wrangle_act

July 26, 2018

```
In [1]: import pandas as pd
        import requests
        import os
        import tweepy
        import re
        import requests
        from bs4 import BeautifulSoup
        import urllib

        import nltk
        from nltk.corpus import stopwords
        nltk.download('punkt')
        nltk.download('averaged_perceptron_tagger')
        nltk.download('stopwords')


        from keras.applications.resnet50 import ResNet50
        from keras.preprocessing import image
        from keras.applications.resnet50 import preprocess_input, decode_predictions
        model = ResNet50(weights='imagenet')

        import numpy as np
        import datetime
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
Using TensorFlow backend.
```

```
Downloading data from https://github.com/fchollet/deep-learning-models/releases/download/v0.2/re
102858752/102853048 [==============================] - 5s 0us/step
```

1 1. Gather Data

Load the twitter data in hand through twitter-archive-enhanced.csv

Load the dog name prediction data availed in image-predictions.tsv

```
In [2]: #load the twitter info available at hand
        twitter_archive_enhanced = pd.read_csv("twitter-archive-enhanced.csv")
        twitter_archive_cleaned = twitter_archive_enhanced.copy()
        twitter_archive_cleaned
```

```
Out[2]:                 tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
        0     892420643555336193                    NaN                  NaN
        1     892177421306343426                    NaN                  NaN
        2     891815181378084864                    NaN                  NaN
        3     891689557279858688                    NaN                  NaN
        4     891327558926688256                    NaN                  NaN
        5     891087950875897856                    NaN                  NaN
        6     890971913173991426                    NaN                  NaN
        7     890729181411237888                    NaN                  NaN
        8     890609185150312448                    NaN                  NaN
        9     890240255349198849                    NaN                  NaN
        10    890006608113172480                    NaN                  NaN
        11    889880896479866881                    NaN                  NaN
        12    889665388333682689                    NaN                  NaN
        13    889638837579907072                    NaN                  NaN
        14    889531135344209921                    NaN                  NaN
        15    889278841981685760                    NaN                  NaN
        16    888917238123831296                    NaN                  NaN
        17    888804989199671297                    NaN                  NaN
        18    888554962724278272                    NaN                  NaN
        19    888202515573088257                    NaN                  NaN
        20    888078434458587136                    NaN                  NaN
        21    887705289381826560                    NaN                  NaN
        22    887517139158093824                    NaN                  NaN
        23    887473957103951883                    NaN                  NaN
        24    887343217045368832                    NaN                  NaN
        25    887101392804085760                    NaN                  NaN
        26    886983233522544640                    NaN                  NaN
        27    886736880519319552                    NaN                  NaN
        28    886680336477933568                    NaN                  NaN
        29    886366144734445568                    NaN                  NaN
        ...                  ...                    ...                  ...
        2326  666411507551481857                    NaN                  NaN
        2327  666407126856765440                    NaN                  NaN
        2328  666396247373291520                    NaN                  NaN
        2329  666373753744588802                    NaN                  NaN
        2330  666362758909284353                    NaN                  NaN
```

```
2331  666353288456101888                    NaN                  NaN
2332  666345417576210432                    NaN                  NaN
2333  666337882303524864                    NaN                  NaN
2334  666293911632134144                    NaN                  NaN
2335  666287406224695296                    NaN                  NaN
2336  666273097616637952                    NaN                  NaN
2337  666268910803644416                    NaN                  NaN
2338  666104133288665088                    NaN                  NaN
2339  666102155909144576                    NaN                  NaN
2340  666099513787052032                    NaN                  NaN
2341  666094000022159362                    NaN                  NaN
2342  666082916733198337                    NaN                  NaN
2343  666073100786774016                    NaN                  NaN
2344  666071193221509120                    NaN                  NaN
2345  666063827256086533                    NaN                  NaN
2346  666058600524156928                    NaN                  NaN
2347  666057090499244032                    NaN                  NaN
2348  666055525042405380                    NaN                  NaN
2349  666051853826850816                    NaN                  NaN
2350  666050758794694657                    NaN                  NaN
2351  666049248165822465                    NaN                  NaN
2352  666044226329800704                    NaN                  NaN
2353  666033412701032449                    NaN                  NaN
2354  666029285002620928                    NaN                  NaN
2355  666020888022790149                    NaN                  NaN

                      timestamp  \
0     2017-08-01 16:23:56 +0000
1     2017-08-01 00:17:27 +0000
2     2017-07-31 00:18:03 +0000
3     2017-07-30 15:58:51 +0000
4     2017-07-29 16:00:24 +0000
5     2017-07-29 00:08:17 +0000
6     2017-07-28 16:27:12 +0000
7     2017-07-28 00:22:40 +0000
8     2017-07-27 16:25:51 +0000
9     2017-07-26 15:59:51 +0000
10    2017-07-26 00:31:25 +0000
11    2017-07-25 16:11:53 +0000
12    2017-07-25 01:55:32 +0000
13    2017-07-25 00:10:02 +0000
14    2017-07-24 17:02:04 +0000
15    2017-07-24 00:19:32 +0000
16    2017-07-23 00:22:39 +0000
17    2017-07-22 16:56:37 +0000
18    2017-07-22 00:23:06 +0000
19    2017-07-21 01:02:36 +0000
20    2017-07-20 16:49:33 +0000
```

```
21    2017-07-19 16:06:48 +0000
22    2017-07-19 03:39:09 +0000
23    2017-07-19 00:47:34 +0000
24    2017-07-18 16:08:03 +0000
25    2017-07-18 00:07:08 +0000
26    2017-07-17 16:17:36 +0000
27    2017-07-16 23:58:41 +0000
28    2017-07-16 20:14:00 +0000
29    2017-07-15 23:25:31 +0000
...                         ...
2326  2015-11-17 00:24:19 +0000
2327  2015-11-17 00:06:54 +0000
2328  2015-11-16 23:23:41 +0000
2329  2015-11-16 21:54:18 +0000
2330  2015-11-16 21:10:36 +0000
2331  2015-11-16 20:32:58 +0000
2332  2015-11-16 20:01:42 +0000
2333  2015-11-16 19:31:45 +0000
2334  2015-11-16 16:37:02 +0000
2335  2015-11-16 16:11:11 +0000
2336  2015-11-16 15:14:19 +0000
2337  2015-11-16 14:57:41 +0000
2338  2015-11-16 04:02:55 +0000
2339  2015-11-16 03:55:04 +0000
2340  2015-11-16 03:44:34 +0000
2341  2015-11-16 03:22:39 +0000
2342  2015-11-16 02:38:37 +0000
2343  2015-11-16 01:59:36 +0000
2344  2015-11-16 01:52:02 +0000
2345  2015-11-16 01:22:45 +0000
2346  2015-11-16 01:01:59 +0000
2347  2015-11-16 00:55:59 +0000
2348  2015-11-16 00:49:46 +0000
2349  2015-11-16 00:35:11 +0000
2350  2015-11-16 00:30:50 +0000
2351  2015-11-16 00:24:50 +0000
2352  2015-11-16 00:04:52 +0000
2353  2015-11-15 23:21:54 +0000
2354  2015-11-15 23:05:30 +0000
2355  2015-11-15 22:32:08 +0000

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

```
6       <a href="http://twitter.com/download/iphone" r...
7       <a href="http://twitter.com/download/iphone" r...
8       <a href="http://twitter.com/download/iphone" r...
9       <a href="http://twitter.com/download/iphone" r...
10      <a href="http://twitter.com/download/iphone" r...
11      <a href="http://twitter.com/download/iphone" r...
12      <a href="http://twitter.com/download/iphone" r...
13      <a href="http://twitter.com/download/iphone" r...
14      <a href="http://twitter.com/download/iphone" r...
15      <a href="http://twitter.com/download/iphone" r...
16      <a href="http://twitter.com/download/iphone" r...
17      <a href="http://twitter.com/download/iphone" r...
18      <a href="http://twitter.com/download/iphone" r...
19      <a href="http://twitter.com/download/iphone" r...
20      <a href="http://twitter.com/download/iphone" r...
21      <a href="http://twitter.com/download/iphone" r...
22      <a href="http://twitter.com/download/iphone" r...
23      <a href="http://twitter.com/download/iphone" r...
24      <a href="http://twitter.com/download/iphone" r...
25      <a href="http://twitter.com/download/iphone" r...
26      <a href="http://twitter.com/download/iphone" r...
27      <a href="http://twitter.com/download/iphone" r...
28      <a href="http://twitter.com/download/iphone" r...
29      <a href="http://twitter.com/download/iphone" r...
...                                                   ...
2326    <a href="http://twitter.com/download/iphone" r...
2327    <a href="http://twitter.com/download/iphone" r...
2328    <a href="http://twitter.com/download/iphone" r...
2329    <a href="http://twitter.com/download/iphone" r...
2330    <a href="http://twitter.com/download/iphone" r...
2331    <a href="http://twitter.com/download/iphone" r...
2332    <a href="http://twitter.com/download/iphone" r...
2333    <a href="http://twitter.com/download/iphone" r...
2334    <a href="http://twitter.com/download/iphone" r...
2335    <a href="http://twitter.com/download/iphone" r...
2336    <a href="http://twitter.com/download/iphone" r...
2337    <a href="http://twitter.com/download/iphone" r...
2338    <a href="http://twitter.com/download/iphone" r...
2339    <a href="http://twitter.com/download/iphone" r...
2340    <a href="http://twitter.com/download/iphone" r...
2341    <a href="http://twitter.com/download/iphone" r...
2342    <a href="http://twitter.com/download/iphone" r...
2343    <a href="http://twitter.com/download/iphone" r...
2344    <a href="http://twitter.com/download/iphone" r...
2345    <a href="http://twitter.com/download/iphone" r...
2346    <a href="http://twitter.com/download/iphone" r...
2347    <a href="http://twitter.com/download/iphone" r...
2348    <a href="http://twitter.com/download/iphone" r...
```

```
2349  <a href="http://twitter.com/download/iphone" r...
2350  <a href="http://twitter.com/download/iphone" r...
2351  <a href="http://twitter.com/download/iphone" r...
2352  <a href="http://twitter.com/download/iphone" r...
2353  <a href="http://twitter.com/download/iphone" r...
2354  <a href="http://twitter.com/download/iphone" r...
2355  <a href="http://twitter.com/download/iphone" r...

                                                   text  retweeted_status_id  \
0     This is Phineas. He's a mystical boy. Only eve...                  NaN
1     This is Tilly. She's just checking pup on you...                  NaN
2     This is Archie. He is a rare Norwegian Pouncin...                  NaN
3     This is Darla. She commenced a snooze mid meal...                  NaN
4     This is Franklin. He would like you to stop ca...                  NaN
5     Here we have a majestic great white breaching ...                  NaN
6     Meet Jax. He enjoys ice cream so much he gets ...                  NaN
7     When you watch your owner call another dog a g...                  NaN
8     This is Zoey. She doesn't want to be one of th...                  NaN
9     This is Cassie. She is a college pup. Studying...                  NaN
10    This is Koda. He is a South Australian decksha...                  NaN
11    This is Bruno. He is a service shark. Only get...                  NaN
12    Here's a puppo that seems to be on the fence a...                  NaN
13    This is Ted. He does his best. Sometimes that'...                  NaN
14    This is Stuart. He's sporting his favorite fan...                  NaN
15    This is Oliver. You're witnessing one of his m...                  NaN
16    This is Jim. He found a fren. Taught him how t...                  NaN
17    This is Zeke. He has a new stick. Very proud o...                  NaN
18    This is Ralphus. He's powering up. Attempting ...                  NaN
19    RT @dog_rates: This is Canela. She attempted s...         8.874740e+17
20    This is Gerald. He was just told he didn't get...                  NaN
21    This is Jeffrey. He has a monopoly on the pool...                  NaN
22    I've yet to rate a Venezuelan Hover Wiener. Th...                  NaN
23    This is Canela. She attempted some fancy porch...                  NaN
24    You may not have known you needed to see this ...                  NaN
25    This... is a Jubilant Antarctic House Bear. We...                  NaN
26    This is Maya. She's very shy. Rarely leaves he...                  NaN
27    This is Mingus. He's a wonderful father to his...                  NaN
28    This is Derek. He's late for a dog meeting. 13...                  NaN
29    This is Roscoe. Another pupper fallen victim t...                  NaN
...                                                 ...                  ...
2326  This is quite the dog. Gets really excited whe...                  NaN
2327  This is a southern Vesuvius bumblegruff. Can d...                  NaN
2328  Oh goodness. A super rare northeast Qdoba kang...                  NaN
2329  Those are sunglasses and a jean jacket. 11/10 ...                  NaN
2330  Unique dog here. Very small. Lives in containe...                  NaN
2331  Here we have a mixed Asiago from the Galápagos...                  NaN
2332  Look at this jokester thinking seat belt laws ...                  NaN
2333  This is an extremely rare horned Parthenon. No...                  NaN
```

```
2334  This is a funny dog. Weird toes. Won't come do...                  NaN
2335  This is an Albanian 3 1/2 legged  Episcopalian...                  NaN
2336     Can take selfies 11/10 https://t.co/ws2AMaNwPW                  NaN
2337  Very concerned about fellow dog trapped in com...                  NaN
2338  Not familiar with this breed. No tail (weird)...                  NaN
2339  Oh my. Here you are seeing an Adobe Setter giv...                  NaN
2340  Can stand on stump for what seems like a while...                  NaN
2341  This appears to be a Mongolian Presbyterian mi...                  NaN
2342  Here we have a well-established sunblockerspan...                  NaN
2343  Let's hope this flight isn't Malaysian (lol). ...                  NaN
2344  Here we have a northern speckled Rhododendron...                  NaN
2345  This is the happiest dog you will ever see. Ve...                  NaN
2346  Here is the Rand Paul of retrievers folks! He'...                  NaN
2347  My oh my. This is a rare blond Canadian terrie...                  NaN
2348  Here is a Siberian heavily armored polar bear ...                  NaN
2349  This is an odd dog. Hard on the outside but lo...                  NaN
2350  This is a truly beautiful English Wilson Staff...                  NaN
2351  Here we have a 1949 1st generation vulpix. Enj...                  NaN
2352  This is a purebred Piers Morgan. Loves to Netf...                  NaN
2353  Here is a very happy pup. Big fan of well-main...                  NaN
2354  This is a western brown Mitsubishi terrier. Up...                  NaN
2355  Here we have a Japanese Irish Setter. Lost eye...                  NaN

