WeRateDogs®

Wrangling and analyzing data

Abstract:

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson, and has received international media attention both for its popularity and for the attention drawn to social media copyright law when it was suspended by Twitter for breaking these aforementioned laws.

Mohamed Shaaban
Mohamed.shaaban89@gmail.com

Introduction:

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user dog rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. In this project I had applied the data analysis from:

1. **Data gathering**
2. **Data assessing**
3. **Data cleaning**
4. **Data storing**
5. **Asking questions and provide insights/visualization**

“In the below section I will describe my effort through each step one by one”

1. Data gathering:

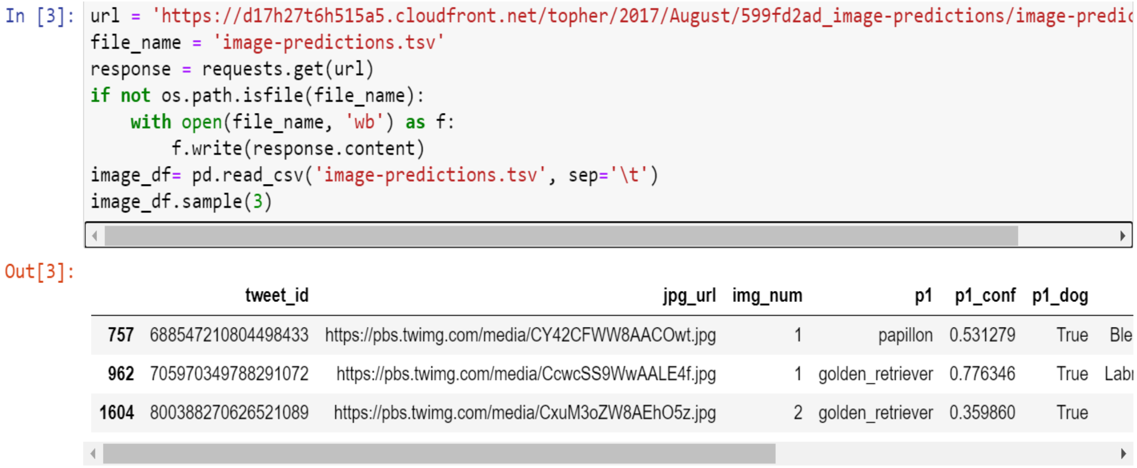
In this step I had extracted data sets from 3 different sources which was a (.csv file, .tsv file and JSON text). And I can go through each file and the line of code I applied:

- **Enhanced Twitter Archive:**

This file had been downloaded manually from Udacity classroom, no need for code to do so.

- **Image Predictions File:**

I used the provided URL in the project details and by requests



In [3]:

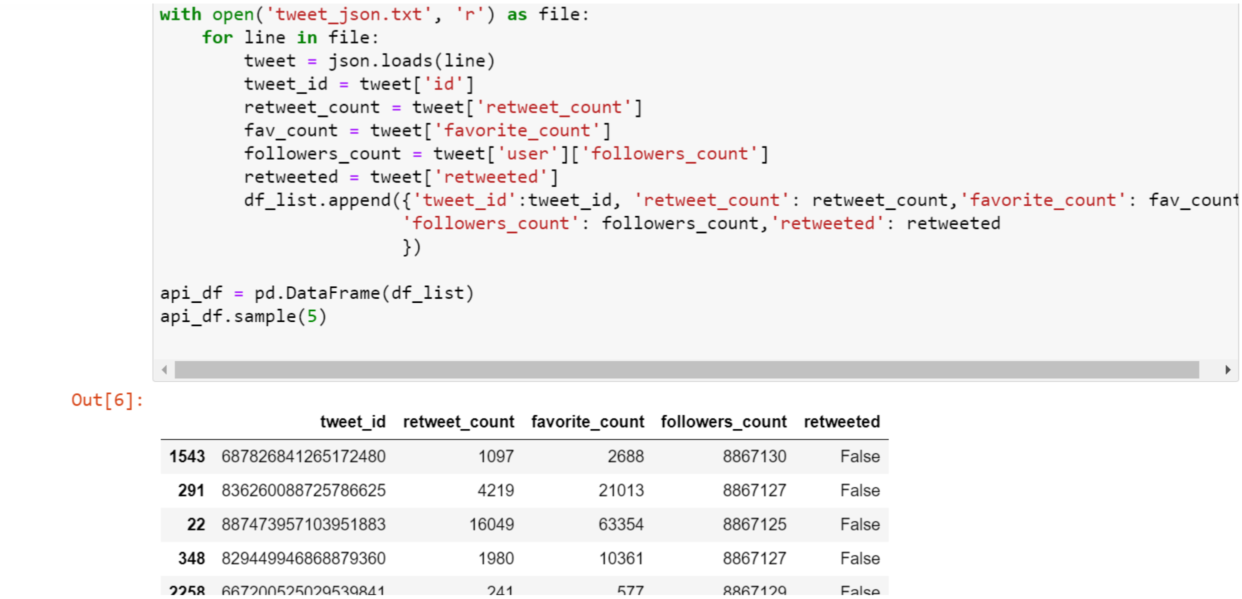
```
url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predic
file_name = 'image-predictions.tsv'
response = requests.get(url)
if not os.path.isfile(file_name):
    with open(file_name, 'wb') as f:
        f.write(response.content)
image_df= pd.read_csv('image-predictions.tsv', sep='\t')
image_df.sample(3)
```

Out[3]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	
757	688547210804498433	https://pbs.twimg.com/media/CY42CFWW8AACOwt.jpg	1	papillon	0.531279	True	Ble
962	705970349788291072	https://pbs.twimg.com/media/CcwcSS9WwAALE4f.jpg	1	golden_retriever	0.776346	True	Labı
1604	800388270626521089	https://pbs.twimg.com/media/CxuM3oZW8AEhO5z.jpg	2	golden_retriever	0.359860	True	

- **Twitter API:**

I had created twitter developer account and used secret keys and token to extract Tweet.JSON text file:



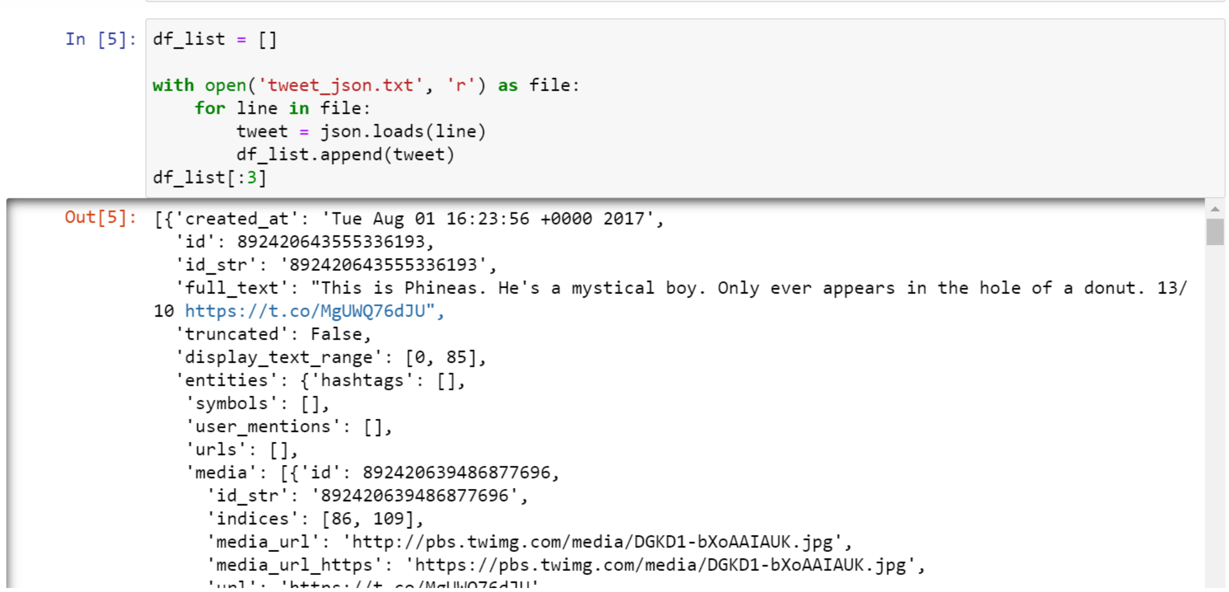
```
with open('tweet_json.txt', 'r') as file:
    for line in file:
        tweet = json.loads(line)
        tweet_id = tweet['id']
        retweet_count = tweet['retweet_count']
        fav_count = tweet['favorite_count']
        followers_count = tweet['user']['followers_count']
        retweeted = tweet['retweeted']
        df_list.append({'tweet_id':tweet_id, 'retweet_count': retweet_count,'favorite_count': fav_count
                       'followers_count': followers_count,'retweeted': retweeted
                       })

api_df = pd.DataFrame(df_list)
api_df.sample(5)
```

