计算机科学与技术学院大数据分析与实践课程实验报告

实验题目：数据质量实践		学号：202322130197
日期：	班级：数据	姓名：崔嘉铭
Email：cjm13969665900@gmail.com		

实验目的：

本实验以 Pokémon 数据集为研究对象，围绕真实数据中常见的数据质量问题展开实践，重点学习并掌握以下内容：

1. 识别数据集中存在的缺失值、重复值、异常值和错误数据；
2. 使用 Pandas 等数据分析工具对数据进行清洗和预处理；
3. 建立一套基本的数据质量处理流程，为后续数据分析和建模奠定可靠的数据基础；
4. 加深对“数据质量直接影响分析结果可靠性”的理解。

实验软件和硬件环境：

- 操作系统：Windows
- 编程语言：Python 3.9
- 开发环境：Jupyter Notebook
- 主要使用库：
  - pandas
  - numpy
  - matplotlib（用于辅助可视化分析）

实验原理和方法：

在实际数据分析过程中，原始数据往往并不“干净”，可能包含缺失值、重复记录、异常值或人为录入错误。本实验基于数据质量管理的基本思想，按照以下流程进行处理：

1. **数据整体结构检查**
   使用 info()、describe() 等方法了解数据规模、字段类型及基本分布情况。
2. **缺失值与无效数据处理**
   对明显无意义或缺失严重的数据进行删除或修正，避免其干扰统计分析。
3. **重复值检测与处理**
   通过去重操作保证每条记录的唯一性，防止重复样本对统计结果造成偏置。
4. **异常值检测与分析**
   结合统计指标和常识判断异常值的合理性，对明显不合理的数据进行修正或剔除。
5. **逻辑错误修正**
   对字段之间存在逻辑冲突的数据进行人工校验和修复

实验步骤：（不要求罗列完整源代码）

1. 数据读取与初步检查

使用 Pandas 读取 Pokémon 数据集，通过查看前后几行数据发现：

- 数据集最后两行数据为空或无实际意义；
- 部分字段存在缺失或异常取值。

因此，首先对无意义的尾部数据进行删除，保证数据结构完整。

---

2. 缺失值处理

在 Type2 字段中，存在异常取值和缺失情况。考虑到该字段表示 Pokémon 的第二属性，部分 Pokémon 本身只有一种属性，因此将异常或非法取值统一处理为缺失值（NaN），并保留记录本身，以免丢失有效样本。

---

3. 重复值检测

通过 duplicated() 方法检测数据集中是否存在完全重复的记录。实验发现数据集中确实存在重复行，若不处理将导致统计结果偏向重复样本，因此使用 drop_duplicates() 方法对数据进行去重处理。

---

4. 异常值分析与处理

在 Attack（攻击力）属性中发现存在明显高于正常范围的极端值。

通过结合箱线图和统计分布分析，判断该值明显偏离总体分布，可能来源于录入错误或数据异常。为避免其对后续分析造成过大影响，本实验选择将该异常值删除。

---

5. 数据逻辑错误修正

实验过程中发现有两条记录中 Generation 与 Legendary 字段被错误置换，导致字段含义不一致。

通过人工检查数据分布和字段取值范围，确认错误后对这两条记录进行了字段值交换修正，恢复其正确含义。

---

6. 清洗结果确认

在完成上述步骤后，再次对数据集进行整体检查，确认：

- 不存在明显缺失的关键字段；
- 无重复记录；
- 异常值已被合理处理；
- 数据逻辑一致性良好。

至此，数据质量处理流程完成。

结论分析与体会：

通过本次实验，我对数据质量问题有了更加直观和深入的认识。实验表明，真实世界的数据往往存在多种问题，如果直接进行分析或建模，可能得到完全错误的结论。

在数据分析之前，系统性地进行数据清洗和质量控制是非常必要的步骤。本实验不仅提升了我使用 Pandas 进行数据处理的熟练度，也让我认识到数据分析

并非只依赖算法，前期的数据准备同样至关重要。
就实验过程中遇到和出现的问题，你是如何解决和处理的，自拟 1－3 道问答题： **问题一：是否所有异常值都应该直接删除？** 答：不一定。异常值需要结合具体业务背景判断。有些异常值可能代表极端但真实的情况，而明显违反常识或数据分布规律的异常值才适合删除或修正。