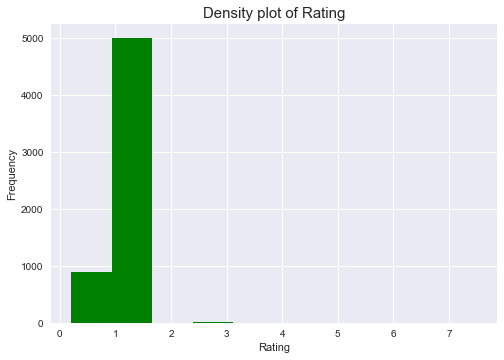
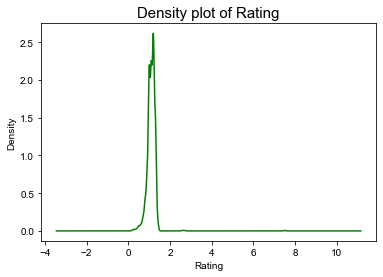
**Exploratory Data Analysis Documentation**



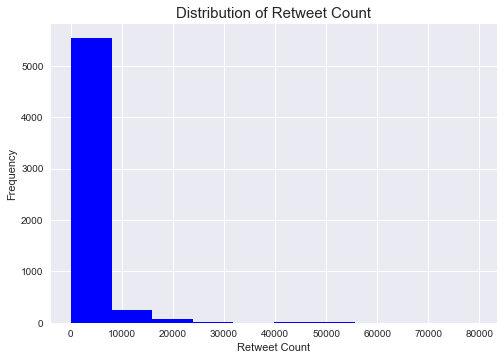
**Question 1:** What is the distribution of some Quantitative Features



From the histogram above, we see that ratings between 1.20 to almost 1.25 occur the most, while ratings greater than 2 but not up to 3.5 seems to be an outlier with almost 0 occurence

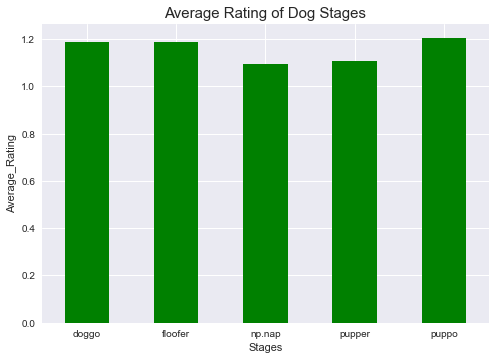


This is a Density plot in order to buttress the histogram of rating feature above in a curve pattern.



We see from the histogram above that it is Right Skewed which means that the mean of Retweet count is greater than its median.

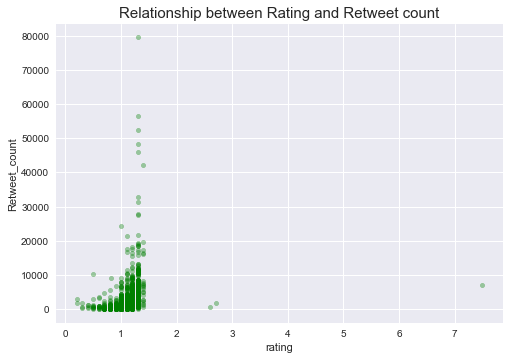
# Question 2: Which Dog stage has the highest average rating

****

The above bar graph shows that Puppo with an average rating of

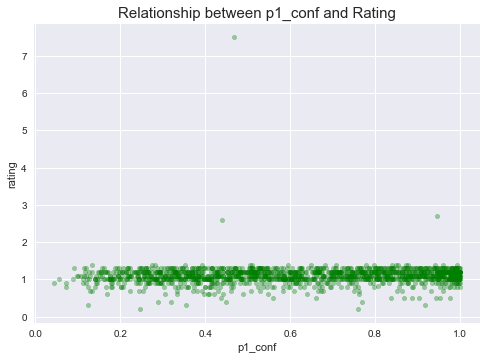
1.2 has the highest average rating.

# Question 3(What is the relationship between Rating and Retweet Count?)

****

From the scatter plot above, there in no particular negative nor positive relationship between both features. Hence, I notice that there are some extreme values which are obviously outliers.

**Question 4: (What is the relationship between p1\_conf and rating)**



**Conclusion:**

I have successfully cleaned the datasets, i have combined all the data sets into one based on common column "tweet\_id".

From my analysis, I can see that Pupp0 seems to have higher average rating.

The mean of Retweet count is greater than the median.

0 to 2 has the highest rating occurrence.

**Limitations:**

I discovered that some columns had wrong data types so I had to deal with that, I noticed that certain features have to be dropped and some has to be renamed so I also dealt with that.

I also saw that it is better to work with situations where our p1 are all true so I dropped all the false cases or else we will be having wrong analysis.

I also noticed some tidy and quality issues which I had to deal with.

References:

Udacity lessons, Python documentation, Stack overflow