实验 1 数据采样方法实践

陆云
202300130239
23 数据
2025.9.19

实验内容以及结果：

[4]:
```
import pandas as pd
from pandas import DataFrame
import numpy as np

primitive_data = pd.read_csv("data.csv", encoding = "gbk")
primitive_data
```

[4]:

	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

[7]:
```
primitive_data_1=primitive_data.dropna(how='any')
primitive_data_1
```

[7]:

	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

[8]:
```
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
```

[8]:

	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

[9]:
```
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[columns]
weight_sample_finish
```

[9]:

	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

[10]:
```
# 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
```

[10]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
39	96	114	呼和浩特	一般节点	2473	769	吉林	一般节点	50350633304	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
1035	36036	54	长春	一般节点	591	23	绥化	一般节点	50638071722	1.000000e+11
437	591	1274	绥化	一般节点	1997	250	天津	网络核心	50278361522	1.000000e+11
881	591	586	绥化	一般节点	1997	39	天津	网络核心	49268870810	1.000000e+11
140	591	56	绥化	一般节点	36036	52	长春	一般节点	48627355195	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
97	474	416	哈尔滨	一般节点	1257	178	上海	网络核心	50599061005	1.000000e+11
888	36036	20	长春	一般节点	1997	85	天津	网络核心	48987594976	1.000000e+11
300	63	70	通辽	一般节点	3643	831	武汉	网络核心	50635697563	1.000000e+11
88	180	254	呼和浩特	一般节点	235	1663	北京	网络核心	51477333650	1.000000e+11
599	474	672	哈尔滨	一般节点	2050	336	石家庄	网络核心	51340689424	1.000000e+11
333	96	383	呼和浩特	一般节点	1536	766	广州	网络核心	50062726803	1.000000e+11
76	180	90	呼和浩特	一般节点	235	1958	北京	网络核心	50714891315	1.000000e+11
644	47	314	通辽	一般节点	96	152	呼和浩特	一般节点	51384841448	1.000000e+11
340	180	20	呼和浩特	一般节点	2841	102	郑州	网络核心	51392475128	1.000000e+11
1021	2473	762	吉林	一般节点	1997	464	天津	网络核心	47991126091	1.000000e+11
49	96	152	呼和浩特	一般节点	47	314	通辽	一般节点	51981076188	1.000000e+11

[11]:
```
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
```

[11]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
59	96	391	呼和浩特	一般节点	47	417	通辽	一般节点	51570663870	1.000000e+11
979	2473	1043	吉林	一般节点	63	282	通辽	一般节点	49176857434	1.000000e+11
1104	63	6	通辽	一般节点	36036	20	长春	一般节点	50355678076	1.000000e+11
1079	63	224	通辽	一般节点	4069	1196	宁波	一般节点	50209459772	1.000000e+11
86	180	226	呼和浩特	一般节点	36036	20	长春	一般节点	49248544673	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
997	36036	52	长春	一般节点	63	12	通辽	一般节点	50822505842	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
49	96	152	呼和浩特	一般节点	47	314	通辽	一般节点	51981076188	1.000000e+11
387	474	677	哈尔滨	一般节点	474	672	哈尔滨	一般节点	50850714694	1.000000e+11
445	787	60	玉溪	一般节点	47	314	通辽	一般节点	49484495071	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
180	787	360	玉溪	一般节点	3615	191	长沙	一般节点	49629725686	1.000000e+11
381	474	475	哈尔滨	一般节点	2473	941	吉林	一般节点	49402590822	1.000000e+11
360	180	218	呼和浩特	一般节点	4069	1195	宁波	一般节点	50955164303	1.000000e+11
913	2473	799	吉林	一般节点	47	243	通辽	一般节点	50993016382	1.000000e+11
441	591	1300	绥化	一般节点	47	252	通辽	一般节点	50817586398	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