山东大学计算机科学与技术学院

大数据分析与实践课程实验报告

学号：202300130100	姓名：王玺源	班级：23 级数据
实验题目：数据采样方法实践		
实验学时：2	实验日期：2025/9/20	

实验目标：

利用 Pandas 库实现多种数据采样和过滤的方法

实验内容：

先导入库并读取数据，发现空行后用 dropna(how='any')删除；

再筛选 traffic≠0 且 from_level='一般节点'的数据；

最后实现 3 种采样（加权采样：“一般节点” 与 “网络核心” 权重 1:5；随机采样；分层采样：“一般节点” 抽 17 个、“网络核心” 抽 33 个），各抽 50 个样本对比。

[11]:
```
import pandas as pd
from pandas import DataFrame
import numpy as np
primitive_data=pd.read_csv("data.csv",encoding='gbk')
primitive_data
```

[11]:

	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

[9]:
```
primitive_data_1=primitive_data.dropna(how='any')
primitive_data_1
```

[9]:

	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

[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

550 rows × 10 columns

[25]:
```
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
```

[25]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
52	96	157	呼和浩特	一般节点	2050	443	石家庄	网络核心	50096366926	1.000000e+11
121	474	1269	哈尔滨	一般节点	2549	1430	沈阳	网络核心	50312177853	1.000000e+11
37	96	108	呼和浩特	一般节点	2360	236	太原	网络核心	48210462086	1.000000e+11
382	474	614	哈尔滨	一般节点	1536	2226	广州	网络核心	51241236810	1.000000e+11
341	180	26	呼和浩特	一般节点	1756	796	北京	网络核心	48797633450	1.000000e+11
308	63	286	通辽	一般节点	47	258	通辽	一般节点	50067368970	1.000000e+11
357	180	205	呼和浩特	一般节点	2360	341	太原	网络核心	50595793729	1.000000e+11
407	591	11	绥化	一般节点	235	112	北京	网络核心	50766117914	1.000000e+11
296	63	58	通辽	一般节点	2549	922	沈阳	网络核心	49092144382	1.000000e+11
97	474	416	哈尔滨	一般节点	1257	178	上海	网络核心	50599061005	1.000000e+11
318	96	124	呼和浩特	一般节点	1536	1891	广州	网络核心	49479386359	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
337	96	460	呼和浩特	一般节点	2050	313	石家庄	网络核心	50175772267	1.000000e+11
0	47	71	通辽	一般节点	1756	585	北京	网络核心	49636052613	1.000000e+11
550	63	74	通辽	一般节点	2549	1461	沈阳	网络核心	49909937131	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
64	180	18	呼和浩特	一般节点	1536	26	鄂尔多斯	网络核心	51722488070	1.000000e+11
349	180	52	呼和浩特	一般节点	3227	449	济南	网络核心	47569937466	1.000000e+11
361	180	226	呼和浩特	一般节点	5242	763	西安	网络核心	49270522752	1.000000e+11
47	96	136	呼和浩特	一般节点	2360	215	太原	网络核心	49292630301	1.000000e+11
366	180	264	呼和浩特	一般节点	2360	195	太原	网络核心	47435896137	1.000000e+11
55	96	336	呼和浩特	一般节点	1756	1029	北京	网络核心	51600306541	1.000000e+11
31	63	278	通辽	一般节点	235	1649	北京	网络核心	50882530855	1.000000e+11
705	47	242	通辽	一般节点	63	286	通辽	一般节点	49144860439	1.000000e+11
544	63	54	通辽	一般节点	2050	336	石家庄	网络核心	51911829933	1.000000e+11
406	474	1473	哈尔滨	一般节点	36422	394	天津	网络核心	48378712039	1.000000e+11
128	474	1409	哈尔滨	一般节点	1756	1067	北京	网络核心	49473981680	1.000000e+11
850	474	422	哈尔滨	一般节点	591	638	绥化	一般节点	51214123797	1.000000e+11
111	474	673	哈尔滨	一般节点	2473	799	吉林	一般节点	48852033101	1.000000e+11
925	4360	472	南京	一般节点	1997	251	天津	网络核心	48414179107	1.000000e+11
107	474	614	哈尔滨	一般节点	3227	724	济南	网络核心	51504522549	1.000000e+11
564	96	117	呼和浩特	一般节点	2194	506	唐山	网络核心	49468205759	1.000000e+11
115	474	683	哈尔滨	一般节点	1997	84	天津	网络核心	49446798762	1.000000e+11
93	180	276	呼和浩特	一般节点	36272	235	太原	网络核心	51775514286	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
168	787	52	玉溪	一般节点	3213	246	重庆	网络核心	50468642387	1.000000e+11
280	47	243	通辽	一般节点	3213	562	重庆	网络核心	49512830312	1.000000e+11
53	96	158	呼和浩特	一般节点	2841	545	郑州	网络核心	51342500152	1.000000e+11

[21]:
```
random_sample=data_before_sample
random_sample_finish=random_sample.sample(n=50)
random_sample_finish=random_sample_finish[columns]
random_sample_finish
```

