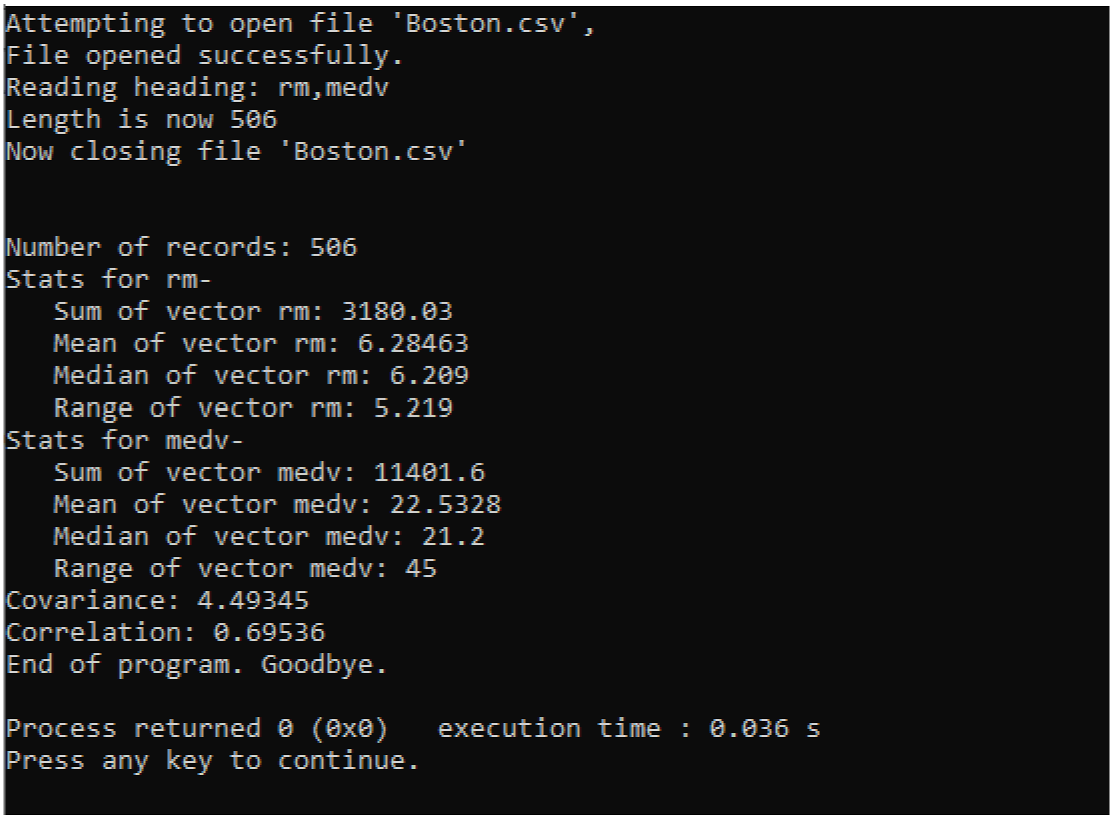
Gray Simpson
Machine Learning Assignment 2 - Data Exploration

Runs of Code



```
Attempting to open file 'Boston.csv',
File opened successfully.
Reading heading: rm,medv
Length is now 506
Now closing file 'Boston.csv'


Number of records: 506
Stats for rm-
   Sum of vector rm: 3180.03
   Mean of vector rm: 6.28463
   Median of vector rm: 6.209
   Range of vector rm: 5.219
Stats for medv-
   Sum of vector medv: 11401.6
   Mean of vector medv: 22.5328
   Median of vector medv: 21.2
   Range of vector medv: 45
Covariance: 4.49345
Correlation: 0.69536
End of program. Goodbye.

Process returned 0 (0x0)   execution time : 0.036 s
Press any key to continue.
```

Experience Using R vs C++ For Data Exploration

In comparison to C++, R is much simpler and straightforward for looking through data and gathering information on it. Everything is centralized, and the data can be looked at fairly simply instead of having to open up the data in an external file when it could be quite long. Coding it from scratch in C++ wasn't necessarily difficult, or particularly time-consuming, but it was enough so to see how much more straightforward R is for this task. R is prepared to do many tasks so there are less algorithms to remember (though understanding them will always help with understanding data and what everything means).

Usefulness of Mean, Median, and Range

The mean, median, and range are straightforward values good for giving someone an idea of what standard values look like in a data set. The mean would be the sort of expected value, and both the median and range help towards understanding how much that expected value can vary, as well as hint towards whether or not there are outliers. With or without machine learning, these are important to know to understand data. Prior to machine learning, this was important to figure out what needed to be done next, and what uses the data could have– via characteristics, commonalities, and the like, so trends can be understood.

Covariance and Correlation

Covariance is a number to represent the linear trends of two pieces of data. It is important in calculating the correlation coefficient, which is normalized and therefore more straightforward to interpret. It will always be between -1 and 1. A number such as 1 would correspond to a perfect correlation– both go up equally. A number like -1 would mean they correspond perfectly inverse. This lets us know the trends between lot of data at a glance. The closer to 0 it is, the less the data has to do with each other at all.