Data Wrangling Report for WeRateDogs

By Shaimaa Badawy

The Wrangling Process consists of three stages: Gathering Data, Assessing Data, and Cleaning Data.

1- Gathering Data:

I gathered the required data for this project from three different resources:

1- WeRateDogs Twitter archive, which an enhanced version delivered to us “Udacity Students” by Udacity team.

2- image_predictions.tsv hosted on Udacity’s servers, I downloaded it using requests library.

3- More data about the tweets gathered by using Twitter API.

2- Assessing Data:

I assessed the gathered data both visually and programmatically.

1. Visual Assessment:

To get sense of the gathered data, done by using head(), tail(), and sample() functions.
Here, you will find a description for columns in each Dataframe

df_twitter_arch columns:

- **tweet_id**: unique identifier for each tweet.
- **in_reply_to_status_id**: if the tweet is a reply, it will contain the original tweet’s id.
- **in_reply_to_user_id**: if the tweet is a reply, it will contain the original tweet’s user id.
- **timestamp**: time when this tweet was created.
- **source**: utility used to post the tweet: Android app, iPhone app, or Web Client.
- **text**: actual UTF-8 text of the status update.
- **retweeted_status_id**: if the tweet is a retweet, it will contain the original tweet’s id.
- **retweeted_status_user_id**: if the tweet is a retweet, it will contain the original tweet’s user id.
- **retweeted_status_timestamp**: time of retweet.
- **expanded_urls**: tweet url.
- **rating_numerator**: numerator of the rating of a dog (ratings should have a numerator greater than 10).
- **rating_denominator**: denominator of the rating of a dog (ratings should have a denominator of 10).
- **name**: dog's name.
- **doggo**: one of the dog stages.
- **floofer**: one of the dog stages.
- **pupper**: one of the dog stages.
- **puppo**: one of the dog stages.

df_image_pred columns:

- **tweet_id**: the unique identifier for each tweet
- **jpg_url**: dog's image URL
- **img_num**: the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images)
- **p1**: algorithm's #1 prediction for the image in the tweet
- **p1_conf**: how confident the algorithm is in its #1 prediction
- **p1_dog**: whether or not the #1 prediction is a breed of dog
- **p2**: algorithm's #2 prediction for the image in the tweet
- **p2_conf**: how confident the algorithm is in its #2 prediction
- **p2_dog**: whether or not the #2 prediction is a breed of dog
- **p3**: algorithm's #3 prediction for the image in the tweet
- **p3_conf**: how confident the algorithm is in its #3 prediction
- **p3_dog**: whether or not the #3 prediction is a breed of dog

df_extra_data columns:

- **tweet_id**: the unique identifier for each tweet
- **retweet_count**: number of how many times this tweet has been retweeted
- **favorite_count**: number of how many times this tweet has been liked
- **create_date**: time when this tweet has been created

2. **Programmatically Assessment:**
   Done by using info(), value_counts(), max() functions.

After assessing data I found some problems stated in the following assessing report:

Assessing Report

Quality Issues:

- twitter-archive-enhanced.csv

- some tweets are retweets (Solved)
- some tweets are reply (Solved)
- some tweets missing in image_predictions.tsv (Solved)
- source column has HTML Tags. (Solved)
- timestamp column in wrong datatype, it should be datetime. (Solved)
- tweet_id column in 3 dataframes is in wrong datatype, it should be string. (Solved)
- rating_denominator has different values other than 10. (Solved)
- rating_numerator has values less than 6.(my rule here that it is okay if the rating less than 10, that may point to dislike to that dog but less than 6 may be due to different reasons) (Not Solved)
- some of dogs' names are wrong (some are not even a name and some are in lowercase) (Solved)

- image-predictions.tsv

- columns p1, p1_conf, p1_dog, p2, p2_conf, p2_dog are not descriptive. (Solved)

- twitter_json.txt

- retweet_count and favorite_count are in wrong datatype (float), they should be int (Solved)

Tidiness Issues:

- twitter-archive-enhanced.csv

- four columns for dog stage, it should be one column. (Solved)

- image-predictions.tsv

- three columns for prediction, three columns for confidence, three columns for the breed. (Solved)
- this dataset contains extra data related to Tweets, it should be merged with Archive dataset

- twitter_json.txt

- time column for the tweets repeated in twitter-archive-enhanced.tsv, also this data should be merged in same table represented in twitter-archive-enhanced.csv (Solved)

3- Cleaning Data:

First, I made a copy from all the gathered dataframes. Then, I cleaned the data programmatically using the 3 steps process: define, code, test.

I used some functions during the cleaning such as, drop(), replace(), astype(). And I used regex patterns to find Information.

After the wrangling process, I stored the final cleaned dataframe into csv file to apply the analysis process on it. “twitter_archive_master.csv”