第一句：

Indicate whether the similarity data, as reported in the *data\_influence* data set, suggest that the identified influencers in fact influence the respective artists.

说明*data\_influence* 数据集中报告的相似性数据是否表明识别出的影响者实际上影

响了各自的艺术家。

首先通过并查集算法找所有节点中的祖宗节点

“祖宗节点”的概念：可以在每个领域中找到少部分艺术家，这部分艺术家影响到了全部艺术家，所有艺术家都可以找到他们唯一的“祖宗节点”

并查集算法：

主要用于解决一些**元素分组**的问题。它管理一系列**不相交的集合**

并查集顾名思义就是有“合并集合”和“查找集合中的元素”两种操作的关于数据结构的一种算法。通常是在开始时让每个元素构成一个单元素的集合，然后按一定顺序将属于同一组的元素所在的集合合并，其间要反复查找一个元素在哪个集合中。

主要操作：

1. 初始化

把每个点所在[集合](https://baike.baidu.com/item/%E9%9B%86%E5%90%88)初始化为其自身。

通常来说，这个步骤在每次使用该数据结构时只需要执行一次，无论何种实现方式，[时间复杂度](https://baike.baidu.com/item/%E6%97%B6%E9%97%B4%E5%A4%8D%E6%9D%82%E5%BA%A6)均为O(N)。

一个集合内的所有元素组织成以代表元为根的树形结构

1. 查找

查找元素所在的集合，即根节点。

要确定x所在的的集合，也就是确定集合的代表元。可以沿着parent[x]不断在树形结构中向上移动，直到到达根节点。

1. 合并

将两个元素所在的集合合并为一个集合。

通常来说，合并之前，应先判断两个元素是否属于同一集合，这可用上面的“查找”操作实现。

对于每一个元素 parent[x]指向x在树形结构上的父亲节点。如果x是根节点，则令parent[x] = x。

通过以上步骤，我们得到了2220个祖宗节点，各领域祖宗节点数目如下：

Pop/Rock 1117

Electronic 339

Reggae 98

Jazz 263

Country 297

Comedy/Spoken 53

R&B; 412

Classical 63

Latin 251

Vocal 206

Folk 199

Easy Listening 38

International 138

Avant-Garde 23

Blues 103

Stage & Screen 44

New Age 90

Religious 135

Children's 4

Unknown 9

部分祖宗节点的子节点数目如下：

934243: 402,

25462: 7,

898336: 869,

816890: 222,

2993710: 7333,

934069: 196,

378288: 18,

988386: 2084,

864257: 646,

175775: 2,

688804: 24,

829643: 268,

798792: 32,

151290: 11,

286662: 4,

12654: 8,

959770: 1104,

为了数值化表现出祖宗节点的6项指标的特征与其子节点6项指标特征的相似性，我们定义子节点的某一特征值与父节点对应特征值的差在如下范围内时，称该子节点与其祖宗节点在该指标上相似：

和influencer的指标之差在以下范围内认为是相似的

'emotion', +-0.5

'rhythm', +-5

'purity',+-0.25

'duration\_ms', +-20000

'popularity',+-10

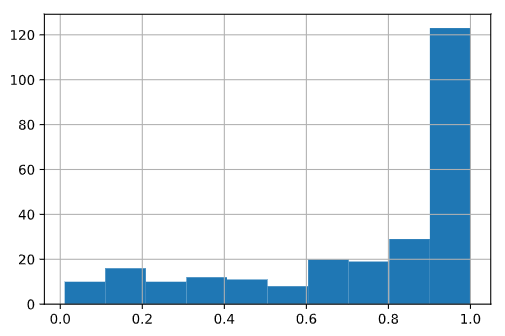
'instrumentalness'+-0.01

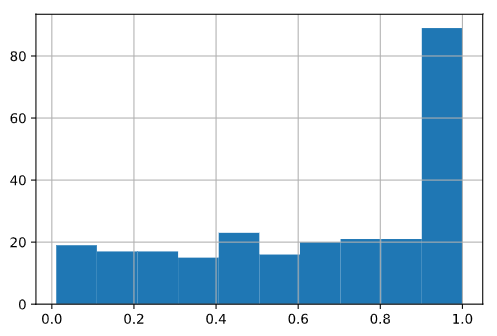
然后我们分别统计各个祖宗节点的所有子节点中，各个特征与其祖宗节点相似的子节点数占所有子节点数的比例，得到每个祖宗节点的6项指标对其所有子节点的影响力（体现在其子节点的某项指标和祖宗节点该项指标相似的比例大，则该祖宗节点的该指标influence the respective artists。

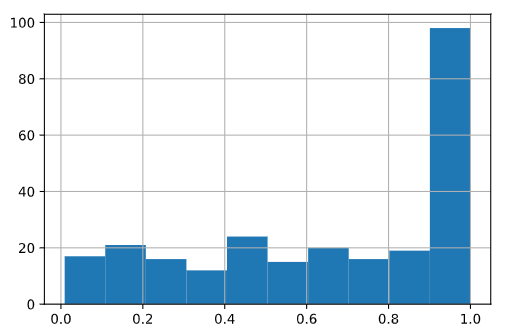
在统计结果中，我们去掉6项比例之和为6的祖宗节点，因为出现这种情况的原因是该祖宗节点仅有它自己一个追随者

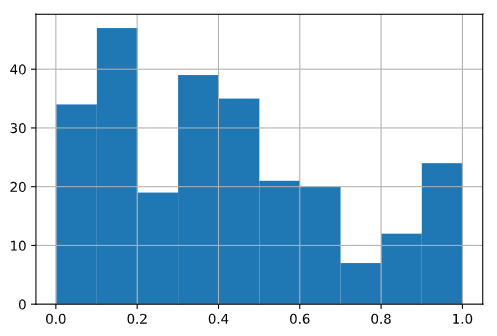
统计结果如下：

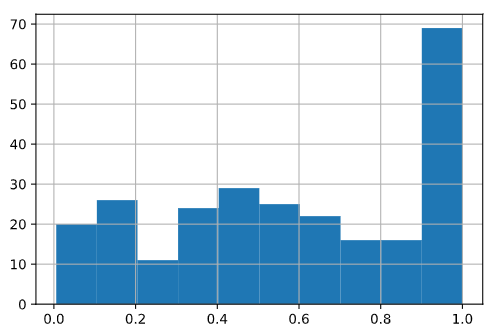
分别是'emotion',  'rhythm',   'purity',   'duration\_ms',  'popularity',   'instrumentalness'的数值分布

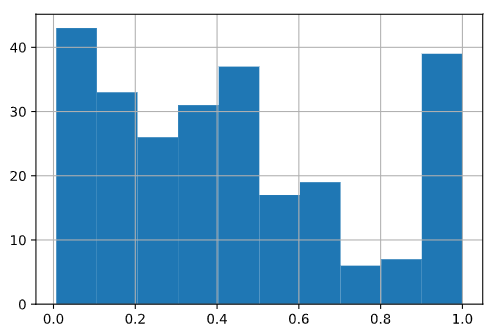












由以上图片可知，

the identified influencers（祖宗节点） in fact influence the respective artists（祖宗节点的子节点）.

具体体现在在'emotion',  'rhythm',   'purity' 3个指标上，祖宗节点与子节点的相似度较高，在这四个特性上祖宗节点是会影响子节点的