How WeRateDogs™ rates the dogs?

Project: Wrangle and Analyze Data

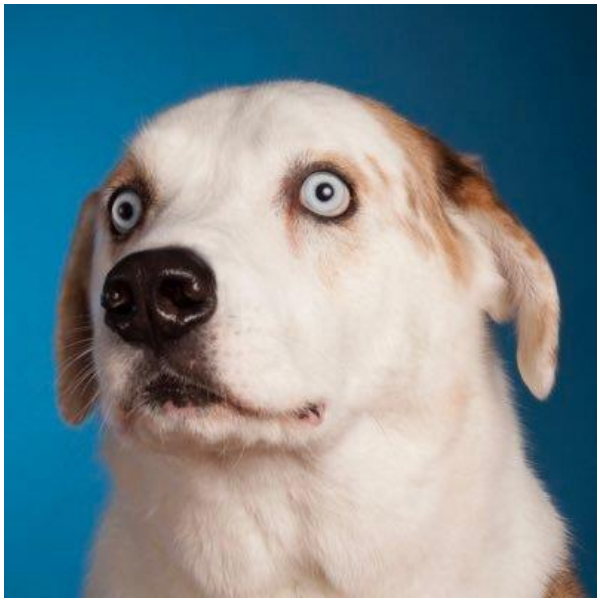


Photo from WeRateDogs™

By Nattacha Paksung

HOW WERATEDOGS™ RATES THE DOGS?

INTRODUCTION

This project aims at wrangling Twitter data to create interesting and trustworthy analyses and visualizations. The dataset is the tweet archive of Twitter user @dog_rates, aka WeRateDogs™. WeRateDogs™ is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings always have a denominator of 10, but the ratings are always greater than 10, i.e. 11/10, 12/10, 13/10, etc.

DATA WRANGLING

First thing to do in this project is "Data Wrangling", which consists of "Data Gathering", "Data Assessing", and "Data Cleaning". Oftentimes, the raw data is rarely clean. Data wrangling is a process to prepare datasets for easy access and analysis.

Data from WeRateDogs™ was gathered by query the Twitter API for each tweet using Python's Tweepy library. In addition, data from every image in the WeRateDogs™ Twitter archive ran through a neural network that can classify breeds of dog was also used.

The datasets were assessed, cleaned, and stored in prior to analysis.

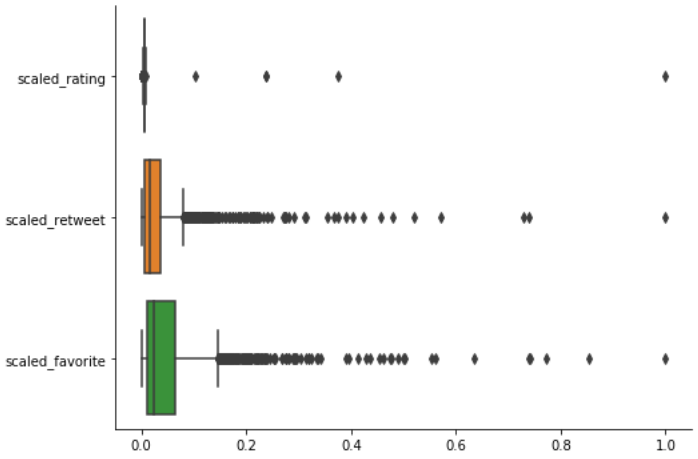
DATA ANALYSIS

The uniqueness of WeRateDogs™'s rating system is that numerators are often larger than denominators. Can the rating even tell how popular each dog is?! In this part, let's find some interesting insights from this fun dataset.

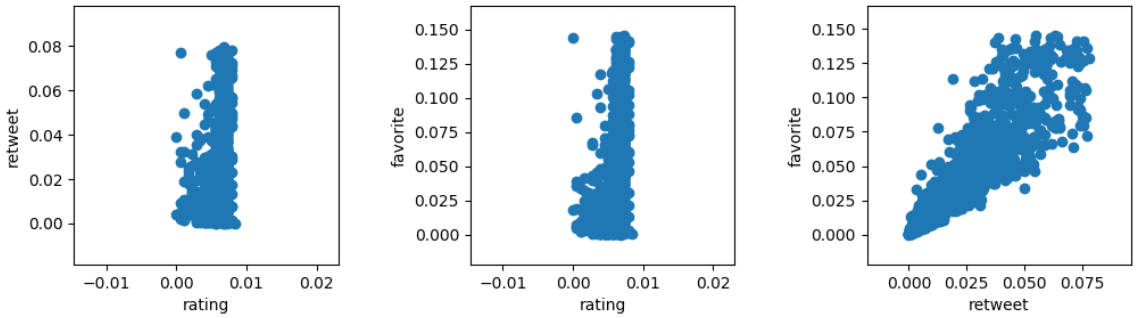
Question 1

First, it is questionable that the rating system is only for fun or it can really be a measurement for the dogs' popularity. Numbers of retweet and favorite should reflect more realistic popularity than the ratings. Let's see how the rating actually is when comparing with numbers of retweet and favorite!

These three variables have a large difference of mean values. Feature scaling is used here to standardize the values into a range of [0,1].



The distribution of all of measurement is highly right-skewed. The outliers on the right side should be removed. After removing scaled values of rating were plotted against scaled numbers of retweet and favorite.



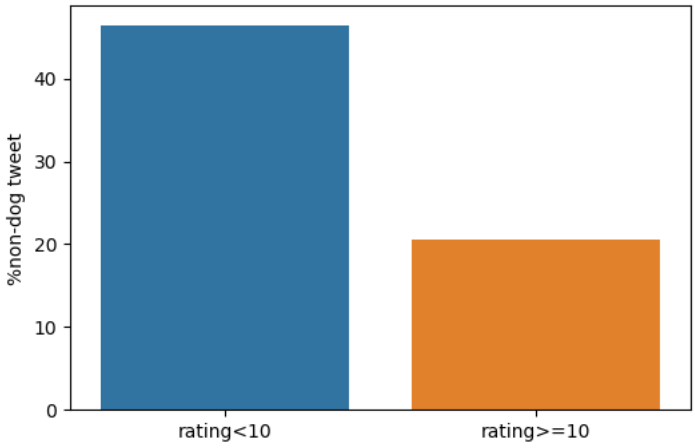
The first two scatterplots reveal that there is no correlation between rating and retweet count nor favorite count. Thus, we can confirm that WeRateDog™ rates dogs just for fun. The rating should not be the measurement to evaluate the popularity of dogs.

In addition, positive correlation can be seen from the last scatter plot. It is also noteworthy that people tend to retweet rather than mark favorite as there are many retweets without favorite count.

Question 2

So, what the ratings tell? Most of the rating numerator is larger than the denominator. What makes WeRateDog™ do it this way?

The data was divided into two groups according to the rating. The ratings lower than denominator (10) and the ratings greater or equal to 10. I looked into details of the low-rating group's data. I found some clues in the tweet's text! For example, text in row 730 says "*Who keeps sending in pictures without dogs in them? This needs to stop...*". It might be a lot of tasks to open the url of image to prove it one by one. Instead, let's count the prediction by neural network if it predicts a dog or not!

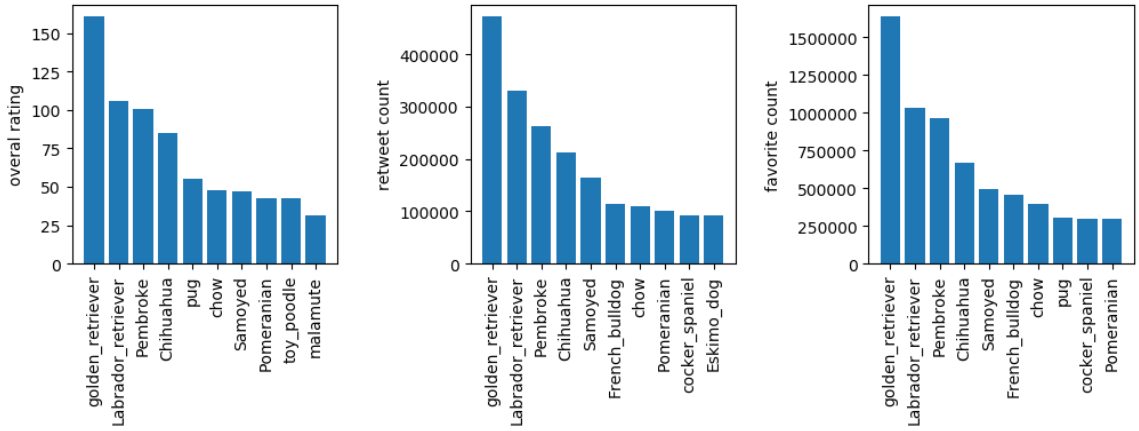


It was found that there was around 47% and only 21% of images that are not dogs in the low-rating group and the high-rating group, respectively. **So most likely, WeRateDogs™ rated images that are no dogs in them under 10.**

Question 3

What are the most popular dog breeds in WeRateDog™ dataset?

The measurements to evaluate the popularity here are overall rating of all tweets, total counts of all retweet and total counts of all favorite marks of the same breed.



With all three measurements including overall rating, retweet counts, and favorite counts, the results show that golden retriever, labrador retriever, and pembroke are the top three most popular dogs from this dataset.

CONCLUSION

From data analysis, some insights regarding WeRateDog™ dataset were found:

- *WeRateDog™ rates dogs just for fun. The rating should not be the measurement to evaluate the popularity of dogs.*
- *Either retweet_count or favorite_count would give similar interpretation when using them to measure the popularity of the dogs.*
- *It is likely that WeRateDogs™ rated images that are no dogs in them under 10.*
- *Golden retriever, labrador retriever, and pembroke are the top three most popular dogs from this dataset.*