      retweeted_status_user_id retweeted_status_timestamp  \
0                          NaN                        NaN
1                          NaN                        NaN
2                          NaN                        NaN
3                          NaN                        NaN
4                          NaN                        NaN
5                          NaN                        NaN
6                          NaN                        NaN
7                          NaN                        NaN
8                          NaN                        NaN
9                          NaN                        NaN
10                         NaN                        NaN
11                         NaN                        NaN
12                         NaN                        NaN
13                         NaN                        NaN
14                         NaN                        NaN
15                         NaN                        NaN
16                         NaN                        NaN
17                         NaN                        NaN
18                         NaN                        NaN
19                4.196984e+09  2017-07-19 00:47:34 +0000
20                         NaN                        NaN
21                         NaN                        NaN
22                         NaN                        NaN
23                         NaN                        NaN
```

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

                                          expanded_urls  rating_numerator  \
0     https://twitter.com/dog_rates/status/892420643...                13
1     https://twitter.com/dog_rates/status/892177421...                13
2     https://twitter.com/dog_rates/status/891815181...                12
3     https://twitter.com/dog_rates/status/891689557...                13
4     https://twitter.com/dog_rates/status/891327558...                12
5     https://twitter.com/dog_rates/status/891087950...                13
6     https://gofundme.com/ydvmve-surgery-for-jax,ht...                13
7     https://twitter.com/dog_rates/status/890729181...                13
8     https://twitter.com/dog_rates/status/890609185...                13
```

```
9     https://twitter.com/dog_rates/status/890240255...                14
10    https://twitter.com/dog_rates/status/890006608...                13
11    https://twitter.com/dog_rates/status/889880896...                13
12    https://twitter.com/dog_rates/status/889665388...                13
13    https://twitter.com/dog_rates/status/889638837...                12
14    https://twitter.com/dog_rates/status/889531135...                13
15    https://twitter.com/dog_rates/status/889278841...                13
16    https://twitter.com/dog_rates/status/888917238...                12
17    https://twitter.com/dog_rates/status/888804989...                13
18    https://twitter.com/dog_rates/status/888554962...                13
19    https://twitter.com/dog_rates/status/887473957...                13
20    https://twitter.com/dog_rates/status/888078434...                12
21    https://twitter.com/dog_rates/status/887705289...                13
22    https://twitter.com/dog_rates/status/887517139...                14
23    https://twitter.com/dog_rates/status/887473957...                13
24    https://twitter.com/dog_rates/status/887343217...                13
25    https://twitter.com/dog_rates/status/887101392...                12
26    https://twitter.com/dog_rates/status/886983233...                13
27    https://www.gofundme.com/mingusneedsus,https:/...                13
28    https://twitter.com/dog_rates/status/886680336...                13
29    https://twitter.com/dog_rates/status/886366144...                12
...                                                 ...               ...
2326  https://twitter.com/dog_rates/status/666411507...                 2
2327  https://twitter.com/dog_rates/status/666407126...                 7
2328  https://twitter.com/dog_rates/status/666396247...                 9
2329  https://twitter.com/dog_rates/status/666373753...                11
2330  https://twitter.com/dog_rates/status/666362758...                 6
2331  https://twitter.com/dog_rates/status/666353288...                 8
2332  https://twitter.com/dog_rates/status/666345417...                10
2333  https://twitter.com/dog_rates/status/666337882...                 9
2334  https://twitter.com/dog_rates/status/666293911...                 3
2335  https://twitter.com/dog_rates/status/666287406...                 1
2336  https://twitter.com/dog_rates/status/666273097...                11
2337  https://twitter.com/dog_rates/status/666268910...                10
2338  https://twitter.com/dog_rates/status/666104133...                 1
2339  https://twitter.com/dog_rates/status/666102155...                11
2340  https://twitter.com/dog_rates/status/666099513...                 8
2341  https://twitter.com/dog_rates/status/666094000...                 9
2342  https://twitter.com/dog_rates/status/666082916...                 6
2343  https://twitter.com/dog_rates/status/666073100...                10
2344  https://twitter.com/dog_rates/status/666071193...                 9
2345  https://twitter.com/dog_rates/status/666063827...                10
2346  https://twitter.com/dog_rates/status/666058600...                 8
2347  https://twitter.com/dog_rates/status/666057090...                 9
2348  https://twitter.com/dog_rates/status/666055525...                10
2349  https://twitter.com/dog_rates/status/666051853...                 2
2350  https://twitter.com/dog_rates/status/666050758...                10
2351  https://twitter.com/dog_rates/status/666049248...                 5
```

```
2352  https://twitter.com/dog_rates/status/666044226...                 6
2353  https://twitter.com/dog_rates/status/666033412...                 9
2354  https://twitter.com/dog_rates/status/666029285...                 7
2355  https://twitter.com/dog_rates/status/666020888...                 8

      rating_denominator      name  doggo floofer  pupper  puppo
0                     10   Phineas   None    None    None   None
1                     10     Tilly   None    None    None   None
2                     10    Archie   None    None    None   None
3                     10     Darla   None    None    None   None
4                     10  Franklin   None    None    None   None
5                     10      None   None    None    None   None
6                     10       Jax   None    None    None   None
7                     10      None   None    None    None   None
8                     10      Zoey   None    None    None   None
9                     10    Cassie  doggo    None    None   None
10                    10      Koda   None    None    None   None
11                    10     Bruno   None    None    None   None
12                    10      None   None    None    None  puppo
13                    10       Ted   None    None    None   None
14                    10    Stuart   None    None    None  puppo
15                    10    Oliver   None    None    None   None
16                    10       Jim   None    None    None   None
17                    10      Zeke   None    None    None   None
18                    10   Ralphus   None    None    None   None
19                    10    Canela   None    None    None   None
20                    10    Gerald   None    None    None   None
21                    10   Jeffrey   None    None    None   None
22                    10      such   None    None    None   None
23                    10    Canela   None    None    None   None
24                    10      None   None    None    None   None
25                    10      None   None    None    None   None
26                    10      Maya   None    None    None   None
27                    10    Mingus   None    None    None   None
28                    10     Derek   None    None    None   None
29                    10    Roscoe   None    None  pupper   None
...                  ...       ...    ...     ...     ...    ...
2326                  10     quite   None    None    None   None
2327                  10         a   None    None    None   None
2328                  10      None   None    None    None   None
2329                  10      None   None    None    None   None
2330                  10      None   None    None    None   None
2331                  10      None   None    None    None   None
2332                  10      None   None    None    None   None
2333                  10        an   None    None    None   None
2334                  10         a   None    None    None   None
2335                   2        an   None    None    None   None
2336                  10      None   None    None    None   None
```

```
2337                  10      None   None    None    None   None
2338                  10      None   None    None    None   None
2339                  10      None   None    None    None   None
2340                  10      None   None    None    None   None
2341                  10      None   None    None    None   None
2342                  10      None   None    None    None   None
2343                  10      None   None    None    None   None
2344                  10      None   None    None    None   None
2345                  10       the   None    None    None   None
2346                  10       the   None    None    None   None
2347                  10         a   None    None    None   None
2348                  10         a   None    None    None   None
2349                  10        an   None    None    None   None
2350                  10         a   None    None    None   None
2351                  10      None   None    None    None   None
2352                  10         a   None    None    None   None
2353                  10         a   None    None    None   None
2354                  10         a   None    None    None   None
2355                  10      None   None    None    None   None

[2356 rows x 17 columns]
```

```
In [3]: #download the image-predictions file from Udacity Server only if it does not exists
        if not os.path.exists("image-predictions.tsv"):
            with open ("image-predictions.tsv",'wb') as fp:
                response = requests.get("https://d17h27t6h515a5.cloudfront.net/topher/2017/Augus
                fp.write(response.content)

        #load it and visually inspect
        image_predictions = pd.read_csv("image-predictions.tsv",delimiter="\t")
        image_predictions
```

```
Out[3]:               tweet_id                                          jpg_url  \
        0     666020888022790149   https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg
        1     666029285002620928   https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg
        2     666033412701032449   https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg
        3     666044226329800704   https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg
        4     666049248165822465   https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg
        5     666050758794694657   https://pbs.twimg.com/media/CT5Jof1WUAEuVxN.jpg
        6     666051853826850816   https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg
        7     666055525042405380   https://pbs.twimg.com/media/CT5N9tpXIAAifs1.jpg
        8     666057090499244032   https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg
        9     666058600524156928   https://pbs.twimg.com/media/CT5Qw94XAAA_2dP.jpg
        10    666063827256086533   https://pbs.twimg.com/media/CT5Vg_wXIAAXfnj.jpg
        11    666071193221509120   https://pbs.twimg.com/media/CT5cN_3WEAAlOoZ.jpg
        12    666073100786774016   https://pbs.twimg.com/media/CT5d9DZXAAALcwe.jpg
        13    666082916733198337   https://pbs.twimg.com/media/CT5m4VGWEAAtKc8.jpg
        14    666094000022159362   https://pbs.twimg.com/media/CT5w9gUW4AAsBNN.jpg
```

```
15    666099513787052032    https://pbs.twimg.com/media/CT51-JJUEAA6hV8.jpg
16    666102155909144576    https://pbs.twimg.com/media/CT54YGiWUAEZnoK.jpg
17    666104133288665088    https://pbs.twimg.com/media/CT56LSZWoAAlJj2.jpg
18    666268910803644416    https://pbs.twimg.com/media/CT8QCd1WEAADXws.jpg
19    666273097616637952    https://pbs.twimg.com/media/CT8T1mtUwAA3aqm.jpg
20    666287406224695296    https://pbs.twimg.com/media/CT8g3BpUEAAuFjg.jpg
21    666293911632134144    https://pbs.twimg.com/media/CT8mx7KW4AEQu8N.jpg
22    666337882303524864    https://pbs.twimg.com/media/CT9OwFIWEAMuRje.jpg
23    666345417576210432    https://pbs.twimg.com/media/CT9Vn7PWoAA_ZCM.jpg
24    666353288456101888    https://pbs.twimg.com/media/CT9cx0tUEAAhNN_.jpg
25    666362758909284353    https://pbs.twimg.com/media/CT9lXGsUcAAyUFt.jpg
26    666373753744588802    https://pbs.twimg.com/media/CT9vZEYWUAAlZ05.jpg
27    666396247373291520    https://pbs.twimg.com/media/CT-D2ZHWIAA3gK1.jpg
28    666407126856765440    https://pbs.twimg.com/media/CT-NvwmW4AAugGZ.jpg
29    666411507551481857    https://pbs.twimg.com/media/CT-RugiWIAELEaq.jpg
...                  ...                                                  ...
2045  886366144734445568    https://pbs.twimg.com/media/DE0BTnQUwAApKEH.jpg
2046  886680336477933568    https://pbs.twimg.com/media/DE4fEDzWAAAyHMM.jpg
2047  886736880519319552    https://pbs.twimg.com/media/DE5Se8FXcAAJFx4.jpg
2048  886983233522544640    https://pbs.twimg.com/media/DE8yicJW0AAAvBJ.jpg
2049  887101392804085760    https://pbs.twimg.com/media/DE-eAq6UwAA-jaE.jpg
2050  887343217045368832  https://pbs.twimg.com/ext_tw_video_thumb/88734...
2051  887473957103951883    https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg
2052  887517139158093824  https://pbs.twimg.com/ext_tw_video_thumb/88751...
2053  887705289381826560    https://pbs.twimg.com/media/DFHDQBbXgAEqY7t.jpg
2054  888078434458587136    https://pbs.twimg.com/media/DFMWn56WsAAkA7B.jpg
2055  888202515573088257    https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg
2056  888554962724278272    https://pbs.twimg.com/media/DFTH_O-UQAACu20.jpg
2057  888804989199671297    https://pbs.twimg.com/media/DFWra-3VYAA2piG.jpg
2058  888917238123831296    https://pbs.twimg.com/media/DFYRgsOUQAARGhO.jpg
2059  889278841981685760  https://pbs.twimg.com/ext_tw_video_thumb/88927...
2060  889531135344209921    https://pbs.twimg.com/media/DFg_2PVW0AEHN3p.jpg
2061  889638837579907072    https://pbs.twimg.com/media/DFihzFfXsAYGDPR.jpg
2062  889665388333682689    https://pbs.twimg.com/media/DFi579UWsAAatzw.jpg
2063  889880896479866881    https://pbs.twimg.com/media/DFl99B1WsAITKsg.jpg
2064  890006608113172480    https://pbs.twimg.com/media/DFnwSY4WAAAMliS.jpg
2065  890240255349198849    https://pbs.twimg.com/media/DFrEyVuW0AAO3t9.jpg
2066  890609185150312448    https://pbs.twimg.com/media/DFwUU__XcAEpyXI.jpg
2067  890729181411237888    https://pbs.twimg.com/media/DFyBahAVwAAhUTd.jpg
2068  890971913173991426    https://pbs.twimg.com/media/DF1eOmZXUAALUcq.jpg
2069  891087950875897856    https://pbs.twimg.com/media/DF3HwyEWsAABqE6.jpg
2070  891327558926688256    https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg
2071  891689557279858688    https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg
2072  891815181378084864    https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg
2073  892177421306343426    https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg
2074  892420643555336193    https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg

      img_num                           p1   p1_conf  p1_dog  \
0           1        Welsh_springer_spaniel 0.465074    True
1           1                       redbone 0.506826    True
2           1               German_shepherd 0.596461    True
3           1           Rhodesian_ridgeback 0.408143    True
4           1            miniature_pinscher 0.560311    True
5           1          Bernese_mountain_dog 0.651137    True
6           1                    box_turtle 0.933012   False
7           1                          chow 0.692517    True
8           1                 shopping_cart 0.962465   False
9           1              miniature_poodle 0.201493    True
10          1              golden_retriever 0.775930    True
11          1                 Gordon_setter 0.503672    True
12          1                  Walker_hound 0.260857    True
13          1                           pug 0.489814    True
14          1                    bloodhound 0.195217    True
15          1                         Lhasa 0.582330    True
16          1                English_setter 0.298617    True
17          1                           hen 0.965932   False
18          1              desktop_computer 0.086502   False
19          1             Italian_greyhound 0.176053    True
20          1                   Maltese_dog 0.857531    True
21          1              three-toed_sloth 0.914671   False
22          1                            ox 0.416669   False
23          1              golden_retriever 0.858744    True
24          1                      malamute 0.336874    True
25          1                    guinea_pig 0.996496   False
26          1   soft-coated_wheaten_terrier 0.326467    True
27          1                     Chihuahua 0.978108    True
28          1       black-and-tan_coonhound 0.529139    True
29          1                          coho 0.404640   False
...       ...                           ...      ...     ...
2045        1                French_bulldog 0.999201    True
2046        1                   convertible 0.738995   False
2047        1                        kuvasz 0.309706    True
2048        2                     Chihuahua 0.793469    True
2049        1                       Samoyed 0.733942    True
2050        1              Mexican_hairless 0.330741    True
2051        2                      Pembroke 0.809197    True
2052        1                     limousine 0.130432   False
2053        1                        basset 0.821664    True
2054        1                French_bulldog 0.995026    True
2055        2                      Pembroke 0.809197    True
2056        3                Siberian_husky 0.700377    True
2057        1              golden_retriever 0.469760    True
2058        1              golden_retriever 0.714719    True
2059        1                       whippet 0.626152    True
2060        1              golden_retriever 0.953442    True
2061        1                French_bulldog 0.991650    True
```

```
2062        1                     Pembroke  0.966327    True
2063        1               French_bulldog  0.377417    True
2064        1                      Samoyed  0.957979    True
2065        1                     Pembroke  0.511319    True
2066        1                Irish_terrier  0.487574    True
2067        2                   Pomeranian  0.566142    True
2068        1                  Appenzeller  0.341703    True
2069        1     Chesapeake_Bay_retriever  0.425595    True
2070        2                       basset  0.555712    True
2071        1                  paper_towel  0.170278   False
2072        1                    Chihuahua  0.716012    True
2073        1                    Chihuahua  0.323581    True
2074        1                       orange  0.097049   False

                            p2   p2_conf  p2_dog                           p3  \
0                       collie  0.156665    True            Shetland_sheepdog
1           miniature_pinscher  0.074192    True          Rhodesian_ridgeback
2                     malinois  0.138584    True                   bloodhound
3                      redbone  0.360687    True           miniature_pinscher
4                   Rottweiler  0.243682    True                     Doberman
5             English_springer  0.263788    True   Greater_Swiss_Mountain_dog
6                   mud_turtle  0.045885   False                     terrapin
7              Tibetan_mastiff  0.058279    True                     fur_coat
8              shopping_basket  0.014594   False             golden_retriever
9                     komondor  0.192305    True  soft-coated_wheaten_terrier
10             Tibetan_mastiff  0.093718    True           Labrador_retriever
11           Yorkshire_terrier  0.174201    True                     Pekinese
12            English_foxhound  0.175382    True                  Ibizan_hound
13                bull_mastiff  0.404722    True               French_bulldog
14             German_shepherd  0.078260    True                     malinois
15                    Shih-Tzu  0.166192    True               Dandie_Dinmont
16                Newfoundland  0.149842    True                       borzoi
17                        cock  0.033919   False                    partridge
18                        desk  0.085547   False                     bookcase
19                 toy_terrier  0.111884    True                      basenji
20                  toy_poodle  0.063064    True             miniature_poodle
21                       otter  0.015250   False               great_grey_owl
22                Newfoundland  0.278407    True                  groenendael
23    Chesapeake_Bay_retriever  0.054787    True           Labrador_retriever
24              Siberian_husky  0.147655    True                   Eskimo_dog
25                       skunk  0.002402   False                      hamster
26                Afghan_hound  0.259551    True                       briard
27                 toy_terrier  0.009397    True                     papillon
28                  bloodhound  0.244220    True        flat-coated_retriever
29                  barracouta  0.271485   False                          gar
...                        ...       ...     ...                          ...
2045                 Chihuahua  0.000361    True                  Boston_bull
2046                sports_car  0.139952   False                    car_wheel
```

```
2047            Great_Pyrenees  0.186136    True               Dandie_Dinmont
2048               toy_terrier  0.143528    True                   can_opener
2049                Eskimo_dog  0.035029    True    Staffordshire_bullterrier
2050                  sea_lion  0.275645   False                   Weimaraner
2051       Rhodesian_ridgeback  0.054950    True                       beagle
2052                 tow_truck  0.029175   False                shopping_cart
2053                   redbone  0.087582    True                   Weimaraner
2054                       pug  0.000932    True                 bull_mastiff
2055       Rhodesian_ridgeback  0.054950    True                       beagle
2056                Eskimo_dog  0.166511    True                     malamute
2057        Labrador_retriever  0.184172    True               English_setter
2058           Tibetan_mastiff  0.120184    True           Labrador_retriever
2059                    borzoi  0.194742    True                       Saluki
2060        Labrador_retriever  0.013834    True                      redbone
2061                     boxer  0.002129    True    Staffordshire_bullterrier
2062                  Cardigan  0.027356    True                      basenji
2063        Labrador_retriever  0.151317    True                       muzzle
2064                Pomeranian  0.013884    True                         chow
2065                  Cardigan  0.451038    True                    Chihuahua
2066              Irish_setter  0.193054    True     Chesapeake_Bay_retriever
2067                Eskimo_dog  0.178406    True                     Pembroke
2068             Border_collie  0.199287    True                    ice_lolly
2069             Irish_terrier  0.116317    True              Indian_elephant
2070          English_springer  0.225770    True  German_short-haired_pointer
2071        Labrador_retriever  0.168086    True                      spatula
2072                  malamute  0.078253    True                       kelpie
2073                  Pekinese  0.090647    True                     papillon
2074                     bagel  0.085851   False                       banana

       p3_conf  p3_dog
0     0.061428    True
1     0.072010    True
2     0.116197    True
3     0.222752    True
4     0.154629    True
5     0.016199    True
6     0.017885   False
7     0.054449   False
8     0.007959    True
9     0.082086    True
10    0.072427    True
11    0.109454    True
12    0.097471    True
13    0.048960    True
14    0.075628    True
15    0.089688    True
16    0.133649    True
17    0.000052   False
```

```
18    0.079480   False
19    0.111152    True
20    0.025581    True
21    0.013207   False
22    0.102643    True
23    0.014241    True
24    0.093412    True
25    0.000461   False
26    0.206803    True
27    0.004577    True
28    0.173810    True
29    0.189945   False
...        ...     ...
2045  0.000076    True
2046  0.044173   False
2047  0.086346    True
2048  0.032253   False
2049  0.029705    True
2050  0.134203    True
2051  0.038915    True
2052  0.026321   False
2053  0.026236    True
2054  0.000903    True
2055  0.038915    True
2056  0.111411    True
2057  0.073482    True
2058  0.105506    True
2059  0.027351    True
2060  0.007958    True
2061  0.001498    True
2062  0.004633    True
2063  0.082981   False
2064  0.008167    True
2065  0.029248    True
2066  0.118184    True
2067  0.076507    True
2068  0.193548   False
2069  0.076902   False
2070  0.175219    True
2071  0.040836   False
2072  0.031379    True
2073  0.068957    True
2074  0.076110   False

[2075 rows x 12 columns]
```

