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：202300130092	姓名：王俊磊	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： 20250912	
实验目的：利用 Pandas 库实现多种数据采样和过滤的方法 数据集：http://storage.amesholland.xyz/data.csv		
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

```
import pandas as pd
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

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

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

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

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
537	47	314	通辽	一般节点	1756	1008	北京	网络核心	49136293957	1.000000e+11
836	180	20	呼和浩特	一般节点	591	27	绥化	一般节点	49701796126	1.000000e+11
116	474	1227	哈尔滨	一般节点	2841	341	郑州	网络核心	48505909225	1.000000e+11
764	2473	941	吉林	一般节点	180	26	呼和浩特	一般节点	49660872427	1.000000e+11
421	591	502	绥化	一般节点	180	264	呼和浩特	一般节点	50790049953	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
427	591	638	绥化	一般节点	235	1649	北京	网络核心	50512261162	1.000000e+11
1059	47	252	通辽	一般节点	1997	250	天津	网络核心	50358481161	1.000000e+11
412	591	23	绥化	一般节点	2701	71	大连	网络核心	50009822342	1.000000e+11
27	63	224	通辽	一般节点	180	20	呼和浩特	一般节点	48761650539	1.000000e+11
371	474	360	哈尔滨	一般节点	3227	530	济南	网络核心	49027966353	1.000000e+11
780	96	391	呼和浩特	一般节点	180	205	呼和浩特	一般节点	50103206178	1.000000e+11
442	787	51	玉溪	一般节点	2549	836	沈阳	网络核心	50594027588	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
1104	63	6	通辽	一般节点	36036	20	长春	一般节点	50355678076	1.000000e+11
168	787	52	玉溪	一般节点	3213	246	重庆	网络核心	50468642387	1.000000e+11
21	63	58	通辽	一般节点	36036	54	长春	一般节点	48363382095	1.000000e+11
1023	96	134	呼和浩特	一般节点	96	124	呼和浩特	一般节点	49523879533	1.000000e+11
309	96	99	呼和浩特	一般节点	2360	76	太原	网络核心	49047882786	1.000000e+11
391	474	1227	哈尔滨	一般节点	2549	839	沈阳	网络核心	51028242209	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

```
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
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
724	4069	1205	宁波	一般节点	36036	52	长春	一般节点	50994646887	1.000000e+11
863	4069	1196	宁波	一般节点	591	1290	绥化	一般节点	48726638175	1.000000e+11
764	2473	941	吉林	一般节点	180	26	呼和浩特	一般节点	49660872427	1.000000e+11
1063	47	314	通辽	一般节点	47	252	通辽	一般节点	49900452417	1.000000e+11
953	180	192	呼和浩特	一般节点	47	249	通辽	一般节点	50233070000	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
1097	2473	1460	吉林	一般节点	591	586	绥化	一般节点	48409925693	1.000000e+11
17	63	6	通辽	一般节点	591	23	绥化	一般节点	50282047691	1.000000e+11
91	180	264	呼和浩特	一般节点	63	70	通辽	一般节点	50106121660	1.000000e+11
638	47	243	通辽	一般节点	2473	762	吉林	一般节点	50544463355	1.000000e+11
766	5058	144	南宁	一般节点	180	30	呼和浩特	一般节点	50481413185	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
441	591	1300	绥化	一般节点	47	252	通辽	一般节点	50817586398	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
773	2473	762	吉林	一般节点	180	84	呼和浩特	一般节点	49702910101	1.000000e+11
583	2473	946	吉林	一般节点	36539	1146	杭州	一般节点	50631070410	1.000000e+11
818	2473	769	吉林	一般节点	474	1259	哈尔滨	一般节点	49274991435	1.000000e+11
497	47	260	通辽	一般节点	36422	350	天津	网络核心	49613775497	1.000000e+11
1059	47	252	通辽	一般节点	1997	250	天津	网络核心	50358481161	1.000000e+11

结论分析与体会：

1.Pandas 库的强大功能：Pandas 提供了丰富的数据操作接口，使得各种复杂的抽样方法都能轻松实现。

2.参数调优的重要性：如加权抽样中的权重设置、分层抽样中的分层比例等参数对结果影响显著。

3.业务理解的关键性：抽样方法的选择需要基于对业务背景的深入理解，不能单纯依赖技术手段。

4.抽样结果的验证：抽样后需要对样本的代表性进行验证，确保样本能够较好地反映总体特征。

5.进一步：引入统计检验方法来量化评估抽样效果，以及增加更多抽样方法等。