wrangle_act

June 18, 2019

1 Gathering Data

In [1]:
```
import tweepy

        consumer_key = 'HIDDEN KEYS'
        consumer_secret = 'HIDDEN KEYS'
        access_token = 'HIDDEN KEYS'
        access_secret = 'HIDDEN KEYS'

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

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

In [7]:
```
import pandas as pd
import numpy as np
import random
import matplotlib.pyplot as plt
import json
import requests
```

In [3]:
```
df = pd.read_csv('twitter-archive-enhanced-2.csv')
```

In [4]:
```
missing_tweet_id = []
data = []
for i in range(len(df['tweet_id'])):
    try:
        tweet = api.get_status(df.iloc[i,0],tweet_mode='extended')
        data.append(tweet._json)
    except:
        print(df.iloc[i,0])
        missing_tweet_id.append(df.iloc[i,0])
```

```
888202515573088257
873697596434513921
872668790621863937
872261713294495745
```

```
869988702071779329
866816280283807744
861769973181624320
856602993587888130
845459076796616705
844704788403113984
842892208864923648
837012587749474308
827228250799742977
812747805718642688
802247111496568832
775096608509886464
770743923962707968
Rate limit reached. Sleeping for: 721
754011816964026368
680055455951884288
Rate limit reached. Sleeping for: 721
```

```
In [5]: with open('tweet_json.txt',mode='w') as f:
                json.dump(data, f)
```

```
In [4]: twitter_additional_archive=pd.read_json('tweet_json.txt')
```

```
In [8]: predictions = requests.get('https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599
        with open('predictions.tsv',mode='wb') as p:
             p.write(predictions.content)
```

```
In [6]: i_predictions = pd.read_csv('predictions.tsv', sep='\t')
```

2 Assesment of data

```
In [175]: #Tidy Issue 1: Removing unnecessary columns
           #Tidy Issue 2: Data needs joining into one master file
```

```
In [10]: df.info()
         #Quality Issue 1: Noticed that tweet_id is a integar but twitter recommends it to be a
         #Quality Issue 2: retweet_status_id is a 181 and so those data points must be removed
         #Quality Issue 3: Once retweeted rows are removed, retweet_status_id, retweeted_status_
                         #retweeted_status_timestamp are obsolete and should be removed.
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                    2356 non-null int64
in_reply_to_status_id       78 non-null float64
in_reply_to_user_id         78 non-null float64
timestamp                   2356 non-null object
```

```
source                          2356 non-null object
text                            2356 non-null object
retweeted_status_id             181 non-null float64
retweeted_status_user_id        181 non-null float64
retweeted_status_timestamp      181 non-null object
expanded_urls                   2297 non-null object
rating_numerator                2356 non-null int64
rating_denominator              2356 non-null int64
name                            2356 non-null object
doggo                           2356 non-null object
floofer                         2356 non-null object
pupper                          2356 non-null object
puppo                           2356 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

```
In [11]: twitter_additional_archive.info()
         #Quality Issue 4: retweet_status_id is a 166 and so those data points must be removed
         #Quality Issue 5: Once retweeted rows are removed, retweet status id and retweeted are
         #Quality Issue 6: Noticed that id is a integar but twitter recommends it to be a string
         #Quality Issue 7: id_str is the same as id and should be removed.

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2337 entries, 0 to 2336
Data columns (total 32 columns):
contributors                    0 non-null float64
coordinates                     0 non-null float64
created_at                      2337 non-null datetime64[ns]
display_text_range              2337 non-null object
entities                        2337 non-null object
extended_entities               2065 non-null object
favorite_count                  2337 non-null int64
favorited                       2337 non-null bool
full_text                       2337 non-null object
geo                             0 non-null float64
id                              2337 non-null int64
id_str                          2337 non-null int64
in_reply_to_screen_name         77 non-null object
in_reply_to_status_id           77 non-null float64
in_reply_to_status_id_str       77 non-null float64
in_reply_to_user_id             77 non-null float64
in_reply_to_user_id_str         77 non-null float64
is_quote_status                 2337 non-null bool
lang                            2337 non-null object
place                           1 non-null object
possibly_sensitive              2203 non-null float64
possibly_sensitive_appealable   2203 non-null float64
```

```
quoted_status                   24 non-null object
quoted_status_id                26 non-null float64
quoted_status_id_str            26 non-null float64
quoted_status_permalink         26 non-null object
retweet_count                   2337 non-null int64
retweeted                       2337 non-null bool
retweeted_status                166 non-null object
source                          2337 non-null object
truncated                       2337 non-null bool
user                            2337 non-null object
dtypes: bool(4), datetime64[ns](1), float64(11), int64(4), object(12)
memory usage: 520.4+ KB
```

```
In [12]: i_predictions.info()
```

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

```
In [13]: i_predictions['p1_dog'].value_counts()
```

```
Out[13]: True     1532
         False     543
         Name: p1_dog, dtype: int64
```

```
In [14]: i_predictions['p2_dog'].value_counts()
```

```
Out[14]: True     1553
         False     522
         Name: p2_dog, dtype: int64
```

```
In [15]: i_predictions['p3_dog'].value_counts()
```

```
Out[15]: True     1499
         False     576
         Name: p3_dog, dtype: int64

In [16]: #Quality Issue number 8: P1 is the most likely prediction for what breed (if any) a dog
         #Therefore any false rows can be removed

In [17]: #checking for duplicates in twitter archive
         df[df['tweet_id'].duplicated()].count()

Out[17]: tweet_id                    0
         in_reply_to_status_id       0
         in_reply_to_user_id         0
         timestamp                   0
         source                      0
         text                        0
         retweeted_status_id         0
         retweeted_status_user_id    0
         retweeted_status_timestamp  0
         expanded_urls               0
         rating_numerator            0
         rating_denominator          0
         name                        0
         doggo                       0
         floofer                     0
         pupper                      0
         puppo                       0
         dtype: int64

In [18]: #Checking for duplicates in the additional archive
         twitter_additional_archive[twitter_additional_archive['id'].duplicated()].count()

Out[18]: contributors                  0
         coordinates                   0
         created_at                    0
         display_text_range            0
         entities                      0
         extended_entities             0
         favorite_count                0
         favorited                     0
         full_text                     0
         geo                           0
         id                            0
         id_str                        0
         in_reply_to_screen_name       0
         in_reply_to_status_id         0
         in_reply_to_status_id_str     0
         in_reply_to_user_id           0
         in_reply_to_user_id_str       0
```

```
is_quote_status                 0
lang                            0
place                           0
possibly_sensitive              0
possibly_sensitive_appealable   0
quoted_status                   0
quoted_status_id                0
quoted_status_id_str            0
quoted_status_permalink         0
retweet_count                   0
retweeted                       0
retweeted_status                0
source                          0
truncated                       0
user                            0
dtype: int64
```

In [19]:
```
#Checking for duplcaites in Predictions.
i_predictions[i_predictions['tweet_id'].duplicated()].count()
```

Out[19]:
```
tweet_id    0
jpg_url     0
img_num     0
p1          0
p1_conf     0
p1_dog      0
p2          0
p2_conf     0
p2_dog      0
p3          0
p3_conf     0
p3_dog      0
dtype: int64
```

In [20]:
```
missing_values = df['tweet_id'].isnull().values
for i in range(len(df['tweet_id'])):
    if missing_values[i] == True:
        i
#Since nothing was out put there is no missing data
```

In [21]:
```
df['rating_numerator'].value_counts()
#Most ratings are between 12 and 13 however some ratings do not follow the
#weratedogs rating system since they are below 10 on the other hand some
#rating are significantly high. Further investigation is required on thier
#legitimacy before they are removed.
```

Out[21]:
```
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
420       2
0         2
15        2
75        2
80        1
20        1
24        1
26        1
44        1
50        1
60        1
165       1
84        1
88        1
144       1
182       1
143       1
666       1
960       1
1776      1
17        1
27        1
45        1
99        1
121       1
204       1
Name: rating_numerator, dtype: int64
```

```
In [22]: df['name'].value_counts()
         #Some of the names are errors, where they may be only one letter or words
         #like 'the'.
```

```
Out[22]: None          745
         a              55
         Charlie        12
         Cooper         11
         Oliver         11
```

Lucy	11
Tucker	10
Lola	10
Penny	10
Winston	9
Bo	9
Sadie	8
the	8
Bailey	7
Daisy	7
an	7
Toby	7
Buddy	7
Rusty	6
Dave	6
Oscar	6
Bella	6
Scout	6
Jax	6
Stanley	6
Jack	6
Milo	6
Leo	6
Koda	6
Phil	5
...	
Kara	1
Lacy	1
Iggy	1
Roscoe	1
DayZ	1
Brandonald	1
Mosby	1
Cal	1
Ole	1
Chadrick	1
Lillie	1
Superpup	1
Stuart	1
Gordon	1
Sprout	1
Spencer	1
Tassy	1
Major	1
Dunkin	1
Combo	1
Nimbus	1
Dwight	1

```
Tess            1
Eleanor         1
Tessa           1
Julio           1
Jiminus         1
Buckley         1
Molly           1
Sweet           1
Name: name, Length: 957, dtype: int64
```

3 Cleaning

Quality Issue 1: Noticed that tweet_id is a integar but twitter recommneds it to be a string

```
In [23]: df_copy = df
```

```
In [24]: #df['tweet_id'] = df['tweet_id'].apply(str)
         df_copy.tweet_id = df_copy.tweet_id.astype(str)
         df_copy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                    2356 non-null object
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(2), object(11)
memory usage: 313.0+ KB
```

```
In [25]: x = df_copy.iloc[0,0]
         type(x)
```

```
Out[25]: str
```

Quality Issue 2: retweet_status_id is a 181 and so those data points must be removed

```
In [26]: df_copy.retweeted_status_id = df_copy.retweeted_status_id.fillna(0)
         a =df_copy.index[df_copy['retweeted_status_id'] > 0]
         df_copy = df_copy.drop(df_copy.index[a])
         df_copy.shape
```

```
Out[26]: (2175, 17)
```

Quality Issue 3: Once retweeted rows are removed, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp are obsolete and should be removed.

```
In [27]: df_copy =df_copy.drop(['retweeted_status_id', 'retweeted_status_user_id', 'retweeted_st
         df_copy.head()
```

```
Out[27]:             tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
         0  892420643555336193                    NaN                  NaN
         1  892177421306343426                    NaN                  NaN
         2  891815181378084864                    NaN                  NaN
         3  891689557279858688                    NaN                  NaN
         4  891327558926688256                    NaN                  NaN

                        timestamp  \
         0  2017-08-01 16:23:56 +0000
         1  2017-08-01 00:17:27 +0000
         2  2017-07-31 00:18:03 +0000
         3  2017-07-30 15:58:51 +0000
         4  2017-07-29 16:00:24 +0000

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

                                                  text  \
         0  This is Phineas. He's a mystical boy. Only eve...
         1  This is Tilly. She's just checking pup on you...
         2  This is Archie. He is a rare Norwegian Pouncin...
         3  This is Darla. She commenced a snooze mid meal...
         4  This is Franklin. He would like you to stop ca...

                                         expanded_urls  rating_numerator  \
         0  https://twitter.com/dog_rates/status/892420643...                13
         1  https://twitter.com/dog_rates/status/892177421...                13
         2  https://twitter.com/dog_rates/status/891815181...                12
         3  https://twitter.com/dog_rates/status/891689557...                13
```

```
4  https://twitter.com/dog_rates/status/891327558...                12

   rating_denominator      name doggo floofer pupper puppo
0                  10   Phineas  None    None   None  None
1                  10     Tilly  None    None   None  None
2                  10    Archie  None    None   None  None
3                  10     Darla  None    None   None  None
4                  10  Franklin  None    None   None  None
```

Quality Issue 4: retweet_status_id is a 166 and so those data points must be removed

```
In [28]: twitter_additional_archive_copy =twitter_additional_archive

In [29]: twitter_additional_archive_copy.retweeted_status = twitter_additional_archive_copy.retw
         b =twitter_additional_archive_copy.index[twitter_additional_archive_copy['retweeted_sta
         twitter_additional_archive_copy = twitter_additional_archive_copy.drop(twitter_addition
         twitter_additional_archive_copy.shape

Out[29]: (2171, 32)
```

Quality Issue 5: Once retweeted rows are removed, retweet status id and retweeted are obsolete and should be removed.

```
In [30]: twitter_additional_archive_copy =twitter_additional_archive_copy.drop(['retweeted_statu

In [31]: twitter_additional_archive_copy.head()

Out[31]:    contributors  coordinates          created_at display_text_range  \
         0           NaN          NaN 2017-08-01 16:23:56            [0, 85]
         1           NaN          NaN 2017-08-01 00:17:27           [0, 138]
         2           NaN          NaN 2017-07-31 00:18:03           [0, 121]
         3           NaN          NaN 2017-07-30 15:58:51            [0, 79]
         4           NaN          NaN 2017-07-29 16:00:24           [0, 138]

                                                 entities  \
         0  {'hashtags': [], 'symbols': [], 'user_mentions...
         1  {'hashtags': [], 'symbols': [], 'user_mentions...
         2  {'hashtags': [], 'symbols': [], 'user_mentions...
         3  {'hashtags': [], 'symbols': [], 'user_mentions...
         4  {'hashtags': [{'text': 'BarkWeek', 'indices': ...

                                       extended_entities  favorite_count  \
         0  {'media': [{'id': 892420639486877696, 'id_str'...          37500
         1  {'media': [{'id': 892177413194625024, 'id_str'...          32245
         2  {'media': [{'id': 891815175371796480, 'id_str'...          24288
         3  {'media': [{'id': 891689552724799489, 'id_str'...          40822
         4  {'media': [{'id': 891327551943041024, 'id_str'...          39048

            favorited                                          full_text  geo  \
```

```
0      False  This is Phineas. He's a mystical boy. Only eve...  NaN
1      False  This is Tilly. She's just checking pup on you...  NaN
2      False  This is Archie. He is a rare Norwegian Pouncin...  NaN
3      False  This is Darla. She commenced a snooze mid meal...  NaN
4      False  This is Franklin. He would like you to stop ca...  NaN

                          ...                          possibly_sensitive  \
0                         ...                                         0.0
1                         ...                                         0.0
2                         ...                                         0.0
3                         ...                                         0.0
4                         ...                                         0.0

   possibly_sensitive_appealable quoted_status  quoted_status_id  \
0                            0.0           NaN               NaN
1                            0.0           NaN               NaN
2                            0.0           NaN               NaN
3                            0.0           NaN               NaN
4                            0.0           NaN               NaN

   quoted_status_id_str  quoted_status_permalink  retweet_count  \
0                   NaN                      NaN           8174
1                   NaN                      NaN           6050
2                   NaN                      NaN           4000
3                   NaN                      NaN           8325
4                   NaN                      NaN           9018

                                             source truncated  \
0  <a href="http://twitter.com/download/iphone" r...     False
1  <a href="http://twitter.com/download/iphone" r...     False
2  <a href="http://twitter.com/download/iphone" r...     False
3  <a href="http://twitter.com/download/iphone" r...     False
4  <a href="http://twitter.com/download/iphone" r...     False

                                                user
0  {'id': 4196983835, 'id_str': '4196983835', 'na...
1  {'id': 4196983835, 'id_str': '4196983835', 'na...
2  {'id': 4196983835, 'id_str': '4196983835', 'na...
3  {'id': 4196983835, 'id_str': '4196983835', 'na...
4  {'id': 4196983835, 'id_str': '4196983835', 'na...

[5 rows x 30 columns]
```

Quality Issue 6: Noticed that id is a integar but twitter recommneds it to be a string

```
In [32]: twitter_additional_archive_copy.id = twitter_additional_archive_copy.id.astype(str)
         twitter_additional_archive_copy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Int64Index: 2171 entries, 0 to 2336
Data columns (total 30 columns):
contributors                     0 non-null float64
coordinates                      0 non-null float64
created_at                       2171 non-null datetime64[ns]
display_text_range               2171 non-null object
entities                         2171 non-null object
extended_entities                1990 non-null object
favorite_count                   2171 non-null int64
favorited                        2171 non-null bool
full_text                        2171 non-null object
geo                              0 non-null float64
id                               2171 non-null object
id_str                           2171 non-null int64
in_reply_to_screen_name          77 non-null object
in_reply_to_status_id            77 non-null float64
in_reply_to_status_id_str        77 non-null float64
in_reply_to_user_id              77 non-null float64
in_reply_to_user_id_str          77 non-null float64
is_quote_status                  2171 non-null bool
lang                             2171 non-null object
place                            1 non-null object
possibly_sensitive               2113 non-null float64
possibly_sensitive_appealable    2113 non-null float64
quoted_status                    24 non-null object
quoted_status_id                 25 non-null float64
quoted_status_id_str             25 non-null float64
quoted_status_permalink          25 non-null object
retweet_count                    2171 non-null int64
source                           2171 non-null object
truncated                        2171 non-null bool
user                             2171 non-null object
dtypes: bool(3), datetime64[ns](1), float64(11), int64(3), object(12)
memory usage: 481.3+ KB
```

Quality Issue 7: id_str is the same as id and so should be removed.

```
In [33]: twitter_additional_archive_copy =twitter_additional_archive_copy.drop(['id_str'], axis=
```

```
In [34]: twitter_additional_archive_copy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2171 entries, 0 to 2336
Data columns (total 29 columns):
contributors                     0 non-null float64
coordinates                      0 non-null float64
created_at                       2171 non-null datetime64[ns]
display_text_range               2171 non-null object
```

```
entities                          2171 non-null object
extended_entities                 1990 non-null object
favorite_count                    2171 non-null int64
favorited                         2171 non-null bool
full_text                         2171 non-null object
geo                               0 non-null float64
id                                2171 non-null object
in_reply_to_screen_name           77 non-null object
in_reply_to_status_id             77 non-null float64
in_reply_to_status_id_str         77 non-null float64
in_reply_to_user_id               77 non-null float64
in_reply_to_user_id_str           77 non-null float64
is_quote_status                   2171 non-null bool
lang                              2171 non-null object
place                             1 non-null object
possibly_sensitive                2113 non-null float64
possibly_sensitive_appealable     2113 non-null float64
quoted_status                     24 non-null object
quoted_status_id                  25 non-null float64
quoted_status_id_str              25 non-null float64
quoted_status_permalink           25 non-null object
retweet_count                     2171 non-null int64
source                            2171 non-null object
truncated                         2171 non-null bool
user                              2171 non-null object
dtypes: bool(3), datetime64[ns](1), float64(11), int64(2), object(12)
memory usage: 464.3+ KB
```

Quality Issue number 8: P1 is the most likely prediction for what breed (if any) a dog maybe. Therefore any false rows can be removed

```
In [35]: i_predictions_copy = i_predictions
```

```
In [36]: c =i_predictions_copy.index[i_predictions_copy['p1_dog'] == False]
         i_predictions_copy = i_predictions_copy.drop(i_predictions_copy.index[c])
         i_predictions_copy.shape
```

```
Out[36]: (1532, 12)
```

4 Tidying

Tidying item 1: Remove unnecessary columns

```
In [37]: df_copy = df_copy.drop(['in_reply_to_status_id','in_reply_to_user_id'],axis=1)
```

```
In [38]: df_copy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 12 columns):
tweet_id              2175 non-null object
timestamp             2175 non-null object
source                2175 non-null object
text                  2175 non-null object
expanded_urls         2117 non-null object
rating_numerator      2175 non-null int64
rating_denominator    2175 non-null int64
name                  2175 non-null object
doggo                 2175 non-null object
floofer               2175 non-null object
pupper                2175 non-null object
puppo                 2175 non-null object
dtypes: int64(2), object(10)
memory usage: 220.9+ KB
```

```
In [39]: twitter_additional_archive_copy = twitter_additional_archive_copy.drop(['contributors',
                                                                                 'extended_entit
                                                                                 'in_reply_to_st
                                                                                 'in_reply_to_us
                                                                                 'quoted_status_
```

```
In [40]: twitter_additional_archive_copy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2171 entries, 0 to 2336
Data columns (total 13 columns):
created_at                      2171 non-null datetime64[ns]
entities                        2171 non-null object
favorite_count                  2171 non-null int64
favorited                       2171 non-null bool
id                              2171 non-null object
is_quote_status                 2171 non-null bool
lang                            2171 non-null object
possibly_sensitive              2113 non-null float64
possibly_sensitive_appealable   2113 non-null float64
retweet_count                   2171 non-null int64
source                          2171 non-null object
truncated                       2171 non-null bool
user                            2171 non-null object
dtypes: bool(3), datetime64[ns](1), float64(2), int64(2), object(5)
memory usage: 192.9+ KB
```

```
In [41]: i_predictions_copy = i_predictions_copy.drop(['img_num'],axis=1)
```

```
In [42]: i_predictions_copy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1532 entries, 0 to 2073
Data columns (total 11 columns):
tweet_id    1532 non-null int64
jpg_url     1532 non-null object
p1          1532 non-null object
p1_conf     1532 non-null float64
p1_dog      1532 non-null bool
p2          1532 non-null object
p2_conf     1532 non-null float64
p2_dog      1532 non-null bool
p3          1532 non-null object
p3_conf     1532 non-null float64
p3_dog      1532 non-null bool
dtypes: bool(3), float64(3), int64(1), object(4)
memory usage: 112.2+ KB
```

Tidying item 2:Joining data

```
In [43]: df_copy['created_at'],df_copy['favorite_count'],df_copy['favorited'],df_copy['is_quote_
         df_copy['possibly_sensitive'],df_copy['possibly_sensitive_appealable'],df_copy['retweet
         df_copy['jpg_url'],df_copy['p1'],df_copy['p1_conf'],df_copy['p1_dog']= ["","","",""]
         df_copy['p2'],df_copy['p2_conf'],df_copy['p2_dog'],df_copy['p3'],df_copy['p3_conf'],df_
```

```
In [44]: df_copy.head(2)
```

```
Out[44]:              tweet_id                   timestamp  \
         0  892420643555336193  2017-08-01 16:23:56 +0000
         1  892177421306343426  2017-08-01 00:17:27 +0000

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

                                                     text  \
         0  This is Phineas. He's a mystical boy. Only eve...
         1  This is Tilly. She's just checking pup on you...

                                            expanded_urls  rating_numerator  \
         0  https://twitter.com/dog_rates/status/892420643...                13
         1  https://twitter.com/dog_rates/status/892177421...                13

            rating_denominator     name doggo floofer  ...   jpg_url p1 p1_conf p1_dog  \
         0                  10  Phineas  None    None  ...
         1                  10    Tilly  None    None  ...
```

```
  p2 p2_conf p2_dog p3 p3_conf p3_dog
0
1

[2 rows x 31 columns]
```

```
In [45]: df_copy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 31 columns):
tweet_id                          2175 non-null object
timestamp                         2175 non-null object
source                            2175 non-null object
text                              2175 non-null object
expanded_urls                     2117 non-null object
rating_numerator                  2175 non-null int64
rating_denominator                2175 non-null int64
name                              2175 non-null object
doggo                             2175 non-null object
floofer                           2175 non-null object
pupper                            2175 non-null object
puppo                             2175 non-null object
created_at                        2175 non-null object
favorite_count                    2175 non-null object
favorited                         2175 non-null object
is_quote_status                   2175 non-null object
lang                              2175 non-null object
possibly_sensitive                2175 non-null object
possibly_sensitive_appealable     2175 non-null object
retweet_count                     2175 non-null object
truncated                         2175 non-null object
jpg_url                           2175 non-null object
p1                                2175 non-null object
p1_conf                           2175 non-null object
p1_dog                            2175 non-null object
p2                                2175 non-null object
p2_conf                           2175 non-null object
p2_dog                            2175 non-null object
p3                                2175 non-null object
p3_conf                           2175 non-null object
p3_dog                            2175 non-null object
dtypes: int64(2), object(29)
memory usage: 543.8+ KB
```

```
In [46]: df_copy = df_copy.reset_index(drop=True)
         df_copy.shape
```

```
Out[46]: (2175, 31)
```

```
In [47]: missing_id = []
         missing_id_index = []
         for i in range(len(df_copy['tweet_id'])):
             try:
                 y=twitter_additional_archive_copy.loc[twitter_additional_archive_copy['id'] ==
                 df_copy.iloc[i,12]= twitter_additional_archive_copy.created_at[y[0]]
                 df_copy.iloc[i,13]= twitter_additional_archive_copy.favorite_count[y[0]]
                 df_copy.iloc[i,14]= twitter_additional_archive_copy.favorited[y[0]]
                 df_copy.iloc[i,15]= twitter_additional_archive_copy.is_quote_status[y[0]]
                 df_copy.iloc[i,16]= twitter_additional_archive_copy.lang[y[0]]
                 df_copy.iloc[i,17]= twitter_additional_archive_copy.possibly_sensitive[y[0]]
                 df_copy.iloc[i,18]= twitter_additional_archive_copy.possibly_sensitive_appealab
                 df_copy.iloc[i,19]= twitter_additional_archive_copy.retweet_count[y[0]]
                 df_copy.iloc[i,20]= twitter_additional_archive_copy.truncated[y[0]]

                 df_copy.iloc[i,21]= i_predictions_copy.jpg_url[y[0]]
                 df_copy.iloc[i,22]= i_predictions_copy.p1[y[0]]
                 df_copy.iloc[i,23]= i_predictions_copy.p1_conf[y[0]]
                 df_copy.iloc[i,24]= i_predictions_copy.p1_dog[y[0]]
                 df_copy.iloc[i,25]= i_predictions_copy.p2[y[0]]
                 df_copy.iloc[i,26]= i_predictions_copy.p2_conf[y[0]]
                 df_copy.iloc[i,27]= i_predictions_copy.p2_dog[y[0]]
                 df_copy.iloc[i,28]= i_predictions_copy.p3[y[0]]
                 df_copy.iloc[i,29]= i_predictions_copy.p3_conf[y[0]]
                 df_copy.iloc[i,30]= i_predictions_copy.p3_dog[y[0]]
             except:
                 missing_id.append(df_copy.tweet_id[i])
                 missing_id_index.append(i)
                 print(df_copy.tweet_id[i])
```

```
890971913173991426
890609185150312448
888804989199671297
888554962724278272
887517139158093824
887473957103951883
886983233522544640
886267009285017600
885528943205470208
883838122936631299
883360690899218434
882268110199369728
882045870035918850
881906580714921986
881666595344535552
881268444196462592
```

879008229531029506
877556246731214848
875097192612077568
874012996292530176
873580283840344065
873213775632977920
872820683541237760
872486979161796608
872261713294495745
871762521631449091
871102520638267392
871032628920680449
870804317367881728
870308999962521604
870063196459192321
869702957897576449
869227993411051520
868880397819494401
867774946302451713
867051520902168576
866334964761202691
865718153858494464
865359393868664832
863427515083354112
863079547188785154
862831371563274240
861288531465048066
861005113778896900
859074603037188101
858471635011153920
858107933456039936
857989990357356544
857393404942143489
857263160327368704
857029823797047296
856282028240666624
855857698524602368
855851453814013952
853760880890318849
853299958564483072
852672615818899456
852553447878664193
852226086759018497
851591660324737024
850380195714523136
850333567704068097
850019790995546112
849776966551130114

849412302885593088
849051919805034497
848324959059550208
847617282490613760
847116187444137987
846514051647705089
846505985330044928
846139713627017216
845812042753855489
845677943972139009
845397057150107648
844979544864018432
844973813909606400
844704788403113984
844580511645339650
844223788422217728
843981021012017153
843856843873095681
842163532590374912
842115215311396866
841077006473256960
840698636975636481
838952994649550848
838476387338051585
838083903487373313
837471256429613056
836753516572119041
836677758902222849
836260088725786625
835574547218894849
835264098648616962
835246439529840640
835152434251116546
834209720923721728
834089966724603904
834086379323871233
833479644947025920
833124694597443584
832757312314028032
832645525019123713
832636094638288896
832397543355072512
831911600680497154
831650051525054464
831262627380748289
829141528400556032
829011960981237760
828650029636317184

828409743546925057
828376505180889089
828046555563323392
828011680017821696
827933404142436356
827324948884643840
826958653328592898
826848821049180160
826476773533745153
825147591692263424
825026590719483904
824297048279236611
824025158776213504
823581115634085888
822610361945911296
821522889702862852
821153421864615936
820078625395449857
819588359383371776
819006400881917954
819004803107983360
818614493328580609
818145370475810820
817827839487737858
817536400337801217
817171292965273600
817056546584727552
816816676327063552
814986499976527872
814638523311648768
814530161257443328
813910438903693312
813812741911748608
813202720496779264
813142292504645637
813112105746448384
812709060537683968
810896069567610880
808838249661788160
808733504066486276
808344865868283904
807010152071229440
806576416489959424
806542213899489280
806219024703037440
805932879469572096
805520635690676224
805207613751304193

804475857670639616
803638050916102144
802952499103731712
802265048156610565
801854953262350336
801115127852503040
799063482566066176
798933969379225600
797545162159308800
797165961484890113
796759840936919040
796149749086875649
796116448414461957
796031486298386433
795464331001561088
794332329137291264
793962221541933056
793845145112371200
793271401113350145
793241302385262592
793135492858580992
792773781206999040
791774931465953280
790987426131050500
790581949425475584
790277117346975746
789986466051088384
789903600034189313
789530877013393408
789137962068021249
788765914992902144
788178268662984705
786363235746385920
785533386513321988
784183165795655680
783821107061198850
783466772167098368
783334639985389568
782969140009107456
781955203444699136
781661882474196992
781308096455073793
780858289093574656
779377524342161408
779056095788752897
778748913645780993
778408200802557953
777621514455814149

777189768882946048
776113305656188928
774757898236878852
773985732834758656
773704687002451968
773247561583001600
772581559778025472
772102971039580160
771908950375665664
771380798096281600
771102124360998913
770069151037685760
768970937022709760
767500508068192258
766693177336135680
766423258543644672
766008592277377025
765395769549590528
764259802650378240
763956972077010945
763837565564780549
761334018830917632
761292947749015552
760656994973933572
760252756032651264
760190180481531904
759923798737051648
758041019896193024
757400162377592832
756303284449767430
756275833623502848
755206590534418437
755110668769038337
754747087846248448
754120377874386944
754011816964026368
752519690950500352
752334515931054080
752173152931807232
751830394383790080
751583847268179968
751205363882532864
751132876104687617
750429297815552001
750383411068534784
750086836815486976
749996283729883136
748932637671223296

748705597323898880
748692773788876800
747885874273214464
747816857231626240
747594051852075008
747204161125646336
746818907684614144
745057283344719872
744334592493166593
743595368194129920
743253157753532416
743210557239623680
742534281772302336
742528092657332225
742423170473463808
742385895052087300
742161199639494656
741793263812808706
741438259667034112
740995100998766593
739623569819336705
739485634323156992
738883359779196928
736365877722001409
735648611367784448
735635087207878657
733828123016450049
733482008106668032
732726085725589504
730211855403241472
729854734790754305
729823566028484608
727685679342333952
727524757080539137
726887082820554753
725842289046749185
725786712245440512
724771698126512129
724046343203856385
724004602748780546
723912936180330496
722974582966214656
722613351520608256
720785406564900865
720389942216527872
718939241951195136
718246886998687744
717841801130979328

717790033953034240
716285507865542656
715220193576927233
715009755312439296
714982300363173890
714214115368108032
714141408463036416
713900603437621249
713761197720473600
712809025985978368
712717840512598017
712668654853337088
711732680602345472
711652651650457602
711363825979756544
710588934686908417
710283270106132480
710269109699739648
710140971284037632
709409458133323776
709179584944730112
709158332880297985
708853462201716736
708810915978854401
708349470027751425
708130923141795840
707610948723478529
706681918348251136
706593038911545345
706538006853918722
706166467411222528
705786532653883392
705591895322394625
705239209544720384
705066031337840642
704859558691414016
704480331685040129
704364645503647744
704347321748819968
704113298707505153
704054845121142784
703774238772166656
703769065844768768
703268521220972544
702276748847800320
702217446468493312
701889187134500865
701545186879471618

700864154249383937
700796979434098688
700747788515020802
700462010979500032
700167517596164096
700029284593901568
698989035503689728
698710712454139905
698703483621523456
697630435728322560
697616773278015490
697516214579523584
697482927769255936
697270446429966336
696488710901260288
696100768806522880
695816827381944320
695409464418041856
694925794720792577
694905863685980160
694352839993344000
694342028726001664
693280720173801472
692905862751522816
692187005137076224
692041934689402880
691820333922455552
691444869282295808
690932576555528194
690649993829576704
690607260360429569
690400367696297985
690360449368465409
689977555533848577
689661964914655233
689289219123089408
689143371370250240
688898160958271489
688828561667567616
688804835492233216
687841446767013888
687826841265172480
687807801670897665
687480748861947905
686947101016735744
686730991906516992
686618349602762752
685532292383666176

684940049151070208
684902183876321280
684880619965411328
684567543613382656
684222868335505415
684147889187209216
683852578183077888
683773439333797890
683357973142474752
683111407806746624
682750546109968385
681891461017812993
681694085539872773
681339448655802368
681297372102656000
681281657291280384
680805554198020098
680440374763077632
680191257256136705
680100725817409536
680055455951884288
679854723806179328
679729593985699840
679722016581222400
679475951516934144
679462823135686656
679158373988876288
678424312106393600
678399652199309312
678396796259975168
678341075375947776
678023323247357953
677716515794329600
677700003327029250
677331501395156992
677228873407442944
676975532580409345
676957860086095872
676237365392908289
676098748976615425
676089483918516224
675888385639251968
675870721063669760
675853064436391936
675845657354215424
675740360753160193
675710890956750848
675707330206547968

675706639471788032
675534494439489536
675501075957489664
675372240448454658
675349384339542016
675334060156301312
675149409102012420
675146535592706048
675113801096802304
675109292475830276
675003128568291329
674781762103414784
674764817387900928
674664755118911488
674606911342424069
674468880899788800
674416750885273600
674410619106390016
674318007229923329
674271431610523648
674269164442398721
674262580978937856
674082852460433408
674075285688614912
674063288070742018
674042553264685056
673688752737402881
673686845050527744
673636718965334016
673612854080196609
673576835670777856
673359818736984064
673352124999274496
673345638550134785
673240798075449344
672997845381865473
672828477930868736
672604026190569472
672538107540070400
672475084225949696
672245253877968896
671768281401958400
671763349865160704
671544874165002241
671542985629241344
671528761649688577
671151324042559489
671115716440031232

670786190031921152
670780561024270336
670778058496974848
670764103623966721
670755717859713024
670733412878163972
670727704916926465
670717338665226240
670704688707301377
670691627984359425
670679630144274432
670676092097810432
670668383499735048
670474236058800128
670468609693655041
670465786746662913
670452855871037440
670449342516494336
670444955656130560
670442337873600512
670435821946826752
670434127938719744
670433248821026816
670428280563085312
670427002554466305
670421925039075328
670420569653809152
670417414769758208
670411370698022913
670408998013820928
670403879788544000
670385711116361728
670374371102445568
670361874861563904
670338931251150849
670319130621435904
670303360680108032
670290420111441920
670093938074779648
670086499208155136
670079681849372674
670073503555706880
670069087419133954
670061506722140161
670055038660800512
670046952931721218
670040295598354432
670037189829525505

670003130994700288
669993076832759809
669972011175813120
669970042633789440
669942763794931712
669926384437997569
669923323644657664
669753178989142016
669749430875258880
669684865554620416
669683899023405056
669682095984410625
669680153564442624
669661792646373376
669625907762618368
669603084620980224
669597912108789760
669583744538451968
669573570759163904
669571471778410496
669567591774625800
669564461267722241
669393256313184256
669375718304980992
669371483794317312
669367896104181761
669363888236994561
669359674819481600
669354382627049472
669353438988365824
669351434509529089
669328503091937280
669327207240699904
669324657376567296
669216679721873412
669214165781868544
669203728096960512
669037058363662336
669015743032369152
669006782128353280
669000397445533696
668994913074286592
668992363537309700
668989615043424256
668988183816871936
668986018524233728
668981893510119424
668979806671884288

668975677807423489
668967877119254528
668960084974809088
668955713004314625
668932921458302977
668902994700836864
668892474547511297
668872652652679168
668852170888998912
668826086256599040
668815180734689280
668779399630725120
668655139528511488
668645506898350081
668643542311546881
668641109086707712
668636665813057536
668633411083464705
668631377374486528
668627278264475648
668625577880875008
668623201287675904
668620235289837568
668614819948453888
668587383441514497
668567822092664832
668544745690562560
668542336805281792
668537837512433665
668528771708952576
668507509523615744
668496999348633600
668484198282485761
668480044826800133
668466899341221888
668297328638447616
668291999406125056
668286279830867968
668274247790391296
668268907921326080
668256321989451776
668248472370458624
668237644992782336
668226093875376128
668221241640230912
668204964695683073
668190681446379520
668171859951755264

668154635664932864
668142349051129856
668113020489474048
667937095915278337
667924896115245057
667915453470232577
667911425562669056
667902449697558528
667886921285246976
667885044254572545
667878741721415682
667873844930215936
667866724293877760
667861340749471744
667832474953625600
667806454573760512
667801013445750784
667793409583771648
667782464991965184
667773195014021121
667766675769573376
667728196545200128
667724302356258817
667549055577362432
667546741521195010
667544320556335104
667538891197542400
667534815156183040
667530908589760512
667524857454854144
667517642048163840
667509364010450944
667502640335572993
667495797102141441
667491009379606528
667470559035432960
667455448082227200
667453023279554560
667443425659232256
667437278097252352
667435689202614272
667405339315146752
667393430834667520
667369227918143488
667211855547486208
667200525029539841
667192066997374976
667188689915760640

667182792070062081
667177989038297088
667176164155375616
667174963120574464
667171260800061440
667165590075940865
667160273090932737
667152164079423490
667138269671505920
667119796878725120
667090893657276420
667073648344346624
667070482143944705
667065535570550784
667062181243039745
667044094246576128
667012601033924608
666996132027977728
666983947667116034
666837028449972224
666835007768551424
666826780179869698
666817836334096384
666804364988780544
666786068205871104
666781792255496192
666776908487630848
666739327293083650
666701168228331520
666691418707132416
666649482315059201
666644823164719104
666454714377183233
666447344410484738
666437273139982337
666435652385423360
666430724426358785
666428276349472768
666421158376562688
666418789513326592
666411507551481857
666407126856765440
666396247373291520
666373753744588802
666362758909284353
666353288456101888
666345417576210432
666337882303524864

```
666293911632134144
666287406224695296
666273097616637952
666268910803644416
666104133288665088
666102155909144576
666099513787052032
666094000022159362
666082916733198337
666073100786774016
666071193221509120
666063827256086533
666058600524156928
666057090499244032
666055525042405380
666051853826850816
666050758794694657
666049248165822465
666044226329800704
666033412701032449
666029285002620928
666020888022790149
```

```
In [48]: df_copy=df_copy.drop(missing_id_index, axis=0)
```

```
In [49]: #reindex
         df_copy = df_copy.reset_index(drop=True)
```

```
In [50]: df_copy['favorite_count'] = df_copy['favorite_count'].astype(int)
         df_copy['favorited'] = df_copy['favorited'].astype('bool')
         df_copy['is_quote_status'] = df_copy['is_quote_status'].astype('bool')
         df_copy['possibly_sensitive'] = df_copy['possibly_sensitive'].astype(float)
         df_copy['possibly_sensitive_appealable'] = df_copy['possibly_sensitive_appealable'].ast
         df_copy['retweet_count'] = df_copy['retweet_count'].astype(int)
         df_copy['truncated'] = df_copy['truncated'].astype('bool')

         df_copy['p1_conf'] = df_copy['p1_conf'].astype(float)
         df_copy['p1_dog'] = df_copy['p1_dog'].astype('bool')
         df_copy['p2_conf'] = df_copy['p2_conf'].astype(float)
         df_copy['p2_dog'] = df_copy['p2_dog'].astype('bool')
         df_copy['p3_conf'] = df_copy['p3_conf'].astype(float)
         df_copy['p3_dog'] = df_copy['p3_dog'].astype('bool')
```

5 Storing, Analyzing, and Visualizing

5.1 Storing

```
In [51]: df_copy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1417 entries, 0 to 1416
Data columns (total 31 columns):
tweet_id                          1417 non-null object
timestamp                         1417 non-null object
source                            1417 non-null object
text                              1417 non-null object
expanded_urls                     1375 non-null object
rating_numerator                  1417 non-null int64
rating_denominator                1417 non-null int64
name                              1417 non-null object
doggo                             1417 non-null object
floofer                           1417 non-null object
pupper                            1417 non-null object
puppo                             1417 non-null object
created_at                        1417 non-null object
favorite_count                    1417 non-null int64
favorited                         1417 non-null bool
is_quote_status                   1417 non-null bool
lang                              1417 non-null object
possibly_sensitive                1375 non-null float64
possibly_sensitive_appealable     1375 non-null float64
retweet_count                     1417 non-null int64
truncated                         1417 non-null bool
jpg_url                           1417 non-null object
p1                                1417 non-null object
p1_conf                           1417 non-null float64
p1_dog                            1417 non-null bool
p2                                1417 non-null object
p2_conf                           1417 non-null float64
p2_dog                            1417 non-null bool
p3                                1417 non-null object
p3_conf                           1417 non-null float64
p3_dog                            1417 non-null bool
dtypes: bool(6), float64(5), int64(4), object(16)
memory usage: 285.1+ KB
```

```
In [52]: import numpy
         df_copy.to_csv("twitter_archive_master.csv", sep=',')
```

5.2 Analyzing, and Visualizing

5.2.1 Insight 1

```
In [53]: df.query('rating_numerator == 1776')
```

```
Out[53]:                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
         979  749981277374128128                    NaN                  NaN
```

```
                       timestamp  \
979  2016-07-04 15:00:45 +0000

                                                 source  \
979  <a href="https://about.twitter.com/products/tw...

                                                   text  retweeted_status_id  \
979  This is Atticus. He's quite simply America af...                  0.0

     retweeted_status_user_id retweeted_status_timestamp  \
979                       NaN                        NaN

                                          expanded_urls  rating_numerator  \
979  https://twitter.com/dog_rates/status/749981277...              1776

     rating_denominator     name doggo floofer pupper puppo
979                  10  Atticus  None    None   None  None
```

```
In [54]: i_predictions.query('tweet_id == 749981277374128128')
```

```
Out[54]:                tweet_id                                          jpg_url  \
1270  749981277374128128  https://pbs.twimg.com/media/CmgBZ7kWcAAlzFD.jpg

      img_num       p1   p1_conf  p1_dog          p2   p2_conf  p2_dog  \
1270        1  bow_tie  0.533941   False  sunglasses  0.080822   False

             p3   p3_conf  p3_dog
1270  sunglass  0.050776   False
```

```
In [55]: print (df.iloc[979,5])
```

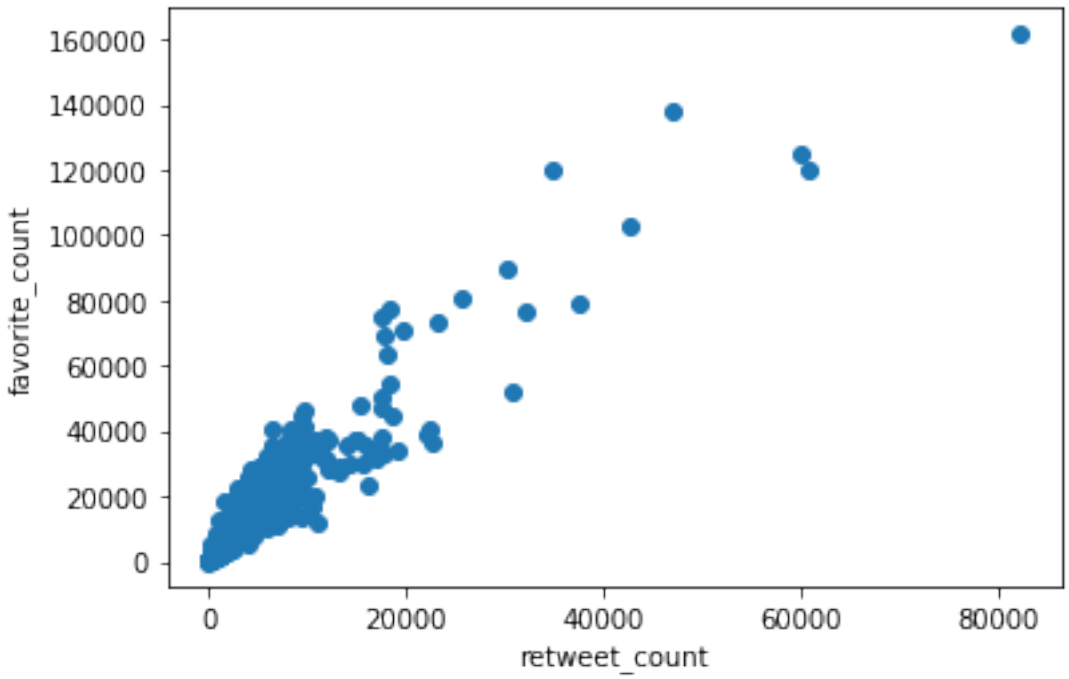
```
This is Atticus. He's quite simply America af. 1776/10 https://t.co/GRXwMxLBkh
```

Having had a look at some of the ratings there was one which stood out to me as being odd. Most of the ratings are around the 11, 12 and 13 mark but this one was at 1776. By looking more in to the Data for that row it can be sen that the post was made on 4th July and by loading in the picture it is obvious the doggo is a patriotic one. Therefore the rating of 1776 is intensional because that is when independance was offically signed and although it is a outlier it should not be disregarded.

5.2.2 Insight 2

```
In [56]: plt.scatter(df_copy['retweet_count'],df_copy['favorite_count'])
         plt.xlabel('retweet_count')
         plt.ylabel('favorite_count')
```

```
Out[56]: Text(0,0.5,'favorite_count')
```

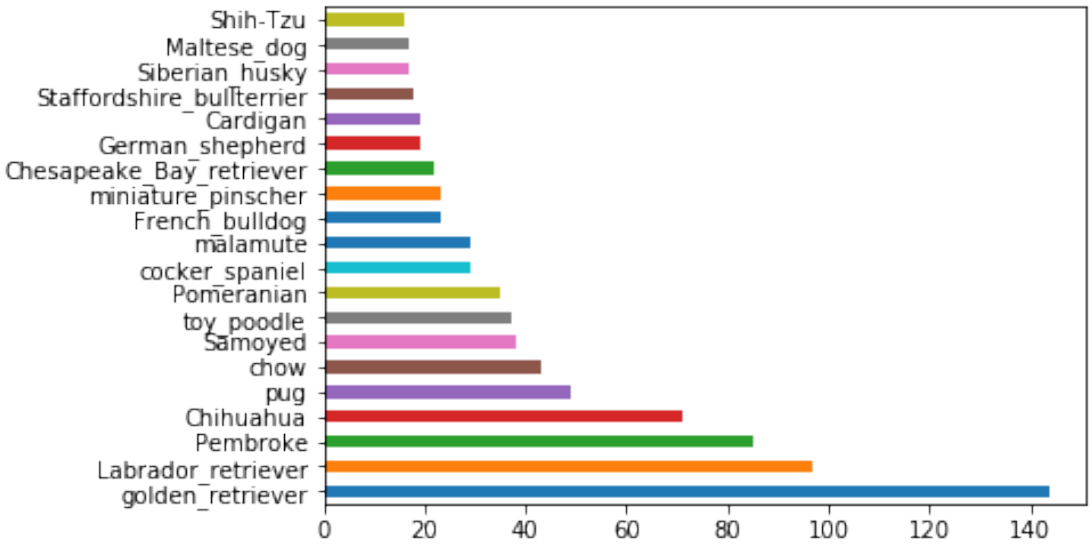


It was interesting to compare if the amount of retweets would corrolate to the amount of favorites a post gets. From the scatter plot above it is clear that there is a positive correlation between retweet_count and favorite_count; the more retweets the more favorites.

5.2.3 Insight 3

```
In [57]: df_copy['p1'].value_counts()[:20].plot(kind='barh')

Out[57]: <matplotlib.axes._subplots.AxesSubplot at 0x7f802d74f3c8>
```



This insight is useful to find which breed is the most popular based on the predictions for what breed the dog is in the image.