REPORT: WRANGLE REPORT

To begin wrangling my data, relevant packages including, pandas, numpy as np, requests,os and json were imported. To gather the relevant data to rate dogs, the twitter-archive-enhanced csv file, using pd.read_csv was downloaded and labelled rate_dogs. The tweet image prediction was downloaded using the Requests library(image_predictions.tsv) and labelled image_prediction. The dataset provided for twitter API was used and the json file read. I set the option using pandas to view all of the columns in rate_dogs_json.

To assess the data, eight quality issues and two issues involving tidiness were found. To execute this, the .info() function was used . value_counts function on various columns on all three datasets to invesstigate further into them .duplicated().sum() and .isnull().sum() functions were used to find duplicated and null values respectively.

To clean the data,

These eight issues below were identified as the quality issues after the assessment:

Missing values in rate_dogs columns. in_reply_to_status_id,in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp,expanded_urls, Missing values in ratedogs_json columns. extended_entities,in_reply_to_status_id,in_reply_to_status_id_str,in_reply_to_user_id,in_reply_to_user_id_str, geo, coordinate, contributors, Change tweet_id datatype from int to object, doggo,fluffer,puppo and pupper columns have "None" in them, Change image prediction column names to make them comprehensible, Change column name on ratedogs_json from id to tweet_id, Retweets have non-null rows in rate_dogs dataset and Retweets have non-null rows in ratedogs_json dataset.

These two issues below were identified as tidiness issues:

doggo, fluffer, puppo and pupper columns should be melted into a dog stage column and There are irrelevant columns in rate_dogs and ratedogs_json dataframes.

To tackle these issues, a copy of each dataframe was made. Naming rate_dogs as rate_dogs_clean, ratedogs_json as ratedogs_json_clean and image_predictions as image_predictions_clean. Retweets were not useful to this dataset. Therefore, any non-null rows regarding retweets in rate_dogs_clean and ratedogs_json_clean datasets were dropped. After this, columns with NaN values that were deemed to be irrelevant to the analysis,i.e, columns regarding retweets and location of the user were dropped. Then, tweet_id in rate_dogs_clean and image_predictions_clean and id column in ratedogs_json_clean, a column common to all three datasets was changed from int to string in order to be able to merge the datasets, if needed. The columns of the image_predictions_clean are changed to make it more comrehensive. The id column in ratedogs_json_clean was changed to tweet_id to make it more comprehensive and to match the columns in the other two datasets in the event of a merge. Afterwards, the doggo,puppo,pupper and floofer columns were added into one column named dog_stages, after the "None" string for dogs with no stages were replaced with empty strings. The doggo,puppo,pupper and floofer columns were then dropped. Some dogs had two stages and these names were formatted together. They were separated with a ","

Finally, a master dataset was created, merging the three datasets, rate_dogs_clean, ratedogs_json_clean and image_predictions_clean into one dataset, on tweet_id with a left join. This master dataset is named twitter_master_archive. The empty dataset in the dog_stages column are then replaced with a string, "No Stage". The dataset is then stored to a csv file using to_csv.

The master dataset twitter_master_archive is further assessed. This is done using the describe function, the isna function and the info function. After this assessment there seemed to be a gap created by the slightly smaller image_predictions_clean dataset. The columns with NaN that were strings in the former image_predictions_clean dataset were changed into "." Whereas the columns that were floats in the former image_predictions_clean were maintained.

After assessing the twitter_master_archive, the count of true in the p1_dog column, which shows the dog from the image prediction, is higher than the count of false. Ratings also seem to be highly suggestive and do not seem to give further insight into dogs. Some dogs did not have stages and some dogs had two. Pupper seems to be the highest dog stage that has only one stage per dog.