Project: Wrangle and Analyze Data by Z. McLaughlin

Gather

In [1]:
```
# Note:
#  The api access info has been removed
#  There is a variable switch = "NO" so that that the api query portion
 of the code will not run
#  To run the twitter api query set switch = "YES" and enter the access
 keys.
```

In [2]:
```
# import statements for all of the packages needed

# Required libraries
import pandas as pd
import numpy as np
import tweepy
import json
import requests

# Other libraries that will be needed
import os
from bs4 import BeautifulSoup
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime
% matplotlib inline
```

In [3]:
```
# Reading in the csv file provided:  twitter-archive-enhanced.csv

df_tae = pd.read_csv('./data/twitter-archive-enhanced.csv')
df_tae.head()
```

Out[3]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http:/ r...
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http:/ r...
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http:/ r...
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="http:/ r...
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href="http:/ r...

In [4]:
```
# Downloading the file image_predictions.tsv

url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad
_image-predictions/image-predictions.tsv'
r=requests.get(url)
open('./data/image-predictions.tsv', 'wb').write(r.content)

# Reading in the tsv file for cleaning

df_tip = pd.read_csv('./data/image-predictions.tsv',sep='\t')
df_tip.head()
```

Out[4]:

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

In [5]:
```
# Setting up switch so it's possible run the entire notebook without
# re-downloading tweepy info.  Set to "NO" if content already downloade
d.
# test

# Switch is set to 'NO' so that the code will run without the access key
s.

switch = 'NO'

if switch == 'YES':

    #Removing old files with api info
    if os.path.isfile("./data/errors.txt"):
        os.remove("./data/errors.txt")
    if os.path.isfile("./data/tweet_json.txt"):
        os.remove("./data/tweet_json.txt")
```

In [6]:
```
# Setup authorization for tweepy - authorization keys will need to be ad
ded to run the notebook

if switch == 'YES':

    consumer_key = ''
    consumer_secret = ''
    access_token = ''
    access_secret = ''

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

    api = tweepy.API(auth)
```

In [7]:
```
# Downloading json via twitter api to text file:  tweet_json.txt

if switch == "YES":
    tweet_errors = []
    tweets = df_tae.tweet_id

    for tweet in tweets:
        try:
            # Getting the tweepy json
            tweepy_info = api.get_status(tweet,wait_on_rate_limit = True
, wait_on_rate_limit_notify = True,tweet_mode='extended')

            # Saving tweepy json off to tweet_json.txt file
            tweepy_file = open("./data/tweet_json.txt","a")
            json.dump(tweepy_info._json, tweepy_file)
            tweepy_file.write("\n")
            tweepy_file.close()

        # Saving errors to a file
        except Exception as e:
            tweet_errors.append(str(tweet))
            print("Error")
            print("This tweet had a problem: " + str(tweet))
            error_file = open("./data/errors.txt","a")
            error_file.write(str(tweet) + ':  ' + str(e) + '\n')
            error_file.close()
            print(e)
```

In [8]:
```
# Reads the contents of tweet_json.txt and extracts the retweet and like
 counts.
# Creates a new dataframe wiht the info and saves it off to: tweet_extra
_info.csv

tweet_id = []
retweet_count = [] # retweet_count
like_count = [] # favourites_count

# Opens the files with json info and reads line by line and extracts inf
o.

with open('./data/tweet_json.txt') as f:
    lines = f.readlines()
for line in lines:
    line=line.rstrip()
    data=json.loads(line)
    tid= data['id']
    rc = data['retweet_count']
    lc = data['favorite_count']
    tweet_id.append(tid)
    retweet_count.append(rc)
    like_count.append(lc)

# Creating new dataframe

columns = ['tweet_id','retweet_count','like_count']
df_ti = pd.DataFrame(columns=columns)

df_ti['tweet_id'] = tweet_id
df_ti['retweet_count'] = retweet_count
df_ti['like_count'] = like_count

# Writing dataframe to csv file

df_ti.to_csv('./data/tweet_extra_info.csv',index=False)

print('The following file has been saved to disk:  tweet_extra_info.csv'
)
```

```
The following file has been saved to disk:  tweet_extra_info.csv
```

In [9]:
```
# Reading in the info file for further use

df_info = pd.read_csv('./data/tweet_extra_info.csv')
df_info.head()

df_info.describe()
```

Out[9]:

	tweet_id	retweet_count	like_count
count	2.344000e+03	2344.000000	2344.000000
mean	7.422890e+17	3010.996587	8037.963311
std	6.835057e+16	5009.803282	12097.967741
min	6.660209e+17	0.000000	0.000000
25%	6.783704e+17	603.750000	1398.500000
50%	7.187854e+17	1403.000000	3524.500000
75%	7.986989e+17	3504.500000	9937.750000
max	8.924206e+17	77042.000000	142895.000000

Assess

The following areas for cleaning were identified:

Project requirement: Assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.

Quality: Issues with content/dirty data:

1. twitter-archive-enhanced.csv: Not all tweets are still valid - While querying the api several of the tweet ids came back with an error. - Cleaned
2. twitter-archive-enhanced.csv: Instead of using null for no content there is "None" under doggo, floofer, pupper,puppo. - Cleaned
3. twitter-archive-enhanced.csv: Contains retweets and replies when we only want original tweets - Cleaned
4. twitter-archive-enhanced.csv: Only want tweets that have images (Some have no images or have videos) - Cleaned
5. twitter-archive-enhanced.csv: Rating denominator min is 0 and max is 170 when should be 10. - Cleaned
6. twitter-archive-enhanced.csv: Source contains whole HTML string instead of just simple source - Cleaned
7. twitter-archive-enhanced.csv: timestamp & retweeted_status_timestamp are not dates - Cleaned (timestamp only)
8. twitter-archive-enhanced.csv: Dog names are sometimes showing as the, a, an - Cleaned

Tidiness: Issues with structure that prevent easy analysis. Messy data.

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

1. twitter-archive-enhanced.csv: Text field includes short links already expanded in another column. - Cleaned
2. Three tables that could all be in one table.

Used a combination of exploring the data manually and programmatically.

- Used google docs to view the tables in csv to get general view of the tables.
- While cleaning each item further investigated other areas leading to the list above.
- Assess order and the cleaning steps were different.

In [10]:
```
df_tae.head()
```

Out[10]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http:/ r...
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http:/ r...
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http:/ r...
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="http:/ r...
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href="http:/ r...

In [11]:
```
df_tae.info()
df_tae.describe()
```

```
<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
```

Out[11]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	retweeted_status_id	retw
count	2.356000e+03	7.800000e+01	7.800000e+01	1.810000e+02	1.81
mean	7.427716e+17	7.455079e+17	2.014171e+16	7.720400e+17	1.24
std	6.856705e+16	7.582492e+16	1.252797e+17	6.236928e+16	9.59
min	6.660209e+17	6.658147e+17	1.185634e+07	6.661041e+17	7.83
25%	6.783989e+17	6.757419e+17	3.086374e+08	7.186315e+17	4.19
50%	7.196279e+17	7.038708e+17	4.196984e+09	7.804657e+17	4.19
75%	7.993373e+17	8.257804e+17	4.196984e+09	8.203146e+17	4.19
max	8.924206e+17	8.862664e+17	8.405479e+17	8.874740e+17	7.87

In [12]:
```
df_tip.head()
```

Out[12]:

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

In [13]:
```
df_tip.info()
df_tip.describe()
```

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

Out[13]:

	tweet_id	img_num	p1_conf	p2_conf	p3_conf
count	2.075000e+03	2075.000000	2075.000000	2.075000e+03	2.075000e+03
mean	7.384514e+17	1.203855	0.594548	1.345886e-01	6.032417e-02
std	6.785203e+16	0.561875	0.271174	1.006657e-01	5.090593e-02
min	6.660209e+17	1.000000	0.044333	1.011300e-08	1.740170e-10
25%	6.764835e+17	1.000000	0.364412	5.388625e-02	1.622240e-02
50%	7.119988e+17	1.000000	0.588230	1.181810e-01	4.944380e-02
75%	7.932034e+17	1.000000	0.843855	1.955655e-01	9.180755e-02
max	8.924206e+17	4.000000	1.000000	4.880140e-01	2.734190e-01

In [14]:
```
df_info.head()
```

Out[14]:

	tweet_id	retweet_count	like_count
0	892420643555336193	8553	38673
1	892177421306343426	6287	33127
2	891815181378084864	4166	24943
3	891689557279858688	8675	42044
4	891327558926688256	9441	40193

In [15]:
```
df_info.info()
df_info.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2344 entries, 0 to 2343
Data columns (total 3 columns):
tweet_id         2344 non-null int64
retweet_count    2344 non-null int64
like_count       2344 non-null int64
dtypes: int64(3)
memory usage: 55.0 KB
```

Out[15]:

	tweet_id	retweet_count	like_count
count	2.344000e+03	2344.000000	2344.000000
mean	7.422890e+17	3010.996587	8037.963311
std	6.835057e+16	5009.803282	12097.967741
min	6.660209e+17	0.000000	0.000000
25%	6.783704e+17	603.750000	1398.500000
50%	7.187854e+17	1403.000000	3524.500000
75%	7.986989e+17	3504.500000	9937.750000
max	8.924206e+17	77042.000000	142895.000000

Clean

In [16]:
```
# Make a copy of the data

df_tae_clean = df_tae.copy()
df_tae_clean.head()
```

Out[16]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http:/ r...
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http:/ r...
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http:/ r...
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="http:/ r...
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href="http:/ r...

Quality Improvement - Remove replies and retweets

In [17]:
```
# twitter-archive-enhanced.csv: Contains retweets and replies when we on
ly want original tweets
# Remove rows that have content in 'in_reply_to_status_id'

# Idenfity reply rows and review

reply_info = df_tae_clean[(df_tae_clean['in_reply_to_status_id'].notnull
())]
reply_info.info()

# Cut the rows reply rows that are not necessary

df_tae_clean = df_tae_clean[(df_tae_clean['in_reply_to_status_id'].isnul
l())]
df_tae_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 78 entries, 30 to 2298
Data columns (total 17 columns):
tweet_id                      78 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     78 non-null object
source                        78 non-null object
text                          78 non-null object
retweeted_status_id           0 non-null float64
retweeted_status_user_id      0 non-null float64
retweeted_status_timestamp    0 non-null object
expanded_urls                 23 non-null object
rating_numerator              78 non-null int64
rating_denominator            78 non-null int64
name                          78 non-null object
doggo                         78 non-null object
floofer                       78 non-null object
pupper                        78 non-null object
puppo                         78 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 11.0+ KB
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2278 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2278 non-null int64
in_reply_to_status_id         0 non-null float64
in_reply_to_user_id           0 non-null float64
timestamp                     2278 non-null object
source                        2278 non-null object
text                          2278 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                 2274 non-null object
rating_numerator              2278 non-null int64
rating_denominator            2278 non-null int64
name                          2278 non-null object
doggo                         2278 non-null object
floofer                       2278 non-null object
pupper                        2278 non-null object
puppo                         2278 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 320.3+ KB
```

In [18]:
```
# Remove rows that have content in 'retweeted_status_id'

# Idenfity retweet rows and review

retweet_info = df_tae_clean[(df_tae_clean['retweeted_status_id'].notnull
())]
retweet_info.info()

# Cut the rows reply rows that are not necessary

df_tae_clean = df_tae_clean[(df_tae_clean['retweeted_status_id'].isnull
())]
df_tae_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 181 entries, 19 to 2260
Data columns (total 17 columns):
tweet_id                      181 non-null int64
in_reply_to_status_id         0 non-null float64
in_reply_to_user_id           0 non-null float64
timestamp                     181 non-null object
source                        181 non-null object
text                          181 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                 180 non-null object
rating_numerator              181 non-null int64
rating_denominator            181 non-null int64
name                          181 non-null object
doggo                         181 non-null object
floofer                       181 non-null object
pupper                        181 non-null object
puppo                         181 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 25.5+ KB
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2097 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2097 non-null int64
in_reply_to_status_id         0 non-null float64
in_reply_to_user_id           0 non-null float64
timestamp                     2097 non-null object
source                        2097 non-null object
text                          2097 non-null object
retweeted_status_id           0 non-null float64
retweeted_status_user_id      0 non-null float64
retweeted_status_timestamp    0 non-null object
expanded_urls                 2094 non-null object
rating_numerator              2097 non-null int64
rating_denominator            2097 non-null int64
name                          2097 non-null object
doggo                         2097 non-null object
floofer                       2097 non-null object
pupper                        2097 non-null object
puppo                         2097 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 294.9+ KB
```

Tidyness Improvement - Remove columns related to replies and retweets

In [19]:
```
# Now that we've eliminated the replys and retweets we can delete the co
lumns
# that save that information.

# Columns to be deleted - in_reply_to_status_id, in_reply_to_user_id, re
tweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp

columns = ['in_reply_to_status_id', 'in_reply_to_user_id','retweeted_sta
tus_id','retweeted_status_user_id','retweeted_status_timestamp']
df_tae_clean.drop(columns, inplace=True, axis=1)
df_tae_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2097 entries, 0 to 2355
Data columns (total 12 columns):
tweet_id              2097 non-null int64
timestamp             2097 non-null object
source                2097 non-null object
text                  2097 non-null object
expanded_urls         2094 non-null object
rating_numerator      2097 non-null int64
rating_denominator    2097 non-null int64
name                  2097 non-null object
doggo                 2097 non-null object
floofer               2097 non-null object
pupper                2097 non-null object
puppo                 2097 non-null object
dtypes: int64(3), object(9)
memory usage: 213.0+ KB
```

Quality Improvement - Removing tweets that were not found using the api

In [20]:
```
# Need to remove the tweets that couldn't be found using the api

# use the error file created during the api step using tweepy
tweet_errors = pd.read_csv('./data/errors.txt', sep=":", header=None)
tweet_errors.head()
tweets_to_remove = tweet_errors[0].values

# de-duplicate
tweets_to_remove_u = []
for tweet in tweets_to_remove:
    if tweet not in tweets_to_remove_u:
        tweets_to_remove_u.append(tweet)

# Remove any rows from clean data that couldn't be found using the api
for tweet in tweets_to_remove_u:
     # A lot of the tweets that errored out were replies or retweets
     if df_tae_clean[df_tae_clean['tweet_id']== tweet].empty == False:
        df_tae_clean = df_tae_clean[df_tae_clean['tweet_id']!= tweet]

# Test that the tweet in question got removed (turned out there was only
 one.)
df_tae_clean[df_tae_clean['tweet_id']== 754011816964026368]
df_tae_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2096 entries, 0 to 2355
Data columns (total 12 columns):
tweet_id              2096 non-null int64
timestamp             2096 non-null object
source                2096 non-null object
text                  2096 non-null object
expanded_urls         2093 non-null object
rating_numerator      2096 non-null int64
rating_denominator    2096 non-null int64
name                  2096 non-null object
doggo                 2096 non-null object
floofer               2096 non-null object
pupper                2096 non-null object
puppo                 2096 non-null object
dtypes: int64(3), object(9)
memory usage: 212.9+ KB
```

Quality - Change timestamp to date format

In [21]:
```
# timestamp to datetime
df_tae_clean.timestamp = pd.to_datetime(df_tae_clean.timestamp)

# test
df_tae_clean.info()

df_tae_clean.head()

# See another possible problem area - The number of exampnded_urls does
n't match the number of tweets
# Investigate

df_tae_clean[(df_tae_clean['expanded_urls'].isnull())]
# Result:  These tweets don't have images so they will be removed via ot
her cleaning
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2096 entries, 0 to 2355
Data columns (total 12 columns):
tweet_id              2096 non-null int64
timestamp             2096 non-null datetime64[ns]
source                2096 non-null object
text                  2096 non-null object
expanded_urls         2093 non-null object
rating_numerator      2096 non-null int64
rating_denominator    2096 non-null int64
name                  2096 non-null object
doggo                 2096 non-null object
floofer               2096 non-null object
pupper                2096 non-null object
puppo                 2096 non-null object
dtypes: datetime64[ns](1), int64(3), object(8)
memory usage: 212.9+ KB
```

Out[21]:

	tweet_id	timestamp	source	text
375	828361771580813312	2017-02-05 21:56:51	<a href="http://twitter.com" rel="nofollow">Tw...	Beebop and Doobert should start a band 12/10 w...
707	785515384317313025	2016-10-10 16:20:36	<a href="http://twitter.com/download/iphone" r...	Today, 10/10, should be National Dog Rates Day
1445	696518437233913856	2016-02-08 02:18:30	<a href="http://twitter.com/download/iphone" r...	Oh my god 10/10 for every little hot dog pupper

Quality Improvement - Change None to Null for name, doggo, floofer, pupper, puppo

In [22]:
```
# Checked each column for unique values.  Replaced all 'None' in table t
o Null.

print(df_tae_clean['doggo'].unique())
print(df_tae_clean['floofer'].unique())
print(df_tae_clean['pupper'].unique())
print(df_tae_clean['puppo'].unique())

df_tae_clean.replace('None', np.nan,inplace = True)
df_tae_clean.head(50)
```

```
['None' 'doggo']
['None' 'floofer']
['None' 'pupper']
['None' 'puppo']
```

Out[22]:

	tweet_id	timestamp	source	tex
0	892420643555336193	2017-08-01 16:23:56	<a href="http://twitter.com/download/iphone" r...	This is Phineas. He's a mystical bo Only eve...
1	892177421306343426	2017-08-01 00:17:27	<a href="http://twitter.com/download/iphone" r...	This is Tilly. She's just checking pup on you....
2	891815181378084864	2017-07-31 00:18:03	<a href="http://twitter.com/download/iphone" r...	This is Archie. He i a rare Norwegian Pouncin...
3	891689557279858688	2017-07-30 15:58:51	<a href="http://twitter.com/download/iphone" r...	This is Darla She commenced a snooze mid meal...
4	891327558926688256	2017-07-29 16:00:24	<a href="http://twitter.com/download/iphone" r...	This is Franklin. He would like you to stop ca...
5	891087950875897856	2017-07-29 00:08:17	<a href="http://twitter.com/download/iphone" r...	Here we have a majestic great white breaching ..
6	890971913173991426	2017-07-28 16:27:12	<a href="http://twitter.com/download/iphone" r...	Meet Jax. He enjoys ice cream s much he gets ...
7	890729181411237888	2017-07-28 00:22:40	<a href="http://twitter.com/download/iphone" r...	When you watch your owner call another dog a g...

	tweet_id	timestamp	source	tex
8	890609185150312448	2017-07-27 16:25:51	<a href="http://twitter.com/download/iphone" r...	This is Zoey She doesn't want to be one of th...
9	890240255349198849	2017-07-26 15:59:51	<a href="http://twitter.com/download/iphone" r...	This is Cassie. She is a college pup. Studying...
10	890066081131724800	2017-07-26 00:31:25	<a href="http://twitter.com/download/iphone" r...	This is Koda He is a South Australian decksha...
11	889880896479866881	2017-07-25 16:11:53	<a href="http://twitter.com/download/iphone" r...	This is Bruno. He is a service shark. Only get...
12	889665388333682689	2017-07-25 01:55:32	<a href="http://twitter.com/download/iphone" r...	Here's a puppo that seems to be on the fence a...
13	889638837579907072	2017-07-25 00:10:02	<a href="http://twitter.com/download/iphone" r...	This is Ted. He does his best. Sometimes that'...
14	889531135344209921	2017-07-24 17:02:04	<a href="http://twitter.com/download/iphone" r...	This is Stuart. He's sporting his favorite fan.
15	889278841981685760	2017-07-24 00:19:32	<a href="http://twitter.com/download/iphone" r...	This is Oliver. You'r witnessing one of his m...
16	888917238123831296	2017-07-23 00:22:39	<a href="http://twitter.com/download/iphone" r...	This is Jim. He found a fren. Taught him how t...

	tweet_id	timestamp	source	tex
17	888804989199671297	2017-07-22 16:56:37	<a href="http://twitter.com/download/iphone" r...	This is Zeke He has a new stick. Very proud o...
18	888554962724278272	2017-07-22 00:23:06	<a href="http://twitter.com/download/iphone" r...	This is Ralphus. He's powering up Attempting ...
20	888078434458587136	2017-07-20 16:49:33	<a href="http://twitter.com/download/iphone" r...	This is Gerald. He was just tol he didn't get...
21	887705289381826560	2017-07-19 16:06:48	<a href="http://twitter.com/download/iphone" r...	This is Jeffrey. He has a monopoly o the pool...
22	887517139158093824	2017-07-19 03:39:09	<a href="http://twitter.com/download/iphone" r...	I've yet to rate a Venezuelan Hover Wiener. Th..
23	887473957103951883	2017-07-19 00:47:34	<a href="http://twitter.com/download/iphone" r...	This is Canela. She attempted some fancy porch...
24	887343217045368832	2017-07-18 16:08:03	<a href="http://twitter.com/download/iphone" r...	You may no have knowr you needed to see this .
25	887101392804085760	2017-07-18 00:07:08	<a href="http://twitter.com/download/iphone" r...	This... is a Jubilant Antarctic House Bear We...

	tweet_id	timestamp	source	tex
26	886983233522544640	2017-07-17 16:17:36	<a href="http://twitter.com/download/iphone" r...	This is Maya She's very shy. Rarely leaves he...
27	886736880519319552	2017-07-16 23:58:41	<a href="http://twitter.com/download/iphone" r...	This is Mingus. He's a wonderful father to his...
28	886680336477933568	2017-07-16 20:14:00	<a href="http://twitter.com/download/iphone" r...	This is Derek. He's late for a dog meeting. 13...
29	886366144734445568	2017-07-15 23:25:31	<a href="http://twitter.com/download/iphone" r...	This is Roscoe. Another pupper falle victim t...
31	886258384151887873	2017-07-15 16:17:19	<a href="http://twitter.com/download/iphone" r...	This is Waffles. His doggles are pupside down....
33	885984800019947520	2017-07-14 22:10:11	<a href="http://twitter.com/download/iphone" r...	Viewer discretion advised. This is Jimbo. He w...
34	885528943205470208	2017-07-13 15:58:47	<a href="http://twitter.com/download/iphone" r...	This is Maisey. She fell asleep mid-excavation.
35	885518971528720385	2017-07-13 15:19:09	<a href="http://twitter.com/download/iphone" r...	I have a new hero and his name is Howard. 14/1...

	tweet_id	timestamp	source	tex
37	885167619883638784	2017-07-12 16:03:00	<a href="http://twitter.com/download/iphone" r...	Here we have a corg undercover as a malamute...
38	884925521741709313	2017-07-12 00:01:00	<a href="http://twitter.com/download/iphone" r...	This is Earl. He found a hat. Nervou about wh...
39	884876753390489601	2017-07-11 20:47:12	<a href="http://twitter.com/download/iphone" r...	This is Lola. It's her first time outside Mus...
40	884562892145688576	2017-07-11 00:00:02	<a href="http://twitter.com/download/iphone" r...	This is Kevin. He's just so happy. 13/1 what ...
41	884441805382717440	2017-07-10 15:58:53	<a href="http://twitter.com/download/iphone" r...	I present to you, Pup in Hat. Pup in Hat is gr...
42	884247878851493888	2017-07-10 03:08:17	<a href="http://twitter.com/download/iphone" r...	OMG HE DIDN'T MEAN TO HE WAS JUST TRYING A LIT...
43	884162670584377345	2017-07-09 21:29:42	<a href="http://twitter.com/download/iphone" r...	Meet Yogi. He doesn't have any important dog m...
44	883838122936631299	2017-07-09 00:00:04	<a href="http://twitter.com/download/iphone" r...	This is Noah He can't believe someone made th...

	tweet_id	timestamp	source	tex
45	883482846933004288	2017-07-08 00:28:19	<a href="http://twitter.com/download/iphone" r...	This is Bella She hopes her smile made you sm...
46	883360690899218434	2017-07-07 16:22:55	<a href="http://twitter.com/download/iphone" r...	Meet Grizzwald. He may be the floofiest floofe...
47	883117836046086144	2017-07-07 00:17:54	<a href="http://twitter.com/download/iphone" r...	Please only send dogs. We don't rate mechanics.
48	882992080364220416	2017-07-06 15:58:11	<a href="http://twitter.com/download/iphone" r...	This is Rusty. He wasn't read for the first p...
49	882762694511734784	2017-07-06 00:46:41	<a href="http://twitter.com/download/iphone" r...	This is Gus. He's quite the cheeky pupper. Alr.
50	882627270321602560	2017-07-05 15:48:34	<a href="http://twitter.com/download/iphone" r...	This is Stanley. He has his first swim lesson ...
51	882268110199369728	2017-07-04 16:01:23	<a href="http://twitter.com/download/iphone" r...	This is Alfy. You're witnessing his first wate...
52	882045870035918850	2017-07-04 01:18:17	<a href="http://twitter.com/download/iphone" r...	This is Koko Her owner, inspired by Barney, r...
53	881906580714921986	2017-07-03 16:04:48	<a href="http://twitter.com/download/iphone" r...	This is Rey. He's a Benebop Cumberfloo 12/10...

