Analysis

Augmentation

The cleaned data is reloaded from the twitter_archive_master.csv file. The data is augmented with 2 columns as to facilitate further analysis.

1. a rating expressed as a single number

2. The combining of all tweets with a dog breed with less than 10 occurences under " other"

The rating expressed as a single number makes it easier to compare different dog breeds and dog types based an their mean rating. The bundling of breeds under “other” allows for a readable plot, since it removes 60 individual dog breeds from the list. These 60 breeds had less than 10 occurrences each and as such would not offer reliable means for comparison.

Preliminary analysis

Preliminary analysis shows some outliers in the `rating` data, to counter this these outliers ar filtered in a new dataframe, the result can be seen in Figure 1Average rating per dog type.

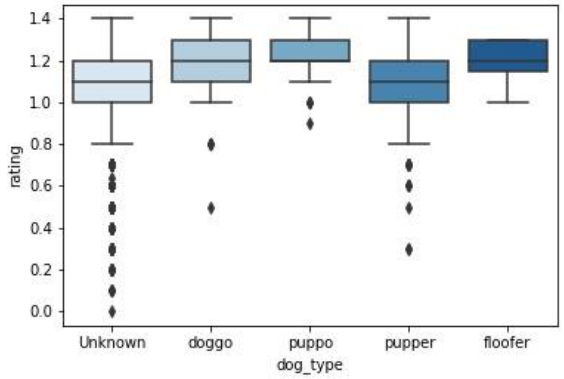

Figure 1Average rating per dog type

Moreover initial plots show relatively little distinction in dog type or dog breed with regards to the average ratings, as can be seen in Figure 2 average rating per dog type and dog breed.

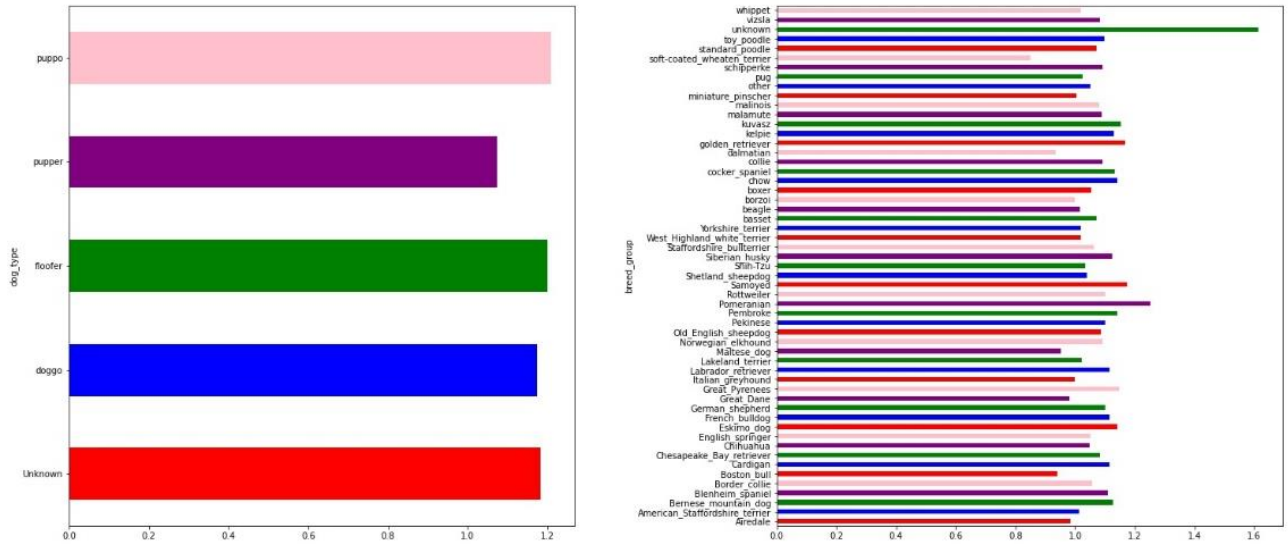

Figure 2 average rating per dog type and dog breed

In the retweet counts and favorite counts we do see distinctions between the dog types and breeds, where several dog breeds score higher on retweet counts (Figure 3 dog types and dog breeds retweet counts) and also on favorite counts (Figure 4 dog types and dog breeds favorite counts)

