Wrangle Report

Wrangling Process:

1. Gathering Data from multiple sources.
2. Assessing
3. Cleaning
4. Storing, Analyzing, and Visualizing the data.

Gathering Data

Three sources of data:

1. **twitter_archive_enhanced.csv**: Downloaded directly and imported using pd.read_csv.
2. **image_predictions.tsv:** Downloaded Programmatically from Udacity's Server.
    - This file has image predictions for every tweet up till August 1, 2017. This was downloaded using the requests library following this url.
        - https://video.udacity-data.com/topher/2018/November/5bf60c69_image-predictions-3/image-predictions-3.tsv
3. **df_api** is data queried through Twitter's API. Using the Tweet IDs in the WeRateDogs archive, we query the API for each tweet's JSON data using Python library Tweepy, and store content of each tweet in JSON format into a text file tweet_json.txt file. This file is then loaded into pandas DataFrame as df_api.
    - The code for this is available in the notebook and is commented out.

Key Points

Key points to keep in mind when data wrangling for this project:

- We only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
- Assessing and cleaning the entire dataset completely would require a lot of time, and is not necessary to practice and demonstrate skills in data wrangling. Therefore, the requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.
- Cleaning includes merging individual pieces of data according to the rules of tidy data.

- The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.
- We do *not* need to gather the tweets beyond August 1st, 2017. You can, but note that we won't be able to gather the image predictions for these tweets since we don't have access to the algorithm used.

Assessing Data

The three datasets were assessed both visually and programmatically. The issues that were discovered are listed below.

Issues Found with Datasets

Data was further inspected visually using Excel and the following issues were spotted:

Inspecting for Quality

Image Predictions Table

```
No issues found.
```

Enhanced Archive Table

```
- We only want Original Ratings, no retweets or replies.
- All denominators should be 10
- Since all denominators should be 10, we can remove this column, and change 'rating_numerator' column to 'rating'.
- Timestamps Data Type is 'Object' and should be converted to Pandas Date-Time Data Type Format.
- Some of the names of dogs were erroneously extracted such as 'a', 'unacceptable', or 'infuriating'.
- To make things simple, we should only take the columns that will be useful.
- Dog nicknames columns have None, instead of NaN.
- Ratings with decimal values are incorrectly extracted. RegEx might be helpful here.
- ID Fields should be converted to strings. Numerical operations aren't supposed to be applicable to them.
```

API Queried Table

```
No issues found.
```

Inspecting for Tidiness (structure)

> All three tables should be merged into one DataFrame.

Image Predictions Table

```
- For each dog, the top 3 predictions of their breed are there along with the True label of whether they actually are a
dog or not. For dog predictions; of the ones that are correctly predicted as a dog, we keep the one with the highest con
fidence.
```

Enhanced Archive Table

```
- Dog nicknames columns (floofer, doggo, pupper, puppo) should all be one column, as a Categorical Data Type.
```

API Queried Table

```
- Source Column in Enhanced Archive Table, and API Queried Table is surrounded by tags. Tags can be removed using RegEx.
```

Cleaning:

The following steps were taken to clean the dataset before analyzing it.

- Inner Merging of three datasets into one table, on Tweet ID.
- Only original tweets were taken.
- Extra columns were dropped.
- Mistakes in Dog Names were fixed.
- Denominators and Numerators were fixed.
- Timestamps were converted to Date-Time format and stripped of extra characters.
- RegEx was used to remove HTML Tags from “Sources” Column.
- Highest Confidence prediction was taken from all three dog breed predictions for each entry.
- Ratings with decimal values were initially incorrectly extracted. RegEx was used to extract them properly, and then they were assigned a float type.
- Multiple dog stages were merged together.
- ID Fields were converted to strings.

Storing

Data was stored into a new CSV file after wrangling, before analysis.

New file is loaded and then analyzed.

Analysis is found in enclosed document Wrangle_Act.pdf