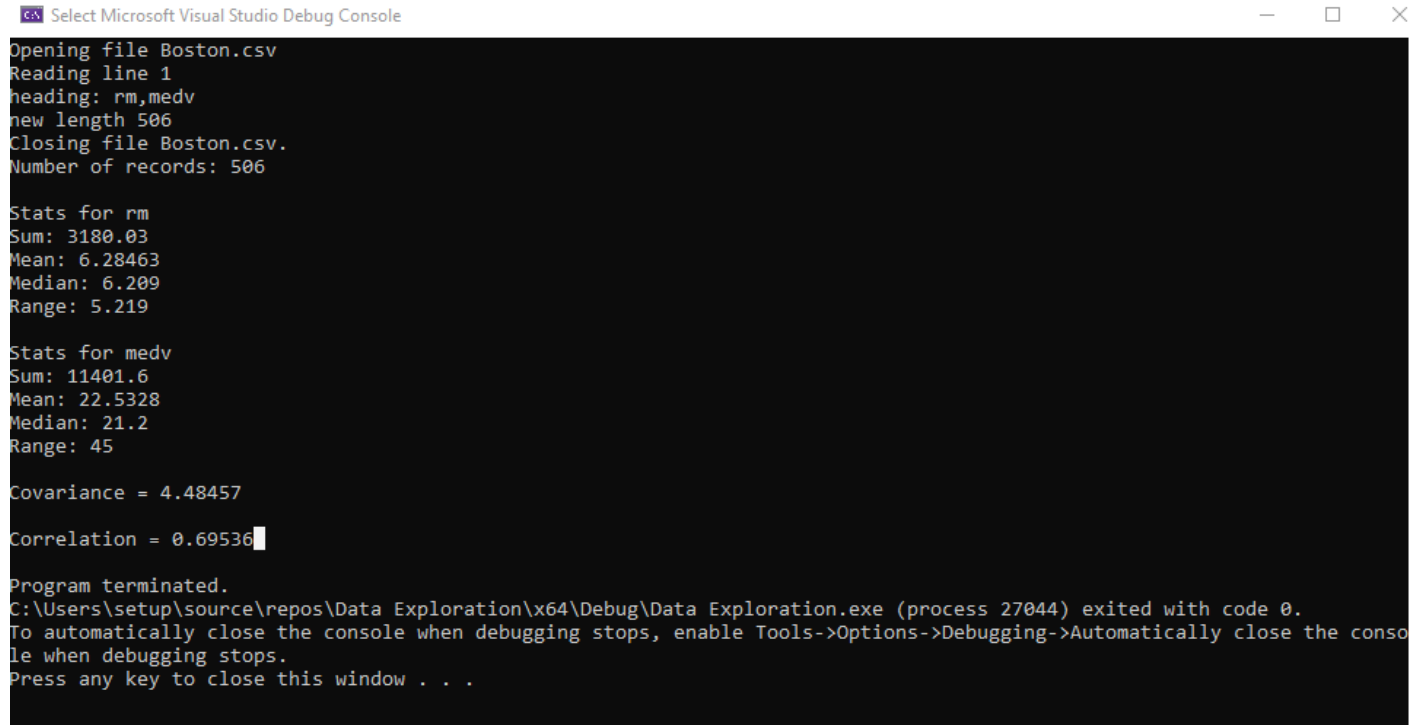


Assignment #1: Data Exploration

a. Below is the output of my C++ code:



```
Select Microsoft Visual Studio Debug Console

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: 6.209
Range: 5.219

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

Covariance = 4.48457
Correlation = 0.69536

Program terminated.
C:\Users\setup\source\repos\Data Exploration\x64\Debug\Data Exploration.exe (process 27044) exited with code 0.
To automatically close the console when debugging stops, enable Tools->Options->Debugging->Automatically close the console when debugging stops.
Press any key to close this window . . .
```

b. Using the built-in functions in R was much easier even though I had experience in C++ programming and no experience in R. The functions are extremely helpful to easily calculate and implement statistical calculations. Even though these were easy tasks to complete in C++, it felt as if I was completing a useless task after doing it in R. I understand the purpose of this lab in showing this, and I definitely understand how helpful R can really be when doing machine learning due to the vast amount of calculations that must be done throughout these processes.

c. All three of these values would be crucial for the initial data analysis to find the patterns and other aspects of the data set that will be fed to the machine. The mean allows us to see the average value of the data set which will most likely be what the algorithm will be focused on. The median can be helpful if we are using data that is too spread out and the outliers are too inconsistent. In some cases, the median may be a more useful statistic to describe the data set than the mean. The range lets us know how significantly the outliers can really affect the data and let us know if the dataset needs to be looked at. All of these statistical measurements are much simpler however they are perfect to use to examine the data prior to any machine learning.

d. The covariance measures the variability between multiple data samples. In machine learning this would be a useful statistic for the machine to use to improve itself. The machine can continuously be comparing the datasets in order to optimize itself. The correlation measures the effect that the change of one variable has on another. This variable is also very important in machine learning because throughout the process, the computer can be aware of the strength of the effect that certain changes have on other values.