Wrangle and Data Project

ANALYSIS AND INSIGHTS OF THIS PROJECT

Oseremen Sandra Osara | Udacity's Junior Data Analysis Nanodegree | May 20, 2022

Introduction:

Real-world data rarely comes clean. The dataset that I worked on (wrangling and analyzing and visualizing) 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.

This project works through the data wrangling process, focusing on the gathering, assessing, and cleaning of data. There are visualization and observation from the analysis provided as well

GATHER:

This project gathered data from the following sources:

• The WeRateDogs Twitter archive. The 'twitter-archive-enhanced-2.csv' file was provided to Udacity Students (Like me). This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

• The Tweet image prediction, i.e., what breed dogs (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students (Like me).

• Twitter API and Python's Tweepy library to gather each tweet's retweet count and favourite ("like") count at minimum and any additional data I find interesting.

ASSESSING DATA:

After gathering all three pieces of data, assess them visually and programmatically for quality and tidiness issues.

I used two types of assessment:

Visual assessment: Each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes. Once displayed, data can additionally be assessed in an external application (e.g., Excel, text editor).

Programmatic assessment: Here, I used different pandas' functions and/or methods to assess the data.

At the end of these, I came out with some quality and Tidiness issues which were taken care of in the cleaning stage.

CLEAN

Cleaning data is tedious and often iterative. Just when data analyst believe they found all quality and tidiness issue, they often found additional issue arises. The cleaning process involves three steps:

In this stage, All the issues that were identified when accessing were cleaned.

Before cleaning, copy of the original data was created. The Define-Code-Test Framework was used for the session:

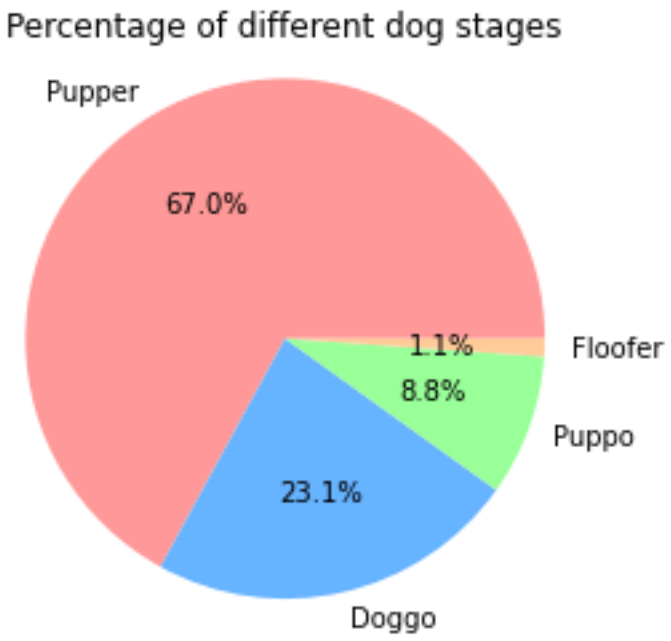
- Define: Determine exactly what needs to be clean and how.
- Code: Programmatically clean the code
- Test: Evaluate the code to ensure the data set was cleaned properly

ANALYSIS AND VISUALIZATION

With the data that was cleaned, a lot of analysis can be done but I worked on just 3 insights.

1. The Percentage of different dog stages
2. Relationship between Favorite count and retweet count
3. The Most Popular Dog Breed.

