1 Project: Wrangling and Analyze Data

1.1 Data Gathering

In the cell below, gather **all** three pieces of data for this project and load them in the notebook. **Note:** the methods required to gather each data are different.

1.1.1 1. Directly download the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv)

```
----------------------------------------
-------------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
None
----------------------------------------
-------------NUNIQUE---------------------
----------------------------------------
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
----------------------------------------
--------------DESCRIBE-----------------
----------------------------------------

           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


----------------------------------------
--------------HEAD--------------------
----------------------------------------

             tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
0  892420643555336193                    NaN                  NaN
```

```
1  892177421306343426                   NaN                 NaN
2  891815181378084864                   NaN                 NaN
3  891689557279858688                   NaN                 NaN
4  891327558926688256                   NaN                 NaN

                  timestamp  \
0  2017-08-01 16:23:56 +0000
1  2017-08-01 00:17:27 +0000
2  2017-07-31 00:18:03 +0000
3  2017-07-30 15:58:51 +0000
4  2017-07-29 16:00:24 +0000

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

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

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

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

   rating_denominator      name doggo floofer pupper puppo
0                  10   Phineas  None    None   None  None
1                  10     Tilly  None    None   None  None
2                  10    Archie  None    None   None  None
3                  10     Darla  None    None   None  None
4                  10  Franklin  None    None   None  None
```

```
----------------------------------------
----------------------------------------
```

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

```
----------------------------------------
-------------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
None
----------------------------------------
-------------NUNIQUE---------------------
----------------------------------------
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
----------------------------------------
-------------DESCRIBE-----------------
----------------------------------------

           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
```

```
----------------------------------------
--------------HEAD----------------------
----------------------------------------
```

```
             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
```

```
----------------------------------------
----------------------------------------
```

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

Not able to get Tweeter API access

```
----------------------------------------
--------------INFO----------------------
----------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 31 columns):
```

```
created_at                          2354 non-null datetime64[ns, UTC]
id                                  2354 non-null int64
id_str                              2354 non-null int64
full_text                           2354 non-null object
truncated                           2354 non-null bool
display_text_range                  2354 non-null object
entities                            2354 non-null object
extended_entities                   2073 non-null object
source                              2354 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
in_reply_to_screen_name             78 non-null object
user                                2354 non-null object
geo                                 0 non-null float64
coordinates                         0 non-null float64
place                               1 non-null object
contributors                        0 non-null float64
is_quote_status                     2354 non-null bool
retweet_count                       2354 non-null int64
favorite_count                      2354 non-null int64
favorited                           2354 non-null bool
retweeted                           2354 non-null bool
possibly_sensitive                  2211 non-null float64
possibly_sensitive_appealable       2211 non-null float64
lang                                2354 non-null object
retweeted_status                    179 non-null object
quoted_status_id                    29 non-null float64
quoted_status_id_str                29 non-null float64
quoted_status                       28 non-null object
dtypes: bool(4), datetime64[ns, UTC](1), float64(11), int64(4), object(11)
memory usage: 505.9+ KB
None
----------------------------------------
--------------NUNIQUE---------------------
----------------------------------------
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
some error at pandas.nunique() occurred
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
----------------------------------------
--------------DESCRIBE-----------------
----------------------------------------

                 id        id_str  in_reply_to_status_id  \
count  2.354000e+03  2.354000e+03           7.800000e+01
mean   7.426978e+17  7.426978e+17           7.455079e+17
std    6.852812e+16  6.852812e+16           7.582492e+16
```

```
min    6.660209e+17  6.660209e+17           6.658147e+17
25%    6.783975e+17  6.783975e+17           6.757419e+17
50%    7.194596e+17  7.194596e+17           7.038708e+17
75%    7.993058e+17  7.993058e+17           8.257804e+17
max    8.924206e+17  8.924206e+17           8.862664e+17

       in_reply_to_status_id_str  in_reply_to_user_id  \
count               7.800000e+01         7.800000e+01
mean                7.455079e+17         2.014171e+16
std                 7.582492e+16         1.252797e+17
min                 6.658147e+17         1.185634e+07
25%                 6.757419e+17         3.086374e+08
50%                 7.038708e+17         4.196984e+09
75%                 8.257804e+17         4.196984e+09
max                 8.862664e+17         8.405479e+17

       in_reply_to_user_id_str  geo  coordinates  contributors  retweet_count  \
count             7.800000e+01  0.0          0.0           0.0    2354.000000
mean              2.014171e+16  NaN          NaN           NaN    3164.797366
std               1.252797e+17  NaN          NaN           NaN    5284.770364
min               1.185634e+07  NaN          NaN           NaN       0.000000
25%               3.086374e+08  NaN          NaN           NaN     624.500000
50%               4.196984e+09  NaN          NaN           NaN    1473.500000
75%               4.196984e+09  NaN          NaN           NaN    3652.000000
max               8.405479e+17  NaN          NaN           NaN   79515.000000

       favorite_count  possibly_sensitive  possibly_sensitive_appealable  \
count     2354.000000              2211.0                         2211.0
mean      8080.968564                 0.0                            0.0
std      11814.771334                 0.0                            0.0
min          0.000000                 0.0                            0.0
25%       1415.000000                 0.0                            0.0
50%       3603.500000                 0.0                            0.0
75%      10122.250000                 0.0                            0.0
max     132810.000000                 0.0                            0.0

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

----------------------------------------
```

```
--------------HEAD----------------------
----------------------------------------

                 created_at                   id              id_str  \
0 2017-08-01 16:23:56+00:00  892420643555336193  892420643555336192
1 2017-08-01 00:17:27+00:00  892177421306343426  892177421306343424
2 2017-07-31 00:18:03+00:00  891815181378084864  891815181378084864
3 2017-07-30 15:58:51+00:00  891689557279858688  891689557279858688
4 2017-07-29 16:00:24+00:00  891327558926688256  891327558926688256

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

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

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

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

   ...  favorite_count  favorited  retweeted possibly_sensitive  \
0  ...           39467      False      False                0.0
1  ...           33819      False      False                0.0
2  ...           25461      False      False                0.0
3  ...           42908      False      False                0.0
4  ...           41048      False      False                0.0

  possibly_sensitive_appealable  lang  retweeted_status quoted_status_id  \
0                           0.0    en               NaN              NaN
1                           0.0    en               NaN              NaN
```

```
2                           0.0    en               NaN              NaN
3                           0.0    en               NaN              NaN
4                           0.0    en               NaN              NaN

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

[5 rows x 31 columns]


----------------------------------------
----------------------------------------
```

