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：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： 2025.9.26	
实验目的：掌握数据清洗技术		
硬件环境： 计算机一台		
软件环境： Linux 或 Windows		
实验步骤与内容： 实验代码：		

```python
import pandas as pd
import matplotlib.pyplot as plt
from pandas import to_numeric

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

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

#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)

#去除重复的行
data.drop_duplicates(inplace=True)

#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")

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"]
```