UDACITY

**Udacity Data Analyst Nanodegree**

# Wrangle and Analyze Data:

Analysis and Visualization

Written by:

**Anastasia Makarevich**

Date: March, 5, 2019

# Introduction

WeRateDogs (@dog_rate) is twitter account that rates dogs. A typical tweet contains a photo of a dog, a comment (often a humorous one) and some rating. Ratings are in the format X/10, where X is typically larger than ten.

The data was gathered from three sources: twitter archive and predictions dataset, provided by Udacity, and also data gathered with Twitter API. It took several days to fully prepare the dataset – identify invalid ratings, fix all formats and unify the data, so that it can be used for the analysis.

This exploratory analysis doesn't cover all possible summaries and variable relationships – I only looked at the most interesting ones. The visualizations were created with ***matplotlib, seaborn*** and **wordcloud** libraries.

# Does the @dog_rates owner have life?

The first thing I was interested in is whether the tweets are published uniformly – every day. And when does then owner of such popular account sleeps (if he sleeps at all)?

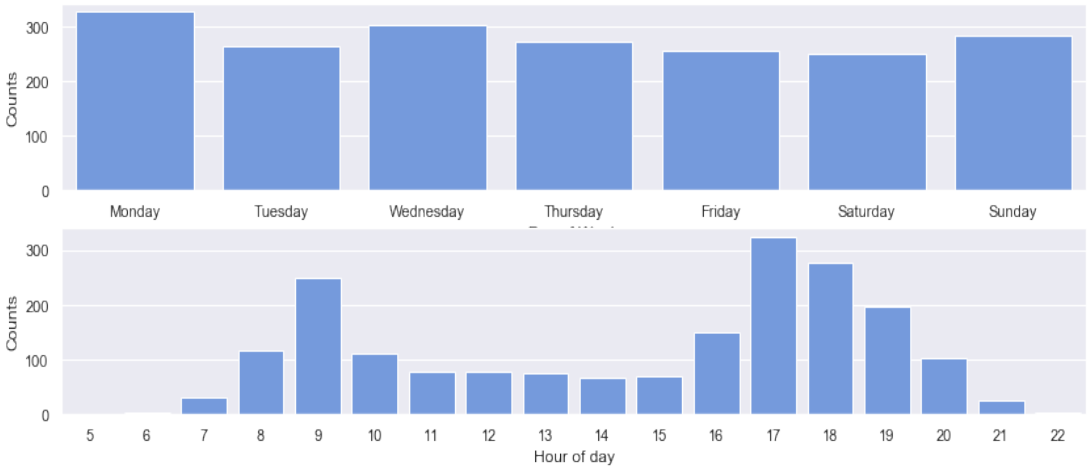


*Figure 1: Tweets Counts by Weekday, by Hour*

Based on the resulting bar plot we can conclude that Monday is the top day for the dog tweets – the author(s) start the week fresh, full of energy. Then, on Wednesday we see what looks like a second breath. The least number of tweets were made on Saturday – we can suspect that the authors… have life. :)

The second bar plot shows us that the author(s) not only have life, but also have some sleep. There are no tweets in the period from 11pm to 4 pm and only a tiny amount of tweets in the periods from 10pm to 11pm and 5pm to 6pm. The most productive periods are either morning (9am-10am) – probably before work and evening – from 5pm to 7pm.

An important thing to note here is that we would have been able to identify that pattern if we hadn't converted the timestamp to the proper timezone when cleaning the dataset.

## Most popular dogs names

Another thing that got me interested is whether we can identify the most popular dog's name. For that purpose I created a word cloud using **wordcloud** library. If we take all dog names found in these tweets, it will look as follows:

