WeRateDogs Data Wrangling Project

22 May 2019

Introduction

This project is part of the data wrangling section of the Udacity Data Analyst Nanodegree program and is primarily focused on wrangling data from the WeRateDogs, which is a Twitter account that rates people's dogs with a humorous comment about the dog. The rating denominator is usually 10, however, the numerators are usually greater than 10. This aspect was not cleaned as it is part of the humor and popularity of WeRateDogs.

Project Goal

Fully assessing and cleaning the entire dataset would require exceptional effort so only a subset of its issues (eight quality issues and two tidiness issues at minimum) needed to be assessed and cleaned.

The tasks for this project were:

- Data wrangling, which consisted of:
    - Gathering data
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing the wrangled data

- Reporting on my data analyses and visualizations (act_report.pdf)

Data Wrangling Process

Gather

I gathered data from three sources:

- The Twitter archive #WeRateDogs, a csv file that contains the tweet, rating, dog name, and dog 'stage' in life (such as puppy).
- An 'image prediction' file, or what breed of dog is in each tweet, according to a neural network. I downloaded the image prediction file programmatically from Udacity's servers using the Requests library.
- Twitter's API to gather retweet count and favorite count,

Assess

I assessed the data based on quality and tidiness.

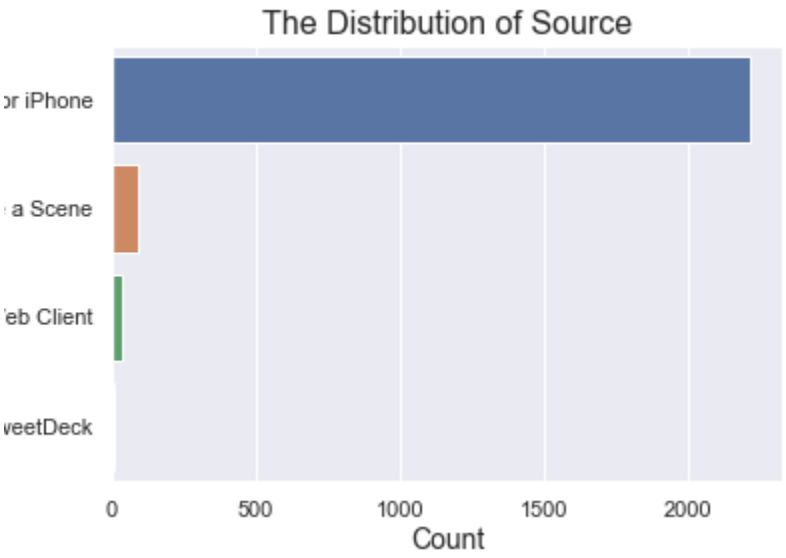
Clean.

- Dropped unnecessary columns
- Converted erroneous data types
- Solved the NaN values in expanded_urls column from twitter_archive dataset
- Solved the Invalid dogs' names in twitter_archive dataset, and capitalized the first letter of dog name for consistence
- Changed the incorrect rating_numerator and rating_denominator values then created a new column rating = rating_numerator / rating_denominator dropped oberservations with extreme ratings (removing outliers.)
- Chose only a dog breed with the highest confidence each row. Tweets which are not predicted as a dog are set as missing value.
- Removed observations which are predicted as non-dog.
- Optimized the source content by ‘Twitter for iphone’, ‘Vine - Make a Scene’, ‘Twitter Web Client’, and ‘TweetDeck’.
- Created a ‘prediction’ column
- Combined Dog stages columns (i.e doggo, floofer, pupper & puppo)
- Merged the 3 data sets into one data set

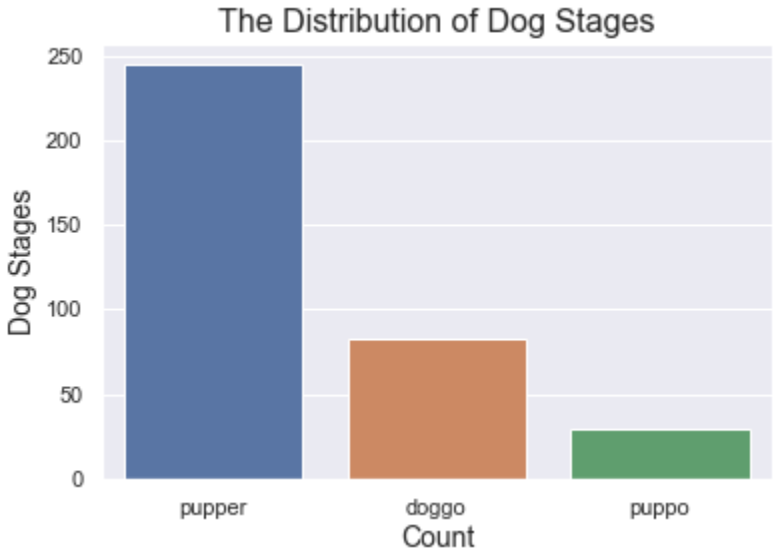
CONCLUSION

I analyzed the information in the clean, combined dataframe and created initial visuals using Matplotlib in Python. Finally, I created a custom visualization of my findings:

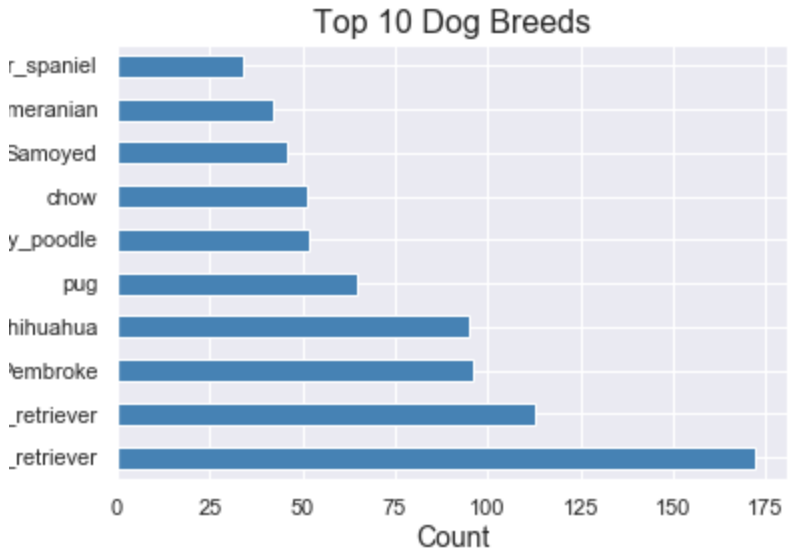
- The dominate source of tweets is from iPhone twitter app, which is 94% in the total. That means the twitter app is the main channel for people using to tweet, retweet, post, and others, while the TweetDesk is pretty rare (less than 1%).



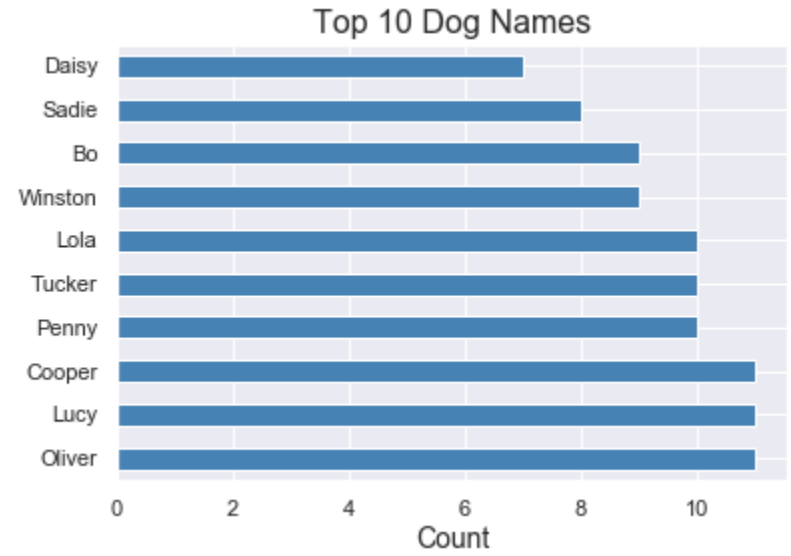
- It shows that ‘pupper’ (a small doggo, usually younger) is the most popular dog stage, followed by ‘doggo’ and ‘puppo’.



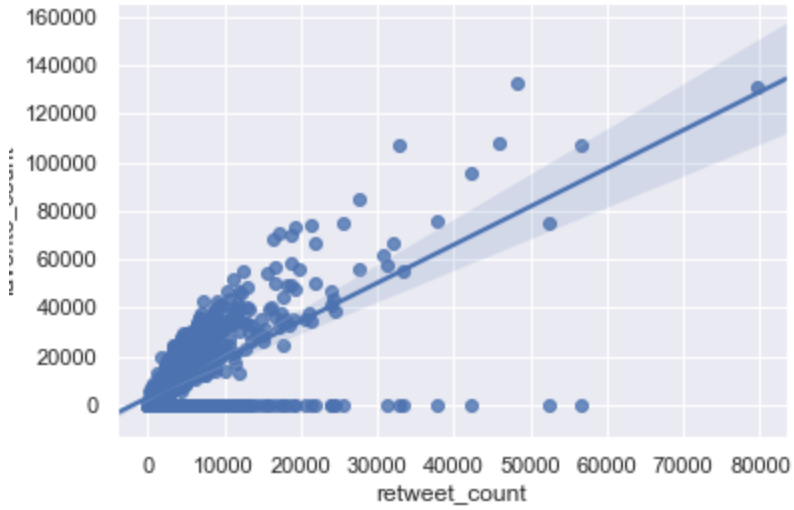
- There are more golden Retrievers than any other dog in the dataset. Labrador Retrievers are the second most common.



- Oliver, Lucy and Cooper are the most popular dog names.



- There is a strong positive correlation between number of retweets and favorite count. That is reasonable, the more a post is retweeted, the more eyes view the post, the more favorites the post receives


160000
140000
120000
100000
80000
60000
40000
20000
0
0
10000
20000
30000
40000
50000
60000
70000
80000
retweet_count