Report: WeRateDogs

—

Giulio Tommasino

Introduction:

The WeRateDogs Twitter archive contains basic tweet data for l 2400 of their tweets.

WeRateDogs is a twitter account that rates people's dogs with some comments about the dog. This project works through the gathering of data (wich consisted of three files) , assessing, cleaning of the data and finally make visualization and observation from the analysis provided. In this report I will show all the visualizations that were made during the project and the results provided.

Gather data:

In order to analyze the archive, three files were gathered:

-The WeRateDogs Twitter archive: (*twitter_archive_enhanced.csv)* t.

-The tweet image predictions: (*image_predictions.tsv*)

-Additional Data via the Twitter API

Asses data:

After gathering those files, they were loaded into separate Pandas dataframes to be assessed.
Each file was evaluated both visually and programmatically, and the following issues were observed during the process:

Quality issues:

1. Timestamp should be in datetime format (The WeRateDogs Twitter archive)
2. Drop columns with too many missing values (The WeRateDogs Twitter archive)
3.Change tweet_id to type str in order to merge with the other 2 tables (The WeRateDogs Twitter archive)
4. Keep original ratings (no retweets) that have images (The WeRateDogs Twitter archive)
5. Errors in name column (a, an,...) in (The WeRateDogs Twitter archive)
6. Keep original tweets (.json file)
7. drop columns (tweet image predictions)
8. Drop jpg_url duplicated (tweet image predictions)

Tidiness issues:

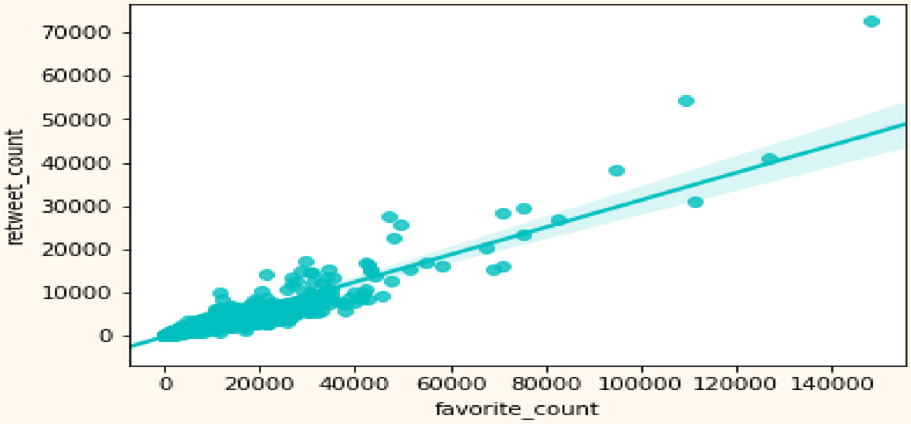
1.Melt the doggo, floofer, pupper and puppo columns (The WeRateDogs Twitter archive)

