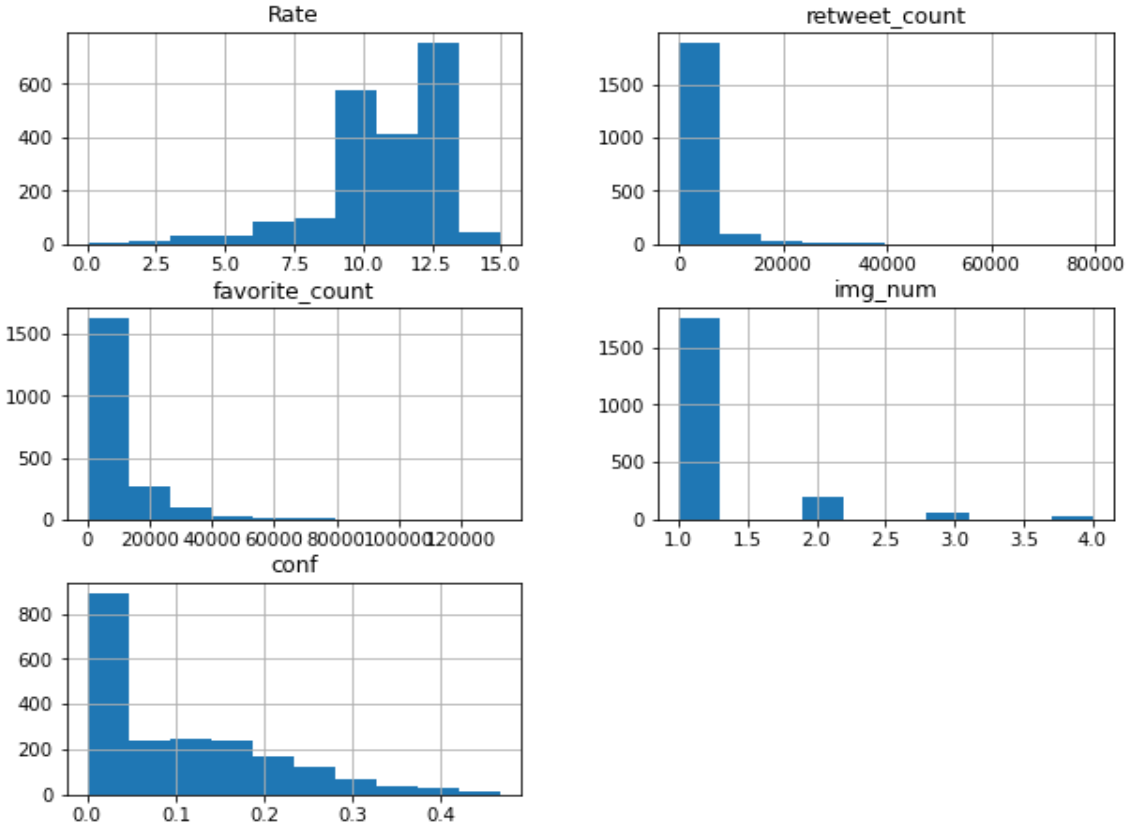
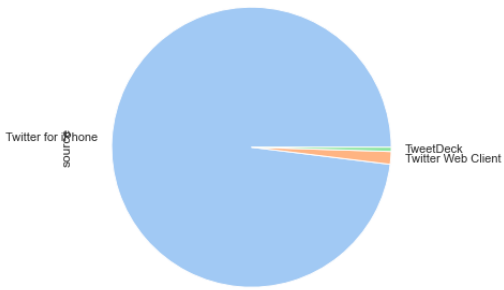
Act report

After data cleaning, the combined data set was analyzed to find the information, visualization, and patterns. The master data contain 14 variables, not all of those columns were necessary for the analysis.

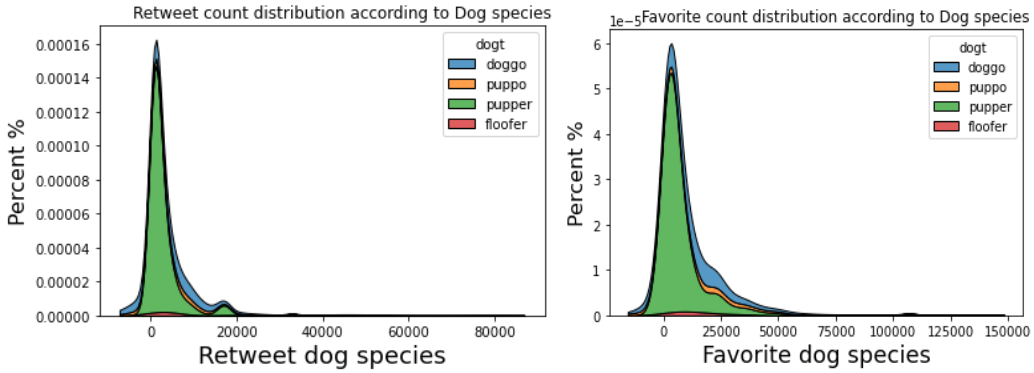
Among all numerical variables, none of them has symmetric distribution according to the below Figure, rate variable has a left-skewed distribution, while retweet, favorite, and conf have right-skewed distribution. Img_num tends to be categorical



