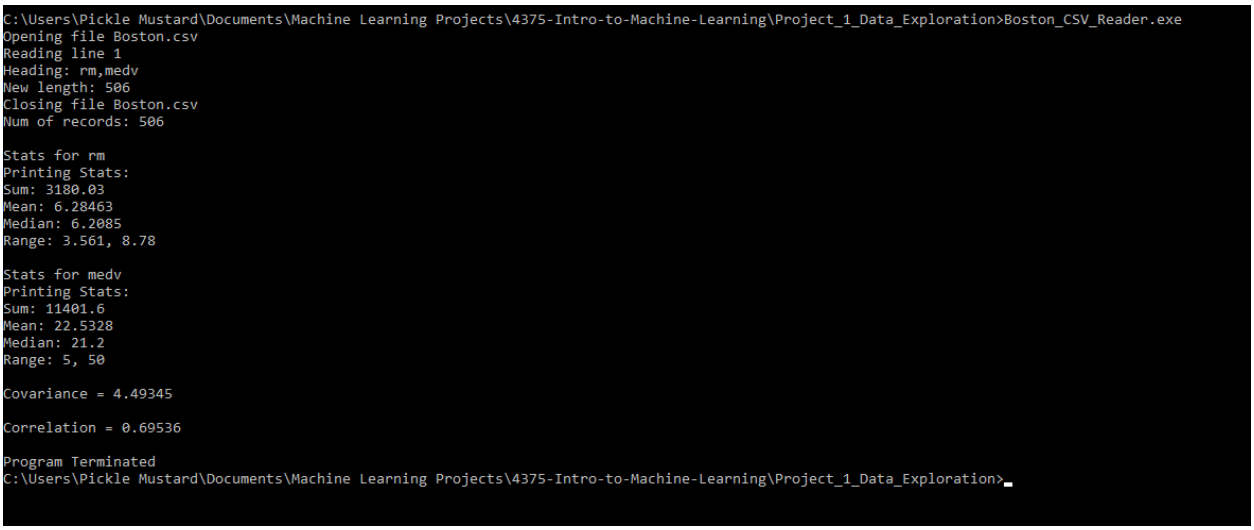
Dillon Carter

Dwc180002

Project 1 Data Exploration Write Up



Using the R functions is so much nicer than having to build them out yourself. The statistical calculations are not difficult themselves to calculate but there is a lot of repetition of actions between each one and it's somewhat tiring making sure that each is properly set out. I did have issues with finding the median but that was exclusively because I forgot the important step of sorting the set of values before finding the median.

The actual statistical measurements are useful for cursory glances at what the data set could provide. The sum isn't all that useful, but the mean and median give two measurements for approximating the center points of the data and the range gives an idea of how wide the data set spreads. Depending on the circumstance, the difference between the min and max of the range and the mean/median of the data can tell an analyst whether the data is useful at all to begin with. This is all for a single data set. When looking at multiple in relation to each other, covariance and correlation are also useful tools for understanding the relationship between two sets at a glance. The covariance between two sets shows the proportional change of variables between two sets and the correlation clamps the values two between -1 and 1. These are great for giving a broad indication of the relationship between the two sets so that if high correlation is found between them, they can be given greater focus over those that do not have as high of a correlation. An analyst cannot be looking at every relationship between every set of data, especially as datasets grow in size. So these cursory looks reduce the amount of time that is needed to understand which set relationships are important. The individual statistics also can tell an analyst if a set of data will be useful by the differences between the stats. Large differences out of place for the expected values of the set can tell the analyst that there are possibly issues with the data and may not be too useful for analysis.