山东大学计算机科学与技术学院

大数据分析与实践实验报告

学号：202300130214	姓名：于博雅	班级：23级数据班
实验题目：数据质量实践		
实验学时：2	实验日期：2025.9.29	

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。

实验步骤与结果：

针对数据集存在的部分问题进行处理：

原数据集展示：

	#	Name	Type 1	Type 2	Total	HP
0	1	Bulbasaur	Grass	Poison	318	45
1	2	Ivysaur	Grass	Poison	405	60
2	3	Venusaur	Grass	Poison	525	80
3	3	VenusaurMega Venusaur	Grass	Poison	625	80
4	4	Charmander	Fire	NaN	309	39
...	...	...	...	...	...	...
805	721	Volcanion	Fire	Water	600	80
806	undefined	undefined	undefined	undefined	undefined	undefined
807	undefined	undefined	undefined	undefined	undefined	undefined
808	NaN	NaN	NaN	NaN	NaN	NaN
809	NaN	NaN	NaN	NaN	NaN	NaN

810 rows × 13 columns

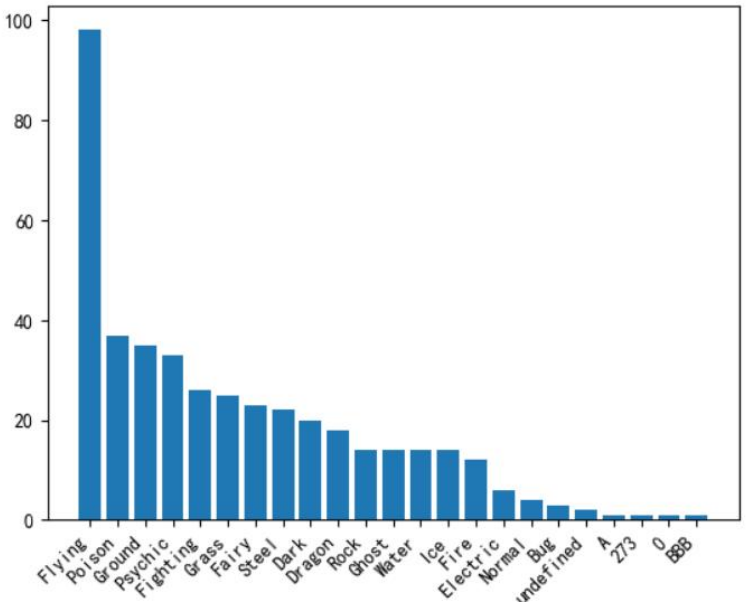
- 最后两行数据无意义，可直接删去

	#	Name	Type 1	Type 2	Total	HP
0	1	Bulbasaur	Grass	Poison	318	45
1	2	Ivysaur	Grass	Poison	405	60
2	3	Venusaur	Grass	Poison	525	80
3	3	VenusaurMega Venusaur	Grass	Poison	625	80
4	4	Charmander	Fire	NaN	309	39
...	...	...	...	...	...	...
803	720	HoopaHoopa Confined	Psychic	Ghost	600	80
804	720	HoopaHoopa Unbound	Psychic	Dark	680	80
805	721	Volcanion	Fire	Water	600	80
806	undefined	undefined	undefined	undefined	undefined	undefined
807	undefined	undefined	undefined	undefined	undefined	undefined

- 808 rows × 13 columns

- type2存在异常的数值取值，可清空

```
Type 2
Flying       98
Poison       37
Ground       35
Psychic      33
Fighting     26
Grass        25
Fairy        23
Steel        22
Dark         20
Dragon       18
Rock         14
Ghost        14
Water        14
Ice          14
Fire         12
Electric      6
Normal        4
Bug           3
undefined     2
A             1
273           1
0             1
BBB           1
Name: count, dtype: int64
```



	#	Name	Type 1	Type 2	Total	HP
0	1	Bulbasaur	Grass	Poison	318	45
1	2	Ivysaur	Grass	Poison	405	60
2	3	Venusaur	Grass	Poison	525	80
3	3	VenusaurMega Venusaur	Grass	Poison	625	80
4	4	Charmander	Fire	NaN	309	39
...	...	...	...	...	...	...
803	720	HoopaHoopa Confined	Psychic	Ghost	600	80
804	720	HoopaHoopa Unbound	Psychic	Dark	680	80
805	721	Volcanion	Fire	Water	600	80
806	undefined	undefined	undefined	undefined	undefined	undefined
807	undefined	undefined	undefined	undefined	undefined	undefined

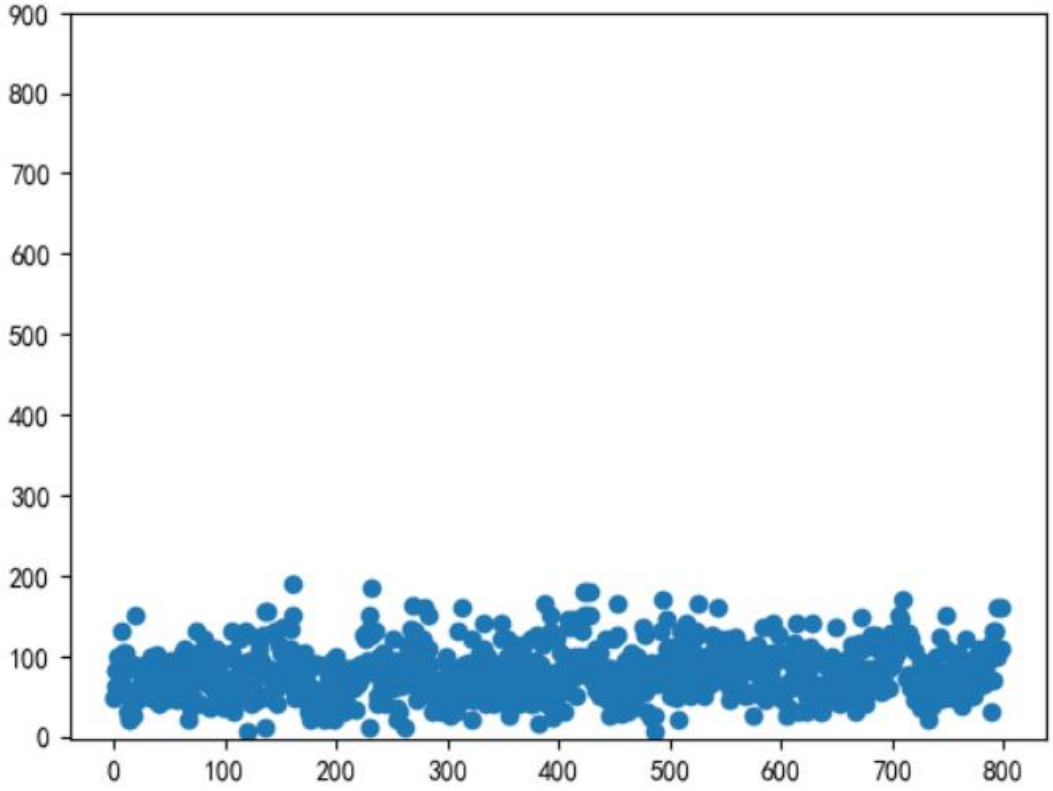
804 rows × 13 columns

- 数据集中存在重复值

	#	Name	Type 1	Type 2	Total	HP
0	1	Bulbasaur	Grass	Poison	318	45
1	2	Ivysaur	Grass	Poison	405	60
2	3	Venusaur	Grass	Poison	525	80
3	3	VenusaurMega Venusaur	Grass	Poison	625	80
4	4	Charmander	Fire	NaN	309	39
...	...	...	...	...	...	...
803	720	HoopaHoopa Confined	Psychic	Ghost	600	80
804	720	HoopaHoopa Unbound	Psychic	Dark	680	80
805	721	Volcanion	Fire	Water	600	80
806	undefined	undefined	undefined	undefined	undefined	undefined
807	undefined	undefined	undefined	undefined	undefined	undefined

804 rows × 13 columns

- Attack属性存在过高的异常值



- 有两条数据的generation与Legendary属性被置换

```
Index([11, 32], dtype='int64')
Index([], dtype='int64')
```

结论分析与体会：

熟悉了分析数据质量的方法，掌握了多种数据清洗的方法。