Wrangling Data from the WeRateDogs Twitter Feed

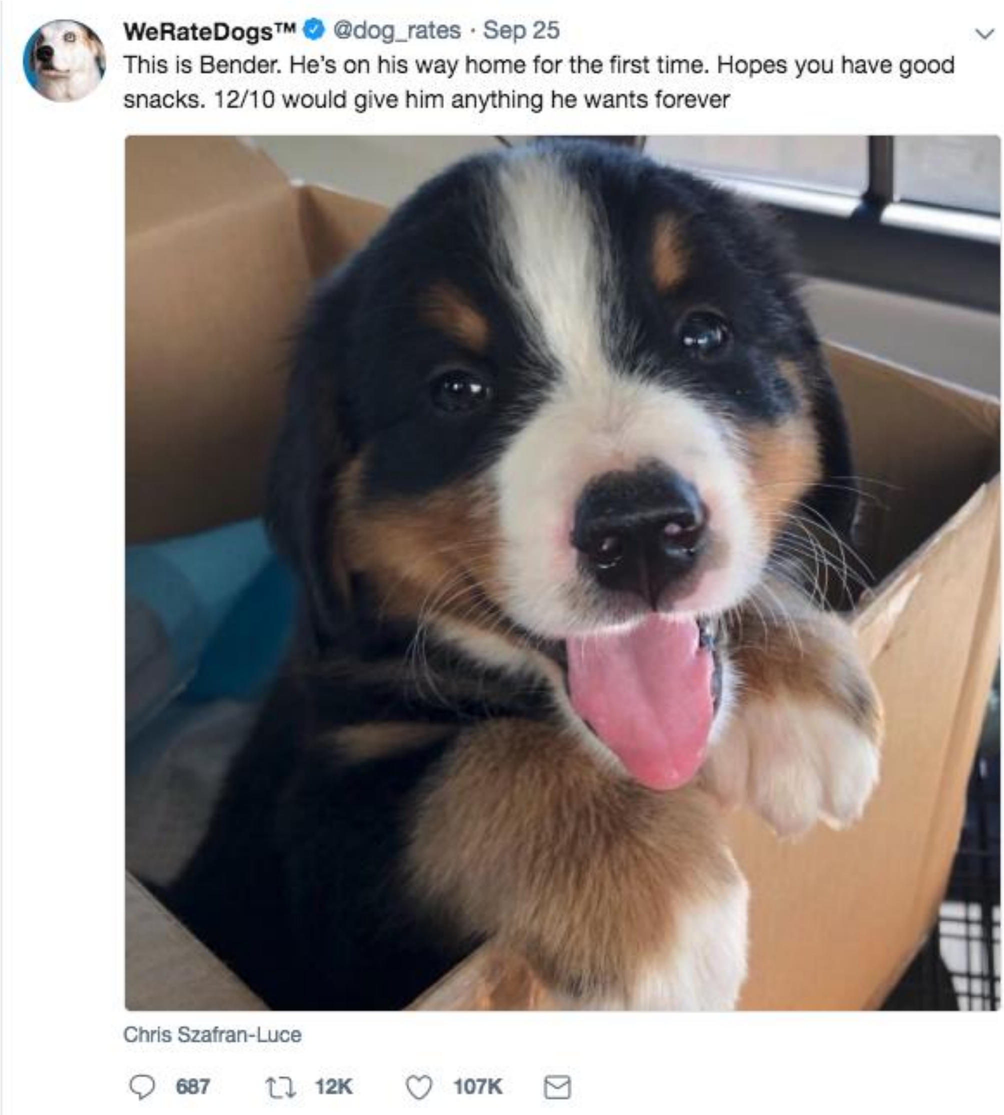
Project Objective

In my Udacity course, I was given a project which involved gathering, assessing, cleaning and analysing data from the WeRateDogs Twitter feed. More specifically, it involved the following:

- Gathering data from 3 different sources using 3 different techniques.
- Detecting and documenting at least 8 data quality issues and 2 data tidiness issues.
- Providing 3 insights and 1 visualization about the clean data

This document details the insights gained from the data after cleaning.

