act_report

WeRateDogs Twitter Archive

Christian Chimezie

Data Analytics Nanodegree 2nd Project
June 2022

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc.

Data Wrangling

Gathering Data

The data for this project was in three different formats:

1.Twitter Archive File: WeRateDogs downloaded and shared their Twitter archive exclusively for this project. The WeRateDogs Enhanced Twitter archive contains data extracted from 2356 of the 5000+ tweets posted between November 15, 2015 and August 1, 2017 on the @dog rates Twitter account.

2. Image Prediction FileThis file stores tweet image predictions, i.e., what breed of dog is present in each tweet based on a neural network. It was hosted in tsv format on Udacity's servers and had to be downloaded programmatically via the Url.

3. Twitter API — JSON File: By using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's tweepy library. The favourite_count and retweet_count was extracted programmatically from this file.

Assessing Data

I visually and programmatically evaluated the datasets, identifying several quality and tidiness issues that needed to be addressed. In the three Datasets, Visual Assessment identified issues such as non-descriptive column headers and repetitive columns.

Most of the quality issues in the three datasets, such as incorrect data types and duplicate data, were identified by Programmatic Assessment.

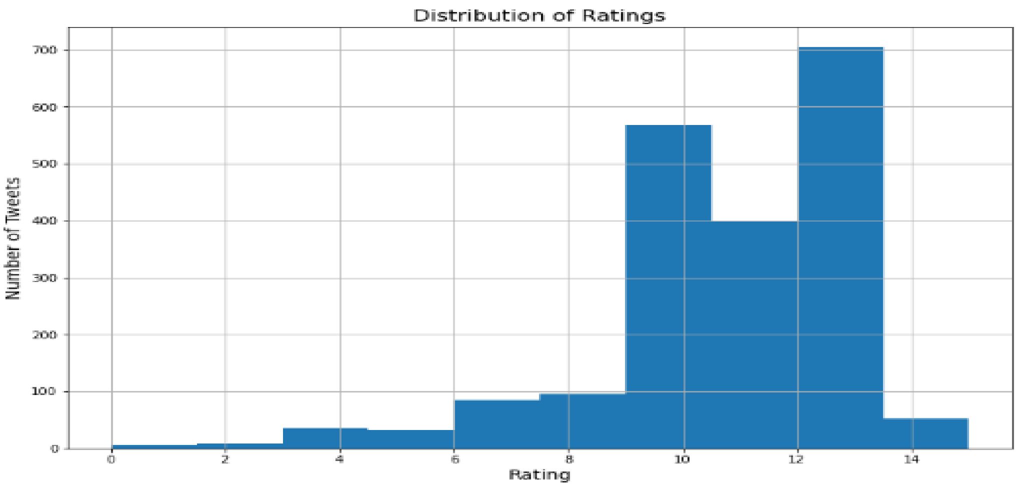
Cleaning Data

I implored the Define, Code and Test framework to programmatically clean the data. I converted my observations from the assess step into defined problems, translated these definitions to sophisticated code to fix these problems, then tested the three datasets to make sure the operations worked.

I merged the cleaned datasets and saved as a csv file (twitter_archive_master).

