# WeRateDog Twitter Dataset – Act Report

Created by Justice Annan

### WeRateDogs Data

The WeRateDogs Twitter Archive Enhanced file contains data extracted from 2356 of the 5000+ tweets from the @dog_rates twitter account, posted between 15 November 2015 and 1 August, 2017. The data comprises of dog ratings that were taken from the test of the tweet along with the dog’s name and dog stage if present.

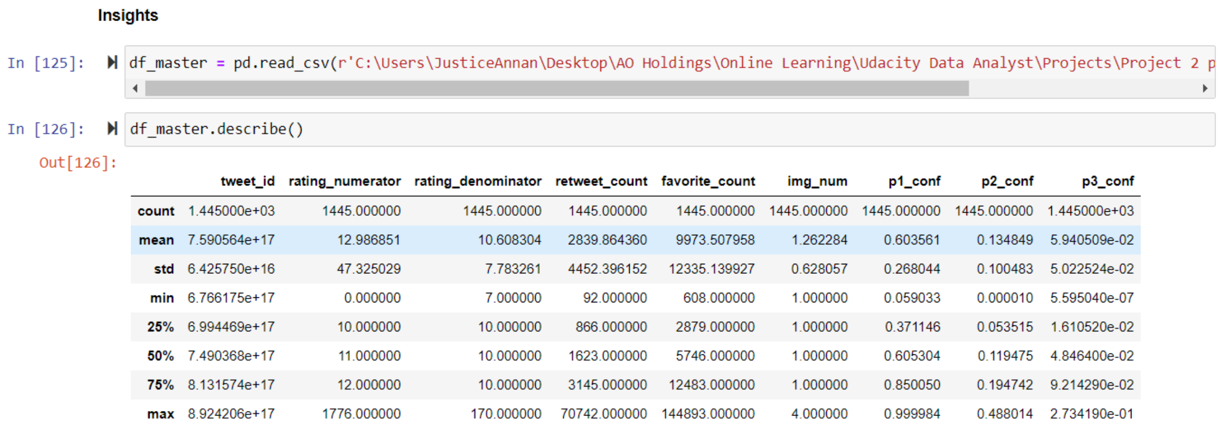
The retweet count and favorite count for each tweet were not included in the enhanced archive, and so I had to download this additional data from the twitter account using the tweet ID from the archive file.

Along with the Twitter data, I also downloaded an image predictions file from Udacity servers containing the image predictions for dog breeds.

### Wrangling Data

Before I could begin the analysis, the data had to be wrangled into shape to make it easier. I assessed the data both visually and programmatically for quality and tidiness. After cleaning many of the issues found during the assessment, there were about 1445 tweets with good quality data.

Insights



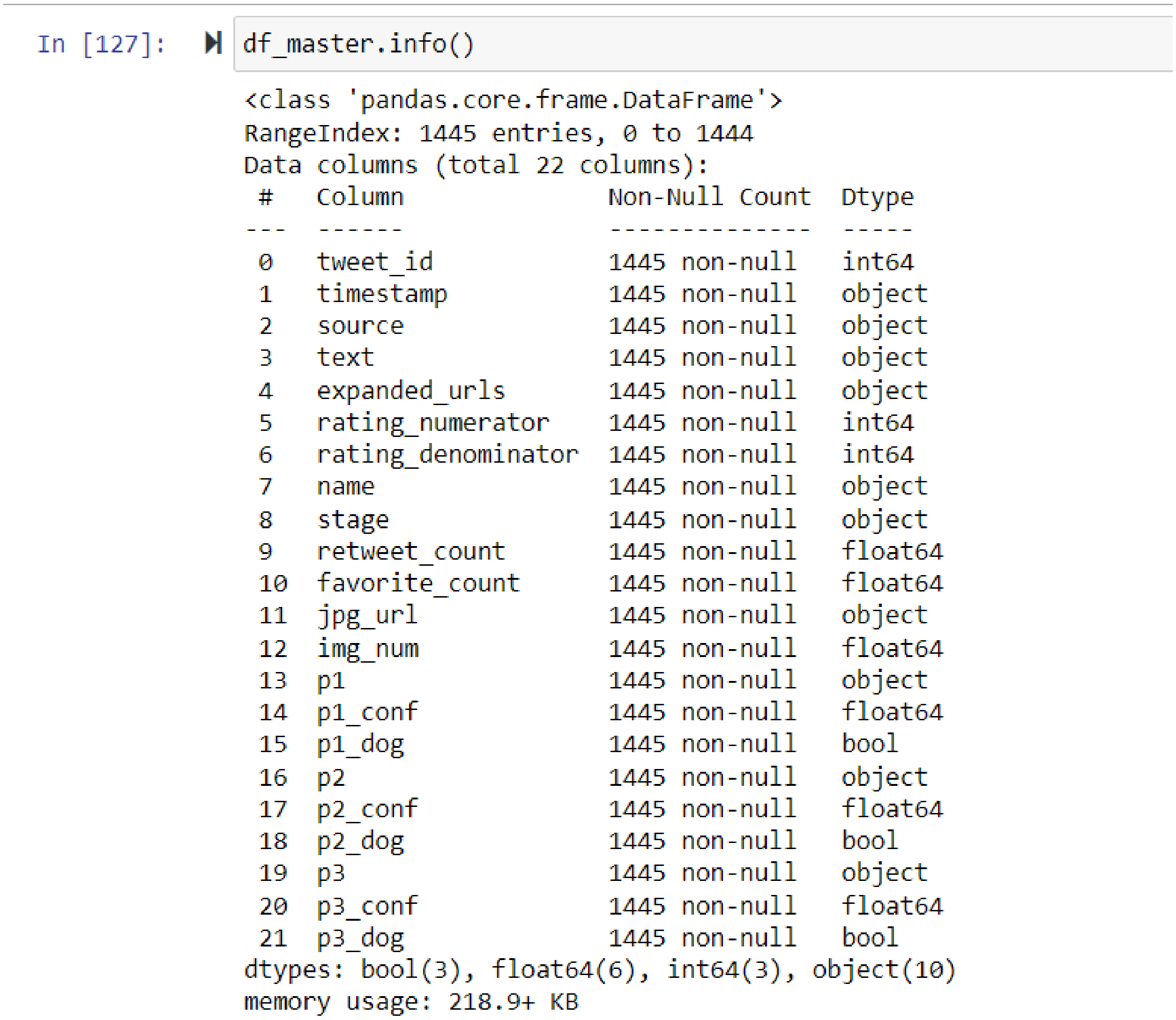
**Insights**

```
In [125]:  df_master = pd.read_csv(r'C:\Users\JusticeAnnan\Desktop\AO Holdings\Online Learning\Udacity Data Analyst\Projects\Project 2 p
```

```
In [126]:  df_master.describe()
```

Out[126]:

| | tweet_id | rating_numerator | rating_denominator | retweet_count | favorite_count | img_num | p1_conf | p2_conf | p3_conf |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 1.445000e+03 | 1445.000000 | 1445.000000 | 1445.000000 | 1445.000000 | 1445.000000 | 1445.000000 | 1445.000000 | 1.445000e+03 |
| **mean** | 7.590564e+17 | 12.986851 | 10.608304 | 2839.864360 | 9973.507958 | 1.262284 | 0.603561 | 0.134849 | 5.940509e-02 |
| **std** | 6.425750e+16 | 47.325029 | 7.783261 | 4452.396152 | 12335.139927 | 0.628057 | 0.268044 | 0.100483 | 5.022524e-02 |
| **min** | 6.766175e+17 | 0.000000 | 7.000000 | 92.000000 | 608.000000 | 1.000000 | 0.059033 | 0.000010 | 5.595040e-07 |
| **25%** | 6.994469e+17 | 10.000000 | 10.000000 | 866.000000 | 2879.000000 | 1.000000 | 0.371146 | 0.053515 | 1.610520e-02 |
| **50%** | 7.490368e+17 | 11.000000 | 10.000000 | 1623.000000 | 5746.000000 | 1.000000 | 0.605304 | 0.119475 | 4.846400e-02 |
| **75%** | 8.131574e+17 | 12.000000 | 10.000000 | 3145.000000 | 12483.000000 | 1.000000 | 0.850050 | 0.194742 | 9.214290e-02 |
| **max** | 8.924206e+17 | 1776.000000 | 170.000000 | 70742.000000 | 144893.000000 | 4.000000 | 0.999984 | 0.488014 | 2.734190e-01 |

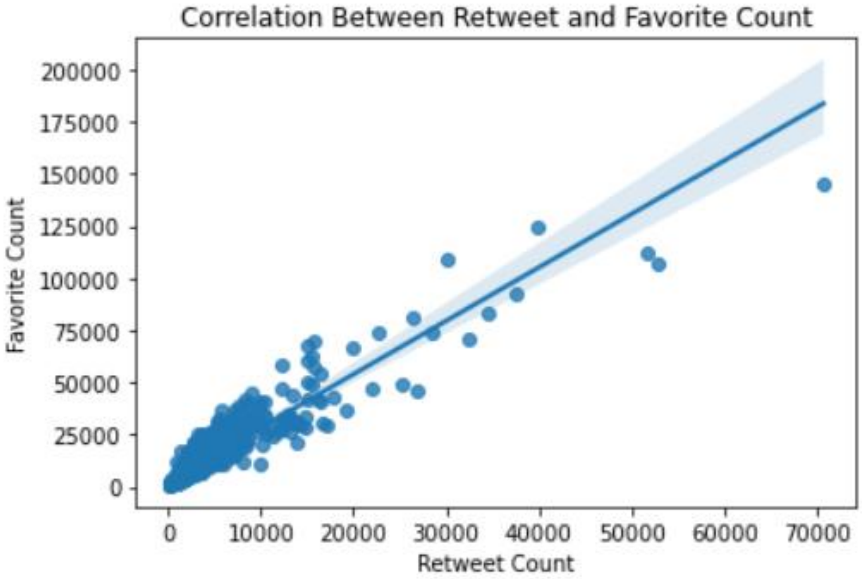


```
In [127]:  df_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1445 entries, 0 to 1444
Data columns (total 22 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   tweet_id            1445 non-null   int64  
 1   timestamp           1445 non-null   object 
 2   source              1445 non-null   object 
 3   text                1445 non-null   object 
 4   expanded_urls       1445 non-null   object 
 5   rating_numerator    1445 non-null   int64  
 6   rating_denominator  1445 non-null   int64  
 7   name                1445 non-null   object 
 8   stage               1445 non-null   object 
 9   retweet_count       1445 non-null   float64
 10  favorite_count      1445 non-null   float64
 11  jpg_url             1445 non-null   object 
 12  img_num             1445 non-null   float64
 13  p1                  1445 non-null   object 
 14  p1_conf             1445 non-null   float64
 15  p1_dog              1445 non-null   bool   
 16  p2                  1445 non-null   object 
 17  p2_conf             1445 non-null   float64
 18  p2_dog              1445 non-null   bool   
 19  p3                  1445 non-null   object 
 20  p3_conf             1445 non-null   float64
 21  p3_dog              1445 non-null   bool   
dtypes: bool(3), float64(6), int64(3), object(10)
memory usage: 218.9+ KB
```

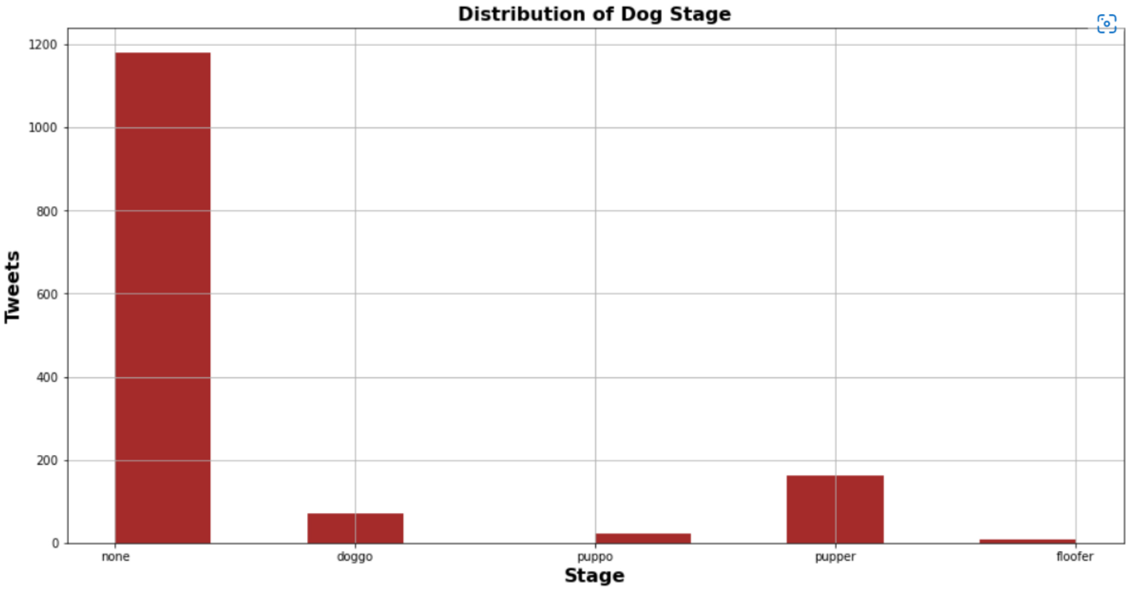
- The minimum retweet count is 7, mean is 2839.86 and maximum is 70742
- The minimum favorite count is 608, mean is 9973.50 and maximum is 144893
- All tweets have higher favorite count than retween count
- The master dataset has 3 boolean, 6 floats, 3 integers and 10 objects datatypes
- The cleaned dataset had a total of 1445 entries and 22 columns.

What is the correlation between Retweet and Favorite Count?



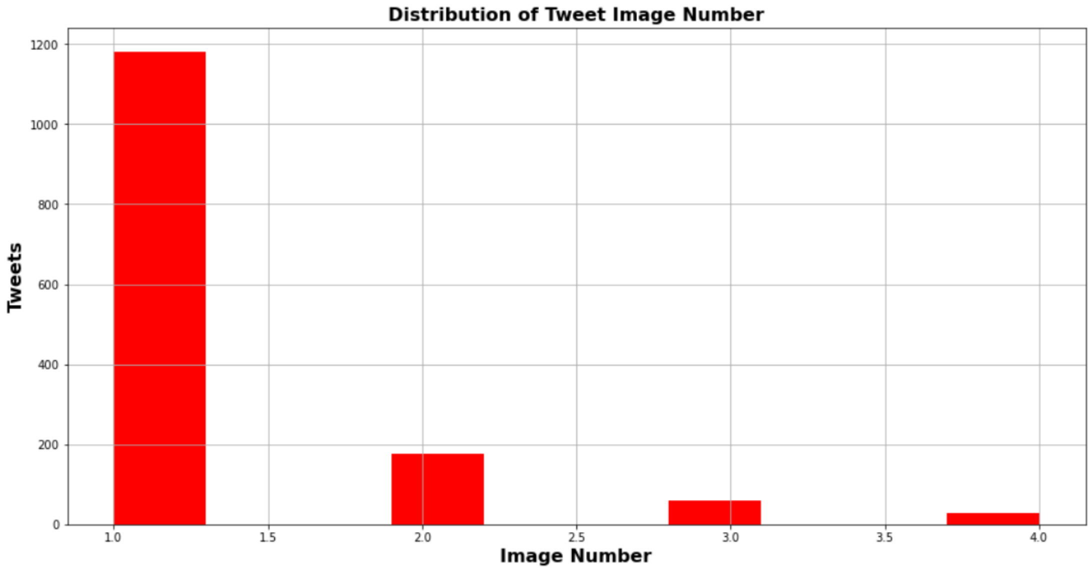
From the Correlation between retweet and favorite count, there is a linear relationship between the two variables. This indicate that there is a very positive correlation between them.

What is the most popular dog stage?



From the distribution diagram, one could clearly see that most tweets were without a dog stage. Irrespective of this, **pupper** stood out as the most popular dog stage.

What is the highest image number?



The distribution of tweet image number shows that image number 1 has the highest counts