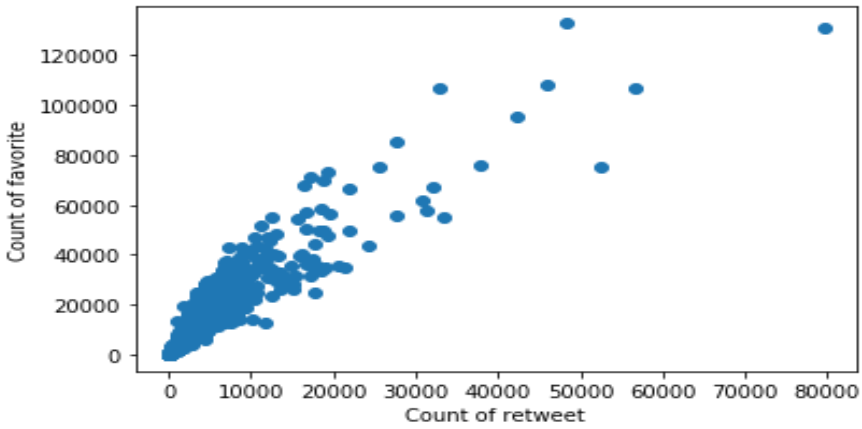
Analyze and Visualize the data

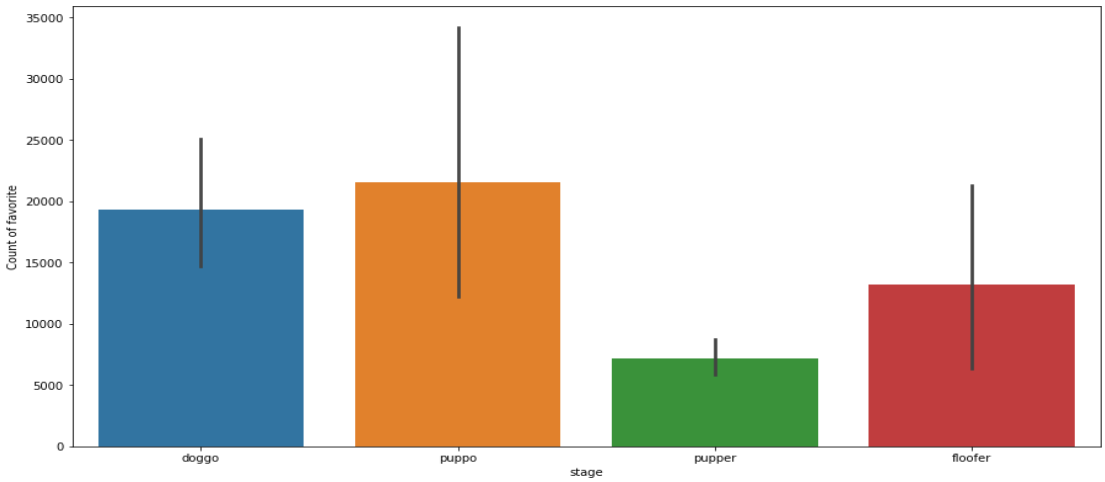
Done by Chen Yi Fei

After the long journey of data wrangling of twitter data, I create a clean output of dataset 'twitter_archive_master.csv' for deep insight analysis. Here are interesting finding.

- Insight-1: both Count of retweet and Count of favorite show very strong positive relationship almost 45 degree of both axies as per observation. It can infer that the more numbers of retweet always follow with more numbers of favorite count in nature. In simple word, if the user doesn't like the photo of dogs, he would not retweet it.



- Insight-2: according to the data, the stage of `puppo` has the highest counts of favorite by each observation which dominates the rest categories. It can infer that the stage of 'puppo' is the most popular one among the others by user's preference. In the future we can try to analyze potential features to explain the behavior.



- Insight-3: I create new variable `ratio` which is defined as rating_numerator divided by rating_denominator. The newly created variable can standardize the scale of rating info

when I compare it within different sub-groups. After that, I also generate descriptive statistics across all 4 groups and observe the distribution. I found all the 4 groups show similar distribution in terms of mean, std and others. Only the pupper shows much more wide spread distribution with higher std=.17. We can infer that the distribution of rating in puppo group is more volatile than the others. But the sample size of both floofer and puppo are not large enough for analysis. I would suggest to ignore these 2 groups of result.

	ratio							
	count	mean	std	min	25%	50%	75%	max
stage								
doggo	63.0	1.188889	0.147135	0.8	1.100	1.2	1.3	1.4
floofer	7.0	1.200000	0.115470	1.0	1.150	1.2	1.3	1.3
pupper	203.0	1.063680	0.172818	0.3	1.000	1.1	1.2	1.4
puppo	22.0	1.200000	0.130931	0.9	1.125	1.2	1.3	1.4