Data Wrangle Report

Background

The dataset that wrangled in the project is the tweet archive of Twitter user *dog_rates*, also known as *WeRateDogs*. It is a Twitter account that rates people's dogs with a humorous comment about the dog.

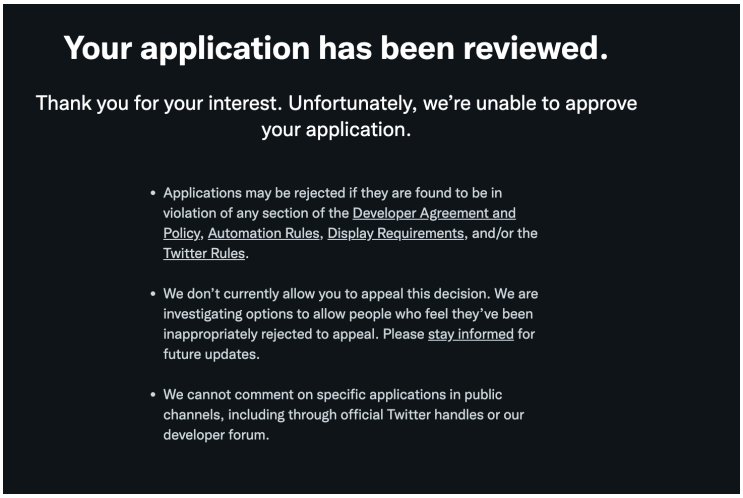
Gathering Data

There are three data tables for this project.

1. **Twitter Archive File**
   This dataset is provided by Udacity and can be download to my work directory directly.

2. **Tweet Image Prediction**
   I used the Request library in Python to download the image predictions data from *https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions* and saved the dataset in a file named *image_predictions.tsv.*

3. **Twitter API**
   The intention of this dataset is to use Tweepy library in Python to request Twitter API for additional likes, retweet counts etc. However, I could not create a Twitter Developer Account to generate user keys and secrets to use the API due to the region I am currently in. Instead, I read the Json file provided by Udacity directly.



Accessing Data

In this part of the project, I firstly evaluated three datasets both by visual assessment and programmatic assessment. And I listed the quality issues and tidiness issues I have discovered.

Quality Issues

1. tweet ids in all three tables are integer type rather than string
2. some dog names in *twitter_archive_df* table don't make sense, for example, names like 'a', 'an', 'the'
3. some rating denominators in *twitter_archive_df* table are not 10
4. rating numerators in *twitter_archive_df* table are wrongly recorded when they're with decimals
5. some dogs in *twitter_archive_df* table are recorded with more than one stages
6. there are retweets in *twitter_archive_df* table where *retweeted_status_id* is not null
7. 66 image URLs are duplicated in *image_predictions_df* table
8. the predicted dog breeds in *image_predictions_df* table are not consistent in terms of upper or lower case for the first letter

Tidiness Issues

1. doggo, floofer, pupper and puppo columns in *twitter_archive_df* table all belong to dog stages, that should be in one column
2. *tweets_selected_df* table should be merged into *twitter_archive_df* table
3. the first predicted breeds in *image_predictions_df* table should be merged into *twitter_archive_df* table

More Issues

There are additional problems in the dataset, which I decide to ignore for now and will assess and clean the data if further analysis is required. Below are some examples:

1. timestamp in *twitter_archive_df* table are string rather than datetime type
2. source column in *twitter_archive_df* table have unnecessary prefixes and suffixes
3. a twitter link and a text are in one column in *twitter_archive_df* table

Cleaning Data

For each issue addressed in the accessing data part, there is a define, code and text procedure in the cleaning data part to solve the issue. I followed this list to clean the data.

1. Change tweet id type from integer to string in all three tables.
2. Remove meaningless dog names in the name column in *twitter_archive_df* table and use null value if no proper name is extracted.
3. Set rating denominator and numerators to null value if the denominator is not 10; correct numerators when decimals are involved.
4. Combine dog stages into one column and keep only one stage for a dog.
5. Remove retweets and remove columns related to retweet ids and reply ids in *twitter_archive_df* table.
6. Remove duplicated image URLs.
7. Change all image prediction values to lower case.
8. Merge three tables into one master table and store the result.

After all the data wrangling processes, my dataset is clean and tidy. I stored the master data table into a new file, kept the original files as references. And the dataset is ready to be analyzed!