WeRateDogs: Wrangle and Analyze Data

Introduction

This article is the result of the WeRateDogs project, from the Data Analyst Nanodegree course by Udacity. During this module, I have learned how to wrangle data, and it became clear to me, thanks to the teacher's insistence, the 3 main stages of data wrangling: gather, assess and clean and all these phases are iterative, so the life of the data wrangler could be simplified as gather, assess, clean, repeat!

To test the skills learned, I had to analyse three datasets associated with the WeRateDogs Twitter account. This account basically scores dogs with an unconventional scoring system, in which the numerator is almost always higher than the denominator. In addition, Matt Nelson, the owner of the account, gives a touch of humour with witty comments about the dogs.

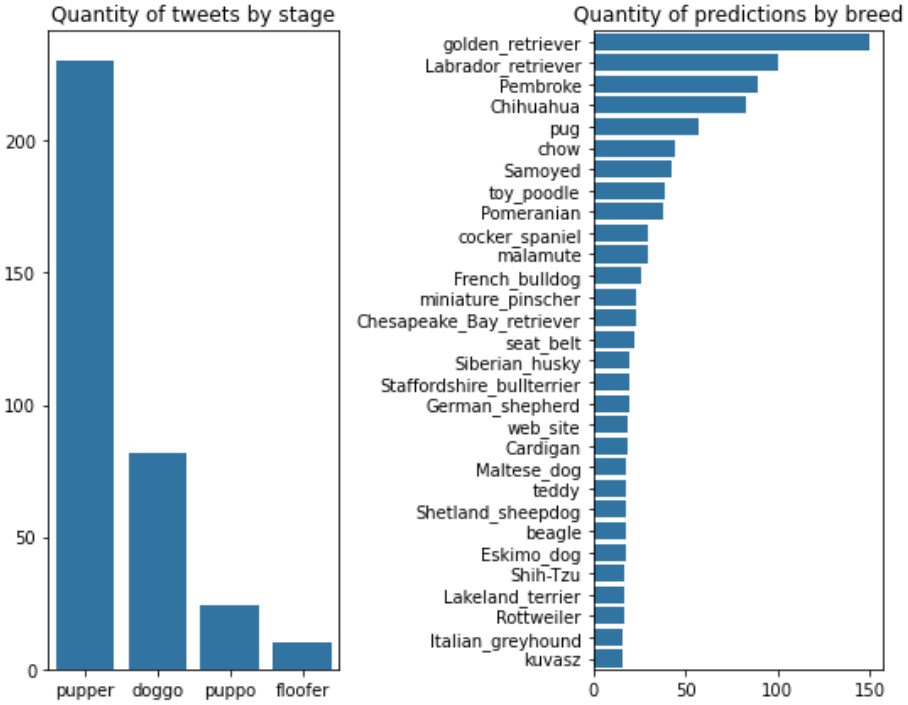
Picture 1.- Profile of WeRateDogs

Okay, but you may be wondering how I got this information, and even if you don't, I'm going to tell you anyway. For me, that was the most difficult part of the project, since I had to learn how encodings work, how various types of files are internally, how to communicate with APIs and how to do web scraping. Challenging. But once the hard work is done, the rest of the way has been lighter. And once I had all the nice, clean data, all I had to do was to ask questions and answer them.

Which dogs are the most popular?

