wrangle_act

July 2, 2019

1 Data wrangling project

1.0.1 Context

Wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

[1]:
```
import pandas as pd
import numpy as np
import twitter_credetials as twc
import requests
import tweepy
import json
import matplotlib.pyplot as plt

%matplotlib inline
```

1.1 Gathering Data

[2]:
```
twitter = pd.read_csv('twitter-archive-enhanced.csv')
```

[3]:
```
url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/
 ↪599fd2ad_image-predictions/image-predictions.tsv'

response = requests.get(url)
with open('image_predictions.tsv', mode='wb') as file:
        file.write(response.content)
```

[4]:
```
image_prd = pd.read_csv('image_predictions.tsv', sep='\t')
```

[5]:
```
auth = tweepy.OAuthHandler(twc.consumer_key, twc.consumer_secret)
auth.set_access_token(twc.access_token, twc.access_token_secret)
api = tweepy.API(auth)
```

[12]:
```
with open('tweet_json.txt', 'w') as file:
    for tweet_id in twitter['tweet_id']:
        try:
            json.dump(api.get_status(tweet_id, wait_on_rate_limit=True)._json,␣
↪file)
            file.write('\n')
        except tweepy.TweepError as e:
            print('{} : {}'.format(e.args[0][0]['message'], tweet_id))
```

```
No status found with that ID. : 888202515573088257
No status found with that ID. : 873697596434513921
No status found with that ID. : 872668790621863937
No status found with that ID. : 872261713294495745
No status found with that ID. : 869988702071779329
No status found with that ID. : 866816280283807744
No status found with that ID. : 861769973181624320
No status found with that ID. : 856602993587888130
No status found with that ID. : 851953902622658560
No status found with that ID. : 845459076796616705
No status found with that ID. : 844704788403113984
No status found with that ID. : 842892208864923648
No status found with that ID. : 837366284874571778
No status found with that ID. : 837012587749474308
No status found with that ID. : 829374341691346946
No status found with that ID. : 827228250799742977
No status found with that ID. : 812747805718642688
No status found with that ID. : 802247111496568832
No status found with that ID. : 775096608509886464
No status found with that ID. : 770743923962707968
No status found with that ID. : 754011816964026368
No status found with that ID. : 680055455951884288
```

[6]:

```
df = []
with open('tweet_json.txt', 'r') as file:
    for line in file:
        tweet_dict = dict(json.loads(line))
        tweet_id = tweet_dict['id']
        favorite = tweet_dict['favorite_count']
        retweet = tweet_dict['retweet_count']

        df.append({'tweet_id': tweet_id,
                   'favorite_count': favorite,
                   'retweet_count': retweet})

twitter_plus = pd.DataFrame(df, columns = ['tweet_id', 'favorite_count',␣
↪'retweet_count'])
```

1.2 Assessing Data

[7]:
```
twitter
```

[7]:
```
                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]
```

[8]:
```
image_prd
```

[8]:
```
                 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]
```

[9]:

```
twitter_plus
```

[9]:

```
                tweet_id  favorite_count  retweet_count
0     892420643555336193           37414           8129
1     892177421306343426           32152           6025
```

2	891815181378084864	24248	3988
3	891689557279858688	40724	8294
4	891327558926688256	38952	8983
5	891087950875897856	19566	2984
6	890971913173991426	11432	1977
7	890729181411237888	63087	18092
8	890609185150312448	26921	4093
9	890240255349198849	30862	7071
10	890006608113172480	29673	7021
11	889880896479866881	26920	4778
12	889665388333682689	46493	9611
13	889638837579907072	26200	4340
14	889531135344209921	14625	2162
15	889278841981685760	24411	5155
16	888917238123831296	28151	4325
17	888804989199671297	24732	4104
18	888554962724278272	19168	3384
19	888078434458587136	21043	3337
20	887705289381826560	29197	5158
21	887517139158093824	44791	11221
22	887473957103951883	66613	17391
23	887343217045368832	32567	10000
24	887101392804085760	29575	5714
25	886983233522544640	33945	7422
26	886736880519319552	11627	3126
27	886680336477933568	21696	4280
28	886366144734445568	20471	3064
29	886267009285017600	116	4
...	...	...	...
2304	666411507551481857	426	319
2305	666407126856765440	103	37
2306	666396247373291520	161	81
2307	666373753744588802	179	87
2308	666362758909284353	753	552
2309	666353288456101888	211	71
2310	666345417576210432	282	131
2311	666337882303524864	190	88
2312	666293911632134144	484	341
2313	666287406224695296	141	63
2314	666273097616637952	167	75
2315	666268910803644416	99	34
2316	666104133288665088	13879	6281
2317	666102155909144576	75	11
2318	666099513787052032	148	64
2319	666094000022159362	158	69
2320	666082916733198337	110	43
2321	666073100786774016	311	155

```
2322  666071193221509120          140          55
2323  666063827256086533          458         209
2324  666058600524156928          108          56
2325  666057090499244032          282         135
2326  666055525042405380          422         230
2327  666051853826850816         1184         822
2328  666050758794694657          129          56
2329  666049248165822465          103          41
2330  666044226329800704          287         135
2331  666033412701032449          120          43
2332  666029285002620928          124          46
2333  666020888022790149         2491         489

[2334 rows x 3 columns]
```

