Data Wrangling Project: **WeRateDogs** by Natália M. S. Costa

In this report, I discuss some of the findings/insights of my data wrangling project. The content of the wrangled data is related to tweets of a twitter account named WeRateDogs. This account shares via tweets a comment about the dogs and the rating each one of them gets. Below I explain the results I obtained after gathering, assessing and cleaning the dataset.

One of the aspects I analyzed is how the favorite counts and the number of retweets are related to each other. As shown in **Fig.1**, first by looking at the plot on the left side, it seems that these two variables have a positive and linear relationship. By using the `stats` library I obtained the linear regression model, which is represented by the red line and has the following features: slope ~ 2.52, and intercept ~ 2060.06. With the slope and intercept we can predict the values for the favorite counts given the retweet counts. I also confirmed the strong correlation between these variables by computing the r-value, which is given by 0.93. This result indicates, for example, that a tweet that is highly retweeted will most likely be also highly favorited.

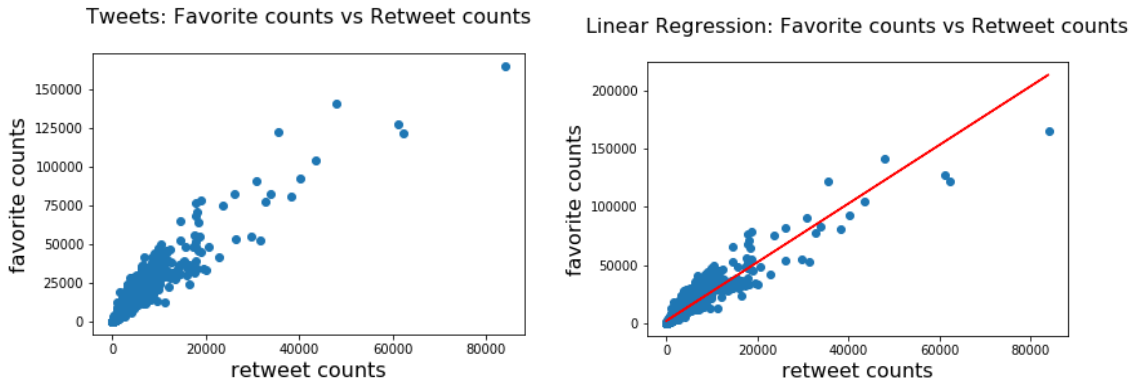

Fig.1: The blue dots in the scatter plot represent the observations (each one is a tweet). The red line in the right plot is a linear regression obtained by fitting the data in blue.

Another aspect of this scatter plot I investigated was the comparison between tweets (or dogs) that received ratings below or equal to 1.0, and greater than 1.0. Most of the tweets in the WeRateDogs twitter account use an unusual rating system in which the rating denominator may be smaller than the rating numerator. We interpret this as an indication that the dog is so amazing that it deserves a grade above the top possible. However, as shown in **Fig.2**, not all dogs were lucky enough to receive such higher grades. Therefore, I created two categories $\textbf{rating numerator} \leq \textbf{rating denominator}$ and $\textbf{rating numerator} > \textbf{rating denominator}$, where for the first the maximum value a dog can receive is 1.0, and for the second the minimum value a dog can receive is above 1.0. In **Fig.2** we can also see that most of the dogs that got ratings above 1.0 received more retweets and were more favorited than the dogs that got ratings below or equal to 1.0.

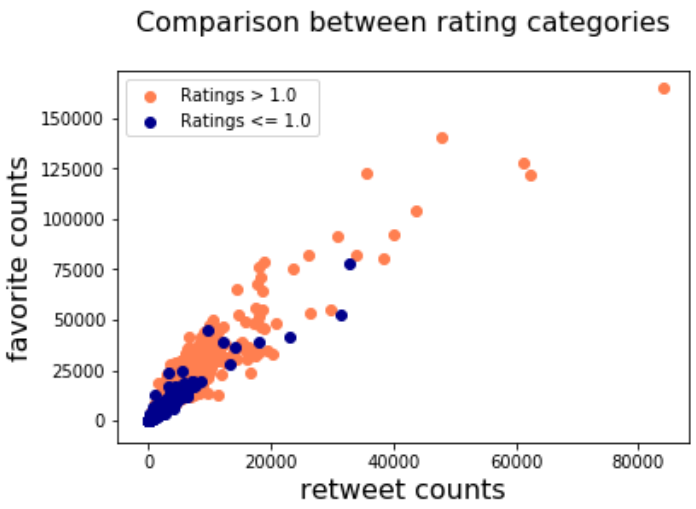

Fig.2: The dark blue and coral dots in the scatter plot represent the tweets where the dogs received ratings greater than 1.0 and below or equal to 1.0, respectively.

After spending some time on tweets texts to find and extract the names of dogs that had missing data, I got interested in knowing how people were naming their dogs. I plotted the pie chart below to illustrate the 10 most popular dogs' names in the dataset.

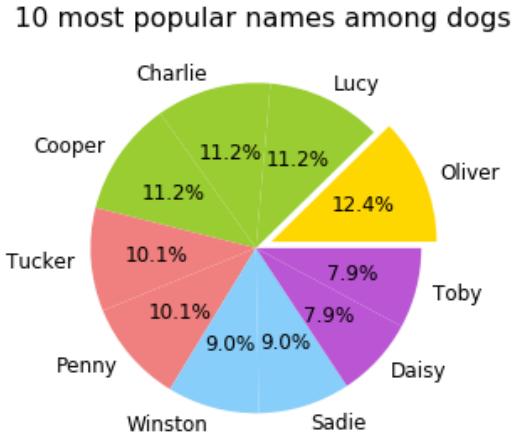

Fig.2: Pie chart with the 10 most popular names among dogs. The different colors are representing a different percentage, or proportion, of dogs with a given name among this sample of 10 different dog names. Oliver, for example, is the most common name for a dog in the dataset.

Besides dogs' names, we may also find in the tweets a briefly description them according to a 'dogtionary'. This dogtionary labels dogs as doggo, pupper, puppo and floofer. Basically, a doggo is an older dog that has more living experience; a pupper is a young dog still unexperienced; and the pupper is a dog that is transitioning between these pupper and doggo. Floofer refers to dogs that have excess fur. I checked how many dogs were in each of these stages, and the result is shown in **Fig.3**. We can see that there are more dogs in the pupper category, while floofer are quite rare in the dataset.

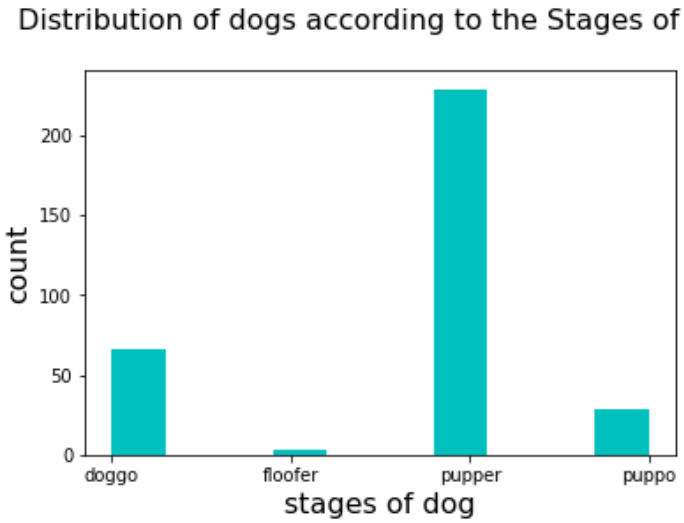

Fig.3: The different stages of dog and the number of them in the dataset. The number of pupper dogs is greater than the sum of the other three categories altogether.

In this report, I discussed some variables of the dataset and provided data visualizations to gain insights that help to understand a bit more about the twitter account WeRateDogs.