wrangle_act

May 4, 2021

0.1 Gather

```
In [1]: import pandas as pd
        df_1 = pd.read_csv('twitter-archive-enhanced.csv')
        df_1.head()
```

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

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

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

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

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

```
3                       NaN                         NaN
4                       NaN                         NaN

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

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

In [2]:
```
import tweepy
from tweepy import OAuthHandler
import json
from timeit import default_timer as timer

if False:
    consumer_key = 'HIDDEN'
    consumer_secret = 'HIDDEN'
    access_token = 'HIDDEN'
    access_secret = 'HIDDEN'

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

    api = tweepy.API(auth, wait_on_rate_limit=True)

    tweet_ids = df_1.tweet_id.values
    len(tweet_ids)

    # Query Twitter's API for JSON data for each tweet ID in the Twitter archive
    count = 0
    fails_dict = {}
    start = timer()
    # Save each tweet's returned JSON as a new line in a .txt file
    with open('tweet_json.txt', 'w') as outfile:
        # This loop will likely take 20-30 minutes to run because of Twitter's rate limi
        for tweet_id in tweet_ids:
            count += 1
            print(str(count) + ": " + str(tweet_id))
            try:
                tweet = api.get_status(tweet_id, tweet_mode='extended')
```

```
                    print("Success")
                    json.dump(tweet._json, outfile)
                    outfile.write('\n')
                except tweepy.TweepError as e:
                    print("Fail")
                    fails_dict[tweet_id] = e
                    pass
        end = timer()
        print(end - start)
        print(fails_dict)
```

In [3]:
```
df_image=pd.read_csv('image-predictions.tsv',sep='\t')
```

In [4]:
```
with open('tweet_json.txt', 'r') as file:
    tweet_json = pd.read_json('tweet_json.txt', lines=True)
```

In [5]:
```
tweet_json
```

Out[5]:

	contributors	coordinates	created_at	display_text_range	\
0	NaN	NaN	2017-08-01 16:23:56	[0, 85]	
1	NaN	NaN	2017-08-01 00:17:27	[0, 138]	
2	NaN	NaN	2017-07-31 00:18:03	[0, 121]	
3	NaN	NaN	2017-07-30 15:58:51	[0, 79]	
4	NaN	NaN	2017-07-29 16:00:24	[0, 138]	
5	NaN	NaN	2017-07-29 00:08:17	[0, 138]	
6	NaN	NaN	2017-07-28 16:27:12	[0, 140]	
7	NaN	NaN	2017-07-28 00:22:40	[0, 118]	
8	NaN	NaN	2017-07-27 16:25:51	[0, 122]	
9	NaN	NaN	2017-07-26 15:59:51	[0, 133]	
10	NaN	NaN	2017-07-26 00:31:25	[0, 130]	
11	NaN	NaN	2017-07-25 16:11:53	[0, 107]	
12	NaN	NaN	2017-07-25 01:55:32	[0, 106]	
13	NaN	NaN	2017-07-25 00:10:02	[0, 91]	
14	NaN	NaN	2017-07-24 17:02:04	[0, 118]	
15	NaN	NaN	2017-07-24 00:19:32	[0, 138]	
16	NaN	NaN	2017-07-23 00:22:39	[0, 86]	
17	NaN	NaN	2017-07-22 16:56:37	[0, 128]	
18	NaN	NaN	2017-07-22 00:23:06	[0, 87]	
19	NaN	NaN	2017-07-20 16:49:33	[0, 127]	
20	NaN	NaN	2017-07-19 16:06:48	[0, 127]	
21	NaN	NaN	2017-07-19 03:39:09	[0, 108]	
22	NaN	NaN	2017-07-19 00:47:34	[0, 99]	
23	NaN	NaN	2017-07-18 16:08:03	[0, 88]	
24	NaN	NaN	2017-07-18 00:07:08	[0, 129]	
25	NaN	NaN	2017-07-17 16:17:36	[0, 101]	
26	NaN	NaN	2017-07-16 23:58:41	[0, 121]	
27	NaN	NaN	2017-07-16 20:14:00	[0, 71]	
28	NaN	NaN	2017-07-15 23:25:31	[0, 131]	
29	NaN	NaN	2017-07-15 16:51:35	[27, 105]	

```
...            ...          ...                  ...                ...
2324           NaN          NaN 2015-11-17 00:24:19           [0, 140]
2325           NaN          NaN 2015-11-17 00:06:54           [0, 139]
2326           NaN          NaN 2015-11-16 23:23:41           [0, 137]
2327           NaN          NaN 2015-11-16 21:54:18            [0, 81]
2328           NaN          NaN 2015-11-16 21:10:36           [0, 140]
2329           NaN          NaN 2015-11-16 20:32:58           [0, 135]
2330           NaN          NaN 2015-11-16 20:01:42           [0, 112]
2331           NaN          NaN 2015-11-16 19:31:45           [0, 139]
2332           NaN          NaN 2015-11-16 16:37:02           [0, 138]
2333           NaN          NaN 2015-11-16 16:11:11           [0, 136]
2334           NaN          NaN 2015-11-16 15:14:19            [0, 46]
2335           NaN          NaN 2015-11-16 14:57:41            [0, 82]
2336           NaN          NaN 2015-11-16 04:02:55           [0, 134]
2337           NaN          NaN 2015-11-16 03:55:04           [0, 128]
2338           NaN          NaN 2015-11-16 03:44:34           [0, 140]
2339           NaN          NaN 2015-11-16 03:22:39           [0, 132]
2340           NaN          NaN 2015-11-16 02:38:37           [0, 125]
2341           NaN          NaN 2015-11-16 01:59:36           [0, 137]
2342           NaN          NaN 2015-11-16 01:52:02           [0, 137]
2343           NaN          NaN 2015-11-16 01:22:45           [0, 107]
2344           NaN          NaN 2015-11-16 01:01:59           [0, 135]
2345           NaN          NaN 2015-11-16 00:55:59           [0, 124]
2346           NaN          NaN 2015-11-16 00:49:46           [0, 140]
2347           NaN          NaN 2015-11-16 00:35:11           [0, 138]
2348           NaN          NaN 2015-11-16 00:30:50           [0, 140]
2349           NaN          NaN 2015-11-16 00:24:50           [0, 120]
2350           NaN          NaN 2015-11-16 00:04:52           [0, 137]
2351           NaN          NaN 2015-11-15 23:21:54           [0, 130]
2352           NaN          NaN 2015-11-15 23:05:30           [0, 139]
2353           NaN          NaN 2015-11-15 22:32:08           [0, 131]

                                                    entities  \
0     {'hashtags': [], 'symbols': [], 'user_mentions...
1     {'hashtags': [], 'symbols': [], 'user_mentions...
2     {'hashtags': [], 'symbols': [], 'user_mentions...
3     {'hashtags': [], 'symbols': [], 'user_mentions...
4     {'hashtags': [{'text': 'BarkWeek', 'indices': ...
5     {'hashtags': [{'text': 'BarkWeek', 'indices': ...
6     {'hashtags': [], 'symbols': [], 'user_mentions...
7     {'hashtags': [], 'symbols': [], 'user_mentions...
8     {'hashtags': [{'text': 'BarkWeek', 'indices': ...
9     {'hashtags': [], 'symbols': [], 'user_mentions...
10    {'hashtags': [{'text': 'BarkWeek', 'indices': ...
11    {'hashtags': [], 'symbols': [], 'user_mentions...
12    {'hashtags': [], 'symbols': [], 'user_mentions...
13    {'hashtags': [], 'symbols': [], 'user_mentions...
14    {'hashtags': [{'text': 'BarkWeek', 'indices': ...
```

```
15    {'hashtags': [{'text': 'BarkWeek', 'indices': ...
16    {'hashtags': [], 'symbols': [], 'user_mentions...
17    {'hashtags': [], 'symbols': [], 'user_mentions...
18    {'hashtags': [], 'symbols': [], 'user_mentions...
19    {'hashtags': [], 'symbols': [], 'user_mentions...
20    {'hashtags': [], 'symbols': [], 'user_mentions...
21    {'hashtags': [], 'symbols': [], 'user_mentions...
22    {'hashtags': [], 'symbols': [], 'user_mentions...
23    {'hashtags': [], 'symbols': [], 'user_mentions...
24    {'hashtags': [], 'symbols': [], 'user_mentions...
25    {'hashtags': [], 'symbols': [], 'user_mentions...
26    {'hashtags': [], 'symbols': [], 'user_mentions...
27    {'hashtags': [], 'symbols': [], 'user_mentions...
28    {'hashtags': [], 'symbols': [], 'user_mentions...
29    {'hashtags': [], 'symbols': [], 'user_mentions...
...                                                 ...
2324  {'hashtags': [], 'symbols': [], 'user_mentions...
2325  {'hashtags': [], 'symbols': [], 'user_mentions...
2326  {'hashtags': [], 'symbols': [], 'user_mentions...
2327  {'hashtags': [], 'symbols': [], 'user_mentions...
2328  {'hashtags': [], 'symbols': [], 'user_mentions...
2329  {'hashtags': [], 'symbols': [], 'user_mentions...
2330  {'hashtags': [], 'symbols': [], 'user_mentions...
2331  {'hashtags': [], 'symbols': [], 'user_mentions...
2332  {'hashtags': [], 'symbols': [], 'user_mentions...
2333  {'hashtags': [], 'symbols': [], 'user_mentions...
2334  {'hashtags': [], 'symbols': [], 'user_mentions...
2335  {'hashtags': [], 'symbols': [], 'user_mentions...
2336  {'hashtags': [], 'symbols': [], 'user_mentions...
2337  {'hashtags': [], 'symbols': [], 'user_mentions...
2338  {'hashtags': [], 'symbols': [], 'user_mentions...
2339  {'hashtags': [], 'symbols': [], 'user_mentions...
2340  {'hashtags': [], 'symbols': [], 'user_mentions...
2341  {'hashtags': [], 'symbols': [], 'user_mentions...
2342  {'hashtags': [], 'symbols': [], 'user_mentions...
2343  {'hashtags': [], 'symbols': [], 'user_mentions...
2344  {'hashtags': [], 'symbols': [], 'user_mentions...
2345  {'hashtags': [], 'symbols': [], 'user_mentions...
2346  {'hashtags': [], 'symbols': [], 'user_mentions...
2347  {'hashtags': [], 'symbols': [], 'user_mentions...
2348  {'hashtags': [], 'symbols': [], 'user_mentions...
2349  {'hashtags': [], 'symbols': [], 'user_mentions...
2350  {'hashtags': [], 'symbols': [], 'user_mentions...
2351  {'hashtags': [], 'symbols': [], 'user_mentions...
2352  {'hashtags': [], 'symbols': [], 'user_mentions...
2353  {'hashtags': [], 'symbols': [], 'user_mentions...

                                      extended_entities  favorite_count  \
```

0	{'media': [{'id': 892420639486877696, 'id_str'...	39467
1	{'media': [{'id': 892177413194625024, 'id_str'...	33819
2	{'media': [{'id': 891815175371796480, 'id_str'...	25461
3	{'media': [{'id': 891689552724799489, 'id_str'...	42908
4	{'media': [{'id': 891327551943041024, 'id_str'...	41048
5	{'media': [{'id': 891087942176911360, 'id_str'...	20562
6	{'media': [{'id': 890971906207338496, 'id_str'...	12041
7	{'media': [{'id': 890729118844600320, 'id_str'...	56848
8	{'media': [{'id': 890609177319665665, 'id_str'...	28226
9	{'media': [{'id': 890240245463175168, 'id_str'...	32467
10	{'media': [{'id': 890006600089468928, 'id_str'...	31166
11	{'media': [{'id': 889880888800096258, 'id_str'...	28268
12	{'media': [{'id': 889665366129029120, 'id_str'...	38818
13	{'media': [{'id': 889638825424826374, 'id_str'...	27672
14	{'media': [{'id': 889531127467266049, 'id_str'...	15359
15	{'media': [{'id': 889278779352338437, 'id_str'...	25652
16	{'media': [{'id': 888917229776945152, 'id_str'...	29611
17	{'media': [{'id': 888804981515575296, 'id_str'...	26080
18	{'media': [{'id': 888554915546542081, 'id_str'...	20290
19	{'media': [{'id': 888078426338406400, 'id_str'...	22201
20	{'media': [{'id': 887705281597243393, 'id_str'...	30779
21	{'media': [{'id': 887517108413886465, 'id_str'...	46959
22	{'media': [{'id': 887473949361045505, 'id_str'...	69871
23	{'media': [{'id': 887343120832229379, 'id_str'...	34222
24	{'media': [{'id': 887101385971384320, 'id_str'...	31061
25	{'media': [{'id': 886983218871902208, 'id_str'...	35859
26	{'media': [{'id': 886736868116754432, 'id_str'...	12306
27	{'media': [{'id': 886680331239161856, 'id_str'...	22798
28	{'media': [{'id': 886366138128449536, 'id_str'...	21524
29	NaN	117
...	...	...
2324	{'media': [{'id': 666411498068123649, 'id_str'...	459
2325	{'media': [{'id': 666407121513275392, 'id_str'...	113
2326	{'media': [{'id': 666396240351993856, 'id_str'...	172
2327	{'media': [{'id': 666373746337402880, 'id_str'...	194
2328	{'media': [{'id': 666362717482020864, 'id_str'...	804
2329	{'media': [{'id': 666353280906170368, 'id_str'...	229
2330	{'media': [{'id': 666345414279471104, 'id_str'...	307
2331	{'media': [{'id': 666337857791987715, 'id_str'...	204
2332	{'media': [{'id': 666293909010702337, 'id_str'...	522
2333	{'media': [{'id': 666287399580733440, 'id_str'...	152
2334	{'media': [{'id': 666273081518768128, 'id_str'...	184
2335	{'media': [{'id': 666268904428277760, 'id_str'...	108
2336	{'media': [{'id': 666104129232740352, 'id_str'...	14765
2337	{'media': [{'id': 666102150364286977, 'id_str'...	81
2338	{'media': [{'id': 666099505364733952, 'id_str'...	164
2339	{'media': [{'id': 666093996847063040, 'id_str'...	169
2340	{'media': [{'id': 666082912819875840, 'id_str'...	121

