Display twitter-archive-enhanced Dataset

In [1]:
```
import pandas as pd
import numpy as np
from scipy import stats
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import re
import tweepy
from tweepy import OAuthHandler
import json
from timeit import default_timer as timer
```

In [2]:
```
archive_df = pd.read_csv('twitter-archive-enhanced.csv')
archive_df.head()
```

Out[2]:

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

In [7]:
```
archive_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   tweet_id                    2356 non-null   int64  
 1   in_reply_to_status_id       78 non-null     float64
 2   in_reply_to_user_id         78 non-null     float64
 3   timestamp                   2356 non-null   object 
 4   source                      2356 non-null   object 
 5   text                        2356 non-null   object 
 6   retweeted_status_id         181 non-null    float64
 7   retweeted_status_user_id    181 non-null    float64
 8   retweeted_status_timestamp  181 non-null    object 
 9   expanded_urls               2297 non-null   object 
 10  rating_numerator            2356 non-null   int64  
 11  rating_denominator          2356 non-null   int64  
 12  name                        2356 non-null   object 
 13  doggo                       2356 non-null   object 
 14  floofer                     2356 non-null   object 
 15  pupper                      2356 non-null   object 
 16  puppo                       2356 non-null   object 
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

Display image-predictions Dataset

In [3]:
```
image_predictions_df = pd.read_table("image-predictions.tsv")
image_predictions_df
```

Out[3]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springer_spaniel	0.465074	True	collie	0.156665	True	Shetland_sheepdog	0.061428	True
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	redbone	0.506826	True	miniature_pinscher	0.074192	True	Rhodesian_ridgeback	0.072010	True
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_shepherd	0.596461	True	malinois	0.138584	True	bloodhound	0.116197	True
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_ridgeback	0.408143	True	redbone	0.360687	True	miniature_pinscher	0.222752	True
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_pinscher	0.560311	True	Rottweiler	0.243682	True	Doberman	0.154629	True
...	...	...	...	...	...	...	...	...	...	...	...	...
2070	891327558926688256	https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg	2	basset	0.555712	True	English_springer	0.225770	True	German_short-haired_pointer	0.175219	True
2071	891689557279858688	https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg	1	paper_towel	0.170278	False	Labrador_retriever	0.168086	True	spatula	0.040836	False
2072	891815181378084864	https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg	1	Chihuahua	0.716012	True	malamute	0.078253	True	kelpie	0.031379	True
2073	892177421306343426	https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg	1	Chihuahua	0.323581	True	Pekinese	0.090647	True	papillon	0.068957	True
2074	892420643555336193	https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg	1	orange	0.097049	False	bagel	0.085851	False	banana	0.076110	False

2075 rows × 12 columns

In [9]:
```
image_predictions_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   tweet_id  2075 non-null   int64  
 1   jpg_url   2075 non-null   object 
 2   img_num   2075 non-null   int64  
 3   p1        2075 non-null   object 
 4   p1_conf   2075 non-null   float64
 5   p1_dog    2075 non-null   bool   
 6   p2        2075 non-null   object 
 7   p2_conf   2075 non-null   float64
 8   p2_dog    2075 non-null   bool   
 9   p3        2075 non-null   object 
 10  p3_conf   2075 non-null   float64
 11  p3_dog    2075 non-null   bool   
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

In [4]:
```
image_predictions_df[image_predictions_df['jpg_url'] == 'None']
```

Out[4]:

tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog

In [ ]:
```
# 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 = df_1.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 [5]:
```
lst = []
with open('tweet-json.txt') as json_file:
    for line in json_file:
        tweet = json.loads(line)
        tweet_id = tweet['id']
        retweet_count = tweet['retweet_count']
        fav_count = tweet['favorite_count']
        lst.append({'tweet_id':tweet_id,
                    'retweet_count': retweet_count,
                    'favorite_count': fav_count})

api_df = pd.DataFrame(lst)
```

Display tweet-json Dataset

In [6]:
```
api_df
```

Out[6]:

	tweet_id	retweet_count	favorite_count
0	892420643555336193	8853	39467
1	892177421306343426	6514	33819
2	891815181378084864	4328	25461
3	891689557279858688	8964	42908
4	891327558926688256	9774	41048
...	...	...	...
2349	666049248165822465	41	111
2350	666044226329800704	147	311
2351	666033412701032449	47	128
2352	666029285002620928	48	132
2353	666020888022790149	532	2535

2354 rows × 3 columns

In [7]:
```
api_df.describe()
```

Out[7]:

	tweet_id	retweet_count	favorite_count
count	2.354000e+03	2354.000000	2354.000000
mean	7.426978e+17	3164.797366	8080.968564
std	6.852812e+16	5284.770364	11814.771334
min	6.660209e+17	0.000000	0.000000
25%	6.783975e+17	624.500000	1415.000000
50%	7.194596e+17	1473.500000	3603.500000
75%	7.993058e+17	3652.000000	10122.250000
max	8.924206e+17	79515.000000	132810.000000

In [8]:
```
api_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 3 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   tweet_id        2354 non-null   int64
 1   retweet_count   2354 non-null   int64
 2   favorite_count  2354 non-null   int64
dtypes: int64(3)
memory usage: 55.3 KB
```

In [9]:
```
api_df.count()
```

Out[9]:
```
tweet_id          2354
retweet_count     2354
favorite_count    2354
dtype: int64
```

Assessing

In [10]:
```
archive_df['name'].value_counts()
```

Out[10]:
```
None       745
a           55
Charlie     12
Oliver      11
Cooper      11
           ...
Eugene       1
Duddles      1
Fabio        1
Colin        1
Dixie        1
Name: name, Length: 957, dtype: int64
```

In [11]:
```
#a should be null
archive_df['text'][archive_df['name'] == 'a']
```

Out[11]:
```
56      Here is a pupper approaching maximum borkdrive...
649     Here is a perfect example of someone who has t...
801     Guys this is getting so out of hand. We only r...
1002    This is a mighty rare blue-tailed hammer sherk...
1004    Viewer discretion is advised. This is a terrib...
1017    This is a carrot. We only rate dogs. Please on...
1049    This is a very rare Great Alaskan Bush Pupper....
1193    People please. This is a Deadly Mediterranean ...
1207    This is a taco. We only rate dogs. Please only...
1340    Here is a heartbreaking scene of an incredible...
1351    Here is a whole flock of puppers.  60/50 I'll ...
1361    This is a Butternut Cumberfloof. It's not wind...
1368    This is a Wild Tuscan Poofwiggle. Careful not ...
1382    "Pupper is a present to world. Here is a bow f...
1499    This is a rare Arctic Wubberfloof. Unamused by...
1737    Guys this really needs to stop. We've been ove...
1785    This is a dog swinging. I really enjoyed it so...
1853    This is a Sizzlin Menorah spaniel from Brookly...
1854    Seriously guys?! Only send in dogs. I only rat...
1877    C'mon guys. We've been over this. We only rate...
1878    This is a fluffy albino Bacardi Columbia mix. ...
1923    This is a Sagitariot Baklava mix. Loves her ne...
1941    This is a heavily opinionated dog. Loves walls...
1955    This is a Lofted Aphrodisiac Terrier named Kip...
1994    This is a baby Rand Paul. Curls for days. 11/1...
2034    This is a Tuscaloosa Alcatraz named Jacob (Yac...
2066    This is a Helvetica Listerine named Rufus. Thi...
2116    This is a Deciduous Trimester mix named Spork....
2125    This is a Rich Mahogany Seltzer named Cherokee...
2128    This is a Speckled Cauliflower Yosemite named ...
2146    This is a spotted Lipitor Rumpelstiltskin name...
2153    This is a brave dog. Excellent free climber. T...
2161    This is a Coriander Baton Rouge named Alfredo....
2191    This is a Slovakian Helter Skelter Feta named ...
2198    This is a wild Toblerone from Papua New Guinea...
2211    Here is a horned dog. Much grace. Can jump ove...
2218    This is a Birmingham Quagmire named Chuk. Love...
2222    Here is a mother dog caring for her pups. Snaz...
2235    This is a Trans Siberian Kellogg named Alfonso...
2249    This is a Shotokon Macadamia mix named Cheryl....
2255    This is a rare Hungarian Pinot named Jessiga. ...
2264    This is a southwest Coriander named Klint. Hat...
2273    This is a northern Wahoo named Kohl. He runs t...
2287    This is a Dasani Kingfisher from Maine. His na...
2304    This is a curly Ticonderoga named Pepe. No fee...
2311    This is a purebred Bacardi named Octaviath. Ca...
2314    This is a golden Buckminsterfullerene named Jo...
2327    This is a southern Vesuvius bumblegruff. Can d...
2334    This is a funny dog. Weird toes. Won't come do...
2347    My oh my. This is a rare blond Canadian terrie...
2348    Here is a Siberian heavily armored polar bear ...
2350    This is a truly beautiful English Wilson Staff...
2352    This is a purebred Piers Morgan. Loves to Netf...
```

