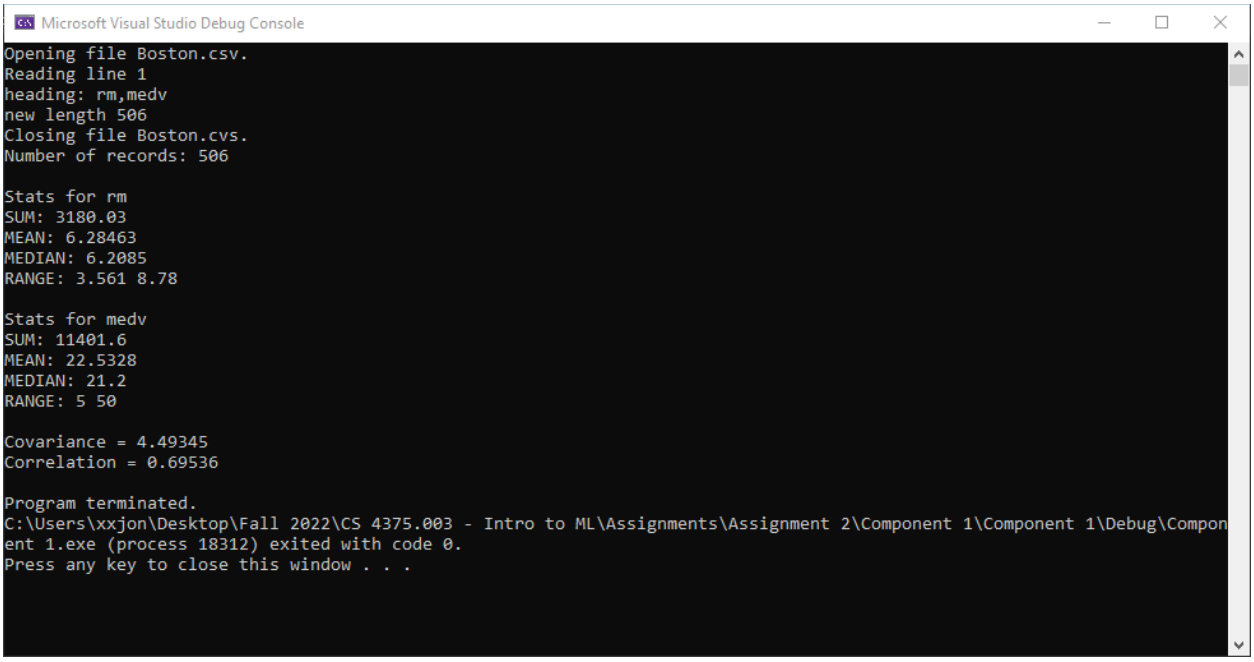
Jonathan Ho

CS 4375.003

Component 1 Write-Up

1. Output of the code:



2. Using the built-in functions of R is much more convenient than coding the functions in C++. However, the actual coding itself was ok. While coding the functions, there were two things I had to pay attention to. First, the Covariance used was the sample version, which required the summation to be divided by the number of data points minus one. The other was how the range was setup. Basing how the range function works in R, it returns the maximum and the minimum of the data set, and the data can be stored into another variable if desired. I had to make the function return a vector just incase a user wished to store the range into a vector.

3. Mean is the average of all the data given for one attribute of data. Seeing the mean helps to understand what is the data you would expect to see typically. Median is finding the “middle” of the data once it is all sorted. This would help to perceive what is the data most often recorded when the data set is potentially favoring one range of data over another. Finally, there is range, which given how it is used in R, it is the minimum and maximum of the data set. The range can be used to see not only the boundaries of the data set, but how consistent your data could be. If the range is not far off from the usual data set, then it potentially helps with determining the precision of the data. However, if the range is wide, there is the possibility the data may not be as consistent or not as precise.

4. Correlation and covalence help to determine the relationship between two attributes. Correlation determines the linear correlation between two attributes. For example, a positive correlation would be if kids read more books, then they score higher grades on tests. Correlation is just a scaled version of covariance. Covariance looks at the proportionality of one attribute in relation with another. When using these statistical measurements in ML, this could look at the relationship between two variables in the data set. Correlation and covalence may help in creating the output model to try and predict the target.