wrangle_report

February 25, 2019

1 dataset

in this project we used WeRateDogsdataset which is a Twitter account that rates people's dogs with a humorous comment about the dog.

2 gathering

we gather from 3 different resource 1- csv file 2- by url 3- by twitter API

3 problems

3.0.1 Quality

1- many coulums are empty in twitter-archive-enhanced 2- tweet_id must be string 3- timestamp must be a timestamp not object 4- some retweeted_status_user_id are duplicated 5- some tweet_id in twitter-archive-enhanced are missingin image-predictions 6- numerator and denominator has incorrect value 7- the number of df1 diffrent than the number of df2 8- dog type must be category 9- some dogs name are not valid rename it to None 10-a> html tages in source column

3.0.2 Tidiness

1- doggo, floofer, pupper and puppo,should be merged into one (new) column 2- timestamp has date and time

4 Solving problems

to solve the problem I have use 3 technique : 1- by using properties of dataframe 2- by using pandas 3- by using regex

5 clean data set

the new or (clean dataset) has this attributes: tweet_id source text expanded_urls rating_numerator rating_denominator name DogType date time and has 3 types of data: 1- category(1) 2- int64(2) 3- object(7)

5.0.1 Visualization

i have Visualize the data using basic matplotlib

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```