Analysing and Visualizing WeRateDogs Dataset

Introduction:

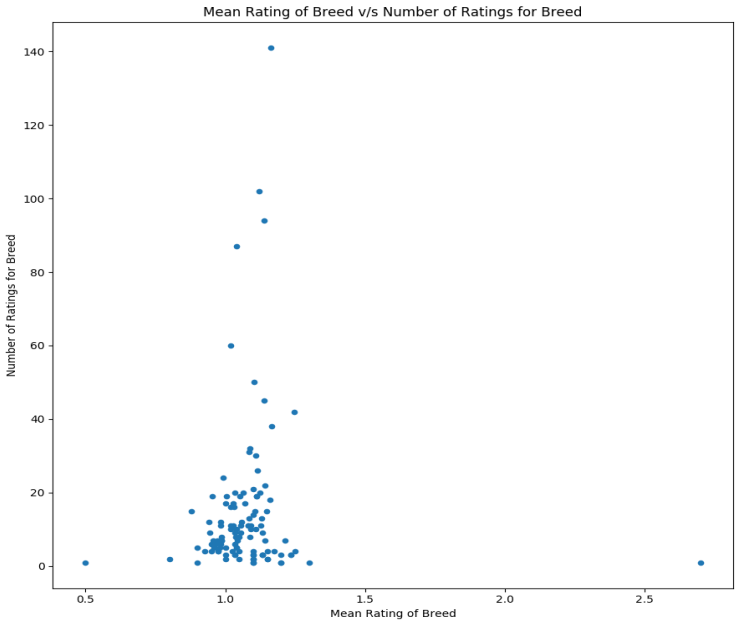
WeRateDogs is a twitter account that rates people's dogs in a humorous way. It was launched 3 years ago on 15th November 2015 and has over 7 million followers. The rates are calculated mostly having a denominator of 10.

With the help of Udacity and Twitter API a dataset is put into place to answer a few questions about dogs. What are typical ratings for dogs and what is the rating distribution for breed? What are various breeds of dogs that this dataset holds? The source of the tweets (medium used to tweet)? What is the Prediction confidence per breed?

Analysing and Visualizing :

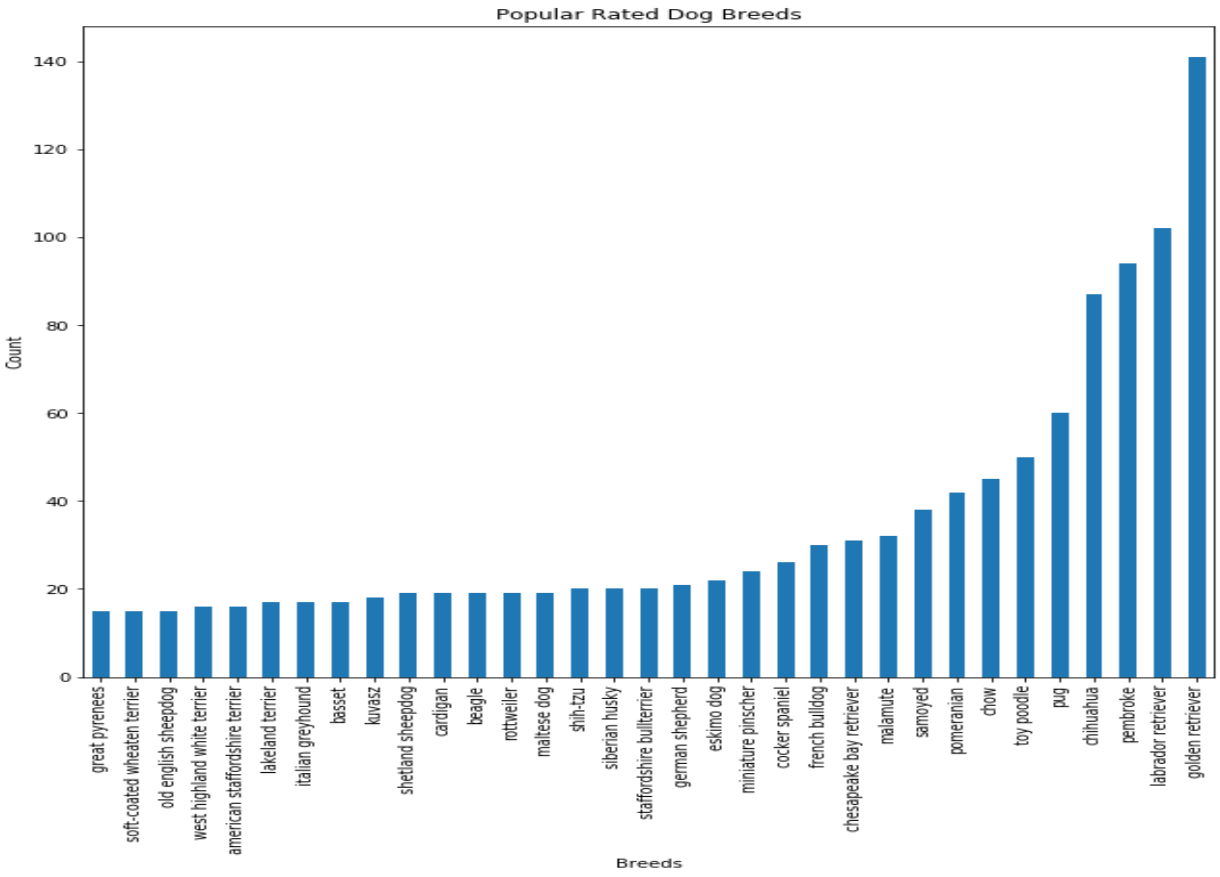
What are typical ratings for dogs and what is the rating distribution for breed?