Quality Improvement - Change obvious non-name items (a, an, the) to Null for name

In [23]:
```
# Changing non-names in name column to Null (testing if first chacter is
 lower case)

df_tae_clean['name'] = df_tae_clean['name'].astype(str)
names = df_tae_clean['name'].unique()
names_to_change_to_null = []
for nam in names:
    if nam[0].islower():
        names_to_change_to_null.append(nam)
print(names_to_change_to_null)

for ncn in names_to_change_to_null:
    df_tae_clean['name'].replace(ncn, np.nan,inplace = True)

print(df_tae_clean['name'].unique())

df_tae_clean.head(100)
```

```
['nan', 'such', 'a', 'quite', 'not', 'one', 'incredibly', 'very', 'my',
'his', 'an', 'actually', 'just', 'getting', 'mad', 'this', 'unacceptabl
e', 'all', 'old', 'infuriating', 'the', 'by', 'officially', 'life', 'li
ght', 'space']
['Phineas' 'Tilly' 'Archie' 'Darla' 'Franklin' nan 'Jax' 'Zoey' 'Cassi
e'
 'Koda' 'Bruno' 'Ted' 'Stuart' 'Oliver' 'Jim' 'Zeke' 'Ralphus' 'Gerald'
 'Jeffrey' 'Canela' 'Maya' 'Mingus' 'Derek' 'Roscoe' 'Waffles' 'Jimbo'
 'Maisey' 'Earl' 'Lola' 'Kevin' 'Yogi' 'Noah' 'Bella' 'Grizzwald' 'Rust
y'
 'Gus' 'Stanley' 'Alfy' 'Koko' 'Rey' 'Gary' 'Elliot' 'Louis' 'Jesse'
 'Romeo' 'Bailey' 'Duddles' 'Jack' 'Steven' 'Beau' 'Snoopy' 'Shadow' 'E
mmy'
 'Aja' 'Penny' 'Dante' 'Nelly' 'Ginger' 'Benedict' 'Venti' 'Goose' 'Nug
get'
 'Cash' 'Jed' 'Sebastian' 'Sierra' 'Monkey' 'Harry' 'Kody' 'Lassie' 'Ro
ver'
 'Napolean' 'Boomer' 'Cody' 'Rumble' 'Clifford' 'Dewey' 'Scout' 'Gizmo'
 'Walter' 'Cooper' 'Harold' 'Shikha' 'Lili' 'Jamesy' 'Coco' 'Sammy'
 'Meatball' 'Paisley' 'Albus' 'Neptune' 'Belle' 'Quinn' 'Zooey' 'Dave'
 'Jersey' 'Hobbes' 'Burt' 'Lorenzo' 'Carl' 'Jordy' 'Milky' 'Trooper'
 'Sophie' 'Wyatt' 'Rosie' 'Thor' 'Oscar' 'Callie' 'Cermet' 'Marlee' 'Ar
ya'
 'Einstein' 'Alice' 'Rumpole' 'Benny' 'Aspen' 'Jarod' 'Wiggles' 'Genera
l'
 'Sailor' 'Iggy' 'Snoop' 'Kyle' 'Leo' 'Riley' 'Noosh' 'Odin' 'Jerry'
 'Georgie' 'Rontu' 'Cannon' 'Furzey' 'Daisy' 'Tuck' 'Barney' 'Vixen'
 'Jarvis' 'Mimosa' 'Pickles' 'Brady' 'Luna' 'Charlie' 'Margo' 'Sadie'
 'Hank' 'Tycho' 'Indie' 'Winnie' 'George' 'Bentley' 'Max' 'Dawn' 'Maddi
e'
 'Monty' 'Sojourner' 'Winston' 'Odie' 'Arlo' 'Vincent' 'Lucy' 'Clark'
 'Mookie' 'Meera' 'Ava' 'Eli' 'Ash' 'Tucker' 'Tobi' 'Chester' 'Wilson'
 'Sunshine' 'Lipton' 'Bronte' 'Poppy' 'Gidget' 'Rhino' 'Willow' 'Orion'
 'Eevee' 'Smiley' 'Miguel' 'Emanuel' 'Kuyu' 'Dutch' 'Pete' 'Scooter'
 'Reggie' 'Lilly' 'Samson' 'Mia' 'Astrid' 'Malcolm' 'Dexter' 'Alfie'
 'Fiona' 'Mutt' 'Bear' 'Doobert' 'Beebop' 'Alexander' 'Sailer' 'Brutus'
 'Kona' 'Boots' 'Ralphie' 'Loki' 'Cupid' 'Pawnd' 'Pilot' 'Ike' 'Mo' 'To
by'
 'Sweet' 'Pablo' 'Nala' 'Crawford' 'Gabe' 'Jimison' 'Duchess' 'Harlso'
 'Sundance' 'Luca' 'Flash' 'Sunny' 'Howie' 'Jazzy' 'Anna' 'Finn' 'Bo'
 'Wafer' 'Tom' 'Florence' 'Autumn' 'Buddy' 'Dido' 'Eugene' 'Ken' 'Strud
el'
 'Tebow' 'Chloe' 'Timber' 'Binky' 'Moose' 'Dudley' 'Comet' 'Akumi' 'Tit
an'
 'Olivia' 'Alf' 'Oshie' 'Chubbs' 'Sky' 'Atlas' 'Eleanor' 'Layla' 'Rock
y'
 'Baron' 'Tyr' 'Bauer' 'Swagger' 'Brandi' 'Mary' 'Moe' 'Halo' 'Augie'
 'Craig' 'Sam' 'Hunter' 'Pavlov' 'Phil' 'Kyro' 'Wallace' 'Ito' 'Seamus'
 'Ollie' 'Stephan' 'Lennon' 'Major' 'Duke' 'Sansa' 'Shooter' 'Django'
 'Diogi' 'Sonny' 'Marley' 'Severus' 'Ronnie' 'Milo' 'Bones' 'Mauve' 'Ch
ef'
 'Doc' 'Peaches' 'Sobe' 'Longfellow' 'Mister' 'Iroh' 'Pancake' 'Snicku'
 'Ruby' 'Brody' 'Mack' 'Nimbus' 'Laika' 'Maximus' 'Dobby' 'Moreton' 'Ju
no'
 'Maude' 'Lily' 'Newt' 'Benji' 'Nida' 'Robin' 'Monster' 'BeBe' 'Remus'
 'Levi' 'Mabel' 'Misty' 'Betty' 'Mosby' 'Maggie' 'Bruce' 'Happy' 'Ralph
y'
```

```
 'Brownie' 'Rizzy' 'Stella' 'Butter' 'Frank' 'Tonks' 'Lincoln' 'Rory'
 'Logan' 'Dale' 'Rizzo' 'Arnie' 'Mattie' 'Pinot' 'Dallas' 'Hero' 'Frank
ie'
 'Stormy' 'Reginald' 'Balto' 'Mairi' 'Loomis' 'Godi' 'Cali' 'Deacon'
 'Timmy' 'Sampson' 'Chipson' 'Combo' 'Oakley' 'Dash' 'Hercules' 'Jay'
'Mya'
 'Strider' 'Wesley' 'Solomon' 'Huck' 'O' 'Blue' 'Anakin' 'Finley'
 'Sprinkles' 'Heinrich' 'Shakespeare' 'Chelsea' 'Bungalo' 'Chip' 'Grey'
 'Roosevelt' 'Willem' 'Davey' 'Dakota' 'Fizz' 'Dixie' 'Al' 'Jackson'
 'Carbon' 'Klein' 'DonDon' 'Kirby' 'Lou' 'Chevy' 'Tito' 'Philbert' 'Lou
ie'
 'Rupert' 'Rufus' 'Brudge' 'Shadoe' 'Angel' 'Brat' 'Tove' 'Gromit' 'Aub
ie'
 'Kota' 'Leela' 'Glenn' 'Shelby' 'Sephie' 'Bonaparte' 'Albert' 'Wishes'
 'Rose' 'Theo' 'Rocco' 'Fido' 'Emma' 'Spencer' 'Lilli' 'Boston'
 'Brandonald' 'Corey' 'Leonard' 'Beckham' 'Devón' 'Gert' 'Watson' 'Keit
h'
 'Dex' 'Ace' 'Tayzie' 'Grizzie' 'Fred' 'Gilbert' 'Meyer' 'Zoe' 'Stewie'
 'Calvin' 'Lilah' 'Spanky' 'Jameson' 'Piper' 'Atticus' 'Blu' 'Dietrich'
 'Divine' 'Tripp' 'Cora' 'Huxley' 'Keurig' 'Bookstore' 'Linus' 'Abby'
 'Shiloh' 'Gustav' 'Arlen' 'Percy' 'Lenox' 'Sugar' 'Harvey' 'Blanket'
 'Geno' 'Stark' 'Beya' 'Kilo' 'Kayla' 'Maxaroni' 'Bell' 'Doug' 'Edmund'
 'Aqua' 'Theodore' 'Baloo' 'Chase' 'Nollie' 'Rorie' 'Simba' 'Charles'
 'Bayley' 'Axel' 'Storkson' 'Remy' 'Chadrick' 'Kellogg' 'Buckley' 'Livv
ie'
 'Terry' 'Hermione' 'Ralpher' 'Aldrick' 'Larry' 'Rooney' 'Crystal' 'Ziv
a'
 'Stefan' 'Pupcasso' 'Puff' 'Flurpson' 'Coleman' 'Enchilada' 'Raymond'
 'Rueben' 'Cilantro' 'Karll' 'Sprout' 'Blitz' 'Bloop' 'Colby' 'Lillie'
 'Ashleigh' 'Kreggory' 'Sarge' 'Luther' 'Ivar' 'Jangle' 'Schnitzel' 'Pa
nda'
 'Berkeley' 'Ralphé' 'Charleson' 'Clyde' 'Harnold' 'Sid' 'Pippa' 'Otis'
 'Carper' 'Bowie' 'Alexanderson' 'Suki' 'Barclay' 'Skittle' 'Ebby' 'Flá
vio'
 'Smokey' 'Link' 'Jennifur' 'Ozzy' 'Bluebert' 'Stephanus' 'Bubbles' 'Ze
us'
 'Bertson' 'Nico' 'Michelangelope' 'Siba' 'Calbert' 'Curtis' 'Travis'
 'Thumas' 'Kanu' 'Lance' 'Opie' 'Stubert' 'Kane' 'Olive' 'Chuckles'
 'Staniel' 'Sora' 'Beemo' 'Gunner' 'Lacy' 'Tater' 'Olaf' 'Cecil' 'Vinc
e'
 'Karma' 'Billy' 'Walker' 'Rodney' 'Klevin' 'Malikai' 'Bobble' 'River'
 'Jebberson' 'Remington' 'Farfle' 'Jiminus' 'Harper' 'Clarkus' 'Finnegu
s'
 'Cupcake' 'Kathmandu' 'Ellie' 'Katie' 'Kara' 'Adele' 'Zara' 'Ambrose'
 'Jimothy' 'Bode' 'Terrenth' 'Reese' 'Chesterson' 'Lucia' 'Bisquick'
 'Ralphson' 'Socks' 'Rambo' 'Rudy' 'Fiji' 'Rilo' 'Bilbo' 'Coopson' 'Yod
a'
 'Millie' 'Chet' 'Crouton' 'Daniel' 'Kaia' 'Murphy' 'Dotsy' 'Eazy' 'Coo
ps'
 'Fillup' 'Miley' 'Charl' 'Reagan' 'Yukon' 'CeCe' 'Cuddles' 'Claude'
 'Jessiga' 'Carter' 'Ole' 'Pherb' 'Blipson' 'Reptar' 'Trevith' 'Berb'
'Bob'
 'Colin' 'Brian' 'Oliviér' 'Grady' 'Kobe' 'Freddery' 'Bodie' 'Dunkin'
 'Wally' 'Tupawc' 'Amber' 'Herschel' 'Edgar' 'Teddy' 'Kingsley' 'Brockl
y'
 'Richie' 'Molly' 'Vinscent' 'Cedrick' 'Hazel' 'Lolo' 'Eriq' 'Phred'
 'Oddie' 'Maxwell' 'Geoff' 'Covach' 'Durg' 'Fynn' 'Ricky' 'Herald' 'Luc
```

```
ky'
 'Ferg' 'Trip' 'Clarence' 'Hamrick' 'Brad' 'Pubert' 'Frönq' 'Derby'
 'Lizzie' 'Ember' 'Blakely' 'Opal' 'Marq' 'Kramer' 'Barry' 'Tyrone'
 'Gordon' 'Baxter' 'Mona' 'Horace' 'Crimson' 'Birf' 'Hammond' 'Lorelei'
 'Marty' 'Brooks' 'Petrick' 'Hubertson' 'Gerbald' 'Oreo' 'Bruiser' 'Per
ry'
 'Bobby' 'Jeph' 'Obi' 'Tino' 'Kulet' 'Sweets' 'Lupe' 'Tiger' 'Jiminy'
 'Griffin' 'Banjo' 'Brandy' 'Lulu' 'Darrel' 'Taco' 'Joey' 'Patrick' 'Kr
eg'
 'Todo' 'Tess' 'Ulysses' 'Toffee' 'Apollo' 'Carly' 'Asher' 'Glacier'
 'Chuck' 'Champ' 'Ozzie' 'Griswold' 'Cheesy' 'Moofasa' 'Hector' 'Goliat
h'
 'Kawhi' 'Emmie' 'Penelope' 'Willie' 'Rinna' 'Mike' 'William' 'Dwight'
 'Evy' 'Hurley' 'Rubio' 'Chompsky' 'Rascal' 'Linda' 'Tug' 'Tango' 'Griz
z'
 'Jerome' 'Crumpet' 'Jessifer' 'Izzy' 'Ralph' 'Sandy' 'Humphrey' 'Tass
y'
 'Juckson' 'Chuq' 'Tyrus' 'Karl' 'Godzilla' 'Vinnie' 'Kenneth' 'Herm'
 'Bert' 'Striker' 'Donny' 'Pepper' 'Bernie' 'Buddah' 'Lenny' 'Arnold'
 'Zuzu' 'Mollie' 'Laela' 'Tedders' 'Superpup' 'Rufio' 'Jeb' 'Rodman'
 'Jonah' 'Chesney' 'Kenny' 'Henry' 'Bobbay' 'Mitch' 'Kaiya' 'Acro' 'Aid
en'
 'Obie' 'Dot' 'Shnuggles' 'Kendall' 'Jeffri' 'Steve' 'Eve' 'Mac' 'Fletc
her'
 'Kenzie' 'Pumpkin' 'Schnozz' 'Gustaf' 'Cheryl' 'Ed' 'Leonidas' 'Norma
n'
 'Caryl' 'Scott' 'Taz' 'Darby' 'Jackie' 'Jazz' 'Franq' 'Pippin' 'Rolf'
 'Snickers' 'Ridley' 'Cal' 'Bradley' 'Bubba' 'Tuco' 'Patch' 'Mojo' 'Bat
dog'
 'Dylan' 'Mark' 'JD' 'Alejandro' 'Scruffers' 'Pip' 'Julius' 'Tanner'
 'Sparky' 'Anthony' 'Holly' 'Jett' 'Amy' 'Sage' 'Andy' 'Mason' 'Trigge
r'
 'Antony' 'Creg' 'Traviss' 'Gin' 'Jeffrie' 'Danny' 'Ester' 'Pluto' 'Blo
o'
 'Edd' 'Paull' 'Willy' 'Herb' 'Damon' 'Peanut' 'Nigel' 'Butters' 'Sandr
a'
 'Fabio' 'Randall' 'Liam' 'Tommy' 'Ben' 'Raphael' 'Julio' 'Andru' 'Kloe
y'
 'Shawwn' 'Skye' 'Kollin' 'Ronduh' 'Billl' 'Saydee' 'Dug' 'Sully' 'Kir
k'
 'Ralf' 'Clarq' 'Jaspers' 'Samsom' 'Terrance' 'Harrison' 'Chaz' 'Jerem
y'
 'Jaycob' 'Lambeau' 'Ruffles' 'Amélie' 'Bobb' 'Banditt' 'Kevon' 'Winifr
ed'
 'Hanz' 'Churlie' 'Zeek' 'Timofy' 'Maks' 'Jomathan' 'Kallie' 'Marvin'
 'Spark' 'Gòrdón' 'Jo' 'DayZ' 'Jareld' 'Torque' 'Ron' 'Skittles'
 'Cleopatricia' 'Erik' 'Stu' 'Tedrick' 'Shaggy' 'Filup' 'Kial' 'Naphani
el'
 'Dook' 'Hall' 'Philippe' 'Biden' 'Fwed' 'Genevieve' 'Joshwa' 'Timison'
 'Bradlay' 'Pipsy' 'Clybe' 'Keet' 'Carll' 'Jockson' 'Josep' 'Lugan'
 'Christoper']
```

Out[23]:

	tweet_id	timestamp	source	
0	892420643555336193	2017-08-01 16:23:56	<a href="http://twitter.com/download/iphone" r...	This is Phi mystical b
1	892177421306343426	2017-08-01 00:17:27	<a href="http://twitter.com/download/iphone" r...	This is Tilly checking p
2	891815181378084864	2017-07-31 00:18:03	<a href="http://twitter.com/download/iphone" r...	This is Arc rare Norwe Pouncin...
3	891689557279858688	2017-07-30 15:58:51	<a href="http://twitter.com/download/iphone" r...	This is Dar commenc mid meal..
4	891327558926688256	2017-07-29 16:00:24	<a href="http://twitter.com/download/iphone" r...	This is Fra like you to
5	891087950875897856	2017-07-29 00:08:17	<a href="http://twitter.com/download/iphone" r...	Here we h great white
6	890971913173991426	2017-07-28 16:27:12	<a href="http://twitter.com/download/iphone" r...	Meet Jax. cream so ...
7	890729181411237888	2017-07-28 00:22:40	<a href="http://twitter.com/download/iphone" r...	When you owner call g...
8	890609185150312448	2017-07-27 16:25:51	<a href="http://twitter.com/download/iphone" r...	This is Zoe want to be
9	890240255349198849	2017-07-26 15:59:51	<a href="http://twitter.com/download/iphone" r...	This is Cas college pu
10	890006608113172480	2017-07-26 00:31:25	<a href="http://twitter.com/download/iphone" r...	This is Koo South Aus decksha...
11	889880896479866881	2017-07-25 16:11:53	<a href="http://twitter.com/download/iphone" r...	This is Bru service sh
12	889665388333682689	2017-07-25 01:55:32	<a href="http://twitter.com/download/iphone" r...	Here's a p seems to b a...

	tweet_id	timestamp	source	
13	889638837579907072	2017-07-25 00:10:02	<a href="http://twitter.com/download/iphone" r...	This is Ted best. Som
14	889531135344209921	2017-07-24 17:02:04	<a href="http://twitter.com/download/iphone" r...	This is Stu sporting h
15	889278841981685760	2017-07-24 00:19:32	<a href="http://twitter.com/download/iphone" r...	This is Oliv witnessing
16	888917238123831296	2017-07-23 00:22:39	<a href="http://twitter.com/download/iphone" r...	This is Jim fren. Taugh
17	888804989199671297	2017-07-22 16:56:37	<a href="http://twitter.com/download/iphone" r...	This is Zek new stick.
18	888554962724278272	2017-07-22 00:23:06	<a href="http://twitter.com/download/iphone" r...	This is Ral powering ...
20	888078434458587136	2017-07-20 16:49:33	<a href="http://twitter.com/download/iphone" r...	This is Ger just told he
21	887705289381826560	2017-07-19 16:06:48	<a href="http://twitter.com/download/iphone" r...	This is Jef monopoly
22	887517139158093824	2017-07-19 03:39:09	<a href="http://twitter.com/download/iphone" r...	I've yet to Venezuela Wiener. Th
23	887473957103951883	2017-07-19 00:47:34	<a href="http://twitter.com/download/iphone" r...	This is Car attempted porch...
24	887343217045368832	2017-07-18 16:08:03	<a href="http://twitter.com/download/iphone" r...	You may n you neede
25	887101392804085760	2017-07-18 00:07:08	<a href="http://twitter.com/download/iphone" r...	This... is a Antarctic H We...
26	886983233522544640	2017-07-17 16:17:36	<a href="http://twitter.com/download/iphone" r...	This is Ma shy. Rarely

	tweet_id	timestamp	source	
27	886736880519319552	2017-07-16 23:58:41	<a href="http://twitter.com/download/iphone" r...	This is Mir wonderful
28	886680336477933568	2017-07-16 20:14:00	<a href="http://twitter.com/download/iphone" r...	This is Der for a dog r
29	886366144734445568	2017-07-15 23:25:31	<a href="http://twitter.com/download/iphone" r...	This is Ros pupper fal
31	886258384151887873	2017-07-15 16:17:19	<a href="http://twitter.com/download/iphone" r...	This is Wa doggles ar down....
...	...	...	...	...
80	877316821321428993	2017-06-21 00:06:44	<a href="http://twitter.com/download/iphone" r...	Meet Dant wasn't a fa
81	877201837425926144	2017-06-20 16:29:50	<a href="http://twitter.com/download/iphone" r...	This is Nel graduated dogtorate.
82	876838120628539392	2017-06-19 16:24:33	<a href="http://twitter.com/download/iphone" r...	This is Gin having a ru To...
83	876537666061221889	2017-06-18 20:30:39	<a href="http://twitter.com/download/iphone" r...	I can say v pupmost c the...
84	876484053909872640	2017-06-18 16:57:37	<a href="http://twitter.com/download/iphone" r...	This is Ber wants to tl th...
85	876120275196170240	2017-06-17 16:52:05	<a href="http://twitter.com/download/iphone" r...	Meet Vent caffeinated
86	875747767867523072	2017-06-16 16:11:53	<a href="http://twitter.com/download/iphone" r...	This is Goo womanize h*c...
87	875144289856114688	2017-06-15 00:13:52	<a href="http://twitter.com/download/iphone" r...	Meet Nugg Nugget too bone....
88	875097192612077568	2017-06-14 21:06:43	<a href="http://twitter.com/download/iphone" r...	You'll get y when that man...

	tweet_id	timestamp	source	
89	875021211251597312	2017-06-14 16:04:48	<a href="http://twitter.com/download/iphone" r...	Guys pleas sending pi any ...
90	874680097055178752	2017-06-13 17:29:20	<a href="http://twitter.com/download/iphone" r...	Meet Cash acquired a go...
92	874296783580663808	2017-06-12 16:06:11	<a href="http://twitter.com/download/iphone" r...	This is Jec the fancies
93	874057562936811520	2017-06-12 00:15:36	<a href="http://twitter.com/download/iphone" r...	I can't beli happening
94	874012996292530176	2017-06-11 21:18:31	<a href="http://twitter.com/download/iphone" r...	This is Seb can't see a
96	873580283840344065	2017-06-10 16:39:04	<a href="http://twitter.com/download/iphone" r...	We usually Deck-boun Bla...
98	873213775632977920	2017-06-09 16:22:42	<a href="http://twitter.com/download/iphone" r...	This is Sie precious p
99	872967104147763200	2017-06-09 00:02:31	<a href="http://twitter.com/download/iphone" r...	Here's a v He has a c
100	872820683541237760	2017-06-08 14:20:41	<a href="http://twitter.com/download/iphone" r...	Here are m #dogsatpo \n...
102	872620804844003328	2017-06-08 01:06:27	<a href="http://twitter.com/download/iphone" r...	This is Mo supporting everyw...
103	872486979161796608	2017-06-07 16:14:40	<a href="http://twitter.com/download/iphone" r...	We. Only. not send i
104	872261713294495745	2017-06-07 01:19:32	<a href="http://twitter.com/download/iphone" r...	This is Har activated
105	872122724285648897	2017-06-06 16:07:15	<a href="http://twitter.com/download/iphone" r...	This is Koo baller. Wis l...

