Act_report

October 29, 2019

1 Analysing and Visualsing WeRate Dogs

```
In [18]: import matplotlib
         import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as sns
         %matplotlib inline

In [19]: matplotlib.style.use('ggplot')

In [28]: df = pd.read_csv('twitter_archive_master.csv')
         df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1994 entries, 0 to 1993
Data columns (total 43 columns):
tweet_id                        1994 non-null int64
source_x                        1994 non-null object
text_x                          1994 non-null object
expanded_urls                   1994 non-null object
rating_numerator                1994 non-null float64
rating_denominator              1994 non-null float64
name                            1994 non-null object
contributors                    0 non-null float64
coordinates                     0 non-null float64
created_at                      1994 non-null object
entities                        1994 non-null object
extended_entities               1749 non-null object
favorite_count                  1994 non-null int64
favorited                       1994 non-null bool
geo                             0 non-null float64
id                              1994 non-null int64
id_str                          1994 non-null int64
in_reply_to_screen_name         23 non-null object
in_reply_to_status_id_y         23 non-null float64
in_reply_to_status_id_str       23 non-null float64
```

```
in_reply_to_user_id_y            23 non-null float64
in_reply_to_user_id_str          23 non-null float64
is_quote_status                  1994 non-null bool
lang                             1994 non-null object
place                            1 non-null object
possibly_sensitive               1994 non-null float64
possibly_sensitive_appealable    1994 non-null float64
quoted_status                    0 non-null float64
quoted_status_id                 0 non-null float64
quoted_status_id_str             0 non-null float64
retweet_count                    1994 non-null int64
retweeted                        1994 non-null bool
retweeted_status                 0 non-null float64
source_y                         1994 non-null object
text_y                           1994 non-null object
truncated                        1994 non-null bool
user                             1994 non-null object
jpg_url                          1994 non-null object
dog_stage                        1994 non-null object
prediction_algorithm             1686 non-null object
confidence_level                 1994 non-null float64
source                           1994 non-null object
dog_gender                       862 non-null object
dtypes: bool(4), float64(16), int64(5), object(18)
memory usage: 615.4+ KB
```

```
In [29]: df.head()
```

Out[29]:

	tweet_id	source_x	text_x	expanded_urls	rating_numerator
0	667405339315146752	<a href="http://twitter.com/download/iphone" r...	This is Biden. Biden just tripped... 7/10 http...	https://twitter.com/dog_rates/status/667405339...	7.0
1	667435689202614272	<a href="http://twitter.com/download/iphone" r...	Ermergerd 12/10 https://t.co/PQni2sjPsm	https://twitter.com/dog_rates/status/667435689...	12.0
2	667437278097252352	<a href="http://twitter.com/download/iphone" r...	Never seen this breed before. Very pointy pup...	https://twitter.com/dog_rates/status/667437278...	10.0
3	667443425659232256	<a href="http://twitter.com/download/iphone" r...	Exotic dog here. Long neck. Weird paws. Obsess...	https://twitter.com/dog_rates/status/667443425...	6.0
4	667453023279554560	<a href="http://twitter.com" rel="nofollow">Tw...	Meet Cupcake. I would do unspeakable things fo...		

```
4  https://twitter.com/dog_rates/status/667453023...              11.0

   rating_denominator     name  contributors  coordinates  \
0                10.0    Biden           NaN          NaN
1                10.0     None           NaN          NaN
2                10.0     None           NaN          NaN
3                10.0     None           NaN          NaN
4                10.0  Cupcake           NaN          NaN

           created_at     ...      \
0  2015-11-19 18:13:27     ...
1  2015-11-19 20:14:03     ...
2  2015-11-19 20:20:22     ...
3  2015-11-19 20:44:47     ...
4  2015-11-19 21:22:56     ...

                                           source_y  \
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" rel="nofollow">Tw...

                                             text_y  truncated  \
0  This is Biden. Biden just tripped... 7/10 http...      False
1            Ermergerd 12/10 https://t.co/PQni2sjPsm      False
2  Never seen this breed before. Very pointy pup...      False
3  Exotic dog here. Long neck. Weird paws. Obsess...      False
4  Meet Cupcake. I would do unspeakable things fo...      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...

                                         jpg_url  dog_stage  \
0  https://pbs.twimg.com/media/CUMZnmhUEAEbtis.jpg       None
1  https://pbs.twimg.com/media/CUM1OHCW4AEgGSi.jpg       None
2  https://pbs.twimg.com/media/CUM2qWaWoAUZO6L.jpg       None
3  https://pbs.twimg.com/media/CUM8QZwW4AAVsBl.jpg       None
4  https://pbs.twimg.com/media/CUNE_OSUwAAdHhX.jpg       None

   prediction_algorithm confidence_level              source  dog_gender
0         Saint_Bernard         0.381377  Twitter for iPhone         NaN
1            Rottweiler         0.999091  Twitter for iPhone         NaN
2                   NaN         0.000000  Twitter for iPhone         NaN
```

```
3                    NaN         0.000000  Twitter for iPhone         NaN
4    Labrador_retriever         0.825670  Twitter Web Client         NaN

[5 rows x 43 columns]
```

