Wrangling and Analysis Data

(Wrangling Report)

1- Gathering:

I gathered the data from three resources, the first file is twitter-archive-enhanced.csv which is csv file I downloaded it from Udacity project page, the second file is image-predictions.tsv I downloaded it programmatically from the internet using python requests library and basics of HTTP, then read it in Jupyter Notebook, the last file is tweet-json.txt, I downloaded it first time by using API application on twitter and tweetpy library it took a long time, after while I'm working on the project load file become failed then I had to use Udacity file..

2- Assessing:

First, to assessing data and extract the data issues either are quality or tidiness issues, I used many functions with all three dataframes like head(), isnull(), duplicated(), notnull(), sum(), value_counts(), sample() and info() function, info() function used to explore all information about dataframes such as number of observations and columns, missing data, and datatypes.

Data Quality issues:

1. tweet_id in archive_df is a float not a object.
2. Timestamp in archive_df is object not datetime.
3. retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id and in_reply_to_user_id in `archive_df` have wrong data type. but not necessary to convert them we need only the original data not retweeted so, will be dropped.
4. 59 missing values in expanded_urls variable in archive_df.
5. In archive_df many outliers values in rating_numerator.
6. In archive_df many observations have rating_denominator more or less than 10, because ratings almost always have a denominator of 10.
7. In archive_df name variable contain words that are not a names like: a, an, the, my, such, by, this, all, old, very. All the words begin with lower case.
8. In archive_df name, doggo, floofer, popper and puppo variables contain a lot of None values that express the missing values.
9. I replaced unclear text in source variable with more clear and short text.

Tidiness issues:

- doggo, floofer, popper and puppo variables merge in one column.
- Merge the three dataframes in one dataframe.

3- Cleaning:

Data Quality:

1. I have converted tweet_id in archive_df_clean and images_df_clean also id in tweet_df_clean to object datatype by using astype() function.
2. I have converted timestamp datatype to datetime.
3. I Dropped missing value in expanded_urls by using dropna() function.
4. I dropped all rows contains outliers in rating_denominator by calculating outliers online then store them in an array and reassign the dataframe with new values.
5. I dropped all rows that contain less or more than 10 in rating_denominator, because ratings almost always have a denominator of 10.
6. I dropped all retweeted tweets.
7. I have replaced wrong names with NaN value by using str.islower() function.
8. I have replaced the None values in the name variable to NaN by using np.nan function.
9. I have replaced unclear text in source variable with more clear and short text.

Tidiness:

1. I dropped in_reply_to_status_id and in_reply_to_user_id columns because they are not useful all observations the tweets contain image and rate.
2. I have merged doggo, floofer, popper and puppo variables in new column stage.
3. I have merged the three dataframes in a new dataframe by using inner join method.