Dog Rates Analysis and Visualization

Introduction

The dataset that is chosen for wrangling (and analysing 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.

The goal is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The data wrangling process has been outlined in another article. This article focuses solely on the analysis and visualization after the data wrangling is complete.

Questions

Generally, when the motive is to analyse a dataset, we either get the questions to be answered from a stakeholder or we have to pose them ourselves. The data wrangling process completely depends on the questions to be answered. This is because only the fields required for the analysis needs to be wrangled. The rest of the fields can be dropped.

Visualizations serve many purposes. They can be used to portray a clear and concise picture of the data to the stakeholder(s). They can also be used to solidify the understanding of the data as a whole. Visualizations, when chosen wisely can give more information that pages together of text might fail to give.

In the case of this analysis, the questions were not provided by the stakeholder. They were posed based on the available data and the data wrangling effort was in an effort in the same direction. The questions posed are in no way exhaustive. They are a small subset of all the questions that can be posed for the given dataset.

Question 1: What is the monthly trend of the followers count?

In order to answer this question, we need to plot the month versus the average favorite count for that month regardless of the year. First, we need to extract the month when the tweets were created. We can extract the year as well in order to find out the time span of the tweets. We observe that there is at least one tweet corresponding to every month. Also, the tweets span from 2015 to 2017.

Next, we prepare the data for visualization. We convert the extracted month field into an ordered categorical field and get the average of the favorite count after grouping the data by month.

Finally, we use this data to plot a line graph complete with appropriate tick labels on both axes, chart title and axes labels as shown below. [Figure 1]