Analysis and Visualization

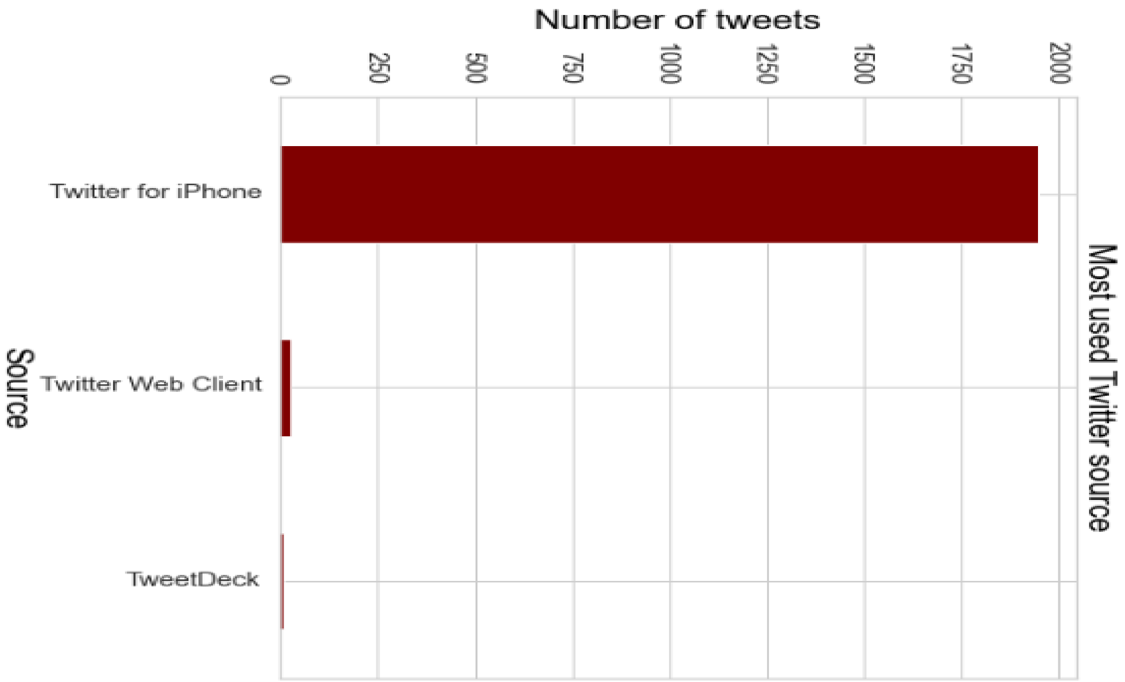
Insight 1: Distribution of rating



They are good dogs Brent.!

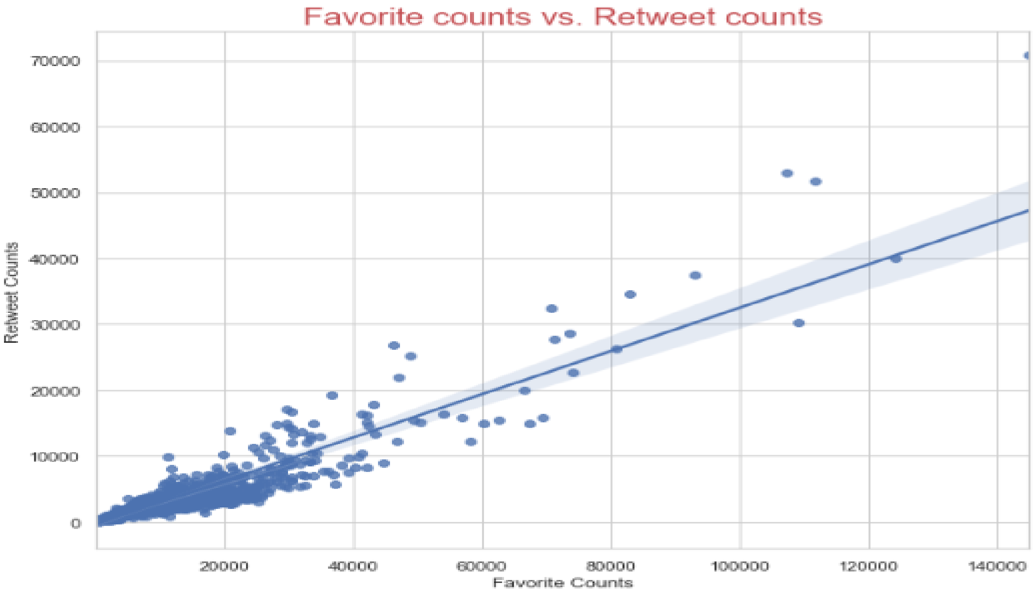
WeRateDogs ratings are almost always greater than 10. 11/10, 12/10, 13/10, etc.

Insight 2: Most used twitter source



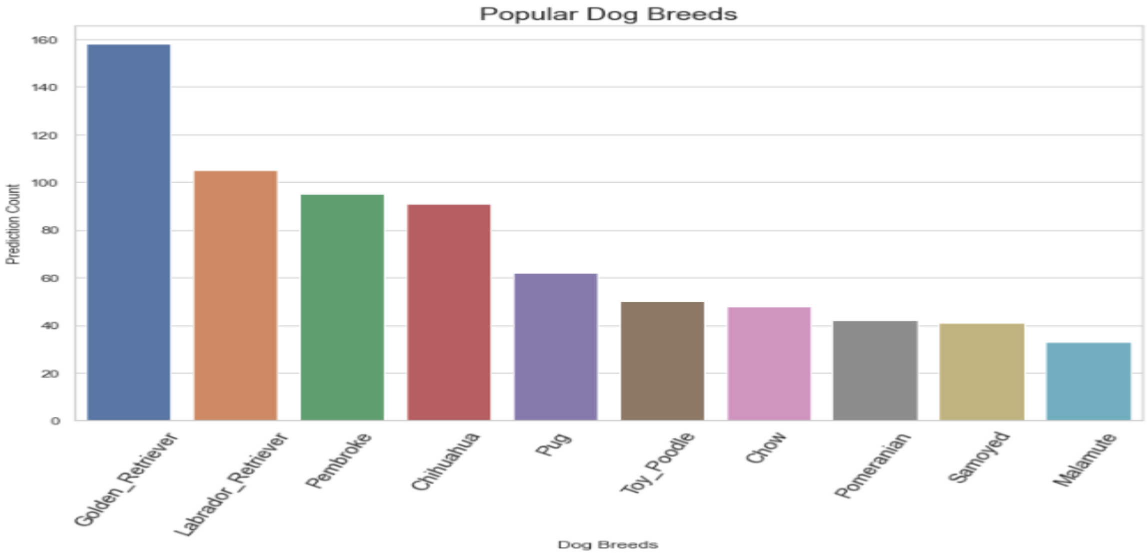
The users make use of Twitter for iPhone, Twitter web client and TwitterDeck. From analysis and visualisation, Twitter for iPhone is the most used by users with over 1800 tweets.

Insight 3: The relationship between favorite counts and retweets counts



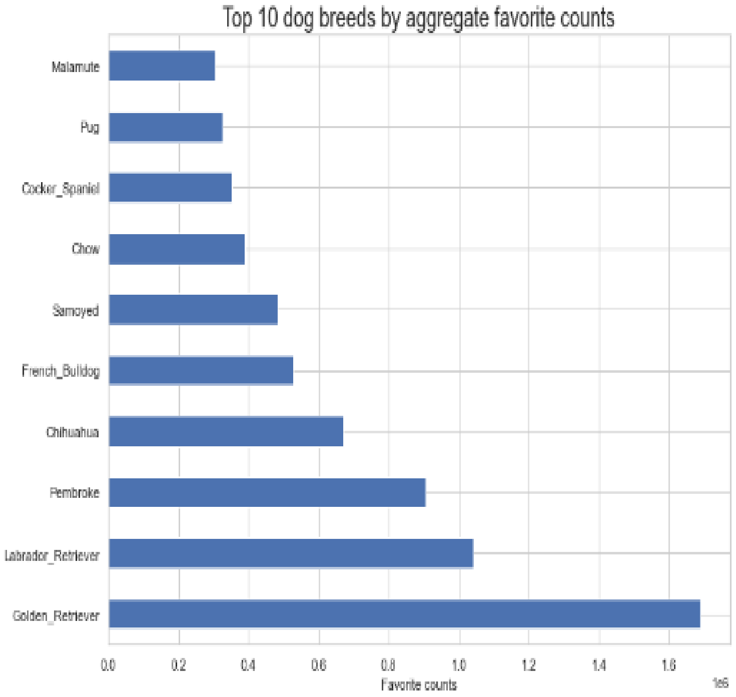
There is strong positive correlation between favorites and retweets. The correlation coefficient between favorites and retweets count is 0.92, which is close to 1 and positive demonstrating a strong positive correlation between those two metrics. This makes sense because if you liked the tweet, you are most likely to retweet it.

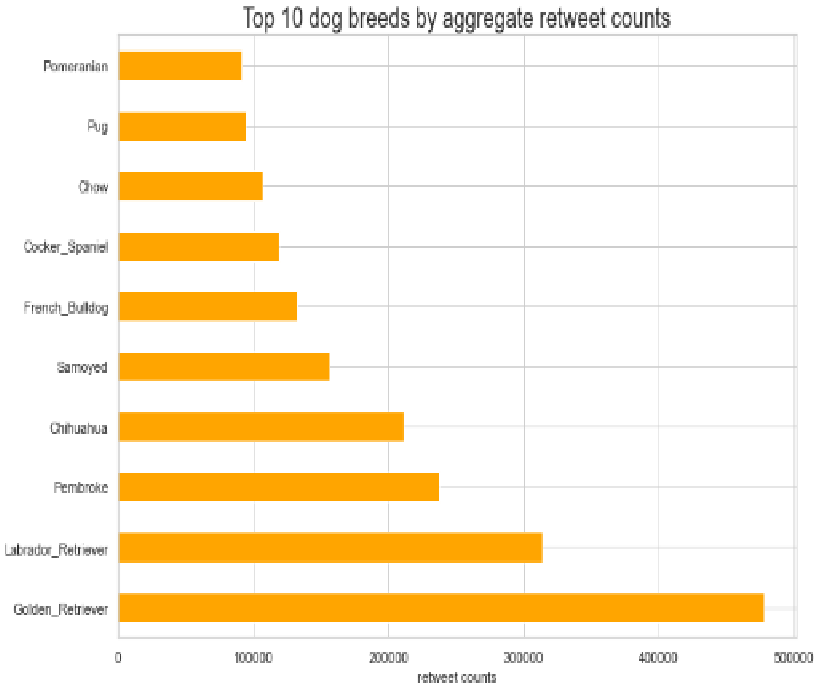
Insight 4: Most popular dog breeds



The 10 most popular dog breeds are Golden Retriever, Labrador Retriever, Pembroke, Chihuahua, Pug, Toy Poodle, Chow, Pomeranian, Samoyed and Malamute with Golden Retriever being the most popular with 158 mentions**.**

