Wrangle report (Udacity Data Analysis Project Wrangle and Analyze Data)

Introduction

Wrangle and Analyze data project involves gathering data from various recourses. It is associated with tweets from the Twitter user @dog_rates, also known as WeRateDogs. Three main steps were processed in order to complete this project including, gathering data from different sources, assessing gathered data and clean data, since the data you need comes in different formats and ways, we had to assess this data and clean it in order to make it able to be analyzed and visualized. At the end, two visualizations were created from the dataset which will be discussed below along.

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. Figure 1 shows 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.

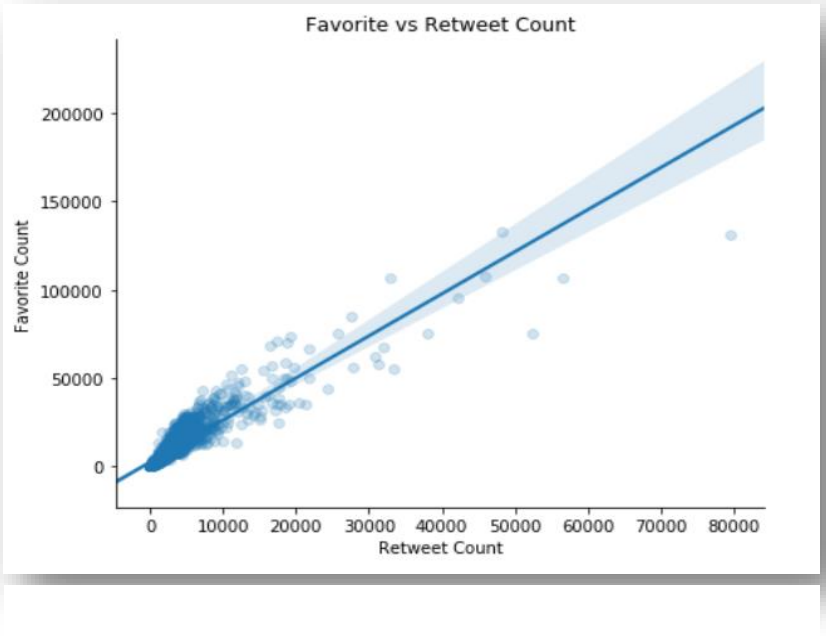

Figure 1 Favorite VS Retweet Count

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 out of 10 with humorous comments. although, they are often given ratings higher than 10, some dogs were given a rating below 10, in addition many ratings have no denominator of 10. Therefore, to standardize the ratings by calculating a value of numerator divided by denominator. It was noticed that 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. Figure 2 shows that overtime the frequency of ratings below 1 decreases. Before 2016-11 there are many ratings below 1, while after that time there are very few.

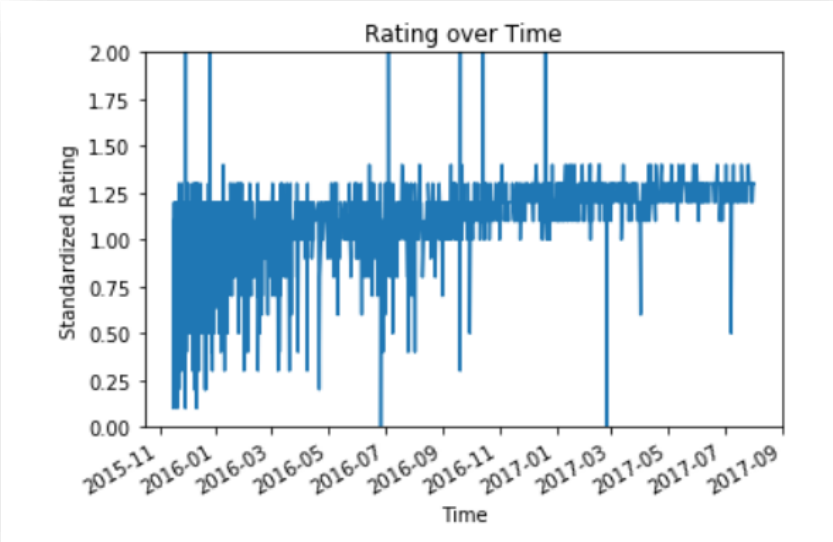

Figure 2 Rating over Time

Dogs Types

Among dogs we have in the dataset we found that there are 4 dog's types, the majority of them are of the type Pupper with 223 dogs (more than half), Duggo comes second place in terms of capacity with 72 dogs, leaving Puppo type at the third place with 28 dogs. Finally, there are only three dogs of type floofer.

```
# How many dogs we have from each type in our dataset
twitter_Archive_Clean.dog_stage.value_counts()
```

```
pupper     223
doggo       72
puppo       28
floofer      3
```