wrangle_report

September 28, 2022

0.1 Reporting: wrangle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

0.1.1 About the Datasets Used

In the course of this project, three datasets were wrangled and analyzed and visualized. Viz:

1. twitter_archived_enhanced.csv: The tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage. These dataset was gotten from the Udacity project section. According to Udacity, WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively to be used solely for this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
2. The image_prediction.tsv: This file (image_predictions.tsv) is present in each tweet according to a neural network. It is hosted on Udacity's servers and was downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. tweet_json.txt: All efforts to get the Twitter developer account was abortive. hence, I was unable to get my twitter developer account activated by twitter, so I opted for the alternative as given by Udacity. Hence the twitter_json file as well as the lines of code given in the additional resoiurces section of the classroom was adopted by me.

0.1.2 Wrangling Datasets

Gathering Data

- The twitter_archived_enhanced.csv dataset was gather by directly downloading and saved into pandas DataFramean named df_archive.
- The image_prediction.tsv dataset was downloaded programmatically using the Requests library and the following URL:

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv and saved into pandas DataFrame named df_img_pred.

- the tweet_json.txt dataset was downloaded and read and written line by line to contain the 'id', 'favorite_count' and 'retweet_count' columns

Assessing Data After all three datasets have been gathered, they were assessed visually using python df.head() function and also scrolling through it in excel and programmatically for quality and tidiness issues using functions like df.describe, df.info, df.column_name.count_count(), etc to detect and document some quality and tidiness issues. More than eight (8) quality issues and two (2) tidiness issues was detected and documented

Cleaning Data Some of the issues documented were addressed and cleaned here. Some of the issues address include

1. timestamp column is in the object dtype instead of datetime
2. The name column contains some entries that are not names in the the real sense. E.g None, a, an, etc. Meanwhile, the real names has a pattern of Proper nouns. That is, their initial letter is being capitalised. It's better to replace them with 'None'
3. the column names p1, p2, p3, p1_conf, p2_conf, p3_conf, p1_dog, p2_dog, and p3_dog are confusing as they are not descriptive enough to be understood. Hence let's make them descriptive by renaming them.
4. id column in all the datasets have integer dtype intstead of string
5. The source column in twitter_archive_enhanced is ambiguous as it's values are being embedded into an html tag. this needs to be cleaned
6. the tweet id column name of tweet_json is different from it's corresponding name in the other two datasets, it's 'id' in tweet_json whereas 'tweet_id' in other two. This need to be changed to enhance consistency in column names across all dataset.
7. The four dog stages are all in the object dtype instead of strings even after they've been combined
8. The p1, p2, p3, columns in image prrediction dataset cuntains underscores and some valuesd starts with capital letter
9. There are about 181 retweets and since as instructed retweets are not needed, they need to be removed

The complete steps can be seen in the wrangle_act.ipynb notebook using this https://viewf6b31853.udacity-student-workspaces.com/notebooks/wrangle_act.ipynb

0.1.3 Analizing and visualization

After analysis was done, the following sights were given

1. For a tweet to be retweeted, there's a very high tendency that it has to be first liked.

2. Most retweets and likes occurs in the month of July and in year 2017.

3. The modal source is ’Twitter for iPhone’