Act Report for WeRateDogs Tweet Archive

Date

Farnaz Zarian

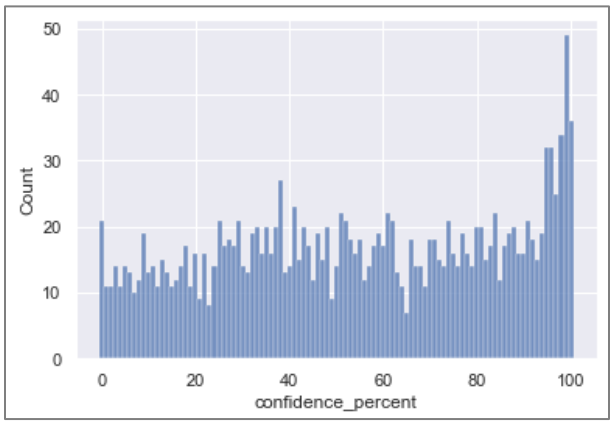
January 8, 2021

This document includes a summary of the Exploratory Data Analysis (EDA) that I performed on 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. These ratings almost always have a denominator of 10 and a numerator of greater than 10. Why? Because "they're good dogs Brent." WeRateDogs has over 8.9 million followers and has received international media coverage.

Insights

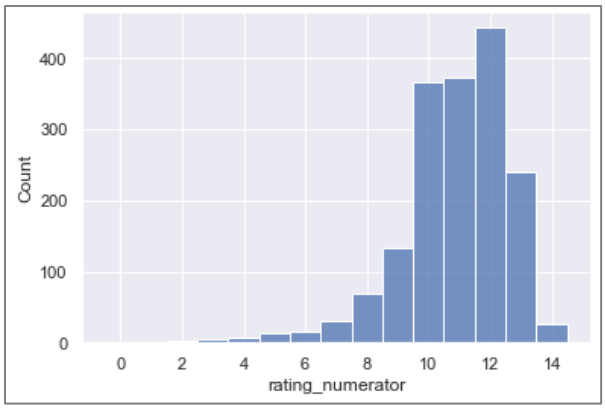
Using Python libraries, I gathered data from a variety of sources and in a variety of formats, then performed data wrangling on these datasets by assessing their quality and tidiness, then cleaned them and made them ready for analysis.

Based on the univariate distribution of all variables in the dataset, we can see that confidence and rating_numerator are skewed to the left. This observation indicates that a significant fraction of high confidence data is above 50%, and the majority of dog ratings are in the 10-12 rating range. Similarly, favorite_count and retweet_count histograms clearly show a big skew to the right indicating the majority of the data ~75% fall below 10,000 favorite_count and ~75% fall below 2,900 retweet_count.



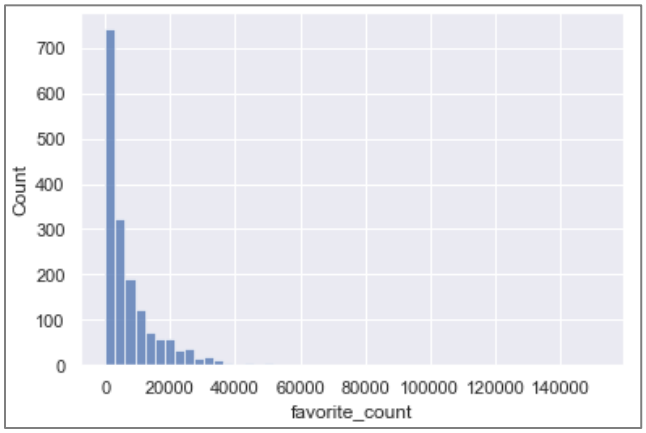

```
1 WeRateDogs_twitter_arch_clean['confidence_percent'].describe()
```

