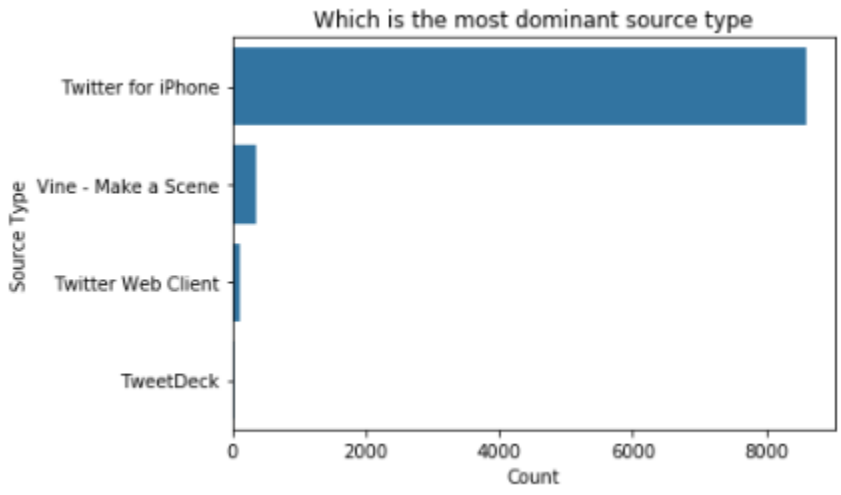
Data Visualization insights

Project Motivation:

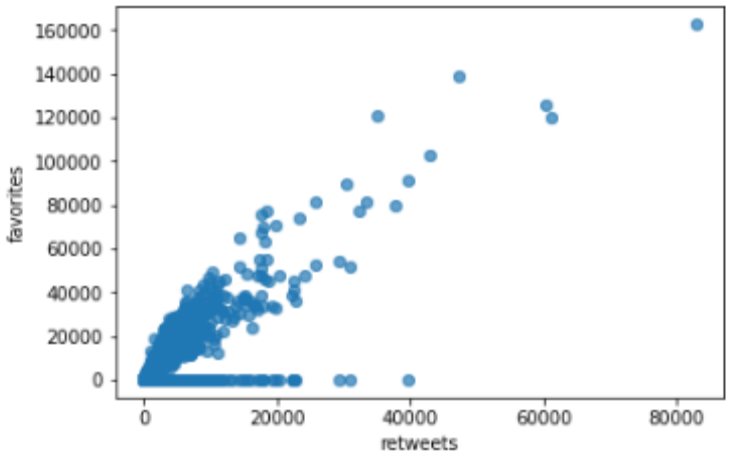
My goal in this project is to wrangle (gather, assess and clean) WeRateDogs Twitter data to create interesting and trustworthy data visualizations. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. I have also included additional gathering of data through twitter APIs to provide better insights and visualizations. The dataset provides information regarding the dog's breed, stage, name, image etc.

I have done the necessary cleaning of data by fixing the issues with the quality of data and its tidiness, for example: Dropping of redundant columns, correcting data types, renaming columns for better user readability etc. Below, I have listed down few of the interesting insights that I have seen while exploring the dataset.

1. The below graph shows the most dominant source type i.e. from where most tweets have been seen. We can easily identify that "Twitter for iPhone" dominates the others by a large margin.



2. Here, I have checked for any correlation between the retweets and favorites count. Both seem to have a positive correlation i.e tweets having for retweet counts also have more favorite counts and tweets having less retweet counts also have less favorite counts.



3. The ratings seen in WeRateDogs page almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc.

Though it's said that the denominator rating is always 10, we can see that the denominator has many other values as well such as below:

```
Out[30]: 12     543
         11     455
         10     453
         13     332
         9      155
         8      102
         7       53
         14      47
         5       35
         6       32
         3       19
         4       16
         2        9
         1        6
         0        2
         75       2
         60       1
         24       1
         26       1
         44       1
         50       1
         165      1
         80       1
         84       1
         88       1
         144      1
```

Here I haven't performed any cleaning operations on the ratings.

4. The breed of the dogs also can be identified based on the prediction and the confidence level marked against it.

5. Many more interesting insights about the dog's likes, dislikes and hobbies etc can be extracted from the text column available in the dataset.