摘要：

针对问题一，我们建立从影响者到追求者的子系，以不同音乐genre为一个单位。根据三方面直接追求者人数，影响时间范围，影响的音乐genre数量指标，建立一个影响度的灰色关联综合指标，选出每个音乐genre中影响度评分最高的influencer。对于选出的影响者，绘制树状图去表示不同音乐genre的子系并进行分析。其中POP/ROCK中代表人物为The Beatles，其余结果在正文中展示。

针对问题二，我们建立关于音乐相似度的模型，并判断流派内的艺术家与流派间的艺术家哪种更相似。首先我们利用主成分分析法将data\_by\_artist中的14指标降维为5个主成分。其次，我们用余弦相似性对音乐家进行相关性分析。比较流派内与流派间艺术家的相似性，其中有6个genres的平均相似度大于流派间的平均相似度。

针对问题三，我们比较流派之间和流派内部的相似之处和影响并揭示流派是如何随着时间的推移而变化的。首先利用问题二的数据以及相似性热力图找出流派之间和流派内部影响力较大的音乐家进行比较分析，具体内容见正文。 其次我们绘制了盛行genre随时间的变化图，最后我们用Sen’s趋势估计法判断总体趋势的变化情况。其中POP/Rock这个种类内人数的总体趋势呈上升状态，上升的趋势程度较大。

针对问题四，我们要判断影响者是否直接影响追求者，具体从哪些方面进行影响；判断出音乐的那些特征更具有感染力；说明更具有感染力的特征对某个特定的音乐家成为一个著名的流行者中扮演的角色。首先，影响程度可以通过指标之间的相似性反映，我们选择一country genre的代表人物Hank Williams进行分析。对代表人物与追求者利用问题二中的五个主成分进行余弦相似性分析，发现不是直接影响的，再计算指标之间的余弦相似性。其次我们可以通过计算其余指标与popularity的 Spearman秩相关相关系数，选出与popularity指标相关性高。感染力的指标：populrarity, valence,tempo,loudness,acousticness,duration。最后，我们选择country genre子网为主要研究对象。对选出的指标做灰色关联综合性分析，找出综合指标突出的人物。分析综合指数高的音乐家的影响力。结果说明感染力指标对一个音乐家成为影响者并没有太大关系。

问题一分析：

问题一要求我们建立从影响者到追求者的子系。利用influence\_data中的数据对数据进行分析。首先我们观察数据发现影响者的追求者包括直接追求者和间接追求者。其中直接追求者是被影响者直接影响从而成为追求者；间接追求者是被直接追求者所影响的追求者。

其次，音乐的genre随着时代的前进而增加，产生新的genre。其中音乐genre是指音乐的风格和特点，能本质地反映出时代、民族、音乐家个人的思想观念、审美理想、精神气质等内在特性的外部印记。因此我们将音乐的genre作为划分音乐人的重要标签，对音乐人进行划分。

再次，我们从音乐genre中找出具有代表性的影响者，建立一个影响度的综合指标。从表格数据可以看出影响者从三个方面体现影响度，分别是直接追求者人数，追求者所处的时间区间，影响的音乐genre数量。其中对于追求者，由于间接追求者可能是因为上一级影响者的个人风格而去追随该影响者，并不是因为最顶级的影响者，因此我们只考虑直接追求者。由于这三个因素之间关系既相互联系，又各有特点，因此我们选用灰色关联法进行综合评价，得出影响度评分的综合指标。计算出每个音乐种类中影响者的影响度评分，选出具有代表性的影响者。

最后，对于选出的影响者，绘制树状图去表示不同音乐genre的子系并进行分析。

5.1.1模型的建立与求解：

5.1.1.1名词解释：

* 直接追求者：被影响者直接影响从而成为追求者。
* 间接追求者：被直接追求者所影响的追求者。

5.1.1.2建立影响者影响力的综合评价模型

*Step1* 建立分类标准

根据influence\_data中的数据类型进行分析，我们发现音乐人的属性有姓名，编号ID，音乐genre，influencer—active—start。分析以上数据类型，由于音乐genre是指音乐的风格和特点，能本质地反映出时代、民族、音乐家个人的思想观念、审美理想、精神气质等内在特性的外部印记。因此我们将音乐的genre作为划分音乐人的重要标签，对音乐人进行划分。

*Step2* 指标的选取

为了体现出一个音乐人的影响力，我们从以上三个方面去分析：

* 直接追求者数量(N)：追求者的数量可以直接的反映该影响者受欢迎的程度，也可以反映出该影响者的影响力。由于间接追求者可能是因为上一级影响者的个人风格而去追随该影响者，并不是因为最顶级的影响者，因此在这里我们只考虑直接追求者的数量。
* 影响者的影响时长(T)：追求者的时间属性范围反映了该影响者的影响的时间范围。范围越大表明该影响者影响力深远（形容时间），范围越小表示该影响者只是昙花一现，是一时的潮流。

其中是followers—active—start数据中最大的，其中是 followers—active—start数据中最小的。

* 影响的音乐genre(P)：虽然每个音乐家的genre不一样，但是影响者所影响的曲风范围也可以反映出该影响者的影响力。影响的范围越大说明该影响者对音乐的造诣很深，可以从细枝末节去渗透到各个层面。

由于音乐genre的随着时代的前进而增加，所以在计算这项指标的时候我们根据追求者所在年代音乐genre数量进行加权求和，具体公式如下所示：

其中表示为追求者所在时代存在的音乐genre总数。

*Step3* 建立综合评价模型

根据上述我们选择了直接追求者数量、影响者的影响时长、影响的音乐genre这三个指标。由于各个指标之间相互联系但是有反应不同方面的性质，因此要建立一个科学合理的综合评价模型，我们选用灰色关联模型，具体模型如下所示：

* **选取母指标**

由于直接追随者人数最直观的反映影响者影响力的程度，因此直接追求者人数作为母指标。

* **计算关联系数**

定义母指标与在第点的关联系数为：

其中的计算公式如下所示：

的计算公式如下所示：

的取值范围为,这里我们取。

* **计算关联度**
* **构建综合评价模型**

从上一步中求出的关联度，计算各个指标的权重：

作为各个指标的权重，得到综合指标.

|  |  |  |  |
| --- | --- | --- | --- |
| 指标 | 直接追求者数量(N) | 影响者的影响时长(T) | 影响的音乐genre(P) |
| 权重 | 0.963244 | 0.915657 | 0.8207 |

最后综合指数为影响力的评分，评分越高表示影响力越大。

*Step4* 选出每个genre中的代表人物

根据上述评价模型计算出每个影响者的影响力的综合评分，选出每个领域评分最高的人作为该音乐genre的代表性人物。选出的19个代表人物的信息如下表所示：

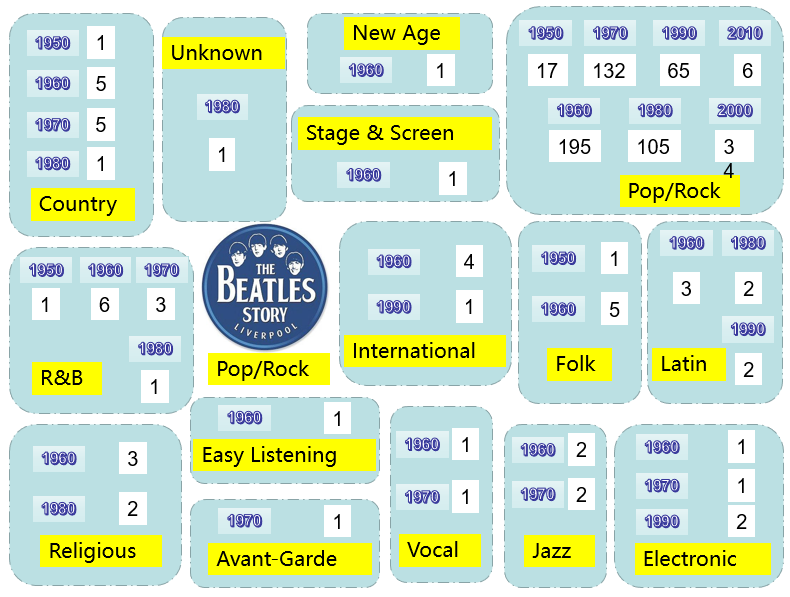
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| artist\_name | artist\_id | genre | T | N | P |
| Terry Riley | 750519 | Avant-Garde | 40 | 36 | 5 |
| Muddy Waters | 608701 | Blues | 60 | 113 | 6 |
| Alvin & the Chipmunks | 744969 | Children's | 40 | 3 | 2 |
| Steve Reich | 37658 | Classical | 40 | 24 | 5 |
| Spike Jones | 744548 | Comedy/Spoken | 60 | 20 | 9 |
| Hank Williams | 549797 | Country | 80 | 184 | 6 |
| Henry Mancini | 678009 | Easy Listening | 50 | 19 | 6 |
| Kraftwerk | 104714 | Electronic | 40 | 108 | 4 |
| Woody Guthrie | 577531 | Folk | 80 | 80 | 7 |
| Fela Kuti | 138833 | International | 50 | 23 | 5 |
| Miles Davis | 423829 | Jazz | 60 | 114 | 12 |
| Tito Puente | 607283 | Latin | 50 | 21 | 6 |
| Mannheim Steamroller | 675495 | New Age | 40 | 9 | 5 |
| The Beatles | 754032 | Pop/Rock | 50 | 615 | 15 |
| Marvin Gaye | 316834 | R&B; | 60 | 169 | 7 |
| Desmond Dekker | 819344 | Reggae | 30 | 21 | 2 |
| Amy Grant | 24944 | Religious | 40 | 16 | 5 |
| Ennio Morricone | 798662 | Stage & Screen | 50 | 30 | 3 |
| Billie Holiday | 79016 | Vocal | 80 | 106 | 9 |

5.1.1.3代表性人物为中心的关系图

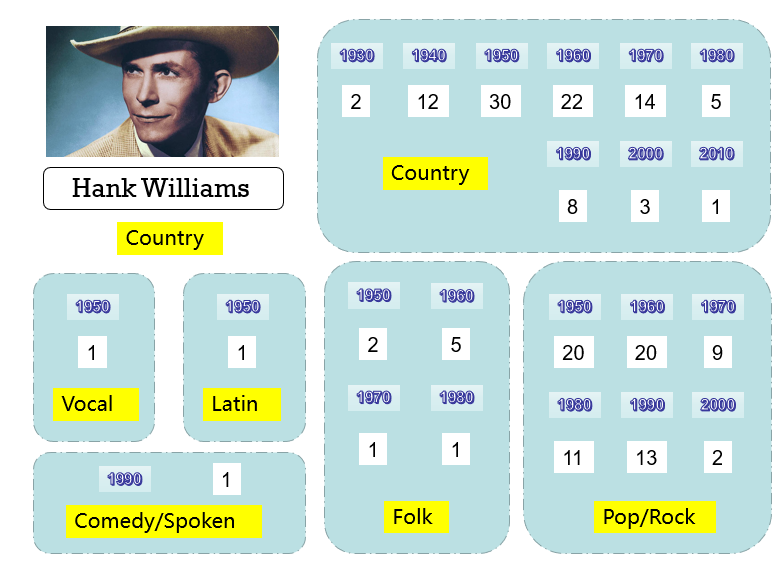
对于每个音乐genre中的代表性人物绘制出关系网图，其中包括直接追求者和间接追求者树状图，并标明追求者的年代属性和音乐genre属性,由于数据过于复杂，我们选择The Beatles乐队和Hank Williams进行可视化子网分析，其余主要影响者我们用表格数据进行展示。

5.1.2 结果分析

下面展示有关于The Beatles的直接影响者，影响类别和年份的关系图：

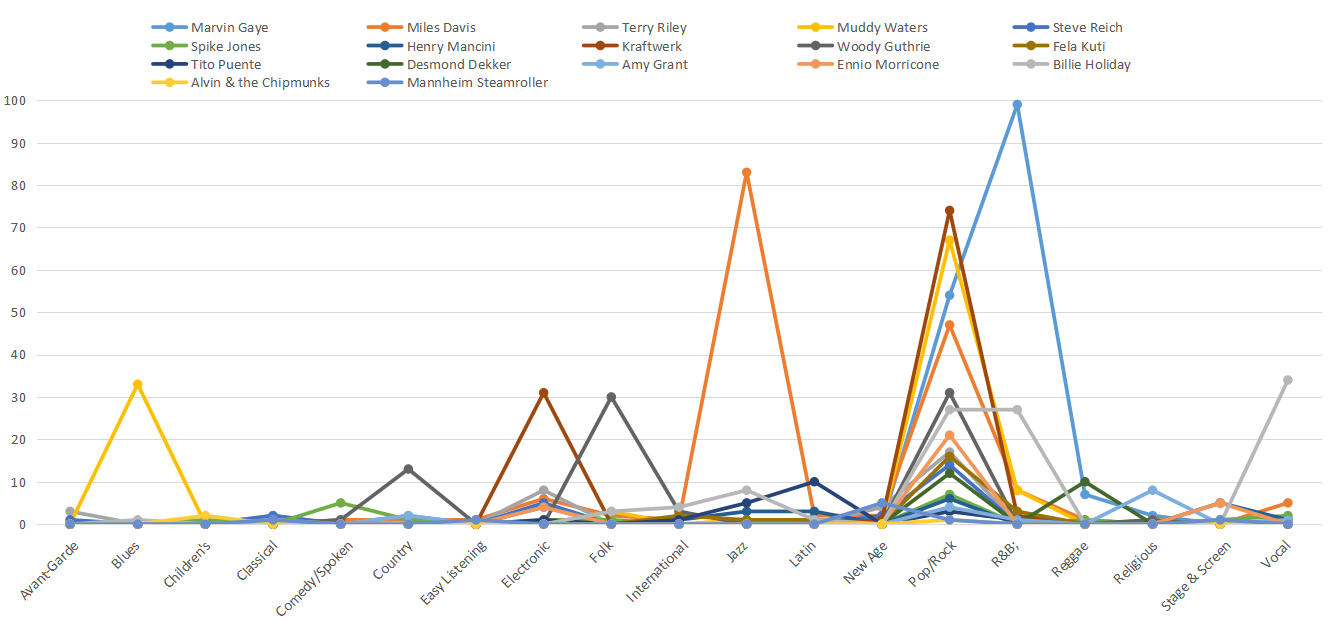


The Beatles是Pop/Rock gerne的乐队，它们的追求者数多，影响时间长，从1960开始它们便在世界上拥有大量的followers，尤其是在60，70，80年代，它们拥有超过200人的直接追随者，但随着时代的变迁它们所影响的人数也在下降，影响力也在衰退。它的影响的music gerne范围十分广泛，从它们自身所属的Pop/Rock到Country、Folk、R&B、Latin等gernes都有它们的直接影响者，但在这其中它们主要的影响力还是在Pop/Rock上，超过90%的直接追随者都来自于Pop/Rock gernes中不同年代的人。



Hank Williams, Country gerne早期的代表人物，影响力很大，有很多在60，70年代具有较大影响力的artists都是它的直接follower，但由于他所属的年代距今久远，导致属于它的音乐风格在历史的车轮中逐渐淡出了舞台，影响力越来越小，直接追随者也呈明显的下降趋势。他所影响的音乐gernes有6个，直接追随者的音乐genre以Country和Pop/Rock gernes为主。

上述做出两个音乐genre可视化子网，由于代表人物信息众多，其余代表人物信息由下面图像展示：



上图中每条折线代表了一位在他所属的music genre中有很大影响力的artist。我们通过观察上图发现，每条折线都有一个顶峰，并且这个顶峰基本都出现在他们自身所处的music genre中；其次每一位artist在Pop/Rock genre中都有较大的影响力和较多的direct followers，由此判断，Pop/Rock genre比较大众，能广泛被人们所接受；另外Jazz genre的Miles Davis和R&B genre的Marvin Gaye他们在各自所处的genre中的direct followers多于其它的influencer，从这点可以看出，他们的song认可度较高，符合时代的潮流。

问题二分析：

问题二要求我们建立关于音乐相似度的模型，并判断流派内的艺术家与流派间的艺术家哪种更相似。首先通过观察data\_by\_artist的数据，我们观察到有artist\_name,artist\_id,energy等14个指标，很多指标之间相似性较高，反映的信息在一定程度上有重叠，在研究相关性时变量太多增加问题的复杂性，因此我们选择主成分分析法将指标降维，用较少的指标综合反映信息。用主成分分析法计算出指标的信息贡献率和累计贡献率，并对指标进行排序，选择贡献率在90%以上的指标为主成分，对主要成分加权求和得到音乐家的综合指数去反映其特征。

其次，我们对音乐家进行相关性分析。由于每个音乐家的特征信息是相同维度的因此我们选用余弦相似度对每类歌曲进行相关性计算。

最后，对不同音乐genre的歌曲之间，我们选择问题一中做出的子系中的代表音乐人的作品进行相关性的分析，并与相同genre音乐人之间平均相关性进行比较分析。

5.2.1模型的建立

5.2.1.1 主成分分析

***Step1***指标的选取

将所有数据进行分类，挑选有代表性的指标。

* 基本信息：

音乐家姓名、音乐家ID。

* Characteristics of the music:

danceability,energy,valance,tempo,loudness,mode,key

* Type of vocals:  
  acousticness,instrumentalness,liveness,speechiness.
* Descriiption:

duration\_ms,popularity,count.

其中基本信息作为描述音乐家信息，因此基本信息不是指标；Characteristics of the music, Type of vocals, Descriiption中的数据均可以作为反映歌曲的指标。

最后我们选取的指标有以下14个：

danceability,energy,valance,tempo,loudness,mode,key，acousticness instrumentalness,liveness,speechiness,explicit, duration\_ms,popularity,count.

***Step2*** 主成分分析法

由于数据给出的15个指标可以多方位的反映歌曲主要特征，指标与指标之间具有一定的关联性，反映问题的信息会有所重叠。选用主成分分析法对15个指标变为少数几个综合性更强的指标，计算出各个指标的信息贡献率与累计贡献率，以主成分的贡献率为权重，建立主成分的综合评价模型。

15个指标数据构成的序列为。

* + 对于数据预处理以后的15个指标的序列进行归一化处理：
  + 计算相关系数矩阵：
  + 计算特征值与特征向量：

计算相关矩阵R的特征值，和对应的标准化特征向量。

* + 计算信息贡献率和累计贡献率：

14个指标的信息贡献率和累计贡献率如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 指标 | 信息贡献率 | 累计贡献率 | 指标 | 信息贡献率 | 累计贡献率 |
| 1 | 0.227378746 | 0.227378746 | 8 | 0.06040273 | 0.810419479 |
| 2 | 0.122550481 | 0.349929227 | 9 | 0.053311569 | 0.863731048 |
| 3 | 0.100385594 | 0.450314821 | 10 | 0.046942923 | 0.910673971 |
| 4 | 0.091267136 | 0.541581957 | 11 | 0.039310617 | 0.949984588 |
| 5 | 0.073334632 | 0.614916589 | 12 | 0.02339942 | 0.973384008 |
| 6 | 0.0702334 | 0.685149989 | 13 | 0.020041915 | 0.993425923 |
| 7 | 0.06486676 | 0.750016749 | 14 | 0.006574077 | 1 |

根据以上表格选出5个主成分，使特征向量累计贡献率超过90%（五个主成分数据如下所示）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 指标 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 第一主成分 | -0.164 | 0.5499 | -0.21686 | 0.305966 | 0.172624 | -0.06488 | 0.0587 |
| 第二主成分 | -0.492 | -0.172 | 0.086142 | -0.02197 | -0.00688 | -0.15297 | 0.0641 |
| 第三主成分 | -0.196 | 0.5667 | 0.024154 | 0.08834 | -0.11204 | -0.39041 | 0.107816 |
| 第四主成分 | -0.211 | -0.129 | 0.154015 | -0.28158 | -0.35194 | -0.48096 | -0.00034 |
| 第五主成分 | -0.487 | -0.066 | 0.04583 | -0.0733 | 0.059896 | 0.087899 | 0.002733 |
| 指标 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 第一主成分 | 0.139356 | 0.088082 | 0.090994 | 0.280688 | 0.127535 | -0.57008 | -0.19167 |
| 第二主成分 | 0.12906 | -0.23191 | 0.092694 | -0.20371 | 0.034249 | 0.121726 | -0.7453 |
| 第三主成分 | 0.161702 | -0.08132 | -0.10595 | 0.030528 | -0.0844 | 0.597468 | 0.21478 |
| 第四主成分 | -0.22257 | 0.586166 | -0.00349 | 0.246121 | 0.056696 | -0.15186 | -0.01281 |
| 第五主成分 | -0.02634 | -0.1414 | -2E-05 | -0.18506 | 0.697935 | -0.08118 | 0.435905 |

利用上述权重记为矩阵A，将14个指标降维到5个主成分，利用下面的公式进行计算：

由于计算后数据量较大，在正文中不做展示。

5.2.1.2 相关性分析

***Step1*** 上述计算出的所有音乐家五项主成分的部分数据如下表所示：  
***Step2*** 计算两两音乐家关于五个主成分之间的余弦相似度，将音乐家的主成分序列记为，关于余弦相似度的计算公式如下所示：

得到相似度矩阵部分数据如下所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Frank Sinatra | Vladimir Horowitz | Johnny Cash | Billie Holiday | Bob Dylan | The Rolling Stones |
| Frank Sinatra | 1 | 0.956 | 0.938 | 0.964 | 0.976 | 0.979 |
| Vladimir Horowitz | 0 | 1 | 0.967 | 0.99 | 0.979 | 0.979 |
| Johnny Cash | 0 | 0 | 1 | 0.988 | 0.979 | 0.981 |
| Billie Holiday | 0 | 0 | 0 | 1 | 0.990 | 0.991 |
| Bob Dylan | 0 | 0 | 0 | 0 | 1 | 0.999 |
| The Rolling Stones | 0 | 0 | 0 | 0 | 0 | 1 |

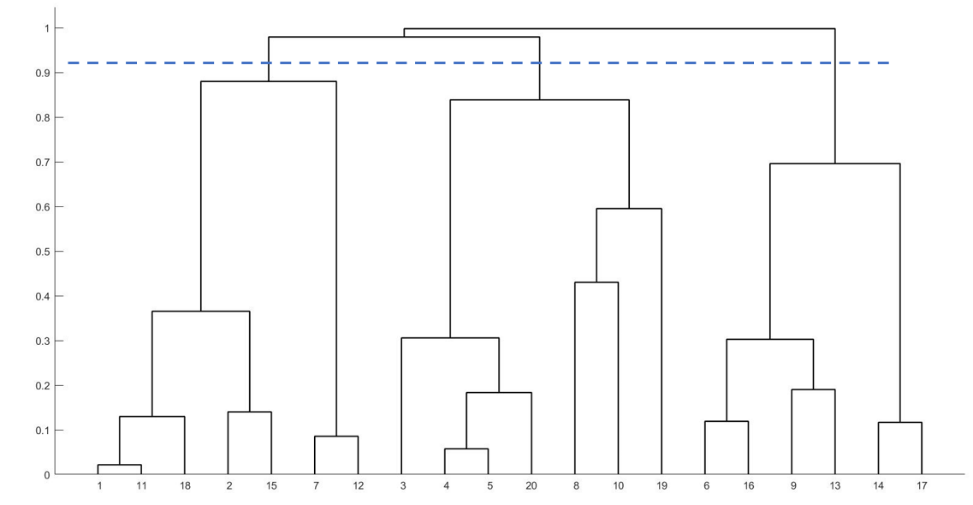
**余弦值越接近1，就表明夹角越接近0度，也就是两个向量越相似，就表明两个音乐人之间的相关性越高。**  
5.2.2 流派内音乐家的相似性

对流派内音乐家进行分析时，我们主要了Pop/Rock和R&B两种音乐genre进行分析。找出上述相关矩阵中关于Pop/Rock和R&B两种音乐genre中音乐人之间的相关性数据，重新制作两个相关性矩阵，并将两种音乐genre的相关性信息通过聚类分析和热力学图像进行可视化表示。

Pop/Rock流派聚类分析如下所示，由于我们已经求出相关性矩阵，因此我们选用R型聚类分析。其中我们用数字代表音乐家名字，对应关系如下表所示：

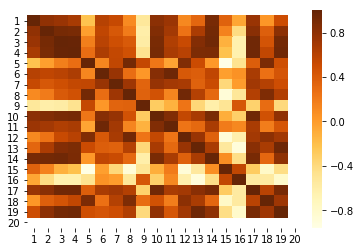
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | The Beatles | 8 | Casting Crowns | 15 | Be Bop Deluxe |
| 2 | Soundgarden | 9 | Teenage Fanclub | 16 | Klaatu |
| 3 | System of a Down | 10 | The Apples in Stereo | 17 | BarlowGirl |
| 4 | T. Rex | 11 | The Beach Boys | 18 | Don McLean |
| 5 | Big Star | 12 | Blind Melon | 19 | John Fogerty |
| 6 | Sloan | 13 | Badfinger | 20 | Bob Weir |
| 7 | Toto | 14 | Christine McVie |  |  |

Pop/Rock流派聚类分析如下所示：



从图中我们可以看出，聚类结果将音乐人编号为（1，11，18，2，15，7，12），（3，4，5，20，8，10，19），（6，16，9，13，14，17）聚为3类。相同类中的音乐家的相似性较高。

Pop/Rock流派相关性的热力图如下所示：

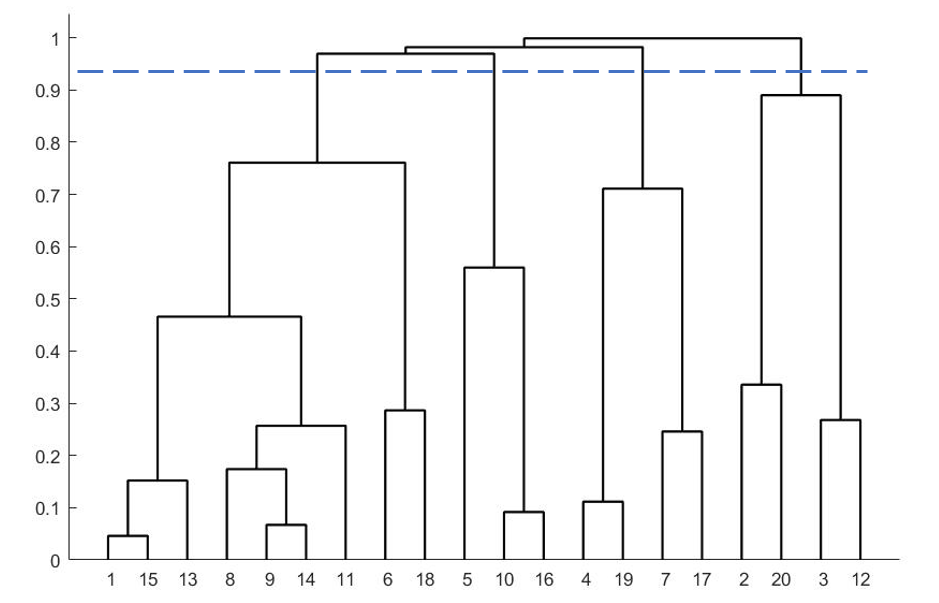


从上述热力图我们可以直观的看出音乐家之间的相似性，颜色越深表示相似性越高

有关于R&B流派，数字与音乐家名字对应关系如下表所示：

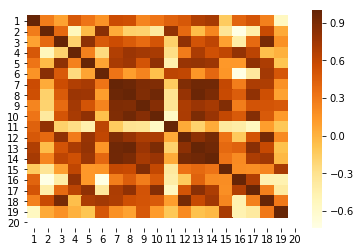
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | Prince | 8 | D'Angelo | 15 | Al Green |
| 2 | Patti LaBelle | 9 | Lisa Stansfield | 16 | Don Covay |
| 3 | Trey Songz | 10 | Ready for the World | 17 | The Flamingos |
| 4 | Morris Day | 11 | Chaka Khan | 18 | The Mills Brothers |
| 5 | The-Dream | 12 | Dionne Warwick | 19 | The Stylistics |
| 6 | The Weekend | 13 | Commodores | 20 | Lamont Dozier |
| 7 | TLC | 14 | Dinah Washington |  |  |

R&B流派聚类分析如下所示：



从图中我们可以看出，聚类结果将音乐人编号为（1，15,13,8,9,14,11,6,18），（5,10,16），（4,19,7,17）,(2,20,3,12)聚类成4类。

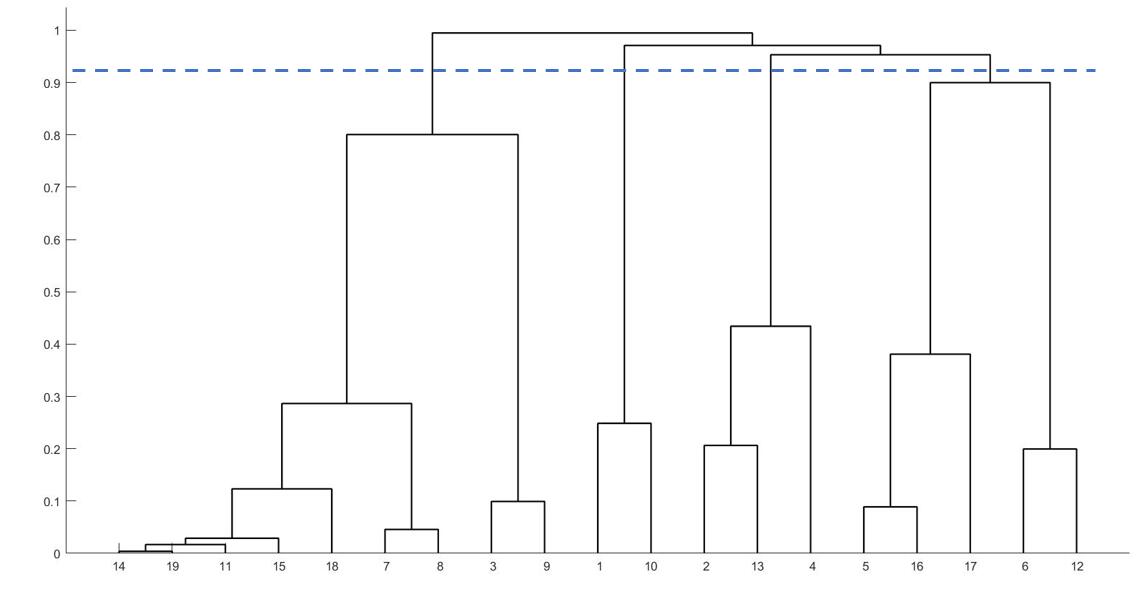
R&B流派相关性的热力图如下所示：



5.2.3 流派间音乐家相似性

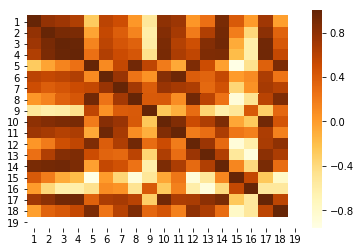
对流派间的音乐家，我们选用第一问中各个流派中代表音乐人之间的相似性进行分析，找出上述相关矩阵中各genre代表音乐人之间的相关性数据，重新制作相关性矩阵，并进行聚类分析和绘制热力学图像。

聚类分析如下图：



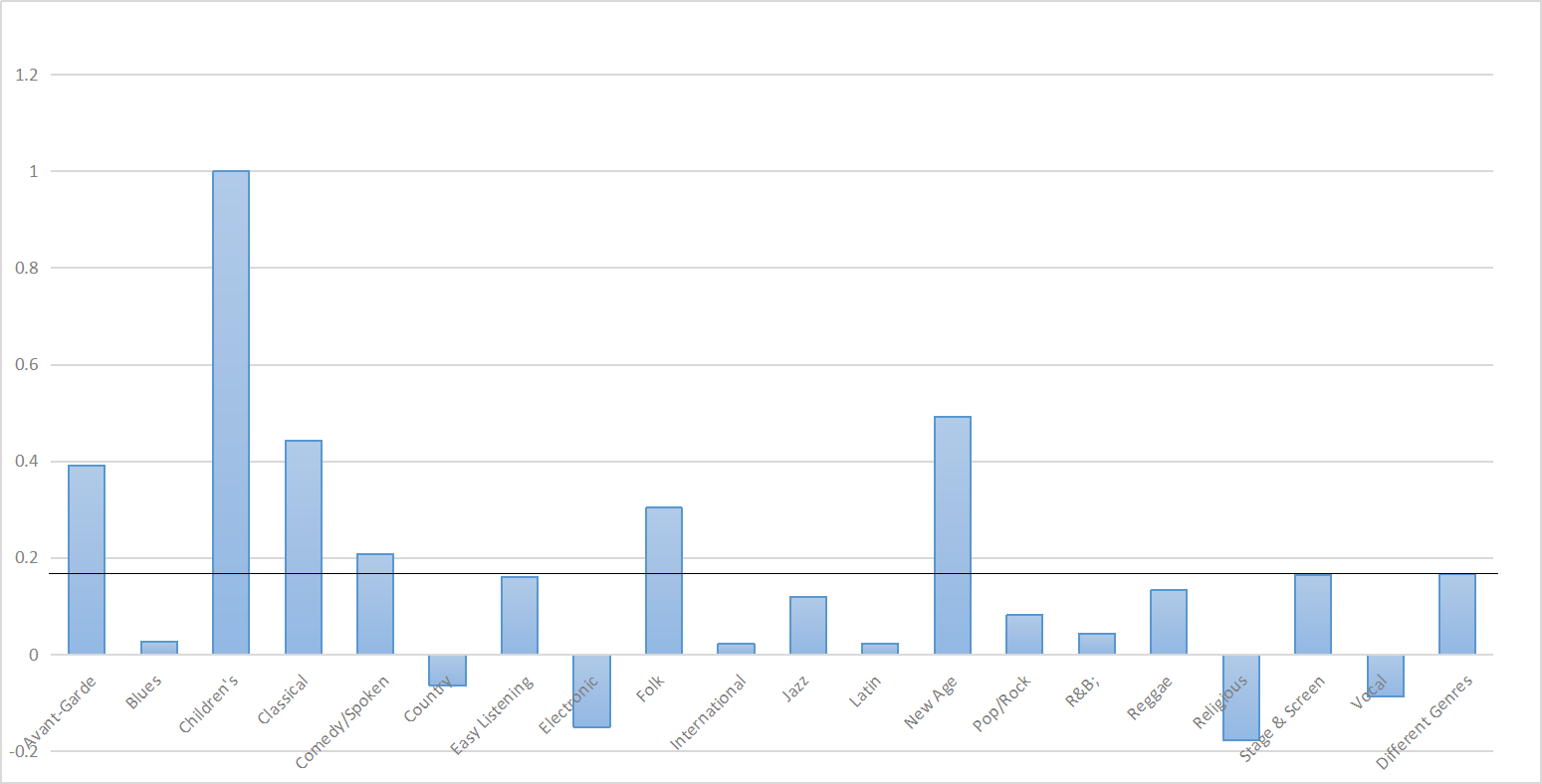
从图中我们可以看出，聚类结果将音乐人编号为（14，19，11，15，18，7，8，3，9），（1，10），（2，13，4），（5，16，17，6，12）聚成4类。相同类中的音乐家的相似性较高。

热力图如下所示：



5.2.4 两类相似性数据之间的比较

求出所有音乐genre的相似性的平均值和不同genre相似性的平均值为方便观察，我们用柱状图进行表示：



通过相似度平均值的图像可以明显地观察到6个genres内的平均相似度大于不同genres间的平均相似度，剩下的genres内的平均值小于不同genres间的平均值。其中Children’s genre的相似度平均值最高，达到了1，这是因为该genre中只有一个artist，故而为1，但由于样本数据过少，所以该genre内的相似度平均值不具备任何参考价值。

问题三分析：

问题三首先要求我们比较流派之间和流派内部的相似之处和影响并说明有些流派与其他流派关系，其次要求我们揭示流派是如何随着时间的推移而变化的。

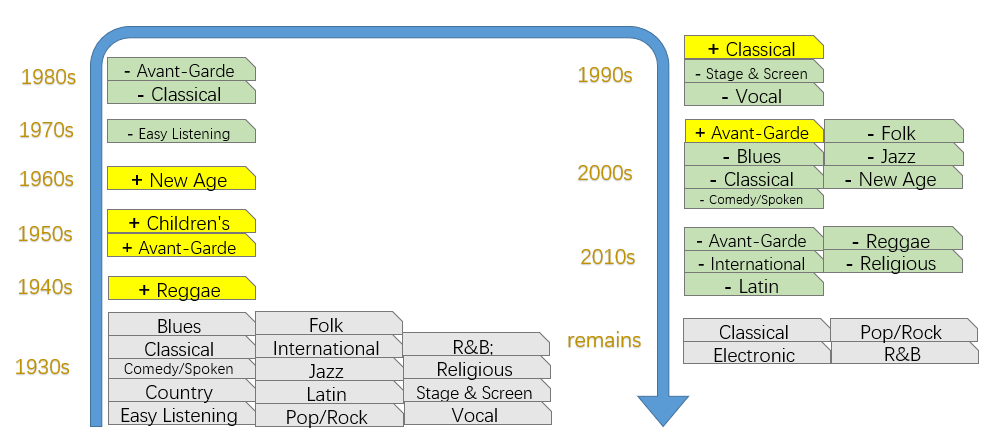
首先利用问题二的数据以及相似性热力图找出流派之间和流派内部影响力较大的音乐家，通过对比这些音乐家的各类信息，分析出流派内部和流派之间的关系。

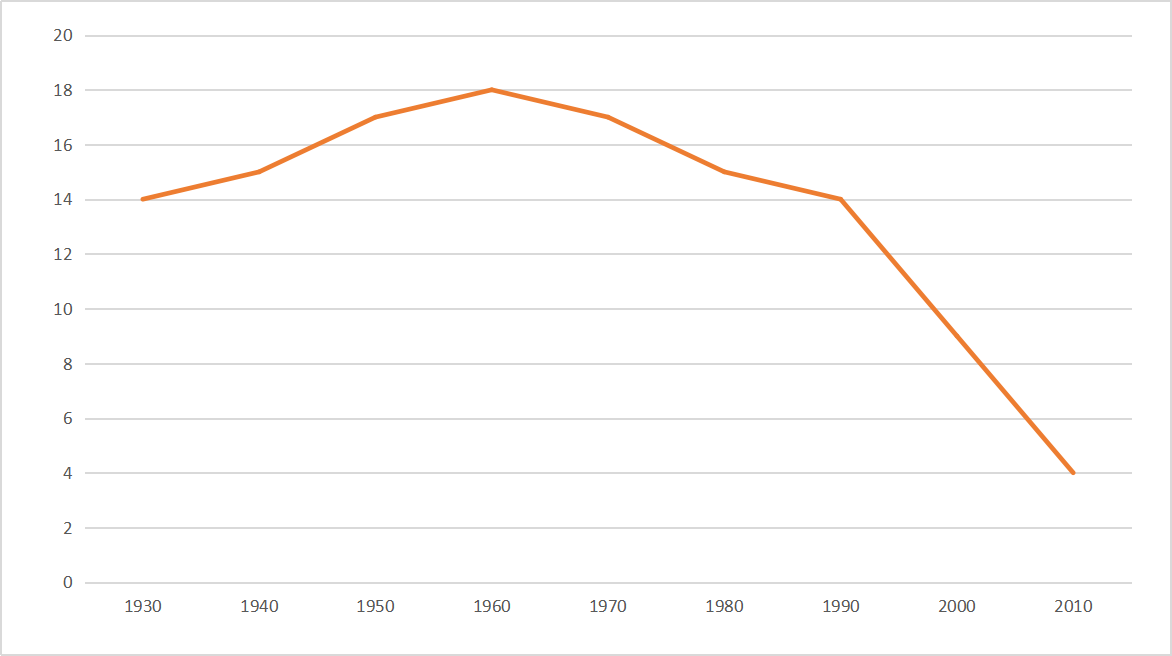
其次我们分析流派是如何随着时间的推移而变化，对该问题的分析我们分为以下两个部分。第一部分，绘制盛行流派随时间变化的关系图，反映出随着时代变化有哪些新的genre 出现；第二部分，分析某一种流派内人数的变化趋势，这里我们选择具体分析Pop/Rock这个种类的数据。对题目中给出的数据进行整理，整理出各个时间段Pop/Rock中的人数。为了反映出人数的趋势变化，用Sen’s趋势估计法判断总体趋势的变化情况，做出折线图对斜率变化较大的点进行分析。

5.3.1 流派间与流派内的关系

5.3.2 盛行流派随着时间演变

我们用图文结合的方式直观的表现出盛行流派随着时间演变：





通过图片可以知道在1930年，世界上有Pop/Rock等14盛行的个music genres，随着时间推移，到1960为止，流行music genres呈明显地上升趋势，1960年流行music genres数量达到顶峰，有18个，其中从1940到1950，上升趋势最大。但在1960年以后流行的music genres呈现了下降地趋势，1990到2010，盛行music genres减少了10个。从折线图可以直观看到，在世界发展历程中，早期和现在流行的music genres的种类较为单一，大众审美趋向一致，而在中期流行music genres多元、丰富。

5.3.2.1 流派随时间增加的关系图

5.3.2.2 Pop/Rock流派内人数变化趋势

对于整体趋势的分析，要表现趋势的变化应从人数的变化上看，我们使用Sen’s趋势检验。选取一个可以反映整体变化率的指标--倾率，倾率是各个时间节点向后计算出斜率，最后将所求出的斜率取均值得到，因此倾率可以表示整体趋势的总变化。若倾率为正值，则说明总体趋势呈上升趋势；反之若倾率为负值，则说明总体趋势是下降的。倾率绝对值的大小可以反映出趋势变化的程度。

倾率的计算公式如下：

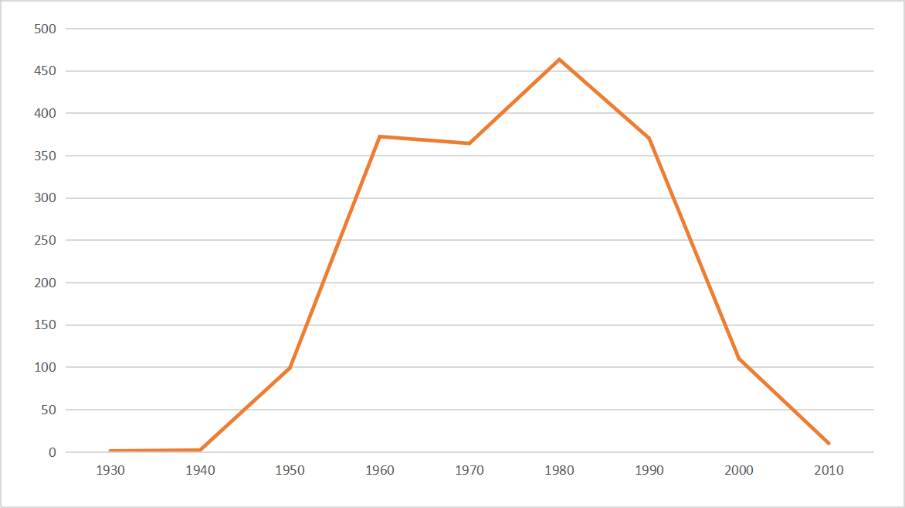
其中为每十年人数，其中表示向后面的年份计算斜率。

计算得到人数变化的倾斜率为：

49.4889762

得到倾斜率的值为正，说明Pop/Rock这个种类内人数的总体趋势呈上升状态，上升的趋势程度较大。

每个节点的人数我们用折线图进行表示：



各个节点的数据如下表所示：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1930 | 1940 | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010 |
| 1 | 2 | 99 | 372 | 364 | 463 | 370 | 110 | 10 |

从图中我们发现在1950年到1960年间该genre的人数增长飞快，斜率达到27.3，结合问题一我们发现这个时期著名The Beatles开始盛行，该乐队是所有影响者中直接追求者最多的音乐人,因此增长率较快。Pop/Rock人数自1980到达顶峰后开始下降，下降年份最快的时间段是1990年到2000年，观察Influence\_data中的数据可以发现，从1990年开始R&B genre开始盛行，我们判断其他的音乐genre人数的增加导致Pop/Rock 人数的减少。

问题四分析：  
 问题四要求对三个子问题进行分析，第一个是根据 data\_influence 中的数据判断影响者是否直接影响追求者，具体从哪些方面进行影响。第二个是判断出音乐的那些特征更具有感染力。第三个是更具有感染力的特征对某个特定的音乐家成为一个著名的流行者中扮演着相似的角色。

首先，我们对影响者是否直接影响追求者进行分析。影响程度可以通过指标之间的相似性反映。我们选择一个流派进行分析，结合问题一给出的详细子网，我们对country genre的代表人物Hank Williams进行分析。对代表人物与追求者利用问题二中的五个主成分进行余弦相似性分析，做出折线图观察影响者是否直接影响追求者；再分别计算追求者在data\_by\_artist 中的14个指标与影响者的14个指标之间的相似性，通过相似性数据说明影响者在哪些方面对追求者的影响较大。

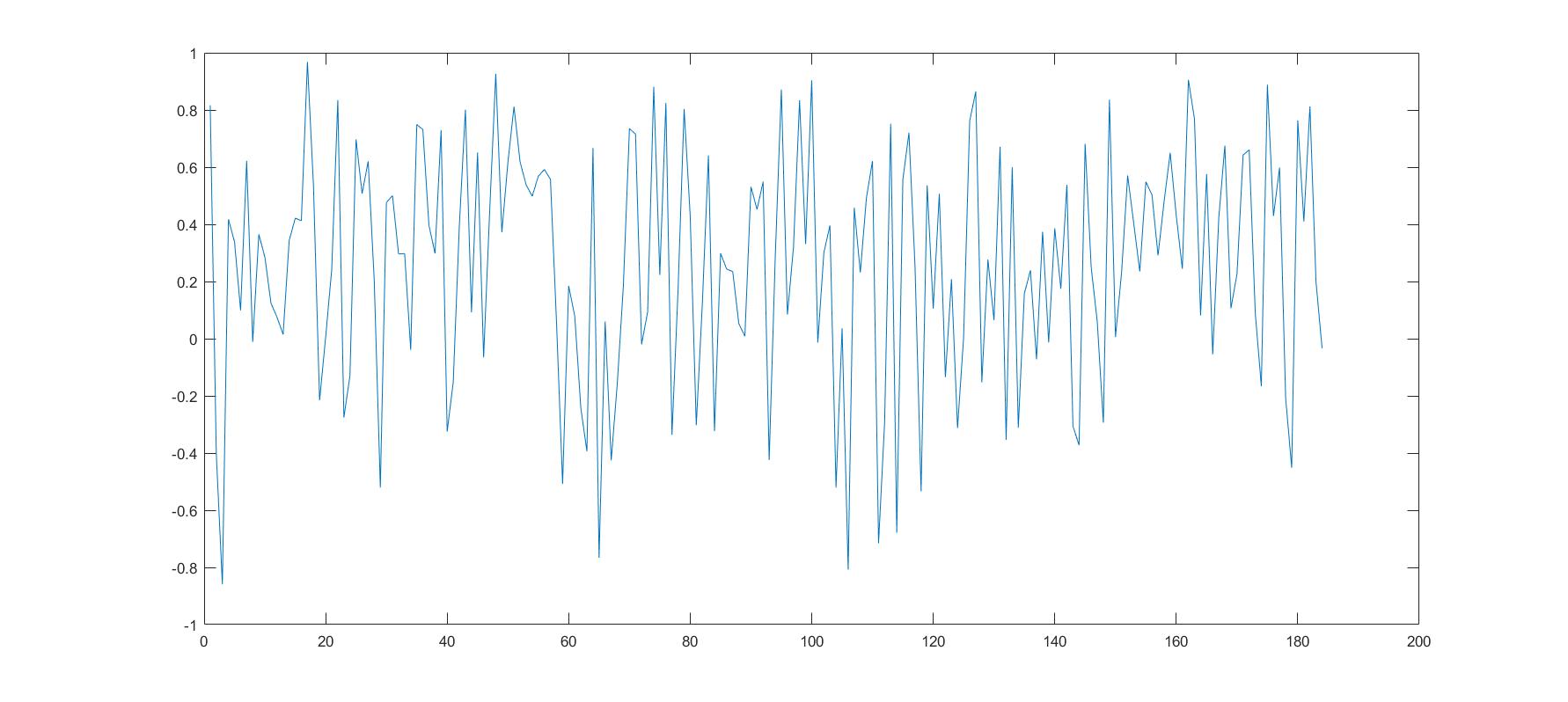
其次，我们要判断出音乐的哪些特征更具有感染力。其中感染力我们可以通过指标popularity来反映，popularity表示大众对音乐的喜爱程度。对full\_music\_data 中所有指标的数据与popularity指标数据利用Spearman秩相关进行相关性计算，选出与popularity指标相关性较大的指标。

最后，我们要反映出上述选出的指标对一个音乐家成为流行者是否有影响。我们通过问题一中做出的流派的网络，选择country genre 为主要研究对象。研究该流派内所有音乐家具有感染力指标，对选出的指标做灰色关联综合性分析，找出综合指标突出的人物。如果这些人物在问题一中影响力指标较大，则说明选出的指标在country genre音乐家成为一个著名的流行者中扮演着重要的角色。反之，则说明不重要。

5.4.1模型的建立与求解

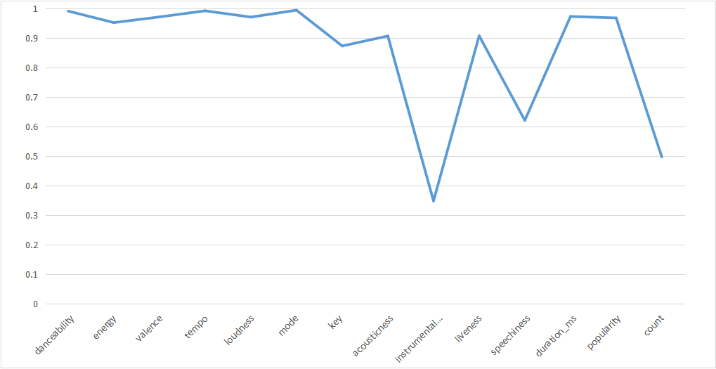
5.4.1.1 影响者对追求者影响关系模型

首先，我们分析影响者是否直接影响追求者，影响程度可以通过指标之间的相似性反映。我们选择country genre的代表人物Hank Williams进行分析，利用问题二中关于音乐家所作的主成分分析的数据，记Hank Williams的主成分数据为,记184个直接追求者的主成分数据矩阵为，计算出关于矩阵为中音乐家主成分数据之间的余弦相似度，公式如下所示：

影响者Hank Williams与其流派内所有音乐家之间的相似系数由下折线图所示：

通过该图片，我们可以明显看出Hank Williams和他的direct followers之间相似度波动较大，起伏明显，对于某些direct followers而言，Hank Williams对他们的综合影响较大，而对于另一部分人来说，综合影响力较小。

为了得到一个直接的结果，我们对每个指标都进行相似性分析，选取data\_by\_artist中的Hank Williams和他的direct followers的各项具体数据，分别计算出direct followers和Hank Williams之间14个指标的相似度。记Hank Williams的指标数据为,记184个直接追求者的指标数据矩阵为,其中各项指标的数据为的列向量,利用余弦相似性进行计算，结果如下图所示：



折线图是Hank Williams的各项音乐特征指标与其追随者指标的相似度。由图我们可以知道，相似度在0.9以上的指标有danceability，energy，valence，tempo，loudness，mode，acousticness，liveness，duration\_ms，popularity，相似度在0.9以下的指标有key，instrumentalness，speechiness，count。我们认为相似度在0.9以上的是相似度较高的指标，在这些指标中，Hank对追随者的影响较大。值得一提的是在instrumentalness, specchiness, count指标中，Hank Williams对追随者的影响较小，这也与Hank Williams所处的时代环境和个人观点有较大的关系。因为他是活跃在1930年代的美国country genre的早期歌手，时代与现在相差久远，大众的审美也发生了较大的变化，不同时代的人喜欢不同风格的artists，因此artists也会跟随时代选择他们所认可的instrumentalness，speechiness形式来搭配他们的表演，与Hank Williams没有较大的关系，所以导致了他在这方面的影响力较小。而在count指标中，他影响力小的原因则主要受他自身的观点所影响，与他人无关，他的followers并不会因为他所发的count数量的多少而改变他们自身的count数量。

5.4.1.2 Spearman秩相关——选出具有感染力的指标

感染力我们可以通过指标popularity来反映，popularity表示大众对音乐的喜爱程度。对full\_music\_data 中所有指标的数据与popularity指标数据利用Spearman秩相关进行相关性计算，相关系数如下所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| danceability | energy | valence | tempo | loudness | mode | key |
| 0.0754 | 0.1881 | -0.2231 | 0.2275 | 0.2182 | -0.1401 | 0.1195 |
| acousticness | instrumentalness | liveness | speechiness | duration\_ms | popularity | count |
| -0.3996 | -0.0811 | -0.0917 | -0.1524 | 0.4809 | 1 | -0.1366 |

选出与popularity指标相关性较大的指标：

根据文献资料[3]可得,当斯皮尔曼相关系数会在绝对值上增加时，两个变量的相似程度越高。因此我们选出感染力较高的指标为：

|  |  |  |
| --- | --- | --- |
| valence | tempo | loudness |
| acousticness | duration\_ms | popularity |

5.4.1.3 综合评价感染力指标模型

我们选择country genre 为主要研究对象。研究该流派内所有音乐家具有感染力指标，对选出的指标做灰色关联综合性分析，找出综合指标突出的人物。

***Step 1* 对感染力指标进行灰色关联分析**

利用问题一种关于灰色关联求解方法，得出各个指标的权重，得到综合指标：

权重如下表所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 指标 | valence | tempo | loudness | acousticness | duration\_ms | popularity |
| 权重 | 0.523797 | 0.523944 | 0.523783 | 0.523797 | 1 | 0.523838 |

选出综合指标较大的音乐人如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| artist\_name | 综合评分 | artist\_name | 综合评分 |
| Jeff Foxworthy | 406957.6 | The The | 322790.5 |
| James McMurtry | 355484.3 | The Marshall Tucker Band | 313573.3 |
| Grateful Dead | 337784.3 | Jeff Buckley | 307734.7 |
| Jerry Garcia | 335233.9 | Greg Brown | 283397.2 |
| The Waterboys | 329112.8 | Don Henley | 281026.7 |

***Step 2* 分析综合指标较大的音乐家的影响度指标**

上述综合指标较大的音乐家的直接追求者，影响时间范围，影响genre数量以及问题一中的影响度指标如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| name | followers | genre | con | 影响力 |
| Jeff Foxworthy | 0 | 0 | 0 | 0 |
| James McMurtry | 2 | 1 | 20 | 0 |
| Grateful Dead | 60 | 5 | 50 | 21.06033 |
| Jerry Garcia | 10 | 3 | 40 | 107.681 |
| The Waterboys | 11 | 4 | 20 | 48.72083 |
| The The | 7 | 1 | 20 | 32.19163 |
| The Marshall Tucker Band | 12 | 2 | 30 | 25.87655 |
| Jeff Buckley | 24 | 2 | 20 | 40.67005 |
| Greg Brown | 8 | 3 | 20 | 43.07241 |
| Don Henley | 12 | 2 | 20 | 28.4812 |

观察上表我们发现，表中音乐家的影响力很小，因此说明具有感染力的指标在音乐家成为一个著名的流行者中并不起什么影响。

第五题分析：

第五问要求我们从题目中给出的数据中找出音乐演进过程中是否有可能革命性的重大飞跃的特征。并找出哪些音乐家是重大变革的影响者。

首先我们要找出历史性的重大变化，因此我们考虑到音乐随时间的变化，利用题目中给出的data\_by\_year 的数据进行分析。由于该数据中指标类型较多，很多指标之间相似性较高，因此我们选择主成分分析法将指标降维，用较少的指标综合反映信息。用主成分分析法计算出指标的信息贡献率和累计贡献率，并对指标进行排序，选择贡献率在99.5%以上的指标为主成分，并计算出各个主成分的数值。

其次要找出从1921年到2021年音乐特点发生突变的年份进行分析。通过统计学的原理，用Mann-Kendall趋势检验法，时间进行的正相关序列与负相关相交的时间节点表示发生突变的年份。

最后对时间节点表示发生突变的年份附近进行分析，找出问题一中的关于突发年份的代表人物并进行分析。

5.5.1 模型的建立与求解

5.5.1.1 主成分分析

选用主成分分析法对13个指标变为少数几个综合性更强的指标，计算出各个指标的信息贡献率与累计贡献率，以主成分的贡献率为权重，建立主成分的综合评价模型。计算公式方法和问题二一致。

13个指标数据构成的序列为。

13个指标的信息贡献率和累计贡献率如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 指标 | 信息贡献率 | 累计贡献率 | 指标 | 信息贡献率 | 累计贡献率 |
| 1 | 54.97443276 | 54.97443276 | 8 | 0.827175288 | 98.53133182 |
| 2 | 13.93677281 | 68.91120557 | 9 | 0.692320971 | 99.22365279 |
| 3 | 11.75770216 | 80.66890773 | 10 | 0.601372029 | 99.82502482 |
| 4 | 7.27659598 | 87.94550371 | 11 | 0.142655703 | 99.96768053 |
| 5 | 5.271383484 | 93.2168872 | 12 | 0.032319474 | 100 |
| 6 | 2.587363089 | 95.80425028 | 13 | 3.6538E-16 | 100 |
| 7 | 1.89990625 | 97.70415654 |  |  |  |

根据以上表格选出4个主成分，使特征向量累计贡献率超过90%（五个主成分数据如下所示）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 指标 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 第一主成分 | 0.245818 | -0.3660 | 0.455273 | -0.05066 | -0.17391 | 0.030812 | -0.37842 |
| 第二主成分 | 0.381396 | -0.0324 | -0.06587 | -0.02532 | 0.095413 | -0.16167 | -0.05901 |
| 第三主成分 | 0.026553 | -0.6275 | -0.189 | 0.496019 | -0.17628 | 0.348094 | 0.101292 |
| 第四主成分 | 0.355149 | -0.1612 | -0.04336 | 0.247088 | 0.020982 | 0.017019 | -0.14276 |
| 指标 | 8 | 9 | 10 | 11 | 12 | 13 |  |
| 第一主成分 | 0.302304 | -0.01557 | -0.53498 | -0.20846 | 0.00665 | 2.65E-17 |  |
| 第二主成分 | -0.26103 | 0.211549 | 0.162426 | -0.41597 | 0.705324 | 3.59E-15 |  |
| 第三主成分 | -0.0182 | 0.338749 | 0.166514 | 0.135678 | 0.00233 | 7.56E-17 |  |
| 第四主成分 | -0.29999 | -0.80141 | 0.107191 | 0.140541 | -0.01798 | -4E-16 |  |

利用上述权重记为矩阵，将13个指标降维到4个主成分，利用下面的公式进行计算：

由于计算后数据量较大，在正文中不做展示。

5.5.1.2 Mann-Kendall趋势检验

音乐特征变化会出现突变，因此我们想要找出从1921年到2021年音乐特征发生突变的年份进行分析。通过统计学的原理，用Mann-Kendall趋势检验法，用时间进行的正相关序列与负相关相关作图，两条线相交的时间节点即为发生突变的年份。本问对主成分分析得出的四个主成分进行Mann-Kendall趋势检测法。

Mann-Kendall趋势检验法的步骤[4]如下：

* 对于影响夏季极端降水特征的4个主成分分别构造秩序列：

其中取值如下：

* 计算出序列的均值和方差

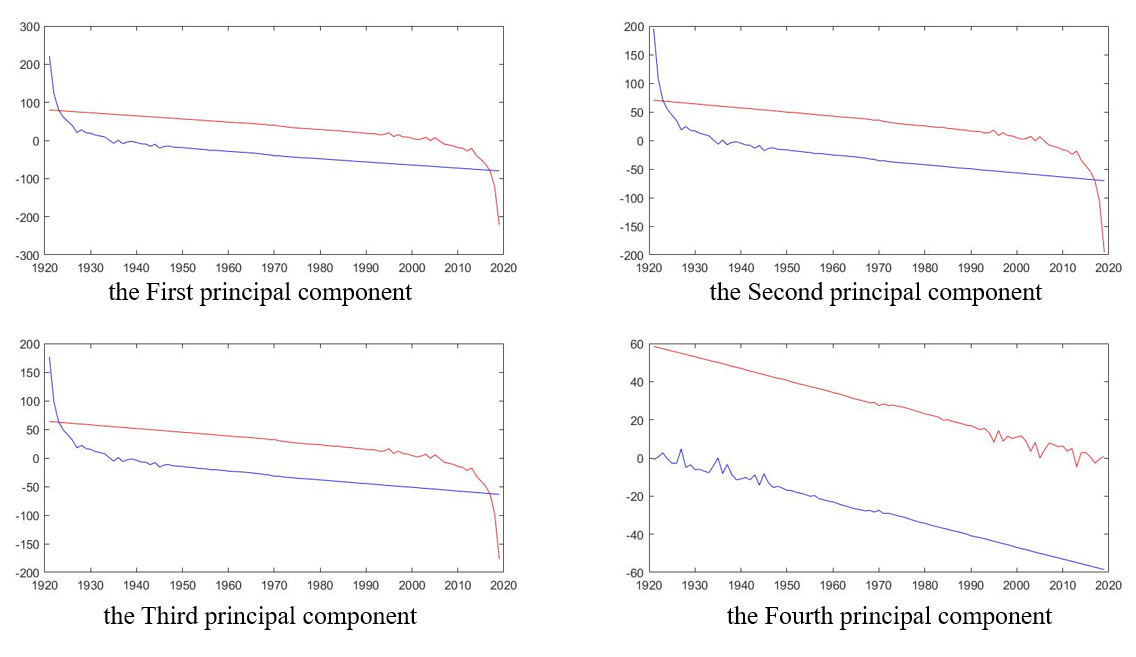
是独立的序列，均值和方差的计算公式如下:

* 计算

根据我们可以看出是大于时数值的所有的数值个数的累和。定义统计量为：

* 计算：

将计算出四个主成分的，根据年份做出图像如下：



得到前三主成分发生突变的年份为1923，2017。

5.5.1.3 重大变革影响者