wrangle_act

April 26, 2018

1 Data Wrangling

2 Created by Benjamin Fundelits

2.1 Gather

2.1.1 Local Disk

In [191]:
```
import pandas as pd
import numpy as np
import requests
import csv
import tweepy
import json
import matplotlib.pyplot as plt
import re
```

In [192]:
```
# Import archive data
archive = pd.read_csv("twitter-archive-enhanced.csv")
```

In [193]:
```
# Check to see if the file was imported correctly
archive.head()
```

Out[193]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	

```
0  <a href="http://twitter.com/download/iphone" r...
1  <a href="http://twitter.com/download/iphone" r...
2  <a href="http://twitter.com/download/iphone" r...
3  <a href="http://twitter.com/download/iphone" r...
4  <a href="http://twitter.com/download/iphone" r...

                                                text  retweeted_status_id  \
0  This is Phineas. He's a mystical boy. Only eve...                   NaN
1  This is Tilly. She's just checking pup on you...                   NaN
2  This is Archie. He is a rare Norwegian Pouncin...                   NaN
3  This is Darla. She commenced a snooze mid meal...                   NaN
4  This is Franklin. He would like you to stop ca...                   NaN

   retweeted_status_user_id retweeted_status_timestamp  \
0                       NaN                        NaN
1                       NaN                        NaN
2                       NaN                        NaN
3                       NaN                        NaN
4                       NaN                        NaN

                                       expanded_urls  rating_numerator  \
0  https://twitter.com/dog_rates/status/892420643...                13
1  https://twitter.com/dog_rates/status/892177421...                13
2  https://twitter.com/dog_rates/status/891815181...                12
3  https://twitter.com/dog_rates/status/891689557...                13
4  https://twitter.com/dog_rates/status/891327558...                12

   rating_denominator      name doggo floofer pupper puppo
0                  10   Phineas  None    None   None  None
1                  10     Tilly  None    None   None  None
2                  10    Archie  None    None   None  None
3                  10     Darla  None    None   None  None
4                  10  Franklin  None    None   None  None
```

2.1.2 Udacity Website

```
In [194]: # Import image predictions
          url = "https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predi

          with requests.Session() as s:
              download = s.get(url)

              decoded_content = download.content.decode("utf-8")

              reader = csv.reader(decoded_content.splitlines(), delimiter = '\t')

              downloaded_ip_list = list(reader)
```

In [195]:
```
# Convert downloaded csv into a dataframe
image_prediction = pd.DataFrame(data = downloaded_ip_list[1:], columns = downloaded_
```

In [196]:
```
# Check to see if file was imported correctly
image_prediction.head(10)
```

Out[196]:

	tweet_id	jpg_url	\
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	
5	666050758794694657	https://pbs.twimg.com/media/CT5Jof1WUAEuVxN.jpg	
6	666051853826850816	https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg	
7	666055525042405380	https://pbs.twimg.com/media/CT5N9tpXIAAifs1.jpg	
8	666057090499244032	https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg	
9	666058600524156928	https://pbs.twimg.com/media/CT5Qw94XAAA_2dP.jpg	

	img_num	p1	p1_conf	p1_dog	\
0	1	Welsh_springer_spaniel	0.465074	True	
1	1	redbone	0.506826	True	
2	1	German_shepherd	0.596461	True	
3	1	Rhodesian_ridgeback	0.408143	True	
4	1	miniature_pinscher	0.560311	True	
5	1	Bernese_mountain_dog	0.651137	True	
6	1	box_turtle	0.9330120000000001	False	
7	1	chow	0.692517	True	
8	1	shopping_cart	0.962465	False	
9	1	miniature_poodle	0.201493	True	

	p2	p2_conf	p2_dog	\
0	collie	0.156665	True	
1	miniature_pinscher	0.07419169999999999	True	
2	malinois	0.13858399999999998	True	
3	redbone	0.360687	True	
4	Rottweiler	0.243682	True	
5	English_springer	0.263788	True	
6	mud_turtle	0.04588540000000001	False	
7	Tibetan_mastiff	0.058279399999999995	True	
8	shopping_basket	0.014593799999999999	False	
9	komondor	0.192305	True	

	p3	p3_conf	p3_dog
0	Shetland_sheepdog	0.0614285	True
1	Rhodesian_ridgeback	0.07201	True
2	bloodhound	0.11619700000000001	True
3	miniature_pinscher	0.222752	True
4	Doberman	0.154629	True

```
5   Greater_Swiss_Mountain_dog             0.0161992   True
6                      terrapin  0.017885299999999996  False
7                      fur_coat             0.0544486  False
8             golden_retriever            0.00795896   True
9  soft-coated_wheaten_terrier   0.08208610000000001   True
```

2.1.3 API

```
In [197]: # text file name
          text_file_name = "tweet_json.txt"
```

Load Data from API If you do not want to load all data from the API again, check down below the "Load Data from Text" title

```
In [189]: #Hide this code part
          #My authentication informations
          my_secret_consumer_key = "Private_information"
          my_secret_consumer_secret = "Private_information"
          my_secret_access_token = "Private_information"
          my_secret_access_token_secret = "Private_information"
```

```
In [ ]: # Authentication
        auth = tweepy.OAuthHandler(consumer_key = my_secret_consumer_key,
                                   consumer_secret = my_secret_consumer_secret)
        auth.set_access_token(my_secret_access_token, my_secret_access_token_secret)
```

```
In [ ]: # Get Tweepy API
        api = tweepy.API(auth, wait_on_rate_limit = True, wait_on_rate_limit_notify = True)
```

```
In [ ]: # Gather json tweets into a list
        counter = 0
        tweet_json_list_from_api = []
        for _id in archive['tweet_id']:
            try:
                tweet = api.get_status(_id)
                tweet_json_list.append(tweet._json)
            except:
                continue
```

```
In [ ]: # Write tweet json objects into a text
        with open(text_file_name, "a", encoding="utf-8") as text_file:
            for tweet in tweet_json_list_from_api:
                json.dump(tweet, text_file)
                text_file.write("\n")
```

Load Data from Text Load json from text, if we shutted down the Kernel. In this case we do not have to load all data from the API again

In [198]:
```
# Load json from text
tweet_json_list = []
with open(text_file_name, "r", encoding = 'utf-8') as text_file:
    for line in text_file:
        tweet_line = line
        tweet_json_list.append(tweet_line)
```

In [199]:
```
# Gather important informations from tweets into a list of dictionaries
tweet_list = []

for token in tweet_json_list:
    tweet = json.loads(token)
    tweet_id = tweet["id_str"]
    retweet_count = tweet["retweet_count"]
    favorite_count = tweet["favorite_count"]

    tweet_list.append({
        "id" : tweet_id,
        "retweet_count" : retweet_count,
        "favorite_count" : favorite_count,
    })
```

In [200]:
```
# Create Data Frame from list of dictionaries
tweet_count = pd.DataFrame(tweet_list, columns = ["id", "retweet_count",  "favorite_
```

In [201]:
```
# Check to see if counts were gathered correctly
tweet_count.sample(10)
```

Out[201]:

	id	retweet_count	favorite_count
420	821107785811234820	2405	10463
434	819347104292290561	1354	7871
1410	698195409219559425	6586	18053
1638	683742671509258241	3690	7023
1867	675047298674663426	356	1118
574	800141422401830912	2909	16797
55	881536004380872706	16212	49668
712	783085703974514689	2501	8934
965	750026558547456000	873	2936
1124	728409960103686147	2217	5284

2.2 Assess

In [202]:
```
archive.head()
```

Out[202]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	\
0	892420643555336193	NaN	NaN	
1	892177421306343426	NaN	NaN	

```
2  891815181378084864                    NaN                  NaN
3  891689557279858688                    NaN                  NaN
4  891327558926688256                    NaN                  NaN

                   timestamp  \
0  2017-08-01 16:23:56 +0000
1  2017-08-01 00:17:27 +0000
2  2017-07-31 00:18:03 +0000
3  2017-07-30 15:58:51 +0000
4  2017-07-29 16:00:24 +0000

                                              source  \
0  <a href="http://twitter.com/download/iphone" r...
1  <a href="http://twitter.com/download/iphone" r...
2  <a href="http://twitter.com/download/iphone" r...
3  <a href="http://twitter.com/download/iphone" r...
4  <a href="http://twitter.com/download/iphone" r...

                                                text  retweeted_status_id  \
0  This is Phineas. He's a mystical boy. Only eve...                  NaN
1  This is Tilly. She's just checking pup on you...                  NaN
2  This is Archie. He is a rare Norwegian Pouncin...                  NaN
3  This is Darla. She commenced a snooze mid meal...                  NaN
4  This is Franklin. He would like you to stop ca...                  NaN

   retweeted_status_user_id retweeted_status_timestamp  \
0                       NaN                       NaN
1                       NaN                       NaN
2                       NaN                       NaN
3                       NaN                       NaN
4                       NaN                       NaN

                                        expanded_urls  rating_numerator  \
0  https://twitter.com/dog_rates/status/892420643...                13
1  https://twitter.com/dog_rates/status/892177421...                13
2  https://twitter.com/dog_rates/status/891815181...                12
3  https://twitter.com/dog_rates/status/891689557...                13
4  https://twitter.com/dog_rates/status/891327558...                12

   rating_denominator      name doggo floofer pupper puppo
0                  10   Phineas  None    None   None  None
1                  10     Tilly  None    None   None  None
2                  10    Archie  None    None   None  None
3                  10     Darla  None    None   None  None
4                  10  Franklin  None    None   None  None
```

```
In [203]: image_prediction.head(10)
```

```
Out[203]:              tweet_id                                          jpg_url  \
```

```
0  666020888022790149  https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg
1  666029285002620928  https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg
2  666033412701032449  https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg
3  666044226329800704  https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg
4  666049248165822465  https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg
5  666050758794694657  https://pbs.twimg.com/media/CT5Jof1WUAEuVxN.jpg
6  666051853826850816  https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg
7  666055525042405380  https://pbs.twimg.com/media/CT5N9tpXIAAifs1.jpg
8  666057090499244032  https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg
9  666058600524156928  https://pbs.twimg.com/media/CT5Qw94XAAA_2dP.jpg

  img_num                     p1             p1_conf p1_dog  \
0       1  Welsh_springer_spaniel            0.465074   True
1       1                redbone            0.506826   True
2       1         German_shepherd            0.596461   True
3       1     Rhodesian_ridgeback            0.408143   True
4       1      miniature_pinscher            0.560311   True
5       1    Bernese_mountain_dog            0.651137   True
6       1              box_turtle  0.9330120000000001  False
7       1                   chow            0.692517   True
8       1           shopping_cart            0.962465  False
9       1        miniature_poodle            0.201493   True

                 p2              p2_conf p2_dog  \
0             collie             0.156665   True
1  miniature_pinscher  0.07419169999999999   True
2           malinois  0.13858399999999998   True
3            redbone             0.360687   True
4         Rottweiler             0.243682   True
5   English_springer             0.263788   True
6         mud_turtle  0.04588540000000001  False
7    Tibetan_mastiff  0.058279399999999995   True
8    shopping_basket  0.014593799999999999  False
9           komondor             0.192305   True

                          p3              p3_conf p3_dog
0           Shetland_sheepdog            0.0614285   True
1         Rhodesian_ridgeback              0.07201   True
2                   bloodhound  0.11619700000000001   True
3          miniature_pinscher             0.222752   True
4                    Doberman             0.154629   True
5  Greater_Swiss_Mountain_dog            0.0161992   True
6                      terrapin  0.017885299999999996  False
7                     fur_coat            0.0544486  False
8            golden_retriever           0.00795896   True
9  soft-coated_wheaten_terrier  0.08208610000000001   True
```

In [204]: `tweet_count.head(15)`

Out[204]:

	id	retweet_count	favorite_count
0	892420643555336193	8646	38976
1	892177421306343426	6351	33358
2	891815181378084864	4213	25144
3	891689557279858688	8762	42314
4	891327558926688256	9527	40485
5	891087950875897856	3158	20297
6	890971913173991426	2105	11904
7	890729181411237888	19161	65840
8	890609185150312448	4321	27885
9	890240255349198849	7525	32076
10	890006608113172480	7440	30782
11	889880896479866881	5031	27907
12	889665388333682689	10194	48272
13	889638837579907072	4602	27297
14	889531135344209921	2262	15152

In [205]:
```
# Display a text
archive["text"][0]
```

Out[205]: "This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/1

2.2.1 Call Info on every table

In [206]:
```
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
```

```
In [207]: image_prediction.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null object
jpg_url     2075 non-null object
img_num     2075 non-null object
p1          2075 non-null object
p1_conf     2075 non-null object
p1_dog      2075 non-null object
p2          2075 non-null object
p2_conf     2075 non-null object
p2_dog      2075 non-null object
p3          2075 non-null object
p3_conf     2075 non-null object
p3_dog      2075 non-null object
dtypes: object(12)
memory usage: 194.6+ KB
```

```
In [208]: tweet_count.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2345 entries, 0 to 2344
Data columns (total 3 columns):
id                2345 non-null object
retweet_count     2345 non-null int64
favorite_count    2345 non-null int64
dtypes: int64(2), object(1)
memory usage: 55.0+ KB
```

2.2.2 Describe tables

```
In [209]: archive.describe()
```

```
Out[209]:            tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
          count  2.356000e+03           7.800000e+01         7.800000e+01
          mean   7.427716e+17           7.455079e+17         2.014171e+16
          std    6.856705e+16           7.582492e+16         1.252797e+17
          min    6.660209e+17           6.658147e+17         1.185634e+07
          25%    6.783989e+17           6.757419e+17         3.086374e+08
          50%    7.196279e+17           7.038708e+17         4.196984e+09
          75%    7.993373e+17           8.257804e+17         4.196984e+09
          max    8.924206e+17           8.862664e+17         8.405479e+17

                 retweeted_status_id  retweeted_status_user_id  rating_numerator  \
          count         1.810000e+02              1.810000e+02       2356.000000
```

```
mean          7.720400e+17              1.241698e+16         13.126486
std           6.236928e+16              9.599254e+16         45.876648
min           6.661041e+17              7.832140e+05          0.000000
25%           7.186315e+17              4.196984e+09         10.000000
50%           7.804657e+17              4.196984e+09         11.000000
75%           8.203146e+17              4.196984e+09         12.000000
max           8.874740e+17              7.874618e+17       1776.000000

       rating_denominator
count         2356.000000
mean            10.455433
std              6.745237
min              0.000000
25%             10.000000
50%             10.000000
75%             10.000000
max            170.000000
```

What is that 0 rating at enumareator and denominator? They don't really give less then 10 points for dogs.

```
In [210]: # Gather values with minimum numerator or denominator
          minimum_rating = archive[(archive["rating_numerator"] == 0) | (archive["rating_denom
          minimum_rating
```

Out[210]:

```
                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
313   835246439529840640           8.352460e+17         2.625958e+07
315   835152434251116546                    NaN                  NaN
1016  746906459439529985           7.468859e+17         4.196984e+09

                     timestamp  \
313   2017-02-24 21:54:03 +0000
315   2017-02-24 15:40:31 +0000
1016  2016-06-26 03:22:31 +0000

                                                 source  \
313   <a href="http://twitter.com/download/iphone" r...
315   <a href="http://twitter.com/download/iphone" r...
1016  <a href="http://twitter.com/download/iphone" r...

                                                   text  retweeted_status_id  \
313   @jonnysun @Lin_Manuel ok jomny I know you're e...                  NaN
315   When you're so blinded by your systematic plag...                  NaN
1016  PUPDATE: can't see any. Even if I could, I cou...                  NaN

      retweeted_status_user_id retweeted_status_timestamp  \
313                        NaN                        NaN
```

```
315                         NaN                         NaN
1016                        NaN                         NaN

                                                expanded_urls  rating_numerator  \
313                                                       NaN               960
315   https://twitter.com/dog_rates/status/835152434...                   0
1016  https://twitter.com/dog_rates/status/746906459...                   0

      rating_denominator  name doggo floofer pupper puppo
313                    0  None  None    None   None  None
315                   10  None  None    None   None  None
1016                  10  None  None    None   None  None
```

In [211]:
```
# DIsplay an url
archive["expanded_urls"][1]
```

Out[211]:
```
'https://twitter.com/dog_rates/status/892177421306343426/photo/1'
```

In [212]:
```
image_prediction.describe()
```

Out[212]:
```
                   tweet_id                                       jpg_url  \
count                  2075                                          2075
unique                 2075                                          2009
top     667119796878725120  https://pbs.twimg.com/media/Co-hmcYXYAASkiG.jpg
freq                      1                                             2

       img_num                p1   p1_conf p1_dog                  p2  \
count     2075              2075      2075   2075                2075
unique       4               378      2006      2                 405
top          1  golden_retriever  0.978833   True  Labrador_retriever
freq      1780               150         2   1532                 104

          p2_conf p2_dog                  p3    p3_conf p3_dog
count        2075   2075                2075       2075   2075
unique       2004      2                 408       2006      2
top     0.0693617   True  Labrador_retriever  0.0174919   True
freq            3   1553                  79          2   1499
```

In [213]:
```
image_prediction.p1_conf.sort_values()
```

Out[213]:
```
38                 0.0443334
136     0.055379399999999995
1093    0.059032600000000005
1370     0.06315180000000001
246                 0.070076
250      0.07112439999999999
145                 0.0715361
680                 0.0728852
701      0.08110099999999999
```

```
1831     0.08248860000000001
18                  0.086502
109                0.0885297
568                  0.08854
301                0.0891647
1627               0.0903414
1503               0.0903414
954                0.0905082
277                0.0960627
2074     0.09704860000000001
664                 0.097232
789                0.0974997
515                0.0982826
1664               0.0998035
731                 0.100499
1723                0.100896
1037                0.105171
247      0.10731700000000001
876                 0.107948
866      0.11058699999999999
1245     0.11149300000000001
                ...
76                  0.999091
1988     0.9991200000000001
2045                0.999201
863                 0.999223
1872                0.999281
1548     0.9993059999999999
95       0.9993350000000001
611                 0.999365
1711                0.999403
512                 0.999484
168      0.9996139999999999
107                 0.999647
1796     0.9997149999999999
1455     0.9998229999999999
1687                0.999828
1725                0.999833
331                 0.999834
1014     0.9998370000000001
594      0.9998459999999999
475                 0.999876
865                 0.999885
45       0.9998879999999999
1447                0.999916
242      0.9999239999999999
230      0.9999450000000001
1372     0.9999530000000001
```

```
149                  0.999956
1229      0.9999620000000001
1299                 0.999984
106                       1.0
Name: p1_conf, Length: 2075, dtype: object
```

