WeRateDogs Tweets Archive

By Muhammed Jimoh

18th February 2023

The method used to wrangle the WeRateDogs Twitter archive is summarized in this report. The steps involved: Gathering, Assessing, and Cleaning Data.

Gathering Data

The initial step in the process of data analysis is data gathering. Only occasionally is the data accessible in a single location. It typically needs to be acquired from various sources because it is dispersed. The information for this project was acquired from three separate sources.

- The WeRateDogs Twitter archive data was downloaded from Udacity.
- The image dataset, which was downloaded remotely from Udacity servers
- The Tweepy library was used to extract tweet data in JSON format from the Twitter API.

Assessing Data

The process of evaluating the data to identify organization and quality problems in the existing data is known as data assessment. The results of the individual assessments made for each of the three datasets are listed below:

- Twitter Archive Dataset
    - Columns like doggo, floofer, pupper, and puppo are not necessary. Instead, one column category(a different name can be used as well)can serve the purpose
    - Missing values are expressed as "none". (name, duppo, flopper, etc.)
    - The source column includes HTML tags, these tags should be stripped.
    - The names column should be cleaned, there are invalid records like a, the, an, the, very, and unacceptable which start in lowercase.
    - The timestamp and retweeted_status_timestamp column type should be datetime instead of object
    - Not all columns are necessary for the analysis.
- Image Predictions Dataset
    - Column names p1, p2, and p3 columns should be standardized as all lowercase and "-" expressions should be removed.
    - The tweet_id data type should be a string
- Tweets Dataset
    - For joining and uniformity purposes, the id column must be renamed to tweet_id. Moreover, the id column appears twice (id and id str).
    - Contributors, coordinates, and geo are examples of columns that have no significant use in this analysis.
    - The created_at type is an object (str) rather than a datetime. I will change it to datetime.

Cleaning Data

While cleaning the dataset, all of the problems mentioned above for the various datasets were resolved.

Though the aforementioned examination and cleaning technique may not have covered all the issues, we can take the data cleaning a step further.