Analysis Report For the WeRateDogs Project

In this article I'll be sharing the insights gotten from a data wrangling project I carried out.

About the project

The aim of this project is to demonstrate my proficiency in collecting data using different methods, accessing data, and cleaning data for use. The datasets used in this project contains information about tweets from the WeRateDogs Twitter account. In order to complete the project these were the objectives I achieved

1. Collect data about the WeRateDogs Twitter account using three methods.

- Manual download.
- Programatic download.
- Web scraping using Tweepy.

2. Access the collected data to discover data cleanliness issues.

3. Clean identified issues using the 'Define, Code, and Test' model.

The projects was done in Python programming language. The different steps followed to achieve the goal of the project were iterative and overlapping. This is typical in any data analysis task.

Data Analysis Findings

My analysis was focused on answering to the following questions.

1. What dog class is the most popular

2. What name is the most common dog name in our data

3. What is the average retweet_count and favorite per dog class

4. What is the distribution of the favorite_count for WeRateDogs tweets

5. What is the distribution of the favorite_count for WeRateDogs tweets

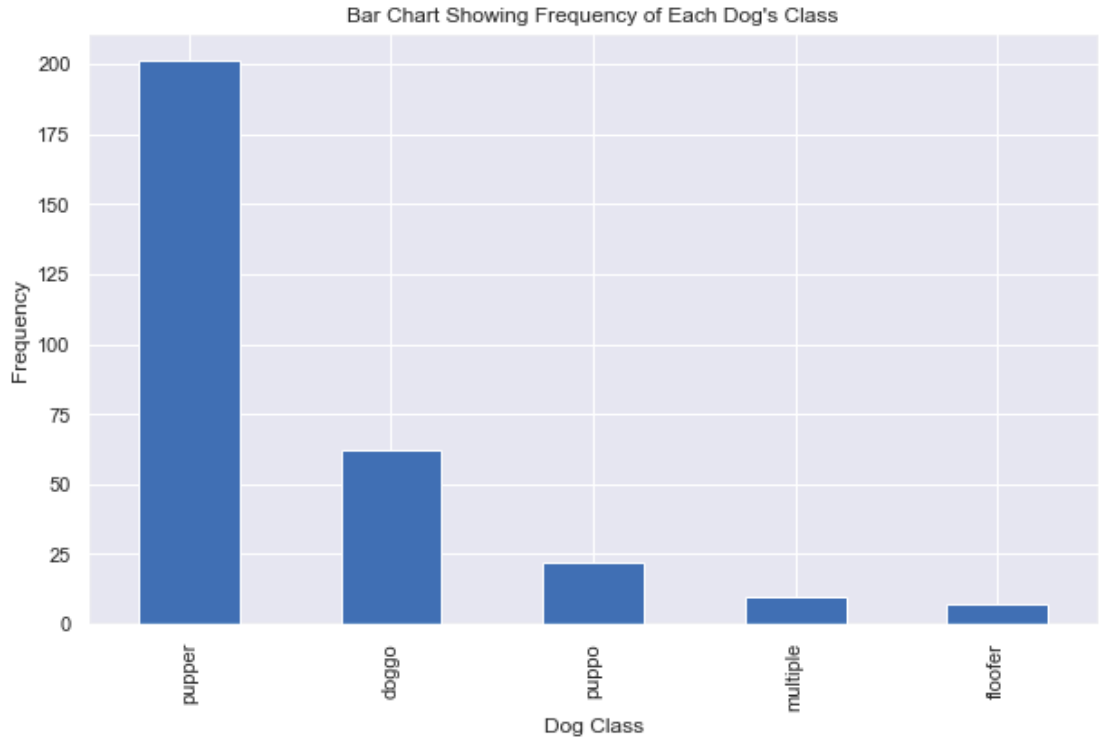
6. Which dogs had the highest retweet and favorite counts

