**目标：**

通过对比全球各国新冠疫情爆发以来的数据，来找到各国疫情之间的相关性，并试图对比它们地理上的相对位置，找到疫情与各国地理位置的相关性。

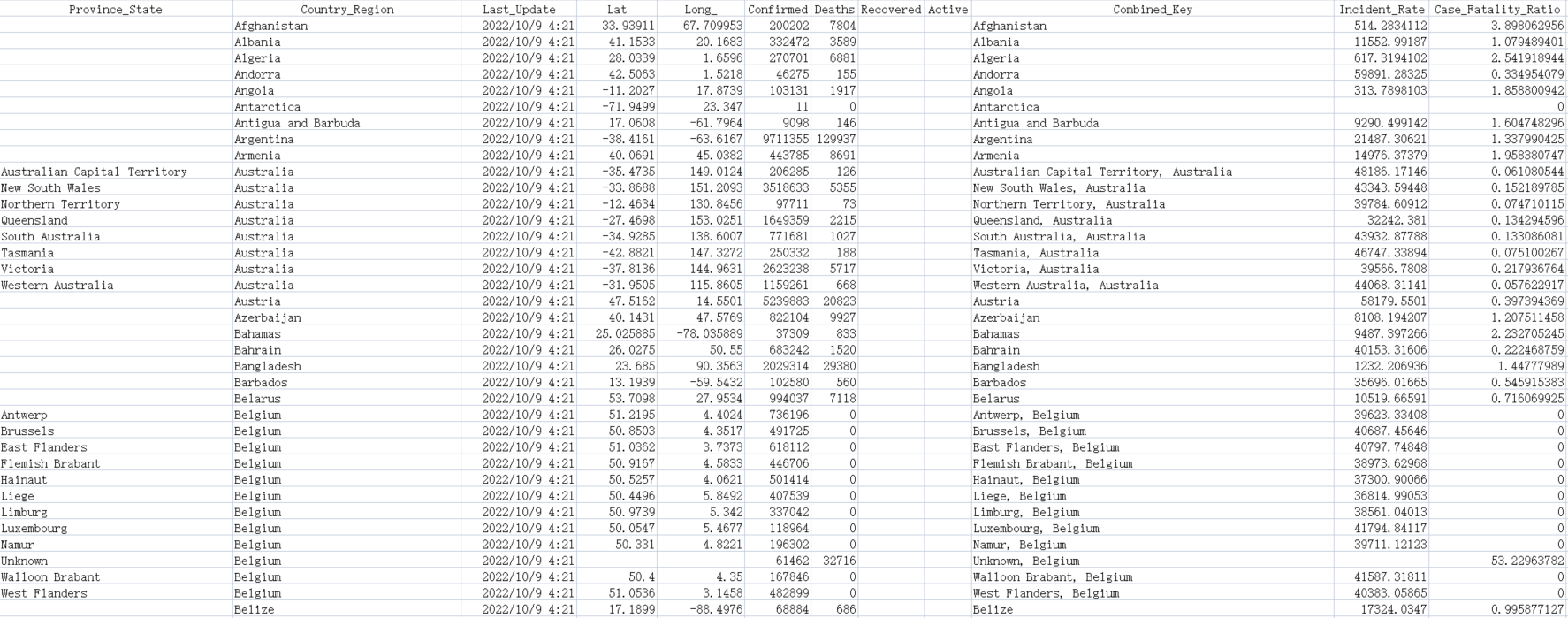
**数据来源：**

约翰霍普金斯大学系统科学与工程中心（CSSE）的COVID-19数据存储库<https://github.com/CSSEGISandData/COVID-19>

维基百科各国家和地区人口列表

<https://zh.wikipedia.org/wiki/%E5%90%84%E5%9B%BD%E5%AE%B6%E5%92%8C%E5%9C%B0%E5%8C%BA%E4%BA%BA%E5%8F%A3%E5%88%97%E8%A1%A8>

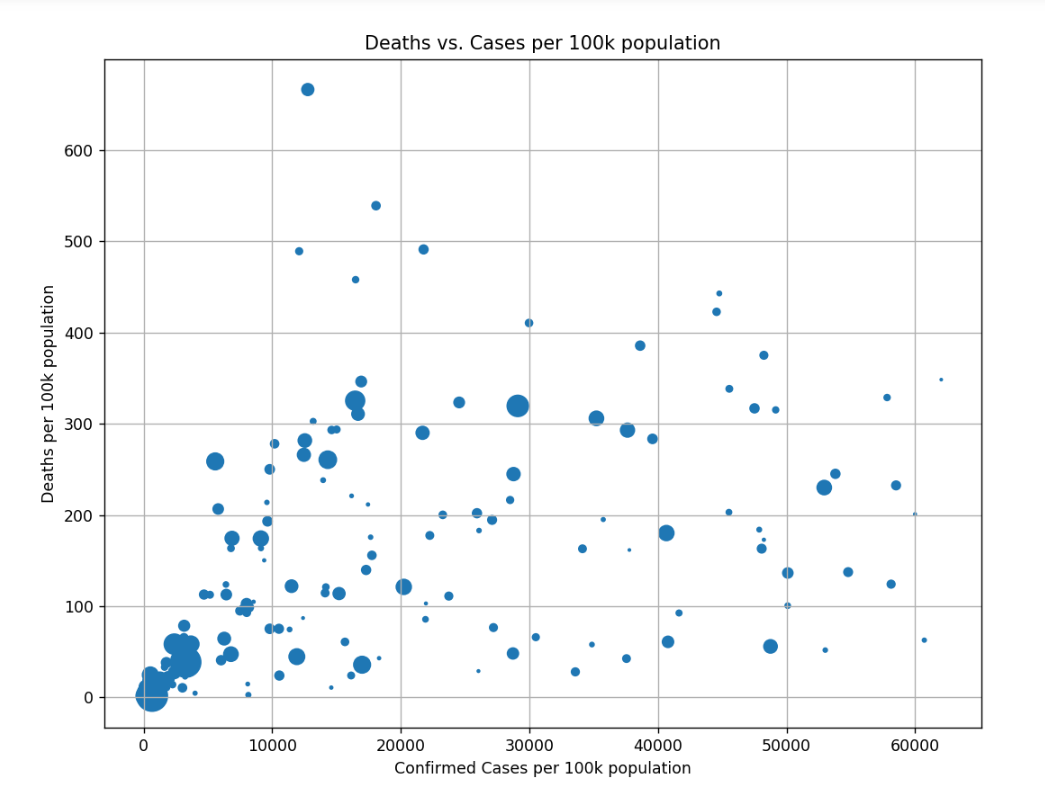
数据式例：



**数据处理与分析：**

在新冠众多数据中，我们选择Confirmed cases和Deaths作为指标来分析各国疫情的关联性。我们从数据库中选取截至2022.10.9的全球各地区的新冠疫情数据，并以国家为单位处理数据。

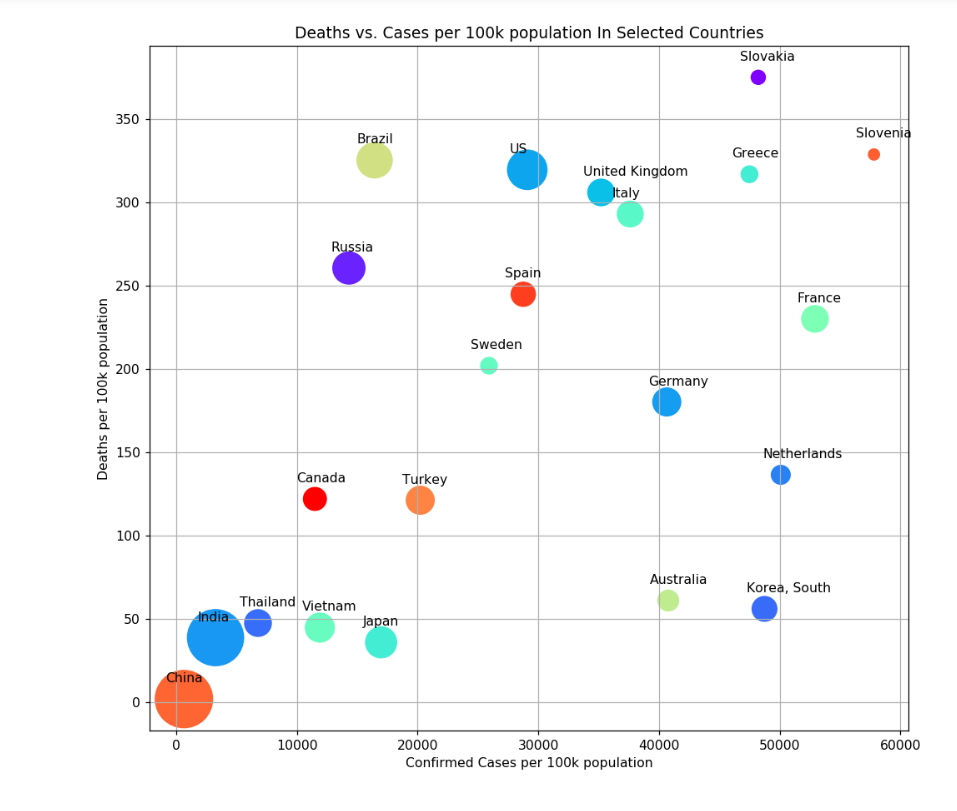
考虑到各国的人口基数差距过大，导致各国的Confirmed cases和Deaths也有数量级的差距。因此我们认为Confirmed cases和Deaths占该国总人口的比例更能说明该国疫情的严重程度。以它们作为横纵坐标，我们得到如下散点图：