There are multiple rows with the same jpg_url. Let's see them.

```
In [214]: # Unique jpg_url counts
          image_prediction.jpg_url.value_counts()
```

```
Out[214]: https://pbs.twimg.com/media/Co-hmcYXYAASkiG.jpg
          https://pbs.twimg.com/media/CvaYgDOWgAEfjls.jpg
          https://pbs.twimg.com/media/CmoPdmHW8AAi8BI.jpg
          https://pbs.twimg.com/media/CW88XN4WsAAlo8r.jpg
          https://pbs.twimg.com/media/CtKHLuCWYAA2TTs.jpg
          https://pbs.twimg.com/media/CrXhIqBW8AA6Bse.jpg
          https://pbs.twimg.com/media/CuRDF-XWcAIZSer.jpg
          https://pbs.twimg.com/media/Cs_DYr1XEAA54Pu.jpg
          https://pbs.twimg.com/media/CtzKC7zXEAALfSo.jpg
          https://pbs.twimg.com/media/CvJCabcWgAIoUxW.jpg
          https://pbs.twimg.com/ext_tw_video_thumb/817423809049493505/pu/img/5OFW0yueFu9oTUiQ.j
          https://pbs.twimg.com/media/Cp6db4-XYAAMmqL.jpg
          https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg
          https://pbs.twimg.com/media/Ck2d7tJWUAEPTL3.jpg
          https://pbs.twimg.com/media/C2kzTGxWEAEOpPL.jpg
          https://pbs.twimg.com/ext_tw_video_thumb/807106774843039744/pu/img/8XZg1xW35Xp2J6JW.j
          https://pbs.twimg.com/media/CkjMx99UoAM2B1a.jpg
          https://pbs.twimg.com/media/CYLDikFWEAAIy1y.jpg
          https://pbs.twimg.com/media/CwiuEJmW8AAZnit.jpg
          https://pbs.twimg.com/media/CvoBPWRWgAA4het.jpg
          https://pbs.twimg.com/media/C4bTH6nWMAAX_bJ.jpg
          https://pbs.twimg.com/media/CiibOMzUYAA9Mxz.jpg
          https://pbs.twimg.com/media/C12x-JTVIAAzdfl.jpg
          https://pbs.twimg.com/media/Ct72q9jWcAAhlnw.jpg
          https://pbs.twimg.com/media/C12whDoVEAALRxa.jpg
          https://pbs.twimg.com/media/CVgdFjNWEAAxmbq.jpg
          https://pbs.twimg.com/media/CwJR1okWIAA6XMp.jpg
          https://pbs.twimg.com/media/Cwx99rpW8AMk_Ie.jpg
          https://pbs.twimg.com/media/CcG07BYW0AErrC9.jpg
          https://pbs.twimg.com/media/CdHwZd0VIAA4792.jpg

          https://pbs.twimg.com/media/CUym4Y5WsAEiI9_.jpg
          https://pbs.twimg.com/media/CZv13u5WYAA6wQe.jpg
          https://pbs.twimg.com/media/Cf4bcm8XEAAX4xV.jpg
          https://pbs.twimg.com/media/CavjCdJW0AIB5Oz.jpg
          https://pbs.twimg.com/media/DEF2-_hXoAAs62q.jpg
          https://pbs.twimg.com/media/CyIgaTEVEAA-9zS.jpg
```

```
https://pbs.twimg.com/media/CUObvUJVEAAnYPF.jpg
https://pbs.twimg.com/media/Cmjzc-oWEAESFCm.jpg
https://pbs.twimg.com/media/Cmjlsh1XgAEvhq_.jpg
https://pbs.twimg.com/media/Ce6b4MPWwAA22Xm.jpg
https://pbs.twimg.com/media/CUJK18UWEAEg7AR.jpg
https://pbs.twimg.com/media/CU3d0azWUAA38FD.jpg
https://pbs.twimg.com/media/CmU2DVWWgAArvp3.jpg
https://pbs.twimg.com/media/CUoSjTnWwAANNak.jpg
https://pbs.twimg.com/media/Cp8k6oRWcAUL78U.jpg
https://pbs.twimg.com/ext_tw_video_thumb/698341973569245184/pu/img/Sj3A2vSfbKWSv61T.j
https://pbs.twimg.com/media/CoAqwPTW8AAiJlz.jpg
https://pbs.twimg.com/media/CbJRrigW0AIcJ2N.jpg
https://pbs.twimg.com/media/Cf3sH62VAAA-LiP.jpg
https://pbs.twimg.com/media/DFg_2PVW0AEHN3p.jpg
https://pbs.twimg.com/media/CwJEIKTWYAAvL-T.jpg
https://pbs.twimg.com/media/CW-dU34WQAANBGy.jpg
https://pbs.twimg.com/media/Cf4qRcmWEAA9V4h.jpg
https://pbs.twimg.com/media/CWFFt3_XIAArIYK.jpg
https://pbs.twimg.com/media/Crdhh_1XEAAHKHi.jpg
https://pbs.twimg.com/media/CfxcKU6W8AE-wEx.jpg
https://pbs.twimg.com/media/CZ5entwWYAAocEg.jpg
https://pbs.twimg.com/media/Ctc_-BTWEAAQpZh.jpg
https://pbs.twimg.com/media/C0DhpcrUAAAnx88.jpg
https://pbs.twimg.com/media/CYyucekVAAESj8K.jpg
Name: jpg_url, Length: 2009, dtype: int64
```

In [215]:
```
#duplicated jpg_url
image_prediction[image_prediction.jpg_url.duplicated()]
```

Out[215]:

	tweet_id	jpg_url
1297	752309394570878976	https://pbs.twimg.com/ext_tw_video_thumb/67535...
1315	754874841593970688	https://pbs.twimg.com/media/CWza7kpWcAAdYLc.jpg
1333	757729163776290825	https://pbs.twimg.com/media/CWyD2HGUYAQ1Xa7.jpg
1345	759159934323924993	https://pbs.twimg.com/media/CU1zsMSUAAAS0qW.jpg
1349	759566828574212096	https://pbs.twimg.com/media/CkNjahBXAAQ2kWo.jpg
1364	761371037149827077	https://pbs.twimg.com/tweet_video_thumb/CeBym7...
1368	761750502866649088	https://pbs.twimg.com/media/CYLDikFWEAAIy1y.jpg
1387	766078092750233600	https://pbs.twimg.com/media/ChK1tdBWwAQ1flD.jpg
1407	770093767776997377	https://pbs.twimg.com/media/CkjMx99UoAM2B1a.jpg
1417	771171053431250945	https://pbs.twimg.com/media/CVgdFjNWEAAxmbq.jpg
1427	772615324260794368	https://pbs.twimg.com/media/Cp6db4-XYAAMmqL.jpg
1446	775898661951791106	https://pbs.twimg.com/media/CiyHLocU4AI2pJu.jpg
1453	776819012571455488	https://pbs.twimg.com/media/CW88XN4WsAAlo8r.jpg
1456	777641927919427584	https://pbs.twimg.com/media/CmoPdmHW8AAi8BI.jpg
1463	778396591732486144	https://pbs.twimg.com/media/CcG07BYW0AErrC9.jpg
1476	780496263422808064	https://pbs.twimg.com/media/Ck2d7tJWUAEPTL3.jpg
1487	782021823840026624	https://pbs.twimg.com/media/CdHwZd0VIAA4792.jpg
1495	783347506784731136	https://pbs.twimg.com/media/CVuQ2LeUsAAIe3s.jpg

```
1510  786036967502913536    https://pbs.twimg.com/media/CtKHLuCWYAA2TTs.jpg
1522  788070120937619456    https://pbs.twimg.com/media/Co-hmcYXYAASkiG.jpg
1538  790723298204217344    https://pbs.twimg.com/media/CvaYgDOWgAEfjls.jpg
1541  791026214425268224    https://pbs.twimg.com/media/CpmyNumW8AAAJGj.jpg
1564  793614319594401792    https://pbs.twimg.com/media/CvyVxQRWEAAdSZS.jpg
1569  794355576146903043    https://pbs.twimg.com/media/CvJCabcWgAIoUxW.jpg
1571  794983741416415232    https://pbs.twimg.com/media/CvT6IV6WEAQhhV5.jpg
1579  796177847564038144    https://pbs.twimg.com/media/Cwx99rpW8AMk_Ie.jpg
1588  798340744599797760    https://pbs.twimg.com/media/CrXhIqBW8AA6Bse.jpg
1589  798628517273620480    https://pbs.twimg.com/media/CUN4Or5UAAAa5K4.jpg
1590  798644042770751489    https://pbs.twimg.com/media/CU3mITUWIAAfyQS.jpg
1591  798665375516884993    https://pbs.twimg.com/media/CVMOlMiWwAA4Yxl.jpg
...                  ...                                               ...
1619  802624713319034886    https://pbs.twimg.com/media/CsrjryzWgAAZY00.jpg
1624  803692223237865472    https://pbs.twimg.com/media/CZhn-QAWwAASQan.jpg
1627  804413760345620481    https://pbs.twimg.com/media/CuRDF-XWcAIZSer.jpg
1634  805958939288408065    https://pbs.twimg.com/media/CtzKC7zXEAALfSo.jpg
1636  806242860592926720    https://pbs.twimg.com/media/Ct72q9jWcAAhlnw.jpg
1640  807059379405148160    https://pbs.twimg.com/media/Ct2qO5PXEAE6eB0.jpg
1645  808134635716833280    https://pbs.twimg.com/media/Cx5R8wPVEAALa9r.jpg
1652  809808892968534016    https://pbs.twimg.com/media/CwS4aqZXUAAe3IO.jpg
1683  813944609378369540    https://pbs.twimg.com/media/Cveg1-NXgAASaaT.jpg
1693  816014286006976512    https://pbs.twimg.com/media/CiibOMzUYAA9Mxz.jpg
1699  816829038950027264    https://pbs.twimg.com/media/CvoBPWRWgAA4het.jpg
1703  817181837579653120  https://pbs.twimg.com/ext_tw_video_thumb/81596...
1712  818588835076603904    https://pbs.twimg.com/media/Crwxb5yWgAAX5P_.jpg
1717  819015331746349057    https://pbs.twimg.com/media/C12x-JTVIAAzdfl.jpg
1718  819015337530290176    https://pbs.twimg.com/media/C12whDoVEAALRxa.jpg
1727  820446719150292993    https://pbs.twimg.com/media/CxqsX-8XUAAEvjD.jpg
1736  821813639212650496    https://pbs.twimg.com/media/CtVAvX-WIAAcGTf.jpg
1742  822647212903690241    https://pbs.twimg.com/media/C2oRbOuWEAAbVSl.jpg
1746  823269594223824897    https://pbs.twimg.com/media/C2kzTGxWEAEOpPL.jpg
1755  824796380199809024    https://pbs.twimg.com/media/CwiuEJmW8AAZnit.jpg
1789  829878982036299777    https://pbs.twimg.com/media/C3nygbBWQAAjwcW.jpg
1803  832040443403784192    https://pbs.twimg.com/media/Cq9guJ5WgAADfpF.jpg
1804  832215726631055365    https://pbs.twimg.com/media/CwJR1okWIAA6XMp.jpg
1858  841833993020538882  https://pbs.twimg.com/ext_tw_video_thumb/81742...
1864  842892208864923648  https://pbs.twimg.com/ext_tw_video_thumb/80710...
1903  851953902622658560    https://pbs.twimg.com/media/C4KHj-nWQAA3poV.jpg
1944  861769973181624320    https://pbs.twimg.com/media/CzG425nWgAAnP7P.jpg
1992  873697596434513921    https://pbs.twimg.com/media/DA7iHL5U0AA1OQo.jpg
2041  885311592912609280    https://pbs.twimg.com/media/C4bTH6nWMAAX_bJ.jpg
2055  888202515573088257    https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg

     img_num                  p1              p1_conf p1_dog  \
1297       1            upright             0.303415  False
1315       1                pug  0.27220500000000003   True
1333       2        cash_machine             0.802333  False
```

1345	1	Irish_terrier	0.25485599999999997	True
1349	1	Labrador_retriever	0.967397	True
1364	1	brown_bear	0.7132930000000001	False
1368	1	golden_retriever	0.586937	True
1387	1	toy_poodle	0.42046300000000003	True
1407	1	golden_retriever	0.843799	True
1417	3	Samoyed	0.978833	True
1427	1	dalmatian	0.556595	True
1446	1	golden_retriever	0.945523	True
1453	3	Chihuahua	0.346545	True
1456	1	golden_retriever	0.964929	True
1463	1	hippopotamus	0.581403	False
1476	1	pug	0.99731	True
1487	1	golden_retriever	0.383223	True
1495	1	Cardigan	0.611525	True
1510	1	golden_retriever	0.9938299999999999	True
1522	1	golden_retriever	0.735163	True
1538	1	tub	0.47947700000000004	False
1541	1	malamute	0.375098	True
1564	1	golden_retriever	0.705092	True
1569	1	cocker_spaniel	0.500509	True
1571	3	schipperke	0.363272	True
1579	1	golden_retriever	0.600276	True
1588	1	papillon	0.53318	True
1589	1	beagle	0.636169	True
1590	1	English_springer	0.403698	True
1591	1	chow	0.243529	True
...	...	...	...	...
1619	1	cocker_spaniel	0.253442	True
1624	1	Lakeland_terrier	0.530104	True
1627	1	chow	0.0903414	True
1634	1	Irish_setter	0.574557	True
1636	2	Cardigan	0.593858	True
1640	1	seat_belt	0.474292	False
1645	1	cocker_spaniel	0.74022	True
1652	1	Labrador_retriever	0.8616510000000001	True
1683	1	Labrador_retriever	0.427742	True
1693	1	English_setter	0.677408	True
1699	1	dishwasher	0.700466	False
1703	1	Tibetan_mastiff	0.506312	True
1712	1	Norwegian_elkhound	0.372202	True
1717	4	prison	0.907083	False
1718	1	standard_poodle	0.351308	True
1727	3	golden_retriever	0.938048	True
1736	1	Saint_Bernard	0.995143	True
1742	1	Samoyed	0.41676899999999995	True
1746	1	Samoyed	0.585441	True
1755	2	gas_pump	0.676439	False

1789	1	golden_retriever	0.6173890000000001	True
1803	1	miniature_pinscher	0.796313	True
1804	1	Afghan_hound	0.27463699999999996	True
1858	1	ice_bear	0.3362	False
1864	1	Chihuahua	0.50537	True
1903	1	Staffordshire_bullterrier	0.757547	True
1944	2	Arabian_camel	0.366248	False
1992	1	laptop	0.153718	False
2041	1	Labrador_retriever	0.908703	True
2055	2	Pembroke	0.809197	True

	p2	p2_conf	p2_dog	\
1297	golden_retriever	0.181351	True	
1315	bull_mastiff	0.25153000000000003	True	
1333	schipperke	0.0455186	True	
1345	briard	0.22771599999999997	True	
1349	golden_retriever	0.0166414	True	
1364	Indian_elephant	0.172844	False	
1368	Labrador_retriever	0.39826	True	
1387	miniature_poodle	0.13264	True	
1407	Labrador_retriever	0.0529559	True	
1417	Pomeranian	0.012763	True	
1427	whippet	0.15104700000000001	True	
1446	Labrador_retriever	0.0423191	True	
1453	dalmatian	0.166246	True	
1456	Labrador_retriever	0.0115837	True	
1463	doormat	0.152445	False	
1476	Brabancon_griffon	0.00118563	True	
1487	cocker_spaniel	0.16593	True	
1495	Pembroke	0.368566	True	
1510	cocker_spaniel	0.00314271	True	
1522	Sussex_spaniel	0.064897	True	
1538	bathtub	0.325106	False	
1541	jean	0.0693617	False	
1564	Labrador_retriever	0.219721	True	
1569	golden_retriever	0.27273400000000003	True	
1571	kelpie	0.197021	True	
1579	Labrador_retriever	0.140798	True	
1588	collie	0.192031	True	
1589	Labrador_retriever	0.119256	True	
1590	Brittany_spaniel	0.347609	True	
1591	hamster	0.22715	False	
...	...	...	...	
1619	golden_retriever	0.16285	True	
1624	Irish_terrier	0.19731400000000002	True	
1627	binoculars	0.08349880000000001	False	
1634	golden_retriever	0.339251	True	
1636	Shetland_sheepdog	0.13061099999999998	True	

1640	golden_retriever	0.17139300000000002	True
1645	Dandie_Dinmont	0.06160449999999999	True
1652	golden_retriever	0.0444618	True
1683	Great_Dane	0.190503	True
1693	Border_collie	0.052724	True
1699	golden_retriever	0.245773	True
1703	Tibetan_terrier	0.29569	True
1712	Chesapeake_Bay_retriever	0.13718699999999998	True
1717	palace	0.0200891	False
1718	toy_poodle	0.271929	True
1727	kuvasz	0.0251195	True
1736	Cardigan	0.00304359	True
1742	malamute	0.252706	True
1746	Pomeranian	0.193654	True
1755	harvester	0.049995300000000006	False
1789	Labrador_retriever	0.33705300000000005	True
1803	Chihuahua	0.15541300000000002	True
1804	borzoi	0.142204	True
1858	Samoyed	0.20135799999999998	True
1864	Pomeranian	0.12035799999999999	True
1903	American_Staffordshire_terrier	0.14995	True
1944	house_finch	0.209852	False
1992	French_bulldog	0.0999839	True
2041	seat_belt	0.0570909	False
2055	Rhodesian_ridgeback	0.05495	True

	p3	p3_conf	p3_dog
1297	Brittany_spaniel	0.16208399999999998	True
1315	bath_towel	0.116806	False
1333	German_shepherd	0.0233535	True
1345	soft-coated_wheaten_terrier	0.22326300000000002	True
1349	ice_bear	0.014857599999999999	False
1364	water_buffalo	0.0389022	False
1368	kuvasz	0.00540969	True
1387	Chesapeake_Bay_retriever	0.121523	True
1407	kelpie	0.035711099999999996	True
1417	Eskimo_dog	0.00185305	True
1427	American_Staffordshire_terrier	0.09643550000000001	True
1446	doormat	0.00395626	False
1453	toy_terrier	0.117502	True
1456	refrigerator	0.00749862	False
1463	sea_lion	0.026364299999999997	False
1476	French_bulldog	0.00042798900000000004	True
1487	Chesapeake_Bay_retriever	0.118199	True
1495	Chihuahua	0.00332957	True
1510	Great_Pyrenees	0.000917414	True
1522	Labrador_retriever	0.047703699999999995	True
1538	golden_retriever	0.07853049999999999	True

```
1541                      keeshond              0.0505276   True
1564                        kuvasz               0.015965   True
1569                  jigsaw_puzzle    0.04147580000000001  False
1571              Norwegian_elkhound              0.151024   True
1579                     seat_belt              0.0873548  False
1588                  Border_collie              0.121626   True
1589              golden_retriever    0.08254919999999999   True
1590          Welsh_springer_spaniel              0.137186   True
1591                     Pomeranian              0.0560567   True
...                             ...                   ...    ...
1619                     otterhound              0.110921   True
1624                       Airedale    0.08251460000000001   True
1627                   Irish_setter              0.0774556   True
1634                     seat_belt              0.0461082  False
1636                       Pembroke               0.100842   True
1640              Labrador_retriever    0.11059200000000001   True
1645                  English_setter    0.041331400000000004   True
1652       Staffordshire_bullterrier    0.016496700000000003   True
1683          curly-coated_retriever               0.146427   True
1693                  cocker_spaniel              0.0485719   True
1699                            chow    0.039011699999999996   True
1703                     otterhound              0.0362507   True
1712                       malamute    0.07143619999999999   True
1717                       umbrella             0.00784954  False
1718                 Tibetan_terrier              0.0947592   True
1727              Labrador_retriever              0.0229773   True
1736                English_springer             0.00104955   True
1742                          kuvasz               0.157028   True
1746                     Arctic_fox              0.0716476  False
1755                           swing              0.0446596  False
1789                     tennis_ball             0.00855442  False
1803       Staffordshire_bullterrier              0.0309433   True
1804                         doormat               0.109677  False
1858                      Eskimo_dog               0.186789   True
1864                     toy_terrier              0.0770081   True
1903        Chesapeake_Bay_retriever    0.047522699999999994   True
1944                  cocker_spaniel              0.0464032   True
1992                         printer              0.0771299  False
2041                             pug              0.0119335   True
2055                          beagle              0.0389148   True

[66 rows x 12 columns]
```

In [216]:
```
tweet_count.describe()
```

Out[216]:

	retweet_count	favorite_count
count	2345.000000	2345.000000
mean	3043.350959	8099.741151

```
std      5059.974621    12183.656173
min         0.000000        0.000000
25%       609.000000     1408.000000
50%      1424.000000     3553.000000
75%      3550.000000    10022.000000
max     77746.000000   143985.000000
```

```
In [217]: dup = tweet_count[tweet_count.id.duplicated()]
```

2.2.3 View some samples

```
In [218]: archive.sample(30)
```

```
Out[218]:
          tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
113   870726314365509632           8.707262e+17            16487760.0
430   821044531881721856                    NaN                   NaN
1676  682088079302213632                    NaN                   NaN
70    879008229531029506                    NaN                   NaN
316   834931633769889797                    NaN                   NaN
1921  674262580978937856                    NaN                   NaN
683   788412144018661376                    NaN                   NaN
2063  671159727754231808                    NaN                   NaN
706   785533386513321988                    NaN                   NaN
1493  692752401762250755                    NaN                   NaN
1927  674051556661161984                    NaN                   NaN
1105  734912297295085568                    NaN                   NaN
1138  728035342121635841                    NaN                   NaN
815   771004394259247104                    NaN                   NaN
539   806576416489959424                    NaN                   NaN
484   814638523311648768                    NaN                   NaN
375   828361771580813312                    NaN                   NaN
2153  669661792646373376                    NaN                   NaN
1186  718540630683709445                    NaN                   NaN
996   748337862848962560                    NaN                   NaN
1890  674767892831932416                    NaN                   NaN
392   826115272272650244                    NaN                   NaN
1485  693155686491000832                    NaN                   NaN
115   870374049280663552                    NaN                   NaN
1373  701981390485725185                    NaN                   NaN
329   833479644947025920                    NaN                   NaN
868   761750502866649088                    NaN                   NaN
922   756275833623502848                    NaN                   NaN
1578  687317306314240000                    NaN                   NaN
515   811386762094317568                    NaN                   NaN

                      timestamp  \
113   2017-06-02 19:38:25 +0000
430   2017-01-16 17:20:45 +0000
```

```
1676  2015-12-30 06:37:25 +0000
70    2017-06-25 16:07:47 +0000
316   2017-02-24 01:03:08 +0000
1921  2015-12-08 16:21:41 +0000
683   2016-10-18 16:11:17 +0000
2063  2015-11-30 02:52:03 +0000
706   2016-10-10 17:32:08 +0000
1493  2016-01-28 16:53:37 +0000
1927  2015-12-08 02:23:09 +0000
1105  2016-05-24 01:02:00 +0000
1138  2016-05-05 01:35:26 +0000
815   2016-08-31 15:19:06 +0000
539   2016-12-07 19:09:37 +0000
484   2016-12-30 01:05:33 +0000
375   2017-02-05 21:56:51 +0000
2153  2015-11-25 23:39:47 +0000
1186  2016-04-08 20:46:50 +0000
996   2016-06-30 02:10:24 +0000
1890  2015-12-10 01:49:36 +0000
392   2017-01-30 17:10:04 +0000
1485  2016-01-29 19:36:08 +0000
115   2017-06-01 20:18:38 +0000
1373  2016-02-23 04:06:20 +0000
329   2017-02-20 00:53:27 +0000
868   2016-08-06 02:27:27 +0000
922   2016-07-21 23:53:04 +0000
1578  2016-01-13 16:56:30 +0000
515   2016-12-21 01:44:13 +0000

                                                    source  \
113   <a href="http://twitter.com/download/iphone" r...
430   <a href="http://twitter.com/download/iphone" r...
1676  <a href="http://vine.co" rel="nofollow">Vine -...
70    <a href="http://twitter.com/download/iphone" r...
316   <a href="http://twitter.com/download/iphone" r...
1921  <a href="http://twitter.com/download/iphone" r...
683   <a href="http://twitter.com/download/iphone" r...
2063  <a href="http://twitter.com/download/iphone" r...
706   <a href="http://twitter.com/download/iphone" r...
1493  <a href="http://twitter.com/download/iphone" r...
1927  <a href="http://twitter.com/download/iphone" r...
1105  <a href="http://twitter.com/download/iphone" r...
1138  <a href="http://twitter.com/download/iphone" r...
815   <a href="http://twitter.com/download/iphone" r...
539   <a href="http://twitter.com/download/iphone" r...
484   <a href="http://twitter.com/download/iphone" r...
375   <a href="http://twitter.com" rel="nofollow">Tw...
2153  <a href="http://twitter.com/download/iphone" r...
```

```
1186  <a href="http://twitter.com/download/iphone" r...
996   <a href="http://vine.co" rel="nofollow">Vine -...
1890  <a href="http://twitter.com/download/iphone" r...
392   <a href="http://twitter.com/download/iphone" r...
1485  <a href="http://twitter.com/download/iphone" r...
115   <a href="http://twitter.com/download/iphone" r...
1373  <a href="http://twitter.com/download/iphone" r...
329   <a href="http://twitter.com/download/iphone" r...
868   <a href="http://twitter.com/download/iphone" r...
922   <a href="http://twitter.com/download/iphone" r...
1578  <a href="http://twitter.com/download/iphone" r...
515   <a href="http://twitter.com/download/iphone" r...

                                                  text  retweeted_status_id  \
113   @ComplicitOwl @ShopWeRateDogs &gt;10/10 is res...                   NaN
430   This is Flash. He went way too hard celebratin...                   NaN
1676  I'm not sure what this dog is doing but it's p...                   NaN
70    This is Beau. That is Beau's balloon. He takes...                   NaN
316   This is Tucker. He decided it was time to part...                   NaN
1921  This is Gus. He's super stoked about being an ...                   NaN
683   This is Dexter. He breaks hearts for a living...                   NaN
2063  This is Anthony. He just finished up his maste...                   NaN
706   This is Dallas. Her tongue is ridiculous. 11/1...                   NaN
1493  "Hello yes could I get one pupper to go please...                   NaN
1927  This is Lucy. She knits. Specializes in tobogg...                   NaN
1105  This is Jax. He's a literal fluffball. Sneaky ...                   NaN
1138  This is all I want in my life. 12/10 for super...                   NaN
815   RT @katieornah: @dog_rates learning a lot at c...         7.710021e+17
539   Hooman catch successful. Massive hit by dog. F...                   NaN
484   This is Olivia. She's a passionate advocate of...                   NaN
375   Beebop and Doobert should start a band 12/10 w...                   NaN
2153  This is a brave dog. Excellent free climber. T...                   NaN
1186  Get you a pup that can do both. 10/10 https://...                   NaN
996   SWIM AWAY PUPPER SWIM AWAY 13/10 #BarkWeek  ht...                   NaN
1890  This pup was carefully tossed to make it look ...                   NaN
392   This is Ike. He's demonstrating the pupmost re...                   NaN
1485  This is Dunkin. He can only see when he's wet ...                   NaN
115   This is Zoey. She really likes the planet. Wou...                   NaN
1373  This is Fiji. She's a Powdered Stegafloof. Ver...                   NaN
329   This is Poppy. She just arrived. 13/10 would s...                   NaN
868   RT @dog_rates: "Tristan do not speak to me wit...         6.853251e+17
922   When ur older siblings get to play in the deep...                   NaN
1578  This is Tyrone. He's a leaf wizard. Self-motiv...                   NaN
515   This is Craig. That's actually a normal sized ...                   NaN

      retweeted_status_user_id retweeted_status_timestamp  \
113                        NaN                        NaN
430                        NaN                        NaN
```

```
1676                        NaN                          NaN
70                          NaN                          NaN
316                         NaN                          NaN
1921                        NaN                          NaN
683                         NaN                          NaN
2063                        NaN                          NaN
706                         NaN                          NaN
1493                        NaN                          NaN
1927                        NaN                          NaN
1105                        NaN                          NaN
1138                        NaN                          NaN
815               1.732729e+09  2016-08-31 15:10:07 +0000
539                         NaN                          NaN
484                         NaN                          NaN
375                         NaN                          NaN
2153                        NaN                          NaN
1186                        NaN                          NaN
996                         NaN                          NaN
1890                        NaN                          NaN
392                         NaN                          NaN
1485                        NaN                          NaN
115                         NaN                          NaN
1373                        NaN                          NaN
329                         NaN                          NaN
868               4.196984e+09  2016-01-08 05:00:14 +0000
922                         NaN                          NaN
1578                        NaN                          NaN
515                         NaN                          NaN

                                              expanded_urls  rating_numerator  \
113                                                     NaN                10
430   https://twitter.com/dog_rates/status/821044531...                12
1676                      https://vine.co/v/iqMjlxULzbn                12
70    https://twitter.com/dog_rates/status/879008229...                13
316   https://twitter.com/dog_rates/status/834931633...                12
1921  https://twitter.com/dog_rates/status/674262580...                 9
683   https://twitter.com/dog_rates/status/788412144...                11
2063  https://twitter.com/dog_rates/status/671159727...                 5
706   https://twitter.com/dog_rates/status/785533386...                11
1493  https://twitter.com/dog_rates/status/692752401...                13
1927  https://twitter.com/dog_rates/status/674051556...                10
1105  https://twitter.com/dog_rates/status/734912297...                10
1138  https://twitter.com/dog_rates/status/728035342...                12
815   https://twitter.com/katieornah/status/77100213...                12
539   https://twitter.com/deadspin/status/8065709331...                13
484   https://twitter.com/dog_rates/status/814638523...                12
375                                                     NaN                12
2153  https://twitter.com/dog_rates/status/669661792...                 5
```

```
1186  https://twitter.com/dog_rates/status/718540630...               10
996                        https://vine.co/v/h5aDaFthX6O               13
1890  https://twitter.com/dog_rates/status/674767892...               12
392   https://twitter.com/dog_rates/status/826115272...               13
1485  https://twitter.com/dog_rates/status/693155686...               12
115   https://twitter.com/dog_rates/status/870374049...               13
1373  https://twitter.com/dog_rates/status/701981390...               12
329   https://twitter.com/dog_rates/status/833479644...               13
868   https://twitter.com/dog_rates/status/685325112...               10
922   https://twitter.com/dog_rates/status/756275833...               10
1578  https://twitter.com/dog_rates/status/687317306...               11
515   https://twitter.com/dog_rates/status/811386762...               11

      rating_denominator     name doggo floofer  pupper  puppo
113                   10     None  None    None    None   None
430                   10    Flash  None    None    None   None
1676                  10     None  None    None    None   None
70                    10     Beau  None    None    None   None
316                   10   Tucker  None    None    None   None
1921                  10      Gus  None    None  pupper   None
683                   10   Dexter  None    None    None   None
2063                  10  Anthony  None    None    None   None
706                   10   Dallas  None    None    None   None
1493                  10     None  None    None  pupper   None
1927                  10     Lucy  None    None    None   None
1105                  10      Jax  None    None    None   None
1138                  10      all  None    None  pupper   None
815                   10     None  None    None  pupper   None
539                   10     None  None    None    None   None
484                   10   Olivia  None    None    None   None
375                   10     None  None    None    None   None
2153                  10        a  None    None    None   None
1186                  10     None  None    None    None   None
996                   10     None  None    None  pupper   None
1890                  10     None  None    None    None   None
392                   10      Ike  None    None    None   None
1485                  10   Dunkin  None    None    None   None
115                   10     Zoey  None    None    None   None
1373                  10     Fiji  None    None    None   None
329                   10    Poppy  None    None    None   None
868                   10     None  None    None    None   None
922                   10     None  None    None    None  puppo
1578                  10   Tyrone  None    None    None   None
515                   10    Craig  None    None  pupper   None
```

In [219]:
```
image_prediction.sample(30)
```

Out[219]:
```
                tweet_id                                          jpg_url  \
293   671347597085433856    https://pbs.twimg.com/media/CVEbFDRWsAAkN_7.jpg
```

1969	868622495443632128	https://pbs.twimg.com/media/DA33i0XXsAEQtCA.jpg
1067	715733265223708672	https://pbs.twimg.com/media/Ce7LlUeUUAEQkQl.jpg
295	671357843010908160	https://pbs.twimg.com/media/CVEkZaPXIAEw5vr.jpg
1320	756288534030475264	https://pbs.twimg.com/media/Cn7gaHrWIAAZJMt.jpg
142	668614819948453888	https://pbs.twimg.com/media/CUdloW8WEAAxB_Y.jpg
519	676496375194980353	https://pbs.twimg.com/media/CWNl3S9WcAARN34.jpg
1849	839990271299457024	https://pbs.twimg.com/media/C6g-sX-VsAAHfJ9.jpg
11	666071193221509120	https://pbs.twimg.com/media/CT5cN_3WEAAlOoZ.jpg
1474	780459368902959104	https://pbs.twimg.com/media/CtS_p9kXEAE2nh8.jpg
1536	790581949425475584	https://pbs.twimg.com/media/Cvi2FiKWgAAif1u.jpg
267	670807719151067136	https://pbs.twimg.com/media/CU8v-rdXIAId12Z.jpg
1432	773308824254029826	https://pbs.twimg.com/media/CrtYRMEWIAAUkCl.jpg
1668	813051746834595840	https://pbs.twimg.com/media/C0iKPZIXUAAbDYV.jpg
1853	840696689258311684	https://pbs.twimg.com/media/C6rBLenU0AAr8MN.jpg
1681	813812741911748608	https://pbs.twimg.com/media/C0s-XtzWgAAp1W-.jpg
913	700864154249383937	https://pbs.twimg.com/media/Cbn4OqKWwAADGWt.jpg
1252	747963614829678593	https://pbs.twimg.com/media/CmFM7ngXEAEitfh.jpg
566	678334497360859136	https://pbs.twimg.com/media/CWntoDVWcAEl3NB.jpg
2006	877611172832227328	https://pbs.twimg.com/media/DCszHgmW0AAmIpT.jpg
547	677331501395156992	https://pbs.twimg.com/media/CWZdaGxXAAAjGjb.jpg
1923	857029823797047296	https://pbs.twimg.com/media/C-TIEwMW0AEjb55.jpg
1751	824297048279236611	https://pbs.twimg.com/media/C3B9ypNWEAM1bVs.jpg
161	668932921458302977	https://pbs.twimg.com/media/CUiG6_ZXAAAPaw_.jpg
657	682303737705140231	https://pbs.twimg.com/media/CXgHoLnWAAA8i52.jpg
789	690597161306841088	https://pbs.twimg.com/media/CZV-c9NVIAEWtiU.jpg
1306	753398408988139520	https://pbs.twimg.com/ext_tw_video_thumb/75339...
1756	825026590719483904	https://pbs.twimg.com/media/C3MVTeHWcAAGNfx.jpg
291	671182547775299584	https://pbs.twimg.com/media/CVCE9uYXIAEtSzR.jpg
1979	870804317367881728	https://pbs.twimg.com/media/DBW35ZsVoAEWZUU.jpg

	img_num	p1	p1_conf	p1_dog	\
293	1	picket_fence	0.382918	False	
1969	1	Labrador_retriever	0.868107	True	
1067	1	Dandie_Dinmont	0.740229	True	
295	1	Italian_greyhound	0.831757	True	
1320	3	conch	0.9256209999999999	False	
142	1	bustard	0.380772	False	
519	1	pug	0.9853870000000001	True	
1849	2	Staffordshire_bullterrier	0.604938	True	
11	1	Gordon_setter	0.503672	True	
1474	1	Great_Dane	0.38249099999999997	True	
1536	2	refrigerator	0.998886	False	
267	1	Old_English_sheepdog	0.9580350000000001	True	
1432	1	shopping_cart	0.572349	False	
1668	1	golden_retriever	0.9148040000000001	True	
1853	1	web_site	0.8417680000000001	False	
1681	1	French_bulldog	0.709146	True	
913	1	kuvasz	0.8058569999999999	True	

```
1252       1                    kelpie             0.307672   True
566        1            Norfolk_terrier             0.378643   True
2006       1               Irish_setter  0.36472899999999997   True
547        1                     beagle  0.31346399999999996   True
1923       2           golden_retriever             0.968623   True
1751       2                     teddy              0.58823  False
161        1            standard_poodle  0.23763800000000002   True
657        1                  seat_belt             0.997659  False
789        1                      Lhasa            0.0974997   True
1306       1                    whippet             0.163794   True
1756       2                 Eskimo_dog             0.524454   True
291        1                 Rottweiler             0.331179   True
1979       1               home_theater  0.16829000000000002  False

                                p2              p2_conf p2_dog  \
293                    rain_barrel   0.10880899999999999  False
1969                 Great_Pyrenees             0.060973   True
1067               miniature_poodle            0.0819151   True
295                    toy_terrier  0.043305800000000005   True
1320                French_bulldog  0.032492200000000006   True
142                        pelican   0.10055399999999999  False
519             Norwegian_elkhound            0.0044169   True
1849  American_Staffordshire_terrier              0.31154   True
11                Yorkshire_terrier             0.174201   True
1474                German_shepherd             0.312026   True
1536                       malinois          0.000152999   True
267               Sealyham_terrier            0.0138922   True
1432            Labrador_retriever   0.15140599999999999   True
1668            Labrador_retriever              0.08355   True
1853                          rule           0.00708731  False
1681                   Boston_bull             0.247621   True
913                  Great_Pyrenees             0.187272   True
1252                  Irish_terrier             0.197486   True
566                golden_retriever            0.0955939   True
2006               golden_retriever             0.202907   True
547                           boxer   0.21850300000000003   True
1923            Labrador_retriever  0.010325200000000001   True
1751                 jigsaw_puzzle            0.0289096  False
161            Old_English_sheepdog             0.195573   True
657                Lakeland_terrier           0.00173092   True
789                           koala            0.0919339  False
1306              Italian_greyhound             0.157192   True
1756                 Siberian_husky             0.467678   True
291                           kelpie   0.21860100000000002   True
1979                         sandbar            0.0980404  False

                  p3              p3_conf p3_dog
293       plastic_bag            0.0388782  False
```

1969	Saint_Bernard	0.0334889	True
1067	toy_poodle	0.0637485	True
295	Chihuahua	0.036773	True
1320	tiger_cat	0.00667908	False
142	crane	0.0847135	False
519	French_bulldog	0.00389287	True
1849	Boston_bull	0.0371591	True
11	Pekinese	0.109454	True
1474	bull_mastiff	0.0332719	True
1536	kelpie	0.000130817	True
267	Border_collie	0.00460114	True
1432	shopping_basket	0.107102	False
1668	kuvasz	0.000453224	True
1853	envelope	0.0068203000000000005	False
1681	boxer	0.0188551	True
913	Samoyed	0.0034909	True
1252	dingo	0.105475	False
566	kelpie	0.0853092	True
2006	Irish_terrier	0.107473	True
547	French_bulldog	0.10646199999999999	True
1923	Saluki	0.00414842	True
1751	doormat	0.0222507	False
161	toy_poodle	0.14465799999999998	True
657	Airedale	0.00020361	True
789	sunglasses	0.0915048	False
1306	English_foxhound	0.142995	True
1756	malamute	0.00497584	True
291	Appenzeller	0.18252000000000002	True
1979	television	0.07972939999999999	False

In [220]: `tweet_count.sample(30)`

Out[220]:

	id	retweet_count	favorite_count
2240	667806454573760512	524	1088
101	872486979161796608	9224	41034
1644	683391852557561860	2636	8187
402	823269594223824897	11100	0
666	789314372632018944	2609	9466
1280	708119489313951744	1074	2881
653	790946055508652032	5339	18258
1754	678410210315247616	2011	4543
1309	706346369204748288	1013	3697
656	790581949425475584	8127	22726
289	837366284874571778	5859	22677
1783	677328882937298944	1643	3870
703	784183165795655680	9142	22104
1636	683828599284170753	1164	3000
1977	672834301050937345	613	1371

```
279   838831947270979586          12269              0
2123  670069087419133954            263            663
1524  689977555533848577            491           1465
453   817502432452313088           3823              0
893   758355060040593408           1209           3723
1367  701570477911896070           1030           3027
2342  666033412701032449             45            125
742   778990705243029504           8233          21933
1002  747219827526344708           1744           5698
2170  668994913074286592            249            459
276   838952994649550848           4398          21022
576   799774291445383169           5334              0
2061  671109016219725825            468           1193
132   866334964761202691          15091          53758
2203  668544745690562560            241            544
```

In [221]:
```
image_prediction.jpg_url.count()
```

Out[221]:
```
2075
```

In [222]:
```
archive.name.value_counts()
```

Out[222]:
```
None         745
a             55
Charlie       12
Oliver        11
Lucy          11
Cooper        11
Penny         10
Tucker        10
Lola          10
Winston        9
Bo             9
the            8
Sadie          8
Toby           7
Buddy          7
Daisy          7
an             7
Bailey         7
Jax            6
Bella          6
Stanley        6
Scout          6
Rusty          6
Leo            6
Jack           6
Oscar          6
Koda           6
```

```
Milo           6
Dave           6
Alfie          5
              ...
Joshwa         1
Dallas         1
Kawhi          1
Shelby         1
this           1
Rueben         1
Akumi          1
Fwed           1
Jebberson      1
Shawwn         1
Berb           1
Flash          1
Eazy           1
Timber         1
Kayla          1
Todo           1
Lulu           1
Shadoe         1
Horace         1
Ronduh         1
Benny          1
Coopson        1
Sid            1
Comet          1
Bradley        1
Strudel        1
Dot            1
Pubert         1
Aqua           1
Ralph          1
Name: name, Length: 957, dtype: int64
```

examine numerator

```
In [223]: archive[archive["rating_numerator"] == 5]
```

```
Out[223]:                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
          45    883482846933004288                    NaN                  NaN
          730   781661882474196992                    NaN                  NaN
          956   751583847268179968                    NaN                  NaN
          1399  699691744225525762                    NaN                  NaN
          1461  694925794720792577                    NaN                  NaN
          1508  691483041324204033                    NaN                  NaN
          1583  687102708889812993                    NaN                  NaN
```

1618	684969860808454144	6.849598e+17	4.196984e+09
1619	684959798585110529	NaN	NaN
1624	684880619965411328	NaN	NaN
1645	683849932751646720	NaN	NaN
1680	682003177596559360	NaN	NaN
1689	681340665377193984	6.813394e+17	4.196984e+09
1727	679877062409191424	NaN	NaN
1796	677301033169788928	NaN	NaN
1808	676897532954456065	NaN	NaN
1820	676588346097852417	NaN	NaN
1861	675483430902214656	NaN	NaN
1874	675135153782571009	NaN	NaN
1901	674646392044941312	NaN	NaN
1904	674632714662858753	NaN	NaN
1925	674063288070742018	NaN	NaN
1979	672980819271634944	NaN	NaN
2013	672231046314901505	NaN	NaN
2026	671879137494245376	NaN	NaN
2063	671159727754231808	NaN	NaN
2092	670782429121134593	NaN	NaN
2109	670449342516494336	NaN	NaN
2134	670069087419133954	NaN	NaN
2139	670037189829525505	NaN	NaN
2153	669661792646373376	NaN	NaN
2181	668994913074286592	NaN	NaN
2206	668631377374486528	NaN	NaN
2242	667911425562669056	NaN	NaN
2260	667550882905632768	NaN	NaN
2312	666776908487630848	NaN	NaN
2351	666049248165822465	NaN	NaN

	timestamp \
45	2017-07-08 00:28:19 +0000
730	2016-09-30 01:08:10 +0000
956	2016-07-09 01:08:47 +0000
1399	2016-02-16 20:28:06 +0000
1461	2016-02-03 16:49:55 +0000
1508	2016-01-25 04:49:38 +0000
1583	2016-01-13 02:43:46 +0000
1618	2016-01-07 05:28:35 +0000
1619	2016-01-07 04:48:36 +0000
1624	2016-01-06 23:33:58 +0000
1645	2016-01-04 03:18:23 +0000
1680	2015-12-30 01:00:03 +0000
1689	2015-12-28 05:07:27 +0000
1727	2015-12-24 04:11:37 +0000
1796	2015-12-17 01:35:24 +0000
1808	2015-12-15 22:52:02 +0000

```
1820  2015-12-15 02:23:26 +0000
1861  2015-12-12 01:12:54 +0000
1874  2015-12-11 02:08:58 +0000
1901  2015-12-09 17:46:48 +0000
1904  2015-12-09 16:52:27 +0000
1925  2015-12-08 03:09:46 +0000
1979  2015-12-05 03:28:25 +0000
2013  2015-12-03 01:49:05 +0000
2026  2015-12-02 02:30:43 +0000
2063  2015-11-30 02:52:03 +0000
2092  2015-11-29 01:52:48 +0000
2109  2015-11-28 03:49:14 +0000
2134  2015-11-27 02:38:14 +0000
2139  2015-11-27 00:31:29 +0000
2153  2015-11-25 23:39:47 +0000
2181  2015-11-24 03:29:51 +0000
2206  2015-11-23 03:25:17 +0000
2242  2015-11-21 03:44:27 +0000
2260  2015-11-20 03:51:47 +0000
2312  2015-11-18 00:36:17 +0000
2351  2015-11-16 00:24:50 +0000

                                                    source  \
45    <a href="http://twitter.com/download/iphone" r...
730   <a href="http://twitter.com/download/iphone" r...
956   <a href="http://twitter.com/download/iphone" r...
1399  <a href="http://twitter.com/download/iphone" r...
1461  <a href="http://vine.co" rel="nofollow">Vine -...
1508  <a href="http://twitter.com/download/iphone" r...
1583  <a href="http://twitter.com/download/iphone" r...
1618  <a href="http://twitter.com/download/iphone" r...
1619  <a href="http://twitter.com/download/iphone" r...
1624  <a href="http://twitter.com/download/iphone" r...
1645  <a href="http://twitter.com/download/iphone" r...
1680  <a href="http://twitter.com/download/iphone" r...
1689  <a href="http://twitter.com/download/iphone" r...
1727  <a href="http://twitter.com/download/iphone" r...
1796  <a href="http://twitter.com/download/iphone" r...
1808  <a href="http://twitter.com/download/iphone" r...
1820  <a href="http://twitter.com/download/iphone" r...
1861  <a href="http://twitter.com/download/iphone" r...
1874  <a href="http://twitter.com/download/iphone" r...
1901  <a href="http://twitter.com/download/iphone" r...
1904  <a href="http://twitter.com/download/iphone" r...
1925  <a href="http://twitter.com/download/iphone" r...
1979  <a href="http://twitter.com/download/iphone" r...
2013  <a href="http://twitter.com/download/iphone" r...
2026  <a href="http://twitter.com/download/iphone" r...
```

```
2063  <a href="http://twitter.com/download/iphone" r...
2092  <a href="http://twitter.com/download/iphone" r...
2109  <a href="http://twitter.com/download/iphone" r...
2134  <a href="http://twitter.com/download/iphone" r...
2139  <a href="http://twitter.com/download/iphone" r...
2153  <a href="http://twitter.com/download/iphone" r...
2181  <a href="http://twitter.com/download/iphone" r...
2206  <a href="http://twitter.com/download/iphone" r...
2242  <a href="http://twitter.com/download/iphone" r...
2260  <a href="http://twitter.com" rel="nofollow">Tw...
2312  <a href="http://twitter.com/download/iphone" r...
2351  <a href="http://twitter.com/download/iphone" r...

                                                 text  retweeted_status_id  \
45    This is Bella. She hopes her smile made you sm...                  NaN
730   Who keeps sending in pictures without dogs in ...                  NaN
956   Please stop sending it pictures that don't eve...                  NaN
1399  This is Dave. He's a tropical pup. Short lil l...                  NaN
1461  Please only send in dogs. This t-rex is very s...                  NaN
1508  When bae says they can't go out but you see th...                  NaN
1583  Army of water dogs here. None of them know whe...                  NaN
1618  For those who claim this is a goat, u are wron...                  NaN
1619  This is Jerry. He's a neat dog. No legs (tragi...                  NaN
1624  Here we have a basking dino pupper. Looks powe...                  NaN
1645  This is Jiminy. He's not the brightest dog. Ne...                  NaN
1680  Unique dog here. Wrinkly as hell. Weird segmen...                  NaN
1689  I've been told there's a slight possibility he...                  NaN
1727  Meet Penelope. She's a bacon frise. Total babe...                  NaN
1796  This is Juckson. He's totally on his way to a ...                  NaN
1808  Exotic handheld dog here. Appears unathletic. ...                  NaN
1820  This is Bubbles. He kinda resembles a fish. Al...                  NaN
1861  Rare shielded battle dog here. Very happy abou...                  NaN
1874  This is Steven. He got locked outside. Damn it...                  NaN
1901  Two gorgeous dogs here. Little waddling dog is...                  NaN
1904  Rare submerged pup here. Holds breath for a lo...                  NaN
1925  This is Earl. Earl is lost. Someone help Earl...                   NaN
1979  Extraordinary dog here. Looks large. Just a he...                  NaN
2013  Exotic underwater dog here. Very shy. Wont ret...                  NaN
2026  This is Brad. He's a chubby lil pup. Doesn't r...                  NaN
2063  This is Anthony. He just finished up his maste...                  NaN
2092  This dude slaps your girl's ass what do you do...                  NaN
2109  Vibrant dog here. Fabulous tail. Only 2 legs t...                  NaN
2134  This is Randall. He's from Chernobyl. Built pl...                  NaN
2139  Awesome dog here. Not sure where it is tho. Sp...                  NaN
2153  This is a brave dog. Excellent free climber. T...                  NaN
2181  Two gorgeous pups here. Both have cute fake ho...                  NaN
2206  Meet Zeek. He is a grey Cumulonimbus. Zeek is ...                  NaN
2242  Wow. Armored dog here. Ready for battle. Face ...                  NaN
```

```
2260  RT @dogratingrating: Unoriginal idea. Blatant ...         6.675484e+17
2312  This is Josep. He is a Rye Manganese mix. Can ...                  NaN
2351  Here we have a 1949 1st generation vulpix. Enj...                  NaN

      retweeted_status_user_id retweeted_status_timestamp  \
45                         NaN                         NaN
730                        NaN                         NaN
956                        NaN                         NaN
1399                       NaN                         NaN
1461                       NaN                         NaN
1508                       NaN                         NaN
1583                       NaN                         NaN
1618                       NaN                         NaN
1619                       NaN                         NaN
1624                       NaN                         NaN
1645                       NaN                         NaN
1680                       NaN                         NaN
1689                       NaN                         NaN
1727                       NaN                         NaN
1796                       NaN                         NaN
1808                       NaN                         NaN
1820                       NaN                         NaN
1861                       NaN                         NaN
1874                       NaN                         NaN
1901                       NaN                         NaN
1904                       NaN                         NaN
1925                       NaN                         NaN
1979                       NaN                         NaN
2013                       NaN                         NaN
2026                       NaN                         NaN
2063                       NaN                         NaN
2092                       NaN                         NaN
2109                       NaN                         NaN
2134                       NaN                         NaN
2139                       NaN                         NaN
2153                       NaN                         NaN
2181                       NaN                         NaN
2206                       NaN                         NaN
2242                       NaN                         NaN
2260              4.296832e+09  2015-11-20 03:41:59 +0000
2312                       NaN                         NaN
2351                       NaN                         NaN

                                          expanded_urls  rating_numerator  \
45    https://twitter.com/dog_rates/status/883482846...                 5
730   https://twitter.com/dog_rates/status/781661882...                 5
956   https://twitter.com/dog_rates/status/751583847...                 5
1399  https://twitter.com/dog_rates/status/699691744...                 5
```

```
1461                      https://vine.co/v/iJvUqWQ166L                5
1508  https://twitter.com/dog_rates/status/691483041...                5
1583  https://twitter.com/dog_rates/status/687102708...                5
1618                                                 NaN               5
1619  https://twitter.com/dog_rates/status/684959798...                5
1624  https://twitter.com/dog_rates/status/684880619...                5
1645  https://twitter.com/dog_rates/status/683849932...                5
1680  https://twitter.com/dog_rates/status/682003177...                5
1689                                                 NaN               5
1727  https://twitter.com/dog_rates/status/679877062...                5
1796  https://twitter.com/dog_rates/status/677301033...                5
1808  https://twitter.com/dog_rates/status/676897532...                5
1820  https://twitter.com/dog_rates/status/676588346...                5
1861  https://twitter.com/dog_rates/status/675483430...                5
1874  https://twitter.com/dog_rates/status/675135153...                5
1901  https://twitter.com/dog_rates/status/674646392...                5
1904  https://twitter.com/dog_rates/status/674632714...                5
1925  https://twitter.com/dog_rates/status/674063288...                5
1979  https://twitter.com/dog_rates/status/672980819...                5
2013  https://twitter.com/dog_rates/status/672231046...                5
2026  https://twitter.com/dog_rates/status/671879137...                5
2063  https://twitter.com/dog_rates/status/671159727...                5
2092  https://twitter.com/dog_rates/status/670782429...                5
2109  https://twitter.com/dog_rates/status/670449342...                5
2134  https://twitter.com/dog_rates/status/670069087...                5
2139  https://twitter.com/dog_rates/status/670037189...                5
2153  https://twitter.com/dog_rates/status/669661792...                5
2181  https://twitter.com/dog_rates/status/668994913...                5
2206  https://twitter.com/dog_rates/status/668631377...                5
2242  https://twitter.com/dog_rates/status/667911425...                5
2260  https://twitter.com/dogratingrating/status/667...                5
2312  https://twitter.com/dog_rates/status/666776908...                5
2351  https://twitter.com/dog_rates/status/666049248...                5

      rating_denominator      name  doggo floofer  pupper puppo
45                    10     Bella   None    None    None  None
730                   10      None   None    None    None  None
956                   10      None  doggo    None  pupper  None
1399                  10      Dave   None    None    None  None
1461                  10      None   None    None    None  None
1508                  10      None   None    None    None  None
1583                  10      None   None    None    None  None
1618                  10      None   None    None    None  None
1619                  10     Jerry   None    None    None  None
1624                  10      None   None    None  pupper  None
1645                  10    Jiminy   None    None    None  None
1680                  10      None   None    None    None  None
1689                  10      None   None    None    None  None
```

1727	10	Penelope	None	None	None	None
1796	10	Juckson	None	None	None	None
1808	10	None	None	None	None	None
1820	10	Bubbles	None	None	None	None
1861	10	None	None	None	None	None
1874	10	Steven	None	None	None	None
1901	10	None	None	None	None	None
1904	10	None	None	None	None	None
1925	10	Earl	None	None	None	None
1979	10	None	None	None	None	None
2013	10	None	None	None	None	None
2026	10	Brad	None	None	None	None
2063	10	Anthony	None	None	None	None
2092	10	None	None	None	None	None
2109	10	None	None	None	None	None
2134	10	Randall	None	None	None	None
2139	10	None	None	None	None	None
2153	10	a	None	None	None	None
2181	10	None	None	None	None	None
2206	10	Zeek	None	None	None	None
2242	10	None	None	None	None	None
2260	10	None	None	None	None	None
2312	10	Josep	None	None	None	None
2351	10	None	None	None	None	None

```
In [224]: text_list = archive.text.str.split(r"[ .\n](?=http)", expand = True)[0]

In [225]: text_list[45]

Out[225]: 'This is Bella. She hopes her smile made you smile. If not, she is also offering you
```

Quality

'archive' table

- Incorrect names (a, an, the, my)
- Erronous data types (rating_numerator, tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id columns, retweeted_status_timestamp and timestamp)
- Retweets
- Bad representation for missing tweet_links
- Wrong decimal numerators

'image_prediction' table

- Non descriptive column headers(p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog)
- Multiple records for the same jpg_url
- snake case dog predictions
- Missing images

'tweet_count' table

Tidiness

'archive' table

- Text and twitter link in the same column (text column)
- doggo, floofer, pupper, puppo columns are variable names

'image_prediction' table These information should be in the archive table

'tweet_count' table

- This is an unuseful table

2.3 Clean

```
In [226]: # Copy all tables
          archive_clean = archive.copy()
          image_prediction_clean = image_prediction.copy()
          tweet_count_clean = tweet_count.copy()
```

Define Fill missing tweet links with None to have good missing value representations

Code

```
In [227]: # Gather missing tweet source ids
          archive_clean["expanded_urls"] = archive_clean["expanded_urls"].fillna("None")
```

Test

```
In [228]: archive_clean["expanded_urls"].value_counts()
```

```
Out[228]: None
          https://twitter.com/dog_rates/status/773308824254029826/photo/1
          https://twitter.com/dog_rates/status/810254108431155201/photo/1
          https://twitter.com/dog_rates/status/841077006473256960/photo/1
          https://www.gofundme.com/help-lorenzo-beat-cancer,https://twitter.com/dog_rates/statu
          https://twitter.com/dog_rates/status/786709082849828864/photo/1
          https://twitter.com/dog_rates/status/683391852557561860/photo/1
          https://twitter.com/dog_rates/status/786233965241827333/photo/1
          https://twitter.com/dog_rates/status/832369877331693569/photo/1
          https://www.gofundme.com/lolas-life-saving-surgery-funds,https://twitter.com/dog_rate
          https://twitter.com/dog_rates/status/673295268553605120/photo/1
          https://twitter.com/dog_rates/status/820749716845686786/photo/1,https://twitter.com/d
          https://twitter.com/dog_rates/status/694669722378485760/photo/1,https://twitter.com/d
          https://twitter.com/dog_rates/status/750719632563142656/photo/1
          https://twitter.com/dog_rates/status/839549326359670784/photo/1
```

```
https://twitter.com/dog_rates/status/771380798096281600/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/786963064373534720/photo/1
https://twitter.com/dog_rates/status/740373189193256964/photo/1,https://twitter.com/d
https://www.gofundme.com/surgeryforjacktheminpin,https://twitter.com/dog_rates/status
https://www.gofundme.com/help-my-baby-sierra-get-better,https://twitter.com/dog_rates
https://www.gofundme.com/3ti3nps,https://twitter.com/dog_rates/status/86855227852483
https://twitter.com/dog_rates/status/844704788403113984/photo/1
https://twitter.com/dog_rates/status/667866724293877760/photo/1
https://twitter.com/dog_rates/status/774314403806253056/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/718631497683582976/photo/1
https://twitter.com/dog_rates/status/768193404517830656/photo/1
https://www.gofundme.com/servicedogoliver,https://twitter.com/dog_rates/status/81995
https://twitter.com/dog_rates/status/780931614150983680/photo/1
https://twitter.com/dog_rates/status/878057613040115712/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/700143752053182464/photo/1

https://twitter.com/dog_rates/status/862096992088072192/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/875021211251597312/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/825535076884762624/photo/1
https://twitter.com/dog_rates/status/716802964044845056/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/766008592277377025/photo/1
https://twitter.com/dog_rates/status/668466899341221888/photo/1
https://twitter.com/dog_rates/status/799063482566066176/photo/1,https://twitter.com/d
https://vine.co/v/ienexVMZgi5
https://twitter.com/dog_rates/status/770414278348247044/photo/1
https://twitter.com/dog_rates/status/675853064436391936/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/880095782870896641/photo/1
https://twitter.com/dog_rates/status/772152991789019136/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/700518061187723268/photo/1
https://twitter.com/dog_rates/status/720059472081784833/photo/1
https://twitter.com/dog_rates/status/744971049620602880/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/674271431610523648/photo/1
https://twitter.com/dog_rates/status/871515927908634625/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/696405997980676096/photo/1
https://twitter.com/dog_rates/status/836989968035819520/photo/1
https://twitter.com/dog_rates/status/792883833364439040/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/691444869282295808/photo/1,https://twitter.com/d
https://twitter.com/dog_rates/status/739932936087216128/photo/1
https://twitter.com/dog_rates/status/796031486298386433/photo/1
https://twitter.com/dog_rates/status/667517642048163840/photo/1
https://twitter.com/dog_rates/status/801538201127157760/photo/1
https://twitter.com/dog_rates/status/670778058496974848/photo/1
https://twitter.com/dog_rates/status/674053186244734976/photo/1
https://twitter.com/dog_rates/status/730924654643314689/photo/1
https://twitter.com/dog_rates/status/720785406564900865/photo/1
https://twitter.com/dog_rates/status/813112105746448384/photo/1
Name: expanded_urls, Length: 2219, dtype: int64
```