The answers to the questions are shown in the visualisations below

1. Most popular dog class.

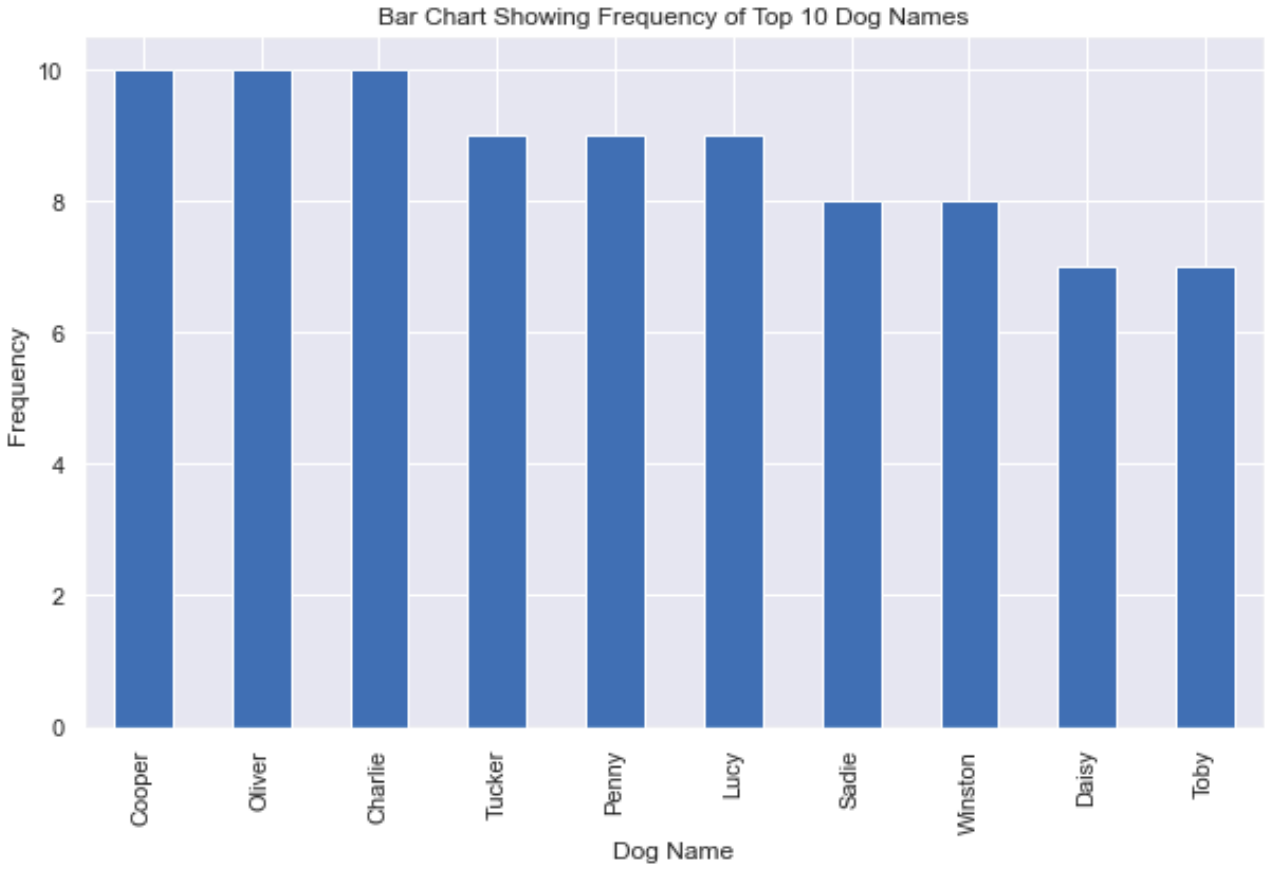
The Pupper class is the most popular dog class in the dataset. The distribution of the classes are given below.