1.2 Assessing Data

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

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

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

1.2.1 Quality issues

1. Wrong Data types: (df1)

Column	type
in_reply_to_status_id	float64
in_reply_to_user_id	float64
timestamp	object
retweeted_status_id	float64
retweeted_status_user_id	float64
retweeted_status_timestamp	object

2. Rows with "Stage of Dogs" with more than one classification

3. Dog names column as 'None' (string) istead of Nan (null object) for missing Data
4. Retweets rows
5. In reply rows
6. Error getting the rate numbers like 5 instead 13.5, 75 instead 9.75, etc
7. Tweets with "This is a —" geting Dog name as "a"
8. Not necessary Columns (df1) - source - in_reply_to_status_id - in_reply_to_user_id - retweeted_status_id
   - retweeted_status_user_id
   - retweeted_status_timestamp
9. Remove unnecessary columns 'doggo', 'floofer', 'pupper', 'puppo', 'stg_count'
10. Missing values handling

1.2.2 Tidiness issues

1. Text Column with text and pictures URL (df1)
2. Stage of dogs in colums (df1)

1.3 Cleaning Data

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

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

```
<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
<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
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 31 columns):
created_at                      2354 non-null datetime64[ns, UTC]
id                              2354 non-null int64
id_str                          2354 non-null int64
full_text                       2354 non-null object
truncated                       2354 non-null bool
display_text_range              2354 non-null object
entities                        2354 non-null object
extended_entities               2073 non-null object
source                          2354 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
in_reply_to_screen_name         78 non-null object
user                            2354 non-null object
geo                             0 non-null float64
coordinates                     0 non-null float64
place                           1 non-null object
contributors                    0 non-null float64
is_quote_status                 2354 non-null bool
retweet_count                   2354 non-null int64
favorite_count                  2354 non-null int64
favorited                       2354 non-null bool
retweeted                       2354 non-null bool
possibly_sensitive              2211 non-null float64
possibly_sensitive_appealable   2211 non-null float64
```

```
lang                                2354 non-null object
retweeted_status                    179 non-null object
quoted_status_id                    29 non-null float64
quoted_status_id_str                29 non-null float64
quoted_status                       28 non-null object
dtypes: bool(4), datetime64[ns, UTC](1), float64(11), int64(4), object(11)
memory usage: 505.9+ KB
```

1.3.1 General Issue

[12]:
```
2354
```

** All df3 is same in df1 so leave it out.

1.3.2 Merge the two dataframes df1 and df2 into one and clean

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 28 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
jpg_url                       2075 non-null object
img_num                       2075 non-null float64
p1                            2075 non-null object
p1_conf                       2075 non-null float64
p1_dog                        2075 non-null object
p2                            2075 non-null object
p2_conf                       2075 non-null float64
p2_dog                        2075 non-null object
p3                            2075 non-null object
p3_conf                       2075 non-null float64
p3_dog                        2075 non-null object
dtypes: float64(8), int64(3), object(17)
```

```
memory usage: 533.8+ KB
```

1.3.3 Issue #1:

Wrong Data types:

Column	type
in_reply_to_status_id	float64
in_reply_to_user_id	float64
timestamp	object
retweeted_status_id	float64
retweeted_status_user_id	float64
retweeted_status_timestamp	object

Define: Change Column type:

- in_reply_to_status_id to int
- in_reply_to_user_id to int
- timestamp to datatime
- retweeted_status_id to int
- retweeted_status_user_id to int
- retweeted_status_timestamp to datatime

Code

Test

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 28 columns):
tweet_id                      2356 non-null int64
in_reply_to_status_id         78 non-null Int64
in_reply_to_user_id           78 non-null Int64
timestamp                     2356 non-null datetime64[ns]
source                        2356 non-null object
text                          2356 non-null object
retweeted_status_id           181 non-null Int64
retweeted_status_user_id      181 non-null Int64
retweeted_status_timestamp    181 non-null datetime64[ns]
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
jpg_url                       2075 non-null object
```

```
img_num                     2075 non-null float64
p1                          2075 non-null object
p1_conf                     2075 non-null float64
p1_dog                      2075 non-null object
p2                          2075 non-null object
p2_conf                     2075 non-null float64
p2_dog                      2075 non-null object
p3                          2075 non-null object
p3_conf                     2075 non-null float64
p3_dog                      2075 non-null object
dtypes: Int64(4), datetime64[ns](2), float64(4), int64(3), object(15)
memory usage: 543.0+ KB
```

1.3.4 Issue #2:

Rows with "Stage of Dogs" with more than one classification

Define Select rows with more than 1 dog stage classification, if more than 1 delete entry.

Code

[18]:
```
14
```

[20]:
```
0
```

Test

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2342 entries, 0 to 2341
Data columns (total 30 columns):
index                       2342 non-null int64
tweet_id                    2342 non-null int64
in_reply_to_status_id       77 non-null Int64
in_reply_to_user_id         77 non-null Int64
timestamp                   2342 non-null datetime64[ns]
source                      2342 non-null object
text                        2342 non-null object
retweeted_status_id         179 non-null Int64
retweeted_status_user_id    179 non-null Int64
retweeted_status_timestamp  179 non-null datetime64[ns]
expanded_urls               2283 non-null object
rating_numerator            2342 non-null int64
rating_denominator          2342 non-null int64
name                        2342 non-null object
doggo                       83 non-null object
floofer                     9 non-null object
pupper                      245 non-null object
puppo                       29 non-null object
jpg_url                     2062 non-null object
```

```
img_num                       2062 non-null float64
p1                            2062 non-null object
p1_conf                       2062 non-null float64
p1_dog                        2062 non-null object
p2                            2062 non-null object
p2_conf                       2062 non-null float64
p2_dog                        2062 non-null object
p3                            2062 non-null object
p3_conf                       2062 non-null float64
p3_dog                        2062 non-null object
stg_count                     2342 non-null int64
dtypes: Int64(4), datetime64[ns](2), float64(4), int64(5), object(15)
memory usage: 558.2+ KB
None
```

1.3.5 Issue #3:

Missing Data at Dog names column as 'None' (string) istead of None (null object)

Define Select rows with more than 1 dog stage classification, if more than 1 delete entry.

Test

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2342 entries, 0 to 2341
Data columns (total 30 columns):
index                         2342 non-null int64
tweet_id                      2342 non-null int64
in_reply_to_status_id         77 non-null Int64
in_reply_to_user_id           77 non-null Int64
timestamp                     2342 non-null datetime64[ns]
source                        2342 non-null object
text                          2342 non-null object
retweeted_status_id           179 non-null Int64
retweeted_status_user_id      179 non-null Int64
retweeted_status_timestamp    179 non-null datetime64[ns]
expanded_urls                 2283 non-null object
rating_numerator              2342 non-null int64
rating_denominator            2342 non-null int64
name                          1605 non-null object
doggo                         83 non-null object
floofer                       9 non-null object
pupper                        245 non-null object
puppo                         29 non-null object
jpg_url                       2062 non-null object
img_num                       2062 non-null float64
p1                            2062 non-null object
p1_conf                       2062 non-null float64
```

```
p1_dog                          2062 non-null object
p2                              2062 non-null object
p2_conf                         2062 non-null float64
p2_dog                          2062 non-null object
p3                              2062 non-null object
p3_conf                         2062 non-null float64
p3_dog                          2062 non-null object
stg_count                       2342 non-null int64
dtypes: Int64(4), datetime64[ns](2), float64(4), int64(5), object(15)
memory usage: 558.2+ KB
None
```

1.3.6 Issue #4:

Retwites rows

Define Mask not empty "retweeted_status_id" column and Drop Rows

Code

Test

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2163 entries, 0 to 2162
Data columns (total 30 columns):
index                           2163 non-null int64
tweet_id                        2163 non-null int64
in_reply_to_status_id           77 non-null Int64
in_reply_to_user_id             77 non-null Int64
timestamp                       2163 non-null datetime64[ns]
source                          2163 non-null object
text                            2163 non-null object
retweeted_status_id             0 non-null Int64
retweeted_status_user_id        0 non-null Int64
retweeted_status_timestamp      0 non-null datetime64[ns]
expanded_urls                   2105 non-null object
rating_numerator                2163 non-null int64
rating_denominator              2163 non-null int64
name                            1490 non-null object
doggo                           75 non-null object
floofer                         9 non-null object
pupper                          224 non-null object
puppo                           24 non-null object
jpg_url                         1983 non-null object
img_num                         1983 non-null float64
p1                              1983 non-null object
p1_conf                         1983 non-null float64
p1_dog                          1983 non-null object
```

```
p2                          1983 non-null object
p2_conf                     1983 non-null float64
p2_dog                      1983 non-null object
p3                          1983 non-null object
p3_conf                     1983 non-null float64
p3_dog                      1983 non-null object
stg_count                   2163 non-null int64
dtypes: Int64(4), datetime64[ns](2), float64(4), int64(5), object(15)
memory usage: 515.5+ KB
```

1.3.7 Issue #5:

Mask not empty "In reply rows" column and Drop Rows

Define Mask not empty "rin_reply_to_status_id" column and Drop Rows

Code

Test

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2086 entries, 0 to 2085
Data columns (total 30 columns):
index                       2086 non-null int64
tweet_id                    2086 non-null int64
in_reply_to_status_id       0 non-null Int64
in_reply_to_user_id         0 non-null Int64
timestamp                   2086 non-null datetime64[ns]
source                      2086 non-null object
text                        2086 non-null object
retweeted_status_id         0 non-null Int64
retweeted_status_user_id    0 non-null Int64
retweeted_status_timestamp  0 non-null datetime64[ns]
expanded_urls               2083 non-null object
rating_numerator            2086 non-null int64
rating_denominator          2086 non-null int64
name                        1489 non-null object
doggo                       72 non-null object
floofer                     9 non-null object
pupper                      221 non-null object
puppo                       23 non-null object
jpg_url                     1961 non-null object
img_num                     1961 non-null float64
p1                          1961 non-null object
p1_conf                     1961 non-null float64
p1_dog                      1961 non-null object
p2                          1961 non-null object
p2_conf                     1961 non-null float64
```

```
p2_dog                        1961 non-null object
p3                            1961 non-null object
p3_conf                       1961 non-null float64
p3_dog                        1961 non-null object
stg_count                     2086 non-null int64
dtypes: Int64(4), datetime64[ns](2), float64(4), int64(5), object(15)
memory usage: 497.2+ KB
```

1.3.8 Issue #6:

6. Error geting the numbers like 5 intead 13.5

Define Select rows and round values.

13.5 –> 13

9.75 –> 10

11.27–>11

11.26–>11

Code

```
                                                                    text  \
313  This is Scooter. His lack of opposable thumbs is rendering his resistance
to tickling embarrassingly moot. 12/10 would keep tickling https://t.co/…

     rating_numerator
313                12
```

[31]:

```
       index            tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
41        45  883482846933004288                    NaN                  NaN
523      695  786709082849828864                    NaN                  NaN
579      763  778027034220126208                    NaN                  NaN
1463    1712  680494726643068929                    NaN                  NaN

                timestamp  \
41   2017-07-08 00:28:19
523  2016-10-13 23:23:56
579  2016-09-20 00:24:34
1463 2015-12-25 21:06:00

  source  \
41    <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
523   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
579   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
```

```
1463  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>

                                                                      text  \
41                 This is Bella. She hopes her smile made you smile. If not,
she is also offering you her favorite monkey. 13.5/10 https://t.co/qjrljjt948
523                         This is Logan, the Chow who lived. He solemnly swears
he's up to lots of good. H*ckin magical af 9.75/10 https://t.co/yBO5wuqaPS
579   This is Sophie. She's a Jubilant Bush Pupper. Super h*ckin rare. Appears
at random just to smile at the locals. 11.27/10 would smile back https://…
1463                                          Here we have uncovered an
entire battalion of holiday puppers. Average of 11.26/10 https://t.co/eNm2S6p9BD

      retweeted_status_id  retweeted_status_user_id  \
41                    NaN                       NaN
523                   NaN                       NaN
579                   NaN                       NaN
1463                  NaN                       NaN

     retweeted_status_timestamp  …                p1   p1_conf  p1_dog  \
41                          NaT  …  golden_retriever  0.943082    True
523                         NaT  …        Pomeranian  0.467321    True
579                         NaT  …           clumber  0.946718    True
1463                        NaT  …            kuvasz  0.438627    True

                      p2   p2_conf p2_dog              p3   p3_conf p3_dog  \
41    Labrador_retriever  0.032409   True          kuvasz  0.005501   True
523          Persian_cat  0.122978  False            chow  0.102654   True
579       cocker_spaniel  0.015950   True           Lhasa  0.006519   True
1463             Samoyed  0.111622   True  Great_Pyrenees  0.064061   True

      stg_count
41            0
523           0
579           1
1463          0

[4 rows x 30 columns]
```

Test

[33]:
```
      index            tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
41       45  883482846933004288                    NaN                  NaN
523     695  786709082849828864                    NaN                  NaN
579     763  778027034220126208                    NaN                  NaN
1463   1712  680494726643068929                    NaN                  NaN
```

```
              timestamp  \
41   2017-07-08 00:28:19
523  2016-10-13 23:23:56
579  2016-09-20 00:24:34
1463 2015-12-25 21:06:00

  source  \
41    <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
523   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
579   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
1463  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>

                                                                    text  \
41                 This is Bella. She hopes her smile made you smile. If not,
she is also offering you her favorite monkey. 13.5/10 https://t.co/qjrljjt948
523                       This is Logan, the Chow who lived. He solemnly swears
he's up to lots of good. H*ckin magical af 9.75/10 https://t.co/yBO5wuqaPS
579   This is Sophie. She's a Jubilant Bush Pupper. Super h*ckin rare. Appears
at random just to smile at the locals. 11.27/10 would smile back https://…
1463                                          Here we have uncovered an
entire battalion of holiday puppers. Average of 11.26/10 https://t.co/eNm2S6p9BD

     retweeted_status_id  retweeted_status_user_id  \
41                   NaN                       NaN
523                  NaN                       NaN
579                  NaN                       NaN
1463                 NaN                       NaN

     retweeted_status_timestamp  …                p1   p1_conf  p1_dog  \
41                          NaT  …  golden_retriever  0.943082    True
523                         NaT  …        Pomeranian  0.467321    True
579                         NaT  …           clumber  0.946718    True
1463                        NaT  …            kuvasz  0.438627    True

                    p2   p2_conf p2_dog             p3   p3_conf p3_dog  \
41    Labrador_retriever  0.032409   True          kuvasz  0.005501   True
523          Persian_cat  0.122978  False            chow  0.102654   True
579       cocker_spaniel  0.015950   True           Lhasa  0.006519   True
1463             Samoyed  0.111622   True  Great_Pyrenees  0.064061   True

      stg_count
41            0
523           0
```

```
579           1
1463          0

[4 rows x 30 columns]
```

Define Mask not empty “rin_reply_to_status_id” column and Drop Rows

Code

1.3.9 Issue #7:

Tweets with “This is a —” geting Dog name as “a”

Define Mask rows witn with name column entry equal ‘a’ and change for NaN

Code

Test

[35]: 0

1.3.10 Issue #8:

Not necessary Columns:

source in_reply_to_status_id in_reply_to_user_id retweeted_status_id retweeted_status_user_id retweeted_status_timestamp

Define Drop the columns:

source

in_reply_to_status_id

in_reply_to_user_id

retweeted_status_id

retweeted_status_user_id

retweeted_status_timestamp

Code

Test

```
Index(['index', 'tweet_id', 'timestamp', 'text', 'expanded_urls',
       'rating_numerator', 'rating_denominator', 'name', 'doggo', 'floofer',
       'pupper', 'puppo', 'jpg_url', 'img_num', 'p1', 'p1_conf', 'p1_dog',
       'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog', 'stg_count'],
      dtype='object')
```

1.4 Tidiness issues

1.4.1 Issue #1

Text Column with text and URL

Define 1 - At DataSet 1 create new column twt_url

2 - Extract from DataSet 1 text column the url, and copy to colugn twt_url

Code

Test

[39]:
```
   index            tweet_id           timestamp  \
0      0  892420643555336193 2017-08-01 16:23:56
1      1  892177421306343426 2017-08-01 00:17:27
2      2  891815181378084864 2017-07-31 00:18:03
3      3  891689557279858688 2017-07-30 15:58:51
4      4  891327558926688256 2017-07-29 16:00:24

                                                                text  \
0                                                              This is Phineas. He's a
mystical boy. Only ever appears in the hole of a donut. 13/10
1  This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not,
she's available for pats, snugs, boops, the whole bit. 13/10
2                   This is Archie. He is a rare Norwegian Pouncing Corgo. Lives
in the tall grass. You never know when one may strike. 12/10
3                                                                    This is Darla. She
commenced a snooze mid meal. 13/10 happens to the best of us
4  This is Franklin. He would like you to stop calling him "cute." He is a very
fierce shark and should be respected as such. 12/10 #BarkWeek

                                      expanded_urls  \
0
https://twitter.com/dog_rates/status/892420643555336193/photo/1
1
https://twitter.com/dog_rates/status/892177421306343426/photo/1
2
https://twitter.com/dog_rates/status/891815181378084864/photo/1
3
https://twitter.com/dog_rates/status/891689557279858688/photo/1
4  https://twitter.com/dog_rates/status/891327558926688256/photo/1,https://twitt
er.com/dog_rates/status/891327558926688256/photo/1

   rating_numerator  rating_denominator      name doggo floofer  …  \
0                13                  10   Phineas   NaN     NaN  …
1                13                  10     Tilly   NaN     NaN  …
2                12                  10    Archie   NaN     NaN  …
```

```
3               13                  10     Darla   NaN     NaN  …
4               12                  10  Franklin   NaN     NaN  …

    p1_conf p1_dog                  p2   p2_conf p2_dog  \
0  0.097049  False               bagel  0.085851  False
1  0.323581   True            Pekinese  0.090647   True
2  0.716012   True           malamute  0.078253   True
3  0.170278  False  Labrador_retriever  0.168086   True
4  0.555712   True    English_springer  0.225770   True

                           p3   p3_conf p3_dog  stg_count  \
0                      banana  0.076110  False          0
1                    papillon  0.068957   True          0
2                      kelpie  0.031379   True          0
3                     spatula  0.040836  False          0
4  German_short-haired_pointer  0.175219   True          0

                   twt_url
0  https://t.co/MgUWQ76dJU
1  https://t.co/0Xxu71qeIV
2  https://t.co/wUnZnhtVJB
3  https://t.co/tD36da7qLQ
4  https://t.co/AtUZn91f7f

[5 rows x 25 columns]
```

1.4.2 Issue #2

Stage of dogs in columns

Define 1- Create column dog_stage.

2- Copy Dog from individual colugn to dog_stage column leaving missing values as Nan.

3- Remove individual dog stage column.

Code

Test

[41]:
```
   index             tweet_id           timestamp  \
0      0  892420643555336193 2017-08-01 16:23:56
1      1  892177421306343426 2017-08-01 00:17:27
2      2  891815181378084864 2017-07-31 00:18:03
3      3  891689557279858688 2017-07-30 15:58:51
4      4  891327558926688256 2017-07-29 16:00:24

                                                text  \
0                                 This is Phineas. He's a
```

```
mystical boy. Only ever appears in the hole of a donut. 13/10
1  This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not,
she's available for pats, snugs, boops, the whole bit. 13/10
2                   This is Archie. He is a rare Norwegian Pouncing Corgo. Lives
in the tall grass. You never know when one may strike. 12/10
3                                                             This is Darla. She
commenced a snooze mid meal. 13/10 happens to the best of us
4  This is Franklin. He would like you to stop calling him "cute." He is a very
fierce shark and should be respected as such. 12/10 #BarkWeek

                                    expanded_urls  \
0
https://twitter.com/dog_rates/status/892420643555336193/photo/1
1
https://twitter.com/dog_rates/status/892177421306343426/photo/1
2
https://twitter.com/dog_rates/status/891815181378084864/photo/1
3
https://twitter.com/dog_rates/status/891689557279858688/photo/1
4  https://twitter.com/dog_rates/status/891327558926688256/photo/1,https://twitt
er.com/dog_rates/status/891327558926688256/photo/1

   rating_numerator  rating_denominator      name doggo floofer  … p1_dog  \
0                13                  10   Phineas   NaN     NaN  …  False
1                13                  10     Tilly   NaN     NaN  …   True
2                12                  10    Archie   NaN     NaN  …   True
3                13                  10     Darla   NaN     NaN  …  False
4                12                  10  Franklin   NaN     NaN  …   True

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

    p3_conf p3_dog stg_count                 twt_url dog_stage
0  0.076110  False         0  https://t.co/MgUWQ76dJU       NaN
1  0.068957   True         0  https://t.co/0Xxu71qeIV       NaN
2  0.031379   True         0  https://t.co/wUnZnhtVJB       NaN
3  0.040836  False         0  https://t.co/tD36da7qLQ       NaN
4  0.175219   True         0  https://t.co/AtUZn91f7f       NaN

[5 rows x 26 columns]
```

1.4.3 Clean Issue #9

Remove unnecessary columns 'doggo', 'floofer', 'pupper', 'puppo', 'stg_count'

define Drop columns: 'doggo', 'floofer', 'pupper', 'puppo', 'stg_count'

code

test

```
Index(['tweet_id', 'timestamp', 'text', 'expanded_urls', 'rating_numerator',
       'rating_denominator', 'name', 'jpg_url', 'img_num', 'p1', 'p1_conf',
       'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog',
       'twt_url', 'dog_stage'],
      dtype='object')
```

1.4.4 Clean Issue #10

Filling missing values

[44]:
```
tweet_id                0
timestamp               0
text                    0
expanded_urls           3
rating_numerator        0
rating_denominator      0
name                  652
jpg_url               125
img_num               125
p1                    125
p1_conf               125
p1_dog                125
p2                    125
p2_conf               125
p2_dog                125
p3                    125
p3_conf               125
p3_dog                125
twt_url                 3
dog_stage            1761
dtype: int64
```

define Fill all missing values with 'Unkown'

code

test

[46]:
```
tweet_id              0
timestamp             0
text                  0
expanded_urls         0
rating_numerator      0
rating_denominator    0
name                  0
jpg_url               0
img_num               0
p1                    0
p1_conf               0
p1_dog                0
p2                    0
p2_conf               0
p2_dog                0
p3                    0
p3_conf               0
p3_dog                0
twt_url               0
dog_stage             0
dtype: int64
```

1.5 Storing Data

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2086 entries, 0 to 2085
Data columns (total 20 columns):
tweet_id              2086 non-null int64
timestamp             2086 non-null datetime64[ns]
text                  2086 non-null object
expanded_urls         2086 non-null object
rating_numerator      2086 non-null int64
rating_denominator    2086 non-null int64
name                  2086 non-null object
jpg_url               2086 non-null object
img_num               2086 non-null object
p1                    2086 non-null object
p1_conf               2086 non-null object
p1_dog                2086 non-null object
p2                    2086 non-null object
p2_conf               2086 non-null object
p2_dog                2086 non-null object
p3                    2086 non-null object
p3_conf               2086 non-null object
p3_dog                2086 non-null object
```

```
twt_url               2086 non-null object
dog_stage             2086 non-null object
dtypes: datetime64[ns](1), int64(3), object(16)
memory usage: 326.1+ KB
```

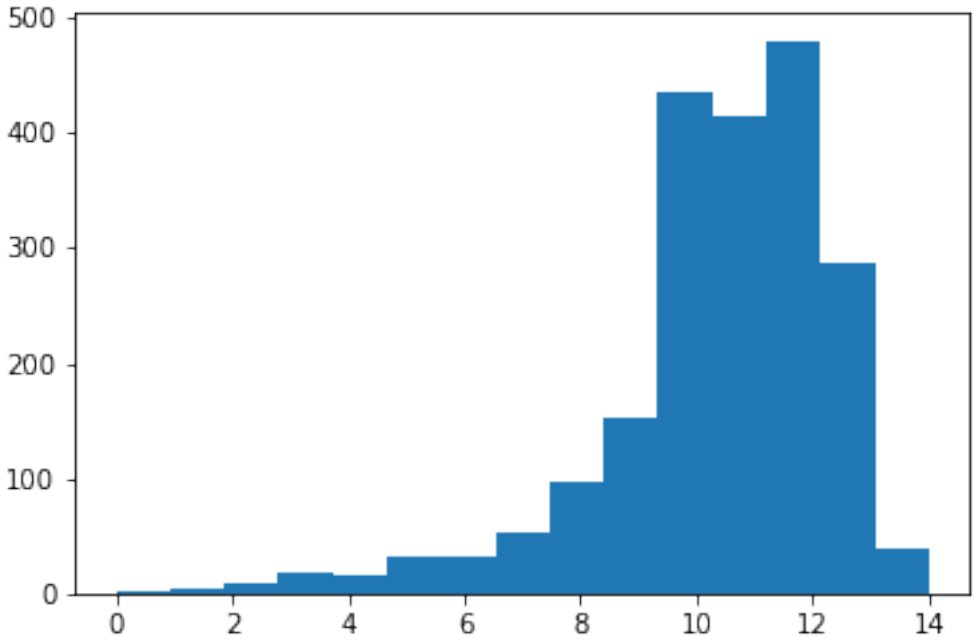
1.6 Analyzing and Visualizing Data

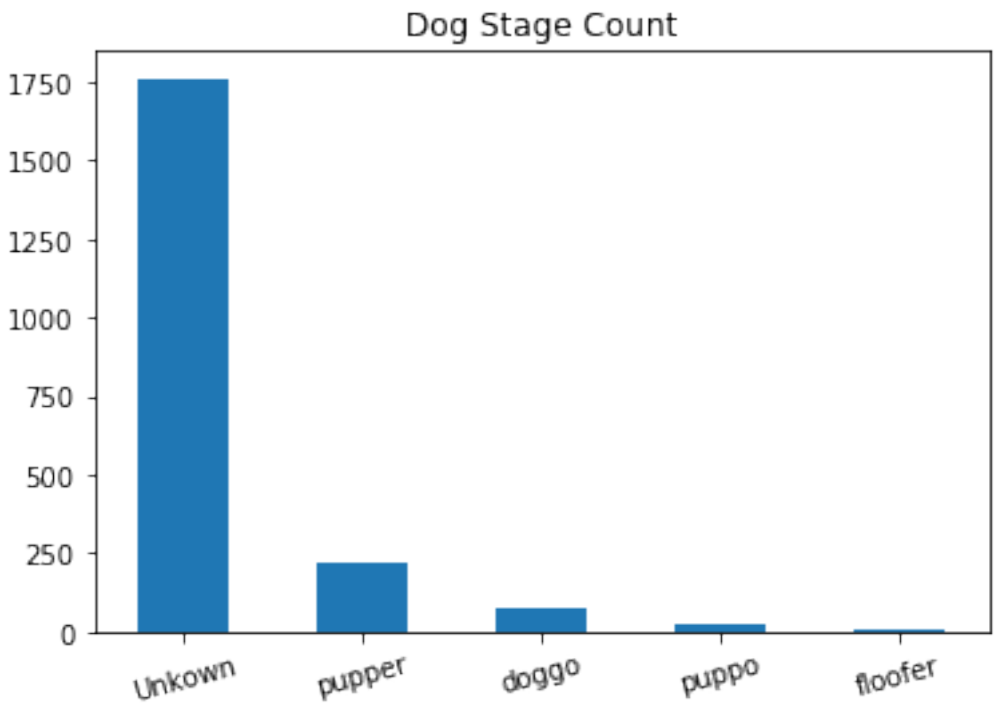
In this section, analyze and visualize your wrangled data. You must produce at least **three (3) insights and one (1) visualization.**

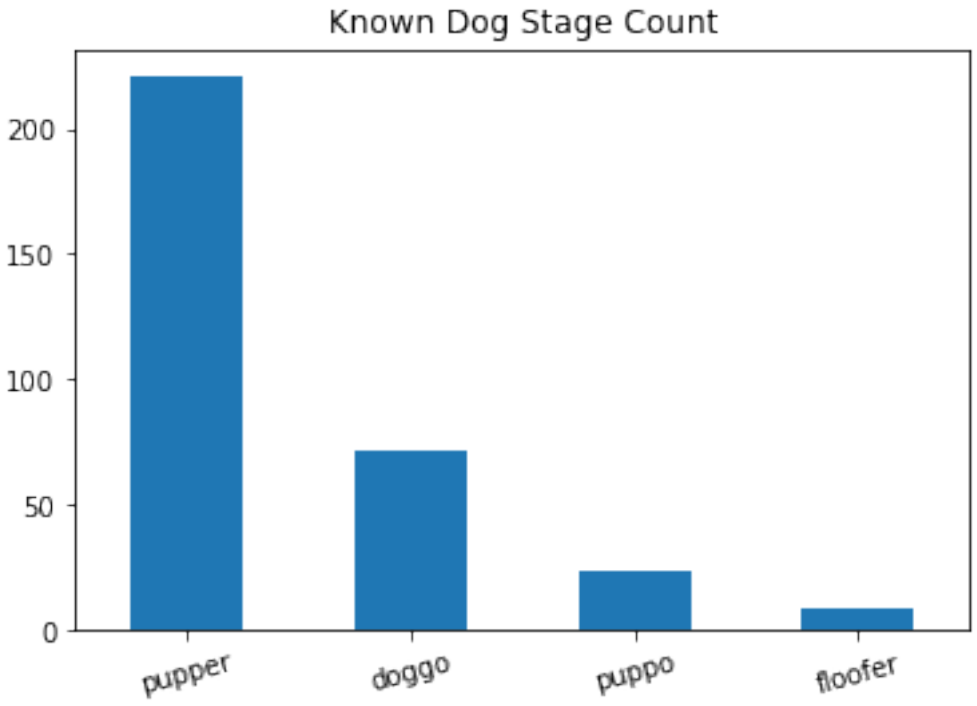
```
Dog Stage Nan ratio: 0.8441994247363375
Dog Names Nan ratio: 0.31255992329817833
```

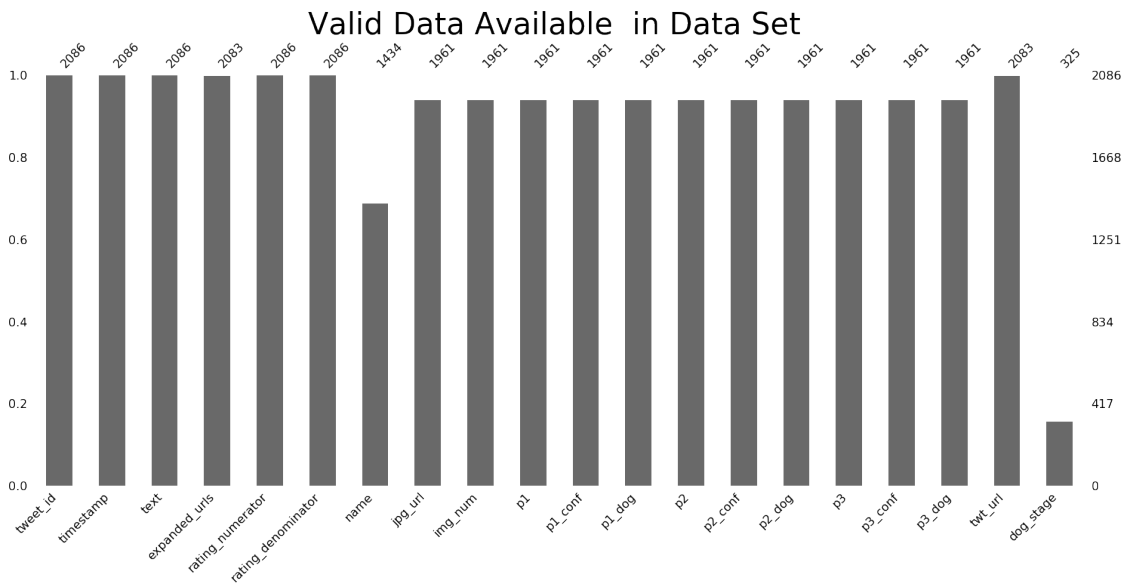
[50]:
```
count    2086.000000
mean       12.151965
std        40.444893
min         0.000000
25%        10.000000
50%        11.000000
75%        12.000000
max      1776.000000
Name: rating_numerator, dtype: float64
```

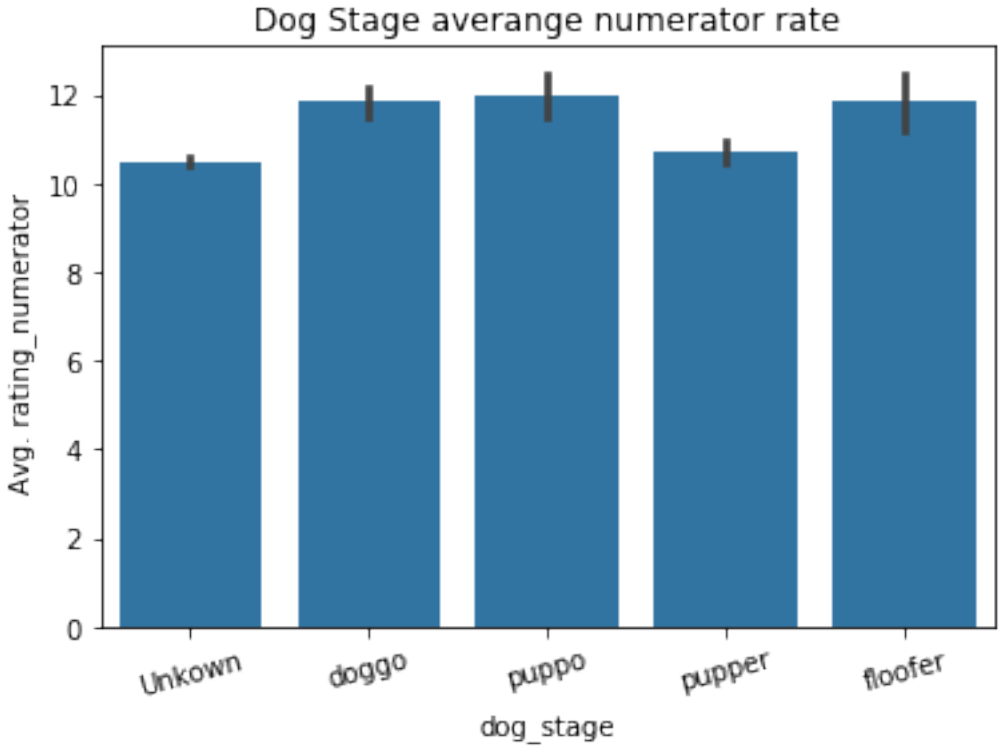
1.6.1 Filtering outliers at the rating_numerator column




Dog Stage Count
1750
1500
1250
1000
750
500
250
0
Unkown
pupper
doggo
puppo
floofer


Known Dog Stage Count
200
150
100
50
0
pupper
doggo
puppo
floofer





1.6.2 Insights:

1. Dog Stage and Dog Name columns are sparce.

Dog Stage Nan ratio: 0.8441994247363375

Dog Names Nan ratio: 0.31255992329817833

2. For the known dog stages the most common is the pupper followed by doggo, puppo and floofer.

3. Puppers dog has the less numerator rating averange.

1.6.3 Visualization

Visualizing cleaned dataframe sparsity