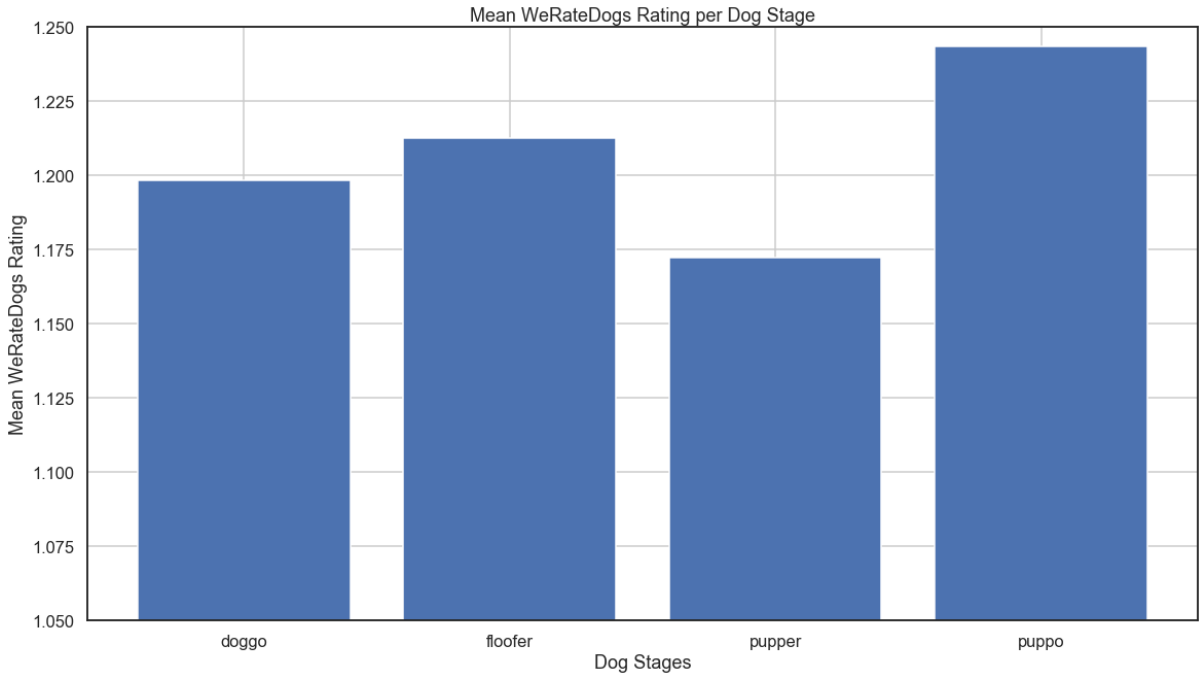
Wrangle and Analyze Data of a Twitter Account

This project is an analysis of the WeRateDog account. Main goal of this project is to follow through the complete Data Analysis process. Firstly data is gathered from different sources: a local csv-file, a programmatically downloaded file and data downloaded via the twitter API. Secondly the data from the different sources is visually and programmatically assessed to be cleaned in the next step for the final analysis. The analysis compares the ratings, retweets and favorites of different dog stages and races.

