# wrangle_act

December 14, 2018

## 1 Project - WeRateDogs

In [69]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import requests
import json
import os
import io
import time
import re
import tweepy
```

In [70]:
```
#### File no.1 - 'WeRateDogs Twitter archive'
#Read CSV file into DataFrame
twitter_archive_df = pd.read_csv('twitter-archive.csv')
twitter_archive_df.head(2)
```

Out[70]:

| | tweet_id | in_reply_to_status_id | in_reply_to_user_id | timestamp | source | text | retweeted_status_id | retweeted_status_user_id | retweeted_status_timestamp |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 892420643555336193 | NaN | NaN | 2017-08-01 16:23:56 +0000 | <a href="http://twitter.com/download/iphone" r... | This is Phineas. He's a mystical boy. Only eve... | NaN | NaN | NaN |
| 1 | 892177421306343426 | NaN | NaN | 2017-08-01 00:17:27 +0000 | <a href="http://twitter.com/download/iphone" r... | This is Tilly. She's just checking pup on you... | NaN | NaN | NaN |

```
1                       NaN                         NaN

                                          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 pupper puppo
0                  10  Phineas  None    None   None  None
1                  10    Tilly  None    None   None  None
```

In [71]:
```
####  File no.2 - 'Tweet image predictions'
#Programmatically download image_predictions.tsv to a folder called image_predictions u
```

In [72]:
```
# Make directory if it doesn't already exist
import requests
import os

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

url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predicti
urls_Data = requests.get(url).content
image_df = pd.read_table(io.StringIO(urls_Data.decode('utf-8')))
image_df.head(2)
```

Out[72]:
```
             tweet_id                                          jpg_url  \
0  666020888022790149  https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg
1  666029285002620928  https://pbs.twimg.com/media/CT42GRgUYAA5iDo.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

    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
```

In [73]:
```
####  File no.3 -  Additional data -> 'tweet_json.txt'  -> df
#info - for accessing API

consumer_key = 'DLF6T49XhH6FFT0p8nsXhS0j9'
consumer_secret = '7B4t8BiZ6GFoyoXYnsxeHqneJH0pnVg24SUF93J8fsAAVkuoBs'
access_token = '1071060638322049024-UdKhy0MMnLUPfbHRCyRak1wLI6x8ow'
access_token_secret = 'UbfJTtfvSiC2dXr574OjFYpy0lDjRECAiwTpE4t5hp2iZ'


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

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

In [74]:
```
tweet_id_ = list(twitter_archive_df['tweet_id'])

processed =[]
not_processed = []
with open('tweet_json.txt', 'w') as file:
      for tweet_id in tweet_id_:
        try:
            tweet_details = api.get_status(tweet_id,tweet_mode='extended')
            json.dump(tweet_details._json, file)
            file.write('\n')
            processed.append(tweet_id)
        except Exception as e:
            print(tweet_id,e)
            not_processed.append(tweet_id)
```

```
888202515573088257 [{'code': 144, 'message': 'No status found with that ID.'}]
873697596434513921 [{'code': 144, 'message': 'No status found with that ID.'}]
872668790621863937 [{'code': 144, 'message': 'No status found with that ID.'}]
869988702071779329 [{'code': 144, 'message': 'No status found with that ID.'}]
866816280283807744 [{'code': 144, 'message': 'No status found with that ID.'}]
861769973181624320 [{'code': 144, 'message': 'No status found with that ID.'}]
845459076796616705 [{'code': 144, 'message': 'No status found with that ID.'}]
842892208864923648 [{'code': 144, 'message': 'No status found with that ID.'}]
837012587749474308 [{'code': 144, 'message': 'No status found with that ID.'}]
827228250799742977 [{'code': 144, 'message': 'No status found with that ID.'}]
802247111496568832 [{'code': 144, 'message': 'No status found with that ID.'}]
775096608509886464 [{'code': 144, 'message': 'No status found with that ID.'}]
770743923962707968 [{'code': 144, 'message': 'No status found with that ID.'}]
Rate limit reached. Sleeping for: 738
754011816964026368 [{'code': 144, 'message': 'No status found with that ID.'}]
Rate limit reached. Sleeping for: 738
```