	tweet_id	timestamp	source	
106	871879754684805121	2017-06-06 00:01:46	<a href="http://twitter.com/download/iphone" r...	Say hello t celebratin
107	871762521631449091	2017-06-05 16:15:56	<a href="http://twitter.com/download/iphone" r...	This is Ro pupper pr
108	871515927908634625	2017-06-04 23:56:03	<a href="http://twitter.com/download/iphone" r...	This is Na Raggedy E
110	871102520638267392	2017-06-03 20:33:19	<a href="http://twitter.com/download/iphone" r...	Never dou 14/10 https://t.c
111	871032628920680449	2017-06-03 15:55:36	<a href="http://twitter.com/download/iphone" r...	This is Bo doing an a t...
112	870804317367881728	2017-06-03 00:48:22	<a href="http://twitter.com/download/iphone" r...	Real funny in a pic wi
114	870656317836468226	2017-06-02 15:00:16	<a href="http://twitter.com/download/iphone" r...	This is Co too aggres
115	870374049280663552	2017-06-01 20:18:38	<a href="http://twitter.com/download/iphone" r...	This is Zoe likes the p

100 rows × 12 columns

Quality Improvement - Change Source to HTML value instead of the whole html string

In [24]:
```
# Use Extract to extract the text between the HTML markers.

df_tae_clean['source'].unique()
df_tae_clean['source'] = df_tae_clean.source.str.extract('>(.*)<', expan
d=True)

df_tae_clean.head()
```

Out[24]:

	tweet_id	timestamp	source	text	
0	892420643555336193	2017-08-01 16:23:56	Twitter for iPhone	This is Phineas. He's a mystical boy. Only eve...	https://twitter.com/dog_rates/statu
1	892177421306343426	2017-08-01 00:17:27	Twitter for iPhone	This is Tilly. She's just checking pup on you....	https://twitter.com/dog_rates/statu
2	891815181378084864	2017-07-31 00:18:03	Twitter for iPhone	This is Archie. He is a rare Norwegian Pouncin...	https://twitter.com/dog_rates/statu
3	891689557279858688	2017-07-30 15:58:51	Twitter for iPhone	This is Darla. She commenced a snooze mid meal...	https://twitter.com/dog_rates/statu
4	891327558926688256	2017-07-29 16:00:24	Twitter for iPhone	This is Franklin. He would like you to stop ca...	https://twitter.com/dog_rates/statu

In [25]:
```
# Test - In separate cell because running the extract twice removes the
 content altogether.
df_tae_clean['source'].unique()
```

Out[25]:
```
array(['Twitter for iPhone', 'Twitter Web Client', 'Vine - Make a Scen
e',
       'TweetDeck'], dtype=object)
```

Tidyness - Remove links from text since they have been put into expanded url section

In [26]:
```
# Use extract with regext to get rid of the URLs at the end of the text

df_tae_clean['text'] = df_tae_clean.text.str.extract('(.*)[^https://]',
expand=True)
```

In [27]:
```
# Test individual value that was proving to be tricky because it had two
 links

print(df_tae_clean['text'][6])
```

```
Meet Jax. He enjoys ice cream so much he gets nervous around it. 13/10
help Jax enjoy more things by clicking below
```

In [28]:
```
df_tae_clean.head()
```

Out[28]:

	tweet_id	timestamp	source	text	
0	892420643555336193	2017-08-01 16:23:56	Twitter for iPhone	This is Phineas. He's a mystical boy. Only eve...	https://twitter.com/dog_rates/statu
1	892177421306343426	2017-08-01 00:17:27	Twitter for iPhone	This is Tilly. She's just checking pup on you....	https://twitter.com/dog_rates/statu
2	891815181378084864	2017-07-31 00:18:03	Twitter for iPhone	This is Archie. He is a rare Norwegian Pouncin...	https://twitter.com/dog_rates/statu
3	891689557279858688	2017-07-30 15:58:51	Twitter for iPhone	This is Darla. She commenced a snooze mid meal...	https://twitter.com/dog_rates/statu
4	891327558926688256	2017-07-29 16:00:24	Twitter for iPhone	This is Franklin. He would like you to stop ca...	https://twitter.com/dog_rates/statu

Quality Improvement - Remove rows where the rating denominator is not 10

In [29]:
```
# Check the rows that have a rating_demoninator not equal to 10.
# option to drop those rows or to adjust the data - Decide to drop the d
ata because it would end up with
# non-interger numbers

df_tae_clean.head()
df_tae_clean.info()

df_rd = df_tae_clean[df_tae_clean['rating_denominator'] != 10]

print(df_rd)

df_tae_clean=df_tae_clean[df_tae_clean['rating_denominator'] == 10]
df_tae_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2096 entries, 0 to 2355
Data columns (total 12 columns):
tweet_id              2096 non-null int64
timestamp             2096 non-null datetime64[ns]
source                2096 non-null object
text                  2096 non-null object
expanded_urls         2093 non-null object
rating_numerator      2096 non-null int64
rating_denominator    2096 non-null int64
name                  1389 non-null object
doggo                 83 non-null object
floofer               10 non-null object
pupper                230 non-null object
puppo                 24 non-null object
dtypes: datetime64[ns](1), int64(3), object(8)
memory usage: 292.9+ KB
                tweet_id           timestamp              source  \
433   820690176645140481 2017-01-15 17:52:40  Twitter for iPhone
516   810984652412424192 2016-12-19 23:06:23  Twitter for iPhone
902   758467244762497024 2016-07-28 01:00:57  Twitter for iPhone
1068  740373189193256964 2016-06-08 02:41:38  Twitter for iPhone
1120  731156023742988288 2016-05-13 16:15:54  Twitter for iPhone
1165  722974582966214656 2016-04-21 02:25:47  Twitter for iPhone
1202  716439118184652801 2016-04-03 01:36:11  Twitter for iPhone
1228  713900603437621249 2016-03-27 01:29:02  Twitter for iPhone
1254  710658690886586372 2016-03-18 02:46:49  Twitter for iPhone
1274  709198395643068416 2016-03-14 02:04:08  Twitter for iPhone
1351  704054845121142784 2016-02-28 21:25:30  Twitter for iPhone
1433  697463031882764288 2016-02-10 16:51:59  Twitter for iPhone
1635  684222868335505415 2016-01-05 04:00:18  Twitter for iPhone
1662  682962037429899265 2016-01-01 16:30:13  Twitter for iPhone
1779  677716515794329600 2015-12-18 05:06:23  Twitter for iPhone
1843  675853064436391936 2015-12-13 01:41:41  Twitter for iPhone
2335  666287406224695296 2015-11-16 16:11:11  Twitter for iPhone

                                                    text  \
433   The floofs have been released I repeat the flo...
516   Meet Sam. She smiles 24/7 &amp;amp; secretly aspir...
902   Why does this never happen at my front door.....
1068  After so many requests, this is Bretagne. She ...
1120  Say hello to this unbelievably well behaved sq...
1165  Happy 4/20 from the squad! 13/10 for all https...
1202  This is Bluebert. He just saw that both #Final...
1228  Happy Saturday here's 9 puppers on a bench. 99...
1254  Here's a brigade of puppers. All look very pre...
1274                                 From left to right:
1351  Here is a whole flock of puppers.  60/50 I'll ...
1433  Happy Wednesday here's a bucket of pups. 44/40...
1635  Someone help the girl is being mugged. Several...
1662  This is Darrel. He just robbed a 7/11 and is i...
1779  IT'S PUPPERGEDDON. Total of 144/120 ...I think...
1843  Here we have an entire platoon of puppers. Tot...
2335  This is an Albanian 3 1/2 legged  Episcopalian...

                                          expanded_urls  rating_numerat
or  \
```

```
433   https://twitter.com/dog_rates/status/820690176...
84
516   https://www.gofundme.com/sams-smile,https://tw...
24
902   https://twitter.com/dog_rates/status/758467244...              1
65
1068  https://twitter.com/dog_rates/status/740373189...
9
1120  https://twitter.com/dog_rates/status/731156023...              2
04
1165  https://twitter.com/dog_rates/status/722974582...
4
1202  https://twitter.com/dog_rates/status/716439118...
50
1228  https://twitter.com/dog_rates/status/713900603...
99
1254  https://twitter.com/dog_rates/status/710658690...
80
1274  https://twitter.com/dog_rates/status/709198395...
45
1351  https://twitter.com/dog_rates/status/704054845...
60
1433  https://twitter.com/dog_rates/status/697463031...
44
1635  https://twitter.com/dog_rates/status/684222868...              1
21
1662  https://twitter.com/dog_rates/status/682962037...
7
1779  https://twitter.com/dog_rates/status/677716515...              1
44
1843  https://twitter.com/dog_rates/status/675853064...
88
2335  https://twitter.com/dog_rates/status/666287406...
1

      rating_denominator      name doggo floofer pupper puppo
433                   70       NaN   NaN     NaN    NaN   NaN
516                    7       Sam   NaN     NaN    NaN   NaN
902                  150       NaN   NaN     NaN    NaN   NaN
1068                  11       NaN   NaN     NaN    NaN   NaN
1120                 170       NaN   NaN     NaN    NaN   NaN
1165                  20       NaN   NaN     NaN    NaN   NaN
1202                  50  Bluebert   NaN     NaN    NaN   NaN
1228                  90       NaN   NaN     NaN    NaN   NaN
1254                  80       NaN   NaN     NaN    NaN   NaN
1274                  50       NaN   NaN     NaN    NaN   NaN
1351                  50       NaN   NaN     NaN    NaN   NaN
1433                  40       NaN   NaN     NaN    NaN   NaN
1635                 110       NaN   NaN     NaN    NaN   NaN
1662                  11    Darrel   NaN     NaN    NaN   NaN
1779                 120       NaN   NaN     NaN    NaN   NaN
1843                  80       NaN   NaN     NaN    NaN   NaN
2335                   2       NaN   NaN     NaN    NaN   NaN
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2079 entries, 0 to 2355
Data columns (total 12 columns):
tweet_id              2079 non-null int64
```

```
timestamp             2079 non-null datetime64[ns]
source                2079 non-null object
text                  2079 non-null object
expanded_urls         2076 non-null object
rating_numerator      2079 non-null int64
rating_denominator    2079 non-null int64
name                  1386 non-null object
doggo                 83 non-null object
floofer               10 non-null object
pupper                230 non-null object
puppo                 24 non-null object
dtypes: datetime64[ns](1), int64(3), object(8)
memory usage: 211.1+ KB
```

Quality Improvement - Remove rows where there are no photos

In [30]:
```
# Made a copy while working out the code.

df_eu = df_tae_clean.copy()

# Checking to see if there are photos in expanded urls and removing rows
 that do not have photo

df_tae_clean = df_eu[df_eu["expanded_urls"].str.contains('photo')==True]

df_tae_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1882 entries, 0 to 2355
Data columns (total 12 columns):
tweet_id              1882 non-null int64
timestamp             1882 non-null datetime64[ns]
source                1882 non-null object
text                  1882 non-null object
expanded_urls         1882 non-null object
rating_numerator      1882 non-null int64
rating_denominator    1882 non-null int64
name                  1313 non-null object
doggo                 64 non-null object
floofer               7 non-null object
pupper                199 non-null object
puppo                 23 non-null object
dtypes: datetime64[ns](1), int64(3), object(8)
memory usage: 191.1+ KB
```

Tidiness - Merging the three tables into one table

In [31]:
```
# For reference:
# df_tae_clean = pd.read_csv('twitter-archive-enhanced.csv') - cleaned u
p
# df_tip = pd.read_csv('image-predictions.tsv',sep='\t')
# df_info = pd.read_csv('tweet_extra_info.csv')

# Double Checking the data

df_tae_clean.info()
df_tae_clean.head()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1882 entries, 0 to 2355
Data columns (total 12 columns):
tweet_id              1882 non-null int64
timestamp             1882 non-null datetime64[ns]
source                1882 non-null object
text                  1882 non-null object
expanded_urls         1882 non-null object
rating_numerator      1882 non-null int64
rating_denominator    1882 non-null int64
name                  1313 non-null object
doggo                 64 non-null object
floofer               7 non-null object
pupper                199 non-null object
puppo                 23 non-null object
dtypes: datetime64[ns](1), int64(3), object(8)
memory usage: 191.1+ KB
```

Out[31]:

	tweet_id	timestamp	source	text	
0	892420643555336193	2017-08-01 16:23:56	Twitter for iPhone	This is Phineas. He's a mystical boy. Only eve...	https://twitter.com/dog_rates/statu
1	892177421306343426	2017-08-01 00:17:27	Twitter for iPhone	This is Tilly. She's just checking pup on you....	https://twitter.com/dog_rates/statu
2	891815181378084864	2017-07-31 00:18:03	Twitter for iPhone	This is Archie. He is a rare Norwegian Pouncin...	https://twitter.com/dog_rates/statu
3	891689557279858688	2017-07-30 15:58:51	Twitter for iPhone	This is Darla. She commenced a snooze mid meal...	https://twitter.com/dog_rates/statu
4	891327558926688256	2017-07-29 16:00:24	Twitter for iPhone	This is Franklin. He would like you to stop ca...	https://twitter.com/dog_rates/statu

In [32]:
```
# Double Checking the data

df_tip.info()
df_tip.head()
```

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

Out[32]:

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

In [33]:
```
# Double Checking the data

df_info.info()
df_info.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2344 entries, 0 to 2343
Data columns (total 3 columns):
tweet_id         2344 non-null int64
retweet_count    2344 non-null int64
like_count       2344 non-null int64
dtypes: int64(3)
memory usage: 55.0 KB
```

Out[33]:

	tweet_id	retweet_count	like_count
0	892420643555336193	8553	38673
1	892177421306343426	6287	33127
2	891815181378084864	4166	24943
3	891689557279858688	8675	42044
4	891327558926688256	9441	40193

In [34]:
```
# Left merge to get values from the likes and retweets table into the ma
in cleaned table.

df_merged=pd.merge(df_tae_clean, df_info, on='tweet_id', how='left')
df_merged.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1882 entries, 0 to 1881
Data columns (total 14 columns):
tweet_id              1882 non-null int64
timestamp             1882 non-null datetime64[ns]
source                1882 non-null object
text                  1882 non-null object
expanded_urls         1882 non-null object
rating_numerator      1882 non-null int64
rating_denominator    1882 non-null int64
name                  1313 non-null object
doggo                 64 non-null object
floofer               7 non-null object
pupper                199 non-null object
puppo                 23 non-null object
retweet_count         1882 non-null int64
like_count            1882 non-null int64
dtypes: datetime64[ns](1), int64(5), object(8)
memory usage: 220.5+ KB
```

In [35]:
```
# Left merge to get values from the photo analysis table into the main c
leaned table.

df_merged=pd.merge(df_merged, df_tip, on='tweet_id', how='left')
df_merged.info()
df_merged.head()

df_cleaned = df_merged.copy()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1882 entries, 0 to 1881
Data columns (total 25 columns):
tweet_id              1882 non-null int64
timestamp             1882 non-null datetime64[ns]
source                1882 non-null object
text                  1882 non-null object
expanded_urls         1882 non-null object
rating_numerator      1882 non-null int64
rating_denominator    1882 non-null int64
name                  1313 non-null object
doggo                 64 non-null object
floofer               7 non-null object
pupper                199 non-null object
puppo                 23 non-null object
retweet_count         1882 non-null int64
like_count            1882 non-null int64
jpg_url               1882 non-null object
img_num               1882 non-null int64
p1                    1882 non-null object
p1_conf               1882 non-null float64
p1_dog                1882 non-null bool
p2                    1882 non-null object
p2_conf               1882 non-null float64
p2_dog                1882 non-null bool
p3                    1882 non-null object
p3_conf               1882 non-null float64
p3_dog                1882 non-null bool
dtypes: bool(3), datetime64[ns](1), float64(3), int64(6), object(12)
memory usage: 343.7+ KB
```

Storing

In [36]:
```
# Saving final cleaned data off to csv

df_cleaned.to_csv('./data/twitter_archive_master.csv',index=False)
```

Analyzing and Visualizing Data

Requirements:

Analyze and visualize your wrangled data in your wrangle_act.ipynb Jupyter Notebook. At least three (3) insights and one (1) visualization must be produced.

In [37]:
```
# Reading in the cleaned data and looking at the basic info for the fina
l product.

df_t = pd.read_csv("./data/twitter_archive_master.csv")
df_t.head()
```

Out[37]:

	tweet_id	timestamp	source	text	
0	892420643555336193	2017-08-01 16:23:56	Twitter for iPhone	This is Phineas. He's a mystical boy. Only eve...	https://twitter.com/dog_rates/statu
1	892177421306343426	2017-08-01 00:17:27	Twitter for iPhone	This is Tilly. She's just checking pup on you....	https://twitter.com/dog_rates/statu
2	891815181378084864	2017-07-31 00:18:03	Twitter for iPhone	This is Archie. He is a rare Norwegian Pouncin...	https://twitter.com/dog_rates/statu
3	891689557279858688	2017-07-30 15:58:51	Twitter for iPhone	This is Darla. She commenced a snooze mid meal...	https://twitter.com/dog_rates/statu
4	891327558926688256	2017-07-29 16:00:24	Twitter for iPhone	This is Franklin. He would like you to stop ca...	https://twitter.com/dog_rates/statu

5 rows × 25 columns

In [38]:
```
df_t.describe()
```

Out[38]:

	tweet_id	rating_numerator	rating_denominator	retweet_count	like_cou
count	1.882000e+03	1882.000000	1882.0	1882.000000	1882.000000
mean	7.350763e+17	11.691817	10.0	2466.648247	8417.953773
std	6.751862e+16	41.857148	0.0	3611.245684	11378.795038
min	6.660209e+17	0.000000	10.0	13.000000	80.000000
25%	6.753650e+17	10.000000	10.0	591.000000	1830.000000
50%	7.077175e+17	11.000000	10.0	1284.500000	3919.500000
75%	7.868996e+17	12.000000	10.0	3012.500000	10829.750000
max	8.924206e+17	1776.000000	10.0	48908.000000	142895.00000

In [39]:
```
# Most popular in terms of retweet counts

df_t[df_t['retweet_count']==df_t['retweet_count'].max()]
```

Out[39]:

	tweet_id	timestamp	source	text	
292	822872901745569793	2017-01-21 18:26:02	Twitter for iPhone	Here's a super supportive puppo participating ...	https://twitter.com/dog_rates/st

1 rows × 25 columns

In [40]:
```
df_t['retweet_count'].max()
```

Out[40]:
```
48908
```

In [41]:
```
# Most popular in terms of like counts

df_t[df_t['like_count']==df_t['like_count'].max()]
```

Out[41]:

	tweet_id	timestamp	source	text	
292	822872901745569793	2017-01-21 18:26:02	Twitter for iPhone	Here's a super supportive puppo participating ...	https://twitter.com/dog_rates/st

1 rows × 25 columns

In [42]:
```
df_t['like_count'].max()
```

Out[42]:
```
142895
```

In [43]:
```
df_t.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1882 entries, 0 to 1881
Data columns (total 25 columns):
tweet_id              1882 non-null int64
timestamp             1882 non-null object
source                1882 non-null object
text                  1882 non-null object
expanded_urls         1882 non-null object
rating_numerator      1882 non-null int64
rating_denominator    1882 non-null int64
name                  1313 non-null object
doggo                 64 non-null object
floofer               7 non-null object
pupper                199 non-null object
puppo                 23 non-null object
retweet_count         1882 non-null int64
like_count            1882 non-null int64
jpg_url               1882 non-null object
img_num               1882 non-null int64
p1                    1882 non-null object
p1_conf               1882 non-null float64
p1_dog                1882 non-null bool
p2                    1882 non-null object
p2_conf               1882 non-null float64
p2_dog                1882 non-null bool
p3                    1882 non-null object
p3_conf               1882 non-null float64
p3_dog                1882 non-null bool
dtypes: bool(3), float64(3), int64(6), object(13)
memory usage: 329.1+ KB
```

In [44]:
```
df_t.sample(100)
```

Out[44]:

	tweet_id	timestamp	source	text	
225	834209720923721728	2017-02-22 01:14:30	Twitter for iPhone	This is Wilson. He's aware that he has somethi...	https://twitter.com/dog_rate
751	739485634323156992	2016-06-05 15:54:48	Twitter for iPhone	This is Kyle. He's a heavy drinker and an avid...	https://twitter.com/dog_rate
1719	668979806671884288	2015-11-24 02:29:49	Twitter for iPhone	This is Chaz. He's an X Games half pipe supers...	https://twitter.com/dog_rate
1508	673240798075449344	2015-12-05 20:41:29	Twitter for iPhone	Magical floating dog here. Very calm. Always h...	https://twitter.com/dog_rate
1637	670676092097810432	2015-11-28 18:50:15	Twitter for iPhone	This is Bloo. He's a Westminster Cîroc. Doesn'...	https://twitter.com/dog_rate
1388	676098748976615425	2015-12-13 17:57:57	Twitter for iPhone	Extremely rare pup here. Very religious. Alway...	https://twitter.com/dog_rate
1479	673709992831262724	2015-12-07 03:45:53	Twitter for iPhone	I know a lot of you are studying for finals. G...	https://twitter.com/dog_rate
1460	674082852460433408	2015-12-08 04:27:30	Twitter for iPhone	This is a Sagitariot Baklava mix. Loves her ne...	https://twitter.com/dog_rate
883	713411074226274305	2016-03-25 17:03:49	Twitter for iPhone	Here we see an extremely rare Bearded Floofmal...	https://twitter.com/dog_rate
841	719551379208073216	2016-04-11 15:43:12	Twitter for iPhone	This is Harnold. He accidentally opened the fr...	https://twitter.com/dog_rate

	tweet_id	timestamp	source	text	
380	805487436403003392	2016-12-04 19:02:24	Twitter for iPhone	Meet Sansa and Gary. They run along the fence ...	https://twitter.com/dog_rate
1399	675710890956750848	2015-12-12 16:16:45	Twitter for iPhone	This is Lenny. He was just told that he couldn...	https://twitter.com/dog_rate
1053	697482927769255936	2016-02-10 18:11:03	Twitter for iPhone	Meet Blipson. He's a Doowap Hufflepuff. That U...	https://twitter.com/dog_rate
120	860524505164394496	2017-05-05 16:00:04	Twitter for iPhone	This is Carl. He likes to dance. Doesn't care ...	https://twitter.com/dog_rate
939	707776935007539200	2016-03-10 03:55:45	Twitter for iPhone	This is Sadie. She's a Bohemian Rhapsody. Rema...	https://twitter.com/dog_rate
648	752917284578922496	2016-07-12 17:27:23	Twitter for iPhone	This is Grizzie. She's a semi-submerged Bahrai...	https://twitter.com/dog_rate
657	751538714308972544	2016-07-08 22:09:27	Twitter for iPhone	This is Max. She has one ear that's always sli...	https://twitter.com/dog_rate
1380	676430933382295552	2015-12-14 15:57:56	Twitter for iPhone	Meet Duke. He's an Urban Parmesan. They know h...	https://twitter.com/dog_rate
1267	681242418453299201	2015-12-27 22:37:04	Twitter for iPhone	This is Champ. He's being sacrificed to the Az...	https://twitter.com/dog_rate
390	802239329049477120	2016-11-25 19:55:35	Twitter for iPhone	This is Loki. He'll do your taxes for you. Can...	https://twitter.com/dog_rate

	tweet_id	timestamp	source	text	
960	706291001778950144	2016-03-06 01:31:11	Twitter for iPhone	When you're just relaxin and having a swell ti...	https://twitter.com/dog_rate
526	776218204058357768	2016-09-15 00:36:55	Twitter for iPhone	Atlas rolled around in some chalk and now he's...	https://twitter.com/dog_rate
1393	675878199931371520	2015-12-13 03:21:34	Twitter for iPhone	Ok, I'll admit this is a pretty adorable bunny...	https://twitter.com/dog_rate
1843	666649482315059201	2015-11-17 16:09:56	Twitter for iPhone	Cool dog. Enjoys couch. Low monotone bark. Ver...	https://twitter.com/dog_rate
1419	675135153782571009	2015-12-11 02:08:58	Twitter for iPhone	This is Steven. He got locked outside. Damn it...	https://twitter.com/dog_rate
1691	669573570759163904	2015-11-25 17:49:14	Twitter for iPhone	This is Linda. She just looked up and saw you ...	https://twitter.com/dog_rate
1245	682389078323662849	2015-12-31 02:33:29	Twitter for iPhone	Meet Brody. He's a Downton Abbey Falsetto. Add...	https://twitter.com/dog_rate
125	859607811541651456	2017-05-03 03:17:27	Twitter for iPhone	Sorry for the lack of posts today. I came home...	https://twitter.com/dog_rate
1029	699423671849451520	2016-02-16 02:42:52	Twitter for iPhone	"Don't ever talk to me or my son again." ...bo...	https://twitter.com/dog_rate
810	727286334147182592	2016-05-02 23:59:09	Twitter for iPhone	I swear to god if we get sent another Blue Mad...	https://twitter.com/dog_rate
...	...	...	...	...	...