```
2353    Here is a very happy pup. Big fan of well-main...
2354    This is a western brown Mitsubishi terrier. Up...
Name: text, dtype: object
```

In [12]:
```
archive_df.iloc[2347].text
```

Out[12]:
```
'My oh my. This is a rare blond Canadian terrier on wheels. Only $8.98. Rather docile. 9/10 very rare https://t.co/yWBqbrzy8O' (https://t.co/yWBqbrzy8O')
```

In [13]:
```
archive_df.iloc[1017]
```

Out[13]:
```
tweet_id                                                746872823977771008
in_reply_to_status_id                                                  NaN
in_reply_to_user_id                                                    NaN
timestamp                                        2016-06-26 01:08:52 +0000
source                    <a href="http://twitter.com/download/iphone" r...
text                      This is a carrot. We only rate dogs. Please on...
retweeted_status_id                                                    NaN
retweeted_status_user_id                                               NaN
retweeted_status_timestamp                                             NaN
expanded_urls             https://twitter.com/dog_rates/status/746872823... (https://twitter.com/dog_rates/status/746872823...)
rating_numerator                                                        11
rating_denominator                                                      10
name                                                                     a
doggo                                                                 None
floofer                                                               None
pupper                                                                None
puppo                                                                 None
Name: 1017, dtype: object
```

In [14]:
```
# color of carrot is orange, so the image has a dog in it who is wearing orange closes. Funny :D
image_predictions_df[image_predictions_df['tweet_id'] ==746872823977771008]
```

Out[14]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
1239	746872823977771008	https://pbs.twimg.com/media/CI1s1p7WMAA44Vk.jpg	1	Pembroke	0.540201	True	beagle	0.207835	True	Italian_greyhound	0.043565	True

In [15]:
```
#Taking 5 samples from archive_df
archive_df.text.sample(5)
```

Out[15]:
```
560     This is Marley. She's having a ruff day. Prett...
252     This is Brady. He's a recovering alcoholic. De...
2214    It is an honor to rate this pup. He is a Snork...
1505    We usually don't rate penguins but this one is...
2270    Say hello to Hall and Oates. Oates is winking ...
Name: text, dtype: object
```

In [16]:
```
archive_df.text[45]
```

Out[16]:
```
'This is Bella. She hopes her smile made you smile. If not, she is also offering you her favorite monkey. 13.5/10 https://t.co/qjrljjt948' (https://t.co/qjrljjt948')
```

In [17]:
```
#rating_numerator is 5 instaied of 13.5
archive_df.iloc[45][10:]
```

Out[17]:
```
rating_numerator          5
rating_denominator       10
name                  Bella
doggo                  None
floofer                None
pupper                 None
puppo                  None
Name: 45, dtype: object
```

In [18]:
```
archive_df.text.sample(5)
```

Out[18]:
```
2089    Two obedient dogs here. Left one has extra leg...
1655    Say hello to Jack (pronounced "Kevin"). He's a...
1769    This is Crumpet. He underestimated the snow. Q...
794     RT @dog_rates: Meet Fizz. She thinks love is a...
1226    This is Kanu. He's a Freckled Ticonderoga. Sim...
Name: text, dtype: object
```

In [19]:
```
#Name is missing "Cannon"
archive_df.text[234]
```

Out[19]:
```
'.@breaannanicolee PUPDATE: Cannon has a heart on his nose. Pupgraded to a 13/10'
```

In [20]:
```
archive_df.iloc[234][10:]
```

Out[20]:
```
rating_numerator        13
rating_denominator      10
name                  None
doggo                 None
floofer               None
pupper                None
puppo                 None
Name: 234, dtype: object
```

In [21]:
```
#wrong name 'a'
archive_df.text[1854]
```

Out[21]:
```
'Seriously guys?! Only send in dogs. I only rate dogs. This is a baby black bear... 11/10 https://t.co/H7kpabTfLj' (https://t.co/H7kpabTfLj')
```

In [22]:
```
archive_df.iloc[1854]
```

Out[22]:
```
tweet_id                                                675534494439489536
in_reply_to_status_id                                                  NaN
in_reply_to_user_id                                                    NaN
timestamp                                        2015-12-12 04:35:48 +0000
source                    <a href="http://twitter.com/download/iphone" r...
text                      Seriously guys?! Only send in dogs. I only rat...
retweeted_status_id                                                    NaN
retweeted_status_user_id                                               NaN
retweeted_status_timestamp                                             NaN
expanded_urls             https://twitter.com/dog_rates/status/675534494... (https://twitter.com/dog_rates/status/675534494...)
rating_numerator                                                        11
rating_denominator                                                      10
name                                                                     a
doggo                                                                 None
floofer                                                               None
pupper                                                                None
puppo                                                                 None
Name: 1854, dtype: object
```

In [23]:
```
image_predictions_df[image_predictions_df['tweet_id'] ==675534494439489536]
```

Out[23]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
491	675534494439489536	https://pbs.twimg.com/media/CV_7CV6XIAEV05u.jpg	1	chow	0.749368	True	schipperke	0.133738	True	Newfoundland	0.049914	True

In [24]:
```
#In the expanded_url column of the archive_df, the missing values are for tweets without photos so those entries can be dropped safely.
archive_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   tweet_id                    2356 non-null   int64  
 1   in_reply_to_status_id       78 non-null     float64
 2   in_reply_to_user_id         78 non-null     float64
 3   timestamp                   2356 non-null   object 
 4   source                      2356 non-null   object 
 5   text                        2356 non-null   object 
 6   retweeted_status_id         181 non-null    float64
 7   retweeted_status_user_id    181 non-null    float64
 8   retweeted_status_timestamp  181 non-null    object 
 9   expanded_urls               2297 non-null   object 
 10  rating_numerator            2356 non-null   int64  
 11  rating_denominator          2356 non-null   int64  
 12  name                        2356 non-null   object 
 13  doggo                       2356 non-null   object 
 14  floofer                     2356 non-null   object 
 15  pupper                      2356 non-null   object 
 16  puppo                       2356 non-null   object 
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

In [25]:
```
archive_df['rating_numerator'].value_counts()
```

Out[25]:
```
12      558
11      464
10      461
13      351
9       158
8       102
7        55
14       54
5        37
6        32
3        19
4        17
1         9
2         9
0         2
15        2
75        2
420       2
182       1
204       1
143       1
121       1
99        1
20        1
45        1
27        1
17        1
24        1
26        1
44        1
50        1
60        1
80        1
84        1
88        1
1776      1
960       1
666       1
144       1
165       1
Name: rating_numerator, dtype: int64
```

In [26]:
```
archive_df['rating_denominator'].value_counts()
```

Out[26]:
```
10     2333
11        3
50        3
20        2
80        2
0         1
120       1
7         1
170       1
150       1
130       1
90        1
110       1
2         1
70        1
40        1
16        1
15        1
Name: rating_denominator, dtype: int64
```

In [27]:
```
image_predictions_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   tweet_id  2075 non-null   int64  
 1   jpg_url   2075 non-null   object 
 2   img_num   2075 non-null   int64  
 3   p1        2075 non-null   object 
 4   p1_conf   2075 non-null   float64
 5   p1_dog    2075 non-null   bool   
 6   p2        2075 non-null   object 
 7   p2_conf   2075 non-null   float64
 8   p2_dog    2075 non-null   bool   
 9   p3        2075 non-null   object 
 10  p3_conf   2075 non-null   float64
 11  p3_dog    2075 non-null   bool   
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

In [28]:
```
image_predictions_df.count()
```

Out[28]:
```
tweet_id    2075
jpg_url     2075
img_num     2075
p1          2075
p1_conf     2075
p1_dog      2075
p2          2075
p2_conf     2075
p2_dog      2075
p3          2075
p3_conf     2075
p3_dog      2075
dtype: int64
```

Quality

- resert the indexes in all tables after droping some rows*

Archive Enhanced Tabel

- wrong datatype in (tweet_id) (int >>> str)*
- (2356 - 78) Missing recordes in (in_reply_to_status_id & in_reply_to_user_id) columns*
- wrong datatype in (in_reply_to_status_id & in_reply_to_user_id) columns (float >>> str)*
- wrong datatype in (timestamp) column (object >> date)*
- (2356 - 181) Missing recordes in (retweeted_status_id & retweeted_status_user_id & retweeted_status_timestamp) columns*
- wrong datatype in (retweeted_status_id & retweeted_status_user_id) columns (float >>> str)*
- wrong datatype in (retweeted_status_timestamp) columns (object >>> date)*
- (2356 - 2297) Missing recordes in (expanded_urls) columns which can be droped(data with no images)*
- rating isn't alawys correct (like in Bella at index 45)*
- wrong datatype in (rating_numerator) column (int >>> float)*
- missing &wrong data in name column*
- missing &wrong data in (doggo, floofer, pupper, puppo) columns*
- wrong representation of null value in (name,doggo, floofer, pupper, puppo) columns (None >> Nan)*
- retweetes & replies should be removed*
- tweet ids with no images*
- tweet id = 670842764863651840 is not a dog, numerator & denominator >> Null*
- tweet id = 749981277374128128 is a dog but with no rating, numerator & denominator >> Null*
- numerator = 24 is wrong >> null*
- (dog_stage) Dealing with (doggopupper,doggopuppo,doggofloofer)*

Image Predictions Tabel

- wrong datatype in (tweet_id) (int >>> str)*

- bad column names*
- number of entries = 2075 (<2356 in archive) >>> some tweets without images will be deleted*
- retweets & replies should be removed*

Api Tabel

- wrong datatype in (tweet_id) (int >>> str)*
- number of entries = 2354 >>> some tweets will be deleted*

Tidiness

- (doggo, floofer, pupper, puppo) columns in Archive Enhanced Tabel should be compined into one column (stage)*
- (in_reply_to_status_id & in_reply_to_user_id & retweeted_status_id & retweeted_status_user_id & retweeted_status_timestamp) columns in Archive Enhanced Tabel need to be removed.*
- api tabel should be with the archive table in one table*

Cleaning

In [29]:
```
archive_clean = archive_df.copy()
image_predictions_clean = image_predictions_df.copy()
api_clean = api_df.copy()
```

(Archive) retweetes & replies should be removed

Define

- drop rows in archive which have retweetes & replies

Code

In [30]:
```
#remove rows that have non null data in retweeted_status_id & in_reply_to_status_id
archive_clean = archive_clean[~archive_clean['retweeted_status_id'].notnull()]
archive_clean = archive_clean[~archive_clean['in_reply_to_status_id'].notnull()]
```

Test

In [31]:
```
#should be zero
archive_clean['retweeted_status_id'].count()
```

Out[31]:
```
0
```

In [32]:
```
#should be zero
archive_clean['in_reply_to_status_id'].count()
```

Out[32]:
```
0
```

(archive)(2356 - 2297) Missing recordes in (expanded_urls) columns which can be droped(data with no images)

Define

- drop the rows which have missing values

Code

In [33]:
```
archive_clean = archive_clean[archive_clean['expanded_urls'].notnull()]
```

Test

In [34]:
```
#Should be zero
archive_clean['expanded_urls'].isnull().sum()
```

Out[34]:
```
0
```

(Archive) tweet ids with no images

Define

- drop ids in acrhive which don't have images guided by Image Prediction Tabel

Code

In [35]:
```
# creating a list of tweet_ids with images
tweets_with_image = list(image_predictions_clean.tweet_id)
#droping ids which aren't in image prediction
archive_clean = archive_clean[archive_clean.tweet_id.isin(tweets_with_image)]
```

Test

In [36]:
```
archive_clean.shape[0]
```

Out[36]:
```
1971
```

In [37]:
```
image_predictions_clean.shape[0]
```

Out[37]:
```
2075
```

(Image Prediction) retweets & replies should be removed

Define

- drop retweets & replies guided by Archive Table

Code

In [38]:
```
# creating a list of tweet_ids that are in Archive_clean
tweets_in_arc = list(archive_clean.tweet_id)
#droping ids which aren't in Archive_clean
image_predictions_clean = image_predictions_clean[image_predictions_clean.tweet_id.isin(tweets_in_arc)]
```

Test

In [39]:
```
image_predictions_clean.shape[0]
```

Out[39]:
```
1971
```

(Archive) (in_reply_to_status_id & in_reply_to_user_id & retweeted_status_id & retweeted_status_user_id & retweeted_status_timestamp & expanded_urls & source) columns in Archive Enhanced Tabel need to be removed.

Define

- drop those columns, no longer needed

Code

In [40]:
```
archive_clean = archive_clean.drop(['in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id'
                                    ,'retweeted_status_user_id','retweeted_status_timestamp', 'source', 'expanded_urls'],axis=1)
```

Test

In [41]:
```
archive_clean
```

Out[41]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10	Phineas	None	None	None	None
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10	Tilly	None	None	None	None
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12	10	Archie	None	None	None	None
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13	10	Darla	None	None	None	None
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12	10	Franklin	None	None	None	None
...	...	...	...	...	...	...	...	...	...	...
2351	666049248165822465	2015-11-16 00:24:50 +0000	Here we have a 1949 1st generation vulpix. Enj...	5	10	None	None	None	None	None
2352	666044226329800704	2015-11-16 00:04:52 +0000	This is a purebred Piers Morgan. Loves to Netf...	6	10	a	None	None	None	None
2353	666033412701032449	2015-11-15 23:21:54 +0000	Here is a very happy pup. Big fan of well-main...	9	10	a	None	None	None	None
2354	666029285002620928	2015-11-15 23:05:30 +0000	This is a western brown Mitsubishi terrier. Up...	7	10	a	None	None	None	None
2355	666020888022790149	2015-11-15 22:32:08 +0000	Here we have a Japanese Irish Setter. Lost eye...	8	10	None	None	None	None	None

1971 rows × 10 columns

(Api Table) number of entries = 2354 >>> some tweets will be deleted

Define

- Deleting tweets that aren't in archive table

Code

In [42]:
```
# creating a list of tweet_ids that are in Archive_clean
tweets_in_arc = list(archive_clean.tweet_id)
#droping ids which aren't in Archive_clean
api_clean = api_clean[api_clean.tweet_id.isin(tweets_in_arc)]
```

Test

In [43]:
```
api_clean.shape[0]
```

Out[43]: 1971

resert the indexes in all tables after droping some rows

Define

- reset indexes

Code

In [44]:
```
archive_clean = archive_clean.reset_index(drop=True)
image_predictions_clean = image_predictions_clean.reset_index(drop=True)
api_clean = api_clean.reset_index(drop=True)
```

Test

In [45]:
```
archive_clean
```

Out[45]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10	Phineas	None	None	None	None
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10	Tilly	None	None	None	None
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12	10	Archie	None	None	None	None
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13	10	Darla	None	None	None	None
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12	10	Franklin	None	None	None	None
...	...	...	...	...	...	...	...	...	...	...
1966	666049248165822465	2015-11-16 00:24:50 +0000	Here we have a 1949 1st generation vulpix. Enj...	5	10	None	None	None	None	None
1967	666044226329800704	2015-11-16 00:04:52 +0000	This is a purebred Piers Morgan. Loves to Netf...	6	10	a	None	None	None	None
1968	666033412701032449	2015-11-15 23:21:54 +0000	Here is a very happy pup. Big fan of well-main...	9	10	a	None	None	None	None
1969	666029285002620928	2015-11-15 23:05:30 +0000	This is a western brown Mitsubishi terrier. Up...	7	10	a	None	None	None	None
1970	666020888022790149	2015-11-15 22:32:08 +0000	Here we have a Japanese Irish Setter. Lost eye...	8	10	None	None	None	None	None

1971 rows × 10 columns

In [46]:
```
image_predictions_clean
```

Out[46]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springer_spaniel	0.465074	True	collie	0.156665	True	Shetland_sheepdog	0.061428	True
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	redbone	0.506826	True	miniature_pinscher	0.074192	True	Rhodesian_ridgeback	0.072010	True
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_shepherd	0.596461	True	malinois	0.138584	True	bloodhound	0.116197	True
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_ridgeback	0.408143	True	redbone	0.360687	True	miniature_pinscher	0.222752	True
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_pinscher	0.560311	True	Rottweiler	0.243682	True	Doberman	0.154629	True
...	...	...	...	...	...	...	...	...	...	...	...	...
1966	891327558926688256	https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg	2	basset	0.555712	True	English_springer	0.225770	True	German_short-haired_pointer	0.175219	True
1967	891689557279858688	https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg	1	paper_towel	0.170278	False	Labrador_retriever	0.168086	True	spatula	0.040836	False
1968	891815181378084864	https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg	1	Chihuahua	0.716012	True	malamute	0.078253	True	kelpie	0.031379	True
1969	892177421306343426	https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg	1	Chihuahua	0.323581	True	Pekinese	0.090647	True	papillon	0.068957	True
1970	892420643555336193	https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg	1	orange	0.097049	False	bagel	0.085851	False	banana	0.076110	False

1971 rows × 12 columns

In [47]:
```
api_clean
```

Out[47]:

	tweet_id	retweet_count	favorite_count
0	892420643555336193	8853	39467
1	892177421306343426	6514	33819
2	891815181378084864	4328	25461
3	891689557279858688	8964	42908
4	891327558926688256	9774	41048
...	...	...	...
1966	666049248165822465	41	111
1967	666044226329800704	147	311
1968	666033412701032449	47	128
1969	666029285002620928	48	132
1970	666020888022790149	532	2535

1971 rows × 3 columns

(Archive) rating isn't alawys correct

Define

- convert numerator to float
- scrap the text for the right rating value
- for big values of numerator & denominator, get the average

In [48]:
```
archive_clean.iloc[1696]
```

Out[48]:
```
tweet_id                                            670842764863651840
timestamp                                 2015-11-29 05:52:33 +0000
text                 After so many requests... here you go.\n\nGood...
rating_numerator                                                   420
rating_denominator                                                  10
name                                                              None
doggo                                                             None
floofer                                                           None
pupper                                                            None
puppo                                                             None
Name: 1696, dtype: object
```

Code

In [49]:
```
# convert to float
archive_clean['rating_numerator'] = archive_clean['rating_numerator'].astype(float)

# scrap the text for the right numerator rating value
archive_clean['rating_numerator'] = archive_clean['text'].str.extract('(\d+\.?\d?\d?)\/\d{1,3}', expand = False).astype('float')

# getting num of dogs for big values of numerator & denominator
dogs_num = archive_clean['rating_denominator'][archive_clean['rating_denominator'] >= 20]/10

# deviding every rating_numerator & rating_denominator that have many dogs included in the rating, to get the average
for i in dogs_num:
    idx = dogs_num.index[dogs_num == i]
    archive_clean['rating_numerator'][idx] = archive_clean['rating_numerator'][idx]/dogs_num[idx]
    archive_clean['rating_denominator'][idx] = 10
```

```
c:\users\eslam\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:13: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  del sys.path[0]
c:\users\eslam\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:14: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  del sys.path[0]
c:\users\eslam\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:14: SettingWithCopyWarning:
```

Test

In [ ]:
```
```

In [50]:
```
archive_clean.sample(5)
```

Out[50]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
167	850145622816686080	2017-04-07 00:38:06 +0000	This is Riley. He's making new friends. Jubila...	11.0	10	Riley	None	None	None	None
721	749996283729883136	2016-07-04 16:00:22 +0000	This is Bo. He emanates happiness. 12/10 I cou...	12.0	10	Bo	None	None	None	None
411	803380650405482500	2016-11-28 23:30:47 +0000	Meet Sonny. He's an in-home movie critic. That...	12.0	10	Sonny	None	None	None	None
1634	672248013293752320	2015-12-03 02:56:30 +0000	10/10 for dog. 7/10 for cat. 12/10 for human. ...	10.0	10	None	None	None	None	None
735	748575535303884801	2016-06-30 17:54:50 +0000	This is one of the most reckless puppers I've ...	6.0	10	one	None	None	None	None

In [51]:
```
archive_clean[archive_clean['name'] == 'Bella']
```

Out[51]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
39	883482846933004288	2017-07-08 00:28:19 +0000	This is Bella. She hopes her smile made you sm...	13.5	10	Bella	None	None	None	None
53	880465832366813184	2017-06-29 16:39:47 +0000	This is Bella. She had her first beach experie...	12.0	10	Bella	None	None	None	None
813	737800304142471168	2016-06-01 00:17:54 +0000	This is Bella. She's ubering home after a few ...	10.0	10	Bella	None	None	None	None
1048	703631701117943808	2016-02-27 17:24:05 +0000	This is Bella. Based on this picture she's at ...	11.0	10	Bella	None	None	None	None
1411	678389028614488064	2015-12-20 01:38:42 +0000	This is Bella. She just learned that her final...	11.0	10	Bella	None	None	pupper	None
1588	673350198937153538	2015-12-06 03:56:12 +0000	This is Bella. She's a Genghis Flopped Canuck....	9.0	10	Bella	None	None	None	None

In [52]:
```
archive_clean['rating_denominator'].value_counts()
```

Out[52]:
```
10    1967
11       2
2        1
7        1
Name: rating_denominator, dtype: int64
```

In [53]:
```
archive_clean['rating_numerator'].value_counts()
```

Out[53]:
```
12.000      449
10.000      417
11.000      396
13.000      253
9.000       150
8.000        95
7.000        52
14.000       33
5.000        32
6.000        32
3.000        19
4.000        15
2.000        10
1.000         5
420.000       1
11.260        1
0.480         1
0.000         1
0.400         1
1.375         1
1776.000      1
0.360         1
13.500        1
11.270        1
24.000        1
1.250         1
9.750         1
Name: rating_numerator, dtype: int64
```

Assis

In [54]:
```
# outlier, tweet id = 670842764863651840 is not a dog, numerator & denominator >> Null
archive_clean[archive_clean['rating_numerator'] == 420.000]
```

Out[54]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
1696	670842764863651840	2015-11-29 05:52:33 +0000	After so many requests... here you go.\n\nGood...	420.0	10	None	None	None	None	None

In [55]:
```
image_predictions_clean[image_predictions_clean['tweet_id'] == 670842764863651840]
```

Out[55]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
274	670842764863651840	https://pbs.twimg.com/media/CU9P717W4AAOlKx.jpg	1	microphone	0.096063	False	accordion	0.094075	False	drumstick	0.061113	False

In [56]:
```
# tweet id = 749981277374128128 doesn't have a raiting, numerator & denominator >> Null
archive_clean[archive_clean['rating_numerator'] == 1776.000]
```

Out[56]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
722	749981277374128128	2016-07-04 15:00:45 +0000	This is Atticus. He's quite simply America af....	1776.0	10	Atticus	None	None	None	None

In [57]:
```
# it's a dog but with no rating, numerator & denominator >> Null
image_predictions_clean[image_predictions_clean['tweet_id'] == 749981277374128128]
```

Out[57]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
1248	749981277374128128	https://pbs.twimg.com/media/CmgBZ7kWcAAIzFD.jpg	1	bow_tie	0.533941	False	sunglasses	0.080822	False	sunglass	0.050776	False

In [58]:
```
archive_clean['name']
```

Out[58]:
```
0        Phineas
1          Tilly
2         Archie
3          Darla
4       Franklin
          ...
1966        None
1967           a
1968           a
1969           a
1970        None
Name: name, Length: 1971, dtype: object
```

In [59]:
```
archive_clean['name'].value_counts()
```

Out[59]:
```
None       524
a           55
Charlie     11
Cooper      10
Oliver      10
           ...
Lulu         1
Chesney      1
Cupid        1
Bruiser      1
Dixie        1
Name: name, Length: 935, dtype: int64
```

missing & wrong data in name column

tweet id = 670842764863651840 is not a dog, numerator & denominator >> Null
tweet id = 749981277374128128 is a dog but with no rating, numerator & denominator >> Null
numerator = 24 is wrong >> null

Define

- replase those ratings with null

Code

In [60]:
```
archive_clean.rating_numerator[archive_clean['tweet_id'] == 670842764863651840] = np.nan
archive_clean.rating_numerator[archive_clean['tweet_id'] == 749981277374128128] = np.nan
archive_clean.rating_numerator[archive_clean['rating_numerator'] == 24.000] = np.nan
```

```
c:\users\eslam\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  """Entry point for launching an IPython kernel.
c:\users\eslam\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

c:\users\eslam\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  This is separate from the ipykernel package so we can avoid doing imports until
```

Test

In [61]:
```
archive_clean['rating_numerator'].value_counts()
```

Out[61]:
```
12.000    449
10.000    417
11.000    396
13.000    253
9.000     150
8.000      95
7.000      52
14.000     33
5.000      32
6.000      32
3.000      19
4.000      15
2.000      10
1.000       5
0.400       1
9.750       1
1.375       1
11.260      1
0.360       1
13.500      1
11.270      1
0.480       1
1.250       1
0.000       1
Name: rating_numerator, dtype: int64
```

Define

- scrap the right name from text column
- replace missing names and None with NaN

Code

In [62]:
```
pattern_2 = re.compile(r'(?:name(?:d)?)\s{1}(?:is\s)?([A-Za-z]+)')
for index, row in archive_clean.iterrows():
    if row['name'][0].islower() or row['name'] == 'None':
        try:
            c_name = re.findall(pattern_2, row['text'])[0]
            archive_clean.loc[index,'name'] = archive_clean.loc[index,'name'].replace(row['name'], c_name)

        except IndexError:
            archive_clean.loc[index,'name'] = np.nan
```

Test

In [63]:
```
archive_clean
```

Out[63]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13.0	10	Phineas	None	None	None	None
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13.0	10	Tilly	None	None	None	None
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12.0	10	Archie	None	None	None	None
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13.0	10	Darla	None	None	None	None
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12.0	10	Franklin	None	None	None	None
...	...	...	...	...	...	...	...	...	...	...
1966	666049248165822465	2015-11-16 00:24:50 +0000	Here we have a 1949 1st generation vulpix. Enj...	5.0	10	NaN	None	None	None	None
1967	666044226329800704	2015-11-16 00:04:52 +0000	This is a purebred Piers Morgan. Loves to Netf...	6.0	10	NaN	None	None	None	None
1968	666033412701032449	2015-11-15 23:21:54 +0000	Here is a very happy pup. Big fan of well-main...	9.0	10	NaN	None	None	None	None
1969	666029285002620928	2015-11-15 23:05:30 +0000	This is a western brown Mitsubishi terrier. Up...	7.0	10	NaN	None	None	None	None
1970	666020888022790149	2015-11-15 22:32:08 +0000	Here we have a Japanese Irish Setter. Lost eye...	8.0	10	NaN	None	None	None	None

1971 rows × 10 columns

In [64]:
```
archive_clean['name'].value_counts()
```

Out[64]:
```
Charlie    11
Cooper     10
Lucy       10
Oliver     10
Penny       9
           ..
Chesney     1
Cupid       1
Bruiser     1
Dotsy       1
Dixie       1
Name: name, Length: 936, dtype: int64
```

In [65]:
```
archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1971 entries, 0 to 1970
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   tweet_id            1971 non-null   int64  
 1   timestamp           1971 non-null   object 
 2   text                1971 non-null   object 
 3   rating_numerator    1968 non-null   float64
 4   rating_denominator  1971 non-null   int64  
 5   name                1379 non-null   object 
 6   doggo               1971 non-null   object 
 7   floofer             1971 non-null   object 
 8   pupper              1971 non-null   object 
 9   puppo               1971 non-null   object 
dtypes: float64(1), int64(2), object(7)
memory usage: 154.1+ KB
```

(Archive) wrong representation of null value in (doggo, floofer, pupper, puppo) columns (None >> Nan)

(Archive) (doggo, floofer, pupper, puppo) columns in Archive Enhanced Tabel should be compined into one column (stage)

Define

- replase every none with "" empty
- compine the three columns into one (dog_stage)
- replace "" with nan

In [66]:
```
archive_clean['doggo'].value_counts()
```

Out[66]:
```
None     1898
doggo      73
Name: doggo, dtype: int64
```

In [67]:
```
archive_clean['floofer'].value_counts()
```

Out[67]:
```
None       1963
floofer       8
Name: floofer, dtype: int64
```

In [68]:
```
archive_clean['pupper'].value_counts()
```

Out[68]:
```
None      1762
pupper     209
Name: pupper, dtype: int64
```

In [69]:
```
archive_clean['puppo'].value_counts()
```

Out[69]:
```
None     1948
puppo      23
Name: puppo, dtype: int64
```

Code

In [70]:
```
# replase none with ""
archive_clean['doggo'] = archive_clean['doggo'].replace('None','')
archive_clean['floofer'] = archive_clean['floofer'].replace('None','')
archive_clean['pupper'] = archive_clean['pupper'].replace('None','')
archive_clean['puppo'] = archive_clean['puppo'].replace('None','')

# Compine 4 columns into one
archive_clean['dog_stage'] = archive_clean['doggo'] + archive_clean['floofer'] + archive_clean['pupper'] + archive_clean['puppo']

# replace "" with nan
archive_clean['dog_stage'] = archive_clean['dog_stage'].replace('', np.nan)

# drop (doggo, floofer, pupper, puppo) columns
columns = ['doggo', 'floofer', 'pupper', 'puppo']
archive_clean.drop(columns, inplace=True, axis=1)
```

Test

In [71]:
```
archive_clean
```

Out[71]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	dog_stage
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13.0	10	Phineas	NaN
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13.0	10	Tilly	NaN
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12.0	10	Archie	NaN
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13.0	10	Darla	NaN
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12.0	10	Franklin	NaN
...	...	...	...	...	...	...	...
1966	666049248165822465	2015-11-16 00:24:50 +0000	Here we have a 1949 1st generation vulpix. Enj...	5.0	10	NaN	NaN
1967	666044226329800704	2015-11-16 00:04:52 +0000	This is a purebred Piers Morgan. Loves to Netf...	6.0	10	NaN	NaN
1968	666033412701032449	2015-11-15 23:21:54 +0000	Here is a very happy pup. Big fan of well-main...	9.0	10	NaN	NaN
1969	666029285002620928	2015-11-15 23:05:30 +0000	This is a western brown Mitsubishi terrier. Up...	7.0	10	NaN	NaN
1970	666020888022790149	2015-11-15 22:32:08 +0000	Here we have a Japanese Irish Setter. Lost eye...	8.0	10	NaN	NaN

1971 rows × 7 columns

In [72]:
```
archive_clean['dog_stage'].value_counts()
```

Out[72]:
```
pupper          201
doggo            63
puppo            22
doggopupper       8
floofer           7
doggofloofer      1
doggopuppo        1
Name: dog_stage, dtype: int64
```

(dog_stage) Dealing with (doggopupper,doggopuppo,doggofloofer)

Define

- seprate them by -

Code

In [73]:
```
for i in range (1971):
    if archive_clean['dog_stage'][i]=='doggopupper':
        archive_clean['dog_stage'][i] = 'doggo-pupper'

    if archive_clean['dog_stage'][i]=='doggopuppo':
        archive_clean['dog_stage'][i] = 'doggo-puppo'

    if archive_clean['dog_stage'][i]=='doggofloofer':
        archive_clean['dog_stage'][i] = 'doggo-floofer'
```

```
c:\users\eslam\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

c:\users\eslam\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:9: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  if __name__ == '__main__':
c:\users\eslam\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  This is separate from the ipykernel package so we can avoid doing imports until
```

Test

In [74]:
```
archive_clean['dog_stage'].value_counts()
```

Out[74]:
```
pupper           201
doggo             63
puppo             22
doggo-pupper       8
floofer            7
doggo-floofer      1
doggo-puppo        1
Name: dog_stage, dtype: int64
```

(Archive & api) api tabel should be with the archive table in one table

Define

- merg two tables into one master table

Code

In [75]:
```
master_df = pd.merge(archive_clean,api_clean, on ='tweet_id' , how = 'left')
```

Test

In [76]:
```
master_df
```

Out[76]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	dog_stage	retweet_count	favorite_count
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13.0	10	Phineas	NaN	8853	39467
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13.0	10	Tilly	NaN	6514	33819
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12.0	10	Archie	NaN	4328	25461
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13.0	10	Darla	NaN	8964	42908
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12.0	10	Franklin	NaN	9774	41048
...	...	...	...	...	...	...	...	...	...
1966	666049248165822465	2015-11-16 00:24:50 +0000	Here we have a 1949 1st generation vulpix. Enj...	5.0	10	NaN	NaN	41	111
1967	666044226329800704	2015-11-16 00:04:52 +0000	This is a purebred Piers Morgan. Loves to Netf...	6.0	10	NaN	NaN	147	311
1968	666033412701032449	2015-11-15 23:21:54 +0000	Here is a very happy pup. Big fan of well-main...	9.0	10	NaN	NaN	47	128
1969	666029285002620928	2015-11-15 23:05:30 +0000	This is a western brown Mitsubishi terrier. Up...	7.0	10	NaN	NaN	48	132
1970	666020888022790149	2015-11-15 22:32:08 +0000	Here we have a Japanese Irish Setter. Lost eye...	8.0	10	NaN	NaN	532	2535

1971 rows × 9 columns

(api) wrong datatype in (tweet_id) (int >>> str)
(archive) wrong datatype in (tweet_id) (int >>> str)

they both now in master_df, so only deal with it one time
(image_predictions_clean) wrong datatype in (tweet_id) (int >>> str)

Define

- convert tweet id to data datatype

Code

In [77]:
```
master_df['tweet_id'] = master_df['tweet_id'].astype('str')
image_predictions_clean['tweet_id'] = image_predictions_clean['tweet_id'].astype('str')
```

Test

In [78]:
```
master_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1971 entries, 0 to 1970
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   tweet_id            1971 non-null   object 
 1   timestamp           1971 non-null   object 
 2   text                1971 non-null   object 
 3   rating_numerator    1968 non-null   float64
 4   rating_denominator  1971 non-null   int64  
 5   name                1379 non-null   object 
 6   dog_stage           303 non-null    object 
 7   retweet_count       1971 non-null   int64  
 8   favorite_count      1971 non-null   int64  
dtypes: float64(1), int64(3), object(5)
memory usage: 154.0+ KB
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1971 entries, 0 to 1970
Data columns (total 12 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   tweet_id  1971 non-null   object 
 1   jpg_url   1971 non-null   object 
 2   img_num   1971 non-null   int64  
 3   p1        1971 non-null   object 
 4   p1_conf   1971 non-null   float64
 5   p1_dog    1971 non-null   bool   
 6   p2        1971 non-null   object 
 7   p2_conf   1971 non-null   float64
 8   p2_dog    1971 non-null   bool   
 9   p3        1971 non-null   object 
 10  p3_conf   1971 non-null   float64
 11  p3_dog    1971 non-null   bool   
dtypes: bool(3), float64(3), int64(1), object(5)
memory usage: 144.5+ KB
```

(archive)(master) wrong datatype in (timestamp) column (object >> date)

Define

- change datatype to date

Code

In [80]:
```
master_df['timestamp'] = pd.to_datetime(master_df['timestamp'])
```

Test

In [81]:
```
master_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1971 entries, 0 to 1970
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   tweet_id            1971 non-null   object
 1   timestamp           1971 non-null   datetime64[ns, UTC]
 2   text                1971 non-null   object
 3   rating_numerator    1968 non-null   float64
 4   rating_denominator  1971 non-null   int64
 5   name                1379 non-null   object
 6   dog_stage           303 non-null    object
 7   retweet_count       1971 non-null   int64
 8   favorite_count      1971 non-null   int64
dtypes: datetime64[ns, UTC](1), float64(1), int64(3), object(4)
memory usage: 154.0+ KB
```

In [82]:
```
master_df
```

Out[82]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	dog_stage	retweet_count	favorite_count
0	892420643555336193	2017-08-01 16:23:56+00:00	This is Phineas. He's a mystical boy. Only eve...	13.0	10	Phineas	NaN	8853	39467
1	892177421306343426	2017-08-01 00:17:27+00:00	This is Tilly. She's just checking pup on you....	13.0	10	Tilly	NaN	6514	33819
2	891815181378084864	2017-07-31 00:18:03+00:00	This is Archie. He is a rare Norwegian Pouncin...	12.0	10	Archie	NaN	4328	25461
3	891689557279858688	2017-07-30 15:58:51+00:00	This is Darla. She commenced a snooze mid meal...	13.0	10	Darla	NaN	8964	42908
4	891327558926688256	2017-07-29 16:00:24+00:00	This is Franklin. He would like you to stop ca...	12.0	10	Franklin	NaN	9774	41048
...	...	...	...	...	...	...	...	...	...
1966	666049248165822465	2015-11-16 00:24:50+00:00	Here we have a 1949 1st generation vulpix. Enj...	5.0	10	NaN	NaN	41	111
1967	666044226329800704	2015-11-16 00:04:52+00:00	This is a purebred Piers Morgan. Loves to Netf...	6.0	10	NaN	NaN	147	311
1968	666033412701032449	2015-11-15 23:21:54+00:00	Here is a very happy pup. Big fan of well-main...	9.0	10	NaN	NaN	47	128
1969	666029285002620928	2015-11-15 23:05:30+00:00	This is a western brown Mitsubishi terrier. Up...	7.0	10	NaN	NaN	48	132
1970	666020888022790149	2015-11-15 22:32:08+00:00	Here we have a Japanese Irish Setter. Lost eye...	8.0	10	NaN	NaN	532	2535

1971 rows × 9 columns

(image_predictions) bad column names

Define

- change the column names to somthing representive

Code

In [83]:
```
image_predictions_clean = image_predictions_clean.rename(columns={'p1': '1st_prediction'
                                                                  ,'p1_conf': '1st_prediction_confidence%'
                                                                  ,'p1_dog' : 'is_dog_breed1'

                                                                  ,'p2': '2nd_prediction'
                                                                  ,'p2_conf': '2nd_prediction_confidence%'
                                                                  ,'p2_dog' : 'is_dog_breed2'

                                                                  ,'p3': '3rd_prediction'
                                                                  ,'p3_conf': '3rd_prediction_confidence%'
                                                                  ,'p3_dog' : 'is_dog_breed3'

                                                                  ,'jpg_url' : 'image_url'
                                                                  ,'img_num' : 'image_number'})
```

Test

In [84]:
```
image_predictions_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1971 entries, 0 to 1970
Data columns (total 12 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   tweet_id                   1971 non-null   object 
 1   image_url                  1971 non-null   object 
 2   image_number               1971 non-null   int64  
 3   1st_prediction             1971 non-null   object 
 4   1st_prediction_confidence%  1971 non-null   float64
 5   is_dog_breed1              1971 non-null   bool   
 6   2nd_prediction             1971 non-null   object 
 7   2nd_prediction_confidence%  1971 non-null   float64
 8   is_dog_breed2              1971 non-null   bool   
 9   3rd_prediction             1971 non-null   object 
 10  3rd_prediction_confidence%  1971 non-null   float64
 11  is_dog_breed3              1971 non-null   bool   
dtypes: bool(3), float64(3), int64(1), object(5)
memory usage: 144.5+ KB
```

Store clean dataframes

In [85]:
```
master_df.to_csv('twitter_archive_master.csv')
image_predictions_clean.to_csv('image_predictions_clean.csv')
```

Analysis & Visualization

In [100]:
```
master_df.describe()
```

Out[100]:

	rating_numerator	rating_denominator	retweet_count	favorite_count	year
count	1968.000000	1971.000000	1971.000000	1971.000000	1971.000000
mean	10.509474	9.995434	2784.449518	8949.106545	2015.847286
std	2.238268	0.195065	4697.662893	12267.799790	0.699178
min	0.000000	2.000000	16.000000	81.000000	2015.000000
25%	10.000000	10.000000	628.500000	1997.000000	2015.000000
50%	11.000000	10.000000	1367.000000	4147.000000	2016.000000
75%	12.000000	10.000000	3239.000000	11402.500000	2016.000000
max	14.000000	11.000000	79515.000000	132810.000000	2017.000000

In [99]:
```
image_predictions_clean.describe()
```

Out[99]:

	image_number	1st_prediction_confidence%	2nd_prediction_confidence%	3rd_prediction_confidence%
count	1971.000000	1971.000000	1.971000e+03	1.971000e+03
mean	1.201928	0.594558	1.345850e-01	6.016556e-02
std	0.559020	0.272126	1.010527e-01	5.094156e-02
min	1.000000	0.044333	1.011300e-08	1.740170e-10
25%	1.000000	0.363091	5.339800e-02	1.608055e-02
50%	1.000000	0.587764	1.173970e-01	4.944380e-02
75%	1.000000	0.847827	1.955655e-01	9.153815e-02
max	4.000000	1.000000	4.880140e-01	2.734190e-01

In [86]:
```
master_df
```

Out[86]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	dog_stage	retweet_count	favorite_count
0	892420643555336193	2017-08-01 16:23:56+00:00	This is Phineas. He's a mystical boy. Only eve...	13.0	10	Phineas	NaN	8853	39467
1	892177421306343426	2017-08-01 00:17:27+00:00	This is Tilly. She's just checking pup on you....	13.0	10	Tilly	NaN	6514	33819
2	891815181378084864	2017-07-31 00:18:03+00:00	This is Archie. He is a rare Norwegian Pouncin...	12.0	10	Archie	NaN	4328	25461
3	891689557279858688	2017-07-30 15:58:51+00:00	This is Darla. She commenced a snooze mid meal...	13.0	10	Darla	NaN	8964	42908
4	891327558926688256	2017-07-29 16:00:24+00:00	This is Franklin. He would like you to stop ca...	12.0	10	Franklin	NaN	9774	41048
...	...	...	...	...	...	...	...	...	...
1966	666049248165822465	2015-11-16 00:24:50+00:00	Here we have a 1949 1st generation vulpix. Enj...	5.0	10	NaN	NaN	41	111
1967	666044226329800704	2015-11-16 00:04:52+00:00	This is a purebred Piers Morgan. Loves to Netf...	6.0	10	NaN	NaN	147	311
1968	666033412701032449	2015-11-15 23:21:54+00:00	Here is a very happy pup. Big fan of well-main...	9.0	10	NaN	NaN	47	128
1969	666029285002620928	2015-11-15 23:05:30+00:00	This is a western brown Mitsubishi terrier. Up...	7.0	10	NaN	NaN	48	132
1970	666020888022790149	2015-11-15 22:32:08+00:00	Here we have a Japanese Irish Setter. Lost eye...	8.0	10	NaN	NaN	532	2535

In [87]:
```
image_predictions_clean
```

Out[87]:

	tweet_id	image_url	image_number	1st_prediction	1st_prediction_confidence%	is_dog_breed1	2nd_prediction	2nd_prediction_confidence%	is_dog_breed2	
