Wrangle and Analyze « WeRateDogs » Data

Act Report

Kévin Péricart

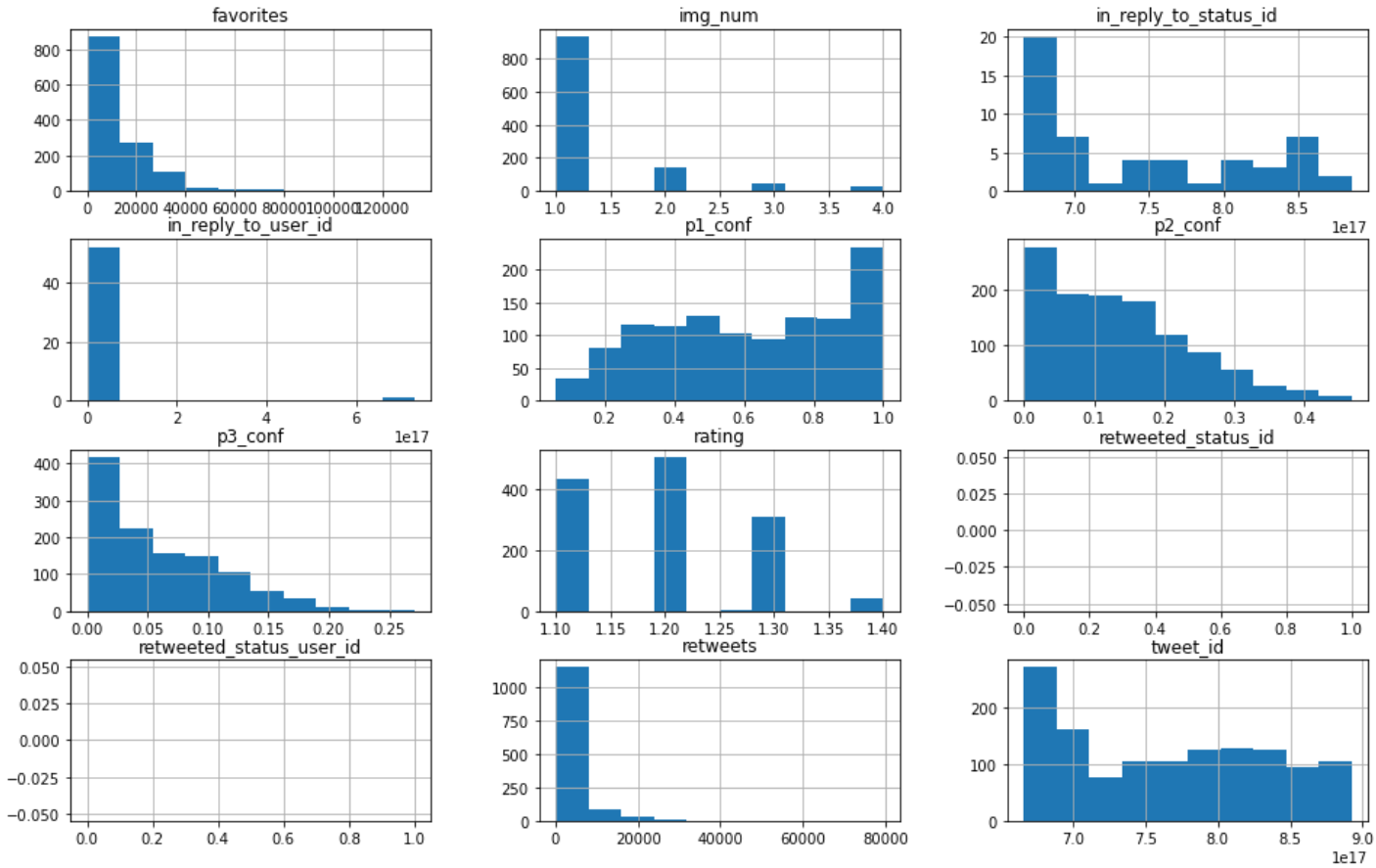
Introduction

The dataset that we will bre wrangling (and analyzind 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 as over 4 million followers and has received international media coverage.

