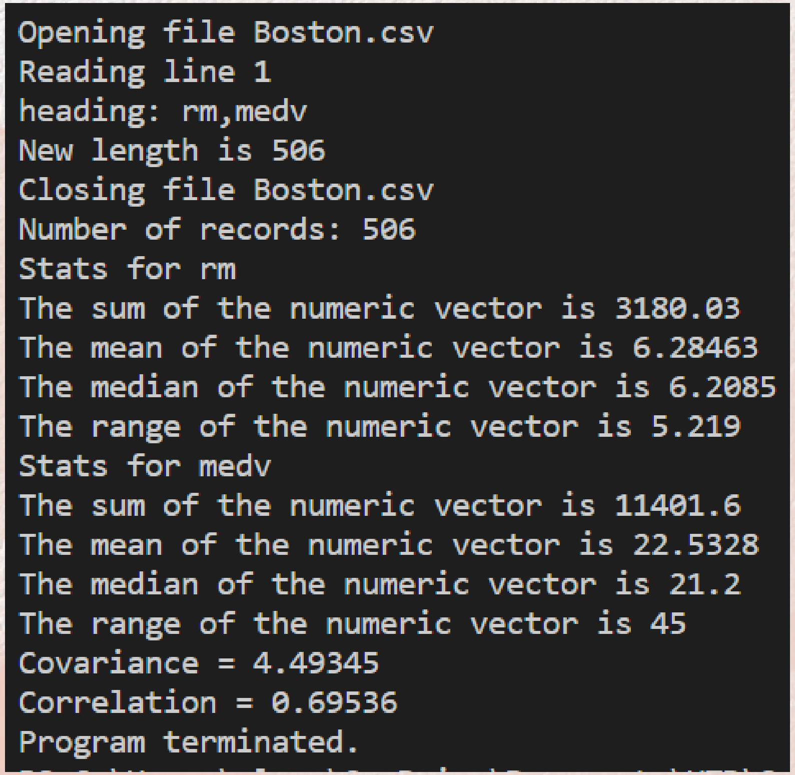
Portfolio Component 1: Data Exploration



```
Opening file Boston.csv
Reading line 1
heading: rm,medv
New length is 506
Closing file Boston.csv
Number of records: 506
Stats for rm
The sum of the numeric vector is 3180.03
The mean of the numeric vector is 6.28463
The median of the numeric vector is 6.2085
The range of the numeric vector is 5.219
Stats for medv
The sum of the numeric vector is 11401.6
The mean of the numeric vector is 22.5328
The median of the numeric vector is 21.2
The range of the numeric vector is 45
Covariance = 4.49345
Correlation = 0.69536
Program terminated.
```

For the last two semesters, I have been using R which has a lot of built-in functions that can be used to calculate statistical measures such as the sum, mean, median, range, covariance, and correlation. I noticed that using built-in functions was very easy in comparison to having to write your own functions. If we were to have to write a function each time, it would be very inefficient. This assignment made me realize how useful built-in functions in R can be.

There are various statistical measures that can be used in machine learning:

Mean: also known as the average, it is the sum of all the elements of a vector divided by the number of elements

Median: the middle number, or the average of the middle two, when the vector is sorted into ascending or descending order.

Range: the difference between the highest and lowest value. Can tell us how spread out the data is.

Although the mean is more susceptible to outliers(extremely large or small values), the median is not. If the mean and median are similar it tells us that there is an absence of outliers. However, if they are very different that tells us that there might be outliers in the data. These statistical measures can be used in normalization or feature scaling. Sometimes we can replace N/A values in a data set with the mean or median.

Covariance and correlation tell us about two variables in comparison to one another, rather than just one variable. The covariance tells us the direction of a linear relationship such as positive or negative. The correlation depends on the covariance. It tells us how positively or negatively correlated two variables are. For example, a correlation of -1 is perfectly negatively correlated, 1 is perfectly positively correlated but 0 shows no correlation. These can be used in feature selection and used to decide which machine learning algorithm to use.