Act Report

Project: Wrangle and Analyze "WeRateDogs" Data

The aim of this report is to present the results of analyzing the clean data previously obtained in the data wrangling stage of this project. The dataset provides information from the Twitter account **WeRateDogs**, which rates people's dogs with humorous comments about them.

In the following, I present some of the questions posed to gain insights about the data. In order to answer them, I made use of the visualizations functionalities of the Pandas and Matplotlib libraries.

The clean data frame **df_twitter_master** consists of 22 columns and 1948 non-null value entries. Below I describe the variables that each column represents:

- tweet_id (int): Tweet ID
- timestamp (datetime64): Timestamp of tweet
- source (string): Twitter source
- text (string): Text of tweet
- expanded_urls (string): URL of tweet
- rating_numerator (int): Numerator of rating for this dog
- rating_denominator (int): Denominator of rating for this dog (always 10)
- name (string): Name of dog
- dog_stage (string): corresponding dog stage (None, doggo, floofer, pupper or puppo)
- retweet_count (int): number of retweets this tweet has
- favorite_count (int): number of favorites (likes) this tweet has
- jpg_url (string): URL of dog image
- img_num (float): most confident prediction for this image
- p1 (string): first prediction of classification algorithm
- p1_conf (float): confidence of first prediction
- p1_dog (string): whether image actually represents a dog or not according to first prediction
- p2 (string): second prediction of classification algorithm
- p2_conf (float): confidence of second prediction
- p2_dog (string): whether image actually represents a dog or not according to second prediction
- p3 (string): third prediction of classification algorithm
- p3_conf (float): confidence of third prediction
- p3_dog (string): whether image actually represents a dog or not according to third prediction

Q1: What are the 15 most common dog breeds in the dataset?

To answer this I investigated the values of the **p1** column (corresponding to the first prediction of the classification algorithm), which seems to be the most reliable in

general. I plotted the bar chart shown in Figure 1, where the fifteen dog breeds with the most counts in the dataset are presented.

As it can be seen, the most common dog breed predicted by the neural network in this dataset is by far the Golden Retriever.

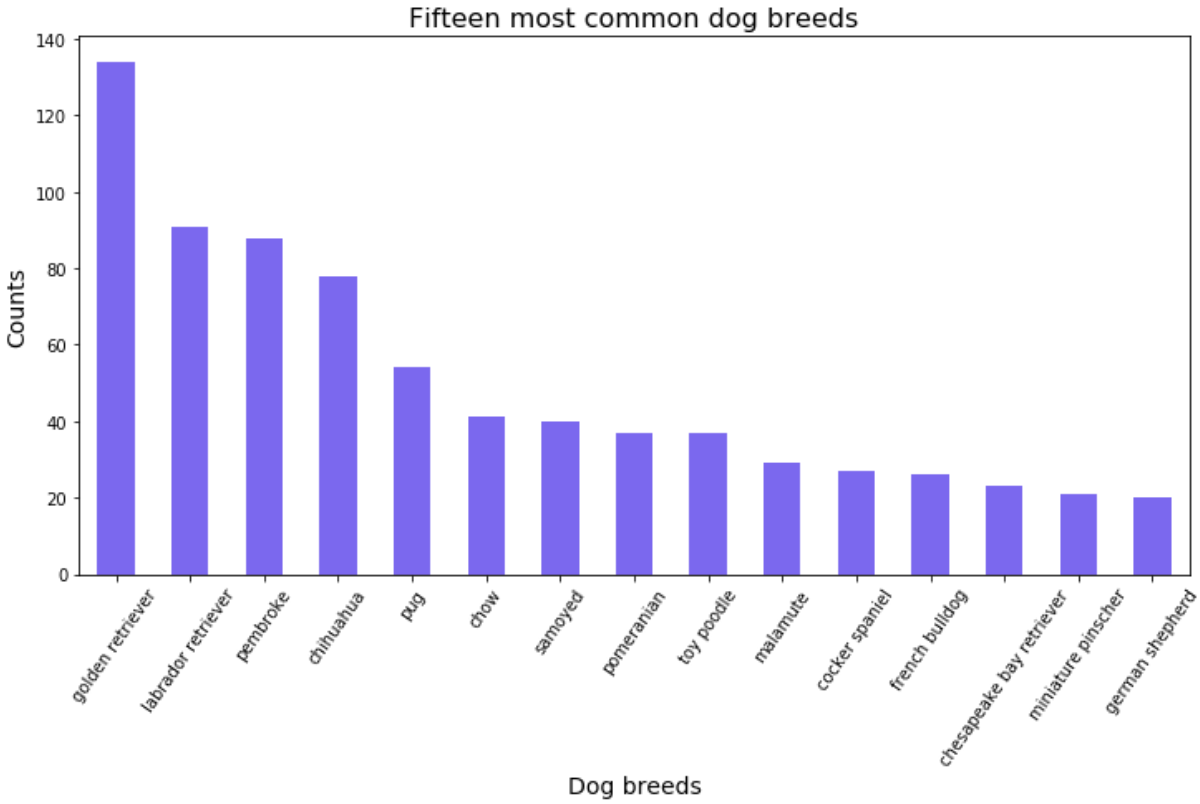


Figure 1. Bar plot with 15 most common dog breeds

Q2: Is there a correlation between Twitter favorites (likes) and retweets?

In order to investigate this, I plotted the two variables **favorite_count** and **retweet_count** together in a scatter plot, which is shown in Figure 2. Then, I calculated the Pearson's correlation coefficient.

The correlation coefficient I obtained (**r = 0.9134**) is consistent with what is observed in the graph. There seems to be a correlation between faving (liking) a tweet and retweeting it.

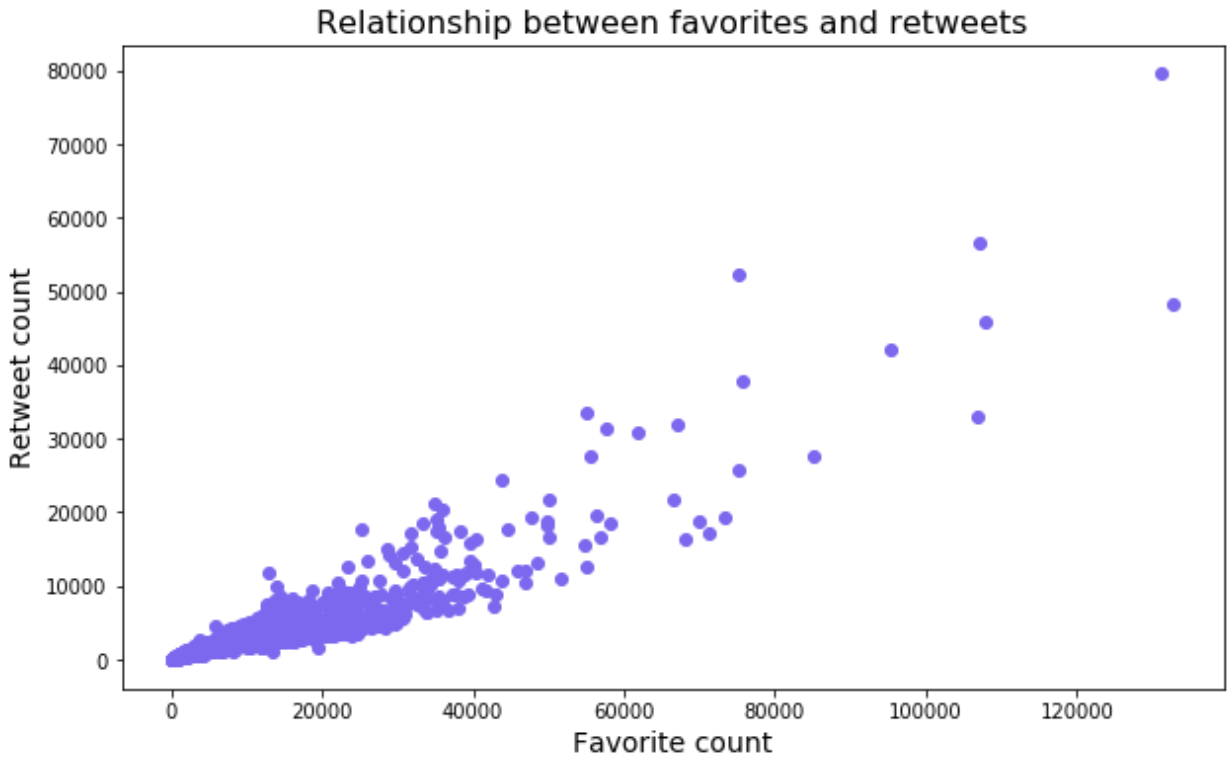


Figure 2. Scatter plot between favorite count and retweet count

Q3: What are the 15 most common dog names?

In order to answer this, I plotted in a bar chart the 15 dog names that appear most frequently in the dataset, which is shown in Figure 3. The corresponding variable in the data frame is the **name** column.

We can see that the most common dog name in this dataset is "Charlie".

