Data Wrangling Report

By, Md Rahamat Ullah

Introduction:

This project is primarily focused on wrangling data from the '**WeRateDogs'** Twitter account using Python. This Twitter account rates dogs with humorous commentary. The rating denominator is usually 10, however, the numerators are usually greater than 10. This aspect was not cleaned as it is part of the humor and popularity of this twitter account. In data wrangling part of the project I mainly focus on fixing the data quality and tidiness issues.

Data Gathering:

There are 3 different sources of data in this project. These datasets were in different formats.

Data source-01: 'WeRateDogs' enhanced twitter archive having 2356 tweet which was provided by in the resource section of this course. So, I downloaded the file and loaded it from my local file system using panda's library. The file was in .csv format.

Data source-02: The tweet image predictions file, i.e., what breed of dog (or other objects, animal, etc.) is present in each tweet according to a neural network. This file ('image_predictions.tsv') is hosted on Udacity's servers and downloaded programmatically using the requests library and the provided URL. The file is in .tsv format.

Data source-03: Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called 'tweet_json.txt' file. Each tweet's JSON data is written to its own line. Each tweet's retweet count and favorite ("like") count is important which I collected using tweepy (for accessing the twitter API). I created a developer account in twitter to get the key and access token which I used to call the twitter API through the tweepy module of python.

The above three datasets were loaded in three different dataframes.

- `df_twitter_archive` has the tweet's text, rating, dog name, dog category etc.
- `df_img_predictions` has the prediction results of a neural network(nn) trying to identify the dog breed in a tweet's picture.
- `df_twitter_api` has retweet, favorite counts, place, language.

Data Assessing - Visually and Programmatically:

I need to detect and document at least 8 quality issues and 2 tidiness issues. Also, I only want original ratings (no retweets) that have images as it is said in project motivation section. Here, I am inspecting the three dataframes for two things:

- Data quality issues and
- Tidiness issues

Quality Issues means content issues like missing, duplicate, or incorrect data
Tidiness issues means structural issues.

There are four dimensions of data quality assessment.

Completeness ~ Are there any missing data in specific rows or columns?
Validity ~ Are there any records not correct due to any reason?
Accuracy ~ Are there any extreme data or unusual data?
Consistency ~ Are they keep the consistence of scale standard or data type?

Quality Issues:

Issue 01: I need those tweets having images. But there are some tweets having no image which I find from the 'expanded_urls' column. There are 2297 entries in this column. So, missing images for 59 tweets. Any tweets without images should be removed.

Issue 02: The datatype of the column timestamp is not correct.

Issue 03: retweeted_status_id has 181 entries which means these 181 tweets are retweets (not original tweets). I should remove these tweets from my analysis.

Issue 04: Some columns are not necessary for analysis purpose. e.g. retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp. I prefer to remove these columns

Issue 05: The value for the column rating_denominator should be always 10. But I see that the mean is 10.46 and the max value is 170 which means there are some wrong entries in this column

Issue 06: The max value for the column rating_numerator is 1776 and mean value is 13.13 which means the max value is an outlier. The rating 1776 is incorrect. So, the column has some incorrect values.

Issue 07: Make the contents of the column source human readable by change the long url links to specific words which are part of the anchor tag. From the value distribution, I find that there are only 4 types of sources:

- Twitter for iPhone
- Vine - Make a Scene
- Twitter Web Client and
- TweetDeck.

Issue 08: The name column has some incorrect names. e.g. 'a' for 55 times !

Issue 09: Also, some names start with capital letter while some start with small letter in the column name.

Issue 10: Column names are not clear and straightforward in df_img_predictions dataframes such as p1, p2.

Issue 11: Prediction of dog breeds involve both uppercase and lowercase for the first letter. e.g. Labrador_retriever, golden_retriever etc.

Tidiness Issues:

Issue 01: There are 4 columns (doggo, floofer, pupper, puppo) for dog stage which are actually values of the dog stage. So, these 4 columns can be replaced with a single column - **dog_stage**

Issue 02: I have total 3 dataframes which have one common column tweet id (or id). So, I can merge three dataframes into a single dataframe.

Data Cleaning:

Tidiness- Issue 01: Create a new column dog_stage, remove individual dog stage columns and fill the empty value with NaN.

Tidiness- Issue 02: Merge 3 dataframes into a single master dataframe using the common column twitter_id (or id)

Quality- Issue 01: Remove the tweets where there are no images ('expanded_urls' column has 'NaN' value).

Quality- Issue 02: Update the datatype of 'timestamp' to datetime

Quality- Issue 03: Remove the retweets

Quality- Issue 04: Remove all unnecessary columns. For example: 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'in_reply_to_status_id', 'in_reply_to_user_id', 'in_reply_to_user_id. 'retweeted_status_id' is not need any more as I already removed the retweets using this column.

Quality- Issue 05: The value for the column rating_denominator should be always 10. But I see that the mean is 10.46 and the max value is 170 which means there are some wrong entries in this column. Correct the incorrect values based on the corresponding text of the tweet.

Quality- Issue 06: Correct the 'rating_numerator' values from the tweet text

Quality- Issue 07: Make the contents of the column ‘source’ human readable by change the long url links to specific words which are part of the anchor tag.

Quality- Issue 08: Some dog names are incorrect. Replace the frequent incorrect names with 'None'.

Quality-Issue-09: Capitalize the name of the dog for consistency.

Quality- Issue 10: Change names of the columns from 'df_img_predictions' datframe for better readability

Quality- Issue 11: Prediction of dog breeds involve both uppercase and lowercase for the first letter. e.g. Labrador_retriever, golden_retriever etc. Capitalize the dog breed predictions.

Finally, I store the df_merged to the file 'twitter_archive_master.csv'.