wrangle_act

September 4, 2022

1 Project: Wrangling and Analyze Data

1.1 Data Gathering

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

```
In [803]: # libraries needed
          import pandas as pd
          import numpy as np
          import seaborn as sns
          import os
          import matplotlib.pyplot as plt
          %matplotlib inline
          import requests
          import warnings
          warnings.simplefilter('ignore')
```

```
In [804]: #reading the twitter data
          t_archive = pd.read_csv('twitter_archive_enhanced.csv')
```

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

```
In [805]: #Dowloading the tweet image using requests
          url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predict
          response = requests.get(url)

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

          image_pred = pd.read_csv('image_predictions.tsv', sep = '\t')
```

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

```
In [806]: #reading the tweet_json data
          import tweepy
          from tweepy import OAuthHandler
```

```
         import json
         from timeit import default_timer as timer

         with open('tweet_json.txt', encoding='utf8') as file:
             lst=[]
             for line in file:
                 lst.append(json.loads(line))
```

```
In [807]: t_list = []

          for line in open('tweet_json.txt', 'r'):
              t_data = json.loads(line)
              t_list.append({'tweet_id': t_data['id_str'],
                             'retweet_count': t_data['retweet_count'],
                             'favorite_count': t_data['favorite_count'],
                             'followers_count': t_data['user']['followers_count']})
```

```
In [808]: # convert our dictionaries to a dataframe
          t_data = pd.DataFrame( t_list , columns = ['tweet_id' , 'retweet_count', 'favorite_cou
```

1.2 Assessing Data

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

Visual Assesment

```
In [809]: # we are just going to view the entire data manually
          t_archive
```

```
Out[809]:                 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 [810]: # we are doing a virtual assesment of image data
          image_pred
```

```
Out[810]:                 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]
```

```
In [811]: # we are just gonna do the virtual assesment here
          t_data
```

```
Out[811]:                 tweet_id  retweet_count  favorite_count  followers_count
          0     892420643555336193           8853           39467          3200889
          1     892177421306343426           6514           33819          3200889
          2     891815181378084864           4328           25461          3200889
          3     891689557279858688           8964           42908          3200889
          4     891327558926688256           9774           41048          3200889
          5     891087950875897856           3261           20562          3200889
          6     890971913173991426           2158           12041          3200889
          7     890729181411237888          16716           56848          3200889
          8     890609185150312448           4429           28226          3200889
          9     890240255349198849           7711           32467          3200889
          10    890006608113172480           7624           31166          3200889
          11    889880896479866881           5156           28268          3200889
          12    889665388333682689           8538           38818          3200889
          13    889638837579907072           4735           27672          3200889
          14    889531135344209921           2321           15359          3200889
          15    889278841981685760           5637           25652          3200889
          16    888917238123831296           4709           29611          3200889
          17    888804989199671297           4559           26080          3200889
          18    888554962724278272           3732           20290          3200889
          19    888078434458587136           3653           22201          3200889
          20    887705289381826560           5609           30779          3200890
          21    887517139158093824          12082           46959          3200890
          22    887473957103951883          18781           69871          3200890
          23    887343217045368832          10737           34222          3200890
          24    887101392804085760           6167           31061          3200890
          25    886983233522544640           8084           35859          3200890
          26    886736880519319552           3443           12306          3200890
          27    886680336477933568           4610           22798          3200890
          28    886366144734445568           3316           21524          3200890
          29    886267009285017600              4             117          3200890
          ...                  ...            ...             ...              ...
          2324  666411507551481857            339             459          3201017
          2325  666407126856765440             44             113          3201018
          2326  666396247373291520             92             172          3201018
          2327  666373753744588802            100             194          3201018
```

```
2328  666362758909284353            595             804          3201018
2329  666353288456101888             77             229          3201018
2330  666345417576210432            146             307          3201018
2331  666337882303524864             96             204          3201018
2332  666293911632134144            368             522          3201018
2333  666287406224695296             71             152          3201018
2334  666273097616637952             82             184          3201018
2335  666268910803644416             37             108          3201018
2336  666104133288665088           6871           14765          3201018
2337  666102155909144576             16              81          3201018
2338  666099513787052032             73             164          3201018
2339  666094000022159362             79             169          3201018
2340  666082916733198337             47             121          3201018
2341  666073100786774016            174             335          3201018
2342  666071193221509120             67             154          3201018
2343  666063827256086533            232             496          3201018
2344  666058600524156928             61             115          3201018
2345  666057090499244032            146             304          3201018
2346  666055525042405380            261             448          3201018
2347  666051853826850816            879            1253          3201018
2348  666050758794694657             60             136          3201018
2349  666049248165822465             41             111          3201018
2350  666044226329800704            147             311          3201018
2351  666033412701032449             47             128          3201018
2352  666029285002620928             48             132          3201018
2353  666020888022790149            532            2535          3201018

[2354 rows x 4 columns]
```

Programmatic Assesment

```
In [812]: # viewing  sample of the t_archive data
          t_archive.sample()
```

```
Out[812]:                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
          898  758854675097526272                    NaN                  NaN   

                                timestamp  \
          898  2016-07-29 02:40:28 +0000   

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

                                                            text  retweeted_status_id  \
          898  This is Lilli Bee &amp; Honey Bear. Unfortunat...                  NaN   

               retweeted_status_user_id retweeted_status_timestamp  \
          898                       NaN                        NaN   
```

```
                                          expanded_urls  rating_numerator  \
898  https://twitter.com/dog_rates/status/758854675...                11

     rating_denominator   name doggo floofer pupper puppo
898                  10  Lilli  None    None   None  None
```

```
In [813]: #doing a programmatic assesment on the t_archive data
          t_archive.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2356 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2356 non-null object
source                        2356 non-null object
text                          2356 non-null object
retweeted_status_id           181 non-null float64
retweeted_status_user_id      181 non-null float64
retweeted_status_timestamp    181 non-null object
expanded_urls                 2297 non-null object
rating_numerator              2356 non-null int64
rating_denominator            2356 non-null int64
name                          2356 non-null object
doggo                         2356 non-null object
floofer                       2356 non-null object
pupper                        2356 non-null object
puppo                         2356 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

```
In [814]: # seeing the different counts of the numerator rating
          t_archive.rating_numerator.value_counts()
```

```
Out[814]: 12      558
          11      464
          10      461
          13      351
          9       158
          8       102
          7        55
          14       54
          5        37
          6        32
          3        19
          4        17
```

```
1          9
2          9
420        2
0          2
15         2
75         2
80         1
20         1
24         1
26         1
44         1
50         1
60         1
165        1
84         1
88         1
144        1
182        1
143        1
666        1
960        1
1776       1
17         1
27         1
45         1
99         1
121        1
204        1
Name: rating_numerator, dtype: int64
```

```
In [815]: # counts of the denominator ratings
          t_archive.rating_denominator.value_counts()
```

```
Out[815]: 10     2333
          11        3
          50        3
          80        2
          20        2
          2         1
          16        1
          40        1
          70        1
          15        1
          90        1
          110       1
          120       1
          130       1
          150       1
```

```
170      1
7        1
0        1
Name: rating_denominator, dtype: int64
```

```
In [816]: # the number of times different names has appered in our data
          t_archive.name.value_counts()
```

```
Out[816]: None         745
          a             55
          Charlie       12
          Oliver        11
          Lucy          11
          Cooper        11
          Lola          10
          Penny         10
          Tucker        10
          Winston        9
          Bo             9
          Sadie          8
          the            8
          Buddy          7
          Daisy          7
          Toby           7
          an             7
          Bailey         7
          Oscar          6
          Rusty          6
          Jack           6
          Koda           6
          Milo           6
          Jax            6
          Scout          6
          Stanley        6
          Leo            6
          Bella          6
          Dave           6
          Phil           5
                      ...
          Scott          1
          Jeremy         1
          Durg           1
          Goose          1
          Aiden          1
          Erik           1
          Aubie          1
          Kuyu           1
          Boston         1
          Butters        1
          Sonny          1
          Todo           1
          Tino           1
          Batdog         1
          Dotsy          1
          Willy          1
          Rizzo          1
          Adele          1
          Sprinkles      1
          Storkson       1
          Anna           1
          Karl           1
          Grady          1
          Saydee         1
          Iggy           1
          Apollo         1
          Mojo           1
          Tebow          1
          Shnuggles      1
          Kloey          1
          Name: name, Length: 957, dtype: int64
```

```
In [817]: # view some statistics of the data
          t_archive.describe()
```

```
Out[817]:            tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
          count  2.356000e+03           7.800000e+01         7.800000e+01
          mean   7.427716e+17           7.455079e+17         2.014171e+16
          std    6.856705e+16           7.582492e+16         1.252797e+17
          min    6.660209e+17           6.658147e+17         1.185634e+07
          25%    6.783989e+17           6.757419e+17         3.086374e+08
          50%    7.196279e+17           7.038708e+17         4.196984e+09
          75%    7.993373e+17           8.257804e+17         4.196984e+09
          max    8.924206e+17           8.862664e+17         8.405479e+17

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

                 rating_denominator
          count         2356.000000
```

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

```
In [818]: # image dataset
          image_pred.head()
```

```
Out[818]:              tweet_id                                          jpg_url  \
          0  666020888022790149  https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg
          1  666029285002620928  https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg
          2  666033412701032449  https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg
          3  666044226329800704  https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg
          4  666049248165822465  https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg

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

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

```
In [819]: #lets look at one of the images which belongs to tweet_id 666020888022790149
          from IPython.display import Image
          Image(url = 'https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg')
```

```
Out[819]: <IPython.core.display.Image object>
```

```
In [820]: # info of the image data
          image_pred.info()
```

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

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

```
In [821]: # value counts of p1
          image_pred.p1.value_counts()
```

```
Out[821]: golden_retriever             150
          Labrador_retriever           100
          Pembroke                      89
          Chihuahua                     83
          pug                           57
          chow                          44
          Samoyed                       43
          toy_poodle                    39
          Pomeranian                    38
          malamute                      30
          cocker_spaniel                30
          French_bulldog                26
          miniature_pinscher            23
          Chesapeake_Bay_retriever      23
          seat_belt                     22
          Staffordshire_bullterrier     20
          German_shepherd               20
          Siberian_husky                20
          Cardigan                      19
          web_site                      19
          Maltese_dog                   18
          Eskimo_dog                    18
          Shetland_sheepdog             18
          beagle                        18
          teddy                         18
          Rottweiler                    17
          Shih-Tzu                      17
          Lakeland_terrier              17
          Italian_greyhound             16
          kuvasz                        16
                                      ...
          snowmobile                     1
          microphone                     1
          nail                           1
```

```
beach_wagon                    1
trombone                       1
polecat                        1
dhole                          1
zebra                          1
lynx                           1
sandbar                        1
padlock                        1
rapeseed                       1
china_cabinet                  1
sea_urchin                     1
pedestal                       1
restaurant                     1
lorikeet                       1
robin                          1
American_black_bear            1
school_bus                     1
walking_stick                  1
syringe                        1
washer                         1
lacewing                       1
African_grey                   1
African_hunting_dog            1
leaf_beetle                    1
suit                           1
mailbox                        1
radio_telescope                1
Name: p1, Length: 378, dtype: int64
```

```
In [822]: # value counts of p2
          image_pred.p2.value_counts()
```

```
Out[822]: Labrador_retriever                104
          golden_retriever                   92
          Cardigan                           73
          Chihuahua                          44
          Pomeranian                         42
          Chesapeake_Bay_retriever           41
          French_bulldog                     41
          toy_poodle                         37
          cocker_spaniel                     34
          miniature_poodle                   33
          Siberian_husky                     33
          beagle                             28
          collie                             27
          Pembroke                           27
          Eskimo_dog                         27
          kuvasz                             26
```

```
Italian_greyhound                 22
Pekinese                          21
American_Staffordshire_terrier    21
chow                              20
malinois                          20
miniature_pinscher                20
toy_terrier                       20
Samoyed                           20
Boston_bull                       19
Norwegian_elkhound                19
Staffordshire_bullterrier         18
Irish_terrier                     17
pug                               17
Shih-Tzu                          16
                                ...
purse                              1
grey_whale                         1
Japanese_spaniel                   1
umbrella                           1
giant_panda                        1
necklace                           1
crutch                             1
oxygen_mask                        1
apron                              1
ashcan                             1
stingray                           1
sunglass                           1
space_heater                       1
cannon                             1
mosquito_net                       1
can_opener                         1
mud_turtle                         1
crate                              1
cowboy_boot                        1
sweatshirt                         1
komondor                           1
Kerry_blue_terrier                 1
chain_mail                         1
shower_curtain                     1
streetcar                          1
spindle                            1
Windsor_tie                        1
tiger                              1
projectile                         1
laptop                             1
Name: p2, Length: 405, dtype: int64
```

```
In [823]: # value counts of p3
```

```
          image_pred.p3.value_counts()

Out[823]: Labrador_retriever                79
          Chihuahua                         58
          golden_retriever                  48
          Eskimo_dog                        38
          kelpie                            35
          kuvasz                            34
          Staffordshire_bullterrier         32
          chow                              32
          cocker_spaniel                    31
          beagle                            31
          toy_poodle                        29
          Pomeranian                        29
          Pekinese                          29
          Pembroke                          27
          Chesapeake_Bay_retriever          27
          Great_Pyrenees                    27
          malamute                          26
          French_bulldog                    26
          American_Staffordshire_terrier    24
          Cardigan                          23
          pug                               23
          basenji                           21
          toy_terrier                       20
          bull_mastiff                      20
          Siberian_husky                    19
          Shetland_sheepdog                 17
          Boston_bull                       17
          boxer                             16
          Lakeland_terrier                  16
          doormat                           16
                                            ..
          power_drill                        1
          rhinoceros_beetle                  1
          cup                                1
          cuirass                            1
          shoji                              1
          steam_locomotive                   1
          bannister                          1
          pot                                1
          loggerhead                         1
          restaurant                         1
          stinkhorn                          1
          hammerhead                         1
          jaguar                             1
          mosquito_net                       1
          chest                              1
          crayfish                           1
          partridge                          1
          wok                                1
          cliff                              1
          chimpanzee                         1
          tiger_cat                          1
          grocery_store                      1
          snail                              1
          otter                              1
          shovel                             1
          African_grey                       1
          chickadee                          1
          shower_cap                         1
          bell_cote                          1
          banana                             1
          Name: p3, Length: 408, dtype: int64
```

```
In [824]: # info of the t_data for programmatic assessment
          t_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 4 columns):
tweet_id           2354 non-null object
retweet_count      2354 non-null int64
favorite_count     2354 non-null int64
followers_count    2354 non-null int64
dtypes: int64(3), object(1)
memory usage: 73.6+ KB
```

```
In [825]: # some statistics
          t_data.describe()
```

```
Out[825]:        retweet_count  favorite_count  followers_count
          count    2354.000000     2354.000000     2.354000e+03
          mean     3164.797366     8080.968564     3.200942e+06
          std      5284.770364    11814.771334     4.457302e+01
          min         0.000000        0.000000     3.200799e+06
          25%       624.500000     1415.000000     3.200898e+06
          50%      1473.500000     3603.500000     3.200945e+06
          75%      3652.000000    10122.250000     3.200953e+06
          max     79515.000000   132810.000000     3.201018e+06
```

1.2.1 Quality issues

Archive table 1. Remove the retweets and keep the original tweets

2. Remove columns that we dont need for this analysis

3. code tweet_id with the right data type code
4. code timestamp with the right data type code
5. Missing names that are represented as none
6. Remove the hyperlinks from the text
7. Source column is in form of HTML but not a normal string format

Image table

8. Diffreent dogs in columns p1, p2, and p3 had some uppercase and lowercase letters.

1.2.2 Tidiness issues

1. create one column for the different dog stages do away with the previous ones
2. All the three tables should be merged
3. Drop the tweets that doesnt have images

1.3 Cleaning Data

```
In [826]: # Make copies of original pieces of data
          t_archive_clean = t_archive.copy()
          image_pred_clean = image_pred.copy()
          t_data_clean = t_data.copy()
```

1.3.1 1. Define: since we are just going to use the original tweets we have to remove all the retweets

Code

```
In [827]: # removing the retweets
          t_archive_clean = t_archive_clean[pd.isnull(t_archive_clean['retweeted_status_user_id'
```

Test

```
In [828]: print(sum(t_archive_clean.retweeted_status_user_id.value_counts()))
```

```
0
```

1.3.2 2. Define: We are going to remove the columns that we are not going to use

Code

```
In [829]: # dropping some columns
          t_archive_clean = t_archive_clean.drop(['in_reply_to_status_id', 'in_reply_to_user_id'
```

Test

```
In [830]: t_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 11 columns):
tweet_id              2175 non-null int64
timestamp             2175 non-null object
source                2175 non-null object
text                  2175 non-null object
rating_numerator      2175 non-null int64
rating_denominator    2175 non-null int64
name                  2175 non-null object
doggo                 2175 non-null object
floofer               2175 non-null object
pupper                2175 non-null object
puppo                 2175 non-null object
dtypes: int64(3), object(8)
memory usage: 203.9+ KB
```

1.3.3 3. Define : Changing the tweet_id data type from integer to string

Code

```
In [831]: # changing the data types
          t_archive_clean.tweet_id = t_archive_clean.tweet_id.astype(str)
          t_data_clean.tweet_id = t_data_clean.tweet_id.astype(str)
          image_pred_clean.tweet_id = image_pred_clean.tweet_id.astype(str)
```

