Alekhya Pinnamaneni

CS 4375.003

C++ Data Exploration

1.

```
Opening file Boston.csv.
Reading line 1
heading: rm,medv
new length 506
Closing file Boston.csv.

Stats for rm
        Sum of numeric vector = 3180.03
        Mean of numeric vector = 6.28463
        Median of numeric vector = 6.2085
        Range of numeric vector = 5.219

Stats for medv
        Sum of numeric vector = 11401.6
        Mean of numeric vector = 22.5328
        Median of numeric vector = 21.2
        Range of numeric vector = 45

 Covariance = 4.49345

 Correlation = 0.69536

Program terminated.%
```

2. In my experience with this assignment, using R's built-in functions for data exploration was much easier than coding your own functions in C++. Using R's built-in functions allowed me to view statistics about the data without requiring any knowledge about how the statistics are actually calculated. However, coding the functions on my own in C++ gave me the opportunity to learn how important statistical measures, such as covariance, standard deviation, and correlation, are calculated and the significance of these statistics.

3. The mean is the average of multiple data points, and it is calculated by taking the sum of all the data and dividing the sum by the number of data points there are in the sample. The mean can be used to determine how accurate a prediction is based on how far it deviates from the mean. The median is the data point in the middle of a stored set of data. If the sample has an even number of data points, the median is calculated by taking the average of the middle two data points. The median can be used before machine learning to establish a common, base data value that can be used later during machine learning. The range is the largest number in the data range minus the smallest number. The range can be used in data exploration before machine learning to get a sense of what the bounds of the predicted data should look like.

4. Covariance measures the direction (positive or negative) of the relationship between two variables. A positive covariance means that when one variable increases the other one increases as well. A negative covariance means that when one variable increases the other one decreases in response. This information can be useful in machine learning to make a more accurate prediction. For example, if the machine knows that there is a negative covariance between variables x and y and that the value of variable x tends to be low, then the machine can predict that the value of variable y will be high. Correlation measures the strength of the relationship between two variables and can be any value between 1 and -1. If the correlation between two variables is closer to 1 or -1, there is a strong relationship between those two variables. If the correlation is closer to 0, there is a weak relationship between the two variables. Correlation can also be used to make more accurate predictions in machine learning by determining to wat extent one variable can affect the value of the variable the machine is trying to predict.