## 山东大学<u>计算机科学与技术</u>学院

## 大数据分析实践 课程实验报告

学号: 202300130086 | 姓名: 张珈恺 | 班级: 23 数据

实验题目:数据采样方法实践

实验目标:

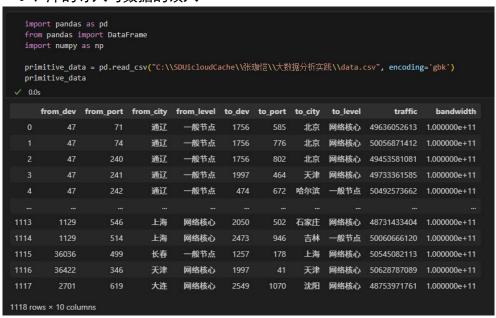
利用 Pandas 库实现多种数据采样和过滤的方法

实验环境:

Python3.9, Jupyter notebook

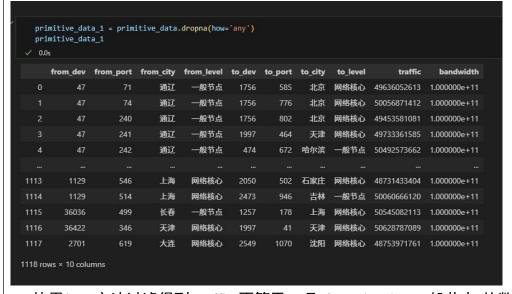
实验步骤与内容:

1 、库的导入与数据的读入

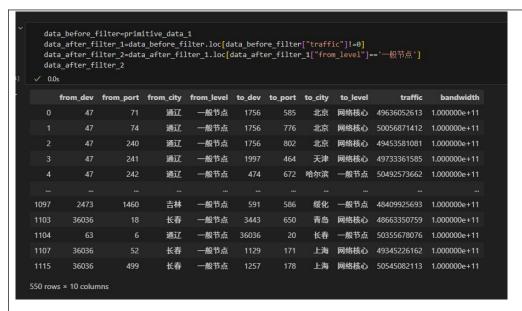


2 、删除多余的空行并进行过滤

使用 dropna 方法并指定参数为'any'删除多余的空行



使用loc 方法过滤得到'traffic'不等于0 且'from level'='一般节点'的数据



## 3、对数据进行抽样

使用 sample 方法采取不同的采样方式采取 50 个样本并比较采样结果 加权采样: to level 的值为一般节点与网络核心的数据采样权重之比为 1:5

co we we fo we	lumns=dat ight_samp ight_samp r i in we if weigl else: weight_ ight_samp ight_samp ight_samp	ght=1 ght=5 sample.at[ le_finish= le_finish=	ample.col fore_samp ']=0 e.index: at[i,'to_  i,'weight weight_sam	umns le.copy() level']=='	⊵ <mark>(n=50,</mark> ı	weights=	'weight'			
	from_dev	from_port				to_port	to_city	to_level	traffic	bandwidth
142	591	64	绥化	一般节点	36272	105	太原	网络核心	51256753219	1.000000e+11
321	96	135	呼和浩特	一般节点	2050	553	石家庄	网络核心	51921872375	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
147	591	526	绥化	一般节点	1129	514	上海	网络核心	49318922185	1.000000e+11
91	180	264	呼和浩特	一般节点	63	70		一般节点	50106121660	1.000000e+11
8	47	251	通辽	一般节点	2549	839	沈阳	网络核心	50755299504	1.000000e+11
22	63	60	通辽	一般节点	36422	258	天津	网络核心	49920786706	1.000000e+11
492	47	250	通辽	一般节点	4515	652	西安	网络核心	49014089485	1.000000e+11
929	4360	468	南京	一般节点	1997	464	天津	网络核心	49145116989	1.000000e+11
305	63	232	通辽	一般节点	2549	1066	沈阳	网络核心	49269663214	1.000000e+11
54	96	159	呼和浩特	一般节点	2360	266	太原	网络核心	51625089370	1.000000e+11
717	2473	1043	吉林	一般节点	36422	324	天津	网络核心	50594312992	1.000000e+11
	787	52	玉溪	一般节点	2360	215	太原	网络核心	49322809158	1.000000e+11
443			通辽	一般节点	235	106	北京	网络核心	52195591947	1.000000e+11
	63	10	AMAZ							
443	63 47	10 252	通辽	一般节点	591	560	绥化	一般节点	51065218921	1.000000e+11

随机采样

random_sample=data_before_sample random_sample_finish=random_sample.sample(n=50) random_sample_finish=random_sample_finish[columns] random_sample_finish  ✓ 0.0s										
	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
80	180	200	呼和浩特	一般节点	2701	300	大连	网络核心	51884294458	1.000000e+11
436	591	1266	绥化	一般节点	2050	505	石家庄	网络核心	51285397493	1.000000e+11
1039	180	264	呼和浩特	一般节点	36036	54	长春	一般节点	49124032697	1.000000e+11
75	180	84	呼和浩特	一般节点	1536	86	鄂尔多斯	网络核心	49100967003	1.000000e+11
131	474	1473	哈尔滨	一般节点	2549	1461	沈阳	网络核心	53304989080	1.000000e+11
40	96	117	呼和浩特	一般节点	2050	505	石家庄	网络核心	48814619370	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
150	591	582	绥化	一般节点	2701	619	大连	网络核心	50838395442	1.000000e+11
276	47	74	通辽	一般节点	4561	1033	成都	网络核心	50819524115	1.000000e+11
824	47	252	通辽	一般节点	474	1389	哈尔滨	一般节点	51228405663	1.000000e+11
47	96	136	呼和浩特	一般节点	2360	215	太原	网络核心	49292630301	1.000000e+11
363	180	254	呼和浩特	一般节点	2360	62	太原	网络核心	50252917820	1.000000e+11
372	474	416	哈尔滨	一般节点	3227	512	济南	网络核心	49544939922	1.000000e+11
309	96	99	呼和浩特	一般节点	2360	76	太原	网络核心	49047882786	1.000000e+11
932	2473	1460	吉林	一般节点	1997	467	天津	网络核心	50151515116	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
110	474	672	哈尔滨	一般节点	47	242	通辽	一般节点	51555817613	1.000000e+11
448	787	307	玉溪	一般节点	36422	258	天津	网络核心	51727332383	1.000000e+11

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

```
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
✓ 0.0s
    from_dev from_port from_city from_level to_dev to_port
                                                    to_city
                                                           to_level
                        通辽
                              一般节点 36539
                                                      杭州
                                                           一般节点 50888438116 1.000000e+11
        47
                249
                                              1140
                       哈尔滨
                              一般节点
                                        96
                                               124 呼和浩特 一般节点 49289354051 1.000000e+11
308
                        通辽
                              一般节点
                                                      通辽 一般节点 50067368970
                              一般节点 2473
                                                      吉林 一般节点 49735704801
                       通辽
         63
                              一般节点 36036
                                                      长春 一般节点 48363382095 1.000000e+11
                 60
                        通辽
                              一般节点
                                      4360
                                               468
                                                      南京
                                                           一般节点
                                                                            1.000000e+11
546
                                                                  47970715088
                       哈尔滨
                              一般节点
                                              18 呼和浩特
                                                           一般节点
                                                                   48043608658
                        宁波
                                                           一般节点
                              一般节点
                                                   呼和浩特
                                                                  50099141709
                              一般节点
                                             256 呼和浩特 一般节点 51915256521 1.000000e+11
                        长春
                                        180
       36036
                        绥化.
                              一般节点 180
                                               218 呼和浩特 一般节点 50522034278 1.000000e+11
                     呼和浩特
                              一般节点 474
                                                    哈尔滨 一般节点 49006847943 1.000000e+11
                               一般节点 474
                                                     哈尔滨 一般节点 48667628783
779
         96
                 152 呼和浩特
                               一般节点
                                      180
                                               202 呼和浩特 一般节点 51162997127
                                                                            1.000000e+11
383
                       哈尔滨
                               一般节点
                                        5058
                                               144
                                                      南宁
                                                           一般节点 50998204735
                                                                             1.000000e+11
                               一般节点
                                                      吉林
                                                           一般节点
                                                                  49803820036
                 52 呼和浩特
                               一般节点
                                                    哈尔滨 一般节点 50713290427
                                                                             1.000000e+11
                                                      无锡 一般节点 50322958171 1.000000e+11
                         绥化
                              一般节点
                                      4448
160
1086
       36539
                1140
                         杭州
                              一般节点
                                               1661
                                                      北京 网络核心 51411580502 1.000000e+11
                              一般节点
                                                      大连 网络核心 49891276242 1.000000e+11
                       玉溪 一般节点 3213
                                               246 重庆 网络核心 50468642387 1.000000e+11
```

结论与体会:
本次实验成功用 Pandas 实现数据清洗与三种采样。删除空行并筛选出有效数据后,加权、随机、分层采样各有特点:加权突出网络核心样本,分层按比例保留类别特征,随机则无偏向性。实验让我掌握了数据预处理关键步骤,也认识到需依需求选合适采样方法,为后续分析打牢基础。