Insight Report

on

Data Wrangled

from

a

Twitter User

(WeRateDogs)

Introduction

Data Scientists and Data Analysts almost never work with clean data. Understanding this concept, they best equipped themselves with techniques for cleaning data in order to gain insight from the data. This is what inspired this project. It focuses on data collection, cleaning, and analysis through visualization. However, this report focuses more on analysis aspect of it.

I worked on datasets gathered from a Twitter user called **WeRateDogs** for this project. What **WeRateDogs** does on Twitter as a platform is to rate people's dogs. Each tweet from them comes with image of a dog, rating of the dog, and their humorous comment. Each of their content gets people reaction. And in reaction, we mean, likes, retweets, and etc.

From the attached image, it's obvious that Lucy is a very pretty Dog. She was rated 12/10. Such is the sample of WeRateDogs contents on Twitter

Going forward, I intend to analyze WeRateDogs data on Twitter. To analyze this, different datasets had been gathered, assessed, and cleaned (I've already gave a detailed data wrangling process in **wrangle_report** document).

Data Analysis and Visualization

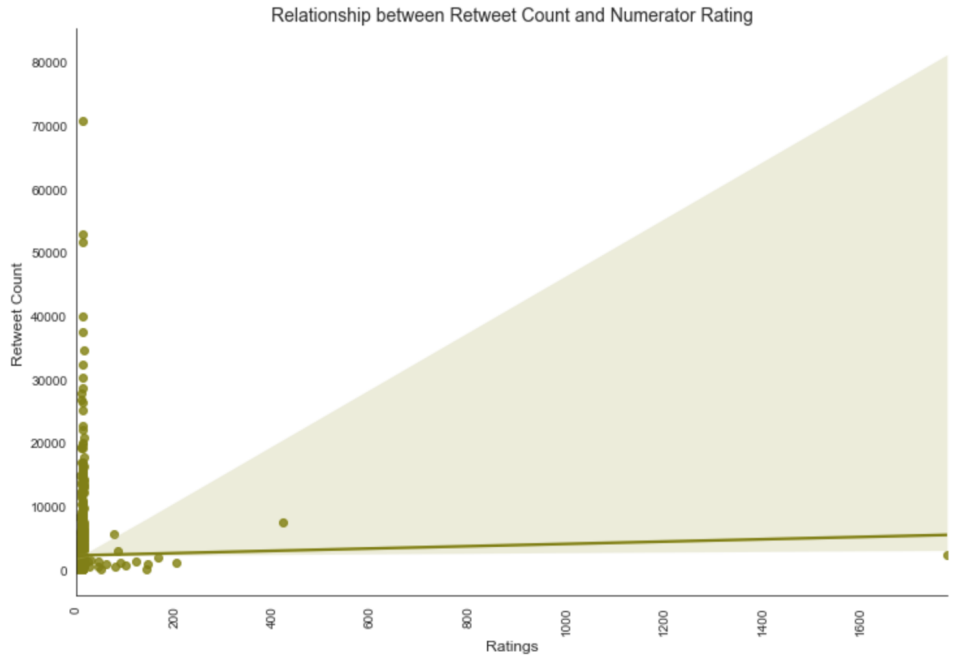
After cleaning all of the datasets, the byproduct of these processes was a single dataset. We want to know what insights we could glean from the dataset. We asked questions to gain these insights. Statistical techniques were used to answer each question. Follow me as I detail my analysis processes, visualizations, and insights.

First Insight

Question: Do higher numerator ratings lead to more retweet_count?

The numerator scores range from 0 to 1776. We want to know if the ratings influence retweets. Could a high rating result in a high retweet count? That is the query!

To solve this, I used a scatter plot to plot the values of the numerator ratings and retweet count variables. This process yielded the following visual:



Visual Interpretation:

- [ ] Most of the data points are more concentrated between **0 - 20**
- [ ] This also shows that we have outliers in the data
- [ ] We can see clearly that most of the retweet that had high retweet hovers arount rdtings between **0 - 20**
- [ ] No rating above **20** had retweet greater than **10,000**

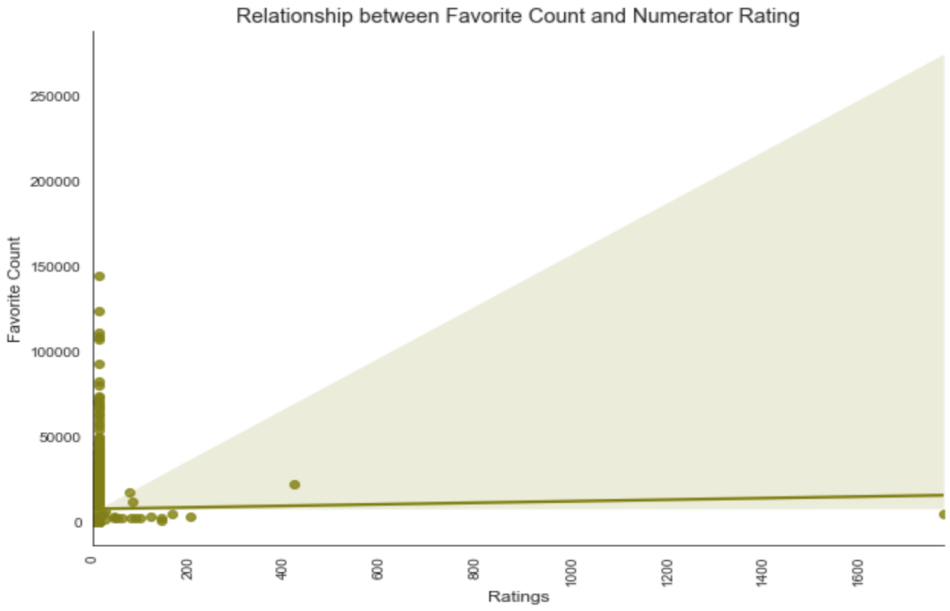
- [ ] There is really no obvious relationship betweet **numerator ratings and retweet_count**

Second Insight

Question: Do higher numerator ratings lead to more favorite_count?

This question arose as a result of wanting to know if ratings have an effect on favorite count. Is there a chance that Twitter users will click on like button if the dog image is rated high? This is was we intend to analyze

To answer this, we follow the same procedure as the first insight visual. This resulted to the below visual:



Visual Interpretation:

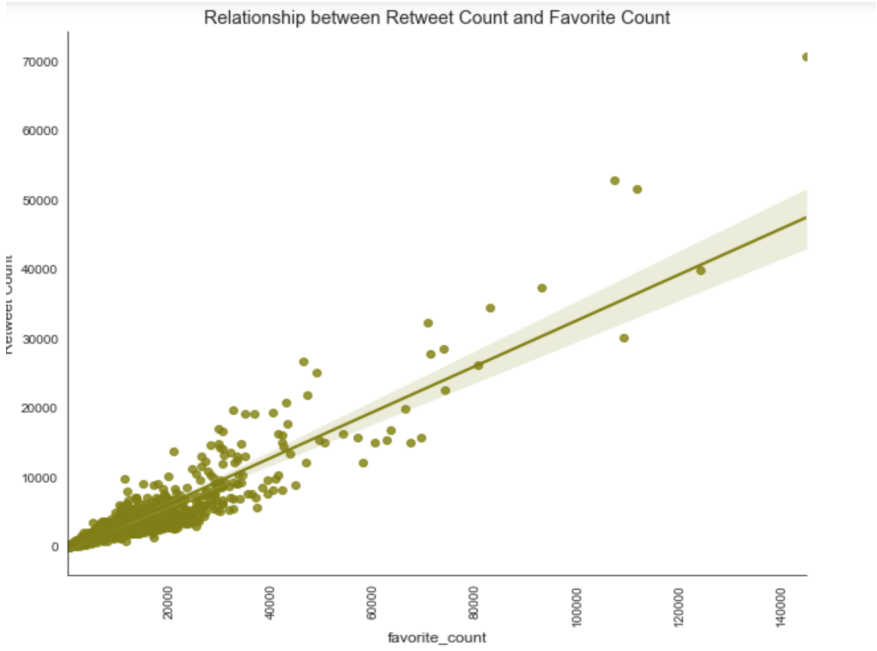
Like retweet count in insight 1;

- [ ] Most of the data points are more concentrated between **0 - 20**
- [ ] This also shows that we have outliers in the data
- [ ] We can see clearly that most of the tweet that had high favorite count were more concentrated where ratings are between **0 - 20**
- [ ] No rating above **20** had retweet greater than **30,000**
- [ ] There is really no obvious relationship between **numerator ratings and favourite_count**

Third Insight

Question: What is the relationship between favorite_count and retweet_count?

This question was prompted by a desire to learn whether people who liked a tweet were likelier to retweet it. Would Dog lovers adore a Dog to the point of having an image of the Dog on their Timeline? This question demanded an answer. Loving would imply that they liked the image of the Dog and also retweeted the Dog's tweet so that it appears on their Timeline. The image below provides an answer to the following question:



Visual Interpretation:

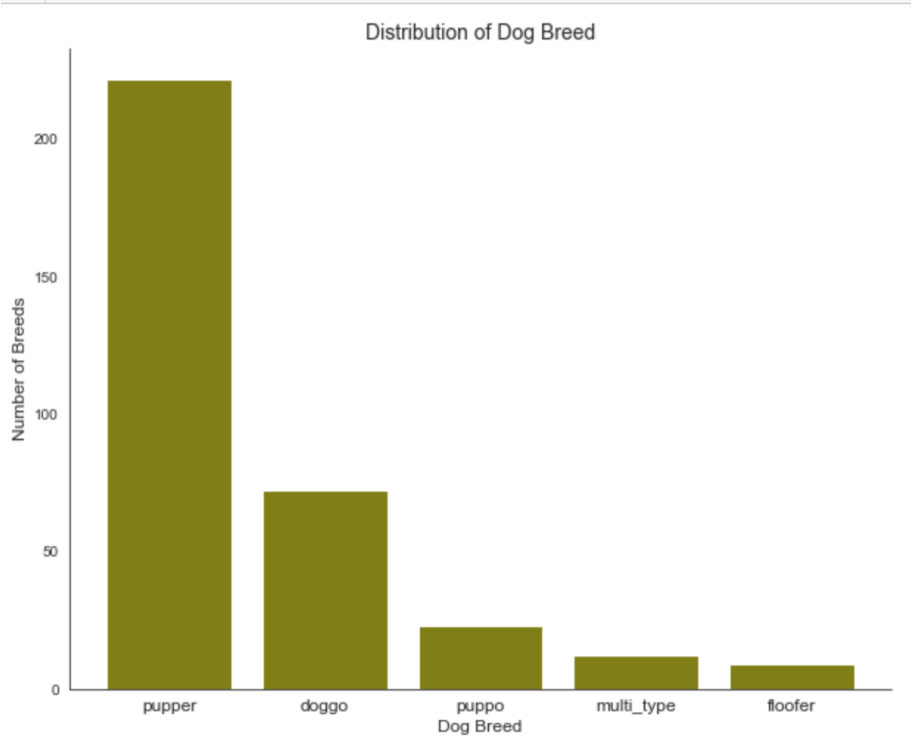
- ☐ There is a positive relationship between favorite_count and retweet_count
- ☐ The visual shows that there is high possibility of WeRateDogs tweet get retweeted by users when the users liked the tweet

Fourth Insight

Question: Which one of the dog breeds is the most famous of all the dog breeds?

Which of the Dog Breeds appeared the most in all of the Dog images with different Dog Breeds? In this context, the most famous means "the one with the most representation." This will assist us in determining the location of our data. Do we have an unfair representation of dog breeds? Will the algorithm be biased towards a section of a dog breed because it has a high

representation if this feature is presented to it? The diagram below will assist us in determining the answers to the following questions:



Visual Interpretation:

☐ The most famous breed of all is **PUPPER** followed by **DOGGO**

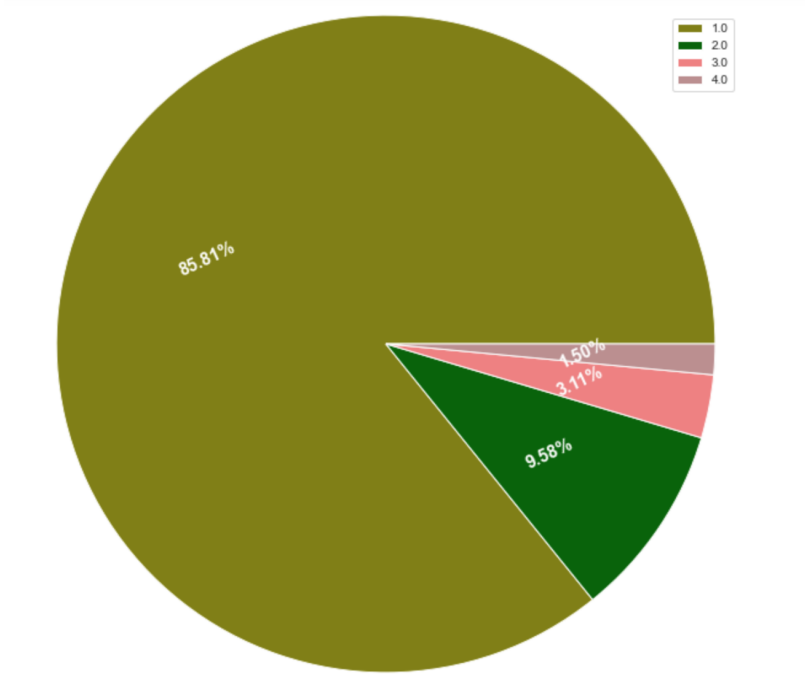
☐ **DOGGO** can a distance second to **PUPPER**

☐ If this feature is supplied to an algorithm, the algorithm will be biased towards **PUPPER** because It had the highest representation.

Fifth Insight

Question: Of all the images supplied to each algorithm, which category of image numbers supplied to the algorithms have the highest percentage?

Img_num variable is the number of the Dog images supplied to the different algorithms for prediction. We marked this variable as categorical variable. We want know which of image numbers had the highest supply to the algorithms. This can be further Analyzed by a Data Scientist to get more insight as regards the performance of the algorithms. We depicted this in a Pie Chart, which will check the percentage of total images of different category supplied to the algorithms. The visual below answers this question:



Visual Interpretation:

- ☐ **85.81%** of the images supplied to the algorithms were single **(1.0)** image
- ☐ So, the image category **1.0** is the most frequent
- ☐ **14.19%** of the total images had more than one image supplied to the algorithms

Conclusion

This report highlighted **Data Analysis** processes in detail. Questions were outlined, and visuals to answer the questions were created. Overall, this report offers an overview of the **WeRateDogs** dataset gathered from **Twitter.**