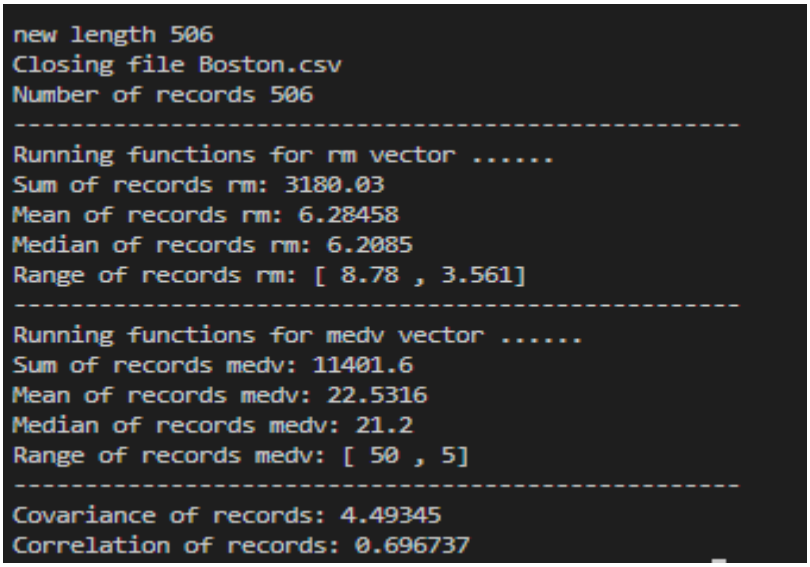
Portfolio C++ Data Exploration



Describe your experience using built-in functions in R versus coding your own functions in C++.

Generally, in my experience, it seems R is much more user-friendly than C++. The vectorization built-in to R's backend processing met with its built-in functions are really convenient especially when trying to make things efficient. I noticed in C++, it was a bit more work to go off and redefine these built-in measures from scratch, I found myself referencing statistical definitions from the slides frequently to ensure they were correct. Though C++ is very efficient and effective at decreasing run time and space time complexity in general, it comes down to the setup of how the transformative calculations are defined to get it right. In other words, while the performance benefit of C++ can sometimes outweigh the cost of building tools to perform the calculations, the setup cost of implementing things as efficiently as they can be expressed is sometimes high. In which case, practical languages such as R come to the rescue as its building upon a set of already efficient infrastructure in C.

As for which one was preferential over the other, it depends. I enjoy writing C++ code and I like breaking down problems to their atomic levels and manipulating operations there to see if there's any performance benefit hiding in plain sight. However, that takes more time and thoughtful planning in order to do. If there isn't a situation, where that time is absolutely necessary, I'd prefer to use other languages such as R solely for their convenience and efficient architecture that's already built in to the design of the language. For that reason, to me, one is not better than the other in an absolute way, it really depends on the situation I'm dealing with and what I plan to do with the language.

Describe the descriptive statistical measures mean, median, and range and how these values might be useful in data exploration prior to machine learning.

The descriptive statistical measures mean, median, and range can tell a lot about the data during exploratory data analysis phase prior to machine learning. The mean is the sum of the numerical data points divided by the number of data points there are. When sorted, the median is the value in the middle of a sequence of data. The range is a span between the minimal value and the maximal value in a sequence of data.

These three statistical measures can be used to explore more about what a data set has within it. For example, in a given numerical column, the mean represents the average observational value. From that we can see how far and near other examples sit relative to the average in a distribution. The median could be useful for that same purpose as well. We can see the counts of data relative to where they sit close or far from the middle value. The range can be useful as well for similar purposes.

Describe the covariance and correlation statistics, and what information they give about two attributes. How might this information be useful for machine learning?

Covariance describes a linear relationship between attributes based on the measures of how much they vary between the two. If the two attributes larger values correspond to one another, then their fewer values do too, and the covariance is positive. If the two attributes inversely correspond, where one attributes larger values correspond to the others smaller values, then the covariance is negative.

Once a linear relationship has been established by covariance, correlation measures the degree to which that relationship exists. The correlation coefficient is a ratio of the two attributes' covariance measure divided by the product of their standard deviations. The correlation coefficient represents a best-fit line through the dataset by projecting the expected values and seeing how close those come to the actual values.

So with covariance a linear relationship between two attributes can be established but the magnitude of this measure is hard to tell since it's not normalized. Correlation establishes to what degree is that relationship present in the data.

This information can be really useful for machine learning when trying to predict something. For example, in a regression problem, suppose there is a single target which is numerical. Covariance between an attribute and that target can express if there is a positive or negative relationship between the two. Correlation can then express whether or not the attribute is useful in predicting the target.