Image via Boston Magazine

Wrangle and Analyze Data 🤠

Report on the insights gathered from the data

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that you will be wrangling (and analyzing and visualizing) 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. These ratings almost always have 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." WeRateDogs has over 4 million followers and has received international media coverage.

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. More on this soon.

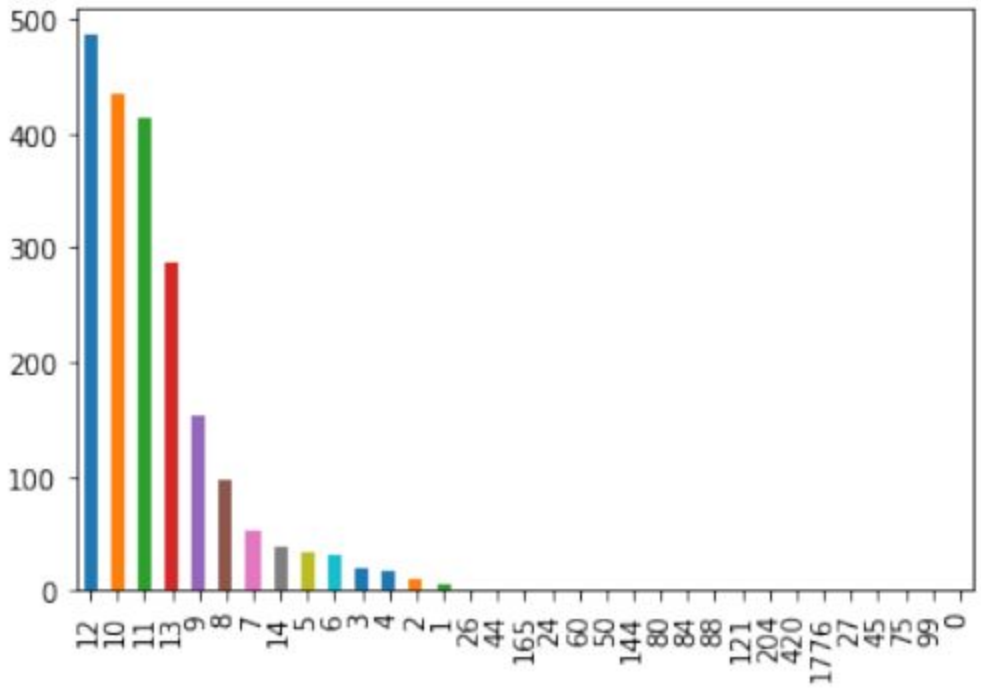
What did we learn from the data?

1. The Rating System

The WeRateDogs twitter account follows a weird and rather unscientific method for rating dogs, but clearly this is because of the fact that *all dogs are good dogs* :)

This makes it really difficult to get more insights about the actual dogs from the ratings alone. Also, it was observed that in most cases the ratings are allotted in denominations of 10, with the numerator being larger than the denominator. But, in some instances the numerator was upwards of 1000, this tips the scales in undesirable directions when we try to use this data.

A better way to make sense of the rating system would be to try to find the correlation between the ratings and the number of retweets and favorite a post gets if there is one.



2. The Dog Breeds

Plotting the dog breeds on the histogram it revealed that a majority of the dogs were Golden Retrievers, closely followed by Labrador Retrievers.

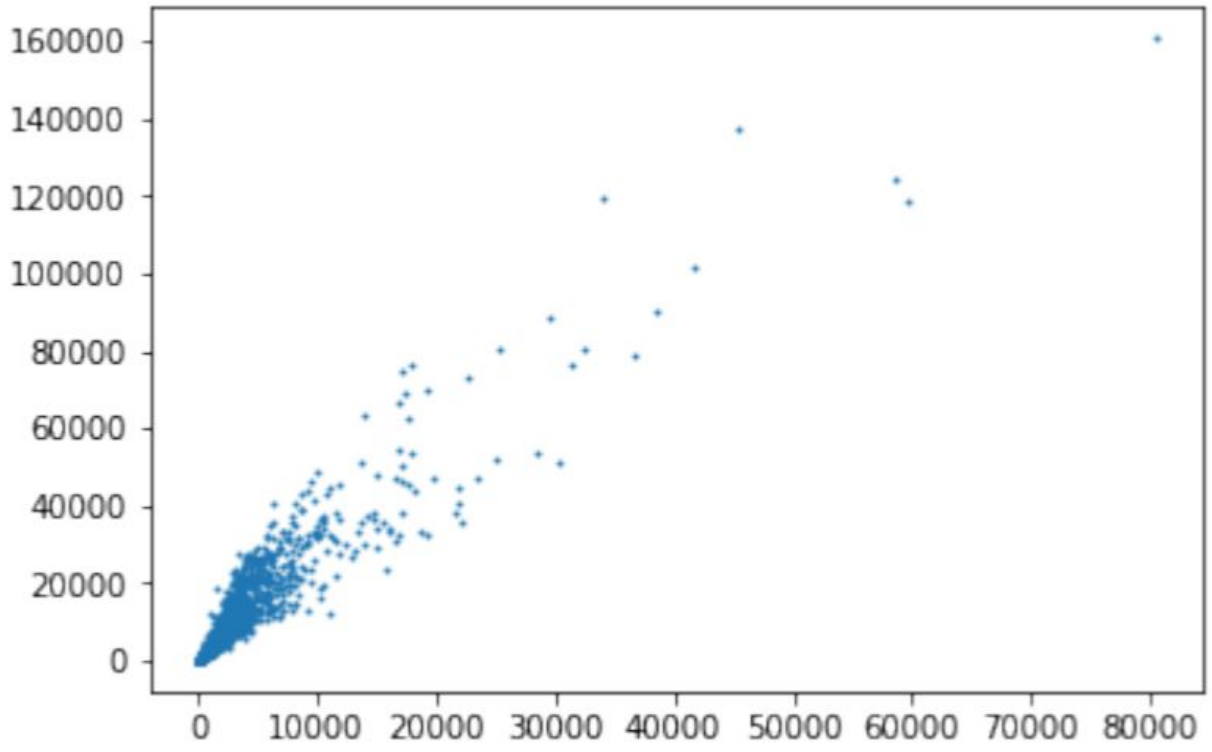
The reason for this is simple - Retrievers are super-friendly, cute and cuddly. 🐱

Golden Retriever	137
Labrador Retriever	94

Pembroke	88
Chihuahua	78
Pug	54
Chow	41
Samoyed	40

3. Retweets V/s Favourites

We plotted the number of retweets and the number of favorites each post got, and the results were exactly what we expected. It was revealed that if a post has more retweets it is more likely to have more people marking it as favorites as well.



4. Has the account grown over time?

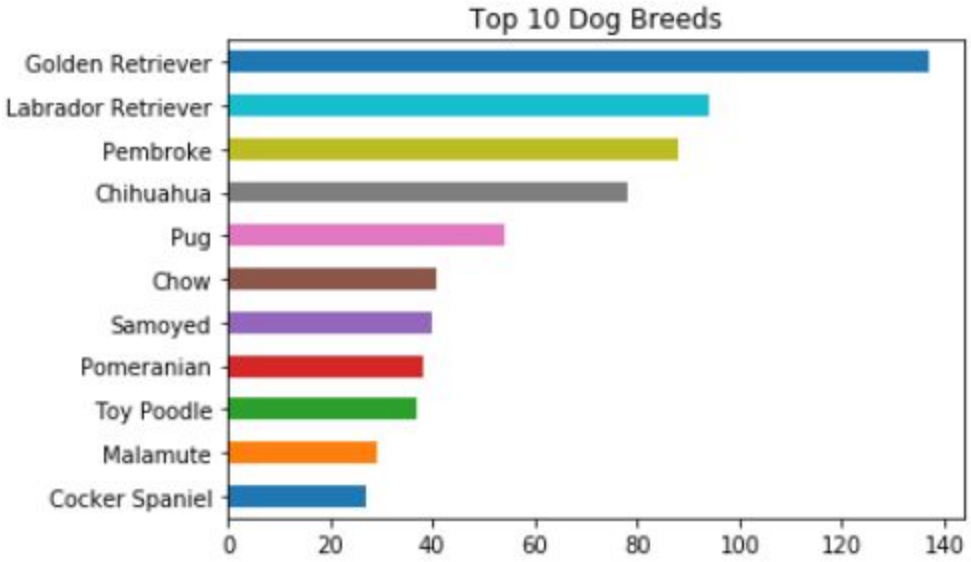
From the graph, we can clearly see that the twitter account in quest has grown consistently over the years.

This growth might be due to the fact that the target audience of the twitter account is actually quite large, as almost everyone loves dogs. The account also has been posting tweets frequently since

2015 this also is one of the reasons for the growth.

5. The Names

On counting the number of times every name appeared in the data, we discovered that Lucy and Charlie were the most common names, closely followed by Oliver which appeared 10 times.



Conclusion

Thus we conclude from all of the data that - people love Golden Retrievers! And all dogs in general. We also understand that consistent posting on twitter will lead to bigger followings overtime