[10]:

```
twitter.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
```

[11]:

```
image_prd.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
```

[12]:
```
twitter_plus.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2334 entries, 0 to 2333
Data columns (total 3 columns):
tweet_id          2334 non-null int64
favorite_count    2334 non-null int64
retweet_count     2334 non-null int64
dtypes: int64(3)
memory usage: 54.8 KB
```

[13]:
```
twitter.duplicated().sum(), image_prd.duplicated().sum(), twitter_plus.
 ↪duplicated().sum()
```

[13]:
```
(0, 0, 0)
```

[14]:
```
image_prd['p1'].sample(20)
```

[14]:
```
1912           miniature_pinscher
1903    Staffordshire_bullterrier
578                      malamute
1394                         chow
1265      wire-haired_fox_terrier
1496           Norwegian_elkhound
1311                       beagle
871                    Great_Dane
994                French_bulldog
1019                     Pekinese
1134                 window_shade
1520                          pug
1532                         chow
391                         teddy
1647                    seat_belt
832                     washbasin
741                      Shih-Tzu
460               giant_schnauzer
```

```
50                      triceratops
1935                 French_bulldog
Name: p1, dtype: object
```

[15]:
```
twitter.shape[0], image_prd.shape[0], twitter_plus.shape[0]
```

[15]:
```
(2356, 2075, 2334)
```

1.2.1 Quality

twitter - Missing some `expanded_urls`. - Many names aren't real names, sometimes is just a letter or syllable. - Data type is wrong in `timestamp` column, should be date type. - Discard retweets because we are interested just in original ratings. - Discard `source`, `in_reply_to_status_id`, `in_reply_to_user_id` columns, because isn't necessary for this analysis.

image_prd - Standardize `p1`,`p2`,`p3` columns to lowercase. - Use underline as the standard separator.

twitter_plus - Missing data, **twitter** has 2356 and **twitter_plus** has 2334.

1.2.2 Tidiness

- Join the dataframe **twitter** with **twitter_plus**.

twitter - Convert the four columns `doggo`,`floofer`,`pupper` and `puppo` in just one called `stage`.

1.3 Cleaning Data

[16]:
```
twitter_clean = twitter.copy()
image_prd_clean = image_prd.copy()
twitter_plus_clean = twitter_plus.copy()
```

1.3.1 Tidiness

Join the dataframe **twitter** with **twitter_plus**

Define

Using the `merge` method to handle **twitter** and **twitter_plus**

Code

[17]:
```
twitter_clean = pd.merge(twitter_clean, twitter_plus_clean, on='tweet_id',␣
 ↪how='right')
```

Test

[18]:
```
twitter_clean.info(), twitter_clean.shape
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2334 entries, 0 to 2333
Data columns (total 19 columns):
tweet_id                      2334 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
```

```
timestamp                     2334 non-null object
source                        2334 non-null object
text                          2334 non-null object
retweeted_status_id           165 non-null float64
retweeted_status_user_id      165 non-null float64
retweeted_status_timestamp    165 non-null object
expanded_urls                 2275 non-null object
rating_numerator              2334 non-null int64
rating_denominator            2334 non-null int64
name                          2334 non-null object
doggo                         2334 non-null object
floofer                       2334 non-null object
pupper                        2334 non-null object
puppo                         2334 non-null object
favorite_count                2334 non-null int64
retweet_count                 2334 non-null int64
dtypes: float64(4), int64(5), object(10)
memory usage: 364.7+ KB
```

[18]:
```
(None, (2334, 19))
```

[19]:
```
twitter_clean.head(2)
```

[19]:
```
             tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
0  892420643555336193                    NaN                  NaN
1  892177421306343426                    NaN                  NaN

                   timestamp  \
0  2017-08-01 16:23:56 +0000
1  2017-08-01 00:17:27 +0000

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

   retweeted_status_user_id retweeted_status_timestamp  \
0                       NaN                        NaN
1                       NaN                        NaN

                                       expanded_urls  rating_numerator  \
0  https://twitter.com/dog_rates/status/892420643...                13
1  https://twitter.com/dog_rates/status/892177421...                13

   rating_denominator     name doggo floofer pupper puppo  favorite_count  \
0                  10  Phineas  None     None    None  None             37414
1                  10    Tilly  None     None    None  None             32152

   retweet_count
0           8129
1           6025
```

[20]:
```python
twitter_clean.tail(2)
```

[20]:
```
                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
2332  666029285002620928                    NaN                  NaN
2333  666020888022790149                    NaN                  NaN

                      timestamp  \
2332  2015-11-15 23:05:30 +0000
2333  2015-11-15 22:32:08 +0000

                                                 source  \
2332  <a href="http://twitter.com/download/iphone" r...
2333  <a href="http://twitter.com/download/iphone" r...

                                                   text  retweeted_status_id  \
2332  This is a western brown Mitsubishi terrier. Up...                  NaN
2333  Here we have a Japanese Irish Setter. Lost eye...                  NaN

      retweeted_status_user_id retweeted_status_timestamp  \
2332                       NaN                        NaN
2333                       NaN                        NaN

                                          expanded_urls  rating_numerator  \
2332  https://twitter.com/dog_rates/status/666029285...                 7
2333  https://twitter.com/dog_rates/status/666020888...                 8

      rating_denominator  name doggo floofer pupper puppo  favorite_count  \
2332                  10     a  None     None    None  None             124
2333                  10  None  None     None    None  None            2491

      retweet_count
2332             46
2333            489
```

