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：202300130030	姓名：赵汉哲	班级：数据 23
实验题目：		
实验学时：4	实验日期：	

实验目标：

利用 Pandas 库实现多种数据采样和过滤的方法

实验描述：

1. 导入库并读取数据、删除多余的空行

```
import pandas as pd
from pandas import DataFrame
import numpy as np

primitive_data = pd.read_csv( filepath_or_buffer: "data.csv", encoding='gbk')
primitive_data_1=primitive_data.dropna(how='any')
print(primitive_data_1)
```

```
      from_dev  from_port from_city  ... to_level      traffic     bandwidth
0           47         71        通辽  ...     网络核心  49636052613  1.000000e+11
1           47         74        通辽  ...     网络核心  50056871412  1.000000e+11
2           47        240        通辽  ...     网络核心  49453581081  1.000000e+11
3           47        241        通辽  ...     网络核心  49733361585  1.000000e+11
4           47        242        通辽  ...     一般节点  50492573662  1.000000e+11
```

2. 对数据进行过滤，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']
print(data_after_filter_2)
```

```
      from_dev  from_port from_city  ... to_level      traffic     bandwidth
0           47         71        通辽  ...     网络核心  49636052613  1.000000e+11
1           47         74        通辽  ...     网络核心  50056871412  1.000000e+11
2           47        240        通辽  ...     网络核心  49453581081  1.000000e+11
3           47        241        通辽  ...     网络核心  49733361585  1.000000e+11
4           47        242        通辽  ...     一般节点  50492573662  1.000000e+11
```

3. 对数据进行抽样

   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

```python
# 加权采样
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_finish[columns]
print(weight_sample_finish)
```

随机采样：

```python
# 随机采样
random_sample=data_before_sample
random_sample_finish=random_sample.sample(n=50)
#random_sample_finish=random_sample_finish[columns]
print(random_sample_finish)
```

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

```python
# 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(after_sample)
```

系统抽样：从 0 到 550 的数据中采样，每隔 10 个数据抽样一个数据

```python
# 系统抽样
system_sample = data_before_sample.sort_values(by="from_dev")
indexes = np.arange(0, 550, step=10)
system_sample_finish=system_sample.iloc[indexes]
print(system_sample_finish)
```

结果图片：

加权采样：

547	63	62	通辽	...	网络核心	49632977575	1.000000e+11
300	63	70	通辽	...	网络核心	50635697563	1.000000e+11
550	63	74	通辽	...	网络核心	49909937131	1.000000e+11
732	96	141	呼和浩特	...	一般节点	47474335913	1.000000e+11
417	591	64	绥化	...	网络核心	50045645266	1.000000e+11
454	787	359	玉溪	...	网络核心	51542253485	1.000000e+11
353	180	188	呼和浩特	...	网络核心	50649912010	1.000000e+11
701	2473	803	吉林	...	网络核心	48906180396	1.000000e+11
867	63	224	通辽	...	一般节点	49892262893	1.000000e+11

[50 rows x 10 columns]

随机抽样：

92	180	272	呼和浩特	...	网络核心	52854391127	1.000000e+11
1023	96	134	呼和浩特	...	一般节点	49523879533	1.000000e+11
559	96	102	呼和浩特	...	一般节点	49483965391	1.000000e+11
376	474	460	哈尔滨	...	一般节点	48394911971	1.000000e+11
23	63	62	通辽	...	网络核心	50322780029	1.000000e+11
289	47	417	通辽	...	一般节点	50099712071	1.000000e+11
962	4448	127	无锡	...	一般节点	50961073987	1.000000e+11
79	180	192	呼和浩特	...	一般节点	49504348509	1.000000e+11
298	63	62	通辽	...	一般节点	50533229665	1.000000e+11
638	47	243	通辽	...	一般节点	50544463355	1.000000e+11
593	2473	803	吉林	...	网络核心	49383348895	1.000000e+11

[50 rows x 10 columns]

分层抽样：

358	180	210	呼和浩特	...	网络核心	49636949412	1.000000e+11
1	47	74	通辽	...	网络核心	50056871412	1.000000e+11
176	787	324	玉溪	...	网络核心	48712502205	1.000000e+11
442	787	51	玉溪	...	网络核心	50594027588	1.000000e+11
399	474	1311	哈尔滨	...	网络核心	50081963602	1.000000e+11
363	180	254	呼和浩特	...	网络核心	50252917820	1.000000e+11
342	180	28	呼和浩特	...	网络核心	50028471161	1.000000e+11
557	63	286	通辽	...	网络核心	50247988397	1.000000e+11
371	474	360	哈尔滨	...	网络核心	49027966353	1.000000e+11
881	591	586	绥化	...	网络核心	49268870810	1.000000e+11
356	180	202	呼和浩特	...	网络核心	50231972607	1.000000e+11

[50 rows x 10 columns]

系统抽样：

152	591	638	绥化	...	网络核心	49178187887	1.000000e+11
133	591	13	绥化	...	网络核心	50085514419	1.000000e+11
174	787	316	玉溪	...	网络核心	51407063255	1.000000e+11
452	787	325	玉溪	...	网络核心	49891276242	1.000000e+11
181	787	418	玉溪	...	一般节点	50699123305	1.000000e+11
725	2473	1460	吉林	...	网络核心	49869730875	1.000000e+11
898	2473	946	吉林	...	网络核心	50778035219	1.000000e+11
757	3615	179	长沙	...	一般节点	51467597716	1.000000e+11
718	4360	472	南京	...	网络核心	48195505413	1.000000e+11
950	36036	499	长春	...	网络核心	50524728588	1.000000e+11
616	36036	54	长春	...	网络核心	50464551947	1.000000e+11

[55 rows x 10 columns]

、