```
PS C:\Users\sunni\Documents\GitHub\CS4375-Portfolio\component_1> g++ data_exploration.cpp
PS C:\Users\sunni\Documents\GitHub\CS4375-Portfolio\component_1> ./a.exe
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
Program terminated.
PS C:\Users\sunni\Documents\GitHub\CS4375-Portfolio\component_1>
```

CLI output of data_exploration.cpp

I prefer this C++ implementation than R's built in functions. This is because I am more familiar with C++ and this allows me to better understand and conceptualize the math behind the models.

The mean and median can help indicate where the data is centered. Although they are depicting similar insights in the data, they are may not always be so similar. The mean for example, is very sensitive to outliers in the data which will lead to skewed data. Similarly, if the midrange data is uniformly distributed, the median will inaccurately estimate the average of the data. While not directly useful in machine learning, they are often part of higher level mathematical models.

The range of the data helps show the spread of the data, however, this is also highly susceptible to outliers, which may cause inaccurate estimates.

Covariance and correlation both tell us if and how much two data sets are related to each other. Covariance indicated how the two variables are related to each other. Correlation is a standardize metric of covariance with +1 being a perfect dependent relationship. Covariance and correlation provide a large amount of insight into the data which can be used in many different machine learning models. For example recommendation models rely highly on the amount of correlation between two examples.