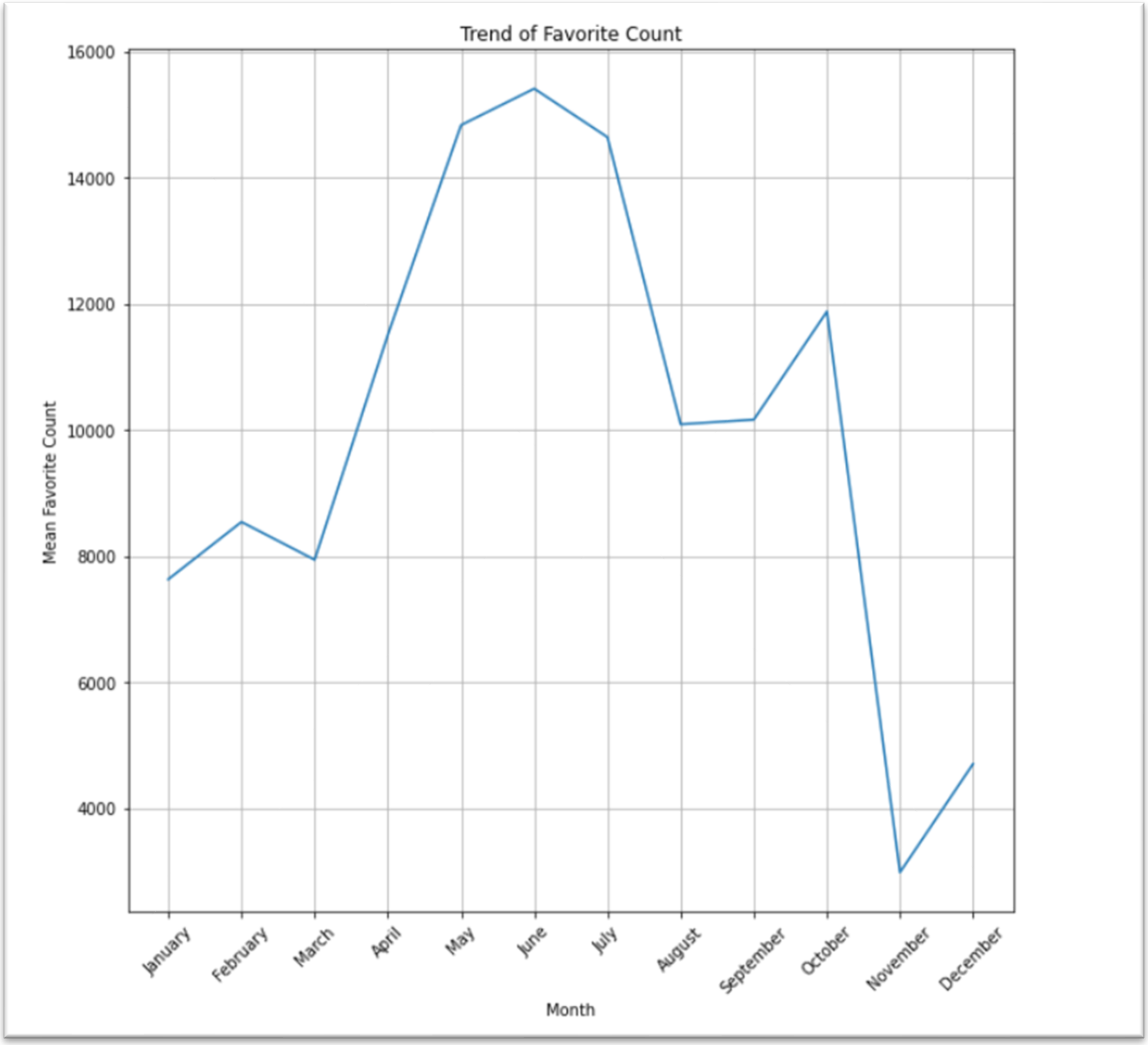

Figure 1

From the above line plot, we can observe that the month of June has the highest number of average favorite count while November has the lowest number. There is however, no consistent trend over the months.

Question 2: What are the details of the tweet that was retweeted the most?

The details of the tweet with the greatest number of retweets is required to be obtained. The five most retweeted tweets can be easily obtained using a pandas operation. Each tweet is identified using a tweet ID that is a unique eighteen-digit number. This number cannot be used to identify the tweet in a graph since it would make the graph crowded. So, we can create a new identifier that follows the format of the alphabet T followed by the next number in the sequence.

The identifier of the five most retweeted tweets and their corresponding retweet count can be visualized using a bar graph as shown below. [Figure 2]

