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：202300130046	姓名：李江涛	班级：23 数据
实验题目：实验 1		
实验学时：	实验日期：	

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：

1. 库的导入与数据读入

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

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

1118 rows × 10 columns

2. 删除多余的空行并进行过滤

```
primitive_data_1=primitive_data.dropna(how='any')
primitive_data_1
```

	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

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

```
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
```

	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

3.对数据进行抽样

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

```
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
```

	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：

[8]:
```
random_sample=data_before_sample
random_sample_finish=random_sample.sample(n=50)
random_sample_finish=random_sample_finish[columns]
random_sample_finish
```

[8]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
583	2473	946	吉林	一般节点	36539	1146	杭州	一般节点	50631070410	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
377	474	467	哈尔滨	一般节点	5058	70	南宁	一般节点	51745421052	1.000000e+11
555	63	278	通辽	一般节点	36036	18	长春	一般节点	50478302302	1.000000e+11
96	474	360	哈尔滨	一般节点	2473	946	吉林	一般节点	51819320173	1.000000e+11
295	63	54	通辽	一般节点	3227	493	济南	网络核心	49566827928	1.000000e+11
486	47	74	通辽	一般节点	1385	133	广州	网络核心	49136084036	1.000000e+11
23	63	62	通辽	一般节点	36422	394	天津	网络核心	50322780029	1.000000e+11
275	47	71	通辽	一般节点	3443	1022	青岛	网络核心	50975030653	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
1035	36036	54	长春	一般节点	591	23	绥化	一般节点	50638071722	1.000000e+11
993	36036	18	长春	一般节点	2194	450	唐山	网络核心	49826827167	1.000000e+11
562	96	111	呼和浩特	一般节点	3443	101	青岛	网络核心	51065224623	1.000000e+11
119	474	1246	哈尔滨	一般节点	3227	705	济南	网络核心	50954049859	1.000000e+11

分层抽样：

[9]:
```
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
```

[9]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
388	474	678	哈尔滨	一般节点	96	124	呼和浩特	一般节点	49289354051	1.000000e+11
962	4448	127	无锡	一般节点	47	425	通辽	一般节点	50961073987	1.000000e+11
498	47	314	通辽	一般节点	591	586	绥化	一般节点	50043006782	1.000000e+11
444	787	54	玉溪	一般节点	474	422	哈尔滨	一般节点	50571503467	1.000000e+11
339	180	18	呼和浩特	一般节点	47	241	通辽	一般节点	51793025548	1.000000e+11
812	180	52	呼和浩特	一般节点	474	682	哈尔滨	一般节点	50713290427	1.000000e+11
381	474	475	哈尔滨	一般节点	2473	941	吉林	一般节点	49402590822	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
96	474	360	哈尔滨	一般节点	2473	946	吉林	一般节点	51819320173	1.000000e+11
164	591	1286	绥化	一般节点	36539	1146	杭州	一般节点	50089116753	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
423	591	558	绥化	一般节点	180	20	呼和浩特	一般节点	48364223310	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
441	591	1300	绥化	一般节点	47	252	通辽	一般节点	50817586398	1.000000e+11
157	591	1106	绥化	一般节点	36036	939	长春	一般节点	50954337724	1.000000e+11

系统抽样：

[9]:
```
systematic_sample=data_before_sample
k=int(len(systematic_sample)/50)
random=5
sample_index=[random-1+i*k for i in range(50)]
systematic_sample_finished=systematic_sample.iloc[sample_index]
systematic_sample_finished
```

[9]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
4	47	242	通辽	一般节点	474	672	哈尔滨	一般节点	50492573662	1.000000e+11
15	47	425	通辽	一般节点	1756	1018	北京	网络核心	50796899329	1.000000e+11
26	63	74	通辽	一般节点	2701	181	大连	网络核心	50364636480	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
48	96	141	呼和浩特	一般节点	474	422	哈尔滨	一般节点	49429192047	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
70	180	36	呼和浩特	一般节点	2194	406	唐山	网络核心	50973267302	1.000000e+11
81	180	202	呼和浩特	一般节点	36272	247	太原	网络核心	49867223584	1.000000e+11
92	180	272	呼和浩特	一般节点	3443	316	青岛	网络核心	52854391127	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
118	474	1238	哈尔滨	一般节点	1756	1008	北京	网络核心	51270474683	1.000000e+11
129	474	1410	哈尔滨	一般节点	4069	1205	宁波	一般节点	46523775334	1.000000e+11
140	591	56	绥化	一般节点	36036	52	长春	一般节点	48627355195	1.000000e+11
151	591	586	绥化	一般节点	180	192	呼和浩特	一般节点	49061517661	1.000000e+11
165	591	1290	绥化	一般节点	2194	180	唐山	网络核心	49758461056	1.000000e+11

整群抽样：

```python
cluster_stats = data_before_sample.groupby("to_city").size().reset_index(name="cluster_size")#按照to_city分群
avg_cluster_size = cluster_stats["cluster_size"].mean()  # 平均每个城市的数据量
num_clusters_to_sample = int(np.ceil(50 / avg_cluster_size))  # 需抽取的城市数量（向上取整）
sampled_cities = cluster_stats.sample(n=num_clusters_to_sample, random_state=19)["to_city"].tolist()#抽取城市群
cluster_sample_initial = data_before_sample[data_before_sample["to_city"].isin(sampled_cities)]
if len(cluster_sample_initial) > 50:
    cluster_sample_finish = cluster_sample_initial.sample(n=50, random_state=19)#大于50就在这其中随机抽取
else:
    cluster_sample_finish = cluster_sample_initial  # 若不足50则直接保留
cluster_sample_finish = cluster_sample_finish[columns]
cluster_sample_finish
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
17	63	6	通辽	一般节点	591	23	绥化	一般节点	50282047691	1.000000e+11
79	180	192	呼和浩特	一般节点	591	586	绥化	一般节点	49504348509	1.000000e+11
167	787	51	玉溪	一般节点	4561	1033	成都	网络核心	51033155364	1.000000e+11
170	787	60	玉溪	一般节点	4561	1025	成都	网络核心	49992676292	1.000000e+11
175	787	317	玉溪	一般节点	5058	118	南宁	一般节点	49579743371	1.000000e+11
177	787	325	玉溪	一般节点	4561	1087	成都	网络核心	48864832885	1.000000e+11
276	47	74	通辽	一般节点	4561	1033	成都	网络核心	50819524115	1.000000e+11
279	47	242	通辽	一般节点	4561	1025	成都	网络核心	49436367939	1.000000e+11
284	47	252	通辽	一般节点	5058	118	南宁	一般节点	49295040137	1.000000e+11
286	47	259	通辽	一般节点	4561	1087	成都	网络核心	49068568496	1.000000e+11
314	96	114	呼和浩特	一般节点	4561	1086	成都	网络核心	49729944227	1.000000e+11

结果分析：

若需无偏差呈现总体特征，优先选择分层抽样（代表性最优）或系统抽样（均匀性好）；若需重点分析某类数据，选择加权抽样，刻意提升目标类别样本量；若需快速获取无偏向样本，选择随机抽样，代码最简单，无需额外参数设置；若需按区域维度分析，选择整群抽样，按 to_city 划分群，直接反映区域特征。