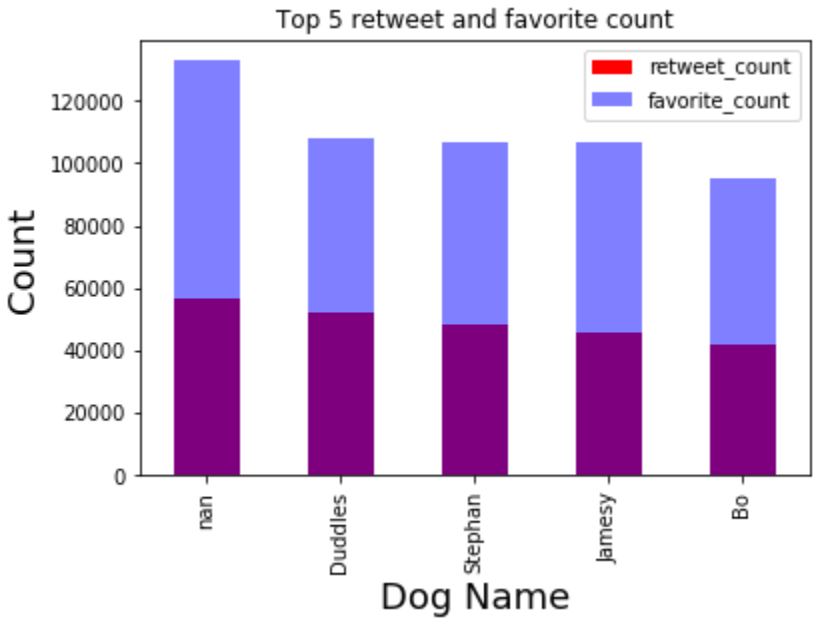
Data Analysis and Visualization

Following the completion of the cleaning exercise to create a master dataset called twitter_archive_master.csv. some analysis was done on the cleaned data to reveal some insights.

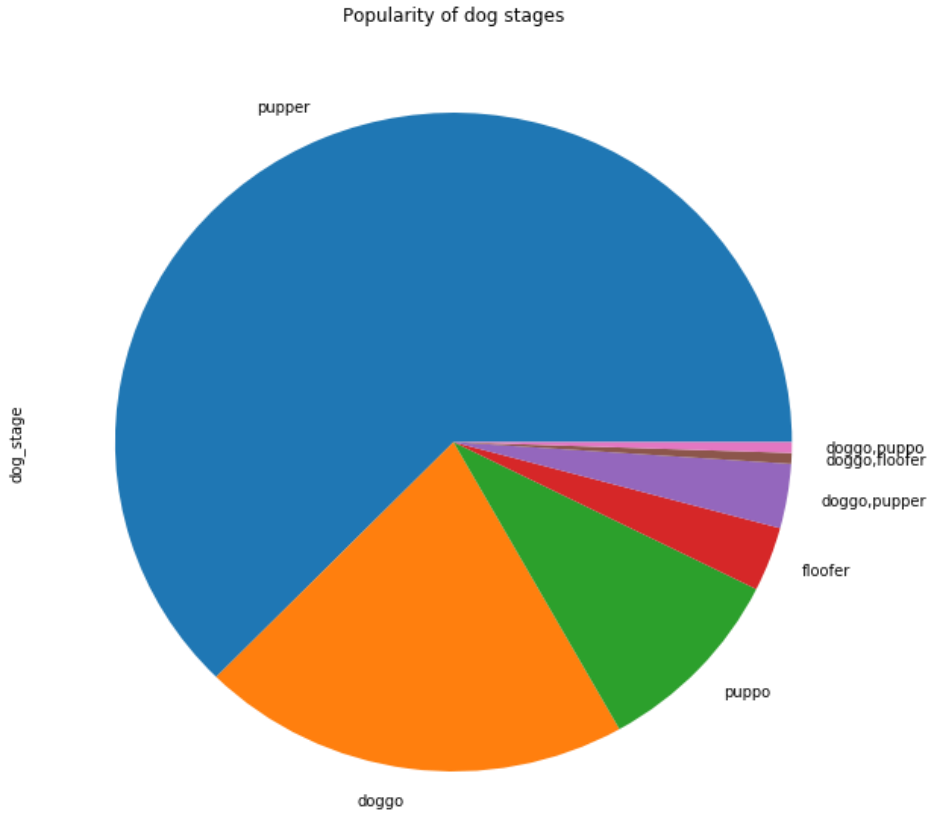
Firstly, the retweet count and favorite count columns with the 5 highest numbers were retrieved to see the corresponding dogs with these attributes. This data was stored in dataframes called highest_retweets and highest_fav_counts. Dog names that featured in this were Stephan, Doodles, Jamesy and Bo were part of the 5 dogs with the highest retweet and favourite counts. However, the bar graph also plotted shows that Nan values in the name columns had a high retweet count and favorite counts. See visualization below:



It is believed that if additional dog names were given in the dataset, the values currently associated with the nans at the moment will be spread out, giving us a true picture of dogs with the highest retweet count and favorite count.

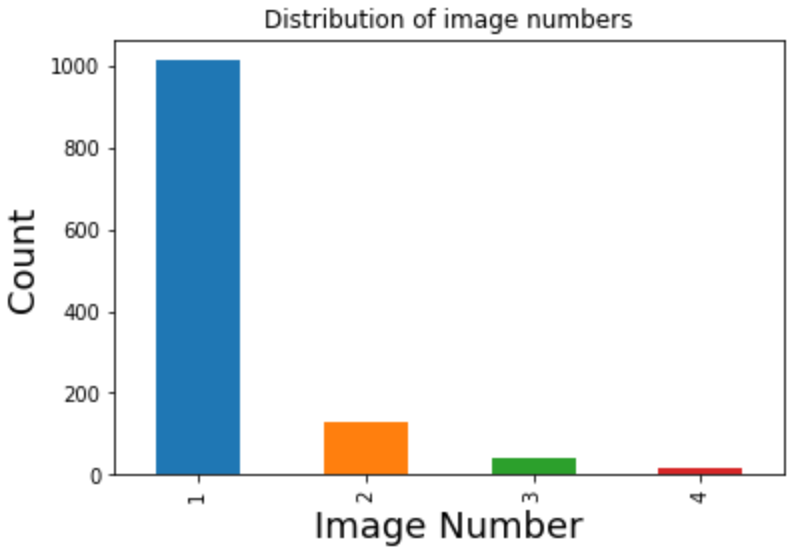
Secondly, using the value_counts() function on the dog_stage column, we were able to determine that Pupper was the most popular dog stage,

followed by 'doggo', 'puppo', 'doggo,pupper', 'floofer', 'doggo,floofer', 'doggo,puppo'. See visualization below:



According to the Dogtionary, a Pupper is a small Doggo, usually younger. Can be equally, if not more mature than some doggos. From the above over 50% of the ratings belong to the Puppers.

Lastly, for the image number criteria, it was noticed that 1 was highly frequent in the dataset. See visualization below:



:

From the above, we can also see that in most cases, the first image is the image that corresponds with the most confident prediction.