分析

tweet_id
timestamp
text
retweet_count
favorite_count
jpg_url：是预测的图像资源链接
img_num：最可信的预测结果对应的图像编号 → 1 推特中的第一张图片
p1：是算法对推特中图片的一号预测 → 金毛犬
p1_conf：是算法的一号预测的可信度 → 95%
p1_dog：是一号预测该图片是否属于“狗”（有可能是其他物种，比如熊、马等） → True 真
p2：是算法对推特中图片预测的第二种可能性 → 拉布拉多犬
p2_conf：是算法的二号预测的可信度 → 1%
p2_dog：是二号预测该图片是否属于“狗” → True 真
p3：是算法对推特中图片预测的第三种可能性
p3_conf：是算法的三号预测的可信度
p3_dog：是三号预测该图片是否属于“狗”
status：狗狗的地位（doggo、floofer、puppet、puppo）
rating_numerator：评级的分子
rating_denominator：评级的分母
name：狗狗的名字
rating_score：评级的分子除以分母所得的值

数据可视化

1、参与weratedog活动中什么地位的狗狗最多?

In [75]:

```
twitter_archive_enhanced_clean.status.value_counts()
```

Out[75]:

```
None       1652
pupper      246
doggo        68
puppo        30
floofer       9
Name: status, dtype: int64
```

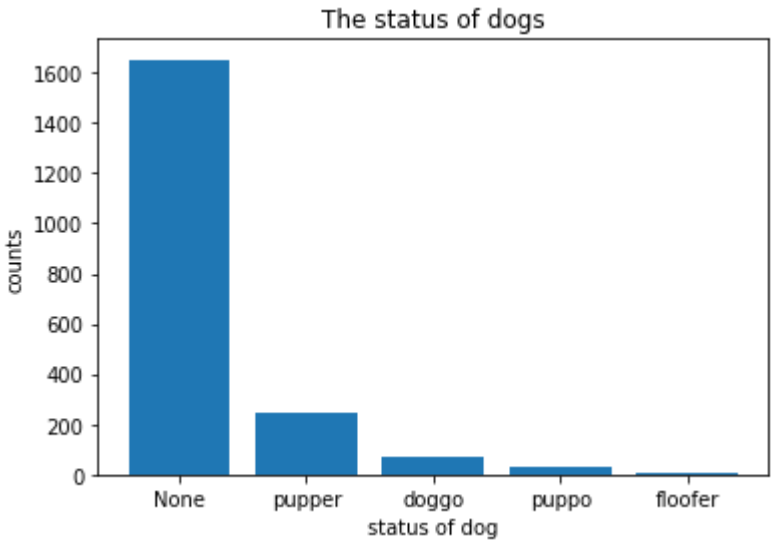
In [76]:

```
x = ['None','pupper','doggo','puppo','floofer']
y = twitter_archive_enhanced_clean.status.value_counts()

plt.bar(x,y)
plt.xlabel("status of dog")
plt.ylabel("counts")
plt.title("The status of dogs")
```

Out[76]:

```
Text(0.5,1,'The status of dogs')
```



In [77]:

```
twitter_archive_enhanced_clean.groupby('status')['rating_score'].mean()
```

Out[77]:

```
status
None       1.177910
doggo      1.180882
floofer    1.177778
pupper     1.063630
puppo      1.220000
Name: rating_score, dtype: float64
```

2、不同地位的狗狗的平均评分值

In [78]:

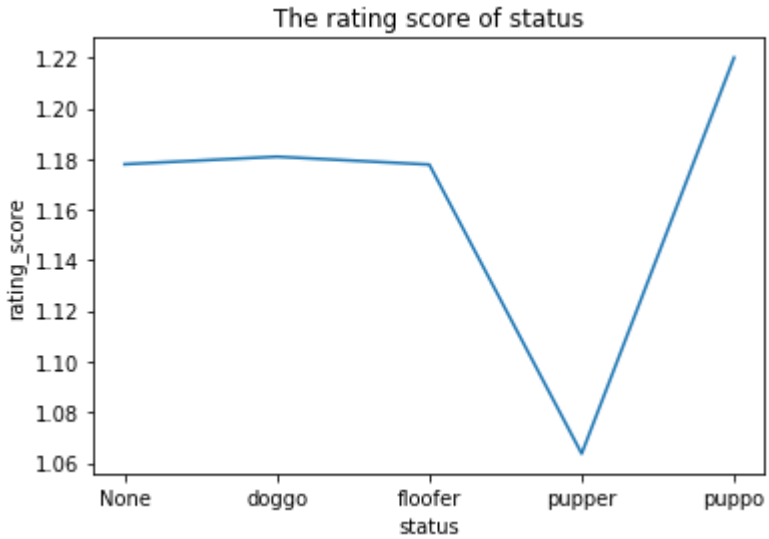
```
from matplotlib import pyplot
```

In [79]:

```
x = ['None','doggo','floofer','pupper','puppo']
y = twitter_archive_enhanced_clean.groupby('status')['rating_score'].mean()

pyplot.plot(x,y)
pyplot.xlabel('status')
pyplot.ylabel('rating_score')
pyplot.title('The rating score of status')

pyplot.show()
```



评分对应的转发数量和喜爱数量

3、是否评分越高，喜爱和转发数量越高?

In [80]:

```
twitter_archive_enhanced_clean.rating_score.value_counts()
```

Out[80]:

```
1.200000      459
1.000000      422
1.100000      404
1.300000      263
0.900000      151
0.800000       95
0.700000       51
1.400000       35
0.500000       34
0.600000       32
0.300000       19
0.400000       15
0.200000       10
0.100000        4
0.000000        2
1.127000        1
1.350000        1
3.428571        1
0.636364        1
0.818182        1
1.126000        1
0.975000        1
42.000000       1
177.600000      1
Name: rating_score, dtype: int64
```

In [81]:

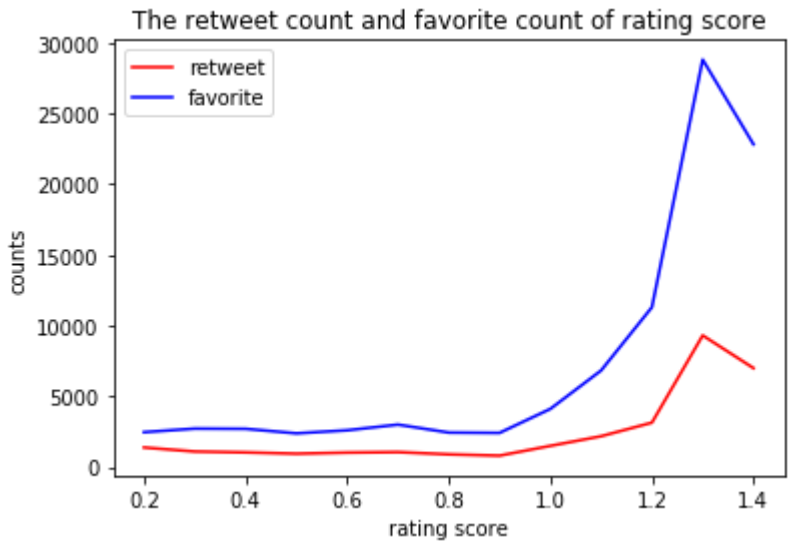
```
#因为部分评分计数值很小，当作特殊数据，绘制图形的的时候不包含进去
x = [0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0,1.1,1.2,1.3,1.4]
y1 = []
y2 = []
for a in set(x):
    b = twitter_archive_enhanced_clean[twitter_archive_enhanced_clean['rating_sc
ore']==a].retweet_count.mean()
    c = twitter_archive_enhanced_clean[twitter_archive_enhanced_clean['rating_sc
ore']==a].favorite_count.mean()
    y1.append(b)
    y2.append(c)
```

In [82]:

```
plt.plot(x,y1,'r',label='retweet')
plt.plot(x,y2,'b',label='favorite')

plt.title('The retweet count and favorite count of rating score')
plt.xlabel('rating score')
plt.ylabel('counts')

plt.legend()
plt.show()
```



结论

1、由可视化图表("The status of dogs")可以看出，数据集中有很多狗狗缺失地位信息，从推文中无法获取，显示为None。除了没有地位信息的狗狗之外，地位是pupper的狗狗最多，第二是doggo,第三和第四分别是puppo和floofer。 2、不同地位的狗狗评分值差异，由可视化图表('The rating score of status')看出，地位是pupper的狗狗平均评分最低，puppo地位的狗狗平均评分最高，doggo和floofer几乎一致。 3、查看评分和喜爱转发之间是否一定的关联度，由可视化图表('The retweet count and favorite count of rating score')看出，转发数量和喜爱数量，从评分0.2～1.3区间，有个明显的上升趋势。在本数据集中可以认为评分和转发、喜爱数量之间有一定的正相关，但是分析有一定的局限性，因为数据量不够大，可能对数据结果造成了一定影响。