We Rate Dogs Data Insights Report

by Lei Pei

In this project, we will be investigating the data from We Rate Dogs Twitter account. We Rate Dogs rates people's dogs with a humorous comment about the dog on Twitter, and we gathered their data from various sources, including the basic tweet data (tweet ID, timestamp, text, etc.). This report outlined the Data Wrangling process as well as the data visualizations built on the cleaned data.

1. Data Sources

The raw data was collected from three different sources using various techniques.

- Download WeRateDogs Twitter archive: Contains basic twitter information.
- Tweet image predictions: The breed of dog is present in each tweet according to a neural network. We downloaded programmatically using the Requests library and the URL provided in this project.
- Data scraped from Twitter API: Contains tweet ID, retweet count, favorite count, and follower counts. Query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data.

2. Data Assessment and Cleaning

Real-World data rarely comes clean. Before stepping into the analysis, the raw data was assessed manually and programmatically for both *quality* and *tidiness* issues. The quality of data is assessed against four criteria: completeness, validity, accuracy, and consistency. Data tidiness is a structural issue. According to the regulations by Hadley Wikham, a tidy dataset is defined as 1. Each variable forms a column. 2. Each observation forms a row. 3. Each type of observational unit forms a table.

After the data was cleaned, we obtained 2093 records of useful quality data for further visualization and analysis.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2093 entries, 0 to 2092
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   tweet_id          2093 non-null   int64  
 1   timestamp         2093 non-null   object 
 2   source            2093 non-null   object 
 3   text              2093 non-null   object 
 4   expanded_urls     2090 non-null   object 
 5   rating_numerator  2093 non-null   int64  
 6   name              2093 non-null   object 
 7   stage             2093 non-null   object 
 8   retweet_count     2093 non-null   int64  
 9   favorite_count    2093 non-null   int64  
 10  followers_count   2093 non-null   int64  
 11  jpg_url           1967 non-null   object 
 12  img_num           1967 non-null   float64
 13  breed             1967 non-null   object 
dtypes: float64(1), int64(5), object(8)
memory usage: 229.0+ KB
```

Figure 1 Columns of the Final Data Frame

3. Insights

1) Insight 1: What is the most common stage of the dog?

In this question, we will be looking at the most common stage stored in the stage column. We generated the data column from the original data set to make it tidier for this analysis.

The results show 83.9% of dogs were not assigned with a stage. In the remaining dog that comes with valid data, 72 are doggo, 230 are pupper, 24 are puppo and only 10 are floofer. Thus, we can conclude that **pupper** is the most common among the dogs that come with stage information.

2) Insight 2: What are the top 5 most common breed of dogs?

We investigate the breed of the dog using the neural network information provided in the images_prediction dataset. After cleaning the data, we found the top 5 common breed of the dogs are: 303 are not identified with the breed information, 155 golden_retriever, Labrador_retriever comes third with 106 tweets, and Pembroke and Chow are the fourth and fifth common breed.

