山东大学 计算机科学与技术 学院

大数据分析与实践 课程实验报告

学号： 202300130242	姓名：王启源	班级：数据 23
实验题目：实验 2		
实验学时：2	实验日期： 2025.12.18	

实验目的：

利用 pandas 库实现多种数据采样和过滤的方法

硬件环境：

计算机一台

软件环境：

Linux 或 Windows

实验步骤与内容：

1.将数据进行一个导入

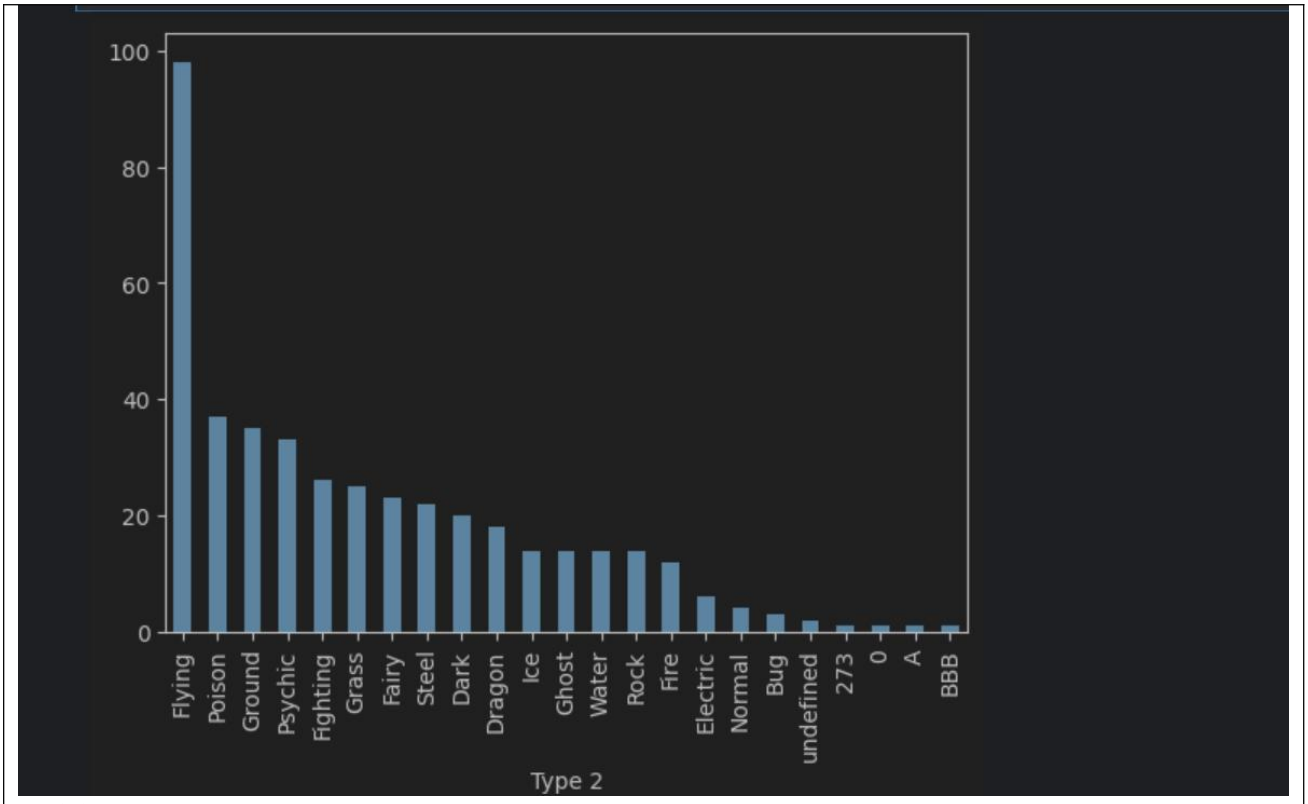
```
data = pd.read_csv("../lab2/Pokemon.csv",encoding="Windows-1252")
```

2.对数据删除多余的行和列

```
#删除多余的行和列
data.drop(columns="#", inplace=True)
data.drop(index=[808, 809], inplace=True)
```

3.检查这一列数据的各个类别的数量，发现异常值

```
#去除type2列的异常值
data['Type 2'].value_counts().plot(kind='bar')
drop_types = ['Bug', 'A', '273', '0', 'BBB']
for type in drop_types:
    index = data[data["Type 2"] == type].index
    data.drop(index=index, inplace=True)
```



4.删除重复的行

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

5.将 Generation 和 Legendary 列进行一个向前填充

```
#将Generation列和Legendary列进行一个向前填充
data["Generation"] = data["Generation"].fillna(method="ffill")
data["Legendary"] = data["Legendary"].fillna(method="ffill")
print(data)
```

6.对 Generation 列和 Legendary 列进行一个向前填充，然后打印

```
nums = ["1", "2","3","4","5","6","7","8","9"]

data = data.reset_index(drop=True)

#将generation列和Legendary列的异常值修改
for index, value in data["Generation"].items():
    if value not in nums:
        #print(index, value)
        data.at[index, "Generation"] = data.at[index-1, "Generation"]

bools = ["TRUE", "FALSE"]
for index, value in data["Legendary"].items():
    if value not in bools:
        #print(index, value)
        data.at[index, "Legendary"] = data.at[index-1, "Legendary"]
print(data)
```

```
                       Name     Type 1     Type 2      Total         HP  \
0                 Bulbasaur      Grass     Poison        318         45
1                   Ivysaur      Grass     Poison        405         60
2                  Venusaur      Grass     Poison        525         80
3    VenusaurMega Venusaur      Grass     Poison        625         80
4                Charmander       Fire        NaN        309         39
..                      ...        ...        ...        ...        ...
790    DiancieMega Diancie       Rock      Fairy        700         50
791    HoopaHoopa Confined    Psychic      Ghost        600         80
792     HoopaHoopa Unbound    Psychic       Dark        680         80
793              Volcanion       Fire      Water        600         80
794              undefined  undefined  undefined  undefined  undefined

          Attack    Defense    Sp. Atk    Sp. Def      Speed Generation  \
```

7.将每一列的数据类型进行一个修正

```
#将列的数据类型修正
to_numeric_columns = ['Total', 'HP', 'Attack', 'Sp. Atk', 'Sp. Def', 'Speed','Defense', 'Generation']
data.drop(index=794,inplace=True)
for col in to_numeric_columns:
    data[col] = pd.to_numeric(data[col], errors="coerce")
```

8.填充缺失值，将数据进行一个打印

```
#填充缺失值
data.dropna(subset=["Name"], inplace=True)
data["HP"] = data["HP"].fillna(data['HP'].mean())
data["Type 2"] = data["Type 2"].fillna(method="ffill")
print(data)
```

	Name	Type 1	Type 2	Total	HP	Attack	Defense	\
0	Bulbasaur	Grass	Poison	318.0	45.0	49.0	49.0	
1	Ivysaur	Grass	Poison	405.0	60.0	62.0	63.0	
2	Venusaur	Grass	Poison	525.0	80.0	82.0	83.0	
3	VenusaurMega Venusaur	Grass	Poison	625.0	80.0	100.0	123.0	
4	Charmander	Fire	Poison	309.0	39.0	52.0	43.0	
..	...	...	...	...	...	...	...	
789	Diancie	Rock	Fairy	600.0	50.0	100.0	150.0	
790	DiancieMega Diancie	Rock	Fairy	700.0	50.0	160.0	110.0	
791	HoopaHoopa Confined	Psychic	Ghost	600.0	80.0	110.0	60.0	
792	HoopaHoopa Unbound	Psychic	Dark	680.0	80.0	160.0	60.0	
793	Volcanion	Fire	Water	600.0	80.0	110.0	120.0	

	Sp. Atk	Sp. Def	Speed	Generation	Legendary

结论分析与体会：

熟悉掌握了使用对数据进行分析，清洗的方法。了解了如何使用 pandas 和 numpy 对数据进行一个分析和清洗