PROJECT: WRANGLING AND ANALYZING TWITTER DATA

AUTHOR: ANDREA CLAUDIA VILLANCA ROSALES

JUNE 2020

Boston Magazine

Introduction

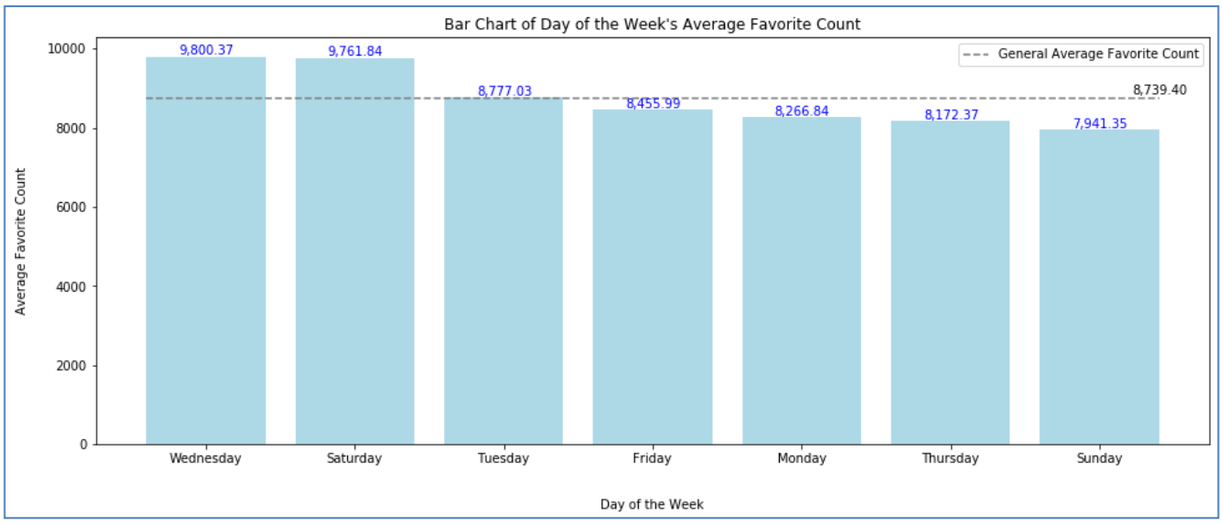
Real-world data rarely comes clean... that's why data wrangling becomes an important part of the data analysis process.

The dataset that we wrangled, analyzed and visualized in 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. 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 8 million followers and has received international media coverage.

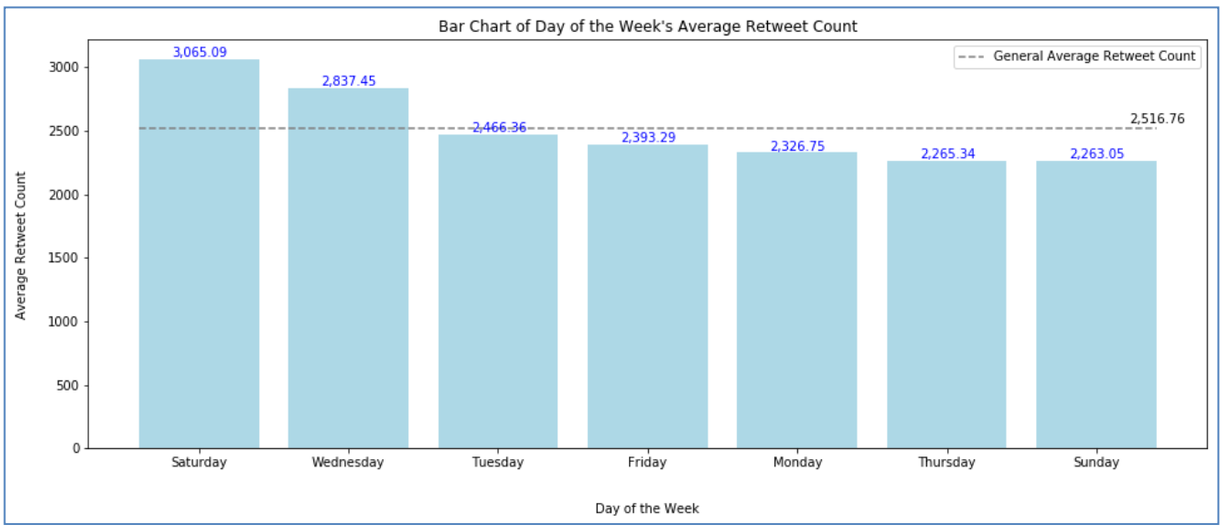
The goal of the project was to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. Using Python and its libraries, we gathered data from a variety of sources and in a variety of formats, assessed its quality and tidiness, then cleaned it before proceeding with the analyses. In this document we will explain the most valuable insights obtained.

Main Insights

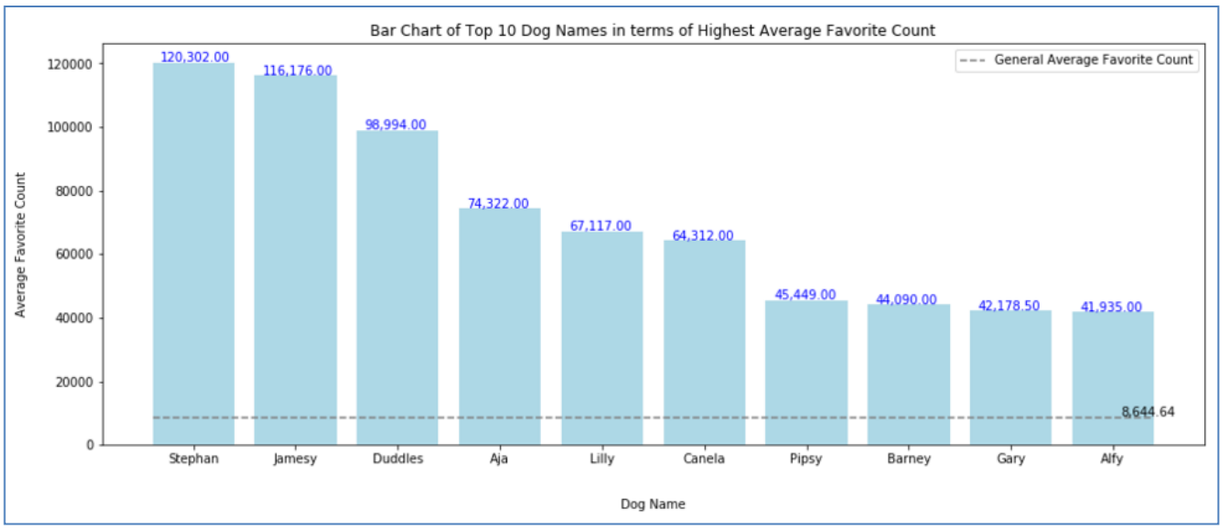
For our clean dataset (Rating Tweets of WeRateDog of one single dog and between November 15, 2015 and August 1st, 2017), we can conclude the following:

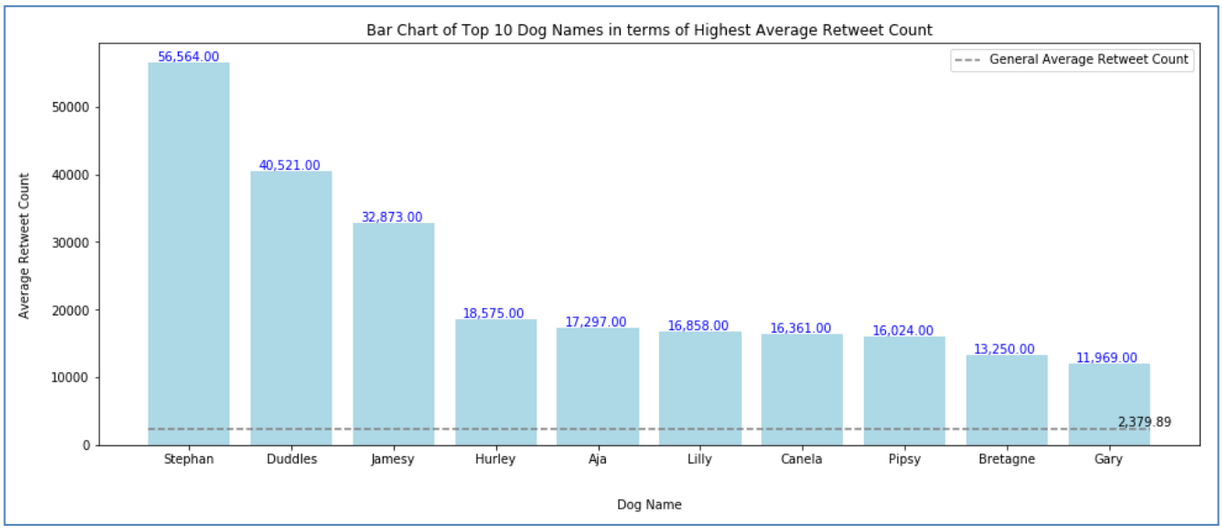


➔ Wednesday is the day with the Highest Average Favorite Count, 12% higher than the general mean.

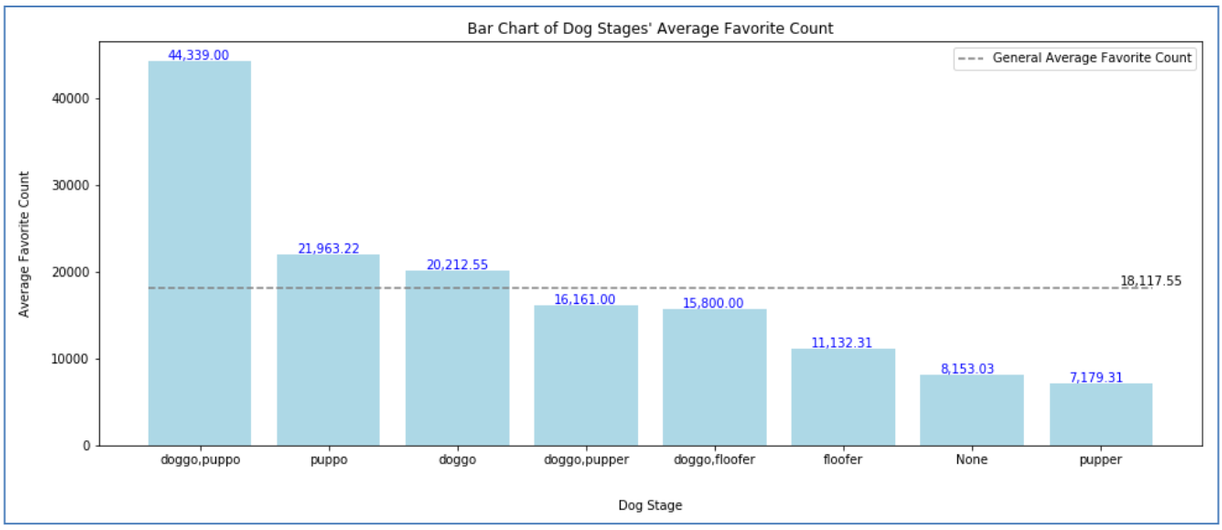


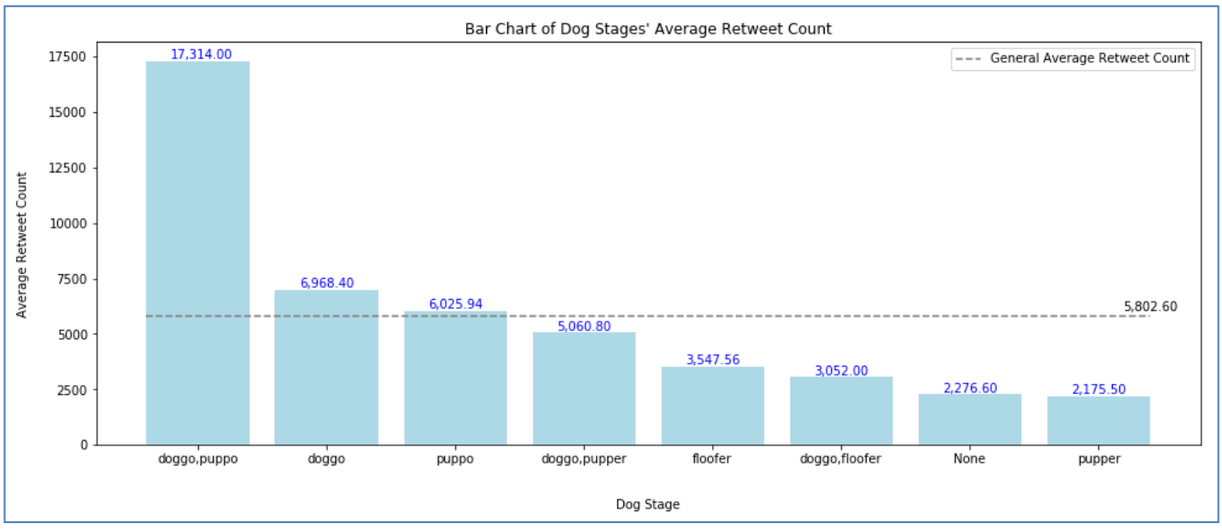
➔ Saturday is the day with the Highest Average Retweet Count, 22% higher than the general mean.



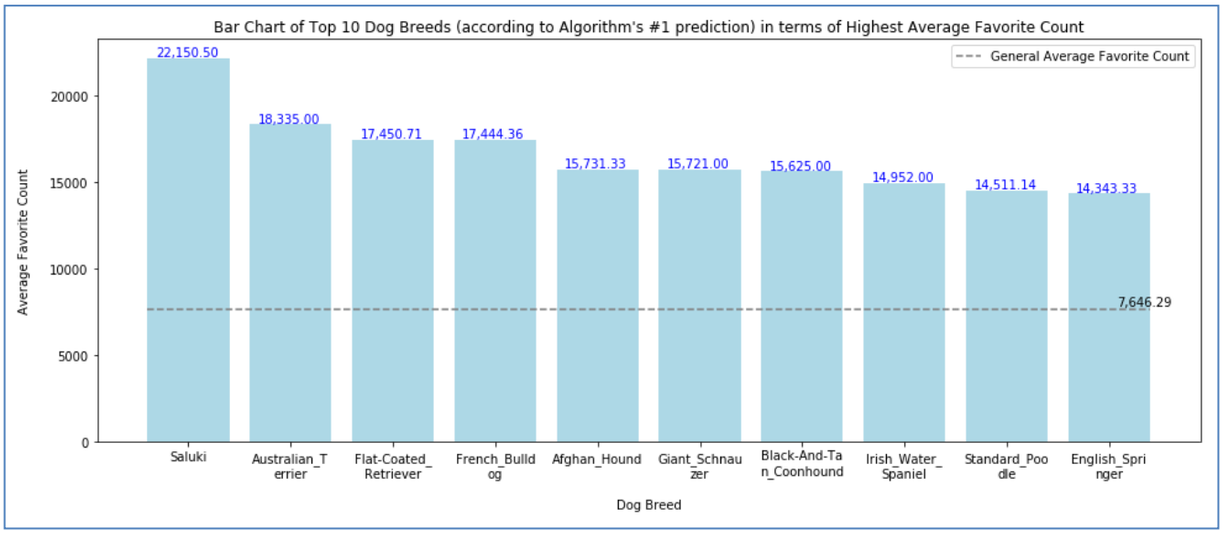


➔ Stephan is the dog name that received the Highest Average Favorite Count (12 times higher than the general mean) and the Highest Average Retweet Count (22 times higher than the general mean).

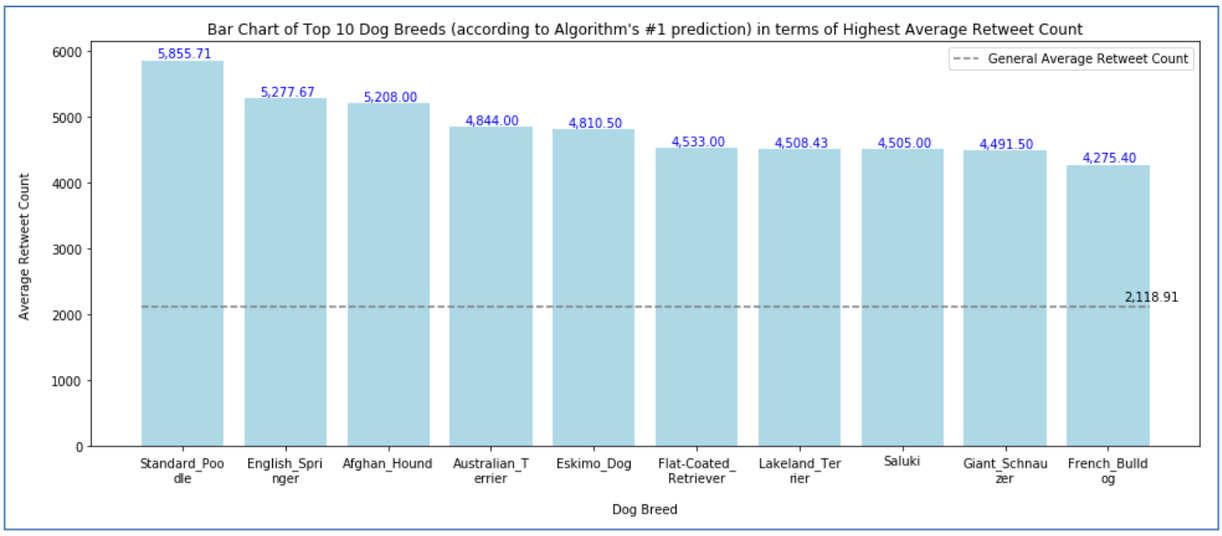




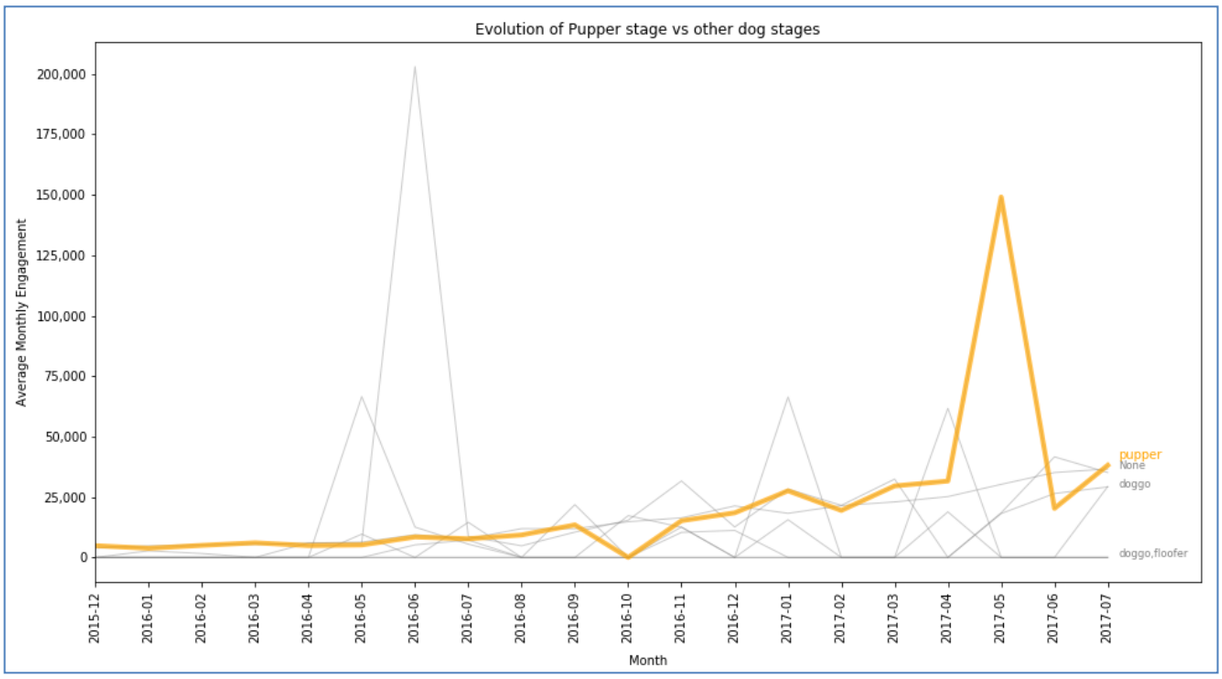
➔ The combination of doggo and puppo is the dog stage that received the Highest Average Favorite Count (144% higher than the general mean) and the Highest Average Retweet Count (198% higher than the general mean).



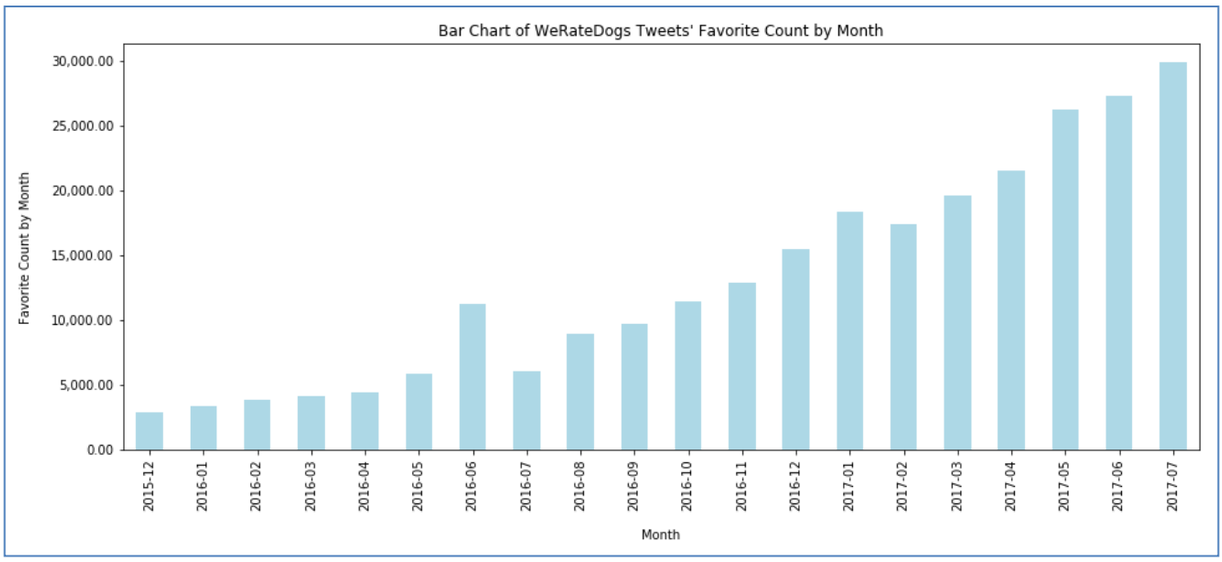
➔ Saluki (according to Algorithm's #1 prediction) is the dog breed that received the Highest Average Favorite Count, even 189% higher than the general mean.



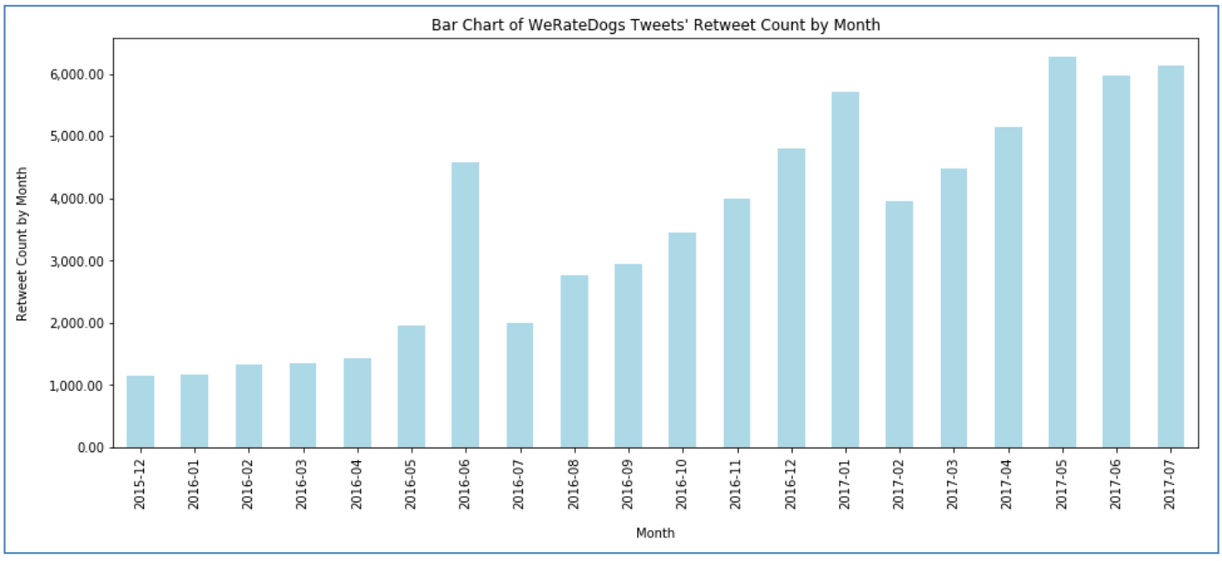
➔ Standard Poodle (according to Algorithm's #1 prediction) is the dog breed that received the Highest Average Retweet Count, even 176% higher than the general mean.



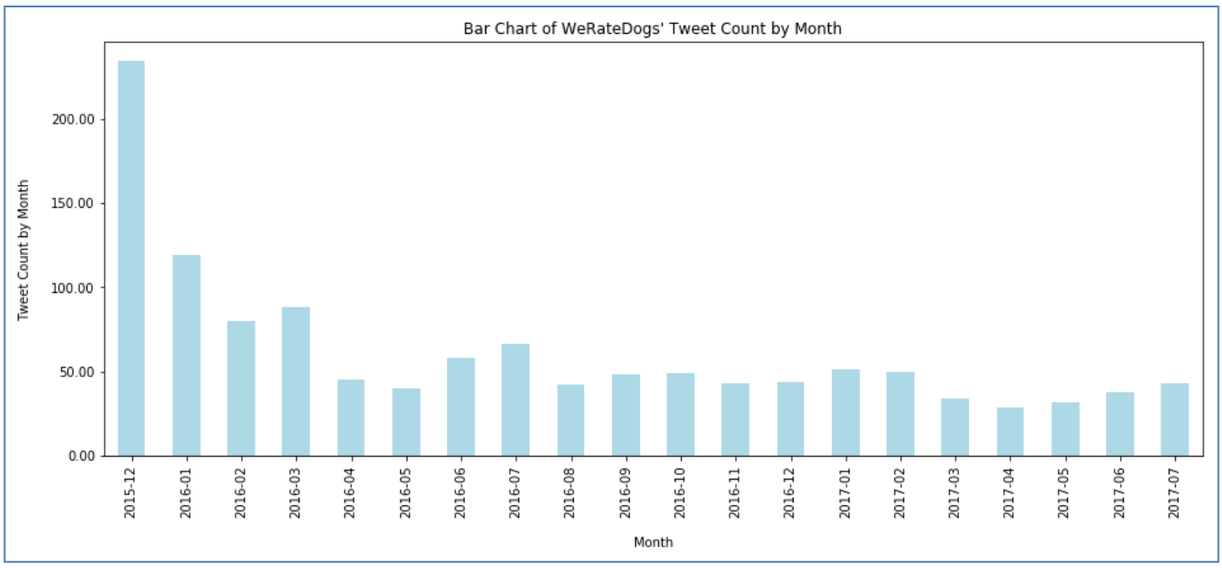
➔ "Pupper" stage has received the greatest Average Monthly Engagement (retweet count plus favorite count) among other dog stages mentioned in WeRateDogs rating Tweets through the months.



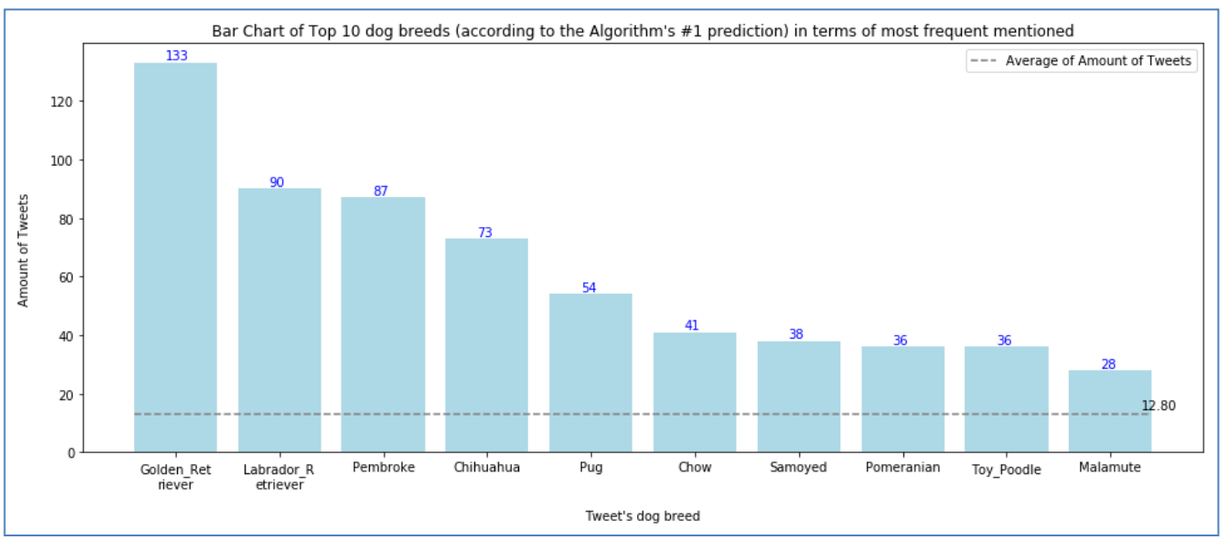
➔ WeRateDogs Tweets' Monthly Favorite Count has increased through the time, increasing its monthly count from 2,883 in December 2015 to 29,820 in July 2017 (9 times higher).



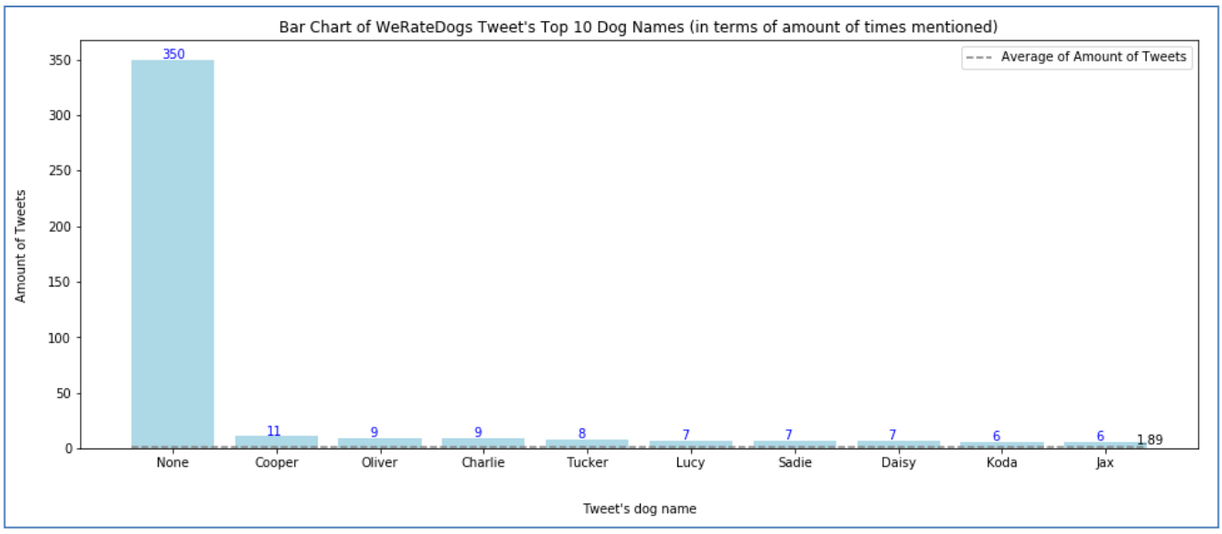
➔ WeRateDogs Tweets' Monthly Retweet Count has increased through the time, increasing its monthly count from 1,159 in December 2015 to 6,135 in July 2017 (4 times higher).



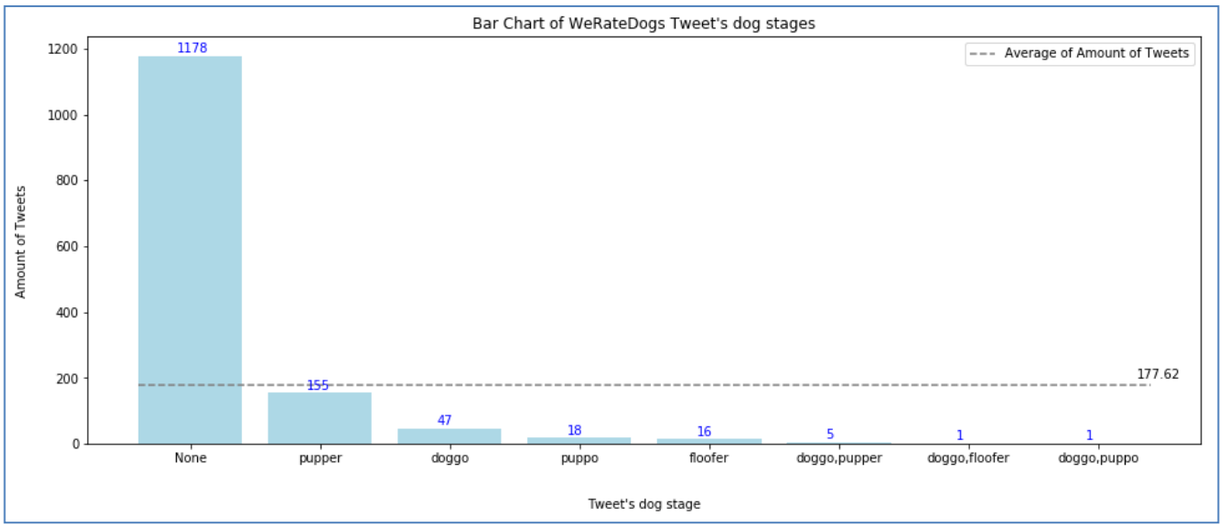
➔ WeRateDogs' Monthly Tweet Count has decreased through the time, decreasing its monthly count from 234 in December 2015 to 43 in July 2017 (82% less).



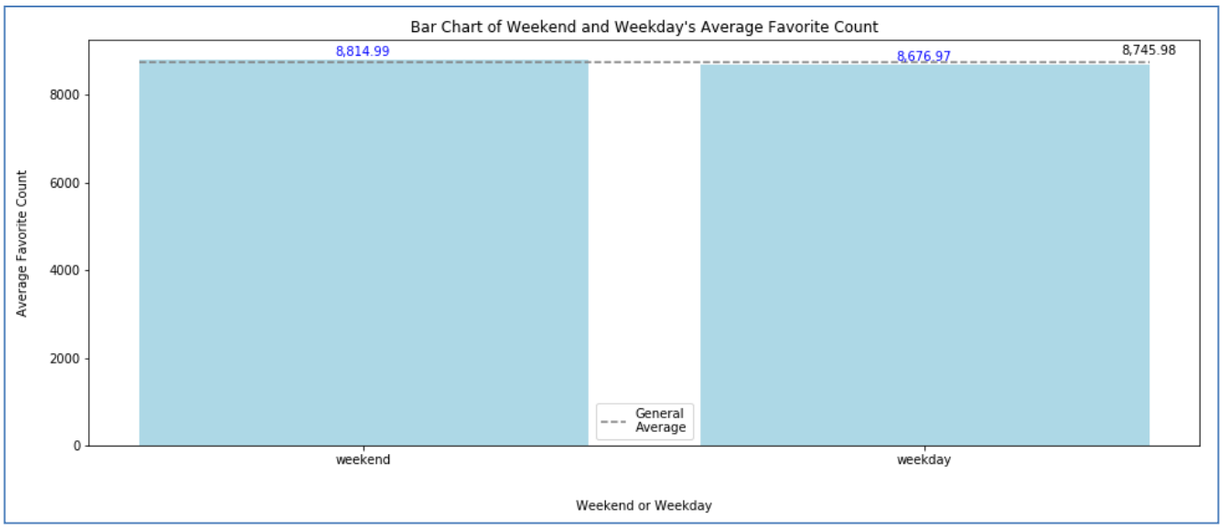
➔ The dog breed most included in the WeRateDogs' rating Tweets, according to the Algorithm's #1 prediction, was Golden Retriever with 133 Tweets (9%). This dog breed was mentioned 9 times more than the general average, according to the algorithm.



