wrangle_act

December 4, 2023

1 Project: Wrangling and Analyze Data

1.1 Data Gathering

In the cell below, gather **all** three pieces of data for this project and load them in the notebook. **Note:** the methods required to gather each data are different. 1. Directly download the WeRate-Dogs Twitter archive data (twitter_archive_enhanced.csv)

```
In [1]: import pandas as pd
        import numpy as np
        import os
        import requests
        from PIL import Image
        from io import BytesIO
        import tweepy
        from tweepy import OAuthHandler
        import json
        from timeit import default_timer as timer
        import matplotlib.pyplot as plt
        import seaborn as sns

        %matplotlib inline
```

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

2. Use the Requests library to download the tweet image prediction (image_predictions.tsv)

```
In [3]: url ="https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-prediction
        r = requests.get(url)
        with open(url.split('/')[-1], mode = 'wb') as file:
            file.write(r.content)
```

```
In [4]: images_df = pd.read_csv('image-predictions.tsv', sep='\t')
```

3. Use the Tweepy library to query additional data via the Twitter API (tweet_json.txt)

```
In [5]: # Read the JSON file into a DataFrame
        tweets_df = pd.read_json('tweet-json.txt', lines=True)
```

```
# Extract relevant columns
tweets_df = tweets_df[['id', 'retweet_count', 'favorite_count']]

# Rename columns for clarity
tweets_df.columns = ['tweet_id', 'retweet_count', 'favorite_count']
```

1.2 Assessing Data

In this section, detect and document at least **eight (8) quality issues and two (2) tidiness issue**. You must use **both** visual assessment programmatic assessement to assess the data.

Note: pay attention to the following key points when you access the data.

- You only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
- Assessing and cleaning the entire dataset completely would require a lot of time, and is not necessary to practice and demonstrate your skills in data wrangling. Therefore, the requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.
- The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.
- You do not need to gather the tweets beyond August 1st, 2017. You can, but note that you won't be able to gather the image predictions for these tweets since you don't have access to the algorithm used.

1.2.1 Visual Assessment

```
In [6]: twitter_df
```

Out[6]:

	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]
```

1.2.2 twitter_df columns description:

- tweet_id: the unique identifier for each tweet
- in_reply_to_status_id: if the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's ID
- in_reply_to_user_id: if the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's author ID

- timestamp: time when this Tweet was created
- source: utility used to post the Tweet, as an HTML-formatted string. e.g. Twitter for Android, Twitter for iPhone, Twitter Web Client
- text: actual UTF-8 text of the status update
- retweeted_status_id: if the represented Tweet is a retweet, this field will contain the integer representation of the original Tweet's ID
- retweeted_status_user_id: if the represented Tweet is a retweet, this field will contain the integer representation of the original Tweet's author ID
- retweeted_status_timestamp: time of retweet
- expanded_urls: tweet URL
- rating_numerator: numerator of the rating of a dog. Note: ratings almost always greater than 10
- rating_denominator: denominator of the rating of a dog. Note: ratings almost always have a denominator of 10
- name: name of the dog
- doggo: dog stage
- floofer: dog stage
- pupper: dog stage
- puppo: dog stage

```
In [7]: images_df
```

Out[7]:

	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/CT9cxOtUEAAhNN_.jpg
25    666362758909284353    https://pbs.twimg.com/media/CT9lXGsUcAAyUFt.jpg
26    666373753744588802    https://pbs.twimg.com/media/CT9vZEYWUAAlZO5.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-UQAACu2O.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  890066608113172480    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]
```

1.2.3 images_df columns description:

- tweet_id is the last part of the tweet URL after "status/".

- p1 is the algorithm's #1 prediction for the image in the tweet.
- p1_conf is how confident the algorithm is in its #1 prediction.
- p1_dog is whether or not the #1 prediction is a breed of dog.
- p2 is the algorithm's second most likely prediction.
- p2_conf is how confident the algorithm is in its #2 prediction.
- p2_dog is whether or not the #2 prediction is a breed of dog.
- p3 is the algorithm's third most likely prediction.
- p3_conf is how confident the algorithm is in its #3 prediction.
- p3_dog is whether or not the #3 prediction is a breed of dog.

```
In [8]: tweets_df
```

Out[8]:

	tweet_id	retweet_count	favorite_count
0	892420643555336193	8853	39467
1	892177421306343426	6514	33819
2	891815181378084864	4328	25461
3	891689557279858688	8964	42908
4	891327558926688256	9774	41048
5	891087950875897856	3261	20562
6	890971913173991426	2158	12041
7	890729181411237888	16716	56848
8	890609185150312448	4429	28226
9	890240255349198849	7711	32467
10	890006608113172480	7624	31166
11	889880896479866881	5156	28268
12	889665388333682689	8538	38818
13	889638837579907072	4735	27672
14	889531135344209921	2321	15359
15	889278841981685760	5637	25652
16	888917238123831296	4709	29611
17	888804989199671297	4559	26080
18	888554962724278272	3732	20290
19	888078434458587136	3653	22201
20	887705289381826560	5609	30779
21	887517139158093824	12082	46959
22	887473957103951883	18781	69871
23	887343217045368832	10737	34222
24	887101392804085760	6167	31061
25	886983233522544640	8084	35859
26	886736880519319552	3443	12306
27	886680336477933568	4610	22798
28	886366144734445568	3316	21524
29	886267009285017600	4	117

```
...                   ...            ...             ...
2324  666411507551481857            339             459
2325  666407126856765440             44             113
2326  666396247373291520             92             172
2327  666373753744588802            100             194
2328  666362758909284353            595             804
2329  666353288456101888             77             229
2330  666345417576210432            146             307
2331  666337882303524864             96             204
2332  666293911632134144            368             522
2333  666287406224695296             71             152
2334  666273097616637952             82             184
2335  666268910803644416             37             108
2336  666104133288665088           6871           14765
2337  666102155909144576             16              81
2338  666099513787052032             73             164
2339  666094000022159362             79             169
2340  666082916733198337             47             121
2341  666073100786774016            174             335
2342  666071193221509120             67             154
2343  666063827256086533            232             496
2344  666058600524156928             61             115
2345  666057090499244032            146             304
2346  666055525042405380            261             448
2347  666051853826850816            879            1253
2348  666050758794694657             60             136
2349  666049248165822465             41             111
2350  666044226329800704            147             311
2351  666033412701032449             47             128
2352  666029285002620928             48             132
2353  666020888022790149            532            2535

[2354 rows x 3 columns]
```

1.2.4 tweet_json columns description

- id: the unique identifier for each tweet
- retweet_count: the number of times the original tweet was retweeted
- favorite_count: the number of times the the original tweet was loved or liked

```
In [9]: twitter_df.describe()
```

```
Out[9]:            tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
        count  2.356000e+03           7.800000e+01         7.800000e+01
        mean   7.427716e+17           7.455079e+17         2.014171e+16
        std    6.856705e+16           7.582492e+16         1.252797e+17
        min    6.660209e+17           6.658147e+17         1.185634e+07
```

```
25%    6.783989e+17           6.757419e+17         3.086374e+08
50%    7.196279e+17           7.038708e+17         4.196984e+09
75%    7.993373e+17           8.257804e+17         4.196984e+09
max    8.924206e+17           8.862664e+17         8.405479e+17

       retweeted_status_id  retweeted_status_user_id  rating_numerator  \
count         1.810000e+02              1.810000e+02       2356.000000
mean          7.720400e+17              1.241698e+16         13.126486
std           6.236928e+16              9.599254e+16         45.876648
min           6.661041e+17              7.832140e+05          0.000000
25%           7.186315e+17              4.196984e+09         10.000000
50%           7.804657e+17              4.196984e+09         11.000000
75%           8.203146e+17              4.196984e+09         12.000000
max           8.874740e+17              7.874618e+17       1776.000000

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

In [10]:
```
twitter_df[twitter_df['rating_numerator'] > 12].rating_numerator.value_counts()
```

Out[10]:
```
13      351
14       54
75        2
15        2
420       2
17        1
20        1
24        1
26        1
27        1
44        1
45        1
50        1
60        1
960       1
84        1
88        1
99        1
121       1
143       1
144       1
```

```
666       1
165       1
182       1
204       1
1776      1
80        1
Name: rating_numerator, dtype: int64
```

```
In [11]: twitter_df[twitter_df['rating_numerator'] ==1776]
```

```
Out[11]:                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
979  749981277374128128                    NaN                  NaN

                      timestamp  \
979  2016-07-04 15:00:45 +0000

                                                source  \
979  <a href="https://about.twitter.com/products/tw...

                                                  text  retweeted_status_id  \
979  This is Atticus. He's quite simply America af...                  NaN

     retweeted_status_user_id retweeted_status_timestamp  \
979                       NaN                        NaN

                                          expanded_urls  rating_numerator  \
979  https://twitter.com/dog_rates/status/749981277...              1776

     rating_denominator     name doggo floofer pupper puppo
979                  10  Atticus  None    None   None  None
```

Issue : rating_numerator of 1776 is too oulier for the rest of data

1.2.5 Programmatic Assessment

```
In [12]: twitter_df.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 [13]: images_df.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 [14]: tweets_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 3 columns):
tweet_id          2354 non-null int64
retweet_count     2354 non-null int64
favorite_count    2354 non-null int64
dtypes: int64(3)
memory usage: 55.2 KB
```

```
In [15]: all_columns = pd.Series(list(twitter_df) + list(images_df) + list(tweets_df))
```

```
In [16]: all_columns
```

```
Out[16]: 0                       tweet_id
         1            in_reply_to_status_id
         2              in_reply_to_user_id
         3                        timestamp
         4                           source
         5                             text
         6              retweeted_status_id
         7         retweeted_status_user_id
         8       retweeted_status_timestamp
         9                    expanded_urls
         10                rating_numerator
         11              rating_denominator
         12                            name
         13                           doggo
         14                          floofer
         15                           pupper
         16                           puppo
         17                        tweet_id
         18                         jpg_url
         19                         img_num
         20                              p1
         21                         p1_conf
         22                          p1_dog
         23                              p2
         24                         p2_conf
         25                          p2_dog
         26                              p3
         27                         p3_conf
         28                          p3_dog
         29                        tweet_id
         30                  retweet_count
         31                 favorite_count
         dtype: object

In [17]: all_columns[all_columns.duplicated()]

Out[17]: 17    tweet_id
         29    tweet_id
         dtype: object

In [18]: twitter_df.source.value_counts()

Out[18]: <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
         <a href="http://vine.co" rel="nofollow">Vine - Make a Scene</a>
         <a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>
         <a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>
         Name: source, dtype: int64

