



**《Python数据分析及可视化》课程大作业**

**实**

**施**

**报**

**告**

|  |  |
| --- | --- |
| **项目名称：** | **基于新冠疫情数据的传染病区域地理性分析** |
| **小组组长：** | **戴棋** |
| **小组成员：** | **张青松 刘浩天** |
|  | **刘泽康 刘俊杰** |

**二O二三年六月**

**《基于新冠疫情数据的传染病区域地理性分析》实施报告**

**摘要：** 在新冠疫情的地理性分析项目中，通过收集和准备疫情数据，进行数据分析和可视化，以及区域和地域分析，可以深入了解疫情的地域分布和趋势。该项目的主要步骤包括数据收集和准备、数据分析和可视化、区域和地域分析，结果解释和报告，以及持续更新和改进。通过这些步骤，可以得出对疫情的结论和洞察，为决策者和利益相关者提供科学依据。关键字：新冠疫情、地理性分析、数据收集、数据准备、数据分析、可视化、区域分析、地域分析、结果解释、报告、持续更新。

**关键字：**新冠疫情、地理性分析、数据收集、数据准备、数据分析、可视化、区域分析、地域分析、结果解释、报告

**一、项目简介**

**本项目旨在通过对新冠疫情数据的统计分析和地理可视化，深入探索疫情的地域分布和传播模式，为相关决策和防控措施提供科学依据。项目将收集2020年到2022年中国各个省市的新冠疫情时间序列数据，包括确诊人数、治愈人数和疑似症状人数等指标。通过数据清洗、整理和规范化，确保数据的准确性和一致性。**

**在数据分析阶段，项目将使用Python编程语言和相关的数据分析库，如Pandas、NumPy和Matplotlib等，计算每个省市在不同时间点的累计确诊人数、治愈人数和疑似症状人数，并绘制相应的折线图和柱状图，以展示疫情的总体趋势和变化。**

**针对地理性分析，项目将利用地理信息可视化库，如Geopandas、Folium和Plotly，绘制中国地图，并根据各个省市的确诊人数、治愈人数和疑似症状人数的大小，展示疫情的地理分布和密度，以便观察不同地区的疫情情况和变化趋势。**

**最终，项目将对疫情数据进行汇总统计和区域比较分析，探索不同地区之间的关联性和趋势。通过解释和解读分析结果，形成结论和洞察，并撰写项目报告，提供给决策者和相关利益相关者，以支持相关决策和未来研究方向。**

**通过该项目的实施，我们将能够全面了解新冠疫情的地域特征和趋势，为疫情防控策略的制定和调整提供科学支持，以促进公共卫生和社会稳定的维护。**

**二、项目环境及资源的配置**

**项目实施所需要的环境及条件包括以下方面：**

**1. 软件资源：**

**- Python编程语言：项目的实施需要有Python的开发环境，我们所使用的是pycharm编译器**

**- 数据分析和可视化库：我们安装了必要的Python库，如Pandas、NumPy、Matplotlib和Seaborn等，以支持对疫情数据的处理、分析和可视化。**

**- 数据库管理系统：如果在项目后期我们需要处理大量的疫情数据，可以考虑使用数据库管理系统（如MySQL、PostgreSQL等）来存储和查询数据。**

**2. 硬件资源：**

**- 计算机：需要一台配置良好的计算机或服务器，以确保能够高效处理大规模的数据集和计算任务。**

**- 存储空间：有足够的存储空间来存储疫情数据和项目所生成的分析结果。**

**- 内存和处理器：用于支持大规模数据处理和计算任务，且拥有足够的内存和处理器性能。**

**3. 网络资源：**

**- 数据获取和共享：确保能够连接到互联网，以获取最新的疫情数据和相关数据集。还可以通过网络资源与其他研究人员、机构或开源社区进行合作和知识共享。**

**- 在线文档和资料：访问在线文档、技术资料和开源社区论坛等网络资源，以获取关于Python编程、数据分析和可视化的指导和支持。**

**4. 数据资源：**

**- 疫情数据：确保能够获取到较完整且准确的新冠疫情数据。这些数据来自公共卫生部门、卫生组织、学术机构、国际组织或公开数据平台等。**

**总体而言，项目实施所需的环境和条件包括适当的软件资源（Python、数据分析库、数据可视化库等）、良好的硬件配置（计算机、存储空间、内存和处理器）以及可靠的网络资源（数据获取和共享、在线文档和资料）。这些资源将为项目的进行提供支持，并确保能够高效处理和分析疫情数据。**

**三、项目实施过程**

**1、数据集的准备**

**从获取原始数据，再到数据的初步清洗、整理、规范化、标准化……**

**1.1原始数据的环境因素**

**依托与疫情时代下的大数据分析,原始疫情数据具有以下特点:**

**- 数据可用性：新冠疫情数据在全球范围内得到广泛收集和共享。大量的疫情数据的可用性为项目提供了坚实的基础，可以支持对疫情的深入分析。**

**- 数据质量：确保疫情数据的准确性、完整性和时效性至关重要。有关机构和组织已经建立了数据收集和报告的标准，这有助于保证数据的可靠性，从而提高项目的成功实施概率。**

**1.2 应对网站的反扒机制**

**为了获取原始的疫情数据,我们准备从丁香园疫情数据平台以及网络上各种开源的疫情数据接口等爬取了原始的疫情数据,在爬取的过程遭遇到了不同程度上的困难, 其中为了防止被过度爬取或保护数据的安全性，丁香园疫情数据平台等数据接口会施加一些阻碍:**

**- 网站的反爬机制：为了防止被过度爬取或保护数据的安全性，网站可能会实施一些反爬机制。例如，限制对特定页面的访问频率，要求进行人机验证（如验证码），或者通过检测爬虫的请求头信息来阻止爬取。**

**- 登录和身份验证：如果目标网站要求用户登录才能访问数据，可能需要编写代码来模拟登录过程，包括提交登录表单、处理会话管理和保持登录状态等。**

**- 动态生成的内容：一些网站使用JavaScript或AJAX等技术来动态生成页面内容。如果使用基于文本的爬取工具（如BeautifulSoup）进行爬取，可能无法获取到动态生成的内容。在这种情况下，可以考虑使用Selenium等工具，模拟浏览器行为，从而获取完整的页面内容。**

**- Robots.tx文件限制：网站通常会在根目录下提供一个名为 Robots.txt 的文件，用于指定哪些页面可以被爬取，哪些页面禁止被爬取。在进行网络爬取时，应该尊重 Robots.txt 文件中的规定，遵循网站的爬取规则。**

**- IP封锁和限制：如果爬取请求频率过高或过于频繁，可能会导致IP被封锁或受到访问限制。为了避免这种情况，可以使用代理IP轮换、限制请求频率、使用随机的用户代理等方法来减少被封锁的风险。**

**- 数据格式和结构变化：网站的数据格式和结构可能会发生变化，导致爬虫代码无法正确解析和提取数据。在进行网络爬取时，需要保持对目标网站的监测，及时更新和调整爬虫代码以适应变化。**

**为了应对这些阻碍，我们通过查询网络来获取一些常用的技术手段,其中我们通过使用代理IP、随机延迟访问时间、模拟人类行为, 使用Selenium模拟浏览器行为等技术手段来获取我们所需的原始数据,再获取到我们所需要原始数据**

**1.3数据的初步清洗、整理、规范化、标准化**

**在进行数据的初步清洗、整理、规范化和标准化过程中，我们使用到了以下 Python 知识和技术：**

**- 数据加载与查看：使用 Pandas 库中的函数读取数据文件（如 CSV、Excel、数据库等），将数据加载到 DataFrame 中，并使用 DataFrame 的方法和属性查看数据的结构、样本等信息。**

**- 缺失值处理：使用 Pandas 提供的函数和方法，如 `isnull()`、`fillna()`，对缺失值进行检测、填充或删除。我们会根据数据的特点和分析需求，选择适当的缺失值处理方法。**

**- 重复值处理：使用 Pandas 的 `duplicated()` 和 `drop\_duplicates()` 函数，检测和删除数据中的重复值。这有助于保持数据的一致性和准确性。**

**- 数据类型转换：使用 Pandas 的 `astype()` 方法，将数据列转换为特定的数据类型，如将字符串转换为数字、日期等。这有助于进行后续的计算和分析操作。**

**- 数据筛选与子集提取：使用 Pandas 的条件筛选、切片和索引功能，根据特定条件选择数据行或列。我们会使用布尔索引、`loc[]` 或 `iloc[]` 进行数据子集的提取。**

**- 数据排序与排名：使用 Pandas 的排序函数，如 `sort\_values()` 和 `sort\_index()`，对数据进行排序操作。此外，还可以使用 `rank()` 函数对数据进行排名，以便进一步分析和比较。**

**- 数据规范化与标准化：使用 Scikit-learn 或其他统计库中的函数，对数据进行规范化或标准化处理。其中我们使用 `StandardScaler` 进行标准化（均值为0，方差为1）或使用 `MinMaxScaler` 进行数据归一化（缩放到特定范围）。**

**- 数据聚合与分组操作：使用 Pandas 的 `groupby()` 函数对数据进行分组，然后应用聚合函数（如求和、平均值、计数等）进行数据汇总与分析。这对于数据的统计摘要和汇总非常有用。**

**- 数据去重与合并：使用 Pandas 的 `drop\_duplicates()` 函数去除数据中的重复项，并使用 `merge()` 函数根据某个共同的键合并不同的数据集。这有助于整合和合并来自不同来源的数据。**

**- 文本处理与正则表达式：使用 Python 的内置字符串方法、正则表达式和文本处理库（如 re、nltk）对文本数据进行清洗、提取关键信息、分词等操作。这在文本挖掘和自然语言处理中经常用到。**

**通过以上数据清洗,我们以时间序列为排序,分为以下几列题头continentName,** **continentEnglishName,** **countryName,** **countryEnglishName,** **provinceName,provinceEnglishName,province\_zipCode,** **province\_confirmedCount,** **province\_suspectedCount,** **province\_curedCount,** **province\_deadCount,** **cityName,** **cityEnglishName,** **city\_zipCode,** **city\_confirmedCount,** **city\_suspectedCount,** **city\_curedCount,** **city\_deadCount以及updateTime**

