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：数据抽样		
实验学时：2	实验日期： 2025-9-19	
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： 代码截图：		

```python
import pandas as pd
import numpy as np

data = pd.read_csv( filepath_or_buffer: "./data.csv",encoding="gbk")
data_1 = data.dropna(how="any")
print(data_1)

data_before_filter = data_1
data_after_filter1 = data_1.loc[data_before_filter["traffic"] != 0]
data_after_filter2 = data_after_filter1.loc[data_after_filter1["from_level"] == '一般节点']

print(data_after_filter2)

data_before_sample = data_after_filter2
columns = data_before_sample.columns
weight_sample = data_before_sample.copy()
weight_sample['weight'] = 0

for i in weight_sample.index:
    if weight_sample.at[i, 'to_level'] == '一般节点':
        weight = 1
    else:
        weight = 5
    weight_sample.at[i, 'weight'] = weight

weight_sample_finish = weight_sample.sample(n=50, weights='weight')
print(weight_sample_finish)
weight_sample_finish = weight_sample[columns]
print(weight_sample_finish)

random_sample = data_before_sample
random_sample_finish = random_sample.sample(n=50)
print(random_sample_finish)
random_sample_finish = random_sample_finish[columns]
print(random_sample_finish)

ybjd = data_before_sample.loc[data_before_sample['to_level'] == '一般节点']
wlhx = data_before_sample.loc[data_before_sample['to_level'] == '网络核心']

after_sample = pd.concat([ybjd.sample(17), wlhx.sample(33)])
print(after_sample)
```

结论分析与体会： 在进行数据的抽样和过滤前我们需要先判断数据是不是存在空值，将空值删除 在进行数据抽样时，有随机抽样，分层抽样，加权抽样