wrangle_act

June 27, 2022

1 Project: Wrangling and Analyze Data

Author ---> Edwin Kihara

1.1 Introduction

Using Python and its libraries, we will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called **data wrangling**. We will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python and its libraries.

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and shared it exclusively for use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

1.2 Files Required for this project

- **wrangle_act.ipynb**: code for gathering, assessing, cleaning, analyzing, and visualizing data
- **wrangle_report.pdf or wrangle_report.html**: documentation for data wrangling steps: gather, assess, and clean
- **act_report.pdf or act_report.html**: documentation of analysis and insights into final data
- **twitter_archive_enhanced.csv**: file as given
- **image_predictions.tsv**: file downloaded programmatically
- **tweet_json.txt**: file constructed via API
- **twitter_archive_master.csv**: combined and cleaned data

1.3 Data Gathering

Gathering Data for this Project composed from three pieces of data as described below:

1. The WeRateDogs Twitter archive. We will download this file manually by clicking the following link: twitter_archive_enhanced.csv

2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Each tweet's retweet count and favorite (i.e. "like") count at minimum, and any additional data we will find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, we will query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then we will read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

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

In [2]:
```
# Import the libraries that we will need in this project
import pandas as pd
import datetime as dt
import numpy as np
import requests
import tweepy
import json
import re
import time
from nltk import pos_tag
```

In [3]:
```
# Read the twitter-archive-enhanced.csv file and store it as dataframe in archive
archive = pd.read_csv('twitter-archive-enhanced.csv', encoding = 'utf-8')
# Quick check to the file content and structure
archive
```

Out[3]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	\
0	892420643555336193	NaN	NaN	
1	892177421306343426	NaN	NaN	
2	891815181378084864	NaN	NaN	
3	891689557279858688	NaN	NaN	
4	891327558926688256	NaN	NaN	
5	891087950875897856	NaN	NaN	
6	890971913173991426	NaN	NaN	
7	890729181411237888	NaN	NaN	
8	890609185150312448	NaN	NaN	
9	890240255349198849	NaN	NaN	

10	890006608113172480	NaN	NaN
11	889880896479866881	NaN	NaN
12	889665388333682689	NaN	NaN
13	889638837579907072	NaN	NaN
14	889531135344209921	NaN	NaN
15	889278841981685760	NaN	NaN
16	888917238123831296	NaN	NaN
17	888804989199671297	NaN	NaN
18	888554962724278272	NaN	NaN
19	888202515573088257	NaN	NaN
20	888078434458587136	NaN	NaN
21	887705289381826560	NaN	NaN
22	887517139158093824	NaN	NaN
23	887473957103951883	NaN	NaN
24	887343217045368832	NaN	NaN
25	887101392804085760	NaN	NaN
26	886983233522544640	NaN	NaN
27	886736880519319552	NaN	NaN
28	886680336477933568	NaN	NaN
29	886366144734445568	NaN	NaN
...	...	...	...
2326	666411507551481857	NaN	NaN
2327	666407126856765440	NaN	NaN
2328	666396247373291520	NaN	NaN
2329	666373753744588802	NaN	NaN
2330	666362758909284353	NaN	NaN
2331	666353288456101888	NaN	NaN
2332	666345417576210432	NaN	NaN
2333	666337882303524864	NaN	NaN
2334	666293911632134144	NaN	NaN
2335	666287406224695296	NaN	NaN
2336	666273097616637952	NaN	NaN
2337	666268910803644416	NaN	NaN
2338	666104133288665088	NaN	NaN
2339	666102155909144576	NaN	NaN
2340	666099513787052032	NaN	NaN
2341	666094000022159362	NaN	NaN
2342	666082916733198337	NaN	NaN
2343	666073100786774016	NaN	NaN
2344	666071193221509120	NaN	NaN
2345	666063827256086533	NaN	NaN
2346	666058600524156928	NaN	NaN
2347	666057090499244032	NaN	NaN
2348	666055525042405380	NaN	NaN
2349	666051853826850816	NaN	NaN
2350	666050758794694657	NaN	NaN
2351	666049248165822465	NaN	NaN
2352	666044226329800704	NaN	NaN

```
2353  666033412701032449                    NaN                    NaN
2354  666029285002620928                    NaN                    NaN
2355  666020888022790149                    NaN                    NaN

                      timestamp  \
0     2017-08-01 16:23:56 +0000
1     2017-08-01 00:17:27 +0000
2     2017-07-31 00:18:03 +0000
3     2017-07-30 15:58:51 +0000
4     2017-07-29 16:00:24 +0000
5     2017-07-29 00:08:17 +0000
6     2017-07-28 16:27:12 +0000
7     2017-07-28 00:22:40 +0000
8     2017-07-27 16:25:51 +0000
9     2017-07-26 15:59:51 +0000
10    2017-07-26 00:31:25 +0000
11    2017-07-25 16:11:53 +0000
12    2017-07-25 01:55:32 +0000
13    2017-07-25 00:10:02 +0000
14    2017-07-24 17:02:04 +0000
15    2017-07-24 00:19:32 +0000
16    2017-07-23 00:22:39 +0000
17    2017-07-22 16:56:37 +0000
18    2017-07-22 00:23:06 +0000
19    2017-07-21 01:02:36 +0000
20    2017-07-20 16:49:33 +0000
21    2017-07-19 16:06:48 +0000
22    2017-07-19 03:39:09 +0000
23    2017-07-19 00:47:34 +0000
24    2017-07-18 16:08:03 +0000
25    2017-07-18 00:07:08 +0000
26    2017-07-17 16:17:36 +0000
27    2017-07-16 23:58:41 +0000
28    2017-07-16 20:14:00 +0000
29    2017-07-15 23:25:31 +0000
...                        ...
2326  2015-11-17 00:24:19 +0000
2327  2015-11-17 00:06:54 +0000
2328  2015-11-16 23:23:41 +0000
2329  2015-11-16 21:54:18 +0000
2330  2015-11-16 21:10:36 +0000
2331  2015-11-16 20:32:58 +0000
2332  2015-11-16 20:01:42 +0000
2333  2015-11-16 19:31:45 +0000
2334  2015-11-16 16:37:02 +0000
2335  2015-11-16 16:11:11 +0000
2336  2015-11-16 15:14:19 +0000
2337  2015-11-16 14:57:41 +0000
```

```
2338  2015-11-16 04:02:55 +0000
2339  2015-11-16 03:55:04 +0000
2340  2015-11-16 03:44:34 +0000
2341  2015-11-16 03:22:39 +0000
2342  2015-11-16 02:38:37 +0000
2343  2015-11-16 01:59:36 +0000
2344  2015-11-16 01:52:02 +0000
2345  2015-11-16 01:22:45 +0000
2346  2015-11-16 01:01:59 +0000
2347  2015-11-16 00:55:59 +0000
2348  2015-11-16 00:49:46 +0000
2349  2015-11-16 00:35:11 +0000
2350  2015-11-16 00:30:50 +0000
2351  2015-11-16 00:24:50 +0000
2352  2015-11-16 00:04:52 +0000
2353  2015-11-15 23:21:54 +0000
2354  2015-11-15 23:05:30 +0000
2355  2015-11-15 22:32:08 +0000

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

```
28    <a href="http://twitter.com/download/iphone" r...
29    <a href="http://twitter.com/download/iphone" r...
...                                                 ...
2326  <a href="http://twitter.com/download/iphone" r...
2327  <a href="http://twitter.com/download/iphone" r...
2328  <a href="http://twitter.com/download/iphone" r...
2329  <a href="http://twitter.com/download/iphone" r...
2330  <a href="http://twitter.com/download/iphone" r...
2331  <a href="http://twitter.com/download/iphone" r...
2332  <a href="http://twitter.com/download/iphone" r...
2333  <a href="http://twitter.com/download/iphone" r...
2334  <a href="http://twitter.com/download/iphone" r...
2335  <a href="http://twitter.com/download/iphone" r...
2336  <a href="http://twitter.com/download/iphone" r...
2337  <a href="http://twitter.com/download/iphone" r...
2338  <a href="http://twitter.com/download/iphone" r...
2339  <a href="http://twitter.com/download/iphone" r...
2340  <a href="http://twitter.com/download/iphone" r...
2341  <a href="http://twitter.com/download/iphone" r...
2342  <a href="http://twitter.com/download/iphone" r...
2343  <a href="http://twitter.com/download/iphone" r...
2344  <a href="http://twitter.com/download/iphone" r...
2345  <a href="http://twitter.com/download/iphone" r...
2346  <a href="http://twitter.com/download/iphone" r...
2347  <a href="http://twitter.com/download/iphone" r...
2348  <a href="http://twitter.com/download/iphone" r...
2349  <a href="http://twitter.com/download/iphone" r...
2350  <a href="http://twitter.com/download/iphone" r...
2351  <a href="http://twitter.com/download/iphone" r...
2352  <a href="http://twitter.com/download/iphone" r...
2353  <a href="http://twitter.com/download/iphone" r...
2354  <a href="http://twitter.com/download/iphone" r...
2355  <a href="http://twitter.com/download/iphone" r...

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

13	This is Ted. He does his best. Sometimes that'...	NaN
14	This is Stuart. He's sporting his favorite fan...	NaN
15	This is Oliver. You're witnessing one of his m...	NaN
16	This is Jim. He found a fren. Taught him how t...	NaN
17	This is Zeke. He has a new stick. Very proud o...	NaN
18	This is Ralphus. He's powering up. Attempting ...	NaN
19	RT @dog_rates: This is Canela. She attempted s...	8.874740e+17
20	This is Gerald. He was just told he didn't get...	NaN
21	This is Jeffrey. He has a monopoly on the pool...	NaN
22	I've yet to rate a Venezuelan Hover Wiener. Th...	NaN
23	This is Canela. She attempted some fancy porch...	NaN
24	You may not have known you needed to see this ...	NaN
25	This... is a Jubilant Antarctic House Bear. We...	NaN
26	This is Maya. She's very shy. Rarely leaves he...	NaN
27	This is Mingus. He's a wonderful father to his...	NaN
28	This is Derek. He's late for a dog meeting. 13...	NaN
29	This is Roscoe. Another pupper fallen victim t...	NaN
...	...	...
2326	This is quite the dog. Gets really excited whe...	NaN
2327	This is a southern Vesuvius bumblegruff. Can d...	NaN
2328	Oh goodness. A super rare northeast Qdoba kang...	NaN
2329	Those are sunglasses and a jean jacket. 11/10 ...	NaN
2330	Unique dog here. Very small. Lives in containe...	NaN
2331	Here we have a mixed Asiago from the Galápagos...	NaN
2332	Look at this jokester thinking seat belt laws ...	NaN
2333	This is an extremely rare horned Parthenon. No...	NaN
2334	This is a funny dog. Weird toes. Won't come do...	NaN
2335	This is an Albanian 3 1/2 legged Episcopalian...	NaN
2336	Can take selfies 11/10 https://t.co/ws2AMaNwPW	NaN
2337	Very concerned about fellow dog trapped in com...	NaN
2338	Not familiar with this breed. No tail (weird)...	NaN
2339	Oh my. Here you are seeing an Adobe Setter giv...	NaN
2340	Can stand on stump for what seems like a while...	NaN
2341	This appears to be a Mongolian Presbyterian mi...	NaN
2342	Here we have a well-established sunblockerspan...	NaN
2343	Let's hope this flight isn't Malaysian (lol). ...	NaN
2344	Here we have a northern speckled Rhododendron...	NaN
2345	This is the happiest dog you will ever see. Ve...	NaN
2346	Here is the Rand Paul of retrievers folks! He'...	NaN
2347	My oh my. This is a rare blond Canadian terrie...	NaN
2348	Here is a Siberian heavily armored polar bear ...	NaN
2349	This is an odd dog. Hard on the outside but lo...	NaN
2350	This is a truly beautiful English Wilson Staff...	NaN
2351	Here we have a 1949 1st generation vulpix. Enj...	NaN
2352	This is a purebred Piers Morgan. Loves to Netf...	NaN
2353	Here is a very happy pup. Big fan of well-main...	NaN
2354	This is a western brown Mitsubishi terrier. Up...	NaN
2355	Here we have a Japanese Irish Setter. Lost eye...	NaN

```
      retweeted_status_user_id retweeted_status_timestamp  \
0                          NaN                         NaN
1                          NaN                         NaN
2                          NaN                         NaN
3                          NaN                         NaN
4                          NaN                         NaN
5                          NaN                         NaN
6                          NaN                         NaN
7                          NaN                         NaN
8                          NaN                         NaN
9                          NaN                         NaN
10                         NaN                         NaN
11                         NaN                         NaN
12                         NaN                         NaN
13                         NaN                         NaN
14                         NaN                         NaN
15                         NaN                         NaN
16                         NaN                         NaN
17                         NaN                         NaN
18                         NaN                         NaN
19                4.196984e+09  2017-07-19 00:47:34 +0000
20                         NaN                         NaN
21                         NaN                         NaN
22                         NaN                         NaN
23                         NaN                         NaN
24                         NaN                         NaN
25                         NaN                         NaN
26                         NaN                         NaN
27                         NaN                         NaN
28                         NaN                         NaN
29                         NaN                         NaN
...                        ...                         ...
2326                       NaN                         NaN
2327                       NaN                         NaN
2328                       NaN                         NaN
2329                       NaN                         NaN
2330                       NaN                         NaN
2331                       NaN                         NaN
2332                       NaN                         NaN
2333                       NaN                         NaN
2334                       NaN                         NaN
2335                       NaN                         NaN
2336                       NaN                         NaN
2337                       NaN                         NaN
2338                       NaN                         NaN
2339                       NaN                         NaN
2340                       NaN                         NaN
```

```
2341                      NaN                      NaN
2342                      NaN                      NaN
2343                      NaN                      NaN
2344                      NaN                      NaN
2345                      NaN                      NaN
2346                      NaN                      NaN
2347                      NaN                      NaN
2348                      NaN                      NaN
2349                      NaN                      NaN
2350                      NaN                      NaN
2351                      NaN                      NaN
2352                      NaN                      NaN
2353                      NaN                      NaN
2354                      NaN                      NaN
2355                      NaN                      NaN

                                           expanded_urls  rating_numerator  \
0     https://twitter.com/dog_rates/status/892420643...                13
1     https://twitter.com/dog_rates/status/892177421...                13
2     https://twitter.com/dog_rates/status/891815181...                12
3     https://twitter.com/dog_rates/status/891689557...                13
4     https://twitter.com/dog_rates/status/891327558...                12
5     https://twitter.com/dog_rates/status/891087950...                13
6     https://gofundme.com/ydvmve-surgery-for-jax,ht...                13
7     https://twitter.com/dog_rates/status/890729181...                13
8     https://twitter.com/dog_rates/status/890609185...                13
9     https://twitter.com/dog_rates/status/890240255...                14
10    https://twitter.com/dog_rates/status/890006608...                13
11    https://twitter.com/dog_rates/status/889880896...                13
12    https://twitter.com/dog_rates/status/889665388...                13
13    https://twitter.com/dog_rates/status/889638837...                12
14    https://twitter.com/dog_rates/status/889531135...                13
15    https://twitter.com/dog_rates/status/889278841...                13
16    https://twitter.com/dog_rates/status/888917238...                12
17    https://twitter.com/dog_rates/status/888804989...                13
18    https://twitter.com/dog_rates/status/888554962...                13
19    https://twitter.com/dog_rates/status/887473957...                13
20    https://twitter.com/dog_rates/status/888078434...                12
21    https://twitter.com/dog_rates/status/887705289...                13
22    https://twitter.com/dog_rates/status/887517139...                14
23    https://twitter.com/dog_rates/status/887473957...                13
24    https://twitter.com/dog_rates/status/887343217...                13
25    https://twitter.com/dog_rates/status/887101392...                12
26    https://twitter.com/dog_rates/status/886983233...                13
27    https://www.gofundme.com/mingusneedsus,https:/...                13
28    https://twitter.com/dog_rates/status/886680336...                13
29    https://twitter.com/dog_rates/status/886366144...                12
...                                                 ...               ...
```

```
2326  https://twitter.com/dog_rates/status/666411507...                2
2327  https://twitter.com/dog_rates/status/666407126...                7
2328  https://twitter.com/dog_rates/status/666396247...                9
2329  https://twitter.com/dog_rates/status/666373753...               11
2330  https://twitter.com/dog_rates/status/666362758...                6
2331  https://twitter.com/dog_rates/status/666353288...                8
2332  https://twitter.com/dog_rates/status/666345417...               10
2333  https://twitter.com/dog_rates/status/666337882...                9
2334  https://twitter.com/dog_rates/status/666293911...                3
2335  https://twitter.com/dog_rates/status/666287406...                1
2336  https://twitter.com/dog_rates/status/666273097...               11
2337  https://twitter.com/dog_rates/status/666268910...               10
2338  https://twitter.com/dog_rates/status/666104133...                1
2339  https://twitter.com/dog_rates/status/666102155...               11
2340  https://twitter.com/dog_rates/status/666099513...                8
2341  https://twitter.com/dog_rates/status/666094000...                9
2342  https://twitter.com/dog_rates/status/666082916...                6
2343  https://twitter.com/dog_rates/status/666073100...               10
2344  https://twitter.com/dog_rates/status/666071193...                9
2345  https://twitter.com/dog_rates/status/666063827...               10
2346  https://twitter.com/dog_rates/status/666058600...                8
2347  https://twitter.com/dog_rates/status/666057090...                9
2348  https://twitter.com/dog_rates/status/666055525...               10
2349  https://twitter.com/dog_rates/status/666051853...                2
2350  https://twitter.com/dog_rates/status/666050758...               10
2351  https://twitter.com/dog_rates/status/666049248...                5
2352  https://twitter.com/dog_rates/status/666044226...                6
2353  https://twitter.com/dog_rates/status/666033412...                9
2354  https://twitter.com/dog_rates/status/666029285...                7
2355  https://twitter.com/dog_rates/status/666020888...                8
```

	rating_denominator	name	doggo	floofer	pupper	puppo
0	10	Phineas	None	None	None	None
1	10	Tilly	None	None	None	None
2	10	Archie	None	None	None	None
3	10	Darla	None	None	None	None
4	10	Franklin	None	None	None	None
5	10	None	None	None	None	None
6	10	Jax	None	None	None	None
7	10	None	None	None	None	None
8	10	Zoey	None	None	None	None
9	10	Cassie	doggo	None	None	None
10	10	Koda	None	None	None	None
11	10	Bruno	None	None	None	None
12	10	None	None	None	None	puppo
13	10	Ted	None	None	None	None
14	10	Stuart	None	None	None	puppo
15	10	Oliver	None	None	None	None

16	10	Jim	None	None	None	None
17	10	Zeke	None	None	None	None
18	10	Ralphus	None	None	None	None
19	10	Canela	None	None	None	None
20	10	Gerald	None	None	None	None
21	10	Jeffrey	None	None	None	None
22	10	such	None	None	None	None
23	10	Canela	None	None	None	None
24	10	None	None	None	None	None
25	10	None	None	None	None	None
26	10	Maya	None	None	None	None
27	10	Mingus	None	None	None	None
28	10	Derek	None	None	None	None
29	10	Roscoe	None	None	pupper	None
...	...	...	...	...	...	...
2326	10	quite	None	None	None	None
2327	10	a	None	None	None	None
2328	10	None	None	None	None	None
2329	10	None	None	None	None	None
2330	10	None	None	None	None	None
2331	10	None	None	None	None	None
2332	10	None	None	None	None	None
2333	10	an	None	None	None	None
2334	10	a	None	None	None	None
2335	2	an	None	None	None	None
2336	10	None	None	None	None	None
2337	10	None	None	None	None	None
2338	10	None	None	None	None	None
2339	10	None	None	None	None	None
2340	10	None	None	None	None	None
2341	10	None	None	None	None	None
2342	10	None	None	None	None	None
2343	10	None	None	None	None	None
2344	10	None	None	None	None	None
2345	10	the	None	None	None	None
2346	10	the	None	None	None	None
2347	10	a	None	None	None	None
2348	10	a	None	None	None	None
2349	10	an	None	None	None	None
2350	10	a	None	None	None	None
2351	10	None	None	None	None	None
2352	10	a	None	None	None	None
2353	10	a	None	None	None	None
2354	10	a	None	None	None	None
2355	10	None	None	None	None	None

