wrangle_act

September 5, 2022

1 Project: Wrangling and Analyze Data

1.0.1 Table Of Contents:

1. Data Gathering
2. Assessing Data
3. Cleaning Data
4. Storing Data
5. Analysis and Visualization

1.1 Data Gathering

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

```
In [1]: import pandas as pd
        import numpy as np
        import tweepy
        import matplotlib.pyplot as plt
        from functools import reduce
        %matplotlib inline
```

```
In [2]: twitter_df = pd.read_csv('twitter_archive_enhanced.csv', sep= ',')
```

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

```
In [3]: import requests
        import os
        url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictic
        response = requests.get(url)
        response
```

```
Out[3]: <Response [200]>
```

```
In [4]: # Making a directory if it doesn't already exist
        folder_name = 'image_predictions.tsv'
        if not os.path.exists(folder_name):
            os.makedirs(folder_name)
```

```
In [5]: with open(os.path.join(folder_name, url.split('/')[-1]), mode = 'wb') as file:
            file.write(response.content)
```

Confirming Directory just created

```
In [8]: os.listdir(folder_name)
```

```
Out[8]: ['image-predictions.tsv']
```

```
In [6]: #using image predictions dataframe as df_ip
        image_df = pd.read_csv('image-predictions.tsv', sep = '\t')
```

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

```
In [13]: import tweepy
         import json
         from timeit import default_timer as timer
         from tweepy import OAuthHandler

         CONSUMER_KEY = 'confidential'
         CONSUMER_SECRET = 'confidential'
         OAUTH_TOKEN = 'confidential'
         OAUTH_TOKEN_SECRET = 'confidential'

         auth = tweepy.OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET)
         auth.set_access_token(OAUTH_TOKEN, OAUTH_TOKEN_SECRET)
         api = tweepy.API(auth_handler=auth, wait_on_rate_limit= True, wait_on_rate_limit_notify
```

```
In [14]: # tweet_ids = twitter_df.tweet_id.values
         # len(tweet_ids)

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

```
#             pass
# end = timer()
# print(end - start)
# print(fails_dict)
```

In [15]:
```
#Reading the Json files into texts and reading each tweet_id line by line
df_list = []

with open('tweet_json.txt', 'r', encoding= 'utf-8') as tweet_data:
    for line in tweet_data:
        data = (json.loads(line))
        tweet_id = data['id']
        created_at = data['created_at']
        favorite_count = data['favorite_count']
        retweet_count = data['retweet_count']

        # Append to list of dictionaries
        df_list.append({'tweet_id' : tweet_id,

                'date_created' : created_at,

                'favorite_count' : favorite_count,

                'retweet_count' : retweet_count })


API_df = pd.DataFrame(df_list, columns = ['tweet_id', 'date_created', 'favorite_count',
```

1.2 Assessing Data

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

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

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

In [16]:
```
# Assessing the three different datasets visually
# WeRateDogs Twitter Archive Data
twitter_df.head()
```

Out[16]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	retweeted_status_id	retweeted_status_user_id	retweeted_status_timestamp	expanded_urls	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" r...	This is Phineas. He's a mystical boy. Only eve...	NaN	NaN	NaN	https://twitter.com/dog_rates/status/892420643...	13	10	Phineas	None	None	None	None
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" r...	This is Tilly. She's just checking pup on you...	NaN	NaN	NaN	https://twitter.com/dog_rates/status/892177421...	13	10	Tilly	None	None	None	None
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" r...	This is Archie. He is a rare Norwegian Pouncin...	NaN	NaN	NaN	https://twitter.com/dog_rates/status/891815181...	12	10	Archie	None	None	None	None
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="http://twitter.com/download/iphone" r...	This is Darla. She commenced a snooze mid meal...	NaN	NaN	NaN	https://twitter.com/dog_rates/status/891689557...	13	10	Darla	None	None	None	None
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href="http://twitter.com/download/iphone" r...	This is Franklin. He would like you to stop ca...	NaN	NaN	NaN	https://twitter.com/dog_rates/status/891327558...	12	10	Franklin	None	None	None	None

```
In [17]: #checking for duplicated tweets
         twitter_df.duplicated().sum()

Out[17]: 0

In [18]: #checking programmatically
         twitter_df.info()

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


In [19]: # Assessing TWeet image prediction dataset visually
         image_df.head()

Out[19]:              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 [20]: image_df.duplicated().sum()
```

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

```
In [21]: # Assessing Tweet image prediction programmatically
         image_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null int64
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

```
In [22]: # Assessing additional data gotten from querying twitter API visually
         API_df
```

```
Out[22]:                tweet_id                   date_created  favorite_count  \
         0     892420643555336193  Tue Aug 01 16:23:56 +0000 2017           33697
         1     892177421306343426  Tue Aug 01 00:17:27 +0000 2017           29222
         2     891815181378084864  Mon Jul 31 00:18:03 +0000 2017           21978
         3     891689557279858688  Sun Jul 30 15:58:51 +0000 2017           36791
         4     891327558926688256  Sat Jul 29 16:00:24 +0000 2017           35182
         5     891087950875897856  Sat Jul 29 00:08:17 +0000 2017           17749
         6     890971913173991426  Fri Jul 28 16:27:12 +0000 2017           10331
         7     890729181411237888  Fri Jul 28 00:22:40 +0000 2017           56670
         8     890609185150312448  Thu Jul 27 16:25:51 +0000 2017           24427
         9     890240255349198849  Wed Jul 26 15:59:51 +0000 2017           27848
         10    890006608113172480  Wed Jul 26 00:31:25 +0000 2017           26947
         11    889880896479866881  Tue Jul 25 16:11:53 +0000 2017           24488
```

12	889665388333682689	Tue Jul 25 01:55:32 +0000 2017	41878
13	889638837579907072	Tue Jul 25 00:10:02 +0000 2017	23580
14	889531135344209921	Mon Jul 24 17:02:04 +0000 2017	13313
15	889278841981685760	Mon Jul 24 00:19:32 +0000 2017	22026
16	888917238123831296	Sun Jul 23 00:22:39 +0000 2017	25533
17	888804989199671297	Sat Jul 22 16:56:37 +0000 2017	22392
18	888554962724278272	Sat Jul 22 00:23:06 +0000 2017	17252
19	888078434458587136	Thu Jul 20 16:49:33 +0000 2017	19101
20	887705289381826560	Wed Jul 19 16:06:48 +0000 2017	26530
21	887517139158093824	Wed Jul 19 03:39:09 +0000 2017	40548
22	887473957103951883	Wed Jul 19 00:47:34 +0000 2017	59978
23	887343217045368832	Tue Jul 18 16:08:03 +0000 2017	29490
24	887101392804085760	Tue Jul 18 00:07:08 +0000 2017	26897
25	886983233522544640	Mon Jul 17 16:17:36 +0000 2017	30239
26	886736880519319552	Sun Jul 16 23:58:41 +0000 2017	10455
27	886680336477933568	Sun Jul 16 20:14:00 +0000 2017	19655
28	886366144734445568	Sat Jul 15 23:25:31 +0000 2017	18489
29	886267009285017600	Sat Jul 15 16:51:35 +0000 2017	105
...	...	...	...
2297	666411507551481857	Tue Nov 17 00:24:19 +0000 2015	371
2298	666407126856765440	Tue Nov 17 00:06:54 +0000 2015	93
2299	666396247373291520	Mon Nov 16 23:23:41 +0000 2015	147
2300	666373753744588802	Mon Nov 16 21:54:18 +0000 2015	162
2301	666362758909284353	Mon Nov 16 21:10:36 +0000 2015	649
2302	666353288456101888	Mon Nov 16 20:32:58 +0000 2015	179
2303	666345417576210432	Mon Nov 16 20:01:42 +0000 2015	242
2304	666337882303524864	Mon Nov 16 19:31:45 +0000 2015	168
2305	666293911632134144	Mon Nov 16 16:37:02 +0000 2015	425
2306	666287406224695296	Mon Nov 16 16:11:11 +0000 2015	123
2307	666273097616637952	Mon Nov 16 15:14:19 +0000 2015	151
2308	666268910803644416	Mon Nov 16 14:57:41 +0000 2015	99
2309	666104133288665088	Mon Nov 16 04:02:55 +0000 2015	12844
2310	666102155909144576	Mon Nov 16 03:55:04 +0000 2015	66
2311	666099513787052032	Mon Nov 16 03:44:34 +0000 2015	134
2312	666094000022159362	Mon Nov 16 03:22:39 +0000 2015	142
2313	666082916733198337	Mon Nov 16 02:38:37 +0000 2015	92
2314	666073100786774016	Mon Nov 16 01:59:36 +0000 2015	273
2315	666071193221509120	Mon Nov 16 01:52:02 +0000 2015	127
2316	666063827256086533	Mon Nov 16 01:22:45 +0000 2015	396
2317	666058600524156928	Mon Nov 16 01:01:59 +0000 2015	99
2318	666057090499244032	Mon Nov 16 00:55:59 +0000 2015	247
2319	666055525042405380	Mon Nov 16 00:49:46 +0000 2015	367
2320	666051853826850816	Mon Nov 16 00:35:11 +0000 2015	1023
2321	666050758794694657	Mon Nov 16 00:30:50 +0000 2015	115
2322	666049248165822465	Mon Nov 16 00:24:50 +0000 2015	88
2323	666044226329800704	Mon Nov 16 00:04:52 +0000 2015	246
2324	666033412701032449	Sun Nov 15 23:21:54 +0000 2015	100
2325	666029285002620928	Sun Nov 15 23:05:30 +0000 2015	112

```
2326  666020888022790149  Sun Nov 15 22:32:08 +0000 2015            2283

      retweet_count
0              6969
1              5272
2              3464
3              7191
4              7717
5              2586
6              1647
7             15679
8              3604
9              6068
10             6111
11             4141
12             8302
13             3700
14             1874
15             4418
16             3742
17             3513
18             2863
19             2885
20             4516
21             9810
22            14952
23             8775
24             4969
25             6287
26             2615
27             3720
28             2613
29                4
...             ...
2297            261
2298             30
2299             68
2300             73
2301            467
2302             56
2303            121
2304             79
2305            288
2306             55
2307             66
2308             38
2309           5454
2310             11
```

```
2311             53
2312             63
2313             37
2314            130
2315             51
2316            180
2317             47
2318            111
2319            196
2320            699
2321             50
2322             36
2323            115
2324             36
2325             39
2326            419

[2327 rows x 4 columns]
```

```
In [23]: #checking for any duplicate values
         API_df.duplicated().sum()
```

```
Out[23]: 0
```

```
In [24]: #Assessing additional data gotten from querying twitter API visually programmatically
         API_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2327 entries, 0 to 2326
Data columns (total 4 columns):
tweet_id          2327 non-null int64
date_created      2327 non-null object
favorite_count    2327 non-null int64
retweet_count     2327 non-null int64
dtypes: int64(3), object(1)
memory usage: 72.8+ KB
```

1.2.1 Quality issues

1.column 'retweeted_status_user_id' and name is not descriptive enough

2.There are retweets in the dataframe.

