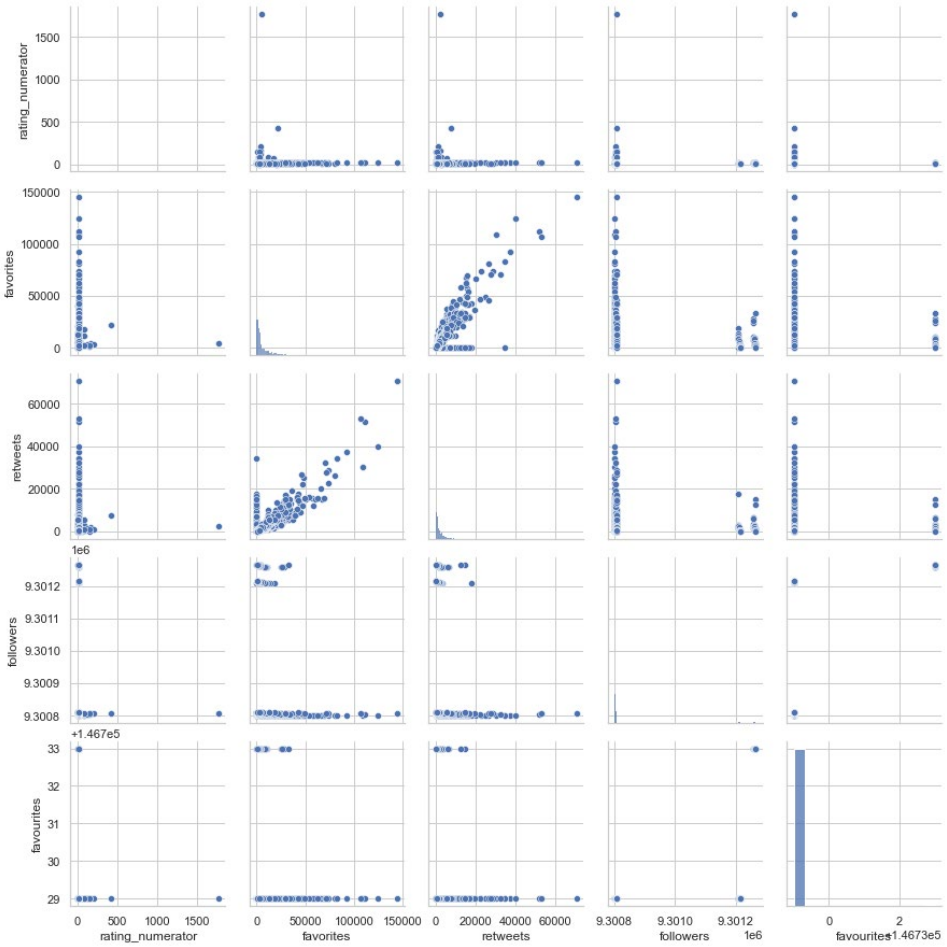
# Analysis Report

**Ednalyn C. De Dios**
**D309, WGU**
**07/09/2022**

## Pairplot

I used a pairplot to compare how variables compare with each other.

A plot that looks like a Southwest to Northeast line suggests a positive trend or positive correlation. There is strong positive correlation between favorites and retweets. This is helpful to know if we are to isolate truly independent variables to avoid multicollinearity. Multicollinearity is bad because “it undermines the statistical significance of an independent variable.”

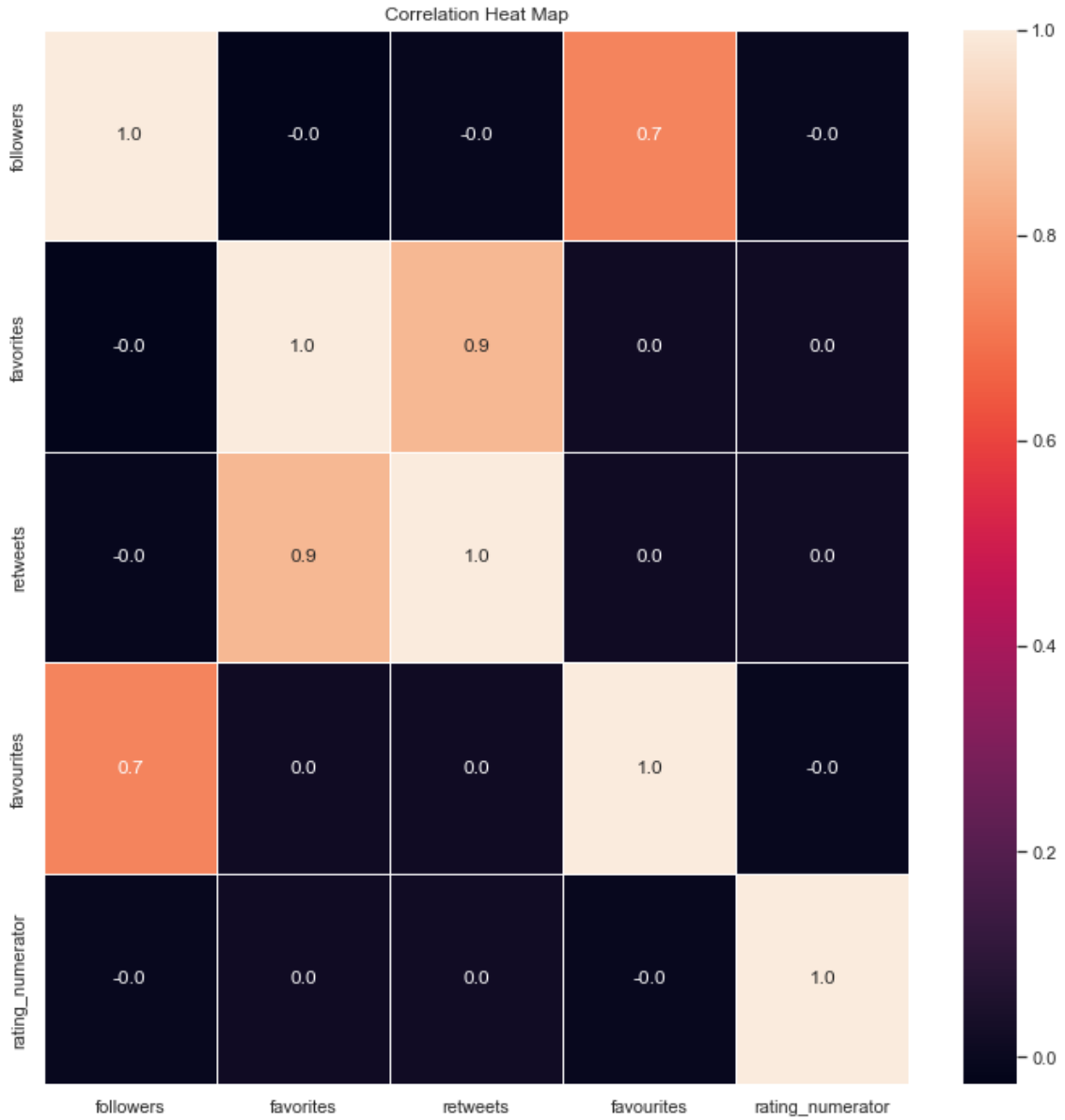


# Heat Map

Another visualization I used to compare the variables to each other is the heat map.

The darker shade (almost black) suggests a low correlation while the lighter shade suggest the opposite. The Northwest to Southeast diagonal squares represents perfect multicollinearity because they are identical to each other and should not be included in the analysis.

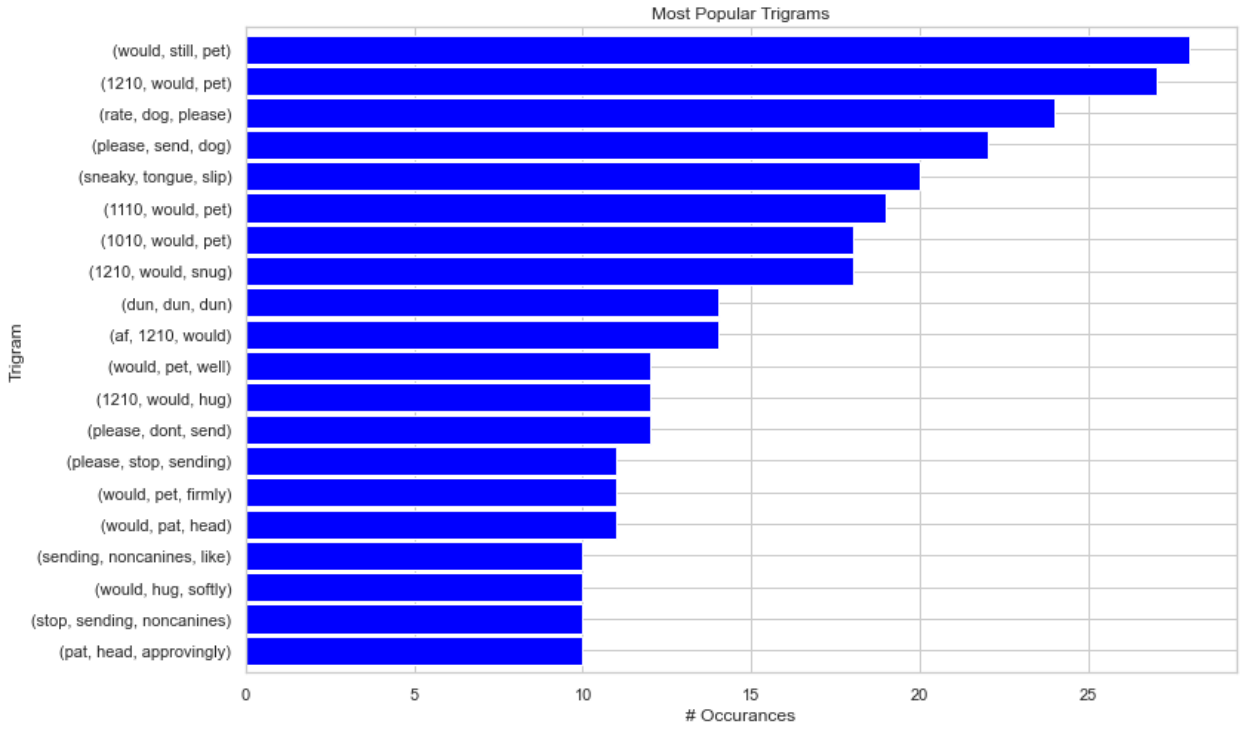
Below, we see a strong correlation between favorites and retweets, corroborating the findings above. Also, we observe that favourites and retweets are also correlated albeit not as strong as the previous one.



## NLP

A simple plot of the most popular trigrams is shown below.

From cursory examination, we can see that the textual data needs most cleaning. The graph also suggests that there are possible word combinations to investigate like "would still pet", "rate dog please", and "please don't send." In addition, the numbers 1210 and 1110 seems to repeated significantly and this warrants a deeper examination so it could be determine whether they should be excluded (added to the Stop Words list) or kept.




# Conclusion

The WeRateDogs is fun to play with. Since it has textual freeform data, it also presents a good opportunity to try out some NLP techniques.

# Reference

https://link.springer.com/chapter/10.1007/978-0-585-25657-3_37