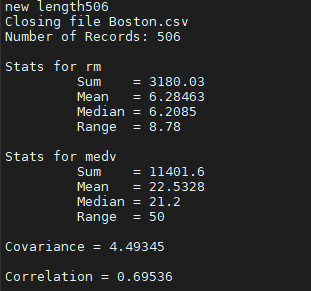
Subash Chandra – SXC200027

CS4375 – Machine Learning Homework 1

C++ Data Exploration



It was fairly simple to code my own functions in C++. The hardest part of this whole homework was debugging a multithreading issue which made it impossible to read the CSV. It was definitely easier to use the built in functions in R.

Mean is the average value of all the houses, but this value can be skewed heavily by outliers. This can be neutralized by taking into account the Median, which is less skewed by outliers. The range can help us narrow down our certainty as well as identify where the outliers are. Using all three of these values allows us to accurately deduce the tendency of the data.

Covariance is a way to measure how two variables differ, and correlation shows how well they are related. 0.695 means that the Number of rooms in the houses in Boston are ~69.5% correlated with the median value of that house. This allows us to better narrow down our parameters to better remove outliers and get better accuracy from our machine learning models.