Dog Class	Population
pupper	201
doggo	62
puppo	22
Multiple	10
floofer	7



2. Ten most popular dog names.

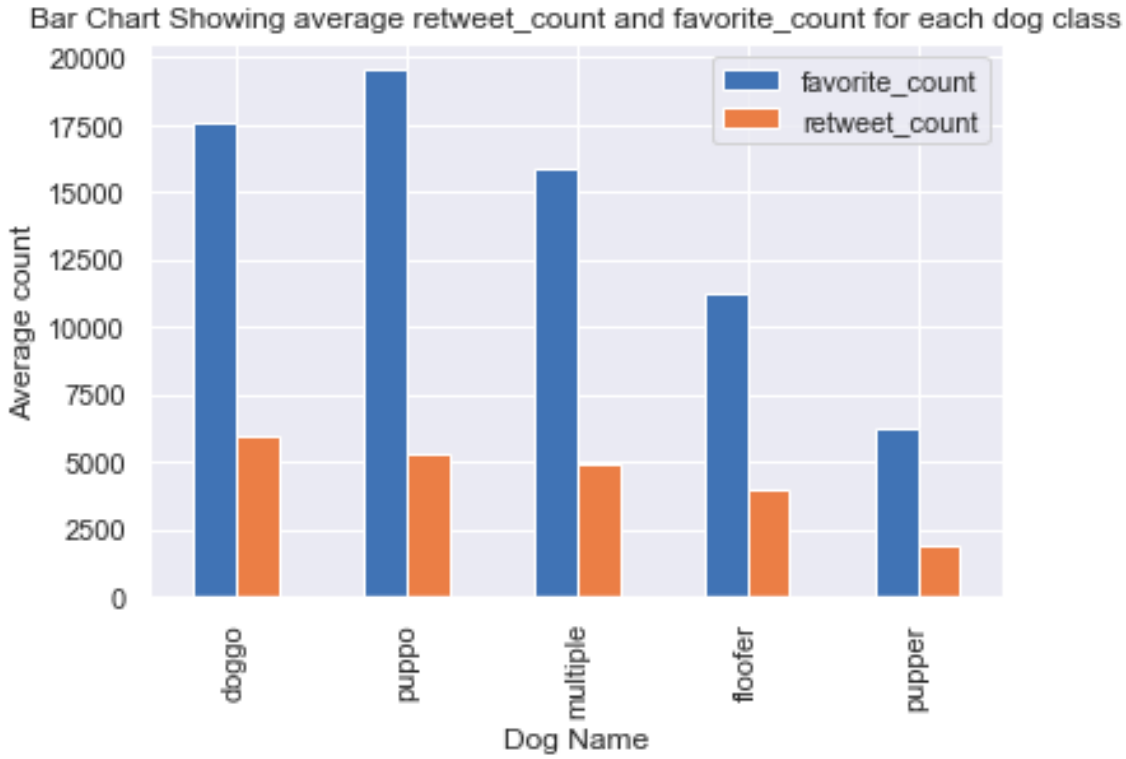
Dog Name	Population
Oliver	10
Cooper	10
Charlie	10
Tucker	9
Penny	9
Lucy	9
Sadie	8
Winston	8
Daisy	7
Toby	7



3. Average retweet_count and favorite_count for each dog class.

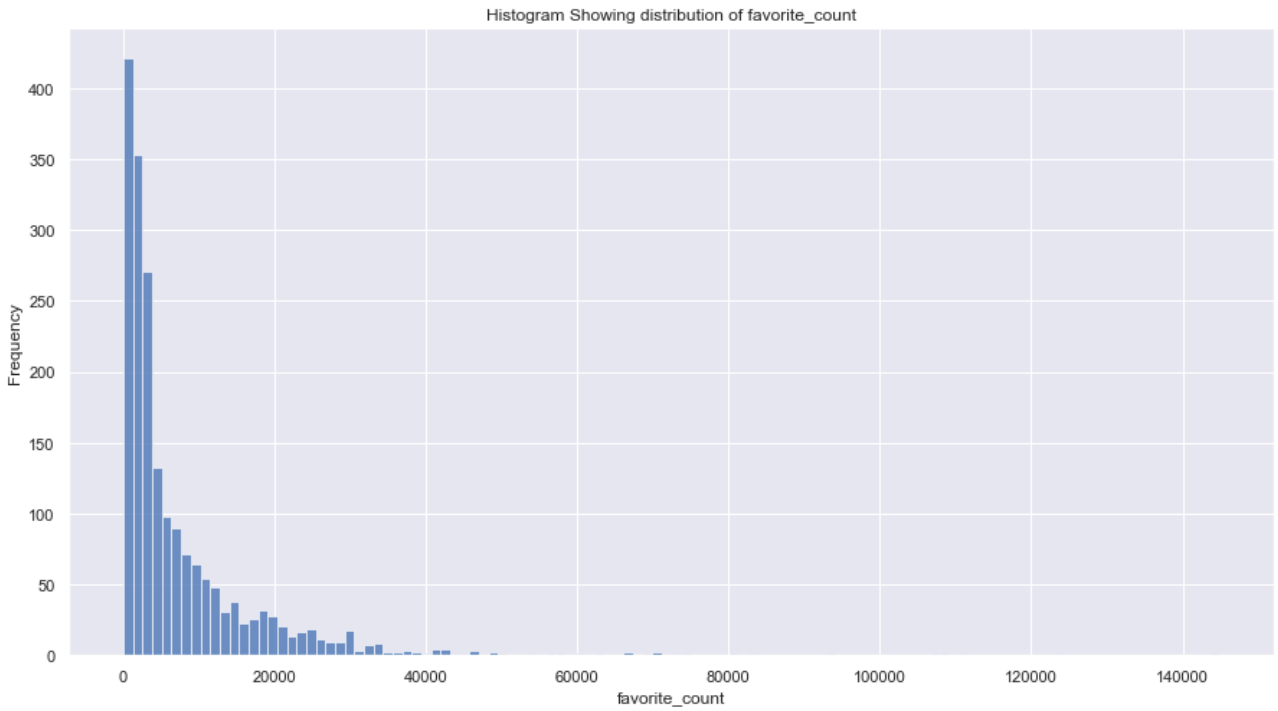
Doggo dogs seem to be the most loved class of dogs. They rank first in retweet_count and second in the favorite_count.

