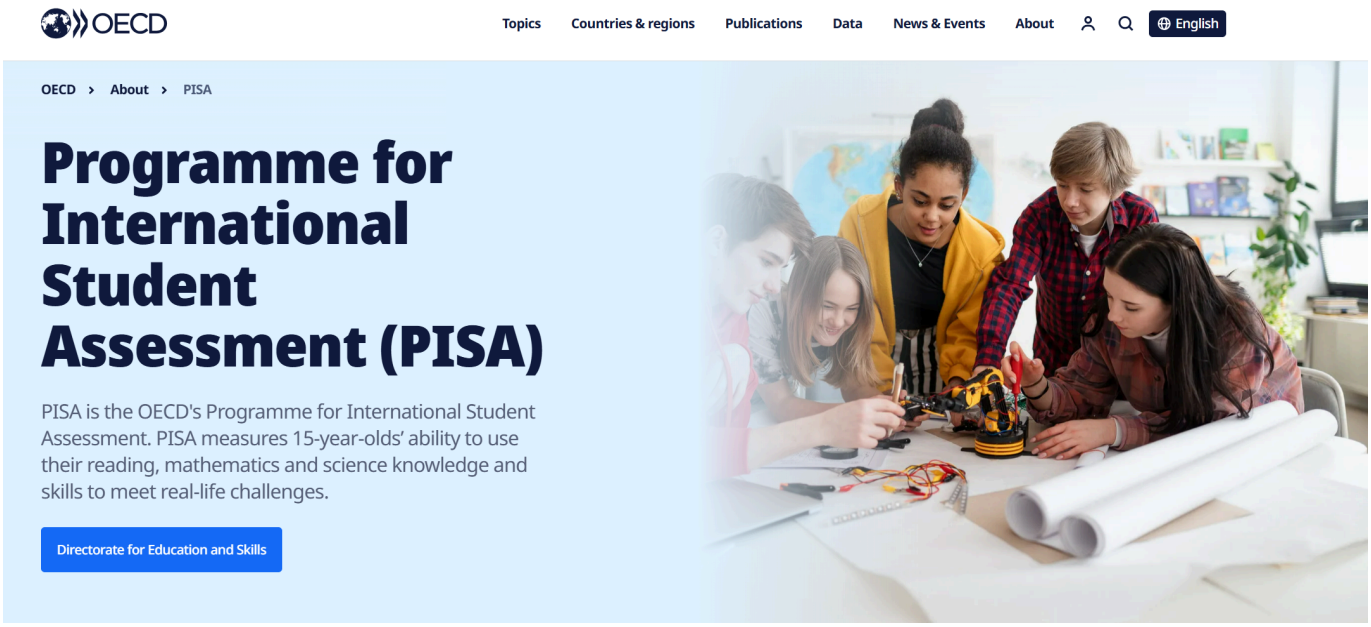
实验三

中国科大2025年春季学期"数据分析及实践"课程 - 实验三说明文档。

任务概述

PISA（Programme for International Student Assessment）是由经济合作与发展组织发起的一项全球性的教育评估项目，旨在评估15岁学生运用阅读、数学和科学领域的核心素养，并通过大规模的跨国数据收集和分析，为各国教育政策的制定提供参考依据。知识和技能应对现实挑战的能力。

本实验数据集取自PISA 2018 School Questionaire，它包含对学校进行的调查结果。现要求基于此数据集，进行探索性分析、可视化分析和特征关系分析，请按要求完成实验任务。



任务列表

1.（30%）使用`pandas`库的相关方法，进行数据集读取、信息处理和探索性分析。

- Q1.（4%）读取数据集`data.csv`（将首列作为索引）至变量`df`，展示该数据集的前10行内容，并展示数据集有多少行和多少列。
- Q2.（4%）数据集存在很多缺失值，输出各列缺失值的个数，并删除数据集的最后一列。基于更新后的数据，展示哪一列的缺失值最多，哪些列没有缺失值。
- Q3.（4%）数据集的有些列在所有记录上均有相同取值，作为独立的一列是相对冗余的。请查找并输出这些列的名称和取值，并阐述这些列代表的含义，最后删除这些列。
- Q4.（4%）观察`PRIVATESCH`特征列，统计所有取值及其出现的次数。其中有一部分取值含义一致但形式不同（如`private`和`PRIVATE`），试对它们进行归并，随后展示所有取值及其出现的次数。
- Q5.（5%）选取特征`STUBEHA,TEACHBEHA,EDUSHORT,STAFFSHORT`，展示它们的基本统计特征信息（平均值、标准差、四分位点、最小值、最大值、Pearson相关系数矩阵）。
- Q6.（4%）Q5中所得Pearson相关系数矩阵显示，特征`STUBEHA`与`TEACHBEHA`之间、`EDUSHORT`与`STAFFSHORT`的相关系数较高，请通过特征定义推测可能导致相关性的原因。
- Q7.（5%）执行以下子表提取操作：

```
df1 = df[['PRIVATESCH', 'EDUSHORT', 'STAFFSHORT']]
```

并基于df1，以特征PRIVATESCH为先验条件，对其余各特征中可能存在的缺失值进行均值填补。

2.（17%）导入numpy和matplotlib库，对数据集df进行一定数据可视化分析。

- Q1.（5%）选择两个连续数值型特征，绘制其分布散点图，要求合理设置散点颜色和大小，并配上合适的标题和图例注记。
- Q2.（5%）选择一个离散数值型特征（建议所有取值数量不超过10），绘制饼图，要求设置合理配色和比例，并配上合适的标题和图例注记。
- Q3.（7%）对T1-Q5中的Pearson相关系数矩阵，绘制热力矩阵图，要求为每个位置增添对应数值表示（保留三位小数），设置数值与颜色的对应关系条，并配上合适的标题和坐标表示。

3.（23%）现欲对数据集特征STUBEHA,TEACHBEHA进行分布校验。执行以下子表提取和缺失记录删除操作：

```
df2 = df[['STUBEHA', 'TEACHBEHA']].dropna()
```

并基于df2完成以下任务：

- Q1.（6%）以区间数为10，分别绘制两个特征的频数直方图，基于频数直方图的结果，是否可以认为两特征近似服从正态分布？
- Q2.（8%）Q-Q图（Quantile-Quantial Plot）又称为分位图，可简单直观地判断一组数据是否服从某种理论分布。例如，若一容量为n的样本X服从标准正态分布，则其各分位点应与标准正态分布对应分位点完全一致。若从标准正态分布抽样得到容量为n的样本Y，则散点$(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)$将近似分布于直线$y=x$上，其中$x_1\leq x_2\leq\dots\leq x_n$，$y_1\leq y_2\leq\dots\leq y_n$。单个特征样本Q-Q图可通过以下代码绘制：

  ```
  import statsmodels.api as sm
  import matplotlib.pyplot as plt

  sm.qqplot(data, line='45')
  plt.show()
  ```

  参考上述代码，分别绘制这两个特征样本的Q-Q图，基于Q-Q图的结果，是否可以认为两特征近似服从正态分布？将该结论与Q1所得结论进行对比，你有什么感想？
- Q3.（9%）特征STUBEHA与TEACHBEHA的理论分布是否具有一致性？请自行编写代码绘制两特征样本的Q-Q图和直线$y=x$，并基于可视化结果简述你的发现。

4.（13%）基于正态分布假设，对特征STUBEHA,TEACHBEHA的总体分布进行参数估计。

- Q1.（8%）假设特征STUBEHA和TEACHBEHA样本分别独立同分布于正态分布$N(\mu_1,\sigma_1^2),N(\mu_2,\sigma_2^2)$，请分别求均值参数$\mu_1,\mu_2$和方差参数$\sigma_1^2,\sigma_2^2$的极大似然估计。作为总体均值和总体方差的估计，样本均值和样本方差均具有无偏性，请问上述极大似然估计是否也同样具有无偏性？
- Q2.（5%）设特征STUBEHA的样本为$Y=(y_1,y_2,\dots,y_n)$，对该特征进行常数估计$y=a$，求参数a的最小二乘解，并比较其与Q1中所得总体均值极大似然估计的结果。

5.（17%）基于T4的假设，现需对特征STUBEHA,TEACHBEHA的总体均值差异进行检验。请阅读本文档的内容，并导入scipy库，完成以下任务：

- Q1.（5%）简述本情景下应使用成组检验还是成对检验，并写出单侧检验原假设。
- Q2.（4%）使用`scipy.stats`中的相关方法，执行相应的假设检验。
- Q3.（4%）基于Q2所得结果，请仔细斟酌并叙述你所得到的结论。
- Q4.（4%）上述结论隐含了犯哪一类错误的可能？相应犯错概率是多少？

格式和提交要求

1. 请按具体任务分步编写代码，存储于`.ipynb`格式文件中用于复现，必要时可增加注释。
2. 实验报告必须涵盖任务列表中的所有内容和相应结果，并请存储于`.pdf`格式文件中。
3. 提交时，请将实验源代码和实验报告保存至一个压缩包中，命名为"学号-姓名-实验三.zip"，并于2025年4月24日之前发送至USTC_AD2025@163.com。

参考资料

以下资料可能会对你顺利完成实验有所帮助。

1. Kaggle 数据挖掘与预测竞赛平台网站：点击这里
2. 数据分析库 pandas 官方网站：点击这里
3. 科学计算库 numpy 官方网站：点击这里
4. 数据可视化库 matplotlib 官方网站：点击这里
5. 科学计算库 scipy 官方网站：点击这里