```
2341  {'media': [{'id': 666073098362486784, 'id_str'...            335
2342  {'media': [{'id': 666071190449033216, 'id_str'...            154
2343  {'media': [{'id': 666063820255862784, 'id_str'...            496
2344  {'media': [{'id': 666058597072306176, 'id_str'...            115
2345  {'media': [{'id': 666057085227016192, 'id_str'...            304
2346  {'media': [{'id': 666055517517848576, 'id_str'...            448
2347  {'media': [{'id': 666051848592334848, 'id_str'...           1253
2348  {'media': [{'id': 666050754986266625, 'id_str'...            136
2349  {'media': [{'id': 666049244999131136, 'id_str'...            111
2350  {'media': [{'id': 666044217047650304, 'id_str'...            311
2351  {'media': [{'id': 666033409081393153, 'id_str'...            128
2352  {'media': [{'id': 666029276303482880, 'id_str'...            132
2353  {'media': [{'id': 666020881337073664, 'id_str'...           2535

      favorited                                          full_text  geo  \
0         False  This is Phineas. He's a mystical boy. Only eve...  NaN
1         False  This is Tilly. She's just checking pup on you...  NaN
2         False  This is Archie. He is a rare Norwegian Pouncin...  NaN
3         False  This is Darla. She commenced a snooze mid meal...  NaN
4         False  This is Franklin. He would like you to stop ca...  NaN
5         False  Here we have a majestic great white breaching ...  NaN
6         False  Meet Jax. He enjoys ice cream so much he gets ...  NaN
7         False  When you watch your owner call another dog a g...  NaN
8         False  This is Zoey. She doesn't want to be one of th...  NaN
9         False  This is Cassie. She is a college pup. Studying...  NaN
10        False  This is Koda. He is a South Australian decksha...  NaN
11        False  This is Bruno. He is a service shark. Only get...  NaN
12        False  Here's a puppo that seems to be on the fence a...  NaN
13        False  This is Ted. He does his best. Sometimes that'...  NaN
14        False  This is Stuart. He's sporting his favorite fan...  NaN
15        False  This is Oliver. You're witnessing one of his m...  NaN
16        False  This is Jim. He found a fren. Taught him how t...  NaN
17        False  This is Zeke. He has a new stick. Very proud o...  NaN
18        False  This is Ralphus. He's powering up. Attempting ...  NaN
19        False  This is Gerald. He was just told he didn't get...  NaN
20        False  This is Jeffrey. He has a monopoly on the pool...  NaN
21         True  I've yet to rate a Venezuelan Hover Wiener. Th...  NaN
22        False  This is Canela. She attempted some fancy porch...  NaN
23        False  You may not have known you needed to see this ...  NaN
24        False  This... is a Jubilant Antarctic House Bear. We...  NaN
25        False  This is Maya. She's very shy. Rarely leaves he...  NaN
26        False  This is Mingus. He's a wonderful father to his...  NaN
27        False  This is Derek. He's late for a dog meeting. 13...  NaN
28        False  This is Roscoe. Another pupper fallen victim t...  NaN
29        False  @NonWhiteHat @MayhewMayhem omg hello tanner yo...  NaN
...         ...                                                ...  ...
2324      False  This is quite the dog. Gets really excited whe...  NaN
2325      False  This is a southern Vesuvius bumblegruff. Can d...  NaN
```

```
2326      False  Oh goodness. A super rare northeast Qdoba kang...  NaN
2327      False  Those are sunglasses and a jean jacket. 11/10 ...  NaN
2328      False  Unique dog here. Very small. Lives in containe...  NaN
2329      False  Here we have a mixed Asiago from the Galápagos...  NaN
2330      False  Look at this jokester thinking seat belt laws ...  NaN
2331      False  This is an extremely rare horned Parthenon. No...  NaN
2332      False  This is a funny dog. Weird toes. Won't come do...  NaN
2333      False  This is an Albanian 3 1/2 legged  Episcopalian...  NaN
2334      False     Can take selfies 11/10 https://t.co/ws2AMaNwPW  NaN
2335      False  Very concerned about fellow dog trapped in com...  NaN
2336      False  Not familiar with this breed. No tail (weird)...  NaN
2337      False  Oh my. Here you are seeing an Adobe Setter giv...  NaN
2338      False  Can stand on stump for what seems like a while...  NaN
2339      False  This appears to be a Mongolian Presbyterian mi...  NaN
2340      False  Here we have a well-established sunblockerspan...  NaN
2341      False  Let's hope this flight isn't Malaysian (lol). ...  NaN
2342      False  Here we have a northern speckled Rhododendron...  NaN
2343      False  This is the happiest dog you will ever see. Ve...  NaN
2344      False  Here is the Rand Paul of retrievers folks! He'...  NaN
2345      False  My oh my. This is a rare blond Canadian terrie...  NaN
2346      False  Here is a Siberian heavily armored polar bear ...  NaN
2347      False  This is an odd dog. Hard on the outside but lo...  NaN
2348      False  This is a truly beautiful English Wilson Staff...  NaN
2349      False  Here we have a 1949 1st generation vulpix. Enj...  NaN
2350      False  This is a purebred Piers Morgan. Loves to Netf...  NaN
2351      False  Here is a very happy pup. Big fan of well-main...  NaN
2352      False  This is a western brown Mitsubishi terrier. Up...  NaN
2353      False  Here we have a Japanese Irish Setter. Lost eye...  NaN

                                ...                             \
0                               ...
1                               ...
2                               ...
3                               ...
4                               ...
5                               ...
6                               ...
7                               ...
8                               ...
9                               ...
10                              ...
11                              ...
12                              ...
13                              ...
14                              ...
15                              ...
16                              ...
17                              ...
```

```
18                          ...
19                          ...
20                          ...
21                          ...
22                          ...
23                          ...
24                          ...
25                          ...
26                          ...
27                          ...
28                          ...
29                          ...
...                         ...
2324                        ...
2325                        ...
2326                        ...
2327                        ...
2328                        ...
2329                        ...
2330                        ...
2331                        ...
2332                        ...
2333                        ...
2334                        ...
2335                        ...
2336                        ...
2337                        ...
2338                        ...
2339                        ...
2340                        ...
2341                        ...
2342                        ...
2343                        ...
2344                        ...
2345                        ...
2346                        ...
2347                        ...
2348                        ...
2349                        ...
2350                        ...
2351                        ...
2352                        ...
2353                        ...

      possibly_sensitive_appealable  quoted_status quoted_status_id  \
0                               0.0            NaN              NaN
1                               0.0            NaN              NaN
2                               0.0            NaN              NaN
```

3	0.0	NaN	NaN
4	0.0	NaN	NaN
5	0.0	NaN	NaN
6	0.0	NaN	NaN
7	0.0	NaN	NaN
8	0.0	NaN	NaN
9	0.0	NaN	NaN
10	0.0	NaN	NaN
11	0.0	NaN	NaN
12	0.0	NaN	NaN
13	0.0	NaN	NaN
14	0.0	NaN	NaN
15	0.0	NaN	NaN
16	0.0	NaN	NaN
17	0.0	NaN	NaN
18	0.0	NaN	NaN
19	0.0	NaN	NaN
20	0.0	NaN	NaN
21	0.0	NaN	NaN
22	0.0	NaN	NaN
23	0.0	NaN	NaN
24	0.0	NaN	NaN
25	0.0	NaN	NaN
26	0.0	NaN	NaN
27	0.0	NaN	NaN
28	0.0	NaN	NaN
29	NaN	NaN	NaN
...	...	...	...
2324	0.0	NaN	NaN
2325	0.0	NaN	NaN
2326	0.0	NaN	NaN
2327	0.0	NaN	NaN
2328	0.0	NaN	NaN
2329	0.0	NaN	NaN
2330	0.0	NaN	NaN
2331	0.0	NaN	NaN
2332	0.0	NaN	NaN
2333	0.0	NaN	NaN
2334	0.0	NaN	NaN
2335	0.0	NaN	NaN
2336	0.0	NaN	NaN
2337	0.0	NaN	NaN
2338	0.0	NaN	NaN
2339	0.0	NaN	NaN
2340	0.0	NaN	NaN
2341	0.0	NaN	NaN
2342	0.0	NaN	NaN
2343	0.0	NaN	NaN

```
2344                          0.0            NaN               NaN
2345                          0.0            NaN               NaN
2346                          0.0            NaN               NaN
2347                          0.0            NaN               NaN
2348                          0.0            NaN               NaN
2349                          0.0            NaN               NaN
2350                          0.0            NaN               NaN
2351                          0.0            NaN               NaN
2352                          0.0            NaN               NaN
2353                          0.0            NaN               NaN

      quoted_status_id_str  retweet_count  retweeted  retweeted_status  \
0                      NaN           8853      False               NaN
1                      NaN           6514      False               NaN
2                      NaN           4328      False               NaN
3                      NaN           8964      False               NaN
4                      NaN           9774      False               NaN
5                      NaN           3261      False               NaN
6                      NaN           2158      False               NaN
7                      NaN          16716      False               NaN
8                      NaN           4429      False               NaN
9                      NaN           7711      False               NaN
10                     NaN           7624      False               NaN
11                     NaN           5156      False               NaN
12                     NaN           8538      False               NaN
13                     NaN           4735      False               NaN
14                     NaN           2321      False               NaN
15                     NaN           5637      False               NaN
16                     NaN           4709      False               NaN
17                     NaN           4559      False               NaN
18                     NaN           3732      False               NaN
19                     NaN           3653      False               NaN
20                     NaN           5609      False               NaN
21                     NaN          12082      False               NaN
22                     NaN          18781      False               NaN
23                     NaN          10737      False               NaN
24                     NaN           6167      False               NaN
25                     NaN           8084      False               NaN
26                     NaN           3443      False               NaN
27                     NaN           4610      False               NaN
28                     NaN           3316      False               NaN
29                     NaN              4      False               NaN
...                    ...            ...        ...               ...
2324                   NaN            339      False               NaN
2325                   NaN             44      False               NaN
2326                   NaN             92      False               NaN
2327                   NaN            100      False               NaN
2328                   NaN            595      False               NaN
```

```
2329                   NaN             77      False               NaN
2330                   NaN            146      False               NaN
2331                   NaN             96      False               NaN
2332                   NaN            368      False               NaN
2333                   NaN             71      False               NaN
2334                   NaN             82      False               NaN
2335                   NaN             37      False               NaN
2336                   NaN           6871      False               NaN
2337                   NaN             16      False               NaN
2338                   NaN             73      False               NaN
2339                   NaN             79      False               NaN
2340                   NaN             47      False               NaN
2341                   NaN            174      False               NaN
2342                   NaN             67      False               NaN
2343                   NaN            232      False               NaN
2344                   NaN             61      False               NaN
2345                   NaN            146      False               NaN
2346                   NaN            261      False               NaN
2347                   NaN            879      False               NaN
2348                   NaN             60      False               NaN
2349                   NaN             41      False               NaN
2350                   NaN            147      False               NaN
2351                   NaN             47      False               NaN
2352                   NaN             48      False               NaN
2353                   NaN            532      False               NaN

                                                    source truncated  \
0     <a href="http://twitter.com/download/iphone" r...     False
1     <a href="http://twitter.com/download/iphone" r...     False
2     <a href="http://twitter.com/download/iphone" r...     False
3     <a href="http://twitter.com/download/iphone" r...     False
4     <a href="http://twitter.com/download/iphone" r...     False
5     <a href="http://twitter.com/download/iphone" r...     False
6     <a href="http://twitter.com/download/iphone" r...     False
7     <a href="http://twitter.com/download/iphone" r...     False
8     <a href="http://twitter.com/download/iphone" r...     False
9     <a href="http://twitter.com/download/iphone" r...     False
10    <a href="http://twitter.com/download/iphone" r...     False
11    <a href="http://twitter.com/download/iphone" r...     False
12    <a href="http://twitter.com/download/iphone" r...     False
13    <a href="http://twitter.com/download/iphone" r...     False
14    <a href="http://twitter.com/download/iphone" r...     False
15    <a href="http://twitter.com/download/iphone" r...     False
16    <a href="http://twitter.com/download/iphone" r...     False
17    <a href="http://twitter.com/download/iphone" r...     False
18    <a href="http://twitter.com/download/iphone" r...     False
19    <a href="http://twitter.com/download/iphone" r...     False
20    <a href="http://twitter.com/download/iphone" r...     False
```

```
21    <a href="http://twitter.com/download/iphone" r...     False
22    <a href="http://twitter.com/download/iphone" r...     False
23    <a href="http://twitter.com/download/iphone" r...     False
24    <a href="http://twitter.com/download/iphone" r...     False
25    <a href="http://twitter.com/download/iphone" r...     False
26    <a href="http://twitter.com/download/iphone" r...     False
27    <a href="http://twitter.com/download/iphone" r...     False
28    <a href="http://twitter.com/download/iphone" r...     False
29    <a href="http://twitter.com/download/iphone" r...     False
...                                                  ...       ...
2324  <a href="http://twitter.com/download/iphone" r...     False
2325  <a href="http://twitter.com/download/iphone" r...     False
2326  <a href="http://twitter.com/download/iphone" r...     False
2327  <a href="http://twitter.com/download/iphone" r...     False
2328  <a href="http://twitter.com/download/iphone" r...     False
2329  <a href="http://twitter.com/download/iphone" r...     False
2330  <a href="http://twitter.com/download/iphone" r...     False
2331  <a href="http://twitter.com/download/iphone" r...     False
2332  <a href="http://twitter.com/download/iphone" r...     False
2333  <a href="http://twitter.com/download/iphone" r...     False
2334  <a href="http://twitter.com/download/iphone" r...     False
2335  <a href="http://twitter.com/download/iphone" r...     False
2336  <a href="http://twitter.com/download/iphone" r...     False
2337  <a href="http://twitter.com/download/iphone" r...     False
2338  <a href="http://twitter.com/download/iphone" r...     False
2339  <a href="http://twitter.com/download/iphone" r...     False
2340  <a href="http://twitter.com/download/iphone" r...     False
2341  <a href="http://twitter.com/download/iphone" r...     False
2342  <a href="http://twitter.com/download/iphone" r...     False
2343  <a href="http://twitter.com/download/iphone" r...     False
2344  <a href="http://twitter.com/download/iphone" r...     False
2345  <a href="http://twitter.com/download/iphone" r...     False
2346  <a href="http://twitter.com/download/iphone" r...     False
2347  <a href="http://twitter.com/download/iphone" r...     False
2348  <a href="http://twitter.com/download/iphone" r...     False
2349  <a href="http://twitter.com/download/iphone" r...     False
2350  <a href="http://twitter.com/download/iphone" r...     False
2351  <a href="http://twitter.com/download/iphone" r...     False
2352  <a href="http://twitter.com/download/iphone" r...     False
2353  <a href="http://twitter.com/download/iphone" r...     False

                                                   user
0     {'id': 4196983835, 'id_str': '4196983835', 'na...
1     {'id': 4196983835, 'id_str': '4196983835', 'na...
2     {'id': 4196983835, 'id_str': '4196983835', 'na...
3     {'id': 4196983835, 'id_str': '4196983835', 'na...
4     {'id': 4196983835, 'id_str': '4196983835', 'na...
5     {'id': 4196983835, 'id_str': '4196983835', 'na...
```

```
6     {'id': 4196983835, 'id_str': '4196983835', 'na...
7     {'id': 4196983835, 'id_str': '4196983835', 'na...
8     {'id': 4196983835, 'id_str': '4196983835', 'na...
9     {'id': 4196983835, 'id_str': '4196983835', 'na...
10    {'id': 4196983835, 'id_str': '4196983835', 'na...
11    {'id': 4196983835, 'id_str': '4196983835', 'na...
12    {'id': 4196983835, 'id_str': '4196983835', 'na...
13    {'id': 4196983835, 'id_str': '4196983835', 'na...
14    {'id': 4196983835, 'id_str': '4196983835', 'na...
15    {'id': 4196983835, 'id_str': '4196983835', 'na...
16    {'id': 4196983835, 'id_str': '4196983835', 'na...
17    {'id': 4196983835, 'id_str': '4196983835', 'na...
18    {'id': 4196983835, 'id_str': '4196983835', 'na...
19    {'id': 4196983835, 'id_str': '4196983835', 'na...
20    {'id': 4196983835, 'id_str': '4196983835', 'na...
21    {'id': 4196983835, 'id_str': '4196983835', 'na...
22    {'id': 4196983835, 'id_str': '4196983835', 'na...
23    {'id': 4196983835, 'id_str': '4196983835', 'na...
24    {'id': 4196983835, 'id_str': '4196983835', 'na...
25    {'id': 4196983835, 'id_str': '4196983835', 'na...
26    {'id': 4196983835, 'id_str': '4196983835', 'na...
27    {'id': 4196983835, 'id_str': '4196983835', 'na...
28    {'id': 4196983835, 'id_str': '4196983835', 'na...
29    {'id': 4196983835, 'id_str': '4196983835', 'na...
...                                                 ...
2324  {'id': 4196983835, 'id_str': '4196983835', 'na...
2325  {'id': 4196983835, 'id_str': '4196983835', 'na...
2326  {'id': 4196983835, 'id_str': '4196983835', 'na...
2327  {'id': 4196983835, 'id_str': '4196983835', 'na...
2328  {'id': 4196983835, 'id_str': '4196983835', 'na...
2329  {'id': 4196983835, 'id_str': '4196983835', 'na...
2330  {'id': 4196983835, 'id_str': '4196983835', 'na...
2331  {'id': 4196983835, 'id_str': '4196983835', 'na...
2332  {'id': 4196983835, 'id_str': '4196983835', 'na...
2333  {'id': 4196983835, 'id_str': '4196983835', 'na...
2334  {'id': 4196983835, 'id_str': '4196983835', 'na...
2335  {'id': 4196983835, 'id_str': '4196983835', 'na...
2336  {'id': 4196983835, 'id_str': '4196983835', 'na...
2337  {'id': 4196983835, 'id_str': '4196983835', 'na...
2338  {'id': 4196983835, 'id_str': '4196983835', 'na...
2339  {'id': 4196983835, 'id_str': '4196983835', 'na...
2340  {'id': 4196983835, 'id_str': '4196983835', 'na...
2341  {'id': 4196983835, 'id_str': '4196983835', 'na...
2342  {'id': 4196983835, 'id_str': '4196983835', 'na...
2343  {'id': 4196983835, 'id_str': '4196983835', 'na...
2344  {'id': 4196983835, 'id_str': '4196983835', 'na...
2345  {'id': 4196983835, 'id_str': '4196983835', 'na...
2346  {'id': 4196983835, 'id_str': '4196983835', 'na...
```

```
2347  {'id': 4196983835, 'id_str': '4196983835', 'na...
2348  {'id': 4196983835, 'id_str': '4196983835', 'na...
2349  {'id': 4196983835, 'id_str': '4196983835', 'na...
2350  {'id': 4196983835, 'id_str': '4196983835', 'na...
2351  {'id': 4196983835, 'id_str': '4196983835', 'na...
2352  {'id': 4196983835, 'id_str': '4196983835', 'na...
2353  {'id': 4196983835, 'id_str': '4196983835', 'na...

[2354 rows x 31 columns]
```

1 Assess

Assess df_1

```
In [6]: df_1.head()
```

```
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

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

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

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

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

```
3                       NaN                       NaN
4                       NaN                       NaN

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

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

```
In [7]: df_1.info()
        # tweet id should be object
        # retweeted_status_id  should be object
        # retweeted_status_user_id should be object
        # timestamp should be date time
        # retweeted_status_timestamp should be date time
        # expanded_urls should contain only the link not html command
```

```
<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 [8]: list(df_1)
```

```
Out[8]: ['tweet_id',
         'in_reply_to_status_id',
         'in_reply_to_user_id',
         'timestamp',
         'source',
         'text',
         'retweeted_status_id',
         'retweeted_status_user_id',
         'retweeted_status_timestamp',
         'expanded_urls',
         'rating_numerator',
         'rating_denominator',
         'name',
         'doggo',
         'floofer',
         'pupper',
         'puppo']
```

```
In [9]: df_1.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]: # Check quality issue on the name column
         df_1['name'].value_counts()
         # Remove "a", "an", "the" in df_1['name']
         # Quality issue
```

```
Out[10]: None          745
         a              55
         Charlie        12
         Cooper         11
         Oliver         11
         Lucy           11
         Lola           10
         Tucker         10
         Penny          10
         Bo              9
         Winston         9
         Sadie           8
         the             8
         Bailey          7
         an              7
         Daisy           7
         Toby            7
         Buddy           7
         Bella           6
         Rusty           6
         Oscar           6
         Jack            6
         Milo            6
         Jax             6
         Stanley         6
         Dave            6
         Leo             6
         Scout           6
         Koda            6
         Alfie           5
                        ...
         Eevee           1
         Andru           1
         Rodman          1
         Beya            1
         Sundance        1
         Harvey          1
         Odin            1
         Tobi            1
```

```
Adele           1
Huck            1
Pete            1
Divine          1
Ester           1
Carbon          1
Shooter         1
Bobble          1
Sweet           1
Genevieve       1
Gordon          1
Spanky          1
Derby           1
Christoper      1
Maude           1
Gunner          1
Ivar            1
Jo              1
Jarvis          1
Tripp           1
Gabby           1
Huxley          1
Name: name, Length: 957, dtype: int64
```

```
In [11]: list(df_1['name'])
         # "his", "one" are not names
         # Quality issue
```

```
Out[11]: ['Phineas',
          'Tilly',
          'Archie',
          'Darla',
          'Franklin',
          'None',
          'Jax',
          'None',
          'Zoey',
          'Cassie',
          'Koda',
          'Bruno',
          'None',
          'Ted',
          'Stuart',
          'Oliver',
          'Jim',
          'Zeke',
          'Ralphus',
          'Canela',
```

```
'Gerald',
'Jeffrey',
'such',
'Canela',
'None',
'None',
'Maya',
'Mingus',
'Derek',
'Roscoe',
'None',
'Waffles',
'None',
'Jimbo',
'Maisey',
'None',
'Lilly',
'None',
'Earl',
'Lola',
'Kevin',
'None',
'None',
'Yogi',
'Noah',
'Bella',
'Grizzwald',
'None',
'Rusty',
'Gus',
'Stanley',
'Alfy',
'Koko',
'Rey',
'Gary',
'None',
'a',
'Elliot',
'Louis',
'None',
'Bella',
'Jesse',
'None',
'Romeo',
'None',
'Bailey',
'Duddles',
'Jack',
```

```
'Emmy',
'Steven',
'Beau',
'Snoopy',
'None',
'Shadow',
'Terrance',
'Shadow',
'Emmy',
'Aja',
'None',
'Penny',
'Dante',
'Nelly',
'Ginger',
'None',
'Benedict',
'Venti',
'Goose',
'Nugget',
'None',
'None',
'Cash',
'Coco',
'Jed',
'None',
'Sebastian',
'Walter',
'None',
'Sierra',
'Sierra',
'None',
'None',
'None',
'Monkey',
'None',
'Harry',
'Kody',
'Lassie',
'Rover',
'Napolean',
'Dawn',
'None',
'Boomer',
'None',
'None',
'Cody',
'Zoey',
```

```
'Rumble',
'Clifford',
'quite',
'Dewey',
'Stanley',
'Scout',
'Gizmo',
'Walter',
'Cooper',
'None',
'Cooper',
'None',
'Harold',
'Shikha',
'None',
'None',
'Jamesy',
'None',
'Lili',
'Jamesy',
'Coco',
'None',
'Boomer',
'Sammy',
'Nelly',
'None',
'Meatball',
'Paisley',
'Albus',
'Neptune',
'Quinn',
'Belle',
'None',
'None',
'Quinn',
'Zooey',
'Dave',
'Jersey',
'None',
'None',
'Hobbes',
'None',
'Burt',
'Lorenzo',
'None',
'Lorenzo',
'Carl',
'Jordy',
```

```
'None',
'None',
'Milky',
'Trooper',
'None',
'quite',
'None',
'Winston',
'None',
'Sophie',
'Wyatt',
'Rosie',
'Thor',
'None',
'Oscar',
'None',
'None',
'Zeke',
'Luna',
'Callie',
'None',
'None',
'None',
'Cermet',
'None',
'None',
'None',
'None',
'None',
'quite',
'George',
'None',
'Marlee',
'Arya',
'Einstein',
'None',
'None',
'Alice',
'None',
'Rumpole',
'None',
'Benny',
'Aspen',
'Jarod',
'Wiggles',
'General',
'Sailor',
'Astrid',
```

```
'None',
'None',
'Iggy',
'Snoop',
'Kyle',
'Leo',
'None',
'Riley',
'Boomer',
'None',
'Gidget',
'Noosh',
'None',
'Kevin',
'None',
'Odin',
'None',
'Jerry',
'Charlie',
'None',
'Georgie',
'Rontu',
'None',
'Cannon',
'Furzey',
'Daisy',
'None',
'Tuck',
'Barney',
'None',
'Vixen',
'None',
'Jarvis',
'None',
'None',
'None',
'Mimosa',
'Pickles',
'Bungalo',
'None',
'Brady',
'Luna',
'Charlie',
'Margo',
'None',
'Sadie',
'Hank',
'Tycho',
```

```
'Stephan',
'Charlie',
'Indie',
'Winnie',
'George',
'Bentley',
'Ken',
'Penny',
'None',
'None',
'Max',
'Dawn',
'Maddie',
'Pipsy',
'None',
'None',
'Maddie',
'None',
'Monty',
'Sojourner',
'Winston',
'None',
'Odie',
'None',
'Arlo',
'None',
'Riley',
'Walter',
'Stanley',
'Sunny',
'None',
'None',
'Daisy',
'None',
'Waffles',
'Vincent',
'Lucy',
'Clark',
'None',
'Mookie',
'Meera',
'Oliver',
'None',
'Buddy',
'Ava',
'Lucy',
'None',
'Rory',
```

```
'Eli',
'Lola',
'None',
'Ash',
'Lola',
'None',
'None',
'None',
'Tucker',
'Tobi',
'None',
'Leo',
'Chester',
'Wilson',
'Sunshine',
'None',
'Lipton',
'Bentley',
'Charlie',
'Gabby',
'Bronte',
'Poppy',
'Gidget',
'Rhino',
'None',
'Willow',
'None',
'not',
'Orion',
'Eevee',
'Charlie',
'Smiley',
'Logan',
'Moreton',
'None',
'Klein',
'Miguel',
'Emanuel',
'None',
'Kuyu',
'Daisy',
'None',
'Dutch',
'Pete',
'None',
'Scooter',
'Tucker',
'Reggie',
```

```
'Lilly',
'Kyro',
'Samson',
'Loki',
'Mia',
'Leo',
'None',
'Astrid',
'Malcolm',
'Dexter',
'Gus',
'Alfie',
'Fiona',
'one',
'Mutt',
'Bear',
'Doobert',
'Beebop',
'Alexander',
'None',
'Sailer',
'Brutus',
'Kona',
'Boots',
'Tucker',
'Ralphie',
'Phil',
'Charlie',
'Loki',
'Cupid',
'None',
'None',
'Pawnd',
'Pilot',
'None',
'None',
'Ike',
'Mo',
'Toby',
'None',
'Sweet',
'Pablo',
'Pablo',
'Bailey',
'Scooter',
'Wilson',
'None',
'Nala',
```

```
'None',
'Cash',
'Balto',
'Winston',
'Crawford',
'None',
'Wyatt',
'None',
'Albus',
'None',
'Hobbes',
'Paisley',
'None',
'Paisley',
'Gabe',
'None',
'Mattie',
'Jimison',
'Hercules',
'Duchess',
'Harlso',
'Sampson',
'Sundance',
'None',
'Luca',
'None',
'Flash',
'Finn',
'Sunny',
'None',
'None',
'Peaches',
'None',
'None',
'Oliver',
'Oliver',
'None',
'Howie',
'Jazzy',
'Anna',
'None',
'Finn',
'Bo',
'Sunny',
'Sunny',
'Bo',
'Seamus',
'Wafer',
```

```
'Bear',
'Chelsea',
'Tom',
'Moose',
'Florence',
'Autumn',
'None',
'Buddy',
'Dido',
'Eugene',
'Herschel',
'Ken',
'Strudel',
'None',
'Tebow',
'None',
'Chloe',
'Betty',
'Timber',
'Binky',
'Moose',
'Dudley',
'Comet',
'Jack',
'Larry',
'Jack',
'None',
'Levi',
'Akumi',
'Titan',
'None',
'Cooper',
'Olivia',
'Beau',
'Alf',
'Oshie',
'Bruce',
'Chubbs',
'Gary',
'Sky',
'Atlas',
'None',
'None',
'Eleanor',
'Layla',
'None',
'None',
'None',
```

```
'Toby',
'Rocky',
'Baron',
'Tyr',
'Bauer',
'Swagger',
'Sammy',
'Brandi',
'None',
'Mary',
'Moe',
'Ted',
'Halo',
'None',
'Augie',
'Craig',
'Sam',
'Hunter',
'Pavlov',
'Phil',
'Gus',
'None',
'Maximus',
'None',
'Kyro',
'Wallace',
'Ito',
'None',
'Koda',
'Seamus',
'Milo',
'None',
'Cooper',
'Ollie',
'Stephan',
'Cali',
'Lennon',
'None',
'None',
'None',
'Waffles',
'Dave',
'incredibly',
'Penny',
'Major',
'Duke',
'Reginald',
'Zeke',
```

```
'Sansa',
'Shooter',
'Django',
'None',
'Rusty',
'Bo',
'Diogi',
'None',
'None',
'Sonny',
'Philbert',
'Winston',
'Marley',
'None',
'Bailey',
'Winnie',
'Severus',
'None',
'None',
'Loki',
'None',
'Ronnie',
'None',
'Wallace',
'None',
'Milo',
'Anakin',
'Bones',
'None',
'None',
'Mauve',
'Chef',
'None',
'Sampson',
'Doc',
'Bo',
'Peaches',
'None',
'Tucker',
'Sobe',
'Longfellow',
'None',
'Jeffrey',
'Mister',
'Iroh',
'Shadow',
'Baloo',
'None',
```

```
'Stubert',
'None',
'Jack',
'None',
'None',
'Lola',
'Paull',
'None',
'Timison',
'None',
'Davey',
'Cooper',
'None',
'Cassie',
'Pancake',
'None',
'Tyrone',
'Tyr',
'Romeo',
'None',
'None',
'Snicku',
'Ruby',
'Ruby',
'None',
'None',
'Yogi',
'Daisy',
'None',
'Brody',
'Bailey',
'Rizzy',
'Mack',
'Butter',
'Nimbus',
'Laika',
'Maximus',
'Clark',
'None',
'Dobby',
'Fiona',
'Moreton',
'Dave',
'None',
'Tucker',
'Juno',
'Maude',
'Lily',
```

```
'Newt',
'Benji',
'Nida',
'None',
'Robin',
'a',
'Bailey',
'Monster',
'BeBe',
'Remus',
'None',
'None',
'Maddie',
'None',
'None',
'Levi',
'Mabel',
'Alfie',
'Misty',
'Betty',
'Happy',
'Mosby',
'Duke',
'Maggie',
'Bruce',
'Leela',
'Happy',
'Buddy',
'Ralphy',
'Eli',
'Brownie',
'Rizzy',
'None',
'Meyer',
'Stella',
'Bo',
'Lucy',
'Butter',
'mad',
'Dexter',
'None',
'Leo',
'Bo',
'None',
'Frank',
'Tonks',
'Moose',
'Lincoln',
```

```
'Carl',
'Rory',
'Oakley',
'Logan',
'None',
'Dale',
'Rizzo',
'Arnie',
'Mattie',
'None',
'Scout',
'Lucy',
'Rusty',
'Pinot',
'Dallas',
'None',
'Doc',
'Hero',
'Rusty',
'Frankie',
'Stormy',
'Reginald',
'Balto',
'Riley',
'Mairi',
'Loomis',
'Finn',
'Godi',
'Kenny',
'Dave',
'Earl',
'Cali',
'Deacon',
'Penny',
'Timmy',
'Sampson',
'Harper',
'Chipson',
'None',
'Combo',
'None',
'None',
'Oakley',
'None',
'None',
'Dash',
'Koda',
'Hercules',
```

```
'None',
'Bell',
'None',
'Bear',
'None',
'Hank',
'None',
'Scout',
'None',
'Hurley',
'Reggie',
'None',
'Jay',
'None',
'None',
'Mya',
'Strider',
'Penny',
'None',
'an',
'Nala',
'Stanley',
'None',
'Sophie',
'Gerald',
'Wesley',
'None',
'Arnie',
'Derek',
'Jeffrey',
'None',
'Solomon',
'Huck',
'very',
'None',
'O',
'Sampson',
'None',
'None',
'Blue',
'Anakin',
'None',
'Finley',
'Maximus',
'None',
'Tucker',
'Finley',
'Sprinkles',
```

```
'None',
'Winnie',
'Heinrich',
'Loki',
'Shakespeare',
'Chelsea',
'Fizz',
'Bungalo',
'Chip',
'Grey',
'None',
'Roosevelt',
'Gromit',
'a',
'Willem',
'None',
'Jack',
'Finn',
'Penny',
'None',
'Davey',
'Dakota',
'Fizz',
'Frankie',
'Dixie',
'Charlie',
'None',
'None',
'Winston',
'Sebastian',
'None',
'very',
'Al',
'Jackson',
'just',
'Carbon',
'Klein',
'Titan',
'None',
'DonDon',
'Kirby',
'None',
'Jesse',
'Lou',
'Oakley',
'Nollie',
'Chevy',
'Gerald',
```

```
'Tito',
'Philbert',
'Louie',
'None',
'Rupert',
'None',
'Rufus',
'None',
'Brudge',
'Shadoe',
'Oscar',
'Colby',
'Juno',
'Angel',
'Brat',
'Tove',
'my',
'Louie',
'Gromit',
'Aubie',
'Kota',
'None',
'Alfie',
'Clark',
'Eve',
'Belle',
'Leela',
'Glenn',
'Buddy',
'Scout',
'None',
'Shelby',
'None',
'None',
'None',
'Sephie',
'None',
'Bruce',
'Bonaparte',
'Albert',
'Bo',
'Wishes',
'Rose',
'Theo',
'Atlas',
'None',
'Rocco',
'Fido',
```

```
'Sadie',
'None',
'None',
'None',
'Kirby',
'Maggie',
'None',
'Emma',
'Oakley',
'None',
'Luna',
'None',
'Toby',
'Spencer',
'Lilli',
'None',
'Boston',
'Brandonald',
'None',
'Odie',
'Corey',
'None',
'None',
'Leonard',
'Chompsky',
'Beckham',
'Cooper',
'None',
'None',
'None',
'None',
'Devón',
'Oliver',
'Jax',
'Gert',
'None',
'None',
'None',
'None',
'None',
'one',
'Watson',
'Rubio',
'Winnie',
'Keith',
'Milo',
'Dex',
'None',
```

```
'Charlie',
'None',
'None',
'Scout',
'Hank',
'Carly',
'Ace',
'None',
'Tayzie',
'Carl',
'Grizzie',
'None',
'None',
'None',
'None',
'None',
'None',
'None',
'Brody',
'Lola',
'Ruby',
'Tucker',
'Fred',
'Toby',
'None',
'Max',
'None',
'Gilbert',
'None',
'Cooper',
'Milo',
'Meyer',
'Malcolm',
'Arnie',
'Zoe',
'None',
'None',
'Stewie',
'Calvin',
'Lilah',
'Spanky',
'None',
'Jameson',
'Beau',
'Jax',
'Piper',
'Bo',
'Atticus',
```

```
 'Lucy',
 'Finn',
 'None',
 'George',
 'Blu',
 'Boomer',
 'Winston',
 'Dietrich',
 'not',
 'Divine',
 'None',
 'Tripp',
 'his',
 'one',
 'Cora',
 'None',
 'None',
 'Duke',
 'None',
 'None',
 ...]
```

```
In [12]: # Check to see if there is any quality issue in the doggo column
         # doggo == "None" or "doggo"
         df_1['doggo'].value_counts()
         # No quality issue here
```

```
Out[12]: None     2259
         doggo      97
         Name: doggo, dtype: int64
```

```
In [13]: # Check to see if there is any quality issue in the floofer column
         # floofer == "None" or "floofer"
         df_1['floofer'].value_counts()
         # No quality issue here
```

```
Out[13]: None       2346
         floofer      10
         Name: floofer, dtype: int64
```

```
In [14]: # Check to see if there is any quality issue in the pupper column
         # pupper == "None" or "pupper"
         df_1['pupper'].value_counts()
         # No quality issue here
```

```
Out[14]: None      2099
         pupper     257
         Name: pupper, dtype: int64
```

```
In [15]: # Check to see if there is any quality issue in the puppo column
         # puppo == "None" or "puppo"
         df_1['puppo'].value_counts()
         # No quality issue here

Out[15]: None     2326
         puppo      30
         Name: puppo, dtype: int64

In [16]: # Check to see if there is any quality issue in the rating_denominator column
         # Rating_denominator == 10
         df_1['rating_denominator'].value_counts()
         # There are other values in this column rather than 10
         # Quality issue

Out[16]: 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 [17]: # Check if there is any duplicate tweet
         sum(df_1['tweet_id'].duplicated())
         # All clear

Out[17]: 0
```

Assess df image

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

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null int64
```

```
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

```
In [19]: df_image.describe()
```

Out[19]:

	tweet_id	img_num	p1_conf	p2_conf	p3_conf
count	2.075000e+03	2075.000000	2075.000000	2.075000e+03	2.075000e+03
mean	7.384514e+17	1.203855	0.594548	1.345886e-01	6.032417e-02
std	6.785203e+16	0.561875	0.271174	1.006657e-01	5.090593e-02
min	6.660209e+17	1.000000	0.044333	1.011300e-08	1.740170e-10
25%	6.764835e+17	1.000000	0.364412	5.388625e-02	1.622240e-02
50%	7.119988e+17	1.000000	0.588230	1.181810e-01	4.944380e-02
75%	7.932034e+17	1.000000	0.843855	1.955655e-01	9.180755e-02
max	8.924206e+17	4.000000	1.000000	4.880140e-01	2.734190e-01

```
In [20]: # Check if there is any duplicate tweet
         sum(df_image['tweet_id'].duplicated())
         # All clear
```

```
Out[20]: 0
```

```
In [21]: # p1_conf <= 1
         df_image['p1_conf'].max()
         # No quality issue here
```

```
Out[21]: 1.0
```

```
In [22]: # p2_conf <= 1
         df_image['p2_conf'].max()
         # No quality issue here
```

```
Out[22]: 0.48801400000000011
```

```
In [23]: # p3_conf <= 1
         df_image['p3_conf'].max()
         # No quality issue here
```

```
Out[23]: 0.27341900000000002
```

```
In [24]: # p1_dog == "True" or "False"
         df_image['p1_dog'].value_counts()
         # No quality issue here

Out[24]: True     1532
         False     543
         Name: p1_dog, dtype: int64

In [25]: # p2_dog == "True" or "False"
         df_image['p2_dog'].value_counts()
         # No quality issue here

Out[25]: True     1553
         False     522
         Name: p2_dog, dtype: int64

In [26]: # p3_dog == "True" or "False"
         df_image['p3_dog'].value_counts()
         # No quality issue here

Out[26]: True     1499
         False     576
         Name: p3_dog, dtype: int64
```

Assess df_tweet

```
In [27]: tweet_json.head()

Out[27]:    contributors  coordinates          created_at display_text_range  \
         0           NaN          NaN 2017-08-01 16:23:56            [0, 85]
         1           NaN          NaN 2017-08-01 00:17:27           [0, 138]
         2           NaN          NaN 2017-07-31 00:18:03           [0, 121]
         3           NaN          NaN 2017-07-30 15:58:51            [0, 79]
         4           NaN          NaN 2017-07-29 16:00:24           [0, 138]

                                                 entities  \
         0  {'hashtags': [], 'symbols': [], 'user_mentions...
         1  {'hashtags': [], 'symbols': [], 'user_mentions...
         2  {'hashtags': [], 'symbols': [], 'user_mentions...
         3  {'hashtags': [], 'symbols': [], 'user_mentions...
         4  {'hashtags': [{'text': 'BarkWeek', 'indices': ...

                                       extended_entities  favorite_count  \
         0  {'media': [{'id': 892420639486877696, 'id_str'...           39467
         1  {'media': [{'id': 892177413194625024, 'id_str'...           33819
         2  {'media': [{'id': 891815175371796480, 'id_str'...           25461
         3  {'media': [{'id': 891689552724799489, 'id_str'...           42908
         4  {'media': [{'id': 891327551943041024, 'id_str'...           41048
```

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

                          ...                            \
0                         ...
1                         ...
2                         ...
3                         ...
4                         ...

   possibly_sensitive_appealable  quoted_status quoted_status_id  \
0                            0.0            NaN              NaN
1                            0.0            NaN              NaN
2                            0.0            NaN              NaN
3                            0.0            NaN              NaN
4                            0.0            NaN              NaN

   quoted_status_id_str  retweet_count  retweeted  retweeted_status  \
0                   NaN           8853      False               NaN
1                   NaN           6514      False               NaN
2                   NaN           4328      False               NaN
3                   NaN           8964      False               NaN
4                   NaN           9774      False               NaN

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

                                                user
0  {'id': 4196983835, 'id_str': '4196983835', 'na...
1  {'id': 4196983835, 'id_str': '4196983835', 'na...
2  {'id': 4196983835, 'id_str': '4196983835', 'na...
3  {'id': 4196983835, 'id_str': '4196983835', 'na...
4  {'id': 4196983835, 'id_str': '4196983835', 'na...

[5 rows x 31 columns]
```

In [28]:
```
tweet_json.info()
# contributors does not have any data
# coordinates does not have any data
# geo does not have any data
```

```
        # place has only one row
        # favorite_count should be int
        # id should be object

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 31 columns):
contributors                    0 non-null float64
coordinates                     0 non-null float64
created_at                      2354 non-null datetime64[ns]
display_text_range              2354 non-null object
entities                        2354 non-null object
extended_entities               2073 non-null object
favorite_count                  2354 non-null int64
favorited                       2354 non-null bool
full_text                       2354 non-null object
geo                             0 non-null float64
id                              2354 non-null int64
id_str                          2354 non-null int64
in_reply_to_screen_name         78 non-null object
in_reply_to_status_id           78 non-null float64
in_reply_to_status_id_str       78 non-null float64
in_reply_to_user_id             78 non-null float64
in_reply_to_user_id_str         78 non-null float64
is_quote_status                 2354 non-null bool
lang                            2354 non-null object
place                           1 non-null object
possibly_sensitive              2211 non-null float64
possibly_sensitive_appealable   2211 non-null float64
quoted_status                   28 non-null object
quoted_status_id                29 non-null float64
quoted_status_id_str            29 non-null float64
retweet_count                   2354 non-null int64
retweeted                       2354 non-null bool
retweeted_status                179 non-null object
source                          2354 non-null object
truncated                       2354 non-null bool
user                            2354 non-null object
dtypes: bool(4), datetime64[ns](1), float64(11), int64(4), object(11)
memory usage: 505.8+ KB


In [29]: tweet_json['quoted_status_id_str']

Out[29]: 0      NaN
         1      NaN
         2      NaN
         3      NaN
```

```
4       NaN
5       NaN
6       NaN
7       NaN
8       NaN
9       NaN
10      NaN
11      NaN
12      NaN
13      NaN
14      NaN
15      NaN
16      NaN
17      NaN
18      NaN
19      NaN
20      NaN
21      NaN
22      NaN
23      NaN
24      NaN
25      NaN
26      NaN
27      NaN
28      NaN
29      NaN
         ..
2324    NaN
2325    NaN
2326    NaN
2327    NaN
2328    NaN
2329    NaN
2330    NaN
2331    NaN
2332    NaN
2333    NaN
2334    NaN
2335    NaN
2336    NaN
2337    NaN
2338    NaN
2339    NaN
2340    NaN
2341    NaN
2342    NaN
2343    NaN
2344    NaN
```

```
2345   NaN
2346   NaN
2347   NaN
2348   NaN
2349   NaN
2350   NaN
2351   NaN
2352   NaN
2353   NaN
Name: quoted_status_id_str, Length: 2354, dtype: float64
```

```
In [30]: tweet_json['extended_entities'].sample(10)
         # tidiness issue
         # seperate tweet_json['extended_entities'] into columns by comma
```

```
Out[30]: 28      {'media': [{'id': 886366138128449536, 'id_str'...
         1163    {'media': [{'id': 722974578855804928, 'id_str'...
         1101    {'media': [{'id': 735256011242541056, 'id_str'...
         1208    {'media': [{'id': 715680780459098112, 'id_str'...
         300     {'media': [{'id': 836677701918392320, 'id_str'...
         143     {'media': [{'id': 863907404156518400, 'id_str'...
         1099    {'media': [{'id': 735635075396599808, 'id_str'...
         1670    {'media': [{'id': 682388986111913984, 'id_str'...
         358     {'media': [{'id': 826958645422342144, 'id_str'...
         1929    {'media': [{'id': 674036081864609793, 'id_str'...
         Name: extended_entities, dtype: object
```

```
In [31]: tweet_json['user'].sample(10)
         # tidiness issue
         # seperate tweet_json['user'] into columns by comma
```

```
Out[31]: 1247    {'id': 4196983835, 'id_str': '4196983835', 'na...
         1832    {'id': 4196983835, 'id_str': '4196983835', 'na...
         256     {'id': 4196983835, 'id_str': '4196983835', 'na...
         942     {'id': 4196983835, 'id_str': '4196983835', 'na...
         786     {'id': 4196983835, 'id_str': '4196983835', 'na...
         1394    {'id': 4196983835, 'id_str': '4196983835', 'na...
         835     {'id': 4196983835, 'id_str': '4196983835', 'na...
         435     {'id': 4196983835, 'id_str': '4196983835', 'na...
         156     {'id': 4196983835, 'id_str': '4196983835', 'na...
         1416    {'id': 4196983835, 'id_str': '4196983835', 'na...
         Name: user, dtype: object
```

```
In [32]: tweet_json['quoted_status'].sample(10)
         # tidiness issue
         # value should not be a dictionary
         # seperate values that contain a dictionary in tweet_json['quoted_status']
```

```
Out[32]: 936     NaN
         2076    NaN
```

```
1648    NaN
218     NaN
635     NaN
1914    NaN
613     NaN
51      NaN
107     NaN
88      NaN
Name: quoted_status, dtype: object
```

```
In [33]: tweet_json['retweeted_status'].sample(10)
         # tidiness issue
         # value should not be a dictionary
         # seperate values that contain a dictionary in tweet_json['retweeted_status']
```

```
Out[33]: 2341    NaN
         1288    NaN
         1874    NaN
         21      NaN
         338     NaN
         900     NaN
         1873    NaN
         792     NaN
         460     NaN
         1273    NaN
         Name: retweeted_status, dtype: object
```

```
In [34]: tweet_json['source'].sample(10)
         # quality issue
         # Remove html tags in tweet_json['source']
```

```
Out[34]: 1978    <a href="http://twitter.com/download/iphone" r...
         1685    <a href="http://twitter.com/download/iphone" r...
         313     <a href="http://twitter.com/download/iphone" r...
         774     <a href="http://twitter.com/download/iphone" r...
         240     <a href="http://twitter.com/download/iphone" r...
         2015    <a href="http://twitter.com/download/iphone" r...
         976     <a href="https://about.twitter.com/products/tw...
         1052    <a href="http://twitter.com/download/iphone" r...
         550     <a href="http://twitter.com/download/iphone" r...
         392     <a href="http://twitter.com/download/iphone" r...
         Name: source, dtype: object
```

```
In [35]: tweet_json.describe()
```

```
Out[35]:        contributors  coordinates  favorite_count  geo            id  \
         count           0.0          0.0     2354.000000  0.0  2.354000e+03
         mean            NaN          NaN     8080.968564  NaN  7.426978e+17
         std             NaN          NaN    11814.771334  NaN  6.852812e+16
```

```
min             NaN          NaN        0.000000  NaN  6.660209e+17
25%             NaN          NaN     1415.000000  NaN  6.783975e+17
50%             NaN          NaN     3603.500000  NaN  7.194596e+17
75%             NaN          NaN    10122.250000  NaN  7.993058e+17
max             NaN          NaN   132810.000000  NaN  8.924206e+17

            id_str  in_reply_to_status_id  in_reply_to_status_id_str  \
count  2.354000e+03           7.800000e+01               7.800000e+01
mean   7.426978e+17           7.455079e+17               7.455079e+17
std    6.852812e+16           7.582492e+16               7.582492e+16
min    6.660209e+17           6.658147e+17               6.658147e+17
25%    6.783975e+17           6.757419e+17               6.757419e+17
50%    7.194596e+17           7.038708e+17               7.038708e+17
75%    7.993058e+17           8.257804e+17               8.257804e+17
max    8.924206e+17           8.862664e+17               8.862664e+17

       in_reply_to_user_id  in_reply_to_user_id_str  possibly_sensitive  \
count         7.800000e+01             7.800000e+01              2211.0
mean          2.014171e+16             2.014171e+16                 0.0
std           1.252797e+17             1.252797e+17                 0.0
min           1.185634e+07             1.185634e+07                 0.0
25%           3.086374e+08             3.086374e+08                 0.0
50%           4.196984e+09             4.196984e+09                 0.0
75%           4.196984e+09             4.196984e+09                 0.0
max           8.405479e+17             8.405479e+17                 0.0

       possibly_sensitive_appealable  quoted_status_id  quoted_status_id_str  \
count                         2211.0      2.900000e+01          2.900000e+01
mean                             0.0      8.162686e+17          8.162686e+17
std                              0.0      6.164161e+16          6.164161e+16
min                              0.0      6.721083e+17          6.721083e+17
25%                              0.0      7.888183e+17          7.888183e+17
50%                              0.0      8.340867e+17          8.340867e+17
75%                              0.0      8.664587e+17          8.664587e+17
max                              0.0      8.860534e+17          8.860534e+17

       retweet_count
count    2354.000000
mean     3164.797366
std      5284.770364
min         0.000000
25%       624.500000
50%      1473.500000
75%      3652.000000
max     79515.000000
```

2 Quality issues

df_1

1) Change tweet id to object

2) Change retweeted_status_id to object

3) Change timestamp to date time

4) Change retweeted_status_timestamp to date time

5) Remove "a", "an", "the", "his", "her", "one" in df_1['name']

6) Change all rows whose values != 10 in df_1['rating_denominator'] to 10

tweet_json

7) Drop columns that do not contain values: contributors, coordinates, geo, place, possibly_sensitive, possibly_sensitive_id, possibly_sensitive_id_string

8) favorite_count should be int

9) id should be object

df_image

8) Remove values that are not dog breeds in p1, p2 and p3

3 Tidiness issues

df_1

12) Remove html command in the expanded_urls

tweet_json.txt

13) Remove html tags in tweet_json['source']

4 Clean

df_1

```
In [36]: # 1) Change tweet id to object
         df_1['tweet_id'] = df_1['tweet_id'].astype(object)
```

```
In [37]: # 2) Change retweeted_status_id to object
         df_1['retweeted_status_id'] = df_1['retweeted_status_id'].astype(object)
```

```
In [38]: # 3) Change timestamp to date time
         df_1['timestamp'] = df_1['timestamp'].astype('datetime64[ns]')

In [39]: # 4) Change retweeted_status_timestamp to date time
         df_1['retweeted_status_timestamp'] = df_1['retweeted_status_timestamp'].astype('datetim

In [40]: # 6) Remove "a", "an", "the", "his", "her", "one" in df_1['name']
         df_1['name'] = df_1['name'].replace(['a','an','the','his','her','one','not'],'None')
         list(df_1['name'])

Out[40]: ['Phineas',
          'Tilly',
          'Archie',
          'Darla',
          'Franklin',
          'None',
          'Jax',
          'None',
          'Zoey',
          'Cassie',
          'Koda',
          'Bruno',
          'None',
          'Ted',
          'Stuart',
          'Oliver',
          'Jim',
          'Zeke',
          'Ralphus',
          'Canela',
          'Gerald',
          'Jeffrey',
          'such',
          'Canela',
          'None',
          'None',
          'Maya',
          'Mingus',
          'Derek',
          'Roscoe',
          'None',
          'Waffles',
          'None',
          'Jimbo',
          'Maisey',
          'None',
          'Lilly',
          'None',
```

```
'Earl',
'Lola',
'Kevin',
'None',
'None',
'Yogi',
'Noah',
'Bella',
'Grizzwald',
'None',
'Rusty',
'Gus',
'Stanley',
'Alfy',
'Koko',
'Rey',
'Gary',
'None',
'None',
'Elliot',
'Louis',
'None',
'Bella',
'Jesse',
'None',
'Romeo',
'None',
'Bailey',
'Duddles',
'Jack',
'Emmy',
'Steven',
'Beau',
'Snoopy',
'None',
'Shadow',
'Terrance',
'Shadow',
'Emmy',
'Aja',
'None',
'Penny',
'Dante',
'Nelly',
'Ginger',
'None',
'Benedict',
'Venti',
```

```
'Goose',
'Nugget',
'None',
'None',
'Cash',
'Coco',
'Jed',
'None',
'Sebastian',
'Walter',
'None',
'Sierra',
'Sierra',
'None',
'None',
'None',
'Monkey',
'None',
'Harry',
'Kody',
'Lassie',
'Rover',
'Napolean',
'Dawn',
'None',
'Boomer',
'None',
'None',
'Cody',
'Zoey',
'Rumble',
'Clifford',
'quite',
'Dewey',
'Stanley',
'Scout',
'Gizmo',
'Walter',
'Cooper',
'None',
'Cooper',
'None',
'Harold',
'Shikha',
'None',
'None',
'Jamesy',
'None',
```

```
'Lili',
'Jamesy',
'Coco',
'None',
'Boomer',
'Sammy',
'Nelly',
'None',
'Meatball',
'Paisley',
'Albus',
'Neptune',
'Quinn',
'Belle',
'None',
'None',
'Quinn',
'Zooey',
'Dave',
'Jersey',
'None',
'None',
'Hobbes',
'None',
'Burt',
'Lorenzo',
'None',
'Lorenzo',
'Carl',
'Jordy',
'None',
'None',
'Milky',
'Trooper',
'None',
'quite',
'None',
'Winston',
'None',
'Sophie',
'Wyatt',
'Rosie',
'Thor',
'None',
'Oscar',
'None',
'None',
'Zeke',
```

```
'Luna',
'Callie',
'None',
'None',
'None',
'Cermet',
'None',
'None',
'None',
'None',
'None',
'quite',
'George',
'None',
'Marlee',
'Arya',
'Einstein',
'None',
'None',
'Alice',
'None',
'Rumpole',
'None',
'Benny',
'Aspen',
'Jarod',
'Wiggles',
'General',
'Sailor',
'Astrid',
'None',
'None',
'Iggy',
'Snoop',
'Kyle',
'Leo',
'None',
'Riley',
'Boomer',
'None',
'Gidget',
'Noosh',
'None',
'Kevin',
'None',
'Odin',
'None',
'Jerry',
```

```
'Charlie',
'None',
'Georgie',
'Rontu',
'None',
'Cannon',
'Furzey',
'Daisy',
'None',
'Tuck',
'Barney',
'None',
'Vixen',
'None',
'Jarvis',
'None',
'None',
'None',
'Mimosa',
'Pickles',
'Bungalo',
'None',
'Brady',
'Luna',
'Charlie',
'Margo',
'None',
'Sadie',
'Hank',
'Tycho',
'Stephan',
'Charlie',
'Indie',
'Winnie',
'George',
'Bentley',
'Ken',
'Penny',
'None',
'None',
'Max',
'Dawn',
'Maddie',
'Pipsy',
'None',
'None',
'Maddie',
'None',
```

```
'Monty',
'Sojourner',
'Winston',
'None',
'Odie',
'None',
'Arlo',
'None',
'Riley',
'Walter',
'Stanley',
'Sunny',
'None',
'None',
'Daisy',
'None',
'Waffles',
'Vincent',
'Lucy',
'Clark',
'None',
'Mookie',
'Meera',
'Oliver',
'None',
'Buddy',
'Ava',
'Lucy',
'None',
'Rory',
'Eli',
'Lola',
'None',
'Ash',
'Lola',
'None',
'None',
'None',
'Tucker',
'Tobi',
'None',
'Leo',
'Chester',
'Wilson',
'Sunshine',
'None',
'Lipton',
'Bentley',
```

```
'Charlie',
'Gabby',
'Bronte',
'Poppy',
'Gidget',
'Rhino',
'None',
'Willow',
'None',
'None',
'Orion',
'Eevee',
'Charlie',
'Smiley',
'Logan',
'Moreton',
'None',
'Klein',
'Miguel',
'Emanuel',
'None',
'Kuyu',
'Daisy',
'None',
'Dutch',
'Pete',
'None',
'Scooter',
'Tucker',
'Reggie',
'Lilly',
'Kyro',
'Samson',
'Loki',
'Mia',
'Leo',
'None',
'Astrid',
'Malcolm',
'Dexter',
'Gus',
'Alfie',
'Fiona',
'None',
'Mutt',
'Bear',
'Doobert',
'Beebop',
```

```
'Alexander',
'None',
'Sailer',
'Brutus',
'Kona',
'Boots',
'Tucker',
'Ralphie',
'Phil',
'Charlie',
'Loki',
'Cupid',
'None',
'None',
'Pawnd',
'Pilot',
'None',
'None',
'Ike',
'Mo',
'Toby',
'None',
'Sweet',
'Pablo',
'Pablo',
'Bailey',
'Scooter',
'Wilson',
'None',
'Nala',
'None',
'Cash',
'Balto',
'Winston',
'Crawford',
'None',
'Wyatt',
'None',
'Albus',
'None',
'Hobbes',
'Paisley',
'None',
'Paisley',
'Gabe',
'None',
'Mattie',
'Jimison',
```

```
'Hercules',
'Duchess',
'Harlso',
'Sampson',
'Sundance',
'None',
'Luca',
'None',
'Flash',
'Finn',
'Sunny',
'None',
'None',
'Peaches',
'None',
'None',
'Oliver',
'Oliver',
'None',
'Howie',
'Jazzy',
'Anna',
'None',
'Finn',
'Bo',
'Sunny',
'Sunny',
'Bo',
'Seamus',
'Wafer',
'Bear',
'Chelsea',
'Tom',
'Moose',
'Florence',
'Autumn',
'None',
'Buddy',
'Dido',
'Eugene',
'Herschel',
'Ken',
'Strudel',
'None',
'Tebow',
'None',
'Chloe',
'Betty',
```

```
'Timber',
'Binky',
'Moose',
'Dudley',
'Comet',
'Jack',
'Larry',
'Jack',
'None',
'Levi',
'Akumi',
'Titan',
'None',
'Cooper',
'Olivia',
'Beau',
'Alf',
'Oshie',
'Bruce',
'Chubbs',
'Gary',
'Sky',
'Atlas',
'None',
'None',
'Eleanor',
'Layla',
'None',
'None',
'None',
'Toby',
'Rocky',
'Baron',
'Tyr',
'Bauer',
'Swagger',
'Sammy',
'Brandi',
'None',
'Mary',
'Moe',
'Ted',
'Halo',
'None',
'Augie',
'Craig',
'Sam',
'Hunter',
```

```
'Pavlov',
'Phil',
'Gus',
'None',
'Maximus',
'None',
'Kyro',
'Wallace',
'Ito',
'None',
'Koda',
'Seamus',
'Milo',
'None',
'Cooper',
'Ollie',
'Stephan',
'Cali',
'Lennon',
'None',
'None',
'None',
'Waffles',
'Dave',
'incredibly',
'Penny',
'Major',
'Duke',
'Reginald',
'Zeke',
'Sansa',
'Shooter',
'Django',
'None',
'Rusty',
'Bo',
'Diogi',
'None',
'None',
'Sonny',
'Philbert',
'Winston',
'Marley',
'None',
'Bailey',
'Winnie',
'Severus',
'None',
```

```
'None',
'Loki',
'None',
'Ronnie',
'None',
'Wallace',
'None',
'Milo',
'Anakin',
'Bones',
'None',
'None',
'Mauve',
'Chef',
'None',
'Sampson',
'Doc',
'Bo',
'Peaches',
'None',
'Tucker',
'Sobe',
'Longfellow',
'None',
'Jeffrey',
'Mister',
'Iroh',
'Shadow',
'Baloo',
'None',
'Stubert',
'None',
'Jack',
'None',
'None',
'Lola',
'Paull',
'None',
'Timison',
'None',
'Davey',
'Cooper',
'None',
'Cassie',
'Pancake',
'None',
'Tyrone',
'Tyr',
```

```
'Romeo',
'None',
'None',
'Snicku',
'Ruby',
'Ruby',
'None',
'None',
'Yogi',
'Daisy',
'None',
'Brody',
'Bailey',
'Rizzy',
'Mack',
'Butter',
'Nimbus',
'Laika',
'Maximus',
'Clark',
'None',
'Dobby',
'Fiona',
'Moreton',
'Dave',
'None',
'Tucker',
'Juno',
'Maude',
'Lily',
'Newt',
'Benji',
'Nida',
'None',
'Robin',
'None',
'Bailey',
'Monster',
'BeBe',
'Remus',
'None',
'None',
'Maddie',
'None',
'None',
'Levi',
'Mabel',
'Alfie',
```

```
'Misty',
'Betty',
'Happy',
'Mosby',
'Duke',
'Maggie',
'Bruce',
'Leela',
'Happy',
'Buddy',
'Ralphy',
'Eli',
'Brownie',
'Rizzy',
'None',
'Meyer',
'Stella',
'Bo',
'Lucy',
'Butter',
'mad',
'Dexter',
'None',
'Leo',
'Bo',
'None',
'Frank',
'Tonks',
'Moose',
'Lincoln',
'Carl',
'Rory',
'Oakley',
'Logan',
'None',
'Dale',
'Rizzo',
'Arnie',
'Mattie',
'None',
'Scout',
'Lucy',
'Rusty',
'Pinot',
'Dallas',
'None',
'Doc',
'Hero',
```

```
'Rusty',
'Frankie',
'Stormy',
'Reginald',
'Balto',
'Riley',
'Mairi',
'Loomis',
'Finn',
'Godi',
'Kenny',
'Dave',
'Earl',
'Cali',
'Deacon',
'Penny',
'Timmy',
'Sampson',
'Harper',
'Chipson',
'None',
'Combo',
'None',
'None',
'Oakley',
'None',
'None',
'Dash',
'Koda',
'Hercules',
'None',
'Bell',
'None',
'Bear',
'None',
'Hank',
'None',
'Scout',
'None',
'Hurley',
'Reggie',
'None',
'Jay',
'None',
'None',
'Mya',
'Strider',
'Penny',
```

```
'None',
'None',
'Nala',
'Stanley',
'None',
'Sophie',
'Gerald',
'Wesley',
'None',
'Arnie',
'Derek',
'Jeffrey',
'None',
'Solomon',
'Huck',
'very',
'None',
'O',
'Sampson',
'None',
'None',
'Blue',
'Anakin',
'None',
'Finley',
'Maximus',
'None',
'Tucker',
'Finley',
'Sprinkles',
'None',
'Winnie',
'Heinrich',
'Loki',
'Shakespeare',
'Chelsea',
'Fizz',
'Bungalo',
'Chip',
'Grey',
'None',
'Roosevelt',
'Gromit',
'None',
'Willem',
'None',
'Jack',
'Finn',
```

```
'Penny',
'None',
'Davey',
'Dakota',
'Fizz',
'Frankie',
'Dixie',
'Charlie',
'None',
'None',
'Winston',
'Sebastian',
'None',
'very',
'Al',
'Jackson',
'just',
'Carbon',
'Klein',
'Titan',
'None',
'DonDon',
'Kirby',
'None',
'Jesse',
'Lou',
'Oakley',
'Nollie',
'Chevy',
'Gerald',
'Tito',
'Philbert',
'Louie',
'None',
'Rupert',
'None',
'Rufus',
'None',
'Brudge',
'Shadoe',
'Oscar',
'Colby',
'Juno',
'Angel',
'Brat',
'Tove',
'my',
'Louie',
```

```
'Gromit',
'Aubie',
'Kota',
'None',
'Alfie',
'Clark',
'Eve',
'Belle',
'Leela',
'Glenn',
'Buddy',
'Scout',
'None',
'Shelby',
'None',
'None',
'None',
'Sephie',
'None',
'Bruce',
'Bonaparte',
'Albert',
'Bo',
'Wishes',
'Rose',
'Theo',
'Atlas',
'None',
'Rocco',
'Fido',
'Sadie',
'None',
'None',
'None',
'Kirby',
'Maggie',
'None',
'Emma',
'Oakley',
'None',
'Luna',
'None',
'Toby',
'Spencer',
'Lilli',
'None',
'Boston',
'Brandonald',
```

```
'None',
'Odie',
'Corey',
'None',
'None',
'Leonard',
'Chompsky',
'Beckham',
'Cooper',
'None',
'None',
'None',
'None',
'Devón',
'Oliver',
'Jax',
'Gert',
'None',
'None',
'None',
'None',
'None',
'None',
'Watson',
'Rubio',
'Winnie',
'Keith',
'Milo',
'Dex',
'None',
'Charlie',
'None',
'None',
'Scout',
'Hank',
'Carly',
'Ace',
'None',
'Tayzie',
'Carl',
'Grizzie',
'None',
'None',
'None',
'None',
'None',
'None',
'None',
```

```
'Brody',
'Lola',
'Ruby',
'Tucker',
'Fred',
'Toby',
'None',
'Max',
'None',
'Gilbert',
'None',
'Cooper',
'Milo',
'Meyer',
'Malcolm',
'Arnie',
'Zoe',
'None',
'None',
'Stewie',
'Calvin',
'Lilah',
'Spanky',
'None',
'Jameson',
'Beau',
'Jax',
'Piper',
'Bo',
'Atticus',
'Lucy',
'Finn',
'None',
'George',
'Blu',
'Boomer',
'Winston',
'Dietrich',
'None',
'Divine',
'None',
'Tripp',
'None',
'None',
'Cora',
'None',
'None',
'Duke',
```

```
 'None',
 'None',
 ...]
```

```
In [41]: # 7) Change all rows whose values != 10 in df_1['rating_denominator'] to 10
         df_1['rating_denominator'] = df_1['rating_denominator'].replace([11,50,80,20,2,16,40,70
         df_1['rating_denominator'].value_counts()
```

```
Out[41]: 10    2356
         Name: rating_denominator, dtype: int64
```

```
In [42]: # 12) Remove html command in the expanded_urls
         df_1['source'] = df_1['source'].str.strip("<a href=")
         df_1['source'] = df_1['source'].str.strip("</a>")
         df_1['source'] = df_1['source'].str.split(pat="rel=", expand=True)
         df_1['source'].sample(5)
```

```
Out[42]: 903     "http://twitter.com/download/iphone"
         1245    "http://twitter.com/download/iphone"
         1200    "http://twitter.com/download/iphone"
         19      "http://twitter.com/download/iphone"
         1693    "http://twitter.com/download/iphone"
         Name: source, dtype: object
```

```
In [43]: df_1
```

```
Out[43]:                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    887978434458587136                    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                                source  \
0   2017-08-01 16:23:56  "http://twitter.com/download/iphone"
1   2017-08-01 00:17:27  "http://twitter.com/download/iphone"
2   2017-07-31 00:18:03  "http://twitter.com/download/iphone"
3   2017-07-30 15:58:51  "http://twitter.com/download/iphone"
4   2017-07-29 16:00:24  "http://twitter.com/download/iphone"
5   2017-07-29 00:08:17  "http://twitter.com/download/iphone"
6   2017-07-28 16:27:12  "http://twitter.com/download/iphone"
```

```
7    2017-07-28 00:22:40  "http://twitter.com/download/iphone"
8    2017-07-27 16:25:51  "http://twitter.com/download/iphone"
9    2017-07-26 15:59:51  "http://twitter.com/download/iphone"
10   2017-07-26 00:31:25  "http://twitter.com/download/iphone"
11   2017-07-25 16:11:53  "http://twitter.com/download/iphone"
12   2017-07-25 01:55:32  "http://twitter.com/download/iphone"
13   2017-07-25 00:10:02  "http://twitter.com/download/iphone"
14   2017-07-24 17:02:04  "http://twitter.com/download/iphone"
15   2017-07-24 00:19:32  "http://twitter.com/download/iphone"
16   2017-07-23 00:22:39  "http://twitter.com/download/iphone"
17   2017-07-22 16:56:37  "http://twitter.com/download/iphone"
18   2017-07-22 00:23:06  "http://twitter.com/download/iphone"
19   2017-07-21 01:02:36  "http://twitter.com/download/iphone"
20   2017-07-20 16:49:33  "http://twitter.com/download/iphone"
21   2017-07-19 16:06:48  "http://twitter.com/download/iphone"
22   2017-07-19 03:39:09  "http://twitter.com/download/iphone"
23   2017-07-19 00:47:34  "http://twitter.com/download/iphone"
24   2017-07-18 16:08:03  "http://twitter.com/download/iphone"
25   2017-07-18 00:07:08  "http://twitter.com/download/iphone"
26   2017-07-17 16:17:36  "http://twitter.com/download/iphone"
27   2017-07-16 23:58:41  "http://twitter.com/download/iphone"
28   2017-07-16 20:14:00  "http://twitter.com/download/iphone"
29   2017-07-15 23:25:31  "http://twitter.com/download/iphone"
...                  ...                                    ...
2326 2015-11-17 00:24:19  "http://twitter.com/download/iphone"
2327 2015-11-17 00:06:54  "http://twitter.com/download/iphone"
2328 2015-11-16 23:23:41  "http://twitter.com/download/iphone"
2329 2015-11-16 21:54:18  "http://twitter.com/download/iphone"
2330 2015-11-16 21:10:36  "http://twitter.com/download/iphone"
2331 2015-11-16 20:32:58  "http://twitter.com/download/iphone"
2332 2015-11-16 20:01:42  "http://twitter.com/download/iphone"
2333 2015-11-16 19:31:45  "http://twitter.com/download/iphone"
2334 2015-11-16 16:37:02  "http://twitter.com/download/iphone"
2335 2015-11-16 16:11:11  "http://twitter.com/download/iphone"
2336 2015-11-16 15:14:19  "http://twitter.com/download/iphone"
2337 2015-11-16 14:57:41  "http://twitter.com/download/iphone"
2338 2015-11-16 04:02:55  "http://twitter.com/download/iphone"
2339 2015-11-16 03:55:04  "http://twitter.com/download/iphone"
2340 2015-11-16 03:44:34  "http://twitter.com/download/iphone"
2341 2015-11-16 03:22:39  "http://twitter.com/download/iphone"
2342 2015-11-16 02:38:37  "http://twitter.com/download/iphone"
2343 2015-11-16 01:59:36  "http://twitter.com/download/iphone"
2344 2015-11-16 01:52:02  "http://twitter.com/download/iphone"
2345 2015-11-16 01:22:45  "http://twitter.com/download/iphone"
2346 2015-11-16 01:01:59  "http://twitter.com/download/iphone"
2347 2015-11-16 00:55:59  "http://twitter.com/download/iphone"
2348 2015-11-16 00:49:46  "http://twitter.com/download/iphone"
2349 2015-11-16 00:35:11  "http://twitter.com/download/iphone"
```

```
2350 2015-11-16 00:30:50  "http://twitter.com/download/iphone"
2351 2015-11-16 00:24:50  "http://twitter.com/download/iphone"
2352 2015-11-16 00:04:52  "http://twitter.com/download/iphone"
2353 2015-11-15 23:21:54  "http://twitter.com/download/iphone"
2354 2015-11-15 23:05:30  "http://twitter.com/download/iphone"
2355 2015-11-15 22:32:08  "http://twitter.com/download/iphone"

                                                   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...         8.87474e+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                         NaT
1                          NaN                         NaT
2                          NaN                         NaT
3                          NaN                         NaT
4                          NaN                         NaT
5                          NaN                         NaT
6                          NaN                         NaT
7                          NaN                         NaT
8                          NaN                         NaT
9                          NaN                         NaT
10                         NaN                         NaT
11                         NaN                         NaT
12                         NaN                         NaT
13                         NaN                         NaT
14                         NaN                         NaT
15                         NaN                         NaT
16                         NaN                         NaT
17                         NaN                         NaT
18                         NaN                         NaT
19                4.196984e+09        2017-07-19 00:47:34
20                         NaN                         NaT
21                         NaN                         NaT
22                         NaN                         NaT
23                         NaN                         NaT
24                         NaN                         NaT
```

```
25                       NaN                       NaT
26                       NaN                       NaT
27                       NaN                       NaT
28                       NaN                       NaT
29                       NaN                       NaT
...                      ...                       ...
2326                     NaN                       NaT
2327                     NaN                       NaT
2328                     NaN                       NaT
2329                     NaN                       NaT
2330                     NaN                       NaT
2331                     NaN                       NaT
2332                     NaN                       NaT
2333                     NaN                       NaT
2334                     NaN                       NaT
2335                     NaN                       NaT
2336                     NaN                       NaT
2337                     NaN                       NaT
2338                     NaN                       NaT
2339                     NaN                       NaT
2340                     NaN                       NaT
2341                     NaN                       NaT
2342                     NaN                       NaT
2343                     NaN                       NaT
2344                     NaN                       NaT
2345                     NaN                       NaT
2346                     NaN                       NaT
2347                     NaN                       NaT
2348                     NaN                       NaT
2349                     NaN                       NaT
2350                     NaN                       NaT
2351                     NaN                       NaT
2352                     NaN                       NaT
2353                     NaN                       NaT
2354                     NaN                       NaT
2355                     NaN                       NaT

                                          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      None   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      None   None    None    None   None
2334                  10      None   None    None    None   None
2335                  10      None   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      None    None    None    None    None
2346                  10      None    None    None    None    None
2347                  10      None    None    None    None    None
2348                  10      None    None    None    None    None
2349                  10      None    None    None    None    None
2350                  10      None    None    None    None    None
2351                  10      None    None    None    None    None
2352                  10      None    None    None    None    None
2353                  10      None    None    None    None    None
2354                  10      None    None    None    None    None
2355                  10      None    None    None    None    None

[2356 rows x 17 columns]
```

tweet_json

```
In [44]: # 9) Remove html tags in tweet_json['source']
         tweet_json['source'] = tweet_json['source'].str.strip("<a href=")
         tweet_json['source'] = tweet_json['source'].str.strip("</a>")
         tweet_json['source'] = tweet_json['source'].str.split(pat="rel=", expand=True)
         tweet_json['source'].sample(5)

Out[44]: 341     "http://twitter.com/download/iphone"
         543     "http://twitter.com/download/iphone"
         1023    "http://twitter.com/download/iphone"
         110     "http://twitter.com/download/iphone"
         2118    "http://twitter.com/download/iphone"
         Name: source, dtype: object

In [45]: # 10) Drop columns that do not contain values: contributors, coordinates, geo, place, p
         tweet_json = tweet_json.drop(columns=['contributors','coordinates','geo','place','possi
         tweet_json.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 11 columns):
created_at            2354 non-null datetime64[ns]
display_text_range    2354 non-null object
favorite_count        2354 non-null int64
favorited             2354 non-null bool
full_text             2354 non-null object
id                    2354 non-null int64
```

```
is_quote_status       2354 non-null bool
retweet_count         2354 non-null int64
retweeted             2354 non-null bool
source                2354 non-null object
truncated             2354 non-null bool
dtypes: bool(4), datetime64[ns](1), int64(3), object(3)
memory usage: 138.0+ KB
```

```
In [46]: tweet_json.head()
```

Out[46]:

	created_at	display_text_range	favorite_count	favorited	\
0	2017-08-01 16:23:56	[0, 85]	39467	False	
1	2017-08-01 00:17:27	[0, 138]	33819	False	
2	2017-07-31 00:18:03	[0, 121]	25461	False	
3	2017-07-30 15:58:51	[0, 79]	42908	False	
4	2017-07-29 16:00:24	[0, 138]	41048	False	

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

	is_quote_status	retweet_count	retweeted	\
0	False	8853	False	
1	False	6514	False	
2	False	4328	False	
3	False	8964	False	
4	False	9774	False	

	source	truncated
0	"http://twitter.com/download/iphone"	False
1	"http://twitter.com/download/iphone"	False
2	"http://twitter.com/download/iphone"	False
3	"http://twitter.com/download/iphone"	False
4	"http://twitter.com/download/iphone"	False

```
In [47]: tweet_json = tweet_json.rename(columns={'id': 'tweet_id', 'full_text': 'text', 'created

In [48]: # change the tweet_id data type from int64 to object
         tweet_json['tweet_id'] = tweet_json['tweet_id'].astype(object)

In [49]: tweet_json.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 11 columns):
```

```
timestamp             2354 non-null datetime64[ns]
display_text_range    2354 non-null object
favorite_count        2354 non-null int64
favorited             2354 non-null bool
text                  2354 non-null object
tweet_id              2354 non-null object
is_quote_status       2354 non-null bool
retweet_count         2354 non-null int64
retweeted             2354 non-null bool
source                2354 non-null object
truncated             2354 non-null bool
dtypes: bool(4), datetime64[ns](1), int64(2), object(4)
memory usage: 138.0+ KB
```

```
In [50]: df_1 = df_1.drop(columns=['in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_st
```

```
In [51]: df_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 12 columns):
tweet_id              2356 non-null object
timestamp             2356 non-null datetime64[ns]
source                2356 non-null object
text                  2356 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: datetime64[ns](1), int64(2), object(9)
memory usage: 221.0+ KB
```

```
In [52]: df_clean = df_1.merge(tweet_json, how='outer', on=['tweet_id','text', 'timestamp', 'sou
```

```
In [53]: df_clean.head()
```

```
Out[53]:              tweet_id           timestamp  \
         0  892420643555336193 2017-08-01 16:23:56
         1  892177421306343426 2017-08-01 00:17:27
         2  891815181378084864 2017-07-31 00:18:03
         3  891689557279858688 2017-07-30 15:58:51
         4  891327558926688256 2017-07-29 16:00:24
```

```
                                  source  \
0  "http://twitter.com/download/iphone"
1  "http://twitter.com/download/iphone"
2  "http://twitter.com/download/iphone"
3  "http://twitter.com/download/iphone"
4  "http://twitter.com/download/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 doggo floofer pupper puppo display_text_range  \
0                  10   Phineas  None    None   None  None            [0, 85]
1                  10     Tilly  None    None   None  None           [0, 138]
2                  10    Archie  None    None   None  None           [0, 121]
3                  10     Darla  None    None   None  None            [0, 79]
4                  10  Franklin  None    None   None  None           [0, 138]

   favorite_count favorited is_quote_status  retweet_count retweeted truncated
0         39467.0     False           False         8853.0     False     False
1         33819.0     False           False         6514.0     False     False
2         25461.0     False           False         4328.0     False     False
3         42908.0     False           False         8964.0     False     False
4         41048.0     False           False         9774.0     False     False
```

In [54]: `df_image.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 [55]: # change the tweet_id data type from int64 to object
         df_image['tweet_id'] = df_image['tweet_id'].astype(object)

In [56]: df_clean = df_clean.merge(df_image, how='outer', on=['tweet_id'])

In [57]: df_clean.head()
```

Out[57]:

	tweet_id	timestamp	source	text	expanded_urls	rating_numerator	rating_denominator	name	doggo	floofer	...	img_num	p1
0	892420643555336193	2017-08-01 16:23:56	"http://twitter.com/download/iphone"	This is Phineas. He's a mystical boy. Only eve...	https://twitter.com/dog_rates/status/892420643...	13	10	Phineas	None	None	...	1.0	orange
1	892177421306343426	2017-08-01 00:17:27	"http://twitter.com/download/iphone"	This is Tilly. She's just checking pup on you...	https://twitter.com/dog_rates/status/892177421...	13	10	Tilly	None	None	...	1.0	Chihuahua
2	891815181378084864	2017-07-31 00:18:03	"http://twitter.com/download/iphone"	This is Archie. He is a rare Norwegian Pouncin...	https://twitter.com/dog_rates/status/891815181...	12	10	Archie	None	None	...	1.0	Chihuahua
3	891689557279858688	2017-07-30 15:58:51	"http://twitter.com/download/iphone"	This is Darla. She commenced a snooze mid meal...	https://twitter.com/dog_rates/status/891689557...	13	10	Darla	None	None	...	1.0	paper_towel
4	891327558926688256	2017-07-29 16:00:24	"http://twitter.com/download/iphone"	This is Franklin. He would like you to stop ca...	https://twitter.com/dog_rates/status/891327558...	12	10	Franklin	None	None	...	2.0	basset

```
    p1_conf  p1_dog                    p2   p2_conf  p2_dog  \
0  0.097049   False                 bagel  0.085851   False
1  0.323581    True              Pekinese  0.090647    True
2  0.716012    True              malamute  0.078253    True
3  0.170278   False  Labrador_retriever  0.168086    True
4  0.555712    True    English_springer  0.225770    True

                              p3   p3_conf p3_dog
0                         banana  0.076110  False
1                       papillon  0.068957   True
2                         kelpie  0.031379   True
3                        spatula  0.040836  False
4  German_short-haired_pointer  0.175219   True

[5 rows x 30 columns]
```

```
In [58]: df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 30 columns):
tweet_id              2356 non-null object
timestamp             2356 non-null datetime64[ns]
source                2356 non-null object
text                  2356 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
display_text_range    2354 non-null object
favorite_count        2354 non-null float64
favorited             2354 non-null object
is_quote_status       2354 non-null object
retweet_count         2354 non-null float64
retweeted             2354 non-null object
truncated             2354 non-null object
jpg_url               2075 non-null object
img_num               2075 non-null float64
p1                    2075 non-null object
p1_conf               2075 non-null float64
p1_dog                2075 non-null object
p2                    2075 non-null object
p2_conf               2075 non-null float64
```

```
p2_dog                2075 non-null object
p3                    2075 non-null object
p3_conf               2075 non-null float64
p3_dog                2075 non-null object
dtypes: datetime64[ns](1), float64(6), int64(2), object(21)
memory usage: 570.6+ KB
```

5 Visualization

Question 1: Is there a correlation between nickname and number of favorites or retweets?

In [59]:
```
# create a new dataframe that has only the tweet_id, favorite_count, retweet_count and
df_nickname = pd.DataFrame(df_clean, columns = ['tweet_id', 'favorite_count', 'retweet_
```

In [60]:
```
# I change values that contain the nick name to 1 and the ones that don't contain any r
df_nickname['doggo'] = df_nickname['doggo'].replace(['doggo','None'],['1','0'])
df_nickname['floofer'] = df_nickname['floofer'].replace(['floofer','None'],['1','0'])
df_nickname['pupper'] = df_nickname['pupper'].replace(['pupper','None'],['1','0'])
df_nickname['puppo'] = df_nickname['puppo'].replace(['puppo','None'],['1','0'])
df_nickname['doggo'].astype(int)
df_nickname['floofer'].astype(int)
df_nickname['puppo'].astype(int)
df_nickname['pupper'].astype(int)
```

Out[60]:
```
0     0
1     0
2     0
3     0
4     0
5     0
6     0
7     0
8     0
9     0
10    0
11    0
12    0
13    0
14    0
15    0
16    0
17    0
18    0
19    0
20    0
21    0
22    0
```

```
23      0
24      0
25      0
26      0
27      0
28      0
29      1
       ..
2326    0
2327    0
2328    0
2329    0
2330    0
2331    0
2332    0
2333    0
2334    0
2335    0
2336    0
2337    0
2338    0
2339    0
2340    0
2341    0
2342    0
2343    0
2344    0
2345    0
2346    0
2347    0
2348    0
2349    0
2350    0
2351    0
2352    0
2353    0
2354    0
2355    0
Name: pupper, Length: 2356, dtype: int64
```

```
In [61]: df_nickname.loc[(df_nickname['doggo'] == '0') & (df_nickname['floofer'] == '0') & (df_n
         df_nickname.loc[(df_nickname['doggo'] == '1') | (df_nickname['floofer'] == '1') | (df_n
```

```
In [62]: df_nickname['no_nickname'] = df_nickname['no_nickname'].astype(object)
```

```
In [63]: df_nickname.sample(10)
```

```
Out[63]:                tweet_id  favorite_count  retweet_count doggo floofer pupper  \
         1089  737800304142471168         10943.0         3904.0     0       0      0
```

```
2150  669683899023405056          412.0          119.0     0       0       0
722   783085703974514689         9112.0         2565.0     0       0       0
807   771908950375665664         7298.0         2181.0     1       0       0
1441  696877980375769088         2689.0          802.0     0       0       1
333   832757312314028032        18423.0         4127.0     0       0       0
1012  747242308580548608            0.0         3257.0     0       0       1
2125  670361874861563904          344.0           71.0     0       0       0
1103  735256018284875776         3675.0          993.0     1       0       0
1075  739623569819336705         4185.0         1547.0     1       0       0

      puppo no_nickname
1089      0           1
2150      0           1
722       0           1
807       0           0
1441      0           0
333       0           1
1012      0           0
2125      0           1
1103      0           0
1075      0           0
```

```
In [64]: df_nickname.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 8 columns):
tweet_id          2356 non-null object
favorite_count    2354 non-null float64
retweet_count     2354 non-null float64
doggo             2356 non-null object
floofer           2356 non-null object
pupper            2356 non-null object
puppo             2356 non-null object
no_nickname       2356 non-null object
dtypes: float64(2), object(6)
memory usage: 165.7+ KB
```

```
In [65]: # Only keep rows that contain the nickname
         doggo = df_nickname.drop(columns = ['tweet_id', 'floofer', 'pupper', 'puppo', 'no_nickn
         doggo = doggo[doggo.doggo != 'O']
         doggo.describe()
```

```
Out[65]:        favorite_count  retweet_count
         count       97.000000      97.000000
         mean     15345.010309    7295.494845
         std      19785.868648   12378.265124
         min          0.000000      39.000000
```

```
25%       5272.000000    2042.000000
50%      10042.000000    3327.000000
75%      16304.000000    5757.000000
max     131075.000000   79515.000000
```

In [66]:
```
# Only keep rows that contain the nickname
floofer = df_nickname.drop(columns = ['tweet_id', 'doggo', 'pupper', 'puppo', 'no_nickn
floofer = floofer[floofer.floofer != '0']
floofer.describe()
```

Out[66]:
```
       favorite_count  retweet_count
count        10.00000      10.000000
mean      11674.90000    4083.600000
std       10253.00493    5275.497664
min        1618.00000     496.000000
25%        4391.25000    1381.750000
50%        8689.00000    2887.000000
75%       15990.75000    3727.000000
max       33345.00000   18497.000000
```

In [67]:
```
# Only keep rows that contain the nickname
pupper = df_nickname.drop(columns = ['tweet_id', 'floofer', 'doggo', 'puppo', 'no_nickn
pupper = pupper[pupper.pupper != '0']
pupper.describe()
```

Out[67]:
```
       favorite_count  retweet_count
count      256.000000     256.000000
mean      6750.996094    2982.199219
std      10321.533383    4592.451363
min          0.000000      26.000000
25%       2123.750000     737.500000
50%       3194.000000    1375.000000
75%       7474.250000    3264.000000
max     106827.000000   32883.000000
```

In [68]:
```
# Only keep rows that contain the nickname
puppo = df_nickname.drop(columns = ['tweet_id', 'floofer', 'pupper', 'doggo', 'no_nickn
puppo = puppo[puppo.puppo != '0']
puppo.describe()
```

Out[68]:
```
       favorite_count  retweet_count
count       30.000000      30.000000
mean     18225.900000    6581.133333
std      25987.198406    9290.067361
min          0.000000     179.000000
25%       4861.750000    1538.250000
50%      10667.500000    3230.000000
75%      19691.250000    8536.000000
max     132810.000000   48265.000000
```

```
In [69]: # Only keep rows that does not contain a nickname
         no_nickname = df_nickname.drop(columns = ['tweet_id', 'floofer', 'pupper', 'puppo', 'do
         no_nickname = no_nickname[no_nickname.no_nickname != 0]
         no_nickname.describe()
```

Out[69]:

	favorite_count	retweet_count
count	1975.000000	1975.000000
mean	7761.082025	2948.880506
std	10965.312058	4582.093147
min	0.000000	0.000000
25%	1266.500000	581.000000
50%	3504.000000	1374.000000
75%	9734.000000	3538.000000
max	107956.000000	56625.000000

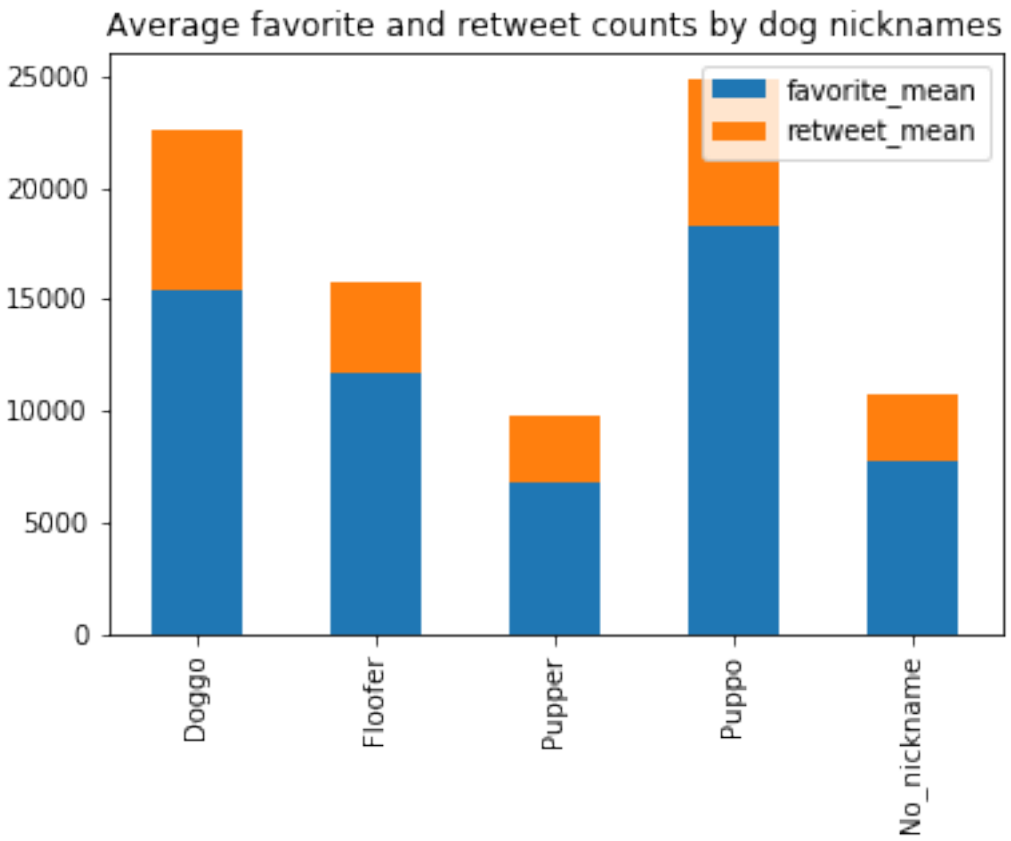
```
In [70]: # create a new dataframe with favorite_mean and retweet_mean
         fav_retweet = pd.DataFrame(data=[['15345', '7295'], ['11674', '4083'], ['6750', '2982']
                                          columns= ['favorite_mean', 'retweet_mean'],
                                          index=['Doggo', 'Floofer', 'Pupper', 'Puppo', 'No_n

In [71]: fav_retweet.info()

<class 'pandas.core.frame.DataFrame'>
Index: 5 entries, Doggo to No_nickname
Data columns (total 2 columns):
favorite_mean    5 non-null object
retweet_mean     5 non-null object
dtypes: object(2)
memory usage: 120.0+ bytes


In [72]: fav_retweet['favorite_mean'] = fav_retweet['favorite_mean'].astype(int)
         fav_retweet['retweet_mean'] = fav_retweet['retweet_mean'].astype(int)

In [73]: fav_retweet
```

Out[73]:

	favorite_mean	retweet_mean
Doggo	15345	7295
Floofer	11674	4083
Pupper	6750	2982
Puppo	18225	6581
No_nickname	7761	2948

```
In [91]: fav_retweet.plot.bar(title='Average favorite and retweet counts by dog nicknames', stac
```

Out[91]: <matplotlib.axes._subplots.AxesSubplot at 0x7fba28a8a198>



Tweets contain dog nicknames "puppo" ranked the highest average favorite count with 18225. In contrast, tweets with dog nickanme "pupper" in them receivw the lowest favorite count with only 6750 on average. So go for "puppo" not "pupper".

Tweets contain dog nicknames "doggo" that has the highest average retwwet count with 7295. On contraty, tweets with no dog nicknames at all ranked lowest retweet count at 2948. So use a dog slang for more retweet.

Question 2: What are the top 10 most popular dog breed get recognized by image scanner and their average confidence interval?

```
In [75]: df_breed =  pd.DataFrame(df_clean, columns = ['tweet_id', 'p1','p1_conf','p1_dog','p2',

In [76]: df_breed.head(10)

Out[76]:              tweet_id                          p1   p1_conf p1_dog  \
         0  892420643555336193                      orange  0.097049  False
         1  892177421306343426                   Chihuahua  0.323581   True
         2  891815181378084864                   Chihuahua  0.716012   True
         3  891689557279858688                 paper_towel  0.170278  False
         4  891327558926688256                      basset  0.555712   True
         5  891087950875897856  Chesapeake_Bay_retriever  0.425595   True
         6  890971913173991426                Appenzeller  0.341703   True
         7  890729181411237888                  Pomeranian  0.566142   True
```

```
8  890609185150312448            Irish_terrier  0.487574   True
9  890240255349198849                 Pembroke  0.511319   True

                   p2   p2_conf p2_dog                        p3   p3_conf  \
0              bagel  0.085851  False                    banana  0.076110
1           Pekinese  0.090647   True                  papillon  0.068957
2           malamute  0.078253   True                    kelpie  0.031379
3  Labrador_retriever  0.168086   True                   spatula  0.040836
4    English_springer  0.225770   True  German_short-haired_pointer  0.175219
5       Irish_terrier  0.116317   True           Indian_elephant  0.076902
6       Border_collie  0.199287   True                 ice_lolly  0.193548
7          Eskimo_dog  0.178406   True                  Pembroke  0.076507
8        Irish_setter  0.193054   True  Chesapeake_Bay_retriever  0.118184
9            Cardigan  0.451038   True                 Chihuahua  0.029248

  p3_dog  dog_breed
0  False        NaN
1   True        NaN
2   True        NaN
3  False        NaN
4   True        NaN
5  False        NaN
6  False        NaN
7   True        NaN
8   True        NaN
9   True        NaN
```

```
In [77]: df_breed.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 11 columns):
tweet_id     2356 non-null object
p1           2075 non-null object
p1_conf      2075 non-null float64
p1_dog       2075 non-null object
p2           2075 non-null object
p2_conf      2075 non-null float64
p2_dog       2075 non-null object
p3           2075 non-null object
p3_conf      2075 non-null float64
p3_dog       2075 non-null object
dog_breed    0 non-null float64
dtypes: float64(4), object(7)
memory usage: 220.9+ KB
```

```
In [88]: dog_breed = []
```

```
for p1, p2 p3, p1_dog, p2_dog, p3_dog, p1_conf, p2_conf, p3_conf in df_breed:
    if (p1_dog=='False' & p2_dog=='False' & p3_dog=='False'):
        dog_breed = Null

    # scenario 2: one True case
    elif (p1_dog=='True' & p2_dog=='False' & p3_dog =='False'):
        dog_breed = p1
    elif (p1_dog=='False' & p2_dog=='True' & p3_dog =='False'):
        dog_breed = p2
    elif (p1_dog=='True' & p2_dog=='False' & p3_dog =='True'):
        dog_breed = p3

    # scenario 3: two True cases
    # p1_dog and p2_dog == 'True'
    elif (p1_dog=='True' & p2_dog=='True' & p3_dog=='False' & p1_conf >= p2_conf):
        dog_breed = p1
    elif (p1_dog=='True' & p2_dog=='True' & p3_dog=='False' & p1_conf < p2_conf):
        dog_breed = p2

    # p1_dog and p3_dog == 'True'
    elif (p1_dog=='True' & p3_dog=='True' & p2_dog=='False' & p1_conf >= p3_conf):
        dog_breed = p1
    elif (p1_dog=='True' & p3_dog=='True' & p2_dog=='False' & p1_conf < p3_conf):
        dog_breed = p3

    # p2_dog and p3_dog == 'True'
    elif (p2_dog=='True' & p3_dog=='True' & p1_dog=='False' & p2_conf >= p3_conf):
        dog_breed = p2
    elif (p2_dog=='True' & p3_dog=='True' & p1_dog=='False' & p2_conf < p3_conf):
        dog_breed = p3


    # scenario 4: all cases are True
    # p1 >= p2 & p3
    elif (p1_dog=='True' & p2_dog=='True' & p3_dog=='True' & p1 >= p2 & p1 >= p3):
        dog_breed = p1
    #p2 > p1 & p2 >= P3
    elif (p1_dog=='True' & p2_dog=='True' & p3_dog=='True' & p2 > p1 & p2 >= p3):
        dog_breed = p2

    # p3 > p1 & p2
    else:
        dog_breed = p3


  File "<ipython-input-88-9d628b0f39ad>", line 3
for p1, p2 p3, p1_dog, p2_dog, p3_dog, p1_conf, p2_conf, p3_conf in df_breed:
          ^
```

```
    SyntaxError: invalid syntax
```

In [ ]: