Wrangle Report

➢ About dataset

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage. WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

- Tasks in the data wrangling consists of:
  - Gathering data
  - Assessing data
  - Cleaning data

Gathering Data : we will gathering data from the following resources.

- The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students.
- The tweet image predictions. what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students.
- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data I find interesting.

Assessing Data :

After gathering each of the above pieces of data, we assess them visually and programmatically for quality and tidiness issues. We detect and document at least **eight (8) quality issues** and **two (2) tidiness issues** .

Quality Issues

df:

- Completeness:
  - missing data in the following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
  - tweet_id is an int (applies to all tables)
- Validity:
  - dog names: some dogs have 'None' as a name, or 'a', or 'an.'
  - this dataset includes retweets, which means there is duplicated data (as a result, these columns will be empty: retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp)
- Accuracy:
  - timestamp is an object
  - retweeted_status_timestamp is also an object (the other retweeted statuses are floats)
  - rating_numerator goes up to 1776
- Consistency:
  - rating_denominator should be a standard 10, but there are a multitude of other values

images_df:

- Validity:
  - p1, p2 and p3 columns have invalid data.
- Consistency:
  - p1, p2 and p3 columns aren't consistent when it comes to capitalization: sometimes the dog breed listed is all lowercase, sometimes it is written in Sentence Case.
  - in p1, p2 and p3 columns there is an underscore for multi-word dog breeds

tweets_df:

- Completeness:
  - missing some data

Tidiness Issues

df:

- four columns all relate to the same variable (dogoo, floofer, pupper, puppo)

Images_df:

- this data set is part of the same observational unit as the data in the archive - one table with all basic information about the dog ratings

Cleaning Data: Wrangling process will consists of the following:

- Define

  (1) Merge the clean versions of df, images, and tweets_df dataframes
  Correct the dog types
  (2) Create one column for the various dog types: doggo, floofer, pupper, puppo
  (3) Delete retweets
  (4) Remove columns no longer needed columns
  (5) Change tweet_id from an integer to a string
  (6) Change the timestamp to correct datetime format
  (7) Correct naming issues
- Code
- Test