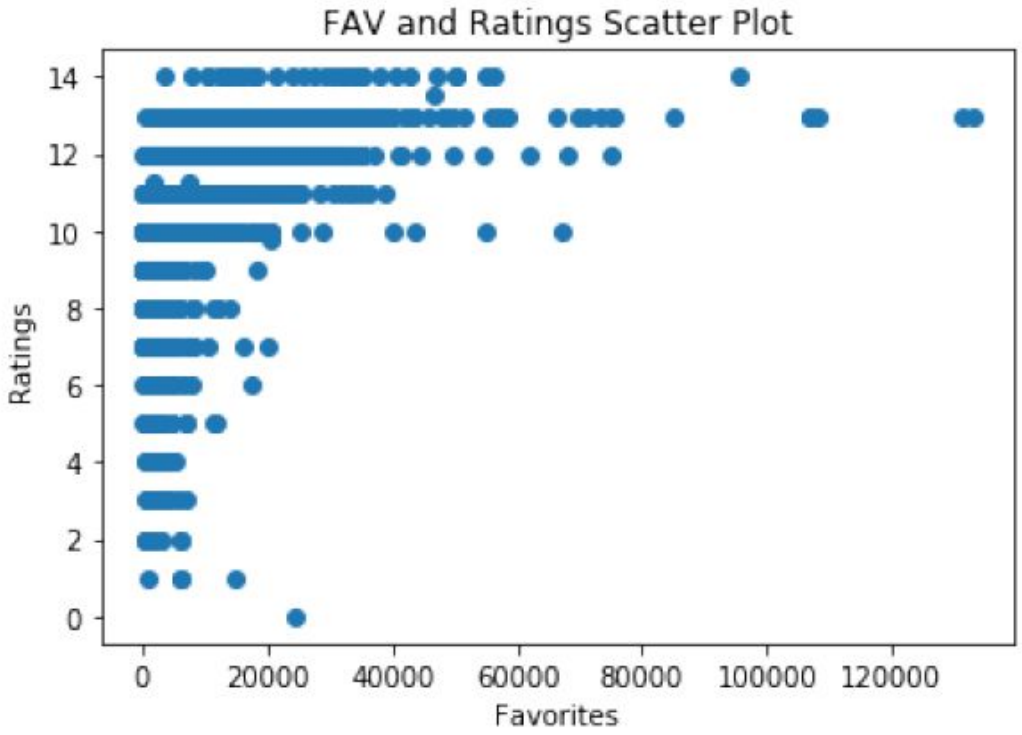
Udacity Project # 2
Data Wrangling: We Rate Dogs
By: Maryam Mohammed

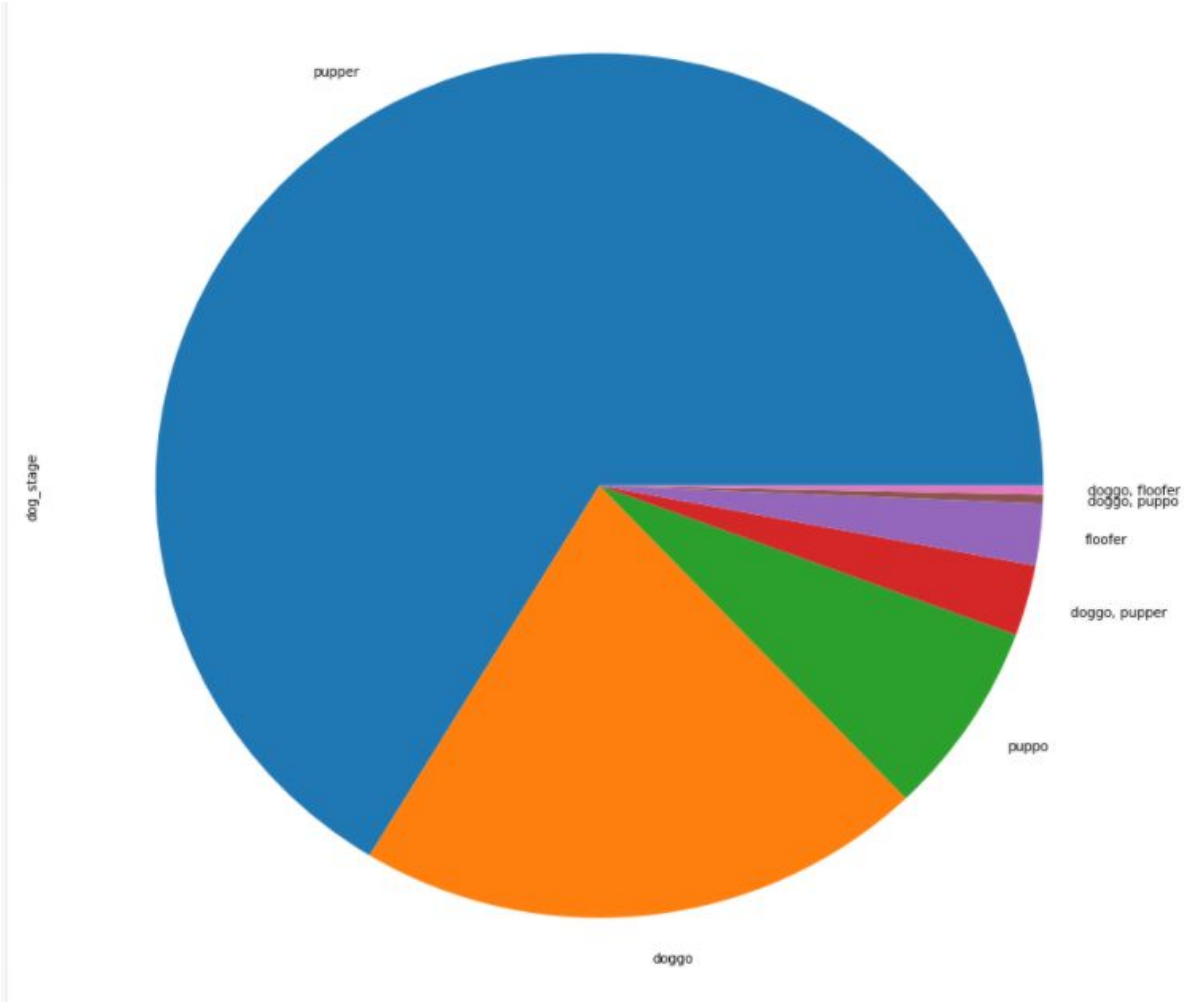
Welcome to my second project with udacity in the Data Analysis Nanodegree. In this project, I applied what I have learnt throughout the course. I gathered the needed data. I assessed the data and spotted some issues that I have worked on resolving them later in the cleaning process. My project consists of 5 parts. Gathering, Assessing, Cleaning, Storing, Visualizations and Conclusions. The following you are some key findings of my project:

Figure no. 1: Are the dogs who receive high ratings also favorited by a high number of other users?



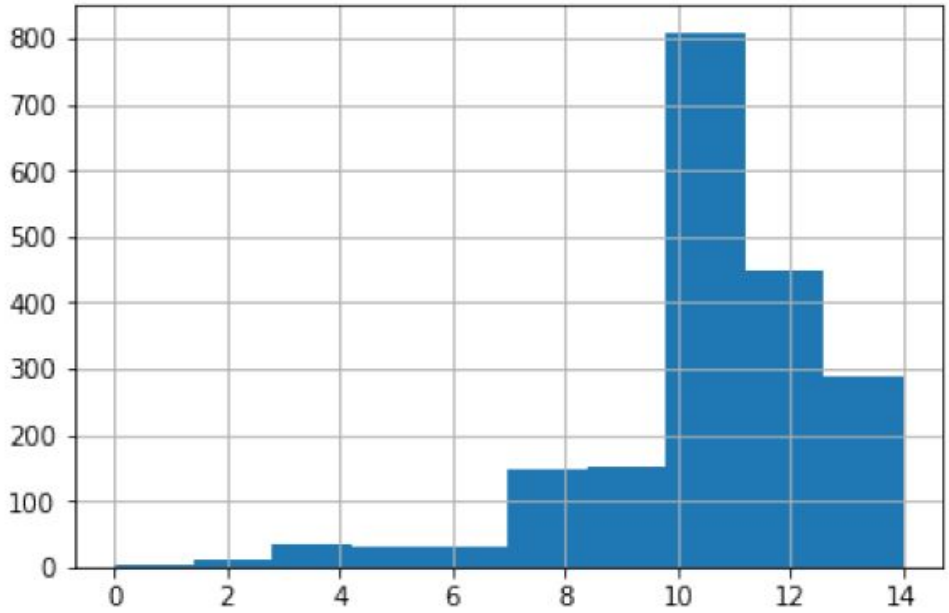
The above scatter plot shows that the frequencies of favorites are high when the rating is also high.

Figure no. 2: What is the most common dog stage in its life cycle among the users?



The above pie chart shows that the majority of tweets is concerned with the pupper stage of dogs.

Figure no. 3 - Numerator ratings frequencies:
What does the distribution look like for the dog ratings?



The histogram shows that the distribution is skewed to the left meaning that people tend to give high scores for dogs in general.

Figure no. 4: How are favorites and retweets correlated together?

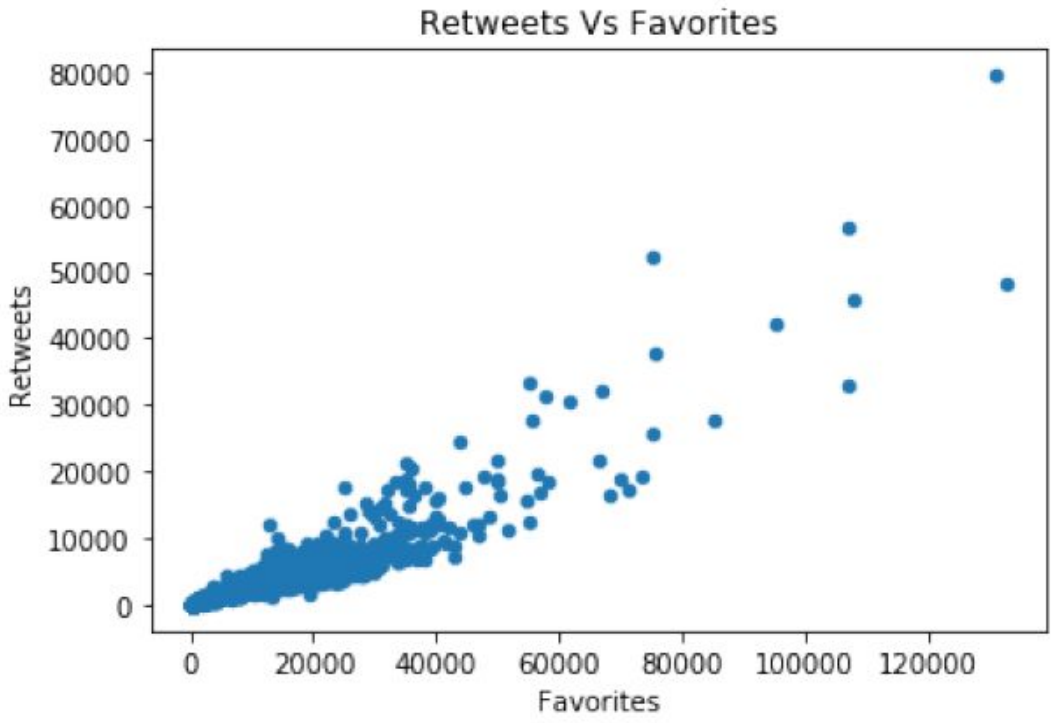


Figure no. 5

	retweet_count	favorite_count	p1_conf	p1_dog	\
tweet_id	0.408362	0.657249	0.103995	0.121065	
rating_numerator	0.312199	0.418245	0.102997	0.291527	
rating_denominator	NaN	NaN	NaN	NaN	
retweet_count	1.000000	0.913182	0.055848	0.009500	
favorite_count	0.913182	1.000000	0.079702	0.054800	
p1_conf	0.055848	0.079702	1.000000	0.124663	

Not all favorited tweets are retweeted.There is a clear linear relationship between the two variables. There is a strong positive correlation between favorite count and retweet counts with coefficient of 0.913182 and the correlation between retweets, favorites with the numerator ranking is weak.