Act Report For WeRateDogs Twitter Archive

David Onwachukwu

DAND Project 2

May 2022

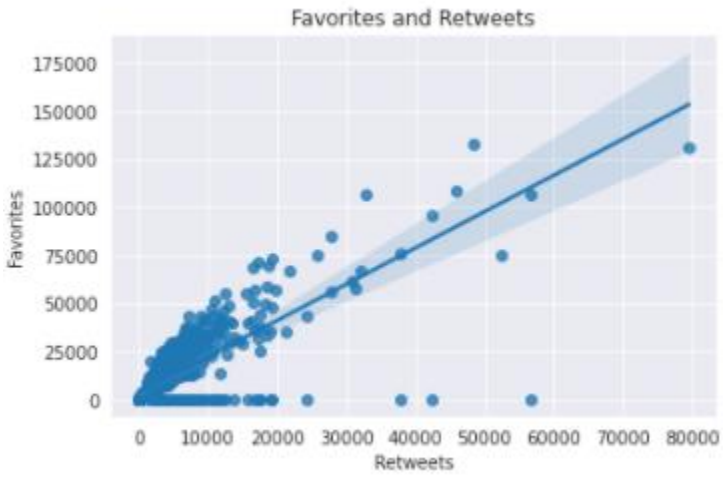
An interesting fact about the dog rating system is that the score can be granted above 10 out 10, i.e. it can be 15 out of 10 or 1000 out of 10. Why dogs are given such outrange number is because "they're good dogs Brent" as it is stated in the article: WeRateDogs

The popularity of the page led to this project and this report documents my analysis and visualizations.

I started my analysis by checking information of the data using the .info function, this pushed me to want to check for the relationship among some columns.

Starting with the Retweets and Favorites columns, I checked for the overall correlation of the whole data using the .corr() function and I saw that there was a strong positive correlation of 0.7.

I plotted a regplot using seaborn and matplotlib to confirm.



Next I decided to check the most retweeted and most liked pictures and view the images using the links in the jpg_url column with the skimage library.

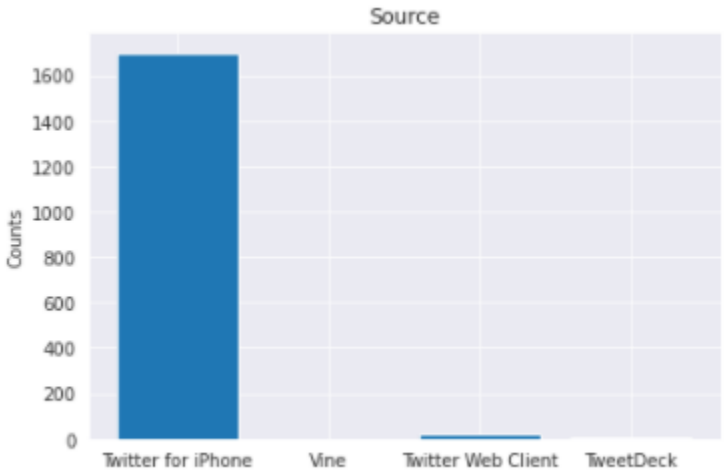
The most liked picture had 132,810 likes, below you can see the code and the image of the dog retrieved with it.

```python
from skimage import io
import matplotlib.pyplot as plt

plt.rcParams["figure.figsize"] = [7.50, 3.50]
plt.rcParams["figure.autolayout"] = True

f = "https://pbs.twimg.com/media/C2tugXLXgAArJO4.jpg"
a = io.imread(f)

plt.imshow(a)
plt.axis('off')

plt.show()
```

The most retweeted picture had 79,515 retweets and this picture was also the second most liked one.

The most retweeted picture is a picture of a Labrador retriever swimming, no surprise with how many retweets and likes it ended up getting.

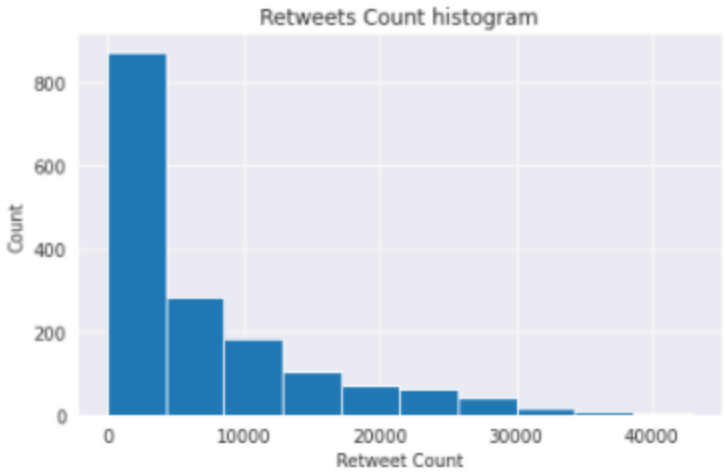
Of the four sources; Twitter for iPhone, Vine, Twitter web client and TweetDeck, I wanted to know the most used source and visualize with a bar graph.

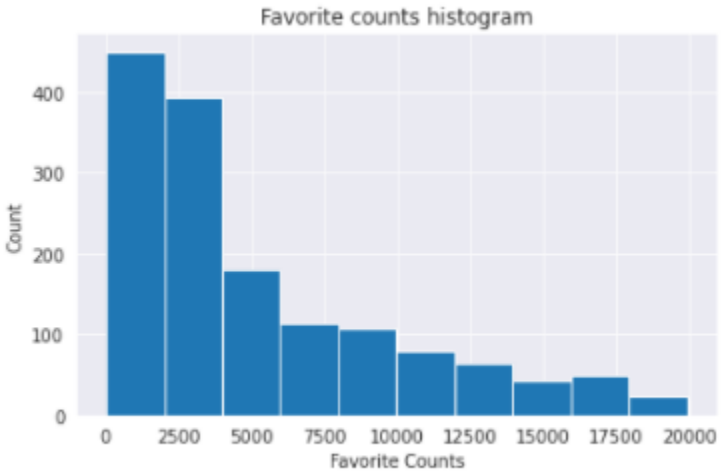


Twitter for iPhone was the most used source among all the tweets in the WeRateDogs archive.

Finally, I decided to check if the retweet and favorite counts of the data were normally distributed and to decide on the best measure of spread to report for each data.

Both Retweet and Favorite counts were right skewed, as seen below.





Both counts are right skewed, therefore the mean cannot be reported as the best measure of spread for these two. Instead the median would be a better measure of spread to report for both retweets and favorites.