Wrangle Report : WeRateDogs Twitter Archive Dataset

by Laila Shahreen(November 2020)

This report concisely portrays what exploration and data wrangling have been performed in order to extract insights from twitter data set.

Data Gathering

- The twitter_archive_enhanced.csv file was downloaded manually from the Udacity Project section.
- The image_predictions.tsv file is hosted on Udacity's servers and has been downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

- By querying Twitter API for each tweet's JSON data using Python's Tweepy library, each tweet's entire set of JSON data has been stored in a file called tweet_json.txt file. Each tweet's retweet count and favorite ("like") count has been extracted later and merged with twitter_archive_df

Data Assessing and Cleaning

At first I started to look into the dataframe visually in excel file and also via programmatic display. Summarized info gave us how the entire twitter archive file was structured. I started to look into individual columns and checked their properties.Tidiness and quality issues have been identified then.

The main steps of cleaning are formulated below.

- Since no retweets are desired, all rows with non_null value in retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns were dropped.
- The rows with non_null value in_reply_to_status_id and in_reply_to_user_id columns were eliminated.
- Finally the above columns were also dropped.
- Tweets with missing data in the expanded_urls were dropped.
- The datatype of timestamp was changed from string to datetime.
- The invalid names in the name column were replaced with 'none'.
- The rating_numerator and rating_denominator columns have been revised.Several rows had wrong numerator rating and replaced with correct values.Also the data type was changed from int to float.Only denominators of 10 and numerator upto 14 were accumulated after intermediate cleaning and manual fixing. Then we dropped the denominator column and renamed the rating_numerator as rating.

- The display string portion has been extracted from the html string in the source column to make it visually better and concise.
- The four dog_stage columns have also been minimized into one column and tweets without stages were set to none. Multiple dog stages were separated and fixed manually.
- The dog breed prediction with the highest confidence level was kept only and the table was combined with the archive table.
- Then the json_clean data frame was merged with twitter_archive_clean dataframe.retweet_count and favorite_count data type were converted to int type.
- All the columns were then reordered for better structure and saved a twitter_archive_master.csv file.
- The image_predictions file and json_data file have not been clean. So no additional master files were created and saved.