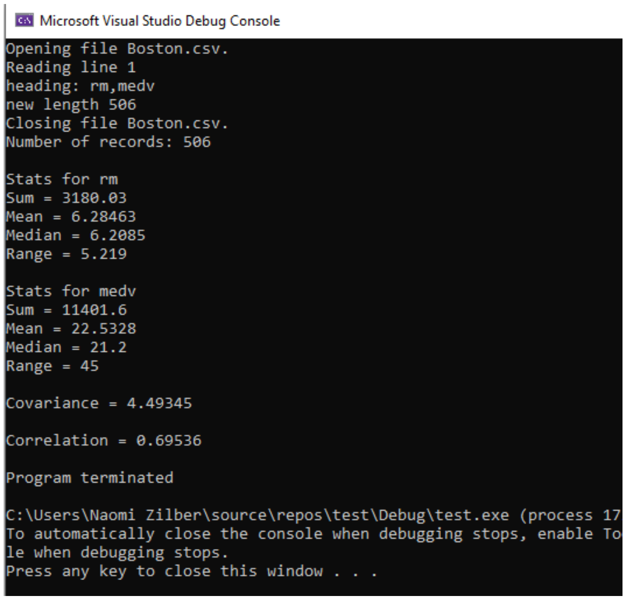
Data Exploration

Output of 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

Program terminated

C:\Users\Naomi Zilber\source\repos\test\Debug\test.exe (process 17
To automatically close the console when debugging stops, enable To
le when debugging stops.
Press any key to close this window . . .
```

Reflection:

Overall, the built-in functions in R are more convenient than having to code the same functions in C++. Coding the basic functions for sum, mean, median, and range is simple enough, but functions such as covariance and correlation take more time/code, unlike R which has a built-in function for each of them. Therefore, though I didn’t find it too challenging to code the functions, I grew to appreciate how R already has them built in.

Mean is the mathematical average of all the values in a data set, the median is the middle element of a set of data, and the range is the difference between the largest and smallest elements in a set of data. These values might be useful in data exploration because the mean can be used to summarize a data set into a single value that can represent the typical value for that data set, the median allows for another way to find the typical/central value for a data set, and the range tells about the spread of the data set, so it allows us to see how spread out the values in a data set are from each other. Therefore, the mean and median let us see the central tendency of a data set while the range tells us how far apart the values in the data set are.

Covariance measures how changes in one variable are associated with changes in a second variable, while correlation measures the linear correlation between two variables, which ranges from -1 to +1, where these correspond to a perfect negative correlation and a perfect positive correlation, respectively, and values close to 0 indicate little to no correlation. This information might be useful in machine learning because it helps gain an understanding of how one variable is related to another, which is an important step during data analysis. This information then lays the foundation for creating a model to represent that data set and build predictions from it.