Data Wrangling

Gathering Data

Data for this project was collected from three different sources:

- The **WeRateDogs Twitter Archive** in CSV format (twitter_archive_enhanced.csv) was manually downloaded from the following link: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv.
- The **Tweet image predictions** file (image_predictions.tsv) was hosted on Udacity's servers and 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 data and storing it as an entire set of JSON data in a file called `tweet_json.txt`

Assessing Data

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

- **Visual assessment** involved displaying each piece of gathered data in the Jupyter Notebook using the `head()` and `sample()` functions to display the first 5 rows and a random sample of 5 rows respectively. Data was also assessed in an external application, Microsoft Excel.
- **Programmatic assessment** involved using pandas' functions and methods to assess the data for quality and tidiness issues.

Assessing Observations

Quality

General

- Also, tweet_id and id column is an int dtype instead of object or string across dataset

Enhanced Twitter Archive(`df`)
- Some columns are mostly empty and not needed for the assessing data objectives. This include; `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp`,`in_reply_to_status_id`, and `in_reply_to_user_id`.
- `timestamp` is an object dtype instead of datetime.
- `tweet_id` is an int dtype instead of object or string.
- Nulls represented as 'None' in the `name` column.
- Duplicated and unusual dog names in the `name` column like 'a' and 'an'.
- Unnecessary HTML tags in the `source` column instead of utility name.

Image Predictions File(`img_df`)
- Refining p1, p2 and p3 columns and confidence associated with them by combining.
- Inconsistent capitalization in the prediction column.

Additional Data via the Twitter API(`tweet_df`)
- Column named `id` instead of `tweet_id`. For easier merging

Tidiness
- All tables should be merged into a single dataset.
- The `doggo`, `floofer`, `pupper`, and `puppo` columns in `df` could be combined into a single column.

Cleaning Data
During the cleaning process, the following steps were taken to address the issues detected earlier:

Enhanced Twitter Archive(`df`)
1. Remove rows in `rating_numerator` that were not correctly extracted
2. Change `rating_numerator` and `rating_denominator` from int dtype to object or float dtype
3. Remove the columns `retweeted_status_id`, `retweeted_status_user_id`,`retweeted_status_timestamp`,`in_reply_to_status_id`, and `in_reply_to_user_id`.
4. Change the `timestamp` column from object dtype to datetime dtype.
5. Change the `tweet_id` column from int dtype to object or string dtype.
6. Change the 'None' values in the `name` column to NaN.
7. Change the rows with very unusual dog names in the `name` column like 'a' and 'an' to NaN.

8. Remove the anchor link and retain only the text for the `source` column.

Image Predictions File(`img_df`)

1. Change `tweet_id` from int dtype to object or string
2. Refine p1, p2 and p3 columns and confidence associated with them by combining.
3. Inconsistent capitalization in the prediction column.

Additional Data via the Twitter API(`tweet_df`)

1. Change column name from `id` to `tweet_id`
2. Change the `tweet_id` column from int dtype to string dtype.

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”.