In [33]:
```
# Convert columns to their appropriate types and set the timestamp as an index

df['tweet_id'] = df['tweet_id'].astype(object)
df['timestamp'] = pd.to_datetime(df.created_at)
df['source'] = df['source'].astype('category')
df['dog_stage'] = df['dog_stage'].astype('category')

df.set_index('timestamp', inplace=True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1994 entries, 2015-11-19 18:13:27 to 2016-06-16 01:25:36
Data columns (total 43 columns):
tweet_id                          1994 non-null object
source_x                          1994 non-null object
text_x                            1994 non-null object
expanded_urls                     1994 non-null object
rating_numerator                  1994 non-null float64
rating_denominator                1994 non-null float64
name                              1994 non-null object
contributors                      0 non-null float64
coordinates                       0 non-null float64
created_at                        1994 non-null object
entities                          1994 non-null object
extended_entities                 1749 non-null object
favorite_count                    1994 non-null int64
favorited                         1994 non-null bool
geo                               0 non-null float64
id                                1994 non-null int64
id_str                            1994 non-null int64
in_reply_to_screen_name           23 non-null object
in_reply_to_status_id_y           23 non-null float64
in_reply_to_status_id_str         23 non-null float64
in_reply_to_user_id_y             23 non-null float64
in_reply_to_user_id_str           23 non-null float64
is_quote_status                   1994 non-null bool
lang                              1994 non-null object
place                             1 non-null object
possibly_sensitive                1994 non-null float64
possibly_sensitive_appealable     1994 non-null float64
quoted_status                     0 non-null float64
quoted_status_id                  0 non-null float64
quoted_status_id_str              0 non-null float64
```

```
retweet_count                    1994 non-null int64
retweeted                        1994 non-null bool
retweeted_status                 0 non-null float64
source_y                         1994 non-null object
text_y                           1994 non-null object
truncated                        1994 non-null bool
user                             1994 non-null object
jpg_url                          1994 non-null object
dog_stage                        1994 non-null category
prediction_algorithm             1686 non-null object
confidence_level                 1994 non-null float64
source                           1994 non-null category
dog_gender                       862 non-null object
dtypes: bool(4), category(2), float64(16), int64(4), object(17)
memory usage: 603.9+ KB
```

```
In [34]: df.describe()
```

Out[34]:

	rating_numerator	rating_denominator	contributors	coordinates
count	1994.000000	1994.000000	0.0	0.0
mean	12.212528	10.510030	NaN	NaN
std	41.463532	7.261522	NaN	NaN
min	0.000000	7.000000	NaN	NaN
25%	10.000000	10.000000	NaN	NaN
50%	11.000000	10.000000	NaN	NaN
75%	12.000000	10.000000	NaN	NaN
max	1776.000000	170.000000	NaN	NaN

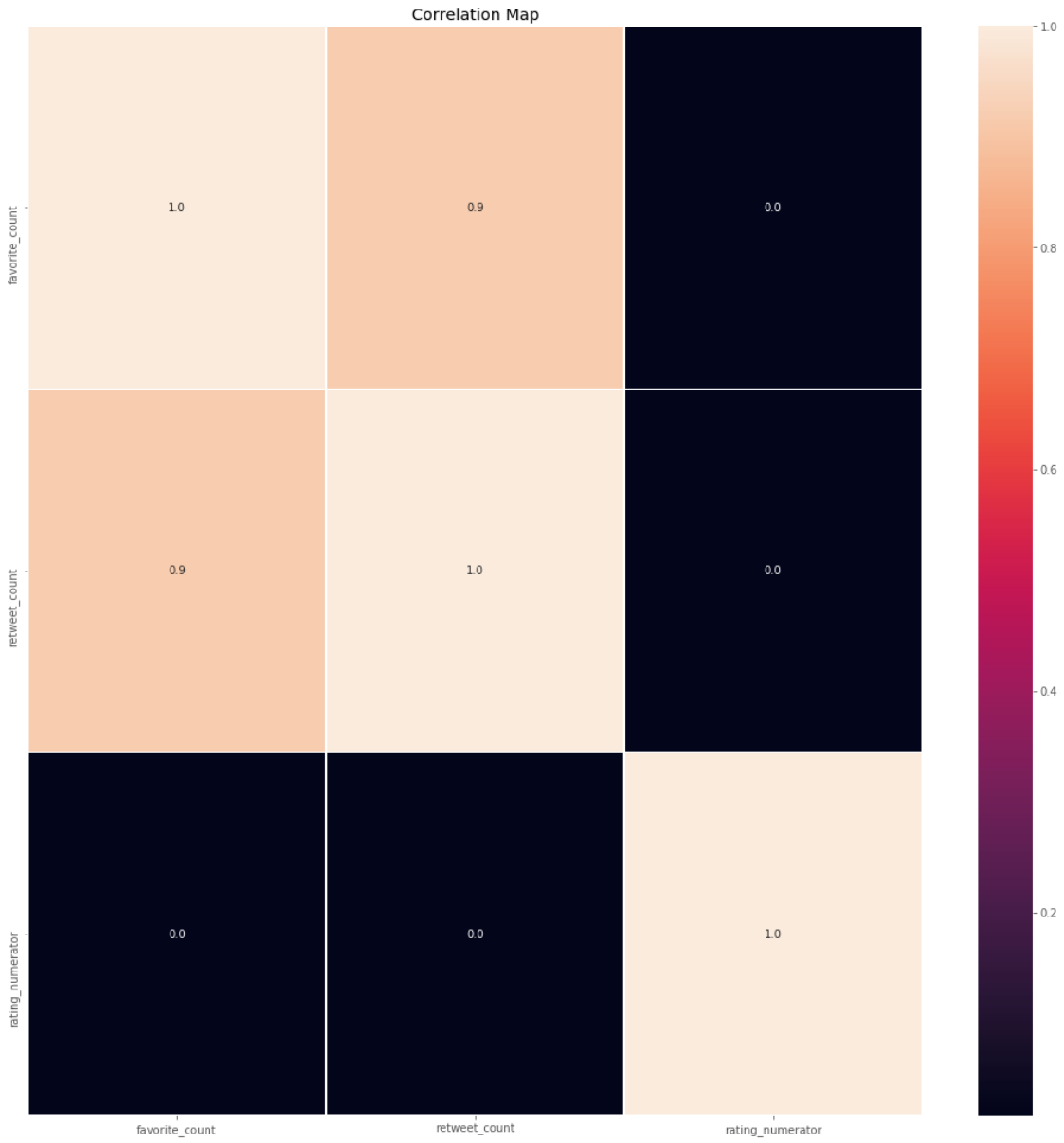
	favorite_count	geo	id	id_str
count	1994.000000	0.0	1.994000e+03	1.994000e+03
mean	8858.264293	NaN	7.358508e+17	7.358508e+17
std	12577.702272	NaN	6.747816e+16	6.747816e+16
min	79.000000	NaN	6.660209e+17	6.660209e+17
25%	1927.500000	NaN	6.758475e+17	6.758475e+17
50%	4048.500000	NaN	7.084748e+17	7.084748e+17
75%	11160.500000	NaN	7.877873e+17	7.877873e+17
max	143519.000000	NaN	8.924206e+17	8.924206e+17

	in_reply_to_status_id_y	in_reply_to_status_id_str
count	2.300000e+01	2.300000e+01
mean	6.978112e+17	6.978112e+17
std	4.359384e+16	4.359384e+16
min	6.671522e+17	6.671522e+17
25%	6.732411e+17	6.732411e+17
50%	6.757073e+17	6.757073e+17
75%	7.031489e+17	7.031489e+17
max	8.558181e+17	8.558181e+17

```
       in_reply_to_user_id_y  in_reply_to_user_id_str  possibly_sensitive  \
count           2.300000e+01             2.300000e+01              1994.0
mean            4.196984e+09             4.196984e+09                 0.0
std             0.000000e+00             0.000000e+00                 0.0
min             4.196984e+09             4.196984e+09                 0.0
25%             4.196984e+09             4.196984e+09                 0.0
50%             4.196984e+09             4.196984e+09                 0.0
75%             4.196984e+09             4.196984e+09                 0.0
max             4.196984e+09             4.196984e+09                 0.0

       possibly_sensitive_appealable  quoted_status  quoted_status_id  \
count                         1994.0            0.0               0.0
mean                             0.0            NaN               NaN
std                              0.0            NaN               NaN
min                              0.0            NaN               NaN
25%                              0.0            NaN               NaN
50%                              0.0            NaN               NaN
75%                              0.0            NaN               NaN
max                              0.0            NaN               NaN

       quoted_status_id_str  retweet_count  retweeted_status  confidence_level
count                   0.0    1994.000000               0.0       1994.000000
mean                    NaN    2720.340020               NaN          0.464991
std                     NaN    4697.005583               NaN          0.339470
min                     NaN      13.000000               NaN          0.000000
25%                     NaN     608.250000               NaN          0.140466
50%                     NaN    1313.000000               NaN          0.459130
75%                     NaN    3126.750000               NaN          0.776387
max                     NaN   77435.000000               NaN          0.999956
```

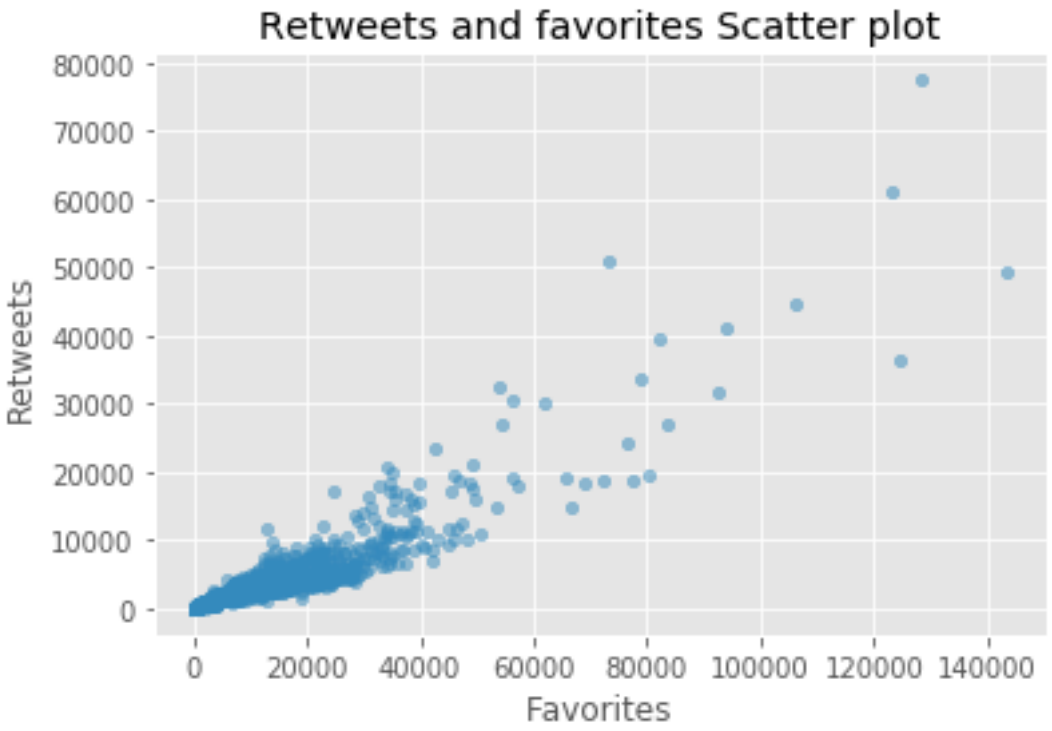
In [37]:
```
f,ax = plt.subplots(figsize=(18, 18))
sns.heatmap(df[['source', 'favorite_count','retweet_count',
                'rating_numerator']].corr(), annot=True, linewidths=.5, fmt= '.1f',ax=a
plt.title('Correlation Map')
```

Out[37]:
```
Text(0.5,1,'Correlation Map')
```



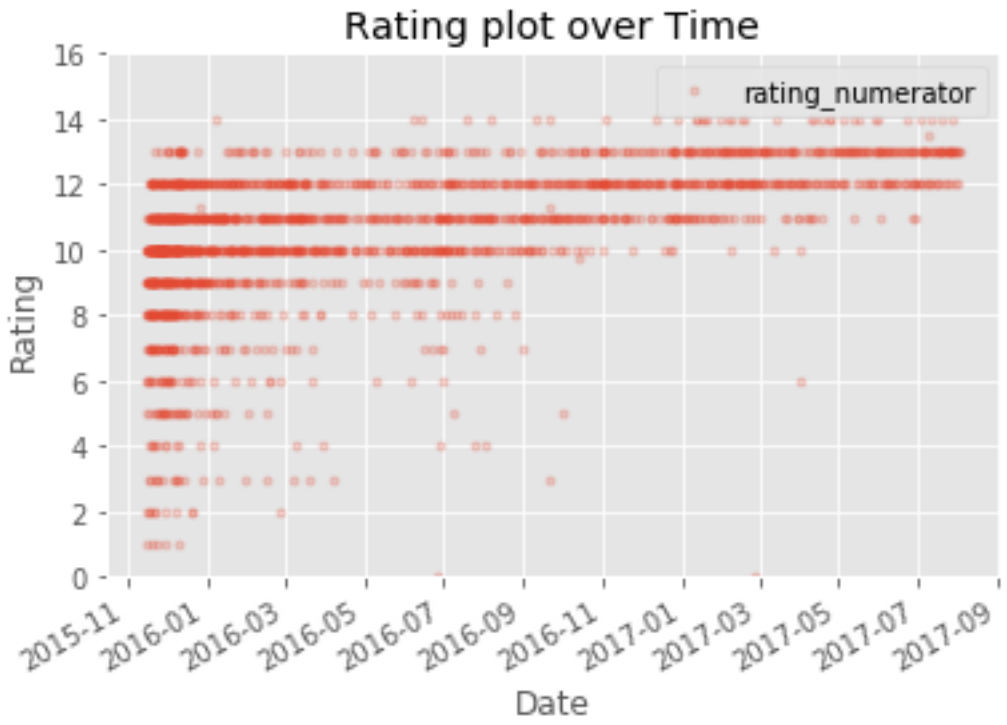
We can see that there is high corelation between (faviorate_count & rating_numerator) and also between (retweet_count & rating_numerator)

```
In [38]: df.plot(kind='scatter',x='favorite_count',y='retweet_count', alpha = 0.5)
         plt.xlabel('Favorites')
         plt.ylabel('Retweets')
         plt.title('Retweets and favorites Scatter plot')

Out[38]: Text(0.5,1,'Retweets and favorites Scatter plot')
```



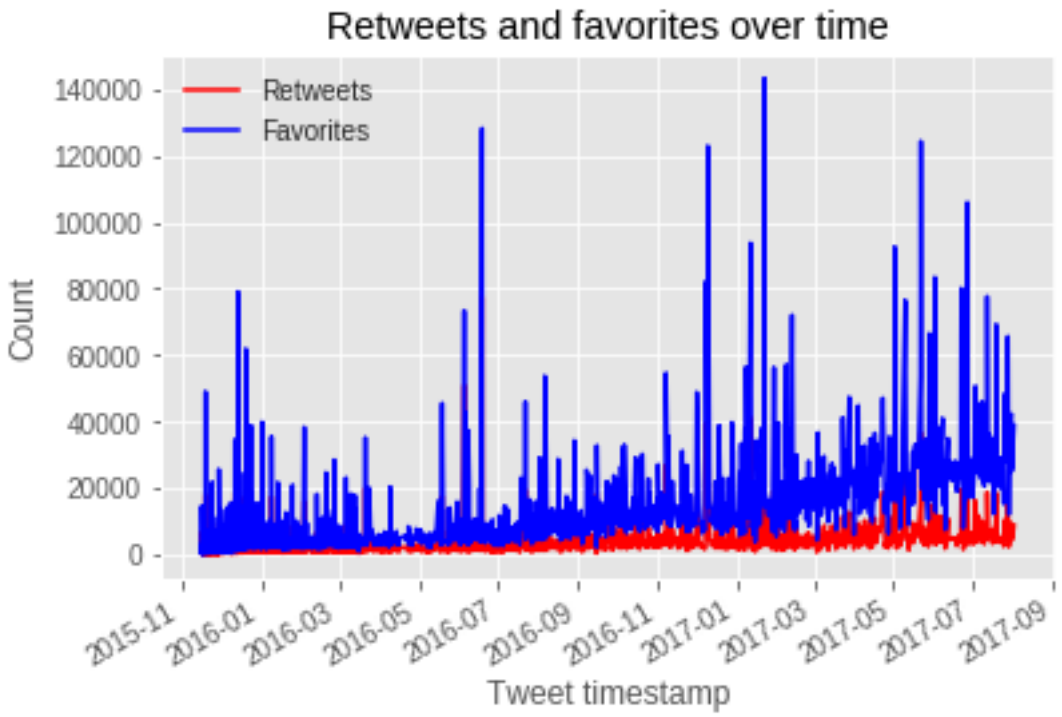
we noticed that there is a positive correlation between Retweets & Favorites.

```
In [39]: # Our range will be [0,16] taking of the two ouliers (1776 and 420)
         df.plot(y ='rating_numerator', ylim=[0,16], style = '.', alpha = .2)
         plt.title('Rating plot over Time')
         plt.xlabel('Date')
         plt.ylabel('Rating')

Out[39]: Text(0,0.5,'Rating')
```



We can see that people are giving rating above 8 over time increases. most of the rating are of 10 to 12

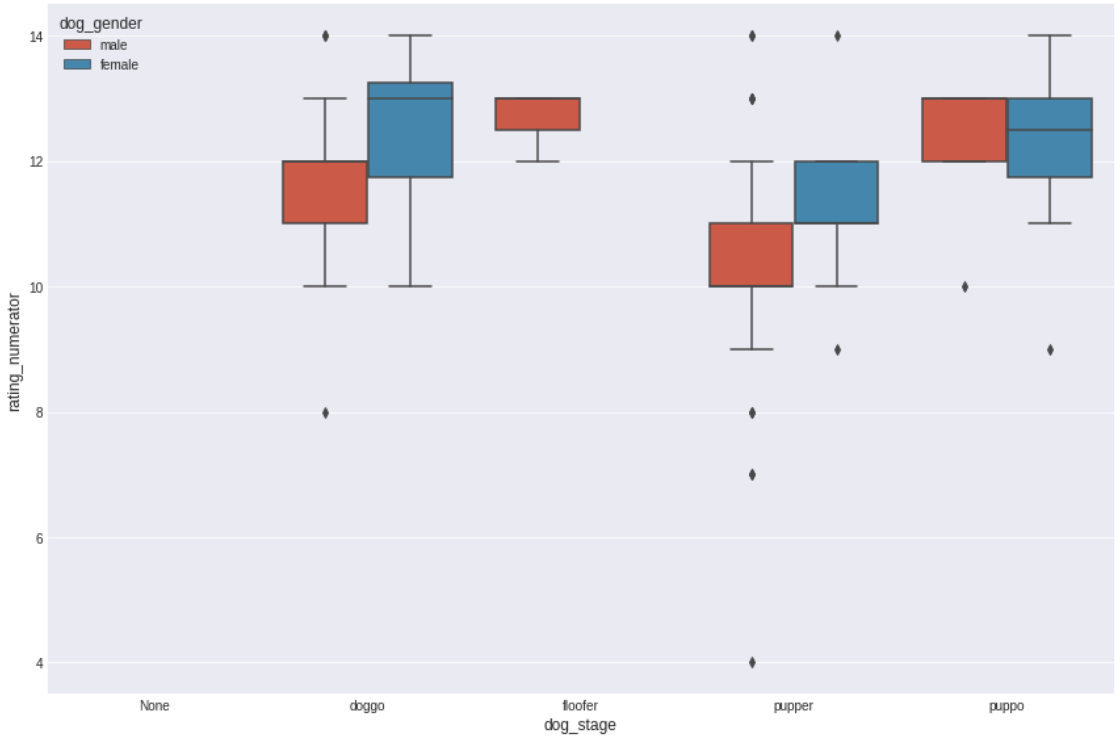
```
In [40]: df['retweet_count'].plot(color = 'red', label='Retweets')
         df['favorite_count'].plot(color = 'blue', label='Favorites')
         plt.style.use('seaborn-darkgrid')
         plt.legend(loc='upper left')
         plt.xlabel('Tweet timestamp')
         plt.ylabel('Count')
         plt.title('Retweets and favorites over time')
         plt.savefig('retweets_favorites.png')
         plt.show()
```



By the following plot we conclude that faviorate count is more than number of retweets and both gradually increases over time

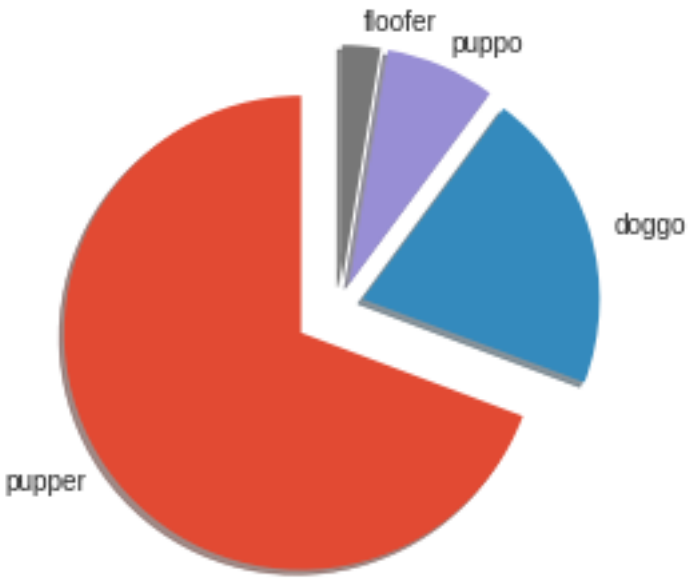
```
In [41]: sns.factorplot(kind='box',
                        y='rating_numerator',
                        x='dog_stage',
                        hue='dog_gender',
                        data=df[df['dog_stage'] != 'None'],
                        size=8,
                        aspect=1.5,
                        legend_out=False)
```

```
Out[41]: <seaborn.axisgrid.FacetGrid at 0x7f5af16e4048>
```



This is Gender Count of 5 different types of dog stages

```
In [42]: # Plot the data partitioned by dog stage
         dog_stage_count = list(df[df['dog_stage'] != 'None']['dog_stage'].value_counts())[0:4]
         dog_stages = df[df['dog_stage'] != 'None']['dog_stage'].value_counts().index.tolist()[0
         explode = (0.2, 0.1, 0.1, 0.1)

         fig1, ax1 = plt.subplots()
         ax1.pie(dog_stage_count, explode = explode, labels = dog_stages, shadow = True, startan
         ax1.axis('equal')

Out[42]: (-1.288268191449591,
          1.2310305760057396,
          -1.2401381220397572,
          1.2110819987279693)
```



We find that many pictures are of pupper through weRateDogs Twitter account

```
In [43]: df[df['dog_stage'] != 'None'].groupby('dog_stage')['rating_numerator'].mean()

Out[43]: dog_stage
         None              NaN
         doggo       11.888889
         floofer     11.875000
         pupper      10.638066
         puppo       12.043478
         Name: rating_numerator, dtype: float64
```