Convert the four columns doggo, `floofer`, pupper and puppo in just one called stage

Define

Using the `melt` method to join columns in one column called stage and after putting in order, drop duplicate data

Code

[21]:
```python
twitter_clean = pd.melt(twitter_clean, id_vars=['tweet_id',
                                                 'in_reply_to_status_id',
                                                 'in_reply_to_user_id',
                                                 'timestamp',
                                                 'source',
                                                 'text',
                                                 'retweeted_status_id',
                                                 'retweeted_status_user_id',
                                                 'retweeted_status_timestamp',
                                                 'expanded_urls',
                                                 'rating_numerator',
                                                 'rating_denominator',
                                                 'name',
                                                 'favorite_count',
                                                 'retweet_count'],␣
  ↪value_vars=['doggo','floofer','pupper', 'puppo'],␣
  ↪var_name='tab',value_name='stage')
 twitter_clean = twitter_clean.drop('tab', axis=1)
```

[22]:
```
twitter_clean = twitter_clean.sort_values('stage').
 ↪drop_duplicates(subset='tweet_id', keep='last')
```

[23]:
```
twitter_clean = twitter_clean.sort_values('tweet_id', ascending=False).
 ↪reset_index(drop=True)
```

Test

[24]:
```
twitter_clean['stage'].value_counts()
```

[24]:
```
None       1958
pupper      256
doggo        80
puppo        30
floofer      10
Name: stage, dtype: int64
```

[25]:
```
twitter_clean['stage'].sample(20)
```

[25]:
```
1653       None
2128       None
1007       None
1371       None
1724       None
2198       None
1825       None
231       doggo
474        None
1614       None
2227       None
353        None
722        None
74         None
1934     pupper
```

```
195       None
1444      None
1854      None
10        None
119       None
Name: stage, dtype: object
```

[26]:

```
twitter_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2334 entries, 0 to 2333
Data columns (total 16 columns):
tweet_id                      2334 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2334 non-null object
source                        2334 non-null object
text                          2334 non-null object
retweeted_status_id           165 non-null float64
retweeted_status_user_id      165 non-null float64
retweeted_status_timestamp    165 non-null object
expanded_urls                 2275 non-null object
rating_numerator              2334 non-null int64
rating_denominator            2334 non-null int64
name                          2334 non-null object
favorite_count                2334 non-null int64
retweet_count                 2334 non-null int64
stage                         2334 non-null object
dtypes: float64(4), int64(5), object(7)
memory usage: 291.8+ KB
```

1.3.2 Quality

1.3.3 twitter

Data type is wrong in `timestamp` column, should be date type.

Define

Using the known method `to_datetime` to convert `timestamp` column.

Code

[27]:

```
twitter_clean['timestamp'] = pd.to_datetime(twitter_clean['timestamp'])
```

Test

[28]:

```
twitter_clean['timestamp'].dtype
```

[28]:

```
datetime64[ns, UTC]
```

Missing some `expanded_urls`

Define

Replace all `expanded_urls` using the standard url with `tweet_id` column.

Code

```
[29]: twitter_clean['expanded_urls'] = ['https://twitter.com/dog_rates/status/{}/
       ↪photo/1'.format(tweet_id) for tweet_id in twitter_clean['tweet_id']]
```

Test

```
[30]: pd.set_option('display.max_colwidth', -1)
      twitter_clean['expanded_urls'].head(), twitter_clean['expanded_urls'].tail()
```

```
[30]: (0    https://twitter.com/dog_rates/status/892420643555336193/photo/1
       1    https://twitter.com/dog_rates/status/892177421306343426/photo/1
       2    https://twitter.com/dog_rates/status/891815181378084864/photo/1
       3    https://twitter.com/dog_rates/status/891689557279858688/photo/1
       4    https://twitter.com/dog_rates/status/891327558926688256/photo/1
       Name: expanded_urls, dtype: object,
       2329    https://twitter.com/dog_rates/status/666049248165822465/photo/1
       2330    https://twitter.com/dog_rates/status/666044226329800704/photo/1
       2331    https://twitter.com/dog_rates/status/666033412701032449/photo/1
       2332    https://twitter.com/dog_rates/status/666029285002620928/photo/1
       2333    https://twitter.com/dog_rates/status/666020888022790149/photo/1
       Name: expanded_urls, dtype: object)
```

```
[31]: twitter['expanded_urls'].isnull().sum(), twitter_clean['expanded_urls'].
       ↪isnull().sum()
```

```
[31]: (59, 0)
```

Many names aren't real names, sometimes is just a letter or syllable.

Define

Replace all names that are lowercase to 'None', because in assessing data we can look that mostly of wrong names are lowercase.

Code

```
[32]: twitter_clean['name'] = ['None' if name.islower() else name for name in␣
       ↪twitter_clean['name']]
```

Test

```
[33]: twitter_clean['name'].head(), twitter_clean['name'].tail()
```

```
[33]: (0    Phineas
       1    Tilly
       2    Archie
       3    Darla
       4    Franklin
       Name: name, dtype: object, 2329    None
       2330    None
       2331    None
       2332    None
       2333    None
       Name: name, dtype: object)
```

```
[34]: twitter_clean['name'].value_counts()[:30]
```

[34]:
```
None        846
Charlie      11
Oliver       11
Cooper       11
Tucker       10
Lucy         10
Penny        10
Lola         10
Bo            9
Winston       9
Sadie         8
Toby          7
Buddy         7
Daisy         7
Bailey        7
Bella         6
Rusty         6
Oscar         6
Leo           6
Jack          6
Scout         6
Jax           6
Stanley       6
Dave          6
Milo          6
Koda          6
Bentley       5
Finn          5
Larry         5
Chester       5
Name: name, dtype: int64
```

Discard retweets because we are interested just in original ratings.

Define

Capture all row that retweeted is null, then we have just original tweets

Code

[35]:
```
twitter_clean = twitter_clean[twitter_clean['retweeted_status_id'].isnull()]
```

Test

[36]:
```
twitter_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2169 entries, 0 to 2333
Data columns (total 16 columns):
tweet_id                      2169 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2169 non-null datetime64[ns, UTC]
```

```
source                        2169 non-null object
text                          2169 non-null object
retweeted_status_id           0 non-null float64
retweeted_status_user_id      0 non-null float64
retweeted_status_timestamp    0 non-null object
expanded_urls                 2169 non-null object
rating_numerator              2169 non-null int64
rating_denominator            2169 non-null int64
name                          2169 non-null object
favorite_count                2169 non-null int64
retweet_count                 2169 non-null int64
stage                         2169 non-null object
dtypes: datetime64[ns, UTC](1), float64(4), int64(5), object(6)
memory usage: 288.1+ KB
```

Discard source, in_reply_to_status_id, in_reply_to_user_id columns, because isn't necessary for this analysis.

Define

Just drop the useless columns

code

[37]:

```
bad_columns = ['in_reply_to_status_id',
               'in_reply_to_user_id',
               'source', 'retweeted_status_id',
               'retweeted_status_user_id',
               'retweeted_status_timestamp']

twitter_clean = twitter_clean.drop(bad_columns, axis=1)
```

Test

[38]:

```
twitter_clean.columns
```

[38]:

```
Index(['tweet_id', 'timestamp', 'text', 'expanded_urls', 'rating_numerator',
       'rating_denominator', 'name', 'favorite_count', 'retweet_count',
       'stage'],
      dtype='object')
```

1.3.4 image_prd

Standardize p1,p2,p3 columns to lowercase.

Define

Transform all data in p1, p2 and p3 to lowercase **Code**

[39]:

```
image_prd_clean['p1'] = image_prd_clean['p1'].str.lower()
image_prd_clean['p2'] = image_prd_clean['p2'].str.lower()
image_prd_clean['p3'] = image_prd_clean['p3'].str.lower()
```

Test

[40]:

```
image_prd_clean.head()
```

```
[40]:            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
```

```
[41]: image_prd_clean.tail()
```

```
[41]:               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
```

Use underline as the standard separator.

Define

Replace all '-' to underline

Code

[42]:

```
image_prd_clean['p1'] = image_prd_clean['p1'].str.replace('-', '_')
image_prd_clean['p2'] = image_prd_clean['p2'].str.replace('-', '_')
image_prd_clean['p3'] = image_prd_clean['p3'].str.replace('-', '_')
```

