Weratedog 数据可视化分析报告

总体概况：

由评分箱线图和直方图（图 1）可知：

1.大多数的评分集中在 0.75-1.5 之间（满分 1 分）。评分都较高，接近 1，甚至大于 1

2.大于 2 分的狗狗也不少，但是分布较散，最大可到 8+分

3.由表 1 可得 1.1 分只是在 50%的位置，拿到 1.1+分也只是及格分数。

所以总体来看，就算是拿了超过 1.1 分的评分也不算很高，排名在数据集中的中间。所以，这些评分有一定程度偏高，不过评分项目本来就不是什么专业评审，只是一个娱乐而已，所以如果要给自己发的照片打分做对比的话，50%分位线 1.1 分和 75%分位线 1.2 分进行对比吧。

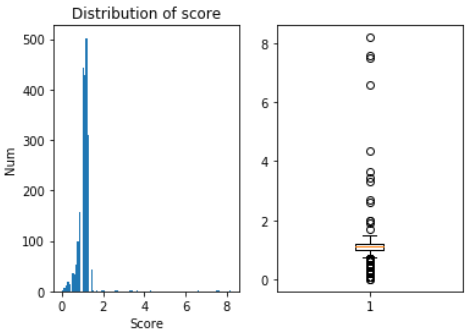


图 1 直方图和箱线图

表 1 整体描述统计

	favorite_count	retweet_count	score	status
count	2175.000000	2175.000000	2175.000000	2175.000000
mean	8769.120000	2759.735632	1.219011	0.180690
std	12224.417865	4682.731540	4.244435	0.396622
min	52.000000	0.000000	0.000000	0.000000
25%	1904.500000	605.500000	1.000000	0.000000
50%	4023.000000	1336.000000	1.100000	0.000000
75%	11072.000000	3202.000000	1.200000	0.000000
max	132318.000000	79116.000000	177.600000	2.000000

有身份的狗：

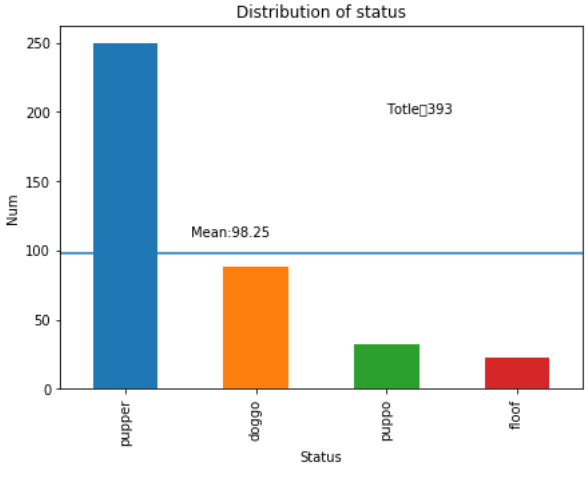
由狗狗身份数量描述统计和条形图可得：

1.狗狗身份难得，在 2000+只狗狗中，只有 383 只有身份，393 的总数是因为部分狗狗有多重身份！

2.其中 pupper 最多，远超其它的身份的狗狗。定义是初入狗世，未经历多少狗狗间的勾爪斗脚的未成熟的，还没准备好当一个大狗的小狗狗。

3.由描述统计可知，其 50%和 75%分位线和整体一致，均分反而比整体低了 0.1 分。

所以，身份并不是狗狗评分的一个主要依据，有身份并没有对分数带来多大的变化。推选过来评分的有身份的狗狗中，大多数都是幼狗，也许是在推特上晒照片要求评分的人们，相比于其它更成熟的大狗对幼狗会有特殊的偏好。



	favorite_count	retweet_count	score	status
count	383.000000	383.000000	383.000000	383.00000
mean	10203.331593	3408.973890	1.105561	1.02611
std	14688.636332	6494.427094	0.176936	0.15967
min	195.000000	3.000000	0.300000	1.00000
25%	2684.000000	831.500000	1.000000	1.00000
50%	4909.000000	1598.000000	1.100000	1.00000
75%	12116.500000	3550.000000	1.200000	1.00000
max	132318.000000	79116.000000	1.400000	2.00000

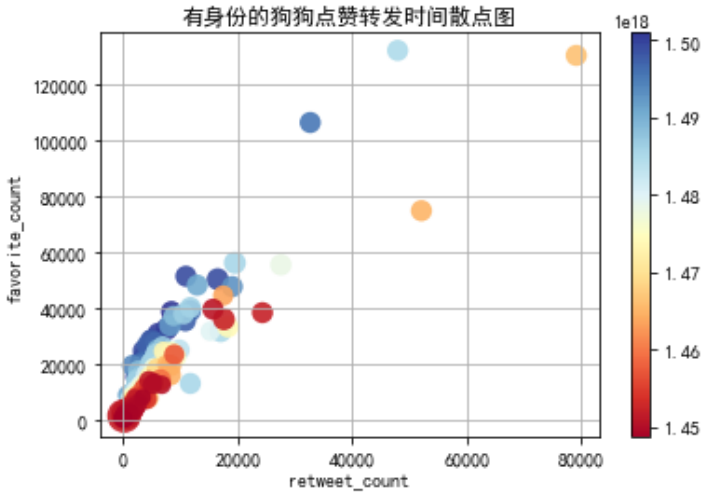
有身份的 VS 总体：

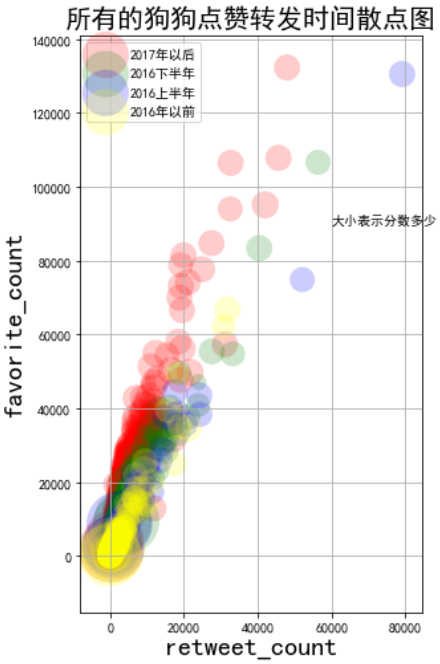
一.点赞转发时间的关系：

1.转发数量和点赞数量基本成正比
2.2017 年以后发布的 tweet，每一个转发带来的点赞数更多。可能原因，项目关注度高了，关注人群结构变化。
3.从对比两图可以看出。点赞和转发都高的狗狗评分不一定高,也不一定有身份
4.部分评分高的狗狗，点赞数却很低，数据没问题，一个是有国家意义的象征的狗，点赞数不为 0 但是也不高。一个是一个叫做史努比狗狗的名人，给了很高的分，42 分（满分 1 分）。

由有身份的狗狗点赞转发发布时间散点图可知
5.有身份的狗狗，分布和整个数据集相似
6.所有的有身份狗狗，评分相差不大
7.点赞数和转发数量最高的的几只狗狗，都是较晚发布的（橙黄色以后）。

总体而言，推特的转发数量和点赞数基本成正比，但是它们和分数没有明显关系，每一个转发对应远点赞数大于 1，有身份和没有身份，整体分布差别不大。点赞数和转发数最高的几个推特都是较晚发布的，2016 年上半年及以后，更多的是在 2017 年以后。





二.评分时间的关系

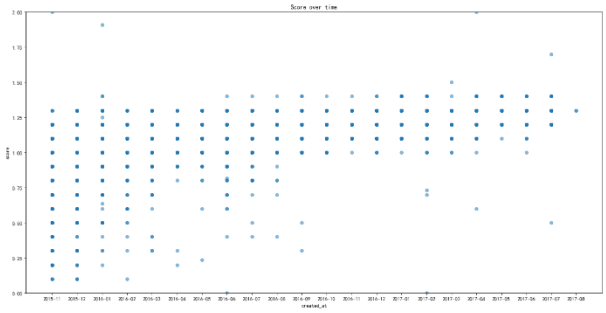
1.评分大多分布在 1.0-1.4 之间，在 2016-10 之后更明显，低于 1.0 分出现得更少了。
猜测原因：

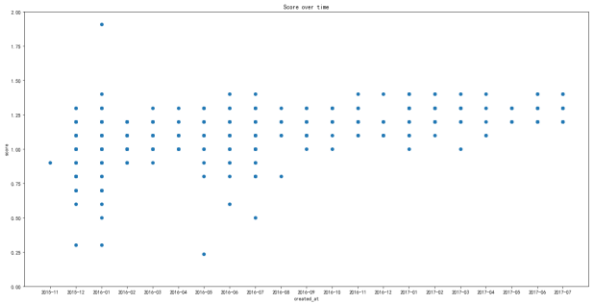
主观原因：项目受到更多的关注，低分对狗狗及主人也许会有较大的影响，所以倾向于给一个相对之前较高的分数

客观原因：项目越来越火，受到的关注更多，发照片的网友形成了一些共识，例如发些可以引人捧腹照片，而这些共识里是主办方打分的关键因素

2.可以发现，有身份的狗狗的评分比较稳定，整体都维持在 0.8 到 1.5 之间，符合之前箱线图的情况。

总体而言，随着时间的推移，打低分的情况更少了，高分的分布并没有更高，所以分数随着时间的推移，分数在贬值，就是说以前的相同分数会更有含量一些。

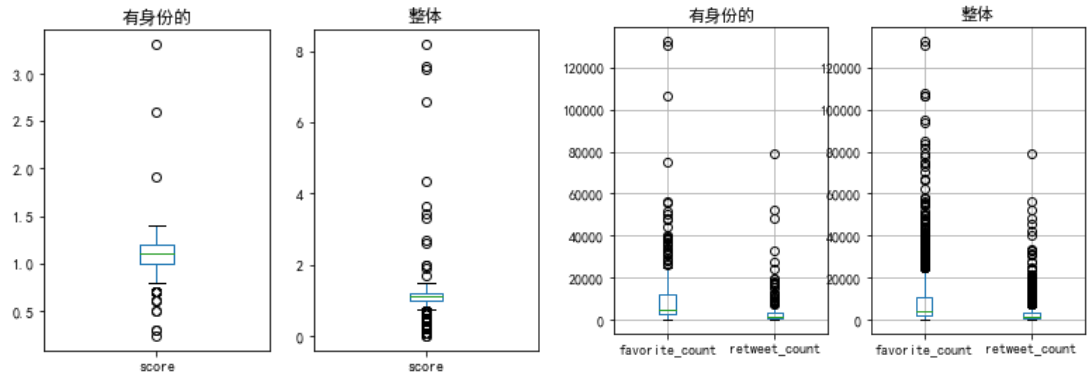




三.转发数点赞数对比：

有身份的狗狗，在评分上似乎比整体要高，但这个结论和之前的描述统计作出的结论相悖，因为 50%和 75%分位线是相同的，原来是刻度比例尺不同，疏忽疏忽。

所以总体而言，有身份和没身份（在 tweet 中会有说明），在评分、点赞数和转发数并没有明显区别。



四.神经网络模型预测图片为狗的数据

由于预测模型的准确率有待提高，所以就简略分析下。预测给出的结果中，占据最多的狗品种为，golden_retriever,就是右边这货，叫做金毛寻回猎犬又称金色猎犬（俗称金毛）。百度百科说它，体态匀称、稳重大方、性情乖巧、活泼机警，善良易驯。还可以被训练成导盲犬，好看还可以实用。第二和第三分别是拉布拉多寻回犬和彭布罗克威尔士柯基犬。第四个是吉娃娃。百度搜索给出的价位均在1000-2000 元之间。