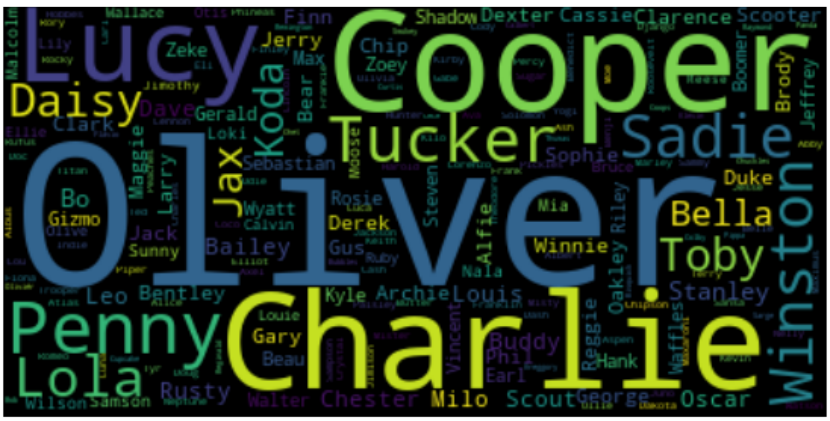


*Figure 2: Word Cloud for Dog Names*

On one hand, it looks like indeed there are some top names for the dogs. To be specific, from 1345 extracted names there are 10 dogs named Oliver, 10 – Cooper, 10 – Charlie, 10 – Lucy. All other names have less than 10 counts. So in general the dog owners prefer to be original and give their pets individual names.

## Typical Words to Praise Dogs

I also created the word cloud for the most typical words used to praise dogs – i.e., the most frequent words in tweets texts.


*Figure 3: Word Cloud for Tweets Texts*

When visually inspecting the data I noticed that some of the tweets contained expressions with "af". However, I didn't expect that it would be one of the top words! In fact, 140 tweets out of 1962 do contain "af". Some examples:

- "inspirational af"
- "paw-inspiring af" (2x)
- "intellectual af"
- "environmentally savvy af"

## Most popular dog breed

It was also interesting to check what are the most popular dog breed posted on this Twitter account.


*Figure 4: Most Popular Breeds*

Looks like retrievers are by far the most popular dogs (12% of all recognized dogs are either Labrador retriever or Golden retriever.

## Most popular rating

Since the whole Twitter account is about dog rates we couldn't avoid comparing the ratings:

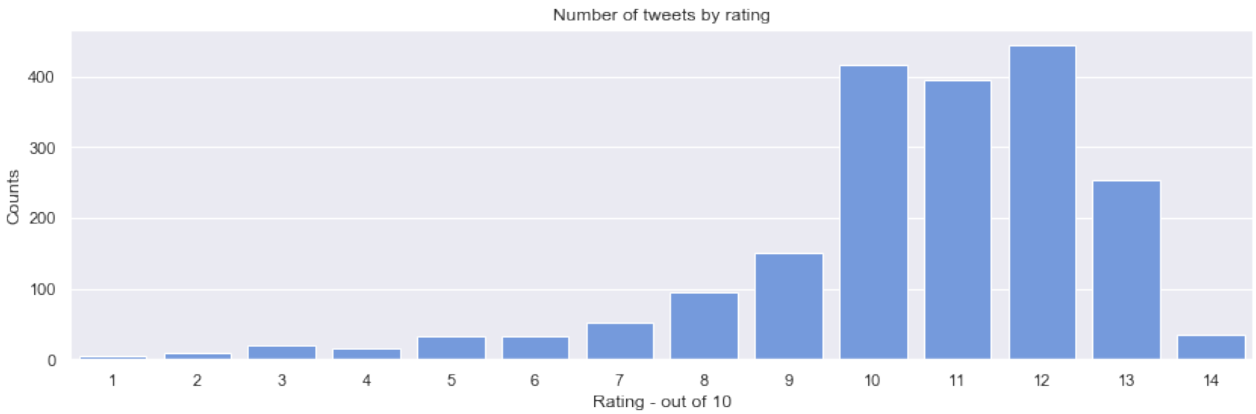


*Figure 5: Distribution of Ratings*

The bar plot covers the "typical" range from 1/10 to 20/10. Everything above that range are normally puppies where rating is scaled by the number of puppies (for example 88/80) or some special dogs (e.g. 1776/10 – the Declaration of Independence dog).

As we can see, most dogs are rated from 10/10 to 13/10. Because most dogs are "good dogs, Brent". Some low rating are due to the fact that it's not a dog at all, but we couldn't filter them out because the predictions were wrong in too many cases.

## Most Liked Dog Stages

Some tweets also contain information about the dog stage. It's one of doggo, floofer, pupper or puppo. It's important to note there, that only 328 tweets where the dog stage is explicitly named. For these dogs – what dog stages get more favorites on average?

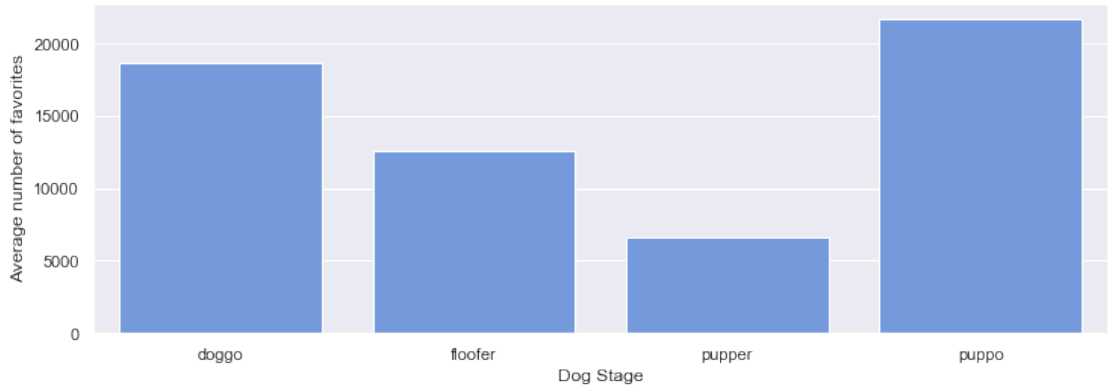


*Figure 6: Average Number of Favorites by Dog Stage*

We can notice that puppo is by far the most liked stage of the dog followed by doggo. However, we can't make any broad conclusions here given that only a tiny fraction of dogs have the stage identified.

## Strong but Useless Relationship

Finally, I decided to look at various relationships between variables, but the only strong one I was able to discover was the most obvious one – between the number of favorites and the number of retweets with correlation coefficient of 0.93.

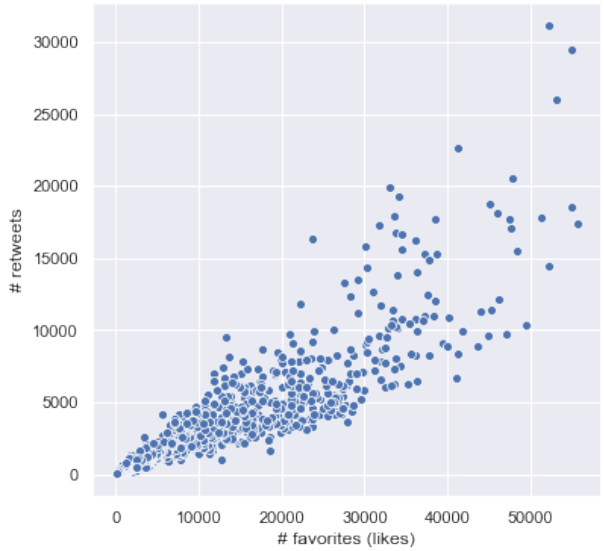


*Figure 7: Retweets vs Favoritees*

## Conclusion

In this analysis we did some basic exploratory data analysis that allowed us to understand the general characteristics of the majority of the tweets. We discovered that WeRateDogs twitter typically tweets in the intervals of 9-10am or 5-7pm and that Monday is the most productive day in terms of twitting.

We also found out that dog owner don't like to give their pets too common names, but when it comes to praising dogs, a lot of post start with "Meet", "look" and describe the dog with the characteristic ending with "af". And we also found out that most ratings are in the interval from 10/10 to 13/10.

Some ideas regarding future analysis can include trying to predict the dog stage both from text and from the picture. It can also be interesting to perform the sentiment analysis of the tweets, although the word cloud suggest that most of the tweets will have mostly positive sentiment polarity. Other text characteristics like text length, number of words are likely to be irrelevant (see my exploration of online news in Project 6).