Kelly Trinh
Dr. Karen Mazidi
CS 4375.004

**Link to Portfolio : https://github.com/KellyTranT/KellyTrinh_CS4347_ML_Portfolio**

**a) Output of Code :**

```
Opening file Boston.csv.
Readling line 1
heading: rm,medv
new length 506
Closing Boston.csv file
Number of records: 506

Stats for rm:
Sum = 3180.03
Mean = 6.28463
Median = 6.2085
Range = 5.219

Stats for medv:
Sum = 11401.6
Mean = 22.5328
Median = 21.2
Range = 45

Covariance between rm and medv: 4.49345
Correlation between rm and medv: 0.69536
```

**b)** After creating my own functions to output the statistics of data using C++, I saw that there were numerical similarities between the functions that I wrote myself and the built-in functions provided in R through RStudio. However, the greatest difference between creating and using my own functions in C++ and using the functions within RStudio is the amount of time that can be saved. In C++, it took me a significant amount of time to make sure that all of my equations were correct and that they would output the correct value according to the data. Through R, I was able to simply use the built-in function, thus saving me a significant amount of time. In general, using the built-in functions that R provides is much easier as well, as I didn't need to worry about knowing the equation or making sure I'm coding it correctly.

**c)** The first few statistical measures calculated in this program are the mean, median, and range of each data set (rm and medv). The mean represents an average of all the numbers in the set, and the median represents the middle-most value in the data set when it is sorted from least to greatest. The range is the difference between the largest and the smallest number in the set of data. Both the mean and the median allow an individual to interpret what value the data sets are centered around. The range can help us see the spread of the data. This essentially means that we are able to interpret by how much the values are distributed within the data set. These three statistical measures are useful in data exploration prior to machine learning because the user is able to have a greater understanding of what values the data sets actually consist of, which would allow them to create algorithms best suited to the data.

**d)** Covariance and correlation were also determined in this assignment. Covariance essentially measures how changes in one data set are associated with changes in another data set, scaled from [-inf, inf]. When the covariance value is larger, in either the negative or positive direction, that means that the relationship between the two data sets are more reliant on each other. Correlation is the same as covariance, but scaled from [-1, 1]. Correlation also shows how reliant the data are on each other, with the relationship being stronger when the correlation value is closer to -1 or 1. In regards to machine learning, knowing the correlation and covariance is useful because if the two data sets are strongly reliant on each other, then machine learning algorithms can use one data set to predict values in the other. However, if they are not reliant on each other, or if the correlation and/or covariance is close to 0, then machine learning algorithms know that the data sets are not helpful in regards to predicting each other.