Data Analysis and Visualizations

Introduction

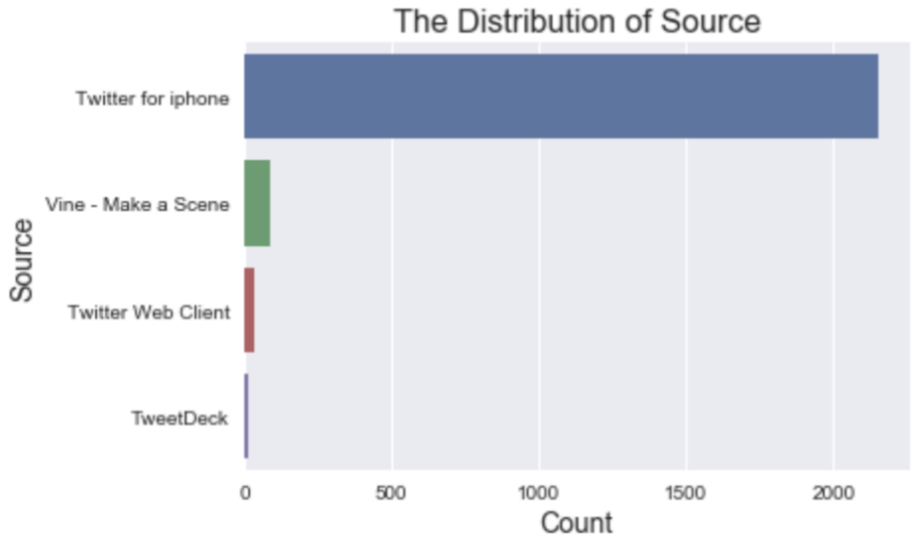
This act report includes the basic data analysis of WeRateDogs twitter account data from two datasets: 'twitter_archive_clean' and 'image_predictions'. It provides four insights from the analysis and visualization results.

Data Analysis and Visualizations

1. The Distribution of Source

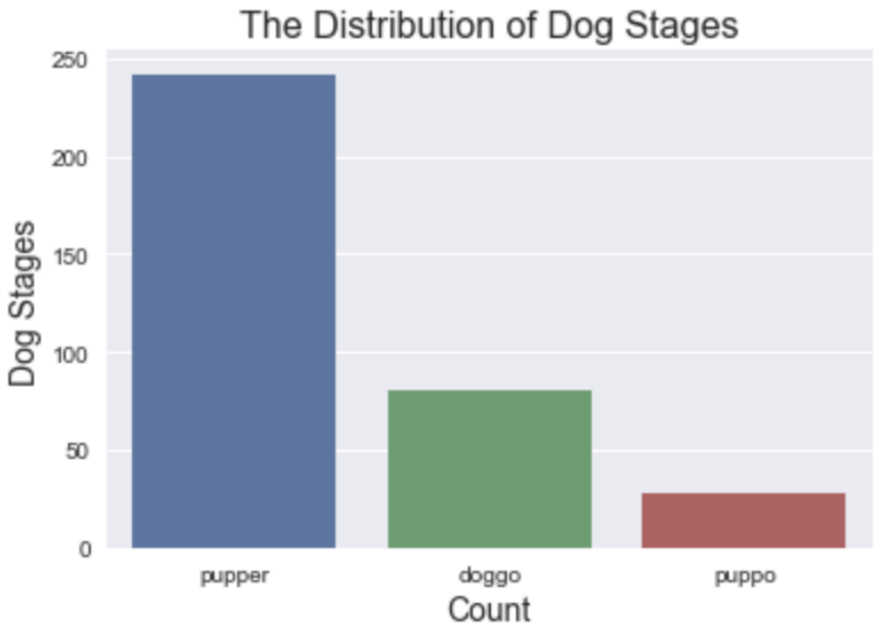
This plot above shows the distribution of source. We can see that the dominate source of tweets is from iPhone twitter app, which is 94% in the total. That means the twitter app is the main channel for people using to tweet, retweet, post, and others, while the TweetDesk is pretty rare (less than 1%).

```
Twitter for iphone     2152
Vine - Make a Scene    91
Twitter Web Client     32
TweetDeck              11
Name: source, dtype: int64
```



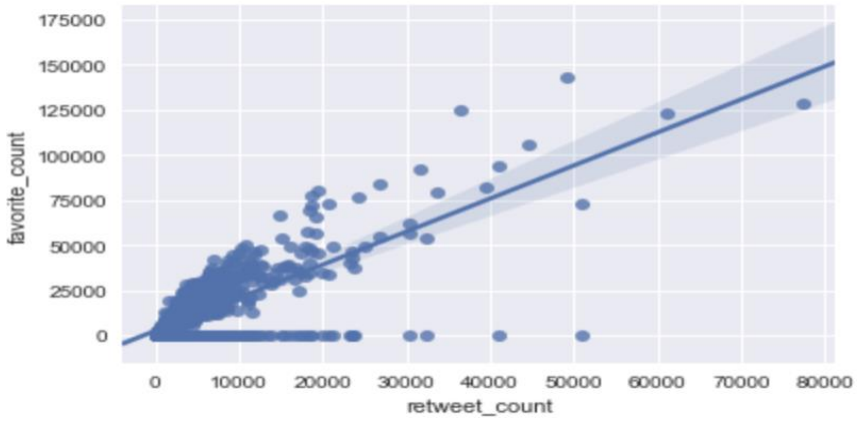
2. The Distribution of Dog Stages

Similarly, I check the distribution of dog stages. It shows that 'pupper' (a small doggo, usually younger) is the most popular dog stage, followed by 'doggo' and 'puppo'. It could be due to the young and unmatured dog is usually cuter than the adult dog. It should also be noticed that there's huge amount missing data in dog stages, thus the distribution may not reflect the truth.



3. Retweet_count and favorite_count Correlation

A reasonable hypothesis is that most popular tweets usually get a large number of retweets and favorite counts. I test the correlation between 'retweet_count' and 'favorite_count'. The pearson r^2 is 0.759, that is a high value showing a strong positive relationship between 'retweet_count' and 'favorite_count'. The plot below confirms this hypothesis.

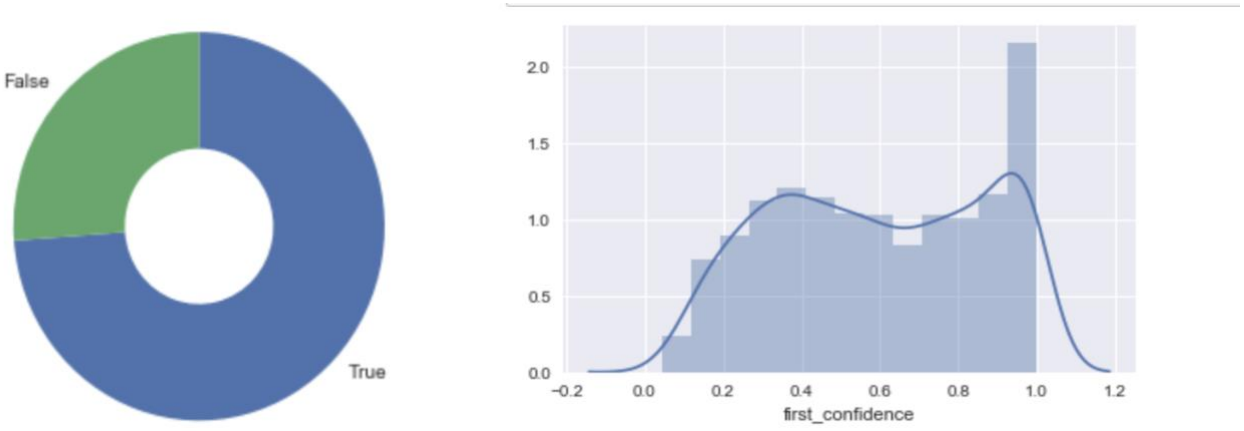


4. The Classification of Dogs Results Analysis

The 'image_predictions' table stores the result of a classification of dog breeds through a neural network. I am curious about the how this model works? What's the accuracy of this model? Therefore, I analyze and visualize the results in below.

Golden_retriever	150
Labrador_retriever	100
Pembroke	89
Chihuahua	83
Pug	57
Chow	44
Samoyed	43
Toy_poodle	39
Pomeranian	38
Cocker_spaniel	30

These breeds above are the top 10 dog breeds this model predicted. Golden retriever and Labrador retriever are top 2 and both over 100 predictions. It could be because those two are most common breeds in U.S. We have more image data on those breeds, and thus trained a better result.



The first plot above shows the prediction success rate of whether or not first prediction is a breed of dog. The pie chart indicates almost 2/3 situations the predictions are correct, even though this result is not good enough for a deep learning model. The second plot shows how confident the algorithm is in its first prediction. We can see 100% is the most cases, however the amounts of 0.1 to 0.8 dominate the entire distribution. That also could suggest that the model is not good enough.