Test

```
In [832]: t_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 11 columns):
tweet_id              2175 non-null object
timestamp             2175 non-null object
source                2175 non-null object
text                  2175 non-null object
rating_numerator      2175 non-null int64
rating_denominator    2175 non-null int64
name                  2175 non-null object
doggo                 2175 non-null object
floofer               2175 non-null object
pupper                2175 non-null object
puppo                 2175 non-null object
```

```
dtypes: int64(2), object(9)
memory usage: 203.9+ KB
```

```
In [833]: t_data_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 4 columns):
tweet_id           2354 non-null object
retweet_count      2354 non-null int64
favorite_count     2354 non-null int64
followers_count    2354 non-null int64
dtypes: int64(3), object(1)
memory usage: 73.6+ KB
```

```
In [834]: image_pred_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null object
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(1), object(5)
memory usage: 152.1+ KB
```

1.3.4 4. Define : Changing the data type of timestamp

Code

```
In [835]: # changing the timestamp datatyoe
          t_archive_clean.timestamp= pd.to_datetime(t_archive_clean.timestamp)
```

Test

```
In [836]: t_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 11 columns):
tweet_id              2175 non-null object
timestamp             2175 non-null datetime64[ns]
source                2175 non-null object
text                  2175 non-null object
rating_numerator      2175 non-null int64
rating_denominator    2175 non-null int64
name                  2175 non-null object
doggo                 2175 non-null object
floofer               2175 non-null object
pupper                2175 non-null object
puppo                 2175 non-null object
dtypes: datetime64[ns](1), int64(2), object(8)
memory usage: 203.9+ KB
```

1.3.5 5. Define: Correcting the names that are represented as none

Code

```
In [837]: t_archive_clean['name'] = t_archive_clean['name'].replace('None', np.nan)
```

Test

```
In [838]: t_archive_clean.name.value_counts()
```

```
Out[838]: a              55
          Lucy           11
          Charlie        11
          Oliver         10
          Cooper         10
          Penny           9
          Tucker          9
          Sadie           8
          the             8
          Lola            8
          Winston         8
          Daisy           7
          Toby            7
          Bailey          6
          Stanley         6
          Bella           6
          an              6
          Jax             6
          Oscar           6
          Koda            6
          Bo              6
```

```
Buddy         5
Bentley       5
Chester       5
Scout         5
Louis         5
Rusty         5
Leo           5
Milo          5
Dave          5
             ..
Kenzie        1
Milky         1
Christoper    1
Bubba         1
Nollie        1
Mojo          1
Apollo        1
Iggy          1
Rover         1
Anakin        1
Tilly         1
Sierra        1
DonDon        1
Kuyu          1
Butters       1
Sonny         1
Todo          1
Tino          1
Batdog        1
Dotsy         1
Willy         1
Rizzo         1
Adele         1
Boston        1
Sprinkles     1
Anna          1
Karl          1
Grady         1
Saydee        1
Kloey         1
Name: name, Length: 955, dtype: int64
```

1.3.6 6. Define: Remove the hyperlinks from the text

Code

```
In [839]: #define function and apply to t_archive_clean table
          def htmlink(x):
```

```
    http_pos = x.find("http")
    # If no link, retain row
    if http_pos == -1:
      x = x
    else:
      # Remove space before link to end
      x = x[:http_pos - 1]
    return x

t_archive_clean.text = t_archive_clean.text.apply(htmlink)
```

Test

```
In [840]: # confirm the code
          for row in t_archive_clean.text[:4]:
              print(row)
```

```
This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10
This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available fo
This is Archie. He is a rare Norwegian Pouncing Corgo. Lives in the tall grass. You never know w
This is Darla. She commenced a snooze mid meal. 13/10 happens to the best of us
```

1.3.7 7.Define: Source column is in form of HTML but not a normal string format

Code

```
In [841]: # extract the values
          t_archive_clean.source = t_archive_clean.source.str.extract('>([\w\W\s]*)<', expand=Tr
```

Test

```
In [842]: # confirm chages
          t_archive_clean.source.value_counts()
```

```
Out[842]: Twitter for iPhone     2042
          Vine - Make a Scene      91
          Twitter Web Client       31
          TweetDeck                11
          Name: source, dtype: int64
```

1.3.8 8. Define Changing the case of P1 , p2 and p3 to lowercase

Code

```
In [843]: # change to lower case
          image_pred_clean.p1 = image_pred_clean.p1.str.lower()
          image_pred_clean.p2 = image_pred_clean.p2.str.lower()
          image_pred_clean.p3 = image_pred_clean.p3.str.lower()
```

Test

```
In [844]: # confirm changes
          image_pred_clean.p1

Out[844]: 0            welsh_springer_spaniel
          1                           redbone
          2                   german_shepherd
          3               rhodesian_ridgeback
          4                miniature_pinscher
          5              bernese_mountain_dog
          6                        box_turtle
          7                              chow
          8                     shopping_cart
          9                  miniature_poodle
          10                 golden_retriever
          11                    gordon_setter
          12                     walker_hound
          13                              pug
          14                       bloodhound
          15                            lhasa
          16                   english_setter
          17                              hen
          18                 desktop_computer
          19                italian_greyhound
          20                      maltese_dog
          21                 three-toed_sloth
          22                               ox
          23                 golden_retriever
          24                         malamute
          25                       guinea_pig
          26      soft-coated_wheaten_terrier
          27                        chihuahua
          28          black-and-tan_coonhound
          29                             coho
                             ...
          2045                 french_bulldog
          2046                    convertible
          2047                         kuvasz
          2048                      chihuahua
          2049                        samoyed
          2050               mexican_hairless
          2051                       pembroke
          2052                      limousine
          2053                         basset
          2054                 french_bulldog
          2055                       pembroke
          2056                 siberian_husky
          2057               golden_retriever
          2058               golden_retriever
          2059                        whippet
          2060               golden_retriever
          2061                 french_bulldog
          2062                       pembroke
          2063                 french_bulldog
          2064                        samoyed
          2065                       pembroke
          2066                  irish_terrier
          2067                     pomeranian
          2068                    appenzeller
          2069       chesapeake_bay_retriever
          2070                         basset
          2071                    paper_towel
          2072                      chihuahua
          2073                      chihuahua
          2074                         orange
          Name: p1, Length: 2075, dtype: object
```

```
In [845]: # confirm changes
          image_pred_clean.p2
```

```
Out[845]: 0                         collie
          1             miniature_pinscher
          2                       malinois
          3                        redbone
          4                     rottweiler
          5               english_springer
          6                     mud_turtle
          7                tibetan_mastiff
          8                shopping_basket
          9                       komondor
          10               tibetan_mastiff
          11             yorkshire_terrier
          12              english_foxhound
          13                  bull_mastiff
          14               german_shepherd
          15                      shih-tzu
          16                  newfoundland
          17                          cock
          18                          desk
          19                   toy_terrier
          20                    toy_poodle
          21                         otter
          22                  newfoundland
          23      chesapeake_bay_retriever
          24                siberian_husky
          25                         skunk
          26                  afghan_hound
          27                   toy_terrier
          28                    bloodhound
          29                    barracouta
                            ...
          2045                   chihuahua
          2046                  sports_car
          2047              great_pyrenees
          2048                 toy_terrier
          2049                  eskimo_dog
          2050                    sea_lion
          2051         rhodesian_ridgeback
          2052                   tow_truck
          2053                     redbone
          2054                         pug
          2055         rhodesian_ridgeback
          2056                  eskimo_dog
          2057          labrador_retriever
          2058             tibetan_mastiff
          2059                      borzoi
          2060          labrador_retriever
          2061                       boxer
          2062                    cardigan
          2063          labrador_retriever
          2064                  pomeranian
          2065                    cardigan
          2066                irish_setter
          2067                  eskimo_dog
          2068               border_collie
          2069               irish_terrier
          2070            english_springer
          2071          labrador_retriever
          2072                    malamute
          2073                    pekinese
          2074                       bagel
          Name: p2, Length: 2075, dtype: object
```

```
In [846]: # confirm changes
          image_pred_clean.p3
```

```
Out[846]: 0                 shetland_sheepdog
          1               rhodesian_ridgeback
          2                        bloodhound
          3                miniature_pinscher
          4                          doberman
          5        greater_swiss_mountain_dog
          6                          terrapin
          7                          fur_coat
```

```
8                  golden_retriever
9       soft-coated_wheaten_terrier
10               labrador_retriever
11                         pekinese
12                     ibizan_hound
13                   french_bulldog
14                         malinois
15                    dandie_dinmont
16                           borzoi
17                        partridge
18                         bookcase
19                          basenji
20                 miniature_poodle
21                   great_grey_owl
22                      groenendael
23               labrador_retriever
24                       eskimo_dog
25                          hamster
26                           briard
27                         papillon
28            flat-coated_retriever
29                              gar
                   ...
2045                    boston_bull
2046                      car_wheel
2047                 dandie_dinmont
2048                     can_opener
2049    staffordshire_bullterrier
2050                     weimaraner
2051                         beagle
2052                  shopping_cart
2053                     weimaraner
2054                    bull_mastiff
2055                         beagle
2056                       malamute
2057                 english_setter
2058             labrador_retriever
2059                         saluki
2060                        redbone
2061    staffordshire_bullterrier
2062                        basenji
2063                         muzzle
2064                           chow
2065                      chihuahua
2066      chesapeake_bay_retriever
2067                       pembroke
2068                      ice_lolly
2069                 indian_elephant
```

```
2070    german_short-haired_pointer
2071                        spatula
2072                         kelpie
2073                       papillon
2074                         banana
Name: p3, Length: 2075, dtype: object
```

1.3.9 Tidiness

1.3.10 1. Define. Creating a column to represent the different dog stages

```
In [847]: #replace the none
          t_archive_clean.doggo.replace('None', '', inplace=True)
          t_archive_clean.floofer.replace('None', '', inplace=True)
          t_archive_clean.pupper.replace('None', '', inplace=True)
          t_archive_clean.puppo.replace('None', '', inplace=True)

          # Then combine stage columns.
          t_archive_clean['dog_stage'] = t_archive_clean.doggo + t_archive_clean.floofer + t_arc

          # drop the unused columns
          t_archive_clean.drop(['doggo', 'floofer', 'pupper', 'puppo'], axis=1, inplace=True)

          # Then format entries with multiple dog stages which appeared like doggopupper.
          t_archive_clean.loc[t_archive_clean.dog_stage == 'doggopupper', 'dog_stage'] = 'doggo,
          t_archive_clean.loc[t_archive_clean.dog_stage == 'doggopuppo', 'dog_stage'] = 'doggo,p
          t_archive_clean.loc[t_archive_clean.dog_stage == 'doggofloofer', 'dog_stage'] = 'doggo
```

1.3.11 Test

```
In [848]: # change the dog datatype
          t_archive_clean.dog_stage = t_archive_clean.dog_stage.astype('category')
```

```
In [849]: # confirm changes
          t_archive_clean.dog_stage.value_counts()
```

```
Out[849]:                  1831
          pupper            224
          doggo              75
          puppo              24
          doggo,pupper       10
          floofer             9
          doggo,puppo         1
          doggo,floofer       1
          Name: dog_stage, dtype: int64
```

1.3.12 2. Define: Merge the three tables

Code

```
In [850]: # merging the tables
          t_archive_clean = pd.merge(left=t_archive_clean, right=t_data_clean, how='left', on='t
          t_archive_clean = pd.merge(left=t_archive_clean, right=image_pred_clean, how='left', o
```

