Springboard Capstone Project #1 – Data Wrangling Write-up

Now that you have a basic ideas of the various data wrangling steps and techniques available, let's apply it to your Capstone Project! By now, you probably have a data set in mind for your project (If you don't have a data set yet, come back to this assignment once you have one). Apply some of the data wrangling techniques you have learned to this data set.

Submission: Create a short document (1-2 pages) in your github describing the data wrangling steps that you undertook to clean your capstone project data set. What kind of cleaning steps did you perform? How did you deal with missing values, if any? Were there outliers, and how did you decide to handle them? This document will eventually become part of your milestone report.

In my case, the first phase of my project was centered around several CSV files. Each file was very simple in that it included a date column and a single column of values. The challenge was finding an efficient method to consistently load in the files and aggregate into a single coherent data frame. When I was struggling to do this properly, it impaired my ability to move forward. However, once I got it sorted out, I was able to quickly take off on the analysis.

In the GC_BTC_Capstone1_EDA notebook, I found a tool called glob from another user on github in a similar situation of importing multiple similar files. This was instrumental in quickly loading in my data and was agile enough when I later added in additional data sources.

With the other notebooks, I had to be cognizant of the different formats of data because I was now using a variety of sources. It was a worthwhile exercise in utilizing the pd.read_csv function to resolve a gamut of issues (such as date-time formatting and order, skipping rows, etc.). Also resolving formatting issues during the import process was instrumental in avoiding unnecessary cleaning steps later down the line.

As for missing values, thankfully the time-series data is robust enough and my avenues of analysis were such that missing or NaN values were not an issue for me. If it was a few missing days, it would not detract from the broader trends. In my total return analysis, it would just be skipped. I will need to resample and interpolate my tulip data (since data is limited) for plotting purposes.

My data is also fraught with outliers. To reconcile this, I used resampling methods to smooth out data for plotting. Because of a number of extreme outliers, I resampled using the median of the period. In a plotting example, I resampled daily Bitcoin prices to a monthly basis and plotted the median price within that period. This helped “smooth” out the plot as to not detract from the bigger and broader trends.