Wrangle Report

By Mukund Kulkarni

OVERVIEW

WeRateDogs is a twitter handle that rates photos of dogs. It has its own unique rating system. Our task is to gather, clean, assess its data. .

GOALS

1. Gather the Data
2. Assess and Clean the Data
3. Store, Analyze and Visualize the Data

KEY POINTS

- We need only original tweets, no re-tweets.
- Clean at least 8 Quality issues and 2 Tidiness issues
- Produce at least 3 Insights and 1 Visualization.

MILESTONES

Gathering the Data

- The “The "twitter-archive-enhanced.csv" is a Twitter archive containing the data about the tweets of user @dog_rates also known as WeRateDogs. This file was provided directly.
- The "image-predictions.tsv" is a file that was generated using a model to detect the dog breed in the tweet's photos. This file is supposed to be downloaded using the requests module.
- The "tweet-json.txt" is created by querying the Twitter API using the tweepy module. The Twitter API is used to gather additional data about the tweet_id's in twitter-archive-enhanced.csv.

Assessing and Cleaning the Data

The data gathered was assessed both manually and programmatically. Various issues were discovered,

Quality issues cleaned

1. Re-tweets were removed by selecting rows where the “re_tweeted_status_id” was null.
2. Columns like “in_reply_to_status_id" and "in_reply_to_user_id" which were unnecessary were dropped.
3. The type of “timestamp” was changed to datetime.
4. Useless data from the “source” column was removed.
5. The problem with “rating_denominator” and “rating_numerator” was solved by either correcting the value or dropping the row itself where the data was inconsistent.
6. In the “names” column only the values representing actual names were kept, rest were changed to NaN.
7. The “image-predictions” file was cleaned , to contain only the tweet_id and the breed of the dog.
8. The “breed” names were made more readable.
9. Unnecessary columns from “tweet-json” were dropped.

Tidiness issues cleaned

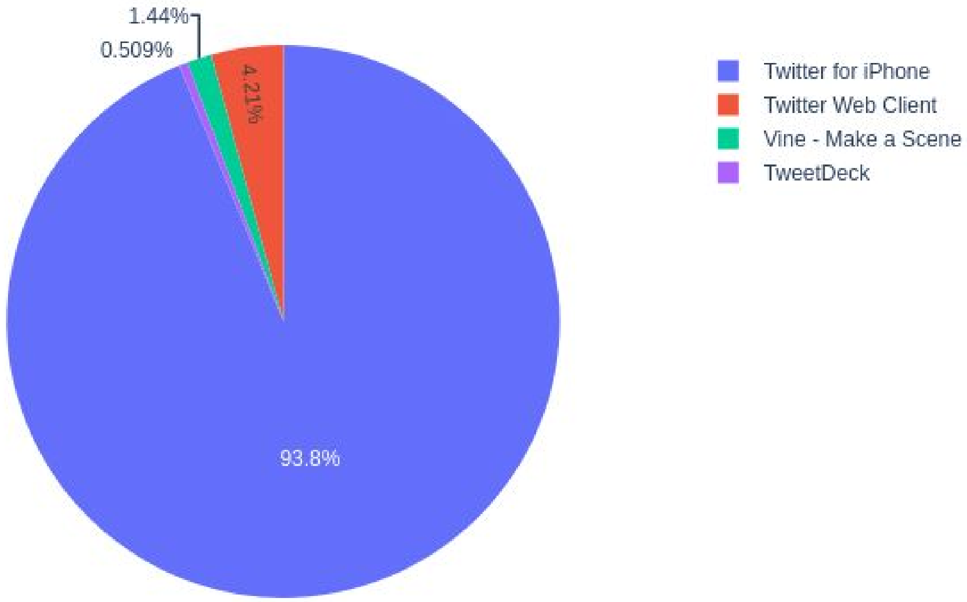
1. Multiple columns containing similar value in “twitter_archive_enhanced” , ie. "dog_stage", was changed to a single column containing either "doggo", "floofer", "pupper" or "puppo"
2. The column "id" was renamed to "tweet_id" to have a common column across all data frames.
3. All the data frames were merged since they all contain data about the same tweet.

Storing, Analyzing and Visualizing the Data

The final data frame was stored for analyzing and visualizing its data. I used the plotly to create the graphs as it is easy to use and produces interactive visualizations.

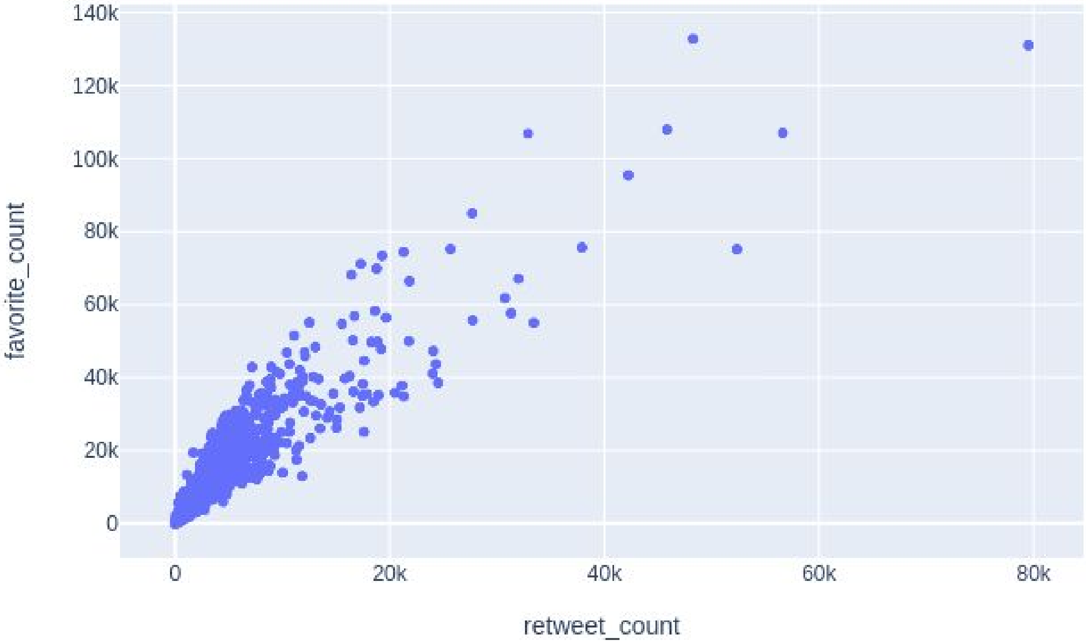
Insights after analyzing the data

1. Almost all the viewers of “WeRateDogs” used an iPhone.

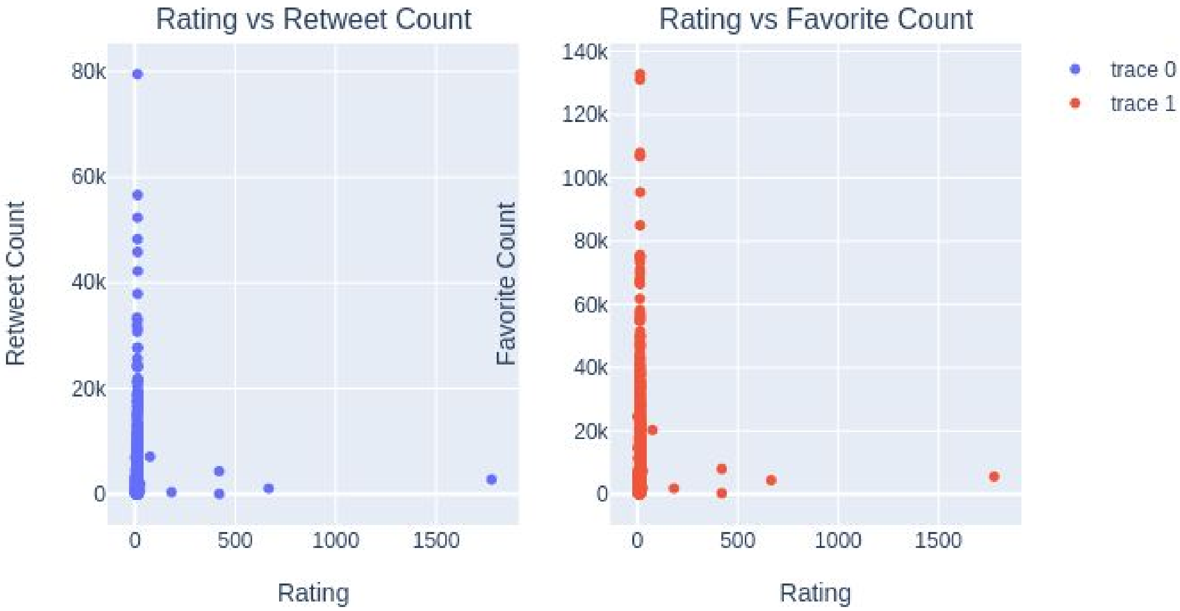


2. The top 5 breeds owned by people are,
    a. Golden Retriever
    b. Labrador Retriever
    c. Pembroke
    d. Chihuahua
    e. Pug

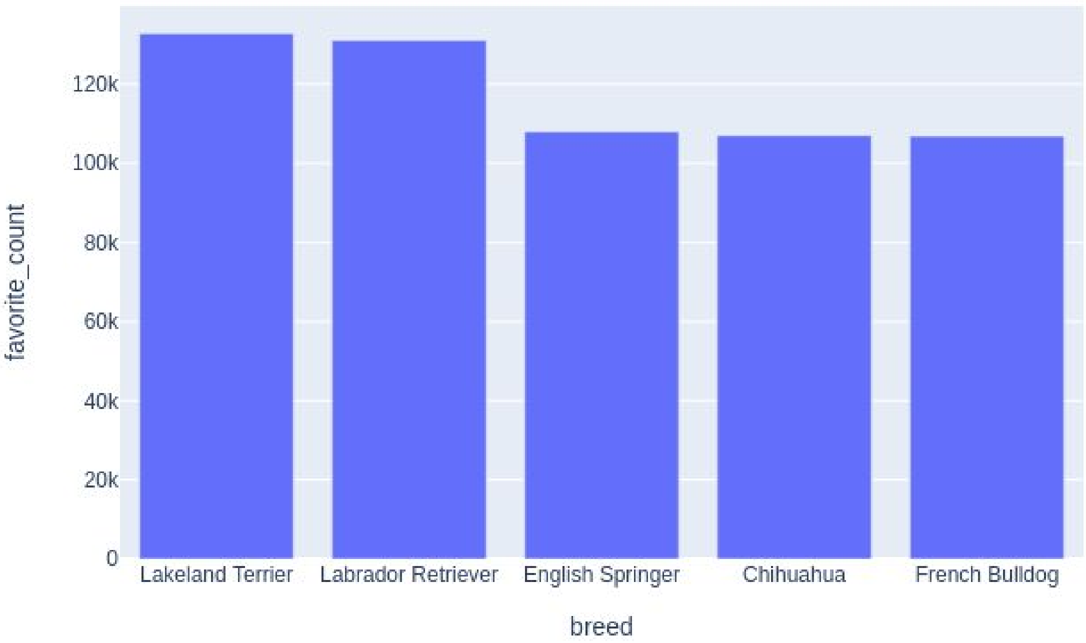
3. A tweet having a high retweet count probably has a high favorite count.



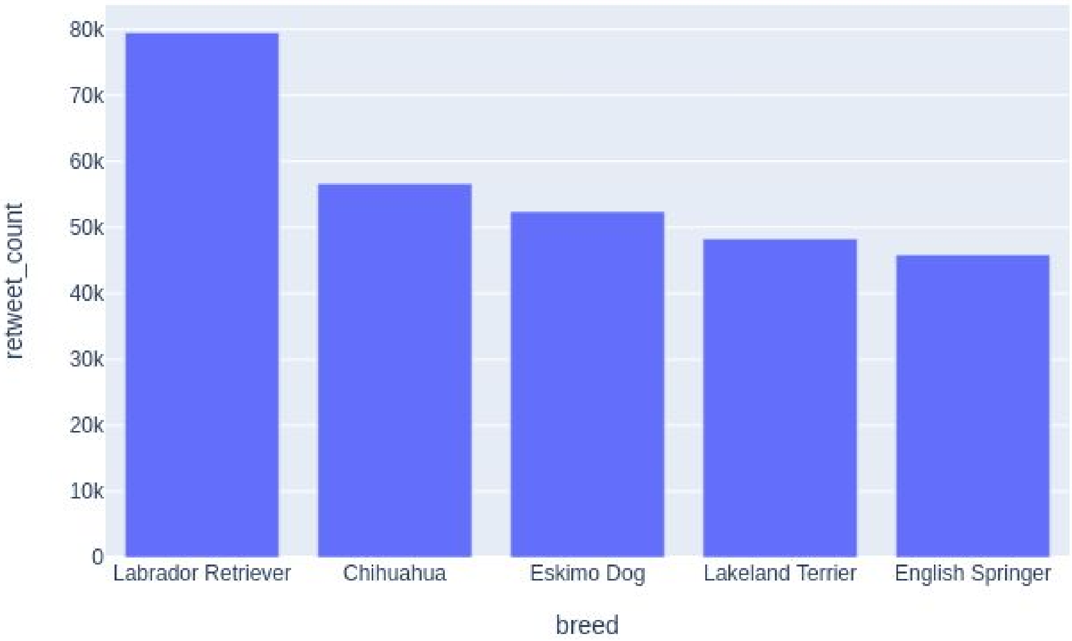
4. Tweets with a high rating doesn't necessary mean a high favorite count or retweet count.



5. Top 5 breed of dogs according to favorite_count



6. Top 5 breed of dogs according to retweet_count


retweet_count
80k
70k
60k
50k
40k
30k
20k
10k
Labrador Retriever
Chihuahua
Eskimo Dog
Lakeland Terrier
English Springer
breed