Test

```
In [851]: # confirm changes
          t_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2174
Data columns (total 22 columns):
tweet_id              2175 non-null object
timestamp             2175 non-null datetime64[ns]
source                2175 non-null object
text                  2175 non-null object
rating_numerator      2175 non-null int64
rating_denominator    2175 non-null int64
name                  1495 non-null object
dog_stage             2175 non-null category
retweet_count         2175 non-null int64
favorite_count        2175 non-null int64
followers_count       2175 non-null int64
jpg_url               1994 non-null object
img_num               1994 non-null float64
p1                    1994 non-null object
p1_conf               1994 non-null float64
p1_dog                1994 non-null object
p2                    1994 non-null object
p2_conf               1994 non-null float64
p2_dog                1994 non-null object
p3                    1994 non-null object
p3_conf               1994 non-null float64
p3_dog                1994 non-null object
dtypes: category(1), datetime64[ns](1), float64(4), int64(5), object(11)
memory usage: 376.3+ KB
```

1.3.13 3 Define : Drop the tweets with zero images

Code

```
In [852]: # dropping the tweets with zero image
          t_archive_clean.dropna(axis=0 ,inplace= True)
```

Test

```
In [853]: t_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1448 entries, 0 to 2173
Data columns (total 22 columns):
tweet_id              1448 non-null object
timestamp             1448 non-null datetime64[ns]
source                1448 non-null object
text                  1448 non-null object
rating_numerator      1448 non-null int64
rating_denominator    1448 non-null int64
name                  1448 non-null object
dog_stage             1448 non-null category
retweet_count         1448 non-null int64
favorite_count        1448 non-null int64
followers_count       1448 non-null int64
jpg_url               1448 non-null object
img_num               1448 non-null float64
p1                    1448 non-null object
p1_conf               1448 non-null float64
p1_dog                1448 non-null object
p2                    1448 non-null object
p2_conf               1448 non-null float64
p2_dog                1448 non-null object
p3                    1448 non-null object
p3_conf               1448 non-null float64
p3_dog                1448 non-null object
dtypes: category(1), datetime64[ns](1), float64(4), int64(5), object(11)
memory usage: 250.7+ KB
```

1.3.14 Save the clean data as a master dataset

```
In [854]: # first we make a copy of the clean dataset
          t_archive_clean1 = t_archive_clean.copy()
```

```
In [855]: t_archive_clean1.to_csv('twitter_archive_master.csv', index= False)
```

Test

```
In [856]: twitter_archive_master = pd.read_csv('twitter_archive_master.csv')
```

```
In [857]: twitter_archive_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1448 entries, 0 to 1447
Data columns (total 22 columns):
tweet_id              1448 non-null int64
timestamp             1448 non-null object
source                1448 non-null object
text                  1448 non-null object
```

```
rating_numerator      1448 non-null int64
rating_denominator    1448 non-null int64
name                  1448 non-null object
dog_stage             189 non-null object
retweet_count         1448 non-null int64
favorite_count        1448 non-null int64
followers_count       1448 non-null int64
jpg_url               1448 non-null object
img_num               1448 non-null float64
p1                    1448 non-null object
p1_conf               1448 non-null float64
p1_dog                1448 non-null bool
p2                    1448 non-null object
p2_conf               1448 non-null float64
p2_dog                1448 non-null bool
p3                    1448 non-null object
p3_conf               1448 non-null float64
p3_dog                1448 non-null bool
dtypes: bool(3), float64(4), int64(6), object(9)
memory usage: 219.3+ KB
```

1.4 Analyzing and Visualizing Data

```
In [858]: twitter_archive_master = pd.read_csv('twitter_archive_master.csv')
```

```
In [859]: twitter_archive_master. info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1448 entries, 0 to 1447
Data columns (total 22 columns):
tweet_id              1448 non-null int64
timestamp             1448 non-null object
source                1448 non-null object
text                  1448 non-null object
rating_numerator      1448 non-null int64
rating_denominator    1448 non-null int64
name                  1448 non-null object
dog_stage             189 non-null object
retweet_count         1448 non-null int64
favorite_count        1448 non-null int64
followers_count       1448 non-null int64
jpg_url               1448 non-null object
img_num               1448 non-null float64
p1                    1448 non-null object
p1_conf               1448 non-null float64
p1_dog                1448 non-null bool
p2                    1448 non-null object
```

```
p2_conf                 1448 non-null float64
p2_dog                  1448 non-null bool
p3                      1448 non-null object
p3_conf                 1448 non-null float64
p3_dog                  1448 non-null bool
dtypes: bool(3), float64(4), int64(6), object(9)
memory usage: 219.3+ KB
```

```
In [860]: # We have to change our data types again
          twitter_archive_master.tweet_id = twitter_archive_master.tweet_id.astype(str)
          twitter_archive_master.timestamp = pd.to_datetime(twitter_archive_master.timestamp)
          twitter_archive_master.source = twitter_archive_master.source.astype("category")
```

```
In [861]: twitter_archive_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1448 entries, 0 to 1447
Data columns (total 22 columns):
tweet_id                1448 non-null object
timestamp               1448 non-null datetime64[ns]
source                  1448 non-null category
text                    1448 non-null object
rating_numerator        1448 non-null int64
rating_denominator      1448 non-null int64
name                    1448 non-null object
dog_stage               189 non-null object
retweet_count           1448 non-null int64
favorite_count          1448 non-null int64
followers_count         1448 non-null int64
jpg_url                 1448 non-null object
img_num                 1448 non-null float64
p1                      1448 non-null object
p1_conf                 1448 non-null float64
p1_dog                  1448 non-null bool
p2                      1448 non-null object
p2_conf                 1448 non-null float64
p2_dog                  1448 non-null bool
p3                      1448 non-null object
p3_conf                 1448 non-null float64
p3_dog                  1448 non-null bool
dtypes: bool(3), category(1), datetime64[ns](1), float64(4), int64(5), object(8)
memory usage: 209.5+ KB
```

