

WeRateDogs

Data Analysis and Visualization



In this project we were interested in wrangling, analyzing and apply visualization techniques the tweet archive of Twitter user [@dog_rates](#), also known as [WeRateDogs](#) in order to extract more valuable information and insights. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

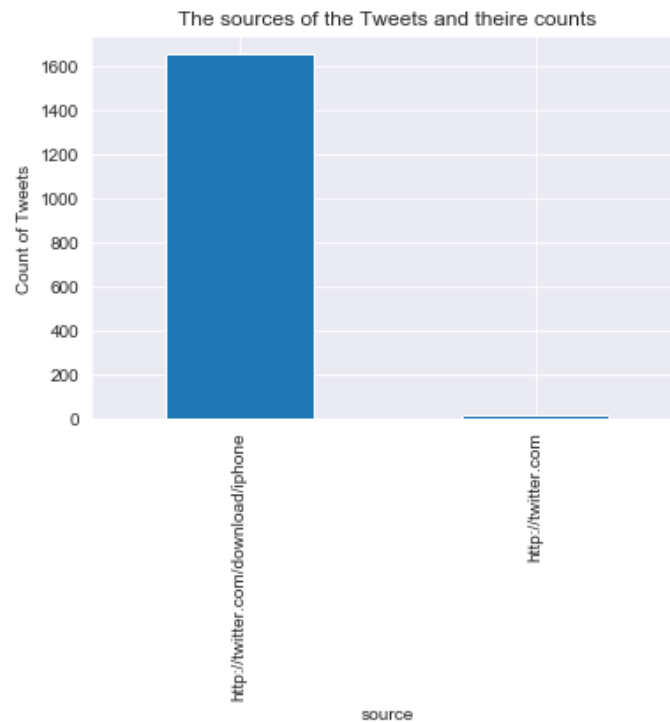
In this document we will clarify the efforts made in the analysis and visualization phase. The analysis is implemented on clean data.

We are interested in the following topics:

- The sources of the tweets
- Exploring numbers of ratings in the period of time between Nov, 2015 and Aug 2017. (in moth-year format).
- Exploring the Number of Ratings in Each year from 2015 to 2017.
- Exploring Number of tweets in the various dog stages.
- Exploring the Favorite numbers and retweetings among the various stages of dogs.
- The mean value of the confidence values of the first image prediction.
- Exploring the most frequently used names of dogs.

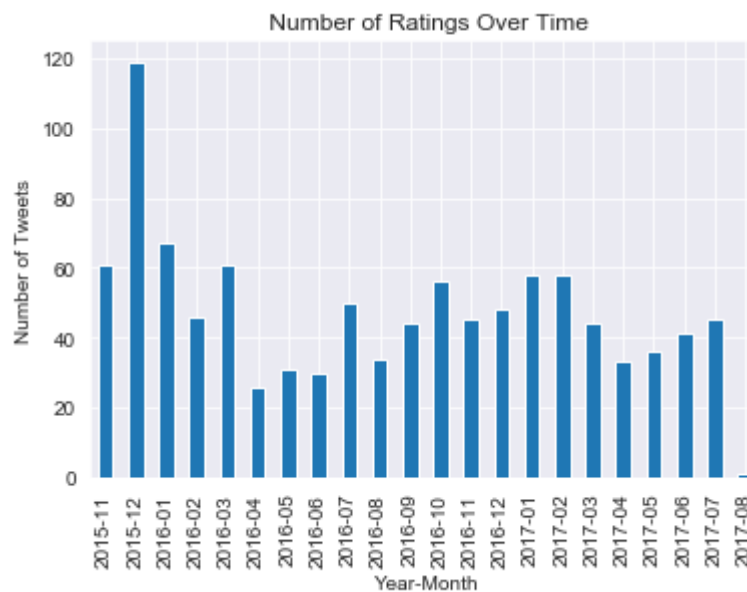
The analysis was made in order to extract information related to each of these topics and these information were visualized. In the following, the obtained figures are explored and conclusions are derived.

1) The sources of the tweets



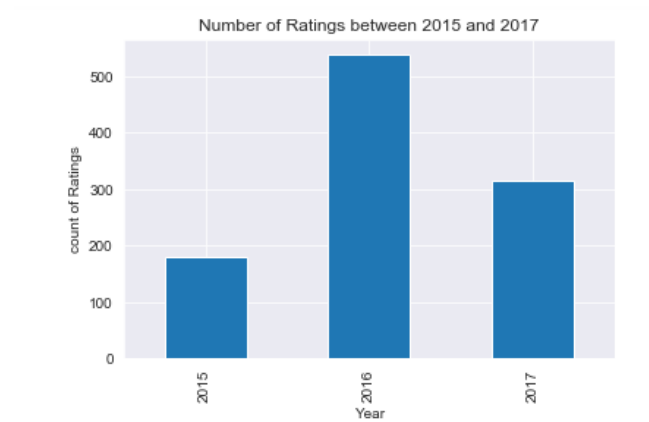
⇒ We see that the most frequently used source is <http://twitter.com/download/iphone>

2) Numbers of ratings in the period of time between Nov, 2015 and Aug 2017 in month- Year format.



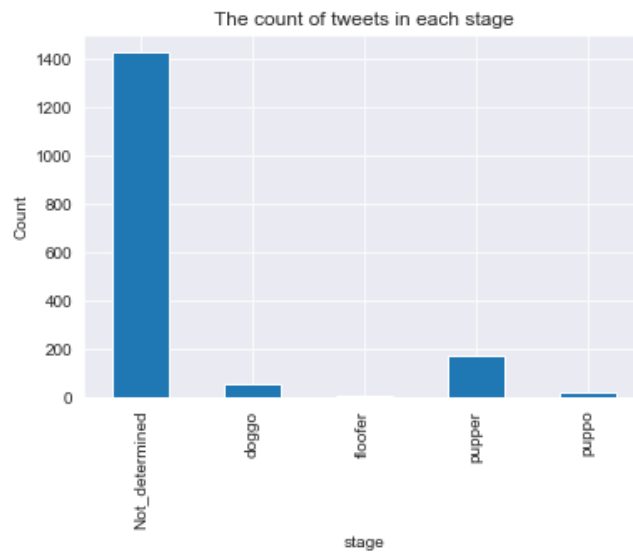
⇒ We can see that the largest number of ratings was in Dec 2015.

3) Number of Ratings in Each year from 2015 to 2017



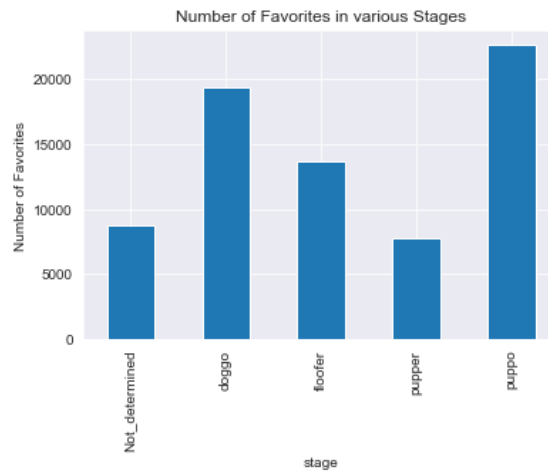
⇒ We see that year 2016 has the largest number of ratings

4) Exploring the number of tweets among the various dog sages:



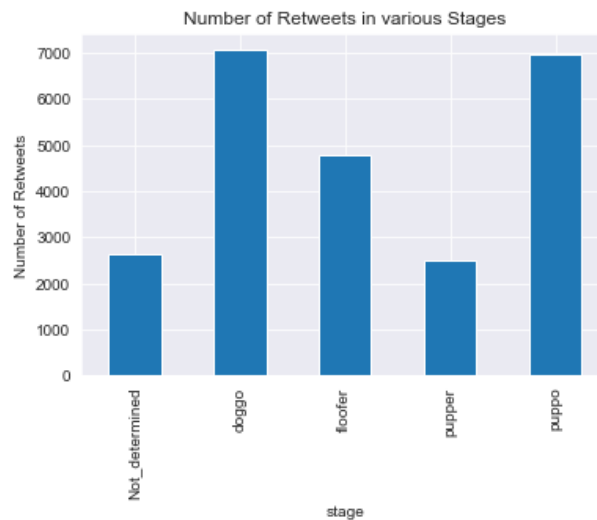
⇒ By comparing between doggo, floofer, pupper, and puppo, we see that pupper stage takes a larger number of tweets than doggo, floofer, and puppo.

