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：202300130153	姓名：吴宇轩	班级：数据 23
实验题目：数据质量实践		
实验学时：2	实验日期：2025.9.26	

实验目的：

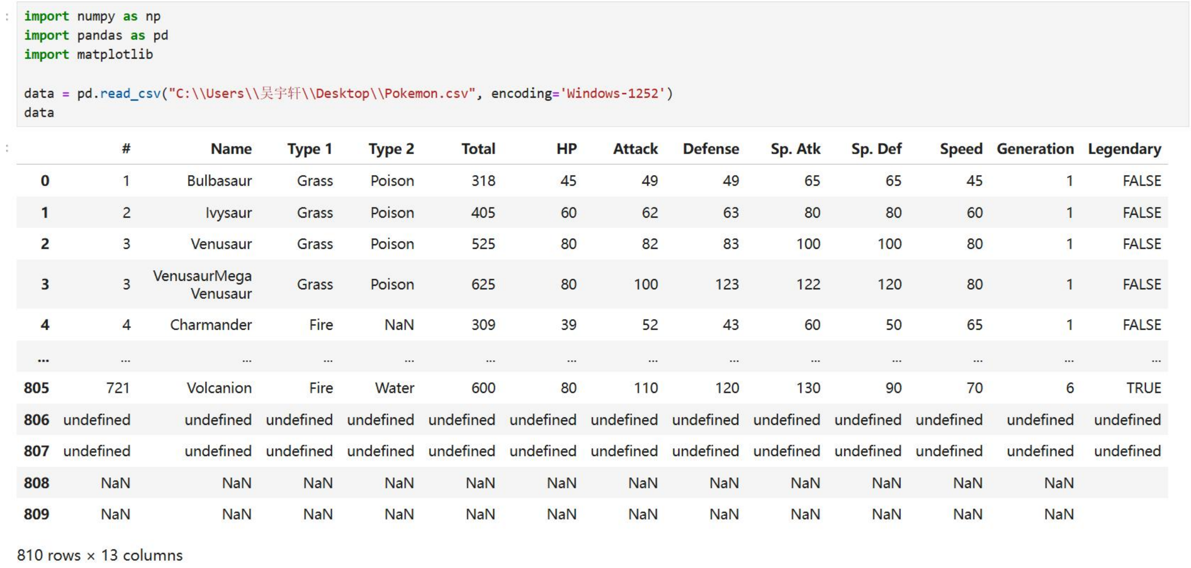
本次实验主要围绕宝可梦数据集进行分析，考察在拿到数据后如何对现有的数据进行预处理清洗操作，建立起对于脏数据、缺失数据等异常情况的一套完整流程的认识。

实验环境：

Python3.9, Jupyter notebook

实验步骤与内容：

1、导入数据集



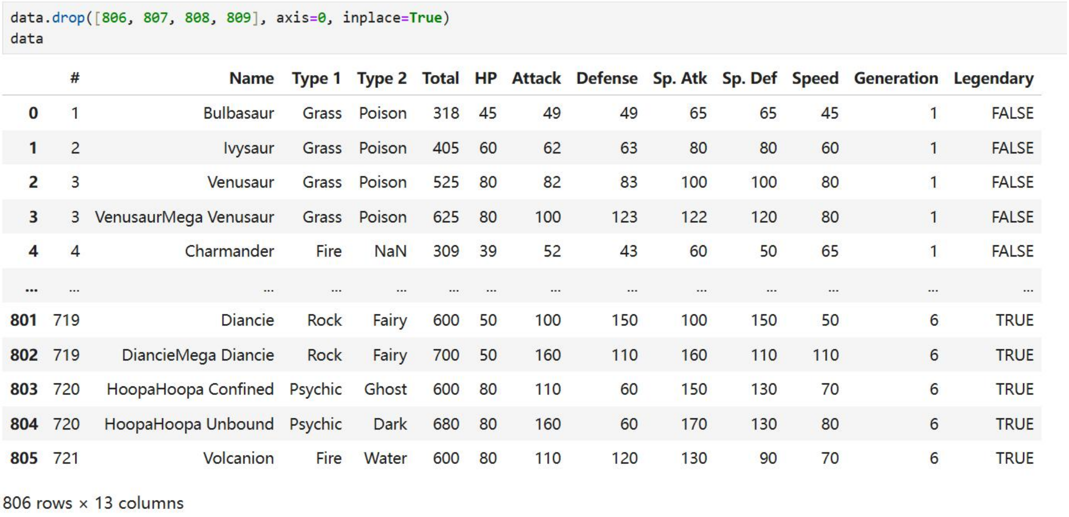
```
import numpy as np
import pandas as pd
import matplotlib

data = pd.read_csv("C:\\Users\\吴宇轩\\Desktop\\Pokemon.csv", encoding='Windows-1252')
data
```

	#	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...
805	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	TRUE
806	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
807	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
808	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
809	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

810 rows × 13 columns

2、删除无意义数据

最后四行数据无意义，直接删去。



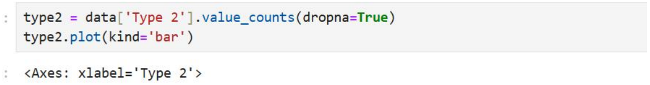
```
data.drop([806, 807, 808, 809], axis=0, inplace=True)
data
```

	#	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...
801	719	Diancie	Rock	Fairy	600	50	100	150	100	150	50	6	TRUE
802	719	DiancieMega Diancie	Rock	Fairy	700	50	160	110	160	110	110	6	TRUE
803	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110	60	150	130	70	6	TRUE
804	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160	60	170	130	80	6	TRUE
805	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	TRUE

806 rows × 13 columns

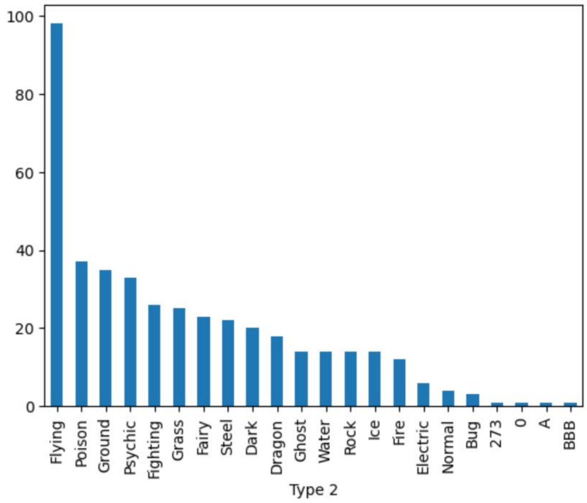
3、删除存在异常值的数据

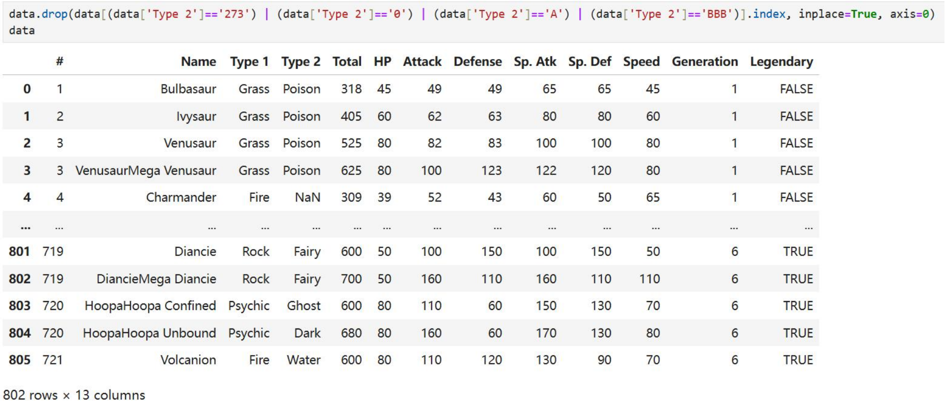
对 Type2 列的取值频次进行可视化，发现存在少数无意义异常值，删除。



```
type2 = data['Type 2'].value_counts(dropna=True)
type2.plot(kind='bar')
```

```
<Axes: xlabel='Type 2'>
```



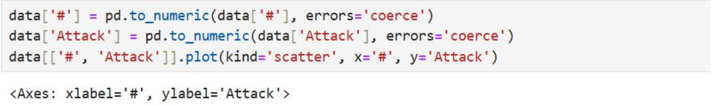


```
data.drop(data[(data['Type 2']=='273') | (data['Type 2']=='0') | (data['Type 2']=='A') | (data['Type 2']=='BBB')].index, inplace=True, axis=0)
data
```

	#	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...
801	719	Diancie	Rock	Fairy	600	50	100	150	100	150	50	6	TRUE
802	719	DiancieMega Diancie	Rock	Fairy	700	50	160	110	160	110	110	6	TRUE
803	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110	60	150	130	70	6	TRUE
804	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160	60	170	130	80	6	TRUE
805	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	TRUE

