Data Gathering

For this wrangling project, I worked with 3 datasets (Enhanced Twitter Archive, Additional Data via the Twitter API, and Image Predictions File) generated from Twitter user @dog_rates, also known as WeRateDogs. The goal of this project is to show my ability in gathering data, assessing data, cleaning data, storing data, analyzing data and visualizing data.

I started by importing the libraries needed for the project, then I imported the Enhanced Twitter Archive, which is a CSV file. It had 2356 entries and 17 columns. Then, I downloaded the tweet image prediction file hosted in Udacity servers using the Request library, and created a dataset, it had 2075 entries and 12 columns. I created a Twitter developer account, which was used in gathering data via Twitter API using Tweepy and created a dataset, it had 2327 entries and 3 columns.

Assessing Data

I did a quick analysis of the 3 datasets. I observed that the data had retweets, the numerator and denominator values had outliners, some image predictions were not dogs and there were also duplicate image URLs. I completed my assessment and came up with the following quality and tidiness issues:

1. There are retweets and replies in the dataset

2. Some rating denominators have outliers

3. The timestamp datatype should be datetime.

4. In twt_df (Enhanced Twitter Archive), the column name floofer should be "floof", which is a dog stage

5. Some dog names are wrong

6. There are predictions where the prediction images are not dog breeds

7. Some rows have 'None' values, which can be replaced with NaN, to indicate the missing values

8. The dog breed names capitalization is not consistent, the first letter should be capitalized.

9. In the twt_df dataset, the doggo, floofer, pupper and puppo columns table should be merged into one column.

10. All three datasets should be merged

Cleaning Data

First, I created a copy of the existing datasets before I started my cleaning process. I performed the programmatic data cleaning process in 3 stages - Define, Code and Test, for each of the stated quality and tidiness issues. After which, I merged the cleaned datasets using the "tweet_id" of each dataset, to create a master dataset. I had previously converted the "tweet_id" column name of the Twitter API dataset from "ID", while gathering my data.

Storing Data

After completing the cleaning process, I stored the master dataset in a .csv format and named it "twitter_archive_master.csv".