Test

[43]:

```
image_prd_clean.sample(20)
```

[43]:

```
                tweet_id  \
1508  785872687017132033
1302  752917284578922496
1118  725842289046749185
613   680145970311643136
1864  842892208864923648
1517  787322443945877504
770   689289219123089408
2025  881906580714921986
996   708349470027751425
344   672267570918129665
650   681981167097122816
566   678334497360859136
402   673697980713705472
1591  798665375516884993
1743  822859134160621569
1387  766078092750233600
986   707693576495472641
66    667176164155375616
1380  765222098633691136
698   684567543613382656

      jpg_url  \
1508  https://pbs.twimg.com/ext_tw_video_thumb/785872596088811520/pu/img/5O-_Bgq
dFQu_2Bt7.jpg
1302  https://pbs.twimg.com/media/CnLmRiYXEAAO_8f.jpg
1118  https://pbs.twimg.com/media/ChK1tdBWwAQ1flD.jpg
613   https://pbs.twimg.com/media/CXBdJxLUsAAWql2.jpg
1864  https://pbs.twimg.com/ext_tw_video_thumb/807106774843039744/pu/img/8XZg1xW
35Xp2J6JW.jpg
1517  https://pbs.twimg.com/media/Cu0hlfwWYAEdnXO.jpg
770   https://pbs.twimg.com/ext_tw_video_thumb/689289176076959744/pu/img/hEFkFtm
Mu_hkTlxK.jpg
2025  https://pbs.twimg.com/media/DD0pWm9XcAAeSBL.jpg
996   https://pbs.twimg.com/media/CdSQFWOWAAApgfq.jpg
344   https://pbs.twimg.com/media/CVRfyZxWUAAFIQR.jpg
650   https://pbs.twimg.com/media/CXbiQHmWcAAt6Lm.jpg
566   https://pbs.twimg.com/media/CWntoDVWcAEl3NB.jpg
402   https://pbs.twimg.com/media/CVl0vFeWoAAMTfg.jpg
1591  https://pbs.twimg.com/media/CVMOlMiWwAA4Yxl.jpg
1743  https://pbs.twimg.com/media/C2tiAzGXgAIFdqi.jpg
```

```
1387  https://pbs.twimg.com/media/ChK1tdBWwAQ1flD.jpg
986   https://pbs.twimg.com/media/CdI7jDnW0AA2dtO.jpg
66    https://pbs.twimg.com/media/CUJJLtWWsAE-go5.jpg
1380  https://pbs.twimg.com/media/Cp6db4-XYAAMmqL.jpg
698   https://pbs.twimg.com/media/CYASi6FWQAEQMW2.jpg

      img_num                           p1   p1_conf  p1_dog  \
1508  1        great_pyrenees               0.392108  True
1302  1        german_shepherd              0.609283  True
1118  1        toy_poodle                   0.420463  True
613   1        miniature_poodle             0.457117  True
1864  1        chihuahua                    0.505370  True
1517  1        seat_belt                    0.747739  False
770   1        snowmobile                   0.254642  False
2025  1        weimaraner                   0.291539  True
996   1        muzzle                       0.243890  False
344   1        irish_terrier                0.716932  True
650   1        labrador_retriever           0.452577  True
566   1        norfolk_terrier              0.378643  True
402   1        porcupine                    0.151876  False
1591  1        chow                         0.243529  True
1743  1        malinois                     0.332897  True
1387  1        toy_poodle                   0.420463  True
986   1        bathtub                      0.499525  False
66    1        soft_coated_wheaten_terrier  0.318981  True
1380  1        dalmatian                    0.556595  True
698   1        minibus                      0.401942  False

                            p2   p2_conf  p2_dog  \
1508  golden_retriever          0.198358  True
1302  malinois                  0.352460  True
1118  miniature_poodle          0.132640  True
613   toy_poodle                0.226481  True
1864  pomeranian                0.120358  True
1517  golden_retriever          0.105703  True
770   assault_rifle             0.129558  False
2025  chesapeake_bay_retriever  0.278966  True
996   basenji                   0.187158  True
344   miniature_pinscher        0.051234  True
650   golden_retriever          0.403420  True
566   golden_retriever          0.095594  True
402   hen                       0.111380  False
1591  hamster                   0.227150  False
1743  chihuahua                 0.104116  True
1387  miniature_poodle          0.132640  True
986   tub                       0.488014  False
66    lakeland_terrier          0.215218  True
```

```
1380  whippet                       0.151047  True
698   llama                         0.229145  False

                                        p3   p3_conf  p3_dog
1508  pekinese                            0.143328  True
1302  kelpie                              0.016105  True
1118  chesapeake_bay_retriever            0.121523  True
613   maltese_dog                         0.067682  True
1864  toy_terrier                         0.077008  True
1517  dingo                               0.017257  False
770   rifle                               0.110875  False
2025  koala                               0.127017  False
996   boston_bull                         0.092727  True
344   airedale                            0.044381  True
650   beagle                              0.069486  True
566   kelpie                              0.085309  True
402   doormat                             0.058934  False
1591  pomeranian                          0.056057  True
1743  staffordshire_bullterrier           0.047745  True
1387  chesapeake_bay_retriever            0.121523  True
986   washbasin                           0.009298  False
66    toy_poodle                          0.106014  True
1380  american_staffordshire_terrier      0.096435  True
698   seat_belt                           0.209393  False
```

1.4 Storing Data

[44]:

```
twitter_clean.to_csv('twitter_archive_master.csv', index=False)
image_prd_clean.to_csv('image_predictions_master.csv', index=False)
```

1.5 Analyzing and Visualizing Data

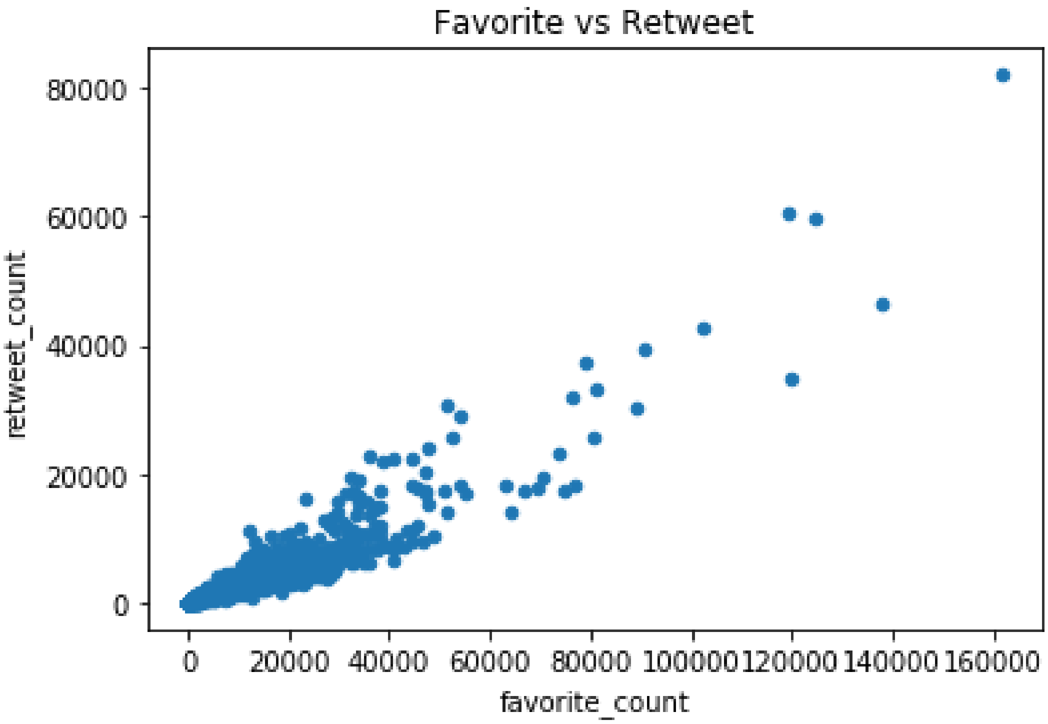
[2]:

```
# Read updated files
df_twitter = pd.read_csv('twitter_archive_master.csv')
df_img_predic = pd.read_csv('image_predictions_master.csv')
```

Visualizing
Has some correlation with the quantity of favorited and retweeted

[4]:

```
df_twitter.plot(kind='scatter', x='favorite_count', y='retweet_count',␣
 ↪title='Favorite vs Retweet');
```



Result: We can see that an ascending line, this mean when that post was retweeted more people liked.

Insight

Which dogs name is the most popular?

[5]:
```
df_twitter['name'].value_counts()[0:10]
```

[5]:
```
None        784
Cooper       10
Oliver       10
Charlie      10
Lucy         10
Tucker        9
Penny         9
Sadie         8
Winston       8
Lola          8
Name: name, dtype: int64
```

Result: In our data the most popular name is Cooper, but don't have great significance seen that has 784 None values.

Insight

Following the predictions, which are the most common dog breeds in WeRateDogs.

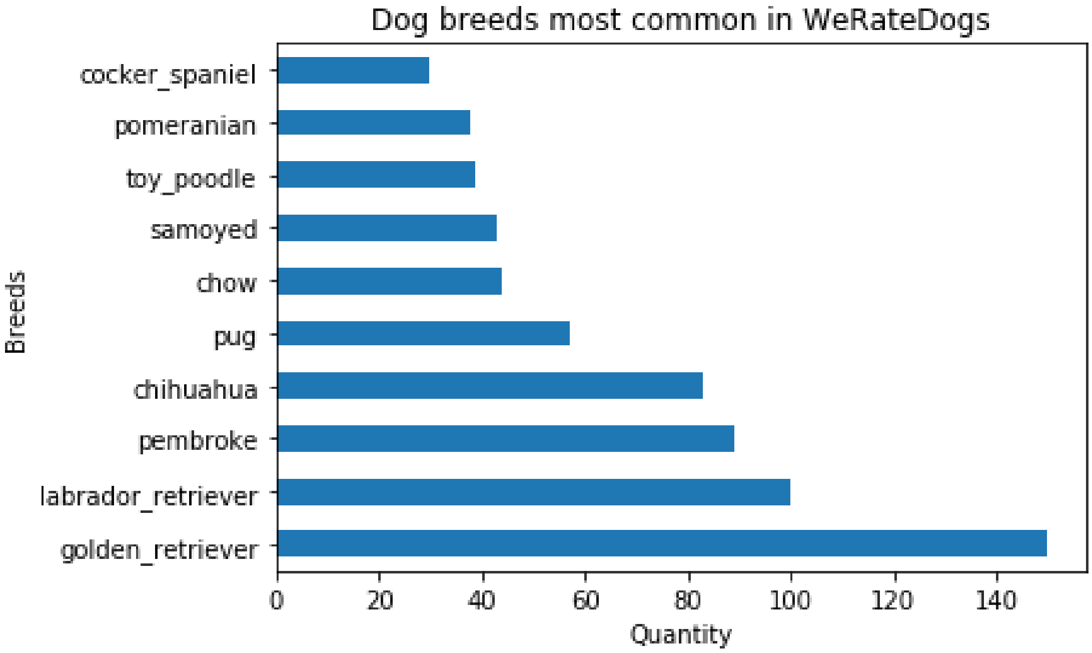
[6]:
```
def best_predict(row):
    indx = np.argmax([row['p1_conf'], row['p2_conf'], row['p3_conf']])
    dog = row['p{}_dog'.format(indx+1)]
    if dog:
        best = row['p{}'.format(indx+1)]
        return best
    else:
        return None
```

