Data Wrangling

Gathering Data

The Data for this project was gathered from three(3) different sources.

- The **WeRateDogs Twitter archive** in csv format(twitter_archive_enhanced.csv) was downloaded manually from udacity from this link https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv
- The **tweet image predictions** file (image_predictions.tsv) which is hosted on Udacity's servers, was downloaded programmatically using the Requests library and the following URL:https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- **Additional Data from the Twitter API** was obtained by querying Twitter's API to gather this data and store as an entire set of JSON data in a file called tweet_json.txt.

Assessing Data

Each piece of gathered data was assessed both visually and programmatically for quality and tidiness issues.

- **Visual assessment**: each piece of gathered data is displayed in the Jupyter Notebook with pandas' head() and sample() function to display the first 50 rows and random 50 rows of each piece of gathered data. Each piece of data was also assessed in an external application, Microsoft Excel.
- **Programmatic assessment:** pandas' functions and methods were used to assess the data.

Assessing Observations

Below are some issues that detected during the assessment:

Quality

Enhanced Twitter Archive(df)

- From the assessing data objectives it was stated that only original data would be needed, hence, some columns won't be needed. For example, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp
- timestamp is object dtype instead of datetime
- tweet_id is int dtype instead of object or string
- Nulls represented as 'None' in the name column
- Duplicated and very unusual dog names like 'a' and 'an'

- Unnecessary html tags in source column in place of utility name.
- Remove rows in rating_numerator that were not correctly extracted
- Change rating_numerator and rating_denominator from int dtype to object or float dtype

Image Predictions File(img_df)

- tweet_id is int dtype instead of object or string

Additional Data via the Twitter API(tweet_data)

- Column named id instead of tweet_id
- id as int type instead of string type

Tidiness

- All tables should be part of one dataset
- Column doggo, floofer, pupper, and puppo in the df could be in one column not 4

Cleaning Data

The following was done with code during the cleaning process to solve the issues detected earlier

Quality

Enhanced Twitter Archive(df)

- Remove rows in rating_numerator were not correctly extracted
- Change rating_numerator and rating_denominator from int dtype to object or float dtype
- Remove retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp column.
- Change timestamp from object dtype to datetime dtype
- Change tweet_id from int dtype to object or string dtype
- Change Nulls represented as None in the name column to NaN
- Change rows with very unusual dog names like 'a' and 'an' to NaN
- Removing the anchor link and retaining only the text for source column

Image Predictions File(img_df)

- Change tweet_id from int dtype to object or string

Additional Data via the Twitter API(tweet_data)

- Change column name from id to tweet_id
- Change column id (now tweet_id) from int dtype to string dtype

Tidiness

- Join column doggo, floofer, pupper, and puppo in df into one column not 4
- Join all tables into of one dataset

Storing Data

Cleaned data was joined and stored in a cleaned master DataFrame in a CSV file with the main name **twitter_archive_master.csv**.