3.columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_timestamp, source, text, expanded_url, 'jpg_url', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog', 'date_created') are of little or no importance.

4.False predictions in the image prediction dataframe indicates predictions contain animals other than dogs

5.p1, p1_conf, p1_dog are not descriptive enough

6.Datas contained in the 'timestamp' not in the right format

7.Name column in the twitter_archive dataset contains incosistent data
8.Dog names are inconsistent

1.2.2 Tidiness issues

1.The dog "stage" (i.e. doggo, floofer, pupper, and puppo) should be one column
2.Rating should be a single column instead of two

1.3 Cleaning Data

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

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

```
In [25]: twitter_df_clean = twitter_df.copy(deep = True)
         API_df_clean = API_df.copy(deep = True)
         image_df_clean = image_df.copy(deep =  True)

In [26]: twitter_df_clean.head()
```

Out[26]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	retweeted_status_id
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" r...	This is Phineas. He's a mystical boy. Only eve...	NaN
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" r...	This is Tilly. She's just checking pup on you...	NaN
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" r...	This is Archie. He is a rare Norwegian Pouncin...	NaN
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="http://twitter.com/download/iphone" r...	This is Darla. She commenced a snooze mid meal...	NaN
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href="http://twitter.com/download/iphone" r...	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
```

1.3.1 Issue #1: columns 'retweeted_status_user_id' and 'name' is not descriptive enough

Define: column "retweeted_status_user_id" and "name" is renamed to "retweet_id" and "dog_name" respectivly to make it descriptive.

Code

```
In [27]: twitter_df_clean  = twitter_df_clean.rename(columns={"retweeted_status_user_id" : "retw
```

Test

```
In [28]: twitter_df_clean
```

```
Out[28]:                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

        retweet_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	dog_name	doggo	floofer	pupper	puppo
0	10	Phineas	None	None	None	None
1	10	Tilly	None	None	None	None
2	10	Archie	None	None	None	None
3	10	Darla	None	None	None	None
4	10	Franklin	None	None	None	None
5	10	None	None	None	None	None
6	10	Jax	None	None	None	None
7	10	None	None	None	None	None
8	10	Zoey	None	None	None	None
9	10	Cassie	doggo	None	None	None
10	10	Koda	None	None	None	None
11	10	Bruno	None	None	None	None
12	10	None	None	None	None	puppo
13	10	Ted	None	None	None	None
14	10	Stuart	None	None	None	puppo
15	10	Oliver	None	None	None	None
16	10	Jim	None	None	None	None
17	10	Zeke	None	None	None	None
18	10	Ralphus	None	None	None	None

```
19                   10    Canela   None    None    None   None
20                   10    Gerald   None    None    None   None
21                   10   Jeffrey   None    None    None   None
22                   10      such   None    None    None   None
23                   10    Canela   None    None    None   None
24                   10      None   None    None    None   None
25                   10      None   None    None    None   None
26                   10      Maya   None    None    None   None
27                   10    Mingus   None    None    None   None
28                   10     Derek   None    None    None   None
29                   10    Roscoe   None    None  pupper   None
...                 ...       ...    ...     ...     ...    ...
2326                 10     quite   None    None    None   None
2327                 10         a   None    None    None   None
2328                 10      None   None    None    None   None
2329                 10      None   None    None    None   None
2330                 10      None   None    None    None   None
2331                 10      None   None    None    None   None
2332                 10      None   None    None    None   None
2333                 10        an   None    None    None   None
2334                 10         a   None    None    None   None
2335                  2        an   None    None    None   None
2336                 10      None   None    None    None   None
2337                 10      None   None    None    None   None
2338                 10      None   None    None    None   None
2339                 10      None   None    None    None   None
2340                 10      None   None    None    None   None
2341                 10      None   None    None    None   None
2342                 10      None   None    None    None   None
2343                 10      None   None    None    None   None
2344                 10      None   None    None    None   None
2345                 10       the   None    None    None   None
2346                 10       the   None    None    None   None
2347                 10         a   None    None    None   None
2348                 10         a   None    None    None   None
2349                 10        an   None    None    None   None
2350                 10         a   None    None    None   None
2351                 10      None   None    None    None   None
2352                 10         a   None    None    None   None
2353                 10         a   None    None    None   None
2354                 10         a   None    None    None   None
2355                 10      None   None    None    None   None

[2356 rows x 17 columns]
```

1.3.2 Issue #2: There are retweets in the dataframe.

Define: There are retweets in the dataset, and we do not need them as we only need original tweets. Hence I mask the retweet column to only contain rows that are null for the retweet column

Code

```
In [29]: twitter_df_clean =twitter_df_clean[twitter_df_clean['retweet_id'].isna()]
```

Test

```
In [30]: twitter_df_clean
```

Out[30]:

	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	
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	
30	886267009285017600	8.862664e+17	2.281182e+09	
...	...	...	...	
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
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
30    2017-07-15 16:51:35 +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...
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...
30    <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
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
30    @NonWhiteHat @MayhewMayhem omg hello tanner yo...                   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

      retweet_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
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
30           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
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
30	NaN	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	dog_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
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
30	10	None	None	None	None	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

[2175 rows x 17 columns]
```

```
In [31]: twitter_df_clean.columns
```

```
Out[31]: Index(['tweet_id', 'in_reply_to_status_id', 'in_reply_to_user_id', 'timestamp',
                'source', 'text', 'retweeted_status_id', 'retweet_id',
                'retweeted_status_timestamp', 'expanded_urls', 'rating_numerator',
                'rating_denominator', 'dog_name', 'doggo', 'floofer', 'pupper',
                'puppo'],
               dtype='object')
```

1.3.3 Issue #3: There are some irrelevant columns

Define: The columns: 'in_reply_to_status_id', 'in_reply_to_user_id', 'source', 'text', 'retweeted_status_id', 'retweeted_status_timestamp', 'expanded_urls' are not useful for our analysis, hence they are dropped

Code

```
In [32]: import numpy as np
         idx = np.r_ [1:3, 4:7, 8:10]
         twitter_df_clean.drop(twitter_df_clean.columns[idx], axis= 1, inplace= True)
```

```
/opt/conda/lib/python3.6/site-packages/pandas/core/frame.py:3697: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#
  errors=errors)
```

Test

```
In [33]: twitter_df_clean.head()
```

```
Out[33]:              tweet_id                  timestamp  retweet_id  \
         0  892420643555336193  2017-08-01 16:23:56 +0000         NaN
         1  892177421306343426  2017-08-01 00:17:27 +0000         NaN
         2  891815181378084864  2017-07-31 00:18:03 +0000         NaN
         3  891689557279858688  2017-07-30 15:58:51 +0000         NaN
         4  891327558926688256  2017-07-29 16:00:24 +0000         NaN

            rating_numerator  rating_denominator  dog_name doggo floofer pupper puppo
         0                13                  10   Phineas  None    None   None  None
         1                13                  10     Tilly  None    None   None  None
         2                12                  10    Archie  None    None   None  None
         3                13                  10     Darla  None    None   None  None
         4                12                  10  Franklin  None    None   None  None
```

```
In [34]: twitter_df_clean.columns
```

```
Out[34]: Index(['tweet_id', 'timestamp', 'retweet_id', 'rating_numerator',
                'rating_denominator', 'dog_name', 'doggo', 'floofer', 'pupper',
                'puppo'],
               dtype='object')
```

The "date_created" column will be deleted in the twitter API dataset.

Code

```
In [35]: API_df_clean = API_df_clean.drop(columns={"date_created"})
```

Test

