Wrangle Report – WeRateDogs Twitter Archive

By Emmanuel Chisom Egwuonwu

Introduction

In this project, a couple of wrangling efforts will be used to wrangle the WeRateDogs Twitter archive datasets. The data wrangling steps involves Data Gathering, Data Assessing, Data Cleaning, Data Analyzing and Visualizations. Three different datasets will be gathered, assessed, cleaned, merged into one and analyzed.

Data Gathering

In this section of the wrangling, the three datasets were gathered using different libraries and different methods.

- The Twitter Archive file (twitter_archive_enhanced.csv) was provided by the Udacity and it was manually downloaded and read into a pandas dataframe.
- The tweets image prediction file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests Library from this url: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- The Tweets JSON file was downloaded programmatically by querying the Twitter API using the Tweepy library. The retweet count and favorite count were extracted from this file and read line by line into pandas dataframe.

Data assessing

In this section, the datasets gathered were assessed carefully using both visual and programmatic assessment technique. The quality and tidiness issues observed from these datasets were documented. A number of the quality and tidiness issues noted from the assessment are listed below.

Quality Issues

1. The 'expanded_urls' column contains null entries and duplicated entries
2. The twitter archive dataset contains retweets that needs to be removed
3. The twitter_archive dataframe contains irrelevant columns that won't be used for the analysis such as 'in_reply_to_status_id', 'in_reply_to_user_id', 'source', 'retweeted_status_id', 'retweeted_status_user_id','retweeted_status_timestamp'.
4. The timestamp datatype is incorrect
5. Some texts in the text column contains '&amp' instead of '&'
6. Incorrect dog names
7. Incorrect extraction of ratings from the text column
8. Incorrect entries for ratings with decimal numerators
9. We need only one column each for the dog_breed and confidence level
10. The image_prediction dataframe contains duplicated jpg_urls
11. Incorrect datatypes for the retweet count and favorite count

Tidiness Issues

1. There are four columns showing dog stages that could be melted into one
2. The three dataframes needs to be merged into one main dataframe

Data Cleaning

In this section, the quality and tidiness issues noted from the Data Assessing section were carefully cleaned. The first step taken in this cleaning process was making a copy of the three original dataframes. Some of the Data cleaning efforts employed are listed below:

- Replace entries with 'None' or 'NaN' with an empty string ' ', then, concatenates dog stages entries from the four columns into a new column 'dog_stage', thereafter separate double dog stages with a comma.
- Drop all the null entries in the 'expanded_urls' column and drop the duplicated urls in the 'expanded_urls' column

- Drop all rows containing retweets, where these columns will be non-null: retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp.
- Drop all the irrelevant columns such as 'in_reply_to_status_id', 'in_reply_to_user_id', 'source','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp' from the twitter_archive_clean dataframe.
- Convert the 'timestamp' datatype to datetime.
- Replace all entries that contain '&' in the 'text' column with '&'
- Change all the incorrect dog names to the correct names extracted from the 'text' column using a regex function.
- Correct all the ratings being extracted incorrectly from the 'text' column using a regex function.
- Fix rating numerators that contain decimals by manually setting the correct numerators of those ratings.
- Merge the twitter_archive_clean dataset with the tweets_info_clean dataset and image prediction dataset.

Conclusion

In the course of undertaking this Data wrangling project, a number of data wrangling processes have been used, thereby exploring python and its libraries. Data wrangling is an essential skill for all Data analysts today, therefore this project presented a sterling opportunity to hone my data wrangling skills.