Analysis Report

Byron Kilian

(1) Library used in the analysis

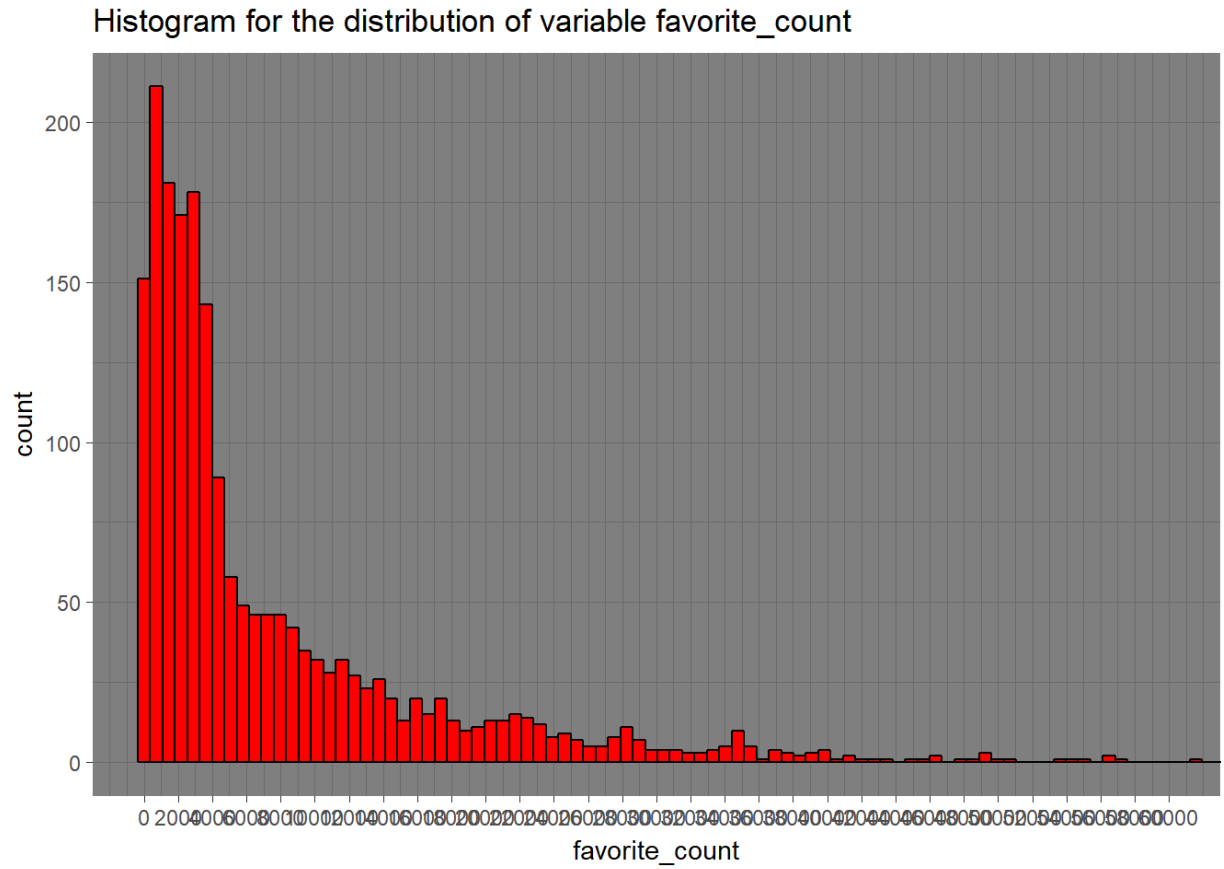
The library I used for creating the visuals is ggplot2 in R.

