Data Wrangle Report

Data Gathering:

I think data gathering was one of the hard steps for this project. In this project, I need to gather data from three different sources.

1. I would get data from the existing twtter-archive-enhanced.csv file by using pd.read_csv() and save the data as df table.

2. I would download the image_prediction.tsv programmatically by using request library and given URL and save the data as image_predict table.

3. I would gather data from twitter API by using tweepy library and store the JSON data as tweet_json.txt file. Then, we read the file and store data as tweet_df table.

Data Assessing:

By investigating all three tables and looking at their summary, I realized that we have below quality and tidiness issues that are needed to be fixed.

Quality Issues:

Df Table:

- Remove retweeted rows
- Change incorrect dog names ('a','an',and 'the') to 'None'
- Convert 'tweet_id' from int to str type
- Convert 'timestamp' from object to datetime

Image_predict Table:

- Capitalize the first letter of each word in p1,p2,p3 columns
- Remove the "_" between words in p1,p2,p3 columns
- Convert 'tweet_id' from int to str type

Tweet_df Table:

- Remove retweeted rows
- Rename "id" to "tweet_id' to match with other tables
- Convert 'tweet_id' from int to str type

Tidiness issue

- Combine 'doggo', 'floofer', 'pupper', and 'puppo' to one 'stage' column
- Merge df, image_predict and tweet_df tables with 'tweet_id' column
- Drop unuseful columns for each table

Data Cleaning:

1. For df table, I only selected rows with null retweeted_status_id value in order to remove retweeted rows. Then, I used replace to change incorrect dog names ('a', 'an', and 'the') to 'None'. For converting 'tweet_id' and 'timestamp' to str and datetime, I used astype(str) and pd.to_datetime. In addition, I combined values from 'doggo','floofer','pupper' and 'puppo' to column 'stage' by using replace and str.cat.

2. For image_predict table, I used str.replace to change '_' to ' ' in p1,p2,p3 columns. Also, I capitalized the first letter of each word by using str.title. At last, I converted tweet_id from int to str by using astype(str).

3. For tweet_df table, I renamed 'id' to 'tweet_id' and used astype(str) to convert it to string as well. Then, I removed retweeted rows by only selecting rows with null retweeted_status values.

Before doing data visualization, I dropped unuseful columns for each table and merge three tables into 'Final_df' table for data analysis and further investigation .