In [836]:

```
import pandas as pd
import numpy as np
import requests
import os
import json
import matplotlib.pyplot as plt
import seaborn as sns
pd.set_option('max_colwidth',100)
```

Data Wrangling

Gather

- Gathering Data from csv file (Source No: 1).

In [837]:

```
twitter_archive_df = pd.read_csv('twitter-archive-enhanced.csv')
twitter_archive_df
```

Out[837]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Phinea mystical boy. (appears in the donut. 13/10 h
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Tilly. S checking pup Hopes you're do not, she's avail
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Archie rare Norwegian F Corgo. Lives i grass. You ne
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Da commenced a sno meal. 13/10 ha the b https://t
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Franklin. H like you to stop ca "cute." He is a ve shark
5	891087950875897856	NaN	NaN	2017-07-29 00:08:17 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have a great white brea South Africa Absolutely h*ckin
6	890971913173991426	NaN	NaN	2017-07-28 16:27:12 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Meet Jax. He e cream so much nervous around help Jax en
7	890729181411237888	NaN	NaN	2017-07-28 00:22:40 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	When you wa owner call anoth good boy but t turn back to you
8	890609185150312448	NaN	NaN	2017-07-27 16:25:51 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Zoey. Sh want to be o scary sharks. Ju to be a snuggly
9	890240255349198849	NaN	NaN	2017-07-26	<a href="http://twitter.com/download/iphone"	This is Cassie. college pup. internation

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	
				15:59:51 +0000	rel="nofollow">Twitter for iPhone</a>	communication
10	890006608113172480	NaN	NaN	2017-07-26 00:31:25 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Koda South A deckshark. De deadly. Frigl majes
11	889880896479866881	NaN	NaN	2017-07-25 16:11:53 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Brunc service shark. C out of the water you. 13/10 te
12	889665388333682689	NaN	NaN	2017-07-25 01:55:32 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here's a pu seems to be on t about something but seriously s
13	889638837579907072	NaN	NaN	2017-07-25 00:10:02 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Ted. He best. Sometimes t enough. But it's c would ass
14	889531135344209921	NaN	NaN	2017-07-24 17:02:04 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Stu sporting his favor pack. Secretly f bones only. 1
15	889278841981685760	NaN	NaN	2017-07-24 00:19:32 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Olive witnessing c many bruta Seems to be pla
16	888917238123831296	NaN	NaN	2017-07-23 00:22:39 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Jim. He fren. Taught him h like the good boy for both h
17	888804989199671297	NaN	NaN	2017-07-22 16:56:37 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Zeke. new stick. Very pr Would like you t for him w
18	888554962724278272	NaN	NaN	2017-07-22 00:23:06 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Ralph powering up. At maximum borkdriv inspirational af
19	888202515573088257	NaN	NaN	2017-07-21 01:02:36 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	RT @dog_rate Canela. She a some fancy pc They were unsu
20	888078434458587136	NaN	NaN	2017-07-20 16:49:33 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Gerald just told he didn job he interview h*ckin inju
21	887705289381826560	NaN	NaN	2017-07-19 16:06:48 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Jeffrey. monopoly on noodles. Currently a 'boop for two'
22	887517139158093824	NaN	NaN	2017-07-19 03:39:09 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	I've yet Venezuela Wiener. This is honor. 14/10 paw- a
23	887473957103951883	NaN	NaN	2017-07-19 00:47:34 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Car attempted sor porch pics. Th unsuccessf someon
24	887343217045368832	NaN	NaN	2017-07-18 16:08:03 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	You may not hav you needed to today. 13/10 plea (IG: emmylouroo)
25	887101392804085760	NaN	NaN	2017-07-18 00:07:08 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This... is a Antarctic House B only rate dogs only send dog
26	886983233522544640	NaN	NaN	2017-07-17 16:17:36 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Maya. Sh shy. Rarely le cup. 13/10 would an environmen
				2017-07-		This is Mingu

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	
27	886736880519319552	NaN	NaN	2017-07-16 23:58:41 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	wonderful fath smol pup. C 13/10, but he ne
28	886680336477933568	NaN	NaN	2017-07-16 20:14:00 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Derek. He' a dog meetir pet...al to t https://t.co/BCoW
29	886366144734445568	NaN	NaN	2017-07-15 23:25:31 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Roscoe. pupper fallen spontaneou ejections Ble
...	...	...	...	...	...	
2326	666411507551481857	NaN	NaN	2015-11-17 00:24:19 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is quite the c really excited wh water. Not very Bad at f
2327	666407126856765440	NaN	NaN	2015-11-17 00:06:54 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is a Vesuvius bum Can drive a truc Made friends wit
2328	666396247373291520	NaN	NaN	2015-11-16 23:23:41 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Oh goodness rare northeas kangaroo mix. feet. N (disappoi
2329	666373753744588802	NaN	NaN	2015-11-16 21:54:18 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Those are sunglas a jean jacket. 1 https://t.co/uHX
2330	666362758909284353	NaN	NaN	2015-11-16 21:10:36 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Unique dog h small. Lives in co Frosted Flakes (legs. Must b
2331	666353288456101888	NaN	NaN	2015-11-16 20:32:58 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have Asiago Galápagos Islan one ear working of r
2332	666345417576210432	NaN	NaN	2015-11-16 20:01:42 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Look at this thinking seat don't apply to hi tongue tho 10/10
2333	666337882303524864	NaN	NaN	2015-11-16 19:31:45 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is an extrer horned Parthe amused. Wear Overall very n
2334	666293911632134144	NaN	NaN	2015-11-16 16:37:02 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is a funny do toes. Won't com Loves branch. Re eat his food. I
2335	666287406224695296	NaN	NaN	2015-11-16 16:11:11 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is an Alban legged Episc Loves well- hardwood floorin
2336	666273097616637952	NaN	NaN	2015-11-16 15:14:19 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Can take selfi https://t.co/ws2AM
2337	666268910803644416	NaN	NaN	2015-11-16 14:57:41 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Very concern fellow dog tr comput https://t.co/0y
2338	666104133288665088	NaN	NaN	2015-11-16 04:02:55 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Not familiar breed. No tai Only 2 legs. Does Surprisingly qu
2339	666102155909144576	NaN	NaN	2015-11-16 03:55:04 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Oh my. Here seeing an Adol giving birth to twi world is an
				2015-11-	<a	Can stand on s what seems like

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	
2340	666099513787052032	NaN	NaN	16 03:44:34 +0000	href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Built that bir Impressive. Mad
2341	666094000022159362	NaN	NaN	2015-11-16 03:22:39 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This appear Mongolian Pres mix. Very tired slip confirmed. 9/
2342	666082916733198337	NaN	NaN	2015-11-16 02:38:37 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we hav est sunblockerspa his other flip-flop. ver
2343	666073100786774016	NaN	NaN	2015-11-16 01:59:36 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Let's hope this f Malaysian (lol) dog! Almost co camouflaged.
2344	666071193221509120	NaN	NaN	2015-11-16 01:52:02 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have a speckled Rhodo Much sass. Gives Good tong
2345	666063827256086533	NaN	NaN	2015-11-16 01:22:45 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is the happ you will ever s committed own couch. 10/10 http
2346	666058600524156928	NaN	NaN	2015-11-16 01:01:59 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here is the Ran retrievers fo probably good Can drink beer (
2347	666057090499244032	NaN	NaN	2015-11-16 00:55:59 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	My oh my. This blond Canadian wheels. On Rather docile. 9/1
2348	666055525042405380	NaN	NaN	2015-11-16 00:49:46 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here is a Siberia armored polar b Strong owne would do unsp
2349	666051853826850816	NaN	NaN	2015-11-16 00:35:11 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is an odd d on the outside b on the inside. Pe fun. Does
2350	666050758794694657	NaN	NaN	2015-11-16 00:30:50 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is a truly English Wils retriever. Ha phone. Privilege
2351	666049248165822465	NaN	NaN	2015-11-16 00:24:50 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have a generation vulpix sweat tea and Fo Cannot be phase
2352	666044226329800704	NaN	NaN	2015-11-16 00:04:52 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is a purebr Morgan. Loves and chill. Always l he forgot
2353	666033412701032449	NaN	NaN	2015-11-15 23:21:54 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here is a very ha Big fan of well-ma decks. Just loo tongue. 9/10
2354	666029285002620928	NaN	NaN	2015-11-15 23:05:30 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is a weste Mitsubishi terrie about leaf. Actuall here. 7/10 w
2355	666020888022790149	NaN	NaN	2015-11-15 22:32:08 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have a J Irish Setter. Lo Vietnam (?). B relaxing on st

2356 rows × 17 columns

- Gathering Data from tsv file by downloading it programmatically (Source No: 2).

