Report on Analysis, Visualization and Insights of the Final Data.

By:

Michael A. Olojo

Introduction:

The dataset that was wrangled is the tweet archive of Twitter user@dog_rates, also known as WeRateDogs. This is a Twitter account that rated people's dogs with different comment about the dogs.

The project focuses on data wrangling which involves gathering, assessing and cleaning. It's also involves analysis, visualization and insight about the report.

Gathering Data:

My wrangling effort for the WeRateDogs project involves gathering data from the following sources:

- The WeRateDogs Twitter Archive. The twitter_archive_enhanced.csv file was provided to students by Udacity, which contains tweet data e.g tweet ID, timestamp, text etc for all 5000+ of their tweet.

- The tweet image prediction. The image prediction file contains the breed of dogs that is present in each tweet according to a neural network. This was also provided by Udacity to the students.

- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite count at minimum. Alternatively, the data was provided on the online content by Udacity, if students are not satisfied or comfortable with the first option.

Assessing Data:

There are four main issues about quality of data which are:

- Completeness: Missing Data
- Validity: Data making sense
- Consistency: Standardization
- Accuracy: Inaccurate data

There are three main requirements of tidiness:

- Each variable forms a column
- Each observation forms a row
- Each type of observation unit forms a table.

Cleaning Data:

This involves three steps which are:

Define: Determine exactly what need to be cleaned.

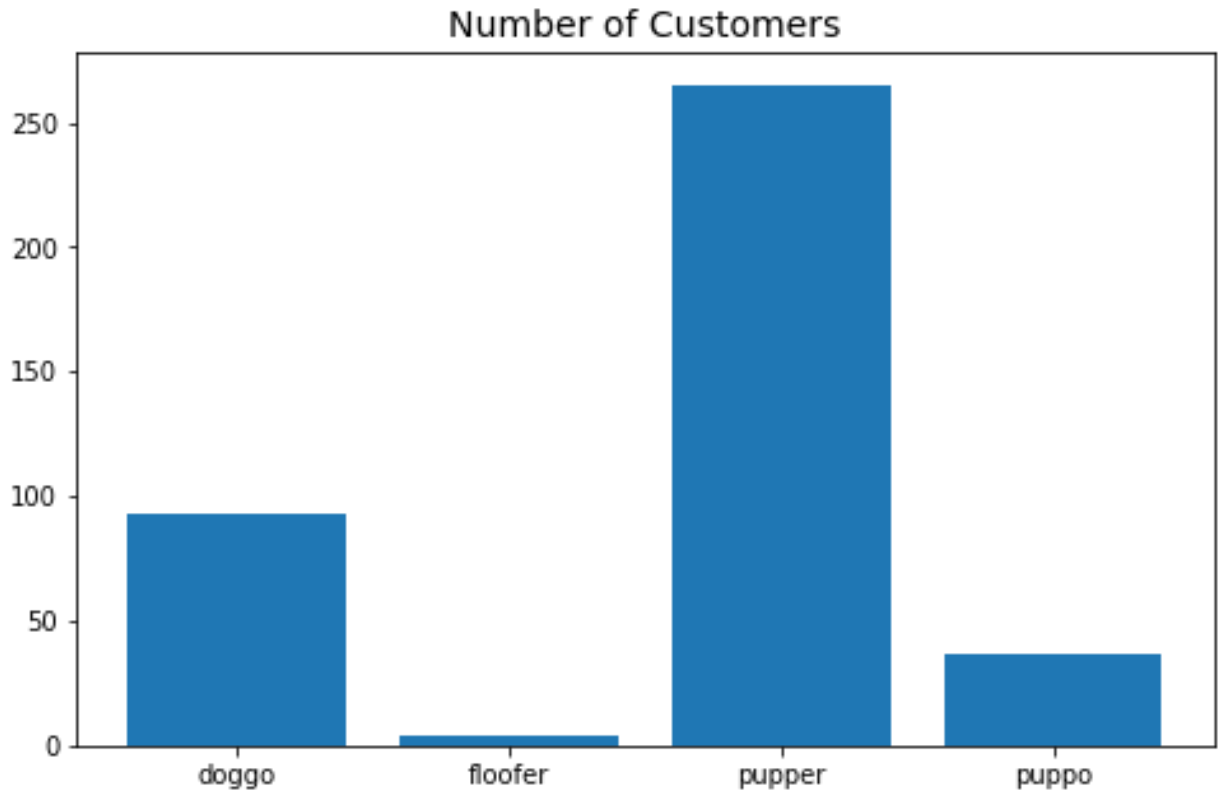
Code: Clean the dataset programmatically.

Test: Evaluate the code to know if the dataset has been properly cleaned.

