WeRateDogs

Human most beloved friend

Name: Mark George Louiz

Gathering Data:

1: Download the twitter-archive-enhanced manually then opened in the notebook using pandas method read_csv .

2: Download image-predictions.tsv programmatically

using requests package .

3: tweets_jason.txt with read line by line constructed using twitter api .

Assessing Data and cleaning:

A: Quality issues:

First : twitter archive enhanced data set :

1: timestampe wrong data type (object type) .

Solution: converting to Datetime type using pd.to_datetime method .

2: doggo,floofer,pupperand puppo columns wrong description Nan value as None .

Solution: converting None values to Nan using np.nan .

3: name row wrong names as a , an etc .

Solution : open text and search about right names manually and correct it programmatically for the images with names and the rest convert it to nan values using np.nan method

4: missing value in expanded_urls column .

Solution : drop the missing values from the datasets.

5: rating_denominator wrong values for dogs images.

Solution : searching about the correct values in the text of tweet manuely and correct it .

6: retweets and replies in data set .

Solution : one of the requirement of data analysis according to the rubric is to use only tweets in analysis and visualization so we drop them using comparison with image prediction data set using tweet id column .

7 : tweet id int type

Solution : knowing that we won’t do any mathematical operations on tweet id column so according to data quality it’s better to change its type to object (string) type using astype method in pandas .

Second : image prediction data set :

1 : small letters and capital letters in p1,p2,p3 columns .

Solution : to make sure that every value in the data column is consistent we capitalize every value using capitalize method

2 : tweet id int type

Solution : knowing that we won’t do any mathematical operations on tweet id column so according to data quality it’s better to change its type to object (string) type using astype method in pandas .

Third : twitter Api query data set :

1: tweet id int type

Solution : knowing that we won’t do any mathematical operations on tweet id column so according to data quality it’s better to change its type to object (string) type using astype method in pandas .

B: tidiness issues :

First : twitter archive enhanced data set :

1 : doggo , floofer , pupper and puppo are values not variables .

Solution : according to tidy data rules(structure data) is every variables values exist in one column so using concatenate operator (+) into new column .

2 : the columns 'retweeted_status_id' 'retweeted_status_user_id' 'retweeted_status_timestamp','in_reply_to_status_id'

and 'in_reply_to_user_id' .

Solution : After we got rid of retweets and replies there is no need for these columns so we drop it using drop method .

Second : twitter Api query data set :

1: has no new observation so it shouldn't be a table .

Solution : according to tidy data rules every observation in one table so we merge twitter Api query data set with twitter archive data set based on tweet id .

Conclusion :

We get two master data sets :

1: archive master data set with 1971 entries .

2: image prediction clean with 1971 entries .