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： 202300130242	姓名：王启源	班级：数据 23
实验题目：实验 1		
实验学时：2	实验日期： 2025.12.18	

实验目的：

利用 pandas 库实现多种数据采样和过滤的方法

硬件环境：

计算机一台

软件环境：

Linux 或 Windows

实验步骤与内容：

1.将数据导入并输出数据集

```
data = pd.read_csv("../lab1/data.csv",encoding="gbk")
data_1 = data.dropna(how="any")
print(data_1)
```

```
      from_dev  from_port from_city from_level  to_dev  to_port to_city  \
0           47         71       通辽       一般节点    1756      585      北京
1           47         74       通辽       一般节点    1756      776      北京
2           47        240       通辽       一般节点    1756      802      北京
3           47        241       通辽       一般节点    1997      464      天津
4           47        242       通辽       一般节点     474      672     哈尔滨
...        ...        ...      ...        ...     ...      ...     ...
1113      1129        546       上海       网络核心    2050      502     石家庄
1114      1129        514       上海       网络核心    2473      946      吉林
1115     36036        499       长春       一般节点    1257      178      上海
1116     36422        346       天津       网络核心    1997       41      天津
1117      2701        619       大连       网络核心    2549     1070      沈阳

      to_level      traffic     bandwidth
```

2.将 traffic 不为 0，以及 from_level 为一般节点的数据取出

```
##得到traffic属性值不为0的数据，在该数据中挑选出from_level为一般节点的数据
data_before_filter = data_1
data_after_filter1 = data_1.loc[data_before_filter["traffic"] != 0]
data_after_filter2 = data_after_filter1.loc[data_after_filter1["from_level"] == '一般节点']

print(data_after_filter2)
```

```
      from_dev  from_port from_city from_level  to_dev  to_port to_city  \
0           47         71       通辽       一般节点    1756      585      北京
1           47         74       通辽       一般节点    1756      776      北京
2           47        240       通辽       一般节点    1756      802      北京
3           47        241       通辽       一般节点    1997      464      天津
4           47        242       通辽       一般节点     474      672     哈尔滨
...        ...        ...      ...        ...     ...      ...     ...
1097      2473       1460       吉林       一般节点     591      586      绥化
1103     36036         18       长春       一般节点    3443      650      青岛
1104        63          6       通辽       一般节点   36036       20      长春
1107     36036         52       长春       一般节点    1129      171      上海
1115     36036        499       长春       一般节点    1257      178      上海

      to_level      traffic     bandwidth
```

3.为数据赋予权重，一般节点赋予 weight 为 1，网络核心权重为 5

```
data_before_sample = data_after_filter2
columns = data_before_sample.columns
weight_sample = data_before_sample.copy()
weight_sample['weight'] = 0

##如果该节点的to_level为一般节点的话，weight为1，否则为5
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
```

4.按照指定的权重进行抽样，权重高的被抽到的概率大，权重小的被抽到的概率小

```
##按照weight的值抽样50个
weight_sample_finish = weight_sample.sample(n=50, weights='weight')
print(weight_sample_finish)
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city
345	180	36	呼和浩特	一般节点	2841	519	郑州
1018	474	672	哈尔滨	一般节点	1756	585	北京
405	474	1470	哈尔滨	一般节点	36422	258	天津
532	47	251	通辽	一般节点	1997	124	天津
314	96	114	呼和浩特	一般节点	4561	1086	成都
612	591	23	绥化	一般节点	3443	117	青岛
488	47	241	通辽	一般节点	3443	503	青岛
431	591	1104	绥化	一般节点	2549	852	沈阳
1063	47	314	通辽	一般节点	47	252	通辽
69	180	34	呼和浩特	一般节点	3443	503	青岛
780	96	391	呼和浩特	一般节点	180	205	呼和浩特
321	96	135	呼和浩特	一般节点	2050	553	石家庄
1053	2473	769	吉林	一般节点	2549	852	沈阳

5.

随机抽样 50 个

```
##随机抽样50个
random_sample = data_before_sample
random_sample_finish = random_sample.sample(n=50)
print(random_sample_finish)
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city
124	474	1311	哈尔滨	一般节点	2549	1570	沈阳
14	47	417	通辽	一般节点	96	391	呼和浩特
497	47	260	通辽	一般节点	36422	350	天津
383	474	670	哈尔滨	一般节点	5058	144	南宁
358	180	210	呼和浩特	一般节点	1756	642	北京
876	63	286	通辽	一般节点	787	326	玉溪
851	47	314	通辽	一般节点	591	1028	绥化
371	474	360	哈尔滨	一般节点	3227	530	济南
291	47	427	通辽	一般节点	3227	766	济南
912	47	242	通辽	一般节点	47	242	通辽
437	591	1274	绥化	一般节点	1997	250	天津
604	96	134	呼和浩特	一般节点	2473	1460	吉林
547	63	62	通辽	一般节点	1756	1067	北京

6.

分层抽样

```
##对一般节点抽样17个，网络核心抽样33个，然后拼接在一起
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)])
print(after_sample)
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	\
828	47	314	通辽	一般节点	474	1470	哈尔滨	
111	474	673	哈尔滨	一般节点	2473	799	吉林	
785	180	252	呼和浩特	一般节点	180	252	呼和浩特	
556	63	282	通辽	一般节点	63	6	通辽	
614	180	252	呼和浩特	一般节点	36036	52	长春	
1023	96	134	呼和浩特	一般节点	96	124	呼和浩特	
851	47	314	通辽	一般节点	591	1028	绥化	
822	47	243	通辽	一般节点	474	1311	哈尔滨	
980	4360	472	南京	一般节点	63	286	通辽	
997	36036	52	长春	一般节点	63	12	通辽	
396	474	1269	哈尔滨	一般节点	96	152	呼和浩特	
830	36036	54	长春	一般节点	591	11	绥化	
136	591	19	绥化	一般节点	36036	18	长春	

结论分析与体会：

掌握了对数据进行抽样的几种方法，比如随机抽样，分层抽样，以及按权重抽样