2 2. Assess the Data

2.0.1 Quality isssues:

1. Completeness:

- 1976 entries do not have ‘Stage’ info.
- 281 entries of dog breed names are missing.

2. Validity:

- Name do not comply to the naming standard; there are 55 entries as “a”.
- There are ratings given on reply text message. we are trying find a trend in the root message not the subsequent discussion for the tweet. The reason being, one we do not have access to all reply messages. Two, the study is on weratedogs tweet analysis and not replies :)
- Erraneous datatype -
    - timestamp, retweeted_status_timestamp (to_datetime),
    - all “_id" to int64 as integer operations are faster than string. And also makes like easy to do simple logical operator in subset selection.

3. Accuracy:

- p2_* & p3_* are not required as we have p1 > p2 > p3
- Dog breed names also has names other than dog breed names. p1_dog indicates that whether the name identified is dog name or not.
- rating_numerator is not float type and hence 13.5, 11.27, 11.26, 9.75 & 9.5 rating are missing in the data set.

4. Consistency:

- source variable do not have the variables stored along with html tag; like other variable object variables should have the core value and not the residual of the extract.
- Dog breed names are not consistent with respect to usage of capital letters.

2.0.2 Tidiness Issues:

- Twitter archive and image prediction can be combined to store only dog breed name.
- Dog stages can be melted to a column and can be considered as category variable

```
In [4]: #1.a 1976 entries do not have 'Stage' info.

        temp_un = twitter_archive_cleaned[twitter_archive_cleaned.pupper == 'None']
        temp_un = temp_un[temp_un.puppo == "None"]
        temp_un = temp_un[temp_un.floofer == "None"]
        temp_un = temp_un[temp_un.doggo == 'None']
        print("In uncleaned dataset: \nUnknown Count :",temp_un.shape[0])
        print("Pupper Count :", twitter_archive_cleaned[twitter_archive_cleaned.pupper != 'None'
        print("Doggo Count :", twitter_archive_cleaned[twitter_archive_cleaned.doggo != 'None'].
        print("Floof Count :", twitter_archive_cleaned[twitter_archive_cleaned.floofer != 'None'
        print("Puppo Count :", twitter_archive_cleaned[twitter_archive_cleaned.puppo != 'None'].
In uncleaned dataset:
Unknown Count : 1976
Pupper Count : 257
Doggo Count : 97
Floof Count : 10
Puppo Count : 30
```

```
In [5]: #1.b number of entries with dog breed names are missing.
        twitter_archive_cleaned = twitter_archive_cleaned.merge(image_predictions[['tweet_id','p
        print("Need dog breed name prediction for :",twitter_archive_cleaned[twitter_archive_cle
```

```
Need dog breed name prediction for : 281  entries
```

```
In [6]: #2.a Name do not comply to the naming standard; there are 55 entries as "a".
        twitter_archive_cleaned.name.value_counts()[:11]
```

```
Out[6]: None       745
        a           55
        Charlie     12
        Lucy        11
        Cooper      11
        Oliver      11
        Penny       10
        Tucker      10
        Lola        10
        Bo           9
        Winston      9
        Name: name, dtype: int64
```

```
In [7]: #2.b there are ratings given on reply text message. we are trying find a trend in the ra
        # discussion for the tweet. The reason being, one we do not have access to all reply mes
        #tweet analysis and not replies :)
        print("Rating discussion on reply or retweets : \n", list(twitter_archive_cleaned[twitte
        print("5 High ratings are : ", sorted(list(twitter_archive_cleaned.rating_numerator))[-5
        print("5 low  ratings are : ", sorted(list(twitter_archive_cleaned.rating_numerator))[:5
```

```
Rating discussion on reply or retweets :
 ['@RealKentMurphy 14/10 confirmed', '@ComplicitOwl @ShopWeRateDogs &gt;10/10 is reserved for do
5 High ratings are :  [420, 420, 666, 960, 1776]
5 low  ratings are :  [0, 0, 1, 1, 1]
```

```
In [8]: #2.c Erraneous datatype - timestamp, retweeted_status_timestamp (to_datetime), tweet id,
        twitter_archive_cleaned.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
```

```
Data columns (total 18 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
p1                            2075 non-null object
dtypes: float64(4), int64(3), object(11)
memory usage: 349.7+ KB
```

```
In [9]: #3.a p2_* & p3_* are not required as we have p1 > p2 > p3
        if (image_predictions[image_predictions.p1_conf < image_predictions.p2_conf].empty):
            print("p1_conf is always greater than p2_conf")
        if (image_predictions[image_predictions.p2_conf < image_predictions.p3_conf].empty):
            print("p2_conf is always greater than p3_conf")
```

```
p1_conf is always greater than p2_conf
p2_conf is always greater than p3_conf
```

```
In [10]: #let us load the kaggle dog breed name list (120 names) : https://www.kaggle.com/c/dog-
         dog_breed_name_list = []
         with open('dog_breed_names.txt','r') as fp:
             for line in fp:
                 dog_breed_name_list.append(line[:-1])

         print ("Total number of breed names taken from Kaggle :", len(dog_breed_name_list))
         dog_breed_name_list = dog_breed_name_list + list(image_predictions[image_predictions.p1

         dog_breed_name_list = list(set([dog.lower() for dog in dog_breed_name_list]))
         print ("Total number of breed names along with in hand dataset and kaggle : ", len(dog_
         dog_breed_name_list[5:10]
```

```
Total number of breed names taken from Kaggle : 120
Total number of breed names along with in hand dataset and kaggle :  122
```

```
Out[10]: ['dhole', 'borzoi', 'irish_water_spaniel', 'eskimo_dog', 'japanese_spaniel']
```

```
In [11]: #3.b Dog breed names also has other than dog breed names. p1_dog indicates that whether
         twitter_archive_cleaned[~twitter_archive_cleaned.p1.str.lower().isin(dog_breed_name_lis
```

```
Out[11]: seat_belt      22
         web_site       19
         teddy          18
         tennis_ball     9
         doormat         8
         tub             7
         swing           7
         bath_towel      7
         hamster         7
         Siamese_cat     7
         Name: p1, dtype: int64
```

```
In [12]: #3.c rating_numerator is not float type and hence 13.5, 11.27, 9.5 & 9.75 rating are mi
         list(twitter_archive_cleaned[twitter_archive_cleaned.text.str.contains("[0-9]*[.][0-9]*
```

```
Out[12]: ['This is Bella. She hopes her smile made you smile. If not, she is also offering you h
          "RT @dog_rates: This is Logan, the Chow who lived. He solemnly swears he's up to lots
          "This is Logan, the Chow who lived. He solemnly swears he's up to lots of good. H*ckin
          "This is Sophie. She's a Jubilant Bush Pupper. Super h*ckin rare. Appears at random ju
          "This is Finn. He's very nervous for the game. Has a lot of money riding on it.10/10 w
          'RT @dog_rates: This... is a Tyrannosaurus rex. We only rate dogs. Please only send in
          'What jokester sent in a pic without a dog in it? This is not @rock_rates. This is @do
          'Again w the sharks guys. This week is about dogs ACTING or DRESSING like sharks. NOT
          "Guys pls stop sending actual sharks. It's too dangerous for me and the people taking
          'Guys... I said DOGS with "shark qualities" or "costumes." Not actual sharks. This did
          'This is a carrot. We only rate dogs. Please only send in dogs. You all really should
          "This is an Iraqi Speed Kangaroo. It is not a dog. Please only send in dogs. I'm very
          'This is getting incredibly frustrating. This is a Mexican Golden Beaver. We only rate
          'This... is a Tyrannosaurus rex. We only rate dogs. Please only send in dogs. Thank yo
          '"Don\'t talk to me or my son ever again" ...10/10 for both https://t.co/s96OYXZIfK',
          "Right after you graduate vs when you remember you're on your own now and can barely w
          '"Ello this is dog how may I assist" ...10/10 https://t.co/jeAENpjH7L',
          '*lets out a tiny whimper and then collapses* ...12/10 https://t.co/BNdVZEHRow',
          "When you're just relaxin and having a swell time but then remember you have to fill o
          "This is Layla. She's giving you a standing ovation.13/10 just magnificent (vid by @CS
          '"Yes hi could I get a number 4 with no pickles" ...12/10 https://t.co/kQPVxqA3gq',
          "I know it's tempting, but please stop sending in pics of Donald Trump. Thank you ...9
          'Please stop sending in saber-toothed tigers. This is getting ridiculous. We only rate
          'For the last time, WE. DO. NOT. RATE. BULBASAUR. We only rate dogs. Please only send
          '"FOR THE LAST TIME I DON\'T WANNA PLAY TWISTER ALL THE SPOTS ARE GREY DAMN IT CINDY"
          "I've been told there's a slight possibility he's checking his mirror. We'll bump to 9
          'Here we have uncovered an entire battalion of holiday puppers. Average of 11.26/10 ht
```

```
In [13]: #4.a source variable do not have the variables stored along with html tag; like other v
         twitter_archive_cleaned.source.value_counts()
```

```
Out[13]: <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
         <a href="http://vine.co" rel="nofollow">Vine - Make a Scene</a>
         <a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>
         <a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>
         Name: source, dtype: int64
```

```
In [14]: #4.b Dog breed names are not consistent with respect to usage of capital letters.
         temp_df_name = twitter_archive_cleaned[~twitter_archive_cleaned.p1.isnull()]
         print ("Number of Dog name having one or more Capital letters : ",len(list(temp_df_name
         print ("Number of Dog name in Lower Case : ",len(list(temp_df_name[temp_df_name.p1.str.
```

```
Number of Dog name having one or more Capital letters :  940
Number of Dog name in Lower Case :  1135
```

3 Clean the Data

Let us first build couple of procedures which would help us later on in our cleaning job

