The WeRateDogs Project

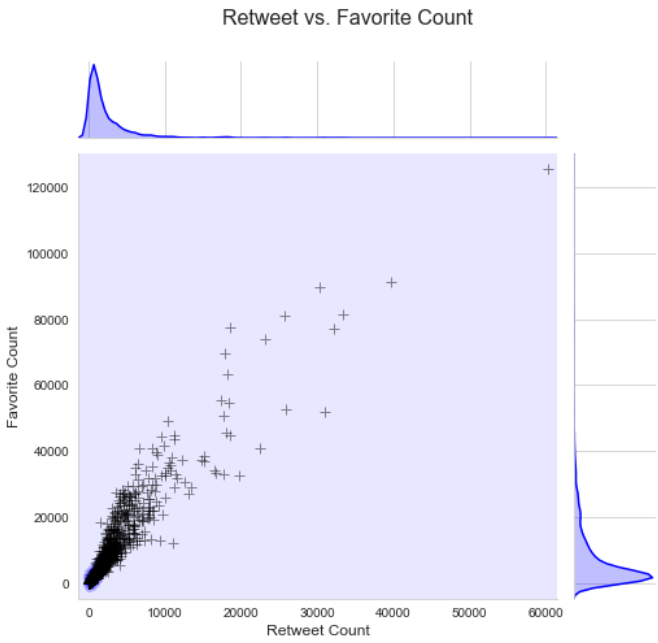
A journey from messy data to auspicious analysis

By Jeremy Sung
Date: May 1, 2019

WeRateDogs (@dog_rates) is a Twitter account that rates people's dogs with a humorous comment about the dog. This Twitter account has over 7.6 million followers, as of Dec 2018, since its debut in 2015 and has received international media coverage for its popularity and for the attention drawn to social media copyright law when it was suspended by Twitter. [1]

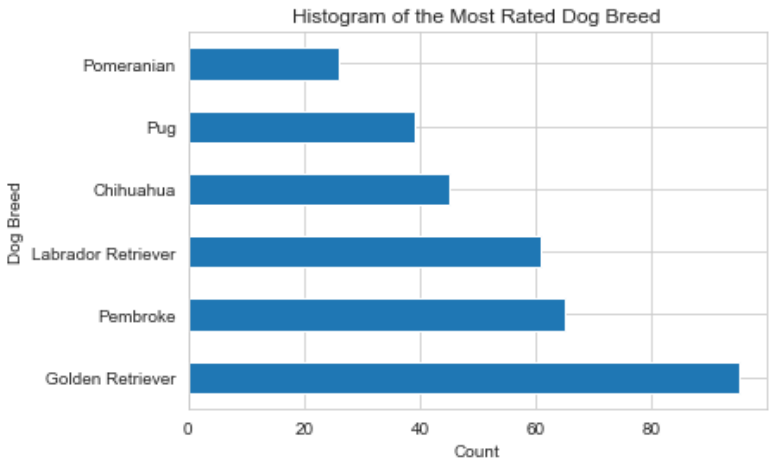
The purpose of this project is to retrieve insightful information from a carefully curated, though still messy, dataset, the WeRateDogs Twitter archive. Throughout the process I had a chance to exercise data wrangling related tasks and practice techniques including
- data gathering via Twitter API and via an URL with assistance of Python requests module,
- handling JSON format to store locally and load to dataframes,
- cleaning data, and visualizing some aspects of the data.

I learned some of the industry best practices in preparing data for further interesting and trustworthy analyses. I also realized that, while the evolution in machine learning algorithms and techniques such as CNN and PyTorch, etc. have made quite a huge progress in recent years meaningful results from analyses still heavily rely on the quality of data. To deliver trustworthy analyses human intervention is often inevitable especially in key decision to ensure the delivery of trustworthy results. With that said, systematic approach to a problem including messy data is still the idea and automation is the future of any operations in the big data and AI era.



Favorite vs Retweet Count

At the time of this dataset, WeRateDogs had over 4 million followers. **As shown in this figure the retweet and favorite counts are strongly correlated.** Roughly every 3 to 4 favorite tweets triggered 1 retweet.



The Top 6 most rated Dog Breeds are shown in this diagram:

Golden Retriever (95),
Pembroke (65),
Labrador Retriever (61),
Chihuahua (45),
Pug (39),
Pomeranian (26), etc.

Human intervention is needed before certain decisions can be handled more accurately by complicate algorithms. For example, while looking into the content of some least popular tweets I noticed that, among 10 lowest rated tweets, only 3 tweets talked about dogs. What caught my attention is one of them in tweet_id 667549055577362432 was actually classified as “Paper Towel”, instead of a dog, in its p1 prediction though the confident level (p1_conf) was indeed not very high, only 32.80%.

As you may see above, I tend to believe the object in the center of the picture look more like a Maltese lying on a carpet with the same color as its hair. It could be a stuffed toy but it is in a puppet shape.

This little surprise led me to expand my filter criteria to include those pictures not classified as Dogs yet with low confident levels of their prediction falls, such as < 40%.
It turns out my “linear” approach to this misclassification issue was not very effective. Although the new criteria did include the misclassified Maltese dog (False Negative) it also introduced a cricket which was classified as a Tick but with an even lower Confident Level, 24.25%, on its p1 prediction.

In summary, data wrangling is essential in data analyses. While we invest a great amount of resources in gathering and cleaning data the effort can be more efficient and optimized along with our domain knowledge and continuous systematic approach to the tasks. And, before more accurate algorithms or tools are furnished visual inspection still plays an authentic factor in the process.