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：202300130003	姓名：肖皓天	班级：数据 23
实验题目：实验 1. 数据采样方法实践		
实验学时：32	实验日期：2025/9/23	

实验目标：

利用 Pandas 库实现多种数据采样和过滤的方法

流程描述：

一、库的导入与数据的读入

```
[2]: import pandas as pd
     from pandas import DataFrame
     import numpy as np

     primitive_data=pd.read_csv(r"D:\0_aaa大三上\大数据分析实践\data.csv",encoding='gbk')
     primitive_data
```

[2]:

	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

二、删除多余的空行并进行过滤

删空行

```
primitive_data_1=primitive_data.dropna(how='any')
primitive_data_1
```

	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

过滤

```
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
```

	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

三、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

```
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]
weight_sample_finish
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
146	591	502	绥化	一般节点	1129	546	上海	网络核心	49465128399	1.000000e+11
419	591	98	绥化	一般节点	3227	781	济南	网络核心	50666845945	1.000000e+11
97	474	416	哈尔滨	一般节点	1257	178	上海	网络核心	50599061005	1.000000e+11
1073	47	417	通辽	一般节点	1756	1029	北京	网络核心	49459363742	1.000000e+11
365	180	260	呼和浩特	一般节点	1756	788	北京	网络核心	48917626581	1.000000e+11
414	591	29	绥化	一般节点	235	1649	北京	网络核心	49268934149	1.000000e+11
22	63	60	通辽	一般节点	36422	258	天津	网络核心	49920786706	1.000000e+11
371	474	360	哈尔滨	一般节点	3227	530	济南	网络核心	49027966353	1.000000e+11
1018	474	672	哈尔滨	一般节点	1756	585	北京	网络核心	48132652830	1.000000e+11
432	591	1106	绥化	一般节点	3227	781	济南	网络核心	48568999606	1.000000e+11
174	787	316	玉溪	一般节点	1257	177	上海	网络核心	51407063255	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
669	63	286	通辽	一般节点	3227	468	济南	网络核心	50318390185	1.000000e+11
178	787	326	玉溪	一般节点	3213	597	重庆	网络核心	48608499709	1.000000e+11
295	63	54	通辽	一般节点	3227	493	济南	网络核心	49566827928	1.000000e+11
561	96	108	呼和浩特	一般节点	36272	105	太原	网络核心	49739592973	1.000000e+11

2. 随机抽样

```
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
738	2473	803	吉林	一般节点	235	1621	北京	网络核心	49820870504	1.000000e+11
551	63	224	通辽	一般节点	2994	488	洛阳	网络核心	50811142728	1.000000e+11
142	591	64	绥化	一般节点	36272	105	太原	网络核心	51256753219	1.000000e+11
620	180	264	呼和浩特	一般节点	1129	546	上海	网络核心	50207994896	1.000000e+11
93	180	276	呼和浩特	一般节点	36272	235	太原	网络核心	51775514286	1.000000e+11
987	2473	1460	吉林	一般节点	96	117	呼和浩特	一般节点	50920098518	1.000000e+11
72	180	42	呼和浩特	一般节点	36539	1140	杭州	一般节点	49293665157	1.000000e+11
400	474	1374	哈尔滨	一般节点	591	23	绥化	一般节点	49461593438	1.000000e+11
1035	36036	54	长春	一般节点	591	23	绥化	一般节点	50638071722	1.000000e+11
178	787	326	玉溪	一般节点	3213	597	重庆	网络核心	48608499709	1.000000e+11
1079	63	224	通辽	一般节点	4069	1196	宁波	一般节点	50209459772	1.000000e+11
538	47	417	通辽	一般节点	3227	705	济南	网络核心	49998156282	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
1063	47	314	通辽	一般节点	47	252	通辽	一般节点	49900452417	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
145	591	100	绥化	一般节点	2194	506	唐山	网络核心	51437026945	1.000000e+11
725	2473	1460	吉林	一般节点	36422	446	天津	网络核心	49869730875	1.000000e+11
310	96	102	呼和浩特	一般节点	474	678	哈尔滨	一般节点	49006847943	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

```
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
793	180	20	呼和浩特	一般节点	474	359	哈尔滨	一般节点	50601340670	1.000000e+11
308	63	286	通辽	一般节点	47	258	通辽	一般节点	50067368970	1.000000e+11
822	47	243	通辽	一般节点	474	1311	哈尔滨	一般节点	49029906488	1.000000e+11
402	474	1399	哈尔滨	一般节点	180	252	呼和浩特	一般节点	49271182579	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
59	96	391	呼和浩特	一般节点	47	417	通辽	一般节点	51570663870	1.000000e+11
555	63	278	通辽	一般节点	36036	18	长春	一般节点	50478302302	1.000000e+11
491	47	249	通辽	一般节点	36539	1140	杭州	一般节点	50888438116	1.000000e+11
282	47	250	通辽	一般节点	4953	686	贵阳	一般节点	50250217535	1.000000e+11
780	96	391	呼和浩特	一般节点	180	205	呼和浩特	一般节点	50103206178	1.000000e+11
284	47	252	通辽	一般节点	5058	118	南宁	一般节点	49295040137	1.000000e+11
297	63	60	通辽	一般节点	2473	1053	吉林	一般节点	49803473764	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
1057	47	243	通辽	一般节点	2473	769	吉林	一般节点	49117847542	1.000000e+11
140	591	56	绥化	一般节点	36036	52	长春	一般节点	48627355195	1.000000e+11
732	96	141	呼和浩特	一般节点	36036	499	长春	一般节点	47474335913	1.000000e+11
377	474	467	哈尔滨	一般节点	5058	70	南宁	一般节点	51745421052	1.000000e+11
378	474	472	哈尔滨	一般节点	3643	902	武汉	网络核心	50470657254	1.000000e+11
283	47	251	通辽	一般节点	1257	177	上海	网络核心	49380010804	1.000000e+11

4. 系统抽样

```python
import numpy as np
import pandas as pd

def perform_stratified_selection(dataset_frame, desired_count, *, seed_value=None):

    total_rows = len(dataset_frame)
    if desired_count >= total_rows:
        return dataset_frame.copy()
    random_generator = np.random.default_rng(seed_value)
    sampling_interval = int(total_rows / desired_count) # 直接使用整除代替math.floor
    initial_offset = random_generator.integers(0, sampling_interval)
    selected_indices = np.arange(initial_offset, total_rows, sampling_interval)

    if len(selected_indices) < desired_count:
        all_possible_indices = np.arange(total_rows)
        remaining_indices_pool = np.setdiff1d(all_possible_indices, selected_indices)

        additional_indices = random_generator.choice(
            remaining_indices_pool,
            size=desired_count - len(selected_indices),
            replace=False
        )
        selected_indices = np.concatenate([selected_indices, additional_indices])
    elif len(selected_indices) > desired_count:
        selected_indices = random_generator.choice(selected_indices, size=desired_count, replace=False)

    return dataset_frame.iloc[np.sort(selected_indices)].copy()

result_sample_df = perform_stratified_selection(data_after_filter_2, desired_count=50, seed_value=42)

print(f"抽样后的样本数量: {len(result_sample_df)}")
print(f"抽样后的DataFrame形状: {result_sample_df.shape}")
print("抽样结果前10行:")
display(result_sample_df.head(10))
print("\n完整抽样结果:")
display(result_sample_df)
```

```
抽样后的样本数量: 50
抽样后的DataFrame形状: (50, 10)
抽样结果前10行:
```

	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
11	47	259	通辽	一般节点	1756	245	北京	网络核心	50703793815	1.000000e+11
22	63	60	通辽	一般节点	36422	258	天津	网络核心	49920786706	1.000000e+11
33	63	286	通辽	一般节点	180	52	呼和浩特	一般节点	49725190236	1.000000e+11
44	96	127	呼和浩特	一般节点	1756	1027	北京	网络核心	50087522340	1.000000e+11
55	96	336	呼和浩特	一般节点	1756	1029	北京	网络核心	51600306541	1.000000e+11
66	180	26	呼和浩特	一般节点	36272	133	太原	网络核心	51023900961	1.000000e+11

5. 整群抽样

```python
import numpy as np
import pandas as pd

def get_clusterized_sample(data: pd.DataFrame,
                           category_col: str,
                           sample_size: int,
                           *,
                           random_seed=None):

    if category_col not in data.columns:
        raise KeyError(f"列名 '{category_col}' 不正确。请从 {data.columns.tolist()} 中选择。")

    all_groups = data.loc[data[category_col].notna(), category_col].unique()

    n_selected = min(sample_size, len(all_groups))

    chosen_groups = pd.Series(all_groups).sample(n=n_selected, replace=False, random_state=random_seed).tolist()

    sampled_data = data.query(f"`{category_col}` in @chosen_groups").copy()

    return sampled_data

final_sample = get_clusterized_sample(
    data=data_after_filter_2,
    category_col='to_city',
    sample_size=5,
    random_seed=42
)

record_count = final_sample.shape[0]
final_sample.head()
```

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
26	63	74	通辽	一般节点	2701	181	大连	网络核心	50364636480	1.000000e+11
29	63	230	通辽	一般节点	2701	71	大连	网络核心	50037668767	1.000000e+11
50	96	155	呼和浩特	一般节点	1536	681	广州	网络核心	51538493830	1.000000e+11
61	96	407	呼和浩特	一般节点	4069	1196	宁波	一般节点	49745162804	1.000000e+11
67	180	28	呼和浩特	一般节点	1385	133	广州	网络核心	52708223188	1.000000e+11

结论分析与体会：

通过本次实验，我了解并实践了不同种类的抽样方法。并注意到了不同抽样方法导致结果对应的不同。