Report Project

“Wrangle and Analyze Data”

By Wael Al harbi

Udacity Data Analysis Nanodegree

Friday, January 22, 2021

Wrangling “WeRateDogs” Twitter Data

Wangle and Analyze Data Project include three part :

1. ***Gathering Data***
2. ***Assessing Data***
3. ***Cleaning Data***

1. *Gathering Data:*

I Downloaded File Directly From The Resources Tab, And I Apologize   For Not Downloading Them Programmatically Due To Lack Of Time.

This Part Of The Project Includes Collecting The Files Required To Complete The Project, Which Are Three Files:

a. *twitter_archive_enhanced.csv*

THIS FILE IS READ NORMALLY AS WE ARE ACCUSTOMED TO VIA A FUNCTION “READ_CSV” , USING THE PANDAS LIBRARY.

b. *image_predictions.tsv*

I Downloaded the image prediction file using the link provided https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv.

And read it via a function “read_csv , separate by tab "\t"

c. tweet_json.txt

I downloaded file directly from the resources tab, and I apologize for not downloading them programmatically due to lack of time.

And read it via a function “read_json.

2- *Assessing Data*

Data Assessing summary :

Function use in code Assessing :

info()
duplicated()
sum()
sort_values()
head()
tail()
sample()
value_counts()
isnull()
rename()

❖ *Quality Issues :*

a. twitter_archive table

- *Datatype of ‘tweet_id and ‘timestemp*
- *Retweet duplicate actual tweet*
- *Format of ‘ Source’*
- *rating_numerator and ‘rating_denominator’ not all of them are correct*
- *expanded_urls’ has missing value (tweet without URL)*
- *Datatype of dog stage*
- *There is unneeded Column*

b. df_image_predictions table

- *jpg_url has missing value (tweet without image)*
- *p1 , p2 and p3 not all of them lowercase*

c. tweet table

- *Datatype of 'tweet_id*

❖ *Tidiness Issues*

- *Merge dataframes together*
- *Create new column for dogs stage*

3- *Cleaning data*

Data cleaning summary :

This section consists of the cleaning portion of the data wrangling process

include three part :

1. *Define*
2. *Code*
3. *Test*

Define:

- *Merge the three dataframe.*
- *Delate retweet in 'retweeted_status_id'.*
- *Drop unneeded columns.*
- *Delate tweet without URL (remove null value in 'expanded_urls').*
- *Change datatype of 'tweet_id' to String*
- *Change datatype of 'timestamp' to datetime.*
- *Delate tweet without image (remove null value in 'jpg_url').*
- *Create new column named dog_type , And fill the null values with None.*

- *Change datatype of 'dog_type to category*
- *convert p1,p2 and p3 to lowercase.*
- *Correct wrong values in rating numerator , rating denominator.*

Code:

Function use in code part :

Merge()

Isnan()

Drop()

Notnull()

Astype()

to_datetime()

str.extract()

str.lower()

replace()