wrangle_act

April 2, 2019

1

- WeRateDogs
-

In [ ]:

In [226]:
```
import requests
import pandas as pd
pd.options.display.max_columns=200
pd.set_option('max_colwidth',100)
```

WeRateDogs

In [227]:
```
twitter_archive_enhanced=pd.read_csv("twitter-archive-enhanced.csv")
```

In [228]:
```
twitter_archive_enhanced.head(5)
```

Out[228]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	tim
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24

In [229]:
```
import requests
with open("image-predictions.tsv",mode='wb') as f:
    file = requests.get('https://raw.githubusercontent.com/udacity/new-dand-advanced-c
    f.write(file.content)
```

In [230]:
```
from io import BytesIO
with open("image-predictions.tsv",mode='wb') as f:
    f.write(file.content)
```

In [231]:
```
image_predictions=pd.read_csv("image-predictions.tsv",sep="\t")
```

In [232]:
```
image_predictions.head(5)
```

Out[232]:

	tweet_id	jpg_url	img_num	
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rh
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	m

In [233]:
```
import json
with open("tweet_json.txt") as f:
        data={"id":[],"retweet_count":[],"favorite_count":[]}
        for line in f.readlines():
                json_obj=json.loads(line)
                data['id'].append(json_obj['id'])
                data['retweet_count'].append(json_obj['retweet_count'])
                data['favorite_count'].append(json_obj['favorite_count'])
        tweet_json=pd.DataFrame(data)
```

In [234]:
```
tweet_json.head(5)
```

Out[234]:

	id	retweet_count	favorite_count
0	892420643555336193	8842	39492
1	892177421306343426	6480	33786
2	891815181378084864	4301	25445
3	891689557279858688	8925	42863
4	891327558926688256	9721	41016

In [ ]:
```
```

2

2.1 twitter_archive_enhanced

In [235]:
```
twitter_archive_enhanced.head(5)
```

Out[235]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	tim
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24

In [236]:
```
twitter_archive_enhanced.tail(5)
```

Out[236]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	
2351	666049248165822465	NaN	NaN	2015-11-16 00:24
2352	666044226329800704	NaN	NaN	2015-11-16 00:04
2353	666033412701032449	NaN	NaN	2015-11-15 23:21
2354	666029285002620928	NaN	NaN	2015-11-15 23:05
2355	666020888022790149	NaN	NaN	2015-11-15 22:32

```
In [237]: twitter_archive_enhanced.columns.values.tolist()
```

```
Out[237]: ['tweet_id',
           'in_reply_to_status_id',
           'in_reply_to_user_id',
           'timestamp',
           'source',
           'text',
           'retweeted_status_id',
           'retweeted_status_user_id',
           'retweeted_status_timestamp',
           'expanded_urls',
           'rating_numerator',
           'rating_denominator',
           'name',
           'doggo',
           'floofer',
           'pupper',
           'puppo']
```

```
In [238]: twitter_archive_enhanced.source.sample(8)
```

```
Out[238]: 215     <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone
          1460    <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone
          1765    <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone
          1671    <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone
          847     <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone
          1521    <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone
          461     <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone
          631     <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone
          Name: source, dtype: object
```

twitter_archive_enhanced
tweet_idID
in_reply_to_status_idID
in_reply_to_user_idID
timestamptweet
sourcetweetweb
text
retweeted_status_idID
retweeted_status_user_idID
retweeted_status_timestamp
expanded_urls
rating_numerator
rating_denominator
name
doggo
floofer

pupper
puppo:

```
In [239]: ###
          twitter_archive_enhanced.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
```

infotwitter_archive_enhanced2356
in_reply_to_status_idin_reply_to_user_id78
timestamp
retweeted_status_idretweeted_status_user_idretweeted_status_timestamp 181

```
In [240]: twitter_archive_enhanced.tweet_id.value_counts().max()
```

```
Out[240]: 1
```