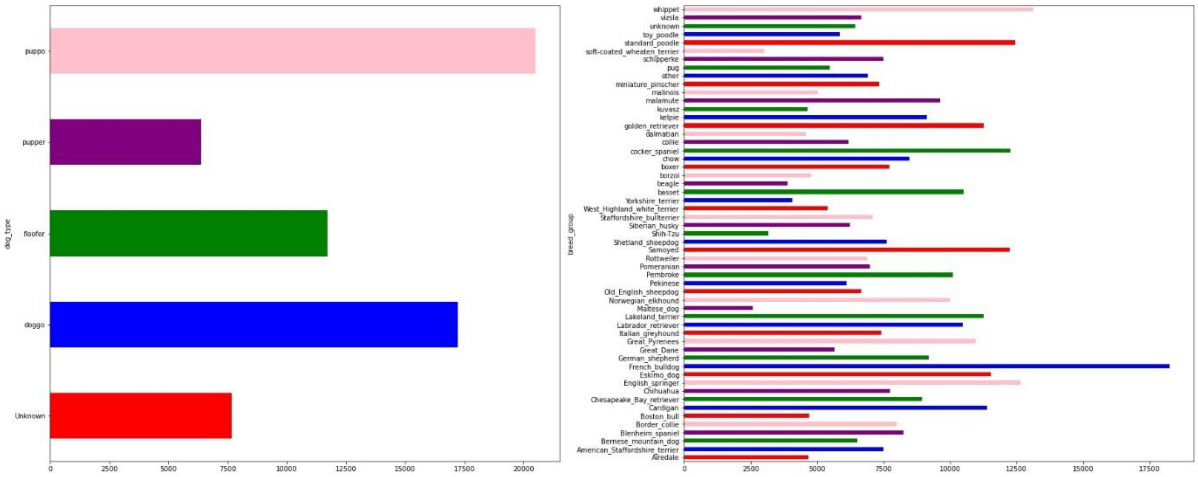

Figure 3 dog types and dog breeds retweet counts

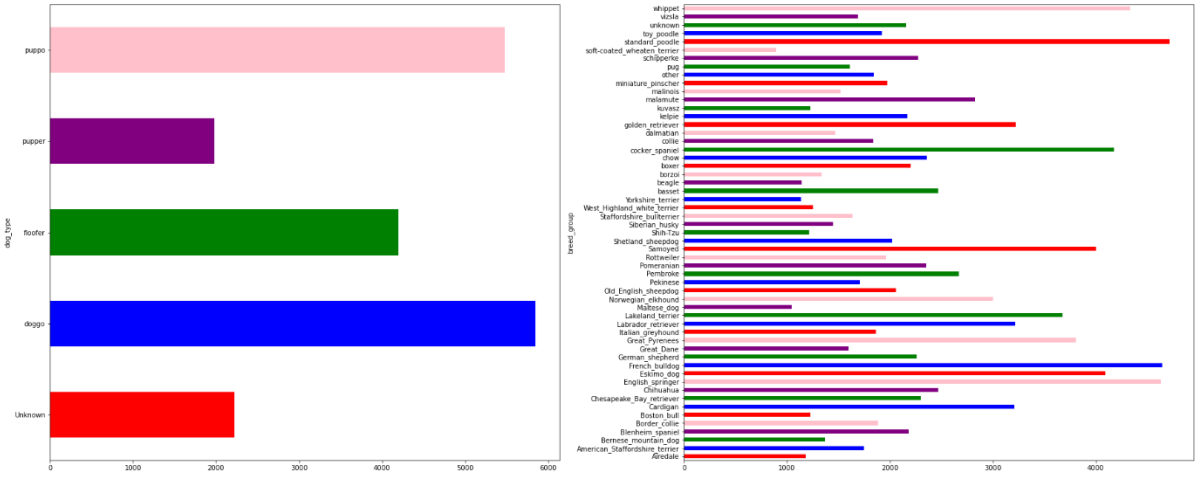

Figure 4 dog types and dog breeds favorite counts

Because of this initial analysis the favorite and retweet counts are further studied statistically.

