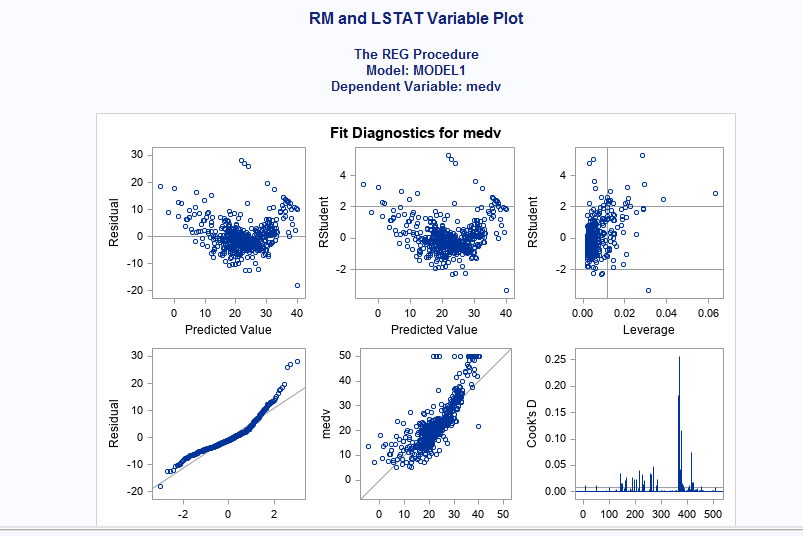
**proc** **reg** data=work.kevin plots(maxpoints=none);

model medv = rm lstat;

**run**;

ods graphics off;





A.3.9 Interpret the two coefficients from the model produced in A.3.8. Are the coefficients intuitive? Why/why not?

* You do not want patterns in residuals, you want the graph to essentially look like noise. These coefficients do not seem to be in a pattern, thus the coefficients are intuitive

# **Part B: Time Series**

B.1: Use the getSymbols() function from the quantmod package to pull the stock prices for your chosen companies from Yahoo!. Use the from= and to= arguments to ensure the time series are of the same length. Use the cbind() function to combine the series into a matrix. Finally, pass this matrix to the head() and tail() functions, and screenshot the output.

* Install.packages(“quantmod”)

Library(quantmod)

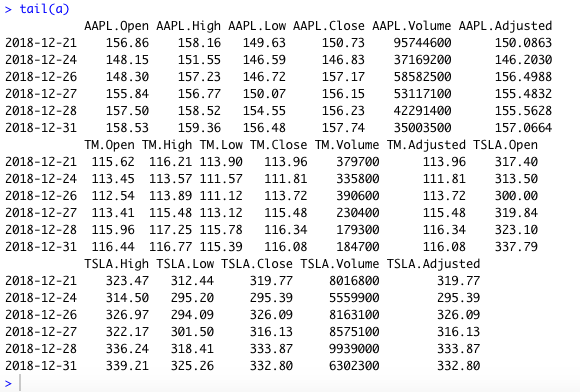
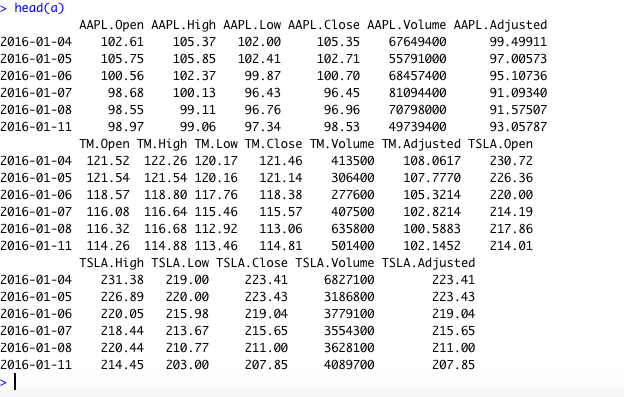
getSymbols("TSLA", from = "2016-01-01", to = "2019-01-01", src = "yahoo")

getSymbols("TM", from = "2016-01-01", to = "2019-01-01", src = "yahoo")

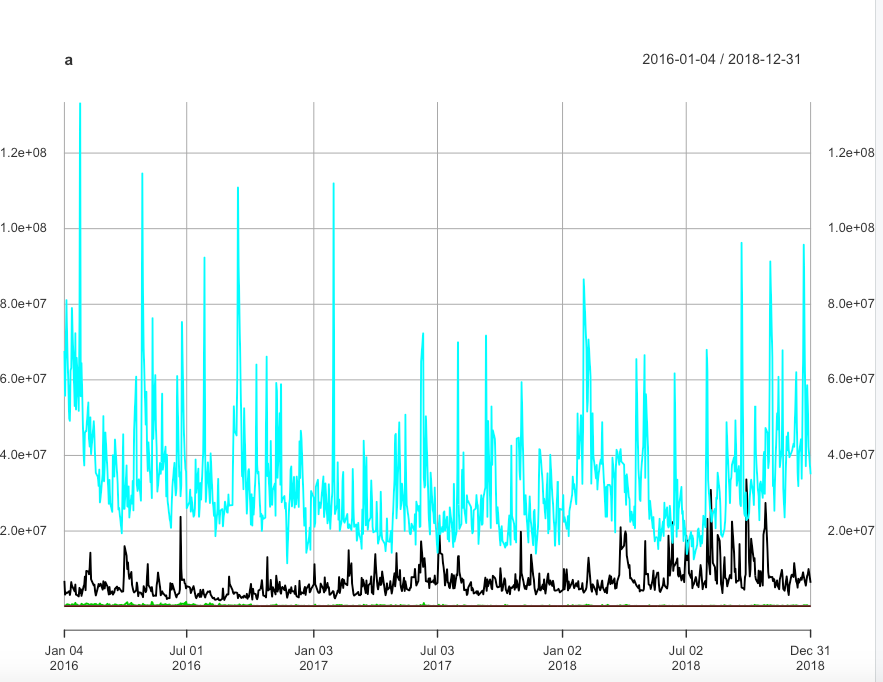
getSymbols("AAPL", from = "2016-01-01", to = "2019-01-01", src = "yahoo")

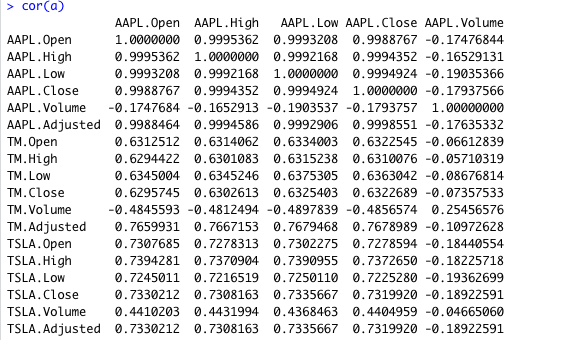
a<-cbind(AAPL,TM,TSLA)

head(a)

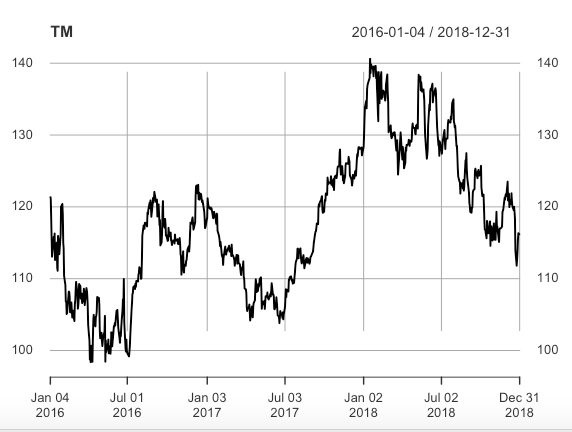


B.2: Plot the series, and screenshot the output.



B.3: Produce a correlation matrix with Pearson’s r. Interpret the values and screenshot the output. 

B.4: By measures of correlation, do the stock prices appear to be related with one another?