5) Exploring the Favorite numbers among the various stages of dogs.



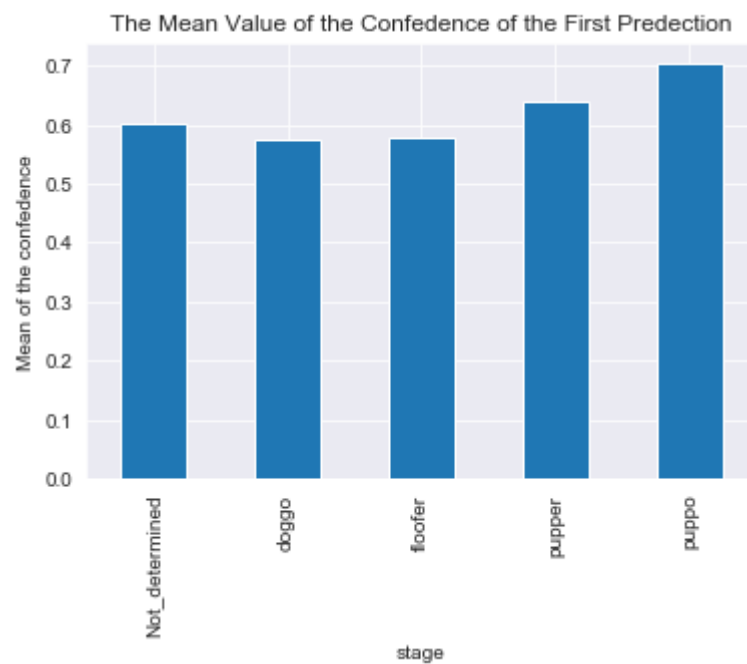
⇒ It is obvious that puppo stage has more favorites than doggo, but in general puppo and doggo dogs are more liked than the dogs in other stages. This can be assured using the retweeting counts shown in the next figure.

6) Exploring the Retweeting counts among the various stages of dogs



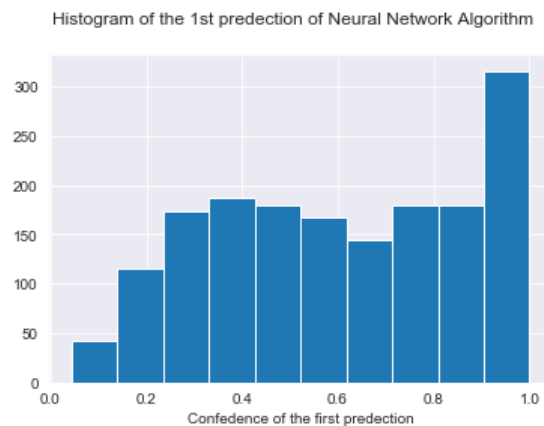
⇒ So we see that doggo and puppo stages have larger number of retweets than the other dog stages

7) The mean value of the confidence values of the first image prediction



⇒ We can see that the means of the first prediction confidence have approximately equal values among the various dog stages.

8) Exploring the statistical distribution of the first result of the prediction algorithm



⇒ We see that the statistical distribution (nearly) takes the shape of increasing exponential function. The performance of the prediction algorithm is enhanced with time due to the enhancements made in the neural network used in this algorithm. By time, I expect that the confidence of the first prediction will approach 1.

Additional insights:

- ⇒ **The mean of rating numerator = 11..237 with standard deviation (std)= 5.180, the maximum rating numerator equals 90.0 and the minimum is 2. We see that there are some outliers.**
- ⇒ **The total number of valid ratings= 1686 (after cleaning)**
- ⇒ **Values of the first prediction result of the prediction algorithm always has the largest confidence value (p1_conf)**
- ⇒ **The p1_conf has mean value 0.605 with std is about 0.266**
- ⇒ **The most common dog names are: A, Lucy, cooper, Charlie, and Oliver.**