DATA GATHERING, WRANGLING, AND EXPLORATION WITH WERATEDOGS TWITTER ARCHIVE DATA.

Introduction

Pet lovers have one thing in common, and that is doting on their pets. Dog owners and lovers are not left out. So being thoughtful, WeRateDogs created a niche account focused on commenting on pictures and videos of dogs posted by their owners, while rating them with their unique rating system. Given the attention that WeRateDogs has received by the dog-loving public, their Twitter account contains vast amount of information relating to dogs and how they score on their rating system. Using their Twitter archive data, I performed data gathering, wrangling and exploration to distill some useful insights. The following contains a summary of my work on the WeRateDogs data. For all the codes used for this analysis, refer to the jupyter notebook link below.

Data Gathering

Data was gathered from three sources for the project. First, I manually downloaded already retrieved WeRateDog Twitter archive data from this **link**. Then using ``Requests`` library, data about image predictions of dogs in the Twitter archive was also downloaded. Finally, using Twitter API, additional data were programmatically retrieved from WeRateDogs Twitter archive, to ensure we have more information for analysis.

Assessing Data

A visual and programmatic assessment of the data was carried out to identify any data quality or tidiness issues. Some of the problems identified include duplicate values, incorrect datatypes, missing values, and excess data columns. Each quality and tidiness issue were documented during the assessment process to guide the data cleaning efforts.

Cleaning

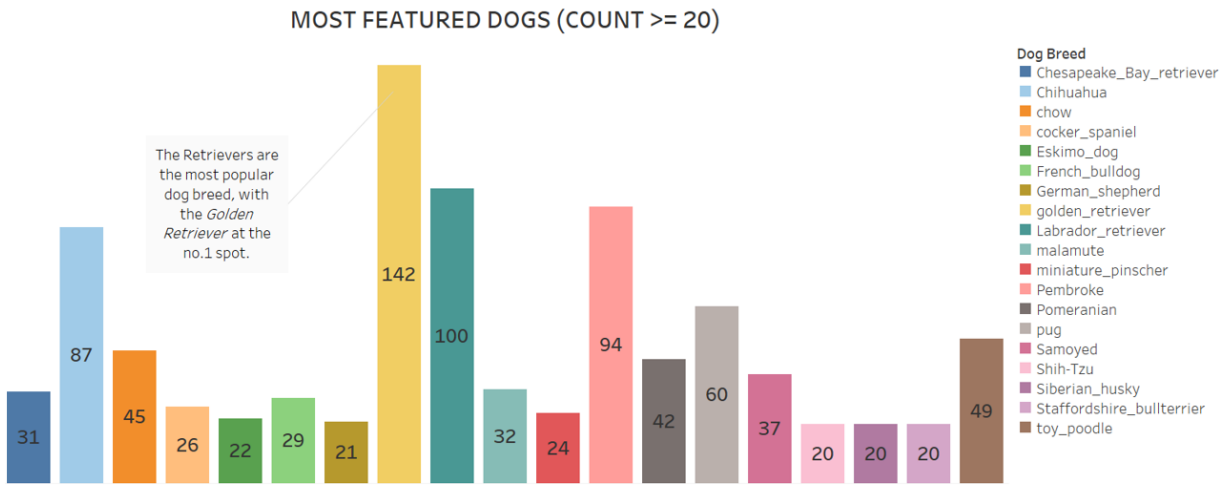
Using a “define”, “code” and “test” approach, each cleaning process was properly documented and carried completed. Define states the approach to cleaning a data quality or tidiness issue, a code to implement followed, and finally a test to validate the cleaning. In total, 11 data quality and 5 tidiness issues were cleaned and documented. A master dataframe containing merged data columns from the three dataframes we gathered earlier was produced. A copy was exported as a CSV file and we proceeded to analysis and visualization.

Insights

We set out to discover which dog breed get the most ratings on average and which breed is commonly posted by the owners.

1. On average, the Japanese Spaniel gets the least rating, compared to the Soft-Coated-Wheaten-Terrier which gets the most rating among all the dog breeds.
2. The retrievers in our dataset seem to be the people's favorite. They are the most common dog breed in the data. Golden retrievers are number one, followed by the Labrador retriever. The Japanese Spaniel is among the least common dog breed in the dataset, even though they get the most rating on average by WeRateDogs; there must be something special about them.
3. In terms of most liked and retweeted dog post in the period covered in our data, a Labrador retriever comes first. The dog at a "doggo" dog stage, got a rating of 13/10, was liked 144247 times, and retweeted 70331 times.

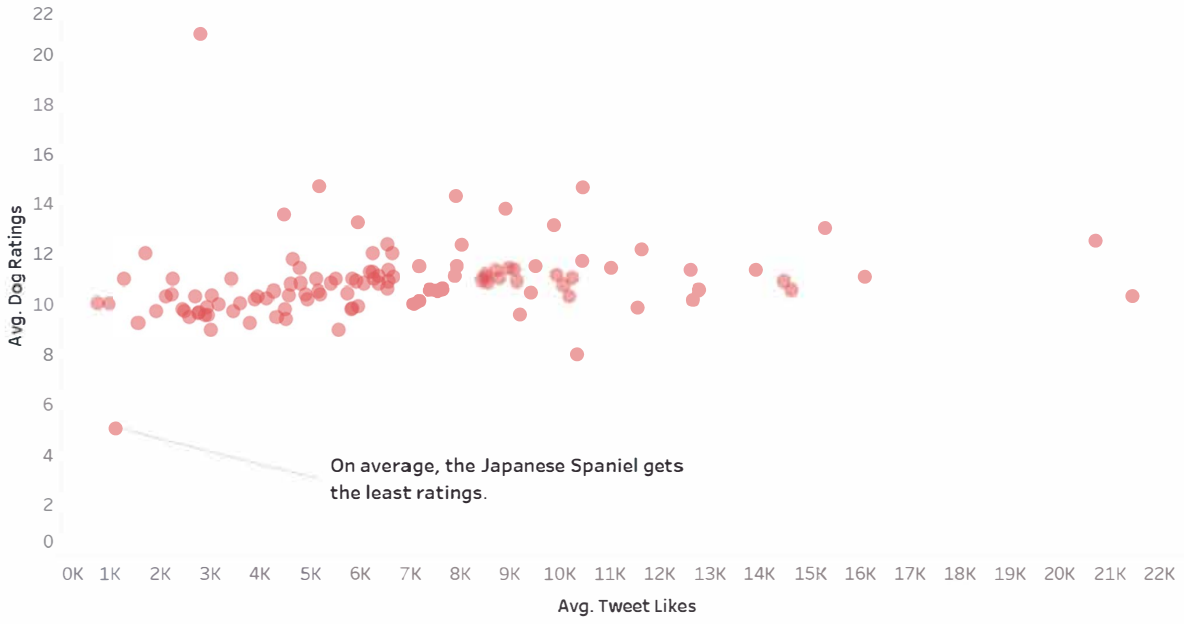
Viz



The first viz captures the number of most featured dog breeds in the dataset. Using Tableau's filter function, we selected dog breeds that appeared 20 or more times in the data. As we already saw in our analysis above, the vizualizations shows clearly that the retrievers are the most popular breeds in the dataset.

The visuals below shows the average rating, tweet likes and retweets for each dog breed. We can see there is a cluster of the ratings around 10-12. The diagrams also point to evidence that high dog ratings does not equate to high likes or retweets for a dog post.

DOG RATINGS & TWEET LIKES



DOG RATINGS & RETWEETS