```
In [36]: API_df_clean
```

```
Out[36]:                 tweet_id  favorite_count  retweet_count
         0     892420643555336193           33697           6969
         1     892177421306343426           29222           5272
         2     891815181378084864           21978           3464
         3     891689557279858688           36791           7191
         4     891327558926688256           35182           7717
         5     891087950875897856           17749           2586
         6     890971913173991426           10331           1647
         7     890729181411237888           56670          15679
         8     890609185150312448           24427           3604
         9     890240255349198849           27848           6068
         10    890006608113172480           26947           6111
         11    889880896479866881           24488           4141
         12    889665388333682689           41878           8302
```

13	889638837579907072	23580	3700
14	889531135344209921	13313	1874
15	889278841981685760	22026	4418
16	888917238123831296	25533	3742
17	888804989199671297	22392	3513
18	888554962724278272	17252	2863
19	888078434458587136	19101	2885
20	887705289381826560	26530	4516
21	887517139158093824	40548	9810
22	887473957103951883	59978	14952
23	887343217045368832	29490	8775
24	887101392804085760	26897	4969
25	886983233522544640	30239	6287
26	886736880519319552	10455	2615
27	886680336477933568	19655	3720
28	886366144734445568	18489	2613
29	886267009285017600	105	4
...	...	...	...
2297	666411507551481857	371	261
2298	666407126856765440	93	30
2299	666396247373291520	147	68
2300	666373753744588802	162	73
2301	666362758909284353	649	467
2302	666353288456101888	179	56
2303	666345417576210432	242	121
2304	666337882303524864	168	79
2305	666293911632134144	425	288
2306	666287406224695296	123	55
2307	666273097616637952	151	66
2308	666268910803644416	99	38
2309	666104133288665088	12844	5454
2310	666102155909144576	66	11
2311	666099513787052032	134	53
2312	666094000022159362	142	63
2313	666082916733198337	92	37
2314	666073100786774016	273	130
2315	666071193221509120	127	51
2316	666063827256086533	396	180
2317	666058600524156928	99	47
2318	666057090499244032	247	111
2319	666055525042405380	367	196
2320	666051853826850816	1023	699
2321	666050758794694657	115	50
2322	666049248165822465	88	36
2323	666044226329800704	246	115
2324	666033412701032449	100	36
2325	666029285002620928	112	39
2326	666020888022790149	2283	419

```
[2327 rows x 3 columns]
```

2 The columns ('jpg_url', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog') are dropped from the image prediction dataset because they're irrelevant

Code

```
In [37]: #Drop ('jpg_url', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog') column

         idx = np.r_ [1:2, 6:12]
         image_df_clean.drop(image_df_clean.columns[idx], axis= 1, inplace= True)
```

Test

```
In [38]: image_df_clean
```

Out[38]:

	tweet_id	img_num	p1	p1_conf
0	666020888022790149	1	Welsh_springer_spaniel	0.465074
1	666029285002620928	1	redbone	0.506826
2	666033412701032449	1	German_shepherd	0.596461
3	666044226329800704	1	Rhodesian_ridgeback	0.408143
4	666049248165822465	1	miniature_pinscher	0.560311
5	666050758794694657	1	Bernese_mountain_dog	0.651137
6	666051853826850816	1	box_turtle	0.933012
7	666055525042405380	1	chow	0.692517
8	666057090499244032	1	shopping_cart	0.962465
9	666058600524156928	1	miniature_poodle	0.201493
10	666063827256086533	1	golden_retriever	0.775930
11	666071193221509120	1	Gordon_setter	0.503672
12	666073100786774016	1	Walker_hound	0.260857
13	666082916733198337	1	pug	0.489814
14	666094000022159362	1	bloodhound	0.195217
15	666099513787052032	1	Lhasa	0.582330
16	666102155909144576	1	English_setter	0.298617
17	666104133288665088	1	hen	0.965932
18	666268910803644416	1	desktop_computer	0.086502
19	666273097616637952	1	Italian_greyhound	0.176053
20	666287406224695296	1	Maltese_dog	0.857531
21	666293911632134144	1	three-toed_sloth	0.914671
22	666337882303524864	1	ox	0.416669
23	666345417576210432	1	golden_retriever	0.858744
24	666353288456101888	1	malamute	0.336874
25	666362758909284353	1	guinea_pig	0.996496
26	666373753744588802	1	soft-coated_wheaten_terrier	0.326467
27	666396247373291520	1	Chihuahua	0.978108

```
28    666407126856765440        1      black-and-tan_coonhound  0.529139
29    666411507551481857        1                         coho  0.404640
...                  ...      ...                          ...       ...
2045  886366144734445568        1               French_bulldog  0.999201
2046  886680336477933568        1                  convertible  0.738995
2047  886736880519319552        1                       kuvasz  0.309706
2048  886983233522544640        2                    Chihuahua  0.793469
2049  887101392804085760        1                      Samoyed  0.733942
2050  887343217045368832        1            Mexican_hairless  0.330741
2051  887473957103951883        2                     Pembroke  0.809197
2052  887517139158093824        1                    limousine  0.130432
2053  887705289381826560        1                       basset  0.821664
2054  888078434458587136        1               French_bulldog  0.995026
2055  888202515573088257        2                     Pembroke  0.809197
2056  888554962724278272        3               Siberian_husky  0.700377
2057  888804989199671297        1            golden_retriever  0.469760
2058  888917238123831296        1            golden_retriever  0.714719
2059  889278841981685760        1                     whippet  0.626152
2060  889531135344209921        1            golden_retriever  0.953442
2061  889638837579907072        1               French_bulldog  0.991650
2062  889665388333682689        1                     Pembroke  0.966327
2063  889880896479866881        1               French_bulldog  0.377417
2064  890066608113172480        1                      Samoyed  0.957979
2065  890240255349198849        1                     Pembroke  0.511319
2066  890609185150312448        1                Irish_terrier  0.487574
2067  890729181411237888        2                   Pomeranian  0.566142
2068  890971913173991426        1                  Appenzeller  0.341703
2069  891087950875897856        1    Chesapeake_Bay_retriever  0.425595
2070  891327558926688256        2                       basset  0.555712
2071  891689557279858688        1                  paper_towel  0.170278
2072  891815181378084864        1                    Chihuahua  0.716012
2073  892177421306343426        1                    Chihuahua  0.323581
2074  892420643555336193        1                       orange  0.097049
```

```
      p1_dog
0       True
1       True
2       True
3       True
4       True
5       True
6      False
7       True
8      False
9       True
10      True
11      True
12      True
```

```
13      True
14      True
15      True
16      True
17     False
18     False
19      True
20      True
21     False
22     False
23      True
24      True
25     False
26      True
27      True
28      True
29     False
...      ...
2045    True
2046   False
2047    True
2048    True
2049    True
2050    True
2051    True
2052   False
2053    True
2054    True
2055    True
2056    True
2057    True
2058    True
2059    True
2060    True
2061    True
2062    True
2063    True
2064    True
2065    True
2066    True
2067    True
2068    True
2069    True
2070    True
2071   False
2072    True
2073    True
2074   False
```

```
[2075 rows x 5 columns]
```

2.0.1 Issue #4 : The dog "stage" (i.e. doggo, floofer, pupper, and puppo) should be one column

Define: Make one column that includes the type of dog "stages" and rename it dog_stages

Code

```
In [39]: twitter_df_clean['dog_stages'] = twitter_df_clean[
             ['doggo', 'floofer','pupper','puppo']].apply(lambda x: ', '.join(x), axis=1)
```

```
/opt/conda/lib/python3.6/site-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#
```

```
In [40]: twitter_df_clean = twitter_df_clean.replace(regex=r'(None,? ?)', value='').replace(rege

In [41]: twitter_df_clean = twitter_df_clean.replace(regex=r'', value= np.nan)

In [42]: twitter_df_clean.drop(columns = ['doggo', 'floofer','pupper','puppo'], inplace = True)

         idx = np.r_ [-4:]
         twitter_df_clean.drop(twitter_df_clean.columns[idx], axis= 1, inplace= True)
```

Test

```
In [43]: twitter_df_clean
```

Out[43]:

	tweet_id	timestamp	retweet_id	\
0	892420643555336193	2017-08-01 16:23:56 +0000	NaN	
1	892177421306343426	2017-08-01 00:17:27 +0000	NaN	
2	891815181378084864	2017-07-31 00:18:03 +0000	NaN	
3	891689557279858688	2017-07-30 15:58:51 +0000	NaN	
4	891327558926688256	2017-07-29 16:00:24 +0000	NaN	
5	891087950875897856	2017-07-29 00:08:17 +0000	NaN	
6	890971913173991426	2017-07-28 16:27:12 +0000	NaN	
7	890729181411237888	2017-07-28 00:22:40 +0000	NaN	
8	890609185150312448	2017-07-27 16:25:51 +0000	NaN	
9	890240255349198849	2017-07-26 15:59:51 +0000	NaN	
10	890006608113172480	2017-07-26 00:31:25 +0000	NaN	
11	889880896479866881	2017-07-25 16:11:53 +0000	NaN	
12	889665388333682689	2017-07-25 01:55:32 +0000	NaN	
13	889638837579907072	2017-07-25 00:10:02 +0000	NaN	
14	889531135344209921	2017-07-24 17:02:04 +0000	NaN	