	favorite_count	retweet_count
dog_class		
doggo	17600.403226	5971.967742
puppo	19574.090909	5325.090909
multiple	15847.600000	4873.300000
floofer	11222.857143	3984.714286
pupper	6250.303483	1924.721393



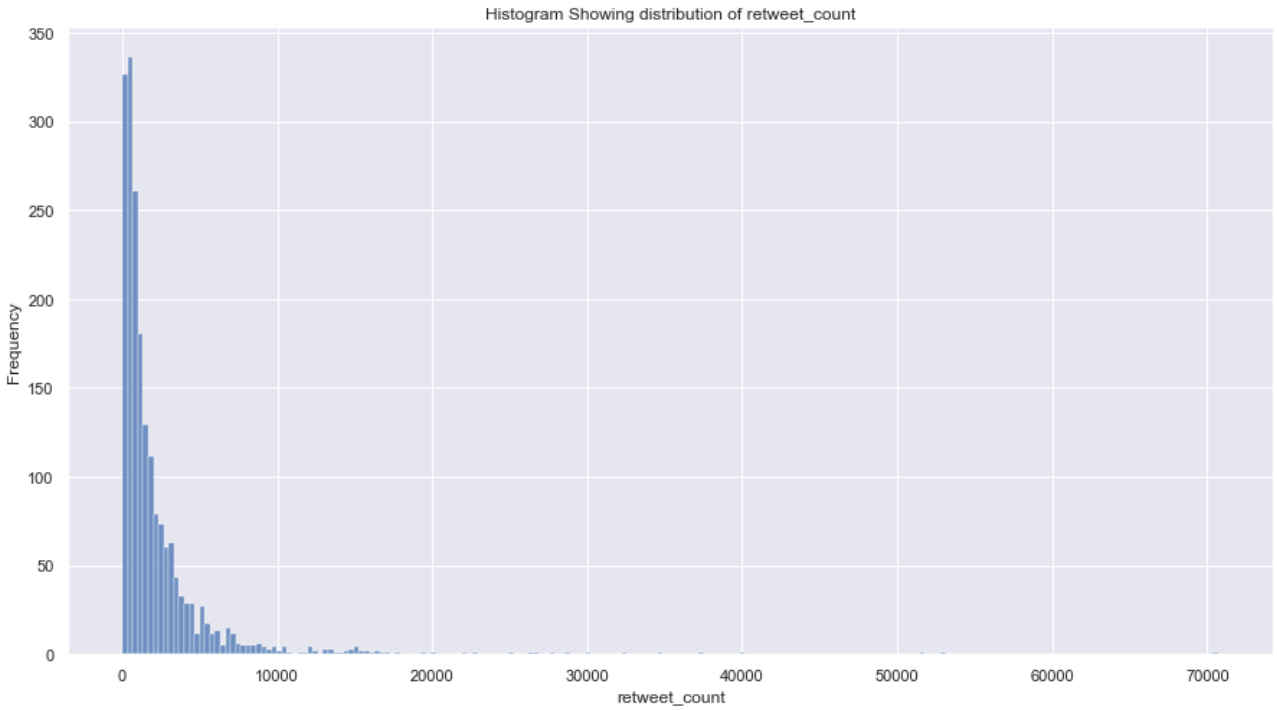
4. Distribution of favorite_count.

