wrangle_act

May 18, 2020

```
In [105]: import pandas as pd
          import numpy as np
          import requests
          import json
          import statsmodels.api as sm
          import matplotlib.pyplot as plt
          %matplotlib inline
```

0.0.1 Gathering Data

1. reading csv

```
In [106]: twitter_archive = pd.read_csv('twitter_archive_enhanced.csv', sep=',')
```

2. Image prediction data

```
In [107]: url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predict
          r = requests.get(url)
          open('image_predictions.tsv', 'wb').write(r.content)

Out[107]: 335079
```

```
In [108]: image_predictions = pd.read_csv('image_predictions.tsv', sep = '\t')
```

3. Tweeter retweet data

```
In [109]: tweets_list =[]

          with open('tweet_json.txt') as json_file:
              for line in json_file:

                  tweets_dict = {}
                  tweets_json = json.loads(line)

                  try:
                      tweets_dict['tweet_id'] = tweets_json['extended_entities']['media'][0]['id
                  except:
                      tweets_dict['tweet_id'] = 'na'
                  tweets_dict['retweet_count'] = tweets_json['retweet_count']
                  tweets_dict['favorite_count'] = tweets_json['favorite_count']

                  tweets_list.append(tweets_dict)
```

```
In [110]: tweets_df = pd.DataFrame(tweets_list)
```

0.0.2 Assessing Data

```
In [111]: twitter_archive.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 [112]: twitter_archive.tail()
```

```
Out[112]:                 tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
          2351  666049248165822465                    NaN                  NaN
          2352  666044226329800704                    NaN                  NaN
          2353  666033412701032449                    NaN                  NaN
          2354  666029285002620928                    NaN                  NaN
          2355  666020888022790149                    NaN                  NaN

                                timestamp  \
          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  \
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  \
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  \
2351                       NaN                        NaN
2352                       NaN                        NaN
2353                       NaN                        NaN
2354                       NaN                        NaN
2355                       NaN                        NaN

                                          expanded_urls  rating_numerator  \
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
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
```

```
In [113]: twitter_archive.isnull().sum()
```

```
Out[113]: tweet_id                         0
          in_reply_to_status_id         2278
          in_reply_to_user_id           2278
          timestamp                        0
          source                           0
          text                             0
          retweeted_status_id           2175
          retweeted_status_user_id      2175
          retweeted_status_timestamp    2175
          expanded_urls                   59
          rating_numerator                 0
          rating_denominator               0
          name                             0
          doggo                            0
          floofer                          0
          pupper                           0
          puppo                            0
          dtype: int64
```

```
In [114]: twitter_archive.rating_numerator.value_counts().head()
```

```
Out[114]: 12    558
          11    464
          10    461
          13    351
          9     158
          Name: rating_numerator, dtype: int64
```

```
In [115]: print(twitter_archive.doggo.value_counts())
          print()
          print(twitter_archive.floofer.value_counts())
          print()
          print(twitter_archive.pupper.value_counts())
          print()
          print(twitter_archive.puppo.value_counts())
```

```
None     2259
doggo      97
Name: doggo, dtype: int64

None       2346
floofer      10
Name: floofer, dtype: int64

None      2099
pupper     257
Name: pupper, dtype: int64

None     2326
puppo      30
Name: puppo, dtype: int64
```

```
In [116]: twitter_archive.retweeted_status_id.notnull().sum()
```

```
Out[116]: 181
```

```
In [117]: twitter_archive.in_reply_to_status_id.notnull().sum()

Out[117]: 78

In [118]: twitter_archive.duplicated().sum()

Out[118]: 0

In [119]: tweets_df.head()

Out[119]:    favorite_count  retweet_count            tweet_id
          0           39467           8853  892420639486877696
          1           33819           6514  892177413194625024
          2           25461           4328  891815175371796480
          3           42908           8964  891689552724799489
          4           41048           9774  891327551943041024

In [120]: tweets_df.duplicated().sum()

Out[120]: 0

In [121]: tweets_df.retweet_count[-10:]

Out[121]: 2344     61
          2345    146
          2346    261
          2347    879
          2348     60
          2349     41
          2350    147
          2351     47
          2352     48
          2353    532
          Name: retweet_count, dtype: int64

In [122]: image_predictions.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null int64
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
```

```
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

```
In [123]: image_predictions.head()
```

```
Out[123]:              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 [124]: image_predictions.duplicated().sum()
```

```
Out[124]: 0
```

```
In [125]: image_predictions.p1.value_counts().head()
```

```
Out[125]: golden_retriever      150
          Labrador_retriever    100
          Pembroke               89
          Chihuahua              83
          pug                    57
          Name: p1, dtype: int64
```

Quality

1) twitter_archive table

- 'name' is all written by the string including Nones
- 'name' sometimes do not have the right name i.e. "a"
- 'doggo' 'floofer' 'pupper' 'puppo' are written by string including Nones

- 'rating_numerator' values sometimes have outliers
- 'tweet_id' is integer
- we do not need the tweets beyond August 1st, 2017
- delete unnecessary columns regarding to image_predictions and tweets_df table

2) tweets_df table

- 'tweet_id' is integer

3) image_predictions table

- 'p1' 'p2' 'p3' have lower cases
- 'tweet_id' is integer
- it is easier to read the names of 'p1', 'p2', 'p3' without deliminators

Tidiness

- Merge all 3 tables
- Make a new column 'dogs' to show the types of dogs at once

0.0.3 Cleaning

```
In [126]: twitter_archive_clean = twitter_archive.copy()
          tweets_df_clean = tweets_df.copy()
          image_predictions_clean = image_predictions.copy()
```

Quality

Define:

1 'name' is all written by the string including Nones

Code:

```
In [127]: #'name' is all written by the string including Nones
          twitter_archive_clean.name.replace("None", "a", inplace=True)
          twitter_archive_clean.name.tail()

Out[127]: 2351    a
          2352    a
          2353    a
          2354    a
          2355    a
          Name: name, dtype: object
```

```
In [128]: #'name' sometimes do not have the right name i.e. "a"
          bad_names = twitter_archive_clean[twitter_archive_clean['name'].str.islower()]
          bad_names = bad_names['name'].unique()
          bad_names
```

```
Out[128]: array(['a', 'such', 'quite', 'not', 'one', 'incredibly', 'mad', 'an',
                 'very', 'just', 'my', 'his', 'actually', 'getting', 'this',
                 'unacceptable', 'all', 'old', 'infuriating', 'the', 'by',
                 'officially', 'life', 'light', 'space'], dtype=object)
```

```
In [129]: #replace values equals to invalid names with none
          twitter_archive_clean['name'].replace(bad_names, np.nan, inplace=True)
```

Test:

```
In [130]: twitter_archive_clean.name.tail()
```

```
Out[130]: 2351    NaN
          2352    NaN
          2353    NaN
          2354    NaN
          2355    NaN
          Name: name, dtype: object
```

Define:

2 ’doggo’ ’floofer’ ’pupper’ ’puppo’ are written by string including Nones

Code:

```
In [131]: #'doggo' 'floofer' 'pupper' 'puppo' are written by string including Nones
          twitter_archive_clean.doggo.replace("None", np.nan, inplace=True)
          twitter_archive_clean.floofer.replace("None", np.nan, inplace=True)
          twitter_archive_clean.pupper.replace("None", np.nan, inplace=True)
          twitter_archive_clean.puppo.replace("None", np.nan, inplace=True)
```

Test:

```
In [132]: twitter_archive_clean.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                          1502 non-null object
doggo                         97 non-null object
floofer                       10 non-null object
pupper                        257 non-null object
puppo                         30 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

Define:

3 'rating_numerator' values sometimes have outliers

Code:

```
In [133]: #'rating_numerator' values sometimes have outliers
          twitter_archive_clean.rating_numerator.describe()

Out[133]: count    2356.000000
          mean       13.126486
          std        45.876648
          min         0.000000
          25%        10.000000
          50%        11.000000
          75%        12.000000
          max      1776.000000
          Name: rating_numerator, dtype: float64

In [134]: IQR = 12-10
          Lower_Outlier = 10 - (1.5*IQR)
          Higher_Outlier = 12 + (1.5*IQR)
          print(Lower_Outlier)
          print(Higher_Outlier)

7.0
15.0


In [135]: #Using the Outlier Formula, we omit the outliers(<7.0 or >15.0) as None and make a new
          twitter_archive_clean['rating_numerator'] = twitter_archive_clean.query('rating_numera
```

Test:

```
In [136]: twitter_archive_clean['rating_numerator'].value_counts()
```

```
Out[136]: 12.0    558
          11.0    464
          10.0    461
          13.0    351
          9.0     158
          8.0     102
          7.0      55
          14.0     54
          15.0      2
          Name: rating_numerator, dtype: int64
```

Define:

4 switching the dtype of ’tweet_id’ in twitter_archive table

Code:

```
In [137]: #switching the dtype of 'tweet_id' in twitter_archive table
          twitter_archive_clean.tweet_id = twitter_archive_clean.tweet_id.astype(str)
```

Test:

```
In [138]: twitter_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2356 non-null object
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2356 non-null object
source                        2356 non-null object
text                          2356 non-null object
retweeted_status_id           181 non-null float64
retweeted_status_user_id      181 non-null float64
retweeted_status_timestamp    181 non-null object
expanded_urls                 2297 non-null object
rating_numerator              2205 non-null float64
rating_denominator            2356 non-null int64
name                          1502 non-null object
doggo                         97 non-null object
floofer                       10 non-null object
pupper                        257 non-null object
puppo                         30 non-null object
dtypes: float64(5), int64(1), object(11)
memory usage: 313.0+ KB
```

Define:

5 switching the dtype of 'tweet_id' in tweets_df table

Code:

```
In [139]: #switching the dtype of 'tweet_id' in tweets_df table
          tweets_df_clean.tweet_id = tweets_df_clean.tweet_id.astype(str)
```

Test:

```
In [140]: tweets_df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 3 columns):
favorite_count    2354 non-null int64
retweet_count     2354 non-null int64
tweet_id          2354 non-null object
dtypes: int64(2), object(1)
memory usage: 55.2+ KB
```

Define:

6 'p1', 'p2', 'p3' have lower cases in image_predictions table

Code:

```
In [141]: #'p1', 'p2', 'p3' have lower cases in image_predictions table
          image_predictions_clean['p1'] = image_predictions_clean['p1'].str.title()
          image_predictions_clean['p2'] = image_predictions_clean['p2'].str.title()
          image_predictions_clean['p3'] = image_predictions_clean['p3'].str.title()
```

Test:

```
In [142]: image_predictions_clean.head()
```

```
Out[142]:              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
```

Define:

7 switching the dtype of 'tweet_id' in image_predictions table

Code:

```
In [143]: #switching the dtype of 'tweet_id' in image_predictions table
          image_predictions_clean.tweet_id = image_predictions_clean.tweet_id.astype(str)
```

Test:

```
In [144]: image_predictions_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null object
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(1), object(5)
memory usage: 152.1+ KB
```

Define:

8 we do not need the tweets beyond August 1st, 2017

Code:

```
In [145]: #we do not need the tweets beyond August 1st, 2017
          twitter_archive_clean.timestamp = pd.to_datetime(twitter_archive_clean.timestamp, form
          twitter_archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2356 non-null object
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2356 non-null datetime64[ns]
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              2205 non-null float64
rating_denominator            2356 non-null int64
name                          1502 non-null object
doggo                         97 non-null object
floofer                       10 non-null object
pupper                        257 non-null object
puppo                         30 non-null object
dtypes: datetime64[ns](1), float64(5), int64(1), object(10)
memory usage: 313.0+ KB
```

Test:

```
In [146]: #There has already been sorted by August 1st, 2017
          twitter_archive_clean.timestamp.head()
```

```
Out[146]: 0   2017-08-01 16:23:56
          1   2017-08-01 00:17:27
          2   2017-07-31 00:18:03
          3   2017-07-30 15:58:51
          4   2017-07-29 16:00:24
          Name: timestamp, dtype: datetime64[ns]
```

Define:

9 it is easier to read the names of 'p1', 'p2', 'p3' without deliminators

Code:

```
In [147]: #it is easier to read the names of 'p1', 'p2', 'p3' without deliminators
          image_predictions_clean['p1'] = image_predictions_clean['p1'].str.split('_')
          image_predictions_clean['p2'] = image_predictions_clean['p2'].str.split('_')
          image_predictions_clean['p3'] = image_predictions_clean['p3'].str.split('_')
```

Test:

```
In [148]: image_predictions_clean.head()
```

```
Out[148]:
             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  \
0        1  [Welsh, Springer, Spaniel]  0.465074    True
1        1                   [Redbone]  0.506826    True
2        1          [German, Shepherd]  0.596461    True
3        1      [Rhodesian, Ridgeback]  0.408143    True
4        1       [Miniature, Pinscher]  0.560311    True

                      p2   p2_conf  p2_dog                      p3   p3_conf  \
0               [Collie]  0.156665    True    [Shetland, Sheepdog]  0.061428
1  [Miniature, Pinscher]  0.074192    True  [Rhodesian, Ridgeback]  0.072010
2             [Malinois]  0.138584    True            [Bloodhound]  0.116197
3              [Redbone]  0.360687    True   [Miniature, Pinscher]  0.222752
4           [Rottweiler]  0.243682    True              [Doberman]  0.154629

   p3_dog
0    True
1    True
2    True
3    True
4    True
```

Define:

10 delete unnecessary columns regarding to image_predictions and tweets_df table

Code:

```
In [149]: #delete unnecessary columns regarding to image_predictions and tweets_df table
          twitter_archive_clean = twitter_archive_clean[['tweet_id', 'timestamp', 'source', 'tex
```

Test:

```
In [150]: twitter_archive_clean.head()
```

```
Out[150]:              tweet_id           timestamp  \
          0  892420643555336193 2017-08-01 16:23:56
          1  892177421306343426 2017-08-01 00:17:27
          2  891815181378084864 2017-07-31 00:18:03
          3  891689557279858688 2017-07-30 15:58:51
          4  891327558926688256 2017-07-29 16:00:24

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

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

                                                 expanded_urls  rating_numerator  \
          0  https://twitter.com/dog_rates/status/892420643...              13.0
          1  https://twitter.com/dog_rates/status/892177421...              13.0
          2  https://twitter.com/dog_rates/status/891815181...              12.0
          3  https://twitter.com/dog_rates/status/891689557...              13.0
          4  https://twitter.com/dog_rates/status/891327558...              12.0

             rating_denominator      name doggo floofer pupper puppo
          0                  10   Phineas   NaN     NaN    NaN   NaN
          1                  10     Tilly   NaN     NaN    NaN   NaN
          2                  10    Archie   NaN     NaN    NaN   NaN
          3                  10     Darla   NaN     NaN    NaN   NaN
          4                  10  Franklin   NaN     NaN    NaN   NaN
```

Tidiness

Define

11 We need the new combined DataFrames to 'twitter_archive_master'

Code

```
In [151]: #Merge three DataFrames to 'twitter_archive_master'
          twitter_image = pd.merge(twitter_archive_clean, image_predictions_clean)
          tweets_df_clean['tweet_id_retweet'] = tweets_df_clean['tweet_id']
          twitter_archive_master = pd.merge(twitter_archive_clean, tweets_df_clean, how="outer")
```

Test

```
In [152]: twitter_archive_master
```

```
Out[152]:                 tweet_id           timestamp  \
          0     892420643555336193 2017-08-01 16:23:56
          1     892177421306343426 2017-08-01 00:17:27
          2     891815181378084864 2017-07-31 00:18:03
          3     891689557279858688 2017-07-30 15:58:51
          4     891327558926688256 2017-07-29 16:00:24
          5     891087950875897856 2017-07-29 00:08:17
          6     890971913173991426 2017-07-28 16:27:12
          7     890729181411237888 2017-07-28 00:22:40
          8     890609185150312448 2017-07-27 16:25:51
          9     890240255349198849 2017-07-26 15:59:51
          10    890006608113172480 2017-07-26 00:31:25
          11    889880896479866881 2017-07-25 16:11:53
          12    889665388333682689 2017-07-25 01:55:32
          13    889638837579907072 2017-07-25 00:10:02
          14    889531135344209921 2017-07-24 17:02:04
          15    889278841981685760 2017-07-24 00:19:32
          16    888917238123831296 2017-07-23 00:22:39
          17    888804989199671297 2017-07-22 16:56:37
          18    888554962724278272 2017-07-22 00:23:06
          19    888202515573088257 2017-07-21 01:02:36
          20    888078434458587136 2017-07-20 16:49:33
          21    887705289381826560 2017-07-19 16:06:48
          22    887517139158093824 2017-07-19 03:39:09
          23    887473957103951883 2017-07-19 00:47:34
          24    887343217045368832 2017-07-18 16:08:03
          25    887101392804085760 2017-07-18 00:07:08
          26    886983233522544640 2017-07-17 16:17:36
          27    886736880519319552 2017-07-16 23:58:41
          28    886680336477933568 2017-07-16 20:14:00
          29    886366144734445568 2017-07-15 23:25:31
          ...                  ...                 ...
          4680  666411498068123649                 NaT
          4681  666407121513275392                 NaT
          4682  666396240351993856                 NaT
          4683  666373746337402880                 NaT
          4684  666362717482020864                 NaT
          4685  666353280906170368                 NaT
          4686  666345414279471104                 NaT
```

```
4687  666337857791987715                 NaT
4688  666293909010702337                 NaT
4689  666287399580733440                 NaT
4690  666273081518768128                 NaT
4691  666268904428277760                 NaT
4692  666104129232740352                 NaT
4693  666102150364286977                 NaT
4694  666099505364733952                 NaT
4695  666093996847063040                 NaT
4696  666082912819875840                 NaT
4697  666073098362486784                 NaT
4698  666071190449033216                 NaT
4699  666063820255862784                 NaT
4700  666058597072306176                 NaT
4701  666057085227016192                 NaT
4702  666055517517848576                 NaT
4703  666051848592334848                 NaT
4704  666050754986266625                 NaT
4705  666049244999131136                 NaT
4706  666044217047650304                 NaT
4707  666033409081393153                 NaT
4708  666029276303482880                 NaT
4709  666020881337073664                 NaT

                                                 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...
5     <a href="http://twitter.com/download/iphone" r...
6     <a href="http://twitter.com/download/iphone" r...
7     <a href="http://twitter.com/download/iphone" r...
8     <a href="http://twitter.com/download/iphone" r...
9     <a href="http://twitter.com/download/iphone" r...
10    <a href="http://twitter.com/download/iphone" r...
11    <a href="http://twitter.com/download/iphone" r...
12    <a href="http://twitter.com/download/iphone" r...
13    <a href="http://twitter.com/download/iphone" r...
14    <a href="http://twitter.com/download/iphone" r...
15    <a href="http://twitter.com/download/iphone" r...
16    <a href="http://twitter.com/download/iphone" r...
17    <a href="http://twitter.com/download/iphone" r...
18    <a href="http://twitter.com/download/iphone" r...
19    <a href="http://twitter.com/download/iphone" r...
20    <a href="http://twitter.com/download/iphone" r...
21    <a href="http://twitter.com/download/iphone" r...
22    <a href="http://twitter.com/download/iphone" r...
```

```
23    <a href="http://twitter.com/download/iphone" r...
24    <a href="http://twitter.com/download/iphone" r...
25    <a href="http://twitter.com/download/iphone" r...
26    <a href="http://twitter.com/download/iphone" r...
27    <a href="http://twitter.com/download/iphone" r...
28    <a href="http://twitter.com/download/iphone" r...
29    <a href="http://twitter.com/download/iphone" r...
...                                                 ...
4680                                                NaN
4681                                                NaN
4682                                                NaN
4683                                                NaN
4684                                                NaN
4685                                                NaN
4686                                                NaN
4687                                                NaN
4688                                                NaN
4689                                                NaN
4690                                                NaN
4691                                                NaN
4692                                                NaN
4693                                                NaN
4694                                                NaN
4695                                                NaN
4696                                                NaN
4697                                                NaN
4698                                                NaN
4699                                                NaN
4700                                                NaN
4701                                                NaN
4702                                                NaN
4703                                                NaN
4704                                                NaN
4705                                                NaN
4706                                                NaN
4707                                                NaN
4708                                                NaN
4709                                                NaN

                                                   text  \
0     This is Phineas. He's a mystical boy. Only eve...
1     This is Tilly. She's just checking pup on you...
2     This is Archie. He is a rare Norwegian Pouncin...
3     This is Darla. She commenced a snooze mid meal...
4     This is Franklin. He would like you to stop ca...
5     Here we have a majestic great white breaching ...
6     Meet Jax. He enjoys ice cream so much he gets ...
7     When you watch your owner call another dog a g...
```

```
8       This is Zoey. She doesn't want to be one of th...
9       This is Cassie. She is a college pup. Studying...
10      This is Koda. He is a South Australian decksha...
11      This is Bruno. He is a service shark. Only get...
12      Here's a puppo that seems to be on the fence a...
13      This is Ted. He does his best. Sometimes that'...
14      This is Stuart. He's sporting his favorite fan...
15      This is Oliver. You're witnessing one of his m...
16      This is Jim. He found a fren. Taught him how t...
17      This is Zeke. He has a new stick. Very proud o...
18      This is Ralphus. He's powering up. Attempting ...
19      RT @dog_rates: This is Canela. She attempted s...
20      This is Gerald. He was just told he didn't get...
21      This is Jeffrey. He has a monopoly on the pool...
22      I've yet to rate a Venezuelan Hover Wiener. Th...
23      This is Canela. She attempted some fancy porch...
24      You may not have known you needed to see this ...
25      This... is a Jubilant Antarctic House Bear. We...
26      This is Maya. She's very shy. Rarely leaves he...
27      This is Mingus. He's a wonderful father to his...
28      This is Derek. He's late for a dog meeting. 13...
29      This is Roscoe. Another pupper fallen victim t...
                              ...                        
4680                                                  NaN
4681                                                  NaN
4682                                                  NaN
4683                                                  NaN
4684                                                  NaN
4685                                                  NaN
4686                                                  NaN
4687                                                  NaN
4688                                                  NaN
4689                                                  NaN
4690                                                  NaN
4691                                                  NaN
4692                                                  NaN
4693                                                  NaN
4694                                                  NaN
4695                                                  NaN
4696                                                  NaN
4697                                                  NaN
4698                                                  NaN
4699                                                  NaN
4700                                                  NaN
4701                                                  NaN
4702                                                  NaN
4703                                                  NaN
4704                                                  NaN
```

```
4705                                                NaN
4706                                                NaN
4707                                                NaN
4708                                                NaN
4709                                                NaN

                                          expanded_urls  rating_numerator  \
0     https://twitter.com/dog_rates/status/892420643...              13.0
1     https://twitter.com/dog_rates/status/892177421...              13.0
2     https://twitter.com/dog_rates/status/891815181...              12.0
3     https://twitter.com/dog_rates/status/891689557...              13.0
4     https://twitter.com/dog_rates/status/891327558...              12.0
5     https://twitter.com/dog_rates/status/891087950...              13.0
6     https://gofundme.com/ydvmve-surgery-for-jax,ht...              13.0
7     https://twitter.com/dog_rates/status/890729181...              13.0
8     https://twitter.com/dog_rates/status/890609185...              13.0
9     https://twitter.com/dog_rates/status/890240255...              14.0
10    https://twitter.com/dog_rates/status/890006608...              13.0
11    https://twitter.com/dog_rates/status/889880896...              13.0
12    https://twitter.com/dog_rates/status/889665388...              13.0
13    https://twitter.com/dog_rates/status/889638837...              12.0
14    https://twitter.com/dog_rates/status/889531135...              13.0
15    https://twitter.com/dog_rates/status/889278841...              13.0
16    https://twitter.com/dog_rates/status/888917238...              12.0
17    https://twitter.com/dog_rates/status/888804989...              13.0
18    https://twitter.com/dog_rates/status/888554962...              13.0
19    https://twitter.com/dog_rates/status/887473957...              13.0
20    https://twitter.com/dog_rates/status/888078434...              12.0
21    https://twitter.com/dog_rates/status/887705289...              13.0
22    https://twitter.com/dog_rates/status/887517139...              14.0
23    https://twitter.com/dog_rates/status/887473957...              13.0
24    https://twitter.com/dog_rates/status/887343217...              13.0
25    https://twitter.com/dog_rates/status/887101392...              12.0
26    https://twitter.com/dog_rates/status/886983233...              13.0
27    https://www.gofundme.com/mingusneedsus,https:/...              13.0
28    https://twitter.com/dog_rates/status/886680336...              13.0
29    https://twitter.com/dog_rates/status/886366144...              12.0
...                                                 ...               ...
4680                                                NaN               NaN
4681                                                NaN               NaN
4682                                                NaN               NaN
4683                                                NaN               NaN
4684                                                NaN               NaN
4685                                                NaN               NaN
4686                                                NaN               NaN
4687                                                NaN               NaN
4688                                                NaN               NaN
4689                                                NaN               NaN
```

```
4690                                                NaN               NaN
4691                                                NaN               NaN
4692                                                NaN               NaN
4693                                                NaN               NaN
4694                                                NaN               NaN
4695                                                NaN               NaN
4696                                                NaN               NaN
4697                                                NaN               NaN
4698                                                NaN               NaN
4699                                                NaN               NaN
4700                                                NaN               NaN
4701                                                NaN               NaN
4702                                                NaN               NaN
4703                                                NaN               NaN
4704                                                NaN               NaN
4705                                                NaN               NaN
4706                                                NaN               NaN
4707                                                NaN               NaN
4708                                                NaN               NaN
4709                                                NaN               NaN

      rating_denominator      name  doggo floofer  pupper  puppo  \
0                   10.0   Phineas    NaN     NaN     NaN    NaN
1                   10.0     Tilly    NaN     NaN     NaN    NaN
2                   10.0    Archie    NaN     NaN     NaN    NaN
3                   10.0     Darla    NaN     NaN     NaN    NaN
4                   10.0  Franklin    NaN     NaN     NaN    NaN
5                   10.0       NaN    NaN     NaN     NaN    NaN
6                   10.0       Jax    NaN     NaN     NaN    NaN
7                   10.0       NaN    NaN     NaN     NaN    NaN
8                   10.0      Zoey    NaN     NaN     NaN    NaN
9                   10.0    Cassie  doggo     NaN     NaN    NaN
10                  10.0      Koda    NaN     NaN     NaN    NaN
11                  10.0     Bruno    NaN     NaN     NaN    NaN
12                  10.0       NaN    NaN     NaN     NaN  puppo
13                  10.0       Ted    NaN     NaN     NaN    NaN
14                  10.0    Stuart    NaN     NaN     NaN  puppo
15                  10.0    Oliver    NaN     NaN     NaN    NaN
16                  10.0       Jim    NaN     NaN     NaN    NaN
17                  10.0      Zeke    NaN     NaN     NaN    NaN
18                  10.0   Ralphus    NaN     NaN     NaN    NaN
19                  10.0    Canela    NaN     NaN     NaN    NaN
20                  10.0    Gerald    NaN     NaN     NaN    NaN
21                  10.0   Jeffrey    NaN     NaN     NaN    NaN
22                  10.0       NaN    NaN     NaN     NaN    NaN
23                  10.0    Canela    NaN     NaN     NaN    NaN
24                  10.0       NaN    NaN     NaN     NaN    NaN
25                  10.0       NaN    NaN     NaN     NaN    NaN
```

```
26                  10.0      Maya    NaN     NaN     NaN    NaN
27                  10.0    Mingus    NaN     NaN     NaN    NaN
28                  10.0     Derek    NaN     NaN     NaN    NaN
29                  10.0    Roscoe    NaN     NaN  pupper    NaN
...                  ...       ...    ...     ...     ...    ...
4680                 NaN       NaN    NaN     NaN     NaN    NaN
4681                 NaN       NaN    NaN     NaN     NaN    NaN
4682                 NaN       NaN    NaN     NaN     NaN    NaN
4683                 NaN       NaN    NaN     NaN     NaN    NaN
4684                 NaN       NaN    NaN     NaN     NaN    NaN
4685                 NaN       NaN    NaN     NaN     NaN    NaN
4686                 NaN       NaN    NaN     NaN     NaN    NaN
4687                 NaN       NaN    NaN     NaN     NaN    NaN
4688                 NaN       NaN    NaN     NaN     NaN    NaN
4689                 NaN       NaN    NaN     NaN     NaN    NaN
4690                 NaN       NaN    NaN     NaN     NaN    NaN
4691                 NaN       NaN    NaN     NaN     NaN    NaN
4692                 NaN       NaN    NaN     NaN     NaN    NaN
4693                 NaN       NaN    NaN     NaN     NaN    NaN
4694                 NaN       NaN    NaN     NaN     NaN    NaN
4695                 NaN       NaN    NaN     NaN     NaN    NaN
4696                 NaN       NaN    NaN     NaN     NaN    NaN
4697                 NaN       NaN    NaN     NaN     NaN    NaN
4698                 NaN       NaN    NaN     NaN     NaN    NaN
4699                 NaN       NaN    NaN     NaN     NaN    NaN
4700                 NaN       NaN    NaN     NaN     NaN    NaN
4701                 NaN       NaN    NaN     NaN     NaN    NaN
4702                 NaN       NaN    NaN     NaN     NaN    NaN
4703                 NaN       NaN    NaN     NaN     NaN    NaN
4704                 NaN       NaN    NaN     NaN     NaN    NaN
4705                 NaN       NaN    NaN     NaN     NaN    NaN
4706                 NaN       NaN    NaN     NaN     NaN    NaN
4707                 NaN       NaN    NaN     NaN     NaN    NaN
4708                 NaN       NaN    NaN     NaN     NaN    NaN
4709                 NaN       NaN    NaN     NaN     NaN    NaN

      favorite_count  retweet_count    tweet_id_retweet
0                NaN            NaN                 NaN
1                NaN            NaN                 NaN
2                NaN            NaN                 NaN
3                NaN            NaN                 NaN
4                NaN            NaN                 NaN
5                NaN            NaN                 NaN
6                NaN            NaN                 NaN
7                NaN            NaN                 NaN
8                NaN            NaN                 NaN
9                NaN            NaN                 NaN
10               NaN            NaN                 NaN
```

```
11               NaN            NaN                  NaN
12               NaN            NaN                  NaN
13               NaN            NaN                  NaN
14               NaN            NaN                  NaN
15               NaN            NaN                  NaN
16               NaN            NaN                  NaN
17               NaN            NaN                  NaN
18               NaN            NaN                  NaN
19               NaN            NaN                  NaN
20               NaN            NaN                  NaN
21               NaN            NaN                  NaN
22               NaN            NaN                  NaN
23               NaN            NaN                  NaN
24               NaN            NaN                  NaN
25               NaN            NaN                  NaN
26               NaN            NaN                  NaN
27               NaN            NaN                  NaN
28               NaN            NaN                  NaN
29               NaN            NaN                  NaN
...              ...            ...                  ...
4680           459.0          339.0   666411498068123649
4681           113.0           44.0   666407121513275392
4682           172.0           92.0   666396240351993856
4683           194.0          100.0   666373746337402880
4684           804.0          595.0   666362717482020864
4685           229.0           77.0   666353280906170368
4686           307.0          146.0   666345414279471104
4687           204.0           96.0   666337857791987715
4688           522.0          368.0   666293909010702337
4689           152.0           71.0   666287399580733440
4690           184.0           82.0   666273081518768128
4691           108.0           37.0   666268904428277760
4692         14765.0         6871.0   666104129232740352
4693            81.0           16.0   666102150364286977
4694           164.0           73.0   666099505364733952
4695           169.0           79.0   666093996847063040
4696           121.0           47.0   666082912819875840
4697           335.0          174.0   666073098362486784
4698           154.0           67.0   666071190449033216
4699           496.0          232.0   666063820255862784
4700           115.0           61.0   666058597072306176
4701           304.0          146.0   666057085227016192
4702           448.0          261.0   666055517517848576
4703          1253.0          879.0   666051848592334848
4704           136.0           60.0   666050754986266625
4705           111.0           41.0   666049244999131136
4706           311.0          147.0   666044217047650304
4707           128.0           47.0   666033409081393153
```

```
4708           132.0           48.0  666029276303482880
4709          2535.0          532.0  666020881337073664

[4710 rows x 15 columns]
```

Define

12 make a new column 'dogs' to show the types of dogs at once

Code

```
In [153]: #make a new column 'dogs' to show the types of dogs at once
          twitter_archive_clean['dogs'] = twitter_archive_clean['doggo']
          twitter_archive_clean['dogs'] = twitter_archive_clean['dogs'].fillna(twitter_archive_c
          twitter_archive_clean['dogs'] = twitter_archive_clean['dogs'].fillna(twitter_archive_c
          twitter_archive_clean['dogs'] = twitter_archive_clean['dogs'].fillna(twitter_archive_c
```

Test

```
In [154]: twitter_archive_clean['dogs'].value_counts()

Out[154]: pupper     245
          doggo       97
          puppo       29
          floofer      9
          Name: dogs, dtype: int64
```

12.0.1 Storing

```
In [155]: twitter_archive_master.to_csv('twitter_archive_master.csv', sep=',')
```

12.0.2 Insights

- According to the rating numerator column, most of the users know that there is no maximum rating system(usually 10 points) and rate more than 10. (visualized data with 'Visualization')

- As we refer the first data below, Yorkshire Terrier with more than one number photos had the highest confident algorithm in the #1 predicion. It means that is is adventageous to apply the algorithm with more than one number of photos with Yorkshire Terrier.

```
In [156]: image_predictions_clean.query('p1_dog == True').max()

Out[156]: tweet_id                                    892177421306343426
          jpg_url     https://pbs.twimg.com/tweet_video_thumb/CtTFZZ...
          img_num                                                      4
          p1                                       [Yorkshire, Terrier]
          p1_conf                                               0.999956
          p1_dog                                                    True
          p2                                       [Yorkshire, Terrier]
```

```
p2_conf                                             0.467678
p2_dog                                                  True
p3                                                   [Zebra]
p3_conf                                             0.273419
p3_dog                                                  True
dtype: object
```

- There was a significant relationship between favorite count and retweet counts, having the p-value ’0’. As the number of retweet increase one unit, the favorite also increases 1.57.

```
In [157]: tweets_df['intercept'] = 1
          lm = sm.OLS(tweets_df['favorite_count'], tweets_df[['intercept', 'retweet_count']])
          results = lm.fit()
          results.summary()
```

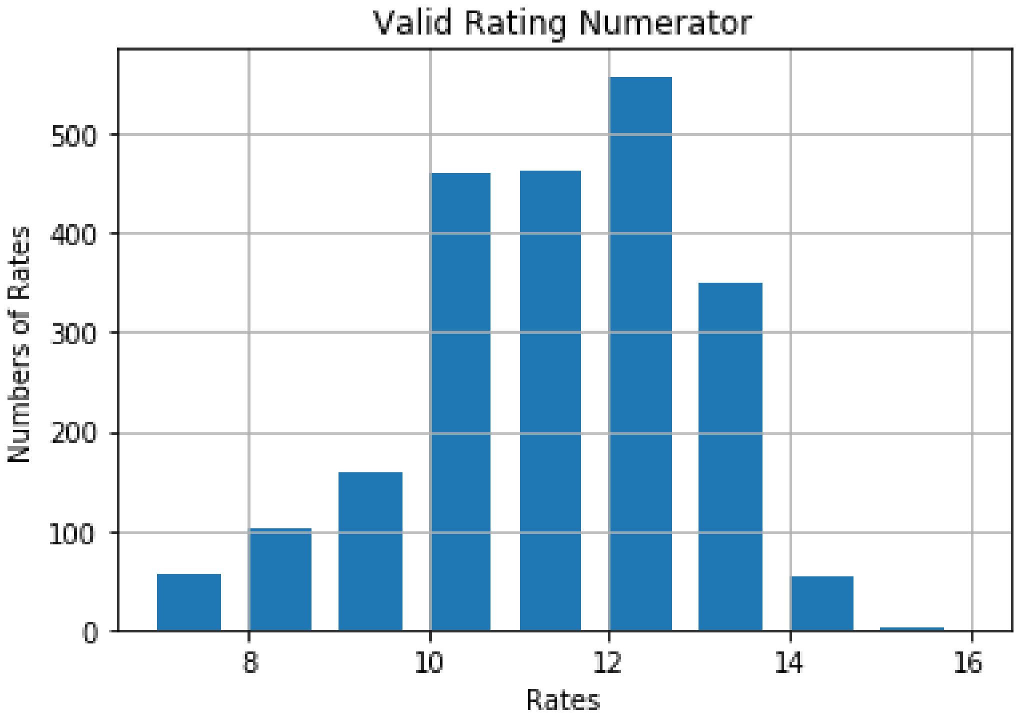
```
Out[157]: <class 'statsmodels.iolib.summary.Summary'>
          """
                                      OLS Regression Results
          ==============================================================================
          Dep. Variable:         favorite_count   R-squared:                       0.494
          Model:                            OLS   Adj. R-squared:                  0.494
          Method:                 Least Squares   F-statistic:                     2297.
          Date:                Mon, 18 May 2020   Prob (F-statistic):               0.00
          Time:                        14:22:05   Log-Likelihood:                -24611.
          No. Observations:                2354   AIC:                         4.923e+04
          Df Residuals:                    2352   BIC:                         4.924e+04
          Df Model:                           1
          Covariance Type:            nonrobust
          =================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
          ---------------------------------------------------------------------------------
          intercept      3107.8692    201.951     15.389      0.000    2711.849    3503.889
          retweet_count     1.5714      0.033     47.923      0.000       1.507       1.636
          ==============================================================================
          Omnibus:                     1034.735   Durbin-Watson:                   1.655
          Prob(Omnibus):                  0.000   Jarque-Bera (JB):            42336.254
          Skew:                          -1.368   Prob(JB):                         0.00
          Kurtosis:                      23.595   Cond. No.                     7.18e+03
          ==============================================================================

          Warnings:
          [1] Standard Errors assume that the covariance matrix of the errors is correctly speci
          [2] The condition number is large, 7.18e+03. This might indicate that there are
          strong multicollinearity or other numerical problems.
          """
```

12.0.3 Visualization

```
In [158]: twitter_archive_clean['rating_numerator'].value_counts()
```

```
Out[158]: 12.0    558
          11.0    464
          10.0    461
          13.0    351
          9.0     158
          8.0     102
          7.0      55
          14.0     54
          15.0      2
          Name: rating_numerator, dtype: int64
```

```
In [159]: twitter_archive_clean['rating_numerator'].hist(bins = np.arange(7,17,1), width=0.7)
          plt.title("Valid Rating Numerator")
          plt.xlabel("Rates")
          plt.ylabel("Numbers of Rates")
          plt.show()
```



- As most of the rates for the posts are positioned between 10 and 13, most of the users tend to give more than 10 points (when it is usually easy to think that the maximum rate would be out of 10).
- we can understand how generously the users rate the posts.