[2356 rows x 17 columns]

```
In [11]: archive.info()

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

** The file twitter-archive-enhanced.csv successfully stored in archive data frame, it has 17 columns and 2356 entries **

```
In [4]: # Using Requests library to download a file then store it in a tsv file
        url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictic
        response = requests.get(url)

        with open(url.split('/')[-1], mode = 'wb') as outfile:
            outfile.write(response.content)

        # Read the downloaded file into a dataframe 'images'
        images = pd.read_csv('image-predictions.tsv', sep = '\t', encoding = 'utf-8')
        # Quick check to the file content and structure
        images
```

Out[4]:

	tweet_id	jpg_url	\
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	
5	666050758794694657	https://pbs.twimg.com/media/CT5Jof1WUAEuVxN.jpg	
6	666051853826850816	https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg	

7	666055525042405380	https://pbs.twimg.com/media/CT5N9tpXIAAifs1.jpg
8	666057090499244032	https://pbs.twimg.com/media/CT5PY9OWoAAQGLo.jpg
9	666058600524156928	https://pbs.twimg.com/media/CT5Qw94XAAA_2dP.jpg
10	666063827256086533	https://pbs.twimg.com/media/CT5Vg_wXIAAXfnj.jpg
11	666071193221509120	https://pbs.twimg.com/media/CT5cN_3WEAAlOoZ.jpg
12	666073100786774016	https://pbs.twimg.com/media/CT5d9DZXAAALcwe.jpg
13	666082916733198337	https://pbs.twimg.com/media/CT5m4VGWEAAtKc8.jpg
14	666094000022159362	https://pbs.twimg.com/media/CT5w9gUW4AAsBNN.jpg
15	666099513787052032	https://pbs.twimg.com/media/CT51-JJUEAA6hV8.jpg
16	666102155909144576	https://pbs.twimg.com/media/CT54YGiWUAEZnoK.jpg
17	666104133288665088	https://pbs.twimg.com/media/CT56LSZWoAAlJj2.jpg
18	666268910803644416	https://pbs.twimg.com/media/CT8QCd1WEAADXws.jpg
19	666273097616637952	https://pbs.twimg.com/media/CT8T1mtUwAA3aqm.jpg
20	666287406224695296	https://pbs.twimg.com/media/CT8g3BpUEAAuFjg.jpg
21	666293911632134144	https://pbs.twimg.com/media/CT8mx7KW4AEQu8N.jpg
22	666337882303524864	https://pbs.twimg.com/media/CT9OwFIWEAMuRje.jpg
23	666345417576210432	https://pbs.twimg.com/media/CT9Vn7PWoAA_ZCM.jpg
24	666353288456101888	https://pbs.twimg.com/media/CT9cx0tUEAAhNN_.jpg
25	666362758909284353	https://pbs.twimg.com/media/CT9lXGsUcAAyUFt.jpg
26	666373753744588802	https://pbs.twimg.com/media/CT9vZEYWUAAlZO5.jpg
27	666396247373291520	https://pbs.twimg.com/media/CT-D2ZHWIAA3gK1.jpg
28	666407126856765440	https://pbs.twimg.com/media/CT-NvwmW4AAugGZ.jpg
29	666411507551481857	https://pbs.twimg.com/media/CT-RugiWIAELEaq.jpg
...	...	...
2045	886366144734445568	https://pbs.twimg.com/media/DE0BTnQUwAApKEH.jpg
2046	886680336477933568	https://pbs.twimg.com/media/DE4fEDzWAAAyHMM.jpg
2047	886736880519319552	https://pbs.twimg.com/media/DE5Se8FXcAAJFx4.jpg
2048	886983233522544640	https://pbs.twimg.com/media/DE8yicJW0AAAvBJ.jpg
2049	887101392804085760	https://pbs.twimg.com/media/DE-eAq6UwAA-jaE.jpg
2050	887343217045368832	https://pbs.twimg.com/ext_tw_video_thumb/88734...
2051	887473957103951883	https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg
2052	887517139158093824	https://pbs.twimg.com/ext_tw_video_thumb/88751...
2053	887705289381826560	https://pbs.twimg.com/media/DFHDQBbXgAEqY7t.jpg
2054	888078434458587136	https://pbs.twimg.com/media/DFMWn56WsAAkA7B.jpg
2055	888202515573088257	https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg
2056	888554962724278272	https://pbs.twimg.com/media/DFTH_O-UQAACu2O.jpg
2057	888804989199671297	https://pbs.twimg.com/media/DFWra-3VYAA2piG.jpg
2058	888917238123831296	https://pbs.twimg.com/media/DFYRgsOUQAARGhO.jpg
2059	889278841981685760	https://pbs.twimg.com/ext_tw_video_thumb/88927...
2060	889531135344209921	https://pbs.twimg.com/media/DFg_2PVW0AEHN3p.jpg
2061	889638837579907072	https://pbs.twimg.com/media/DFihzFfXsAYGDPR.jpg
2062	889665388333682689	https://pbs.twimg.com/media/DFi579UWsAAatzw.jpg
2063	889880896479866881	https://pbs.twimg.com/media/DFl99B1WsAITKsg.jpg
2064	890006608113172480	https://pbs.twimg.com/media/DFnwSY4WAAAMliS.jpg
2065	890240255349198849	https://pbs.twimg.com/media/DFrEyVuW0AAO3t9.jpg
2066	890609185150312448	https://pbs.twimg.com/media/DFwUU__XcAEpyXI.jpg
2067	890729181411237888	https://pbs.twimg.com/media/DFyBahAVwAAhUTd.jpg
2068	890971913173991426	https://pbs.twimg.com/media/DF1eOmZXUAALUcq.jpg

```
2069  891087950875897856    https://pbs.twimg.com/media/DF3HwyEWsAABqE6.jpg
2070  891327558926688256    https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg
2071  891689557279858688    https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg
2072  891815181378084864    https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg
2073  892177421306343426    https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg
2074  892420643555336193    https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg
```

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

2054	1	French_bulldog	0.995026	True
2055	2	Pembroke	0.809197	True
2056	3	Siberian_husky	0.700377	True
2057	1	golden_retriever	0.469760	True
2058	1	golden_retriever	0.714719	True
2059	1	whippet	0.626152	True
2060	1	golden_retriever	0.953442	True
2061	1	French_bulldog	0.991650	True
2062	1	Pembroke	0.966327	True
2063	1	French_bulldog	0.377417	True
2064	1	Samoyed	0.957979	True
2065	1	Pembroke	0.511319	True
2066	1	Irish_terrier	0.487574	True
2067	2	Pomeranian	0.566142	True
2068	1	Appenzeller	0.341703	True
2069	1	Chesapeake_Bay_retriever	0.425595	True
2070	2	basset	0.555712	True
2071	1	paper_towel	0.170278	False
2072	1	Chihuahua	0.716012	True
2073	1	Chihuahua	0.323581	True
2074	1	orange	0.097049	False

	p2	p2_conf	p2_dog	p3	\
0	collie	0.156665	True	Shetland_sheepdog	
1	miniature_pinscher	0.074192	True	Rhodesian_ridgeback	
2	malinois	0.138584	True	bloodhound	
3	redbone	0.360687	True	miniature_pinscher	
4	Rottweiler	0.243682	True	Doberman	
5	English_springer	0.263788	True	Greater_Swiss_Mountain_dog	
6	mud_turtle	0.045885	False	terrapin	
7	Tibetan_mastiff	0.058279	True	fur_coat	
8	shopping_basket	0.014594	False	golden_retriever	
9	komondor	0.192305	True	soft-coated_wheaten_terrier	
10	Tibetan_mastiff	0.093718	True	Labrador_retriever	
11	Yorkshire_terrier	0.174201	True	Pekinese	
12	English_foxhound	0.175382	True	Ibizan_hound	
13	bull_mastiff	0.404722	True	French_bulldog	
14	German_shepherd	0.078260	True	malinois	
15	Shih-Tzu	0.166192	True	Dandie_Dinmont	
16	Newfoundland	0.149842	True	borzoi	
17	cock	0.033919	False	partridge	
18	desk	0.085547	False	bookcase	
19	toy_terrier	0.111884	True	basenji	
20	toy_poodle	0.063064	True	miniature_poodle	
21	otter	0.015250	False	great_grey_owl	
22	Newfoundland	0.278407	True	groenendael	
23	Chesapeake_Bay_retriever	0.054787	True	Labrador_retriever	
24	Siberian_husky	0.147655	True	Eskimo_dog	

25	skunk	0.002402	False	hamster
26	Afghan_hound	0.259551	True	briard
27	toy_terrier	0.009397	True	papillon
28	bloodhound	0.244220	True	flat-coated_retriever
29	barracouta	0.271485	False	gar
...	...	...	...	...
2045	Chihuahua	0.000361	True	Boston_bull
2046	sports_car	0.139952	False	car_wheel
2047	Great_Pyrenees	0.186136	True	Dandie_Dinmont
2048	toy_terrier	0.143528	True	can_opener
2049	Eskimo_dog	0.035029	True	Staffordshire_bullterrier
2050	sea_lion	0.275645	False	Weimaraner
2051	Rhodesian_ridgeback	0.054950	True	beagle
2052	tow_truck	0.029175	False	shopping_cart
2053	redbone	0.087582	True	Weimaraner
2054	pug	0.000932	True	bull_mastiff
2055	Rhodesian_ridgeback	0.054950	True	beagle
2056	Eskimo_dog	0.166511	True	malamute
2057	Labrador_retriever	0.184172	True	English_setter
2058	Tibetan_mastiff	0.120184	True	Labrador_retriever
2059	borzoi	0.194742	True	Saluki
2060	Labrador_retriever	0.013834	True	redbone
2061	boxer	0.002129	True	Staffordshire_bullterrier
2062	Cardigan	0.027356	True	basenji
2063	Labrador_retriever	0.151317	True	muzzle
2064	Pomeranian	0.013884	True	chow
2065	Cardigan	0.451038	True	Chihuahua
2066	Irish_setter	0.193054	True	Chesapeake_Bay_retriever
2067	Eskimo_dog	0.178406	True	Pembroke
2068	Border_collie	0.199287	True	ice_lolly
2069	Irish_terrier	0.116317	True	Indian_elephant
2070	English_springer	0.225770	True	German_short-haired_pointer
2071	Labrador_retriever	0.168086	True	spatula
2072	malamute	0.078253	True	kelpie
2073	Pekinese	0.090647	True	papillon
2074	bagel	0.085851	False	banana

	p3_conf	p3_dog
0	0.061428	True
1	0.072010	True
2	0.116197	True
3	0.222752	True
4	0.154629	True
5	0.016199	True
6	0.017885	False
7	0.054449	False
8	0.007959	True
9	0.082086	True

```
10    0.072427   True
11    0.109454   True
12    0.097471   True
13    0.048960   True
14    0.075628   True
15    0.089688   True
16    0.133649   True
17    0.000052  False
18    0.079480  False
19    0.111152   True
20    0.025581   True
21    0.013207  False
22    0.102643   True
23    0.014241   True
24    0.093412   True
25    0.000461  False
26    0.206803   True
27    0.004577   True
28    0.173810   True
29    0.189945  False
...        ...    ...
2045  0.000076   True
2046  0.044173  False
2047  0.086346   True
2048  0.032253  False
2049  0.029705   True
2050  0.134203   True
2051  0.038915   True
2052  0.026321  False
2053  0.026236   True
2054  0.000903   True
2055  0.038915   True
2056  0.111411   True
2057  0.073482   True
2058  0.105506   True
2059  0.027351   True
2060  0.007958   True
2061  0.001498   True
2062  0.004633   True
2063  0.082981  False
2064  0.008167   True
2065  0.029248   True
2066  0.118184   True
2067  0.076507   True
2068  0.193548  False
2069  0.076902  False
2070  0.175219   True
2071  0.040836  False
```

```
2072  0.031379   True
2073  0.068957   True
2074  0.076110  False

[2075 rows x 12 columns]
```

```
In [10]: images.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
```

The file image-predictions.tsv successfully downloaded and stored in images data frame, it has 12 columns and 2075 entries

```
In [5]: # Autontification to twetter API

        # Generate your own at https://apps.twitter.com/app
        # CONSUMER_KEY = 'Consumer Key (API key)'
        # CONSUMER_SECRET = 'Consumer Secret (API Secret)'
        # OAUTH_TOKEN = 'Access Token'
        # OAUTH_TOKEN_SECRET = 'Access Token Secret'

        consumer_key = ''
        consumer_secret = ''
        access_token = ''
        access_token_secret = ''

        auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
        auth.set_access_token(access_token, access_token_secret)

        # Construct the API instance
        api = tweepy.API(auth,
                         parser = tweepy.parsers.JSONParser(), # Parse the result to Json Object
```

```
                        wait_on_rate_limit = True, # Automatically wait for rate limits to repl
                        wait_on_rate_limit_notify = True) # Print a notification when Tweepy is
```

In [8]:
```
# Liste where we will store the dictionaries of our result
df_list = []
# Liste frame where we will store the tweet_id of the errors
error_list = []

# Calculate the time of excution
start = time.time()

# Get the tweet object for all the teweets in archive dataframe
for tweet_id in archive['tweet_id']:
    try:
        page = api.get_status(tweet_id, tweet_mode = 'extended')
        # Print one page to look at the structure of the returned file
        # and the names of attributes
        # print(json.dumps(page, indent = 4))
        #break

        favorites = page['favorite_count'] # How many favorites the tweet had
        retweets = page['retweet_count'] # Count of the retweet
        user_followers = page['user']['followers_count'] # How many followers the user h
        user_favourites = page['user']['favourites_count'] # How many favorites the user
        date_time = page['created_at'] # The date and time of the creation

        df_list.append({'tweet_id': int(tweet_id),
                        'favorites': int(favorites),
                        'retweets': int(retweets),
                        'user_followers': int(user_followers),
                        'user_favourites': int(user_favourites),
                        'date_time': pd.to_datetime(date_time)})

    # Catch the exceptions of the TweepError
    except Exception as e:
        print(str(tweet_id)+ " _ " + str(e))
        error_list.append(tweet_id)

# Calculate the time of excution
end = time.time()
print(end - start)
# 888202515573088257 _ [{'code': 144, 'message': 'No status found with that ID.'}]
# 873697596434513921 _ [{'code': 144, 'message': 'No status found with that ID.'}]
# 869988702071779329 _ [{'code': 144, 'message': 'No status found with that ID.'}]
# 861769973181624320 _ [{'code': 144, 'message': 'No status found with that ID.'}]
# 842892208864923648 _ [{'code': 144, 'message': 'No status found with that ID.'}]
# 802247111496568832 _ [{'code': 144, 'message': 'No status found with that ID.'}]
# 775096608509886464 _ [{'code': 144, 'message': 'No status found with that ID.'}]
```

```
        # Rate limit reached. Sleeping for: 212
        # Rate limit reached. Sleeping for: 532
        # 1980.119999885559
```

```
888202515573088257 _ [{'code': 144, 'message': 'No status found with that ID.'}]
873697596434513921 _ [{'code': 144, 'message': 'No status found with that ID.'}]
869988702071779329 _ [{'code': 144, 'message': 'No status found with that ID.'}]
861769973181624320 _ [{'code': 144, 'message': 'No status found with that ID.'}]
842892208864923648 _ [{'code': 144, 'message': 'No status found with that ID.'}]
802247111496568832 _ [{'code': 144, 'message': 'No status found with that ID.'}]
775096608509886464 _ [{'code': 144, 'message': 'No status found with that ID.'}]
Rate limit reached. Sleeping for: 212
Rate limit reached. Sleeping for: 532
1980.119999885559
```

```
In [9]: # lengh of the result
        print("The lengh of the result", len(df_list))
        # The tweet_id of the errors
        print("The lengh of the errors", len(error_list))
```

```
The lengh of the result 2349
The lengh of the errors 7
```

From the above results: - We reached the limit of the tweepy API twice but wait_on_rate_limit automatically wait for rate limits to replenish and wait_on_rate_limit_notify print a notification when Tweepy is waiting - The total time was about 1980 seconds (~ 33 min) - We could get 2349 tweet_id correctly with 7 errors (we will query those 7 errors separately)

```
In [12]: # We repeat the same operation for the tweet_ids that we coudln't get and append the re
         ee_list = []
         for e in error_list:
             try:
                 favorites = page['favorite_count']
                 retweets = page['retweet_count']
                 user_followers = page['user']['followers_count']
                 user_favourites = page['user']['favourites_count']
                 date_time = page['created_at']

                 df_list.append({'tweet_id': int(tweet_id),
                                 'favorites': int(favorites),
                                 'retweets': int(retweets),
                                 'user_followers': int(user_followers),
                                 'user_favourites': int(user_favourites),
                                 'date_time': pd.to_datetime(date_time)})

             except Exception:
                 print(str(tweet_id)+ " _ " + str(e))
                 ee_list.append(e)
```

In [15]:
```
# We can see that now the 7 errors saved in the list
# lengh of the result
print("The lengh of the result after Querying the errors separately", len(df_list))
```

```
The lengh of the result after Querying the errors separately 2356
```

In [16]:
```
# Create DataFrames from list of dictionaries
json_tweets = pd.DataFrame(df_list, columns = ['tweet_id', 'favorites', 'retweets',
                                               'user_followers', 'user_favourites', 'da

# Save the dataFrame in file
json_tweets.to_csv('tweet_json.txt', encoding = 'utf-8', index=False)
```

In [18]:
```
# Read the saved tweet_json.txt file into a dataframe
json_tweets = pd.read_csv('tweet_json.txt', encoding = 'utf-8')
json_tweets
```

Out[18]:

	tweet_id	favorites	retweets	user_followers	\
0	892420643555336193	39361	8793	4484668	
1	892177421306343426	33693	6447	4484669	
2	891815181378084864	25383	4272	4484669	
3	891689557279858688	42731	8880	4484669	
4	891327558926688256	40897	9661	4484669	
5	891087950875897856	20496	3217	4484669	
6	890971913173991426	12026	2130	4484669	
7	890729181411237888	66500	19466	4484669	
8	890609185150312448	28129	4370	4484669	
9	890240255349198849	32365	7632	4484669	
10	890006608113172480	31022	7535	4484669	
11	889880896479866881	28134	5091	4484669	
12	889665388333682689	38609	8462	4484669	
13	889638837579907072	27548	4673	4484669	
14	889531135344209921	15302	2295	4484669	
15	889278841981685760	25648	5597	4484669	
16	888917238123831296	29474	4647	4484669	
17	888804989199671297	25947	4501	4484670	
18	888554962724278272	20213	3700	4484670	
19	888078434458587136	22081	3612	4484670	
20	887705289381826560	30613	5552	4484670	
21	887517139158093824	46813	11991	4484670	
22	887473957103951883	70100	18785	4484670	
23	887343217045368832	34123	10664	4484670	
24	887101392804085760	30966	6118	4484670	
25	886983233522544640	35675	7995	4484670	
26	886736880519319552	12254	3394	4484670	
27	886680336477933568	22733	4583	4484670	
28	886366144734445568	21433	3274	4484670	
29	886267009285017600	117	4	4484670	

