wrangle_act

September 18, 2022

1 Project: Wrangling and Analyze Data

1.1 Data Gathering

In the cell below, gather **all** three pieces of data for this project and load them in the notebook. **Note:** the methods required to gather each data are different. 1. Directly download the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv)

[46]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb

%matplotlib inline
```

[ ]:

```

```

[4]:

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

[4]:

```
                 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
…                    …                    …                  …
2351  666049248165822465                    NaN                  NaN
2352  666044226329800704                    NaN                  NaN
2353  666033412701032449                    NaN                  NaN
2354  666029285002620928                    NaN                  NaN
2355  666020888022790149                    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
…                          …
2351  2015-11-16 00:24:50 +0000
2352  2015-11-16 00:04:52 +0000
2353  2015-11-15 23:21:54 +0000
2354  2015-11-15 23:05:30 +0000
2355  2015-11-15 22:32:08 +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…
…                                                 …
2351  <a href="http://twitter.com/download/iphone" r…
2352  <a href="http://twitter.com/download/iphone" r…
2353  <a href="http://twitter.com/download/iphone" r…
2354  <a href="http://twitter.com/download/iphone" r…
2355  <a href="http://twitter.com/download/iphone" r…

                                                   text  retweeted_status_id  \
0     This is Phineas. He's a mystical boy. Only eve…                    NaN
1     This is Tilly. She's just checking pup on you…                    NaN
2     This is Archie. He is a rare Norwegian Pouncin…                    NaN
3     This is Darla. She commenced a snooze mid meal…                    NaN
4     This is Franklin. He would like you to stop ca…                    NaN
…                                                 …                   …
2351  Here we have a 1949 1st generation vulpix. Enj…                    NaN
2352  This is a purebred Piers Morgan. Loves to Netf…                    NaN
2353  Here is a very happy pup. Big fan of well-main…                    NaN
2354  This is a western brown Mitsubishi terrier. Up…                    NaN
2355  Here we have a Japanese Irish Setter. Lost eye…                    NaN

      retweeted_status_user_id retweeted_status_timestamp  \
0                          NaN                        NaN
1                          NaN                        NaN
2                          NaN                        NaN
3                          NaN                        NaN
4                          NaN                        NaN
…                         …                         …
2351                       NaN                        NaN
2352                       NaN                        NaN
2353                       NaN                        NaN
2354                       NaN                        NaN
2355                       NaN                        NaN
```

```
                                        expanded_urls  rating_numerator  \
0     https://twitter.com/dog_rates/status/892420643…                13
1     https://twitter.com/dog_rates/status/892177421…                13
2     https://twitter.com/dog_rates/status/891815181…                12
3     https://twitter.com/dog_rates/status/891689557…                13
4     https://twitter.com/dog_rates/status/891327558…                12
…                                                   …               …
2351  https://twitter.com/dog_rates/status/666049248…                 5
2352  https://twitter.com/dog_rates/status/666044226…                 6
2353  https://twitter.com/dog_rates/status/666033412…                 9
2354  https://twitter.com/dog_rates/status/666029285…                 7
2355  https://twitter.com/dog_rates/status/666020888…                 8

      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
…                    …         …     …       …     …    …
2351                  10      None  None    None   None  None
2352                  10         a  None    None   None  None
2353                  10         a  None    None   None  None
2354                  10         a  None    None   None  None
2355                  10      None  None    None   None  None

[2356 rows x 17 columns]
```

2. Use the Requests library to download the tweet image prediction (image_predictions.tsv)

[4]:

```
import requests as rq
import os
```

[5]:

```
folder_name = 'twitter_images'
if not os.path.exists(folder_name):
    os.makedirs(folder_name)
```

[6]:

```
url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/
 ↪599fd2ad_image-predictions/image-predictions.tsv'
response = requests.get(url)
response
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [6], in <cell line: 2>()
      1 url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/
 ↪599fd2ad_image-predictions/image-predictions.tsv'
```

```
----> 2 response = requests.get(url)
      3 response

NameError: name 'requests' is not defined
```

```
[3]: tweetimages=pd.read_csv('image-predictions.tsv',sep="\t")
     tweetimages.head()
```

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

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

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

```
[8]: with open(os.path.join(folder_name, url.split('/')[-1]), mode = 'wb') as file:
         file.write(response.content)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [8], in <cell line: 1>()
      1 with open(os.path.join(folder_name, url.split('/')[-1]), mode = 'wb') a␣
  ↪file:
----> 2     file.write(response.content)

NameError: name 'response' is not defined
```

3. Use the Tweepy library to query additional data via the Twitter API (tweet_json.txt)

```
[2]: import json

     twitterjson = [json.loads(line) for line in open('tweet-json.txt','r')]
```

```python
#print(pizzaJson)
print(type(twitterjson))

tweet_json=pd.DataFrame(twitterjson)

tweet_json.head()
```

```
<class 'list'>
```

[2]:

```
                       created_at                  id              id_str  \
0  Tue Aug 01 16:23:56 +0000 2017  892420643555336193  892420643555336193
1  Tue Aug 01 00:17:27 +0000 2017  892177421306343426  892177421306343426
2  Mon Jul 31 00:18:03 +0000 2017  891815181378084864  891815181378084864
3  Sun Jul 30 15:58:51 +0000 2017  891689557279858688  891689557279858688
4  Sat Jul 29 16:00:24 +0000 2017  891327558926688256  891327558926688256

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

  display_text_range                                           entities  \
0            [0, 85]  {'hashtags': [], 'symbols': [], 'user_mentions…
1           [0, 138]  {'hashtags': [], 'symbols': [], 'user_mentions…
2           [0, 121]  {'hashtags': [], 'symbols': [], 'user_mentions…
3            [0, 79]  {'hashtags': [], 'symbols': [], 'user_mentions…
4           [0, 138]  {'hashtags': [{'text': 'BarkWeek', 'indices': …

                                   extended_entities  \
0  {'media': [{'id': 892420639486877696, 'id_str'…
1  {'media': [{'id': 892177413194625024, 'id_str'…
2  {'media': [{'id': 891815175371796480, 'id_str'…
3  {'media': [{'id': 891689552724799489, 'id_str'…
4  {'media': [{'id': 891327551943041024, 'id_str'…

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

   … favorite_count  favorited retweeted possibly_sensitive  \
0  …          39467      False     False              False
1  …          33819      False     False              False
```

```
2  …          25461      False     False                  False
3  …          42908      False     False                  False
4  …          41048      False     False                  False

  possibly_sensitive_appealable lang retweeted_status quoted_status_id  \
0                         False    en              NaN              NaN
1                         False    en              NaN              NaN
2                         False    en              NaN              NaN
3                         False    en              NaN              NaN
4                         False    en              NaN              NaN

  quoted_status_id_str  quoted_status
0                  NaN            NaN
1                  NaN            NaN
2                  NaN            NaN
3                  NaN            NaN
4                  NaN            NaN

[5 rows x 31 columns]
```

1.2 Assessing Data

In this section, detect and document at least **eight (8) quality issues and two (2) tidiness issue**. You must use **both** visual assessment programmatic assessement to assess the data.

Note: pay attention to the following key points when you access the data.

- You only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
- Assessing and cleaning the entire dataset completely would require a lot of time, and is not necessary to practice and demonstrate your skills in data wrangling. Therefore, the requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.
- The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.
- You do not need to gather the tweets beyond August 1st, 2017. You can, but note that you won't be able to gather the image predictions for these tweets since you don't have access to the algorithm used.

[10]:
```
WeRateDogs.dtypes
```

[10]:
```
tweet_id                        int64
in_reply_to_status_id         float64
in_reply_to_user_id           float64
timestamp                      object
source                         object
text                           object
retweeted_status_id           float64
```

```
retweeted_status_user_id      float64
retweeted_status_timestamp     object
expanded_urls                  object
rating_numerator                int64
rating_denominator              int64
name                           object
doggo                          object
floofer                        object
pupper                         object
puppo                          object
dtype: object
```

[11]:

```
WeRateDogs.describe()
```

[11]:

```
           tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
count  2.356000e+03           7.800000e+01         7.800000e+01
mean   7.427716e+17           7.455079e+17         2.014171e+16
std    6.856705e+16           7.582492e+16         1.252797e+17
min    6.660209e+17           6.658147e+17         1.185634e+07
25%    6.783989e+17           6.757419e+17         3.086374e+08
50%    7.196279e+17           7.038708e+17         4.196984e+09
75%    7.993373e+17           8.257804e+17         4.196984e+09
max    8.924206e+17           8.862664e+17         8.405479e+17

       retweeted_status_id  retweeted_status_user_id  rating_numerator  \
count         1.810000e+02              1.810000e+02       2356.000000
mean          7.720400e+17              1.241698e+16         13.126486
std           6.236928e+16              9.599254e+16         45.876648
min           6.661041e+17              7.832140e+05          0.000000
25%           7.186315e+17              4.196984e+09         10.000000
50%           7.804657e+17              4.196984e+09         11.000000
75%           8.203146e+17              4.196984e+09         12.000000
max           8.874740e+17              7.874618e+17       1776.000000

       rating_denominator
count         2356.000000
mean            10.455433
std              6.745237
min              0.000000
25%             10.000000
50%             10.000000
75%             10.000000
max            170.000000
```

[14]:

```
WeRateDogs['tweet_id'].duplicated().sum()
```

[14]:

```
0
```

```
[15]: WeRateDogs.doggo.unique()
```

```
[15]: array(['None', 'doggo'], dtype=object)
```

```
[16]: WeRateDogs.floofer.unique()
```

```
[16]: array(['None', 'floofer'], dtype=object)
```

```
[17]: WeRateDogs.pupper.unique()
```

```
[17]: array(['None', 'pupper'], dtype=object)
```

```
[18]: WeRateDogs.puppo.unique()
```

```
[18]: array(['None', 'puppo'], dtype=object)
```

```
[19]: WeRateDogs.query('text.str.contains("RT")')
```

```
[19]:                 tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
19    888202515573088257                    NaN                  NaN
32    886054160059072513                    NaN                  NaN
36    885311592912609280                    NaN                  NaN
68    879130579576475649                    NaN                  NaN
73    878404777348136964                    NaN                  NaN
…                    …                    …                  …
1766  678399652199309312                    NaN                  NaN
1860  675489971617296384                    NaN                  NaN
1991  672622327801233409                    NaN                  NaN
2259  667550904950915073                    NaN                  NaN
2260  667550882905632768                    NaN                  NaN

                      timestamp  \
19    2017-07-21 01:02:36 +0000
32    2017-07-15 02:45:48 +0000
36    2017-07-13 01:35:06 +0000
68    2017-06-26 00:13:58 +0000
73    2017-06-24 00:09:53 +0000
…                          …
1766  2015-12-20 02:20:55 +0000
1860  2015-12-12 01:38:53 +0000
1991  2015-12-04 03:43:54 +0000
2259  2015-11-20 03:51:52 +0000
2260  2015-11-20 03:51:47 +0000

                                                 source  \
19    <a href="http://twitter.com/download/iphone" r…
32    <a href="http://twitter.com/download/iphone" r…
```

```
36    <a href="http://twitter.com/download/iphone" r…
68    <a href="http://twitter.com/download/iphone" r…
73    <a href="http://twitter.com/download/iphone" r…
…                                                …
1766  <a href="http://twitter.com/download/iphone" r…
1860  <a href="http://twitter.com/download/iphone" r…
1991  <a href="http://twitter.com/download/iphone" r…
2259  <a href="http://twitter.com" rel="nofollow">Tw…
2260  <a href="http://twitter.com" rel="nofollow">Tw…

                                                  text  retweeted_status_id  \
19    RT @dog_rates: This is Canela. She attempted s…         8.874740e+17
32    RT @Athletics: 12/10 #BATP https://t.co/WxwJmv…         8.860537e+17
36    RT @dog_rates: This is Lilly. She just paralle…         8.305833e+17
68    RT @dog_rates: This is Emmy. She was adopted t…         8.780576e+17
73    RT @dog_rates: Meet Shadow. In an attempt to r…         8.782815e+17
…                                                …                  …
1766  This made my day. 12/10 please enjoy https://t…                  NaN
1860  RT until we find this dog. Clearly a cool dog …                  NaN
1991  This lil pupper is sad because we haven't foun…                  NaN
2259  RT @dogratingrating: Exceptional talent. Origi…         6.675487e+17
2260  RT @dogratingrating: Unoriginal idea. Blatant …         6.675484e+17

      retweeted_status_user_id retweeted_status_timestamp  \
19                4.196984e+09  2017-07-19 00:47:34 +0000
32                1.960740e+07  2017-07-15 02:44:07 +0000
36                4.196984e+09  2017-02-12 01:04:29 +0000
68                4.196984e+09  2017-06-23 01:10:23 +0000
73                4.196984e+09  2017-06-23 16:00:04 +0000
…                         …                        …
1766                       NaN                        NaN
1860                       NaN                        NaN
1991                       NaN                        NaN
2259              4.296832e+09  2015-11-20 03:43:06 +0000
2260              4.296832e+09  2015-11-20 03:41:59 +0000

                                         expanded_urls  rating_numerator  \
19    https://twitter.com/dog_rates/status/887473957…                13
32    https://twitter.com/dog_rates/status/886053434…                12
36    https://twitter.com/dog_rates/status/830583320…                13
68    https://twitter.com/dog_rates/status/878057613…                14
73    https://www.gofundme.com/3yd6y1c,https://twitt…                13
…                                                …               …
1766  https://twitter.com/dog_rates/status/678399652…                12
1860  https://twitter.com/dog_rates/status/675489971…                10
1991  https://twitter.com/dog_rates/status/672622327…                12
2259  https://twitter.com/dogratingrating/status/667…                12
```

```
2260  https://twitter.com/dogratingrating/status/667…               5

      rating_denominator    name doggo floofer  pupper puppo
19                    10  Canela  None    None    None  None
32                    10    None  None    None    None  None
36                    10   Lilly  None    None    None  None
68                    10    Emmy  None    None    None  None
73                    10  Shadow  None    None    None  None
…                  …     …   …     …     …   …
1766                  10    None  None    None    None  None
1860                  10    None  None    None    None  None
1991                  10    None  None    None  pupper  None
2259                  10    None  None    None    None  None
2260                  10    None  None    None    None  None

[192 rows x 17 columns]
```

[20]:

```
tweetimages.dtypes
```

[20]:

```
tweet_id      int64
jpg_url      object
img_num       int64
p1           object
p1_conf     float64
p1_dog         bool
p2           object
p2_conf     float64
p2_dog         bool
p3           object
p3_conf     float64
p3_dog         bool
dtype: object
```

[21]:

```
tweetimages['tweet_id'].duplicated().sum()
```

[21]:

```
0
```

[5]:

```
tweetimages.describe()
```

[5]:

```
           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
```

[23]:

```
tweetimages.p1.unique()
```

[23]:

```
array(['Welsh_springer_spaniel', 'redbone', 'German_shepherd',
       'Rhodesian_ridgeback', 'miniature_pinscher',
       'Bernese_mountain_dog', 'box_turtle', 'chow', 'shopping_cart',
       'miniature_poodle', 'golden_retriever', 'Gordon_setter',
       'Walker_hound', 'pug', 'bloodhound', 'Lhasa', 'English_setter',
       'hen', 'desktop_computer', 'Italian_greyhound', 'Maltese_dog',
       'three-toed_sloth', 'ox', 'malamute', 'guinea_pig',
       'soft-coated_wheaten_terrier', 'Chihuahua',
       'black-and-tan_coonhound', 'coho', 'toy_terrier',
       'Blenheim_spaniel', 'Pembroke', 'llama',
       'Chesapeake_Bay_retriever', 'curly-coated_retriever', 'dalmatian',
       'Ibizan_hound', 'Border_collie', 'Labrador_retriever', 'seat_belt',
       'snail', 'miniature_schnauzer', 'Airedale', 'triceratops', 'swab',
       'hay', 'hyena', 'jigsaw_puzzle', 'West_Highland_white_terrier',
       'toy_poodle', 'giant_schnauzer', 'vizsla', 'vacuum', 'Rottweiler',
       'Siberian_husky', 'teddy', 'papillon', 'Saint_Bernard',
       'porcupine', 'goose', 'Tibetan_terrier', 'borzoi', 'beagle',
       'hare', 'Yorkshire_terrier', 'Pomeranian', 'electric_fan',
       'web_site', 'ibex', 'kuvasz', 'fire_engine', 'lorikeet',
       'flat-coated_retriever', 'toyshop', 'common_iguana',
       'Norwegian_elkhound', 'frilled_lizard', 'leatherback_turtle',
       'hamster', 'Angora', 'Arctic_fox', 'trombone', 'canoe',
       'king_penguin', 'shopping_basket', 'standard_poodle',
       'Staffordshire_bullterrier', 'basenji', 'Lakeland_terrier',
       'American_Staffordshire_terrier', 'bearskin', 'Shih-Tzu',
       'bustard', 'crash_helmet', 'French_bulldog', 'Pekinese',
       'komondor', 'ski_mask', 'malinois', 'kelpie', 'Brittany_spaniel',
       'cocker_spaniel', 'shower_curtain', 'basset', 'jellyfish',
       'doormat', 'Arabian_camel', 'lynx', 'hog', 'comic_book', 'minivan',
       'seashore', 'cuirass', 'Brabancon_griffon', 'candle', 'Eskimo_dog',
       'weasel', 'Christmas_stocking', 'washbasin', 'car_mirror',
       'piggy_bank', 'pot', 'boathouse', 'mud_turtle',
       'German_short-haired_pointer', 'Shetland_sheepdog',
       'Irish_terrier', 'cairn', 'platypus', 'English_springer',
       'whippet', 'ping-pong_ball', 'sea_urchin', 'bow_tie',
       'window_shade', "jack-o'-lantern", 'sorrel', 'Sussex_spaniel',
       'peacock', 'axolotl', 'wool', 'banana', 'Dandie_Dinmont',
       'Norwich_terrier', 'wood_rabbit', 'dhole', 'keeshond',
       'Norfolk_terrier', 'lacewing', 'dingo', 'brown_bear',
       'Old_English_sheepdog', 'scorpion', 'flamingo', 'microphone',
       'Samoyed', 'pitcher', 'African_hunting_dog', 'refrigerator',
       'picket_fence', 'tub', 'zebra', 'hermit_crab', 'swing', 'Doberman',
       'park_bench', 'feather_boa', 'Loafer', 'stone_wall', 'ice_bear',
       'prayer_rug', 'chimpanzee', 'china_cabinet', 'bee_eater',
       'tennis_ball', 'carton', 'killer_whale', 'ostrich', 'terrapin',
       'Siamese_cat', 'gondola', 'Great_Pyrenees', 'microwave',
       'starfish', 'sandbar', 'tusker', 'motor_scooter', 'ram',
       'leaf_beetle', 'wombat', 'schipperke', 'Newfoundland',
       'bull_mastiff', 'water_bottle', 'suit', 'toilet_seat', 'collie',
       'robin', 'Cardigan', 'Greater_Swiss_Mountain_dog', 'slug',
       'toilet_tissue', 'acorn_squash', 'soccer_ball',
       'African_crocodile', 'tick', 'ocarina', 'boxer', 'street_sign',
       'bow', 'stove', 'paper_towel', 'upright', 'dough',
       'Scottish_deerhound', 'bath_towel', 'standard_schnauzer',
       'walking_stick', 'Irish_water_spaniel', 'bubble', 'Boston_bull',
       'book_jacket', 'rain_barrel', 'black-footed_ferret', 'guenon',
       'Japanese_spaniel', 'water_buffalo', 'patio', 'cowboy_hat',
       'dogsled', 'maze', 'harp', 'panpipe', 'cash_machine', 'mailbox',
       'wallaby', 'EntleBucher', 'earthstar', 'pillow', 'bluetick',
       'space_heater', 'carousel', 'Irish_setter', 'birdhouse', 'snorkel',
       'bald_eagle', 'koala', 'Leonberg', 'cheetah', 'minibus',
       'Weimaraner', 'clog', 'dishwasher', 'white_wolf', 'sliding_door',
       'damselfly', 'Great_Dane', 'Tibetan_mastiff', 'cheeseburger',
       'fiddler_crab', 'bannister', 'crane', 'Scotch_terrier',
       'snowmobile', 'badger', 'bighorn', 'geyser', 'barrow', 'bison',
       'Mexican_hairless', 'ice_lolly', 'sea_lion', 'dining_table',
       'groenendael', 'Australian_terrier', 'beaver', 'briard',
       'Appenzeller', 'grey_fox', 'mousetrap', 'hippopotamus',
       'Border_terrier', 'hummingbird', 'tailed_frog', 'otter',
       'Egyptian_cat', 'four-poster', 'wild_boar', 'bathtub', 'agama',
       'muzzle', 'hotdog', 'bib', 'espresso', 'timber_wolf', 'meerkat',
       'nail', 'hammer', 'home_theater', 'alp', 'bonnet', 'handkerchief',
       'hand_blower', 'polecat', 'lakeside', 'studio_couch', 'cup',
       'cliff', 'Bedlington_terrier', 'lawn_mower', 'balloon',
       'sunglasses', 'rapeseed', 'traffic_light', 'coil', 'binoculars',
       'paddle', 'tiger_shark', 'sulphur-crested_cockatoo',
       'wire-haired_fox_terrier', 'Saluki', 'American_black_bear',
       'rotisserie', 'conch', 'skunk', 'bookshop', 'radio_telescope',
       'cougar', 'African_grey', 'coral_reef', 'lion', 'maillot',
       'Madagascar_cat', 'tabby', 'silky_terrier', 'giant_panda',
       'long-horned_beetle', 'Afghan_hound', 'clumber', 'sundial',
       'padlock', 'pool_table', 'quilt', 'beach_wagon', 'remote_control',
       'bakery', 'pedestal', 'gas_pump', 'bookcase', 'shield', 'loupe',
       'restaurant', 'prison', 'school_bus', 'cowboy_boot', 'jersey',
       'wooden_spoon', 'leopard', 'mortarboard', 'teapot',
       'military_uniform', 'washer', 'coffee_mug', 'fountain',
       'pencil_box', 'barbell', 'grille', 'revolver', 'envelope',
       'syringe', 'marmot', 'pole', 'laptop', 'basketball', 'tricycle',
       'convertible', 'limousine', 'orange'], dtype=object)
```

[24]:

```
tweet_json.head()
```

[24]:

```
                       created_at                  id              id_str  \
0  Tue Aug 01 16:23:56 +0000 2017  892420643555336193  892420643555336193
1  Tue Aug 01 00:17:27 +0000 2017  892177421306343426  892177421306343426
2  Mon Jul 31 00:18:03 +0000 2017  891815181378084864  891815181378084864
3  Sun Jul 30 15:58:51 +0000 2017  891689557279858688  891689557279858688
4  Sat Jul 29 16:00:24 +0000 2017  891327558926688256  891327558926688256

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

  display_text_range                                           entities  \
0            [0, 85]  {'hashtags': [], 'symbols': [], 'user_mentions…
1           [0, 138]  {'hashtags': [], 'symbols': [], 'user_mentions…
2           [0, 121]  {'hashtags': [], 'symbols': [], 'user_mentions…
3            [0, 79]  {'hashtags': [], 'symbols': [], 'user_mentions…
4           [0, 138]  {'hashtags': [{'text': 'BarkWeek', 'indices': …

                                   extended_entities  \
0  {'media': [{'id': 892420639486877696, 'id_str'…
1  {'media': [{'id': 892177413194625024, 'id_str'…
2  {'media': [{'id': 891815175371796480, 'id_str'…
3  {'media': [{'id': 891689552724799489, 'id_str'…
4  {'media': [{'id': 891327551943041024, 'id_str'…

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

  … favorite_count  favorited retweeted possibly_sensitive  \
0  …          39467      False     False              False
1  …          33819      False     False              False
2  …          25461      False     False              False
3  …          42908      False     False              False
4  …          41048      False     False              False

  possibly_sensitive_appealable lang retweeted_status quoted_status_id  \
0                         False   en              NaN              NaN
1                         False   en              NaN              NaN
```

```
2                          False   en              NaN              NaN
3                          False   en              NaN              NaN
4                          False   en              NaN              NaN

  quoted_status_id_str  quoted_status
0                  NaN            NaN
1                  NaN            NaN
2                  NaN            NaN
3                  NaN            NaN
4                  NaN            NaN

[5 rows x 31 columns]
```

[25]:

```
tweet_json.dtypes
```

[25]:

```
created_at                            object
id                                     int64
id_str                                object
full_text                             object
truncated                               bool
display_text_range                    object
entities                              object
extended_entities                     object
source                                object
in_reply_to_status_id                float64
in_reply_to_status_id_str             object
in_reply_to_user_id                  float64
in_reply_to_user_id_str               object
in_reply_to_screen_name               object
user                                  object
geo                                   object
coordinates                           object
place                                 object
contributors                          object
is_quote_status                         bool
retweet_count                          int64
favorite_count                         int64
favorited                               bool
retweeted                               bool
possibly_sensitive                    object
possibly_sensitive_appealable         object
lang                                  object
retweeted_status                      object
quoted_status_id                     float64
quoted_status_id_str                  object
quoted_status                         object
dtype: object
```

[26]:
```
tweet_json.describe()
```

[26]:
```
                 id  in_reply_to_status_id  in_reply_to_user_id  \
count  2.354000e+03           7.800000e+01         7.800000e+01
mean   7.426978e+17           7.455079e+17         2.014171e+16
std    6.852812e+16           7.582492e+16         1.252797e+17
min    6.660209e+17           6.658147e+17         1.185634e+07
25%    6.783975e+17           6.757419e+17         3.086374e+08
50%    7.194596e+17           7.038708e+17         4.196984e+09
75%    7.993058e+17           8.257804e+17         4.196984e+09
max    8.924206e+17           8.862664e+17         8.405479e+17

       retweet_count  favorite_count  quoted_status_id
count    2354.000000     2354.000000      2.900000e+01
mean     3164.797366     8080.968564      8.162686e+17
std      5284.770364    11814.771334      6.164161e+16
min         0.000000        0.000000      6.721083e+17
25%       624.500000     1415.000000      7.888183e+17
50%      1473.500000     3603.500000      8.340867e+17
75%      3652.000000    10122.250000      8.664587e+17
max     79515.000000   132810.000000      8.860534e+17
```

[27]:
```
tweet_json.is_quote_status.unique()
```

[27]:
```
array([False,  True])
```

[28]:
```
tweet_json.favorited.unique()
```

[28]:
```
array([False,  True])
```

[29]:
```
tweet_json.retweeted.unique()
```

[29]:
```
array([False])
```

[30]:
```
tweet_json.possibly_sensitive.unique()
```

[30]:
```
array([False, nan], dtype=object)
```

1.2.1 Quality issues

1. remove the tweets with 'RT' in it's begining from WeRateDogs
2. keep only the tweets that has dog in either 1 of the 3 photos in tweetimages
3. Having Pivot column for the dog type instead of 4 columns as each dog is only 1 of the 4 types
4. Columns in Tweet-Json is repeated one time integer and another string , whcih we need to make it only 1 type

5. As described in the project overview the rating_numerator should be greater than the denominator so we are making it applicable by multipying with 10
6. denominator in WeRateDogs can't be 0 as this will make issue so we need to update it to 10 as the default value
7. We need to join data together so we can remove not dogs Tweets and keep only what we intersted in
8. Find the correct name for the dogs with Name 'None' or 'a'

1.2.2 Tidiness issues

1.Time STamp need to be changed. to datetime in WeRateDogs

2.change created_dt to date instead of object

1.3 Cleaning Data

In this section, clean **all** of the issues you documented while assessing.

Note: Make a copy of the original data before cleaning. Cleaning includes merging individual pieces of data according to the rules of tidy data. The result should be a high-quality and tidy master pandas DataFrame (or DataFrames, if appropriate).

[6]:
```
# Make copies of original pieces of data
tweet_json_clean=tweet_json
tweetimages_clean=tweetimages
WeRateDogs_clean=WeRateDogs
```

1.3.1 Issue #1:

Define: Remove the tweets with 'RT' in it's begining from WeRateDogs

Code

[7]:
```
WeRateDogs_clean=WeRateDogs
#Create List with tweets that HAs RT in the begining
rt_list=WeRateDogs_clean.query('text.str.contains("RT")')['tweet_id']

# Exclude those Tweets from the DataFrame we are working with
WeRateDogs_clean=WeRateDogs_clean[~WeRateDogs_clean.tweet_id.isin(rt_list)]
WeRateDogs_clean.describe()
```

[7]:
```
           tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
count  2.164000e+03           7.800000e+01         7.800000e+01
mean   7.371741e+17           7.455079e+17         2.014171e+16
std    6.753662e+16           7.582492e+16         1.252797e+17
min    6.660209e+17           6.658147e+17         1.185634e+07
25%    6.768214e+17           6.757419e+17         3.086374e+08
50%    7.097095e+17           7.038708e+17         4.196984e+09
75%    7.896066e+17           8.257804e+17         4.196984e+09
```

```
max    8.924206e+17           8.862664e+17         8.405479e+17

       retweeted_status_id  retweeted_status_user_id  rating_numerator  \
count                  0.0                       0.0       2164.000000
mean                   NaN                       NaN         13.226433
std                    NaN                       NaN         47.846578
min                    NaN                       NaN          0.000000
25%                    NaN                       NaN         10.000000
50%                    NaN                       NaN         11.000000
75%                    NaN                       NaN         12.000000
max                    NaN                       NaN       1776.000000

       rating_denominator
count         2164.000000
mean            10.495379
std              7.036821
min              0.000000
25%             10.000000
50%             10.000000
75%             10.000000
max            170.000000
```

Test

```
[8]: WeRateDogs_clean.query('text.str.contains("RT")')['tweet_id']
```

```
[8]: Series([], Name: tweet_id, dtype: int64)
```

1.3.2 Issue #2:

Define keep only the tweets that has dog in either 1 of the 3 photos in tweetimages

Code

```
[9]: # keep only the tweets with any of the 3 pictures has dog on it
     tweetimages_clean=tweetimages_clean.query('p1_dog == True or p2_dog == True or␣
      ↪p3_dog== True')
```

Test

```
[10]: tweetimages_clean.query('p1_dog == False and p2_dog == False and p3_dog==␣
       ↪False')
```

```
[10]: Empty DataFrame
      Columns: [tweet_id, jpg_url, img_num, p1, p1_conf, p1_dog, p2, p2_conf, p2_dog,
      p3, p3_conf, p3_dog]
      Index: []
```

1.3.3 Issue #3:

Define Having Pivot column for the dog type instead of 4 columns as each dog is only 1 of the 4 types in WeRateDogs

Code

[11]:
```
WeRateDogs_clean.head()
```

[11]:
```
             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  retweeted_status_id  \
0  This is Phineas. He's a mystical boy. Only eve…                  NaN
1  This is Tilly. She's just checking pup on you…                  NaN
2  This is Archie. He is a rare Norwegian Pouncin…                  NaN
3  This is Darla. She commenced a snooze mid meal…                  NaN
4  This is Franklin. He would like you to stop ca…                  NaN

   retweeted_status_user_id retweeted_status_timestamp  \
0                       NaN                        NaN
1                       NaN                        NaN
2                       NaN                        NaN
3                       NaN                        NaN
4                       NaN                        NaN

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

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

   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
```

[12]:

```
#changing the values in the 4 columms first into 0s and 1s
WeRateDogs_clean.loc[WeRateDogs_clean['doggo']== 'doggo','doggo']=1
WeRateDogs_clean.loc[WeRateDogs_clean['doggo']== 'None','doggo']=0

WeRateDogs_clean.loc[WeRateDogs_clean['floofer']== 'floofer','floofer']=1
WeRateDogs_clean.loc[WeRateDogs_clean['floofer']== 'None','floofer']=0


WeRateDogs_clean.loc[WeRateDogs_clean['pupper']== 'pupper','pupper']=1
WeRateDogs_clean.loc[WeRateDogs_clean['pupper']== 'None','pupper']=0


WeRateDogs_clean.loc[WeRateDogs_clean['puppo']== 'puppo','puppo']=1
WeRateDogs_clean.loc[WeRateDogs_clean['puppo']== 'None','puppo']=0
```

[13]:

```
    WeRateDogs_clean['doggo']=WeRateDogs_clean['doggo'].astype('int64')
    WeRateDogs_clean['floofer']=WeRateDogs_clean['floofer'].astype('int64')
    WeRateDogs_clean['pupper']=WeRateDogs_clean['pupper'].astype('int64')
    WeRateDogs_clean['puppo']=WeRateDogs_clean['puppo'].astype('int64')
```

```
/var/folders/gg/1s_1cfr929d6_g735t8pv1k40000gn/T/ipykernel_18836/457195937.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  WeRateDogs_clean['doggo']=WeRateDogs_clean['doggo'].astype('int64')
/var/folders/gg/1s_1cfr929d6_g735t8pv1k40000gn/T/ipykernel_18836/457195937.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  WeRateDogs_clean['floofer']=WeRateDogs_clean['floofer'].astype('int64')
/var/folders/gg/1s_1cfr929d6_g735t8pv1k40000gn/T/ipykernel_18836/457195937.py:3:
SettingWithCopyWarning:
```

```
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  WeRateDogs_clean['pupper']=WeRateDogs_clean['pupper'].astype('int64')
/var/folders/gg/1s_1cfr929d6_g735t8pv1k40000gn/T/ipykernel_18836/457195937.py:4:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  WeRateDogs_clean['puppo']=WeRateDogs_clean['puppo'].astype('int64')
```

[14]:
```
WeRateDogs_clean.dtypes
```

[14]:
```
tweet_id                          int64
in_reply_to_status_id           float64
in_reply_to_user_id             float64
timestamp                        object
source                           object
text                             object
retweeted_status_id             float64
retweeted_status_user_id        float64
retweeted_status_timestamp       object
expanded_urls                    object
rating_numerator                  int64
rating_denominator                int64
name                             object
doggo                             int64
floofer                           int64
pupper                            int64
puppo                             int64
dtype: object
```

[15]:
```
WeRateDogs_clean['dog_type']=WeRateDogs_clean[['doggo','floofer','pupper','puppo']].
↪apply(lambda x: x.idxmax(), axis=1)
```

```
/var/folders/gg/1s_1cfr929d6_g735t8pv1k40000gn/T/ipykernel_18836/416427393.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  WeRateDogs_clean['dog_type']=WeRateDogs_clean[['doggo','floofer','pupper','pup
po']].apply(lambda x: x.idxmax(), axis=1)
```

[16]:
```
WeRateDogs_clean.drop(columns={'doggo','floofer','pupper','puppo'},inplace=True)
```

```
/var/folders/gg/1s_1cfr929d6_g735t8pv1k40000gn/T/ipykernel_18836/775026658.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
WeRateDogs_clean.drop(columns={'doggo','floofer','pupper','puppo'},inplace=True)
```

Test:

[17]:
```
WeRateDogs_clean.head()
```

[17]:
```
             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  retweeted_status_id  \
0  This is Phineas. He's a mystical boy. Only eve…                  NaN
1  This is Tilly. She's just checking pup on you…                  NaN
2  This is Archie. He is a rare Norwegian Pouncin…                  NaN
3  This is Darla. She commenced a snooze mid meal…                  NaN
4  This is Franklin. He would like you to stop ca…                  NaN

   retweeted_status_user_id retweeted_status_timestamp  \
0                       NaN                        NaN
1                       NaN                        NaN
2                       NaN                        NaN
3                       NaN                        NaN
4                       NaN                        NaN
```

```
                                       expanded_urls  rating_numerator  \
0  https://twitter.com/dog_rates/status/892420643…                13   
1  https://twitter.com/dog_rates/status/892177421…                13   
2  https://twitter.com/dog_rates/status/891815181…                12   
3  https://twitter.com/dog_rates/status/891689557…                13   
4  https://twitter.com/dog_rates/status/891327558…                12   

   rating_denominator      name dog_type  
0                  10   Phineas    doggo  
1                  10     Tilly    doggo  
2                  10    Archie    doggo  
3                  10     Darla    doggo  
4                  10  Franklin    doggo  
```

1.3.4 Issue #4:

Define Columns in Tweet-Json is repeated one time integer and another string , whcih we need to make it only 1 type

Code

```
[18]: tweet_json_clean.head()
```

```
[18]:                        created_at                  id              id_str  \
0  Tue Aug 01 16:23:56 +0000 2017  892420643555336193  892420643555336193   
1  Tue Aug 01 00:17:27 +0000 2017  892177421306343426  892177421306343426   
2  Mon Jul 31 00:18:03 +0000 2017  891815181378084864  891815181378084864   
3  Sun Jul 30 15:58:51 +0000 2017  891689557279858688  891689557279858688   
4  Sat Jul 29 16:00:24 +0000 2017  891327558926688256  891327558926688256   

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

  display_text_range                                           entities  \
0            [0, 85]  {'hashtags': [], 'symbols': [], 'user_mentions…   
1           [0, 138]  {'hashtags': [], 'symbols': [], 'user_mentions…   
2           [0, 121]  {'hashtags': [], 'symbols': [], 'user_mentions…   
3            [0, 79]  {'hashtags': [], 'symbols': [], 'user_mentions…   
4           [0, 138]  {'hashtags': [{'text': 'BarkWeek', 'indices': …   

                                   extended_entities  \
0  {'media': [{'id': 892420639486877696, 'id_str'…   
1  {'media': [{'id': 892177413194625024, 'id_str'…   
```

```
2  {'media': [{'id': 891815175371796480, 'id_str'…
3  {'media': [{'id': 891689552724799489, 'id_str'…
4  {'media': [{'id': 891327551943041024, 'id_str'…

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

   … favorite_count  favorited retweeted possibly_sensitive  \
0  …          39467      False     False              False
1  …          33819      False     False              False
2  …          25461      False     False              False
3  …          42908      False     False              False
4  …          41048      False     False              False

  possibly_sensitive_appealable lang retweeted_status quoted_status_id  \
0                         False   en              NaN              NaN
1                         False   en              NaN              NaN
2                         False   en              NaN              NaN
3                         False   en              NaN              NaN
4                         False   en              NaN              NaN

  quoted_status_id_str  quoted_status
0                  NaN            NaN
1                  NaN            NaN
2                  NaN            NaN
3                  NaN            NaN
4                  NaN            NaN

[5 rows x 31 columns]
```

[19]:
```
tweet_json_clean.dtypes
```

[19]:
```
created_at                        object
id                                 int64
id_str                            object
full_text                         object
truncated                           bool
display_text_range                object
entities                          object
extended_entities                 object
source                            object
in_reply_to_status_id            float64
in_reply_to_status_id_str         object
```

```
in_reply_to_user_id             float64
in_reply_to_user_id_str          object
in_reply_to_screen_name          object
user                             object
geo                              object
coordinates                      object
place                            object
contributors                     object
is_quote_status                    bool
retweet_count                     int64
favorite_count                    int64
favorited                          bool
retweeted                          bool
possibly_sensitive               object
possibly_sensitive_appealable    object
lang                             object
retweeted_status                 object
quoted_status_id                float64
quoted_status_id_str             object
quoted_status                    object
dtype: object
```

[20]:
```
tweet_json_clean.
 ↪drop(columns={'in_reply_to_status_id_str','in_reply_to_user_id_str','quoted_status_id_str',
```

Test:

[21]:
```
tweet_json_clean.dtypes
```

[21]:
```
created_at                       object
id                                int64
full_text                        object
truncated                          bool
display_text_range               object
entities                         object
extended_entities                object
source                           object
in_reply_to_status_id           float64
in_reply_to_user_id             float64
in_reply_to_screen_name          object
user                             object
geo                              object
coordinates                      object
place                            object
contributors                     object
is_quote_status                    bool
retweet_count                     int64
```

```
favorite_count                       int64
favorited                             bool
retweeted                             bool
possibly_sensitive                  object
possibly_sensitive_appealable       object
lang                                object
retweeted_status                    object
quoted_status_id                   float64
quoted_status                       object
dtype: object
```

1.3.5 Issue 5:

Define: As described in the project overview the rating_numerator should be greater than the denominator so we are making it applicable by multipying with 10

Code:

[22]:
```
WeRateDogs_clean.dog_type.nunique()
```

[22]:
```
4
```

[23]:
```
WeRateDogs_clean['rating_numerator']=np.
  ↪where(WeRateDogs_clean['rating_numerator'] <␣
  ↪WeRateDogs_clean['rating_denominator'],WeRateDogs_clean['rating_numerator']*10,WeRateDogs_c
```

```
/var/folders/gg/1s_1cfr929d6_g735t8pv1k40000gn/T/ipykernel_18836/731236152.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  WeRateDogs_clean['rating_numerator']=np.where(WeRateDogs_clean['rating_numerat
or'] < WeRateDogs_clean['rating_denominator'],WeRateDogs_clean['rating_numerator
']*10,WeRateDogs_clean['rating_numerator'])
```

Test:

[24]:
```
WeRateDogs_clean.query('rating_numerator < rating_denominator')
```

[24]:
```
                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
315   835152434251116546                    NaN                  NaN
1016  746906459439529985           7.468859e+17         4.196984e+09

                       timestamp  \
315   2017-02-24 15:40:31 +0000
1016  2016-06-26 03:22:31 +0000
```

```
                                                 source  \
315   <a href="http://twitter.com/download/iphone" r…
1016  <a href="http://twitter.com/download/iphone" r…

                                                   text  retweeted_status_id  \
315   When you're so blinded by your systematic plag…                  NaN
1016  PUPDATE: can't see any. Even if I could, I cou…                  NaN

      retweeted_status_user_id retweeted_status_timestamp  \
315                        NaN                        NaN
1016                       NaN                        NaN

                                          expanded_urls  rating_numerator  \
315   https://twitter.com/dog_rates/status/835152434…                 0
1016  https://twitter.com/dog_rates/status/746906459…                 0

      rating_denominator  name dog_type
315                   10  None    doggo
1016                  10  None    doggo
```

1.3.6 Issue 6:

Define: denominator in WeRateDogs can't be 0 as this will make issue so we need to update it to 10 as the default value

Code:

[25]:
```
WeRateDogs_clean.query('rating_denominator==0')
```

[25]:
```
               tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
313  835246439529840640           8.352460e+17           26259576.0

                     timestamp  \
313  2017-02-24 21:54:03 +0000

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

                                                  text  retweeted_status_id  \
313  @jonnysun @Lin_Manuel ok jomny I know you're e…                  NaN

     retweeted_status_user_id retweeted_status_timestamp expanded_urls  \
313                       NaN                        NaN           NaN

     rating_numerator  rating_denominator  name dog_type
313               960                   0  None    doggo
```

[26]:
```
WeRateDogs_clean['rating_denominator']=np.
↪where(WeRateDogs_clean['rating_denominator']==0,10,WeRateDogs_clean['rating_denominator'])
```

```
/var/folders/gg/1s_1cfr929d6_g735t8pv1k40000gn/T/ipykernel_18836/1860325895.py:1
: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  WeRateDogs_clean['rating_denominator']=np.where(WeRateDogs_clean['rating_denom
inator']==0,10,WeRateDogs_clean['rating_denominator'])
```

Test:

[27]:
```
WeRateDogs_clean.query('rating_denominator==0')
```

[27]:
```
Empty DataFrame
Columns: [tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp,
source, text, retweeted_status_id, retweeted_status_user_id,
retweeted_status_timestamp, expanded_urls, rating_numerator, rating_denominator,
name, dog_type]
Index: []
```

[ ]:
```

```

1.3.7 Issue 7:

Define: Find the correct name for the dogs with Name ‘None’ or ‘a’

Code:

[28]:
```
WeRateDogs_clean.dtypes
```

[28]:
```
tweet_id                        int64
in_reply_to_status_id         float64
in_reply_to_user_id           float64
timestamp                      object
source                         object
text                           object
retweeted_status_id           float64
retweeted_status_user_id      float64
retweeted_status_timestamp     object
expanded_urls                  object
rating_numerator                int64
rating_denominator              int64
name                           object
dog_type                       object
dtype: object
```

```
[29]: tweetimages_clean.dtypes
```

```
[29]: tweet_id      int64
      jpg_url      object
      img_num       int64
      p1           object
      p1_conf     float64
      p1_dog         bool
      p2           object
      p2_conf     float64
      p2_dog         bool
      p3           object
      p3_conf     float64
      p3_dog         bool
      dtype: object
```

```
[30]: tweet_json_clean.dtypes
```

```
[30]: created_at                        object
      id                                 int64
      full_text                         object
      truncated                           bool
      display_text_range                object
      entities                          object
      extended_entities                 object
      source                            object
      in_reply_to_status_id            float64
      in_reply_to_user_id              float64
      in_reply_to_screen_name           object
      user                              object
      geo                               object
      coordinates                       object
      place                             object
      contributors                      object
      is_quote_status                     bool
      retweet_count                      int64
      favorite_count                     int64
      favorited                           bool
      retweeted                           bool
      possibly_sensitive                object
      possibly_sensitive_appealable     object
      lang                              object
      retweeted_status                  object
      quoted_status_id                 float64
      quoted_status                     object
      dtype: object
```

[31]:

```
WeRateDogs_clean1=pd.merge(WeRateDogs_clean,tweetimages_clean,how='inner')

WeRateDogs_clean_anlyz=pd.
 ↪merge(WeRateDogs_clean1,tweet_json_clean,how='left',left_on=WeRateDogs_clean1['tweet_id'],r
WeRateDogs_clean_anlyz
```

[31]:

```
                   key_0            tweet_id  in_reply_to_status_id_x  \
0     892177421306343426  892177421306343426                      NaN
1     891815181378084864  891815181378084864                      NaN
2     891689557279858688  891689557279858688                      NaN
3     891327558926688256  891327558926688256                      NaN
4     891087950875897856  891087950875897856                      NaN
…                  …                 …                      …
1673  666049248165822465  666049248165822465                      NaN
1674  666044226329800704  666044226329800704                      NaN
1675  666033412701032449  666033412701032449                      NaN
1676  666029285002620928  666029285002620928                      NaN
1677  666020888022790149  666020888022790149                      NaN

      in_reply_to_user_id_x                  timestamp  \
0                       NaN  2017-08-01 00:17:27 +0000
1                       NaN  2017-07-31 00:18:03 +0000
2                       NaN  2017-07-30 15:58:51 +0000
3                       NaN  2017-07-29 16:00:24 +0000
4                       NaN  2017-07-29 00:08:17 +0000
…                     …                        …
1673                    NaN  2015-11-16 00:24:50 +0000
1674                    NaN  2015-11-16 00:04:52 +0000
1675                    NaN  2015-11-15 23:21:54 +0000
1676                    NaN  2015-11-15 23:05:30 +0000
1677                    NaN  2015-11-15 22:32:08 +0000

                                               source_x  \
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…
…                                                 …
1673  <a href="http://twitter.com/download/iphone" r…
1674  <a href="http://twitter.com/download/iphone" r…
1675  <a href="http://twitter.com/download/iphone" r…
1676  <a href="http://twitter.com/download/iphone" r…
1677  <a href="http://twitter.com/download/iphone" r…

                                                   text  retweeted_status_id  \
0     This is Tilly. She's just checking pup on you…                  NaN
```

```
1     This is Archie. He is a rare Norwegian Pouncin…                  NaN
2     This is Darla. She commenced a snooze mid meal…                  NaN
3     This is Franklin. He would like you to stop ca…                  NaN
4     Here we have a majestic great white breaching …                  NaN
…                                                 …                  …
1673  Here we have a 1949 1st generation vulpix. Enj…                  NaN
1674  This is a purebred Piers Morgan. Loves to Netf…                  NaN
1675  Here is a very happy pup. Big fan of well-main…                  NaN
1676  This is a western brown Mitsubishi terrier. Up…                  NaN
1677  Here we have a Japanese Irish Setter. Lost eye…                  NaN

      retweeted_status_user_id retweeted_status_timestamp  … retweet_count  \
0                          NaN                        NaN  …          6514
1                          NaN                        NaN  …          4328
2                          NaN                        NaN  …          8964
3                          NaN                        NaN  …          9774
4                          NaN                        NaN  …          3261
…                        …                        …  …           …
1673                       NaN                        NaN  …            41
1674                       NaN                        NaN  …           147
1675                       NaN                        NaN  …            47
1676                       NaN                        NaN  …            48
1677                       NaN                        NaN  …           532

      favorite_count  favorited retweeted possibly_sensitive  \
0              33819      False     False              False
1              25461      False     False              False
2              42908      False     False              False
3              41048      False     False              False
4              20562      False     False              False
…              …        …       …                …
1673             111      False     False              False
1674             311      False     False              False
1675             128      False     False              False
1676             132      False     False              False
1677            2535      False     False              False

     possibly_sensitive_appealable  lang retweeted_status  quoted_status_id  \
0                            False    en              NaN               NaN
1                            False    en              NaN               NaN
2                            False    en              NaN               NaN
3                            False    en              NaN               NaN
4                            False    en              NaN               NaN
…                            …   …              …               …
1673                         False    en              NaN               NaN
1674                         False    en              NaN               NaN
1675                         False    en              NaN               NaN
```

```
1676                         False    en              NaN               NaN
1677                         False    en              NaN               NaN

      quoted_status
0               NaN
1               NaN
2               NaN
3               NaN
4               NaN
…             …
1673            NaN
1674            NaN
1675            NaN
1676            NaN
1677            NaN

[1678 rows x 53 columns]
```

1.3.8 Issue 8:

Define: We need to join data together so we can remove not dogs Tweets and keep only what we intersted in

Code:

[32]:
```
WeRateDogs_clean_anlyz.dtypes
```

[32]:
```
key_0                                  int64
tweet_id                               int64
in_reply_to_status_id_x              float64
in_reply_to_user_id_x                float64
timestamp                             object
source_x                              object
text                                  object
retweeted_status_id                  float64
retweeted_status_user_id             float64
retweeted_status_timestamp            object
expanded_urls                         object
rating_numerator                       int64
rating_denominator                     int64
name                                  object
dog_type                              object
jpg_url                               object
img_num                                int64
p1                                    object
p1_conf                              float64
p1_dog                                  bool
p2                                    object
```

```
p2_conf                          float64
p2_dog                              bool
p3                                object
p3_conf                          float64
p3_dog                              bool
created_at                        object
id                                 int64
full_text                         object
truncated                           bool
display_text_range                object
entities                          object
extended_entities                 object
source_y                          object
in_reply_to_status_id_y          float64
in_reply_to_user_id_y            float64
in_reply_to_screen_name           object
user                              object
geo                               object
coordinates                       object
place                             object
contributors                      object
is_quote_status                     bool
retweet_count                      int64
favorite_count                     int64
favorited                           bool
retweeted                           bool
possibly_sensitive                object
possibly_sensitive_appealable     object
lang                              object
retweeted_status                  object
quoted_status_id                 float64
quoted_status                     object
dtype: object
```

[33]:

```
#after joining all the data we found some of records tagged as having dogs␣
 ↪meanwhile it'snt so we have to remove them
WeRateDogs_clean_anlyz.
 ↪drop(WeRateDogs_clean_anlyz[WeRateDogs_clean_anlyz['text'].str.contains('We␣
 ↪only rate dogs')].index,axis=0,inplace=True)
```

[34]:

```
WeRateDogs_clean_anlyz['text'].str.contains('We only rate dogs').sum()
```

[34]: 0

[35]:

```
WeRateDogs_clean_anlyz
```

```
[35]:                    key_0            tweet_id  in_reply_to_status_id_x  \
      0     892177421306343426  892177421306343426                      NaN
      1     891815181378084864  891815181378084864                      NaN
      2     891689557279858688  891689557279858688                      NaN
      3     891327558926688256  891327558926688256                      NaN
      4     891087950875897856  891087950875897856                      NaN
      …                  …                 …                      …
      1673  666049248165822465  666049248165822465                      NaN
      1674  666044226329800704  666044226329800704                      NaN
      1675  666033412701032449  666033412701032449                      NaN
      1676  666029285002620928  666029285002620928                      NaN
      1677  666020888022790149  666020888022790149                      NaN

            in_reply_to_user_id_x                  timestamp  \
      0                       NaN  2017-08-01 00:17:27 +0000
      1                       NaN  2017-07-31 00:18:03 +0000
      2                       NaN  2017-07-30 15:58:51 +0000
      3                       NaN  2017-07-29 16:00:24 +0000
      4                       NaN  2017-07-29 00:08:17 +0000
      …                     …                        …
      1673                    NaN  2015-11-16 00:24:50 +0000
      1674                    NaN  2015-11-16 00:04:52 +0000
      1675                    NaN  2015-11-15 23:21:54 +0000
      1676                    NaN  2015-11-15 23:05:30 +0000
      1677                    NaN  2015-11-15 22:32:08 +0000

