

# wrangle\_act

December 16, 2018

## 0.1 Introduction

Real-world data can be disorderly and unclean, making analyses difficult. In this project, I have collected tweet data and details for the twitter archive of the user @dog\_rates, also known as WeRateDogs. WeRateDogs rates dogs with some humorous content about them. The ratings are out of 10, but these fun dogs all manage to score above 10. Since the twitter data consists primarily of tweets, additional data including favorite/retweet count data, and image predictions to identify dog breeds, were collected by querying Twitter's API and from alternate sources.

```
In [72]: import pandas as pd
import tweepy
import numpy as np
import matplotlib.pyplot as plt
import requests
import os
import json
```

## 0.2 Gathering

Sources: The data were collected from three sources. The basic tweet texts from 2015 to 2017, along with the information gleaned from it including the dog's name, stage and ratings were downloaded from the Udacity server as a csv file and saved as twitter-archive-enhanced.csv, and loaded into a dataframe using Pandas. Additional data were obtained by querying Twitter's API using the Tweepy library after creating a developer account, using the associated secure keys. The tweet\_id indicated json files were downloaded and stored in a text file tweet\_json.txt, and parsed to extract the favorite and retweet counts into a dataframe. Image\_based breed predictions for the dogs, openly available from a url was downloaded as a tsv (image-predictions.tsv) and loaded into a third dataframe.

```
In [73]: # Load the twitter archive downloaded from the Udacity server, into a dataframe
df_archive = pd.read_csv('twitter-archive-enhanced.csv')
```

```
In [74]: # Download the image-predictions file from the url below
url = "https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions.tsv"
file_name = url.split('/')[-1]
if not os.path.exists(file_name):
    response = requests.get(url).content
    with open(file_name, mode = 'wb') as file:
```

```

        file.write(response)
    # Load the tsv file into a dataframe
    df_impreds = pd.read_csv('image-predictions.tsv', sep = '\t')

In [75]: # Connect to the Twitter API using the Tweepy library (after setting up a developer a
    consumer_key = ''
    consumer_secret = ''
    access_token = ''
    access_secret = ''

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

    api = tweepy.API(auth)

In [76]: # Collect additional information (retweet count, favorite count) from the Twitter API
    file_name = 'tweet_json.txt'
    if not os.path.exists(file_name):
        tweet_list = []
        # get tweet data for each tweet id
        for count, tweet_id in enumerate(df_archive.tweet_id):
            # Catch exceptions
            try:
                tweet = api.get_status(id = tweet_id, tweet_mode='extended')
                tweet_list.append({
                    "tweet_id": tweet_id,
                    "retweet_count": tweet.retweet_count,
                    "favorite_count": tweet.favorite_count,
                })
            except:
                print('{} Tweet id not found'.format(tweet_id))

        # save tweet details into a text file
        with open(file_name, 'w') as outfile:
            json.dump(tweet_list, outfile)

In [77]: # Create the dataframe from the text file
    with open('tweet_json.txt') as handle:
        json_file = json.load(handle)
    df_add1 = pd.DataFrame(json_file)

