Wrangle Data Report

For this wrangle data project, we have three preparation steps: gather, assess and clean the data.

For the gathering data part, we extracted three datasets from three data sources. One was read from the csv, one was pulled from the server, and one was pull through the Tweepy API. With the practice of Tweepy use, it gives more understanding and ideas of how to use those data in the future.

The most important part is to identify the quality and tidiness issues in those datasets. In total, the notebook documented 8 quality issues and 2 tidiness issues. From the very beginning, we always first check the missing values and duplicated value that we can remove accordingly. Then we checked the data type and correct them to proceed our further analysis. Some of the issues can be visually identified as unreal names or place holder names. Sometimes is impossible to correct all of them without too much effort, but in this project, this issue is just set up as an example for point to look at. Also, to ensure the connections between different datasets, we need to merge and join datasets based on common column to create the complete dataset for visualizations. Another good practice is that we found there are found different columns about the dog stages in the twitter achieve dataset, and what we did is to combine those four columns into one with appropriate values from each column.

For the last cleaning part, we tackled each issue one by one and using various skills like drop_duplicates(), capitalize(), replace(), lambda function, for loop datetime conversion and merge to achieve the cleaning purpose of this project. The final step is to create the master cleaned dataset in csv file for the next stage of data analysis and visualizations.