wrangle_report

July 4, 2021

1 Wrangle and Analyze Data

2 1. Import libraries

```
[1]: import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
     import json
     import datetime
     import requests
     import tweepy
     from tweepy import OAuthHandler
     from timeit import default_timer as timer
     from subprocess import call
     from tqdm import trange,tqdm
```

3 2. Gathering Data

3.1 2.1 WeRateDogs Twitter archive

```
[2]: df1 = pd.read_csv('twitter-archive-enhanced.csv')
     df1.shape
```

```
[2]: (2356, 17)
```

```
[3]: df1.head(2)
```

```
[3]:              tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
     0  892420643555336193                    NaN                  NaN
     1  892177421306343426                    NaN                  NaN

                        timestamp  \
     0  2017-08-01 16:23:56 +0000
     1  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  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

   retweeted_status_user_id retweeted_status_timestamp  \
0                       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
```

3.2 2.2 Tweet image predictions

[4]:
```
URL = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/
 ↪599fd2ad_image-predictions/image-predictions.tsv'
get_requests = requests.get(URL)
with open (URL.split('/')[-1], mode='wb') as file:
    file.write(get_requests.content)
```

[5]:
```
df2 = pd.read_csv('image-predictions.tsv', sep='\t')
df2.shape
```

[5]:
```
(2075, 12)
```

[6]:
```
df2.head(2)
```

[6]:
```
             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
```

3.3 2.3 tweet's retweet count and favorite

[7]:

```
with open('creds.json') as f:
    twitter_creds = json.load(f)
```

[8]:

```
### Authentication
auth = OAuthHandler(twitter_creds['API_KEY'], twitter_creds['API_SECRET_KEY'])
auth.set_access_token(twitter_creds['ACCESS_TOKEN'],␣
 ↪twitter_creds['ACCESS_TOKEN_SECRET'])
api = tweepy.API(auth, parser = tweepy.parsers.JSONParser(),␣
 ↪wait_on_rate_limit=True)
```

[45]:

```
### Twitter query
with open('tweet_json.txt', 'w') as json_file:
    for tweet_id in tqdm(df1.tweet_id.unique()):
        try:
            status = api.get_status(tweet_id)
            json_file.write(json.dumps(status))
            json_file.write('\n')
        except tweepy.TweepError as e:
            pass
```

```
100%|      | 2356/2356 [35:02<00:00,  1.12it/s]
```

[9]:

```
### Read json file
df3 = pd.
 ↪DataFrame(columns=['tweet_id','retweet_count','favorite_count','followers_count','retweeted
with open('tweet_json.txt', encoding='utf-8') as json_file:
    for status in json_file:
        data = json.loads(status)
        tweet_id = data['id']
        retweet_count = data['retweet_count']
        favorite_count = data['favorite_count']
        followers_count = data['user']['followers_count']
        full_text = data['text']
        original_url = full_text[full_text.find('https'):]
        retweeted_status = data['retweeted_status'] = data.
 ↪get('retweeted_status', 'Original tweet')
        if retweeted_status == 'Original tweet':
            url = original_url
        else:
            retweeted_status = 'Retweet'
            url = 'Retweet'
        tweet_dict = {'tweet_id': tweet_id,'retweet_count':␣
 ↪retweet_count,'favorite_count': favorite_count,'followers_count':␣
 ↪followers_count,'retweeted_status': retweeted_status,'url': url}
        df3 = df3.append(tweet_dict, ignore_index=True)
```

```
df3.shape
```

```
[9]: (2328, 6)
```

```
[10]: df3.head(2)
```

```
[10]:              tweet_id retweet_count favorite_count followers_count  \
      0  892420643555336193          7237          34675         9001806
      1  892177421306343426          5420          30046         9001806

        retweeted_status                      url
      0   Original tweet  https://t.co/MgUWQ76dJU
      1   Original tweet  https://t.co/aQFSeaCu9L
```

4 3. Assessing Data

4.0.1 Head

```
[11]: df1.head(2)
```

```
[11]:              tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
      0  892420643555336193                    NaN                  NaN
      1  892177421306343426                    NaN                  NaN

                         timestamp  \
      0  2017-08-01 16:23:56 +0000
      1  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  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

         retweeted_status_user_id retweeted_status_timestamp  \
      0                       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
```

[12]:
```
df2.head(2)
```

[12]:
```
             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
```

[13]:
```
df3.head(2)
```

[13]:
```
             tweet_id retweet_count favorite_count followers_count  \
0  892420643555336193          7237          34675         9001806
1  892177421306343426          5420          30046         9001806

  retweeted_status                      url
0   Original tweet  https://t.co/MgUWQ76dJU
1   Original tweet  https://t.co/aQFSeaCu9L
```

4.0.2 Null values

[14]:
```
df1.info()
```

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

[15]:
```
df2.info()
```

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

[16]:
```
df3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2328 entries, 0 to 2327
Data columns (total 6 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   tweet_id          2328 non-null   object
 1   retweet_count     2328 non-null   object
 2   favorite_count    2328 non-null   object
 3   followers_count   2328 non-null   object
 4   retweeted_status  2328 non-null   object
 5   url               2328 non-null   object
dtypes: object(6)
memory usage: 109.2+ KB
```

—> We can see that only df1 has null values, so we have to make an analysis

4.0.3 Duplicates

```
[17]: df1[df1.duplicated()].shape
```

```
[17]: (0, 17)
```

```
[18]: df2[df2.duplicated()].shape
```

```
[18]: (0, 12)
```

```
[19]: df3[df3.duplicated()].shape
```

```
[19]: (0, 6)
```

4.0.4 Statistical description

```
[20]: df1.describe()
```

```
[20]:            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
```

[21]:
```
df2.describe()
```

[21]:
```
           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
```

[22]:
```
df3.describe()
```

[22]:
```
                  tweet_id  retweet_count  favorite_count  followers_count  \
count                 2328           2328            2328             2328
unique                2328           1670            1958               36
top     891815181378084864            494               0          9001814
freq                     1              6             161              295

       retweeted_status      url
count              2328     2328
unique                2     2132
top      Original tweet  Retweet
freq               2167      161
```

4.0.5 Unique values

[23]:
```
df1.nunique()
```

[23]:
```
tweet_id                      2356
in_reply_to_status_id           77
in_reply_to_user_id             31
timestamp                     2356
source                           4
text                          2356
retweeted_status_id            181
retweeted_status_user_id        25
retweeted_status_timestamp     181
expanded_urls                 2218
rating_numerator                40
rating_denominator              18
name                           957
doggo                            2
floofer                          2
pupper                           2
puppo                            2
```

```
dtype: int64
```

[24]:
```
df2.nunique()
```

[24]:
```
tweet_id    2075
jpg_url     2009
img_num        4
p1           378
p1_conf     2006
p1_dog         2
p2           405
p2_conf     2004
p2_dog         2
p3           408
p3_conf     2006
p3_dog         2
dtype: int64
```

[25]:
```
df3.nunique()
```

[25]:
```
tweet_id            2328
retweet_count       1670
favorite_count      1958
followers_count       36
retweeted_status       2
url                 2132
dtype: int64
```

[26]:
```
df1.name.value_counts()
```

[26]:
```
None       745
a           55
Charlie     12
Lucy        11
Oliver      11
          …
Bertson      1
Rover        1
Bloo         1
Chuck        1
Kirk         1
Name: name, Length: 957, dtype: int64
```

[28]:
```
df1.doggo.value_counts()
```

[28]:
```
None     2259
doggo      97
```

```
      Name: doggo, dtype: int64
```

```
[29]: df1.floofer.value_counts()
```

```
[29]: None       2346
      floofer      10
      Name: floofer, dtype: int64
```

```
[30]: df1.pupper.value_counts()
```

```
[30]: None      2099
      pupper     257
      Name: pupper, dtype: int64
```

```
[31]: df1.puppo.value_counts()
```

```
[31]: None     2326
      puppo      30
      Name: puppo, dtype: int64
```

```
[37]: df2.p1.value_counts()
```

```
[37]: golden_retriever      150
      Labrador_retriever    100
      Pembroke               89
      Chihuahua              83
      pug                    57
                           …
      African_crocodile       1
      clog                    1
      walking_stick           1
      crash_helmet            1
      book_jacket             1
      Name: p1, Length: 378, dtype: int64
```

```
[38]: df3.retweeted_status.value_counts()
```

```
[38]: Original tweet    2167
      Retweet            161
      Name: retweeted_status, dtype: int64
```

```
[ ]:
```

After this analysis we can say: 1) **df1 (Archive data) - Quality Issue:** 1) **Null values** in [in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls] 2) **tweet_id** column has *int* type, but the other “id” columns have float type, so to be consistent we have to convert all “id” columns to *float* type. 3) **timestamp** and **retweeted_status_timestamp** columns have *object* type, but that columns must have *timestamp* type. 4) **name** column has 745 rows with “None” value and 55 values with “a” value. - **Tidiness Issue:** 1) **doggo**, **floofer**, **pupper** and **puppo** almost their values are “None” and few other cases hace the same value, so it would be better remove these columns. 2) **df2 (Image data) - Quality Issue:** 1) **p1**, **p2** and **p3** have invalid data, there are rows with cases such as laptop, restaurant, basketball, tricycle, etc. 2) **p1_conf**, **p2_conf** and **p3_conf** have lower values (near to zero), so this indicates that there are predictions made with underestimation. 3) In order to merge with the other dataframes, we have to change the **tweet_id** column’s type to *float*. - **Tidiness Issue:** 1) The **predictions** could be combined in two columns, the label with the higher value and the condifidence with this higher value. 3) **df3 (tweet’s retweet info) - Quality Issue:** 1) There are 161 retweets. 2) In order to merge with the other dataframes, we have to change the **tweet_id** column’s type to *float*. - **Tidiness Issue:** 1) We have to merge the three datasets.

5 4. Cleaning Data

5.1 4.1 Merge the three datasets

5.1.1 4.1.1 Define

- Merge with *concat* method

5.1.2 4.1.2 Code

```
[59]: df_merge = pd.concat([df1, df2, df3], join='outer', axis=1)
```

5.1.3 4.1.3 Test

```
[60]: df_merge.shape
```

```
[60]: (2356, 35)
```

```
[61]: df_merge.head(2)
```

```
[61]:              tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
      0  892420643555336193                    NaN                  NaN
      1  892177421306343426                    NaN                  NaN

                        timestamp  \
      0  2017-08-01 16:23:56 +0000
      1  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  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

   retweeted_status_user_id retweeted_status_timestamp  \
0                       NaN                        NaN
1                       NaN                        NaN

                                     expanded_urls  …  p2_dog  \
0  https://twitter.com/dog_rates/status/892420643…  …    True
1  https://twitter.com/dog_rates/status/892177421…  …    True

                  p3   p3_conf p3_dog            tweet_id retweet_count  \
0   Shetland_sheepdog  0.061428   True  892420643555336193          7237
1  Rhodesian_ridgeback  0.072010   True  892177421306343426          5420

  favorite_count  followers_count retweeted_status                      url
0          34675          9001806   Original tweet  https://t.co/MgUWQ76dJU
1          30046          9001806   Original tweet  https://t.co/aQFSeaCu9L

[2 rows x 35 columns]
```

5.2 4.2 Remove null values

5.2.1 4.2.1 Define

- Remove columns with missing data

5.2.2 4.2.2 Code

[62]:
```
df_merge = df_merge.drop(['in_reply_to_status_id', 'in_reply_to_user_id',␣
↪'retweeted_status_id', 'retweeted_status_user_id',␣
↪'retweeted_status_timestamp', 'expanded_urls'], axis=1)
```

5.2.3 4.1.3 Test

[63]:
```
df_merge.shape
```

[63]:
```
(2356, 29)
```

[64]:
```
df_merge.head(2)
```

[64]:
```
             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  rating_numerator  \
0  This is Phineas. He's a mystical boy. Only eve…                13
1  This is Tilly. She's just checking pup on you…                13

   rating_denominator     name doggo floofer pupper  … p2_dog  \
0                  10  Phineas  None    None   None  …   True
1                  10    Tilly  None    None   None  …   True

                  p3   p3_conf  p3_dog            tweet_id  retweet_count  \
0   Shetland_sheepdog  0.061428    True  892420643555336193           7237
1  Rhodesian_ridgeback  0.072010    True  892177421306343426           5420

  favorite_count followers_count  retweeted_status                      url
0          34675         9001806    Original tweet  https://t.co/MgUWQ76dJU
1          30046         9001806    Original tweet  https://t.co/aQFSeaCu9L

[2 rows x 29 columns]
```

5.3 4.3 Remove doggo, floofer, pupper and puppo columns

5.3.1 4.3.1 Define

- Remove these columns because almost their values are “None” and few other cases hace the same value

5.3.2 4.2.2 Code

[65]:
```
df_merge = df_merge.drop(['doggo', 'floofer', 'pupper', 'puppo'], axis=1)
```

5.3.3 4.1.3 Test

[66]:
```
df_merge.shape
```

[66]:
```
(2356, 25)
```

[67]:
```
df_merge.head(2)
```

[67]:
```
             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  rating_numerator  \
0  This is Phineas. He's a mystical boy. Only eve…                13
```

```
1  This is Tilly. She's just checking pup on you…               13

   rating_denominator     name      tweet_id  \
0                  10  Phineas  6.660209e+17
1                  10    Tilly  6.660293e+17

                                        jpg_url  img_num  … p2_dog  \
0  https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg      1.0  …   True
1  https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg      1.0  …   True

                   p3   p3_conf p3_dog            tweet_id retweet_count  \
0    Shetland_sheepdog  0.061428   True  892420643555336193          7237
1  Rhodesian_ridgeback  0.072010   True  892177421306343426          5420

  favorite_count  followers_count retweeted_status                      url
0          34675          9001806   Original tweet  https://t.co/MgUWQ76dJU
1          30046          9001806   Original tweet  https://t.co/aQFSeaCu9L

[2 rows x 25 columns]
```

5.4 4.4 Column types

5.4.1 4.4.1 Define

- Convert columns to their correct types

5.4.2 4.4.2 Code

[72]:

```
df_merge['tweet_id'] = df_merge['tweet_id'].astype(str)
df_merge['timestamp'] = pd.to_datetime(df_merge['timestamp'])
df_merge['retweet_count'] = df_merge['retweet_count'].astype(float)
df_merge['favorite_count'] = df_merge['favorite_count'].astype(float)
df_merge['followers_count'] = df_merge['followers_count'].astype(float)
```

5.4.3 4.4.3 Test

[73]:

```
df_merge.dtypes
```

[73]:

```
tweet_id                            object
timestamp              datetime64[ns, UTC]
source                              object
text                                object
rating_numerator                     int64
rating_denominator                   int64
name                                object
tweet_id                            object
jpg_url                             object
```

```
img_num                              float64
p1                                    object
p1_conf                              float64
p1_dog                                object
p2                                    object
p2_conf                              float64
p2_dog                                object
p3                                    object
p3_conf                              float64
p3_dog                                object
tweet_id                              object
retweet_count                        float64
favorite_count                       float64
followers_count                      float64
retweeted_status                      object
url                                   object
dtype: object
```

5.5 4.5 Duplicated columns

5.5.1 4.5.1 Define

- *tweet_id* column appears three times due to the merge of dataframes, so we have to remove duplicated columns

5.5.2 4.5.2 Code

```
[85]: df_merge = df_merge.loc[:,~df_merge.columns.duplicated()]
```

5.5.3 4.5.3 Test

```
[86]: df_merge.columns
```

```
[86]: Index(['tweet_id', 'timestamp', 'source', 'text', 'rating_numerator',
             'rating_denominator', 'name', 'jpg_url', 'img_num', 'p1', 'p1_conf',
             'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog',
             'retweet_count', 'favorite_count', 'followers_count',
             'retweeted_status', 'url'],
            dtype='object')
```

```
[87]: df_merge.head(2)
```

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

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

```
1  <a href="http://twitter.com/download/iphone" r…

                                                text  rating_numerator  \
0  This is Phineas. He's a mystical boy. Only eve…                13
1  This is Tilly. She's just checking pup on you…                13

   rating_denominator     name  \
0                  10  Phineas
1                  10    Tilly

                                           jpg_url  img_num  \
0  https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg      1.0
1  https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg      1.0

                       p1  …   p2_conf p2_dog                   p3  \
0  Welsh_springer_spaniel  …  0.156665    True    Shetland_sheepdog
1                 redbone  …  0.074192    True  Rhodesian_ridgeback

    p3_conf p3_dog retweet_count  favorite_count followers_count  \
0  0.061428   True        7237.0         34675.0       9001806.0
1  0.072010   True        5420.0         30046.0       9001806.0

   retweeted_status                     url
0    Original tweet  https://t.co/MgUWQ76dJU
1    Original tweet  https://t.co/aQFSeaCu9L

[2 rows x 23 columns]
```

5.6 4.6 Quality issue

5.6.1 4.6.1 Define

- *name* column has 745 rows with “None” value and 55 values with “a” value.

5.6.2 4.6.2 Code

[74]:

```
df_merge['name'] = df_merge['name'].replace(['None', 'a'], np.nan)
```

5.6.3 4.6.3 Test

[75]:

```
df_merge[df_merge['name'] == 'None'].shape
```

[75]:

```
(0, 25)
```

[76]:

```
df_merge[df_merge['name'] == 'a'].shape
```

[76]:

```
(0, 25)
```

[ ]:

```
```