wrangle_report

September 7, 2022

0.1 Reporting: wragle_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.

In this report I outline the wrangling efforts to collect the data required for analysis of the WeRateDogs Twitter Archive.

0.2 Data Gathering:

- Data was gathered from 3 different sources:

  1. WeRateDogs Twitter Enhanced archive, programmatically downloaded from the Udacity servers
  2. The image predictions file, programmatically downloaded from the Udacity servers.
  3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library. The tweet_id, favorite_count, retweet_count and followers_count were extracted programmatically.

0.3 Assessment & Cleaning:

- After visual inspection and checking the data of the three files programmatically using Jupyter Notebook and pandas functions, I identified several quality issues and tidiness as follows:

Quality:

1. timestamp is not in correctly applied to columns in the twitter_archive dataframe.

2. Some of the tweet_id's id not present in the twitter_api_data dataframe meaning some did not pull through when the data was downloaded from twitter, so the data will not be complete

3. In the image_predictions dataframe, the dog names are not a consistent case it needs to be be either converted to lowercase, uppercase or title also Replace underscores with whitespace for readability

4. Column names p1_conf, p2_conf and p3_conf and img_num is not very descriptive, even jpg_url can be changed, since pictures can be in different format and extensions

5. in_reply_to_status_id and in_reply_to_user_id can be removed from twitter_archive since it only has 78 rows of none null values, also retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, since it won't be needed for analysis

6. twitter_archive names column contains names that is either typos, e.g 'None', 'a', 'an', 'such', 'the', will be replaced by NaN

7. expanded_url in twitter_archive has 59 empty values and the empty rows can be dropped, since the requirement says that all empty image cells can be dropped

8. tweet_id will be converted to a string since the length of the tweet_id is larger then the ranges for integers

Tidiness Issues

1. The columns, doggo-floofer-pupper-puppo, should be one column as they are just different stages to identify a dog

2. Column names p1_conf, p2_conf and p3_conf and img_num is not very descriptive, even jpg_url can be changed, since pictures can be in different format and extensions

0.4 Data Cleaning:

- For each quality and tidiness issue, the work flow was

1. Define the cleaning steps.
2. Write code to do cleaning.
3. Test the results.
4. I have documented all appropriate steps in the cleaning process

- After cleaning and combinig all the datasets, the data was saved in a .CSV file

```
In [ ]:
```