Insights Report

Introduction:

The Wrangle and Analyze Data project is part of Udacity's Data Analyst Nanodegree program. This project included gathering data from the Twitter page WeRateDogs. The data was gathered from several sources and was then assessed, cleaned, and stored, and I then created some visualizations. Below are the insights from my visualizations.

Descriptive Statistics:

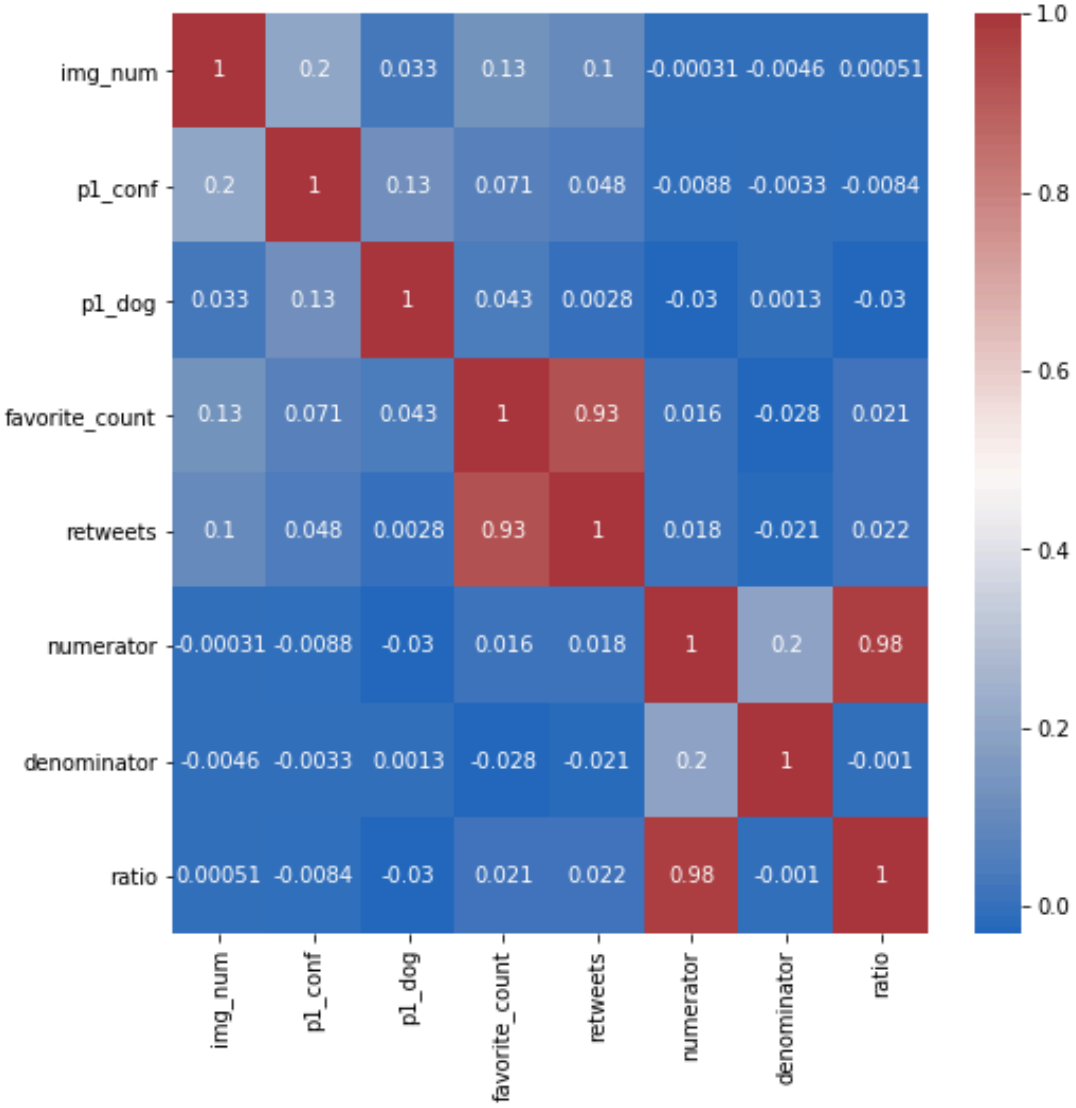
The first step I took when analyzing the data was to gather the descriptive statistics for the data set. Below is the summary of the data:

	img_num	p1_conf	favorite_count	retweets	numerator	denominator	ratio
count	1986.000000	1986.000000	1986.000000	1986.000000	1986.000000	1986.000000	1986.000000
mean	1.203424	0.593177	8179.125881	2417.440584	12.235539	10.538771	1.163836
std	0.561492	0.271956	12046.000424	4325.191726	41.553063	7.332462	4.072127
min	1.000000	0.044333	70.000000	11.000000	0.000000	10.000000	0.000000
25%	1.000000	0.360998	1746.750000	540.000000	10.000000	10.000000	1.000000
50%	1.000000	0.587222	3677.500000	1171.500000	11.000000	10.000000	1.100000
75%	1.000000	0.843883	10160.750000	2754.500000	12.000000	10.000000	1.200000
max	4.000000	1.000000	154210.000000	76476.000000	1776.000000	170.000000	177.600000

A lot of insights can be gained from this data set. There appear to be some outliers in every category, and this would need to be addressed if any future work would be done with this dataset. The p1_conf column has an average of ~.593, and this means that the primary prediction is only about 59.3% sure that a prediction is a dog, on average. There may be room to improve here. The ratio value of 1.163836 means that most dogs are on average ~11.6/10, and this make sense because they're all very good dogs!

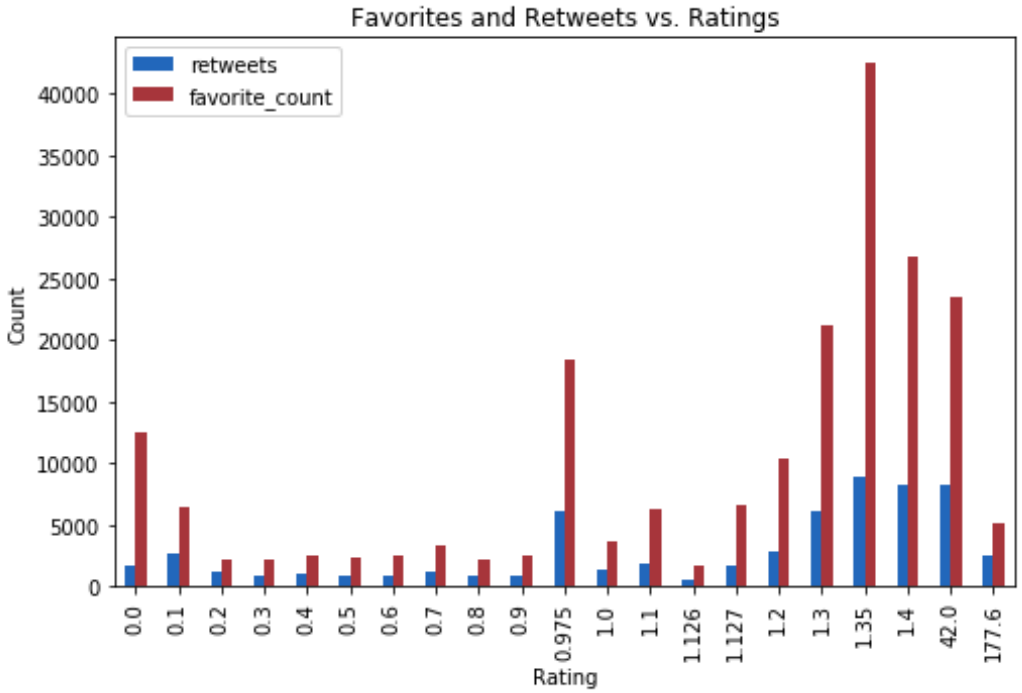
Heatmap:

The heatmap will tell us how each variable is correlated to the other variables. This is a good step to take if you are going to make a regression or other prediction model because if you have too many variables that are correlated, your model could be prone to overfitting. The heatmap below shows that there are two significant correlations: 1) between favorite_counts and retweets and 2) between numerator and ratio. Dropping one of each of these would be a good idea when creating a prediction model, but further testing would still be necessary to determine the features of any prediction model. Below is the heatmap:


img_num
p1_conf
p1_dog
favorite_count
retweets
numerator
denominator
ratio
1 0.2 0.033 0.13 0.1 -0.00031 -0.0046 0.00051
0.2 1 0.13 0.071 0.048 -0.0088 -0.0033 -0.0084
0.033 0.13 1 0.043 0.0028 -0.03 0.0013 -0.03
0.13 0.071 0.043 1 0.93 0.016 -0.028 0.021
0.1 0.048 0.0028 0.93 1 0.018 -0.021 0.022
-0.00031 -0.0088 -0.03 0.016 0.018 1 0.2 0.98
-0.0046 -0.0033 0.0013 -0.028 -0.021 0.2 1 -0.001
0.00051 -0.0084 -0.03 0.021 0.022 0.98 -0.001 1
1.0
0.8
0.6
0.4
0.2
0.0

Favorites and Retweets:

Another interesting insight is seeing how ranking affected the favorite and retweet count. Most of the favorites and retweets occurred between 1.3 and 1.4, and you can also see that that retweets occur far less often than favorites. Above we saw that retweets and favorites have a high correlation (.93), and this graph confirms that relationship. Below is the graph:



Conclusion:

In conclusion, this analysis barely scratches the surface of what can be done with this dataset. These are just a few graphs and insights that I found interesting, but as I learn more about data analysis and machine learning, I will likely come back to this dataset to practice.