In [838]:

```
url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-p
url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-p
redictions.tsv'
response = requests.get(url)
file_name = '/image-predictions.tsv'
if response.status_code == 200:
    #print(response.content)
    with open(os.getcwd() + file_name,mode='wb') as file:
        file.write(response.content)
```

In [839]:

```
image_predictions_df = pd.read_csv('image-predictions.tsv',sep='\t')
image_predictions_df
```

Out[839]:

	tweet_id	jpg_url	img_num	
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	C
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rho
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	m
5	666050758794694657	https://pbs.twimg.com/media/CT5Jof1WUAEuVxN.jpg	1	Berne
6	666051853826850816	https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg	1	
7	666055525042405380	https://pbs.twimg.com/media/CT5N9tpXIAAifs1.jpg	1	
8	666057090499244032	https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg	1	
9	666058600524156928	https://pbs.twimg.com/media/CT5Qw94XAAA_2dP.jpg	1	
10	666063827256086533	https://pbs.twimg.com/media/CT5Vg_wXIAAXfnj.jpg	1	
11	666071193221509120	https://pbs.twimg.com/media/CT5cN_3WEAAlOoZ.jpg	1	
12	666073100786774016	https://pbs.twimg.com/media/CT5d9DZXAAALcwe.jpg	1	
13	666082916733198337	https://pbs.twimg.com/media/CT5m4VGWEAAtKc8.jpg	1	
14	666094000022159362	https://pbs.twimg.com/media/CT5w9gUW4AAsBNN.jpg	1	
15	666099513787052032	https://pbs.twimg.com/media/CT51-JJUEAA6hV8.jpg	1	
16	666102155909144576	https://pbs.twimg.com/media/CT54YGiWUAEZnoK.jpg	1	
17	666104133288665088	https://pbs.twimg.com/media/CT56LSZWoAAlJj2.jpg	1	
18	666268910803644416	https://pbs.twimg.com/media/CT8QCd1WEAADXws.jpg	1	c
19	666273097616637952	https://pbs.twimg.com/media/CT8T1mtUwAA3aqm.jpg	1	
20	666287406224695296	https://pbs.twimg.com/media/CT8g3BpUEAAuFjg.jpg	1	
21	666293911632134144	https://pbs.twimg.com/media/CT8mx7KW4AEQu8N.jpg	1	
22	666337882303524864	https://pbs.twimg.com/media/CT9OwFIWEAMuRje.jpg	1	
23	666345417576210432	https://pbs.twimg.com/media/CT9Vn7PWoAA_ZCM.jpg	1	
24	666353288456101888	https://pbs.twimg.com/media/CT9cx0tUEAAhNN_.jpg	1	
25	666362758909284353	https://pbs.twimg.com/media/CT9lXGsUcAAyUFt.jpg	1	
26	666373753744588802	https://pbs.twimg.com/media/CT9vZEYWUAAlZ05.jpg	1	coate
27	666396247373291520	https://pbs.twimg.com/media/CT-D2ZHWIAA3gK1.jpg	1	
28	666407126856765440	https://pbs.twimg.com/media/CT-NvwmW4AAugGZ.jpg	1	black-an
29	666411507551481857	https://pbs.twimg.com/media/CT-RugiWIAELEaq.jpg	1	
...	...	...	...	
2045	886366144734445568	https://pbs.twimg.com/media/DE0BTnQUwAApKEH.jpg	1	
2046	886680336477933568	https://pbs.twimg.com/media/DE4fEDzWAAAyHMM.jpg	1	
2047	886736880519319552	https://pbs.twimg.com/media/DE5Se8FXcAAJFx4.jpg	1	
2048	886983233522544640	https://pbs.twimg.com/media/DE8yicJW0AAAvBJ.jpg	2	
2049	887101392804085760	https://pbs.twimg.com/media/DE-eAq6UwAA-jaE.jpg	1	
2050	887343217045368832	https://pbs.twimg.com/ext_tw_video_thumb/887343120832229379/pu/img/6HSuFrW1lzI_9Mht.jpg	1	
2051	887473957103951883	https://pbs.twimg.com/media/DEDw2tyUQAAAFke.jpg	2	

	tweet_id	jpg_url	img_num	
2051	887473957103951883	https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg	2	
2052	887517139158093824	https://pbs.twimg.com/ext_tw_video_thumb/887517108413886465/pu/img/WanJKwssZj4VJvL9.jpg	1	
2053	887705289381826560	https://pbs.twimg.com/media/DFHDQBbXgAEqY7t.jpg	1	
2054	888078434458587136	https://pbs.twimg.com/media/DFMWn56WsAAkA7B.jpg	1	
2055	888202515573088257	https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg	2	
2056	888554962724278272	https://pbs.twimg.com/media/DFTH_O-UQAACu20.jpg	3	
2057	888804989199671297	https://pbs.twimg.com/media/DFWra-3VYAA2piG.jpg	1	
2058	888917238123831296	https://pbs.twimg.com/media/DFYRgsOUQAARGhO.jpg	1	
2059	889278841981685760	https://pbs.twimg.com/ext_tw_video_thumb/889278779352338437/pu/img/VlbFB3v8H8VwzVNY.jpg	1	
2060	889531135344209921	https://pbs.twimg.com/media/DFg_2PVW0AEHN3p.jpg	1	
2061	889638837579907072	https://pbs.twimg.com/media/DFihzFfXsAYGDPR.jpg	1	
2062	889665388333682689	https://pbs.twimg.com/media/DFi579UWsAAatzw.jpg	1	
2063	889880896479866881	https://pbs.twimg.com/media/DFl99B1WsAITKsg.jpg	1	
2064	890006608113172480	https://pbs.twimg.com/media/DFnwSY4WAAAMliS.jpg	1	
2065	890240255349198849	https://pbs.twimg.com/media/DFrEyVuW0AAO3t9.jpg	1	
2066	890609185150312448	https://pbs.twimg.com/media/DFwUU__XcAEpyXI.jpg	1	
2067	890729181411237888	https://pbs.twimg.com/media/DFyBahAVwAAhUTd.jpg	2	
2068	890971913173991426	https://pbs.twimg.com/media/DF1eOmZXUAALUcq.jpg	1	
2069	891087950875897856	https://pbs.twimg.com/media/DF3HwyEWsAABqE6.jpg	1	Chesapea
2070	891327558926688256	https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg	2	
2071	891689557279858688	https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg	1	
2072	891815181378084864	https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg	1	
2073	892177421306343426	https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg	1	
2074	892420643555336193	https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg	1	

2075 rows × 12 columns

- Gathering Data from API Call (Source No: 3).

In [840]:

```
import tweepy
from tweepy import OAuthHandler
import json
from timeit import default_timer as timer

# Query Twitter API for each tweet in the Twitter archive and save JSON in a text file
# These are hidden to comply with Twitter's API terms and conditions
consumer_key = 'HIDDEN'
consumer_secret = 'HIDDEN'
access_token = 'HIDDEN'
access_secret = 'HIDDEN'

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

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

# NOTE TO STUDENT WITH MOBILE VERIFICATION ISSUES:
# df_1 is a DataFrame with the twitter_archive_enhanced.csv file. You may have to
# change line 17 to match the name of your DataFrame with twitter_archive_enhanced.csv
# NOTE TO REVIEWER: this student had mobile verification issues so the following
# Twitter API code was sent to this student from a Udacity instructor
# Tweet IDs for which to gather additional data via Twitter's API
tweet_ids = twitter_archive_df.tweet_id.values
len(tweet_ids)

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

In [841]:

```
df_lists = []
with open('tweet_json.txt', 'r') as file:
    while True:
        data =file.readline()
        if data:
            data = json.loads(data)
#             print(data['id'])
#             print(data['retweet_count'])
#             print(data['favorite_count'])
            df_lists.append({
                    'tweet_id': data['id'],
                    'retweet_count': data['retweet_count'],
                    'favorite_count': data['favorite_count']
                })
        else:
            break
```

In [842]:

```
tweet_json_df = pd.DataFrame(df_lists)
tweet_json_df.shape
```

Out[842]:

```
(2354, 3)
```

Assess

- **Visual assessments.**

In [843]:

```
twitter_archive_df.head(20)
```

Out[843]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	re
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t....	
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available fo...	

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	re
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Archie. He is a rare Norwegian Pouncing Corgo. Lives in the tall grass. You never know w...	
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Darla. She commenced a snooze mid meal. 13/10 happens to the best of us https://t.co/tD3...	
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Franklin. He would like you to stop calling him "cute." He is a very fierce shark and sh...	
5	891087950875897856	NaN	NaN	2017-07-29 00:08:17 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have a majestic great white breaching off South Africa's coast. Absolutely h*ckin breath...	
6	890971913173991426	NaN	NaN	2017-07-28 16:27:12 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Meet Jax. He enjoys ice cream so much he gets nervous around it. 13/10 help Jax enjoy more thing...	
7	890729181411237888	NaN	NaN	2017-07-28 00:22:40 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	When you watch your owner call another dog a good boy but then they turn back to you and say you...	
8	890609185150312448	NaN	NaN	2017-07-27 16:25:51 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Zoey. She doesn't want to be one of the scary sharks. Just wants to be a snuggly pettabl...	
9	890240255349198849	NaN	NaN	2017-07-26 15:59:51 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Cassie. She is a college pup. Studying international doggo communication and stick theor...	
10	890006608113172480	NaN	NaN	2017-07-26 00:31:25 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Koda. He is a South Australian deckshark. Deceptively deadly. Frighteningly majestic. 13...	
11	889880896479866881	NaN	NaN	2017-07-25 16:11:53 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Bruno. He is a service shark. Only gets out of the water to assist you. 13/10 terrifying...	
12	889665388333682689	NaN	NaN	2017-07-25 01:55:32 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here's a puppo that seems to be on the fence about something haha no but seriously someone help ...	

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	re
13	889638837579907072	NaN	NaN	2017-07-25 00:10:02 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Ted. He does his best. Sometimes that's not enough. But it's ok. 12/10 would assist http...	
14	889531135344209921	NaN	NaN	2017-07-24 17:02:04 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Stuart. He's sporting his favorite fanny pack. Secretly filled with bones only. 13/10 pu...	
15	889278841981685760	NaN	NaN	2017-07-24 00:19:32 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Oliver. You're witnessing one of his many brutal attacks. Seems to be playing with his v...	
16	888917238123831296	NaN	NaN	2017-07-23 00:22:39 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Jim. He found a fren. Taught him how to sit like the good boys. 12/10 for both https://t...	
17	888804989199671297	NaN	NaN	2017-07-22 16:56:37 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Zeke. He has a new stick. Very proud of it. Would like you to throw it for him without t...	
18	888554962724278272	NaN	NaN	2017-07-22 00:23:06 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Ralphus. He's powering up. Attempting maximum borkdrive. 13/10 inspirational af https://...	
19	888202515573088257	NaN	NaN	2017-07-21 01:02:36 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	RT @dog_rates: This is Canela. She attempted some fancy porch pics. They were unsuccessful. 13/1...	

In [844]:

```
twitter_archive_df.tail(20)
```

Out[844]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	
2336	666273097616637952	NaN	NaN	2015-11-16 15:14:19 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Can take selfi https://t.co/ws2AM
2337	666268910803644416	NaN	NaN	2015-11-16 14:57:41 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Very concern fellow dog tr comput https://t.co/0y
2338	666104133288665088	NaN	NaN	2015-11-16 04:02:55 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Not familiar breed. No tai Only 2 legs. Does Surprisingly qu
2339	666102155909144576	NaN	NaN	2015-11-16 03:55:04 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Oh my. Here seeing an Adol giving birth to twi world is an
2340	666099513787052032	NaN	NaN	2015-11-16 03:44:34	<a href="http://twitter.com/download/iphone"	Can stand on s what seems like Built that bir

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	
				03:44:54	rel="nofollow">Twitter for iPhone</a>	Impressive. Mad
2341	666094000022159362	NaN	NaN	2015-11-16 03:22:39 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This appear Mongolian Pres mix. Very tired slip confirmed. 9/
2342	666082916733198337	NaN	NaN	2015-11-16 02:38:37 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we hav est sunblockerspa his other flip-flop. ver
2343	666073100786774016	NaN	NaN	2015-11-16 01:59:36 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Let's hope this f Malaysian (lol) dog! Almost cc camouflaged.
2344	666071193221509120	NaN	NaN	2015-11-16 01:52:02 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have a speckled Rhodo Much sass. Gives Good tong
2345	666063827256086533	NaN	NaN	2015-11-16 01:22:45 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is the happ you will ever s committed owi couch. 10/10 http
2346	666058600524156928	NaN	NaN	2015-11-16 01:01:59 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here is the Ran retrievers fc probably good Can drink beer (
2347	666057090499244032	NaN	NaN	2015-11-16 00:55:59 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	My oh my. This blond Canadian wheels. Or Rather docile. 9/1
2348	666055525042405380	NaN	NaN	2015-11-16 00:49:46 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here is a Siberia armored polar l Strong owne would do uns
2349	666051853826850816	NaN	NaN	2015-11-16 00:35:11 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is an odd d on the outside b on the inside. Pe fun. Doe
2350	666050758794694657	NaN	NaN	2015-11-16 00:30:50 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is a truly English Wil retriever. H phone. Privilege
2351	666049248165822465	NaN	NaN	2015-11-16 00:24:50 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have a generation vulpi sweat tea and Fc Cannot be phase
2352	666044226329800704	NaN	NaN	2015-11-16 00:04:52 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is a purebı Morgan. Loves and chill. Always l he forgot
2353	666033412701032449	NaN	NaN	2015-11-15 23:21:54 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here is a very ha Big fan of well-ma decks. Just loc tongue. 9/1(
2354	666029285002620928	NaN	NaN	2015-11-15 23:05:30 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is a weste Mitsubishi terri about leaf. Actuall here. 7/10 w
2355	666020888022790149	NaN	NaN	2015-11-15 22:32:08 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have a J Irish Setter. Lc Vietnam (?). E relaxing on st

- **Programmatic assessments.**

In [845]:

```
twitter_archive_df.shape
```

Out[845]:

```
(2356, 17)
```

In [846]:

```
twitter_archive_df.info()
```

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

In [847]:

```
twitter_archive_df.isna().sum()
```

Out[847]:

```
tweet_id                         0
in_reply_to_status_id         2278
in_reply_to_user_id           2278
timestamp                        0
source                           0
text                             0
retweeted_status_id           2175
retweeted_status_user_id      2175
retweeted_status_timestamp    2175
expanded_urls                   59
rating_numerator                 0
rating_denominator               0
name                             0
doggo                            0
floofer                          0
pupper                           0
puppo                            0
dtype: int64
```

In [848]:

```
#Looking for rows where denominator is not equal to 10.
for i ,j in twitter_archive_df[twitter_archive_df.rating_denominator != 10].iterrows():
    print(j['text'],j['rating_numerator'],j['rating_denominator'])
    print("")
```

```
@jonnysun @Lin_Manuel ok jomny I know you're excited but 960/00 isn't a valid rating, 13/10 is tho
960 0

@docmisterio account started on 11/15/15 11 15

The floofs have been released I repeat the floofs have been released. 84/70
https://t.co/NIYC820tmd 84 70

Meet Sam. She smiles 24/7 &amp; secretly aspires to be a reindeer.
Keep Sam smiling by clicking and sharing this link:
https://t.co/98tB8y7y7t https://t.co/LouL5vdvxx 24 7
```

```
RT @dog_rates: After so many requests, this is Bretagne. She was the last surviving 9/11 search do
g, and our second ever 14/10. RIP https:/… 9 11

Why does this never happen at my front door... 165/150 https://t.co/HmwrdfEfUE 165 150

After so many requests, this is Bretagne. She was the last surviving 9/11 search dog, and our seco
nd ever 14/10. RIP https://t.co/XAVDNDaVgQ 9 11

Say hello to this unbelievably well behaved squad of doggos. 204/170 would try to pet all at once 
https://t.co/yGQI3He3xv 204 170

Happy 4/20 from the squad! 13/10 for all https://t.co/eV1diwds8a 4 20

This is Bluebert. He just saw that both #FinalFur match ups are split 50/50. Amazed af. 11/10 http
s://t.co/Kky1DPG4iq 50 50

Happy Saturday here's 9 puppers on a bench. 99/90 good work everybody https://t.co/mpvaVxKmc1 99 9
0

Here's a brigade of puppers. All look very prepared for whatever happens next. 80/80 
https://t.co/0eb7R1Om12 80 80

From left to right:
Cletus, Jerome, Alejandro, Burp, &amp; Titson
None know where camera is. 45/50 would hug all at once https://t.co/sedre1ivTK 45 50

Here is a whole flock of puppers.  60/50 I'll take the lot https://t.co/9dpcw6MdWa 60 50

Happy Wednesday here's a bucket of pups. 44/40 would pet all at once https://t.co/HppvrYuamZ 44 40

Yes I do realize a rating of 4/20 would've been fitting. However, it would be unjust to give these
cooperative pups that low of a rating 4 20

Two sneaky puppers were not initially seen, moving the rating to 143/130. Please forgive us. Thank
you https://t.co/kRK51Y5ac3 143 130

Someone help the girl is being mugged. Several are distracting her while two steal her shoes. Clev
er puppers 121/110 https://t.co/1zfnTJLt55 121 110

This is Darrel. He just robbed a 7/11 and is in a high speed police chase. Was just spotted by the
helicopter 10/10 https://t.co/7EsP8LmSp5 7 11

I'm aware that I could've said 20/16, but here at WeRateDogs we are very professional. An 
inconsistent rating scale is simply irresponsible 20 16

IT'S PUPPERGEDDON. Total of 144/120 ...I think https://t.co/ZanVtAtvIq 144 120

Here we have an entire platoon of puppers. Total score: 88/80 would pet all at once 
https://t.co/y93p6FLvVw 88 80

This is an Albanian 3 1/2 legged  Episcopalian. Loves well-polished hardwood flooring. Penis on th
e collar. 9/10 https://t.co/d9NcXFKwLv 1 2
```

In [849]:

```
twitter_archive_df.duplicated().sum()
```

Out[849]:

```
0
```

- **Visual assessments.**

In [850]:

```
image_predictions_df.head(20)
```

Out[850]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springer_spaniel	0.465074	True	
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	redbone	0.506826	True	miniatur
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_shepherd	0.596461	True	
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_ridgeback	0.408143	True	
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_pinscher	0.560311	True	
5	666050758794694657	https://pbs.twimg.com/media/CT5Jof1WUAEuVxN.jpg	1	Bernese_mountain_dog	0.651137	True	Englis
6	666051853826850816	https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg	1	box_turtle	0.933012	False	
7	666055525042405380	https://pbs.twimg.com/media/CT5N9tpXIAAifs1.jpg	1	chow	0.692517	True	Tibe
8	666057090499244032	https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg	1	shopping_cart	0.962465	False	shopp
9	666058600524156928	https://pbs.twimg.com/media/CT5Qw94XAAA_2dP.jpg	1	miniature_poodle	0.201493	True	
10	666063827256086533	https://pbs.twimg.com/media/CT5Vg_wXIAAXfnj.jpg	1	golden_retriever	0.775930	True	Tibe
11	666071193221509120	https://pbs.twimg.com/media/CT5cN_3WEAAlOoZ.jpg	1	Gordon_setter	0.503672	True	Yorks
12	666073100786774016	https://pbs.twimg.com/media/CT5d9DZXAAALcwe.jpg	1	Walker_hound	0.260857	True	English
13	666082916733198337	https://pbs.twimg.com/media/CT5m4VGWEAAtKc8.jpg	1	pug	0.489814	True	I
14	666094000022159362	https://pbs.twimg.com/media/CT5w9gUW4AAsBNN.jpg	1	bloodhound	0.195217	True	Germar
15	666099513787052032	https://pbs.twimg.com/media/CT51-JJUEAA6hV8.jpg	1	Lhasa	0.582330	True	
16	666102155909144576	https://pbs.twimg.com/media/CT54YGiWUAEZnoK.jpg	1	English_setter	0.298617	True	Nev
17	666104133288665088	https://pbs.twimg.com/media/CT56LSZWoAAlJj2.jpg	1	hen	0.965932	False	
18	666268910803644416	https://pbs.twimg.com/media/CT8QCd1WEAADXws.jpg	1	desktop_computer	0.086502	False	
19	666273097616637952	https://pbs.twimg.com/media/CT8T1mtUwAA3aqm.jpg	1	Italian_greyhound	0.176053	True	

In [851]:

```
image_predictions_df.tail(20)
```

Out[851]:

	tweet_id	jpg_url	img_num	
2055	888202515573088257	https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg	2	
2056	888554962724278272	https://pbs.twimg.com/media/DFTH_O-UQAACu20.jpg	3	
2057	888804989199671297	https://pbs.twimg.com/media/DFWra-3VYAA2piG.jpg	1	
2058	888917238123831296	https://pbs.twimg.com/media/DFYRgsOUQAARGhO.jpg	1	
2059	889278841981685760	https://pbs.twimg.com/ext_tw_video_thumb/889278779352338437/pu/img/VlbFB3v8H8VwzVNY.jpg	1	
2060	889531135344209921	https://pbs.twimg.com/media/DFg_2PVW0AEHN3p.jpg	1	
2061	889638837579907072	https://pbs.twimg.com/media/DFihzFfXsAYGDPR.jpg	1	
2062	889665388333682689	https://pbs.twimg.com/media/DFi579UWsAAatzw.jpg	1	
2063	889880896479866881	https://pbs.twimg.com/media/DFl99B1WsAITKsg.jpg	1	
2064	890006608113172480	https://pbs.twimg.com/media/DFnwSY4WAAAMliS.jpg	1	
2065	890240255349198849	https://pbs.twimg.com/media/DFrEyVuW0AAO3t9.jpg	1	
2066	890609185150312448	https://pbs.twimg.com/media/DFwUU__XcAEpyXI.jpg	1	
2067	890729181411237888	https://pbs.twimg.com/media/DFyBahAVwAAhUTd.jpg	2	
2068	890971913173991426	https://pbs.twimg.com/media/DF1eOmZXUAALUcq.jpg	1	
2069	891087950875897856	https://pbs.twimg.com/media/DF3HwyEWsAABqE6.jpg	1	Chesapea
2070	891327558926688256	https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg	2	
2071	891689557279858688	https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg	1	
2072	891815181378084864	https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg	1	
2073	892177421306343426	https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg	1	
2074	892420643555336193	https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg	1	

- **Programmatic assessments.**

In [852]:

```
image_predictions_df.info()
```

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

In [853]:

```
image_predictions_df.isna().sum()
```

Out[853]:

```
tweet_id    0
jpg_url     0
img_num     0
p1          0
p1_conf     0
p1_dog      0
p2          0
p2_conf     0
p2_dog      0
p3          0
p3_conf     0
p3_dog      0
dtype: int64
```

In [854]:

```
image_predictions_df.duplicated().sum()
```

Out[854]:

```
0
```

- **Visual assessments.**

In [855]:

```
tweet_json_df.head(20)
```

Out[855]:

	favorite_count	retweet_count	tweet_id
0	39467	8853	892420643555336193
1	33819	6514	892177421306343426
2	25461	4328	891815181378084864
3	42908	8964	891689557279858688
4	41048	9774	891327558926688256
5	20562	3261	891087950875897856
6	12041	2158	890971913173991426

	favorite_count	retweet_count	tweet_id
7	56848	16716	890729181411237888
8	28226	4429	890609185150312448
9	32467	7711	890240255349198849
10	31166	7624	890006608113172480
11	28268	5156	889880896479866881
12	38818	8538	889665388333682689
13	27672	4735	889638837579907072
14	15359	2321	889531135344209921
15	25652	5637	889278841981685760
16	29611	4709	888917238123831296
17	26080	4559	888804989199671297
18	20290	3732	888554962724278272
19	22201	3653	888078434458587136

In [856]:

```
tweet_json_df.tail(20)
```

Out[856]:

	favorite_count	retweet_count	tweet_id
2334	184	82	666273097616637952
2335	108	37	666268910803644416
2336	14765	6871	666104133288665088
2337	81	16	666102155909144576
2338	164	73	666099513787052032
2339	169	79	666094000022159362
2340	121	47	666082916733198337
2341	335	174	666073100786774016
2342	154	67	666071193221509120
2343	496	232	666063827256086533
2344	115	61	666058600524156928
2345	304	146	666057090499244032
2346	448	261	666055525042405380
2347	1253	879	666051853826850816
2348	136	60	666050758794694657
2349	111	41	666049248165822465
2350	311	147	666044226329800704
2351	128	47	666033412701032449
2352	132	48	666029285002620928
2353	2535	532	666020888022790149

- **Programmatic assessments.**

In [857]:

```
tweet_json_df.shape
```

Out[857]:

```
(2354, 3)
```

In [858]:

```
tweet_json_df.duplicated().sum()
```

Out[858]:

```
0
```

In [859]:

```
tweet_json_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 3 columns):
favorite_count    2354 non-null int64
retweet_count     2354 non-null int64
tweet_id          2354 non-null int64
dtypes: int64(3)
memory usage: 55.2 KB
```

In [860]:

```
tweet_json_df.isna().sum()
```

Out[860]:

```
favorite_count    0
retweet_count     0
tweet_id          0
dtype: int64
```

- ### Quality Issues

- **twitter_archive_df**

> - Timestamp,retweeted_status_timestamp columns should be of datetime datatype instead of strings.
> - Missing values in columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls).
> - Invalid values (like a, an, the) in name column.
> - Interpretation of None as a non-null value in the columns (name, all four stages).
> - Invalid values (other than 10) in rating_denominator column.
> - Need to extract original ratings only (No retweets).
> - Drop irrelevant columns.

- **image_predictions_df**

> - Non Descriptive column headers (p1,p1_conf ,p1_dog p2 p2_conf p2_dog,p3 p3_conf,p3_dog).
> - 2075 rows instead of 2356, means missing data.
> - Drop irrelevant columns.

- ### Tidiness Issues

> - Columns (doggo floofer pupper puppo) should be merged in a single column indicating dog stage.
> - Merge tweet_json_df and image_predictions_df with twitter_archive_df so that one master df can be created.

Clean

In [861]:

```
#Creating copies of dataframes for cleaning purpose.
twitter_archive_df_clean = twitter_archive_df.copy()
image_predictions_df_clean = image_predictions_df.copy()
tweet_json_df_clean = tweet_json_df.copy()
```

- ### Quality Issues

Define

Change datatype of timestamp column to datetime from strings. No need to change datatype of retweeted_status_timestamp because we are going to drop it.

Code

In [862]:

```
twitter_archive_df_clean['timestamp']  = pd.to_datetime(twitter_archive_df_clean.timestamp)
```

Test

In [863]:

```
twitter_archive_df_clean.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 datetime64[ns]
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: datetime64[ns](1), float64(4), int64(3), object(9)
memory usage: 313.0+ KB
```

Define

Drop columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp after extracting original tweets (means rows having null in retweeted_status_id column).

Delete rows where expanded_urls column is null.

Code

In [864]:

```
#Dropping retweets.
print('Number of retweets = ' ,
twitter_archive_df_clean[twitter_archive_df_clean.retweeted_status_id.notnull()].shape[0])
twitter_archive_df_clean =
twitter_archive_df_clean[twitter_archive_df_clean.retweeted_status_id.isnull()]
```

```
twitter_archive_df_clean[twitter_archive_df_clean.retweeted_status_id.isnull()]
```

```
Number of retweets =  181
```

In [865]:

```
#Dropping reply tweets.
print('Number of reply tweets = ' ,
twitter_archive_df_clean[twitter_archive_df_clean.in_reply_to_status_id.notnull()].shape[0])
twitter_archive_df_clean = twitter_archive_df_clean[twitter_archive_df_clean.in_reply_to_status_id
.isnull()]
```

```
Number of reply tweets =  78
```

In [866]:

```
#Getting indices of rows where expanded_url is null.
print('Number of rows where expanded urls is null = ' ,
twitter_archive_df_clean[twitter_archive_df_clean.expanded_urls.isnull()].shape[0])
indices = list (twitter_archive_df_clean[twitter_archive_df_clean.expanded_urls.isnull()].index)
twitter_archive_df_clean.drop(indices,inplace=True)#Dropping rows at above indices.
```

```
Number of rows where expanded urls is null =  3
```

In [867]:

```
#Now dropping Columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,
retweeted_status_user_id,retweeted_status_timestamp

cols_to_drop =['in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id',
'retweeted_status_user_id', 'retweeted_status_timestamp']

twitter_archive_df_clean.drop(cols_to_drop,axis=1,inplace=True)
```

Test

So now we should have 2356 - 181 - 78 - 3 = 2094 rows and we should have 17 - 5 = 12 columns.

In [868]:

```
twitter_archive_df_clean.info() #So we can clearly see we now have 2094 rows and 12 columns.
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2094 entries, 0 to 2355
Data columns (total 12 columns):
tweet_id              2094 non-null int64
timestamp             2094 non-null datetime64[ns]
source                2094 non-null object
text                  2094 non-null object
expanded_urls         2094 non-null object
rating_numerator      2094 non-null int64
rating_denominator    2094 non-null int64
name                  2094 non-null object
doggo                 2094 non-null object
floofer               2094 non-null object
pupper                2094 non-null object
puppo                 2094 non-null object
dtypes: datetime64[ns](1), int64(3), object(8)
memory usage: 212.7+ KB
```

Define

Get all names starting with lower case and repalce them by None.
Next, replace all None by NaN.

Code

In [869]:

```
#Getting all names starting with lower case and replacing them with None.
def invalid_name(row):
    if row['name'].islower():
        #print (row['name'])
        return 'None'
    else:
        return row['name']

twitter_archive_df_clean.name = twitter_archive_df_clean.apply(invalid_name,axis=1)
```

In [870]:

```
print("Total None Values in name column = ", twitter_archive_df_clean[twitter_archive_df_clean.nam
e == 'None'].shape[0])
```

```
Total None Values in name column =  704
```

In [871]:

```
twitter_archive_df_clean.name = twitter_archive_df_clean.name.replace('None',np.nan)
```

Test

In [872]:

```
twitter_archive_df_clean.isna().sum() #We can clearly see now we have 704 Nulls in name columns.
```

Out[872]:

```
tweet_id                0
timestamp               0
source                  0
text                    0
expanded_urls           0
rating_numerator        0
rating_denominator      0
name                  704
doggo                   0
floofer                 0
pupper                  0
puppo                   0
dtype: int64
```

In [873]:

```
#Also now total none values in name column is 0.
print("Total None Values in name column = ", twitter_archive_df_clean[twitter_archive_df_clean.nam
e == 'None'].shape[0])
```

```
Total None Values in name column =  0
```

Define

We will only deal with stages columns and replace all None with NaN.
Notice: We have already fixed this issue above for column name.

Code

In [874]:

```
print (twitter_archive_df_clean.doggo.value_counts())
print("\n\n")
print (twitter_archive_df_clean.floofer.value_counts())
print("\n\n")
print (twitter_archive_df_clean.pupper.value_counts())
print (twitter_archive_df_clean.pupper.value_counts())
print("\n\n")
twitter_archive_df_clean.puppo.value_counts()
```

```
None      2011
doggo       83
Name: doggo, dtype: int64



None       2084
floofer      10
Name: floofer, dtype: int64



None      1865
pupper     229
Name: pupper, dtype: int64
```

Out[874]:

```
None      2070
puppo       24
Name: puppo, dtype: int64
```

In [875]:

```
twitter_archive_df_clean.doggo = twitter_archive_df_clean.doggo.map({'None':np.NaN, 'doggo':'doggo'
})
twitter_archive_df_clean.floofer = twitter_archive_df_clean.floofer.map({'None':np.NaN, 'floofer':'
floofer'})
twitter_archive_df_clean.pupper = twitter_archive_df_clean.pupper.map({'None':np.NaN, 'pupper':'pup
per'})
twitter_archive_df_clean.puppo = twitter_archive_df_clean.puppo.map({'None':np.NaN, 'puppo':'puppo'
})
```

Test

In [876]:

```
print (twitter_archive_df_clean.doggo.value_counts())
print("\n\n")
print (twitter_archive_df_clean.floofer.value_counts())
print("\n\n")
print (twitter_archive_df_clean.pupper.value_counts())
print("\n\n")
twitter_archive_df_clean.puppo.value_counts()
#We can see, Now there are only actual values and not None.
```

```
doggo    83
Name: doggo, dtype: int64



floofer    10
Name: floofer, dtype: int64



pupper    229
Name: pupper, dtype: int64
```

Out[876]:

```
puppo    24
Name: puppo, dtype: int64
```

In [877]:

```
#Also Visualizing the same.
twitter_archive_df_clean.head(20)
```

Out[877]:

	tweet_id	timestamp	source	text	
0	892420643555336193	2017-08-01 16:23:56	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t....	https://twitter.com
1	892177421306343426	2017-08-01 00:17:27	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available fo...	https://twitter.com
2	891815181378084864	2017-07-31 00:18:03	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Archie. He is a rare Norwegian Pouncing Corgo. Lives in the tall grass. You never know w...	https://twitter.com
3	891689557279858688	2017-07-30 15:58:51	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Darla. She commenced a snooze mid meal. 13/10 happens to the best of us https://t.co/tD3...	https://twitter.com
4	891327558926688256	2017-07-29 16:00:24	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Franklin. He would like you to stop calling him "cute." He is a very fierce shark and sh...	https://twitter.com/dog_rates/status/891327558926
5	891087950875897856	2017-07-29 00:08:17	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have a majestic great white breaching off South Africa's coast. Absolutely h*ckin breath...	https://twitter.com
6	890971913173991426	2017-07-28 16:27:12	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Meet Jax. He enjoys ice cream so much he gets nervous around it. 13/10 help Jax enjoy more thing...	jax,https://t
7	890729181411237888	2017-07-28 00:22:40	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	When you watch your owner call another dog a good boy but then they turn back to you and say you...	https://twitter.com/dog_rates/status/890729181411
8	890609185150312448	2017-07-27 16:25:51	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Zoey. She doesn't want to be one of the scary sharks. Just wants to be a snuggly pettabl...	https://twitter.com
9	890240255349198849	2017-07-26 15:59:51	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Cassie. She is a college pup. Studying international doggo	https://twitter.com

	tweet_id	timestamp	source	text	
				communication and stick theor...	
10	890006608113172480	2017-07-26 00:31:25	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Koda. He is a South Australian deckshark. Deceptively deadly. Frighteningly majestic. 13...	https://twitter.com/dog_rates/status/890006608113
11	889880896479866881	2017-07-25 16:11:53	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Bruno. He is a service shark. Only gets out of the water to assist you. 13/10 terrifying...	https://twitter.com
12	889665388333682689	2017-07-25 01:55:32	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here's a puppo that seems to be on the fence about something haha no but seriously someone help ...	https://twitter.com
13	889638837579907072	2017-07-25 00:10:02	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Ted. He does his best. Sometimes that's not enough. But it's ok. 12/10 would assist http...	https://twitter.com/dog_rates/status/889638837579
14	889531135344209921	2017-07-24 17:02:04	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Stuart. He's sporting his favorite fanny pack. Secretly filled with bones only. 13/10 pu...	https://twitter.com
15	889278841981685760	2017-07-24 00:19:32	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Oliver. You're witnessing one of his many brutal attacks. Seems to be playing with his v...	https://twitter.com
16	888917238123831296	2017-07-23 00:22:39	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Jim. He found a fren. Taught him how to sit like the good boys. 12/10 for both https://t...	https://twitter.com
17	888804989199671297	2017-07-22 16:56:37	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Zeke. He has a new stick. Very proud of it. Would like you to throw it for him without t...	https://twitter.com/dog_rates/status/888804989199
18	888554962724278272	2017-07-22 00:23:06	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Ralphus. He's powering up. Attempting maximum borkdrive. 13/10 inspirational af https://...	https://twitter.com/dog_rates/status/888554962724
20	888078434458587136	2017-07-20 16:49:33	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Gerald. He was just told he didn't get the job he interviewed for. A h*ckin injustice. 1...	https://twitter.com/dog_rates/status/888078434458

Define

Replace rating_denominator with correct rating by extracting it from text column.

Code

In [878]:

```
twitter_archive_df_clean.shape #Notice there are 2094 rows.
```

Out[878]:

```
(2094, 12)
```

In [879]:

```
print ('No of rows where denominator is not equal to 10 = ',
twitter_archive_df_clean[twitter_archive_df_clean.rating_denominator != 10].shape[0])
df_10 =  twitter_archive_df_clean[twitter_archive_df_clean.rating_denominator != 10]
df_10_indicies = df_10.index
```

```
No of rows where denominator is not equal to 10 =  17
```

In [880]:

```
df_10
```

Out[880]:

	tweet_id	timestamp	source	text	
433	820690176645140481	2017-01-15 17:52:40	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	The floofs have been released I repeat the floofs have been released. 84/70 https://t.co/NIYC820tmd	https://twitter.com/dog_rates/status/820
516	810984652412424192	2016-12-19 23:06:23	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Meet Sam. She smiles 24/7 &amp; secretly aspires to be a reindeer. \nKeep Sam smiling by clickin...	smile,https:/
902	758467244762497024	2016-07-28 01:00:57	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Why does this never happen at my front door... 165/150 https://t.co/HmwrdfEfUE	https:
1068	740373189193256964	2016-06-08 02:41:38	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	After so many requests, this is Bretagne. She was the last surviving 9/11 search dog, and our se...	https://twitter.com/dog_rates/status/740
1120	731156023742988288	2016-05-13 16:15:54	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Say hello to this unbelievably well behaved squad of doggos. 204/170 would try to pet all at onc...	https:/
1165	722974582966214656	2016-04-21 02:25:47	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Happy 4/20 from the squad! 13/10 for all https://t.co/eV1diwds8a	https:/
1202	716439118184652801	2016-04-03 01:36:11	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Bluebert. He just saw that both #FinalFur match ups are split 50/50. Amazed af. 11/10 ht...	https:/
1228	713900603437621249	2016-03-27 01:29:02	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Happy Saturday here's 9 puppers on a bench. 99/90 good work everybody https://t.co/mpvaVxKmc1	https:/
1254	710658690886586372	2016-03-18 02:46:49	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here's a brigade of puppers. All look very prepared for whatever happens next. 80/80 https://t.c...	https:/
1274	709198395643068416	2016-03-14 02:04:08	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	From left to right:\nCletus, Jerome, Alejandro, Burp, &amp; Titson\nNone know where camera is. 4...	https:/
				Here is a whole flock of	

	tweet_id	timestamp	source	text	
1351	704054845121142784	2016-02-28 21:25:30	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here is a whole flock of puppers. 60/50 I'll take the lot https://t.co/9dpcw6MdWa	https:/
1433	697463031882764288	2016-02-10 16:51:59	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Happy Wednesday here's a bucket of pups. 44/40 would pet all at once https://t.co/HppvrYuamZ	https:/
1635	684222868335505415	2016-01-05 04:00:18	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Someone help the girl is being mugged. Several are distracting her while two steal her shoes. Cl...	https:/
1662	682962037429899265	2016-01-01 16:30:13	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Darrel. He just robbed a 7/11 and is in a high speed police chase. Was just spotted by t...	https:/
1779	677716515794329600	2015-12-18 05:06:23	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	IT'S PUPPERGEDDON. Total of 144/120 ...I think https://t.co/ZanVtAtvIq	https:/
1843	675853064436391936	2015-12-13 01:41:41	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have an entire platoon of puppers. Total score: 88/80 would pet all at once https://t.co...	https://twitter.com/dog_rates/status/675
2335	666287406224695296	2015-11-16 16:11:11	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is an Albanian 3 1/2 legged Episcopalian. Loves well-polished hardwood flooring. Penis on ...	https:/

In [881]:

```
from fractions import *
df_10_dict = dict(df_10.text.str.findall('(\d+[/]\d+)')) #Applying regex to get ratings from text column.
```

In [882]:

```
#Making separate dictionaries for numerators and denominators.
numerators_dict ={}
denominators_dict = {}

for key,value in df_10_dict.items():
    #print(key," ",value)
    for i in value:
        i=i.split("/")
        fraction = (Fraction(int(i[0]),int(i[1])))
        num = str(fraction).split("/")[0]
        if num == '6' or num == '7':
            numerators_dict[key]= int(num) * 2 #Multiplying numerator 6 or 7 by 2 so that denominator can be multiplied as well,so that it will become 10.
        else:
            numerators_dict[key]=int(num)

        if len(str(fraction).split("/")) == 2:
            den = str(fraction).split("/")[1]
            if den == '5':
                denominators_dict[key]= 10
            else:
                denominators_dict[key]=int(den)

        else:
            denominators_dict[key]=1

    #print("\n\n")
```

In [883]:

```
df_10.rating_numerator = numerators_dict.values()
df_10.rating_denominator = denominators_dict.values()
```

```
/opt/conda/lib/python3.6/site-packages/pandas/core/generic.py:4405: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/indexing.html#indexing-view-versus-copy
  self[name] = value
```

In [884]:

```
df_10
```

Out[884]:

	tweet_id	timestamp	source	text	
433	820690176645140481	2017-01-15 17:52:40	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	The floofs have been released I repeat the floofs have been released. 84/70 https://t.co/NIYC820tmd	https://twitter.com/dog_rates/status/820
516	810984652412424192	2016-12-19 23:06:23	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Meet Sam. She smiles 24/7 &amp; secretly aspires to be a reindeer. \nKeep Sam smiling by clickin...	smile,https:/
902	758467244762497024	2016-07-28 01:00:57	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Why does this never happen at my front door... 165/150 https://t.co/HmwrdfEfUE	https:
1068	740373189193256964	2016-06-08 02:41:38	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	After so many requests, this is Bretagne. She was the last surviving 9/11 search dog, and our se...	https://twitter.com/dog_rates/status/740
1120	731156023742988288	2016-05-13 16:15:54	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Say hello to this unbelievably well behaved squad of doggos. 204/170 would try to pet all at onc...	https:/
1165	722974582966214656	2016-04-21 02:25:47	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Happy 4/20 from the squad! 13/10 for all https://t.co/eV1diwds8a	https:/
1202	716439118184652801	2016-04-03 01:36:11	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Bluebert. He just saw that both #FinalFur match ups are split 50/50. Amazed af. 11/10 ht...	https:/
1228	713900603437621249	2016-03-27 01:29:02	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Happy Saturday here's 9 puppers on a bench. 99/90 good work everybody https://t.co/mpvaVxKmc1	https:/
1254	710658690886586372	2016-03-18 02:46:49	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here's a brigade of puppers. All look very prepared for whatever happens next. 80/80 https://t.c...	https:/
1274	709198395643068416	2016-03-14 02:04:08	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	From left to right:\nCletus, Jerome, Alejandro, Burp, &amp; Titson\nNone know where camera is. 4...	https:/
1351	704054845121142784	2016-02-28 21:25:30	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here is a whole flock of puppers. 60/50 I'll take the lot https://t.co/9dpcw6MdWa	https:/
1433	697463031882764288	2016-02-10 16:51:59	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Happy Wednesday here's a bucket of pups. 44/40 would pet all at once https://t.co/HppvrYuamZ	https:/
1635	684222868335505415	2016-01-05 04:00:18	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Someone help the girl is being mugged. Several are distracting her while two steal her shoes. Cl...	https:/
1662	682962037429899265	2016-01-01 16:30:13	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Darrel. He just robbed a 7/11 and is in a high speed police chase. Was just spotted by t...	https:/

	tweet_id	timestamp	source	text	
1779	677716515794329600	2015-12-18 05:06:23	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	IT'S PUPPERGEDDON. Total of 144/120 ...I think https://t.co/ZanVtAtvIq	https:/
1843	675853064436391936	2015-12-13 01:41:41	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have an entire platoon of puppers. Total score: 88/80 would pet all at once https://t.co...	https://twitter.com/dog_rates/status/675
2335	666287406224695296	2015-11-16 16:11:11	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is an Albanian 3 1/2 legged Episcopalian. Loves well-polished hardwood flooring. Penis on ...	https:/

In [885]:

```
twitter_archive_df_clean.drop(df_10_indicies,inplace=True) #Now Deleting orginal rows where denominator was not equal to 10.
```

In [886]:

```
#Making sure all rows with denominator not equal to 10 are gone.
print ('No of rows where denominator is not equal to 10 = ',
twitter_archive_df_clean[twitter_archive_df_clean.rating_denominator != 10].shape[0])
print(twitter_archive_df_clean.shape)#Also notice now there are 2077 rows, because 2094 - 17 = 2077 rows.
```

```
No of rows where denominator is not equal to 10 =  0
(2077, 12)
```

In [887]:

```
twitter_archive_df_clean = twitter_archive_df_clean.append(df_10)#Now appending dropped rows with correct numerator and denominator.
twitter_archive_df_clean.shape #Notice now rows are 2094 means deleted rows are now successfully appended.
```

Out[887]:

```
(2094, 12)
```

In [888]:

```
print ('No of rows where denominator is not equal to 10 = ',
twitter_archive_df_clean[twitter_archive_df_clean.rating_denominator != 10].shape[0])
#Now we are only left with 3 rows where denominator is not equal to 10. Let's visualize them.
twitter_archive_df_clean[twitter_archive_df_clean.rating_denominator != 10]
```

```
No of rows where denominator is not equal to 10 =  3
```

Out[888]:

	tweet_id	timestamp	source	text	
516	810984652412424192	2016-12-19 23:06:23	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Meet Sam. She smiles 24/7 &amp; secretly aspires to be a reindeer. \nKeep Sam smiling by clickin...	https://www.got smile,https://twitter.com/dog_rates/status/8109846524
1254	710658690886586372	2016-03-18 02:46:49	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here's a brigade of puppers. All look very prepared for whatever happens	https://twitter.com/dog_rates/status/7106586908

	tweet_id	timestamp	source	text	
				happens next. 80/80 https://t.c...	
1662	682962037429899265	2016-01-01 16:30:13	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Darrel. He just robbed a 7/11 and is in a high speed police chase. Was just spotted by t...	https://twitter.com/dog_rates/status/682962037

In [889]:

```
#These 3 rows have invalid denominator rating. I think it will be a better idea to delete these ro
ws.
index_delete = twitter_archive_df_clean[twitter_archive_df_clean.rating_denominator != 10].index
twitter_archive_df_clean.drop(index_delete,0,inplace=True)
```

Test

In [890]:

```
print ('No of rows where denominator is not equal to 10 = ',
twitter_archive_df_clean[twitter_archive_df_clean.rating_denominator != 10].shape[0])
#Now we can see, there are no more rows with denominator is not equal to 10.
```

```
No of rows where denominator is not equal to 10 =  0
```

Define

Drop all retweet related columns.

Code

In [891]:

```
#Note, Columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,
retweeted_status_user_id,
#retweeted_status_timestamp are already dropped after extracting original tweeets ,above while cle
aning 2nd quality issue.
#So this means, technically we have already cleaned this issue.
```

Test

In [892]:

```
twitter_archive_df_clean.info() #Making sure there are no columns left of retweets.
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2091 entries, 0 to 2335
Data columns (total 12 columns):
tweet_id                 2091 non-null int64
timestamp                2091 non-null datetime64[ns]
source                   2091 non-null object
text                     2091 non-null object
expanded_urls            2091 non-null object
rating_numerator         2091 non-null int64
rating_denominator       2091 non-null int64
name                     1388 non-null object
doggo                    83 non-null object
floofer                  10 non-null object
pupper                   229 non-null object
puppo                    24 non-null object
dtypes: datetime64[ns](1), int64(3), object(8)
memory usage: 212.4+ KB
```

Define

Change Columns headers p1 to prediction_1, p1_conf to prediction1_confidence, p1_dog to prediction1_dog_breed and so on.

Code

In [893]:

```
new_columns = {'p1': 'prediction1', 'p1_conf': 'prediction1_confidence', 'p1_dog': 'prediction1_dog_breed',
               'p2': 'prediction2', 'p2_conf': 'prediction2_confidence', 'p2_dog': 'prediction2_dog_breed',
               'p3': 'prediction3', 'p3_conf': 'prediction3_confidence', 'p3_dog': 'prediction3_dog_breed'
              }
image_predictions_df_clean.rename(new_columns,axis=1,inplace=True)
```

Test

In [894]:

```
image_predictions_df_clean.info() #we can see columns name now have been changed.
```

```
<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
prediction1                2075 non-null object
prediction1_confidence     2075 non-null float64
prediction1_dog_breed      2075 non-null bool
prediction2                2075 non-null object
prediction2_confidence     2075 non-null float64
prediction2_dog_breed      2075 non-null bool
prediction3                2075 non-null object
prediction3_confidence     2075 non-null float64
prediction3_dog_breed      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

Define

Drop Columns that are not useful for analyzing the data from image_predictions_df.

Code

In [895]:

```
image_predictions_df_clean.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
prediction1                2075 non-null object
prediction1_confidence     2075 non-null float64
prediction1_dog_breed      2075 non-null bool
prediction2                2075 non-null object
prediction2_confidence     2075 non-null float64
prediction2_dog_breed      2075 non-null bool
prediction3                2075 non-null object
prediction3_confidence     2075 non-null float64
prediction3_dog_breed      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
```

```
memory usage: 152.1+ KB
```

In [896]:

```
image_predictions_df_clean.drop(['jpg_url','img_num'],1,inplace=True) #I think there is no need of
jpg_url and img_num to analyze the data.
```

Test

In [897]:

```
image_predictions_df_clean.info() #Columns dropped successfully.
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 10 columns):
tweet_id                  2075 non-null int64
prediction1               2075 non-null object
prediction1_confidence    2075 non-null float64
prediction1_dog_breed     2075 non-null bool
prediction2               2075 non-null object
prediction2_confidence    2075 non-null float64
prediction2_dog_breed     2075 non-null bool
prediction3               2075 non-null object
prediction3_confidence    2075 non-null float64
prediction3_dog_breed     2075 non-null bool
dtypes: bool(3), float64(3), int64(1), object(3)
memory usage: 119.6+ KB
```

- ### Tidiness Issues

Define

Columns (doggo floofer pupper puppo) should be merged in a single column indicating dog stage.
Tweets with multiple dog stages are to placed as multiple in their respective cell.

Code

In [898]:

```
twitter_archive_df_clean.loc[(twitter_archive_df_clean[['doggo', 'floofer', 'pupper', 'puppo']].no
tna()
                  ).sum(axis=1) > 1]
```

Out[898]:

	tweet_id	timestamp	source	text	
191	855851453814013952	2017-04-22 18:31:02	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here's a puppo participating in the #ScienceMarch. Cleverly disguising her own doggo agenda. 13/...	https:/
200	854010172552949760	2017-04-17 16:34:26	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	At first I thought this was a shy doggo, but it's actually a Rare Canadian Floofer Owl. Amateurs...	https://twitter.com/dog_rates/status/854
460	817777686764523521	2017-01-07 16:59:28	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Dido. She's playing the lead role in "Pupper Stops to Catch Snow Before Resuming Shadow ...	https:
531	808106460588765185	2016-12-12 00:29:28	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here we have Burke (pupper) and Dexter (doggo). Pupper wants to be exactly like doggo. Both 12/1...	https:/
		2016-11-	<a	This is Bones. He's being haunted by another	

	tweet_id	timestamp	source	text	
575	801115127852503040	22 17:28:25	href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	doggo of roughly the same size. 12/10 deep breaths ...	https://twitter.com/dog_rates/status/801
705	785639753186217984	2016-10-11 00:34:48	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is Pinot. He's a sophisticated doggo. You can tell by the hat. Also pointier than your aver...	https://twitter.com/dog_rates/status/785
733	781308096455073793	2016-09-29 01:42:20	<a href="http://vine.co" rel="nofollow">Vine - Make a Scene</a>	Pupper butt 1, Doggo 0. Both 12/10 https://t.co/WQvcPEpH2u	
889	759793422261743616	2016-07-31 16:50:42	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Meet Maggie &amp; Lila. Maggie is the doggo, Lila is the pupper. They are sisters. Both 12/10 wo...	https://twitter.com/dog_rates/status/759
956	751583847268179968	2016-07-09 01:08:47	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Please stop sending it pictures that don't even have a doggo or pupper in them. Churlish af. 5/1...	https:/
1063	741067306818797568	2016-06-10 00:39:48	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	This is just downright precious af. 12/10 for both pupper and doggo https://t.co/o5J479bZUC	https:/
1113	733109485275860992	2016-05-19 01:38:16	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Like father (doggo), like son (pupper). Both 12/10 https://t.co/pG2inLaOda	https:/

In [900]:

```
def dogStage(row):
    return(','.join(row.dropna().astype(str)))


twitter_archive_df_clean['dog_stage'] = twitter_archive_df_clean.iloc[: , -4:].apply(dogStage,axis=
1)
```

In [901]:

```
twitter_archive_df_clean.dog_stage.value_counts() #so there are 1756 empty cells that need to be r
eplace by np.nan
```

Out[901]:

```
                 1756
pupper            220
doggo              72
puppo              23
floofer             9
doggo,pupper        9
doggo,floofer       1
doggo,puppo         1
Name: dog_stage, dtype: int64
```

In [902]:

```
twitter_archive_df_clean.dog_stage = 
twitter_archive_df_clean.dog_stage.replace('',np.nan)#replacing empty cells with np.nan
```

In [903]:

```
twitter_archive_df_clean.dog_stage.value_counts()
#Now we don't have any empty cell. Let's merge multiple dog stages to value multiple. Note, at the 
end we should have
# 11 cells with value multiple.
```

Out[903]:

```
pupper           220
doggo             72
puppo             23
floofer            9
doggo,pupper       9
doggo,floofer      1
doggo,puppo        1
```

```
doggo,puppo         1
Name: dog_stage, dtype: int64
```

In [904]:

```
#Now merging multiple dog stages to value multiple.
def multiple(stage):
    if ',' in str(stage):
        return 'Multiple'
    else:
        return stage


twitter_archive_df_clean.dog_stage = twitter_archive_df_clean.dog_stage.apply(multiple)
```

In [905]:

```
twitter_archive_df_clean.drop(['doggo','floofer','pupper','puppo'],1,inplace=True) #dropping all t
hese columns.
```

Test

In [906]:

```
twitter_archive_df_clean.info() #As we can see, dog_stage has been added successfully and the colu
mns
#('doggo','floofer','pupper','puppo') have been deleted successfully.
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2091 entries, 0 to 2335
Data columns (total 9 columns):
tweet_id              2091 non-null int64
timestamp             2091 non-null datetime64[ns]
source                2091 non-null object
text                  2091 non-null object
expanded_urls         2091 non-null object
rating_numerator      2091 non-null int64
rating_denominator    2091 non-null int64
name                  1388 non-null object
dog_stage             335 non-null object
dtypes: datetime64[ns](1), int64(3), object(5)
memory usage: 163.4+ KB
```

In [908]:

```
twitter_archive_df_clean.dog_stage.value_counts() #As expected, Now we have 11 cells with value Mu
ltiple that means our cleaning
#was successful.
```

Out[908]:

```
pupper      220
doggo        72
puppo        23
Multiple     11
floofer       9
Name: dog_stage, dtype: int64
```

In [910]:

```
twitter_archive_df_clean.loc[[191]] #we can also visualize, dog_stage column contains Multiple.
```

Out[910]:

	tweet_id	timestamp	source	text	
191	855851453814013952	2017-04-22 18:31:02	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Here's a puppo participating in the #ScienceMarch. Cleverly	https://twitter.com/dog_rates/status/85585145381

tweet_id	timestamp	source	text
	18:31:02	rel="nofollow">Twitter for iPhone</a>	disguising her own doggo agenda. 13/...

Define

Merge tweet_json_df and image_predictions_df with twitter_archive_df so that one master df can be created.

Code

In [67]:

```
twitter_archive_df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2091 entries, 0 to 2335
Data columns (total 9 columns):
tweet_id              2091 non-null int64
timestamp             2091 non-null datetime64[ns]
source                2091 non-null object
text                  2091 non-null object
expanded_urls         2091 non-null object
rating_numerator      2091 non-null int64
rating_denominator    2091 non-null int64
name                  1388 non-null object
dog_stage             335 non-null object
dtypes: datetime64[ns](1), int64(3), object(5)
memory usage: 163.4+ KB
```

In [68]:

```
tweet_json_df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 3 columns):
favorite_count    2354 non-null int64
retweet_count     2354 non-null int64
tweet_id          2354 non-null int64
dtypes: int64(3)
memory usage: 55.2 KB
```

In [911]:

```
image_predictions_df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 10 columns):
tweet_id                  2075 non-null int64
prediction1               2075 non-null object
prediction1_confidence    2075 non-null float64
prediction1_dog_breed     2075 non-null bool
prediction2               2075 non-null object
prediction2_confidence    2075 non-null float64
prediction2_dog_breed     2075 non-null bool
prediction3               2075 non-null object
prediction3_confidence    2075 non-null float64
prediction3_dog_breed     2075 non-null bool
dtypes: bool(3), float64(3), int64(1), object(3)
memory usage: 119.6+ KB
```

In [933]:

```
twitter_archive_master = pd.merge(twitter_archive_df_clean,tweet_json_df_clean,on='tweet_id')
twitter_archive_master = pd.merge(twitter_archive_master,image_predictions_df_clean,on='tweet_id',
how='right')
```

In [934]:

```
#Now we have merged dataframe, Let's drop columns that are not useful for analysis.
#Note we have also dropped rating_denominator as this column contains value only 10 and that is implicit.
cols = ['source','text','expanded_urls','rating_denominator']
twitter_archive_master.drop(cols,axis=1,inplace=True)
```

In [935]:

```
twitter_archive_master.dropna(inplace=True)
```

Test

In [936]:

```
twitter_archive_master.info() #Now we can see we have successfully merged and dropped columns.
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 177 entries, 9 to 1625
Data columns (total 16 columns):
tweet_id                  177 non-null int64
timestamp                 177 non-null datetime64[ns]
rating_numerator          177 non-null float64
name                      177 non-null object
dog_stage                 177 non-null object
favorite_count            177 non-null float64
retweet_count             177 non-null float64
prediction1               177 non-null object
prediction1_confidence    177 non-null float64
prediction1_dog_breed     177 non-null bool
prediction2               177 non-null object
prediction2_confidence    177 non-null float64
prediction2_dog_breed     177 non-null bool
prediction3               177 non-null object
prediction3_confidence    177 non-null float64
prediction3_dog_breed     177 non-null bool
dtypes: bool(3), datetime64[ns](1), float64(6), int64(1), object(5)
memory usage: 19.9+ KB
```

Saving Master/Cleaned DataFrames as .csv

In [947]:

```
twitter_archive_master.to_csv('twitter_archive_master.csv',index=False)
#image_predictions_df_clean.to_csv('image_predictions_clean.csv',index=False)
twitter_archive_master.info()
```