```

### 0.3 Assessment

```

In [78]: # Preliminary analyses
    df_archive.head(100)

```

```

Out[78]:
      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
..	...	...	...
70	879008229531029506	NaN	NaN
71	878776093423087618	NaN	NaN
72	878604707211726852	NaN	NaN
73	878404777348136964	NaN	NaN
74	878316110768087041	NaN	NaN
75	878281511006478336	NaN	NaN
76	878057613040115712	NaN	NaN
77	877736472329191424	NaN	NaN
78	877611172832227328	NaN	NaN
79	877556246731214848	NaN	NaN
80	877316821321428993	NaN	NaN
81	877201837425926144	NaN	NaN
82	876838120628539392	NaN	NaN
83	876537666061221889	NaN	NaN
84	876484053909872640	NaN	NaN
85	876120275196170240	NaN	NaN
86	875747767867523072	NaN	NaN
87	875144289856114688	NaN	NaN
88	875097192612077568	NaN	NaN

89	875021211251597312	NaN	NaN
90	874680097055178752	NaN	NaN
91	874434818259525634	NaN	NaN
92	874296783580663808	NaN	NaN
93	874057562936811520	NaN	NaN
94	874012996292530176	NaN	NaN
95	873697596434513921	NaN	NaN
96	873580283840344065	NaN	NaN
97	873337748698140672	NaN	NaN
98	873213775632977920	NaN	NaN
99	872967104147763200	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
..	...
70	2017-06-25 16:07:47 +0000
71	2017-06-25 00:45:22 +0000
72	2017-06-24 13:24:20 +0000
73	2017-06-24 00:09:53 +0000

74 2017-06-23 18:17:33 +0000  
 75 2017-06-23 16:00:04 +0000  
 76 2017-06-23 01:10:23 +0000  
 77 2017-06-22 03:54:17 +0000  
 78 2017-06-21 19:36:23 +0000  
 79 2017-06-21 15:58:08 +0000  
 80 2017-06-21 00:06:44 +0000  
 81 2017-06-20 16:29:50 +0000  
 82 2017-06-19 16:24:33 +0000  
 83 2017-06-18 20:30:39 +0000  
 84 2017-06-18 16:57:37 +0000  
 85 2017-06-17 16:52:05 +0000  
 86 2017-06-16 16:11:53 +0000  
 87 2017-06-15 00:13:52 +0000  
 88 2017-06-14 21:06:43 +0000  
 89 2017-06-14 16:04:48 +0000  
 90 2017-06-13 17:29:20 +0000  
 91 2017-06-13 01:14:41 +0000  
 92 2017-06-12 16:06:11 +0000  
 93 2017-06-12 00:15:36 +0000  
 94 2017-06-11 21:18:31 +0000  
 95 2017-06-11 00:25:14 +0000  
 96 2017-06-10 16:39:04 +0000  
 97 2017-06-10 00:35:19 +0000  
 98 2017-06-09 16:22:42 +0000  
 99 2017-06-09 00:02:31 +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...  
 .. ..  
 70 <a href="http://twitter.com/download/iphone" r...  
 71 <a href="http://twitter.com/download/iphone" r...  
 72 <a href="http://twitter.com/download/iphone" r...  
 73 <a href="http://twitter.com/download/iphone" r...  
 74 <a href="http://twitter.com/download/iphone" r...  
 75 <a href="http://twitter.com/download/iphone" r...  
 76 <a href="http://twitter.com/download/iphone" r...  
 77 <a href="http://twitter.com/download/iphone" r...  
 78 <a href="http://twitter.com/download/iphone" r...  
 79 <a href="http://twitter.com/download/iphone" r...  
 80 <a href="http://twitter.com/download/iphone" r...  
 81 <a href="http://twitter.com/download/iphone" r...  
 82 <a href="http://twitter.com/download/iphone" r...  
 83 <a href="http://twitter.com/download/iphone" r...  
 84 <a href="http://twitter.com/download/iphone" r...  
 85 <a href="http://twitter.com/download/iphone" r...  
 86 <a href="http://twitter.com/download/iphone" r...  
 87 <a href="http://twitter.com/download/iphone" r...  
 88 <a href="http://twitter.com/download/iphone" r...  
 89 <a href="http://twitter.com/download/iphone" r...  
 90 <a href="http://twitter.com/download/iphone" r...  
 91 <a href="http://twitter.com/download/iphone" r...  
 92 <a href="http://twitter.com/download/iphone" r...  
 93 <a href="http://twitter.com/download/iphone" r...  
 94 <a href="http://twitter.com/download/iphone" r...  
 95 <a href="http://twitter.com/download/iphone" r...  
 96 <a href="http://twitter.com/download/iphone" r...  
 97 <a href="http://twitter.com/download/iphone" r...  
 98 <a href="http://twitter.com/download/iphone" r...  
 99 <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
..	...	...
70	This is Beau. That is Beau's balloon. He takes...	NaN
71	This is Snoopy. He's a proud #PrideMonthPuppo...	NaN
72	Martha is stunning how h*ckin dare you. 13/10 ...	NaN
73	RT @dog_rates: Meet Shadow. In an attempt to r...	8.782815e+17
74	RT @dog_rates: Meet Terrance. He's being yelle...	6.690004e+17
75	Meet Shadow. In an attempt to reach maximum zo...	NaN
76	This is Emmy. She was adopted today. Massive r...	NaN
77	This is Aja. She was just told she's a good do...	NaN
78	RT @rachel2195: @dog_rates the boyfriend and h...	8.768508e+17
79	This is Penny. She's both pupset and fired pup...	NaN
80	Meet Dante. At first he wasn't a fan of his ne...	NaN
81	This is Nelly. He graduated with his dogtorate...	NaN
82	This is Ginger. She's having a ruff Monday. To...	NaN
83	I can say with the pupmost confidence that the...	NaN
84	This is Benedict. He wants to thank you for th...	NaN
85	Meet Venti, a seemingly caffeinated puppoccino...	NaN
86	This is Goose. He's a womanizer. Cheeky as h*c...	NaN
87	Meet Nugget and Hank. Nugget took Hank's bone...	NaN
88	You'll get your package when that precious man...	NaN
89	Guys please stop sending pictures without any ...	NaN
90	Meet Cash. He hath acquired a stick. A very go...	NaN
91	RT @dog_rates: This is Coco. At first I though...	8.663350e+17

92	This is Jed. He may be the fanciest pupper in ...	NaN
93	I can't believe this keeps happening. This, is...	NaN
94	This is Sebastian. He can't see all the colors...	NaN
95	RT @dog_rates: This is Walter. He won't start ...	8.688804e+17
96	We usually don't rate Deck-bound Saskatoon Bla...	NaN
97	RT @dog_rates: This is Sierra. She's one preci...	8.732138e+17
98	This is Sierra. She's one precious pupper. Abs...	NaN
99	Here's a very large dog. He has a date later. ...	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	
..	...	...	
70	NaN	NaN	
71	NaN	NaN	
72	NaN	NaN	
73	4.196984e+09	2017-06-23 16:00:04	+0000
74	4.196984e+09	2015-11-24 03:51:38	+0000
75	NaN	NaN	
76	NaN	NaN	



77	NaN	NaN
78	5.128045e+08	2017-06-19 17:14:49 +0000
79	NaN	NaN
80	NaN	NaN
81	NaN	NaN
82	NaN	NaN
83	NaN	NaN
84	NaN	NaN
85	NaN	NaN
86	NaN	NaN
87	NaN	NaN
88	NaN	NaN
89	NaN	NaN
90	NaN	NaN
91	4.196984e+09	2017-05-21 16:48:45 +0000
92	NaN	NaN
93	NaN	NaN
94	NaN	NaN
95	4.196984e+09	2017-05-28 17:23:24 +0000
96	NaN	NaN
97	4.196984e+09	2017-06-09 16:22:42 +0000
98	NaN	NaN
99	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
..	...	...
70	https://twitter.com/dog_rates/status/879008229...	13
71	https://twitter.com/dog_rates/status/878776093...	13
72	https://twitter.com/bbcworld/status/8785998685...	13
73	https://www.gofundme.com/3yd6y1c,https://twitt...	13
74	https://twitter.com/dog_rates/status/669000397...	11
75	https://www.gofundme.com/3yd6y1c,https://twitt...	13
76	https://twitter.com/dog_rates/status/878057613...	14
77	https://twitter.com/dog_rates/status/877736472...	13
78	https://twitter.com/rachel2195/status/87685077...	14
79	https://twitter.com/dog_rates/status/877556246...	12
80	https://twitter.com/dog_rates/status/877316821...	13
81	https://twitter.com/dog_rates/status/877201837...	12
82	https://twitter.com/dog_rates/status/876838120...	12
83	https://twitter.com/mpstowerham/status/8761629...	14
84	https://twitter.com/dog_rates/status/876484053...	13
85	https://twitter.com/dog_rates/status/876120275...	13
86	https://twitter.com/dog_rates/status/875747767...	13
87	https://twitter.com/dog_rates/status/875144289...	13
88	https://twitter.com/drboondoc/status/874413398...	13
89	https://twitter.com/dog_rates/status/875021211...	12
90	https://twitter.com/dog_rates/status/874680097...	12
91	https://twitter.com/dog_rates/status/866334964...	12
92	https://twitter.com/dog_rates/status/874296783...	13
93	https://twitter.com/dog_rates/status/874057562...	12
94	https://twitter.com/dog_rates/status/874012996...	13
95	https://twitter.com/dog_rates/status/868880397...	14
96	https://twitter.com/dog_rates/status/873580283...	13
97	https://www.gofundme.com/help-my-baby-sierra-g...	12
98	https://www.gofundme.com/help-my-baby-sierra-g...	12
99	https://twitter.com/dog_rates/status/872967104...	12

	rating_denominator	name	doggo	floofer	pupper	puppo
0	10	Phineas	None	None	None	None
1	10	Tilly	None	None	None	None
2	10	Archie	None	None	None	None
3	10	Darla	None	None	None	None
4	10	Franklin	None	None	None	None
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
..	...	...	...	...	...	...
70	10	Beau	None	None	None	None
71	10	Snoopy	None	None	None	puppo
72	10	None	None	None	None	None
73	10	Shadow	None	None	None	None
74	10	Terrance	None	None	None	None
75	10	Shadow	None	None	None	None
76	10	Emmy	None	None	None	None
77	10	Aja	None	None	None	None
78	10	None	None	None	pupper	None
79	10	Penny	None	None	None	None
80	10	Dante	None	None	None	None
81	10	Nelly	None	None	None	None
82	10	Ginger	None	None	pupper	None
83	10	None	None	None	None	None
84	10	Benedict	None	None	None	None
85	10	Venti	None	None	None	None
86	10	Goose	None	None	None	None
87	10	Nugget	None	None	None	None
88	10	None	None	None	None	None
89	10	None	None	None	None	None
90	10	Cash	None	None	None	None
91	10	Coco	None	None	None	None
92	10	Jed	None	None	pupper	None
93	10	None	None	None	None	None
94	10	Sebastian	None	None	None	puppo

95	10	Walter	None	None	None	None
96	10	None	None	None	None	None
97	10	Sierra	None	None	pupper	None
98	10	Sierra	None	None	pupper	None
99	10	None	doggo	None	None	None

[100 rows x 17 columns]

```
In [79]: df_archive.info()
```

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

```
In [80]: df_archive.doggo.nunique(), df_archive.floofer.nunique(), df_archive.pupper.nunique()
```

```
Out[80]: (2, 2, 2, 2)
```

```
In [81]: df_archive.expanded_urls[900]
```

```
Out[81]: 'https://twitter.com/dog_rates/status/758740312047005698/photo/1'
```

```
In [82]: df_archive.nunique()
```

```
Out[82]: tweet_id                2356
in_reply_to_status_id            77
in_reply_to_user_id              31
timestamp                        2356
source                            4
```

text	2356
retweeted_status_id	181
retweeted_status_user_id	25
retweeted_status_timestamp	181
expanded_urls	2218
rating_numerator	40
rating_denominator	18
name	957
doggo	2
floofer	2
pupper	2
puppo	2
dtype: int64	

In [83]: df\_impreds.head()

```
Out [83]:
```

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

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

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

In [84]: df\_impreds.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null int64
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
```

```

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

```

```
In [85]: df_add1.head()
```

```

Out[85]:   favorite_count  retweet_count      tweet_id
0           38149           8345  892420643555336193
1           32717           6170  892177421306343426
2           24637           4082  891815181378084864
3           41491           8481  891689557279858688
4           39658           9184  891327558926688256

```

```
In [86]: df_add1.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 887 entries, 0 to 886
Data columns (total 3 columns):
favorite_count      887 non-null int64
retweet_count       887 non-null int64
tweet_id            887 non-null int64
dtypes: int64(3)
memory usage: 20.9 KB

```

```

In [87]: # Are any rows duplicated?
         df_archive.duplicated().sum()

```

```
Out[87]: 0
```

```
In [88]: df_impreds.duplicated().sum()
```

```
Out[88]: 0
```

```
In [89]: df_add1.duplicated().sum()
```

```
Out[89]: 0
```

## 0.4 Cleaning

Upon assessment using the Pandas and Numpy functions, the following data tidiness and quality issues were identified. The prospective solutions are noted alongside.

