Wrangle Data Report

1- Data wrangling:

1- I read the three datasets (archive dataset, tweet dataset and images dataset).
2- I use JOSN file tweet_json.text from Udacity.

2- Assessing data:
Quality Issues and Tidiness for archive_df

Quality:

- 1- Null values recorded as None and NaN(missing values).
- 2- tweet_id type is int64.
- 3 - convert timestamp to be datetime and rename the column to tweet_date.
- 4 - source mixed html tag.
- 5 - some columns that contain unnecessary data we need to delete it.
- we have 179 tweets that are retweeted we need to drop it and drop the coulmns.

Tidiness:

- 1 - Extract doggo, floofer, pupper and puppo columns into new 'stages_of_dogs' column.

Wrangle Data Report

Quality Issues and **Tidiness for images_df**

Quality:

- 1 - tweet_id should be string.
- 2 - The types of dogs in columns p1, p2, and p3 had some uppercase \ lowercase letters

Tidiness:

- 1 extract breed of dog from columns p, p_conf and p_dog.

Quality Issues and **Tidiness for tweet_df Dataset**

Quality:

- 1 - there is a Missing values in some cloumns.
- 2 - id column should be named 'tweet_id' as the others dataset, and the change the data type to string.
- 3 - sourse data type should be category
- 4 - We have some columns that contain unnecessary data and we need to drop it.
- 5 - source mixed html tag ,Rewrite the tweet source, from iphone ,web...etc.

Wrangle Data Report

After I clean and merge the three datasets, I stored the (merge_df_clean) into csv file, twitter_archive_master_.csv.

Analyzing and Visualizing Data

Here some questions will be answered through the Act report about account:

Questions:

- 1 what is the number of Tweets per month?

- 2 What is the most used source?

- 3 what is the Top 3 frequent dog breed?

- 4 What is the most stage of doges?