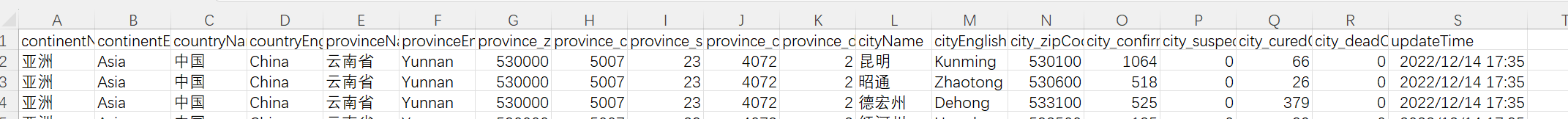


图1-1数据图预览

**2、数据统计分析处理**

**在我们的数据统计分析处理中,我们主要是运用到了pyhton语言中的pandas库**

**在数据科学和分析项目中，Pandas库在数据处理、数据分析和数据清洗方面发挥了重要作用。以下是Pandas库在项目中的主要作用：**

**1. 数据读取和写入：Pandas提供了读取和写入各种数据格式的函数，如CSV、Excel、SQL数据库、JSON等。这使得从不同来源获取数据变得简单，可以轻松加载和保存数据。**

**2. 数据清洗和预处理：Pandas提供了丰富的功能和方法来处理数据中的缺失值、异常值和重复值。通过使用Pandas的函数和方法，可以进行数据过滤、填充缺失值、删除异常值、去除重复值等操作，以确保数据的质量和一致性。**

**3. 数据探索和分析：Pandas提供了灵活的数据结构，如Series和DataFrame，使得数据的探索和分析变得简单高效。可以使用Pandas的函数和方法进行数据的切片、索引、过滤、排序、分组、聚合等操作，以便对数据进行深入分析。**

**4. 数据转换和处理：Pandas提供了丰富的数据转换和处理方法，如数据类型转换、字符串处理、日期时间处理、数值计算、数据合并和重塑等。这些功能使得对数据进行计算、转换和整理变得方便，有助于为后续的统计分析和建模做好准备。**

**5. 数据可视化：Pandas结合Matplotlib等数据可视化库，可以轻松生成各种图表和可视化图像，如折线图、柱状图、散点图等。这有助于直观地展示数据的分布、趋势和关系，提供洞察和可视化结果。**

**6. 数据导出和报告：Pandas可以将处理和分析后的数据导出为各种格式，如CSV、Excel、HTML、PDF等。这方便了数据的共享和报告的生成，使得项目的成果能够以可视化和可交互的方式呈现给相关的利益相关者。**

**总而言之，Pandas库为数据处理、数据分析和数据清洗提供了强大而灵活的工具和方法，能够帮助项目团队高效地处理和分析数据，从而取得准确的结果和洞察。**

**借助pandas库,我们首先会对处理过的数据进行汇总统计：**

**为了了解新冠疫情的影响范围,累计确诊人数是一个十分重要的统计点,通过统计各个省市在同一时间点的确诊人数的增加量,可以结合各个省市的地理关系初步统计出其中的影响:**

**- 累计确诊人数：计算每个省市在不同时间点的累计确诊人数**

**我们将会通过以下步骤:**

**1. 导入库：使用 `import pandas as pd` 导入 Pandas 库，以便进行数据处理和分析。**

**2. 读取数据：使用 `pd.read\_csv()` 函数读取名为 "data.csv" 的 CSV 文件，并将数据存储在 DataFrame (`df`) 中。**

**3. 数据筛选：利用 Pandas 的逻辑条件筛选，通过对列进行条件判断，选择包含同样年份且 countryName 为 "中国" 的数据子集。使用逻辑运算符 `&` 连接多个条件。**

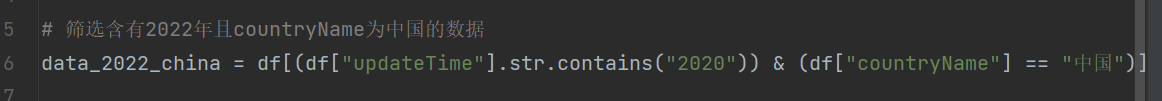


图2-1 数据筛选代码预览

**4. 分组计算：使用 `groupby()` 方法根据 "provinceName" 列进行分组，将数据按照不同的省份进行分组。**

**5. 循环迭代：通过 `for` 循环迭代每个分组，依次处理每个省份的数据。**

**6. DataFrame 的行操作：通过 `iloc[]` 方法，通过索引位置选择第一个和最后一个数据行，并提取对应的 "province\_confirmedCount" 列的值。**

**7. 计算差值：将第一个值减去最后一个值，得到差值后处理成为该年的确诊人数。**

**8. 创建和追加 DataFrame：使用 `pd.concat()` 函数将每个省份的结果追加到结果 DataFrame (`result`) 中。**

**9. 结果输出：使用 `print()` 函数打印输出结果 DataFrame。**

**这些步骤涵盖了数据读取、筛选、分组、循环迭代、DataFrame 的行操作、数据计算和结果输出等方面的操作。通过组合运用这些知识，可以实现对数据的处理、分析和结果展示。**

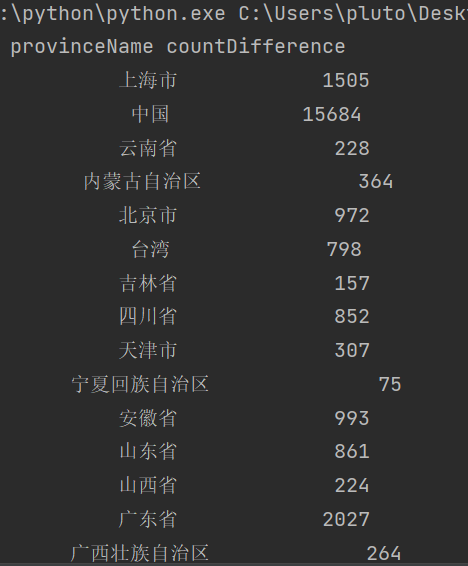


图2-2 2020年累计确诊人数数据预览

**通过以上相同步骤我们可以处理得到在不同年份的治愈人数的各省份统计**



图2-3 2020年累计治愈人数统计预览

**接下来我们会从网上爬取中国的行政区划地图的 Shapefile 文件。Shapefile 是一种常见的地理信息数据格式，它由多个文件组成，包括 .shp、.shx、.dbf 等。其中，.shp 文件存储地理几何数据，.dbf 文件存储属性数据，.shx 文件是索引文件。在读取 Shapefile 时，只需要指定 .shp 文件的路径即可，读取函数会自动关联其他相关文件。通过以上数据统计处理,为接下来的数据可视化提供了坚实的基础**

**3、数据处理结果可视化**

**我们通过以上数据的处理,绘制出了柱状图,折线图,柱状折线图,雷达图,饼状图以及地理热力图,以下为我们绘制数据可视化图的过程:**

**3.1 柱状及折线图**

**柱状图（Bar Chart）是一种常用的数据可视化图表，用于展示分类数据的数量或频率。柱状图的基本构成包括横轴（X轴）和纵轴（Y轴）。横轴表示不同的分类变量，如时间、地区或其他类别，纵轴表示对应分类变量的数量或频率。**

**在柱状图中，每个分类变量对应一个矩形条（柱子），其高度表示该分类的数量或频率。柱子的宽度一般保持一致，可以根据需求进行调整。柱状图常用于比较不同分类之间的数据大小或呈现数据随时间或其他变量的变化趋势。**

**折线图（Line Chart）是一种展示数据随着变量的变化而呈现的趋势的图表。折线图也由横轴（X轴）和纵轴（Y轴）构成，横轴表示自变量，纵轴表示因变量。**

**在折线图中，数据点按照自变量的顺序连接起来，形成一条或多条折线。每个数据点表示对应自变量取值时的因变量的数值。通过连接数据点的折线，可以观察数据随自变量的变化而呈现的趋势，如上升、下降、波动等。**

**折线图常用于显示时间序列数据的变化趋势，可以清晰地观察到数据的趋势和周期性变化。折线图还可以用于比较不同组或不同条件下的数据变化情况，以便观察它们之间的差异和关系。**

**在新冠疫情数据分析中，柱状图可以用来比较不同省市的累计确诊人数、治愈人数和疑似症状人数，以展示不同地区之间的差异。折线图可以用来显示这些数据随时间的变化趋势，如疫情的爆发、峰值和下降趋势，以便了解疫情的发展情况和变化趋势。**

**首先我们先根据所处理过后的数据按照年份的排序,分为了2020年,2021年,2022年。分别统计在这三年时间当中中国的确诊人数的增加量对比**

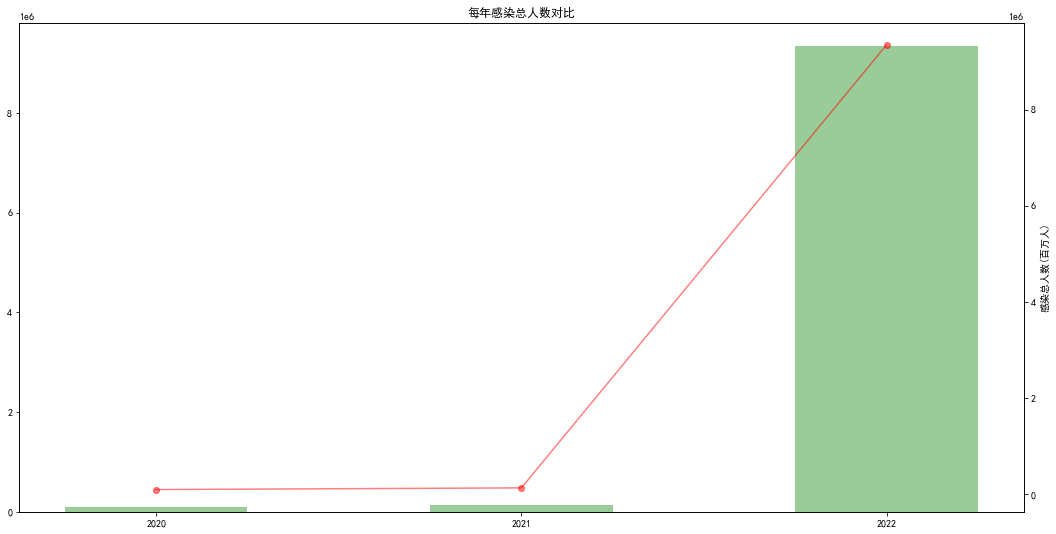


图3-1 中国2020年-2022年确诊人数对比

**其中我们对三年中各省市的确诊人数进行统计并绘制柱状图:**

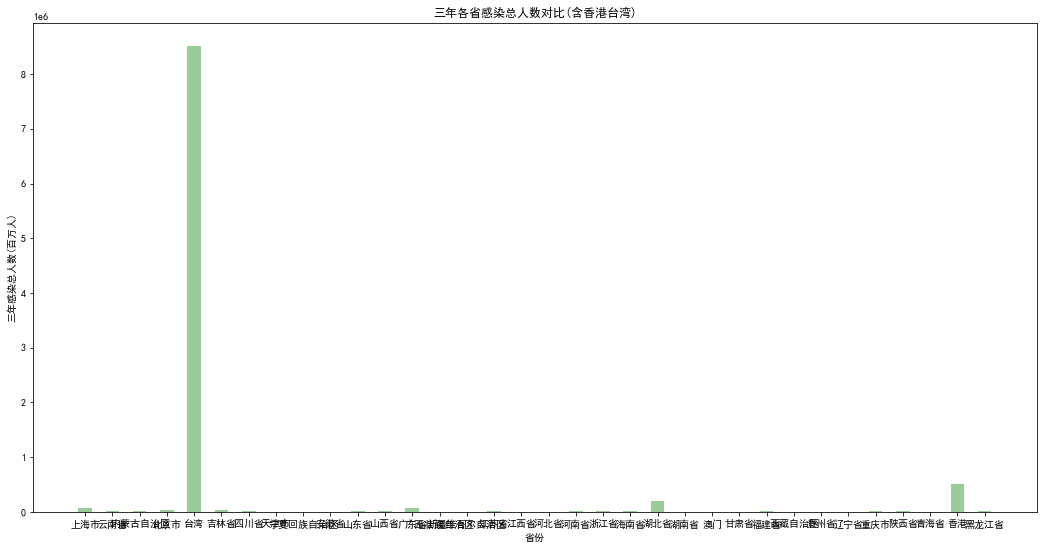
****

图3-2 各省市2020年-2022年确诊人数对比(含香港台湾)

