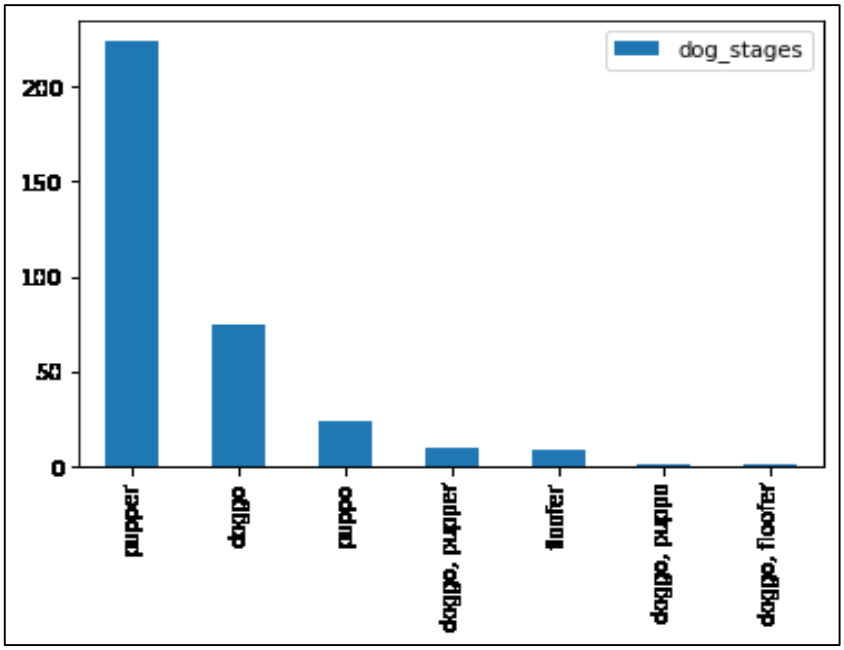
ANALYZING AND VISUALIZING DATA

I recently undertook a data wrangling project by Udacity that involved analysing the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The first step involved loading the data into my Jupyter notebook via the *pd.read_csv* function. My cleaned dataset (cleaned_data) had a total of 2127 rows and 20 columns.

I then analysed and visualized the data to derive a few insights, as shown below:

1) The most common dog stage is the pupper, followed by the doggo and puppo, each with 224, 75 and 24 dogs respectively. The distribution is as below:



Looking further into the dog breeds under the pupper dog stage, the Golden retriever, the Pembroke and Labrador retriever were the most common, with 16, 10 and 8 dogs respectively.

2) On average, the standard poodle dogbreed received the most retweets when compared to other breeds under the first prediction (p1). On the other hand, the black-and-tan coonhound dogbreed received the most likes.

```
In [107]: #Checking for the most retweeted dog breed based on the image predictions
          cleaned_data.groupby('p1').mean().sort_values(by='retweet_count', ascending=False)
```

Out[107]:

p1	tweet_id	rating_numerator	rating_denominator	retweet_count	favorite_count	img_num	p1_conf	p2_conf	p3_conf
standard_poodle	7.262901e+17	10.500000	10.0	8762.750000	21084.250000	1.000000	0.423407	0.204231	0.072459
black-and-tan_coonhound	7.602637e+17	10.500000	10.0	6584.000000	29155.000000	2.500000	0.692000	0.147506	0.097786
Saluki	8.315403e+17	12.500000	10.0	5220.333333	26667.333333	1.000000	0.523054	0.208351	0.118570
English_springer	7.204467e+17	11.111111	10.0	4850.666667	13410.444444	1.000000	0.546486	0.200781	0.080672
Afghan_hound	8.041621e+17	9.666667	10.0	4774.000000	14671.000000	1.000000	0.433959	0.099034	0.081650
...	...	...	...	...	...	...	...	...	...
Tibetan_terrier	6.973258e+17	9.250000	10.0	386.000000	1393.333333	1.000000	0.408551	0.144144	0.069139
Japanese_spaniel	6.773010e+17	5.000000	10.0	354.000000	1109.000000	1.000000	0.661178	0.150119	0.119720
groenendael	6.939424e+17	10.000000	10.0	328.000000	1621.000000	1.000000	0.550796	0.154770	0.080802
Brabancon_griffon	6.725689e+17	10.000000	10.0	227.333333	742.333333	1.333333	0.369981	0.247691	0.102877
Sussex_spaniel	7.151443e+17	11.000000	10.0	175.000000	579.000000	1.000000	0.379473	0.198698	0.139757

111 rows × 9 columns

```
In [108]: #Checking for the most liked dog breed based on the image predictions
          cleaned_data.groupby('p1').mean().sort_values(by='favorite_count', ascending=False)
```

Out[108]:

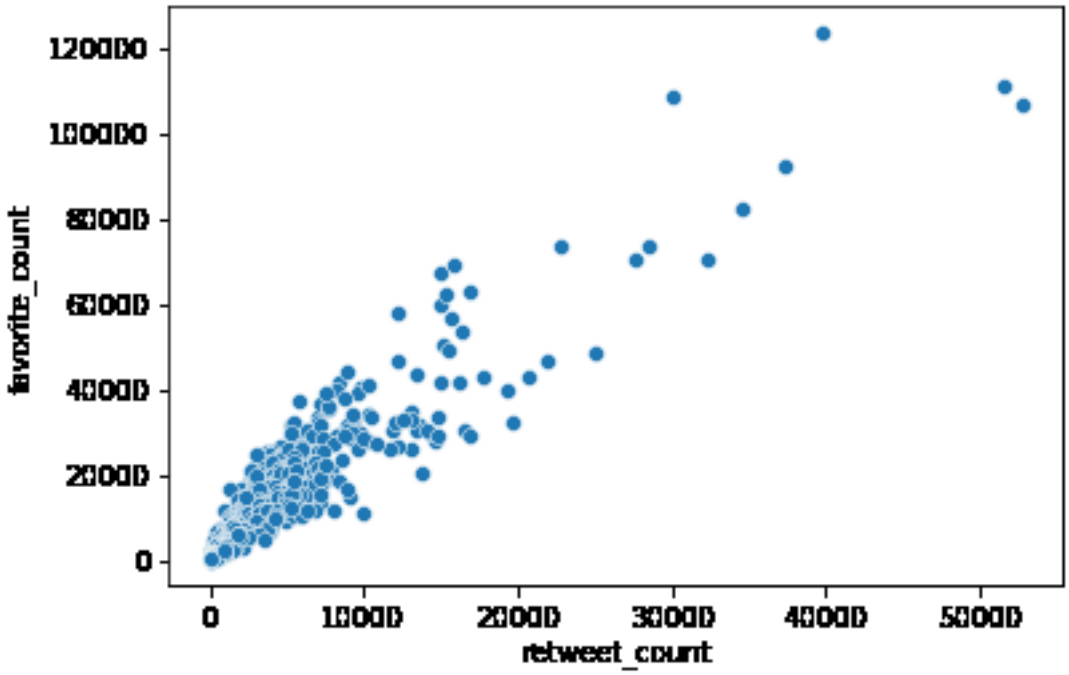
p1	tweet_id	rating_numerator	rating_denominator	retweet_count	favorite_count	img_num	p1_conf	p2_conf	p3_conf
black-and-tan_coonhound	7.602637e+17	10.500000	10.0	6584.000000	29155.000000	2.500000	0.692000	0.147506	0.097786
Saluki	8.315403e+17	12.500000	10.0	5220.333333	26667.333333	1.000000	0.523054	0.208351	0.118570
standard_poodle	7.262901e+17	10.500000	10.0	8762.750000	21084.250000	1.000000	0.423407	0.204231	0.072459
French_bulldog	7.855027e+17	11.291667	10.0	4114.217391	17041.217391	1.125000	0.767621	0.096307	0.033183
Afghan_hound	8.041621e+17	9.666667	10.0	4774.000000	14671.000000	1.000000	0.433959	0.099034	0.081650
...	...	...	...	...	...	...	...	...	...
Tibetan_terrier	6.973258e+17	9.250000	10.0	386.000000	1393.333333	1.000000	0.408551	0.144144	0.069139
Ibizan_hound	6.754075e+17	9.000000	10.0	475.000000	1293.500000	1.000000	0.413412	0.104057	0.059385
Japanese_spaniel	6.773010e+17	5.000000	10.0	354.000000	1109.000000	1.000000	0.661178	0.150119	0.119720
Brabancon_griffon	6.725689e+17	10.000000	10.0	227.333333	742.333333	1.333333	0.369981	0.247691	0.102877
Sussex_spaniel	7.151443e+17	11.000000	10.0	175.000000	579.000000	1.000000	0.379473	0.198698	0.139757

111 rows × 9 columns

3) On average, the pupper dog stage received the least likes and retweets, despite being the most common dog stage.
4) The dog with the highest retweets is an Eskimo dog with 52,724 retweets. Photo shown below:

5) The dog with the highest likes is a Lakeland terrier with 123,717 likes. Photo shown below:

I then looked into whether there was any correlation between the retweet count and favourite count of my dataset. To this, there was a positive correlation of 0.965. This implies that there a likelihood that a tweet with a high retweet count, will also have a high favourite count. Although this does not mean that one causes the other.



The data that was used in my analysis and visualisation is from the ‘twitter_archive_master.csv’.