We Rate Dogs Act Report

Tachin Volker Felvic Katche

13-06-2022

As part of my Data Analyst Nanodegree program at Udacity, I was tasked with analysing data from the We Rate Dogs twitter page. This process involved gathering data from various sources, assessing and cleaning the data into a master dataset.
Furthermore, I was to generate at least 3 insights from this data set, along with at least 1 visualisation.. Here were my findings.

Questions asked:

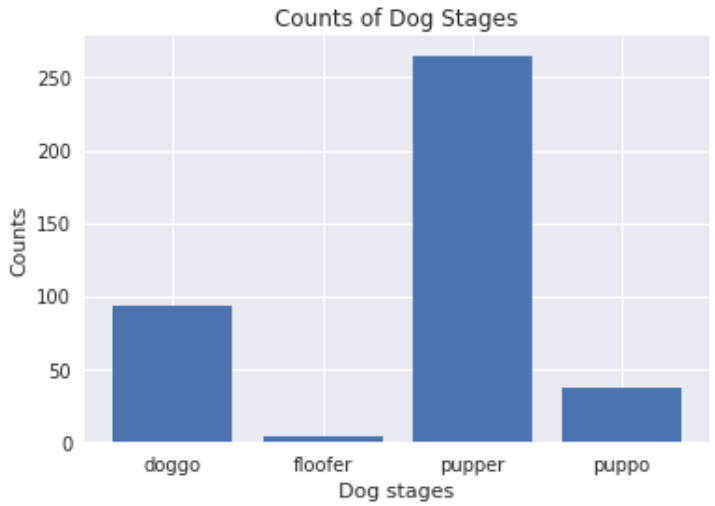
- What is the most common dog stage and the least popular
- How many images are usually associated with the tweets
- Highest and lowest ratings of the dogs
- Were the algorithms able to predict the dog breed of the highest rated dog?

Insights from this data:

1. To answer this question, the a bar chart could be plotted using matplotlib's pyplot as below:

```
plt.title("Counts of Dog Stages")
plt.xlabel("Dog stages")
plt.ylabel("Counts")
plt.bar(x = df.dog_stage.value_counts().index, height = df.dog_stage.value_counts().values)
```

   This generates the following output



It is apparent that the most common dog stage is the "pupper". While the least common is the floofer

2. To find out the number of images usually attached to the tweets. The code below provides us with the answer.

```
df.img_num.value_counts()
```

And this is the output

```
1.0    1780
2.0     198
3.0      66
4.0      31
Name: img_num, dtype: int64
```

From here, we can see that most tweets have just 1 image attached to them.

3. We can also check for the highest and lowest rated tweets as such;

```
df.describe()
```

This gives us the summary statistics of the numerical values in the dataset and from here we can see the max and minimum rated tweets

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	rating_numerator	rating_denominator	img_num
count	2.356000e+03	7.800000e+01	7.800000e+01	2356.000000	2356.0	2075.000000
mean	7.427716e+17	7.455079e+17	2.014171e+16	13.656197	10.0	1.203855
std	6.856705e+16	7.582492e+16	1.252797e+17	45.811761	0.0	0.561875
min	6.660209e+17	6.658147e+17	1.185634e+07	10.000000	10.0	1.000000
25%	6.783989e+17	6.757419e+17	3.086374e+08	10.000000	10.0	1.000000
50%	7.196279e+17	7.038708e+17	4.196984e+09	11.000000	10.0	1.000000
75%	7.993373e+17	8.257804e+17	4.196984e+09	12.000000	10.0	1.000000
max	8.924206e+17	8.862664e+17	8.405479e+17	1776.000000	10.0	4.000000

As you can see from the `rating_numerator` column, the max value is 1776. That is a really good rating.

4. The last question can be answered by querying the row which has the highest rating as shown below:

```
df.query("rating_numerator == 1776")
```

ımerator	rating_denominator	name	...	p1_dog	p2	p2_confidence	p2_dog	p3	p3_confidence	p3_dog	favorite_count	retweet_count	dog_stage
1776	10	Atticus	...	False	sunglasses	0.080822	False	sunglass	0.050776	False	4790.0	2281.0	NaN

By taking a look at the three algorithms, `p1_dog`, `p2_dog`, and `p3_dog`, we can see that none of these algorithms were able to predict the dog breed