# WeRateDogs: Wrangling, analyzing, and visualizing

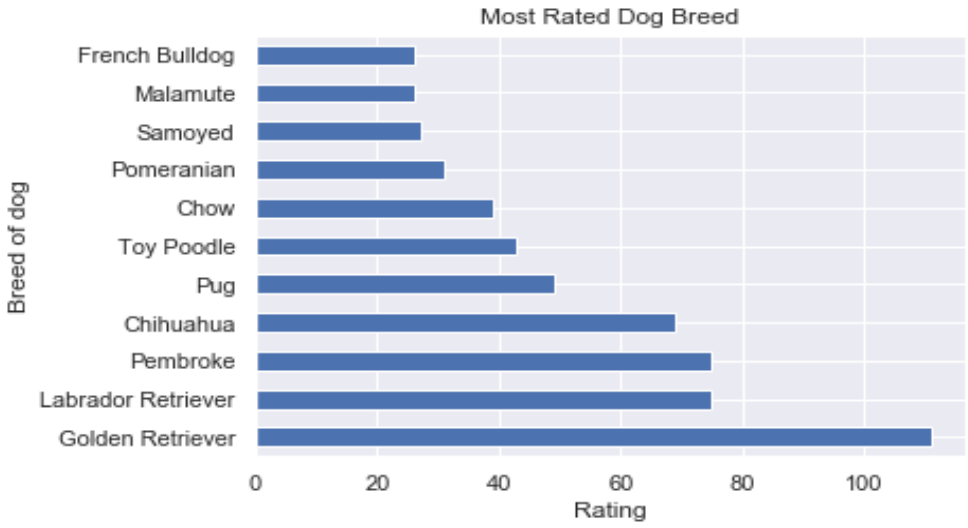
**By Zakariya Boutayeb**

**WeRateDogs** is a popular Twitter account where followers can post humorous comments and photos about a dog and its rating. Active since November 2015 with over 7.6 million followers by January 2019, WeRateDogs has grown beyond being a twitter account to launch their own store with related dog products.

Udacity made the data relating to WeRateDogs available as part of Data Analyst Nanodegree Program which include only tweets up to August 2017, and these tweets were filtered to remove retweets which are tweets with dog breed missing and tweets where there was no image.

Our data analysis process for this project will provide a step-by-step guidance, starting by gathering data from several different sources and in a variety of formats, assessing their quality and tidiness, then cleaning them and combining them all to create a master dataset which will be used to answer a series of interesting questions through analysis and produce stunning visualization and/or modeling using Python and its libraries.

### What is the most popular dog breed?

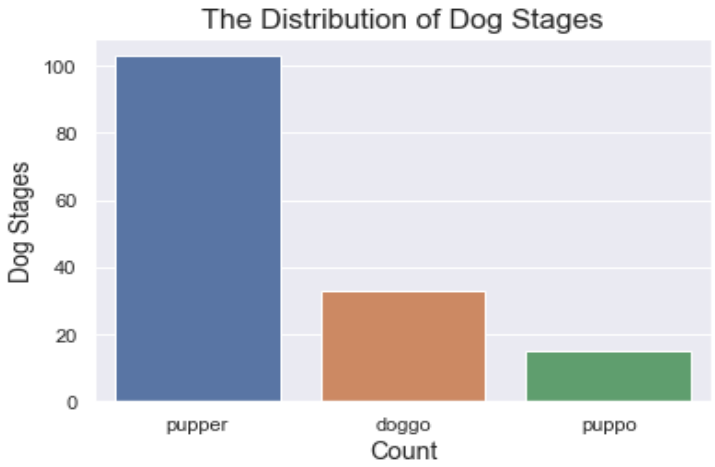



The most popular dog breed is a Golden Retriever (111), followed by Labrador retriever and Pembroke (75), then comes Chihuahua in the third place (69).

And the winner is Sam.
Sam is a Golder Retriever.
Sam has the best rating with 3.4, who
smiles 24/7 and secretly aspires to be a
reindeer.
Keep Sam smiling by clicking and
sharing his link:
https://t.co/LouL5vdvxx

