Twitter Data Analysis and Visualization Report

This report is to communicate the insights and display visualization(s) produced from the wrangled data from WeRateDogs twitter archive.

Retweets and replies have been excluded from the dataset. Only original ratings with images in the tweets have been included. The data wrangling report details the gathering, assessing and cleaning steps performed to prepare the data for the following analysis.

Analysis of distributions for some parameters of interest

Distribution of rating numerator

Using describe() shows us that the max value is something unlikely for the numerator

```
twitter_clean.rating_numerator.describe()
```

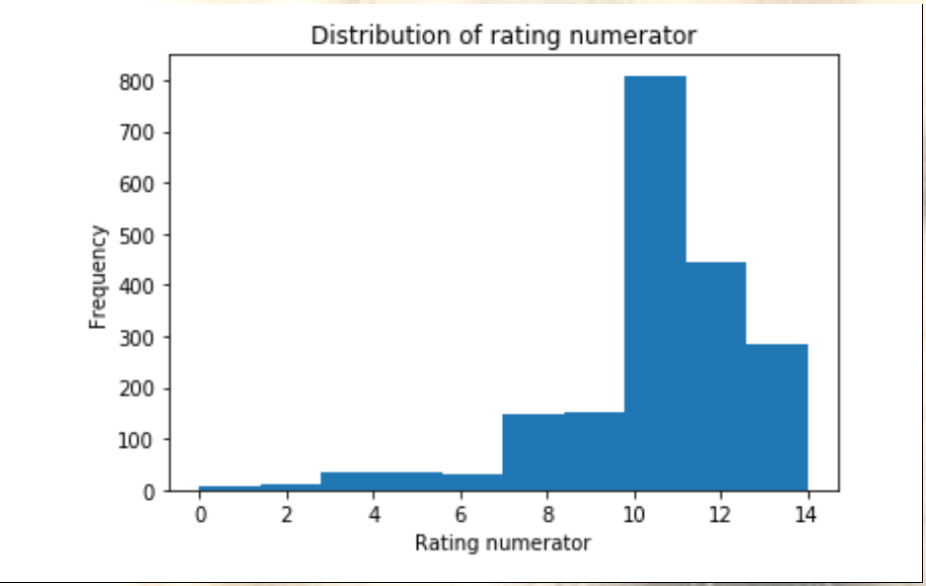
```
count    1971.000000
mean       12.223237
std        41.634034
min         0.000000
25%        10.000000
50%        11.000000
75%        12.000000
max      1776.000000
Name: rating_numerator, dtype: float64
```

```
: sum(twitter_clean.rating_numerator > 15)
```

```
: 18
```

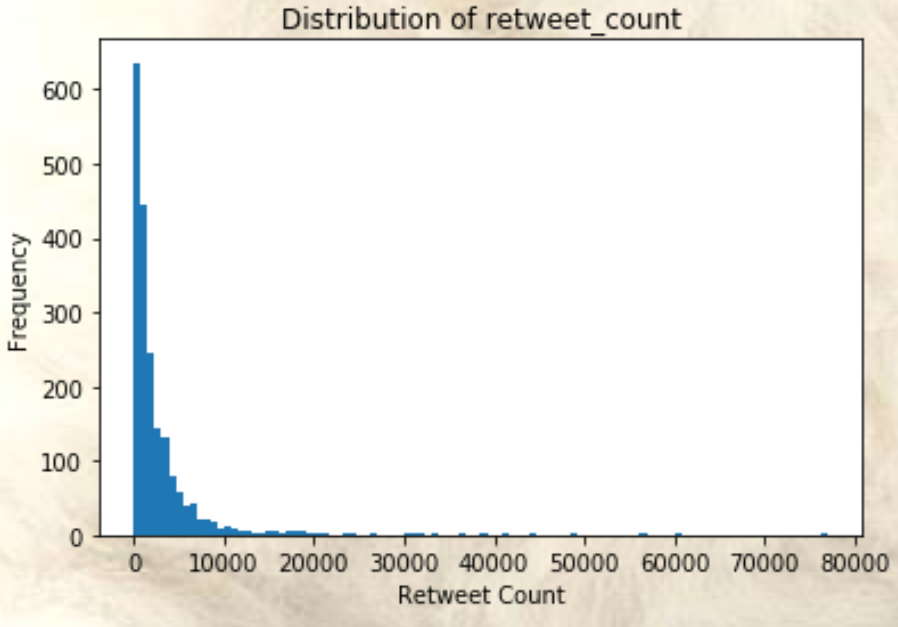
There are 18 entries where the numerator is greater than 15. So I decided to exclude them from the analysis.

