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：202300130083	姓名：王乐临	班级：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，掌握数据预处理的基本技能，包括数据清洗、数据过滤以及五种不同的抽样方法的实现与应用。		

实验步骤与内容：

一、数据加载与初步探索

数据加载：使用 Pandas 读取本地保存的 CSV 文件，通过查看数据形状确认数据是否完整加载，同时观察数据头部和尾部内容，定位核心问题。

问题：空行占用无效行数、存在流量为 0 的无效记录、需筛选来源节点级别为一般节点的目标样本。

二、数据预处理

1. 空行删除

删除逻辑：基于 dropna(how='any')方法，设定只要某一行存在任意一列空值即删除的规则。

效果验证：删除后查看数据形状（约 1118 行），确认空行已完全清除，仅保留所有字段均有值的有效数据行。

2. 数据过滤

第一步，过滤 traffic≠0 的记录；第二步，过滤 from_level='一般节点'的记录，通过该条件筛选出目标样本群体。

结果：最终得到 550 行符合条件的有效数据，作为后续抽样的总体数据集，确保抽样对象均为有价值的样本。

三、四种抽样方法实现

1. 加权抽样

抽样原理：根据 to_level 目标节点级别赋予不同权重，因实验可能更关注网络核心节点数据，故设定权重比例一般节点：网络核心 = 1:5，让网络核心样本被选中的概率更高，贴合业务重点。

操作：先给总体数据新增权重列，循环判断每条数据的目标节点级别并赋值对应权重；再基于权重列抽取 50 个样本，确保抽样结果中网络核心节点样本占比符合预期。

2. 随机抽样

抽样原理：遵循总体中每个样本被选中概率相等的原则，快速获取无偏样本。

操作：直接从预处理后的总体数据中随机抽取 50 个样本，无需额外处理，仅需确保抽样后样本数量准确。

3. 分层抽样

抽样原理：按 to_level 将总体分为一般节点层和网络核心层，避免单一随机抽样可能导致某一层样本缺失的问题，确保每层均有代表性样本。

操作：先拆分两层数据，再按固定比例从每层抽样，最后合并两层样本得到 50 个总样本，保证两层在抽样结果中的占比合理。

4. 系统抽样

① 计算抽样间隔，总体数量 554÷样本数量 50≈11；② 从 0 到间隔值 11 之间随机生成起始

索引；③ 按起始索引+间隔×n 的等距规则选取样本，确保样本在总体中均匀分布，避免集中在某一区间。

5. 整群抽样

① 按 to_level 将总体分为两个一般节点层、网络核心层；② 随机选取所有层；③ 从选中的层中抽取 50 个样本，若群内样本数超 50 则随机抽，不足则全取。

结果：

加权抽样：

1	63	12	通辽	一般节点	2549	1570	沈阳	网络核心	5.073E+10	1E+11	加权抽样
2	63	232	通辽	一般节点	3443	186	青岛	网络核心	5.031E+10	1E+11	加权抽样
3	474	416	哈尔滨	一般节点	1257	178	上海	网络核心	5.06E+10	1E+11	加权抽样
4	180	52	呼和浩特	一般节点	235	1621	北京	网络核心	4.923E+10	1E+11	加权抽样
5	96	152	呼和浩特	一般节点	47	314	通辽	一般节点	5.198E+10	1E+11	加权抽样
6	180	30	呼和浩特	一般节点	235	1661	北京	网络核心	4.96E+10	1E+11	加权抽样
7	63	278	通辽	一般节点	235	1649	北京	网络核心	5.088E+10	1E+11	加权抽样
8	36036	20	长春	一般节点	1536	681	广州	网络核心	4.932E+10	1E+11	加权抽样
9	180	264	呼和浩特	一般节点	1129	546	上海	网络核心	5.021E+10	1E+11	加权抽样
10	47	251	通辽	一般节点	1997	124	天津	网络核心	5.116E+10	1E+11	加权抽样

随机抽样：

51	787	307	玉溪	一般节点	36422	258	天津	网络核心	5.173E+10	1E+11	随机抽样
52	180	252	呼和浩特	一般节点	1997	724	天津	网络核心	4.903E+10	1E+11	随机抽样
53	96	124	呼和浩特	一般节点	1536	1891	广州	网络核心	4.948E+10	1E+11	随机抽样
54	180	42	呼和浩特	一般节点	4360	406	南京	一般节点	5.018E+10	1E+11	随机抽样
55	180	485	呼和浩特	一般节点	1756	806	北京	网络核心	4.979E+10	1E+11	随机抽样
56	2473	803	吉林	一般节点	2194	406	唐山	网络核心	4.891E+10	1E+11	随机抽样
57	591	1112	绥化	一般节点	2360	236	太原	网络核心	5.067E+10	1E+11	随机抽样
58	180	34	呼和浩特	一般节点	2050	295	石家庄	网络核心	5.035E+10	1E+11	随机抽样
59	787	316	玉溪	一般节点	36422	394	天津	网络核心	5.088E+10	1E+11	随机抽样
60	47	243	通辽	一般节点	2473	762	吉林	一般节点	5.054E+10	1E+11	随机抽样

分层抽样：

101	63	224	通辽	一般节点	180	20	呼和浩特	一般节点	4.876E+10	1E+11	分层抽样
102	474	1410	哈尔滨	一般节点	36036	54	长春	一般节点	4.949E+10	1E+11	分层抽样
103	4069	1205	宁波	一般节点	96	114	呼和浩特	一般节点	4.941E+10	1E+11	分层抽样
104	474	473	哈尔滨	一般节点	474	1374	哈尔滨	一般节点	5.059E+10	1E+11	分层抽样
105	787	307	玉溪	一般节点	4953	686	贵阳	一般节点	4.94E+10	1E+11	分层抽样
106	63	286	通辽	一般节点	47	258	通辽	一般节点	5.007E+10	1E+11	分层抽样
107	47	243	通辽	一般节点	591	526	绥化	一般节点	4.863E+10	1E+11	分层抽样
108	63	278	通辽	一般节点	36036	18	长春	一般节点	5.048E+10	1E+11	分层抽样
109	591	23	绥化	一般节点	180	218	呼和浩特	一般节点	5.052E+10	1E+11	分层抽样
110	180	20	呼和浩特	一般节点	591	27	绥化	一般节点	4.97E+10	1E+11	分层抽样

系统抽样：

151	47	74	通辽	一般节点	1756	776	北京	网络核心	5.006E+10	1E+11	系统抽样
152	47	260	通辽	一般节点	2549	835	沈阳	网络核心	5.022E+10	1E+11	系统抽样
153	63	62	通辽	一般节点	36422	394	天津	网络核心	5.032E+10	1E+11	系统抽样
154	96	99	呼和浩特	一般节点	1257	560	上海	网络核心	4.975E+10	1E+11	系统抽样
155	96	134	呼和浩特	一般节点	47	252	通辽	一般节点	4.942E+10	1E+11	系统抽样
156	96	346	呼和浩特	一般节点	1257	138	上海	网络核心	4.776E+10	1E+11	系统抽样
157	180	28	呼和浩特	一般节点	1385	133	广州	网络核心	5.28E+10	1E+11	系统抽样
158	180	188	呼和浩特	一般节点	36422	350	天津	网络核心	4.905E+10	1E+11	系统抽样
159	180	256	呼和浩特	一般节点	1129	171	上海	网络核心	4.951E+10	1E+11	系统抽样
160	474	467	哈尔滨	一般节点	1257	174	上海	网络核心	4.999E+10	1E+11	系统抽样

整群抽样：

201	36036	54	长春	一般节点	180	256	呼和浩特	一般节点	5.192E+10	1E+11	整群抽样
202	2473	762	吉林	一般节点	1997	464	天津	网络核心	4.799E+10	1E+11	整群抽样
203	96	152	呼和浩特	一般节点	3643	559	武汉	网络核心	4.967E+10	1E+11	整群抽样
204	591	17	绥化	一般节点	180	20	呼和浩特	一般节点	4.992E+10	1E+11	整群抽样
205	474	1470	哈尔滨	一般节点	2473	1460	吉林	一般节点	4.905E+10	1E+11	整群抽样
206	47	417	通辽	一般节点	3615	191	长沙	一般节点	5.01E+10	1E+11	整群抽样
207	96	399	呼和浩特	一般节点	1756	1117	北京	网络核心	5.024E+10	1E+11	整群抽样
208	180	42	呼和浩特	一般节点	4360	406	南京	一般节点	5.018E+10	1E+11	整群抽样
209	787	54	玉溪	一般节点	474	422	哈尔滨	一般节点	5.057E+10	1E+11	整群抽样
210	47	250	通辽	一般节点	4953	686	贵阳	一般节点	5.025E+10	1E+11	整群抽样

结论分析：
加权抽样：适合需重点关注某类样本的场景，通过权重调整抽样概率；
随机抽样：适合总体分布均匀、无特殊关注点的场景，操作简便但可能存在偏差；
分层抽样：适合总体存在明显分层（如 to_level 的两类）的场景，确保每层代表性；
系统抽样：适合总体量大、需均匀抽样的场景，效率高且样本分布均匀；
整群抽样：适合群内差异大、群间差异小的场景，抽样成本低但群选择可能影响结果；