wrangle_act

December 29, 2017

1 Wrangle and Analyze Data Project

1.1 1. Gather Data

```
In [4]: # Load packages
        import tweepy
        import numpy as np
        import pandas as pd
        from bs4 import BeautifulSoup
        import os
        import requests
        from PIL import Image
        from io import BytesIO
        import json
```

1.1.1 The WeRateDogs Twitter archive

```
In [5]: twitter_archive_enhanced = pd.read_csv('twitter-archive-enhanced.csv')
```

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

```
Out[6]:              tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
        0  892420643555336193                    NaN                  NaN
        1  892177421306343426                    NaN                  NaN
        2  891815181378084864                    NaN                  NaN
        3  891689557279858688                    NaN                  NaN
        4  891327558926688256                    NaN                  NaN

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

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

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

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

   retweeted_status_user_id retweeted_status_timestamp  \
0                       NaN                        NaN
1                       NaN                        NaN
2                       NaN                        NaN
3                       NaN                        NaN
4                       NaN                        NaN

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

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

1.1.2 Image Predictions

```
In [7]: tweet_image_prediction_url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/
        response = requests.get(tweet_image_prediction_url)
        with open('image_predictions.tsv', mode = 'wb') as file:
            file.write(response.content)
```

1.1.3 Tweet's Data Stats

```
In [8]: # consumer_key, consumer_secret, access_token and access_secret should be achieved
        # by creating a twitter account and twitter app on https://apps.twitter.com/app

        # consumer_key = 'YOUR CONSUMER KEY'
        # consumer_secret = 'YOUR CONSUMER SECRET'
        # access_token = 'YOUR ACCESS TOKEN'
        # access_secret = 'YOUR ACCESS SECRET'
        # auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
        # auth.set_access_token(access_token, access_secret)

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

In [9]: # Use this for retrive json data
        # with open('tweet_json.txt', 'w') as outfile:
        #     data = []
        #     for id in twitter_archive_enhanced['tweet_id']:
        #         try:
        #             tweet = api.get_status(id, tweet_mode='extended')
        #             data.append(tweet._json)
        #         except:
        #             pass
        #     json.dump(data, outfile)

In [10]: # Test if data was written successfully
         with open('tweet_json.txt') as json_file:
             data = json.load(json_file)
             for tweet in data[:4]:
                 print(tweet)

{'created_at': 'Tue Aug 01 16:23:56 +0000 2017', 'id': 892420643555336193, 'id_str': '89242064
{'created_at': 'Tue Aug 01 00:17:27 +0000 2017', 'id': 892177421306343426, 'id_str': '89217742
{'created_at': 'Mon Jul 31 00:18:03 +0000 2017', 'id': 891815181378084864, 'id_str': '89181518
{'created_at': 'Sun Jul 30 15:58:51 +0000 2017', 'id': 891689557279858688, 'id_str': '89168955


In [11]: # read json file
         tweets = pd.read_json('tweet_json.txt')

In [12]: tweets.tail()

Out[12]:       contributors  coordinates          created_at display_text_range  \
         2344           NaN          NaN 2015-11-16 00:24:50           [0, 120]
         2345           NaN          NaN 2015-11-16 00:04:52           [0, 137]
         2346           NaN          NaN 2015-11-15 23:21:54           [0, 130]
         2347           NaN          NaN 2015-11-15 23:05:30           [0, 139]
         2348           NaN          NaN 2015-11-15 22:32:08           [0, 131]

                                                        entities  \
         2344  {'hashtags': [], 'symbols': [], 'user_mentions...
         2345  {'hashtags': [], 'symbols': [], 'user_mentions...
         2346  {'hashtags': [], 'symbols': [], 'user_mentions...
         2347  {'hashtags': [], 'symbols': [], 'user_mentions...
         2348  {'hashtags': [], 'symbols': [], 'user_mentions...

                                              extended_entities  favorite_count  \
         2344  {'media': [{'id': 666049244999131136, 'id_str'...             110
         2345  {'media': [{'id': 666044217047650304, 'id_str'...             306
         2346  {'media': [{'id': 666033409081393153, 'id_str'...             127
         2347  {'media': [{'id': 666029276303482880, 'id_str'...             131
         2348  {'media': [{'id': 666020881337073664, 'id_str'...            2531

               favorited                                          full_text  geo  \
         2344      False  Here we have a 1949 1st generation vulpix. Enj...  NaN
         2345      False  This is a purebred Piers Morgan. Loves to Netf...  NaN
         2346      False  Here is a very happy pup. Big fan of well-main...  NaN
         2347      False  This is a western brown Mitsubishi terrier. Up...  NaN
         2348      False  Here we have a Japanese Irish Setter. Lost eye...  NaN

                                     ...                          \
         2344                        ...
         2345                        ...
         2346                        ...
         2347                        ...
         2348                        ...

               possibly_sensitive_appealable  quoted_status quoted_status_id  \
         2344                            0.0            NaN              NaN
         2345                            0.0            NaN              NaN
         2346                            0.0            NaN              NaN
         2347                            0.0            NaN              NaN
         2348                            0.0            NaN              NaN

               quoted_status_id_str  retweet_count  retweeted  retweeted_status  \
         2344                   NaN             40      False               NaN
         2345                   NaN            143      False               NaN
         2346                   NaN             46      False               NaN
         2347                   NaN             47      False               NaN
         2348                   NaN            527      False               NaN

                                                          source truncated  \
         2344  <a href="http://twitter.com/download/iphone" r...     False
         2345  <a href="http://twitter.com/download/iphone" r...     False
         2346  <a href="http://twitter.com/download/iphone" r...     False
         2347  <a href="http://twitter.com/download/iphone" r...     False
         2348  <a href="http://twitter.com/download/iphone" r...     False

                                                            user
         2344  {'id': 4196983835, 'id_str': '4196983835', 'na...
         2345  {'id': 4196983835, 'id_str': '4196983835', 'na...
         2346  {'id': 4196983835, 'id_str': '4196983835', 'na...
         2347  {'id': 4196983835, 'id_str': '4196983835', 'na...
         2348  {'id': 4196983835, 'id_str': '4196983835', 'na...
```