(2) Data used in the analysis

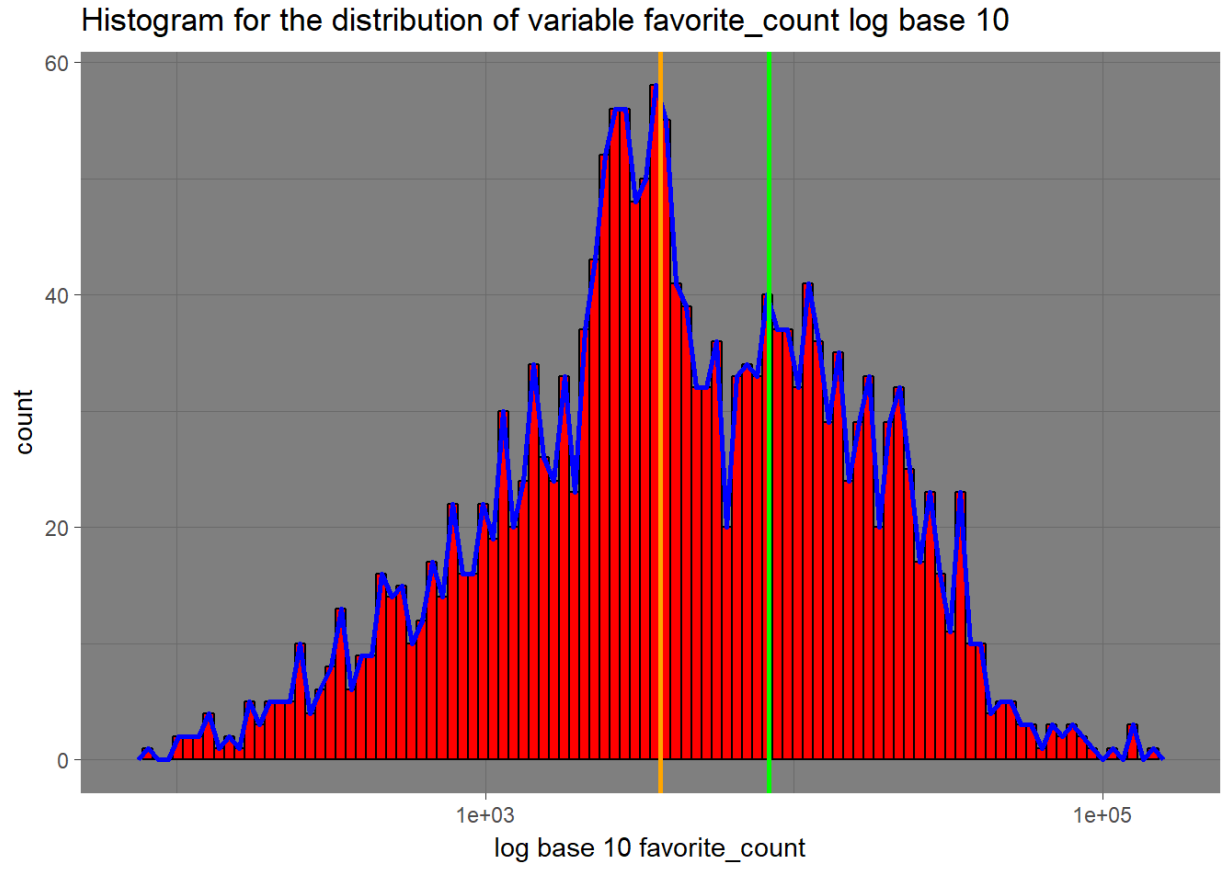
After the gathering, assessing and cleaning phase of the project I created a master view in SQL of which the content I wrote to a csv file. This master file was then read into R using the ‘read.csv’ command.

(3) Univariate Plots

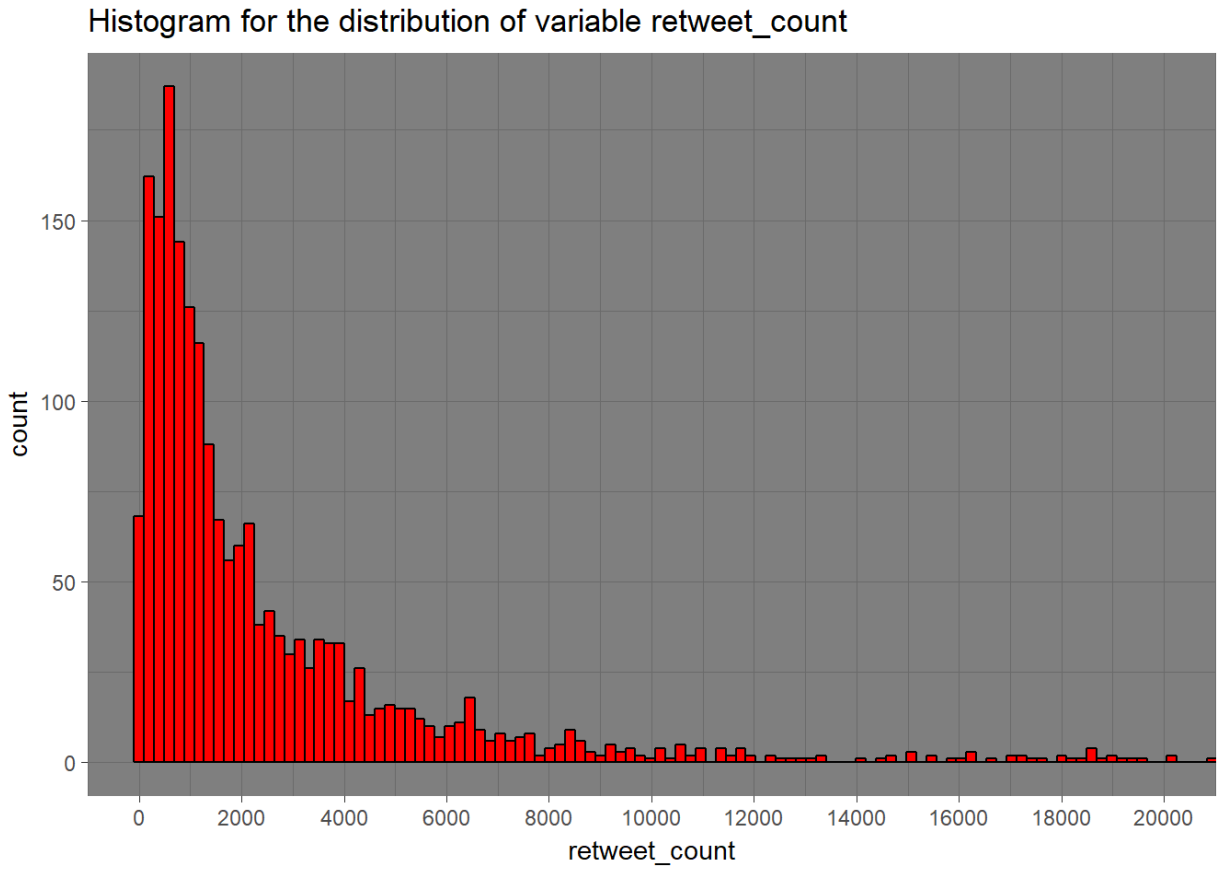
For the investigation of the data I first start by viewing univariate plots to get an understanding of the distribution for some of the variables. The first variable I looked at was the histrogram for the favorite_count variable.



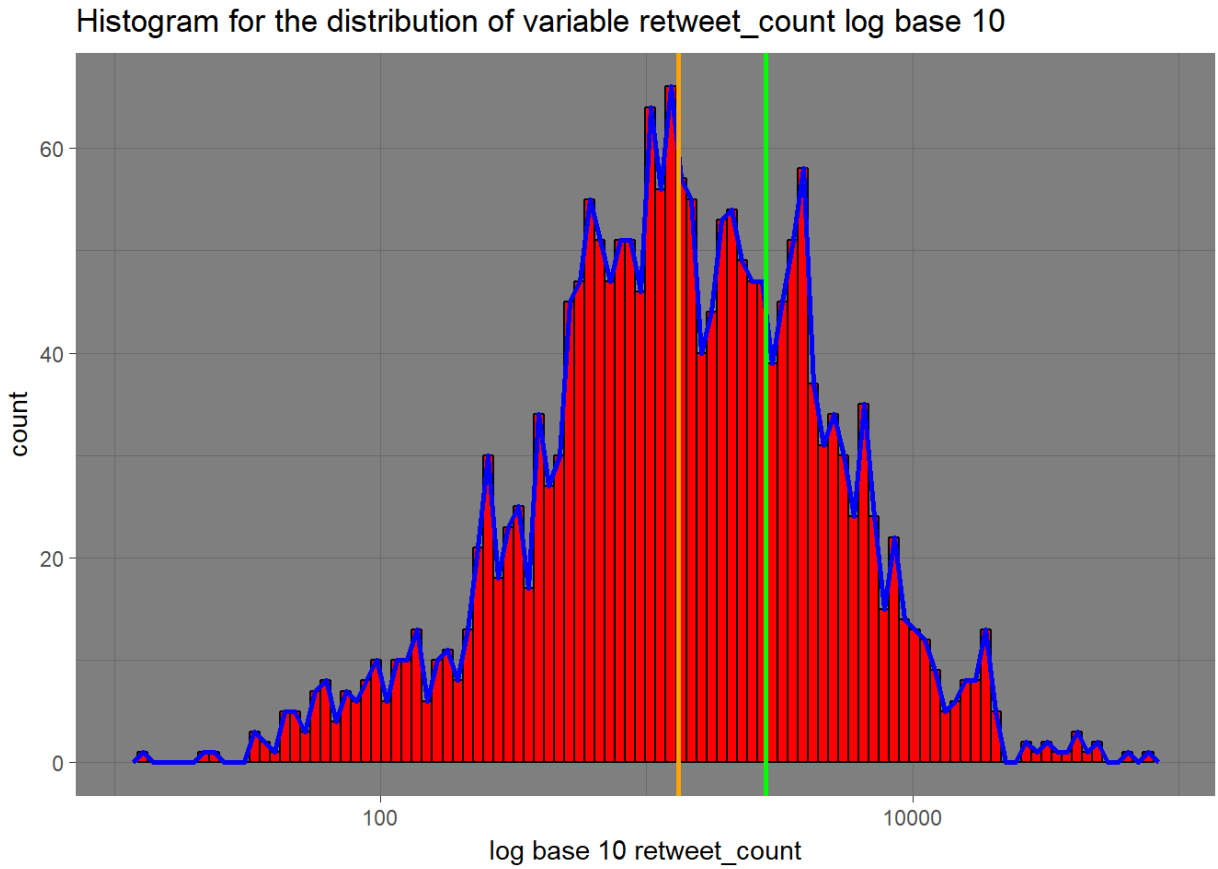
The distribution of this variable has a long tail towards the right and this is usually a good sign to try converting the axis to log scale. After the conversion the image below was produced.



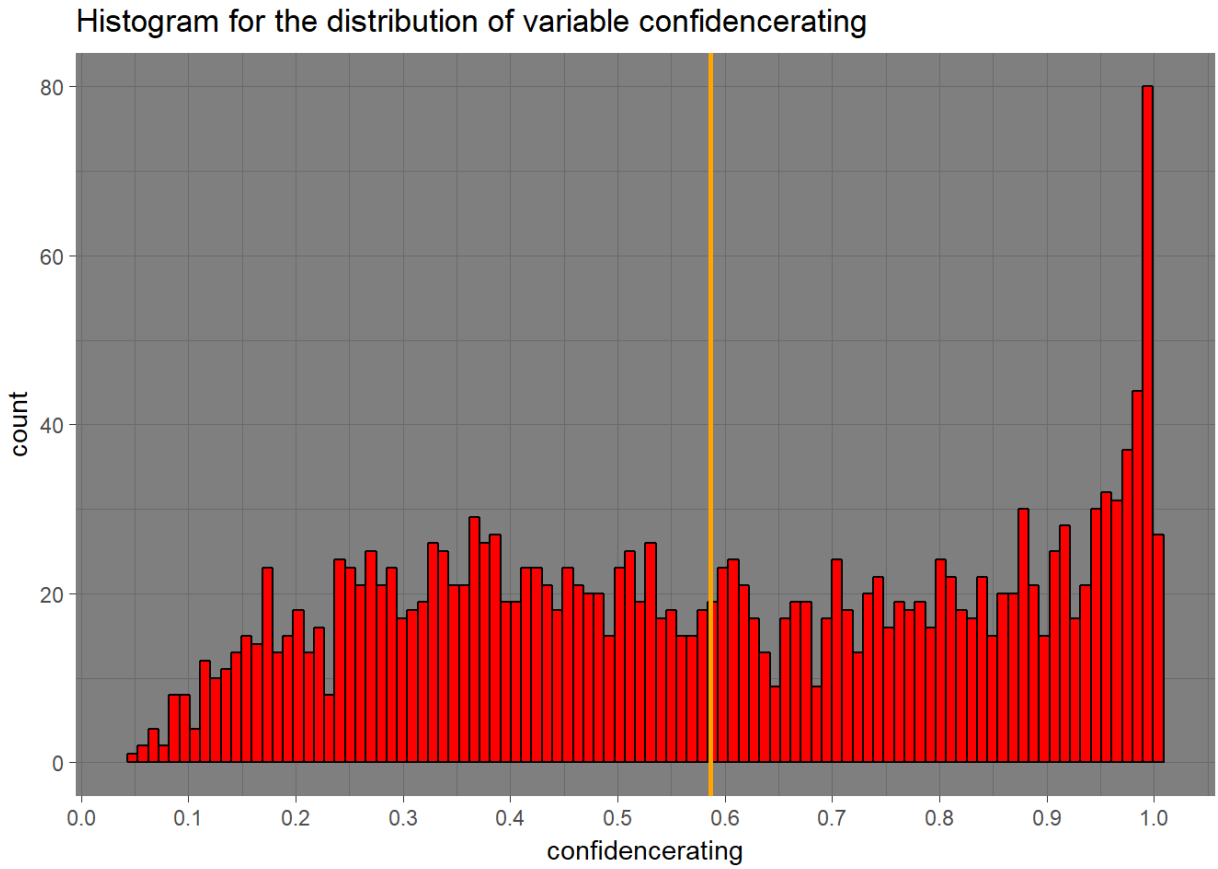
The blue freqpoly shows some up and down spark lines which indicates that this is not a stable variable. The vertical green line presents the mean and the vertical orange line presents the median. Since the mean excedes median this distribution is right skewed. The second variable I looked at is the histogram for the retweet_count variable



The distribution for this variable also has a long tail towards the right. To get a better feel for the distribution of this variable I converted the scale of the axis to log.



The image produced seems more symmetric, but it is also a right skewed distribution since the mean excedes the median. The blue freqpoly shows lots of spark lines which indicate that this variable is also unstable. The final univariate plot I looked at is the histogram for the optimal (best) confidence rating for each tweet.

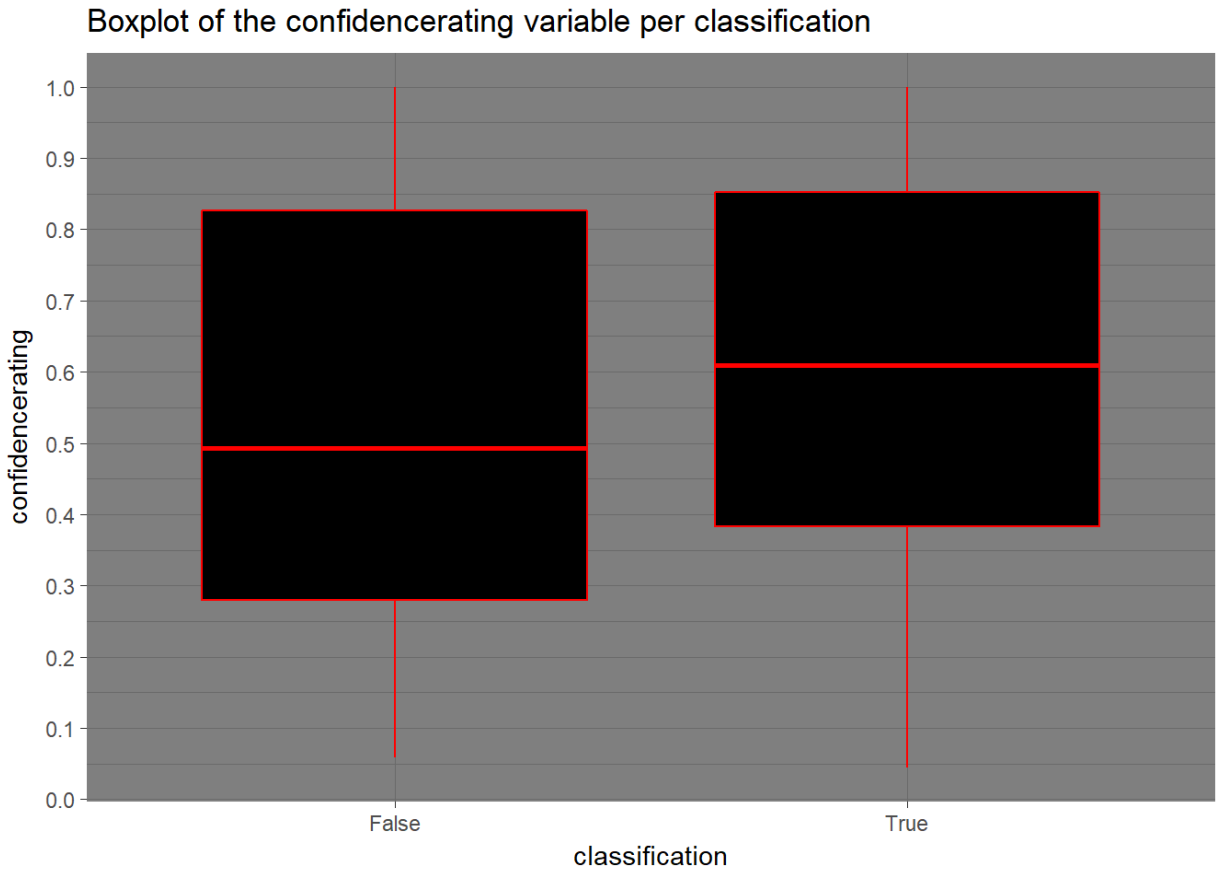


There seems to be a farely wide distribution of the confidence in the prediction ‘power’ of the classification technique applied. The median predicted confidence comes in at just under 60% while most of the observations had a confidence rating of 98% and up.

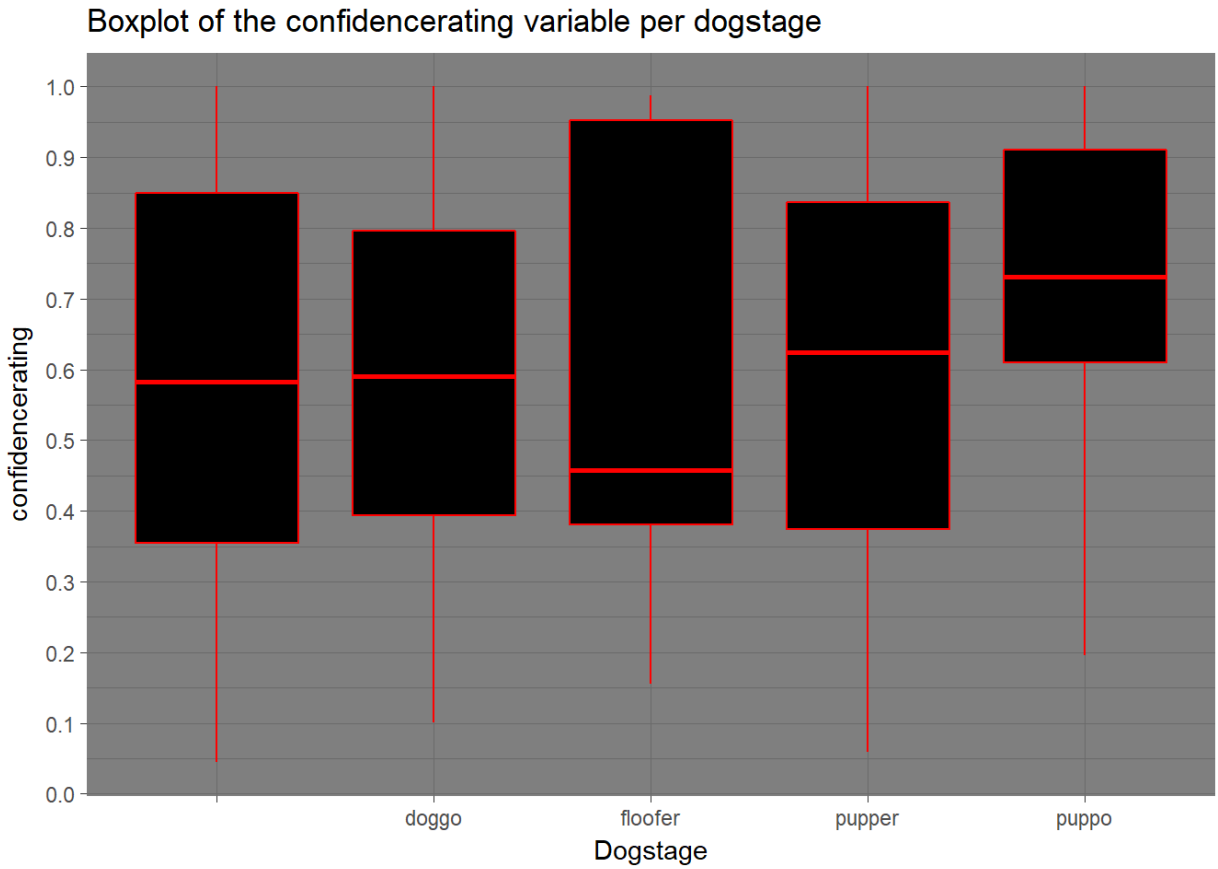
(4) Bivariate Plots

Under this sections I consider bivariate plots to see if there are potentially any other hidden insights by observing a combination of variables. The visual below is a box plot of variable classification against confidence rating

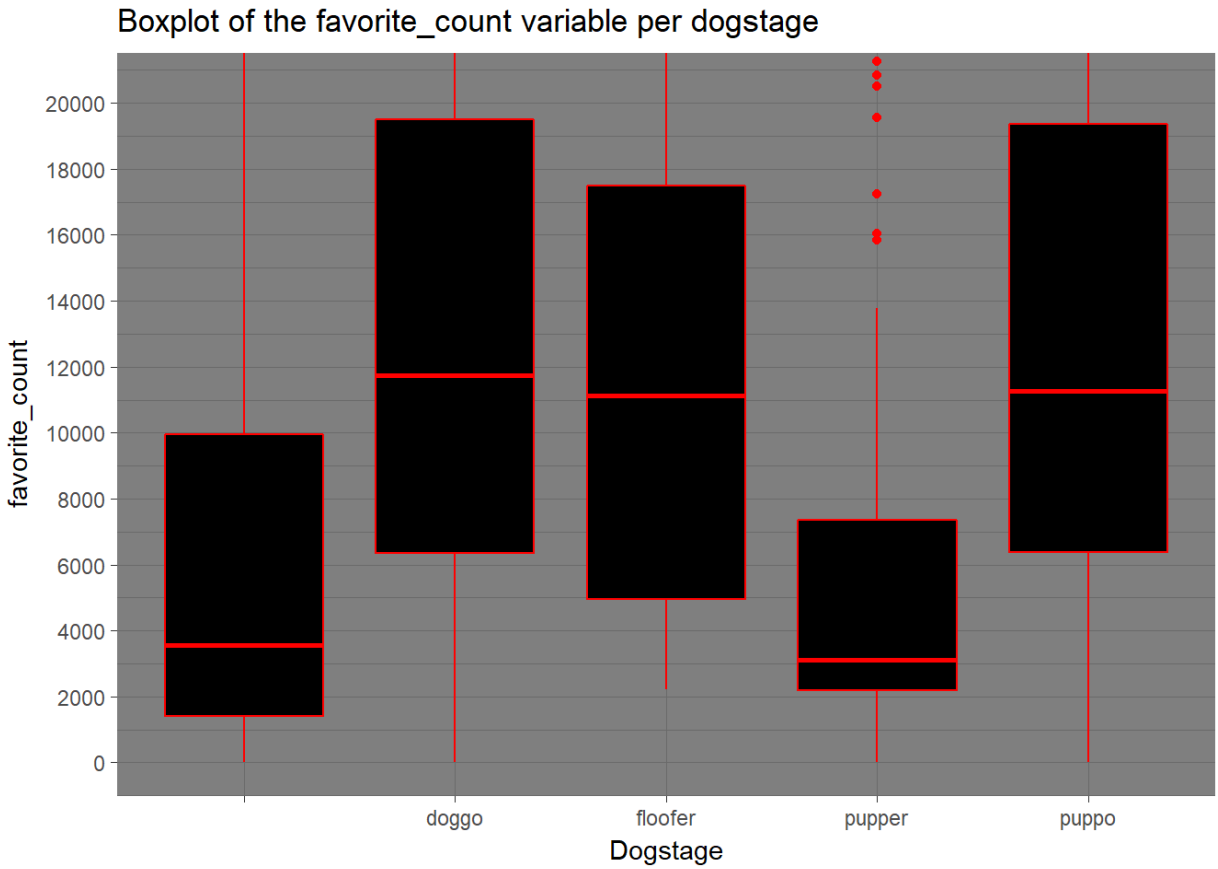
The ‘True’ value presents the concept that the classification is of breed dog and ‘False’ is not a breed of dog. Interestingly the median confidence rating for the classification of identifying a breed of dog is higher than the opposite.



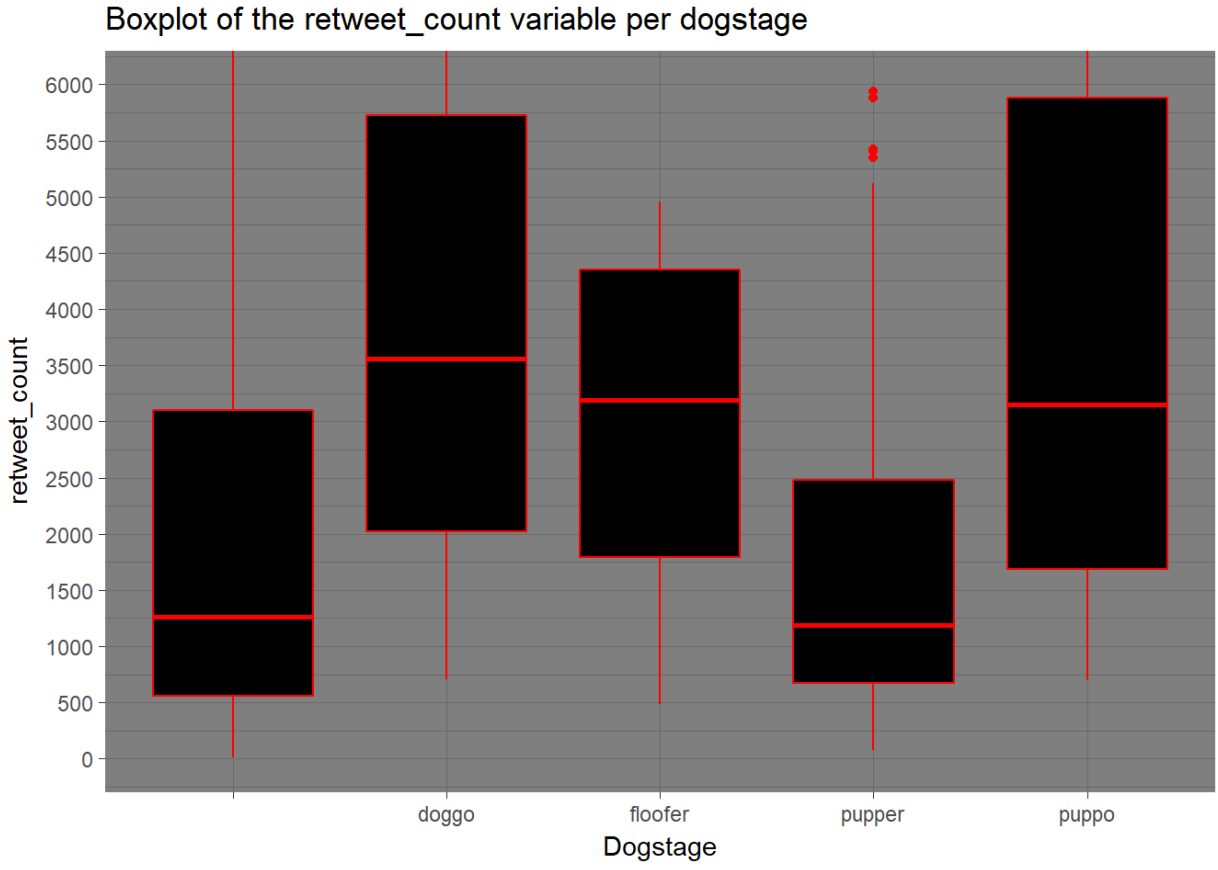
The box plot below considers the categorical variable (dog stage) against the confidence rating. The ‘empty’ bucket on the x axis present those observations who did not have a dog stage. Out of the dogs that did get a dog stage the median confidence rating for a puppo is the highest (even though there weren’t many observations for this category) and that of a floofer is the lowest (understandable since it had the lowest representation within the dataset).



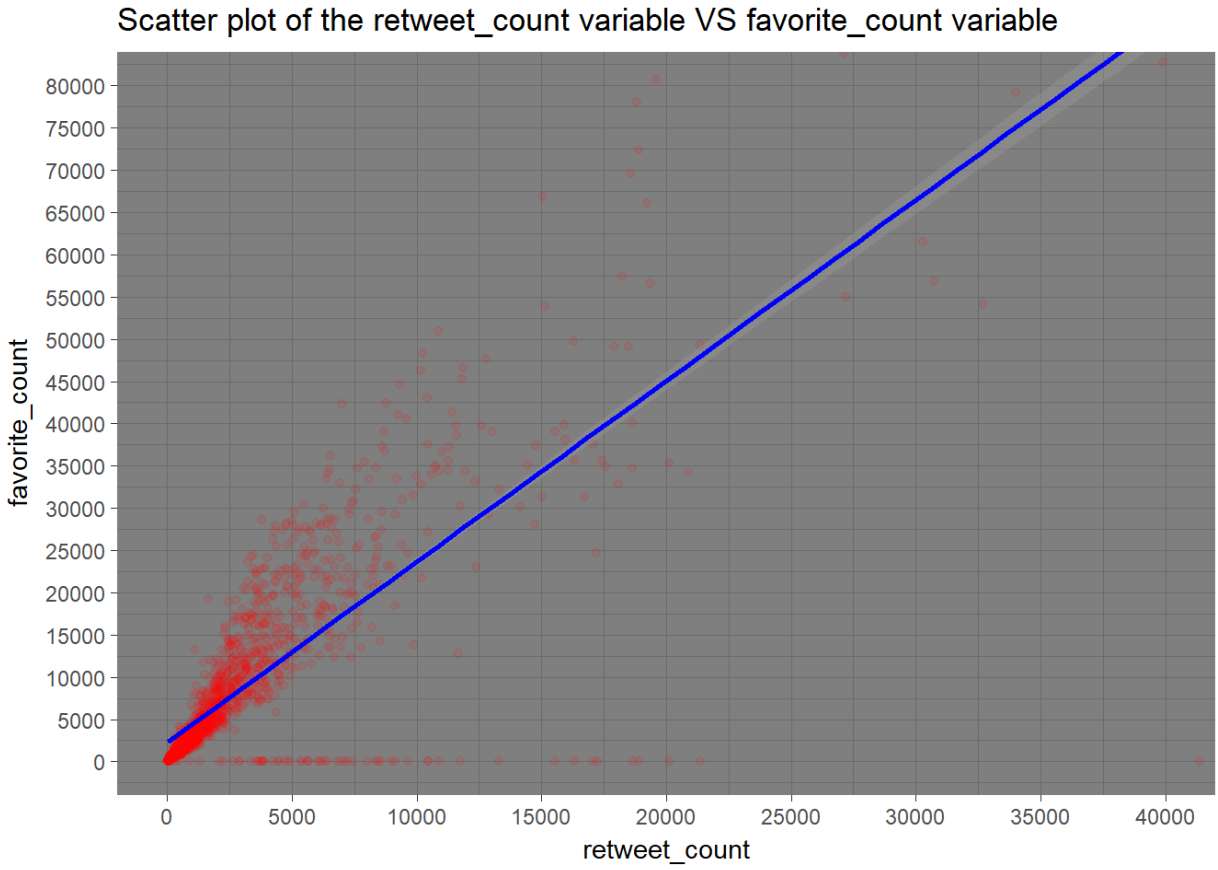
The differnt hights in the box plot below was very interesting. Regarless of the proportion of representation of the various categories within the dataset the median favorite_count for categories: doggo, floofer and pupper are very close, which seems that this could be the stage of a dogs life that most people relate with. The pupper has a few outliers as indicated by the red dots, and as expected the pupper is a lot of hard work and very naughty. This is evident from the fact that it received a lower median favorite_count.



Similar to the box plot above the retweet_count shows similar patterns. The variance observed is also very wide.



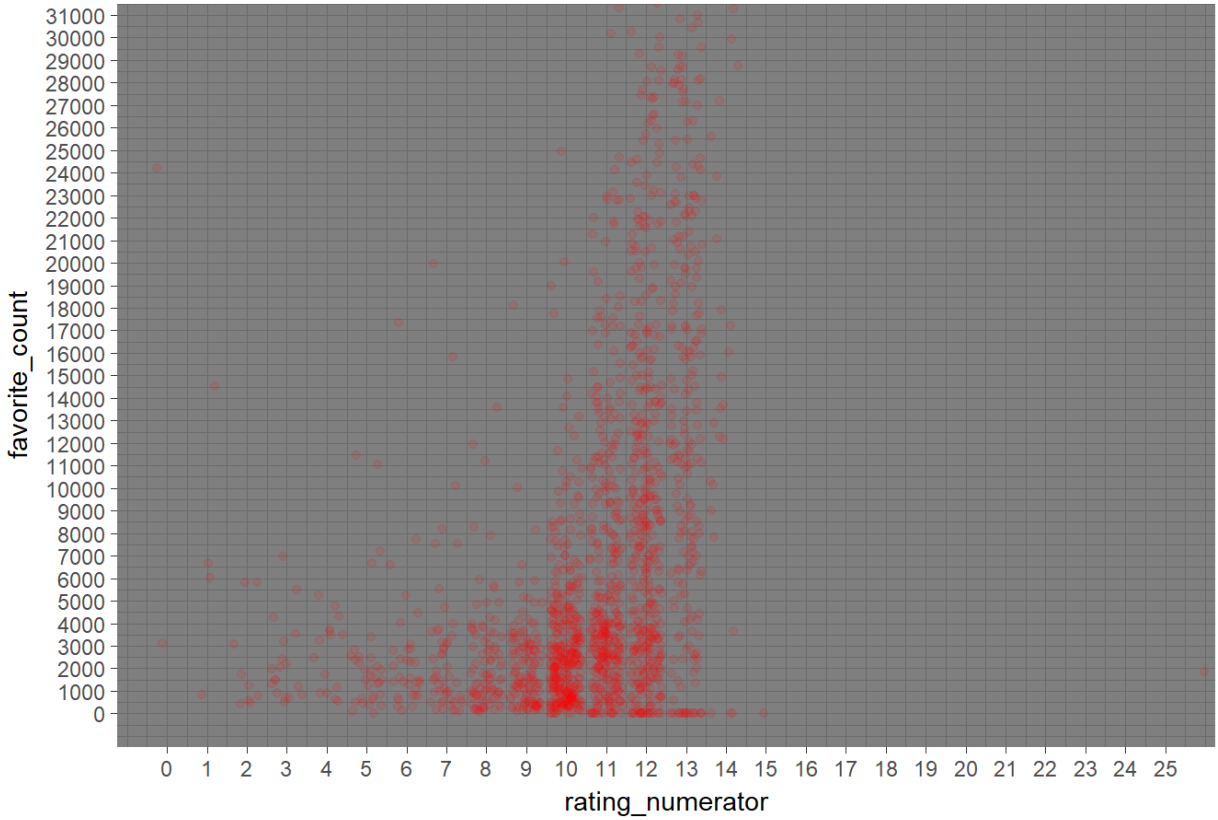
The scatter plot below shows the linear relationship between the favorite_count and retweet_count variables. After performing a correlation test it came back with a correlation coefficient of 85% which indicates that there is a strong positive relationship between these two variables




```
##
##  Pearson's product-moment correlation
##
## data:  twitter_archive_master$retweet_count and twitter_archive_master$favorite_
count
## t = 69.695, df = 1960, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8308796 0.8563610
## sample estimates:
##       cor
## 0.8440962
```

The scatter plots below show that the relationship between the rating numerator and favorite_count or retweet_count is not linear, but rather of a higher order polynomial. For the numerator between 10 and 14 the count variables shoot high up and for the lower numerator values the count variables are low aswell.

Scatter plot of the rating_numerator variable VS favorite_count variable



Scatter plot of the rating_numerator variable VS retweet_count variable