wrangle_act

November 21, 2020

```
In [94]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import requests
         import os
         import tweepy
         import json
         import time
         import datetime
         import matplotlib.pyplot as plt
         import seaborn as sns
         % matplotlib inline
```

0.0.1 Data Gathering

```
In [2]: twitter_archive= pd.read_csv("twitter-archive-enhanced (2).csv")
        twitter_archive.head(20)
```

```
Out[2]:               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

                    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

                                               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...

                                                 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

    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

                                        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

    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
```

```
In [3]: r = requests.get("https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_imag
```

```
In [4]: with open('image-predictions.tsv', mode ='wb') as file:
            file.write(r.content)
        image_prediction= pd.read_csv("image-predictions.tsv", sep='\t')
```

```
In [5]: image_prediction.head()
```

```
Out[5]:              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

   img_num                      p1   p1_conf  p1_dog                  p2  \
0        1  Welsh_springer_spaniel  0.465074    True              collie
1        1                 redbone  0.506826    True  miniature_pinscher
2        1         German_shepherd  0.596461    True            malinois
3        1     Rhodesian_ridgeback  0.408143    True             redbone
4        1      miniature_pinscher  0.560311    True          Rottweiler

    p2_conf  p2_dog                   p3   p3_conf  p3_dog
0  0.156665    True    Shetland_sheepdog  0.061428    True
1  0.074192    True  Rhodesian_ridgeback  0.072010    True
2  0.138584    True           bloodhound  0.116197    True
3  0.360687    True   miniature_pinscher  0.222752    True
4  0.243682    True             Doberman  0.154629    True
```

```
In [8]: key = "XXX"
        key_secret = "XXX"
        token = "XXX"
        token_secret = "XXX"

        auth = tweepy.OAuthHandler(key, key_secret)
        auth.set_access_token(token, token_secret)

        api = tweepy.API(auth)
```

```
In [9]: tweet_ids = list(twitter_archive.tweet_id)

        tweet_data = {}
        for tweet in tweet_ids:
            try:
                tweet_status = api.get_status(tweet, wait_on_rate_limit=True, wait_on_rate_limit
                tweet_data[str(tweet)] = tweet_status._json
            except:
                print("Error for: " + str(tweet))
```

```
Error for: 888202515573088257
Error for: 873697596434513921
Error for: 872668790621863937
Error for: 872261713294495745
Error for: 869988702071779329
Error for: 866816280283807744
Error for: 861769973181624320
Error for: 856602993587888130
Error for: 851953902622658560
Error for: 845459076796616705
Error for: 844704788403113984
Error for: 842892208864923648
Error for: 837366284874571778
Error for: 837012587749474308
Error for: 829374341691346946
Error for: 827228250799742977
Error for: 812747805718642688
Error for: 802247111496568832
Error for: 779123168116150273
Error for: 775096608509886464
Error for: 771004394259247104
Error for: 770743923962707968
Error for: 759566828574212096
Rate limit reached. Sleeping for: 741
Error for: 754011816964026368
Error for: 680055455951884288
Rate limit reached. Sleeping for: 741
```

```
In [10]: with open('tweet_json.txt', 'w') as file:
             json.dump(tweet_data, file)
```

```
In [11]: with open('tweet_json.txt') as file:
             data = json.load(file)

         df_list = []

         for tweet_id in data.keys():
             retweets = data[tweet_id]['retweet_count']
             favorites = data[tweet_id]['favorite_count']
             df_list.append({'tweet_id': tweet_id,
                                 'retweets': retweets,
                                 'favorites': favorites})

         tweets_df = pd.DataFrame(df_list, columns = ['tweet_id', 'retweets', 'favorites'])
         tweets_df.head(30)
```

```
Out[11]:               tweet_id  retweets  favorites
         0   892420643555336193      7484      35437
         1   892177421306343426      5551      30653
         2   891815181378084864      3678      23060
         3   891689557279858688      7657      38733
         4   891327558926688256      8262      36991
         5   891087950875897856      2765      18645
         6   890971913173991426      1796      10838
         7   890729181411237888     16748      59710
         8   890609185150312448      3819      25666
```

```
9   890240255349198849      6503      29299
10  890006608113172480      6508      28230
11  889880896479866881      4421      25682
12  889665388333682689      8872      44132
13  889638837579907072      3972      24836
14  889531135344209921      2004      13962
15  889278841981685760      4727      23172
16  888917238123831296      3979      26786
17  888804989199671297      3754      23504
18  888554962724278272      3068      18122
19  888078434458587136      3076      20034
20  887705289381826560      4790      27852
21  887517139158093824     10455      42650
22  887473957103951883     15955      63167
23  887343217045368832      9341      30965
24  887101392804085760      5291      28175
25  886983233522544640      6786      32014
26  886736880519319552      2830      10995
27  886680336477933568      3976      20686
28  886366144734445568      2808      19429
29  886267009285017600         4        110
```

0.0.2 Assessment

```
In [12]: twitter_archive.shape
```

```
Out[12]: (2356, 17)
```

```
In [13]: 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 [15]: sum(twitter_archive.duplicated())
```

```
Out[15]: 0
```

```
In [16]: twitter_archive.isnull().sum()
```

```
Out[16]: tweet_id                         0
         in_reply_to_status_id         2278
         in_reply_to_user_id           2278
         timestamp                        0
         source                           0
         text                             0
         retweeted_status_id           2175
         retweeted_status_user_id      2175
         retweeted_status_timestamp    2175
         expanded_urls                   59
         rating_numerator                 0
         rating_denominator               0
         name                             0
         doggo                            0
         floofer                          0
         pupper                           0
         puppo                            0
         dtype: int64
```

```
In [17]: image_prediction.shape
```

```
Out[17]: (2075, 12)
```

```
In [18]: image_prediction.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null int64
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
```

```
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

```
In [19]: sum(image_prediction.duplicated())
```

```
Out[19]: 0
```

```
In [20]: image_prediction.isnull().sum()
```

```
Out[20]: tweet_id    0
         jpg_url     0
         img_num     0
         p1          0
         p1_conf     0
         p1_dog      0
         p2          0
         p2_conf     0
         p2_dog      0
         p3          0
         p3_conf     0
         p3_dog      0
         dtype: int64
```

```
In [21]: image_prediction.tail(30)
```

```
Out[21]:                 tweet_id                                           jpg_url  \
         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  \
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  \
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
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
```

```
In [22]: tweets_df.shape
```

```
Out[22]: (2331, 3)
```

```
In [23]: sum(tweets_df.duplicated())
```

```
Out[23]: 0
```

```
In [24]: tweets_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2331 entries, 0 to 2330
Data columns (total 3 columns):
tweet_id     2331 non-null object
retweets     2331 non-null int64
favorites    2331 non-null int64
dtypes: int64(2), object(1)
memory usage: 54.7+ KB
```

```
In [25]: tweets_df.isnull().sum()
```

```
Out[25]: tweet_id     0
         retweets     0
         favorites    0
         dtype: int64
```

```
In [26]: twitter_archive.tail()
```

```
Out[26]:                  tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
        2351  666049248165822465                    NaN                  NaN
        2352  666044226329800704                    NaN                  NaN
        2353  666033412701032449                    NaN                  NaN
        2354  666029285002620928                    NaN                  NaN
        2355  666020888022790149                    NaN                  NaN

                              timestamp  \
        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  \
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  \
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  \
2351                       NaN                        NaN
2352                       NaN                        NaN
2353                       NaN                        NaN
2354                       NaN                        NaN
2355                       NaN                        NaN

                                          expanded_urls  rating_numerator  \
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
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
```

```
In [33]: image_prediction.tail()
```

```
Out[33]:                 tweet_id                                          jpg_url  \
         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                  p2   p2_conf  \
2070        2      basset  0.555712    True    English_springer  0.225770
2071        1  paper_towel  0.170278   False  Labrador_retriever  0.168086
2072        1   Chihuahua  0.716012    True            malamute  0.078253
2073        1   Chihuahua  0.323581    True            Pekinese  0.090647
2074        1      orange  0.097049   False               bagel  0.085851

      p2_dog                           p3   p3_conf  p3_dog
2070    True  German_short-haired_pointer  0.175219    True
2071    True                      spatula  0.040836   False
2072    True                       kelpie  0.031379    True
2073    True                     papillon  0.068957    True
2074   False                       banana  0.076110   False
```