tweet_id

```
In [241]: twitter_archive_enhanced.timestamp.sample(7)
```

```
Out[241]: 1262    2016-03-16 16:29:35 +0000
          1048    2016-06-16 01:25:36 +0000
          1422    2016-02-12 16:16:41 +0000
          732     2016-09-29 16:03:01 +0000
          1991    2015-12-04 03:43:54 +0000
          1431    2016-02-10 20:23:19 +0000
          1363    2016-02-25 19:04:13 +0000
          Name: timestamp, dtype: object
```

timestamp

```
In [242]: twitter_archive_enhanced.source.value_counts()
```

```
Out[242]: <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
          <a href="http://vine.co" rel="nofollow">Vine - Make a Scene</a>
          <a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>
          <a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>
          Name: source, dtype: int64
```

sourcehtml,html

```
In [243]: twitter_archive_enhanced.text.sample(7)
```

```
Out[243]: 1940                The millennials have spoken and we've decided to immediately demote
          1761    Exotic pup here. Tail long af. Throat looks swollen. Might breathe fire. Excep
          2112    Two unbelievably athletic dogs here. Great form. Perfect execution. 10/10 for
          1965    This is Gerald. He's a fluffy lil yellow pup. Always looks like his favorite t
          969     This is Stewie. He will roundhouse kick anyone who questions his independence.
          665     This is Mosby. He appears to be rather h*ckin snuggable af. 12/10 keep it up M
          54      This is Gary. He couldn't miss this puppertunity for a selfie. Flawless focusi
          Name: text, dtype: object
```

```
In [244]: twitter_archive_enhanced.text.value_counts().max()
```

```
Out[244]: 1
```

```
In [245]: # pd.set_option('display.height', 1000)
          pd.set_option('display.max_rows', 500)
          pd.set_option('display.max_columns', 500)
          pd.set_option('display.width', 1000)
          pd.option_context('display.max_rows', None, 'display.max_columns', None)
```

```
Out[245]: <pandas.core.config.option_context at 0x7fd877746be0>
```

```
In [246]: import random
          random.randint(0,10)
          twitter_archive_enhanced.shape
```

```
Out[246]: (2356, 17)
```

```
In [247]: for i in range(5):
              print(twitter_archive_enhanced.text.loc[random.randint(0,twitter_archive_enhanced.
              print("\n")
```

```
Here's a doggo blowing bubbles. It's downright legendary. 13/10 would watch on repeat forever (v


This is Wesley. He's clearly trespassing. Seems rather h*ckin violent too. Weaponized forehead.
```

```
Yea I can't handle the cuteness anymore. Curls for days. 12/10 for all https://t.co/sAI6gCGZYX

Say hello to Levi. He's a Madagascan Butterbop. One of the more docile Butterbops I've seen. 12/

This is Toby. He asked for chocolate cake for his birthday but was given vanilla instead. 8/10 i
```

```
In [248]: twitter_archive_enhanced.expanded_urls.sample(7)

Out[248]: 2170                                          https://twitter.com/dog_rates/status/66935
          333     https://twitter.com/dog_rates/status/832757312314028032/photo/1,https://twitte
          186
          1344                                          https://twitter.com/dog_rates/status/70449
          1279    https://twitter.com/dog_rates/status/708845821941387268/photo/1,https://twitte
          72                                                  https://twitter.com/bbcworld/stat
          1186    https://twitter.com/dog_rates/status/718540630683709445/photo/1,https://twitte
          Name: expanded_urls, dtype: object

In [249]: twitter_archive_enhanced.rating_denominator.value_counts()

Out[249]: 10     2333
          11        3
          50        3
          80        2
          20        2
          2         1
          16        1
          40        1
          70        1
          15        1
          90        1
          110       1
          120       1
          130       1
          150       1
          170       1
          7         1
          0         1
          Name: rating_denominator, dtype: int64
```

10

```
In [ ]:
```

```
In [250]: rating_denominator=twitter_archive_enhanced.rating_denominator.value_counts()
          twitter_archive_enhanced.query("rating_denominator!=10")[["rating_numerator","rating_d
```

Out[250]:

	rating_numerator	rating_denominator
313	960	0
342	11	15
433	84	70
516	24	7
784	9	11
902	165	150
1068	9	11
1120	204	170
1165	4	20
1202	50	50
1228	99	90
1254	80	80
1274	45	50
1351	60	50
1433	44	40
1598	4	20
1634	143	130
1635	121	110
1662	7	11
1663	20	16
1779	144	120
1843	88	80
2335	1	2

tweet10

```
In [251]: index=twitter_archive_enhanced.query("rating_denominator!=10")[["rating_numerator","ra
          for i in index:
              print(twitter_archive_enhanced.text.loc[i])
              print(twitter_archive_enhanced.rating_numerator.loc[i])
              print(twitter_archive_enhanced.rating_denominator.loc[i])
              print("\n")
```

```
@jonnysun @Lin_Manuel ok jomny I know you're excited but 960/00 isn't a valid rating, 13/10 is t
960
0


@docmisterio account started on 11/15/15
11
15


The floofs have been released I repeat the floofs have been released. 84/70 https://t.co/NIYC82C
84
```

70

Meet Sam. She smiles 24/7 & secretly aspires to be a reindeer.
Keep Sam smiling by clicking and sharing this link:
https://t.co/98tB8y7y7t https://t.co/LouL5vdvxx
24
7

RT @dog_rates: After so many requests, this is Bretagne. She was the last surviving 9/11 search
9
11

Why does this never happen at my front door... 165/150 https://t.co/HmwrdfEfUE
165
150

After so many requests, this is Bretagne. She was the last surviving 9/11 search dog, and our se
9
11

Say hello to this unbelievably well behaved squad of doggos. 204/170 would try to pet all at onc
204
170

Happy 4/20 from the squad! 13/10 for all https://t.co/eV1diwds8a
4
20

This is Bluebert. He just saw that both #FinalFur match ups are split 50/50. Amazed af. 11/10 ht
50
50

Happy Saturday here's 9 puppers on a bench. 99/90 good work everybody https://t.co/mpvaVxKmc1
99
90

Here's a brigade of puppers. All look very prepared for whatever happens next. 80/80 https://t.c
80
80

```
From left to right:
Cletus, Jerome, Alejandro, Burp, &amp;amp; Titson
None know where camera is. 45/50 would hug all at once https://t.co/sedre1ivTK
45
50


Here is a whole flock of puppers.  60/50 I'll take the lot https://t.co/9dpcw6MdWa
60
50


Happy Wednesday here's a bucket of pups. 44/40 would pet all at once https://t.co/HppvrYuamZ
44
40


Yes I do realize a rating of 4/20 would've been fitting. However, it would be unjust to give the
4
20


Two sneaky puppers were not initially seen, moving the rating to 143/130. Please forgive us. Tha
143
130


Someone help the girl is being mugged. Several are distracting her while two steal her shoes. Cl
121
110


This is Darrel. He just robbed a 7/11 and is in a high speed police chase. Was just spotted by t
7
11


I'm aware that I could've said 20/16, but here at WeRateDogs we are very professional. An incons
20
16


IT'S PUPPERGEDDON. Total of 144/120 ...I think https://t.co/ZanVtAtvIq
144
120
```

```
Here we have an entire platoon of puppers. Total score: 88/80 would pet all at once https://t.co
88
80


This is an Albanian 3 1/2 legged  Episcopalian. Loves well-polished hardwood flooring. Penis on
1
2
```

In [252]:
```
twitter_archive_enhanced.rating_numerator.value_counts()
```

Out[252]:
```
12      558
11      464
10      461
13      351
9       158
8       102
7        55
14       54
5        37
6        32
3        19
4        17
1         9
2         9
420       2
0         2
15        2
75        2
80        1
20        1
24        1
26        1
44        1
50        1
60        1
165       1
84        1
88        1
144       1
182       1
143       1
666       1
960       1
```