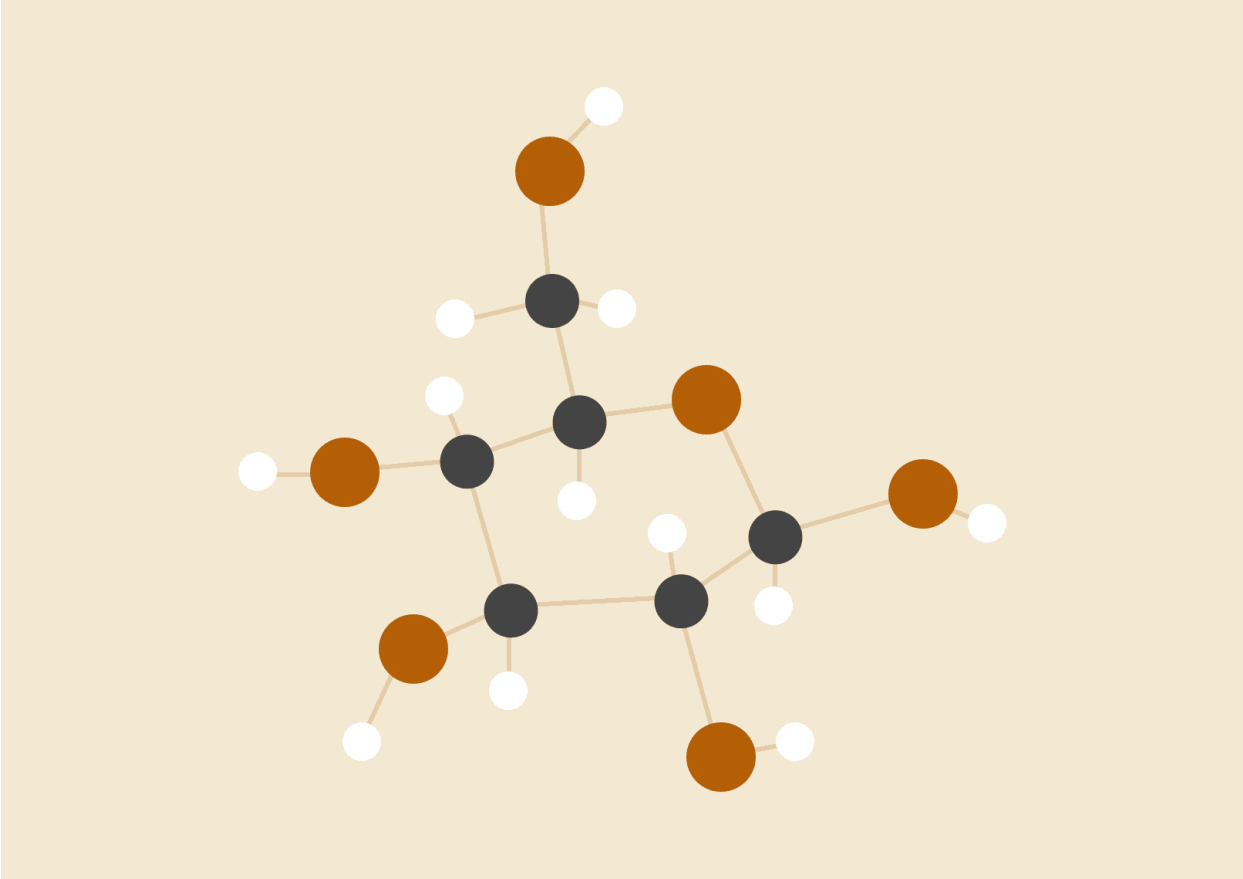
WRANGLING AND ANALYSING DATA

Report 2: Analysis and visualisation



Alham O. Hotaki

@Masterschool

Dec 14, 2022

INTRODUCTION

Real-world data rarely comes clean. Using Python and its libraries and gathering data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. The wrangling efforts are documented in a Jupyter Notebook, plus showcase them through analyses and visualisations using Python (and its libraries) and/or SQL.

