5/9/2022

Wrangle_report

Dihia Mezghiche

Project: Wrangling

In this project we will be studying the tweets of the twitter account @dog_rates.

My wrangling efforts are as follows:

Data Gathering

In the Gathering of my data, I used three different methods to get my data as asked during the project.

1- I downloaded the data that is in twitter_archive_enhanced.csv using the basic pd.read_csv function that gives us the possibility to read csv files.
2- I downloaded the image prediction tsv file this time I used the second method learn from this course which is the use of the request library that helps us download data directly from an url.
3- I tried to use the tweeter api as asked but not having a response from tweeter concerning my request to signup of a tweeter developer account, I had to directly use the tweet_json.txt file that gave us access to. To get the data from this file I read eachline and loeaded using json.loads and got the id of the tweet, favorite count and retweet count.

Assesing Data

To highlight and detect the problems of this dataset I had to make the following efforts:

1- Visual Assessment:
Through the visual evaluation I noticed the following issues:
Archive dataset

- Missing values in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id et retweeted_status_timestamp

- The age stage of the dogs are in 4 different columns
- Wrong values in the name column: dog named 'the'

Prediction dataset

- From a visual observation the dataset looks fine except the columns should have more understandable name.

2- Programmatic Assessment:
The few issues that I noticed through the programmatic assessment are as follows:

Archive dataset

- in_reply_to_status_id and in_reply_to_user_id type float instead of int.
- timestamp type string instead of datatime
- Some tweets are retweets so they could be considered as duplicates
- Missing values in expanded_urls
- Some dogs have denominator rating above and under 10

Prediction dataset

- Duplicated images(jpg_url) of dogs that have been rated, we can see that in the jpg_url column some photos that have been used to predict the breed of the dog have been used multiple times.

Json dataset

- Some tweets have zero favorite count and zero retweet.
- Some tweets have a huge number of retweets like 19297 and zero favorite count that means not all tweets that have zero favorite counts are clean.

Lastly the last issue that i detected is the there are 3 datasets without the need of having them separated so merging the 3 datasets will be a necessity

Cleaning Data

Most of the efforts that I made during the project are here where I had to solve every single one of the issues that I have detected:

Issue 1: timestamp attribute is of type string

Solution: convert the type using to_datetime function.

Issue 2: Some tweets are retweets

Solution: Delete the tweets using the drop function

Issue 3: rating_denominator above and under 10

Solution: Retrieve the id of the the tweets and delete.

Issue 4: Wrong values in the name column

Solution: Replace the wrong name by none.

Issue 5: Missing values in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id

Solution: Fill the NaN values with zero.

Issue 6: in_reply_to_status_id and in_reply_to_user_id type float instead of int.

Solution: Convert using astype()

Issue 7: each type of dog is a column

Solution: Concat 4 columns

Issue 8: duplicated images of dogs that have been rated

Solution: Drop duplicates

Issue 9: 3 datasets

Solution: Merge the three datasets

Issue 10: Rename columns

Solution: Rename columns to have significate names