Wrangling and Analyze Data

● Dataset Description

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with humorous comments about the dog.

● Data Gathering

We will gather data from 3 different sources:-

1- CSV file: WeRateDogs Twitter archive data (twitter_archive_enhanced.csv).

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, It filtered for tweets with ratings only (there are 2356).

2- Requests library to download file: tweet image prediction (image_predictions.tsv).

A table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

3- Tweepy library to query additional data via the Twitter API: (tweet_json.txt).

Retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API. Well, "anyone" who has access to data for the 3000 most recent tweets, at least.

● Assessing Data

1- **Visual assessment:**
It is simple. It involves looking at the dataset in its entirety in a jupyter notebook.

2- **Programmatic assessment:**
Using functions and methods to reveal something about your data's quality and tidiness.

Quality issues

- **df_twitter_archive_enhanced**

1. There were 181 non-null values in [retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp] columns.
2. Unnecessary columns [in_reply_to_status_id, in_reply_to_user_id]
3. In the [name] column, it has invalid values.
4. [timestamp] is 'str' but should be 'datetime'.
5. There were dogs with more than one stage like (doggo and floofer, doggo and puppo, doggo and pupper).

- **df_image_predictions**

6. [jpg_url] has 66 duplicated values.
7. [p1], [p2], [p3] have inconsistently values as they have capital and small letters.

- **tweet_df**

8. [created_at] is 'str' but should be 'datetime'

- **General**

9. [tweet_id] is 'int' but should be 'str'.
10. The number of rows is different in the 3 tables.
11. Delete the [text] column from the merged dataset as it is found as [full_text] column in tweet_df.
12. Delete the [timestamp] column from the merged dataset as it is found as [created_at] column in the tweet_df.

Tidiness issues

1. The 3 tables should be in a one dataset, wil merge it.
2. The dog stage is one variable and hence should form single column. But this variable is spread across 4 columns - doggo, floofer, pupper, puppo.

● Cleaning Data

1- Merge the 3 tables in only one dataset using INNER join

2- Remove retweets (text column starts with RT @) and Delete unnecessary columns [retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id, in_reply_to_user_id, timestamp, text]

3- The dog stage is one variable and hence should form single column. But this variable is spread across 4 columns (doggo, floofer, pupper, puppo)

4- Convert the data type of [created_at] to 'datetime'

5- Convert the data type of [tweet_id] to 'str'

6- Capitalize and remove underscore in [p1], [p2], [p3] columns

7- In the [name] column, there were invalid values.

Save gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv".