wrangle_act

November 30, 2017

0.1 Gather

```
In [67]: import pandas as pd
         import numpy as np
         import requests
         import os
         import tweepy
         import json
         import matplotlib.pyplot as plt
```

```
In [68]: twitter_archive = pd.read_csv('Source_Files/twitter-archive-enhanced.csv')

         folder_name = 'Source_Files'
         url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predicti
         response = requests.get(url)
         with open(os.path.join(folder_name, url.split('/')[-1]), mode = 'wb') as file:
                 file.write(response.content)

         image_predictions = pd.read_csv('Source_Files/image-predictions.tsv', sep = '\t')
```

```
In [69]: consumer_key = 'consumer_key'
         consumer_secret = 'consumer_secret'
         access_token = 'access_token'
         access_secret = 'access_secret'

         auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
         auth.set_access_token(access_token, access_secret)

         api = tweepy.API(auth_handler=auth, wait_on_rate_limit=True, wait_on_rate_limit_notify=

         tweet_id = twitter_archive['tweet_id']
         data = {}
         data['twitter_data'] = []

         for id_of_tweet in tweet_id:
             try:
                 tweet = api.get_status(id_of_tweet, tweet_mode='extended')
                 twitter_id = tweet._json['id']
```

```
        favorite_count = tweet._json['favorite_count']
        retweet_count = tweet._json['retweet_count']
        data['twitter_data'].append({'tweet_id' : twitter_id,
                                    'favorite_count' : int(favorite_count),
                                    'retweet_count' : int(retweet_count)})

    except:
        data['twitter_data'].append({'tweet_id' : 'id not found',
                                    'favorite_count' : int(favorite_count),
                                    'retweet_count' : int(retweet_count)})

with open('Source_Files/tweet_json.txt', 'w') as outfile:
    json.dump(data,outfile)
```

```
Rate limit reached. Sleeping for: 727
Rate limit reached. Sleeping for: 729
```

```
In [70]: df_list = []
         with open('Source_Files/tweet_json.txt') as json_file:
             data = json.load(json_file)
             for tweet in data['twitter_data']:
                 df_list.append({'tweet_id': tweet['tweet_id'],
                                 'favorite_count': tweet['favorite_count'],
                                 'retweet_count': tweet['retweet_count']})
```

```
In [71]: tweet_archive = pd.DataFrame(df_list, columns = ['tweet_id', 'favorite_count', 'retweet
```

```
In [72]: twitter_archive
```

```
Out[72]:               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
         5     891087950875897856                    NaN                  NaN
         6     890971913173991426                    NaN                  NaN
         7     890729181411237888                    NaN                  NaN
         8     890609185150312448                    NaN                  NaN
         9     890240255349198849                    NaN                  NaN
         10    890006608113172480                    NaN                  NaN
         11    889880896479866881                    NaN                  NaN
         12    889665388333682689                    NaN                  NaN
         13    889638837579907072                    NaN                  NaN
         14    889531135344209921                    NaN                  NaN
         15    889278841981685760                    NaN                  NaN
         16    888917238123831296                    NaN                  NaN
         17    888804989199671297                    NaN                  NaN
```

```
18    888554962724278272                    NaN                  NaN
19    888202515573088257                    NaN                  NaN
20    888078434458587136                    NaN                  NaN
21    887705289381826560                    NaN                  NaN
22    887517139158093824                    NaN                  NaN
23    887473957103951883                    NaN                  NaN
24    887343217045368832                    NaN                  NaN
25    887101392804085760                    NaN                  NaN
26    886983233522544640                    NaN                  NaN
27    886736880519319552                    NaN                  NaN
28    886680336477933568                    NaN                  NaN
29    886366144734445568                    NaN                  NaN
...                  ...                    ...                  ...
2326  666411507551481857                    NaN                  NaN
2327  666407126856765440                    NaN                  NaN
2328  666396247373291520                    NaN                  NaN
2329  666373753744588802                    NaN                  NaN
2330  666362758909284353                    NaN                  NaN
2331  666353288456101888                    NaN                  NaN
2332  666345417576210432                    NaN                  NaN
2333  666337882303524864                    NaN                  NaN
2334  666293911632134144                    NaN                  NaN
2335  666287406224695296                    NaN                  NaN
2336  666273097616637952                    NaN                  NaN
2337  666268910803644416                    NaN                  NaN
2338  666104133288665088                    NaN                  NaN
2339  666102155909144576                    NaN                  NaN
2340  666099513787052032                    NaN                  NaN
2341  666094000022159362                    NaN                  NaN
2342  666082916733198337                    NaN                  NaN
2343  666073100786774016                    NaN                  NaN
2344  666071193221509120                    NaN                  NaN
2345  666063827256086533                    NaN                  NaN
2346  666058600524156928                    NaN                  NaN
2347  666057090499244032                    NaN                  NaN
2348  666055525042405380                    NaN                  NaN
2349  666051853826850816                    NaN                  NaN
2350  666050758794694657                    NaN                  NaN
2351  666049248165822465                    NaN                  NaN
2352  666044226329800704                    NaN                  NaN
2353  666033412701032449                    NaN                  NaN
2354  666029285002620928                    NaN                  NaN
2355  666020888022790149                    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
5     2017-07-29 00:08:17 +0000
6     2017-07-28 16:27:12 +0000
7     2017-07-28 00:22:40 +0000
8     2017-07-27 16:25:51 +0000
9     2017-07-26 15:59:51 +0000
10    2017-07-26 00:31:25 +0000
11    2017-07-25 16:11:53 +0000
12    2017-07-25 01:55:32 +0000
13    2017-07-25 00:10:02 +0000
14    2017-07-24 17:02:04 +0000
15    2017-07-24 00:19:32 +0000
16    2017-07-23 00:22:39 +0000
17    2017-07-22 16:56:37 +0000
18    2017-07-22 00:23:06 +0000
19    2017-07-21 01:02:36 +0000
20    2017-07-20 16:49:33 +0000
21    2017-07-19 16:06:48 +0000
22    2017-07-19 03:39:09 +0000
23    2017-07-19 00:47:34 +0000
24    2017-07-18 16:08:03 +0000
25    2017-07-18 00:07:08 +0000
26    2017-07-17 16:17:36 +0000
27    2017-07-16 23:58:41 +0000
28    2017-07-16 20:14:00 +0000
29    2017-07-15 23:25:31 +0000
...                         ...
2326  2015-11-17 00:24:19 +0000
2327  2015-11-17 00:06:54 +0000
2328  2015-11-16 23:23:41 +0000
2329  2015-11-16 21:54:18 +0000
2330  2015-11-16 21:10:36 +0000
2331  2015-11-16 20:32:58 +0000
2332  2015-11-16 20:01:42 +0000
2333  2015-11-16 19:31:45 +0000
2334  2015-11-16 16:37:02 +0000
2335  2015-11-16 16:11:11 +0000
2336  2015-11-16 15:14:19 +0000
2337  2015-11-16 14:57:41 +0000
2338  2015-11-16 04:02:55 +0000
2339  2015-11-16 03:55:04 +0000
2340  2015-11-16 03:44:34 +0000
2341  2015-11-16 03:22:39 +0000
2342  2015-11-16 02:38:37 +0000
2343  2015-11-16 01:59:36 +0000
2344  2015-11-16 01:52:02 +0000
2345  2015-11-16 01:22:45 +0000
```

```
2346  2015-11-16 01:01:59 +0000
2347  2015-11-16 00:55:59 +0000
2348  2015-11-16 00:49:46 +0000
2349  2015-11-16 00:35:11 +0000
2350  2015-11-16 00:30:50 +0000
2351  2015-11-16 00:24:50 +0000
2352  2015-11-16 00:04:52 +0000
2353  2015-11-15 23:21:54 +0000
2354  2015-11-15 23:05:30 +0000
2355  2015-11-15 22:32:08 +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...
5     <a href="http://twitter.com/download/iphone" r...
6     <a href="http://twitter.com/download/iphone" r...
7     <a href="http://twitter.com/download/iphone" r...
8     <a href="http://twitter.com/download/iphone" r...
9     <a href="http://twitter.com/download/iphone" r...
10    <a href="http://twitter.com/download/iphone" r...
11    <a href="http://twitter.com/download/iphone" r...
12    <a href="http://twitter.com/download/iphone" r...
13    <a href="http://twitter.com/download/iphone" r...
14    <a href="http://twitter.com/download/iphone" r...
15    <a href="http://twitter.com/download/iphone" r...
16    <a href="http://twitter.com/download/iphone" r...
17    <a href="http://twitter.com/download/iphone" r...
18    <a href="http://twitter.com/download/iphone" r...
19    <a href="http://twitter.com/download/iphone" r...
20    <a href="http://twitter.com/download/iphone" r...
21    <a href="http://twitter.com/download/iphone" r...
22    <a href="http://twitter.com/download/iphone" r...
23    <a href="http://twitter.com/download/iphone" r...
24    <a href="http://twitter.com/download/iphone" r...
25    <a href="http://twitter.com/download/iphone" r...
26    <a href="http://twitter.com/download/iphone" r...
27    <a href="http://twitter.com/download/iphone" r...
28    <a href="http://twitter.com/download/iphone" r...
29    <a href="http://twitter.com/download/iphone" r...
...                                                 ...
2326  <a href="http://twitter.com/download/iphone" r...
2327  <a href="http://twitter.com/download/iphone" r...
2328  <a href="http://twitter.com/download/iphone" r...
2329  <a href="http://twitter.com/download/iphone" r...
2330  <a href="http://twitter.com/download/iphone" r...
```

```
2331  <a href="http://twitter.com/download/iphone" r...
2332  <a href="http://twitter.com/download/iphone" r...
2333  <a href="http://twitter.com/download/iphone" r...
2334  <a href="http://twitter.com/download/iphone" r...
2335  <a href="http://twitter.com/download/iphone" r...
2336  <a href="http://twitter.com/download/iphone" r...
2337  <a href="http://twitter.com/download/iphone" r...
2338  <a href="http://twitter.com/download/iphone" r...
2339  <a href="http://twitter.com/download/iphone" r...
2340  <a href="http://twitter.com/download/iphone" r...
2341  <a href="http://twitter.com/download/iphone" r...
2342  <a href="http://twitter.com/download/iphone" r...
2343  <a href="http://twitter.com/download/iphone" r...
2344  <a href="http://twitter.com/download/iphone" r...
2345  <a href="http://twitter.com/download/iphone" r...
2346  <a href="http://twitter.com/download/iphone" r...
2347  <a href="http://twitter.com/download/iphone" r...
2348  <a href="http://twitter.com/download/iphone" r...
2349  <a href="http://twitter.com/download/iphone" r...
2350  <a href="http://twitter.com/download/iphone" r...
2351  <a href="http://twitter.com/download/iphone" r...
2352  <a href="http://twitter.com/download/iphone" r...
2353  <a href="http://twitter.com/download/iphone" r...
2354  <a href="http://twitter.com/download/iphone" r...
2355  <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
5     Here we have a majestic great white breaching ...                  NaN
6     Meet Jax. He enjoys ice cream so much he gets ...                  NaN
7     When you watch your owner call another dog a g...                  NaN
8     This is Zoey. She doesn't want to be one of th...                  NaN
9     This is Cassie. She is a college pup. Studying...                  NaN
10    This is Koda. He is a South Australian decksha...                  NaN
11    This is Bruno. He is a service shark. Only get...                  NaN
12    Here's a puppo that seems to be on the fence a...                  NaN
13    This is Ted. He does his best. Sometimes that'...                  NaN
14    This is Stuart. He's sporting his favorite fan...                  NaN
15    This is Oliver. You're witnessing one of his m...                  NaN
16    This is Jim. He found a fren. Taught him how t...                  NaN
17    This is Zeke. He has a new stick. Very proud o...                  NaN
18    This is Ralphus. He's powering up. Attempting ...                  NaN
19    RT @dog_rates: This is Canela. She attempted s...         8.874740e+17
20    This is Gerald. He was just told he didn't get...                  NaN
```

```
21    This is Jeffrey. He has a monopoly on the pool...                 NaN
22    I've yet to rate a Venezuelan Hover Wiener. Th...                 NaN
23    This is Canela. She attempted some fancy porch...                 NaN
24    You may not have known you needed to see this ...                 NaN
25    This... is a Jubilant Antarctic House Bear. We...                 NaN
26    This is Maya. She's very shy. Rarely leaves he...                 NaN
27    This is Mingus. He's a wonderful father to his...                 NaN
28    This is Derek. He's late for a dog meeting. 13...                 NaN
29    This is Roscoe. Another pupper fallen victim t...                 NaN
...                                                 ...                 ...
2326  This is quite the dog. Gets really excited whe...                 NaN
2327  This is a southern Vesuvius bumblegruff. Can d...                 NaN
2328  Oh goodness. A super rare northeast Qdoba kang...                 NaN
2329  Those are sunglasses and a jean jacket. 11/10 ...                 NaN
2330  Unique dog here. Very small. Lives in containe...                 NaN
2331  Here we have a mixed Asiago from the Galápagos...                 NaN
2332  Look at this jokester thinking seat belt laws ...                 NaN
2333  This is an extremely rare horned Parthenon. No...                 NaN
2334  This is a funny dog. Weird toes. Won't come do...                 NaN
2335  This is an Albanian 3 1/2 legged  Episcopalian...                 NaN
2336     Can take selfies 11/10 https://t.co/ws2AMaNwPW                 NaN
2337  Very concerned about fellow dog trapped in com...                 NaN
2338  Not familiar with this breed. No tail (weird)...                 NaN
2339  Oh my. Here you are seeing an Adobe Setter giv...                 NaN
2340  Can stand on stump for what seems like a while...                 NaN
2341  This appears to be a Mongolian Presbyterian mi...                 NaN
2342  Here we have a well-established sunblockerspan...                 NaN
2343  Let's hope this flight isn't Malaysian (lol). ...                 NaN
2344  Here we have a northern speckled Rhododendron...                 NaN
2345  This is the happiest dog you will ever see. Ve...                 NaN
2346  Here is the Rand Paul of retrievers folks! He'...                 NaN
2347  My oh my. This is a rare blond Canadian terrie...                 NaN
2348  Here is a Siberian heavily armored polar bear ...                 NaN
2349  This is an odd dog. Hard on the outside but lo...                 NaN
2350  This is a truly beautiful English Wilson Staff...                 NaN
2351  Here we have a 1949 1st generation vulpix. Enj...                 NaN
2352  This is a purebred Piers Morgan. Loves to Netf...                 NaN
2353  Here is a very happy pup. Big fan of well-main...                 NaN
2354  This is a western brown Mitsubishi terrier. Up...                 NaN
2355  Here we have a Japanese Irish Setter. Lost eye...                 NaN

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

```
6                          NaN                        NaN
7                          NaN                        NaN
8                          NaN                        NaN
9                          NaN                        NaN
10                         NaN                        NaN
11                         NaN                        NaN
12                         NaN                        NaN
13                         NaN                        NaN
14                         NaN                        NaN
15                         NaN                        NaN
16                         NaN                        NaN
17                         NaN                        NaN
18                         NaN                        NaN
19                4.196984e+09  2017-07-19 00:47:34 +0000
20                         NaN                        NaN
21                         NaN                        NaN
22                         NaN                        NaN
23                         NaN                        NaN
24                         NaN                        NaN
25                         NaN                        NaN
26                         NaN                        NaN
27                         NaN                        NaN
28                         NaN                        NaN
29                         NaN                        NaN
...                        ...                        ...
2326                       NaN                        NaN
2327                       NaN                        NaN
2328                       NaN                        NaN
2329                       NaN                        NaN
2330                       NaN                        NaN
2331                       NaN                        NaN
2332                       NaN                        NaN
2333                       NaN                        NaN
2334                       NaN                        NaN
2335                       NaN                        NaN
2336                       NaN                        NaN
2337                       NaN                        NaN
2338                       NaN                        NaN
2339                       NaN                        NaN
2340                       NaN                        NaN
2341                       NaN                        NaN
2342                       NaN                        NaN
2343                       NaN                        NaN
2344                       NaN                        NaN
2345                       NaN                        NaN
2346                       NaN                        NaN
2347                       NaN                        NaN
2348                       NaN                        NaN
```

```
2349                       NaN                     NaN
2350                       NaN                     NaN
2351                       NaN                     NaN
2352                       NaN                     NaN
2353                       NaN                     NaN
2354                       NaN                     NaN
2355                       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
5     https://twitter.com/dog_rates/status/891087950...                13
6     https://gofundme.com/ydvmve-surgery-for-jax,ht...                13
7     https://twitter.com/dog_rates/status/890729181...                13
8     https://twitter.com/dog_rates/status/890609185...                13
9     https://twitter.com/dog_rates/status/890240255...                14
10    https://twitter.com/dog_rates/status/890006608...                13
11    https://twitter.com/dog_rates/status/889880896...                13
12    https://twitter.com/dog_rates/status/889665388...                13
13    https://twitter.com/dog_rates/status/889638837...                12
14    https://twitter.com/dog_rates/status/889531135...                13
15    https://twitter.com/dog_rates/status/889278841...                13
16    https://twitter.com/dog_rates/status/888917238...                12
17    https://twitter.com/dog_rates/status/888804989...                13
18    https://twitter.com/dog_rates/status/888554962...                13
19    https://twitter.com/dog_rates/status/887473957...                13
20    https://twitter.com/dog_rates/status/888078434...                12
21    https://twitter.com/dog_rates/status/887705289...                13
22    https://twitter.com/dog_rates/status/887517139...                14
23    https://twitter.com/dog_rates/status/887473957...                13
24    https://twitter.com/dog_rates/status/887343217...                13
25    https://twitter.com/dog_rates/status/887101392...                12
26    https://twitter.com/dog_rates/status/886983233...                13
27    https://www.gofundme.com/mingusneedsus,https:/...                13
28    https://twitter.com/dog_rates/status/886680336...                13
29    https://twitter.com/dog_rates/status/886366144...                12
...                                                 ...               ...
2326  https://twitter.com/dog_rates/status/666411507...                 2
2327  https://twitter.com/dog_rates/status/666407126...                 7
2328  https://twitter.com/dog_rates/status/666396247...                 9
2329  https://twitter.com/dog_rates/status/666373753...                11
2330  https://twitter.com/dog_rates/status/666362758...                 6
2331  https://twitter.com/dog_rates/status/666353288...                 8
2332  https://twitter.com/dog_rates/status/666345417...                10
2333  https://twitter.com/dog_rates/status/666337882...                 9
```

```
2334  https://twitter.com/dog_rates/status/666293911...                  3
2335  https://twitter.com/dog_rates/status/666287406...                  1
2336  https://twitter.com/dog_rates/status/666273097...                 11
2337  https://twitter.com/dog_rates/status/666268910...                 10
2338  https://twitter.com/dog_rates/status/666104133...                  1
2339  https://twitter.com/dog_rates/status/666102155...                 11
2340  https://twitter.com/dog_rates/status/666099513...                  8
2341  https://twitter.com/dog_rates/status/666094000...                  9
2342  https://twitter.com/dog_rates/status/666082916...                  6
2343  https://twitter.com/dog_rates/status/666073100...                 10
2344  https://twitter.com/dog_rates/status/666071193...                  9
2345  https://twitter.com/dog_rates/status/666063827...                 10
2346  https://twitter.com/dog_rates/status/666058600...                  8
2347  https://twitter.com/dog_rates/status/666057090...                  9
2348  https://twitter.com/dog_rates/status/666055525...                 10
2349  https://twitter.com/dog_rates/status/666051853...                  2
2350  https://twitter.com/dog_rates/status/666050758...                 10
2351  https://twitter.com/dog_rates/status/666049248...                  5
2352  https://twitter.com/dog_rates/status/666044226...                  6
2353  https://twitter.com/dog_rates/status/666033412...                  9
2354  https://twitter.com/dog_rates/status/666029285...                  7
2355  https://twitter.com/dog_rates/status/666020888...                  8

      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
5                     10      None   None    None    None   None
6                     10       Jax   None    None    None   None
7                     10      None   None    None    None   None
8                     10      Zoey   None    None    None   None
9                     10    Cassie  doggo    None    None   None
10                    10      Koda   None    None    None   None
11                    10     Bruno   None    None    None   None
12                    10      None   None    None    None  puppo
13                    10       Ted   None    None    None   None
14                    10    Stuart   None    None    None  puppo
15                    10    Oliver   None    None    None   None
16                    10       Jim   None    None    None   None
17                    10      Zeke   None    None    None   None
18                    10   Ralphus   None    None    None   None
19                    10    Canela   None    None    None   None
20                    10    Gerald   None    None    None   None
21                    10   Jeffrey   None    None    None   None
22                    10      such   None    None    None   None
23                    10    Canela   None    None    None   None
```

```
24                    10      None   None    None    None   None
25                    10      None   None    None    None   None
26                    10      Maya   None    None    None   None
27                    10    Mingus   None    None    None   None
28                    10     Derek   None    None    None   None
29                    10    Roscoe   None    None  pupper   None
...                  ...       ...    ...     ...     ...    ...
2326                  10     quite   None    None    None   None
2327                  10         a   None    None    None   None
2328                  10      None   None    None    None   None
2329                  10      None   None    None    None   None
2330                  10      None   None    None    None   None
2331                  10      None   None    None    None   None
2332                  10      None   None    None    None   None
2333                  10        an   None    None    None   None
2334                  10         a   None    None    None   None
2335                   2        an   None    None    None   None
2336                  10      None   None    None    None   None
2337                  10      None   None    None    None   None
2338                  10      None   None    None    None   None
2339                  10      None   None    None    None   None
2340                  10      None   None    None    None   None
2341                  10      None   None    None    None   None
2342                  10      None   None    None    None   None
2343                  10      None   None    None    None   None
2344                  10      None   None    None    None   None
2345                  10       the   None    None    None   None
2346                  10       the   None    None    None   None
2347                  10         a   None    None    None   None
2348                  10         a   None    None    None   None
2349                  10        an   None    None    None   None
2350                  10         a   None    None    None   None
2351                  10      None   None    None    None   None
2352                  10         a   None    None    None   None
2353                  10         a   None    None    None   None
2354                  10         a   None    None    None   None
2355                  10      None   None    None    None   None

[2356 rows x 17 columns]
```

```
In [73]: image_predictions
```

```
Out[73]:                 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
10    666063827256086533    https://pbs.twimg.com/media/CT5Vg_wXIAAXfnj.jpg
11    666071193221509120    https://pbs.twimg.com/media/CT5cN_3WEAAlOoZ.jpg
12    666073100786774016    https://pbs.twimg.com/media/CT5d9DZXAAALcwe.jpg
13    666082916733198337    https://pbs.twimg.com/media/CT5m4VGWEAAtKc8.jpg
14    666094000022159362    https://pbs.twimg.com/media/CT5w9gUW4AAsBNN.jpg
15    666099513787052032    https://pbs.twimg.com/media/CT51-JJUEAA6hV8.jpg
16    666102155909144576    https://pbs.twimg.com/media/CT54YGiWUAEZnoK.jpg
17    666104133288665088    https://pbs.twimg.com/media/CT56LSZWoAAlJj2.jpg
18    666268910803644416    https://pbs.twimg.com/media/CT8QCd1WEAADXws.jpg
19    666273097616637952    https://pbs.twimg.com/media/CT8T1mtUwAA3aqm.jpg
20    666287406224695296    https://pbs.twimg.com/media/CT8g3BpUEAAuFjg.jpg
21    666293911632134144    https://pbs.twimg.com/media/CT8mx7KW4AEQu8N.jpg
22    666337882303524864    https://pbs.twimg.com/media/CT9OwFIWEAMuRje.jpg
23    666345417576210432    https://pbs.twimg.com/media/CT9Vn7PWoAA_ZCM.jpg
24    666353288456101888    https://pbs.twimg.com/media/CT9cx0tUEAAhNN_.jpg
25    666362758909284353    https://pbs.twimg.com/media/CT9lXGsUcAAyUFt.jpg
26    666373753744588802    https://pbs.twimg.com/media/CT9vZEYWUAAlZ05.jpg
27    666396247373291520    https://pbs.twimg.com/media/CT-D2ZHWIAA3gK1.jpg
28    666407126856765440    https://pbs.twimg.com/media/CT-NvwmW4AAugGZ.jpg
29    666411507551481857    https://pbs.twimg.com/media/CT-RugiWIAELEaq.jpg
...                  ...                                                ...
2045  886366144734445568    https://pbs.twimg.com/media/DE0BTnQUwAApKEH.jpg
2046  886680336477933568    https://pbs.twimg.com/media/DE4fEDzWAAAyHMM.jpg
2047  886736880519319552    https://pbs.twimg.com/media/DE5Se8FXcAAJFx4.jpg
2048  886983233522544640    https://pbs.twimg.com/media/DE8yicJW0AAAvBJ.jpg
2049  887101392804085760    https://pbs.twimg.com/media/DE-eAq6UwAA-jaE.jpg
2050  887343217045368832  https://pbs.twimg.com/ext_tw_video_thumb/88734...
2051  887473957103951883    https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg
2052  887517139158093824  https://pbs.twimg.com/ext_tw_video_thumb/88751...
2053  887705289381826560    https://pbs.twimg.com/media/DFHDQBbXgAEqY7t.jpg
2054  888078434458587136    https://pbs.twimg.com/media/DFMWn56WsAAkA7B.jpg
2055  888202515573088257    https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg
2056  888554962724278272    https://pbs.twimg.com/media/DFTH_O-UQAACu20.jpg
2057  888804989199671297    https://pbs.twimg.com/media/DFWra-3VYAA2piG.jpg
2058  888917238123831296    https://pbs.twimg.com/media/DFYRgsOUQAARGhO.jpg
2059  889278841981685760  https://pbs.twimg.com/ext_tw_video_thumb/88927...
2060  889531135344209921    https://pbs.twimg.com/media/DFg_2PVW0AEHN3p.jpg
2061  889638837579907072    https://pbs.twimg.com/media/DFihzFfXsAYGDPR.jpg
2062  889665388333682689    https://pbs.twimg.com/media/DFi579UWsAAatzw.jpg
2063  889880896479866881    https://pbs.twimg.com/media/DFl99B1WsAITKsg.jpg
2064  890006608113172480    https://pbs.twimg.com/media/DFnwSY4WAAAMliS.jpg
2065  890240255349198849    https://pbs.twimg.com/media/DFrEyVuW0AAO3t9.jpg
2066  890609185150312448    https://pbs.twimg.com/media/DFwUU__XcAEpyXI.jpg
```

```
2067  890729181411237888    https://pbs.twimg.com/media/DFyBahAVwAAhUTd.jpg
2068  890971913173991426    https://pbs.twimg.com/media/DF1eOmZXUAALUcq.jpg
2069  891087950875897856    https://pbs.twimg.com/media/DF3HwyEWsAABqE6.jpg
2070  891327558926688256    https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg
2071  891689557279858688    https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg
2072  891815181378084864    https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg
2073  892177421306343426    https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg
2074  892420643555336193    https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.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  0.933012   False
7           1                         chow  0.692517    True
8           1                shopping_cart  0.962465   False
9           1             miniature_poodle  0.201493    True
10          1             golden_retriever  0.775930    True
11          1                Gordon_setter  0.503672    True
12          1                 Walker_hound  0.260857    True
13          1                          pug  0.489814    True
14          1                   bloodhound  0.195217    True
15          1                        Lhasa  0.582330    True
16          1               English_setter  0.298617    True
17          1                          hen  0.965932   False
18          1             desktop_computer  0.086502   False
19          1            Italian_greyhound  0.176053    True
20          1                  Maltese_dog  0.857531    True
21          1             three-toed_sloth  0.914671   False
22          1                           ox  0.416669   False
23          1             golden_retriever  0.858744    True
24          1                     malamute  0.336874    True
25          1                   guinea_pig  0.996496   False
26          1  soft-coated_wheaten_terrier  0.326467    True
27          1                    Chihuahua  0.978108    True
28          1      black-and-tan_coonhound  0.529139    True
29          1                         coho  0.404640   False
...       ...                          ...       ...     ...
2045        1               French_bulldog  0.999201    True
2046        1                  convertible  0.738995   False
2047        1                       kuvasz  0.309706    True
2048        2                    Chihuahua  0.793469    True
2049        1                      Samoyed  0.733942    True
2050        1             Mexican_hairless  0.330741    True
2051        2                     Pembroke  0.809197    True
```

```
2052        1                    limousine  0.130432   False
2053        1                       basset  0.821664    True
2054        1               French_bulldog  0.995026    True
2055        2                     Pembroke  0.809197    True
2056        3               Siberian_husky  0.700377    True
2057        1             golden_retriever  0.469760    True
2058        1             golden_retriever  0.714719    True
2059        1                      whippet  0.626152    True
2060        1             golden_retriever  0.953442    True
2061        1               French_bulldog  0.991650    True
2062        1                     Pembroke  0.966327    True
2063        1               French_bulldog  0.377417    True
2064        1                      Samoyed  0.957979    True
2065        1                     Pembroke  0.511319    True
2066        1                Irish_terrier  0.487574    True
2067        2                   Pomeranian  0.566142    True
2068        1                  Appenzeller  0.341703    True
2069        1     Chesapeake_Bay_retriever  0.425595    True
2070        2                       basset  0.555712    True
2071        1                  paper_towel  0.170278   False
2072        1                    Chihuahua  0.716012    True
2073        1                    Chihuahua  0.323581    True
2074        1                       orange  0.097049   False

                            p2   p2_conf  p2_dog                           p3  \
0                       collie  0.156665    True            Shetland_sheepdog
1           miniature_pinscher  0.074192    True          Rhodesian_ridgeback
2                     malinois  0.138584    True                   bloodhound
3                      redbone  0.360687    True           miniature_pinscher
4                   Rottweiler  0.243682    True                     Doberman
5             English_springer  0.263788    True   Greater_Swiss_Mountain_dog
6                   mud_turtle  0.045885   False                     terrapin
7              Tibetan_mastiff  0.058279    True                     fur_coat
8              shopping_basket  0.014594   False             golden_retriever
9                     komondor  0.192305    True  soft-coated_wheaten_terrier
10             Tibetan_mastiff  0.093718    True           Labrador_retriever
11           Yorkshire_terrier  0.174201    True                     Pekinese
12            English_foxhound  0.175382    True                 Ibizan_hound
13                bull_mastiff  0.404722    True               French_bulldog
14             German_shepherd  0.078260    True                     malinois
15                    Shih-Tzu  0.166192    True               Dandie_Dinmont
16                Newfoundland  0.149842    True                       borzoi
17                        cock  0.033919   False                    partridge
18                        desk  0.085547   False                     bookcase
19                 toy_terrier  0.111884    True                      basenji
20                  toy_poodle  0.063064    True             miniature_poodle
21                       otter  0.015250   False               great_grey_owl
22                Newfoundland  0.278407    True                  groenendael
```

```
23    Chesapeake_Bay_retriever  0.054787    True           Labrador_retriever
24              Siberian_husky  0.147655    True                   Eskimo_dog
25                       skunk  0.002402   False                      hamster
26                Afghan_hound  0.259551    True                       briard
27                 toy_terrier  0.009397    True                     papillon
28                  bloodhound  0.244220    True        flat-coated_retriever
29                  barracouta  0.271485   False                          gar
...                        ...       ...     ...                          ...
2045                 Chihuahua  0.000361    True                  Boston_bull
2046                sports_car  0.139952   False                    car_wheel
2047            Great_Pyrenees  0.186136    True               Dandie_Dinmont
2048               toy_terrier  0.143528    True                   can_opener
2049                Eskimo_dog  0.035029    True    Staffordshire_bullterrier
2050                  sea_lion  0.275645   False                   Weimaraner
2051       Rhodesian_ridgeback  0.054950    True                       beagle
2052                 tow_truck  0.029175   False                shopping_cart
2053                   redbone  0.087582    True                   Weimaraner
2054                       pug  0.000932    True                 bull_mastiff
2055       Rhodesian_ridgeback  0.054950    True                       beagle
2056                Eskimo_dog  0.166511    True                     malamute
2057        Labrador_retriever  0.184172    True               English_setter
2058           Tibetan_mastiff  0.120184    True           Labrador_retriever
2059                    borzoi  0.194742    True                       Saluki
2060        Labrador_retriever  0.013834    True                      redbone
2061                     boxer  0.002129    True    Staffordshire_bullterrier
2062                  Cardigan  0.027356    True                      basenji
2063        Labrador_retriever  0.151317    True                       muzzle
2064                Pomeranian  0.013884    True                         chow
2065                  Cardigan  0.451038    True                    Chihuahua
2066              Irish_setter  0.193054    True     Chesapeake_Bay_retriever
2067                Eskimo_dog  0.178406    True                     Pembroke
2068             Border_collie  0.199287    True                    ice_lolly
2069             Irish_terrier  0.116317    True              Indian_elephant
2070          English_springer  0.225770    True  German_short-haired_pointer
2071        Labrador_retriever  0.168086    True                      spatula
2072                  malamute  0.078253    True                       kelpie
2073                  Pekinese  0.090647    True                     papillon
2074                     bagel  0.085851   False                       banana

       p3_conf  p3_dog
0     0.061428    True
1     0.072010    True
2     0.116197    True
3     0.222752    True
4     0.154629    True
5     0.016199    True
6     0.017885   False
7     0.054449   False
```

```
8      0.007959    True
9      0.082086    True
10     0.072427    True
11     0.109454    True
12     0.097471    True
13     0.048960    True
14     0.075628    True
15     0.089688    True
16     0.133649    True
17     0.000052   False
18     0.079480   False
19     0.111152    True
20     0.025581    True
21     0.013207   False
22     0.102643    True
23     0.014241    True
24     0.093412    True
25     0.000461   False
26     0.206803    True
27     0.004577    True
28     0.173810    True
29     0.189945   False
...         ...     ...
2045   0.000076    True
2046   0.044173   False
2047   0.086346    True
2048   0.032253   False
2049   0.029705    True
2050   0.134203    True
2051   0.038915    True
2052   0.026321   False
2053   0.026236    True
2054   0.000903    True
2055   0.038915    True
2056   0.111411    True
2057   0.073482    True
2058   0.105506    True
2059   0.027351    True
2060   0.007958    True
2061   0.001498    True
2062   0.004633    True
2063   0.082981   False
2064   0.008167    True
2065   0.029248    True
2066   0.118184    True
2067   0.076507    True
2068   0.193548   False
2069   0.076902   False
```

```
2070  0.175219    True
2071  0.040836   False
2072  0.031379    True
2073  0.068957    True
2074  0.076110   False

[2075 rows x 12 columns]
```

```
In [74]: tweet_archive
```

```
Out[74]:             tweet_id  favorite_count  retweet_count
      0     892420643555336193           39373           8796
      1     892177421306343426           33696           6451
      2     891815181378084864           25391           4276
      3     891689557279858688           42741           8885
      4     891327558926688256           40902           9670
      5     891087950875897856           20504           3218
      6     890971913173991426           12029           2132
      7     890729181411237888           66534          19478
      8     890609185150312448           28134           4371
      9     890240255349198849           32377           7638
      10    890006608113172480           31036           7537
      11    889880896479866881           28143           5095
      12    889665388333682689           38624           8465
      13    889638837579907072           27546           4675
      14    889531135344209921           15303           2296
      15    889278841981685760           25646           5601
      16    888917238123831296           29482           4651
      17    888804989199671297           25950           4505
      18    888554962724278272           20219           3702
      19          id not found           20219           3702
      20    888078434458587136           22085           3613
      21    887705289381826560           30614           5554
      22    887517139158093824           46825          11997
      23    887473957103951883           69812          18696
      24    887343217045368832           34123          10667
      25    887101392804085760           30972           6121
      26    886983233522544640           35680           7997
      27    886736880519319552           12249           3395
      28    886680336477933568           22736           4582
      29    886366144734445568           21437           3277
      ...                  ...             ...            ...
      2326  666411507551481857             457            337
      2327  666407126856765440             113             42
      2328  666396247373291520             171             90
      2329  666373753744588802             194             95
      2330  666362758909284353             799            588
      2331  666353288456101888             228             74
```

```
        2332  666345417576210432          307           145
        2333  666337882303524864          203            95
        2334  666293911632134144          515           365
        2335  666287406224695296          152            70
        2336  666273097616637952          182            80
        2337  666268910803644416          108            36
        2338  666104133288665088        14652          6808
        2339  666102155909144576           81            14
        2340  666099513787052032          161            72
        2341  666094000022159362          167            77
        2342  666082916733198337          121            46
        2343  666073100786774016          333           172
        2344  666071193221509120          154            65
        2345  666063827256086533          492           226
        2346  666058600524156928          117            60
        2347  666057090499244032          304           145
        2348  666055525042405380          448           260
        2349  666051853826850816         1247           874
        2350  666050758794694657          136            59
        2351  666049248165822465          111            40
        2352  666044226329800704          307           144
        2353  666033412701032449          128            46
        2354  666029285002620928          132            47
        2355  666020888022790149         2529           527

        [2356 rows x 3 columns]
```

```
In [75]: twitter_archive.info()
```

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

```
puppo                         2356 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

```
In [76]: twitter_archive['rating_numerator'].value_counts()
```

```
Out[76]: 12      558
         11      464
         10      461
         13      351
         9       158
         8       102
         7        55
         14       54
         5        37
         6        32
         3        19
         4        17
         1         9
         2         9
         420       2
         0         2
         15        2
         75        2
         80        1
         20        1
         24        1
         26        1
         44        1
         50        1
         60        1
         165       1
         84        1
         88        1
         144       1
         182       1
         143       1
         666       1
         960       1
         1776      1
         17        1
         27        1
         45        1
         99        1
         121       1
         204       1
         Name: rating_numerator, dtype: int64
```

```
In [77]: twitter_archive.rating_denominator.value_counts()

Out[77]: 10     2333
         11        3
         50        3
         80        2
         20        2
         2         1
         16        1
         40        1
         70        1
         15        1
         90        1
         110       1
         120       1
         130       1
         150       1
         170       1
         7         1
         0         1
         Name: rating_denominator, dtype: int64
```

```
In [78]: image_predictions.info()

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

```
In [79]: image_predictions[image_predictions.jpg_url.duplicated()]

Out[79]:                  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/CcGO7BYW0AErrC9.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/CsrjryzWgAAZYO0.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  0.272205    True
1333        2               cash_machine  0.802333   False
1345        1              Irish_terrier  0.254856    True
1349        1         Labrador_retriever  0.967397    True
1364        1                 brown_bear  0.713293   False
1368        1           golden_retriever  0.586937    True
1387        1                 toy_poodle  0.420463    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  0.997310    True
1487        1           golden_retriever  0.383223    True
1495        1                   Cardigan  0.611525    True
1510        1           golden_retriever  0.993830    True
1522        1           golden_retriever  0.735163    True
1538        1                        tub  0.479477   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  0.533180    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  0.090341    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  0.740220    True
1652        1         Labrador_retriever  0.861651    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  0.416769    True
1746        1                    Samoyed  0.585441    True
1755        2                   gas_pump  0.676439   False
1789        1           golden_retriever  0.617389    True
1803        1         miniature_pinscher  0.796313    True
1804        1               Afghan_hound  0.274637    True
1858        1                   ice_bear  0.336200   False
1864        1                  Chihuahua  0.505370    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  0.251530    True
1333                      schipperke  0.045519    True
1345                          briard  0.227716    True
1349                golden_retriever  0.016641    True
1364                 Indian_elephant  0.172844   False
1368              Labrador_retriever  0.398260    True
1387                miniature_poodle  0.132640    True
1407              Labrador_retriever  0.052956    True
1417                      Pomeranian  0.012763    True
1427                         whippet  0.151047    True
1446              Labrador_retriever  0.042319    True
1453                       dalmatian  0.166246    True
1456              Labrador_retriever  0.011584    True
1463                         doormat  0.152445   False
1476               Brabancon_griffon  0.001186    True
1487                  cocker_spaniel  0.165930    True
1495                        Pembroke  0.368566    True
1510                  cocker_spaniel  0.003143    True
1522                  Sussex_spaniel  0.064897    True
1538                         bathtub  0.325106   False
```

```
1541                            jean  0.069362   False
1564              Labrador_retriever  0.219721    True
1569                golden_retriever  0.272734    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  0.227150   False
...                              ...       ...     ...
1619                golden_retriever  0.162850    True
1624                   Irish_terrier  0.197314    True
1627                      binoculars  0.083499   False
1634                golden_retriever  0.339251    True
1636               Shetland_sheepdog  0.130611    True
1640                golden_retriever  0.171393    True
1645                  Dandie_Dinmont  0.061604    True
1652                golden_retriever  0.044462    True
1683                      Great_Dane  0.190503    True
1693                   Border_collie  0.052724    True
1699                golden_retriever  0.245773    True
1703                 Tibetan_terrier  0.295690    True
1712        Chesapeake_Bay_retriever  0.137187    True
1717                          palace  0.020089   False
1718                      toy_poodle  0.271929    True
1727                          kuvasz  0.025119    True
1736                        Cardigan  0.003044    True
1742                        malamute  0.252706    True
1746                      Pomeranian  0.193654    True
1755                       harvester  0.049995   False
1789              Labrador_retriever  0.337053    True
1803                       Chihuahua  0.155413    True
1804                          borzoi  0.142204    True
1858                         Samoyed  0.201358    True
1864                      Pomeranian  0.120358    True
1903  American_Staffordshire_terrier  0.149950    True
1944                     house_finch  0.209852   False
1992                  French_bulldog  0.099984    True
2041                       seat_belt  0.057091   False
2055             Rhodesian_ridgeback  0.054950    True

                                  p3   p3_conf  p3_dog
1297                Brittany_spaniel  0.162084    True
1315                      bath_towel  0.116806   False
1333                 German_shepherd  0.023353    True
1345     soft-coated_wheaten_terrier  0.223263    True
1349                        ice_bear  0.014858   False
1364                   water_buffalo  0.038902   False
```

```
1368                          kuvasz  0.005410    True
1387        Chesapeake_Bay_retriever  0.121523    True
1407                          kelpie  0.035711    True
1417                      Eskimo_dog  0.001853    True
1427  American_Staffordshire_terrier  0.096435    True
1446                         doormat  0.003956   False
1453                     toy_terrier  0.117502    True
1456                    refrigerator  0.007499   False
1463                        sea_lion  0.026364   False
1476                  French_bulldog  0.000428    True
1487        Chesapeake_Bay_retriever  0.118199    True
1495                       Chihuahua  0.003330    True
1510                  Great_Pyrenees  0.000917    True
1522              Labrador_retriever  0.047704    True
1538                golden_retriever  0.078530    True
1541                        keeshond  0.050528    True
1564                          kuvasz  0.015965    True
1569                   jigsaw_puzzle  0.041476   False
1571              Norwegian_elkhound  0.151024    True
1579                       seat_belt  0.087355   False
1588                   Border_collie  0.121626    True
1589                golden_retriever  0.082549    True
1590          Welsh_springer_spaniel  0.137186    True
1591                      Pomeranian  0.056057    True
...                              ...       ...     ...
1619                      otterhound  0.110921    True
1624                        Airedale  0.082515    True
1627                    Irish_setter  0.077456    True
1634                       seat_belt  0.046108   False
1636                        Pembroke  0.100842    True
1640              Labrador_retriever  0.110592    True
1645                  English_setter  0.041331    True
1652       Staffordshire_bullterrier  0.016497    True
1683          curly-coated_retriever  0.146427    True
1693                  cocker_spaniel  0.048572    True
1699                            chow  0.039012    True
1703                      otterhound  0.036251    True
1712                        malamute  0.071436    True
1717                        umbrella  0.007850   False
1718                 Tibetan_terrier  0.094759    True
1727              Labrador_retriever  0.022977    True
1736                English_springer  0.001050    True
1742                          kuvasz  0.157028    True
1746                      Arctic_fox  0.071648   False
1755                           swing  0.044660   False
1789                     tennis_ball  0.008554   False
1803       Staffordshire_bullterrier  0.030943    True
1804                         doormat  0.109677   False
```

```
1858                      Eskimo_dog  0.186789    True
1864                     toy_terrier  0.077008    True
1903        Chesapeake_Bay_retriever  0.047523    True
1944                   cocker_spaniel  0.046403    True
1992                         printer  0.077130   False
2041                             pug  0.011933    True
2055                          beagle  0.038915    True

[66 rows x 12 columns]
```

```
In [80]: tweet_archive.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 3 columns):
tweet_id          2356 non-null object
favorite_count    2356 non-null int64
retweet_count     2356 non-null int64
dtypes: int64(2), object(1)
memory usage: 55.3+ KB
```

```
In [81]: tweet_archive['tweet_id'].value_counts()
```

```
Out[81]: id not found          6
         667495797102141441    1
         710296729921429505    1
         727644517743104000    1
         739932936087216128    1
         753298634498793472    1
         759793422261743616    1
         743835915802583040    1
         870656317836468226    1
         744223424764059648    1
         676897532954456065    1
         756288534030475264    1
         768909767477751808    1
         746369468511756288    1
         757741869644341248    1
         689623661272240129    1
         747242308580548608    1
         838201503651401729    1
         736392552031657984    1
         783334639985389568    1
         742465774154047488    1
         726224900189511680    1
         726887082820554753    1
         717421804990701568    1
         741099773336379392    1
```

```
760539183865880579    1
719367763014393856    1
718939241951195136    1
751830394383790080    1
722613351520608256    1
                     ..
694925794720792577    1
682389078323662849    1
710588934686908417    1
708810915978854401    1
832769181346996225    1
667062181243039745    1
667724302356258817    1
710844581445812225    1
676440007570247681    1
827228250799742977    1
666396247373291520    1
692142790915014657    1
673612854080196609    1
855862651834028034    1
889278841981685760    1
886680336477933568    1
832757312314028032    1
802572683846291456    1
679844490799091713    1
805932879469572096    1
889665388333682689    1
857746408056729600    1
691459709405118465    1
886736880519319552    1
828650029636317184    1
690400367696297985    1
887705289381826560    1
678675843183484930    1
666020888022790149    1
891815181378084864    1
Name: tweet_id, Length: 2351, dtype: int64
```

0.2 Assess

0.2.1 Quality

`twitter_archive` **table**

- Erroneous datatypes(tweet id, timestamp, retweeted_status_timestamp, rating_numerator).
- Irregular values in columns retweeted_status_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id.
- Rating is not provided for *810984652412424192, 832088576586297345*.

- Rating denominator is sometimes wrong because that particular instance has multiple ratings.
- Rating numerator is sometimes wrong because that particular instance has multiple ratings(666287406224695296, 695064344191721472, 674646392044941312, 691483041324204033).
- Some rating numerators are wrong because the rating is given in float.(883482846933004288, 681340665377193984,786709082849828864, 680494726643068929, 778027034220126208).
- Multiple values in same rows of columns(doggo, pupper, puppo, floofer).

`image_predictions` **table**

- Duplicated images.
- Erroneous datatypes(tweet id)

`tweet_archive` **table**

- Some tweets doesn't have a tweet id.
- Erroneous datatypes(tweet id)

0.2.2 Tidiness

- Retweets with ratings must be removed.
- One variable in four columns in twitter_archive table (dog_stage, doggo, floofer, puppo and pupper).
- tweet_archive table must be a part of twitter_archive table.
- Drop the columns which aren't necessary for analysis in the cleaned table.

0.3 Clean

```
In [82]: twitter_archive_clean = twitter_archive.copy()
         image_predictions_clean = image_predictions.copy()
         tweet_archive_clean = tweet_archive.copy()
```

0.3.1 Tidiness

1. Delete retweeted data in `twitter archive` table.

i. Irregular values in columns retweeted_status_id, retweeted_status_user_id.

Define Remove the data which has been retweeted. Convert irregular values by changing NaN to zero(0) and changing the datatype into int.

Code

```
In [83]: twitter_archive_clean.retweeted_status_id = twitter_archive_clean.retweeted_status_id.f
         twitter_archive_clean.retweeted_status_user_id = twitter_archive_clean.retweeted_status
         twitter_archive_clean = twitter_archive_clean[~(twitter_archive_clean.retweeted_status_
```

Test

```
In [84]: twitter_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2175 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2175 non-null object
source                        2175 non-null object
text                          2175 non-null object
retweeted_status_id           2175 non-null int64
retweeted_status_user_id      2175 non-null int64
retweeted_status_timestamp    0 non-null object
expanded_urls                 2117 non-null object
rating_numerator              2175 non-null int64
rating_denominator            2175 non-null int64
name                          2175 non-null object
doggo                         2175 non-null object
floofer                       2175 non-null object
pupper                        2175 non-null object
puppo                         2175 non-null object
dtypes: float64(2), int64(5), object(10)
memory usage: 305.9+ KB
```

Delete retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns in `twitter archive` table.

Define Drop the columns(retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp).

Code

```
In [85]: twitter_archive_clean = twitter_archive_clean.drop('retweeted_status_id', axis = 1)
         twitter_archive_clean = twitter_archive_clean.drop('retweeted_status_user_id', axis = 1
         twitter_archive_clean = twitter_archive_clean.drop('retweeted_status_timestamp', axis =
```

Test

```
In [86]: twitter_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 14 columns):
tweet_id                 2175 non-null int64
```

```
in_reply_to_status_id    78 non-null float64
in_reply_to_user_id      78 non-null float64
timestamp                2175 non-null object
source                   2175 non-null object
text                     2175 non-null object
expanded_urls            2117 non-null object
rating_numerator         2175 non-null int64
rating_denominator       2175 non-null int64
name                     2175 non-null object
doggo                    2175 non-null object
floofer                  2175 non-null object
pupper                   2175 non-null object
puppo                    2175 non-null object
dtypes: float64(2), int64(3), object(9)
memory usage: 254.9+ KB
```

2. One variable in four columns in `twitter_archive` table (dog_stage, doggo, floofer, puppo and pupper).

ii. Multiple values in same rows of columns(doggo, pupper, puppo, floofer).

Define Combine the columns pupper, doggo, puppo and floofer to dog_stage column. Then drop the pupper, doggo, puppo and floofer columns.

Code

```
In [87]: twitter_archive_clean.doggo = twitter_archive_clean.doggo.replace('None','')
         twitter_archive_clean.floofer = twitter_archive_clean.floofer.replace('None','')
         twitter_archive_clean.pupper = twitter_archive_clean.pupper.replace('None','')
         twitter_archive_clean.puppo = twitter_archive_clean.puppo.replace('None','')
         twitter_archive_clean['dog_stage'] = twitter_archive_clean[['doggo','floofer','pupper',
         twitter_archive_clean.dog_stage = twitter_archive_clean.dog_stage.str.strip()
         twitter_archive_clean.dog_stage = twitter_archive_clean.dog_stage.replace('doggo  puppe
         twitter_archive_clean.dog_stage = twitter_archive_clean.dog_stage.replace('doggo floofe
         twitter_archive_clean.dog_stage = twitter_archive_clean.dog_stage.replace('doggo   pupp
         twitter_archive_clean.dog_stage = twitter_archive_clean.dog_stage.replace('','None')
```

```
In [88]: twitter_archive_clean = twitter_archive_clean.drop('doggo', axis = 1)
         twitter_archive_clean = twitter_archive_clean.drop('floofer', axis = 1)
         twitter_archive_clean = twitter_archive_clean.drop('pupper', axis = 1)
         twitter_archive_clean = twitter_archive_clean.drop('puppo', axis = 1)
```

Test

```
In [89]: twitter_archive_clean.dog_stage.value_counts()
```

```
Out[89]: None              1831
         pupper             224
         doggo               75
         puppo               24
         doggo, pupper       10
         floofer              9
         doggo, puppo         1
         doggo, floofer       1
         Name: dog_stage, dtype: int64
```

```
In [90]: twitter_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 11 columns):
tweet_id                 2175 non-null int64
in_reply_to_status_id    78 non-null float64
in_reply_to_user_id      78 non-null float64
timestamp                2175 non-null object
source                   2175 non-null object
text                     2175 non-null object
expanded_urls            2117 non-null object
rating_numerator         2175 non-null int64
rating_denominator       2175 non-null int64
name                     2175 non-null object
dog_stage                2175 non-null object
dtypes: float64(2), int64(3), object(6)
memory usage: 203.9+ KB
```

0.3.2 Quality

iii. `twitter_archive`: Erroneous datatypes

iii. `image_predictions`: Erroneous datatypes

iii. `tweet_archive`: Erroneous datatypes

i. Irregular values in columns in_reply_to_status_id, in_reply_to_user_id.

Define Convert timestamp to datetime data type. Convert in_reply_to_status_id and in_reply_to_status_id to int data type. Convert NaN to zero(0) so as to convert the data type into int. Convert tweet_id to string data type.

Code

```
In [91]: twitter_archive_clean.timestamp = pd.to_datetime(twitter_archive_clean.timestamp)
         twitter_archive_clean.in_reply_to_user_id = twitter_archive_clean.in_reply_to_user_id.f
         twitter_archive_clean.in_reply_to_status_id = twitter_archive_clean.in_reply_to_status_
         twitter_archive_clean.tweet_id = twitter_archive_clean.tweet_id.astype(str)
         tweet_archive_clean.tweet_id = tweet_archive_clean.tweet_id.astype(str)
         image_predictions_clean.tweet_id = image_predictions_clean.tweet_id.astype(str)
```

Test

```
In [92]: twitter_archive_clean.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 11 columns):
tweet_id                 2175 non-null object
in_reply_to_status_id    2175 non-null int64
in_reply_to_user_id      2175 non-null int64
timestamp                2175 non-null datetime64[ns]
source                   2175 non-null object
text                     2175 non-null object
expanded_urls            2117 non-null object
rating_numerator         2175 non-null int64
rating_denominator       2175 non-null int64
name                     2175 non-null object
dog_stage                2175 non-null object
dtypes: datetime64[ns](1), int64(4), object(6)
memory usage: 203.9+ KB
```

```
In [93]: tweet_archive_clean.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 3 columns):
tweet_id          2356 non-null object
favorite_count    2356 non-null int64
retweet_count     2356 non-null int64
dtypes: int64(2), object(1)
memory usage: 55.3+ KB
```

```
In [94]: image_predictions_clean.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 int64
```

```
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(1), object(5)
memory usage: 152.1+ KB
```

iv. `twitter_archive`: Rating not provided in some tweets.

Define Remove tweets without ratings(810984652412424192, 832088576586297345).

Code

```
In [95]: twitter_archive_clean = twitter_archive_clean[twitter_archive_clean.tweet_id != 8109846
         twitter_archive_clean = twitter_archive_clean[twitter_archive_clean.tweet_id != 8320885
```

Test

```
In [96]: twitter_archive_clean[twitter_archive_clean.tweet_id == 832088576586297345]

Out[96]: Empty DataFrame
         Columns: [tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, source, text
         Index: []

In [97]: twitter_archive_clean[twitter_archive_clean.tweet_id == 810984652412424192]

Out[97]: Empty DataFrame
         Columns: [tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, source, text
         Index: []
```

v. `twitter_archive`: Rating numerator is sometimes wrong because that particular instance has multiple ratings

vi. Some rating numerators are wrong because the rating is given in float

vii. Rating denominator is sometimes wrong because that particular instance has multiple ratings

iii. Erroneous datatypes (Rating numerator)

Define Extract the *rating numerator* and *rating denominator* variables from the *text* column using regular expressions and pandas' `str.extract` method. Drop any intermediate columns.

Code

```
In [98]: twitter_archive_clean['text_rev'] = twitter_archive_clean.text.apply(lambda x : ', '.jo
         twitter_archive_clean['rating_numerator'] = twitter_archive_clean.text_rev.str.extract(
         twitter_archive_clean['rating_denominator'] = twitter_archive_clean.text.str.extract('(
```

```
In [99]: twitter_archive_clean = twitter_archive_clean.drop('text_rev', axis = 1)
```

```
In [100]: twitter_archive_clean.rating_numerator = twitter_archive_clean.rating_numerator.astype
          twitter_archive_clean.rating_denominator = twitter_archive_clean.rating_denominator.as
```

Test

```
In [101]: twitter_archive_clean[twitter_archive_clean.tweet_id == 786709082849828864]
```

```
Out[101]: Empty DataFrame
          Columns: [tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, source, tex
          Index: []
```

```
In [102]: twitter_archive_clean[twitter_archive_clean.tweet_id == 691483041324204033]
```

```
Out[102]: Empty DataFrame
          Columns: [tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, source, tex
          Index: []
```

```
In [103]: twitter_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 11 columns):
tweet_id                 2175 non-null object
in_reply_to_status_id    2175 non-null int64
in_reply_to_user_id      2175 non-null int64
timestamp                2175 non-null datetime64[ns]
source                   2175 non-null object
text                     2175 non-null object
expanded_urls            2117 non-null object
rating_numerator         2175 non-null float64
rating_denominator       2175 non-null int64
name                     2175 non-null object
dog_stage                2175 non-null object
dtypes: datetime64[ns](1), float64(1), int64(3), object(6)
memory usage: 203.9+ KB
```

viii. image_predictions: Duplicated images

Define Remove the duplicated images.

Code

```
In [104]: image_predictions_clean = image_predictions_clean[~(image_predictions_clean.jpg_url.du
```

Test

```
In [105]: image_predictions_clean[image_predictions_clean.jpg_url.duplicated()]

Out[105]: Empty DataFrame
          Columns: [tweet_id, jpg_url, img_num, p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3
          Index: []
```

```
In [106]: image_predictions_clean.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 2009 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2009 non-null object
jpg_url     2009 non-null object
img_num     2009 non-null int64
p1          2009 non-null object
p1_conf     2009 non-null float64
p1_dog      2009 non-null bool
p2          2009 non-null object
p2_conf     2009 non-null float64
p2_dog      2009 non-null bool
p3          2009 non-null object
p3_conf     2009 non-null float64
p3_dog      2009 non-null bool
dtypes: bool(3), float64(3), int64(1), object(5)
memory usage: 162.8+ KB
```

ix. tweet_archive: tweet ID missing

Define Drop rows without tweet_id.

Code

```
In [107]: tweet_archive_clean = tweet_archive_clean[~(tweet_archive_clean.tweet_id.duplicated())
```

Test

```
In [108]: tweet_archive_clean['tweet_id'].value_counts()

Out[108]: 690607260360429569    1
          852912242202992640    1
          667119796878725120    1
          685321586178670592    1
```

```
888917238123831296      1
666102155909144576      1
866686824827068416      1
730924654643314689      1
845397057150107648      1
793845145112371200      1
713761197720473600      1
769335591808995329      1
689659372465688576      1
883482846933004288      1
686606069955735556      1
689623661272240129      1
719551379208073216      1
692158366030913536      1
680221482581123072      1
870374049280663552      1
771004394259247104      1
817171292965273600      1
701570477911896070      1
816336735214911488      1
669214165781868544      1
704847917308362754      1
689599056876867584      1
749417653287129088      1
752173152931807232      1
733109485275860992      1
                       ..
711306686208872448      1
814578408554463233      1
768909767477751808      1
680609293079592961      1
743545585370791937      1
691459709405118465      1
769695466921623552      1
753398408988139520      1
723673163800948736      1
687807801670897665      1
747816857231626240      1
780092040432480260      1
674024893172875264      1
669367896104181761      1
666057090499244032      1
889638837579907072      1
706169069255446529      1
666430724426358785      1
755955933503782912      1
671147085991960577      1
826476773533745153      1
```

```
679405845277462528    1
852189679701164033    1
775842724423557120    1
684188786104872960    1
811647686436880384    1
666776908487630848    1
707738799544082433    1
712309440758808576    1
718971898235854848    1
Name: tweet_id, Length: 2351, dtype: int64
```

0.3.3 Tidiness

3. tweet_archive table must be a part of twitter_archive table.

Define Merge tweet_archive_clean table with twitter_archive_clean table.

Code

```
In [109]: twitter_archive_clean = pd.merge(twitter_archive_clean, tweet_archive_clean,
                                           on = ['tweet_id'], how = 'left')
```

Test

```
In [110]: twitter_archive_clean.head()
```

```
Out[110]:             tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
          0  892420643555336193                      0                    0
          1  892177421306343426                      0                    0
          2  891815181378084864                      0                    0
          3  891689557279858688                      0                    0
          4  891327558926688256                      0                    0

                     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...

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

                                                 expanded_urls  rating_numerator  \
0  https://twitter.com/dog_rates/status/892420643...             13.0
1  https://twitter.com/dog_rates/status/892177421...             13.0
2  https://twitter.com/dog_rates/status/891815181...             12.0
3  https://twitter.com/dog_rates/status/891689557...             13.0
4  https://twitter.com/dog_rates/status/891327558...             12.0

   rating_denominator      name dog_stage  favorite_count  retweet_count
0                  10   Phineas      None           39373           8796
1                  10     Tilly      None           33696           6451
2                  10    Archie      None           25391           4276
3                  10     Darla      None           42741           8885
4                  10  Franklin      None           40902           9670
```

4. Drop columns which aren't necessary for analysis in `image_predictions` table.

Define Drop img_num column in image_predictions table.

Code

```
In [111]: image_predictions_clean = image_predictions_clean.drop('img_num', axis = 1)
```

Test

```
In [112]: image_predictions_clean.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 2009 entries, 0 to 2074
Data columns (total 11 columns):
tweet_id    2009 non-null object
jpg_url     2009 non-null object
p1          2009 non-null object
p1_conf     2009 non-null float64
p1_dog      2009 non-null bool
p2          2009 non-null object
p2_conf     2009 non-null float64
p2_dog      2009 non-null bool
p3          2009 non-null object
p3_conf     2009 non-null float64
p3_dog      2009 non-null bool
dtypes: bool(3), float64(3), object(5)
memory usage: 147.1+ KB
```

```
In [113]: image_predictions_clean.head()

Out[113]:             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

                        p1   p1_conf  p1_dog                  p2   p2_conf  \
0  Welsh_springer_spaniel  0.465074    True              collie  0.156665
1                 redbone  0.506826    True  miniature_pinscher  0.074192
2         German_shepherd  0.596461    True            malinois  0.138584
3     Rhodesian_ridgeback  0.408143    True             redbone  0.360687
4      miniature_pinscher  0.560311    True          Rottweiler  0.243682

  p2_dog                   p3   p3_conf  p3_dog
0   True     Shetland_sheepdog  0.061428    True
1   True  Rhodesian_ridgeback  0.072010    True
2   True           bloodhound  0.116197    True
3   True   miniature_pinscher  0.222752    True
4   True             Doberman  0.154629    True
```

5. Drop columns which aren't necessary for analysis like in_reply_to_status_id, in_reply_to_user_id, expanded_urls, timestamp and source in `twitter archive` table.

Define Drop columns in_reply_to_status_id, in_reply_to_user_id, expanded_urls, timestamp and source in twitter archive table.

Code

```
In [114]: twitter_archive_clean = twitter_archive_clean.drop('in_reply_to_status_id', axis = 1)
          twitter_archive_clean = twitter_archive_clean.drop('in_reply_to_user_id', axis = 1)
          twitter_archive_clean = twitter_archive_clean.drop('expanded_urls', axis = 1)
          twitter_archive_clean = twitter_archive_clean.drop('timestamp', axis = 1)
          twitter_archive_clean = twitter_archive_clean.drop('source', axis = 1)
```

Test

```
In [115]: twitter_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2174
Data columns (total 8 columns):
tweet_id              2175 non-null object
text                  2175 non-null object
rating_numerator      2175 non-null float64
rating_denominator    2175 non-null int64
name                  2175 non-null object
dog_stage             2175 non-null object
favorite_count        2175 non-null int64
retweet_count         2175 non-null int64
dtypes: float64(1), int64(3), object(4)
```

```
memory usage: 152.9+ KB
```

```
In [116]: twitter_archive_clean.head()
```

```
Out[116]:             tweet_id                                              text  \
          0  892420643555336193  This is Phineas. He's a mystical boy. Only eve...
          1  892177421306343426  This is Tilly. She's just checking pup on you...
          2  891815181378084864  This is Archie. He is a rare Norwegian Pouncin...
          3  891689557279858688  This is Darla. She commenced a snooze mid meal...
          4  891327558926688256  This is Franklin. He would like you to stop ca...

             rating_numerator  rating_denominator      name dog_stage  favorite_count  \
          0              13.0                  10   Phineas      None           39373
          1              13.0                  10     Tilly      None           33696
          2              12.0                  10    Archie      None           25391
          3              13.0                  10     Darla      None           42741
          4              12.0                  10  Franklin      None           40902

             retweet_count
          0           8796
          1           6451
          2           4276
          3           8885
          4           9670
```

6. Merge columns from `image predictions` **which are necessary for analysis to** `twitter archive` **table.**

Define Merge columns p2, p2_conf and p2_dog from image predictions table to twitter archive table. p2_dog has more True values than p1_dog and p3_dog

Code

```
In [117]: twitter_archive_clean = pd.merge(twitter_archive_clean,
                                           image_predictions_clean[['tweet_id', 'p2', 'p2_conf',
                                           on='tweet_id', how='left')
          twitter_archive_clean = twitter_archive_clean.rename(columns = {'p2' : 'p', 'p2_conf'
```

Test

```
In [118]: twitter_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2174
Data columns (total 11 columns):
tweet_id              2175 non-null object
text                  2175 non-null object
```

```
rating_numerator      2175 non-null float64
rating_denominator    2175 non-null int64
name                  2175 non-null object
dog_stage             2175 non-null object
favorite_count        2175 non-null int64
retweet_count         2175 non-null int64
p                     1994 non-null object
p_conf                1994 non-null float64
p_dog                 1994 non-null object
dtypes: float64(2), int64(3), object(6)
memory usage: 203.9+ KB
```

```
In [119]: twitter_archive_clean.head()
```

```
Out[119]:              tweet_id                                              text  \
          0  892420643555336193  This is Phineas. He's a mystical boy. Only eve...
          1  892177421306343426  This is Tilly. She's just checking pup on you...
          2  891815181378084864  This is Archie. He is a rare Norwegian Pouncin...
          3  891689557279858688  This is Darla. She commenced a snooze mid meal...
          4  891327558926688256  This is Franklin. He would like you to stop ca...

             rating_numerator  rating_denominator      name dog_stage  favorite_count  \
          0              13.0                  10   Phineas      None           39373
          1              13.0                  10     Tilly      None           33696
          2              12.0                  10    Archie      None           25391
          3              13.0                  10     Darla      None           42741
          4              12.0                  10  Franklin      None           40902

             retweet_count                   p    p_conf  p_dog
          0           8796               bagel  0.085851  False
          1           6451            Pekinese  0.090647   True
          2           4276            malamute  0.078253   True
          3           8885  Labrador_retriever  0.168086   True
          4           9670    English_springer  0.225770   True
```

0.4 Storing

```
In [120]: twitter_archive_clean.to_csv('Source_Files/twitter_archive_master.csv', index=False)
          image_predictions_clean.to_csv('Source_Files/image_predictions_master.csv', index=Fals
```

```
In [121]: twitter_archive_master = pd.read_csv('Source_Files/twitter_archive_master.csv')
          image_predictions_master = pd.read_csv('Source_Files/image_predictions_master.csv')
```

```
In [122]: twitter_archive_master.head()
```

```
Out[122]:              tweet_id                                              text  \
          0  892420643555336193  This is Phineas. He's a mystical boy. Only eve...
          1  892177421306343426  This is Tilly. She's just checking pup on you...
          2  891815181378084864  This is Archie. He is a rare Norwegian Pouncin...
          3  891689557279858688  This is Darla. She commenced a snooze mid meal...
          4  891327558926688256  This is Franklin. He would like you to stop ca...

             rating_numerator  rating_denominator      name dog_stage  favorite_count  \
          0              13.0                  10   Phineas      None           39373
          1              13.0                  10     Tilly      None           33696
          2              12.0                  10    Archie      None           25391
          3              13.0                  10     Darla      None           42741
          4              12.0                  10  Franklin      None           40902

             retweet_count                   p    p_conf  p_dog
          0           8796               bagel  0.085851  False
          1           6451            Pekinese  0.090647   True
          2           4276            malamute  0.078253   True
          3           8885  Labrador_retriever  0.168086   True
          4           9670    English_springer  0.225770   True
```

```
In [123]: image_predictions_master.head()
```

```
Out[123]:              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

                                 p1   p1_conf  p1_dog                  p2   p2_conf  \
          0  Welsh_springer_spaniel  0.465074    True              collie  0.156665
          1                 redbone  0.506826    True  miniature_pinscher  0.074192
          2         German_shepherd  0.596461    True            malinois  0.138584
          3     Rhodesian_ridgeback  0.408143    True             redbone  0.360687
          4      miniature_pinscher  0.560311    True          Rottweiler  0.243682

             p2_dog                   p3   p3_conf  p3_dog
          0    True    Shetland_sheepdog  0.061428    True
          1    True  Rhodesian_ridgeback  0.072010    True
          2    True           bloodhound  0.116197    True
          3    True   miniature_pinscher  0.222752    True
          4    True             Doberman  0.154629    True
```

```
In [141]: twitter_archive_master.tweet_id = twitter_archive_master.tweet_id.astype(str)
          image_predictions_master.tweet_id = image_predictions_master.tweet_id.astype(str)
```

```
In [142]: twitter_archive_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2175 entries, 0 to 2174
Data columns (total 11 columns):
```

```
tweet_id               2175 non-null object
text                   2175 non-null object
rating_numerator       2175 non-null float64
rating_denominator     2175 non-null int64
name                   2175 non-null object
dog_stage              2175 non-null object
favorite_count         2175 non-null int64
retweet_count          2175 non-null int64
p                      1994 non-null object
p_conf                 1994 non-null float64
p_dog                  1994 non-null object
dtypes: float64(2), int64(3), object(6)
memory usage: 187.0+ KB
```

```
In [143]: image_predictions_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2009 entries, 0 to 2008
Data columns (total 11 columns):
tweet_id    2009 non-null object
jpg_url     2009 non-null object
p1          2009 non-null object
p1_conf     2009 non-null float64
p1_dog      2009 non-null bool
p2          2009 non-null object
p2_conf     2009 non-null float64
p2_dog      2009 non-null bool
p3          2009 non-null object
p3_conf     2009 non-null float64
p3_dog      2009 non-null bool
dtypes: bool(3), float64(3), object(5)
memory usage: 131.5+ KB
```

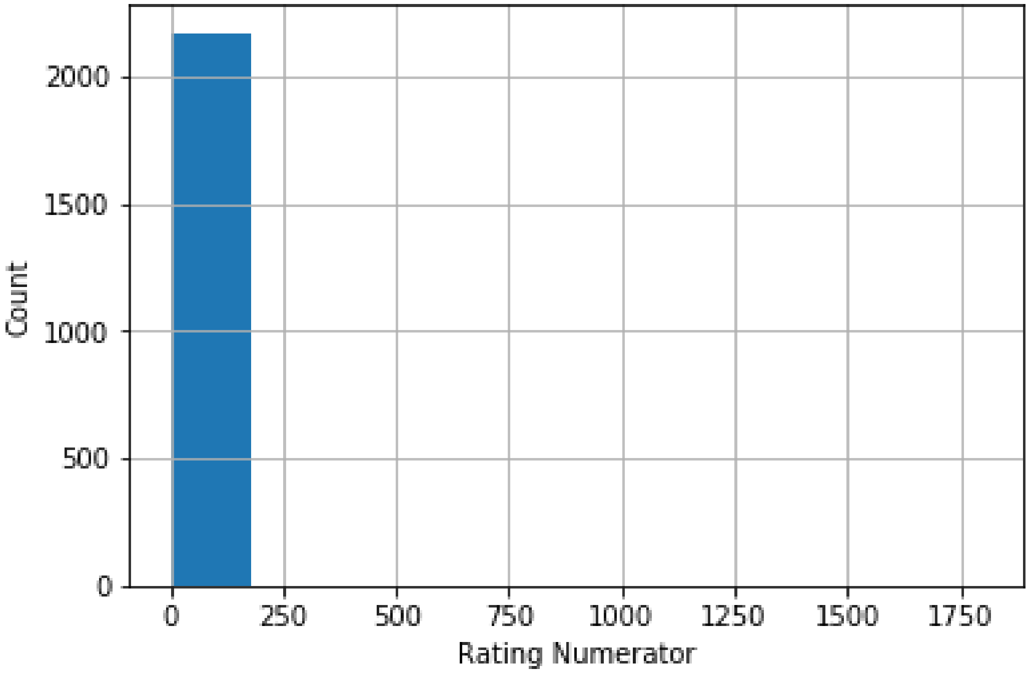
0.5 Analysis

1. rating_numerator

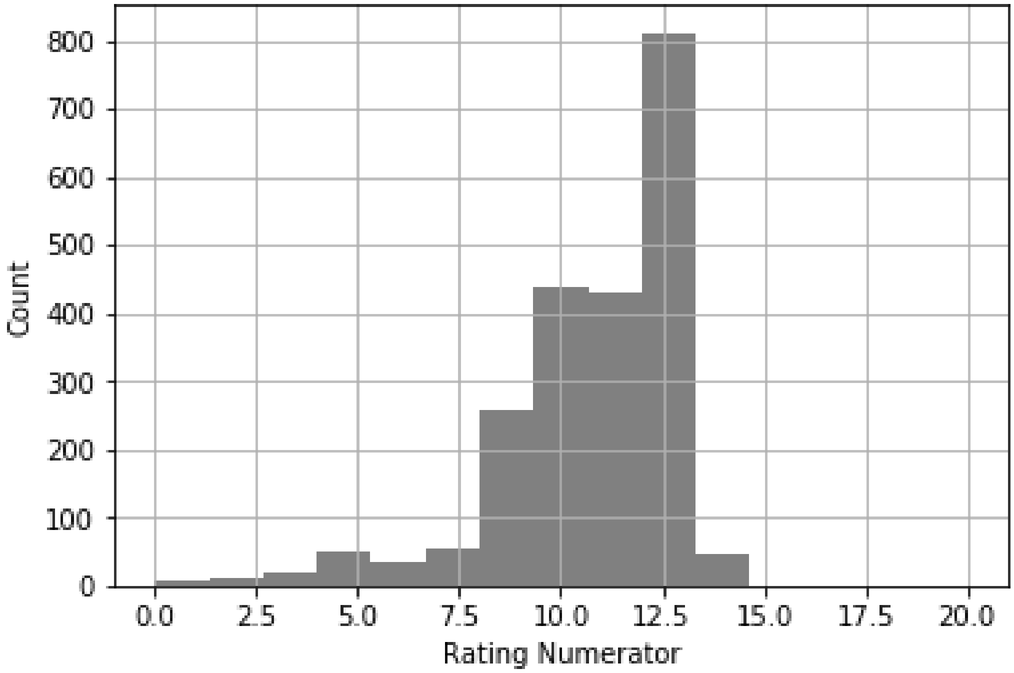
```
In [144]: twitter_archive_master.rating_numerator.describe()
```

```
Out[144]: count    2175.000000
          mean       12.722887
          std        43.157715
          min         0.000000
          25%        10.000000
          50%        11.000000
          75%        12.000000
          max      1776.000000
          Name: rating_numerator, dtype: float64
```

```
In [145]: %matplotlib inline

          twitter_archive_master.rating_numerator.hist(bins = 10, range=[0, 1800], align='mid')
          plt.xlabel('Rating Numerator')
          plt.ylabel('Count')

Out[145]: Text(0,0.5,'Count')
```



```
In [146]: %matplotlib inline
          plt.xlabel('Rating Numerator')
          plt.ylabel('Count')
          twitter_archive_master.rating_numerator.hist(bins = 15, range=[0, 20], facecolor='gray

Out[146]: <matplotlib.axes._subplots.AxesSubplot at 0x7efef759f860>
```



As we can observe the 75% quantile of rating_numerator is 12. But the maximum value is 1776. This indicates that the rating numerator has some outliers in the top 25% quantile. So we don't have to pay attention to them. On viewing the histogram of rating_numerator we can see confirm our suspicion of the outliers being present. It shows that maximum of values lies between 10 and 13 which explains the mean value of 12.718491.

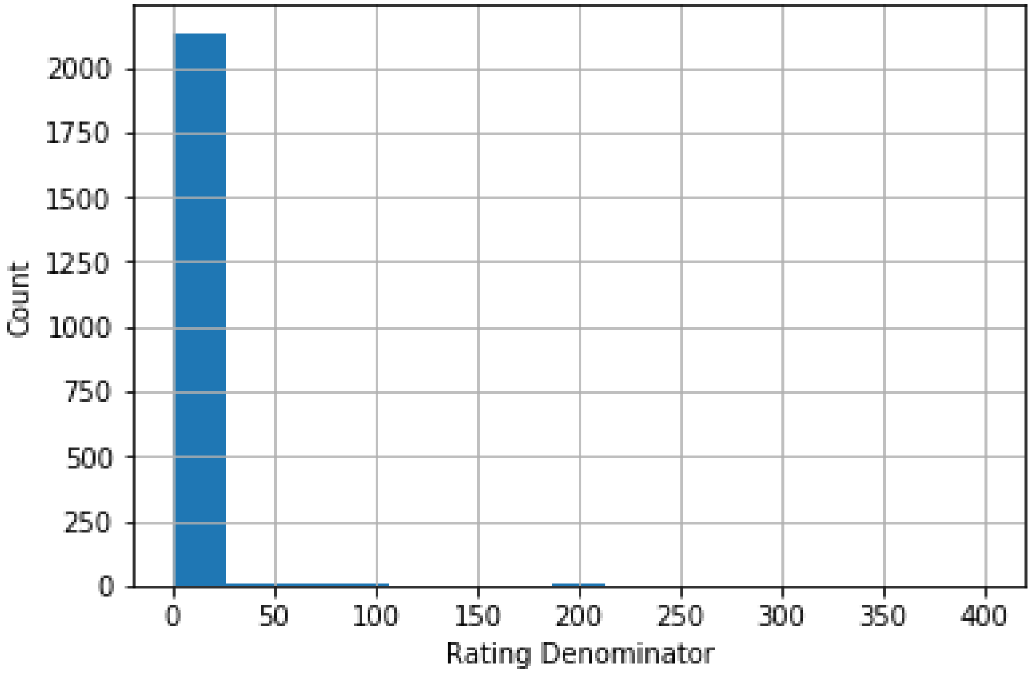
2. rating_denominator

```
In [147]: twitter_archive_master.rating_denominator.describe()
```

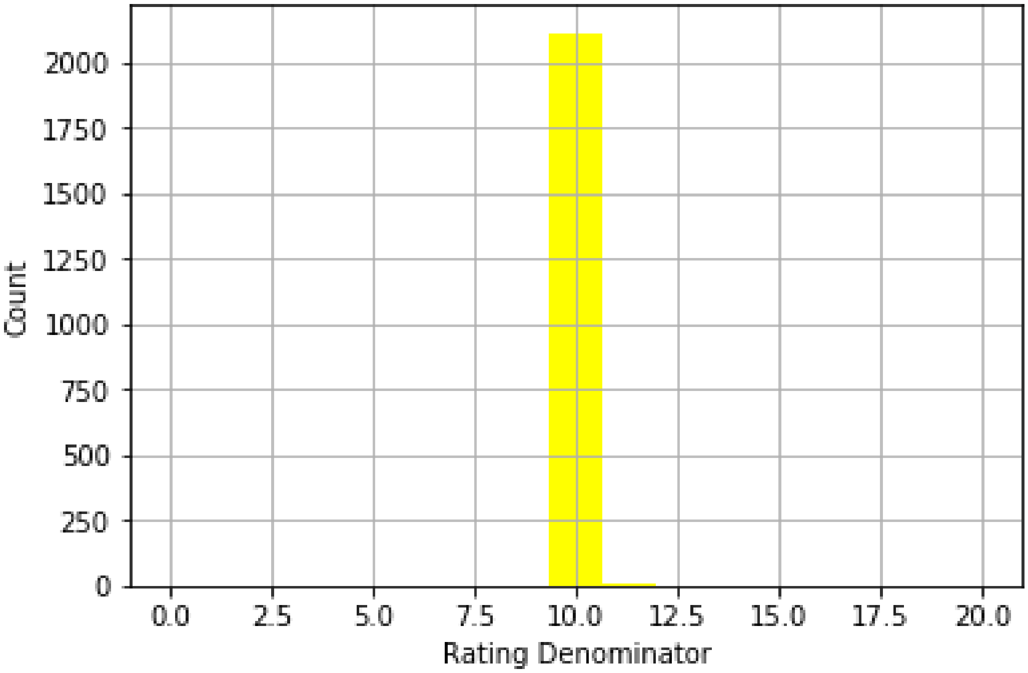
```
Out[147]: count    2175.000000
          mean       11.920000
          std        17.152464
          min         0.000000
          25%        10.000000
          50%        10.000000
          75%        10.000000
          max       400.000000
          Name: rating_denominator, dtype: float64
```

```
In [148]: %matplotlib inline
          plt.xlabel('Rating Denominator')
          plt.ylabel('Count')
          twitter_archive_master.rating_denominator.hist(bins = 15, range=[0, 400], align='mid')
```

```
Out[148]: <matplotlib.axes._subplots.AxesSubplot at 0x7efef72edb00>
```



```
In [149]: %matplotlib inline
          plt.xlabel('Rating Denominator')
          plt.ylabel('Count')
          twitter_archive_master.rating_denominator.hist(bins = 15, range=[0, 20], facecolor='ye

Out[149]: <matplotlib.axes._subplots.AxesSubplot at 0x7efef73128d0>
```



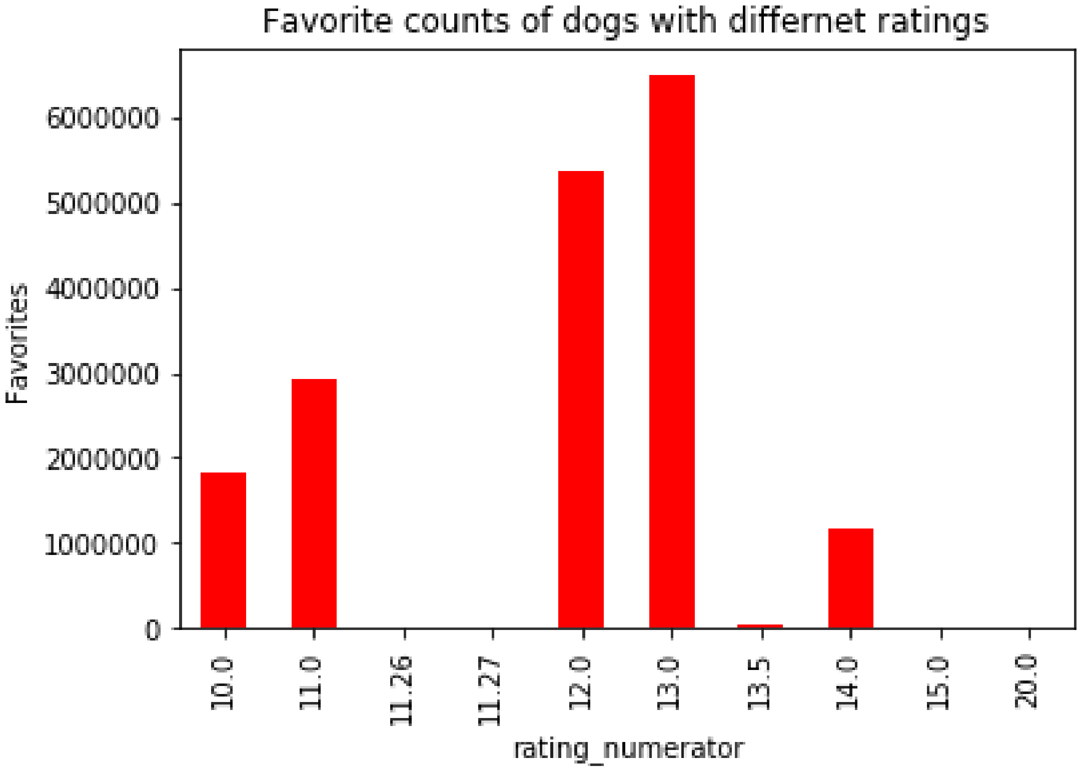
We can see that 75% quantile of rating_denominator is equal to it's 25% quantile. So most of the values is a constant 10. There are indeed some deviations and a limited number of outliers which we can observe from the mean and max of the rating_denominator. The above two plots confirm our predictions.

3. rating_numerator and favourite_count: Most common rating

```
In [150]: rate_1 = twitter_archive_master.where(twitter_archive_master.rating_numerator >= 10)
          rate_2 = rate_1.where(twitter_archive_master.rating_numerator <= 20)
          num_rate = rate_2.groupby(['rating_numerator'])['favorite_count'].sum()
```

We left out the rating less than 10 and greater than 100. Ratings greater than 20 is usually an outlier and ratings less than 10 is so low that we can ignore them.

