Kevin Cordano
Data Exploration
CS 4375.004

Program Output:

```
C:\Users\kervi\Desktop\CS 4375\Portfolio - C++ Data Exploration>"Data Exploration.exe"
Number of records: 506

Stats for rm
        Sum:     3180.03
        Mean:    6.28463
        Median:  6.209
        Range:   5.219

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

Covariance = 4.49345

Correlation = 0.69536

Program terminated
```

Naturally, it was much easier to retrieve the statistics in R than in C++ since R already has built in functions. However, it is beneficial to have control over the function and being able to manipulate the logic within.

Mean is the average value of a data set, which is calculated by summing all elements and dividing by the number of elements. Median is the middle value of a data set, which is calculated by sorting the set and selecting the element that is at the N/2 index. Range is the difference between the maximum and minimum value of a data set. These statistics are useful outside of machine learning for many reasons. They provide a short but descriptive overview of a data set; one can get an idea of the size and spread of data based on these 3 values

Covariance is the value that describes the directional relationship between two variables; variables that move in the same direction have positive covariance, and variables that move in opposite directions have negative covariance. Correlation is a form of covariance that is scaled to be between -1 and +1. -1 correlations are perfectly inversely correlated; as one variable increases, the other decreases. +1 correlations are perfectly correlated; both variables increase/decrease together. This is useful in machine learning because significant correlations can identify patterns in data that allows for intelligent processing.