Dog Rating Data Analysis

In the dog rating data analysis project for Udacity, we were provided three datasets. The first one consists of 5000+ dog rating data which was filtered to 2,356 containing data on twitter id, source, url, name and the dog stage.

The second dataset scraped directly from Twitter consists of retweets that tracks retweet count, favorite counts among many other fields.

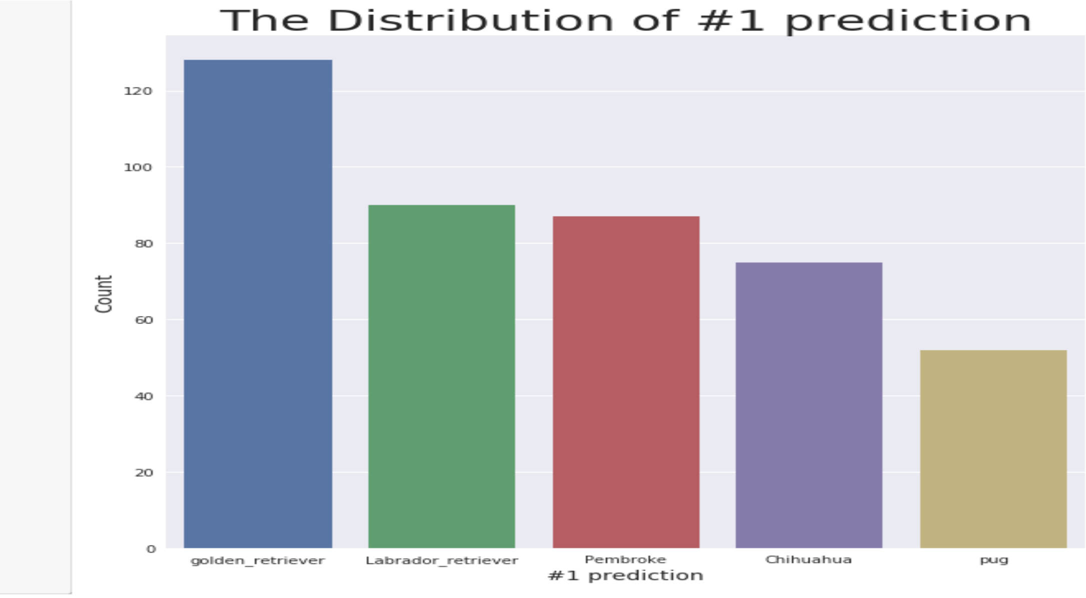
The third dataset contains image prediction (named predictions) consists of twitter id, top three predictions and their probabilities. This had to be downloaded programmatically from Udacity’s website.

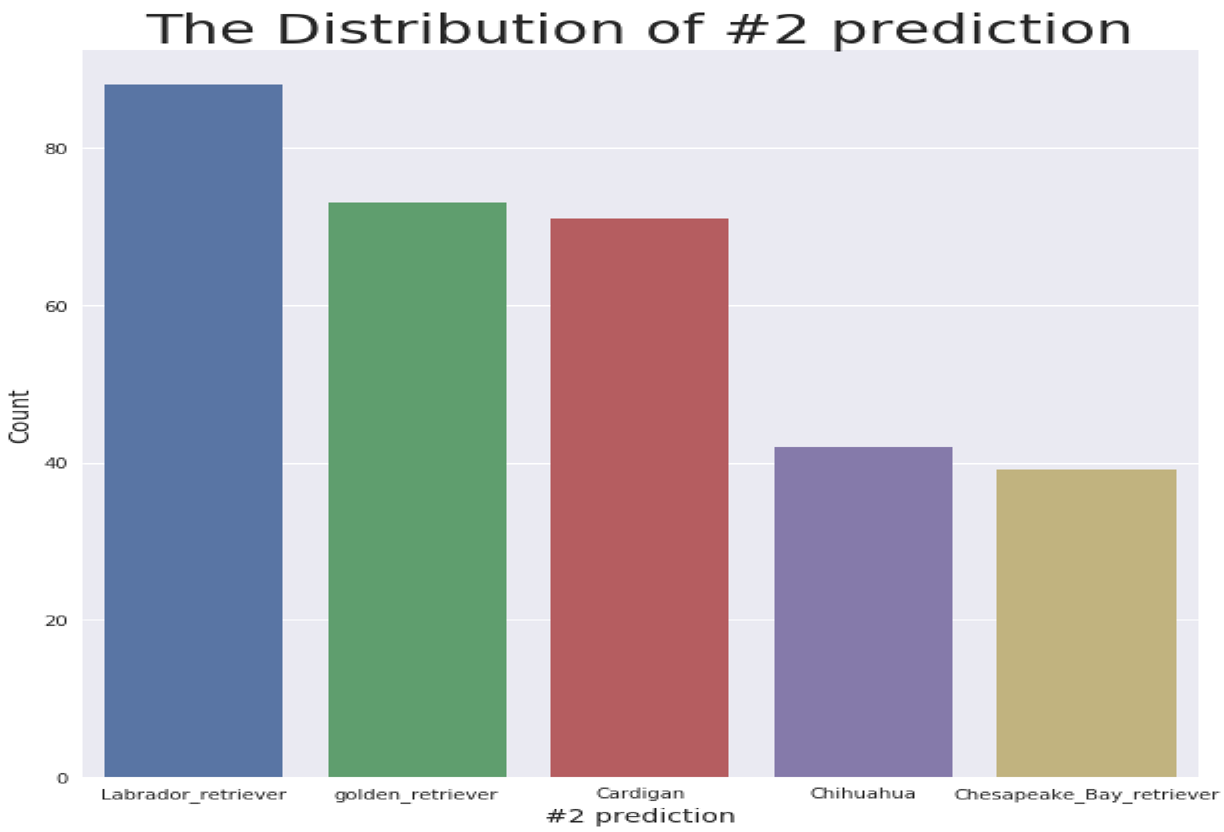
As expected, the datasets contained a couple of quality and tidiness issues that needed to be fixed before the data is analyzed. I assessed the data both visually and programmatically in order to discover all issues and I fixed them.

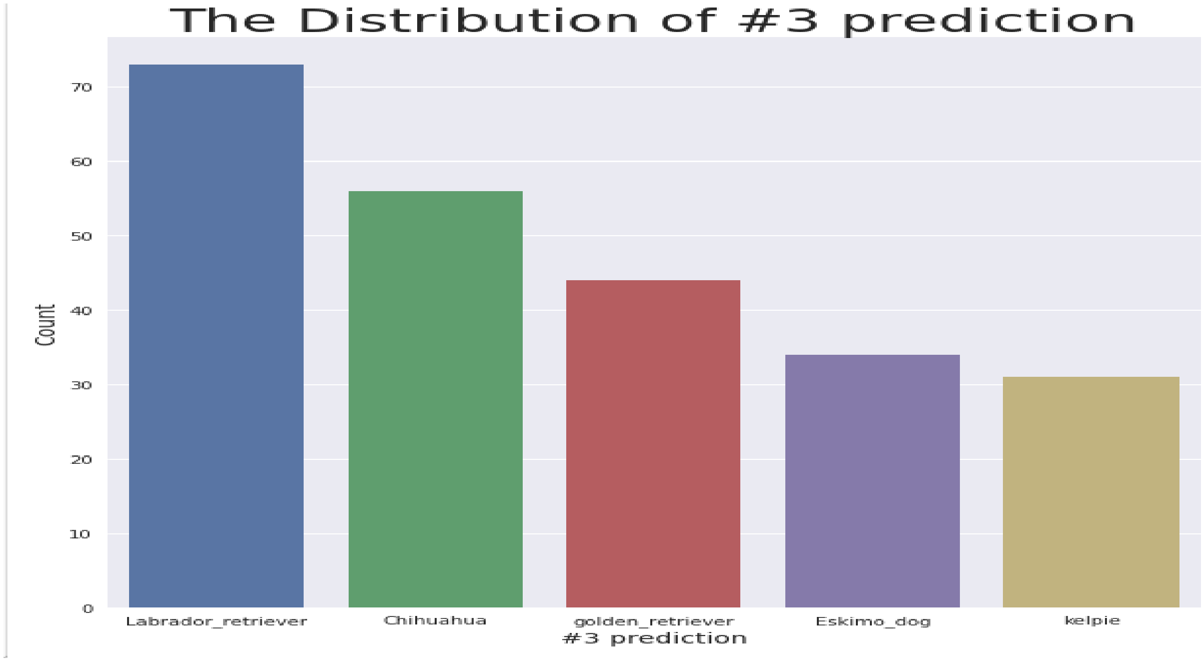
Having fixed the issues, I merged the first dataset with the dataset containing prediction. Then I saved both the merged dataset and the second dataset containing retweet.

In the merged dataset (which I named TwitterAM), I got interested in columns p1, p2 and p3 which show the first, second and the third prediction of the content of the image downloaded from Twitter.

I plotted the histograms of p1, p2 and p3 respectively.




The Distribution of #2 prediction
Count
0
20
40
60
80
Labrador_retriever
golden_retriever
Cardigan
Chihuahua
Chesapeake_Bay_retriever
#2 prediction


The Distribution of #3 prediction
Count
0
10
20
30
40
50
60
70
Labrador_retriever
Chihuahua
golden_retriever
Eskimo_dog
kelpie
#3 prediction

Insights from the charts:

1. Golden Retriever has the highest #1 prediction, followed by Labrador retriever.

2. Golden Retriever and Labrador Retriever are in the top 5 for the #1 prediction, #2 prediction and #3 prediction.

3. Only Labrador Retriever was predicted for more than 70 dogs as a #1, #2 or #3 prediction.