Wrangling and Analyzing WeRateDogs Data

Data wrangling consists of three steps:

1. Gathering
2. Assessing
3. Cleaning

In our project we will go over these steps.

1. Gathering

We have data from three different resources, so we had to gather the data using different techniques.

The Data resources we had:

1. Enhanced Twitter Archive (.csv file)
    - It can be downloaded manually and imported using pd.read_csv().
    - The data is read into df_twitter_archive dataframe
2. Image Predictions File
    - It can be downloaded programmatically using requests liberary
    - The downloaded file is .tsv file.
    - The data is read into df_image_predections dataframe using pd.read_csv() taking into consideration that the separator is \t
3. Additional Data via the Twitter API (tweepy)
    - Using Tweepy to query Twitter's API for additional data.
    - This additional data will include retweet count and favorite count.
    - The data is read from a json file into a txt file then into a dataframe.

2. Assessing

- df_twitter_archive
    - **Quality issues**:
        1. Source format cannot be read easily.
        2. in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be integers instead of float.
        3. retweeted_status_timestamp, timestamp should be datetime instead of object.

4. Alot of missing Values (NaN) in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and expanded_urls.
5. name column has inaccurate data.
6. Some tweets are actually retweets and replies not original tweets that have to be deleted.
7. In the rating_denominator column any value below or above 10 is removed.
8. In the rating_nomenator column values most probably be more than 10.

- **Tidiness issues**:
  1. Dog stage is in 4 columns (doggo, floofer, pupper, puppo), no need for that.

- df_image_predections
  - **Quality issues**
    1. Undescriptive column headers
    2. Missing values from images dataset (2075 rows instead of 2356).
    3. Some tweets have two different tweet_ids, which is retweets or replies.

- df_dataapi
  - **Tidiness issues**
    1. The id column is the same as tweet_id column in the other two dataframes

3. cleaning

The steps I followed are:

1. Replacing the None values with empty string then NaN in columns (doggo, floofer, pupper, puppo).
2. Replaced the 4 columns (doggo, floofer, pupper, puppo) with one column dog_stage
3. Dropping useless rows and columns.
4. Removed the retweets and replies.
5. Edited the source column to be human readable.
6. Changed the data type of timestamp column to datetime.
7. Changed the data type of dog_stage to category.
8. Gave the columns more descriptive headers.

9. Dropped the tweets which have no image url.
10. Concatenated 1st_prediction, 2nd_prediction, and 3rd_prediction in one column and 1st pred conf, 2nd pred conf, and 3rd pred conf columns in one column
11. Dropped duplicated tweets.
12. Concatenated the data frames then removed the duplicated columns.

Storing

I saved the data after cleaning in one csv file (twitter_archive_master.csv)