C++_Data_Exploration

4375.004
Chris Talley

Running Code

I wrote my code in a single c++ file. I compiled it using the following command:

```
g++ data_exploration.cpp -o data_exploration.out -Wall
```

This is terminal output after executing the .out file:

```
➜ ./data_exploration.out
Opening file Boston.csv
Reading line 1
heading: rm,medv
new length 506
Closing file Boston.csv
Number of records: 506

Stats for rm
Sum:    3180.03
Mean:   6.28463
Median: 6.209
Range:  5.219
***************

Stats for medv
Sum:    11401.6
Mean:   22.5328
Median: 21.2
Range:  45
***************

 Covariance = 4.49345

 Correlation = 0.696737
```

R vs C++

In my experience, running functions in R is much more efficient. This efficiency allows for less reused code and frees up brain power to analyze the data. On the other hand, C++ has it's benefits. When implementing my own functions I gained a deeper understanding of the values. My intuition for the concepts of covariance and correlation has improved greatly through this exercise.

Statistical Measures

The following are basic statistical measures. They are useful in traditional statistical applications to gain an overview of the data set's shape. This allows for more insight to be gained from visualizing and plotting the data at a glance. These values are also the basis for more advanced measures such as Variance and Sigma.

Mean

The average of all values in the data set.

Median

The middle value of the data set when sorted.

Range

The largest value - smallest value, that is the distance covered by the data set.

Covariance and Correlation

Covariance is a measure of how changes in one value affect another. For example, there exists a covariance between class attendance and final exam grades. Correlation is covariance scaled to the interval [-1, 1]. This provides a much more useable and readable measure of the relationship between two values. These are applied in machine learning algorithms to determine the effect that changing weights has on the output accuracy and percision.