UDACITY

Project5: Wrangle and Analyze Data

Introduction:

This project involves wrangling data from three different sources, all of which are connected to the popular WeRateDogs (@dog rates) Twitter account. WeRateDogs is a Twitter account that posts photos of dogs sent by their owners, along with a humorous caption and a rating that nearly always surpasses 10/10.

needed packages :

The following packages (libraries) need to be installed. pandas

- NumPy
- requests
- tweepy
- json
- matplotlib

Gathering Data:

This project encompass three dataset:

- Twitter archive (csv file)
- Image predictions for dogs(tsv file).
- Twitter API .

Assess:

Quality:

df_archive:

1. Change the timestamp's datatype from str to datetime.
2. Convert the rating numerator and rating denominator datatypes to float.
3. remove columns with too many missing values such as in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'
4. some dog's name are weird , replace the unclear dog name with nan
5. Remove all rating denominator values below 10
6. Make the source column's content more readable by cleaning it up.
7. Delete retweets

df_image:

1. change the Datatype img_num Column to string

df_tweets_API:

1. convert retweets, and favorites to int datatyp and convert tweet_id,to str datatype

Tidiness:

1. Make a master data set out of three distinct dataframes.(df_archive, df_image, df_tweets_API)
2. Create one column for the various dog types: doggo, floofer, pupper, puppo then remove the columns since there's no need for it

The structured as following:

1. Make a master data set out of three distinct dataframes.(df_archive, df_image, df_tweets_API)
2. Change the timestamp's datatype from str to datetime
3. Convert the rating numerator and rating denominator datatypes to float.
4. change the Datatype img_num Column to string
5. convert retweets, and favorites to int datatyp and convert tweet_id,to str datatype
6. remove columns with too many missing values such as in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'
7. some dog’s name are weird , replace the unclear dog name with nan
8. Create one column for the various dog types: doggo, floofer, pupper, puppo then remove the columns since there’s no need for it
9. Remove all rating denominator values below 10
10. Make the source column's content more readable by cleaning it up.
11. Delete retweets

sources

- http://ocrpsychology2015.blogspot.com/2015/06/
- https://video.udacity-data.com/topher/2018/November/5be5fb4c_twitter-api/twitter-api.py
- https://towardsdatascience.com/twitter-analytics-weratedogs-a441be7d4a85
- https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/twitter-data-in-python/