```
[5 rows x 31 columns]
```

```
In [13]: tweets.describe()
```

```
Out[13]:        contributors  coordinates  favorite_count  geo            id  \
         count           0.0          0.0     2349.000000  0.0  2.349000e+03
         mean            NaN          NaN     8109.720732  NaN  7.424674e+17
         std             NaN          NaN    12051.729881  NaN  6.840987e+16
         min             NaN          NaN        0.000000  NaN  6.660209e+17
         25%             NaN          NaN     1415.000000  NaN  6.783890e+17
         50%             NaN          NaN     3585.000000  NaN  7.193325e+17
         75%             NaN          NaN    10100.000000  NaN  7.989257e+17
         max             NaN          NaN   131820.000000  NaN  8.924206e+17

                     id_str  in_reply_to_status_id  in_reply_to_status_id_str  \
         count  2.349000e+03           7.800000e+01               7.800000e+01
         mean   7.424674e+17           7.455079e+17               7.455079e+17
         std    6.840987e+16           7.582492e+16               7.582492e+16
         min    6.660209e+17           6.658147e+17               6.658147e+17
         25%    6.783890e+17           6.757419e+17               6.757419e+17
         50%    7.193325e+17           7.038708e+17               7.038708e+17
         75%    7.989257e+17           8.257804e+17               8.257804e+17
         max    8.924206e+17           8.862664e+17               8.862664e+17

                in_reply_to_user_id  in_reply_to_user_id_str  possibly_sensitive  \
         count         7.800000e+01             7.800000e+01              2208.0
         mean          2.014171e+16             2.014171e+16                 0.0
         std           1.252797e+17             1.252797e+17                 0.0
         min           1.185634e+07             1.185634e+07                 0.0
         25%           3.086374e+08             3.086374e+08                 0.0
         50%           4.196984e+09             4.196984e+09                 0.0
         75%           4.196984e+09             4.196984e+09                 0.0
         max           8.405479e+17             8.405479e+17                 0.0

                possibly_sensitive_appealable  quoted_status_id  quoted_status_id_str  \
         count                         2208.0      2.900000e+01          2.900000e+01
         mean                             0.0      8.162686e+17          8.162686e+17
         std                              0.0      6.164161e+16          6.164161e+16
         min                              0.0      6.721083e+17          6.721083e+17
         25%                              0.0      7.888183e+17          7.888183e+17
         50%                              0.0      8.340867e+17          8.340867e+17
         75%                              0.0      8.664587e+17          8.664587e+17
         max                              0.0      8.860534e+17          8.860534e+17

                retweet_count
         count    2349.000000
         mean     3097.263942
         std      5133.834333
         min         0.000000
         25%       615.000000
         50%      1445.000000
         75%      3606.000000
         max     78754.000000
```

```
In [14]: tweets.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2349 entries, 0 to 2348
Data columns (total 31 columns):
contributors                       0 non-null float64
coordinates                        0 non-null float64
created_at                         2349 non-null datetime64[ns]
display_text_range                 2349 non-null object
entities                           2349 non-null object
extended_entities                  2070 non-null object
favorite_count                     2349 non-null int64
favorited                          2349 non-null bool
full_text                          2349 non-null object
geo                                0 non-null float64
id                                 2349 non-null int64
id_str                             2349 non-null int64
in_reply_to_screen_name            78 non-null object
in_reply_to_status_id              78 non-null float64
in_reply_to_status_id_str          78 non-null float64
in_reply_to_user_id                78 non-null float64
in_reply_to_user_id_str            78 non-null float64
is_quote_status                    2349 non-null bool
lang                               2349 non-null object
place                              1 non-null object
possibly_sensitive                 2208 non-null float64
possibly_sensitive_appealable      2208 non-null float64
quoted_status                      28 non-null object
quoted_status_id                   29 non-null float64
quoted_status_id_str               29 non-null float64
retweet_count                      2349 non-null int64
retweeted                          2349 non-null bool
retweeted_status                   174 non-null object
source                             2349 non-null object
truncated                          2349 non-null bool
user                               2349 non-null object
dtypes: bool(4), datetime64[ns](1), float64(11), int64(4), object(11)
memory usage: 523.0+ KB
```

```
In [15]: # Keep only tweet ID, retweet count, and favorite count
         tweets = tweets[['user', 'favorite_count', 'retweet_count']]
```

```
In [16]: tweets.head()
```

```
Out[16]:                                                 user  favorite_count  \
         0  {'id': 4196983835, 'id_str': '4196983835', 'na...           39334
         1  {'id': 4196983835, 'id_str': '4196983835', 'na...           33671
         2  {'id': 4196983835, 'id_str': '4196983835', 'na...           25363
         3  {'id': 4196983835, 'id_str': '4196983835', 'na...           42708
         4  {'id': 4196983835, 'id_str': '4196983835', 'na...           40850

            retweet_count
         0           8786
         1           6439
         2           4270
         3           8877
         4           9654
```

1.2 2. Assessing Data

```
In [17]: twitter_archive_enhanced.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
```

1.2.1 Quality Issues

twitter_archive_enhanced: - We don't need reply tweets - We don't need retweets - Wrong time format - Some expanded urls are missing - Uncorrect rating compared with text - Missing Name - Unclear tag(doggo, floofer, pupper, puppo) - Source should not include link

1.2.2 Tidyness

twitter_archive_enhanced: - Unuseful dog tag
tweets - User id displayed in wrong format

1.3 Cleaning Data

1.3.1 Quality

```
In [18]: # Make a copy of twitter_archive_enhanced
         twitter_archive_enhanced_clean = twitter_archive_enhanced.copy()
```

```
In [19]: # 1. Remove the reply tweets and drop in_reply_to_status_id and in_reply_to_user_id c
         # only keep the tweets that in_reply_to_user_id is NA then drop the two columns
         twitter_archive_enhanced_clean = twitter_archive_enhanced_clean[twitter_archive_enhanc
         twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.drop(['in_reply_to_sta
```

```
In [20]: # 2. Remove the reply retweets and drop retweeted_status_id,
         # retweeted_status_user_id and retweeted_status_timestamp column
         # only keep the tweets that retweeted_status_id is NA then drop the two columns
         twitter_archive_enhanced_clean = twitter_archive_enhanced_clean[twitter_archive_enhanc
         twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.drop(['retweeted_statu
```

```
In [21]: # 3. Correct format of timestamp
         # use to_datetime function to clean timestamp column
         twitter_archive_enhanced_clean['timestamp'] = pd.to_datetime(twitter_archive_enhanced_
```

```
In [22]: # 4 & 5. Uncorrect rating and extract short url instead of using the expanded url
         from fractions import Fraction
         def clean_ratings_and_urls(twitter_archive_enhanced_clean):
             # achieve the text from text column
             text = twitter_archive_enhanced_clean.text
             # replace \n \t , with space
             text = text.replace('\n', ' ')
             text = text.replace('\t', ' ')
             text = text.replace(',', ' ')
             # split the text with space for further extraction
             text = text.split(' ')
             # set up the numerator and denominator for extract ratings
             numerator, denominator, url = 0, 0, None
             for s in text:
             # for string in text list, if 't.co' in string means it's a url
                 if 't.co' in s:
                     url = s
                 else:
                     try:
                         # try to put string into fucntion fraction, if the string can be
                         # evaluated as a rational number, then we can get numerator and denom
                         num = Fraction(s)
```