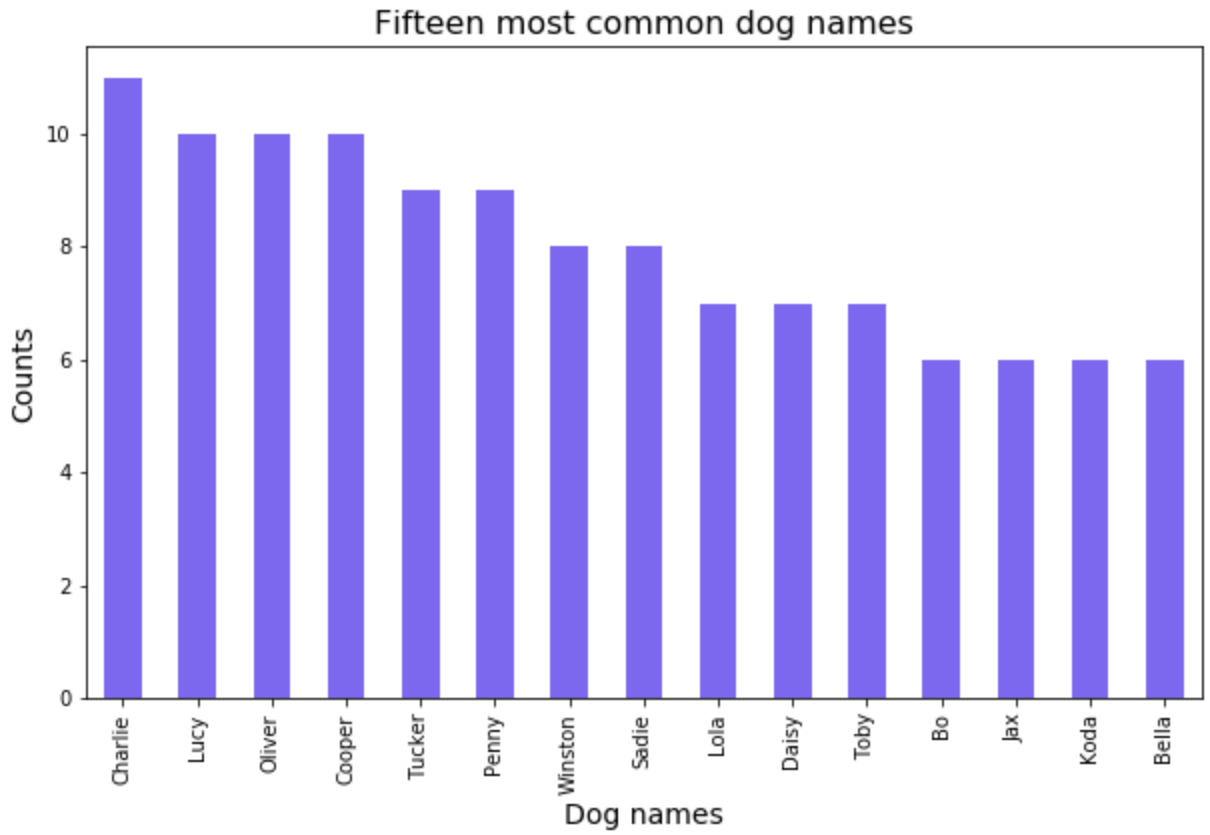
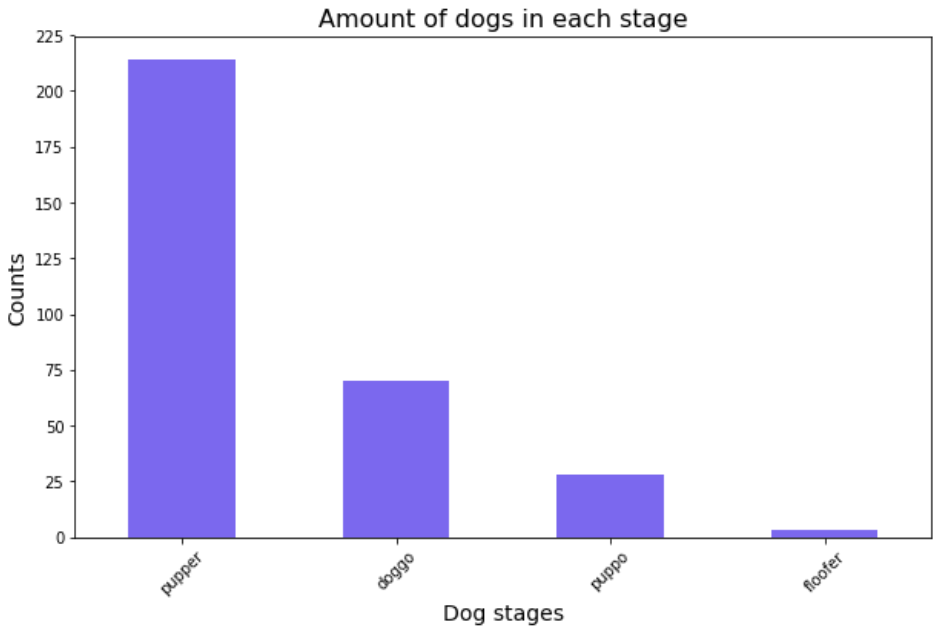


Figure 3. Bar plot of fifteen most common dog names

Q4: How is the distribution of dog stages?

Finally, I plotted in a bar chart the number of dogs corresponding to each stage: **doggo, floofer, pupper and puppo**. This is shown in Figure 4.

As we can observe, the most common dog stage by far is **pupper**, while the least common is **floofer**.


Amount of dogs in each stage
Counts
225
200
175
150
125
100
75
50
25
0
pupper
doggo
puppo
floofer
Dog stages