WeRateDogs

Data Analysis

By Mukund Kulkarni

Introduction

We are going to wrangle the data of a twitter user @dog_rates also known as WeRateDogs. WeRateDogs is a twitter account that rates people's dog with a humorous comment about the dog. The rating is out of 10 but the unique thing about the rating system is that the numerator of these ratings is almost always greater than 10. Our task is to gather, assess, store, analyze and visualize the data.

Gathering the Data

While gathering the data some of the data is already provided, while some data needs to be scraped from various sources like websites, images, databases, texts, etc. In our case we gathered data from 3 sources,

- The “The "twitter-archive-enhanced.csv" is a Twitter archive. 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 was downloaded using the requests module.
- The "tweet-json.txt" was created by querying the Twitter API using the tweepy module.

Assessing and Cleaning the Data

As the data is gathered from various sources, it is not always clean. Hence it needs to be assessed and cleaned. According to wikipedia a clean data has the following properties,

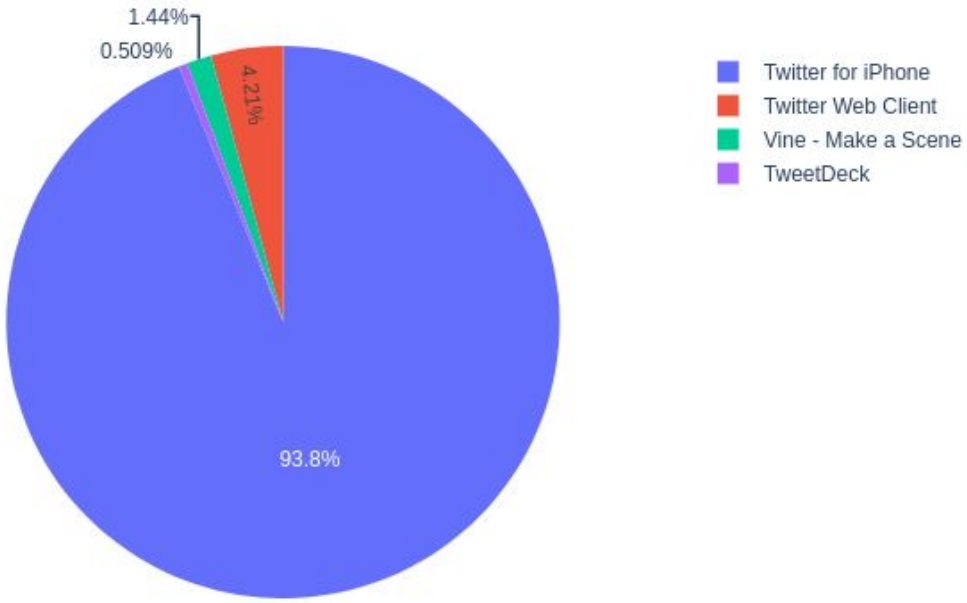
- Validity
- Accuracy
- Completeness
- Consistency
- Uniformity

We assessed the data both manually and programmatically using Pandas. Various issues were found and were solved. The result of cleaning, a dataset that is ready to be analyzed and visualized.

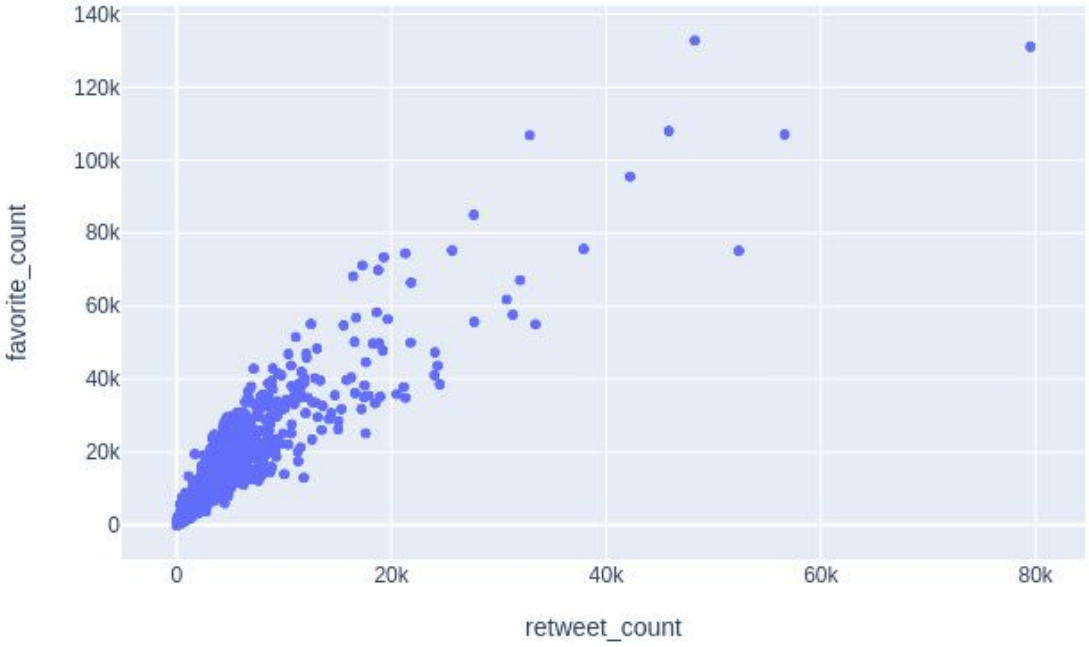
Analyzing and Visualizing the Data

After analyzing and visualizing the data we discover some wonderful insights,

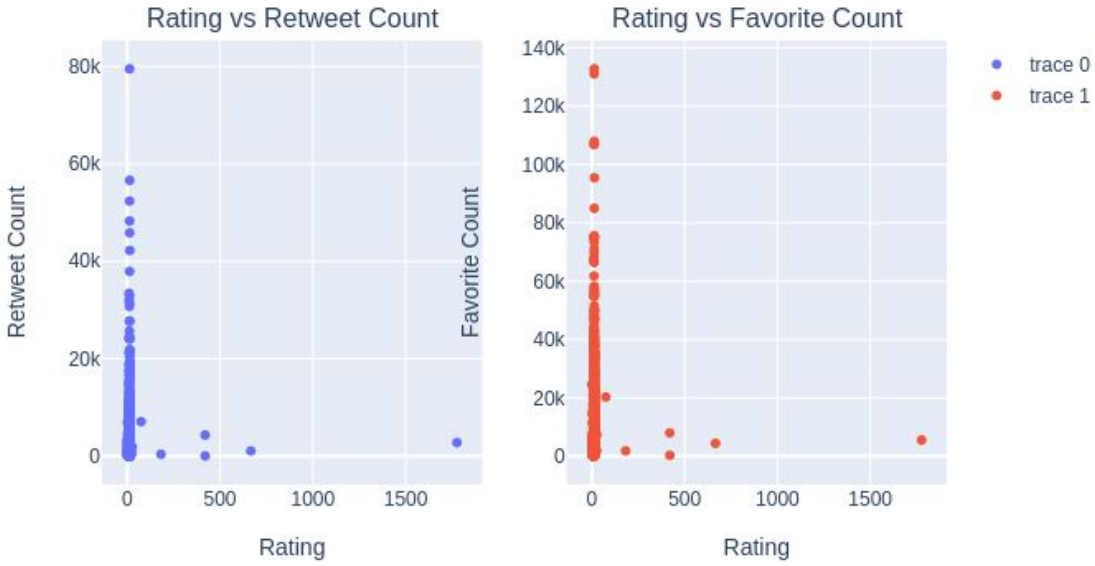
1. Almost all the viewers of “WeRateDogs” used an iPhone.



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



3. Tweets with a high rating doesn't necessarily mean a high favorite count or retweet count.



4. The top 5 breeds owned by people are, Golden Retriever, Labrador Retriever, Pembroke Chihuahua and Pug. But the most favorite top 5 breeds are,

So only, Labrador Retriever and Chihuahua are common in both the lists, hence there are breeds of dogs that people don't own a lot, but like very much.

Conclusion

Data Wrangling is an important part of Data Analysis. It cleans and tidies the data, for inconsistencies and inaccuracies, without which analysis and visualization is not possible.