```
...                   ...         ...       ...             ...
2326  666337882303524864         203        95         4484780
2327  666293911632134144         515       365         4484780
2328  666287406224695296         152        70         4484781
2329  666273097616637952         182        80         4484781
2330  666268910803644416         108        36         4484781
2331  666104133288665088       14650      6804         4484781
2332  666102155909144576          81        14         4484781
2333  666099513787052032         161        72         4484781
2334  666094000022159362         167        77         4484781
2335  666082916733198337         121        46         4484781
2336  666073100786774016         333       172         4484781
2337  666071193221509120         154        65         4484781
2338  666063827256086533         492       226         4484781
2339  666058600524156928         117        60         4484781
2340  666057090499244032         304       145         4484781
2341  666055525042405380         448       260         4484781
2342  666051853826850816        1246       873         4484782
2343  666050758794694657         136        59         4484782
2344  666049248165822465         111        40         4484784
2345  666044226329800704         307       144         4484784
2346  666033412701032449         128        46         4484784
2347  666029285002620928         132        47         4484784
2348  666020888022790149        2532       527         4484784
2349  666020888022790149        2532       527         4484784
2350  666020888022790149        2532       527         4484784
2351  666020888022790149        2532       527         4484784
2352  666020888022790149        2532       527         4484784
2353  666020888022790149        2532       527         4484784
2354  666020888022790149        2532       527         4484784
2355  666020888022790149        2532       527         4484784

      user_favourites            date_time
0              124451  2017-08-01 16:23:56
1              124451  2017-08-01 00:17:27
2              124451  2017-07-31 00:18:03
3              124451  2017-07-30 15:58:51
4              124451  2017-07-29 16:00:24
5              124451  2017-07-29 00:08:17
6              124451  2017-07-28 16:27:12
7              124451  2017-07-28 00:22:40
8              124451  2017-07-27 16:25:51
9              124451  2017-07-26 15:59:51
10             124451  2017-07-26 00:31:25
11             124451  2017-07-25 16:11:53
12             124451  2017-07-25 01:55:32
13             124451  2017-07-25 00:10:02
14             124451  2017-07-24 17:02:04
```

```
15             124451  2017-07-24 00:19:32
16             124451  2017-07-23 00:22:39
17             124451  2017-07-22 16:56:37
18             124451  2017-07-22 00:23:06
19             124451  2017-07-20 16:49:33
20             124451  2017-07-19 16:06:48
21             124451  2017-07-19 03:39:09
22             124451  2017-07-19 00:47:34
23             124451  2017-07-18 16:08:03
24             124451  2017-07-18 00:07:08
25             124451  2017-07-17 16:17:36
26             124451  2017-07-16 23:58:41
27             124451  2017-07-16 20:14:00
28             124451  2017-07-15 23:25:31
29             124451  2017-07-15 16:51:35
...               ...                  ...
2326           124450  2015-11-16 19:31:45
2327           124450  2015-11-16 16:37:02
2328           124450  2015-11-16 16:11:11
2329           124450  2015-11-16 15:14:19
2330           124450  2015-11-16 14:57:41
2331           124450  2015-11-16 04:02:55
2332           124450  2015-11-16 03:55:04
2333           124450  2015-11-16 03:44:34
2334           124450  2015-11-16 03:22:39
2335           124450  2015-11-16 02:38:37
2336           124450  2015-11-16 01:59:36
2337           124450  2015-11-16 01:52:02
2338           124450  2015-11-16 01:22:45
2339           124450  2015-11-16 01:01:59
2340           124450  2015-11-16 00:55:59
2341           124450  2015-11-16 00:49:46
2342           124450  2015-11-16 00:35:11
2343           124450  2015-11-16 00:30:50
2344           124450  2015-11-16 00:24:50
2345           124450  2015-11-16 00:04:52
2346           124450  2015-11-15 23:21:54
2347           124450  2015-11-15 23:05:30
2348           124450  2015-11-15 22:32:08
2349           124450  2015-11-15 22:32:08
2350           124450  2015-11-15 22:32:08
2351           124450  2015-11-15 22:32:08
2352           124450  2015-11-15 22:32:08
2353           124450  2015-11-15 22:32:08
2354           124450  2015-11-15 22:32:08
2355           124450  2015-11-15 22:32:08

[2356 rows x 6 columns]
```

```
In [19]: json_tweets.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 6 columns):
tweet_id           2356 non-null int64
favorites          2356 non-null int64
retweets           2356 non-null int64
user_followers     2356 non-null int64
user_favourites    2356 non-null int64
date_time          2356 non-null object
dtypes: int64(5), object(1)
memory usage: 110.5+ KB
```

The file tweet_json.txt successfully saved in our working directory contains the result of the API Querying then stored in json_tweets data frame, it has 6 columns and 2356 entries

1.3.1 Gather: Summary

Gathering is the first step in the data wrangling process. We could finish the high-level gathering process:

- Obtaining data
  - Getting data from an existing file (twitter-archive-enhanced.csv) *Reading from csv file using pandas*
  - Downloading a file from the internet (image-predictions.tsv) *Downloading file using requests*
  - Querying an API (tweet_json.txt) *Get JSON object of all the tweet_ids using Tweepy*
- Importing that data into our programming environment (Jupyter Notebook)

1.4 Assess

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues will be our newt step. We will detect and document at quality issues and tidiness issues.

```
In [21]: # Print all archive dataset to assess it visually
         archive
```

```
Out[21]:                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
         0     892420643555336193                    NaN                  NaN
         1     892177421306343426                    NaN                  NaN
         2     891815181378084864                    NaN                  NaN
         3     891689557279858688                    NaN                  NaN
         4     891327558926688256                    NaN                  NaN
         5     891087950875897856                    NaN                  NaN
         6     890971913173991426                    NaN                  NaN
```

7	890729181411237888	NaN	NaN
8	890609185150312448	NaN	NaN
9	890240255349198849	NaN	NaN
10	890006608113172480	NaN	NaN
11	889880896479866881	NaN	NaN
12	889665388333682689	NaN	NaN
13	889638837579907072	NaN	NaN
14	889531135344209921	NaN	NaN
15	889278841981685760	NaN	NaN
16	888917238123831296	NaN	NaN
17	888804989199671297	NaN	NaN
18	888554962724278272	NaN	NaN
19	888202515573088257	NaN	NaN
20	888078434458587136	NaN	NaN
21	887705289381826560	NaN	NaN
22	887517139158093824	NaN	NaN
23	887473957103951883	NaN	NaN
24	887343217045368832	NaN	NaN
25	887101392804085760	NaN	NaN
26	886983233522544640	NaN	NaN
27	886736880519319552	NaN	NaN
28	886680336477933568	NaN	NaN
29	886366144734445568	NaN	NaN
...	...	...	...
2326	666411507551481857	NaN	NaN
2327	666407126856765440	NaN	NaN
2328	666396247373291520	NaN	NaN
2329	666373753744588802	NaN	NaN
2330	666362758909284353	NaN	NaN
2331	666353288456101888	NaN	NaN
2332	666345417576210432	NaN	NaN
2333	666337882303524864	NaN	NaN
2334	666293911632134144	NaN	NaN
2335	666287406224695296	NaN	NaN
2336	666273097616637952	NaN	NaN
2337	666268910803644416	NaN	NaN
2338	666104133288665088	NaN	NaN
2339	666102155909144576	NaN	NaN
2340	666099513787052032	NaN	NaN
2341	666094000022159362	NaN	NaN
2342	666082916733198337	NaN	NaN
2343	666073100786774016	NaN	NaN
2344	666071193221509120	NaN	NaN
2345	666063827256086533	NaN	NaN
2346	666058600524156928	NaN	NaN
2347	666057090499244032	NaN	NaN
2348	666055525042405380	NaN	NaN
2349	666051853826850816	NaN	NaN

```
2350  666050758794694657                    NaN                    NaN
2351  666049248165822465                    NaN                    NaN
2352  666044226329800704                    NaN                    NaN
2353  666033412701032449                    NaN                    NaN
2354  666029285002620928                    NaN                    NaN
2355  666020888022790149                    NaN                    NaN

                      timestamp  \
0     2017-08-01 16:23:56 +0000
1     2017-08-01 00:17:27 +0000
2     2017-07-31 00:18:03 +0000
3     2017-07-30 15:58:51 +0000
4     2017-07-29 16:00:24 +0000
5     2017-07-29 00:08:17 +0000
6     2017-07-28 16:27:12 +0000
7     2017-07-28 00:22:40 +0000
8     2017-07-27 16:25:51 +0000
9     2017-07-26 15:59:51 +0000
10    2017-07-26 00:31:25 +0000
11    2017-07-25 16:11:53 +0000
12    2017-07-25 01:55:32 +0000
13    2017-07-25 00:10:02 +0000
14    2017-07-24 17:02:04 +0000
15    2017-07-24 00:19:32 +0000
16    2017-07-23 00:22:39 +0000
17    2017-07-22 16:56:37 +0000
18    2017-07-22 00:23:06 +0000
19    2017-07-21 01:02:36 +0000
20    2017-07-20 16:49:33 +0000
21    2017-07-19 16:06:48 +0000
22    2017-07-19 03:39:09 +0000
23    2017-07-19 00:47:34 +0000
24    2017-07-18 16:08:03 +0000
25    2017-07-18 00:07:08 +0000
26    2017-07-17 16:17:36 +0000
27    2017-07-16 23:58:41 +0000
28    2017-07-16 20:14:00 +0000
29    2017-07-15 23:25:31 +0000
...                        ...
2326  2015-11-17 00:24:19 +0000
2327  2015-11-17 00:06:54 +0000
2328  2015-11-16 23:23:41 +0000
2329  2015-11-16 21:54:18 +0000
2330  2015-11-16 21:10:36 +0000
2331  2015-11-16 20:32:58 +0000
2332  2015-11-16 20:01:42 +0000
2333  2015-11-16 19:31:45 +0000
2334  2015-11-16 16:37:02 +0000
```

```
2335  2015-11-16 16:11:11 +0000
2336  2015-11-16 15:14:19 +0000
2337  2015-11-16 14:57:41 +0000
2338  2015-11-16 04:02:55 +0000
2339  2015-11-16 03:55:04 +0000
2340  2015-11-16 03:44:34 +0000
2341  2015-11-16 03:22:39 +0000
2342  2015-11-16 02:38:37 +0000
2343  2015-11-16 01:59:36 +0000
2344  2015-11-16 01:52:02 +0000
2345  2015-11-16 01:22:45 +0000
2346  2015-11-16 01:01:59 +0000
2347  2015-11-16 00:55:59 +0000
2348  2015-11-16 00:49:46 +0000
2349  2015-11-16 00:35:11 +0000
2350  2015-11-16 00:30:50 +0000
2351  2015-11-16 00:24:50 +0000
2352  2015-11-16 00:04:52 +0000
2353  2015-11-15 23:21:54 +0000
2354  2015-11-15 23:05:30 +0000
2355  2015-11-15 22:32:08 +0000

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

```
25    <a href="http://twitter.com/download/iphone" r...
26    <a href="http://twitter.com/download/iphone" r...
27    <a href="http://twitter.com/download/iphone" r...
28    <a href="http://twitter.com/download/iphone" r...
29    <a href="http://twitter.com/download/iphone" r...
...                                                 ...
2326  <a href="http://twitter.com/download/iphone" r...
2327  <a href="http://twitter.com/download/iphone" r...
2328  <a href="http://twitter.com/download/iphone" r...
2329  <a href="http://twitter.com/download/iphone" r...
2330  <a href="http://twitter.com/download/iphone" r...
2331  <a href="http://twitter.com/download/iphone" r...
2332  <a href="http://twitter.com/download/iphone" r...
2333  <a href="http://twitter.com/download/iphone" r...
2334  <a href="http://twitter.com/download/iphone" r...
2335  <a href="http://twitter.com/download/iphone" r...
2336  <a href="http://twitter.com/download/iphone" r...
2337  <a href="http://twitter.com/download/iphone" r...
2338  <a href="http://twitter.com/download/iphone" r...
2339  <a href="http://twitter.com/download/iphone" r...
2340  <a href="http://twitter.com/download/iphone" r...
2341  <a href="http://twitter.com/download/iphone" r...
2342  <a href="http://twitter.com/download/iphone" r...
2343  <a href="http://twitter.com/download/iphone" r...
2344  <a href="http://twitter.com/download/iphone" r...
2345  <a href="http://twitter.com/download/iphone" r...
2346  <a href="http://twitter.com/download/iphone" r...
2347  <a href="http://twitter.com/download/iphone" r...
2348  <a href="http://twitter.com/download/iphone" r...
2349  <a href="http://twitter.com/download/iphone" r...
2350  <a href="http://twitter.com/download/iphone" r...
2351  <a href="http://twitter.com/download/iphone" r...
2352  <a href="http://twitter.com/download/iphone" r...
2353  <a href="http://twitter.com/download/iphone" r...
2354  <a href="http://twitter.com/download/iphone" r...
2355  <a href="http://twitter.com/download/iphone" r...

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

10	This is Koda. He is a South Australian decksha...	NaN
11	This is Bruno. He is a service shark. Only get...	NaN
12	Here's a puppo that seems to be on the fence a...	NaN
13	This is Ted. He does his best. Sometimes that'...	NaN
14	This is Stuart. He's sporting his favorite fan...	NaN
15	This is Oliver. You're witnessing one of his m...	NaN
16	This is Jim. He found a fren. Taught him how t...	NaN
17	This is Zeke. He has a new stick. Very proud o...	NaN
18	This is Ralphus. He's powering up. Attempting ...	NaN
19	RT @dog_rates: This is Canela. She attempted s...	8.874740e+17
20	This is Gerald. He was just told he didn't get...	NaN
21	This is Jeffrey. He has a monopoly on the pool...	NaN
22	I've yet to rate a Venezuelan Hover Wiener. Th...	NaN
23	This is Canela. She attempted some fancy porch...	NaN
24	You may not have known you needed to see this ...	NaN
25	This... is a Jubilant Antarctic House Bear. We...	NaN
26	This is Maya. She's very shy. Rarely leaves he...	NaN
27	This is Mingus. He's a wonderful father to his...	NaN
28	This is Derek. He's late for a dog meeting. 13...	NaN
29	This is Roscoe. Another pupper fallen victim t...	NaN
...	...	...
2326	This is quite the dog. Gets really excited whe...	NaN
2327	This is a southern Vesuvius bumblegruff. Can d...	NaN
2328	Oh goodness. A super rare northeast Qdoba kang...	NaN
2329	Those are sunglasses and a jean jacket. 11/10 ...	NaN
2330	Unique dog here. Very small. Lives in containe...	NaN
2331	Here we have a mixed Asiago from the Galápagos...	NaN
2332	Look at this jokester thinking seat belt laws ...	NaN
2333	This is an extremely rare horned Parthenon. No...	NaN
2334	This is a funny dog. Weird toes. Won't come do...	NaN
2335	This is an Albanian 3 1/2 legged Episcopalian...	NaN
2336	Can take selfies 11/10 https://t.co/ws2AMaNwPW	NaN
2337	Very concerned about fellow dog trapped in com...	NaN
2338	Not familiar with this breed. No tail (weird)...	NaN
2339	Oh my. Here you are seeing an Adobe Setter giv...	NaN
2340	Can stand on stump for what seems like a while...	NaN
2341	This appears to be a Mongolian Presbyterian mi...	NaN
2342	Here we have a well-established sunblockerspan...	NaN
2343	Let's hope this flight isn't Malaysian (lol). ...	NaN
2344	Here we have a northern speckled Rhododendron...	NaN
2345	This is the happiest dog you will ever see. Ve...	NaN
2346	Here is the Rand Paul of retrievers folks! He'...	NaN
2347	My oh my. This is a rare blond Canadian terrie...	NaN
2348	Here is a Siberian heavily armored polar bear ...	NaN
2349	This is an odd dog. Hard on the outside but lo...	NaN
2350	This is a truly beautiful English Wilson Staff...	NaN
2351	Here we have a 1949 1st generation vulpix. Enj...	NaN
2352	This is a purebred Piers Morgan. Loves to Netf...	NaN

```
2353  Here is a very happy pup. Big fan of well-main...                NaN
2354  This is a western brown Mitsubishi terrier. Up...                NaN
2355  Here we have a Japanese Irish Setter. Lost eye...                 NaN

      retweeted_status_user_id retweeted_status_timestamp  \
0                          NaN                         NaN
1                          NaN                         NaN
2                          NaN                         NaN
3                          NaN                         NaN
4                          NaN                         NaN
5                          NaN                         NaN
6                          NaN                         NaN
7                          NaN                         NaN
8                          NaN                         NaN
9                          NaN                         NaN
10                         NaN                         NaN
11                         NaN                         NaN
12                         NaN                         NaN
13                         NaN                         NaN
14                         NaN                         NaN
15                         NaN                         NaN
16                         NaN                         NaN
17                         NaN                         NaN
18                         NaN                         NaN
19                4.196984e+09  2017-07-19 00:47:34 +0000
20                         NaN                         NaN
21                         NaN                         NaN
22                         NaN                         NaN
23                         NaN                         NaN
24                         NaN                         NaN
25                         NaN                         NaN
26                         NaN                         NaN
27                         NaN                         NaN
28                         NaN                         NaN
29                         NaN                         NaN
...                        ...                         ...
2326                       NaN                         NaN
2327                       NaN                         NaN
2328                       NaN                         NaN
2329                       NaN                         NaN
2330                       NaN                         NaN
2331                       NaN                         NaN
2332                       NaN                         NaN
2333                       NaN                         NaN
2334                       NaN                         NaN
2335                       NaN                         NaN
2336                       NaN                         NaN
2337                       NaN                         NaN
```

```
2338                        NaN                         NaN
2339                        NaN                         NaN
2340                        NaN                         NaN
2341                        NaN                         NaN
2342                        NaN                         NaN
2343                        NaN                         NaN
2344                        NaN                         NaN
2345                        NaN                         NaN
2346                        NaN                         NaN
2347                        NaN                         NaN
2348                        NaN                         NaN
2349                        NaN                         NaN
2350                        NaN                         NaN
2351                        NaN                         NaN
2352                        NaN                         NaN
2353                        NaN                         NaN
2354                        NaN                         NaN
2355                        NaN                         NaN

                                            expanded_urls  rating_numerator  \
0     https://twitter.com/dog_rates/status/892420643...                  13
1     https://twitter.com/dog_rates/status/892177421...                  13
2     https://twitter.com/dog_rates/status/891815181...                  12
3     https://twitter.com/dog_rates/status/891689557...                  13
4     https://twitter.com/dog_rates/status/891327558...                  12
5     https://twitter.com/dog_rates/status/891087950...                  13
6     https://gofundme.com/ydvmve-surgery-for-jax,ht...                  13
7     https://twitter.com/dog_rates/status/890729181...                  13
8     https://twitter.com/dog_rates/status/890609185...                  13
9     https://twitter.com/dog_rates/status/890240255...                  14
10    https://twitter.com/dog_rates/status/890006608...                  13
11    https://twitter.com/dog_rates/status/889880896...                  13
12    https://twitter.com/dog_rates/status/889665388...                  13
13    https://twitter.com/dog_rates/status/889638837...                  12
14    https://twitter.com/dog_rates/status/889531135...                  13
15    https://twitter.com/dog_rates/status/889278841...                  13
16    https://twitter.com/dog_rates/status/888917238...                  12
17    https://twitter.com/dog_rates/status/888804989...                  13
18    https://twitter.com/dog_rates/status/888554962...                  13
19    https://twitter.com/dog_rates/status/887473957...                  13
20    https://twitter.com/dog_rates/status/888078434...                  12
21    https://twitter.com/dog_rates/status/887705289...                  13
22    https://twitter.com/dog_rates/status/887517139...                  14
23    https://twitter.com/dog_rates/status/887473957...                  13
24    https://twitter.com/dog_rates/status/887343217...                  13
25    https://twitter.com/dog_rates/status/887101392...                  12
26    https://twitter.com/dog_rates/status/886983233...                  13
27    https://www.gofundme.com/mingusneedsus,https:/...                  13
```