Insight 5: Top 10 dog breeds by aggregate favorite and retweet count

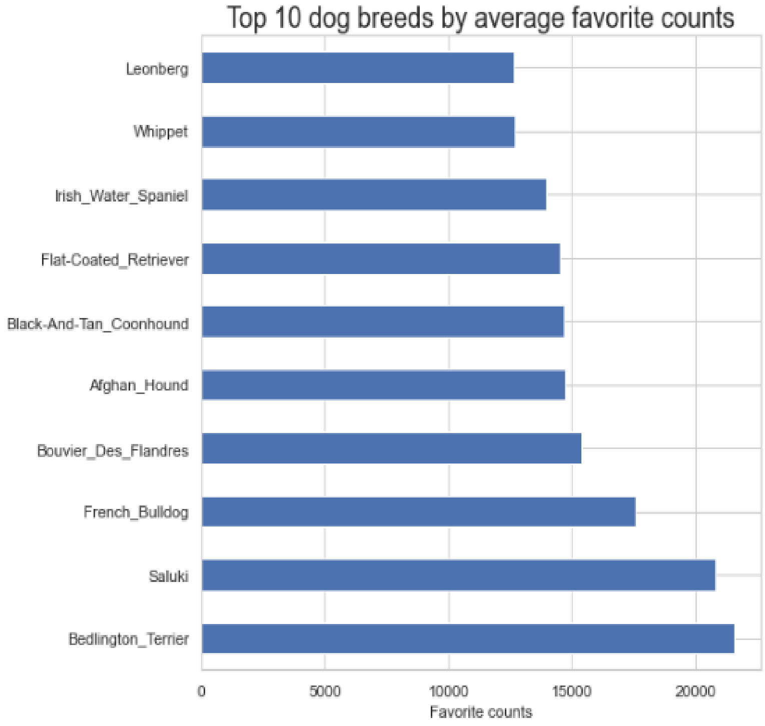


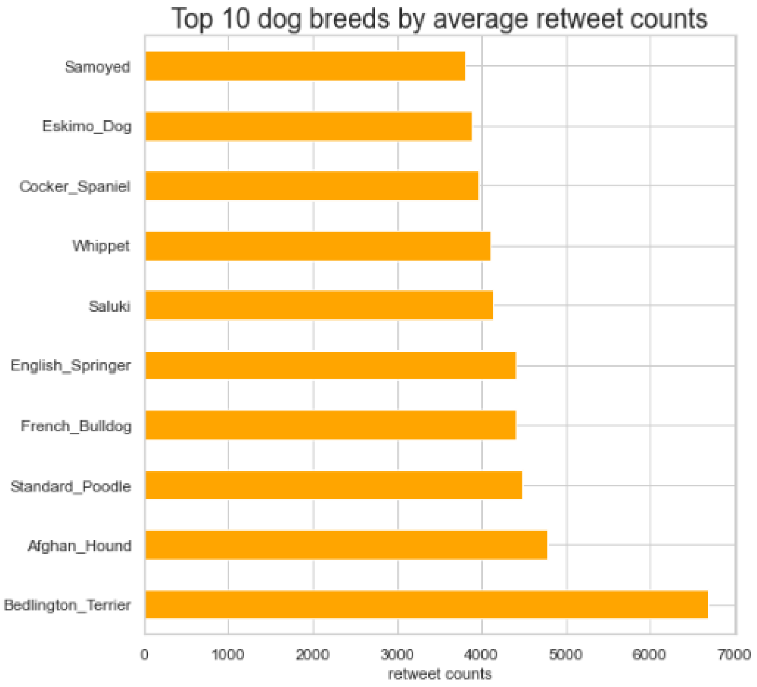


People love Retrievers!

The images of a Golden Retriever are the most liked and retweeted tweets. The Labrador Retriever and Pembroke are the second and third breeds. If we only extract tweets about Golden Retrievers, they will undoubtedly be the breed with the highest retweet and favorite count because they are the most popular breeds.

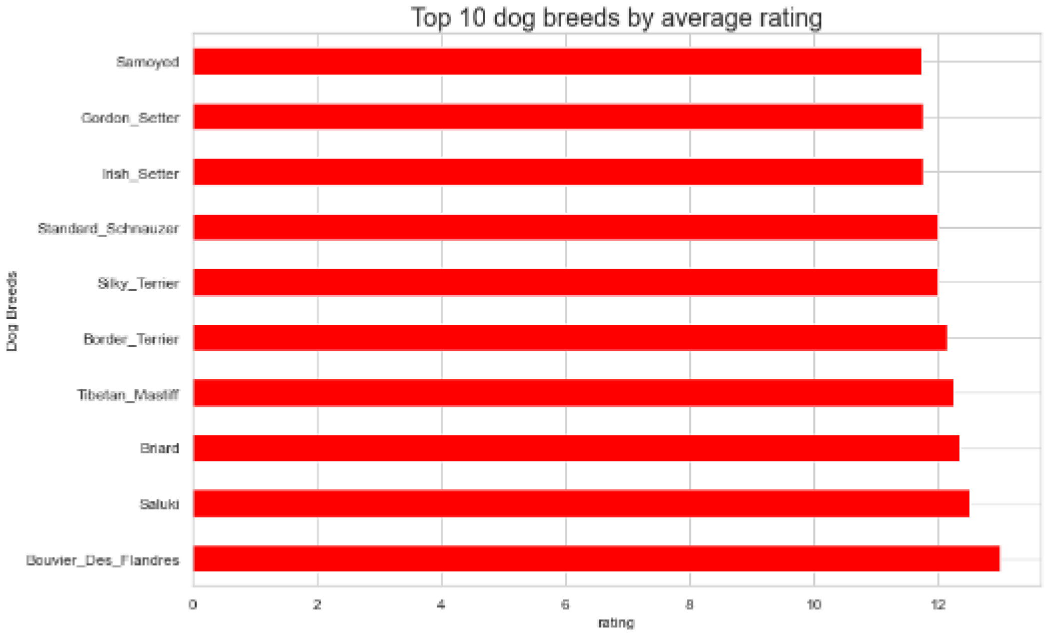
Insight 6: Top 10 dog breeds by average favorite and retweet counts