## Tidiness:

1. Type of dogs split into 4 columns in df\_archive; melt to one
2. jpg\_url and img\_num columns in df\_impreds are not necessary for the planned analyses; remove these columns
3. df\_addl should be part of df\_archive; merge to the archive
4. Add df\_impreds also to the df\_archive for easier analyses

## Quality:

1. in\_reply\_to\_status\_id and in\_reply\_to\_user\_id columns are reply tweets; remove rows
2. retweeted\_status\_id, retweeted\_user\_id and retweeted\_time\_stamps columns represent retweets; remove rows
3. Timestamp as string; change to datetime format
4. Source column has 4 unique values which represent the actual sources, lost in the html format; Pick out the source from the html link
5. expanded\_urls column doesn't make sense; remove
6. Rating denominator should be 10 for all; change to 10
7. Rating numerator has low and high values; begin at 11, cap at 15
8. Datatype of tweet\_id is int; change to string

```
In [90]: # Copy the dataframes
df_archive_clean = df_archive.copy()
df_impreds_clean = df_impreds.copy()
df_addl_clean = df_addl.copy()
```

### 0.4.1 Tidiness

Define: Type of dogs split into 4 columns in df\_archive; melt to one

```
In [91]: # Code
df_archive_clean.replace('None', '', inplace = True)
df_archive_clean['dog_type'] = df_archive_clean[['doggo', 'floofer', 'pupper', 'puppo']]
df_archive_clean = df_archive_clean.drop(['doggo', 'floofer', 'pupper', 'puppo'], axis=1)
df_archive_clean.replace('', np.nan, inplace = True)
```

```
In [92]: # Test
df_archive_clean.dog_type.value_counts()
```

```
Out[92]: pupper          245
         doggo           83
         puppo           29
         doggopupper     12
         floofer          9
         doggopuppo       1
         doggofloofer      1
         Name: dog_type, dtype: int64
```

Define: jpg\_url and img\_num in df\_impreds are not necessary for the planned analyses; remove these columns

```
In [93]: # Code
df_impreds_clean = df_impreds_clean.drop(['jpg_url', 'img_num'], axis = 1)
```

```
In [94]: # Test
df_impreds_clean.head(1)
```

```
Out[94]:
```

	tweet_id		p1	p1_conf	p1_dog	p2	\
0	666020888022790149	Welsh_springer_spaniel	0.465074	True	collie		
	p2_conf	p2_dog	p3	p3_conf	p3_dog		
0	0.156665	True	Shetland_sheepdog	0.061428	True		

Define: df\_addl should be part of archive; merge to the archive

```
In [95]: # Code: Merge the additional data (favorite and retweet counts) to the archive dataframe
df_archive_clean = df_archive_clean.merge(df_addl_clean, on = 'tweet_id', how = 'left')
```

```
In [96]: # Test
df_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 16 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                     1611 non-null object
dog_type                  380 non-null object
favorite_count            887 non-null float64
retweet_count             887 non-null float64
dtypes: float64(6), int64(3), object(7)
memory usage: 312.9+ KB
```

```
In [97]: # Test that the null and real values in the favorite_count add up to the total number
df_archive_clean['favorite_count'].isnull().sum() + df_addl.shape[0]
```

```
Out[97]: 2356
```

```
In [98]: # Test
df_archive_clean.head(100)
```



```

Out[98]:
      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
..      ...
70   879008229531029506              NaN              NaN
71   878776093423087618              NaN              NaN
72   878604707211726852              NaN              NaN
73   878404777348136964              NaN              NaN
74   878316110768087041              NaN              NaN
75   878281511006478336              NaN              NaN
76   878057613040115712              NaN              NaN
77   877736472329191424              NaN              NaN
78   877611172832227328              NaN              NaN
79   877556246731214848              NaN              NaN
80   877316821321428993              NaN              NaN
81   877201837425926144              NaN              NaN
82   876838120628539392              NaN              NaN
83   876537666061221889              NaN              NaN
84   876484053909872640              NaN              NaN
85   876120275196170240              NaN              NaN

```

86	875747767867523072	NaN	NaN
87	875144289856114688	NaN	NaN
88	875097192612077568	NaN	NaN
89	875021211251597312	NaN	NaN
90	874680097055178752	NaN	NaN
91	874434818259525634	NaN	NaN
92	874296783580663808	NaN	NaN
93	874057562936811520	NaN	NaN
94	874012996292530176	NaN	NaN
95	873697596434513921	NaN	NaN
96	873580283840344065	NaN	NaN
97	873337748698140672	NaN	NaN
98	873213775632977920	NaN	NaN
99	872967104147763200	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
..	...
70	2017-06-25 16:07:47 +0000

71 2017-06-25 00:45:22 +0000  
 72 2017-06-24 13:24:20 +0000  
 73 2017-06-24 00:09:53 +0000  
 74 2017-06-23 18:17:33 +0000  
 75 2017-06-23 16:00:04 +0000  
 76 2017-06-23 01:10:23 +0000  
 77 2017-06-22 03:54:17 +0000  
 78 2017-06-21 19:36:23 +0000  
 79 2017-06-21 15:58:08 +0000  
 80 2017-06-21 00:06:44 +0000  
 81 2017-06-20 16:29:50 +0000  
 82 2017-06-19 16:24:33 +0000  
 83 2017-06-18 20:30:39 +0000  
 84 2017-06-18 16:57:37 +0000  
 85 2017-06-17 16:52:05 +0000  
 86 2017-06-16 16:11:53 +0000  
 87 2017-06-15 00:13:52 +0000  
 88 2017-06-14 21:06:43 +0000  
 89 2017-06-14 16:04:48 +0000  
 90 2017-06-13 17:29:20 +0000  
 91 2017-06-13 01:14:41 +0000  
 92 2017-06-12 16:06:11 +0000  
 93 2017-06-12 00:15:36 +0000  
 94 2017-06-11 21:18:31 +0000  
 95 2017-06-11 00:25:14 +0000  
 96 2017-06-10 16:39:04 +0000  
 97 2017-06-10 00:35:19 +0000  
 98 2017-06-09 16:22:42 +0000  
 99 2017-06-09 00:02:31 +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...
.. ..
70 <a href="http://twitter.com/download/iphone" r...
71 <a href="http://twitter.com/download/iphone" r...
72 <a href="http://twitter.com/download/iphone" r...
73 <a href="http://twitter.com/download/iphone" r...
74 <a href="http://twitter.com/download/iphone" r...
75 <a href="http://twitter.com/download/iphone" r...
76 <a href="http://twitter.com/download/iphone" r...
77 <a href="http://twitter.com/download/iphone" r...
78 <a href="http://twitter.com/download/iphone" r...
79 <a href="http://twitter.com/download/iphone" r...
80 <a href="http://twitter.com/download/iphone" r...
81 <a href="http://twitter.com/download/iphone" r...
82 <a href="http://twitter.com/download/iphone" r...
83 <a href="http://twitter.com/download/iphone" r...
84 <a href="http://twitter.com/download/iphone" r...
85 <a href="http://twitter.com/download/iphone" r...
86 <a href="http://twitter.com/download/iphone" r...
87 <a href="http://twitter.com/download/iphone" r...
88 <a href="http://twitter.com/download/iphone" r...
89 <a href="http://twitter.com/download/iphone" r...
90 <a href="http://twitter.com/download/iphone" r...
91 <a href="http://twitter.com/download/iphone" r...
92 <a href="http://twitter.com/download/iphone" r...
93 <a href="http://twitter.com/download/iphone" r...
94 <a href="http://twitter.com/download/iphone" r...
95 <a href="http://twitter.com/download/iphone" r...
96 <a href="http://twitter.com/download/iphone" r...
97 <a href="http://twitter.com/download/iphone" r...
98 <a href="http://twitter.com/download/iphone" r...
99 <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
..	...	...
70	This is Beau. That is Beau's balloon. He takes...	NaN
71	This is Snoopy. He's a proud #PrideMonthPuppo...	NaN
72	Martha is stunning how h*ckin dare you. 13/10 ...	NaN
73	RT @dog_rates: Meet Shadow. In an attempt to r...	8.782815e+17
74	RT @dog_rates: Meet Terrance. He's being yelle...	6.690004e+17
75	Meet Shadow. In an attempt to reach maximum zo...	NaN
76	This is Emmy. She was adopted today. Massive r...	NaN
77	This is Aja. She was just told she's a good do...	NaN
78	RT @rachel2195: @dog_rates the boyfriend and h...	8.768508e+17
79	This is Penny. She's both pupset and fired pup...	NaN
80	Meet Dante. At first he wasn't a fan of his ne...	NaN
81	This is Nelly. He graduated with his dogtorate...	NaN
82	This is Ginger. She's having a ruff Monday. To...	NaN
83	I can say with the pupmost confidence that the...	NaN
84	This is Benedict. He wants to thank you for th...	NaN
85	Meet Venti, a seemingly caffeinated puppoccino...	NaN
86	This is Goose. He's a womanizer. Cheeky as h*c...	NaN
87	Meet Nugget and Hank. Nugget took Hank's bone...	NaN
88	You'll get your package when that precious man...	NaN