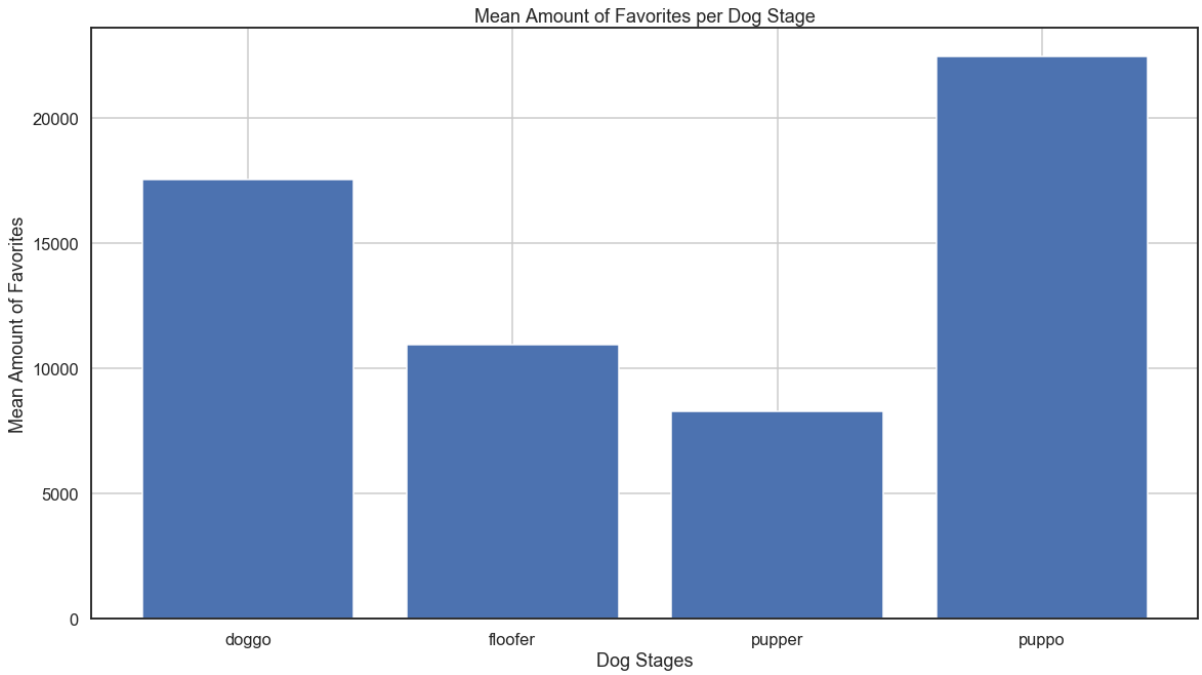
The resulting data from the gathering process contains information of the WeRateDogs Twitter archive, the tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network and each tweet's retweet count and favorite ("like") count.

Now analyzing the data I wanted to find out if the owners of the WeRateDogs account have a preference of one specific dog stage? In other words: is there a difference in mean rating between the different dog stages `doggo`, `floofer`, `pupper`and `puppo`? The rating is an evaluation of the dog photo given by the owners of the twitter account WeRateDogs for each dog depicted in their twitter post. A rating consists of a numerator, usually between 10 and 14, and a denominator of 10. Dividing numerator and the denominator results in a value between 1.0 and 1.4. Outliers were removed.

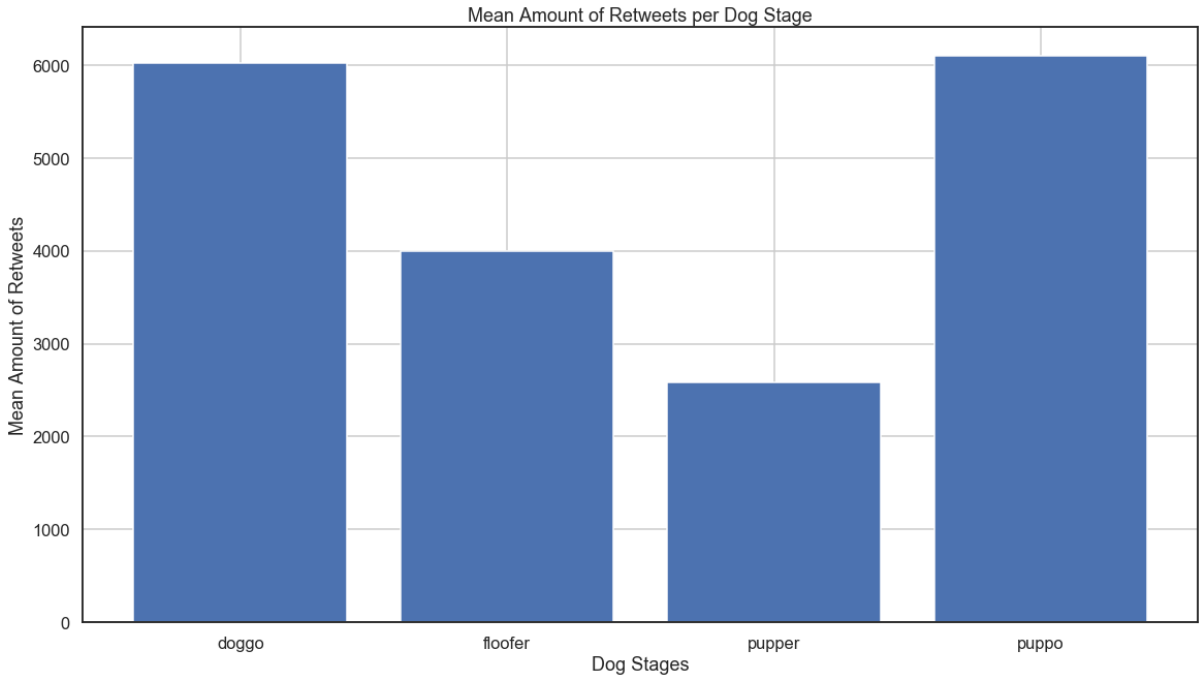


The graphic above “Mean WeRateDogs Rating per Dog Stage” depics that the dog stage `pupppo` in mean has the highest rating with 1.210714. The stages `doggo` and `floofer` are close to that with mean ratings of 1.172222 and 1.188889. The lowest mean rating with a signifcant difference has the stage `pupper` with 1.068841.

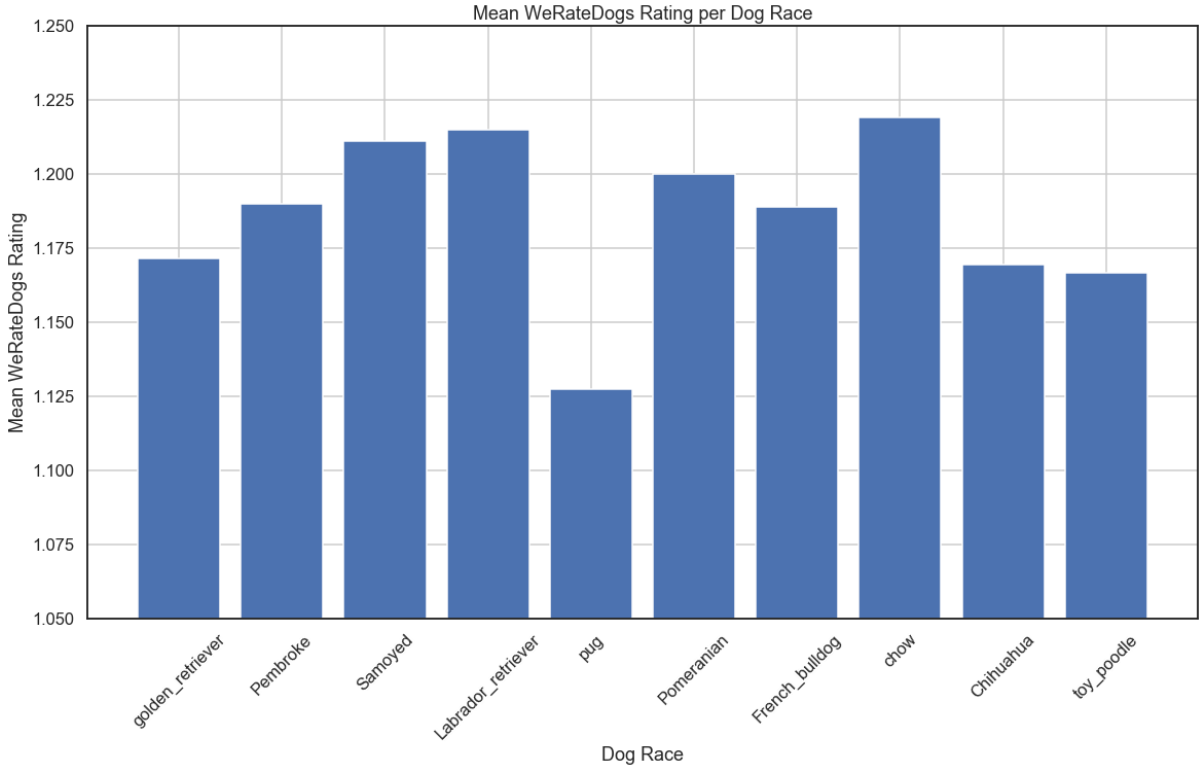
How does the rating given by the WeRateDogs twitter account correspond to the rating given by followers of that account. We can only measure amount of retweets and favorite clicks. Does the observed difference in mean rating also apply to the favorite count?