In [19]: twitter_df.rating_denominator.value_counts()
```

```
Out[19]: 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 [20]: twitter_df[twitter_df.rating_denominator==0]

Out[20]:               tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
         313  835246439529840640           8.352460e+17           26259576.0

                             timestamp  \
         313  2017-02-24 21:54:03 +0000

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

                                                      text  retweeted_status_id  \
         313  @jonnysun @Lin_Manuel ok jomny I know you're e...                  NaN

              retweeted_status_user_id retweeted_status_timestamp expanded_urls  \
         313                       NaN                        NaN           NaN

              rating_numerator  rating_denominator  name doggo floofer pupper puppo
         313               960                   0  None  None    None   None  None

In [21]: images_df.p1.value_counts()

Out[21]: golden_retriever             150
         Labrador_retriever           100
         Pembroke                      89
         Chihuahua                     83
         pug                           57
         chow                          44
         Samoyed                       43
```

```
toy_poodle                  39
Pomeranian                  38
malamute                    30
cocker_spaniel              30
French_bulldog              26
Chesapeake_Bay_retriever    23
miniature_pinscher          23
seat_belt                   22
German_shepherd             20
Siberian_husky              20
Staffordshire_bullterrier   20
Cardigan                    19
web_site                    19
Shetland_sheepdog           18
beagle                      18
Eskimo_dog                  18
teddy                       18
Maltese_dog                 18
Lakeland_terrier            17
Shih-Tzu                    17
Rottweiler                  17
Italian_greyhound           16
kuvasz                      16
                           ...
standard_schnauzer           1
water_buffalo                1
rain_barrel                  1
skunk                        1
desktop_computer             1
cowboy_boot                  1
loupe                        1
microphone                   1
wooden_spoon                 1
marmot                       1
cougar                       1
hay                          1
harp                         1
bannister                    1
ocarina                      1
long-horned_beetle           1
rapeseed                     1
hammer                       1
flamingo                     1
African_hunting_dog          1
African_crocodile            1
coral_reef                   1
bib                          1
fire_engine                  1
```

```
guenon                          1
damselfly                       1
grey_fox                        1
bee_eater                       1
trombone                        1
carousel                        1
Name: p1, Length: 378, dtype: int64
```

```
In [22]: images_df.p2.value_counts()
```

```
Out[22]: Labrador_retriever               104
         golden_retriever                  92
         Cardigan                          73
         Chihuahua                         44
         Pomeranian                        42
         Chesapeake_Bay_retriever          41
         French_bulldog                    41
         toy_poodle                        37
         cocker_spaniel                    34
         Siberian_husky                    33
         miniature_poodle                  33
         beagle                            28
         collie                            27
         Pembroke                          27
         Eskimo_dog                        27
         kuvasz                            26
         Italian_greyhound                 22
         Pekinese                          21
         American_Staffordshire_terrier    21
         malinois                          20
         toy_terrier                       20
         miniature_pinscher                20
         chow                              20
         Samoyed                           20
         Boston_bull                       19
         Norwegian_elkhound                19
         Staffordshire_bullterrier         18
         Irish_terrier                     17
         pug                               17
         kelpie                            16
                                          ...
         EntleBucher                        1
         hyena                              1
         wombat                             1
         promontory                         1
         water_bottle                       1
         snail                              1
         spotlight                          1
```

```
web_site                        1
bearskin                        1
desk                            1
horse_cart                      1
giant_panda                     1
screw                           1
ashcan                          1
comic_book                      1
lifeboat                        1
lawn_mower                      1
affenpinscher                   1
umbrella                        1
toaster                         1
hotdog                          1
dining_table                    1
shower_cap                      1
Kerry_blue_terrier              1
patio                           1
maillot                         1
crutch                          1
snorkel                         1
menu                            1
hand-held_computer              1
Name: p2, Length: 405, dtype: int64
```

```
In [23]: twitter_df[twitter_df['retweeted_status_id'].notnull()]
```

Out[23]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	\
19	888202515573088257	NaN	NaN	
32	886054160059072513	NaN	NaN	
36	885311592912609280	NaN	NaN	
68	879130579576475649	NaN	NaN	
73	878404777348136964	NaN	NaN	
74	878316110768087041	NaN	NaN	
78	877611172832227328	NaN	NaN	
91	874434818259525634	NaN	NaN	
95	873697596434513921	NaN	NaN	
97	873337748698140672	NaN	NaN	
101	872668790621863937	NaN	NaN	
109	871166179821445120	NaN	NaN	
118	869988702071779329	NaN	NaN	
124	868639477480148993	NaN	NaN	
130	867072653475098625	NaN	NaN	
132	866816280283807744	NaN	NaN	
137	866094527597207552	NaN	NaN	
146	863471782782697472	NaN	NaN	
155	861769973181624320	NaN	NaN	
159	860981674716409858	NaN	NaN	

```
160   860924035999428608                  NaN                  NaN
165   860177593139703809                  NaN                  NaN
171   858860390427611136                  NaN                  NaN
180   857062103051644929                  NaN                  NaN
182   856602993587888130                  NaN                  NaN
185   856330835276025856                  NaN                  NaN
194   855245323840757760                  NaN                  NaN
195   855138241867124737                  NaN                  NaN
204   852936405516943360                  NaN                  NaN
211   851953902622658560                  NaN                  NaN
...                  ...                  ...                  ...
784   775096608509886464                  NaN                  NaN
794   773336787167145985                  NaN                  NaN
800   772615324260794368                  NaN                  NaN
811   771171053431250945                  NaN                  NaN
815   771004394259247104                  NaN                  NaN
818   770743923962707968                  NaN                  NaN
822   770093767776997377                  NaN                  NaN
826   769335591808995329                  NaN                  NaN
829   768909767477751808                  NaN                  NaN
833   768554158521745409                  NaN                  NaN
841   766864461642756096                  NaN                  NaN
847   766078092750233600                  NaN                  NaN
860   763167063695355904                  NaN                  NaN
868   761750502866649088                  NaN                  NaN
872   761371037149827077                  NaN                  NaN
885   760153949710192640                  NaN                  NaN
890   759566828574212096                  NaN                  NaN
895   759159934323924993                  NaN                  NaN
908   757729163776290825                  NaN                  NaN
911   757597904299253760                  NaN                  NaN
926   754874841593970688                  NaN                  NaN
937   753298634498793472                  NaN                  NaN
943   752701944171524096                  NaN                  NaN
949   752309394570878976                  NaN                  NaN
1012  747242308580548608                  NaN                  NaN
1023  746521445350707200                  NaN                  NaN
1043  743835915802583040                  NaN                  NaN
1242  711998809858043904                  NaN                  NaN
2259  667550904950915073                  NaN                  NaN
2260  667550882905632768                  NaN                  NaN

                     timestamp  \
19    2017-07-21 01:02:36 +0000
32    2017-07-15 02:45:48 +0000
36    2017-07-13 01:35:06 +0000
68    2017-06-26 00:13:58 +0000
73    2017-06-24 00:09:53 +0000
```

```
74    2017-06-23 18:17:33 +0000
78    2017-06-21 19:36:23 +0000
91    2017-06-13 01:14:41 +0000
95    2017-06-11 00:25:14 +0000
97    2017-06-10 00:35:19 +0000
101   2017-06-08 04:17:07 +0000
109   2017-06-04 00:46:17 +0000
118   2017-05-31 18:47:24 +0000
124   2017-05-28 01:26:04 +0000
130   2017-05-23 17:40:04 +0000
132   2017-05-23 00:41:20 +0000
137   2017-05-21 00:53:21 +0000
146   2017-05-13 19:11:30 +0000
155   2017-05-09 02:29:07 +0000
159   2017-05-06 22:16:42 +0000
160   2017-05-06 18:27:40 +0000
165   2017-05-04 17:01:34 +0000
171   2017-05-01 01:47:28 +0000
180   2017-04-26 02:41:43 +0000
182   2017-04-24 20:17:23 +0000
185   2017-04-24 02:15:55 +0000
194   2017-04-21 02:22:29 +0000
195   2017-04-20 19:16:59 +0000
204   2017-04-14 17:27:40 +0000
211   2017-04-12 00:23:33 +0000
...                          ...
784   2016-09-11 22:20:06 +0000
794   2016-09-07 01:47:12 +0000
800   2016-09-05 02:00:22 +0000
811   2016-09-01 02:21:21 +0000
815   2016-08-31 15:19:06 +0000
818   2016-08-30 22:04:05 +0000
822   2016-08-29 03:00:36 +0000
826   2016-08-27 00:47:53 +0000
829   2016-08-25 20:35:48 +0000
833   2016-08-24 21:02:45 +0000
841   2016-08-20 05:08:29 +0000
847   2016-08-18 01:03:45 +0000
860   2016-08-10 00:16:21 +0000
868   2016-08-06 02:27:27 +0000
872   2016-08-05 01:19:35 +0000
885   2016-08-01 16:43:19 +0000
890   2016-07-31 01:50:18 +0000
895   2016-07-29 22:53:27 +0000
908   2016-07-26 00:08:05 +0000
911   2016-07-25 15:26:30 +0000
926   2016-07-18 03:06:01 +0000
937   2016-07-13 18:42:44 +0000
```

```
943   2016-07-12 03:11:42 +0000
949   2016-07-11 01:11:51 +0000
1012  2016-06-27 01:37:04 +0000
1023  2016-06-25 01:52:36 +0000
1043  2016-06-17 16:01:16 +0000
1242  2016-03-21 19:31:59 +0000
2259  2015-11-20 03:51:52 +0000
2260  2015-11-20 03:51:47 +0000

                                                  source  \
19    <a href="http://twitter.com/download/iphone" r...
32    <a href="http://twitter.com/download/iphone" r...
36    <a href="http://twitter.com/download/iphone" r...
68    <a href="http://twitter.com/download/iphone" r...
73    <a href="http://twitter.com/download/iphone" r...
74    <a href="http://twitter.com/download/iphone" r...
78    <a href="http://twitter.com/download/iphone" r...
91    <a href="http://twitter.com/download/iphone" r...
95    <a href="http://twitter.com/download/iphone" r...
97    <a href="http://twitter.com/download/iphone" r...
101   <a href="http://twitter.com/download/iphone" r...
109   <a href="http://twitter.com/download/iphone" r...
118   <a href="http://twitter.com/download/iphone" r...
124   <a href="http://twitter.com/download/iphone" r...
130   <a href="http://twitter.com/download/iphone" r...
132   <a href="http://twitter.com/download/iphone" r...
137   <a href="http://twitter.com/download/iphone" r...
146   <a href="http://twitter.com/download/iphone" r...
155   <a href="http://twitter.com/download/iphone" r...
159   <a href="http://twitter.com/download/iphone" r...
160   <a href="http://twitter.com/download/iphone" r...
165   <a href="http://twitter.com/download/iphone" r...
171   <a href="http://twitter.com/download/iphone" r...
180   <a href="http://twitter.com/download/iphone" r...
182   <a href="http://twitter.com/download/iphone" r...
185   <a href="http://twitter.com/download/iphone" r...
194   <a href="http://twitter.com/download/iphone" r...
195   <a href="http://twitter.com/download/iphone" r...
204   <a href="http://twitter.com/download/iphone" r...
211   <a href="http://twitter.com/download/iphone" r...
...                                                 ...
784   <a href="http://twitter.com/download/iphone" r...
794   <a href="http://twitter.com/download/iphone" r...
800   <a href="http://twitter.com/download/iphone" r...
811   <a href="http://twitter.com/download/iphone" r...
815   <a href="http://twitter.com/download/iphone" r...
818   <a href="http://twitter.com/download/iphone" r...
822   <a href="http://twitter.com/download/iphone" r...
```

