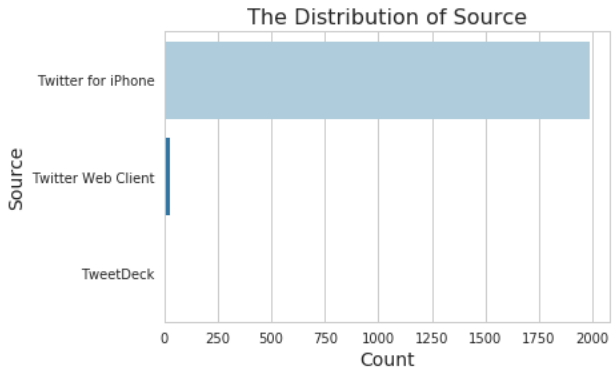
Analysis and Visualization

Introduction

The following analysis is based on the collection of three different sources: twitter_archive from @dog_rates provided as a .csv file, twitter API data and Udacity's neural network to classify the breed of dogs.

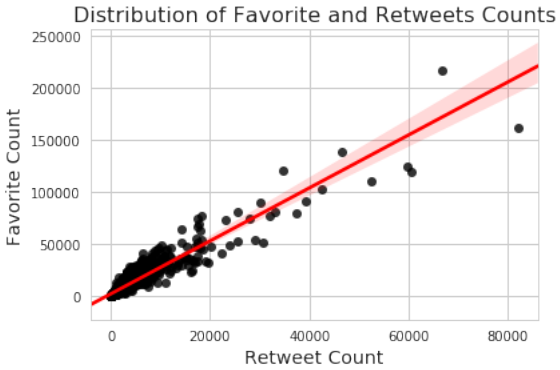
Distribution Source

The following plot shows the distribution of source, and illustrates that contributions from users to "We Rate Dogs" are mostly from the iPhone twitter app (98% in total), while the use of TweetDesk is pretty rare (less than 1%).



Distribution of Favorite and Retweets Counts

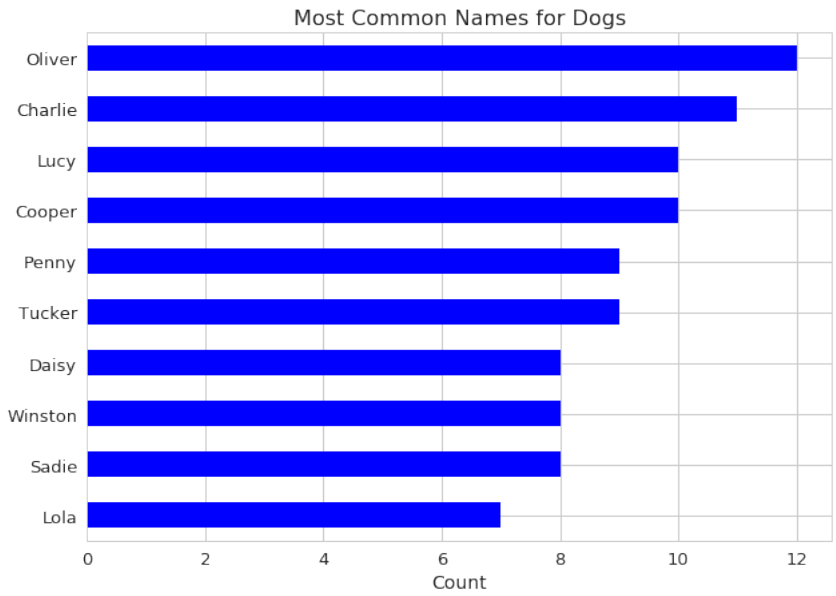
There is a strong positive correlation between 'retweet_count' and 'favorite_count'. The computed result between both variables is close to 0.935 using the "Pearson" method to obtain the standard correlation coefficient, which suggest that the most popular tweets usually get a larger number of retweets and favorite from users of "We Rate Dogs". The plot below confirms this hypothesis.



Most Common Names

According to our data, most owners seem to forget to add the name of their dog, some of those 605 'None' names had previously determiners "a", "an" or "the". However, for those who remember

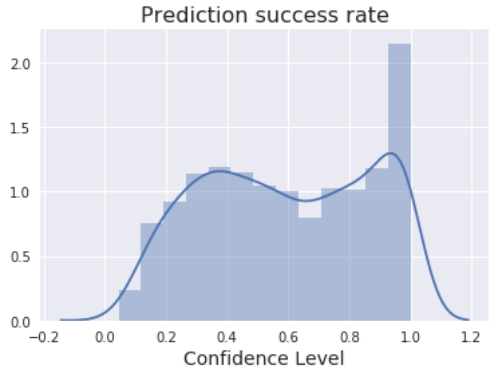
to add the name of their dogs, the data shows that "Oliver" is the most popular name, followed by “Charlie”, “Lucy” and “Cooper”.



Classification of Dogs

```
golden_retriever             139
Labrador_retriever            93
Pembroke                      88
Chihuahua                     79
pug                           54
chow                          41
Samoyed                       39
toy_poodle                    38
Pomeranian                    38
malamute                      29
cocker_spaniel                27
French_bulldog                25
Chesapeake_Bay_retriever      23
miniature_pinscher            22
seat_belt                     21
Name: predicted_breed, dtype: int64
```

The neural network from Udacity seems to suggest that "Golden_retriever" and "Labrador_retriever" are the two of the most common breeds of dogs, followed by "Pembroke" and "Chihuahua". However the plot shows that the entire distribution is dominated by amounts between 0.1 and 0.8, which could suggest that the model used still requires some work since its information might not be accurate.


Prediction success rate
2.0
1.5
1.0
0.5
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Confidence Level