act_report

June 6, 2021

In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

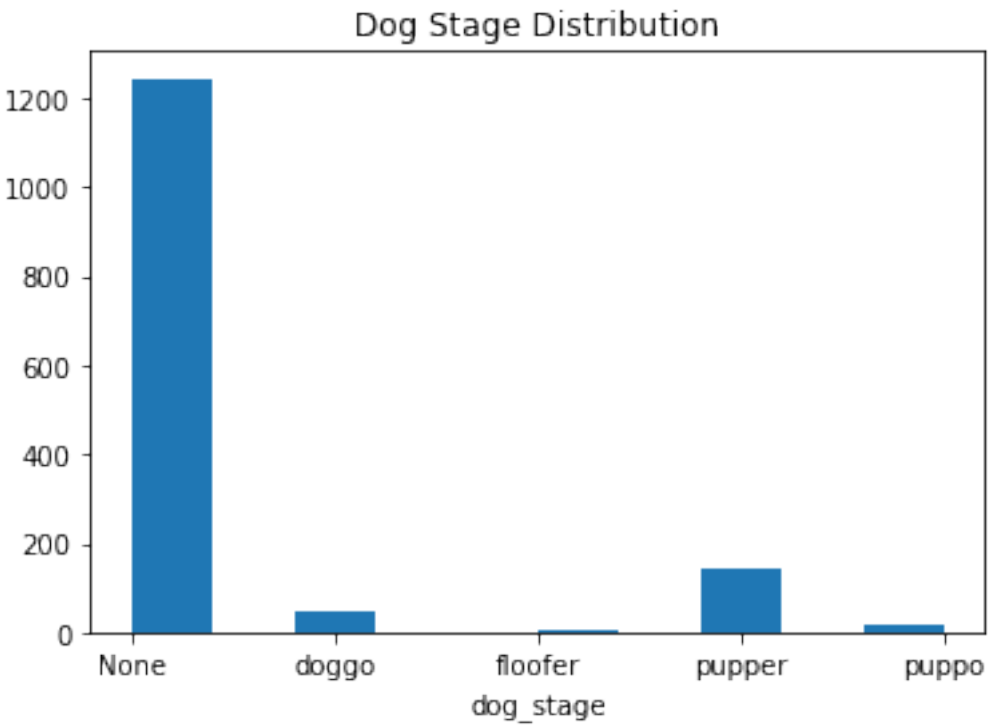
In [3]:
```
df = pd.read_csv('twitter_archive_master.csv')
```

Our dataset in composed by a lot of information about the WeRateDogs trend in twitter. We have information about the tweets ids in which dogs were rated, the rate they were given, the dog's stage, url, text of the tweeet, predicted breed of the dog based on the image and counts of retweets and favorite.

We can easily take a look at some trens of the dataset, such as: which are the most frequent dog stages, and which one of them has the higher retweet count or favorite count; Which is the most common breed of dog posted in this twitter trend; Which prediction model was best in terms of confidence intervals to predict the dog breeds. We can also see if there is any prediction model that is best for a certain type of breed.

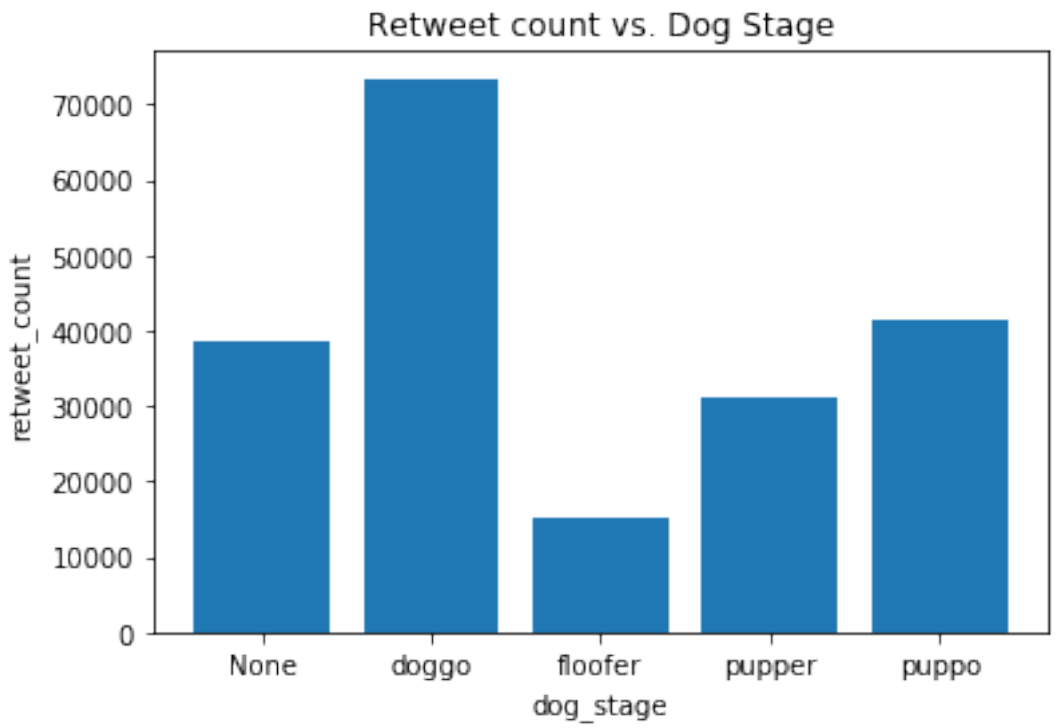
In [4]:
```
plt.hist(df['stage'])
plt.xlabel('dog_stage')
plt.title('Dog Stage Distribution')
```

Out[4]:
```
Text(0.5,1,'Dog Stage Distribution')
```



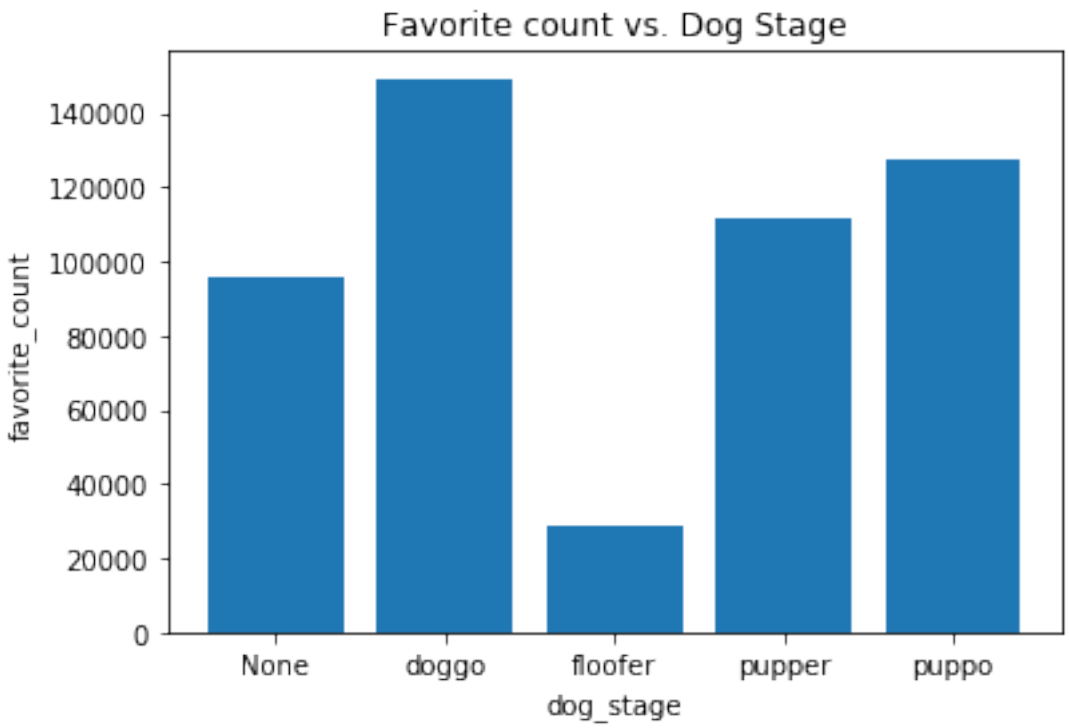
From this plot we can see that to most of our data the stage of the dog obtained from the tweet was 'None', but of those which wer not, the category that appears most is 'pupper', followed by 'doggo'.

```
In [5]: plt.bar(data = df, x='stage', height='retweet_count')
        plt.xlabel('dog_stage')
        plt.ylabel('retweet_count')
        plt.title('Retweet count vs. Dog Stage')

Out[5]: Text(0.5,1,'Retweet count vs. Dog Stage')
```



Here we can see that in terms of retweet_count, users usually retweet more the tweets with the stage 'doggo', followe by 'puppo'.

```
In [6]: plt.bar(data = df, x='stage', height='favorite_count')
        plt.xlabel('dog_stage')
        plt.ylabel('favorite_count')
        plt.title('Favorite count vs. Dog Stage')

Out[6]: Text(0.5,1,'Favorite count vs. Dog Stage')
```



Here we can see that in terms of favorite_count, 'doggo' ins the winning category, followed by 'puppo' and 'pupper'. People seem to favorite more tweets with the 'doggo' dog stage. That happened to the retweets two. Probably, 'doggo' is the most common stage for dogs.

```
In [14]: df['dog_breed'].value_counts().head(10)
```

```
Out[14]: golden_retriever      139
         Labrador_retriever     94
         Pembroke               86
         Chihuahua              80
         pug                    55
         chow                   42
         Samoyed                40
         toy_poodle             38
         Pomeranian             36
         malamute               30
         Name: dog_breed, dtype: int64
```

The most common dog breed in our database is Golden Retriever, followed by Labrador Retriever.

```
In [19]: print(df['best_model'].value_counts())
         print(df[df['dog_breed'] == 'golden_retriever']['best_model'].value_counts())
         print(df[df['dog_breed'] == 'Labrador_retriever']['best_model'].value_counts())
         print(df[df['dog_breed'] == 'Pembroke']['best_model'].value_counts())
```

```
p1    1454
Name: best_model, dtype: int64
p1    139
Name: best_model, dtype: int64
p1    94
Name: best_model, dtype: int64
p1    86
Name: best_model, dtype: int64
```

Apparently, the p1 model is the best one to predict dog breeds in terms of confidence interval. I was chosen as the best in all of the cases. I was trying to see if there was a different best model for each breed, but apparently the p1 wins every time.