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：202300130051	姓名： 汤冉	班级：数据班
实验题目：实验二		
实验学时： 2	实验日期：2025. 9. 16	

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

软件环境：

python3.9，jupyter notebook

实验步骤与内容：

1、导入数据并规范列名，展示部分数据，可以看到该数据集带上表头共有 810 行，13 列。

```
((810, 13),
    #                  Name Type 1  Type 2 Total  HP Attack Defense Sp. Atk
 0  1             Bulbasaur  Grass  Poison   318  45     49      49      65
 1  2               Ivysaur  Grass  Poison   405  60     62      63      80
 2  3              Venusaur  Grass  Poison   525  80     82      83     100
 3  3  VenusaurMega Venusaur  Grass  Poison   625  80    100     123     122
 4  4            Charmander   Fire     NaN   309  39     52      43      60

   Sp. Def Speed Generation Legendary
 0      65    45          1     FALSE
 1      80    60          1     FALSE
 2     100    80          1     FALSE
```

2、最后两行无意义，删去，结果如下：

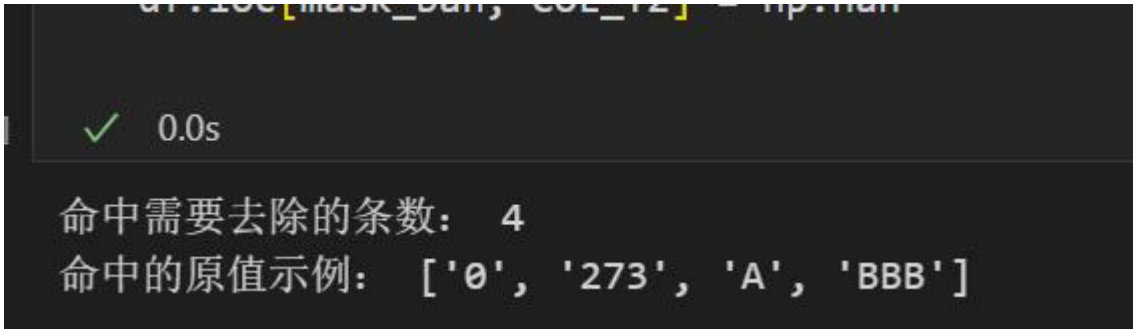
```
before = len(df)
if before >= 2:
    df = df.iloc[:-2].copy()   # 只保留到倒数第3行
else:
    df = df.iloc[0:0].copy()   # 不足2行时结果设为空表
print(f"Rows: {before} -> {len(df)}")

✓ 0.0s

Rows: 810 -> 808
```

3、对数据进行统计发现出现了这几类异常数据

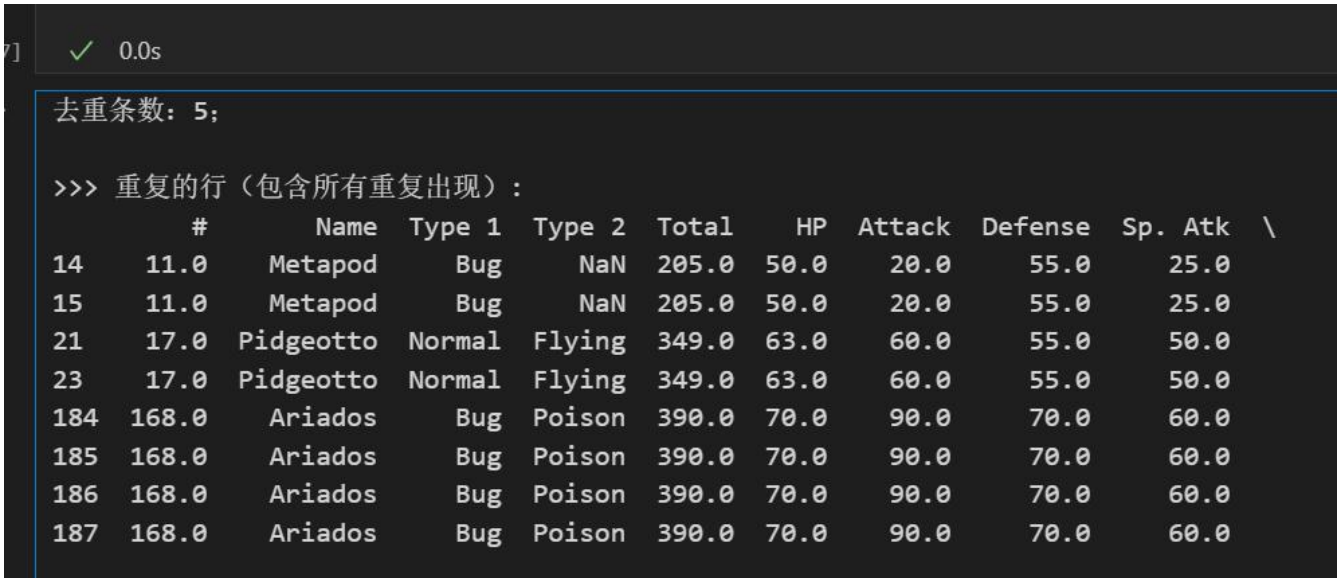
```
Type 2
[Missing]    384
Flying        98
Poison        37
Ground        35
Psychic       33
Fighting      26
Grass         25
Fairy         23
Steel         22
Dark          20
Dragon        18
Ghost         14
Water         14
Rock          14
Ice           14
Fire          12
Electric       6
Normal         4
Bug            3
273            1
0              1
A              1
BBB            1
```

可以看出 273、0、A、BBB 是异常值，进行去除。

结果如下：



4、去重

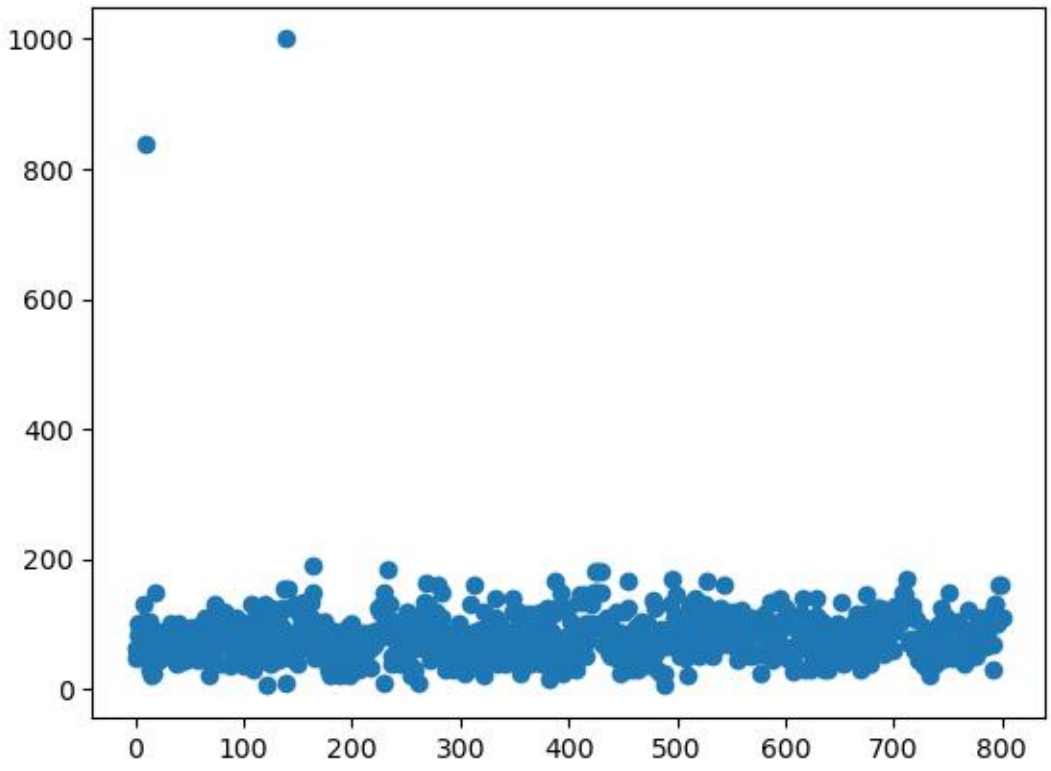


结果

```
>>> 重复的行（每组只保留一次）：
       #       Name  Type 1  Type 2  Total    HP  Attack  Defense  Sp. Atk  \
14    11.0    Metapod     Bug     NaN  205.0  50.0    20.0     55.0     25.0
21    17.0  Pidgeotto  Normal  Flying  349.0  63.0    60.0     55.0     50.0
184  168.0    Ariados     Bug  Poison  390.0  70.0    90.0     70.0     60.0

     Sp. Def  Speed Generation Legendary
14      25.0   30.0          1     FALSE
21      50.0   71.0          1     FALSE
184     60.0   40.0          2     FALSE
```

5、Attack 属性存在过高的异常值，可视化结果如下所示：

```
import matplotlib.pyplot as plt
plt.scatter(range(0,df_cleaned.shape[0]),df_cleaned.iloc[:,6])
```



可以看到有两个超过 800 的异常值，去除。

```
df_cleaned_6 = df[df['Attack'] <= 200]
print(df_cleaned_6)
```

6、有两条数据的 generation 与 Legendary 属性被置换，找到倒置的行号，再使用函数进行交换。

```
    condition = df['Generation'].apply(lambda x: isinstance(x, (int, float)))
    row_n = df.index[condition== False ].tolist()
    print(row_n)
```

```
[771]
```

```
def swap_values(df, row1, col1, col2):
    # 交换 col1 列的值
    df.loc[row1, col1], df.loc[row1, col2] = df.loc[row1, col2], df.loc[row1, col1]


# 使用函数来交换行 1 和行 2 中 'Age' 和 'City' 的值
swap_values(df, 771,  'Generation', 'Legendary')
```

结论分析与体会：

建立了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。

注：实验报告的命名规则：学号_姓名_实验 n_班级