STEP 1: DATA GATHERING

First, I imported all the required libraries that I needed for this project. I gathered data from three different sources for this analysis.

- I downloaded twitter_archive_enhanced.csv and read it into a pandas data frame I called twitter.
- I used the request library to download image-prediction.tsv file using the request library and programmatically loaded it into a data frame I called image.
- I read tweet_json.txt file line by line into a data frame I called twitter_extra as I was not given access to query the twitter API directly.

STEP 2: ACCESSING DATA

In this stage, I viewed the data frames visually and programmatically while taking note of quality and tidiness issues affecting the data frames. These are the Quality issues I identified

1. The data type for the timestamp column is object when it should be a datetime.

2. In several columns null objects are represented as 'None' instead of NaN

3. Dog Name column have invalid names i.e 'None', 'quite', 'such', 'the 'a', 'an' etc

4. these columns type (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and tweet_id) are floats instead of strings

5. Some rows have several identical values in the expanded_url column

6. Some tweet_ids have the same jpg_url in the image prediction data

7. Tweet_id fields in the three datasets are stored as numeric values and should be strings.

8. Create_date is object instead of datetime.

Below are the Tidiness Issues I identified,

9. We are only interested in “original tweets”, no “retweets”; the retweet data is in columns like retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp.

10. Reply tweets are not “original tweets” either; this data is stored in the columns in_reply_to_status_id and in_reply_to_user_id

11. Dog stages (doggo, floofer, pupper, puppo) are spread in different columns.

12. Breed Predictions, Confidence intervals and Dog tests are spread in three columns.

13. All data frames will be merged into 1 using tweet_id as the primary key.

STEP 3: DATA CLEANING

This is the final stage of data wrangling. Here, I cleaned all the data quality and tidiness issues using the Define Code, Test frame work which I clearly documented each step. After, cleaning the three data frames, I merged them into one data frame which I called twitter master data frame using the tweet_id column which is the primary key.

STEP 4: STORING DATA

In this phase, I just proceeded to copy the twitter master data frame into the twitter_archive_master.csv file and store it there.

My data frame is now ready to draw insights, analyze and visualize.
I’ll be documenting my analysis and visualization process in a separate document called Act Report.