其中每个点都代表一个国家（由于此处国家数量过多，未显示各国名字），圆点的半径越大表示其人口基数越大。

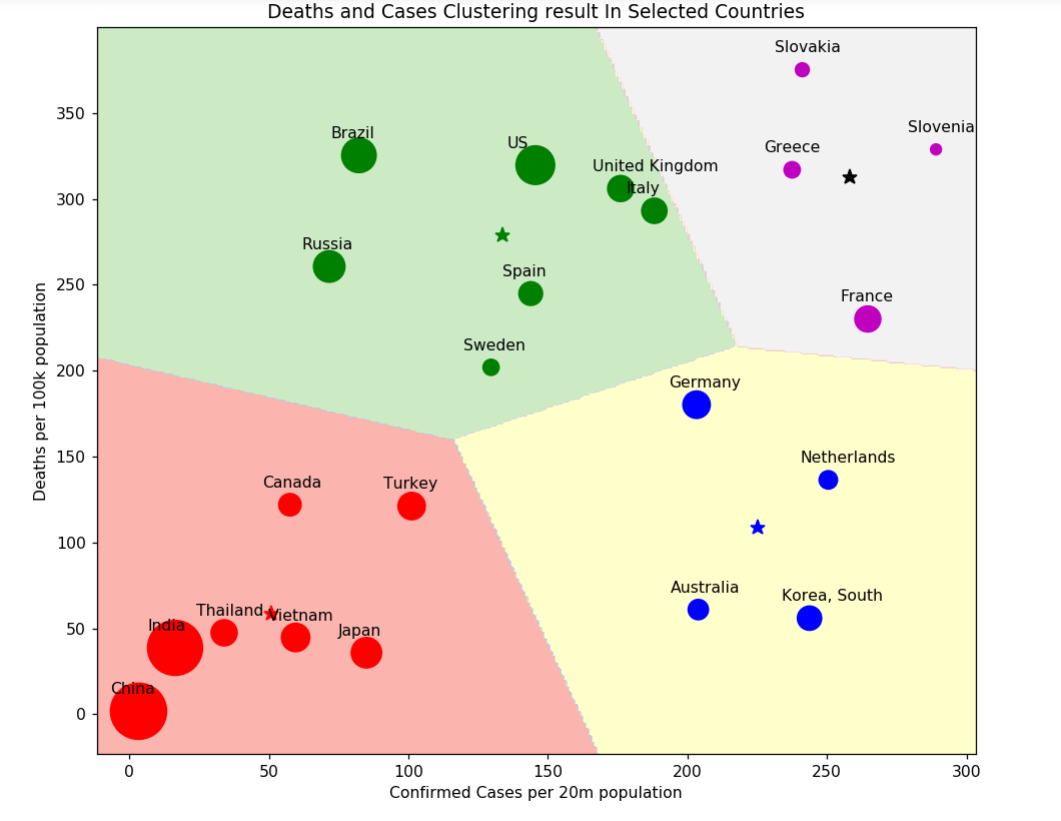
可以粗略地看出，左下角的点较为密集，其中不乏几个较大的圆点。可以推断出几个人口基数较大和数个人口基数较小的国家的Deaths和Cases比例较小，它们对疫情的控制较好。其余大部分的点分布较为均匀，只有少数国家有着过高的确证率和死亡率。

不过，由于该图包含了全球所有的国家，因此会包含较多噪音和冗余。比如说，很多极少人口的国家也被包含进去，但是如果人口数量过小，疫情影响有较大随机性；一些不发达国家的数据真实性不好确定；而很多人口接近，地理位置接近的国家也有着很近似的数据。 因此我们选取了22个比较有代表性的国家，着重对它们进行分析：



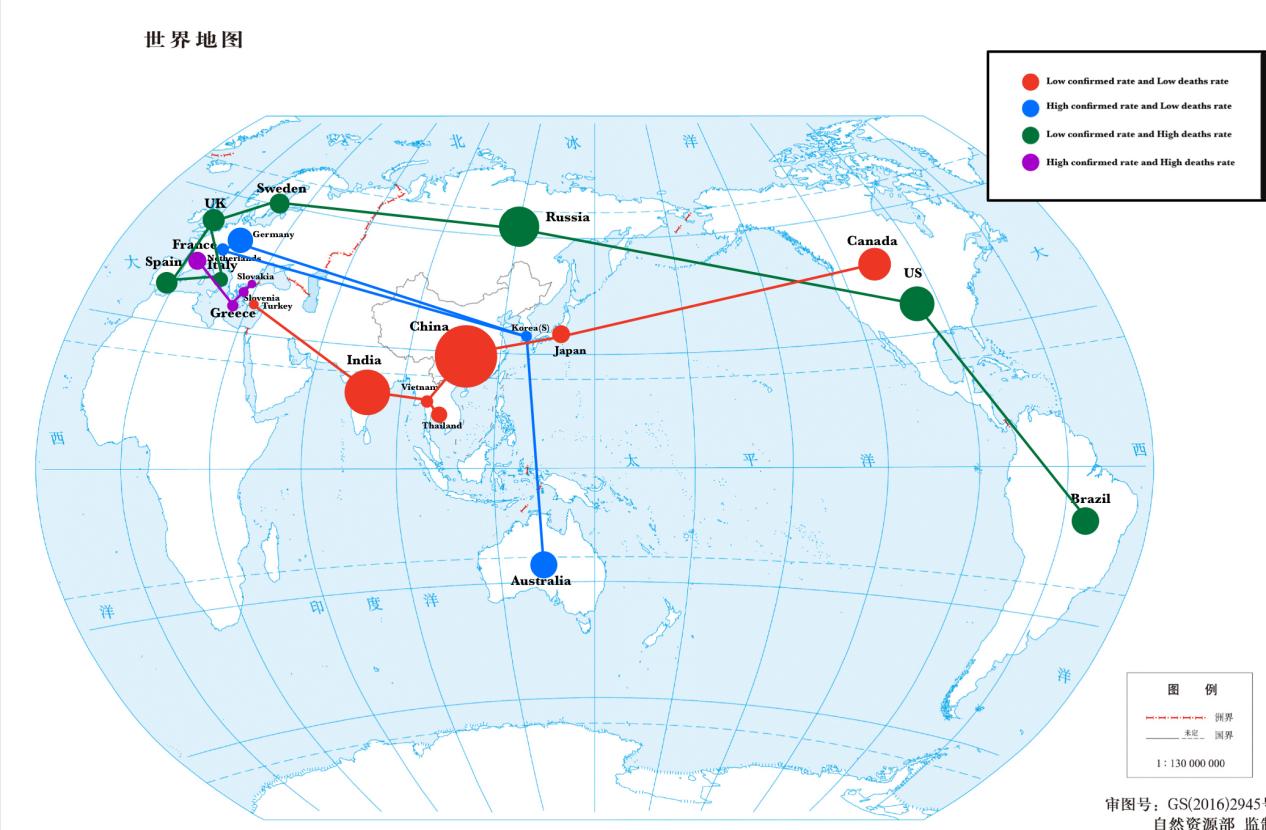
可以看到，这些国家的分布较为均匀。为了找到各个国家在Deaths和Confirmed cases比例的数据上的相关性。我们选择用简单的K-means聚类算法对这些国家进行聚类。

由于Confirmed cases的数值与Deaths的数值差距极大，对聚类效果影响很大，所以我们把它们进行归一化处理，给予两者表现国家疫情的严重程度相同的权重。最后聚类效果如下：



可以看到，这些代表性国家被分成四类，且由其分布的形状，恰好对应于四种情况：Low confirmed rate and Low deaths rate;High confirmed rate and Low deaths rate;Low confirmed rate and High deaths rate;High confirmed rate and High deaths rate。

我们把以上得到的结果可视化在世界地图上，期望找到其聚类结果与地理位置的关系。



我们将同一类的点用同一种颜色标出，并依次连线，从而得到一个复杂网络。这个复杂网络的每个连通分量都代表一个类，表示它们具有相似的Confirmed cases rate和Deaths rate。

**结论：**

可视化的结果一定程度上可以验证我们的猜测。Low confirmed rate and Low deaths rate的国家主要分布于东南亚；Low confirmed rate and High deaths rate的国家大部分分布在西欧；High confirmed rate and High deaths rate的国家位于中东附近。而High confirmed rate and Low deaths rate的国家分布广泛，并未集中。

不过也可以看到，每个类都存在地理位置差距较远的国家，这可能与每个国家的国家性质、防疫政策和经济医疗实力等各种因素相关。

**未来可行的工作：**

因为可视化是人工手动进行的，所以选取的国家数量不足。未来可以设计出自动化的算法进行可视化，便可以加入更多的国家数据，得到更全面的数据。同时把全球各个国家按照地理位置进行分类。对比两种类别的分布，以更好地判断疫情严重程度与地理位置的关系。