wrangle_report

September 7, 2022

0.1 Reporting: wragle_report

The dataset that I wrangled, analyzed, and visualized is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs, WeRateDogs is a Twitter account that rates people's dogs with over 4 million followers and has received international media coverage.

-To wrangle this dataset I follow the below process:

1- Gathering Data: In this step,There are three pieces of data, the first one ('Twitter_archive_enhanced.csv') I downloaded this file manually and read it to be ready for analysis, the second data file is ('image_predictions.tsv') I downloaded this file programmatically using the following URL:'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv', and read it to be ready for analysis, finally, the last file ('tweet_json.txt') I downloaded programmatically using Twitter Developer account (using the key and token to get Twitter API) and read it to be ready for analysis.

2- Assessing Data: After gathering all three pieces of data, assess them visually (displayed and looking into issues) and programmatically (using pandas functions and methods are used to assess the data) for quality and tidiness issues by detecting and documenting nine (10) quality issues and this is not all quality issues and two (2) tidiness issues as below:

Quality issues

1. Delete (181) retweets and (78) replies values.
2. Change timeestamp dtype to be datetime.
3. Correct invalide names in 'name' column like('a', 'th', 'an',...) and fix the format to uppercase letter.
4. Ratings of numerator and denominator containts outliers values.
5. Clean 'source' column to be readable easily.
6. Drop duplicated image in jpg_url column.
7. Change tweet_id dtype to be int.
8. Drop unnecessary columns (in_reply,..,retweet_status, ..., expanded, text,..).
9. Marge columns p1,p2,p3 to one column and p1_conf,.. to be one column as well.

10. Remove the tags ’_’ form dogs type and fix the format to uppercase letter.

Tidiness issues

1. Marge the dog stage from 4 columns (doggo, floofer, pupper, puppo) to be one column.
2. Marge the 3 dataframes to be one dataset.

3- Cleaning Data: After assessing all the data, it’s time to clean all of the issues I detected and documented with write clear code, and code testing to confirm what we did?, after I made a copy of each piece of a data file from the original data file.

-After that previous process I stored all data files to be one main master dataset to be ready to the next step of the project (Analyze (using the functions to ask some questions related the dataset analysis to provide answers that will explore the dataset) and Visualize (using python functions to display the charts to more expline and clear the dataset by visual)).

```
In [ ]:
```