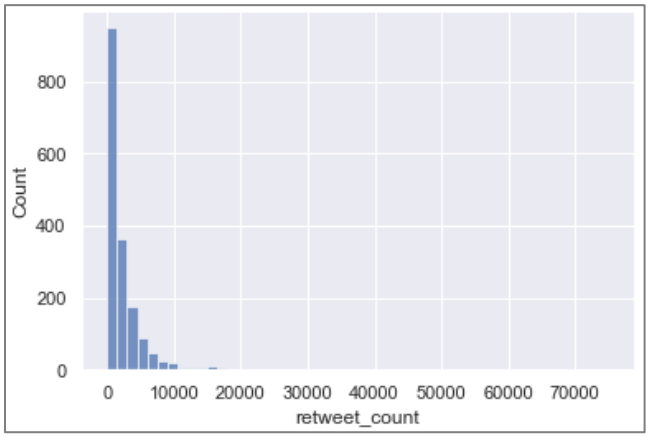
```
count    1725.000000
mean       54.749715
std        29.853321
min         0.001003
25%        29.996600
50%        54.740100
75%        81.795300
max        99.995600
Name: confidence_percent, dtype: float64
```




```
1 WeRateDogs_twitter_arch_clean['rating_numerator'].describe()
```

```
count    1725.000000
mean       10.863768
std         1.771943
min         0.000000
25%        10.000000
50%        11.000000
75%        12.000000
max        14.000000
Name: rating_numerator, dtype: float64
```

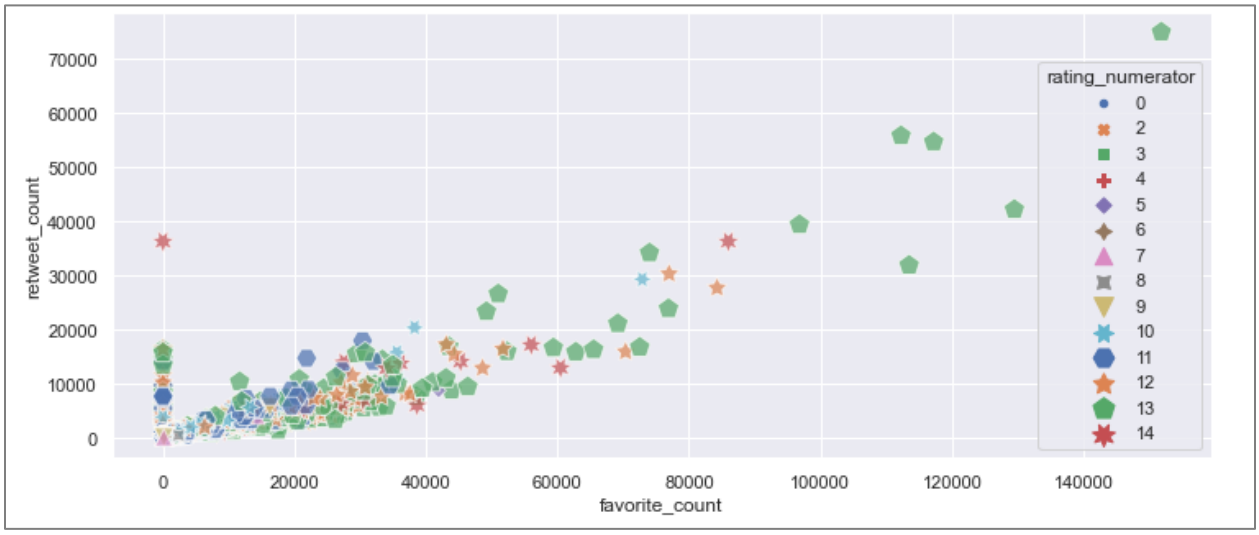




```
[WeRateDogs_twitter_arch_clean['favorite_count'].describe(),
 WeRateDogs_twitter_arch_clean['retweet_count'].describe()]
```

```
[count     1725.000000
 mean      8112.731594
 std      12227.907903
 min          0.000000
 25%       1638.000000
 50%       3685.000000
 75%      10021.000000
 max     151887.000000
 Name: favorite_count, dtype: float64,
 count     1725.000000
 mean      2551.209855
 std       4508.617270
 min         10.000000
 25%        553.000000
 50%       1257.000000
 75%       2903.000000
 max      75036.000000
 Name: retweet_count, dtype: float64]
```

The number of retweets has a strong positive correlation with the number of favorites a tweet receives. Both of these variables are also positively correlated with the rating that the dogs have received (although not as strong of a correlation).



Based on the data available, *pupper* is the most popular dog stage, followed by *doggo*, *puppo* and *floofer*. It's worth noting that since the majority of the dog stage data were missing, (1457 none values), we cannot infer that this distribution is statistically conclusive.