Wrangle Report

We Rate Dogs Dataset

By:

Daniel Cucinotta

Introduction:

Data was gathered, assessed, cleaned, and analyzed as well as utilized for insights with visualizations. The finalized dataset was compiled from several sources, which includes Twitter's API, and Udacity's servers.

Data Gathering:

- The twitter_archive_enhanced.csv file was directly downloaded from Udacity's servers.
- The Requests library was imported to download the image_predictions.tsv content.
    - The image_predictions.tsv content was downloaded and stored in a variable, then read into a Pandas dataframe.
- The Tweepy library was used to query additional data via the Twitter API.
    - The data was saved as tweet_json.txt, and then converted into a Pandas dataframe.

Assessing Data:

The data assessment process revealed many errors, inconsistencies, and incorrect/invalid entries. 9 quality and 3 tidiness issues were defined to be resolved, to create a clean master dataset.

Quality Issues:

1. There are several incorrect datatypes for columns in df_twit2, img_predictns2, and twts_cnts2.

2. There are many incorrect or mislabeled dog names in df_twit2.

3. The 'floofer' stage name is incorrect in df_twit2.

4. Only original ratings should be in df_twit2.

5. There are inconsistent numerators and denominators in the rating system of df_twit2.

6. There are non-dog entries in the img_predictns2 dataset.

7. HTML tags are in the source column text of the df_twit2 dataset.

8. Dog breeds in 'p1', 'p2', and 'p3' columns of the img_predictns2 dataframe have inconsistent capitalization.

Tidiness Issues:

1. Merge the df_twit2, img_predictns2, and twts_cnts2 dataframes.

2. Condense the 4 specific dog stage columns into 1.

3. Remove duplicate columns after joining/merging dataframes.

4. Remove unnecessary columns.

Cleaning Data:

1. Change datatypes of some columns in 3 datasets that would be more appropriate for cleaning and analysis.

2. Remove incorrect names in df_twit2.

3. Change the column name in df_twit2.

4. Remove non-original ratings in df_twit2.

5. Remove incorrect/outlier values in df_twit2.

6. remove non-dog entries in img_predictns2 dataframe.

7. Remove the HTML tags in df_twit2.

8. Change all uppercase letters to lowercase in img_predictns2 dataframe.

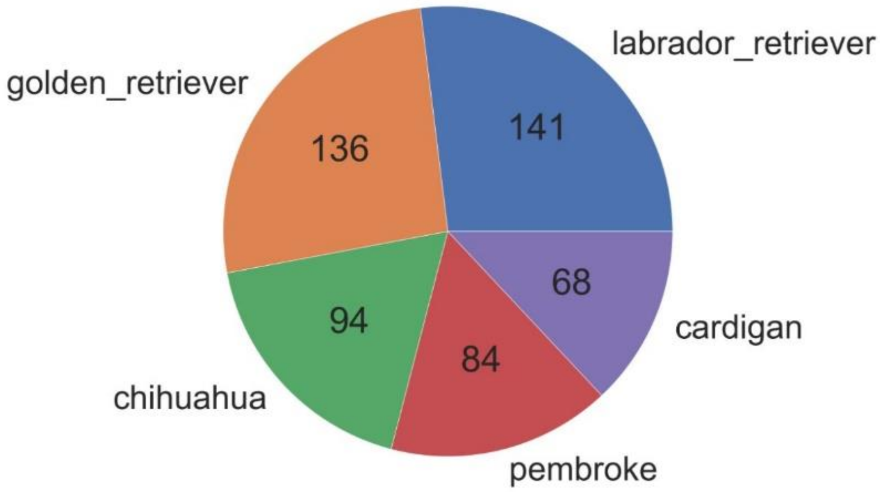
WeRateDogs Twitter Account Analysis Insights and Visualizations Report

Daniel Cucinotta

The WeRateDogs Twitter account data archive has a vast array of data to be utilized for analyses and visualizations. 5 interesting insights have been defined, discerned, and displayed from this dataset, after successfully wrangling, analyzing, and cleaning.

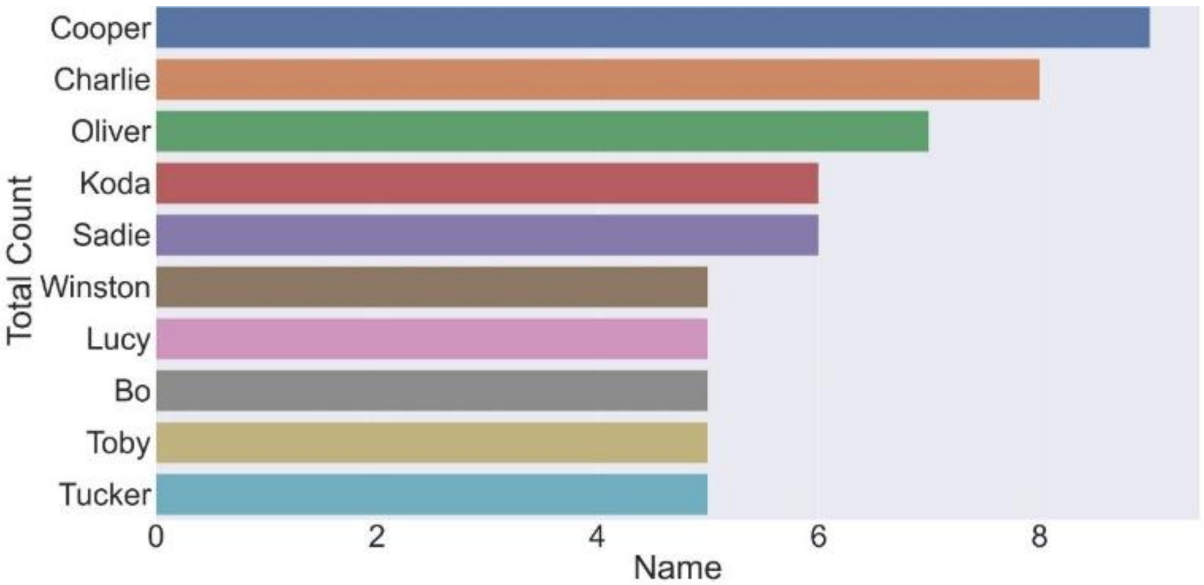
Insights and Visualizations:

1. Top 5 Most Common of the Highest Rated Dog Breeds.



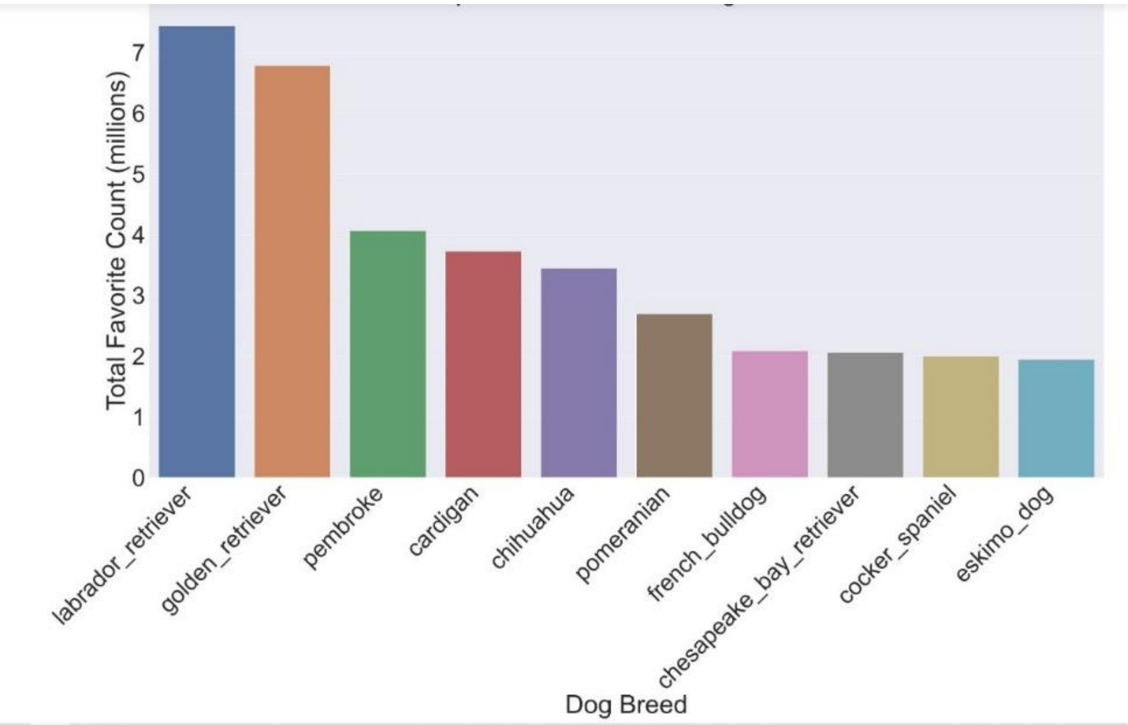
The Labrador Retriever and the Golden Retriever breeds have significantly higher values, respectively, than the next 3 most common of the highest rated breeds on the list.

