# UDACITY – PROJECT 4 – DATA WRANGLE AND ANALYZE

Kardelen Oğlakcıoğlu

## Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson, and has received international media attention both for its popularity and for the attention drawn to social media copyright law when it was suspended by Twitter for breaking these aforementioned laws.

WeRateDogs asks people to send photos of their dogs, then tweets selected photos rating and a humorous comment. Dogs are rated on a scale of one to ten, but are invariably given ratings in excess of the maximum, such as "13/10". Popular posts are re-posted on Instagram and Facebook. In 2017, Nelson started a spin-off Twitter account, Thoughts of Dog.

My main goal was: gather, assess and clean these data to analyze it better. Data had missing or incorrect information such as incorrect ratings or missing breeds of dog etc. To find out most favorited breed or best rated dog, I firstly assess and cleaned this data. Finally I shared results below.

### 1. Most Favorited Dog


Most favorited dog is a puppo without a name! Owner stated that this puppo supported Womens March day in Toronto. The special day picture may have caused more interaction with people. There is a sign celebrating women's day on him. His breed was presumed to be the lakeland terrier. But he looked like a golden puppy to me.

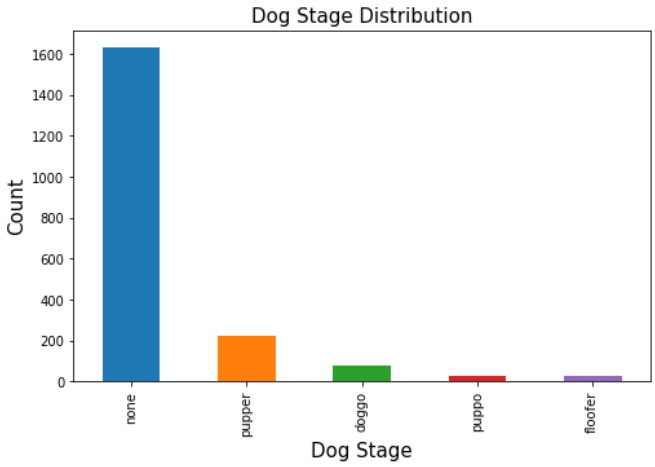
### 2. Most Retweeted Dog


Most retweeted dog is a doggo withot a name! He is swimming in a pool at first time and having fun! Owner shared a video of him. It is interesting to see dogs while they are swimming. So this may have caused more interaction. His breed was presumed to be labrador retriever.

### 3. Distribution of Stage Of Dog Category

It was interesting that people were tweeting with puns like the example in the tweet. There was hundred of tweets like below.

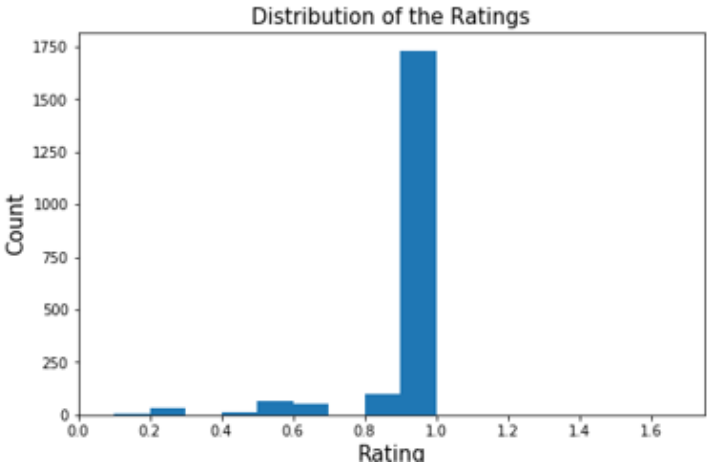
Distribution of categories is shared below. Pupper category has more tweets!

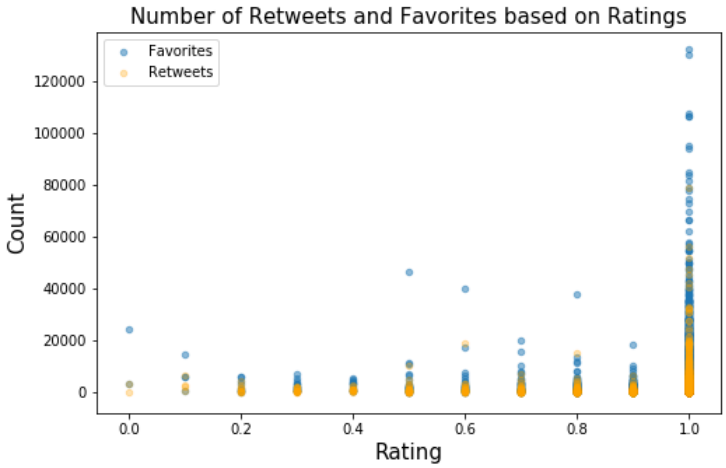




### 4. Distribution of Ratings

Even though the rating scale was specified between 0-10, people voted outside of this range. Example like 1275/3000. I assumed that if numerator/denominator is equal or more than 1, rating should be 10/10. If less, I scaled it between 0-10. First table shows that most of dogs are rated as 10/10. Really low percentage rated them less than 10/10.

Second table shows distribution of favorites(blue) and retweets(orange) regarding to ratings. Its obvious that 10/10 rated dogs have more retweet or favorite. But there are outliers for 0.5,0.6,0,8 ratings. Even though the owner rated his/her dog less than 10/10, the tweet was favored or retweeted more than expected.



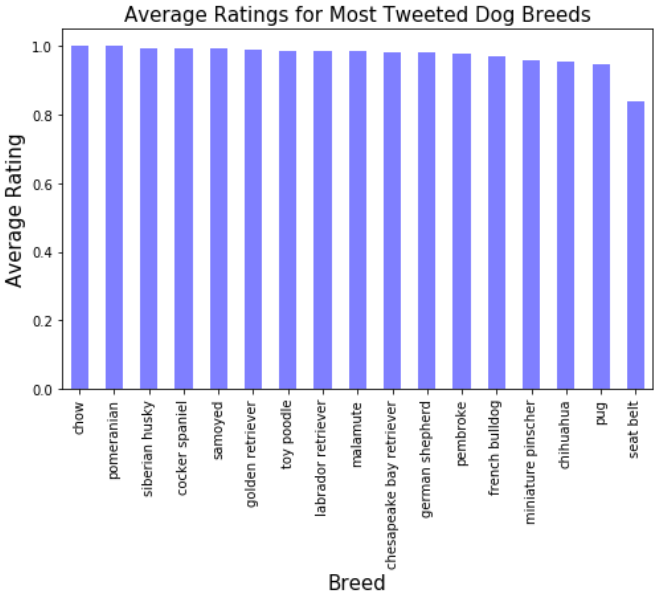



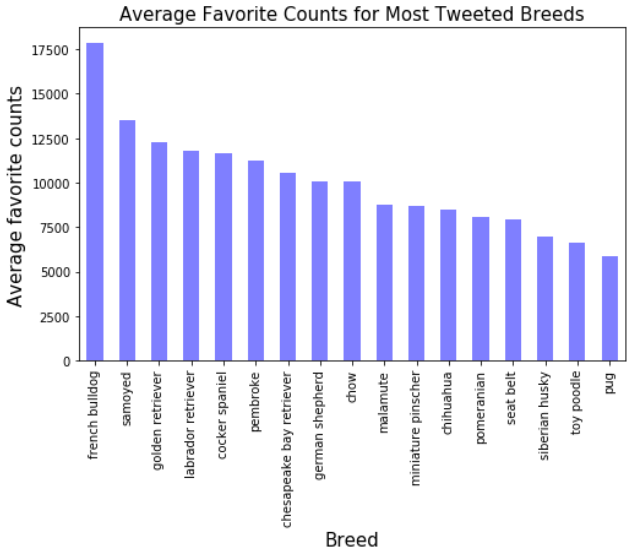
## 5. Distribution of Breeds

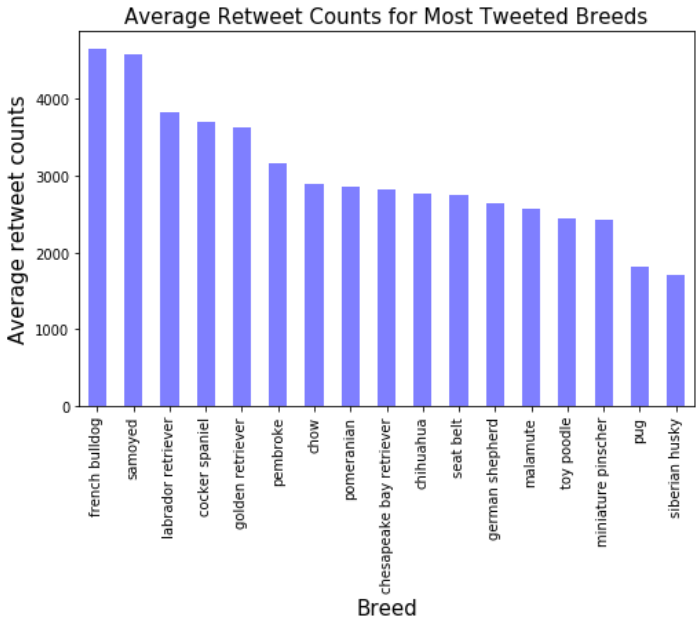
Breeds are predicted with machine learning algorithm. I assumed that it is true which has more confident level.

If more than 200 tweets were posted for a breed, the table of voting, favoritism and retweet numbers for these breeds is as follows.

First table shows that most of breed has 10/10 rating. Second table shows that most favorited breed is french bulldog!Third table shows that most retweeted breed is again french buldog!




**RESULTS**

It was interesting report for me to complete this project. Data was complicated and people usually did not follow rules to have clean data. There should be more tweets which are not even about dogs! But still report shows clear understanding for most favorited and retweeted tweets.

I tried to apply Twitter Developer account but it has been rejected by Twitter team. So I could not gather data from API.