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：202300130041	姓名： 徐守政	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	

实验目的：

本实验旨在通过 Python 的 Pandas 库，系统地学习和实践多种数据采样与过滤方法，全面提升数据处理能力。首先，在数据预处理方面，实验着重培养处理原始数据中缺失值和异常值的技能，能够有效识别和清理数据质量问题，为后续分析提供干净、完整的数据基础。其次，在数据过滤技术层面，实验训练掌握基于多条件组合的数据筛选方法，能够根据具体分析需求精准提取目标数据子集。在采样方法实现方面，实验涵盖多种重要采样技术的实践，包括加权采样、随机抽样、分层抽样、系统抽样和整群抽样等，深入理解各种采样方法的原理、适用场景和实现细节。通过采样效果对比分析环节，学会评估不同采样方法的结果差异，分析各种方法对数据分布特征的影响，并能够根据具体应用场景选择最合适的采样策略。

硬件环境：

计算机一台

软件环境：

Linux 或 Windows

实验步骤与内容：

加载数据：

因为下载的数据集并非 utf-8 格式，而 pandas 默认读取 utf-8，导致加载失败，我们利用 vscode 打开文件，将文件的格式重置为 utf-8 格式，便可正常提取。

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

primitive_data = pd.read_csv("D:\Desktop\data\data1.csv")
primitive_data
```

	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

删除多余的空行并进行过滤

[10]:
```
primitive_data_1 = primitive_data.dropna(how='any')

primitive_data_1
```

[10]:

	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

过滤部分：
过滤出 traffic 不为 0 的一般节点

[12]:
```
data_before_filter = primitive_data_1

data_after_filter_1 = data_before_filter.loc[data_before_filter["traffic"] != 0]

data_after_filter_2 = data_after_filter_1.loc[data_after_filter_1["from_level"] == '一般节点']

data_after_filter_2
```

[12]:

	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

加权采样：
首先对原始数据进行清洗和过滤，然后根据 to_level 字段为每个样本分配权重（’一般节点 ’权重为 1，其他节点权重为 5），最后使用 sample()方法按权重抽取 50 个样本。这种抽样 方法确保了非一般节点有更高的被抽中概率，适用于需要重点考察特定类别数据的场景

[17]:
```
data_before_sample = data_after_filter_2
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')

weight_sample_finish = weight_sample_finish[columns]

weight_sample_finish
```

[17]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
42	96	123	呼和浩特	一般节点	2841	237	郑州	网络核心	48077485179	1.000000e+11
358	180	210	呼和浩特	一般节点	1756	642	北京	网络核心	49636949412	1.000000e+11
34	96	99	呼和浩特	一般节点	1257	560	上海	网络核心	49753614568	1.000000e+11
90	180	260	呼和浩特	一般节点	2994	486	洛阳	网络核心	48006842653	1.000000e+11
304	63	230	通辽	一般节点	3227	77	济南	网络核心	50504074996	1.000000e+11
138	591	27	绥化	一般节点	3443	117	青岛	网络核心	49213859972	1.000000e+11
68	180	30	呼和浩特	一般节点	235	1661	北京	网络核心	49596659754	1.000000e+11
164	591	1286	绥化	一般节点	36539	1146	杭州	一般节点	50089116753	1.000000e+11
73	180	50	呼和浩特	一般节点	4515	652	西安	网络核心	50640954639	1.000000e+11
148	591	558	绥化	一般节点	36036	499	长春	一般节点	49953028308	1.000000e+11
495	47	258	通辽	一般节点	235	1958	北京	网络核心	48574009525	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
114	474	682	哈尔滨	一般节点	1536	585	广州	网络核心	50262691915	1.000000e+11
342	180	28	呼和浩特	一般节点	1536	1901	广州	网络核心	50028471161	1.000000e+11
372	474	416	哈尔滨	一般节点	3227	512	济南	网络核心	49544939922	1.000000e+11

随机抽样：
直接使用 sample()方法无放回地随机抽取 50 个样本，每个样本被抽中的概率均等。

[18]:
```
random_sample=data_before_sample
random_sample_finish=random_sample.sample(n=50)
random_sample_finish=random_sample_finish[columns]
random_sample_finish
```

[18]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
1104	63	6	通辽	一般节点	36036	20	长春	一般节点	50355678076	1.000000e+11
682	63	12	通辽	一般节点	3227	103	济南	网络核心	52079990489	1.000000e+11
1086	36539	1140	杭州	一般节点	235	1661	北京	网络核心	51411580502	1.000000e+11
554	63	232	通辽	一般节点	3443	186	青岛	网络核心	50311811210	1.000000e+11
336	96	407	呼和浩特	一般节点	3227	188	济南	网络核心	50219393940	1.000000e+11
691	2473	946	吉林	一般节点	1756	1117	北京	网络核心	48978564669	1.000000e+11
94	180	485	呼和浩特	一般节点	36422	102	天津	网络核心	52460156321	1.000000e+11
434	591	1250	绥化	一般节点	3227	468	济南	网络核心	50478302588	1.000000e+11
69	180	34	呼和浩特	一般节点	3443	503	青岛	网络核心	49811891470	1.000000e+11
443	787	52	玉溪	一般节点	2360	215	太原	网络核心	49322809158	1.000000e+11
278	47	241	通辽	一般节点	4953	725	贵阳	一般节点	50008939996	1.000000e+11
177	787	325	玉溪	一般节点	4561	1087	成都	网络核心	48864832885	1.000000e+11
166	591	1300	绥化	一般节点	3443	1022	青岛	网络核心	49657631257	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
669	63	286	通辽	一般节点	3227	468	济南	网络核心	50318390185	1.000000e+11
326	96	156	呼和浩特	一般节点	4561	1031	成都	网络核心	50272713910	1.000000e+11
315	96	117	呼和浩特	一般节点	1257	581	上海	网络核心	50502305163	1.000000e+11
321	96	135	呼和浩特	一般节点	2050	553	石家庄	网络核心	51921872375	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。可以精确控制每个类别的抽样数量，确保抽样结果中各类别的比例符合预期（17:33），可以根据需要调整不同类别的抽样数量、

```
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)])

after_sample
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
387	474	677	哈尔滨	一般节点	474	672	哈尔滨	一般节点	50850714694	1.000000e+11
435	591	1258	绥化	一般节点	2473	769	吉林	一般节点	49855631239	1.000000e+11
759	3757	122	福州	一般节点	96	407	呼和浩特	一般节点	47597054356	1.000000e+11
310	96	102	呼和浩特	一般节点	474	678	哈尔滨	一般节点	49006847943	1.000000e+11
830	36036	54	长春	一般节点	591	11	绥化	一般节点	49794381448	1.000000e+11
21	63	58	通辽	一般节点	36036	54	长春	一般节点	48363382095	1.000000e+11
421	591	502	绥化	一般节点	180	264	呼和浩特	一般节点	50790049953	1.000000e+11
614	180	252	呼和浩特	一般节点	36036	52	长春	一般节点	49966571450	1.000000e+11
45	96	134	呼和浩特	一般节点	47	252	通辽	一般节点	49416652053	1.000000e+11
836	180	20	呼和浩特	一般节点	591	27	绥化	一般节点	49701796126	1.000000e+11
743	4069	1195	宁波	一般节点	96	134	呼和浩特	一般节点	50099141709	1.000000e+11
913	2473	799	吉林	一般节点	47	243	通辽	一般节点	50993016382	1.000000e+11
804	180	264	呼和浩特	一般节点	474	475	哈尔滨	一般节点	49012460413	1.000000e+11
827	474	422	哈尔滨	一般节点	474	1410	哈尔滨	一般节点	49998657939	1.000000e+11
86	180	226	呼和浩特	一般节点	36036	20	长春	一般节点	49248544673	1.000000e+11
180	787	360	玉溪	一般节点	3615	191	长沙	一般节点	49629725686	1.000000e+11
1057	47	243	通辽	一般节点	2473	769	吉林	一般节点	49117847542	1.000000e+11
276	47	74	通辽	一般节点	4561	1033	成都	网络核心	50819524115	1.000000e+11

系统抽样：

系统抽样核心思想是按照固定的间隔从总体中抽取样本。具体操作过程是：首先将总体中的 N 个单元按某种顺序排列，根据需要的样本量 n 计算出抽样间隔 k（k = N/n），然后从第一个间隔内随机选择一个起始单元，之后每隔 k 个单元抽取一个样本，直到抽满所需的 n 个样本为止。

```
# 系统抽样 - 按固定间隔抽取样本
n_samples = 50  # 需要抽取的样本数量
N = len(data_before_sample)  # 总体本量

# 计算抽样间隔
sampling_interval = N // n_samples

# 随机选择起始点（0到sampling_interval-1之间的随机数）
start_index = np.random.randint(0, sampling_interval)

# 生成抽样索引
systematic_indices = [start_index + i * sampling_interval for i in range(n_samples) if (start_index + i * sampling_interval) < N]

# 进行系统抽样
systematic_sample = data_before_sample.iloc[systematic_indices]

systematic_sample
```

	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
124	474	1311	哈尔滨	一般节点	2549	1570	沈阳	网络核心	49783212426	1.000000e+11
135	591	17	绥化	一般节点	3443	186	青岛	网络核心	49474305249	1.000000e+11

整群抽样：

整群抽样是一种将总体分成若干集群，然后随机选择部分群组，并对选中群组中的所有个体进行全面调查的抽样方法。与分层抽样不同，整群抽样中的每个群组应尽可能代表总体的多样性，而不是群组内部同质、群组间异质。

[24]:
```
clusters = data_before_sample['to_city'].unique()

# 随机选择几个群进行抽样（这里假设选择2个群）
n_clusters_to_select = 2
selected_clusters = np.random.choice(clusters, n_clusters_to_select, replace=False)
# 从选择的群中抽取所有样本
cluster_sample = data_before_sample[data_before_sample['to_city'].isin(selected_clusters)]

cluster_sample
```

[24]:

	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
5	47	243	通辽	一般节点	96	124	呼和浩特	一般节点	49942713747	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
...	...	...	...	...	...	...	...	...	...	...
1033	180	252	呼和浩特	一般节点	1756	1018	北京	网络核心	51084158075	1.000000e+11
1062	474	422	哈尔滨	一般节点	1756	1027	北京	网络核心	49590902097	1.000000e+11
1073	47	417	通辽	一般节点	1756	1029	北京	网络核心	49459363742	1.000000e+11
1075	4069	1196	宁波	一般节点	1756	1187	北京	网络核心	50488255524	1.000000e+11
1086	36539	1140	杭州	一般节点	235	1661	北京	网络核心	51411580502	1.000000e+11

结论分析与体会：

通过本次数据采样方法实践，我深刻体会到数据预处理和抽样技术在数据分析中的重要性。实验过程中，我发现不同的抽样方法会显著影响最终样本的分布特征和代表性：加权抽样能够根据业务需求调整特定群体的采样概率，分层抽样有效保持了总体分布结构，系统抽样操作简便且样本分布均匀，而整群抽样在大规模数据中展现出明显的效率优势。同时，我也认识到数据清洗的重要性，原始数据中的缺失值和异常值会直接影响抽样质量。这次实验让我明白了在实际项目中需要根据数据特征和分析目标灵活选择合适的抽样方法，既要保证样本的代表性，又要考虑操作成本和效率，为后续的统计分析和机器学习建模奠定可靠的数据基础。