Data Wrangling Project [Data-Analyst-Nanodegree]: WeRateDogs®

By: Lina AlKhodair

Introduction:

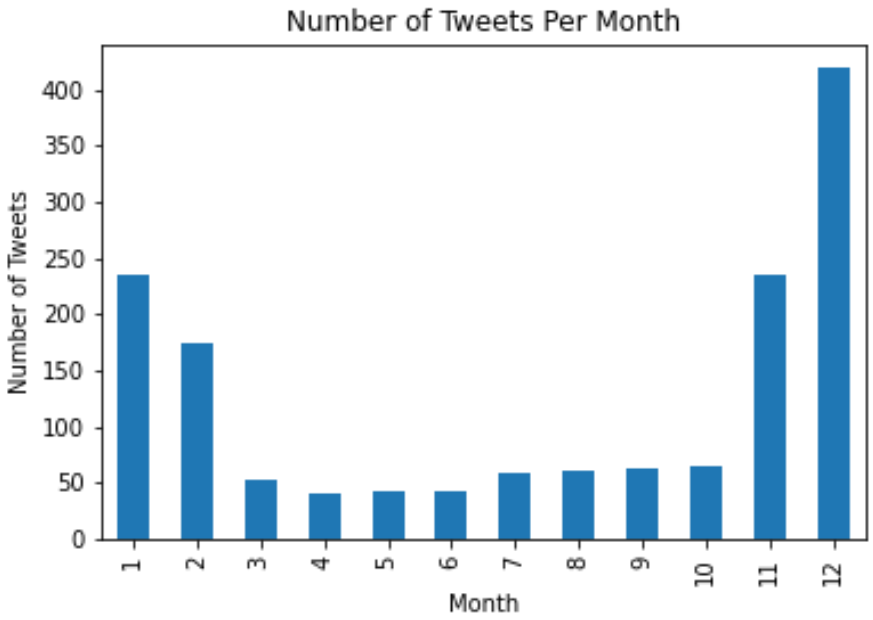
Have you ever thought of creating a Twitter account that posts humorous content about your dog? No, okay same. However, @WeRateDogs have come up with this idea and they did a great job of rating dogs and making us laugh. The twitter account allows you to send pictures of your dog and later on they will be rated and posted with funny comments. So, my motivation struck from there.

The goal is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations. In order to do so, multiple datasets have been used. Firstly, enhanced Twitter archive, The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets. Additional data via the Twitter API. Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API. But, because you have the WeRateDogs Twitter archive and specifically the tweet IDs within it, can gather this data for all 5000+. So I have queried Twitter API to gather this valuable data. One more dataset, an image predictions file that contains every image in the WeRateDogs Twitter archive that has been run through a neural network that can classify breeds of dogs. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

So that's all fun and good. But the goal is to gain useful insights and data visualizations. So, after assessing and cleaning the data some questions have been raised that I have decided to take on and analyze them. I will be posing each question along with its visualization and the insight.

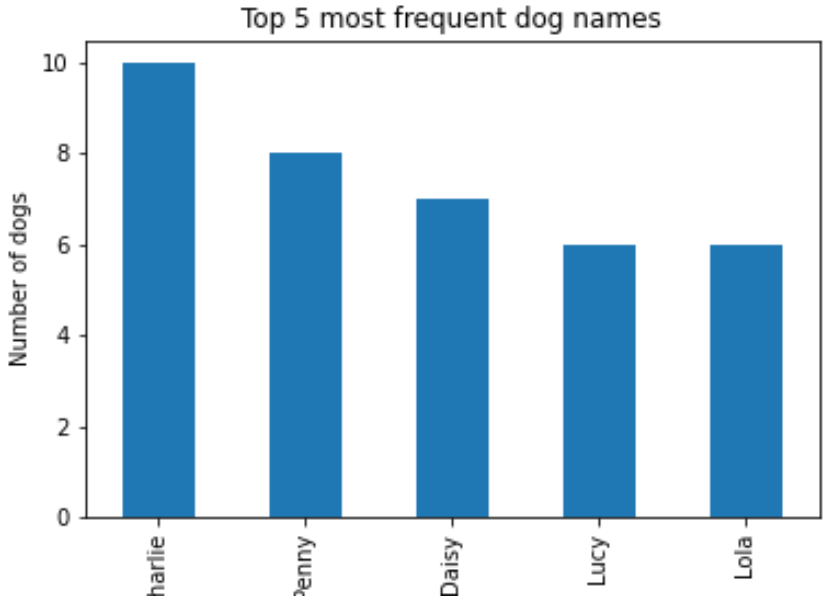
Analysis and Visualizations:

1. **Which month has the highest number of Tweets ?**



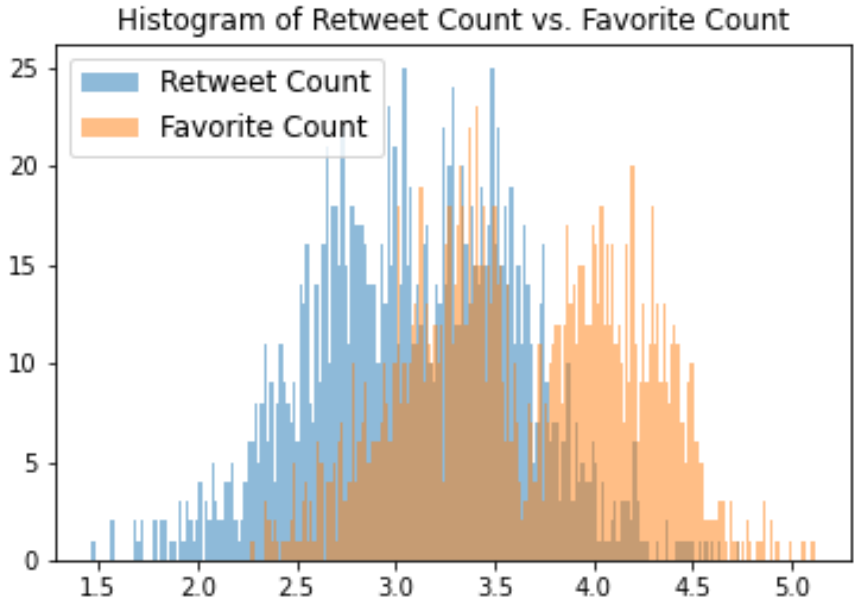
- As shown in the figure, we can understand that December has the greatest number of tweets, following that would be November.

2. What are the top 5 most frequent dog names ?



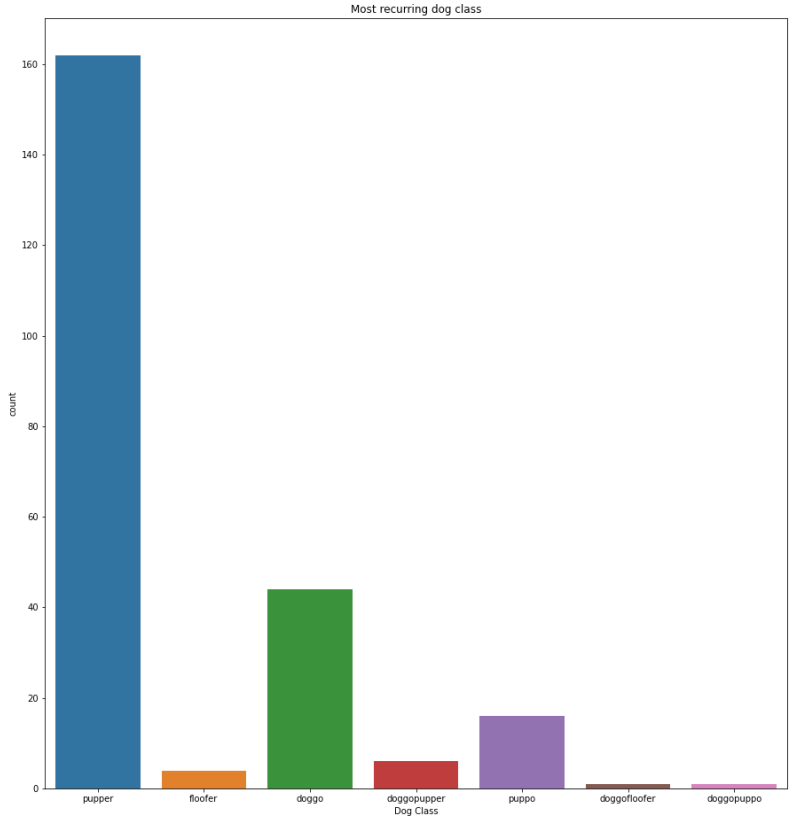
- As shown in the figure, we can observe that The name Charlie is the most frequent name for dogs, along with Penny coming very close.

3. What is Retweet count vs. Favorite count distribution ?



- As shown above, we can see favorite count increasing. So, we can observe that users favor tweets more than retweeting.

4. What is the most recurring dog class ?



- As shown in the figure above, the class ‘pupper’ is the most recurring dog class with high variance across other classes.