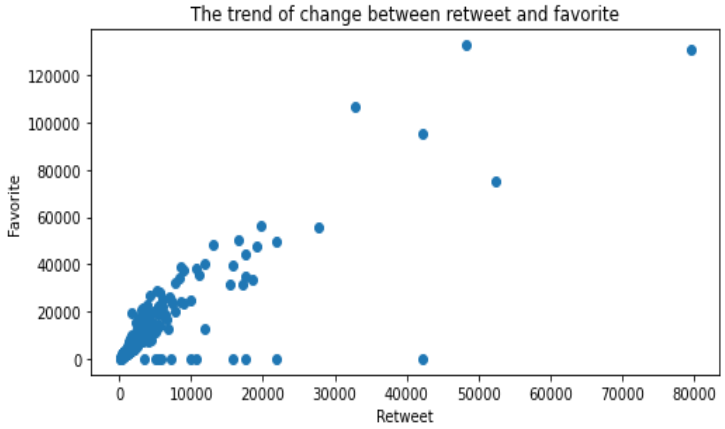
Since the source of the dog rate on the Twitter website is a categorical variable, its distribution showed that the highest percent of rate is from iPhone, as other values occupy the lowest percent.



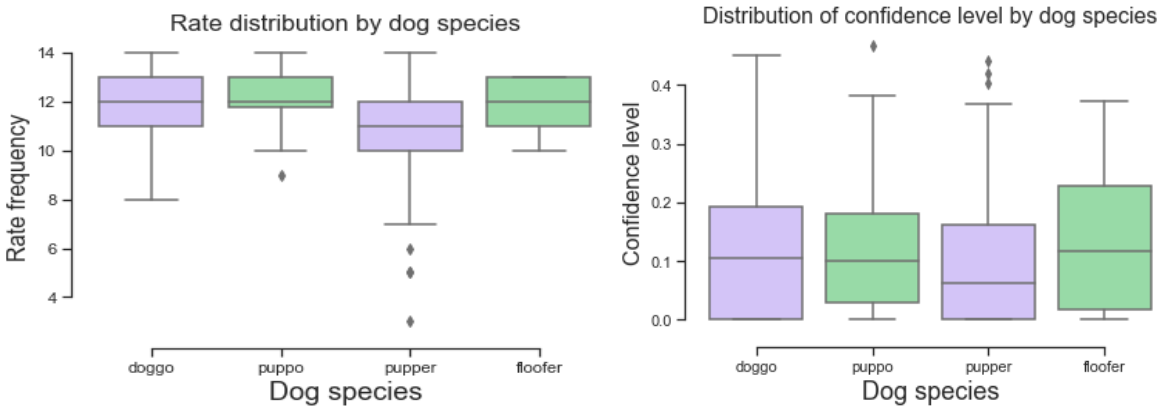
From the bivariate analysis, the distribution plot of retweet and favorite by dog species tend to be approximately similar as both shapes of the distributions are almost similar. In terms of the mean distribution, pupper has the lowest mean in both distributions, which are 7120.9 and 2610.5 for f avorite and retweet respectively. The highest mean values are 21777 on puppo from favorite and 7901.5 on doggo from retweet.



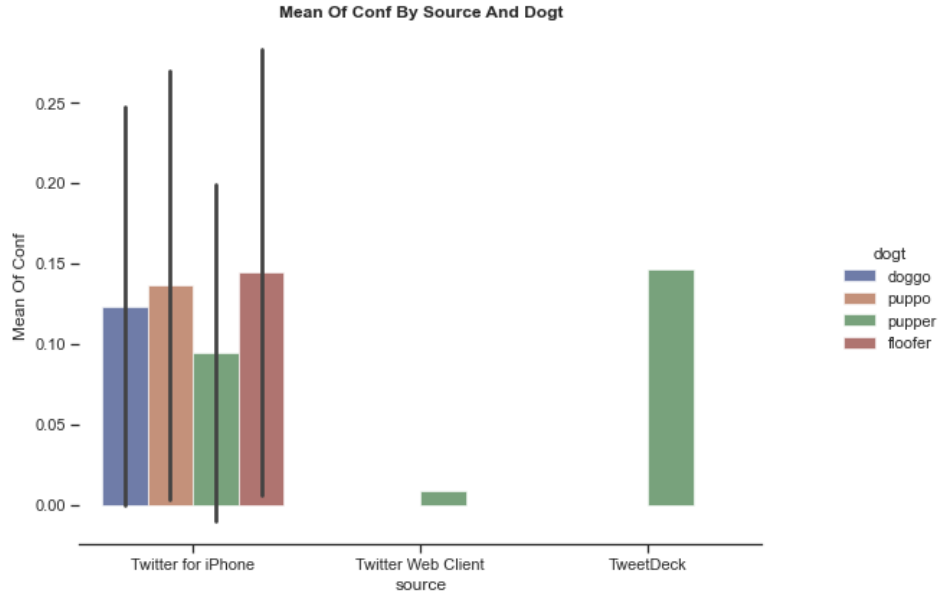
By correlation between favorite and retweet was accessed using a scatter plot, which indicates that there is a high positive association between them, as an increase is equivalent to an increase in other vice verse. The relation is not perfect as some points fall apart from the center, but it is strong.

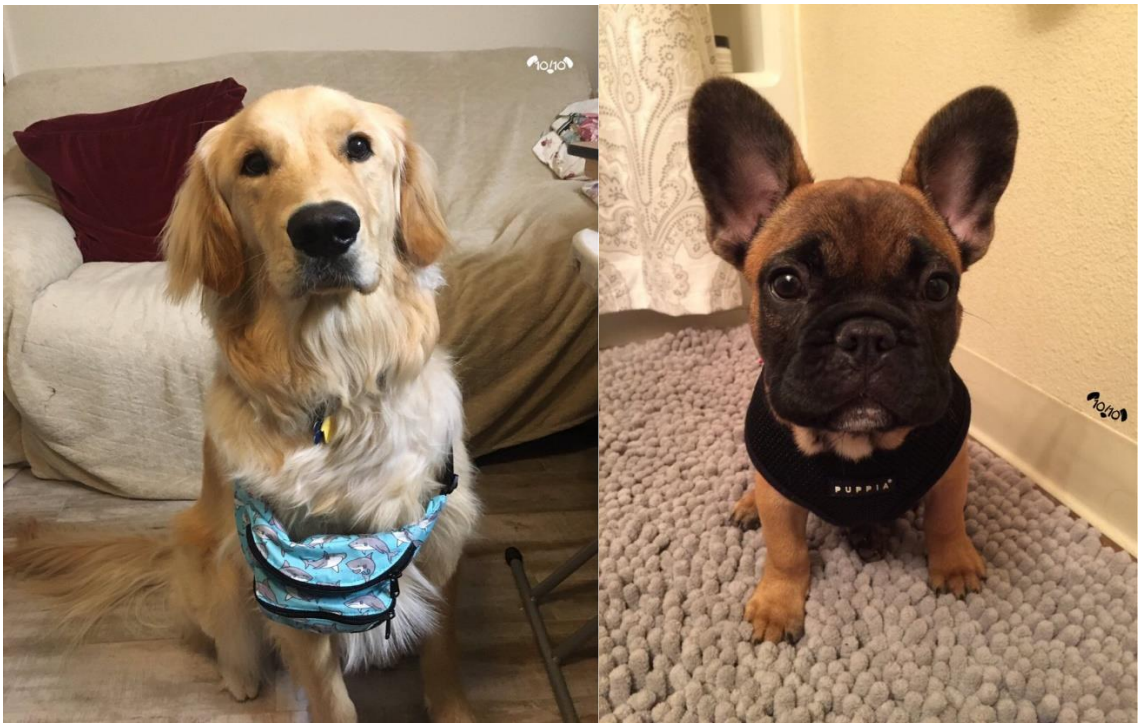


During the data collection, the dogs were rated 10, and the rate was based on the species type, the below Figure showed that the rate column has outliers, also the distribution of rate variable in dog species categories differ. The rate distribution in pupper is somehow normal compared to the rest. The confidence level distribution was not symmetric in all dog species categories, since there presence of the outliers values.



By combining the source and dog species distribution according to confidence level, the Figure below indicates that the majority of the distribution falls under iPhone





This is a sample of the dogs which got the rate great than 10

The data set for this project contain 2050 observations from three combined datasets on the Twitter dog rate with similar key identifier “tweet id”. The main variables for the analysis are dog species type, dog rate, retweet, favorite, and source of Twitter information.

The distribution of the data indicate many people use their phone when they are visiting Twitter to give a rate to the dogs' page, also the data have many outliers since none of the distribution was symmetric.

The findings indicated that there is a strong relation between retweet and favorite since one increases with the other, and also the shape of their distribution is similar.

In conclusion, it was notable that the distribution of rate, retweet, favorite, source, and confidence was affected by dog species type.