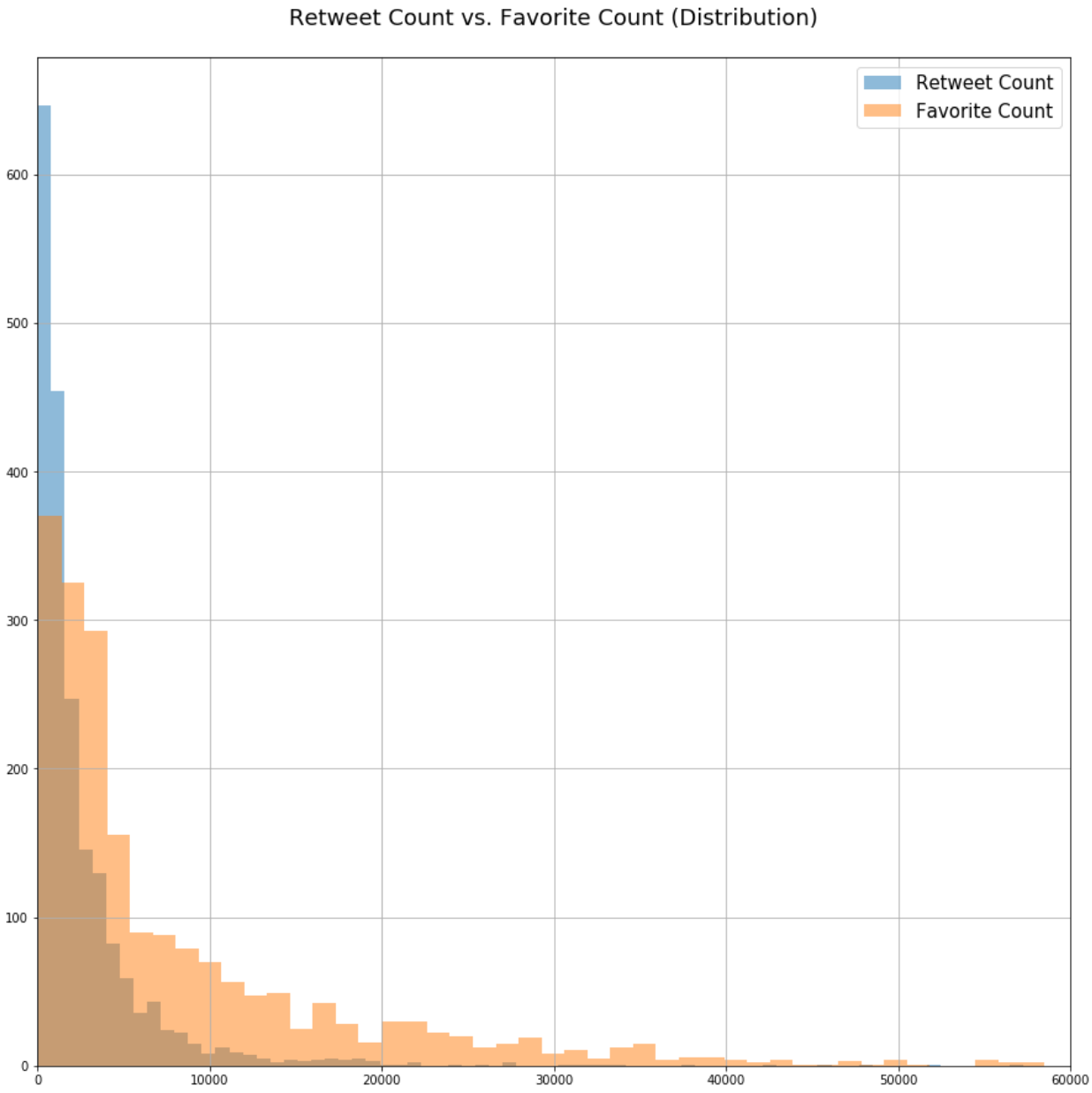
Data Wrangling Report

1. Take care of the imports needed for the analysis
2. Load the data and get a quick preview of the data

For the first analysis I will attempt to identify the distribution of the retweet count as compared to the distribution on favourite count.

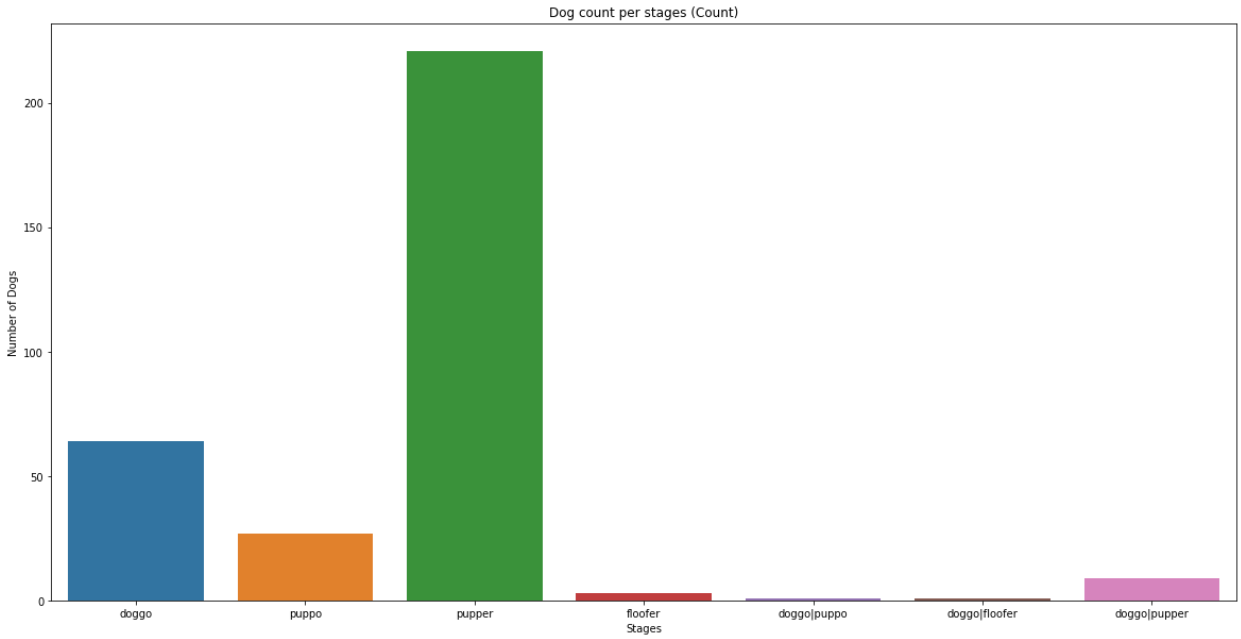


From the histogram above, we note that the distribution of both retweets and favourite is significantly right-skewed, however, the mean of favourite count (8895.7) is greater

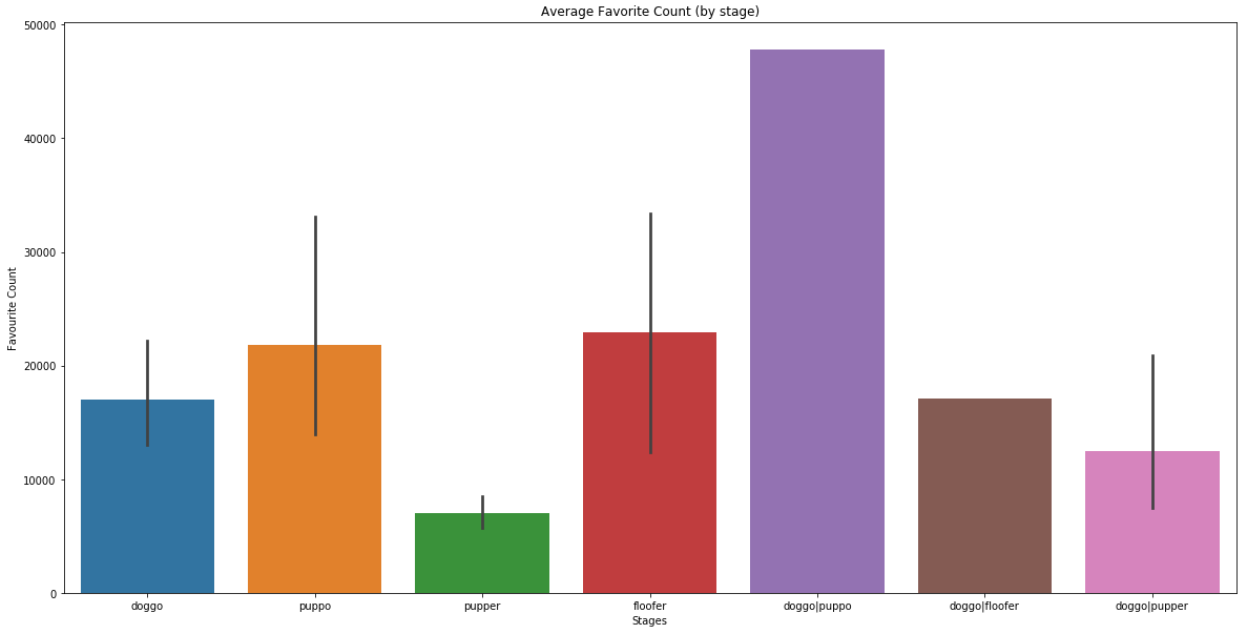
than retweet count (2766.7). We also note that the maximum number of favourite a tweet received was 132810, where the maximum number of retweets a tweet received was 79,515.

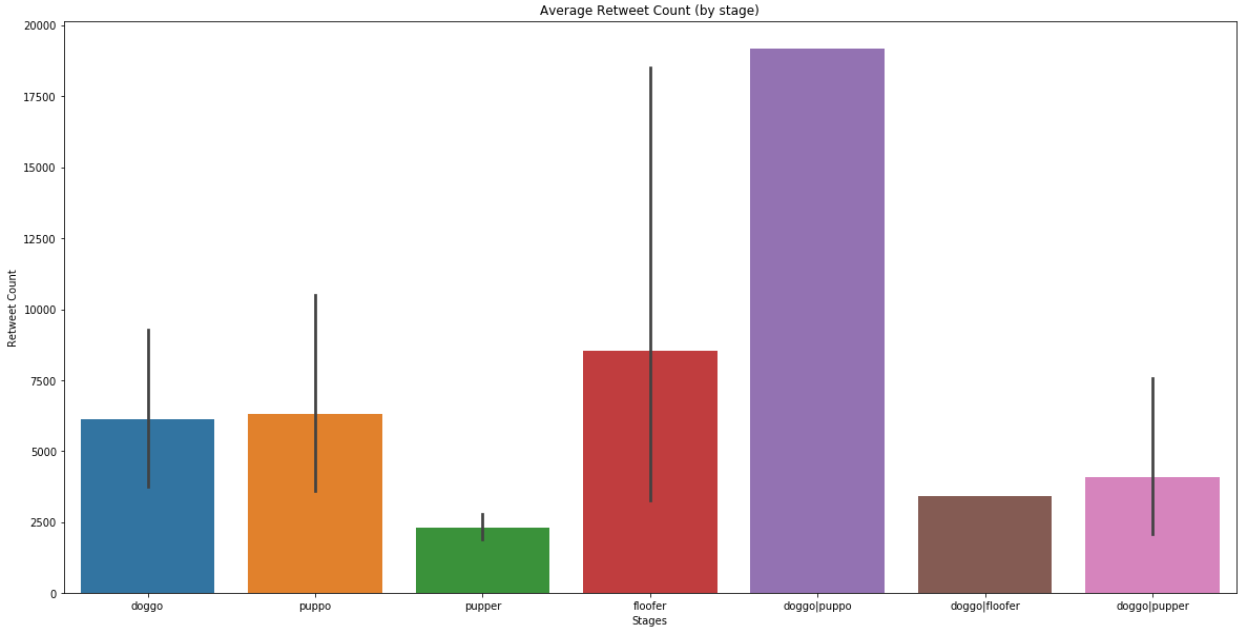
For the next step I will attempt to find the count per stage.

We



We can note that pupper have the count but we need to dig deeper. For this ill find the average count for each stage in relation to retweet count and favourite count.





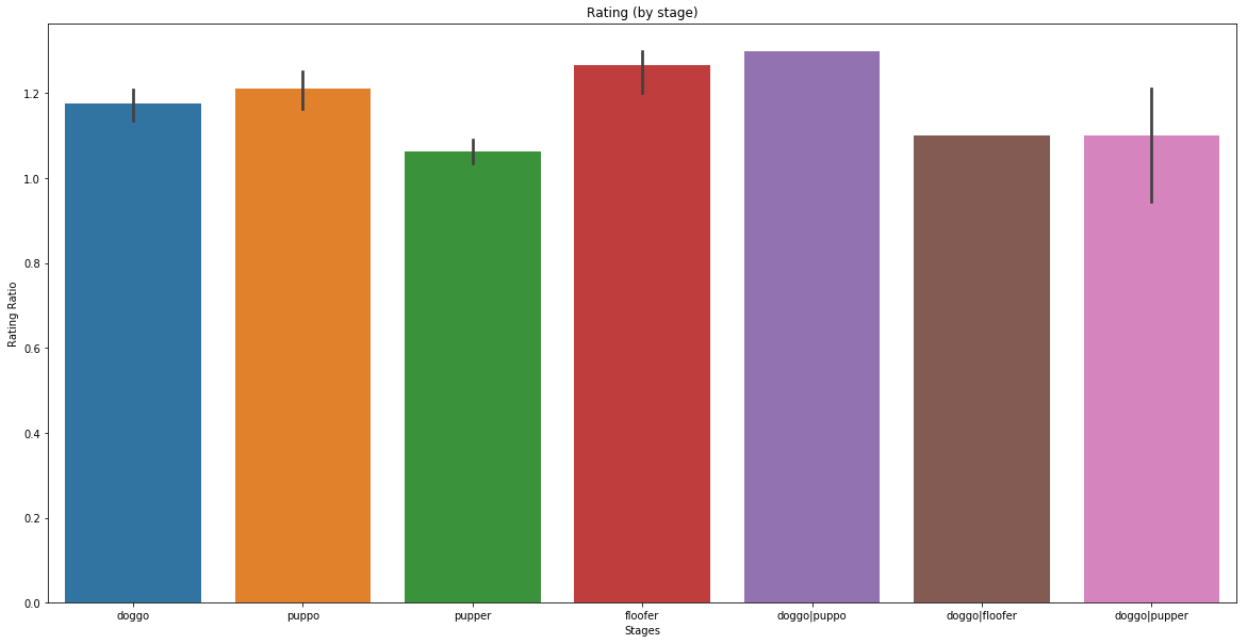
From the three plots above, we can note the following:

- The pupper stage has the highest count where the (doggo|floofer) and (doggo|puppo) has the lowest count.

- The average retweet_count for (doggo|puppo) is the highest, where the average retweet_count for the pupper is the lowest (less than 2,500).

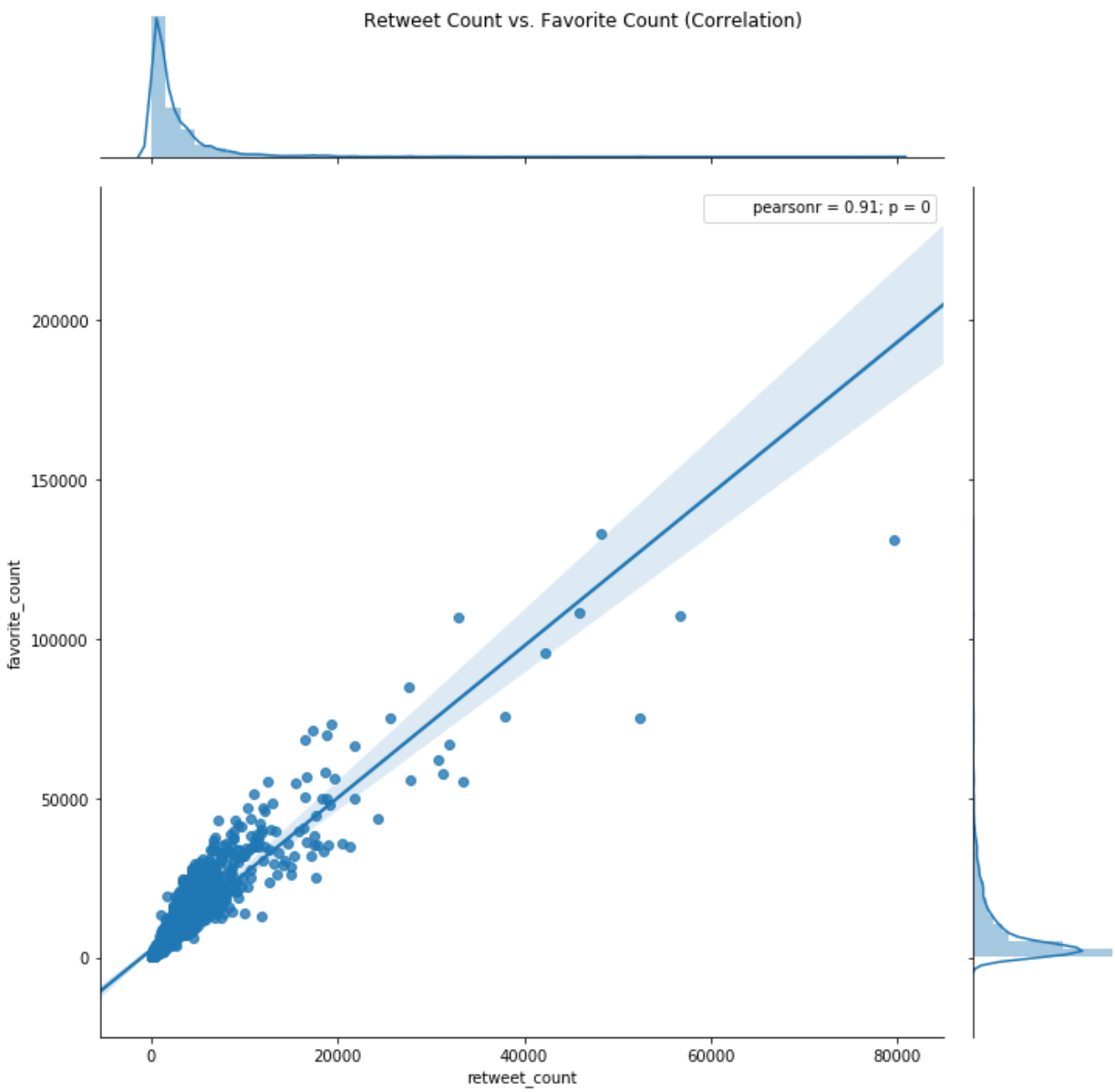
- Similarly, the average favorite_count for (doggo|puppo) stage has the highest average favorite_count, where the pupper has the lowest average favorite count.

We can also find out which stage has the highest rating.



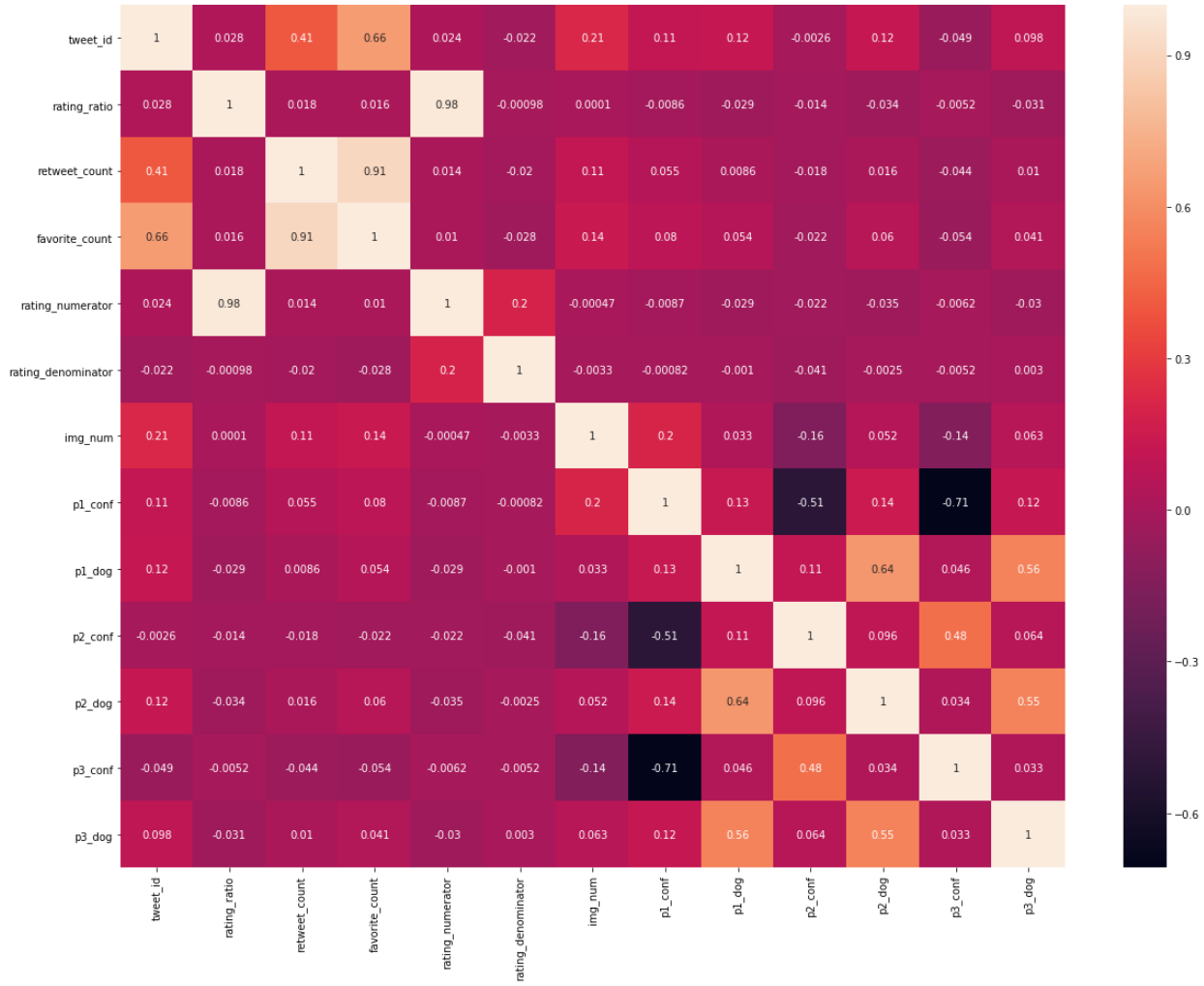
From the plot above, we can note that the doggo|puppo stage has the highest average rating where the pupper stage has the lowest rating.

We can also identify whether there is correlation between the retweet count and the favourite count.



From the plot above, we can note that there is a positive correlation between the favorite_count and the retweet_count since the number of retweets increase as the number of favorites increase, which is expected as most people who would mark a tweet as a favorite are most likely to retweet it.

For the last visualization I will attempt to derive the relations in the data using a heatmap.



From the heatmap above, we can note the following:

- There is a strong positive correlation (0.98) between `rating_ratio` and `rating_numerator` which is given (rating_ratio is based on the rating_numerator and rating denominator)

- There is a relatively strong negative correlation (-0.71) between `p3_conf` and `p1_conf` which is interesting as different predictions may have opposing confidence levels.

-There is a strong correlation (0.91) between `favourite_count` and `retweet_count`.