# Act Report

## Udacity

## Data Analyst Nanodegree Program

## Tariq Ibrahim Alhajri

**For @Dograte**s **tweets from 2017 onwards.**

First, let’s describe the datasets we have, all taken from Udacity: -

### Twitter_archive dataset:

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (doggo, puppo,..etc) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

### Tweet_json dataset:

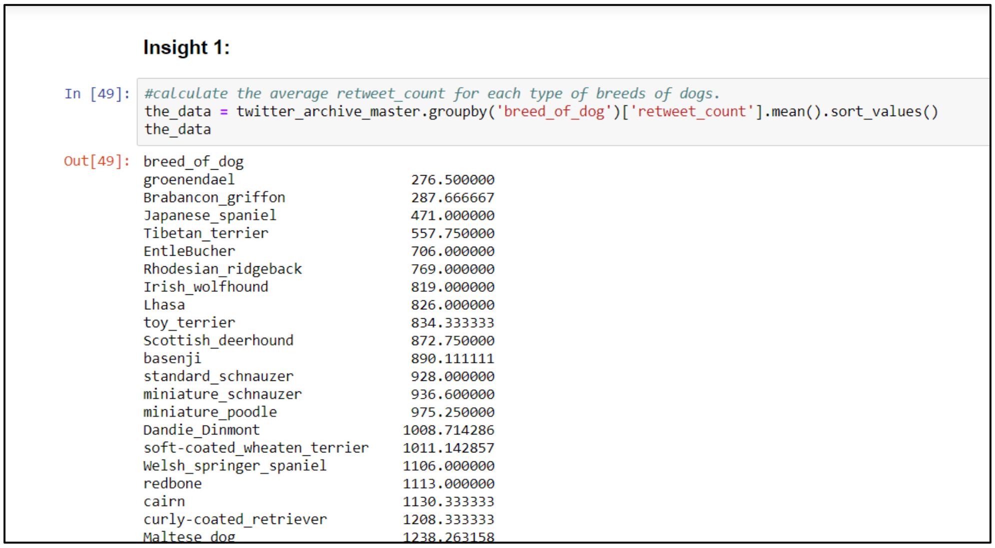
Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API. Well, "anyone" who has access to data for the 3000 most recent tweets, at least. But you, because you have the WeRateDogs Twitter archive and specifically the tweet IDs within it, can gather this data for all 5000+. And guess what? You're going to query Twitter's API to gather this valuable data.

### Image_predictions dataset:

One more cool thing: I ran every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs*. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

After cleaning the datasets, and then merge it, here is my analysis conclusion:-

**Insight 1)**



**Insight 1:**

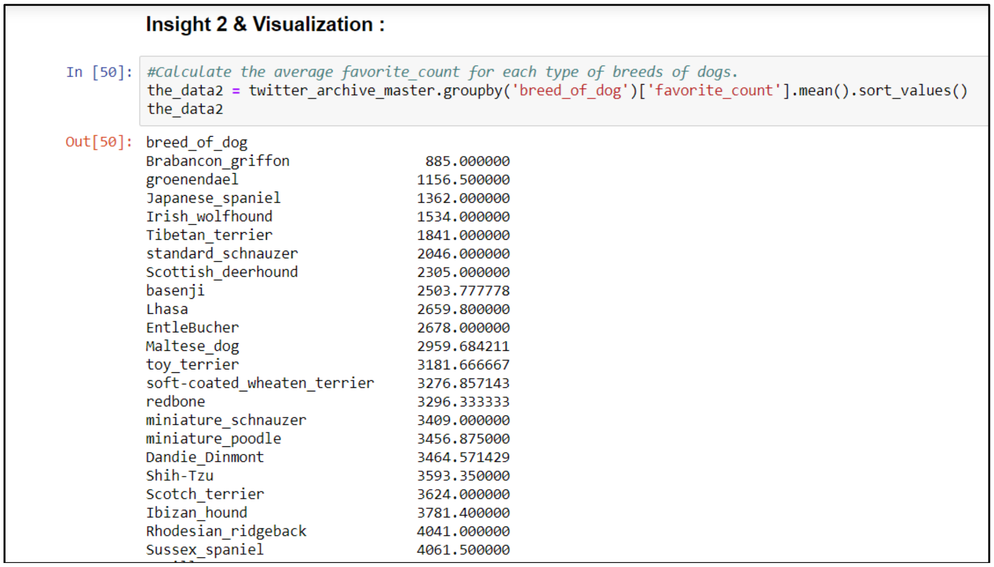
```
In [49]: #calculate the average retweet_count for each type of breeds of dogs.
         the_data = twitter_archive_master.groupby('breed_of_dog')['retweet_count'].mean().sort_values()
         the_data
```

```
Out[49]: breed_of_dog
         groenendael                    276.500000
         Brabancon_griffon              287.666667
         Japanese_spaniel               471.000000
         Tibetan_terrier                557.750000
         EntleBucher                    706.000000
         Rhodesian_ridgeback            769.000000
         Irish_wolfhound                819.000000
         Lhasa                          826.000000
         toy_terrier                    834.333333
         Scottish_deerhound             872.750000
         basenji                        890.111111
         standard_schnauzer             928.000000
         miniature_schnauzer            936.600000
         miniature_poodle               975.250000
         Dandie_Dinmont                1008.714286
         soft-coated_wheaten_terrier   1011.142857
         Welsh_springer_spaniel        1106.000000
         redbone                       1113.000000
         cairn                         1130.333333
         curly-coated_retriever        1208.333333
         Maltese_dog                   1238.263158
```

Figure 1 (mean of retweets for each type of breed)

- The breed of dogs that got the highest average in retweets is Bedlington_terrier with 7510 retweet in average.
- Whereas the breed of dog that got the lowest retweet average is groenendael with 276 retweet in average.

**Insight 2)**

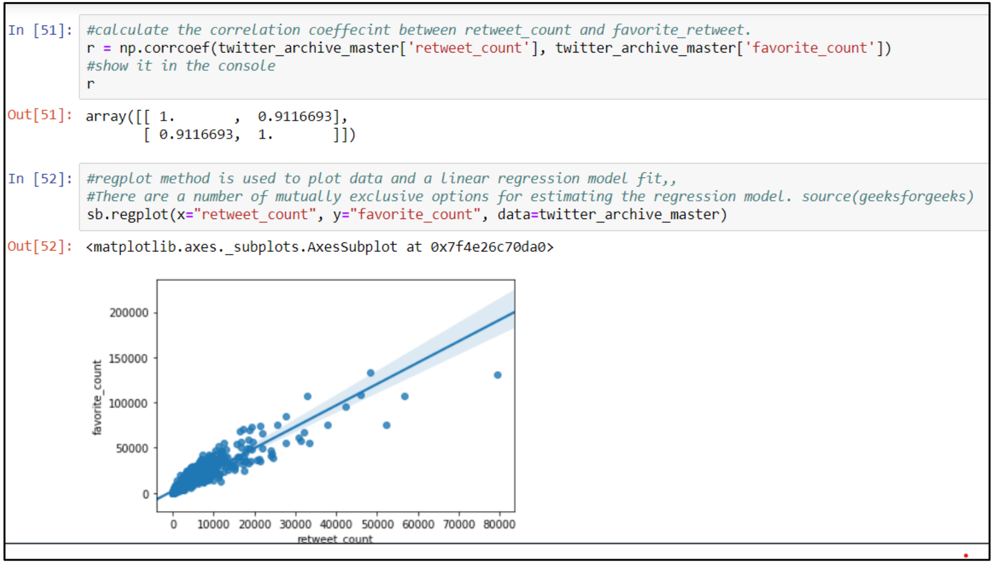


**Insight 2 & Visualization :**

```
In [50]: #Calculate the average favorite_count for each type of breeds of dogs.
         the_data2 = twitter_archive_master.groupby('breed_of_dog')['favorite_count'].mean().sort_values()
         the_data2

Out[50]: breed_of_dog
         Brabancon_griffon                 885.000000
         groenendael                      1156.500000
         Japanese_spaniel                 1362.000000
         Irish_wolfhound                  1534.000000
         Tibetan_terrier                  1841.000000
         standard_schnauzer               2046.000000
         Scottish_deerhound               2305.000000
         basenji                          2503.777778
         Lhasa                            2659.800000
         EntleBucher                      2678.000000
         Maltese_dog                      2959.684211
         toy_terrier                      3181.666667
         soft-coated_wheaten_terrier      3276.857143
         redbone                          3296.333333
         miniature_schnauzer              3409.000000
         miniature_poodle                 3456.875000
         Dandie_Dinmont                   3464.571429
         Shih-Tzu                         3593.350000
         Scotch_terrier                   3624.000000
         Ibizan_hound                     3781.400000
         Rhodesian_ridgeback              4041.000000
         Sussex_spaniel                   4061.500000
```

Figure 2 (mean of favorites for each type of breed)

- The breed of dogs that got the highest average in favorites is Saluki with 24060 favorites in average.
- Whereas the breed of dog that got the lowest retweet average is brabancon_griffon with 885 retweet in average.
- it appears that there is a strong relationship between retweet_count and favorite_count for the tweet since that the breed with least average in retweets came second to last for favorite_count and same for the highest average in tweets came the second highest average in favorite count. will try to confirm that in the next two cells.

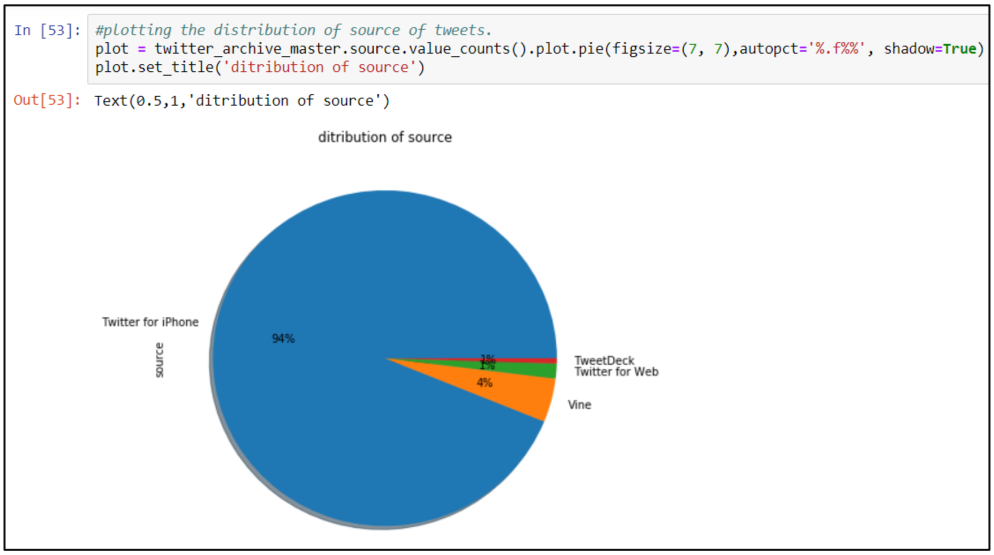


```
In [51]: #calculate the correlation coeffecint between retweet_count and favorite_retweet.
         r = np.corrcoef(twitter_archive_master['retweet_count'], twitter_archive_master['favorite_count'])
         #show it in the console
         r

Out[51]: array([[ 1.       ,  0.9116693],
                [ 0.9116693,  1.       ]])

In [52]: #regplot method is used to plot data and a linear regression model fit,,
         #There are a number of mutually exclusive options for estimating the regression model. source(geeksforgeeks)
         sb.regplot(x="retweet_count", y="favorite_count", data=twitter_archive_master)

Out[52]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4e26c70da0>
```

Figure 3 (linear regression model showing the relationship between retweets and favorites)

- With correlation coefficient equals to 0.91 and a positive strong relationship between retweet_count and favorite_count, there is definitely a relationship between these two column.

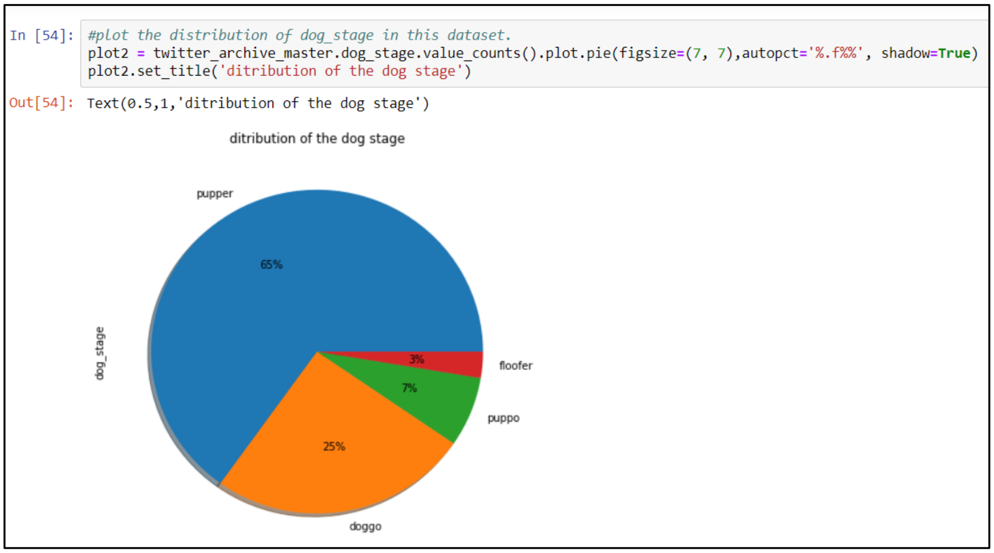
## Insight 3 & Visualization)



```
In [53]: #plotting the distribution of source of tweets.
         plot = twitter_archive_master.source.value_counts().plot.pie(figsize=(7, 7),autopct='%.f%%', shadow=True)
         plot.set_title('ditribution of source')

Out[53]: Text(0.5,1,'ditribution of source')
```

Figure 4 (distribution of source)

- As we can see in pie plot above, 94% of the tweets in this dataset came from twitter for iPhone, which is an indicator that twitter in mobiles in general is where most of users uses the application and perhaps it is easier.

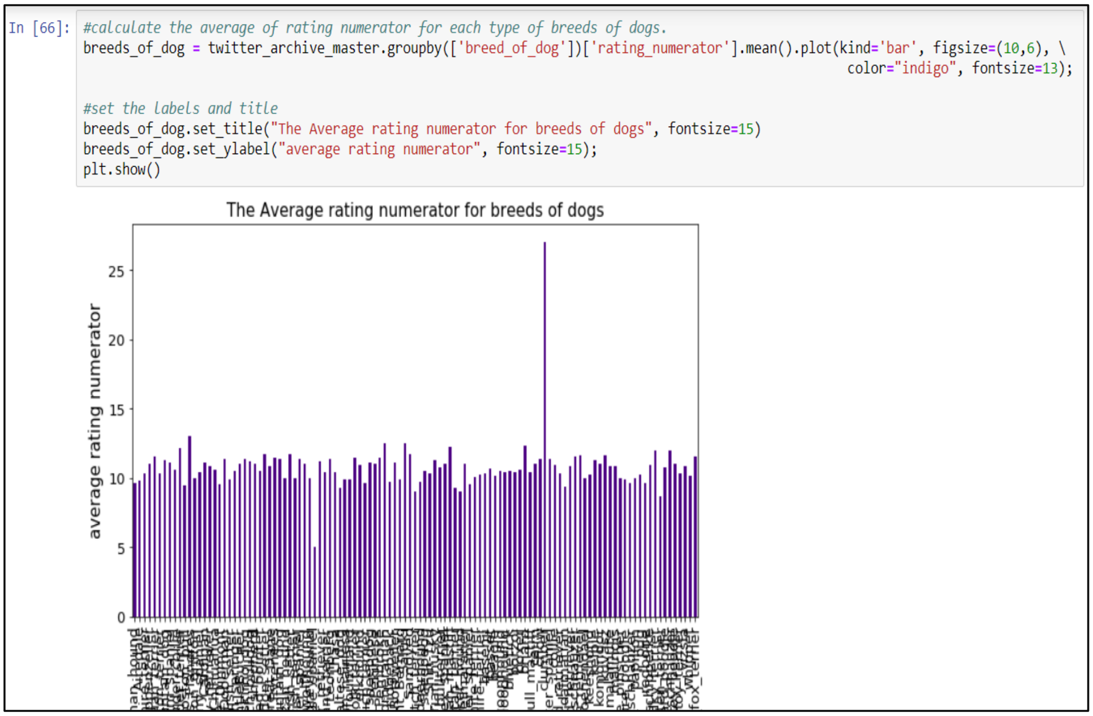


```
In [54]: #plot the distribution of dog_stage in this dataset.
         plot2 = twitter_archive_master.dog_stage.value_counts().plot.pie(figsize=(7, 7),autopct='%.f%%', shadow=True)
         plot2.set_title('ditribution of the dog stage')

Out[54]: Text(0.5,1,'ditribution of the dog stage')
```

Figure 5 (distribution of the dog stage in the dataset)

- 65% of dogs are pupper dogs, and 25% are doggo dogs.

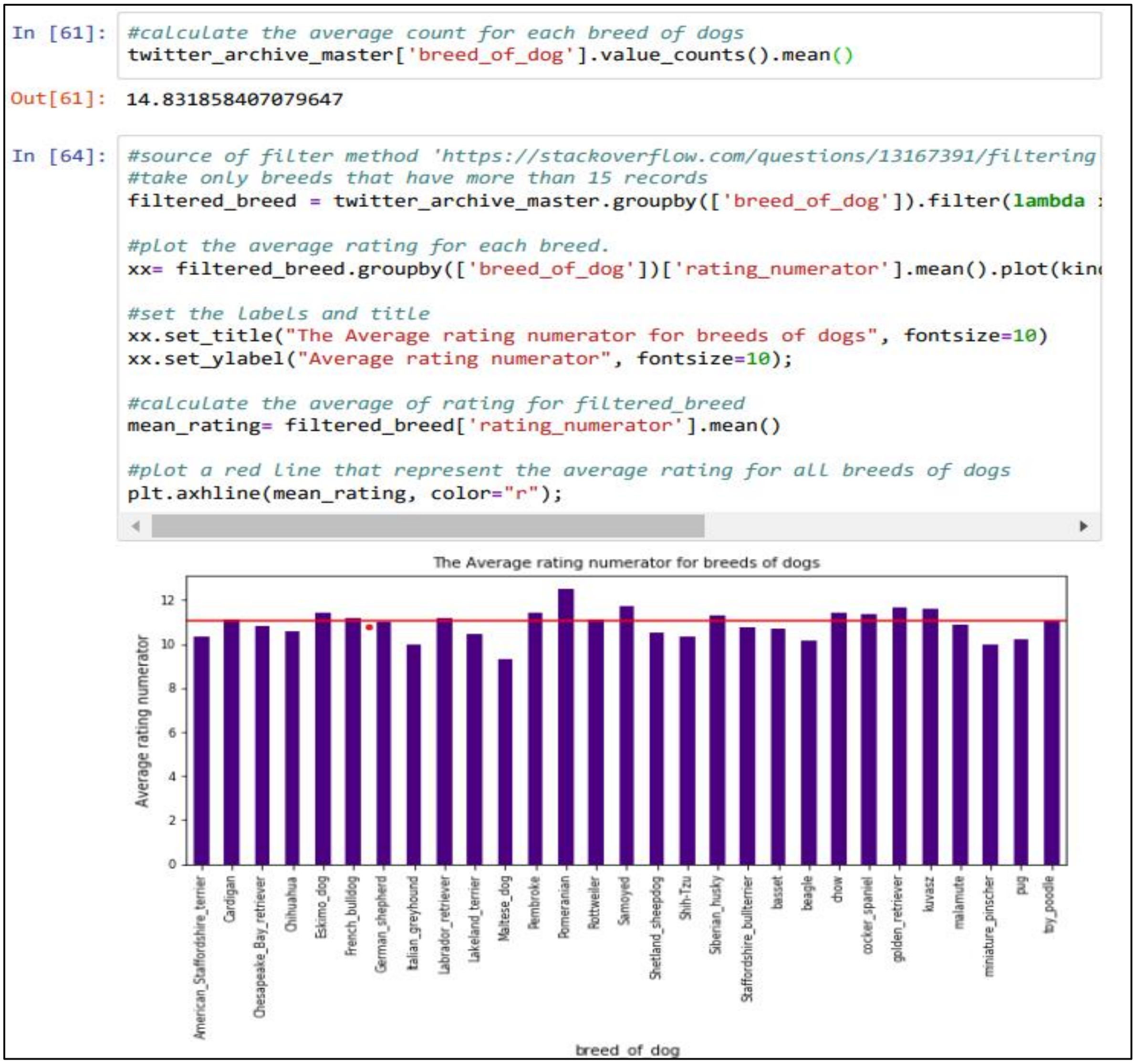
## Insight 4 & Visualization)



```
In [66]: #calculate the average of rating numerator for each type of breeds of dogs.
         breeds_of_dog = twitter_archive_master.groupby(['breed_of_dog'])['rating_numerator'].mean().plot(kind='bar', figsize=(10,6), \
                                                                                                            color="indigo", fontsize=13);

         #set the labels and title
         breeds_of_dog.set_title("The Average rating numerator for breeds of dogs", fontsize=15)
         breeds_of_dog.set_ylabel("average rating numerator", fontsize=15);
         plt.show()
```

Figure 6 (rating average for each type of breeds )

- Since we have so many breeds, we need to simplify the graph in order to see it probaly and derive some information.

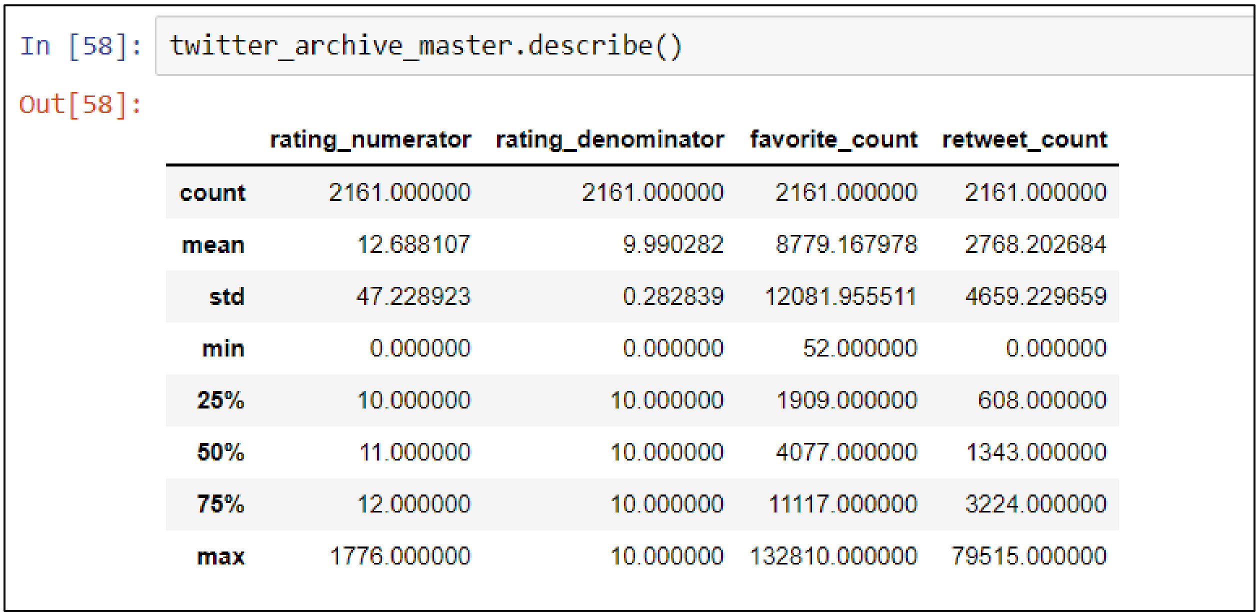


```
In [61]: #calculate the average count for each breed of dogs
         twitter_archive_master['breed_of_dog'].value_counts().mean()

Out[61]: 14.831858407079647

In [64]: #source of filter method 'https://stackoverflow.com/questions/13167391/filtering
         #take only breeds that have more than 15 records
         filtered_breed = twitter_archive_master.groupby(['breed_of_dog']).filter(lambda :

         #plot the average rating for each breed.
         xx= filtered_breed.groupby(['breed_of_dog'])['rating_numerator'].mean().plot(kin

         #set the labels and title
         xx.set_title("The Average rating numerator for breeds of dogs", fontsize=10)
         xx.set_ylabel("Average rating numerator", fontsize=10);

         #calculate the average of rating for filtered_breed
         mean_rating= filtered_breed['rating_numerator'].mean()

         #plot a red line that represent the average rating for all breeds of dogs
         plt.axhline(mean_rating, color="r");
```

Figure 7 (rating average for each type of breeds )

- After filtering , we can clearly see the average rating for each type of breeds.
- Also, most of breeds are close to the average with an exception for Maltese_dog.

## Insight 5)



```
In [58]: twitter_archive_master.describe()
```

Out[58]:

| | rating_numerator | rating_denominator | favorite_count | retweet_count |
|---|---|---|---|---|
| **count** | 2161.000000 | 2161.000000 | 2161.000000 | 2161.000000 |
| **mean** | 12.688107 | 9.990282 | 8779.167978 | 2768.202684 |
| **std** | 47.228923 | 0.282839 | 12081.955511 | 4659.229659 |
| **min** | 0.000000 | 0.000000 | 52.000000 | 0.000000 |
| **25%** | 10.000000 | 10.000000 | 1909.000000 | 608.000000 |
| **50%** | 11.000000 | 10.000000 | 4077.000000 | 1343.000000 |
| **75%** | 12.000000 | 10.000000 | 11117.000000 | 3224.000000 |
| **max** | 1776.000000 | 10.000000 | 132810.000000 | 79515.000000 |

Figure 7 (describe method’s output for our dataset)

- The average favorites is 8779 whereas the retweets is 2768 which shows that users tend to press the like button more often than the retweets, which is obvious since it an account for dog’s fans and like button is a gesture to indicate you like the dog.
- The average rating numerator is 12.68/10.