- *get_text_from_url* : provided a list of url, it identifies the twitter url and gets the full text from the tweet.
- *get_rating_from_text* : provided a tweet text message, it identifies rating given in the form of xx.yy/10. Here, xx and yy are numbers.
- *get_dog_stage_from_text* : Given a tweet text message, it tries to identify 'doggo', 'pupper', 'puppo', 'blep', 'snoot' & 'floof' stages as defined by weratedogs.
- *get_name_from_text* : Given a tweet text message, it tries to identify a name based on common patterns in writing tweets and also leverages the nltk.pos_tag to identify Nouns.
- *get_dog_breed_from_local_image* : tries to identify a dog breed name from an image stored locally as 'local-image.jpg'. This procedure has to be called from get_dog_breed_name_from_url.
- *get_dog_breed_name_from_url*: given a list of url, identifies the images in the url. It then iterates over all image to identify a dog image. The dog image should have at least 70% confidence when it was predicted or else the image is ignored and subsequent image is used for prediction.
- *get_dict_from_url* : given the list of urls; identifies the retweet, favorite and reply count for a tweet.

```
In [15]: #let the Data gathering begin because we are the good data scientist , brent :)
         #let us define the resuable components

         def get_text_from_url(url) :

             tweet_text = ""
             twitter_url = ""
```

```
    try:
        #print(type(url))
        url_list = url.split(',')

        for url_i in url_list:
            if "twitter.com" in url_i:
                #print("Setting url: ",url_i)
                twitter_url = url_i
                break

        #if there are shrinked url or twitter url does not exists; use the available on
        if twitter_url == "":
            twitter_url = url_list[0]

        print("twitter_url :", twitter_url)
        response = requests.get(twitter_url)
        #print("response message got")
        soup = BeautifulSoup(response.content, 'lxml')
        #print("soup cooked")

        try:
            tweet_text = \
            soup.find('p',
                      class_='TweetTextSize TweetTextSize--jumbo js-tweet-text tweet-te
            #print("loaded the tweet text")
        except:
            #print("tweet not found....ignored !!!")
            pass

    except:
        #no point in going further as we dont have breed of the dog.
        #tweet_text = ""
        #above 2 are risky because we can have dog breed name of dog or dog name which
        #handy in our analysis.
        print("Error in url : ", url)
        pass
    print(tweet_text)
    return tweet_text
#test the function: get_text_from_url
print(get_text_from_url('https://t.co/Vm7oRwuRK9'))
```

```
twitter_url : https://t.co/Vm7oRwuRK9
This is Badger. Today he embarks on his first plane ride. Already checked a bag containing a sin
This is Badger. Today he embarks on his first plane ride. Already checked a bag containing a sin
```

```
In [16]: #let us get the rating out of text
         def get_rating_from_text(text):
             #to include his first rating of 8/10
             try:
                 rating_str = re.search('[0-9][0-9[0-9]?[0-9]?.?[0-9]?[0-9]?/[0-9][0-9][0-9]',te
                 numerator_denom = rating_str.split("/")
                 return [float(numerator_denom[0]), float(numerator_denom[0])]
             except:
                 return [-1.0,-1.0]

         #test the function: get_rating_from_text
         print(get_rating_from_text(get_text_from_url('https://t.co/Vm7oRwuRK9')))
```

```
twitter_url : https://t.co/Vm7oRwuRK9
This is Badger. Today he embarks on his first plane ride. Already checked a bag containing a sin
[-1.0, -1.0]
```

```
In [17]: #let us search the dog stage from the text
         def get_dog_stage_from_text(text):
             dog_stages = ['doggo', 'pupper', 'puppo', 'blep', 'snoot', 'floof']

             stage = ""
             for word in nltk.word_tokenize(text):
                 for s in dog_stages:
                     #print(s,word)
                     if s in word.lower():
                         stage = s
                         break
                 if len(stage) > 0:
                     break

             #print("stage:",stage)
             return stage


         #test the function: get_dog_stage_from_text
         print(get_dog_stage_from_text(get_text_from_url('https://t.co/Vm7oRwuRK9')))
         print(get_dog_stage_from_text(get_text_from_url('https://twitter.com/dog_rates/status/1
         print(get_dog_stage_from_text(get_text_from_url('https://t.co/r28jFx9uyF')))
```

```
twitter_url : https://t.co/Vm7oRwuRK9
This is Badger. Today he embarks on his first plane ride. Already checked a bag containing a sin

twitter_url : https://twitter.com/dog_rates/status/1014563241572397061
This is Misty. She accidentally dropped her ball in the pool. Not a problem. 14/10 for the deep
doggo
twitter_url : https://t.co/r28jFx9uyF
This is Walter. He won't start hydrotherapy without his favorite floatie. 14/10 keep it pup Walt
```

```
In [18]: # Identify a name in the tweet text
         def get_name_from_text(text):
             dog_stages = ['doggo', 'pupper', 'puppo', 'blep', 'snoot', 'floof']
             nouns_tag = ['NN','NNS','NNP','NNPS']
             dog_stopwords = stopwords.words('english')
             custom_stopwards = ['please','thank', 'rt','pupdate','everyone','prefers','V-DAY','
                                 'people'
                                 ]
             for w in text.split(" "):
                 if "@" in w or "#" in w:
                     custom_stopwards.append(w[1:])
             custom_stopwards = [x.lower() for x in custom_stopwards]

             dog_stopwords = dog_stopwords + custom_stopwards
             name = ""

             #now check if the sentense starts with This is X. or Meet X. remove RT
             first_sent = nltk.sent_tokenize(text)[0]
             #print("first sentence :", first_sent)
             first_sent1 = first_sent.split(":")
             if len(first_sent1) > 1:
                 first_sent = first_sent1[1][1:]
             else:
                 first_sent = first_sent1[0]
             #print("first sentence now:", first_sent,"\npost split: ", first_sent.split(" "))
             if len(first_sent.split(" ")) == 3:
                 if first_sent.split(" ")[0] == "This":
                     if first_sent.split(" ")[1] == "is":
                         name = first_sent.split(" ")[2][:-1]
             elif len(text.split("Meet ")) > 1:
                 name = text.split("Meet ")[1].split(".")[0]
             elif len(text.split("Say hello to ")) > 1:
                 name = text.split("Say hello to ")[1].split(".")[0]
             else:
                 if len(text.split(":")) > 2 :
                     text = text.split(":")[1]
                 elif len(text.split("https:")) > 1 :
                     text = text.split(":")[0]

                 tagged = nltk.pos_tag(nltk.word_tokenize(text))
                 #print(len(tagged))
                 tagged_lower = nltk.pos_tag(nltk.word_tokenize(text.lower()))
                 #print(len(tagged_lower))
                 if len(tagged) == len(tagged_lower):
                     for iterate in range(len(tagged)):
                         if tagged_lower[iterate][0] not in dog_stopwords:
                             #print(tagged[iterate],tagged_lower[iterate])
                             #if tag[1] == 'NNP' and tag[0].lower() not in stopwords.words('engl
                    if tagged[iterate][1] in nouns_tag and \
                    tagged_lower[iterate][1] in nouns_tag and \
                    tagged[iterate][0] != tagged_lower[iterate][0]:
                        name = tagged[iterate][0]
                        break
        else:
            for iterate in range(len(tagged)):
                if tagged[iterate][0].lower() not in dog_stopwords:
                    if tagged[iterate][1] in nouns_tag:
                        name = tagged[iterate][0]
                        break

    print("Tweet Message : ", text, "\nDog Name : ", name)

    return name

#test get_name_from_text
sample_text = '''RT @dog_rates: This is a Emmy. She was adopted today. Massive round of
'''
print(get_name_from_text(sample_text))
sample_text = '''"This is an odd dog. Hard on the outside but loving on the inside. Pet
print(get_name_from_text(sample_text))
sample_text = '''This is a purebred Piers Morgan. pupper Loves to Netflix and chill. Al
print(get_name_from_text(sample_text))
sample_text = '''Please only send dogs. We don't rate mechanics, no matter how h*ckin g
print(get_name_from_text(sample_text))
sample_text = '''I couldn't make it to the #WKCDogShow BUT I have people there on the g
print(get_name_from_text(sample_text))
```

```
Tweet Message :   This is a Emmy. She was adopted today. Massive round of pupplause for Emmy and
Dog Name :  Emmy
Emmy
Tweet Message :  "This is an odd dog. Hard on the outside but loving on the inside. Petting stil
Dog Name :

Tweet Message :  This is a purebred Piers Morgan. pupper Loves to Netflix and chill. Always look
Dog Name :  Piers
Piers
Tweet Message :  Please only send dogs. We don't rate mechanics, no matter how h*ckin good. Than
Dog Name :

Tweet Message :  I couldn't make it to the #WKCDogShow BUT I have people there on the ground rel
Dog Name :
```

```
In [19]: #let us get the last item: dog name
         def get_dog_breed_from_local_image():
             img_path = r'local-image.jpg'
             if os.path.exists(img_path):
                 img = image.load_img(img_path, target_size=(224, 224))
                 x = image.img_to_array(img)
                 x = np.expand_dims(x, axis=0)
                 x = preprocess_input(x)
                 preds = model.predict(x)
                 # decode the results into a list of tuples (class, description, probability)
                 # (one such list for each sample in the batch)
                 #print("Prediction Done !!",decode_predictions(preds, top=3)[0])
                 if decode_predictions(preds, top=3)[0][0][2] > 0.7 and \
                    decode_predictions(preds, top=3)[0][0][1].lower() in dog_breed_name_list:
                         return decode_predictions(preds, top=3)[0][0][1]
             return ""

         #test the function: get_dog_breed_from_local_image
         get_text_from_url('https://t.co/Vm7oRwuRK9')
         print(get_dog_breed_from_local_image())
         get_text_from_url('https://twitter.com/dog_rates/status/1014563241572397061')
         print(get_dog_breed_from_local_image())
```

```
twitter_url : https://t.co/Vm7oRwuRK9
This is Badger. Today he embarks on his first plane ride. Already checked a bag containing a sin
Downloading data from https://s3.amazonaws.com/deep-learning-models/image-models/imagenet_class_
40960/35363 [=================================] - 0s 2us/step
Labrador_retriever
twitter_url : https://twitter.com/dog_rates/status/1014563241572397061
This is Misty. She accidentally dropped her ball in the pool. Not a problem. 14/10 for the deep
Labrador_retriever
```

```
In [20]: def get_images_from_url(url):
             images_url_list = []
             try:
                 for u in url.split(","):
                     response = requests.get(u)
                     soup = BeautifulSoup(response.content, 'lxml')
                     try:
                         image_tag_list = \
                             soup.find('div', class_="AdaptiveMedia-container").find_all('img')
                         #image_loc = []
                         for igl in image_tag_list:
                             images_url_list.append(igl['src'])
                         #print(image_loc)
                     except:
                         pass
             except:
                 print("Url has issue : ", url)
             return images_url_list

         get_images_from_url('https://twitter.com/dog_rates/status/879862464715927552/photo/1')

Out[20]: ['https://pbs.twimg.com/media/DDXmPreXUAA3QGJ.jpg',
          'https://pbs.twimg.com/media/DDXmPrrXoAI42eo.jpg',
          'https://pbs.twimg.com/media/DDXmPrbWAAEKMvy.jpg']

In [21]: #define a function to know the dog breed name from the url
         def get_dog_breed_name_from_url(url):
             images_list = get_images_from_url(url)
             #print(images_list)
             for image in images_list:
                 #let us also foresee if breed recognition through image is feasible
                 if os.path.exists("local-image.jpg"):
                     os.remove("local-image.jpg")

                 try:
                     urllib.request.urlretrieve(image, "local-image.jpg")

                     dog_bread_name = get_dog_breed_from_local_image()
                     #print(dog_bread_name)
                     if  dog_bread_name != "":
                         return dog_bread_name
                 except:
                     pass

             return ""

         #test get_dog_breed_name_from_url
         get_dog_breed_name_from_url('https://t.co/Vm7oRwuRK9')

Out[21]: 'Labrador_retriever'

In [22]: #let us load the reply count, retweet count and favorite count
         def get_dict_from_url(url) :

             dict_out = {}
             dict_out['retweet_count'] = dict_out['favorite_count'] = dict_out['reply_count'] =

             try:
                 for url in url.split(','):
                     try:
                         response = requests.get(url)
                         #print("response message got")
                         soup = BeautifulSoup(response.content, 'lxml')
                         #print("soup cooked")
```

```
                dict_out['retweet_count'] = \
                soup.find('li',
                      class_='js-stat-count js-stat-retweets stat-count').find('a')['da

                dict_out['favorite_count'] = \
                soup.find('li',
                      class_='js-stat-count js-stat-favorites stat-count').find('a')['d

                dict_out['reply_count'] = 0
                dict_out['reply_count'] = \
                    soup.find('span',
                      class_='ProfileTweet-actionCountForPresentation').get_text()
                if dict_out['reply_count'] == '':
                    #print("Fetching reply count Again")
                    dict_out['reply_count'] = \
                        soup.find('span',
                          class_='ProfileTweet-actionCount')['data-tweet-stat-count']

                if dict_out['reply_count'] > 0 or dict_out['favorite_count'] > 0 or dic
                    break

                    #print("Fetched :",dict_out['reply_count'])
            except:
                pass

    except:
        print("Something went wrong ; returning empty dict:",url)
        return [0,0,0]

    #print(dict_out)
    return [dict_out['retweet_count'], dict_out['favorite_count'], dict_out['reply_coun

#test get_dict_from_url
print(get_dict_from_url('https://twitter.com/dog_rates/status/709901256215666688/photo/
print(get_dict_from_url('https://twitter.com/dog_rates/status/888078434458587136/photo/
print(get_dict_from_url('https://twitter.com/dog_rates/status/879862464715927552/photo/
```

```
['108', '714', '30']
['3485', '21661', '111']
['3504', '22230', '78']
```

Let us first fix the following 2 observations

- 1.a 1976 entries do not have 'Stage' info.
- Dog stages can be melted to a column and can be considered as category variable

Define: Extract the stages from the text and check if the all the stages are there in the tweets we have.

Code:

```
In [23]: twitter_archive_cleaned["stage"] = twitter_archive_cleaned.text.apply(lambda x:get_dog_
         twitter_archive_cleaned = twitter_archive_cleaned.drop('doggo', 1)
         twitter_archive_cleaned = twitter_archive_cleaned.drop('floofer', 1)
         twitter_archive_cleaned = twitter_archive_cleaned.drop('pupper', 1)
         twitter_archive_cleaned = twitter_archive_cleaned.drop('puppo', 1)
```

Test:

```
In [24]: twitter_archive_cleaned.stage.value_counts()
```

```
Out[24]:           1902
         pupper     275
         doggo      100
         floof       38
         puppo       38
         blep         3
         Name: stage, dtype: int64
```

The results are better than the initial data. Now we have less than 1976 missing entries for dog stage.

Let us now fix the dog breed name related issues

- 1.b 281 entries of dog breed names are missing.
- 3.b Dog breed names also has other than dog breed names. p1_dog indicates that whether the name identified is dog name or not.
- 4.b Dog breed names are not consistent with respect to usage of capital letters.
- Twitter archive and image prediction can be combined to store only dog breed name.

Define:

- left join twitter_archive_cleaned with image_predictions because image_predictions has less samples. Take only p1 because p1_conf > p2_conf > p3_conf
- join twitter_archive_cleaned with image_recognition_keras (which has image prediction of a dog for each tweet with a confidence more than 70%). Take only image_recognition_keras 'breed_name' column as we have rest of the column are same as that of twitter_archive_cleaned.
- remove non dog breed name from the p1
- replace p1 with breed_name value where p1 = null
- replace p1 with lower case breed names

```
In [25]: #check if image recognition by keras model does the better job
```

```
if os.path.exists('twitter_archive_cleaned_images.csv'):
    image_recognition_keras = pd.read_csv('twitter_archive_cleaned_images.csv')
else:
    twitter_archive_cleaned['breed_name'] = twitter_archive_cleaned.expanded_urls.apply
    twitter_archive_cleaned.to_csv('twitter_archive_cleaned_images.csv')
    #twitter_archive_cleaned['re_fav_reply'] = twitter_archive_cleaned.expanded_urls.ap
    #twitter_archive_cleaned.to_csv('twitter_archive_cleaned_images.csv')
    image_recognition_keras = twitter_archive_cleaned.copy()

image_recognition_keras.breed_name.value_counts()
```

```
Out[25]: golden_retriever                  102
         Labrador_retriever                 48
         Pembroke                           45
         Samoyed                            32
         pug                                32
         Chihuahua                          22
         French_bulldog                     17
         Pomeranian                         16
         malamute                           13
         chow                               12
         toy_poodle                         11
         German_shepherd                    10
         cocker_spaniel                      9
         American_Staffordshire_terrier      9
         Maltese_dog                         7
         Bernese_mountain_dog                7
         Shetland_sheepdog                   7
         Old_English_sheepdog                7
         basset                              7
         Blenheim_spaniel                    7
         dalmatian                           7
         Brittany_spaniel                    7
         Pekinese                            6
         Yorkshire_terrier                   6
         Chesapeake_Bay_retriever            6
         beagle                              6
         miniature_pinscher                  5
         Italian_greyhound                   5
         Doberman                            5
         Shih-Tzu                            5
                                          ...
         Tibetan_mastiff                     2
         Norwich_terrier                     2
         Cardigan                            2
         soft-coated_wheaten_terrier         2
         collie                              2
         Irish_terrier                       2
```

```
Afghan_hound                    2
Greater_Swiss_Mountain_dog      2
Scotch_terrier                  2
boxer                           2
flat-coated_retriever           2
Weimaraner                      2
kuvasz                          1
Sussex_spaniel                  1
giant_schnauzer                 1
Eskimo_dog                      1
kelpie                          1
Welsh_springer_spaniel          1
miniature_schnauzer             1
redbone                         1
Japanese_spaniel                1
malinois                        1
miniature_poodle                1
dingo                           1
African_hunting_dog             1
bloodhound                      1
Tibetan_terrier                 1
Ibizan_hound                    1
bluetick                        1
Lakeland_terrier                1
Name: breed_name, Length: 84, dtype: int64
```

```
In [26]: #let us now check how may dogs are there in 2 dfs
         #image_df = pd.DataFrame(image_predictions[image_predictions.p1.isin(dog_breed_name_lis
         image_df = image_predictions.copy()
         image_df['p1'] = image_df.p1.apply(lambda x: x.lower())
         image_df = image_df[image_df.p1.isin(dog_breed_name_list)]
         print("available data set has : ", image_df.shape[0], " dog breed names")
         image_df = image_df[image_df.p1_conf > 0.7].p1.value_counts().reset_index()
         print("available data set has : ", image_df.shape[0], " dog breed names with greater th

         image_df = image_df.sort_values('index').reset_index(drop=True)
         #image_df
         image_df2 = image_recognition_keras.copy()
         image_df2 = image_df2[~image_df2['breed_name'].isnull()]
         image_df2['breed_name'] = image_df2['breed_name'].apply(lambda x: x.lower())
         image_df2 = image_df2[image_df2.breed_name.isin(dog_breed_name_list)].breed_name.value_
         #image_df2 = pd.DataFrame(image_recognition_keras[image_recognition_keras.breed_name.is
         image_df2 = image_df2.sort_values('index').reset_index(drop=True)
         print("prediction algo has : ", image_df2.shape[0], " dog breed names with greater than
         #image_df2
         image_df3 = image_df.merge( image_df2, how="left")
         image_df3.columns = [['breed name', 'available count','predicted count']]
         image_df3
```

```
        #Seems like there can be improvement to custom made function; mostly related to how the
```

```
available data set has :  1543  dog breed names
available data set has :  86  dog breed names with greater than 70% confidence
prediction algo has :  84  dog breed names with greater than 70% confidence
```

```
Out[26]:                      breed name  available count  predicted count
        0                    afghan_hound                1              2.0
        1              african_hunting_dog                1              1.0
        2                        airedale                4              NaN
        3   american_staffordshire_terrier                5              9.0
        4                         basenji                3              4.0
        5                          basset                6              7.0
        6                          beagle                5              6.0
        7             bernese_mountain_dog                7              7.0
        8          black-and-tan_coonhound                1              NaN
        9                  blenheim_spaniel                6              7.0
        10                      bloodhound                2              1.0
        11                        bluetick                1              1.0
        12                   border_collie                1              4.0
        13                          borzoi                4              3.0
        14                     boston_bull                4              3.0
        15                           boxer                5              2.0
        16                          briard                1              NaN
        17                brittany_spaniel                7              7.0
        18                     bull_mastiff                4              3.0
        19                        cardigan                6              2.0
        20         chesapeake_bay_retriever                8              6.0
        21                       chihuahua               34             22.0
        22                            chow               19             12.0
        23                         clumber                1              NaN
        24                   cocker_spaniel               11              9.0
        25                          collie                3              2.0
        26           curly-coated_retriever                1              NaN
        27                       dalmatian                5              7.0
        28                  dandie_dinmont                1              NaN
        29                           dingo                1              1.0
        ..                             ...              ...              ...
        56               miniature_pinscher                8              5.0
        57                 miniature_poodle                1              1.0
        58              miniature_schnauzer                1              1.0
        59               norwegian_elkhound                3              3.0
        60              old_english_sheepdog                5              7.0
        61                        papillon                2              4.0
        62                        pekinese                5              6.0
        63                        pembroke               52             45.0
        64                      pomeranian               23             16.0
```

```
65                             pug               37             32.0
66                         redbone                1              1.0
67             rhodesian_ridgeback                1              NaN
68                      rottweiler                5              3.0
69                   saint_bernard                3              4.0
70                         samoyed               25             32.0
71                      schipperke                4              2.0
72               shetland_sheepdog                8              7.0
73                        shih-tzu                6              5.0
74                  siberian_husky                5              3.0
75     soft-coated_wheaten_terrier                4              2.0
76       staffordshire_bullterrier                3              3.0
77                 tibetan_mastiff                2              2.0
78                      toy_poodle               15             11.0
79                     toy_terrier                1              NaN
80                          vizsla                6              3.0
81                      weimaraner                2              2.0
82          welsh_springer_spaniel                1              1.0
83     west_highland_white_terrier                7              3.0
84                         whippet                3              NaN
85               yorkshire_terrier                2              6.0

[86 rows x 3 columns]
```

