

REPORT

ON

Wrangle 'WeRateDogs' Twitter Data Project by Adeniran Kayode

Data Wrangling Steps:

1. Gathering
2. Accessing
3. Cleaning

Project Details

Our tasks in this project are as follows:

-
- Data wrangling, which consists of:
- Gathering data.
- Assessing data
- Cleaning data
- Storing, analyzing, and visualizing our wrangled data
- Reporting on
- 1) our data wrangling efforts and
- 2) our data analyses and visualizations

Gathering Data for this Project

1. The WeRateDogs Twitter archive is given this file to us. I Downloaded this file manually by clicking the following link: `twitter_archive_enhanced.csv`
2. The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is presented in each tweet according to a neural network. This file (`image_predictions.tsv`) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: `https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv`
3. Each tweet's retweet count and favorite ("like") count at minimum. Using the tweet IDs in the WeRateDogs Twitter archive, I query the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called `tweet_json.txt` file.
4. I was not given access by twitter, so I just followed through the instruction guide to solve this issue and called the `tweet_json.txt` file which already have the twitter api details.

Assessing Data for this Project

These are the following Quality and Tidiness Issues what I found.

Quality issues.

1. removing retweets rows.
2. doggo, floofer, pupper, puppo can also be dog type or description i think
3. The `tweet_id` should be string not integers
4. Some tweets have more than 2 ratings
5. Timestamp is not of a date format
6. The `img_num` column should be a string
7. Correct columns contain some invalid name
8. Wrong Datatype Name Column should be in string

9. Tidiness Issue
10. remove columns with too many missing values. (retweeted_status_user_id, retweeted_status_id, retweeted_status_timestamp, in_reply_to_user_id, in_reply_to_status_id)
11. Merge the dataframe twitter_archive, dataframe image_predictions, and tweet_json dataframes.
12. doggo, floofer, pupper, puppo these 4 variables should be combined into one categorical variable Dog Type.

Cleaning Data for this Project

Tidiness Issue

13. Merge the dataframe twitter_archive, dataframe image_predictions, and tweet_json dataframes.

Define

I used pandas merge functions to join our datasets

Test

14. Confirmed if the three datasets have been merged.

Define

Remove columns with too many missing values.

we can delete the column using Command DROP, for column Axis =1

Code

Test

for testing we can display our dataset using. HEAD() method

doggo, floofer, pupper, puppo these 4 variables should be combined into one categorical variable Dog Type.

Define

- we can extract the data of the column using .EXTRACT() method

Code

Test

- we can check the values of our newly created column

Quality Issue

- remove doggo, floofer, pupper, puppo

Define

- we can delete any column using .DROP() Method

Code

Test

- check it by displaying the dataset
- Wrong Datatype img_num Column should be in string

Define

- we can convert the datatype using .ASTYPE() method

Code

Test

- we can check the datatype using type method
- Change tweet_id from an integer to a string.

Define

- we can convert the datatype using .ASTYPE() method

Code

Test

- we can check the datatype using type method
- Wrong Datatype Source Column should be in Category

Define

- we can convert the datatype using .ASTYPE() method

Code

Test

- we can check the datatype using type method
- Timestamp is not of datetime format .

Define

- for datetime format

Code

Test

- we can display some dataset for the testing purpose
- Correct name Column it contain some invalid name

Define

- we can replace incorrect names with None using .REPLACE() Method

Code

Test

- we can check if any row with name equals to these is present or not
- Delete retweets.

Define

- we can delete any column using .DROP() method

Code

Test

- we can check if any column named retweeted_status_id is present or not
- Checking for duplicate values and deleting them

Define

- we can check duplicate value using .DUPLICATED method

Code

Test

- if any value present

Storing Data for this Project

Storing our Cleaned data into CSV File

- `df2.to_csv('twitter_archive_master.csv', encoding='utf-8')`

