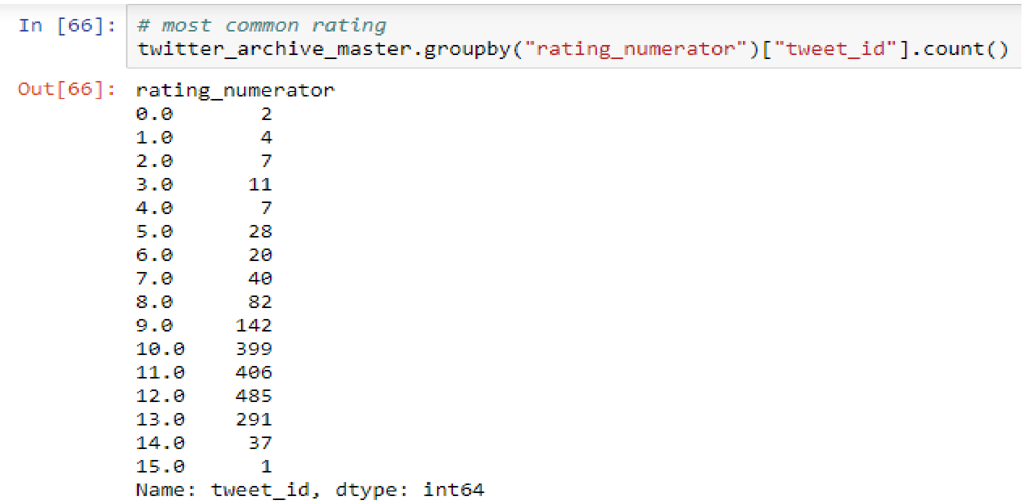
DATA WRANGLING PROJECT (WeRateDogs)

WeRateDogs (@dog_rates) is a twitter handle that rates pictures of people’s dogs. This project focused on the 3 key aspects of Data wrangling: data gathering, assessing, and cleaning.

Data was gathered by downloading files manually(file containing tweet archive’s made available by WeRateDogs) and programmatically. Missing data were obtained using tweepy to query twitter’s application programming interface (API). The gathered data were thereafter assessed visually and programmatically for structural and quality issues before they were cleaned.

Insights: A number of insights were generated while exploring the cleaned dataset:

1. Most dogs were rated a 12/10 with the most common rating range being between 8 and 13.

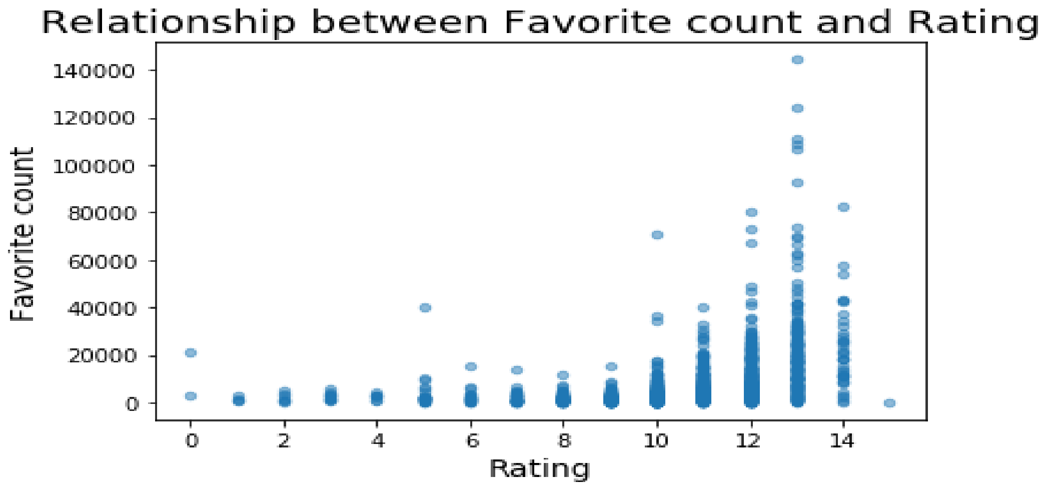


```
In [66]: # most common rating
         twitter_archive_master.groupby("rating_numerator")["tweet_id"].count()

Out[66]: rating_numerator
         0.0       2
         1.0       4
         2.0       7
         3.0      11
         4.0       7
         5.0      28
         6.0      20
         7.0      40
         8.0      82
         9.0     142
         10.0    399
         11.0    406
         12.0    485
         13.0    291
         14.0     37
         15.0      1
         Name: tweet_id, dtype: int64
```

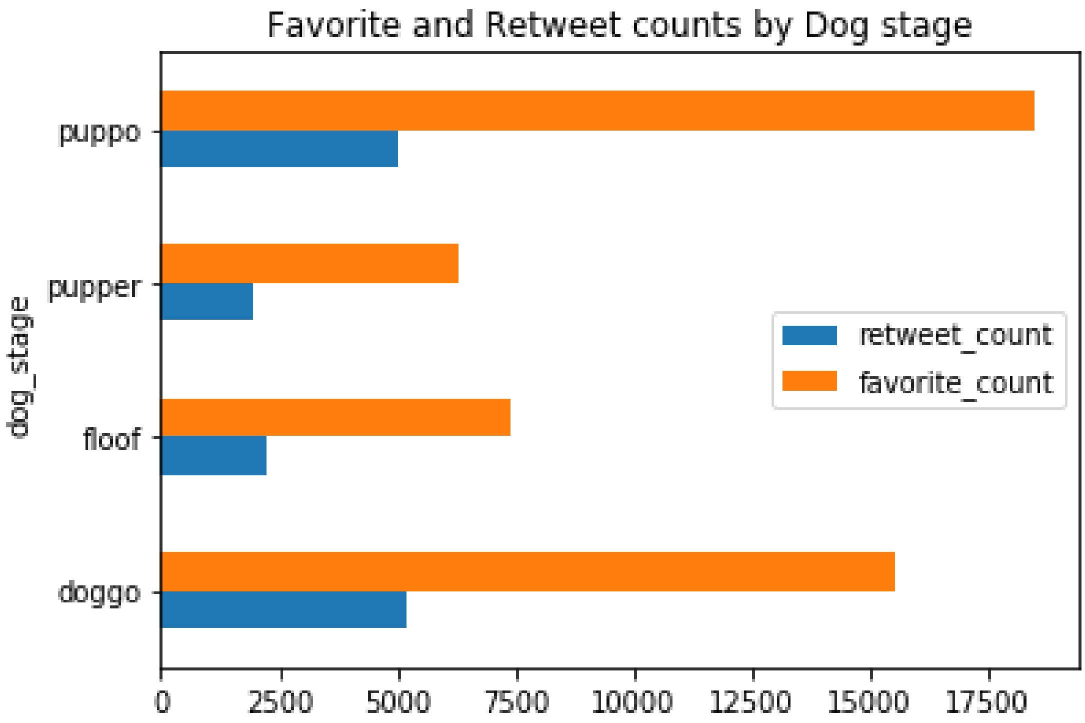
I analyzed the ratings from the 1962 original tweets in the dataset and this showed that 24.7% of the times, dogs were rated a 12 out of 10.

2. Visualization of the Relationship between Favorite count and Rating showed that overall, there is a positive correlation between the rating and the number of likes the tweet gets

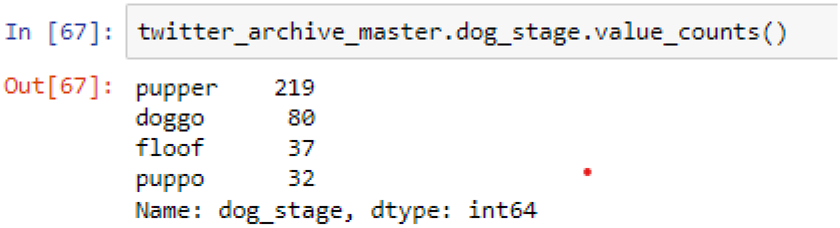


Dogs rated highly are generally more likely to be favorited than those with a lower rating with the favorite count peaking at about 140,000 at a rating of 13.

3. The visualization of the Favorite and Retweet counts by Dog stage shows doggo and puppo being the most common having similar average retweet counts but the puppo is the most liked exceeding the doggo in average favorite count by about 20%.



Of the 1962 tweets available after cleaning, 368 tweets included the dog’s stage: puppo, pupper, floof, doggo in their tweets. The count of each dog stage is as shown below:



```
In [67]: twitter_archive_master.dog_stage.value_counts()

Out[67]: pupper     219
         doggo       80
         floof       37
         puppo       32
         Name: dog_stage, dtype: int64
```

The details of the data wrangling process can be obtained in the wrangle-act notebook.