WeRateDogs
Findings & Analysis

"This is Atticus. He's quite simply America af. 1776/10."

Introduction

In this project I gathered, assessed and cleaned data from the WeRateDogs Twitter archive data, before analysing and visualising the data.

Data

1. The first dataset ('twitter-archive-enhanced.csv) is the **enhanced archive data** provided by Udacity, which contains extra columns for dog stages.
2. The second is an **image predictions** file (image-predictions.tsv) that is a result of the WeRateDogs Twitter archive data being run through a neural network that can classify dog breeds.
3. The third is **Twitter's API data** (tweet-json.txt), from where I extracted the retweet_count and favorite_count columns for the WeRateDogs archive data.
4. After cleaning and merging the data into one dataframe called master_df, I saved it to a **master dataset** titled 'twitter_arc

Summary of Findings

1. There is a positive correlation between retweet counts and favourite tweet counts.
2. Pupper is the most common dog stage while dogs in both doggo and floofer stages have the highest mean number of retweets.
3. Golden Retriever is the most predicted dog breed while the Standard Poodle is the most popular breed wrt mean number of retweets.

Summary Statistics

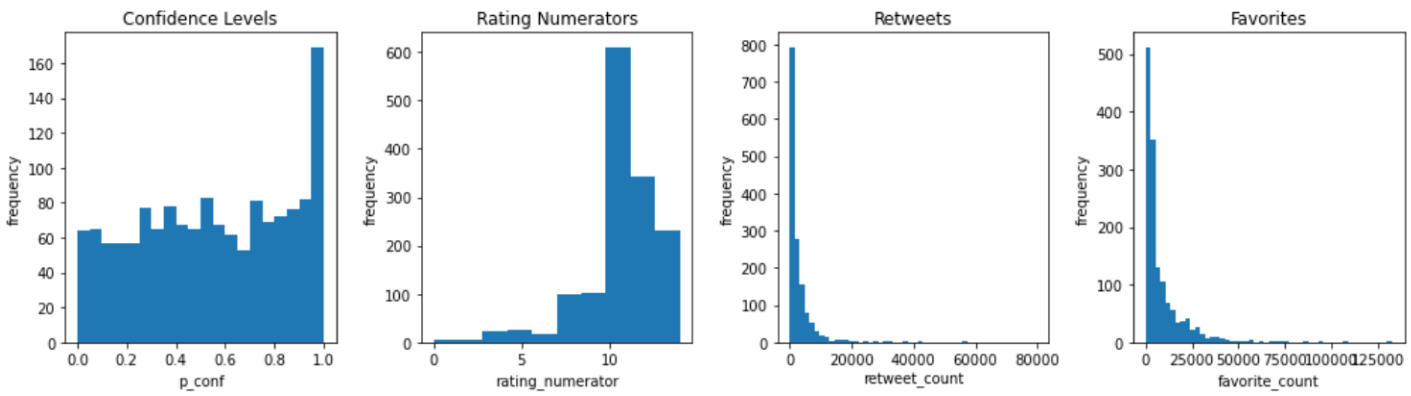
a) For both retweet_count and favorite_count, mean > than median, so the distributions are skewed to the right. This is most likely because the data has a lot of extreme outliers, which makes sense given how the Twitter algorithm can promote certain tweets more than others, the likelihood of bot accounts engaging with certain tweets more than others, and also the fact that some tweets are older and have more time to accumulate likes and retweets.

	p_conf	rating_numerator	retweet_count	favorite_count
count	1466.000000	1466.000000	1466.000000	1466.000000
mean	0.547648	11.864256	3065.960437	8771.723056
std	0.301375	46.187125	5512.480105	12886.857409
min	0.000010	0.000000	16.000000	0.000000
25%	0.292723	10.000000	652.250000	1834.250000
50%	0.547977	11.000000	1421.500000	3897.500000
75%	0.821962	12.000000	3433.000000	10635.250000
max	0.999956	1776.000000	79515.000000	132810.000000

b) The distributions of retweet_count and favorite_count below confirm point a), showing extreme skewness.

The image prediction confidence level (p_conf) has a more or less uniform distribution, with one prominent peak at the end. This is most likely because the variable is a combination of multiple sources of variation i.e., p_conf1, p2_conf and p3_conf, which would need to be analysed separately to reveal any interesting patterns.

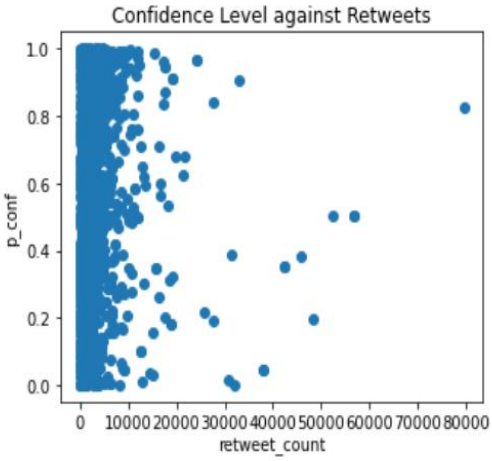
rating_numerators are concentrated around 10 to 13. Dogs are generally adored, so most dogs are likely to receive exaggerated ratings. This is also part of WeRateDog's unique rating system. Further analysis shows that there's only one extreme rating (1776) left after cleaning.

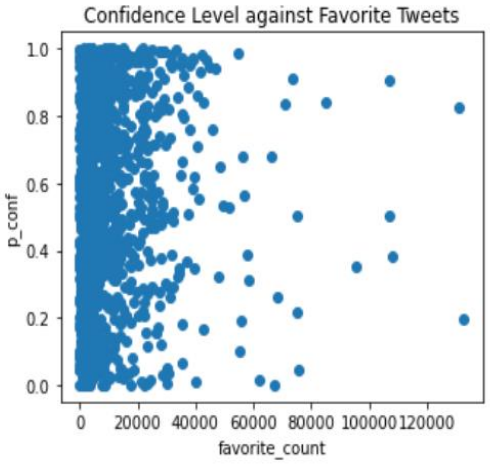


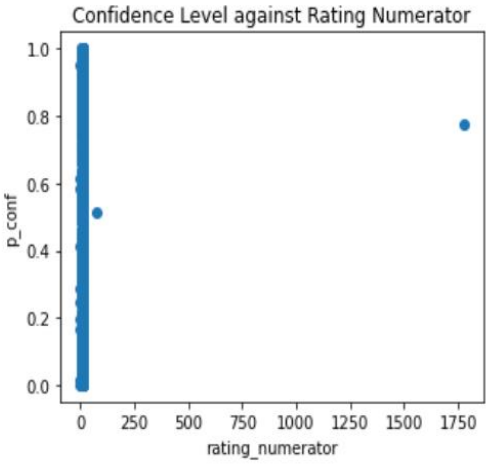
After reading the image url for the 1776 rating, it pulled up an image of the dog Atticus. The tweet caption "This is Atticus. He's quite simply America af. 1776/10" shows that the rating was intentional and it was not a mistake.

c) I inspected more outliers and read the url for the image with a 0 rating and it showed an image of an unnamed dog by a pool. The caption "When you're so blinded by your systematic plagiarism that you forget what day it is. 0/10" shows that the rating was intentionally made in jest.

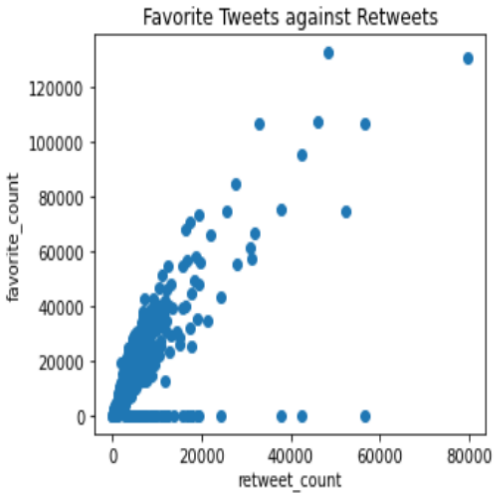
d) There is no correlation between confidence level and retweet_count, favorite_count and rating_numerator, respectively. This means that there is no relationship between breed popularity (as it pertains to ratings and retweets) and likelihood of of being predicted correctly by the Neural Networks algorithm.

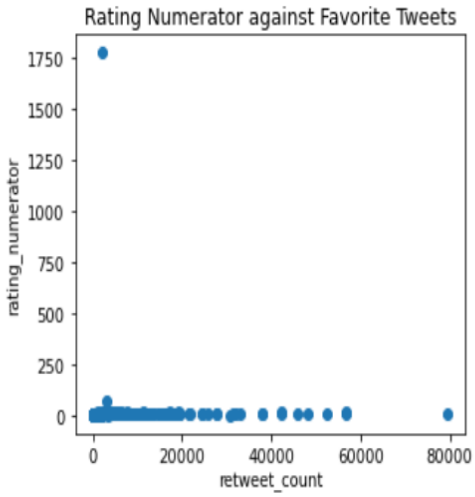


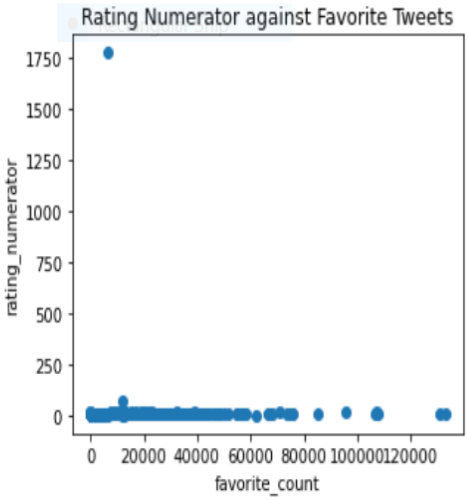




e) There is however a positive correlation between favorite count and retweet count. There is no correlation between rating_numerator and either retweet_count or favorite_count because rating_numerator is generally concentrated around similar values, making the graphs appear constant.

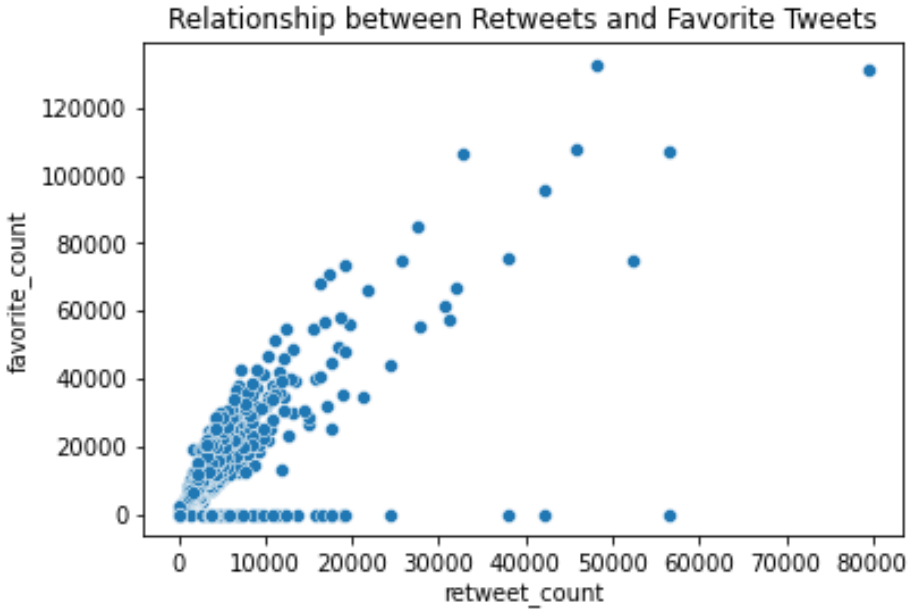






Findings and Visualisations

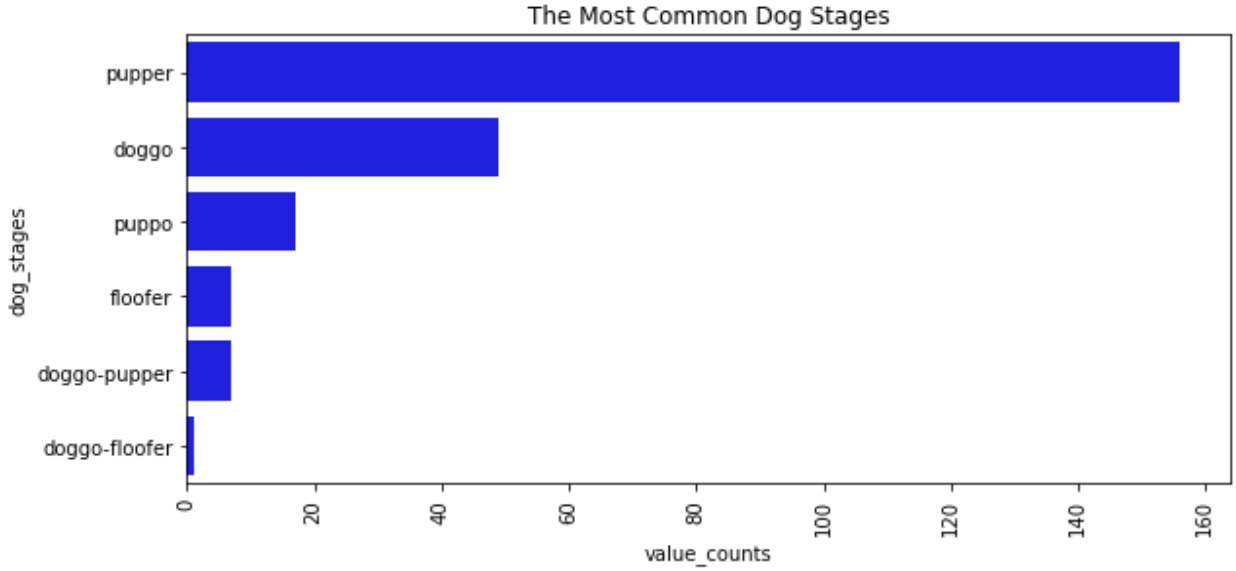
- As mentioned, there is a positive correlation between retweet count and favorite count. There's also an interesting horizontal line along 0 favorite_count, showing 0 favorite_counts for retweet_counts that are surprisingly high. Further investigation is needed to check if these zero values are errors, which is too complicated for this project. A statistical test is also needed to test the significance of the relationship, which is also out of scope for this project.

