蓝桥杯全国大学生软件和信息技术大赛组委会

第十八届蓝桥杯全国大学生软件和信息技术大赛
人工智能赛（数据分析与可视化科目）
竞赛规则及说明

一、参赛对象

具有正式全日制学籍并且符合相关科目报名要求的研究生、本科及高职高专学生（以报名时状态为准）。

每位选手配备一名指导教师，同一名指导教师可指导多位选手。全国选拔赛和全国总决赛比赛后指导教师原则上不能更改。

二、组别设置

该科目设置大学组。

三、竞赛赛制

本届大赛采用校赛、全国选拔赛、全国总决赛三级竞赛体系。

校赛：由参赛院校自行组织并进行评审。

全国选拔赛：由大赛组委会统一组织。比赛时长为 4 小时。

全国总决赛：由大赛组委会统一组织。比赛时长为 4 小时。

详细赛程安排以组委会公布信息为准。

四、竞赛形式

1. 全国选拔赛、全国总决赛均采用封闭、限时方式举办。选手以个人为单位，独立进行作答。

2. 采取线上比赛的形式。选手需访问互联网，通过大赛组委会指定的比赛系统进行答题。

3. 选手将答案提交到比赛系统中，超过比赛时间将无法提交。

五、参赛选手机器环境

1. 选手机器配置：

CPU：x86_64 双核≥2.0 GHz（如 Intel Core i3-3xxx/AMD 同级）

内存（RAM）：8 GB 及以上

存储空间：可用空间≥10 GB（系统盘或安装盘）

操作系统：Windows8、Windows10 或 Windows11

显示分辨率：1920 × 1080 及以上

2.编程环境：

序号	软件名称及版本号
1	Visual Studio Code 软件版本： v1.36 扩展：Chinese、Python、Jupyter
2	Anaconda Anaconda3-2023.03
3	Python 3.10.10
4	Notebook 7.2
5	Numpy 2.0.1
6	Torch 2.3.1
7	Scikit-learn 1.5.1
8	Pandas 2.2.2
9	Matplotlib 3.9.0
10	Scipy 1.14.0
11	Opencv-python 4.9.0.80
12	Transformers 4.42.3
13	Jieba 0.42.1
14	Gensim 4.3.2
15	Pillow 10.3.0
16	BeautifulSoup4 4.12.3
17	XGBoost 2.0.3
18	Spacy 3.7.4
19	Plotly 5.22.0
20	Onnx 1.16.2
21	Onnxruntime 1.18.0
22	Flask 3.0.3

23	Seaborn 0.13.2
24	Statsmodels 0.14.2
25	Bokeh 3.4.1
26	Altair 5.3.0
27	Polars 1.1.0
28	Dask 2024.10.0
29	Xarray 2024.10.0
30	WPS
31	Chrome 浏览器 v100 以上版本

六、赛题形式

赛题均为场景实战题（数据分析与可视化项目实操），具体题型及题目数量以正式比赛时赛题为准。选手需依据任务说明，使用比赛环境预置的软件工具完成预期目标，按题目需求提供最终答案。

比赛评分除特别说明外，均以结果为评分依据，选手需根据题面要求，合理选择分析工具，完成题目，最终根据选手提交的答案文件，机器自动评分。

所有题目均提供完整的题面 PDF 文档及必要的基础资源包（如数据集、基础代码/模板）。题面文档详细说明背景、需求与目标。选手需认真阅读题意，结合所给资源，通过撰写代码与文档，达成题目规定的最终结果。

部分题目可能包含前序准备步骤，例如解压相应数据文件、在试题文档中预览可视化结果等。一般情况下，默认选手已掌握数据清洗、可视化绘图等基础方法与工具使用，题面不再单独提供软件/库函数的使用指导。

特别说明：在无明确说明的情况下，不得随意更改基础资源中的文件名称、文件夹名称及其存放结构；请严格按题意规范操作，否则将影响最终评分。

七、赛题考察范围

本次比赛主要考查数据清洗与特征构造、数据加工、分析建模与评估、数据预测、数据可视化设计等数据分析与可视化技术应用相关领域职业能力，要求选手依据题目背景及要求，进行业务需求分析，在指定比赛环境内完成数据预处理、数据加工、数据分析与可视化、数据挖掘等工作任务。

