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：202300130067	姓名：罗艺超	班级：数据班
实验题目：实验 1：数据采样方法实践		
实验学时：2 学时	实验日期：9.18	
实验目标：利用 Pandas 库实现多种数据采样和过滤的方法		

实验过程：

1. 读取数据并直接过滤掉空行，显示出头部的部分数据和尾部的部分数据

```
import pandas as pd
import chardet

# 先检测文件编码
with open('data.csv', 'rb') as f:
    result = chardet.detect(f.read())
    encoding = result['encoding']

print(f"检测到的编码: {encoding}")

# 使用检测到的编码读取文件
primitive_data = pd.read_csv( filepath_or_buffer: "data.csv", encoding=encoding)
primitive_data1=primitive_data.dropna(how='any')
print(primitive_data1)
```

```
D:\doubao\lab1\.venv\Scripts\python.exe D:\doubao\lab1\main1.py
检测到的编码: GB2312
      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
...        ...        ...       ...  ...      ...          ...           ...
1113      1129        546         上海           网络核心  48731433604  1.000000e+11
```

2. 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']
print(data_after_filter_2)
```

得到 550 行数据

```
D:\doubao\lab1\.venv\Scripts\python.exe D:\doubao\lab1\main1.py
      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
...        ...       ...       ...  ...      ...          ...           ...
1097      2473      1460        吉林  ...     一般节点  48409925693  1.000000e+11
1103     36036        18        长春  ...     网络核心  48663350759  1.000000e+11
1104        63         6        通辽  ...     一般节点  50355678076  1.000000e+11
1107     36036        52        长春  ...     网络核心  49345226162  1.000000e+11
1115     36036       499        长春  ...     网络核心  50545082113  1.000000e+11

[550 rows x 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

（1 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

看出来网络核心节点明显比一般节点多

```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')
weight_sample_finish=weight_sample_finish[columns]
#print(weight_sample_finish)
```

```
D:\doubao\lab1\.venv\Scripts\python.exe D:\doubao\lab1\main1.py
      from_dev  from_port from_city from_level  to_dev  to_port to_city to_level      traffic     bandwidth
414        591         29       绥化       一般节点     235     1649      北京     网络核心  49268934149  1.000000e+11
682         63         12       通辽       一般节点    3227      103      济南     网络核心  52079990489  1.000000e+11
160        591       1258       绥化       一般节点    4448      127      无锡     一般节点  50322958171  1.000000e+11
159        591       1250       绥化       一般节点     235     1749      北京     网络核心  49636424242  1.000000e+11
306         63        278       通辽       一般节点    3227       70      济南     网络核心  51091741717  1.000000e+11
313         96        111      呼和浩特      一般节点    2360      197      太原     网络核心  49309667295  1.000000e+11
300         63         70       通辽       一般节点    3643      831      武汉     网络核心  50635697563  1.000000e+11
1093       591        586       绥化       一般节点    1536       86     鄂尔多斯     网络核心  47929885030  1.000000e+11
307         63        282       通辽       一般节点    1756       18      北京     网络核心  49252024885  1.000000e+11
545         63         58       通辽       一般节点    1756     1127      北京     网络核心  51132553467  1.000000e+11
1028        96        391      呼和浩特      一般节点    1997      122      天津     网络核心  49100896137  1.000000e+11
530         47        249       通辽       一般节点    2473      799      吉林     一般节点  49803820036  1.000000e+11
558         96         99      呼和浩特      一般节点    2701      227      大连     网络核心  49166600948  1.000000e+11
117        474       1228       哈尔滨      一般节点    1997      468      天津     网络核心  49556923953  1.000000e+11
1005     36036        499       长春       一般节点    2050      502      石家庄     网络核心  49116324777  1.000000e+11
120        474       1259       哈尔滨      一般节点    3227      787      济南     网络核心  49591440488  1.000000e+11
21          63         58       通辽       一般节点   36036       54      长春     一般节点  48363382095  1.000000e+11
876         63        286       通辽       一般节点     787      326      玉溪     一般节点  51447111269  1.000000e+11
14          47        417       通辽       一般节点      96      391      呼和浩特     一般节点  49358372500  1.000000e+11
```

（2 随机抽样:随机抽样 50 个样本

```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)
```

```
D:\doubao\lab1\.venv\Scripts\python.exe D:\doubao\lab1\main1.py
      from_dev  from_port from_city from_level  to_dev  to_port to_city to_level      traffic     bandwidth
574         63         12       通辽       一般节点    2994      430      洛阳     网络核心  49165115912  1.000000e+11
799        180         52      呼和浩特      一般节点     474      460      哈尔滨      一般节点  49553070694  1.000000e+11
728       2473        946       吉林       一般节点    2701      195      大连     网络核心  52184126133  1.000000e+11
392        474       1228       哈尔滨      一般节点      96      134      呼和浩特      一般节点  51278220999  1.000000e+11
101        474        460       哈尔滨      一般节点    3227      344      济南     网络核心         0  1.000000e+11
171        787         61       玉溪       一般节点    3213      562      重庆     网络核心  50063136706  1.000000e+11
550         63         74       通辽       一般节点    2549     1461      沈阳     网络核心  49909937131  1.000000e+11
138        591         27       绥化       一般节点    3443      117      青岛     网络核心  49213859972  1.000000e+11
494         47        252       通辽       一般节点    1536       86     鄂尔多斯     网络核心  50478868327  1.000000e+11
130        474       1470       哈尔滨      一般节点    2473     1460      吉林      一般节点  49047884661  1.000000e+11
1035     36036         54       长春       一般节点     591       23      绥化      一般节点  50638071722  1.000000e+11
362        180        252      呼和浩特      一般节点    1997      724      天津     网络核心  49033191620  1.000000e+11
39          96        114      呼和浩特      一般节点    2473      769      吉林      一般节点  50350633304  1.000000e+11
37          96        108      呼和浩特      一般节点    2360      236      太原     网络核心  48210462086  1.000000e+11
60          96        399      呼和浩特      一般节点    1756     1117      北京     网络核心  50243694923  1.000000e+11
380        474        474       哈尔滨      一般节点    3227      701      济南     网络核心  51078772989  1.000000e+11
8           47        251       通辽       一般节点    2549      839      沈阳     网络核心  50755299504  1.000000e+11
1063        47        314       通辽       一般节点      47      252      通辽      一般节点  49900452417  1.000000e+11
178        787        326       玉溪       一般节点    3213      597      重庆     网络核心  48608499709  1.000000e+11
113        474        678       哈尔滨      一般节点    1997      124      天津     网络核心  49044545927  1.000000e+11
330         96        336      呼和浩特      一般节点    1756     1106      北京     网络核心  51277669375  1.000000e+11
324         96        152      呼和浩特      一般节点    3643      559      武汉     网络核心  49665987866  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
```