1.4.1 Insights:

1. Look at the most popular dog name

2. favorites count over time
3. Retweet count overtime
4. correlation between favorite count and retweet

1.4.2 Analysis and Visualization

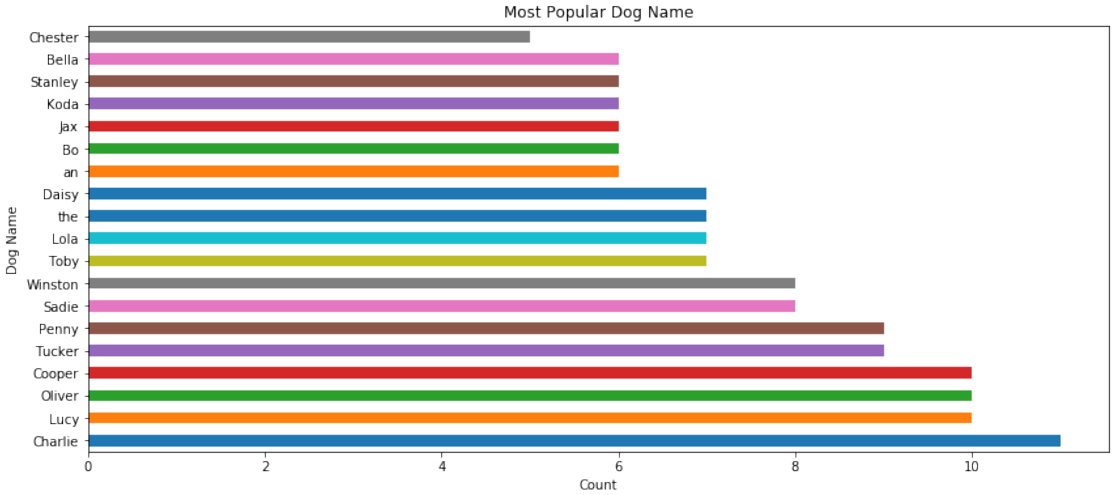
1.Most popular dog name

```
In [862]: twitter_archive_master.name.value_counts()
```

```
Out[862]: a              55
          Charlie        11
          Lucy           10
          Oliver         10
          Cooper         10
          Tucker          9
          Penny           9
          Sadie           8
          Winston         8
          Toby            7
          Lola            7
          the             7
          Daisy           7
          an              6
          Bo              6
          Jax             6
          Koda            6
          Stanley         6
          Bella           6
          Chester         5
          Leo             5
          Buddy           5
          Rusty           5
          Bailey          5
          Oscar           5
          Dave            5
          Scout           5
          Milo            5
          Louis           5
          Maggie          4
                         ..
          Milky           1
          Christoper      1
          Bubba           1
          Scott           1
          Shnuggles       1
          Nollie          1
```

```
Tilly          1
Mojo           1
Sierra         1
DonDon         1
Kuyu           1
Butters        1
Sonny          1
Todo           1
Tino           1
Batdog         1
Dotsy          1
Divine         1
Willy          1
Rizzo          1
Adele          1
Boston         1
Sprinkles      1
Anna           1
Karl           1
Grady          1
Saydee         1
Iggy           1
Apollo         1
Kloey          1
Name: name, Length: 935, dtype: int64
```

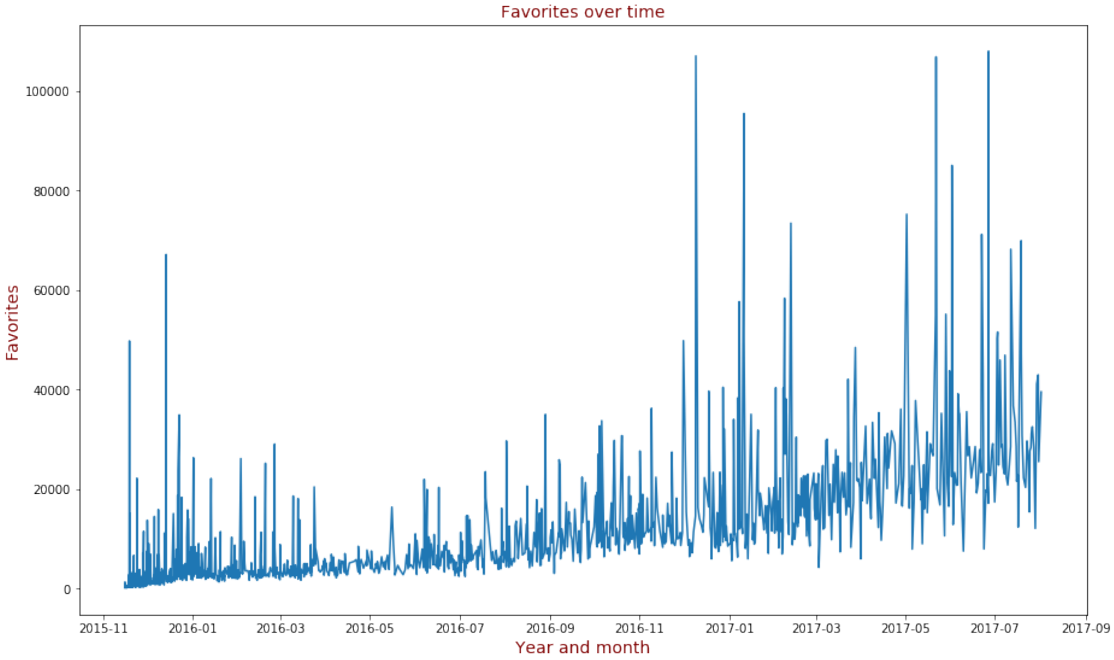
```
In [863]: twitter_archive_master.name.value_counts()[1:20].plot('barh', figsize=(14,6))
          plt.title('Most Popular Dog Name')
          plt.xlabel('Count')
          plt.ylabel('Dog Name');
```



