Wrangle Report

1. Gather

Data was gathered from different sources. Below are the datasets we used for this project.

- twitter-archive-enhanced.csv
- image-predictions.tsv
- tweet_json.txt

First of all, 'twitter-archive-enhanced.csv' was imported from the folder the project is saved in. Secondly, image-predictions was extracted programmatically from URL: "https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv". We named the dataframe 'img_predict'. Thirdly, we should extracted from Twitter API by using Python's library, tweepy. The dataset was saved as JASON file and UTF-8 encoding.

2. Assess

In this step, I tried to use different angles to observe these three data frames and to see if there is anything should be cleaned. First of all, I used sample() function to check the content of archive_enhanced randomly. Using info() to check if there is any missing values and data type for each columns. And I used the same way to check the rest of the two data frames. After checking the content, I made a list for data quality and data tidiness issues as below,

- Quality
  - archive_enhanced
    - in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls have null values
    - The numberator annd denominator columns have invalid values
    - in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and tweet_id shoud be string
    - timestamp column should be datetime type
    - Name column has invalid names such as 'None', 'a', 'an' and less than 3 characters.
  - img_predict
    - All columns have missing values (2075 rows)
    - jpg_url column has duplicate
  - tweet_data
    - tweet_id should not be integer but str
- Tidiness
  - Remove Unnamed column
  - Put "dogoo", "floofer", "pupper" and "puppo" columns into one column "dog_stage"
  - Join archive_enhanced, img_predict and tweet_data datasets

3. Cleaning

First of all, I copied these three data frames so then we can keep the original data frames

as our backup. And then I joined three data frames and named it df. Then I merged "dogoo", "floofer", "pupper" and "puppo" columns into one column "dog_stage". Then I removed extra columns ‘Unnamed: 0’, ‘retweets’, duplicated ‘tweet_id’ and tweets with no pictures. Then I changed the column timestamp to datetime type. After that, I put prediction algorithms and level of confidence into one column and removed original columns. And I clean the content of souse column and fixed numerators and denominators which are invalid values. Before save the dataframe to csv file, I used info() to check the data frame again to see if there is anything I miss or should be fixed. Finally, I saved the data frame to csv and named it ‘twitter_archive_master’.