Abigail Smith

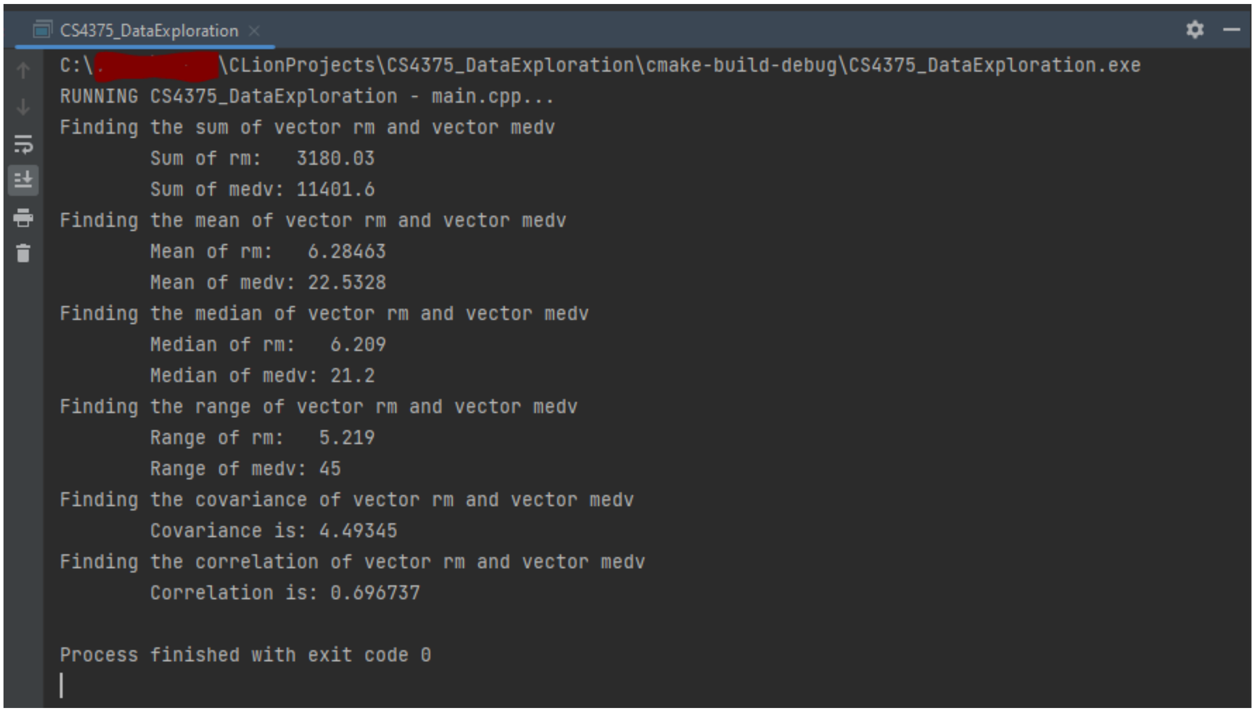
ARS190011

CS 4375.004

Dr. Mazidi

C++ Data Exploration

Output screenshot:



```
C:\[redacted]\CLionProjects\CS4375_DataExploration\cmake-build-debug\CS4375_DataExploration.exe
RUNNING CS4375_DataExploration - main.cpp...
Finding the sum of vector rm and vector medv
        Sum of rm:   3180.03
        Sum of medv: 11401.6
Finding the mean of vector rm and vector medv
        Mean of rm:   6.28463
        Mean of medv: 22.5328
Finding the median of vector rm and vector medv
        Median of rm:   6.209
        Median of medv: 21.2
Finding the range of vector rm and vector medv
        Range of rm:   5.219
        Range of medv: 45
Finding the covariance of vector rm and vector medv
        Covariance is: 4.49345
Finding the correlation of vector rm and vector medv
        Correlation is: 0.696737

Process finished with exit code 0
```

Written responses:

In terms of data exploration, there is a noticeable difference in how R and C++ are used to calculate statistical measures. While I was able to use built-in functions in R, calculating the sum, mean, median, range, covariance, and correlation for Boston.csv required a thorough understanding of each measurement's formula in order to recreate what R could do through built-in function calls. While the convenience of R is a notable bonus in terms of data exploration, I found that completing this assignment in C++ was a great hands-on refresher on not only how these measurements are calculated, but their importance in understanding the potential relationships that exist in a data set. Overall, C++ required much more code compared to R to calculate the sum, mean, median, range, covariance, and correlation for Boston.csv.

Prior to machine learning, the statistical measurements mean, median, and range both give insight into a data set as well as provide the foundation for more complex and insightful statistical measurements. Mean, or the average of a data set found by summing all data points and dividing by the total number of observations, provides an expectation for the values represented in the data. In contrast, the median, calculated by selecting the middle-most element when all observations are ordered from smallest to largest, represents the exact center of a dataset. Comparing the median and mean shows whether the data set is balanced (mean and median are equal), or if the dataset is skewed (mean and median are not equal). Lastly, the range, or the difference between the highest value and the lowest value, can be used to understand the range of values in the data set and highlight extreme values when utilized in graphs like boxplot. Each of these measurements have held prominence in statistical analysis well before machine learning was developed as they gave quick insight into data from a wholistic view. By having a thorough understanding of each of these statistical measurements prior to training a machine learning application, programmers can foster a better understanding of the data set in question and the potential relationships existing in the data.

The statistical measurements covariance and correlation reflect how much change occurs in one variable that corresponds to changes in another. Both correlation and covariance reflect the same sentiment; however, while correlation is scaled [-1, 1], covariance is scaled [-inf., +inf.]. Though scaled differently, both correlation and covariance can be read the same way – the further away from zero, the more dependent the relationship. If covariance and correlation are negative, that means there is an inverse relationship between the values. In contrast, if correlation and covariance and positive, there is a direct relationship. If covariance and correlation are equal to or close to 0, there is a weak or lack of relationship.

In relation to machine learning, covariance and correlation help developers understand the relationships that exist within a given dataset. By using correlation and covariance, programmers can better train their selected machine learning scenario to be more accurate by having a better understanding of patterns that exist in the used data set. Without exploring covariance and correlation, developers may miss the relationships, or lack therefore, that exist in the data they are exploring, and thus, will not be able to develop a well-rounded machine learning application.