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：202300130083	姓名：王乐临	班级：数据
实验题目：数据质量实践		
实验学时：2	实验日期：2025/9/26	
实验目标：围绕宝可梦数据集，开展数据预处理清洗操作，建立对脏数据、缺失数据等异常情况的完整处理流程认知。		

实验步骤与内容：

一、数据加载与初步探查

按文档指引加载数据集，重点关注数据的行列数量、列名及数据类型；查看数据首尾部分，初步识别无效行、列取值异常等问题；记录初步发现的问题，为后续清洗提供方向。

二、无效行删除

1.依据文档明确提示最后两行数据无意义，可直接删去，先定位末尾无意义行；
2.进一步识别全表中的全空行；
3.统一删除上述两类无效行。

三、分类列（Type 2）异常值处理

1.聚焦文档提及的异常列 Type 2，查看其所有唯一值，筛选出非合理属性的异常值；
2.考虑到 Type 2 为宝可梦次要属性，空值属于正常情况，因此将异常值转为 NaN；
3.验证处理后 Type 2 的唯一值，仅保留合理属性与 NaN，确保分类列取值符合逻辑。

四、重复值检测与清除

1.参考文档示例，检测数据中的完全重复行，统计重复行数量；
2.采用保留首次出现行、删除后续重复行的策略，执行重复值删除操作；
3.再次检测重复行，确认重复数据已清除。

五、数值列（Attack）异常值处理

1.针对文档指出的 Attack 属性存在过高的异常值，先通过统计描述定位异常值；
2.结合宝可梦能力值常识，设定合理阈值；
3.选择中位数替换，用 Attack 列的中位数替换异常值，确保数值列分布符合实际情况；
4.验证处理后 Attack 列的统计指标，确认异常值已修正。

六、属性置换问题修正

1.依据文档提示有两条数据的 generation 与 Legendary 属性被置换，先检查两列取值类型：Generation 应为数值 1-6，Legendary 应为布尔值 TRUE/FALSE；
2.识别置换行即 Generation 为 TRUE/FALSE、Legendary 为数字的行，统计置换行数；
3.交换置换行中两列的数值，再分别修正两列的数据类型，确保属性与数值匹配正确。

七、清洗结果验证与数据保存

全面验证清洗后数据：检查是否仍有无效行、重复值、异常值、属性置换问题，统计清洗前后数据量变化，确认数据完整性后，将清洗后的数据集保存为新文件。

结果：

#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generatior	Legendary
1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	FALSE
2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	FALSE
3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	FALSE
3	VenusaurMe	Grass	Poison	625	80	100	123	122	120	80	1	FALSE
4	Charmander	Fire	Flying	309	39	52	43	60	50	65	1	FALSE
5	Charmeleor	Fire	Flying	405	58	64	58	80	65	80	1	FALSE
6	Charizard	Fire	Flying	534	78	84	78	109	85	100	1	FALSE
6	CharizardM	Fire	Dragon	634	78	130	111	130	85	100	1	FALSE
6	CharizardM	Fire	Flying	634	78	104	78	159	115	100	1	FALSE
7	Squirtle	Water	Flying	314	44	75	65	50	64	43	1	FALSE
8	Wartortle	Water	Flying	405	59	63	80	65	80	58	1	FALSE
9	Blastoise	Water	Flying	530	79	83	100	85	105	78	3	FALSE
9	BlastoiseM	Water	Flying	630	79	103	120	135	115	78	1	FALSE
10	Caterpie	Bug	Flying	195	45	30	35	20	20	45	1	FALSE
11	Metapod	Bug	Flying	205	50	20	55	25	25	30	1	FALSE
12	Butterfree	Bug	Flying	395	60	45	50	90	80	70	1	FALSE
13	Weedle	Bug	Poison	195	65	35	30	20	20	50	1	FALSE
14	Kakuna	Bug	Poison	205	45	25	50	25	25	35	1	FALSE
15	Beedrill	Bug	Poison	395	65	90	40	45	80	75	1	FALSE
15	BeedrillMe	Bug	Poison	495	65	150	40	15	80	145	1	FALSE
17	Pidgeotto	Normal	Flying	349	63	60	55	50	50	71	1	FALSE
16	Pidgey	Normal	Flying	251	40	45	40	35	35	56	1	FALSE
18	Pidgeot	Normal	Flying	479	83	80	75	70	70	101	1	FALSE
18	PidgeotMeg	Normal	Flying	579	83	80	80	135	80	121	1	FALSE
19	Rattata	Normal	Flying	253	30	56	35	25	35	72	1	FALSE
20	Raticate	Normal	Flying	413	55	81	60	50	70	97	1	FALSE
21	Spearow	Normal	Flying	262	40	60	30	31	31	70	1	FALSE

结果分析：

1. 所有清洗步骤均围绕文档明确指出的宝可梦数据集问题展开，不额外新增无关处理；
2. 对异常数据优先采用修正，如中位数替换、属性交换，而非删除，避免样本量过度减少；
3. 分类列（Type 2）、数值列（Attack）、标识列（Generation/Legendary）的处理方式均贴合字段业务含义，如次要属性为空合理、宝可梦能力值有常识范围，确保清洗后数据符合实际逻辑。