Reporting: wrangle_report

Gathering data

Three different datasets were gathered:

- twitter_archive_enhanced.csv data which was downloaded manually
- image_predictions.tsv data was downloaded programmatically using the url by udacity
- tweet_json.txt data was also queryed via the Twitter API using the tweepy library

Assessing Data

All the three datasets were assessed visually and programmatically for quality and tidiness issues.

Quality issues

- The column 'tweet_id' is an integer.
- The 'rating_denominator' column contains ratings more than 10.
- The 'rating_numerator' column contains ratings more than 20.
- Names format is not consistent. some having starting with lowercase and otherr too with uppercase.
- The column 'name' contains 'None' instead of NaN and some values have unusual names character lengths.
- Not appropriate column name for 'text'.
- There are some duplicate values in the 'jpg_url' column.
- The columns 'p1', 'p2', and 'p3' contains underscores instead of spaces in some of the values.
- The columns 'p1', 'p2', and 'p3' names are not descriptive and lowercase for some
- Retweets not needed for analysis.
- Some unused columns for analysis.

Tidiness Issues

- All three different data frames should be in a single data set.
- There are four different columns (doggo, floofer, pupper, and puppo) for dog stages.

Cleaning Data

First, all the three datasets were merged into a single dataset then using the various cleaning techniques the data was nicely cleaned.

Storing Data

The data was stored was stored to a csv file, 'twitter_archive_master.csv'

Analysis and Visualizations

A few insights and analysis were made on the cleaned and also visually the insights.