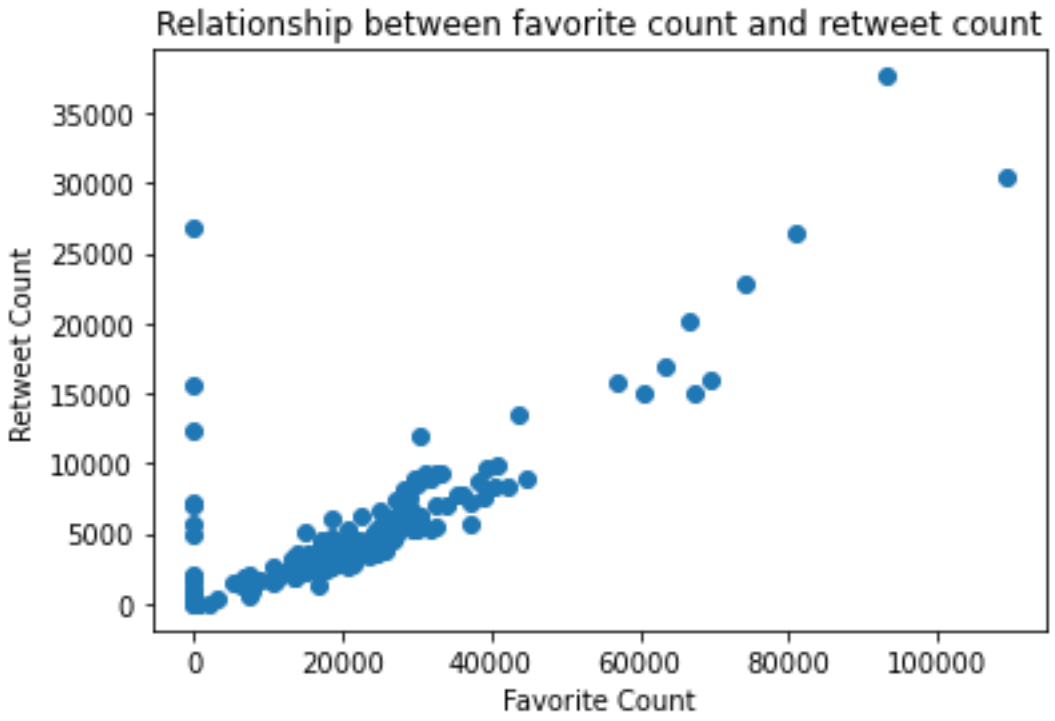
The Percentage of different dog stages



INSIGHTS(A)

1. Pupper is the dog stage with the highest percentage(67%)
2. Floofer has the lowest percentage (1.1%)

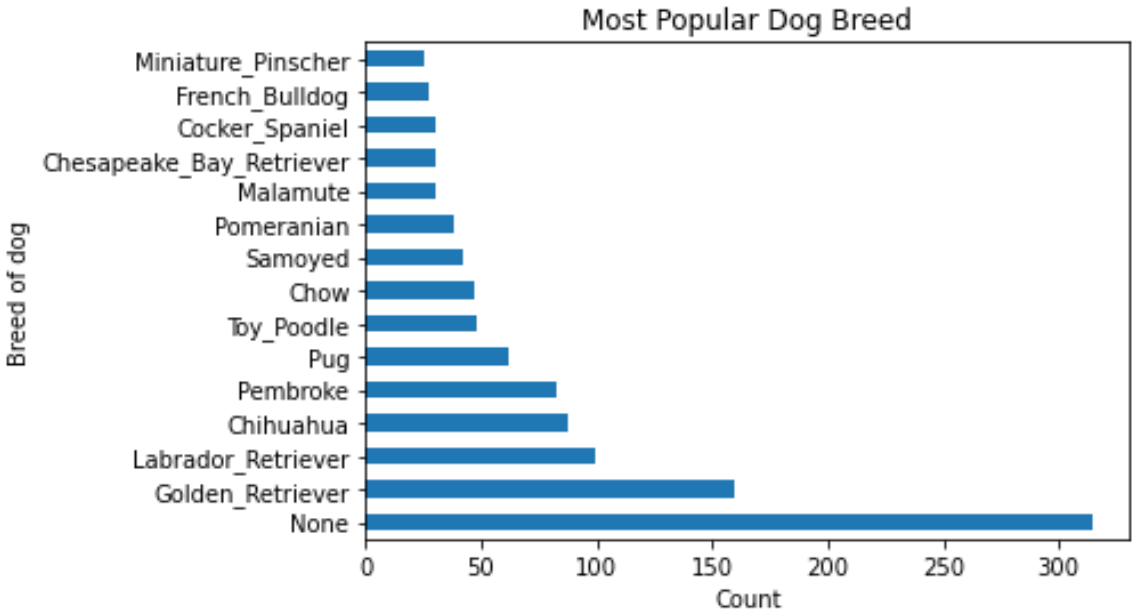
Relationship between Favorite count and retweet count



INSIGHTS(B)

There is a high positive correlation between retweet count and the favourite count. The favourite count increases as the retweet counts increases. This correlation is important for the owner of the WeRateDogs twitter account to understand when determining method to increase users' traffic on the page. A data analysis team could recommend previous posts with either a high retweet counts or high favorite count so that page owner could model future posts off historically popular posts.

The Most Popular Dog Breed



INSIGHTS(C)

From the Bar Chart, the most popular dog breed is a Golden Retriever, with a Labrador Retriever coming in as the second most popular breed.

The owner of the WeRateDogs twitter account could use this information to create targeted marketing efforts for certain breed that aren't popular to increase their popularity, but also utilize the breeds that are proven to be popular to drive user traffic to the page.