

## Portfolio Component 1: Data Exploration

The output of the program:

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

Stats for rm
Sum: 3180.03
Mean: 6.28463
Medium: 6.209
Range: [3.561, 8.78]

Stats for medv
Sum: 11401.6
Mean: 22.5328
Medium: 21.2
Range: [5, 50]

Covariance = 4.49345

Correlation = 0.69536

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

### Experience using built-in functions versus coding my own in C++:

I found that it's much easier to use the built-in function that R provides to us. In C++, I had to manually look up formulas as well as write code for them which definitely can take some time if an equation is complex enough.

**Mean** - describes the average number in a given data.

**Median** - the number that is located in the middle of sorted data.

**Range** - the difference between the maximum value and the minimum value in data.

Mean, median, and range are very useful in data exploration because they act as the fundamental values in a lot of formulas that are used for finding any insights about data. For example, they help us detect outliers in data, where it's centered, or if it's skewed in one direction or another.

**Covariance** - measures how two variables are related to each other.

**Correlation** - measures how several variables in the dataset are related to each other.

Covariance and correlation are important in machine learning for detecting patterns in data sets. For example, you need to find the correlation of data sets in order to make a linear regression model to be able to make predictions from data.