Part A

~/school/ML/homework/01-data-exploration $ ./data-exploration.out
Opening file
Reading line 1
Heading: rm,medv
New length: 506
Statistics for
Sum: 3180.03
Mean: 6.28463
Median: 6.2085
Range: 5.219

Statistics for
Sum: 11401.6
Mean: 22.5328
Median: 21.2
Range: 45

Covariance: 4.49345

Correlation: 0.69536

Part B

Using built-in functions in R provided more abstraction when thinking about how the statistical measures were calculated. I did not have to consider how the numbers were being calculated in R, and instead could focus on the implications of those numbers. In C++, since I had to write the statistical functions myself, I was less focused on the implications of the statistical measurements.

Part C

The mean is the average of all the data points in a vector. The mean gives insight into a typical value for the data points. The mean is susceptible to skew from outliers. The median is the middle data point when the data points are sorted. The median also gives an idea of a typical value for the data points, but is significantly less affected by left and right skew. The range is the difference between the largest data point and the smallest data point. The range can give insight into how widely a specific data vector can vary from either the mean or the median. These statistical measures are useful prior to machine learning because they provide a general understanding of the characteristics of the data set.

Part D

Covariance and correlation both give insight into how closely two data vectors can predict the other. Covariance is dependent on the scale of the numbers in the data vectors, whereas correlation is scaled from -1 to +1. When correlation is close to -1 or +1 it indicates that the data vectors are good predictors of the other, and when the correlation is close to 0 it indicates that

the data vectors are poor predictors of each other. Covariance and correlation are useful in machine learning because they guide us in the construction of models, by making it more obvious which data columns can be used to predict the target column.