实验报告

1. 原始数据

```
primitive_data = pd.read_csv("D:/vscode/BigDataAnalysisPractice/lab1/data.csv", encoding='gbk')
print("原始数据：")
print(primitive_data.head())
```

2. 删除空行

```
primitive_data_1 = primitive_data.dropna(how='any')
print("删除空行后：")
print(primitive_data_1.head())
```

3. 过滤后数据

```
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']
print("过滤后数据：")
print(data_after_filter_2.head())
```

```
原始数据：
   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
删除空行后：
   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
过滤后数据：
   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
```

4. 加权采样

```
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]
print("加权采样结果：")
```

```
print(weight_sample_finish.head())
```

5. 随机抽样

```
random_sample_finish = data_before_sample.sample(n=50)
random_sample_finish = random_sample_finish[columns]
print("随机抽样结果：")
print(random_sample_finish.head())
```

6. 分层抽样

```
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("分层抽样结果：")
print(after_sample.head())
```

```
加权采样结果：
     from_dev  from_port from_city from_level  to_dev  to_port to_city to_level      traffic     bandwidth
170       787         60       玉溪       一般节点    4561     1025      成都     网络核心  49992676292  1.000000e+11
654        47        417       通辽       一般节点   36422      394      天津     网络核心  50110808318  1.000000e+11
888     36036         20       长春       一般节点    1997       85      天津     网络核心  48987594976  1.000000e+11
443       787         52       玉溪       一般节点    2360      215      太原     网络核心  49322809158  1.000000e+11
121       474       1269      哈尔滨       一般节点    2549     1430      沈阳     网络核心  50312177853  1.000000e+11
随机抽样结果：
     from_dev  from_port from_city from_level  to_dev  to_port to_city to_level      traffic     bandwidth
177       787        325       玉溪       一般节点    4561     1087      成都     网络核心  48864832885  1.000000e+11
2          47        240       通辽       一般节点    1756      802      北京     网络核心  49453581081  1.000000e+11
85        180        218      呼和浩特       一般节点    3443      650      青岛     网络核心  50106572586  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
448       787        307       玉溪       一般节点   36422      258      天津     网络核心  51727332383  1.000000e+11
分层抽样结果：
      from_dev  from_port from_city from_level  to_dev  to_port to_city to_level      traffic     bandwidth
762        474       1374      哈尔滨       一般节点     180       18     呼和浩特     一般节点  48043608658  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
1104        63          6       通辽       一般节点   36036       20      长春     一般节点  50355678076  1.000000e+11
966      36539       1146       杭州       一般节点      63       12      通辽     一般节点  49520418698  1.000000e+11
787      36036         54       长春       一般节点     180      256     呼和浩特     一般节点  51915256521  1.000000e+11
```