数据分析报告

一、 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$\mu_1 = 1.097$其, 而所有非狗的平均分为$\mu_2 = 1.361$，为这两组数的评分为做以下假设:

H0: ratings are inrelated to dogs or not, $\mu_1 = \mu_2$
H1: ratings are related to dogs or not, $\mu_1 \neq \mu_2$

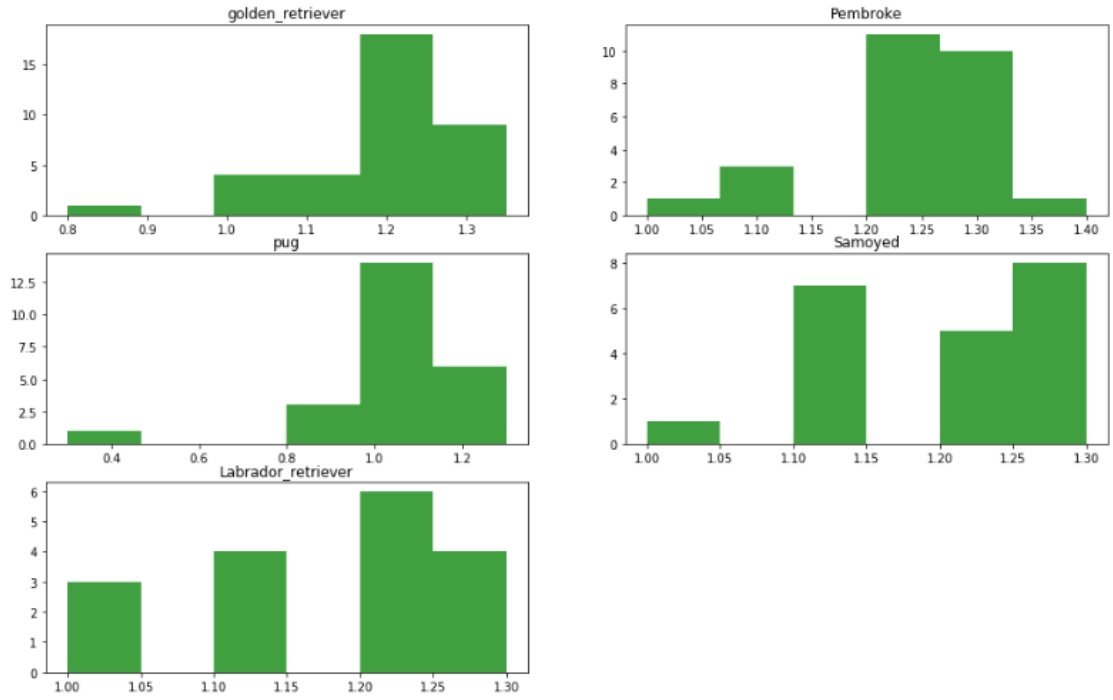
通过使用 t-test 的独立检验，发现两组数据的 t 值为-0.8， p 值为 0.43， p 值远大于我们设定的置信水平$\alpha = 0.05$。我们无法推翻 H0 假设，因此我们可以认为是否是狗对评分并无统计意义上的影响。

二、 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：

golden_retriever	36
Pembroke	26
pug	24
Samoyed	21
Labrador_retriever	17
Pomeranian	16
Chihuahua	13
French_bulldog	12
chow	9
Blenheim_spaniel	6
toy_poodle	5
Shetland_sheepdog	4
German_shepherd	4
Brittany_spaniel	4
Bernese_mountain_dog	4
vizsla	4
American_Staffordshire_terrier	4
Shih-Tzu	4
beagle	3

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：

H0: ratings of five dogs are same. $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
H1: not all ratings are same

通过 F 检验可以发现 p 值为 2.95e-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μ_3从检验中踢除，

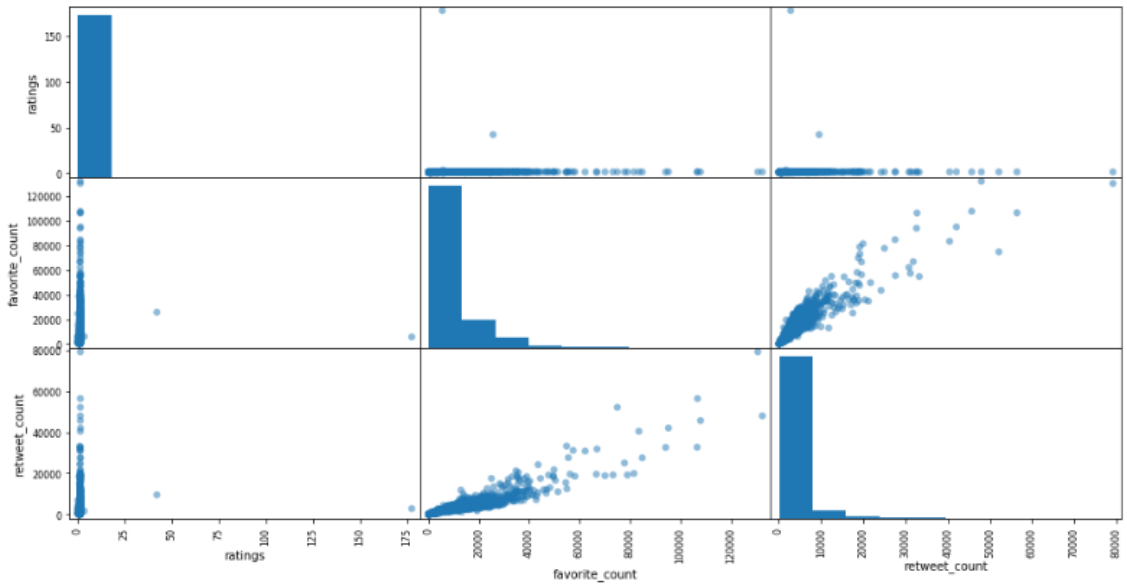
H0: ratings of five dogs are same. $\mu_1 = \mu_2 = \mu_4 = \mu_5$
H1: not all ratings are same

通过 F 检验可以发现 p 值为 0.214，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。

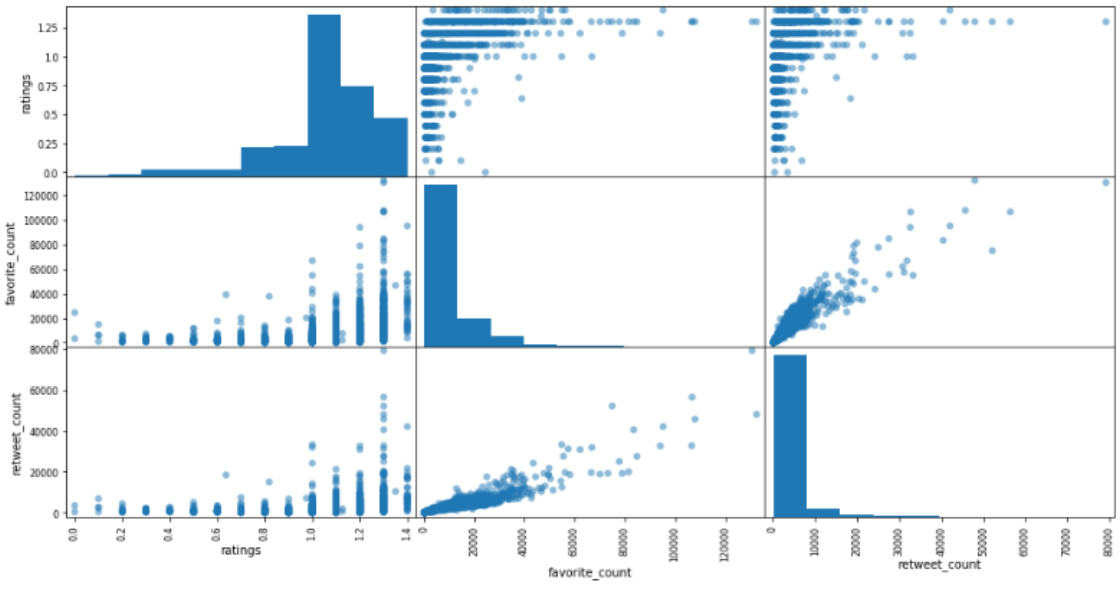
因此我们可以认为在**最流行的 5 种狗中，WeRateDog 对 Pug 这种狗的评分低于其他四种狗。**

三、 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