In [75]:
```
tweet_df = pd.read_json('tweet_json.txt', lines = True,encoding='utf-8')
tweet_df.head(2)
```

Out[75]:

| | 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] | |

| | entities | \ |
|---|---|---|
| 0 | {'hashtags': [], 'symbols': [], 'user_mentions... | |
| 1 | {'hashtags': [], 'symbols': [], 'user_mentions... | |

```
                                      extended_entities  favorite_count  \
0  {'media': [{'id': 892420639486877696, 'id_str'...           38151
1  {'media': [{'id': 892177413194625024, 'id_str'...           32715

   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

                          ...                          quoted_status  \
0                         ...                                    NaN
1                         ...                                    NaN

   quoted_status_id quoted_status_id_str  quoted_status_permalink  \
0               NaN                  NaN                      NaN
1               NaN                  NaN                      NaN

   retweet_count  retweeted  retweeted_status  \
0           8347      False               NaN
1           6170      False               NaN

                                             source truncated  \
0  <a href="http://twitter.com/download/iphone" r...     False
1  <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 rows x 32 columns]
```

```
In [76]: # different lengths
         len(twitter_archive_df) , len(image_df), len(tweet_df)
```

```
Out[76]: (2356, 2075, 2342)
```

### 1.1 Assessing Data

Note:for visual and programmatic quality and tidiness issues. Detect and document at least eight (8) quality issues and two (2) tidiness issues

```
In [77]: twitter_archive_df.head()
```

```
Out[77]:             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
```

In [78]:
```
image_df.head()
```

Out[78]:
```
             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
```

```
In [79]: tweet_df.head(2)
```

```
Out[79]:    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]

                                             entities  \
0  {'hashtags': [], 'symbols': [], 'user_mentions...
1  {'hashtags': [], 'symbols': [], 'user_mentions...

                                    extended_entities  favorite_count  \
0  {'media': [{'id': 892420639486877696, 'id_str'...           38151
1  {'media': [{'id': 892177413194625024, 'id_str'...           32715

   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

                          ...                          quoted_status  \
0                         ...                                    NaN
1                         ...                                    NaN

   quoted_status_id quoted_status_id_str  quoted_status_permalink  \
0               NaN                  NaN                      NaN
1               NaN                  NaN                      NaN

   retweet_count  retweeted  retweeted_status  \
0           8347      False               NaN
1           6170      False               NaN
```

```
                                                    source truncated  \
0  <a href="http://twitter.com/download/iphone" r...     False
1  <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 rows x 32 columns]
```

```
In [80]: twitter_archive_df.info()
```

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

```
In [81]: twitter_archive_df.sample(2)
```

```
Out[81]:                tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
1453  695629776980148225                    NaN                  NaN
607   798209839306514432                    NaN                  NaN

                      timestamp  \
1453  2016-02-05 15:27:17 +0000
607   2016-11-14 17:03:50 +0000

                                                    source  \
1453  <a href="http://twitter.com/download/iphone" r...
607   <a href="http://twitter.com/download/iphone" r...
```

```
                                                    text  retweeted_status_id  \
1453  Meet Calvin. He's proof that degrees mean abso...                  NaN
607   This is Cooper. His bow tie was too heavy for ...                  NaN

      retweeted_status_user_id retweeted_status_timestamp  \
1453                       NaN                       NaN
607                        NaN                       NaN

                                          expanded_urls  rating_numerator  \
1453  https://twitter.com/dog_rates/status/695629776...                 8
607   https://twitter.com/dog_rates/status/798209839...                13

      rating_denominator    name doggo floofer pupper puppo
1453                  10  Calvin  None    None   None  None
607                   10  Cooper  None    None   None  None
```

In [82]: `image_df.info()`

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

In [83]: `tweet_df.info()`

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2342 entries, 0 to 2341
Data columns (total 32 columns):
contributors                    0 non-null float64
coordinates                     0 non-null float64
created_at                      2342 non-null datetime64[ns]
display_text_range              2342 non-null object
entities                        2342 non-null object
extended_entities               2068 non-null object
```

```
favorite_count                  2342 non-null int64
favorited                       2342 non-null bool
full_text                       2342 non-null object
geo                             0 non-null float64
id                              2342 non-null int64
id_str                          2342 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                 2342 non-null bool
lang                            2342 non-null object
place                           1 non-null object
possibly_sensitive              2206 non-null float64
possibly_sensitive_appealable   2206 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                   2342 non-null int64
retweeted                       2342 non-null bool
retweeted_status                168 non-null object
source                          2342 non-null object
truncated                       2342 non-null bool
user                            2342 non-null object
dtypes: bool(4), datetime64[ns](1), float64(11), int64(4), object(12)
memory usage: 539.8+ KB
```

```
In [84]: # Number of tweet_id duplicates
         all_columns = pd.Series(list(twitter_archive_df) + list(image_df) + list(tweet_df))
         all_columns[all_columns.duplicated()]
```

```
Out[84]: 17                 tweet_id
         42    in_reply_to_status_id
         44      in_reply_to_user_id
         58                   source
         dtype: object
```

**Quality**

- 17 tweet_id duplicates

**twitter table**

- retweets, in_reply_to - not important
- mising values(NaNs)

- columns with zero or NaN(i.e in_reply_to_status_id, in_reply_to_user_id, in_reply_to_status_id, in_reply_to_user_id)
- IDs as objects
- missing rows - in expanded_urls
- timestamp as object

**image table**

- id as int
- column headers - non descriptive

**tweet table**

- empty (Nan or zero) columns - contributors, coordinates, geo,place,quoted status, in_reply to,entities, favorited, id_str, quoted_status, quoted_status_id, retweeted, retweeted_status, source, truncated, user
- missing values (extended_entities)
- change id to tweet_id

**Tidy**

- combine the 3 dfs together
- dfs have different lengths(2356, 2075, 2341)

I fixed the data as follows coz i assumed....

```
In [85]: ## Clean

In [86]: # make copies
         twitter_ach_clean = twitter_archive_df.copy()
         image_clean = image_df.copy()
         tweet_clean = tweet_df.copy()
```

**Define** 1.Drop empty or columns with zero or NaN values in twitter_clean and tweet_clean using pandas .drop() method.

```
In [87]: tweet_clean = tweet_clean.drop(['contributors','coordinates','entities', 'favorited','g
         tweet_clean = tweet_clean.drop(['in_reply_to_screen_name', 'in_reply_to_status_id', 'in

In [88]: twitter_clean = twitter_ach_clean.drop(['in_reply_to_status_id', 'in_reply_to_user_id',

In [89]: #### Test
         twitter_clean.head(2)

Out[89]:              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 pupper puppo
0                  10  Phineas  None    None   None  None
1                  10    Tilly  None    None   None  None
```

In [90]:
```
#Test
tweet_clean.sample(5)
```

Out[90]:
```
              created_at display_text_range  \
2214 2015-11-22 02:34:57           [0, 138]
1587 2016-01-10 01:54:44           [0, 133]
2042 2015-11-30 15:59:17            [0, 68]
510  2016-12-17 22:43:27           [0, 118]
2103 2015-11-28 02:00:17            [0, 40]

                                     extended_entities  favorite_count  \
2214  {'media': [{'id': 668256315861504000, 'id_str'...            1325
1587  {'media': [{'id': 686003200739115008, 'id_str'...            1919
2042  {'media': [{'id': 671357838413996033, 'id_str'...             406
510   {'media': [{'id': 810254102546489344, 'id_str'...           15737
2103  {'media': [{'id': 670421917862649857, 'id_str'...            1338

