Data Wrangling Efforts

In order for me to successfully get this project done, I followed the three steps that are required for the data wrangling. They are:

- Data Gathering;
- Data Assessment;
- Data Cleaning.

For the data gathering, I acquired the data from three (3) different sources and ended having files with different format; one in CSV, another one in TSV, and the last one in .TXT extension but was a JSON file. I eventually converted the JSON file into a csv so I can easily manipulate all the files together.

The gathering process required that one of the data needed to be downloaded using the tweeter api. Unfortunately, it could not be possible for me to do this and meet that requirement. But that's okay for now, we were given a chance to go on without using the api just in case there could be a problem (like my case, I was unable to have a tweeter developer account on time …). So, I had to just read the given api code then copy and paste it into my notebook just for meeting the requirements then took the tweet_json.txt file to use it in place.

After, the gathering, i dealt with the assessment which was just about looking at the raw data in different prospective by observing the data in visual form and also in programmatical form. To assess the data programmatically, I used a various of Pandas methods. I order to prepare for the next step, I had to document the issues in found from the data during this step. Some of the methods used during the assessment were (*df.head(), df.sample(20), df.info(), df.name.value_counts(), df.columns, df.column_name.nunique(), df.duplicated()*, …)

Then the last step of data wrangling was to clean the data based on the issues I found during the assessment. Cleaning is the hardest and longest step during the wrangling and it didn't deceive. I went through a lot throughout this process and it was said in the course, the cleaning process just list the other two processes is iterative. Most of the times, I went back and add more cleaning data during this process and even after the process, during the analtsis and visualization, I could go back and redo the cleaning after observing a mistake.

After doing all the steps above, I stored the cleaned data to a new csv using the **to_csv** pandas method and also using the **to_sql** method. Then, I eventually gave a little conclusion from the cleaned data.

On the conclusion is where the insights are found thanks to the analysis and simple visualization we had here.