wrangle_report

August 11, 2022

0.1 WeRateDogs: wrangle_report

0.1.1 Introduction

In this report, I will be documenting my steps in wrangling WeRateDogs data. WeRateDogs is a Twitter account that rates dogs and comments on them. There are three different datasets involved in my analysis and I will be walking us through the processes involved from beginning to the end as it all started from gathering the data.

0.1.2 Data Gathering

As aforementioned, three different data were used for this analysis, of which each was gathered separately with different approaches.

The three data and the approaches involved are:

1. **The WeRateDog Twitter archive:** This data (twitter-archive-enhanced.csv) was available on hand and thereby downloaded manually before being uploaded into Jupyter Notebook and read into DataFrame.
2. **The tweet image predictions:** This file (subsequently stored as tweet_image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library . It contains the modeling outcome of each image using Neural Network.
3. **The Tweet_json data:** This was created after querying the Twitter API for additional data such as favorite count, retweet count etc. These additional data were scraped from Twitter, read line by line before it was later stored as 'tweet_json.txt'.

0.1.3 Data Assessment

After all the data has been gathered and loaded into the project workspace, I moved on to assessing the data visually and programmatically. Along this process, some quality and tidiness issues were observed and listed below.

Quality issues

Twitter-archive-enhanced data

1. Erroneous datatype (column: timestamp and retweeted_status_timestamp)
2. Some numerators ratings are more than expected.
3. Some denominators are not 10.
4. Some ratings are not tweets but retweets so remove them.
5. Some ratings are not tweets but replies so remove them.
6. The columns retweeted_status_user_id, retweeted_status_user, in_reply_to_status_id are not relevant.

Tweet_image_predictions data

7. The 'tweet_id' column should not be integer but string/object.
8. The predicted names of the dogs are not consisted case-wise under the columns p1, p2 and p3.

Tidiness issues

9. Melt the last 4 columns in twitter archive data into one column since they are redundant.
10. Different tables for the same observation so we merge.

0.1.4 Data Cleaning

All the observations above were not just documented but augmented with the necessary step which is cleaning the data. Each problem was dealt with separately by applying Define-Code-Test approach. This means that each observation was restated, cleaned and tested to ascertain correct output.

After ensuring that all the quality and tidiness issues have been cleaned, a single dataframe was henceforth created for the three tables. This new dataset is stored as 'twitter_archive_master.csv' and it is from this that all our analysis and visualizations were created.