The dataset used for wrangling (and analysing and visualising) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with humorous comments 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. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for this report to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

Project Steps

Steps in this project are as follows:

Step 1: Gathering data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Storing data

Step 5: Analysing, and visualising data

Step 6: Reporting

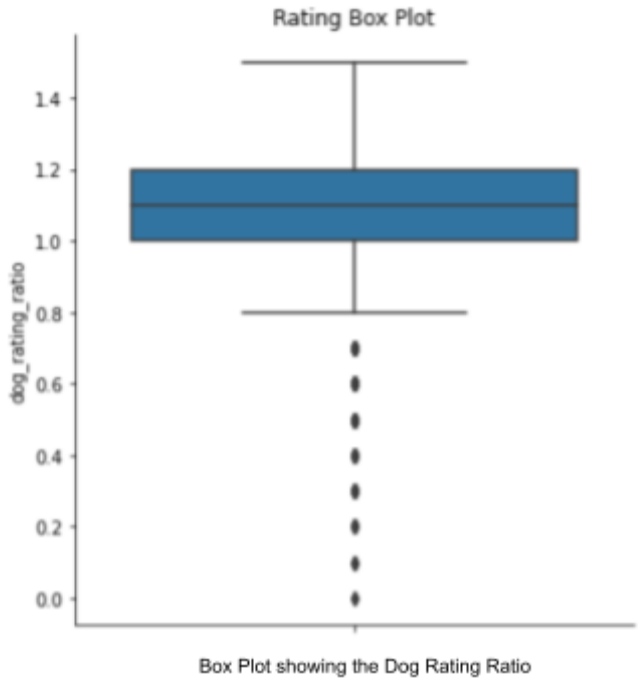
- data wrangling efforts
- data analyses and visualisations (current report)

Insight 1: Rating Analysis

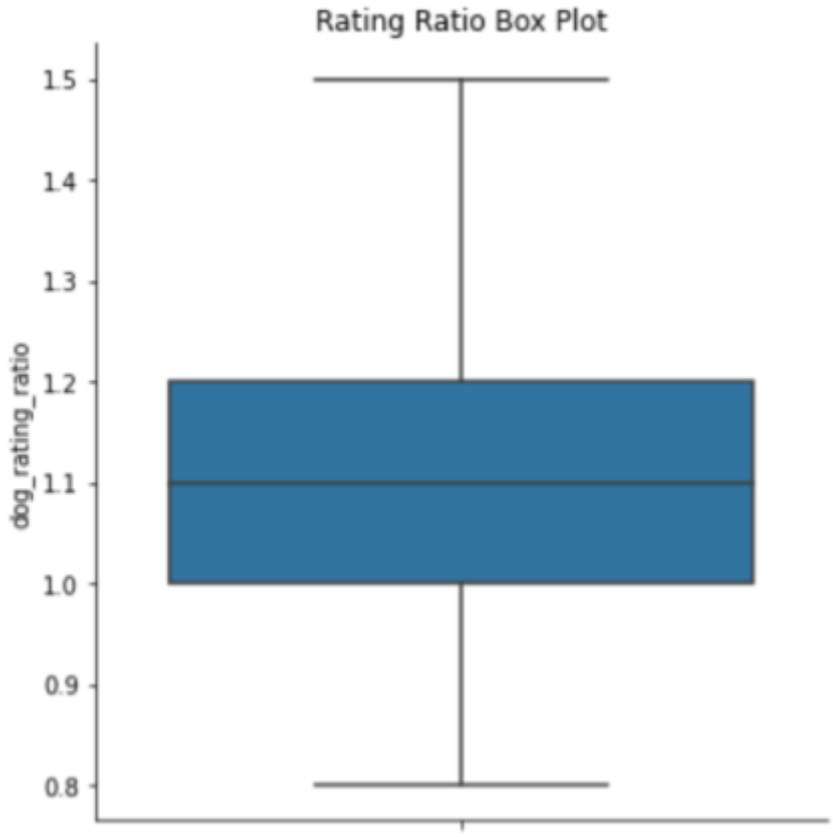
In order to have a better understanding of the rating, a new column was created by dividing numerator and denominator.