The graphic above “Mean Amount of Favorites per Dog Stage” shows that as in mean stage rating, the mean stage favorite count for stage `puppo` has higher mean values. The stage `pupper` has, like in mean stage rating, the lowest mean favorite count. Only the stage `floofer` has higher ranking than in mean stage ranking. Does the observed difference in mean rating and mean favorite count also apply to the retweet count? Retweet count, just like favorite count, is a indicator of popularity among the followers of the twitter account.

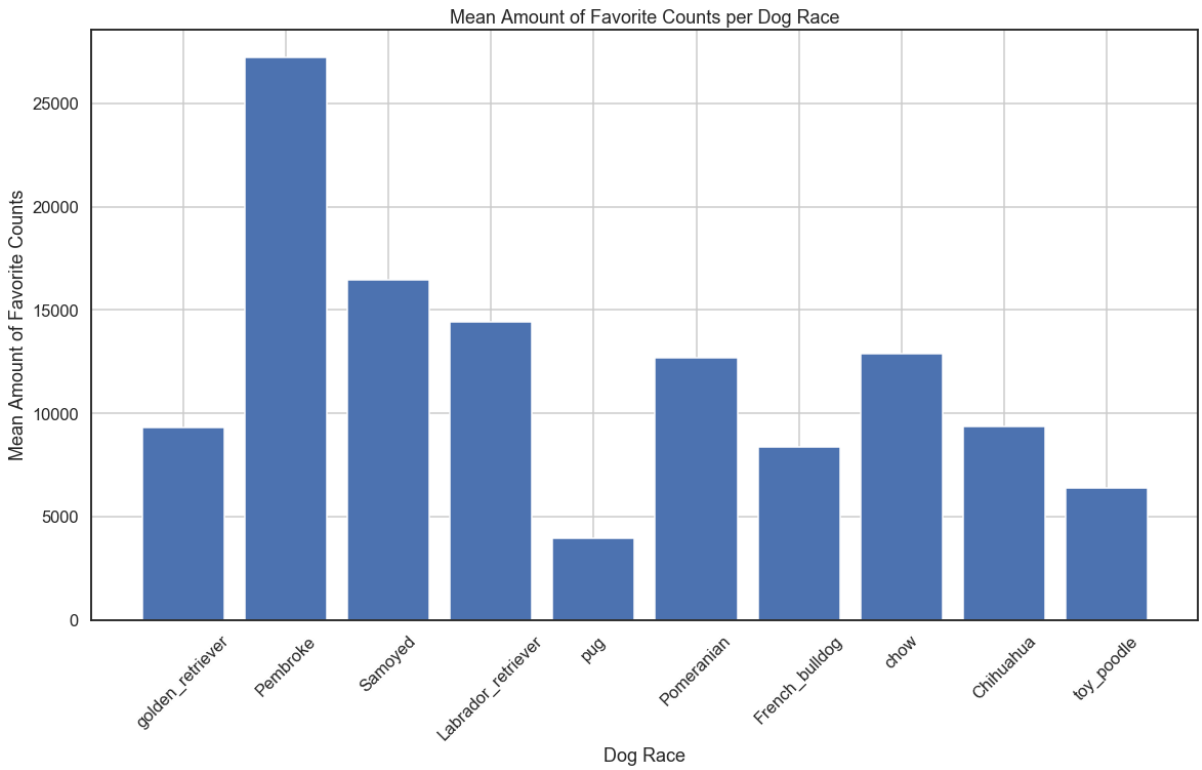


In all three indicators, mean stage rating, mean stage favorite count and mean stage retweet count, the stage `puppo`performs best and the stage `pupper` performs worst, as seen in the graphic above “Mean Amount of Retweets per Dog Stage”.
Another category we can order our dog tweets is by dog races. Do certain dog races get rated higher than others?



