Analysis Report

WeRateDogs is a Twitter account that rates people's dogs. It has over 4 million followers and has received international media coverage. WeRateDogs almost always gives rating greater than 10 (11/10, 12/10, 13/10, etc.) Such unique rating system plays big part of the popularity of WeRateDogs.

To analyze the tweets form WeRateDogs, I gathered data from three different sources. The first source is the archive of past tweet provide by Udacity in the form of a CSV file. The second source is dog breed predicted by a neural network from tweet image, this information is stored on Udacity server. The last source is extra tweets information via Twitter API. After fixing the quality and tidiness issues of three datasets, they are combined to analyze the most common dog breeds, rating and dog stage, popularity of dog breeds and timestamp analysis.

Most tweeted dog breeds

Figure 1 shows top 10 common dog breeds. The largest percentage (more than 20%) of the tweets were predicted as not a dog breed. The possible reasons could be 1) images for these tweets were not dogs, or 2) neural network used for prediction has low accuracy which failed to predict the dog breeds. I quickly checked two images. It turns out the predictions were correct. Finding out the actual reason requires manually examine each of these imag. Due to the time limitation, I will ignore this issue in this report.

The most popular dog breed is the golden retriever (~ 7%), followed by the Labrador retriever, the Prembroke and chihuahua with very close percentage of around 4%-5%.

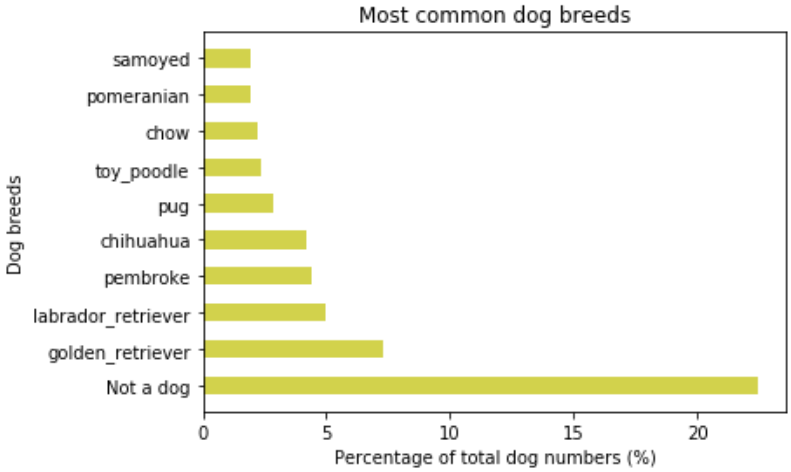


Figure 1Most common dog breeds

Rating and Dog Stage

The second question is whether WeRateDogs tends to give higher/lower rate to specific dog stage. I found that some dog stages only contain very small number of tweets. For example (doggo, floofer) and (doggo, puppo) only contains one data points, floofer only contains 9 data points. The potential problem is the small sample may not represent the distribution of population correctly. More data points are required to support our conclusion.

stage	tweet_id
	1831
doggo	75
doggo, floofer	1
doggo, pupper	10
doggo, puppo	1
floofer	9
pupper	224
puppo	24

A boxplot of rating numerator is used to examine the distribution of each dog stage. The plot clearly shows large outliers which needs to be removed first.
Figure 3 shows the boxplot without outliers. We can see that the pupper received the lowest rating and has the same distribution as the dogs without any stage.