We can see that the most popular dog is Charlie with a value of 11

2. Favorites count overtime

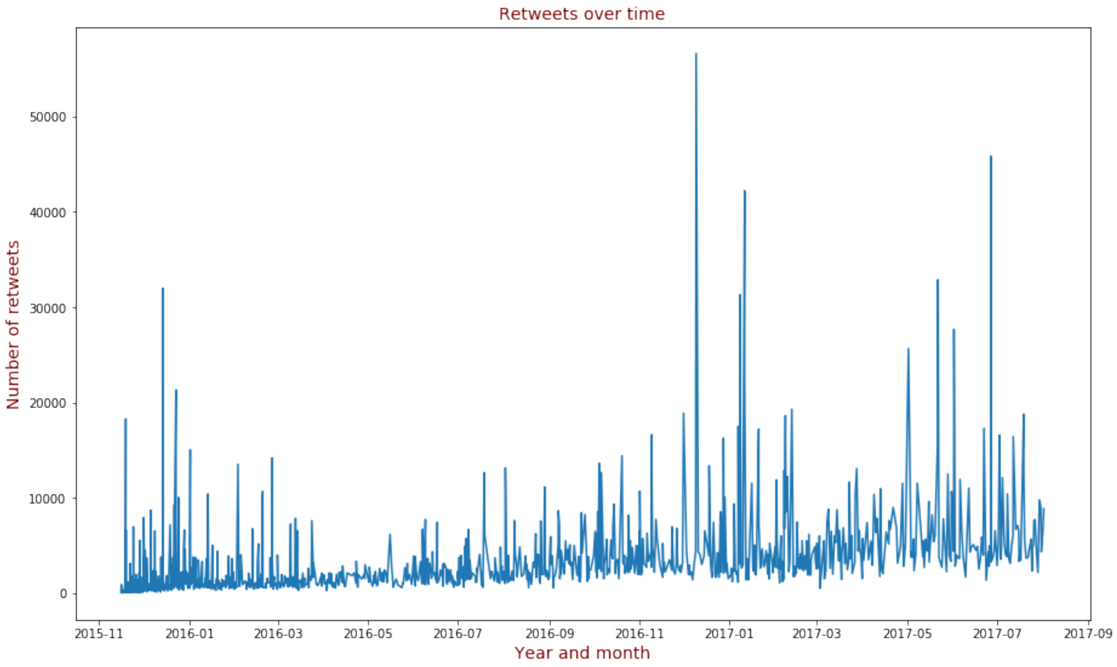
```
In [864]: sns.set_context()
          plt.subplots(figsize=(15, 9))
          plt.plot(twitter_archive_master.timestamp, twitter_archive_master.favorite_count)
          plt.title('Favorites over time', color = 'maroon', fontsize = '14')
          plt.xlabel('Year and month', color = 'maroon', fontsize = '14')
          plt.ylabel('Favorites', color = 'maroon', fontsize = '14');
```



From our data we can see that favorites spiked in the year 2016 and 2017

3. Retweets overtime

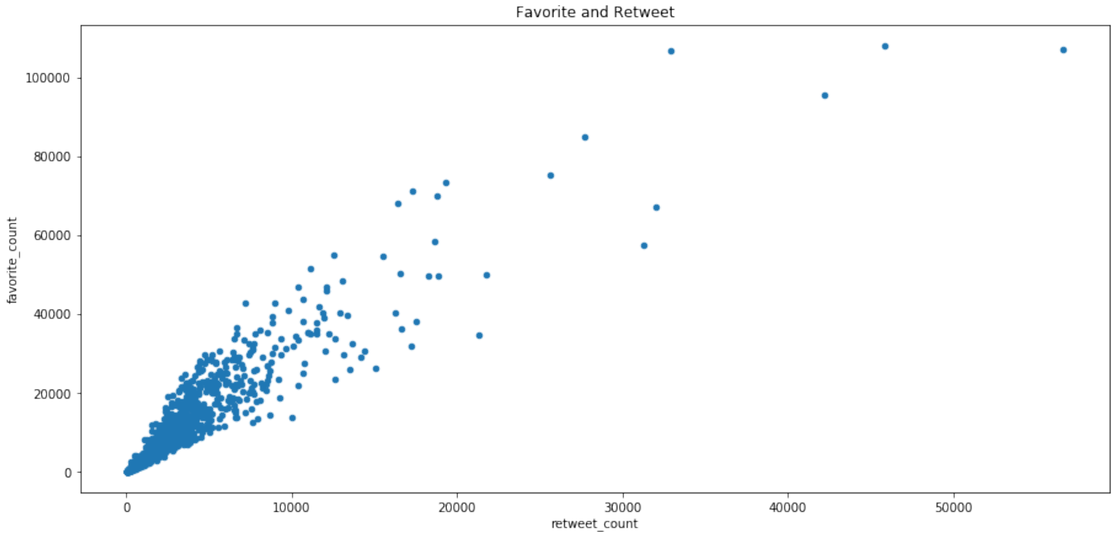
```
In [865]: sns.set_context()
          plt.subplots(figsize=(15, 9))
          plt.plot(twitter_archive_master.timestamp, twitter_archive_master.retweet_count)
          plt.title('Retweets over time', color = 'maroon', fontsize = '14')
          plt.xlabel('Year and month', color = 'maroon', fontsize = '14')
          plt.ylabel('Number of retweets', color = 'maroon', fontsize = '14');
```



Retweets really increased in the year 2016 towrds 2017

4. Correlation between favorite count and retweets

```
In [866]: twitter_archive_master.plot(x='retweet_count', y='favorite_count', kind='scatter',figs
```



```
In [867]: twitter_archive_master['retweet_count'].corr(twitter_archive_master['favorite_count'])
```

```
Out[867]: 0.9240070692143515
```

From the above data we can see there is a positive correlation between retweets and favorites which has a value of 0.924

```
In [ ]:
```