Data Exploration in C++

The purpose of this program is to read in a 2-column CSV file, store its data into 2 vectors, run various statistical functions on them, and display the results. Here is an example showing the code output:

```
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: 6.2085
Range: 5.219

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

Covariance = 6.31527

Correlation = 0.977287

Program terminated.
```

R vs. C++

In R, vector operations are built-in, easy to understand, and simple to use. In C++, however, these operations need to be hard-coded to be used. The basic vector operations, sum, mean, median, and range, are easy enough to write functions for using `for` loops and C++'s `sort()` function. More complex statistical functions like correlation and covariance are less simple, however. These functions involve numerous sub-calculations that aren't easy to code without a thorough understanding of their statistical meaning. Furthermore, testing whether these functions have been coded properly is a chore as well, since data frames typically comprise of hundreds of observations. Calculating covariance and correlation by hand to check the code's accuracy is virtually impossible, and is still arduous even when specifically testing on a smaller data frame. Meanwhile in R, programmers can be assured of the accuracy and efficiency of these built-in vector operations, and instead can focus their programming efforts on manipulating and exploring the data.

Mean, Median, and Range

Mean, median, and range are descriptive statistical measures that all describe the average of the data: mean is the mathematical average, median is the middle element of a sorted dataset, and range is the difference between the maximum and minimum values. The average of a dataset is valuable information because it represents the most typical, most normal value to be expected. This makes the average useful for making predictions on new or hypothetical data points. These statistical measures for calculating the average were likely useful prior to machine learning because they allowed for fairly accurate predictions to be made without necessitating particularly advanced algorithms.

Covariance and Correlation

While the aforementioned statistical values describe averages, it is also useful to be able to quantify how data varies from the average, because that allows for more accurate predictions on new data. This is what covariance and correlation signify. Covariance measures how much changes to one variable are related to changes in another variable. This can imply positive, negative, or no relationship between the variables. However, because the actual value of covariance is sensitive to the scale it is calculated on, a scaled version of covariance is preferred for easier interpretation. This is correlation, which uses the scale [-1, +1]: the closer the correlation is to +1 or -1, the more respectively positive or negative the relationship is between the variables, and the closer the correlation is to 0, the more random and unrelated the variables are. The formula for correlation between two variables x and y comprises of the individual variances of each variable and the covariance between them: $cor(x, y) = \frac{covar(x, y)}{\sqrt{var(x)} \bullet \sqrt{var(y)}}$. Covariance and correlation are useful in machine learning for how they quantify broader distribution patterns across data.