I plotted all the companies just to see how it would look. It is easy to see that Apple has been on a steady increase primarily, while Tesla and Toyota fluctuate more together. The stocks could be somewhat related, but not too similar. Apple has always had high stock, however Tesla and Toyota seem to be more related. I believe they are more related because of the same business. I was interested to see the Tesla stock after the cars on fire.

B.5: Transform the data by taking the difference of the log-transformed stock price.

* getSymbols("AAPL", from = "2016-01-01", to = "2019-01-01", src = "yahoo")

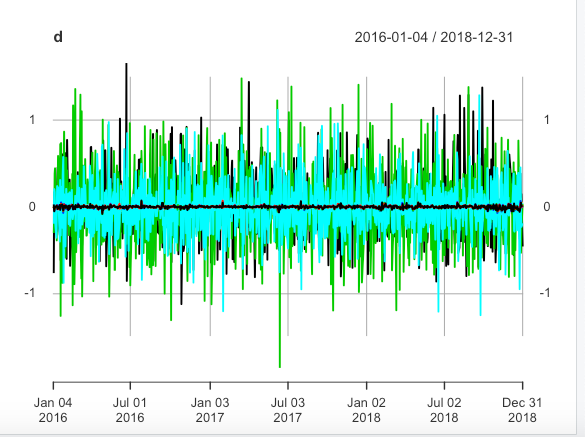
getSymbols("TM", from = "2016-01-01", to = "2019-01-01", src = "yahoo")

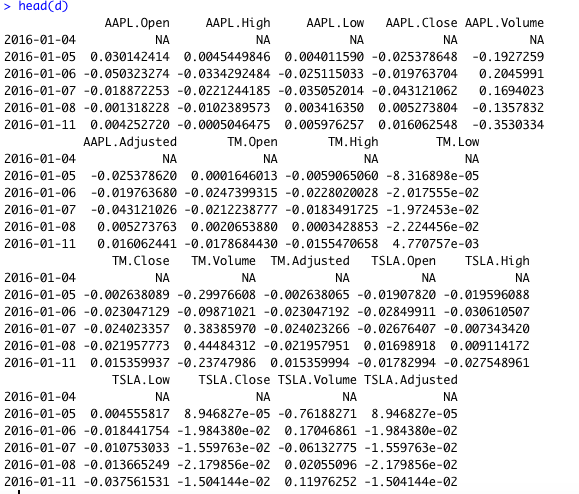
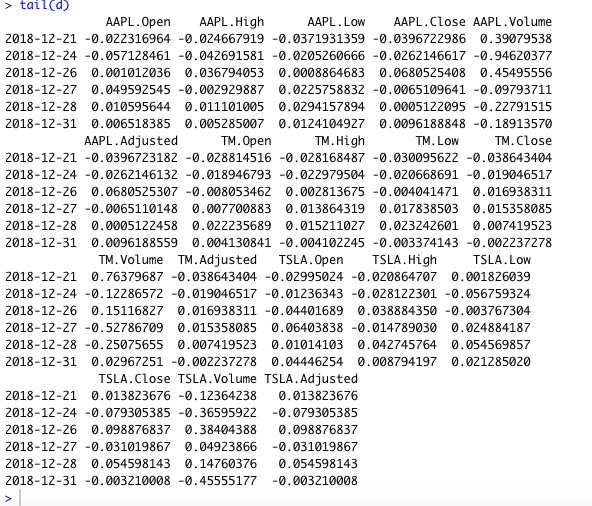
getSymbols("TSLA", from = "2016-01-01", to = "2019-01-01", src = "yahoo")

a<-cbind(AAPL,TM,TSLA)

d<-diff(log(a))

plot(d)

