We Rate Dogs

This is a project analysing the data from the archive of a twitter account named WeRateDogs

We Rate Dogs 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. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "**they're good dogs Brent**." WeRateDogs has over 4 million followers and has received international media coverage.

The analysing is divided into Five main steps.

- Gathering
- Assessing
- Cleaning
- Storing
- Data Analysis and Visualization

Gathering :

The project starts with gathering data and saving it in three different files

First file is a csv file named twitter-archive-enhanced.csv which is provided by Udacity classroom, it contains the twitter archive of the account We Rate Dogs

Second file is a tsv file named image-predictions which is hosted on Udacity's servers and downloaded programmatically using the Requests library.

Third file is a txt json file contains each Tweet Json Data collected via the query of the twitter Api using tweepy library based on the tweets' ids in the twitter-archive-enhanced.csv file, it contains additional information about different tweets of the We Rate Dogs account, Each tweet's retweet count and favourite ("like") count.

Assessing :

The three files are loaded into three DataFrames, named api, df_image, df_archive, Different methods like head,tail,describe,info are used to assess the three data frames and get more information about each of them.

A list of the different Quality and Tidiness issues found in the DataFrames is made to represent them.

Assessment Summary

Quality

1. 'id' column is integer not string

3. some values in names column are missing(none) or incorrectly extracted (a,an)

4. 259 (181+78) rows contain retweets and replies

6. There are 59 null expanded_urls values only three of them with null retweets and replies

7. some rating_numerator values are incorrectly extracted from tweets text that has decimals

8. rating numerator max 1776 and min 0

9. rating denominator max 170,min 0,and many other invalid values,all values should be 10

10. timestamp is in string format not datetime

11. remove unrequired columns from archive

12. tweet_id column is integer format not string

13. there are more rows in archive and api tables that have no data in images table

Tidiness

1. dog stages are in four separate columns

2. image table headers names

3. all three dataFrames could be merged into one

Cleaning

A copy of each DataFrame is made,and each quality and tidiness issue is resolved one by one , the cleaned Data Frames are merged into one DataFrame called twitter-master-cleaned.

Storing

DataFrame is stored into a csv file with the same name “twitter-master-cleaned.csv”

Data Analysis and Visualization

In this final step Data Analysis and Visualization to give different insights.