Statistical analysis

A statistical analysis of the favorite and retweet counts can be levied by first assessing the possible correlation between favorite and retweet counts through a scatterplot.

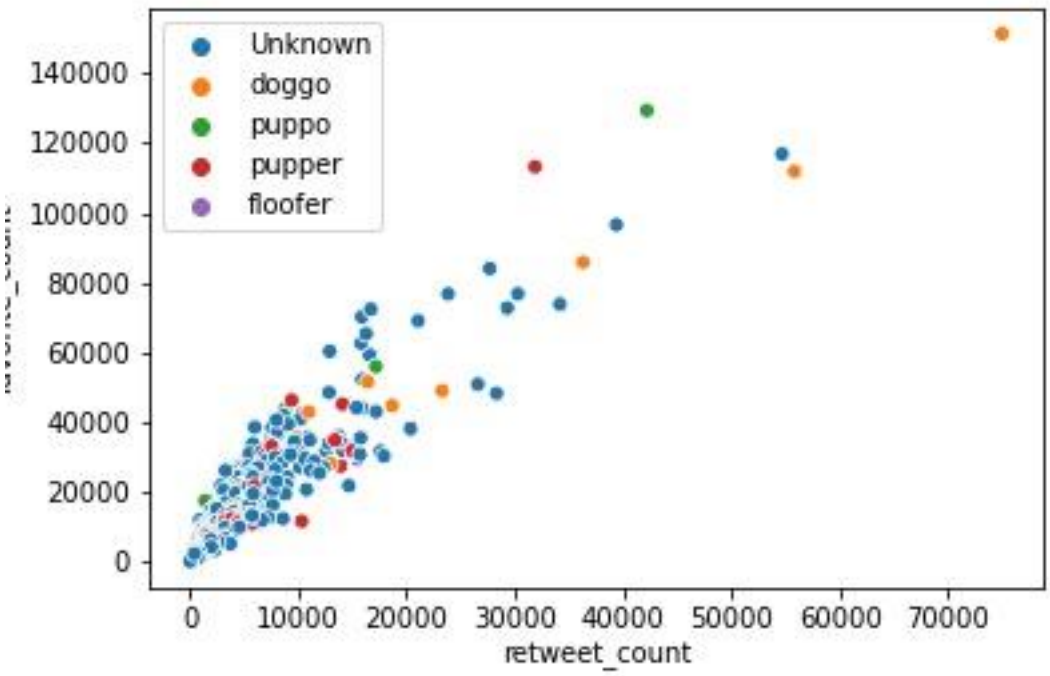

Figure 5 scatterplot favorite count vs retweet count

This shows that favorite and retweet counts appear to be correlated, and furthermore that the dog type does not seem to be of influence in the correlation.

To check the correlation we can see in the following heatmap:

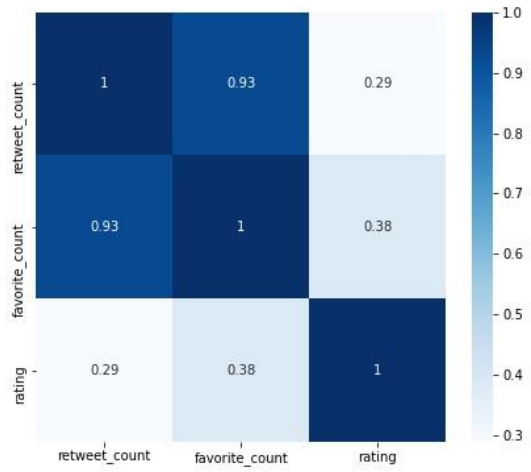

Figure 6 correlation heatmap

That favorite count and retweet counts are indeed highly correlated. This seems logical as twitter users are most likely inclined to both like and retweet any post they come across that they actually like.

Conclusions

The analysis confirms that favorite count and retweets are highly correlated and we can see from the scatterplot that there is no large distinction between dog types as to this correlation.