Data Wrangling report

I started this project by gathering the data from three different sources:

- **The first data frame is “twitter_archive_enhanced.csv”, this data frame was downloaded manually. This data frame has 2356 entries and 17 columns, these columns are (tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp,source,text, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls, rating_numerator, rating_denominator, name, doggo, floofer, pupper, and puppo). It’s represented as “twt_archive”**

- **The second data frame was hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv. It was called “image_predict”**

- **The third data frame was a bit challenging, I had to query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet-json.csv file and the data frame named “tweet_info”.**

After that all the data frames where assessed manually and programmatically and then I found these issues in quality and tidiness:

Issues with quality

- There is no need for the columns 'retweeted_status_id' 'retweeted_status_user_id' and 'retweeted_status_timestamp' since they are not useful.
- Remove retweets
- Incorrect dog names
- Tweets with no images
- Extra characters after '&' which are not important.
- Sources are hard to read.
- The data type for the source can be changed.
- The data type for the column 'timestamp' is incorrect.

Issues with the tidiness

- Dog "stage" variable in four columns: doggo, floofer, pupper, puppo.
- All tables should be part of one dataset.

I created a copy of each data frame so, “twt_archive_clean”, “image_predict_clean”, and “tweet_info_clean” where created to save the cleaned data without changing the original data. Then after cleaning the issues stated above, a new data frame was created under the name of “unified_df”. This master data set contains information from all the three data frames. Finally, few plots where created to drive insight and they are presented in the jupyter notebook and in the file “act_report”.