```
In [27]: twitter_archive['doggo'].nunique()
```

```
Out[27]: 2
```

```
In [28]: twitter_archive['floofer'].nunique()
```

```
Out[28]: 2
```

```
In [29]: twitter_archive['name'].unique()
```

```
Out[29]: array(['Phineas', 'Tilly', 'Archie', 'Darla', 'Franklin', 'None', 'Jax',
                'Zoey', 'Cassie', 'Koda', 'Bruno', 'Ted', 'Stuart', 'Oliver', 'Jim',
                'Zeke', 'Ralphus', 'Canela', 'Gerald', 'Jeffrey', 'such', 'Maya',
                'Mingus', 'Derek', 'Roscoe', 'Waffles', 'Jimbo', 'Maisey', 'Lilly',
                'Earl', 'Lola', 'Kevin', 'Yogi', 'Noah', 'Bella', 'Grizzwald',
                'Rusty', 'Gus', 'Stanley', 'Alfy', 'Koko', 'Rey', 'Gary', 'a',
                'Elliot', 'Louis', 'Jesse', 'Romeo', 'Bailey', 'Duddles', 'Jack',
                'Emmy', 'Steven', 'Beau', 'Snoopy', 'Shadow', 'Terrance', 'Aja',
                'Penny', 'Dante', 'Nelly', 'Ginger', 'Benedict', 'Venti', 'Goose',
                'Nugget', 'Cash', 'Coco', 'Jed', 'Sebastian', 'Walter', 'Sierra',
                'Monkey', 'Harry', 'Kody', 'Lassie', 'Rover', 'Napolean', 'Dawn',
                'Boomer', 'Cody', 'Rumble', 'Clifford', 'quite', 'Dewey', 'Scout',
                'Gizmo', 'Cooper', 'Harold', 'Shikha', 'Jamesy', 'Lili', 'Sammy',
                'Meatball', 'Paisley', 'Albus', 'Neptune', 'Quinn', 'Belle',
                'Zooey', 'Dave', 'Jersey', 'Hobbes', 'Burt', 'Lorenzo', 'Carl',
                'Jordy', 'Milky', 'Trooper', 'Winston', 'Sophie', 'Wyatt', 'Rosie',
                'Thor', 'Oscar', 'Luna', 'Callie', 'Cermet', 'George', 'Marlee',
                'Arya', 'Einstein', 'Alice', 'Rumpole', 'Benny', 'Aspen', 'Jarod',
                'Wiggles', 'General', 'Sailor', 'Astrid', 'Iggy', 'Snoop', 'Kyle',
                'Leo', 'Riley', 'Gidget', 'Noosh', 'Odin', 'Jerry', 'Charlie',
                'Georgie', 'Rontu', 'Cannon', 'Furzey', 'Daisy', 'Tuck', 'Barney',
                'Vixen', 'Jarvis', 'Mimosa', 'Pickles', 'Bungalo', 'Brady', 'Margo',
                'Sadie', 'Hank', 'Tycho', 'Stephan', 'Indie', 'Winnie', 'Bentley',
                'Ken', 'Max', 'Maddie', 'Pipsy', 'Monty', 'Sojourner', 'Odie',
'Arlo', 'Sunny', 'Vincent', 'Lucy', 'Clark', 'Mookie', 'Meera',
'Buddy', 'Ava', 'Rory', 'Eli', 'Ash', 'Tucker', 'Tobi', 'Chester',
'Wilson', 'Sunshine', 'Lipton', 'Gabby', 'Bronte', 'Poppy', 'Rhino',
'Willow', 'not', 'Orion', 'Eevee', 'Smiley', 'Logan', 'Moreton',
'Klein', 'Miguel', 'Emanuel', 'Kuyu', 'Dutch', 'Pete', 'Scooter',
'Reggie', 'Kyro', 'Samson', 'Loki', 'Mia', 'Malcolm', 'Dexter',
'Alfie', 'Fiona', 'one', 'Mutt', 'Bear', 'Doobert', 'Beebop',
'Alexander', 'Sailer', 'Brutus', 'Kona', 'Boots', 'Ralphie', 'Phil',
'Cupid', 'Pawnd', 'Pilot', 'Ike', 'Mo', 'Toby', 'Sweet', 'Pablo',
'Nala', 'Balto', 'Crawford', 'Gabe', 'Mattie', 'Jimison',
'Hercules', 'Duchess', 'Harlso', 'Sampson', 'Sundance', 'Luca',
'Flash', 'Finn', 'Peaches', 'Howie', 'Jazzy', 'Anna', 'Bo',
'Seamus', 'Wafer', 'Chelsea', 'Tom', 'Moose', 'Florence', 'Autumn',
'Dido', 'Eugene', 'Herschel', 'Strudel', 'Tebow', 'Chloe', 'Betty',
'Timber', 'Binky', 'Dudley', 'Comet', 'Larry', 'Levi', 'Akumi',
'Titan', 'Olivia', 'Alf', 'Oshie', 'Bruce', 'Chubbs', 'Sky',
'Atlas', 'Eleanor', 'Layla', 'Rocky', 'Baron', 'Tyr', 'Bauer',
'Swagger', 'Brandi', 'Mary', 'Moe', 'Halo', 'Augie', 'Craig', 'Sam',
'Hunter', 'Pavlov', 'Maximus', 'Wallace', 'Ito', 'Milo', 'Ollie',
'Cali', 'Lennon', 'incredibly', 'Major', 'Duke', 'Reginald',
'Sansa', 'Shooter', 'Django', 'Diogi', 'Sonny', 'Philbert',
'Marley', 'Severus', 'Ronnie', 'Anakin', 'Bones', 'Mauve', 'Chef',
'Doc', 'Sobe', 'Longfellow', 'Mister', 'Iroh', 'Baloo', 'Stubert',
'Paull', 'Timison', 'Davey', 'Pancake', 'Tyrone', 'Snicku', 'Ruby',
'Brody', 'Rizzy', 'Mack', 'Butter', 'Nimbus', 'Laika', 'Dobby',
'Juno', 'Maude', 'Lily', 'Newt', 'Benji', 'Nida', 'Robin',
'Monster', 'BeBe', 'Remus', 'Mabel', 'Misty', 'Happy', 'Mosby',
'Maggie', 'Leela', 'Ralphy', 'Brownie', 'Meyer', 'Stella', 'mad',
'Frank', 'Tonks', 'Lincoln', 'Oakley', 'Dale', 'Rizzo', 'Arnie',
'Pinot', 'Dallas', 'Hero', 'Frankie', 'Stormy', 'Mairi', 'Loomis',
'Godi', 'Kenny', 'Deacon', 'Timmy', 'Harper', 'Chipson', 'Combo',
'Dash', 'Bell', 'Hurley', 'Jay', 'Mya', 'Strider', 'an', 'Wesley',
'Solomon', 'Huck', 'very', 'O', 'Blue', 'Finley', 'Sprinkles',
'Heinrich', 'Shakespeare', 'Fizz', 'Chip', 'Grey', 'Roosevelt',
'Gromit', 'Willem', 'Dakota', 'Dixie', 'Al', 'Jackson', 'just',
'Carbon', 'DonDon', 'Kirby', 'Lou', 'Nollie', 'Chevy', 'Tito',
'Louie', 'Rupert', 'Rufus', 'Brudge', 'Shadoe', 'Colby', 'Angel',
'Brat', 'Tove', 'my', 'Aubie', 'Kota', 'Eve', 'Glenn', 'Shelby',
'Sephie', 'Bonaparte', 'Albert', 'Wishes', 'Rose', 'Theo', 'Rocco',
'Fido', 'Emma', 'Spencer', 'Lilli', 'Boston', 'Brandonald', 'Corey',
'Leonard', 'Chompsky', 'Beckham', 'Devón', 'Gert', 'Watson',
'Rubio', 'Keith', 'Dex', 'Carly', 'Ace', 'Tayzie', 'Grizzie',
'Fred', 'Gilbert', 'Zoe', 'Stewie', 'Calvin', 'Lilah', 'Spanky',
'Jameson', 'Piper', 'Atticus', 'Blu', 'Dietrich', 'Divine', 'Tripp',
'his', 'Cora', 'Huxley', 'Keurig', 'Bookstore', 'Linus', 'Abby',
'Shaggy', 'Shiloh', 'Gustav', 'Arlen', 'Percy', 'Lenox', 'Sugar',
'Harvey', 'Blanket', 'actually', 'Geno', 'Stark', 'Beya', 'Kilo',
'Kayla', 'Maxaroni', 'Doug', 'Edmund', 'Aqua', 'Theodore', 'Chase',
'getting', 'Rorie', 'Simba', 'Charles', 'Bayley', 'Axel',
'Storkson', 'Remy', 'Chadrick', 'Kellogg', 'Buckley', 'Livvie',
'Terry', 'Hermione', 'Ralpher', 'Aldrick', 'this', 'unacceptable',
'Rooney', 'Crystal', 'Ziva', 'Stefan', 'Pupcasso', 'Puff',
'Flurpson', 'Coleman', 'Enchilada', 'Raymond', 'all', 'Rueben',
'Cilantro', 'Karll', 'Sprout', 'Blitz', 'Bloop', 'Lillie',
'Ashleigh', 'Kreggory', 'Sarge', 'Luther', 'Ivar', 'Jangle',
'Schnitzel', 'Panda', 'Berkeley', 'Ralphé', 'Charleson', 'Clyde',
'Harnold', 'Sid', 'Pippa', 'Otis', 'Carper', 'Bowie',
'Alexanderson', 'Suki', 'Barclay', 'Skittle', 'Ebby', 'Flávio',
'Smokey', 'Link', 'Jennifur', 'Ozzy', 'Bluebert', 'Stephanus',
'Bubbles', 'old', 'Zeus', 'Bertson', 'Nico', 'Michelangelope',
'Siba', 'Calbert', 'Curtis', 'Travis', 'Thumas', 'Kanu', 'Lance',
'Opie', 'Kane', 'Olive', 'Chuckles', 'Staniel', 'Sora', 'Beemo',
'Gunner', 'infuriating', 'Lacy', 'Tater', 'Olaf', 'Cecil', 'Vince',
'Karma', 'Billy', 'Walker', 'Rodney', 'Klevin', 'Malikai', 'Bobble',
'River', 'Jebberson', 'Remington', 'Farfle', 'Jiminus', 'Clarkus',
'Finnegus', 'Cupcake', 'Kathmandu', 'Ellie', 'Katie', 'Kara',
'Adele', 'Zara', 'Ambrose', 'Jimothy', 'Bode', 'Terrenth', 'Reese',
'Chesterson', 'Lucia', 'Bisquick', 'Ralphson', 'Socks', 'Rambo',
'Rudy', 'Fiji', 'Rilo', 'Bilbo', 'Coopson', 'Yoda', 'Millie',
'Chet', 'Crouton', 'Daniel', 'Kaia', 'Murphy', 'Dotsy', 'Eazy',
'Coops', 'Fillup', 'Miley', 'Charl', 'Reagan', 'Yukon', 'CeCe',
'Cuddles', 'Claude', 'Jessiga', 'Carter', 'Ole', 'Pherb', 'Blipson',
'Reptar', 'Trevith', 'Berb', 'Bob', 'Colin', 'Brian', 'Oliviér',
'Grady', 'Kobe', 'Freddery', 'Bodie', 'Dunkin', 'Wally', 'Tupawc',
'Amber', 'Edgar', 'Teddy', 'Kingsley', 'Brockly', 'Richie', 'Molly',
'Vinscent', 'Cedrick', 'Hazel', 'Lolo', 'Eriq', 'Phred', 'the',
'Oddie', 'Maxwell', 'Geoff', 'Covach', 'Durg', 'Fynn', 'Ricky',
'Herald', 'Lucky', 'Ferg', 'Trip', 'Clarence', 'Hamrick', 'Brad',
'Pubert', 'Frönq', 'Derby', 'Lizzie', 'Ember', 'Blakely', 'Opal',
'Marq', 'Kramer', 'Barry', 'Gordon', 'Baxter', 'Mona', 'Horace',
'Crimson', 'Birf', 'Hammond', 'Lorelei', 'Marty', 'Brooks',
'Petrick', 'Hubertson', 'Gerbald', 'Oreo', 'Bruiser', 'Perry',
'Bobby', 'Jeph', 'Obi', 'Tino', 'Kulet', 'Sweets', 'Lupe', 'Tiger',
'Jiminy', 'Griffin', 'Banjo', 'Brandy', 'Lulu', 'Darrel', 'Taco',
'Joey', 'Patrick', 'Kreg', 'Todo', 'Tess', 'Ulysses', 'Toffee',
'Apollo', 'Asher', 'Glacier', 'Chuck', 'Champ', 'Ozzie', 'Griswold',
'Cheesy', 'Moofasa', 'Hector', 'Goliath', 'Kawhi', 'by', 'Emmie',
'Penelope', 'Willie', 'Rinna', 'Mike', 'William', 'Dwight', 'Evy',
'officially', 'Rascal', 'Linda', 'Tug', 'Tango', 'Grizz', 'Jerome',
'Crumpet', 'Jessifer', 'Izzy', 'Ralph', 'Sandy', 'Humphrey',
'Tassy', 'Juckson', 'Chuq', 'Tyrus', 'Karl', 'Godzilla', 'Vinnie',
'Kenneth', 'Herm', 'Bert', 'Striker', 'Donny', 'Pepper', 'Bernie',
'Buddah', 'Lenny', 'Arnold', 'Zuzu', 'Mollie', 'Laela', 'Tedders',
'Superpup', 'Rufio', 'Jeb', 'Rodman', 'Jonah', 'Chesney', 'life',
'Henry', 'Bobbay', 'Mitch', 'Kaiya', 'Acro', 'Aiden', 'Obie', 'Dot',
'Shnuggles', 'Kendall', 'Jeffri', 'Steve', 'Mac', 'Fletcher',
'Kenzie', 'Pumpkin', 'Schnozz', 'Gustaf', 'Cheryl', 'Ed',
'Leonidas', 'Norman', 'Caryl', 'Scott', 'Taz', 'Darby', 'Jackie',
'light', 'Jazz', 'Franq', 'Pippin', 'Rolf', 'Snickers', 'Ridley',
'Cal', 'Bradley', 'Bubba', 'Tuco', 'Patch', 'Mojo', 'Batdog',
'Dylan', 'space', 'Mark', 'JD', 'Alejandro', 'Scruffers', 'Pip',
'Julius', 'Tanner', 'Sparky', 'Anthony', 'Holly', 'Jett', 'Amy',
'Sage', 'Andy', 'Mason', 'Trigger', 'Antony', 'Creg', 'Traviss',
'Gin', 'Jeffrie', 'Danny', 'Ester', 'Pluto', 'Bloo', 'Edd', 'Willy',
'Herb', 'Damon', 'Peanut', 'Nigel', 'Butters', 'Sandra', 'Fabio',
'Randall', 'Liam', 'Tommy', 'Ben', 'Raphael', 'Julio', 'Andru',
'Kloey', 'Shawwn', 'Skye', 'Kollin', 'Ronduh', 'Billl', 'Saydee',
'Dug', 'Tessa', 'Sully', 'Kirk', 'Ralf', 'Clarq', 'Jaspers',
'Samsom', 'Harrison', 'Chaz', 'Jeremy', 'Jaycob', 'Lambeau',
'Ruffles', 'Amélie', 'Bobb', 'Banditt', 'Kevon', 'Winifred', 'Hanz',
'Churlie', 'Zeek', 'Timofy', 'Maks', 'Jomathan', 'Kallie', 'Marvin',
'Spark', 'Gòrdón', 'Jo', 'DayZ', 'Jareld', 'Torque', 'Ron',
'Skittles', 'Cleopatricia', 'Erik', 'Stu', 'Tedrick', 'Filup',
'Kial', 'Naphaniel', 'Dook', 'Hall', 'Philippe', 'Biden', 'Fwed',
'Genevieve', 'Joshwa', 'Bradlay', 'Clybe', 'Keet', 'Carll',
'Jockson', 'Josep', 'Lugan', 'Christoper'], dtype=object)
```

Quality: 1. The columns on the file called twitter_archive have to many missing values. They should be dropped. 2. remove +0000 from timestamp 3. No need the imege number coulmn 4. Timestamp is object, it should be date 5. The collumn called name has invalid records such as a and none. It should be cleaned. 6. drop the text column 7. Clean the source list 8. Clean rating_numerator′ values if they higher than 20

Tideness: 1. The file has doggo, floofer, pupper, and puppo columns. It can be combined and created a new column called dog_type. 2. Combined the files called twitter_archive_cpy and tweets_df_cpy, and image_prediction_cpy

0.0.3 Cleaning Data

```
In [31]: twitter_archive_cpy= twitter_archive.copy()
         image_prediction_cpy=image_prediction.copy()
         tweets_df_cpy=tweets_df.copy()
```

```
In [32]: twitter_archive_cpy.drop(['in_reply_to_status_id','in_reply_to_user_id','retweeted_stat
```

```
In [33]: twitter_archive_cpy.head(3)
```

```
Out[33]:              tweet_id                  timestamp  \
         0  892420643555336193  2017-08-01 16:23:56 +0000
         1  892177421306343426  2017-08-01 00:17:27 +0000
         2  891815181378084864  2017-07-31 00:18:03 +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...
```

```
                                                text  rating_numerator  \
0  This is Phineas. He's a mystical boy. Only eve...                13
1  This is Tilly. She's just checking pup on you...                13
2  This is Archie. He is a rare Norwegian Pouncin...                12

   rating_denominator     name doggo floofer pupper puppo
0                  10  Phineas  None    None   None  None
1                  10    Tilly  None    None   None  None
2                  10   Archie  None    None   None  None
```

```
In [35]: twitter_archive_cpy.timestamp = twitter_archive_cpy.timestamp.str[:-5].str.strip()
```

```
In [37]: twitter_archive_cpy.timestamp.head(3)
```

```
Out[37]: 0    2017-08-01 16:23:56
         1    2017-08-01 00:17:27
         2    2017-07-31 00:18:03
         Name: timestamp, dtype: object
```

