Act Report

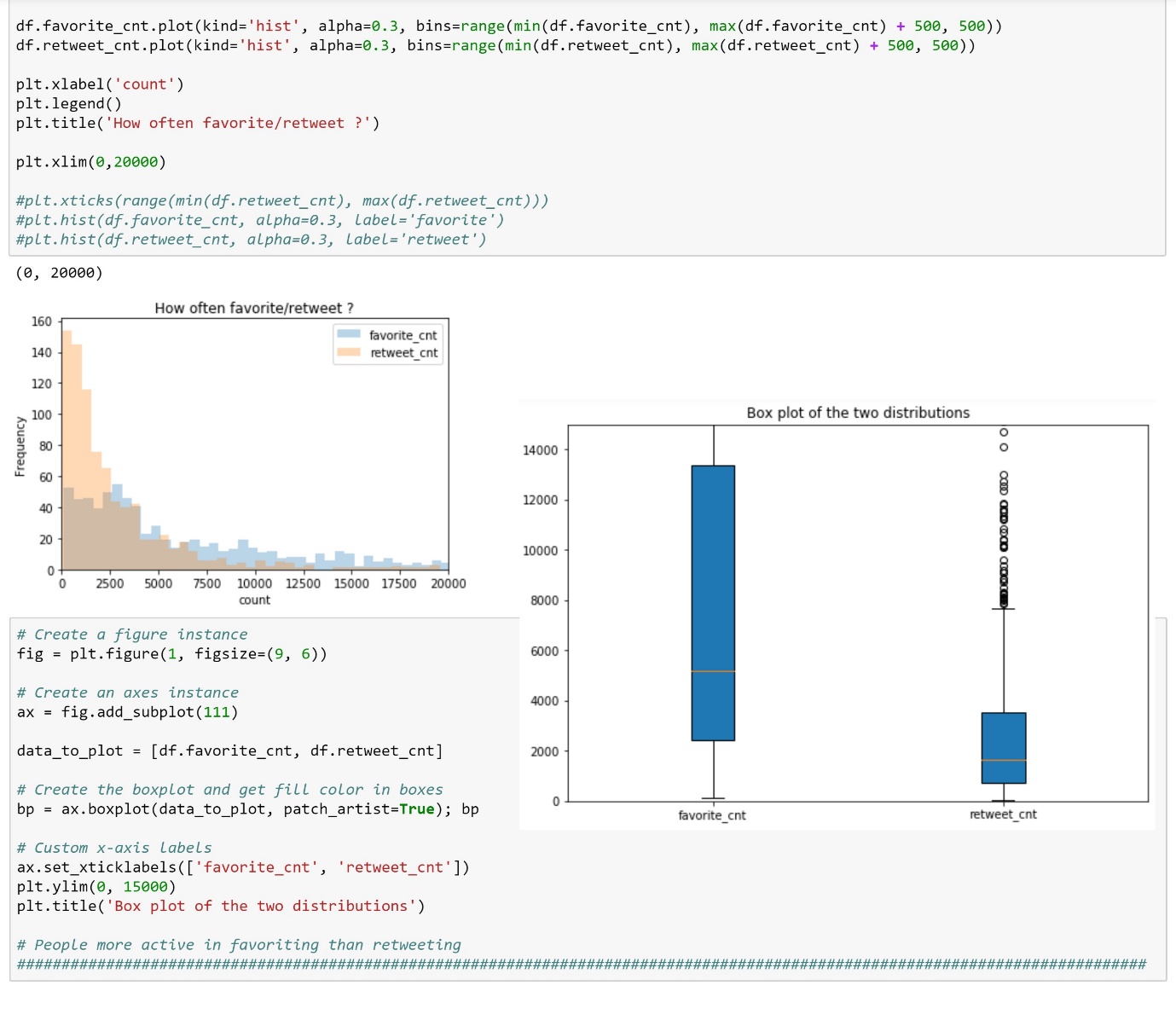
**Analyzing & Visualizing**

**Note: I eliminated all rows which have nothing to do with dog (Based on the probability of ‘Dog-True’ greater than 0.6).**

Q. favorite VS retweet

>Insight\_01: people do favorite more often than retweet.

- People are more active in doing favorite than retweet. This becomes clear when ploting a histogram to compare the frequencies of favorite count and retweet count. For example, in the shape of each distribution, within the count range from 0 to 2500, the frequency of retweet counts significantly outnumber that of favorite counts, then when the range goes beyond 2500, the frequency of favorite counts keep exceeding its counterpart. This trend can be reaffirmed through the boxplots as well. The major Interquartile range of the favorite count relatively spread over the full frequency intervals while that of retweet concentrates among a few intervals which registers relatively lower frequencies.

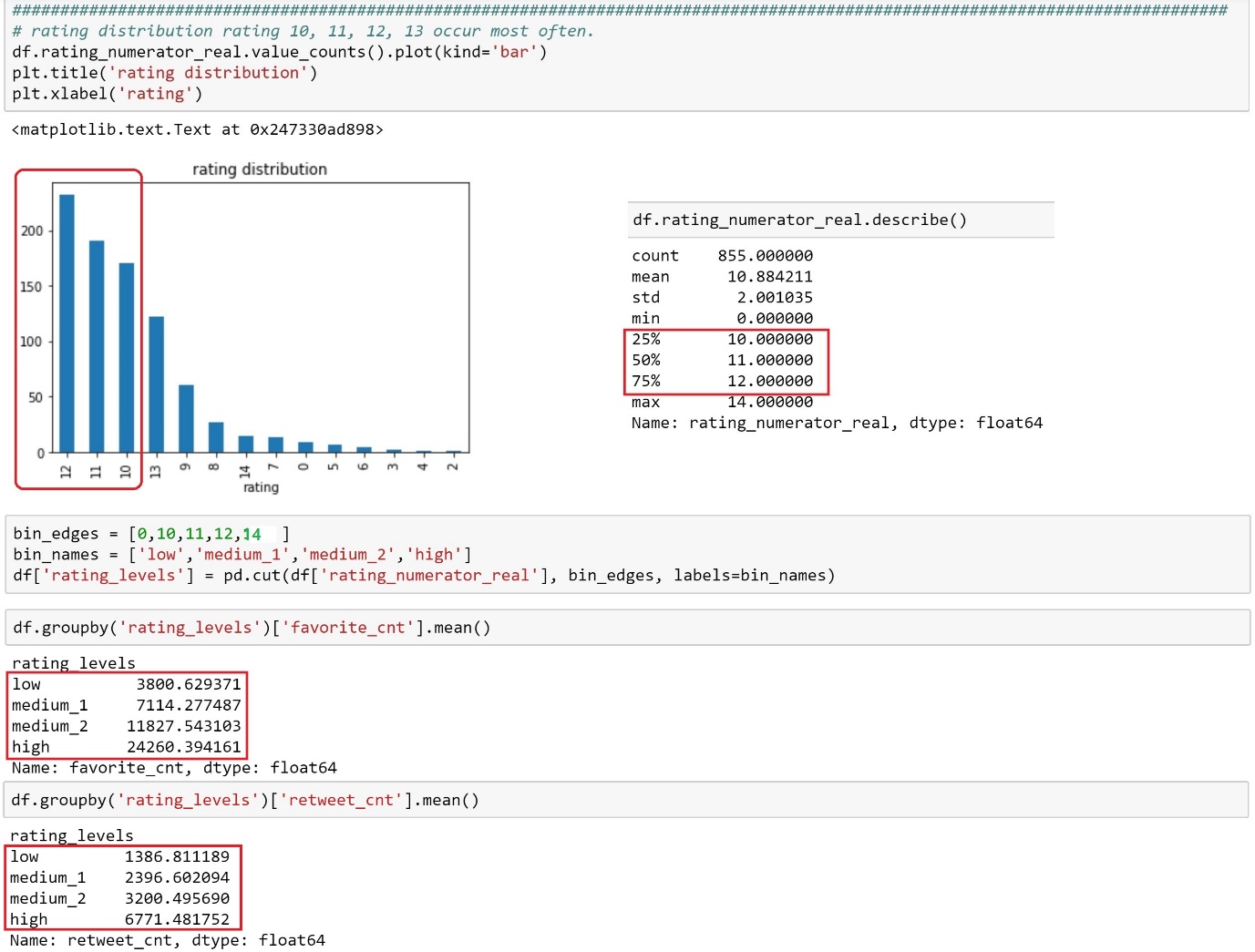


Q. Rating VS favorite & retweet

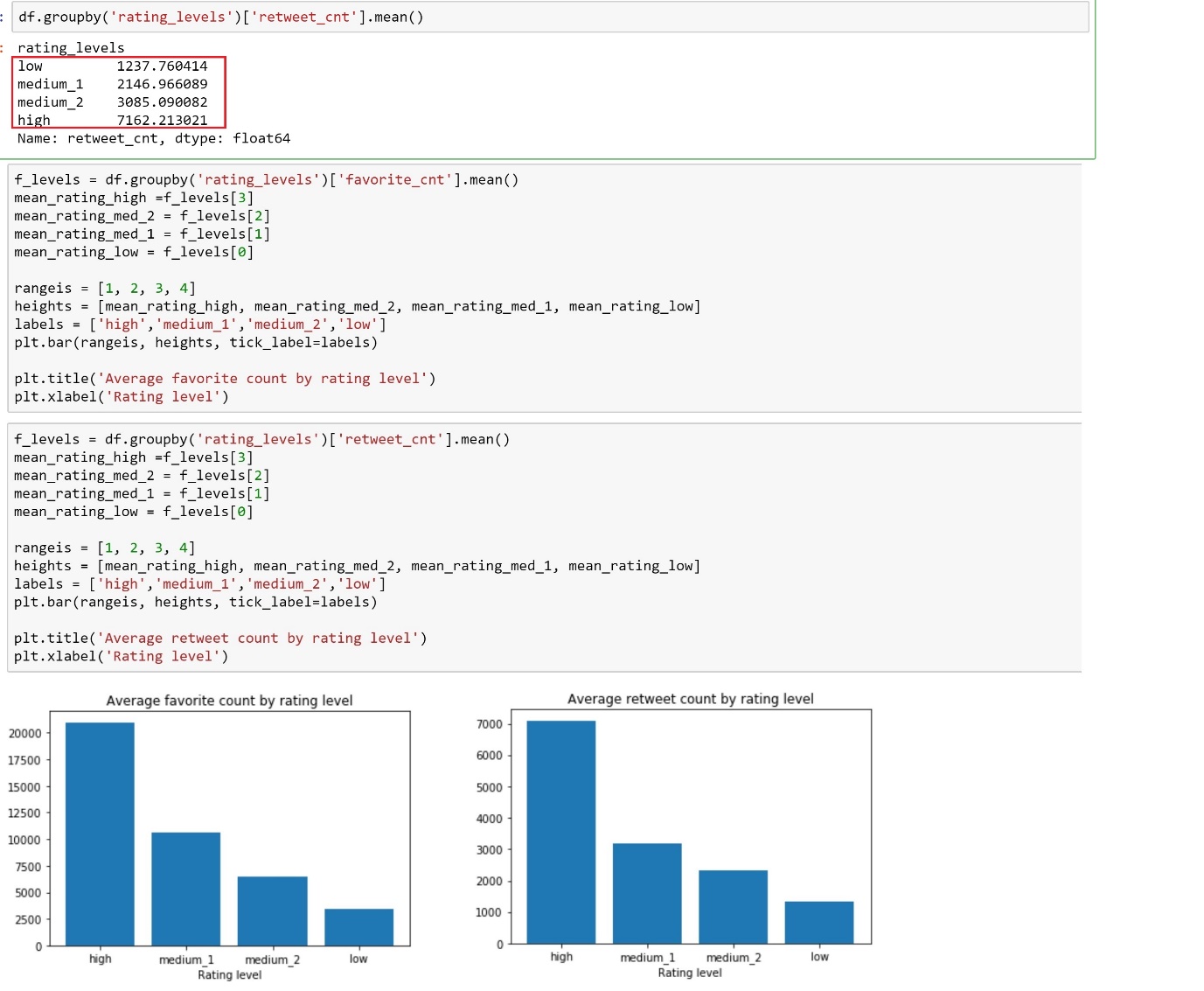
>Insight\_02: they have a positive relationship.

- Obviously, there are some irrelevant tweets that have nothing to do with dogs. And what do ‘favorite & retweet counts’ response to this? We expected that rating-behaviour also comes along with ‘favorite & retweet counts’ by their nature. But it is interesting to note that some higher ratings are seemingly somewhat related to such funny, irrelevant, abnormal tweets disassociated with dogs. In our analysis, we rule out such records that have nothing to do with dogs. We want to see if there are any negative relationships between ‘ratings’ and ‘favorite, retweet counts’. Let’s investigate their relationships.

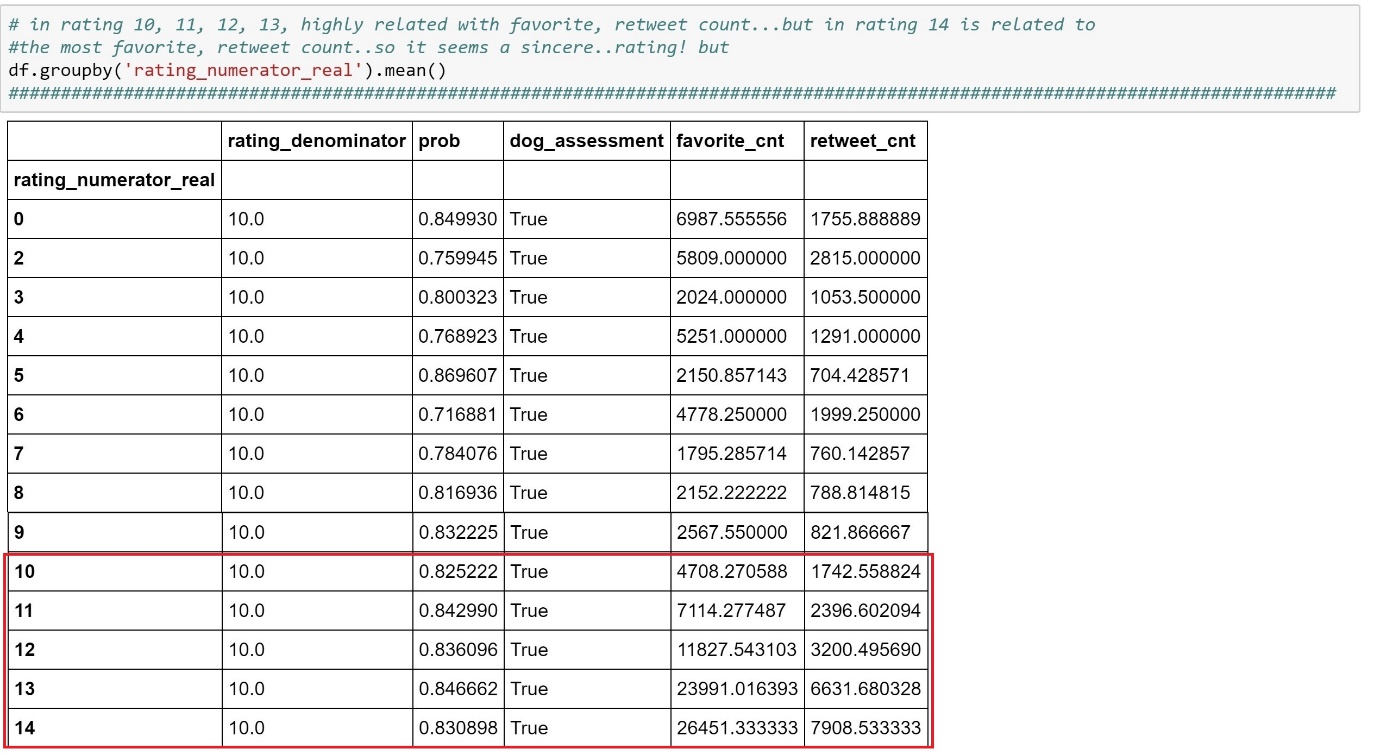
Here, we can notice that some absurd ratings or outliers(such as 1776/10, 420/10, 182/10, 666/10, etc) are disappeared when we got rid of ‘Dog-False’ data. From the distribution of ratings, rating 10, 11, 12 occur most often. Creating 4 levels based on this, we can see there is a positive relation going on between ratings VS favorite, retweet count.



So, this means when a certain tweet receives higher ratings, it also tends to get higher favorite & retweet count and vice versa.



Having seen previously, the ratings 10, 11, 12, 13 constitute the major interquartile range of all ratings…so I’d say they are reliable data. In this case we should take the ratings 10,11,12,13 into more account when we investigate their relationship with ‘favorite & retweet counts’.

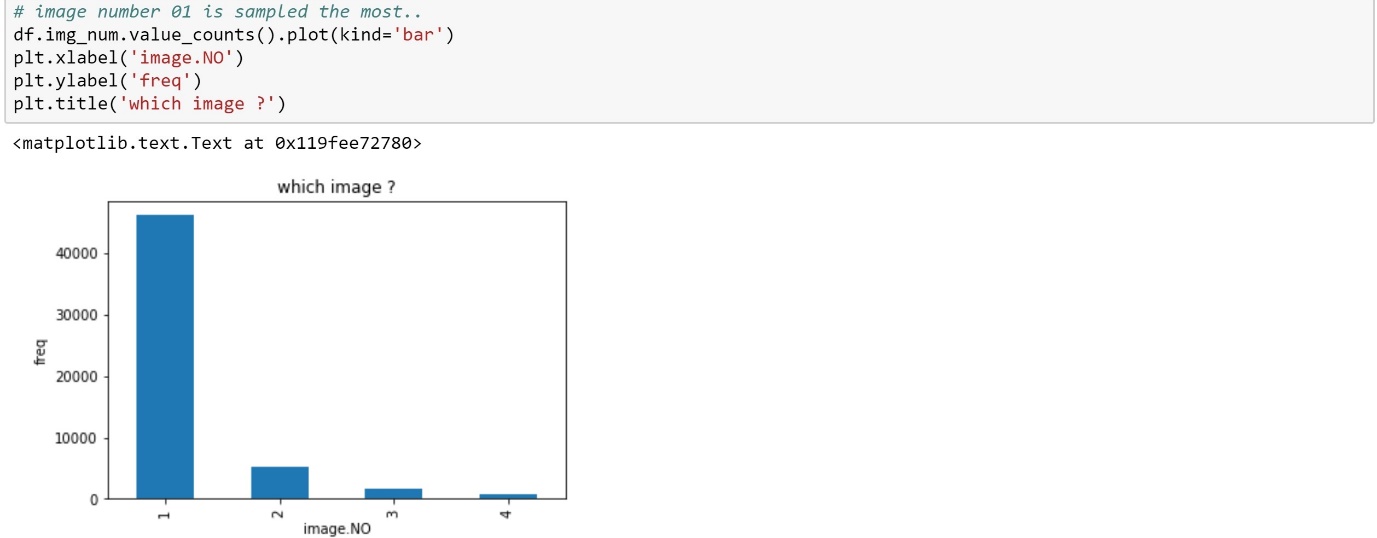


Here, we can clearly see that the positive relationship between ‘ratings’ and ‘favorite&retweet counts’ is significant within a range of ‘ratings’ from 10 to 14, so we concluded ‘ratings’ tend to come along with ‘favorite & retweet counts’.

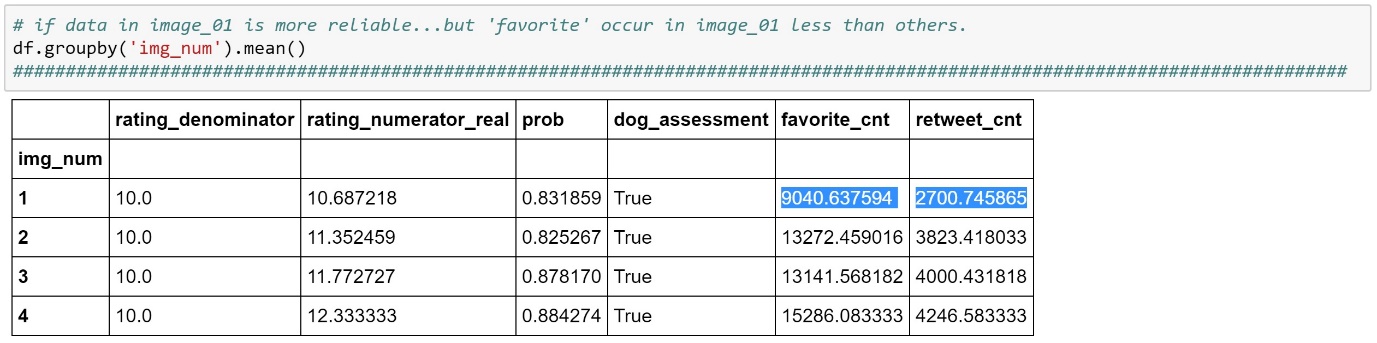
Q. Image + Pet\_status VS favorite & retweet

>Insight\_03: If you want to get more favorite, retweet counts, don’t tweet ‘pupper’ and tweet ‘floofer !

- The ‘image\_01’ has the largest samples. We can expect that of all images, ‘image\_01’ would yield the most reliable result.

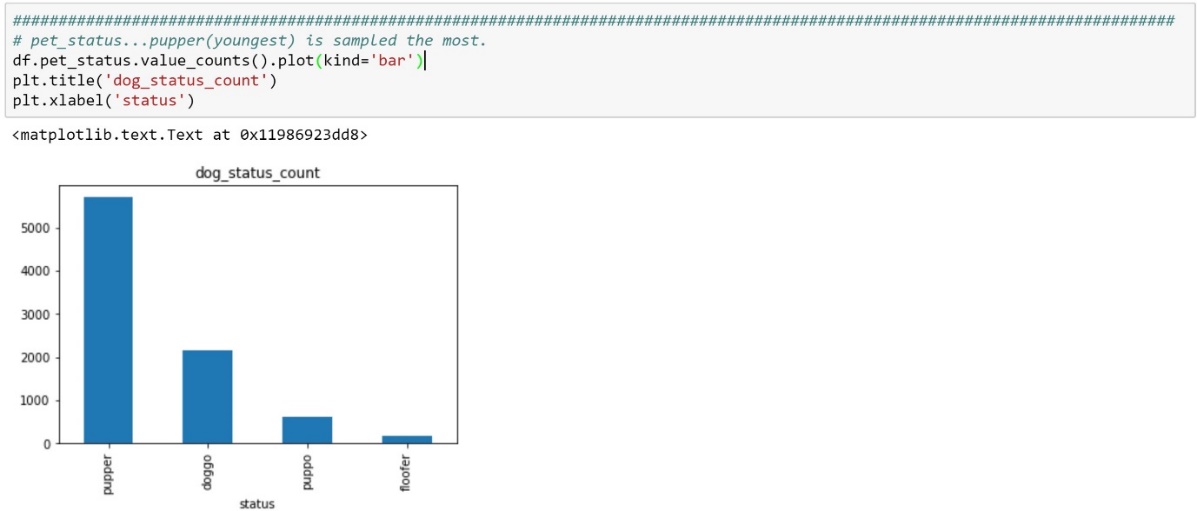


But, if this is the case, why does the mean 'favorite & retweet counts' in image\_01 show smaller value than others ? Can we blame this on outliers ?

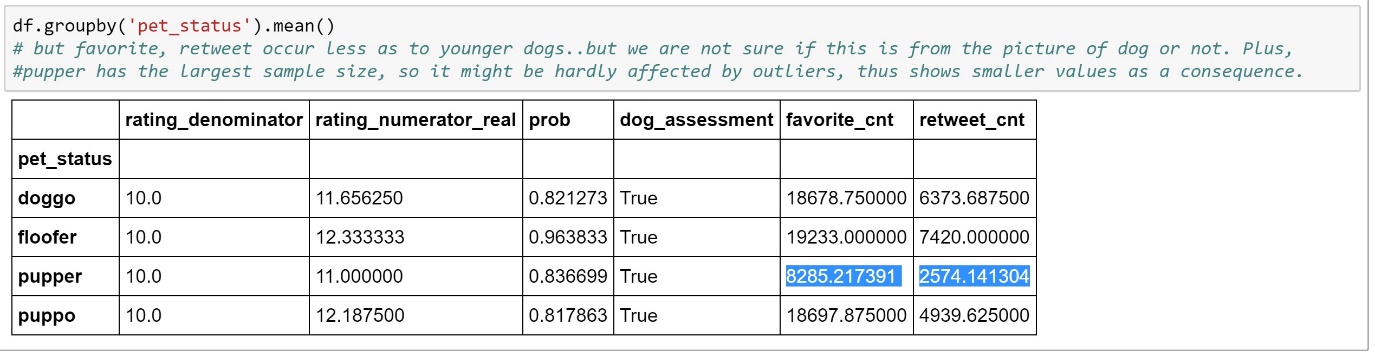


- Let’s see the pet\_status.

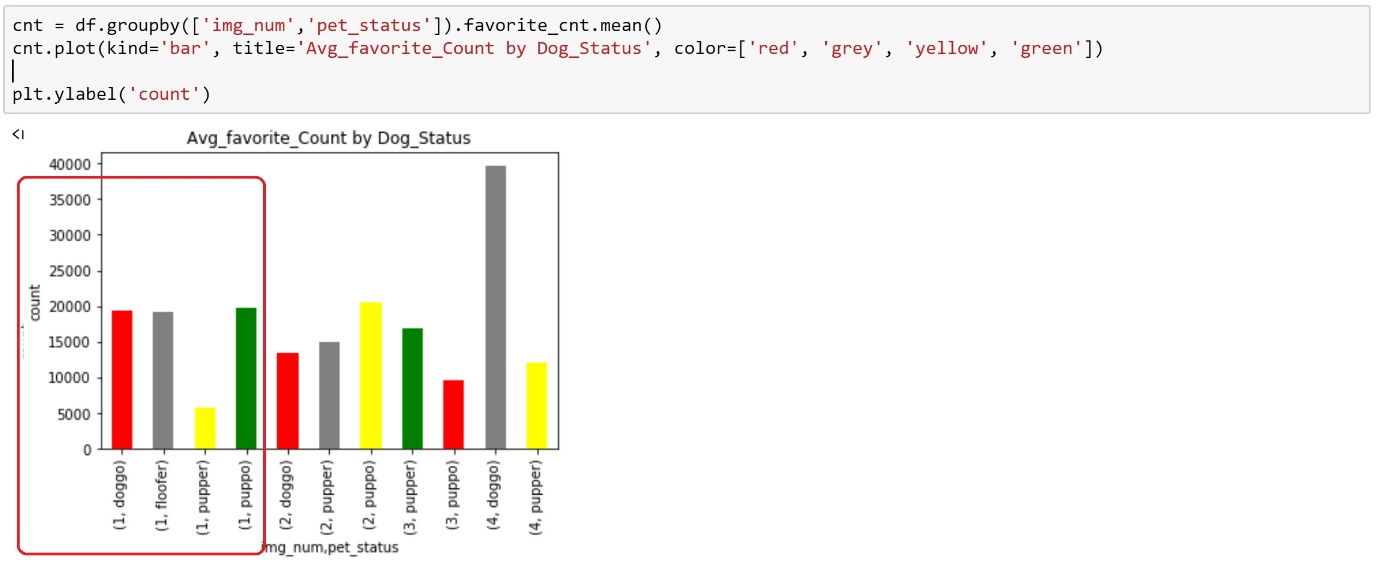
The ‘pupper’ has the largest samples. We can expect that of all pet\_status, ‘pupper’ would yield the most reliable results.



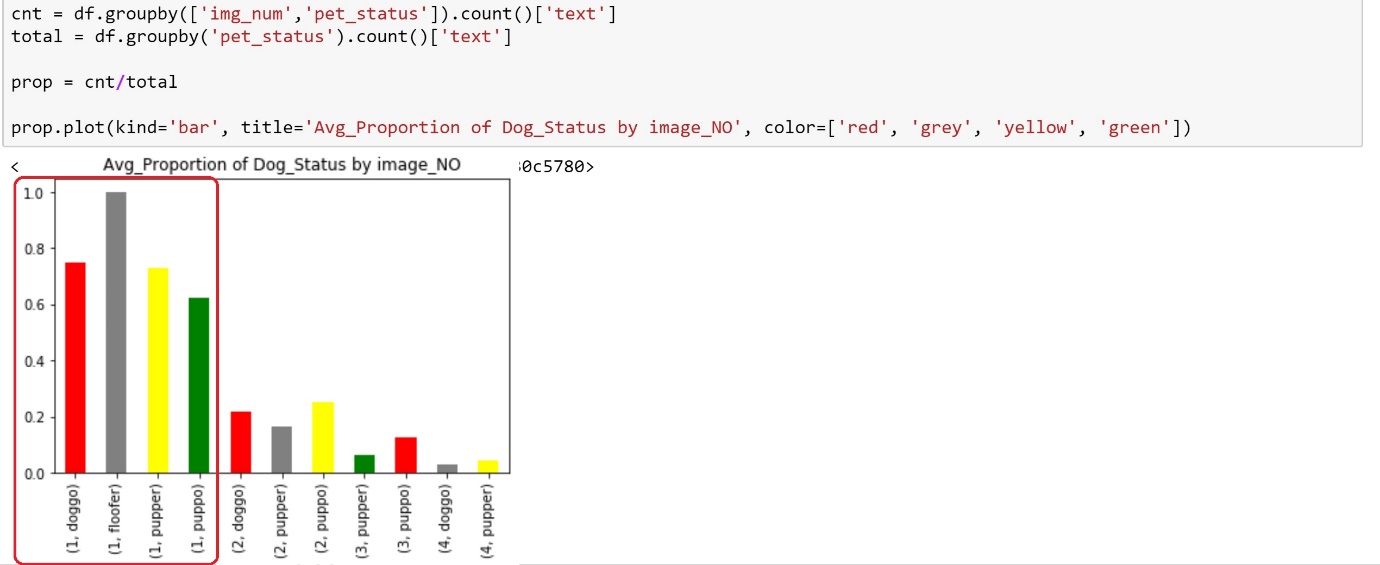
But, if this is the case, why does the mean 'favorite & retweet counts' in ‘pupper’ show smaller value than others ? Can we blame this on outliers ?



Now we tried to plot the mean ‘favorite counts’ grouping by ‘image and pet\_status’. We believe only the result on image\_01 can be reliable.



There are clearly more samples on ‘pupper’ and on ‘image\_01’ than other categories. So, it's hard to make a fair comparison, using out plot above. To balance this out, we tried to divide each count by the total count for that 'pet\_status' in order to use "proportions" instead.



We believe the largest sample size is translated into better reliability; therefore, can conclude that the tweets associated with ‘pupper’ is the least popular(the largest size of their samples underpin this). We can, accordingly, assume that tweets of ‘floofer’ would get the most frequent high ‘favorite, retweet counts’.