INT104 Check List:

在本次作业中，我们需要处理一份包含学生信息的电子表格数据，进行数据分析。该数据包括学生的索引、性别、所在的课程（Programme）、年级、总分数以及五个考试题目的分数。课程项目的目的是提取数据特征，并分析这些特征与学生所在课程之间的关系。

下面是我们必须要达到的四个任务：

1. 观察原始数据分布：使用箱型图查看数据分布，并讨论如何通过缩放原始数据来减少数据规模的影响。

2. 执行主成分分析（PCA）：对数据执行 PCA，观察各主成分的分布情况，找出可以更容易地分类课程的主成分集合。

3. 提取特征：独立地提取能够方便分类学生所属课程的特征。

4. 可视化比较：对原始特征、缩放后的特征、PCA 特征及你提取的特征进行可视化和比较。

要求使用 Python 完成这些任务，并在实验室会议期间进行现场演示，向教学助理展示你的工作。完成后，需要撰写一份实验报告，总结你的实验设计、结果和分析。

当然，这里是对课程作业评分标准的详细翻译：

实验报告编辑与语言问题（共 10 分）

- 10 分：没有格式问题。

- 8 分：轻微的语言问题或轻微的格式问题。

- 6 分：报告基本上符合要求，有一些语言和格式问题。

- 4 分：报告勉强可以阅读。

- 2 分：报告难以阅读，但可以理解。

- 0 分：报告无法理解。

任务 1、2、3 和 4（总共 60 分，每个任务 15 分）

- 15 分：通过结果展示出科学假设。

- 12 分：比较并分析了不同实验配置的结果。

- 9 分：深入比较了不同实验配置的结果。

- 6 分：完全达到了任务目标。

- 3 分：部分达到了任务目标，做出了良好尝试。

- 0 分：未达到任务目标，没有合理的尝试。

现场演示：

回答问题（共 15 分，每个问题 5 分）

- 5 分：完全理解概念并提供了满意的答案。

- 4 分：提供了满意的答案。

- 3 分：提供了满意的答案，但有轻微的误解。

- 2 分：答案勉强可以接受。

- 1 分：答案不正确。

- 0 分：学生无法回答问题。

代码运行（15 分）

- 15 分：代码执行效率高，并能预测结果，对算法有很好的理解。

- 12 分：按要求实现代码，并深入讨论了结果。

- 9 分：可能需要协助来实现代码，并显示对结果的理解。

- 6 分：在一定时间内可能需要协助来实现代码，并表现出一些对结果的理解。

- 3 分：无法实现所需更改，并对结果有合理的预期。

- 0 分：不理解所需更改的意图。

奖励分数（总分不超过 100 分，无单项上限）

- +10 分：在任何任务中展示了新颖的科学假设。

- +5 分 ： 以便他人轻松重现实验的方式呈现实验。

- +5 分：报告的格式可发布。

扣分：

- -10 分：引用不当。

- -20 分：严重的引用不当（多次引用不当或复制整段文字）。

- 根据学校的学术诚信政策，可能会有其他惩罚。

以上是详细的评分标准，确保你在准备报告和现场演示时注意这些细节。

上面是题目的要求，下面简单的要点：

简单的流程如下：

Task1:数据观察->数据清洗->数据正则化

Task2:使用 pca 降维，并分析分布特征

Task3:进行自己的特征分析，根据上文所获的信息，来进行特征选取

卷分细节：

在做数据观察的时候可以把每一列的特征都分析一下，说说 index 是干什么的，每个数据代表着什么意思。(这个部分不是重点，简单提一嘴就好了。

数据清洗就比较需要技巧了，这里面根张洪斌一样对数据进行了降噪处理。就是将数据集当中 programme 当中不同但是 feature 完全相同的点。然后对数据进行了 Shapiro-Wilk normal distribution test 来检测每个特征的分布是否符合正态分布。然后画出箱图，看数据当中有哪些点有异常值，对有异常值的列（箱外有点）进行 IQR 数据处理。就进行了数据清理的全过程。

接下来就用数据降维：这里面因为所有的特征都不是正态分布，所以理论上应该用 min-max 来进行正则化，但是这里面，效果并不好，我就不用这个了（你们感兴趣可以把实验做完）。我就用 z-score 进行了一个简单处理。

接下来就把数据放到 PCA 里面跑，得到的图中可以看见，programme 3 被分出来了，简单分析一下即可。

然后画出热力图，用 spearman 相关系数。简单分析一下，犯下 MCQ 和 Total 相关度

太高了，这里可以去掉 MCQ，然后降维，所以你的 resulting feature 就是“Gender Grade Total Q1 Q2 Q3 Q4 Q5”,在对比一下就可以了，为以后的模型简化了计算。