The distribution of rating numerator values less than 15 is as follows:



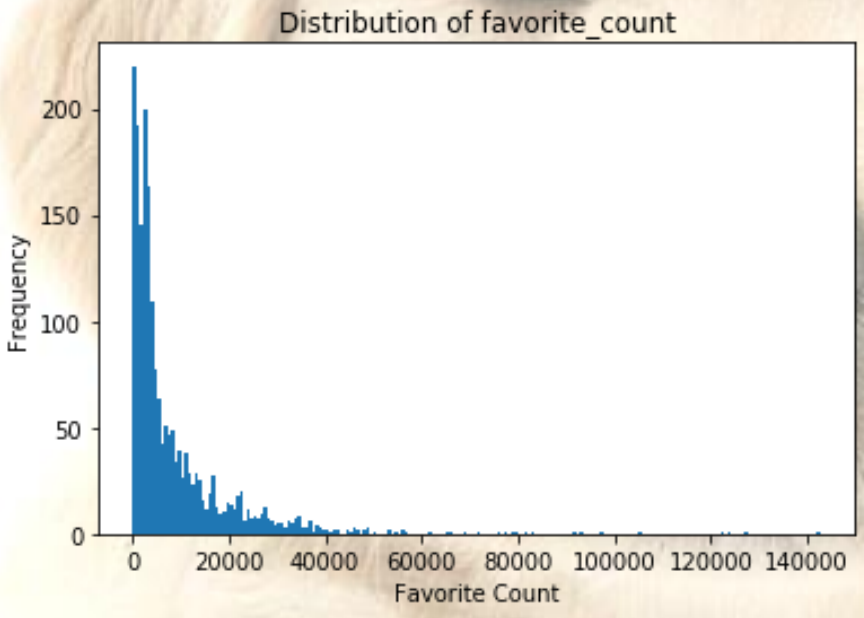
The bulk of rating numerators falls in the range of 7 - 14

Distribution of retweet count



The bulk of retweet counts fall below 10000 with just 83 tweets being retweeted more than 10000 times. The distribution is skewed to the right.

Distribution of favorite count



The bulk of favorite counts lies under 40000 with only 46 tweets having a favorite count above 10000. The distribution is skewed to the right.

Breeds that are most tweeted about based on top dog breed prediction (p1)

Since p1 is the top prediction of dog breed from the neural network let's look closer at that column and try to find some interesting insights

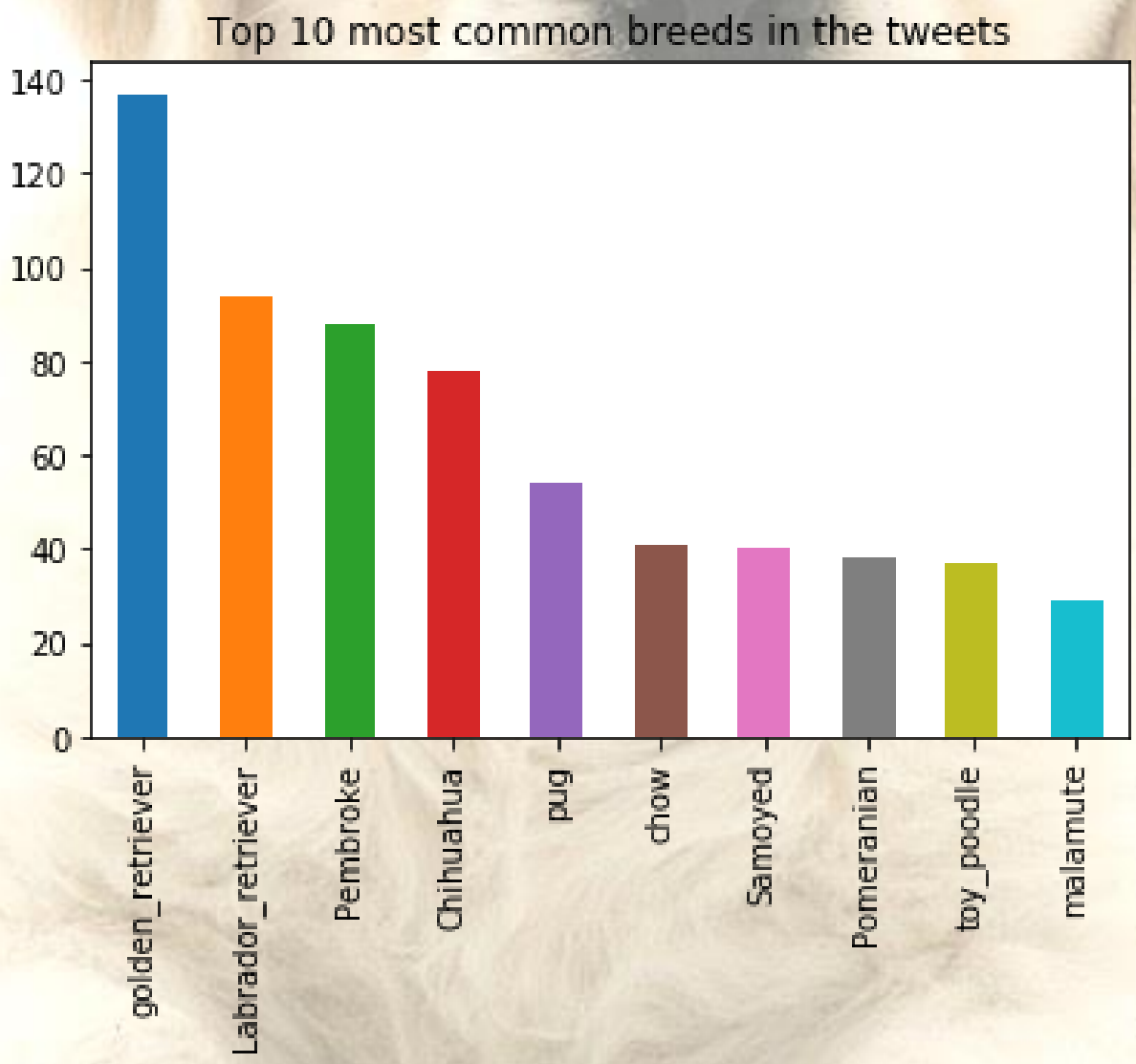
```
twitter_clean.p1_dog.value_counts()
```

```
True     1463
False     508
Name: p1_dog, dtype: int64
```

There are 508 predictions which are not dog breeds. Let's exclude them and focus on predictions that are actually dog breeds and find the top 10 most commonly occurring breeds

```
# Top 10 most commonly occurring dog breeds in p1 where p1_dog is TRUE
most_common = (twitter_clean.p1 [twitter_clean.p1_dog]).value_counts().sort_values(ascending= False)[:10]
most_common
```

```
golden_retriever      137
Labrador_retriever     94
Pembroke               88
Chihuahua              78
pug                    54
chow                   41
Samoyed                40
Pomeranian             38
toy_poodle             37
malamute               29
Name: p1, dtype: int64
```



Golden retriever and Labrador retriever which are supposed to be great family dogs and very friendly are two of the breeds that are most tweeted about!!

Most highly rated dog breeds

To find the most highly rated dog breeds let us subset the dataframe and take a look at the ratings less than 15 (exclude the 18 unlikely rating numerators that look like errors)

Grouping by the breed and taking the median of the rating numerator for each breed we get,

Most highly rated dog breeds

```
# Filter the dataframe and get only the breed and numerator rating (<15)

breed_rating = twitter_clean[['p1','p1_dog','rating_numerator']]
breed_rating = breed_rating [breed_rating.p1_dog]
breed_rating = breed_rating [breed_rating.rating_numerator < 15]
rated_breeds = breed_rating.groupby(by= ['p1'])['rating_numerator'].median().sort_values(ascending = False)[:10]
rated_breeds
```

```
p1
Afghan_hound             13.0
Saluki                   13.0
Great_Pyrenees           12.0
briard                   12.0
Tibetan_mastiff          12.0
Samoyed                  12.0
flat-coated_retriever    12.0
golden_retriever         12.0
Rottweiler               12.0
Pembroke                 12.0
Name: rating_numerator, dtype: float64
```

Afghan hound and Saluki are the top rated breeds. The Golden retriever which is the most commonly occurring breed in tweets is not in the top five but certainly features in the top 10 breeds in terms of rating! For a Golden retriever lover that's some heart warming insight 😊

Most retweeted dog breeds

To find the most highly retweeted dog breeds let us subset the dataframe and take a look at the retweet counts

Grouping by the breed, taking the median of the retweet count for each breed and sorting by descending we get,

Most highly retweeted dog breeds

```
br = twitter_clean[['p1','p1_dog','retweet_count']]
br = br [br.p1_dog]
br = br [~br.retweet_count.isnull()]
retweet_breeds = br.groupby(by= ['p1'])['retweet_count'].median().sort_values(ascending = False)[:20]
retweet_breeds
```

```
p1
Irish_water_spaniel         5839.0
Afghan_hound                5121.0
giant_schnauzer             4947.5
Saluki                      4057.5
black-and-tan_coonhound     3993.5
Irish_setter                3229.5
Australian_terrier          2967.5
Leonberg                    2878.0
wire-haired_fox_terrier     2822.5
Tibetan_mastiff             2815.5
flat-coated_retriever       2792.0
Cardigan                    2720.0
French_bulldog              2608.0
Samoyed                     2587.0
basset                      2585.0
Norwegian_elkhound          2529.0
Bedlington_terrier          2459.0
Weimaraner                  2452.0
kelpie                      2366.0
Border_terrier              2266.0
Name: retweet_count, dtype: float64
```

Breeds with the most favorite counts

To find the most favorite dog breeds favorite let us subset the dataframe and take a look at the favorite counts

Grouping by the breed, taking the median of the favorite count for each breed and sorting by descending we get,

Breeds with high favorite counts

```
bf = twitter_clean[['p1','p1_dog','favorite_count']]
bf = bf [bf.p1_dog]
bf = bf [~bf.favorite_count.isnull()]
favorite_breeds = bf.groupby(by= ['p1'])['favorite_count'].median().sort_values(ascending = False)[:20]
favorite_breeds
```

```
p1
Irish_water_spaniel        21266.0
Saluki                     20219.5
giant_schnauzer            16831.5
Afghan_hound               16831.0
black-and-tan_coonhound    16601.5
flat-coated_retriever      14700.5
Bedlington_terrier         13537.0
Border_terrier             13117.0
Norwegian_elkhound         12871.0
Leonberg                   11612.0
Australian_terrier         10854.5
French_bulldog             10824.0
Cardigan                   10332.0
Irish_setter                9948.0
Tibetan_mastiff             9684.0
kelpie                      9266.0
Weimaraner                  8863.5
basset                      8630.0
wire-haired_fox_terrier     8295.0
cocker_spaniel              8254.0
Name: favorite_count, dtype: float64
```

It's interesting to note that the top 5 breeds with the most retweets and the top breeds with the most favorite counts are almost the same ones! These seem like breeds that are liked by people and are popular though they aren't the most commonly occurring breeds in tweets.

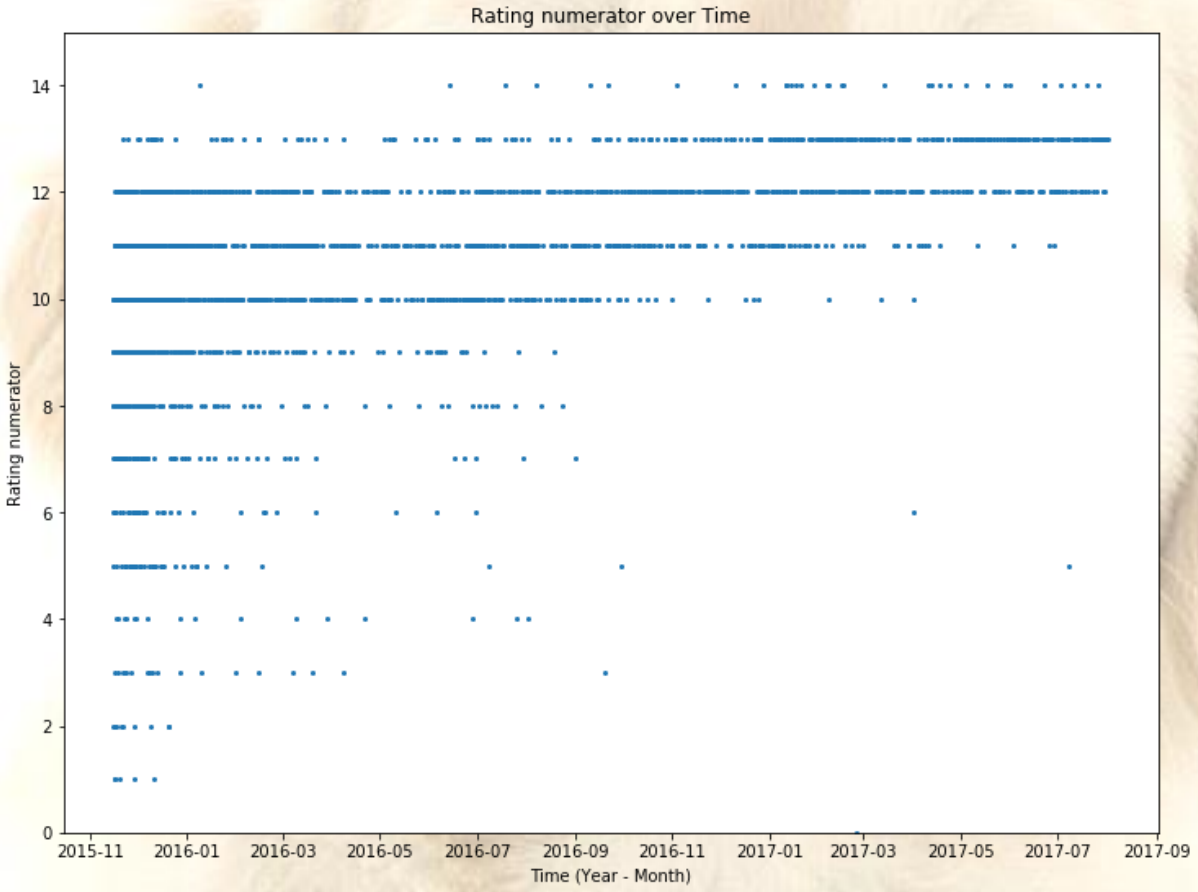
Now moving on to a different parameter for analysis:

The rating system and rating trends over time

We know that WeRateDogs follows a unique rating system ("They're good dogs!") where it's acceptable for the numerator to have values greater than 10. https://knowyourmeme.com/memes/theyre-good-dogs-brent

This piqued my interest to look closely at the ratings

This plot shows the rating numerators over time:

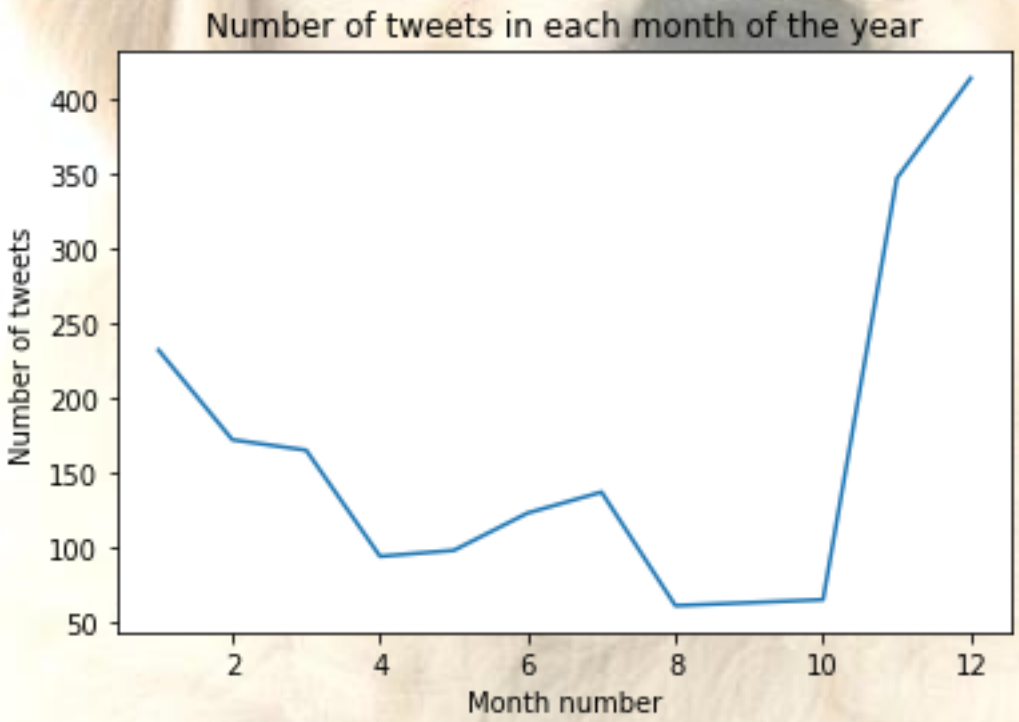


After the incident that made ratings above 10 popular in September 2016, it has become a trend to rate dogs above 10. There are hardly any rating numerators below 10 after 2016 which is interesting. The rise in popularity of the incident and rating system is evident from the visualization.

Number of tweets and time of the year

The graph shows tweet counts grouped by month for 2015, 2016 and 2017.

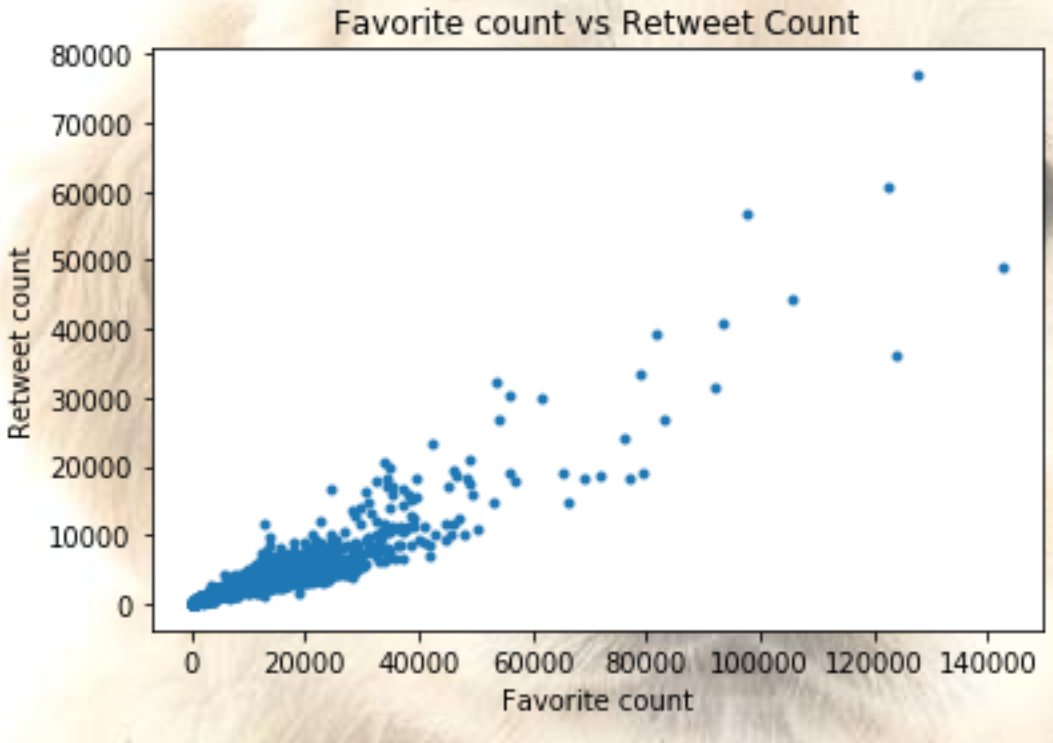
```
twitter_clean['year']=twitter_clean['timestamp'].dt.year
twitter_clean['month']=twitter_clean['timestamp'].dt.month

tweets_time= twitter_clean.groupby(['month'])['tweet_id'].count().plot()

plt.title('Number of tweets in each month of the year');
plt.xlabel('Month number');
plt.ylabel('Number of tweets');
```



The holiday season - Thanksgiving, Christmas , New year seems to have the most number of tweets!

<u>Are the favorite count and retweet counts related??</u>

Plotting the favorite counts and retweet counts on a scatter plot,



Favorite count and retweet count seem positively correlated from the scatterplot. Let's look at the correlation coefficient

```
counts=twitter_clean[['retweet_count','favorite_count']]
counts.corr()
```

	retweet_count	favorite_count
retweet_count	1.000000	0.919902
favorite_count	0.919902	1.000000

A correlation coefficient of 0.91 indicates a pretty strong positive relationship between favorite count and retweet count. It supports the intuition that if people 'like' a tweet they tend to retweet it.

Resources:

https://www.slickremix.com/docs/how-to-get-api-keys-and-tokens-for-twitter

https://stackoverflow.com/questions/28384588/twitter-api-get-tweets-with-specific-id

http://stackabuse.com/reading-and-writing-json-to-a-file-in-python/
https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object

https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html