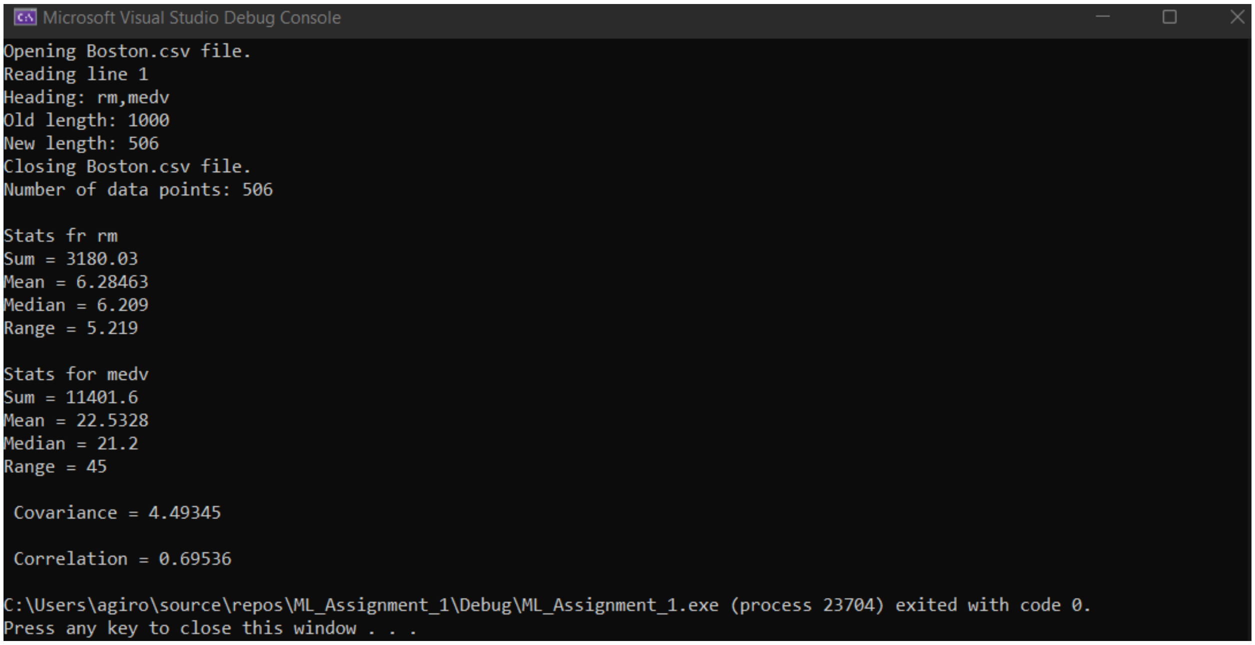
Austin Girouard

Atg180001

Dr. Karen Mazidi

Assignment 1 Breakdown Document

1.



```
Opening Boston.csv file.
Reading line 1
Heading: rm,medv
Old length: 1000
New length: 506
Closing Boston.csv file.
Number of data points: 506

Stats fr 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 = 4.49345

 Correlation = 0.69536

C:\Users\agiro\source\repos\ML_Assignment_1\Debug\ML_Assignment_1.exe (process 23704) exited with code 0.
Press any key to close this window . . .
```

2. The built-in functions in R make calculating more tedious formulas like covariance and correlation MUCH easier. With C++, most of these functionalities are not built-in, and if they are, they are tucked away in a random library that you need to dig for. Having the ability to calculate these types of formulas quickly and efficiently makes R a far more powerful tool when it comes to this type of data analytics.

3.

Mean: The average of all values within a set of numbers. This is useful for finding the “most likely” estimate for a value from a data set without doing deeper statistical analysis. It is also used in important calculations like finding covariance.

Median: The middle element of a set of numbers. This is useful when a data set is skewed in either direction because it can find the middle estimate more accurately than an average which is offset by outliers.

Range: Range: The difference between the highest and lowest values within a data set. This is useful for detailing how wide the values in a data set are and, when used in statistical analysis such as box plots, it is easier to see how far some outliers drift from the middle 50% of data.

4.

Covariance: Covariance measures how changes in one variable associate with changes in a second variable. This is useful for determining how two variables are related as one increases/decreases.

Correlation: Correlation shows how much of a pattern exists in a particular data set, where 0 is “no correlation” and [-1, 1] are “perfect correlation.” Correlation is covariance scaled down to [-1, 1] which becomes more useful as the range in covariance becomes wider, making it the often-preferred statistic.

Both of these statistics are useful in machine learning because one of the main goals of ML is pattern recognition in data. Correlation and covariance help us understand the patterns that may exist between two different data sets and how strongly their relationship is.