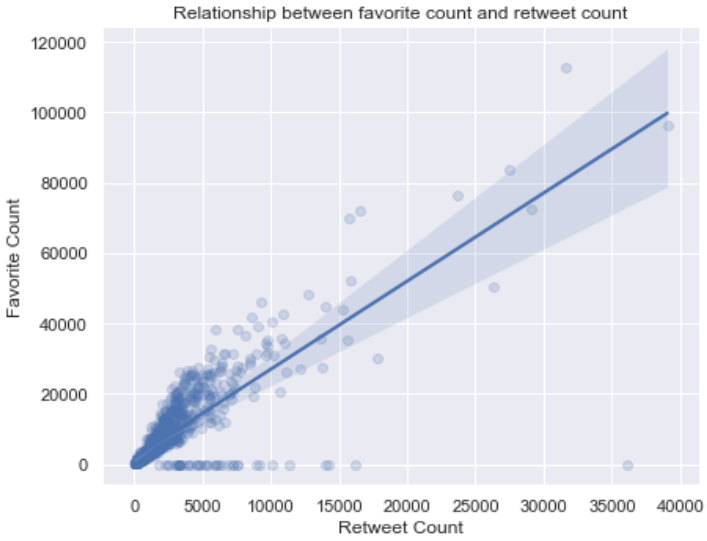
**What is the most frequent Dog stage?**

**WeRateDogs** has developed its own dog development stages based on the age and appearance of a dog. For instance, puppies are **puppers**, older puppers are **puppos**, then comes **doggos** and hairy dogs are **floofers**.




Pupper is the most frequent reported dog stage (245), followed by doggo (83), while puppo (29) and floofer (9).

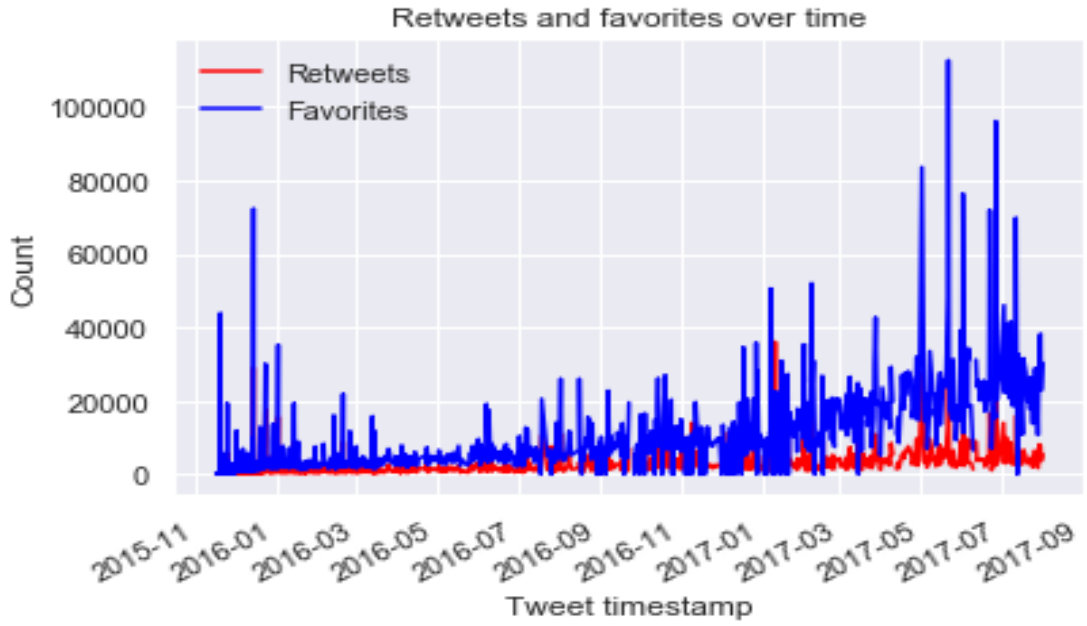
**What's the relationship between the number of retweet and favorite count?**




The scatterplot shows that there is a strong relationship between favorite counts and retweet counts.

The coefficient of correlation between the number of retweet and favorite count is positive and it is around 0.85 which means that a dog breed with the highest retweet tends to have higher favorite account.

The distribution below shows how many retweets as well as favorites (likes) is increasing over time.




We can conclude that favorite counts are higher than retweets and their trends is increasing over time. It worth noting that the highest ratings do not receive the most retweets.

To conclude, I can say that Data Wrangling is incredibly challenging and time-consuming as they are many challenges with gathering, selecting, and transforming data to answer analytical questions.

Bear in mind that one of the limitations to draw perfect conclusion is that a poorer quality of the database can potentially be costing higher price to the final findings. The database was untidy, may be because the data was collected from various sources, the reason why there were many null and missing values, outliers, duplicate values, imputing values, data imbalance, and data encoding. Moreover, our 3 datasets should be cleaned and assessed and then merged in one master DataFrame before being analyzed which leads that many tweets were excluded from our analysis.

However, Data Wrangling is worth the effort, knowing how various data values might impact your final analysis. In addition, Gathering and Cleaning process requires conducting research on the appropriate way to extract data from an URL or Twitter API which was a big challenge for me as a future Data Analyst who value learning, doing research will definitely equips me with knowledge about data quality and consistency.