Wrangling efforts on the three datasets that I worked on

First, the necessary libraries such as pandas, numpy, tweepy, json, requests, and visualization libraries (seaborn and matplotlib) were imported into the notebook. The steps followed for the data wrangling are listed below:

1. Data Gathering

In this data wrangling project, three datasets were gathered from different sources namely:

1. A twitter archive dataset was supplied and provided as a download. The dataset which is a csv file was loaded into the notebook and saved into a dataframe called 'twitter_archive_df '.
2. An image prediction dataset was downloaded from a given url using the request library. The contents were written to file and then read into a dataframe called 'image_predictions', using the read_csv function.
3. Lastly, additional data was gathered from a tweet_json.txt file that was also supplied. Due to Twitter API access issues, I used the txt file supplied and read the data into a dataframe called 'tweets', using the read_json function.

2. Assessing the data

The data from the three sources were assessed for quality (at least 8 of them) and tidiness issues (at least 2 of them). To notice the issues, visual assessing as well as programmatic assessment (using pandas' functions) were carried out. Some of the issues that were noticed include:

A. Quality Issues

1. Column "timestamp" in dataframe "twitter_archive_df" is of type string rather than datetime. Column "retweeted_status_timestamp" in dataframe "twitter_archive_df" is of type string rather than datetime

2. There are rows that are retweets instead of regular tweets

3. Some rows are replies rather than regular tweets

4. Columns "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", and "retweeted_status_timestamp" in the "twitter_archive_df" dataframe have null values.

5. Inaccurate value of dog names such as "a", "an", "not" in the column "names" of the "twitter_archive_df" dataframe

6. Null values for the dog names are set as 'None' instead of numpy 'NaN's.

7. Some of the predictions in the "image_predictions" dataframe could not establish the breed of the dog.

8. Several rows of the "expanded_urls" column in the "twitter_archive_df" dataframe contain the same tweet url multiple times. Also, some of the rows contain urls that are not related to "WeRateDogs".

B. Tidyness Issues

1. The columns "doggo", "floofer", "pupper", and "puppo" in dataframe "twitter_archive_df" should all be in one column because they are different dog stages.

2. All the three dataframes should be combined into one dataset.

3. Cleaning the data

In this section, the issues highlighted above were dealt with. First, I made copies of the three datasets before making any changes. I started with the tidiness issues. A new column was created to accommodate the values of the four columns that were unnecessary. The columns were then deleted. The three datasets were combined into one dataframe using the merge function from pandas. The dataframes were joined on the 'tweet_id' column and an inner join was performed.
For the quality issues, the inappropriate data types for the affected columns were changed to appropriate data types. Rows that are retweets and replies were removed, columns that were not useful for the data analysis were removed, and inaccurate variable values (e.g. dog names) were removed. All of these were done using appropriate pandas functions.

4. Storing the data

The cleaned combined dataset was saved to a csv file

5. Analyzing and visualizing the data

I asked some questions for which the answers can give some insights. Pandas functions, as well as visualization methods from seaborn and matplotlib were used to answer the questions and provide the insights.