

**数据仓库大作业报告**

**课 程 名 称：**  数据仓库

**作 业 题 目：** 新冠肺炎情况的国家影响因素

**小 组 成 员：** 24320182203172 陈俊任

24320182203200 胡先宇

24320182203268 宋彦

24320182203189 冯泊翔

**指 导 教 师：** 王 鸿 吉

目录

[一、 背景和需求分析 3](#_Toc2804)

[二、 需求分析及逻辑设计 4](#_Toc24677)

[1. 总体设计 4](#_Toc13781)

[2. 分析主题域 7](#_Toc12134)

[3. 维度划分 8](#_Toc9089)

[（1） 经济水平维 8](#_Toc31098)

[（2）医疗水平维 9](#_Toc8870)

[（3）检测水平维 10](#_Toc8199)

[（4）民主水平维 11](#_Toc3749)

[（5）人口密度维 12](#_Toc23102)

[三、 物理设计 13](#_Toc3596)

[四、 ETL过程 17](#_Toc22536)

[五、 多维分析及可视化分析结果 21](#_Toc3664)

[（1） 从医疗水平维分析 21](#_Toc8560)

[（2） 从民主水平维分析 23](#_Toc11796)

[（3） 从检测能力维分析 25](#_Toc22756)

[（4） 从经济水平维分析 27](#_Toc32100)

[（5） 从人口密度维分析 29](#_Toc17898)

# **背景和需求分析**

随着2020年初，COVID-19在世界各地集中爆发，世界各国也迅速出台了各项政策以应对疫情的爆发。二十国集团领导人于2020年11月21日强调，国际社会必须继续加强协调与合作，应对新冠疫情。二十国集团领导人第十五次峰会以视频方式拉开帷幕。此次峰会为期两天，由轮值主席国沙特阿拉伯主办。峰会首日，沙特举办了一场“大流行病的防范与应对”边会。二十国集团领导人与国际组织代表在会上强调，必须继续加强协调应对新冠疫情大流行，特别是支持最脆弱群体，同时增加疫情防控相关支出，用于疫苗等的进一步研究和创新。会后发表的声明说，世界正在经历一场规模空前和异常严重的卫生和经济危机。二十国集团领导人在今年3月沙特主持召开的应对新冠肺炎特别峰会上承诺，尽一切努力战胜疫情，与国际社会一道保护生命。今天二十国集团领导人再次聚首，目标是确保采取必要的后续行动，建设一个更安全的世界，塑造一个包容、可持续和有韧性的未来。二十国集团于1999年成立，由中国、阿根廷、澳大利亚、巴西、加拿大、法国、德国、印度、印度尼西亚、意大利、日本、韩国、墨西哥、俄罗斯、沙特阿拉伯、南非、土耳其、英国、美国以及欧盟等20方组成。尽管各国都采取了一定的防控措施，新冠疫情的影响仍在继续，本文从2020年1月至2020年11月各国的新冠疫情数据以及各国的基础数据入手，希望分析新冠疫情的严重程度与国家各方面因素的联系。

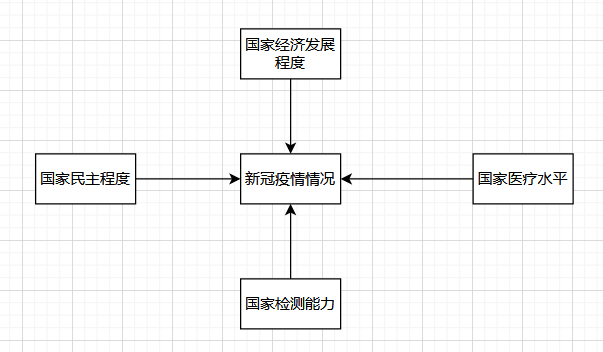
# **需求分析及逻辑设计**

## **总体设计**

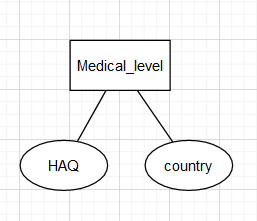
我们认为存储针对新冠疫情情况与国家各方面因素的数据仓库应存放以下数据：

国家名称，日期，每百万病例数，每百万死亡数,每百万检测数，人口密度，人均GDP，HAQ指数，民主指数。

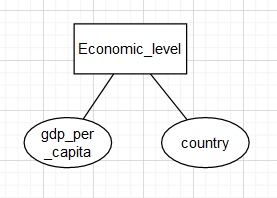
我们将人均GDP作为评定国家经济发展程度的标准，HAQ指数作为评定国家医疗水平的标准，民主指数作为评定国家民主程度的标准，每百万检测数作为评定国家检测能力的标准，分别存放于四个表中，新冠疫情情况表中的每百万病例与死亡数将作为判断严重程度的标准，E-R图如下



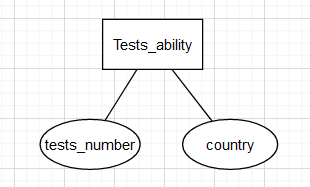
医疗水平：



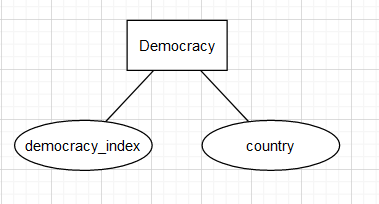
经济发展程度：



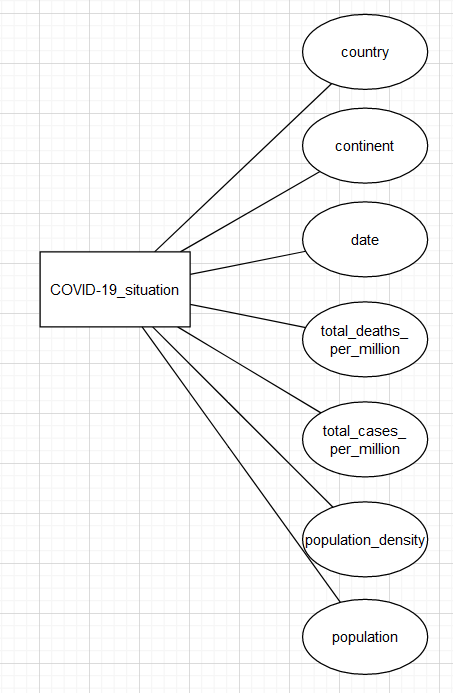
检测能力：



民主程度：

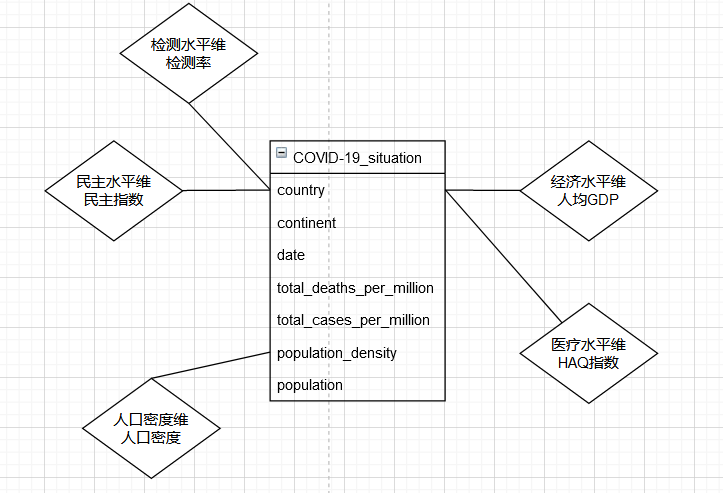


新冠疫情情况表中将存放国家的基本信息以及疫情的重要信息，通过国家与其他四表形成关联



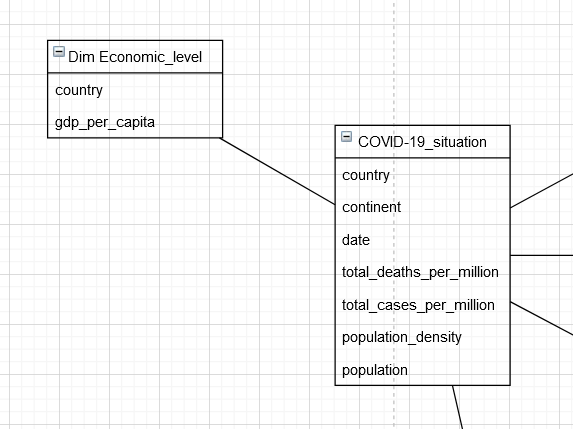
## **分析主题域**

对于新冠肺炎的情况分析，我们需要通过国家的五个评判方向进行多维分析，人口密度维是考虑到COVID-19作为传染性疾病，人口密度会对其传播速率产生较大影响，故提取情况表中作为国家基本信息的人口密度自成一维进行分析，以此来分析国家的各方面实力和性质对新冠疫情的严重程度产生的影响，事实表如下：



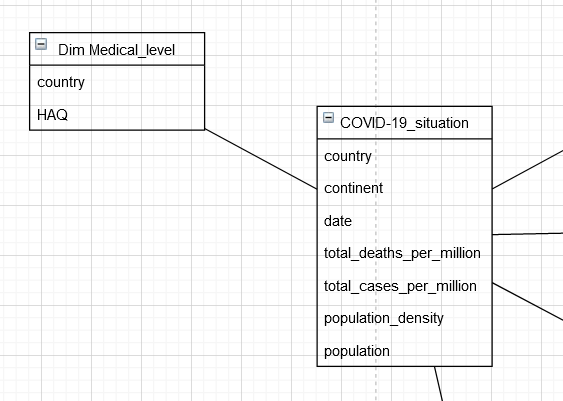
## **维度划分**

### **经济水平维**



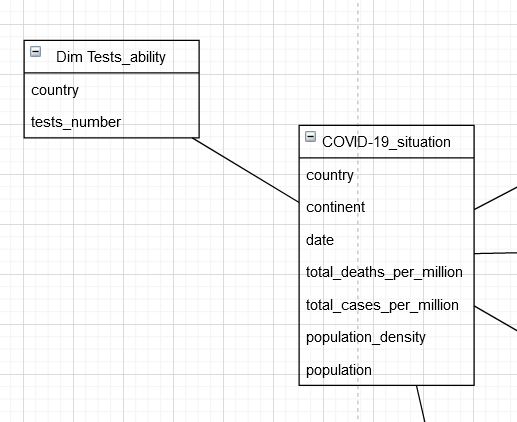
疫情扩散会对国家的经济产生冲击，相对的，国家的经济实力也能在一定程度上影响疫情的发展，故我们以人均GDP为经济水平的衡量标准，尝试分析国家经济水平与疫情发展之间的关系。

### **（2）医疗水平维**



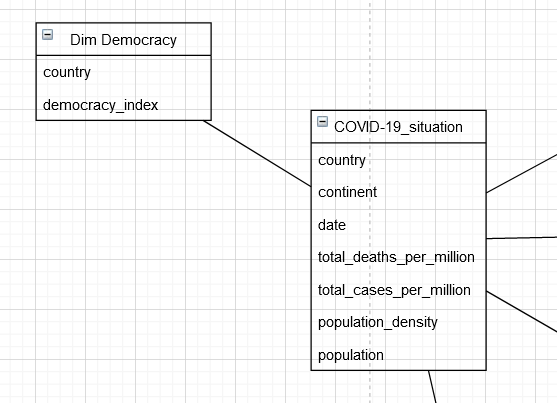
国家医疗水平与传染病的发展有着密切的关系，医疗质量和可及性高低，取决于一个国家对32种可防可治疾病的治愈率与死亡率，包括一系列癌症、传染病、心脏病等。可以说国家的医疗水平的高低在很大程度上影响着疫情发展，我们以HAQ指数作为国家医疗水平的评判标准，尝试分析关联性的高低。

### **（3）检测水平维**



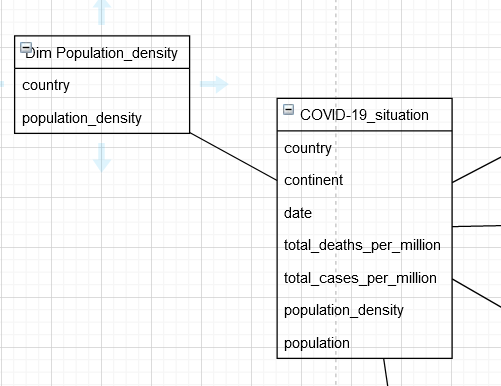
由于目前统计确诊人数，统计病死人数的手段基本局限于核酸检测，由于检测能力的高低，部分国家显现出的疫情情况或许与实际情况有出入，我们以检测人数/总人数作为评判标准，尝试分析检测水平的影响

### **（4）民主水平维**



前日复旦大学的沈逸老师与一场关于中美关系的讲座中提出一个假说：在这场疫情中，一个国家越民主，医疗水平越高，因疫情而死的人越多，我们建立民主维度，以民主水平为参照标准，尝试进行分析证明。

### **（5）人口密度维**

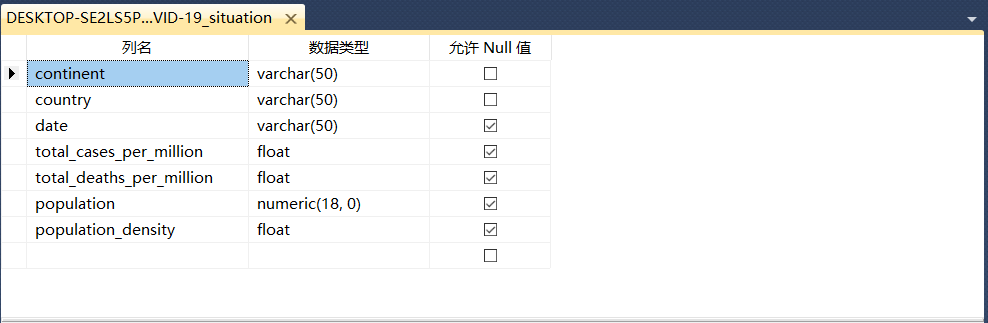


由于此次新冠疫情为传染性疾病，人口密度在一定程度上会影响人处在人口密集场所的几率从而间接影响疫情的发展，故建立人口密度维度分析人口密度对疫情的影响程度

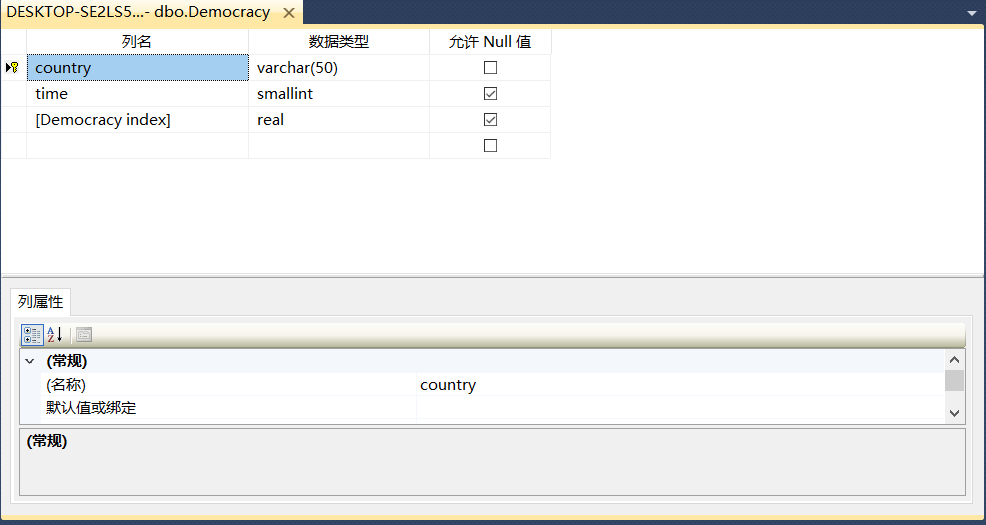
# **物理设计**

数据源在数据库中存放成5张表：COVID-19\_situation,Democracy,Economic\_level, Medical\_level,Test\_ability,数据字典如下：

