Data Exploration

Code Output for dataexploration.cpp

```
Running functions for rm vector ......
Vector size: 506
Sum: 3180.025
Mean: 6.28463439
Median: 6.2085
Range: 8.78 3.561

Running functions for medv vector ......
Vector size: 506
Sum: 11401.6
Mean: 22.5328063
Median: 21.2
Range: 50 5

Running functions for covariance and correlation ......
Covariance: 4.49344588
Correlation: 0.695359947

End of Program
```

Experience in using R and C++

R proved to be the better and easier option in understanding these functions. It also provided the output in a more user friendly manner. For C the time to implement the functions and then test them was much longer. The output was also not as user friendly as R since we were outputing to the console. But all of this is to be expected as R was built for statistics and CPP was not.

Definitions

What is Mean?

The mean is the total average of a list of numbers. The equation is: total sum / num of items

What is Median?

The median is the number right at the middle of the list of number. If the list contains an even number of items than the median is the average of the two numbers left and right of the median index.

What is Range?

The range is basically the smallest and largest number in a list of numbers. The equation is: largest number - smallest number. For this code however we simply listed the min and max.

How are these values useful for ML?

These values lend to us critical information regarding the distribution, spread, and center of the data. This information can be used to determine which algorithm to use that would best fit our data.

What is Covariance?

It is extremely similar to normal variance. Variance, however, only tells you how a single vairable varies. Covarience on the other hand tells you how two variables vary together.

What is Correlation?

It is basically a relation between two variables which describes how related two variables are based on their covariance. It shows a value between -1 and 1. where 0 is basically no correlation.

How are these values useful for ML?

Covariance and Correlation help us understand the spread and relation of our data. It is helpful for algortihms to determine which variables heavily effect the model prediction correctness, and which variables do not.