Michael Surdek

6-3 Journal: Data Analysis Report

Organizing data into structures such as vectors and data frames is a vital part of programming when it comes to the data analysis process. One reason this is important is because in programming languages such as R the types of functions and calculations you can perform on an object is determined by its data structure. For example, a function might call for a logical vector, and if you were to input a data frame it would cause an error in the function. It is important to organize your data in the proper structures and to identify them as such in your code. One technique that I think might be important, but cannot fully describe how or why, is building data structures step by step. In our AAPL and MSFT code, vectors were created with data and then those vectors were bind into a data frame. I do not have a ton of programming experience, but I can imagine that using that method is preferred to just building a data frame from scratch. I think that the technique of building data structures step by step would be helpful in identifying bugs and creating reproducible reports.

Comparing the results of data structures, such as subtracting the MSFT price from the AAPL price in the R code, helps in forming reportable results because it breaks down analysis into easily digestible information. In this example, we originally had the min max and mean of APPL and MSFT stock prices, as well as the min max and mean of the percent change from period to period. These calculated statistics are good information, but it does not mean much until a comparison is made between the two. The comparison itself is the reportable result that the reader would be interested in, so we created a new column that calculated the difference between the original two columns.

Data analysis for reporting is the process of calculating and presenting statistics that a certain party or parties might use to be able to draw conclusions or answer questions. Data analysis for reporting relies on the interaction between raw data and statistical programming for the purpose of communicating relevant information about that data.