Wrangling Efforts

The tasks started with gathering data from three different sources: a link provided by Udacity Class, an URL also provided by Udacity Class and from Twitter's API. While the first source was only a link to download a table with basic information about @dog_rates (stored in the table "twitter_archive"), the second source was accessed with "requests" package and contained information about the images of dogs in the tweets (stored in "image_predictions"). The third source, Twitter API, was accessed with "tweepy" package (stored in "additional_info_tweets"), all used in Python 3.6.

Next step was merging all information in a new table called "twitter_archive_master". The "additional_info_tweets" and "image_predictions" tables contained fewer results compared to the "twitter_archive" table, so after the merge many rows contained NaN values. Before analyzing quality and tidiness problems in the dataset, we first identified the core variables to use in this project, so others could be dropped. The chosen variables were: tweet_id, created_at, full_text, rating_numerator, name, doggo, floofer, pupper, puppo,jpg_url, img_num, p1, p2, p3, favorite_count, retweet_count.

Then, we investigated the data for quality and tidiness problems, using methods like "groupby()", "agg('count')", "head()", as well as indexing with expressions to get subsets of the dataframe, like "twitter_archive_master[twitter_archive_master.name == 'None']". This resulted in identifying 11 quality problems (3 for completeness, 3 for validaty, 4 for accuracy and 1 for consistency) and 2 tidiness problems. Quality problems included duplicated information, missing values, tweets that did not contained tweets about dog ratings, incorrect names, inadequate variable type, lack of standards such as capital letters vs lower letters.

The cleaning step was done based on the problems above, using different approaches. The result was a "twitter_archive_master" table with 804 entries and 10 variables, ready for the next project step: data exploration analysis.