Project: Wrangle and Analyze data

This file contains details about Data wrangling, which consists of:

- Gathering data (Three different sources)
- Assessing data
- Cleaning data

1. Gathering Data
   - twitter-archive-enhanced.csv was given to us as a csv file to be directly downloaded.
   - image_predictions.tsv was to be downloaded from Udacity servers using **http requests** from a given url and saved in a .tsv file.
   - The third data was supposed to gathered from Twitter archive by accessing its API using Tweepy interface and saved it in a .txt file.
     For this I created my developer account and obtained the consumer and access keys and tokens.
     Then using **get_status** method I obtained the required json data and stored it ina .txt file using **json.dump()**

After storing these files into dataframes using pandas **read_csv()** and analyzing using .head() and .info(), I found the following issues:

Data Quality Issues in df_tweets_clean:

- Remove rows with "in reply to.." column values as they are not original tweets
- Drop the columns not required like source, expanded_urls, in_reply_to_status_id, retweeted_status_id, etc,.
- Data type of timestamp column
- Keep data only till August 2017
- Handle tweets that dont exist any longer
- rating_numerator and rating_denominator data type should be float
- The rating_numerator and denominator data have alot of issues, handle them

Data Quality Issues in df_image_clean :

- df_image_clean has unnecesasry columns, drop them
- keep only tweet_id, p1, p1_dog and p1_conf.
- Rename these columns to give more clear info about the contents of columns

Data Tidiness Issues in df_tweets_clean:

- Dog stages to be merged into single column
- Merge two tables on Tweet_id tweet_json_df and df_tweets
- There are multiple Dog stages for same tweet_id. Handle that.

- Handle tweet urls that contain more than one url --- since this column is not required, can ignore this issue
- The "text" column has tweet contents and urls, should be separated --- since this column is not required, can ignore this issue

After this assessment, I have cleaned the data.

And once cleaning process was completed, I have done basic analysis and visualization.