The distribution of favorite_count is highly skewed. The chart shows that a few number of dogs had a favorite_count above 20000.



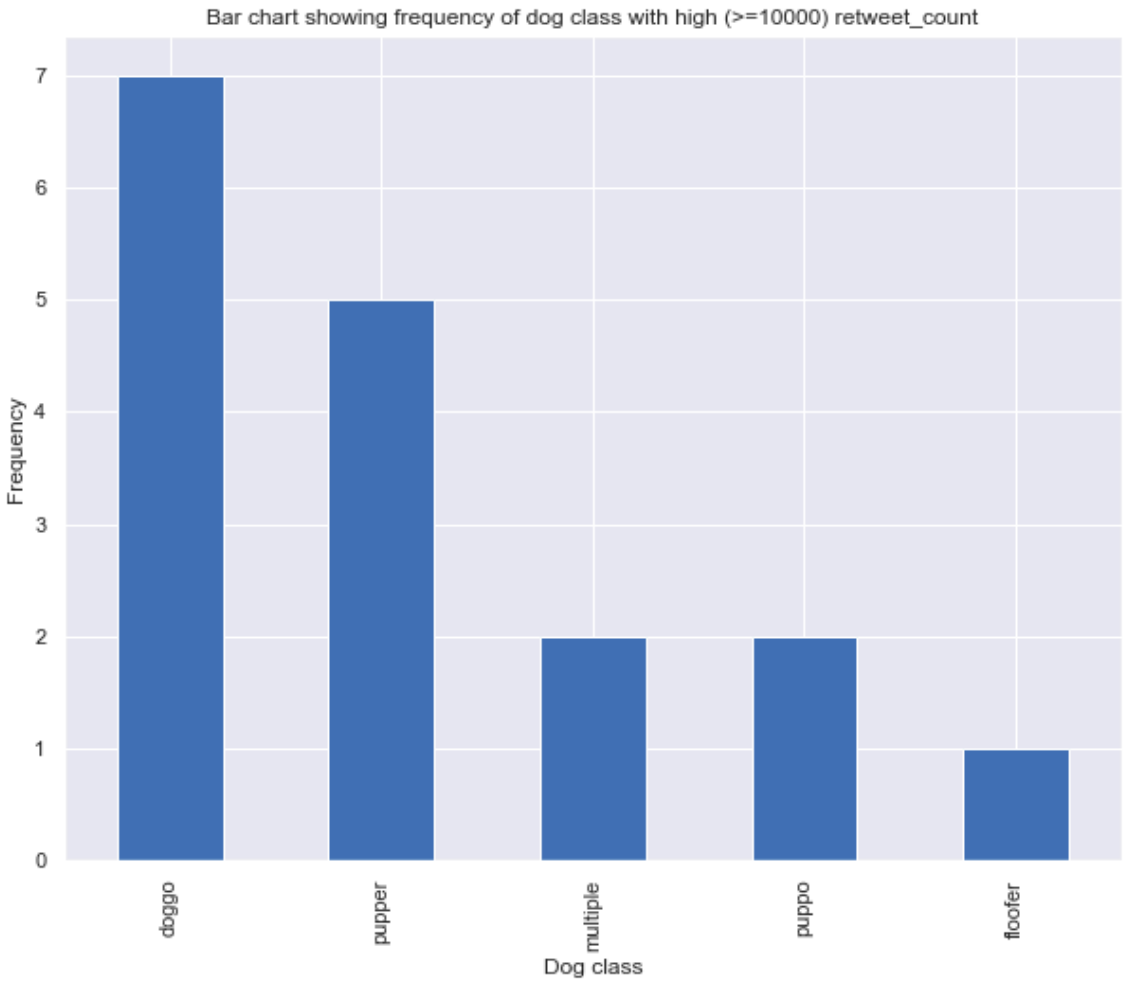
5. Distribution of retweet_count.

The distribution of retweet_count is highly skewed. The chart shows that a few number of dogs had a retweet_count above 10000.



6. Dog class with highest retweet_counts (>10000).

Doggo dogs occupy the largest portion of dogs with retweet_counts greater than 10000.



7. Dog class with highest favorite_counts (>20000)

Doggo and pupper dogs occupy the largest portion of dogs with favorite_counts greater than 20000.