We Rate Dogs

A data wrangling project by Chinomnso Chinedum

We Rate Dogs Data

WeRateDogs is a Twitter account that rates people's dogs with a funny comment about the dog. Their ratings often have a denominator of 10, and a numerator of more than 10. For this project, WeRateDogs provided their twitter archive which contains basic tweet data. The goal of this project was to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. Three datasets were used:

- The twitter archive for WeRateDogs.
- Image predictions dataset that was hosted in Udacity's servers and downloaded programatically,
- Each tweet's retweet count and favorite ("like") count; gathered by querying Twitter's APIs using the tweet IDs in the WeRateDogs twitter archive dataset.

Wrangling the Data

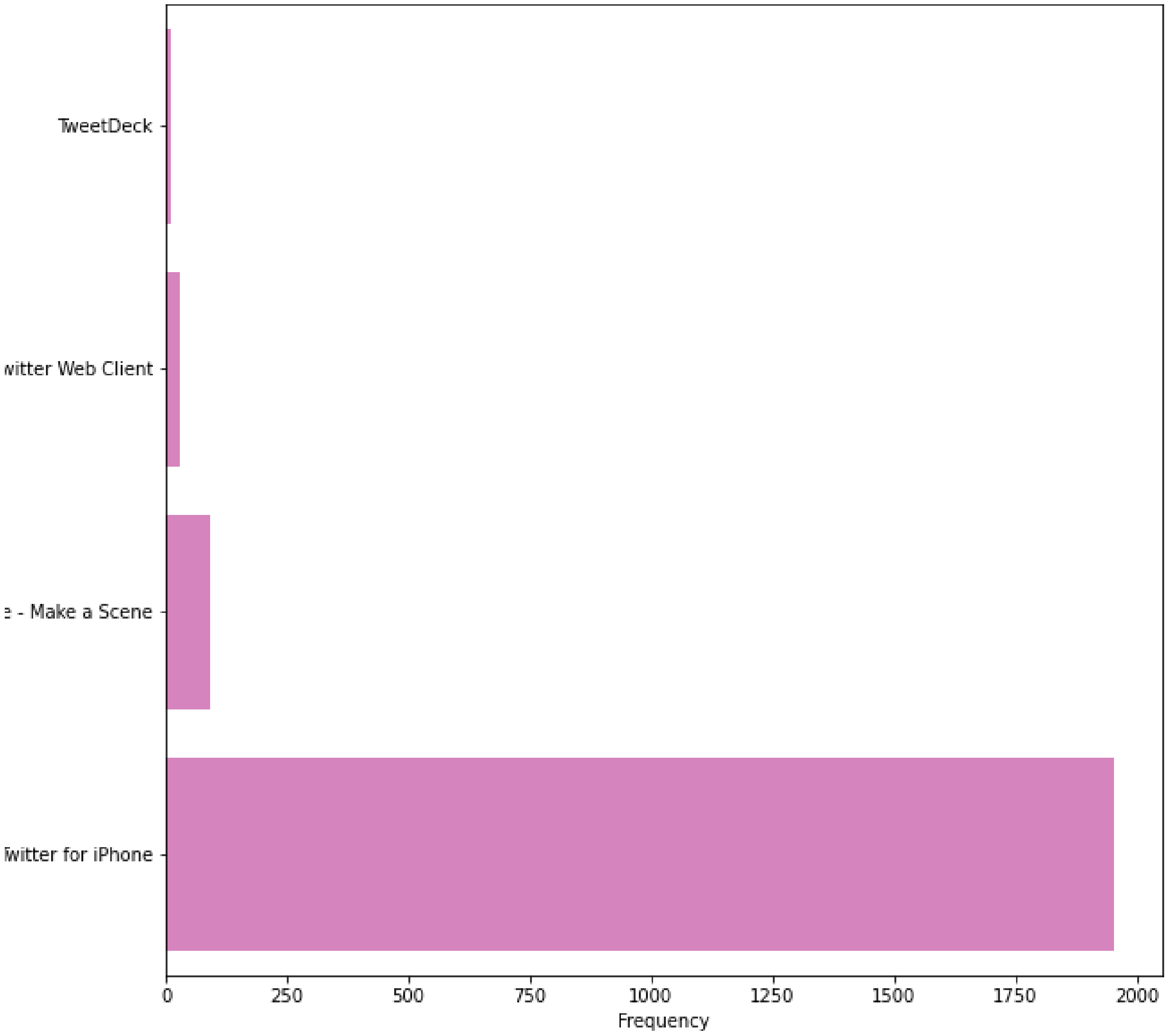
The three datasets were wrangled and 12 quality and tidiness issues were found and sorted. They include:

Quality issues

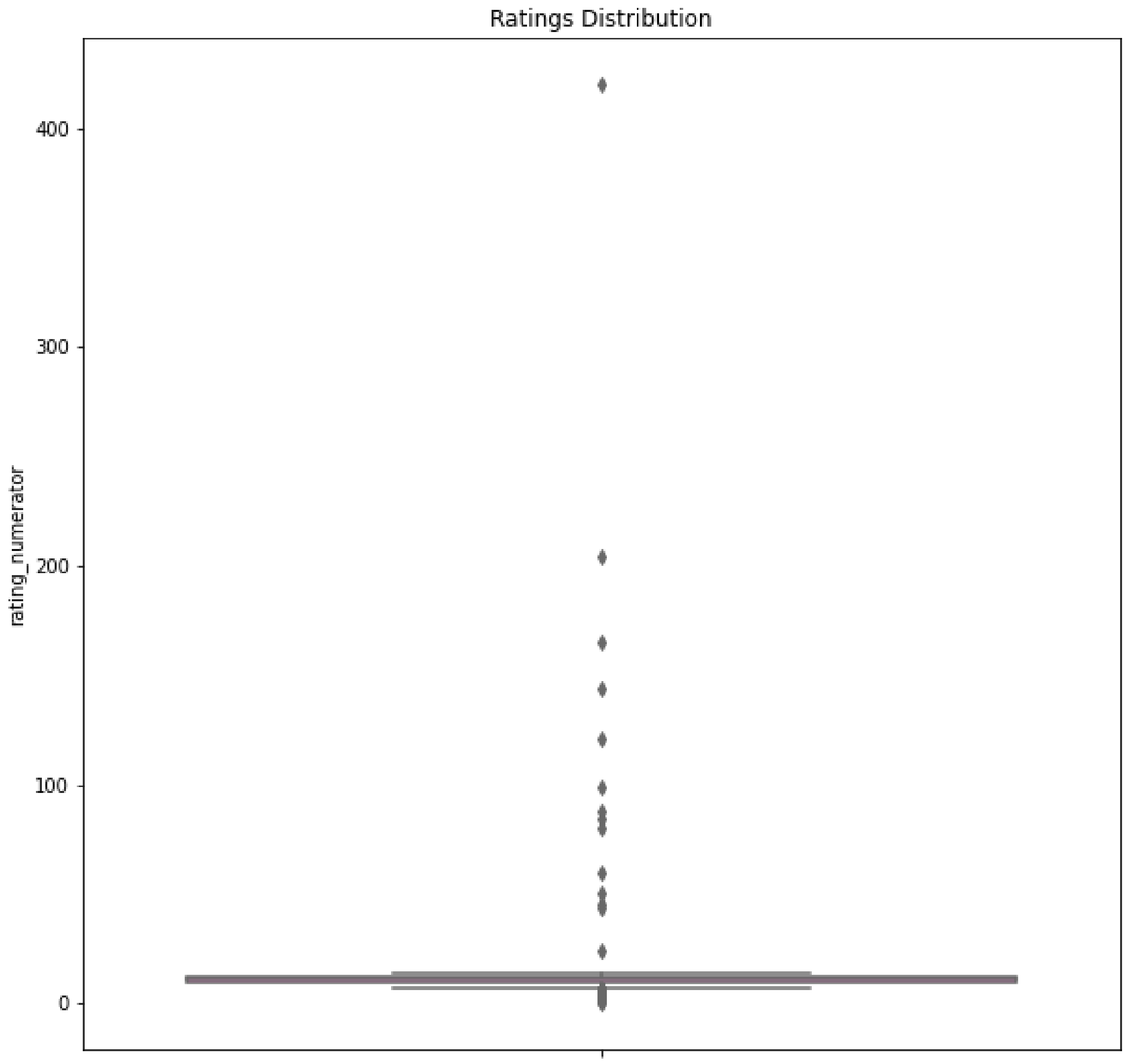
1. Some tweets are replies and retweets, not original tweets.
2. The source_column in `twitter_feed` has some unnecessary HTML code in it.
3. Incorrect rating numerators (decimal issues).
4. Wrong data types for the following columns:
    - a. `timestamp` column in `twitter_feed` has object data type instead of datetime64.
    - b. `tweet_id` in all the datasets
    - c. `source` in `twitter_feed` - should be changed to category data type.
5. Data about dogs and tweets are in the same dataset.
6. Multiple dog stages for some individual records.
7. `create_date` column in fave_retweet_data is not needed (after merge) - see below.
8. Incorrect Names in `dogs_df` .
9. Delete enteries without images #### Tidiness issues
10. Dog stage is split into 4 columns in the `twitter_feed` dataset
11. The predictions, confidence and dog are split into multiple columns in the `image_predictions` dataset.
12. Merge datasets to create a master dataset.

Insights

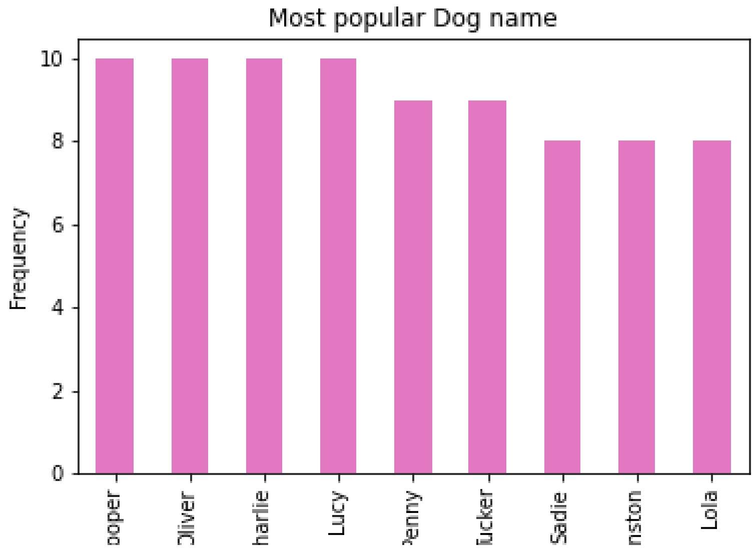
The analysis showed that Twitter for iphone was the most used twitter source for tweets.



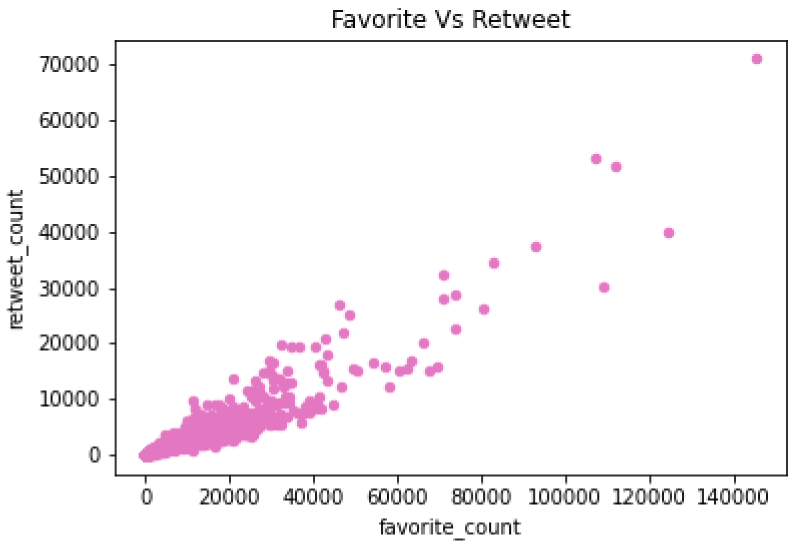
Most of the ratings were between 10 and 14, but there were about 14 outliers whose rating numerators fell between 35/10 and 1776/10



Cooper, Oliver, Charlie, Lucy, Penny, Tucker, Sardie, Winston and Lola are the 10 most popular dog names.



There is a positive relationship between `favorite_count` and `retweet_count`. This relationship is depicted by the correlation value of `0.92`, which suggests a strong positive relationship. This means that when a tweet has a high `retweet_count`, they are likely to have a high `favorite_count` as well.


Favorite Vs Retweet
retweet_count
favorite_count