DATA WRANGLING REPORT

Gathering Data Phase

The initial stage of this project was to gather data from three different sources and use supporting methods to extract the data.

The first step was to download the `twitter_archive_enhanced.csv` file. Then to upload and read it into my working Jupiter notebook using Pandas.

Second was to get my images using the image prediction tsv file provided (`image_predictions.tsv`) and read it using a Request library.

Last step of the Gathering data Phase was the Twitter API which I had to first request for an API access to be able to retrieve the data for my project. Once access was granted, using the Tweepy library I was able to load and run the required information I needed for my project.

Assessing Data Phase and Cleaning Data Phase

In this assessing stage all three previously gathered data was assessed both visually and programmatically for quality and tidiness issues. After all issues had been identified the cleaning phase had to be affected to make the data set tidy.

Issues that were identified and cleaned are:

Issues	Solution
In_reply_to_status_id and in_reply_to_user_id have a lot of NaN values.	The columns of in_reply_to_status_id and in_reply_to_user_id would be dropped from data set

retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp columns also have NaN values, they are also not needed in the final analysis.	Drop retweets from dataset since it is not required for further analysis
Time stamp wrongfully labeled as object.	Change Timestamp from object to datetime using Pandas to-datetime function
Dog category have None as value instead of NaN.	Remove None values and replace with NaN
Dog names should all start with capital and replace other names with actual dog names from text column.	Some dog names do not have actual names of the dogs, so would extract names from text column to fill in those spots. For those labelled as None, would replace with Nan.
Text column name should read as tweet not text	Contents within the text column after visual analysis shows that the column contains Tweets and hence should be labelled as Tweet not Text
Expanded urls has some missing data.	some of the URL links have missing data. Would drop the whole column to make analysis easier.
P1, P2 and P3 have both lower case and upper-case letters at the beginning of the sentence word	Capitalizing each word in the p1, p2 and p3 columns within the Cleaned_Image_Pre dataset.
Dog breeds doggo, floofer, pupper and puppo should be grouped into one column.	Merge all three stages so that data is in unison
Timestamp should be grouped into two (2) columns; Date and Time.	Separate date and time into two columns for better analysis.

Storing Data Phase

After the Gathering, Assessing and Cleaning Data Phases have been completed, all data was stored in a master dataset called the twitter_archive_master.csv.

Three insights were generated from the processes above which are:

1. To find out which name is given the most to dogs.
2. Which dog stage is most documented.
3. Which dog breed is the most popular with people.