```
826   <a href="http://twitter.com/download/iphone" r...
829   <a href="http://twitter.com/download/iphone" r...
833   <a href="http://twitter.com/download/iphone" r...
841   <a href="http://twitter.com/download/iphone" r...
847   <a href="http://twitter.com/download/iphone" r...
860   <a href="http://twitter.com/download/iphone" r...
868   <a href="http://twitter.com/download/iphone" r...
872   <a href="http://twitter.com/download/iphone" r...
885   <a href="http://twitter.com/download/iphone" r...
890   <a href="http://twitter.com/download/iphone" r...
895   <a href="http://twitter.com/download/iphone" r...
908   <a href="http://twitter.com/download/iphone" r...
911   <a href="http://twitter.com/download/iphone" r...
926   <a href="http://twitter.com/download/iphone" r...
937   <a href="http://twitter.com/download/iphone" r...
943   <a href="http://twitter.com/download/iphone" r...
949   <a href="http://twitter.com/download/iphone" r...
1012  <a href="http://twitter.com/download/iphone" r...
1023  <a href="http://twitter.com/download/iphone" r...
1043  <a href="http://twitter.com/download/iphone" r...
1242  <a href="http://twitter.com/download/iphone" r...
2259  <a href="http://twitter.com" rel="nofollow">Tw...
2260  <a href="http://twitter.com" rel="nofollow">Tw...

                                                 text  retweeted_status_id  \
19    RT @dog_rates: This is Canela. She attempted s...         8.874740e+17
32    RT @Athletics: 12/10 #BATP https://t.co/WxwJmv...         8.860537e+17
36    RT @dog_rates: This is Lilly. She just paralle...         8.305833e+17
68    RT @dog_rates: This is Emmy. She was adopted t...         8.780576e+17
73    RT @dog_rates: Meet Shadow. In an attempt to r...         8.782815e+17
74    RT @dog_rates: Meet Terrance. He's being yelle...         6.690004e+17
78    RT @rachel2195: @dog_rates the boyfriend and h...         8.768508e+17
91    RT @dog_rates: This is Coco. At first I though...         8.663350e+17
95    RT @dog_rates: This is Walter. He won't start ...         8.688804e+17
97    RT @dog_rates: This is Sierra. She's one preci...         8.732138e+17
101   RT @loganamnosis: Penelope here is doing me qu...         8.726576e+17
109   RT @dog_rates: This is Dawn. She's just checki...         8.410770e+17
118   RT @dog_rates: We only rate dogs. This is quit...         8.591970e+17
124   RT @dog_rates: Say hello to Cooper. His expres...         8.685523e+17
130   RT @rachaeleasler: these @dog_rates hats are 1...         8.650134e+17
132   RT @dog_rates: This is Jamesy. He gives a kiss...         8.664507e+17
137   RT @dog_rates: Here's a pupper before and afte...         8.378202e+17
146   RT @dog_rates: Say hello to Quinn. She's quite...         8.630625e+17
155   RT @dog_rates: "Good afternoon class today we'...         8.066291e+17
159   RT @dog_rates: Meet Lorenzo. He's an avid nift...         8.605638e+17
160   RT @tallylott: h*ckin adorable promposal. 13/1...         8.609145e+17
165   RT @dog_rates: Ohboyohboyohboyohboyohboyohboyo...         7.616730e+17
171   RT @dog_rates: Meet Winston. He knows he's a l...         8.395493e+17
```

```
180   RT @AaronChewning: First time wearing my @dog_...         8.570611e+17
182   RT @dog_rates: This is Luna. It's her first ti...         8.447048e+17
185   RT @Jenna_Marbles: @dog_rates Thanks for ratin...         8.563302e+17
194   RT @dog_rates: Meet George. He looks slightly ...         8.421635e+17
195   RT @frasercampbell_: oh my... what's that... b...         8.551225e+17
204   RT @dog_rates: I usually only share these on F...         8.316501e+17
211   RT @dog_rates: This is Astrid. She's a guide d...         8.293743e+17
...                                                 ...                  ...
784   RT @dog_rates: After so many requests, this is...         7.403732e+17
794   RT @dog_rates: Meet Fizz. She thinks love is a...         7.713808e+17
800   RT @dog_rates: This is Gromit. He's pupset bec...         7.652221e+17
811   RT @dog_rates: This is Frankie. He's wearing b...         6.733201e+17
815   RT @katieornah: @dog_rates learning a lot at c...         7.710021e+17
818   RT @dog_rates: Here's a doggo blowing bubbles...          7.392382e+17
822   RT @dog_rates: This is just downright precious...         7.410673e+17
826   RT @dog_rates: Ever seen a dog pet another dog...         7.069045e+17
829   RT @dog_rates: When it's Janet from accounting...         7.001438e+17
833   RT @dog_rates: This is Nollie. She's waving at...         7.399792e+17
841   RT @dog_rates: We only rate dogs... this is a ...         7.599238e+17
847   RT @dog_rates: This is Colby. He's currently r...         7.258423e+17
860   RT @dog_rates: Meet Eve. She's a raging alcoho...         6.732953e+17
868   RT @dog_rates: "Tristan do not speak to me wit...         6.853251e+17
872   RT @dog_rates: Oh. My. God. 13/10 magical af h...         7.116948e+17
885   RT @hownottodraw: The story/person behind @dog...         7.601538e+17
890   RT @dog_rates: This... is a Tyrannosaurus rex...          7.395441e+17
895   RT @dog_rates: AT DAWN...\nWE RIDE\n\n11/10 ht...         6.703191e+17
908   RT @dog_rates: This is Chompsky. He lives up t...         6.790626e+17
911   RT @jon_hill987: @dog_rates There is a cunning...         7.575971e+17
926   RT @dog_rates: This is Rubio. He has too much ...         6.791584e+17
937   RT @dog_rates: This is Carly. She's actually 2...         6.815232e+17
943   RT @dog_rates: HEY PUP WHAT'S THE PART OF THE ...         6.835159e+17
949   RT @dog_rates: Everyone needs to watch this. 1...         6.753544e+17
1012  RT @dog_rates: This pupper killed this great w...         7.047611e+17
1023  RT @dog_rates: This is Shaggy. He knows exactl...         6.678667e+17
1043  RT @dog_rates: Extremely intelligent dog here...          6.671383e+17
1242  RT @twitter: @dog_rates Awesome Tweet! 12/10. ...         7.119983e+17
2259  RT @dogratingrating: Exceptional talent. Origi...         6.675487e+17
2260  RT @dogratingrating: Unoriginal idea. Blatant ...         6.675484e+17

      retweeted_status_user_id retweeted_status_timestamp  \
19                4.196984e+09  2017-07-19 00:47:34 +0000
32                1.960740e+07  2017-07-15 02:44:07 +0000
36                4.196984e+09  2017-02-12 01:04:29 +0000
68                4.196984e+09  2017-06-23 01:10:23 +0000
73                4.196984e+09  2017-06-23 16:00:04 +0000
74                4.196984e+09  2015-11-24 03:51:38 +0000
78                5.128045e+08  2017-06-19 17:14:49 +0000
91                4.196984e+09  2017-05-21 16:48:45 +0000
```

95	4.196984e+09	2017-05-28 17:23:24 +0000
97	4.196984e+09	2017-06-09 16:22:42 +0000
101	1.547674e+08	2017-06-08 03:32:35 +0000
109	4.196984e+09	2017-03-13 00:02:39 +0000
118	4.196984e+09	2017-05-02 00:04:57 +0000
124	4.196984e+09	2017-05-27 19:39:34 +0000
130	7.874618e+17	2017-05-18 01:17:25 +0000
132	4.196984e+09	2017-05-22 00:28:40 +0000
137	4.196984e+09	2017-03-04 00:21:08 +0000
146	4.196984e+09	2017-05-12 16:05:02 +0000
155	4.196984e+09	2016-12-07 22:38:52 +0000
159	4.196984e+09	2017-05-05 18:36:06 +0000
160	3.638908e+08	2017-05-06 17:49:42 +0000
165	4.196984e+09	2016-08-05 21:19:27 +0000
171	4.196984e+09	2017-03-08 18:52:12 +0000
180	5.870972e+07	2017-04-26 02:37:47 +0000
182	4.196984e+09	2017-03-23 00:18:10 +0000
185	6.669901e+07	2017-04-24 02:13:14 +0000
194	4.196984e+09	2017-03-16 00:00:07 +0000
195	7.475543e+17	2017-04-20 18:14:33 +0000
204	4.196984e+09	2017-02-14 23:43:18 +0000
211	4.196984e+09	2017-02-08 17:00:26 +0000
...	...	...
784	4.196984e+09	2016-06-08 02:41:38 +0000
794	4.196984e+09	2016-09-01 16:14:48 +0000
800	4.196984e+09	2016-08-15 16:22:20 +0000
811	4.196984e+09	2015-12-06 01:56:44 +0000
815	1.732729e+09	2016-08-31 15:10:07 +0000
818	4.196984e+09	2016-06-04 23:31:25 +0000
822	4.196984e+09	2016-06-10 00:39:48 +0000
826	4.196984e+09	2016-03-07 18:09:06 +0000
829	4.196984e+09	2016-02-18 02:24:13 +0000
833	4.196984e+09	2016-06-07 00:36:02 +0000
841	4.196984e+09	2016-08-01 01:28:46 +0000
847	4.196984e+09	2016-04-29 00:21:01 +0000
860	4.196984e+09	2015-12-06 00:17:55 +0000
868	4.196984e+09	2016-01-08 05:00:14 +0000
872	4.196984e+09	2016-03-20 23:23:54 +0000
885	1.950368e+08	2016-08-01 16:42:51 +0000
890	4.196984e+09	2016-06-05 19:47:03 +0000
895	4.196984e+09	2015-11-27 19:11:49 +0000
908	4.196984e+09	2015-12-21 22:15:18 +0000
911	2.804798e+08	2016-07-25 15:23:28 +0000
926	4.196984e+09	2015-12-22 04:35:49 +0000
937	4.196984e+09	2015-12-28 17:12:42 +0000
943	4.196984e+09	2016-01-03 05:11:12 +0000
949	4.196984e+09	2015-12-11 16:40:19 +0000
1012	4.196984e+09	2016-03-01 20:11:59 +0000

```
1023             4.196984e+09  2015-11-21 00:46:50 +0000
1043             4.196984e+09  2015-11-19 00:32:12 +0000
1242             7.832140e+05  2016-03-21 19:29:52 +0000
2259             4.296832e+09  2015-11-20 03:43:06 +0000
2260             4.296832e+09  2015-11-20 03:41:59 +0000

                                          expanded_urls  rating_numerator  \
19    https://twitter.com/dog_rates/status/887473957...                13
32    https://twitter.com/dog_rates/status/886053434...                12
36    https://twitter.com/dog_rates/status/830583320...                13
68    https://twitter.com/dog_rates/status/878057613...                14
73    https://www.gofundme.com/3yd6y1c,https://twitt...                13
74    https://twitter.com/dog_rates/status/669000397...                11
78    https://twitter.com/rachel2195/status/87685077...                14
91    https://twitter.com/dog_rates/status/866334964...                12
95    https://twitter.com/dog_rates/status/868880397...                14
97    https://www.gofundme.com/help-my-baby-sierra-g...                12
101   https://twitter.com/loganamnosis/status/872657...                14
109   https://twitter.com/dog_rates/status/841077006...                12
118   https://twitter.com/dog_rates/status/859196978...                12
124   https://www.gofundme.com/3ti3nps,https://twitt...                12
130   https://twitter.com/rachaeleasler/status/86501...                13
132   https://twitter.com/dog_rates/status/866450705...                13
137   https://twitter.com/dog_rates/status/837820167...                12
146   https://www.gofundme.com/helpquinny,https://tw...                13
155   https://twitter.com/dog_rates/status/806629075...                13
159   https://www.gofundme.com/help-lorenzo-beat-can...                13
160   https://twitter.com/tallylott/status/860914485...                13
165   https://twitter.com/dog_rates/status/761672994...                10
171   https://twitter.com/dog_rates/status/839549326...                12
180   https://twitter.com/AaronChewning/status/85706...                13
182   https://twitter.com/dog_rates/status/844704788...                13
185                                                 NaN                14
194   https://twitter.com/dog_rates/status/842163532...                12
195   https://twitter.com/frasercampbell_/status/855...                14
204   http://www.gofundme.com/bluethewhitehusky,http...                13
211   https://twitter.com/dog_rates/status/829374341...                13
...                                                 ...               ...
784   https://twitter.com/dog_rates/status/740373189...                 9
794   https://twitter.com/dog_rates/status/771380798...                11
800   https://twitter.com/dog_rates/status/765222098...                10
811   https://twitter.com/dog_rates/status/673320132...                11
815   https://twitter.com/katieornah/status/77100213...                12
818   https://twitter.com/dog_rates/status/739238157...                13
822   https://twitter.com/dog_rates/status/741067306...                12
826   https://vine.co/v/iXQAm5Lrgrh,https://vine.co/...                13
829   https://twitter.com/dog_rates/status/700143752...                10
833   https://twitter.com/dog_rates/status/739979191...                12
```

```
841   https://twitter.com/dog_rates/status/759923798...              10
847   https://twitter.com/dog_rates/status/725842289...              12
860   https://twitter.com/dog_rates/status/673295268...               8
868   https://twitter.com/dog_rates/status/685325112...              10
872   https://twitter.com/dog_rates/status/711694788...              13
885   https://weratedogs.com/pages/about-us,https://...              11
890   https://twitter.com/dog_rates/status/739544079...              10
895   https://twitter.com/dog_rates/status/670319130...              11
908   https://twitter.com/dog_rates/status/679062614...              11
911   https://twitter.com/jon_hill987/status/7575971...              11
926   https://twitter.com/dog_rates/status/679158373...              11
937   https://twitter.com/dog_rates/status/681523177...              12
943   https://vine.co/v/ibvnzrauFuV,https://vine.co/...              11
949   https://twitter.com/dog_rates/status/675354435...              13
1012  https://twitter.com/dog_rates/status/704761120...              13
1023  https://twitter.com/dog_rates/status/667866724...              10
1043  https://twitter.com/dog_rates/status/667138269...              10
1242  https://twitter.com/twitter/status/71199827977...              12
2259  https://twitter.com/dogratingrating/status/667...              12
2260  https://twitter.com/dogratingrating/status/667...               5

      rating_denominator      name  doggo floofer  pupper puppo
19                    10    Canela   None    None    None  None
32                    10      None   None    None    None  None
36                    10     Lilly   None    None    None  None
68                    10      Emmy   None    None    None  None
73                    10    Shadow   None    None    None  None
74                    10  Terrance   None    None    None  None
78                    10      None   None    None  pupper  None
91                    10      Coco   None    None    None  None
95                    10    Walter   None    None    None  None
97                    10    Sierra   None    None  pupper  None
101                   10      None   None    None    None  None
109                   10      Dawn   None    None    None  None
118                   10     quite   None    None    None  None
124                   10    Cooper   None    None    None  None
130                   10      None   None    None    None  None
132                   10    Jamesy   None    None  pupper  None
137                   10      None   None    None  pupper  None
146                   10     Quinn   None    None    None  None
155                   10      None   None    None    None  None
159                   10   Lorenzo   None    None    None  None
160                   10      None   None    None    None  None
165                   10      None   None    None    None  None
171                   10   Winston   None    None    None  None
180                   10      None   None    None    None  None
182                   10      Luna   None    None    None  None
185                   10      None   None    None    None  None
```

```
194                  10    George   None    None    None  None
195                  10      None   None    None    None  None
204                  10      None   None    None    None  None
211                  10    Astrid  doggo    None    None  None
...                 ...       ...    ...     ...     ...   ...
784                  11      None   None    None    None  None
794                  10      Fizz   None    None    None  None
800                  10    Gromit   None    None    None  None
811                  10   Frankie   None    None    None  None
815                  10      None   None    None  pupper  None
818                  10      None  doggo    None    None  None
822                  10      just  doggo    None  pupper  None
826                  10      None   None    None    None  None
829                  10      None   None    None  pupper  None
833                  10    Nollie   None    None    None  None
841                  10      None   None    None    None  None
847                  10     Colby   None    None    None  None
860                  10       Eve   None    None  pupper  None
868                  10      None   None    None    None  None
872                  10      None   None    None    None  None
885                  10      None   None    None    None  None
890                  10      None   None    None    None  None
895                  10      None   None    None    None  None
908                  10  Chompsky   None    None    None  None
911                  10      None   None    None  pupper  None
926                  10     Rubio   None    None    None  None
937                  10     Carly   None    None    None  None
943                  10      None   None    None    None  None
949                  10      None   None    None    None  None
1012                 10      None   None    None  pupper  None
1023                 10    Shaggy   None    None    None  None
1043                 10      None   None    None    None  None
1242                 10      None   None    None    None  None
2259                 10      None   None    None    None  None
2260                 10      None   None    None    None  None

[181 rows x 17 columns]
```

1.2.6 Quality issues

`twitter_df` table:

- `retweeted_status_id`, `retweeted_status_user _id`, and `retweeted_status _timestamp` that are not null are retweeted and won't be used for our analysis.
- Missing values in `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp`, `in_reply_to_status_id`, `in_reply_to_user_id`, `expanded_urls` columns contain mostly missing values.

- Missing rating (rating_numerator/rating_denominator)
- Erroneous datatypes (timestamp should be a datetime, tweet_id should be a string)
- Contains html tags in Source columns
- Some names in name column are not a real name.
- Nulls represented as None in doggo, pupper, puppo, and floofer
- doggo, pupper, puppo, floofer should be category type
- rating_numerator of 1776 is too oulier for the rest of data

`images_df` table: - Values in columns 'p1', 'p2', and 'p3' don't have consistent format

- Erroneous datatypes (tweet_is should be a string)

`tweets_df` table: - Erroneous datatypes (tweet_is should be a string).

- tweet_id 776201521193218049: the name should be O'Malley instead of O

1.2.7 Tidiness issues

`twitter_df` table: - doggo, pupper, puppo, floofer should be in 1 column because it shows the dog stage

- tweet_id columns in 3 tables should be the same for researching purpose.
- We can merge all 3 tables into one.

1.3 Cleaning Data

In this section, clean **all** of the issues you documented while assessing.

Note: Make a copy of the original data before cleaning. Cleaning includes merging individual pieces of data according to the rules of tidy data. The result should be a high-quality and tidy master pandas DataFrame (or DataFrames, if appropriate).

```
In [24]: # Make copies of original pieces of data
         twitter_clean = twitter_df.copy()
         images_clean = images_df.copy()
         tweets_clean = tweets_df.copy()
```

1.3.1 Missing Values:

1.3.2 Issue #1: `retweeted_status_id`, `retweeted_status_user _id`, and `retweeted_status _timestamp` that are not null are retweeted and won't be used for our analysis.

Define: Remove tweet IDs that have 'retweeted_status_id, retweeted_status_user _id, and retweeted_status _timestamp values.

Code

```
In [25]: #Remove tweet IDs that have 'retweeted_status_id', 'retweeted_status_user_id', and 'ret
         twitter_clean = twitter_clean[twitter_clean.retweeted_status_id.isnull()]
         twitter_clean = twitter_clean[twitter_clean.retweeted_status_user_id.isnull()]
         twitter_clean = twitter_clean[twitter_clean.retweeted_status_timestamp.isnull()]
```

Test

```
In [26]: # Check if the retweets have been droped - should be 0 in 3 prints
         print(twitter_clean.retweeted_status_id.notnull().sum())
         print(twitter_clean.retweeted_status_user_id.notnull().sum())
         print(twitter_clean.retweeted_status_timestamp.notnull().sum())
```

```
0
0
0
```

1.3.3 Issue #2: Missing values in retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id, in_reply_to_user_id, expanded_urls columns contain mostly missing values.

Define: Those columns are not used in the anlysis. Remove in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns

Code

```
In [27]: # Make a list of the columns to be dropped
         drop_list = ['in_reply_to_status_id', 'in_reply_to_user_id',
         'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp']
```

```
In [28]: # Drop the columns
         twitter_clean.drop(drop_list, axis=1, inplace=True)
```

Test

