# **Twitter Data Analysis and Visualization Report**

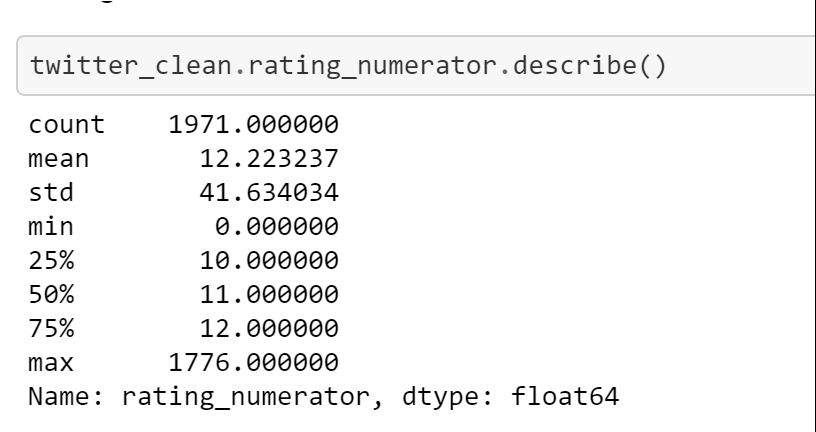
**This report is to communicate the insights and display visualization(s) produced from the wrangled data from WeRateDogs twitter archive.**

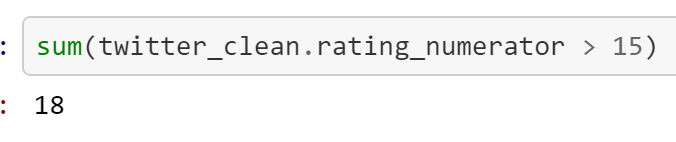
**Retweets and replies have been excluded from the dataset. Only original ratings with images in the tweets have been included. The data wrangling report details the gathering, assessing and cleaning steps performed to prepare the data for the following analysis.**

**Analysis of distributions for some parameters of interest**

**Distribution of rating numerator**

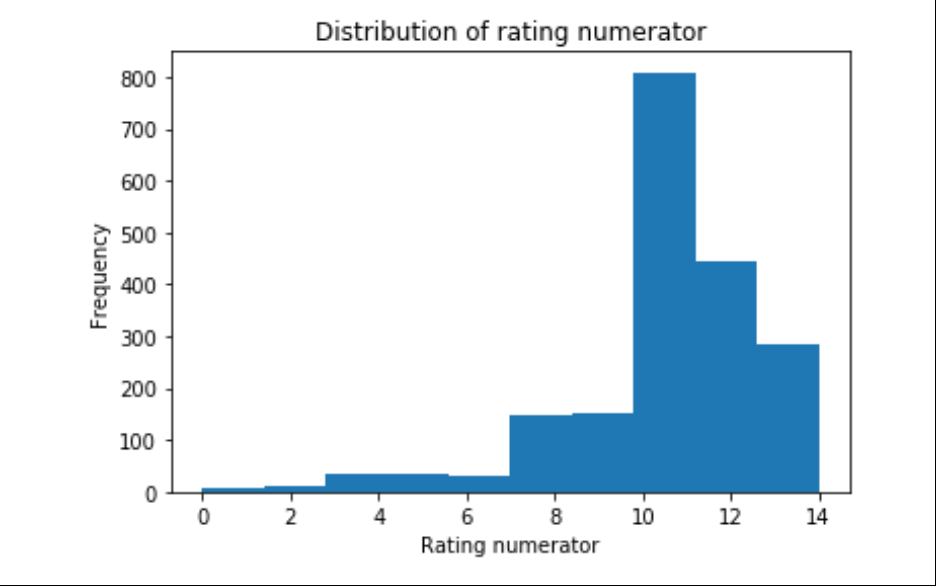
Using describe() shows us that the max value is something unlikely for the numerator





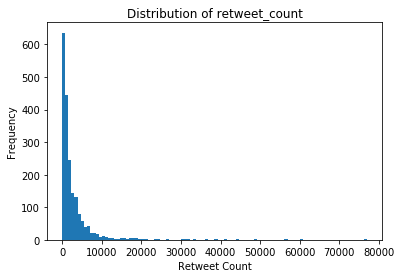
There are 18 entries where the numerator is greater than 15. So I decided to exclude them from the analysis.

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



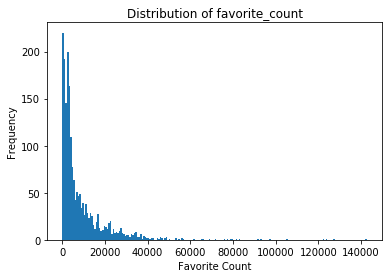
The bulk of rating numerators falls in the range of 7 - 14

**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.

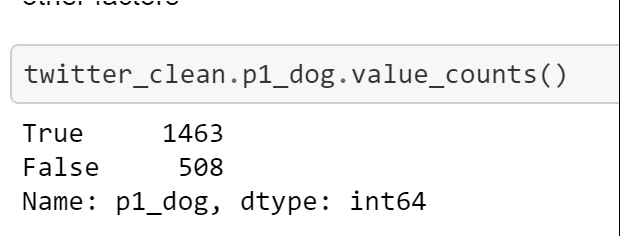
**Distribution of favorite count**



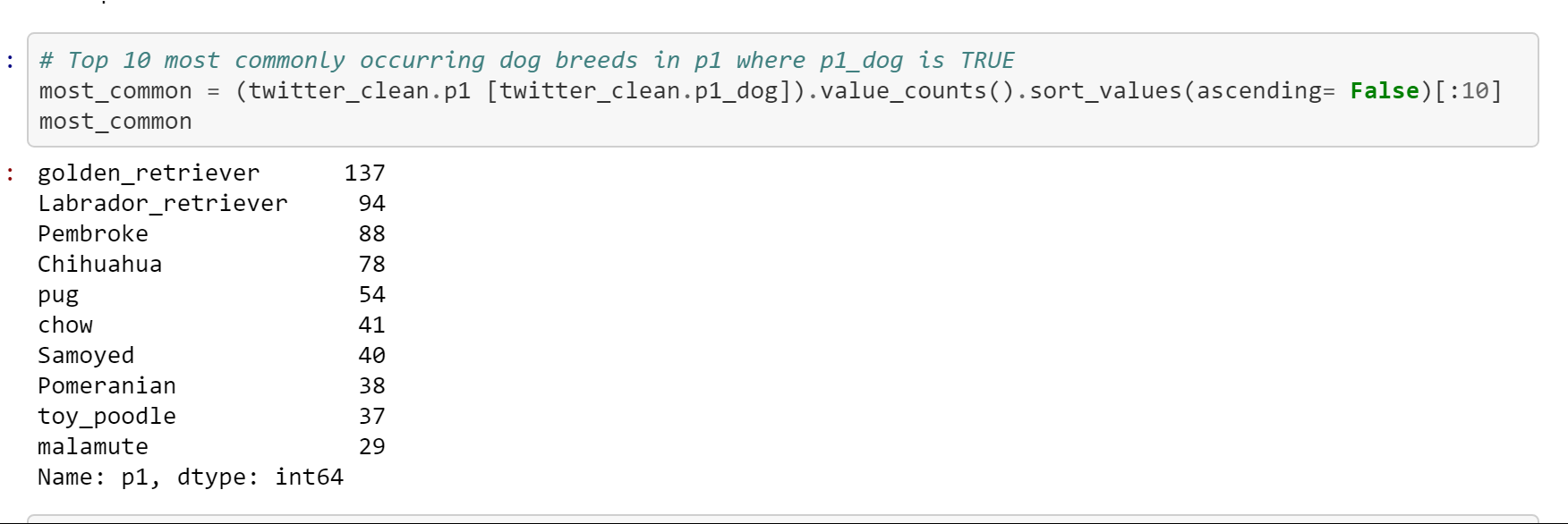
The bulk of favorite counts lies under 40000 with only 46 tweets having a favorite count above 10000. The distribution is skewed to the right.

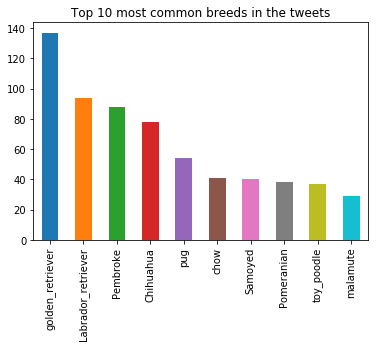
**Breeds that are most tweeted about based on top dog breed prediction (p1)**

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



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



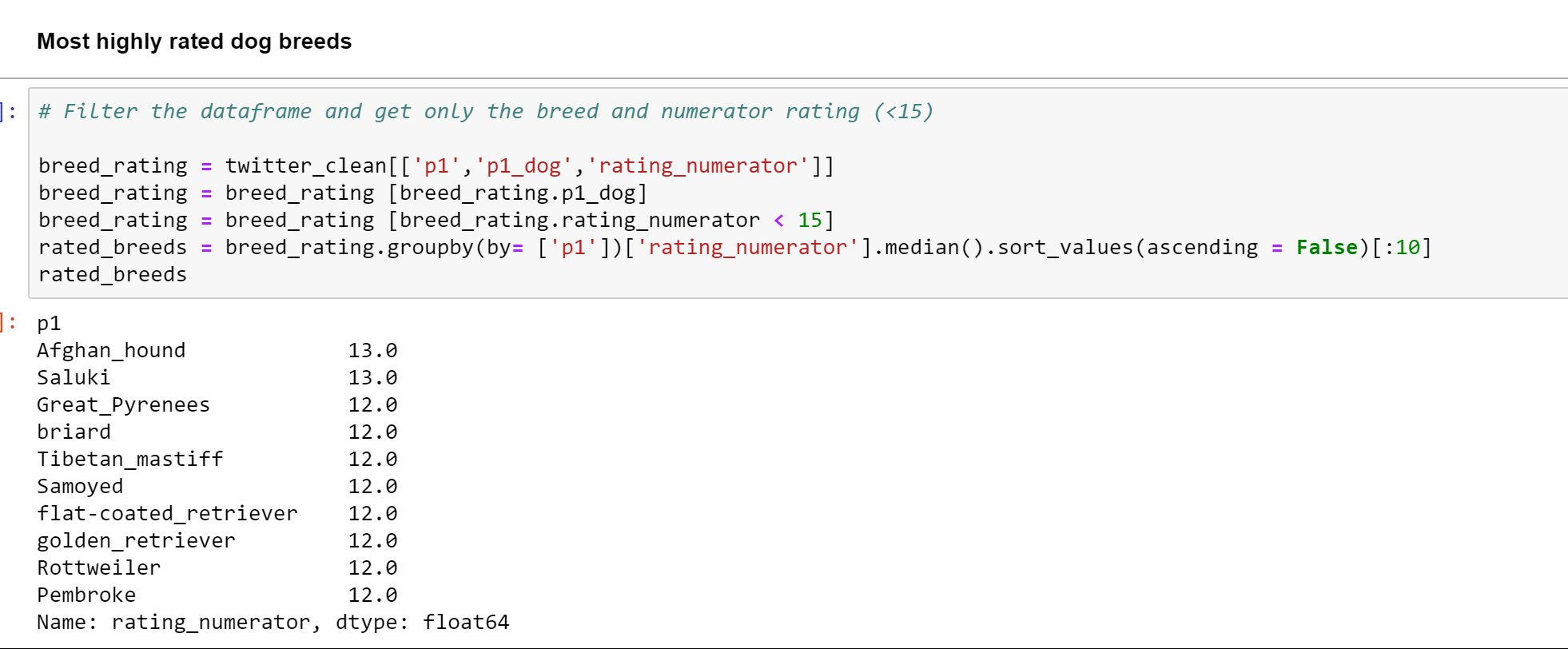


**Golden retriever and Labrador retriever which are supposed to be great family dogs and very friendly are two of the breeds that are most tweeted about!!**

**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,



**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

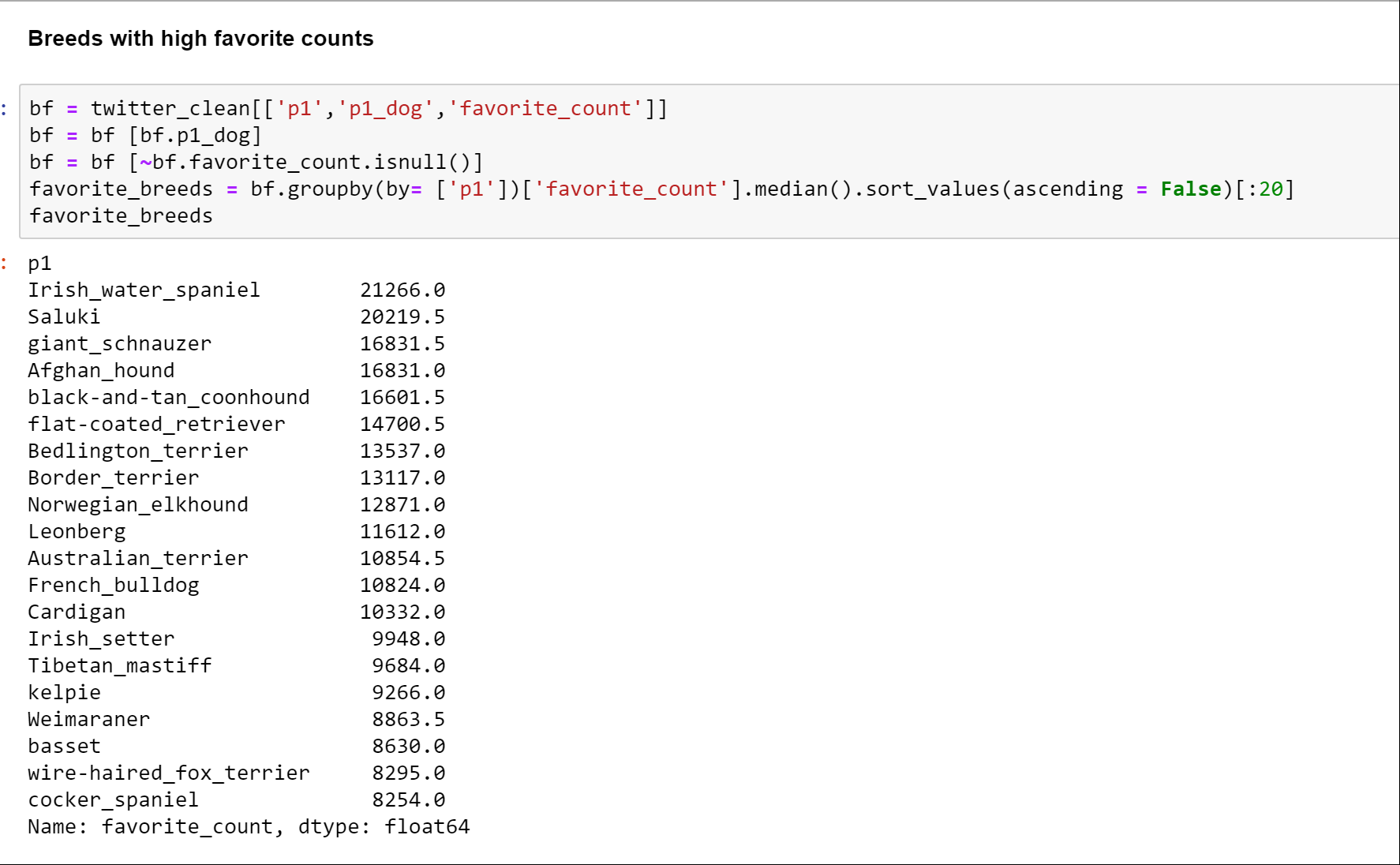
Grouping by the breed, taking the median of the retweet count for each breed and sorting by descending we get,



**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,



**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.**

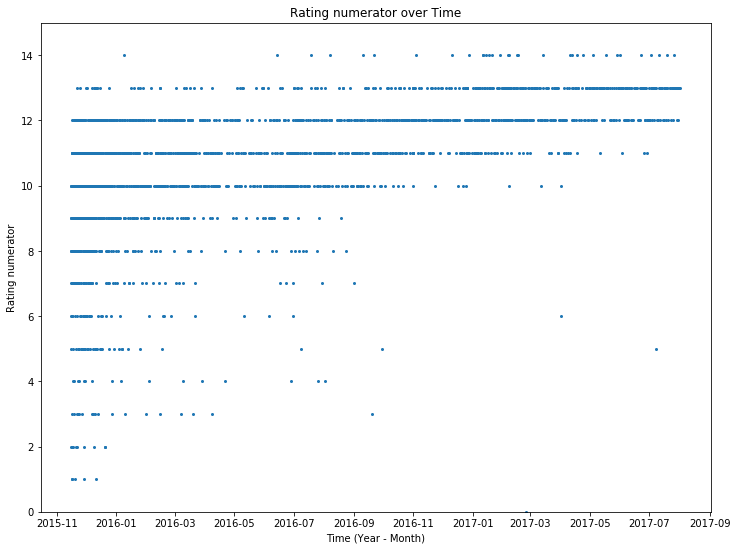
Now moving on to a different parameter for analysis:

**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**](https://knowyourmeme.com/memes/theyre-good-dogs-brent)

**This piqued my interest to look closely at the ratings**

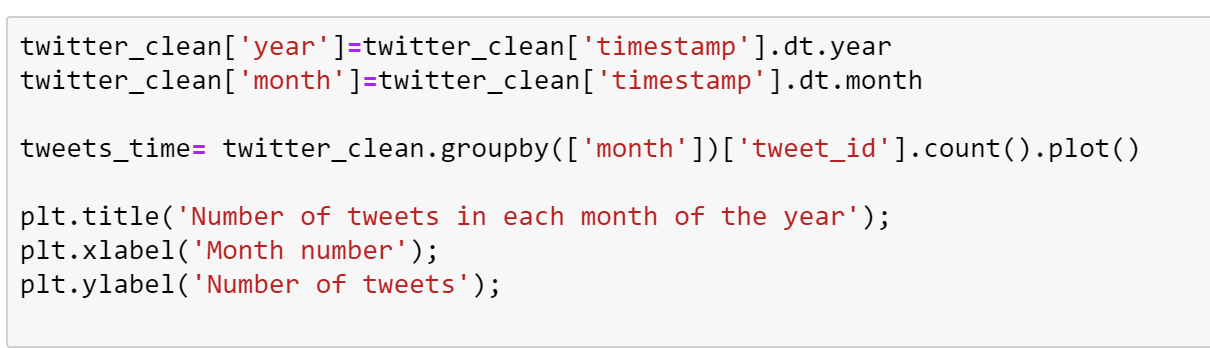
This plot shows the rating numerators over time:

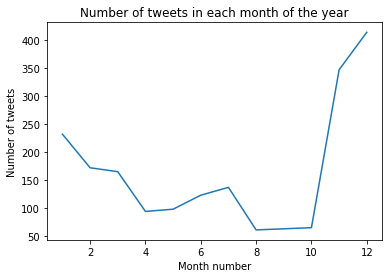
****

**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**.**

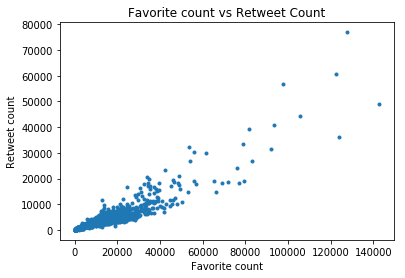


****

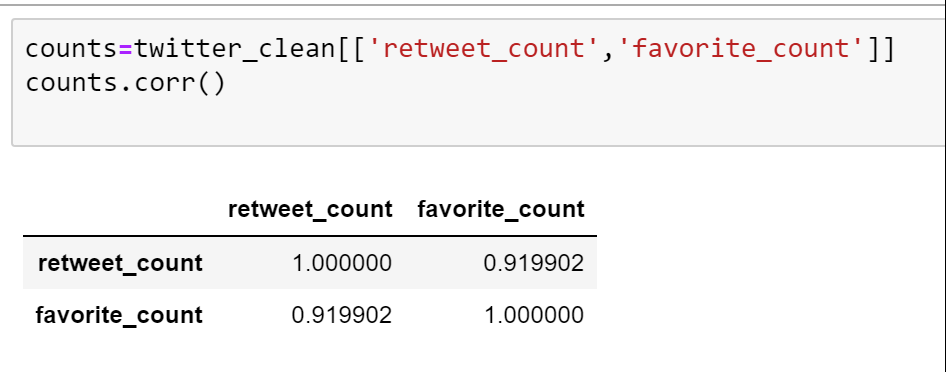
**The holiday season - Thanksgiving, Christmas , New year seems to have the most number of tweets!**

**Are the favorite count and retweet counts related??**

**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



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

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

[**https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html**](https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html)