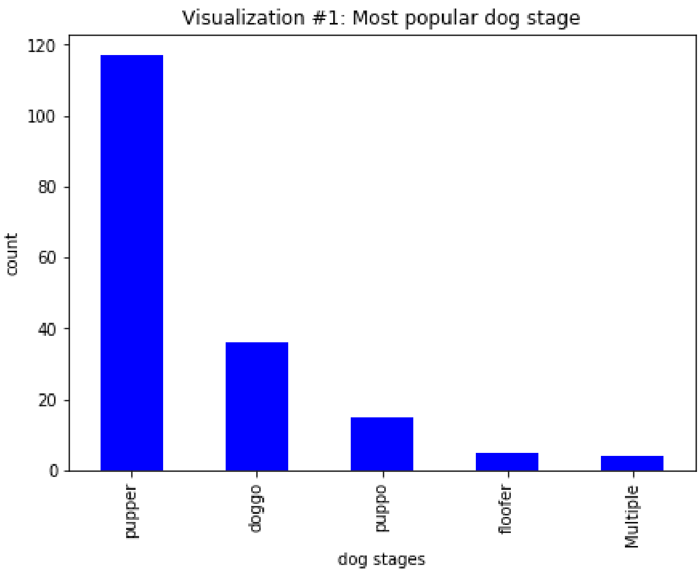
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 177 entries, 9 to 1625
Data columns (total 17 columns):
tweet_id                  177 non-null int64
timestamp                 177 non-null datetime64[ns]
rating_numerator          177 non-null float64
name                      177 non-null object
dog_stage                 177 non-null object
favorite_count            177 non-null float64
retweet_count             177 non-null float64
prediction1               177 non-null object
prediction1_confidence    177 non-null float64
prediction1_dog_breed     177 non-null bool
prediction2               177 non-null object
prediction2_confidence    177 non-null float64
prediction2_dog_breed     177 non-null bool
prediction3               177 non-null object
prediction3_confidence    177 non-null float64
prediction3_dog_breed     177 non-null bool
year                      177 non-null int64
dtypes: bool(3), datetime64[ns](1), float64(6), int64(2), object(5)
memory usage: 26.3+ KB
```

Data Analyzation and Visualization

Insight #1 What is the most popular dog stage?

In [938]:

```
twitter_archive_master.dog_stage.value_counts().plot(kind='bar',color=(0.0, 0.0, 1.0),figsize=(7,5)
);
plt.xlabel('dog stages')
plt.ylabel('count')
plt.title('Visualization #1: Most popular dog stage');
```



We can see most popular dog stage is pupper.

Insight #2 Who has the highest favorite and retweet counts?

In [939]:

```
twitter_archive_master[twitter_archive_master.favorite_count
==twitter_archive_master.favorite_count.max() ]
```

Out[939]:

	tweet_id	timestamp	rating_numerator	name	dog_stage	favorite_count	retweet_count	prediction1	prediction1_
108	866450705531457537	2017-05-22 00:28:40	13.0	Jamesy	pupper	106827.0	32883.0	French_bulldog	

As we can see most favorite dog is ***Jamesy*** at stage ***pupper*** having 106827 total favorite counts.

In [940]:

```
twitter_archive_master[twitter_archive_master.retweet_count ==twitter_archive_master.retweet_count
.max() ]
```

Out[940]:

	tweet_id	timestamp	rating_numerator	name	dog_stage	favorite_count	retweet_count	prediction1	prediction1_c
329	819004803107983360	2017-01-11 02:15:36	14.0	Bo	doggo	95450.0	42228.0	standard_poodle	

So most retweet goes to the ***doggo*** dog stage.

Insight #3 Who got the highest rating in each year?

In [941]:

```
twitter_archive_master['year'] = twitter_archive_master.timestamp.dt.year #extracting year from ti
mestamp.
```

In [942]:

```
twitter_archive_master.groupby(['year']).max()
```

Out[942]:

	tweet_id	timestamp	rating_numerator	name	dog_stage	favorite_count	retweet_count	prediction1	prediction1_c
year									
2015	682406705142087680	2015-12-31 03:43:31	12.0	Zuzu	pupper	14010.0	4581.0	wombat	
2016	814986499976527872	2016-12-31 00:08:17	27.0	Zoe	puppo	24553.0	7724.0	wood_rabbit	
2017	890240255349198849	2017-07-26 15:59:51	14.0	Yogi	puppo	106827.0	42228.0	wooden_spoon	

In the year 2015, **Zuzu** was the highest rated dog. In 2016, **Zoe** was the highest rated dog, while in 2017, **Yogi** got the highest ratings.

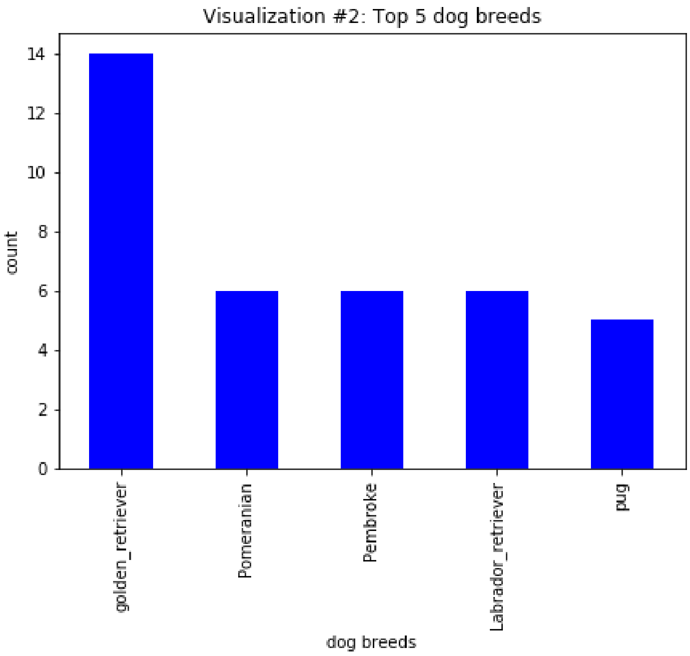
Insight #4 Most common dog breeds (top 5)

In [943]:

```
twitter_archive_master[twitter_archive_master.prediction1_dog_breed ==
True].prediction1.value_counts()[:5].plot(kind='bar',

color=(0.0, 0.0, 1.0),

figsize=(7,5));
plt.xlabel('dog breeds')
plt.ylabel('count')
plt.title('Visualization #2: Top 5 dog breeds');
```

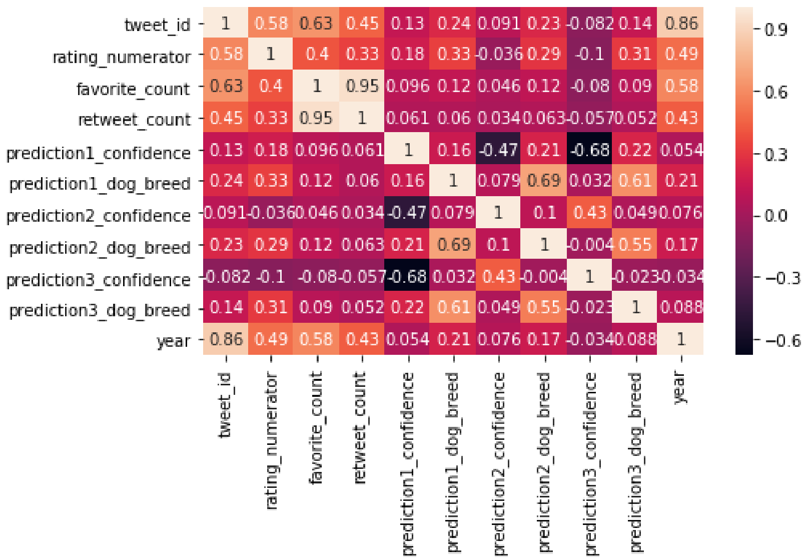


So we can visualize clearly, golden_retriever is the most common dog breed followed by Labrador_retriever.

Insight #5 Relationship between favorite count and retweet count.

In [944]:

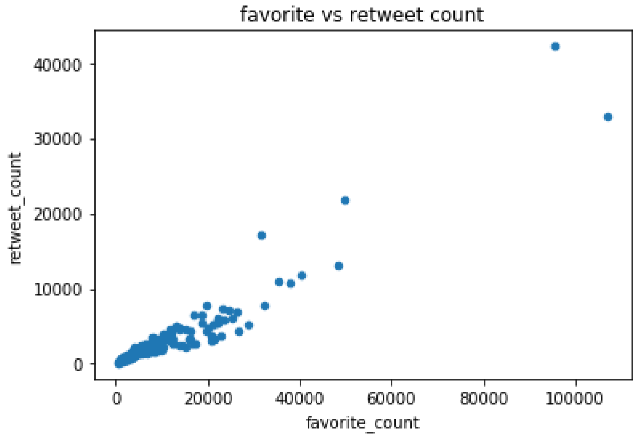
```
plt.subplots(figsize=(7,4))
sns.heatmap(twitter_archive_master.corr(),annot=True);
```



We can see clearly, there is a strong positive correlation bewteen favorite count and retweet count.

In [945]:

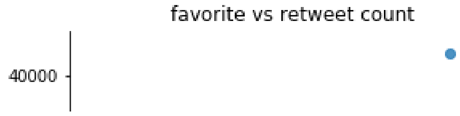
```
twitter_archive_master.plot(kind='scatter', x='favorite_count', y='retweet_count'); #Plotting scatter plot
plt.title("favorite vs retweet count");
```

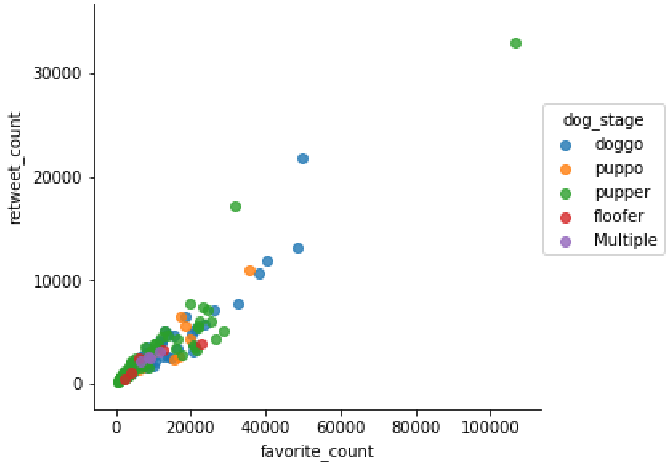


This suggests favorite dogs are more retweeted, this make sense.

In [946]:

```
sns.lmplot(data=twitter_archive_master, x='favorite_count', y='retweet_count',hue='dog_stage',fit_reg=False);
plt.title("favorite vs retweet count");
```




retweet_count
30000
20000
10000
0
favorite_count
0
20000
40000
60000
80000
100000
dog_stage
● doggo
● puppo
● pupper
● floofer
● Multiple