```
In [29]: #show the number of columns after dropped 5 columns - the number should be 12
         twitter_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 12 columns):
tweet_id              2175 non-null int64
timestamp             2175 non-null object
source                2175 non-null object
text                  2175 non-null object
expanded_urls         2117 non-null object
```

```
rating_numerator      2175 non-null int64
rating_denominator    2175 non-null int64
name                  2175 non-null object
doggo                 2175 non-null object
floofer               2175 non-null object
pupper                2175 non-null object
puppo                 2175 non-null object
dtypes: int64(3), object(9)
memory usage: 220.9+ KB
```

1.3.4 Issue #3: Missing rating (rating_numerator/rating_denominator)

Define: Remove rows which have the rating_denominator is smaller or equal 0 Calculate the rate column: rating_numerator divided by rating_denominator

Code

```
In [30]: #Remove rows which have the rating_denominator is smaller or equal 0
         twitter_clean = twitter_clean[twitter_clean.rating_denominator > 0]
```

```
In [31]: #Calculate the rate column: rating_numerator divided by rating_denominator
         twitter_clean['rate'] = (twitter_clean.rating_numerator / twitter_clean.rating_denomina
```

Test

```
In [32]: #show the descriptive statistic of rate column
         twitter_clean.rate.describe()
```

```
Out[32]: count    2174.000000
         mean        1.223398
         std         4.247731
         min         0.000000
         25%         1.000000
         50%         1.100000
         75%         1.200000
         max       177.600000
         Name: rate, dtype: float64
```

1.3.5 Tidiness:

1.3.6 Issue #4: doggo, pupper, puppo, floofer should be in 1 column because it shows the dog stage.

1.3.7 Issue #5: Nulls represented as None in doggo, pupper, puppo, and floofer

Define

- First replace the None values to nan in doggo, pupper, puppo, floofer
- Merge 4 columns: doggo, pupper, puppo, floofer into one named dog_stage
- Drop 4 previous columns: doggo, pupper, puppo, floofer.
- Convert empty cell to undefined in dog_stage column

Code

```
In [33]: #First replace the None values to nan in doggo, pupper, puppo, floofer
         twitter_clean.doggo = twitter_clean.doggo.replace('None',np.nan)
         twitter_clean.pupper = twitter_clean.pupper.replace('None',np.nan)
         twitter_clean.puppo = twitter_clean.puppo.replace('None',np.nan)
         twitter_clean.floofer = twitter_clean.floofer.replace('None',np.nan)

In [34]: # Merge 4 columns: doggo, pupper, puppo, floofer into one named dog_stage
         twitter_clean['dog_stage'] = twitter_clean[['doggo', 'pupper', 'puppo','floofer']].appl

In [35]: #Drop 4 previous columns: doggo, pupper, puppo, floofer.
         twitter_clean.drop(['doggo', 'pupper', 'puppo','floofer'], axis=1, inplace=True)

In [36]: #Convert empty cell to None in dog_stage column
         twitter_clean['dog_stage'] = twitter_clean['dog_stage'].replace('','None')
```

Test

```
In [37]: #show the result
         twitter_clean.head()

Out[37]:             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
         3  891689557279858688  2017-07-30 15:58:51 +0000
         4  891327558926688256  2017-07-29 16:00:24 +0000

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

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

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

```
   rating_denominator      name  rate dog_stage
0                  10   Phineas   1.3      None
1                  10     Tilly   1.3      None
2                  10    Archie   1.2      None
3                  10     Darla   1.3      None
4                  10  Franklin   1.2      None
```

```
In [38]: twitter_clean.dog_stage.value_counts()
```

```
Out[38]: None              1830
         pupper             224
         doggo               75
         puppo               24
         doggo, pupper       10
         floofer              9
         doggo, floofer       1
         doggo, puppo         1
         Name: dog_stage, dtype: int64
```

```
In [39]: twitter_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2174 entries, 0 to 2355
Data columns (total 10 columns):
tweet_id              2174 non-null int64
timestamp             2174 non-null object
source                2174 non-null object
text                  2174 non-null object
expanded_urls         2117 non-null object
rating_numerator      2174 non-null int64
rating_denominator    2174 non-null int64
name                  2174 non-null object
rate                  2174 non-null float64
dog_stage             2174 non-null object
dtypes: float64(1), int64(3), object(6)
memory usage: 186.8+ KB
```

Issue #6 and #7: tweet_id columns in 3 tables should be the same for researching purpose, we can merge all 3 tables into one.

Define Use merge function to merge 3 tables by the inner join

Code

```
In [40]: #Use merge function to merge 3 tables by the inner join
         Merge_df = pd.merge(twitter_clean, images_clean, on = 'tweet_id', how = 'inner').merge(
```

Test

```
In [41]: #check if the result data have all the columns of 3 previous datasets.
         Merge_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1994 entries, 0 to 1993
Data columns (total 23 columns):
tweet_id              1994 non-null int64
timestamp             1994 non-null object
source                1994 non-null object
text                  1994 non-null object
expanded_urls         1994 non-null object
rating_numerator      1994 non-null int64
rating_denominator    1994 non-null int64
name                  1994 non-null object
rate                  1994 non-null float64
dog_stage             1994 non-null object
jpg_url               1994 non-null object
img_num               1994 non-null int64
p1                    1994 non-null object
p1_conf               1994 non-null float64
p1_dog                1994 non-null bool
p2                    1994 non-null object
p2_conf               1994 non-null float64
p2_dog                1994 non-null bool
p3                    1994 non-null object
p3_conf               1994 non-null float64
p3_dog                1994 non-null bool
retweet_count         1994 non-null int64
favorite_count        1994 non-null int64
dtypes: bool(3), float64(4), int64(6), object(10)
memory usage: 333.0+ KB
```

```
In [42]: #show few rows for checking result
         Merge_df.head()
```

Out[42]:

	tweet_id	timestamp	source
0	892420643555336193	2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" r...
1	892177421306343426	2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" r...
2	891815181378084864	2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" r...
3	891689557279858688	2017-07-30 15:58:51 +0000	
4	891327558926688256	2017-07-29 16:00:24 +0000	

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

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

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

   rating_denominator      name  rate dog_stage       ...         p1_conf  \
0                  10   Phineas   1.3      None       ...        0.097049
1                  10     Tilly   1.3      None       ...        0.323581
2                  10    Archie   1.2      None       ...        0.716012
3                  10     Darla   1.3      None       ...        0.170278
4                  10  Franklin   1.2      None       ...        0.555712

   p1_dog                  p2   p2_conf  p2_dog                         p3  \
0   False               bagel  0.085851   False                     banana
1    True            Pekinese  0.090647    True                   papillon
2    True            malamute  0.078253    True                     kelpie
3   False  Labrador_retriever  0.168086    True                    spatula
4    True    English_springer  0.225770    True  German_short-haired_pointer

    p3_conf  p3_dog retweet_count  favorite_count
0  0.076110   False          8853           39467
1  0.068957    True          6514           33819
2  0.031379    True          4328           25461
3  0.040836   False          8964           42908
4  0.175219    True          9774           41048

[5 rows x 23 columns]
```

1.3.8 Quality:

1.3.9 Issue #7, #8 and #9: Erroneous datatypes (timestamp should be a datetime, tweet_id should be a string)

Define

- Because we already merged 3 table into one. So we can ignore issue 9 and 10 (change tweet_id datatype in `image` and `tweets`.

- Change the timestamp datatype to datetime and then change the tweet_id datatype to string in `Merge_df`

Code

```
In [43]: #Change the timestamp datatype to datetime
         Merge_df.timestamp = pd.to_datetime(Merge_df.timestamp)

In [44]: #Change the tweet_id datatype to string
         Merge_df.tweet_id = Merge_df.tweet_id.astype(str)
```

Test

```
In [45]: #check the data types - timestamp should be datetime and tweet_id should be string (obj
         Merge_df.dtypes

Out[45]: tweet_id                       object
         timestamp             datetime64[ns]
         source                         object
         text                           object
         expanded_urls                  object
         rating_numerator                int64
         rating_denominator              int64
         name                           object
         rate                          float64
         dog_stage                      object
         jpg_url                        object
         img_num                         int64
         p1                             object
         p1_conf                       float64
         p1_dog                           bool
         p2                             object
         p2_conf                       float64
         p2_dog                           bool
         p3                             object
         p3_conf                       float64
         p3_dog                           bool
         retweet_count                   int64
         favorite_count                  int64
         dtype: object
```

1.3.10 Issue #10: Contains html tags in Source columns

Define Remove all the string starts with the "<" and ends with the ">"

Code

```
In [46]: #Check the source value first.
         Merge_df.source.value_counts()
```

```
Out[46]: <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
         <a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>
         <a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>
         Name: source, dtype: int64

In [47]: #Remove all the string starts with the "<" and ends with the ">"
         Merge_df.source = Merge_df.source.str.replace(r'<[^>]+>', '', regex=True)
```

Test

```
In [48]: #check the source value again.
         Merge_df.source.value_counts()

Out[48]: Twitter for iPhone    1955
         Twitter Web Client      28
         TweetDeck               11
         Name: source, dtype: int64
```

1.3.11 Issue #11: Some names in name column are not a real name.

Define The valid name's first letter should be uppercase, so replace all the lowercase string in name to "None"

Code

```
In [49]: #replace all the lowercase string in name to "None"
         Merge_df.loc[Merge_df.name.str.islower(),'name'] = 'None'
```

Test

```
In [50]: #show a few rows to check
         Merge_df.sample(20)

Out[50]:                tweet_id           timestamp              source  \
         231   835574547218894849 2017-02-25 19:37:50  Twitter for iPhone
         1027  706346369204748288 2016-03-06 05:11:12  Twitter for iPhone
         662   758467244762497024 2016-07-28 01:00:57  Twitter for iPhone
         925   716439118184652801 2016-04-03 01:36:11  Twitter for iPhone
         650   759923798737051648 2016-08-01 01:28:46  Twitter for iPhone
         522   783334639985389568 2016-10-04 15:55:06  Twitter for iPhone
         150   854732716440526848 2017-04-19 16:25:34  Twitter for iPhone
         1551  674644256330530816 2015-12-09 17:38:19  Twitter for iPhone
         1054  704054845121142784 2016-02-28 21:25:30  Twitter for iPhone
         1728  670807719151067136 2015-11-29 03:33:17  Twitter for iPhone
         516   784431430411685888 2016-10-07 16:33:21  Twitter for iPhone
         1219  689661964914655233 2016-01-20 04:13:20  Twitter for iPhone
         1232  688908934925697024 2016-01-18 02:21:04  Twitter for iPhone
         791   741793263812808706 2016-06-12 00:44:30  Twitter for iPhone
         354   815736392542261248 2017-01-02 01:48:06  Twitter for iPhone
```

