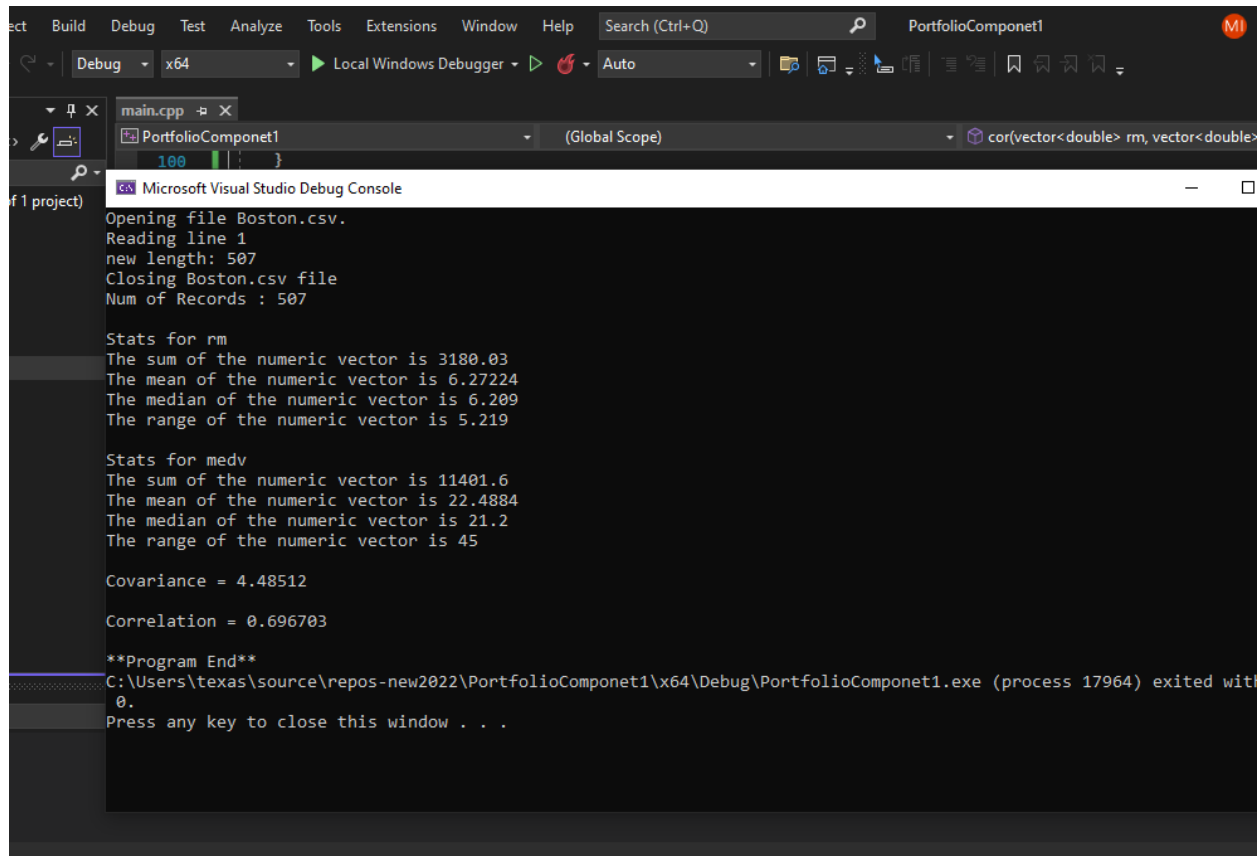


## Portfolio Component 1: Data Exploration

The image is a screenshot of the Microsoft Visual Studio IDE. The top menu bar includes 'File', 'Edit', 'Build', 'Debug', 'Test', 'Analyze', 'Tools', 'Extensions', 'Window', and 'Help'. The 'Debug' menu is open, showing options like 'Local Windows Debugger', 'Auto', and 'Run'. The 'x64' architecture is selected. The 'PortfolioComponet1' project is open, and the 'Global Scope' is selected. The 'Microsoft Visual Studio Debug Console' is open, displaying the following output:

```
Opening file Boston.csv.  
Reading line 1  
new length: 507  
Closing Boston.csv file  
Num of Records : 507  
  
Stats for rm  
The sum of the numeric vector is 3180.03  
The mean of the numeric vector is 6.27224  
The median of the numeric vector is 6.209  
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 22.4884  
The median of the numeric vector is 21.2  
The range of the numeric vector is 45  
  
Covariance = 4.48512  
Correlation = 0.696703  
  
**Program End**  
C:\Users\texas\source\repos-new2022\PortfolioComponet1\x64\Debug\PortfolioComponet1.exe (process 17964) exited with 0.  
Press any key to close this window . . .
```

When it comes to finding statistical observations, I heavily prefer using R rather than coding my own functions in C++. This is because with R it is much easier and quicker to get statistical observations such as correlation and covariance. I found it more challenging to make functions in C++ rather than using R, but it was made easier by fact that there are some vector functions in C++ which I used. But R's superiority when it comes to statistical coding can not be denied especially when it comes to graph making.

The three descriptive statical measures: mean, median, and range are very useful in data exploration prior to machine learning. To give context: mean is the average of a set of values, which is found taking the sum and dividing it by the number of values. The median is the middle value of a sorted set of values. And the range is the difference between the maximum value and the minimum value. These three measures are useful in data exploration because they provide keen insight into the characteristics of the data. The mean and median allow people to see the 'average' characteristics of the data and the range allows us to see the data's disbursement along the entire data segment.

The covariance statistic shows the relationship between two different variables and how the change in one variable effects the other. It shows how much they change as a pair. Correlation on the other hand, simply refers to the relationship between two variables. With correlation we can

see to what extent two variables are related to each other. This information is useful in machine learning because machine learning works by analyzing data and its patterns and making predictions. Covariance and Correlation are both important to this because they show relationships between two separate variables and help establish data patterns. Covariance can show the degree to which two different variables change together and Correlation shows the relation levels between two different variables, meaning that with this information a ML program can look at 1<sup>st</sup> variable and give a general prediction of how the 2<sup>nd</sup> will look. This makes these two statistics very important.

### References

Mazidi, Karen. “. The Craft of Machine Learning.” Machine Learning Handbook Using R and Python, 2nd ed.