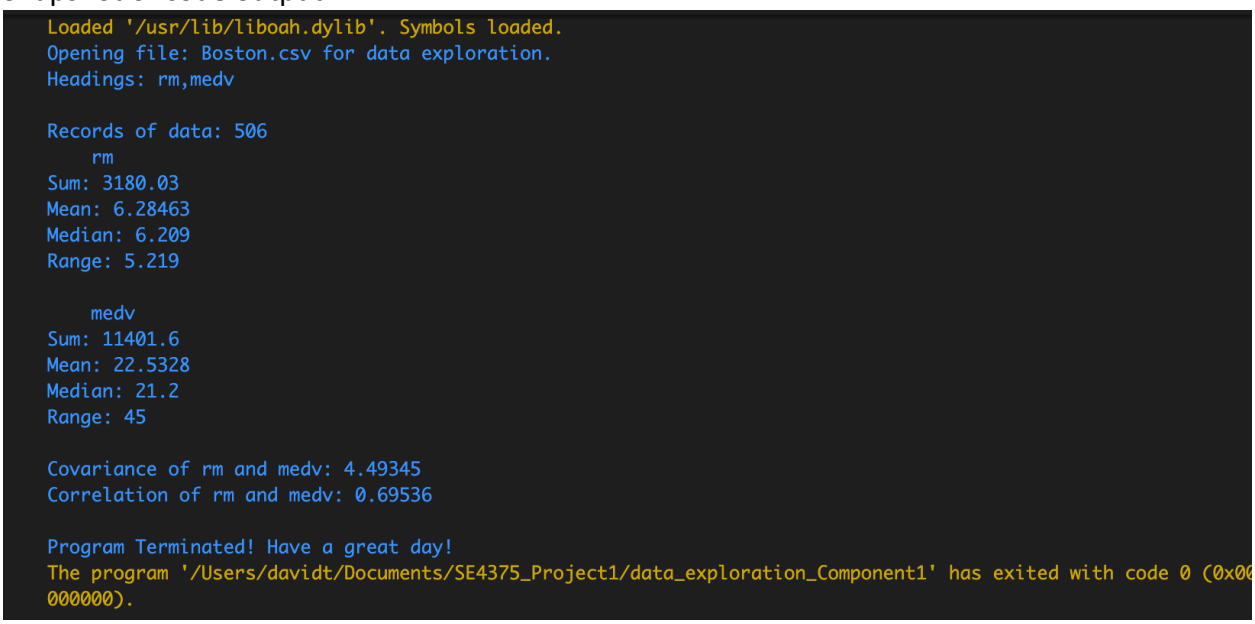
David Teran
February 4, 2023
CS 4375.004

Portfolio Component 1: Data Exploration

This program will read the Boston.csv file and reproduce some of the same functions as the built-in functions in R Studio. There are 7 functions in total, with six of them calculating the following statistics: sum, mean, median, range, correlation, and covariance. The remaining function will print out the results of the calculations for the data.

The Boston.csv file will be read in by the program and will be parsed and set as part of 2 different vectors, for both columns of data: rm and medv. Once parsed and assigned into their respective vectors, the data is calculated for both vectors separately. The covariance and correlation is calculated using both vectors. Once finished, the program will print the results onto the console.

Snapshot of code output



Experience Using R Built-in Functions vs. Coding Functions in C++:
A good refresher in programming using C++, but also not too bad to program. Using the built-in functions in R is much easier and convenient than programming the functions themselves in C++. It was still a good way to get back into remembering how to program in C++ after using Java for a while.

Statistical Measures:
The mean, or the average of values in a set of data, is useful in data exploration to determine patterns in sets of data, whether the average is rising or falling. The range, or the difference between the minimum and maximum values in a data set, can also help in determining how far apart the data values. The median, or the middle values of a set of data, can also determine the average values as well. All three descriptive statistics can be useful in data exploration for

determining how varied the values are in a data set, how apart they are, and can help determine a pattern or relationship between data sets.

Covariance and Correlation:
The covariance helps determine how changes in one variable can be associated with the changes of a second or other variable, and how these variables are related to each other. Correlation determines the extent of how two variables or data sets are related to each other. The information from both covariance and correlation can be used to predict the future behavior of data sets or variables, which is useful in machine learning.