```
1442  677573743309385728 2015-12-17 19:39:03  Twitter for iPhone
543   779123168116150273 2016-09-23 01:00:13  Twitter for iPhone
1117  698355670425473025 2016-02-13 03:59:01  Twitter for iPhone
1028  706310011488698368 2016-03-06 02:46:44  Twitter for iPhone
89    871515927908634625 2017-06-04 23:56:03  Twitter for iPhone

                                                    text  \
231   This is Eli. He works backstage at Bone Jovi c...
1027  This is Koda. She's a Beneboom Cumberwiggle. 1...
662   Why does this never happen at my front door...
925   This is Bluebert. He just saw that both #Final...
650   We only rate dogs... this is a Taiwanese Guide...
522   This is Dave. He's currently in a predicament...
150   This is Marlee. She fetched a flower and immed...
1551  When you see sophomores in high school driving...
1054  Here is a whole flock of puppers.  60/50 I'll ...
1728  Say hello to Andy. He can balance on one foot,...
516   This is Stormy. He's curly af. Already pupared...
1219  Meet Luca. He's a Butternut Scooperfloof. Glor...
1232  Meet Clarence. He does parkour. 8/10 very tale...
791   When your crush won't pay attention to you. Bo...
354   This is Akumi. It's his birthday. He received ...
1442  This is Sandy. He's sexually confused. Thinks ...
543   This is Reggie. He hugs everyone he meets. 12/...
1117  This is Jessiga. She's a Tasmanian McCringlebe...
1028  Here's a very sleepy pupper. Thinks it's an ai...
89    This is Napolean. He's a Raggedy East Nicaragu...

                                          expanded_urls  rating_numerator  \
231   https://twitter.com/dog_rates/status/835574547...                11
1027  https://twitter.com/dog_rates/status/706346369...                12
662   https://twitter.com/dog_rates/status/758467244...               165
925   https://twitter.com/dog_rates/status/716439118...                50
650   https://twitter.com/dog_rates/status/759923798...                10
522   https://twitter.com/dog_rates/status/783334639...                12
150   https://twitter.com/dog_rates/status/854732716...                12
1551  https://twitter.com/dog_rates/status/674644256...                11
1054  https://twitter.com/dog_rates/status/704054845...                60
1728  https://twitter.com/dog_rates/status/670807719...                11
516   https://twitter.com/dog_rates/status/784431430...                12
1219  https://twitter.com/dog_rates/status/689661964...                12
1232  https://twitter.com/dog_rates/status/688908934...                 8
791   https://twitter.com/dog_rates/status/741793263...                10
354   https://twitter.com/dog_rates/status/815736392...                11
1442  https://twitter.com/dog_rates/status/677573743...                10
543   https://twitter.com/dog_rates/status/779123168...                12
1117  https://twitter.com/dog_rates/status/698355670...                10
1028  https://twitter.com/dog_rates/status/706310011...                12
```

```
89    https://twitter.com/dog_rates/status/871515927...               12

      rating_denominator      name  rate dog_stage       ...          p1_conf  \
231                   10       Eli   1.1      None       ...         0.610655
1027                  10      Koda   1.2      None       ...         0.956462
662                  150      None   1.1      None       ...         0.436377
925                   50   Bluebert   1.0      None       ...         0.396495
650                   10      None   1.0      None       ...         0.324579
522                   10      Dave   1.2      None       ...         0.593858
150                   10    Marlee   1.2      None       ...         0.695548
1551                  10      None   1.1      None       ...         0.398102
1054                  50      None   1.2      None       ...         0.667939
1728                  10      Andy   1.1      None       ...         0.958035
516                   10    Stormy   1.2      None       ...         0.744819
1219                  10      Luca   1.2      None       ...         0.322818
1232                  10   Clarence   0.8      None       ...         0.158859
791                   10      None   1.0      None       ...         0.311325
354                   10     Akumi   1.1      None       ...         0.548907
1442                  10     Sandy   1.0      None       ...         0.535070
543                   10    Reggie   1.2      None       ...         0.431080
1117                  10   Jessiga   1.0      None       ...         0.990191
1028                  10      None   1.2    pupper       ...         0.698165
89                    10   Napolean   1.2     doggo       ...         0.974781

      p1_dog                         p2   p2_conf  p2_dog  \
231     True                    muzzle  0.132138   False
1027    True                Rottweiler  0.025381    True
662     True                 Chihuahua  0.113956    True
925     True                  malamute  0.317053    True
650     True                 seat_belt  0.109168   False
522     True          Shetland_sheepdog  0.130611    True
150     True                  Cardigan  0.058902    True
1551   False                    basset  0.335692    True
1054    True                    kuvasz  0.228764    True
1728    True           Sealyham_terrier  0.013892    True
516     True                 toy_poodle  0.243192    True
1219    True                   whippet  0.246966    True
1232   False                      pier  0.130016   False
791     True             French_bulldog  0.115349    True
354     True                  Cardigan  0.178523    True
1442   False              folding_chair  0.080419   False
543     True  soft-coated_wheaten_terrier  0.060365    True
1117    True                  Pekinese  0.002799    True
1028    True                 Chihuahua  0.105834    True
89      True                    briard  0.020041    True

                                 p3   p3_conf  p3_dog retweet_count  \
231   American_Staffordshire_terrier  0.109544    True          4121
```

```
1027                      Appenzeller  0.008679    True          1035
662   American_Staffordshire_terrier  0.099689    True          2539
925                      Eskimo_dog  0.273419    True           247
650                             pug  0.102466    True          6521
522                        Pembroke  0.100842    True         13616
150                            chow  0.028411    True          6690
1551                 cocker_spaniel  0.072941    True           311
1054               golden_retriever  0.043885    True          1028
1728                  Border_collie  0.004601    True           546
516                 standard_poodle  0.010920    True          1491
1219                      Chihuahua  0.122541    True          1052
1232                      bell_cote  0.087741   False           874
791              Labrador_retriever  0.068533    True          1698
354                          collie  0.146351    True          2625
1442                 parallel_bars  0.034796   False           819
543                  cocker_spaniel  0.059845    True          4207
1117                     sunglasses  0.001310   False           516
1028                     bloodhound  0.062030    True          9034
89                             swab  0.003228   False          3628

      favorite_count
231            19447
1027            3768
662             5316
925             2574
650            16284
522            32651
150            24188
1551            1111
1054            3201
1728            1234
516             6329
1219            3501
1232            2310
791             4982
354            10937
1442            2322
543            13206
1117            2046
1028           23443
89             20730

[20 rows x 23 columns]
```

Noticed another issue when look at the sample above: tweet_id 776201521193218049: the name should be O'Malley instead of O

```
In [51]: Merge_df[Merge_df.name.str.len() == 1]
```

```
Out[51]:              tweet_id          timestamp              source  \
         561  776201521193218049 2016-09-14 23:30:38  Twitter for iPhone

                                                              text  \
         561  This is O'Malley. That is how he sleeps. Doesn...

                                               expanded_urls  rating_numerator  \
         561  https://twitter.com/dog_rates/status/776201521...                10

              rating_denominator name  rate dog_stage       ...         p1_conf  \
         561                  10    O   1.0      None       ...        0.502228

              p1_dog                        p2   p2_conf  p2_dog          p3   p3_conf  \
         561    True  black-and-tan_coonhound  0.154594    True  bloodhound  0.135176

              p3_dog retweet_count  favorite_count
         561    True          2919           10681

         [1 rows x 23 columns]
```

1.3.12 Issue #12: tweet_id 776201521193218049: the name should be O'Malley instead of O

Define Change row of tweet_id 776201521193218049's name from O to O'Malley

Code

```
In [52]: #Change row of tweet_id 776201521193218049's name from O to O'Malley
         Merge_df.loc[Merge_df.tweet_id == '776201521193218049', 'name'] = "O'Malley"
```

Test

```
In [53]: #Check if there still exist a name which has 1 letter
         Merge_df[Merge_df.name.str.len() == 1]
```

```
Out[53]: Empty DataFrame
         Columns: [tweet_id, timestamp, source, text, expanded_urls, rating_numerator, rating_de
         Index: []

         [0 rows x 23 columns]
```

```
In [54]: #check if the row of tweet_id 776201521193218049's name has changed
         Merge_df[Merge_df.tweet_id == '776201521193218049']
```

```
Out[54]:              tweet_id          timestamp              source  \
         561  776201521193218049 2016-09-14 23:30:38  Twitter for iPhone

                                                              text  \
         561  This is O'Malley. That is how he sleeps. Doesn...
```

```
                                              expanded_urls  rating_numerator  \
561  https://twitter.com/dog_rates/status/776201521...                 10

     rating_denominator      name  rate dog_stage       ...          p1_conf  \
561                  10  O'Malley   1.0      None       ...         0.502228

     p1_dog                       p2   p2_conf  p2_dog          p3   p3_conf  \
561    True  black-and-tan_coonhound  0.154594    True  bloodhound  0.135176

     p3_dog retweet_count  favorite_count
561    True          2919           10681

[1 rows x 23 columns]
```

1.3.13 Issue #13: doggo, pupper, puppo, floofer should be category type

Define Convert dog_stage to categorical data types

Code

```
In [55]: #Convert dog_stage to categorical data types
         Merge_df.dog_stage = Merge_df.dog_stage.astype('category')
```

Test

```
In [56]: #show the data information to check if the dog_stage data types is category
         Merge_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1994 entries, 0 to 1993
Data columns (total 23 columns):
tweet_id              1994 non-null object
timestamp             1994 non-null datetime64[ns]
source                1994 non-null object
text                  1994 non-null object
expanded_urls         1994 non-null object
rating_numerator      1994 non-null int64
rating_denominator    1994 non-null int64
name                  1994 non-null object
rate                  1994 non-null float64
dog_stage             1994 non-null category
jpg_url               1994 non-null object
img_num               1994 non-null int64
p1                    1994 non-null object
p1_conf               1994 non-null float64
p1_dog                1994 non-null bool
p2                    1994 non-null object
p2_conf               1994 non-null float64
```

```
p2_dog                1994 non-null bool
p3                    1994 non-null object
p3_conf               1994 non-null float64
p3_dog                1994 non-null bool
retweet_count         1994 non-null int64
favorite_count        1994 non-null int64
dtypes: bool(3), category(1), datetime64[ns](1), float64(4), int64(5), object(9)
memory usage: 319.7+ KB
```

1.3.14 Issue #14: rating_numerator of 1776 is too oulier for the rest of data

Define Remove the row which has the rating_numerator of 1776

Code

