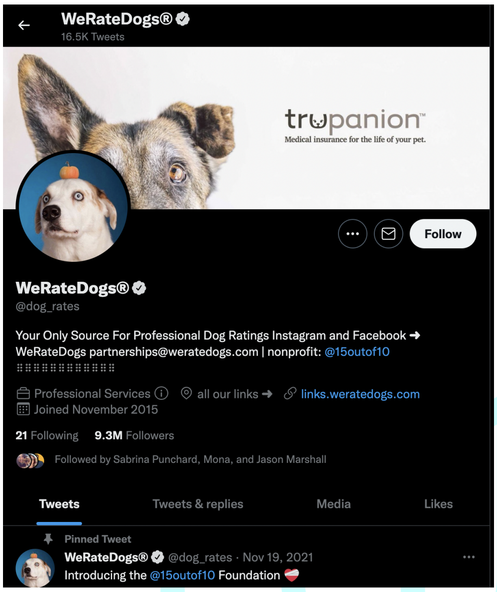
Data Wrangling - WeRate Dogs Twitter Data

This report details the process that went into wrangling (gathering, assessing and cleaning) data from twitter user's @dog_rates page.



WeRateDogs®
@dog_rates

Your Only Source For Professional Dog Ratings Instagram and Facebook ➜ WeRateDogs partnerships@weratedogs.com | nonprofit: @15outof10

Professional Services ⓘ all our links ➜ links.weratedogs.com
Joined November 2015

21 Following 9.3M Followers

Followed by Sabrina Punchard, Mona, and Jason Marshall

Pinned Tweet
WeRateDogs® @dog_rates · Nov 19, 2021
Introducing the @15outof10 Foundation

The wrangling process was carried out in three stages:

1. Gathering
2. Assessing
3. Cleaning

Gathering

The following data was sourced and used for this project:

1. The WeRateDogs twitter archive:
   WeRateDogs downloaded their Twitter archive and sent it over to Udacity, which was provided for this project to download manually. The archive was stored as a csv file, this was imported with pandas read_csv.

2. Image Predictions File:
   This file is hosted in udacity's server, it is a tsv file a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction.

This file was imported using the requests library from this url : https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Additional data from the Twitter API:
   Access to the twitter API was denied, so additional data was provided. This data was gotten with the twitter API as a JSON file, the json was read line by line into a pandas dataframe. This data contained information like; tweet_id, retweets_count and favorite_count.

Assessing

The data on all tables was loaded into the jupyter notebook and assessed visually and also programmatically, using various methods like df.info(), df.describe(). df.duplicated() e.t.c. Various issues were identified during the assessing process, which were later cleaned.

Cleaning

Before cleaning the data, copies of the data were made. The define, code, test framework was used for the cleaning process.

The following cleaning activities were carried out:

1. The tweet source was embedded in html tags, this was resolved using string slicing.
2. The column name source was changed to device in the twitter archive data.
3. Null values stored as ‘None’ were replaced with ‘np.nan’ in the twitter archive data.
4. ‘id’ column in the additional data from API was renamed to ‘tweet_id’.
5. Retweets were removed from the data
6. The data type for tweet_id was changed to string across all tables.
7. Timestamp datatype was changed to datetime.
8. Tweets without images were removed.
9. Multiword dog names in the image prediction table had underscores instead of spaces were replaced with space.(i.e ‘German_shepherd’ replaced with ‘German Shepherd’)
10. All predictions in the image prediction table were changed to title-case.
11. Irregular rating denominator was regularised by creating a new rating column(i.e numerator / denominator).

12. The most confident prediction in the image prediction table was stored as a new column ‘breed’.
13. Dog stages were recorded in three different columns, this was solved by making a dog stage column.
14. Some columns were renamed to be more descriptive, while unwanted columns were dropped.
15. Invalid names in twitter archive tables were changed to null.
16. All tables were merged.

The wrangled data was saved as csv file : ‘`twitter_archive_master.csv`’ and few analysis was carried out.