Analysis and Insights

Wrangle and Analyze Data project

5/24/2020

Suryaday Nath

INTRODUCTION

This Wrangle and Analyze Data Project forms part of Udacity's Nanodegree Data Analyst.

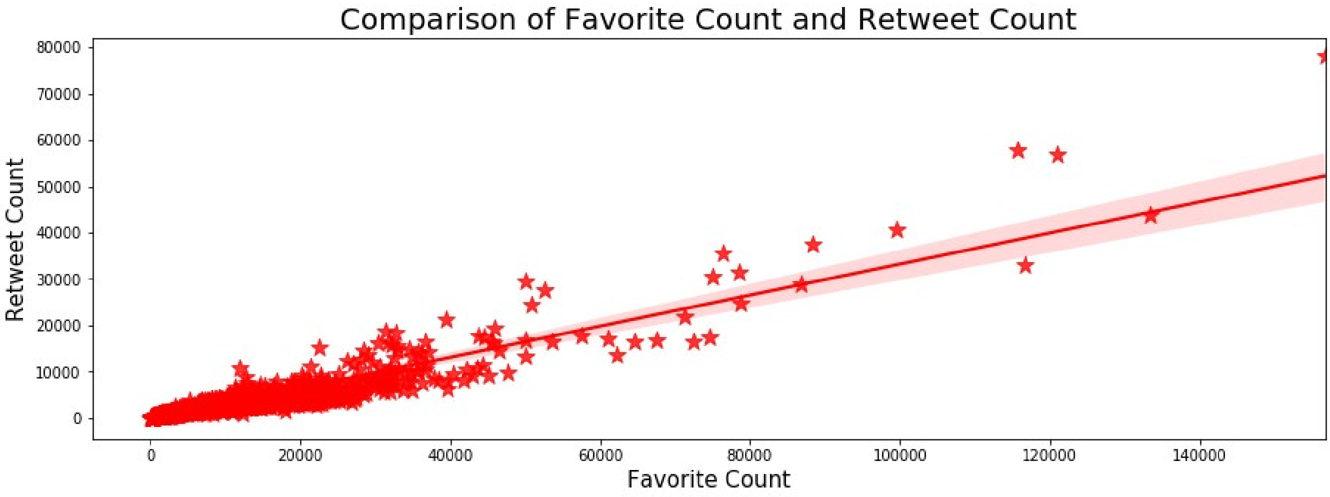
The project involves wrangling data from the Twitter user @WeRateDogs, from various sources associated with the tweets. WeRateDogs rates photos of people's dogs in a humorous way, most frequently offering scores above 10/10 because they are good dogs Brent!

Performance and tidiness problems were evaluated and then cleaned after the data had been scraped together. Finally, two visualizations were created and insights can be found below. Also two word clouds were created that show Most popular dog breeds and dog names.

FAVORITE VS RETWEET COUNT

WeRateDogs had more than 4 million followers at the time this data was collected; thus, their tweets are likely to get many favorites and retweets.

In fact, if they are part of international news coverage or go viral, there might be some tweets which are highly common. Figure 1 shows that favourite and retweet counts are highly correlated positively. There is 1 retweet for about every 4 favourites. Most data dropped below 40000 favourites and 10000 retweets. There are about 130,000 favorites and 80,000 retweets on the most popular tweet.



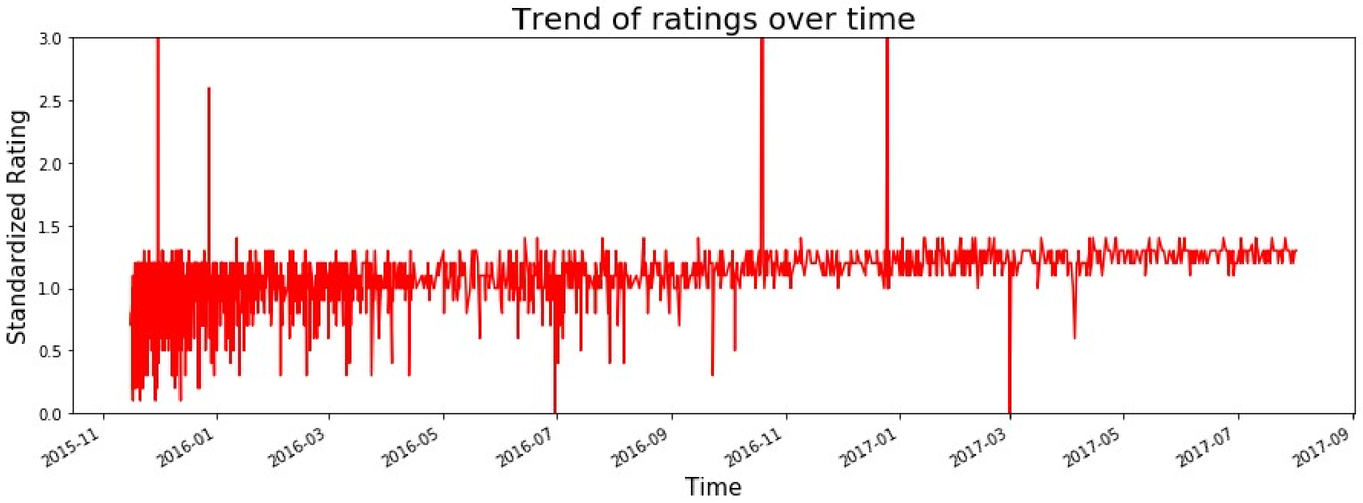
STANDARDIZED RATING OVER TIME

The premise behind the WeRateDogs account is that they invite users to send them pictures of their dogs and, with funny remarks, they will score them on a scale of 1-10; however, they are also given scores above 10.

It was assumed that nearly all the dogs had a rating higher than 10/10 but there was many ratings below that. Additionally, there were not many ratings with a denominator of 10 or numerator with 10.

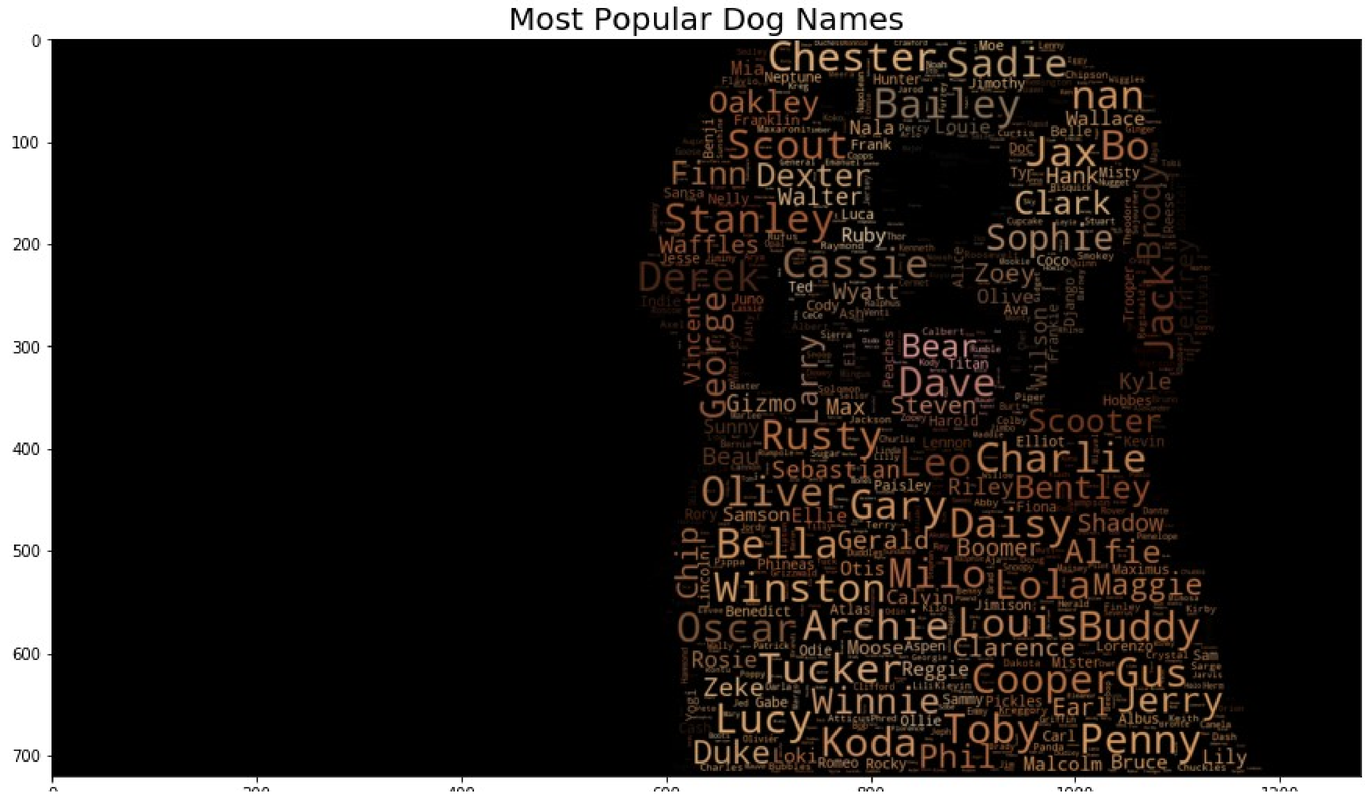
It could be interesting to see if overtime, with the account becoming more popular and people associating the above 10/10 ratings with being funny would make the higher ratings more predictable.

Indeed, as shown in Figure 2, the frequency of ratings below 1 decreases with overtime. There are many ratings below 1, before 2016-11, while there are barely any after this time. The average uniform rating is roughly 1.3 except for a few outliers, including the joke scores 1776/10 and 420/10, which were not cleaned up.

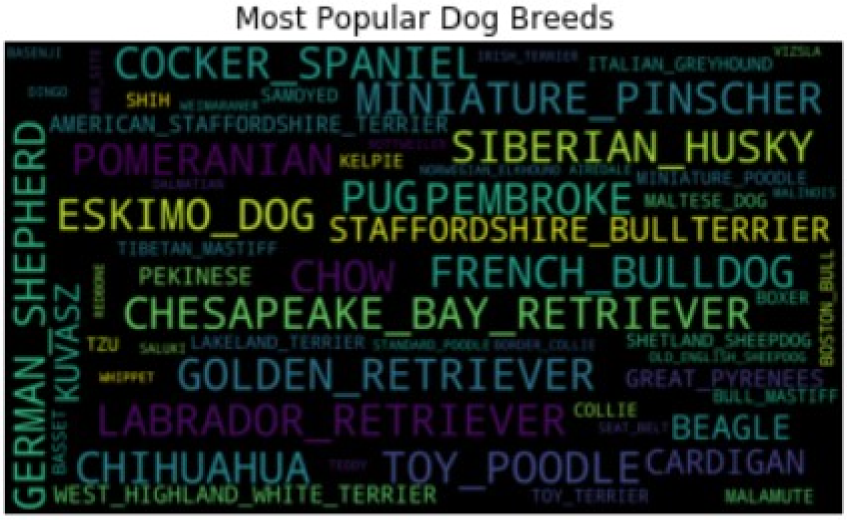


Thanks to Andreas Mueller and his easy code for creating Word Clouds, I was able to create two word clouds – one which shows most popular Dog names and another which shows most popular dog breeds.

Word Cloud showing most popular Dog Names



Word Cloud showing most popular Dog Breeds



Conclusion:

This project was important to understand that real life data doesn’t usually come from a single source but has to be extracted from various sources to even start analyzing for some information. Also, it emphasized on Tidyness and Quality issues that occur in datasets and how to deal with those issues.