UDACITY

UDACITY DATA ANALYST NANODEGREE 2020

PROJECT

WRANGLE AND ANALYZE DATA

Qasim Hassan

May 1st, 2020

WeRateDogs – Insights into the @dog_rates Twitter page

A. INTRODUCTION AND BACKGROUND:

Real-world data rarely comes clean. The dataset wrangled for this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Here's an example:

(Source: https://twitter.com/dog_rates)

This project works through the data wrangling process, focusing on the gathering, assessing, and cleaning of data. There are visualizations and observations from the analysis provided as well.

Please feel free to reference my Jupyter Notebook titled "wrangle_act.ipynb" on my Github account: https://github.com/qasim1020/Wrangle-and-Analyze-Data/tree/master to follow the data wrangling process!

B. GATHER

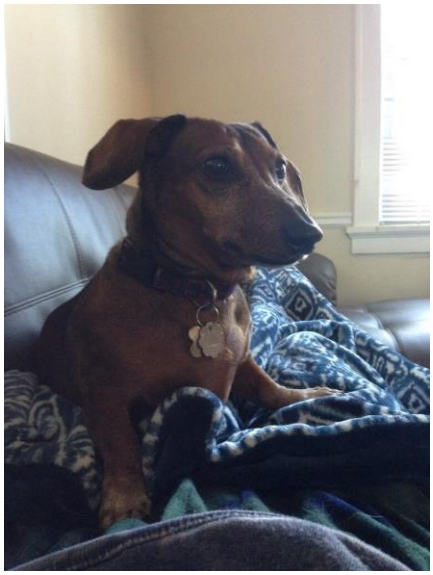
This project gathered data from the following sources:

- The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students.

    - WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project.
    - This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

- The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students.

- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count.



Did you get all the data we need??
(Source: https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg)

C. ASSESS

Assessing data requires data analysts to evaluate a data set on quality and tidiness issues.

The four (4) main data quality dimensions are:

- Completeness: missing data?
- Validity: does the data make sense?
- Accuracy: inaccurate data? (wrong data can still show up as valid)
- Consistency: standardization?

And there are three (3) requirements for tidiness:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

As you look at the data gathered, keep the final product in mind – what kind of data should be presented visually vs. which portions of data only require programmatically analyzing in order to convey insights into the data set?

D. CLEAN

Cleaning data is tedious, and often iterative. Just when an analyst believes they have found all quality and tidiness issues, there are often additional issues that arise. The cleaning process involves three steps:

1. Define: determine exactly what needs to be cleaned, and how

2. Code: programmatically clean the code

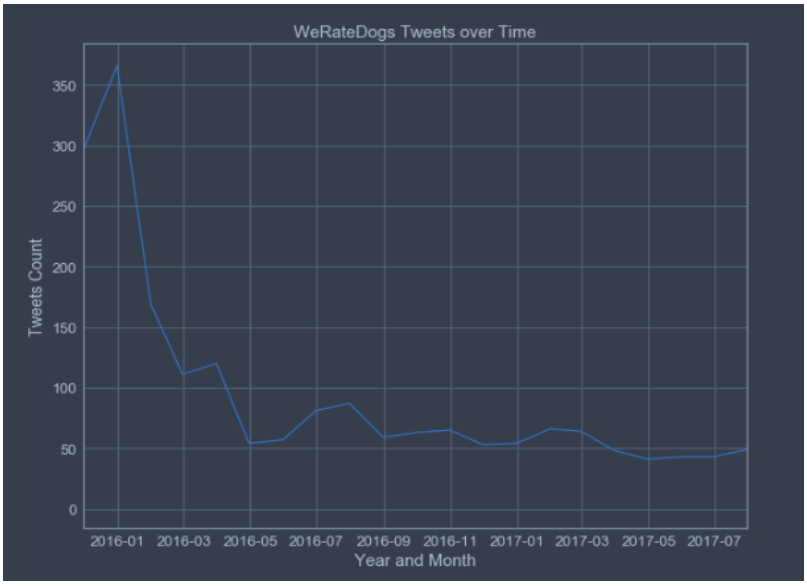
3. Test: evaluate the code to ensure the data set was cleaned properly

E. ANALYSIS AND VISUALIZATION

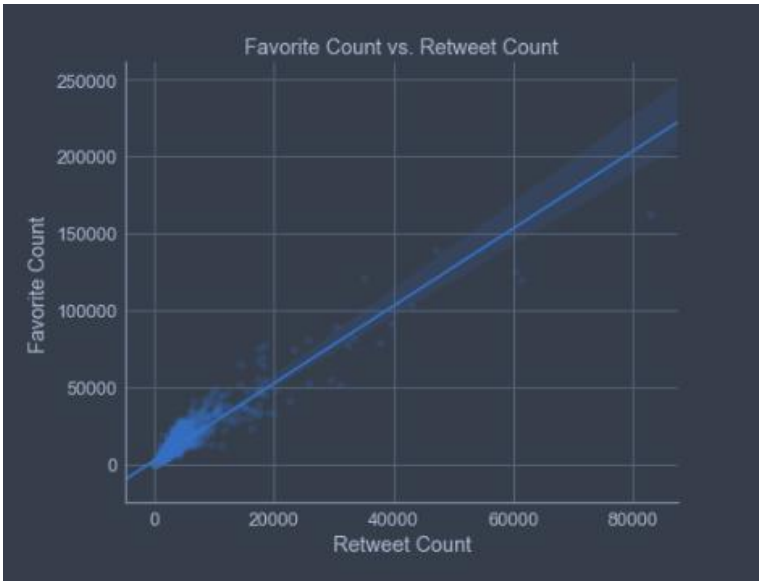
I chose to analyze and present on four different pieces of the WeRateDogs data set.

