WRANGLE REPORT

The following libraries were imported my jupyter notebook;

pandas

numpy

matplotlib

'twitter_archive_enhanced.csv' read into a pandas dataframe, as **df_01**, with the following are the columns;

'tweet_id' - User ID of Twitter account holder

'in_reply_to_status_id' - ID of reply tweet

'in_reply_to_user_id' - User ID of the Twitter subscriber replying to tweet

'timestamp' - Date and time of tweet upload

'source' - Url source of tweet

'text' - Written information on tweet

'retweeted_status_id' - ID of the retweet

'retweeted_status_user_id' - User ID of subscriber retweeting

'retweeted_status_timestamp' - Date and time of retweet

'expanded_urls' -

'rating_numerator' -

'rating_denominator' -

'name' - Given name of pet

'doggo' - A mature dog, in age and character

'floofer' - Furry dog

'pupper' - An immature dog, in age and character

'puppo' - Transitioning dog, between pupper and doggo

Requests library was imported to fetch data from the url;

"https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'pages"

and download as 'image-predictions.tsv'', and read into pandas dataframe as **df_02**.

The following are the columns;

'tweet_id' - User ID of Twitter subscriber
'jpg_url' - Image url
'img_num' - Image number.
'p1' - Algorithm's #1 prediction
'p1_conf' - %tage confidence of #1 prediction
'p1_dog' - Whether or not #1 is a dog breed
'p2'- #2 most likely prediction
'p2_conf' - %tage confidence of #2 prediction
'p2_dog' - Whether or not #2 is a dog breed
'p3' - #3 most likely prediction
'p3_conf' - %tage confidence of #3 prediction
'p3_dog' - Whether or not #3 is a dog breed

After repeated unsucceesful attempts to apply for developer status, I was unable to query the API using **Tweepy**. However, a dataframe was built for the data using the **'tweet_json.txt'** file provided in the Udacity workspace, as **df_03**. The columns are as follows;

'tweet_id' - ID of Twitter account holder.
'retweet_count' - A count of retweeting by other Twitter subscribers.
'favorite_count' - A count of likes by other Twitter subscribers.

Each dataframe was assessed, both manually and programmatically, and the following issues identified and systematically cleaned;

Quality issues:

DF_01 DataFrame

1. Replies and retweet rows should be removed.
2. Excess columns with less than 50 % No-Null cells and inconsistent data in df_01 dataframe.

3. Surplus data in multiple columns;

'doggo', 'floofer', 'pupper', 'puppo' in df_01 dataframe, and

'p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf'and 'p3_dog' in df_02 dataframe.

4. ‘timestamp' column is datetime, not object.
5. Missing values in 'name' column represented as 'None' and 'a'.
6. 'retweeted_status_timestamp' column is datetime, not object.
7. Missing values in the 'doggo', 'floofer', 'pupper', 'puppo' columns represented as 'none'.

DF_02 DataFrame

1. Data in 'tweet_id' column should be rearranged to match the order of df_03.
2. The varied capitalization of the values in some columns.
3. Misrepresented values in 'p1', 'p2', 'p3' columns.
4. Missing records (2075 instead of 2356).

DF_03 DataFrame

1. Missing records (2354 instaed of 2356).
2. Data type of 'tweet_id', 'retweet_count' and 'favorite_count' columns should be int64 (integer) not object.

Tidiness issues;

1. . The 'timestamp' column contains both date and time values in the df_01 dataframe.
2. 'tweet_id' column in df_01 dataframe is common both df_02 and df_03 dataframes.

Copies of the dataframes were first made for the data cleaning process, with **‘df_01’**, **‘df_02’** and **‘df_03’** saved as **‘df_01_clean’**, **‘df_02_clean’** and **‘df_03_clean’** respectively.
Each dataframe was cleaned as much as possible, with cotinuous reiteration done during the process, to produce the final document, ‘merged_df_2’,with the following columns(prior to being saved as as 'twitter_archive_master.csv');

'tweet_id',

'name',

'dog_attribute' - Denotes either doggo, pupper, fluffer or puppo

'retweet_count',

'favorite_count',

'rating_numerator',

'rating_denominator',

'jpg_url',

'img_num',

'breed_prediction' - Highest voted prediction by algorithm

'confidence_%' - %tage confidence in image by algorithm

'is_dog' - Denotes True/False, whether prediction is dog breed

'source',

'date'

'time'

'text'

'expanded_urls'.