Out[6]:

	tweet_id	retweet_count	favorite_count	followers_count	retweeted
1543	687826841265172480	1097	2688	8867130	False
291	836260088725786625	4219	21013	8867127	False
22	887473957103951883	16049	63354	8867125	False
348	829449946868879360	1980	10361	8867127	False
2258	667200525029539841	241	577	8867129	False

- Also I tried to go deep in each tweet to extract some useful information:



```
In [5]: df_list = []

        with open('tweet_json.txt', 'r') as file:
            for line in file:
                tweet = json.loads(line)
                df_list.append(tweet)
        df_list[:3]
```

```
Out[5]: [{'created_at': 'Tue Aug 01 16:23:56 +0000 2017',
          'id': 892420643555336193,
          'id_str': '892420643555336193',
          'full_text': "This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/
        10 https://t.co/MgUWQ76dJU",
          'truncated': False,
          'display_text_range': [0, 85],
          'entities': {'hashtags': [],
           'symbols': [],
           'user_mentions': [],
           'urls': [],
           'media': [{'id': 892420639486877696,
             'id_str': '892420639486877696',
             'indices': [86, 109],
             'media_url': 'http://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg',
             'media_url_https': 'https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg',
             'url': 'https://t.co/MgUWQ76dJU'
```

“now I have got data gathered from 3 different types of data sources”. We can go to the next phase now:

- **Data assessing:**

In this process I had investigated each data set to try to figure out some quality and tidiness issues so as to get clean and accurate data for our analysis and below quality and tidiness issues I had discovered in each data set:

- **Enhanced Twitter Archive:**

○ **Quality issues:**

▪ time_stamp column need to be date time dtype.

▪ Remove zone from time_stamp +0000

▪ tweet_id need to be in string format

▪ rating_numerator need to be in float dtype

▪ rating_denominator need to be in float dtype

▪ Drop tweet_id not matched with image_prediction table

▪ Need to investigate below or above 10 rating_denominator

▪ 'Name' column need investigation for the extracted names from text column

▪ -drop rows ' expanded_urls' with null values

▪ -drop retweets with not null values

▪ -drop replies with not null values

○ **Tidiness issues:**

- Need to create new column called "dog_stage" to define the each dog stage as (doggo, floofer, pupper, and puppo)

- Need to drop columns: in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id ,retweeted_status_user_id, retweeted_status_timestamp has no meaningful usage for them in my analysi

- **Image Predictions File:**

- **Quality issues:**

- tweet_id need to be in string dtype

- Delete duplicated images in JPEG URLs

- **Tidiness issues:**

- 1. P (1,2,3)_dog which may be prediction only , p(1,2,3)_conf may be column name “confidence” only as example

- **Additional Data via the Twitter API:**

- **Quality issues:**

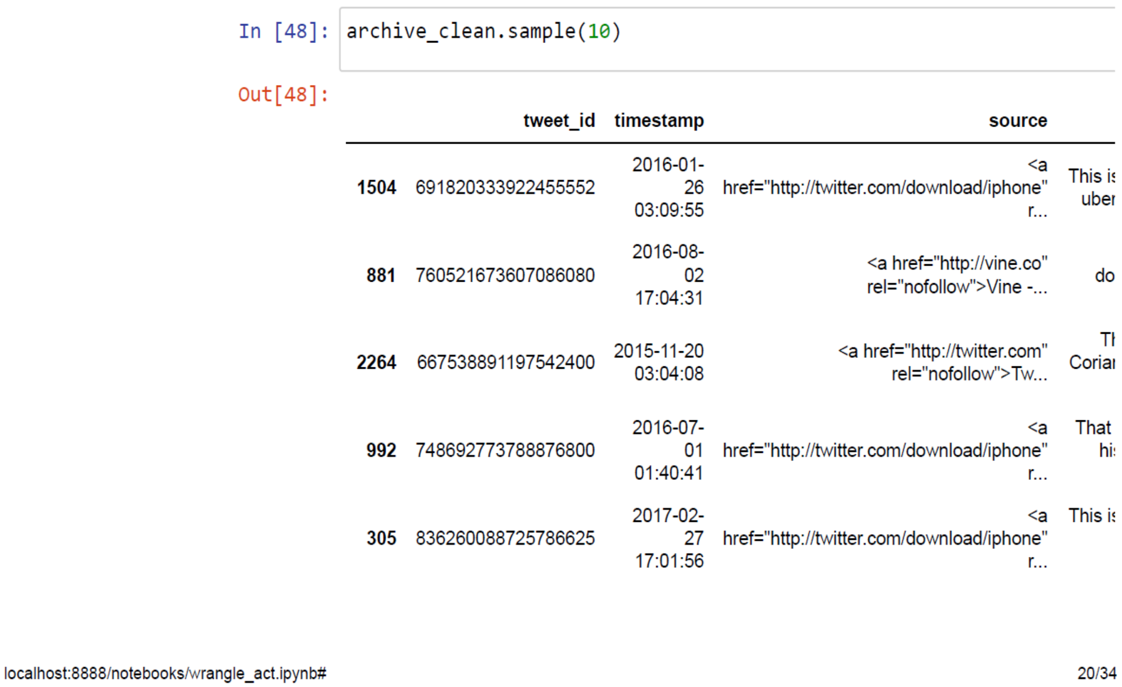
- tweet_id need to be in string dtype

- -retweet_count ,favorite_count ,followers_count need to be integer

Data cleaning:

In this part I had investigated each data set and applied cleaning process to resolve each quality/tidiness issue in order to create master csv file:

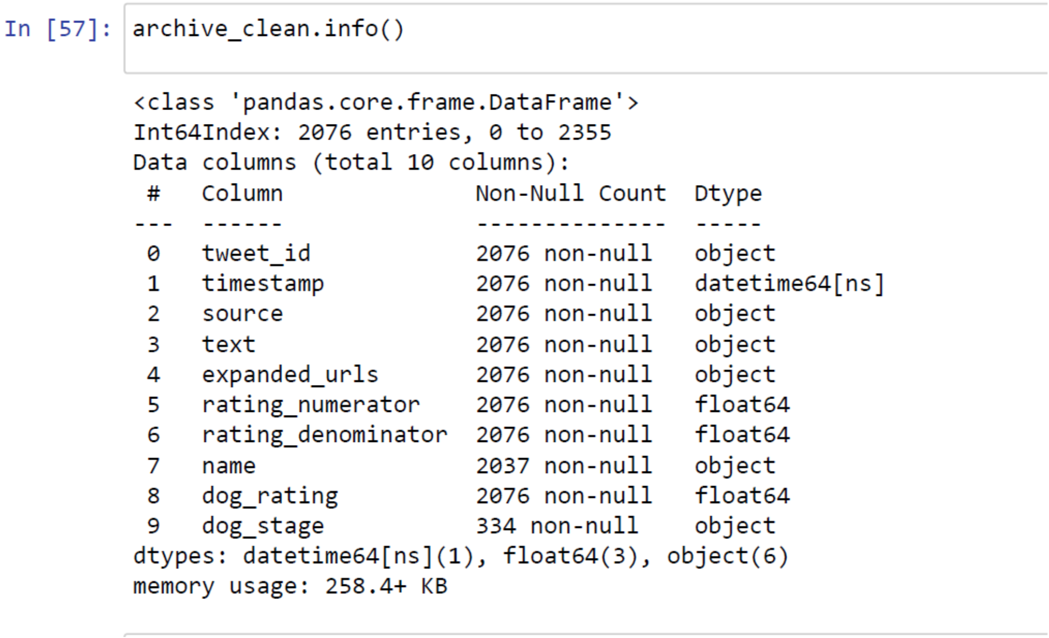
Enhanced Twitter Archive:



```
In [48]: archive_clean.sample(10)
```

Out[48]:

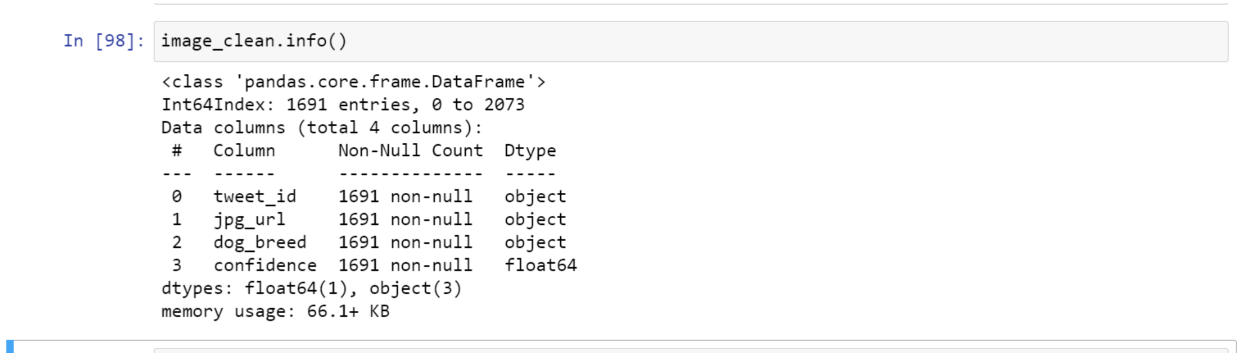
	tweet_id	timestamp	source	
1504	691820333922455552	2016-01-26 03:09:55	<a href="http://twitter.com/download/iphone" r...	This is uber
881	760521673607086080	2016-08-02 17:04:31	<a href="http://vine.co" rel="nofollow">Vine -...	do
2264	667538891197542400	2015-11-20 03:04:08	<a href="http://twitter.com" rel="nofollow">Tw...	Th Coriar
992	748692773788876800	2016-07-01 01:40:41	<a href="http://twitter.com/download/iphone" r...	That his
305	836260088725786625	2017-02-27 17:01:56	<a href="http://twitter.com/download/iphone" r...	This is



```
In [57]: archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2076 entries, 0 to 2355
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   tweet_id            2076 non-null   object
 1   timestamp           2076 non-null   datetime64[ns]
 2   source              2076 non-null   object
 3   text                2076 non-null   object
 4   expanded_urls       2076 non-null   object
 5   rating_numerator    2076 non-null   float64
 6   rating_denominator  2076 non-null   float64
 7   name                2037 non-null   object
 8   dog_rating          2076 non-null   float64
 9   dog_stage           334 non-null    object
dtypes: datetime64[ns](1), float64(3), object(6)
memory usage: 258.4+ KB
```

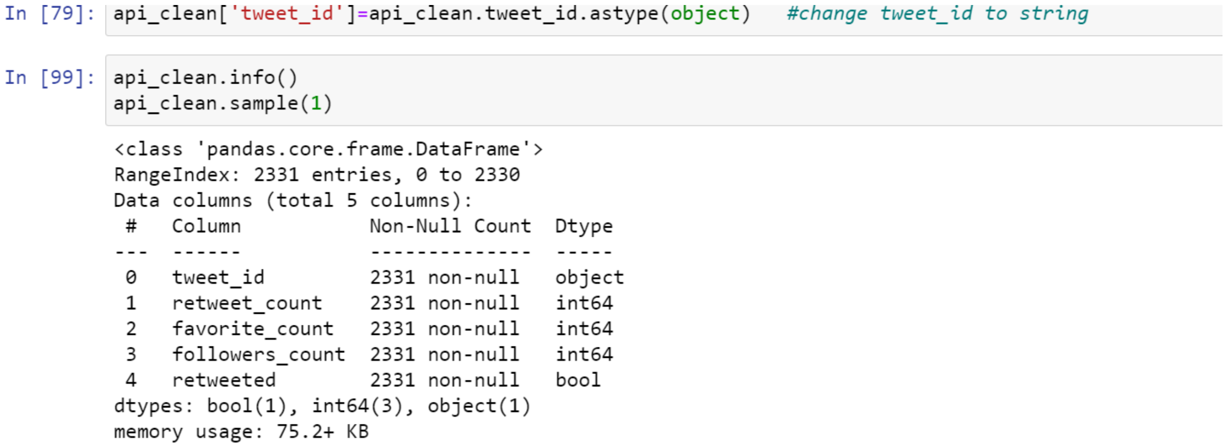
Image Predictions File:



```
In [98]: image_clean.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1691 entries, 0 to 2073
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   tweet_id    1691 non-null   object 
 1   jpg_url     1691 non-null   object 
 2   dog_breed   1691 non-null   object 
 3   confidence  1691 non-null   float64
dtypes: float64(1), object(3)
memory usage: 66.1+ KB
```

Additional Data via the Twitter API:



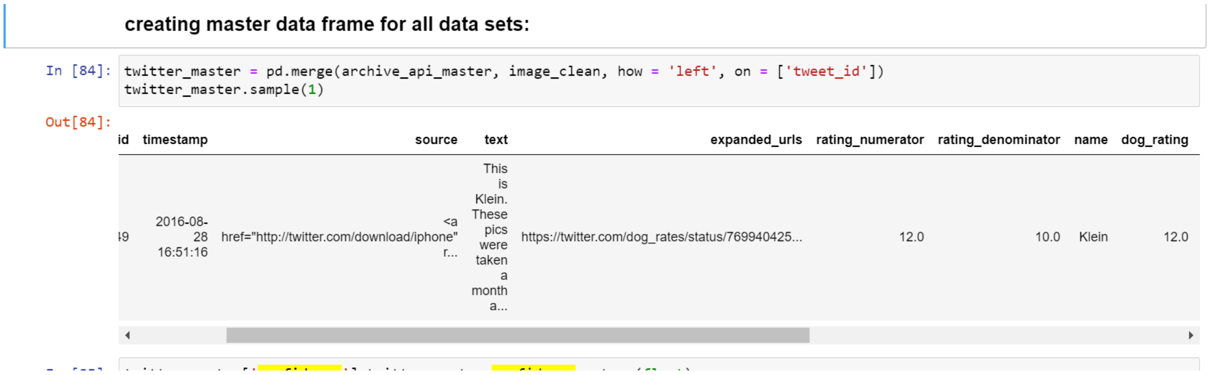
```
In [79]: api_clean['tweet_id']=api_clean.tweet_id.astype(object)   #change tweet_id to string

In [99]: api_clean.info()
         api_clean.sample(1)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2331 entries, 0 to 2330
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   tweet_id         2331 non-null   object
 1   retweet_count    2331 non-null   int64 
 2   favorite_count   2331 non-null   int64 
 3   followers_count  2331 non-null   int64 
 4   retweeted        2331 non-null   bool  
dtypes: bool(1), int64(3), object(1)
memory usage: 75.2+ KB
```

Creating master archive csv file for all data sets:

In this part I have combined and merged data sets in only one data frame in order to start analyze, asking questions, finding insights and creating visitations for insights:



creating master data frame for all data sets:

```
In [84]: twitter_master = pd.merge(archive_api_master, image_clean, how = 'left', on = ['tweet_id'])
         twitter_master.sample(1)
```

Out[84]:

id	timestamp	source	text	expanded_urls	rating_numerator	rating_denominator	name	dog_rating
49	2016-08-28 16:51:16	<a href="http://twitter.com/download/iphone" r...	This is Klein. These pics were taken a month a...	https://twitter.com/dog_rates/status/769940425...	12.0	10.0	Klein	12.0