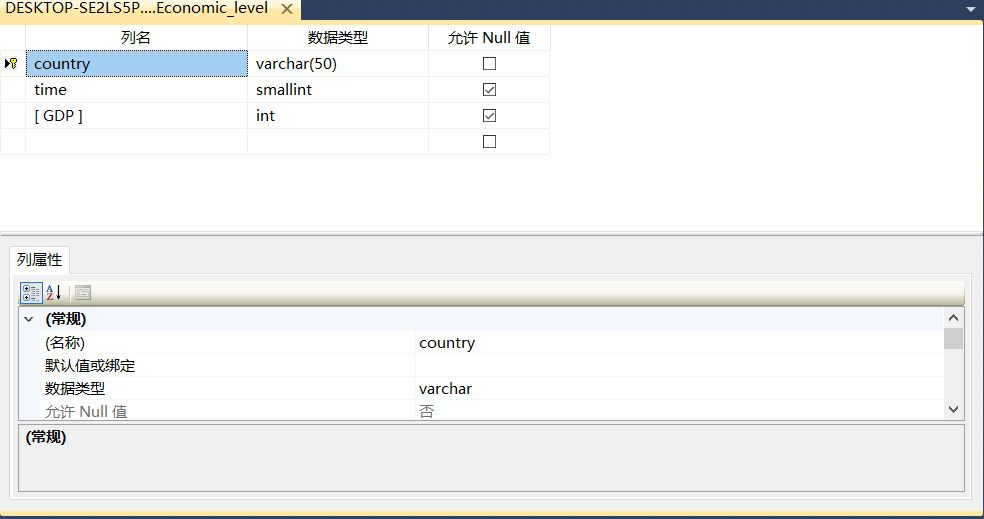
COVID-19\_situation表：



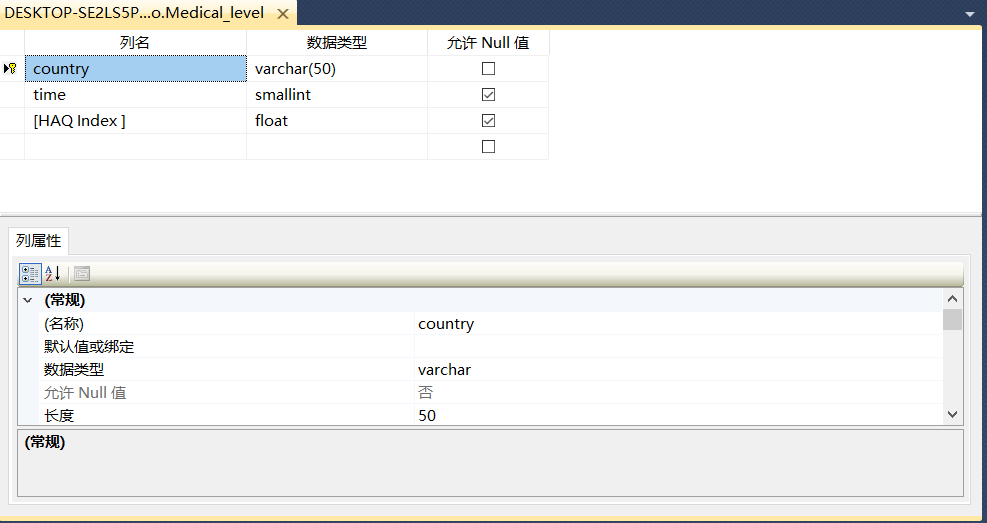
Democracy表：



Economic\_level表：



Medical\_level表：

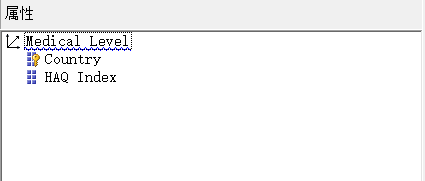


Test\_ability表：

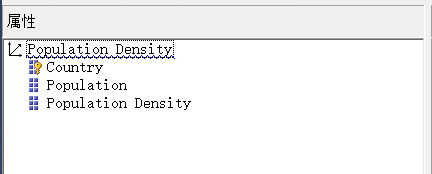


建立的五个维度如下：

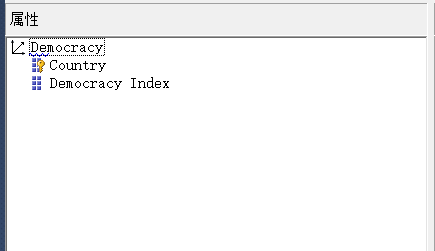
医疗水平维度：



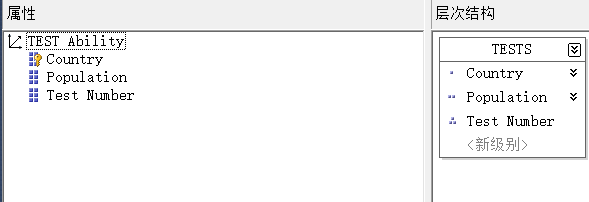
人口密度维度：



民主水平维度：



检测水平维度：

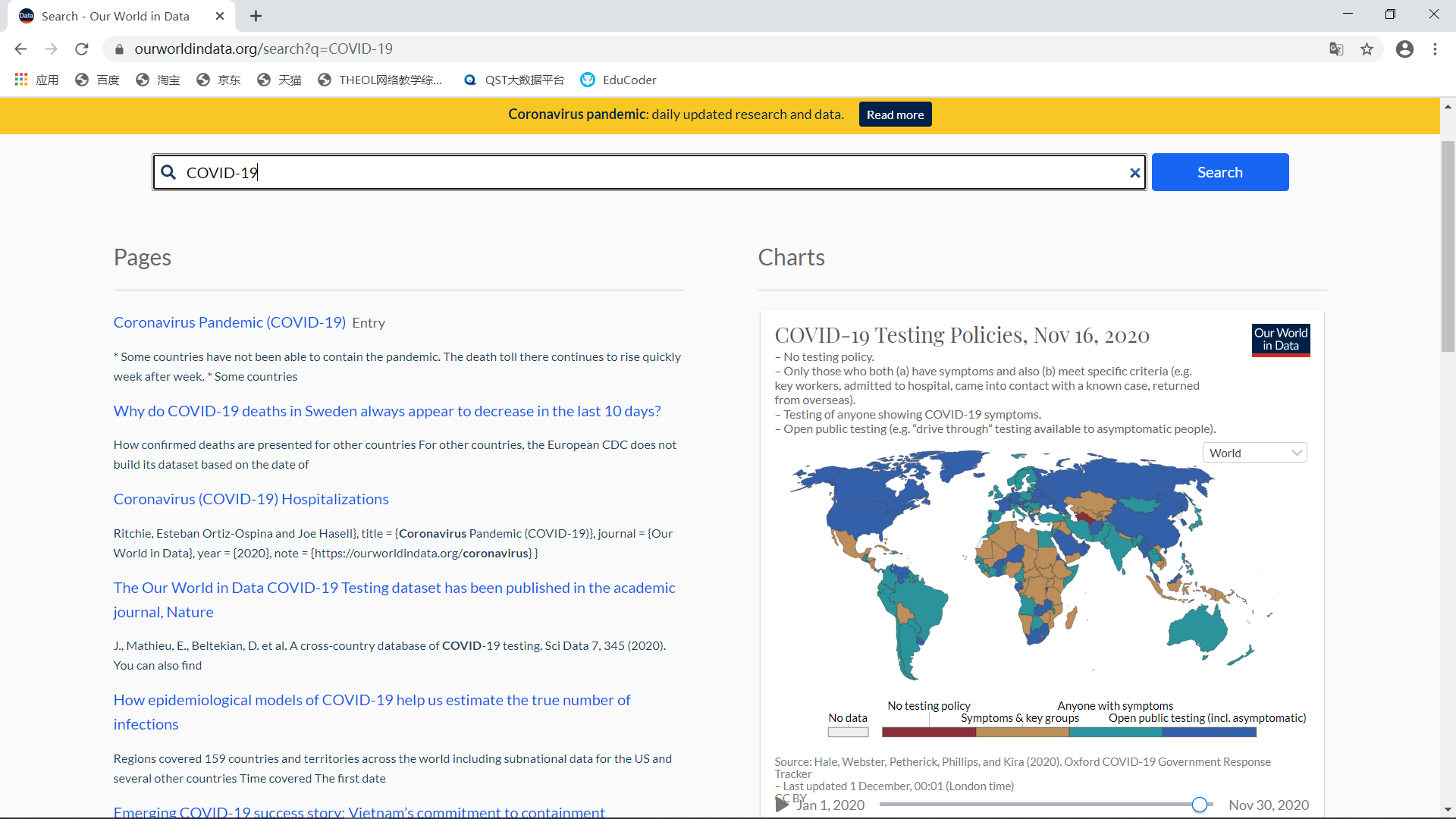


经济水平维度：

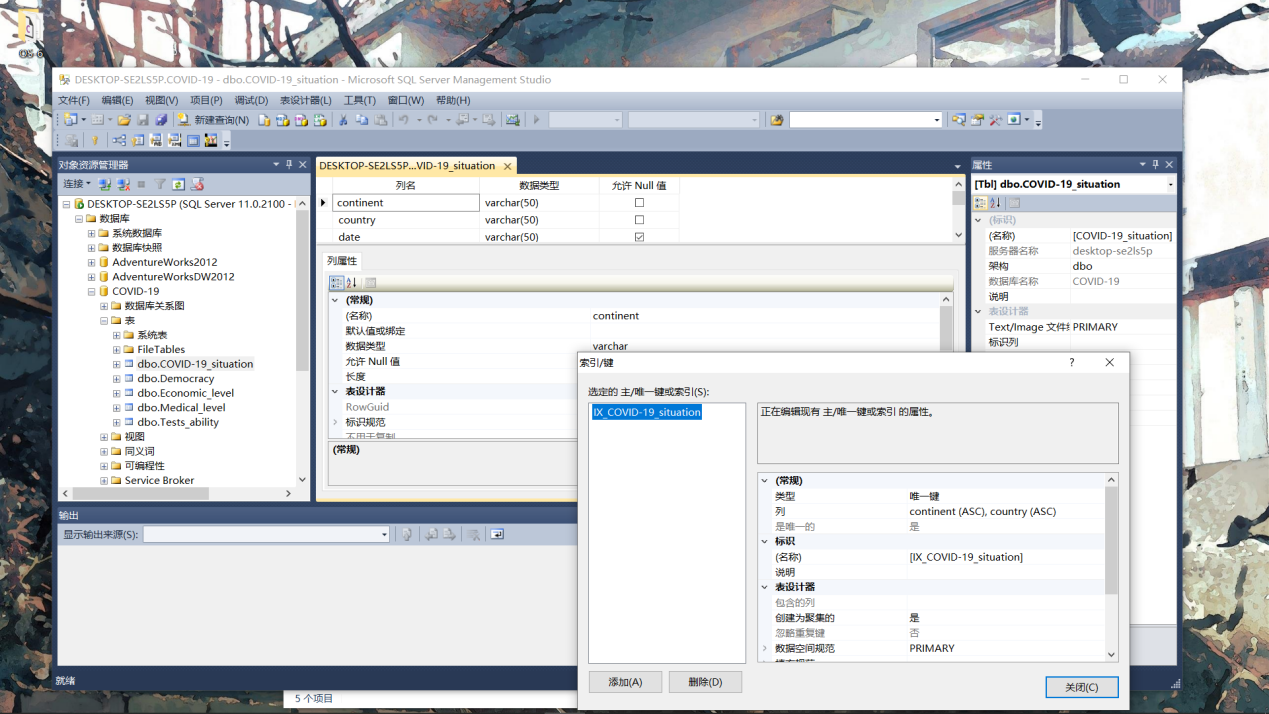


# **ETL过程**

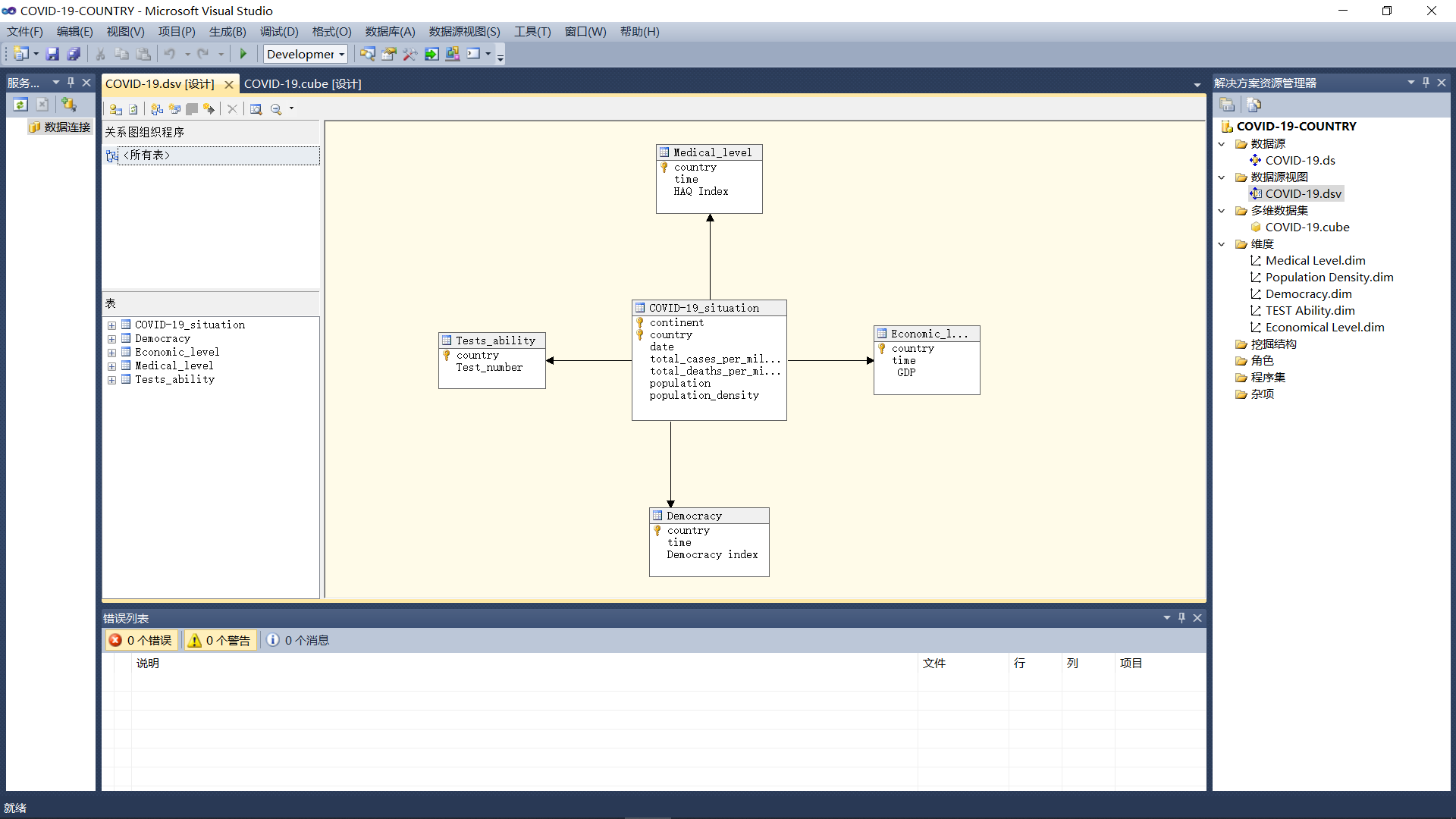
本项目数据来自Our World in Data,源数据文件格式为csv

****

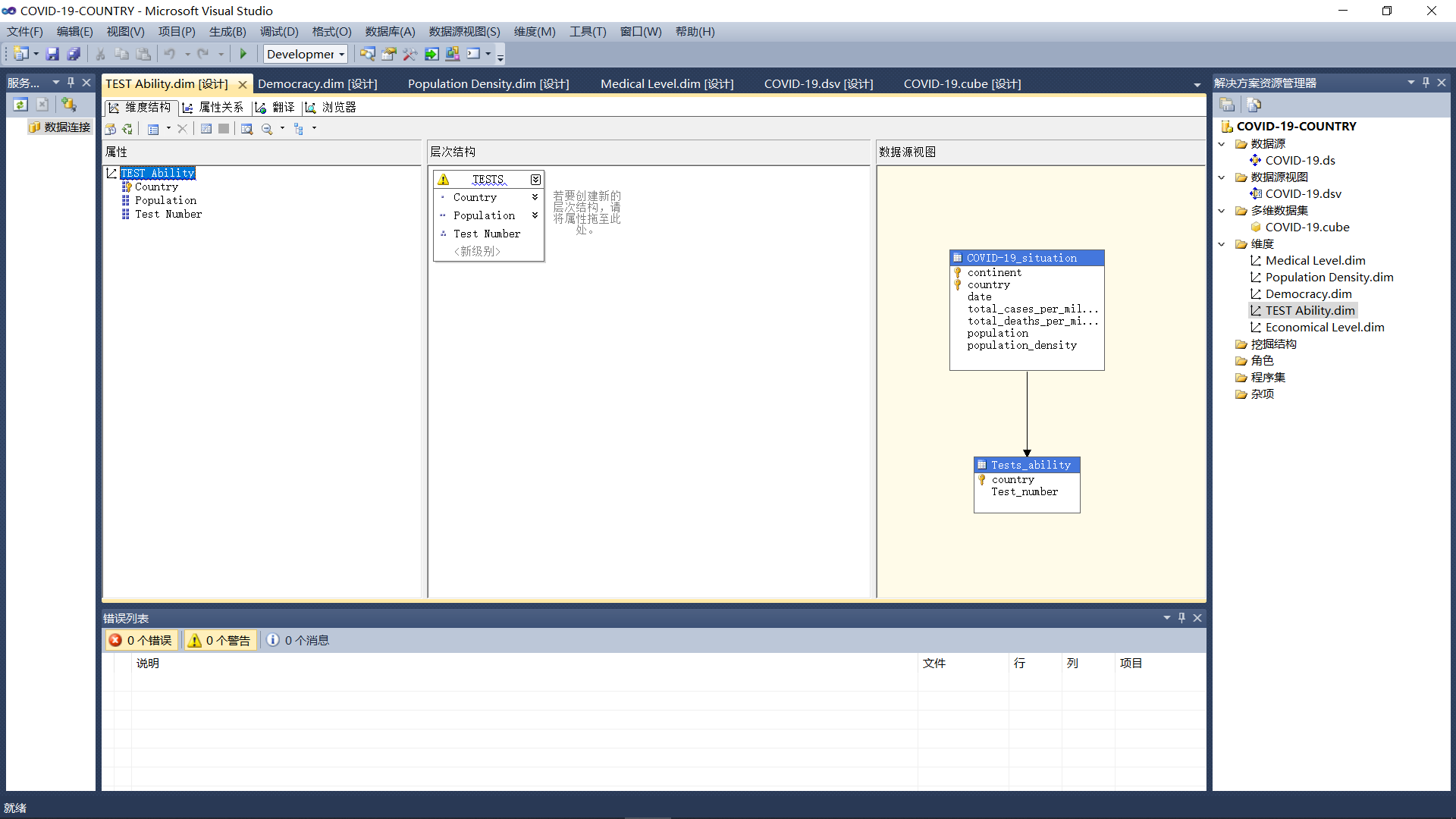
经过物理设计，数据处理（删除不合规范的数据与无法形成完整主外键关系的数据）等完成数据库配置



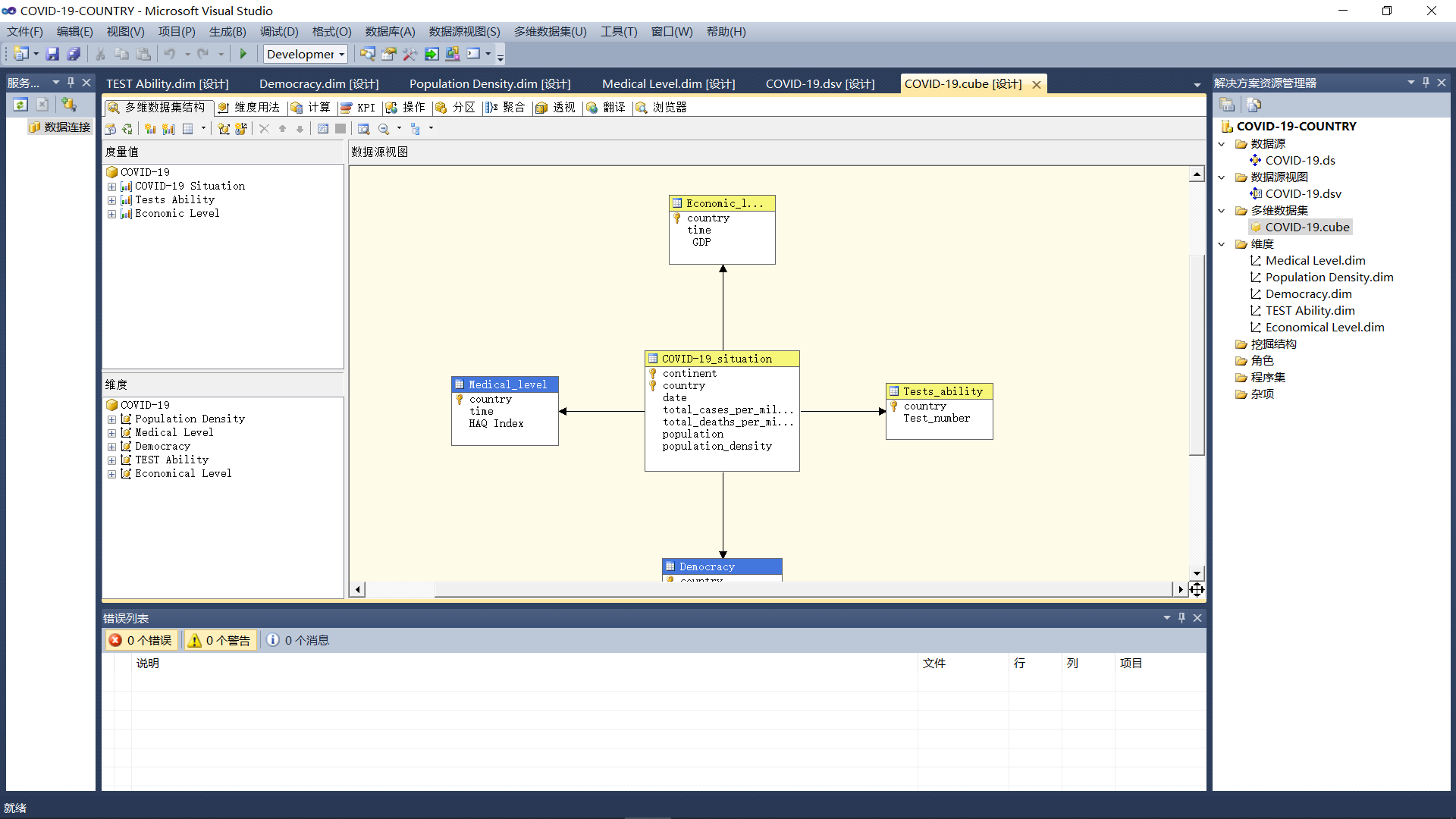
在SQL Server Data Tools新建项目，连接数据源并新建数据源视图：



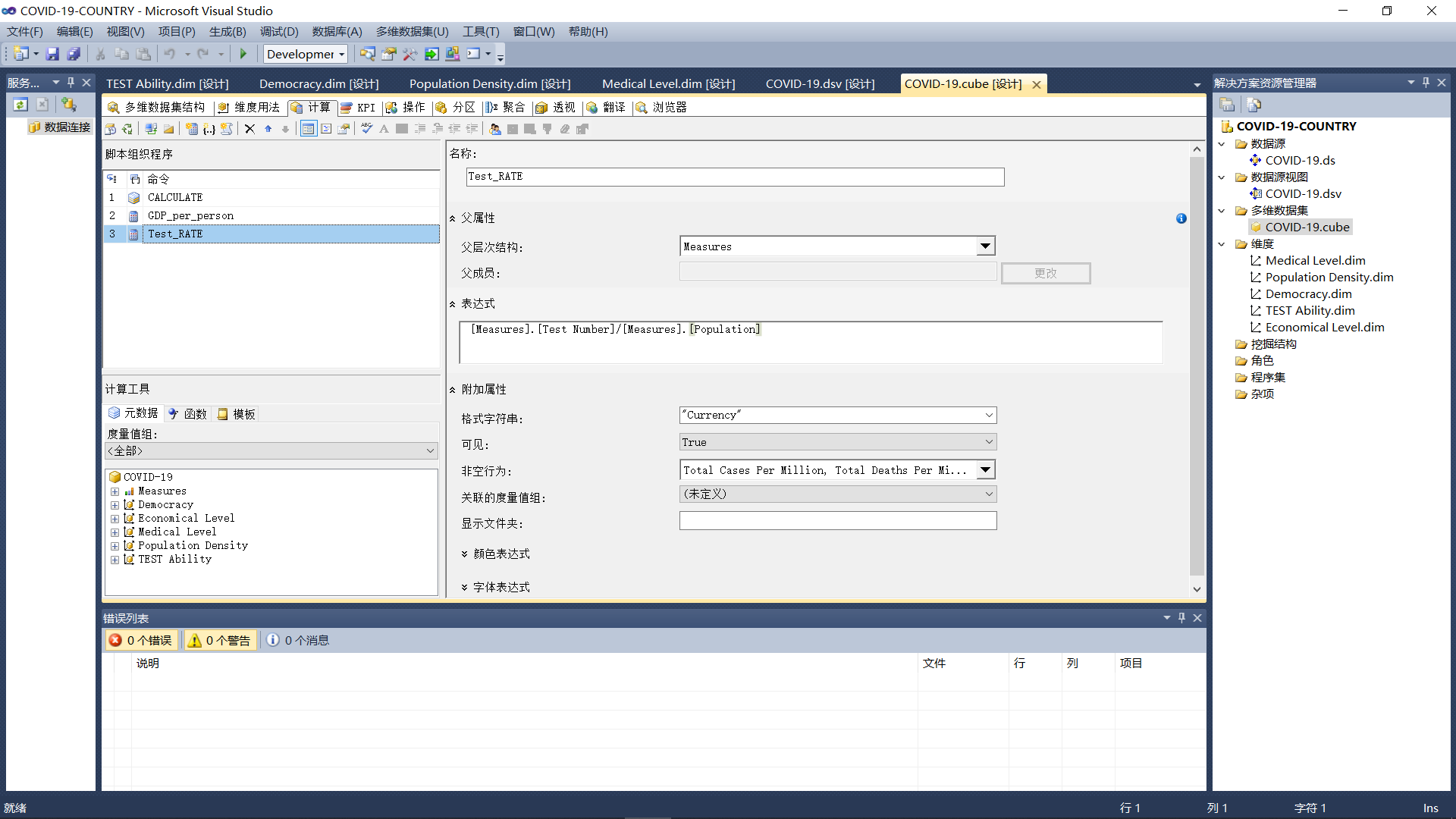
新建维度并建立层次结构：



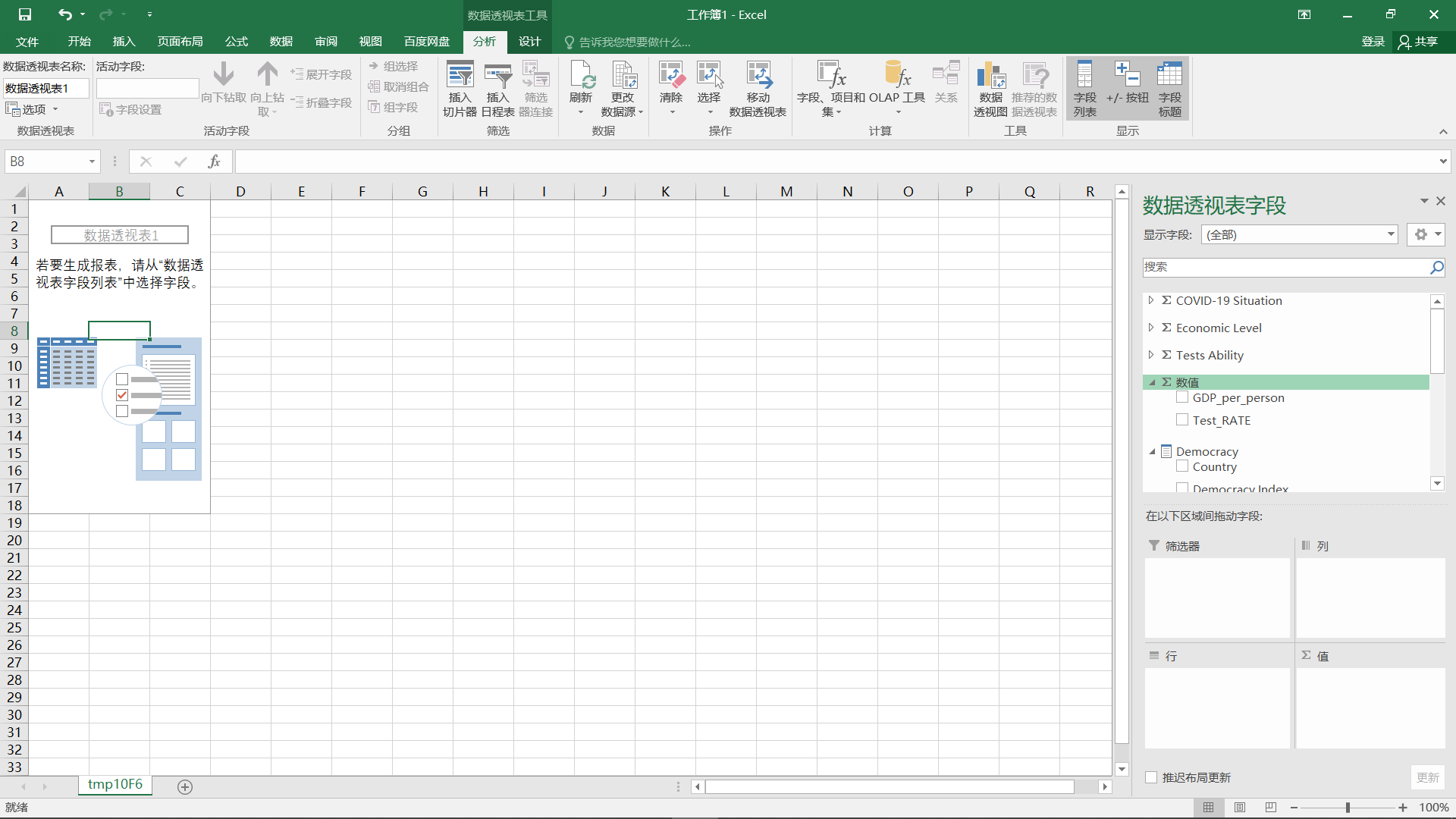
完成多维数据集建立与配置：



新建计算等：



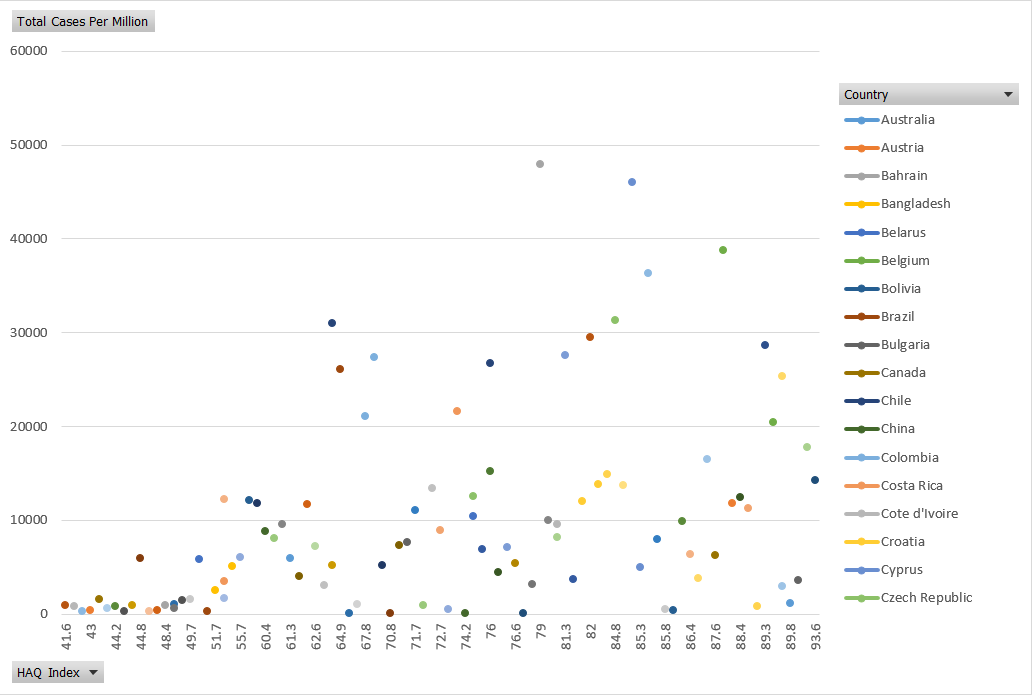
在数据透视表字段中勾选需要的字段就可以进行分析：



# **多维分析及可视化分析结果**

## 从医疗水平维分析

**COVID-19感染情况与医疗水平的关系**



散点图可直观看出病例数的聚合度

由图可知，低医疗水平的国家检测病例人数也较低，其中原因可能有：

1. 感染者可能并无症状，而这自然会造成检测到的病例数少于实际感染人数。
2. 低医疗水平的国家检测能力也较低，可能是案例数量少的因素之一。
3. 国家位置较偏，国际访客量小，疫情传播速度较慢

除此之外，从图中可以看出，随着医疗水平的提高各个国家的总病例也聚合在10000左右，有些国家总病例数较高的缘故可能是地域因素导致了病毒的扩散和传播以及未采取防疫措施、不够重视疫情等行为导致了病例数较高。总体来看自图中央开始，随着医疗水平的提高，国家的总病例数也随之降低，病例数逐渐控制在一定范围内。说明医疗水平高的国家发现病例，控制疫情的能力也较高。

综合而看医疗水平对于国家疫情的感染情况的影响程度较高，关联性较强。

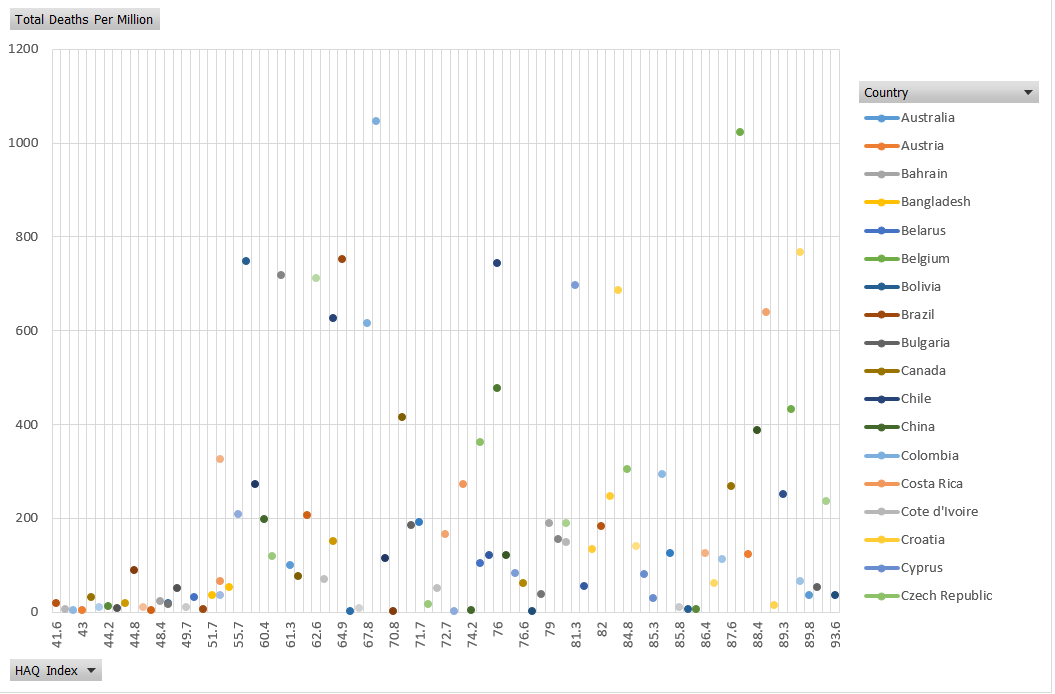
**COVID-19死亡情况与医疗水平的关系**

散点图可直观看出总死亡病例的聚合度。

由图可得，与上图类似，低医疗水平的国家死亡病例也较少，其中原因与上图类似，疫情尚未在低医疗水平国家传播开以及国家检测疫情死亡人数能力较低等原因都可能导致该国家记录的死亡人数较低。

除此之外，随着医疗水平的提高，可以看出图中总死亡人数聚合在0-200的区域内，少有一些国家的死亡人数较高，由此可得医疗水平高的国家治疗感染病例的能力也较高，对疫情的控制力也较强。

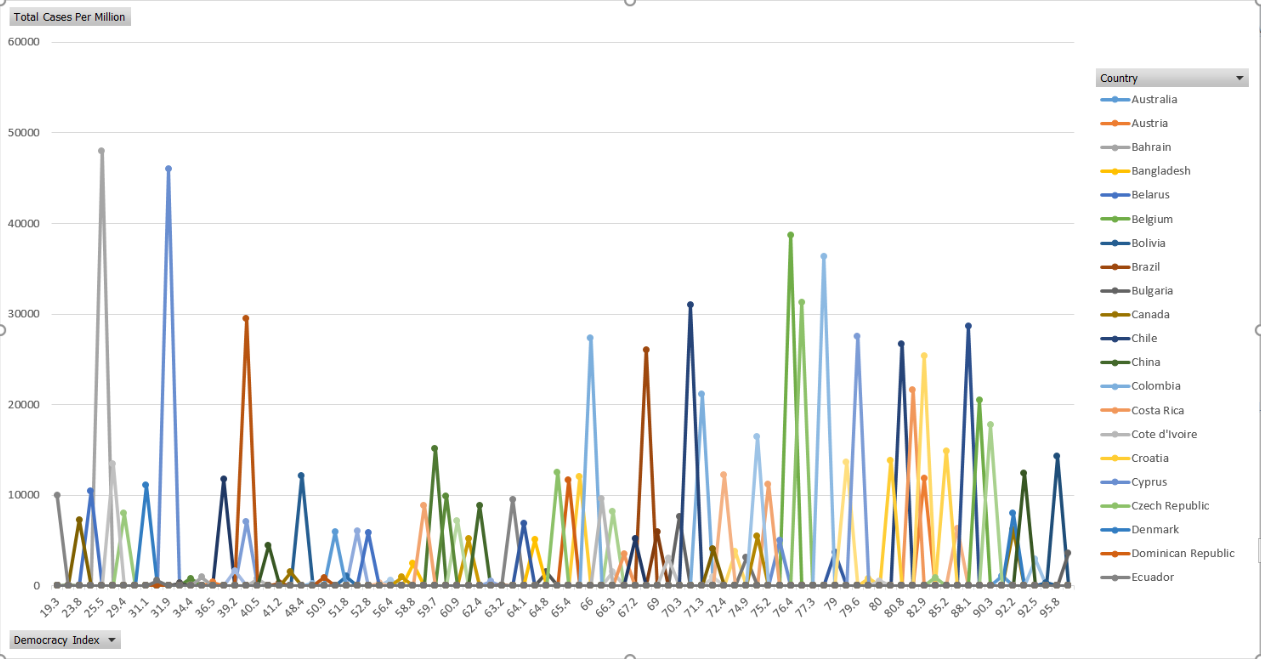
综合而看医疗水平对于国家疫情的治愈情况的影响程度较高，关联性较强。



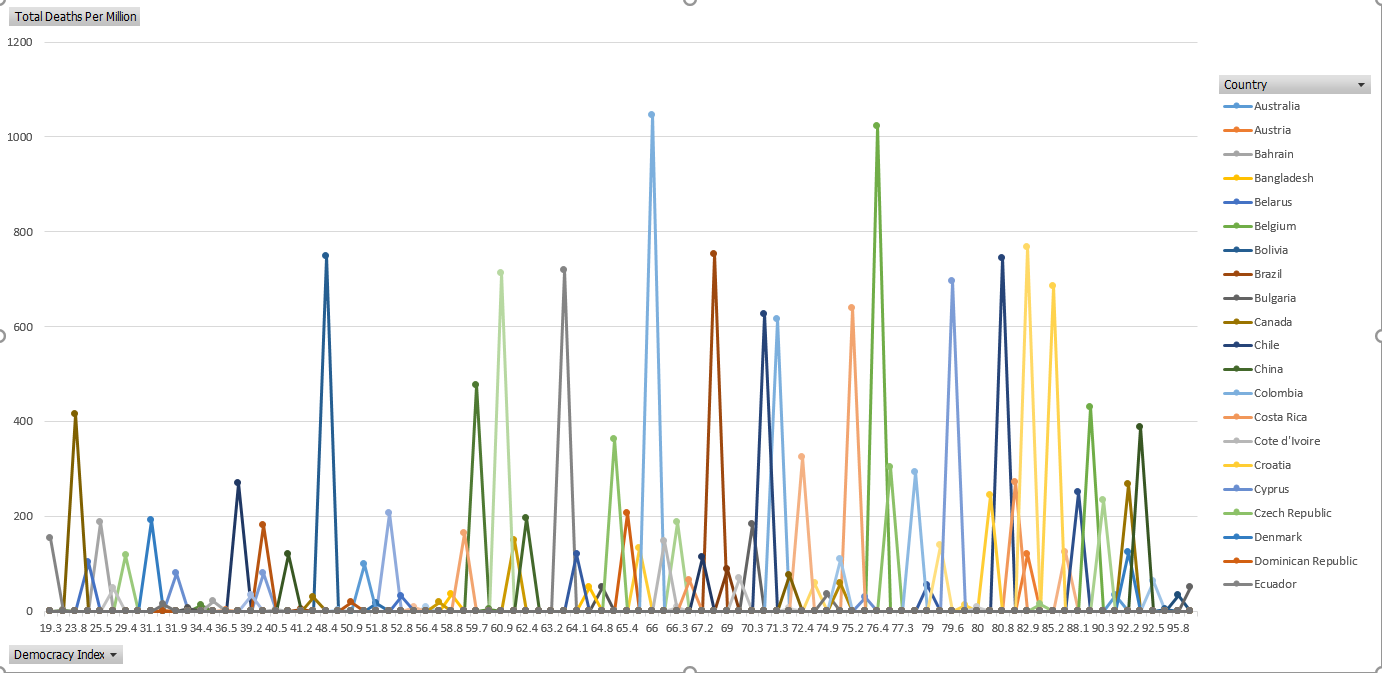
## 从民主水平维分析

由折线图可看出感染情况较为严重的国家集中分布在民主指数极低与民主指数较高的区段，而死亡情况的分布则较为均匀，但民主指数较高的区段的严重程度明显更甚。将民主指数分区段取中位数，则可得到较为显著的上升曲线显示随民主程度的提高COVID-19死亡人数也在增加。

**COVID-19感染情况与民主指数的关系**

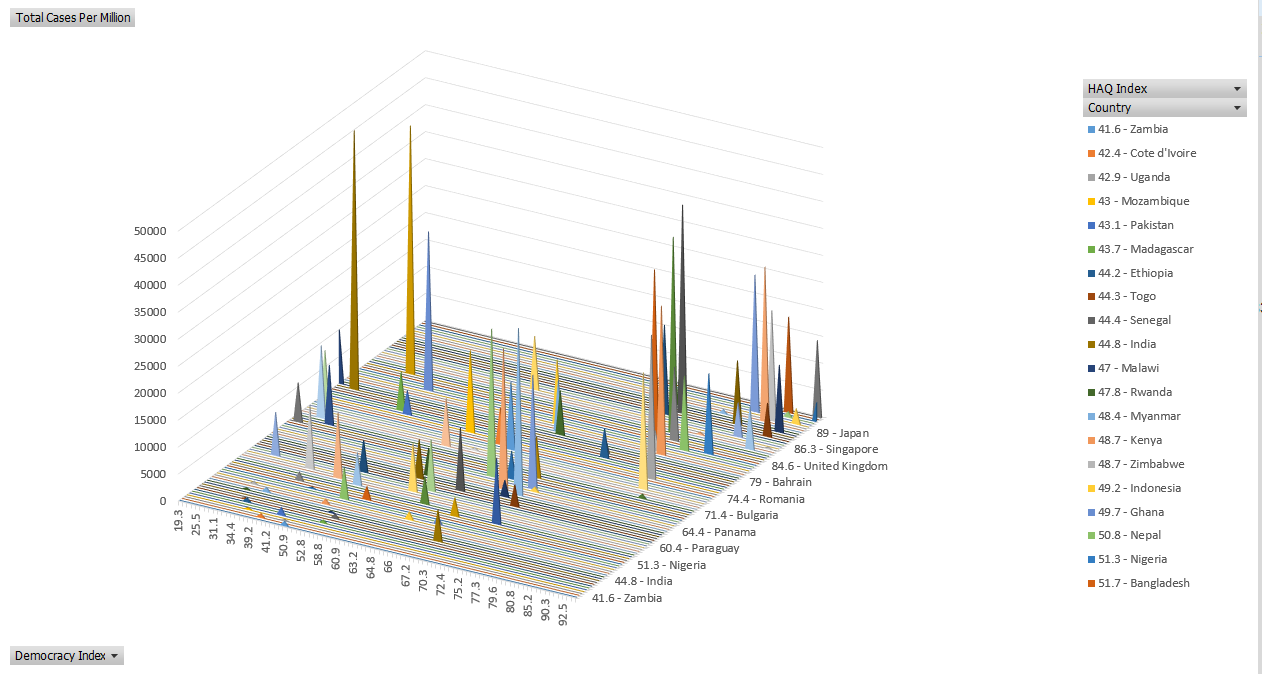


**COVID-19死亡情况与民主指数的关系**

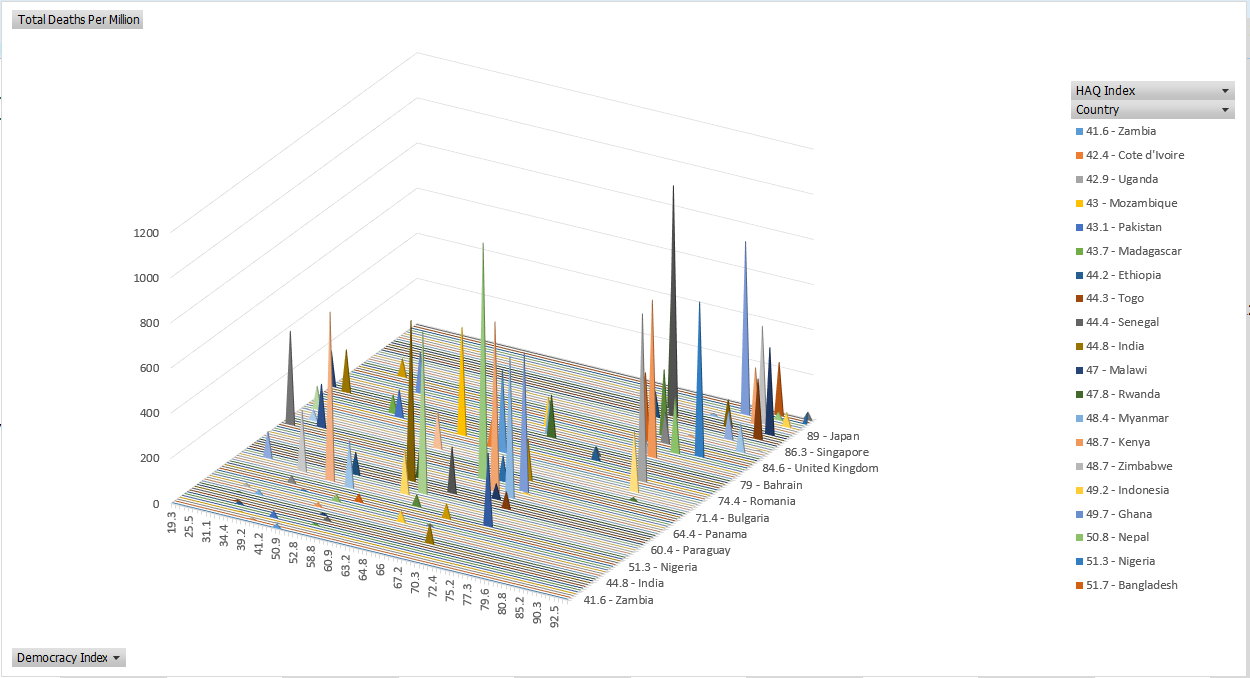


考虑到民主指数的影响不够显著，我们引入HAQ指数进行综合分析，同时尝试检验假说“一个国家越民主，医疗水平越高，因疫情而死的人越多”。由图可得感染较为严重的国家集中在高民主指数高HAQ区域，低民主指数高HAQ区域也有显著个例，但民主指数与HAQ都较为中等的区间更趋向于随机分布。而从死亡情况看，高民主指数与高HAQ区域和民主指数与HAQ指数都较为中等区域都存在聚集的较为严重的国家，由其分布规律较难得出此假说。

**COVID-19感染情况与民主指数和HAQ的关系**



**COVID-19死亡情况与民主指数和HAQ的关系**



## 从检测能力维分析

从图表中可以看出，在检测数目比较少的时候，感染的数目也比较少，当检测数目逐渐增加，感染的数目也开始增加，但是当检测数目增加到一定程度，感染的数目会浮动并且朝着减少的势头发展，其中的原因是：