```
28    https://twitter.com/dog_rates/status/886680336...                13
29    https://twitter.com/dog_rates/status/886366144...                12
...                                                  ...               ...
2326  https://twitter.com/dog_rates/status/666411507...                 2
2327  https://twitter.com/dog_rates/status/666407126...                 7
2328  https://twitter.com/dog_rates/status/666396247...                 9
2329  https://twitter.com/dog_rates/status/666373753...                11
2330  https://twitter.com/dog_rates/status/666362758...                 6
2331  https://twitter.com/dog_rates/status/666353288...                 8
2332  https://twitter.com/dog_rates/status/666345417...                10
2333  https://twitter.com/dog_rates/status/666337882...                 9
2334  https://twitter.com/dog_rates/status/666293911...                 3
2335  https://twitter.com/dog_rates/status/666287406...                 1
2336  https://twitter.com/dog_rates/status/666273097...                11
2337  https://twitter.com/dog_rates/status/666268910...                10
2338  https://twitter.com/dog_rates/status/666104133...                 1
2339  https://twitter.com/dog_rates/status/666102155...                11
2340  https://twitter.com/dog_rates/status/666099513...                 8
2341  https://twitter.com/dog_rates/status/666094000...                 9
2342  https://twitter.com/dog_rates/status/666082916...                 6
2343  https://twitter.com/dog_rates/status/666073100...                10
2344  https://twitter.com/dog_rates/status/666071193...                 9
2345  https://twitter.com/dog_rates/status/666063827...                10
2346  https://twitter.com/dog_rates/status/666058600...                 8
2347  https://twitter.com/dog_rates/status/666057090...                 9
2348  https://twitter.com/dog_rates/status/666055525...                10
2349  https://twitter.com/dog_rates/status/666051853...                 2
2350  https://twitter.com/dog_rates/status/666050758...                10
2351  https://twitter.com/dog_rates/status/666049248...                 5
2352  https://twitter.com/dog_rates/status/666044226...                 6
2353  https://twitter.com/dog_rates/status/666033412...                 9
2354  https://twitter.com/dog_rates/status/666029285...                 7
2355  https://twitter.com/dog_rates/status/666020888...                 8

      rating_denominator      name  doggo floofer  pupper  puppo
0                     10   Phineas   None    None    None   None
1                     10     Tilly   None    None    None   None
2                     10    Archie   None    None    None   None
3                     10     Darla   None    None    None   None
4                     10  Franklin   None    None    None   None
5                     10      None   None    None    None   None
6                     10       Jax   None    None    None   None
7                     10      None   None    None    None   None
8                     10      Zoey   None    None    None   None
9                     10    Cassie  doggo    None    None   None
10                    10      Koda   None    None    None   None
11                    10     Bruno   None    None    None   None
12                    10      None   None    None    None  puppo
```

13	10	Ted	None	None	None	None
14	10	Stuart	None	None	None	puppo
15	10	Oliver	None	None	None	None
16	10	Jim	None	None	None	None
17	10	Zeke	None	None	None	None
18	10	Ralphus	None	None	None	None
19	10	Canela	None	None	None	None
20	10	Gerald	None	None	None	None
21	10	Jeffrey	None	None	None	None
22	10	such	None	None	None	None
23	10	Canela	None	None	None	None
24	10	None	None	None	None	None
25	10	None	None	None	None	None
26	10	Maya	None	None	None	None
27	10	Mingus	None	None	None	None
28	10	Derek	None	None	None	None
29	10	Roscoe	None	None	pupper	None
...	...	...	...	...	...	...
2326	10	quite	None	None	None	None
2327	10	a	None	None	None	None
2328	10	None	None	None	None	None
2329	10	None	None	None	None	None
2330	10	None	None	None	None	None
2331	10	None	None	None	None	None
2332	10	None	None	None	None	None
2333	10	an	None	None	None	None
2334	10	a	None	None	None	None
2335	2	an	None	None	None	None
2336	10	None	None	None	None	None
2337	10	None	None	None	None	None
2338	10	None	None	None	None	None
2339	10	None	None	None	None	None
2340	10	None	None	None	None	None
2341	10	None	None	None	None	None
2342	10	None	None	None	None	None
2343	10	None	None	None	None	None
2344	10	None	None	None	None	None
2345	10	the	None	None	None	None
2346	10	the	None	None	None	None
2347	10	a	None	None	None	None
2348	10	a	None	None	None	None
2349	10	an	None	None	None	None
2350	10	a	None	None	None	None
2351	10	None	None	None	None	None
2352	10	a	None	None	None	None
2353	10	a	None	None	None	None
2354	10	a	None	None	None	None
2355	10	None	None	None	None	None

```
[2356 rows x 17 columns]
```

In [50]:
```
# Print some random expamles from columns values
# random number just to check if we can find something sucpisious
print(archive['text'][900])
print(archive['name'][12])
```

```
Meet Boston. He's worried because his tongue won't fit all the way in his mouth. 12/10 it'll be
None
```

In [69]:
```
# Assessing the data programmaticaly
archive.info()
archive.describe()
archive['rating_numerator'].value_counts()
archive['rating_denominator'].value_counts()
archive['name'].value_counts()
```

In [52]:
```
images
```

Out[52]:

	tweet_id	jpg_url
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg
5	666050758794694657	https://pbs.twimg.com/media/CT5Jof1WUAEuVxN.jpg
6	666051853826850816	https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg
7	666055525042405380	https://pbs.twimg.com/media/CT5N9tpXIAAifs1.jpg
8	666057090499244032	https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg
9	666058600524156928	https://pbs.twimg.com/media/CT5Qw94XAAA_2dP.jpg
10	666063827256086533	https://pbs.twimg.com/media/CT5Vg_wXIAAXfnj.jpg
11	666071193221509120	https://pbs.twimg.com/media/CT5cN_3WEAAl0oZ.jpg
12	666073100786774016	https://pbs.twimg.com/media/CT5d9DZXAAALcwe.jpg
13	666082916733198337	https://pbs.twimg.com/media/CT5m4VGWEAAtKc8.jpg
14	666094000022159362	https://pbs.twimg.com/media/CT5w9gUW4AAsBNN.jpg
15	666099513787052032	https://pbs.twimg.com/media/CT51-JJUEAA6hV8.jpg
16	666102155909144576	https://pbs.twimg.com/media/CT54YGiWUAEZnoK.jpg
17	666104133288665088	https://pbs.twimg.com/media/CT56LSZWoAAlJj2.jpg
18	666268910803644416	https://pbs.twimg.com/media/CT8QCd1WEAADXws.jpg
19	666273097616637952	https://pbs.twimg.com/media/CT8T1mtUwAA3aqm.jpg
20	666287406224695296	https://pbs.twimg.com/media/CT8g3BpUEAAuFjg.jpg
21	666293911632134144	https://pbs.twimg.com/media/CT8mx7KW4AEQu8N.jpg
22	666337882303524864	https://pbs.twimg.com/media/CT9OwFIWEAMuRje.jpg
23	666345417576210432	https://pbs.twimg.com/media/CT9Vn7PWoAA_ZCM.jpg
24	666353288456101888	https://pbs.twimg.com/media/CT9cx0tUEAAhNN_.jpg
25	666362758909284353	https://pbs.twimg.com/media/CT9lXGsUcAAyUFt.jpg
26	666373753744588802	https://pbs.twimg.com/media/CT9vZEYWUAAlZ05.jpg

```
27    666396247373291520    https://pbs.twimg.com/media/CT-D2ZHWIAA3gK1.jpg
28    666407126856765440    https://pbs.twimg.com/media/CT-NvwmW4AAugGZ.jpg
29    666411507551481857    https://pbs.twimg.com/media/CT-RugiWIAELEaq.jpg
...                  ...                                                ...
2045  886366144734445568    https://pbs.twimg.com/media/DE0BTnQUwAApKEH.jpg
2046  886680336477933568    https://pbs.twimg.com/media/DE4fEDzWAAAyHMM.jpg
2047  886736880519319552    https://pbs.twimg.com/media/DE5Se8FXcAAJFx4.jpg
2048  886983233522544640    https://pbs.twimg.com/media/DE8yicJW0AAAvBJ.jpg
2049  887101392804085760    https://pbs.twimg.com/media/DE-eAq6UwAA-jaE.jpg
2050  887343217045368832  https://pbs.twimg.com/ext_tw_video_thumb/88734...
2051  887473957103951883    https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg
2052  887517139158093824  https://pbs.twimg.com/ext_tw_video_thumb/88751...
2053  887705289381826560    https://pbs.twimg.com/media/DFHDQBbXgAEqY7t.jpg
2054  888078434458587136    https://pbs.twimg.com/media/DFMWn56WsAAkA7B.jpg
2055  888202515573088257    https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg
2056  888554962724278272    https://pbs.twimg.com/media/DFTH_0-UQAACu20.jpg
2057  888804989199671297    https://pbs.twimg.com/media/DFWra-3VYAA2piG.jpg
2058  888917238123831296    https://pbs.twimg.com/media/DFYRgsOUQAARGhO.jpg
2059  889278841981685760  https://pbs.twimg.com/ext_tw_video_thumb/88927...
2060  889531135344209921    https://pbs.twimg.com/media/DFg_2PVW0AEHN3p.jpg
2061  889638837579907072    https://pbs.twimg.com/media/DFihzFfXsAYGDPR.jpg
2062  889665388333682689    https://pbs.twimg.com/media/DFi579UWsAAatzw.jpg
2063  889880896479866881    https://pbs.twimg.com/media/DFl99B1WsAITKsg.jpg
2064  890006608113172480    https://pbs.twimg.com/media/DFnwSY4WAAAMliS.jpg
2065  890240255349198849    https://pbs.twimg.com/media/DFrEyVuW0AAO3t9.jpg
2066  890609185150312448    https://pbs.twimg.com/media/DFwUU__XcAEpyXI.jpg
2067  890729181411237888    https://pbs.twimg.com/media/DFyBahAVwAAhUTd.jpg
2068  890971913173991426    https://pbs.twimg.com/media/DF1eOmZXUAALUcq.jpg
2069  891087950875897856    https://pbs.twimg.com/media/DF3HwyEWsAABqE6.jpg
2070  891327558926688256    https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg
2071  891689557279858688    https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg
2072  891815181378084864    https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg
2073  892177421306343426    https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg
2074  892420643555336193    https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg

      img_num                          p1   p1_conf p1_dog  \
0           1       Welsh_springer_spaniel  0.465074   True
1           1                     redbone  0.506826   True
2           1              German_shepherd  0.596461   True
3           1          Rhodesian_ridgeback  0.408143   True
4           1           miniature_pinscher  0.560311   True
5           1         Bernese_mountain_dog  0.651137   True
6           1                   box_turtle  0.933012  False
7           1                         chow  0.692517   True
8           1                shopping_cart  0.962465  False
9           1             miniature_poodle  0.201493   True
10          1             golden_retriever  0.775930   True
11          1                Gordon_setter  0.503672   True
```

12	1	Walker_hound	0.260857	True
13	1	pug	0.489814	True
14	1	bloodhound	0.195217	True
15	1	Lhasa	0.582330	True
16	1	English_setter	0.298617	True
17	1	hen	0.965932	False
18	1	desktop_computer	0.086502	False
19	1	Italian_greyhound	0.176053	True
20	1	Maltese_dog	0.857531	True
21	1	three-toed_sloth	0.914671	False
22	1	ox	0.416669	False
23	1	golden_retriever	0.858744	True
24	1	malamute	0.336874	True
25	1	guinea_pig	0.996496	False
26	1	soft-coated_wheaten_terrier	0.326467	True
27	1	Chihuahua	0.978108	True
28	1	black-and-tan_coonhound	0.529139	True
29	1	coho	0.404640	False
...	...	...	...	...
2045	1	French_bulldog	0.999201	True
2046	1	convertible	0.738995	False
2047	1	kuvasz	0.309706	True
2048	2	Chihuahua	0.793469	True
2049	1	Samoyed	0.733942	True
2050	1	Mexican_hairless	0.330741	True
2051	2	Pembroke	0.809197	True
2052	1	limousine	0.130432	False
2053	1	basset	0.821664	True
2054	1	French_bulldog	0.995026	True
2055	2	Pembroke	0.809197	True
2056	3	Siberian_husky	0.700377	True
2057	1	golden_retriever	0.469760	True
2058	1	golden_retriever	0.714719	True
2059	1	whippet	0.626152	True
2060	1	golden_retriever	0.953442	True
2061	1	French_bulldog	0.991650	True
2062	1	Pembroke	0.966327	True
2063	1	French_bulldog	0.377417	True
2064	1	Samoyed	0.957979	True
2065	1	Pembroke	0.511319	True
2066	1	Irish_terrier	0.487574	True
2067	2	Pomeranian	0.566142	True
2068	1	Appenzeller	0.341703	True
2069	1	Chesapeake_Bay_retriever	0.425595	True
2070	2	basset	0.555712	True
2071	1	paper_towel	0.170278	False
2072	1	Chihuahua	0.716012	True
2073	1	Chihuahua	0.323581	True

```
2074        1                      orange  0.097049  False

                          p2   p2_conf p2_dog                          p3  \
0                     collie  0.156665   True            Shetland_sheepdog
1           miniature_pinscher  0.074192   True          Rhodesian_ridgeback
2                    malinois  0.138584   True                   bloodhound
3                     redbone  0.360687   True           miniature_pinscher
4                  Rottweiler  0.243682   True                     Doberman
5             English_springer  0.263788   True   Greater_Swiss_Mountain_dog
6                  mud_turtle  0.045885  False                      terrapin
7              Tibetan_mastiff  0.058279   True                     fur_coat
8              shopping_basket  0.014594  False             golden_retriever
9                     komondor  0.192305   True  soft-coated_wheaten_terrier
10             Tibetan_mastiff  0.093718   True           Labrador_retriever
11           Yorkshire_terrier  0.174201   True                     Pekinese
12            English_foxhound  0.175382   True                 Ibizan_hound
13                bull_mastiff  0.404722   True               French_bulldog
14             German_shepherd  0.078260   True                     malinois
15                    Shih-Tzu  0.166192   True               Dandie_Dinmont
16                Newfoundland  0.149842   True                       borzoi
17                        cock  0.033919  False                    partridge
18                        desk  0.085547  False                     bookcase
19                toy_terrier  0.111884   True                      basenji
20                toy_poodle  0.063064   True             miniature_poodle
21                      otter  0.015250  False               great_grey_owl
22                Newfoundland  0.278407   True                  groenendael
23    Chesapeake_Bay_retriever  0.054787   True           Labrador_retriever
24              Siberian_husky  0.147655   True                   Eskimo_dog
25                        skunk  0.002402  False                      hamster
26                Afghan_hound  0.259551   True                      briard
27                toy_terrier  0.009397   True                     papillon
28                 bloodhound  0.244220   True         flat-coated_retriever
29                 barracouta  0.271485  False                          gar
...                       ...       ...    ...                          ...
2045                Chihuahua  0.000361   True                  Boston_bull
2046                sports_car  0.139952  False                    car_wheel
2047           Great_Pyrenees  0.186136   True                Dandie_Dinmont
2048              toy_terrier  0.143528   True                    can_opener
2049                Eskimo_dog  0.035029   True    Staffordshire_bullterrier
2050                  sea_lion  0.275645  False                   Weimaraner
2051       Rhodesian_ridgeback  0.054950   True                        beagle
2052                 tow_truck  0.029175  False                shopping_cart
2053                   redbone  0.087582   True                   Weimaraner
2054                       pug  0.000932   True                 bull_mastiff
2055       Rhodesian_ridgeback  0.054950   True                        beagle
2056                Eskimo_dog  0.166511   True                     malamute
2057        Labrador_retriever  0.184172   True                English_setter
2058           Tibetan_mastiff  0.120184   True           Labrador_retriever
```

2059	borzoi	0.194742	True	Saluki
2060	Labrador_retriever	0.013834	True	redbone
2061	boxer	0.002129	True	Staffordshire_bullterrier
2062	Cardigan	0.027356	True	basenji
2063	Labrador_retriever	0.151317	True	muzzle
2064	Pomeranian	0.013884	True	chow
2065	Cardigan	0.451038	True	Chihuahua
2066	Irish_setter	0.193054	True	Chesapeake_Bay_retriever
2067	Eskimo_dog	0.178406	True	Pembroke
2068	Border_collie	0.199287	True	ice_lolly
2069	Irish_terrier	0.116317	True	Indian_elephant
2070	English_springer	0.225770	True	German_short-haired_pointer
2071	Labrador_retriever	0.168086	True	spatula
2072	malamute	0.078253	True	kelpie
2073	Pekinese	0.090647	True	papillon
2074	bagel	0.085851	False	banana

	p3_conf	p3_dog
0	0.061428	True
1	0.072010	True
2	0.116197	True
3	0.222752	True
4	0.154629	True
5	0.016199	True
6	0.017885	False
7	0.054449	False
8	0.007959	True
9	0.082086	True
10	0.072427	True
11	0.109454	True
12	0.097471	True
13	0.048960	True
14	0.075628	True
15	0.089688	True
16	0.133649	True
17	0.000052	False
18	0.079480	False
19	0.111152	True
20	0.025581	True
21	0.013207	False
22	0.102643	True
23	0.014241	True
24	0.093412	True
25	0.000461	False
26	0.206803	True
27	0.004577	True
28	0.173810	True
29	0.189945	False

```
...         ...    ...
2045  0.000076   True
2046  0.044173  False
2047  0.086346   True
2048  0.032253  False
2049  0.029705   True
2050  0.134203   True
2051  0.038915   True
2052  0.026321  False
2053  0.026236   True
2054  0.000903   True
2055  0.038915   True
2056  0.111411   True
2057  0.073482   True
2058  0.105506   True
2059  0.027351   True
2060  0.007958   True
2061  0.001498   True
2062  0.004633   True
2063  0.082981  False
2064  0.008167   True
2065  0.029248   True
2066  0.118184   True
2067  0.076507   True
2068  0.193548  False
2069  0.076902  False
2070  0.175219   True
2071  0.040836  False
2072  0.031379   True
2073  0.068957   True
2074  0.076110  False

[2075 rows x 12 columns]
```

```
In [57]: images.info()
         images['jpg_url'].value_counts()
         images[images['jpg_url'] == 'https://pbs.twimg.com/media/CZhn-QAWwAASQan.jpg']
```

```
<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
```

Out[57]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
800	691416866452082688	https://pbs.twimg.com/media/CZhn-QAWwAASQan.jpg	1	Lakeland_terrier	0.530104	True	Irish_terrier	0.197314	True	Airedale	0.082515	True
1624	803692223237865472	https://pbs.twimg.com/media/CZhn-QAWwAASQan.jpg	1	Lakeland_terrier	0.530104	True	Irish_terrier	0.197314	True	Airedale	0.082515	True

In [65]:
```
json_tweets
json_tweets.info()
# json_tweets['tweet_id'].value_counts() count tweet_ids
# json_tweets['user_followers'].value_counts() check if querying the use_followers had
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 6 columns):
tweet_id           2356 non-null int64
favorites          2356 non-null int64
retweets           2356 non-null int64
user_followers     2356 non-null int64
user_favourites    2356 non-null int64
date_time          2356 non-null object
dtypes: int64(5), object(1)
memory usage: 110.5+ KB
```

Out[65]:

4484699	55
4484638	50
4484676	50
4484509	48
4484670	46
4484741	45
4484688	44
4484752	42
4484449	42
4484456	41

```
4484685    40
4484693    38
4484497    37
4484690    37
4484770    36
4484642    36
4484530    32
4484513    30
4484773    28
4484491    28
4484680    27
4484504    26
4484492    26
4484687    25
4484691    24
4484754    23
4484698    23
4484751    22
4484535    20
4484760    20
           ..
4484602     2
4484761     2
4484505     2
4484607     2
4484782     2
4484665     2
4484458     2
4484697     2
4484455     1
4484758     1
4484598     1
4484597     1
4484668     1
4484532     1
4484516     1
4484612     1
4484494     1
4484634     1
4484759     1
4484735     1
4484704     1
4484605     1
4484730     1
4484649     1
4484647     1
4484635     1
4484625     1
```

```
4484619     1
4484611     1
4484450     1
Name: user_followers, dtype: int64
```

Some description of the variables in the dataset, you can find more in this link - **tweet_id**: The integer representation of the unique identifier for this Tweet. - **in_reply_to_status_id**: If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's ID - **in_reply_to_user_id_str**: If the represented Tweet is a reply, this field will contain the string representation of the original Tweet's author ID. This will not necessarily always be the user directly mentioned in the Tweet - **retweet_count**: Number of times this Tweet has been retweeted - **favorite_count**: Indicates approximately how many times this Tweet has been liked by Twitter users

1.4.1 Quality

Completeness, Validity, Accuracy, Consistency => a.k.a content issues

`archive` **dataset** - `in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id` should be `intergs` instead of `float` - `retweeted_status_timestamp, timestamp` should be `datetime` instead of `object` (string) - The numerator and denominator columns have invalid values - In several columns null objects are non-null (None to NaN) - `Name` column have invalid names i.e 'None', 'a', 'an' - We only want original ratings (no retweets) that have images - We may want to change this columns type (`in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id` and `tweet_id`) to `string` because We don't want any operations on them

`images` **dataset** - Missing values from images dataset (2075 rows instead of 2356) - Some tweet_ids have the same jpg_url - Some tweets are have 2 different tweet_id one redirect to the other

`json_tweeets` **dataset** - This tweet_id (666020888022790149) duplicated 8 times

1.4.2 Tidiness

Untidy data => a.k.a structural issues - No need to all the informations in `images dataset`, (tweet_id and jpg_url what matters) - Various stages of dogs in columns instead of rows `archives dataset` - We may want to add a gender column from the text columns in `archives dataset` - All tables should be part of one dataset

NB : We could add a column called `jpg_url_api` contain the query of the api `media_url_https` it will have the same result as the images dataset

1.5 Clean

Cleaning our data is the third step in data wrangling. It is where we will fix the quality and tidiness issues that we identified in the assess step.

```
In [585]: # Since we want to create one high quality and tidy master pandas DataFrame
          # we will start by merging our dataframe in one
          # the we save the result in file as backup
          df_master = pd.merge(archive, images, how = 'left', on = ['tweet_id'] )
          df_master = pd.merge(df_master, json_tweets, how = 'left', on = ['tweet_id'])
```

```
df_master.to_csv('df_master.csv', encoding = 'utf-8')
df_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2363 entries, 0 to 2362
Data columns (total 33 columns):
tweet_id                      2363 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2363 non-null object
source                        2363 non-null object
text                          2363 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                 2304 non-null object
rating_numerator              2363 non-null int64
rating_denominator            2363 non-null int64
name                          2363 non-null object
doggo                         2363 non-null object
floofer                       2363 non-null object
pupper                        2363 non-null object
puppo                         2363 non-null object
jpg_url                       2082 non-null object
img_num                       2082 non-null float64
p1                            2082 non-null object
p1_conf                       2082 non-null float64
p1_dog                        2082 non-null object
p2                            2082 non-null object
p2_conf                       2082 non-null float64
p2_dog                        2082 non-null object
p3                            2082 non-null object
p3_conf                       2082 non-null float64
p3_dog                        2082 non-null object
favorites                     2356 non-null float64
retweets                      2356 non-null float64
user_followers                2356 non-null float64
user_favourites               2356 non-null float64
date_time                     2356 non-null object
dtypes: float64(12), int64(3), object(18)
memory usage: 627.7+ KB
```

We notice 7 additional rows, it maybe because of the sencond query for those tweet_id that tweeepy except throw. We will investigate on this more while cleaning our data.

1.5.1 Quick Clean to rows and columns that we will not need

In [586]:
```
# Delete the retweets
df_master = df_master[pd.isnull(df_master.retweeted_status_id)]
# Delete duplicated tweet_id
df_master = df_master.drop_duplicates()
# Delete tweets with no pictures
df_master = df_master.dropna(subset = ['jpg_url'])

# test
len(df_master)
```

Out[586]:
```
1994
```

In [587]:
```
# Delete columns related to retweet we don't need anymore
df_master = df_master.drop('retweeted_status_id', 1)
df_master = df_master.drop('retweeted_status_user_id', 1)
df_master = df_master.drop('retweeted_status_timestamp', 1)

# Delete column date_time we imported from the API, it has the same values as timestan
df_master = df_master.drop('date_time', 1)

# test
list(df_master)
```

Out[587]:
```
['tweet_id',
 'in_reply_to_status_id',
 'in_reply_to_user_id',
 'timestamp',
 'source',
 'text',
 'expanded_urls',
 'rating_numerator',
 'rating_denominator',
 'name',
 'doggo',
 'floofer',
 'pupper',
 'puppo',
 'jpg_url',
 'img_num',
 'p1',
 'p1_conf',
 'p1_dog',
 'p2',
 'p2_conf',
 'p2_dog',
 'p3',
 'p3_conf',
```

```
            'p3_dog',
            'favorites',
            'retweets',
            'user_followers',
            'user_favourites']
```

1.5.2 Melt the 'doggo', 'floofer', 'pupper' and 'puppo' columns into one column 'dog_stage'

```
In [588]: # Check the values in those columns by excuting those columns
          print(df_master.doggo.value_counts())
          print(df_master.floofer.value_counts())
          print(df_master.pupper.value_counts())
          print(df_master.puppo.value_counts())
```

```
None     1920
doggo      74
Name: doggo, dtype: int64
None       1986
floofer       8
Name: floofer, dtype: int64
None      1782
pupper     212
Name: pupper, dtype: int64
None     1971
puppo      23
Name: puppo, dtype: int64
```

```
In [589]: # Select the columns to melt and to remain
          columns_to_melt = ['doggo', 'floofer', 'pupper', 'puppo']
          columns_to_stay = [x for x in df_master.columns.tolist() if x not in columns_to_melt]

          # Mlet the the columns into values
          df_master = pd.melt(df_master, id_vars = columns_to_stay, value_vars = columns_to_melt
                                      var_name = 'stages', value_name = 'dog_stage')

          # Delete column 'stages'
          df_master = df_master.drop('stages', 1)

          # Filter for unique values then remove duplicate values based on 'dog_stage' values

          # This part for test *
          print(df_master.dog_stage.value_counts())

          df_master = df_master.sort_values('dog_stage').drop_duplicates('tweet_id', keep = 'las

          # This part for test
          print(df_master.dog_stage.value_counts())
          print(len(df_master))
```

```
None       7659
pupper      212
doggo        74
puppo        23
floofer       8
Name: dog_stage, dtype: int64
None       1688
pupper      212
doggo        63
puppo        23
floofer       8
Name: dog_stage, dtype: int64
1994
```

1.5.3 Get rid of image prediction columns

```
In [233]: # We will store the fisrt true algorithm with it's level of confidence
          prediction_algorithm = []
          confidence_level = []

          # Get_prediction_confidence function:
          # search the first true algorithm and append it to a list with it's level of confidenc
          # if flase prediction_algorthm will have a value of NaN
          def get_prediction_confidence(dataframe):
              if dataframe['p1_dog'] == True:
                  prediction_algorithm.append(dataframe['p1'])
                  confidence_level.append(dataframe['p1_conf'])
              elif dataframe['p2_dog'] == True:
                  prediction_algorithm.append(dataframe['p2'])
                  confidence_level.append(dataframe['p2_conf'])
              elif dataframe['p3_dog'] == True:
                  prediction_algorithm.append(dataframe['p3'])
                  confidence_level.append(dataframe['p3_conf'])
              else:
                  prediction_algorithm.append('NaN')
                  confidence_level.append(0)

          df_master.apply(get_prediction_confidence, axis=1)
          df_master['prediction_algorithm'] = prediction_algorithm
          df_master['confidence_level'] = confidence_level

          # Test
          list(df_master)

Out[233]: ['tweet_id',
           'in_reply_to_status_id',
           'in_reply_to_user_id',
```

```
 'timestamp',
 'source',
 'text',
 'expanded_urls',
 'rating_numerator',
 'rating_denominator',
 'name',
 'jpg_url',
 'img_num',
 'p1',
 'p1_conf',
 'p1_dog',
 'p2',
 'p2_conf',
 'p2_dog',
 'p3',
 'p3_conf',
 'p3_dog',
 'favorites',
 'retweets',
 'user_followers',
 'user_favourites',
 'dog_stage',
 'prediction_algorithm',
 'confidence_level']
```

In [244]:
```
# Delete the columns of image prediction information
df_master = df_master.drop(['img_num', 'p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2

# Test
list(df_master)
```

Out[244]:
```
['tweet_id',
 'in_reply_to_status_id',
 'in_reply_to_user_id',
 'timestamp',
 'source',
 'text',
 'expanded_urls',
 'rating_numerator',
 'rating_denominator',
 'name',
 'jpg_url',
 'favorites',
 'retweets',
 'user_followers',
 'user_favourites',
 'dog_stage',
```

```
          'prediction_algorithm',
          'confidence_level']
```

1.5.4 Check for the duplicated values and delete useless informations

```
In [247]: # Print the count of the unique elements in all columns
          df_master.apply(lambda x: len(x.unique()))
```

```
Out[247]: tweet_id                 1994
          in_reply_to_status_id      23
          in_reply_to_user_id         2
          timestamp                1994
          source                      3
          text                     1994
          expanded_urls            1994
          rating_numerator           34
          rating_denominator         15
          name                      936
          jpg_url                  1994
          favorites                1874
          retweets                 1624
          user_followers            201
          user_favourites             2
          dog_stage                   5
          prediction_algorithm      114
          confidence_level         1684
          dtype: int64
```

```
In [257]: # let's concentrate on low values.. let's dig more
          df_master.info()
          df_master['in_reply_to_user_id'].value_counts()
          df_master['source'].value_counts()
          df_master['user_favourites'].value_counts()
```

The result of the above lines says: - One value in **in_reply_to_user_id** so we will delete the columns of reply all of them replying to one user id '4196983835' @dog_rates - **source** has 3 types, we will clean that column and made them clear - **user_favourites** has 2 values and they are close, (that information meant to calculate the % of the like by favorites of the user; to see if the number of the folowers or the favorites of the page in general affect the number of the favorites or likes a tweet will get. Since this dataset gathered by querying the API in one date this information has no meaning. we will delete this column

```
In [259]: # drop the following columns 'in_reply_to_status_id', 'in_reply_to_user_id', 'user_fav
          df_master = df_master.drop(['in_reply_to_status_id', 'in_reply_to_user_id', 'user_favo
```

```
In [268]: # Clean the content of source column
          df_master['source'] = df_master['source'].apply(lambda x: re.findall(r'>(.*)<', x)[0])

          # Test
          df_master
```

Out[268]:

	tweet_id	timestamp	source
1918	667405339315146752	2015-11-19 18:13:27 +0000	Twitter for iPhone
1917	667435689202614272	2015-11-19 20:14:03 +0000	Twitter for iPhone
1916	667437278097252352	2015-11-19 20:20:22 +0000	Twitter for iPhone
1915	667443425659232256	2015-11-19 20:44:47 +0000	Twitter for iPhone
1914	667453023279554560	2015-11-19 21:22:56 +0000	Twitter Web Client
1913	667455448082227200	2015-11-19 21:32:34 +0000	Twitter Web Client
1912	667470559035432960	2015-11-19 22:32:36 +0000	Twitter Web Client
1911	667491009379606528	2015-11-19 23:53:52 +0000	Twitter Web Client
1910	667495797102141441	2015-11-20 00:12:54 +0000	Twitter Web Client
1909	667502640335572993	2015-11-20 00:40:05 +0000	Twitter Web Client
1908	667509364010450944	2015-11-20 01:06:48 +0000	Twitter Web Client
1907	667517642048163840	2015-11-20 01:39:42 +0000	Twitter Web Client
1906	667524857454854144	2015-11-20 02:08:22 +0000	Twitter Web Client
1905	667530908589760512	2015-11-20 02:32:25 +0000	Twitter Web Client
1904	667534815156183040	2015-11-20 02:47:56 +0000	Twitter Web Client
1903	667538891197542400	2015-11-20 03:04:08 +0000	Twitter Web Client
1902	667544320556335104	2015-11-20 03:25:43 +0000	Twitter Web Client
1901	667546741521195010	2015-11-20 03:35:20 +0000	Twitter Web Client
1900	667549055577362432	2015-11-20 03:44:31 +0000	Twitter Web Client
1899	667724302356258817	2015-11-20 15:20:54 +0000	Twitter Web Client
1898	667728196545200128	2015-11-20 15:36:22 +0000	Twitter Web Client
1897	667766675769573376	2015-11-20 18:09:16 +0000	Twitter Web Client
1896	667773195014021121	2015-11-20 18:35:10 +0000	Twitter Web Client
1895	667782464991965184	2015-11-20 19:12:01 +0000	Twitter for iPhone
1894	667793409583771648	2015-11-20 19:55:30 +0000	Twitter for iPhone
1893	667801013445750784	2015-11-20 20:25:43 +0000	Twitter for iPhone
1892	667806454573760512	2015-11-20 20:47:20 +0000	Twitter for iPhone
1919	667393430834667520	2015-11-19 17:26:08 +0000	Twitter for iPhone
1891	667832474953625600	2015-11-20 22:30:44 +0000	Twitter for iPhone
1920	667369227918143488	2015-11-19 15:49:57 +0000	Twitter for iPhone
...	...	...	...
5200	690015576308211712	2016-01-21 03:38:27 +0000	Twitter for iPhone
5513	675113801096802304	2015-12-11 00:44:07 +0000	Twitter for iPhone
5204	689905486972461056	2016-01-20 20:21:00 +0000	Twitter for iPhone
4064	874296783580663808	2017-06-12 16:06:11 +0000	Twitter for iPhone
5543	674447403907457024	2015-12-09 04:36:06 +0000	Twitter for iPhone
5057	702598099714314240	2016-02-24 20:56:55 +0000	Twitter for iPhone
5209	689623661272240129	2016-01-20 01:41:08 +0000	Twitter for iPhone
6281	825026590719483904	2017-01-27 17:04:02 +0000	Twitter for iPhone
5996	889531135344209921	2017-07-24 17:02:04 +0000	Twitter for iPhone
6680	752519690950500352	2016-07-11 15:07:30 +0000	Twitter for iPhone
6130	855851453814013952	2017-04-22 18:31:02 +0000	Twitter for iPhone
6291	822872901745569793	2017-01-21 18:26:02 +0000	Twitter for iPhone
6691	751132876104687617	2016-07-07 19:16:47 +0000	Twitter for iPhone
6448	793195938047070209	2016-10-31 21:00:23 +0000	Twitter for iPhone
6060	874012996292530176	2017-06-11 21:18:31 +0000	Twitter for iPhone
5994	889665388333682689	2017-07-25 01:55:32 +0000	Twitter for iPhone

```
6116  859607811541651456  2017-05-03 03:17:27 +0000  Twitter for iPhone
6481  787717603741622272  2016-10-16 18:11:26 +0000  Twitter for iPhone
6308  819952236453363712  2017-01-13 17:00:21 +0000  Twitter for iPhone
6795  738537504001953792  2016-06-03 01:07:16 +0000  Twitter for iPhone
6753  744995568523612160  2016-06-20 20:49:19 +0000  Twitter for iPhone
6661  756275833623502848  2016-07-21 23:53:04 +0000  Twitter for iPhone
6516  780931614150983680  2016-09-28 00:46:20 +0000  Twitter for iPhone
6402  802239329049477120  2016-11-25 19:55:35 +0000  Twitter for iPhone
6394  803773340896923648  2016-11-30 01:31:12 +0000  Twitter for iPhone
6463  790946055508652032  2016-10-25 16:00:09 +0000  Twitter for iPhone
6087  867421006826221569  2017-05-24 16:44:18 +0000  Twitter for iPhone
6044  878776093423087618  2017-06-25 00:45:22 +0000  Twitter for iPhone
6279  825535076884762624  2017-01-29 02:44:34 +0000  Twitter for iPhone
6764  743253157753532416  2016-06-16 01:25:36 +0000  Twitter for iPhone

                                                    text  \
1918  This is Biden. Biden just tripped... 7/10 http...
1917            Ermergerd 12/10 https://t.co/PQni2sjPsm
1916  Never seen this breed before. Very pointy pup...
1915  Exotic dog here. Long neck. Weird paws. Obsess...
1914  Meet Cupcake. I would do unspeakable things fo...
1913  This is Reese and Twips. Reese protects Twips...
1912  This is a northern Wahoo named Kohl. He runs t...
1911  Two dogs in this one. Both are rare Jujitsu Py...
1910  This is Philippe from Soviet Russia. Commandin...
1909  Say hello to Hall and Oates. Oates is winking ...
1908  This a Norwegian Pewterschmidt named Tickles. ...
1907  This is Dook &amp;amp; Milo. Dook is struggling to...
1906  Another topnotch dog. His name is Big Jumpy Ra...
1905  Meet Naphaniel. He doesn't necessarily enjoy h...
1904  This is Frank (pronounced "Fronq"). Too many b...
1903  This is a southwest Coriander named Klint. Hat...
1902  This is Kial. Kial is either wearing a cape, w...
1901  Here is George. George took a selfie of his ne...
1900  Never seen dog like this. Breathes heavy. Tilt...
1899  What a dog to start the day with. Very calm. L...
1898  Meet Olive. He comes to spot by tree to remini...
1897  This is Calvin. He is a Luxembourgian Mayo. Ha...
1896  This is a rare Hungarian Pinot named Jessiga. ...
1895  Super rare dog. Endangered (?). Thinks it's fu...
1894  Dogs only please. Small cows and other non can...
1893               OMIGOD 12/10 https://t.co/SVMF4Frf1w
1892  This is Filup. He is overcome with joy after f...
1919  This is Fwed. He is a Canadian Asian Taylormad...
1891  THE EYES 12/10\n\nI'm sorry. These are suppose...
1920  Here we have a neat pup. Very white. Cool shad...
...                                                 ...
5200  This pupper can only sleep on shoes. It's a cr...
```

```
5513  Meet Zuzu. He just graduated college. Astute p...
5204  Say hello to Gizmo. He's quite the pupper. Con...
4064  This is Jed. He may be the fanciest pupper in ...
5543  This pupper just wants a belly rub. This puppe...
5057  This is Sansa. She's gotten too big for her ch...
5209  This is Lucy. She's terrified of the stuffed b...
6281  Say hello to Pablo. He's one gorgeous puppo. A...
5996  This is Stuart. He's sporting his favorite fan...
6680  Hopefully this puppo on a swing will help get ...
6130  Here's a puppo participating in the #ScienceMa...
6291  Here's a super supportive puppo participating ...
6691  This is Cooper. He's just so damn happy. 10/10...
6448  Say hello to Lily. She's pupset that her costu...
6060  This is Sebastian. He can't see all the colors...
5994  Here's a puppo that seems to be on the fence a...
6116  Sorry for the lack of posts today. I came home...
6481  This is Tonks. She is a service puppo. Can hea...
6308  This is Oliver. He has dreams of being a servi...
6795  This is Bayley. She fell asleep trying to esca...
6753  This is Abby. She got her face stuck in a glas...
6661  When ur older siblings get to play in the deep...
6516  I want to finally rate this iconic puppo who t...
6402  This is Loki. He'll do your taxes for you. Can...
6394  This is Diogi. He fell in the pool as soon as ...
6463  This is Betty. She's assisting with the dishes...
6087  This is Shikha. She just watched you drop a sk...
6044  This is Snoopy. He's a proud #PrideMonthPuppo...
6279  Here's a very loving and accepting puppo. Appe...
6764  This is Kilo. He cannot reach the snackum. Nif...

                                          expanded_urls  rating_numerator  \
1918  https://twitter.com/dog_rates/status/667405339...                 7
1917  https://twitter.com/dog_rates/status/667435689...                12
1916  https://twitter.com/dog_rates/status/667437278...                10
1915  https://twitter.com/dog_rates/status/667443425...                 6
1914  https://twitter.com/dog_rates/status/667453023...                11
1913  https://twitter.com/dog_rates/status/667455448...                 7
1912  https://twitter.com/dog_rates/status/667470559...                11
1911  https://twitter.com/dog_rates/status/667491009...                 7
1910  https://twitter.com/dog_rates/status/667495797...                 9
1909  https://twitter.com/dog_rates/status/667502640...                11
1908  https://twitter.com/dog_rates/status/667509364...                12
1907  https://twitter.com/dog_rates/status/667517642...                 8
1906  https://twitter.com/dog_rates/status/667524857...                12
1905  https://twitter.com/dog_rates/status/667530908...                10
1904  https://twitter.com/dog_rates/status/667534815...                 8
1903  https://twitter.com/dog_rates/status/667538891...                 9
1902  https://twitter.com/dog_rates/status/667544320...                10
```

```
1901  https://twitter.com/dog_rates/status/667546741...                 9
1900  https://twitter.com/dog_rates/status/667549055...                 1
1899  https://twitter.com/dog_rates/status/667724302...                 7
1898  https://twitter.com/dog_rates/status/667728196...                11
1897  https://twitter.com/dog_rates/status/667766675...                 9
1896  https://twitter.com/dog_rates/status/667773195...                 8
1895  https://twitter.com/dog_rates/status/667782464...                 9
1894  https://twitter.com/dog_rates/status/667793409...                 8
1893  https://twitter.com/dog_rates/status/667801013...                12
1892  https://twitter.com/dog_rates/status/667806454...                10
1919  https://twitter.com/dog_rates/status/667393430...                 8
1891  https://twitter.com/dog_rates/status/667832474...                12
1920  https://twitter.com/dog_rates/status/667369227...                10
...                                                  ...               ...
5200  https://twitter.com/dog_rates/status/690015576...                12
5513  https://twitter.com/dog_rates/status/675113801...                10
5204  https://twitter.com/dog_rates/status/689905486...                11
4064  https://twitter.com/dog_rates/status/874296783...                13
5543  https://twitter.com/dog_rates/status/674447403...                10
5057  https://twitter.com/dog_rates/status/702598099...                11
5209  https://twitter.com/dog_rates/status/689623661...                10
6281  https://www.gofundme.com/my-puppys-double-cata...                12
5996  https://twitter.com/dog_rates/status/889531135...                13
6680  https://twitter.com/dog_rates/status/752519690...                11
6130  https://twitter.com/dog_rates/status/855851453...                13
6291  https://twitter.com/dog_rates/status/822872901...                13
6691  https://twitter.com/dog_rates/status/751132876...                10
6448  https://twitter.com/dog_rates/status/793195938...                12
6060  https://twitter.com/dog_rates/status/874012996...                13
5994  https://twitter.com/dog_rates/status/889665388...                13
6116  https://twitter.com/dog_rates/status/859607811...                13
6481  https://twitter.com/dog_rates/status/787717603...                13
6308  https://www.gofundme.com/servicedogoliver,http...                13
6795  https://twitter.com/dog_rates/status/738537504...                11
6753  https://twitter.com/dog_rates/status/744995568...                 9
6661  https://twitter.com/dog_rates/status/756275833...                10
6516  https://twitter.com/dog_rates/status/780931614...                13
6402  https://twitter.com/dog_rates/status/802239329...                12
6394  https://twitter.com/dog_rates/status/803773340...                12
6463  https://twitter.com/dog_rates/status/790946055...                12
6087  https://twitter.com/dog_rates/status/867421006...                12
6044  https://twitter.com/dog_rates/status/878776093...                13
6279  https://twitter.com/dog_rates/status/825535076...                14
6764  https://twitter.com/dog_rates/status/743253157...                10

      rating_denominator       name  \
1918                  10      Biden
1917                  10       None
```

1916	10	None
1915	10	None
1914	10	Cupcake
1913	10	Reese
1912	10	a
1911	10	None
1910	10	Philippe
1909	10	Hall
1908	10	None
1907	10	Dook
1906	10	None
1905	10	Naphaniel
1904	10	Frank
1903	10	a
1902	10	Kial
1901	10	George
1900	10	None
1899	10	None
1898	10	Olive
1897	10	Calvin
1896	10	a
1895	10	None
1894	10	None
1893	10	None
1892	10	Filup
1919	10	Fwed
1891	10	None
1920	10	None
...	...	...
5200	10	None
5513	10	Zuzu
5204	10	Gizmo
4064	10	Jed
5543	10	None
5057	10	Sansa
5209	10	Lucy
6281	10	Pablo
5996	10	Stuart
6680	10	None
6130	10	None
6291	10	None
6691	10	Cooper
6448	10	Lily
6060	10	Sebastian
5994	10	None
6116	10	None
6481	10	Tonks
6308	10	Oliver

```
6795                  10     Bayley
6753                  10       Abby
6661                  10       None
6516                  10       None
6402                  10       Loki
6394                  10      Diogi
6463                  10      Betty
6087                  10     Shikha
6044                  10     Snoopy
6279                  10       None
6764                  10       Kilo

                                                  jpg_url  favorites  retweets  \
1918  https://pbs.twimg.com/media/CUMZnmhUEAEbtis.jpg      488.0     232.0
1917  https://pbs.twimg.com/media/CUM1OHCW4AEgGSi.jpg      324.0      88.0
1916  https://pbs.twimg.com/media/CUM2qWaWoAUZO6L.jpg      478.0     255.0
1915  https://pbs.twimg.com/media/CUM8QZwW4AAVsBl.jpg      823.0     616.0
1914  https://pbs.twimg.com/media/CUNE_OSUwAAdHhX.jpg      327.0      95.0
1913  https://pbs.twimg.com/media/CUNHMXTU8AAS3HH.jpg      202.0      66.0
1912  https://pbs.twimg.com/media/CUNU78YWEAECmpB.jpg      273.0     101.0
1911  https://pbs.twimg.com/media/CUNniSlUYAEj1Jl.jpg      556.0     237.0
1910  https://pbs.twimg.com/media/CUNr4-7UwAAg2lq.jpg      554.0     293.0
1909  https://pbs.twimg.com/media/CUNyHTMUYAAQVch.jpg      562.0     230.0
1908  https://pbs.twimg.com/media/CUN4Or5UAAAa5K4.jpg     7103.0    2254.0
1907  https://pbs.twimg.com/media/CUN_wiBUkAAakT0.jpg      388.0     203.0
1906  https://pbs.twimg.com/media/CUOGUfJW4AA_eni.jpg     1786.0    1188.0
1905  https://pbs.twimg.com/media/CUOLOuGUkAAx7yh.jpg      496.0     260.0
1904  https://pbs.twimg.com/media/CUOPYI5UcAAj_nO.jpg      859.0     571.0
1903  https://pbs.twimg.com/media/CUOTFZOW4AABsfW.jpg      216.0      72.0
1902  https://pbs.twimg.com/media/CUOYBbbWIAAXQGU.jpg      912.0     560.0
1901  https://pbs.twimg.com/media/CUOaOWXWcAAO_Jy.jpg      350.0     135.0
1900  https://pbs.twimg.com/media/CUOcVCwWsAERUKY.jpg     6079.0    2438.0
1899  https://pbs.twimg.com/media/CUQ7tv3W4AA3KlI.jpg      515.0     340.0
1898  https://pbs.twimg.com/media/CUQ_QahUAAAVQjn.jpg      396.0     160.0
1897  https://pbs.twimg.com/media/CURiQMnUAAAPT2M.jpg      473.0     237.0
1896  https://pbs.twimg.com/media/CURoLrOVEAAaWdR.jpg      243.0      61.0
1895  https://pbs.twimg.com/media/CURwm3cUkAARcO6.jpg      433.0     257.0
1894  https://pbs.twimg.com/media/CUR6jqVWsAEgGot.jpg      731.0     356.0
1893  https://pbs.twimg.com/media/CUSBemVUEAAn-6V.jpg      346.0     101.0
1892  https://pbs.twimg.com/media/CUSGbXeVAAAgztZ.jpg     1102.0     528.0
1919  https://pbs.twimg.com/media/CUMOyd3XIAAl13H.jpg      208.0      60.0
1891  https://pbs.twimg.com/media/CUSeGFNW4AAyyHC.jpg      301.0      68.0
1920  https://pbs.twimg.com/media/CUL4xR9UkAEdlJ6.jpg      385.0     171.0
...                                                 ...        ...       ...
5200  https://pbs.twimg.com/media/CZNtgWhWkAAbq3W.jpg     2725.0     825.0
5513  https://pbs.twimg.com/media/CV58a4nXAAApywo.jpg     2100.0     869.0
5204  https://pbs.twimg.com/media/CZMJYCRVAAE35Wk.jpg     2638.0     782.0
4064  https://pbs.twimg.com/media/DCIgSROXgAANEOY.jpg    26485.0    4264.0
```

5543	https://pbs.twimg.com/media/CVweVUfW4AACPwI.jpg	1128.0	391.0
5057	https://pbs.twimg.com/media/CcAhPevW8AAoknv.jpg	11251.0	3676.0
5209	https://pbs.twimg.com/media/CZIJD2SWIAMJgNI.jpg	2438.0	744.0
6281	https://pbs.twimg.com/media/C3MVTeHWcAAGNfx.jpg	6961.0	1468.0
5996	https://pbs.twimg.com/media/DFg_2PVW0AEHN3p.jpg	15302.0	2295.0
6680	https://pbs.twimg.com/media/CnF8qVDWYAAh0g1.jpg	8097.0	3881.0
6130	https://pbs.twimg.com/media/C-CYWrvWAAU8AXH.jpg	47534.0	19035.0
6291	https://pbs.twimg.com/media/C2tugXLXgAArJO4.jpg	131903.0	47748.0
6691	https://pbs.twimg.com/media/CmyPXNOW8AEtaJ-.jpg	5586.0	1465.0
6448	https://pbs.twimg.com/media/CwH_foYWgAEvTyI.jpg	16959.0	6458.0
6060	https://pbs.twimg.com/media/DCEeLxjXsAAvNSM.jpg	35264.0	10860.0
5994	https://pbs.twimg.com/media/DFi579UWsAAatzw.jpg	38609.0	8462.0
6116	https://pbs.twimg.com/media/C-3wvtxXcAUTuBE.jpg	19351.0	1687.0
6481	https://pbs.twimg.com/media/Cu6I9vvWIAAZG0a.jpg	11333.0	3207.0
6308	https://pbs.twimg.com/media/C2EONHNWQAUWxkP.jpg	5882.0	1353.0
6795	https://pbs.twimg.com/media/Cj_P7rSUgAAYQbz.jpg	5538.0	1746.0
6753	https://pbs.twimg.com/media/ClbBg4WWEAMjwJu.jpg	3256.0	705.0
6661	https://pbs.twimg.com/media/Cn7U2xlW8AI9Pqp.jpg	7066.0	1726.0
6516	https://pbs.twimg.com/media/CtZtJxAXEAAyPGd.jpg	24005.0	8436.0
6402	https://pbs.twimg.com/media/CyIgaTEVEAA-9zS.jpg	10055.0	3011.0
6394	https://pbs.twimg.com/media/CyeTku-XcAALkBd.jpg	11137.0	3181.0
6463	https://pbs.twimg.com/media/CvoBPWRWgAA4het.jpg	18464.0	5432.0
6087	https://pbs.twimg.com/media/DAmyy8FXYAIH8Ty.jpg	16630.0	2663.0
6044	https://pbs.twimg.com/media/DDIKMXzW0AEibje.jpg	19637.0	4259.0
6279	https://pbs.twimg.com/media/C3TjvitXAAAI-QH.jpg	56051.0	19486.0
6764	https://pbs.twimg.com/media/ClCQzFUUYAA5vAu.jpg	4584.0	1355.0

	user_followers	dog_stage	prediction_algorithm	\
1918	4484774.0	None	Saint_Bernard	
1917	4484774.0	None	Rottweiler	
1916	4484774.0	None	NaN	
1915	4484774.0	None	NaN	
1914	4484774.0	None	Labrador_retriever	
1913	4484774.0	None	Tibetan_terrier	
1912	4484774.0	None	toy_poodle	
1911	4484774.0	None	borzoi	
1910	4484773.0	None	Chihuahua	
1909	4484773.0	None	Labrador_retriever	
1908	4484773.0	None	beagle	
1907	4484773.0	None	Italian_greyhound	
1906	4484773.0	None	Chesapeake_Bay_retriever	
1905	4484773.0	None	golden_retriever	
1904	4484773.0	None	Pembroke	
1903	4484773.0	None	Yorkshire_terrier	
1902	4484773.0	None	Pomeranian	
1901	4484773.0	None	toy_poodle	
1900	4484773.0	None	NaN	
1899	4484773.0	None	NaN	

1898	4484773.0	None	kuvasz
1897	4484773.0	None	NaN
1896	4484773.0	None	West_Highland_white_terrier
1895	4484773.0	None	NaN
1894	4484773.0	None	dalmatian
1893	4484773.0	None	flat-coated_retriever
1892	4484773.0	None	Chihuahua
1919	4484775.0	None	papillon
1891	4484773.0	None	miniature_pinscher
1920	4484775.0	None	NaN
...	...	...	...
5200	4484525.0	pupper	malamute
5513	4484610.0	pupper	NaN
5204	4484654.0	pupper	Pomeranian
4064	4484674.0	pupper	cocker_spaniel
5543	4484741.0	pupper	Brabancon_griffon
5057	4484642.0	pupper	kelpie
5209	4484654.0	pupper	toy_poodle
6281	4484689.0	puppo	Eskimo_dog
5996	4484669.0	puppo	golden_retriever
6680	4484497.0	puppo	Labrador_retriever
6130	4484678.0	puppo	flat-coated_retriever
6291	4484690.0	puppo	Lakeland_terrier
6691	4484497.0	puppo	Labrador_retriever
6448	4484701.0	puppo	Labrador_retriever
6060	4484674.0	puppo	Cardigan
5994	4484669.0	puppo	Pembroke
6116	4484676.0	puppo	golden_retriever
6481	4484705.0	puppo	German_shepherd
6308	4484690.0	puppo	American_Staffordshire_terrier
6795	4484503.0	puppo	chow
6753	4484501.0	puppo	Old_English_sheepdog
6661	4484496.0	puppo	Airedale
6516	4484449.0	puppo	NaN
6402	4484698.0	puppo	Eskimo_dog
6394	4484696.0	puppo	miniature_pinscher
6463	4484702.0	puppo	golden_retriever
6087	4484676.0	puppo	Eskimo_dog
6044	4484670.0	puppo	Italian_greyhound
6279	4484688.0	puppo	Rottweiler
6764	4484502.0	puppo	malamute

	confidence_level
1918	0.381377
1917	0.999091
1916	0.000000
1915	0.000000
1914	0.825670

1913	0.676376
1912	0.304175
1911	0.852088
1910	0.143957
1909	0.996709
1908	0.636169
1907	0.125176
1906	0.088122
1905	0.633037
1904	0.435254
1903	0.618957
1902	0.412893
1901	0.787424
1900	0.000000
1899	0.000000
1898	0.360159
1897	0.000000
1896	0.360465
1895	0.000000
1894	0.535073
1893	0.508392
1892	0.187155
1919	0.557009
1891	0.214200
1920	0.000000
...	...
5200	0.949609
5513	0.000000
5204	0.943331
4064	0.437216
5543	0.409909
5057	0.219179
5209	0.279604
6281	0.524454
5996	0.953442
6680	0.000010
6130	0.321676
6291	0.196015
6691	0.929390
6448	0.654762
6060	0.806674
5994	0.966327
6116	0.895529
6481	0.992339
6308	0.925505
6795	0.808737
6753	0.427481
6661	0.602957

```
6516          0.000000
6402          0.482498
6394          0.817066
6463          0.245773
6087          0.616457
6044          0.734684
6279          0.681495
6764          0.442612

[1994 rows x 15 columns]
```

1.5.5 Fill empty rating and Correct the bad ones

```
In [406]: # Print the values and check if there exist in the text
          df_master.rating_numerator.value_counts()
          df_master.rating_denominator.value_counts()
          print(df_master[df_master.rating_denominator == 170]['text'][2842])
          print(df_master[df_master.rating_numerator == 1776]['text'][2720])
          print(df_master[df_master.tweet_id == 786709082849828864]['text'][2497])
          print(df_master['text'][1918])
          print(df_master['text'][1917])
          print(df_master['text'][1916])
          print(df_master['text'][1911])
```

```
Say hello to this unbelievably well behaved squad of doggos. 204/170 would try to pet all at onc
This is Atticus. He's quite simply America af. 1776/10 https://t.co/GRXwMxLBkh
This is Logan, the Chow who lived. He solemnly swears he's up to lots of good. H*ckin magical af
This is Biden. Biden just tripped... 7/10 https://t.co/3Fm9PwLju1
Ermergerd 12/10 https://t.co/PQni2sjPsm
Never seen this breed before. Very pointy pup. Hurts when you cuddle. Still cute tho. 10/10 http
Two dogs in this one. Both are rare Jujitsu Pythagoreans. One slightly whiter than other. Long l
```

- **It turns that the rating maybe thousands not only tens or hundreds**
- **Some tweets have more than one dog so more than one rating**
- **some tweets have no rating**
- **some dogs have float rating as 9.75**

```
In [448]: # Get ratings and treat them depending to their situation
          ratings = df_master['text'].apply(lambda x: re.findall(r'(\d+(\.\d+)|(\d+))\/(\d+0)',
          len(ratings)
```

```
Out[448]: 1994
```

```
In [449]: # Add new columns to store the new ratings and the count of dogs in each tweet
          rating_numerator = []
          rating_denominator = []
          dogs_count = []
```

```
for rate in ratings:
    # Tweets with no rating
    if len(rate) == 0:
        rating_numerator.append('NaN')
        rating_denominator.append('NaN')
        dogs_count.append(1) # It has a picture so it is a dog

    # Tweets with one rate
    elif len(rate) == 1:
        rating_numerator.append((float(rate[0][0]) / (float(rate[0][-1])/10)))
        rating_denominator.append(float(rate[0][-1])
        dogs_count.append(float(rate[0][-1]) / 10)
                                # We assume that the ratings who had rating_denominato
                                # are for group of dogs i.e : https://t.co/yGQI3He3xv
    # we take the average of the tweet with more than one rating
    elif len(rate) > 1 and rate[0][-1] == '10':
        rating_plus = 0
        rating_avg = 0
        for i in range(len(rate)):
            rating_plus = rating_plus + float(rate[i][0])
        result_avg = (rating_plus / len(rate))
        rating_numerator.append(result_avg)
        rating_denominator.append(10)
        dogs_count.append(len(rate))
    else: # without this block I get ValueError: Length of values does not match lengt
          # We will try to catch the errors this  why and see why this happend
        rating_numerator.append('Error')
        rating_denominator.append('Error')
        dogs_count.append('Error')

df_master['new_rating_numerator'] = rating_numerator
df_master['new_rating_denominator'] = rating_denominator
df_master['dogs_count'] = dogs_count
df_master['new_rating_numerator'].value_counts()
```

Out[449]:

12.0	453
10.0	411
11.0	399
13.0	261
9.0	152
8.0	94
7.0	52
14.0	36
6.0	32
5.0	30
3.0	19
4.0	14
2.0	9

```
8.5                    4
1.0                    4
7.5                    3
9.5                    3
10.5                   2
Error                  2
0.0                    2
1776.0                 1
9.75                   1
11.27                  1
4.5                    1
5.5                    1
6.5                    1
11.26                  1
13.5                   1
9.666666666666666      1
420.0                  1
11.5                   1
NaN                    1
Name: new_rating_numerator, dtype: int64
```

```
In [450]: # Test 2919 2885
          print(df_master[df_master.new_rating_numerator == 'Error']['text'][2919])
          print(df_master[df_master.new_rating_numerator == 'Error']['text'][2885])
          print(df_master[df_master.new_rating_numerator == 1776.0]['text'][2720])
          print(df_master[df_master.new_rating_numerator == 420.0]['text'][3712])
```

```
This is Bluebert. He just saw that both #FinalFur match ups are split 50/50. Amazed af. 11/10 ht
Happy 4/20 from the squad! 13/10 for all https://t.co/eV1diwds8a
This is Atticus. He's quite simply America af. 1776/10 https://t.co/GRXwMxLBkh
After so many requests... here you go.

Good dogg. 420/10 https://t.co/yfAAo1gdeY
```

- **We notice that the two errors where for a special situation that has two fractions but only one of them is the valid one. we will could assess this manually and we will clean it manually as well.**
- **We still see two big numbers one of them for Atticus and the other for snop Dog we will exclude them in future visualisations**

```
In [451]: # Correct the errors
          tweet_id_11 = df_master[df_master.new_rating_numerator == 'Error']['tweet_id'][2919]
          tweet_id_13 = df_master[df_master.new_rating_numerator == 'Error']['tweet_id'][2885]

          df_master.loc[df_master['tweet_id'] == tweet_id_11, 'new_rating_numerator'] = 11
          df_master.loc[df_master['tweet_id'] == tweet_id_13, 'new_rating_numerator'] = 13

          df_master.loc[df_master['dogs_count'] == 'Error', 'dogs_count'] = 1
```

```
        df_master.loc[df_master['new_rating_denominator'] == 'Error', 'new_rating_denominator'

        # Test
        print(df_master.new_rating_numerator[df_master.tweet_id == tweet_id_11])
        print(df_master.new_rating_numerator[df_master.tweet_id == tweet_id_13])
```

```
2919    11
Name: new_rating_numerator, dtype: object
2885    13
Name: new_rating_numerator, dtype: object
```

```
In [452]: # Delete the old columns and update the names of the new ones
          df_master = df_master.drop(['rating_numerator', 'rating_denominator'], 1)

          # Rename columns
          df_master.rename(columns = {'new_rating_numerator': 'rating_numerator',
                                      'new_rating_denominator': 'rating_denominator'}, inplace =

          # Test
          list(df_master)
```

```
Out[452]: ['tweet_id',
           'timestamp',
           'source',
           'text',
           'expanded_urls',
           'name',
           'jpg_url',
           'favorites',
           'retweets',
           'user_followers',
           'dog_stage',
           'prediction_algorithm',
           'confidence_level',
           'rating_numerator',
           'rating_denominator',
           'dogs_count']
```

1.5.6 Fill empty names and Correct the bad ones

The examples we could notice from assessing the data visually: - This is [name] .. - Meet [name] .. - Say hallo to [name] .. - Here we have [name] .. - .. named [name] ..

We will treat those cases to get the names from the text of the tweet

```
In [488]: # Loop on all the texts and check if the comment has one of the above conditions
          # and append the result in a list
          dog_names = []
```

```
for text in df_master['text']:
    # Start with 'This is ' and the fisrt letter of the name is uppercase
    if text.startswith('This is ') and re.match(r'[A-Z].*', text.split()[2]):
        dog_names.append(text.split()[2].strip(',').strip('.'))
    # Start with 'Meet ' and the fisrt letter of the name is uppercase
    elif text.startswith('Meet ') and re.match(r'[A-Z].*', text.split()[1]):
        dog_names.append(text.split()[1].strip(',').strip('.'))
    # Start with 'Say hello to ' and the fisrt letter of the name is uppercase
    elif text.startswith('Say hello to ') and re.match(r'[A-Z].*', text.split()[3]):
        dog_names.append(text.split()[3].strip(',').strip('.'))
    # Start with 'Here we have ' and the fisrt letter of the name is uppercase
    elif text.startswith('Here we have ') and re.match(r'[A-Z].*', text.split()[3]):
        dog_names.append(text.split()[3].strip(',').strip('.'))
    # Contain 'named' and the fisrt letter of the name is uppercase
    elif 'named' in text and re.match(r'[A-Z].*', text.split()[text.split().index('nam
        dog_names.append(text.split()[text.split().index('named') + 1].strip(',').stri
    # No name specified or other style
    else:
        dog_names.append('NaN')

# Test
len(dog_names)

# Save the result in a new column 'dog_name'
df_master['dog_name'] = dog_names

# Test
print("New column dog_name count \n", df_master.dog_name.value_counts())
print("Old column name count \n", df_master.name.value_counts())
```

```
New column dog_name count
 NaN         625
Charlie      11
Lucy         10
Oliver       10
Cooper       10
Penny         9
Tucker        9
Winston       8
Sadie         8
Lola          7
Daisy         7
Stanley       6
Koda          6
Jax           6
Bo            6
Bella         6
Toby          6
```

```
Bailey        5
Scout         5
Louis         5
Milo          5
Leo           5
Oscar         5
Chester       5
Buddy         5
Rusty         5
Jerry         4
Finn          4
Clarence      4
Gary          4
             ...
Holly         1
Dixie         1
Kawhi         1
Baloo         1
Bobb          1
Banjo         1
Howie         1
Tyrone        1
Stewie        1
Lugan         1
Tobi          1
Bradlay       1
Emmy          1
Charl         1
Chubbs        1
Rhino         1
Jarod         1
Sandra        1
Trigger       1
Richie        1
Siba          1
Biden         1
Anthony       1
Barney        1
Hanz          1
Skittles      1
Burt          1
Miley         1
Timmy         1
Tater         1
Name: dog_name, dtype: int64
Old column name count
 None         546
a             55
```

```
Charlie       11
Lucy          10
Cooper        10
Oliver        10
Tucker         9
Penny          9
Winston        8
Sadie          8
Toby           7
the            7
Lola           7
Daisy          7
Bo             6
Jax            6
Koda           6
Bella          6
Stanley        6
an             6
Scout          5
Dave           5
Buddy          5
Milo           5
Louis          5
Oscar          5
Bailey         5
Leo            5
Chester        5
Rusty          5
              ...
Kawhi          1
Baloo          1
Pablo          1
Snoop          1
Dido           1
Cermet         1
Lugan          1
Bobb           1
Tobi           1
Bradlay        1
Emmy           1
Charl          1
Chubbs         1
Sprinkles      1
Jarod          1
Sandra         1
Trigger        1
Richie         1
Siba           1
```

```
Biden          1
Anthony        1
Barney         1
Hanz           1
Skittles       1
Burt           1
Miley          1
Timmy          1
Tyrone         1
Banjo          1
Tater          1
Name: name, dtype: int64
```

```
In [490]: # We acn see here that the 'NaN' result for the tweets with two names or no name
          df_master[df_master.dog_name == 'NaN']
```

Out[490]:

	tweet_id	timestamp	source
1917	667435689202614272	2015-11-19 20:14:03 +0000	Twitter for iPhone
1916	667437278097252352	2015-11-19 20:20:22 +0000	Twitter for iPhone
1915	667443425659232256	2015-11-19 20:44:47 +0000	Twitter for iPhone
1911	667491009379606528	2015-11-19 23:53:52 +0000	Twitter Web Client
1906	667524857454854144	2015-11-20 02:08:22 +0000	Twitter Web Client
1901	667546741521195010	2015-11-20 03:35:20 +0000	Twitter Web Client
1900	667549055577362432	2015-11-20 03:44:31 +0000	Twitter Web Client
1899	667724302356258817	2015-11-20 15:20:54 +0000	Twitter Web Client
1895	667782464991965184	2015-11-20 19:12:01 +0000	Twitter for iPhone
1894	667793409583771648	2015-11-20 19:55:30 +0000	Twitter for iPhone
1893	667801013445750784	2015-11-20 20:25:43 +0000	Twitter for iPhone
1891	667832474953625600	2015-11-20 22:30:44 +0000	Twitter for iPhone
1920	667369227918143488	2015-11-19 15:49:57 +0000	Twitter for iPhone
1948	666786068205871104	2015-11-18 01:12:41 +0000	Twitter for iPhone
1945	666826780179869698	2015-11-18 03:54:28 +0000	Twitter for iPhone
1944	666835007768551424	2015-11-18 04:27:09 +0000	Twitter for iPhone
1943	666837028449972224	2015-11-18 04:35:11 +0000	Twitter for iPhone
1939	667044094246576128	2015-11-18 18:17:59 +0000	Twitter for iPhone
1937	667065535570550784	2015-11-18 19:43:11 +0000	Twitter for iPhone
1936	667073648344346624	2015-11-18 20:15:26 +0000	Twitter for iPhone
1933	667138269671505920	2015-11-19 00:32:12 +0000	Twitter for iPhone
1927	667176164155375616	2015-11-19 03:02:47 +0000	Twitter for iPhone
1926	667177989038297088	2015-11-19 03:10:02 +0000	Twitter for iPhone
1924	667188689915760640	2015-11-19 03:52:34 +0000	Twitter for iPhone
1923	667192066997374976	2015-11-19 04:05:59 +0000	Twitter for iPhone
1888	667873844930215936	2015-11-21 01:15:07 +0000	Twitter for iPhone
1883	667911425562669056	2015-11-21 03:44:27 +0000	Twitter for iPhone
1880	667937095915278337	2015-11-21 05:26:27 +0000	Twitter for iPhone
1878	668142349051129856	2015-11-21 19:02:04 +0000	Twitter for iPhone
1877	668154635664932864	2015-11-21 19:50:53 +0000	Twitter for iPhone

```
...                  ...                        ...                  ...
5219  688916208532455424  2016-01-18 02:49:58 +0000  Twitter for iPhone
5051  703268521220972544  2016-02-26 17:20:56 +0000  Twitter for iPhone
5048  703407252292673536  2016-02-27 02:32:12 +0000  Twitter for iPhone
4771  743222593470234624  2016-06-15 23:24:09 +0000  Twitter for iPhone
4774  742465774154047488  2016-06-13 21:16:49 +0000  Twitter for iPhone
5488  675740360753160193  2015-12-12 18:13:51 +0000  Twitter for iPhone
5227  688519176466644993  2016-01-17 00:32:18 +0000  Twitter for iPhone
4290  824325613288833024  2017-01-25 18:38:36 +0000  Twitter for iPhone
4783  741067306818797568  2016-06-10 00:39:48 +0000  Twitter for iPhone
5477  675898130735476737  2015-12-13 04:40:46 +0000  Twitter for iPhone
5036  704761120771465216  2016-03-01 20:11:59 +0000  Twitter for iPhone
5234  687818504314159109  2016-01-15 02:08:05 +0000  Twitter for iPhone
5033  704859558691414016  2016-03-02 02:43:09 +0000  Twitter for iPhone
5468  676263575653122048  2015-12-14 04:52:55 +0000  Twitter for iPhone
4037  881536004380872706  2017-07-02 15:32:16 +0000  Twitter for iPhone
5540  674638615994089473  2015-12-09 17:15:54 +0000  Twitter for iPhone
5070  700864154249383937  2016-02-20 02:06:50 +0000  Twitter for iPhone
4249  831315979191906304  2017-02-14 01:35:49 +0000  Twitter Web Client
5505  675334060156301312  2015-12-11 15:19:21 +0000  Twitter for iPhone
5078  700151421916807169  2016-02-18 02:54:41 +0000  Twitter for iPhone
5200  690015576308211712  2016-01-21 03:38:27 +0000  Twitter for iPhone
5543  674447403907457024  2015-12-09 04:36:06 +0000  Twitter for iPhone
6680  752519690950500352  2016-07-11 15:07:30 +0000  Twitter for iPhone
6130  855851453814013952  2017-04-22 18:31:02 +0000  Twitter for iPhone
6291  822872901745569793  2017-01-21 18:26:02 +0000  Twitter for iPhone
5994  889665388333682689  2017-07-25 01:55:32 +0000  Twitter for iPhone
6116  859607811541651456  2017-05-03 03:17:27 +0000  Twitter for iPhone
6661  756275833623502848  2016-07-21 23:53:04 +0000  Twitter for iPhone
6516  780931614150983680  2016-09-28 00:46:20 +0000  Twitter for iPhone
6279  825535076884762624  2017-01-29 02:44:34 +0000  Twitter for iPhone

                                                  text  \
1917            Ermergerd 12/10 https://t.co/PQni2sjPsm
1916  Never seen this breed before. Very pointy pup...
1915  Exotic dog here. Long neck. Weird paws. Obsess...
1911  Two dogs in this one. Both are rare Jujitsu Py...
1906  Another topnotch dog. His name is Big Jumpy Ra...
1901  Here is George. George took a selfie of his ne...
1900  Never seen dog like this. Breathes heavy. Tilt...
1899  What a dog to start the day with. Very calm. L...
1895  Super rare dog. Endangered (?). Thinks it's fu...
1894  Dogs only please. Small cows and other non can...
1893               OMIGOD 12/10 https://t.co/SVMF4Frf1w
1891  THE EYES 12/10\n\nI'm sorry. These are suppose...
1920  Here we have a neat pup. Very white. Cool shad...
1948  Unfamiliar with this breed. Ears pointy af. Wo...
1945  12/10 simply brilliant pup https://t.co/V6ZzG4...
```

```
1944  These are Peruvian Feldspars. Their names are ...
1943  My goodness. Very rare dog here. Large. Tail d...
1939            12/10 gimme now https://t.co/QZAnwgnOMB
1937  Here we have a Hufflepuff. Loves vest. Eyes wi...
1936  Here is Dave. He is actually just a skinny leg...
1933  Extremely intelligent dog here. Has learned to...
1927  These are strange dogs. All have toupees. Long...
1926  This is a Dasani Kingfisher from Maine. His na...
1924  Quite an advanced dog here. Impressively dress...
1923  *takes several long deep breaths* omg omg oMG ...
1888  Neat dog. Lots of spikes. Always in push-up po...
1883  Wow. Armored dog here. Ready for battle. Face ...
1880  This dog resembles a baked potato. Bed looks u...
1878  This lil pup is Oliver. Hops around. Has wings...
1877  Fun dogs here. Top one clearly an athlete. Bot...
...                                                 ...
5219  This pupper just wants to say hello. 11/10 wou...
5051  Happy Friday here's a sleepy pupper 12/10 http...
5048  This pupper doesn't understand gates. 10/10 so...
4771  This is a very rare Great Alaskan Bush Pupper...
4774  Was just informed about this hero pupper and o...
5488  Here's a pupper licking in slow motion. 12/10 ...
5227  This pupper is sprouting a flower out of her h...
4290  Retweet the h*ck out of this 13/10 pupper #Bel...
4783  This is just downright precious af. 12/10 for ...
5477  I'm sure you've all seen this pupper. Not prep...
5036  This pupper killed this great white in an epic...
5234  With great pupper comes great responsibility. ...
5033  Here is a heartbreaking scene of an incredible...
5468  All this pupper wanted to do was go skiing. No...
4037  Here is a pupper approaching maximum borkdrive...
5540  This pupper is fed up with being tickled. 12/1...
5070  "Pupper is a present to world. Here is a bow f...
4249  I couldn't make it to the #WKCDogShow BUT I ha...
5505  Good morning here's a grass pupper. 12/10 http...
5078  If a pupper gave that to me I'd probably start...
5200  This pupper can only sleep on shoes. It's a cr...
5543  This pupper just wants a belly rub. This puppe...
6680  Hopefully this puppo on a swing will help get ...
6130  Here's a puppo participating in the #ScienceMa...
6291  Here's a super supportive puppo participating ...
5994  Here's a puppo that seems to be on the fence a...
6116  Sorry for the lack of posts today. I came home...
6661  When ur older siblings get to play in the deep...
6516  I want to finally rate this iconic puppo who t...
6279  Here's a very loving and accepting puppo. Appe...

                                          expanded_urls    name  \
```

1917	https://twitter.com/dog_rates/status/667435689...	None
1916	https://twitter.com/dog_rates/status/667437278...	None
1915	https://twitter.com/dog_rates/status/667443425...	None
1911	https://twitter.com/dog_rates/status/667491009...	None
1906	https://twitter.com/dog_rates/status/667524857...	None
1901	https://twitter.com/dog_rates/status/667546741...	George
1900	https://twitter.com/dog_rates/status/667549055...	None
1899	https://twitter.com/dog_rates/status/667724302...	None
1895	https://twitter.com/dog_rates/status/667782464...	None
1894	https://twitter.com/dog_rates/status/667793409...	None
1893	https://twitter.com/dog_rates/status/667801013...	None
1891	https://twitter.com/dog_rates/status/667832474...	None
1920	https://twitter.com/dog_rates/status/667369227...	None
1948	https://twitter.com/dog_rates/status/666786068...	None
1945	https://twitter.com/dog_rates/status/666826780...	None
1944	https://twitter.com/dog_rates/status/666835007...	None
1943	https://twitter.com/dog_rates/status/666837028...	None
1939	https://twitter.com/dog_rates/status/667044094...	None
1937	https://twitter.com/dog_rates/status/667065535...	None
1936	https://twitter.com/dog_rates/status/667073648...	Dave
1933	https://twitter.com/dog_rates/status/667138269...	None
1927	https://twitter.com/dog_rates/status/667176164...	None
1926	https://twitter.com/dog_rates/status/667177989...	a
1924	https://twitter.com/dog_rates/status/667188689...	None
1923	https://twitter.com/dog_rates/status/667192066...	None
1888	https://twitter.com/dog_rates/status/667873844...	None
1883	https://twitter.com/dog_rates/status/667911425...	None
1880	https://twitter.com/dog_rates/status/667937095...	None
1878	https://twitter.com/dog_rates/status/668142349...	None
1877	https://twitter.com/dog_rates/status/668154635...	None
...	...	...
5219	https://twitter.com/dog_rates/status/688916208...	None
5051	https://twitter.com/dog_rates/status/703268521...	None
5048	https://twitter.com/dog_rates/status/703407252...	None
4771	https://twitter.com/dog_rates/status/743222593...	a
4774	https://twitter.com/dog_rates/status/742465774...	None
5488	https://twitter.com/dog_rates/status/675740360...	None
5227	https://twitter.com/dog_rates/status/688519176...	None
4290	https://twitter.com/dog_rates/status/824325613...	None
4783	https://twitter.com/dog_rates/status/741067306...	just
5477	https://twitter.com/dog_rates/status/675898130...	None
5036	https://twitter.com/dog_rates/status/704761120...	None
5234	https://twitter.com/dog_rates/status/687818504...	None
5033	https://twitter.com/dog_rates/status/704859558...	a
5468	https://twitter.com/dog_rates/status/676263575...	None
4037	https://twitter.com/dog_rates/status/881536004...	a
5540	https://twitter.com/dog_rates/status/674638615...	None
5070	https://twitter.com/dog_rates/status/700864154...	a

```
4249  https://twitter.com/dog_rates/status/831315979...    None
5505  https://twitter.com/dog_rates/status/675334060...    None
5078  https://twitter.com/dog_rates/status/700151421...    None
5200  https://twitter.com/dog_rates/status/690015576...    None
5543  https://twitter.com/dog_rates/status/674447403...    None
6680  https://twitter.com/dog_rates/status/752519690...    None
6130  https://twitter.com/dog_rates/status/855851453...    None
6291  https://twitter.com/dog_rates/status/822872901...    None
5994  https://twitter.com/dog_rates/status/889665388...    None
6116  https://twitter.com/dog_rates/status/859607811...    None
6661  https://twitter.com/dog_rates/status/756275833...    None
6516  https://twitter.com/dog_rates/status/780931614...    None
6279  https://twitter.com/dog_rates/status/825535076...    None

                                                  jpg_url  favorites  retweets  \
1917    https://pbs.twimg.com/media/CUM1OHCW4AEgGSi.jpg      324.0      88.0
1916    https://pbs.twimg.com/media/CUM2qWaWoAUZO6L.jpg      478.0     255.0
1915    https://pbs.twimg.com/media/CUM8QZwW4AAVsBl.jpg      823.0     616.0
1911    https://pbs.twimg.com/media/CUNniSlUYAEj1Jl.jpg      556.0     237.0
1906    https://pbs.twimg.com/media/CUOGUfJW4AA_eni.jpg     1786.0    1188.0
1901    https://pbs.twimg.com/media/CUOaOWXWcAAO_Jy.jpg      350.0     135.0
1900    https://pbs.twimg.com/media/CUOcVCwWsAERUKY.jpg     6079.0    2438.0
1899    https://pbs.twimg.com/media/CUQ7tv3W4AA3KlI.jpg      515.0     340.0
1895    https://pbs.twimg.com/media/CURwm3cUkAARcO6.jpg      433.0     257.0
1894    https://pbs.twimg.com/media/CUR6jqVWsAEgGot.jpg      731.0     356.0
1893    https://pbs.twimg.com/media/CUSBemVUEAAn-6V.jpg      346.0     101.0
1891    https://pbs.twimg.com/media/CUSeGFNW4AAyyHC.jpg      301.0      68.0
1920    https://pbs.twimg.com/media/CUL4xR9UkAEdlJ6.jpg      385.0     171.0
1948    https://pbs.twimg.com/media/CUDmZIkWcAAIPPe.jpg      798.0     517.0
1945    https://pbs.twimg.com/media/CUELaONUkAAscGC.jpg      264.0     103.0
1944    https://pbs.twimg.com/media/CUES51dXIAEahyG.jpg      222.0      82.0
1943    https://pbs.twimg.com/media/CUEUva1WsAA2jPb.jpg      849.0     577.0
1939    https://pbs.twimg.com/media/CUHREBXXAAE6A9b.jpg      197.0      52.0
1937    https://pbs.twimg.com/media/CUHkkJpXIAA2w3n.jpg      174.0      50.0
1936    https://pbs.twimg.com/media/CUHr8WbWEAEBPgf.jpg      422.0     131.0
1933    https://pbs.twimg.com/media/CUImtzEVAAAZNJo.jpg     4815.0    2358.0
1927    https://pbs.twimg.com/media/CUJJLtWWsAE-go5.jpg      636.0     482.0
1926    https://pbs.twimg.com/media/CUJK18UWEAEg7AR.jpg      200.0      58.0
1924    https://pbs.twimg.com/media/CUJUk2iWUAAVtOv.jpg      781.0     446.0
1923    https://pbs.twimg.com/media/CUJXpRBXIAANOyz.jpg      411.0     114.0
1888    https://pbs.twimg.com/media/CUTDtyGXIAARxus.jpg      662.0     437.0
1883    https://pbs.twimg.com/media/CUTl5m1WUAAabZG.jpg      520.0     325.0
1880    https://pbs.twimg.com/media/CUT9PuQWwAABQv7.jpg     1348.0     854.0
1878    https://pbs.twimg.com/media/CUW37BzWsAAlJlN.jpg      591.0     306.0
1877    https://pbs.twimg.com/media/CUXDGR2WcAAUQKz.jpg      518.0     335.0
...                                                  ...        ...       ...
5219    https://pbs.twimg.com/media/CY-Fn1FWEAQhzhs.jpg     2976.0     975.0
5051    https://pbs.twimg.com/media/CcKC-5LW4AAK-nb.jpg     2137.0     616.0
```

5048	https://pbs.twimg.com/media/CcMBJODUsAI5-A9.jpg	2670.0	777.0
4771	https://pbs.twimg.com/media/ClBO9zOWYAAA1jz.jpg	6748.0	2141.0
4774	https://pbs.twimg.com/media/Ck3EribXEAAPhZn.jpg	7873.0	4342.0
5488	https://pbs.twimg.com/ext_tw_video_thumb/67574...	1248.0	382.0
5227	https://pbs.twimg.com/media/CY4ciRFUMAAovos.jpg	2507.0	817.0
4290	https://pbs.twimg.com/media/C3CXxaoWQAAiLuC.jpg	12903.0	11748.0
4783	https://pbs.twimg.com/media/CkjMx99UoAM2B1a.jpg	10271.0	3479.0
5477	https://pbs.twimg.com/media/CWFFt3_XIAArIYK.jpg	1758.0	643.0
5036	https://pbs.twimg.com/media/CcfQgHVWoAAxauy.jpg	7217.0	3233.0
5234	https://pbs.twimg.com/media/CYufR8_WQAAWCqo.jpg	2740.0	1072.0
5033	https://pbs.twimg.com/media/CcgqBNVW8AE76lv.jpg	2446.0	607.0
5468	https://pbs.twimg.com/media/CWKSIfUUYAAiOBO.jpg	2223.0	603.0
4037	https://pbs.twimg.com/ext_tw_video_thumb/88153...	50139.0	16477.0
5540	https://pbs.twimg.com/media/CVzMPh1UsAELQ_p.jpg	1785.0	643.0
5070	https://pbs.twimg.com/media/Cbn4OqKWwAADGWt.jpg	2812.0	681.0
4249	https://pbs.twimg.com/media/C4lstObXAAE6MP8.jpg	7061.0	1256.0
5505	https://pbs.twimg.com/media/CV9EvZNUwAAgLCK.jpg	2981.0	1421.0
5078	https://pbs.twimg.com/media/CbdwATgWwAABGID.jpg	2438.0	747.0
5200	https://pbs.twimg.com/media/CZNtgWhWkAAbq3W.jpg	2725.0	825.0
5543	https://pbs.twimg.com/media/CVweVUfW4AACPwI.jpg	1128.0	391.0
6680	https://pbs.twimg.com/media/CnF8qVDWYAAh0g1.jpg	8097.0	3881.0
6130	https://pbs.twimg.com/media/C-CYWrvWAAU8AXH.jpg	47534.0	19035.0
6291	https://pbs.twimg.com/media/C2tugXLXgAArJO4.jpg	131903.0	47748.0
5994	https://pbs.twimg.com/media/DFi579UWsAAatzw.jpg	38609.0	8462.0
6116	https://pbs.twimg.com/media/C-3wvtxXcAUTuBE.jpg	19351.0	1687.0
6661	https://pbs.twimg.com/media/Cn7U2xlW8AI9Pqp.jpg	7066.0	1726.0
6516	https://pbs.twimg.com/media/CtZtJxAXEAAyPGd.jpg	24005.0	8436.0
6279	https://pbs.twimg.com/media/C3TjvitXAAAI-QH.jpg	56051.0	19486.0

	user_followers	dog_stage	prediction_algorithm	confidence_level	\
1917	4484774.0	None	Rottweiler	0.999091	
1916	4484774.0	None	NaN	0.000000	
1915	4484774.0	None	NaN	0.000000	
1911	4484774.0	None	borzoi	0.852088	
1906	4484773.0	None	Chesapeake_Bay_retriever	0.088122	
1901	4484773.0	None	toy_poodle	0.787424	
1900	4484773.0	None	NaN	0.000000	
1899	4484773.0	None	NaN	0.000000	
1895	4484773.0	None	NaN	0.000000	
1894	4484773.0	None	dalmatian	0.535073	
1893	4484773.0	None	flat-coated_retriever	0.508392	
1891	4484773.0	None	miniature_pinscher	0.214200	
1920	4484775.0	None	NaN	0.000000	
1948	4484777.0	None	NaN	0.000000	
1945	4484777.0	None	Maltese_dog	0.359383	
1944	4484777.0	None	Airedale	0.448459	
1943	4484777.0	None	NaN	0.000000	
1939	4484776.0	None	golden_retriever	0.765266	

1937	4484775.0	None	NaN	0.000000
1936	4484775.0	None	Chihuahua	0.483682
1933	4484775.0	None	West_Highland_white_terrier	0.747713
1927	4484646.0	None	soft-coated_wheaten_terrier	0.318981
1926	4484775.0	None	vizsla	0.259249
1924	4484775.0	None	NaN	0.000000
1923	4484775.0	None	Rottweiler	0.283640
1888	4484773.0	None	NaN	0.000000
1883	4484772.0	None	NaN	0.000000
1880	4484772.0	None	NaN	0.000000
1878	4484772.0	None	NaN	0.000000
1877	4484772.0	None	NaN	0.000000
...	...	...	...	...
5219	4484656.0	pupper	Pembroke	0.430544
5051	4484513.0	pupper	kuvasz	0.038243
5048	4484642.0	pupper	NaN	0.000000
4771	4484502.0	pupper	kuvasz	0.350629
4774	4484502.0	pupper	NaN	0.000000
5488	4484738.0	pupper	golden_retriever	0.800495
5227	4484527.0	pupper	Pembroke	0.696372
4290	4484689.0	pupper	Pembroke	0.990793
4783	4484503.0	pupper	golden_retriever	0.843799
5477	4484736.0	pupper	Labrador_retriever	0.407430
5036	4484513.0	pupper	Chihuahua	0.100418
5234	4484527.0	pupper	Lakeland_terrier	0.873029
5033	4484513.0	pupper	pug	0.284428
5468	4484734.0	pupper	toy_poodle	0.098029
4037	4484671.0	pupper	Samoyed	0.281463
5540	4484741.0	pupper	Pomeranian	0.846986
5070	4484642.0	pupper	kuvasz	0.805857
4249	4484687.0	pupper	briard	0.982755
5505	4484739.0	pupper	Pembroke	0.773135
5078	4484642.0	pupper	Italian_greyhound	0.176838
5200	4484525.0	pupper	malamute	0.949609
5543	4484741.0	pupper	Brabancon_griffon	0.409909
6680	4484497.0	puppo	Labrador_retriever	0.000010
6130	4484678.0	puppo	flat-coated_retriever	0.321676
6291	4484690.0	puppo	Lakeland_terrier	0.196015
5994	4484669.0	puppo	Pembroke	0.966327
6116	4484676.0	puppo	golden_retriever	0.895529
6661	4484496.0	puppo	Airedale	0.602957
6516	4484449.0	puppo	NaN	0.000000
6279	4484688.0	puppo	Rottweiler	0.681495

	rating_numerator	rating_denominator	dogs_count	dog_name
1917	12	10	1	NaN
1916	10	10	1	NaN
1915	6	10	1	NaN

1911	7.5	10	2	NaN
1906	12	10	1	NaN
1901	9	10	1	NaN
1900	1	10	1	NaN
1899	7	10	1	NaN
1895	9	10	1	NaN
1894	8	10	1	NaN
1893	12	10	1	NaN
1891	12	10	1	NaN
1920	10	10	1	NaN
1948	2	10	1	NaN
1945	12	10	1	NaN
1944	10	10	2	NaN
1943	3	10	1	NaN
1939	12	10	1	NaN
1937	8	10	1	NaN
1936	10	10	1	NaN
1933	10	10	1	NaN
1927	4	10	1	NaN
1926	8	10	1	NaN
1924	10	10	1	NaN
1923	12	10	1	NaN
1888	10	10	1	NaN
1883	5	10	1	NaN
1880	3	10	1	NaN
1878	2	10	1	NaN
1877	9	10	1	NaN
...	...	...	...	...
5219	11	10	1	NaN
5051	12	10	1	NaN
5048	10	10	1	NaN
4771	12	10	1	NaN
4774	14	10	1	NaN
5488	12	10	1	NaN
5227	12	10	1	NaN
4290	13	10	1	NaN
4783	12	10	1	NaN
5477	10	10	1	NaN
5036	13	10	1	NaN
5234	12	10	1	NaN
5033	10	10	1	NaN
5468	10	10	1	NaN
4037	14	10	1	NaN
5540	12	10	1	NaN
5070	12	10	1	NaN
4249	13	10	1	NaN
5505	12	10	1	NaN
5078	11	10	1	NaN

```
5200                12                  10          1       NaN
5543                10                  10          1       NaN
6680                11                  10          1       NaN
6130                13                  10          1       NaN
6291                13                  10          1       NaN
5994                13                  10          1       NaN
6116                13                  10          1       NaN
6661                10                  10          1       NaN
6516                13                  10          1       NaN
6279                14                  10          1       NaN

[625 rows x 17 columns]
```

In [491]:
```
# Let's delete the old name column now
df_master = df_master.drop(['name'], 1)

# Test
list(df_master)
```

Out[491]:
```
['tweet_id',
 'timestamp',
 'source',
 'text',
 'expanded_urls',
 'jpg_url',
 'favorites',
 'retweets',
 'user_followers',
 'dog_stage',
 'prediction_algorithm',
 'confidence_level',
 'rating_numerator',
 'rating_denominator',
 'dogs_count',
 'dog_name']
```

1.5.7 Get Dogs gender column from text column

In [498]:
```
# Loop on all the texts and check if it has one of pronouns of male or female
# and append the result in a list

male = ['He', 'he', 'him', 'his', "he's", 'himself']
female = ['She', 'she', 'her', 'hers', 'herself', "she's"]

dog_gender = []

for text in df_master['text']:
    # Male
```

```
            if any(map(lambda v:v in male, text.split())):
                dog_gender.append('male')
            # Female
            elif any(map(lambda v:v in female, text.split())):
                dog_gender.append('female')
            # If group or not specified
            else:
                dog_gender.append('NaN')

        # Test
        len(dog_gender)

        # Save the result in a new column 'dog_name'
        df_master['dog_gender'] = dog_gender

        # Test
        print("dog_gender count \n", df_master.dog_gender.value_counts())
```

```
dog_gender count
 NaN       1132
male       636
female     226
Name: dog_gender, dtype: int64
```

1.5.8 Convert the null values to None type

```
In [505]: df_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1994 entries, 1918 to 6764
Data columns (total 17 columns):
tweet_id                1994 non-null object
timestamp               1994 non-null datetime64[ns]
source                  1994 non-null category
text                    1994 non-null object
expanded_urls           1994 non-null object
jpg_url                 1994 non-null object
favorites               1994 non-null int32
retweets                1994 non-null int32
user_followers          1994 non-null int32
dog_stage               1994 non-null category
prediction_algorithm    1994 non-null object
confidence_level        1994 non-null float64
rating_numerator        1994 non-null object
rating_denominator      1994 non-null object
dogs_count              1994 non-null object
dog_name                1994 non-null object
```

```
dog_gender              1994 non-null object
dtypes: category(2), datetime64[ns](1), float64(1), int32(3), object(10)
memory usage: 229.8+ KB
```

```
In [560]: df_master.loc[df_master['prediction_algorithm'] == 'NaN', 'prediction_algorithm'] = No
          df_master.loc[df_master['dog_name'] == 'NaN', 'dog_name'] = None
          df_master.loc[df_master['dog_gender'] == 'NaN', 'dog_gender'] = None
          df_master.loc[df_master['rating_numerator'] == 'NaN', 'rating_numerator'] = 0
          df_master.loc[df_master['rating_denominator'] == 'NaN', 'rating_denominator'] = 0
```

```
In [561]: # Test
          df_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1994 entries, 1918 to 6764
Data columns (total 17 columns):
tweet_id                1994 non-null object
timestamp               1994 non-null datetime64[ns]
source                  1994 non-null category
text                    1994 non-null object
expanded_urls           1994 non-null object
jpg_url                 1994 non-null object
favorites               1994 non-null int32
retweets                1994 non-null int32
user_followers          1994 non-null int32
dog_stage               1994 non-null category
prediction_algorithm    1686 non-null object
confidence_level        1994 non-null float64
rating_numerator        1993 non-null float64
rating_denominator      1993 non-null object
dogs_count              1994 non-null int32
dog_name                1369 non-null object
dog_gender              862 non-null category
dtypes: category(3), datetime64[ns](1), float64(2), int32(4), object(7)
memory usage: 208.4+ KB
```

1.5.9 Convert each column to its appropriate type

```
In [504]: df_master.dtypes
```

```
Out[504]: tweet_id                      object
          timestamp               datetime64[ns]
          source                        category
          text                            object
          expanded_urls                   object
          jpg_url                         object
          favorites                        int32
```

```
retweets                        int32
user_followers                  int32
dog_stage                    category
prediction_algorithm           object
confidence_level              float64
rating_numerator               object
rating_denominator             object
dogs_count                     object
dog_name                       object
dog_gender                     object
dtype: object
```

In [562]:
```
df_master['tweet_id'] = df_master['tweet_id'].astype(object)
df_master['timestamp'] = pd.to_datetime(df_master.timestamp)
df_master['source'] = df_master['source'].astype('category')
df_master['favorites'] = df_master['favorites'].astype(int)
df_master['retweets'] = df_master['retweets'].astype(int)
df_master['user_followers'] = df_master['user_followers'].astype(int)
df_master['dog_stage'] = df_master['dog_stage'].astype('category')
df_master['rating_numerator'] = df_master['rating_numerator'].astype(float)
df_master['rating_denominator'] = df_master['rating_denominator'].astype(float)
df_master['dogs_count'] = df_master['dogs_count'].astype(int)
df_master['dog_gender'] = df_master['dog_gender'].astype('category')

# Test
df_master.dtypes
```

Out[562]:
```
tweet_id                        object
timestamp              datetime64[ns]
source                       category
text                           object
expanded_urls                  object
jpg_url                        object
favorites                       int32
retweets                        int32
user_followers                  int32
dog_stage                    category
prediction_algorithm           object
confidence_level              float64
rating_numerator              float64
rating_denominator            float64
dogs_count                      int32
dog_name                       object
dog_gender                   category
dtype: object
```

1.5.10 Rename columns to be more expressive *and Clean if needed*

```
In [597]: df_master = df_master.rename(columns = {'timestamp': 'tweet_date', 'source': 'tweet_so
                                                  'expanded_urls': 'tweet_url', 'jpg_url': 'twee
                                                  'favorites': 'tweet_favorites', 'retweets': 't
                                                  'prediction_algorithm' : 'dog_breed'})
```

```
In [595]: # All rating_denominator has one value 10
          # We will delete this column
          print(df_master.rating_denominator.value_counts())
          df_master.drop('rating_denominator', 1, inplace = True)
```

```
10    1994
Name: rating_denominator, dtype: int64
```

```
In [601]: # Test
          list(df_master)
```

```
Out[601]: ['tweet_id',
           'tweet_date',
           'tweet_source',
           'tweet_text',
           'tweet_url',
           'tweet_picture_predicted',
           'tweet_favorites',
           'tweet_retweets',
           'user_followers',
           'dog_stage',
           'dog_breed',
           'confidence_level',
           'rating_numerator',
           'dogs_count',
           'dog_name',
           'dog_gender']
```

1.6 Storing, Analyzing, and Visualizing Data

```
In [606]: # Store the clean DataFrame in a CSV file
          df_master.to_csv('twitter_archive_master.csv', index=False, encoding = 'utf-8')
```

```
In [609]: df_master = pd.read_csv('twitter_archive_master.csv')
          df_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1994 entries, 0 to 1993
Data columns (total 16 columns):
tweet_id                  1994 non-null int64
tweet_date                1994 non-null object
```

```
tweet_source               1994 non-null object
tweet_text                 1994 non-null object
tweet_url                  1994 non-null object
tweet_picture_predicted    1994 non-null object
tweet_favorites            1994 non-null int64
tweet_retweets             1994 non-null int64
user_followers             1994 non-null int64
dog_stage                  1994 non-null object
dog_breed                  1686 non-null object
confidence_level           1994 non-null float64
rating_numerator           1993 non-null float64
dogs_count                 1994 non-null int64
dog_name                   1369 non-null object
dog_gender                 862 non-null object
dtypes: float64(2), int64(5), object(9)
memory usage: 249.3+ KB
```