Data Wrangling Report

The dataset that I worked with was a tweet archive of Twitter user @dog_rates, also known as WeRateDogs. This twitter account rates people's dogs with humorous and entertaining comment about the dog.

1. *Gathering Data*

For this project, I had three pieces of data to showcase my wrangling efforts. The first was the WeRateDogs Twitter Archive file provided by Udacity, the second was the tweet image predictions file downloaded by an URL also provided by Udacity, and third was each tweet's retweet count, favorite count, and tweet ID created by the querying of the Twitter API. With using Python's Tweepy library, I stored each tweet's entire set of JSON data in a file called tweet_json.txt file.

2. *Assessing Data*

After gathering the necessary pieces of data, I assessed all files visually and programmatically for quality and tidiness issues. I used pandas' info method for all tables to look for errors with datatypes and any missing data. With this, I noticed I had to remove retweets as I only wanted original tweets, change datatypes for certain columns, remove unnecessary columns, and I had to replace 'None' values for NaN for better read of the data. I also spotted inconsistencies in the image table as there was an '_' and need of capitalization with p1, p2, and p3 columns that provided the breeds of the dogs. This quality issue fixed would provide easier review of data for better reading and understanding. With tidiness issues, I saw that the dog stages should be combined into one to condense the data. I also wanted to merge all three dataframes with the common 'tweet_id' column to see all gathered, assessed, and cleaned data in one single dataframe.

3. *Cleaning Data*

I created copies for all tables to keep originals and to have clean copies. For each quality and tidiness issue, I performed the cleaning process programmatically with three steps: Define, Code, and Test. After cleaning, I went to reassess and iterate on any of the data wrangling steps if needed.

4. *Storing Data*

After the completion of all the steps into data wrangling (gather, assess, clean), I stored the DataFrame into twitter_archive_master.csv file.