Extract

The Six Musketweeters extracted daily Tesla stock (NASDAQ: TSLA) data from Kaggle and Yahoo Finance and combined that data with tweets from Elon Musk’s Twitter profile (@elonmusk).

Kaggle stock data was available in a CSV format, while Yahoo Finance and Twitter required custom BeautifulSoup/Splinter web scrapers. After extracting stock and social media data for the past 5 years (7-28-2015 – 7-28-2020), we were able to begin our data transformation.

Transform

After importing our stock and tweet data into Jupyter notebooks, we began analyzing and cleansing the data for eventual loading into PostgreSQL.

We began by loading the stock and tweet data into Pandas Dataframes and excluding extraneous columns from the data. For the purposes of the project, the we required only the Date, High, Low, Open and Close price from the Tesla stock data, together with the Date and Tweet fields from the Twitter data. We performed a str.contains(“esla”) to limit our dataset to Tweets specifically related to Tesla.

To calculate fluctuations in stock price, we had to strip out non-numeric characters (i.e. “$” and “,”). We then cast the High/Low/Open/Close columns as floats and created new columns to store the calculated values. Differences between open and close price were stored in Open\_Close, while differences in high and low price were stored in High\_Low.

In order to merge the stock and tweet dataframes, we standardized the date fields in each dataframe by casting them each as datetime64[ns].