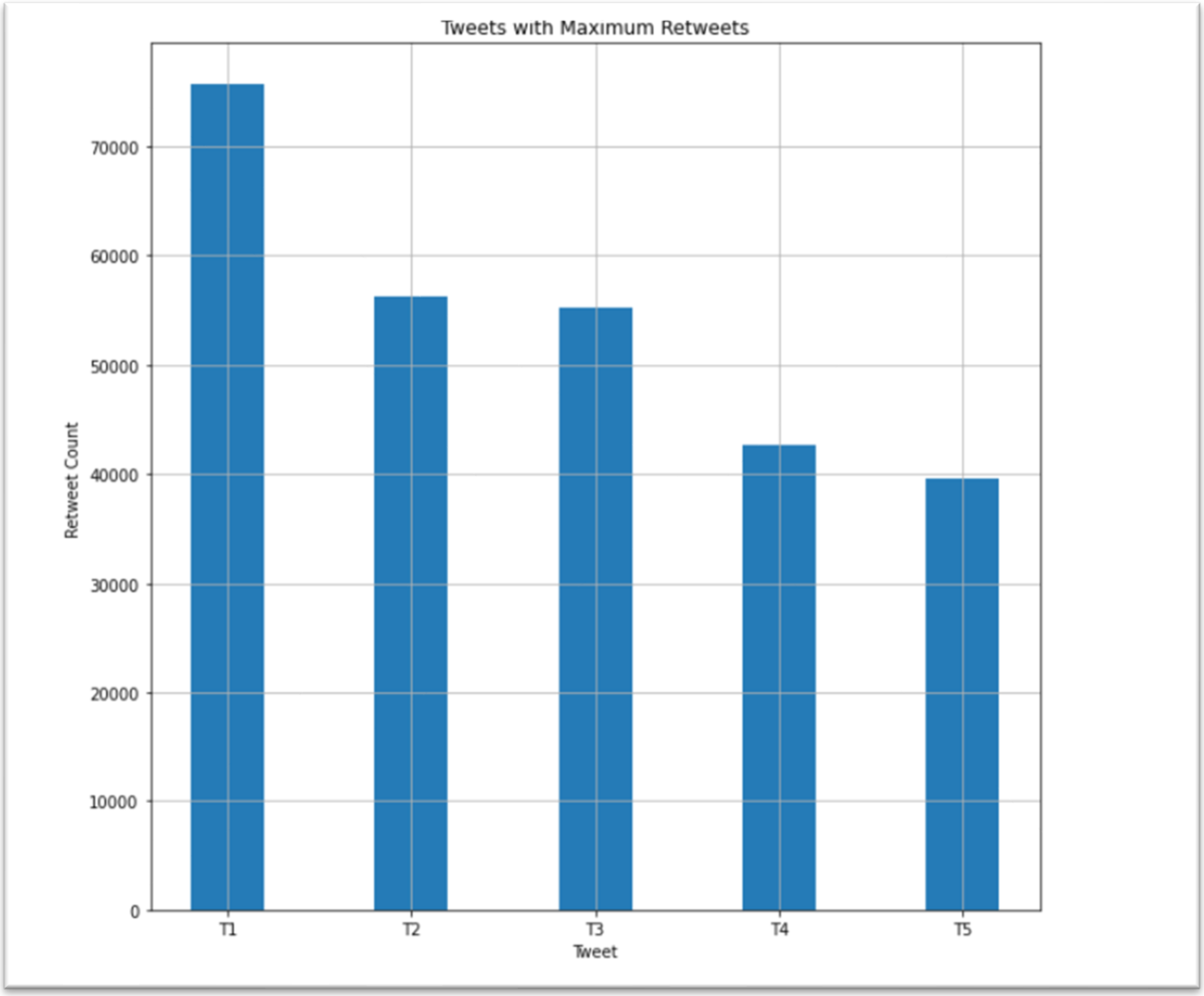

Figure 2

From the result of the operation performed to get the tweet with the most retweets, we get the following details. In the tweet with maximum number of retweets (identified by T1), the dog is rated 13/10. The name is the dog is not available and the stage is 'doggo'. The tweet was created on 18th June, 2016.

Question 3: What is the relationship between rating and number of retweets?

We can use the numerator to represent the rating since the denominator is always 10. The relationship between rating and number of retweets associated with a tweet can be visualized using a scatter plot as shown below. [Figure 3]

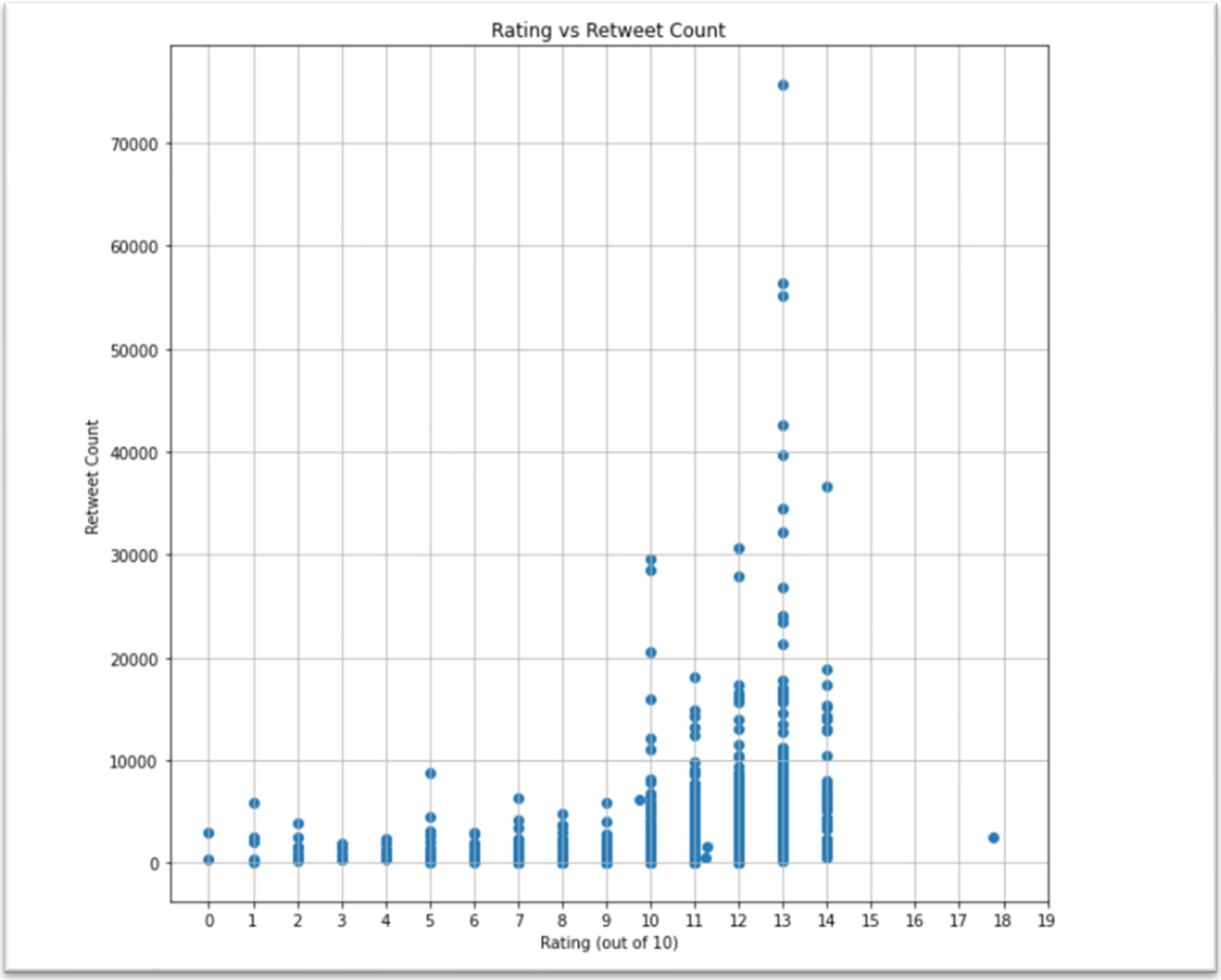

Figure 3

From the scatter plot above, we can make the following inferences:

- The maximum number of retweets for each rating does not demonstrate any strict trends
- The maximum number of retweets were recorded for a dog with a 13/10 rating

Question 4: In how many tweets was the rating given greater than 10?

The denominator of the rating is always 10. The numerator can be greater than, lesser than or equal to 10. We can apply a condition to get the number of tweets that have rating greater than 10. The same can be visualized using a histogram as shown below. [Figure 4]

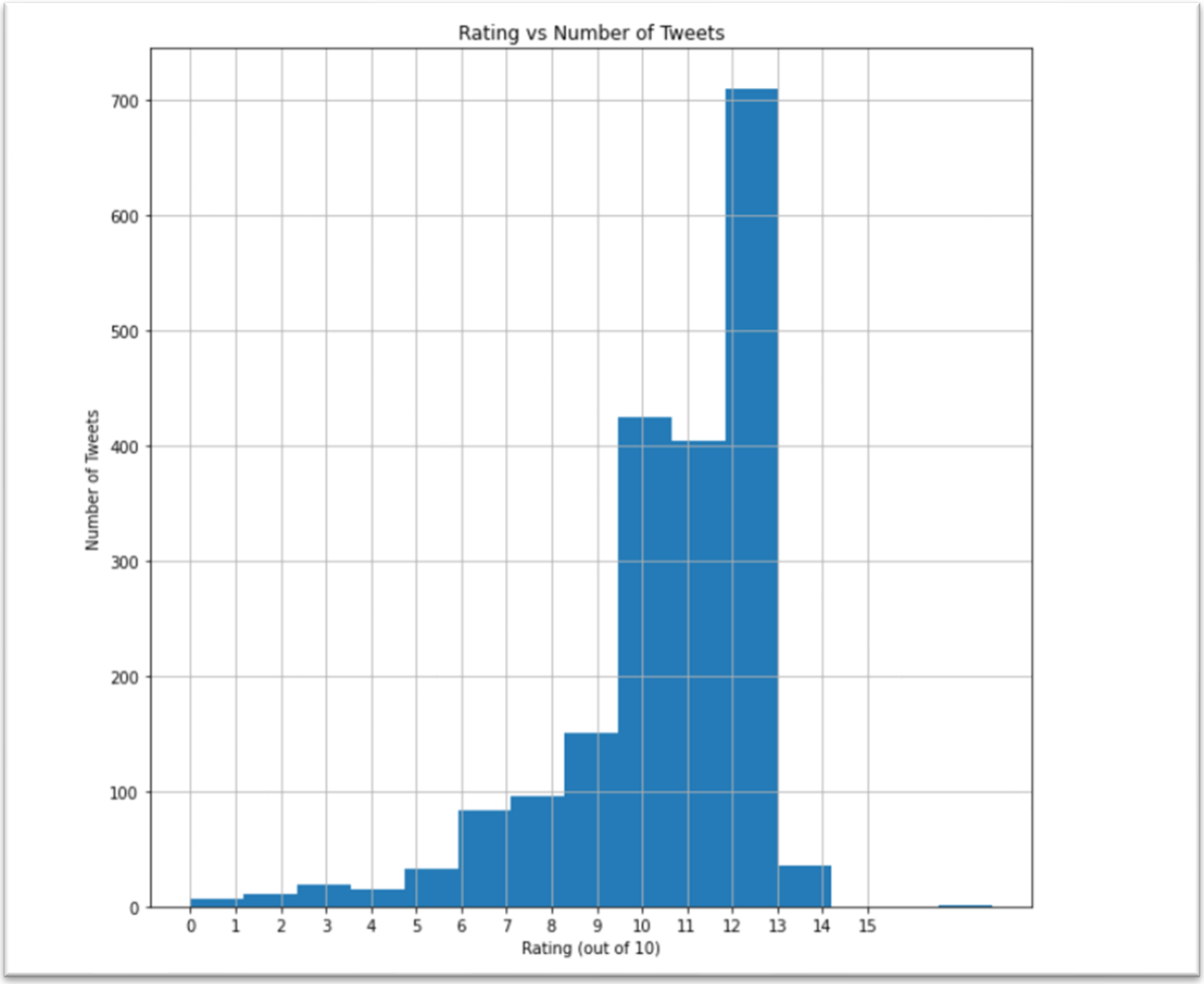

Figure 4

We can observe from the above graph that the majority of the dogs are given a rating above 10, out of 10. This is a unique system of @dog_rates. The graph is heavily left-skewed. From the calculations, 1150 out of the 1987 tweets (approximately 58% of the tweets) have a dog rating above 10.

Question 5: Which dog stage is that occurs most frequently?

There are 4 possible stages that a dog can belong to. They are doggo, floofer, pupper and puppo. It is possible that the stage corresponding to the dog is not available and the value is None. It is also possible that a dog is associated with more than one stages. The stage(s) corresponding to the dog is stored in a list.

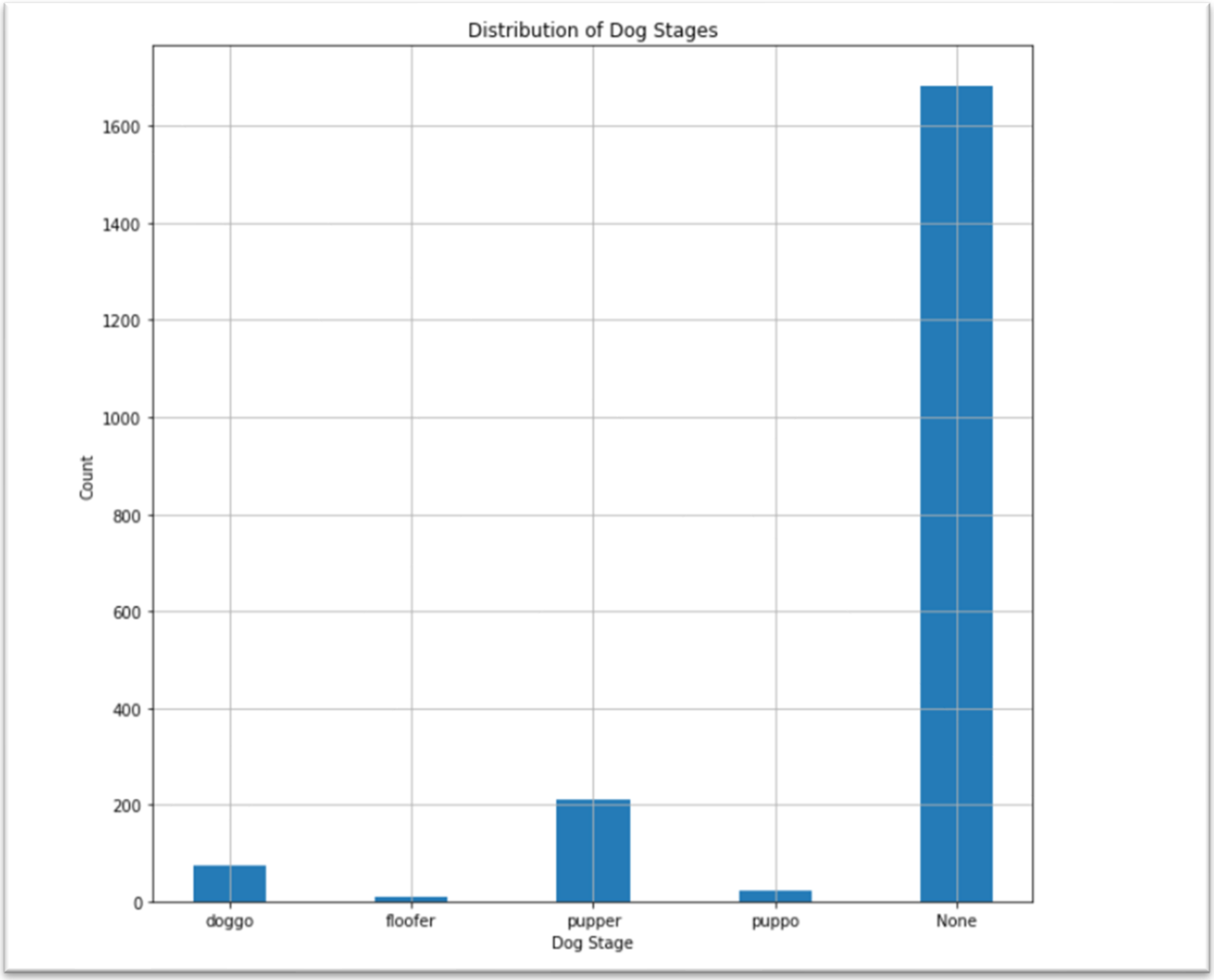

Figure 5

In order to get the count for each of the stages, we loop over the values of the field and update the dictionary with key as the dog stages and value as the count. The results can be visualized effectively using a bar graph as shown above. [Figure 5]

From the plot, we observe that most of the stages corresponding to the dogs are not available. Among the four valid stages, pupper is the most common with 212 occurrences.