**课 程 设 计 任 务 书**

**课程名称**  Python与数据分析课程设计

**题 目**  疫情数据分析

**指导教师**  王灿

**设计起止日期** 2022.12.05至2021.12.13

**学 院**  理学院

**专 业**  信息与计算科学

Python与数据分析课程设计

# 目的

1. 掌握使用Python进行数据搜集、清洗、处理、存储、变换及可视化的方法。
2. 掌握基于数据分析得出结论的科学方法。

# 要求

网上搜集疫情数据，对数据进行清洗、处理、存储、变换及可视化，观察数据中隐含的信息，分析疫情的基本情况、现状、特点。

学号为奇数做国际各国家的，学号为偶数做英国各地区的。

# 步骤

## 数据搜集

### 数据要求

1. 搜集国际疫情数据，符合以下要求
   1. 数据统计以国家为单位，包含至少50个国家
   2. 包含从2020年1月某日开始几乎每天的数据（每个国家启始日期可以不同）
   3. 包含每日新增病例数、每日新增死亡数
   4. 数据源需要来自英文网站
2. 搜集英国疫情数据，符合以下要求
   1. 数据统计以地区为单位，至少包含300个地区
   2. 包含从2020年1月某日开始几乎每天的数据（每个地区启始日期可以不同）
   3. 包含每日新增病例数、每日新增死亡数
   4. 数据源需要来自英文网站

### 撰写报告要求

1. 给出找到数据的具体方式，比如搜索的关键词xxx，然后找到关键网页xxx，找到关键文件xxx，通过什么方式获取数据等（关键步骤截图）
2. 给出数据源中每日新增病例数、每日新增死亡数的统计方法，给出数据描述的原文和翻译

## 数据清洗

### 清洗要求

1. 数据保留四列：日期、区域、每日新增病例数、每日新增死亡数
   1. 日期精确到天，格式处理为2020-01-15样式
   2. 区域名为英文，可以为全称（可以包含空格），也可为缩写
   3. 每日新增病例数为整数
   4. 每日新增死亡数为整数
2. 数据格式处理为CSV格式（即列之间用英文逗号分隔）
3. 需要将CSV保存到本地，以备后续使用

### 撰写报告要求

1. 数据清洗的关键步骤、代码、截图需要完整列出，尽可能使用Python进行清洗
2. 需要给出清洗前后的对比截图

## 数据预处理

### 预处理要求

1. 使用pandas读入清洗后的CSV文件
2. 数据类型处理
   1. 日期：datetime类型（自己了解这个类型）
   2. 区域：str或object类型
   3. 每日新增病例数：int类型
   4. 每日新增死亡数：int类型
3. 重复数据处理
   1. 检测是否存在重复行（思考怎么定义重复行，哪些列重复算重复）
   2. 如果存在重复行，需要去除（思考如何去除是对的，依据是什么）
4. 缺失值
   1. 对于nan值，对照原始数据判断是否应该是nan，是否需要避免pandas自动转换为nan
   2. 对于缺失的日期、地区等信息，思考是否可以用0补全，还是说只能用nan补全

### 撰写报告要求

1. 需要给出关键代码和截图
2. 需要给出处理前后效果对比
3. 需要给出思考结果描述

## 数据库存储

### 存储要求

1. 使用清洗后的DataFrame变量
2. 将预处理后的数据的每行导入数据库
   1. 使用sqlite3来建立数据库
   2. 通过sql语句构建表（covid19），格式要合适，日期和地区两列一起作为主键（搜索如何使用sql完成这个任务）
   3. 对表中数据进行增/删/改/查操作验证
      1. 增加一行数据
      2. 查找某天某个地区的数据
      3. 修改某天某个地区的数据
      4. 删除一行数据
   4. 插入预处理后的DataFrame的所有行到表中，使用for循环批量处理

### 撰写报告要求

1. 需要给出关键代码和截图
2. 需要给出表进行增/查/改/删操作验证的截图，并说明是否符合预期
3. 需要给出最终的表的总行数

## 数据变换

### 数据变换要求

1. 基于清洗后的DataFrame计算并生成一个行是每个地区、列为下面特征的新的DataFrame
   1. 地区
   2. 累积病例数
   3. 累积死亡数
   4. 最大单日新增病例数
   5. 最大单日新增死亡数
   6. 最大单日新增病例数对应日期（如果有多个，取最晚的）
   7. 最大单日新增死亡数对应日期（如果有多个，取最晚的）
2. 分别求出累积病例数、累积死亡数、最大单日新增病例数、最大单日新增死亡数排前5的地区

### 撰写报告要求

1. 需要给出新的DataFrame
2. 需要给出所要求的地区

## 数据可视化

### 可视化要求

1. 对累积病例数最多的5个地区，画出每日新增病例数线图，画在一个图上
2. 对累积死亡数最多的5个地区，画出每日新增死亡数线图，画在一个图上
3. 对累积病例数最多的5个地区，画出累积病例数的柱状图，画在一个图上
4. 对累积死亡数最多的5个地区，画出累积死亡数的柱状图，画在一个图上
5. 对地区数据根据累积病例数画饼图（饼图保留前5项数值最大地区+一项“其他”，“其他”的数据是把所有其他地区数据相加得到的），饼图中各地区按照占比大小顺序排好
6. 对地区数据根据累积死亡数画饼图（饼图保留前5项数值最大地区+一项“其他”，“其他”的数据是把所有其他地区数据相加得到的），饼图中各地区按照占比大小顺序排好

### 撰写报告要求

1. 需要给出作图代码和图

## 得出结论

### 具体要求

1. 需要通过数据结果或可视化结果，给出观察到的现象，得出一些结论，至少5个结论

### 撰写报告要求

1. 每个结论都要有数据支撑，不要脱离数据讲空话
2. 不要为了得到某个结论去扭曲编造数据