Data Wrangling Report

- Gathering Data

- 1. 'df_archive' :- the WeRateDogs Twitter archive whiche is provides by Udacity Course and i use pd.read_csv() to import this data into DataFrame

- 2. 'df_img' :- the tweet image prediction like (what breed of dog) . this file ('image_prediction.tsv') is hosted on Udacity's servers and i downloaded it programmatically using the requests library and provided URL.

- 3. 'df_tweet_json' :- i try to access twitter api and run my code , after run data can't be downloaded so i use the file 'twee_json copy' provides from Udacity Course and read data in it by json and create DataFram with data in the txt file.

- Assessing Data

- We Assess our data based on 2 reasons :

- 1. Quality Issue : Means contents issue like missing,duplicates or incorrect data .

- 2. Untidy data : Means it has a structual issue.

- Tidiness Issue :-

- Columns 'doggo', 'floofer', 'pupper', 'puppo' in df_archive should belong to one colomn -- stage.

- The df_tweet_json table need to merge into the df_archive table.

- Quality Issue :-

df_archive Table

1. Some columns have huge amount of missing values, for example, "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp". Since I don't need in_reply and retweet data in this project, I perfer to delete those columns directly.

2. The varaible "expanded_urls" also has few missing values, which means some records had no images. Any ratings without images should not be taking into account.

3. The datatype of "timestamp" is not correct.

4. change the long url links to certain words.

5. The standard for "rating_denominator" is 10, but it includes some other numbers.

6. The "rating_numerator" also has some incorrect values.

7. incorrect dog names (a, an, the, just, one, very, quite, not, actually, space, infuriating, all, officially, 0, old, life, unacceptable, my, incredibly, by, his, such).

8. The dog names are sometimes first letter capital but sometimes not. Keep the name format consistent

df_img Table

9. The columns'names are not clear and straightforward such as p1,p2.

10. The prediction dog breeds involve both uppercase and lowercase for the first letter.

in three tables

11. Tweet_id int64 , it should be str