[7]:

```
df_twitter['breeds'] = df_img_predic.apply(best_predict, axis=1)
```

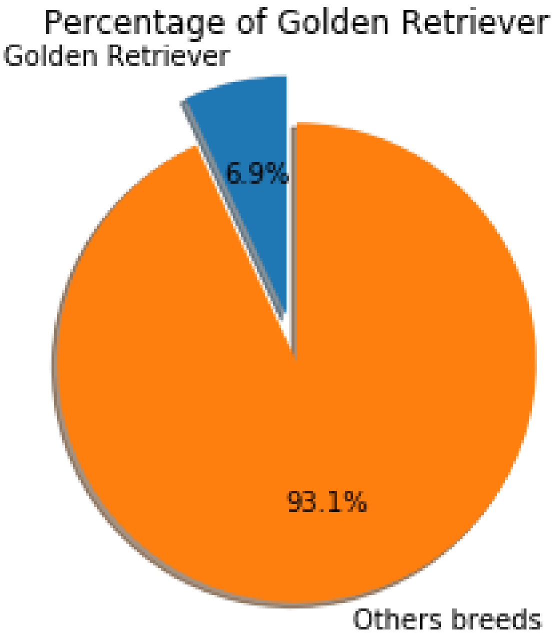
[8]:

```
df_twitter['breeds'].value_counts()[:10].plot(kind='barh')
plt.title('Dog breeds most common in WeRateDogs')
plt.xlabel('Quantity')
plt.ylabel('Breeds');
```



[9]:

```
golden = df_twitter[df_twitter['breeds']=='golden_retriever'].shape[0]
others = df_twitter.shape[0]-golden
sizes = [golden, others]
labels = ['Golden Retriever', 'Others breeds']

explode = (0.2, 0)
fig1, ax1 = plt.subplots()
ax1.pie(sizes, labels=labels,explode=explode, autopct='%1.1f%%', shadow=True,␣
 ↪startangle=90)
ax1.axis('equal')

plt.title('Percentage of Golden Retriever')
plt.show();
```



Result: We can take a look the most commons dog breeds are Golden Retriever appearing in 6.9% of breeds showing some advantage over others, Labrador Retriever, Pembroke, Chihuahua, and Pug.

Insight

Dog breeds less posted on WeRateDogs twitter, and possibly less favored or retweet?

[12]:
```
df_twitter['breeds'].value_counts().sort_values()[:10]
```

[12]:
```
scotch_terrier          1
entlebucher             1
clumber                 1
silky_terrier           1
groenendael             1
japanese_spaniel        1
standard_schnauzer      1
appenzeller             2
toy_terrier             2
sussex_spaniel          2
Name: breeds, dtype: int64
```

[49]:
```
df_twitter.groupby('breeds').min().
 ↪sort_values(['retweet_count','favorite_count'])[:10]
```

[49]:
```
                                        tweet_id                  timestamp  \
breeds
boxer                        672591762242805761  2015-12-04 01:42:26+00:00
dandie_dinmont               671186162933985280  2015-11-30 04:37:05+00:00
kuvasz                       668221241640230912  2015-11-22 00:15:33+00:00
```