89	Guys please stop sending pictures without any ...	NaN
90	Meet Cash. He hath acquired a stick. A very go...	NaN
91	RT @dog_rates: This is Coco. At first I though...	8.663350e+17
92	This is Jed. He may be the fanciest pupper in ...	NaN
93	I can't believe this keeps happening. This, is...	NaN
94	This is Sebastian. He can't see all the colors...	NaN
95	RT @dog_rates: This is Walter. He won't start ...	8.688804e+17
96	We usually don't rate Deck-bound Saskatoon Bla...	NaN
97	RT @dog_rates: This is Sierra. She's one preci...	8.732138e+17
98	This is Sierra. She's one precious pupper. Abs...	NaN
99	Here's a very large dog. He has a date later. ...	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
..	...	...
70	NaN	NaN
71	NaN	NaN
72	NaN	NaN
73	4.196984e+09	2017-06-23 16:00:04 +0000

74	4.196984e+09	2015-11-24 03:51:38	+0000
75	NaN		NaN
76	NaN		NaN
77	NaN		NaN
78	5.128045e+08	2017-06-19 17:14:49	+0000
79	NaN		NaN
80	NaN		NaN
81	NaN		NaN
82	NaN		NaN
83	NaN		NaN
84	NaN		NaN
85	NaN		NaN
86	NaN		NaN
87	NaN		NaN
88	NaN		NaN
89	NaN		NaN
90	NaN		NaN
91	4.196984e+09	2017-05-21 16:48:45	+0000
92	NaN		NaN
93	NaN		NaN
94	NaN		NaN
95	4.196984e+09	2017-05-28 17:23:24	+0000
96	NaN		NaN
97	4.196984e+09	2017-06-09 16:22:42	+0000
98	NaN		NaN
99	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	<a href="https://twitter.com/dog_rates/status/888078434...">https://twitter.com/dog_rates/status/888078434...</a>	12
21	<a href="https://twitter.com/dog_rates/status/887705289...">https://twitter.com/dog_rates/status/887705289...</a>	13
22	<a href="https://twitter.com/dog_rates/status/887517139...">https://twitter.com/dog_rates/status/887517139...</a>	14
23	<a href="https://twitter.com/dog_rates/status/887473957...">https://twitter.com/dog_rates/status/887473957...</a>	13
24	<a href="https://twitter.com/dog_rates/status/887343217...">https://twitter.com/dog_rates/status/887343217...</a>	13
25	<a href="https://twitter.com/dog_rates/status/887101392...">https://twitter.com/dog_rates/status/887101392...</a>	12
26	<a href="https://twitter.com/dog_rates/status/886983233...">https://twitter.com/dog_rates/status/886983233...</a>	13
27	<a href="https://www.gofundme.com/mingusneedsus">https://www.gofundme.com/mingusneedsus</a> , <a href="https://...">https://...</a>	13
28	<a href="https://twitter.com/dog_rates/status/886680336...">https://twitter.com/dog_rates/status/886680336...</a>	13
29	<a href="https://twitter.com/dog_rates/status/886366144...">https://twitter.com/dog_rates/status/886366144...</a>	12
..	...	...
70	<a href="https://twitter.com/dog_rates/status/879008229...">https://twitter.com/dog_rates/status/879008229...</a>	13
71	<a href="https://twitter.com/dog_rates/status/878776093...">https://twitter.com/dog_rates/status/878776093...</a>	13
72	<a href="https://twitter.com/bbcworld/status/8785998685...">https://twitter.com/bbcworld/status/8785998685...</a>	13
73	<a href="https://www.gofundme.com/3yd6y1c">https://www.gofundme.com/3yd6y1c</a> , <a href="https://twitt...">https://twitt...</a>	13
74	<a href="https://twitter.com/dog_rates/status/669000397...">https://twitter.com/dog_rates/status/669000397...</a>	11
75	<a href="https://www.gofundme.com/3yd6y1c">https://www.gofundme.com/3yd6y1c</a> , <a href="https://twitt...">https://twitt...</a>	13
76	<a href="https://twitter.com/dog_rates/status/878057613...">https://twitter.com/dog_rates/status/878057613...</a>	14
77	<a href="https://twitter.com/dog_rates/status/877736472...">https://twitter.com/dog_rates/status/877736472...</a>	13
78	<a href="https://twitter.com/rachel2195/status/87685077...">https://twitter.com/rachel2195/status/87685077...</a>	14
79	<a href="https://twitter.com/dog_rates/status/877556246...">https://twitter.com/dog_rates/status/877556246...</a>	12
80	<a href="https://twitter.com/dog_rates/status/877316821...">https://twitter.com/dog_rates/status/877316821...</a>	13
81	<a href="https://twitter.com/dog_rates/status/877201837...">https://twitter.com/dog_rates/status/877201837...</a>	12
82	<a href="https://twitter.com/dog_rates/status/876838120...">https://twitter.com/dog_rates/status/876838120...</a>	12
83	<a href="https://twitter.com/mpstowerham/status/8761629...">https://twitter.com/mpstowerham/status/8761629...</a>	14
84	<a href="https://twitter.com/dog_rates/status/876484053...">https://twitter.com/dog_rates/status/876484053...</a>	13
85	<a href="https://twitter.com/dog_rates/status/876120275...">https://twitter.com/dog_rates/status/876120275...</a>	13
86	<a href="https://twitter.com/dog_rates/status/875747767...">https://twitter.com/dog_rates/status/875747767...</a>	13
87	<a href="https://twitter.com/dog_rates/status/875144289...">https://twitter.com/dog_rates/status/875144289...</a>	13
88	<a href="https://twitter.com/drboondoc/status/874413398...">https://twitter.com/drboondoc/status/874413398...</a>	13
89	<a href="https://twitter.com/dog_rates/status/875021211...">https://twitter.com/dog_rates/status/875021211...</a>	12
90	<a href="https://twitter.com/dog_rates/status/874680097...">https://twitter.com/dog_rates/status/874680097...</a>	12
91	<a href="https://twitter.com/dog_rates/status/866334964...">https://twitter.com/dog_rates/status/866334964...</a>	12
92	<a href="https://twitter.com/dog_rates/status/874296783...">https://twitter.com/dog_rates/status/874296783...</a>	13
93	<a href="https://twitter.com/dog_rates/status/874057562...">https://twitter.com/dog_rates/status/874057562...</a>	12
94	<a href="https://twitter.com/dog_rates/status/874012996...">https://twitter.com/dog_rates/status/874012996...</a>	13
95	<a href="https://twitter.com/dog_rates/status/868880397...">https://twitter.com/dog_rates/status/868880397...</a>	14
96	<a href="https://twitter.com/dog_rates/status/873580283...">https://twitter.com/dog_rates/status/873580283...</a>	13
97	<a href="https://www.gofundme.com/help-my-baby-sierra-g...">https://www.gofundme.com/help-my-baby-sierra-g...</a>	12
98	<a href="https://www.gofundme.com/help-my-baby-sierra-g...">https://www.gofundme.com/help-my-baby-sierra-g...</a>	12
99	<a href="https://twitter.com/dog_rates/status/872967104...">https://twitter.com/dog_rates/status/872967104...</a>	12

	rating_denominator	name	dog_type	favorite_count	retweet_count
0	10	Phineas	NaN	38149.0	8345.0
1	10	Tilly	NaN	32717.0	6170.0
2	10	Archie	NaN	24637.0	4082.0
3	10	Darla	NaN	41491.0	8481.0
4	10	Franklin	NaN	39658.0	9184.0