	tweet_id	timestamp	source	text	
416	796116448414461957	2016-11-08 22:25:27	Twitter for iPhone	I didn't believe it at first but now I can see...	https://twitter.com/dog_rate
612	759793422261743616	2016-07-31 16:50:42	Twitter for iPhone	Meet Maggie & Lila. Maggie is the doggo, L...	https://twitter.com/dog_rate
251	829501995190984704	2017-02-09 01:27:41	Twitter for iPhone	This is Leo. He was a skater pup. She said see...	https://twitter.com/dog_rate
520	777885040357281792	2016-09-19 15:00:20	Twitter for iPhone	This is Wesley. He's clearly trespassing. Seem...	https://twitter.com/dog_rate
1339	677895101218201600	2015-12-18 16:56:01	Twitter for iPhone	Guys this was terrifying. Really spooked me up...	https://twitter.com/dog_rate
976	704819833553219584	2016-03-02 00:05:17	Twitter for iPhone	This is Chesterson. He's a Bolivian Scoop Dog....	https://twitter.com/dog_rate
1398	675781562965868544	2015-12-12 20:57:34	Twitter for iPhone	Say hello to Buddah. He was Waldo for Hallowee...	https://twitter.com/dog_rate
1001	701981390485725185	2016-02-23 04:06:20	Twitter for iPhone	This is Fiji. She's a Powdered Stegafloof. Ver...	https://twitter.com/dog_rate
1266	681261549936340994	2015-12-27 23:53:05	Twitter for iPhone	Say hello to Panda. He's a Quackadilly Shooste...	https://twitter.com/dog_rate
1505	673317986296586240	2015-12-06 01:48:12	Twitter for iPhone	Take a moment and appreciate how these two dog...	https://twitter.com/dog_rate

	tweet_id	timestamp	source	text	
670	750071704093859840	2016-07-04 21:00:04	Twitter for iPhone	Pause your cookout and admire this pupper's ni...	https://twitter.com/dog_rate
1405	675501075957489664	2015-12-12 02:23:01	Twitter for iPhone	I shall call him squishy and he shall be mine,...	https://twitter.com/dog_rate
1394	675845657354215424	2015-12-13 01:12:15	Twitter for iPhone	This is Pepper. She's not fully comfortable ri...	https://twitter.com/dog_rate
1849	666430724426358785	2015-11-17 01:40:41	Twitter for iPhone	Oh boy what a pup! Sunglasses take this one to...	https://twitter.com/dog_rate
571	768473857036525572	2016-08-24 15:43:39	Twitter for iPhone	Meet Chevy. He had a late breakfast and now ha...	https://twitter.com/dog_rate
1713	668994913074286592	2015-11-24 03:29:51	Twitter for iPhone	Two gorgeous pups here. Both have cute fake ho...	https://twitter.com/dog_rate
1443	674638615994089473	2015-12-09 17:15:54	Twitter for iPhone	This pupper is fed up with being tickled. 12/1...	https://twitter.com/dog_rate
1642	670452855871037440	2015-11-28 04:03:11	Twitter for iPhone	This dog can't see its haters. 11/10 https://t...	https://twitter.com/dog_rate
303	820749716845686786	2017-01-15 21:49:15	Twitter for iPhone	Meet Sunny. He can take down a polar bear in o...	https://twitter.com/dog_rate
553	771770456517009408	2016-09-02 18:03:10	Twitter for iPhone	This is Davey. He'll have your daughter home b...	https://twitter.com/dog_rate

	tweet_id	timestamp	source	text	
299	821765923262631936	2017-01-18 17:07:18	Twitter for iPhone	This is Duchess. She uses dark doggo forces to...	https://twitter.com/dog_rate
163	848324959059550208	2017-04-02 00:03:26	Twitter for iPhone	Meet Odin. He's supposed to be giving directio...	https://twitter.com/dog_rate
1395	675822767435051008	2015-12-12 23:41:18	Twitter for iPhone	🎶 HELLO FROM THE OTHER SIIIIIIIIDE 🎶 10/10s ht...	https://twitter.com/dog_rate
438	793135492858580992	2016-10-31 17:00:11	Twitter for iPhone	Your favorite squad is looking extra h*ckin sp...	https://twitter.com/dog_rate
1463	674053186244734976	2015-12-08 02:29:37	Twitter for iPhone	This is Stanley. Yes he is aware of the spoon'...	https://twitter.com/dog_rate
1255	681694085539872773	2015-12-29 04:31:49	Twitter for iPhone	This is Bo. He's a Benedoop Cumbersnatch. Seem...	https://twitter.com/dog_rate
1490	673656262056419329	2015-12-07 00:12:23	Twitter for iPhone	This is Albert AKA King Banana Peel. He's a ki...	https://twitter.com/dog_rate
460	788908386943430656	2016-10-20 01:03:11	Twitter for iPhone	This is Lucy. She destroyed not one, but two r...	https://twitter.com/dog_rate
1158	688116655151435777	2016-01-15 21:52:49	Twitter for iPhone	Please send dogs. I'm tired of seeing other st...	https://twitter.com/dog_rate
977	704761120771465216	2016-03-01 20:11:59	Twitter for iPhone	This pupper killed this great white in an epic...	https://twitter.com/dog_rate

100 rows × 25 columns

In [45]:
```
# What breed is the most popular to tweet about?

dog_breed = df_t.groupby('p1').count().sort_values('tweet_id',ascending=
False)

dog_breed.head(10)
```

Out[45]:

	tweet_id	timestamp	source	text	expanded_urls	rating_numerator	
p1							
golden_retriever	131	131	131	131	131	131	
Pembroke	87	87	87	87	87	87	
Labrador_retriever	87	87	87	87	87	87	
Chihuahua	76	76	76	76	76	76	
pug	54	54	54	54	54	54	
chow	41	41	41	41	41	41	
toy_poodle	37	37	37	37	37	37	
Pomeranian	37	37	37	37	37	37	
Samoyed	37	37	37	37	37	37	
malamute	29	29	29	29	29	29	

10 rows × 24 columns

In [46]:
```
# Quick check to see if there might be different ways that golden retrie
ver appears

df_t['p1'][df_t['p1'].str.contains('olden')].unique()
df_t['p1'][df_t['p1'].str.contains('nian')].unique()
```

Out[46]:
```
array(['Pomeranian'], dtype=object)
```

What breed is the most popular to tweet about?

Golden Retrievers

Observations:

1. Most of these breeds would be expected to be seen on a list of the most popular breeds. 131 seems kind of low for golden retriever given there are thousands of tweets being analyzed.
2. Looks like to get real information the data would need to be further cleaned. Thought the photo information would be more consistent.
3. After doing a quick it looks like although the breed names could be more consistent in terms of upper case and lower case, not seeing obvious duplicates. Would recommend changing all breed names to all lower case.

In [47]:
```
# What breed of dog got the highest rating?

df_t['rating_numerator'].max()

# What breed of dog got the highest rating?

df_t['p1'][df_t['rating_numerator'] == df_t['rating_numerator'].max()]
```

Out[47]:
```
676    bow_tie
Name: p1, dtype: object
```

In [48]:
```
# Bow tie, what is a bow_tie?  Checking p2 option.

df_t['p2'][df_t['rating_numerator'] == df_t['rating_numerator'].max()]
```

Out[48]:
```
676    sunglasses
Name: p2, dtype: object
```

In [49]:
```
# Sunglasses.  It appears that the image file is classifying things othe
r than breed.

df_t['tweet_id'][df_t['rating_numerator'] == df_t['rating_numerator'].ma
x()]

#  Using tweet_id visit the actual tweet:  https://twitter.com/dog_rate
s/status/749981277374128128
```

Out[49]:
```
676    749981277374128128
Name: tweet_id, dtype: int64
```

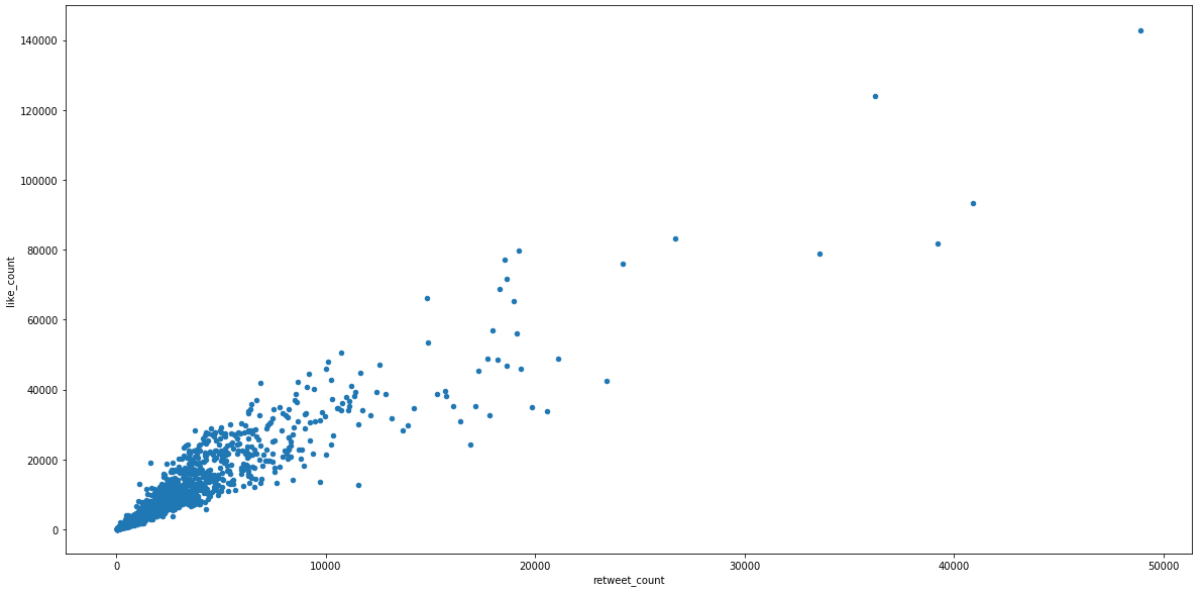
What breed of dog got the highest rating?

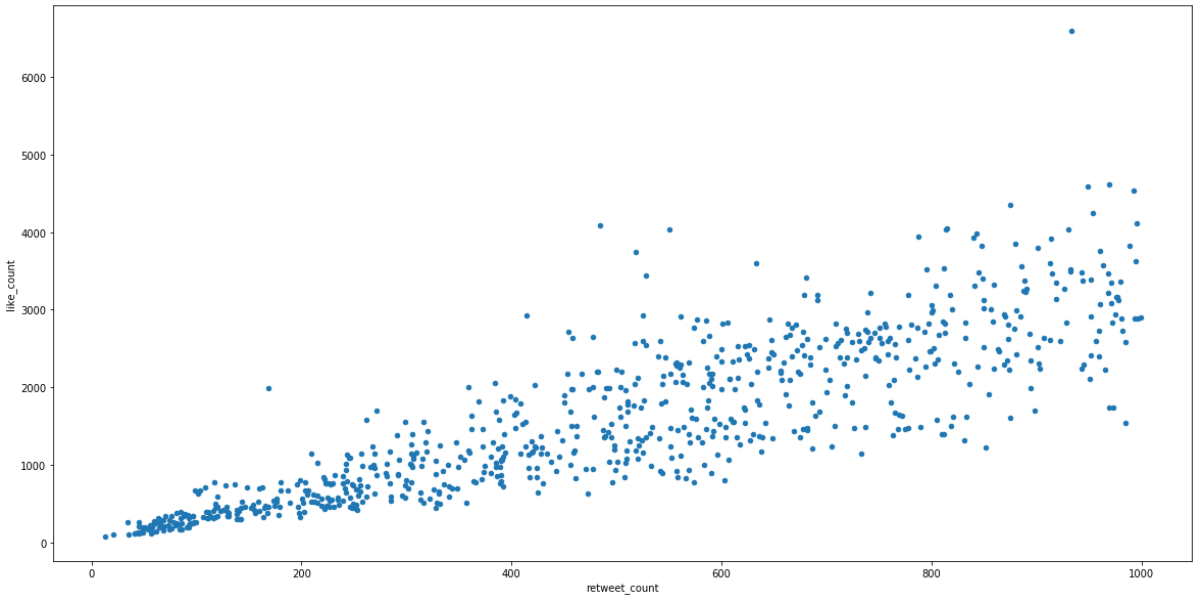
The highest rated dog was not idenfied by breed, but was correctly identified to be wearing a bow_tie and sunglasses: https://twitter.com/dog_rates/status/749981277374128128 (https://twitter.com/dog_rates/status/749981277374128128)

Observations:

1. The highest rated dog was p1 = bow_tie and p2 = sunglasses.
2. The information in the image predictions file contains more than breed information, so can't rely on it to get consistent breed information to compare tweets to breed.

In [50]:
```
# Are number of retweets and likes related?

df_t.plot('retweet_count','like_count',kind='scatter',figsize=(20,10));

df_p = df_t[df_t['retweet_count']<=1000]

df_p.plot('retweet_count','like_count',kind='scatter',figsize=(20,10));
```





In [51]:
```
df_t['tweet_id'][df_t['like_count'] == df_t['like_count'].max()]
```

Out[51]:
```
292    822872901745569793
Name: tweet_id, dtype: int64
```

In [52]:
```
df_t.nlargest(5, 'like_count')
```

Out[52]:

	tweet_id	timestamp	source	text	
292	822872901745569793	2017-01-21 18:26:02	Twitter for iPhone	Here's a super supportive puppo participating ...	https://twitter.com/dog_rates/st
102	866450705531457537	2017-05-22 00:28:40	Twitter for iPhone	This is Jamesy. He gives a kiss to every other...	https://twitter.com/dog_rates/st
312	819004803107983360	2017-01-11 02:15:36	Twitter for iPhone	This is Bo. He was a very good First Doggo. 14...	https://twitter.com/dog_rates/st
87	870374049280663552	2017-06-01 20:18:38	Twitter for iPhone	This is Zoey. She really likes the planet. Wou...	https://twitter.com/dog_rates/st
375	806629075125202948	2016-12-07 22:38:52	Twitter for iPhone	"Good afternoon class today we're going to lea...	https://twitter.com/dog_rates/st

5 rows × 25 columns

Are the number of retweets and likes related?

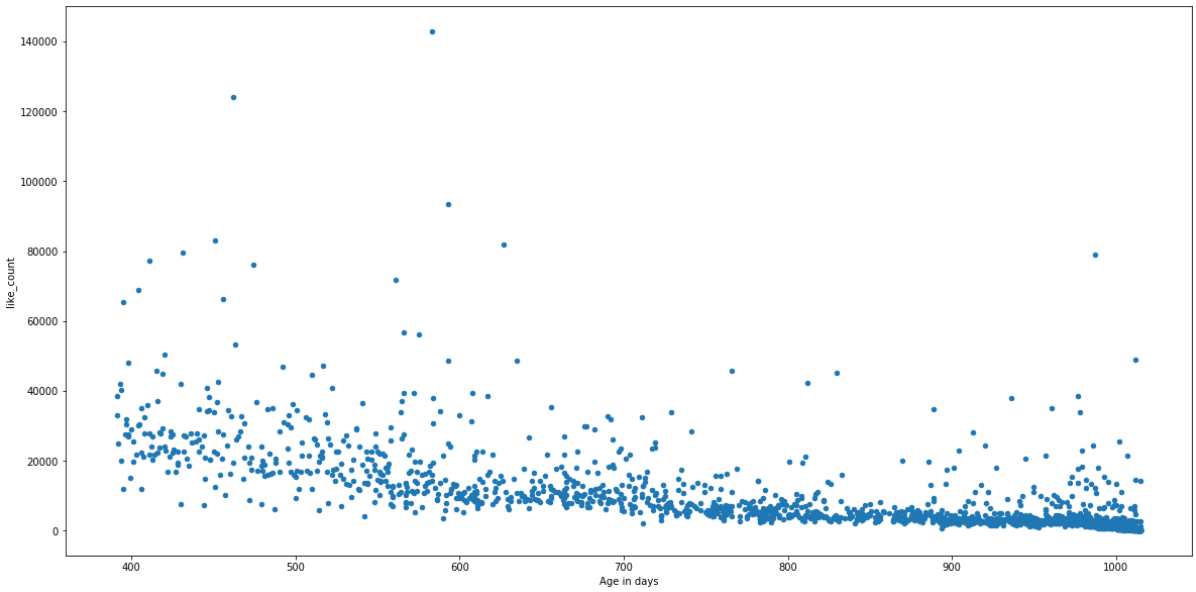
Looking at the plot above they do look related.

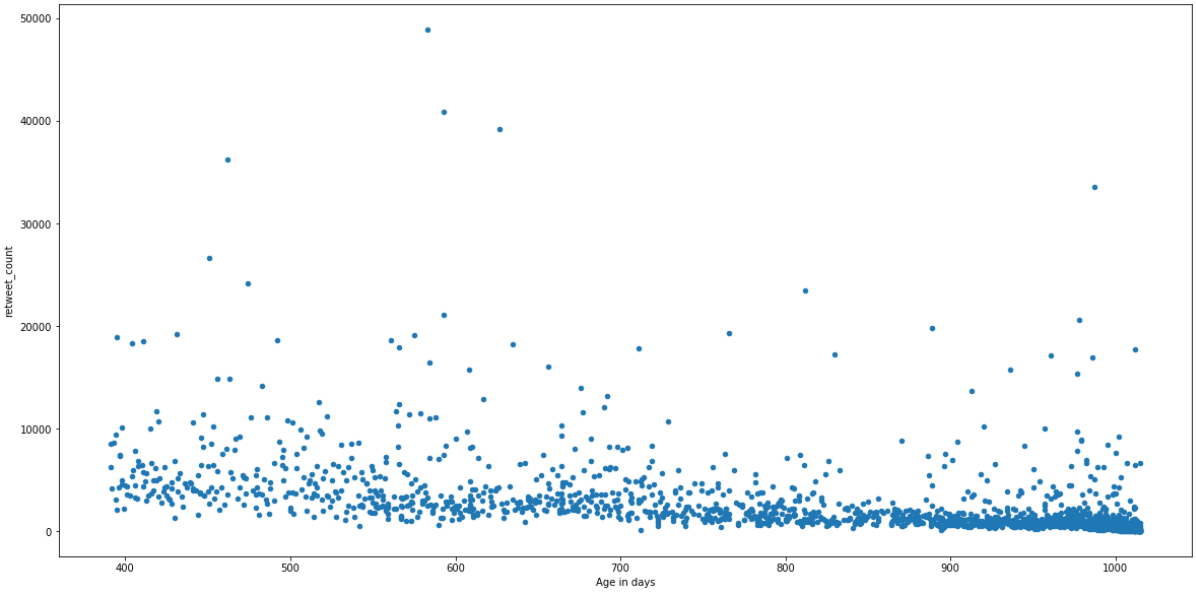
Observations:

1. Manually viewing the top 5 tweets they included the subjects:
2. Cute puppy
3. Political/Women's March dog
4. Obama's dog Bo (famous dog)
5. Environmental issue dog
6. Dog acting as professor

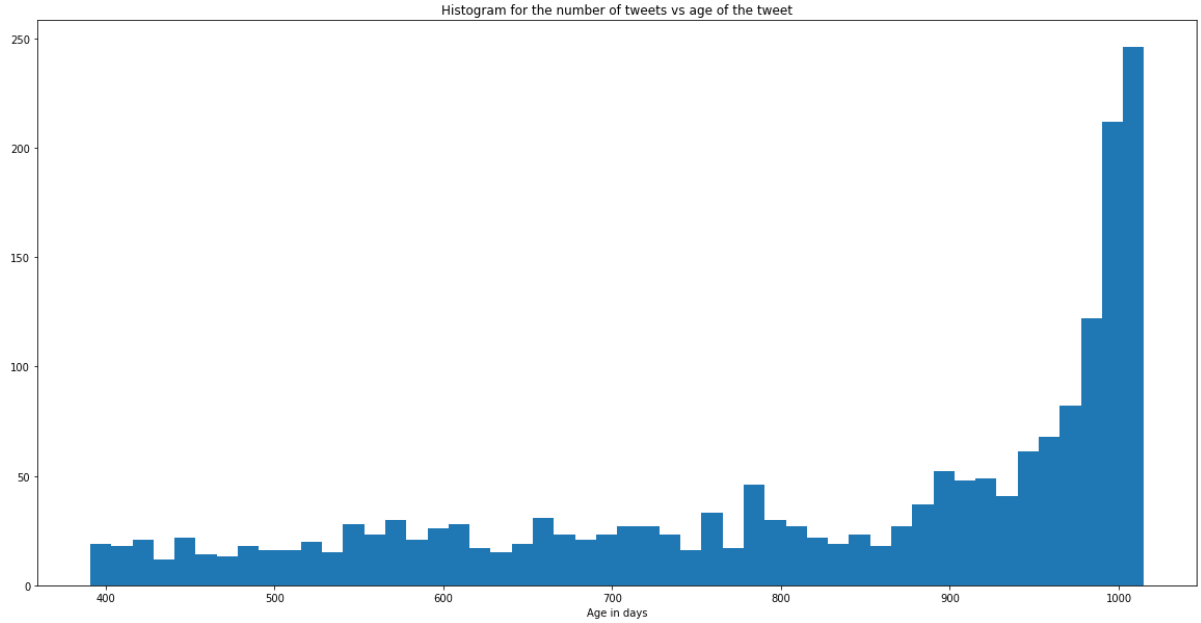
Possibility that tweets that are politically connected or tied to a major issue of the day may be more popular.

In [53]:
```
# Are the number of likes and retweets related to the age of the tweet?

df_a=df_t.copy()

# Can't plot a scatterplot based on the timestamp so need to add age in
 days.

ndates =[]

for d in df_a['timestamp']:
    ndates.append((datetime.now() - pd.to_datetime(d)).days)


df_a['Age in days']=ndates

df_a.head()

# Plotting Age in days vs likes and retweets

df_a.plot('Age in days','like_count',kind='scatter',figsize=(20,10));
df_a.plot('Age in days','retweet_count',kind='scatter',figsize=(20,10));
```


like_count
Age in days
0
20000
40000
60000
80000
100000
120000
140000
400
500
600
700
800
900
1000


retweet_count
Age in days
0
10000
20000
30000
40000
50000
400
500
600
700
800
900
1000

In [54]:
```
# Plotting histogram of Age in days to see tweet volumn over time
plt.figure(figsize=(20,10))
plt.hist(ndates, bins=50)
plt.xlabel("Age in days")
plt.title("Histogram for the number of tweets vs age of the tweet")

plt.show();
```



Are the number of likes and retweets related to the age of the tweet?

Looking at the plots above they do look related.

Observations:

1. There does seem to be a relationship, but it appears that newer tweets are more likely to get more likes and retweets than older tweets.
2. Looks like there were tweets more frequent a few years ago than there are today. Plotted histogram and it does appear that the frequency of tweets was much higher 800 days plus ago.
3. Cleaned data doesn't always provide the data in the right format for graphing.