Basic statistics overview of the rating ratio is as follows:

count	1954.000000
mean	1.054401
std	0.216294
min	0.000000
25%	1.000000
50%	1.100000
75%	1.200000
max	1.400000



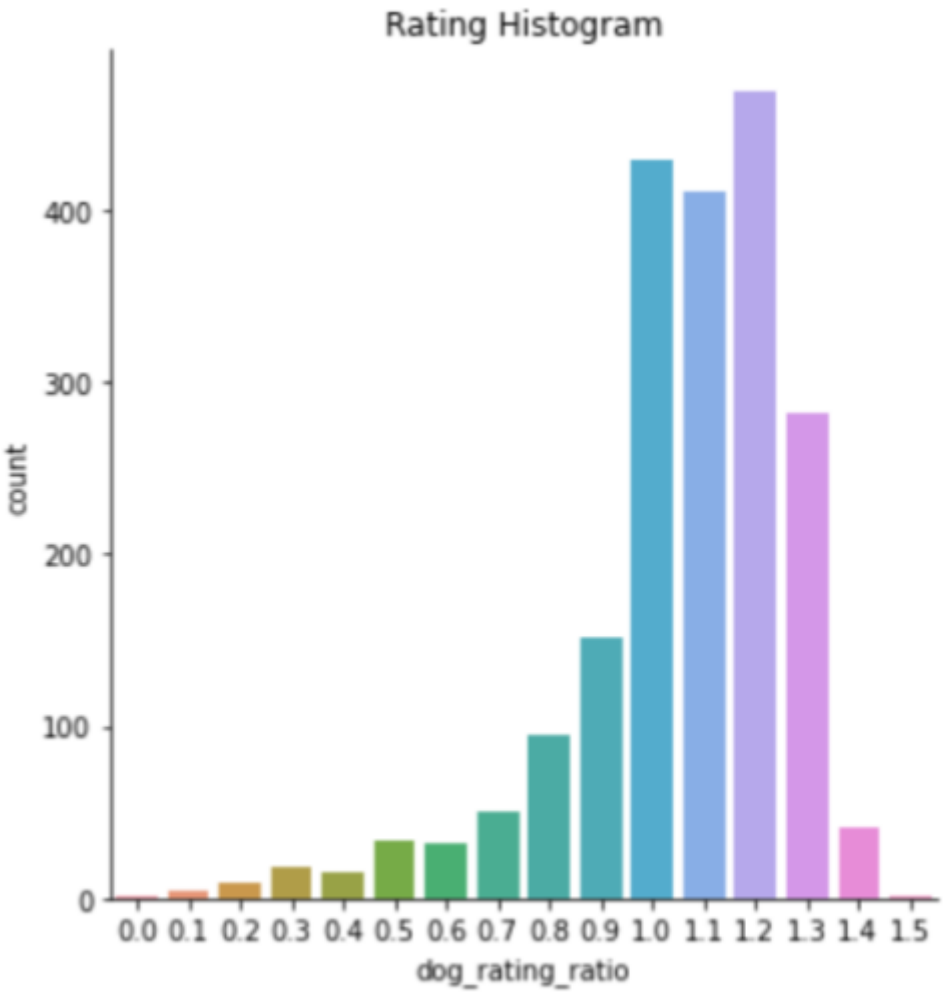
From the chart above, it can be seen that there are outliers, in order to have a proper understanding of the rating ratio, the following chart is without outlier:



Rating Ratio Boxplot without Outliers

It seems that the median of rating is 11 which is more than the denominator of the rating system (10). It shows that it is overrated.