5	10	NaN	NaN	19912.0	3055.0
6	10	Jax	NaN	11642.0	2028.0
7	10	NaN	NaN	64369.0	18526.0
8	10	Zoey	NaN	27359.0	4196.0
9	10	Cassie	doggo	31398.0	7250.0
10	10	Koda	NaN	30182.0	7202.0
11	10	Bruno	NaN	27345.0	4885.0
12	10	NaN	puppo	47340.0	9872.0
13	10	Ted	NaN	26714.0	4462.0
14	10	Stuart	puppo	14867.0	2206.0
15	10	Oliver	NaN	24849.0	5279.0
16	10	Jim	NaN	28627.0	4410.0
17	10	Zeke	NaN	25141.0	4217.0
18	10	Ralphus	NaN	19517.0	3489.0
19	10	Canela	NaN	NaN	NaN
20	10	Gerald	NaN	21415.0	3431.0
21	10	Jeffrey	NaN	29684.0	5289.0
22	10	such	NaN	45557.0	11465.0
23	10	Canela	NaN	67968.0	17845.0
24	10	NaN	NaN	33141.0	10219.0
25	10	NaN	NaN	30092.0	5858.0
26	10	Maya	NaN	34601.0	7632.0
27	10	Mingus	NaN	11860.0	3222.0
28	10	Derek	NaN	22085.0	4383.0
29	10	Roscoe	pupper	20856.0	3141.0
..	...	...	...	...	...
70	10	Beau	NaN	18695.0	2657.0
71	10	Snoopy	puppo	19123.0	4072.0
72	10	NaN	NaN	29810.0	7095.0
73	10	Shadow	NaN	0.0	1270.0
74	10	Terrance	NaN	0.0	6560.0
75	10	Shadow	NaN	7632.0	1270.0
76	10	Emmy	NaN	41419.0	6721.0
77	10	Aja	NaN	78539.0	18817.0
78	10	NaN	pupper	0.0	80.0
79	10	Penny	NaN	22398.0	3756.0
80	10	Dante	NaN	26986.0	5102.0
81	10	Nelly	NaN	26741.0	5513.0
82	10	Ginger	pupper	20394.0	3299.0
83	10	NaN	NaN	23178.0	4572.0
84	10	Benedict	NaN	18497.0	2367.0
85	10	Venti	NaN	27527.0	4633.0
86	10	Goose	NaN	24848.0	4239.0
87	10	Nugget	NaN	21615.0	4868.0
88	10	NaN	NaN	27050.0	5978.0
89	10	NaN	NaN	25119.0	4666.0
90	10	Cash	NaN	27459.0	4596.0
91	10	Coco	NaN	0.0	14575.0

92	10	Jed	pupper	25720.0	4068.0
93	10	NaN	NaN	22355.0	3918.0
94	10	Sebastian	puppo	34149.0	10309.0
95	10	Walter	NaN	NaN	NaN
96	10	NaN	NaN	23893.0	3910.0
97	10	Sierra	pupper	0.0	1572.0
98	10	Sierra	pupper	7146.0	1572.0
99	10	NaN	doggo	27036.0	5364.0

[100 rows x 16 columns]

Define: For ease of analysis, add the df\_impreds also to the archive

In [99]: *# Code*

```
df_archive_clean = df_archive_clean.merge(df_impreds_clean, on = 'tweet_id', how = 'left')
```

In [100]: df\_archive\_clean.head(1)

```
Out[100]:
```

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	\
0	892420643555336193	NaN	NaN	
				timestamp \
0	2017-08-01 16:23:56	+0000		
				source \
0	<a href="http://twitter.com/download/iphone" r...			
				text retweeted_status_id \
0	This is Phineas. He's a mystical boy. Only eve...			NaN
				retweeted_status_user_id retweeted_status_timestamp \
0		NaN	NaN	
				expanded_urls ... retweet_count \
0	https://twitter.com/dog_rates/status/892420643...			8345.0
	p1	p1_conf	p1_dog	p2 p2_conf p2_dog p3 p3_conf p3_dog
0	orange	0.097049	False	bagel 0.085851 False banana 0.07611 False

[1 rows x 25 columns]

## 0.4.2 Quality

Define: in\_reply\_to\_status\_id and in\_reply\_to\_user\_id columns are reply tweets; remove rows

In [101]: *# Code*

```
df_archive_clean = df_archive_clean.drop(['in_reply_to_status_id', 'in_reply_to_user_id'])
```

In [102]: *# Test to check if the attributes are absent*

```
df_archive_clean.info()
```

```

<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 23 columns):
tweet_id                2356 non-null int64
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                    1611 non-null object
dog_type                380 non-null object
favorite_count          887 non-null float64
retweet_count           887 non-null float64
p1                      2075 non-null object
p1_conf                 2075 non-null float64
p1_dog                  2075 non-null object
p2                      2075 non-null object
p2_conf                 2075 non-null float64
p2_dog                  2075 non-null object
p3                      2075 non-null object
p3_conf                 2075 non-null float64
p3_dog                  2075 non-null object
dtypes: float64(7), int64(3), object(13)
memory usage: 441.8+ KB

```

Define: retweeted\_status\_id, retweeted\_user\_id and retweeted\_time\_stamps columns represent retweets; remove rows

```
In [103]: # Code
```

```
df_archive_clean = df_archive_clean.drop(['retweeted_status_id', 'retweeted_status_u...
```

```
In [104]: # Test to check if the columns are deleted
```

```
df_archive_clean.head(1)
```

```
Out[104]:
```

```

          tweet_id          timestamp \
0  892420643555336193  2017-08-01 16:23:56 +0000

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

                                     text \
0  This is Phineas. He's a mystical boy. Only eve...

                                     expanded_urls  rating_numerator \

```

```

0 https://twitter.com/dog_rates/status/892420643... 13
rating_denominator name dog_type favorite_count retweet_count \
0 10 Phineas NaN 38149.0 8345.0
p1 p1_conf p1_dog p2 p2_conf p2_dog p3 p3_conf p3_dog
0 orange 0.097049 False bagel 0.085851 False banana 0.07611 False

```

Define: Timestamp as string; change to datetime format

```

In [105]: # Code
df_archive_clean['timestamp'] = pd.to_datetime(df_archive_clean['timestamp'])

```

```

In [106]: # Test
df_archive_clean.timestamp.head()

```

```

Out[106]: 0 2017-08-01 16:23:56
1 2017-08-01 00:17:27
2 2017-07-31 00:18:03
3 2017-07-30 15:58:51
4 2017-07-29 16:00:24
Name: timestamp, dtype: datetime64[ns]

```

Define: Source column has 4 unique values which represent the actual sources, lost in the html format; Pick out the source from the html link

```

In [107]: # Code
df_archive_clean.source = df_archive_clean.source.apply(lambda x: x.split('>')[-2])

```

```

In [108]: # Test
df_archive_clean.source.head()

```

```

Out[108]: 0 Twitter for iPhone
1 Twitter for iPhone
2 Twitter for iPhone
3 Twitter for iPhone
4 Twitter for iPhone
Name: source, dtype: object