```
golden_retriever      155
Labrador_retriever    106
Pembroke               94
Chihuahua              90
pug                    62
toy_poodle             50
chow                   48
Samoyed                42
Pomeranian             41
Name: breed, dtype: int64
```

3) Insight 3: What is breed in top 10 favorite tweets?

In this question, we will be look at breed in the top 10 tweets with the highest favorite. The results shown as in the table.

	breed	favorite_count
301	Lakeland_terrier	132084
391	Chihuahua	119646
107	French_bulldog	115609
58	English_springer	98576
325	standard_poodle	87467
133	malamute	85754
92	golden_retriever	78095
393	cocker_spaniel	75603
65	Chesapeake_Bay_retriever	73849
33	Italian_greyhound	71660

Table 1: breed in top 10 favorite tweets

Lakeland terrier is the winner here!

4. Visualizations

1) Distribution of the Ratings

The rating of the dog always has a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because “they're good dogs Brent.” Here we look at the distribution of the ratings for each dog in the tweet. The distribution is left-skewed, the majority of them are greater than 11, only very few fall below 10. They are sure good dogs!

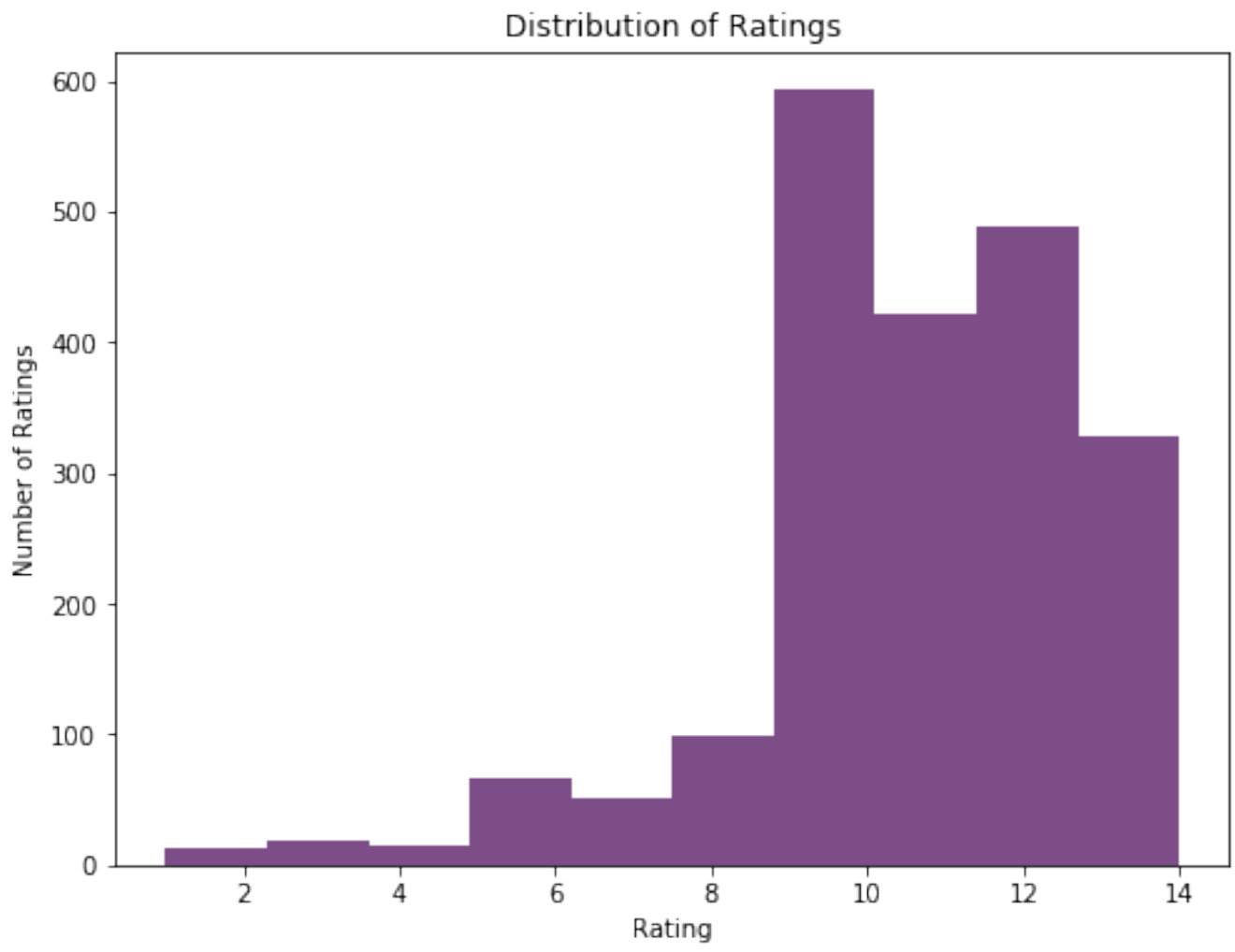

Figure 2: Distribution of the Ratings

2) Relationship between Favorite Count vs. Retweet Count (Log_10 Scale) with color marker of ratings

- Strong Relationship between Favorite Count and Retweet Count

The graph shows a very strong relationship between these two variables. If the tweet gets high favorite counts, it also has high Retweet Count. The scatter plot clearly demonstrate a strong correlation.

- Higher rating indicates higher Favorite/Retweet Count

Another trend that we observed is the higher dog’s rating, the higher Favorite/Retweet Count value will be. This graph demonstrates rating by the gradient of the color - the colder the color, the higher the score. As the line move upward to the top right corner, the line’s color is getting more blue/purple.

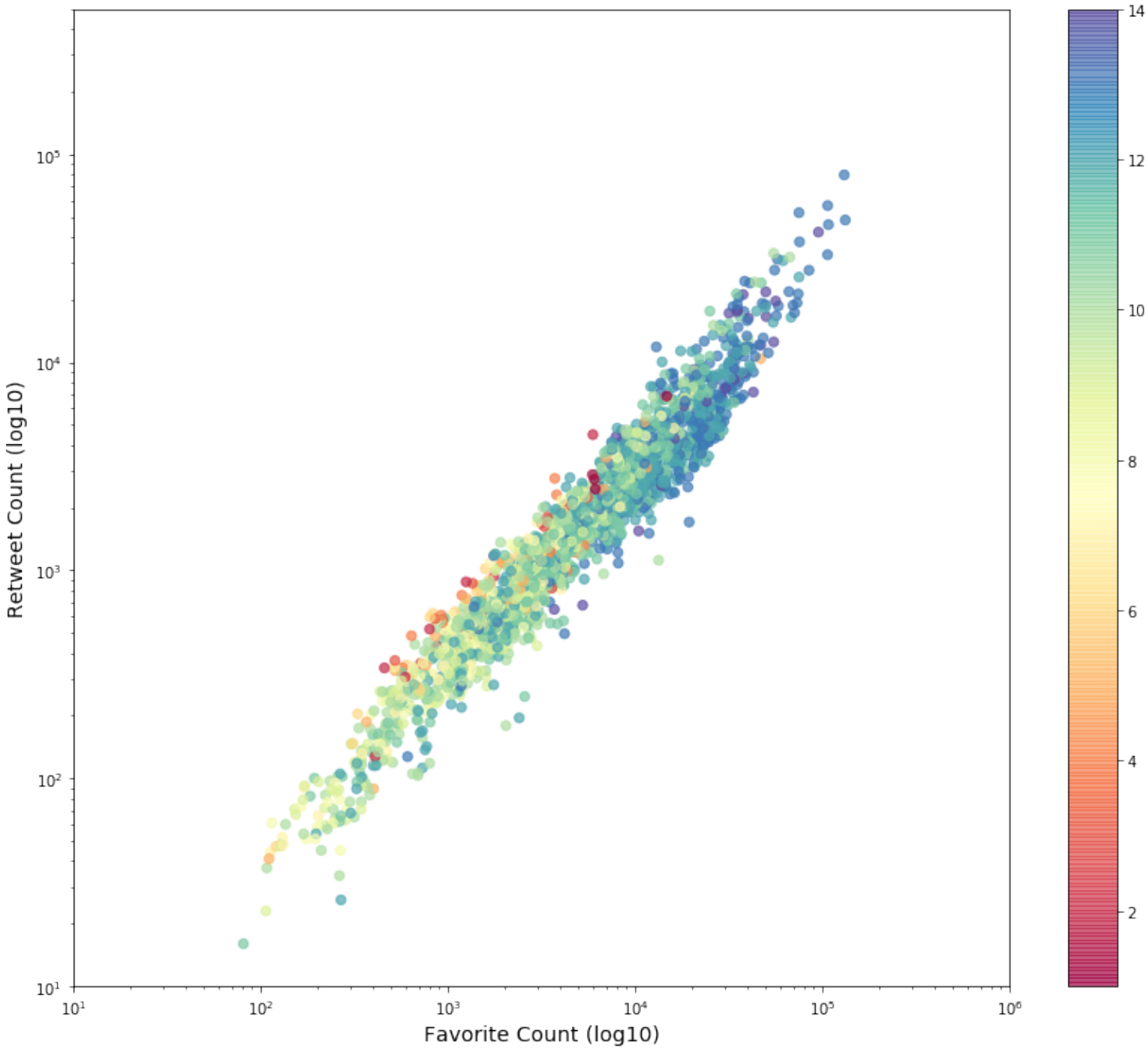

Figure 3: Relationship between Favorite Count vs. Retweet Count with Color Marked Rating