Data Wrangling Report

Introduction

Real-world data rarely come clean. Using Python and its libraries, we gathered data from various sources and formats, assessed its quality and tidiness, and cleaned it. The used data is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc.

Data Gathering

I started the data wrangling process by gathering the necessary data and converting them to Panadas dataframes, which are:

- WeRateDogs Twitter archive data (twitter_archive_enhanced.csv), a CSV file provided by Udacity.
- Tweet image prediction (image_predictions.tsv), a TSV file that was downloaded programmatically using the Requests library.
- Additional data via the Twitter API (tweet_json.txt), a TXT file that was provided by Udacity, since I had issues with my Twitter developers account that prevented me from directly using Twitter API.

Data Assesing

We continue the data wrangling process by assesing the data both visually and programmatically. This process would help us in detecting quality and tidiness issues in our data.

- **Quality issues for Twitter Archive Data**

1. Missing values in columns from in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, and expanded_urls.

2. The columns rating_numerator and rating_denominator have some incosistent values in the numerator and denominator, such as numerators < 10 and denominators other than 10.

3. The column timestamps is in string instead of DateTime data type.

4. Error in dog names, such as (actually, officially, O, a, not, his, etc).

5. The columns which have missing values in doggo, floofer, pupper, and puppo has None instead of NaN.

6. The text contains some 156 retweets which begins with RT @dog_rates: and this is irrelvent to our case since we need original tweets for the dog ratings.

- **Tidiness issues for Twitter Archive Data**

  The columns doggo, floofer, pupper, and puppo are all dog stages and should be merged into one column called dog_stages. This could reduce the dimensionality of the data.

- **Quality issues for Tweet Image Prediction Data**

1. The predicted dog species in columns p1, p2, and p3 are not consistent since some species start with a capital letter and some start with a small letter.
2. The space is represted by an underscore _ in p1, p2, and p3.
3. There are some missing images since twitter_archive has 2356 tweets and image_prediction has 2075 image predictions.

- **Tidiness issues for Tweet Image Prediction Data**

  The image_prediction dataframe should be merged with twitter_archive on the tweet_id.

- **Quality issues for Additional data via the Twitter API**

1. The data type of tweet_id is string and it should be compatible with tweet_id data type in twitter_archive dataframe.
2. There are some missing tweets since twitter_archive contains 2356 tweets and tweet_data contains 2276 tweets.

- **Tidiness issues for Additional data via the Twitter API**

  The tweet_data dataframe should be merged with twitter_archive on tweet_id.

Cleaning Data

In this step and after assesing the data, we need to fix the identified quality and tidiness issues in all the dataframes. This is a crucial step for accuracy and drawing conclusions. We began by making a copy of the original pieces of data to perform the cleaning step on them. We followed the define-code-test framework, which defines the issue and its solution, program the solution, then view and test the outcome.

Cleaning `twitter_archive` dataset

Quality Issues

Issue #1:

Missing values in columns from `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp`, and `expanded_urls`.

Solution:

Drop columns `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp`, and `expanded_urls` since they are irrelevant to our analysis.

Issue #2:

The columns `rating_numerator` and `rating_denominator` have some incosistent values in the numerator and denominator, such as numerators < 10 and denominators other than 10.

Solution:

Drop rows with numerator values less than 10 and denominator values not equal to 10

Issue #3:

The column `timestamps` is in string instead of DateTime data type.

Solution:

Convert `timestamp` data type to DateTime instead of string.

Issue #4:

Error in dog names, such as (actually, officially, O, a, not, his, etc).

Solution:

Store the names in a list and replace them in the dataframe into `None`.

Issue #5:

The columns which have missing values in `doggo`, `floofer`, `pupper`, and `puppo` has None instead of NaN.

Solution:

Replace 'None' string values in any of these four columns with NaN.

Issue #6:

The text contains some 156 retweets which begins with `RT @dog_rates:` and this is irrelvent to our case since we need original tweets for the dog ratings.

Solution:

Drop the rows with text that begins with `RT @dog_rates:`.

Tidiness Issues:

Issue #1:

The `tweet_data` dataframe should be merged with `twitter_archive` on `tweet_id`.

Solution:

As done on the cell above, the two dataframes are merged to `image_archive` dataset.

Cleaning `tweet_data` dataset

Quality Issues

Issue #1:

The predicted dog species in columns `p1`, `p2`, and `p3` are not consistent since some species start with a capital letter and some start with a small letter.

Solution:

Capitalize the first letter of each dog race in `p1`, `p2`, and `p3`.

Issue #2:

The space is represted by an underscore `_` in `p1`, `p2`, and `p3`.

Solution:

Replace the underscore by an empty space in `p1`, `p2`, and `p3`.

Issue #3:

There are some missing images since `twitter_archive` has 2356 tweets and `image_prediction` has 2075 image predictions.

Solution:

Merge the `image_prediction_clean` and `twitter_archive_clean` datasets on `tweet_id` into a new dataframe.

Tidiness Issues:

Issue #1:

The `tweet_data` dataframe should be merged with `twitter_archive` on `tweet_id`.

Solution:

As done on the cell above, the two dataframes are merged to `image_archive` dataset.

Cleaning `tweet_data` dataset

Quality Issues

Issue #1:

The data type of `tweet_id` is string when it should be integer

Solution:

Convert `tweet_id` data type from string to integer

Tidiness Issues:

Issue #1:

The `tweet_data` dataframe should be merged with `twitter_archive` on `tweet_id`.

Solution:

As done on the cell above, the two dataframes are merged to `tweet_archive` dataset.

Storing Data

We saved the gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv". For additional storage, we saved the dataframes in a SQL database.