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：202300130113	姓名：丁正旸	班级：23 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	

实验目标：

利用 Pandas 库实现多种数据采样和过滤的方法

实验过程与内容：

数据集地址：http://storage.ameshol land.xyz/data.csv

1. 库的导入与数据的读入

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

```
[7]: primitive_data = pd.read_csv(r"D:\sdu_study\grade3-1\DA_ex\data.csv", encoding="gbk")
     primitive_data
```

[7]:

	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

本应该看到数据底部有较多的空行，但是提供的数据集似乎是处理空行后的数据集，因此没有显示出来。

2. 删除多余的空行并进行过滤

采用 dropna 方法并指定参数为 any 删除多余的空行

[8]:
```
primitive_data_1=primitive_data.dropna(how='any')
primitive_data_1
```

[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

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

[9]:
```
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
```

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

[14]:
```
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
```

[14]:

	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

[15]:
```
random_sample=data_before_sample
random_sample_finish=random_sample.sample(n=50)
random_sample_finish=random_sample_finish[columns]
random_sample_finish
```

[15]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
423	591	558	绥化	一般节点	180	20	呼和浩特	一般节点	48364223310	1.000000e+11
863	4069	1196	宁波	一般节点	591	1290	绥化	一般节点	48726638175	1.000000e+11
20	63	54	通辽	一般节点	235	100	北京	网络核心	49256234165	1.000000e+11
165	591	1290	绥化	一般节点	2194	180	唐山	网络核心	49758461056	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
681	36036	20	长春	一般节点	1536	681	广州	网络核心	49317137743	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

[16]:
```
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
```

[16]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
913	2473	799	吉林	一般节点	47	243	通辽	一般节点	50993016382	1.000000e+11
61	96	407	呼和浩特	一般节点	4069	1196	宁波	一般节点	49745162804	1.000000e+11
674	591	586	绥化	一般节点	47	243	通辽	一般节点	50565152517	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
959	36036	939	长春	一般节点	47	260	通辽	一般节点	50593921106	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

结论与体会：

熟悉了 Pandas 库的使用，学习了多种数据采样和过滤的方法，针对不同情况采用不同采样方法有了具体的了解。。