WRANGLE REPORT

The entire data wrangling process was carried out in three steps which are:

- ❖ Data Gathering
- ❖ Data Assessing
- ❖ Data Cleaning

Data Gathering

The data used were gathered from three different sources.

The first data was the twitter_archive_enhanced.csv which was manually downloaded from the Udacity classroom.

The second data was the image_predictions.tsv which was downloaded programmatically from the Udacity classroom using the request library.

For the third file, I couldn't set up my twitter developer account so I had to download the tweet-json.txt file using the link provided in the udacity classroom, opened the file, read it line by line and extracted only the columns needed.

All three files were read into pandas dataframe separately as archive, image_pred, and tweet_df respectively.

Data Assessment

I did both virtual and programmatic assessment of the three datasets separately and identified about nine quality issues and two tidiness issues.

Quality Issues:

- ❖ Missing values in some columns related to retweet in the archive dataframe (Eg the retweeted_status_user_id).
- ❖ Timestamp is a string instead of datetime.
- ❖ Some names are invalid (such as a, an, not and none) in archive dataframe.
- ❖ A dog name recorded as 'O' instead of O'Malley as seen in the text.

- Tweet id is an int64 instead of dtype object in archive and image prediction dataframe.
- Columns with very few data such as in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp, should be deleted.
- Missing values in the expanded_urls column.
- Inconsistency in capitalization of dog breed in p1, p2 and p3 columns of image_pred dataframe.

Tidiness Issues:

- The different dog stages should be in one column instead of separate columns.
- The three datasets should be merged into one since they are all related.

Data Cleaning

Before I began cleaning the data, I made copies of the three datasets. I started with the quality issues before the tidiness issues, using the define, code and test pattern. The steps I used are given below.

- The missing values in the retweeted_status_user_id column were extracted by dropping the notnull values of the variable to give us only the original tweets.
- Timestamp was converted from object to datetime dtype using the pandas to_datetime method.
- The invalid names were replaced with NaN using the replace() function.
- The dog's name that was wrongly entered as 'O' was replaced with the original name O'Malley using the replace() function.
- The tweet id for both the archive and image prediction datasets were converted to string using the astype() function.
- The columns with high number of missing values, minimal data and irrelevance to the analysis were dropped. These columns were also related to retweets which were not needed for analysis.

- The missing values in the expanded_urls column were dropped using dropna() function.
- The names of the dog breed were converted to lower case using pandas islower() method.
- The different columns of dog stage (doggo, floofer, pupper and puppo columns) were concatenated into a single column called dog_stage and the individual columns were dropped.
- The archive_clean, the image_pred_clean and the tweet_df_clean dataframes were merged into one dataframe called twitter_master using the merge() function.
- After merging, I checked for duplicates and missing values and noticed there were several missing values. I discovered that majority of them were associated with the image_pred dataframe and they had the same number of missing values. This became a quality issue which I had to address by dropping the missing values.

After cleaning, I saved the data to a csv file named twitter_archive_master.csv