1.Tweets over Time

Over the time period of the tweets collected for this dataset, tweets decreased sharply starting in early 2016. While the tweets continue to decline over time, there are spikes in activity during the early spring of 2016, mid-summer of 2016, but continues to generally decrease from there. This data set did not provide corresponding data that could provide a reason for the sharp decrease in 2016, and steady decrease from there on out. The owner of the WeRateDogs Twitter account should be aware of this trend, and consider ways to increase user traffic on the page.

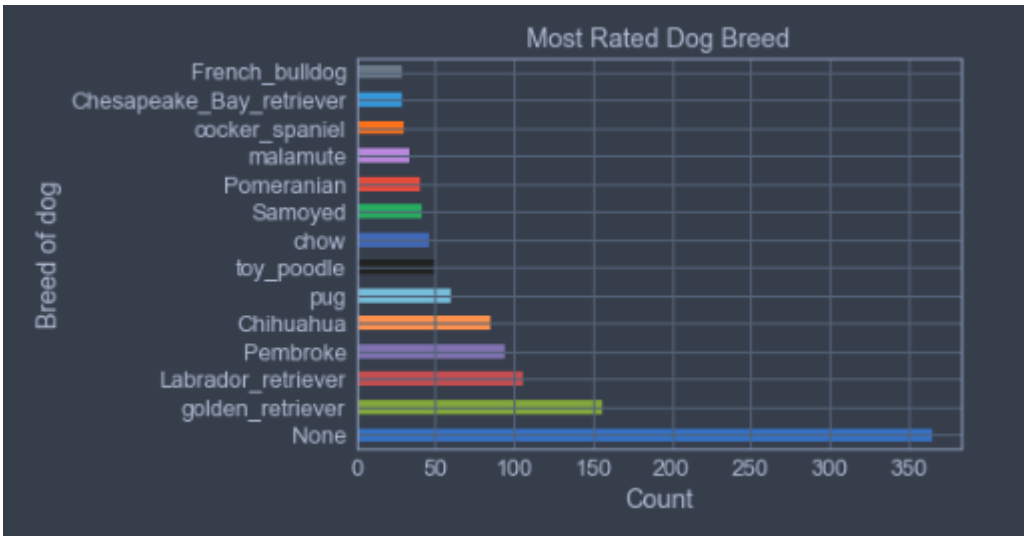


2. Favorite vs. Retweet Counts



There is a positive correlation between favorite (‘like”) counts, and how much a post was retweeted. This correlation is important for the owner of the WeRateDogs Twitter account to understand when determining methods to increase user traffic on the page. A data analyst team could recommend previous posts with either a high retweet count, and/or a high favorite count so the page owner could model future posts off historically popular posts.

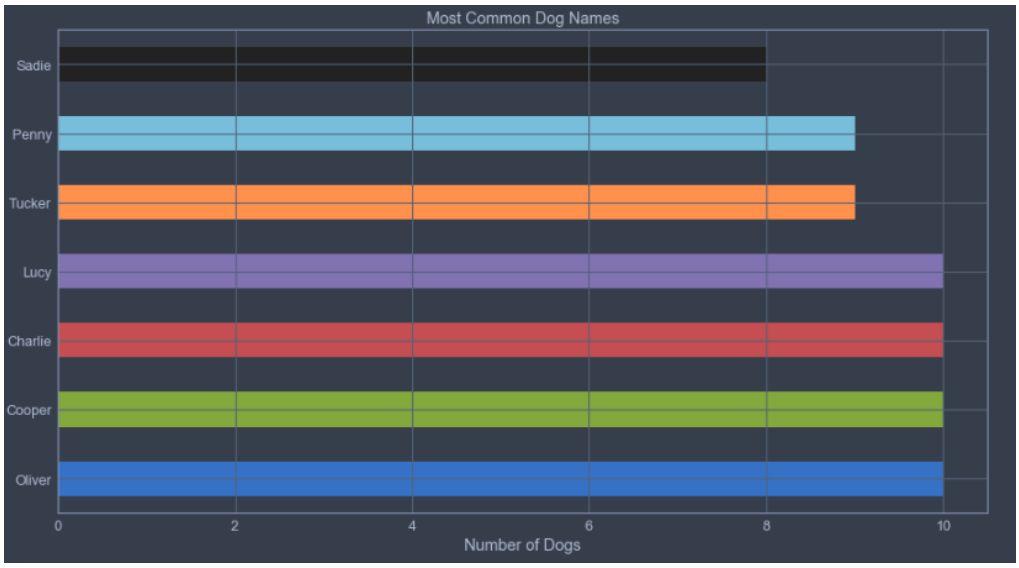
3. Dog Breed Popularity



The most popular dog breed is a golden retriever (ignoring the None label), with a Labrador Retriever coming in as the second most popular breed. Pembroke isn't far behind. The page owner could use this information to create targeted marketing efforts for certain breeds that aren't as popular to increase their popularity, but also utilize the breeds that are proven to be popular to drive user traffic to the page.

4. Dog Name Commonality

Names are important, especially for dogs! The four most popular dog names are Oliver, Cooper, Charlie, and Lucy.



F. CONCLUSION

This write-up offers a straightforward look at the data wrangling process. There is so much more that can be done with this data set, but I encourage aspiring data analysts to dive deep into this data set and see what else you can find!