wrangle_report

June 28, 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.

0.1.1 Wrangle Report

This document briefly describes the wrangling efforts which i carried out for the WeRateDogs Twitter dataset in the wrangle_act.ipynb notebook.

0.1.2 Gathering Data

These were the steps I followed to gather the data

- twitter_archive_enhanced.csv - This file was given buy Udacity
- image_predictions.tsv - I used the request library to download this file
- tweet_json.txt - Unfortunatley my developer account was not approved by twitter and this caused a setback for me so i had to get the data from other source

0.1.3 Assessing Data

The following methods were used to assess the data programmatically

- .info()
- .head()
- .value_counts()
- .query()
- .isna()
- .duplicated()

0.1.4 Quality Issues

For df_archive

1. Some dog names are invalid and duplicated in the `df_archive` table
2. From the assesing data objectives it was stated that retweets are not needed, hence, we'll have to drop the retweeted_statuses also.

3. A lot of missing data in the reply columns in the `df_archive` table are also not needed so we have to drop them
4. The datatype for `timestamp` in the df_archive data is object instead of datetime
5. Removing the anchor link and retaining only the text for source
6. Invalid tweet_id data type (integer instead of string)
7. Some rating denominators are less than 10

for df_images

1. Invalid tweet_id data type (integer instead of string)
2. Upper case and lower case name

for df_tweet

1. id instead of tweet_id

0.1.5 Tidiness issues

1. URLs in some of the text in the text column
2. Breed and predictions to individual columns
3. Merging all the data to one

0.1.6 Cleaning Data

0.1.7 Issue #1:

- Some dog names are invalid and duplicated in the `df_archive` table

Define:

- Rename them to none

0.1.8 Issue #2:

- From the assesing data objectives it was stated that retweets are not needed

Define

- Hence, we'll have to drop the retweeted_statuses also.

0.1.9 Issue #3:

- A lot of missing data in the reply columns in the `df_archive` table

Define

- They're not needed so we have to drop them, but first we drop the row which they reference before we drop the column

0.1.10 Issue #4:

- The datatype for `timestamp` in the df_archive data is object instead of datetime

Define

- Convert to datetime

0.1.11 Issue #5:

- Anchor link in the text

Define:

- Removing the anchor link and retaining only the text for `source`

0.1.12 Issue #6:

- Invalid tweet_id data type (integer instead of string)

Define:

- Change int to str

0.1.13 Issue #1 (For df_images):

- Invalid tweet_id data type (integer instead of string)

Define:

- Change the datatype to string from int

0.1.14 Issue #2:

- Upper case and lower case names in the p1, p2, p3, columns

Define:

- convert all the names to Upper cases

0.1.15 Issue #3:

- Breed and Predictions to individual columns

Define:

- combine the Breed and Predictions to individual columns

0.1.16 Issue #1(df_tweet):

- id instead of tweet_id and datatype

Define:

- Rename id to tweet_id for joining purpose and change datatype

0.2 Tidiness Issues:

0.2.1 Issue #1:

- Dog stages are in multiple columns

Define:

- Put dog stages into one column

0.2.2 Issue #2:

- URLs in some of the text in the text column

Define:

- Remove the URLs in the text

0.2.3 Issue #3:

- Scattered data

Data:

- Merging all the data to one

```
In [ ]:
```