WeRateDogs Twitter Archive Wrangling Report

This report incorporates the wrangling record of WeRateDogs Twitter dataset which I did in wrangle_act.ipynb notebook.

Gathering Data

Data was first gathered from the various sources provided for the analysis, we have three sources for the project:

1. Twitter archive of the WeRateDogs which was provided
2. The Image prediction dataset that was downloaded programmatically from the Udacity servers
3. Query of Twitter API - tweet_json.txt (Even though, I later used the provided tweet-json.txt).

twitter_archive _enhanced.csv

Using pandas library, I read the .csv files directly into a DataFrame

image_prediction.tsv

Using request library, the image prediction file was downloaded programmatically from Udacity servers after which the pandas library was used to read the .tsv into a DataFrame.

tweet_json.txt

Pandas library was used to read the json using the tweepy and json libraries

Accessing Data

The data was access to find tidiness and quality issues of all available data. All three datasets were assessed individually. Some tidiness and quality issues encountered include:

Twitter Archive Dataset

1. Tweets include retweets and replies that might not necessarily be needed for the analysis as it is not the original tweet.
2. Timestamp column is string.
3. The text column contains url attach to it.

4. The source column contain html link.
5. The rating denominator has 3 records that are less than 10 which is below the standard.
6. Columns doggo, floofer, pupper, puppo is not necessary instead it should contain into one category

Image Prediction Dataset

1. There are 324 predictions that are not dogs. As such, will not be needed for the analysis
2. The columns p1,p2,p3 can be given a more descriptive name
3. We need to clean Tweet Dataset

Tweets Dataset

There are some unwanted columns which are not needed for our analysis.

Cleaning Data

The dataset issues above were addressed and cleaned up as follows:

- Retweets and replies were removed because it was not needed for the analysis
- The html link was removed from the source column
- Urls were remove from the Text column in the master
- Timestamp in the master archive was changed from a string to datetime
- The columns doggo, floofer, pupper, puppo was contained in one category
- The records in the master archive that have rating denominator less than 10 (below standard) were removed.
- The columns p1,p2,p3 were merged into a column prediction
- Only columns tweet_id, favorite_count, retweet_count were needed from the tweets dataset for the analysis.