Portfolio Component 1: Data Exploration

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: 6.2085
Range: 5.219

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

 Covariance = 4.49345

 Correlation = 0.69536
```

My experience using built-in functions in R versus coding my own functions in C++

The ability to call functions such as covariance or correlation of course simplifies the entire process, but actually coding them gives you a better understanding of what the code actually does. Most of the functions I coded myself in this program were basic math functions (sum, mean, median, range), which at the end just seemed like a waste of time.

Description of descriptive statistical measures *mean*, *median*, and *range* and how these might be useful in data exploration prior to ML

Mean

The mean is the average of a set of numeric values -> (x1 + x2 + ... + xn) / n
-> {1, 3, 5, 8, 16, 22, 69} -> (1 + 3 + 5 + 8 + 16 + 22 + 69) / 7 = ***17.71***

Median

The median is the middle value of a *sorted* set of numeric values -> {1, 3, 5, ***8***, 16, 22, 69}

Range

The range is the maximum value subtracted from the minimum value of a set of numeric values -> {***1***, 3, 5, 8, 16, 22, ***69***} -> 69 - 1 = ***68***

How these might be useful in data exploration prior to ML

These measures give you a basic idea of how the data looks without having to completely analyze it youself. While the mean just gives a basic average, the median can be more useful since it is less susceptile to outliers, such as in the example above. Here, the mean is 17, which poorly represents the set of numbers, since it is strongly pulled up by the oulier. The range gives a basic understanding of how far the values range in the set of values.

Description of covariance and correlation statistics & what information they give about two attributes.

Covariance

The covariance describes the relationship between two attributes and measures the direction in which it is going. It measures how changes in one variable are associated with changes in the other variable.
Positive covariance is the result of one variable's increase resulting in an increase of the other variable. They both move into the same direction. If one variable increases resulting in the other variable decreasing, the covariance will be negative.

Correlation

The correlation is covariance scaled to [-1,1]. It describes the degree to which two attributes are affecting each other.

How might this information be useful in machine learning?

It makes us better understand data and therefore able to create algorithms to analyze data the way a human mind would.