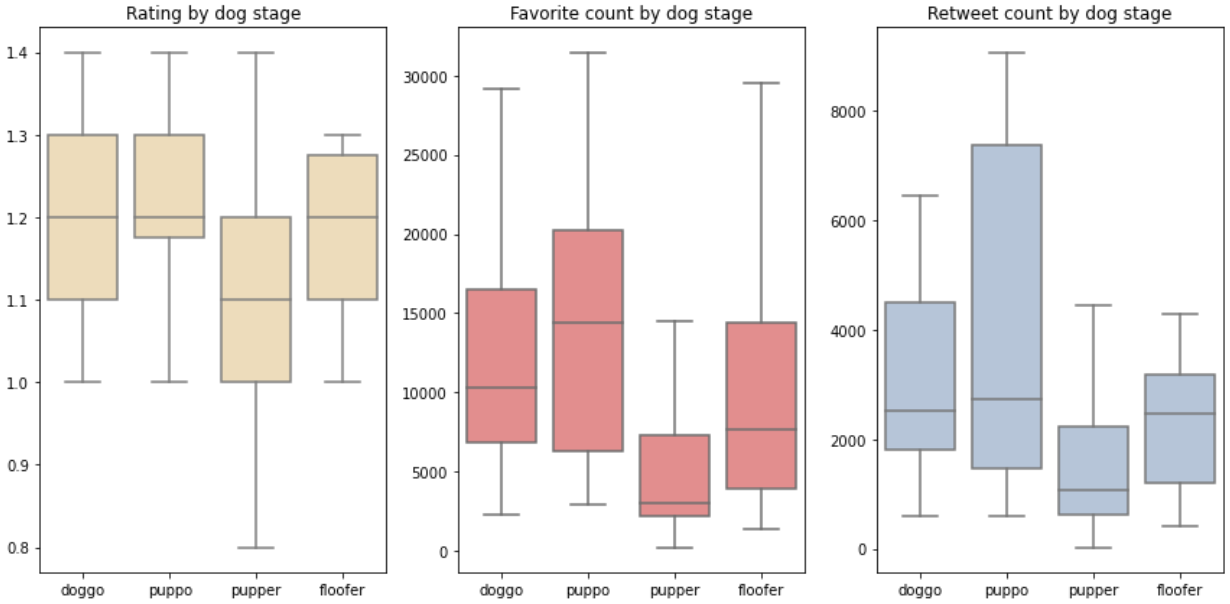
The first question that came to my mind was: Which kind of dog does Matt Nelson like the most? And the rest of humanity?

In order to answer it, I have created two graphs. One of them shows the total number of tweets according to the stage of the dog. The other shows the number of predictions made by a neural network classifier of dog breeds, which is directly related to the number of tweets of each breed, or at least it would be if the classifier was good enough, but I leave this for another point in this article.



Graph 1.- Quantity of tweets by dog stage (left) and by breed (right)

Besides the number of tweets, I found it appropriate to examine Matt Nelson's rating, the number of retweets and the number of favourites according to each dog's stage. For this visualization, I chose a box plot, since it shows in more detail how the distribution of each of the variables is and, to keep the graph within the paper's limits, I eliminated the outliers. Most of the tweets have a numerator lower than 20, but some graceful ones, have a much bigger numerator, like Snoop Dogg, which has a rating no less than 420/10.



Graph 2.- Boxplot of dog stage vs rating (left), favourite count (centre) and retweet count (right)

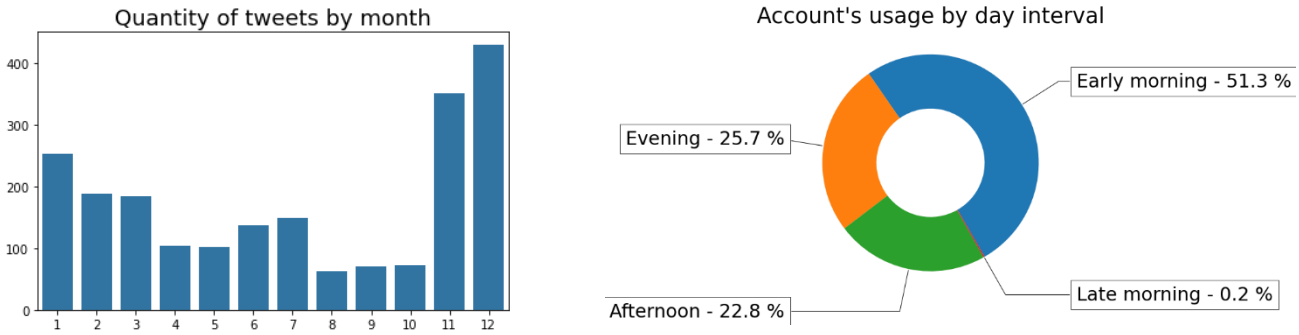
It seems that pupper and golden retriever are the most published stage and breed, respectively. However, the data shows that among the Twitter users, the most appreciated stage is the puppo, since it has a greater number of favourites and retweets, closely followed by the doggo stage. If we take into account the rating given by WeRateDogs, the only stage with a different median to the others is the puppy stage, with a rating of 1.1, slightly lower than the other three stages, with a rating of 1.2 approximately.

Also, it can be observed in the graph on the left that there is a prediction of website that appears almost 20 times, this small problem will also be approached at another point later on.

What kind of person is Matt Nelson?

One issue I find interesting is the differences in work schedules between people, while some are much more productive during the day, others are much more efficient if they start working late in the afternoon, with a huge range of possibilities in the middle. So, one of the questions I found it useful to ask is: What time does Matt Nelson normally publish his tweets?

In addition to variations in hourly productivity, there are also seasonal variations. November is the worst month for some people, while others are slower at the beginning of spring. To quantify WeRateDogs' productivity per month, I have quantified the number of tweets over 3 years that it has posted in each month.

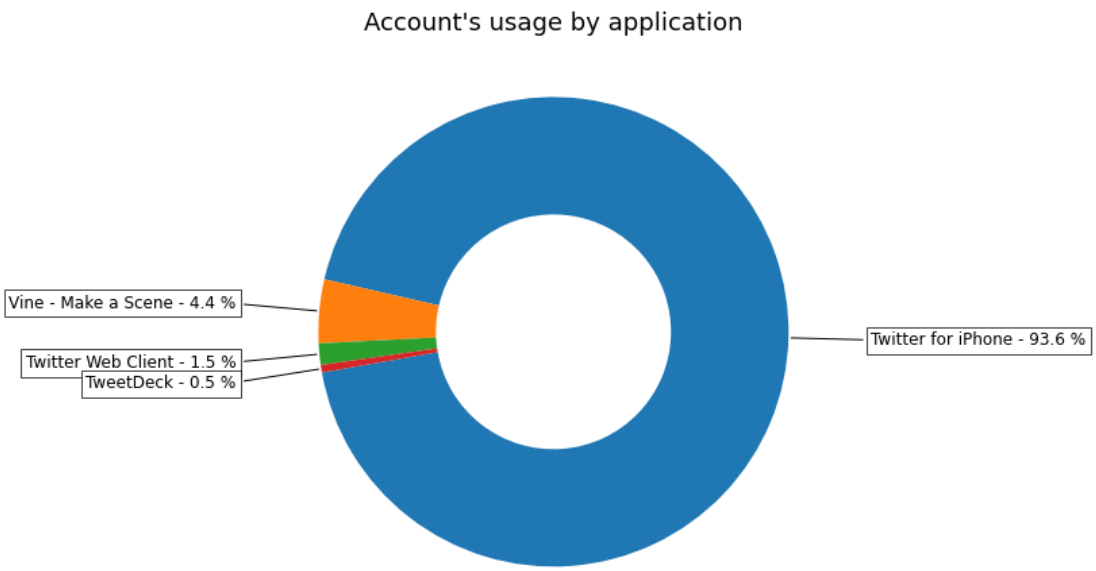


Graph 3.- Quantity of tweets by month (left) and Account's usage by day interval (right)

From the graph 'Account's usage by day interval' it can be inferred that Matt's publications have been made more than 50 percent of the time between 00.00 and 6.00 which suggests that he is a night owl.

The graph 'Quantity of tweets per month' shows that WeRateDogs has published many more tweets in the autumn and winter months than in the summer or spring months, it seems that with the cold many of us tend to stay at home.