考查知识范围详见竞赛大纲。链接：dasai.lanqiao.cn/notices/846。

八、答案提交

选手只有在比赛时间内提交的答案内容是可以用来评测的，比赛之后的任何提交均无效。选手应使用大赛组委会指定的比赛系统来提交答案，任何其他方式的提交（如邮件、U 盘）都不做为评测依据。

选手在比赛系统可重复提交自己的答案，以最后一次提交的答案为准并作为评测的依据。

比赛过程中，赛题分数不会显示给选手，选手应当在没有反馈的情况下自行设计数据调试自己的程序。

选手须仔细阅读并严格遵守赛题指定的答案文件格式或内容。

九、学习资源及样题

1.历届真题

参赛选手可在蓝桥杯官方网站-关于大赛-历届真题里查看往届蓝桥杯大赛的真题。链接：https://www.lanqiao.cn/courses/67522。

2.赛项资源

免费开放供参赛选手备赛使用，包含模拟题、真题解析、课程等资料。链接：https://www.lanqiao.cn/paths/3361/。

3.样题

样题详见文档附录。

十、评分

全部题目将使用机器自动评分。题型及评分标准如下：

1.试题分值及题型分布：

模块	分值
模块一：数据预处理	15 分
模块二：数据分析	30 分
模块三：数据可视化	30 分
模块四：数据挖掘	25 分

2.题型设计及评分方法：

试题按照以上四个模块考点进行出题，每个模块考察题目类型不固定，根据考点特性出

题，详细题型可参考样题。

（1）**计算类：**该类型试题有标准答案，机器自动匹配选手提交答案与标准答案是否相同，若一致即可得分。例如：题面要求学生计算总费用，并将结果填写到指定的文档中，评分标准即为机器自动读取指定文档中的数字，若与标准答案一致，则选手获得本题分数。

（2）**绘图类：**机器自动针对题面提到的图表类型、图表元素等进行评分，题面未提到的内容，均不在评分范围内，选手可自主选择比赛环境中提供的软件工具完成绘图要求。例如：题目要求选手绘制折线图，评分标准即为判定选手提交的文件中，是否包含折线图文本，同时判定折线图的横纵坐标是否与标准答案一致，若一致则选手获得本题分数。

（3）**编程类：**机器自动运行选手提交的代码文件，并通过测试样例验证，若与测试结果一致，获得该题分数。例如：题面要求学生完成总成绩计算，并提交相关 Python 代码，评分标准即为机器自动运行选手提供的代码，对指定文件进行操作，若得到的结果与标准答案一致，则选手获得本题分数。

（4）**操作类：**机器自动匹配选手提交的答案文件中关键点位是否操作正确，若所有点位均正确，则获得该题分数。例如：题面要求删除空白数据，评分标准即为判定提交的文件是否存在空行，若无则选手获得本题分数。

十一、奖项设置及评选办法

1.全国选拔赛

全国选拔赛设立一、二、三等奖，原则上各奖项的获奖比例为 10%、15%、25%，总获奖比例不超过 50%。获奖比例仅作为参考，组委会将根据赛题难易程度及整体答题情况，制定各奖项获奖最低分数线，未达到获奖最低分数线者不得奖。全国选拔赛一等奖选手获得全国总决赛参赛资格。

2.全国总决赛

全国总决赛设立一、二、三等奖，原则上各奖项的获奖比例为 10%、25%、40%，总获奖比例不超过 75%。获奖比例仅作为参考，组委会将根据赛题难易程度及整体答题情况，制定各奖项获奖最低分数线，未达到获奖最低分数线者不得奖。

十二、奖项公示

全国选拔赛及全国总决赛评审完成后，大赛组委会将在大赛官网开放奖项公示。参赛选手若对奖项有异议，可在 3 个工作日内按照大赛组委会相关要求提出复核申请。

十三、监督反馈

详见《蓝桥杯全国大学生软件和信息技术大赛章程》。

十四、其他注意事项

1.选手必须符合参赛资格，不得弄虚作假。资格审查中一旦发现问题，则取消其报名资格；竞赛过程中发现问题，则取消竞赛资格；竞赛后发现问题，则取消竞赛奖项，收回获奖证书及奖品等，并在大赛官方网站上公示。

2.参赛选手应严格遵守蓝桥杯大赛诚信参赛管理办法（办法链接：https://dasai.lanqiao.cn/notices/844/），服从大赛组委会的指挥和安排，爱护竞赛场地的设备。未尽事宜请参照组委会在大赛官方网站公布的通知、章程、比赛管理办法及相关要求并遵照执行。

3.赛项采用智能化评审系统自动评分与人工复核双轨校验机制，确保评审效率与精度。选手要特别注意提交答案的形式。必须仔细阅读题目的要求和示例，不得随意添加不需要的内容。

蓝桥杯大赛组委会

2026年10月8日

附录

样题

一、题目背景

过去六年，我国在线教育行业经历了“高速扩张—疫情催化—智能升级”的三重跃迁。

• 2019年，“互联网+”教育走向纵深，全国在线课程注册人数首次突破3亿，优质课件、直播互动和灵活付费模式重塑了传统课堂。

• 2020—2021年，新冠疫情迫使全国中小学和高校转向“停课不停学”。在线教育用户量在两个月内增加近50%，同时暴露出内容同质化、学习黏性低、数据孤岛等结构性问题。

• 2022年起，在“双减”政策与数字中国战略指引下，行业从“量的增长”转向“质的提升”，重点攻克个性化学习路径、学习效果评估与生态出海三大课题。

• 2024—2025年，大模型与生成式AI大规模落地，在线教育平台加速“数智化运营”：实时行为数据用于精准推送与课程迭代，沉浸式互动与全球化布局同步推进，推动教育公平和产业升级。

为了评估数据驱动策略的可行性，某头部在线教育平台发布了经脱敏处理的“用户学习行为—课程元数据—完成度预测”三大数据集，涵盖420万余条观看日志、近2000门课程属性以及大规模预测样本。

本竞赛旨在让参赛者通过数据预处理、分析和可视化三大任务，挖掘用户留存机制、课程质量指标与全国学习时差等关键洞见。

二、数据集说明

1.数据源说明

本赛题提供三个数据集文件，分别为login.csv、study_information.csv及users.csv。由于篇幅有限，在此仅做数据源文件部分展示：

user_id	login_time	login_place
用户3	2023/9/6 9:32	中国广东广州
用户3	2023/9/7 9:28	中国广东广州
用户3	2023/9/7 9:57	中国广东广州
用户3	2023/9/7 10:55	中国广东广州
用户3	2023/9/7 12:28	中国广东广州
用户3	2023/9/10 9:18	中国广东广州
用户3	2023/9/10 9:53	中国广东广州
用户3	2023/9/10 11:28	中国广东广州
用户3	2023/9/10 14:04	中国北京
用户3	2023/9/10 14:36	中国广东广州

Login.csv文件部分数据及字段展示

user_id	course_id	course_join_time	learn_process	price
用户3	课程106	2025/4/21 10:11	width: 0%;	0
用户3	课程136	2025/3/5 11:44	width: 1%;	0
用户3	课程205	2023/9/10 18:17	width: 63%;	0
用户4	课程26	2025/3/31 10:52	width: 0%;	319
用户4	课程34	2025/3/31 10:52	width: 0%;	299
用户4	课程22	2025/3/31 10:52	width: 0%;	199
用户4	课程17	2025/3/31 10:52	width: 0%;	299
用户4	课程31	2025/3/31 10:52	width: 0%;	109

Study_information.csv 文件部分数据及字段展示

user_id	register_time	recently_logged	number_of_classes_join	number_of_classes_out	learn_time	school
用户44251	2025/6/18 9:49	2025/6/18 9:49	0	0	41.25	北京大学
用户44250	2025/6/18 9:47	2025/6/18 9:48	0	0	0	清华大学
用户44249	2025/6/18 9:43	2025/6/18 9:43	0	0	16.22	复旦大学
用户44248	2025/6/18 9:09	2025/6/18 9:09	0	0	0	中国科学院大学
用户44247	2025/6/18 7:41	2025/6/18 8:15	0	0	1.8	浙江大学城市学院
用户44246	2025/6/17 22:36	2025/6/17 22:36	0	0	48.92	厦门大学
用户44245	2025/6/17 22:16	2025/6/17 22:16	0	0	0.18	福州大学
用户44244	2025/6/17 20:59	2025/6/17 21:34	0	0	0	四川大学

Users.csv 文件部分数据及字段展示

2.字段说明

各文件字段说明如下所示：

1. 统一字段说明

user_id 为用户唯一标识符，三个文档统一

course_id 为课程唯一标识符，三个文档统一

2. users.csv ——注册与学习概况

• 字段说明

register_time、recently_logged 为字符串时间戳（YYYY/M/D H:MM）；

number_of_classes_join/out：已报名/已退出的班级计数，均为整型。

learn_time：累计学习时长（小时）。

school：用户所在学校。

3. study_information.csv —— 报名记录与进度

• 字段说明

learn_process：学习进度，百分比类型。

price：课程结算价。

course_join_time：课程报名时间。

4. login.csv —— 活跃日志

• 字段说明

login_time：登录时间。

login_place：地理位置（“中国省份城市”格式）。

三、任务要求

任务一：数据预处理（15 分）

1.1 填补学习时长缺失数据（5 分）

题型：计算题

【背景】

在用户行为数据分析中，最近登录时间（recently_logged）是衡量平台活跃度与留存情况的关键指标。然而，由于采集延迟、接口异常或用户首次注册后尚未登录等原因，日志中常会出现缺失值。如果不及时处理，这些空缺不仅会影响后续活跃度、留存率等核心指标的计算，还可能导致模型训练偏差。

【要求】

为确保数据完整性与分析准确性，本任务统计针对文档 users.csv 中，字段 recently_logged 数据缺失率（缺失率=无数据个数/该数据集总数据条数），将结果保留两位小数后，写入文档 task0101.txt 文件中。（仅需要填写最后的数字即可，无须其他任何说明）；并将该字段的缺失值，用该用户的账号注册时间填充（将结果写入 users_new.csv 文档中）。

【规定】

请严格按照考试步骤操作，请勿修改考试默认提供项目中的文件名称、文件夹路径等。

【评分标准】

本题完全实现题目目标得满分，“缺失率”计算正确得 5 分，否则不得分；字段 recently_logged 不存在缺失值，否则不得分。

1.2 解析学习进度百分比（5 分）

题型：计算、操作题

【背景】

在在线教育平台的学习行为数据中，学习进度（learn_process）是评估课程完成度、预测学习成果、以及设计个性化推荐的核心指标。由于前端版本迭代数据会存在一定的杂质。这些杂质数据将直接影响后续的课程完成率统计、用户进度分布可视化和预测模型训练。

【要求】

提取 study_information.csv 文档中的字段 learn_process 内的学习进度百分比数据，若数据提取失败，则视为无效数据，并将该条数据删除，同时统计无效数据行数，将结果填

写至考生文件夹中，task0102.txt 文件中。（仅需要填写最后的数字即可，无须其他任何说明）；并将结果填写至考生文件夹中，study_information_new.csv 文件中。

【规定】

请严格按照考试步骤操作，请勿修改考试默认提供项目中的文件名称、文件夹路径等。

【评分标准】

本题完全实现题目目标得满分，“task0102.txt”文件中的结果计算正确得 5 分，否则不得分；task0102.txt 文件中总行数正确得 5 分，否则不得分。

1.3 学习数据去重（5 分）

题型：操作题

【背景】

在分析用户选课行为时，同一名用户因重复报名或课程升级而出现多条相同 course_id 的记录。

【要求】

文档 study_information_new.csv 中记录了用户学习的所有数据，请以字段 user_id 及 course_id 做去重处理，仅保留最新 course_join_time，并将去重后的结果写入文档 clean_new_study_data.csv 中。写出必要的 Python 片段实现去重。

