山东大学计算机科学与技术学院

可视化技术课程实验报告

学号：202300130220	姓名：刘傲宇	班级：数据科学与大数据技术班
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

实验步骤与内容：

导入所需库

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

数据读入

```
[8]: primitive_data=pd.read_csv("D:\\data.csv",encoding='gbk')
     primitive_data
```

[8]:

	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

采用 dropna 方法并指定参数为 any 删除多余的空行

[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

过滤得到 traffic 不等于 0 且 from_level=一般节点的数据

[10]:
```
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
```

[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
...	...	...	...	...	...	...	...	...	...	...
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

采取不同的采样方式采取 50 个样本并比较采样结果
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

[12]:
```
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
```

[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

随机抽样

[13]:
```
random_sample=data_before_sample
random_sample_finish=random_sample.sample(n=50)
random_sample_finish=random_sample_finish[columns]
random_sample_finish
```

[13]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
55	96	336	呼和浩特	一般节点	1756	1029	北京	网络核心	51600306541	1.000000e+11
799	180	52	呼和浩特	一般节点	474	460	哈尔滨	一般节点	49553070694	1.000000e+11
370	474	359	哈尔滨	一般节点	1756	594	北京	网络核心	49659526739	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
554	63	232	通辽	一般节点	3443	186	青岛	网络核心	50311811210	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
18	63	10	通辽	一般节点	235	106	北京	网络核心	52195591947	1.000000e+11
993	36036	18	长春	一般节点	2194	450	唐山	网络核心	49826827167	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
354	180	192	呼和浩特	一般节点	4360	271	南京	一般节点	51828297117	1.000000e+11
950	36036	499	长春	一般节点	2050	293	石家庄	网络核心	50524728588	1.000000e+11
135	591	17	绥化	一般节点	3443	186	青岛	网络核心	49474305249	1.000000e+11
1115	36036	499	长春	一般节点	1257	178	上海	网络核心	50545082113	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

[14]:

```
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
```

[14]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
757	3615	179	长沙	一般节点	96	391	呼和浩特	一般节点	51467597716	1.000000e+11
986	4069	1205	宁波	一般节点	96	114	呼和浩特	一般节点	49413180407	1.000000e+11
148	591	558	绥化	一般节点	36036	499	长春	一般节点	49953028308	1.000000e+11
282	47	250	通辽	一般节点	4953	686	贵阳	一般节点	50250217535	1.000000e+11
908	2473	803	吉林	一般节点	47	71	通辽	一般节点	51423663989	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
770	474	672	哈尔滨	一般节点	180	42	呼和浩特	一般节点	51263599555	1.000000e+11
555	63	278	通辽	一般节点	36036	18	长春	一般节点	50478302302	1.000000e+11
971	4953	725	贵阳	一般节点	63	66	通辽	一般节点	50167347028	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
347	180	42	呼和浩特	一般节点	4360	406	南京	一般节点	50178810628	1.000000e+11
491	47	249	通辽	一般节点	36539	1140	杭州	一般节点	50888438116	1.000000e+11
822	47	243	通辽	一般节点	474	1311	哈尔滨	一般节点	49029906488	1.000000e+11

自行实现：
系统抽样

[15]:
```
sample_size=10
population_size=len(primitive_data)
k=population_size//sample_size
start=np.random.randint(0,k)
systematic_sample=primitive_data.iloc[start::k].reset_index(drop=True)
systematic_sample
```

[15]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
0	180	192	呼和浩特	一般节点	591	586	绥化	一般节点	49504348509	1.000000e+11
1	1997	85	天津	网络核心	47	249	通辽	一般节点	49332647178	1.000000e+11
2	63	74	通辽	一般节点	1756	469	北京	网络核心	49663523668	1.000000e+11
3	591	23	绥化	一般节点	2701	71	大连	网络核心	50009822342	1.000000e+11
4	2050	289	石家庄	网络核心	2549	808	沈阳	网络核心	0	1.000000e+11
5	2473	769	吉林	一般节点	1997	464	天津	网络核心	49319842054	1.000000e+11
6	1997	724	天津	网络核心	96	136	呼和浩特	一般节点	49940892369	1.000000e+11
7	2360	266	太原	网络核心	591	1112	绥化	一般节点	0	1.000000e+11
8	2194	180	唐山	网络核心	63	54	通辽	一般节点	50082229187	1.000000e+11
9	1536	26	鄂尔多斯	网络核心	1756	1117	北京	网络核心	50810839212	1.000000e+11

整群抽样

[16]:
```
sample_size=10
groups=primitive_data['from_city'].dropna().unique()
selected_groups=[]
total=0

while total < sample_size and len(selected_groups) < len(groups):
    group = np.random.choice([g for g in groups if g not in selected_groups])
    selected_groups.append(group)
    total += primitive_data[primitive_data['from_city'] == group].shape[0]

cluster_sample = primitive_data[primitive_data['from_city'].isin(selected_groups)].reset_index(drop=True)
if len(cluster_sample) > sample_size:
    cluster_sample = cluster_sample.sample(sample_size).reset_index(drop=True)

cluster_sample
```

[16]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
0	2841	483	郑州	网络核心	36539	1140	杭州	一般节点	49473859795	1.000000e+11
1	2841	545	郑州	网络核心	591	1106	绥化	一般节点	50138554360	1.000000e+11
2	2841	341	郑州	网络核心	2701	300	大连	网络核心	50218290887	1.000000e+11
3	2841	237	郑州	网络核心	3227	468	济南	网络核心	50906574896	1.000000e+11
4	2841	237	郑州	网络核心	1997	85	天津	网络核心	50411816571	1.000000e+11
5	2841	483	郑州	网络核心	2701	47	大连	网络核心	0	1.000000e+11
6	3757	122	福州	一般节点	96	407	呼和浩特	一般节点	47597054356	1.000000e+11
7	2841	545	郑州	网络核心	47	314	通辽	一般节点	48463318976	1.000000e+11
8	2841	545	郑州	网络核心	591	23	绥化	一般节点	49780271758	1.000000e+11
9	2841	341	郑州	网络核心	47	258	通辽	一般节点	51149366439	1.000000e+11

结论分析与体会：

熟悉了多种数据采样和过滤的方法。