                                                     source_x  \
      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…
      …                                                 …
      1673  <a href="http://twitter.com/download/iphone" r…
      1674  <a href="http://twitter.com/download/iphone" r…
      1675  <a href="http://twitter.com/download/iphone" r…
      1676  <a href="http://twitter.com/download/iphone" r…
      1677  <a href="http://twitter.com/download/iphone" r…

                                                         text  retweeted_status_id  \
      0     This is Tilly. She's just checking pup on you…                  NaN
      1     This is Archie. He is a rare Norwegian Pouncin…                  NaN
      2     This is Darla. She commenced a snooze mid meal…                  NaN
      3     This is Franklin. He would like you to stop ca…                  NaN
      4     Here we have a majestic great white breaching …                  NaN
      …                                                 …                  …
      1673  Here we have a 1949 1st generation vulpix. Enj…                  NaN
```

```
1674  This is a purebred Piers Morgan. Loves to Netf…                 NaN
1675  Here is a very happy pup. Big fan of well-main…                 NaN
1676  This is a western brown Mitsubishi terrier. Up…                 NaN
1677  Here we have a Japanese Irish Setter. Lost eye…                 NaN

      retweeted_status_user_id retweeted_status_timestamp  … retweet_count  \
0                          NaN                        NaN  …          6514
1                          NaN                        NaN  …          4328
2                          NaN                        NaN  …          8964
3                          NaN                        NaN  …          9774
4                          NaN                        NaN  …          3261
…                        …                        …  …           …
1673                       NaN                        NaN  …            41
1674                       NaN                        NaN  …           147
1675                       NaN                        NaN  …            47
1676                       NaN                        NaN  …            48
1677                       NaN                        NaN  …           532

      favorite_count  favorited retweeted possibly_sensitive  \
0              33819      False     False              False
1              25461      False     False              False
2              42908      False     False              False
3              41048      False     False              False
4              20562      False     False              False
…              …        …       …                …
1673             111      False     False              False
1674             311      False     False              False
1675             128      False     False              False
1676             132      False     False              False
1677            2535      False     False              False

     possibly_sensitive_appealable  lang retweeted_status  quoted_status_id  \
0                            False    en              NaN               NaN
1                            False    en              NaN               NaN
2                            False    en              NaN               NaN
3                            False    en              NaN               NaN
4                            False    en              NaN               NaN
…                            …   …              …               …
1673                         False    en              NaN               NaN
1674                         False    en              NaN               NaN
1675                         False    en              NaN               NaN
1676                         False    en              NaN               NaN
1677                         False    en              NaN               NaN

      quoted_status
0               NaN
1               NaN
```

```
2               NaN
3               NaN
4               NaN
…             …
1673            NaN
1674            NaN
1675            NaN
1676            NaN
1677            NaN

[1634 rows x 53 columns]
```

1.3.9 Tidiness issues

1.Time STamp need to be changed. to datetime in WeRateDogs

2.change created_dt to date instead of object

1.3.10 Issue 1& 2:

Define:

1. Time STamp need to be changed. to datetime in WeRateDogs
2. change created_dt to date instead of object

COde :

```
[36]: WeRateDogs_clean_anlyz.dtypes
```

```
[36]: key_0                               int64
      tweet_id                            int64
      in_reply_to_status_id_x           float64
      in_reply_to_user_id_x             float64
      timestamp                          object
      source_x                           object
      text                               object
      retweeted_status_id               float64
      retweeted_status_user_id          float64
      retweeted_status_timestamp         object
      expanded_urls                      object
      rating_numerator                    int64
      rating_denominator                  int64
      name                               object
      dog_type                           object
      jpg_url                            object
      img_num                             int64
      p1                                 object
      p1_conf                           float64
      p1_dog                               bool
```

```
p2                                object
p2_conf                          float64
p2_dog                              bool
p3                                object
p3_conf                          float64
p3_dog                              bool
created_at                        object
id                                 int64
full_text                         object
truncated                           bool
display_text_range                object
entities                          object
extended_entities                 object
source_y                          object
in_reply_to_status_id_y          float64
in_reply_to_user_id_y            float64
in_reply_to_screen_name           object
user                              object
geo                               object
coordinates                       object
place                             object
contributors                      object
is_quote_status                     bool
retweet_count                      int64
favorite_count                     int64
favorited                           bool
retweeted                           bool
possibly_sensitive                object
possibly_sensitive_appealable     object
lang                              object
retweeted_status                  object
quoted_status_id                 float64
quoted_status                     object
dtype: object
```

[37]:
```
# updating the 2 columns into datetime type
WeRateDogs_clean_anlyz['timestamp'] = pd.to_datetime(␣
 ↪WeRateDogs_clean_anlyz['timestamp'])
WeRateDogs_clean_anlyz['created_at']=pd.
 ↪to_datetime(WeRateDogs_clean_anlyz['created_at'])
```

Test

[38]:
```
WeRateDogs_clean_anlyz.dtypes
```

[38]:
```
key_0                                          int64
tweet_id                                       int64
```

```
in_reply_to_status_id_x                      float64
in_reply_to_user_id_x                        float64
timestamp                        datetime64[ns, UTC]
source_x                                      object
text                                          object
retweeted_status_id                          float64
retweeted_status_user_id                     float64
retweeted_status_timestamp                    object
expanded_urls                                 object
rating_numerator                               int64
rating_denominator                             int64
name                                          object
dog_type                                      object
jpg_url                                       object
img_num                                        int64
p1                                            object
p1_conf                                      float64
p1_dog                                          bool
p2                                            object
p2_conf                                      float64
p2_dog                                          bool
p3                                            object
p3_conf                                      float64
p3_dog                                          bool
created_at                       datetime64[ns, UTC]
id                                             int64
full_text                                     object
truncated                                       bool
display_text_range                            object
entities                                      object
extended_entities                             object
source_y                                      object
in_reply_to_status_id_y                      float64
in_reply_to_user_id_y                        float64
in_reply_to_screen_name                       object
user                                          object
geo                                           object
coordinates                                   object
place                                         object
contributors                                  object
is_quote_status                                 bool
retweet_count                                  int64
favorite_count                                 int64
favorited                                       bool
retweeted                                       bool
possibly_sensitive                            object
possibly_sensitive_appealable                 object
```

```
lang                                                object
retweeted_status                                    object
quoted_status_id                                   float64
quoted_status                                       object
dtype: object
```

1.4 Storing Data

Save gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv".

[39]:
```
WeRateDogs_clean_anlyz.to_csv('twitter_archive_master.csv')
```

1.5 Analyzing and Visualizing Data

In this section, analyze and visualize your wrangled data. You must produce at least **three (3) insights and one (1) visualization.**

[40]:
```
# creating new column for retweeting count
WeRateDogs_clean_anlyz['rating_perc']=WeRateDogs_clean_anlyz['rating_numerator']/
 ↪WeRateDogs_clean_anlyz['rating_denominator']
```

[41]:
```
#rounding the Rating percentatige value to the nearst one
WeRateDogs_clean_anlyz['rating_perc']=WeRateDogs_clean_anlyz['rating_perc'].
 ↪round()
```

[42]:
```
# grouping the tweets based on the rating percentaige and Dog type to see which␣
 ↪combination has the most tweets

Rate=WeRateDogs_clean_anlyz.groupby(['dog_type','rating_perc'])['tweet_id'].
 ↪count().reset_index(name='count')
Rate.sort_values(by=['rating_perc'],ascending=False).reset_index()
```

[42]:
```
    index dog_type  rating_perc  count
0       9    doggo          9.0    111
1      15   pupper          9.0     14
2      17    puppo          9.0      1
3       8    doggo          8.0     61
4      14   pupper          8.0      7
5       7    doggo          7.0     27
6      13   pupper          7.0      3
7       6    doggo          6.0     17
8       5    doggo          5.0     15
9       4    doggo          4.0      6
10      3    doggo          3.0      6
11     12   pupper          3.0      1
12      2    doggo          2.0      3
```

```
13      1    doggo          1.0   1192
14     10  floofer          1.0      7
15     11   pupper          1.0    142
16     16    puppo          1.0     20
17      0    doggo          0.0      1
```

[43]:

```
# having the no of dogs in each dog type
WeRateDogs_clean_anlyz.groupby(['dog_type']).size().sort_values()
```

[43]:

```
dog_type
floofer       7
puppo        21
pupper      167
doggo      1439
dtype: int64
```

[80]:

```
# understaing the Heighest dog type which having retweets
WeRateDogs_clean_anlyz.groupby(['retweet_count','dog_type']).size().nlargest(10)
```

[80]:

```
retweet_count  dog_type
61             doggo       3
77             doggo       3
119            doggo       3
146            doggo       3
231            doggo       3
234            doggo       3
242            doggo       3
315            doggo       3
572            doggo       3
602            doggo       3
dtype: int64
```

[51]:

```
# geeting the Min and Max for Retweet count
WeRateDogs_clean_anlyz['retweet_count'].describe()
```

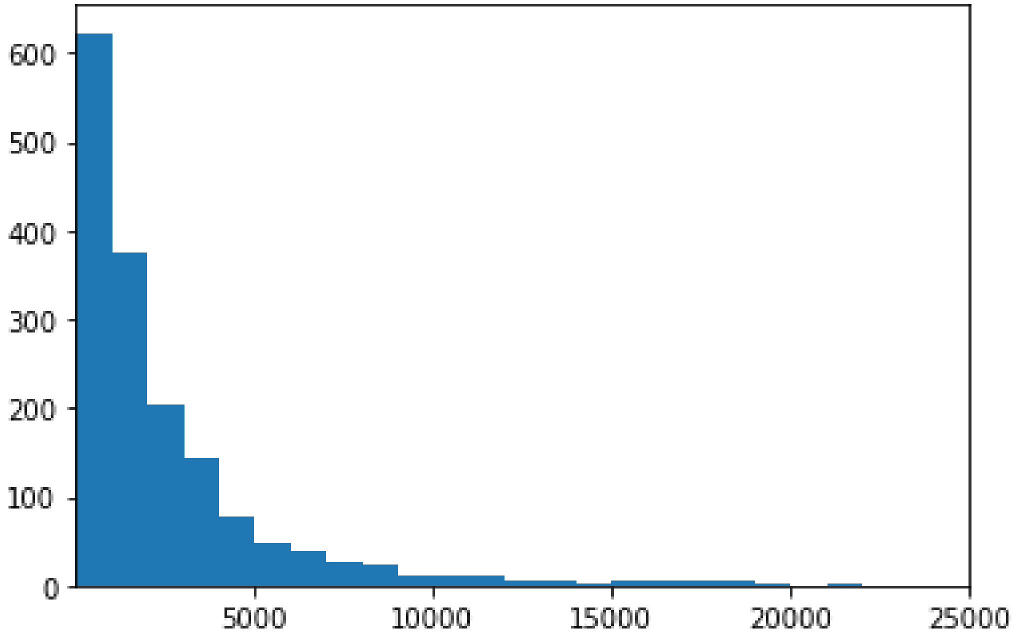
[51]:

```
count     1634.000000
mean      2742.838433
std       4724.113233
min         16.000000
25%        629.000000
50%       1403.000000
75%       3206.250000
max      79515.000000
Name: retweet_count, dtype: float64
```

[57]:

```
# creating proper ranges for the bins to understand the retweet counts
binn=np.arange(16,WeRateDogs_clean_anlyz['retweet_count'].max()+2,1000)
```

```
#plotting Histogram for retweet count
plt.hist(WeRateDogs_clean_anlyz['retweet_count'],bins=binn);
plt.xlim((16,25000));
```



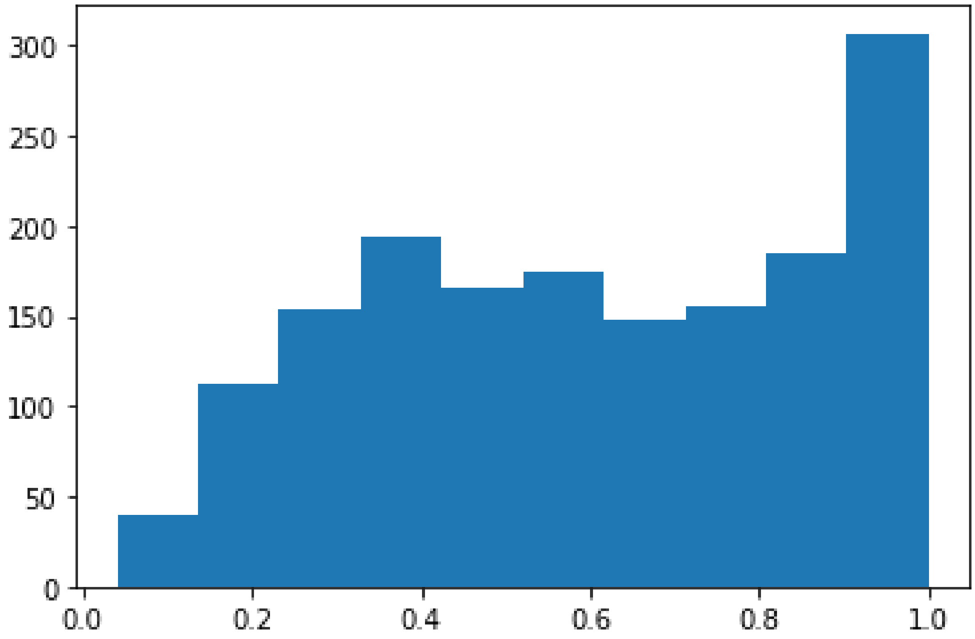
[66]:
```
WeRateDogs_clean_anlyz.groupby(['name']).size().nlargest(10)
```

[66]:
```
name
None       387
a           40
Charlie     10
Cooper      10
Lucy        10
Oliver       9
Tucker       9
Penny        8
Daisy        7
Winston      7
dtype: int64
```

[68]:
```
WeRateDogs_clean_anlyz.groupby(['name']).size().max()/WeRateDogs_clean_anlyz.
 ↪groupby(['name']).size().sum()
```

[68]:
```
0.23684210526315788
```

[77]:
```
plt.hist(WeRateDogs_clean_anlyz['p1_conf'].round(2));
```



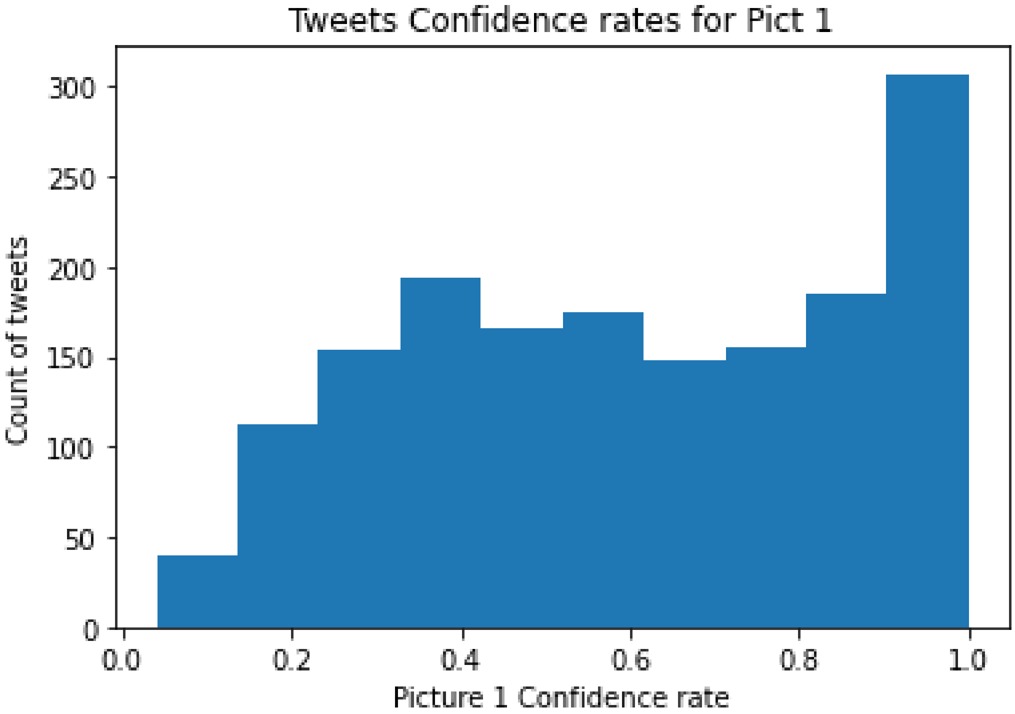
1.5.1 Insights:

1. heighest 5 rating Percentages goes manily to Doggo Type, Then Pupper
2. Most of dogs are doggo with 1439 and the 2nd most is pupper with 167
3. Most of tweets doecn't have dog names as 23% doen't have names

1.5.2 Visualization

```
[79]: plt.hist(WeRateDogs_clean_anlyz['p1_conf'].round(2));
      plt.xlabel('Picture 1 Confidence rate');
      plt.ylabel('Count of tweets');
      plt.title('Tweets Confidence rates for Pict 1');
```

```
[79]: ''
```


Tweets Confidence rates for Pict 1
Count of tweets
300
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0
Picture 1 Confidence rate

[ ]: