Wrangling Efforts

1. Gathering:

Gathering data was done on three steps:

1.1 Twitter archived enhanced: this file was given to me as an internal file in form of csv file that was read into a pandas data frame using the built-in function .read_csv()

1.2 Image predictions file: downloaded programmatically from one of the Udacity's servers using the request library then read into a data frame using .read_csv()

1.3 Tweet json: use the twitter API to gather additional information and store it in a data frame

2. Assessing:

The gathered data was assessed both visually and programmatically and the following issues were detected:

2.1 Quality Issues:

2.1.1 Twitter Archive

2.1.1.1 Missing values in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and expanded_urls columns

2.1.1.2 Inaccurate values in name column

2.1.1.3 Missing values in the dog stage columns represented as None

2.1.1.4 Erroneous data types (timestamp column)

2.1.1.5 Inaccurate values in rating_numerator, rating_denominator columns

2.1.1.6 Include retweets which mean there are duplicates

2.1.1.7 Source column contain HTML tags

2.1.2 Image Predections

2.1.2.1 Multiple columns for dog breeds.

2.1.3 Tweet Json

2.1.3.1 Include retweets which mean there are duplicates

2.2 Tidiness Issues:

2.2.1 Multiple columns representing the dog stage (doggo, floofer, pupper, puppo)

2.2.2 Tweet_id column represented as tweet_id in twitter_archive and image_predictions data frames and as id in tweet_json data frame

3. Cleaning:

The addressed issues was properly cleaned, then the data frames were merged together to form one master data frame that was stored in a csv file called twitter_archive_master and used to perform the analysis and visualizations.