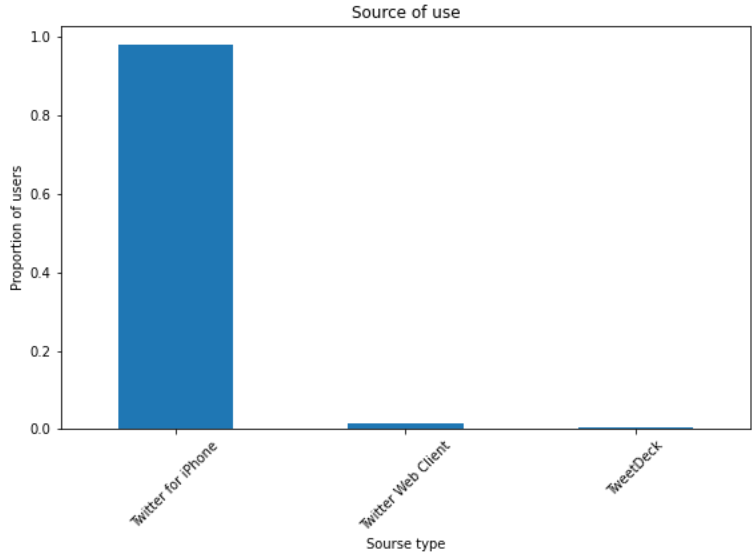
Act Report

Introduction

Real-world data rarely come clean. Using Python and its libraries, we used data wrangling skills to pull real-world data from Twitter, clean it, and did some analysis. The dataset that we wrangled (and also analyzed and visualized) 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.

Below we answered to the following questions:

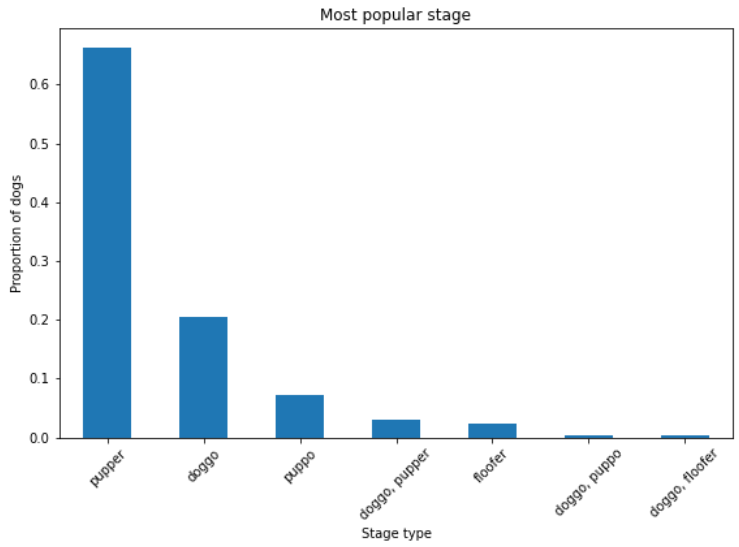
.Question 1: Which kind of source is being used the most?



There is a variable in the dataset named “source” which shows which kind of source people use for accessing Twitter. We used the value_counts function to calculate the proportion of each type of source and then used a bar chart to visualize the results. From the graph, we can see that almost 98% of people use Twitter for iPhone to access Twitter. The share of the other two sources is relatively too small.

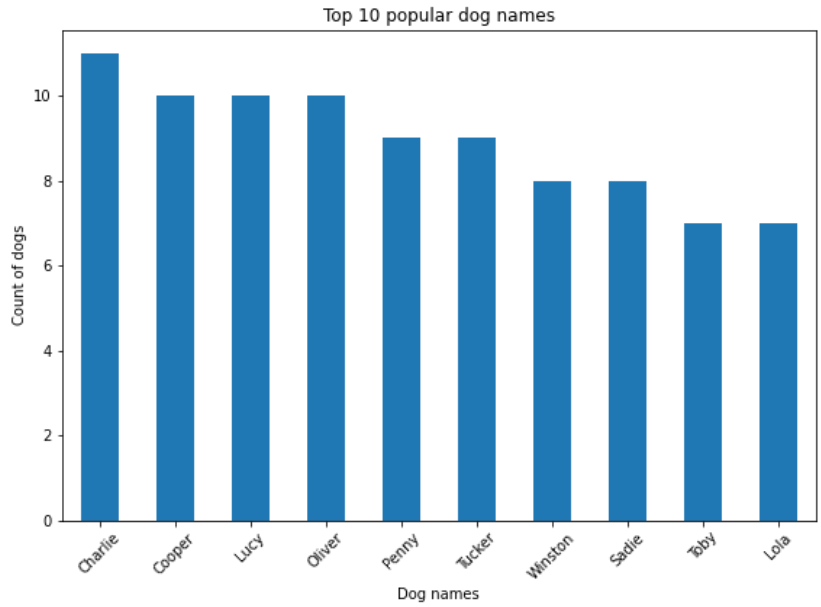
Question 2: Which is the most popular stage for dogs?

Dogs have different stages and this is characterized by the dog_type variable in the dataset. We used the value_counts function to calculate the proportion of each type of source and then used a bar chart to visualize the results. Our analysis showed that the most popular stage is the upper (more than 66% of all the dogs in the dataset).



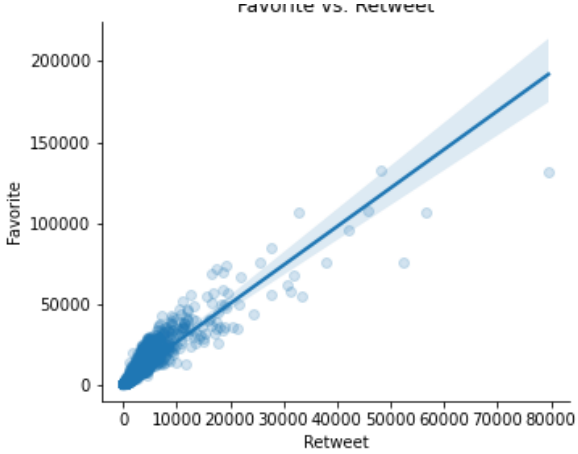
Question 3: What are the most popular 10 names for dogs?

There were more than 900 names for the dogs in the dataset among which the most popular was Charlie.



Question 4: Is there a correlation between Favorite and Retweet counts

We used a scatter plot to show the relationship between favorites and retweets. We used seaborn library and lmplot() function to plot this visualization. From the figure below it can be seen that there is a strong positive correlation between those two variables. For about every 4 favorites there is 1 retweet. We can also notice some outliers here - the most popular tweet has about 130000 favorites and 80000 retweets.


Favorite vs. Retweet
Favorite
200000
150000
100000
50000
0
0
10000
20000
30000
40000
50000
60000
70000
80000
Retweet