All the necessary libraries required for the data wrangling process was loaded.

GATHERING DATA

Three datasets were gathered using 3 different gathering methods.

The first dataset, WeRateDogs Twitter archive was gathered manually using the link provided, then loaded.

The second dataset, **tweet_image_predictions** was gathered programmatically using the **Requests** library and then loaded for wrangling.

The third dataset was gotten from **Twitter API,** but due to trouble getting approval from twitter to access the data, the provided tweet_json.txt file was used. The json file was loaded into a Panda DataFrame including the required keys.

ASSESSING DATA

The twitter_archive dataset was assessed both virtually and programmatically. The max rating was checked, but due to the fact that ratings over 10 were allowed, nothing would be done about this. The .info() method was used to check for enormous datatype and checked for the number of missing values and dimension. Due to the fact that what is needed is the original ratings (no retweets) that have images, the dataset was accessed for the entries that are retweets by using the .notnull() method on the **retweeted_status_id** column. Furthermore, it was accessed for the percentage of missing values for each column. It was also checked for duplicated tweet_id. The name column was checked to see invalid dog names, using both virtual and programmatical accessing method.

The image_pred dataset was also accessed using the .info() method to check for enormous datatypes, number of missing values and dimension. Duplicated tweet_id was also checked for. The prediction on the dog breeds was also checked for different variants (i.e., if we have some with Proper names and some in lowercase). It was checked for entries with no true predictions for all predictions.

The tweet_json was accessed using the .info() method. Duplicated tweet_id was also checked for.

CLEANING DATA

Copies of each dataset was made.

Based on the Quality and Tidiness Issues discovered, the following list the issues and how it was cleaned.

Quality Issues

Issue1: Some of the tweet recorded in the tweet_archive is retweet was cleaned by dropping entries that were retweet.

Issue2: Some Columns in twitter_archive contains missing values greater than 50% was cleaned by dropping columns with missing values greater than 50% of the total entries.

Issue3: Some of the text has the dog stage in it, but absent in the appropriate columns was dealt with by extracting as many as could be extracted from the **text** column. It was tested by comparing an entry that such issue with the before and after.

Issue4: Invalid and incomplete Dog Names in twitter_archive df was resolved by changing incomplete names to the full names and invalid names to None.

Issue5: Enormous in twitter_clean_df was cleaned by converting inappropriate column datatypes to the right ones.

Issue6: Different variants of dog breed in P2 in image_pred was cleaned by changing all entries in p2 to lowercase.

Issue7: Invalid prediction on Dog Breed in p1, p2 and p3 in Image_pred df was resolved by replacing all entries with False in all predictions with not_dog.

Issue8: No True predictions (p1, p2, p3) for some of the tweet_id in image_pred was fixed by dropping all tweet entries with all False prediction for all predictions made.

Tidiness Issues

Issue1: **One column represented as 3 columns in twitter_archive df** was resolved by using the pd.melt() function to combine the 3 columns into one and groupby() to combine entries with more than one dog.

Issue2: one dataframe represented as 2 was cleaned by merging the 2 datasets.

STORING DATA

All datasets were merged together based on the tweet_id and saved into a master dataset called "twitter_archive_master.csv"