The above comparison indicates that the keras model performs almost same as that of available data. Let us now put the predicted values to missing data in the original dataset

```
In [27]: twitter_archive_cleaned= twitter_archive_cleaned.merge(image_recognition_keras[['tweet_

         not_null_entries_before = twitter_archive_cleaned[twitter_archive_cleaned.p1.isnull()].
         twitter_archive_cleaned['p1'] = twitter_archive_cleaned.p1.apply(lambda x: x if type(x)
         valid_breed_name_before = twitter_archive_cleaned[~twitter_archive_cleaned['p1'].isnull
         twitter_archive_cleaned['p1'] = twitter_archive_cleaned.apply( lambda row: row['p1'] if

         valid_breed_name_after = twitter_archive_cleaned[~twitter_archive_cleaned['p1'].isnull(
         not_null_entries_after = twitter_archive_cleaned[twitter_archive_cleaned.p1.isnull()].s

         per_change_null = round((not_null_entries_after - not_null_entries_before)/not_null_ent
         per_change_valid = round((valid_breed_name_after - valid_breed_name_before)/valid_breed

         twitter_archive_cleaned['p1'] = twitter_archive_cleaned['p1'].apply(lambda x : None if
         twitter_archive_cleaned = twitter_archive_cleaned.drop('breed_name', 1)
         twitter_archive_cleaned = twitter_archive_cleaned.rename(index=str, columns={"p1": "bre
```

Test:

```
In [28]: print("There is ", per_change_null, "% increment is missing values")
         print("There is ", per_change_valid, "% increment is valid values")
         tt = twitter_archive_cleaned[~twitter_archive_cleaned['breed_name'].isnull()]
         print("Breed names with upper case names : ", tt[~tt['breed_name'].str.islower()].shape
```

```
There is  1.78 % increment is missing values
There is  0.02 % increment is valid values
Breed names with upper case names :  0
```

```
In [29]: twitter_archive_cleaned
```

```
Out[29]:                 tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
         0     892420643555336193                    NaN                  NaN
         1     892177421306343426                    NaN                  NaN
         2     891815181378084864                    NaN                  NaN
         3     891689557279858688                    NaN                  NaN
         4     891327558926688256                    NaN                  NaN
         5     891087950875897856                    NaN                  NaN
         6     890971913173991426                    NaN                  NaN
         7     890729181411237888                    NaN                  NaN
         8     890609185150312448                    NaN                  NaN
         9     890240255349198849                    NaN                  NaN
         10    890006608113172480                    NaN                  NaN
         11    889880896479866881                    NaN                  NaN
         12    889665388333682689                    NaN                  NaN
         13    889638837579907072                    NaN                  NaN
         14    889531135344209921                    NaN                  NaN
         15    889278841981685760                    NaN                  NaN
         16    888917238123831296                    NaN                  NaN
         17    888804989199671297                    NaN                  NaN
         18    888554962724278272                    NaN                  NaN
         19    888202515573088257                    NaN                  NaN
         20    888078434458587136                    NaN                  NaN
         21    887705289381826560                    NaN                  NaN
         22    887517139158093824                    NaN                  NaN
         23    887473957103951883                    NaN                  NaN
         24    887343217045368832                    NaN                  NaN
         25    887101392804085760                    NaN                  NaN
         26    886983233522544640                    NaN                  NaN
         27    886736880519319552                    NaN                  NaN
         28    886680336477933568                    NaN                  NaN
         29    886366144734445568                    NaN                  NaN
         ...                  ...                    ...                  ...
         2326  666411507551481857                    NaN                  NaN
         2327  666407126856765440                    NaN                  NaN
         2328  666396247373291520                    NaN                  NaN
         2329  666373753744588802                    NaN                  NaN
         2330  666362758909284353                    NaN                  NaN
         2331  666353288456101888                    NaN                  NaN
```

```
2332  666345417576210432                    NaN                  NaN
2333  666337882303524864                    NaN                  NaN
2334  666293911632134144                    NaN                  NaN
2335  666287406224695296                    NaN                  NaN
2336  666273097616637952                    NaN                  NaN
2337  666268910803644416                    NaN                  NaN
2338  666104133288665088                    NaN                  NaN
2339  666102155909144576                    NaN                  NaN
2340  666099513787052032                    NaN                  NaN
2341  666094000022159362                    NaN                  NaN
2342  666082916733198337                    NaN                  NaN
2343  666073100786774016                    NaN                  NaN
2344  666071193221509120                    NaN                  NaN
2345  666063827256086533                    NaN                  NaN
2346  666058600524156928                    NaN                  NaN
2347  666057090499244032                    NaN                  NaN
2348  666055525042405380                    NaN                  NaN
2349  666051853826850816                    NaN                  NaN
2350  666050758794694657                    NaN                  NaN
2351  666049248165822465                    NaN                  NaN
2352  666044226329800704                    NaN                  NaN
2353  666033412701032449                    NaN                  NaN
2354  666029285002620928                    NaN                  NaN
2355  666020888022790149                    NaN                  NaN

                      timestamp  \
0     2017-08-01 16:23:56 +0000
1     2017-08-01 00:17:27 +0000
2     2017-07-31 00:18:03 +0000
3     2017-07-30 15:58:51 +0000
4     2017-07-29 16:00:24 +0000
5     2017-07-29 00:08:17 +0000
6     2017-07-28 16:27:12 +0000
7     2017-07-28 00:22:40 +0000
8     2017-07-27 16:25:51 +0000
9     2017-07-26 15:59:51 +0000
10    2017-07-26 00:31:25 +0000
11    2017-07-25 16:11:53 +0000
12    2017-07-25 01:55:32 +0000
13    2017-07-25 00:10:02 +0000
14    2017-07-24 17:02:04 +0000
15    2017-07-24 00:19:32 +0000
16    2017-07-23 00:22:39 +0000
17    2017-07-22 16:56:37 +0000
18    2017-07-22 00:23:06 +0000
19    2017-07-21 01:02:36 +0000
20    2017-07-20 16:49:33 +0000
21    2017-07-19 16:06:48 +0000
```

```
22    2017-07-19 03:39:09 +0000
23    2017-07-19 00:47:34 +0000
24    2017-07-18 16:08:03 +0000
25    2017-07-18 00:07:08 +0000
26    2017-07-17 16:17:36 +0000
27    2017-07-16 23:58:41 +0000
28    2017-07-16 20:14:00 +0000
29    2017-07-15 23:25:31 +0000
...                         ...
2326  2015-11-17 00:24:19 +0000
2327  2015-11-17 00:06:54 +0000
2328  2015-11-16 23:23:41 +0000
2329  2015-11-16 21:54:18 +0000
2330  2015-11-16 21:10:36 +0000
2331  2015-11-16 20:32:58 +0000
2332  2015-11-16 20:01:42 +0000
2333  2015-11-16 19:31:45 +0000
2334  2015-11-16 16:37:02 +0000
2335  2015-11-16 16:11:11 +0000
2336  2015-11-16 15:14:19 +0000
2337  2015-11-16 14:57:41 +0000
2338  2015-11-16 04:02:55 +0000
2339  2015-11-16 03:55:04 +0000
2340  2015-11-16 03:44:34 +0000
2341  2015-11-16 03:22:39 +0000
2342  2015-11-16 02:38:37 +0000
2343  2015-11-16 01:59:36 +0000
2344  2015-11-16 01:52:02 +0000
2345  2015-11-16 01:22:45 +0000
2346  2015-11-16 01:01:59 +0000
2347  2015-11-16 00:55:59 +0000
2348  2015-11-16 00:49:46 +0000
2349  2015-11-16 00:35:11 +0000
2350  2015-11-16 00:30:50 +0000
2351  2015-11-16 00:24:50 +0000
2352  2015-11-16 00:04:52 +0000
2353  2015-11-15 23:21:54 +0000
2354  2015-11-15 23:05:30 +0000
2355  2015-11-15 22:32:08 +0000

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

```
7     <a href="http://twitter.com/download/iphone" r...
8     <a href="http://twitter.com/download/iphone" r...
9     <a href="http://twitter.com/download/iphone" r...
10    <a href="http://twitter.com/download/iphone" r...
11    <a href="http://twitter.com/download/iphone" r...
12    <a href="http://twitter.com/download/iphone" r...
13    <a href="http://twitter.com/download/iphone" r...
14    <a href="http://twitter.com/download/iphone" r...
15    <a href="http://twitter.com/download/iphone" r...
16    <a href="http://twitter.com/download/iphone" r...
17    <a href="http://twitter.com/download/iphone" r...
18    <a href="http://twitter.com/download/iphone" r...
19    <a href="http://twitter.com/download/iphone" r...
20    <a href="http://twitter.com/download/iphone" r...
21    <a href="http://twitter.com/download/iphone" r...
22    <a href="http://twitter.com/download/iphone" r...
23    <a href="http://twitter.com/download/iphone" r...
24    <a href="http://twitter.com/download/iphone" r...
25    <a href="http://twitter.com/download/iphone" r...
26    <a href="http://twitter.com/download/iphone" r...
27    <a href="http://twitter.com/download/iphone" r...
28    <a href="http://twitter.com/download/iphone" r...
29    <a href="http://twitter.com/download/iphone" r...
...                                                ...
2326  <a href="http://twitter.com/download/iphone" r...
2327  <a href="http://twitter.com/download/iphone" r...
2328  <a href="http://twitter.com/download/iphone" r...
2329  <a href="http://twitter.com/download/iphone" r...
2330  <a href="http://twitter.com/download/iphone" r...
2331  <a href="http://twitter.com/download/iphone" r...
2332  <a href="http://twitter.com/download/iphone" r...
2333  <a href="http://twitter.com/download/iphone" r...
2334  <a href="http://twitter.com/download/iphone" r...
2335  <a href="http://twitter.com/download/iphone" r...
2336  <a href="http://twitter.com/download/iphone" r...
2337  <a href="http://twitter.com/download/iphone" r...
2338  <a href="http://twitter.com/download/iphone" r...
2339  <a href="http://twitter.com/download/iphone" r...
2340  <a href="http://twitter.com/download/iphone" r...
2341  <a href="http://twitter.com/download/iphone" r...
2342  <a href="http://twitter.com/download/iphone" r...
2343  <a href="http://twitter.com/download/iphone" r...
2344  <a href="http://twitter.com/download/iphone" r...
2345  <a href="http://twitter.com/download/iphone" r...
2346  <a href="http://twitter.com/download/iphone" r...
2347  <a href="http://twitter.com/download/iphone" r...
2348  <a href="http://twitter.com/download/iphone" r...
2349  <a href="http://twitter.com/download/iphone" r...
```

```
2350  <a href="http://twitter.com/download/iphone" r...
2351  <a href="http://twitter.com/download/iphone" r...
2352  <a href="http://twitter.com/download/iphone" r...
2353  <a href="http://twitter.com/download/iphone" r...
2354  <a href="http://twitter.com/download/iphone" r...
2355  <a href="http://twitter.com/download/iphone" r...

                                                   text  retweeted_status_id  \
0     This is Phineas. He's a mystical boy. Only eve...                  NaN
1     This is Tilly. She's just checking pup on you...                  NaN
2     This is Archie. He is a rare Norwegian Pouncin...                  NaN
3     This is Darla. She commenced a snooze mid meal...                  NaN
4     This is Franklin. He would like you to stop ca...                  NaN
5     Here we have a majestic great white breaching ...                  NaN
6     Meet Jax. He enjoys ice cream so much he gets ...                  NaN
7     When you watch your owner call another dog a g...                  NaN
8     This is Zoey. She doesn't want to be one of th...                  NaN
9     This is Cassie. She is a college pup. Studying...                  NaN
10    This is Koda. He is a South Australian decksha...                  NaN
11    This is Bruno. He is a service shark. Only get...                  NaN
12    Here's a puppo that seems to be on the fence a...                  NaN
13    This is Ted. He does his best. Sometimes that'...                  NaN
14    This is Stuart. He's sporting his favorite fan...                  NaN
15    This is Oliver. You're witnessing one of his m...                  NaN
16    This is Jim. He found a fren. Taught him how t...                  NaN
17    This is Zeke. He has a new stick. Very proud o...                  NaN
18    This is Ralphus. He's powering up. Attempting ...                  NaN
19    RT @dog_rates: This is Canela. She attempted s...         8.874740e+17
20    This is Gerald. He was just told he didn't get...                  NaN
21    This is Jeffrey. He has a monopoly on the pool...                  NaN
22    I've yet to rate a Venezuelan Hover Wiener. Th...                  NaN
23    This is Canela. She attempted some fancy porch...                  NaN
24    You may not have known you needed to see this ...                  NaN
25    This... is a Jubilant Antarctic House Bear. We...                  NaN
26    This is Maya. She's very shy. Rarely leaves he...                  NaN
27    This is Mingus. He's a wonderful father to his...                  NaN
28    This is Derek. He's late for a dog meeting. 13...                  NaN
29    This is Roscoe. Another pupper fallen victim t...                  NaN
...                                                 ...                  ...
2326  This is quite the dog. Gets really excited whe...                  NaN
2327  This is a southern Vesuvius bumblegruff. Can d...                  NaN
2328  Oh goodness. A super rare northeast Qdoba kang...                  NaN
2329  Those are sunglasses and a jean jacket. 11/10 ...                  NaN
2330  Unique dog here. Very small. Lives in containe...                  NaN
2331  Here we have a mixed Asiago from the Galápagos...                  NaN
2332  Look at this jokester thinking seat belt laws ...                  NaN
2333  This is an extremely rare horned Parthenon. No...                  NaN
2334  This is a funny dog. Weird toes. Won't come do...                  NaN
```