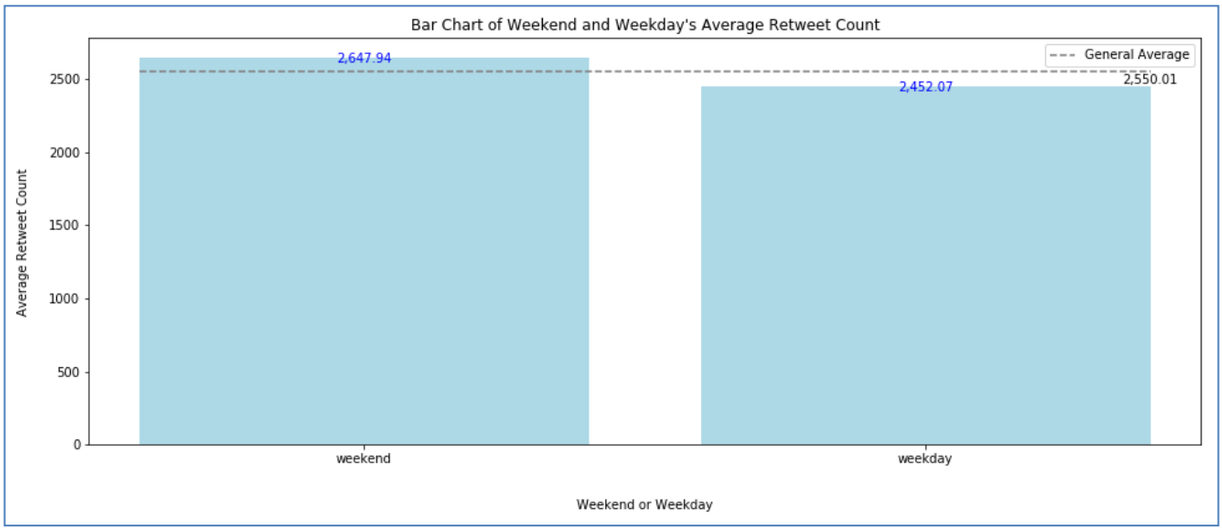
➔ Most WeRateDogs' rating Tweets did not include a dog name (25%). However, "Cooper" is the dog name most included with 11 Tweets (0.77%). This dog name was mentioned 4 times more than the general average.



➔ Most of the Tweets do not have a dog stage mentioned in it (83%). Out of the rating Tweets who do have a dog stage, the most prominent one is pupper with 155 Tweets (11%).



➔ WeRateDogs Weekends' Tweets have a higher Average Favorite Count than Weekdays Tweets. However, just 1.6% higher than Weekdays' average.



➔ WeRateDogs Weekends' Tweets have a higher Average Retweet Count than Weekdays Tweets. However, just 8% higher than Weekdays' average.

Other Insights

- "None" dog stage Tweets were the most predominant (with highest amount of Tweets) through the months.
- WeRateDogs Tweets' Weekly Favorite Count has increased through the time, increasing its weekly count from 880 in the second week of November 2015 to 28,798 in the second week of July 2017.
- WeRateDogs Tweets' Weekly Retweet Count has increased through the time, increasing its weekly count from 181 in the second week of November 2015 to 5,848 in the second week of July 2017.
- The majority of Tweets (between 25% and 75% percentile):
    - had a rating's numerator between 10 and 12.
    - had a rating's denominator of 10.
    - had a rating between 1 and 1.2.
    - had a retweet count between 592 and 2,881.
    - had a favorite count between 2,044 and 10,883.
    - used the first image for the breed prediction.
    - had a Text Length between 94 and 137.
- The majority of Algorithm's Breed #1 Predictions of WeRateDogs Tweets (between 25% and 75% percentile) had a confidence between 39% and 85%.
- Most WeRateDogs' rating Tweets are in English, with 1416 Tweets (99.6%).
- Most WeRateDogs' rating Tweets come from iPhone, with 1398 Tweets (98%).
- The highest favorite count a WeRateDogs Tweet has ever received is 155,733.
- The highest retweet count a WeRateDogs Tweet has ever received is 77,599.
- The highest Rating's Numerator a WeRateDogs Tweet has ever given is 165.
- The highest Rating's Denominator a WeRateDogs Tweet has ever given is 150.
- The highest Rating a WeRateDogs Tweet has ever given is 1.4.
- The average WeRateDogs Tweet's Text Length, in our clean dataset, is 111.29.
- The highest "Algorithm's Breed Prediction Confidence of WeRateDogs Tweets" is 99.9956%.