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：202300130045	姓名：张博文	班级： 数据 23
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：利用 Pandas 库实现多种数据采样和过滤的方法		
硬件环境： 计算机一台		
软件环境： python3.9，jupyter notebook		
实验步骤与内容： 1. 库的导入与数据的读入 2. 删除多余的空行并进行过滤		

In [11]:
```
import pandas as pd
from pandas import DataFrame
import numpy as np

primitive_data=pd.read_csv('data.csv', encoding='gbk')
primitive_data
```

Out[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

In [12]:
```
primitive_data_1=primitive_data.dropna(how='any')
primitive_data_1
```

Out[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
...	...	...	...	...	...	...	...	...	...	...
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

3. 对数据进行抽样

In [13]:
```
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
```

Out[13]:

	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

In [18]:
```
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')
#data_before_sample=data_before_sample[columns]
weight_sample_finish=weight_sample[columns]
weight_sample_finish
```

Out[18]:

	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
...	...	...	...	...	...	...	...	...	...	...

In [19]:
```
random_sample=data_before_sample
random_sample_finish=random_sample.sample(n=50)
random_sample_finish=random_sample_finish[columns]
random_sample_finish
```

Out[19]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
81	180	202	呼和浩特	一般节点	36272	247	太原	网络核心	49867223584	1.000000e+11
313	96	111	呼和浩特	一般节点	2360	197	太原	网络核心	49309667295	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
354	180	192	呼和浩特	一般节点	4360	271	南京	一般节点	51828297117	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
962	4448	127	无锡	一般节点	47	425	通辽	一般节点	50961073987	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
912	47	242	通辽	一般节点	47	242	通辽	一般节点	49820071586	1.000000e+11
434	591	1250	绥化	一般节点	3227	468	济南	网络核心	50478302588	1.000000e+11
530	47	249	通辽	一般节点	2473	799	吉林	一般节点	49803820036	1.000000e+11
942	36036	52	长春	一般节点	2050	272	石家庄	网络核心	49916177327	1.000000e+11
793	180	20	呼和浩特	一般节点	474	359	哈尔滨	一般节点	50601340670	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
383	474	670	哈尔滨	一般节点	5058	144	南宁	一般节点	50998204735	1.000000e+11
282	47	250	通辽	一般节点	4953	686	贵阳	一般节点	50250217535	1.000000e+11
557	63	286	通辽	一般节点	3443	117	青岛	网络核心	50247988397	1.000000e+11
893	63	70	通辽	一般节点	1997	90	天津	网络核心	49056576007	1.000000e+11
109	474	671	哈尔滨	一般节点	2549	919	沈阳	网络核心	50446722135	1.000000e+11
400	474	1374	哈尔滨	一般节点	591	23	绥化	一般节点	49461593438	1.000000e+11
545	63	58	通辽	一般节点	1756	1127	北京	网络核心	51132553467	1.000000e+11
58	96	383	呼和浩特	一般节点	5242	783	西安	网络核心	50609333179	1.000000e+11
21	63	58	通辽	一般节点	36036	54	长春	一般节点	48363382095	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
155	591	1082	绥化	一般节点	2994	430	洛阳	网络核心	49899654326	1.000000e+11
89	180	256	呼和浩特	一般节点	1129	171	上海	网络核心	49512421445	1.000000e+11
331	96	346	呼和浩特	一般节点	1756	1128	北京	网络核心	49834736741	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
307	63	282	通辽	一般节点	1756	18	北京	网络核心	49252024885	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

In [20]:
```
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
```

Out[20]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
157	591	1106	绥化	一般节点	36036	939	长春	一般节点	50954337724	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
33	63	286	通辽	一般节点	180	52	呼和浩特	一般节点	49725190236	1.000000e+11
867	63	224	通辽	一般节点	787	54	玉溪	一般节点	49892262893	1.000000e+11
173	787	307	玉溪	一般节点	4953	686	贵阳	一般节点	49399787960	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
604	96	134	呼和浩特	一般节点	2473	1460	吉林	一般节点	49201392181	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
705	47	242	通辽	一般节点	63	286	通辽	一般节点	49144860439	1.000000e+11
959	36036	939	长春	一般节点	47	260	通辽	一般节点	50593921106	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
1079	63	224	通辽	一般节点	4069	1196	宁波	一般节点	50209459772	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
404	474	1410	哈尔滨	一般节点	36036	54	长春	一般节点	49488245045	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
7	47	250	通辽	一般节点	2473	762	吉林	一般节点	49108721007	1.000000e+11
59	96	391	呼和浩特	一般节点	47	417	通辽	一般节点	51570663870	1.000000e+11
486	47	74	通辽	一般节点	1385	133	广州	网络核心	49136084036	1.000000e+11
573	36036	20	长春	一般节点	235	1621	北京	网络核心	50173217899	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
358	180	210	呼和浩特	一般节点	1756	642	北京	网络核心	49636949412	1.000000e+11
67	180	28	呼和浩特	一般节点	1385	133	广州	网络核心	52798223188	1.000000e+11
417	591	64	绥化	一般节点	1257	560	上海	网络核心	50045645266	1.000000e+11
994	63	6	通辽	一般节点	2701	135	大连	网络核心	50680536460	1.000000e+11
485	47	71	通辽	一般节点	36272	133	太原	网络核心	50529263033	1.000000e+11
306	63	278	通辽	一般节点	3227	70	济南	网络核心	51091741717	1.000000e+11
146	591	502	绥化	一般节点	1129	546	上海	网络核心	49465128399	1.000000e+11
1073	47	417	通辽	一般节点	1756	1029	北京	网络核心	49459363742	1.000000e+11

结论分析与体会：

数据预处理是采样的前提，不同采样方法结果差异显著

不清洗数据会引入无效样本，导致结果偏差。

需按分析目标选采样方法，如对比两类节点用分层采样，快速近似分析用随机采样。

Pandas 工具简化流程，提升效率且减少错误。