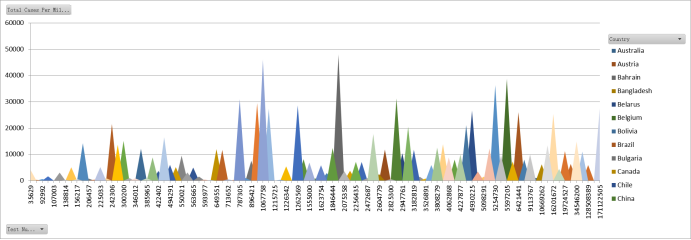
1.在检测数目比较少时，发现的感染数目也会比较少，但当检测数目增加，感染数目会随之增加，此时感染数目少的原因是可能尚有许多感染者未被检测到或者无明显症状。

2.在检测数目的中段（10万-60万），感染数达到最大，这部分感染数多的国家相对比较发达但是医疗水平不高或者防疫措施不够到位，国内外的流通使得感染增加，但是不够发达和普及的医疗并不能遏制疫情。

3.在检测数目很大时，感染数反而减少，因为愿意去检测上百万人的国家，不仅有消灭疫情的决心，还有及时控制疫情、治疗感染者的的疾控实力，检测的人越多，发现感染者就越早，减少了传染的可能，也能够及时治疗患者。

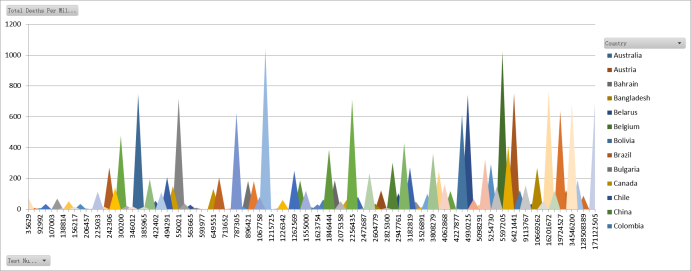
当然其中不乏有医疗水平高并且人口少的国家，能够很好地控制疫情，检测量高而感染数目少（后面的死亡情况也是同理）

**COVID-19感染情况与检测情况的关系**

****

从图表中看出，一开始的一小段检测数目少的国家死亡数目也很少，但是后面检测数目越来越多，死亡数目却都比较高且相差不会很多。这跟每个国家的人口数有关，人口数越少的国家，死亡的数目就会很少；当一个国家人口越多，疫情就越难以控制，死亡数目就会增加，但是检测数目的增加也反映了人口多的国家对疫情的控制力度，所以检测数目越大，因为疫情而死亡的人就越少，两个元素相互制约，导致人口多的国家检测量高，避免了一些死亡，所以死亡情况不会有很大的浮动。

**COVID-19死亡情况与检测情况的关系**

****

## 从经济水平维分析

感染数和死亡数都很少有两种情况：

1.因为国家人口很少，疫情比较容易控制，因此GDP也较少，这样的情况集中在GDP比较的国家（图表的左边）。

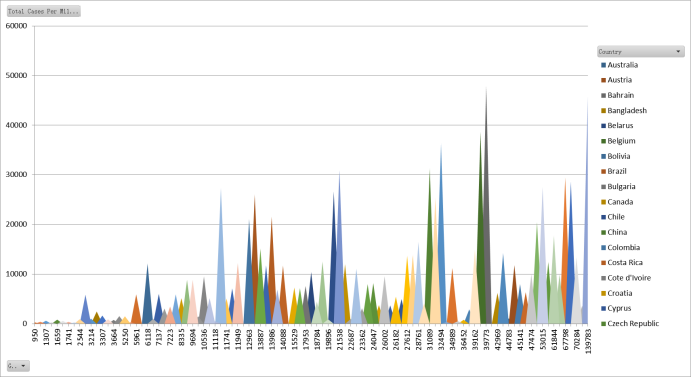
2.国家很发达，GDP较高，医疗水平高并且人口也不是特别多（图表右边的一些国家）。

死亡数比感染数多出很多的国家是一些人口很多的落后国家，因为这种情况发生在图表左边，这些国家的GDP比较少，死亡数目却很高，医疗水平较低，检测到的感染数远少于死亡数目。

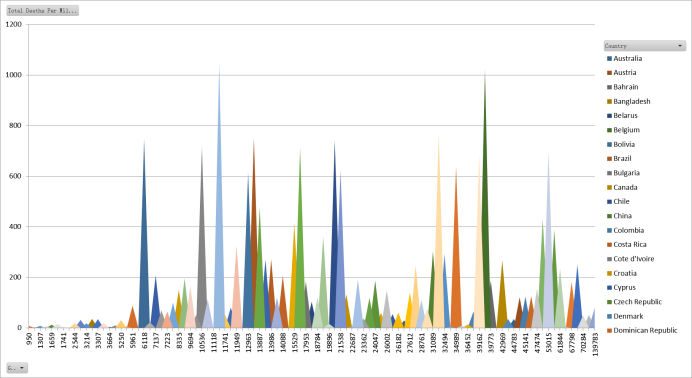
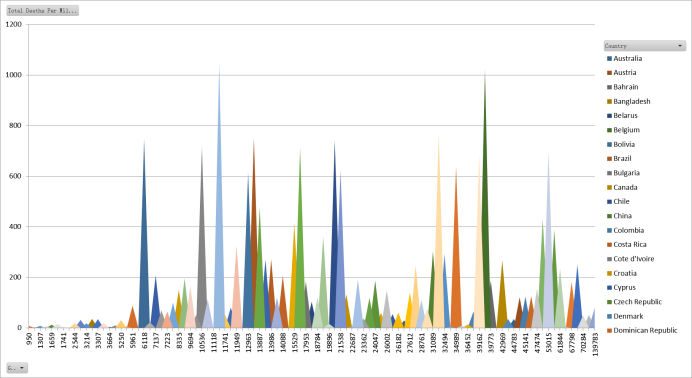
死亡数比感染数多一些的国家的GDP分布在图表的中间，不落后但是也不够发达，人均医疗水平不够高，不能及时发现和治疗一些感染者。

死亡数不大于感染数的国家的GDP比较高，医疗水平比较发达、医疗体系比较完善，感染者通常能被及时发现和治疗。

**COVID-19感染情况与经济水平的关系**

****

**COVID-19死亡情况与经济水平的关系**

****

## 从人口密度维分析

由图可知人口密度较高的国家普遍感染与死亡人数也较多，人口密度较小的国家普遍感染与死亡人数不高，其中的原因可能有几点：

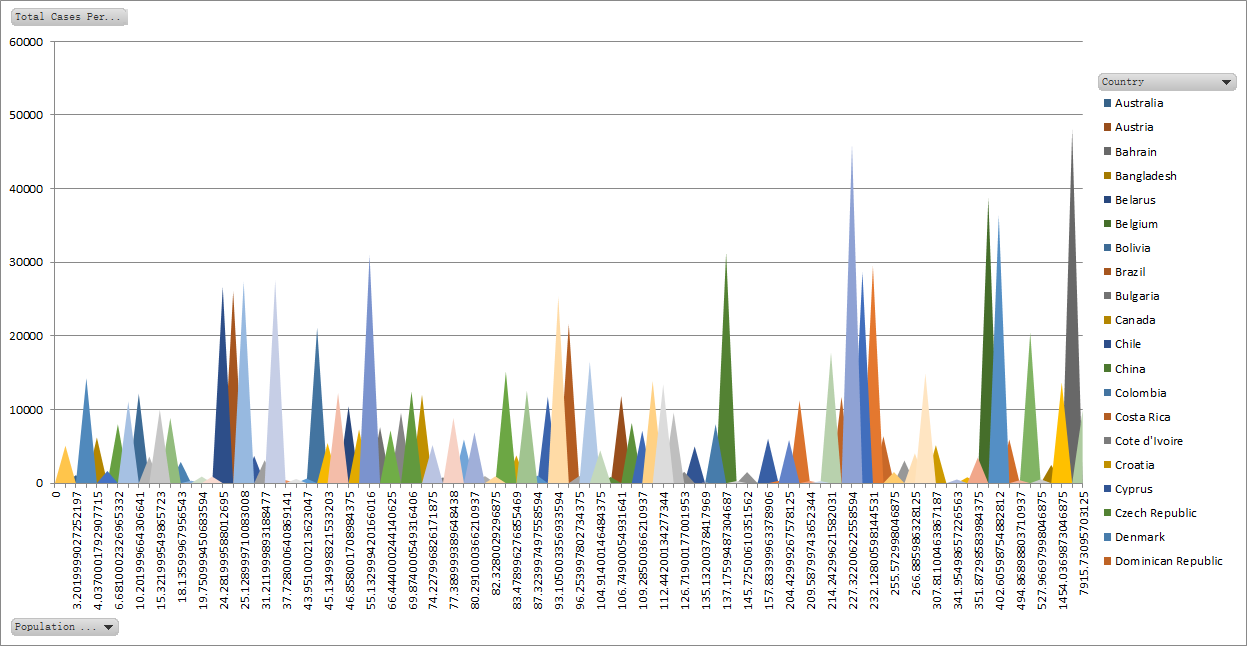
1.人口密度越高国民聚集度越高，在人群聚集的情况下病毒更容易传播

2.人口密度越高的国家国际人流量越高，可能导致国民在外国感染后回国传播的概率变高

3.人口密度较高的国家多数发展水平不高，医疗水平也相对较低，控制疫情的能力较为薄弱

人口密度与感染和死亡数之间的关联性并没有非常高，可以观察到有部分人口密度较高的国家感染与死亡人数较少，人口密度较低的国家感染与死亡人数较多，其中原因可能与医疗水平等其他因素有关。

**COVID-19感染情况与人口密度的关系**



**COVID-19死亡情况与人口密度的关系**