```
15    889278841981685760  2017-07-24 00:19:32 +0000         NaN
16    888917238123831296  2017-07-23 00:22:39 +0000         NaN
17    888804989199671297  2017-07-22 16:56:37 +0000         NaN
18    888554962724278272  2017-07-22 00:23:06 +0000         NaN
20    888078434458587136  2017-07-20 16:49:33 +0000         NaN
21    887705289381826560  2017-07-19 16:06:48 +0000         NaN
22    887517139158093824  2017-07-19 03:39:09 +0000         NaN
23    887473957103951883  2017-07-19 00:47:34 +0000         NaN
24    887343217045368832  2017-07-18 16:08:03 +0000         NaN
25    887101392804085760  2017-07-18 00:07:08 +0000         NaN
26    886983233522544640  2017-07-17 16:17:36 +0000         NaN
27    886736880519319552  2017-07-16 23:58:41 +0000         NaN
28    886680336477933568  2017-07-16 20:14:00 +0000         NaN
29    886366144734445568  2017-07-15 23:25:31 +0000         NaN
30    886267009285017600  2017-07-15 16:51:35 +0000         NaN
...                  ...                        ...         ...
2326  666411507551481857  2015-11-17 00:24:19 +0000         NaN
2327  666407126856765440  2015-11-17 00:06:54 +0000         NaN
2328  666396247373291520  2015-11-16 23:23:41 +0000         NaN
2329  666373753744588802  2015-11-16 21:54:18 +0000         NaN
2330  666362758909284353  2015-11-16 21:10:36 +0000         NaN
2331  666353288456101888  2015-11-16 20:32:58 +0000         NaN
2332  666345417576210432  2015-11-16 20:01:42 +0000         NaN
2333  666337882303524864  2015-11-16 19:31:45 +0000         NaN
2334  666293911632134144  2015-11-16 16:37:02 +0000         NaN
2335  666287406224695296  2015-11-16 16:11:11 +0000         NaN
2336  666273097616637952  2015-11-16 15:14:19 +0000         NaN
2337  666268910803644416  2015-11-16 14:57:41 +0000         NaN
2338  666104133288665088  2015-11-16 04:02:55 +0000         NaN
2339  666102155909144576  2015-11-16 03:55:04 +0000         NaN
2340  666099513787052032  2015-11-16 03:44:34 +0000         NaN
2341  666094000022159362  2015-11-16 03:22:39 +0000         NaN
2342  666082916733198337  2015-11-16 02:38:37 +0000         NaN
2343  666073100786774016  2015-11-16 01:59:36 +0000         NaN
2344  666071193221509120  2015-11-16 01:52:02 +0000         NaN
2345  666063827256086533  2015-11-16 01:22:45 +0000         NaN
2346  666058600524156928  2015-11-16 01:01:59 +0000         NaN
2347  666057090499244032  2015-11-16 00:55:59 +0000         NaN
2348  666055525042405380  2015-11-16 00:49:46 +0000         NaN
2349  666051853826850816  2015-11-16 00:35:11 +0000         NaN
2350  666050758794694657  2015-11-16 00:30:50 +0000         NaN
2351  666049248165822465  2015-11-16 00:24:50 +0000         NaN
2352  666044226329800704  2015-11-16 00:04:52 +0000         NaN
2353  666033412701032449  2015-11-15 23:21:54 +0000         NaN
2354  666029285002620928  2015-11-15 23:05:30 +0000         NaN
2355  666020888022790149  2015-11-15 22:32:08 +0000         NaN

      rating_numerator  rating_denominator  dog_name dog_stages
```

0	13	10	Phineas	NaN
1	13	10	Tilly	NaN
2	12	10	Archie	NaN
3	13	10	Darla	NaN
4	12	10	Franklin	NaN
5	13	10	NaN	NaN
6	13	10	Jax	NaN
7	13	10	NaN	NaN
8	13	10	Zoey	NaN
9	14	10	Cassie	doggo
10	13	10	Koda	NaN
11	13	10	Bruno	NaN
12	13	10	NaN	puppo
13	12	10	Ted	NaN
14	13	10	Stuart	puppo
15	13	10	Oliver	NaN
16	12	10	Jim	NaN
17	13	10	Zeke	NaN
18	13	10	Ralphus	NaN
20	12	10	Gerald	NaN
21	13	10	Jeffrey	NaN
22	14	10	such	NaN
23	13	10	Canela	NaN
24	13	10	NaN	NaN
25	12	10	NaN	NaN
26	13	10	Maya	NaN
27	13	10	Mingus	NaN
28	13	10	Derek	NaN
29	12	10	Roscoe	pupper
30	12	10	NaN	NaN
...	...	...	...	...
2326	2	10	quite	NaN
2327	7	10	a	NaN
2328	9	10	NaN	NaN
2329	11	10	NaN	NaN
2330	6	10	NaN	NaN
2331	8	10	NaN	NaN
2332	10	10	NaN	NaN
2333	9	10	an	NaN
2334	3	10	a	NaN
2335	1	2	an	NaN
2336	11	10	NaN	NaN
2337	10	10	NaN	NaN
2338	1	10	NaN	NaN
2339	11	10	NaN	NaN
2340	8	10	NaN	NaN
2341	9	10	NaN	NaN
2342	6	10	NaN	NaN

```
2343                10                  10       NaN        NaN
2344                 9                  10       NaN        NaN
2345                10                  10       the        NaN
2346                 8                  10       the        NaN
2347                 9                  10         a        NaN
2348                10                  10         a        NaN
2349                 2                  10        an        NaN
2350                10                  10         a        NaN
2351                 5                  10       NaN        NaN
2352                 6                  10         a        NaN
2353                 9                  10         a        NaN
2354                 7                  10         a        NaN
2355                 8                  10       NaN        NaN

[2175 rows x 7 columns]
```

2.0.2 Issue #5 : Rating should be a single column

Define : Numerator and Denominator used for the ratings should be placed on a single column

Code

```
In [44]: # Combining two columns and giving it a new column "rating"

         twitter_df_clean["rating"] = twitter_df_clean["rating_numerator"].map(str) + twitter_df

In [45]: # In the resulting dataframe, we will delete the "rating_denominator" and "rating_numer
         twitter_df_clean = twitter_df_clean.drop(columns={"rating_numerator", "rating_denominat
```

Test

```
In [46]: twitter_df_clean
```

```
Out[46]:                tweet_id                  timestamp  retweet_id  dog_name  \
         0     892420643555336193  2017-08-01 16:23:56 +0000         NaN   Phineas
         1     892177421306343426  2017-08-01 00:17:27 +0000         NaN     Tilly
         2     891815181378084864  2017-07-31 00:18:03 +0000         NaN    Archie
         3     891689557279858688  2017-07-30 15:58:51 +0000         NaN     Darla
         4     891327558926688256  2017-07-29 16:00:24 +0000         NaN  Franklin
         5     891087950875897856  2017-07-29 00:08:17 +0000         NaN       NaN
         6     890971913173991426  2017-07-28 16:27:12 +0000         NaN       Jax
         7     890729181411237888  2017-07-28 00:22:40 +0000         NaN       NaN
         8     890609185150312448  2017-07-27 16:25:51 +0000         NaN      Zoey
         9     890240255349198849  2017-07-26 15:59:51 +0000         NaN    Cassie
         10    890006608113172480  2017-07-26 00:31:25 +0000         NaN      Koda
         11    889880896479866881  2017-07-25 16:11:53 +0000         NaN     Bruno
         12    889665388333682689  2017-07-25 01:55:32 +0000         NaN       NaN
         13    889638837579907072  2017-07-25 00:10:02 +0000         NaN       Ted
         14    889531135344209921  2017-07-24 17:02:04 +0000         NaN    Stuart
```

```
15    889278841981685760  2017-07-24 00:19:32 +0000        NaN    Oliver
16    888917238123831296  2017-07-23 00:22:39 +0000        NaN       Jim
17    888804989199671297  2017-07-22 16:56:37 +0000        NaN      Zeke
18    888554962724278272  2017-07-22 00:23:06 +0000        NaN   Ralphus
20    888078434458587136  2017-07-20 16:49:33 +0000        NaN    Gerald
21    887705289381826560  2017-07-19 16:06:48 +0000        NaN   Jeffrey
22    887517139158093824  2017-07-19 03:39:09 +0000        NaN      such
23    887473957103951883  2017-07-19 00:47:34 +0000        NaN    Canela
24    887343217045368832  2017-07-18 16:08:03 +0000        NaN       NaN
25    887101392804085760  2017-07-18 00:07:08 +0000        NaN       NaN
26    886983233522544640  2017-07-17 16:17:36 +0000        NaN      Maya
27    886736880519319552  2017-07-16 23:58:41 +0000        NaN    Mingus
28    886680336477933568  2017-07-16 20:14:00 +0000        NaN     Derek
29    886366144734445568  2017-07-15 23:25:31 +0000        NaN    Roscoe
30    886267009285017600  2017-07-15 16:51:35 +0000        NaN       NaN
...                  ...                        ...        ...       ...
2326  666411507551481857  2015-11-17 00:24:19 +0000        NaN     quite
2327  666407126856765440  2015-11-17 00:06:54 +0000        NaN         a
2328  666396247373291520  2015-11-16 23:23:41 +0000        NaN       NaN
2329  666373753744588802  2015-11-16 21:54:18 +0000        NaN       NaN
2330  666362758909284353  2015-11-16 21:10:36 +0000        NaN       NaN
2331  666353288456101888  2015-11-16 20:32:58 +0000        NaN       NaN
2332  666345417576210432  2015-11-16 20:01:42 +0000        NaN       NaN
2333  666337882303524864  2015-11-16 19:31:45 +0000        NaN        an
2334  666293911632134144  2015-11-16 16:37:02 +0000        NaN         a
2335  666287406224695296  2015-11-16 16:11:11 +0000        NaN        an
2336  666273097616637952  2015-11-16 15:14:19 +0000        NaN       NaN
2337  666268910803644416  2015-11-16 14:57:41 +0000        NaN       NaN
2338  666104133288665088  2015-11-16 04:02:55 +0000        NaN       NaN
2339  666102155909144576  2015-11-16 03:55:04 +0000        NaN       NaN
2340  666099513787052032  2015-11-16 03:44:34 +0000        NaN       NaN
2341  666094000022159362  2015-11-16 03:22:39 +0000        NaN       NaN
2342  666082916733198337  2015-11-16 02:38:37 +0000        NaN       NaN
2343  666073100786774016  2015-11-16 01:59:36 +0000        NaN       NaN
2344  666071193221509120  2015-11-16 01:52:02 +0000        NaN       NaN
2345  666063827256086533  2015-11-16 01:22:45 +0000        NaN       the
2346  666058600524156928  2015-11-16 01:01:59 +0000        NaN       the
2347  666057090499244032  2015-11-16 00:55:59 +0000        NaN         a
2348  666055525042405380  2015-11-16 00:49:46 +0000        NaN         a
2349  666051853826850816  2015-11-16 00:35:11 +0000        NaN        an
2350  666050758794694657  2015-11-16 00:30:50 +0000        NaN         a
2351  666049248165822465  2015-11-16 00:24:50 +0000        NaN       NaN
2352  666044226329800704  2015-11-16 00:04:52 +0000        NaN         a
2353  666033412701032449  2015-11-15 23:21:54 +0000        NaN         a
2354  666029285002620928  2015-11-15 23:05:30 +0000        NaN         a
2355  666020888022790149  2015-11-15 22:32:08 +0000        NaN       NaN

     dog_stages rating
```

```
0           NaN   1310
1           NaN   1310
2           NaN   1210
3           NaN   1310
4           NaN   1210
5           NaN   1310
6           NaN   1310
7           NaN   1310
8           NaN   1310
9         doggo   1410
10          NaN   1310
11          NaN   1310
12        puppo   1310
13          NaN   1210
14        puppo   1310
15          NaN   1310
16          NaN   1210
17          NaN   1310
18          NaN   1310
20          NaN   1210
21          NaN   1310
22          NaN   1410
23          NaN   1310
24          NaN   1310
25          NaN   1210
26          NaN   1310
27          NaN   1310
28          NaN   1310
29       pupper   1210
30          NaN   1210
...         ...    ...
2326        NaN    210
2327        NaN    710
2328        NaN    910
2329        NaN   1110
2330        NaN    610
2331        NaN    810
2332        NaN   1010
2333        NaN    910
2334        NaN    310
2335        NaN     12
2336        NaN   1110
2337        NaN   1010
2338        NaN    110
2339        NaN   1110
2340        NaN    810
2341        NaN    910
2342        NaN    610
```

```
2343         NaN   1010
2344         NaN    910
2345         NaN   1010
2346         NaN    810
2347         NaN    910
2348         NaN   1010
2349         NaN    210
2350         NaN   1010
2351         NaN    510
2352         NaN    610
2353         NaN    910
2354         NaN    710
2355         NaN    810

[2175 rows x 6 columns]
```

2.0.3 Issue #6: Data contained in the 'timestamp' not in the right format

Define: Timestamp data shows what time each tweet was made, but we do not need so much information about the time, hence extract just the date from it and rename the column "date"

Code

```
In [47]: twitter_df_clean[['date', 'time', 'mins']] = twitter_df_clean['timestamp'].str.split('

In [48]: # In the resulting dataframe, we would be deleting: "timestamp", "time", and "mins" col

         twitter_df_clean = twitter_df_clean.drop(columns={"timestamp", "time", "mins"})
```

Test

```
In [49]: twitter_df_clean
```

Out[49]:

	tweet_id	retweet_id	dog_name	dog_stages	rating	date
0	892420643555336193	NaN	Phineas	NaN	1310	2017-08-01
1	892177421306343426	NaN	Tilly	NaN	1310	2017-08-01
2	891815181378084864	NaN	Archie	NaN	1210	2017-07-31
3	891689557279858688	NaN	Darla	NaN	1310	2017-07-30
4	891327558926688256	NaN	Franklin	NaN	1210	2017-07-29
5	891087950875897856	NaN	NaN	NaN	1310	2017-07-29
6	890971913173991426	NaN	Jax	NaN	1310	2017-07-28
7	890729181411237888	NaN	NaN	NaN	1310	2017-07-28
8	890609185150312448	NaN	Zoey	NaN	1310	2017-07-27
9	890240255349198849	NaN	Cassie	doggo	1410	2017-07-26
10	890006608113172480	NaN	Koda	NaN	1310	2017-07-26
11	889880896479866881	NaN	Bruno	NaN	1310	2017-07-25
12	889665388333682689	NaN	NaN	puppo	1310	2017-07-25
13	889638837579907072	NaN	Ted	NaN	1210	2017-07-25
14	889531135344209921	NaN	Stuart	puppo	1310	2017-07-24

15	889278841981685760	NaN	Oliver	NaN	1310	2017-07-24
16	888917238123831296	NaN	Jim	NaN	1210	2017-07-23
17	888804989199671297	NaN	Zeke	NaN	1310	2017-07-22
18	888554962724278272	NaN	Ralphus	NaN	1310	2017-07-22
20	888078434458587136	NaN	Gerald	NaN	1210	2017-07-20
21	887705289381826560	NaN	Jeffrey	NaN	1310	2017-07-19
22	887517139158093824	NaN	such	NaN	1410	2017-07-19
23	887473957103951883	NaN	Canela	NaN	1310	2017-07-19
24	887343217045368832	NaN	NaN	NaN	1310	2017-07-18
25	887101392804085760	NaN	NaN	NaN	1210	2017-07-18
26	886983233522544640	NaN	Maya	NaN	1310	2017-07-17
27	886736880519319552	NaN	Mingus	NaN	1310	2017-07-16
28	886680336477933568	NaN	Derek	NaN	1310	2017-07-16
29	886366144734445568	NaN	Roscoe	pupper	1210	2017-07-15
30	886267009285017600	NaN	NaN	NaN	1210	2017-07-15
...	...	...	...	...	...	...
2326	666411507551481857	NaN	quite	NaN	210	2015-11-17
2327	666407126856765440	NaN	a	NaN	710	2015-11-17
2328	666396247373291520	NaN	NaN	NaN	910	2015-11-16
2329	666373753744588802	NaN	NaN	NaN	1110	2015-11-16
2330	666362758909284353	NaN	NaN	NaN	610	2015-11-16
2331	666353288456101888	NaN	NaN	NaN	810	2015-11-16
2332	666345417576210432	NaN	NaN	NaN	1010	2015-11-16
2333	666337882303524864	NaN	an	NaN	910	2015-11-16
2334	666293911632134144	NaN	a	NaN	310	2015-11-16
2335	666287406224695296	NaN	an	NaN	12	2015-11-16
2336	666273097616637952	NaN	NaN	NaN	1110	2015-11-16
2337	666268910803644416	NaN	NaN	NaN	1010	2015-11-16
2338	666104133288665088	NaN	NaN	NaN	110	2015-11-16
2339	666102155909144576	NaN	NaN	NaN	1110	2015-11-16
2340	666099513787052032	NaN	NaN	NaN	810	2015-11-16
2341	666094000022159362	NaN	NaN	NaN	910	2015-11-16
2342	666082916733198337	NaN	NaN	NaN	610	2015-11-16
2343	666073100786774016	NaN	NaN	NaN	1010	2015-11-16
2344	666071193221509120	NaN	NaN	NaN	910	2015-11-16
2345	666063827256086533	NaN	the	NaN	1010	2015-11-16
2346	666058600524156928	NaN	the	NaN	810	2015-11-16
2347	666057090499244032	NaN	a	NaN	910	2015-11-16
2348	666055525042405380	NaN	a	NaN	1010	2015-11-16
2349	666051853826850816	NaN	an	NaN	210	2015-11-16
2350	666050758794694657	NaN	a	NaN	1010	2015-11-16
2351	666049248165822465	NaN	NaN	NaN	510	2015-11-16
2352	666044226329800704	NaN	a	NaN	610	2015-11-16
2353	666033412701032449	NaN	a	NaN	910	2015-11-15
2354	666029285002620928	NaN	a	NaN	710	2015-11-15
2355	666020888022790149	NaN	NaN	NaN	810	2015-11-15

[2175 rows x 6 columns]

In [50]: image_df_clean

Out[50]:

	tweet_id	img_num	p1	p1_conf	\
0	666020888022790149	1	Welsh_springer_spaniel	0.465074	
1	666029285002620928	1	redbone	0.506826	
2	666033412701032449	1	German_shepherd	0.596461	
3	666044226329800704	1	Rhodesian_ridgeback	0.408143	
4	666049248165822465	1	miniature_pinscher	0.560311	
5	666050758794694657	1	Bernese_mountain_dog	0.651137	
6	666051853826850816	1	box_turtle	0.933012	
7	666055525042405380	1	chow	0.692517	
8	666057090499244032	1	shopping_cart	0.962465	
9	666058600524156928	1	miniature_poodle	0.201493	
10	666063827256086533	1	golden_retriever	0.775930	
11	666071193221509120	1	Gordon_setter	0.503672	
12	666073100786774016	1	Walker_hound	0.260857	
13	666082916733198337	1	pug	0.489814	
14	666094000022159362	1	bloodhound	0.195217	
15	666099513787052032	1	Lhasa	0.582330	
16	666102155909144576	1	English_setter	0.298617	
17	666104133288665088	1	hen	0.965932	
18	666268910803644416	1	desktop_computer	0.086502	
19	666273097616637952	1	Italian_greyhound	0.176053	
20	666287406224695296	1	Maltese_dog	0.857531	
21	666293911632134144	1	three-toed_sloth	0.914671	
22	666337882303524864	1	ox	0.416669	
23	666345417576210432	1	golden_retriever	0.858744	
24	666353288456101888	1	malamute	0.336874	
25	666362758909284353	1	guinea_pig	0.996496	
26	666373753744588802	1	soft-coated_wheaten_terrier	0.326467	
27	666396247373291520	1	Chihuahua	0.978108	
28	666407126856765440	1	black-and-tan_coonhound	0.529139	
29	666411507551481857	1	coho	0.404640	
...	...	...	...	...	
2045	886366144734445568	1	French_bulldog	0.999201	
2046	886680336477933568	1	convertible	0.738995	
2047	886736880519319552	1	kuvasz	0.309706	
2048	886983233522544640	2	Chihuahua	0.793469	
2049	887101392804085760	1	Samoyed	0.733942	
2050	887343217045368832	1	Mexican_hairless	0.330741	
2051	887473957103951883	2	Pembroke	0.809197	
2052	887517139158093824	1	limousine	0.130432	
2053	887705289381826560	1	basset	0.821664	
2054	888078434458587136	1	French_bulldog	0.995026	
2055	888202515573088257	2	Pembroke	0.809197	
2056	888554962724278272	3	Siberian_husky	0.700377	
2057	888804989199671297	1	golden_retriever	0.469760	
2058	888917238123831296	1	golden_retriever	0.714719	

2059	889278841981685760	1	whippet	0.626152
2060	889531135344209921	1	golden_retriever	0.953442
2061	889638837579907072	1	French_bulldog	0.991650
2062	889665388333682689	1	Pembroke	0.966327
2063	889880896479866881	1	French_bulldog	0.377417
2064	890006608113172480	1	Samoyed	0.957979
2065	890240255349198849	1	Pembroke	0.511319
2066	890609185150312448	1	Irish_terrier	0.487574
2067	890729181411237888	2	Pomeranian	0.566142
2068	890971913173991426	1	Appenzeller	0.341703
2069	891087950875897856	1	Chesapeake_Bay_retriever	0.425595
2070	891327558926688256	2	basset	0.555712
2071	891689557279858688	1	paper_towel	0.170278
2072	891815181378084864	1	Chihuahua	0.716012
2073	892177421306343426	1	Chihuahua	0.323581
2074	892420643555336193	1	orange	0.097049

	p1_dog
0	True
1	True
2	True
3	True
4	True
5	True
6	False
7	True
8	False
9	True
10	True
11	True
12	True
13	True
14	True
15	True
16	True
17	False
18	False
19	True
20	True
21	False
22	False
23	True
24	True
25	False
26	True
27	True
28	True
29	False

```
...      ...
2045    True
2046   False
2047    True
2048    True
2049    True
2050    True
2051    True
2052   False
2053    True
2054    True
2055    True
2056    True
2057    True
2058    True
2059    True
2060    True
2061    True
2062    True
2063    True
2064    True
2065    True
2066    True
2067    True
2068    True
2069    True
2070    True
2071   False
2072    True
2073    True
2074   False

[2075 rows x 5 columns]
```

```
In [51]: image_df_clean.columns
```

```
Out[51]: Index(['tweet_id', 'img_num', 'p1', 'p1_conf', 'p1_dog'], dtype='object')
```

2.0.4 Issue #7: False predictions in the image prediction dataframe indicates predictions contain animals other than dogs

Define: clean dataframe to contain only predictions that are dogs

Code

```
In [52]: image_df_clean = image_df_clean.loc[image_df_clean['p1_dog'] != False]
```

Test

In [53]: `image_df_clean`

Out[53]:

	tweet_id	img_num	p1	p1_conf	\
0	666020888022790149	1	Welsh_springer_spaniel	0.465074	
1	666029285002620928	1	redbone	0.506826	
2	666033412701032449	1	German_shepherd	0.596461	
3	666044226329800704	1	Rhodesian_ridgeback	0.408143	
4	666049248165822465	1	miniature_pinscher	0.560311	
5	666050758794694657	1	Bernese_mountain_dog	0.651137	
7	666055525042405380	1	chow	0.692517	
9	666058600524156928	1	miniature_poodle	0.201493	
10	666063827256086533	1	golden_retriever	0.775930	
11	666071193221509120	1	Gordon_setter	0.503672	
12	666073100786774016	1	Walker_hound	0.260857	
13	666082916733198337	1	pug	0.489814	
14	666094000022159362	1	bloodhound	0.195217	
15	666099513787052032	1	Lhasa	0.582330	
16	666102155909144576	1	English_setter	0.298617	
19	666273097616637952	1	Italian_greyhound	0.176053	
20	666287406224695296	1	Maltese_dog	0.857531	
23	666345417576210432	1	golden_retriever	0.858744	
24	666353288456101888	1	malamute	0.336874	
26	666373753744588802	1	soft-coated_wheaten_terrier	0.326467	
27	666396247373291520	1	Chihuahua	0.978108	
28	666407126856765440	1	black-and-tan_coonhound	0.529139	
30	666418789513326592	1	toy_terrier	0.149680	
31	666421158376562688	1	Blenheim_spaniel	0.906777	
32	666428276349472768	1	Pembroke	0.371361	
34	666435652385423360	1	Chesapeake_Bay_retriever	0.184130	
35	666437273139982337	1	Chihuahua	0.671853	
36	666447344410484738	1	curly-coated_retriever	0.322084	
37	666454714377183233	1	dalmatian	0.278954	
38	666644823164719104	1	Ibizan_hound	0.044333	
...	...	...	...	...	
2041	885311592912609280	1	Labrador_retriever	0.908703	
2042	885528943205470208	1	pug	0.369275	
2043	885984800019947520	1	Blenheim_spaniel	0.972494	
2044	886258384151887873	1	pug	0.943575	
2045	886366144734445568	1	French_bulldog	0.999201	
2047	886736880519319552	1	kuvasz	0.309706	
2048	886983233522544640	2	Chihuahua	0.793469	
2049	887101392804085760	1	Samoyed	0.733942	
2050	887343217045368832	1	Mexican_hairless	0.330741	
2051	887473957103951883	2	Pembroke	0.809197	
2053	887705289381826560	1	basset	0.821664	
2054	888078434458587136	1	French_bulldog	0.995026	
2055	888202515573088257	2	Pembroke	0.809197	

2056	885554962724278272	3	Siberian_husky	0.700377
2057	888804989199671297	1	golden_retriever	0.469760
2058	888917238123831296	1	golden_retriever	0.714719
2059	889278841981685760	1	whippet	0.626152
2060	889531135344209921	1	golden_retriever	0.953442
2061	889638837579907072	1	French_bulldog	0.991650
2062	889665388333682689	1	Pembroke	0.966327
2063	889880896479866881	1	French_bulldog	0.377417
2064	890006608113172480	1	Samoyed	0.957979
2065	890240255349198849	1	Pembroke	0.511319
2066	890609185150312448	1	Irish_terrier	0.487574
2067	890729181411237888	2	Pomeranian	0.566142
2068	890971913173991426	1	Appenzeller	0.341703
2069	891087950875897856	1	Chesapeake_Bay_retriever	0.425595
2070	891327558926688256	2	basset	0.555712
2072	891815181378084864	1	Chihuahua	0.716012
2073	892177421306343426	1	Chihuahua	0.323581

	p1_dog
0	True
1	True
2	True
3	True
4	True
5	True
7	True
9	True
10	True
11	True
12	True
13	True
14	True
15	True
16	True
19	True
20	True
23	True
24	True
26	True
27	True
28	True
30	True
31	True
32	True
34	True
35	True
36	True
37	True

```
38      True
...      ...
2041    True
2042    True
2043    True
2044    True
2045    True
2047    True
2048    True
2049    True
2050    True
2051    True
2053    True
2054    True
2055    True
2056    True
2057    True
2058    True
2059    True
2060    True
2061    True
2062    True
2063    True
2064    True
2065    True
2066    True
2067    True
2068    True
2069    True
2070    True
2072    True
2073    True

[1532 rows x 5 columns]
```

2.0.5 Issue#8: Columns are not descriptive

Define: Rename p1, p1_conf, p1_dog to a more descriptive name

Code

```
In [54]: image_df_clean = image_df_clean.rename(columns = {'p1' : 'dog_breed', 'p1_conf' : 'conf
```

Test

```
In [55]: image_df_clean

Out[55]:                 tweet_id  img_num                    dog_breed  \
         0     666020888022790149        1       Welsh_springer_spaniel
```

1	666029285002620928	1	redbone
2	666033412701032449	1	German_shepherd
3	666044226329800704	1	Rhodesian_ridgeback
4	666049248165822465	1	miniature_pinscher
5	666050758794694657	1	Bernese_mountain_dog
7	666055525042405380	1	chow
9	666058600524156928	1	miniature_poodle
10	666063827256086533	1	golden_retriever
11	666071193221509120	1	Gordon_setter
12	666073100786774016	1	Walker_hound
13	666082916733198337	1	pug
14	666094000022159362	1	bloodhound
15	666099513787052032	1	Lhasa
16	666102155909144576	1	English_setter
19	666273097616637952	1	Italian_greyhound
20	666287406224695296	1	Maltese_dog
23	666345417576210432	1	golden_retriever
24	666353288456101888	1	malamute
26	666373753744588802	1	soft-coated_wheaten_terrier
27	666396247373291520	1	Chihuahua
28	666407126856765440	1	black-and-tan_coonhound
30	666418789513326592	1	toy_terrier
31	666421158376562688	1	Blenheim_spaniel
32	666428276349472768	1	Pembroke
34	666435652385423360	1	Chesapeake_Bay_retriever
35	666437273139982337	1	Chihuahua
36	666447344410484738	1	curly-coated_retriever
37	666454714377183233	1	dalmatian
38	666644823164719104	1	Ibizan_hound
...	...	...	...
2041	885311592912609280	1	Labrador_retriever
2042	885528943205470208	1	pug
2043	885984800019947520	1	Blenheim_spaniel
2044	886258384151887873	1	pug
2045	886366144734445568	1	French_bulldog
2047	886736880519319552	1	kuvasz
2048	886983233522544640	2	Chihuahua
2049	887101392804085760	1	Samoyed
2050	887343217045368832	1	Mexican_hairless
2051	887473957103951883	2	Pembroke
2053	887705289381826560	1	basset
2054	888078434458587136	1	French_bulldog
2055	888202515573088257	2	Pembroke
2056	888554962724278272	3	Siberian_husky
2057	888804989199671297	1	golden_retriever
2058	888917238123831296	1	golden_retriever
2059	889278841981685760	1	whippet
2060	889531135344209921	1	golden_retriever

2061	889638837579907072	1	French_bulldog
2062	889665388333682689	1	Pembroke
2063	889880896479866881	1	French_bulldog
2064	890066608113172480	1	Samoyed
2065	890240255349198849	1	Pembroke
2066	890609185150312448	1	Irish_terrier
2067	890729181411237888	2	Pomeranian
2068	890971913173991426	1	Appenzeller
2069	891087950875897856	1	Chesapeake_Bay_retriever
2070	891327558926688256	2	basset
2072	891815181378084864	1	Chihuahua
2073	892177421306343426	1	Chihuahua

	confidence_percentage	prediction
0	0.465074	True
1	0.506826	True
2	0.596461	True
3	0.408143	True
4	0.560311	True
5	0.651137	True
7	0.692517	True
9	0.201493	True
10	0.775930	True
11	0.503672	True
12	0.260857	True
13	0.489814	True
14	0.195217	True
15	0.582330	True
16	0.298617	True
19	0.176053	True
20	0.857531	True
23	0.858744	True
24	0.336874	True
26	0.326467	True
27	0.978108	True
28	0.529139	True
30	0.149680	True
31	0.906777	True
32	0.371361	True
34	0.184130	True
35	0.671853	True
36	0.322084	True
37	0.278954	True
38	0.044333	True
...	...	...
2041	0.908703	True
2042	0.369275	True
2043	0.972494	True

```
2044              0.943575        True
2045              0.999201        True
2047              0.309706        True
2048              0.793469        True
2049              0.733942        True
2050              0.330741        True
2051              0.809197        True
2053              0.821664        True
2054              0.995026        True
2055              0.809197        True
2056              0.700377        True
2057              0.469760        True
2058              0.714719        True
2059              0.626152        True
2060              0.953442        True
2061              0.991650        True
2062              0.966327        True
2063              0.377417        True
2064              0.957979        True
2065              0.511319        True
2066              0.487574        True
2067              0.566142        True
2068              0.341703        True
2069              0.425595        True
2070              0.555712        True
2072              0.716012        True
2073              0.323581        True

[1532 rows x 5 columns]
```

```
In [56]: twitter_df_clean.sample(5)
```

Out[56]:

	tweet_id	retweet_id	dog_name	dog_stages	rating	date
952	751937170840121344	NaN	Ruby	NaN	1110	2016-07-10
322	834167344700198914	NaN	Sunshine	NaN	1110	2017-02-21
2233	668204964695683073	NaN	Ron	NaN	810	2015-11-21
2051	671488513339211776	NaN	Julius	NaN	810	2015-12-01
951	751950017322246144	NaN	Lola	pupper	1310	2016-07-10

2.0.6 Issue#9: Name column in the twitter_archive dataset contains incosistent data

Define: Change inconsistent names to 'None'

Code

```
In [57]: twitter_df_clean.loc[twitter_df.name.str.islower(),'dog_name'] = None
         twitter_df_clean.dog_name.replace(regex=r'None', value= np.nan, inplace = True)
```

Test

In [58]: `twitter_df_clean`

Out[58]:

	tweet_id	retweet_id	dog_name	dog_stages	rating	date
0	892420643555336193	NaN	Phineas	NaN	1310	2017-08-01
1	892177421306343426	NaN	Tilly	NaN	1310	2017-08-01
2	891815181378084864	NaN	Archie	NaN	1210	2017-07-31
3	891689557279858688	NaN	Darla	NaN	1310	2017-07-30
4	891327558926688256	NaN	Franklin	NaN	1210	2017-07-29
5	891087950875897856	NaN	NaN	NaN	1310	2017-07-29
6	890971913173991426	NaN	Jax	NaN	1310	2017-07-28
7	890729181411237888	NaN	NaN	NaN	1310	2017-07-28
8	890609185150312448	NaN	Zoey	NaN	1310	2017-07-27
9	890240255349198849	NaN	Cassie	doggo	1410	2017-07-26
10	890066608113172480	NaN	Koda	NaN	1310	2017-07-26
11	889880896479866881	NaN	Bruno	NaN	1310	2017-07-25
12	889665388333682689	NaN	NaN	puppo	1310	2017-07-25
13	889638837579907072	NaN	Ted	NaN	1210	2017-07-25
14	889531135344209921	NaN	Stuart	puppo	1310	2017-07-24
15	889278841981685760	NaN	Oliver	NaN	1310	2017-07-24
16	888917238123831296	NaN	Jim	NaN	1210	2017-07-23
17	888804989199671297	NaN	Zeke	NaN	1310	2017-07-22
18	888554962724278272	NaN	Ralphus	NaN	1310	2017-07-22
20	888078434458587136	NaN	Gerald	NaN	1210	2017-07-20
21	887705289381826560	NaN	Jeffrey	NaN	1310	2017-07-19
22	887517139158093824	NaN	None	NaN	1410	2017-07-19
23	887473957103951883	NaN	Canela	NaN	1310	2017-07-19
24	887343217045368832	NaN	NaN	NaN	1310	2017-07-18
25	887101392804085760	NaN	NaN	NaN	1210	2017-07-18
26	886983233522544640	NaN	Maya	NaN	1310	2017-07-17
27	886736880519319552	NaN	Mingus	NaN	1310	2017-07-16
28	886680336477933568	NaN	Derek	NaN	1310	2017-07-16
29	886366144734445568	NaN	Roscoe	pupper	1210	2017-07-15
30	886267009285017600	NaN	NaN	NaN	1210	2017-07-15
...	...	...	...	...	...	...
2326	666411507551481857	NaN	None	NaN	210	2015-11-17
2327	666407126856765440	NaN	None	NaN	710	2015-11-17
2328	666396247373291520	NaN	NaN	NaN	910	2015-11-16
2329	666373753744588802	NaN	NaN	NaN	1110	2015-11-16
2330	666362758909284353	NaN	NaN	NaN	610	2015-11-16
2331	666353288456101888	NaN	NaN	NaN	810	2015-11-16
2332	666345417576210432	NaN	NaN	NaN	1010	2015-11-16
2333	666337882303524864	NaN	None	NaN	910	2015-11-16
2334	666293911632134144	NaN	None	NaN	310	2015-11-16
2335	666287406224695296	NaN	None	NaN	12	2015-11-16
2336	666273097616637952	NaN	NaN	NaN	1110	2015-11-16
2337	666268910803644416	NaN	NaN	NaN	1010	2015-11-16
2338	666104133288665088	NaN	NaN	NaN	110	2015-11-16

```
2339  666102155909144576         NaN       NaN        NaN   1110  2015-11-16
2340  666099513787052032         NaN       NaN        NaN    810  2015-11-16
2341  666094000022159362         NaN       NaN        NaN    910  2015-11-16
2342  666082916733198337         NaN       NaN        NaN    610  2015-11-16
2343  666073100786774016         NaN       NaN        NaN   1010  2015-11-16
2344  666071193221509120         NaN       NaN        NaN    910  2015-11-16
2345  666063827256086533         NaN      None        NaN   1010  2015-11-16
2346  666058600524156928         NaN      None        NaN    810  2015-11-16
2347  666057090499244032         NaN      None        NaN    910  2015-11-16
2348  666055525042405380         NaN      None        NaN   1010  2015-11-16
2349  666051853826850816         NaN      None        NaN    210  2015-11-16
2350  666050758794694657         NaN      None        NaN   1010  2015-11-16
2351  666049248165822465         NaN       NaN        NaN    510  2015-11-16
2352  666044226329800704         NaN      None        NaN    610  2015-11-16
2353  666033412701032449         NaN      None        NaN    910  2015-11-15
2354  666029285002620928         NaN      None        NaN    710  2015-11-15
2355  666020888022790149         NaN       NaN        NaN    810  2015-11-15

[2175 rows x 6 columns]
```

2.0.7 Issue#10: Dogs breeds are inconsistent

Define: Separator in the 'dog_breed' column should be replaced with space

Code

```
In [59]: image_df_clean.dog_breed = image_df_clean.dog_breed.replace(
             regex=r'(_)', value=' ').str.title()
```

Test

```
In [60]: image_df_clean
```

```
Out[60]:                 tweet_id  img_num                    dog_breed  \
         0     666020888022790149        1       Welsh Springer Spaniel
         1     666029285002620928        1                      Redbone
         2     666033412701032449        1             German Shepherd
         3     666044226329800704        1          Rhodesian Ridgeback
         4     666049248165822465        1           Miniature Pinscher
         5     666050758794694657        1         Bernese Mountain Dog
         7     666055525042405380        1                          Chow
         9     666058600524156928        1             Miniature Poodle
         10    666063827256086533        1             Golden Retriever
         11    666071193221509120        1                Gordon Setter
         12    666073100786774016        1                 Walker Hound
         13    666082916733198337        1                          Pug
         14    666094000022159362        1                   Bloodhound
         15    666099513787052032        1                        Lhasa
```

16	666102155909144576	1	English Setter
19	666273097616637952	1	Italian Greyhound
20	666287406224695296	1	Maltese Dog
23	666345417576210432	1	Golden Retriever
24	666353288456101888	1	Malamute
26	666373753744588802	1	Soft-Coated Wheaten Terrier
27	666396247373291520	1	Chihuahua
28	666407126856765440	1	Black-And-Tan Coonhound
30	666418789513326592	1	Toy Terrier
31	666421158376562688	1	Blenheim Spaniel
32	666428276349472768	1	Pembroke
34	666435652385423360	1	Chesapeake Bay Retriever
35	666437273139982337	1	Chihuahua
36	666447344410484738	1	Curly-Coated Retriever
37	666454714377183233	1	Dalmatian
38	666644823164719104	1	Ibizan Hound
...	...	...	...
2041	885311592912609280	1	Labrador Retriever
2042	885528943205470208	1	Pug
2043	885984800019947520	1	Blenheim Spaniel
2044	886258384151887873	1	Pug
2045	886366144734445568	1	French Bulldog
2047	886736880519319552	1	Kuvasz
2048	886983233522544640	2	Chihuahua
2049	887101392804085760	1	Samoyed
2050	887343217045368832	1	Mexican Hairless
2051	887473957103951883	2	Pembroke
2053	887705289381826560	1	Basset
2054	888078434458587136	1	French Bulldog
2055	888202515573088257	2	Pembroke
2056	888554962724278272	3	Siberian Husky
2057	888804989199671297	1	Golden Retriever
2058	888917238123831296	1	Golden Retriever
2059	889278841981685760	1	Whippet
2060	889531135344209921	1	Golden Retriever
2061	889638837579907072	1	French Bulldog
2062	889665388333682689	1	Pembroke
2063	889880896479866881	1	French Bulldog
2064	890006608113172480	1	Samoyed
2065	890240255349198849	1	Pembroke
2066	890609185150312448	1	Irish Terrier
2067	890729181411237888	2	Pomeranian
2068	890971913173991426	1	Appenzeller
2069	891087950875897856	1	Chesapeake Bay Retriever
2070	891327558926688256	2	Basset
2072	891815181378084864	1	Chihuahua
2073	892177421306343426	1	Chihuahua

	confidence_percentage	prediction
0	0.465074	True
1	0.506826	True
2	0.596461	True
3	0.408143	True
4	0.560311	True
5	0.651137	True
7	0.692517	True
9	0.201493	True
10	0.775930	True
11	0.503672	True
12	0.260857	True
13	0.489814	True
14	0.195217	True
15	0.582330	True
16	0.298617	True
19	0.176053	True
20	0.857531	True
23	0.858744	True
24	0.336874	True
26	0.326467	True
27	0.978108	True
28	0.529139	True
30	0.149680	True
31	0.906777	True
32	0.371361	True
34	0.184130	True
35	0.671853	True
36	0.322084	True
37	0.278954	True
38	0.044333	True
...	...	...
2041	0.908703	True
2042	0.369275	True
2043	0.972494	True
2044	0.943575	True
2045	0.999201	True
2047	0.309706	True
2048	0.793469	True
2049	0.733942	True
2050	0.330741	True
2051	0.809197	True
2053	0.821664	True
2054	0.995026	True
2055	0.809197	True
2056	0.700377	True
2057	0.469760	True
2058	0.714719	True

```
2059                0.626152        True
2060                0.953442        True
2061                0.991650        True
2062                0.966327        True
2063                0.377417        True
2064                0.957979        True
2065                0.511319        True
2066                0.487574        True
2067                0.566142        True
2068                0.341703        True
2069                0.425595        True
2070                0.555712        True
2072                0.716012        True
2073                0.323581        True

[1532 rows x 5 columns]
```

2.0.8 Merging all the 3 dataframes together

```
In [61]: dfs = [twitter_df_clean, image_df_clean, API_df_clean]
         twitter_archive_master = reduce(lambda left,right: pd.merge(left,right,on='tweet_id', h

In [62]: twitter_archive_master.head(3)
```

Out[62]:

	tweet_id	retweet_id	dog_name	dog_stages	rating	date	\
0	892177421306343426	NaN	Tilly	NaN	1310	2017-08-01	
1	891815181378084864	NaN	Archie	NaN	1210	2017-07-31	
2	891327558926688256	NaN	Franklin	NaN	1210	2017-07-29	

	img_num	dog_breed	confidence_percentage	prediction	favorite_count	\
0	1	Chihuahua	0.323581	True	29222	
1	1	Chihuahua	0.716012	True	21978	
2	2	Basset	0.555712	True	35182	

	retweet_count
0	5272
1	3464
2	7717

```
In [63]: # checking for null values in just created master dataframe
         twitter_archive_master.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1469 entries, 0 to 1468
Data columns (total 12 columns):
tweet_id                1469 non-null int64
retweet_id              0 non-null float64
dog_name                1037 non-null object
dog_stages              227 non-null object
```

```
rating                    1469 non-null object
date                      1469 non-null object
img_num                   1469 non-null int64
dog_breed                 1469 non-null object
confidence_percentage     1469 non-null float64
prediction                1469 non-null bool
favorite_count            1469 non-null int64
retweet_count             1469 non-null int64
dtypes: bool(1), float64(2), int64(4), object(5)
memory usage: 139.2+ KB
```

2.1 Storing Data

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

```
In [64]: twitter_archive_master.to_csv('twitter_archive_master.csv', index = False)
```

2.2 Analyzing Data

```
In [65]: # The most liked Dog stage
         twitter_archive_master.groupby(['dog_stages'], as_index= False)['favorite_count', 'retw
```

Out[65]:

	dog_stages	favorite_count	retweet_count
0	doggo	18930.978723	6401.382979
1	doggo, floofer	14801.000000	2784.000000
2	doggo, pupper	12565.875000	3606.875000
3	doggo, puppo	41814.000000	16076.000000
4	floofer	11178.714286	3959.857143
5	pupper	6940.111111	2060.402778
6	puppo	20315.842105	5390.684211

```
In [66]: twitter_name = twitter_archive_master.groupby(['dog_name'], as_index= False)['favorite_
         twitter_name.head()
```

Out[66]:

	dog_name	favorite_count	retweet_count
0	Abby	4174.0	1099.5
1	Ace	3117.0	955.0
2	Acro	1012.0	264.0
3	Adele	2858.0	661.0
4	Aiden	1386.0	559.0

```
In [67]: # Getting the most liked name of dog
         top_name = twitter_name.sort_values(by= ['favorite_count', 'retweet_count'], ascending=
         top_name.head()
```

Out[67]:

	dog_name	favorite_count	retweet_count
645	Stephan	111198.0	51426.0

```
320   Jamesy        108499.0        30058.0
208  Duddles         92504.0        37265.0
5        Aja         69047.0        15734.0
406    Lilly         62249.0        15364.0
```

In [68]:
```
# Most popular dog breed
twitter_archive_master.groupby(['dog_breed']).favorite_count.sum().sort_values(ascendin
```

Out[68]:
```
dog_breed
Golden Retriever               1470148
Labrador Retriever              907059
Pembroke                        864393
Chihuahua                       596063
Samoyed                         446309
French Bulldog                  407286
Chow                            352193
Pug                             270198
Cocker Spaniel                  267086
Pomeranian                      263559
Eskimo Dog                      226389
Cardigan                        221058
Malamute                        216139
Chesapeake Bay Retriever        206228
Toy Poodle                      201768
Lakeland Terrier                186422
German Shepherd                 172265
Miniature Pinscher              162260
Basset                          154331
Great Pyrenees                  152031
Shetland Sheepdog               130852
Staffordshire Bullterrier       126521
English Springer                120589
Siberian Husky                  119187
Italian Greyhound               118355
Rottweiler                      116229
Flat-Coated Retriever           115818
Kelpie                           95082
Standard Poodle                  95077
Border Collie                    91652
                                 ...
Soft-Coated Wheaten Terrier      20954
Norwich Terrier                  20503
Australian Terrier               19055
Gordon Setter                    18508
Dandie Dinmont                   17510
Bluetick                         17039
Keeshond                         16495
Redbone                          16466
```

```
Basenji                          15828
Cairn                            15284
Wire-Haired Fox Terrier          14335
Miniature Schnauzer              13952
Rhodesian Ridgeback              13803
Appenzeller                      12496
Curly-Coated Retriever           11816
Lhasa                            11136
Toy Terrier                       7634
Welsh Springer Spaniel            6987
Sussex Spaniel                    6845
Silky Terrier                     6222
Tibetan Terrier                   6211
Scottish Deerhound                6182
Clumber                           6174
Scotch Terrier                    3010
Ibizan Hound                      2788
Entlebucher                       2243
Brabancon Griffon                 2227
Standard Schnauzer                1683
Groenendael                       1619
Japanese Spaniel                  1111
Name: favorite_count, Length: 111, dtype: int64
```

2.2.1 Insights:

1.From the dataset, we can draw the conclusion that doggo dog stage has the highest popularity.
2.The most liked name of dog is Stephan.
3.The most popular dog is the golden_retriever

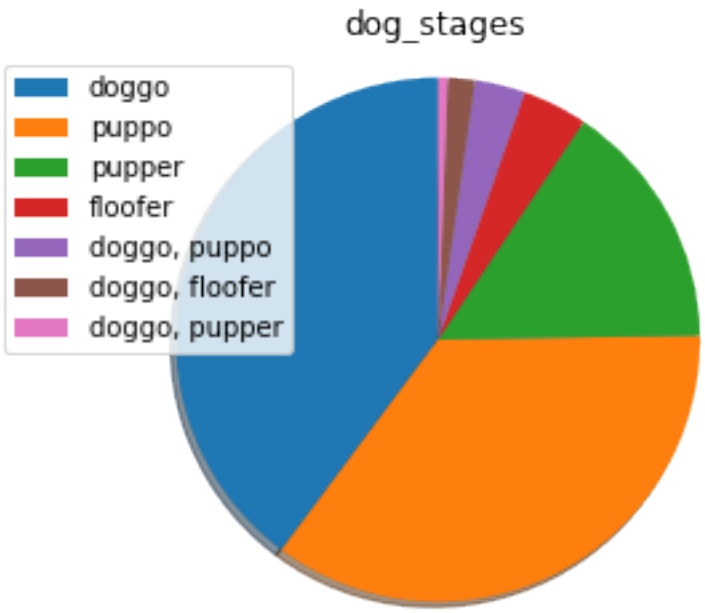
2.2.2 Visualization

```
In [69]: labels = twitter_archive_master.dog_stages.unique().tolist()[1:]
         sizes = twitter_archive_master.groupby(['dog_stages']).favorite_count.sum().sort_values
         # plt.bar(labels, values, width=10)
         #To set the title


         fig1, ax1 = plt.subplots()
         ax1.pie(sizes,shadow=True, startangle=90)

         plt.legend(labels)
         plt.title('dog_stages')
         ax1.axis('equal')

         plt.show()
```



From the visualization above, it is conclusive to say doggo has the is the most liked dog stage amongst the various dog stages

In [ ]: