

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：202300130063	姓名：邱珺	班级：23 数据																																																																																																																																				
实验题目：实验一 数据采样方法实践																																																																																																																																						
实验学时：2	实验日期：2025/9/8																																																																																																																																					
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、库的导入与数据的读入																																																																																																																																						
<div><pre>[4]: import pandas as pd from pandas import DataFrame import numpy as np primitive_data = pd.read_csv("./data.csv") primitive_data</pre></div> <div><table><thead><tr><th></th><th>from_dev</th><th>from_port</th><th>from_city</th><th>from_level</th><th>to_dev</th><th>to_port</th><th>to_city</th><th>to_level</th><th>traffic</th><th>bandwidth</th></tr></thead><tbody><tr><td>0</td><td>47</td><td>71</td><td>通辽</td><td>一般节点</td><td>1756</td><td>585</td><td>北京</td><td>网络核心</td><td>49636052613</td><td>1.000000e+11</td></tr><tr><td>1</td><td>47</td><td>74</td><td>通辽</td><td>一般节点</td><td>1756</td><td>776</td><td>北京</td><td>网络核心</td><td>50056871412</td><td>1.000000e+11</td></tr><tr><td>2</td><td>47</td><td>240</td><td>通辽</td><td>一般节点</td><td>1756</td><td>802</td><td>北京</td><td>网络核心</td><td>49453581081</td><td>1.000000e+11</td></tr><tr><td>3</td><td>47</td><td>241</td><td>通辽</td><td>一般节点</td><td>1997</td><td>464</td><td>天津</td><td>网络核心</td><td>49733361585</td><td>1.000000e+11</td></tr><tr><td>4</td><td>47</td><td>242</td><td>通辽</td><td>一般节点</td><td>474</td><td>672</td><td>哈尔滨</td><td>一般节点</td><td>50492573662</td><td>1.000000e+11</td></tr><tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr><tr><td>1113</td><td>1129</td><td>546</td><td>上海</td><td>网络核心</td><td>2050</td><td>502</td><td>石家庄</td><td>网络核心</td><td>48731433404</td><td>1.000000e+11</td></tr><tr><td>1114</td><td>1129</td><td>514</td><td>上海</td><td>网络核心</td><td>2473</td><td>946</td><td>吉林</td><td>一般节点</td><td>50060666120</td><td>1.000000e+11</td></tr><tr><td>1115</td><td>36036</td><td>499</td><td>长春</td><td>一般节点</td><td>1257</td><td>178</td><td>上海</td><td>网络核心</td><td>50545082113</td><td>1.000000e+11</td></tr><tr><td>1116</td><td>36422</td><td>346</td><td>天津</td><td>网络核心</td><td>1997</td><td>41</td><td>天津</td><td>网络核心</td><td>50628787089</td><td>1.000000e+11</td></tr><tr><td>1117</td><td>2701</td><td>619</td><td>大连</td><td>网络核心</td><td>2549</td><td>1070</td><td>沈阳</td><td>网络核心</td><td>48753971761</td><td>1.000000e+11</td></tr></tbody></table><div>1118 rows × 10 columns</div></div>				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
	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																																																																																																																												
2、删除多余的空行并进行过滤																																																																																																																																						

```
[5]: primitive_data_1=primitive_data.dropna(how='any')
primitive_data_1
```

```
[5]:
```

	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

3、对数据进行抽样

1) 加权抽样

```
[6]: 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
```

```
[6]:
```

	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

2) 随机抽样

```
[11]: 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
444	787	54	玉溪	一般节点	474	422	哈尔滨	一般节点	50571503467	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
1057	47	243	通辽	一般节点	2473	769	吉林	一般节点	49117847542	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
706	2473	799	吉林	一般节点	1536	86	鄂尔多斯	网络核心	49550894885	1.000000e+11
953	180	192	呼和浩特	一般节点	47	249	通辽	一般节点	50233070000	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
1035	36036	54	长春	一般节点	591	23	绥化	一般节点	50638071722	1.000000e+11
416	591	60	绥化	一般节点	180	52	呼和浩特	一般节点	50126205393	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
1079	63	224	通辽	一般节点	4069	1196	宁波	一般节点	50209459772	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
323	96	141	呼和浩特	一般节点	2050	391	石家庄	网络核心	49814111100	1.000000e+11
145	591	100	绥化	一般节点	2194	506	唐山	网络核心	51437026945	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
317	96	123	呼和浩特	一般节点	2360	263	太原	网络核心	50500915133	1.000000e+11
674	591	586	绥化	一般节点	47	243	通辽	一般节点	50565152517	1.000000e+11
322	96	136	呼和浩特	一般节点	3227	389	济南	网络核心	50541979348	1.000000e+11
1075	4069	1196	宁波	一般节点	1756	1187	北京	网络核心	50488255524	1.000000e+11

3) 分层抽样

```
[12]: 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
491	47	249	通辽	一般节点	36539	1140	杭州	一般节点	50888438116	1.000000e+11
750	474	472	哈尔滨	一般节点	96	157	呼和浩特	一般节点	50784665266	1.000000e+11
556	63	282	通辽	一般节点	63	6	通辽	一般节点	49489299594	1.000000e+11
110	474	672	哈尔滨	一般节点	47	242	通辽	一般节点	51555817613	1.000000e+11
827	474	422	哈尔滨	一般节点	474	1410	哈尔滨	一般节点	49998657939	1.000000e+11
656	4069	1196	宁波	一般节点	180	264	呼和浩特	一般节点	49766912004	1.000000e+11
5	47	243	通辽	一般节点	96	124	呼和浩特	一般节点	49942713747	1.000000e+11
1097	2473	1460	吉林	一般节点	591	586	绥化	一般节点	48409925693	1.000000e+11
347	180	42	呼和浩特	一般节点	4360	406	南京	一般节点	50178810628	1.000000e+11
546	63	60	通辽	一般节点	4360	468	南京	一般节点	47970715088	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
804	180	264	呼和浩特	一般节点	474	475	哈尔滨	一般节点	49012460413	1.000000e+11
289	47	417	通辽	一般节点	3615	191	长沙	一般节点	50099712071	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
278	47	241	通辽	一般节点	4953	725	贵阳	一般节点	50008939996	1.000000e+11
861	47	417	通辽	一般节点	591	1284	绥化	一般节点	49276967001	1.000000e+11
376	474	460	哈尔滨	一般节点	3757	122	福州	一般节点	48394911971	1.000000e+11
1	47	74	通辽	一般节点	1756	776	北京	网络核心	50056871412	1.000000e+11
531	47	250	通辽	一般节点	3227	195	济南	网络核心	48844966451	1.000000e+11
26	63	74	通辽	一般节点	2701	181	大连	网络核心	50364636480	1.000000e+11

4) 补充：系统抽样

```
[9]: systematic_sample = data_before_sample.copy()
N = len(systematic_sample) # 总体大小
n = 50 # 样本量, 可以改
k = N // n # 抽样间隔
r = np.random.randint(0, k) # 随机起点

indices = np.arange(r, r + n*k, k)[:n]
systematic_sample_finish = systematic_sample.iloc[indices]
systematic_sample_finish
```

```
[9]:
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
1	47	74	通辽	一般节点	1756	776	北京	网络核心	50056871412	1.000000e+11
12	47	260	通辽	一般节点	2549	835	沈阳	网络核心	50220958279	1.000000e+11
23	63	62	通辽	一般节点	36422	394	天津	网络核心	50322780029	1.000000e+11
34	96	99	呼和浩特	一般节点	1257	560	上海	网络核心	49753614568	1.000000e+11
45	96	134	呼和浩特	一般节点	47	252	通辽	一般节点	49416652053	1.000000e+11
56	96	346	呼和浩特	一般节点	1257	138	上海	网络核心	47759033178	1.000000e+11
67	180	28	呼和浩特	一般节点	1385	133	广州	网络核心	52798223188	1.000000e+11
78	180	188	呼和浩特	一般节点	36422	350	天津	网络核心	49047066099	1.000000e+11
89	180	256	呼和浩特	一般节点	1129	171	上海	网络核心	49512421445	1.000000e+11
102	474	467	哈尔滨	一般节点	1257	174	上海	网络核心	49987703744	1.000000e+11
115	474	683	哈尔滨	一般节点	1997	84	天津	网络核心	49446798762	1.000000e+11
126	474	1389	哈尔滨	一般节点	1756	1127	北京	网络核心	48259332712	1.000000e+11
137	591	23	绥化	一般节点	63	6	通辽	一般节点	49462640634	1.000000e+11
148	591	558	绥化	一般节点	36036	499	长春	一般节点	49953028308	1.000000e+11
162	591	1274	绥化	一般节点	1129	203	上海	网络核心	49321244844	1.000000e+11
173	787	307	玉溪	一般节点	4953	686	贵阳	一般节点	49399787960	1.000000e+11

先统计数据总数和需要的样本数量，计算出抽样间隔；然后在前若干个数据里随机选一个起始点；最后按照这个起点，每隔固定数量依次抽取，直到得到所需样本。

5) 补充：整群抽样

```
[10]: cluster_sample = data_before_sample.copy()
clusters = cluster_sample['to_level'].unique()
chosen_clusters = np.random.choice(clusters, size=1, replace=False)
cluster_sample_finish = cluster_sample[cluster_sample['to_level'].isin(chosen_clusters)]
cluster_sample_finish
```

```
[10]:
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
4	47	242	通辽	一般节点	474	672	哈尔滨	一般节点	50492573662	1.000000e+11
5	47	243	通辽	一般节点	96	124	呼和浩特	一般节点	49942713747	1.000000e+11
7	47	250	通辽	一般节点	2473	762	吉林	一般节点	49108721007	1.000000e+11
9	47	252	通辽	一般节点	96	134	呼和浩特	一般节点	50256475808	1.000000e+11
13	47	314	通辽	一般节点	96	152	呼和浩特	一般节点	50161220081	1.000000e+11
...
1057	47	243	通辽	一般节点	2473	769	吉林	一般节点	49117847542	1.000000e+11
1063	47	314	通辽	一般节点	47	252	通辽	一般节点	49900452417	1.000000e+11
1079	63	224	通辽	一般节点	4069	1196	宁波	一般节点	50209459772	1.000000e+11
1097	2473	1460	吉林	一般节点	591	586	绥化	一般节点	48409925693	1.000000e+11
1104	63	6	通辽	一般节点	36036	20	长春	一般节点	50355678076	1.000000e+11

186 rows × 10 columns

按照 to_level 分为若干个群，再随机抽取一个群，显示结果

结论分析：

本实验围绕数据采样方法展开，旨在通过实际操作掌握数据清洗与多种抽样技术的基本流程。实验中首先利用 Pandas 对原始数据进行预处理，删除空行并过滤掉无效记录，从而得到符合条件的数据集。在此基础上，依次完成了加权抽样、随机抽样和分层抽样等方法的实践，对比了不同抽样策略在样本分布上的差异。实验进一步扩展到系统抽样和整群抽样，以便全面理解各种抽样方式的适用场景与特点。通过本实验的完成，加深了对数据预处理和抽样在大数据分析中重要性的理解，为后续实验的深入学习奠定了基础。