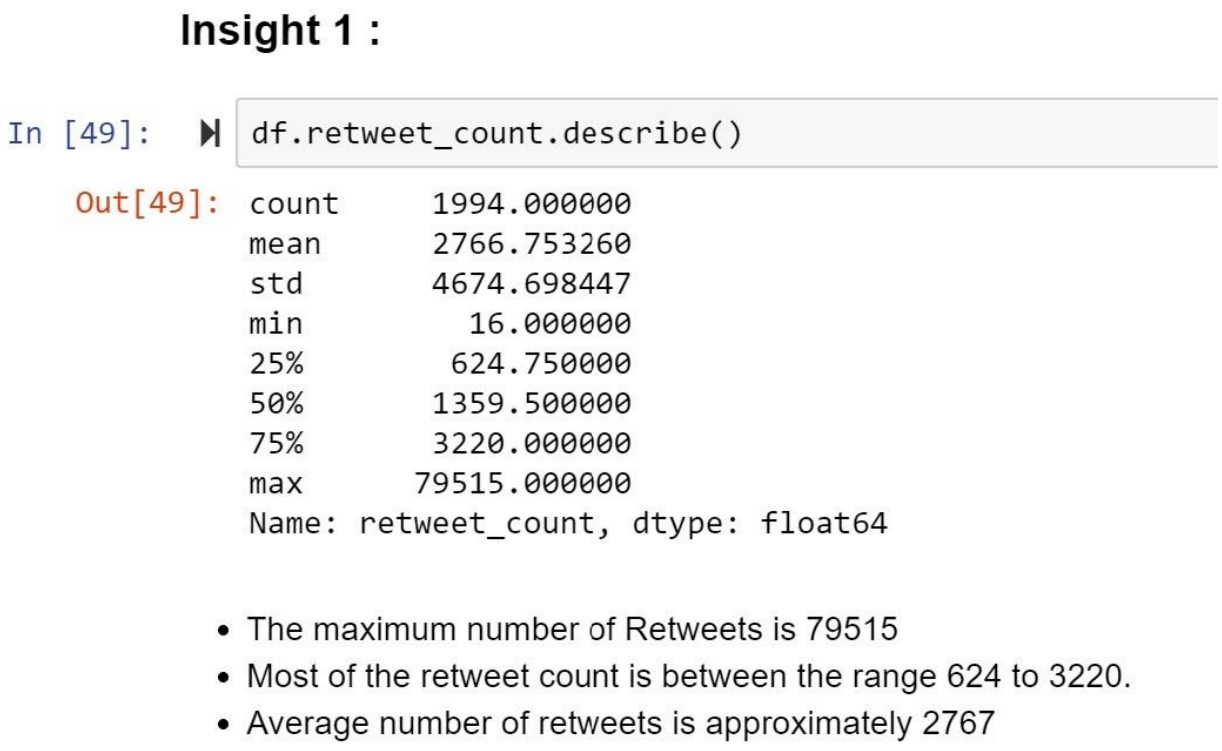
Act Report

-S.Sanjay(26/05/2020)

Once the re-iterative process of data wrangling was completed it was time to act upon the master dataset formed. In this report, I would analyze the effect of wrangling the data on each insight and visualization made at the end of wrangle_act.ipynb.

Insight 1 :

Descriptive factors(Min, Max, and IQR) of retweet count was interpreted. This was made possible only by querying the Twitter API in the Gather process. And then we merged the gathered data with the initial data archive we had by making an inner join based on the tweet ids.



Insight 1 :

```
In [49]: df.retweet_count.describe()

Out[49]: count     1994.000000
         mean      2766.753260
         std       4674.698447
         min         16.000000
         25%        624.750000
         50%       1359.500000
         75%       3220.000000
         max      79515.000000
         Name: retweet_count, dtype: float64
```

- The maximum number of Retweets is 79515
- Most of the retweet count is between the range 624 to 3220.
- Average number of retweets is approximately 2767

Insight 2 :

The counts of various dog stages were calculated and we realized that most dog stages are not identified yet or the data is missing. This counting is possible only due to the melting of four separate columns with high none values to one efficient dog stage column.

Insight 2 :

```
In [50]: df.dog_stage.value_counts()

Out[50]: Stage not known yet    1688
         pupper                  203
         doggo                    63
         puppo                    22
         doggo, pupper             9
         floofer                   7
         doggo, floofer            1
         doggo, puppo              1
         Name: dog_stage, dtype: int64
```

- Most of dog stages haven't been identified yet or there is a lack of data.

Insight 3 :

The various unique clean sources of dog image data were identified. This data was also wrangled to remove the unnecessary HTML tags and also sources of retweets(Vine video) were also excluded when the retweets were dropped.

Insight 3 :

```
In [51]: df.source.unique()

Out[51]: array(['"http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone',
                '"http://twitter.com" rel="nofollow">Twitter Web Client',
                '"https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck'], dtype=object)
```

- The above mentioned are the three unique sources of images(Iphone,Web and Tweetdeck)

Insight 4 and Visualisation 1 :

The dog ratings were generally really high and the analyses prove that “They’re good dogs Brent” with most dog ratings exceeding 100% (i.e greater than 10 in a 10 point scale). This insight is clearly visible after we wrangled the data to standardize the dog rating scores and is also visualized by a pie chart using matplotlib under two classifications :

1) High Rated Dogs (Rating greater than or equal to 10/10)
2) Low Rated Dogs (Rating lesser than 10/10)

Insight 4 :

In [52]:
```
h1 = df.query('rating_numerator >= 10').tweet_id.count()
h2 = df.query('rating_numerator < 10').tweet_id.count()
h1>h2
```

Out[52]: True

- There are more highly-rated dogs than the lower rated ones.

Visualisation 1 :

Note :

- This piechart visualisation is based on the standardised rating scores which we cleaned earlier.

In [53]:
```
import matplotlib.pyplot as plt
%matplotlib inline
plt.pie([h1,h2], labels = ['High Rated dogs','Low Rated dogs']);
```