a. copy/paste runs of your code showing the output

```
Openng file Boston.csv.
getting line one
heading: rm,medv
New length = 506
closing file Boston.csv.
number of records: 506

Stats for rm
Sum: 3180.03
Mean: 6.28463
Median: 6.2085
Range: Min: 3.561000 Max: 8.780000 Difference: 5.219000

Stats for medv
Sum: 11401.6
Mean: 22.5328
Median: 21.2
Range: Min: 5.000000 Max: 50.000000 Difference: 45.000000

Correlation = 0.69536

Covariance = 4.49345

program terminated.
Program ended with exit code: 0
```

b. describing your experience using built-in functions in R versus coding your own functions in C++

- While working on this project I found the sum, mean median, range, covariance, and correlation using built-in functions in R to verify the results of my C++ program. It took me a couple of minutes to find all this information in R, while it took me around half an hour to find the equivalent information using custom made functions in C++. It is clear to me that R is a more practical programming language to use for finding linear regression and comparing data sets. The built-in functions in R save a lot of time and make finding information about huge data sets a lot easier, not to mention in C++ it took over 50 lines of code to read the data from a .csv file while in R it can be done in one line.

c. describe the descriptive statistical measures mean, median, and range, and how these values might be useful in data exploration prior to machine learning

- Mean: is the average of a set of numbers, it can be found by taking the sum of the list and dividing it by the number of items in the list. Having a mean value is useful outside of machine learning as it can show the average value of a dataset as a whole considering all the individual points within the data set. In other words, it is a value that considers all other numbers in the set.
- Median: is the middle value of a set of numbers, it can be found by sorting the list, and taking the middle value if the list has an odd number of items, or the average of the two middle values if the list has an even number of items. Median can be useful outside of machine learning because the median is the central point of a data set, and when it is like the mean it shows that the values in the data set are evenly distributed with few outliers. Rm and Medv are evenly distributed as their means and medians are close together.
- Range: is the difference between the minimum value and maximum value in a set of numbers, it can be found by sorting the set of numbers from lowest to highest and subtracting the first number in the set from the last number in the set. Range was useful in data exploration prior to machine learning as it established a boundary on a set of data, showing the spread of values in a data set. A large range can mean that values spread out widely or there are outliers within the data set while a small range can mean that data values are clumped together. Rm is clumped together, while Medv is more spread out.

d. describe the covariance and correlation statistics, and what information they give about two attributes. How might this information be useful in machine learning?

- The covariance statistic describes how changes in one set of variables affects changes in another set of variables.
- The correlation statistic is the same as the covariance statistic but scaled to a value between -1 and 1, with 1 being complete correlation, 0 being no correlation, and -1 being opposite correlation. Rm and Medv have a positive correlation close to 0.7 which means that as Rm increases Medv tends to also increase.
- This information would be useful in machine learning because, if two data sets have a correlation close to zero or a covariance close to zero then it is unlikely to be useful when training a computer to predict on those data sets as the values in the sets don't relate to each other. When training a machine, it is better to have correlation closer to 1 or -1 as that means the values in the sets are more linearly related either positively or negatively. When the values in a set are linearly related it becomes easier to predict new values added to the sets.