wrangle_report

May 31, 2022

1 Wrangle Report

1.1 We Rate Dogs Data

1.1.1 Introduction

The dataset wrangle in the project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a twitter account that rates people's dogs with humorous comment about the dog.

The WeRateDogs Twitter project goals included: Wrangling the twitter data through the following processes:

Gathering Data, Assessing Data, Cleaning Data, Storing, analyzing and visualizing your wrangled data, Reporting on the data wrangling efforts and data analyse and visualization

Gathering Data The data for this project consist on three different dataset that were obtained as following:

1.Twitter archive file: the twitter archive enhanced.cv was provided by Udacity and downloaded manually. 2.The tweet image predictions, i.e., what breed of is present in each tweet according to a neural network. This file (image predictions.tv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information 3.Twitter API & JSON: by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet son.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

1.1.2 Assessing and Cleaning Data

Some quality and tidiness issues were identified for the three tables and cleaned.

Quality issues 1.Keep original ratings (no retweets) that have images
2.Delete columns that won't be used for analysis in archive_clean table
3.Erroneous datatypes
4.Correct numerators with decimals
5.Error in dog names are not a dog's name
6.Source column is in HTML-formatted string, not a normal string

7.Text columns includes a text and a short link
8.Missing values

Tidiness issues 1.Twitter api table columns(retweet_count, favorite_count, follower_count) and image predictions table should be added to twitter archive table
2.Then dropping tweets with no images

1.1.3 Conclusion

At the end, I stored new cleaned data to the twitter_archive_master.csv file and find out some insights and displayed the visualization(s) produced from my wrangled data.

```
In [ ]:
```

```
In [ ]:
```