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：202300130045	姓名：张博文	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。

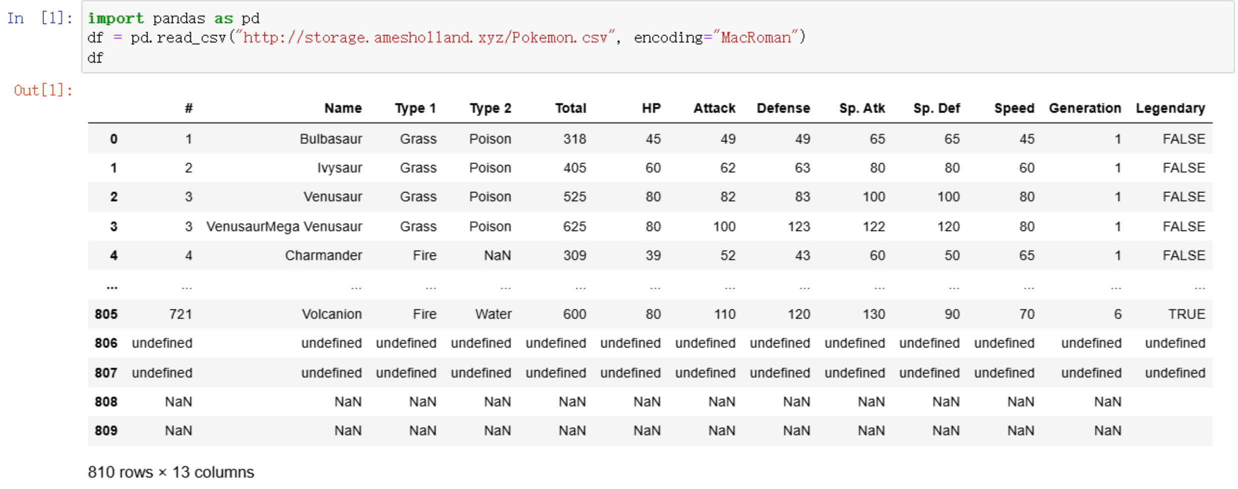
硬件环境：

计算机一台

软件环境：

python3.9，jupyter notebook

实验步骤与内容：

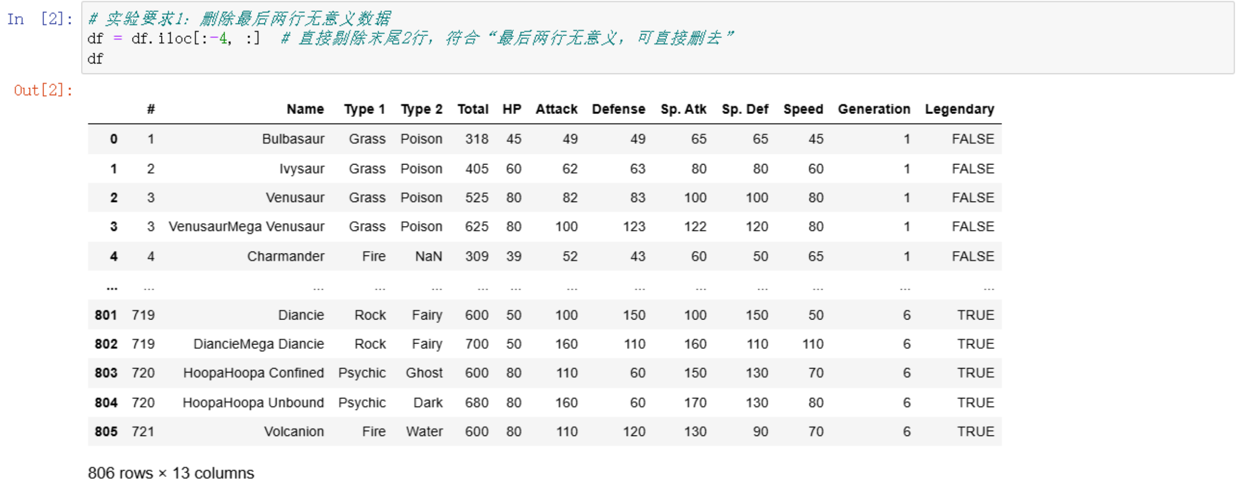


```
In  [1]: import pandas as pd
         df = pd.read_csv("http://storage.amesholland.xyz/Pokemon.csv", encoding="MacRoman")
         df
```

Out[1]:

	#	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

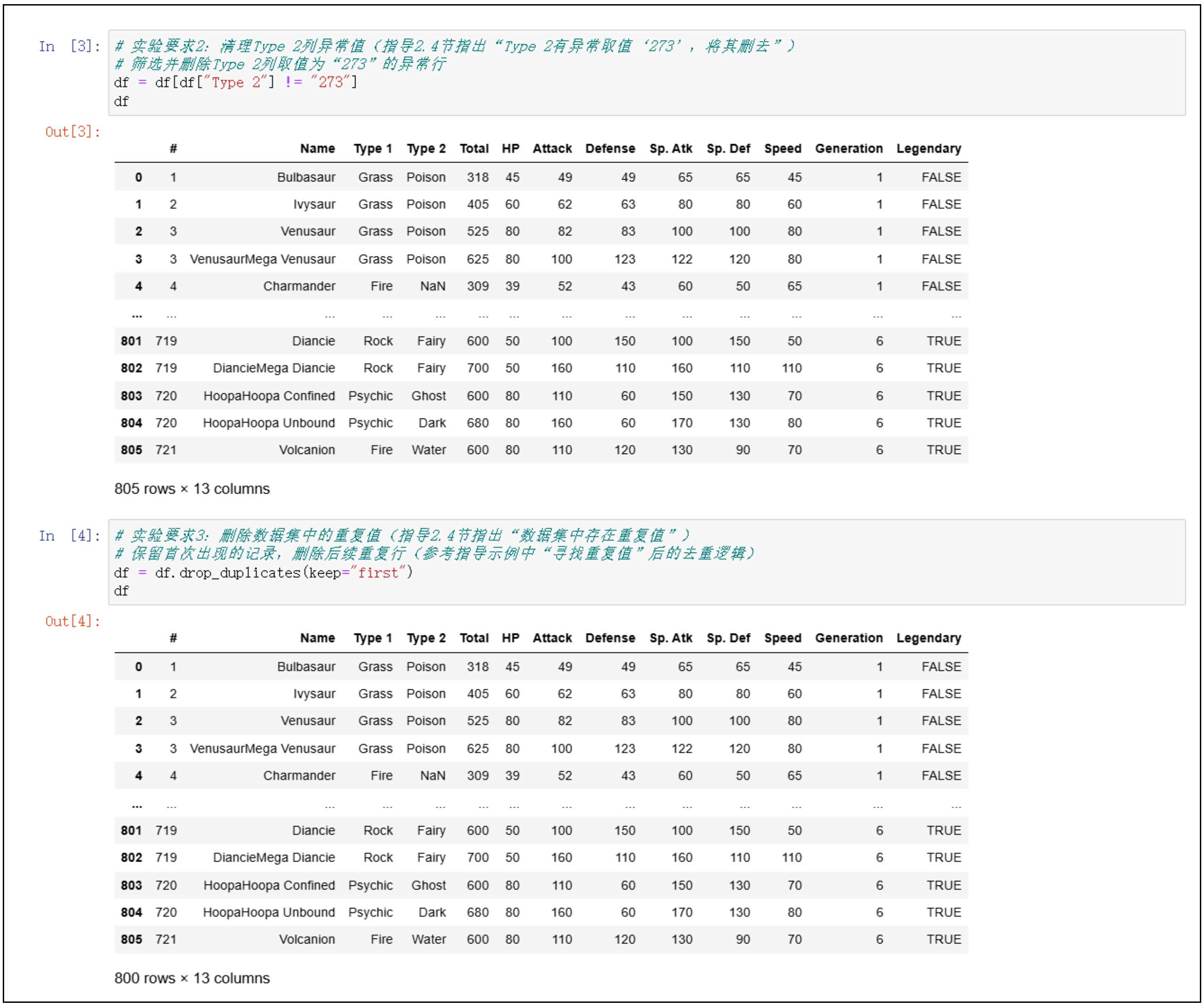


```
In  [2]: # 实验要求1：删除最后两行无意义数据
         df = df.iloc[:-4, :]  # 直接剔除末尾2行，符合“最后两行无意义，可直接删去”
         df
```

Out[2]:

	#	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



```
In [3]: # 实验要求2: 清理Type 2列异常值（指导2.4节指出“Type 2有异常取值‘273’，将其删去”）
        # 筛选并删除Type 2列取值为“273”的异常行
        df = df[df["Type 2"] != "273"]
        df
```

Out[3]:

	#	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

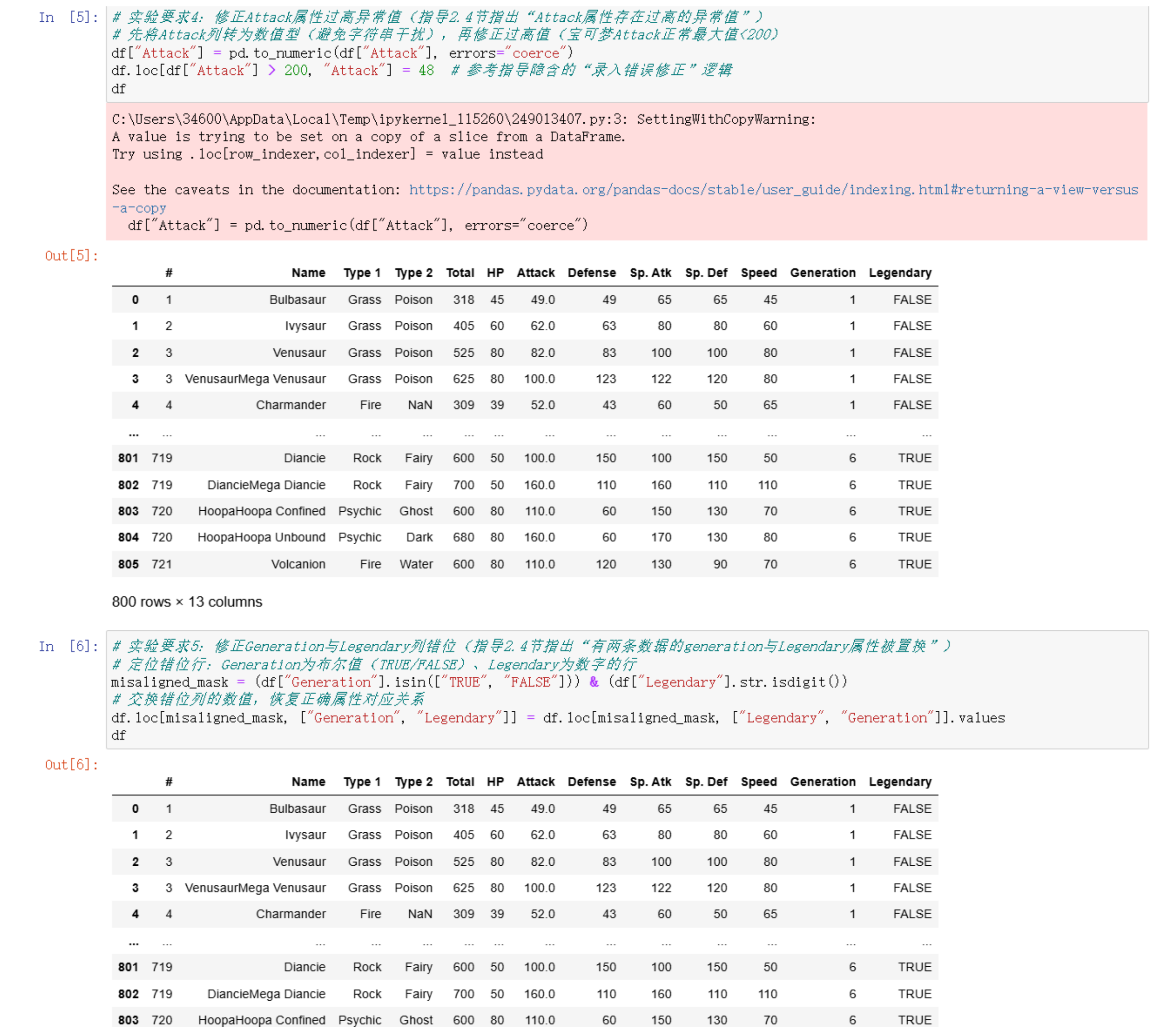
805 rows × 13 columns

```
In [4]: # 实验要求3: 删除数据集中的重复值（指导2.4节指出“数据集中存在重复值”）
        # 保留首次出现的记录，删除后续重复行（参考指导示例中“寻找重复值”后的去重逻辑）
        df = df.drop_duplicates(keep="first")
        df
```

Out[4]:

	#	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

800 rows × 13 columns



```
In  [5]: # 实验要求4: 修正Attack属性过高异常值（指导2.4节指出“Attack属性存在过高的异常值”）
         # 先将Attack列转为数值型（避免字符串干扰），再修正过高值（宝可梦Attack正常最大值<200）
         df["Attack"] = pd.to_numeric(df["Attack"], errors="coerce")
         df.loc[df["Attack"] > 200, "Attack"] = 48  # 参考指导隐含的“录入错误修正”逻辑
         df
```

```
C:\Users\34600\AppData\Local\Temp\ipykernel_115260\249013407.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus
-a-copy
  df["Attack"] = pd.to_numeric(df["Attack"], errors="coerce")
```

Out[5]:

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49.0	49	65	65	45	1	FALSE
1	2	Ivysaur	Grass	Poison	405	60	62.0	63	80	80	60	1	FALSE
2	3	Venusaur	Grass	Poison	525	80	82.0	83	100	100	80	1	FALSE
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100.0	123	122	120	80	1	FALSE
4	4	Charmander	Fire	NaN	309	39	52.0	43	60	50	65	1	FALSE
...	...	...	...	...	...	...	...	...	...	...	...	...	...
801	719	Diancie	Rock	Fairy	600	50	100.0	150	100	150	50	6	TRUE
802	719	DiancieMega Diancie	Rock	Fairy	700	50	160.0	110	160	110	110	6	TRUE
803	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110.0	60	150	130	70	6	TRUE
804	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160.0	60	170	130	80	6	TRUE
805	721	Volcanion	Fire	Water	600	80	110.0	120	130	90	70	6	TRUE

800 rows × 13 columns

```
In  [6]: # 实验要求5: 修正Generation与Legendary列错位（指导2.4节指出“有两条数据的generation与Legendary属性被置换”）
         # 定位错位行: Generation为布尔值（TRUE/FALSE）、Legendary为数字的行
         misaligned_mask = (df["Generation"].isin(["TRUE", "FALSE"])) & (df["Legendary"].str.isdigit())
         # 交换错位列的数值，恢复正确属性对应关系
         df.loc[misaligned_mask, ["Generation", "Legendary"]] = df.loc[misaligned_mask, ["Legendary", "Generation"]].values
         df
```

Out[6]:

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49.0	49	65	65	45	1	FALSE
1	2	Ivysaur	Grass	Poison	405	60	62.0	63	80	80	60	1	FALSE
2	3	Venusaur	Grass	Poison	525	80	82.0	83	100	100	80	1	FALSE
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100.0	123	122	120	80	1	FALSE
4	4	Charmander	Fire	NaN	309	39	52.0	43	60	50	65	1	FALSE
...	...	...	...	...	...	...	...	...	...	...	...	...	...
801	719	Diancie	Rock	Fairy	600	50	100.0	150	100	150	50	6	TRUE
802	719	DiancieMega Diancie	Rock	Fairy	700	50	160.0	110	160	110	110	6	TRUE
803	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110.0	60	150	130	70	6	TRUE

结论分析与体会：

本次实验修正宝可梦数据集 Attack 异常值与 Generation、Legendary 列错位问题，提升数据质量。同时掌握 pandas 实操，深刻认识数据预处理对后续分析的关键作用，为后续工作打基础。