**其中我们通过对比发现,由于香港台湾的防疫政策与大陆有着很大的不同,因此在确诊人数的统计上与大陆的各个省市有着非常大的差距,为了确保我们能够更加直观的了解到大陆各个省市的确诊人数柱状图,我们经过慎重决定,再制作一版数据,将香港台湾暂时先不考量其数据,:**

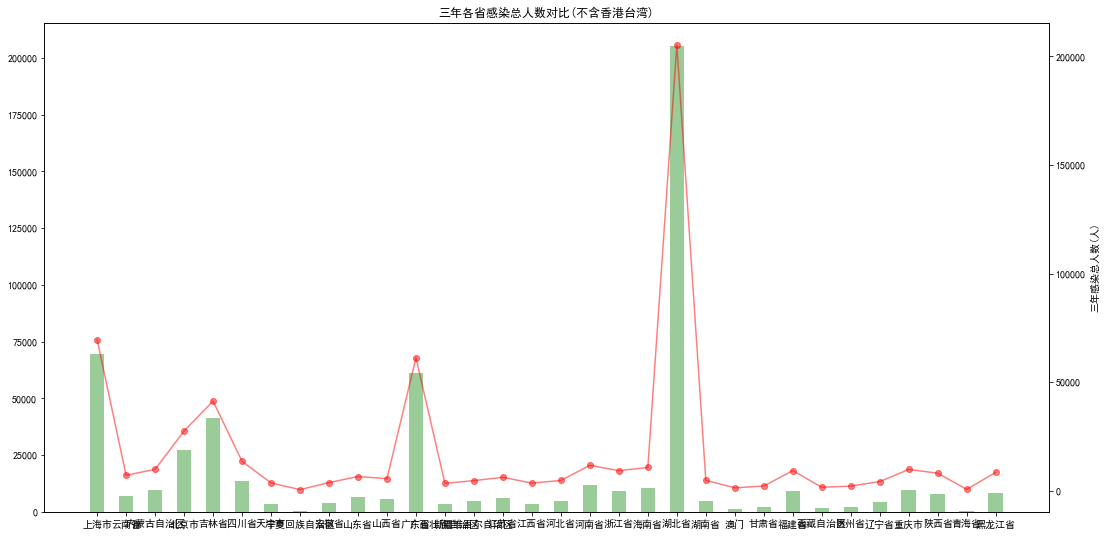
****

图3-3 各省市2020年-2022年确诊人数对比(不含香港台湾)

**绘制完成三年的数据总和之后,我们再分别统计2020年,2021年,2022年的数据,将时间缩短至一年来进行分析:**

**2020年:**

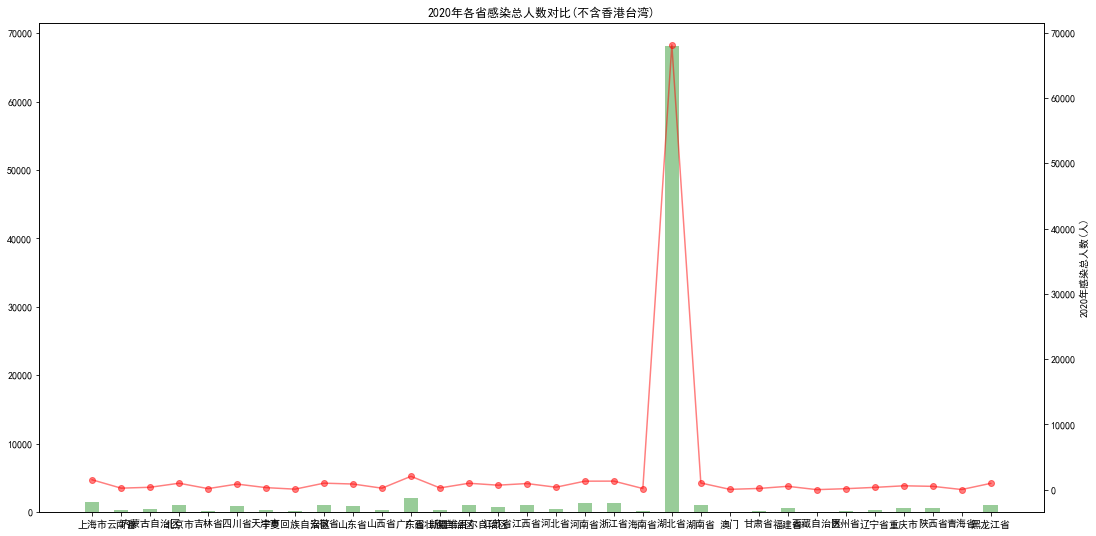
****

图3-4 各省市2020年确诊人数对比(不含香港台湾)

**2021年:**

**文本

描述已自动生成**

图3-5 各省市2021年确诊人数对比(不含香港台湾)

**2022年:**

**图表, 直方图

描述已自动生成**

图3-6 各省市2022年确诊人数对比(不含香港台湾)

**在最后我们还统计的各省市年份的确诊人数变化量对比的:**

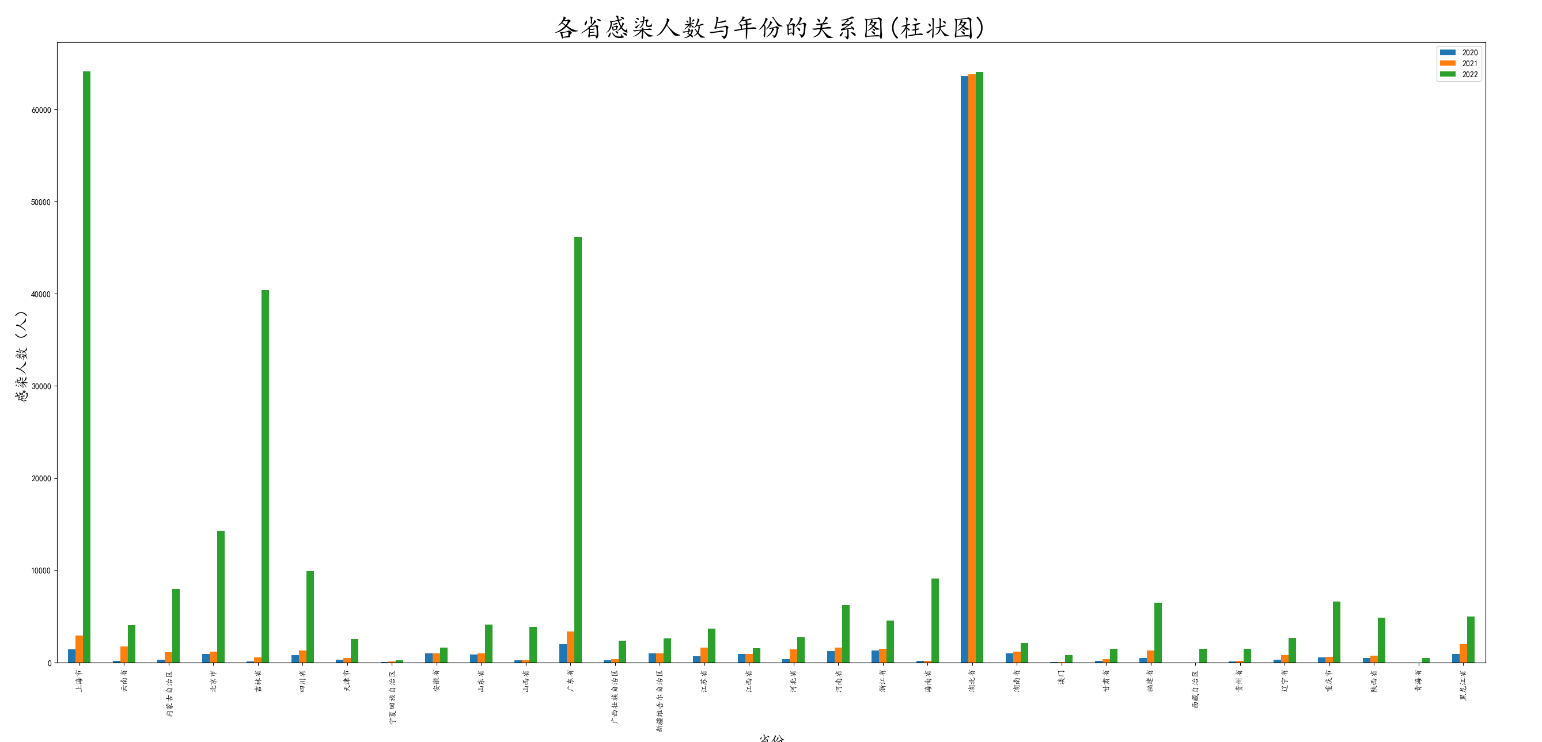
****

图3-7 各省市确诊人数与年份的关系对比(不含香港台湾)

