Summary of The Wrangling Efforts for Data Wrangling Project

The data wrangling efforts involved steps as follows:-

1) Gathering Data
2) Assessing Data
3) Cleaning Data
4) Exploratory Analysis

1) Gathering data

The project contains data from three different data sources. The first was from the "twitter-archive-enhanced.csv"; second from the "image-predictions.tsv" and the third from the "tweet-json.txt" collected via API. For the purpose of this analysis, I did not collect the txt file from Twitter's API. I had to use the one that was already available for our analysis.

2) Assessing data

All three data were assessed one at a time; both visually and programmatically. I also relied on the spreadsheet to visually assess the data specially for csv and tsv file since some of the texts were not easily readable due to the length of the strings. I was able to identify several quality issues visually. Some of them are

df_csv

- Quality: The url links appear duplicated in the column "expanded_urls" and few of the urls contain link to other than twitter.
- Quality: Replace rows with names as "a" with "None"
- Quality: Replacing the values "None" in the columns "doggo", "floofer", "pupper" and "puppo" to "".
- Tidiness: Convert "doggo", "floofer", "pupper" and "puppo" into a single column.
- Quality: Deleting unwanted columns such as "in_reply_to_status_id ", "in_reply_to_user_id", "retweeted_status_user_id", "retweeted_status_timestamp"

df_tweets

- Tidiness: Dataframes df_csv and df_tweets will be merged since they have similar number of records.

Programmatic assessment included identifying any quality issues with the use of functions like "head()", "tail()", "sample()", "info", "describe()" and "value_counts". Some of the issues identified via the programmatic assessment are as follows:

df_csv

- Quality: Datatype for "timestamp" and "retweeted_status_timestamp" is an object instead of datetime.
- Quality: Removing all the rows that are retweets and replies to the tweets.
- Quality: Extracting correct numerator and denominator as some have been extracted incorrectly.
- Tidiness: Not visibie in the notebook but visually assessing the data in the spreadsheet shows that the "text" column has the url link to the tweet.

df_tweets
- Quality: All the records with the "retweet_count" have "retweeted" as False. If the records have retweeted_count, they will be replace with True.

3) **Cleaning the data**
Cleaning the data involved correcting the quality and tidiness issues identified from my preliminary assessment. I had to heavily rely on the Knowledge Forum especially for extracting the correct numerators and denominators from the text column in the csv files. I also used regex patterns heavily to extract the required substring from the columns especially for the issue on url links duplication.

4) **Exploratory Analysis**
Even after having cleaned the data, there were still some quality issues. While conducting exploratory analysis, I discovered that the rating_numerators were numbers other than 10. This was tackled after identification.