Wrangle Report

We were given a tweet archive dataset from Twitter user weratedogs (@dog_rates). WeRateDogs is an account on Twitter that posts images and comments of different dogs.

The major goal of the project was to go through the entire data wrangling process from gathering to assessing to cleaning data after which the cleaned dataset is used to run an analysis.

The high-level process was involved:

- Data gathering
- Assessing of gathered data
- Cleaning of data based on issues identified
- Storing, analyzing and visualizing the data to get insights

Data Gathering

This involved gathering data across three sources. Data gathered was in different formats (csv, via a URL and via twitter APIs)

- The first dataset represented the Twitter archive of WeRateDogs. This wprovided by Udacity in a csv format. The file named twitter_archive_enhanced.csv was read using the pandas read_csv function.
- The second dataset represented the tweet image predictions with information on dog breed based on predictions by a neural network. This was also provided by Udacity and pulled from a URL and was read using the pandas read_csv function while specifying the separator as '\t'
- The third dataset was gathered from Twitter APIs using the Python Tweepy library. The purpose of this was to pull favorite and retweet count.

Assessing of Gathered Data

In this section, the goal was to look through the various datasets to identify issues present in each of them. For this, a programmatic and visual assessment was run on each dataset. The requirements for this were split into quality and tidiness. For quality, I checked for:

- Completeness: Identify missing data
- Validity: Check that data present are valid and adequately represent their fields
- Accuracy: Asides validity, is the data correct?
- Consistency: Is it standardized across fields?

While for tidiness, I checked that:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a taable

The issues identified in this process are listed below:

Quality Issues

twitter_archive:

- Missing data in columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
- Dog names: some dogs have 'None' as a name, or 'a', or 'an.'
- Dataset has retweets which will lead to duplicated data resulting in some empty columns (Columns with the the retweeted tag)
- Datatype of tweet_id is int (on all tables)
- Datatype of timestamp is an object
- Rating_numerator has values up to 1776
- Rating_denominator has values above 10
- The source column still has HTML tags

image_predictions:

- p1, p2 and p3 columns have invalid data
- p1, p2 and p3 columns aren't consistent: It varies between lowercase and sentence case
- Multi word breeds are represented with an underscore in columns p1, p2, p3

action_counts:

- Missing data in columns

Tidiness Issues

twitter_archive:

- Dogoo, Fluffer, Pupper and Puppo all relate to a variable

image_predictions:

- Should form one observational unit with twitter_archive

action_counts:

- Should form one observational unit with twitter_archive

Cleaning of data based on issues identified

In this section, I cleaned the data based on the issues I had identified. Steps taken are:

- Copy the indepedent data sets and merge together to form one table/dataset
- Correct issues in the name column of twitter_archive
- Delete retweets
- Remove columns with missing data not needed for the analysis. These are: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
- Change datatype of tweet_id from integer to string
- Change datatype of timestamp from object to datetime format
- Melt the various dog types into one column
- Remove all columns which are no longer needed
- Standardize dog ratings for consistency
- Create new breed column using data from the image prediction table

This involved coding and testing after which the data was stored and further analysis run.