Wrangle Report 2019

For my data wrangling project, I needed to be able to gather, assess, clean and analyze data that would be sourced from three different locations and in three different formats. The primary dataset was provided as a CSV file and it was based around tweets that were posted on Twitter by WeRateDogs, an account that rates people's dogs based on the images that were provided by users and provides humorous comments about the dog. As this file was provided as a local file, I simply needed to load the csv file into my environment to be able to use it with the other sources. Another source was data that was created through a neural network created to classify what kind of dog was being shown in the submitted posts. This data was to be obtained programmatically from a backend server and the file itself would be in a TSV format. The third source of data I needed to gather was data directly from the Twitter API, where I would leverage the API to gather the tweets from the specific tweet_id in the WeRateDogs data and obtain the retweet and favorite counts. After being able to gather this data, I was then able to have the dataframes with copies of the data to be able to continue to the next step in my process.

While assessing my data, I was able to find a few issues visually and programmatically. Items such as renaming columns and seeing NULL values from my previewing my data led me to have visual assessments versus items that I found to have incorrect column types or column formats that I would find programmatically through pandas. Overall, I found 8 quality issues across the 3 sources of data. Finally, I had to tidy up my data so I could then analyze It. I noticed that there were several columns pertaining to dog stages that were more like their own variables so I combined the several columns into one new column to make it easier to analyze. To conclude my assessment and clean steps, I merged the three data sources into a single dataframe to make it easier to work with for the next steps.