Data Wrangling

This dataset is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

- ## Gathering Data

  In this project have 3 data set from various sources

  1. **twitter_archive_enhanced.csv**
     downloaded from Udacity and opened by read_csv
  2. **image_predictions.tsv:** i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information also read_csv.
  3. **tweet_json.txt:** file constructed via API
     by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library but don't have any response from twitter then I used file called tweet_json.txt. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

- ## Assessing Data

  By Visually and Programmatically by using different methods (e.g. info(), value_counts(), sample(), duplicated(), groupby(), nunique(),…).

Quality issues

- Erroneous datatypes for 'timestamp' and 'retweeted_status_timestamp' columns
- A lot of missing values
- rating_denominator cannot be 0 because These ratings almost always have a denominator of 10
- Remove outliers from rating_numerator
- Calculate the rate into column
- format source column <a href=><\a>
- Erroneous datatypes in two columns 'retweet_count' and 'favorite_count'
- Rename 9 columns ('p1', 'p2','p3','p1_conf', 'p2_conf','p3_conf', 'p1_dog', 'p2_dog','p3_dog')
- Extact short url from expanded_urls column

Tidiness issues

- four column (doggo, floofer, pupper ,puppo) in one column stage
- Marge all 3 dataframe together
- 2 variables in the same column

- ## Cleaning Data

This part divided into 3 parts: define, code, test

- marge 3 datafram 'df_archive','df_tweet', 'df_image' into one datafram Master_df
- convert datatype for 'timestamp' and 'retweeted_status_timestamp' columns from object (str) to datetime by used function to_datetime
- Incorrect value in rating_denominator,not all value was 10 and based on describe data the denominator 10. I
- replaced all value in the 'rating_denominator' column to number 10

- clean 'rating_numerator' column by removed 'outliers' all values > 20 then calculated by devide 'rating_numerator' on 'rating_denominator' and stored into one column called 'Rating'. After that I droped both columns ('rating_numerator','rating_denominator')
- Source link it has html script code in value. so, replace the value in source column to just value between tag a. Here have four from "Twitter for iPhone" and "Vine - Make a Scene", "Twitter Web Client", "TweetDeck"
- 2 variables in the same column,I Separated to two column Date and Time by using datetime package
- A lot of Missing value in the 4 column 'in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id' ,'retweeted_status_user_id', 'retweeted_status_timestamp' and this columns it cause confusion in data analysis and are not useful information for analysis so, dropped.
- Erroneous datatypes in two columns 'retweet_count' and 'favorite_count I converted to int because must be discrete variable
- Change name columns to clearer descriptive name. Change p1 to Prediction_1 and p2 to Prediction_2, Change p3 to Prediction_3. Depending on this change, I changed other columns that is related to the prediction by rename() method.
- Four columns ('doggo' ,'floofer', 'pupper', 'puppo') in dataset into one column called: 'stage' to make this dataset enhanced. To solve None by replaced to space then merge 4 columns into one column 'dog_stage' and to solve problem many dogs have more than one stage som solved by pass on column and replaced then dropped 4 columns.
- By using split method with text column to extact short url from in last expanded_urls column then stored in short_link column and droped the expanded_urls.