In the chart below, we plot (per dog breed that is deduced by image-based prediction) the mean rating for that breed against how many breeds received that same rating (rating being the ratio numerator/denominator). We can see that most breeds received a rating around 1.0 (actually slightly more than one since the ratings seem to favour values more than 1.0 in the typical case). The number of breeds receiving ratings start tapering off as we start going away from 1.0, and the graph indicates a normal distribution. My guess is that if we had more data, then it would resemble more of a normal curve, but you can already see the shape in the current graph.



What are various breeds of dogs that this dataset holds?

I have restricted the breed count to be greater than 14 filtering out the lower numbers. It can be observed that golden retriever has the highest count. There are a bunch retriever, terrier, and sheepdog breeds in the more commonly posted dog breeds.



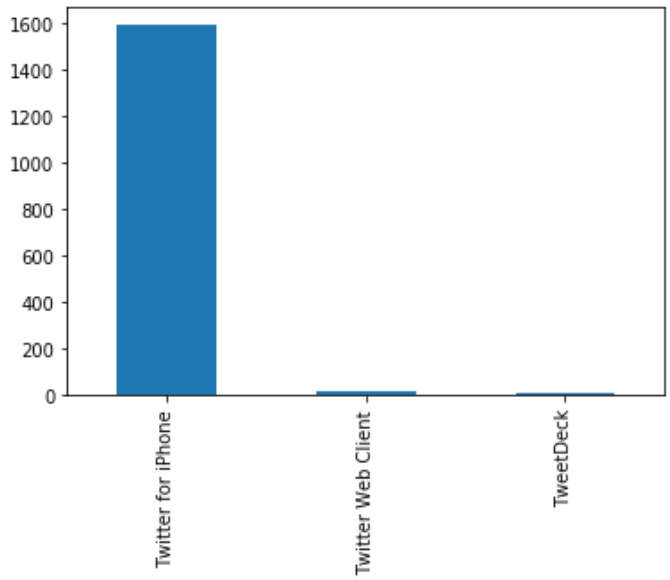
The source of the tweets (medium used to tweet)?

This one is interesting. I wanted to analyse the source of tweets, so I used the source column to get this data using reg expression. After merging three datasets, I observed 4 sources of tweets as shown below:

```
Twitter for iPhone     1978
Vine - Make a Scene      90
Twitter Web Client       30
TweetDeck                11
Name: source, dtype: int64
```

Once I filtered this dataset for ignoring records that were not null for jpg_url column the 90 records for “Vine – Make a Scene” were removed. This may be because the image_predictions file does not account for tweets made using Vine platform.

The chart showing the remaining 3 sources is shown below:



What is the Prediction confidence per breed?

We noticed that the image prediction dataset contains 3 predictions per image. We first cleaned the data to only keep the prediction for the breed with the highest confidence. Then, we plotted the prediction confidence per breed. I'm pasting in photos for the 3 breeds with the highest and 3 breeds with the lowest confidence levels of prediction to be able to see if we can visually determine any striking characteristics about the specific breeds which causes the prediction algorithm to work better or worse for specific breeds.

3 breeds with the highest prediction confidence

3 breeds with the lowest prediction confidence



Based on a cursory visual inspection of the breeds, there is nothing that stands out as particularly interesting, so this would require a more in-depth investigation. It's unclear if the prediction confidence is dependent on breed or on some other aspect of the images unless we deep dive into the specific images used and try to understand how many of the predictions are accurate.