Define Replace incorrect dog names(my, a, an, the) with the default string ("None")

Code

```
In [229]: # make a list about the incorrect names
          incorrect_name_list = ["a", "an", "the", "my"]

In [230]: # Replace incorrect names with "None"
          archive_clean["name"] = archive_clean["name"].replace(incorrect_name_list, "None")
```

Test

```
In [231]: # List unique values
          archive_clean["name"].value_counts()

Out[231]: None         816
          Charlie       12
          Lucy          11
          Cooper        11
          Oliver        11
          Penny         10
          Lola          10
          Tucker        10
          Bo             9
          Winston        9
          Sadie          8
          Toby           7
          Bailey         7
          Daisy          7
          Buddy          7
          Stanley        6
          Leo            6
          Jax            6
          Jack           6
          Oscar          6
          Dave           6
          Rusty          6
          Scout          6
          Koda           6
          Milo           6
          Bella          6
          Louis          5
          Finn           5
          Chester        5
          Oakley         5
                       ...
          Aubie          1
          Geno           1
```

```
Joshwa         1
Dallas         1
Kawhi          1
Akumi          1
Aqua           1
Maisey         1
Lulu           1
Zara           1
Fwed           1
Shawwn         1
Berb           1
Flash          1
Eazy           1
Timber         1
Kayla          1
Todo           1
Shadoe         1
Pubert         1
Jebberson      1
Horace         1
Benny          1
Coopson        1
Sid            1
Comet          1
Bradley        1
Strudel        1
Dot            1
Ralph          1
Name: name, Length: 953, dtype: int64
```

```
In [232]: # This should be emtpy
          archive_clean[(
              archive_clean["name"] == "a") | (
              archive_clean["name"] == "an")| (
              archive_clean["name"] == "the")| (
              archive_clean["name"] == "my")]
```

```
Out[232]: Empty DataFrame
          Columns: [tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, source, te
          Index: []
```

Define Melt "doggo", "floofer", "pupper", "puppo" columns into one, called "meme".

Code

```
In [233]: #make a list to dog memes
          meme_list = []
          for index, row in archive_clean.iterrows():
```

```
        if (row["doggo"] != "None"):
            meme_list.append("doggo")
        elif (row["floofer"] != "None"):
            meme_list.append("floofer")
        elif (row["pupper"] != "None"):
            meme_list.append("pupper")
        elif (row["puppo"] != "None"):
            meme_list.append("puppo")
        else:
            meme_list.append("None")
```

```
In [234]: # Add meme list as a new column to the dataframe
          archive_clean["meme"] = meme_list
```

```
In [235]: # Remove unnecessary columns
          archive_clean = archive_clean.drop(["doggo", "floofer", "pupper", "puppo"], 1)
```

Test

```
In [236]: archive_clean.head()
```

```
Out[236]:             tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
          0  892420643555336193                    NaN                  NaN
          1  892177421306343426                    NaN                  NaN
          2  891815181378084864                    NaN                  NaN
          3  891689557279858688                    NaN                  NaN
          4  891327558926688256                    NaN                  NaN

                        timestamp  \
          0  2017-08-01 16:23:56 +0000
          1  2017-08-01 00:17:27 +0000
          2  2017-07-31 00:18:03 +0000
          3  2017-07-30 15:58:51 +0000
          4  2017-07-29 16:00:24 +0000

                                                  source  \
          0  <a href="http://twitter.com/download/iphone" r...
          1  <a href="http://twitter.com/download/iphone" r...
          2  <a href="http://twitter.com/download/iphone" r...
          3  <a href="http://twitter.com/download/iphone" r...
          4  <a href="http://twitter.com/download/iphone" r...

                                                  text  retweeted_status_id  \
          0  This is Phineas. He's a mystical boy. Only eve...                  NaN
          1  This is Tilly. She's just checking pup on you...                  NaN
          2  This is Archie. He is a rare Norwegian Pouncin...                  NaN
          3  This is Darla. She commenced a snooze mid meal...                  NaN
          4  This is Franklin. He would like you to stop ca...                  NaN
```

```
   retweeted_status_user_id retweeted_status_timestamp  \
0                       NaN                       NaN
1                       NaN                       NaN
2                       NaN                       NaN
3                       NaN                       NaN
4                       NaN                       NaN

                                          expanded_urls  rating_numerator  \
0  https://twitter.com/dog_rates/status/892420643...                13
1  https://twitter.com/dog_rates/status/892177421...                13
2  https://twitter.com/dog_rates/status/891815181...                12
3  https://twitter.com/dog_rates/status/891689557...                13
4  https://twitter.com/dog_rates/status/891327558...                12

   rating_denominator      name  meme
0                  10   Phineas  None
1                  10     Tilly  None
2                  10    Archie  None
3                  10     Darla  None
4                  10  Franklin  None
```

```
In [237]: # list unique meme values
          archive_clean.meme.value_counts()
```

```
Out[237]: None       1976
          pupper      245
          doggo        97
          puppo        29
          floofer       9
          Name: meme, dtype: int64
```

Define Change erronous data types. tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id columns and rating_numerator to String. retweeted_status_timestamp, timestamp to DateTime.

Code

```
In [238]: # Replace NaN values with 0 in float types
          archive_clean.in_reply_to_status_id = archive_clean.in_reply_to_status_id.fillna(0)
          archive_clean.in_reply_to_user_id = archive_clean.in_reply_to_user_id.fillna(0)
          archive_clean.retweeted_status_id = archive_clean.retweeted_status_id.fillna(0)
          archive_clean.retweeted_status_user_id = archive_clean.retweeted_status_user_id.filln
```

```
In [239]: # Change float to integer
          archive_clean.in_reply_to_status_id = archive_clean.in_reply_to_status_id.astype("int
          archive_clean.in_reply_to_user_id = archive_clean.in_reply_to_user_id.astype("int64")
          archive_clean.retweeted_status_id = archive_clean.retweeted_status_id.astype("int64")
          archive_clean.retweeted_status_user_id = archive_clean.retweeted_status_user_id.astyp
```

```
In [240]: # Change integer data types to string
          archive_clean.tweet_id = archive_clean.tweet_id.astype("str")
          archive_clean.in_reply_to_status_id = archive_clean.in_reply_to_status_id.astype("st
          archive_clean.in_reply_to_user_id = archive_clean.in_reply_to_user_id.astype("str")
          archive_clean.retweeted_status_id = archive_clean.retweeted_status_id.astype("str")
          archive_clean.retweeted_status_user_id = archive_clean.retweeted_status_user_id.asty
```

```
In [241]: # Fill missing values with None
          archive_clean.tweet_id.replace("0", "None" , inplace = True)
          archive_clean.in_reply_to_status_id.replace("0", "None" , inplace = True)
          archive_clean.in_reply_to_user_id.replace("0", "None" , inplace = True)
          archive_clean.retweeted_status_id.replace("0", "None" , inplace = True)
          archive_clean.retweeted_status_user_id.replace("0", "None" , inplace = True)
```

```
In [242]: # Change date times from string data type to datetime
          archive_clean.timestamp = pd.to_datetime(archive_clean.timestamp)
          archive_clean.retweeted_status_timestamp = pd.to_datetime(archive_clean.retweeted_st
```

```
In [243]: # Change numerator from int to float
          archive_clean.rating_numerator = archive_clean.rating_numerator.astype("float")
```

Test

```
In [244]: archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 14 columns):
tweet_id                      2356 non-null object
in_reply_to_status_id         2356 non-null object
in_reply_to_user_id           2356 non-null object
timestamp                     2356 non-null datetime64[ns]
source                        2356 non-null object
text                          2356 non-null object
retweeted_status_id           2356 non-null object
retweeted_status_user_id      2356 non-null object
retweeted_status_timestamp    181 non-null datetime64[ns]
expanded_urls                 2356 non-null object
rating_numerator              2356 non-null float64
rating_denominator            2356 non-null int64
name                          2356 non-null object
meme                          2356 non-null object
dtypes: datetime64[ns](2), float64(1), int64(1), object(10)
memory usage: 257.8+ KB
```

```
In [245]: archive_clean.sample(5)
```

```
Out[245]:                tweet_id in_reply_to_status_id in_reply_to_user_id  \
          594   798705661114773508                  None                None
          2194  668892474547511297                  None                None
          2163  669375718304980992                  None                None
          1604  685906723014619143                  None                None
          2307  666826780179869698                  None                None

                           timestamp                                             source  \
          594  2016-11-16 01:54:03  <a href="http://twitter.com/download/iphone" r...
          2194 2015-11-23 20:42:48  <a href="http://twitter.com/download/iphone" r...
          2163 2015-11-25 04:43:02  <a href="http://twitter.com/download/iphone" r...
          1604 2016-01-09 19:31:20  <a href="http://twitter.com/download/iphone" r...
          2307 2015-11-18 03:54:28  <a href="http://twitter.com/download/iphone" r...

                                                          text retweeted_status_id  \
          594   RT @dog_rates: Meet Baloo. He's expecting a fa...  740676976021798912
          2194  This is Ruffles. He is an Albanian Shoop Da Wh...                None
          2163  This is Billl. He's trying to be a ghost but h...                None
          1604  This is Olive. He's stuck in a sleeve. 9/10 da...                None
          2307  12/10 simply brilliant pup https://t.co/V6ZzG4...                None

               retweeted_status_user_id retweeted_status_timestamp  \
          594                 4196983835        2016-06-08 22:48:46
          2194                      None                        NaT
          2163                      None                        NaT
          1604                      None                        NaT
          2307                      None                        NaT

                                                  expanded_urls  rating_numerator  \
          594   https://twitter.com/dog_rates/status/740676976...              11.0
          2194  https://twitter.com/dog_rates/status/668892474...              11.0
          2163  https://twitter.com/dog_rates/status/669375718...               6.0
          1604  https://twitter.com/dog_rates/status/685906723...               9.0
          2307  https://twitter.com/dog_rates/status/666826780...              12.0

                rating_denominator     name    meme
          594                   10    Baloo  pupper
          2194                  10  Ruffles    None
          2163                  10    Billl    None
          1604                  10    Olive    None
          2307                  10     None    None
```

Define

- Separate text and link in the "text" column ("archive" table) by applying a regular expression
- Fix the wrong decimal numerators by the same way as above

Code

In [246]:
```
# Create new lists with the separated values

text = []
link = []


text = archive_clean.text.str.split(r"[ .\n](?=http)", expand = True)[0]

# We do not need this right now, but later, maybe in another analysis we would need
link = archive_clean.text.str.split(r"[ .\n](?=http)", expand = True)[1]
```

In [247]:
```
# Copy table's text column
original_text = archive_clean["text"]
```

In [248]:
```
# Find all numerators
numerator = []
for i in range(len(original_text)):
    splitted_numerator = re.findall(r"\d+\.?\d?(?=\/\d+)", original_text[i])[0]
    numerator.append(splitted_numerator)
```

In [249]:
```
# Fix numerators in the archive table
archive_clean["rating_numerator"] = numerator
```

In [250]:
```
#Change numerator's type back to float
archive_clean.rating_numerator = archive_clean.rating_numerator.astype("float")
```

In [251]:
```
# Replace Text column with the new one
archive_clean["text"] = text
```

Test

In [252]:
```
archive_clean.sample(5)
```

Out[252]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	retweeted_status_id
1172	720389942216527872	None	None	2016-04-13 23:15:21	<a href="http://twitter.com/download/iphone" r...	This is Ralphé. He patrols the lake. Looking f...	None
560	802952499103731712	None	None	2016-11-27 19:09:28	<a href="http://twitter.com/download/iphone" r...		
2302	667012601033924608	None	None	2015-11-18 16:12:51	<a href="http://twitter.com/download/iphone" r...		
784	775096608509886464	None	None	2016-09-11 22:20:06	<a href="http://twitter.com/download/iphone" r...		
498	813130366689148928	813127251579564032	4196983835	2016-12-25 21:12:41	<a href="http://twitter.com/download/iphone" r...		

```
560   This is Marley. She's having a ruff day. Prett...                None
2302  This is Klevin. He laughs a lot. Very cool dog...                None
784   RT @dog_rates: After so many requests, this is...  740373189193256960
498   I've been informed by multiple sources that th...                None

     retweeted_status_user_id retweeted_status_timestamp  \
1172                     None                         NaT
560                      None                         NaT
2302                     None                         NaT
784                 4196983835       2016-06-08 02:41:38
498                      None                         NaT

                                              expanded_urls  rating_numerator  \
1172  https://twitter.com/dog_rates/status/720389942...              11.0
560   https://twitter.com/dog_rates/status/802952499...              12.0
2302  https://twitter.com/dog_rates/status/667012601...               9.0
784   https://twitter.com/dog_rates/status/740373189...               9.0
498                                                 None              12.0

      rating_denominator    name  meme
1172                  10  Ralphé  None
560                   10  Marley  None
2302                  10  Klevin  None
784                   11    None  None
498                   10    None  None
```

text_list[45] numerator had a problem, let's see what its numerator supposed to be:

```
In [253]: text_list[45]

Out[253]: 'This is Bella. She hopes her smile made you smile. If not, she is also offering you
```

It's value now (ID: 883482846933004288) :

```
In [254]: archive_clean[archive_clean["tweet_id"] == "883482846933004288"]

Out[254]:              tweet_id in_reply_to_status_id in_reply_to_user_id  \
45  883482846933004288                  None                None

              timestamp                                              source  \
45 2017-07-08 00:28:19  <a href="http://twitter.com/download/iphone" r...

                                                 text retweeted_status_id  \
45  This is Bella. She hopes her smile made you sm...                None

   retweeted_status_user_id retweeted_status_timestamp  \
45                     None                         NaT

                                          expanded_urls  rating_numerator  \
```

```
45  https://twitter.com/dog_rates/status/883482846...              13.5

    rating_denominator   name  meme
45                  10  Bella  None
```

```
In [255]: archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 14 columns):
tweet_id                    2356 non-null object
in_reply_to_status_id       2356 non-null object
in_reply_to_user_id         2356 non-null object
timestamp                   2356 non-null datetime64[ns]
source                      2356 non-null object
text                        2356 non-null object
retweeted_status_id         2356 non-null object
retweeted_status_user_id    2356 non-null object
retweeted_status_timestamp  181 non-null datetime64[ns]
expanded_urls               2356 non-null object
rating_numerator            2356 non-null float64
rating_denominator          2356 non-null int64
name                        2356 non-null object
meme                        2356 non-null object
dtypes: datetime64[ns](2), float64(1), int64(1), object(10)
memory usage: 257.8+ KB
```

Define Replace "_" characters with " " in dog names

Code

```
In [256]: # Replace "_" characters with blank a space " "
          image_prediction_clean.p1 = image_prediction_clean.p1.str.replace("_", " ")
          image_prediction_clean.p2 = image_prediction_clean.p2.str.replace("_", " ")
          image_prediction_clean.p3 = image_prediction_clean.p3.str.replace("_", " ")
```

Test

```
In [257]: # Display some samples
          image_prediction_clean.sample(10)
```

```
Out[257]:                 tweet_id                                          jpg_url  \
          719   685906723014619143    https://pbs.twimg.com/media/CYTUhn7WkAEXocW.jpg
          1857  841680585030541313    https://pbs.twimg.com/media/C65AA7_WoAEGqA9.jpg
          1126  727524757080539137    https://pbs.twimg.com/media/Chiv6BAW4AAiQvH.jpg
          772   689557536375177216    https://pbs.twimg.com/media/CZHM60BWIAA4AY4.jpg
          832   693647888581312512    https://pbs.twimg.com/media/CaBVE80WAAA8sGk.jpg
```

```
333   672160042234327040    https://pbs.twimg.com/media/CVP9_beUEAAwURR.jpg
1519  787717603741622272    https://pbs.twimg.com/media/Cu6I9vvWIAAZG0a.jpg
697   684538444857667585  https://pbs.twimg.com/ext_tw_video_thumb/68453...
620   680473011644985345    https://pbs.twimg.com/media/CXGGlzvWYAArPfk.jpg
1907  852553447878664193    https://pbs.twimg.com/media/C9Tg1bPW0AkAMDI.jpg

     img_num                 p1   p1_conf p1_dog                         p2  \
719        1  Yorkshire terrier  0.414963   True                     briard
1857       1          Chihuahua  0.547401   True                     bow tie
1126       2         Pomeranian  0.958834   True                   Chihuahua
772        1         Eskimo dog  0.169482   True               Siberian husky
832        1          washbasin  0.272451  False                     doormat
333        1                pug  0.561027   True               French bulldog
1519       3    German shepherd  0.992339   True                     malinois
697        1          Chihuahua  0.702583   True                  Siamese cat
620        1   Lakeland terrier  0.796694   True  West Highland white terrier
1907       1         bloodhound  0.186498   True            Brabancon griffon

                  p2_conf p2_dog                 p3               p3_conf  \
719             0.0635052   True            Pekinese             0.0536822
1857  0.19836099999999998  False            Pembroke   0.058492499999999996
1126            0.0240992   True                chow            0.00394105
772              0.161655   True               dingo    0.15441400000000002
832              0.165871  False             bathtub             0.0663684
333   0.22211399999999998   True  Labrador retriever             0.0654556
1519           0.00492039   True               kelpie  0.0008528019999999999
697             0.0682181  False             macaque             0.0433246
620              0.138709   True     Norwich terrier   0.016253399999999998
1907  0.13902799999999998   True           Rottweiler               0.12594

     p3_dog
719    True
1857   True
1126   True
772   False
832   False
333    True
1519   True
697   False
620    True
1907   True
```

Define Change non descriptive column headers from p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog to prediction1, is_prediction_1_dog, prediction_1_confidence, et cetera.