Analysis and Visualization:

Dog Stage Ratio:

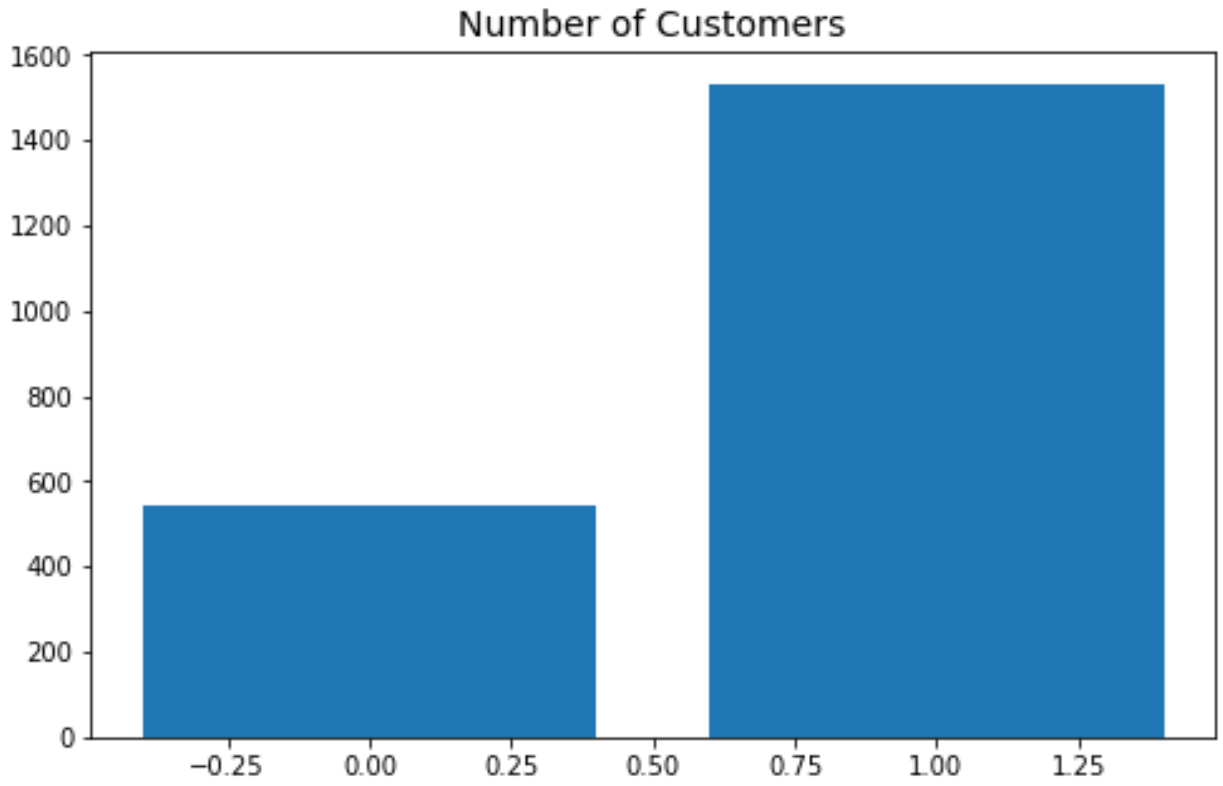
In this analysis, I was able to check the dog stage ratio based on the number of customers. The chart shows that the pupper stage had the highest number, followed by doggo, then puppo and floofer as the last stage.

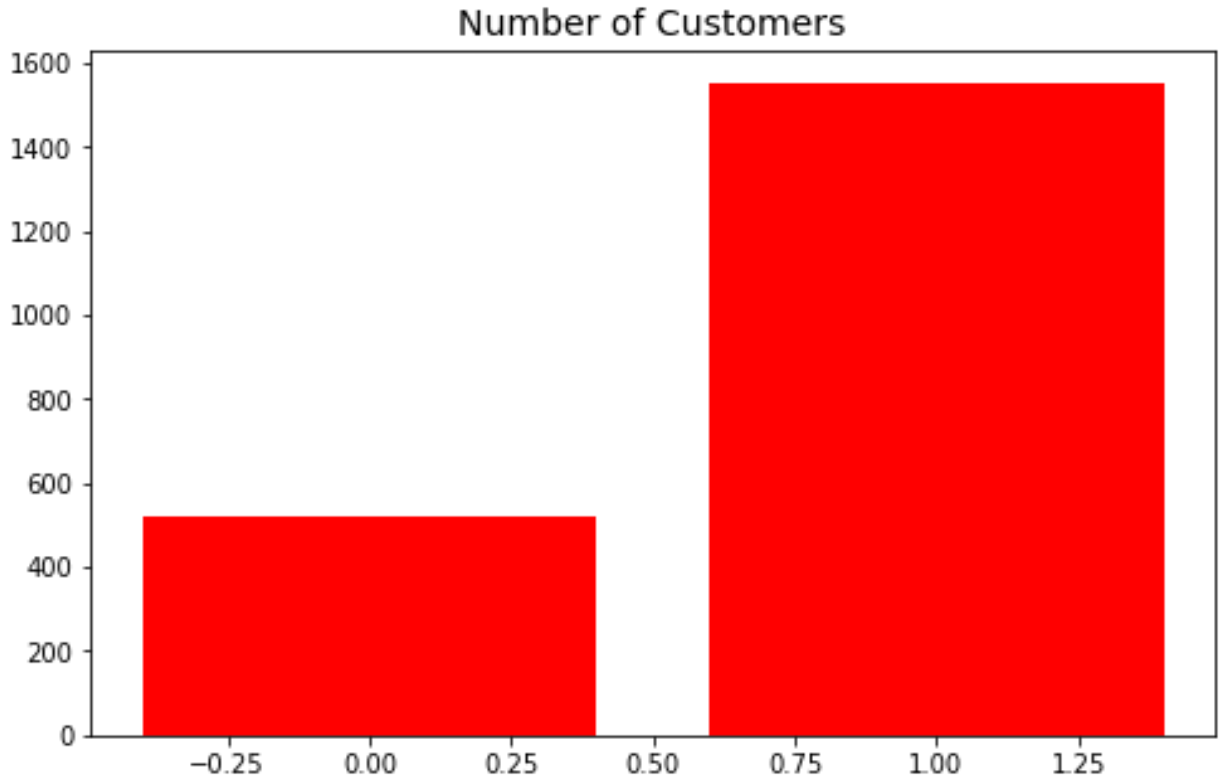


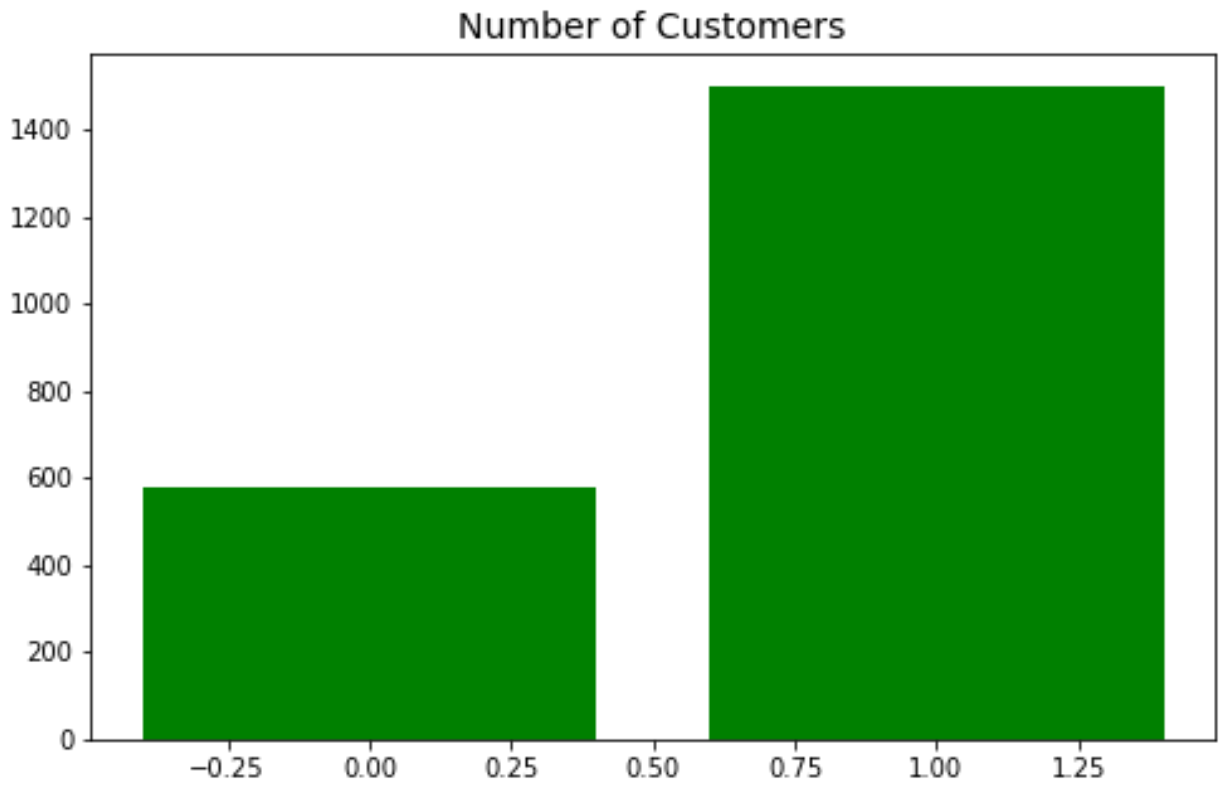
Success Rate of Algorithm:

In this analysis, I was able to figure out the most successful algorithm amongst p1_dog, p2_dog and p3_dog.

Having done all analysis, it was established that p1_dog was 73.8%, p2_dog was 74.8% and p3_dog was 72.2%. This shows that p2_dog was the most successful while p3_dog was the least successful.


Number of Customers
1600
1400
1200
1000
800
600
400
200
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25


Number of Customers
1600
1400
1200
1000
800
600
400
200
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25


Number of Customers
1400
1200
1000
800
600
400
200
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25