Introduction

Real-world data rarely comes clean. For this project, I've downloaded a variety of data from the WeRateDogs Twitter archive. WeRateDogs is a humorous twitter account that rates peoples' photos of their beloved canines. This article will summarize how the data has been wrangled and analysed.

Gathering

For this analysis, 3 sources of data were collected. These sources include:

- WeRateDogs original twitter archive. This CSV file was provided to us by Udacity.
- Tweet image predictions data, which was also provided by Udacity.
- Retweet counts and favorite counts obtained from the Twitter API using Tweepy

Assessing

The data was visually and programmatically assessed for problems. There were multiple issues regarding quality and tidiness of the data that needed to be addressed. Some of quality issues include:

- Missing values in columns
- Values that are duplicated (retweets)
- Values that are non ascii characters
- Inaccurate names and ratings
- Different formatting for the same variable
- Wrong data types

Some of the tidiness issues include:

- Multiple columns for the same variable
- Multiple sources of data in separate datasets

Cleaning

The cleaning process of this data includes:

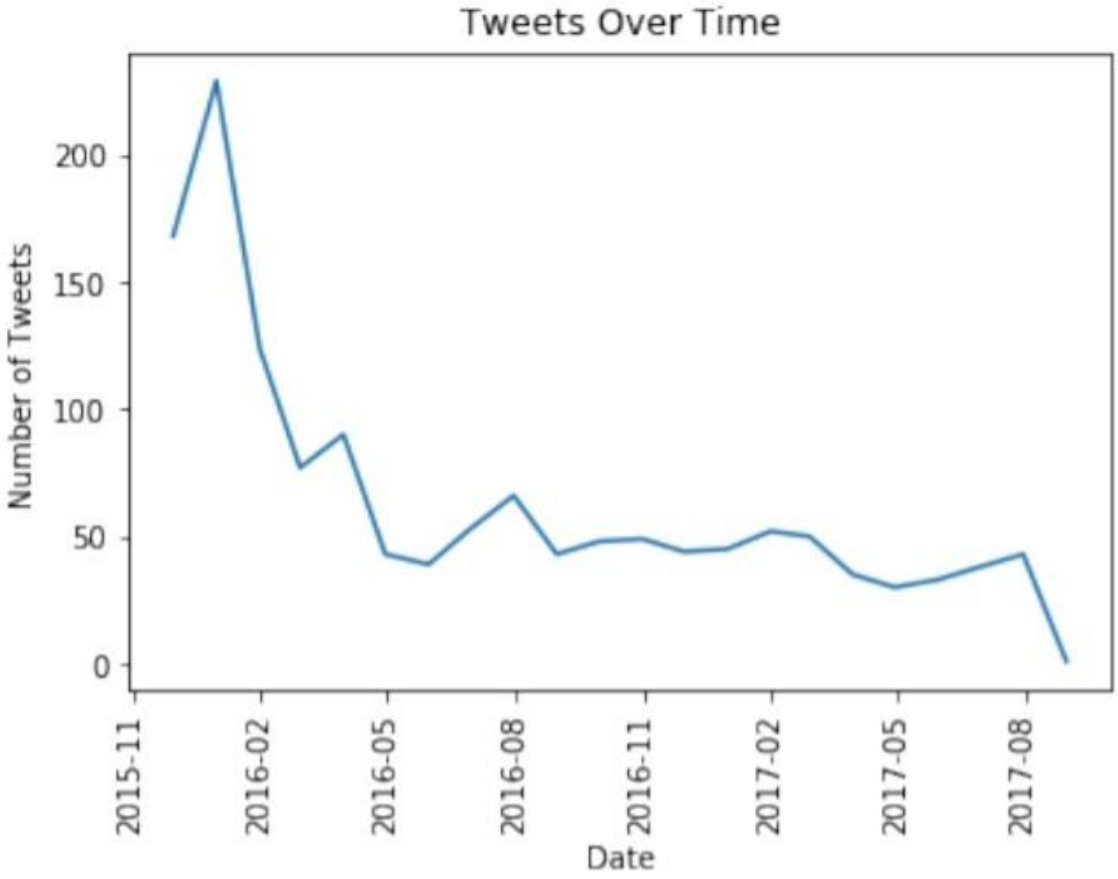
- Filtering out unnecessary, inaccurate, and missing data
- Create a standard format for the same variables
- Creating a standardized rating system
- Changing inaccurate data types
- Merging columns that measure the same variable
- Merging different data tables into one dataset

Analysing and Visualizing

The following includes five areas of analyses from the WeRateDogs data set.

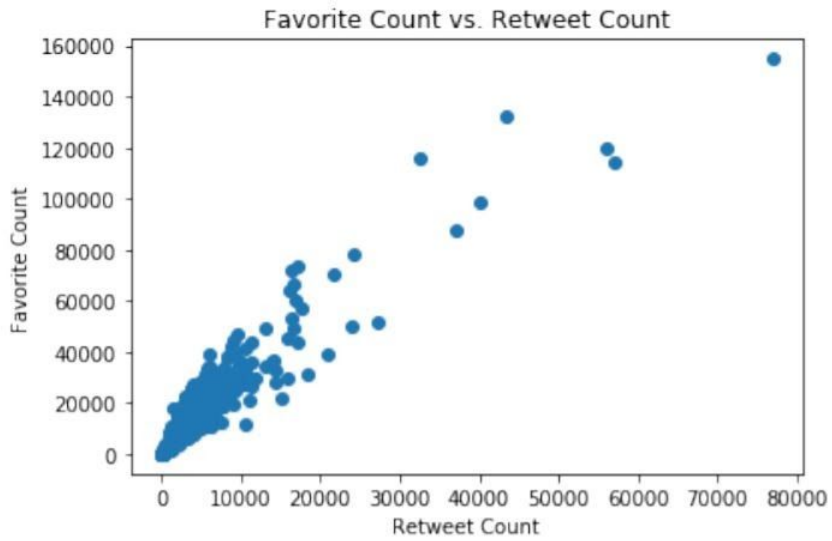
Tweets Over Time

The following diagram shows the number of tweets over time between 2015 and 2017. According to the line graph, the number of tweets spiked between late 2015 and early 2016, then took a sharp decline in early 2016. The number of tweets started to steadily decline beginning around March, 2016. More data is needed to assess whether this trend has continued beyond 2017.



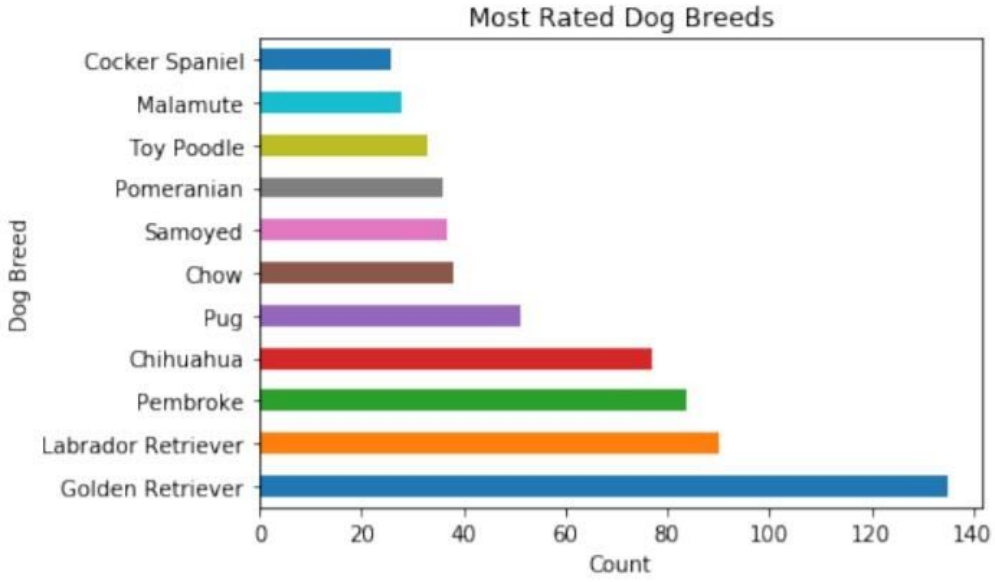
Favorite Counts vs. Retweet Counts

This scatter plot shows the positive correlation between favorite count and retweet count of the WeRateDogs photos. This makes sense considering that “liked” photos have a higher probability of being retweeted.



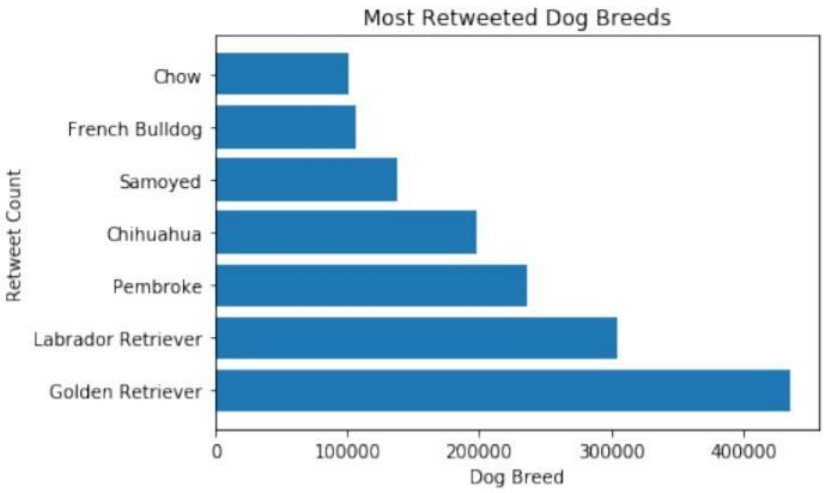
Most Rated Dog Breeds

The following bar chart shows that the most rated dog breed on the WeRateDogs website is the Golden Retriever. Labrador Retriever, Pembroke Corgi, and Chihuahua are the next most rated dog breed.



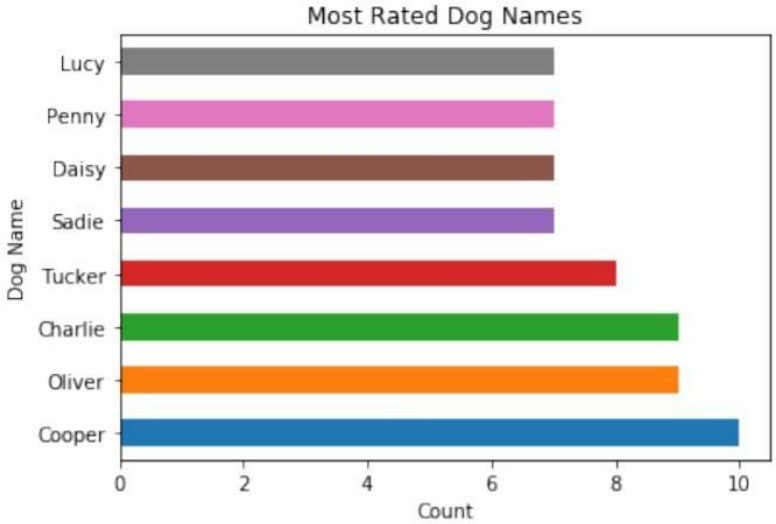
Most Retweeted Dog Breeds

Comparing the most retweeted breeds to the most overall rated breeds (in the above chart) reveals slight variations between the two. The first four breeds are the same between both charts, however the Samoyed seems to be slightly more popular in the retweets than the overall breeds. The French Bulldog doesn't even appear in the top ten most rated breeds, but comes in 6th on the retweeted breeds chart.



Most Rated Dog Names

The most common name on the WeRatedogs website is Cooper. Oliver and Charlie are tied for second most common names and Tucker is the third most common name on the WeRateDogs website. Sadie, Daisy, Penny, and Lucy are all tied for the fourth most common names.



Conclusion

The above analyses offer a small glimpse into the data wrangling and analysis process. The data was quite “messy” when first obtained, but through assessing and cleaning, it produced a clear and accurate summary of the WeRateDogs datasets.