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springer_spaniel	0.465074	True	collie	0.156665	True	She
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	redbone	0.506826	True	miniature_pinscher	0.074192	True	Rhode
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_shepherd	0.596461	True	malinois	0.138584	True	
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_ridgeback	0.408143	True	redbone	0.360687	True	mir
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_pinscher	0.560311	True	Rottweiler	0.243682	True	
...	...	...	...	...	...	...	...	...	...	
1966	891327558926688256	https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg	2	basset	0.555712	True	English_springer	0.225770	True	
1967	891689557279858688	https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg	1	paper_towel	0.170278	False	Labrador_retriever	0.168086	True	
1968	891815181378084864	https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg	1	Chihuahua	0.716012	True	malamute	0.078253	True	
1969	892177421306343426	https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg	1	Chihuahua	0.323581	True	Pekinese	0.090647	True	

Most common dog name is Charlie, people like this name

In [88]:
```
master_df['name'].value_counts()
```

Out[88]:
```
Charlie    11
Cooper     10
Lucy       10
Oliver     10
Penny       9
           ..
Chesney     1
Cupid       1
Bruiser     1
Dotsy       1
Dixie       1
Name: name, Length: 936, dtype: int64
```

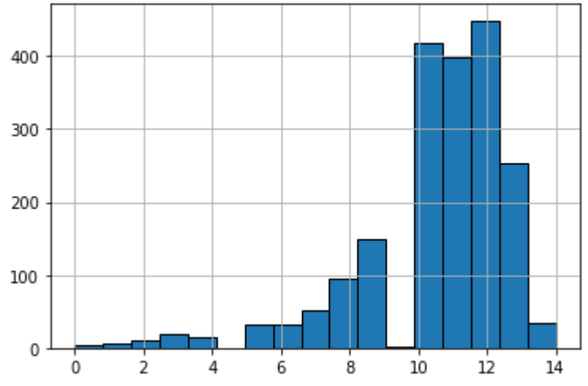
In [89]:
```
master_df['name'].mode()[0]
```

Out[89]:
```
'Charlie'
```

Most common numerator rating is 12

In [90]:
```
master_df['rating_numerator'].hist(edgecolor='black',bins=17)
```

Out[90]:
```
<AxesSubplot:>
```



In [91]:
```
master_df['rating_numerator'].mode()
```

Out[91]:
```
0    12.0
dtype: float64
```

2017 is least year when it comes to number of tweets, tweets decreases with time but the audince interaction (increaes)

that means this page is getting popular and have quality content

In [92]:
```
# creating new column having info about the tweet year
master_df['year'] = master_df['timestamp'].dt.year
```

In [93]:
```
master_df
```

Out[93]:

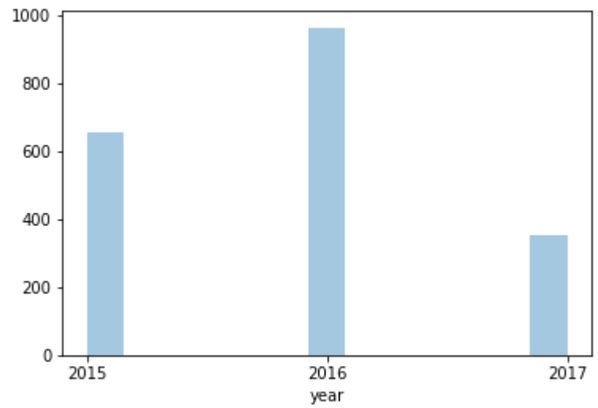
	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	dog_stage	retweet_count	favorite_count	year
0	892420643555336193	2017-08-01 16:23:56+00:00	This is Phineas. He's a mystical boy. Only eve...	13.0	10	Phineas	NaN	8853	39467	2017
1	892177421306343426	2017-08-01 00:17:27+00:00	This is Tilly. She's just checking pup on you....	13.0	10	Tilly	NaN	6514	33819	2017
2	891815181378084864	2017-07-31 00:18:03+00:00	This is Archie. He is a rare Norwegian Pouncin...	12.0	10	Archie	NaN	4328	25461	2017
3	891689557279858688	2017-07-30 15:58:51+00:00	This is Darla. She commenced a snooze mid meal...	13.0	10	Darla	NaN	8964	42908	2017
4	891327558926688256	2017-07-29 16:00:24+00:00	This is Franklin. He would like you to stop ca...	12.0	10	Franklin	NaN	9774	41048	2017
...	...	...	...	...	...	...	...	...	...	...
1966	666049248165822465	2015-11-16 00:24:50+00:00	Here we have a 1949 1st generation vulpix. Enj...	5.0	10	NaN	NaN	41	111	2015
1967	666044226329800704	2015-11-16 00:04:52+00:00	This is a purebred Piers Morgan. Loves to Netf...	6.0	10	NaN	NaN	147	311	2015
1968	666033412701032449	2015-11-15 23:21:54+00:00	Here is a very happy pup. Big fan of well-main...	9.0	10	NaN	NaN	47	128	2015
1969	666029285002620928	2015-11-15 23:05:30+00:00	This is a western brown Mitsubishi terrier. Up...	7.0	10	NaN	NaN	48	132	2015
1970	666020888022790149	2015-11-15 22:32:08+00:00	Here we have a Japanese Irish Setter. Lost eye...	8.0	10	NaN	NaN	532	2535	2015

In [94]:
```
master_df['year'].value_counts()
```

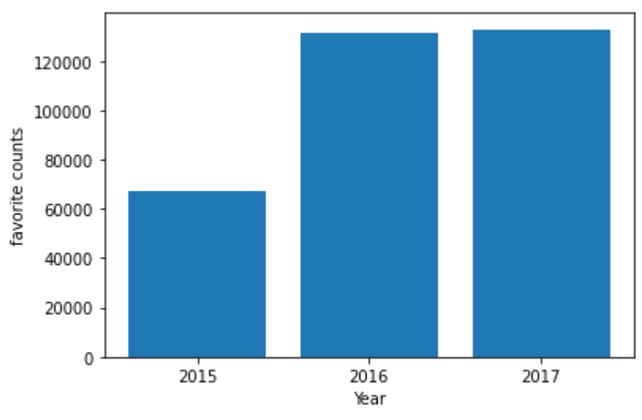
Out[94]:
```
2016    962
2015    655
2017    354
Name: year, dtype: int64
```

In [95]:
```
#2017 is least year when it comes to number of tweets
# tweets decreases with time
sns.distplot(master_df['year'],kde=False);
plt.xticks(range(2015,2018,1));
```

```
c:\users\eslam\appdata\local\programs\python\python37\lib\site-packages\seaborn\distributions.py:2557: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```



In [96]:
```
# the audience interaction (favorite counts) increases with time
plt.bar(master_df['year'], master_df['favorite_count']);
plt.xlabel('Year');
plt.ylabel('favorite counts');
plt.xticks(range(2015,2018,1));
```



---

as rating increases, favorite counts incrases, meaning the audience trust the page ratings

In [97]:
```
# as rating increases, favorite counts incrases, meaning the audience trust the page ratings
plt.scatter(master_df['rating_numerator'], master_df['favorite_count']);
plt.xlabel('numerator ratings');
plt.ylabel('favorite counts');
```