Code

```
In [258]: # Rename image_prediction's non descriptive columns
          image_prediction_clean.rename(columns = {"p1" : "prediction1",
                                                   "p2" : "prediction2",
                                                   "p3" : "prediction3",
                                                  "p1_conf" : "prediction_1_confidence",
                                                  "p2_conf" : "prediction_2_confidence",
                                                  "p3_conf" : "prediction_3_confidence",
                                                  "p1_dog" : "is_prediction_1_dog",
                                                  "p2_dog" : "is_prediction_2_dog",
                                                  "p3_dog" : "is_prediction_3_dog"}, inplace =
```

Test

```
In [259]: image_prediction_clean.head(3)
```

```
Out[259]:              tweet_id                                          jpg_url  \
          0  666020888022790149  https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg
          1  666029285002620928  https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg
          2  666033412701032449  https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg

            img_num             prediction1 prediction_1_confidence is_prediction_1_dog  \
          0       1  Welsh springer spaniel                0.465074                True
          1       1                 redbone                0.506826                True
          2       1         German shepherd                0.596461                True

                    prediction2 prediction_2_confidence is_prediction_2_dog  \
          0              collie                0.156665                True
          1  miniature pinscher     0.07419169999999999                True
          2           malinois     0.13858399999999998                True

                     prediction3 prediction_3_confidence is_prediction_3_dog
          0    Shetland sheepdog               0.0614285                True
          1  Rhodesian ridgeback                 0.07201                True
          2           bloodhound     0.11619700000000001                True
```

Define Gather all retweet ID and delete them from all tables. My assumption is that this will solve our duplicated and missing jpg_url values in the image_prediction table.

Code

```
In [260]: # Collect retweets into a dataframe
          retweet_id_df = archive_clean[(archive_clean["retweeted_status_id"] != "None") |
                                        (archive_clean["retweeted_status_user_id"] !=
          retweet_id_df.head(10)
```

```
Out[260]:              tweet_id in_reply_to_status_id in_reply_to_user_id  \
          19  888202515573088257                 None                None
          32  886054160059072513                 None                None
```

```
36  885311592912609280                 None                 None
68  879130579576475649                 None                 None
73  878404777348136964                 None                 None
74  878316110768087041                 None                 None
78  877611172832227328                 None                 None
91  874434818259525634                 None                 None
95  873697596434513921                 None                 None
97  873337748698140672                 None                 None

             timestamp                                             source  \
19 2017-07-21 01:02:36  <a href="http://twitter.com/download/iphone" r...
32 2017-07-15 02:45:48  <a href="http://twitter.com/download/iphone" r...
36 2017-07-13 01:35:06  <a href="http://twitter.com/download/iphone" r...
68 2017-06-26 00:13:58  <a href="http://twitter.com/download/iphone" r...
73 2017-06-24 00:09:53  <a href="http://twitter.com/download/iphone" r...
74 2017-06-23 18:17:33  <a href="http://twitter.com/download/iphone" r...
78 2017-06-21 19:36:23  <a href="http://twitter.com/download/iphone" r...
91 2017-06-13 01:14:41  <a href="http://twitter.com/download/iphone" r...
95 2017-06-11 00:25:14  <a href="http://twitter.com/download/iphone" r...
97 2017-06-10 00:35:19  <a href="http://twitter.com/download/iphone" r...

                                                 text retweeted_status_id  \
19  RT @dog_rates: This is Canela. She attempted s...  887473957103951872
32                        RT @Athletics: 12/10 #BATP  886053734421102592
36  RT @dog_rates: This is Lilly. She just paralle...  830583320585068544
68  RT @dog_rates: This is Emmy. She was adopted t...  878057613040115712
73  RT @dog_rates: Meet Shadow. In an attempt to r...  878281511006478336
74  RT @dog_rates: Meet Terrance. He's being yelle...  669000397445533696
78  RT @rachel2195: @dog_rates the boyfriend and h...  876850772322988032
91  RT @dog_rates: This is Coco. At first I though...  866334964761202688
95  RT @dog_rates: This is Walter. He won't start ...  868880397819494400
97  RT @dog_rates: This is Sierra. She's one preci...  873213775632977920

   retweeted_status_user_id retweeted_status_timestamp  \
19               4196983835        2017-07-19 00:47:34
32                 19607400        2017-07-15 02:44:07
36               4196983835        2017-02-12 01:04:29
68               4196983835        2017-06-23 01:10:23
73               4196983835        2017-06-23 16:00:04
74               4196983835        2015-11-24 03:51:38
78                512804507        2017-06-19 17:14:49
91               4196983835        2017-05-21 16:48:45
95               4196983835        2017-05-28 17:23:24
97               4196983835        2017-06-09 16:22:42

                                        expanded_urls  rating_numerator  \
19  https://twitter.com/dog_rates/status/887473957...              13.0
32  https://twitter.com/dog_rates/status/886053434...              12.0
```

```
36  https://twitter.com/dog_rates/status/830583320...              13.0
68  https://twitter.com/dog_rates/status/878057613...              14.0
73  https://www.gofundme.com/3yd6y1c,https://twitt...              13.0
74  https://twitter.com/dog_rates/status/669000397...              11.0
78  https://twitter.com/rachel2195/status/87685077...              14.0
91  https://twitter.com/dog_rates/status/866334964...              12.0
95  https://twitter.com/dog_rates/status/868880397...              14.0
97  https://www.gofundme.com/help-my-baby-sierra-g...              12.0

    rating_denominator      name    meme
19                  10    Canela    None
32                  10      None    None
36                  10     Lilly    None
68                  10      Emmy    None
73                  10    Shadow    None
74                  10  Terrance    None
78                  10      None  pupper
91                  10      Coco    None
95                  10    Walter    None
97                  10    Sierra  pupper
```

In [261]:
```
# Make a list of retweet ID
retweet_id = []
retweet_id = retweet_id_df["tweet_id"]
```

In [262]:
```
# Helpful for Test
# length of the current cleaned dataframe and the length of the list of retweet id
current_archive_count = len(archive_clean)

retweet_count = len(retweet_id)
```

In [263]:
```
# Remove retweets from all tables
archive_clean = archive_clean[~archive_clean["tweet_id"].isin(retweet_id)]
tweet_count_clean = tweet_count_clean[~tweet_count_clean["id"].isin(retweet_id)]
image_prediction_clean = image_prediction_clean[~image_prediction_clean["tweet_id"].
```

Test

In [264]:
```
# Check if the retweet count is really missing from the df
# Must be True
print(current_archive_count - retweet_count == len(archive_clean))
```

```
True
```

In [265]:
```
# Check if there is a retweet id in the tables
# Lengths must be 0
test_archive = archive_clean[archive_clean["tweet_id"].isin(retweet_id)]
test_tweet_count = tweet_count_clean[tweet_count_clean["id"].isin(retweet_id)]
```

```
test_image_prediction = image_prediction_clean[image_prediction_clean["tweet_id"].isi

print(len(test_archive))
print(len(test_tweet_count))
print(len(test_image_prediction))
```

```
0
0
0
```

```
In [266]: # Test to see if there are any duplicates in jpg_url column
          # Must be empty
          image_prediction_clean[(image_prediction_clean["jpg_url"].duplicated())]
```

```
Out[266]: Empty DataFrame
          Columns: [tweet_id, jpg_url, img_num, prediction1, prediction_1_confidence, is_predic
          Index: []
```

```
In [267]: # Check to see if there is a missing jpg_url value
          image_prediction_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1994 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id                  1994 non-null object
jpg_url                   1994 non-null object
img_num                   1994 non-null object
prediction1               1994 non-null object
prediction_1_confidence   1994 non-null object
is_prediction_1_dog       1994 non-null object
prediction2               1994 non-null object
prediction_2_confidence   1994 non-null object
is_prediction_2_dog       1994 non-null object
prediction3               1994 non-null object
prediction_3_confidence   1994 non-null object
is_prediction_3_dog       1994 non-null object
dtypes: object(12)
memory usage: 202.5+ KB
```

```
In [268]: # We should not see any strange here like NaN, None, et cetera.
          image_prediction_clean["jpg_url"].value_counts()
```

```
Out[268]: https://pbs.twimg.com/media/CfkG_PMWsAAH0MZ.jpg
          https://pbs.twimg.com/ext_tw_video_thumb/676957802976419840/pu/img/dCj-qlXo73A5hf6Q.j
          https://pbs.twimg.com/media/CjQ4radW0AENP-m.jpg
          https://pbs.twimg.com/media/DCN85nGUwAAzG_q.jpg
          https://pbs.twimg.com/media/CmufLLsXYAAsU0r.jpg
```

https://pbs.twimg.com/media/CXw2jSpWMAAad6V.jpg
https://pbs.twimg.com/media/DBINZcxXgAQ-R6P.jpg
https://pbs.twimg.com/media/Ckd-bqVUkAIiyM7.jpg
https://pbs.twimg.com/media/C9VNNp1XkAEWRFb.jpg
https://pbs.twimg.com/media/DDMD_phXoAQ1qf0.jpg
https://pbs.twimg.com/media/CnLmRiYXEAAO_8f.jpg
https://pbs.twimg.com/media/DBP1asiUAAEKZI5.jpg
https://pbs.twimg.com/media/CzFp3FNW8AAfvV8.jpg
https://pbs.twimg.com/media/C7ztkInW0AEh1CD.jpg
https://pbs.twimg.com/media/C12whDoVEAALRxa.jpg
https://pbs.twimg.com/media/CT5d9DZXAAALcwe.jpg
https://pbs.twimg.com/media/DDcscbXU0AIfDzs.jpg
https://pbs.twimg.com/media/CZNK7NpWwAEAqUh.jpg
https://pbs.twimg.com/media/CYK6kf0WMAAzP-0.jpg
https://pbs.twimg.com/media/DEEEnIqXYAAiJh_.jpg
https://pbs.twimg.com/media/CVqwedgXIAEAT6A.jpg
https://pbs.twimg.com/media/CtzgXgeXYAA1Gxw.jpg
https://pbs.twimg.com/ext_tw_video_thumb/751250895690731520/pu/img/eziHbU1KbgZg-ijN.j
https://pbs.twimg.com/media/Cq_Vy9KWcAIUIuv.jpg
https://pbs.twimg.com/media/CdhUIMSUIAA4wYK.jpg
https://pbs.twimg.com/media/CVltNgxWEAA5sCJ.jpg
https://pbs.twimg.com/media/CVt-SeMWwAAs9HH.jpg
https://pbs.twimg.com/media/CvAr88kW8AEKNAO.jpg
https://pbs.twimg.com/media/CeQVF1eVIAAJaTv.jpg
https://pbs.twimg.com/media/CUb6ebKWcAAJkd0.jpg

https://pbs.twimg.com/media/CmU2DVWWgAArvp3.jpg
https://pbs.twimg.com/media/CZBUO2UWsAAKehS.jpg
https://pbs.twimg.com/media/CUoSjTnWwAANNak.jpg
https://pbs.twimg.com/media/Cp8k6oRWcAUL78U.jpg
https://pbs.twimg.com/media/Cf4bcm8XEAAX4xV.jpg
https://pbs.twimg.com/media/CbJRrigW0AIcJ2N.jpg
https://pbs.twimg.com/media/Cf3sH62VAAA-LiP.jpg
https://pbs.twimg.com/media/DFg_2PVW0AEHN3p.jpg
https://pbs.twimg.com/media/CwJEIKTWYAAvL-T.jpg
https://pbs.twimg.com/media/CW-dU34WQAANBGy.jpg
https://pbs.twimg.com/media/Cf4qRcmWEAA9V4h.jpg
https://pbs.twimg.com/media/CWFFt3_XIAArIYK.jpg
https://pbs.twimg.com/media/Crdhh_1XEAAHKHi.jpg
https://pbs.twimg.com/media/CfxcKU6W8AE-wEx.jpg
https://pbs.twimg.com/media/CZ5entwWYAAocEg.jpg
https://pbs.twimg.com/media/Ctc_-BTWEAAQpZh.jpg
https://pbs.twimg.com/media/C0DhpcrUAAAnx88.jpg
https://pbs.twimg.com/media/CoAqwPTW8AAiJlz.jpg
https://pbs.twimg.com/media/CZv13u5WYAA6wQe.jpg
https://pbs.twimg.com/media/CUym4Y5WsAEiI9_.jpg
https://pbs.twimg.com/media/CsfLUDbXEAAu0VF.jpg
https://pbs.twimg.com/media/CXW4wGHWsAE_eBD.jpg

```
https://pbs.twimg.com/media/CnnKCKNWgAAcOB8.jpg
https://pbs.twimg.com/media/CT9vZEYWUAAlZO5.jpg
https://pbs.twimg.com/media/CWyD2HGUYAQ1Xa7.jpg
https://pbs.twimg.com/ext_tw_video_thumb/859196962498805762/pu/img/-yBpr4-o4GJZECYE.j
https://pbs.twimg.com/media/CUMZnmhUEAEbtis.jpg
https://pbs.twimg.com/media/CrcPjh0WcAA_SPT.jpg
https://pbs.twimg.com/media/Cb2cfd9WAAEL-zk.jpg
https://pbs.twimg.com/media/CYyucekVAAESj8K.jpg
Name: jpg_url, Length: 1994, dtype: int64
```

Define Unuseful "tweet_count" table. "Delete" tweet_count table and insert its columns into the arhive table.

Code

```
In [269]: # Merge these two tables
          archive_clean = pd.merge(archive_clean, tweet_count_clean, how = "left", left_on = "t
```

```
In [270]: # Check if there is a difference between the IDs
          archive_clean[archive_clean["tweet_id"] != archive_clean["id"]]
```

```
Out[270]: Empty DataFrame
          Columns: [tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, source, te
          Index: []
```

```
In [271]: #Remove ID column
          archive_clean.drop("id", axis = 1, inplace = True)
```

Test

```
In [272]: # Check some rows
          archive_clean.sample(5)
```

```
Out[272]:                tweet_id in_reply_to_status_id in_reply_to_user_id  \
          334   825535076884762624                  None                None
          993   720415127506415616                  None                None
          958   728387165835677696                  None                None
          836   747439450712596480                  None                None
          1230  699060279947165696                  None                None

                           timestamp                                             source  \
          334  2017-01-29 02:44:34  <a href="http://twitter.com/download/iphone" r...
          993  2016-04-14 00:55:25  <a href="http://twitter.com/download/iphone" r...
          958  2016-05-06 00:53:27  <a href="http://twitter.com/download/iphone" r...
          836  2016-06-27 14:40:26  <a href="http://vine.co" rel="nofollow">Vine -...
          1230 2016-02-15 02:38:53  <a href="http://vine.co" rel="nofollow">Vine -...

                                                                text retweeted_status_id  \
```

```
334   Here's a very loving and accepting puppo. Appe...                None
993            Garden's coming in nice this year. 10/10                None
958   This is Enchilada (yes, that's her real name)...                 None
836   This is Linus. He just wanted to say hello but...                None
1230  This is Yukon. He pukes rainbows. 12/10 magica...                None

     retweeted_status_user_id retweeted_status_timestamp  \
334                      None                        NaT
993                      None                        NaT
958                      None                        NaT
836                      None                        NaT
1230                     None                        NaT

                                            expanded_urls  rating_numerator  \
334   https://twitter.com/dog_rates/status/825535076...              14.0
993   https://twitter.com/dog_rates/status/720415127...              10.0
958   https://twitter.com/dog_rates/status/728387165...              12.0
836                       https://vine.co/v/5uTVXWvn3Ip              12.0
1230                      https://vine.co/v/inlmMHxtqDD              12.0

      rating_denominator       name   meme  retweet_count  favorite_count
334                   10       None  puppo          19287           56404
993                   10       None   None           1639            4414
958                   10  Enchilada   None           1049            3937
836                   10      Linus   None           2139            5875
1230                  10      Yukon   None           2003            4069
```

Define Merge archive table and image_prediction table

Code

```
In [273]: #Merge two tables on ID
          archive_clean = pd.merge(archive_clean, image_prediction_clean, how = "left", left_o
```

Test

```
In [274]: # Check to see if the columns were imported correctly
          archive_clean
```

```
Out[274]:                tweet_id in_reply_to_status_id in_reply_to_user_id  \
          0     892420643555336193                  None                None
          1     892177421306343426                  None                None
          2     891815181378084864                  None                None
          3     891689557279858688                  None                None
          4     891327558926688256                  None                None
          5     891087950875897856                  None                None
          6     890971913173991426                  None                None
          7     890729181411237888                  None                None
```

8	890609185150312448	None	None
9	890240255349198849	None	None
10	890006608113172480	None	None
11	889880896479866881	None	None
12	889665388333682689	None	None
13	889638837579907072	None	None
14	889531135344209921	None	None
15	889278841981685760	None	None
16	888917238123831296	None	None
17	888804989199671297	None	None
18	888554962724278272	None	None
19	888078434458587136	None	None
20	887705289381826560	None	None
21	887517139158093824	None	None
22	887473957103951883	None	None
23	887343217045368832	None	None
24	887101392804085760	None	None
25	886983233522544640	None	None
26	886736880519319552	None	None
27	886680336477933568	None	None
28	886366144734445568	None	None
29	886267009285017600	886266357075128320	2281181600
...	...	...	...
2145	666411507551481857	None	None
2146	666407126856765440	None	None
2147	666396247373291520	None	None
2148	666373753744588802	None	None
2149	666362758909284353	None	None
2150	666353288456101888	None	None
2151	666345417576210432	None	None
2152	666337882303524864	None	None
2153	666293911632134144	None	None
2154	666287406224695296	None	None
2155	666273097616637952	None	None
2156	666268910803644416	None	None
2157	666104133288665088	None	None
2158	666102155909144576	None	None
2159	666099513787052032	None	None
2160	666094000022159362	None	None
2161	666082916733198337	None	None
2162	666073100786774016	None	None
2163	666071193221509120	None	None
2164	666063827256086533	None	None
2165	666058600524156928	None	None
2166	666057090499244032	None	None
2167	666055525042405380	None	None
2168	666051853826850816	None	None
2169	666050758794694657	None	None

```
2170  666049248165822465                 None                 None
2171  666044226329800704                 None                 None
2172  666033412701032449                 None                 None
2173  666029285002620928                 None                 None
2174  666020888022790149                 None                 None

               timestamp                                              source  \
0    2017-08-01 16:23:56  <a href="http://twitter.com/download/iphone" r...
1    2017-08-01 00:17:27  <a href="http://twitter.com/download/iphone" r...
2    2017-07-31 00:18:03  <a href="http://twitter.com/download/iphone" r...
3    2017-07-30 15:58:51  <a href="http://twitter.com/download/iphone" r...
4    2017-07-29 16:00:24  <a href="http://twitter.com/download/iphone" r...
5    2017-07-29 00:08:17  <a href="http://twitter.com/download/iphone" r...
6    2017-07-28 16:27:12  <a href="http://twitter.com/download/iphone" r...
7    2017-07-28 00:22:40  <a href="http://twitter.com/download/iphone" r...
8    2017-07-27 16:25:51  <a href="http://twitter.com/download/iphone" r...
9    2017-07-26 15:59:51  <a href="http://twitter.com/download/iphone" r...
10   2017-07-26 00:31:25  <a href="http://twitter.com/download/iphone" r...
11   2017-07-25 16:11:53  <a href="http://twitter.com/download/iphone" r...
12   2017-07-25 01:55:32  <a href="http://twitter.com/download/iphone" r...
13   2017-07-25 00:10:02  <a href="http://twitter.com/download/iphone" r...
14   2017-07-24 17:02:04  <a href="http://twitter.com/download/iphone" r...
15   2017-07-24 00:19:32  <a href="http://twitter.com/download/iphone" r...
16   2017-07-23 00:22:39  <a href="http://twitter.com/download/iphone" r...
17   2017-07-22 16:56:37  <a href="http://twitter.com/download/iphone" r...
18   2017-07-22 00:23:06  <a href="http://twitter.com/download/iphone" r...
19   2017-07-20 16:49:33  <a href="http://twitter.com/download/iphone" r...
20   2017-07-19 16:06:48  <a href="http://twitter.com/download/iphone" r...
21   2017-07-19 03:39:09  <a href="http://twitter.com/download/iphone" r...
22   2017-07-19 00:47:34  <a href="http://twitter.com/download/iphone" r...
23   2017-07-18 16:08:03  <a href="http://twitter.com/download/iphone" r...
24   2017-07-18 00:07:08  <a href="http://twitter.com/download/iphone" r...
25   2017-07-17 16:17:36  <a href="http://twitter.com/download/iphone" r...
26   2017-07-16 23:58:41  <a href="http://twitter.com/download/iphone" r...
27   2017-07-16 20:14:00  <a href="http://twitter.com/download/iphone" r...
28   2017-07-15 23:25:31  <a href="http://twitter.com/download/iphone" r...
29   2017-07-15 16:51:35  <a href="http://twitter.com/download/iphone" r...
...                  ...                                                ...
2145 2015-11-17 00:24:19  <a href="http://twitter.com/download/iphone" r...
2146 2015-11-17 00:06:54  <a href="http://twitter.com/download/iphone" r...
2147 2015-11-16 23:23:41  <a href="http://twitter.com/download/iphone" r...
2148 2015-11-16 21:54:18  <a href="http://twitter.com/download/iphone" r...
2149 2015-11-16 21:10:36  <a href="http://twitter.com/download/iphone" r...
2150 2015-11-16 20:32:58  <a href="http://twitter.com/download/iphone" r...
2151 2015-11-16 20:01:42  <a href="http://twitter.com/download/iphone" r...
2152 2015-11-16 19:31:45  <a href="http://twitter.com/download/iphone" r...
2153 2015-11-16 16:37:02  <a href="http://twitter.com/download/iphone" r...
2154 2015-11-16 16:11:11  <a href="http://twitter.com/download/iphone" r...
```

```
2155 2015-11-16 15:14:19  <a href="http://twitter.com/download/iphone" r...
2156 2015-11-16 14:57:41  <a href="http://twitter.com/download/iphone" r...
2157 2015-11-16 04:02:55  <a href="http://twitter.com/download/iphone" r...
2158 2015-11-16 03:55:04  <a href="http://twitter.com/download/iphone" r...
2159 2015-11-16 03:44:34  <a href="http://twitter.com/download/iphone" r...
2160 2015-11-16 03:22:39  <a href="http://twitter.com/download/iphone" r...
2161 2015-11-16 02:38:37  <a href="http://twitter.com/download/iphone" r...
2162 2015-11-16 01:59:36  <a href="http://twitter.com/download/iphone" r...
2163 2015-11-16 01:52:02  <a href="http://twitter.com/download/iphone" r...
2164 2015-11-16 01:22:45  <a href="http://twitter.com/download/iphone" r...
2165 2015-11-16 01:01:59  <a href="http://twitter.com/download/iphone" r...
2166 2015-11-16 00:55:59  <a href="http://twitter.com/download/iphone" r...
2167 2015-11-16 00:49:46  <a href="http://twitter.com/download/iphone" r...
2168 2015-11-16 00:35:11  <a href="http://twitter.com/download/iphone" r...
2169 2015-11-16 00:30:50  <a href="http://twitter.com/download/iphone" r...
2170 2015-11-16 00:24:50  <a href="http://twitter.com/download/iphone" r...
2171 2015-11-16 00:04:52  <a href="http://twitter.com/download/iphone" r...
2172 2015-11-15 23:21:54  <a href="http://twitter.com/download/iphone" r...
2173 2015-11-15 23:05:30  <a href="http://twitter.com/download/iphone" r...
2174 2015-11-15 22:32:08  <a href="http://twitter.com/download/iphone" r...

                                                   text retweeted_status_id  \
0     This is Phineas. He's a mystical boy. Only eve...                 None
1     This is Tilly. She's just checking pup on you...                 None
2     This is Archie. He is a rare Norwegian Pouncin...                 None
3     This is Darla. She commenced a snooze mid meal...                 None
4     This is Franklin. He would like you to stop ca...                 None
5     Here we have a majestic great white breaching ...                 None
6     Meet Jax. He enjoys ice cream so much he gets ...                 None
7     When you watch your owner call another dog a g...                 None
8     This is Zoey. She doesn't want to be one of th...                 None
9     This is Cassie. She is a college pup. Studying...                 None
10    This is Koda. He is a South Australian decksha...                 None
11    This is Bruno. He is a service shark. Only get...                 None
12    Here's a puppo that seems to be on the fence a...                 None
13    This is Ted. He does his best. Sometimes that'...                 None
14    This is Stuart. He's sporting his favorite fan...                 None
15    This is Oliver. You're witnessing one of his m...                 None
16    This is Jim. He found a fren. Taught him how t...                 None
17    This is Zeke. He has a new stick. Very proud o...                 None
18    This is Ralphus. He's powering up. Attempting ...                 None
19    This is Gerald. He was just told he didn't get...                 None
20    This is Jeffrey. He has a monopoly on the pool...                 None
21    I've yet to rate a Venezuelan Hover Wiener. Th...                 None
22    This is Canela. She attempted some fancy porch...                 None
23    You may not have known you needed to see this ...                 None
24    This... is a Jubilant Antarctic House Bear. We...                 None
25    This is Maya. She's very shy. Rarely leaves he...                 None
```

```
26    This is Mingus. He's a wonderful father to his...                None
27    This is Derek. He's late for a dog meeting. 13...                None
28    This is Roscoe. Another pupper fallen victim t...                None
29    @NonWhiteHat @MayhewMayhem omg hello tanner yo...                None
...                                                 ...                 ...
2145  This is quite the dog. Gets really excited whe...                None
2146  This is a southern Vesuvius bumblegruff. Can d...                None
2147  Oh goodness. A super rare northeast Qdoba kang...                None
2148  Those are sunglasses and a jean jacket. 11/10 ...                None
2149  Unique dog here. Very small. Lives in containe...                None
2150  Here we have a mixed Asiago from the Galápagos...                None
2151  Look at this jokester thinking seat belt laws ...                None
2152  This is an extremely rare horned Parthenon. No...                None
2153  This is a funny dog. Weird toes. Won't come do...                None
2154  This is an Albanian 3 1/2 legged  Episcopalian...                None
2155                              Can take selfies 11/10                 None
2156  Very concerned about fellow dog trapped in com...                None
2157  Not familiar with this breed. No tail (weird)...                 None
2158  Oh my. Here you are seeing an Adobe Setter giv...                None
2159  Can stand on stump for what seems like a while...                None
2160  This appears to be a Mongolian Presbyterian mi...                None
2161  Here we have a well-established sunblockerspan...                None
2162  Let's hope this flight isn't Malaysian (lol). ...                None
2163  Here we have a northern speckled Rhododendron...                 None
2164  This is the happiest dog you will ever see. Ve...                None
2165  Here is the Rand Paul of retrievers folks! He'...                None
2166  My oh my. This is a rare blond Canadian terrie...                None
2167  Here is a Siberian heavily armored polar bear ...                None
2168  This is an odd dog. Hard on the outside but lo...                None
2169  This is a truly beautiful English Wilson Staff...                None
2170  Here we have a 1949 1st generation vulpix. Enj...                None
2171  This is a purebred Piers Morgan. Loves to Netf...                None
2172  Here is a very happy pup. Big fan of well-main...                None
2173  This is a western brown Mitsubishi terrier. Up...                None
2174  Here we have a Japanese Irish Setter. Lost eye...                None

     retweeted_status_user_id retweeted_status_timestamp  \
0                        None                        NaT
1                        None                        NaT
2                        None                        NaT
3                        None                        NaT
4                        None                        NaT
5                        None                        NaT
6                        None                        NaT
7                        None                        NaT
8                        None                        NaT
9                        None                        NaT
10                       None                        NaT
```

11	None	NaT
12	None	NaT
13	None	NaT
14	None	NaT
15	None	NaT
16	None	NaT
17	None	NaT
18	None	NaT
19	None	NaT
20	None	NaT
21	None	NaT
22	None	NaT
23	None	NaT
24	None	NaT
25	None	NaT
26	None	NaT
27	None	NaT
28	None	NaT
29	None	NaT
...	...	...
2145	None	NaT
2146	None	NaT
2147	None	NaT
2148	None	NaT
2149	None	NaT
2150	None	NaT
2151	None	NaT
2152	None	NaT
2153	None	NaT
2154	None	NaT
2155	None	NaT
2156	None	NaT
2157	None	NaT
2158	None	NaT
2159	None	NaT
2160	None	NaT
2161	None	NaT
2162	None	NaT
2163	None	NaT
2164	None	NaT
2165	None	NaT
2166	None	NaT
2167	None	NaT
2168	None	NaT
2169	None	NaT
2170	None	NaT
2171	None	NaT
2172	None	NaT

```
2173                    None                    NaT
2174                    None                    NaT

                                          expanded_urls  ...  \
0     https://twitter.com/dog_rates/status/892420643...  ...
1     https://twitter.com/dog_rates/status/892177421...  ...
2     https://twitter.com/dog_rates/status/891815181...  ...
3     https://twitter.com/dog_rates/status/891689557...  ...
4     https://twitter.com/dog_rates/status/891327558...  ...
5     https://twitter.com/dog_rates/status/891087950...  ...
6     https://gofundme.com/ydvmve-surgery-for-jax,ht...  ...
7     https://twitter.com/dog_rates/status/890729181...  ...
8     https://twitter.com/dog_rates/status/890609185...  ...
9     https://twitter.com/dog_rates/status/890240255...  ...
10    https://twitter.com/dog_rates/status/890006608...  ...
11    https://twitter.com/dog_rates/status/889880896...  ...
12    https://twitter.com/dog_rates/status/889665388...  ...
13    https://twitter.com/dog_rates/status/889638837...  ...
14    https://twitter.com/dog_rates/status/889531135...  ...
15    https://twitter.com/dog_rates/status/889278841...  ...
16    https://twitter.com/dog_rates/status/888917238...  ...
17    https://twitter.com/dog_rates/status/888804989...  ...
18    https://twitter.com/dog_rates/status/888554962...  ...
19    https://twitter.com/dog_rates/status/888078434...  ...
20    https://twitter.com/dog_rates/status/887705289...  ...
21    https://twitter.com/dog_rates/status/887517139...  ...
22    https://twitter.com/dog_rates/status/887473957...  ...
23    https://twitter.com/dog_rates/status/887343217...  ...
24    https://twitter.com/dog_rates/status/887101392...  ...
25    https://twitter.com/dog_rates/status/886983233...  ...
26    https://www.gofundme.com/mingusneedsus,https:/...  ...
27    https://twitter.com/dog_rates/status/886680336...  ...
28    https://twitter.com/dog_rates/status/886366144...  ...
29                                                 None  ...
...                                                 ...  ...
2145  https://twitter.com/dog_rates/status/666411507...  ...
2146  https://twitter.com/dog_rates/status/666407126...  ...
2147  https://twitter.com/dog_rates/status/666396247...  ...
2148  https://twitter.com/dog_rates/status/666373753...  ...
2149  https://twitter.com/dog_rates/status/666362758...  ...
2150  https://twitter.com/dog_rates/status/666353288...  ...
2151  https://twitter.com/dog_rates/status/666345417...  ...
2152  https://twitter.com/dog_rates/status/666337882...  ...
2153  https://twitter.com/dog_rates/status/666293911...  ...
2154  https://twitter.com/dog_rates/status/666287406...  ...
2155  https://twitter.com/dog_rates/status/666273097...  ...
2156  https://twitter.com/dog_rates/status/666268910...  ...
2157  https://twitter.com/dog_rates/status/666104133...  ...
```

```
2158  https://twitter.com/dog_rates/status/666102155...         ...
2159  https://twitter.com/dog_rates/status/666099513...         ...
2160  https://twitter.com/dog_rates/status/666094000...         ...
2161  https://twitter.com/dog_rates/status/666082916...         ...
2162  https://twitter.com/dog_rates/status/666073100...         ...
2163  https://twitter.com/dog_rates/status/666071193...         ...
2164  https://twitter.com/dog_rates/status/666063827...         ...
2165  https://twitter.com/dog_rates/status/666058600...         ...
2166  https://twitter.com/dog_rates/status/666057090...         ...
2167  https://twitter.com/dog_rates/status/666055525...         ...
2168  https://twitter.com/dog_rates/status/666051853...         ...
2169  https://twitter.com/dog_rates/status/666050758...         ...
2170  https://twitter.com/dog_rates/status/666049248...         ...
2171  https://twitter.com/dog_rates/status/666044226...         ...
2172  https://twitter.com/dog_rates/status/666033412...         ...
2173  https://twitter.com/dog_rates/status/666029285...         ...
2174  https://twitter.com/dog_rates/status/666020888...         ...

      img_num                 prediction1 prediction_1_confidence  \
0           1                      orange     0.09704860000000001
1           1                   Chihuahua                0.323581
2           1                   Chihuahua                0.716012
3           1                  paper towel     0.17027799999999998
4           2                      basset                0.555712
5           1     Chesapeake Bay retriever                0.425595
6           1                  Appenzeller     0.34170300000000003
7           2                   Pomeranian                0.566142
8           1                Irish terrier     0.48757399999999995
9           1                     Pembroke                0.511319
10          1                      Samoyed      0.9579790000000001
11          1                French bulldog                0.377417
12          1                     Pembroke                0.966327
13          1                French bulldog                 0.99165
14          1              golden retriever                0.953442
15          1                      whippet                0.626152
16          1              golden retriever                0.714719
17          1              golden retriever                 0.46976
18          3                Siberian husky                0.700377
19          1                French bulldog      0.9950260000000001
20          1                      basset                0.821664
21          1                   limousine                0.130432
22          2                     Pembroke                0.809197
23          1              Mexican hairless                0.330741
24          1                      Samoyed                0.733942
25          2                   Chihuahua                0.793469
26          1                      kuvasz                0.309706
27          1                  convertible                0.738995
28          1                French bulldog                0.999201
```

29	NaN	NaN	NaN
...	...	...	...
2145	1	coho	0.40464
2146	1	black-and-tan coonhound	0.529139
2147	1	Chihuahua	0.978108
2148	1	soft-coated wheaten terrier	0.326467
2149	1	guinea pig	0.9964959999999999
2150	1	malamute	0.33687399999999995
2151	1	golden retriever	0.8587440000000001
2152	1	ox	0.41666899999999996
2153	1	three-toed sloth	0.9146709999999999
2154	1	Maltese dog	0.8575309999999999
2155	1	Italian greyhound	0.176053
2156	1	desktop computer	0.086502
2157	1	hen	0.965932
2158	1	English setter	0.298617
2159	1	Lhasa	0.58233
2160	1	bloodhound	0.195217
2161	1	pug	0.489814
2162	1	Walker hound	0.260857
2163	1	Gordon setter	0.503672
2164	1	golden retriever	0.77593
2165	1	miniature poodle	0.201493
2166	1	shopping cart	0.962465
2167	1	chow	0.692517
2168	1	box turtle	0.9330120000000001
2169	1	Bernese mountain dog	0.651137
2170	1	miniature pinscher	0.560311
2171	1	Rhodesian ridgeback	0.408143
2172	1	German shepherd	0.596461
2173	1	redbone	0.506826
2174	1	Welsh springer spaniel	0.465074

	is_prediction_1_dog	prediction2	prediction_2_confidence	\
0	False	bagel	0.08585110000000001	
1	True	Pekinese	0.0906465	
2	True	malamute	0.078253	
3	False	Labrador retriever	0.16808599999999999	
4	True	English springer	0.22576999999999997	
5	True	Irish terrier	0.116317	
6	True	Border collie	0.199287	
7	True	Eskimo dog	0.17840599999999998	
8	True	Irish setter	0.193054	
9	True	Cardigan	0.451038	
10	True	Pomeranian	0.0138835	
11	True	Labrador retriever	0.151317	
12	True	Cardigan	0.0273557	
13	True	boxer	0.00212864	

14	True	Labrador retriever	0.0138341
15	True	borzoi	0.194742
16	True	Tibetan mastiff	0.12018399999999999
17	True	Labrador retriever	0.184172
18	True	Eskimo dog	0.166511
19	True	pug	0.0009319080000000001
20	True	redbone	0.0875815
21	False	tow truck	0.0291754
22	True	Rhodesian ridgeback	0.05495
23	True	sea lion	0.275645
24	True	Eskimo dog	0.0350295
25	True	toy terrier	0.143528
26	True	Great Pyrenees	0.186136
27	False	sports car	0.139952
28	True	Chihuahua	0.00036117800000000003
29	NaN	NaN	NaN
...	...	...	...
2145	False	barracouta	0.271485
2146	True	bloodhound	0.24422
2147	True	toy terrier	0.00939697
2148	True	Afghan hound	0.25955100000000003
2149	False	skunk	0.00240245
2150	True	Siberian husky	0.147655
2151	True	Chesapeake Bay retriever	0.054786800000000004
2152	False	Newfoundland	0.278407
2153	False	otter	0.01525
2154	True	toy poodle	0.0630638
2155	True	toy terrier	0.111884
2156	False	desk	0.0855474
2157	False	cock	0.0339194
2158	True	Newfoundland	0.149842
2159	True	Shih-Tzu	0.166192
2160	True	German shepherd	0.0782598
2161	True	bull mastiff	0.40472199999999997
2162	True	English foxhound	0.17538199999999998
2163	True	Yorkshire terrier	0.174201
2164	True	Tibetan mastiff	0.0937178
2165	True	komondor	0.192305
2166	False	shopping basket	0.014593799999999999
2167	True	Tibetan mastiff	0.058279399999999995
2168	False	mud turtle	0.04588540000000001
2169	True	English springer	0.263788
2170	True	Rottweiler	0.243682
2171	True	redbone	0.360687
2172	True	malinois	0.13858399999999998
2173	True	miniature pinscher	0.07419169999999999
2174	True	collie	0.156665

	is_prediction_2_dog	prediction3	prediction_3_confidence	\
0	False	banana	0.07611	
1	True	papillon	0.0689569	
2	True	kelpie	0.0313789	
3	True	spatula	0.0408359	
4	True	German short-haired pointer	0.175219	
5	True	Indian elephant	0.07690219999999999	
6	True	ice lolly	0.193548	
7	True	Pembroke	0.0765069	
8	True	Chesapeake Bay retriever	0.118184	
9	True	Chihuahua	0.029248200000000002	
10	True	chow	0.00816748	
11	True	muzzle	0.0829811	
12	True	basenji	0.00463323	
13	True	Staffordshire bullterrier	0.00149818	
14	True	redbone	0.00795775	
15	True	Saluki	0.027350700000000002	
16	True	Labrador retriever	0.10550599999999999	
17	True	English setter	0.0734817	
18	True	malamute	0.111411	
19	True	bull mastiff	0.000903211	
20	True	Weimaraner	0.026236400000000003	
21	False	shopping cart	0.0263208	
22	True	beagle	0.0389148	
23	False	Weimaraner	0.134203	
24	True	Staffordshire bullterrier	0.029704700000000004	
25	True	can opener	0.0322529	
26	True	Dandie Dinmont	0.08634630000000001	
27	False	car wheel	0.044172699999999995	
28	True	Boston bull	7.55616e-05	
29	NaN	NaN	NaN	
...	...	...	...	
2145	False	gar	0.189945	
2146	True	flat-coated retriever	0.17381	
2147	True	papillon	0.00457681	
2148	True	briard	0.20680300000000001	
2149	False	hamster	0.00046086300000000005	
2150	True	Eskimo dog	0.09341239999999999	
2151	True	Labrador retriever	0.014240899999999999	
2152	True	groenendael	0.10264300000000001	
2153	False	great grey owl	0.0132072	
2154	True	miniature poodle	0.0255806	
2155	True	basenji	0.111152	
2156	False	bookcase	0.0794797	
2157	False	partridge	5.20658e-05	
2158	True	borzoi	0.133649	
2159	True	Dandie Dinmont	0.0896883	
2160	True	malinois	0.07562780000000001	

2161	True	French bulldog	0.0489595
2162	True	Ibizan hound	0.0974705
2163	True	Pekinese	0.109454
2164	True	Labrador retriever	0.07242660000000001
2165	True	soft-coated wheaten terrier	0.08208610000000001
2166	False	golden retriever	0.00795896
2167	True	fur coat	0.0544486
2168	False	terrapin	0.017885299999999996
2169	True	Greater Swiss Mountain dog	0.0161992
2170	True	Doberman	0.154629
2171	True	miniature pinscher	0.222752
2172	True	bloodhound	0.11619700000000001
2173	True	Rhodesian ridgeback	0.07201
2174	True	Shetland sheepdog	0.0614285

	is_prediction_3_dog
0	False
1	True
2	True
3	False
4	True
5	False
6	False
7	True
8	True
9	True
10	True
11	False
12	True
13	True
14	True
15	True
16	True
17	True
18	True
19	True
20	True
21	False
22	True
23	True
24	True
25	False
26	True
27	False
28	True
29	NaN
...	...
2145	False

```
2146                True
2147                True
2148                True
2149               False
2150                True
2151                True
2152                True
2153               False
2154                True
2155                True
2156               False
2157               False
2158                True
2159                True
2160                True
2161                True
2162                True
2163                True
2164                True
2165                True
2166                True
2167               False
2168               False
2169                True
2170                True
2171                True
2172                True
2173                True
2174                True

[2175 rows x 27 columns]
```

```
In [275]: # Check column names
          archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2174
Data columns (total 27 columns):
tweet_id                      2175 non-null object
in_reply_to_status_id         2175 non-null object
in_reply_to_user_id           2175 non-null object
timestamp                     2175 non-null datetime64[ns]
source                        2175 non-null object
text                          2175 non-null object
retweeted_status_id           2175 non-null object
retweeted_status_user_id      2175 non-null object
retweeted_status_timestamp    0 non-null datetime64[ns]
expanded_urls                 2175 non-null object
```

```
rating_numerator              2175 non-null float64
rating_denominator            2175 non-null int64
name                          2175 non-null object
meme                          2175 non-null object
retweet_count                 2175 non-null int64
favorite_count                2175 non-null int64
jpg_url                       1994 non-null object
img_num                       1994 non-null object
prediction1                   1994 non-null object
prediction_1_confidence       1994 non-null object
is_prediction_1_dog           1994 non-null object
prediction2                   1994 non-null object
prediction_2_confidence       1994 non-null object
is_prediction_2_dog           1994 non-null object
prediction3                   1994 non-null object
prediction_3_confidence       1994 non-null object
is_prediction_3_dog           1994 non-null object
dtypes: datetime64[ns](2), float64(1), int64(3), object(21)
memory usage: 475.8+ KB
```

2.3.1 Storing Data

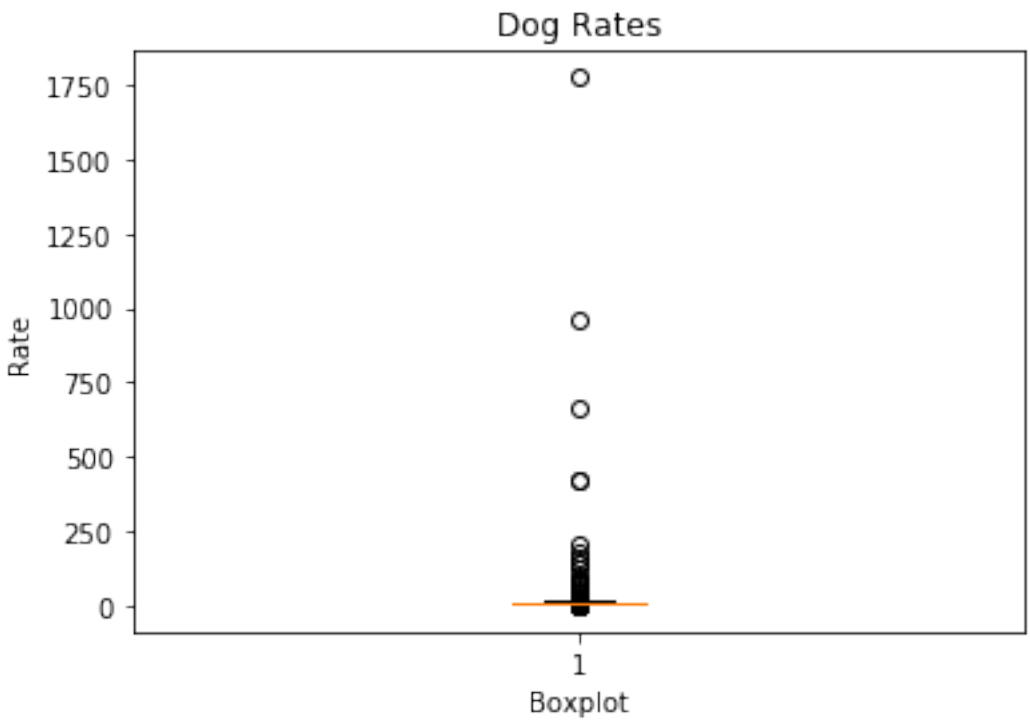
```
In [276]: # Saving file names
          archive_file_name_to_save  = "twitter_archive_master.csv"

In [277]: # Save tables to the local disk as flat (CSV) file
          archive_clean.to_csv(archive_file_name_to_save)
```

2.3.2 Analyze

```
In [278]: #Copy clean datasets to visualize
          archive_analyze = archive_clean.copy()
          image_prediction_analyze = image_prediction_clean.copy()
```

Check the outliers for the picture ratings. I have seen some in the data set, let's visualize them in a boxplot.

```
In [279]: plt.boxplot(archive_analyze["rating_numerator"])
          plt.xlabel("Boxplot")
          plt.ylabel("Rate")
          plt.title("Dog Rates")
          plt.savefig("boxplot.png")
```




```
In [280]: archive_analyze.describe()
```

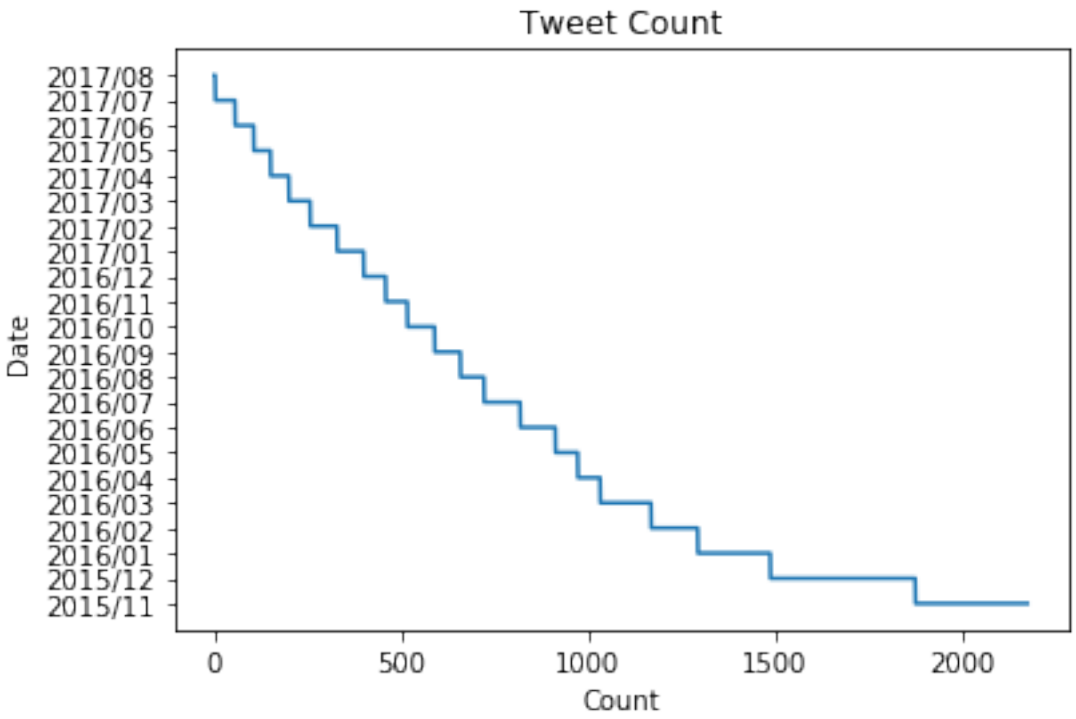
Out[280]:

	rating_numerator	rating_denominator	retweet_count	favorite_count
count	2175.000000	2175.000000	2175.000000	2175.000000
mean	13.221149	10.492874	2724.261609	8732.824368
std	47.725112	7.019084	4686.940457	12430.526513
min	0.000000	0.000000	0.000000	51.000000
25%	10.000000	10.000000	593.000000	1875.500000
50%	11.000000	10.000000	1311.000000	3967.000000
75%	12.000000	10.000000	3136.000000	10925.500000
max	1776.000000	170.000000	77746.000000	143985.000000

There are some very big outliers. Funny how they rate dogs.
How does the tweet count of this WeRateDogs site change over time? Let's see.

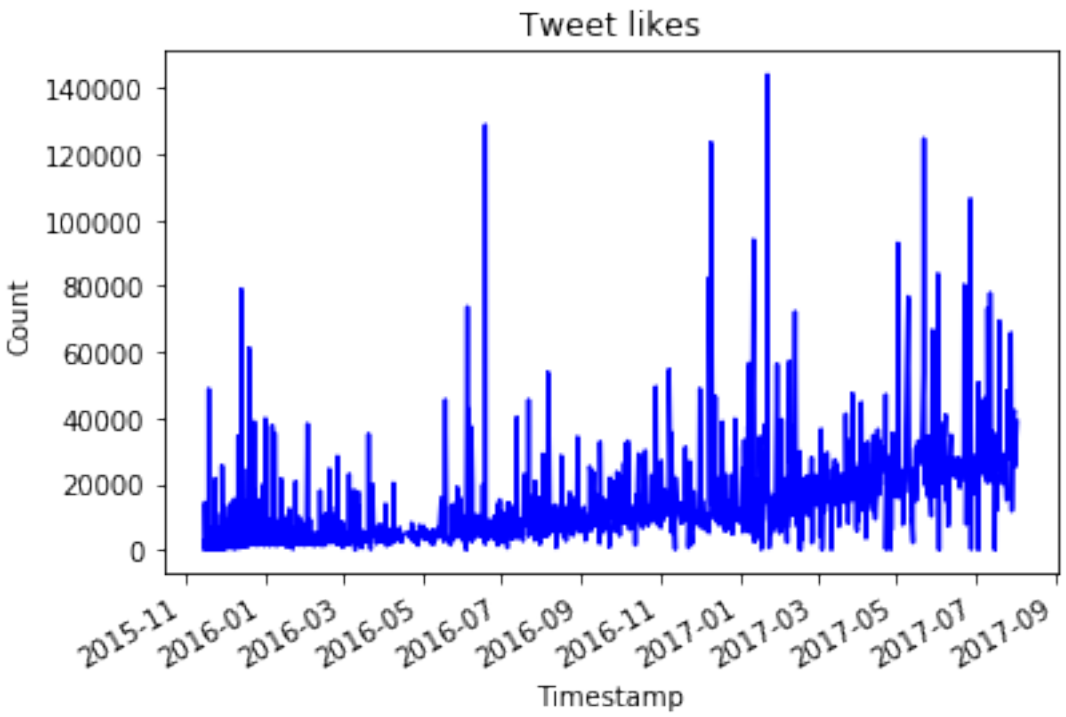
```
In [281]: #remove days from timestamp
          tweet_date_by_year_and_month = archive_analyze["timestamp"].apply(lambda x:x.strftime
```

```
In [282]: #Plot year/month tweet count
          plt.plot((tweet_date_by_year_and_month))
          plt.xlabel('Count')
          plt.ylabel('Date')
          plt.savefig('tweet_count.png')
          plt.title("Tweet Count")
          plt.show()
```



Tweet count is decreasing. What about tweet likes? Is it decreasing too? Maybe the audience is not falling as tweet count does.

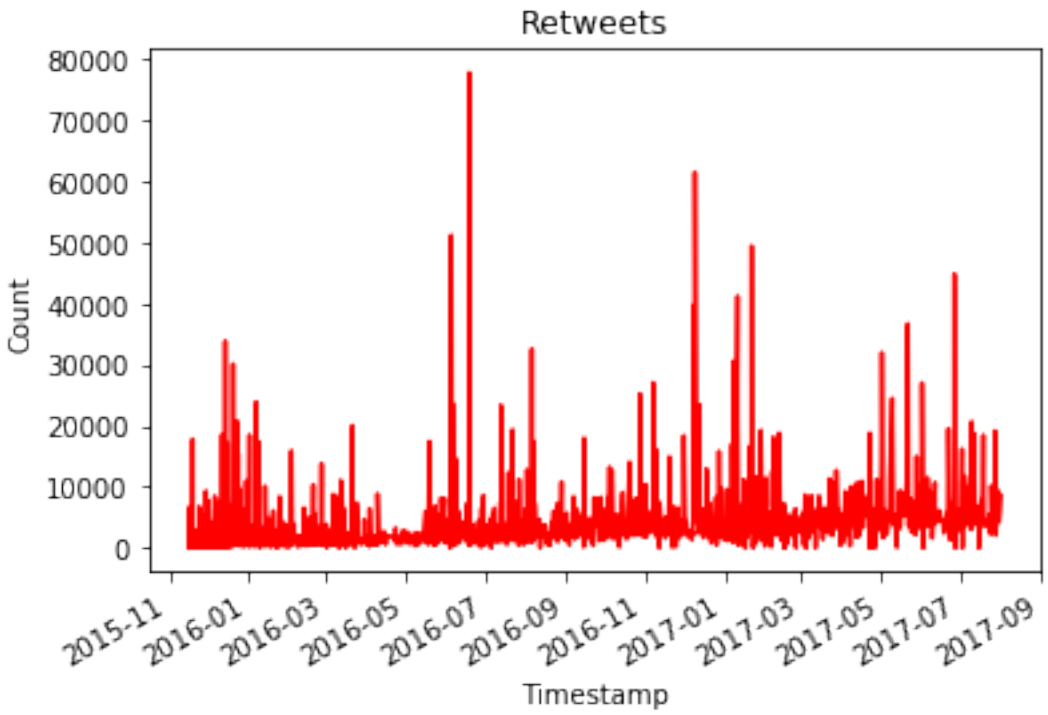
```
In [283]: # Set index to timestampt_by_months
          archive_analyze.set_index("timestamp", inplace = True)

In [284]: archive_analyze["favorite_count"].plot(color = "blue")
          plt.xlabel('Timestamp')
          plt.ylabel('Count')
          plt.savefig('tweet_likes.png')
          plt.title('Tweet likes')
          plt.show()
```



Like count is growing, se people send less pictures to the tweet site, but the audience is growing.

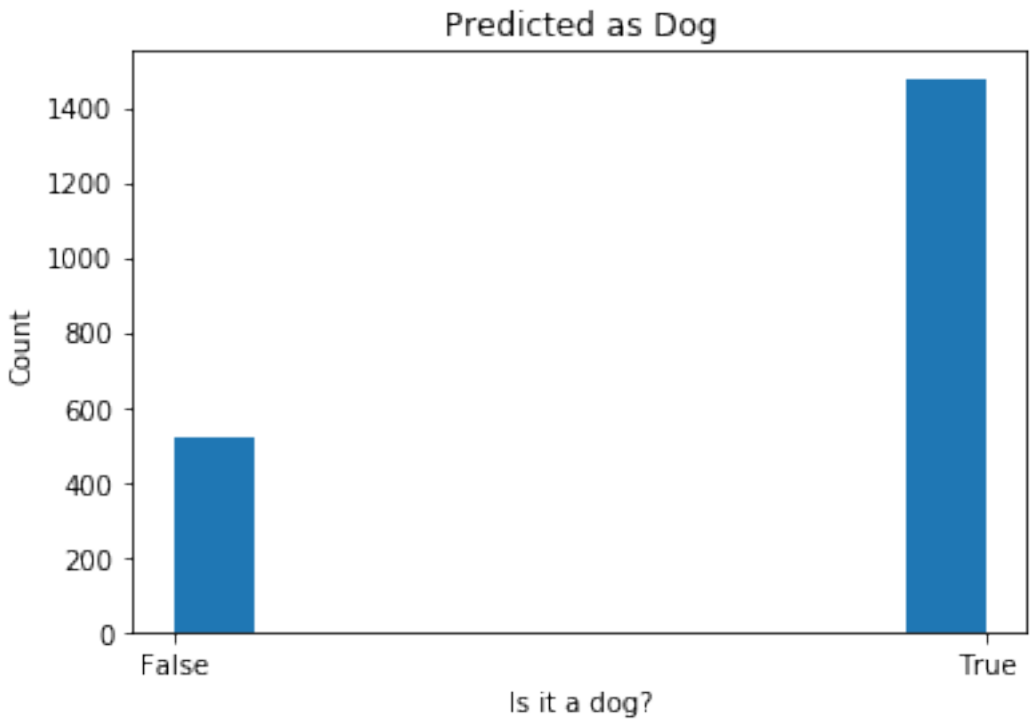
What about retweets?

```
In [285]: archive_analyze["retweet_count"].plot(color = "red")
          plt.xlabel('Timestamp')
          plt.ylabel('Count')
          plt.savefig('retweets.png')
          plt.title('Retweets')
          plt.show()
```



Retweet count is growing too.
What is the proportion of the dogs in the tweets calculated from the first prediction?

```
In [286]: image_prediction_analyze.is_prediction_1_dog = image_prediction_analyze.is_predictio

In [287]: plt.hist(image_prediction_analyze.is_prediction_1_dog, histtype = "barstacked")
          plt.savefig("is_prediction_dog.png")
          plt.title('Predicted as Dog')
          plt.xlabel('Is it a dog?')
          plt.ylabel('Count')
          plt.show()
```



Around 75% of the pictures are predicted as dogs. This is strange, because this is supposed to be a dog rater tweet site.

2.4 Sources

https://stackoverflow.com/
https://pandas.pydata.org/
https://matplotlib.org/
https://eu.udacity.com/