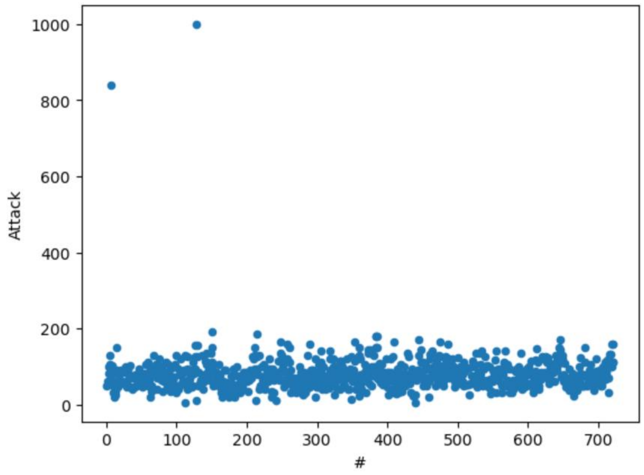
802 rows × 13 columns

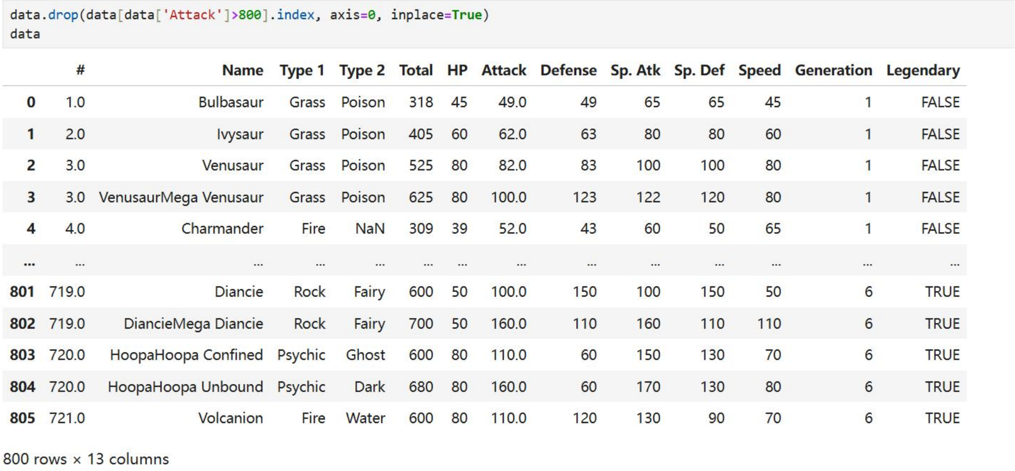
对 Attack 列的值进行可视化，发现存在少数无意义异常值，删除。



```
data['#'] = pd.to_numeric(data['#'], errors='coerce')
data['Attack'] = pd.to_numeric(data['Attack'], errors='coerce')
data[['#', 'Attack']].plot(kind='scatter', x='#', y='Attack')
```

```
<Axes: xlabel='#', ylabel='Attack'>
```



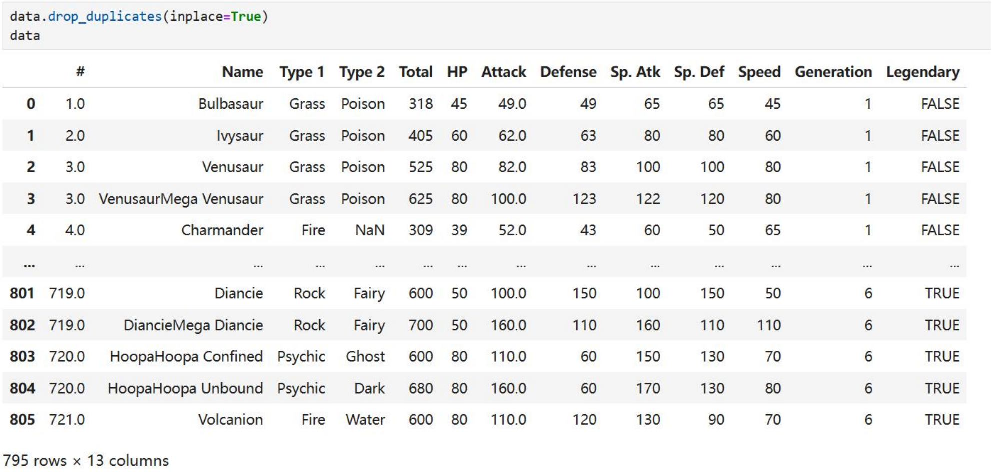


```
data.drop(data[data['Attack']>800].index, axis=0, inplace=True)
data
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1.0	Bulbasaur	Grass	Poison	318	45	49.0	49	65	65	45	1	FALSE
1	2.0	Ivysaur	Grass	Poison	405	60	62.0	63	80	80	60	1	FALSE
2	3.0	Venusaur	Grass	Poison	525	80	82.0	83	100	100	80	1	FALSE
3	3.0	VenusaurMega Venusaur	Grass	Poison	625	80	100.0	123	122	120	80	1	FALSE
4	4.0	Charmander	Fire	NaN	309	39	52.0	43	60	50	65	1	FALSE
...	...	...	...	...	...	...	...	...	...	...	...	...	...
801	719.0	Diancie	Rock	Fairy	600	50	100.0	150	100	150	50	6	TRUE
802	719.0	DiancieMega Diancie	Rock	Fairy	700	50	160.0	110	160	110	110	6	TRUE
803	720.0	HoopaHoopa Confined	Psychic	Ghost	600	80	110.0	60	150	130	70	6	TRUE
804	720.0	HoopaHoopa Unbound	Psychic	Dark	680	80	160.0	60	170	130	80	6	TRUE
805	721.0	Volcanion	Fire	Water	600	80	110.0	120	130	90	70	6	TRUE

800 rows × 13 columns

4、删除重复值

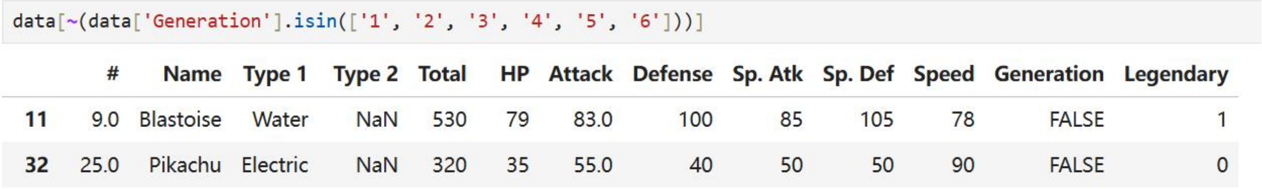


```
data.drop_duplicates(inplace=True)
data
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1.0	Bulbasaur	Grass	Poison	318	45	49.0	49	65	65	45	1	FALSE
1	2.0	Ivysaur	Grass	Poison	405	60	62.0	63	80	80	60	1	FALSE
2	3.0	Venusaur	Grass	Poison	525	80	82.0	83	100	100	80	1	FALSE
3	3.0	VenusaurMega Venusaur	Grass	Poison	625	80	100.0	123	122	120	80	1	FALSE
4	4.0	Charmander	Fire	NaN	309	39	52.0	43	60	50	65	1	FALSE
...	...	...	...	...	...	...	...	...	...	...	...	...	...
801	719.0	Diancie	Rock	Fairy	600	50	100.0	150	100	150	50	6	TRUE
802	719.0	DiancieMega Diancie	Rock	Fairy	700	50	160.0	110	160	110	110	6	TRUE
803	720.0	HoopaHoopa Confined	Psychic	Ghost	600	80	110.0	60	150	130	70	6	TRUE
804	720.0	HoopaHoopa Unbound	Psychic	Dark	680	80	160.0	60	170	130	80	6	TRUE
805	721.0	Volcanion	Fire	Water	600	80	110.0	120	130	90	70	6	TRUE

795 rows × 13 columns

5、有两条数据的 generation 与 Legendary 属性被置换



```
data[~(data['Generation'].isin(['1', '2', '3', '4', '5', '6']))]
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
11	9.0	Blastoise	Water	NaN	530	79	83.0	100	85	105	78	FALSE	1
32	25.0	Pikachu	Electric	NaN	320	35	55.0	40	50	50	90	FALSE	0

结论分析与体会：

本次围绕宝可梦数据集的数据质量实践实验，通过导入数据-清洗异常-优化数据的流程化操作，有效解决了原始数据中的多重质量问题，最终得到了结构完整、逻辑一致、可用于后续分析的高质量数据集。

通过本次实验认识到数据预处理对于数据分析的重要作用，认识到可视化是了解数据、观察数据的重要手段，同时掌握了 pandas 进行数据清洗的主要工具。