wrangle_report

September 21, 2022

0.1 Reporting: wragle_report

This project is based on data wrangling and expalained below are the steps and process followed to wrangle the We Rate Dogs data from twitter.

1. Data Gathering I collected data from three different sources

```
a. from a flat file `twitter_archive_enhavced.csv`
b. from a udacity database `image_predictions.tsv`
c. from twitter API
```

2. Data Accessing The first two datab were pretty much ready for assessment, but the twitter API data was not because it was a in a JSON format. So, I prepared the twitter data for assessment by first dumping it JSON format file in a txt file which is then readable by pandas.

All the three data were accessed both visually and programatically and some quality and tidiness issues were spotted

`Quality Issues`

1. Column header (p1, p1_config, p1_dog, p2, p2_config, p2_dog, p3, p3_config, p3_dog) not descriptive enough.
2. Invalid data representation for percentage in p1_config, p2_config, p3_config
3. 745 None values as dog name
4. Some dog names are not feasibly correct like 'a', 'an'.
5. Text column contains url apart from the text column
6. Time stamp is contians string instead of date time format
7. Rating denominator can't be zero
8. Nondescriptive column header "id"

`Tidiness issues`

1. A single observation unit is stored in multiple column (doggo, floofer, pupper, puppo) shoild all be in dog stages.
2. Delete rows that are not original tweets

3. Data Cleaning In this section, all the issues in the assessment were adrdresses and resolved using the Define-Code-Test approach. Some of the tools I used in the section are

1. Pandas Dataframe rename column
2. Pandas Column mapping and formating
3. Pandas Split string into columns and drop columns
4. Pandas merge
5. Pandas to_datetime
6. Pandas subseting a dataframe
7. Adding columns together to make a string and manipulating the string