Data Wrangling Report for WeRateDogs Dataset

By Hindhuja Gutha

Wrangling is gathering the data from different sources and transforming raw data to appropriate format and creating a clean dataset which is used for Analysis . WeRateDogs twitter account is famous for their unusual dog rates and images.

Step 1 : Gathering the Data

I have gathered three data sets from three different sources.

1. The WeRateDogs Twitter archive Downloaded manually twitter_archive_enhanced.csv . This file contains basic info about tweets like tweet_id,in_reply_to_status_id,in_reply_to_user_id,timestamp,source,text,rating_numerator etc
2. I have also downloaded image_predictions.tsv programmatically using the Requests library of python. This file contains data about tweet image predictions , which is used to predict breed of the dog.
3. Last dataset is downloaded using the tweet IDs in the WeRateDogs Twitter archive and querying the Twitter API for each tweet's JSON data using Python's Tweepy library and storing 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 in tweet_json.txt. After that, tweet_json.txt file is read line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Data gathering using Twitter API is a new way which I have learned in this project and it's also takes some initial time for setup and to create tweet_json.txt took more time than twitter_archive_enhanced.csv and image_predictions.tsv

Step 2 : Accessing the Data

After gathering all datasets , each data set is loaded to a data frame and each data set sample and contents are observed to gather information about dataset. Below are some of the observations.

Quality :

- Total no of rows in 2356
- in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id & retweeted_status_user_id are of type float64 instead of int64 or Object
- timestamp & retweeted_status_timestamp are of type object instead of Date
- expanded_urls has some null values. only 2297 are non-null.
- There are some null values in the data set

- There are some retweeted data, since we're calculating only for original tweets, need to remove them as well as columns related to retweets
- replace multiple columns for prediction algorithm, confidence level with single one
- we have unusual values for rating_numerator & rating_denominator
- rating_denominator has 0 value which is invalid
- Duplicate tweets in image_prediction dataset
- 'id' column in tweet_info.txt should be renamed to 'tweet_id' so that it can be mapped with other datasets
- Number of rows is different in all three datasets. so any duplicates should be removed after merging all

Tidiness :

- remove columns in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id & retweeted_status_user_id in archive_data(Twitter Archive dataset)
- remove image_num column from image_data(Image Prediction dataset)
- Merge archive_data, image_data & api_data based on tweet_id
- combine doggo, floofer , pupper , puppo columns in twitter archive dataset to single column

Step 3 : Cleaning

In this step , issues found while accessing the data are fixed. I was able to fix most of the above mentioned issues except fixing rating_numerator and rating_denominator because of unusual values.