wrangle_report

WRANGLING DATA REPORT

DATA GATHERING:

In this project, all three pieces of data for the project were gathered differently and loaded them in the notebook using the pandas dataframe

The files were:

1. Twitter archive file: twitter_archive_enhanced.csv was manually downloaded and loaded.

2. The tweet image predictions: the dog breed prediction present in each tweet. The file (image_predictions.tsv) is available on Udacity's servers according to a neural network and downloaded programmatically using the Requests library and the url link was also available.

3. Twitter API & JSON: To gather each tweet's id, retweet count and favorite ("like") count at minimum, and any additional interesting data. Utilizing the tweet IDs in the WeRateDogs Twitter archive,to query the Twitter API for each tweet's JSON data using Python's Tweepy library. Each tweet's JSON data was written in its own line and then read as .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Please note that I couldn't get a Twitter developer account, therefore, I accessed the project data without a twitter account using the tweet_json.txt file provided and appending only the id, retweet count and favorite columns.

NOTE: All required libraries were imported before gathering the datasets and loading them into the dataframe.

Assessing Data:

Did a visual and programmatic assessment for the 3 dataset using the assessment methods in pandas and listed 8 quality issues and 2 tidyness issues to be cleaned.

Quality issues were:

Twitter archive dataset issue:

1. Wrong datatype assigned to the timestamp column, supposed to be datetime and not int64

2. Tweet_id column for the 3 datasets has a wrong dataset 3. text column should be renamed to be more descriptive. 1

4. remove unnecessary columns in the dataset

5. some rating denominator are not equal to 10

Image prediction dataset issues

6. some image predictions are not dogs which has to be removed from the dataset

7. some dog name are missing or wierd (like ’a’ and ’the’ which are misspelt), usually those starting with lowercases.

8. change column names such as p1, p2, p3 to be more descriptive

9. the prediction dog breeds are starting with both upper and lower case which is not consistent

Tidiness issues

1. The three dataset should be merged to one.

2. create a single column for all dog names in the twitter archive dataset

Cleaning Data:

- I made a copy of the original data before cleaning.
- Cleaned all of the issues I documented while assessing.
- During cleaning, used the define-code-test framework.
- The result was a high-quality and tidy master pandas DataFrame.

After assessing and cleaning datasets, they were merged to become 1 dataset. Then stored as a csv file called ‘twitter_archive_master.csv’ which was used to make some meaningful insights and visualizations.