                                           full_text                  id  \
2214  This is Jareld. Jareld rules these waters. Lad...  668256321989451776
1587  This is Hammond. He's a peculiar pup. Loves lo...  686003207160610816
2042  Tfw she says hello from the other side. 9/10 h...  671357843010908160
510   This is Gus. He likes to be close to you, whic...  810254108431155201
2103           Meet Herb. 12/10 https://t.co/tLRyYvCci3  670421925039075328

     quoted_status_permalink  retweet_count  \
2214                     NaN            636
1587                     NaN            678
2042                     NaN            151
510                      NaN           3690
2103                     NaN            653

                                                   user
2214  {'id': 4196983835, 'id_str': '4196983835', 'na...
```

```
1587  {'id': 4196983835, 'id_str': '4196983835', 'na...
2042  {'id': 4196983835, 'id_str': '4196983835', 'na...
510   {'id': 4196983835, 'id_str': '4196983835', 'na...
2103  {'id': 4196983835, 'id_str': '4196983835', 'na...
```

In [91]:
```
image_clean.head(2)
```

Out[91]:
```
             tweet_id                                          jpg_url  \
0  666020888022790149  https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg
1  666029285002620928  https://pbs.twimg.com/media/CT42GRgUYAA5iDo.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

    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
```

In [92]:
```
# different lengths
len(twitter_clean) , len(image_clean), len(tweet_clean)
```

Out[92]:
```
(2356, 2075, 2342)
```

**Define**

## 2 Tidy issue

## 3 combine the 3 dfs together - ensure that they all share the same column 'tweet_id'

In [93]:
```
# change 'id' in tweet_clean  to 'tweet_id' & test
tweet_clean = tweet_clean.rename(columns ={'id': 'tweet_id'})
tweet_clean.head(2)
```

Out[93]:
```
           created_at display_text_range  \
0 2017-08-01 16:23:56           [0, 85]
1 2017-08-01 00:17:27          [0, 138]

                                   extended_entities  favorite_count  \
0  {'media': [{'id': 892420639486877696, 'id_str'...           38151
1  {'media': [{'id': 892177413194625024, 'id_str'...           32715

                                           full_text            tweet_id  \
0  This is Phineas. He's a mystical boy. Only eve...  892420643555336193
1  This is Tilly. She's just checking pup on you...  892177421306343426
```

```
  quoted_status_permalink  retweet_count  \
0                     NaN           8347
1                     NaN           6170

                                                  user
0  {'id': 4196983835, 'id_str': '4196983835', 'na...
1  {'id': 4196983835, 'id_str': '4196983835', 'na...
```

### Code

```
In [94]: # to merge the 3 dfs together
         from functools import reduce
         dfs = [twitter_clean, image_clean, tweet_clean]
         df_final = reduce(lambda left,right: pd.merge(left,right,on='tweet_id'), dfs)
```

### Test

```
In [95]: df_final.sample(3)
```

Out[95]:

```
                tweet_id                  timestamp  \
1855  670037189829525505  2015-11-27 00:31:29 +0000
278   830583320585068544  2017-02-12 01:04:29 +0000
1509  677698403548192770  2015-12-18 03:54:25 +0000

                                                source  \
1855  <a href="http://twitter.com/download/iphone" r...
278   <a href="http://twitter.com/download/iphone" r...
1509  <a href="http://twitter.com/download/iphone" r...

                                                   text  \
1855  Awesome dog here. Not sure where it is tho. Sp...
278   This is Lilly. She just parallel barked. Kindl...
1509  This is Sadie. She got her holidays confused. ...