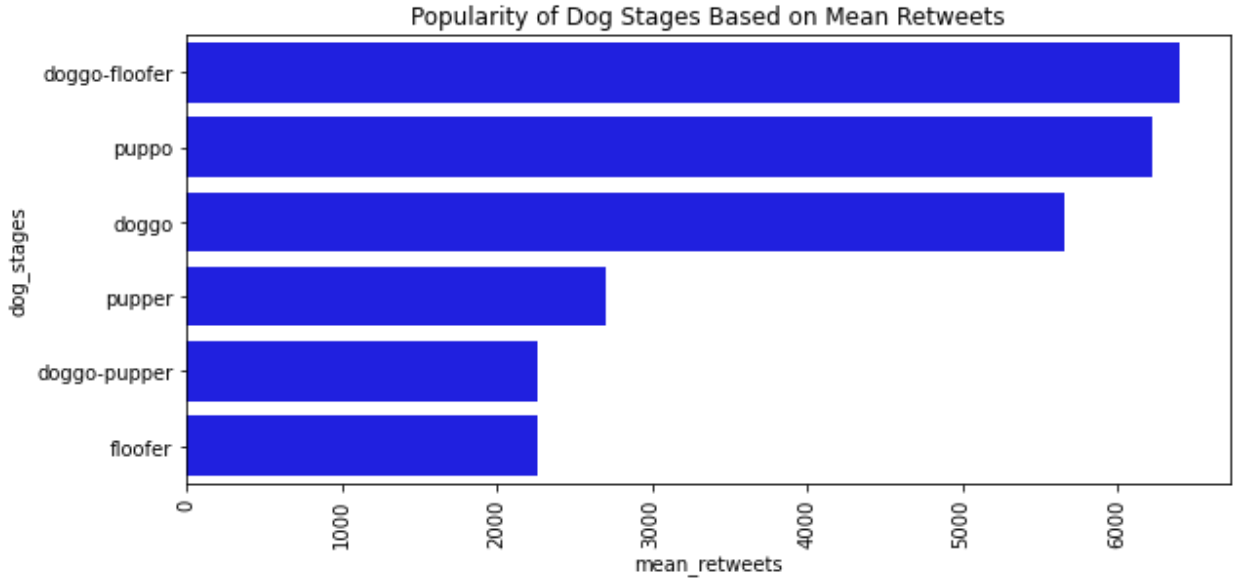


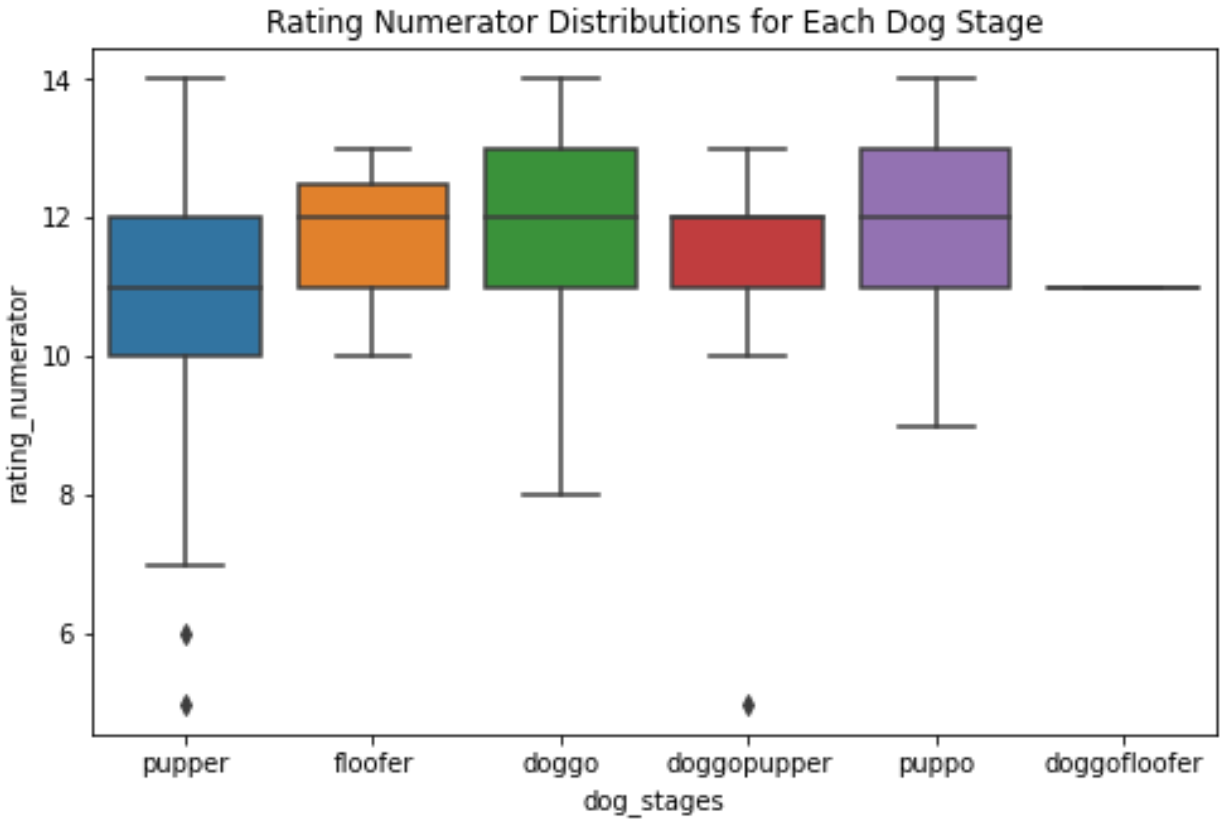
- Pupper is the most common dog stage, yet it has the third lowest mean number of retweets. doggo-floofer is the least common stage yet it has the highest mean number of

retweets. Some limitations to the bar graphs for popular dog stages are that dog_stages has too many null values, and there are dogs with 2 stages, making it difficult to know which of the stages contributes the most to the number of retweets.

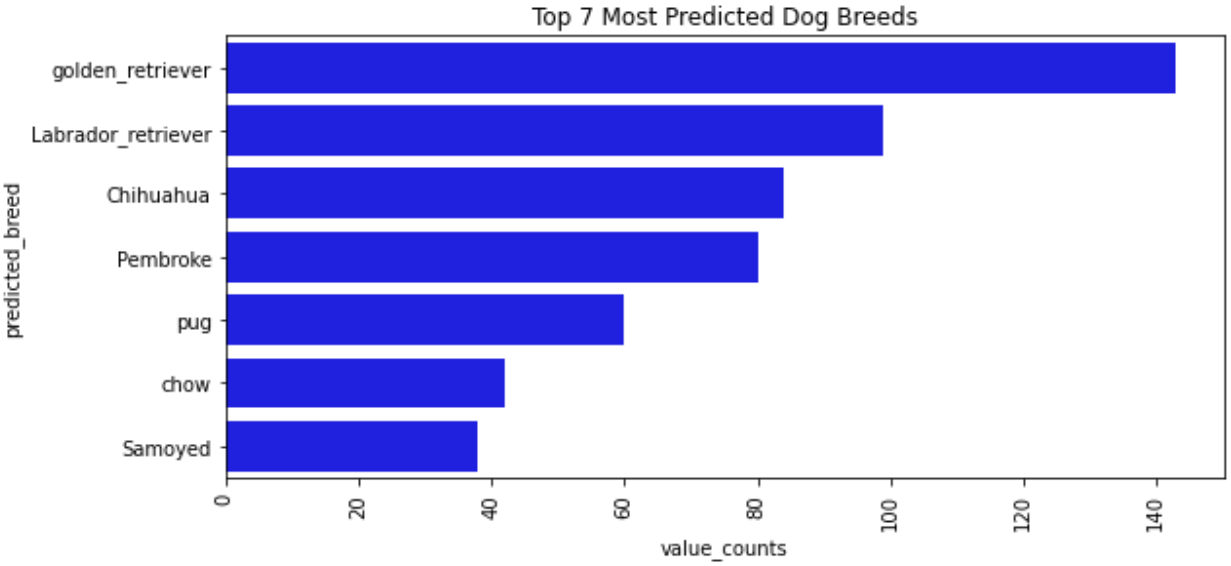
Despite being the most common dog stage, pupper also is amongst dog stages with the lowest median rating numerator of around 11, along with doggo-pupper and doggo-floofer. There are too many null values for dog_stages (there are only 237 non-null values out of 1466 entries) and this posed a limitation to the analysis as a larger set of non-null entries would have probably given clearer and more accurate box plot distributions.

