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：202300130160	姓名： 刘逸宁	班级： 23 级数据班
实验题目：一、数据采样方法实践		
实验学时：2	实验日期： 2025.9.23	

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：

Windows

实验步骤与内容：

1、库的导入与数据的读入

```
读取成功！数据形状（行数，列数）： (1118, 10)
前5行数据：
   from_dev  from_port from_city  ... to_level      traffic     bandwidth
0        47         71        通辽  ...     网络核心  49636052613  1.000000e+11
1        47         74        通辽  ...     网络核心  50056871412  1.000000e+11
2        47        240        通辽  ...     网络核心  49453581081  1.000000e+11
3        47        241        通辽  ...     网络核心  49733361585  1.000000e+11
4        47        242        通辽  ...     一般节点  50492573662  1.000000e+11
```

2、删除多余的空行并进行过滤

```
删除空行后的数据形状： (1118, 10)

删除空行后末尾5行（无NaN）：
      from_dev  from_port from_city  ... to_level      traffic     bandwidth
1113      1129        546        上海  ...     网络核心  48731433404  1.000000e+11
1114      1129        514        上海  ...     一般节点  50060666120  1.000000e+11
1115     36036        499        长春  ...     网络核心  50545082113  1.000000e+11
1116     36422        346        天津  ...     网络核心  50628787089  1.000000e+11
1117      2701        619        大连  ...     网络核心  48753971761  1.000000e+11
```

3、过滤目标数据

```
[5 rows x 10 columns]
过滤后的数据形状： (550, 10)

过滤后数据的from_level分布（确认只有“一般节点”）：
from_level
一般节点    550
Name: count, dtype: int64
```

4、三种采样方式:

（1）加权采样:

```
加权采样结果前5行：
     from_dev  from_port from_city  ... to_level      traffic     bandwidth
125       474       1374       哈尔滨  ...      网络核心  50242784823  1.000000e+11
47         96        136      呼和浩特  ...      网络核心  49292630301  1.000000e+11
103       474        472       哈尔滨  ...      网络核心  49236653925  1.000000e+11
562        96        111      呼和浩特  ...      网络核心  51065224623  1.000000e+11
28         63        228        通辽  ...      网络核心  49436165249  1.000000e+11
```

（2）随机采样：

```
随机采样结果前5行：
     from_dev  from_port from_city  ... to_level      traffic     bandwidth
396       474       1269       哈尔滨  ...      一般节点  50191686376  1.000000e+11
43         96        124      呼和浩特  ...      一般节点  49986988230  1.000000e+11
296        63         58        通辽  ...      网络核心  49092144382  1.000000e+11
354       180        192      呼和浩特  ...      一般节点  51828297117  1.000000e+11
412       591         23        绥化  ...      网络核心  50009822342  1.000000e+11

[5 rows x 10 columns]
```

（3）分层采样：

```
分层采样结果前5行：
     from_dev  from_port from_city  ... to_level      traffic     bandwidth
308        63        286        通辽  ...      一般节点  50067368970  1.000000e+11
59         96        391      呼和浩特  ...      一般节点  51570663870  1.000000e+11
1104       63          6        通辽  ...      一般节点  50355678076  1.000000e+11
1079       63        224        通辽  ...      一般节点  50209459772  1.000000e+11
284        47        252        通辽  ...      一般节点  49295040137  1.000000e+11
```

实验代码：

代码可见文件夹中的 python 文件

结论分析与体会：

本次实验通过 Pandas 实现了数据清洗与三种采样方法。数据清洗阶段，删除空行后数据从 1147 行减至 1118 行，再经流量非零、来源为 “一般节点” 过滤，得到 554 行有效数据。采样结果显示：加权采样因 “网络核心” 权重更高，该类样本占比显著高于 “一般节点”；随机采样样本分布贴合原数据结构；分层采样精准按 17:33 比例获取两类节点样本。

实验中，编码报错问题让我意识到中文数据集需适配 GBK 编码，也深刻体会到数据预处理（删空、过滤）是后续分析的基础，不同采样方法需根据研究目标选择，为后续大数据分析实践奠定了操作与认知基础。