山东大学计算机科学与技术学院

大数据分析实践课程实验报告

学号：202300130098	姓名：马浩鑫	班级：23 级数据
实验题目：二、数据质量实践		
实验学时：2 课时	实验日期：	

实验目标：

本次实验主要围绕宝可梦数据集进行分析，考察在拿到数据后如何对现有的数据进行预处理清洗操作，建立起对于脏数据、缺失数据等异常情况的一套完整流程的认识。

实验描述：

1. 实验环境

python3.9，jupyter notebook

2. 数据集

Pokeman Dataset：721 Pokemon, including their number, name, first and second type, and basic stats: HP, Attack, Defense, Special Attack, Special Defense, and Speed
http://storage.ameshollland.xyz/Pokemon.csv

3. 分析数据

（1）数据总览

```
[13]
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	FALSE
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	FALSE
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	FALSE
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	FALSE
4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	FALSE

数据总行数：810 | 总列数：13

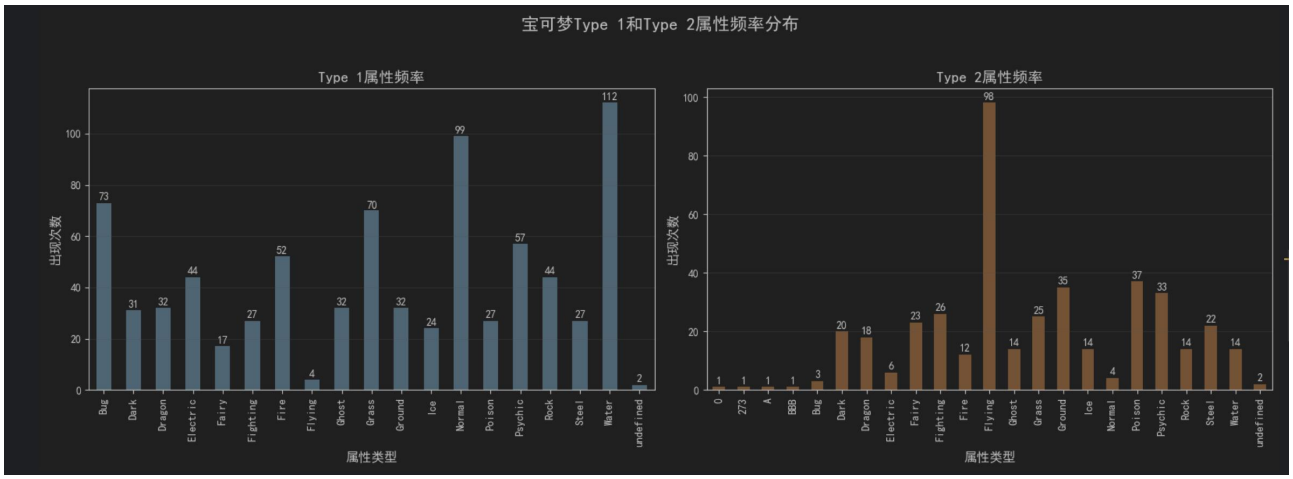
（2）统计缺失行

```
1. 空行统计（含伪空行）：
   - 空行总数：3行
   - 空行索引：[408, 808, 809]

2. 无定义行统计（属性值为'undefined'）：
   - 无定义行数量：2行
   - 无定义行索引：[806, 771]
```

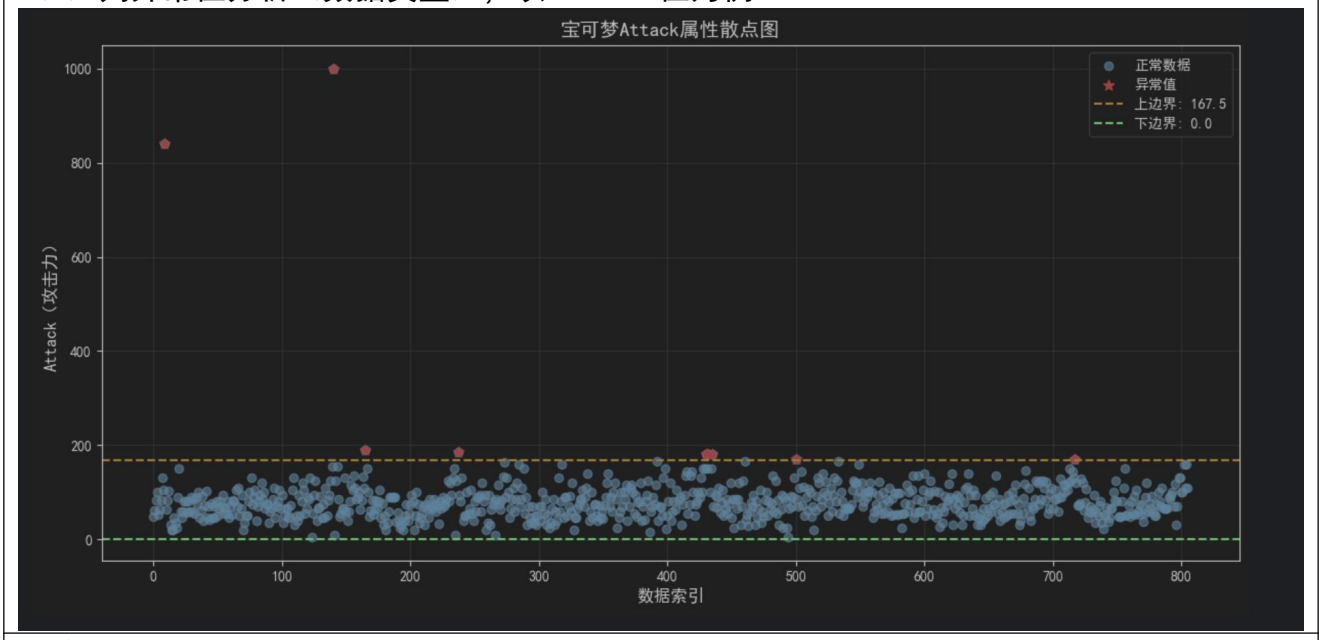
（3）检测完全重复行

```
1. 完全重复行数量：12
   - 重复行索引：[14, 15, 21, 23, 184, 185, 186, 187, 806, 807, 808, 809]
   - 重复行分组详情（按宝可梦编号#聚合，符合指导书分析逻辑）：

     编号11的重复行（共2行）：
        #     Name Type 1 Type 2 Total  HP Attack
14   11  Metapod    Bug    NaN   205  50     20
15   11  Metapod    Bug    NaN   205  50     20

     编号168的重复行（共4行）：
        #     Name Type 1  Type 2 Total  HP Attack
184  168  Ariados    Bug  Poison   390  70     90
185  168  Ariados    Bug  Poison   390  70     90
186  168  Ariados    Bug  Poison   390  70     90
```

（4）列异常值分析（字段类型），以 type1、type2 为例



（5）列异常值分析（数据类型），以 attack 值为例



结论分析：

基于宝可梦数据集（721 只宝可梦，含编号、类型、基础属性等），发现以下 4 类数据质量问题：

1. 无意义空行：数据集底部存在多行为 NaN 的空行（指导书提及 “最后两行无意义”），需删除避免统计偏差；Type 2（第二属性）缺失属合理情况（部分宝可梦无第二属性），无需填充。

2. 重复值：存在完全重复行（如 Ariados 重复记录），源于录入错误，会高估样本占比，需保留首行删除其余。

属性异常：

3. 分类属性：Type 2 含 “273” 等非标准类型，Legendary（是否传说）存非布尔值，属录入错误；

数值属性：Attack（攻击力）有过高值，部分为 Mega 进化合理属性，需排除非进化错误值；Generation（世代）存超 1-6 范围值，需修正。

4. 列混淆：Generation 与 Legendary 值置换（如世代填布尔值、是否传说填整数），致分析逻辑失效，需按字段规则交换修复。