Insights from WeRateDog Dataset

After the data wrangling process of the WeRateDog datasets, a lot of insights were discovered which includes:

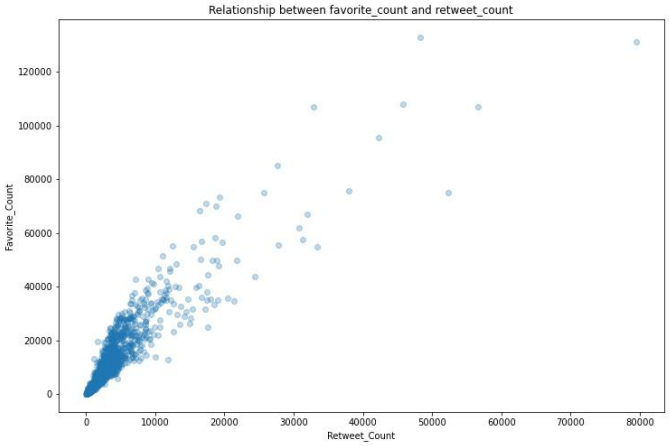
What relationship exist between favorite_count and retweet_count?

On plotting a heatmap of the numerical data, it was evident that there is a strong positive correlation between the retweets count & favorite count.

Also, from the heatmap visualization, it was observed that the number of images has weak correlation with confidence level of algorithm

By visual assessment, it was observed that the favorite_count is always greater than the retweet_count

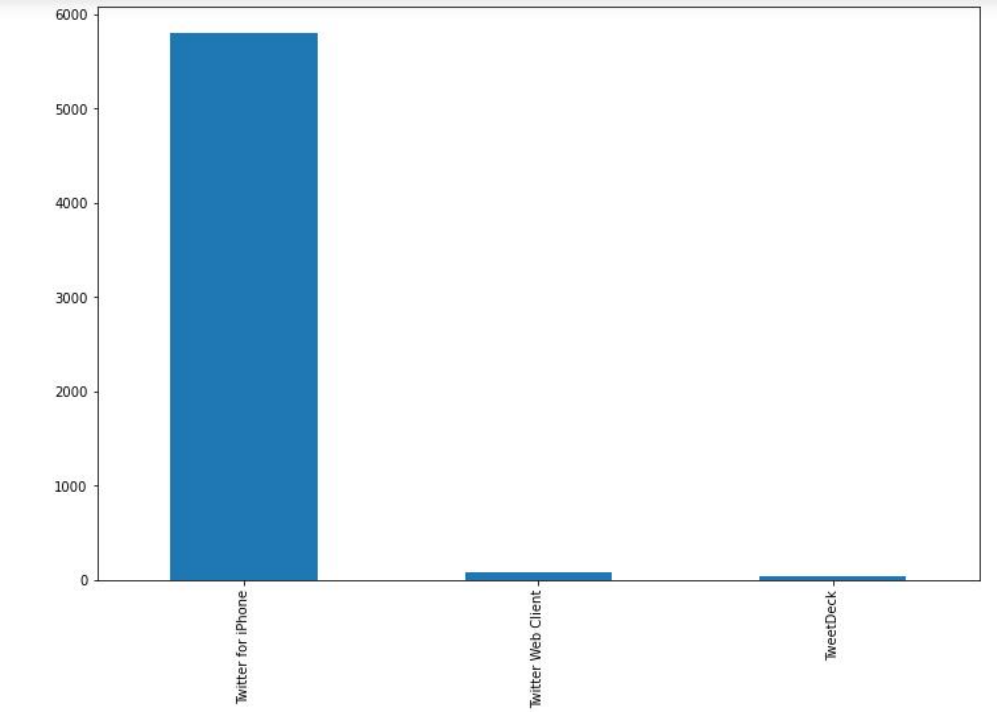
A heatmap of numerical values in the combined dataset



A scatter plot that shows the correlation between favorite_counts and retweet_counts

What's the most used Sources for tweeting?

Using the value_counts() method on the source columns, I noticed that *Twitter for iPhone* is the most used application for Twitter followers of WeRateDogs for tweeting

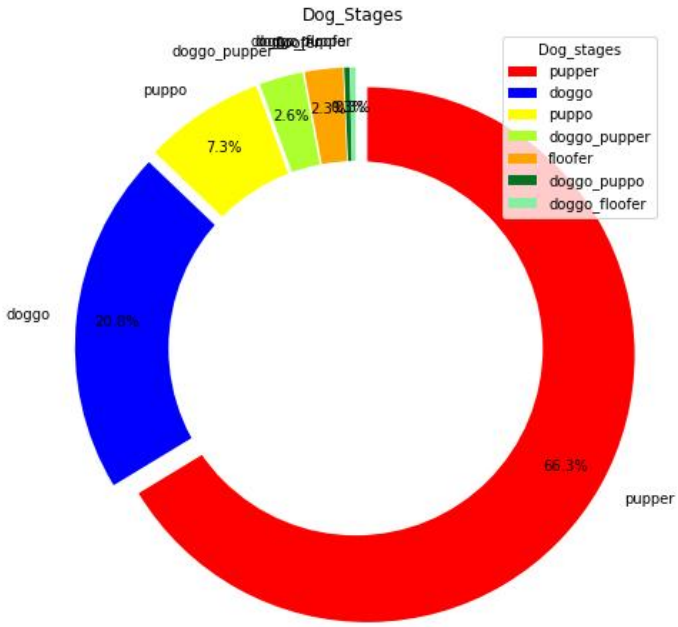


A bar chart showing the sources for tweets

Which dog stage(s) has the highest number of dogs?

Using the value_counts() pandas method, the doggo_puppo and doggo_floofer multistage had same number of dogs(I.e 3) making them have same percentage of 0.3%

By plotting a donut chart of the dog stages, Pupper & Doggo stages were observed to have more than 80% of the dogs.



A donut chart of dog stages

What is the highest retweet count?

Also, using the describe() method on the cleaned combined dataset, the biggest retweet count was about 79,515 and that was in 18th, June 2016(2016-06-18).

Is there any observation on the confidence value?

Using the describe method, I noticed that We the maximum confidence in algorithm was 100%, hence, it must be a clear picture for a specific breed of dog.

Any unique observation on dog ratings?

Among the dog ratings, there were ratings of value zero (0)

Is there any common observation among tweets?

I also noticed that majority of the tweets had only one image.