[21]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
324	96	152	呼和浩特	一般节点	3643	559	武汉	网络核心	49665987866	1.000000e+11
533	47	252	通辽	一般节点	1536	585	广州	网络核心	52135271000	1.000000e+11
293	63	10	通辽	一般节点	1756	595	北京	网络核心	49866815119	1.000000e+11
1	47	74	通辽	一般节点	1756	776	北京	网络核心	50056871412	1.000000e+11
436	591	1266	绥化	一般节点	2050	505	石家庄	网络核心	51285397493	1.000000e+11
1028	96	391	呼和浩特	一般节点	1997	122	天津	网络核心	49100896137	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
498	47	314	通辽	一般节点	591	586	绥化	一般节点	50043006782	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
83	180	210	呼和浩特	一般节点	2194	450	唐山	网络核心	50514699101	1.000000e+11
320	96	134	呼和浩特	一般节点	3643	893	武汉	网络核心	48498103572	1.000000e+11
120	474	1259	哈尔滨	一般节点	3227	787	济南	网络核心	49591440488	1.000000e+11
834	180	264	呼和浩特	一般节点	591	19	绥化	一般节点	50578150343	1.000000e+11
167	787	51	玉溪	一般节点	4561	1033	成都	网络核心	51033155364	1.000000e+11
563	96	114	呼和浩特	一般节点	2701	195	大连	网络核心	51329552752	1.000000e+11
100	474	422	哈尔滨	一般节点	96	141	呼和浩特	一般节点	48084671443	1.000000e+11
108	474	670	哈尔滨	一般节点	2841	483	郑州	网络核心	50632622266	1.000000e+11
301	63	74	通辽	一般节点	1756	469	北京	网络核心	49663523668	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
155	591	1082	绥化	一般节点	2994	430	洛阳	网络核心	49899654326	1.000000e+11
861	47	417	通辽	一般节点	591	1284	绥化	一般节点	49276967001	1.000000e+11
125	474	1374	哈尔滨	一般节点	2050	336	石家庄	网络核心	50242784823	1.000000e+11
423	591	558	绥化	一般节点	180	20	呼和浩特	一般节点	48364223310	1.000000e+11
418	591	96	绥化	一般节点	2549	852	沈阳	网络核心	50439006047	1.000000e+11

[26]:
```
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
```

[26]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
173	787	307	玉溪	一般节点	4953	686	贵阳	一般节点	49399787960	1.000000e+11
732	96	141	呼和浩特	一般节点	36036	499	长春	一般节点	47474335913	1.000000e+11
397	474	1272	哈尔滨	一般节点	96	391	呼和浩特	一般节点	48661563047	1.000000e+11
367	180	272	呼和浩特	一般节点	474	472	哈尔滨	一般节点	49398387251	1.000000e+11
1104	63	6	通辽	一般节点	36036	20	长春	一般节点	50355678076	1.000000e+11
542	63	10	通辽	一般节点	4360	472	南京	一般节点	49716409605	1.000000e+11
39	96	114	呼和浩特	一般节点	2473	769	吉林	一般节点	50350633304	1.000000e+11
760	5058	70	南宁	一般节点	96	460	呼和浩特	一般节点	49703011825	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
779	96	152	呼和浩特	一般节点	180	202	呼和浩特	一般节点	51162997127	1.000000e+11
559	96	102	呼和浩特	一般节点	36036	52	长春	一般节点	49483965391	1.000000e+11
775	96	134	呼和浩特	一般节点	180	98	呼和浩特	一般节点	51993612239	1.000000e+11
164	591	1286	绥化	一般节点	36539	1146	杭州	一般节点	50089116753	1.000000e+11
757	3615	179	长沙	一般节点	96	391	呼和浩特	一般节点	51467597716	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
7	47	250	通辽	一般节点	2473	762	吉林	一般节点	49108721007	1.000000e+11
874	36539	1140	杭州	一般节点	787	324	玉溪	一般节点	48801407907	1.000000e+11
420	591	100	绥化	一般节点	235	112	北京	网络核心	51157112955	1.000000e+11
547	63	62	通辽	一般节点	1756	1067	北京	网络核心	49632977575	1.000000e+11
78	180	188	呼和浩特	一般节点	36422	350	天津	网络核心	49047066099	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
40	96	117	呼和浩特	一般节点	2050	505	石家庄	网络核心	48814619370	1.000000e+11

总结：

数据清洗是基础：通过 dropna 删空行、loc 筛选有效数据，深刻体会到 “脏数据” 会直接影响分析结果，Pandas 的行操作函数是处理数据杂质的关键，也需注意参数精准性。

采样方法需适配场景：打破 “采样仅随机” 的误区 —— 加权采样突出高优先级数据，随机采样操作简单但代表性可能不足，分层采样保证类别均衡，让我明白需根据分析目标选对应方法。

流程思维很重要：实验串联 “数据获取 - 清洗 - 采样” 全流程，模拟真实分析前置环节，也意识到代码逻辑（如避免覆盖抽样结果）和细节（如文件编码）对实验成功的影响，提升了实操严谨性。

、