wrangle_act-Copy1

December 18, 2019

0.1 Ghather

```
In [49]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as sns
         %matplotlib inline
```

```
In [2]: #Downloaded from resources on Udacity
        archive = pd.read_csv('twitter-archive-enhanced.csv')
        image = pd.read_csv('image-predictions.tsv', sep='\t')
        tweet = pd.read_json('tweet-json.txt', lines=True)
```

0.2 Assess

```
In [3]: archive.head()
```

Out[3]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" r...
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" r...
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" r...
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="http://twitter.com/download/iphone" r...
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<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 [4]:
```
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 [5]: sum(archive.duplicated())
```

```
Out[5]: 0
```

```
In [6]: archive.rating_numerator.value_counts()
```

```
Out[6]: 12      558
        11      464
        10      461
        13      351
        9       158
        8       102
        7        55
        14       54
        5        37
        6        32
        3        19
        4        17
        1         9
        2         9
        420       2
        0         2
        15        2
        75        2
        80        1
        20        1
        24        1
        26        1
        44        1
        50        1
        60        1
        165       1
        84        1
        88        1
        144       1
        182       1
        143       1
        666       1
        960       1
        1776      1
        17        1
        27        1
        45        1
```

```
99        1
121       1
204       1
Name: rating_numerator, dtype: int64
```

```
In [7]: archive.rating_denominator.value_counts()
```

```
Out[7]: 10     2333
        11        3
        50        3
        80        2
        20        2
        2         1
        16        1
        40        1
        70        1
        15        1
        90        1
        110       1
        120       1
        130       1
        150       1
        170       1
        7         1
        0         1
        Name: rating_denominator, dtype: int64
```

```
In [8]: archive.name.value_counts()
```

```
Out[8]: None         745
        a             55
        Charlie       12
        Oliver        11
        Cooper        11
        Lucy          11
        Lola          10
        Penny         10
        Tucker        10
        Bo             9
        Winston        9
        Sadie          8
        the            8
        an             7
        Buddy          7
        Toby           7
        Bailey         7
        Daisy          7
        Jax            6
        Rusty          6
```

```
Koda            6
Scout           6
Milo            6
Dave            6
Oscar           6
Bella           6
Stanley         6
Jack            6
Leo             6
Bentley         5
               ...
Patch           1
Beemo           1
all             1
Mo              1
Yukon           1
Marlee          1
Gert            1
Cupid           1
Odin            1
Laika           1
Goose           1
Dewey           1
Cilantro        1
Tayzie          1
Remy            1
Smiley          1
Brudge          1
Major           1
Bonaparte       1
Crouton         1
Shelby          1
Arya            1
Jiminus         1
River           1
Divine          1
Torque          1
Rontu           1
Livvie          1
Mitch           1
Combo           1
Name: name, Length: 957, dtype: int64
```

```
In [9]: image.head()
```

```
Out[9]:              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 [10]:
```
image.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 [11]:
```
image.describe()
```

Out[11]:

	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
```

```
In [12]: sum(image.duplicated())
```

```
Out[12]: 0
```

```
In [13]: tweet.head()
```

```
Out[13]:    contributors  coordinates          created_at display_text_range  \
          0           NaN          NaN 2017-08-01 16:23:56            [0, 85]
          1           NaN          NaN 2017-08-01 00:17:27           [0, 138]
          2           NaN          NaN 2017-07-31 00:18:03           [0, 121]
          3           NaN          NaN 2017-07-30 15:58:51            [0, 79]
          4           NaN          NaN 2017-07-29 16:00:24           [0, 138]

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

                                          extended_entities  favorite_count  \
          0  {'media': [{'id': 892420639486877696, 'id_str'...           39467
          1  {'media': [{'id': 892177413194625024, 'id_str'...           33819
          2  {'media': [{'id': 891815175371796480, 'id_str'...           25461
          3  {'media': [{'id': 891689552724799489, 'id_str'...           42908
          4  {'media': [{'id': 891327551943041024, 'id_str'...           41048

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

                                 ...                                 \
          0                      ...
          1                      ...
          2                      ...
          3                      ...
          4                      ...

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

```
4                           0.0            NaN              NaN

   quoted_status_id_str  retweet_count  retweeted  retweeted_status  \
0                   NaN           8853      False               NaN
1                   NaN           6514      False               NaN
2                   NaN           4328      False               NaN
3                   NaN           8964      False               NaN
4                   NaN           9774      False               NaN

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

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

[5 rows x 31 columns]
```

```
In [14]: tweet.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 31 columns):
contributors                    0 non-null float64
coordinates                     0 non-null float64
created_at                      2354 non-null datetime64[ns]
display_text_range              2354 non-null object
entities                        2354 non-null object
extended_entities               2073 non-null object
favorite_count                  2354 non-null int64
favorited                       2354 non-null bool
full_text                       2354 non-null object
geo                             0 non-null float64
id                              2354 non-null int64
id_str                          2354 non-null int64
in_reply_to_screen_name         78 non-null object
in_reply_to_status_id           78 non-null float64
in_reply_to_status_id_str       78 non-null float64
in_reply_to_user_id             78 non-null float64
in_reply_to_user_id_str         78 non-null float64
is_quote_status                 2354 non-null bool
```

```
lang                            2354 non-null object
place                           1 non-null object
possibly_sensitive              2211 non-null float64
possibly_sensitive_appealable   2211 non-null float64
quoted_status                   28 non-null object
quoted_status_id                29 non-null float64
quoted_status_id_str            29 non-null float64
retweet_count                   2354 non-null int64
retweeted                       2354 non-null bool
retweeted_status                179 non-null object
source                          2354 non-null object
truncated                       2354 non-null bool
user                            2354 non-null object
dtypes: bool(4), datetime64[ns](1), float64(11), int64(4), object(11)
memory usage: 505.8+ KB
```

```
In [15]: tweet.describe()
```

Out[15]:

	contributors	coordinates	favorite_count	geo	id
count	0.0	0.0	2354.000000	0.0	2.354000e+03
mean	NaN	NaN	8080.968564	NaN	7.426978e+17
std	NaN	NaN	11814.771334	NaN	6.852812e+16
min	NaN	NaN	0.000000	NaN	6.660209e+17
25%	NaN	NaN	1415.000000	NaN	6.783975e+17
50%	NaN	NaN	3603.500000	NaN	7.194596e+17
75%	NaN	NaN	10122.250000	NaN	7.993058e+17
max	NaN	NaN	132810.000000	NaN	8.924206e+17

	id_str	in_reply_to_status_id	in_reply_to_status_id_str
count	2.354000e+03	7.800000e+01	7.800000e+01
mean	7.426978e+17	7.455079e+17	7.455079e+17
std	6.852812e+16	7.582492e+16	7.582492e+16
min	6.660209e+17	6.658147e+17	6.658147e+17
25%	6.783975e+17	6.757419e+17	6.757419e+17
50%	7.194596e+17	7.038708e+17	7.038708e+17
75%	7.993058e+17	8.257804e+17	8.257804e+17
max	8.924206e+17	8.862664e+17	8.862664e+17

	in_reply_to_user_id	in_reply_to_user_id_str	possibly_sensitive
count	7.800000e+01	7.800000e+01	2211.0
mean	2.014171e+16	2.014171e+16	0.0
std	1.252797e+17	1.252797e+17	0.0
min	1.185634e+07	1.185634e+07	0.0
25%	3.086374e+08	3.086374e+08	0.0
50%	4.196984e+09	4.196984e+09	0.0
75%	4.196984e+09	4.196984e+09	0.0
max	8.405479e+17	8.405479e+17	0.0

```
       possibly_sensitive_appealable  quoted_status_id  quoted_status_id_str  \
count                         2211.0      2.900000e+01          2.900000e+01
mean                             0.0      8.162686e+17          8.162686e+17
std                              0.0      6.164161e+16          6.164161e+16
min                              0.0      6.721083e+17          6.721083e+17
25%                              0.0      7.888183e+17          7.888183e+17
50%                              0.0      8.340867e+17          8.340867e+17
75%                              0.0      8.664587e+17          8.664587e+17
max                              0.0      8.860534e+17          8.860534e+17

       retweet_count
count    2354.000000
mean     3164.797366
std      5284.770364
min         0.000000
25%       624.500000
50%      1473.500000
75%      3652.000000
max     79515.000000
```

```
In [16]: tweet.lang.value_counts()
```

```
Out[16]: en     2336
         und       7
         nl        3
         in        3
         es        1
         tl        1
         ro        1
         et        1
         eu        1
         Name: lang, dtype: int64
```

1 Quality

1.1 Archive Table

1) Below columns have many null records

i) in_reply_to_status_id
ii) in_reply_to_user_id
iii) retweeted_status_id iV) retweeted_status_id

V) retweeted_status_user_id

VI) retweeted_status_timestamp
VII) expanded_urls

2) Timestamp is object should be date

3) Timestamp has +0000 in the end should be removed to be converted

4) The values in archive dominator and nominator should be from 0 to 10 but some values are way out of range should be capped at 15 for example

1.2 Tweet Table

1) Many columns have null values

2) Nearly all of the values in the lang column is the same 'en' so it doesn't add that much value and will be dropped

Some Columns doesn't add any valueable information in the 3 tables so they should be dropped.

2 Tidniess

1) id column in tweet table should be renamed to tweet_id to match the other tables as they are all refrence to the same thing.

2) The 3 tables should be merged in one table

3) The last 4 columns in archive table should be merged in one column and assesed whether to be dropped or not if most of the values are None

3 Clean

```
In [17]: archive_clean = archive.copy()
         image_clean = image.copy()
         tweet_clean = tweet.copy()
```

3.0.1 Missing Data

Many null values in different columns in archive table

3.0.2 Define

Drop these columns as they don't have to much significant

3.0.3 Code

```
In [18]: archive_clean = archive_clean.drop(['in_reply_to_status_id', 'in_reply_to_user_id', 're
                                'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_
```

3.0.4 Test

```
In [19]: archive_clean.head()
```

```
Out[19]:              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
         3  891689557279858688  2017-07-30 15:58:51 +0000
         4  891327558926688256  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  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
         2  This is Archie. He is a rare Norwegian Pouncin...                12
         3  This is Darla. She commenced a snooze mid meal...                13
         4  This is Franklin. He would like you to stop ca...                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
```

3.0.5 Timestamp cleaning

3.0.6 Define

removing +0000 from the timestamp column

3.0.7 Code

```
In [20]: archive_clean.timestamp = archive_clean.timestamp.str.strip('+0000')
```

3.0.8 Test

```
In [21]: archive_clean.timestamp.head()
```

```
Out[21]: 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: object
```

3.0.9 Define

Timestamp is an object not date

3.0.10 Code

```
In [22]: archive_clean.timestamp = pd.to_datetime(archive_clean.timestamp)
```

3.0.11 Test

```
In [23]: archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 11 columns):
tweet_id              2356 non-null int64
timestamp             2356 non-null datetime64[ns]
source                2356 non-null object
text                  2356 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: datetime64[ns](1), int64(3), object(7)
memory usage: 202.5+ KB
```

3.0.12 Values out of range

Rating values should be from 0 to 10 so will maximize to 15 giving +5 margin

3.0.13 Define

Replacing any values more than 15 with 15

3.0.14 Code

```
In [24]: archive_clean.rating_numerator.head()
```

```
Out[24]: 0    13
         1    13
         2    12
         3    13
```

```
4    12
Name: rating_numerator, dtype: int64
```

```
In [25]: for i in archive_clean.rating_numerator:
             if i > 15:
                 archive_clean.rating_numerator.replace (i, 15, inplace=True)
```

```
In [26]: for i in archive_clean.rating_denominator:
             if i > 15:
                 archive_clean.rating_denominator.replace (i, 15, inplace=True)
```

3.0.15 Test

```
In [27]: archive_clean.rating_numerator.value_counts()
```

```
Out[27]: 12    558
         11    464
         10    461
         13    351
         9     158
         8     102
         7      55
         14     54
         5      37
         6      32
         15     28
         3      19
         4      17
         1       9
         2       9
         0       2
         Name: rating_numerator, dtype: int64
```

```
In [28]: archive_clean.rating_denominator.value_counts()
```

```
Out[28]: 10    2333
         15      17
         11       3
         7        1
         2        1
         0        1
         Name: rating_denominator, dtype: int64
```

3.0.16 Missing data

Many null values in different columns in tweet table

3.0.17 Define

Drop the null columns as it doesn't add much value

3.0.18 Code

```
In [29]: tweet_clean = tweet_clean.drop(['contributors', 'coordinates', 'extended_entities', 'ge
                        'in_reply_to_status_id', 'in_reply_to_status_id_str', 'in_reply_to_use
                        'place', 'possibly_sensitive', 'possibly_sensitive_appealable', 'quote
                       'quoted_status_id_str', 'retweeted_status'], axis=1)
```

3.0.19 Test

```
In [30]: tweet_clean.head()
```

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

                                                 entities  favorite_count  \
         0  {'hashtags': [], 'symbols': [], 'user_mentions...           39467
         1  {'hashtags': [], 'symbols': [], 'user_mentions...           33819
         2  {'hashtags': [], 'symbols': [], 'user_mentions...           25461
         3  {'hashtags': [], 'symbols': [], 'user_mentions...           42908
         4  {'hashtags': [{'text': 'BarkWeek', 'indices': ...           41048

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

                           id              id_str  is_quote_status lang  \
         0  892420643555336193  892420643555336192            False   en
         1  892177421306343426  892177421306343424            False   en
         2  891815181378084864  891815181378084864            False   en
         3  891689557279858688  891689557279858688            False   en
         4  891327558926688256  891327558926688256            False   en

            retweet_count  retweeted  \
         0           8853      False
         1           6514      False
         2           4328      False
         3           8964      False
         4           9774      False

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

```
2  <a href="http://twitter.com/download/iphone" r...      False
3  <a href="http://twitter.com/download/iphone" r...      False
4  <a href="http://twitter.com/download/iphone" r...      False

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

```
In [31]: tweet_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 15 columns):
created_at            2354 non-null datetime64[ns]
display_text_range    2354 non-null object
entities              2354 non-null object
favorite_count        2354 non-null int64
favorited             2354 non-null bool
full_text             2354 non-null object
id                    2354 non-null int64
id_str                2354 non-null int64
is_quote_status       2354 non-null bool
lang                  2354 non-null object
retweet_count         2354 non-null int64
retweeted             2354 non-null bool
source                2354 non-null object
truncated             2354 non-null bool
user                  2354 non-null object
dtypes: bool(4), datetime64[ns](1), int64(4), object(6)
memory usage: 211.6+ KB
```

3.0.20 Dropping Unwanted Columns

Some columns doesn't have valuable information that affects our insights

3.0.21 Define

Dropping the unwanted columns in each table

3.0.22 Code

```
In [32]: archive_clean = archive_clean.drop(['source', 'text'], axis=1)
```

```
In [33]: image_clean = image_clean.drop(['img_num'], axis=1)
```

```
In [34]: tweet_clean = tweet_clean.drop(['created_at', 'display_text_range', 'entities', 'full_t
                        'favorited', 'is_quote_status','retweeted', 'lang' ], axis = 1)
```

3.0.23 Merging Columns

4 columns can be indicated in two instead to be more tidy data

3.0.24 Define

Combinig doggo, floofer, pupper & puppo columns

```
In [35]: melt = pd.melt(archive_clean, id_vars=['tweet_id', 'timestamp', 'rating_numerator', 'na
                                                'rating_denominator'], var_name='dog')
```

3.0.25 Test

```
In [36]: melt.head(10)
         #melt.query('value == "None"')
```

```
Out[36]:              tweet_id           timestamp  rating_numerator      name  \
         0  892420643555336193 2017-08-01 16:23:56                13   Phineas
         1  892177421306343426 2017-08-01 00:17:27                13     Tilly
         2  891815181378084864 2017-07-31 00:18:03                12    Archie
         3  891689557279858688 2017-07-30 15:58:51                13     Darla
         4  891327558926688256 2017-07-29 16:00:24                12  Franklin
         5  891087950875897856 2017-07-29 00:08:17                13      None
         6  890971913173991426 2017-07-28 16:27:12                13       Jax
         7  890729181411237888 2017-07-28 00:22:40                13      None
         8  890609185150312448 2017-07-27 16:25:51                13      Zoey
         9  890240255349198849 2017-07-26 15:59:51                14    Cassie

            rating_denominator    dog  value
         0                  10  doggo   None
         1                  10  doggo   None
         2                  10  doggo   None
         3                  10  doggo   None
         4                  10  doggo   None
         5                  10  doggo   None
         6                  10  doggo   None
         7                  10  doggo   None
         8                  10  doggo   None
         9                  10  doggo  doggo
```

3.1 Re-Assess

```
In [37]: melt.query('value == "None"').count()
```

```
Out[37]: tweet_id              9030
         timestamp             9030
         rating_numerator      9030
         name                  9030
         rating_denominator    9030
```

```
dog                   9030
value                 9030
dtype: int64
```

```
In [38]: melt.query('value != "None"').count()
```

```
Out[38]: tweet_id              394
         timestamp             394
         rating_numerator      394
         name                  394
         rating_denominator    394
         dog                   394
         value                 394
         dtype: int64
```

3.1.1 Will drop these 4 columns from the dataset as it doesn't have that much effect

3.1.2 Code

```
In [39]: archive_clean = archive_clean.drop(['doggo', 'floofer', 'pupper', 'puppo'], axis=1)
```

3.1.3 Test

```
In [40]: archive_clean.head()
```

```
Out[40]:              tweet_id           timestamp  rating_numerator  \
         0  892420643555336193 2017-08-01 16:23:56                13
         1  892177421306343426 2017-08-01 00:17:27                13
         2  891815181378084864 2017-07-31 00:18:03                12
         3  891689557279858688 2017-07-30 15:58:51                13
         4  891327558926688256 2017-07-29 16:00:24                12

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

3.1.4 Renaming Column

id column in tweet table should be tweet_id to match other tables

3.1.5 Define

Renaming id column in tweet table

3.1.6 Code

```
In [41]: tweet_clean = tweet_clean.rename(columns={"id": "tweet_id"})
```

3.1.7 Test

```
In [42]: tweet_clean.head()
```

```
Out[42]:    favorite_count            tweet_id  retweet_count
         0           39467  892420643555336193           8853
         1           33819  892177421306343426           6514
         2           25461  891815181378084864           4328
         3           42908  891689557279858688           8964
         4           41048  891327558926688256           9774
```

3.1.8 Making One Dataset

The 3 datasets should be merged in 1 dataset only

3.1.9 Define

Merging the 3 tables

3.1.10 Code

```
In [43]: df1 = archive_clean.merge(image_clean, on='tweet_id')

In [44]: df = df1.merge(tweet_clean, on='tweet_id')
```

3.1.11 Testing

```
In [45]: df.head()
```

```
Out[45]:              tweet_id           timestamp  rating_numerator  \
         0  892420643555336193 2017-08-01 16:23:56                13
         1  892177421306343426 2017-08-01 00:17:27                13
         2  891815181378084864 2017-07-31 00:18:03                12
         3  891689557279858688 2017-07-30 15:58:51                13
         4  891327558926688256 2017-07-29 16:00:24                12

            rating_denominator      name  \
         0                  10   Phineas
         1                  10     Tilly
         2                  10    Archie
         3                  10     Darla
         4                  10  Franklin

                                                jpg_url           p1   p1_conf  \
         0  https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg       orange  0.097049
         1  https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg    Chihuahua  0.323581
         2  https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg    Chihuahua  0.716012
         3  https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg  paper_towel  0.170278
         4  https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg       basset  0.555712
```

```
   p1_dog                  p2   p2_conf  p2_dog                            p3  \
0   False              bagel  0.085851   False                        banana
1    True           Pekinese  0.090647    True                      papillon
2    True           malamute  0.078253    True                        kelpie
3   False  Labrador_retriever  0.168086    True                       spatula
4    True    English_springer  0.225770    True  German_short-haired_pointer

    p3_conf  p3_dog  favorite_count  retweet_count
0  0.076110   False           39467           8853
1  0.068957    True           33819           6514
2  0.031379    True           25461           4328
3  0.040836   False           42908           8964
4  0.175219    True           41048           9774
```

In [46]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2073 entries, 0 to 2072
Data columns (total 17 columns):
tweet_id              2073 non-null int64
timestamp             2073 non-null datetime64[ns]
rating_numerator      2073 non-null int64
rating_denominator    2073 non-null int64
name                  2073 non-null object
jpg_url               2073 non-null object
p1                    2073 non-null object
p1_conf               2073 non-null float64
p1_dog                2073 non-null bool
p2                    2073 non-null object
p2_conf               2073 non-null float64
p2_dog                2073 non-null bool
p3                    2073 non-null object
p3_conf               2073 non-null float64
p3_dog                2073 non-null bool
favorite_count        2073 non-null int64
retweet_count         2073 non-null int64
dtypes: bool(3), datetime64[ns](1), float64(3), int64(5), object(5)
memory usage: 249.0+ KB
```

3.1.12 Saving The final cleaned data frame to .csv file

In [47]:
```
df.to_csv('tweeter_final.csv')
```