Wrangle Report

Gathering

Three datasets was needed for the analysis

> The first dataset **(twitter_archive_enhanced.csv)** was already provided. The second dataset **(image_predictions.tsv)** was extracted from a url using **requests** library and stored using the **os** library. The third dataset was extracted from twitter using the **tweepy** API library with the help of twitter's consumer tokens and access tokens. The twitter information gotten was written into text file **(tweet_json.txt)** the and the reqiured information for the analysis **(tweet id, retweet count, favorite count)** were extracted from the json file directly. This information was appended to a list and converted to a dataframe which then saved as a csv file using pandas **to_csv()** as **json_tweets.csv**

> All three datasets were read into the notebook using pandas **read_csv()**.

Assessing

Visual assessment:

> Copies of each dataset were made and then they were assessed visually by printing first few rows using the pandas **head()** and **sample()** method.

Quality and Tidiness isuues discovered:

- Column in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id retweeted_status_user_id and retweeted_status_timestamp contains null values
- Column headers doggo, fluffer, pupper and puppo are values not headers
- incorrect dog like *a* and **None**
- p1, p2, p3 in the image prediction dataset having underscore.
- presence of duplicate rows
- missing records due to missing tweet ids

programmatic Assessment

> this was done using the **decribe(), duplicated(), info()** and also searching for specific info about the datasets using the comparison operators

Quality and Tidiness issues discovered

- erroneous data types
- outliers in the rating_numerator e.g minimum value of 0 and maximum of 1776. There are other apart from these.
- rating denominator value of *None* and less than 10
- only one table should exist
- only true prediction with their confidence level are needed

Cleaning

> missing records due to missing tweet ids: the datasets (twittwer_enhanced_df and json_tweet_df) were merged together on the tweet_id using the merge() method to create uniformity

> Column headers doggo, fluffer, pupper and puppo are values not headers: a function was written to extract the count of each dog stage which was appended to a list. The function was applied to the twitter_archive _df (contains the merged data) and a column for the dog_stage was created

> erroneous data types: All wrong data types (timestamp, dog_stage, tweet_id) were converted to the right ones using astype()

> outliers in the rating_numerator e.g minimum value of 0 and maximum of 1776. There are other apart from these: Rows with a rating numerator of 0 in the twitter_enhanced_df were dropped using drop(). Higher values of rating numerator are vaid so they are left as they are.

> rating denominator value of None and less than 10: the values were replaced by 10 using the replace()

> only one table should exist: the image_predictions_df was merged with twitter_enhanced_df on the tweet_id

> only true prediction with their confidence level are needed: a function was written to extract only the true predictions with their respective confidence levels to form new columns called prediction and confidence in the twitter_archive_df

> p1, p2, p3 in the image prediction dataset having underscore: the underscore was replaced with empty string using str.replace()

> incorrect dog like *a* and None: Dogs names with None were left as they are but incorrect dog names like *a, quite etc**. str.contains() together with a regex pattern was used to print them all out before using str.replace() with the same regex pattern to replace those names with **None**

> presence of duplicate rows: the rows with retweet information were isolated by subsetting the twitter_archive_df for row with non_empty values for the specific column of interest. The index for these isolated datafr ame was extracted using index attribute and stored as a list using list(). a for loop was used to iterate through the list and index in the list present in the twitter_archive_df was dropped