```
In [151]: %matplotlib inline
          plt.ylabel('Favorites')
          plt.xlabel('Rating Numerator')
          num_rate.plot(kind = 'bar',
                     title = 'Favorite counts of dogs with differnet ratings',
                   color = 'red')

Out[151]: <matplotlib.axes._subplots.AxesSubplot at 0x7efef7214940>
```



Dogs with a rating of 12 and 13 has the most favorite counts. This states that the common rating for a good dog in WeRateDogs is 13. And that is why the favorite count is at peak in 13.

4. rating_numerator, favorite_counts, p and p_dog: Most popular dog breed Since we found out that the common rating_numerator is 13, we can now see which dog types are popular with a rating of 13.

```
In [152]: rating = twitter_archive_master.where(twitter_archive_master.rating_numerator == 13)
          rating_1 = rating.where(rating.p_dog == True)
          rating_1.favorite_count.describe()

Out[152]: count       204.000000
          mean      23241.289216
          std       20320.819791
          min         608.000000
          25%       10688.000000
          50%       19427.000000
          75%       28493.500000
          max      131942.000000
          Name: favorite_count, dtype: float64

In [153]: rating_2 = rating_1.where(rating_1.favorite_count > 28497)
          rating_2.favorite_count.describe()
```

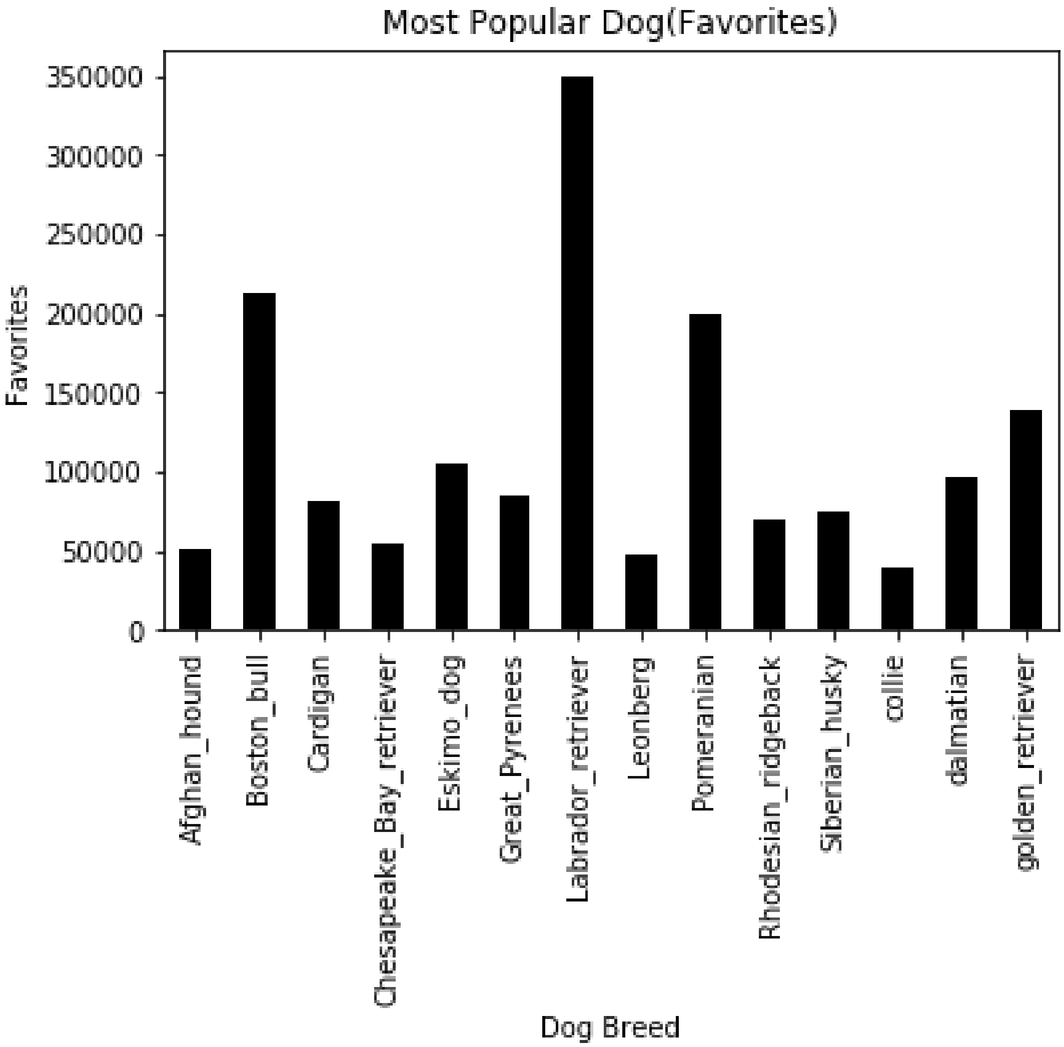
```
Out[153]: count         51.000000
          mean       48755.000000
          std        24712.066062
          min        28912.000000
          25%        34457.000000
          50%        37934.000000
          75%        53320.000000
          max       131942.000000
          Name: favorite_count, dtype: float64
```

```
In [154]: rating_3 = rating_2.where(rating_2.favorite_count > 37953)
          dog = rating_3.groupby(['p'])['favorite_count'].sum()
```

We started by finding those entries in twitter archive master table whose rating numerator is equal to 13. We extracted entries from that dataframe whose predictions are True. Then we found out that the top 25% quantile of favorite count starts from 28497. From those entries we obtained the top 50% quantile of favorite count and assigned them with value greater than that in a final dataframe which we grouped with the dog type(p).

```
In [155]: %matplotlib inline
          dog.plot(kind = 'bar',
                   title = 'Most Popular Dog(Favorites)',
                   color = 'black')
          plt.ylabel('Favorites')
          plt.xlabel('Dog Breed')
```

```
Out[155]: Text(0.5,0,'Dog Breed')
```



From the above visualization, it is evident that the Labrador Retriever is the most popular dog. It has the most favorite counts(more than 350,000). It is followed by Boston Bull, Pomeranian and Golden Retriever

5. rating_numerator, retweet_counts, p and p_dog : Most popular dog breed II

```
In [156]: ratingr = twitter_archive_master.where(twitter_archive_master.rating_numerator == 13)
          rating_1r = ratingr.where(ratingr.p_dog == True)
          rating_1r.retweet_count.describe()
```

```
Out[156]: count      204.000000
          mean      6888.529412
          std       8433.581734
          min        125.000000
          25%       2536.000000
          50%       4303.000000
          75%       7691.000000
          max      61593.000000
          Name: retweet_count, dtype: float64
```

```
In [157]: rating_2r = rating_1r.where(rating_1r.retweet_count > 7695)
          rating_2r.retweet_count.describe()
```

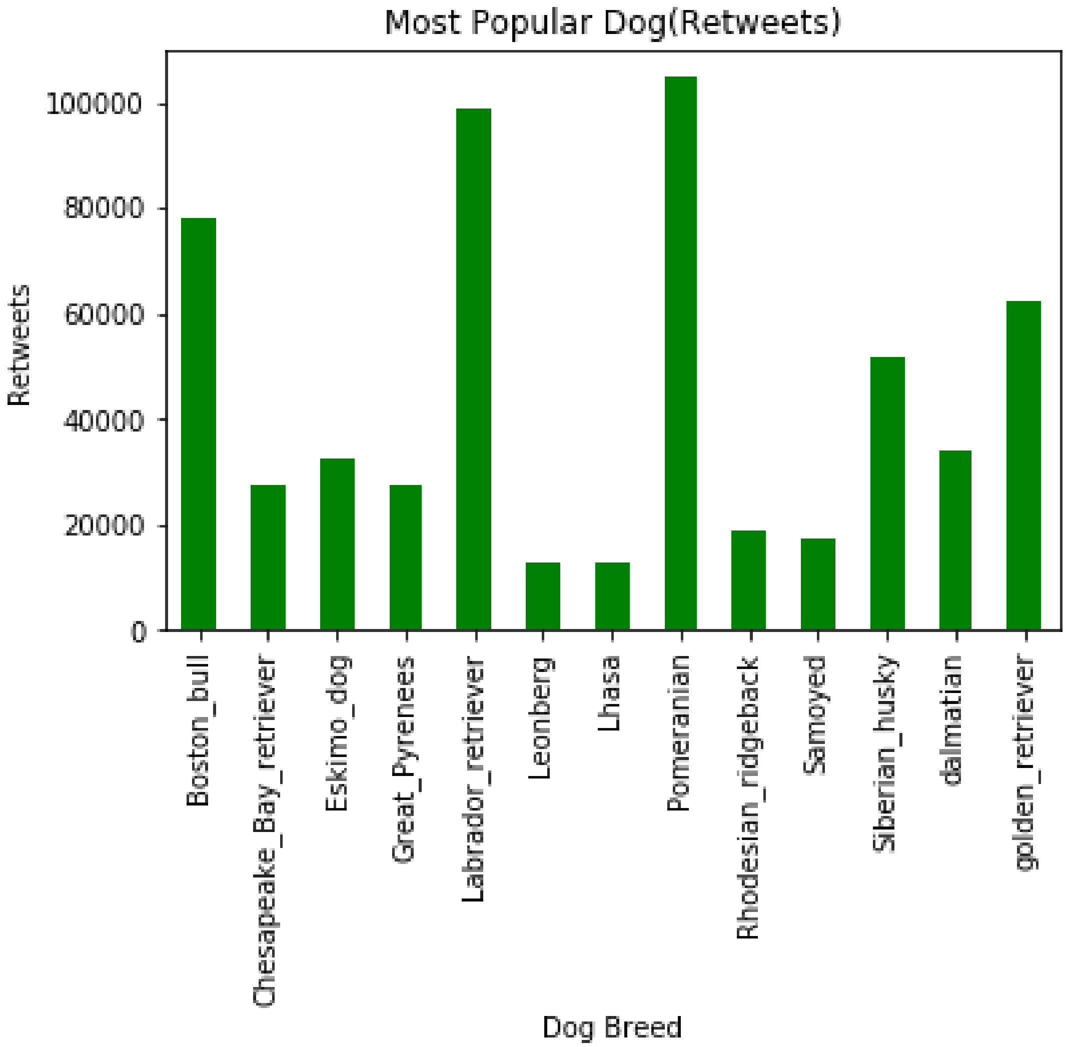
```
Out[157]: count       51.000000
          mean     16780.254902
          std      12070.308389
          min       7715.000000
          25%       9795.000000
          50%      11750.000000
          75%      18767.500000
          max      61593.000000
          Name: retweet_count, dtype: float64
```

```
In [158]: rating_3r = rating_2r.where(rating_2r.retweet_count > 11757)
          dog_1 = rating_3r.groupby(['p'])['retweet_count'].sum()
```

We started by finding those entries in twitter archive master table whose rating numerator is equal to 13. We extracted entries from that dataframe whose predictions are True. Then we found out that the top 25% quantile of favorite count starts from 7695. From those entries we obtained the top 50% quantile of favorite count and assigned them with value greater than that in a final dataframe which we grouped with the dog type(p).

```
In [159]: %matplotlib inline
          dog_1.plot(kind = 'bar',
                       title = 'Most Popular Dog(Retweets)',
                    color = 'green')
          plt.xlabel('Dog Breed')
          plt.ylabel('Retweets')
```

```
Out[159]: Text(0,0.5,'Retweets')
```



The top 4 remained the same as in the previous visualization. But as we ca see Pomeranian has got the most number of retweets(more than 100,000). It is closely followed by our popular dog based on favorite counts 'Labrador Retriever'. Boston Bull and Gloden Retriever follows them.