```
                numerator = num.numerator
                denominator = num.denominator
                # change the base ratings to 10
                if denominator != 10:
                    numerator *= 10/denominator
                    denominator = 10
            except:
                pass
    # update numerator, denominator, and url
    twitter_archive_enhanced_clean['rating_numerator'] = int(numerator)
    twitter_archive_enhanced_clean['rating_denominator'] = denominator
    twitter_archive_enhanced_clean['url'] = url
    return twitter_archive_enhanced_clean

twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.apply(clean_ratings_ar
```

In [23]:

```
# 6. Missing Name
def clean_dog_name(twitter_archive_enhanced_clean):
    # change all the wrong names to unknown
    if twitter_archive_enhanced_clean['name'].islower():
        return 'Unknown'
    elif twitter_archive_enhanced_clean['name'] == 'None':
        return 'Unknown'
    else:
        return twitter_archive_enhanced_clean['name']
twitter_archive_enhanced_clean['name'] = twitter_archive_enhanced_clean.apply(clean_do

twitter_archive_enhanced_clean.tail(50)
```

Out[23]:

```
                tweet_id           timestamp  \
2306  666835007768551424 2015-11-18 04:27:09
2307  666826780179869698 2015-11-18 03:54:28
2308  666817836334096384 2015-11-18 03:18:55
2309  666804364988780544 2015-11-18 02:25:23
2310  666786068205871104 2015-11-18 01:12:41
2311  666781792255496192 2015-11-18 00:55:42
2312  666776908487630848 2015-11-18 00:36:17
2313  666739327293083650 2015-11-17 22:06:57
2314  666701168228331520 2015-11-17 19:35:19
2315  666691418707132416 2015-11-17 18:56:35
2316  666649482315059201 2015-11-17 16:09:56
2317  666644823164719104 2015-11-17 15:51:26
2318  666454714377183233 2015-11-17 03:16:00
2319  666447344410484738 2015-11-17 02:46:43
2320  666437273139982337 2015-11-17 02:06:42
2321  666435652385423360 2015-11-17 02:00:15
2322  666430724426358785 2015-11-17 01:40:41
2323  666428276349472768 2015-11-17 01:30:57
```

```
2324  666421158376562688 2015-11-17 01:02:40
2325  666418789513326592 2015-11-17 00:53:15
2326  666411507551481857 2015-11-17 00:24:19
2327  666407126856765440 2015-11-17 00:06:54
2328  666396247373291520 2015-11-16 23:23:41
2329  666373753744588802 2015-11-16 21:54:18
2330  666362758909284353 2015-11-16 21:10:36
2331  666353288456101888 2015-11-16 20:32:58
2332  666345417576210432 2015-11-16 20:01:42
2333  666337882303524864 2015-11-16 19:31:45
2334  666293911632134144 2015-11-16 16:37:02
2335  666287406224695296 2015-11-16 16:11:11
2336  666273097616637952 2015-11-16 15:14:19
2337  666268910803644416 2015-11-16 14:57:41
2338  666104133288665088 2015-11-16 04:02:55
2339  666102155909144576 2015-11-16 03:55:04
2340  666099513787052032 2015-11-16 03:44:34
2341  666094000022159362 2015-11-16 03:22:39
2342  666082916733198337 2015-11-16 02:38:37
2343  666073100786774016 2015-11-16 01:59:36
2344  666071193221509120 2015-11-16 01:52:02
2345  666063827256086533 2015-11-16 01:22:45
2346  666058600524156928 2015-11-16 01:01:59
2347  666057090499244032 2015-11-16 00:55:59
2348  666055525042405380 2015-11-16 00:49:46
2349  666051853826850816 2015-11-16 00:35:11
2350  666050758794694657 2015-11-16 00:30:50
2351  666049248165822465 2015-11-16 00:24:50
2352  666044226329800704 2015-11-16 00:04:52
2353  666033412701032449 2015-11-15 23:21:54
2354  666029285002620928 2015-11-15 23:05:30
2355  666020888022790149 2015-11-15 22:32:08

                                                 source  \
2306  <a href="http://twitter.com/download/iphone" r...
2307  <a href="http://twitter.com/download/iphone" r...
2308  <a href="http://twitter.com/download/iphone" r...
2309  <a href="http://twitter.com/download/iphone" r...
2310  <a href="http://twitter.com/download/iphone" r...
2311  <a href="http://twitter.com/download/iphone" r...
2312  <a href="http://twitter.com/download/iphone" r...
2313  <a href="http://twitter.com/download/iphone" r...
2314  <a href="http://twitter.com/download/iphone" r...
2315  <a href="http://twitter.com/download/iphone" r...
2316  <a href="http://twitter.com/download/iphone" r...
2317  <a href="http://twitter.com/download/iphone" r...
2318  <a href="http://twitter.com/download/iphone" r...
2319  <a href="http://twitter.com/download/iphone" r...
```

```
2320  <a href="http://twitter.com/download/iphone" r...
2321  <a href="http://twitter.com/download/iphone" r...
2322  <a href="http://twitter.com/download/iphone" r...
2323  <a href="http://twitter.com/download/iphone" r...
2324  <a href="http://twitter.com/download/iphone" r...
2325  <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  \
2306  These are Peruvian Feldspars. Their names are ...
2307  12/10 simply brilliant pup https://t.co/V6ZzG4...
2308  This is Jeph. He is a German Boston Shuttlecoc...
2309  This is Jockson. He is a Pinnacle Sagittarius...
2310  Unfamiliar with this breed. Ears pointy af. Wo...
2311  This is a purebred Bacardi named Octaviath. Ca...
2312  This is Josep. He is a Rye Manganese mix. Can ...
2313  This is Lugan. He is a Bohemian Rhapsody. Very...
2314  This is a golden Buckminsterfullerene named Jo...
2315  This is Christoper. He is a spotted Penne. Can...
```

