第十一届“泰迪杯”数据挖掘挑战赛——

A 题：新冠疫情防控数据的分析

一、问题背景

自 2019 年底至今，全国各地陆续出现不同程度的新冠病毒感染疫情，如何控制疫情蔓延、维持社会生活及经济秩序的正常运行是疫情防控的重要课题。大数据分析为疫情的精准防控提供了高效处置、方便快捷的工具，特别是在人员的分类管理、传播途径追踪、疫情研判等工作中起到了重要作用，为卫生防疫部门的管理决策提供了可靠依据。疫情数据主要包括人员信息、场所信息、个人自查上报信息、场所码扫码信息、核酸采样检测信息、疫苗接种信息等。

本赛题提供了某市新冠疫情防疫系统的相关数据信息，请根据这些数据信息进行综合分析，主要任务包括数据仓库设计、疫情传播途径追踪、传播指数估计及疫情趋势研判等。

二、 解决问题

1. 根据核酸检测中阳性人员的出行时间与场所追踪密接者，将结果保存到“result1.csv”文件中（文件模板见附件 1 中的 result1.csv）。

2. 由问题 1 的结果，根据密接者的出行时间与场所追踪相应的次密接者，将结果保存到“result2.csv”文件中（文件模板见附件 1 中的 result2.csv）。

3. 建立模型，分析接种疫苗对病毒传播指数的影响。

4. 根据阳性人员的数量及辐射范围，分析确定需要重点管控的场所。

5. 为了更精准地进行疫情防控和人员管理，你认为还需要收集哪些相关数据。基于这些数据构建模型，分析其精准防控的效果。

注 在解决上述问题时，要求结合赛题提供的数据信息表建立数据仓库，实现数据治理的内容，请在论文中明确阐述做了哪些数据治理工作，具体是如何实现的。

三、附件说明

附件 1：问题 1 和问题 2 结果的文件模板（result1.csv、result2.csv）

附件 2：人员信息表

序号	字段名	字段说明	字段类型	默认值
1	user_id	人员 ID：人员的唯一标识	bigint(20)	
2	openid	微信 OpenID	varchar(64)	null
3	gender	性别：男、女	varchar(2)	null
4	nation	民族	varchar(20)	null
5	age	年龄	int	null
6	birthdate	出生日期	varchar(20)	null
7	create_time	创建时间	timestamp	null

附件 3：场所信息表

序号	字段名	字段说明	字段类型	默认值
1	grid_point_id	场所 ID：场所的唯一标识	bigint(20)	
2	name	场所名	varchar(255)	null
3	point_type	场所类型	varchar(50)	
4	x_coordinate	X 坐标（单位：米）	decimal(12,2)	null
5	y_coordinate	Y 坐标（单位：米）	decimal(12,2)	null
6	create_time	创建时间	timestamp	null

附件 4：个人自查上报信息表

NO.	字段名	字段说明	字段类型	默认值
1	sno	序列号：自查记录的唯一标识	bigint(20)	
2	user_id	人员 ID：对应于“人员信息表”中的 user_id	bigint(20)	
3	x_coordinate	上报地点的 X 坐标	decimal(12,2)	null
4	y_coordinate	上报地点的 Y 坐标	decimal(12,2)	null
5	symptom	症状：1 发热、2 乏力、3 干咳、4 鼻塞、5 流涕、6 腹泻、7 呼吸困难、8 无症状	varchar(100)	null
6	nucleic_acid_result	核酸检测结果：0 阴性、1 阳性、2 未知（非必填）	varchar(10)	null
7	resident_flag	是否常住居民：0 未知、1 是、2 否	int	null
8	dump_time	上报时间	timestamp	null

附件 5：场所码扫码信息表

序号	字段名	字段说明	字段类型	默认值
1	sno	序列号：扫码记录的唯一标识	bigint(20)	
2	grid_point_id	场所 ID：对应于“场所信息表”中的 grid_point_id	bigint(20)	
3	user_id	人员 ID：对应于“人员信息表”中的 user_id	bigint(20)	
4	temperature	体温	double	null
5	create_time	扫码记录时间	timestamp	null

附件 6：核酸采样检测信息表

序号	字段名	字段说明	字段类型	默认值
1	sno	序列号：核酸采样记录的唯一标识	bigint(20)	
2	user_id	人员 ID：对应于“人员信息表”中的 user_id	bigint(20)	null
3	cysj	采样日期和时间	timestamp	null
4	jcsj	检测日期和时间	timestamp	null
5	jg	检测结果：阴性、阳性、未知	varchar(50)	null
6	grid_point_id	场所 ID：对应于“场所信息表”中的 grid_point_id	bigint(20)	

附件 7：疫苗接种信息表

序号	字段名	字段说明	字段类型	默认值
1	sno	序列号：疫苗接种记录的唯一标识	bigint(20)	
2	inject_sn	接种流水号	varchar(50)	
3	user_id	人员 ID：对应于“人员信息表”中的 user_id	varchar(50)	
4	age	接种者年龄	int	null
5	gender	性别：1 男、2 女	varchar(10)	null
6	birthdate	出生日期	varchar(50)	null
7	inject_date	接种日期	timestamp	null
8	inject_times	针次：1 第一针、2 第二针、3 加强针	varchar(30)	null
9	vaccine_type	疫苗类型：1 灭活疫苗、2 重组蛋白疫苗、3 病毒载体疫苗、4 核酸疫苗、5 减毒疫苗	varchar(30)	null