Sample Tweet:



WeRateDogs™ @dog_rates · Sep 25

This is Bender. He's on his way home for the first time. Hopes you have good snacks. 12/10 would give him anything he wants forever

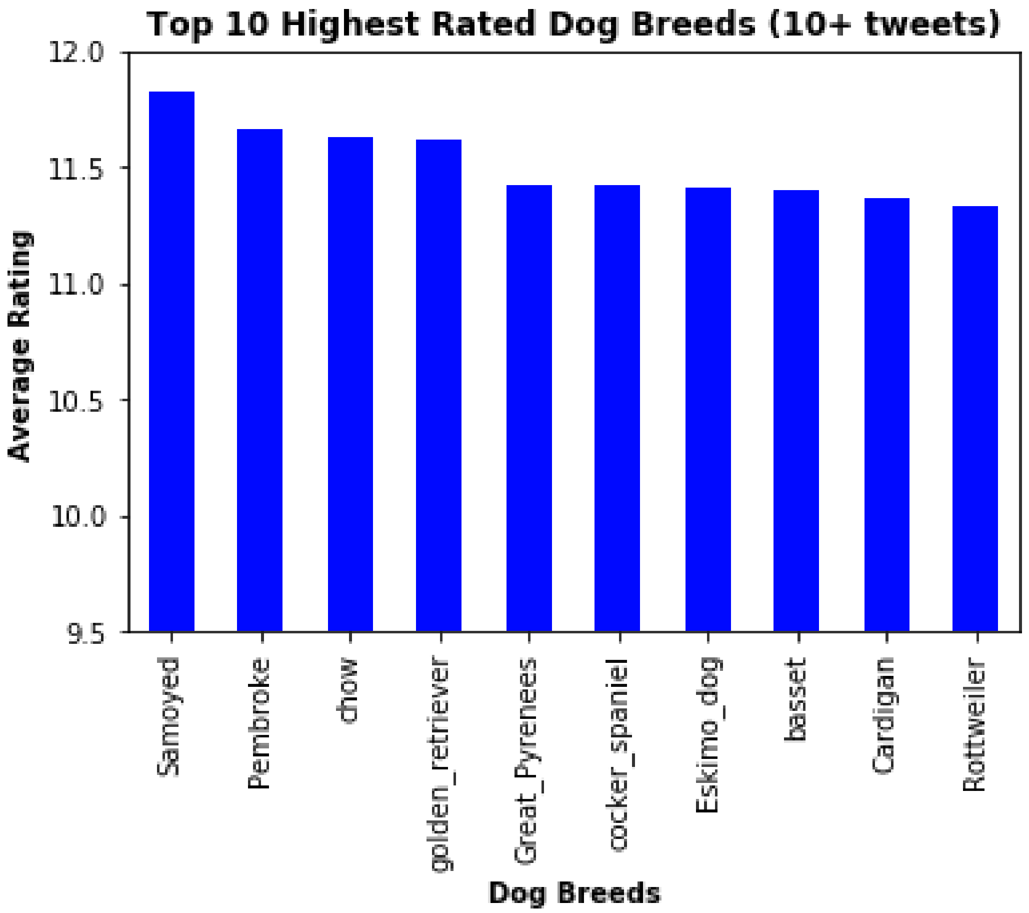
Chris Szafran-Luce

Top 10 Highest Rated Dog Breeds

Information about each dog breed was based on a prediction from a neural network. Three predictions were given and the predictions were not necessarily dog breeds. I decided to take the most probable dog breed prediction and assign that as the dog breed featured in each tweet.

In addition, for this insight I decided to only take into account dog breeds for which there were 10 or more available ratings. This was to avoid anomalies populating the Top 10 list.

