Tweet archive of WeRateDogs

Oduwole Eunice

July 23, 2020

For this project we were interested in wrangling the tweet archive of Twitter user WeRateDogs. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Data Extraction

Three pieces of data was required for the analysis:

1. twitter archive .csv which was provided in the course resourses
2. image predictions.tsv which was downloaded using request library and anmd the given url
3. tweet json.txt was downloaded using tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using python's Tweepy library

Assessing data

In order to ensure the data to the data was viewed both visually from excel sheet and programmatically using various pandas function. Based this accessment, the following was found:

Quality

- 'Al Cabone' written as 'Al' in the name column
- 'Zoey' written as 'my'
- 'O'Malley' wriiten as O
- There are mistakes in the name column such as 'a', 'getting', 'this', 'all' etc.
- Remove coulmns that are not needed for analysis such as the retweets and the in reply to status id
- The rating denominator column has entries greater than 10 (which is the standard for all the dogs)
- Floating number not capture in rating numerator column
- Erroneous datatypes (tweet id) in both 'twitter archive' and 'image predictions' tables
- Erroneous datatypes (timestamp)
- Underscore in the entry for P1, P2 and P3 in the 'image predictions' table

Tidiness

- The dog stage (doggo, floofer, pupper, puppo) should be combined into a single column
- The Favorite count and retweet count should be part of twitterarchive table.

Cleaning data

- Favorite count and retweet count was cut into another dataframe then join it to the twitter archive using what they have in common(twitter id) .
- Each dog was extracted from the text column into a new column. Then drop the columns doggo,floofer,pupper,puppo
- Replace rows that has names mistaken with the correct name
- Extract the floating number from the text column and then replace the affected columns with the right number
- Locate the names using the lowercase and replace them with nan (also replace the existing 'None' in the coulmn with nan such that we are consitent
- Remove the +0000 from the timestamp column and then convert it from string to datetime format
- Ratings denominator was corrected