Data Exploration

```
{cslinux1:~/data_exploration} g++ -Wall -O -g -std=c++11 explore.cpp
{cslinux1:~/data_exploration} ./a.out
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
```

Writing these functions myself rather than use the built-in features of R is clearly worse. Not only did it take extra time to write, but my code is certainly less optimized than that of R. These functions also are not so complex that a student should need hands-on experience to understand them, though the silver lining is that the assignment was quick and easy. Some of the functions in question were for finding the mean, median, and range. Mean refers to the average value, whereas median refers to the centermost value; these are often similar, but if they differ it is likely due to outlier and leverage points that must be taken into account for creating a regression. Range is the difference between the maximum and minimum values; in R, this statistic shows both the maximum and minimum values, though I only implemented the statistical definition. A large range could mean that there is a good amount of diversity in the data, which could allow for a more generalizable model. If the large range is due to leverage or outlier points, it could actually decrease the predictive power instead. From these fundamental functions, the covariance and correlation values can be calculated. Covariance measures the relationship between two variables, while correlation scales covariance to be between -1 and 1. The former can be rather difficult to discern, and the latter more clearly quantifies the relationship; coefficients close to 1 or -1 show strong positive or negative relationships, while values close to 0 show little association between the parameters. High correlations between the values from a linear model and a set of validation data indicate a regression with strong predictive power.