```
2335  This is an Albanian 3 1/2 legged  Episcopalian...                  NaN
2336     Can take selfies 11/10 https://t.co/ws2AMaNwPW                  NaN
2337  Very concerned about fellow dog trapped in com...                  NaN
2338  Not familiar with this breed. No tail (weird)...                  NaN
2339  Oh my. Here you are seeing an Adobe Setter giv...                  NaN
2340  Can stand on stump for what seems like a while...                  NaN
2341  This appears to be a Mongolian Presbyterian mi...                  NaN
2342  Here we have a well-established sunblockerspan...                  NaN
2343  Let's hope this flight isn't Malaysian (lol). ...                  NaN
2344  Here we have a northern speckled Rhododendron...                  NaN
2345  This is the happiest dog you will ever see. Ve...                  NaN
2346  Here is the Rand Paul of retrievers folks! He'...                  NaN
2347  My oh my. This is a rare blond Canadian terrie...                  NaN
2348  Here is a Siberian heavily armored polar bear ...                  NaN
2349  This is an odd dog. Hard on the outside but lo...                  NaN
2350  This is a truly beautiful English Wilson Staff...                  NaN
2351  Here we have a 1949 1st generation vulpix. Enj...                  NaN
2352  This is a purebred Piers Morgan. Loves to Netf...                  NaN
2353  Here is a very happy pup. Big fan of well-main...                  NaN
2354  This is a western brown Mitsubishi terrier. Up...                  NaN
2355  Here we have a Japanese Irish Setter. Lost eye...                  NaN

      retweeted_status_user_id retweeted_status_timestamp  \
0                          NaN                        NaN
1                          NaN                        NaN
2                          NaN                        NaN
3                          NaN                        NaN
4                          NaN                        NaN
5                          NaN                        NaN
6                          NaN                        NaN
7                          NaN                        NaN
8                          NaN                        NaN
9                          NaN                        NaN
10                         NaN                        NaN
11                         NaN                        NaN
12                         NaN                        NaN
13                         NaN                        NaN
14                         NaN                        NaN
15                         NaN                        NaN
16                         NaN                        NaN
17                         NaN                        NaN
18                         NaN                        NaN
19                4.196984e+09  2017-07-19 00:47:34 +0000
20                         NaN                        NaN
21                         NaN                        NaN
22                         NaN                        NaN
23                         NaN                        NaN
24                         NaN                        NaN
```

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

                                          expanded_urls  rating_numerator  \
0     https://twitter.com/dog_rates/status/892420643...                13
1     https://twitter.com/dog_rates/status/892177421...                13
2     https://twitter.com/dog_rates/status/891815181...                12
3     https://twitter.com/dog_rates/status/891689557...                13
4     https://twitter.com/dog_rates/status/891327558...                12
5     https://twitter.com/dog_rates/status/891087950...                13
6     https://gofundme.com/ydvmve-surgery-for-jax,ht...                13
7     https://twitter.com/dog_rates/status/890729181...                13
8     https://twitter.com/dog_rates/status/890609185...                13
9     https://twitter.com/dog_rates/status/890240255...                14
```

```
10    https://twitter.com/dog_rates/status/890006608...        13
11    https://twitter.com/dog_rates/status/889880896...        13
12    https://twitter.com/dog_rates/status/889665388...        13
13    https://twitter.com/dog_rates/status/889638837...        12
14    https://twitter.com/dog_rates/status/889531135...        13
15    https://twitter.com/dog_rates/status/889278841...        13
16    https://twitter.com/dog_rates/status/888917238...        12
17    https://twitter.com/dog_rates/status/888804989...        13
18    https://twitter.com/dog_rates/status/888554962...        13
19    https://twitter.com/dog_rates/status/887473957...        13
20    https://twitter.com/dog_rates/status/888078434...        12
21    https://twitter.com/dog_rates/status/887705289...        13
22    https://twitter.com/dog_rates/status/887517139...        14
23    https://twitter.com/dog_rates/status/887473957...        13
24    https://twitter.com/dog_rates/status/887343217...        13
25    https://twitter.com/dog_rates/status/887101392...        12
26    https://twitter.com/dog_rates/status/886983233...        13
27    https://www.gofundme.com/mingusneedsus,https:/...        13
28    https://twitter.com/dog_rates/status/886680336...        13
29    https://twitter.com/dog_rates/status/886366144...        12
...                                                 ...       ...
2326  https://twitter.com/dog_rates/status/666411507...         2
2327  https://twitter.com/dog_rates/status/666407126...         7
2328  https://twitter.com/dog_rates/status/666396247...         9
2329  https://twitter.com/dog_rates/status/666373753...        11
2330  https://twitter.com/dog_rates/status/666362758...         6
2331  https://twitter.com/dog_rates/status/666353288...         8
2332  https://twitter.com/dog_rates/status/666345417...        10
2333  https://twitter.com/dog_rates/status/666337882...         9
2334  https://twitter.com/dog_rates/status/666293911...         3
2335  https://twitter.com/dog_rates/status/666287406...         1
2336  https://twitter.com/dog_rates/status/666273097...        11
2337  https://twitter.com/dog_rates/status/666268910...        10
2338  https://twitter.com/dog_rates/status/666104133...         1
2339  https://twitter.com/dog_rates/status/666102155...        11
2340  https://twitter.com/dog_rates/status/666099513...         8
2341  https://twitter.com/dog_rates/status/666094000...         9
2342  https://twitter.com/dog_rates/status/666082916...         6
2343  https://twitter.com/dog_rates/status/666073100...        10
2344  https://twitter.com/dog_rates/status/666071193...         9
2345  https://twitter.com/dog_rates/status/666063827...        10
2346  https://twitter.com/dog_rates/status/666058600...         8
2347  https://twitter.com/dog_rates/status/666057090...         9
2348  https://twitter.com/dog_rates/status/666055525...        10
2349  https://twitter.com/dog_rates/status/666051853...         2
2350  https://twitter.com/dog_rates/status/666050758...        10
2351  https://twitter.com/dog_rates/status/666049248...         5
2352  https://twitter.com/dog_rates/status/666044226...         6
```

```
2353  https://twitter.com/dog_rates/status/666033412...                9
2354  https://twitter.com/dog_rates/status/666029285...                7
2355  https://twitter.com/dog_rates/status/666020888...                8

      rating_denominator      name                   breed_name   stage
0                     10   Phineas                         None
1                     10     Tilly                    chihuahua
2                     10    Archie                    chihuahua
3                     10     Darla                         None
4                     10  Franklin                       basset
5                     10      None     chesapeake_bay_retriever
6                     10       Jax                  appenzeller
7                     10      None                   pomeranian
8                     10      Zoey                irish_terrier
9                     10    Cassie                     pembroke   doggo
10                    10      Koda                      samoyed
11                    10     Bruno               french_bulldog
12                    10      None                     pembroke   puppo
13                    10       Ted               french_bulldog
14                    10    Stuart             golden_retriever   puppo
15                    10    Oliver                      whippet
16                    10       Jim             golden_retriever
17                    10      Zeke             golden_retriever
18                    10   Ralphus               siberian_husky
19                    10    Canela                     pembroke
20                    10    Gerald               french_bulldog
21                    10   Jeffrey                       basset
22                    10      such                         None
23                    10    Canela                     pembroke
24                    10      None             mexican_hairless
25                    10      None                      samoyed   floof
26                    10      Maya                    chihuahua
27                    10    Mingus                       kuvasz
28                    10     Derek                         None
29                    10    Roscoe               french_bulldog  pupper
...                  ...       ...                          ...     ...
2326                  10     quite                         None
2327                  10         a      black-and-tan_coonhound
2328                  10      None                    chihuahua
2329                  10      None  soft-coated_wheaten_terrier
2330                  10      None                         None
2331                  10      None                     malamute
2332                  10      None             golden_retriever
2333                  10        an                         None
2334                  10         a                         None
2335                   2        an                  maltese_dog
2336                  10      None            italian_greyhound
2337                  10      None                         None
```

```
2338                  10      None                         None
2339                  10      None               english_setter
2340                  10      None                        lhasa
2341                  10      None                   bloodhound
2342                  10      None                          pug
2343                  10      None                 walker_hound
2344                  10      None                gordon_setter
2345                  10       the             golden_retriever
2346                  10       the             miniature_poodle
2347                  10         a                         None
2348                  10         a                         chow
2349                  10        an                         None
2350                  10         a         bernese_mountain_dog
2351                  10      None           miniature_pinscher
2352                  10         a          rhodesian_ridgeback
2353                  10         a              german_shepherd
2354                  10         a                      redbone
2355                  10      None       welsh_springer_spaniel

[2356 rows x 15 columns]
```

```
In [30]: twitter_archive_cleaned.breed_name.value_counts()
```

```
Out[30]: golden_retriever                  159
         labrador_retriever                101
         pembroke                           91
         chihuahua                          83
         pug                                57
         samoyed                            46
         chow                               45
         toy_poodle                         40
         pomeranian                         39
         cocker_spaniel                     31
         malamute                           30
         french_bulldog                     29
         chesapeake_bay_retriever           24
         miniature_pinscher                 23
         siberian_husky                     20
         german_shepherd                    20
         staffordshire_bullterrier          20
         cardigan                           19
         eskimo_dog                         19
         beagle                             18
         maltese_dog                        18
         shetland_sheepdog                  18
         lakeland_terrier                   17
         rottweiler                         17
         italian_greyhound                  17
```

```
shih-tzu                           17
kuvasz                             16
american_staffordshire_terrier     15
west_highland_white_terrier        14
great_pyrenees                     14
                                 ...
miniature_schnauzer                 4
keeshond                            4
weimaraner                          4
komondor                            4
ibizan_hound                        3
brabancon_griffon                   3
curly-coated_retriever              3
briard                              3
scottish_deerhound                  3
welsh_springer_spaniel              3
scotch_terrier                      3
greater_swiss_mountain_dog          3
irish_water_spaniel                 3
cairn                               3
giant_schnauzer                     3
leonberg                            3
black-and-tan_coonhound             2
wire-haired_fox_terrier             2
toy_terrier                         2
sussex_spaniel                      2
australian_terrier                  2
appenzeller                         2
entlebucher                         1
african_hunting_dog                 1
silky_terrier                       1
dhole                               1
standard_schnauzer                  1
japanese_spaniel                    1
groenendael                         1
clumber                             1
Name: breed_name, Length: 114, dtype: int64
```

2.a Name do not comply to the naming standard; there are 55 entries as “a”.

Define:

- remove stopwords from dog names
- remove any lower case name

Code:

```
In [31]: twitter_archive_cleaned.name = twitter_archive_cleaned.name.apply( lambda x : x if x.lo
         #tt = twitter_archive_cleaned[~twitter_archive_cleaned.name.isnull()]
         twitter_archive_cleaned.loc[twitter_archive_cleaned.name.str.islower(),'name'] = 'None'
```

Test:

```
In [32]: twitter_archive_cleaned.name.value_counts()
```

```
Out[32]: None         855
         Charlie       12
         Cooper        11
         Oliver        11
         Lucy          11
         Tucker        10
         Lola          10
         Penny         10
         Winston        9
         Bo             9
         Sadie          8
         Bailey         7
         Buddy          7
         Toby           7
         Daisy          7
         Scout          6
         Dave           6
         Oscar          6
         Koda           6
         Milo           6
         Bella          6
         Jack           6
         Stanley        6
         Jax            6
         Rusty          6
         Leo            6
         Oakley         5
         Chester        5
         Louis          5
         Sunny          5
                    ...
         Nigel          1
         Leonard        1
         Daniel         1
         Ralphson       1
         Sid            1
         Corey          1
         Franq          1
         Terrenth       1
         Reagan         1
```

```
Aiden          1
Smiley         1
Durg           1
Grey           1
Rooney         1
Amélie         1
Tove           1
Jeffri         1
Sojourner      1
Godi           1
Craig          1
Snicku         1
Rambo          1
Beebop         1
Zoe            1
Blakely        1
Lassie         1
Iroh           1
Ike            1
Jett           1
Mookie         1
Name: name, Length: 931, dtype: int64
```

```
In [33]: twitter_archive_cleaned.sample(10).name
```

```
Out[33]: 830        Jesse
         1902        None
         1981        Chet
         980         Lucy
         557        Sonny
         2222        None
         186         None
         932      Charlie
         26          Maya
         318         None
         Name: name, dtype: object
```

2.b Erraneous datatype -

```
* timestamp, retweeted_status_timestamp (to_datetime),
* all "_id" to int64 as integer operations are faster than string. And also makes like easy to d
```

3.c rating_numerator is not float type and hence 13.5, 11.27, 9.5 & 9.75 rating are missing in the data set.

Define: All _id variants to int; timestamp to datetime

Code:

```
In [34]: if os.path.exists("twitter_archive_cleaned_last.csv"):
             twitter_archive_cleaned = pd.read_csv("twitter_archive_cleaned_last.csv")
             twitter_archive_cleaned = twitter_archive_cleaned.drop('Unnamed: 0', 1)
         else:
             twitter_archive_cleaned.to_csv("twitter_archive_cleaned_last.csv")

         twitter_archive_cleaned.in_reply_to_status_id = pd.to_numeric(twitter_archive_cleaned.i
                                                                .fillna(0).astype(np.int64)
         twitter_archive_cleaned.in_reply_to_user_id = pd.to_numeric(twitter_archive_cleaned.in_
                                                                .fillna(0).astype(np.int64)
         twitter_archive_cleaned.in_reply_to_user_id = pd.to_numeric(twitter_archive_cleaned.in_
                                                                .fillna(0).astype(np.int64)
         twitter_archive_cleaned.retweeted_status_id = pd.to_numeric(twitter_archive_cleaned.ret
                                                                .fillna(0).astype(np.int64)
         twitter_archive_cleaned.retweeted_status_user_id = pd.to_numeric(twitter_archive_cleane
                                                                .fillna(0).astype(np.int64)

         twitter_archive_cleaned.rating_numerator = pd.to_numeric(twitter_archive_cleaned.rating
                                                                .fillna(0).astype(np.float64)

         twitter_archive_cleaned.timestamp = pd.to_datetime(twitter_archive_cleaned.timestamp)
         twitter_archive_cleaned.retweeted_status_timestamp = pd.to_datetime(twitter_archive_cle
         twitter_archive_cleaned = twitter_archive_cleaned.sort_values('timestamp')
```

Test

```
In [35]: twitter_archive_cleaned.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 2355 to 0
Data columns (total 15 columns):
tweet_id                      2356 non-null int64
in_reply_to_status_id         2356 non-null int64
in_reply_to_user_id           2356 non-null int64
timestamp                     2356 non-null datetime64[ns]
source                        2356 non-null object
text                          2356 non-null object
retweeted_status_id           2356 non-null int64
retweeted_status_user_id      2356 non-null int64
retweeted_status_timestamp    181 non-null datetime64[ns]
expanded_urls                 2297 non-null object
rating_numerator              2356 non-null float64
rating_denominator            2356 non-null int64
name                          2356 non-null object
breed_name                    1576 non-null object
stage                         454 non-null object
dtypes: datetime64[ns](2), float64(1), int64(6), object(6)
```

```
memory usage: 294.5+ KB
```

```
In [36]: twitter_archive_cleaned.head()
```

```
Out[36]:                 tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
         2355  666020888022790149                      0                    0
         2354  666029285002620928                      0                    0
         2353  666033412701032449                      0                    0
         2352  666044226329800704                      0                    0
         2351  666049248165822465                      0                    0

                        timestamp                                             source  \
         2355 2015-11-15 22:32:08  <a href="http://twitter.com/download/iphone" r...
         2354 2015-11-15 23:05:30  <a href="http://twitter.com/download/iphone" r...
         2353 2015-11-15 23:21:54  <a href="http://twitter.com/download/iphone" r...
         2352 2015-11-16 00:04:52  <a href="http://twitter.com/download/iphone" r...
         2351 2015-11-16 00:24:50  <a href="http://twitter.com/download/iphone" r...

                                                               text  retweeted_status_id  \
         2355  Here we have a Japanese Irish Setter. Lost eye...                    0
         2354  This is a western brown Mitsubishi terrier. Up...                    0
         2353  Here is a very happy pup. Big fan of well-main...                    0
         2352  This is a purebred Piers Morgan. Loves to Netf...                    0
         2351  Here we have a 1949 1st generation vulpix. Enj...                    0

              retweeted_status_user_id retweeted_status_timestamp  \
         2355                         0                        NaT
         2354                         0                        NaT
         2353                         0                        NaT
         2352                         0                        NaT
         2351                         0                        NaT

                                                      expanded_urls  rating_numerator  \
         2355  https://twitter.com/dog_rates/status/666020888...               8.0
         2354  https://twitter.com/dog_rates/status/666029285...               7.0
         2353  https://twitter.com/dog_rates/status/666033412...               9.0
         2352  https://twitter.com/dog_rates/status/666044226...               6.0
         2351  https://twitter.com/dog_rates/status/666049248...               5.0

              rating_denominator  name               breed_name stage
         2355                  10  None  welsh_springer_spaniel   NaN
         2354                  10  None                 redbone   NaN
         2353                  10  None         german_shepherd   NaN
         2352                  10  None     rhodesian_ridgeback   NaN
         2351                  10  None      miniature_pinscher   NaN
```

2.b there are ratings given on reply text message. we are trying find a trend in the root message not the subsequent discussion for the tweet. The reason being, one we do not have access to all reply messages. Two, the study is on weratedogs tweet analysis and not replies :)

Define: remove the reply messages from the data sets.

