Gather

Data were collected from three different sources.

The first data was “twitter-archive-enhanced.csv” given by Udacity. The csv file was imported into pandas data frame and named as “twitter_archive”.

The second data was “image_predictions.tsv”. It is programmatically extracted from a URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv by using python “request” library, and this file is imported as a data frame using tab as a separator. The name of the data frame is “image_predictions”.

The third data was extracted from Twitter API by using python tweepy library. It was saved as a JSON file and converted into pandas data frame named as “df_tweet”.

Assess

View “twitter_archive”, “image_predictions”, ”df_tweet” by using “head” “info”, “describe”, “value_counts” functions. There are several quality and tidiness issues.

1. Wrong data type: timestamp and retweeted_status_timestamp should be datetime instead of object.
2. Wrong data type: in_reply_to_status_id ,in_reply_to_user_id, retweeted_status_id, and retweeted_status_user_id should be string instead of float64.
3. Retweets: Some entries are retweets. We only want to keep the original.
4. Name: In name column, many entries do not look like names.
5. Rating: The numerator and denominator columns have strange values.
6. Remove columns which are not required for the analysis and store the dataframe to twitter_archive_master.csv
7. Inconsistent capitalization: Some of the first letter in p1/p2/p3 are capital.
8. Missing values: 2075 rows instead of 2356
9. Rename id column to tweet_id to be consistent with twitter_archive and image_predictions.
10. Join image_predictions and df_tweet to twitter_archive.
11. twitter_archive: one variable in four columns (doggo, floofer, pupper, puppo).

Clean

Start by copying the original data frames to clean data frames.

1. Rename id column to tweet_id to be consistent with twitter_archive and image_predictions.
2. Join image_predictions and df_tweet to twitter_archive.
3. Retweets: Some entries are Wrong data type: timestamp and retweeted_status_timestamp should be datetime instead of object.e retweets. We only want to keep the original.
4. Fix the wrong data type problems. Change the date type to datetime64 and int64.
5. Capitalize p1/p2/p3.
6. Melt 'doggo', 'floofer', 'pupper', 'puppo' four columns to one column 'dog_stage'.
7. Replace strange names with None.
8. Remove outliners of numerator and denominator.

9. Remove columns which are not required for the analysis and store the dataframe to twitter_archive_master.csv.