```
In [57]: #Remove the row which has the rating_numerator of 1776
         Merge_df = Merge_df[Merge_df['rating_numerator'] != 1776]
```

Test

```
In [58]: #check if the row which has rating_numertor of 1776 is removed.
         Merge_df[Merge_df['rating_numerator'] ==1776]
```

```
Out[58]: Empty DataFrame
         Columns: [tweet_id, timestamp, source, text, expanded_urls, rating_numerator, rating_de
         Index: []

         [0 rows x 23 columns]
```

```
In [59]: #check for the number of rows after remove 1 row- should be 2072
         Merge_df.shape[0]
```

```
Out[59]: 1993
```

1.3.15 Issue #15: Values in columns 'p1', 'p2', and 'p3' don't have consistent format

Define Change p1, p2, p3 to lowercase

Code

```
In [60]: Merge_df['p1'] = Merge_df['p1'].str.lower()
         Merge_df['p2'] = Merge_df['p2'].str.lower()
         Merge_df['p3'] = Merge_df['p3'].str.lower()
```

Test

In [61]:
```
# Display few rows to check whether names are all lowercase
Merge_df.head()
```

Out[61]:

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

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

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

   rating_denominator      name  rate dog_stage       ...         p1_conf  \
0                  10   Phineas   1.3      None       ...        0.097049
1                  10     Tilly   1.3      None       ...        0.323581
2                  10    Archie   1.2      None       ...        0.716012
3                  10     Darla   1.3      None       ...        0.170278
4                  10  Franklin   1.2      None       ...        0.555712

   p1_dog                  p2   p2_conf  p2_dog                         p3  \
0   False               bagel  0.085851   False                     banana
1    True            pekinese  0.090647    True                   papillon
2    True            malamute  0.078253    True                     kelpie
3   False  labrador_retriever  0.168086    True                    spatula
4    True    english_springer  0.225770    True  german_short-haired_pointer

    p3_conf  p3_dog retweet_count  favorite_count
0  0.076110   False          8853           39467
1  0.068957    True          6514           33819
2  0.031379    True          4328           25461
3  0.040836   False          8964           42908
4  0.175219    True          9774           41048

[5 rows x 23 columns]
```

```
In [62]: #check if the all the names of p1 are lowercase - should be 2072
         Merge_df.p1.str.islower().count()
```

```
Out[62]: 1993
```

```
In [63]: #check if the all the names of p2 are lowercase - should be 2072
         Merge_df.p2.str.islower().count()
```

```
Out[63]: 1993
```

```
In [64]: #check if the all the names of p3 are lowercase - should be 2072
         Merge_df.p3.str.islower().count()
```

```
Out[64]: 1993
```

1.4 Storing Data

Save gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv".

```
In [65]: # Saving the master dataset to a csv file
         Merge_df.to_csv("twitter_archive_master.csv", index=False)
```

1.5 Analyzing and Visualizing Data

In this section, analyze and visualize your wrangled data. You must produce at least **three (3) insights and one (1) visualization.**

```
In [66]: #show statistic description of the dataset
         Merge_df.describe()
```

```
Out[66]:        rating_numerator  rating_denominator         rate      img_num  \
         count       1993.000000         1993.000000  1993.000000  1993.000000
         mean          11.395886           10.532363     1.080724     1.203211
         std           12.670536            7.322538     0.956618     0.560899
         min            0.000000            2.000000     0.000000     1.000000
         25%           10.000000           10.000000     1.000000     1.000000
         50%           11.000000           10.000000     1.100000     1.000000
         75%           12.000000           10.000000     1.200000     1.000000
         max          420.000000          170.000000    42.000000     4.000000

                     p1_conf       p2_conf       p3_conf  retweet_count  \
         count  1993.000000  1.993000e+03  1.993000e+03    1993.000000
         mean      0.593971  1.344463e-01  6.025323e-02    2766.750627
         std       0.272019  1.006988e-01  5.090300e-02    4675.871667
         min       0.044333  1.011300e-08  1.740170e-10      16.000000
         25%       0.362835  5.390140e-02  1.619070e-02     624.000000
         50%       0.587764  1.175080e-01  4.948690e-02    1359.000000
         75%       0.846628  1.952180e-01  9.160200e-02    3220.000000
         max       1.000000  4.880140e-01  2.734190e-01   79515.000000

                favorite_count
         count     1993.000000
         mean      8897.394882
         std      12216.030848
         min         81.000000
         25%       1981.000000
         50%       4134.000000
         75%      11310.000000
         max     132810.000000
```

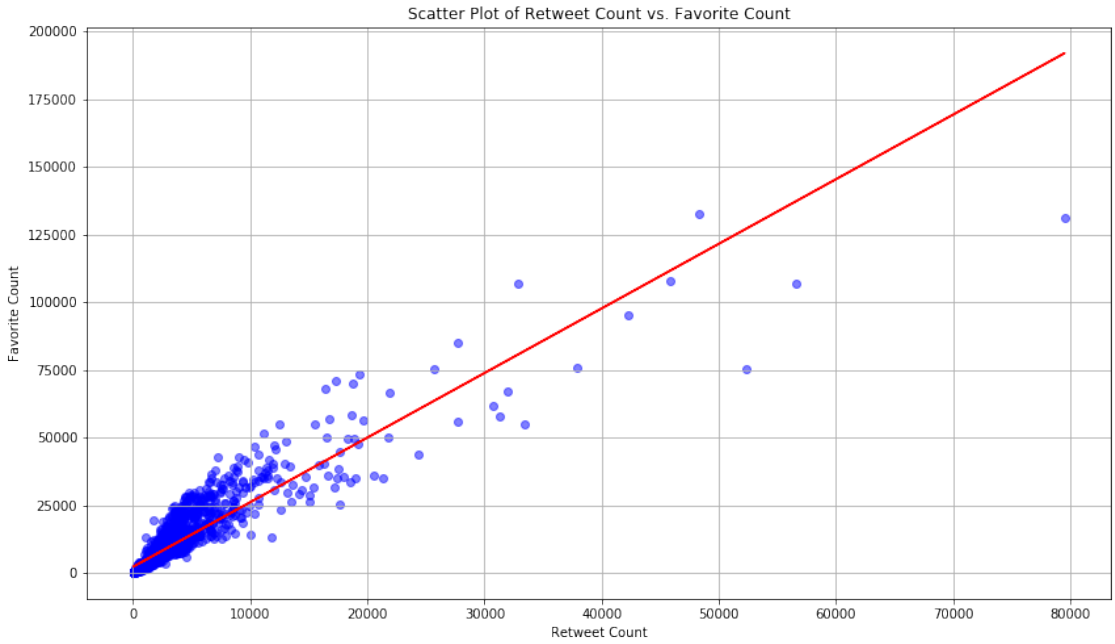
1.5.1 Insight 1: Rates and Confidence Intervals

- Rating distribution: average calculated rating is 1.08, minimum is 0 and maximum is 42. The majority of ratings have denominator of 10, the median of rating_numerator is 11
- The highest confidence is the first prediction (maximum is 100%, minimum is 4.43%, average is 59.39%), second is the second prediction (maximum is 48.8%, minimum is ~0%, and average is 13.44%), the least confidence is third prediction(maximum is 27.34%, minimum is ~0%, average is 6.025%)
- The average favorite count is 8557, range from 0 to 132810, the average retweet count is 2976, range from 16 to 79515.

1.5.2 Insight 2: The correlation between retweet count and favorite count

Visualization

```
In [67]: # Add retweet_count data to x-axis and favorite_count data to y-axis
         x = Merge_df['retweet_count']
         y = Merge_df['favorite_count']
         # Create a scatter plot to compare retweet count and favorite count correlation.
         plt.figure(figsize=(14, 8))
         plt.scatter(x, y, alpha=0.5, color='b')
         # Add a correlation line
         z = np.polyfit(x, y, 1)
         p = np.poly1d(z)
         plt.plot(x, p(x), color='red')
         plt.title('Scatter Plot of Retweet Count vs. Favorite Count')
         #Add title and labels
         plt.xlabel('Retweet Count')
         plt.ylabel('Favorite Count')
         plt.grid(True)
         plt.show()
```




```
In [68]: #Calculate the correlation number between Retweet Count vs. Favorite Count
         Merge_df['retweet_count'].corr(Merge_df['favorite_count'])
```

```
Out[68]: 0.91296378371603659
```

Reasoning: This scatter plot will help us understand whether higher audience engagement leads to higher revenues

Summary: Based on the scatter plot and the calculated correlation number result (0.91296378371603659), it shows the positive correlation between Retweet Count and Favorite Count, suggesting that tweets with more retweet tend to be the favorite.

1.5.3 Insight 3: What is the most popular dog stage?

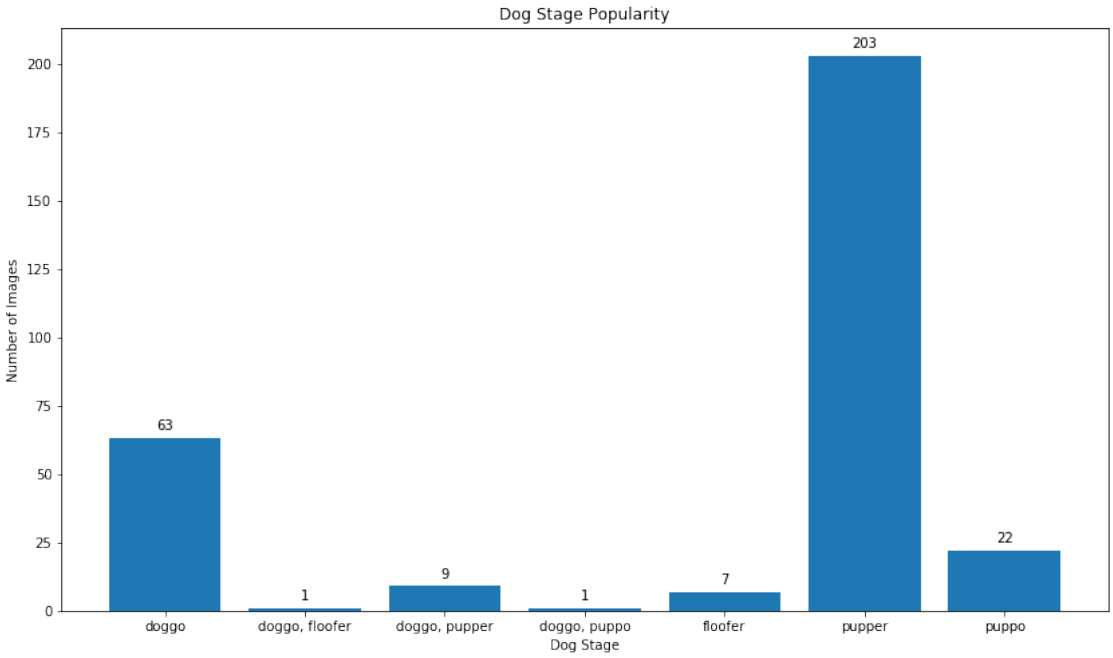
```
In [69]: #Count the values in dog stage column and put it in dog_stage_count dataset (excluding
         dog_stage_count = Merge_df[Merge_df['dog_stage'] != 'None']
```

```
In [70]: # Count each value of the dog_stage
         dog_stage_count= dog_stage_count['dog_stage'].str.strip().value_counts()
```

Visualization

```
In [71]: plt.figure(figsize=(14, 8))
         #create the bar plot
         plt.bar(dog_stage_count.index.tolist(),dog_stage_count)
         # Display count values on each bar
         for stage, count in zip(dog_stage_count.index.tolist(), dog_stage_count):
             plt.text(stage, count + 2, str(count), ha='center', va='bottom')
```

```
plt.title('Dog Stage Popularity')
plt.xlabel('Dog Stage')
plt.ylabel('Number of Images')
plt.show()
```



Reasoning: This bar plot will help us identify which dog stage is the most popular

Summary: The result shows that user usually sends the image of "pupper" dog stage. The highest count is pupper (210), and the smallest counts are "doggo, floofer" and "doggo,puppo" (1)

1.5.4 Analysis Decision 1: Rate 's Outlier Removal

Reasoning: Because the maximum of `rating_numerator` is still 420 after I removed 1776, which is still an extreme outlier compared to the other values. Removing such extreme values could improve the analysis.

Count all the rating_numerator points:

```
In [72]: #show the result of counting all the rating_numerator point
         Merge_df.rating_numerator.value_counts()
```

```
Out[72]: 12     450
         10     419
         11     396
         13     261
         9      151
         8       95
         7       52
         14      35
```

```
5       33
6       32
3       19
4       16
2        9
1        5
0        2
24       1
420      1
204      1
27       1
44       1
45       1
50       1
60       1
75       1
80       1
84       1
88       1
99       1
121      1
143      1
144      1
165      1
26       1
Name: rating_numerator, dtype: int64
```

All the rating_numerators which are higher than 20 is just one unit per rating_numerator point, which are significantly smaller than the majority of values. So I decide to remove all the points which are greater than 20.

```
In [73]: # remove all the rows which are greater than 20 of rating numerator points
         Filtered_df = Merge_df[Merge_df['rating_numerator'] < 15]

In [74]: Filtered_df.describe()
```

Out[74]:

	rating_numerator	rating_denominator	rate	img_num	\
count	1975.000000	1975.000000	1975.000000	1975.000000	
mean	10.538228	10.002025	1.053850	1.204051	
std	2.200311	0.289981	0.219777	0.562582	
min	0.000000	2.000000	0.000000	1.000000	
25%	10.000000	10.000000	1.000000	1.000000	
50%	11.000000	10.000000	1.100000	1.000000	
75%	12.000000	10.000000	1.200000	1.000000	
max	14.000000	20.000000	1.400000	4.000000	

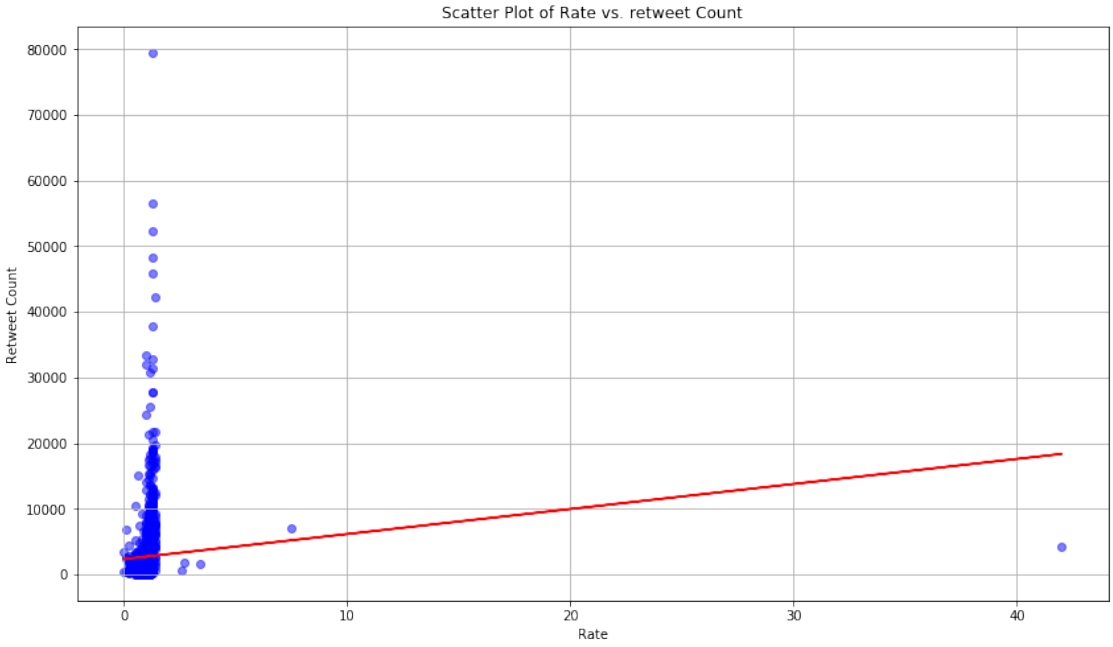
	p1_conf	p2_conf	p3_conf	retweet_count	favorite_count
count	1975.000000	1.975000e+03	1.975000e+03	1975.000000	1975.000000
mean	0.593799	1.348376e-01	6.022196e-02	2775.489114	8929.860759

std	0.272080	1.007948e-01	5.081977e-02	4693.558994	12259.050949
min	0.044333	1.011300e-08	1.740170e-10	16.000000	81.000000
25%	0.360447	5.417505e-02	1.608055e-02	622.500000	1956.500000
50%	0.587764	1.186220e-01	4.948690e-02	1350.000000	4138.000000
75%	0.844583	1.955655e-01	9.164355e-02	3224.500000	11378.500000
max	1.000000	4.880140e-01	2.710420e-01	79515.000000	132810.000000

1.5.5 Visualization

The Correlation relationship between rate and retweet_count before removing outliers

In [75]:
```
# Add rate data to x-axis and retweet_count data to y-axis
x = Merge_df['rate']
y = Merge_df['retweet_count']
# Create a scatter plot to compare retweet count and retweet count correlation.
plt.figure(figsize=(14, 8))
plt.scatter(x, y, alpha=0.5, color='b')
# Add a correlation line
z = np.polyfit(x, y, 1)
p = np.poly1d(z)
plt.plot(x, p(x), color='red')
plt.title('Scatter Plot of Rate vs. retweet Count')
#Add title and labels
plt.xlabel('Rate')
plt.ylabel('Retweet Count')
plt.grid(True)
plt.show()
```

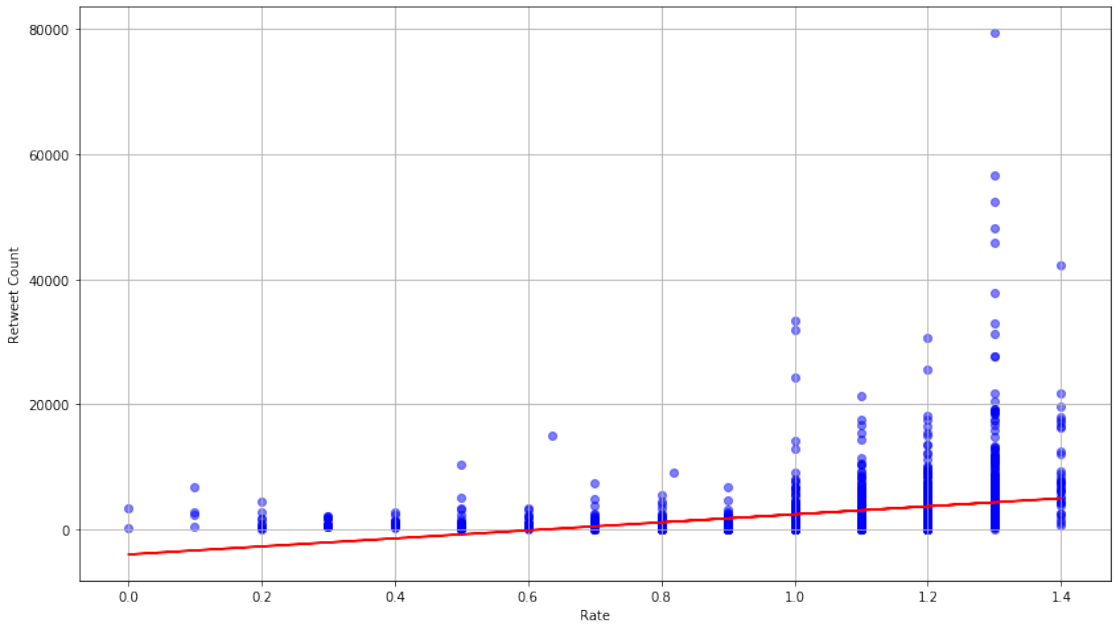


```
In [76]: #Calculate the correlation number between Retweet Count vs. rate
         Merge_df['rate'].corr(Merge_df['retweet_count'])
```

```
Out[76]: 0.077998939766817904
```

The Correlation relationship between rate and retweet_count after removing outliers

```
In [77]: # Add rate data to x-axis and retweet_count data to y-axis
         x = Filtered_df['rate']
         y = Filtered_df['retweet_count']
         # Create a scatter plot to compare retweet count and rate correlation.
         plt.figure(figsize=(14, 8))
         plt.scatter(x, y, alpha=0.5, color='b')
         # Add a correlation line
         z = np.polyfit(x, y, 1)
         p = np.poly1d(z)
         plt.plot(x, p(x), color='red')
         #Add title and labels
         plt.xlabel('Rate')
         plt.ylabel('Retweet Count')
         plt.grid(True)
         plt.show()
```




```
In [78]: Filtered_df['rate'].corr(Filtered_df['retweet_count'])
```

```
Out[78]: 0.30025872122183062
```

Summary: Although the correlation relationship between rate and retweet_count is still a weak positive after I remove the outliers (0.077998939766817904), it still has a significant improvement when comparing the correlation before removing the outliers. The correlation after removing outliers(0.30025872122183062) is 4 times stronger than before removing outliers (0.077998939766817904)

In [ ]: