Analysis_Final_Report

October 21, 2018

1 Analyzing , and Visualizing WeRateDoge,

2 Dataset after Wrangling Process

3 Introduction

Data wrangling is a core skill that everyone who works with data should be familiar with since so much of the world's data isn't clean. We need to wrangle our data for good outcomes, otherwise there could be consequences. If we analyze, visualize, or model our data before we wrangle it, our consequences could be making mistakes, missing out on cool insights, and wasting time. So best practices say wrangle. Always. Through this project could made:

Data wrangling, which consists of: Gathering data Assessing data Cleaning data Storing, analyzing, and visualizing our wrangled data Reporting on - our data wrangling efforts - our data analyses and visualizations

In this Project we are going to use WeRateDogs for the data Wrangling Process. The WeRateDogs Twitter archive. We will download this file manually by clicking the following link: twitter_archive_enhanced.csv The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv Each tweet's retweet count and favorite (i.e. "like") count at minimum, and any additional data we will find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, we will query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then we will read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

4 Now lets see the rating plot over time

```
More than 75% of the data has more than 12/10 as rating.
The page starts with small rating that they adopt the system of rating numerator more than the
```

Let us see the most correlated variable, favorites and retweets.
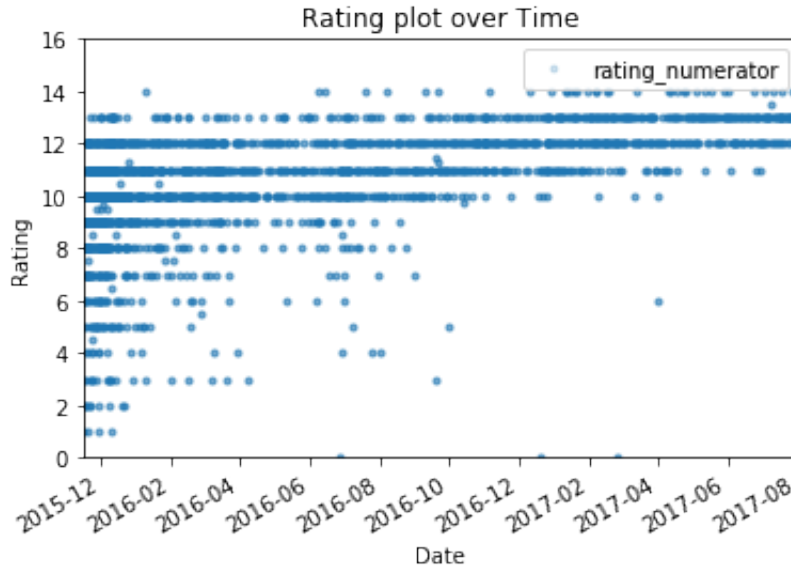
alt text

5 Favorites Vs Retweets

As the correlation map shows if the counts of the retweets is high for a specific tweets the count of the favorites goes high as well.

```
The most retweeted post was a doggo realization that he can stand in pool with 78,809
 retweets.
The most favorites post we a dog marching in the 2017 womens march with 131,903 favorites.
```
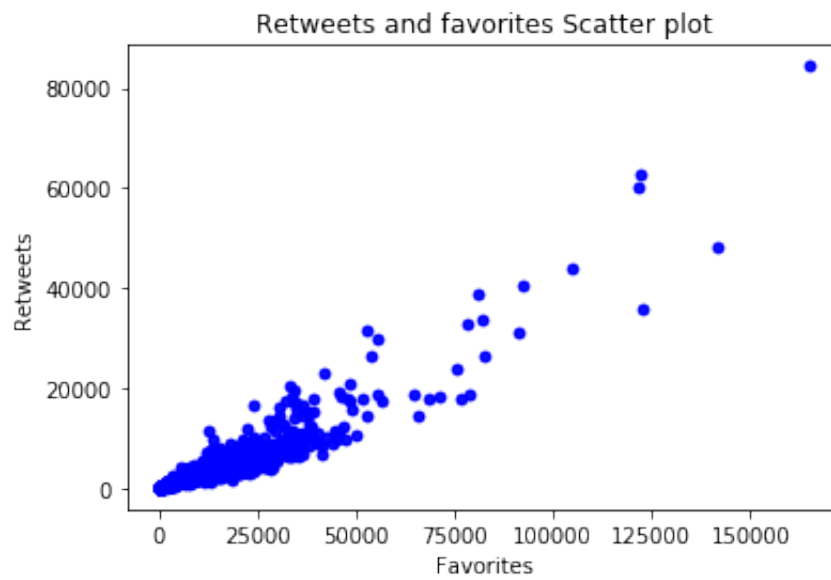
6 Famous dog Breed and gender:

- Top two famous breeds are, Golden_retriver and Labrador according to the neural network we used to classify the breeds of dogs.
- We have male dogs more than female dogs in our dataset ,yet the female rating is more than the male rating.

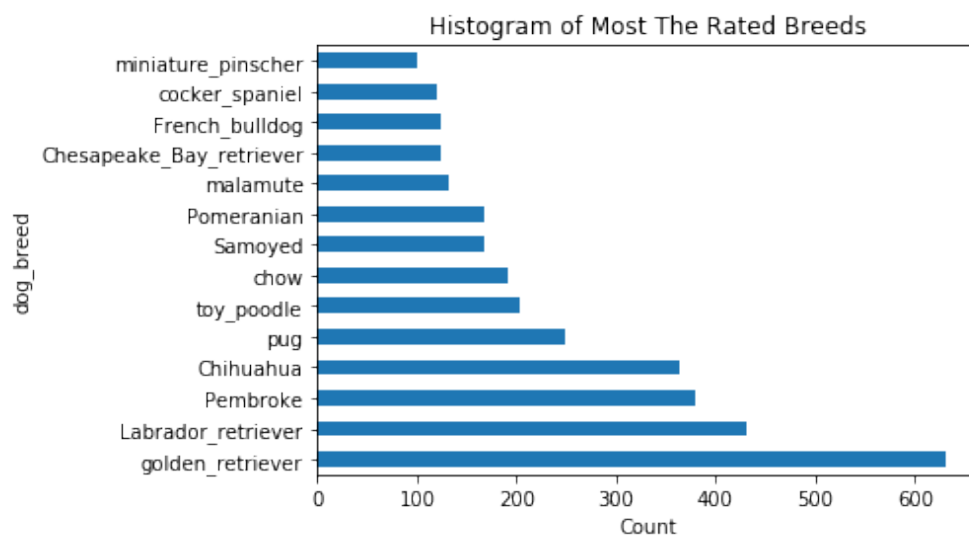## 7 DoggoLingo or Dog Stages :

During your dog's life, bouncing baby pup to elderly matriarch, your dog express different needs and tender a range of rewards at each stage of life. Those four stage are represented in our dataset in a form of a box plot.
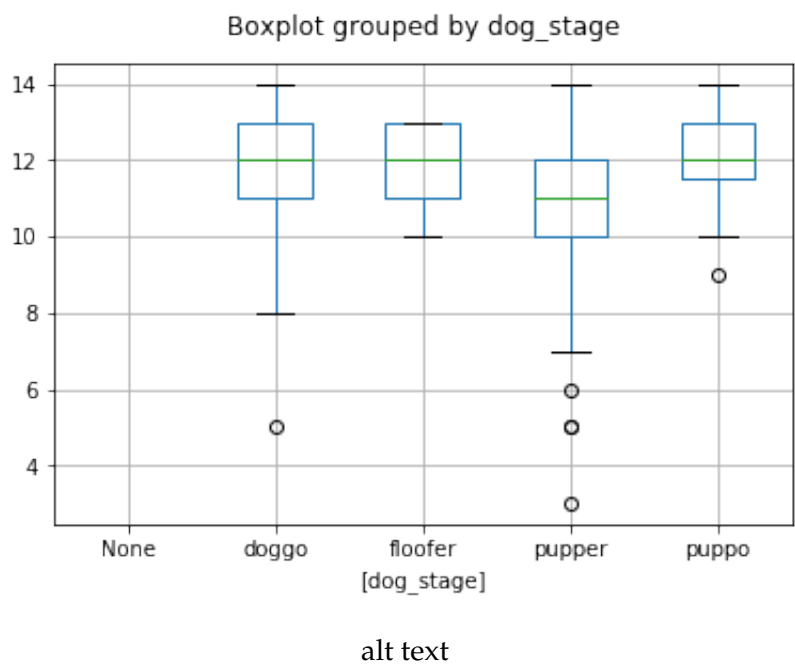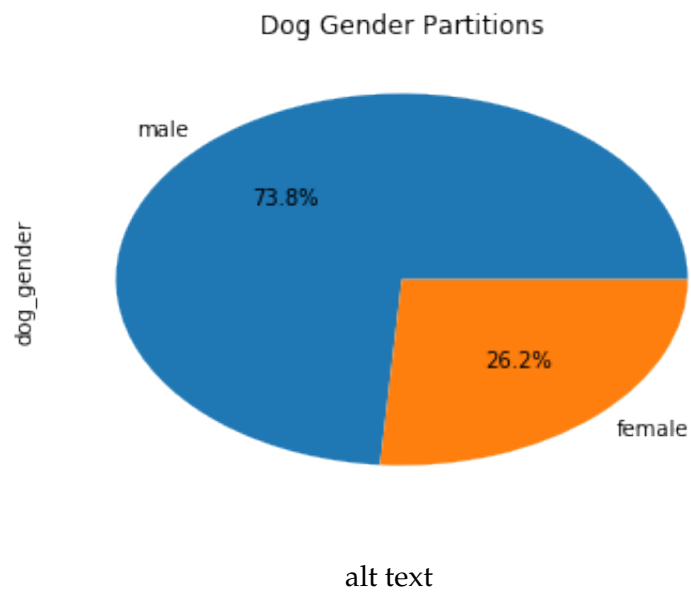
8 Conclusion

The Twitter account WeRateDogs (@dog_rates) is devoted to humorously reviewing pictures of dogs doing adorable poses. Dogs are rated on a scale of one to ten, but are invariably given ratings in excess of the maximum, such as "13/10". It has acquired over 4.50 million followers

Retweets and favorites Scatter plot

alt text

Histogram of Most The Rated Breeds

alt text

Dog Gender Partitions

alt text

Boxplot grouped by dog_stage

alt text

since its debut. If you are thinking of adopting a dog, get a floof(er) - far away from being pupper at least still in the stage of puppo yet not a doggo to live more togother.

Data wrangling indeed is a core skill that everyone who works with data should be familiar with since so much of the world's data isn't clean. If we analyze, visualize, or model our data before we wrangle it, our consequences could be making mistakes, missing out on cool insights, and wasting time. We couldn't be able to make some of the visualization without wrangling (i.e dog gender partition) So best practices say wrangle.

In [ ]: