数据分析报告

# 狗与非狗是否会对打分产生影响

原来的数据中的评分只有分子和分母，在分析时增加一列’rating’:

rating = rating\_numerator / rating\_denominater

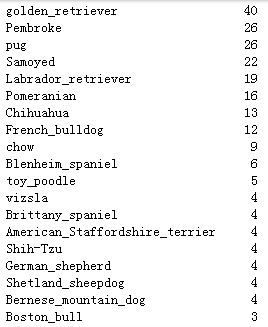
由于神经网络可以判断图片中物体，因此通过神经网络的方法筛选出两组数据，rating\_dog和rating\_nodog。所有狗的平均分其, 而所有非狗的平均分为，为这两组数的评分为做以下假设:

C:\Users\ABSLUT~1\AppData\Local\Temp\1546441840(1).png

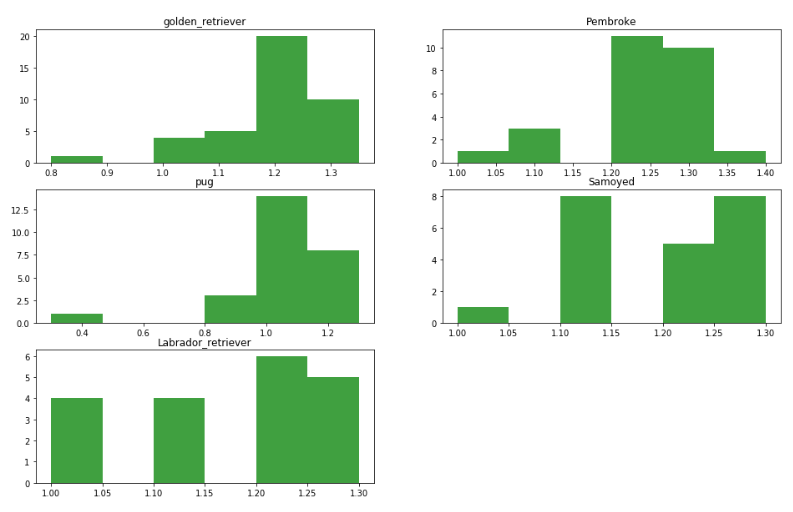
通过使用t-test的独立检验，发现两组数据的t值为-0.8， p值为0.43， p值远大于我们设定的置信水平。我们无法推翻H0假设，因此我们可以认为是否是狗对评分并无统计意义上的影响。

# 不同种类的狗之间评分是否有差异

通过神经网络可以判断狗的类型，在这里分析时，我选取了两个筛选条件，一是判断必须是狗，而是置信水平在0.9以上。通过降序，可以看到：



为了减轻分析的数据量，我们只分析了最热门的5中狗的类型，即上表的前5名。利用直方图可以看到其中只有pug类型的评分与其他四种狗不一样。



利用F-test，我们可以发现，如果假设：

C:\Users\ABSLUT~1\AppData\Local\Temp\1546442678(1).png

通过F检验可以发现p值为6.45e-5，远小于0.05，表明H0假设不成立，5种类型狗的均值不是都相同的。

如果我们将pug也就是从检验中踢除，

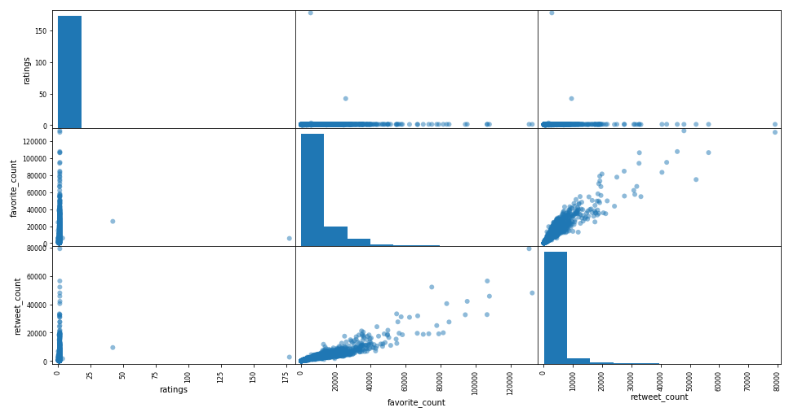
C:\Users\ABSLUT~1\AppData\Local\Temp\1546442841(1).png

通过F检验可以发现p值为0.196，大于0.05，这表明H0假设无法推翻，这四种狗的评分均值并无统计意义上的差别。

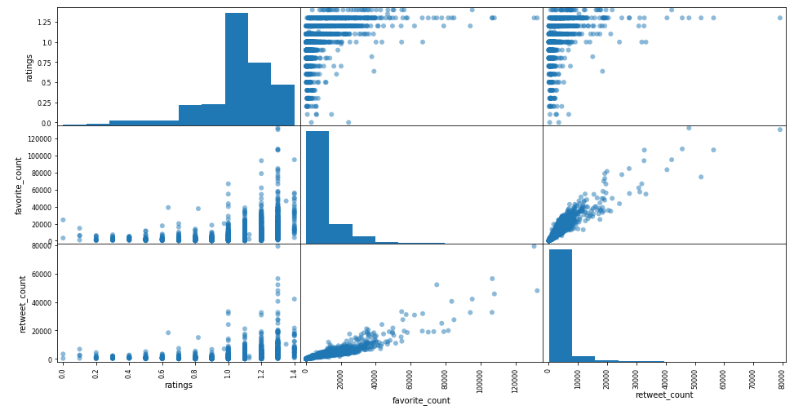
因此我们可以认为在**最流行的5种狗中，WeRateDog对Pug这种狗的评分低于其他四种狗。**

# 评分、点赞数和转发数之间是否有关系

使用散点、直方矩阵可以得到：



通过图片和表格可以发现ratings有三个异常值数值过高，导致无法观察出ratings对点赞数和转发数的影响，把这三个异常值去除后可以得到：



从上面的矩阵图来看，ratings和点赞数与转发数似乎并没有很明显的联系，但是转发数和点赞数之间有明显的线性相关性，通过线性回归可以得到:

转发数 = 0.348\* 点赞数 -336

R^2 为83.7%，表明线性相关性很高。

因此我们可以得到三个结论：

1. 是否是狗对评分并无统计意义上的影响。
2. 最流行的5种狗中，WeRateDog对Pug这种狗的评分低于其他四种狗。
3. 评分与点赞和转发数并无明显联系，但是点赞和转发数之间有很明显的线性相关性