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：二、数据质量实践 | | 实验日期：2025/9/24 | |
| 实验目标：本实验以宝可梦数据集为对象，建立数据预处理清洗流程，处理无意义行、Type 2 列异常值、重复数据、Attack 属性异常高值及属性置换问题，输出规范数据集，提升数据质量问题处理能力。 | | | |
| 实验环境：windows系统 & python3.11 | | | |
| 实验步骤：  1.数据的读入    输出：    2.数据清洗  2.1 删除无意义数据行    2.2 处理Type 2列异常值    2.3 删除重复数据    2.4 处理Attack属性异常高值    2.5 修正属性置换问题    输出：    3.清洗后的数据导出    输出： | | | |
| 实验分析与体会：  一、实验分析  本次实验围绕宝可梦数据集展开，通过多步骤清洗验证了数据预处理的关键逻辑。无意义行删除使数据从 810 行精简至有效行数，Type 2 列通过合理属性集合识别并处理 “0”“A” 等异常值，重复数据清零保障唯一性；Attack 属性经数值转换与 IQR 方法剔除 9 个异常值，属性置换问题修正后两列数据类型分别统一为 int64 和 bool。整个过程中发现，Type 2 列的缺失值包含合理空值与异常值转换结果，需结合业务逻辑区分处理，避免过度清洗破坏数据真实性。  二、实验体会  通过本次实验，深刻认识到数据质量是后续分析的基础。前期仅针对性处理 “273” 异常值导致遗漏其他问题，提醒我需建立全面的异常检测思维；重复数据与属性置换的处理过程，让我掌握了数据一致性校验的实用方法。同时发现，清洗需平衡“完整性”与“合理性”，如宝可梦单属性导致的 Type 2 列空值无需填充，这为后续处理类似数据集提供了宝贵经验，也提升了自身解决实际数据问题的能力。 | | | |