2. merge all the data frames in one Datase

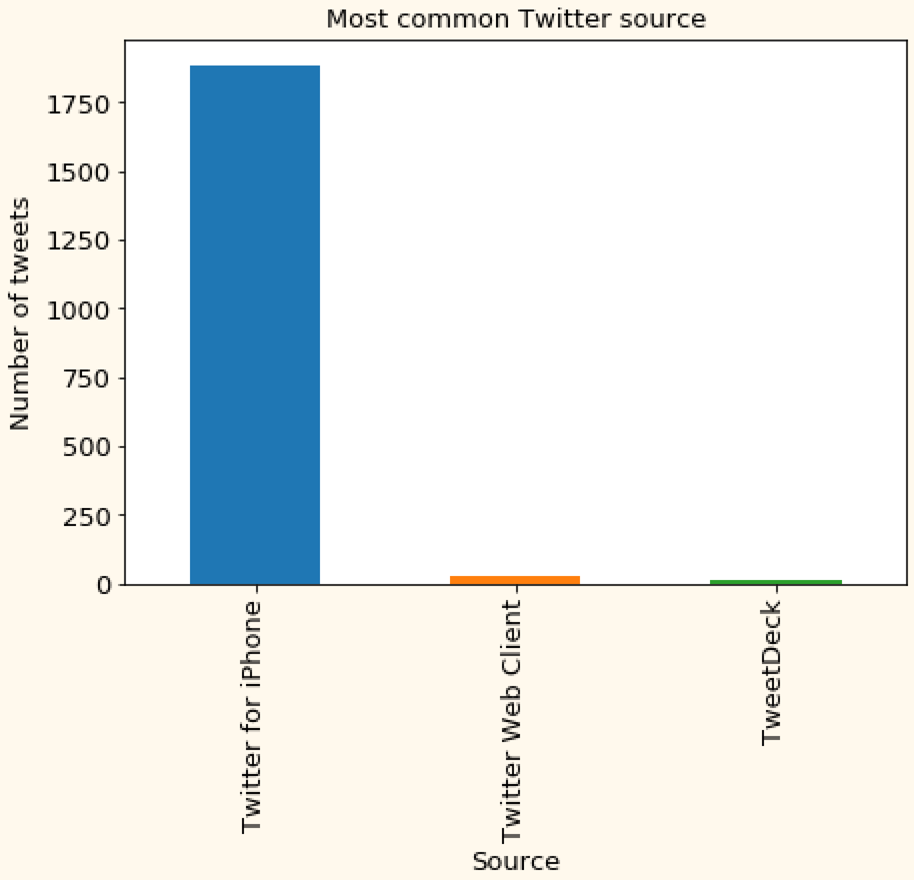
Analysis :

1) Relation between favorites and retweets:

There is a strong correlation between favorites count and retweet counts. This correlation could be useful for WeRateDogs account to understand which method could be useful to increase user’s traffic on the page.



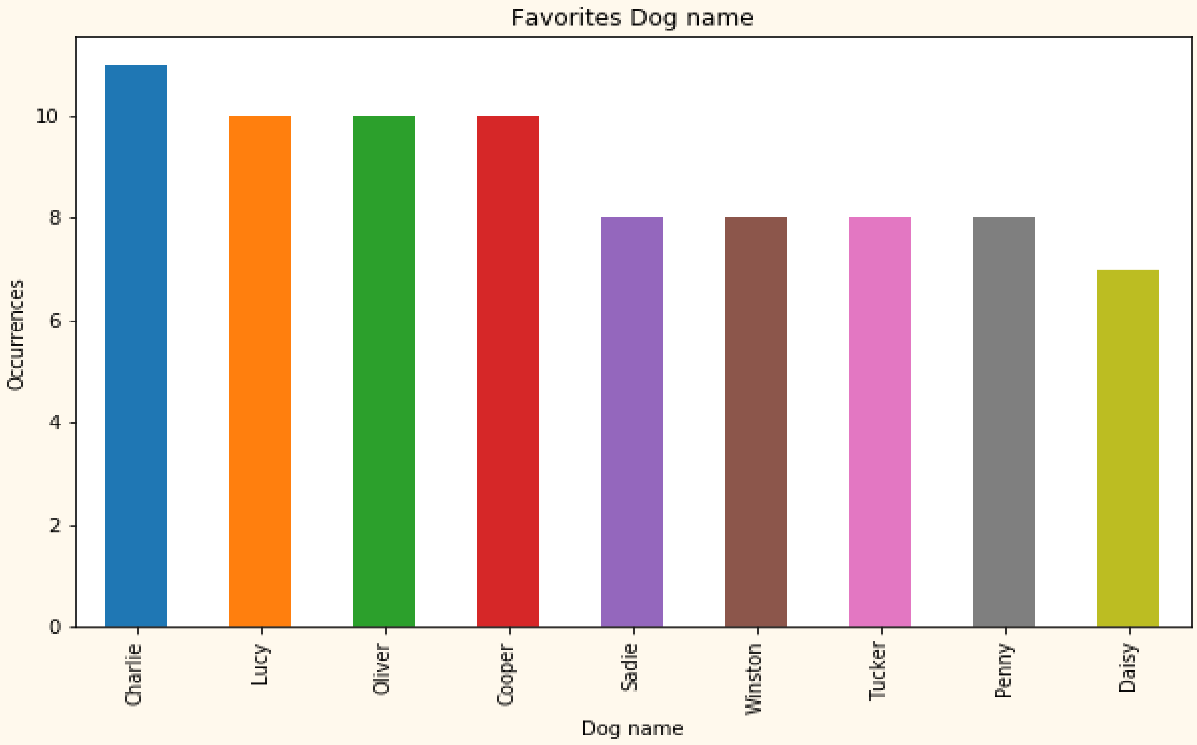
2) Most common Twitter source:



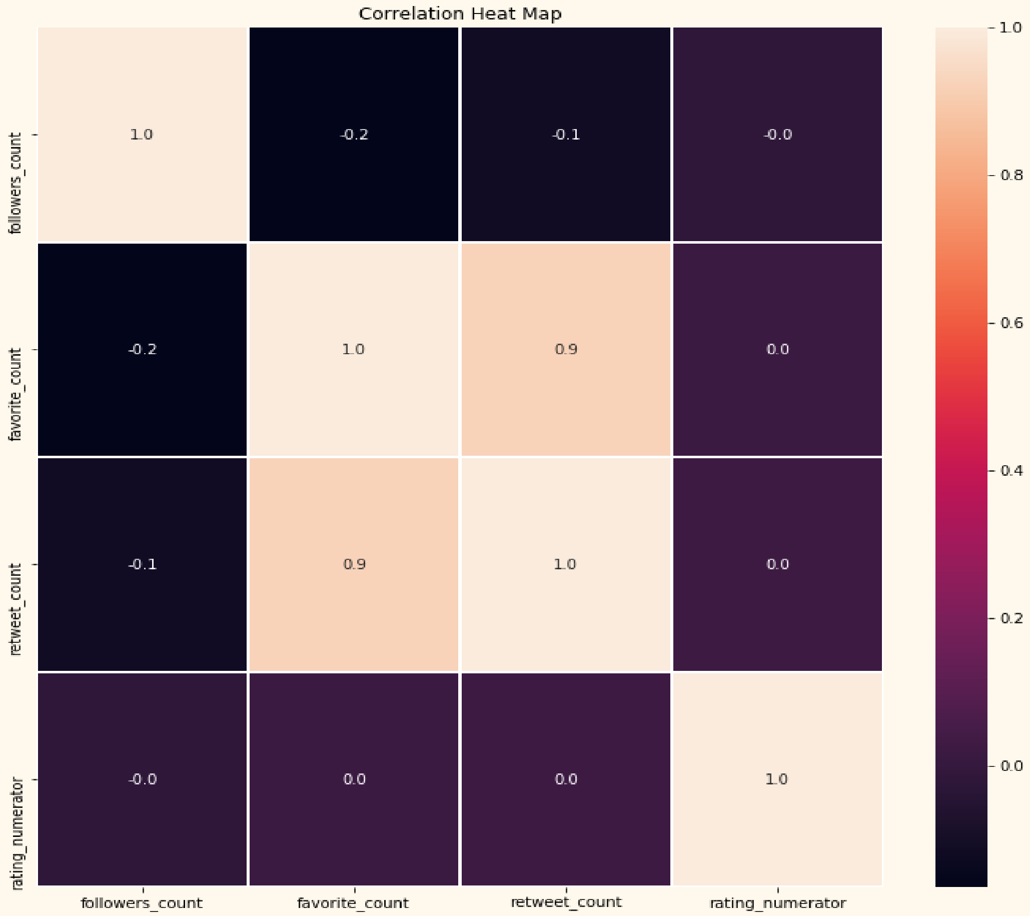
The plot is very clear: the most common Twitter source is iPhone: this is a useful thing to know, because WeRateDogs can create suitable content for the size of the iPhone (eg image sizes).

3) Most common dog name:

This visualization shows that the most popular name for a dog is   Charlie (11) followed by Lucy, Oliver and Cooper (10).



4) Correlation heat map:



This plot shows a strong correlation between followers and retweets (which was expected), and a negative correlation between both followers and favorites and followers and retweets (which wasn't expected).

It's useful to understand the relations between followers, retweets, favorites and ratings.