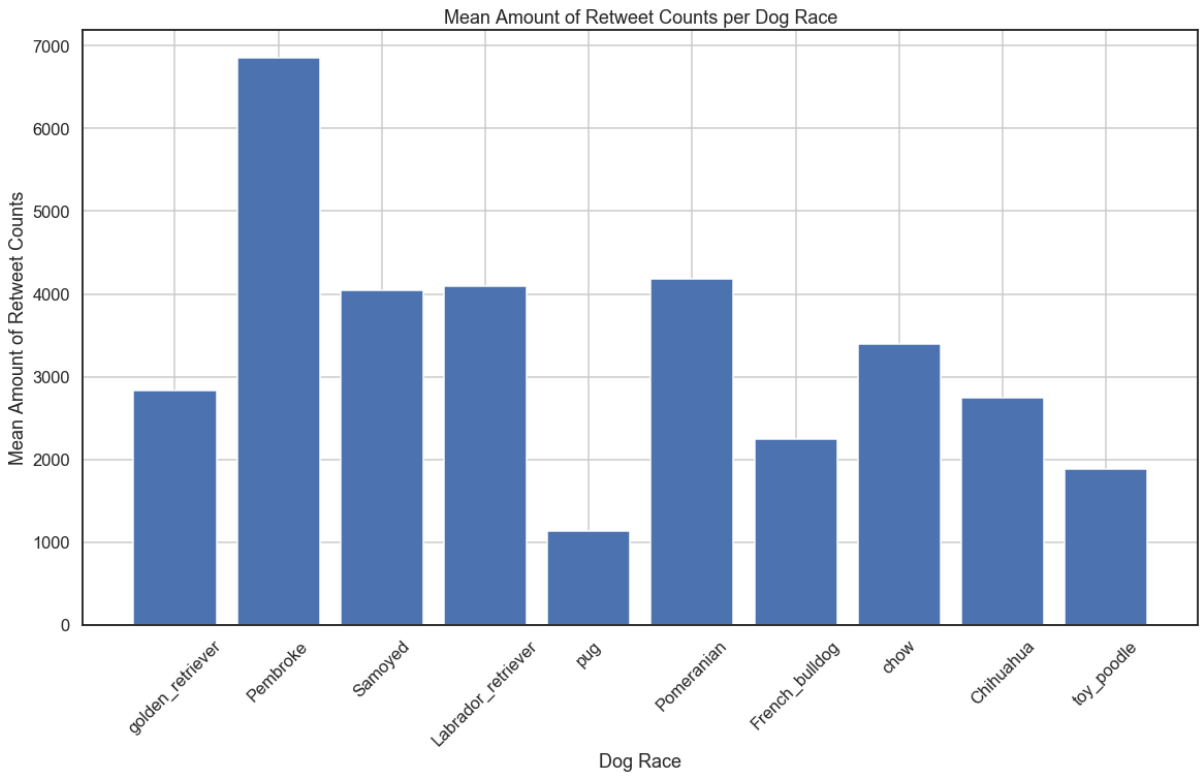
As depicted in the graphic above “Mean WeRateDog Ratings per Dog Race”, among the dog races more than five times predicted with a certainty higher than 85% by the machine learining algorithm, the mean ratings of toy Labrador Retrievers, Chows and Pembrokes are the highest. Pugs, with difference, received the worst mean ratings among the analyzed data. Here are only regarded dog races, where there were more than five tweets, which the machine learning account could predict as the same.