2. Top 10 Most Popular Dog Names.



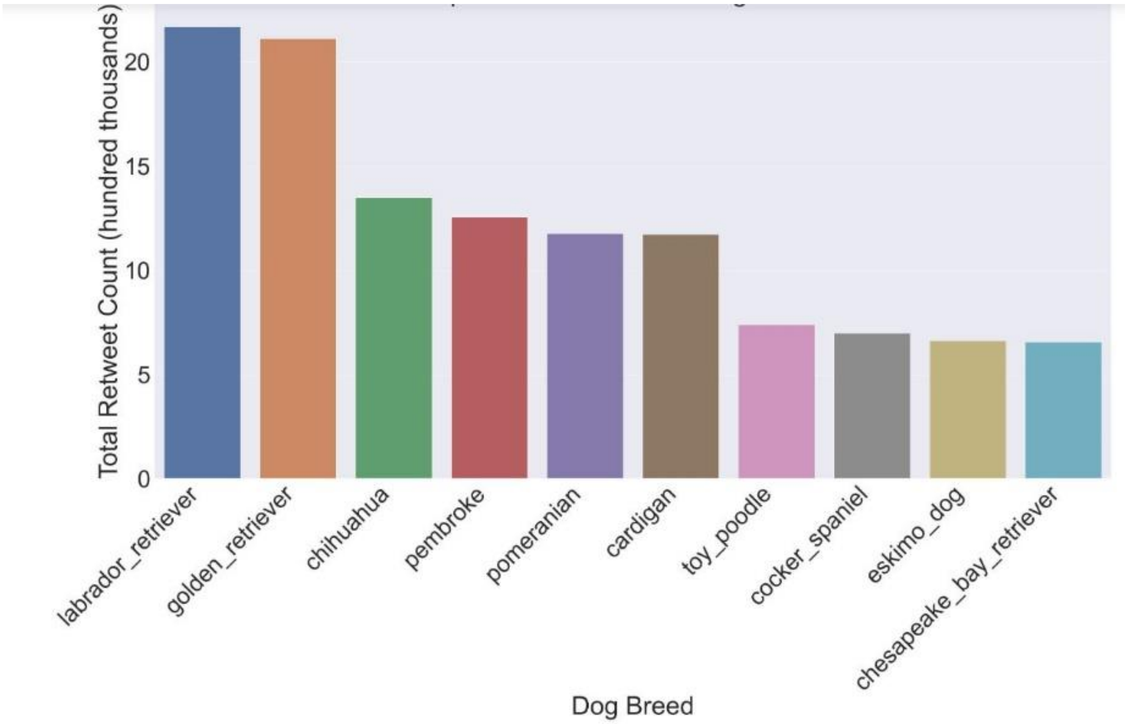
Cooper, Charlie, and Oliver are significantly more popular than all other names (especially last 5 in the list).

3. Top 10 Most Favorited Dog Breeds.



The favorite count for the Labrador Retriever and Golden Retriever breeds, respectively, are significantly higher than all of the other breeds.