```

```

In [109]: df_archive_clean.source.unique()

```

```

Out[109]: array(['Twitter for iPhone', 'Twitter Web Client', 'Vine - Make a Scene',
'TweetDeck'], dtype=object)

```

Define: expanded\_urls column doesn't make sense; remove

```

In [110]: # Code
df_archive_clean = df_archive_clean.drop('expanded_urls', axis = 1)

```

```

In [111]: # Test
df_archive_clean.head(1)

```

```

Out[111]:
      tweet_id      timestamp      source \
0  892420643555336193  2017-08-01 16:23:56  Twitter for iPhone

      text  rating_numerator \
0  This is Phineas. He's a mystical boy. Only eve...      13

      rating_denominator  name  dog_type  favorite_count  retweet_count \
0              10  Phineas      NaN      38149.0      8345.0

      p1  p1_conf  p1_dog      p2  p2_conf  p2_dog      p3  p3_conf  p3_dog
0  orange  0.097049  False  bagel  0.085851  False  banana  0.07611  False

```

Define: Rating denominator should be 10 for all; change to 10

```

In [112]: # Code
          #(df_archive_clean.loc[:, 'rating_denominator'] != 10).sum()
          df_archive_clean.loc[df_archive_clean.loc[:, 'rating_denominator'] != 10, 'rating_denominator'] = 10

```

```

In [113]: # Test
          df_archive_clean.rating_denominator.unique()

```

```

Out[113]: array([10])

```

Define: Rating numerator has high and 0 values; set minimum at 11, maximum at 15

```

In [114]: # Code
          df_archive_clean.loc[df_archive_clean.loc[:, 'rating_numerator'] >= 15, 'rating_numerator'] = 15
          df_archive_clean.loc[df_archive_clean.loc[:, 'rating_numerator'] < 10, 'rating_numerator'] = 10

```

```

In [115]: # Test
          df_archive_clean.rating_numerator.unique()

```

```

Out[115]: array([13, 12, 14, 11, 15, 10])

```

Define: Datatype of tweet\_id is int; change to string

```

In [116]: # Code
          df_archive_clean.tweet_id = df_archive_clean.tweet_id.astype('str')

```

```

In [117]: # Test
          df_archive_clean.tweet_id.dtype

```

```

Out[117]: dtype('O')

```

```

In [118]: df_archive_clean.info()

```

```

<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 19 columns):
tweet_id      2356 non-null object

```

```

timestamp          2356 non-null datetime64[ns]
source             2356 non-null object
text              2356 non-null object
rating_numerator   2356 non-null int64
rating_denominator 2356 non-null int64
name              1611 non-null object
dog_type          380 non-null object
favorite_count     887 non-null float64
retweet_count      887 non-null float64
p1                2075 non-null object
p1_conf           2075 non-null float64
p1_dog            2075 non-null object
p2               2075 non-null object
p2_conf           2075 non-null float64
p2_dog            2075 non-null object
p3               2075 non-null object
p3_conf           2075 non-null float64
p3_dog            2075 non-null object
dtypes: datetime64[ns](1), float64(5), int64(2), object(11)
memory usage: 368.1+ KB

```

## 0.5 Data Analyses

```

In [119]: # How many rows and columns finally?
          df_archive_clean.shape

```

```

Out[119]: (2356, 19)

```

```

In [120]: # Data types
          df_archive_clean.info()

```

```

<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 19 columns):
tweet_id          2356 non-null object
timestamp         2356 non-null datetime64[ns]
source            2356 non-null object
text             2356 non-null object
rating_numerator  2356 non-null int64
rating_denominator 2356 non-null int64
name             1611 non-null object
dog_type         380 non-null object
favorite_count    887 non-null float64
retweet_count     887 non-null float64
p1               2075 non-null object
p1_conf          2075 non-null float64
p1_dog           2075 non-null object
p2              2075 non-null object

```

```

p2_conf          2075 non-null float64
p2_dog           2075 non-null object
p3               2075 non-null object
p3_conf          2075 non-null float64
p3_dog           2075 non-null object
dtypes: datetime64[ns](1), float64(5), int64(2), object(11)
memory usage: 368.1+ KB

```

```

In [121]: # Which source was used most for the images?
          df_archive_clean.source.value_counts()

```

```

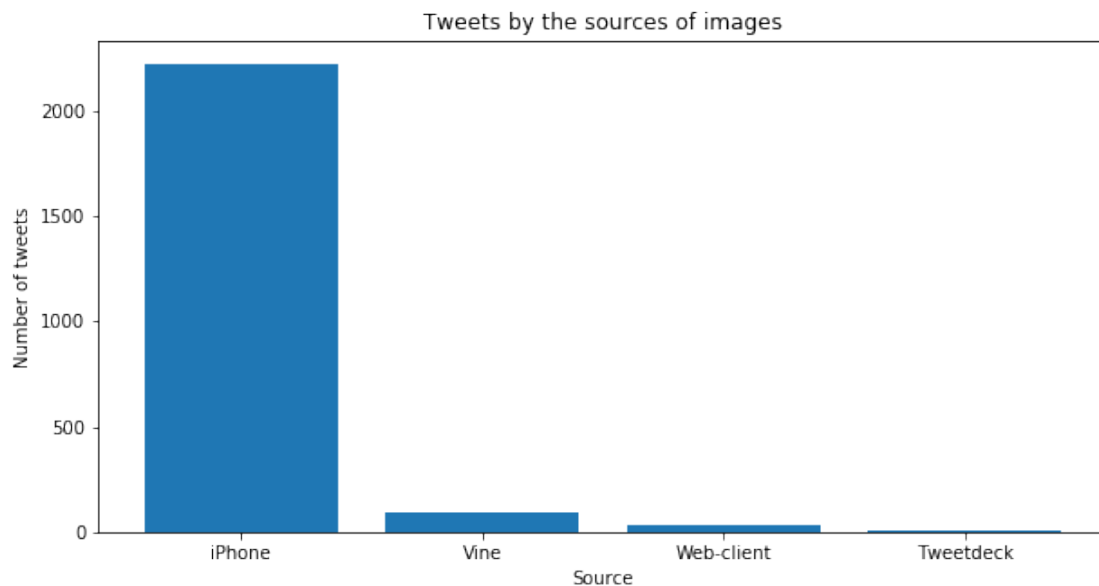
Out[121]: Twitter for iPhone      2221
          Vine - Make a Scene      91
          Twitter Web Client      33
          TweetDeck               11
          Name: source, dtype: int64

```

```

In [122]: x = ['iPhone', 'Vine', 'Web-client', 'Tweetdeck']
          y = df_archive_clean.source.value_counts()
          f, ax = plt.subplots(figsize=(10,5))
          plt.bar(x, y)
          plt.xlabel('Source')
          plt.ylabel('Number of tweets')
          plt.title('Tweets by the sources of images')
          plt.show();

```



```

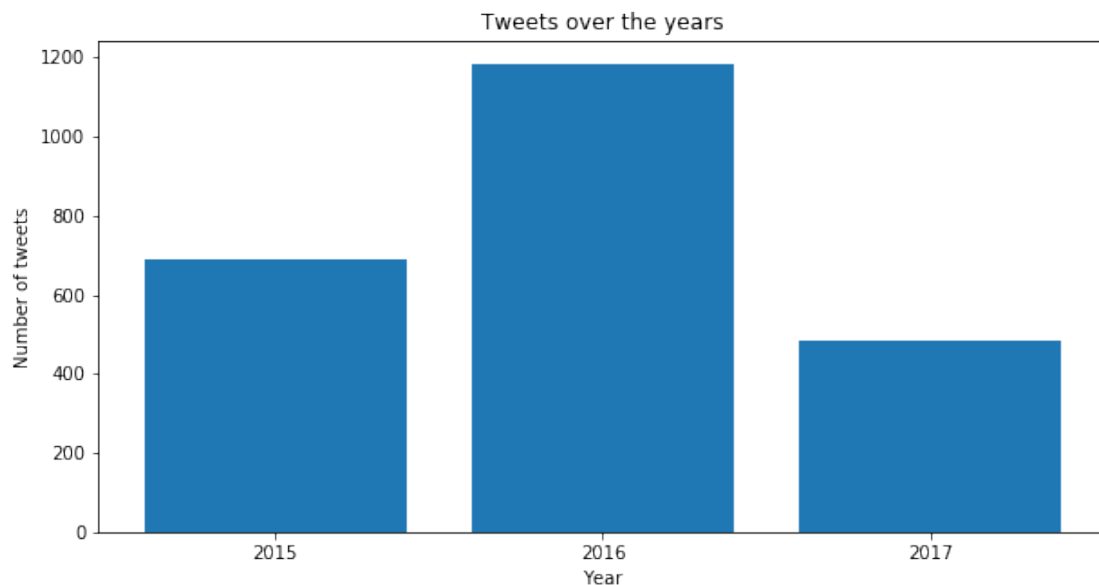
In [123]: # How many tweets were there every year?
          df_archive_clean['tweet_id'].groupby([df_archive_clean['timestamp'].dt.year]).count()