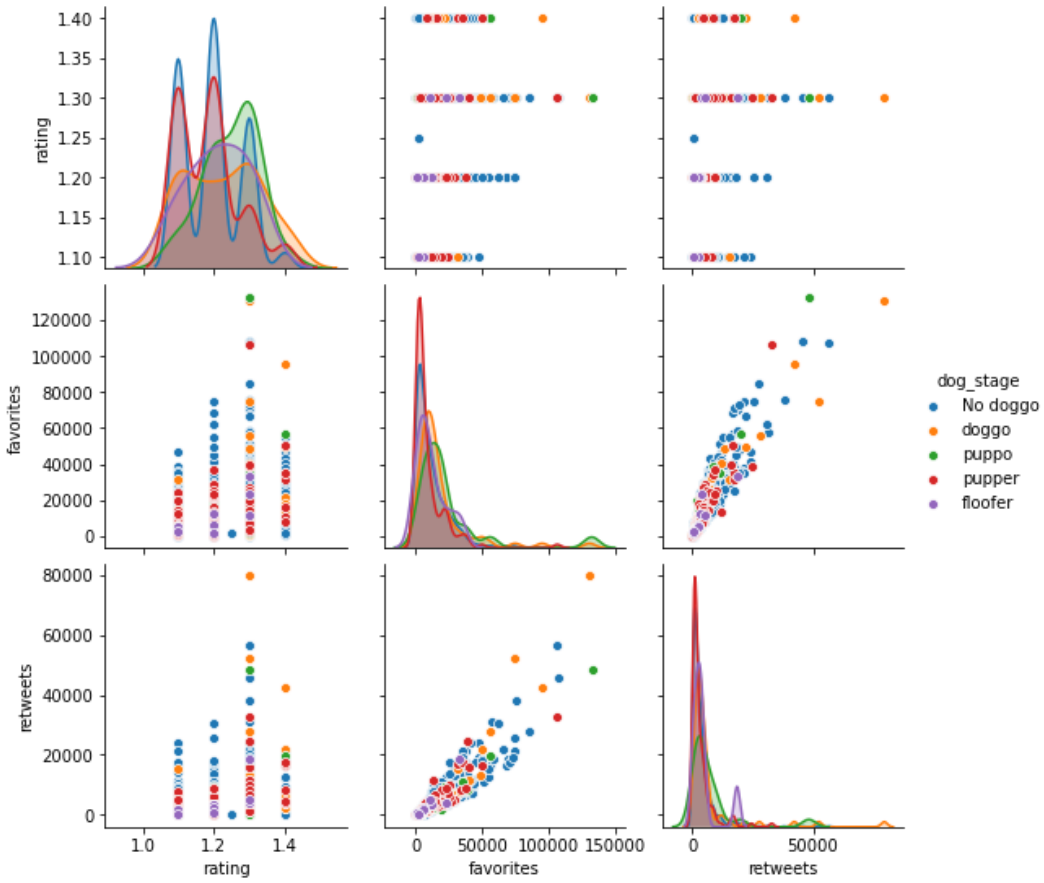
WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for us to us in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Our 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.

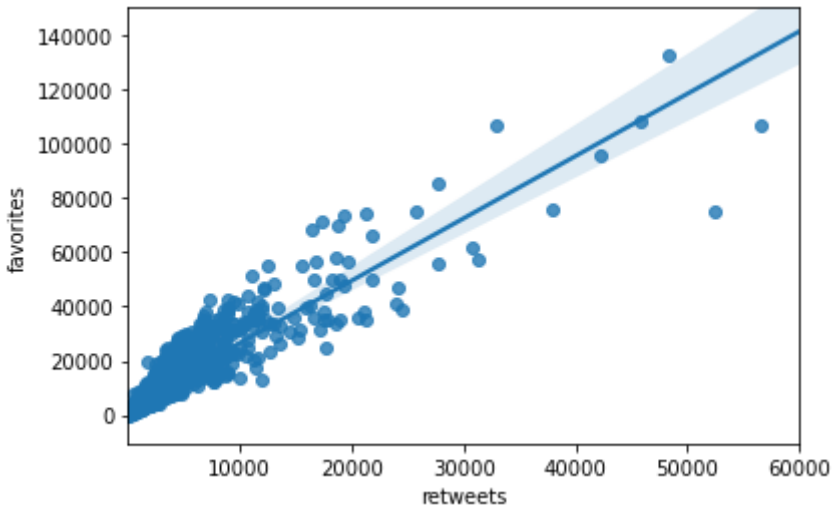
Histograms of the full merge dataset



Plots of ratings, favorites and retweets opposed to « dog_stage »

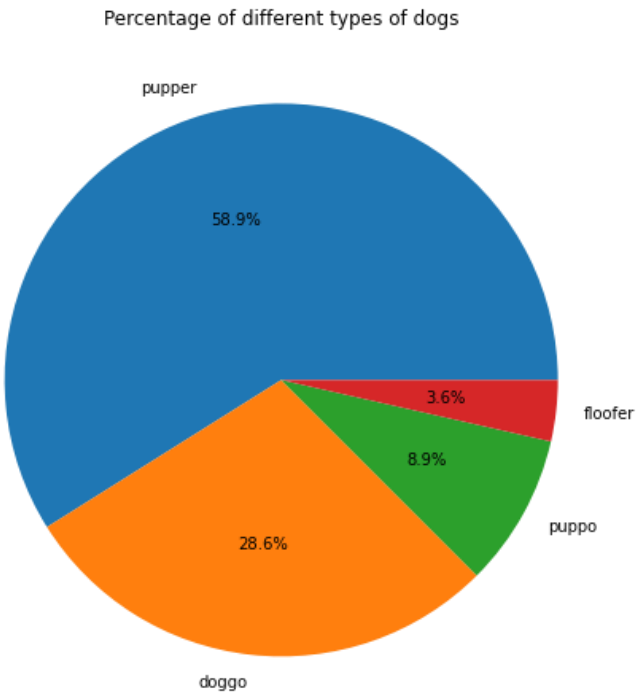


Linear Regression between retweets and favorites



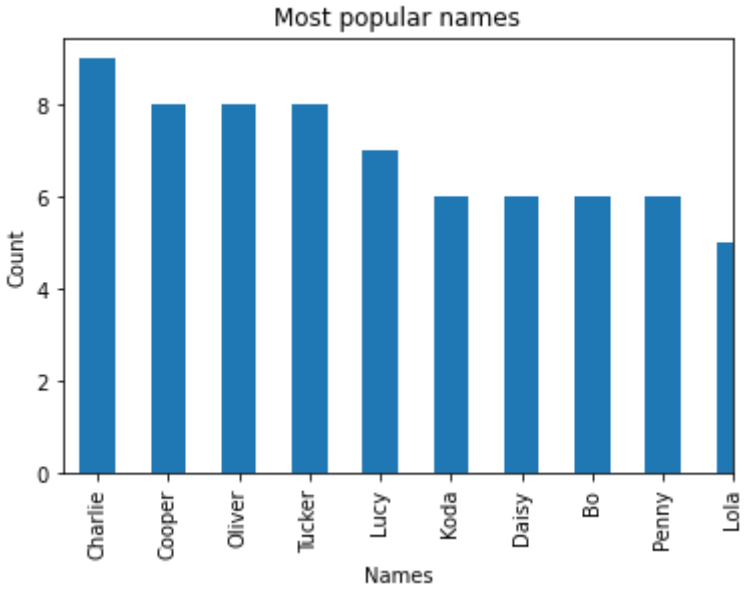
We observed a strong correlation between retweets count and favorites count of 0.9047

Distribution of doggo categories



We observed that pupper category is way more present than others (58.9%). The second category is doggo with 28.6%.

Most popular names



The 10 most popular names in the dataset are :

1. Charlie
2. Cooper
3. Oliver
4. Tucker
5. Lucy
6. Koda
7. Daisy
8. Bo
9. Penny
10. Lola