Wrangling Report

This project involved the wrangling (analyzing and visualizing) of the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. Firstly, the twitter achieve was imported into the jupyter notebook as the first dataframe using the pd.read_csv function. After which the image predictions file was programmatically downloaded into the jupyter notebook as the second dataframe. This was done using the requests library. Also, the Twitter API was queried to give additional data for the analysis, this provided the third dataframe for the analysis, providing data like retweet count and favorite count.

Secondly, the dataframes were visually and programmatically assessed to identify quality and tidiness issues. And the following issues were identified:

1. The name column of the twitter archive dataframe had some values that were not names e.g 'None', 'such', etc. The timestamp column had the datatype of a string

2. There were tweet_ids with values in retweeted_status_id, retweeted_status_user_id, and retweeted_status _timestamp. We were earlier informed that we will not be using retweets for our analysis

3. There were lots of nulls in certain columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls)

4. In twitter_arch dataframe, there were some rating denominators that were greater than 10 when 10 is supposed to be the max denominator

5. In the image dataframe, the dog types listed in the p1,p2 and p3 columns are not consistent as some start with capital letters while others are lower case

6. In the image dataframe, there were some other images there that were not dog pictures

7. In the three dataframes, the tweet_id column should have a datatype of string not int

8. In the twitter_arch dataframe, the dog status should be in one column instead of four

9. The 3 dataframes should be merged into one for analysis

Thirdly, to address the above, the following actions were taken:

- A copy of the 3 dataframes were made before cleaning commenced as was named as df1, df2 and df3
- Values that were not names were removed and replace with nulls using np.nan
- The timestamp column's datatype was change to string using the astype() function
- Tweet Ids with values were eliminated by applying the isnull() function to the dataframe to select only rows with null values in the retweeted_status_id, retweeted_status_user_id, and retweeted_status _timestamp columns
- The drop() method was used to drop columns like in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and expanded_urls as they had too many nulls and will not be useful for the analysis
- The values in the rating_denominator column was set to 10 using the loc function while str.capitalize() function was used to address inconsistencies in names in the df2 dataframe
- To remove images that were not dogs, the dataframe was split into 3 dataframes, after which rows with False values in p1_dog, p2_dog and p3_dog were dropped. These dataframes were merged into one and named dog_df
- The datatype of the tweet id columns was changed to string using the astype() method
- The dog _status column was created by firstly converting Nans in the puppo, doggo, floofer and pupper columns to an empty space. After which all these columns were merged into the doggo column which was renamed dog_status. The other columns were then dropped from the dataframe.
- After all the issues in the above were treated, the three dataframes df1, dog_df and df3 were merged using the tweet_id column. The resulting dtaframe was called data.

Lastly, data was converted to a CSV file named "twitter_archive_master.csv" and stored for further use.