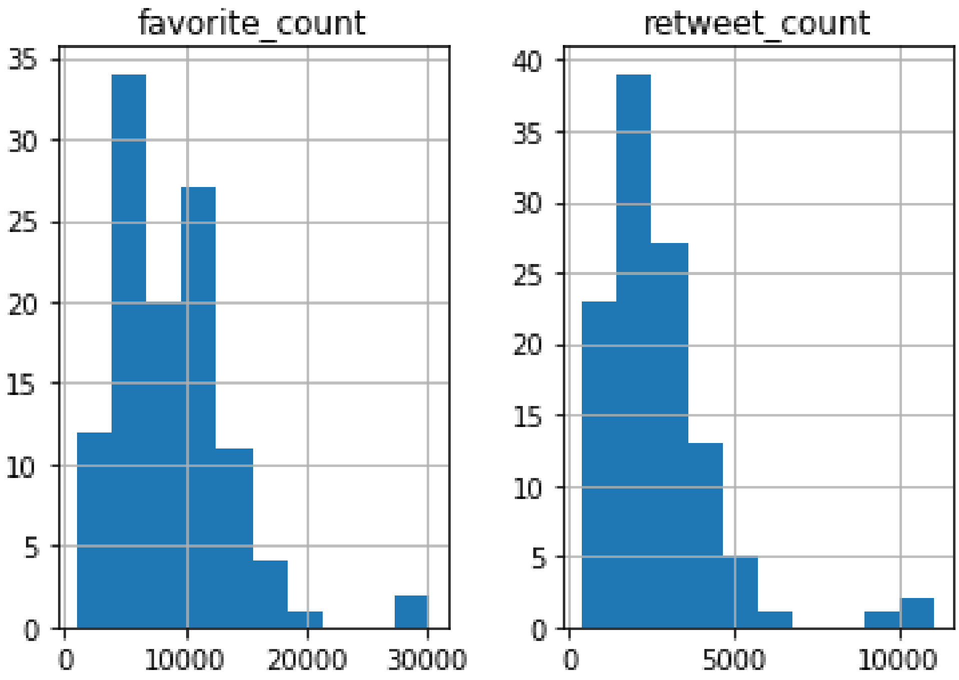
```
toy_poodle                   672997845381865473  2015-12-05 04:36:04+00:00
redbone                      687732144991551489  2016-01-14 20:24:55+00:00
pembroke                     667832474953625600  2015-11-20 22:30:44+00:00
kelpie                       669367896104181761  2015-11-25 04:11:57+00:00
saluki                       668623201287675904  2015-11-23 02:52:48+00:00
soft_coated_wheaten_terrier  707387676719185920  2016-03-09 02:08:59+00:00
standard_poodle              674638615994089473  2015-12-09 17:15:54+00:00

                                                                          text
\
breeds
boxer                        @serial @MrRoles OH MY GOD I listened to all o...
dandie_dinmont                                              @mount_alex3 13/10
kuvasz                                     13/10 such a good doggo\n@spaghemily
toy_poodle                                     @RealKentMurphy 14/10 confirmed
redbone                                           @Marc_IRL pixelated af 12/10
pembroke                     "Challenge completed" \n(pupgraded to 12/10) h...
kelpie                       @xianmcguire @Jenna_Marbles Kardashians wouldn...
saluki                                          12/10 good shit Bubka\n@wane15
soft_coated_wheaten_terrier  Meet Clarkus. He's a Skinny Eastern Worcesters...
standard_poodle              @dhmontgomery We also gave snoop dogg a 420/10...

                                                                 expanded_urls
\
breeds
boxer                        https://twitter.com/dog_rates/status/672591762...
dandie_dinmont               https://twitter.com/dog_rates/status/671186162...
kuvasz                       https://twitter.com/dog_rates/status/668221241...
toy_poodle                   https://twitter.com/dog_rates/status/672997845...
redbone                      https://twitter.com/dog_rates/status/687732144...
pembroke                     https://twitter.com/dog_rates/status/667832474...
kelpie                       https://twitter.com/dog_rates/status/669367896...
saluki                       https://twitter.com/dog_rates/status/668623201...
soft_coated_wheaten_terrier  https://twitter.com/dog_rates/status/707387676...
standard_poodle              https://twitter.com/dog_rates/status/674638615...

                             rating_numerator  rating_denominator       name  \
breeds
boxer                                       7                  10    Frankie
dandie_dinmont                             10                  10       Milo
kuvasz                                     10                  10    Bentley
toy_poodle                                  5                  10       Beya
redbone                                     6                  10      Bilbo
pembroke                                    6                  10  Alejandro
kelpie                                      8                  10    Charlie
saluki                                      4                  10   Jomathan
soft_coated_wheaten_terrier                 7                  10       Beau
```

```
standard_poodle                         4                  10     Chipson

                             favorite_count  retweet_count stage
breeds
boxer                                    51              2  None
dandie_dinmont                          112              6  None
kuvasz                                  265             10  None
toy_poodle                              302             10  None
redbone                                 227             17  None
pembroke                                242             17  None
kelpie                                  459             17  None
saluki                                  146             23  None
soft_coated_wheaten_terrier             247             25  None
standard_poodle                         344             26  None
```

[45]:
```
breeds_mean = df_twitter.groupby('breeds').mean().
 ↪sort_values(['retweet_count','favorite_count'], ascending=False)
breeds_mean[:10]
```

[45]:
```
                                tweet_id  rating_numerator  \
breeds
japanese_spaniel            7.887659e+17         12.000000
wire_haired_fox_terrier     6.877818e+17         10.500000
irish_terrier               7.716884e+17         11.500000
rhodesian_ridgeback         7.648635e+17         11.000000
schipperke                  7.239415e+17         10.300000
miniature_pinscher          7.414337e+17         10.608696
miniature_schnauzer         7.945474e+17         12.000000
basset                      7.307612e+17         10.384615
saint_bernard               7.444181e+17         10.428571
greater_swiss_mountain_dog  7.186492e+17          8.666667

                            rating_denominator  favorite_count  retweet_count
breeds
japanese_spaniel                     10.000000    28891.000000   11065.000000
wire_haired_fox_terrier              10.000000    17282.500000   10186.500000
irish_terrier                        10.000000    30215.500000    9402.166667
rhodesian_ridgeback                  10.000000    20126.000000    5962.250000
schipperke                           10.000000    11720.100000    5302.400000
miniature_pinscher                   10.000000    14006.217391    5137.565217
miniature_schnauzer                  10.000000    16391.250000    5103.000000
basset                               10.000000    12992.153846    4986.000000
saint_bernard                        10.142857    14118.857143    4940.142857
greater_swiss_mountain_dog           10.000000    13539.333333    4607.333333
```

[50]:
```
breeds_mean[['favorite_count','retweet_count']].hist();
```



Result: Well, fist a list of dog breeds less posted on the page, but after we look if had some relation with which has less favorite and retweet, but have no one, and looking at the bigger mean of retweet and favorite we see a dog breed from the list `japanese_spaniel`, but it's because has just one or two posts so the mean is elevated.