```

```
Out[123]: timestamp
2015      690
2016     1183
2017      483
Name: tweet_id, dtype: int64
```

```
In [124]: x = ['2015', '2016', '2017']
y = df_archive_clean['tweet_id'].groupby([df_archive_clean['timestamp'].dt.year]).count()
```

```
In [125]: f, ax = plt.subplots(figsize=(10,5))
plt.bar(x, y)
plt.xlabel('Year')
plt.ylabel('Number of tweets')
plt.title('Tweets over the years')
plt.show();
```

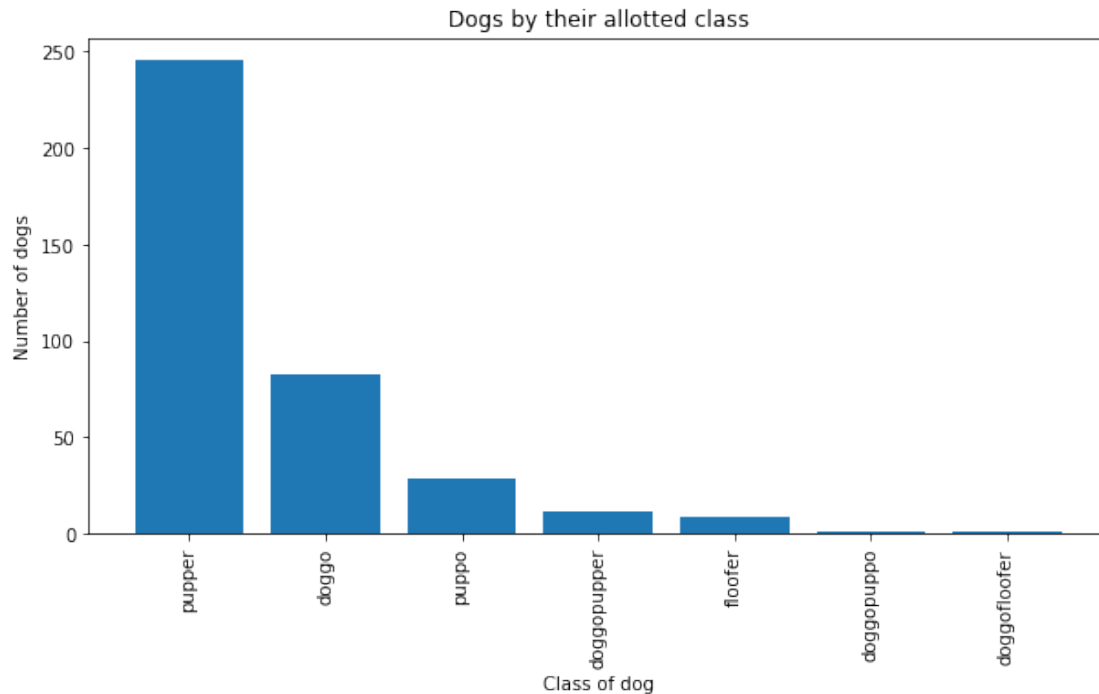


```
In [126]: # Among the dog classes, which was the most frequent?
df_archive_clean.dog_type.value_counts()
```

```
Out[126]: pupper      245
doggo      83
puppo      29
doggopupper  12
floofer     9
doggopuppo   1
doggofloofer  1
Name: dog_type, dtype: int64
```



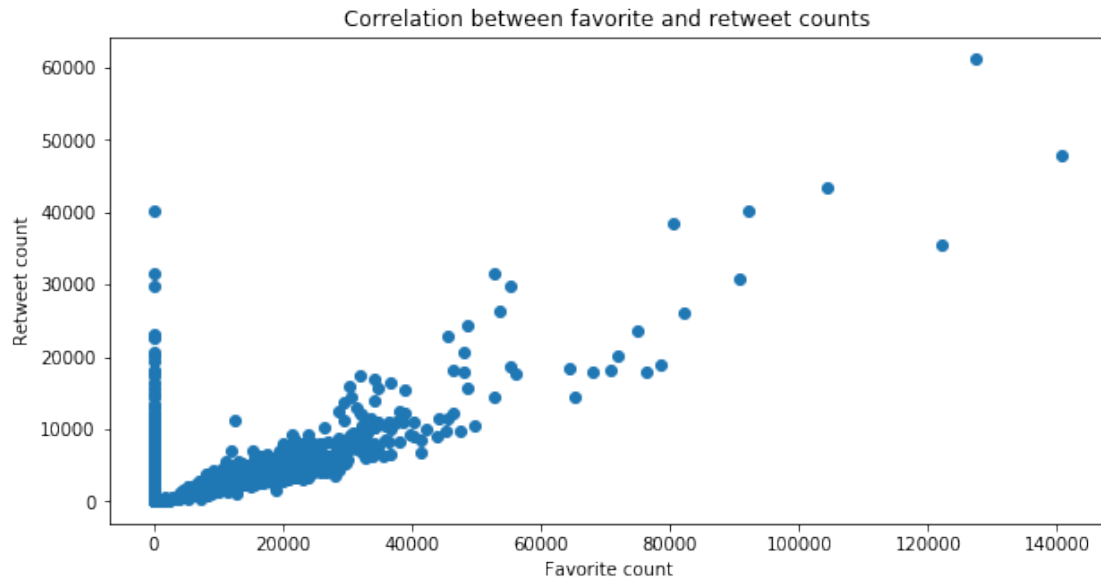
```
In [127]: x = ['pupper', 'doggo', 'puppo', 'doggopupper', 'floofer', 'doggopuppo', 'doggofloof']
y = df_archive_clean.dog_type.value_counts()
f, ax = plt.subplots(figsize=(10,5))
plt.bar(x, y)
plt.xlabel('Class of dog')
plt.ylabel('Number of dogs')
plt.title('Dogs by their allotted class')
plt.xticks(rotation='vertical')
plt.show();
```



```
In [128]: # What were the highest favorite and retweet counts? Which dogs won those?
df_archive_clean['favorite_count'].max(), df_archive_clean['retweet_count'].max()
```

```
Out[128]: (140764.0, 61150.0)
```

```
In [129]: # Correlation between favorite and retweet counts
f, ax = plt.subplots(figsize=(10,5))
plt.scatter(df_archive_clean['favorite_count'], df_archive_clean['retweet_count'])
plt.xlabel('Favorite count')
plt.ylabel('Retweet count')
plt.title('Correlation between favorite and retweet counts')
plt.show();
```



```
In [130]: df_archive_clean[df_archive_clean['favorite_count'] == (df_archive_clean['favorite_count']
```

```
Out[130]:
```

