Weratedogs Data Wrangling Report

In this project, there will be 3 parts included gathering, assessing and cleaning tweets from WeRateDogs on twitter.

1. **Gathering Data**

There are three sources for this project. The first data is provided by Udacity. It was downloaded and imported into Jupyter notebook. The second file's name is image_predictions.tsv. It's hosted on Udacity's server and was programmatically downloaded by requests library. The file's URL is: **https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**. Tweet-json.txt is the last file was in text format and was written line by line into pandas Dataframe. This will take about 15 minutes to run.

2. **Assess Data**

Data was assessed visually and programmatically. Files was visually assessing through Excel, there are columns with no values was found. Also, columns that contain dog breeds was poorly keyed. After visually and programmatically assessed, there are 9 quality issues and 2 Tidiness issues was found.

Quality:

- Re-tweet will not be counted
- Timestamp column value is a string object
- rating_numerator column has values some very large ratings (1776, 960, 666, 420, 204, etc.)
- rating_denominator column has values other than 10
- Columns have empty values:retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
- In image_prediction: p1,p2,p3 are in bad format. Need to capitalize the first letter of each word and remove "_"
- rating_numerator columns shoud be float datatype
- tweet_id and id is integer where it should be a string
- rename id to tweet_id in tweet_into table to match other 2 tables

Tidiness:

- doggo, floofer, pupper, and puppo is 4 different columns where can be in one variable
- merge 3 tables into one on tweet_id

3. **Cleaning**

For each issue, 3 steps were taken: Define the issue, code for the issue and test the result.

Three tables were merged into one master file for assessing and visualizing.

Storing, Analyzing, and Visualizing Data

Final data was stored as twitter_archive_master.csv. It was also analyzing and visualizing.

Conclusion:

I have learned a lot in this project. Before assessing data, it can be confusing by just looking at them. The way data was organized, and other unnecessary information included made them hard for me to picture how merge together into 1 table. After assessing and cleaning, they became easier to read and understand.