（3 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

```
#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)
```

```
D:\doubao\lab1\.venv\Scripts\python.exe D:\doubao\lab1\main1.py
     from_dev  from_port from_city from_level  to_dev  to_port to_city to_level      traffic     bandwidth
602      2473        762       吉林       一般节点    1756     1067      北京     网络核心  49937659339  1.000000e+11
454       787        359       玉溪       一般节点     235     1506      北京     网络核心  51542253485  1.000000e+11
966     36539       1146       杭州       一般节点      63       12      通辽     一般节点  49520418698  1.000000e+11
354       180        192      呼和浩特       一般节点    4360      271      南京     一般节点  51828297117  1.000000e+11
313        96        111      呼和浩特       一般节点    2360      197      太原     网络核心  49309667295  1.000000e+11
10         47        258       通辽       一般节点    1997      122      天津     网络核心  49594312223  1.000000e+11
57         96        379      呼和浩特       一般节点    1756     1187      北京     网络核心  49400869697  1.000000e+11
26         63         74       通辽       一般节点    2701      181      大连     网络核心  50364636480  1.000000e+11
350       180         84      呼和浩特       一般节点    1756      583      北京     网络核心  51561391217  1.000000e+11
622       180         20      呼和浩特       一般节点   36036      499      长春     一般节点  49636788433  1.000000e+11
743      4069       1195       宁波       一般节点      96      134     呼和浩特     一般节点  50099141709  1.000000e+11
128       474       1409      哈尔滨       一般节点    1756     1067      北京     网络核心  49473981680  1.000000e+11
153       591       1028       绥化       一般节点   36422      268      天津     网络核心           0  1.000000e+11
564        96        117      呼和浩特       一般节点    2194      506      唐山     网络核心  49468205759  1.000000e+11
125       474       1374      哈尔滨       一般节点    2050      336     石家庄     网络核心  50242784823  1.000000e+11
328        96        158      呼和浩特       一般节点      47      427      通辽     一般节点  49385366171  1.000000e+11
14         47        417       通辽       一般节点      96      391     呼和浩特     一般节点  49358372500  1.000000e+11
22         63         60       通辽       一般节点   36422      258      天津     网络核心  49920786706  1.000000e+11
87        180        252      呼和浩特       一般节点      63       12      通辽     一般节点  49137975001  1.000000e+11
```

（4 系统采样

采样出 50 个样本，并重新定义序号

```
#系统抽样，抽50个
n=50
N=len(data_before_sample)
k=N // n #整除返回整数
start=random.randint( a: 0,k-1)
indices=[start +i*k for i in range(n) if (start+i*k)<N]
system_sample=data_before_sample.iloc[indices].reset_index(drop=True)
#print(system_sample)
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
0	47	258	通辽	一般节点	1997	122	天津	网络核心	49594312223	1.000000e+11
1	63	58	通辽	一般节点	36036	54	长春	一般节点	48363382095	1.000000e+11
2	63	282	通辽	一般节点	36422	230	天津	网络核心	49455678350	1.000000e+11
3	96	124	呼和浩特	一般节点	47	243	通辽	一般节点	49986988230	1.000000e+11
4	96	159	呼和浩特	一般节点	2360	266	太原	网络核心	51625089370	1.000000e+11
5	180	20	呼和浩特	一般节点	63	224	通辽	一般节点	50551711152	1.000000e+11
6	180	90	呼和浩特	一般节点	235	1958	北京	网络核心	50714891315	1.000000e+11
7	180	252	呼和浩特	一般节点	63	12	通辽	一般节点	49137975001	1.000000e+11
8	474	417	哈尔滨	一般节点	1997	41	天津	网络核心	51874083489	1.000000e+11
9	474	678	哈尔滨	一般节点	1997	124	天津	网络核心	49044545927	1.000000e+11
10	474	1311	哈尔滨	一般节点	2549	1570	沈阳	网络核心	49783212426	1.000000e+11
11	591	17	绥化	一般节点	3443	186	青岛	网络核心	49474305249	1.000000e+11
12	591	502	绥化	一般节点	1129	546	上海	网络核心	49465128399	1.000000e+11
13	591	1258	绥化	一般节点	4448	127	无锡	一般节点	50322958171	1.000000e+11
14	787	61	玉溪	一般节点	3213	562	重庆	网络核心	50063136706	1.000000e+11
15	47	71	通辽	一般节点	3443	1022	青岛	网络核心	50975030653	1.000000e+11
16	47	259	通辽	一般节点	4561	1087	成都	网络核心	49068568496	1.000000e+11
17	63	62	通辽	一般节点	474	683	哈尔滨	一般节点	50533229665	1.000000e+11
18	96	99	呼和浩特	一般节点	2360	76	太原	网络核心	49047882786	1.000000e+11
19	96	135	呼和浩特	一般节点	2050	553	石家庄	网络核心	51921872375	1.000000e+11
20	96	379	呼和浩特	一般节点	1257	519	上海	网络核心	52892882038	1.000000e+11
21	180	30	呼和浩特	一般节点	474	1228	哈尔滨	一般节点	50284244775	1.000000e+11
22	180	192	呼和浩特	一般节点	4360	271	南京	一般节点	51828297117	1.000000e+11
23	180	260	呼和浩特	一般节点	1756	788	北京	网络核心	48917626581	1.000000e+11
24	474	460	哈尔滨	一般节点	3757	122	福州	一般节点	48394911971	1.000000e+11
25	474	677	哈尔滨	一般节点	474	672	哈尔滨	一般节点	50850714694	1.000000e+11
26	474	1311	哈尔滨	一般节点	1997	213	天津	网络核心	50081963402	1.000000e+11

(5 整群抽样:抽取“to_level”值为一般节点的所有元素

```python
#整群抽样
unique_clusters=data_before_sample['to_level'].unique()#获取to_level的所有值
select_clusters=random.sample(list(unique_clusters), k: 1)#随机获取一个to_level的群对应的值
print(select_clusters)
cluster_sample=data_before_sample[data_before_sample['to_level'].isin(select_clusters)].reset_index(drop=True)
print(cluster_sample)
```

```
D:\doubao\lab1\.venv\Scripts\python.exe D:\doubao\lab1\main1.py
['一般节点']
     from_dev  from_port from_city  ... to_level      traffic     bandwidth
0          47        242        通辽  ...     一般节点  50492573662  1.000000e+11
1          47        243        通辽  ...     一般节点  49942713747  1.000000e+11
2          47        250        通辽  ...     一般节点  49108721007  1.000000e+11
3          47        252        通辽  ...     一般节点  50256475808  1.000000e+11
4          47        314        通辽  ...     一般节点  50161220081  1.000000e+11
..        ...        ...       ...  ...      ...          ...           ...
181        47        243        通辽  ...     一般节点  49117847542  1.000000e+11
182        47        314        通辽  ...     一般节点  49900452417  1.000000e+11
183        63        224        通辽  ...     一般节点  50209459772  1.000000e+11
184      2473       1460        吉林  ...     一般节点  48409925693  1.000000e+11
185        63          6        通辽  ...     一般节点  50355678076  1.000000e+11
```

结论分析：
通过本次数据采样方法实践，我深入理解和掌握了多种数据采样技术的原理与实现方法。
实验过程中，我收获颇丰：
首先，在数据预处理阶段，我学会了如何检测文件编码并正确读取数据，以及使用 dropna() 和条件过滤等方法清洗数据，这为后续的采样分析奠定了良好的数据基础。
在采样方法实现方面：

1. **加权采样**让我理解了如何根据业务需求赋予不同样本不同的权重，从而在抽样中体现重要性的差异；
2. **随机抽样**作为基础采样方法，保证了样本的随机性和代表性；
3. **分层抽样**通过按"to_level"分层并保持原比例，确保了样本结构的完整性；
4. **系统抽样**通过等距抽样提供了均匀分布的样本；
5. **整群抽样**则展示了按群体特征进行抽样的实际应用。

通过比较不同采样方法的结果，我深刻认识到每种方法都有其适用的场景：加权采样适合重要性不同的数据，分层抽样适合保持总体结构，系统抽样适合大规模数据的均匀采样，整群抽样则适合群体特征明显的数据。
实验过程中，我也遇到了一些挑战，如编码检测、权重设置逻辑的实现等，通过查阅文档和调试代码，最终成功解决了这些问题。这次实践不仅提升了我的编程能力，更重要的是培养了我对数据采样方法选择和数据质量控制的专业思维。
总之，本次实验让我对数据采样有了更全面的认识，为后续的大数据分析工作打下了坚实的基础。

、