act_report

introduction

I think this data after clean , explain to me that data have known category However, it has some interesting things , and the rating of dogs lose seriousness in data like (13/10,11/10) , they love dogs :)

Grather Data

I use three dataset (twitter-archive-enhanced.csv , image-predictions.tsv , josn file from tweepy from Twitter API)

- twitter-archive-enhanced.csv this file is i was read it Manually by panadas
- image-predictions.tsv this file i was read it by requsts libray i get it from link
- josn file I was download it from twitter API from My account on Twitter developer and get File with information about tweets from this page , we cover it in Assessing data

Assessing data

twitter-archive-enhanced.csv

it has information about tweet and dog stage (doggo,floofer,pupper,puppo) and text of tweets

```
In [4]: twitter_archive.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                    2356 non-null int64
in_reply_to_status_id       78 non-null float64
in_reply_to_user_id         78 non-null float64
timestamp                   2356 non-null object
source                      2356 non-null object
text                        2356 non-null object
retweeted_status_id         181 non-null float64
retweeted_status_user_id    181 non-null float64
retweeted_status_timestamp  181 non-null object
expanded_urls               2297 non-null object
rating_numerator            2356 non-null int64
rating_denominator          2356 non-null int64
name                        2356 non-null object
doggo                       2356 non-null object
floofer                     2356 non-null object
pupper                      2356 non-null object
puppo                       2356 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

image-predictions.tsv

this data set it about prediction of images of dogs and dog bread of three column p1 , p2 , p3 So for the last row in that table:

```
jupyter wrangle_act Last Checkpoint: 22/12/2018 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run Code

In [7]: image_prediction.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null int64
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

- tweet_id is the last part of the tweet URL after "status/" → https://twitter.com/dog_rates/status/889531135344209921
- p1 is the algorithm's #1 prediction for the image in the tweet → golden retriever
- p1_conf is how confident the algorithm is in its #1 prediction → 95%
- p1_dog is whether or not the #1 prediction is a breed of dog → TRUE
- p2 is the algorithm's second most likely prediction → Labrador retriever
- p2_conf is how confident the algorithm is in its #2 prediction → 1%
- p2_dog is whether or not the #2 prediction is a breed of dog → TRUE etc.

josn file

this file As I explained i was download it from twitter API it information tweets but clear than twitter-archive-enhanced.csv

assessing (tweet_json)

In [26]:
```
tweet_json.head()
```

Out[26]:

	tweet_id	favorite_count	retweet_count	followers_count	friends_count	source	retweeted_status	url
0	892420643555336193	34986	7341	8982625	16	Twitter for iPhone	Original tweet	https://t.co/MgUWQ76dJU
1	892177421306343426	30294	5478	8982625	16	Twitter for iPhone	Original tweet	https://t.co/aQFSeaCu9L
2	891815181378084864	22789	3621	8982625	16	Twitter for iPhone	Original tweet	https://t.co/r0YlrsGCgy
3	891689557279858688	38251	7530	8982625	16	Twitter for iPhone	Original tweet	https://t.co/tD36da7qLQ
4	891327558926688256	36532	8110	8982625	16	Twitter for iPhone	Original tweet	https://t.co/0g0KMIVXZ3

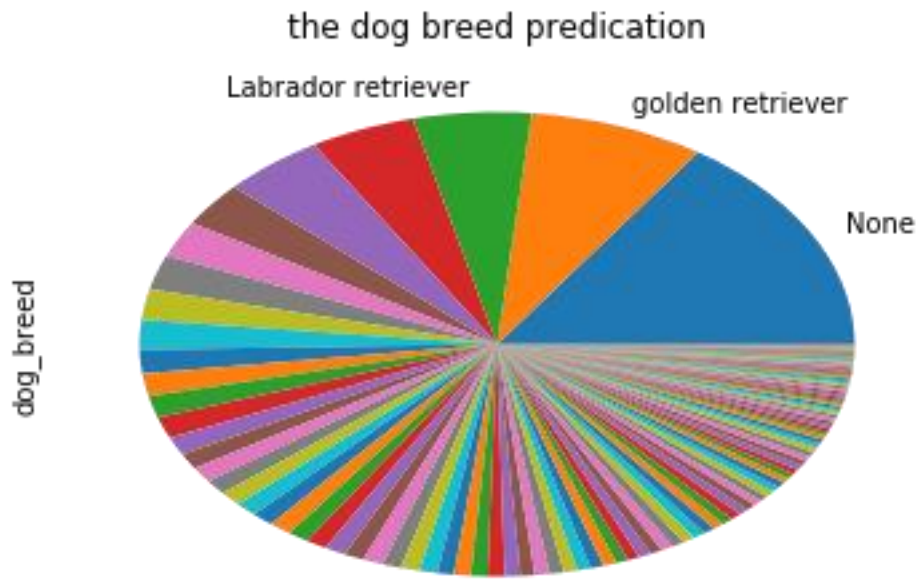
Clean Data

I was try to clean this by useing the issues (Quality,Tidiness) which found it from Assessing data like Create one column for dog stage : (doggo, floofer, pupper, puppo) it Tidiness issue and url has some invalid data like (0 , u , e , y , n , t , elc) it this Quality and At the end I collect the three data sets in one data set

Analyzing & Visualizing Data

the most common dog

when analying the data It is clear that there are many types of dogs, but I have found that the Golden Retriever is more common



the distribution of Data

Also, I found the distribution of the number of tweets and likes in one direction