WeRateDogs™ | @dog_rates
This is Atticus. He's quite simply America af. 1776/10
Reference: https://twitter.com/dog_rates/status/749981277374128128/photo/1

Project: Wrangle and Analyze Data

Dataset: WeRateDogs / Wrangle Report

WeRateDogs is a Twitter account. This account rates people's dogs with entertaining comments. Twitter followers of WeRateDogs submit their dogs funny/cute pictures and are rated on a scale of one to ten. Funnily enough, ratings are mostly more than the maximum, such as "13/10". Favorite posts are retweeted, re-posted on Instagram and Facebook.

Gather:

- Twitter archive file 'twitter-archive-enhanced.csv' is given with 17 columns and 2356 entries. Twitter archive enhanced dataset has columns 'tweet_id', 'in_reply_to_status_id', 'in_reply_to_user_id',

'timestamp', 'source', 'text', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls', 'rating_numerator', 'rating_denominator', 'name', 'doggo', 'floofer', 'pupper', 'puppo'.

- Image prediction file 'image-predictions.tsv' is given and is derived from 'twitter-archive-enhanced.csv' when passed through a neural network that classifies breeds of dogs. Dataset 'image-predictions.tsv' has 12 columns and 2074 entries.
- Programmatically created text file using Tweepy by querying Twitter's API for additional data. Used 'tweet_id' from 'twitter-archive-enhanced.csv' to query Twitter's API using Tweepy and stored text into "tweet_json.txt" in JSON format. CSV file 'tweet_json_text.csv' created from "tweet_json.txt" and extracted columns 'tweet_id', 'retweet_count', and 'favorite_count'.
- Actual rating in Column 'rating_numerator' is different from assigned in column 'text. After decimal, values are selected instead of the whole number

Assess

CSV files 'twitter-archive-enhanced.csv', 'image-predictions.tsv', and 'tweet_json_text.csv' assessed for their quality and tidiness.

Each CSV file read using Pandas in a Dataframe format - a two-dimensional data structure.

Used functions like head(), tail(), sample(), info(), duplicated(), value_count(), unique() etc for assessing the datasets columns and entries.

Following **Quality** and **Tidiness** issues are documented while assessing:

Quality

Twitter archive enhanced Table

- Column 'tweet_id' is of type integer. It should be of type string.
- Missing dog name. 745 entries have dog name column "None" Value. Names given by dog owner(s) is a very subjective matter, but I did take risk suggesting names like 'a,' 'his,' 'all,' ' such,' 'an.' are the typos.
- Two rows of column 'rating_numerator' has zero value. For 'rating_numerator' = 1, Erroneous value at 'rating_numerator' for "tweet_id" = 666287406224695296. Column 'text' show rating '9' were as column 'rating_numerator' is 1.
- Few observation have column "rating_denominator" is not equal to 10
- Column "in_reply_to_status_id" has more than 90% of NaN entries
- Column "in_reply_to_user_id" has more than 90% of NaN entries
- Column "retweeted_status_id" has more than 90% of NaN entries
- Column "retweeted_status_user_id" has more than 90% of NaN entries
- Column "retweeted_status_timestamp" has more than 90% of NaN entries
- Column "expanded_urls" has 59 NaN
- Column "timestamp" has "+0000" on every row and is of type String
- Some observations have listed two stage for a single dog.

Image predictions Table

- Column ‘tweet_id’ is of type integer. It should be of type string.

Tweet via API Table

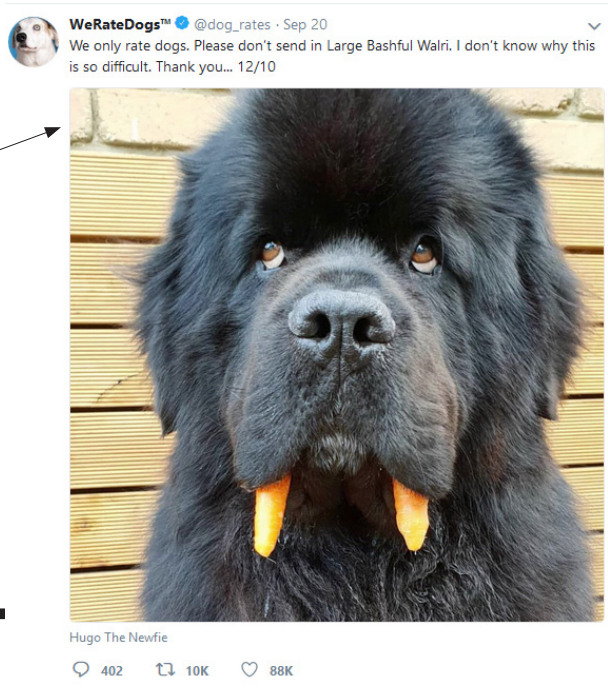
- Column ‘tweet_id’ is of type integer. It should be of type string

Tidiness

- Dog stages in separate columns
- ‘text’ column in the “Twitter archive “ table should split ‘tint_url’ into a separate column.

Clean

All identified quality and tidiness issues were fixed, three dataframes were merged and final CSV file ‘df_master_twitter.csv’ created. File ‘df_master_twitter.csv’ were used for final analysis and insights into data.



This is one of my recent favorite photos. I just cannot stop smiling. I sincerely hope carrots are not hurting dog and dog is also enjoying this funny moment.

Reference: https://twitter.com/dog_rates/status/1042927905507033089