WRANGLE REPORT

The files worked on in the project involve the following which were worked on separately before they were merged into a single file and ready to be used for analysis to generate some insights on the date.

1. The WeRateDogs Twitter archive: This contains an archive of different types of dogs posted by different people that requested their dogs to be rated.
2. The tweet image prediction: This file is present in each tweet according to a neural network which gives a prediction on how related an image relates to a specific dog breed.
3. Data from Twitter API: A collection of each tweets' retweet count and favorite count from the WeRateDogs posts.

The twitter archive file, image prediction file was read into a pandas dataframe as CSV and "sep" parameter was used where necessary. Information about the loaded dataframes were observed to get insight on the cleaning points and tidiness needed to be done. A list was created ''tweet counts'' to create a dataframe including the three columns we needed and append the data for the analysis.

On loading all files into a dataframe, the proper wrangle and data assessing was done using the following steps:

1. Checking the shape of the dataframes for proper insight on what needs to be done on the tidiness aspect
2. It is obvious that in data cleaning, we need to drop some rows. As such, a function was created to drop rows to avoid repetition.
3. We do not need the replies in the twitter archive which we decided to create an index dataframe to be dropped. Likewise, the retweet is not required as we only require tweets by the main account of WeRateDogs.
4. In Line 21, all columns were dropped using the drop column function earlier created for easy accessibility.
5. The code in line 24 is used to clean the tweet count data to filter out any tweet that contains the ancho tag to determine the source of a tweet
6. The first merge involves joining the image prediction dataframe to the twitter archive dataframe to consolidate needed columns that are present in both dataframes
7. Joining the first merge with the tweet counts dataframe and specifying the start point as tweet_id and also the type of joining to be 'inner' since the two dataframes have things in common. Only those items they have in common are returned in the new dataset
8. In line 38, the columns of tweet counts needed to be changed to correspond with the other dataframe before the merging. On completing the wrangle process, the merged dataset was saved as twitter_archive_master.csv

The quality issues noted were

1. tweet Id and tweets Id duplicated, we drop duplicate values
2. Datetime object was in a wrong format and was changed to the correct format.