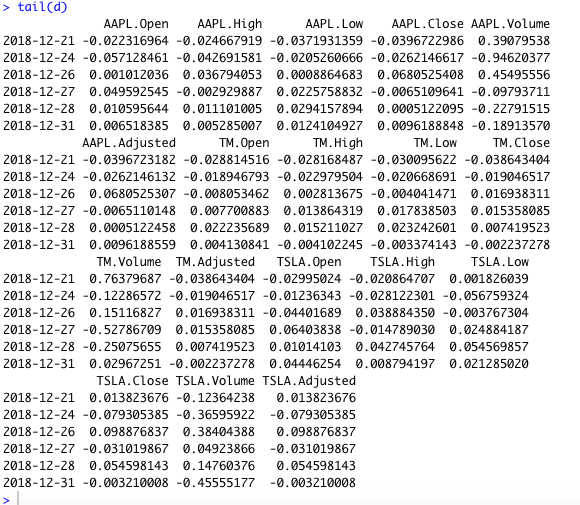
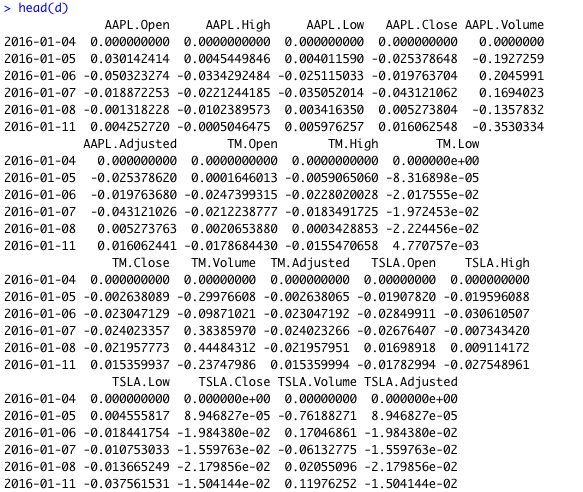
B.6: Plot the transformed series, and screenshot the output. 

B.7: Similar to B.1, use the cbind() function to combine the transformed series into a matrix. Finally, pass this matrix to the head() and tail() functions, and screenshot the output.  

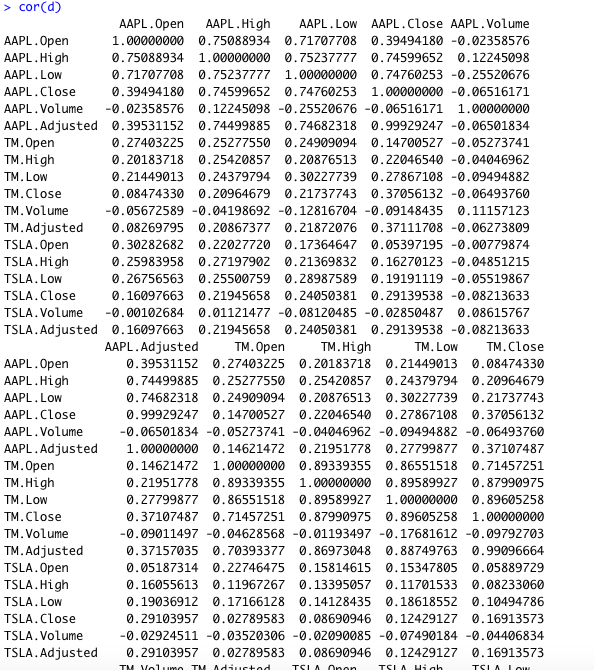
cbind(d)

head(d)

tail(d)

B.8: You should see that the first value is NA for each transformed series. This is because R cannot compute a difference for the first observation. You will create a second matrix that removes the NA. There are two options, described below. After you complete this task, pass the new matrix to the head() and tail() functions, and screenshot the output.  

B.9: Produce a correlation matrix with Pearson’s r for the transformed values. Interpret the values and screenshot the output.



B.10: By measures of correlation, does the daily growth in stock prices appear to be related with one another?

* Yes

B.11: What do your results say about the random walk hypothesis?

* The random walk hypothesis is a financial theory which states the stock market price changes are random and unpredictable. If you were to look at the 3 plots on B4 you could see that these results do NOT seem to be random. While there is fluctuation, the common trends seem to be the same.