【规定】

请严格按照考试步骤操作，请勿修改考试默认提供项目中的文件名称、文件夹路径等。

【评分标准】

本题完全实现题目目标得满分，运行考生提供的 Python 代码，处理测试集中的文档，得到新的文档 clean_new_study_data.csv，并且文件中总行数正确得 10 分，否则若代码成功运行，得 2 分，运行后能够处理样例文档，但处理结果错误，得 5 分。

任务二：数据分析（30 分）

2.1 新增用户 30 天留存率（10 分）

题型：计算题

【要求】

留存率是衡量新用户黏性的核心指标。文档 users_new.csv 中取注册日期为 2025-1-1 至 2025-1-31 注册，且注册后 30 天内至少登录一次的用户为新用户。计算留存率（新用户

/2025 年 1 月新注册的总人数），并将结果填写至考生文件夹中，task0201.txt 文件中。（仅需要填写最后的数字即可，无须其他任何说明）。

【规定】

请严格按照考试步骤操作，请勿修改考试默认提供项目中的文件名称、文件夹路径等。

【评分标准】

本题完全实现题目目标得满分，task0201.txt 文件中的结果计算正确得 10 分，否则不得分。

2.2 学校活跃度计算（10 分）

题型：计算题

【背景】

在线教育平台通常会与多所高校或培训机构合作，为其学生提供课程与服务。登录次数是衡量学生活跃度与课程黏性的关键指标。

【要求】

文档 users_new.csv 中，对于字段 school 中非空内容，计算人均登录次数（**计算单校平均用户数=学生用户总数/学校数**）；并将登录次数最高的 5 所学校，按照登录次数做降序排序，写入 task0202.txt 文件中。（仅需要填写最后的学校名称即可，每一行一所学校，无须其他任何说明）

【规定】

请严格按照考试步骤操作，请勿修改考试默认提供项目中的文件名称、文件夹路径等。

【评分标准】

本题完全实现题目目标得满分，task0202.txt 文件中存在 5 所学校，并且顺序正确得 10 分，否则命中一个学校得 1.5 分。

2.3 课程完成度计算（10 分）

题型：计算题

【背景】

在在线课程运营中，课程完成度直接反映学习内容的吸引力和教学效果。本任务将文档 clean_new_study_data.csv 中学习进度大于 80%视为“完成”课程的标准。

【要求】

请计算每门课程“完成”的学生人数，将“完成”人数最高的 50 门课，视为“人气课程”，并按照人数降序排序后，若人数相同，则按照课程名的拼音首字母升序排序，写入 task0203.txt 文件中。（仅需要填写最后的课程名称即可，每一行一个课程，无须其他任何说明）

【规定】

请严格按照考试步骤操作，请勿修改考试默认提供项目中的文件名称、文件夹路径等。

【评分标准】

本题完全实现题目目标得满分，task0203.txt 文件中存在 50 个课程名字所学校，并且顺序正确得 10 分，否则命中 10 个课程得 1.5 分，不足 10 个按 10 个计算。

任务三：数据可视化（30 分）

3.1 日活跃用户趋势（10 分）

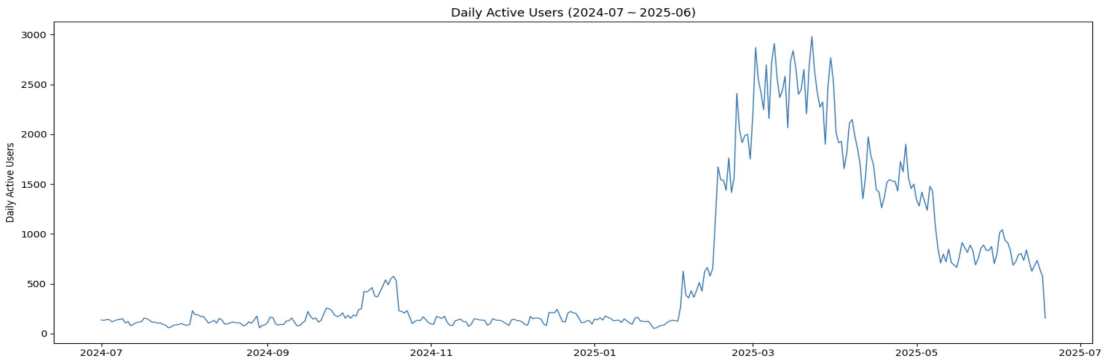
题型：绘图题

【背景】

日活跃用户数（DAU）是衡量平台黏性和运营成效的核心指标。

【要求】

请根据 login.csv 文档中用户数，定义当日登录的用户数为日活跃度，统计 2024-07~2025-06 每日活跃用户数，并绘制折线图。横坐标为日期，纵坐标为登录人数。将绘制的图像放至文档 task0301.xlsx 中。参考效果如下图所示。



【规定】

请严格按照考试步骤操作，请勿修改考试默认提供项目中的文件名称、文件夹路径等。

【评分标准】

本题完全实现题目目标得满分，task0301.xlsx 文件中存在折线图，并且横坐标、纵坐标均正确，得 10 分；否则按照如下评分表计分：

内容	得分
横坐标正确	4 分
纵坐标正确	4 分
特征点数据正确	2 分

3.2 登录热力图（10 分）

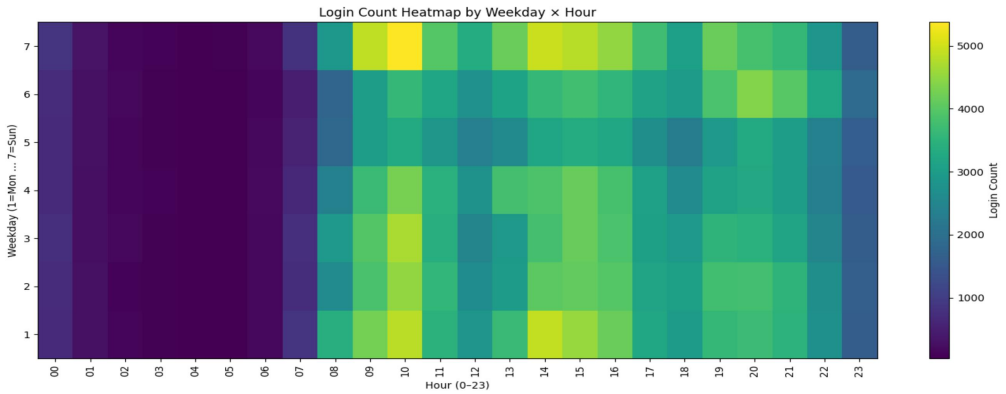
题型：绘图题

【背景】

用户登录的时间分布能够揭示平台流量的“高峰—低谷”规律。

【要求】

请根据 login.csv 文档中数据，按星期（1-7）X 小时（0~23）统计登录次数，并绘制热力图。将绘制的图像放至文档 task0302.xlsx 中。参考效果如下图所示。写出必要的 Python 片段实现热力图绘制。



【规定】

请严格按照考试步骤操作，请勿修改考试默认提供项目中的文件名称、文件夹路径等。

【评分标准】

本题完全实现题目目标得满分，task0302.xlsx 文件中存在热力图，并且横坐标、纵坐标均正确，得 10 分；否则按照如下评分表计分：

内容	得分
存在热力图	2 分（若不存在，本题计 0 分）
横坐标正确	4 分
纵坐标正确	4 分

3.3 课程进度分布柱形图（10 分）

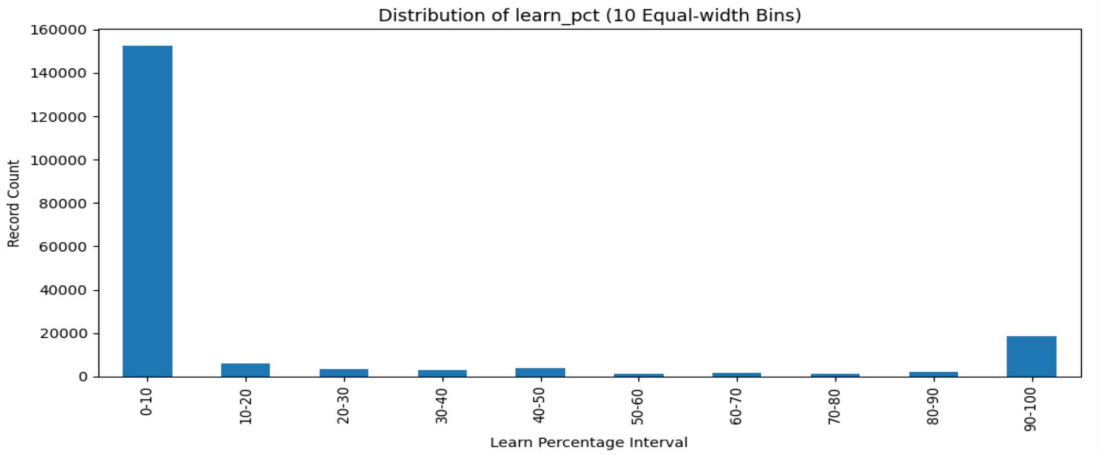
题型：绘图题

【背景】

课程进度的分布揭示了学生在整体学习路径中的聚集点与流失点。

【要求】

请根据 clean_new_study_data.csv 文档中数据，按 learn_pct 取 0-100 分为 10 个等宽的区间，统计排名次数并绘制柱形图。将绘制的图像放至文档 task0303.xlsx 中。参考效果如下图所示。



【规定】

请严格按照考试步骤操作，请勿修改考试默认提供项目中的文件名称、文件夹路径等。

【评分标准】

本题完全实现题目目标得满分，task0303.xlsx 文件中存在柱形图，并且横坐标、纵坐标均正确，得 10 分；否则按照如下评分表计分：

内容	得分
存在柱形图	2 分（若不存在，本题计 0 分）
横坐标正确	4 分
纵坐标正确	4 分

任务四：数据价值挖掘（25 分）

4.1 热门课程推荐（10 分）

题型：计算题

【背景】

高消费用户通常对平台内容黏性强、付费意愿高，是驱动收入与口碑扩散的关键群体。

【要求】

请根据文档 clean_new_study_data.csv，定义用户在平台消费金额之和，记为用户消费价值。为用户价值最高的 10 个用户，根据课程受欢迎程度排行中，推荐该用户 5 个未曾学过的课程。要求用 Python 代码完成，并按照如下格式，将结果写入文档 task0401.txt 中。

写入格式：

```
 用户 ID
--推荐课程 1
--推荐课程 2
--推荐课程 3
--推荐课程 4
--推荐课程 5
```

【规定】

请严格按照考试步骤操作，请勿修改考试默认提供项目中的文件名称、文件夹路径等。

【评分标准】

本题完全实现题目目标得满分，task0401.txt 文件中存在 10 个用户的 ID，并且每个用户推荐 5 门课程 ID 均正确得 10 分，否则每正确匹配一个用户 ID 及对应的推荐课程 ID 正确，得 1 分。

4.2 用户画像（15 分）

题型：计算题

【背景】

为制定差异化运营策略并量化用户价值，需基于无监督学习对全体用户进行客观分群用于指导拉新、促活、召回与增购等精细化运营决策。

【要求】

从 users_new.csv、login.csv 与 clean_new_study_data.csv 提取每位用户的累计登录次数、最近登录间隔（以日志最大时间为基准）与累计消费金额等特征，采用 K-means 完成聚类，固定随机种子确保可复现，最终输出仅含 user_id 与 cluster_id 的 user_cluster.csv。

具体分类要求如下所示：

cluster_id	特征概况	可视化标签
0	登录多、最近活跃、消费高	“高价值常客”（VIP）
1	各项接近平均	“普通稳定用户”
2	登录少、最近登录间隔大、无消费	“潜在流失”
3	登录次数稍高、最近活跃、轻度付费	“活跃试水者”
4	登录不多但花钱多、较久未回访	“高消费低活跃”（挽回重点）

【规定】

请严格按照考试步骤操作，请勿修改考试默认提供项目中的文件名称、文件夹路径等。

【评分标准】

本题完全实现题目目标得满分，user_cluster.csv 文件中，对以下 5 个点位进行判分，每正确匹配一个结果得 3 分。

user_id	cluster_id
用户 31	4
用户 46	3
用户 45	2
用户 50	1
用户 49	0