```
2316  Cool dog. Enjoys couch. Low monotone bark. Ver...
2317  This is Jimothy. He is a Botwanian Gouda. Can ...
2318  I'll name the dogs from now on. This is Kreggo...
2319  This is Scout. She is a black Downton Abbey. I...
2320  Here we see a lone northeastern Cumberbatch. H...
2321  "Can you behave? You're ruining my wedding day...
2322  Oh boy what a pup! Sunglasses take this one to...
2323  Here we have an Austrian Pulitzer. Collectors ...
2324  *internally screaming* 12/10 https://t.co/YMcr...
2325  This is Walter. He is an Alaskan Terrapin. Lov...
2326  This is quite the dog. Gets really excited whe...
2327  This is a southern Vesuvius bumblegruff. Can d...
2328  Oh goodness. A super rare northeast Qdoba kang...
2329  Those are sunglasses and a jean jacket. 11/10 ...
2330  Unique dog here. Very small. Lives in containe...
2331  Here we have a mixed Asiago from the Galápagos...
2332  Look at this jokester thinking seat belt laws ...
2333  This is an extremely rare horned Parthenon. No...
2334  This is a funny dog. Weird toes. Won't come do...
2335  This is an Albanian 3 1/2 legged  Episcopalian...
2336     Can take selfies 11/10 https://t.co/ws2AMaNwPW
2337  Very concerned about fellow dog trapped in com...
2338  Not familiar with this breed. No tail (weird)...
2339  Oh my. Here you are seeing an Adobe Setter giv...
2340  Can stand on stump for what seems like a while...
2341  This appears to be a Mongolian Presbyterian mi...
2342  Here we have a well-established sunblockerspan...
2343  Let's hope this flight isn't Malaysian (lol). ...
2344  Here we have a northern speckled Rhododendron...
2345  This is the happiest dog you will ever see. Ve...
2346  Here is the Rand Paul of retrievers folks! He'...
2347  My oh my. This is a rare blond Canadian terrie...
2348  Here is a Siberian heavily armored polar bear ...
2349  This is an odd dog. Hard on the outside but lo...
2350  This is a truly beautiful English Wilson Staff...
2351  Here we have a 1949 1st generation vulpix. Enj...
2352  This is a purebred Piers Morgan. Loves to Netf...
2353  Here is a very happy pup. Big fan of well-main...
2354  This is a western brown Mitsubishi terrier. Up...
2355  Here we have a Japanese Irish Setter. Lost eye...

                                          expanded_urls  rating_numerator  \
2306  https://twitter.com/dog_rates/status/666835007...                10
2307  https://twitter.com/dog_rates/status/666826780...                12
2308  https://twitter.com/dog_rates/status/666817836...                 9
2309  https://twitter.com/dog_rates/status/666804364...                 8
2310  https://twitter.com/dog_rates/status/666786068...                 2
2311  https://twitter.com/dog_rates/status/666781792...                10
```

```
2312  https://twitter.com/dog_rates/status/666776908...                  5
2313  https://twitter.com/dog_rates/status/666739327...                 10
2314  https://twitter.com/dog_rates/status/666701168...                  8
2315  https://twitter.com/dog_rates/status/666691418...                  8
2316  https://twitter.com/dog_rates/status/666649482...                  4
2317  https://twitter.com/dog_rates/status/666644823...                  9
2318  https://twitter.com/dog_rates/status/666454714...                 10
2319  https://twitter.com/dog_rates/status/666447344...                  9
2320  https://twitter.com/dog_rates/status/666437273...                  7
2321  https://twitter.com/dog_rates/status/666435652...                 10
2322  https://twitter.com/dog_rates/status/666430724...                  6
2323  https://twitter.com/dog_rates/status/666428276...                  7
2324  https://twitter.com/dog_rates/status/666421158...                 12
2325  https://twitter.com/dog_rates/status/666418789...                 10
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...                  9
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  \
2306                  10      Unknown  None    None   None  None
2307                  10      Unknown  None    None   None  None
```

2308	10	Jeph	None	None	None	None
2309	10	Jockson	None	None	None	None
2310	10	Unknown	None	None	None	None
2311	10	Unknown	None	None	None	None
2312	10	Josep	None	None	None	None
2313	10	Lugan	None	None	None	None
2314	10	Unknown	None	None	None	None
2315	10	Christoper	None	None	None	None
2316	10	Unknown	None	None	None	None
2317	10	Jimothy	None	None	None	None
2318	10	Kreggory	None	None	None	None
2319	10	Scout	None	None	None	None
2320	10	Unknown	None	None	None	None
2321	10	Unknown	None	None	None	None
2322	10	Unknown	None	None	None	None
2323	10	Unknown	None	None	None	None
2324	10	Unknown	None	None	None	None
2325	10	Walter	None	None	None	None
2326	10	Unknown	None	None	None	None
2327	10	Unknown	None	None	None	None
2328	10	Unknown	None	None	None	None
2329	10	Unknown	None	None	None	None
2330	10	Unknown	None	None	None	None
2331	10	Unknown	None	None	None	None
2332	10	Unknown	None	None	None	None
2333	10	Unknown	None	None	None	None
2334	10	Unknown	None	None	None	None
2335	10	Unknown	None	None	None	None
2336	10	Unknown	None	None	None	None
2337	10	Unknown	None	None	None	None
2338	10	Unknown	None	None	None	None
2339	10	Unknown	None	None	None	None
2340	10	Unknown	None	None	None	None
2341	10	Unknown	None	None	None	None
2342	10	Unknown	None	None	None	None
2343	10	Unknown	None	None	None	None
2344	10	Unknown	None	None	None	None
2345	10	Unknown	None	None	None	None
2346	10	Unknown	None	None	None	None
2347	10	Unknown	None	None	None	None
2348	10	Unknown	None	None	None	None
2349	10	Unknown	None	None	None	None
2350	10	Unknown	None	None	None	None
2351	10	Unknown	None	None	None	None
2352	10	Unknown	None	None	None	None
2353	10	Unknown	None	None	None	None
2354	10	Unknown	None	None	None	None
2355	10	Unknown	None	None	None	None

```
                 url
2306  https://t.co/ZnEMHBsAs1
2307  https://t.co/V6ZzG45zzG
2308  https://t.co/8whlkYw3mO
2309  https://t.co/RdKbAOEpDK
2310  https://t.co/EIn5kElY1S
2311  https://t.co/uEvsGLOFHa
2312  https://t.co/XNGeDwrtYH
2313  https://t.co/tI3uFLDHBI
2314  https://t.co/uQbZJM2DQB
2315  https://t.co/bg4TqvvkuF
2316  https://t.co/vXMKrJC81s
2317  https://t.co/LEkZjZxESQ
2318  https://t.co/uPqPeXAcua
2319  https://t.co/kH6Ooka1HW
2320  https://t.co/7LtjBS0GPK
2321  https://t.co/GlFZPzqcEU
2322  https://t.co/yECbFrSArM
2323  https://t.co/NMQq6HIglK
2324  https://t.co/YMcrXC2Y6R
2325  https://t.co/qXpcwENTvn
2326  https://t.co/aMCTNWO94t
2327  https://t.co/LopTBkKa8h
2328  https://t.co/Dc7b0E8qFE
2329  https://t.co/uHXrPkUEyl
2330  https://t.co/XMD9CwjEnM
2331  https://t.co/tltQ5w9aUO
2332  https://t.co/VFKG1vxGjB
2333  https://t.co/QpRjllzWAL
2334  https://t.co/IIXis0zta0
2335  https://t.co/d9NcXFKwLv
2336  https://t.co/ws2AMaNwPW
2337  https://t.co/0yxApIikpk
2338  https://t.co/Asgdc6kuLX
2339  https://t.co/11LvqN4WLq
2340  https://t.co/Ri4nMTLq5C
2341  https://t.co/mnioXo3IfP
2342  https://t.co/3RU6x0vHB7
2343  https://t.co/Yk6GHE9tOY
2344  https://t.co/ZoL8kq2XFx
2345  https://t.co/RhUEAloehK
2346  https://t.co/pYAJkAe76p
2347  https://t.co/yWBqbrzy8O
2348  https://t.co/rdivxLiqEt
2349  https://t.co/v5A4vzSDdc
2350  https://t.co/fvIbQfHjIe
2351  https://t.co/4B7cOc1EDq
```

