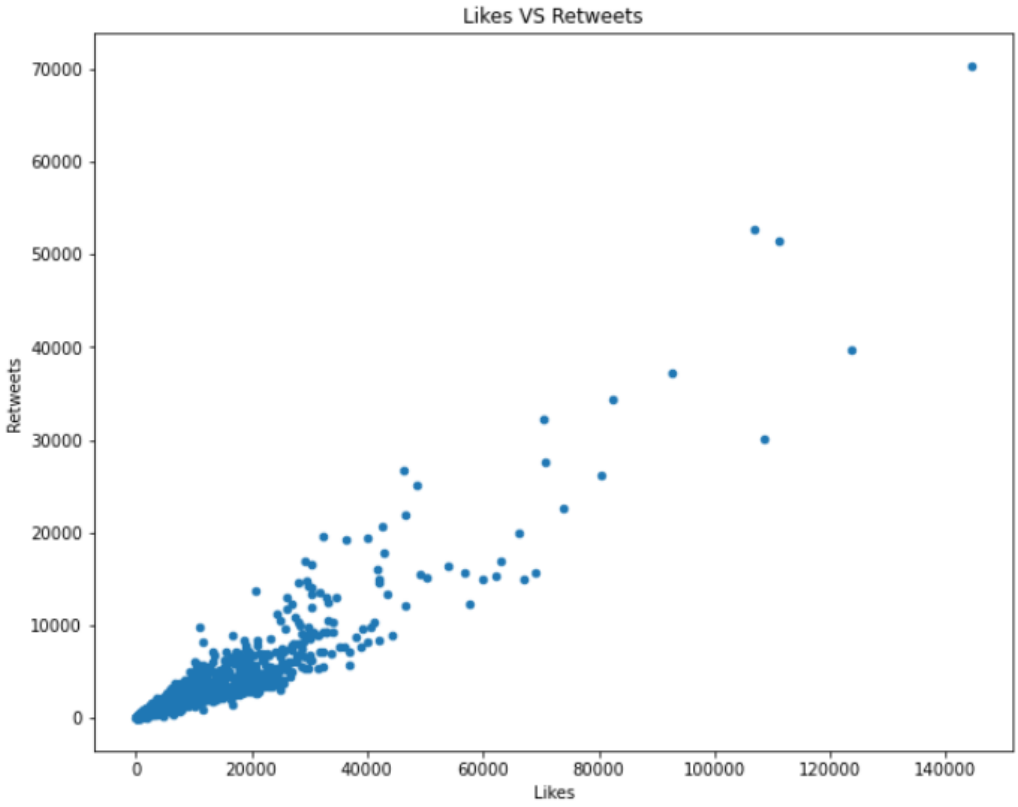
After assessing the data set one after the other, I wrangled and cleaned them programmatically making it ready for analysis.

I started my analysis by first checking the basic info and statistics of the master dataset, from which I moved on to using the pandas describe() function to get a basic statistics of the dataset which revealed some interesting details about the data below:

- The tweet with the lowest number of likes was just **45 likes** & the highest **144,343 likes**
- The tweet with the lowest number of retweets has just **1 retweet** & the highest **70,383 retweets**
- The best/most accurate dog prediction using the first prediction **p1_conf** is **99%** accurate

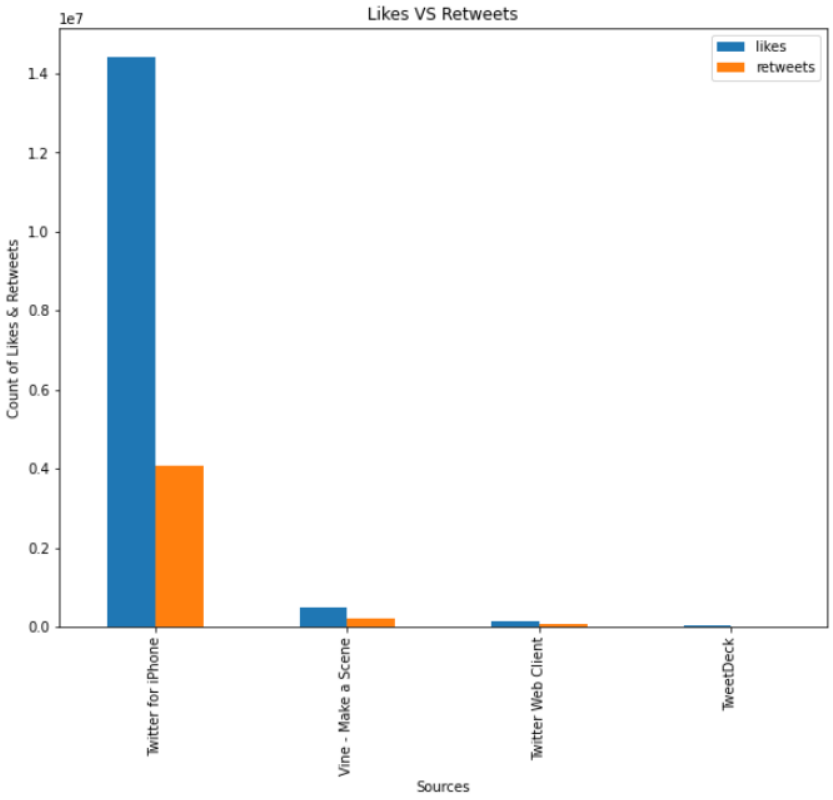
Then I moved on to check if there's a correlation between the number of likes and retweets in the dataset, and as expected there is a strong positive correlation of 0.92 which means as the likes increase the number of retweets increase also.

But we must be careful not to imply that likes causes retweets to grow as this is not enough proof to go on. A chart of a scatterplot with the correlation below:

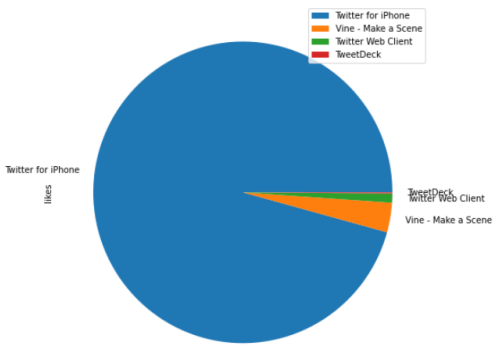


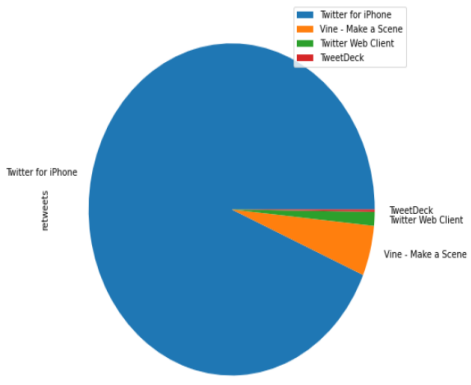
My next analysis was to check the distribution of likes among the various sources of tweets which showed that tweets from iPhone got a greater number of likes and retweets though this is also because most of the tweets were from iPhones. See table and chart below:

source	likes	retweets
Twitter for iPhone	14425527.00000	4060397.00000
Vine - Make a Scene	477881.00000	211887.00000
Twitter Web Client	148057.00000	58362.00000
TweetDeck	17392.00000	10139.00000



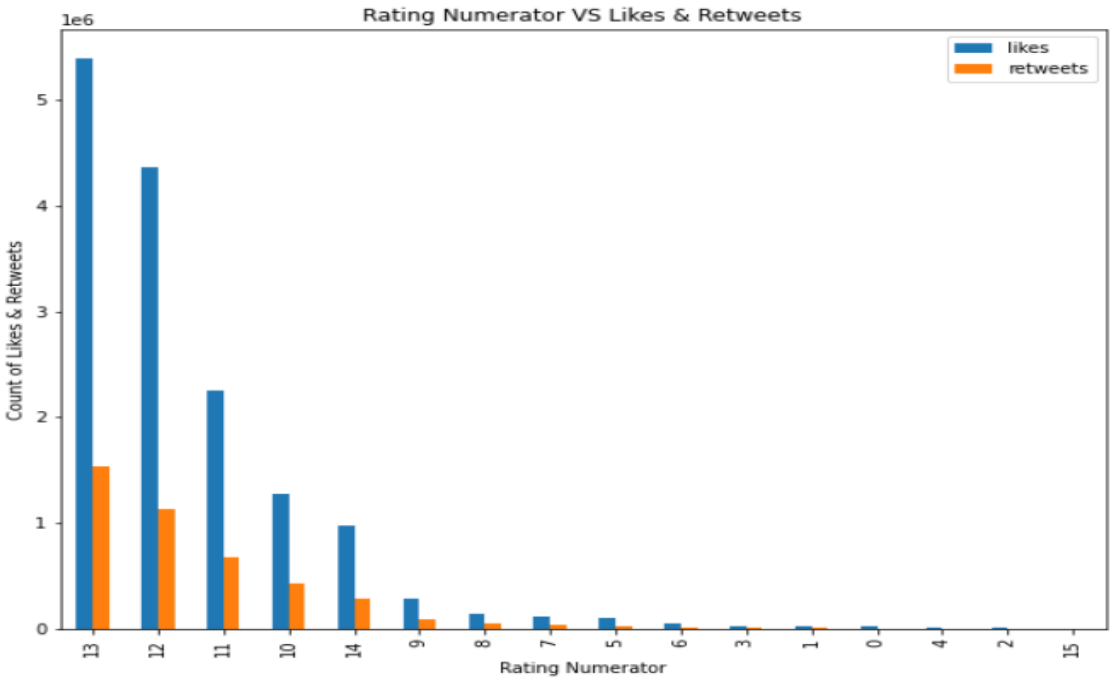
I also used Pie charts:





Then finally my last analysis was checking if the rating_numerator had any effect on the number of likes and retweets, which the analysis showed that the tweets with dogs with higher ratings had more likes and retweets. The chart and table is below:

rating_numerator	likes	retweets
13	5392857.00000	1538713.00000
12	4368787.00000	1135483.00000
11	2256507.00000	678155.00000
10	1272118.00000	435648.00000
14	972993.00000	292989.00000
9	281206.00000	88614.00000
8	142872.00000	49180.00000
7	118923.00000	37559.00000
5	98420.00000	29125.00000
6	57176.00000	19424.00000
3	31990.00000	11511.00000
1	24780.00000	9844.00000
0	20909.00000	2755.00000
4	18694.00000	6495.00000
2	10492.00000	5289.00000
15	133.00000	1.00000



- The above results goes to show that the higher the rating the more likes and retweets the tweet gets except for the case of rating(15). We should also note that this does not neccesarily implies causation as further statistical analysis would need to be done.

This goes to complete my analytical process and insights gotten from the dataset.