	tweet_id	timestamp	source	text	rating_numerator	rating_denominator	name	dog_type	favorite_count	retweet_count	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
413	822872901745569793	2017-01-21 18:26:02	Twitter for iPhone	Here's a super supportive puppo participating ...	13	10	NaN	puppo	140764.0	47855.0	Lakeland_terrier	0.196015	True	Labrador_retriever	0.160329	True	Irish_terrier	0.069126	True

```
In [131]: df_archive_clean[df_archive_clean['favorite_count'] == (df_archive_clean['favorite_count']
```

```
Out[131]: 413    Here's a super supportive puppo participating ...
           Name: text, dtype: object
```

```
In [132]: df_archive_clean[df_archive_clean['retweet_count'] == (df_archive_clean['retweet_count']
```

```
Out[132]:
```

	tweet_id	timestamp	source	text	rating_numerator
534	807106840509214720	2016-12-09 06:17:20	Twitter for iPhone		

```

534 This is Stephan. He just wants to help. 13/10 ... 13

      rating_denominator      name dog_type favorite_count retweet_count \
534              10    Stephan      NaN      127370.0      61150.0

      p1  p1_conf p1_dog      p2  p2_conf p2_dog      p3  \
534 Chihuahua 0.50537   True Pomeranian 0.120358   True  toy_terrier

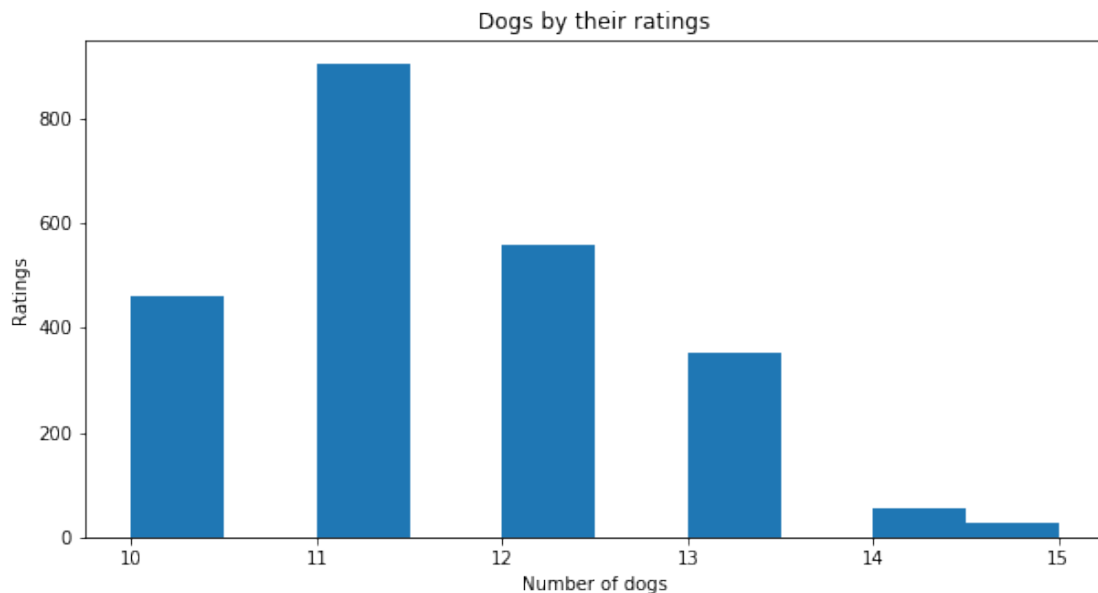
      p3_conf p3_dog
534 0.077008   True

```

```

In [133]: # How did the dog scores vary by the ratings
f, ax = plt.subplots(figsize=(10,5))
plt.hist(df_archive_clean['rating_numerator'])
plt.xlabel('Number of dogs')
plt.ylabel('Ratings')
plt.title('Dogs by their ratings')
plt.show();

```



```

In [134]: p1_conf = df_archive_clean['p1_conf'].mean()
p2_conf = df_archive_clean['p2_conf'].mean()
p3_conf = df_archive_clean['p3_conf'].mean()
p1_err = df_archive_clean['p1_conf'].std()
p2_err = df_archive_clean['p2_conf'].std()
p3_err = df_archive_clean['p3_conf'].std()

```

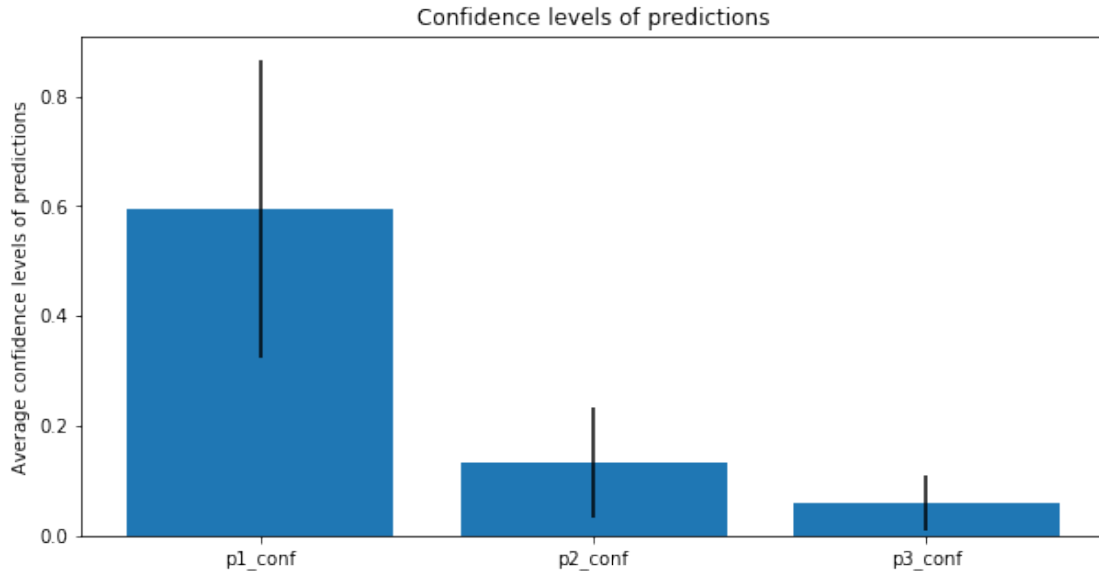
```

In [135]: # How variable were the three top predictions in their confidence levels?
x = ['p1_conf', 'p2_conf', 'p3_conf']

```

```
y = [p1_conf, p2_conf, p3_conf]
z = [p1_err, p2_err, p3_err]
```

```
In [136]: f, ax = plt.subplots(figsize=(10,5))
plt.bar(x, y, yerr=z)
#plt.xlabel('Predictions')
plt.ylabel('Average confidence levels of predictions')
plt.title('Confidence levels of predictions')
plt.show();
```



## 0.6 Conclusions

After merging the three dataframes and fixing the cleanliness issues, it was seen that there were 2356 instances and 19 attributes. These were the tweet\_id, timestamp, source, text, rating\_numerator, rating\_denominator, name, dog\_type, favorite\_count, retweet\_count and the 9 image prediction details columns. Of the 4 sources used to render the images, iPhone, Vine, Web\_client and Tweetdeck, iPhone was found to be the most frequent, with 2221 source listings. There were 690, 1183 and 483 tweets for the years 2015, 2016 and 2017 respectively. The maximum number of tweets during 2016 is expected since the tweets were collected from November 2015 to August 2017. From the information contained in the texts, the dogs were classified into dog stages, ranging from pupper to doggo to floofer, and alternative intermediate or composite stages. The pupper was seen to be the most frequent stage among the classifiable dogs (from the tweet texts) of the categories encountered viz. 'pupper', 'doggo', 'puppo', 'doggopupper', 'floofer', 'doggopuppo' and 'doggofloofer'. The highest favorite count was 140764, for an unnamed puppo, apparently a super supportive one participating in the Women's march in Toronto. The highest retweet count was 61150, for Stephan, a helpful dog predicted to be a Chihuahua with 50% confidence level. A scatter plot analysis showed that there was a strong correlation between the favorite and retweet counts for the dogs. The dogs received ratings between 11 and 15. The

most frequent ratings were in the range of 11-12, numbering around 900. The average confidence levels for dog predictions based on their images were 60% for their first prediction, and lower for the subsequent two breed predictions.