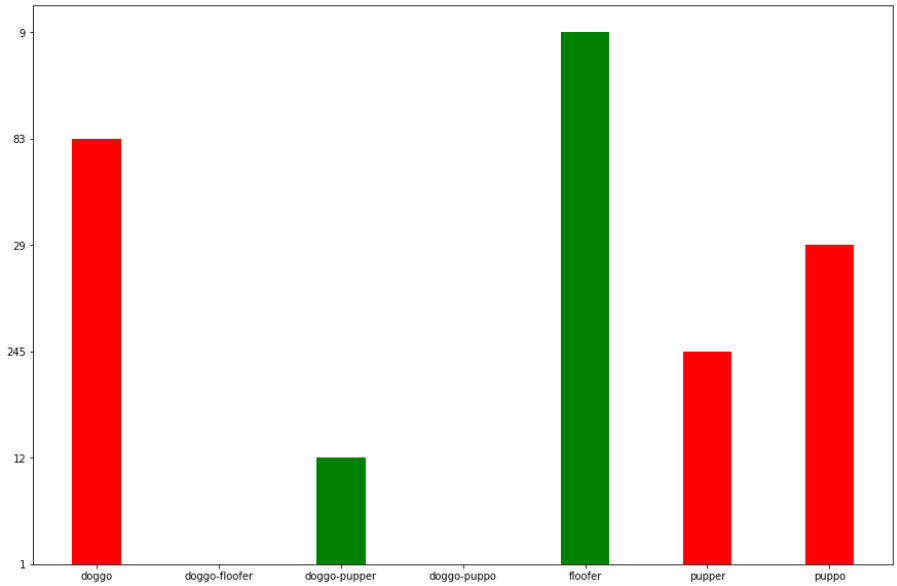
Look at this amazing dog from **WeRateDogs** Twitter account.

I have an archive file contained many tweets like this pretty dog tweet

The most common **breads** founded were floofer, doggo then puppo like the below chart shows.



When we analyze the **rating_numerator** column we find that most values are located around '13' which the mean value.

And the standard deviation is about 45

The minimum value given **for rating_numerator** was 0

The **rating_denominator** mean values are located around 10.

Another issue was the data type for both **numerator, denominator** which was an integer
We change it to float to allow any person to rate with any decimal values.

And the standard deviation is about 6.7 which is a reasonable value and better than the numerator standard deviation.

When looking at the name of dogs

We find the most common **name**s are:

Charlie 12

Oliver 11

Cooper 11

Lucy 11

Tucker 10

Lola 10

Penny 10

Bo 9

Winston 9

Sadie 8

And about 745 names lost or undefined

Tweet_id data type is abject in order not to be calculated when it defined as an integer

the **timestamp** should be defined as datetime64 so we can extract information about definite tweet time from it

we can find that expanded_urls have 59 null values using isna() and sum () functions

that we replace it with ("").

in the **image prediction table**, the names of the columns were abbreviated so we can't understand the meaning of columns names
we change names as follow

['tweet_id', 'jpg_url', 'img_num',

'prediction_1', 'confidence_1', 'breed_1',

'prediction_2', 'confidence_2', 'breed_2',

'prediction_3', 'confidence_3', 'breed_3']