Introduction

wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information.

Part Two: Data visualization:

I have some research questions I need get the answers for:

- **What is the most and the least favorite dog stage according to the dataset?**
- **What are the top 10 favorite dog type according to the dataset?**
- **What are the least 10 favorite dog type according to the dataset?**
- **What are the Top 10 favorite count in total according to the dog type?**
- **What are the least 10 retweet count in total according to the dog type?**
- **What are the Top 10 retweet count in average according to the dog type?**
- **What are the Top 10 favorite count in average according to the dog type?**
- **What are the Top 10 rate in average according to the dog type?**
- **What are the least 10 rate in average according to the dog type?**
- **What are the different statistics for dog_stage according to rate?**
- **What is the type of correlation between favorite_count And retweet_count?**
- **What is the distribution of dogs_stage?**
- **Is there a relation between rate and tweet count?**
- **Is there a relation between rate and favorite count?**

import all package I need

In [559]:

```
# package to be used in the project...
import pandas as pd
import numpy as np
import requests
import os
from PIL import Image
from io import BytesIO
import tweepy
from tweepy import OAuthHandler
import json
from timeit import default_timer as timer
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Insights:

1- The most and the least favorite dog stage according to the dataset:

In [476]:

```
df_master['dogs_stage'].value_counts()
```

Out[476]:

```
pupper      169
doggo        63
puppo        21
floofer       7
```

```
Name: dogs_stage, dtype: int64
```

I found that the most favorite dog stage is the `pupper` with count `169`, that was expected as `pupper` is young, its size small, and better for families.

I found also that the least favorite dog stage is the `floofer (dogs with fur)` , it's just count `7` for our dataset.

2- The top 10 favorite dog type according to the dataset:

In [508]:

```
df_master['dog_type'].value_counts()[0:10]
```

Out[508]:

```
golden_retriever      157
Labrador_retriever    106
Pembroke               95
Chihuahua              91
pug                    63
toy_poodle             50
chow                   48
Pomeranian             42
Samoyed                41
malamute               33
Name: dog_type, dtype: int64
```

I found that the most favorite dog type is the `golden_retriever` with count `175`, and it followed by `Labrador_retriever` with Count `106`

that was expected as `golden_retriever` and `Labrador_retriever` are better for families.

3- The least 10 favorite dog type according to the dataset:

In [506]:

```
df_master['dog_type'].value_counts()[-11:-1]
```

Out[506]:

```
Sussex_spaniel              2
Australian_terrier          2
wire-haired_fox_terrier     2
Bouvier_des_Flandres        1
Scotch_terrier              1
Irish_wolfhound             1
silky_terrier               1
Japanese_spaniel            1
standard_schnauzer          1
clumber                     1
Name: dog_type, dtype: int64
```

As shown above the least 10 favorite dog type.

4- The Top 10 favorite count in total according to the dog type:

In [516]:

```
df_master.groupby(['dog_type'])['favorite_count'].sum().sort_values(ascending=False).hea
d(10)
```

Out[516]:

```
dog_type
golden_retriever      1769933
Labrador_retriever    1110845
Pembroke               953500
Chihuahua              707195
French_bulldog         553362
Samoyed                507433
chow                   410797
cocker_spaniel         371984
pug                    343564
malamute               321318
Name: favorite_count, dtype: int64
```

As Shown above the top 10 favorite count according to dog type.

`golden_retriever` ***takes the highest favorite_count in total =*** `1769933`***.***

Followed by `Labrador_retriever` ***with favorite count in total =*** `1110845` ***.***

5- The Top 10 retweet count in total according to the dog type:

In [528]:

```
df_master.groupby(['dog_type'])['retweet_count'].sum().sort_values(ascending=False).head
(10)
```

Out[528]:

```
dog_type
golden_retriever      508771
Labrador_retriever    341225
Pembroke              253051
Chihuahua             225896
Samoyed               166496
French_bulldog        141189
cocker_spaniel        127124
chow                  114753
pug                   101287
Pomeranian             98175
Name: retweet_count, dtype: int64
```

As Shown above the top 10 retweet count according to dog_type.

`golden_retriever` ***takes the highest retweet_count in total =*** `508771`***.***

Followed by `Labrador_retriever` ***with retweet count in total =*** `341225` ***.***

In [530]:

```
df_master.groupby(['dog_type'])['favorite_count','retweet_count'].mean().\
sort_values(by='favorite_count',ascending=False).head(10)
```

Out[530]:

dog_type	favorite_count	retweet_count
Bedlington_terrier	22731.500000	7165.000000
Saluki	21858.500000	4414.500000
French_bulldog	18445.400000	4706.300000
Bouvier_des_Flandres	16179.000000	3820.000000
Afghan_hound	15501.333333	5104.666667

dog_type	favorite_count	retweet_count
black-and-tan_coonhound	15398.500000	3550.500000
flat-coated_retriever	15229.500000	3953.500000
Irish_water_spaniel	14712.333333	3873.000000
Leonberg	13350.000000	3314.000000
whippet	13279.818182	4393.363636

6- The Top 10 retweet count in average according to the dog type:

In [539]:

```
df_master.groupby(['dog_type'])['retweet_count'].mean().sort_values(ascending=False).hea
d(10)
```

Out[539]:

```
dog_type
Bedlington_terrier    7165.000000
Afghan_hound          5104.666667
standard_poodle       4770.272727
French_bulldog        4706.300000
English_springer      4688.000000
Saluki                4414.500000
whippet               4393.363636
cocker_spaniel        4237.466667
Eskimo_dog            4146.772727
Samoyed               4060.878049
Name: retweet_count, dtype: float64
```

As Shown above the top 10 retweet count according to dog_type.

We can notice here in average the dogtype differed from the retweet count in average , as the number of dog type affect the result in average.

The more the count of dog type , the less the retweet count in average.

`Bedlington_terrier` ***takes the highest retweet_count in average =*** `7165`***.***

Followed by `Afghan_hound` ***with retweet count in average =*** `5104` ***.***

7- The Top 10 favorite count in average according to the dog type:

In [540]:

```
df_master.groupby(['dog_type'])['favorite_count'].mean().sort_values(ascending=False).he
ad(10)
```

Out[540]:

```
dog_type
Bedlington_terrier          22731.500000
Saluki                      21858.500000
French_bulldog              18445.400000
Bouvier_des_Flandres        16179.000000
Afghan_hound                15501.333333
black-and-tan_coonhound     15398.500000
flat-coated_retriever       15229.500000
Irish_water_spaniel         14712.333333
Leonberg                    13350.000000
whippet                     13279.818182
Name: favorite_count, dtype: float64
```

As Shown above the top 10 favorite count according to dog_type.

We can notice here in average the dogtype differed from the favorite count in average, as the number of dog type affect the result in average.

The more the count of dog type , the less the favorite count in average.

`Bedlington_terrier` ***takes the highest favorite_count in average=*** `22731.5`.

Followed by `Afghan_hound` ***with favorite count in average =*** `21858.5` .

8- The Top 10 rate in average according to the dog type:

In [556]:

```
df_master.groupby('dog_type')['rate'].mean().sort_values(ascending=False).head(10)
```

Out[556]:

```
dog_type
Bouvier_des_Flandres    13.000000
Saluki                  12.500000
briard                  12.333333
Tibetan_mastiff         12.250000
Irish_setter            12.200000
Border_terrier          12.142857
standard_schnauzer      12.000000
silky_terrier           12.000000
clumber                 12.000000
Gordon_setter           11.750000
Name: rate, dtype: float64
```

As Shown above the top 10 rate in average according to dog_type.

`Bouvier_des_Flandres` ***takes the highest rate in average =*** `13/10`.

Followed by `Saluki` ***with rate in average =*** `12.5/10` .

9- The least 10 rate in average according to the dog type:

In [558]:

```
df_master.groupby('dog_type')['rate'].mean().sort_values(ascending=True).head(10)
```

Out[558]:

```
dog_type
Japanese_spaniel               5.000000
soft-coated_wheaten_terrier    8.866667
Scotch_terrier                 9.000000
Walker_hound                   9.000000
Tibetan_terrier                9.250000
dalmatian                      9.333333
Boston_bull                    9.416667
Welsh_springer_spaniel         9.500000
Dandie_Dinmont                 9.571429
miniature_schnauzer            9.600000
Name: rate, dtype: float64
```

As Shown above the least 10 rate in average according to dog_type.

`Japanese_spaniel` ***takes the lowest rate in average =*** `5/10`.

Followed by `soft-coated_wheaten_terrier` ***with rate in average =*** `8.87/10` .

11- Calculate different statistic for dog_stage according to rate:

In [552]:

```
df_master.groupby('dogs_stage')['rate'].describe()
```

Out[552]:

dogs_stage	count	mean	std	min	25%	50%	75%	max
doggo	63.0	11.809524	1.564290	5.0	11.0	12.0	13.0	14.0
floofer	7.0	12.000000	1.154701	10.0	11.5	12.0	13.0	13.0
pupper	169.0	10.887574	1.424413	7.0	10.0	11.0	12.0	14.0
puppo	21.0	11.952381	1.321975	9.0	11.0	12.0	13.0	14.0

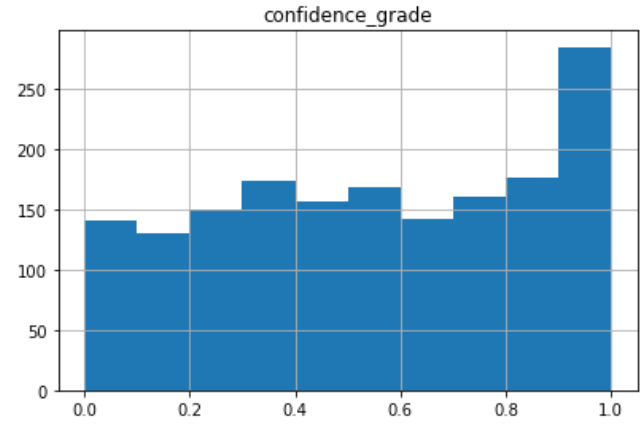
We can notice the following:

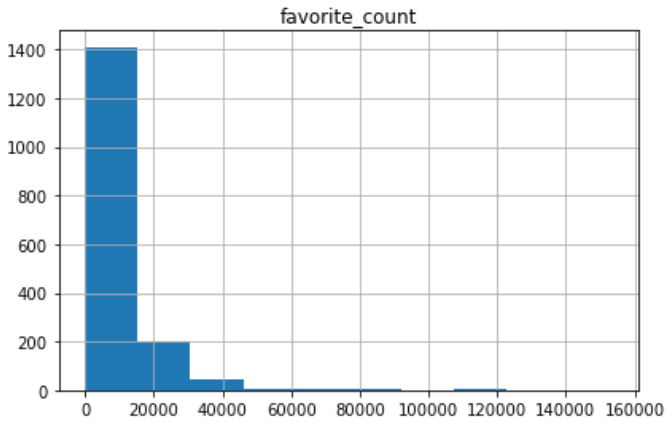
- **Floofer has the highest rate in average (sure it affected by count of floofer (`7`).**
- **Doggo has the minimum rate in average (equal to `5`).**
- **Each of Doggo, Pupper, and Puppo has maximum rate in average equal to `14` .**
- **75% of Doggo, Floofer, and Puppo has rate in average equal to `13` .**
- **50% of Doggo, Floofer, and Puppo has rate in average equal to `12` .**
- **25% of Doggo, Floofer, and Puppo has rate in average equal to `11` .**

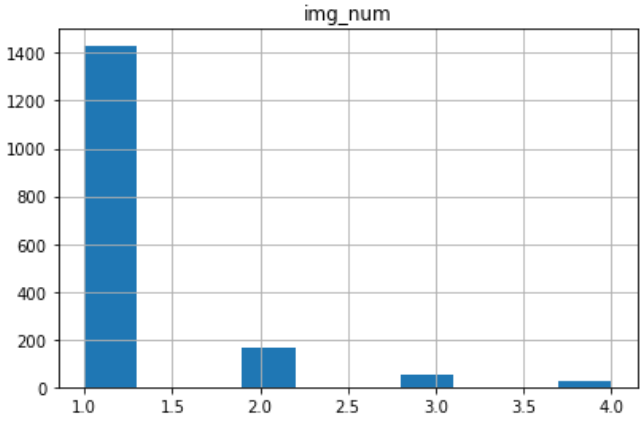
Visualization:

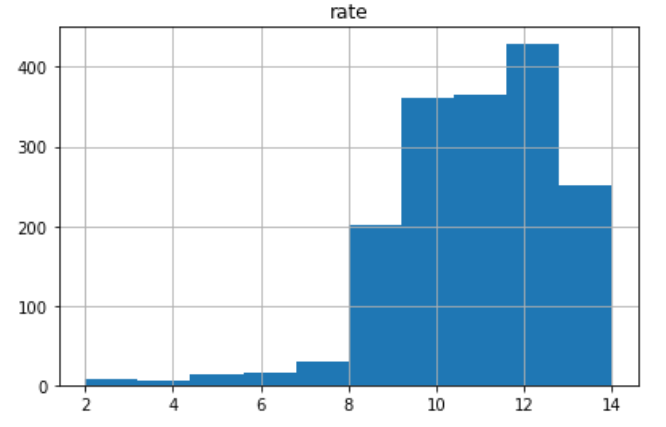
In [564]:

```
df_master.hist(figsize=(15,15));
```

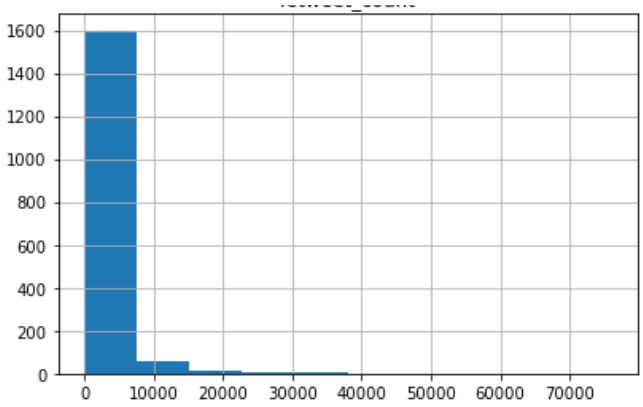








retweet_count



We notice the following:

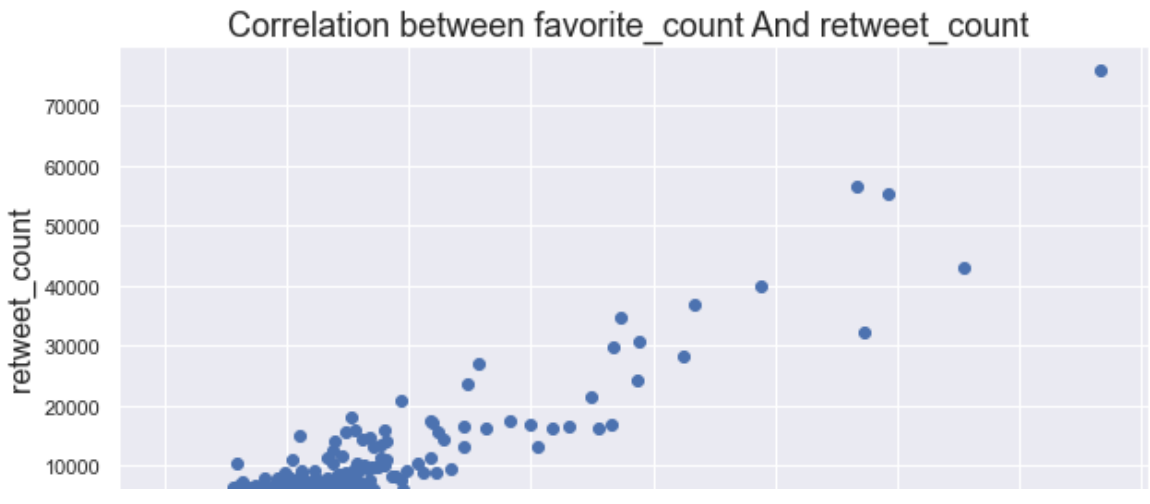
- `confidence_grade` **more skewed to the left. Around** `300 observation` **has** `1.0` **confidence grade.**
- `favorite_count` **most of observation about** `1400 observation` **has favorite_count between** `0: 10000` **, and less than** `100 observation` **has favorite_count between** `30000:50000` **.**
- `retweet_count` **most of observation about** `1600 observation` **has retweet_count between** `0: 8000` **, and less than** `100 observation` **has retweet _count between** `8000: 12000` **.**
- `rate` **most of observation has rate between** `9:14` **.**
- `img_num` **most of observation (** `1400 observation` **) has only** `1 photo` **.**

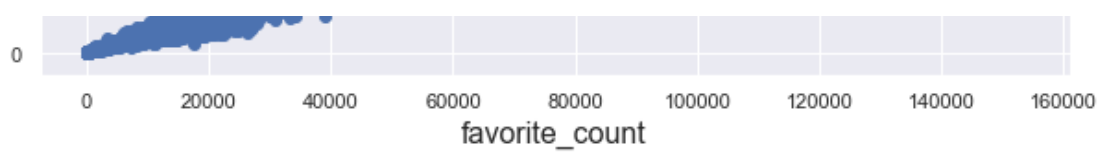
Research Question (What is the type of correlation between favorite_count And retweet_count?)

In [569]:

```
corr = df_master.corr()
print("Correlation Between favorite_count And retweet_count = ",corr.loc['favorite_count'
,'retweet_count'])

print('\n           -->>The plot below show the correlation between favorite_count And r
etweet_count<<--')

#Visualization

plt.scatter(x=df_master['favorite_count'],y=df_master['retweet_count']);
#set the figure size and labels

sns.set(rc={'figure.figsize':(10,10)});
plt.title('Correlation between favorite_count And retweet_count ' ,fontsize = 18);
plt.xlabel('favorite_count',fontsize = 16);
plt.ylabel('retweet_count',fontsize = 16);
```

```
Correlation Between favorite_count And retweet_count =  0.9310570436678284

           -->>The plot below show the correlation between favorite_count And retweet_cou
nt<<--
```





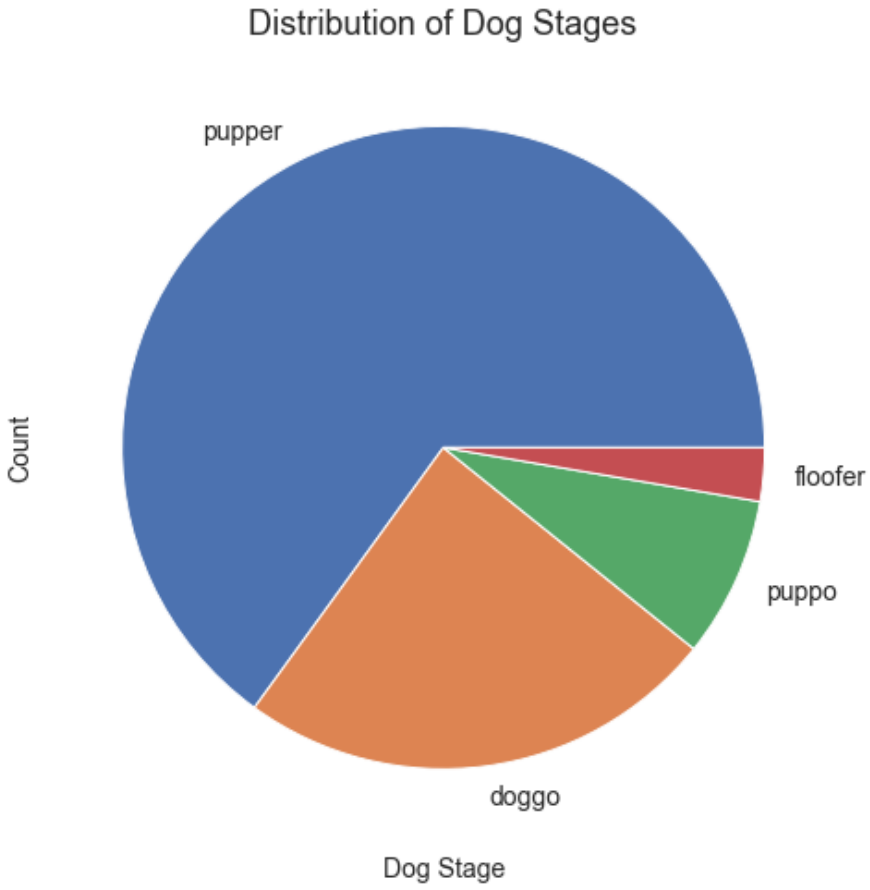
Note: There is strong relation between favorite_count And retweet_count.

- **It is a positive relationship.**
- **While one increasing the other increased too.**
- **That means the more favorite_count increased the more retweet_count increased.**

Research Question (What is the distribution of dogs_stage?)

In [585]:

```
df_master['dogs_stage'].value_counts().plot('pie', figsize=(8,8),fontsize = 14)
plt.title("Distribution of Dog Stages",fontsize = 18)
plt.xlabel('Dog Stage',fontsize = 14)
plt.legend
plt.ylabel('Count',fontsize = 14);
```



As we can notice that `pupper` has the highest count of `169`, and `floofer` has the lowest count of `7`.

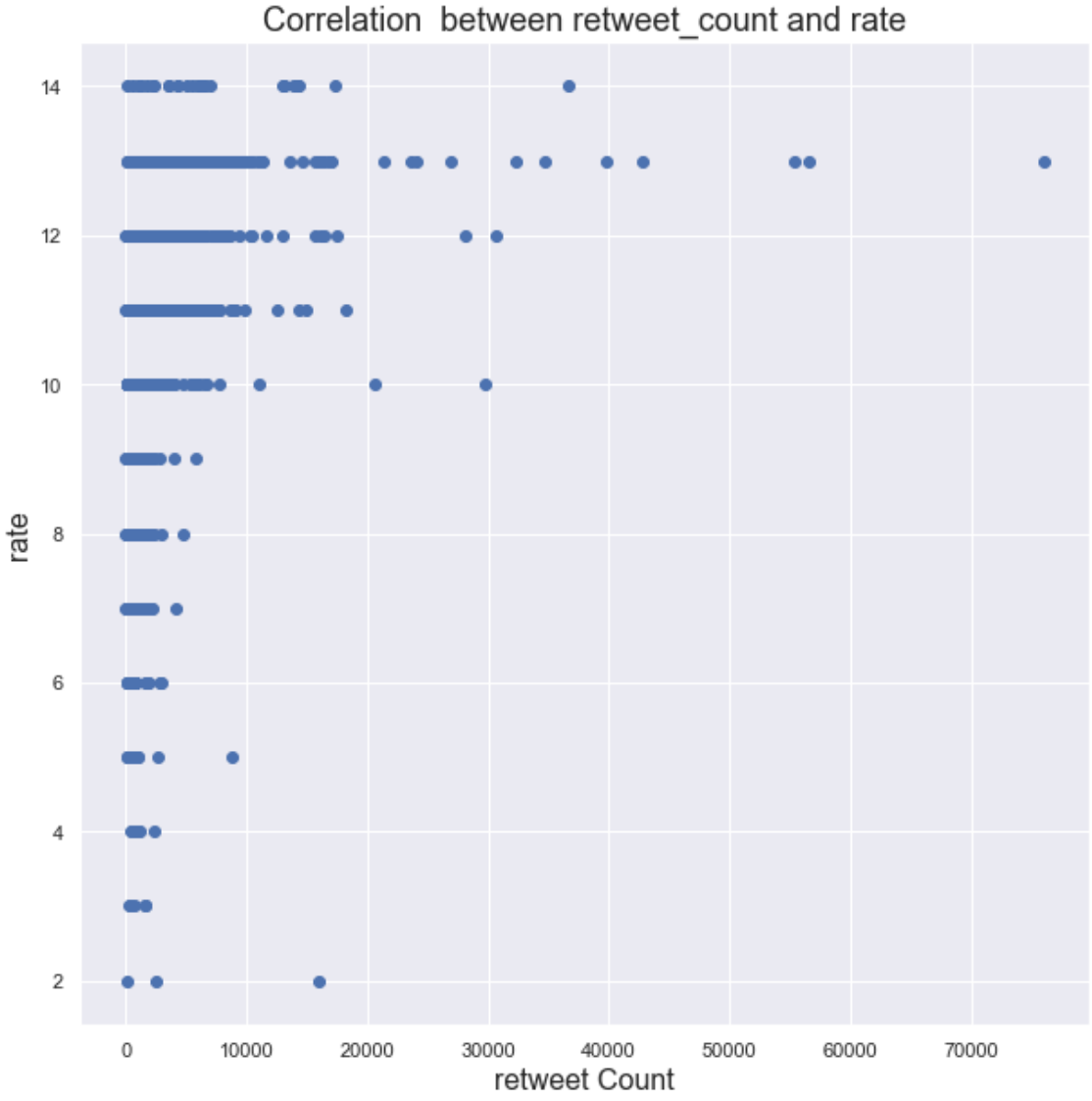
Research Question (Is there a relation between rate and tweet count?)

In [592]:

```
corr = df_master.corr()
print("Correlation  between retweet_count and rate = ",corr.loc['rate','retweet_count'])

print('\n           -->>The plot below show the Correlation  between retweet_count and r
ate <<--')

#Visualization
```

```
plt.scatter(x=df_master['retweet_count'],y=df_master['rate']);
#set the figure size and labels

sns.set(rc={'figure.figsize':(10,10)});
plt.title('Correlation  between retweet_count and rate',fontsize = 18);
plt.xlabel('retweet Count',fontsize = 16);
plt.ylabel('rate',fontsize = 16);
```

```
Correlation  between retweet_count and rate =  0.29283287114441137

           -->>The plot below show the Correlation  between retweet_count and rate <<--
```



===> As we can see the highest ratings do not receive the most retweets.

===> There is weak relation between rate And retweet_count.

Research Question (Is there a relation between rate and favorite count?)

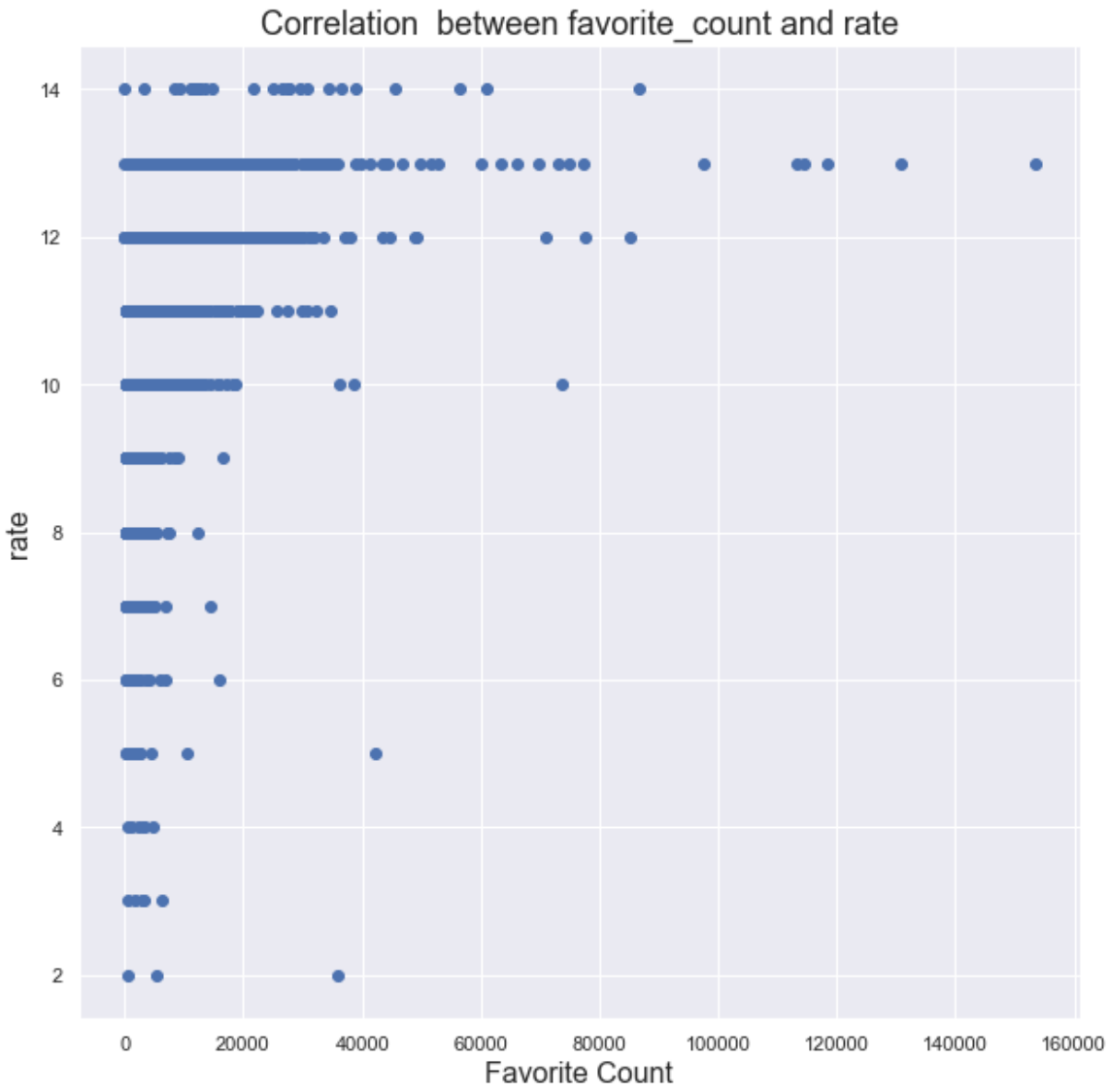
In [593]:

```
corr = df_master.corr()
print("Correlation  between favorite_count and rate = ",corr.loc['favorite_count','rate']
)

print('\n           -->>The plot below show the Correlation  between favorite_count and
rate <<--')

#Visualization

plt.scatter(x=df_master['favorite_count'],y=df_master['rate']);
#set the figure size and labels
```

```
sns.set(rc={'figure.figsize':(10,10)});
plt.title('Correlation  between favorite_count and rate',fontsize = 18);
plt.xlabel('Favorite Count',fontsize = 16);
plt.ylabel('rate',fontsize = 16);
```

```
Correlation  between favorite_count and rate =  0.39316189693676284
```

```
-->>The plot below show the Correlation  between favorite_count and rate <<--
```



===> As we can see the highest ratings do not receive the most favorite count.

===> There is weak relation between rate And favorite_count.

Conclusions:

- **The highest rating doesn't receive the most favorite count or retweet count.**
- **There is strong relation between favorite count and retweet count. The one increased the othher increased too.**
- **Pupper has the highest frequent `(179)` and floofer has the lowest `(7)`.**
- **Japanese_spaniel takes the lowest rate in average = `5/10`.**
- **Bouvier_des_Flandres takes the highest rate in average = `13/10`.**
- **Bedlington_terrier takes the highest favorite_count in average= `22731.5`.**
- **Bedlington_terrier takes the highest retweet_count in average = `7165`**
- **golden_retriever takes the highest retweet_count in total = `508771`.**
- **golden_retriever takes the highest favorite_count in total = `1769933`.**
- **the most favorite dog type is the golden_retriever with count `175`.**

Limitations:

- **This above exploration is not guaranteed 100%**
- **This exploration gives us high expectations and it may affected by other factors that would lead to different results .**