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：202300130262	姓名： 何青青	班级： 23 数据
实验题目：数据采样方法实践		
实验学时：2	实验日期： 2025/9/19	

实验目标：

利用 Pandas 库实现多种数据采样和过滤的方法

实验环境：

Jupyter Notebook，python3.9

实验步骤与内容：

1、库的导入与数据的读入

如图所示，导入实验需要使用的库以及数据集，运行查看读入结果发现报错，经检查发现这个错误是因为读取 CSV 文件时编码问题导致的。UnicodeDecodeError 通常发生在文件包含非 ASCII 字符而 pandas 没有使用正确的编码来读取时，使用自动检测编码来解决该问题。

```
import pandas as pd
from pandas import DataFrame
import numpy as np

primitive_data=pd.read_csv("data.csv")
primitive_data
```

```
---------------------------------------------------------------------------
UnicodeDecodeError                        Traceback (most recent call last)
Cell In[3], line 5
      2 from pandas import DataFrame
      3 import numpy as np
----> 5 primitive_data=pd.read_csv("data.csv")
      6 primitive_data
```

读入结果如下图：

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
0	47	71	通辽	一般节点	1756	585	北京	网络核心	49636052613	1.000000e+11
1	47	74	通辽	一般节点	1756	776	北京	网络核心	50056871412	1.000000e+11
2	47	240	通辽	一般节点	1756	802	北京	网络核心	49453581081	1.000000e+11
3	47	241	通辽	一般节点	1997	464	天津	网络核心	49733361585	1.000000e+11
4	47	242	通辽	一般节点	474	672	哈尔滨	一般节点	50492573662	1.000000e+11
...	...	...	...	...	...	...	...	...	...	...
1113	1129	546	上海	网络核心	2050	502	石家庄	网络核心	48731433404	1.000000e+11
1114	1129	514	上海	网络核心	2473	946	吉林	一般节点	50060666120	1.000000e+11
1115	36036	499	长春	一般节点	1257	178	上海	网络核心	50545082113	1.000000e+11
1116	36422	346	天津	网络核心	1997	41	天津	网络核心	50628787089	1.000000e+11
1117	2701	619	大连	网络核心	2549	1070	沈阳	网络核心	48753971761	1.000000e+11

2、删除多余的空行并进行过滤

采用 dropna 方法并指定参数为 any 删除多余的空行

```
primitive_data_1=primitive_data.dropna(how='any')
primitive_data_1
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
0	47	71	通辽	一般节点	1756	585	北京	网络核心	49636052613	1.000000e+11
1	47	74	通辽	一般节点	1756	776	北京	网络核心	50056871412	1.000000e+11
2	47	240	通辽	一般节点	1756	802	北京	网络核心	49453581081	1.000000e+11
3	47	241	通辽	一般节点	1997	464	天津	网络核心	49733361585	1.000000e+11
4	47	242	通辽	一般节点	474	672	哈尔滨	一般节点	50492573662	1.000000e+11
...	...	...	...	...	...	...	...	...	...	...
1113	1129	546	上海	网络核心	2050	502	石家庄	网络核心	48731433404	1.000000e+11
1114	1129	514	上海	网络核心	2473	946	吉林	一般节点	50060666120	1.000000e+11
1115	36036	499	长春	一般节点	1257	178	上海	网络核心	50545082113	1.000000e+11
1116	36422	346	天津	网络核心	1997	41	天津	网络核心	50628787089	1.000000e+11
1117	2701	619	大连	网络核心	2549	1070	沈阳	网络核心	48753971761	1.000000e+11

1118 rows × 10 columns

接下来过滤得到 traffic 不等于 0 且 from_level=一般节点的数据

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
0	47	71	通辽	一般节点	1756	585	北京	网络核心	49636052613	1.000000e+11
1	47	74	通辽	一般节点	1756	776	北京	网络核心	50056871412	1.000000e+11
2	47	240	通辽	一般节点	1756	802	北京	网络核心	49453581081	1.000000e+11
3	47	241	通辽	一般节点	1997	464	天津	网络核心	49733361585	1.000000e+11
4	47	242	通辽	一般节点	474	672	哈尔滨	一般节点	50492573662	1.000000e+11
...	...	...	...	...	...	...	...	...	...	...
1097	2473	1460	吉林	一般节点	591	586	绥化	一般节点	48409925693	1.000000e+11
1103	36036	18	长春	一般节点	3443	650	青岛	网络核心	48663350759	1.000000e+11
1104	63	6	通辽	一般节点	36036	20	长春	一般节点	50355678076	1.000000e+11
1107	36036	52	长春	一般节点	1129	171	上海	网络核心	49345226162	1.000000e+11
1115	36036	499	长春	一般节点	1257	178	上海	网络核心	50545082113	1.000000e+11

550 rows × 10 columns

可以看到过滤得到的数据量大致缩减为原数据集的一半

3、对数据进行抽样

采取不同的采样方式采取 50 个样本并比较采样结果

①加权采样：to_level 的值为一般节点与网络核心的权重之比为 1 ： 5

在设置权重后进行加权采样，采样 50 个样本，保留展示原始列（去掉添加的 weight 列）

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
292	63	6	通辽	一般节点	2841	418	郑州	网络核心	51392218854	1.000000e+11
1005	36036	499	长春	一般节点	2050	502	石家庄	网络核心	49116324777	1.000000e+11
534	47	258	通辽	一般节点	1997	84	天津	网络核心	50060087433	1.000000e+11
411	591	19	绥化	一般节点	235	1506	北京	网络核心	50171685281	1.000000e+11
80	180	200	呼和浩特	一般节点	2701	300	大连	网络核心	51884294458	1.000000e+11
32	63	282	通辽	一般节点	36422	230	天津	网络核心	49455678350	1.000000e+11
706	2473	799	吉林	一般节点	1536	86	鄂尔多斯	网络核心	49550894885	1.000000e+11
412	591	23	绥化	一般节点	2701	71	大连	网络核心	50009822342	1.000000e+11
495	47	258	通辽	一般节点	235	1958	北京	网络核心	48574009525	1.000000e+11
11	47	259	通辽	一般节点	1756	245	北京	网络核心	50703793815	1.000000e+11

②随机抽样

随机抽取 50 个样本

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
396	474	1269	哈尔滨	一般节点	96	152	呼和浩特	一般节点	50191686376	1.000000e+11
410	591	17	绥化	一般节点	180	20	呼和浩特	一般节点	49921741386	1.000000e+11
365	180	260	呼和浩特	一般节点	1756	788	北京	网络核心	48917626581	1.000000e+11
1115	36036	499	长春	一般节点	1257	178	上海	网络核心	50545082113	1.000000e+11
131	474	1473	哈尔滨	一般节点	2549	1461	沈阳	网络核心	53304989080	1.000000e+11
83	180	210	呼和浩特	一般节点	2194	450	唐山	网络核心	50514699101	1.000000e+11
178	787	326	玉溪	一般节点	3213	597	重庆	网络核心	48608499709	1.000000e+11

③分层抽样：根据 to_level 的值进行分层采样

根据数据比例，一般节点抽 17 个，网络核心抽 33 个。得到的 50 个节点如下所示：

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
785	180	252	呼和浩特	一般节点	180	252	呼和浩特	一般节点	47786098667	1.000000e+11
980	4360	472	南京	一般节点	63	286	通辽	一般节点	49837582425	1.000000e+11
762	474	1374	哈尔滨	一般节点	180	18	呼和浩特	一般节点	48043608658	1.000000e+11
311	96	105	呼和浩特	一般节点	2473	804	吉林	一般节点	51224734473	1.000000e+11
148	591	558	绥化	一般节点	36036	499	长春	一般节点	49953028308	1.000000e+11
791	180	264	呼和浩特	一般节点	180	276	呼和浩特	一般节点	49965760241	1.000000e+11
79	180	192	呼和浩特	一般节点	591	586	绥化	一般节点	49504348509	1.000000e+11
354	180	192	呼和浩特	一般节点	4360	271	南京	一般节点	51828297117	1.000000e+11

④系统抽样

设置抽取 50 个样本，在计算固定抽样间隔后随机选择起始点进行系统抽样

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
10	47	258	通辽	一般节点	1997	122	天津	网络核心	49594312223	1.000000e+11
21	63	58	通辽	一般节点	36036	54	长春	一般节点	48363382095	1.000000e+11
32	63	282	通辽	一般节点	36422	230	天津	网络核心	49455678350	1.000000e+11
43	96	124	呼和浩特	一般节点	47	243	通辽	一般节点	49986988230	1.000000e+11
54	96	159	呼和浩特	一般节点	2360	266	太原	网络核心	51625089370	1.000000e+11
65	180	20	呼和浩特	一般节点	63	224	通辽	一般节点	50551711152	1.000000e+11
76	180	90	呼和浩特	一般节点	235	1958	北京	网络核心	50714891315	1.000000e+11
87	180	252	呼和浩特	一般节点	63	12	通辽	一般节点	49137975001	1.000000e+11
98	474	417	哈尔滨	一般节点	1997	41	天津	网络核心	51874083489	1.000000e+11
113	474	678	哈尔滨	一般节点	1997	124	天津	网络核心	49044545927	1.000000e+11

⑤整群抽样

据“to_level”字段将数据分成不同的群组，然后随机选择一个群，最后将该群内的所有样本都抽取出来作为最终样本。该方法以群为单位进行抽样，选中群内的所有个体都会被纳入样本。根据结果，有 186 个属于该“一般节点”群的样本。

```
可用的群: ['网络核心' '一般节点']
被选中的群: 一般节点
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
4	47	242	通辽	一般节点	474	672	哈尔滨	一般节点	50492573662	1.000000e+11
5	47	243	通辽	一般节点	96	124	呼和浩特	一般节点	49942713747	1.000000e+11
7	47	250	通辽	一般节点	2473	762	吉林	一般节点	49108721007	1.000000e+11
9	47	252	通辽	一般节点	96	134	呼和浩特	一般节点	50256475808	1.000000e+11
13	47	314	通辽	一般节点	96	152	呼和浩特	一般节点	50161220081	1.000000e+11
...	...	...	...	...	...	...	...	...	...	...
1057	47	243	通辽	一般节点	2473	769	吉林	一般节点	49117847542	1.000000e+11
1063	47	314	通辽	一般节点	47	252	通辽	一般节点	49900452417	1.000000e+11
1079	63	224	通辽	一般节点	4069	1196	宁波	一般节点	50209459772	1.000000e+11
1097	2473	1460	吉林	一般节点	591	586	绥化	一般节点	48409925693	1.000000e+11
1104	63	6	通辽	一般节点	36036	20	长春	一般节点	50355678076	1.000000e+11

186 rows × 10 columns

⑥比较五种方法的采样结果

通过生成对比表格来展示各方法在样本量、流量特征和节点类型分布等方面的差异，对比表格如下图所示：

	抽样方法	样本量	流量均值	流量标准差	一般节点数量	网络核心数量	一般节点比例
0	加权抽样	50	5.007086e+10	1.074005e+09	0	50	0.0
1	简单随机抽样	50	5.006416e+10	1.112589e+09	19	31	38.0
2	分层抽样	50	4.988252e+10	1.080898e+09	17	33	34.0
3	系统抽样	50	5.002733e+10	1.073238e+09	19	31	38.0
4	整群抽样	186	4.996351e+10	9.839764e+08	186	0	100.0

从结果可以看出：加权抽样由于设置了 1:5 的权重比例，只抽中了网络核心节点；整群抽样因为随机选中了“一般节点”群，样本量最大且全部为一般节点；而简单随机抽样、分层抽样和系统抽样三种方法的样本结构和流量特征相对接近，都保持了约 35-40%的一般节点比例，且流量均值和标准差较为相似，说明这些方法对总体有较好的代表性。整群抽样的流量标准差最小，显示其抽中的群内流量波动较小。

结论分析与体会：

1、数据预处理是抽样的关键前提：原始数据含空值与无效数据，经删空行、过滤后，数据量从 1118 行减至 550 行，剔除了干扰项，为后续抽样的准确性奠定基础。

2、不同抽样方法适用性差异明显：加权抽样（权重 1:5）：仅抽中“网络核心”节点，适合重点分析特定群体，但缺乏总体代表性；整群抽样：选中 “一般节点” 群（186 个样本），操作简单但样本同质性高，易有偏差；随机、分层、系统抽样：均抽 50 个样本，“一般节点”占比 35%-40%，流量均值与标准差接近总体，其中分层抽样因按比例分配样本，代表性最优，随机与系统抽样更适合快速探索分析。

3、抽样方法需匹配研究目标侧重特定群体选加权抽样，追求低成本选整群抽样，需精准反映总体结构选分层抽样，初步分析可选随机或系统抽样。

4、经过本次实验，了解到遇 UnicodeDecodeError 报错时，通过定位“编码不匹配”问题、能够使用自动检测编码解决。