山东大学计算机科学与技术学院

大数据分析与实践实验报告

学号：202300130214	姓名：于博雅	班级：23级数据班
实验题目：数据采样方法实践		
实验学时：2	实验日期：2025.9.21	

实验目标：

利用 Pandas 库实现多种数据采样和过滤的方法

实验步骤与结果：

1. 导入必需的库和数据读入，因为data.csv是ANSI格式的，所以需要使用GBK编码读取。

Out[3]:

	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

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

Out[4]:

	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

3. 对数据进行采样

- 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

（下图仅为部分数据展示）

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
288	47	314	通辽	一般节点	3213	589	重庆	网络核心	48453452651	1.000000e+11
300	63	70	通辽	一般节点	3643	831	武汉	网络核心	50635697563	1.000000e+11
889	63	12	通辽	一般节点	1997	86	天津	网络核心	49823274555	1.000000e+11
314	96	114	呼和浩特	一般节点	4561	1086	成都	网络核心	49729944227	1.000000e+11
643	474	422	哈尔滨	一般节点	2549	835	沈阳	网络核心	50003053222	1.000000e+11
48	96	141	呼和浩特	一般节点	474	422	哈尔滨	一般节点	49429192047	1.000000e+11
533	47	252	通辽	一般节点	1536	585	广州	网络核心	52135271000	1.000000e+11
449	787	316	玉溪	一般节点	36422	394	天津	网络核心	50880826227	1.000000e+11
489	47	242	通辽	一般节点	2194	406	唐山	网络核心	47826329156	1.000000e+11
628	180	52	呼和浩特	一般节点	235	1621	北京	网络核心	49227038603	1.000000e+11
691	2473	946	吉林	一般节点	1756	1117	北京	网络核心	48978564669	1.000000e+11
1047	180	52	呼和浩特	一般节点	2701	71	大连	网络核心	51851486412	1.000000e+11
593	2473	803	吉林	一般节点	3227	705	济南	网络核心	49383348895	1.000000e+11

- 随机抽样：（下图仅为部分数据展示）

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
331	96	346	呼和浩特	一般节点	1756	1128	北京	网络核心	49834736741	1.000000e+11
533	47	252	通辽	一般节点	1536	585	广州	网络核心	52135271000	1.000000e+11
898	2473	946	吉林	一般节点	2050	331	石家庄	网络核心	50778035219	1.000000e+11
16	47	427	通辽	一般节点	1997	213	天津	网络核心	48476552334	1.000000e+11
11	47	259	通辽	一般节点	1756	245	北京	网络核心	50703793815	1.000000e+11
453	787	326	玉溪	一般节点	180	20	呼和浩特	一般节点	51285240797	1.000000e+11
560	96	105	呼和浩特	一般节点	36422	446	天津	网络核心	51034130435	1.000000e+11
851	47	314	通辽	一般节点	591	1028	绥化	一般节点	50080623466	1.000000e+11
122	474	1272	哈尔滨	一般节点	2473	1043	吉林	一般节点	49735704801	1.000000e+11
0	47	71	通辽	一般节点	1756	585	北京	网络核心	49636052613	1.000000e+11

- 分层抽样：根据to_level的值进行分层采样，根据比例一般节点抽17个，网络核心抽33个（下图仅为部分数据展示）

Out[11]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
39	96	114	呼和浩特	一般节点	2473	769	吉林	一般节点	50350633304	1.000000e+11
74	180	52	呼和浩特	一般节点	63	286	通辽	一般节点	49155371449	1.000000e+11
308	63	286	通辽	一般节点	47	258	通辽	一般节点	50067368970	1.000000e+11
377	474	467	哈尔滨	一般节点	5058	70	南宁	一般节点	51745421052	1.000000e+11
329	96	159	呼和浩特	一般节点	2473	1088	吉林	一般节点	51159730271	1.000000e+11
549	63	70	通辽	一般节点	2473	1460	吉林	一般节点	49551919218	1.000000e+11
164	591	1286	绥化	一般节点	36539	1146	杭州	一般节点	50089116753	1.000000e+11
959	36036	939	长春	一般节点	47	260	通辽	一般节点	50593921106	1.000000e+11
656	4069	1196	宁波	一般节点	180	264	呼和浩特	一般节点	49766912004	1.000000e+11
638	47	243	通辽	一般节点	2473	762	吉林	一般节点	50544463355	1.000000e+11

结论分析与体会：

通过本实验，掌握了数据清洗和多种采样技术的实际应用。不同采样方法适用于不同的分析场景：随机抽样适合一般性分析，分层抽样适合保持群体比例，加权抽样适合突出重点群体。