**3.2 雷达图**

**雷达图，也称为极坐标图或蛛网图，是一种用于显示多个变量之间关系的图表形式。它以一个中心点为基准，将多个变量的取值用射线或线段表示，形成一个多边形或闭合图形。**

**在雷达图中，每个变量对应图表中的一个轴，这些轴平均分布在图表的圆周上。每个轴代表一个特定的变量，并沿着该轴的方向表示变量的取值范围。变量的取值可以通过线段的长度、面积或颜色来表示。**

**雷达图的主要特点是可以同时比较多个变量之间的相对大小和关系。通过观察各个变量在雷达图中所形成的多边形的形状和大小，可以直观地了解各个变量的相对重要性和差异程度。**

**雷达图常用于以下情况：**

**1. 多变量比较：当需要比较多个变量的取值时，雷达图可以将各个变量的取值以直观的方式展示出来，帮助观察者理解各个变量之间的差异和关系。**

**2. 总体评估：通过观察雷达图中形成的多边形的形状，可以对某个对象或个体在各个变量上的表现进行总体评估。例如，在评估一个产品的性能时，可以将不同性能指标作为雷达图的变量，从而快速了解产品在各个指标上的表现情况。**

**3. 目标设定：雷达图可以帮助设定目标或指标，将目标与实际观测值进行对比。通过比较目标值和实际观测值之间的差异，可以识别出存在的问题，并采取相应的改进措施。**

**总结而言，雷达图是一种有效的可视化工具，适用于多变量比较、总体评估和目标设定等情况。它提供了一种直观的方式来展示和理解多个变量之间的关系和差异，帮助人们做出相应的决策和改进。**

**为了**

**我们通过雷达图承载以上数据绘制而成可视化图像:**

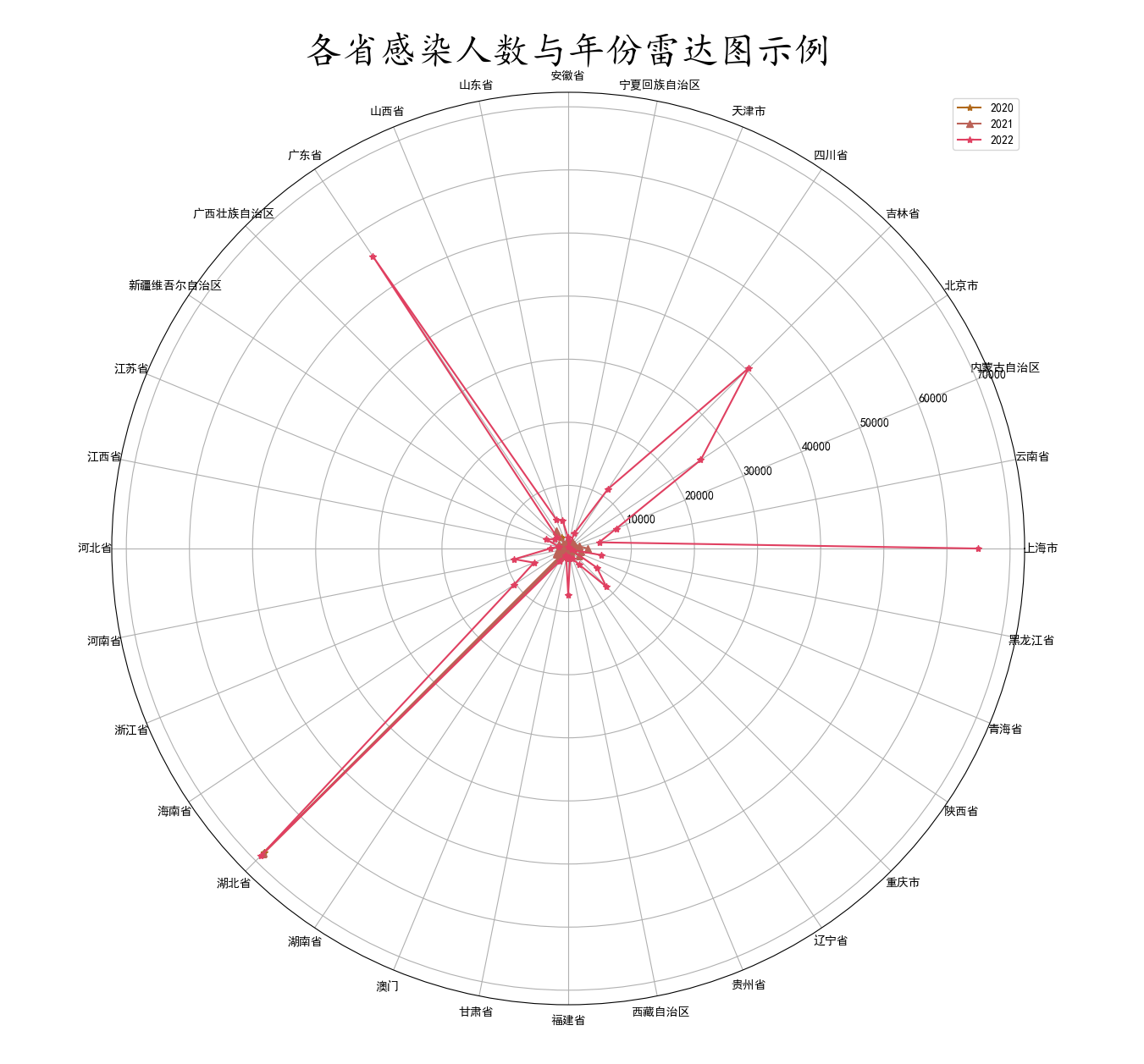
****

图3-7 各省市确诊人数与年份雷达图示例 (不含香港台湾)

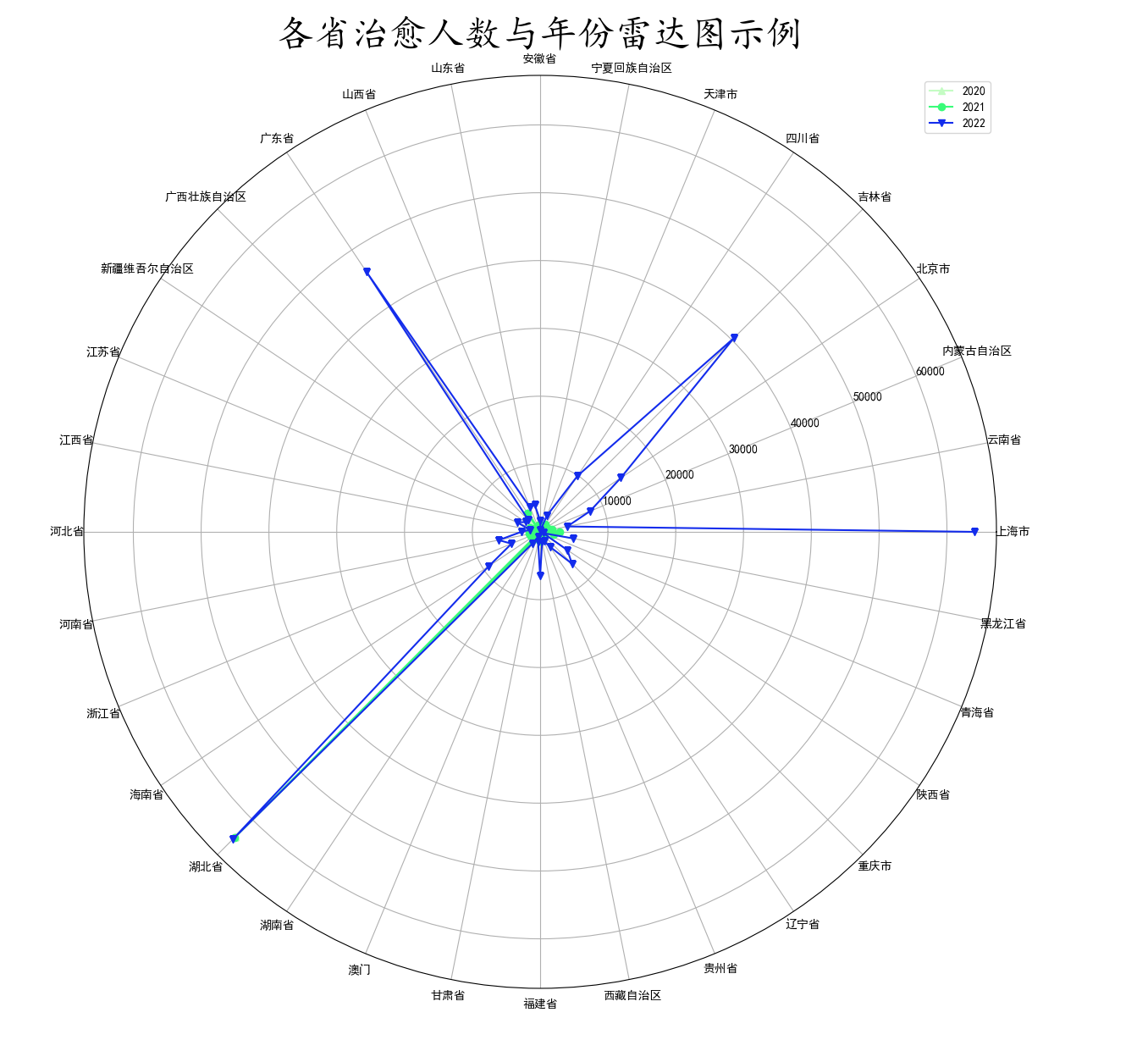
****

图3-8 各省市治愈人数与年份雷达图示例 (不含香港台湾)