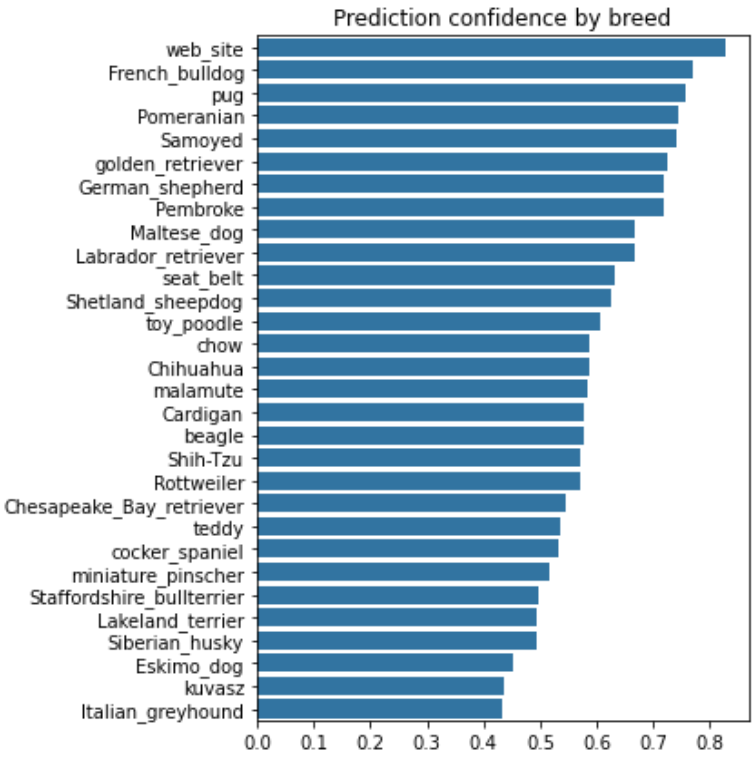
The next question that came to me was: Through which platforms are the tweets of this account published? And it seems that our friend is an Iphone fan since almost 94% of the tweets have been made from the Twitter App for Iphone. The remaining 6% of the tweets have been published between Vine, some web browser and TweetDeck.



Graph 4.- Account's ussage by application

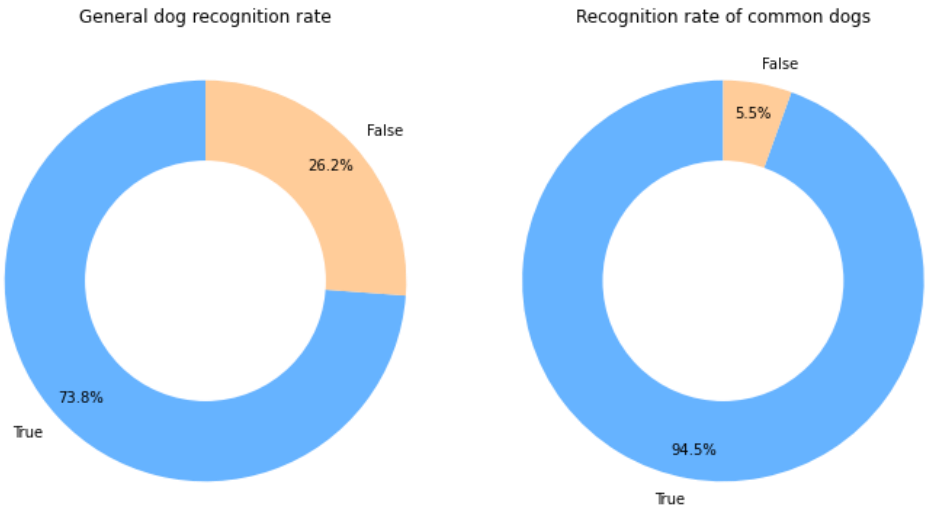
Is the neural network that Udacity has used to classify dog breeds efficient?

Well, it depends on the breed of dog we are talking about. The best prediction occurs with breeds such as: French Bulldog, Pug or Pomeranian, with a success rate of almost 80%, this may be due to the fact that these breeds are among the most published, although it should be studied more thoroughly. However, in breeds such as Kuvasz or Italian Greyhound, it barely exceeds 40% efficiency.



Graph 5.- Prediction confidence by breed

Moreover, the best prediction is not given with any dog, but with websites! One problem that this neural network has is that it is not able to recognise dogs if the image received is a screenshot. This problem has made me ask myself the last question of this article, how many times does the classifier recognise if the image corresponds to a dog?



Graph 6.- Dog recognition by the neural network classifier

I regret to inform you that the answer is: it depends. In the case of the 30 most published dog breeds, recognition is quite efficient, since the neural network concludes that the images correspond to dogs 94.5% of the time. However, for any type of dog in general, this efficiency falls to 73.8%, which is not bad either, it still does better than me.

Given this data, I can already answer the question I left open at the beginning of this post. Is it wise to assume that the number of predictions per race is directly related to the number of publications? I would say so, although the quality of the classification is not perfect, the neural network makes it quite decent.