```
        2352    https://t.co/DWnyCjf2mx
        2353    https://t.co/y671yMhoiR
        2354    https://t.co/r7mOb2m0UI
        2355    https://t.co/BLDqew2Ijj
```

```
In [24]: # 7. Take a look at twitter_archive_enhanced data, we can find all
         # data are either none or same as column name, we can change it to
         # 0 as none or 1 as indicated by column name
         for col in ['doggo', 'floofer', 'pupper', 'puppo']:
             # set all these columns to zero
             twitter_archive_enhanced_clean[col] = 0
             # for those has same column name as entry, create a mask
             mask = twitter_archive_enhanced[col] == col
             # apply the mask
             twitter_archive_enhanced_clean.loc[mask, col] = 1
```

```
In [25]: # 8. Source should not include link
         def clean_source(twitter_archive_enhanced_clean):
             # extract the source text
             text = twitter_archive_enhanced_clean.source
             try:
                 # if '>' is found, means the source is in [left+1:-4]
                 left = text.find('>')
                 return text[left+1:-4]
             except:
                 return 'Unknown'

         twitter_archive_enhanced_clean['source'] = twitter_archive_enhanced_clean.apply(clean_
```

```
In [26]: twitter_archive_enhanced_clean.head(50)
```

```
Out[26]:               tweet_id            timestamp              source  \
         0   892420643555336193 2017-08-01 16:23:56  Twitter for iPhone
         1   892177421306343426 2017-08-01 00:17:27  Twitter for iPhone
         2   891815181378084864 2017-07-31 00:18:03  Twitter for iPhone
         3   891689557279858688 2017-07-30 15:58:51  Twitter for iPhone
         4   891327558926688256 2017-07-29 16:00:24  Twitter for iPhone
         5   891087950875897856 2017-07-29 00:08:17  Twitter for iPhone
         6   890971913173991426 2017-07-28 16:27:12  Twitter for iPhone
         7   890729181411237888 2017-07-28 00:22:40  Twitter for iPhone
         8   890609185150312448 2017-07-27 16:25:51  Twitter for iPhone
         9   890240255349198849 2017-07-26 15:59:51  Twitter for iPhone
         10  890006608113172480 2017-07-26 00:31:25  Twitter for iPhone
         11  889880896479866881 2017-07-25 16:11:53  Twitter for iPhone
         12  889665388333682689 2017-07-25 01:55:32  Twitter for iPhone
         13  889638837579907072 2017-07-25 00:10:02  Twitter for iPhone
         14  889531135344209921 2017-07-24 17:02:04  Twitter for iPhone
         15  889278841981685760 2017-07-24 00:19:32  Twitter for iPhone
         16  888917238123831296 2017-07-23 00:22:39  Twitter for iPhone
```

```
17  888804989199671297 2017-07-22 16:56:37  Twitter for iPhone
18  888554962724278272 2017-07-22 00:23:06  Twitter for iPhone
20  888078434458587136 2017-07-20 16:49:33  Twitter for iPhone
21  887705289381826560 2017-07-19 16:06:48  Twitter for iPhone
22  887517139158093824 2017-07-19 03:39:09  Twitter for iPhone
23  887473957103951883 2017-07-19 00:47:34  Twitter for iPhone
24  887343217045368832 2017-07-18 16:08:03  Twitter for iPhone
25  887101392804085760 2017-07-18 00:07:08  Twitter for iPhone
26  886983233522544640 2017-07-17 16:17:36  Twitter for iPhone
27  886736880519319552 2017-07-16 23:58:41  Twitter for iPhone
28  886680336477933568 2017-07-16 20:14:00  Twitter for iPhone
29  886366144734445568 2017-07-15 23:25:31  Twitter for iPhone
31  886258384151887873 2017-07-15 16:17:19  Twitter for iPhone
33  885984800019947520 2017-07-14 22:10:11  Twitter for iPhone
34  885528943205470208 2017-07-13 15:58:47  Twitter for iPhone
35  885518971528720385 2017-07-13 15:19:09  Twitter for iPhone
37  885167619883638784 2017-07-12 16:03:00  Twitter for iPhone
38  884925521741709313 2017-07-12 00:01:00  Twitter for iPhone
39  884876753390489601 2017-07-11 20:47:12  Twitter for iPhone
40  884562892145688576 2017-07-11 00:00:02  Twitter for iPhone
41  884441805382717440 2017-07-10 15:58:53  Twitter for iPhone
42  884247878851493888 2017-07-10 03:08:17  Twitter for iPhone
43  884162670584377345 2017-07-09 21:29:42  Twitter for iPhone
44  883838122936631299 2017-07-09 00:00:04  Twitter for iPhone
45  883482846933004288 2017-07-08 00:28:19  Twitter for iPhone
46  883360690899218434 2017-07-07 16:22:55  Twitter for iPhone
47  883117836046086144 2017-07-07 00:17:54  Twitter for iPhone
48  882992080364220416 2017-07-06 15:58:11  Twitter for iPhone
49  882762694511734784 2017-07-06 00:46:41  Twitter for iPhone
50  882627270321602560 2017-07-05 15:48:34  Twitter for iPhone
51  882268110199369728 2017-07-04 16:01:23  Twitter for iPhone
52  882045870035918850 2017-07-04 01:18:17  Twitter for iPhone
53  881906580714921986 2017-07-03 16:04:48  Twitter for iPhone