4. Top 10 Most Retweeted Dog Breeds.



The retweet count for the Labrador Retriever and Golden Retriever breeds, respectively, are significantly higher than all other breeds.

Conclusion:

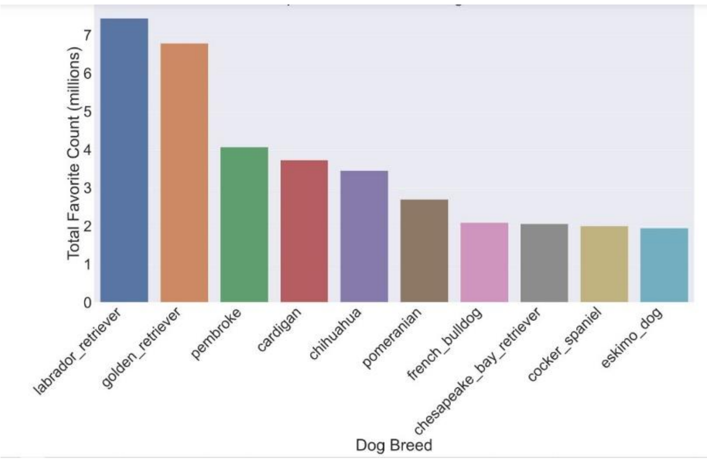
The examination of the insights and visualizations have revealed several correlations well as distinguished the most popular dog names that is contained within the dataset; Cooper, Charlie, and Oliver (respectively).

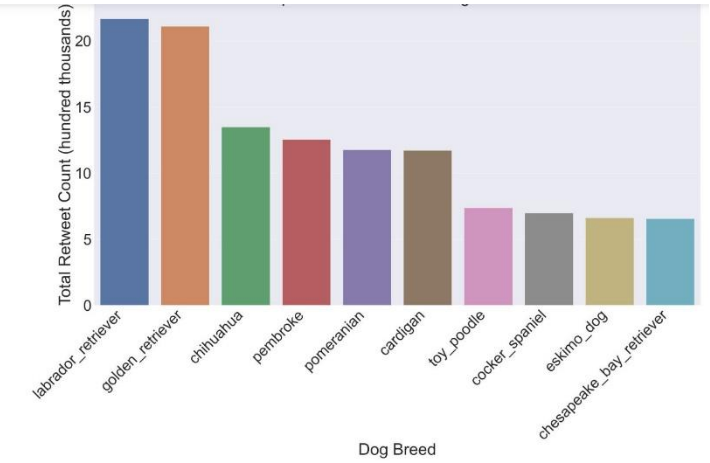
Labrador Retriever and Golden Retriever were number 1 and 2, respectively, for the most common of the highest rated breeds as well as the most retweeted and most favorited dog breeds within the WeRateDogs Twitter account.

For both insights (most retweeted and most favorited), the next 4 dog breed positions on the list (the 3, 4, 5, & 6 positions) are Chihuahua, Pembroke, Pomeranian, and Cardigan (the order varies between the two).

Furthermore, the last 4 positions on the list have nearly identical breeds as well, though, the most favorited breed has French Bulldog in the 7th position and the most retweeted breed has the Toy Poodle in the 7th position – the 8, 9, and 10 positions are the same 3 breeds, in differing order.

In summation, Labrador Retriever and Golden Retriever are the most common of the highest rated dog breeds as well as the most favorited and retweeted dog breeds. The top 10 (most retweeted and most favorited) dog breeds are nearly identical, having 9 out of 10 breeds matching.


Total Favorite Count (millions)
labrador_retriever
golden_retriever
pembroke
cardigan
chihuahua
pomeranian
french_bulldog
chesapeake_bay_retriever
cocker_spaniel
eskimo_dog
Dog Breed


Total Retweet Count (hundred thousands)
labrador_retriever
golden_retriever
chihuahua
pembroke
pomeranian
cardigan
toy_poodle
cocker_spaniel
eskimo_dog
chesapeake_bay_retriever
Dog Breed