Udacity Data wrangling Act report

INTRODUCTION

WeRateDogs is twitter account that rates dogs. The dog ratings are with a denominator of 10 and the numerator can be larger than 10 to signify how highly rated a dog is. The data was gathered from three sources: twitter archive and image predictions dataset, provided by Udacity, and also by querying the Twitter API. This exploratory data analysis contains just a few insights and visualizations as that wasn't the main scope of this project. The visualizations were created with some python libraries, they are matplotlib, seaborn and word cloud.

Insight 1: What are the top rated and lowest rated dogs?

In this insight, I first decided to get the most rated dog which I found out to be Atticus using the *nlargest* pandas' function on the ratings column, but this has a rating of 1776/10 which may be ambiguous, therefore, I decided to get the highest rated dogs in general. After proper wrangling the top-rated dogs are:

Atticus with a rating of 1776/10

Cassie with a rating of 14/10

The lowest rated dogs are:

Happy with a rating of 2/10

Henry with a rating of 2/10

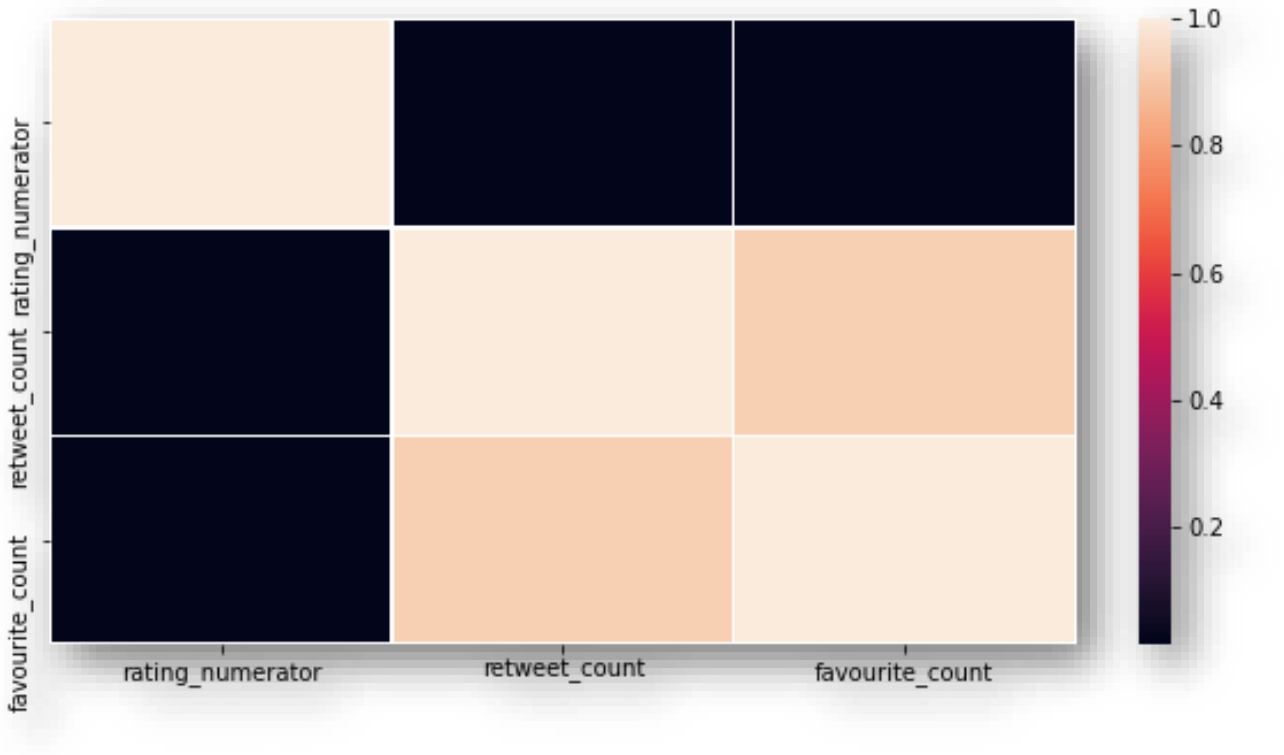
Insight 2: What possible correlations exist between ratings, retweet count and favorite count?

I created a sub data frame that consists of the rating numerator column, the retweet count and favorite count, then I used the pandas' function *corr* which gets the correlation between pairs of numerical data. I did this in order to find out most especially if ratings have an effect on the number of retweets of a dog

```
df_corr.corr()
```

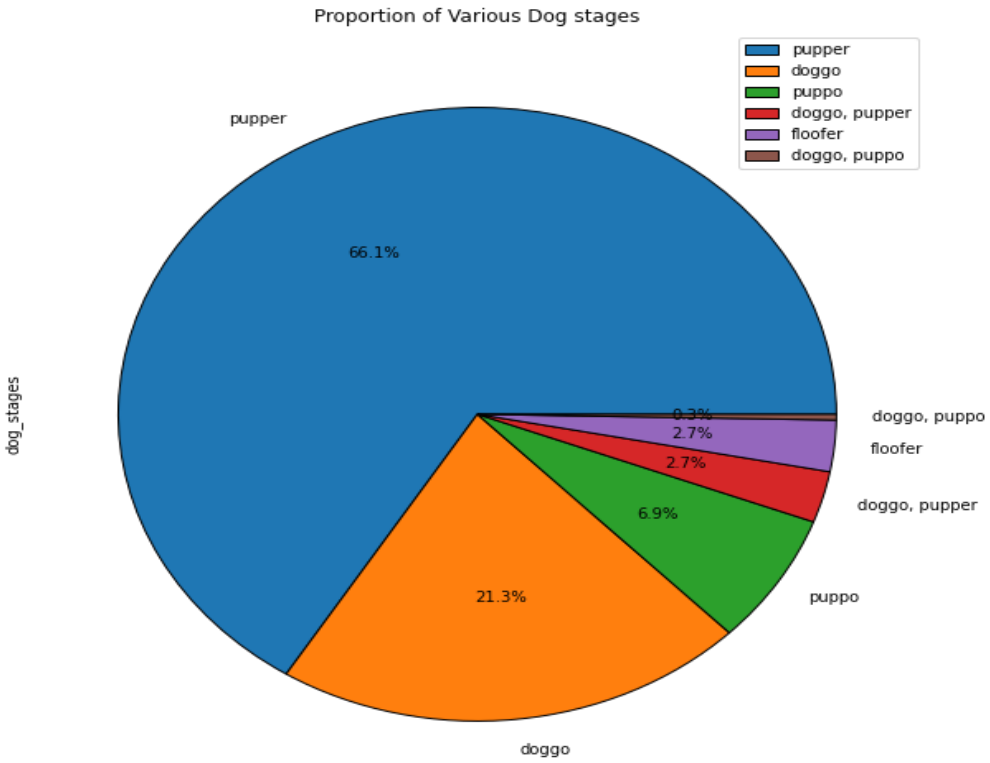
	rating_numerator	retweet_count	favourite_count
rating_numerator	1.000000	0.022638	0.021993
retweet_count	0.022638	1.000000	0.925530
favourite_count	0.021993	0.925530	1.000000

Retweet count varies strongly with favorite count with a correlation of 0.92 Rating and retweet count varies very weekly with a correlation of 0.02. This can also be visualized with a heatmap. If a dog has a high favorite count the tendencies for it to be retweeted are high. It can also be seen that the rating has little effect on the retweets and favorite count. This is understandable because the ratings are "only" part of the site's joke. It can also be random, not reflecting any particular intention. It is based on a particular person's feelings.



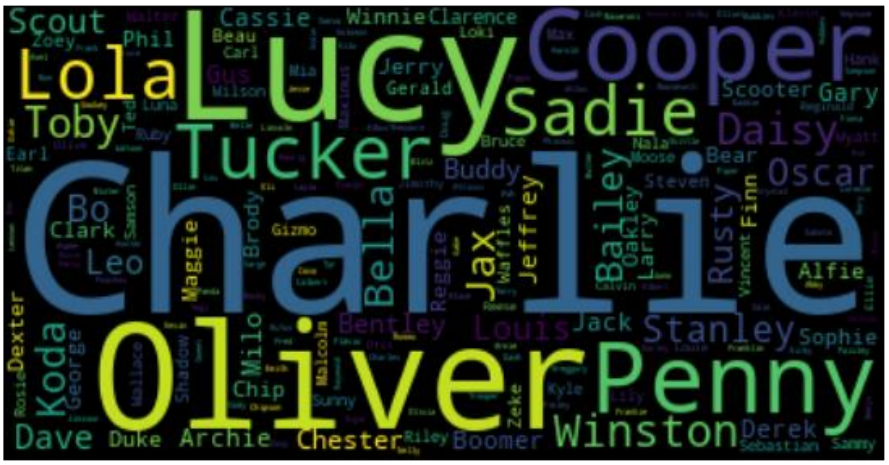
Insight 3: What are the variations of proportions in dog stages?

Using the value counts function,I found the proportions of various dog stages dog stages and they are: pupper the dog stage with 66% , doggo with 21%, puppo with 6.9%, floofer with 2.7%, multiple ratings like doggo, pupper and doggo, puppo with a rating of 2.7% and 0.3% respectively. It can be seen with the pie chart below



Insight 4: The most popular dog names

Since the we rate dogs is a twitter account that has to do with dogs and those dogs have names, we can find out what kind of names are very common among different dogs. In order to find popular dog names, I created a word cloud below



From this word cloud, we can see that most common dog names are: Charlie, Oliver, Cooper and Lucy and they are all names of people. Therefore, it can be inferred that most dog owners give their pets individual names.