                                           expanded_urls  rating_numerator  \
1855  https://twitter.com/dog_rates/status/670037189...                 5
278   https://twitter.com/dog_rates/status/830583320...                13
1509  https://twitter.com/dog_rates/status/677698403...                 9

      rating_denominator   name doggo floofer  \
1855                  10   None  None    None
278                   10  Lilly  None    None
1509                  10  Sadie  None    None

                          ...                          p3_conf p3_dog  \
1855                      ...                         0.050728  False
278                       ...                         0.011933   True
1509                      ...                         0.020126   True
```

```
              created_at  display_text_range  \
1855 2015-11-27 00:31:29           [0, 140]
278  2017-02-12 01:04:29            [0, 94]
1509 2015-12-18 03:54:25            [0, 88]

                                     extended_entities  favorite_count  \
1855  {'media': [{'id': 670037180094488576, 'id_str'...             598
278   {'media': [{'id': 830583314243268608, 'id_str'...           70716
1509  {'media': [{'id': 677698398770888704, 'id_str'...            1277

                                           full_text  \
1855  Awesome dog here. Not sure where it is tho. Sp...
278   This is Lilly. She just parallel barked. Kindl...
1509  This is Sadie. She got her holidays confused. ...

     quoted_status_permalink  retweet_count  \
1855                     NaN            287
278                      NaN          18241
1509                     NaN            339

                                                   user
1855  {'id': 4196983835, 'id_str': '4196983835', 'na...
278   {'id': 4196983835, 'id_str': '4196983835', 'na...
1509  {'id': 4196983835, 'id_str': '4196983835', 'na...

[3 rows x 31 columns]
```

**Define**

## 4 Drop source,text_x, extended_entities, user from the merged df_final

```
In [96]: df_final = df_final.drop(['source',  'extended_entities', 'user',], axis = 1)

In [97]: #### Testing

In [98]: df_final.head(3)
```

Out[98]:

```
             tweet_id                  timestamp  \
0  892420643555336193  2017-08-01 16:23:56 +0000
1  892177421306343426  2017-08-01 00:17:27 +0000
2  891815181378084864  2017-07-31 00:18:03 +0000

                                                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...
```

```
                                        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

   rating_denominator     name doggo floofer pupper      ...      p2_dog  \
0                  10  Phineas  None    None   None      ...       False
1                  10    Tilly  None    None   None      ...        True
2                  10   Archie  None    None   None      ...        True

         p3   p3_conf p3_dog          created_at  display_text_range  \
0    banana  0.076110  False 2017-08-01 16:23:56             [0, 85]
1  papillon  0.068957   True 2017-08-01 00:17:27            [0, 138]
2    kelpie  0.031379   True 2017-07-31 00:18:03            [0, 121]

  favorite_count                                          full_text  \
0          38151  This is Phineas. He's a mystical boy. Only eve...
1          32715  This is Tilly. She's just checking pup on you...
2          24637  This is Archie. He is a rare Norwegian Pouncin...

   quoted_status_permalink retweet_count
0                      NaN          8347
1                      NaN          6170
2                      NaN          4082

[3 rows x 28 columns]
```

## 4.1 Cleaning

Note:for visual and programmatic quality and tidiness issues. Detect and document at least eight (8) quality issues and two (2) tidiness issues

### Quality

1. 17 tweet_id duplicates

#### twitter table

2. retweets, in_reply_to - not important
3. mising values(NaNs)
4. columns with zero or NaN(i.e in_reply_to_status_id, in_reply_to_user_id, in_reply_to_status_id, in_reply_to_user_id)
5. IDs as objects
6. missing rows - in expanded_urls
7. timestamp as object

#### image table

8. id as int

9. column headers - non descriptive

**tweet table**

10. empty (Nan or zero) columns - contributors, coordinates, geo,place,quoted status, in_reply to,entities, favorited, id_str, quoted_status, quoted_status_id, retweeted, retweeted_status, source, truncated, user
11. missing values (extended_entities)
12. change id to tweet_id

**Tidy**

13. combine the 3 dfs together
14. dfs have different lengths(2356, 2075, 2341)

I fixed the data as follows coz i assumed....