```
In [38]: image_prediction_cpy.drop('img_num', axis=1, inplace=True)
```

```
In [39]: image_prediction_cpy.head(2)
```

```
Out[39]:              tweet_id                                          jpg_url  \
         0  666020888022790149  https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg
         1  666029285002620928  https://pbs.twimg.com/media/CT42GRgUYAA5iDo.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

            p2_dog                   p3   p3_conf  p3_dog
         0    True    Shetland_sheepdog  0.061428    True
         1    True  Rhodesian_ridgeback  0.072010    True
```

```
In [41]:  twitter_archive_cpy['timestamp'] = pd.to_datetime(twitter_archive_cpy['timestamp'])
```

```
In [42]: twitter_archive_cpy['timestamp'].head(3)
```

```
Out[42]: 0   2017-08-01 16:23:56
         1   2017-08-01 00:17:27
         2   2017-07-31 00:18:03
         Name: timestamp, dtype: datetime64[ns]
```

```
In [45]: twitter_archive_cpy.name.replace(['None', 'a','an', 'the'], np.nan, inplace=True)
```

```
In [46]: twitter_archive_cpy.name.value_counts()
```

```
Out[46]: Charlie    12
         Cooper     11
         Lucy       11
         Oliver     11
         Lola       10
         Tucker     10
         Penny      10
         Bo          9
         Winston     9
         Sadie       8
         Bailey      7
         Toby        7
         Daisy       7
         Buddy       7
         Leo         6
         Scout       6
         Koda        6
         Milo        6
         Oscar       6
         Rusty       6
         Jack        6
         Jax         6
         Stanley     6
         Bella       6
         Dave        6
         Sunny       5
         George      5
         very        5
         Chester     5
         Larry       5
                    ..
         Ember       1
         Edgar       1
         Moofasa     1
         Arya        1
         Staniel     1
         Brandy      1
         Barclay     1
         Lassie      1
         Stormy      1
         Livvie      1
         Devón       1
         Angel       1
         Liam        1
         Eriq        1
         Hanz        1
         Milky       1
         Shikha      1
```

```
Rueben      1
Newt        1
Rey         1
Clyde       1
Aubie       1
Sailor      1
Bruno       1
Grady       1
Crumpet     1
Scott       1
Goliath     1
Geno        1
Kevon       1
Name: name, Length: 953, dtype: int64
```

```
In [47]: twitter_archive_cpy.drop('text', axis=1, inplace=True)
```

```
In [48]: twitter_archive_cpy.head(3)
```

```
Out[48]:              tweet_id           timestamp  \
         0  892420643555336193 2017-08-01 16:23:56
         1  892177421306343426 2017-08-01 00:17:27
         2  891815181378084864 2017-07-31 00:18:03

                                                        source  rating_numerator  \
         0  <a href="http://twitter.com/download/iphone" r...                13
         1  <a href="http://twitter.com/download/iphone" r...                13
         2  <a href="http://twitter.com/download/iphone" r...                12

            rating_denominator     name doggo floofer pupper puppo
         0                  10  Phineas  None    None   None  None
         1                  10    Tilly  None    None   None  None
         2                  10   Archie  None    None   None  None
```

```
In [49]: sourcelist = ['<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
                       '<a href="http://vine.co" rel="nofollow">Vine - Make a Scene</a>',
                       '<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>',
                       '<a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">Tw
         newsourcelist=['iPhone', 'Vine', 'WebClient', 'TweetDeck']
```

```
In [50]: twitter_archive_cpy.source.replace(sourcelist, newsourcelist, inplace=True)
```

```
In [52]: twitter_archive_cpy.source.sample(10)
```

```
Out[52]: 2195    iPhone
         1741    iPhone
         2147    iPhone
         683     iPhone
         207     iPhone
```

```
128     iPhone
589     iPhone
145     iPhone
1114    iPhone
2347    iPhone
Name: source, dtype: object
```

```
In [53]: twitter_archive_cpy.rating_numerator.value_counts()
```

```
Out[53]: 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 [55]: twitter_archive_cpy.loc[twitter_archive_cpy['rating_numerator']>20, 'rating_numerator']
```

```
In [56]: twitter_archive_cpy.rating_numerator.value_counts()
```

```
Out[56]: 12    558
         11    464
         10    461
         13    351
         9     158
         8     102
         7      55
         14     54
         5      37
         6      32
         20     25
         3      19
         4      17
         1       9
         2       9
         0       2
         15      2
         17      1
         Name: rating_numerator, dtype: int64
```

```
In [57]: twitter_archive_cpy.loc[twitter_archive_cpy['doggo'] == 'doggo', 'dog_class'] = 'doggo'
         twitter_archive_cpy.loc[twitter_archive_cpy['floofer'] == 'floofer', 'dog_class'] = 'fl
         twitter_archive_cpy.loc[twitter_archive_cpy['pupper'] == 'pupper', 'dog_class'] = 'pupp
         twitter_archive_cpy.loc[twitter_archive_cpy['puppo'] == 'puppo', 'dog_class'] = 'puppo'
```

```
In [58]: twitter_archive_cpy.head(10)
```

```
Out[58]:              tweet_id           timestamp  source  rating_numerator  \
         0  892420643555336193 2017-08-01 16:23:56  iPhone                13
         1  892177421306343426 2017-08-01 00:17:27  iPhone                13
         2  891815181378084864 2017-07-31 00:18:03  iPhone                12
         3  891689557279858688 2017-07-30 15:58:51  iPhone                13
         4  891327558926688256 2017-07-29 16:00:24  iPhone                12
         5  891087950875897856 2017-07-29 00:08:17  iPhone                13
         6  890971913173991426 2017-07-28 16:27:12  iPhone                13
         7  890729181411237888 2017-07-28 00:22:40  iPhone                13
         8  890609185150312448 2017-07-27 16:25:51  iPhone                13
         9  890240255349198849 2017-07-26 15:59:51  iPhone                14

            rating_denominator      name  doggo floofer pupper puppo dog_class
         0                  10   Phineas   None    None   None  None       NaN
         1                  10     Tilly   None    None   None  None       NaN
```

```
2                     10    Archie   None    None   None  None       NaN
3                     10     Darla   None    None   None  None       NaN
4                     10  Franklin   None    None   None  None       NaN
5                     10       NaN   None    None   None  None       NaN
6                     10       Jax   None    None   None  None       NaN
7                     10       NaN   None    None   None  None       NaN
8                     10      Zoey   None    None   None  None       NaN
9                     10    Cassie  doggo    None   None  None     doggo
```

```
In [59]: twitter_archive_cpy.drop(['doggo', 'floofer','pupper','puppo'], axis=1, inplace=True)
```

```
In [62]: twitter_archive_cpy.head(10)
```

```
Out[62]:             tweet_id           timestamp  source  rating_numerator  \
         0  892420643555336193 2017-08-01 16:23:56  iPhone                13
         1  892177421306343426 2017-08-01 00:17:27  iPhone                13
         2  891815181378084864 2017-07-31 00:18:03  iPhone                12
         3  891689557279858688 2017-07-30 15:58:51  iPhone                13
         4  891327558926688256 2017-07-29 16:00:24  iPhone                12
         5  891087950875897856 2017-07-29 00:08:17  iPhone                13
         6  890971913173991426 2017-07-28 16:27:12  iPhone                13
         7  890729181411237888 2017-07-28 00:22:40  iPhone                13
         8  890609185150312448 2017-07-27 16:25:51  iPhone                13
         9  890240255349198849 2017-07-26 15:59:51  iPhone                14

            rating_denominator      name dog_class
         0                  10   Phineas       NaN
         1                  10     Tilly       NaN
         2                  10    Archie       NaN
         3                  10     Darla       NaN
         4                  10  Franklin       NaN
         5                  10       NaN       NaN
         6                  10       Jax       NaN
         7                  10       NaN       NaN
         8                  10      Zoey       NaN
         9                  10    Cassie     doggo
```

```
In [66]: twitter_archive_cpy = twitter_archive_cpy.merge(image_prediction_cpy, on='tweet_id', ho
```

```
In [67]: twitter_archive_cpy.head()
```

```
Out[67]:             tweet_id           timestamp  source  rating_numerator  \
         0  892420643555336193 2017-08-01 16:23:56  iPhone                13
         1  892177421306343426 2017-08-01 00:17:27  iPhone                13
         2  891815181378084864 2017-07-31 00:18:03  iPhone                12
         3  891689557279858688 2017-07-30 15:58:51  iPhone                13
         4  891327558926688256 2017-07-29 16:00:24  iPhone                12

            rating_denominator      name dog_class  \
0                  10   Phineas       NaN
1                  10     Tilly       NaN
2                  10    Archie       NaN
3                  10     Darla       NaN
4                  10  Franklin       NaN

                                         jpg_url_x  img_num         p1_x  \
0  https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg      1.0       orange
1  https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg      1.0    Chihuahua
2  https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg      1.0    Chihuahua
3  https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg      1.0  paper_towel
4  https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg      2.0       basset

    ...                                         jpg_url_y         p1_y  \
0   ...     https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg       orange
1   ...     https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg    Chihuahua
2   ...     https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg    Chihuahua
3   ...     https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg  paper_towel
4   ...     https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg       basset

  p1_conf_y  p1_dog_y                p2_y  p2_conf_y  p2_dog_y  \
0  0.097049     False               bagel   0.085851     False
1  0.323581      True            Pekinese   0.090647      True
2  0.716012      True            malamute   0.078253      True
3  0.170278     False  Labrador_retriever   0.168086      True
4  0.555712      True    English_springer   0.225770      True

                         p3_y  p3_conf_y p3_dog_y
0                      banana   0.076110    False
1                    papillon   0.068957     True
2                      kelpie   0.031379     True
3                     spatula   0.040836    False
4  German_short-haired_pointer   0.175219     True

[5 rows x 28 columns]
```

```
In [80]: twitter_archive_cpy.drop('jpg_url_x', axis=1, inplace=True)
```

```
        ---------------------------------------------------------------------------

        KeyError                                  Traceback (most recent call last)

        <ipython-input-80-3deb18a00f8d> in <module>()
    ----> 1 twitter_archive_cpy.drop('jpg_url_x', axis=1, inplace=True)


        /opt/conda/lib/python3.6/site-packages/pandas/core/frame.py in drop(self, labels, axis,
    3695                                               index=index, columns=columns,
    3696                                               level=level, inplace=inplace,
 -> 3697                                               errors=errors)
    3698
    3699     @rewrite_axis_style_signature('mapper', [('copy', True),


     /opt/conda/lib/python3.6/site-packages/pandas/core/generic.py in drop(self, labels, axis
    3109         for axis, labels in axes.items():
    3110             if labels is not None:
 -> 3111                 obj = obj._drop_axis(labels, axis, level=level, errors=errors)
    3112
    3113         if inplace:


     /opt/conda/lib/python3.6/site-packages/pandas/core/generic.py in _drop_axis(self, labels
    3141                 new_axis = axis.drop(labels, level=level, errors=errors)
    3142             else:
 -> 3143                 new_axis = axis.drop(labels, errors=errors)
    3144             result = self.reindex(**{axis_name: new_axis})
    3145


     /opt/conda/lib/python3.6/site-packages/pandas/core/indexes/base.py in drop(self, labels,
    4402             if errors != 'ignore':
    4403                 raise KeyError(
 -> 4404                     '{} not found in axis'.format(labels[mask]))
    4405             indexer = indexer[~mask]
    4406         return self.delete(indexer)


     KeyError: "['jpg_url_x'] not found in axis"
```

```
In [81]: twitter_archive_cpy.head(2)
```

```
Out[81]:              tweet_id           timestamp  source  rating_numerator     name  \
         0  892420643555336193 2017-08-01 16:23:56  iPhone                13  Phineas
         1  892177421306343426 2017-08-01 00:17:27  iPhone                13    Tilly

           dog_class  img_num       p1_x  p1_conf_x p1_dog_x   ...     \
         0       NaN      1.0     orange   0.097049    False   ...
         1       NaN      1.0  Chihuahua   0.323581     True   ...

                                           jpg_url_y       p1_y p1_conf_y  \
         0  https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg     orange  0.097049
         1  https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg  Chihuahua  0.323581
```

```
  p1_dog_y      p2_y  p2_conf_y p2_dog_y      p3_y  p3_conf_y p3_dog_y
0    False     bagel  0.085851    False    banana   0.076110    False
1     True  Pekinese  0.090647     True  papillon   0.068957     True

[2 rows x 26 columns]
```

```
In [77]: twitter_archive_cpy.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]: twitter_archive_cpy.drop('rating_denominator', axis=1, inplace=True)
```

```
In [79]: twitter_archive_cpy.head(2)
```

```
Out[79]:             tweet_id           timestamp  source  rating_numerator     name  \
         0  892420643555336193 2017-08-01 16:23:56  iPhone                13  Phineas
         1  892177421306343426 2017-08-01 00:17:27  iPhone                13    Tilly

           dog_class  img_num       p1_x  p1_conf_x p1_dog_x   ...     \
         0       NaN      1.0     orange   0.097049    False   ...
         1       NaN      1.0  Chihuahua   0.323581     True   ...

                                                 jpg_url_y       p1_y p1_conf_y  \
         0  https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg     orange  0.097049
         1  https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg  Chihuahua  0.323581

           p1_dog_y      p2_y  p2_conf_y p2_dog_y      p3_y  p3_conf_y p3_dog_y
         0    False     bagel  0.085851    False    banana   0.076110    False
         1     True  Pekinese  0.090647     True  papillon   0.068957     True

         [2 rows x 26 columns]
```

```
In [82]: twitter_archive_cpy.drop('jpg_url_y', axis=1, inplace=True)
```

```
In [83]: twitter_archive_cpy.head(2)
```

```
Out[83]:              tweet_id           timestamp  source  rating_numerator     name  \
         0  892420643555336193 2017-08-01 16:23:56  iPhone                13  Phineas
         1  892177421306343426 2017-08-01 00:17:27  iPhone                13    Tilly

           dog_class  img_num       p1_x  p1_conf_x p1_dog_x   ...    p3_dog_x  \
         0       NaN      1.0     orange   0.097049    False   ...       False
         1       NaN      1.0  Chihuahua   0.323581     True   ...        True

                 p1_y p1_conf_y p1_dog_y      p2_y p2_conf_y p2_dog_y      p3_y  \
         0     orange  0.097049    False     bagel  0.085851    False    banana
         1  Chihuahua  0.323581     True  Pekinese  0.090647     True  papillon

           p3_conf_y p3_dog_y
         0  0.076110    False
         1  0.068957     True

         [2 rows x 25 columns]
```

```
In [84]: twitter_archive_cpy.drop('img_num', axis=1, inplace=True)
```

```
In [85]: twitter_archive_cpy.head(2)
```

```
Out[85]:              tweet_id           timestamp  source  rating_numerator     name  \
         0  892420643555336193 2017-08-01 16:23:56  iPhone                13  Phineas
         1  892177421306343426 2017-08-01 00:17:27  iPhone                13    Tilly

           dog_class       p1_x  p1_conf_x p1_dog_x      p2_x    ...     p3_dog_x  \
         0       NaN     orange   0.097049    False     bagel    ...        False
         1       NaN  Chihuahua   0.323581     True  Pekinese    ...         True

                 p1_y p1_conf_y  p1_dog_y      p2_y p2_conf_y  p2_dog_y      p3_y  \
         0     orange  0.097049     False     bagel  0.085851     False    banana
         1  Chihuahua  0.323581      True  Pekinese  0.090647      True  papillon

           p3_conf_y  p3_dog_y
         0  0.076110     False
         1  0.068957      True

         [2 rows x 24 columns]
```

0.0.4 Analyzing Data

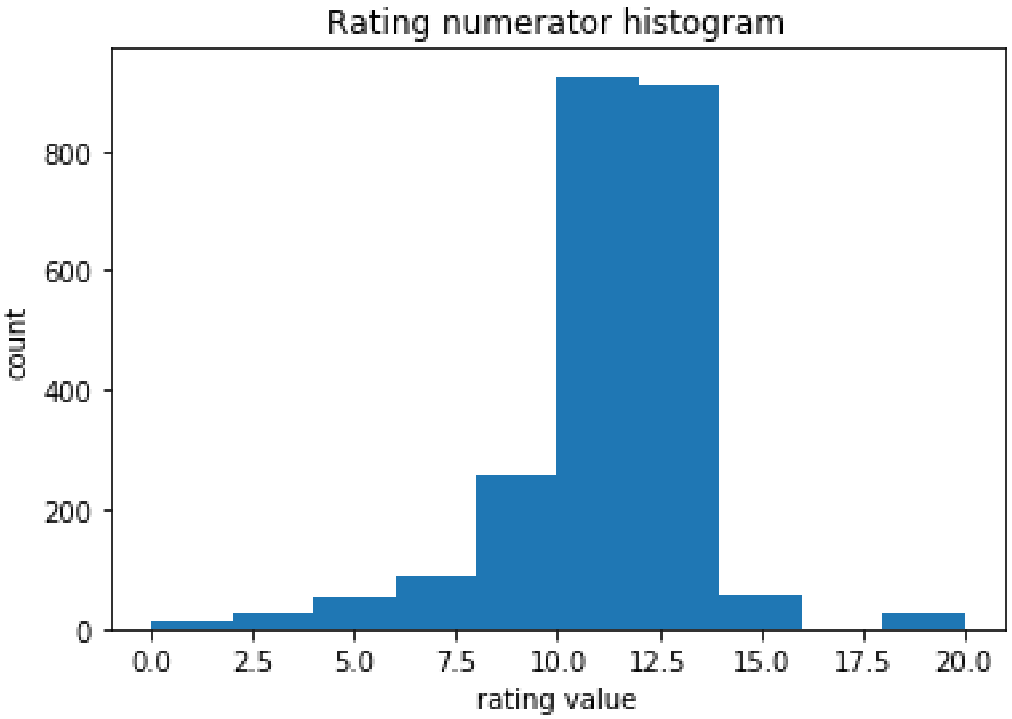
```
In [86]: twitter_archive_cpy.describe()
```

```
Out[86]:            tweet_id  rating_numerator    p1_conf_x     p2_conf_x  \
         count  2.356000e+03       2356.000000  2075.000000  2.075000e+03
```

```
mean   7.427716e+17         10.792869     0.594548  1.345886e-01
std    6.856705e+16          2.383856     0.271174  1.006657e-01
min    6.660209e+17          0.000000     0.044333  1.011300e-08
25%    6.783989e+17         10.000000     0.364412  5.388625e-02
50%    7.196279e+17         11.000000     0.588230  1.181810e-01
75%    7.993373e+17         12.000000     0.843855  1.955655e-01
max    8.924206e+17         20.000000     1.000000  4.880140e-01

