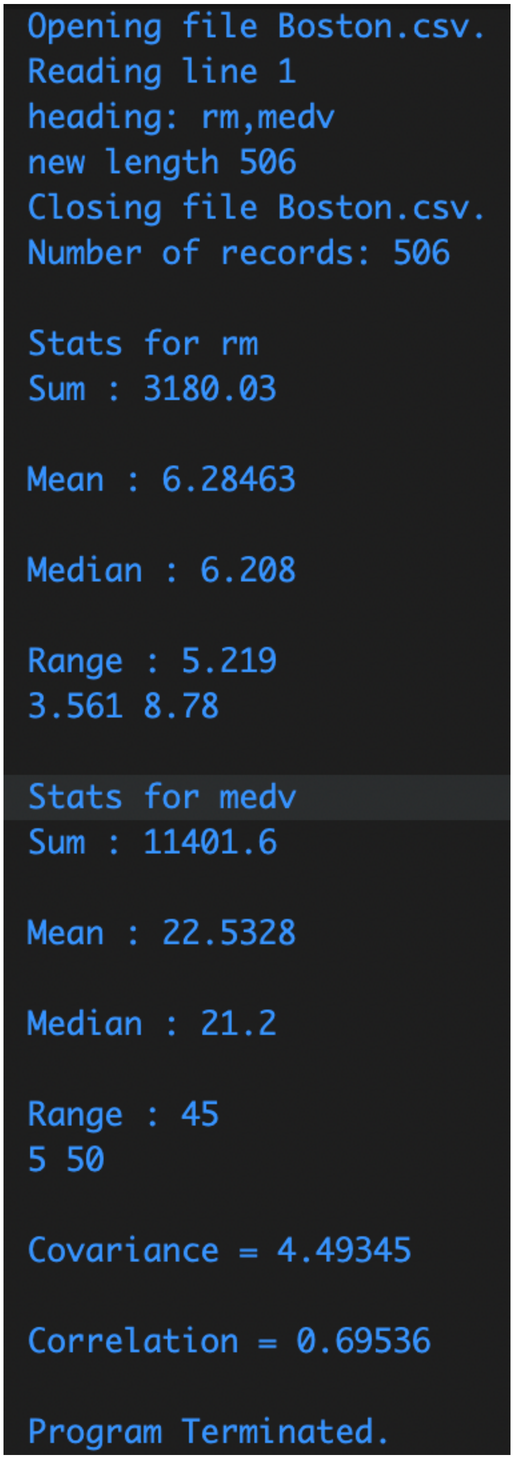
Portfolio Component 1: Data Exploration

Here is the output of my code.



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

Stats for rm
Sum : 3180.03

Mean : 6.28463

Median : 6.208

Range : 5.219
3.561 8.78

Stats for medv
Sum : 11401.6

Mean : 22.5328

Median : 21.2

Range : 45
5 50

Covariance = 4.49345

Correlation = 0.69536

Program Terminated.
```

My Experience Using Built-in Functions in R vs. Coding My Own Functions in C++

Using built-in functions in R is obviously a lot easier to use since everything is already implemented for me. However, implementing own functions to emulate R’s functionality has helped me better understand how R is manipulating the values we give it. I found this especially true when I was implementing the correlation and covariance functions. Even so, none of the functions were difficult to understand/implement.

Mean, Median, Range and How They’re Useful in Data Exploration

The mean is the average of a data set. The median is the middle value of a data set when sorted from least to greatest. Range is the difference between the highest and lowest values in a data set.

As far as data exploration is concerned, the mean is useful since it provides a single number representative of the data set: a typical value. The median is useful since it provides an idea of where the center of a data set is. Comparing the mean and median can help give us an idea of how evenly distributed a data set is. The range gives us a sense of the spread of the data set; larger range meaning the data is more spread out.

Covariance and Correlation Statistics and What Information They Give About the Two Attributes

Covariance is a measure of how much two variables vary with another. This essentially measures the extent the variables change with one another. Correlation is a measure of how strongly related two variables are which tells us how well one variable can predict another. Correlation is just a scaled form of covariance.

This information is useful for machine learning since correlation could imply that one variable is a good predictor for another and ML could use that information to make predictions. Covariance would be used to calculate the actual predicted value.