Wrangling and Analyzing Data

By Israel Imasuen: June 4, 2022

0.1 Reporting: wragle_report

The project's dataset is the collection of tweets from Twitter user @dog rates, commonly known as WeRateDogs. WeRateDogs is a Twitter account that assesses people's pets and comments on them. The following were the objectives of the WeRateDogs Twitter project:

1. Get and Organize Twitter data programmable
2. Data storing, cleansing, and analysing
3. Visual displaying and data insights

0.1.1 Data Gathering

Data was gathered from multiple sources for the WeRateDogs Twitter project, including The Twitter archive enhanced.csv, The WeRateDogs Twitter archive, and The tweet image predictions.

Based on the time period examined, this twitter archive comprises basic tweet data (tweet ID, name, source, timestamp, content, and so on) for all of their tweets.

To obtain the retweet and favourite ("like") counts for each tweet, as well as any other data, the Twitter API and Python's Tweepy package were deployed.

In the tweet image predictions, data such as the kind of breed of dog (German_shepherd, French_bulldog, etc.) were present in each tweet.

0.1.2 Assessing Data

After gathering the data, the next stage is to access and evaluate it for both quality and tidiness concerns. Based on the nature of the data, there are tones of data wrangling activities that can be performed, however, a handful was selected for the purpose of the study. Below are quality and tidiness issues identified.

Quality Issues

1. Some columns are not relevant, especially columns with higher null values
2. Eliminating values without images from data to analyze
3. The name column has 'None' has a name
4. The name has a, an, the, etc. (Observed names that start with the lower case are not the real name)
5. Missing values in expanded_urls

6. Missing values in expanded_urls
7. p1_dog', 'p2_dog', 'p3_dog' having some value set at false. It indicate the false values are not dog
8. HTML syntax among the content of sources rows.
9. tweet_id is having an object data type insteadof int

Tidiness issues

1. Extract values and combined doggo, flooder, pepper, and puppy to a single column form their various columns
2. concate the three dataset and drop duplicate column if any.

0.1.3 Data Cleansing

Following the assessment, the data were processed using the following steps: Define, Code, and Test for all issues identified

Define:

The issues identified were picked one at a time for cleaning

Code:

This section handles the appropriate code needed to cleanse the data

Test:

The data were tested to ensure all the issue been worked on has been resolved

0.1.4 Visualization and Insight

After data has been cleansed, three insights and visualization were also produced based on the analyses from the data.