ANALYSIS AND INSIGHTS

Udacity DAND: Wrangle and Analyze Data Project

By: Himanshu Tripathi

INTRODUCTION

This Wrangle and Analyze Data Project is part of Udacity's Data Analyst Nanodegree Term 2. The project involves wrangling of data from various sources associated with tweets from the Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs rate's pictures of people's dogs in a humorous manner, most often giving ratings higher than 10/10. After scraping together the data, quality and tidiness issues were assessed and then cleaned. Finally, two visualizations were created and insights can be found below.

FAVORITE VS RETWEET COUNT

At the time this data was collected, WeRateDogs had over 4 million followers; therefore, their tweets are likely to get many favorites and retweets. In addition, there may be some tweets that are extremely popular if they become part of international news coverage or go viral. In Figure 1, it can be seen that favorite and retweet counts are highly positively correlated. For about every 4 favorites there is 1 retweet. The majority of the data falls below 40000 favorites and 10000 retweets. The most popular tweet has about 130000 favorites and 80000 retweets.

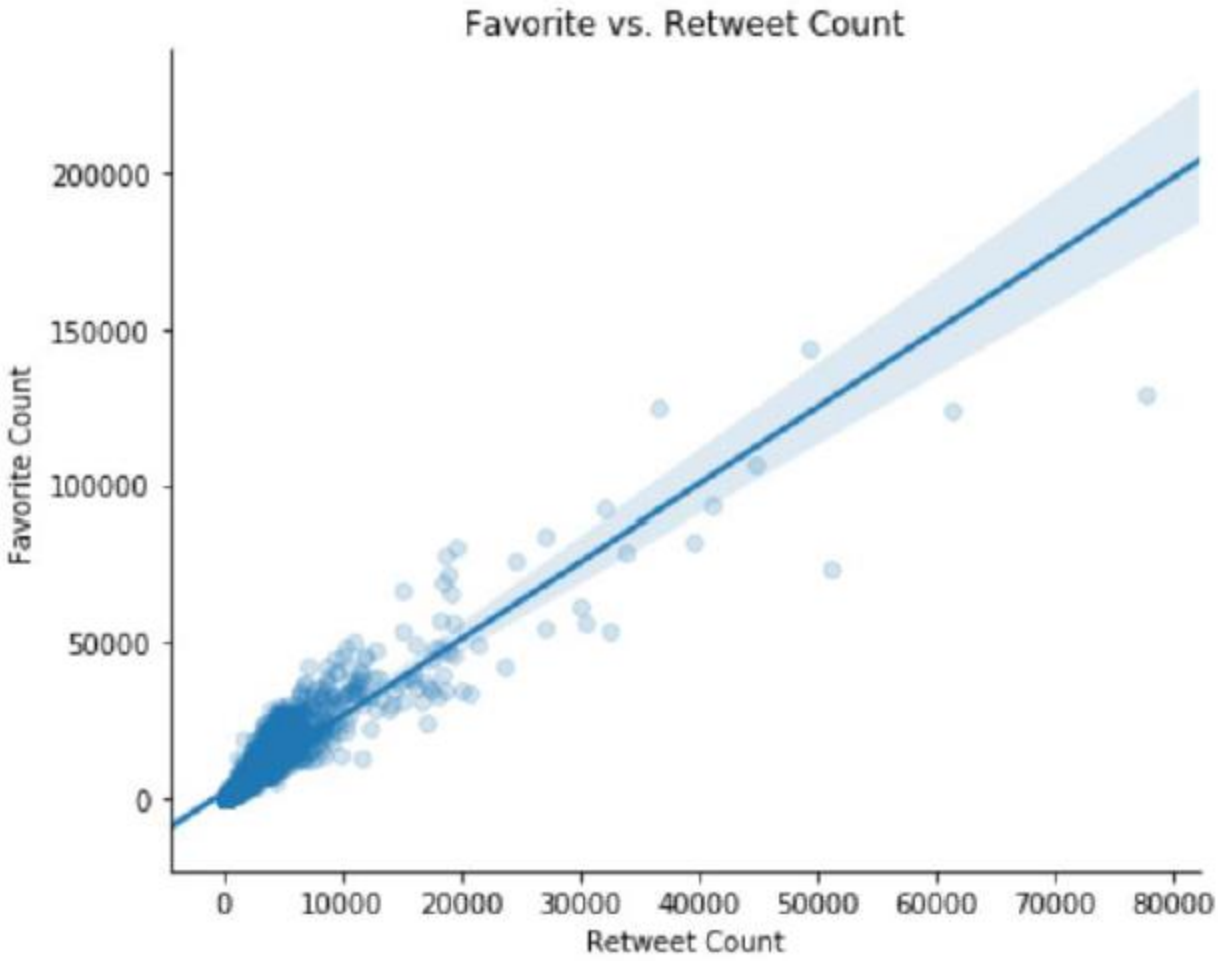

Figure 1. Favorite count and retweet count from 2068 tweets from the Twitter user @dog_rates, also known as WeRateDogs.

STANDARDIZED RATING OVER TIME

The idea behind the WeRateDogs account is that they ask people to send them photos of their dogs, and they will rate them on a scale of 1-10 with humorous comments; however, they are often given ratings higher than 10. I assumed that almost all the dogs were given a rating higher than 10/10 but I was surprised to notice many with ratings lower than 10/10. In addition, many ratings did not have a denominator of 10. Therefore, to standardize the ratings I calculated a value of numerator divided by denominator. I was most curious to see if overtime, as the account became more popular and people associated the above 10/10 ratings with being funny, that the higher ratings would become more prevalent. Indeed, as shown in Figure 2, it appears that overtime the frequency of ratings below 1 decreases. Before 2016-11 there are many ratings below 1, while after that time there are barely any. The maximum standardized rating is about 1.3 except for three outliers including the joke ratings 1776/10 and 420/10 and 1 error that did not end up getting cleaned.

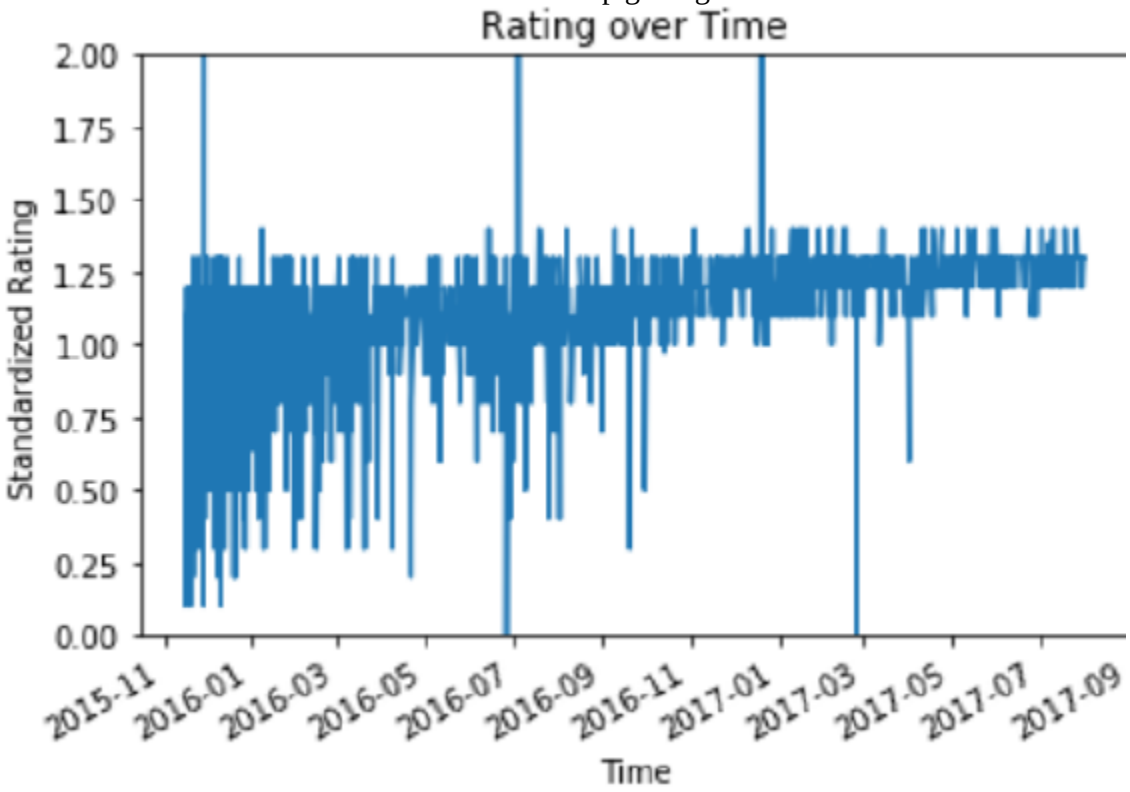

Figure 2. Standardized rating over time from 2068 tweets from the Twitter user @dog_rates, also known as WeRateDogs.