**3.3饼状图**

**饼状图是一种常用的数据可视化图表，用于展示数据中各个部分相对于整体的比例关系。它将整体数据分成多个扇形区域，每个扇形区域的大小表示该部分所占比例的大小。饼状图通常用于展示分类数据，例如不同类别的销售份额、各个部门的预算分配比例等。**

**饼状图的主要特点和用途如下：**

**1. 比例关系展示：饼状图直观地展示了数据中各个部分相对于整体的比例关系。通过比较各个扇形区域的大小，可以快速了解不同部分的重要性和占比情况。**

**2. 类别对比：饼状图可以用于展示分类数据的对比情况。不同类别的扇形区域大小的差异可以直观地显示出各个类别之间的差异程度。**

**3. 相对关系呈现：饼状图还可以显示数据部分之间的相对关系。通过扇形区域的角度，可以比较各个部分之间的相对大小，进一步强调数据的差异和比例关系。**

**4. 可视化传达信息：饼状图以直观的方式传达数据的重要信息，使观众更容易理解和记忆。它常用于演示报告、商业展示和新闻报道等场景，以便清晰地传达数据的关键见解。**

**需要注意的是，在使用饼状图时，应确保数据的准确性和可比性。同时，应避免在饼状图中使用过多的扇形区域，以免造成视觉混乱和信息过载。此外，饼状图并不适用于展示过多的类别或呈现连续数据。**

**综上所述，饼状图是一种常见的数据可视化图表，通过展示数据中各个部分相对于整体的比例关系，提供了直观、清晰的数据展示方式，有助于观众快速理解和分析数据的结构和关系。**

**我们首先绘制了有关确诊人数的饼状图:**

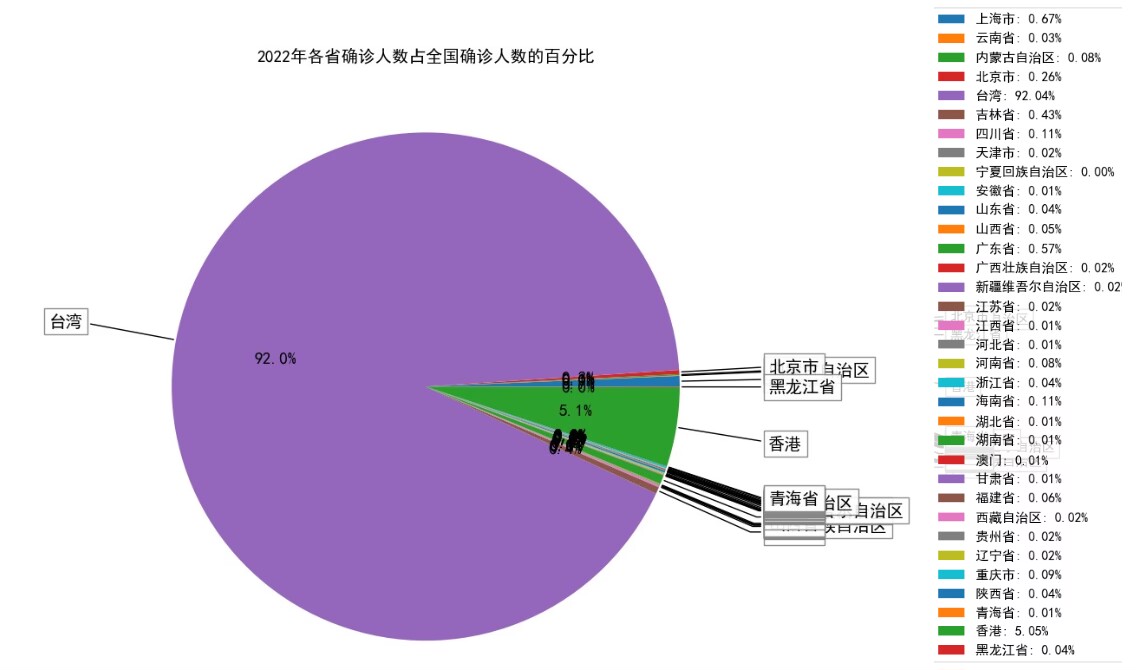
****

图3-9 2022年各省市确诊人数占全国确诊人数百分比(含香港台湾)

**图表, 饼图

描述已自动生成**

图3-10 2021年各省市确诊人数占全国确诊人数百分比(含香港台湾)

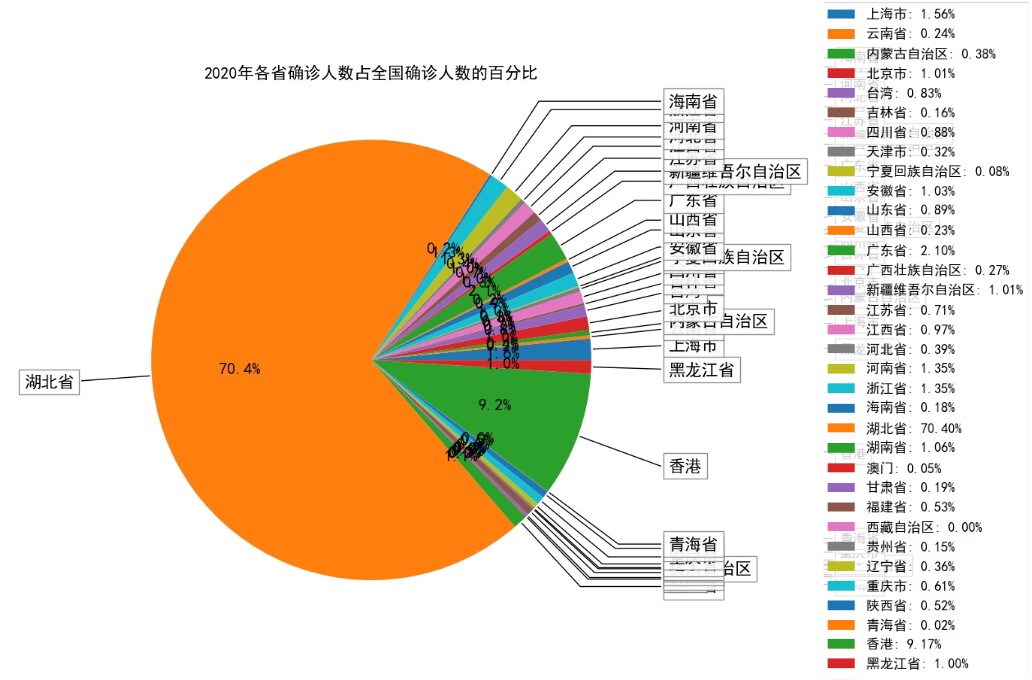
****

图3-11 2020年各省市确诊人数占全国确诊人数百分比(含香港台湾)

**图表, 饼图

描述已自动生成**

图3-12 2022年各省市确诊人数占全国确诊人数百分比(不含香港台湾)

图表, 饼图

描述已自动生成

图3-13 2021年各省市确诊人数占全国确诊人数百分比(不含香港台湾)

图表, 饼图

描述已自动生成

图3-14 2020年各省市确诊人数占全国确诊人数百分比(不含香港台湾)

**图表, 雷达图

描述已自动生成**

图3-15 2020年各省市确诊人数占全国确诊人数百分比(除香港台湾湖北)

**此外,我们还绘制了有关治愈人数的饼状图:**

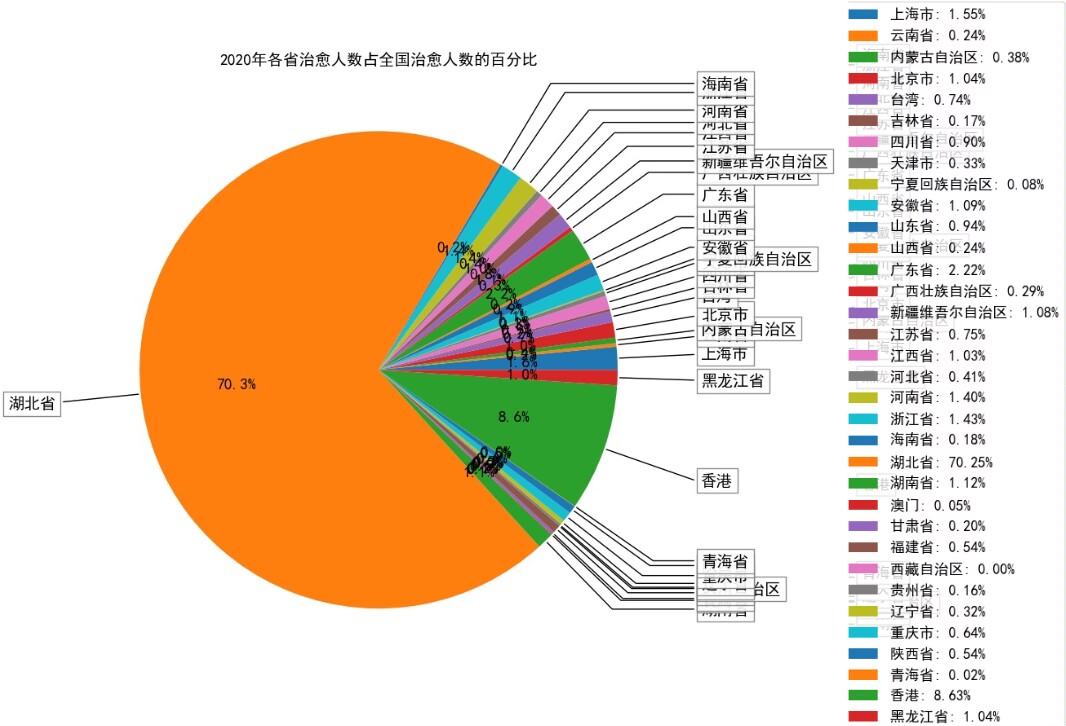
****

图3-16 2020年各省市治愈人数占全国治愈人数百分比

**图表, 饼图

描述已自动生成**

图3-17 2021年各省市治愈人数占全国治愈人数百分比

**图表

描述已自动生成**

图3-18 2022年各省市治愈人数占全国治愈人数百分比

**3.4 地理热力图**

**地理热力图是一种可视化方法，用于显示地理区域上的数据强度或密度分布。在基于新冠疫情数据的传染病区域地理性分析项目中，地理热力图可以用于以下方面的应用：**

**1. 疫情传播分析：通过将每个地理位置上的新冠病例数量映射到地图上的不同颜色或渐变，地理热力图可以清晰地展示疫情的传播情况和热点区域。这可以帮助决策者和研究人员了解疫情在不同地区的传播速度、范围和密度，以便制定相应的防控策略和资源分配。**

**2. 风险评估和预测：通过分析新冠疫情数据并生成地理热力图，可以识别高风险区域和潜在爆发风险。地理热力图可以显示不同地区的感染人数、疑似症状人数和治愈人数的分布情况，帮助评估各地区的风险水平，并为预测未来的传播趋势提供依据。**

**3. 资源调配和决策支持：基于地理热力图的分析结果，可以帮助决策者更好地理解疫情在地理空间上的分布特征，从而更准确地分配防控资源和制定决策。地理热力图可以突出显示疫情的热点区域和高风险地区，帮助决策者重点关注和采取措施。**

**4. 公众沟通和意识提高：地理热力图以直观的方式呈现疫情数据，能够更好地与公众沟通和提高意识。公众可以通过地理热力图了解本地区的疫情情况，从而采取适当的防护措施和行为，同时也能够更好地理解疫情的整体态势和趋势。**

**综上所述，地理热力图在基于新冠疫情数据的传染病区域地理性分析项目中具有重要的应用价值，能够帮助揭示疫情传播模式、辅助决策制定、预测风险和加强公众意识。**

**我们结合中国行政区划以及各省市确诊人数数据绘制出地理热力图:**

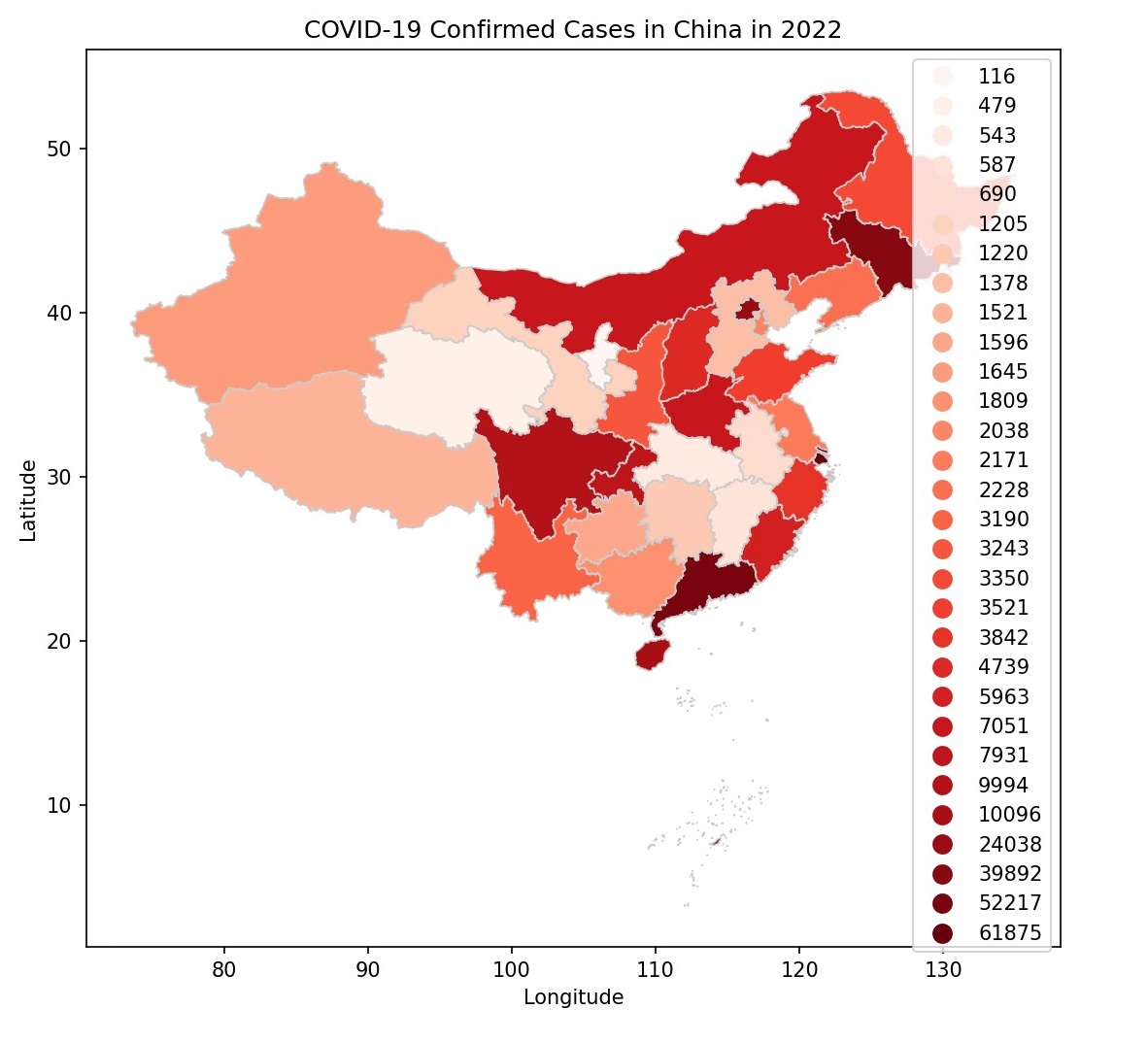
****

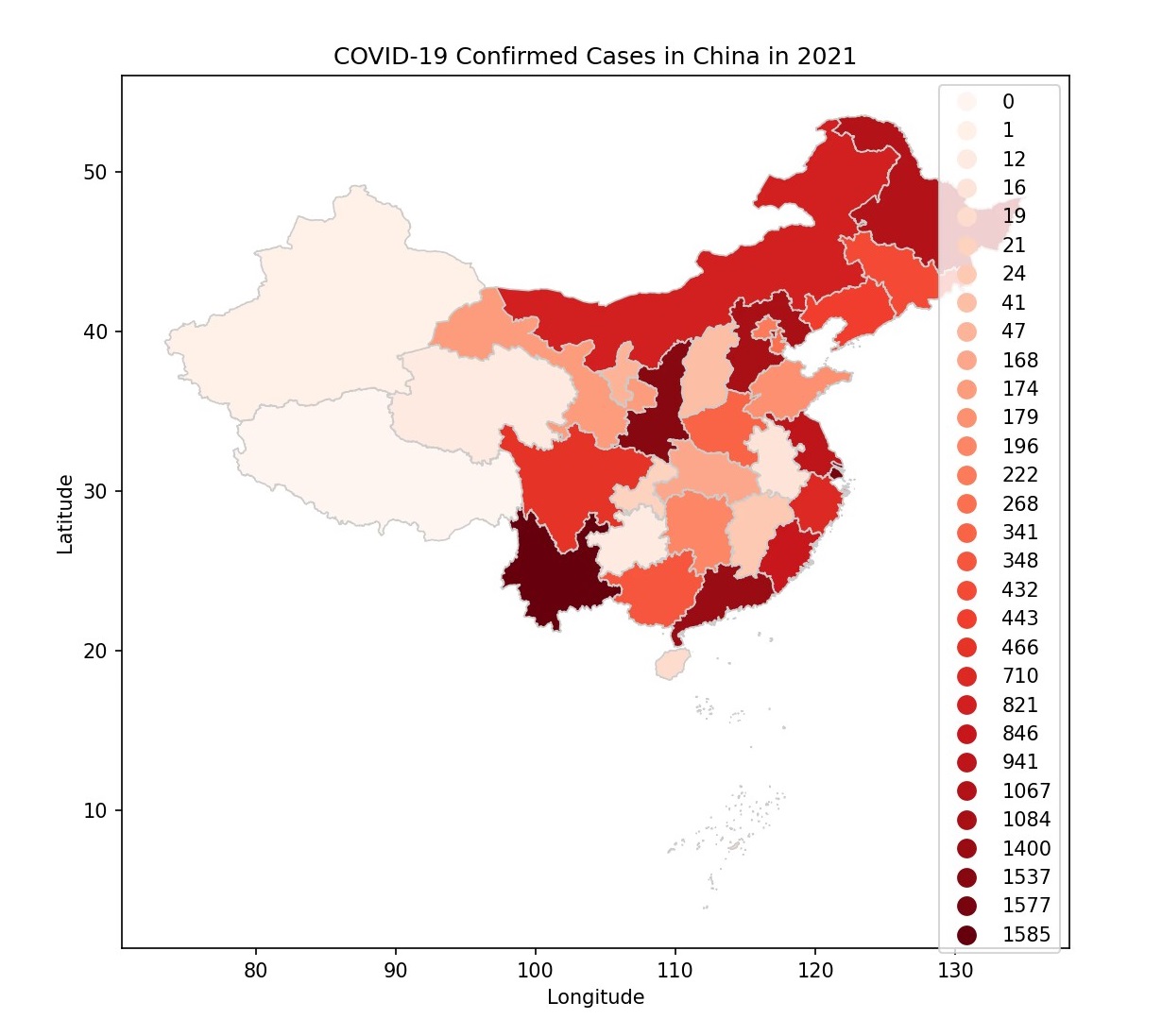
图3-19 2022年各省市确诊人数热力图

图3-20 2021年各省市确诊人数热力图

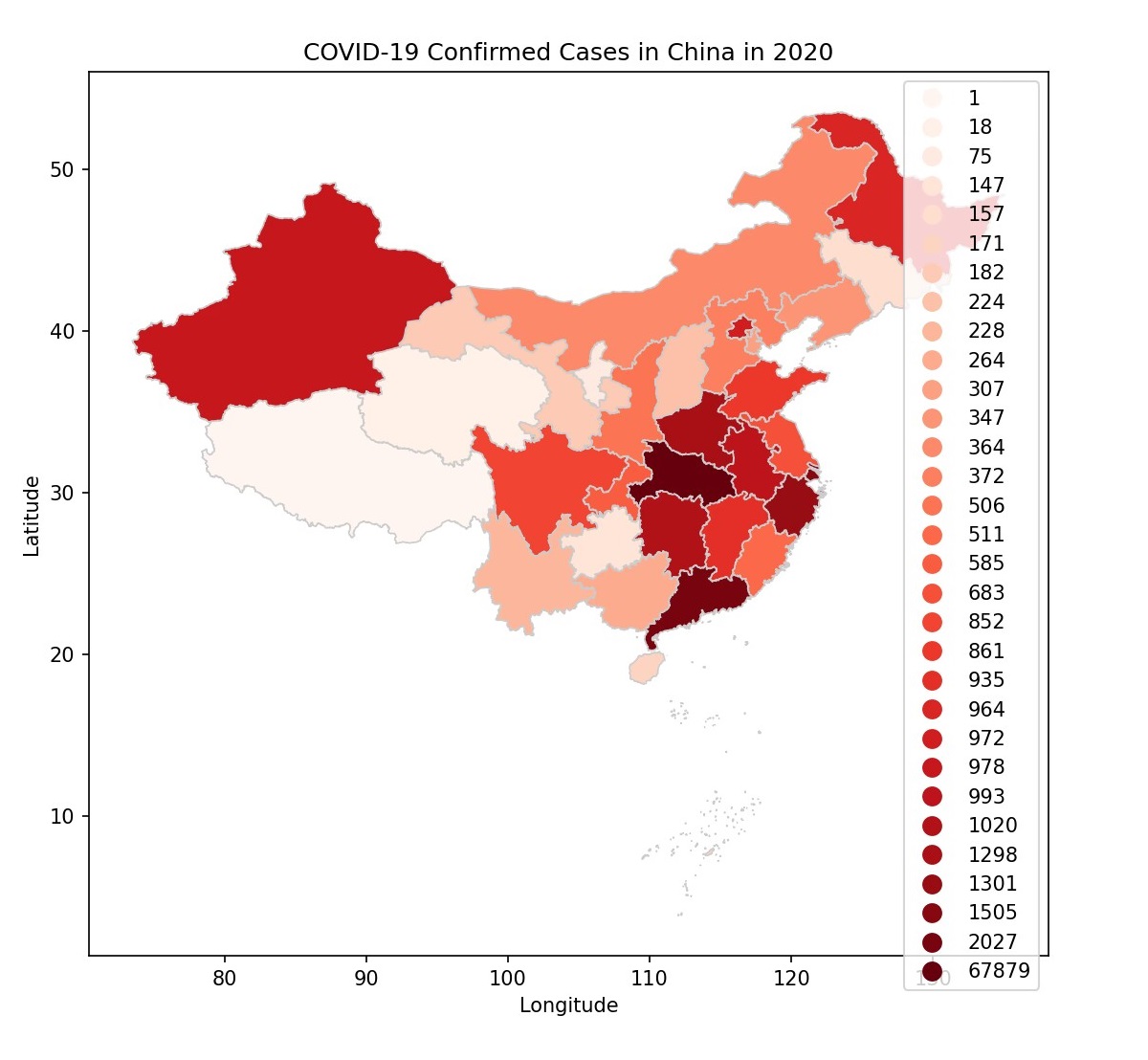


图3-21 2020年各省市确诊人数热力图

**4、数据处理分析的结论或对数据可视化的剖析**

**通过以上数据处理以及可视化分析,以下提出我们的总结报告:**

1.疫情变化趋势:  
 1.1.年度变化:

**比较不同年份的数据，可以观察到每年疫情的整体趋势**

**首先以下是对每年的确诊人数的总结：**

**2020年：**

**- 确诊人数最多的省份是湖北省，达到了高峰，确诊人数为 67,879 人。**

**- 其他高风险地区包括广东省、北京市、上海市和河南省等。**

**- 台湾、香港和澳门作为境外输入和边境地区，也有较高的确诊人数。**

**- 不同地理区域之间存在明显的疫情分布差异。**

**2021年：**

**- 确诊人数最多的省份是台湾，达到了高峰，确诊人数为 16,227 人。**

**- 其他高风险地区包括广东省、北京市、上海市和河南省等。**

**- 台湾、香港和澳门仍然是境外输入和边境地区，确诊人数较高。**

**- 疫情在不同地理区域之间的分布差异持续存在。**

**2022年：**

**- 确诊人数最多的省份是台湾，确诊人数为 8,481,145 人。**

**- 其他高风险地区包括广东省、北京市、上海市和河南省等。**

**- 台湾、香港和澳门仍然是境外输入和边境地区，确诊人数较高。**

**- 不同地理区域之间的疫情分布差异仍然存在。**

1.2.区域比较:

**2020年：**

**- 高风险地区：湖北省是疫情爆发初期的重灾区，拥有最高的确诊人数。广东省、北京市和上海市也是高风险地区。**

**- 部分低风险地区：一些省份如西藏自治区、青海省和内蒙古自治区的确诊人数相对较低。**

**2021年：**

**- 高风险地区：台湾地区、香港和广东省是确诊人数较高的地区。北京市在2021年达到高峰后有所下降。**

**- 部分低风险地区：西藏自治区、青海省和宁夏回族自治区的确诊人数相对较低。**

**2022年：**

**- 高风险地区：台湾地区、香港和广东省继续保持较高的确诊人数。北京市和上海市的确诊人数也较多。**

**- 部分低风险地区：西藏自治区、青海省和甘肃省的确诊人数相对较低。**

2.疫情高风险区域:

2.1.确诊人数比较:

**2020年：**

**- 最高确诊人数：湖北省在2020年拥有最高的确诊人数，达到了惊人的67879人。其次是广东省和北京市，分别有2027人和972人的确诊人数。**

**- 相对较低确诊人数：一些省份如西藏自治区、青海省和宁夏回族自治区的确诊人数相对较低，分别为1人、18人和75人。**

**2021年：**

**- 最高确诊人数：台湾地区在2021年成为确诊人数最高的地区，达到了惊人的16227人。其次是香港和广东省，分别有8846人和1400人的确诊人数。**

**- 相对较低确诊人数：一些省份如新疆维吾尔自治区、西藏自治区和宁夏回族自治区的确诊人数相对较低，分别为1人、0人和47人。**

**2022年：**

**- 最高确诊人数：台湾地区在2022年仍然是确诊人数最高的地区，达到了惊人的8481145人。其次是广东省和北京市，分别有52217人和24038人的确诊人数。**

**- 相对较低确诊人数：一些省份如西藏自治区、青海省和甘肃省的确诊人数相对较低，分别为1521人、479人和1205人。**

2.2.疫情爆发时间和持续时间:

**疫情爆发时间：**

**- 2020年：疫情在湖北省首先爆发，该地区成为疫情的重灾区。其他地区如北京市、广东省等也在早期出现了疫情爆发的迹象。**

**- 2021年：疫情在台湾地区、香港和澳门等地出现较高的确诊人数，显示出疫情的再次爆发。**

**- 2022年：台湾地区仍然是疫情爆发的高风险地区，同时广东省和北京市也出现了较高的确诊人数。**

**疫情持续时间：**

**- 2020年：湖北省的疫情持续时间较长，从疫情爆发到逐渐控制疫情花费了较长的时间。其他地区如北京市、广东省也经历了一段时间的疫情持续期。**

**- 2021年：台湾地区、香港和澳门等地在2021年持续了相当长的时间，尽管疫情在某些地区有所缓解，但仍然存在较高的确诊人数。**

**- 2022年：台湾地区的疫情持续时间较长，广东省和北京市等地也经历了相当长的疫情持续期。**

**总体而言，不同地区的疫情爆发时间和持续时间有所差异。一些地区在早期或后期出现了疫情的爆发，而一些地区经历了较长的疫情持续期。**

2.3.疫情防控措施:

**通过查阅网络以及我们所收集到的数据,我们了解到在疫情高风险地区政府采取了以下防控措施:**

**1. 区域封控措施：在疫情爆发初期，一些重点地区如湖北省采取了封控措施，对疫情严重的城市进行封锁和限制人员流动，以遏制病毒传播。**

**2. 社交隔离和限制措施：在各个省市区域内，疫情期间实施了社交隔离和限制措施，包括封闭小区、限制公共聚集、暂停学校和企事业单位等的正常运作，以减少人员流动和接触，阻断病毒传播链。**

**3. 公共场所管控：为了防止疫情扩散，各个省市区域采取了严格的公共场所管控措施，包括对商场、餐馆、娱乐场所等的限制和规范，要求人员佩戴口罩、测量体温，并保持社交距离。**

**4. 疫苗接种推广：疫苗接种被广泛推广，各地积极组织疫苗接种工作，设立接种点，加强疫苗供应和宣传，鼓励民众主动接种疫苗，以提高群体免疫力。**

**5. 指挥体系和信息公开：各个省市区成立了疫情指挥部或类似机构，负责协调疫情防控工作，及时发布疫情信息和相关政策，加强与公众的沟通和信息公开。**

**6. 医疗资源调配：在疫情严重地区，加强了医疗资源的调配和支援，派遣医疗队到疫情重点地区支援医疗工作，增加床位、医疗设备和物资供应，提高医疗能力和救治水平。**

2.4.疫情高风险区域的人口迁徙和交通情况:

**我们通过查阅资料和分析数据得到以下结论:**

**1. 人口迁徙：在疫情高风险区域，特别是边境地区和经济发达地区，人口流动较为频繁。这可能是由于该地区的经济活动、教育、就业等因素吸引了大量人口。**

**2. 境外输入：疫情高风险区域通常与外界有密切的交流和移民流动。数据显示，台湾、香港和澳门在三年中的确诊人数较高。这些地区与其他国家和地区有较为紧密的联系，因此可能存在较高的境外输入风险。**

**3. 交通情况：疫情高风险区域通常具有发达的交通网络，包括航空、铁路、公路等交通方式。这使得人员的跨地区和跨国家移动更加便捷。特别是国际机场、火车站等交通枢纽地区可能成为疫情传播的重要场所。**

**4. 边境地区：边境地区通常具有特殊的人口流动和交通情况。对于处于边境地区的省市，特别需要关注边境口岸的交通状况和管控措施，以防止疫情通过边境传播。**

**综上所述，疫情高风险区域的人口迁徙和交通情况具有一定的特点，包括人口流动频繁、境外输入风险较高、交通网络发达等。在疫情防控工作中，需要针对这些特点采取相应的措施，加强边境管控、加强交通枢纽的监测和防控，以减少疫情的传播风险。**

3. 地理分布差异:

3.1.行政省市区划:

**我们通过查阅资料和分析数据得到以下结论:**

**1. 疫情高风险地区集中分布：根据数据显示，广东省和北京市在三年中一直是高风险地区。这与它们作为人口密集、经济发达的地区有关。同时，湖北省在疫情爆发初期成为重灾区，这可能与其作为疫情起源地和交通枢纽有关。**

**2. 区域间差异明显：不同行政省市区划之间的疫情表现存在明显的差异。一些偏远地区如西藏自治区和青海省的确诊人数相对较低，这可能与其相对封闭的地理位置和较少的人口流动有关。而经济发达地区和交通枢纽地区的确诊人数相对较高，这与人口流动、密集度以及交通网络的发达程度有关。**

**3. 境外输入和边境地区的影响：台湾、香港和澳门在三年中的确诊人数较高，这些地区通常与外界有密切的交流和移民流动。同时，边境地区如内蒙古自治区和新疆维吾尔自治区的确诊人数也相对较高，这可能与其与邻国接壤、边境口岸繁忙等因素有关。**

**4. 行政省市区划的重要性：中国的行政省市区划在疫情防控中扮演着重要的角色。不同行政区划具有不同的地理、经济和人口特点，需要根据其特点制定相应的防控策略和措施。同时，行政区划之间的协调和合作也至关重要，以实现跨地区的信息共享、资源协调和联防联控。**

**综上所述，地理分布差异中的行政省市区划在疫情分析中具有重要作用。不同行政区划之间存在着明显的疫情表现差异，地理位置、人口流动、经济发展和交通网络等因素都对疫情的传播和防控产生影响。因此，针对不同行政区划的特点，制定相应的疫情防控策略和措施，加强区域间的合作与协调，对于有效控制疫情的传播至关重要。**

3.2 确诊人数比较:

**我们通过查阅资料和分析数据得到以下结论:**

**1. 高风险地区集中分布：根据数据显示，广东省和北京市在三年中一直是高风险地区，这意味着这两个地区的确诊人数相对较高。广东省是中国人口最多的省份之一，且经济发达，人口流动性较高，这可能导致疫情传播的风险增加。北京市作为中国的首都和国际交通枢纽，也面临着较高的人口流动和感染风险。**

**2. 疫情爆发初期的重灾区：湖北省在2020年是疫情爆发的重灾区，这可能与其作为新冠病毒起源地的武汉市有关。湖北省在疫情初期面临着大规模感染和传播，导致了较高的确诊人数。**

**3. 边境地区的影响：根据数据显示，边境地区如内蒙古自治区和新疆维吾尔自治区的确诊人数相对较高。这些地区与邻国接壤，边境口岸繁忙，可能存在较高的境外输入风险和边境地区传播的挑战。**

**4. 人口稠密地区的影响：一些人口密集的地区如上海市、广东省和北京市在三年中的确诊人数较高。这些地区人口众多，人口流动性强，这可能增加了病毒传播的机会。**

**5. 偏远地区的相对较低确诊人数：一些偏远地区如西藏自治区和青海省在三年中的确诊人数相对较低。这些地区人口相对较少，且地理位置偏僻，人口流动性较低，可能有利于控制疫情的传播。**

**综上所述，根据中国的行政省市区划和数据，我们可以观察到地理分布差异中的确诊人数比较。不同地区的确诊人数受到人口密集度、人口流动性、地理位置和经济发展等因素的影响。高风险地区主要集中在人口密集、经济发达和交通枢纽地区，而偏远地区的确诊人数相对较低。这些观察可以为地方政府和相关机构制定疫情防控策略提供参考。**

3.3.地理特点分析:

**我们通过查阅资料和分析数据得到以下结论:**

**1. 沿海地区：广东省、上海市和福建省等沿海地区在三年中的确诊人数相对较高。这些地区地处中国的东部沿海，拥有重要的港口和国际交通枢纽，人口流动性较高，可能容易受到境外输入的影响，同时也面临着较高的感染风险。**

**2. 内陆省份：内蒙古自治区、四川省和陕西省等内陆省份在三年中的确诊人数相对较低。这些地区位于中国的内陆地区，地理相对封闭，人口流动性相对较低，有利于疫情的控制和防控工作的实施。**

**3. 边境地区：新疆维吾尔自治区、西藏自治区和内蒙古自治区等边境地区在三年中的确诊人数相对较高。这些地区与邻国接壤，人口流动性和交流频繁，可能面临着较高的境外输入风险和边境地区传播的挑战。**

**4. 高山地区：青海省和西藏自治区等高山地区在三年中的确诊人数相对较低。这些地区地理环境特殊，海拔较高，气候寒冷，人口稀少，相对封闭，这可能有助于限制疫情的传播。**

**5. 首都和直辖市：北京市和上海市作为中国的首都和直辖市，人口众多，经济发达，国际交流频繁，因此在三年中的确诊人数相对较高。这些地区面临着较高的人口流动性和感染风险，需要加强疫情防控工作。**

**综上所述，根据中国的行政省市区划和数据，我们可以观察到地理分布差异中的地理特点。不同地区的地理环境、人口流动性和经济发展等因素会影响疫情的传播和控制情况。这些地理特点的分析有助于地方政府和相关机构针对不同地区的特点，制定针对性的疫情防控策略。**

3.4.防控策略制定:

**我们通过查阅资料和分析数据得到以下结论:**

1. 高风险地区防控策略：针对高风险地区，如湖北省、台湾地区和香港等，需要采取严格的防控措施。这包括封控疫情爆发的核心区域，加强病毒检测和筛查，提供充足的医疗资源和支持，加强社区防控，限制人员流动等。此外，还需要加强与境外的交流合作，加强边境管控，严防境外输入。

2. 沿海地区防控策略：沿海地区如广东省、上海市和福建省等，由于其交通和人口流动性较高，需要采取措施防止境外输入和传播。这包括加强海关和口岸的检疫和防控措施，严格筛查入境人员，加强航空、海运和陆路交通的防控，加强社区监测和管控等。

3. 内陆省份防控策略：内蒙古自治区、四川省和陕西省等内陆省份由于地理位置相对封闭，人口流动性较低，疫情传播的风险相对较低。针对这些地区，可以加强早期预警和监测体系，加强社区和机构的防控能力，加强病例的追踪和隔离措施，提高公众的防护意识等。

4. 边境地区防控策略：边境地区如新疆维吾尔自治区、西藏自治区和内蒙古自治区等，需要加强边境的管控和监测，防止境外输入的风险。这包括加强边境口岸的检疫和筛查，加强边境地区的监测和隔离措施，加强边境地区人员流动的管控等。

5. 首都和直辖市防控策略：作为人口密集、交通便利的地区，首都和直辖市如北京市和上海市需要采取更加严格的防控策略。这包括加强社区防控，提高人员的自我保护意识，加强公共场所的卫生和消毒措施，加强病例的追踪和隔离措施，加强航空、铁路和公路等交通工具的防控措施等。

综上所述，根据中国的行政省市区划和数据，不同地区的地理特点和疫情传播情况不同，需要制定相应的防控策略。这些策略可以针对不同地理特点和疫情风险，加强边境管控、加强早期预警和监测、加强社区防控、提高公众的防护意识等，以控制疫情的传播和减少风险。

4. 境外输入和边境地区:

4.1. 境外输入风险:

**我们通过查阅资料和分析数据得到以下结论:**

**1. 高风险地区：根据数据显示，台湾、香港和澳门在三年中的确诊人数较高。这些地区通常与外界有密切的交流和移民流动，因此境外输入风险相对较高。**

**2. 交通流动性：这些地区的国际机场、港口和陆路口岸通常是人员和货物流动的重要通道。大量的国际旅客和货物运输可能带来病毒传播的风险。**

**3. 旅行史和接触史：境外输入的风险主要与入境人员的旅行史和接触史有关。如果入境人员来自疫情高风险地区或与确诊病例有过接触，他们可能成为潜在的传播源。**

**4. 境外输入防控措施：为了应对境外输入风险，中国采取了一系列严格的防控措施。这包括加强边境的检疫和筛查，要求入境人员进行核酸检测，实施隔离措施等。此外，还加强了边境地区的监测和管控，防止疫情跨境传播。**

**5. 国际合作：境外输入风险需要国际合作来共同应对。中国与其他国家和地区加强信息交流、数据共享和合作，共同制定防控策略，加强边境卫生合作，以减少境外输入的风险。**

**综上所述，台湾、香港和澳门等地在三年中的确诊人数较高，这提示境外输入风险的存在。为了应对境外输入风险，我们应该采取了严格的防控措施，包括加强边境的检疫和筛查、要求入境人员进行核酸检测、实施隔离措施等，并加强国际合作来共同应对这一挑战。**

4.2.边境地区防控措施:

**我们通过查阅资料和分析数据得到以下结论:**

**1. 边境地区的关注：根据数据显示，西藏自治区和青海省的确诊人数相对较低，这些地区通常是中国的边境地区。由于接近邻国或地区，边境地区对于防控境外输入具有特殊的重要性。**

**2. 边境管控：中国边境地区采取了严格的防控措施，包括加强边境口岸的检疫和筛查，要求入境人员进行核酸检测，并实施隔离措施。这些措施旨在及时发现和隔离潜在的感染者，减少病毒传播的风险。**

**3. 边境监测：边境地区加强了对入境人员和货物的监测和管控。通过加强边境口岸的检查和监测，可以及时发现携带病毒的人员或物品，并采取相应的防控措施。**

**4. 边境合作：中国与邻国或地区开展边境卫生合作，加强信息交流、数据共享和合作，共同应对境外输入的风险。这种跨境合作可以提高边境地区的防控能力，共同应对疫情挑战。**

**综上所述，西藏自治区和青海省等中国的边境地区在三年中的确诊人数相对较低。为了加强边境地区的防控，我们应采取了严格的措施，包括加强边境口岸的检疫和筛查、要求入境人员进行核酸检测、实施隔离措施等，并与邻国或地区开展边境卫生合作。这些措施旨在及时发现和隔离潜在的感染者，减少病毒传播的风险，并共同应对疫情挑战。**

4.3 旅行限制和健康检测:

**我们通过查阅资料和分析数据得到以下结论:**

**1. 旅行限制措施：针对境外输入风险，中国采取了一系列旅行限制措施。这些措施包括限制非必要的国际旅行、暂停某些国家或地区的航班、关闭边境口岸等。通过限制人员的流动，可以减少潜在的病毒传播途径，降低疫情的风险。**

**2. 入境健康监测：入境人员通常需要进行健康监测，包括核酸检测和健康申报。这些措施旨在筛查潜在的感染者，减少病毒传播的风险。入境人员可能需要提供核酸检测阴性证明，并接受隔离观察。健康申报表的填写则有助于追踪和监测入境人员的健康状况。**

**3. 健康监测体系：中国建立了健康监测体系，包括健康码和健康信息登记等工具。通过这些系统，可以追踪和监测人员的健康状况，并及时采取相应的防控措施。健康码通常是一种基于个人健康信息的标识，用于识别和管理人员的健康状态。**

**4. 边境卫生合作：中国与邻国或地区开展边境卫生合作，加强信息交流、数据共享和合作，共同应对境外输入的风险。这种合作包括共享旅行者的健康信息和加强边境口岸的协调和监测等。通过加强合作，可以提高边境地区的防控能力，减少疫情传播的风险。**

**综上所述，我们应采取了旅行限制措施、入境健康监测和建立健康监测体系等措施来应对境外输入的风险。这些措施有助于限制人员流动、筛查潜在感染者并追踪入境人员的健康状况。此外，中国还与邻国或地区展开边境卫生合作，共同应对境外输入的挑战。这些措施的实施旨在减少病毒传播的风险，保护公众的健康安全。**

**四、未来展望**

**项目的意义和对未来发展的促进作用可以从以下几个方面进行阐述：**

**1. 疫情防控决策支持：通过基于新冠疫情数据的传染病区域及地域分析，项目能够提供决策者所需的科学依据。准确了解疫情的传播模式、风险地区和趋势，可以帮助政府和卫生机构制定更有效的疫情防控策略。项目的实施将为决策者提供重要的信息，帮助他们做出明智的决策，保护公众的健康和生命。**

**2. 全球疫情管理能力提升：项目的成功实施将促进全球范围内的疫情管理能力的提升。通过共享分析方法、经验和最佳实践，不同国家和地区可以相互学习和借鉴，加强协作与合作。这有助于改善国际间的信息共享和协调行动，提高全球疫情管理的效能，增强全球卫生安全。**

**3. 传染病预防与控制技术发展：项目将利用Python等技术工具对大规模的疫情数据进行深入分析和可视化，推动传染病预防与控制技术的发展。在实施过程中，可能会开发出一些新的数据处理、分析和可视化方法，为疫情数据的应用提供更多可能性。这将对数据科学和公共卫生领域的技术进步起到促进作用。**

**4. 危机管理与应急响应能力提升：项目的开展将有助于提高各国和地区在危机管理和应急响应方面的能力。通过对疫情数据的深入分析，可以及时发现和评估疫情的风险和威胁，从而采取及时和针对性的措施。这将有助于提高卫生系统的应对能力，增强社会对突发公共卫生事件的抵抗力。**

**综上所述，项目的实施对于疫情防控决策支持、全球疫情管理能力提升、传染病预防与控制技术发展以及危机管理与应急响应能力提升都具有重要意义。项目的成功将为未来的传染病防控和公共卫生领域的发展提供宝贵经验，并为应对类似挑战提供有力支持。**

**五、组长对小组成员的点评**

**首先感谢我们组员的对于我们项目的大力支持,如果没有组员的努力,我想我们的项目也不会完成的如此顺利。**

**感谢张青松在我们进行讨论时出的许多主意,这些主意确实的在一定程度上引导着我们的前进,同时他所完成饼状图可视化也做得十分完美,很大程度上为我们的数据分析提供了坚实的基础。**

**感谢刘浩天和刘泽康在我们为方向而苦恼时他们所提出的建议,这些建议为我们的项目完成推了很大的一份力,同时他们所完成的柱状图和线型图也十分完美, 很大程度上为我们的数据分析提供了坚实的基础。**

**感谢刘俊杰在数据分析上为我们提供的帮助,他的努力推动了我们的数据分析能够更快的进行,并且为之后的可视化提供的很多的帮助**

**再次感谢组员,感谢他们的付出和汗水,这是一次十分愉快的项目合作经历**