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：202300130166	姓名：朱亚宁	班级：23 数据
实验题目：数据采样方法实践		
实验学时：2	实验日期：2025.9.15	
实验目标：利用 Pandas 库实现多种数据采样和过滤的方法		

实验步骤：

一、库的导入与数据的读入

```
import pandas as pd
from pandas import DataFrame
import numpy as np
primitive_data=pd.read_csv("data-sample-and-filter.csv",encoding="gbk")
primitive_data
```

“encoding=”gbk”” 部分：

utf-8：适用于大多数包含中文、英文等字符的文件，现在很多文件默认采用 utf-8 编码。

gbk：主要用于处理包含中文的文件，在中文 Windows 系统环境下生成的文件常使用该编码。

gbk2312：是 gbk 的子集，也用于中文编码，不过现在使用范围相对 gbk 较窄。

latin_1：能够满足显示英文字母、数字、常见标点符号，以及西欧语言中一些带变音符号字符的需求

二、自行实现系统抽样

系统抽样：系统抽样的核心是先确定抽样间隔，再从起始位置开始按间隔抽取样本，确保样本在整体中均匀分布。

代码：

```
#系统抽样

sample_size=50

systematic_sample=data_before_sample

total=len(systematic_sample)

interval = np.ceil(total / sample_size).astype(int)

start = np.random.randint(1, interval + 1)

sample_indices = np.arange(start, total + 1, interval) - 1

sample_indices = sample_indices[sample_indices < total]

after_systematic_sample=systematic_sample.iloc[sample_indices]

after_systematic_sample
```

结果：（部分）

Out[21]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
7	47	250	通辽	一般节点	2473	762	吉林	一般节点	49108721007	1.000000e+11
18	63	10	通辽	一般节点	235	106	北京	网络核心	52195591947	1.000000e+11
29	63	230	通辽	一般节点	2701	71	大连	网络核心	50037668767	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
51	96	156	呼和浩特	一般节点	3227	103	济南	网络核心	50654404568	1.000000e+11
62	96	460	呼和浩特	一般节点	1997	729	天津	网络核心	49455021334	1.000000e+11
73	180	50	呼和浩特	一般节点	4515	652	西安	网络核心	50640954639	1.000000e+11
84	180	214	呼和浩特	一般节点	2701	135	大连	网络核心	48901190886	1.000000e+11
95	474	359	哈尔滨	一般节点	2050	502	石家庄	网络核心	51299508559	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

三、自行实现整群抽样

整群抽样：将总体划分为若干个 “群”，随机抽取部分群，然后对选中的群内所有个体进行调查的抽样方法。

代码：

```
#以出发城市划分群，随机抽取若干城市，将这些城市对应的数据全部选取出来作为样本。
cluster_sample=data_before_sample
group_col='from_city'
n_clusters=3
all_clusters = cluster_sample[group_col].unique()
selected_clusters = np.random.choice(all_clusters, size=n_clusters, replace=False)
sampled_data = cluster_sample[cluster_sample[group_col].isin(selected_clusters)]
sampled_data
```

结果：

Out[28]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
656	4069	1196	宁波	一般节点	180	264	呼和浩特	一般节点	49766912004	1.000000e+11
724	4069	1205	宁波	一般节点	36036	52	长春	一般节点	50994646887	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
757	3615	179	长沙	一般节点	96	391	呼和浩特	一般节点	51467597716	1.000000e+11
759	3757	122	福州	一般节点	96	407	呼和浩特	一般节点	47597054356	1.000000e+11
863	4069	1196	宁波	一般节点	591	1290	绥化	一般节点	48726638175	1.000000e+11
931	4069	1205	宁波	一般节点	1997	466	天津	网络核心	52060473597	1.000000e+11
986	4069	1205	宁波	一般节点	96	114	呼和浩特	一般节点	49413180407	1.000000e+11
1075	4069	1196	宁波	一般节点	1756	1187	北京	网络核心	50488255524	1.000000e+11