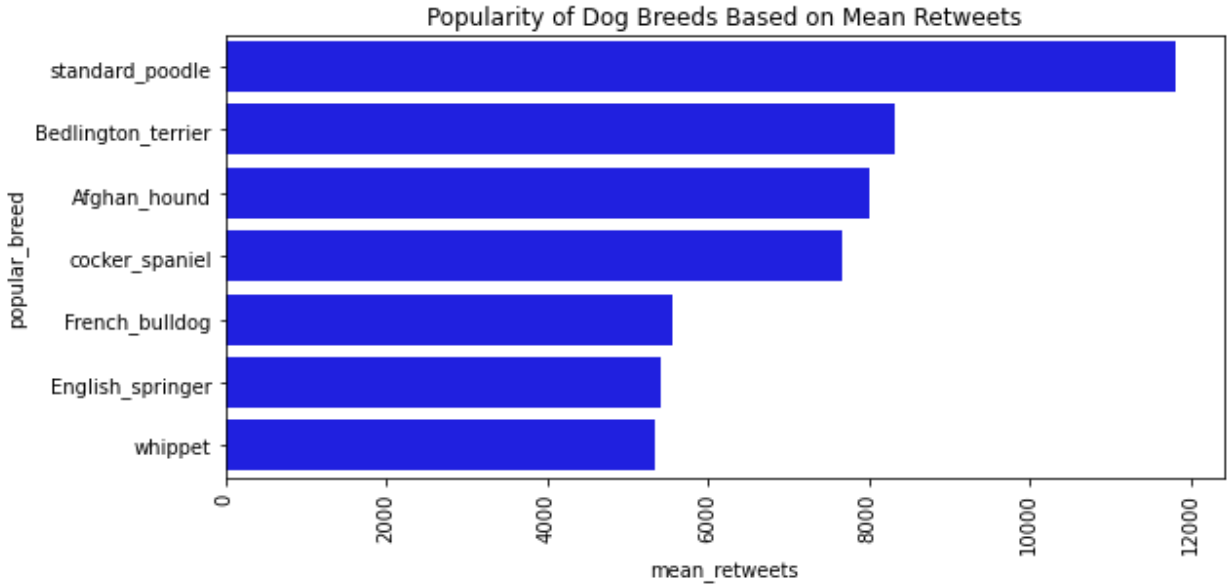






- Golden Retriever is the dog breed most likely to be predicted by the neural network, followed by Labrador Retriever, but it is not necessarily the most popular dog breed based on retweet_count. The most popular dog breed based on mean retweet count is the Standard Poodle. The top 7 dog breeds for the most likely predictions are not the same as top 7 dogs in number of image retweets.


Top 7 Most Predicted Dog Breeds
golden_retriever
Labrador_retriever
Chihuahua
Pembroke
pug
chow
Samoyed
predicted_breed
0
20
40
60
80
100
120
140
value_counts


Popularity of Dog Breeds Based on Mean Retweets
standard_poodle
Bedlington_terrier
Afghan_hound
cocker_spaniel
French_bulldog
English_springer
whippet
popular_breed
0
2000
4000
6000
8000
10000
12000
mean_retweets