DAND - Wrangle and Analyze Data Project

Name of the student : Yousef Al-mutairi

introduction :

Hello , I'm Yousef almutairi , and i will be working on this project required for ND certification from udacity program ,

i will try my best!,

I applied for Twitter Developer Account , but unfortunately my application was not approved

1 - as for the data gathering i downloaded tha files from udacity project home page ,

- tweet_json.txt
- image_predictions.tsv

Data gathering

- Reading the file provided from udacity program project page
- Save it to a data frame.
- Downloading the image-predections.tsv from the URL provided from udacity program project page.

Data Assessing

Accuracy & quality :

Twitter_Arc DataFrame :

- Missing values at most of the cells in ('in_reply_to_status_id', 'in_reply_to_user_id') columns Are 78 instead of 2356
- Missing values at most of the cells in ('retweeted_status_id', 'retweeted_status_user_id' , 'retweeted_status_timestamp') columns Are 181 instead of 2356
- We are only going to use ORIGINAL tweets of dog rating , we will not use nethier retweets or reply tweet as an original tweet .

DAND - Wrangle and Analyze Data Project

- We will use only ORIGINAL tweets of dog rating that held images , we will not use tweets without images .
- wrong datatypes ('timestamp' , 'retweeted_status_timestamp')
- we will change 'tweet_id' datatype to Object for a better quality .
- 'rating_denominator' column values should be of 10 .
- Inaccurate values in 'rating_numerator' colmun , some ratings are not out of 10 - rating_denominator .
- missvalues of the nulls in columns ('doggo','pupper','floofer','puppo') nulls are presented as "None" .
- Content of ('source') column is too long thus its can not be analyzed easily .
- Irregular and illogical values in 'name' column .
- some values in 'name' column are in LOWERCASE .
- missvalues of the nulls in column ('name') nulls are presented as "None".
- underscriptive label for the 'name' column.
- 'rating_numerator' for the 'tweet_id' = 786709082849828864 equals 9.75 instead of 75 .

Img_Pred

- "True" and "False" values in the (`p1_dog`, `p2_dog`, `p3_dog`) columns are not handy.
- Undescriptive columns' labels (`p1`,`p2`,`p3`,`p1_conf`,`p2_conf`,`p3_conf`,`p1_dog`,`p2_dog`,`p3_dog`).
- some images are not dog images , and we are only intrested in dogs rating .
- changing datatype of 'tweet_id' to Object for a better quality.

tweets_json

- 'tweet_id' as a label instead of 'id'.
- Changing 'id' datatype to Object for a better quality .

Tidiness

- (`doggo`,`floofer`,`pupper`,`puppo`) columns values should be represented in one column called 'dog_stage' with a 'category' datatype.
- combine our main dataset with 'tweet_json' and 'img_predict' tables.
- 'rating_numerator' and 'rating_denominator' colmuns should be in one main column which will be called 'rating_out_of_10'.

Resources I've used in order to complete this project

- https://stackoverflow.com/questions/10665889/how-to-take-column-slices-of-dataframe-in-pandas

DAND - **Wrangle and Analyze Data Project**

- https://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.Timestamp.html\
- https://stackoverflow.com/questions/11346283/renaming-columns-in-pandas
- http://pbpython.com/pandas_dtypes.html
- https://www.akc.org/dog-breeds/?letter=O
- https://stackoverflow.com/questions/25125168/array-shape-giving-error-tuple-not-callable
- https://www.geeksforgeeks.org/python-pandas-dataframe-rename/
- https://stackoverflow.com/questions/38101009/changing-multiple-column-names-but-not-all-of-them-pandas-python
- https://stackoverflow.com/questions/29960733/how-to-convert-true-false-values-in-dataframe-as-1-for-true-and-0-for-false
- https://kaijento.github.io/2017/04/22/pandas-create-new-column-sum/
- https://stackoverflow.com/questions/50847374/convert-multiple-columns-to-string-in-pandas-dataframe

and finally I've helped my self learning this amazing project with ideas and resources of this prev project .

https://github.com/Najlaa-Shariefi/Data-Wrangling-of-WeRateDogs-

big thanks to you Najlaa .