```
In [99]: # to re-assess what cleaning is still pending based on the list above
         df_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2068 entries, 0 to 2067
Data columns (total 28 columns):
tweet_id                  2068 non-null int64
timestamp                 2068 non-null object
text                      2068 non-null object
expanded_urls             2068 non-null object
rating_numerator          2068 non-null int64
rating_denominator        2068 non-null int64
name                      2068 non-null object
doggo                     2068 non-null object
floofer                   2068 non-null object
pupper                    2068 non-null object
puppo                     2068 non-null object
jpg_url                   2068 non-null object
img_num                   2068 non-null int64
p1                        2068 non-null object
p1_conf                   2068 non-null float64
p1_dog                    2068 non-null bool
p2                        2068 non-null object
p2_conf                   2068 non-null float64
p2_dog                    2068 non-null bool
p3                        2068 non-null object
p3_conf                   2068 non-null float64
p3_dog                    2068 non-null bool
created_at                2068 non-null datetime64[ns]
display_text_range        2068 non-null object
favorite_count            2068 non-null int64
full_text                 2068 non-null object
```

```
quoted_status_permalink    0 non-null object
retweet_count              2068 non-null int64
dtypes: bool(3), datetime64[ns](1), float64(3), int64(6), object(15)
memory usage: 426.1+ KB
```

**Define** Drop the oly remaining column 'quoted_status_permalink' with Nans using the drop method since 'created at' is the same as 'timestamp', drop created at

```
In [100]: # code
          df_final = df_final.drop(['quoted_status_permalink',], axis = 1)
```

```
In [101]: # Test
          df_final.head(3)
```

Out[101]:

| | tweet_id | timestamp | text | expanded_urls | rating_numerator | rating_denominator | name | doggo | floofer | pupper | ... | p2_conf | p2_dog | p3 | p3_conf | p3_dog | created_at | display_text_range | favorite_count | full_text | retweet_count |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 892420643555336193 | 2017-08-01 16:23:56 +0000 | This is Phineas. He's a mystical boy. Only eve... | https://twitter.com/dog_rates/status/892420643... | 13 | 10 | Phineas | None | None | None | ... | 0.085851 | False | banana | 0.076110 | False | 2017-08-01 16:23:56 | [0, 85] | 38151 | This is Phineas. He's a mystical boy. Only eve... | 8347 |
| 1 | 892177421306343426 | 2017-08-01 00:17:27 +0000 | This is Tilly. She's just checking pup on you... | https://twitter.com/dog_rates/status/892177421... | 13 | 10 | Tilly | None | None | None | ... | 0.090647 | True | papillon | 0.068957 | True | 2017-08-01 00:17:27 | [0, 138] | 32715 | This is Tilly. She's just checking pup on you... | 6170 |
| 2 | 891815181378084864 | 2017-07-31 00:18:03 +0000 | This is Archie. He is a rare Norwegian Pouncin... | https://twitter.com/dog_rates/status/891815181... | 12 | 10 | Archie | None | None | None | ... | 0.078253 | True | kelpie | 0.031379 | True | 2017-07-31 00:18:03 | [0, 121] | 24637 | This is Archie. He is a rare Norwegian Pouncin... | |

```
2          4082

[3 rows x 27 columns]
```

**Define** Remove the 17 tweet_id duplicates

```
In [102]: ####Code & check
          df_final[df_final.tweet_id.duplicated()]

Out[102]: Empty DataFrame
          Columns: [tweet_id, timestamp, text, expanded_urls, rating_numerator, rating_denominat
          Index: []

          [0 rows x 27 columns]
```

**Define** Change the dtypes of columns 'tweet_id' and 'timestamp' from int, object to string and timedate respectively standardise the timeobject by slicing off the time element

```
In [103]: #### code
          df_final['tweet_id'] = df_final['tweet_id'].astype(object)
          df_final['timestamp'] = pd.to_datetime(df_final.timestamp).astype(str).str[:-9]

In [104]: #### Test
          print('tweet_id', df_final.dtypes.tweet_id)
          print('timestamp', df_final.dtypes.timestamp)

tweet_id object
timestamp object


In [105]: df_final.timestamp.head(5)

Out[105]: 0    2017-08-01
          1    2017-08-01
          2    2017-07-31
          3    2017-07-30
          4    2017-07-29
          Name: timestamp, dtype: object

In [106]: df_final.describe()

Out[106]:        rating_numerator  rating_denominator      img_num      p1_conf  \
          count       2068.000000         2068.000000  2068.000000  2068.000000
          mean          12.263056           10.513056     1.203578     0.594944
          std           40.749075            7.189152     0.562191     0.271201
          min            0.000000            2.000000     1.000000     0.044333
          25%           10.000000           10.000000     1.000000     0.364571
          50%           11.000000           10.000000     1.000000     0.588620
          75%           12.000000           10.000000     1.000000     0.845599
```

```
max         1776.000000          170.000000     4.000000     1.000000
```

```
            p2_conf       p3_conf  favorite_count  retweet_count
count  2.068000e+03  2.068000e+03     2068.000000    2068.000000
mean   1.345462e-01  6.028703e-02     8439.850580    2814.643133
std    1.007553e-01  5.094828e-02    12703.546408    4892.323906
min    1.011300e-08  1.740170e-10        0.000000      12.000000
25%    5.352722e-02  1.616933e-02     1604.250000     598.500000
50%    1.181350e-01  4.933745e-02     3713.500000    1324.500000
75%    1.955618e-01  9.198323e-02    10569.000000    3248.750000
max    4.880140e-01  2.734190e-01   164671.000000   83915.000000
```

In [107]:
```
df_final.p1_dog.value_counts()
```

Out[107]:
```
True     1528
False     540
Name: p1_dog, dtype: int64
```

In [108]:
```
df_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2068 entries, 0 to 2067
Data columns (total 27 columns):
tweet_id              2068 non-null object
timestamp             2068 non-null object
text                  2068 non-null object
expanded_urls         2068 non-null object
rating_numerator      2068 non-null int64
rating_denominator    2068 non-null int64
name                  2068 non-null object
doggo                 2068 non-null object
floofer               2068 non-null object
pupper                2068 non-null object
puppo                 2068 non-null object
jpg_url               2068 non-null object
img_num               2068 non-null int64
p1                    2068 non-null object
p1_conf               2068 non-null float64
p1_dog                2068 non-null bool
p2                    2068 non-null object
p2_conf               2068 non-null float64
p2_dog                2068 non-null bool
p3                    2068 non-null object
p3_conf               2068 non-null float64
p3_dog                2068 non-null bool
created_at            2068 non-null datetime64[ns]
display_text_range    2068 non-null object
favorite_count        2068 non-null int64
full_text             2068 non-null object
```

```
retweet_count         2068 non-null int64
dtypes: bool(3), datetime64[ns](1), float64(3), int64(5), object(15)
memory usage: 410.0+ KB
```

**Define** Reduce the number of columns Rename some columns

```
In [109]: #### code
          df_final = df_final.rename(columns = {'rating_numerator':'numerator', 'rating_denomina
          df_final = df_final.drop(['text', 'expanded_urls', 'img_num', 'created_at', 'display_t
```

```
In [110]: #### Test
          df_final.sample(3)
```

```
Out[110]:                tweet_id   timestamp  numerator  denominator  name doggo  \
          1646  674045139690631169  2015-12-08          3           10  None  None
          898   736010884653420544  2016-05-27         10           10  None  None
          1965  667801013445750784  2015-11-20         12           10  None  None

               floofer pupper puppo                                          jpg_url  \
          1646    None   None  None  https://pbs.twimg.com/media/CVqwedgXIAEAT6A.jpg
          898     None   None  None  https://pbs.twimg.com/media/CjbV-lEWgAAr6WY.jpg
          1965    None   None  None  https://pbs.twimg.com/media/CUSBemVUEAAn-6V.jpg

                    ...       p1_dog                        p2   p2_conf p2_dog  \
          1646      ...        False         rhinoceros_beetle  0.110607  False
          898       ...         True        Labrador_retriever  0.119475   True
          1965      ...         True  Chesapeake_Bay_retriever  0.262239   True

                                    p3   p3_conf p3_dog  favorite_count  \
          1646  European_fire_salamander  0.043178  False            1452
          898                   bluetick  0.077475   True            8285
          1965    curly-coated_retriever  0.048920   True             334

                                                  full_text  retweet_count
          1646  Herd of wild dogs here. Not sure what they're ...            664
          898   Right after you graduate vs when you remember ...           3183
          1965               OMIGOD 12/10 https://t.co/SVMF4Frf1w             97

          [3 rows x 22 columns]
```

**Define**

## 5 Correct timestamp above

```
In [111]: #### Define
          #Correct timestamp above
```

```
In [112]: df_final.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 2068 entries, 0 to 2067
Data columns (total 22 columns):
tweet_id          2068 non-null object
timestamp         2068 non-null object
numerator         2068 non-null int64
denominator       2068 non-null int64
name              2068 non-null object
doggo             2068 non-null object
floofer           2068 non-null object
pupper            2068 non-null object
puppo             2068 non-null object
jpg_url           2068 non-null object
p1                2068 non-null object
p1_conf           2068 non-null float64
p1_dog            2068 non-null bool
p2                2068 non-null object
p2_conf           2068 non-null float64
p2_dog            2068 non-null bool
p3                2068 non-null object
p3_conf           2068 non-null float64
p3_dog            2068 non-null bool
favorite_count    2068 non-null int64
full_text         2068 non-null object
retweet_count     2068 non-null int64
dtypes: bool(3), float64(3), int64(4), object(12)
memory usage: 329.2+ KB


In [113]: #### code, test
          df_final['tweet_id'] = df_final['tweet_id'].astype(str)
          df_final['timestamp'] = pd.to_datetime(df_final.timestamp)

In [114]: df_final['timestamp'].head(2)

Out[114]: 0   2017-08-01
          1   2017-08-01
          Name: timestamp, dtype: datetime64[ns]

In [115]: df_final.head(3)

Out[115]:              tweet_id  timestamp  numerator  denominator     name doggo  \
          0  892420643555336193 2017-08-01         13           10  Phineas  None
          1  892177421306343426 2017-08-01         13           10    Tilly  None
          2  891815181378084864 2017-07-31         12           10   Archie  None

            floofer pupper puppo                                       jpg_url  \
```

```
0    None   None  None  https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg
1    None   None  None  https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg
2    None   None  None  https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg

       ...       p1_dog        p2   p2_conf p2_dog        p3   p3_conf p3_dog  \
0      ...        False     bagel  0.085851  False    banana  0.076110  False
1      ...         True  Pekinese  0.090647   True  papillon  0.068957   True
2      ...         True  malamute  0.078253   True    kelpie  0.031379   True

   favorite_count                                          full_text  \
0           38151  This is Phineas. He's a mystical boy. Only eve...
1           32715  This is Tilly. She's just checking pup on you...
2           24637  This is Archie. He is a rare Norwegian Pouncin...

   retweet_count
0           8347
1           6170
2           4082

[3 rows x 22 columns]
```

In [116]:
```
# Jupyter is not so slow and rerunning the api hampers my progress so I got to break i
# note that this is part of the wrangle_act so  I saved it to a temprary file `tempdf_
# hoping that after cleaning thats when I get it to the  twitter_archive_ csv file
df_final.to_csv('tempdf_final.csv', index=False, encoding = 'utf-8')
```

#### 5.0.1 Transfered to another notebook ('wrangle_act_p2') to finalise cleaning

- server slow and
- due to the frequent api calls every time I have to rerun the notebook