数据分析与可视化

结合 wrangle_act.ipynb 文件中数据分析部分,在此给出以下几个方面的见解。

1. favorite_count 与 retweet_count 方面

分别对 favorite_count 与 retweet_count 的数据进行统计，其结果如表 1 所示：

表 1 favorite_count 与 retweet_count 统计信息表

	favorite_count	retweet_count
mean	9867.96	3066.89
std	12896.31	4932.49
max	144166	78187
min	0	0

favorite_count 与 retweet_count 的直方图分别如图 1 和图 2 所示：

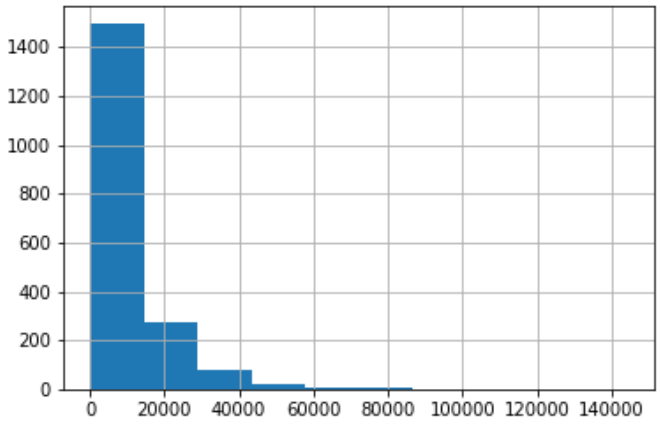

图 1 favorite_count 数据

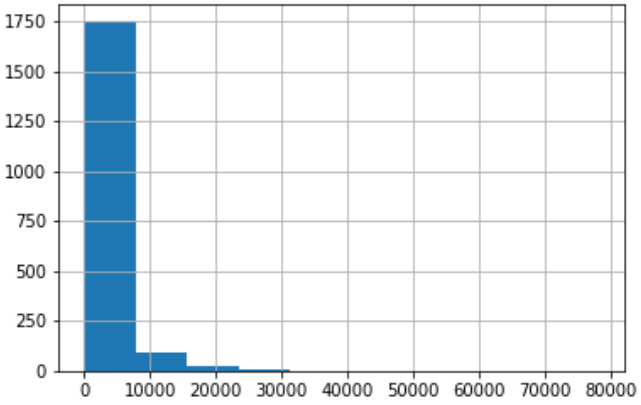

图 2 retweet_count 数据

从上述结果来看，favorite_count 的均值为 9867.96，retweet_count 的均值为 3066.89，以及各自的最大值的来看，favorite_count 的最大值为 144166，retweet_count 的最大值为 78187，总体上 favorite_count 要大于 retweet_count，也就是说人们更加倾向于“点赞”，而不是去转发。

2. 狗的“地位”stage

提取出来有关狗的地位 stage 的数据的统计如表 2 所示：

表 2 stage 统计表

	stage_count
pupper	249
doggo	79
puppo	31
floof	23

从表 2 中可以得知，在 tweet 中有给出狗的地位的信息中，可以得知大部分的狗属于 pupper，结合各 stage 的解释，也就是大部分的狗都是 younger，年纪比较小。

3. 评级

由于在数据中，将 rating_denominator 设为 10，所以下面给出 rating_numerator 的统计信息，如表 3 所示：

表 3 rating_numerator

	value
mean	12.726
std	45.456
max	1776
min	0

从表 3 中可以看出，总体的评级分数分子都是要大于分母的，也就是大于 10 分，说明人们对于狗狗的喜爱，更加倾向给予更多的分数去鼓励人们，说明人们内心都是向往美好的事物。

4. 图片算法预测

分别对 p1_conf，p2_conf 和 p3_conf 做统计分析，如表 4 所示：

表 4 图片预测置信度统计分析

	p1_conf	p2_conf	p3_conf
mean	0.600421	1.340285e-01	5.985643e-02
std	0.269476	1.000261e-01	5.059723e-02
max	1.000000	4.880140e-01	2.734190e-01
min	0.044333	1.011300e-08	1.740170e-10

p1_conf，p2_conf 和 p3_conf 的直方图分别如图 3，图 4 和图 5 所示：

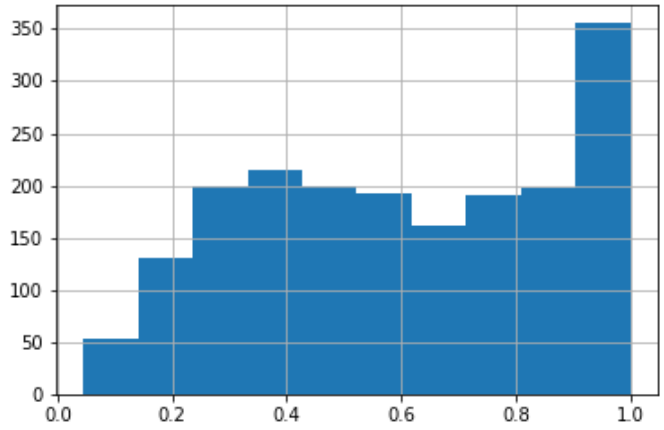

图 3　p1_conf 直方图

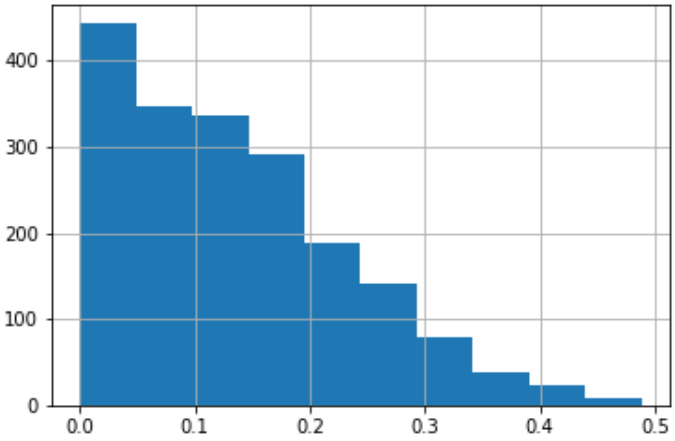

图 4　p2_conf 直方图

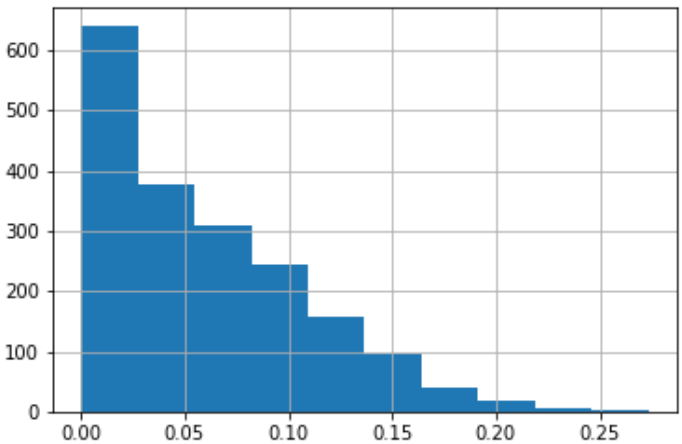

图 5　p3_conf 直方图

从上述结果中可以看出，总体上 p1_conf > p2_conf > p3_conf,这与预期的一致,p2_conf 和 p3_conf 全部在 0.5 以下。但是 p1_conf 的均值只有 0.598，而且 75%的值为 0.848，也就是预测算法的准确度还有待提高。

5. 狗的品种

在这里假设，如果某品种预测可信度大于 0.5 时，可以算作此类品种。则可以得出如表 5 所示的数据。

表 5 部分狗品种的统计数

	stage_count
Golden Retriever	119
Labrador Retriever	65
Pembroke	64
Chihuahua	47
…	…

从表 5 中可以看出，狗的品种中，养的最多的前三个品种分别为 Golden Retriever，Labrador Retriever 和 Pembroke。