wrangle_act-Copy1

March 4, 2018

0.1 Analysis

In [68]:
```
tweet_master_df.head()
```

Out[68]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	expanded_urls
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56	Twitter for iPhone	This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/1	https://twitter.com/dog_rates/status/892420643555336193/photo/1
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27	Twitter for iPhone	This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's	https://twitter.com/dog_rates/status/892177421306343426/photo/1
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03	Twitter for iPhone	This is Archie. He is a rare Norwegian Pouncing Corgo. Lives in the tall grass. You	https://twitter.com/dog_rates/status/891815181378084864/photo/1
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51	Twitter for iPhone	This is Darla. She commenced a snooze mid meal. 13/10 happens to the best of us http	https://twitter.com/dog_rates/status/891689557279858688/photo/1
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24	Twitter for iPhone	This is Franklin. He would like you to stop calling him "cute." He is a very fierce	https://twitter.com/dog_rates/status/891327558926688256/photo/1,https://twitter.com/

```
   rating_numerator  rating_denominator      name  ...   img_num           p1  \
0  13.0              10.0                 Phineas  ...    1       orange
1  13.0              10.0                 Tilly    ...    1       Chihuahua
2  12.0              10.0                 Archie   ...    1       Chihuahua
3  13.0              10.0                 Darla    ...    1       paper_towel
4  12.0              10.0                 Franklin ...    2       basset

    p1_conf  p1_dog                 p2   p2_conf p2_dog  \
0  0.097049  False   bagel               0.085851  False
1  0.323581  True    Pekinese            0.090647  True
2  0.716012  True    malamute            0.078253  True
3  0.170278  False   Labrador_retriever  0.168086  True
4  0.555712  True    English_springer    0.225770  True

                           p3   p3_conf p3_dog
0  banana                      0.076110  False
1  papillon                    0.068957  True
2  kelpie                      0.031379  True
3  spatula                     0.040836  False
4  German_short-haired_pointer  0.175219  True

[5 rows x 25 columns]
```

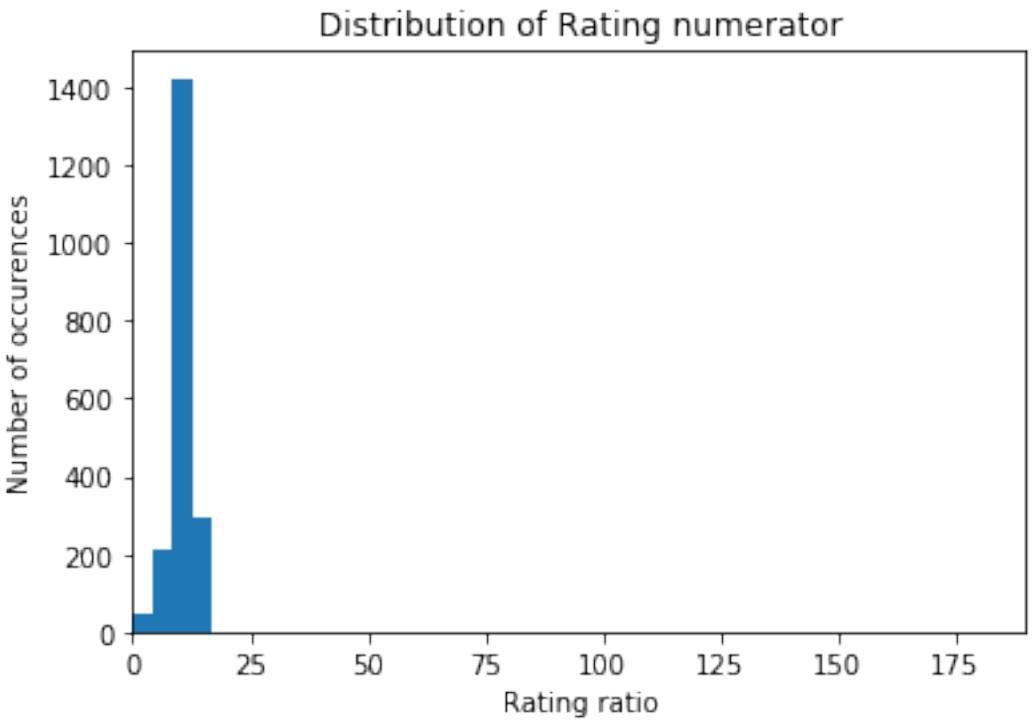
In my data analysis on the We rate dogs tweets I will be mainly focussing on finding realtionship betwen retweet counts and followers count. Distribution of rating_numerator and rating_denominator. I will also like to undertand the various sources of tweets. So let's begin!

At first I want to look at the distribution of rating_numerator and rating_ratio the new variable which I have created. This variable is simply the ration betwen rating_numerator and rating_denominator from the original dataframe

In [69]:
```
#Creating a rating ration variable for further analysis
tweet_master_df['rating_ratio']=tweet_master_df['rating_numerator']/tweet_master_df['ra
```

In [70]:
```
# Rating numerator and denominator distribution
#tweet_master_df.rating_numerator.plot(kind="hist")
bin_size = 100
plt.hist(tweet_master_df.rating_numerator,bin_size)
plt.xlim(0,190)
plt.xlabel("Rating ratio")
plt.ylabel("Number of occurences")
plt.title("Distribution of Rating numerator")
```

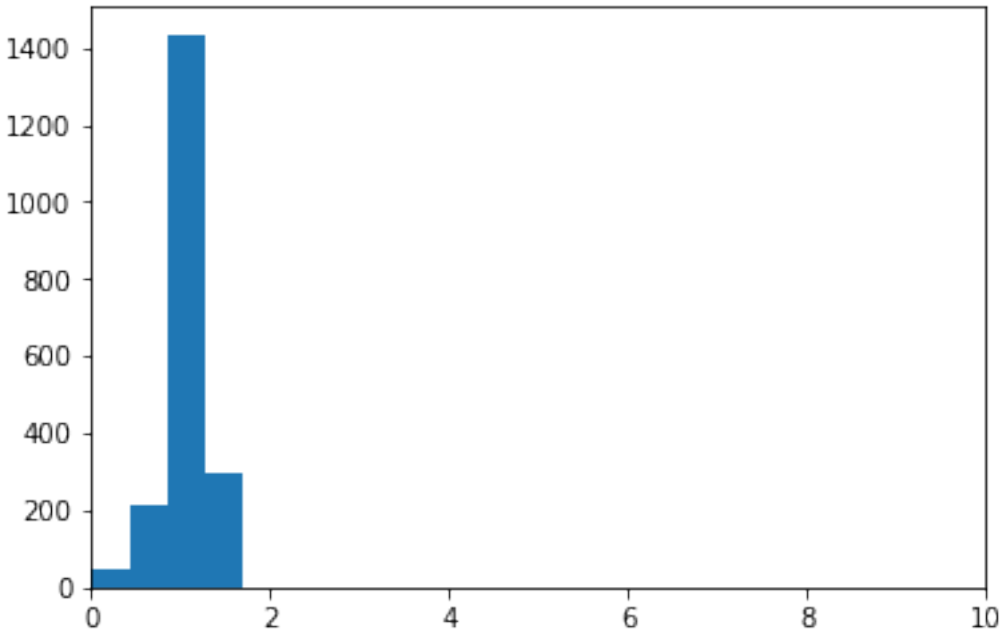
Out[70]:
```
Text(0.5,1,'Distribution of Rating numerator')
```



Seems like most of the dogs have a rating numerator less than 190. Many dogs have rating numerator of around 10. Lets have a look at the rating ration.

```
In [71]: bin_size = 100
         plt.hist(tweet_master_df.rating_ratio,bin_size)
         plt.xlim(0,10)

Out[71]: (0, 10)
```

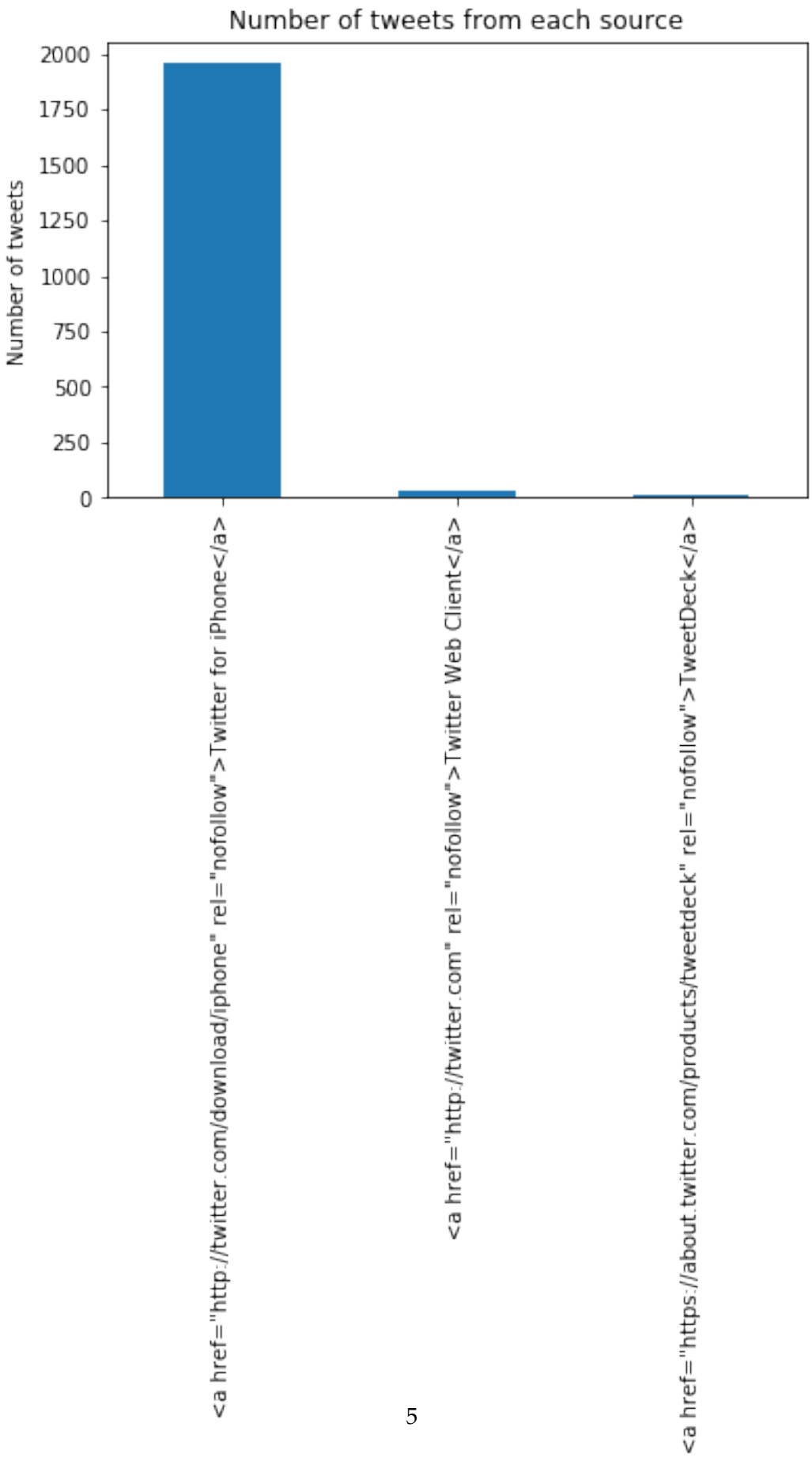


Again, here also there is no notable anamolies. Most of the dogs have a rating ratio of 1.

```
In [72]: #Number of records from each source
         tweet_master_df.source.value_counts().plot(kind="bar")

         plt.ylabel("Number of tweets")
         plt.xlabel("Sources")
         plt.title("Number of tweets from each source")
         #tweet_master_df.info()

Out[72]: Text(0.5,1,'Number of tweets from each source')
```

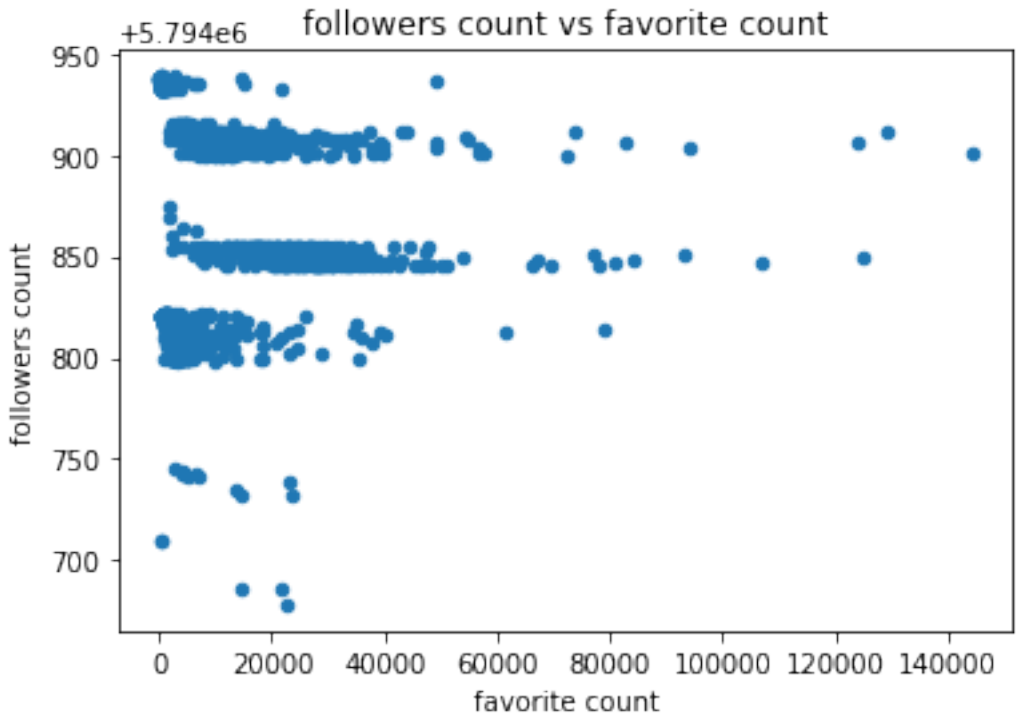

Number of tweets from each source
Number of tweets
2000
1750
1500
1250
1000
750
500
250
0
Twitter for iPhone
Twitter Web Client
TweetDeck
Sources

Source of most of the tweets is from the Twitter for iphone , followed by twitter from web browser, followed by Tweet Deck.

Let's take a look at the realtionship betwen followers count and favorite count. Favorite count is the number of likes each tweet has recieved. I want to see if high number of followers helps in getting high number of likes.

```
In [73]: #Relationship between favorite_count and followers_count
         tweet_master_df.plot(x="favorite_count",y="followers_count",kind="scatter")
         plt.xlabel("favorite count")
         plt.ylabel("followers count")
         plt.title("followers count vs favorite count")
         plt.suptitle("")
         plt.figure(figsize=(10,10))
```

```
Out[73]: <matplotlib.figure.Figure at 0x7f617c41d6d8>
```



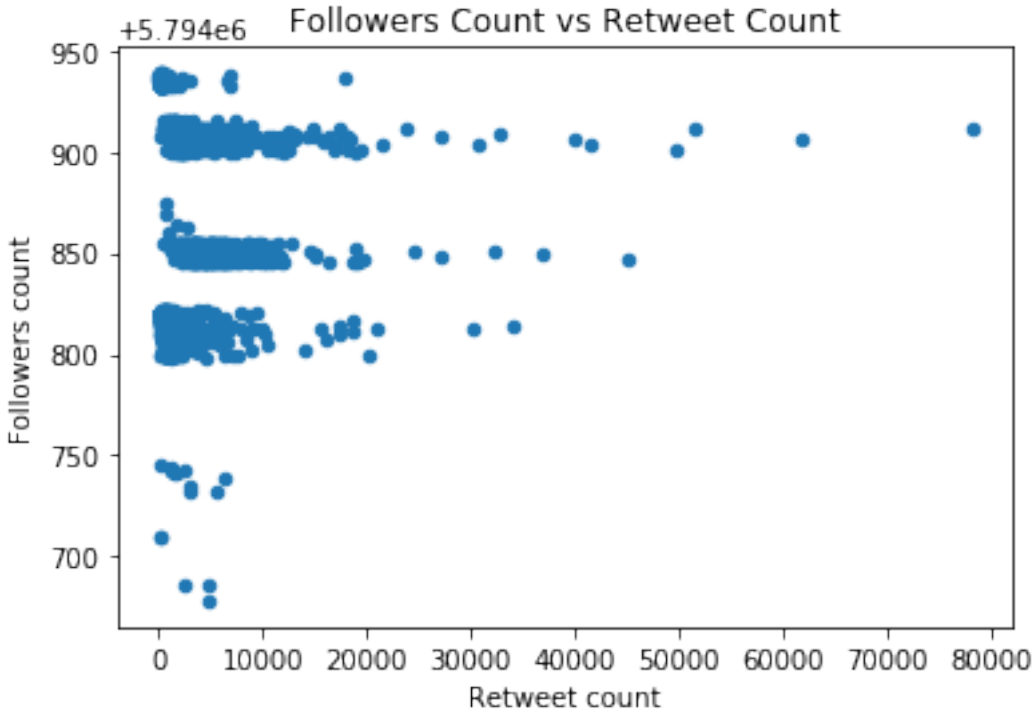

```
<matplotlib.figure.Figure at 0x7f617c41d6d8>
```

As seen above as the number of followers increases the favorite count is also increasing. But the favorite counts is almost nearly the same for dogs with less number of followers.

Now let's have a look at the relationship between followers count and retweet count. As per the common notion more the number of followers higher should be the retweet count. Let's see if this holds true for our dataset

```
In [74]: #Realtionship between retweet_count and followers_count
         tweet_master_df.plot(x="retweet_count",y="followers_count",kind="scatter")
         plt.xlabel("Retweet count")
         plt.ylabel("Followers count")
         plt.title("Followers Count vs Retweet Count")
         plt.figure(figsize=(10,10))

Out[74]: <matplotlib.figure.Figure at 0x7f617c462c18>
```




```
<matplotlib.figure.Figure at 0x7f617c462c18>
```

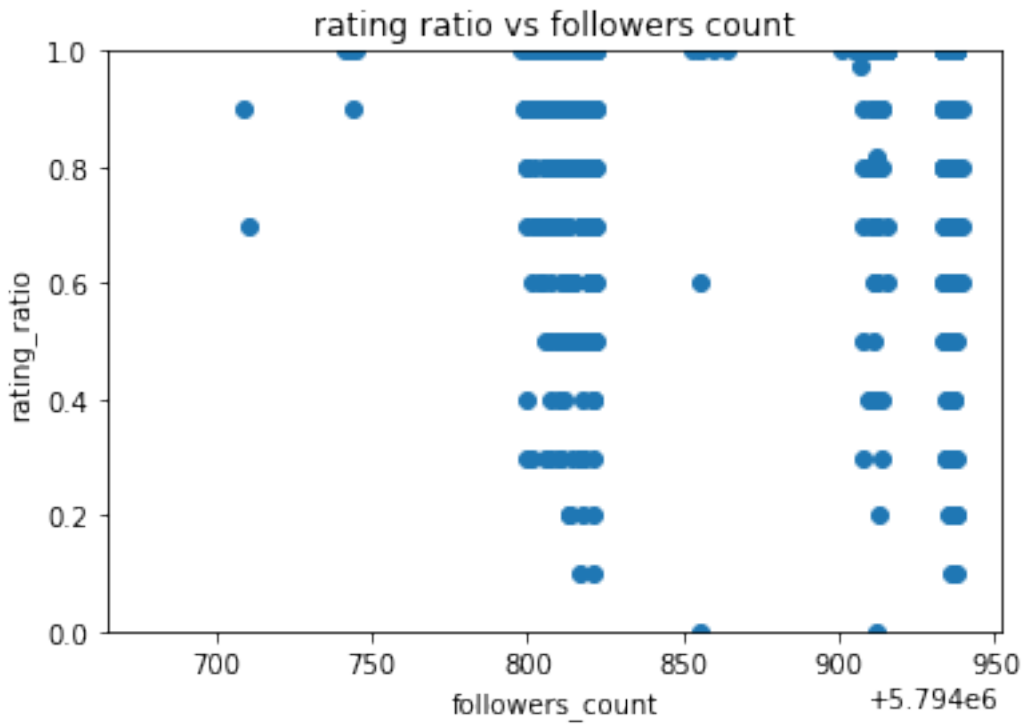
As seen above also as the followers count has increased the retweet counts have increased. Again when the retweet counts are less for dogs who have lesser number of followers. One more notable anamoly here is in case of dogs with high number of followers also the retweet counts are less.

The next relationship I want to see is if the followers_count has any impact on rating_ratio of the dogs. I am expecting that the high rated dogs might be attracting more number of people to follow them. But let's see how the realationship looks like in reality.

```
In [75]: #Relationship between followers count and rating ratio
         plt.scatter(tweet_master_df.followers_count,tweet_master_df.rating_ratio)
         plt.xlabel("followers_count")
         plt.ylabel("rating_ratio")
```

```
plt.title("rating ratio vs followers count")
plt.ylim(0,1)
plt.figure(figsize=(10,10))
```

Out[75]: <matplotlib.figure.Figure at 0x7f617c352160>



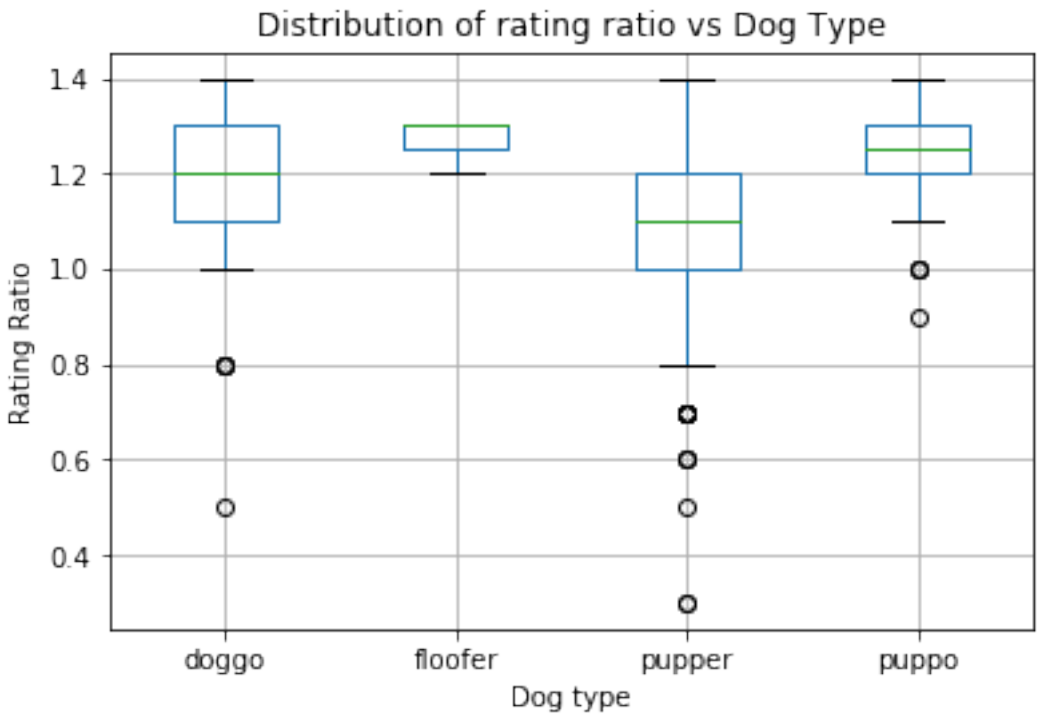
<matplotlib.figure.Figure at 0x7f617c352160>

Well from the plot above there doesn't seem to be any strong relationship between rating ratio and followers count. There are few dogs with hig rating ratio but very few followers. This is possible because it is the owner of the dogs who assign them the ratings not the twitter users.

Now I want to take a look at the rating ration distribution of the different dog types. I want to see if any particulatr dog type is more favorable to get higher rating and how does the mean, median, and IQR range of rating ratio for each of the dog types.

In [76]:
```
tweet_master_df.dog_type.value_counts()
tweet_master_df.boxplot(column='rating_ratio', by='dog_type')
plt.xlabel("Dog type")
plt.ylabel("Rating Ratio")
plt.suptitle("")
plt.title("Distribution of rating ratio vs Dog Type")
```

Out[76]: Text(0.5,1,'Distribution of rating ratio vs Dog Type')



This graph here shows that the median rating ration for puppo dog types is the highest where as for pupper is the lowest. Each of the dog types with exception of floofer have outliers. There is very low variablity in the dog rating for dog type floofer