Reporting: Wrangle Report

My data wrangling in this project was in three phases. I gathered, assessed, and cleaned the data I used.

I gathered the data from three different sources.

The first was a csv file containing the archived tweet information for the user's twitter account. Next, a programmatically downloaded tsv file containing image prediction information already processed by a neural network.

At last, I used the tweepy access library to query the twitter API using tweet IDs in the archive file.I generated my API and Access keys and tokens after creating my Twitter developer account. With my access, I extracted data from the API and converted the status for each tweet ID to JSON format before saving it in a text file. I subsequently went on to read the data from the txt file line by line and extract the desired data points. I made a data frame to organize and store all the information I gathered.

I had to check the data for inaccuracies, errors with completeness, validity, and consistency after I had gathered it.

In evaluating the various data frames, I used both visual and programmatic assessment. In the visual assessment, I opened the data frame with pandas and browsed through it for any anomalies or inconsistencies in the data. In order to have a better look at the data frame, I also opened it in Microsoft Excel.

In contrast, I used code to look for patterns in the data set and to investigate inconsistencies. I listed the characteristics that contributed to the data being dirty and untidy. Once the data had been evaluated and documented, it was time to clean it up and get it ready for analysis.

I made a copy of the original data frames before cleaning could start so I could keep the files before making changes.

During the Cleaning step, I replaced misleading data points, fixed data types issues, merged the data sets, and reorganized the table structure programmatically.

After cleaning, I combined the necessary columns from each of the cleaned tables to produce a master data set. I then created a CSV file and saved it in my working directory. For organization, I also saved the individual data frames.

Ultimately, I extracted some insightful information from the data and produced some visualizations to convey it.