          p3_conf_x    p1_conf_y     p2_conf_y     p3_conf_y
count  2.075000e+03  2075.000000  2.075000e+03  2.075000e+03
mean   6.032417e-02     0.594548  1.345886e-01  6.032417e-02
std    5.090593e-02     0.271174  1.006657e-01  5.090593e-02
min    1.740170e-10     0.044333  1.011300e-08  1.740170e-10
25%    1.622240e-02     0.364412  5.388625e-02  1.622240e-02
50%    4.944380e-02     0.588230  1.181810e-01  4.944380e-02
75%    9.180755e-02     0.843855  1.955655e-01  9.180755e-02
max    2.734190e-01     1.000000  4.880140e-01  2.734190e-01
```

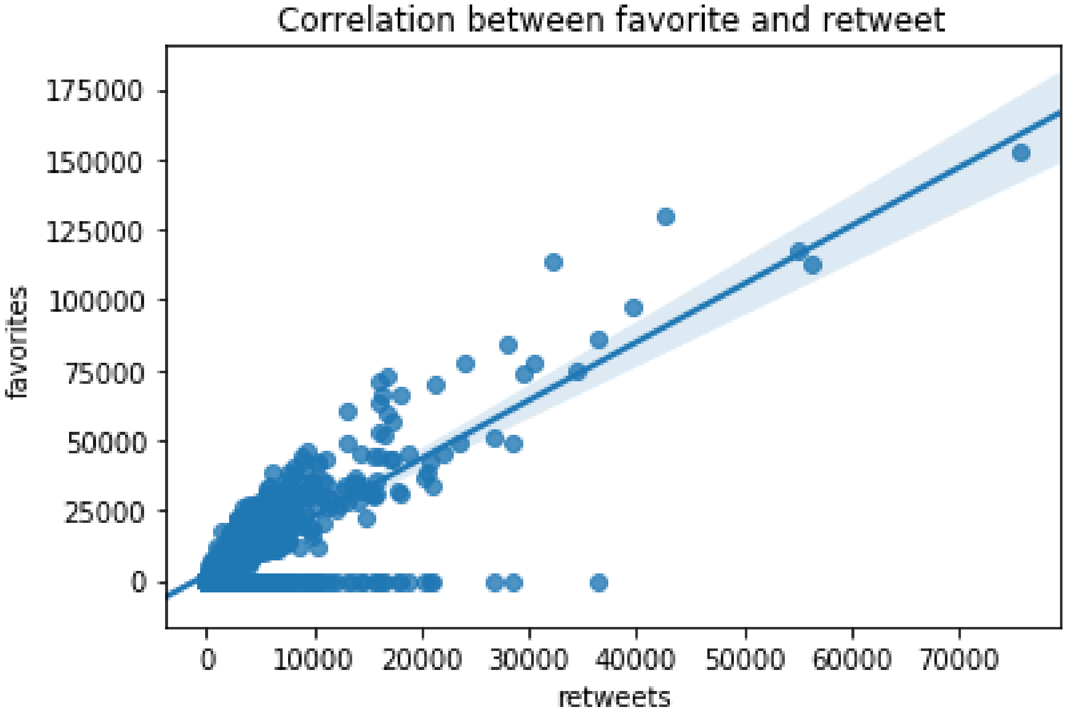
```
In [87]: fig, ax = plt.subplots()
         plt.hist(twitter_archive_cpy.rating_numerator);
         plt.title('Rating numerator histogram');
         ax.set_ylabel('count');
         ax.set_xlabel('rating value');
```



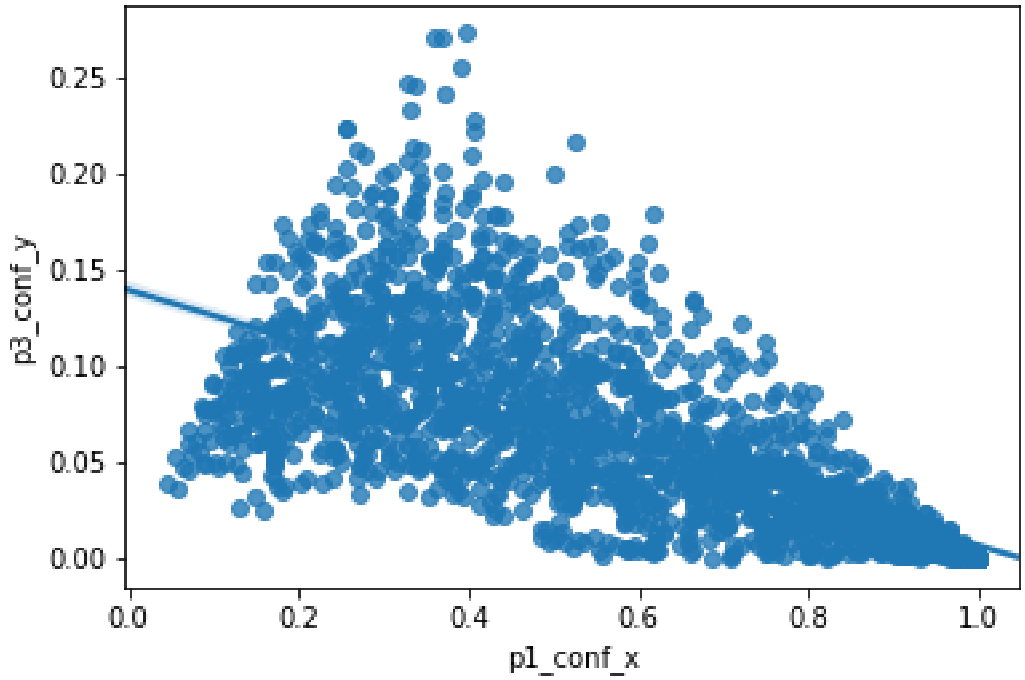
```
In [91]: tweets_df_cpy.corr()

Out[91]:            retweets  favorites
         retweets   1.000000   0.801637
         favorites  0.801637   1.000000

In [97]: sns.regplot(tweets_df_cpy.retweets, tweets_df_cpy.favorites);
         plt.title('Correlation between favorite and retweet');
```



```
In [100]: sns.regplot(data=twitter_archive_cpy, x="p1_conf_x", y='p3_conf_y');
```



```
In [101]: twitter_archive_cpy.corr()

Out[101]:                   tweet_id  rating_numerator  p1_conf_x  p2_conf_x  p3_conf_x  \
          tweet_id          1.000000          0.477239   0.101821   0.002012  -0.043424
          rating_numerator  0.477239          1.000000   0.088171  -0.009263  -0.017357
          p1_conf_x         0.101821          0.088171   1.000000  -0.511298  -0.709449
          p2_conf_x         0.002012         -0.009263  -0.511298   1.000000   0.479027
          p3_conf_x        -0.043424         -0.017357  -0.709449   0.479027   1.000000
          p1_conf_y         0.101821          0.088171   1.000000  -0.511298  -0.709449
          p2_conf_y         0.002012         -0.009263  -0.511298   1.000000   0.479027
          p3_conf_y        -0.043424         -0.017357  -0.709449   0.479027   1.000000

                            p1_conf_y  p2_conf_y  p3_conf_y
          tweet_id           0.101821   0.002012  -0.043424
          rating_numerator   0.088171  -0.009263  -0.017357
          p1_conf_x          1.000000  -0.511298  -0.709449
          p2_conf_x         -0.511298   1.000000   0.479027
          p3_conf_x         -0.709449   0.479027   1.000000
          p1_conf_y          1.000000  -0.511298  -0.709449
          p2_conf_y         -0.511298   1.000000   0.479027
          p3_conf_y         -0.709449   0.479027   1.000000

In [116]: twitter_archive_cpy['name'].describe()

Out[116]: count        1541
          unique        953
```

```
         top       Charlie
         freq           12
         Name: name, dtype: object
```

```
In [117]: twitter_archive_cpy.dog_class.describe()
```

```
Out[117]: count        380
          unique         4
          top       pupper
          freq         257
          Name: dog_class, dtype: object
```

```
In [ ]:
```