In order to have a proper look at the rating ratio, lets see the following chart:



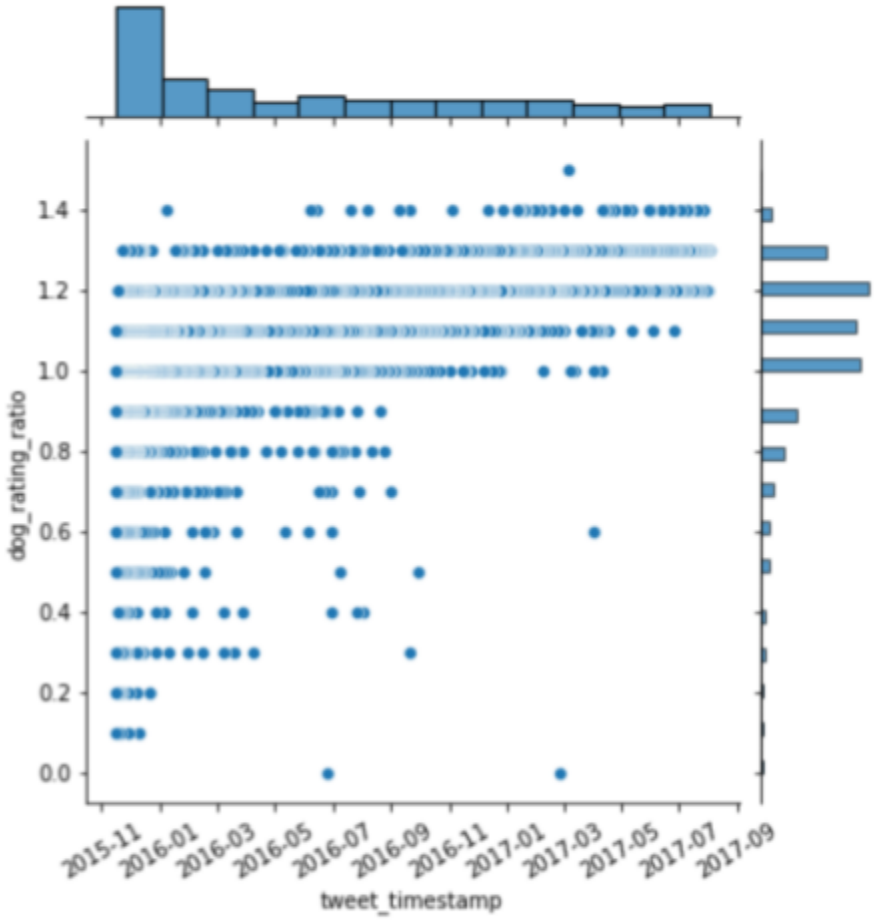
Rating Ratio Histogram

It shows clearly that most of the ratings are above 10 and the highest number of ratings are 12. It shows that the distribution of ratings is left skewed and mean is larger than median.

Suggestion: However, overrated rating seems to be fine, but it is suggested that the rating system be revised in a way that more realistic ratings are generated.

Insight 2: Rating Analysis

Let's have a look at the following chart:



Rating Ratios changes over time

Looking at the chart, it seems that in the beginning years, rating ratios were more realistic, while starting at the end of 2016, ratios got overrated exceeding the 10. It is also seen that in 2015 more ratings were generated in comparison to 2017.

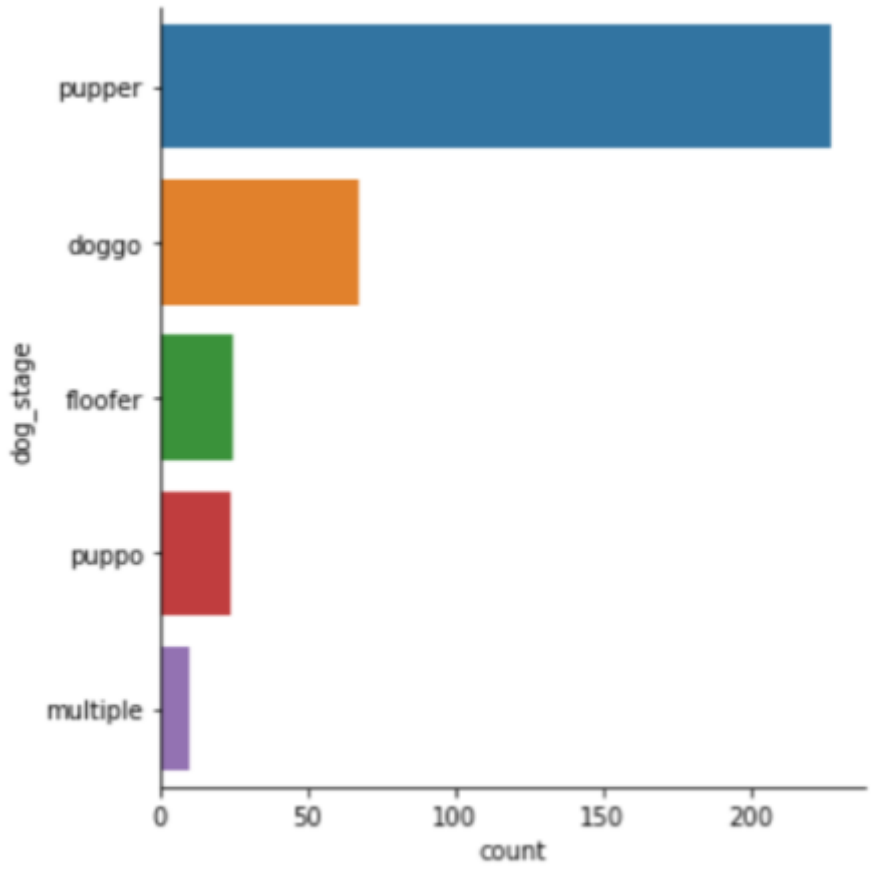
Suggestion: Rating system and its applicability requires hypothetical review and technical adjustments.

Insight 3: Dog Stage Analysis

There are four dog stages are defined in the data frame:

- Puppo
- Pupper
- Doggo
- floofer

There were some dogs categorised in more than one category which during the data cleaning step, they were categorised as ‘multiple’.

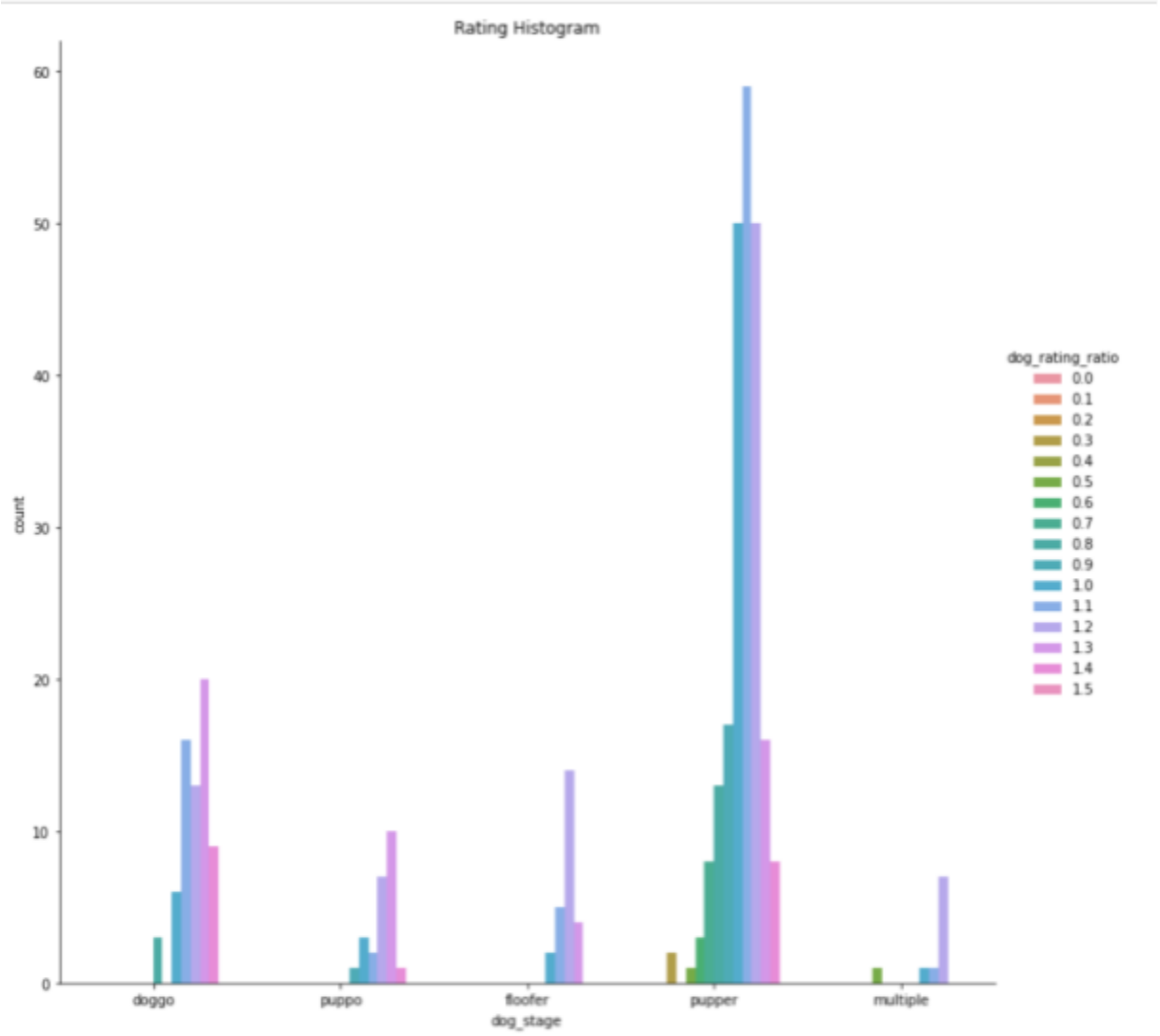


Count of Dog Stage Categories

It seems that puppers are the highest number among all the dogs rated.

Insight 4: Dog Stages and their rating ratios

It showed above that puppers are more popular than other stages of the dogs, while the following chart shows, if they are highly rated or not?

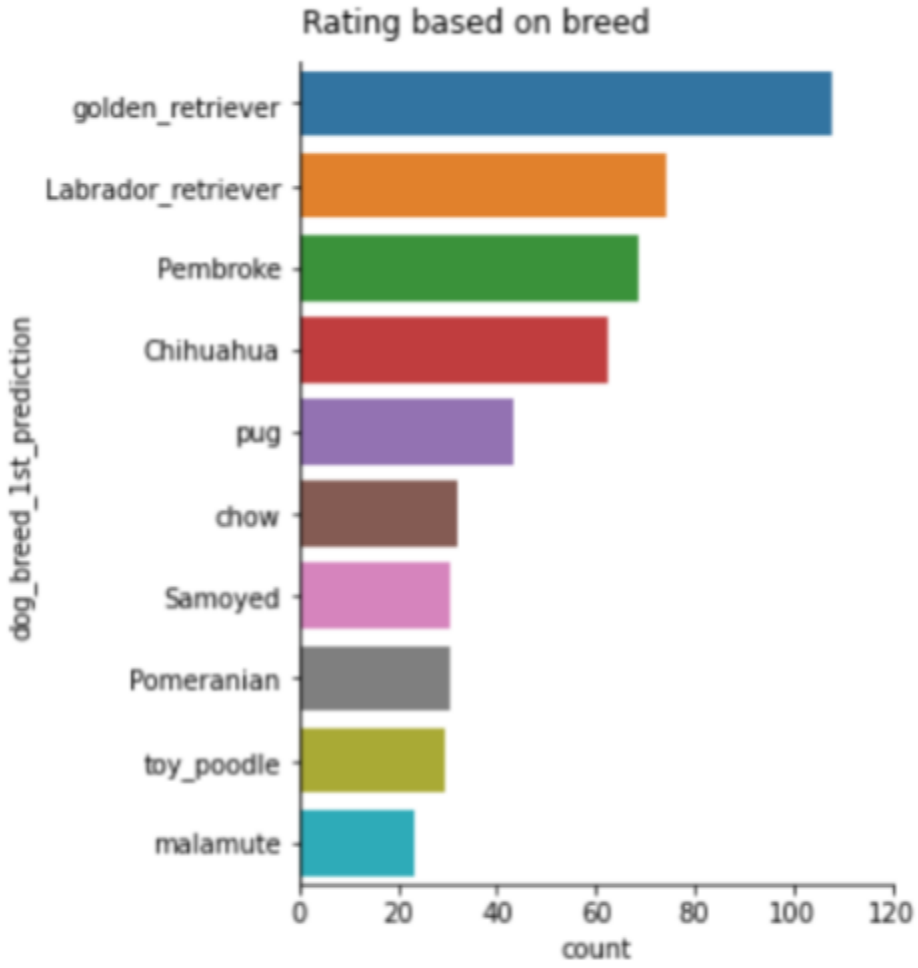


Dog stages histograms and their rating ratios

It seems that despite the popularity of puppers, doggos have the highest rating ratios among other stages of the dogs.

Insight 5: Popular Dog Breeds

There are 375 breeds of dogs listed in the data frame. The following chart shows the top ten breeds of dogs based on the rating ratios:

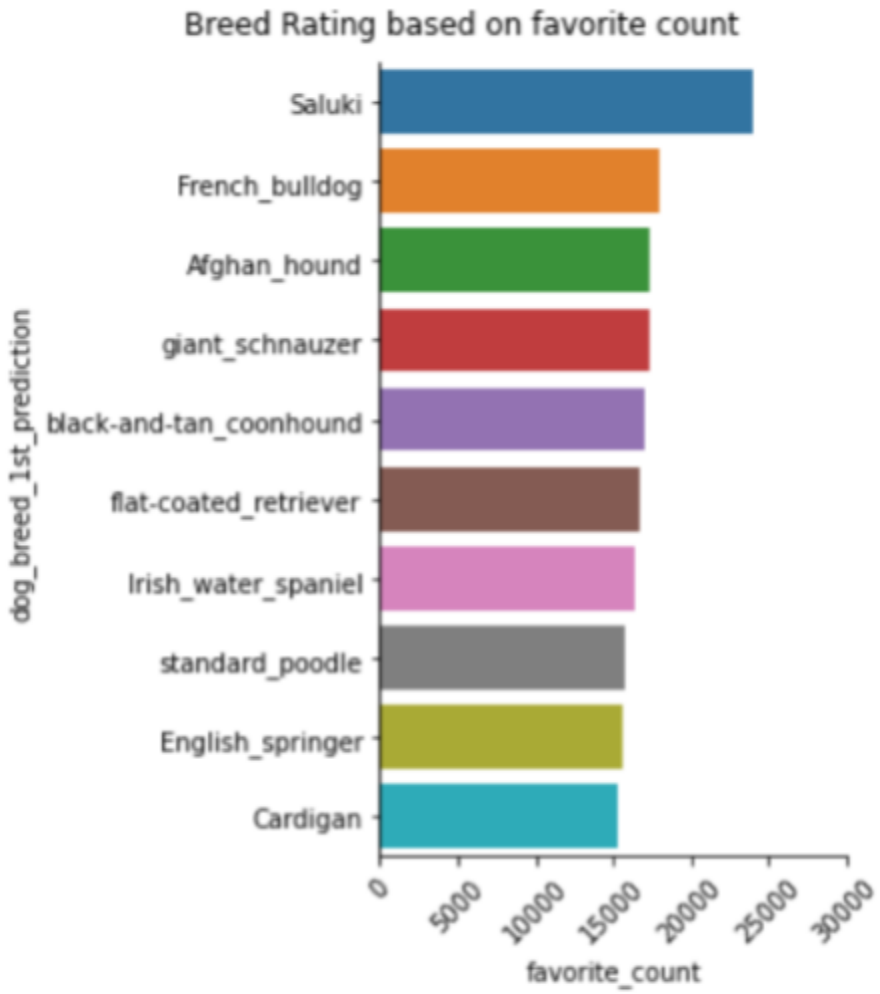


Top 10 breeds of dogs

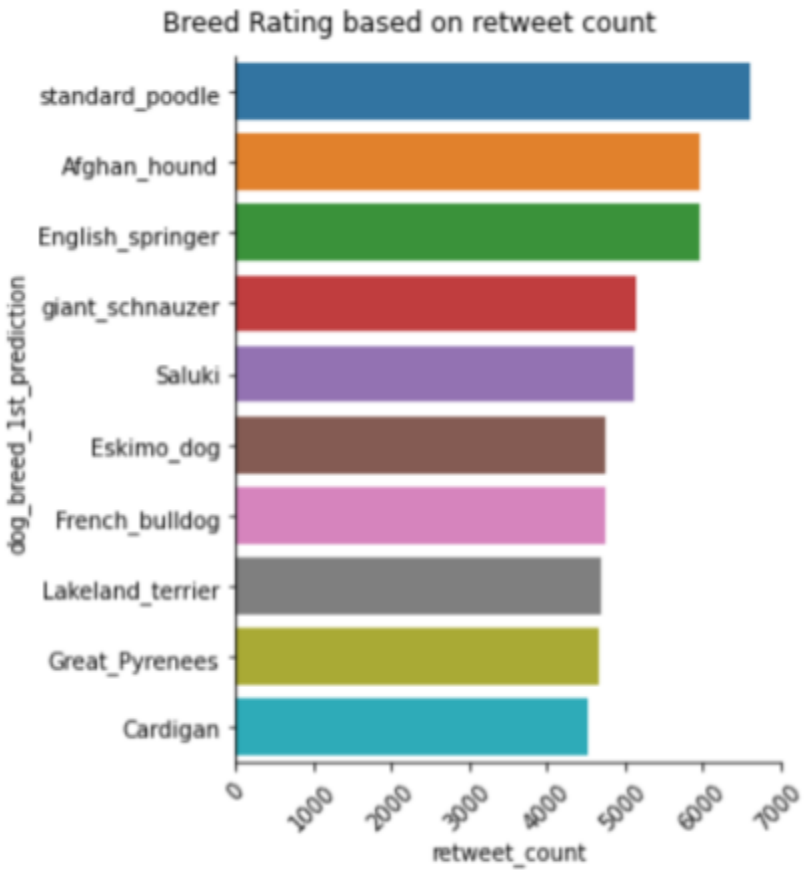
It shows that the retrievers are the top breed.

Insight 6: Favouri`ted and Retweeted Dog Breeds

Despite the number of dogs breeds which have been rated and showed above in insight 5, the following charts shall explain most favourited count of breeds and most retweeted breeds:



Top 10 highest favorited count of dogs breeds



Top 10 highest breed of dogs based on retweet count

It seems that despite popularity of dogs based on the rating ratio counts, popularity of dog breeds differ when compared to retweet count or favourite count.

However there are common breeds among favourite and retweet counts holding different position in top 10 rankings:

Saluki, French Bulldog, Afghan Hound, Giant Schnauzer, Standard Poodle, English springer, and cardigan

End of Report