Wrangling Report

Gathering

Data is gathered from **3 sources**:

1. A file on hand *'twitter-archive-enhanced.csv'* that contains the **WeRateDogs** Twitter archive was read into the *twitter_archive* data frame using read_csv()

2. A file '*image_predictions.tsv'* hosted on Udacity's servers and was downloaded from the following
   URL: **https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv** programmatically using the Requests library and saved locally then read into *image_prediction* data frame using read_csv()

3. A list of twitter ID's from the WeRateDogs Twitter archive was created to perform a query on the Twitter API for each tweet using Python's Tweepy library. Each tweet's JSON data was retrieved and stored in the *'tweet_json.txt'* file and read into the *tweet_status* data frame using read_json()

Assessing

Data was assessed by examining each data frame both visually and programmatically. Both Quality and Tidiness issues were noted.

Quality Issues

twitter_archive table

- Nulls represented as 'None' in name, doggo, floofer, pupper amd puppo columns
- Some tweets are retweets
- Some rating are dates such as 9/11 and 4/20
- Incorrect dog names such as 'a', 'an', 'such', 'the','quite'.
- Tweets that are not ratings
- Erroneous datatypes (tweet_id, timestamp, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, rating_numerator, rating_denomnator)

image_predictions table

- Erroneous datatypes (tweet_id)
- Predictions that are not dogs

tweet_status table

- Erroneous datatypes (id_str, in_reply_to_status_id_str, in_reply_to_user_id_str, quoted_status_id_str, favorite_count, retweet_count)

Tidiness Issues

- Four columns (doggo, floofer, pupper, puppo) in *twitter_archive* table should be represented by one column (stage)
- *image_predictions* table should be merged with *twitter_archive* table. There are 2356 tweets in *twitter_archive* table and 2075 image predictions in *image_predictions* . We only want tweets with images.
- favorite_count and retweet_count should be merged into *twitter_archive* table

Cleaning

Before cleaning was conducted eat data frame was copied to new data frames: *twitter_archive_clean, image_predictions_clean and tweet_status_clean*

Quality Issues

twitter_archive_clean

1. **Issue:** Nulls represented as 'None' in name, doggo, floofer, pupper amd puppo columns
   Define: Replace all None values with NaN

2. **Issue:** Some tweets are retweets
   Define: Only keep rows that are original tweets (i.e. drop retweets) then drop columns related to retweets (retweeted_status_id, retweeted_status_user_id, retweeted_status_id, retweeted_status_timestamp)

3. **Issue:** *Some rating are dates such as 9/11 and 4/20*
   Define: In order to get accurate ratings, read through text column and only select ratings that have denominator of 10

4. **Issue:** *Incorrect dog names such as 'a', 'an', 'such', 'the', 'quite'*
   Define: *Extract dog names from text column following keywords such as 'name is', 'named', 'This is', 'Meet', 'Say hello to'*

5. ***Issue:** Tweets that are not ratings*
   ***Define:** Drop tweets that are not ratings. Keep the rows that have a non-null denominator.*

6. ***Issue:** Erroneous datatypes*
   ***Define:** Convert tweet_id, in_reply_to_status_id and in_reply_to_user_id to string data types. Convert and timestamp to datetime data type.*

image_predictions_clean

7. **Issue:** Erroneous datatypes
   Define: Convert tweet_id to string data type

8. **Issue:** Drop predictions that are not dogs
   Define: Only keep predictions that have number one predictions that are dogs.

tweet_status_clean

9. **Issue:** Erroneous datatypes
   Define:
   - Convert id_str, in_reply_to_status_id_str, in_reply_to_user_id_str and quoted_status_id_str to string data type.
   - Convert favorite_count and retweet_count to int data type.

<u>Tidiness Issues</u>

1. **Issue:** Four columns (doggo, floofer, pupper, puppo) in twitter_archive table should be one column (dog stage)
   Define:
   - Combine the doggo, floofer, pupper and puppo columns into one column called stage.
   - Change the new stage category to a category data type.
   - Drop the doggo, floofer, pupper and puppo columns.

2. **Issue:** image_predictions table should be merged with twitter_archive table.
   There are 2356 tweets in twitter_archive table and 2075 image predictions in image_predictions. We only want tweets with images.
   Define:
   - Merge twitter_archive table with image_predictions to only keep top dog prediction.
   - Drop all other columns from image_predictions table.
   - Rename p1_dog to breed.
3. **Issue:** favorite_count and retweet_count should be merged into twitter_archive table
   Define: Merge only favorite_count and retweet_count columns to master table

The final step in the cleaning process was to drop all the unnecessary columns for data visualization. The following columns were dropped: 'text', 'in_reply_to_status_id', 'in_reply_to_user_id', 'source', 'expanded_urls', 'jpg_url' and 'img_num'

Storing

The cleaned data was stored locally to the ‘twitter_archive_master.csv’ file.