练习五 — 问卷数据分析

某课题组设计并发布了 《大众点评内容营销对消费者行为影响问卷》

收到有效问卷 219 份，保存在“问卷”文件夹内，每个 txt 文件记录了一个问卷

问卷包含 6 道基础信息和 19 道李克特量表题

1、您的性别是:

2、您的年龄是:

3、您的学历是:

4、您的职业:

5、您每个月可支配的生活费为:

6、您是否有用过大众点评APP购买产品/服务的经历:

```
所用时间:35秒
来源:微信
来源详情:梨涡喵??[未知]
来自IP:220.152.161.130(广东-深圳)
1、您的性别是:女
2、您的年龄是:25
3、您的学历是:本科
4、您的职业:全日制学生
5、您每个月可支配的生活费为:1001-2000元
6、您是否有用过大众点评APP购买产品/服务的经历:是
7、内容营销—真实性内容–该平台发布的内容信息是真实的:4
7、该平台发布的内容信息是可靠的:4
7、该平台发布的内容对于产品/服务展示是详细的:4
8、内容营销—娱乐性内容–该平台发布的内容是轻松好玩有趣的:4
8、该平台内发布的内容是让我觉得激动和兴奋的:4
8、该平台内发布的的内容能让我充满想象与好奇:4
8、该平台内发布的内容能让我有沉迷其中的感受:4
9、内容营销—社交互动内容–通过该平台，能够让我遇见与我相似的人:4
9、通过该平台，我能与那些与我相似的人交流互动:4
9、通过该平台，我认识了有意思的人:4
10、内容营销—情感性内容–该平台里发布的内容能让我感同身受:3
10、该平台里发布的内容能让我产生情感共鸣:3
10、该平台里发布的内容接地气让我没有距离感:3
11、感知信任–该平台的商家评分等级，发布者评价的产品/服务值得信任:4
11、该平台发布者/商家的发布内容，让我对产品/服务的质量更加具有可信度:4
11、该平台发布者/商家的推荐的产品/服务对我是有用的:4
12、消费者行为–我会考虑购买该发布者/商家的产品和服务:4
12、该平台发布者/商家影响较大，我购买其推荐的该产品/服务的可能性很大:4
12、该平台发布者/商家发布的内容会对我的消费行为提供一些信息:4
总分:73
```

题目一：数据读取：使用 Python 读取并存储所有问卷的数据

求：可支配生活费在3000元以上的人群中，女性所占的比例

提示：

（1）os.listdir(dir_path)

此函数可以把给定路径下所有文件名保存到列表中，需要 import os

（2）data = pandas.DataFrame( )

这是一种二维表格数据结构，import pandas

使用指南：https://pandas.pydata.org/docs/user_guide/index.html

（3）注意异常值和缺失值

示例：

```
In [8]: data.head()
Out[8]:
```

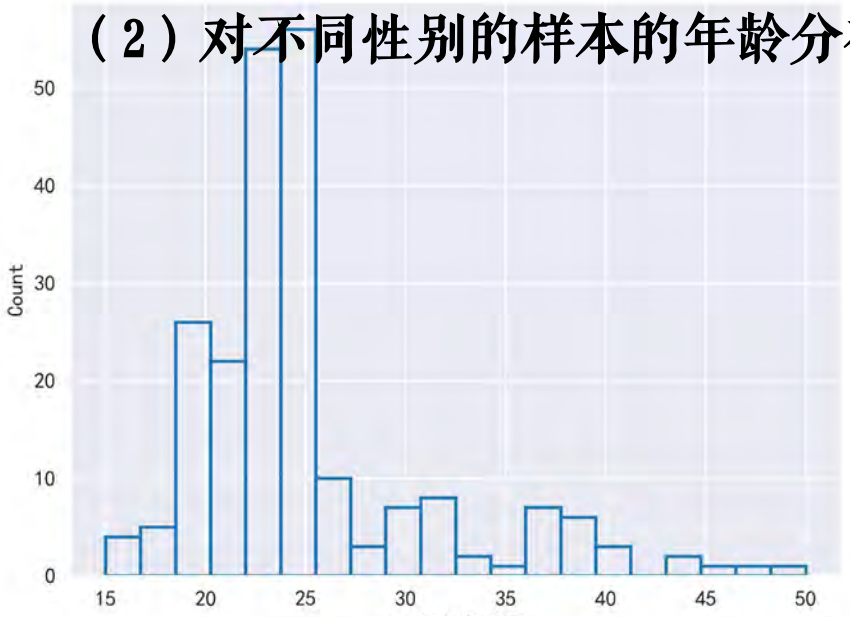
	所用时间	来源	来源详情	来自IP	1、您的性别是	2、您的年龄是	3、您的学历是	4、您的职业	5、您每个月可支配的生活费为	6、您是否有用过大众点评APP购买产品/服务的经历	...	10、内容营销——情感性内容—该平台里发布的内容能让我感同身受	10、该平台里发布的内容能让我产生情感共鸣	10、该平台里发布的内容接地气让我没有距离感	11、感知信任—该平台的商家评分等级，发布者评价的产品/服务值得信任	11、该平台发布者/商家的发布内容，让我对产品/服务的质量更加具有可信度	11、该平台发布者/商家的推荐的产品/服务对我是有用的	12、消费者行为—我会考虑购买该发布者/商家的产品和服务	12、该平台发布者/商家影响较大，我购买其推荐的该产品/服务的可能性很大	12、该平台发布者/商家发布的内容会对我的消费行为提供一些信息	总分
问卷1.txt	35秒	微信	梨涡喵??[未知]	220.152.161.130(广东-深圳)	女	25	本科	全日制学生	1001-2000元	是	...	3	3	3	4	4	4	4	4	4	73
问卷102.txt	20秒	互填问卷	nan	117.136.34.27(广东-广州)	女	20	本科	企业职员	2001-3000元	是	...	5	4	4	4	5	5	4	5	4	83
问卷103.txt	23秒	互填问卷	nan	223.83.93.232(江西-赣州)	男	22	本科	企业职员	3000元以上	是	...	3	4	3	3	4	4	5	4	3	71
问卷105.txt	42秒	互填问卷	nan	27.201.217.250(山东-泰安)	女	38	本科	企业职员	3000元以上	是	...	4	5	4	5	4	5	4	5	4	87
问卷113.txt	20秒	互填问卷	nan	120.43.121.61(福建-漳州)	女	40	本科	企业职员	3000元以上	是	...	3	4	4	3	3	4	3	4	4	69

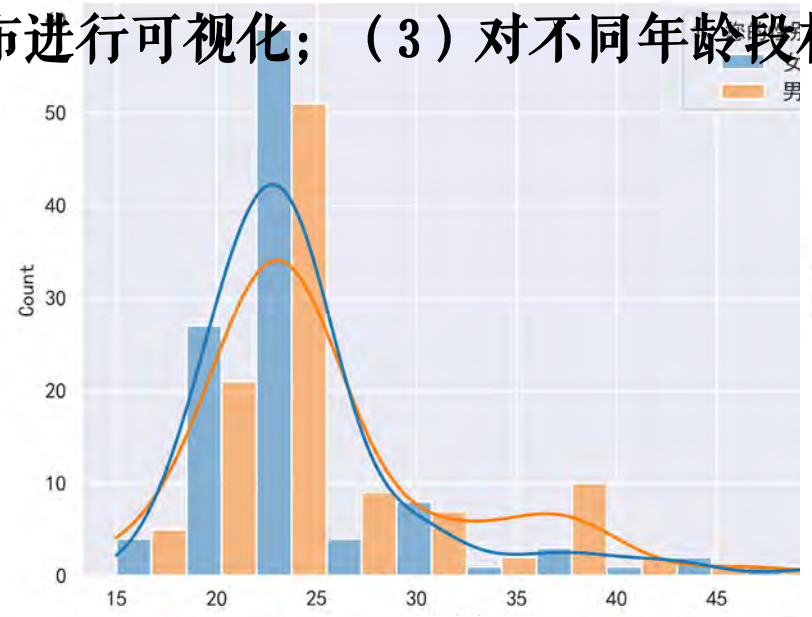
5 rows × 30 columns

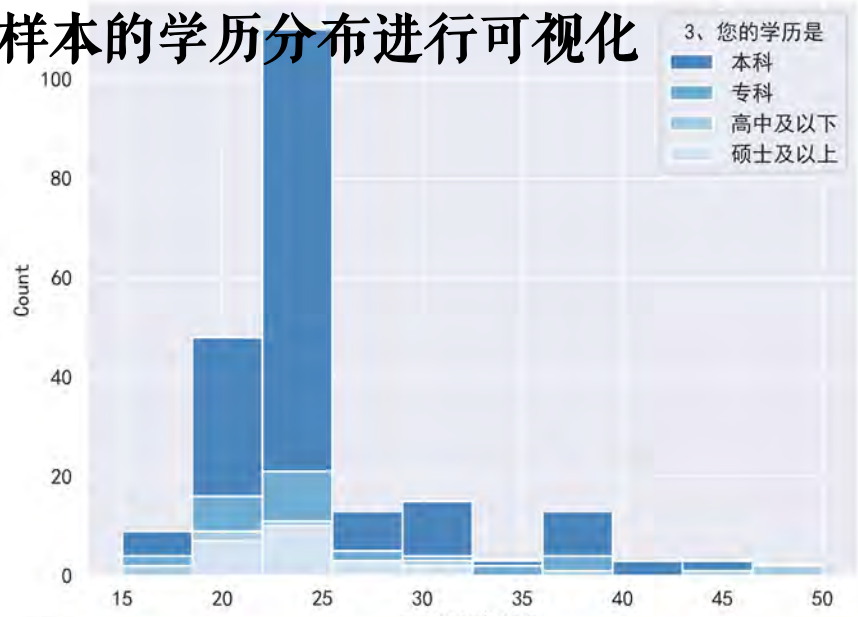
题目二：可视化（1）对样本的年龄分布进行可视化；

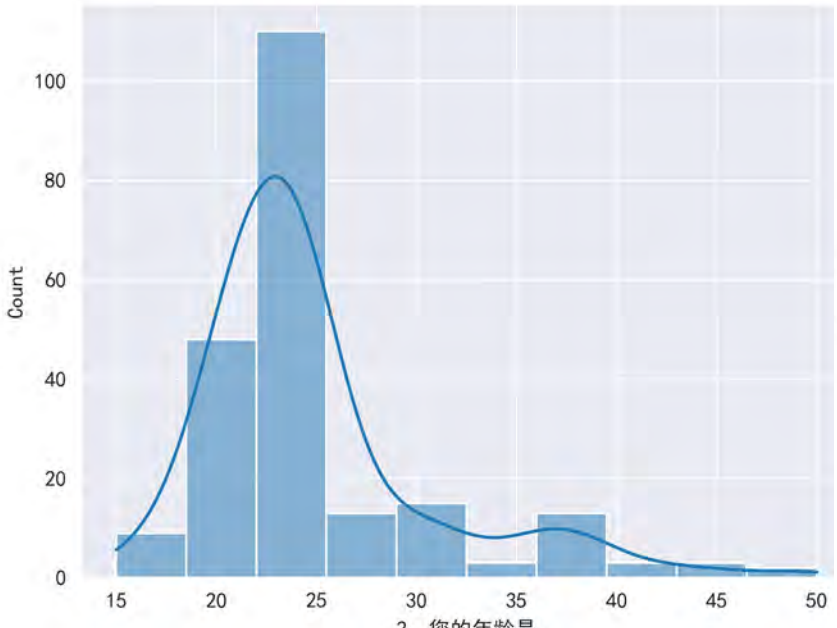
建议使用matplotlib、seaborn或 pyecharts进行绘图

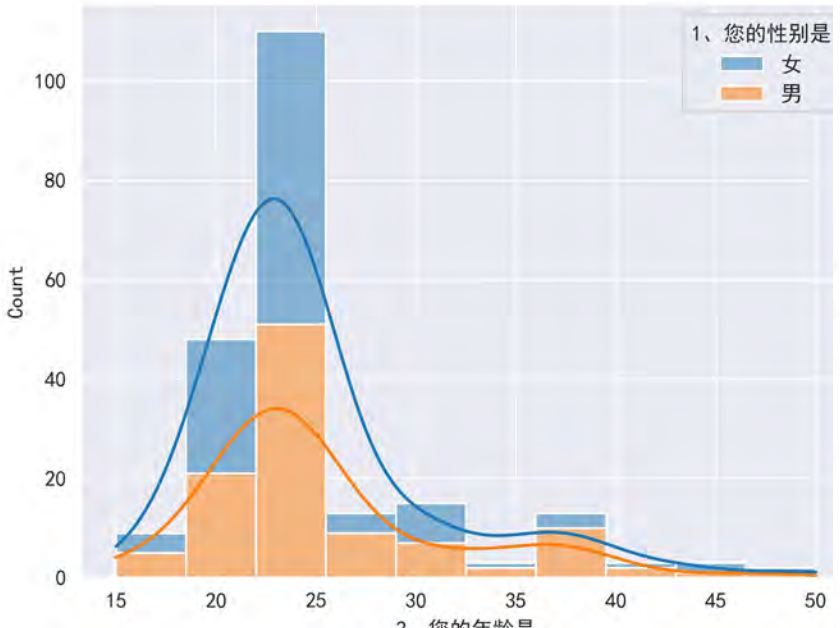
（2）对不同性别的样本的年龄分布进行可视化；（3）对不同年龄段样本的学历分布进行可视化

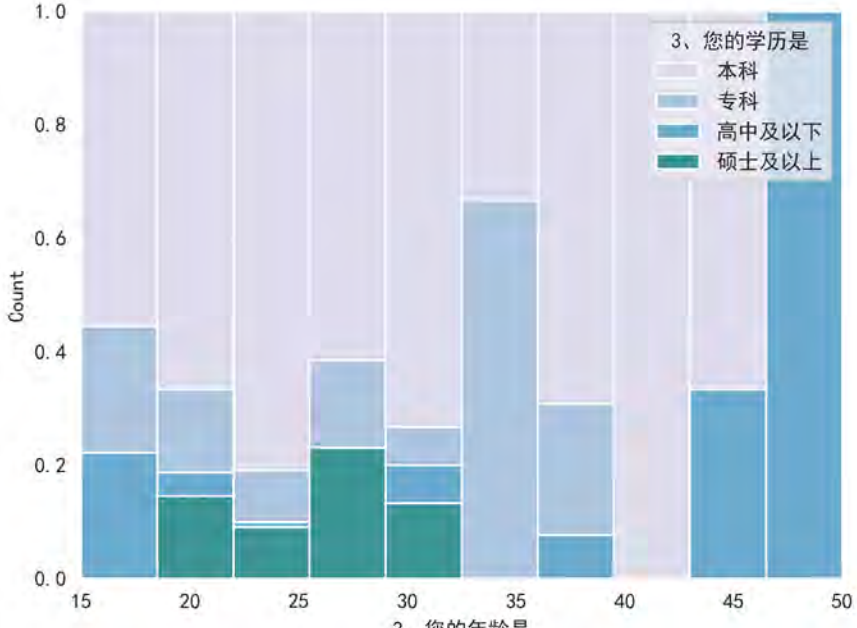






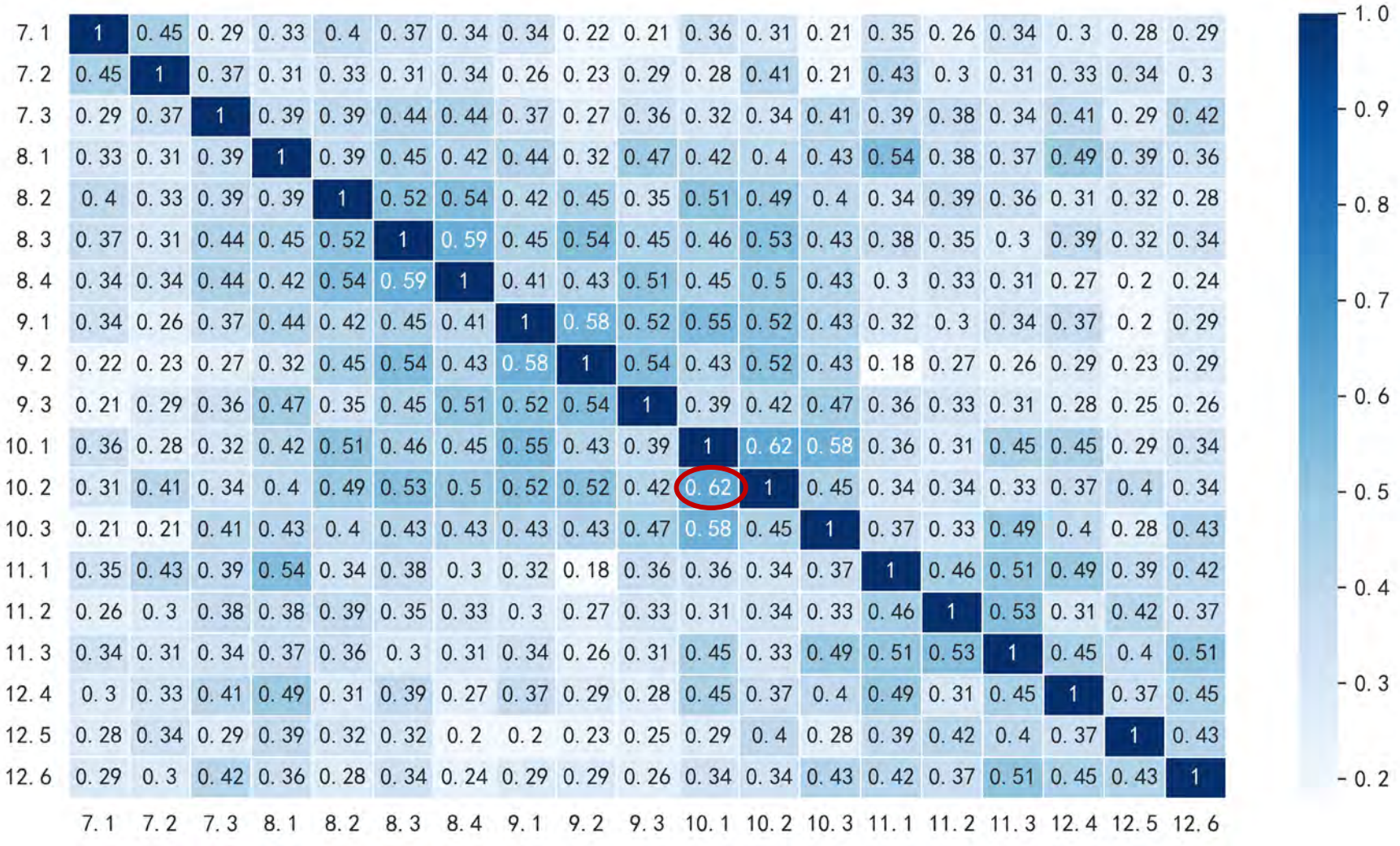






题目三：相关性分析：对 19 道李克特量表题结果进行相关性分析

进行可视化并输出相关性最高的两道题



提示：

可以使用一些库函数来快速的求变量间的相关系数，如：

numpy. corrcoef() ;

pandas.corr() ;

scipy.pearsonr()

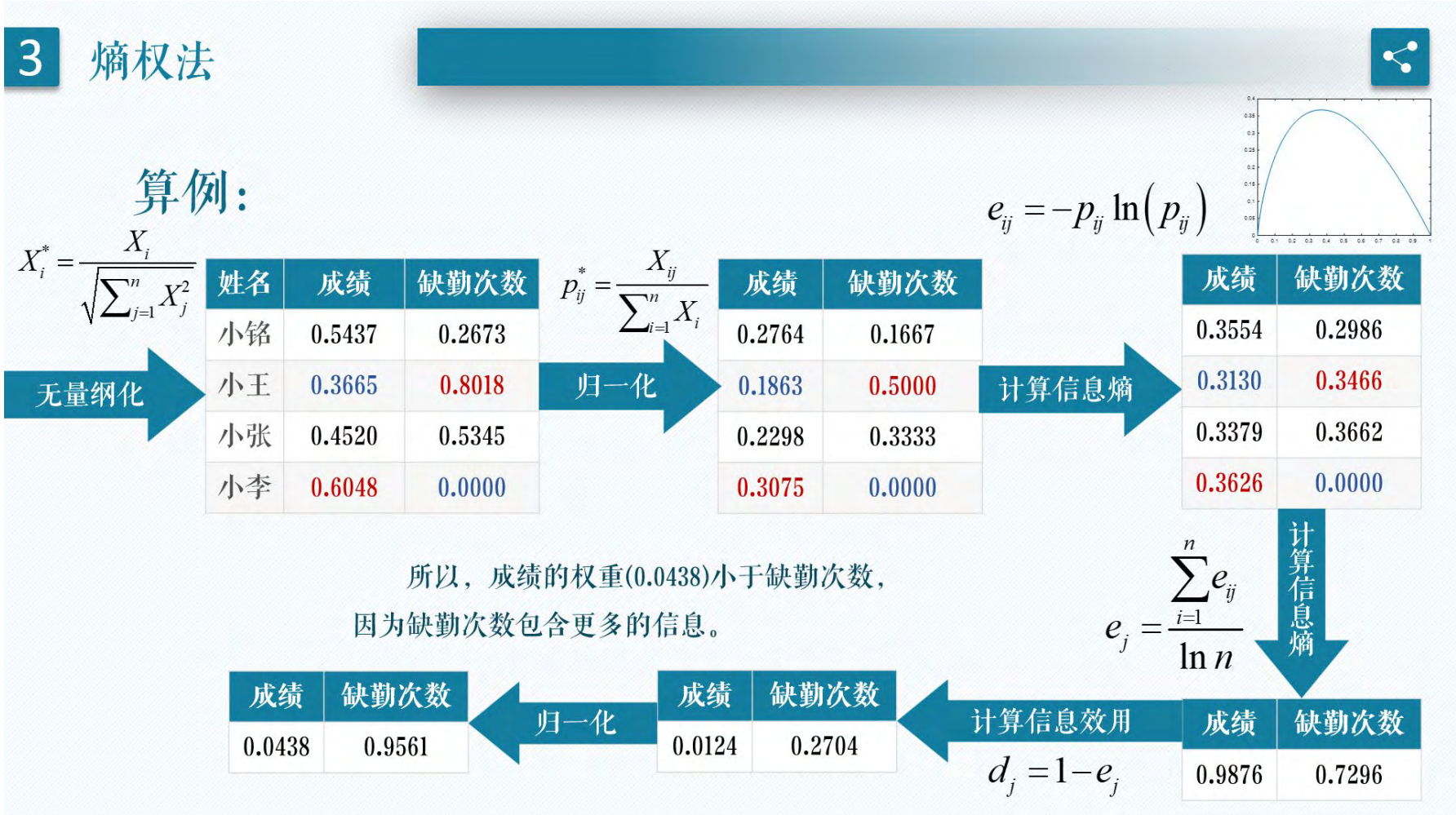
可以使用不同的相关系数，例如pearson、Kendall、spearman等，但要清楚不同系数的使用规则

题目四：评价： 19 道李克特量表代表了 19 项评价指标，请使用线性加权的方式对每个样本进行打分，并输出得分最高和最低的样本（使用熵权法对指标赋权）

提示：

熵权法是一种常用的多指标决策方法，可以用来确定各个指标的权重。

具体计算过程 →



姓名	成绩	缺勤次数
小铭	0.5437	0.2673
小王	0.3665	0.8018
小张	0.4520	0.5345
小李	0.6048	0.0000

成绩	缺勤次数
0.2764	0.1667
0.1863	0.5000
0.2298	0.3333
0.3075	0.0000

成绩	缺勤次数
0.3554	0.2986
0.3130	0.3466
0.3379	0.3662
0.3626	0.0000

成绩	缺勤次数
0.9876	0.7296

成绩	缺勤次数
0.0124	0.2704

成绩	缺勤次数
0.0438	0.9561

题目五：分类：有专家根据问卷结果对大众点评用户群体划分了三个类别：A 接受型，B 尝试型，C 厌恶型，目前已知前 200 个样本的类别类型，存储在 Type.csv 文件中，请训练合适的模型，判断其余 19 个样本的类型。根据你的个人观点填写问卷，并使用分类器判断自己属于什么用户类型。

提示：

（1）分类算法是一种基于一个或多个自变量确定因变量所属类别的机器学习技术。

（2）建议使用 sklearn 库中的函数进行分类，分类方法不限（KNN、朴素贝叶斯、SVM等）

了解如何评价分类的准确程度。