Act Report

Wrangle and Analyze Data

Jiawei He

1. Introduction

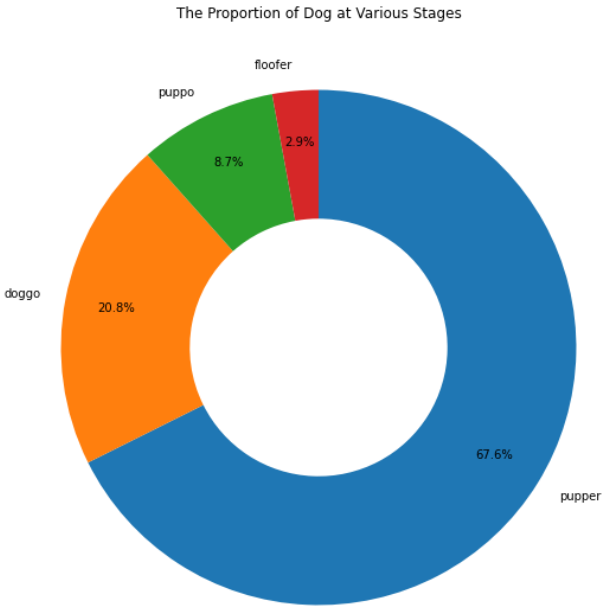
In this project, the following three datasets were worked on: the WeRateDogs twitter archive, image predictions file and additional data via the twitter API. WeRateDogs is a tweeter who rates people's dogs with a humorous way. The image predictions file continues to detect the collected images through a neural network that can classify dog breeds. This data includes the top three image predictions, the ID of each tweet, and the URL of the image. Data obtained from twitter_api.py including tweet ID, retweets and favorites. After data gathering, data assessing and data cleaning, I have visualized and analyzed the data.

2. Analyzing and Visualizing Data

1) The proportion of dog at various stage

The data divides the dogs into 4 stages: doggo, pupper, puppo, and floofer.

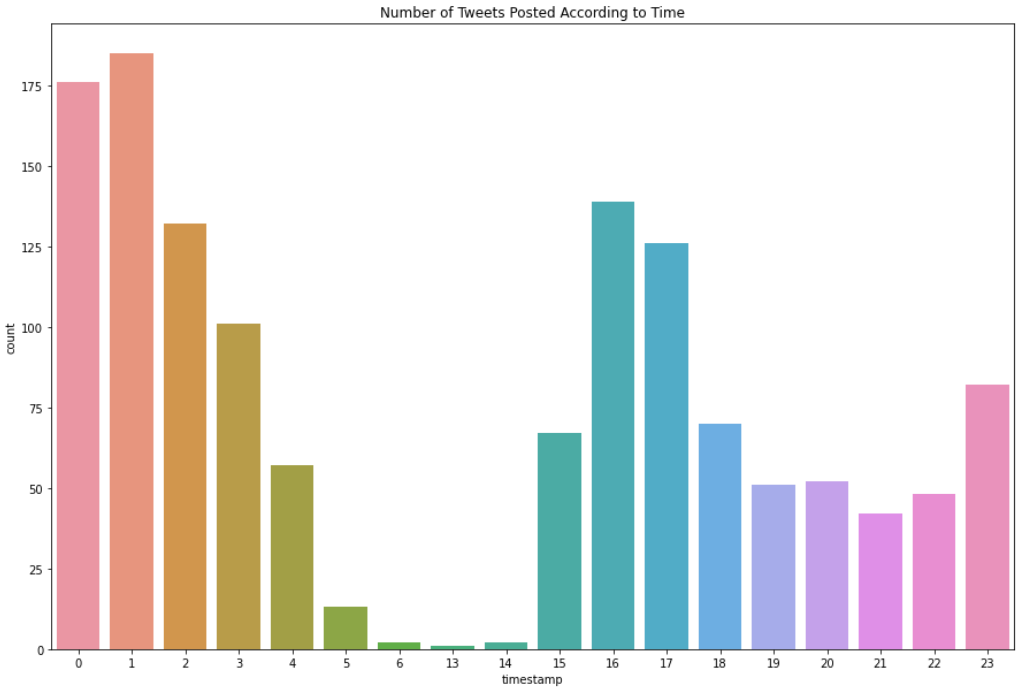
Conclusion 1: The proportion of pupper are the most, next is doggo.



2) Number of Tweets Posted According to Time

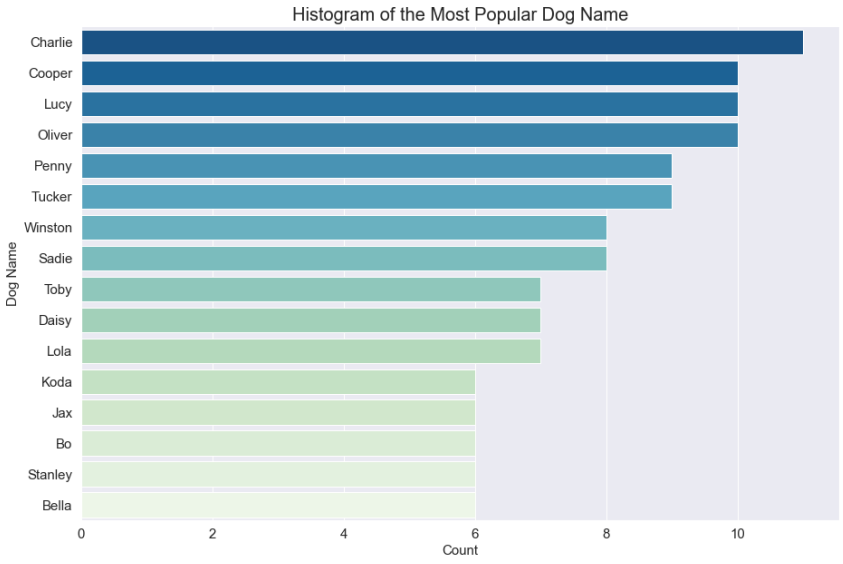
I was curious about what time of day people like to post these tweets, so I counted the number of tweets posted at different times of day.

Conclusion 2: It can be clearly seen that the number of posts at midnight is higher than at other times and the number of tweets posted was the highest at 1 am.



3) The Most Popular Dog Name

Conclusion 3: The most popular dog name is Charlie.



4) Retweet_count and Favorite_count

Conclusion 4: retweet_count and favorite_count have a strong positive correlation, i.e. the larger the retweet_count, the larger the favorite_count in general.