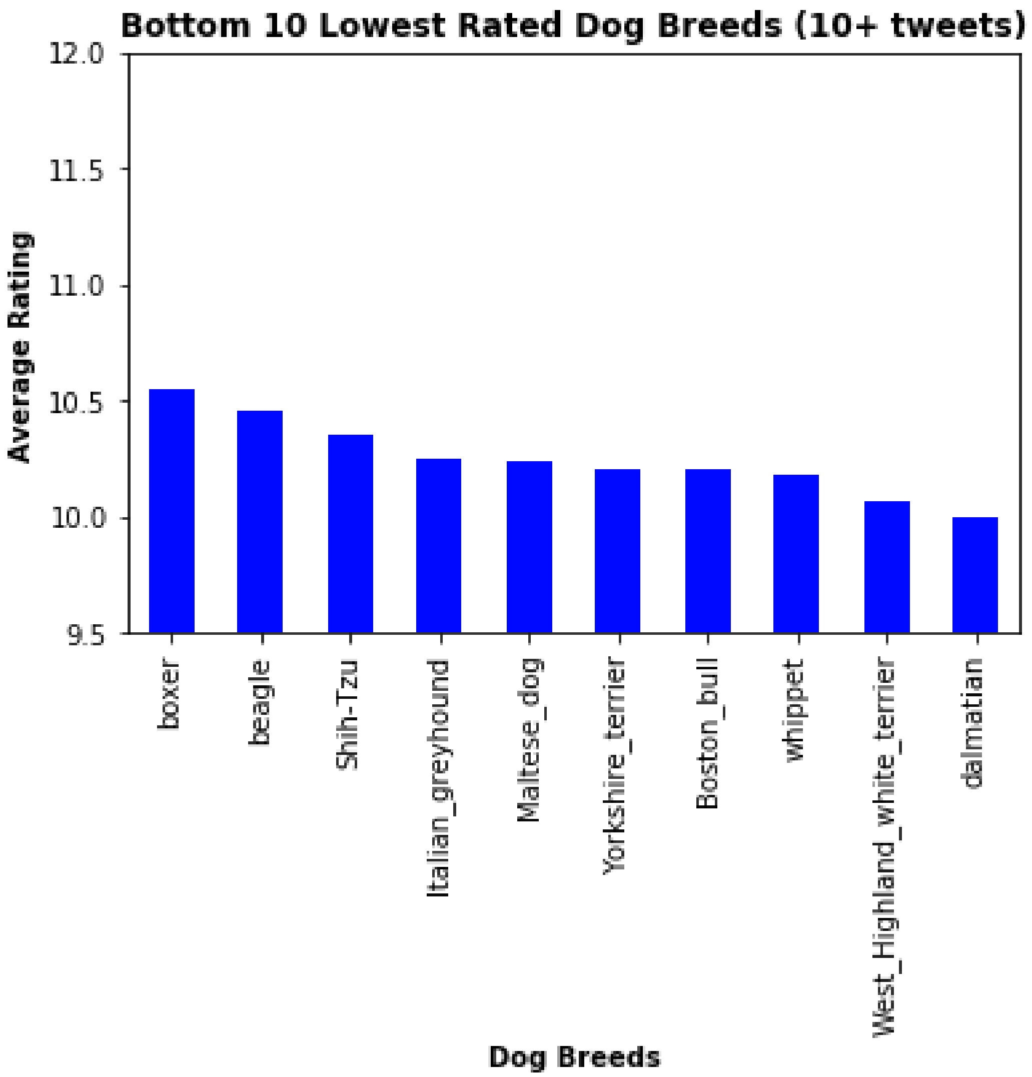
Here are the top 10 Highest Rated Dog Breeds:



In the WeRateDogs Twitter feed, the highest rated dog breeds are Samoyed, Pembroke and Chow.

Bottom 10 Lowest Rated Dog Breeds

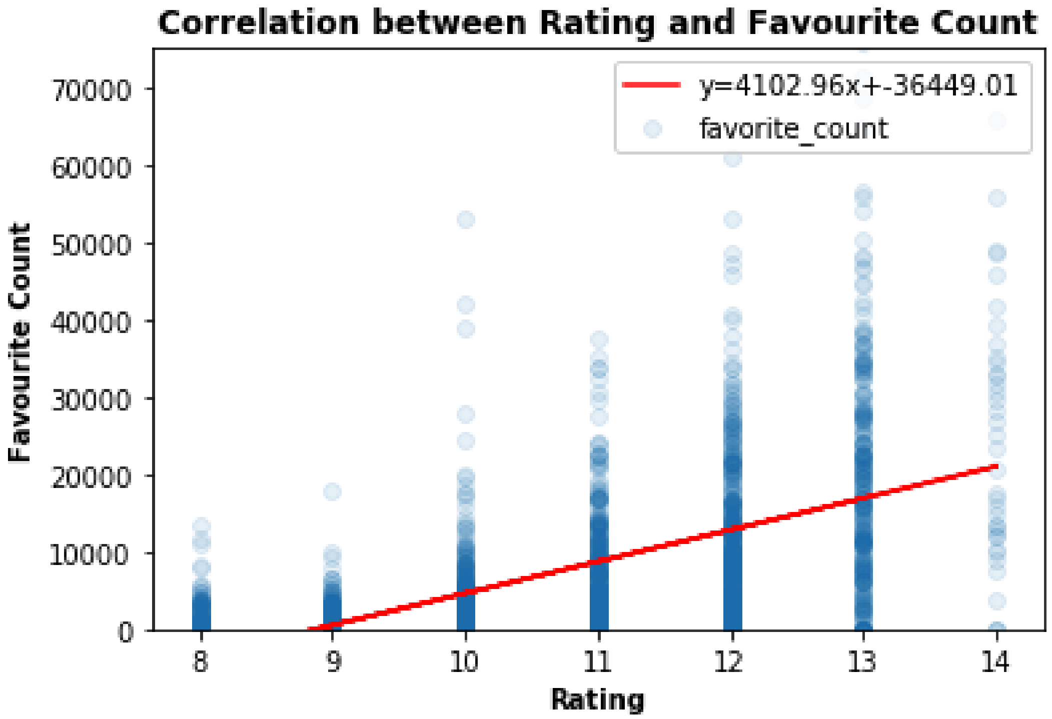
This insight used the same methodology as the Top 10.



The lowest rated breeds are the Dalmatian, West Highland White Terrier and Whippet respectively.

Relationship Between Rating and Favourite Count

Since WeRateDogs isn't a particularly serious review platform, I wasn't expecting there to be any correlation between rating and favourite count.



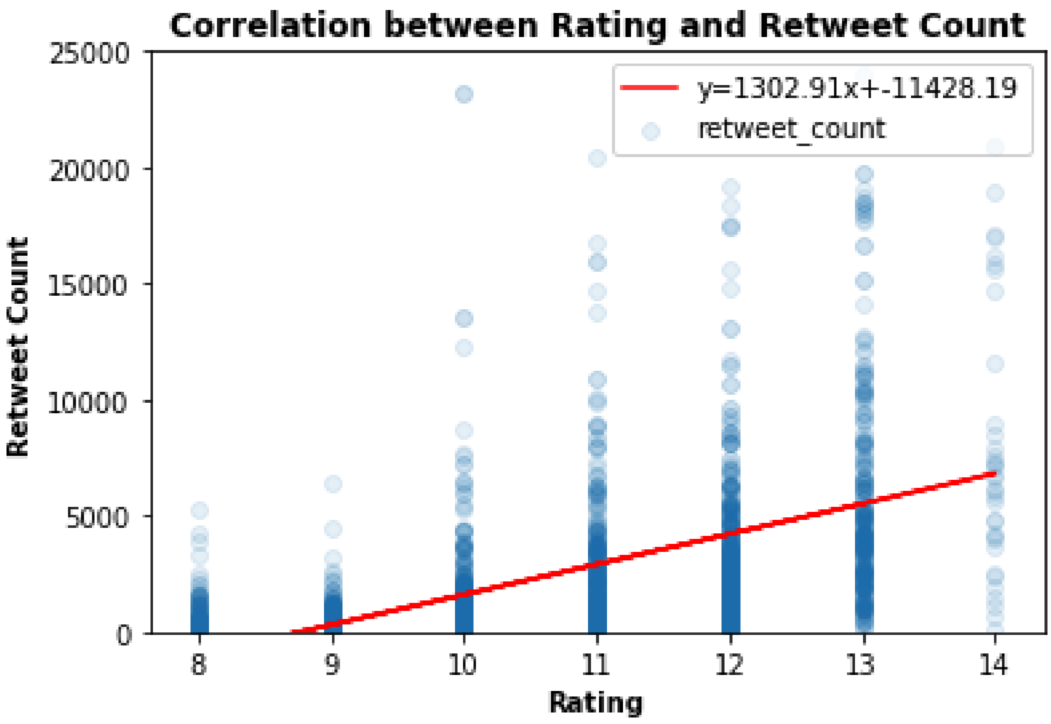
For some reason there does seem to be a correlation. Since this project is primarily to test data wrangling ability I have not conducted a full investigation into why this may be the case. Instead I offer the following theories:

1) Ratings and the number of people who interact with the Twitter account have both increased over time
2) People are genuinely influenced to favourite based on seemingly arbitrary ratings

If this were an analysis-focused project, I would also want to measure whether this correlation is statistically significant.

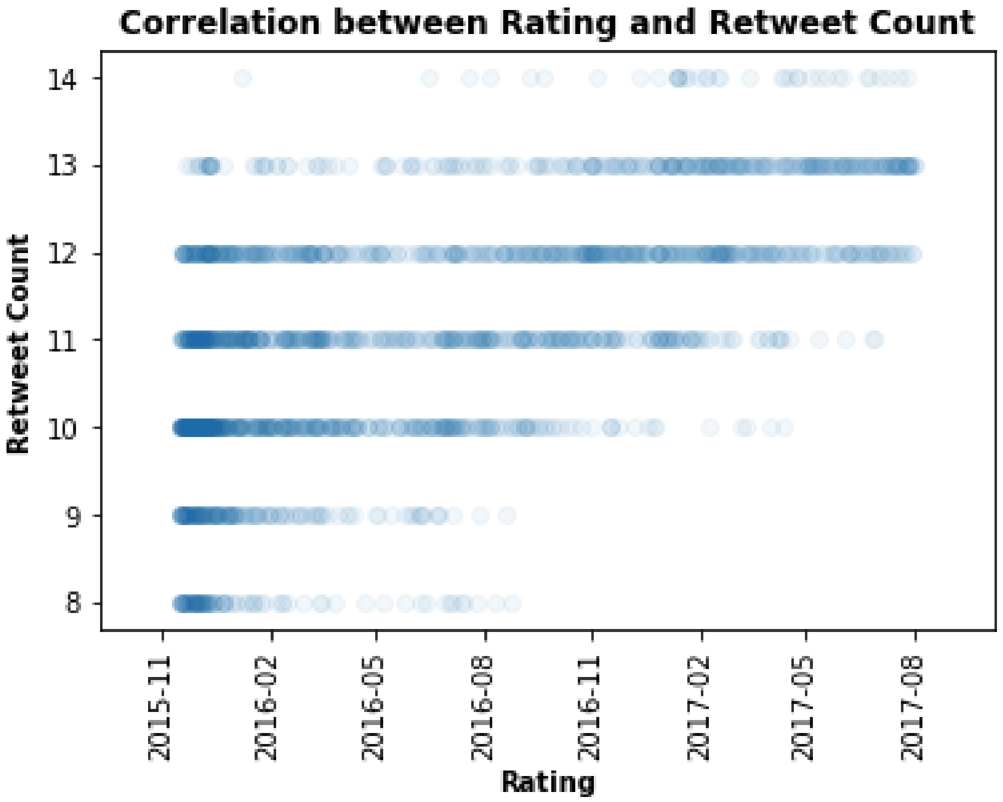
Relationship Between Rating and Retweet Count

The findings here were very similar to the relationship between rating and favourite count.



Changes in Average Rating Over Time

Since there was a correlation between rating and favourite/retweet count, I decided to check whether or not there had been any changes to rating over time.



It seems that ratings have been slowly increasing from November 2015 to August 2017. The degree to which this affects the correlation between rating and favourite/retweet count would be interesting to investigate.