a. Introduction

The final dataframe we had here was stored in twitter_archive_master.csv that consists of 1652 rows x 21 columns.

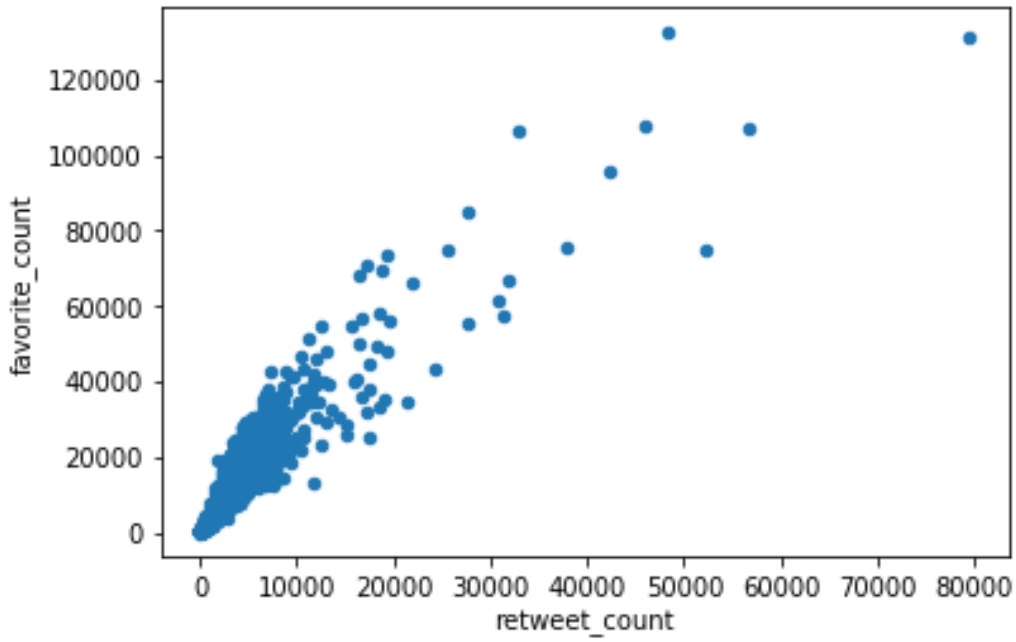
b. Insights

For numerator and denominator columns in order to get the right values I have checked the tweet itself on twitter web page, and referred to that tweet and I found out most of the incorrect values were due to multiple dogs involved in the rating so I dropped the multi-dog rows and some rating was adjusting within the text itself where it was misinterpreted in the rating columns.

And one the things that I acted here which I think it had a big effect on the analysis is that I didn't considered the (simultaneous False prediction at a time) , because I have extract each of them in a list then I included them in a python and checked them one by one to make sure there was no dog breed involved after that I have dropped them all and that's why at the end I got a dataframe with 1652 rows instead of 1954 which was number of rows before dropping those rows.

c. Visualizations

Here I visualized the relation between the retweet_count & favorite_count which shows that is a strong positive relationship which tells the high liked tweet will mostly be retweeted



Here the pie chart below shows the True prediction of dog breed following the algrithm P1 , has a percentage of 88% which I believe it needs to be improved.