Code:

```
In [37]: if os.path.exists("twitter_archive_cleaned_last_1.csv"):
             twitter_archive_cleaned = pd.read_csv("twitter_archive_cleaned_last_1.csv")
             twitter_archive_cleaned = twitter_archive_cleaned.drop('Unnamed: 0', 1)
         else:
             twitter_archive_cleaned.to_csv("twitter_archive_cleaned_last_1.csv")

         #Let us first work on replies
         list_reply_id = list(twitter_archive_cleaned[twitter_archive_cleaned.in_reply_to_status
         print ("How many reply messages are there totally ?\nAns : ", len(list_reply_id))
         print("How many of the original messages are present in the dataset for which a reply i
               twitter_archive_cleaned[twitter_archive_cleaned.tweet_id.isin(list_reply_id)].shap

         list_orig_present = list(twitter_archive_cleaned[twitter_archive_cleaned.tweet_id.isin(
         list_retweet_id = list(twitter_archive_cleaned[twitter_archive_cleaned.retweeted_status
         twitter_archive_cleaned = twitter_archive_cleaned[~twitter_archive_cleaned.in_reply_to_


         #Let us work on retweets
         print("How many retweets are there?\nAns : ", len(list_retweet_id))

         list_orig_present = list(twitter_archive_cleaned[twitter_archive_cleaned.tweet_id.isin(
         twitter_archive_cleaned = twitter_archive_cleaned[~twitter_archive_cleaned.retweeted_st
         print("For each retweet, are the original message present in the sample if yes how many
```

```
How many reply messages are there totally ?
Ans :  78
How many of the original messages are present in the dataset for which a reply is also present ?
Ans:  33
How many retweets are there?
Ans :  181
For each retweet, are the original message present in the sample if yes how many ?
Ans:  112
```

```
In [38]: print (list(twitter_archive_cleaned[twitter_archive_cleaned.text.str.contains("[0-9]*[.
         twitter_archive_cleaned[twitter_archive_cleaned.text.str.contains("[0-9]*[.][0-9]*/")][
```

```
['Here we have uncovered an entire battalion of holiday puppers. Average of 11.26/10 https://t.c
```

Out[38]:

```
                                                   text            tweet_id
643   Here we have uncovered an entire battalion of ...  680494726643068929
728   "FOR THE LAST TIME I DON'T WANNA PLAY TWISTER ...  684594889858887680
745   For the last time, WE. DO. NOT. RATE. BULBASAU...  685532292383666176
920   Please stop sending in saber-toothed tigers. T...  697259378236399616
983   I know it's tempting, but please stop sending ...  702217446468493312
1014  "Yes hi could I get a number 4 with no pickles...  704847917308362754
1029  This is Layla. She's giving you a standing ova...  706153300320784384
1033  When you're just relaxin and having a swell ti...  706291001778950144
1089  *lets out a tiny whimper and then collapses* ...  709852847387627521
1244  "Ello this is dog how may I assist" ...10/10 h...  733482008106668032
1257  Right after you graduate vs when you remember ...  736010884653420544
1271  "Don't talk to me or my son ever again" ...10/...  738402415918125056
1278  This... is a Tyrannosaurus rex. We only rate d...  739544079319588864
1284  This is getting incredibly frustrating. This i...  740214038584557568
1330  This is an Iraqi Speed Kangaroo. It is not a d...  746369468511756288
1338  This is a carrot. We only rate dogs. Please on...  746872823977771008
1340  Guys... I said DOGS with "shark qualities" or ...  747103485104099331
1346  Guys pls stop sending actual sharks. It's too ...  747512671126323200
1347  Again w the sharks guys. This week is about do...  747594051852075008
1367  What jokester sent in a pic without a dog in i...  748977405889503236
1550  This is Finn. He's very nervous for the game. ...  772114945936949249
1592  This is Sophie. She's a Jubilant Bush Pupper. ...  778027034220126208
1660  This is Logan, the Chow who lived. He solemnly...  786709082849828864
2310  This is Bella. She hopes her smile made you sm...  883482846933004288
```

In [39]:

```python
print(list(twitter_archive_cleaned[twitter_archive_cleaned['tweet_id'] == 8834828469330
twitter_archive_cleaned.loc[twitter_archive_cleaned['tweet_id'] == 883482846933004288,'

print(list(twitter_archive_cleaned[twitter_archive_cleaned['tweet_id'] == 7867090828498
twitter_archive_cleaned.loc[twitter_archive_cleaned['tweet_id'] == 786709082849828864,'

print(list(twitter_archive_cleaned[twitter_archive_cleaned['tweet_id'] == 7780270342201
twitter_archive_cleaned.loc[twitter_archive_cleaned['tweet_id'] == 778027034220126208,'

print(list(twitter_archive_cleaned[twitter_archive_cleaned['tweet_id'] == 6804947266430
twitter_archive_cleaned.loc[twitter_archive_cleaned['tweet_id'] == 680494726643068929,'

print("Correcting rating after manually going through tweet text : ", list(twitter_arch
twitter_archive_cleaned.loc[twitter_archive_cleaned.rating_numerator == 17, 'rating_num

print("Correcting rating after manually going through tweet text : ", list(twitter_arch
twitter_archive_cleaned.loc[twitter_archive_cleaned.tweet_id == 835246439529840640, 'ra
twitter_archive_cleaned.loc[twitter_archive_cleaned.tweet_id == 835246439529840640, 'ra

#invalid tweet & hence remove it
twitter_archive_cleaned = twitter_archive_cleaned[twitter_archive_cleaned.tweet_id != 8
```

```
['This is Bella. She hopes her smile made you smile. If not, she is also offering you her favori
```

```
["This is Logan, the Chow who lived. He solemnly swears he's up to lots of good. H*ckin magical
["This is Sophie. She's a Jubilant Bush Pupper. Super h*ckin rare. Appears at random just to smi
['Here we have uncovered an entire battalion of holiday puppers. Average of 11.26/10 https://t.c
Correcting rating after manually going through tweet text :  ['@roushfenway These are good dogs
Correcting rating after manually going through tweet text :  ["@jonnysun @Lin_Manuel ok jomny I
```

Test:

```
In [40]: twitter_archive_cleaned.rating_numerator.value_counts()
```

```
Out[40]: 12.00      515
         10.00      442
         11.00      434
         13.00      321
         9.00       155
         8.00        98
         7.00        54
         14.00       48
         5.00        35
         6.00        32
         3.00        19
         4.00        16
         2.00         9
         1.00         6
         15.00        2
         0.00         2
         60.00        1
         88.00        1
         144.00       1
         44.00        1
         143.00       1
         182.00       1
         666.00       1
         99.00        1
         165.00       1
         9.75         1
         204.00       1
         121.00       1
         45.00        1
         1776.00      1
         50.00        1
         420.00       1
         11.27        1
         11.26        1
         13.50        1
         84.00        1
         24.00        1
```

```
80.00        1
Name: rating_numerator, dtype: int64
```

4.a source variable do not have the variables stored along with html tag; like other variable object variables should have the core value and not the residual of the extract.

Define: remove html tags from it

Code:

```
In [41]: twitter_archive_cleaned.source = twitter_archive_cleaned.source.apply(lambda x: x.split
```

Test

```
In [42]: twitter_archive_cleaned.source.value_counts()

Out[42]: Twitter for iPhone     2075
         Vine - Make a Scene      91
         Twitter Web Client       33
         TweetDeck                11
         Name: source, dtype: int64

In [43]: if os.path.exists("twitter_archive_cleaned_last_2.csv"):
             twitter_archive_cleaned = pd.read_csv("twitter_archive_cleaned_last_2.csv")
             twitter_archive_cleaned = twitter_archive_cleaned.drop('Unnamed: 0', 1)
         else:
             twitter_archive_cleaned['re_fav_reply'] = twitter_archive_cleaned.tweet_id.apply(la
             twitter_archive_cleaned.to_csv("twitter_archive_cleaned_last_2.csv")

         twitter_archive_cleaned['retweet_count'] = twitter_archive_cleaned['re_fav_reply'].appl
         twitter_archive_cleaned['favorite_count'] = twitter_archive_cleaned['re_fav_reply'].app
         twitter_archive_cleaned['reply_count'] = twitter_archive_cleaned['re_fav_reply'].apply(
         twitter_archive_cleaned = twitter_archive_cleaned.drop('re_fav_reply', 1)

In [44]: twitter_archive_cleaned['retweet_count'].value_counts()

Out[44]:         10
         129      6
         305      5
         587      5
         1127     5
         492      4
         1155     4
         389      4
         1340     4
         414      4
         328      4
         585      4
```

```
58       4
244      4
154      4
371      4
1295     4
2        3
298      3
555      3
870      3
1840     3
582      3
954      3
1902     3
542      3
3600     3
6296     3
663      3
421      3
        ..
748      1
3023     1
262      1
3008     1
5574     1
3442     1
890      1
2833     1
1524     1
728      1
1917     1
1624     1
794      1
1647     1
130      1
544      1
2456     1
1705     1
911      1
995      1
8532     1
4706     1
741      1
378      1
4636     1
56       1
128      1
1243     1
2412     1
```

```
         2680     1
         Name: retweet_count, Length: 1740, dtype: int64
```

```
In [45]: twitter_archive_cleaned['favorite_count'].value_counts()
```

```
Out[45]:          10
         761       3
         3472      3
         3838      3
         5400      3
         2311      3
         691       3
         19552     3
         2392      3
         1370      3
         2814      2
         3303      2
         4735      2
         476       2
         35237     2
         1743      2
         148       2
         3215      2
         7095      2
         2093      2
         2999      2
         2628      2
         4678      2
         1978      2
         2950      2
         3133      2
         17292     2
         21649     2
         181       2
         1561      2
                  ..
         5194      1
         3419      1
         19808     1
         6209      1
         3514      1
         256       1
         14925     1
         16303     1
         2735      1
         13394     1
         1367      1
         7128      1
         17839     1
         17391     1
         1172      1
         9434      1
         18342     1
         3409      1
         858       1
         1281      1
         11189     1
         53332     1
         108       1
         1896      1
         930       1
         13688     1
         7802      1
         1497      1
         4949      1
         3500      1
         Name: favorite_count, Length: 2019, dtype: int64
```

```
In [46]: twitter_archive_cleaned['reply_count'].value_counts()
```

```
Out[46]: 2        49
         15       46
         3        43
         18       43
         4        41
         16       41
         13       40
         27       39
         33       39
         6        38
         11       38
         9        37
         10       36
         14       36
         29       35
         5        34
         38       34
         28       33
         25       33
         22       33
         21       33
         8        32
         20       32
         34       31
         1        30
         23       30
```

```
19      30
26      30
31      30
17      30
        ..
235      1
243      1
659      1
814      1
2070     1
588      1
350      1
655      1
601      1
186      1
125      1
178      1
461      1
376      1
232      1
200      1
170      1
110      1
228      1
230      1
129      1
202      1
216      1
381      1
164      1
1008     1
244      1
176      1
422      1
184      1
Name: reply_count, Length: 252, dtype: int64
```

```
In [47]: twitter_archive_cleaned
```

```
Out[47]:                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
         0     666020888022790149                      0                    0
         1     666029285002620928                      0                    0
         2     666033412701032449                      0                    0
         3     666044226329800704                      0                    0
         4     666049248165822465                      0                    0
         5     666050758794694657                      0                    0
         6     666051853826850816                      0                    0
         7     666055525042405380                      0                    0
```

8	666057090499244032	0	0
9	666058600524156928	0	0
10	666063827256086533	0	0
11	666071193221509120	0	0
12	666073100786774016	0	0
13	666082916733198337	0	0
14	666094000022159362	0	0
15	666099513787052032	0	0
16	666102155909144576	0	0
17	666104133288665088	0	0
18	666268910803644416	0	0
19	666273097616637952	0	0
20	666287406224695296	0	0
21	666293911632134144	0	0
22	666337882303524864	0	0
23	666345417576210432	0	0
24	666353288456101888	0	0
25	666362758909284353	0	0
26	666373753744588802	0	0
27	666396247373291520	0	0
28	666407126856765440	0	0
29	666411507551481857	0	0
...	...	...	...
2180	886366144734445568	0	0
2181	886680336477933568	0	0
2182	886736880519319552	0	0
2183	886983233522544640	0	0
2184	887101392804085760	0	0
2185	887343217045368832	0	0
2186	887473957103951883	0	0
2187	887517139158093824	0	0
2188	887705289381826560	0	0
2189	888078434458587136	0	0
2190	888202515573088257	0	0
2191	888554962724278272	0	0
2192	888804989199671297	0	0
2193	888917238123831296	0	0
2194	889278841981685760	0	0
2195	889531135344209921	0	0
2196	889638837579907072	0	0
2197	889665388333682689	0	0
2198	889880896479866881	0	0
2199	890006608113172480	0	0
2200	890240255349198849	0	0
2201	890609185150312448	0	0
2202	890729181411237888	0	0
2203	890971913173991426	0	0
2204	891087950875897856	0	0

```
2205  891327558926688256                      0                    0
2206  891689557279858688                      0                    0
2207  891815181378084864                      0                    0
2208  892177421306343426                      0                    0
2209  892420643555336193                      0                    0

                timestamp              source  \
0     2015-11-15 22:32:08  Twitter for iPhone
1     2015-11-15 23:05:30  Twitter for iPhone
2     2015-11-15 23:21:54  Twitter for iPhone
3     2015-11-16 00:04:52  Twitter for iPhone
4     2015-11-16 00:24:50  Twitter for iPhone
5     2015-11-16 00:30:50  Twitter for iPhone
6     2015-11-16 00:35:11  Twitter for iPhone
7     2015-11-16 00:49:46  Twitter for iPhone
8     2015-11-16 00:55:59  Twitter for iPhone
9     2015-11-16 01:01:59  Twitter for iPhone
10    2015-11-16 01:22:45  Twitter for iPhone
11    2015-11-16 01:52:02  Twitter for iPhone
12    2015-11-16 01:59:36  Twitter for iPhone
13    2015-11-16 02:38:37  Twitter for iPhone
14    2015-11-16 03:22:39  Twitter for iPhone
15    2015-11-16 03:44:34  Twitter for iPhone
16    2015-11-16 03:55:04  Twitter for iPhone
17    2015-11-16 04:02:55  Twitter for iPhone
18    2015-11-16 14:57:41  Twitter for iPhone
19    2015-11-16 15:14:19  Twitter for iPhone
20    2015-11-16 16:11:11  Twitter for iPhone
21    2015-11-16 16:37:02  Twitter for iPhone
22    2015-11-16 19:31:45  Twitter for iPhone
23    2015-11-16 20:01:42  Twitter for iPhone
24    2015-11-16 20:32:58  Twitter for iPhone
25    2015-11-16 21:10:36  Twitter for iPhone
26    2015-11-16 21:54:18  Twitter for iPhone
27    2015-11-16 23:23:41  Twitter for iPhone
28    2015-11-17 00:06:54  Twitter for iPhone
29    2015-11-17 00:24:19  Twitter for iPhone
...                   ...                 ...
2180  2017-07-15 23:25:31  Twitter for iPhone
2181  2017-07-16 20:14:00  Twitter for iPhone
2182  2017-07-16 23:58:41  Twitter for iPhone
2183  2017-07-17 16:17:36  Twitter for iPhone
2184  2017-07-18 00:07:08  Twitter for iPhone
2185  2017-07-18 16:08:03  Twitter for iPhone
2186  2017-07-19 00:47:34  Twitter for iPhone
2187  2017-07-19 03:39:09  Twitter for iPhone
2188  2017-07-19 16:06:48  Twitter for iPhone
2189  2017-07-20 16:49:33  Twitter for iPhone
```

```
2190  2017-07-21 01:02:36  Twitter for iPhone
2191  2017-07-22 00:23:06  Twitter for iPhone
2192  2017-07-22 16:56:37  Twitter for iPhone
2193  2017-07-23 00:22:39  Twitter for iPhone
2194  2017-07-24 00:19:32  Twitter for iPhone
2195  2017-07-24 17:02:04  Twitter for iPhone
2196  2017-07-25 00:10:02  Twitter for iPhone
2197  2017-07-25 01:55:32  Twitter for iPhone
2198  2017-07-25 16:11:53  Twitter for iPhone
2199  2017-07-26 00:31:25  Twitter for iPhone
2200  2017-07-26 15:59:51  Twitter for iPhone
2201  2017-07-27 16:25:51  Twitter for iPhone
2202  2017-07-28 00:22:40  Twitter for iPhone
2203  2017-07-28 16:27:12  Twitter for iPhone
2204  2017-07-29 00:08:17  Twitter for iPhone
2205  2017-07-29 16:00:24  Twitter for iPhone
2206  2017-07-30 15:58:51  Twitter for iPhone
2207  2017-07-31 00:18:03  Twitter for iPhone
2208  2017-08-01 00:17:27  Twitter for iPhone
2209  2017-08-01 16:23:56  Twitter for iPhone

                                                   text  retweeted_status_id  \
0     Here we have a Japanese Irish Setter. Lost eye...                    0   
1     This is a western brown Mitsubishi terrier. Up...                    0   
2     Here is a very happy pup. Big fan of well-main...                    0   
3     This is a purebred Piers Morgan. Loves to Netf...                    0   
4     Here we have a 1949 1st generation vulpix. Enj...                    0   
5     This is a truly beautiful English Wilson Staff...                    0   
6     This is an odd dog. Hard on the outside but lo...                    0   
7     Here is a Siberian heavily armored polar bear ...                    0   
8     My oh my. This is a rare blond Canadian terrie...                    0   
9     Here is the Rand Paul of retrievers folks! He'...                    0   
10    This is the happiest dog you will ever see. Ve...                    0   
11    Here we have a northern speckled Rhododendron...                    0   
12    Let's hope this flight isn't Malaysian (lol). ...                    0   
13    Here we have a well-established sunblockerspan...                    0   
14    This appears to be a Mongolian Presbyterian mi...                    0   
15    Can stand on stump for what seems like a while...                    0   
16    Oh my. Here you are seeing an Adobe Setter giv...                    0   
17    Not familiar with this breed. No tail (weird)...                    0   
18    Very concerned about fellow dog trapped in com...                    0   
19       Can take selfies 11/10 https://t.co/ws2AMaNwPW                    0   
20    This is an Albanian 3 1/2 legged  Episcopalian...                    0   
21    This is a funny dog. Weird toes. Won't come do...                    0   
22    This is an extremely rare horned Parthenon. No...                    0   
23    Look at this jokester thinking seat belt laws ...                    0   
24    Here we have a mixed Asiago from the Galápagos...                    0   
25    Unique dog here. Very small. Lives in containe...                    0   
```

```
26    Those are sunglasses and a jean jacket. 11/10 ...                    0
27    Oh goodness. A super rare northeast Qdoba kang...                    0
28    This is a southern Vesuvius bumblegruff. Can d...                    0
29    This is quite the dog. Gets really excited whe...                    0
...                                                 ...                  ...
2180  This is Roscoe. Another pupper fallen victim t...                    0
2181  This is Derek. He's late for a dog meeting. 13...                    0
2182  This is Mingus. He's a wonderful father to his...                    0
2183  This is Maya. She's very shy. Rarely leaves he...                    0
2184  This... is a Jubilant Antarctic House Bear. We...                    0
2185  You may not have known you needed to see this ...                    0
2186  This is Canela. She attempted some fancy porch...                    0
2187  I've yet to rate a Venezuelan Hover Wiener. Th...                    0
2188  This is Jeffrey. He has a monopoly on the pool...                    0
2189  This is Gerald. He was just told he didn't get...                    0
2190  RT @dog_rates: This is Canela. She attempted s...   887473957103951872
2191  This is Ralphus. He's powering up. Attempting ...                    0
2192  This is Zeke. He has a new stick. Very proud o...                    0
2193  This is Jim. He found a fren. Taught him how t...                    0
2194  This is Oliver. You're witnessing one of his m...                    0
2195  This is Stuart. He's sporting his favorite fan...                    0
2196  This is Ted. He does his best. Sometimes that'...                    0
2197  Here's a puppo that seems to be on the fence a...                    0
2198  This is Bruno. He is a service shark. Only get...                    0
2199  This is Koda. He is a South Australian decksha...                    0
2200  This is Cassie. She is a college pup. Studying...                    0
2201  This is Zoey. She doesn't want to be one of th...                    0
2202  When you watch your owner call another dog a g...                    0
2203  Meet Jax. He enjoys ice cream so much he gets ...                    0
2204  Here we have a majestic great white breaching ...                    0
2205  This is Franklin. He would like you to stop ca...                    0
2206  This is Darla. She commenced a snooze mid meal...                    0
2207  This is Archie. He is a rare Norwegian Pouncin...                    0
2208  This is Tilly. She's just checking pup on you...                    0
2209  This is Phineas. He's a mystical boy. Only eve...                    0

      retweeted_status_user_id retweeted_status_timestamp  \
0                             0                        NaN
1                             0                        NaN
2                             0                        NaN
3                             0                        NaN
4                             0                        NaN
5                             0                        NaN
6                             0                        NaN
7                             0                        NaN
8                             0                        NaN
9                             0                        NaN
10                            0                        NaN
```

```
11                           0                        NaN
12                           0                        NaN
13                           0                        NaN
14                           0                        NaN
15                           0                        NaN
16                           0                        NaN
17                           0                        NaN
18                           0                        NaN
19                           0                        NaN
20                           0                        NaN
21                           0                        NaN
22                           0                        NaN
23                           0                        NaN
24                           0                        NaN
25                           0                        NaN
26                           0                        NaN
27                           0                        NaN
28                           0                        NaN
29                           0                        NaN
...                        ...                        ...
2180                         0                        NaN
2181                         0                        NaN
2182                         0                        NaN
2183                         0                        NaN
2184                         0                        NaN
2185                         0                        NaN
2186                         0                        NaN
2187                         0                        NaN
2188                         0                        NaN
2189                         0                        NaN
2190                4196983835        2017-07-19 00:47:34
2191                         0                        NaN
2192                         0                        NaN
2193                         0                        NaN
2194                         0                        NaN
2195                         0                        NaN
2196                         0                        NaN
2197                         0                        NaN
2198                         0                        NaN
2199                         0                        NaN
2200                         0                        NaN
2201                         0                        NaN
2202                         0                        NaN
2203                         0                        NaN
2204                         0                        NaN
2205                         0                        NaN
2206                         0                        NaN
2207                         0                        NaN
```

```
2208                          0                        NaN
2209                          0                        NaN

                                           expanded_urls  rating_numerator  \
0     https://twitter.com/dog_rates/status/666020888...               8.0
1     https://twitter.com/dog_rates/status/666029285...               7.0
2     https://twitter.com/dog_rates/status/666033412...               9.0
3     https://twitter.com/dog_rates/status/666044226...               6.0
4     https://twitter.com/dog_rates/status/666049248...               5.0
5     https://twitter.com/dog_rates/status/666050758...              10.0
6     https://twitter.com/dog_rates/status/666051853...               2.0
7     https://twitter.com/dog_rates/status/666055525...              10.0
8     https://twitter.com/dog_rates/status/666057090...               9.0
9     https://twitter.com/dog_rates/status/666058600...               8.0
10    https://twitter.com/dog_rates/status/666063827...              10.0
11    https://twitter.com/dog_rates/status/666071193...               9.0
12    https://twitter.com/dog_rates/status/666073100...              10.0
13    https://twitter.com/dog_rates/status/666082916...               6.0
14    https://twitter.com/dog_rates/status/666094000...               9.0
15    https://twitter.com/dog_rates/status/666099513...               8.0
16    https://twitter.com/dog_rates/status/666102155...              11.0
17    https://twitter.com/dog_rates/status/666104133...               1.0
18    https://twitter.com/dog_rates/status/666268910...              10.0
19    https://twitter.com/dog_rates/status/666273097...              11.0
20    https://twitter.com/dog_rates/status/666287406...               1.0
21    https://twitter.com/dog_rates/status/666293911...               3.0
22    https://twitter.com/dog_rates/status/666337882...               9.0
23    https://twitter.com/dog_rates/status/666345417...              10.0
24    https://twitter.com/dog_rates/status/666353288...               8.0
25    https://twitter.com/dog_rates/status/666362758...               6.0
26    https://twitter.com/dog_rates/status/666373753...              11.0
27    https://twitter.com/dog_rates/status/666396247...               9.0
28    https://twitter.com/dog_rates/status/666407126...               7.0
29    https://twitter.com/dog_rates/status/666411507...               2.0
...                                                  ...               ...
2180  https://twitter.com/dog_rates/status/886366144...              12.0
2181  https://twitter.com/dog_rates/status/886680336...              13.0
2182  https://www.gofundme.com/mingusneedsus,https:/...              13.0
2183  https://twitter.com/dog_rates/status/886983233...              13.0
2184  https://twitter.com/dog_rates/status/887101392...              12.0
2185  https://twitter.com/dog_rates/status/887343217...              13.0
2186  https://twitter.com/dog_rates/status/887473957...              13.0
2187  https://twitter.com/dog_rates/status/887517139...              14.0
2188  https://twitter.com/dog_rates/status/887705289...              13.0
2189  https://twitter.com/dog_rates/status/888078434...              12.0
2190  https://twitter.com/dog_rates/status/887473957...              13.0
2191  https://twitter.com/dog_rates/status/888554962...              13.0
2192  https://twitter.com/dog_rates/status/888804989...              13.0
```

```
2193  https://twitter.com/dog_rates/status/888917238...             12.0
2194  https://twitter.com/dog_rates/status/889278841...             13.0
2195  https://twitter.com/dog_rates/status/889531135...             13.0
2196  https://twitter.com/dog_rates/status/889638837...             12.0
2197  https://twitter.com/dog_rates/status/889665388...             13.0
2198  https://twitter.com/dog_rates/status/889880896...             13.0
2199  https://twitter.com/dog_rates/status/890006608...             13.0
2200  https://twitter.com/dog_rates/status/890240255...             14.0
2201  https://twitter.com/dog_rates/status/890609185...             13.0
2202  https://twitter.com/dog_rates/status/890729181...             13.0
2203  https://gofundme.com/ydvmve-surgery-for-jax,ht...             13.0
2204  https://twitter.com/dog_rates/status/891087950...             13.0
2205  https://twitter.com/dog_rates/status/891327558...             12.0
2206  https://twitter.com/dog_rates/status/891689557...             13.0
2207  https://twitter.com/dog_rates/status/891815181...             12.0
2208  https://twitter.com/dog_rates/status/892177421...             13.0
2209  https://twitter.com/dog_rates/status/892420643...             13.0

      rating_denominator      name                   breed_name   stage  \
0                     10      None       welsh_springer_spaniel     NaN
1                     10      None                      redbone     NaN
2                     10      None              german_shepherd     NaN
3                     10      None          rhodesian_ridgeback     NaN
4                     10      None           miniature_pinscher     NaN
5                     10      None         bernese_mountain_dog     NaN
6                     10      None                          NaN     NaN
7                     10      None                         chow     NaN
8                     10      None                          NaN     NaN
9                     10      None             miniature_poodle     NaN
10                    10      None             golden_retriever     NaN
11                    10      None                gordon_setter     NaN
12                    10      None                 walker_hound     NaN
13                    10      None                          pug     NaN
14                    10      None                   bloodhound     NaN
15                    10      None                        lhasa     NaN
16                    10      None               english_setter     NaN
17                    10      None                          NaN     NaN
18                    10      None                          NaN     NaN
19                    10      None            italian_greyhound     NaN
20                     2      None                  maltese_dog     NaN
21                    10      None                          NaN     NaN
22                    10      None                          NaN     NaN
23                    10      None             golden_retriever     NaN
24                    10      None                     malamute     NaN
25                    10      None                          NaN     NaN
26                    10      None  soft-coated_wheaten_terrier     NaN
27                    10      None                    chihuahua     NaN
28                    10      None      black-and-tan_coonhound     NaN
```

```
29                    10      None                          NaN     NaN
...                  ...       ...                          ...     ...
2180                  10    Roscoe               french_bulldog  pupper
2181                  10     Derek                          NaN     NaN
2182                  10    Mingus                       kuvasz     NaN
2183                  10      Maya                    chihuahua     NaN
2184                  10      None                      samoyed   floof
2185                  10      None             mexican_hairless     NaN
2186                  10    Canela                     pembroke     NaN
2187                  10      None                          NaN     NaN
2188                  10   Jeffrey                       basset     NaN
2189                  10    Gerald               french_bulldog     NaN
2190                  10    Canela                     pembroke     NaN
2191                  10   Ralphus               siberian_husky     NaN
2192                  10      Zeke             golden_retriever     NaN
2193                  10       Jim             golden_retriever     NaN
2194                  10    Oliver                      whippet     NaN
2195                  10    Stuart             golden_retriever   puppo
2196                  10       Ted               french_bulldog     NaN
2197                  10      None                     pembroke   puppo
2198                  10     Bruno               french_bulldog     NaN
2199                  10      Koda                      samoyed     NaN
2200                  10    Cassie                     pembroke   doggo
2201                  10      Zoey                irish_terrier     NaN
2202                  10      None                   pomeranian     NaN
2203                  10       Jax                  appenzeller     NaN
2204                  10      None     chesapeake_bay_retriever     NaN
2205                  10  Franklin                       basset     NaN
2206                  10     Darla                          NaN     NaN
2207                  10    Archie                    chihuahua     NaN
2208                  10     Tilly                    chihuahua     NaN
2209                  10   Phineas                          NaN     NaN

     retweet_count favorite_count reply_count
0              515           2562          25
1               47            130           0
2               44            125           1
3              141            299           1
4               40            108           7
5               58            133           0
6              848           1217          11
7              249            434           2
8              141            296           2
9               57            111           4
10             218            476           3
11              60            148           2
12             163            322           3
13              44            119           2
```

14	73	164	2
15	67	155	1
16	12	80	0
17	6604	14317	124
18	35	103	2
19	76	174	1
20	64	148	3
21	354	505	6
22	91	198	2
23	136	295	3
24	72	220	1
25	571	778	7
26	93	189	2
27	84	168	3
28	40	110	1
29	327	447	4
...	...	...	...
2180	3196	21096	82
2181	4457	22316	87
2182	3291	12019	56
2183	7767	34984	158
2184	5957	30394	146
2185	10390	33528	203
2186	18200	68799	253
2187	11661	46040	484
2188	5389	30046	90
2189	3485	21666	111
2190			
2191	3580	19808	84
2192	4339	25459	74
2193	4501	28947	93
2194	5419	25168	124
2195	2236	15030	61
2196	4548	27044	107
2197	10063	47908	202
2198	4976	27651	98
2199	7342	30543	171
2200	7427	31787	170
2201	4270	27670	104
2202	18909	65208	195
2203	2073	11797	63
2204	3115	20127	67
2205	9412	40141	220
2206	8660	42010	168
2207	4156	24910	127
2208	6270	33085	198
2209	8532	38585	166

```
[2210 rows x 18 columns]
```

It seems like to_csv and read_csv changes the datatype. Let us leave this as is and take care of same in visualization.

```
In [48]: twitter_archive_cleaned.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2210 entries, 0 to 2209
Data columns (total 18 columns):
tweet_id                      2210 non-null int64
in_reply_to_status_id         2210 non-null int64
in_reply_to_user_id           2210 non-null int64
timestamp                     2210 non-null object
source                        2210 non-null object
text                          2210 non-null object
retweeted_status_id           2210 non-null int64
retweeted_status_user_id      2210 non-null int64
retweeted_status_timestamp    69 non-null object
expanded_urls                 2169 non-null object
rating_numerator              2210 non-null float64
rating_denominator            2210 non-null int64
name                          2210 non-null object
breed_name                    1512 non-null object
stage                         426 non-null object
retweet_count                 2210 non-null object
favorite_count                2210 non-null object
reply_count                   2210 non-null object
dtypes: float64(1), int64(6), object(11)
memory usage: 310.9+ KB
```

```
In [49]: twitter_archive_cleaned.to_csv("twitter_archive_master.csv")
```

We are done with cleaning. . . . . . Let us move on for visualization :)