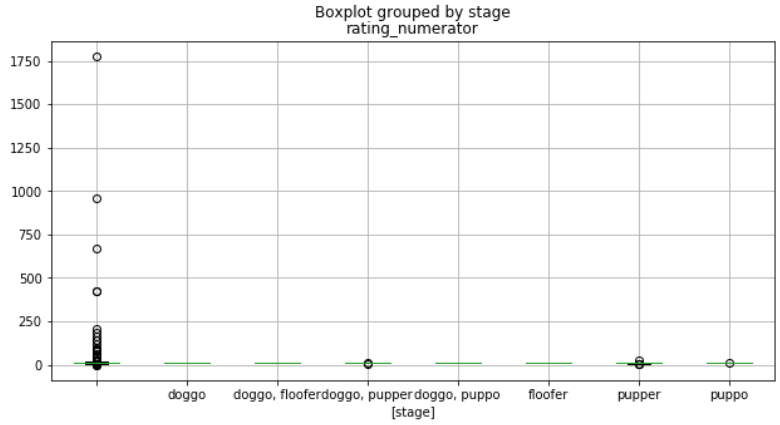


Figure 2 Boxplot of rating numerator of dog stages w/ outliers

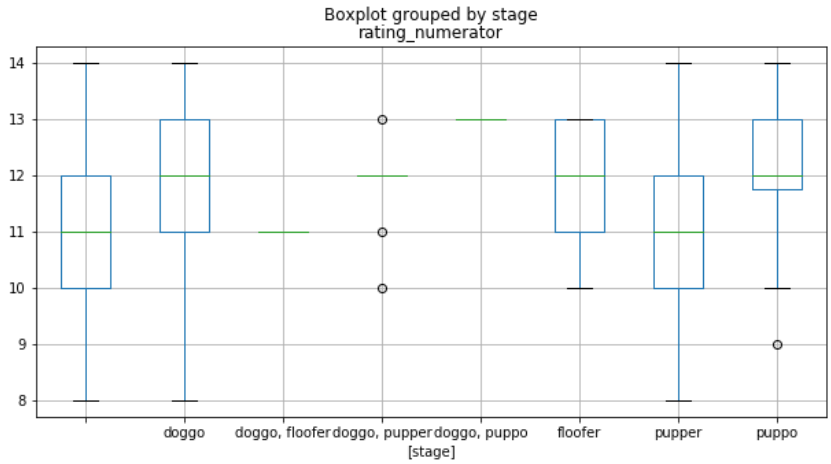
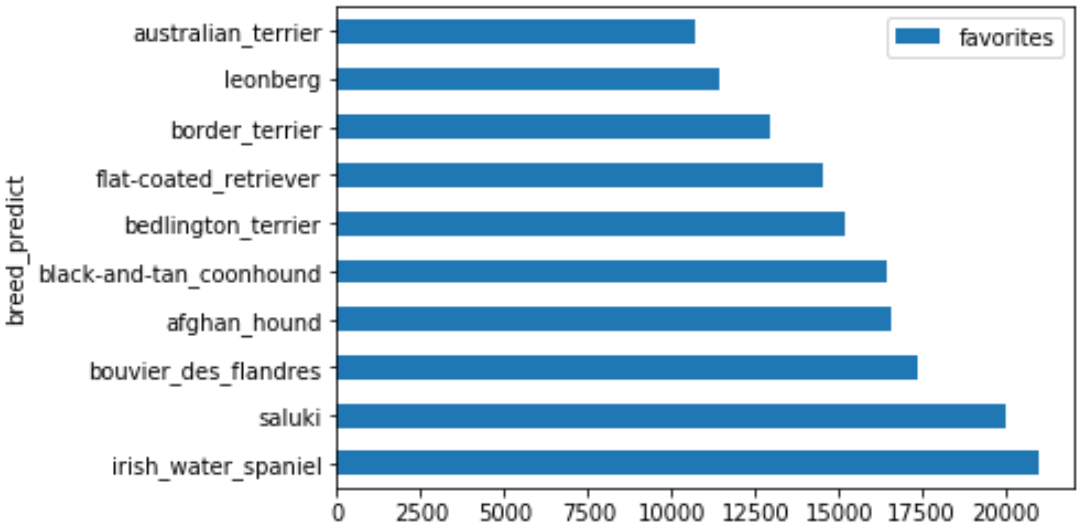


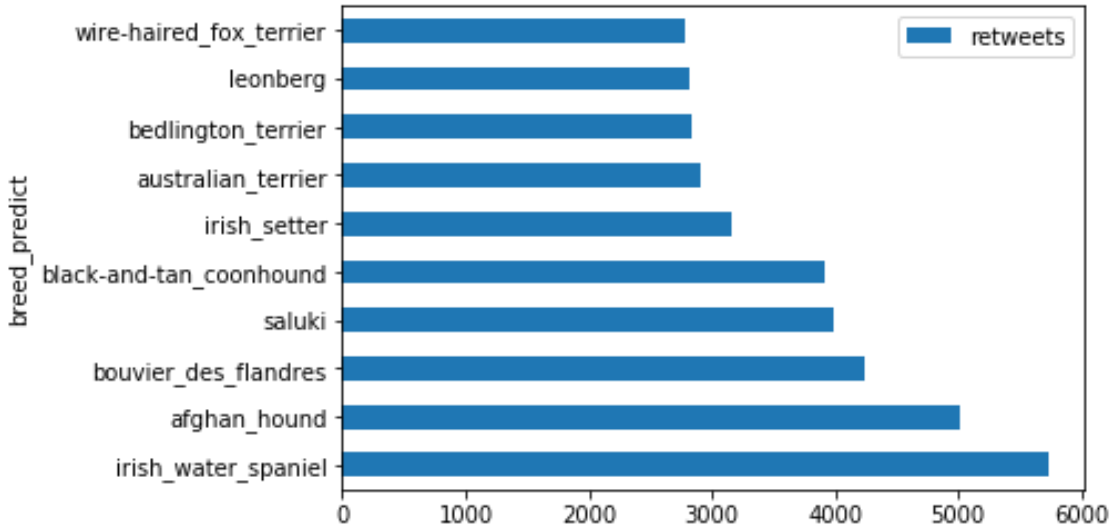
Figure 3Boxplot of rating numerator of dog stages w/o outliers

Popularity of Dog Breeds

Both favorite counts and retweet counts can be used to represent the popularity of dog breeds. Following two plots show the top 10 popular dog breeds based on median favorite counts and median retweet count. (Use of mean count leads to different ranking, but median is used in this analysis to avoid large outliers).

In both plots Irish water spaniel ranks the highest.





Analyze Timestamp

This section is to analyze timestamp of each tweet and when @dog_rates likes to send a tweet throughout the day.

From following figure we can see that @dog_rates actively tweeting at midnight and late afternoon, and never sent a tweet between 7AM to 11AM.

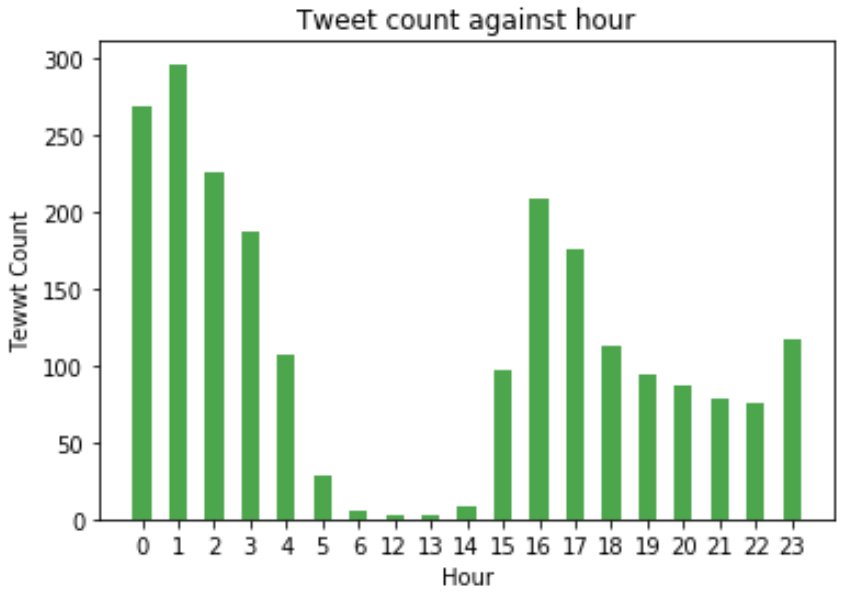


Figure 4 tweet count against hour