                                                 text  \
0   This is Phineas. He's a mystical boy. Only eve...
1   This is Tilly. She's just checking pup on you...
2   This is Archie. He is a rare Norwegian Pouncin...
3   This is Darla. She commenced a snooze mid meal...
4   This is Franklin. He would like you to stop ca...
5   Here we have a majestic great white breaching ...
6   Meet Jax. He enjoys ice cream so much he gets ...
7   When you watch your owner call another dog a g...
8   This is Zoey. She doesn't want to be one of th...
9   This is Cassie. She is a college pup. Studying...
10  This is Koda. He is a South Australian decksha...
11  This is Bruno. He is a service shark. Only get...
12  Here's a puppo that seems to be on the fence a...
```

```
13  This is Ted. He does his best. Sometimes that'...
14  This is Stuart. He's sporting his favorite fan...
15  This is Oliver. You're witnessing one of his m...
16  This is Jim. He found a fren. Taught him how t...
17  This is Zeke. He has a new stick. Very proud o...
18  This is Ralphus. He's powering up. Attempting ...
20  This is Gerald. He was just told he didn't get...
21  This is Jeffrey. He has a monopoly on the pool...
22  I've yet to rate a Venezuelan Hover Wiener. Th...
23  This is Canela. She attempted some fancy porch...
24  You may not have known you needed to see this ...
25  This... is a Jubilant Antarctic House Bear. We...
26  This is Maya. She's very shy. Rarely leaves he...
27  This is Mingus. He's a wonderful father to his...
28  This is Derek. He's late for a dog meeting. 13...
29  This is Roscoe. Another pupper fallen victim t...
31  This is Waffles. His doggles are pupside down...
33  Viewer discretion advised. This is Jimbo. He w...
34  This is Maisey. She fell asleep mid-excavation...
35  I have a new hero and his name is Howard. 14/1...
37  Here we have a corgi undercover as a malamute...
38  This is Earl. He found a hat. Nervous about wh...
39  This is Lola. It's her first time outside. Mus...
40  This is Kevin. He's just so happy. 13/10 what ...
41  I present to you, Pup in Hat. Pup in Hat is gr...
42  OMG HE DIDN'T MEAN TO HE WAS JUST TRYING A LIT...
43  Meet Yogi. He doesn't have any important dog m...
44  This is Noah. He can't believe someone made th...
45  This is Bella. She hopes her smile made you sm...
46  Meet Grizzwald. He may be the floofiest floofe...
47  Please only send dogs. We don't rate mechanics...
48  This is Rusty. He wasn't ready for the first p...
49  This is Gus. He's quite the cheeky pupper. Alr...
50  This is Stanley. He has his first swim lesson ...
51  This is Alfy. You're witnessing his first wate...
52  This is Koko. Her owner, inspired by Barney, r...
53  This is Rey. He's a Benebop Cumberfloof. 12/10...

                                        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
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
31  https://twitter.com/dog_rates/status/886258384...                13
33  https://twitter.com/dog_rates/status/885984800...                12
34  https://twitter.com/dog_rates/status/885528943...                13
35  https://twitter.com/4bonds2carbon/status/88551...                14
37  https://twitter.com/dog_rates/status/885167619...                13
38  https://twitter.com/dog_rates/status/884925521...                12
39  https://twitter.com/dog_rates/status/884876753...                13
40  https://twitter.com/dog_rates/status/884562892...                13
41  https://twitter.com/dog_rates/status/884441805...                14
42  https://twitter.com/kaijohnson_19/status/88396...                13
43  https://twitter.com/dog_rates/status/884162670...                12
44  https://twitter.com/dog_rates/status/883838122...                12
45  https://twitter.com/dog_rates/status/883482846...                 0
46  https://twitter.com/dog_rates/status/883360690...                13
47  https://twitter.com/dog_rates/status/883117836...                13
48  https://twitter.com/dog_rates/status/882992080...                13
49  https://twitter.com/dog_rates/status/882762694...                12
50  https://twitter.com/dog_rates/status/882627270...                13
51  https://twitter.com/dog_rates/status/882268110...                13
52  https://twitter.com/dog_rates/status/882045870...                13
53  https://twitter.com/dog_rates/status/881906580...                12

    rating_denominator       name  doggo  floofer  pupper  puppo  \
0                   10    Phineas      0        0       0      0
1                   10      Tilly      0        0       0      0
2                   10     Archie      0        0       0      0
3                   10      Darla      0        0       0      0
4                   10   Franklin      0        0       0      0
```

```
5                    10    Unknown      0        0       0      0
6                    10        Jax      0        0       0      0
7                    10    Unknown      0        0       0      0
8                    10       Zoey      0        0       0      0
9                    10     Cassie      1        0       0      0
10                   10       Koda      0        0       0      0
11                   10      Bruno      0        0       0      0
12                   10    Unknown      0        0       0      1
13                   10        Ted      0        0       0      0
14                   10     Stuart      0        0       0      1
15                   10     Oliver      0        0       0      0
16                   10        Jim      0        0       0      0
17                   10       Zeke      0        0       0      0
18                   10    Ralphus      0        0       0      0
20                   10     Gerald      0        0       0      0
21                   10    Jeffrey      0        0       0      0
22                   10    Unknown      0        0       0      0
23                   10     Canela      0        0       0      0
24                   10    Unknown      0        0       0      0
25                   10    Unknown      0        0       0      0
26                   10       Maya      0        0       0      0
27                   10     Mingus      0        0       0      0
28                   10      Derek      0        0       0      0
29                   10     Roscoe      0        0       1      0
31                   10    Waffles      0        0       0      0
33                   10      Jimbo      0        0       0      0
34                   10     Maisey      0        0       0      0
35                   10    Unknown      0        0       0      0
37                   10    Unknown      0        0       0      0
38                   10       Earl      0        0       0      0
39                   10       Lola      0        0       0      0
40                   10      Kevin      0        0       0      0
41                   10    Unknown      0        0       0      0
42                   10    Unknown      0        0       0      0
43                   10       Yogi      1        0       0      0
44                   10       Noah      0        0       0      0
45                    0      Bella      0        0       0      0
46                   10  Grizzwald      0        1       0      0
47                   10    Unknown      0        0       0      0
48                   10      Rusty      0        0       0      0
49                   10        Gus      0        0       1      0
50                   10    Stanley      0        0       0      0
51                   10       Alfy      0        0       0      0
52                   10       Koko      0        0       0      0
53                   10        Rey      0        0       0      0

                        url
0   https://t.co/MgUWQ76dJU
```

1	https://t.co/0Xxu71qeIV
2	https://t.co/wUnZnhtVJB
3	https://t.co/tD36da7qLQ
4	https://t.co/AtUZn91f7f
5	https://t.co/kQ04fDDRmh
6	https://t.co/tVJBRMnhxl
7	https://t.co/v0nONBcwxq
8	https://t.co/9TwLuAGH0b
9	https://t.co/t1bfwz5S2A
10	https://t.co/dVPW0BOMme
11	https://t.co/u1XPQMl29g
12	https://t.co/BxvuXk0UCm
13	https://t.co/f8dEDcrKSR
14	https://t.co/y70o6h3isq
15	https://t.co/WpHvrQedPb
16	https://t.co/chxruIOUJN
17	https://t.co/HTQ77yNQ5K
18	https://t.co/YnYAFCTTiK
20	https://t.co/DK7iDPfuRX
21	https://t.co/PhrUk20Q64
22	https://t.co/20VrLAA8ba
23	https://t.co/cLyzpcUcMX
24	https://t.co/WZqNqygEyV
25	https://t.co/4Ad1jzJSdp
26	https://t.co/I6oNy0CgiT
27	https://t.co/ISvKOSkd5b
28	https://t.co/BCoWue0abA
29	https://t.co/RGE08MIJox
31	https://t.co/xZDA9Qsq1O
33	https://t.co/BuveP0uMF1
34	https://t.co/tp1kQ8i9JF
35	https://t.co/gzLHboL7Sk
37	https://t.co/44ItaMubBf
38	https://t.co/MYJvdlNRVa
39	https://t.co/74TKAUsLkO
40	https://t.co/1r4MFCbCX5
41	https://t.co/vvBOcC2VdC
42	https://t.co/uF3pQ8Wubj
43	https://t.co/YSI0OBzTBZ
44	https://t.co/V85xujjDDY
45	https://t.co/qjrljjt948
46	https://t.co/rf661IFEYP
47	https://t.co/Se5fZ9wp5E
48	https://t.co/tyER0KpdXj
49	https://t.co/D43I96SlVu
50	https://t.co/Nx52PGwH94
51	https://t.co/fYP5RlutfA
52	https://t.co/zeDpnsKX7w

```
53  https://t.co/503CgWbhxQ
```

1.3.2 Tidyness

```
In [27]: # 1. Unuseful dog tag doggo floofer pupper puppo, remove those columns
         twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.drop(["doggo", "floofe
         twitter_archive_enhanced_clean.head()
```

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

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

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

            rating_denominator      name                      url
         0                  10   Phineas  https://t.co/MgUWQ76dJU
         1                  10     Tilly  https://t.co/0Xxu71qeIV
         2                  10    Archie  https://t.co/wUnZnhtVJB
         3                  10     Darla  https://t.co/tD36da7qLQ
         4                  10  Franklin  https://t.co/AtUZn91f7f
```

```
In [28]: # 2. Clean user id in tweets data
         tweets_clean = tweets.copy()
```

```
In [29]: from pandas.io.json import json_normalize
         # use json_normalize to extract a tidy form of tweets_clean['user']
         user_info = json_normalize(tweets_clean['user'])
         user_info.columns
```

```
Out[29]: Index(['contributors_enabled', 'created_at', 'default_profile',
                'default_profile_image', 'description', 'entities.description.urls',
                'entities.url.urls', 'favourites_count', 'follow_request_sent',
                'followers_count', 'following', 'friends_count', 'geo_enabled',
                'has_extended_profile', 'id', 'id_str', 'is_translation_enabled',
                'is_translator', 'lang', 'listed_count', 'location', 'name',
                'notifications', 'profile_background_color',
                'profile_background_image_url', 'profile_background_image_url_https',
                'profile_background_tile', 'profile_banner_url', 'profile_image_url',
                'profile_image_url_https', 'profile_link_color',
                'profile_sidebar_border_color', 'profile_sidebar_fill_color',
                'profile_text_color', 'profile_use_background_image', 'protected',
                'screen_name', 'statuses_count', 'time_zone', 'translator_type', 'url',
                'utc_offset', 'verified'],
               dtype='object')
```

```
In [30]: # update tweets_clean['user'] to id only
         tweets_clean['user'] = user_info['id']
         tweets_clean.head()
```

```
Out[30]:          user  favorite_count  retweet_count
         0  4196983835           39334           8786
         1  4196983835           33671           6439
         2  4196983835           25363           4270
         3  4196983835           42708           8877
         4  4196983835           40850           9654
```

1.4 4. Storing, Analyzing, and Visualizing Data for this Project

Store twitter_archive_enhanced_clean to csv

```
In [31]: twitter_archive_enhanced_clean.to_csv('twitter_archive_master.csv')
         tweets_clean.to_csv('tweets_stats.csv')
```

Insights

```
In [32]: twitter_archive_enhanced_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2097 entries, 0 to 2355
Data columns (total 9 columns):
tweet_id              2097 non-null int64
timestamp             2097 non-null datetime64[ns]
source                2097 non-null object
text                  2097 non-null object
expanded_urls         2094 non-null object
rating_numerator      2097 non-null int64
rating_denominator    2097 non-null int64
name                  2097 non-null object
url                   2094 non-null object
dtypes: datetime64[ns](1), int64(3), object(5)
memory usage: 163.8+ KB
```

```
In [33]: twitter_archive_enhanced_clean.timestamp.max
```

```
Out[33]: <bound method Series.max of 0      2017-08-01 16:23:56
         1      2017-08-01 00:17:27
         2      2017-07-31 00:18:03
         3      2017-07-30 15:58:51
         4      2017-07-29 16:00:24
         5      2017-07-29 00:08:17
         6      2017-07-28 16:27:12
         7      2017-07-28 00:22:40
         8      2017-07-27 16:25:51
         9      2017-07-26 15:59:51
         10     2017-07-26 00:31:25
         11     2017-07-25 16:11:53
         12     2017-07-25 01:55:32
         13     2017-07-25 00:10:02
         14     2017-07-24 17:02:04
         15     2017-07-24 00:19:32
         16     2017-07-23 00:22:39
         17     2017-07-22 16:56:37
         18     2017-07-22 00:23:06
         20     2017-07-20 16:49:33
         21     2017-07-19 16:06:48
         22     2017-07-19 03:39:09
         23     2017-07-19 00:47:34
         24     2017-07-18 16:08:03
         25     2017-07-18 00:07:08
         26     2017-07-17 16:17:36
         27     2017-07-16 23:58:41
         28     2017-07-16 20:14:00
         29     2017-07-15 23:25:31
         31     2017-07-15 16:17:19
                        ...
         2326   2015-11-17 00:24:19
         2327   2015-11-17 00:06:54
         2328   2015-11-16 23:23:41
         2329   2015-11-16 21:54:18
         2330   2015-11-16 21:10:36
         2331   2015-11-16 20:32:58
         2332   2015-11-16 20:01:42
         2333   2015-11-16 19:31:45
         2334   2015-11-16 16:37:02
         2335   2015-11-16 16:11:11
         2336   2015-11-16 15:14:19
         2337   2015-11-16 14:57:41
         2338   2015-11-16 04:02:55
         2339   2015-11-16 03:55:04
         2340   2015-11-16 03:44:34
         2341   2015-11-16 03:22:39
         2342   2015-11-16 02:38:37
```

```
2343   2015-11-16 01:59:36
2344   2015-11-16 01:52:02
2345   2015-11-16 01:22:45
2346   2015-11-16 01:01:59
2347   2015-11-16 00:55:59
2348   2015-11-16 00:49:46
2349   2015-11-16 00:35:11
2350   2015-11-16 00:30:50
2351   2015-11-16 00:24:50
2352   2015-11-16 00:04:52
2353   2015-11-15 23:21:54
2354   2015-11-15 23:05:30
2355   2015-11-15 22:32:08
Name: timestamp, Length: 2097, dtype: datetime64[ns]>
```

Insights 1: There are 2097 original tweets from WeRateDogs after we wrangle the data

```
In [34]: tweets_clean.describe()
```

```
Out[34]:                user  favorite_count  retweet_count
         count  2.349000e+03     2349.000000    2349.000000
         mean   4.196984e+09     8109.720732    3097.263942
         std    0.000000e+00    12051.729881    5133.834333
         min    4.196984e+09        0.000000       0.000000
         25%    4.196984e+09     1415.000000     615.000000
         50%    4.196984e+09     3585.000000    1445.000000
         75%    4.196984e+09    10100.000000    3606.000000
         max    4.196984e+09   131820.000000   78754.000000
```

Insights 2: The maximum favorite count is 131820 and retweet count is 78754, and minimum for both is 0.

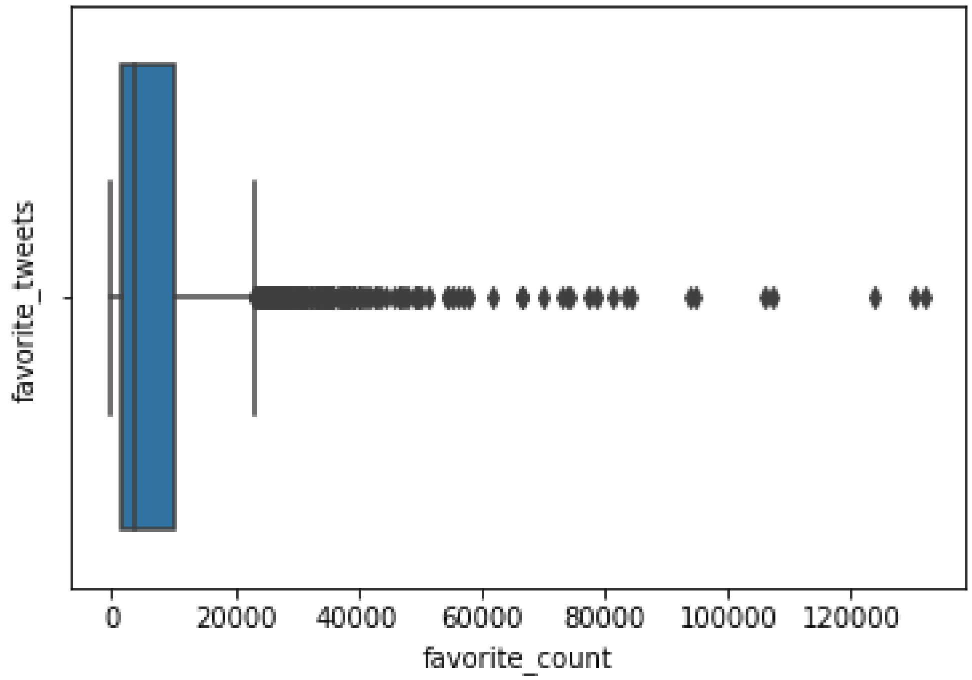
Insights 3: The mean of favorite count is 8110 and mean of retweet count is 3097.

Visualization

```
In [38]: # Use seaborn package to draw boxplot of tweets data to visualize it.
         import seaborn as sns
         import matplotlib as plt
         %matplotlib inline
```

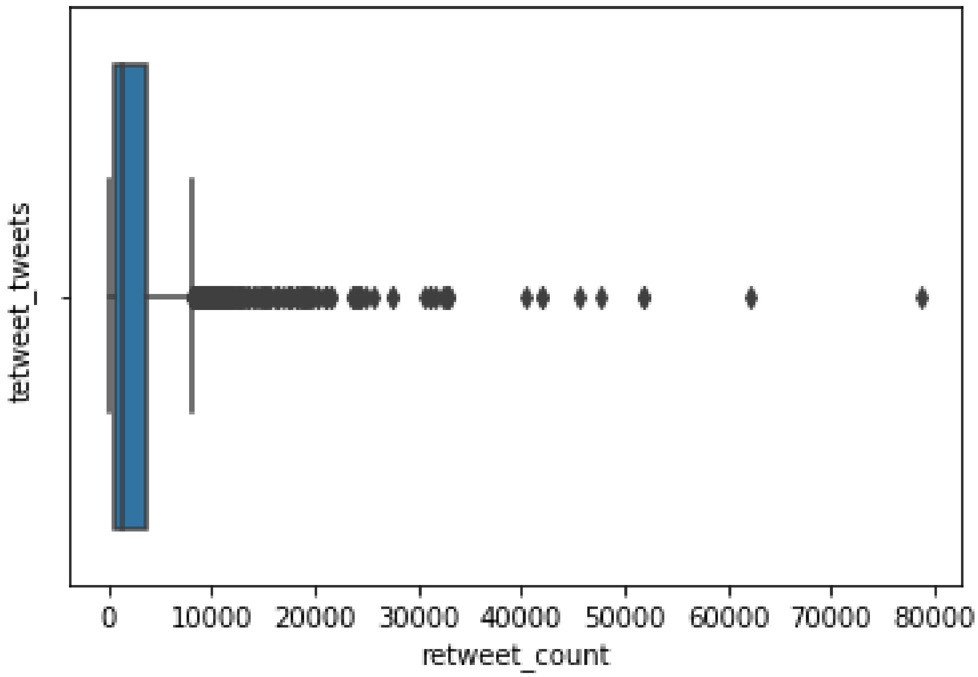
```
In [40]: ax1 = sns.boxplot(x = tweets_clean['favorite_count'])
         ax1.set(ylabel = 'favorite_tweets')
```

```
Out[40]: [Text(0,0.5,'favorite_tweets')]
```



```
In [41]: ax2 = sns.boxplot(x = tweets_clean['retweet_count'])
         ax2.set(ylabel = 'tetweet_tweets')

Out[41]: [Text(0,0.5,'tetweet_tweets')]
```


tetweet_tweets
retweet_count