Do the same dog races, that have higher mean WeRateDog ratings, also have higher mean amounts of favorites? Hence, do the owners of the WeRateDog twitter accounte have the same preferences as the followers of the same account?



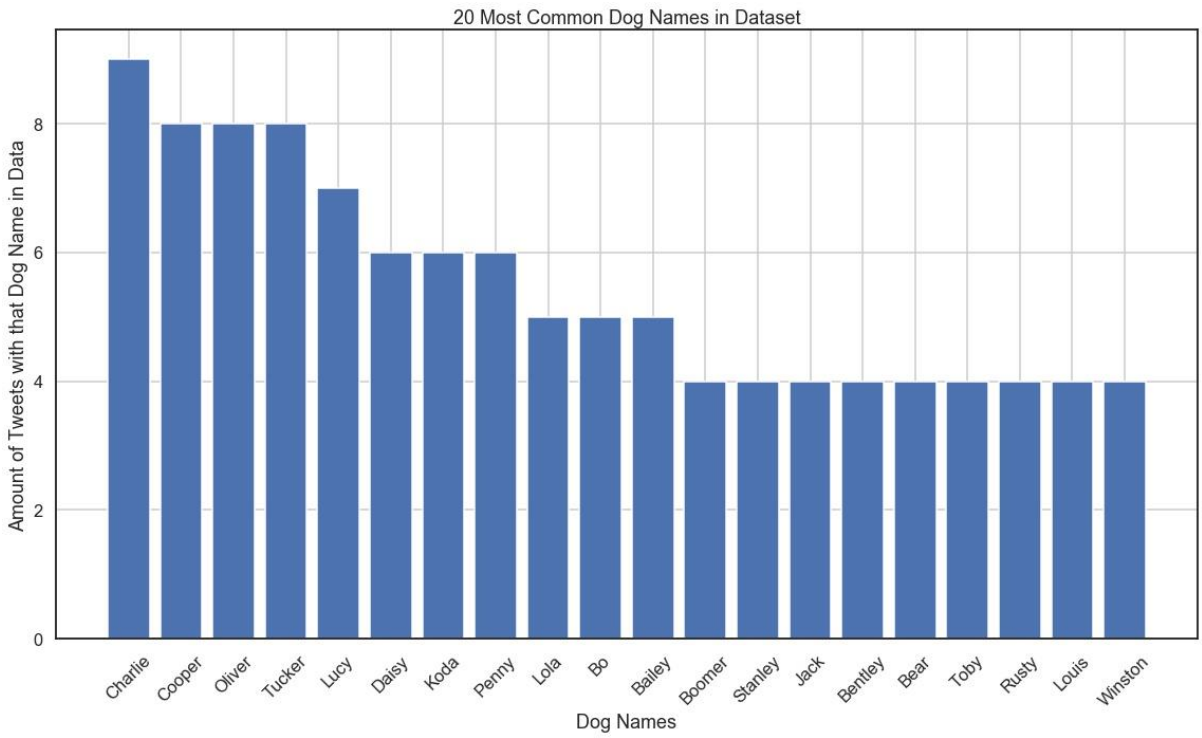
The mean amount of favorite counts per race show a complete different picture than the mean dog rating per race. The mean amount of favorite counts for dogs of Pembroke race are significantly higher than the rest of races. The mean amount of favorite counts for dogs of race pug are significantly lower than for other dog races.

Is the ranking of dog races by mean amount of retweet counts consistent to the ranking by mean amiunt of favorite counts? As both indicators are influenced by the followers of the WeRateDog account the distribution is expected to be similar.



There is no significant difference in the ranking by mean amount of retweets and mean amount of favorites per race. The followers of the WeRateDog account behave consistently in retweeting and clicking favorite of WeRateDogs tweets.

Finally, as for some dog tweets we could extract the name of the dog from the tweets text, we can ask the question: Which is the most common dog name among the data collected from the WeRateDog twitter account?



The most dog name in the dataset is Charlie.