Wangle and Analyze Data

- Gathering Data

    1. Reading (**twitter-archive-enhanced.csv)** file **using pandas library.**

    2. Download and read the (**image-predictions.tsv**) file from the **url (**https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**)** file **using requests library.**

    3. Reading (**tweet-json.txt)** file **using tweepy library** and choosing the needed columns (tweet_id, retweet_count, favorite_count, creat_date, language, display_text_range, source).

- Assessing Data

    Viewed sample of each data set and checking the null values and duplications to understand the nature of every data set then come up with this points:

    - Quality:

        1. 'Name' column has some unrealistic values that need to be cleaned, "none" and "a".
        2. the datatype of timestamp column is object, it should be date.
        3. the datatype of created_at column is object, it should be date.
        4. the rating numerator is integers, it will be more accurate if it is floating and we have to show the double values that changed to decimal because of the data type.
        5. the datatype of tweet_id is integer, it should be object.
        6. some of the jpg urls are duplicated.
        7. the source of the tweet needs to be extracted from source column.
        8. the last 4 columns regarding the dog stage is not easy to analyze, gathering all of it into one new column would be better.

- Tidiness:

    1. Creating the new column for the dog stage because the last 4 columns regarding the dog stage is not easy to analyze, gathering all of it into one new column would be better.
    2. dropping unnecessary columns.
    3. merge the 3 data sets so we can work with one cleaned data set.

- Cleaning

Before starting the cleaning process, we have to create new file for every data set and do the cleaning phase of this files.

Three new files are:

- data_csv_new
- data_img_new
- tweets_data_new

cleaning steps was:

1. Removing records that have “NONE” or “a” as names.
2. Changing the data type of “timestamp” from object to date.
3. Changing the data type of “creat_date” from object to date.
4. Changing the data type of “rating_numerator” from integer to float.
5. Writing the exact flouting rate from the source csv file for the needed records.
6. Changing the data type of “tweet_id” from integer to object.
7. Remove one of the urls for the tweets that has more than 1 picture (I kept the first url).
8. Extracting the exact source from the “source” column which contain link.
9. Create new column for the dog stage.
10. Dropping unnecessary columns from each file.
11. Merging the 3 files into one to use it in the analysis and visualization, file name (all_date).