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：202300130153	姓名：吴宇轩	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

实验环境：

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
from pandas import DataFrame
import numpy as np

primitive_data = pd.read_csv("C:\\Users\\吴宇轩\\Desktop\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

使用 dropna 方法并指定参数为’any’删除多余的空行

```
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

使用 loc 方法过滤得到’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

使用 sample 方法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数据采样权重之比为 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')
weight_sample_finish=weight_sample_finish[columns]
weight_sample_finish
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
888	36036	20	长春	一般节点	1997	85	天津	网络核心	48987594976	1.000000e+11
356	180	202	呼和浩特	一般节点	1257	536	上海	网络核心	50231972607	1.000000e+11
306	63	278	通辽	一般节点	3227	70	济南	网络核心	51091741717	1.000000e+11
319	96	127	呼和浩特	一般节点	3213	606	重庆	网络核心	50687271651	1.000000e+11
1018	474	672	哈尔滨	一般节点	1756	585	北京	网络核心	48132652830	1.000000e+11
546	63	60	通辽	一般节点	4360	468	南京	一般节点	47970715088	1.000000e+11
549	63	70	通辽	一般节点	2473	1460	吉林	一般节点	49551919218	1.000000e+11
881	591	586	绥化	一般节点	1997	39	天津	网络核心	49268870810	1.000000e+11
139	591	29	绥化	一般节点	2701	227	大连	网络核心	49707225860	1.000000e+11

随机采样

```
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
97	474	416	哈尔滨	一般节点	1257	178	上海	网络核心	50599061005	1.000000e+11
329	96	159	呼和浩特	一般节点	2473	1088	吉林	一般节点	51159730271	1.000000e+11
397	474	1272	哈尔滨	一般节点	96	391	呼和浩特	一般节点	48661563047	1.000000e+11
389	474	682	哈尔滨	一般节点	1997	85	天津	网络核心	50053473543	1.000000e+11
181	787	418	玉溪	一般节点	4953	784	贵阳	一般节点	50699123305	1.000000e+11
695	96	141	呼和浩特	一般节点	1536	26	鄂尔多斯	网络核心	48893660398	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
537	47	314	通辽	一般节点	1756	1008	北京	网络核心	49136293957	1.000000e+11
387	474	677	哈尔滨	一般节点	474	672	哈尔滨	一般节点	50850714694	1.000000e+11

分层采样：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
836	180	20	呼和浩特	一般节点	591	27	绥化	一般节点	49701796126	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
766	5058	144	南宁	一般节点	180	30	呼和浩特	一般节点	50481413185	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
19	63	12	通辽	一般节点	180	252	呼和浩特	一般节点	49290094443	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
49	96	152	呼和浩特	一般节点	47	314	通辽	一般节点	51981076188	1.000000e+11
541	63	6	通辽	一般节点	2473	1043	吉林	一般节点	48954016072	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

结论分析与体会：

Pandas 工具为数据处理分析提供了灵活高效的方法，熟练掌握这些方法的参数逻辑是实现精准数据操作的关键。数据预处理通过剔除无效数据、筛选目标数据为后续分析提供保障。采样方法的选择需紧密结合任务目标、数据特点，不可盲目套用。