山东大学计算机科学与技术学院

可视化技术课程实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学号：202302130293 | 姓名：李嘉欣 | | 班级：数据科学与大数据技术 |
| 实验题目：一、数据采样方法实践 | | 实验日期：2025/9/17 | |
| 实验目标：利用Pandas库实现多种数据采样和过滤的方法 | | | |
| 实验环境：python3.11 | | | |
| 实验步骤：  1.库的导入与数据的读入    输出：    2.删除多余的空行并进行过滤  采用dropna方法并指定参数为any删除多余的空行    过滤得到traffic不等于0且from\_level=一般节点的数据    输出：    3.对数据进行抽样  加权采样：to\_level的值为一般节点与网络核心的权重之比为 1 : 5    随机抽样    分层抽样：根据to\_level的值进行分层采样    输出： | | | |
| 实验分析与体会：  一、实验分析  本次实验以网络流量数据集为对象，先通过dropna(how='any')删除原始数据中的空行，再筛选“traffic≠0且from\_level='一般节点'”的有效记录，为后续采样奠定基础。三种采样方法各有适配场景：加权采样按to\_level权重1:5分配，聚焦网络核心节点；随机采样反映数据整体分布；分层采样固定两类节点样本量，避免小类别被忽略。实验中还解决了路径转义（用原始字符串`r""`）、变量拼写错误等问题，凸显细节对代码运行的关键影响。  二、实验体会  此次实验让我实现理论到实践的落地，比如将加权采样的抽象逻辑拆解为“新增权重列—赋值—采样”的具体步骤，明白数据处理需转化业务逻辑。同时，每步操作后验证结果（如查NaN值、样本分布）的过程，让我重视严谨性的重要。此外，思考数据规模扩大后需改用Spark等工具，也让我意识到要根据场景动态调整方案，为后续学习积累了实用经验。 | | | |