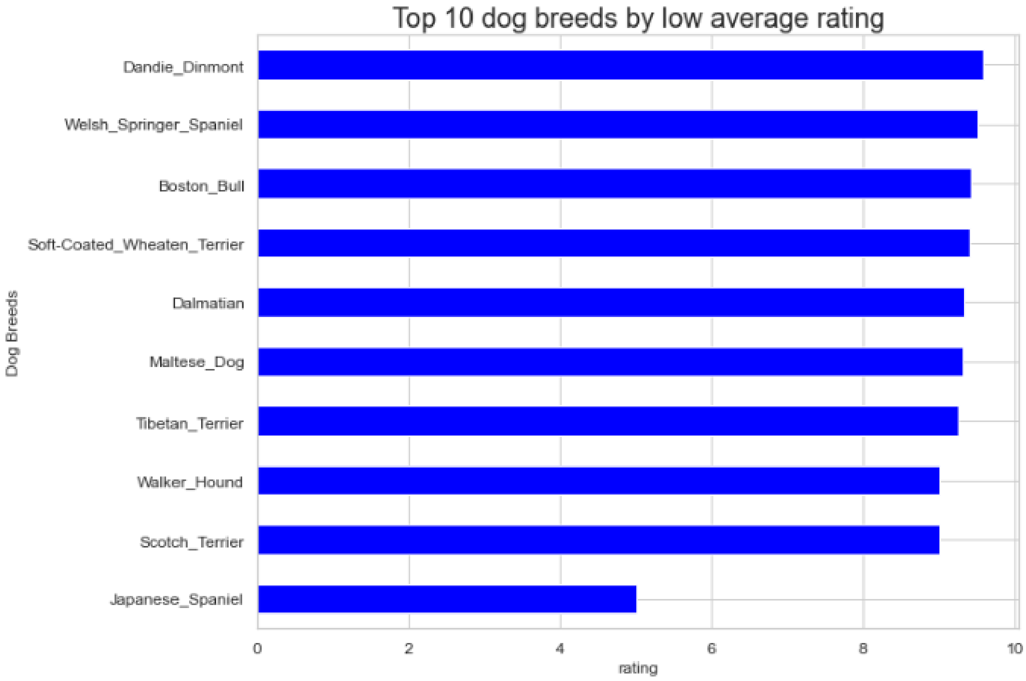
To reduce bias, we use the mean retweet and favorite count. Looking at the average metrics, we can see that the most popular breeds are not Golden Retriever, Labrador Retriever, or Pembroke. The most popular breeds are the Bedlington Terrier, French Bulldog, and Afghan Hound.

Insight 7: Top 10 dog breeds by high average rating



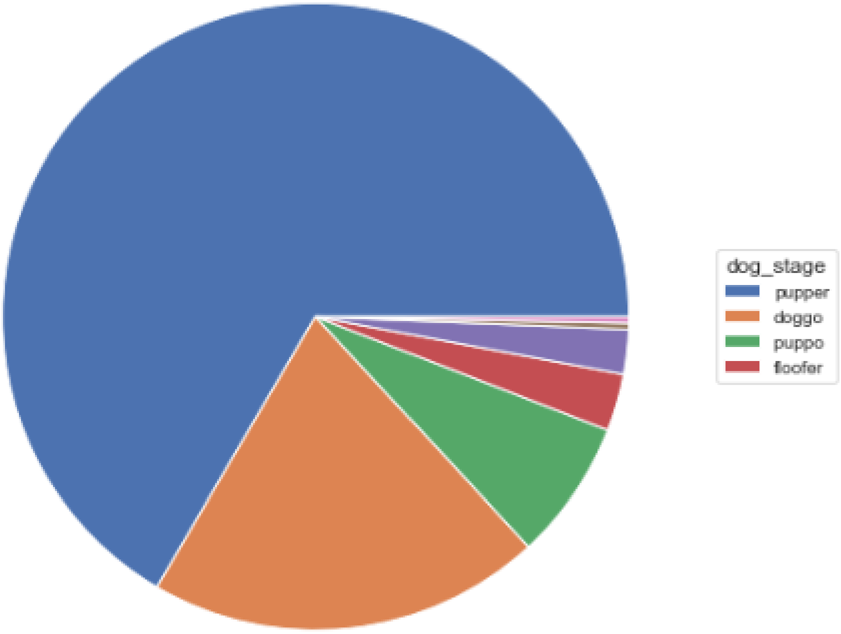
Bouvier Des Flandres, Saluki, Briard, Tibetan Mastiff, Border Terrier, Silky Terrier, Standard Schnauzer, Irish Setter, Gordon Setter, and Samoyed are the dog breeds with the highest average ratings. Bouvier Des Flandres average rating at 13/10.

Insight 8: Top 10 dog breeds by low average rating



The dog breeds with the lowest average ratings are the Japanese Spaniel, Walker Hound, Scotch Terrier, Tibetan Terrier, Maltese Dog, Dalmatian, Soft-Coated Wheaten Terrier, Boston Bull, Welsh Springer Spaniel, and Dandie Dinmont, with the Japanese Spaniel rating at 5/10.

Insight 9: Most Common Dog Stage



I love puppies!

The dog stages are puppies called ‘pupper’ or ‘puppo’, older puppies called ‘doggo’, and older doggo called ‘floofer’. From analysis and visualisation, people love puppies. Puppies are the most retweeted and liked from all.