可视化实验七（大作业二）报告

201900161140 张文浩

实验完成时间：11.7

软件环境：Tableau

实验题目：

Exploratory Data Analysis 探索性数据分析
https://courses.cs.washington.edu/courses/cse512/21sp/a2.html

实验步骤：

第一步：

选择数据集

我选择的数据集是关于新冠疫情的，在 WHO 世界卫生组织官网下载。数据集包含了三个文件，分别是：

①WHO-COVID-19-global-data.csv
②WHO-COVID-19-global-table-data
③vaccination-data

第二步：

分析数据集

观察数据集中包含了哪些变量和信息

①WHO-COVID-19-global-data.csv

Date_reported	Country_code	Country	WHO_region	New_cases	New_deaths
2020/1/3	AF	Afghanistan	EMRO	0	0
2020/1/4	AF	Afghanistan	EMRO	0	0
2020/1/5	AF	Afghanistan	EMRO	0	0
2020/1/6	AF	Afghanistan	EMRO	0	0
2020/1/7	AF	Afghanistan	EMRO	0	0
2020/1/8	AF	Afghanistan	EMRO	0	0
2020/1/9	AF	Afghanistan	EMRO	0	0
2020/1/10	AF	Afghanistan	EMRO	0	0
2020/1/11	AF	Afghanistan	EMRO	0	0

a) 时间：从 2020 年 1 月 3 日开始记录，到 2021 年 1 月 3 日（因为数据集是我在 2021 年 1 月 4 日下载的）。

b) country_code：国家编号，与国家名称完全等价，可以忽略不计

c) country：国家名称
d) WHO_region：世卫组织划分的地区，包括非洲区域 AFRO、美洲区域 AMRO、东南亚区域 SEARO、欧洲区域 EURO、东地中海区域 EMRO、西太平洋区域 WPRO。
e) New_cases：对应日期每个国家当日新增确诊病例人数。
f) New_deaths：对应日期每个国家当日新增死亡病例人数。

②WHO-COVID-19-global-table-data

这个表中属性较多，我只选择了我认为有用的属性。

Name	WHO Region	Cases - cumulative total	Cases - cumulative total per 100000 population	Cases - newly reported in last 7 days
Global		247968227	3181.306602	3045362
United States of America	Americas	45889496	13863.785	527691
India	South-East Asia	34321025	2487.023	89216
Brazil	Americas	21821124	10265.894	72140
The United Kingdom	Europe	9171664	13510.39	274511
Russian Federation	Europe	8673860	5943.668	281163
Turkey	Europe	8121226	9629.258	185219
France	Europe	6956857	10696.38	39087
Iran (Islamic Republic of)	Eastern Mediterrane	5954962	7089.836	66862

Deaths - cumulative total	Deaths - cumulative total per 100000 population	Deaths - newly reported in last 7 days
5020204	64.40667145	49359
743140	224.512	8693
459652	33.308	3266
608071	286.071	1825
141181	207.968	1140
243255	166.688	8198
71298	84.537	1529
115386	177.409	188
126763	150.921	1047
116010	256.683	144
87462	184.782	55

a) Name：国家
b) WHO_region：世卫组织划分的地区
c) cases_cumulative total：累计确诊病例
d) Cases - cumulative total per 100000 population：平均每 100000 人中确诊病例人数
e) Cases - newly reported in last 7 days：七天内（对于 2021.11.4 来说）新增确诊病例人数
f) deaths_cumulative total：累计死亡病例
g) deaths - cumulative total per 100000 population：平均每 100000 人中死亡病例人数
h) deaths - newly reported in last 7 days：七天内（对于 2021.11.4 来说）新增死亡病例人数

③vaccination-data

属性也比较多，我直接去掉了我认为不是很重要的。

COUNTRY	WHO_REGION	DATA_SOURCE	DATE_UPDATED	PERSONS_VACCINATED_1PLUS_DOSE	PERSONS_FULLY_VACCINATED	PERSONS_FULLY_VACCINATED_PER100	VACCINES_USED
Falkland Islanc	AMRO	OWID	2021/4/14	2632	1775	50.962	AstraZeneca - AZ
Saint Helena	AFRO	OWID	2021/5/5	4361	3531	58.162	AstraZeneca - AZ
Greenland	EURO	OWID	2021/11/2	40121	37085	65.323	Moderna - mRN
Faroe Islands	EURO	OWID	2021/11/2	40002	38006	77.778	Moderna - mRN
Jersey	EURO	OWID	2021/10/27	78809	74589	69.195	Moderna - mRN
Guernsey	EURO	OWID	2021/11/1				Moderna - mRN
Liechtenstein	OTHER	OWID	2021/11/1	25070	24440	63.076	Moderna - mRN
Gibraltar	EURO	OWID	2021/10/31	40583	39814	118.174	Pfizer BioNTech
Isle of Man	EURO	OWID	2021/11/2	67894	64532	75.891	Moderna - mRN

a) country：国家
b) WHO_region：世卫组织划分的地区
c) Data_sourse：数据来源，本次任务不关心这个属性
d) DATE_UPDATED：数据更新时间，本次任务不关心这个属性

e) PERSONS_VACCINATED_1PLUS_DOSE：至少注射一针疫苗的人数
f) PERSONS_FULLY_VACCINATED：完全注射疫苗的人数
g) PERSONS_FULLY_VACCINATED_PER100：平均每 100 人中完全注射疫苗的人数
h) VACCINES_USED：疫苗的种类，本次任务不关心这个属性

第三步：

数据集健全性检查

经过检查，本数据集不存在质量问题。

第四步：

想要调查的问题。

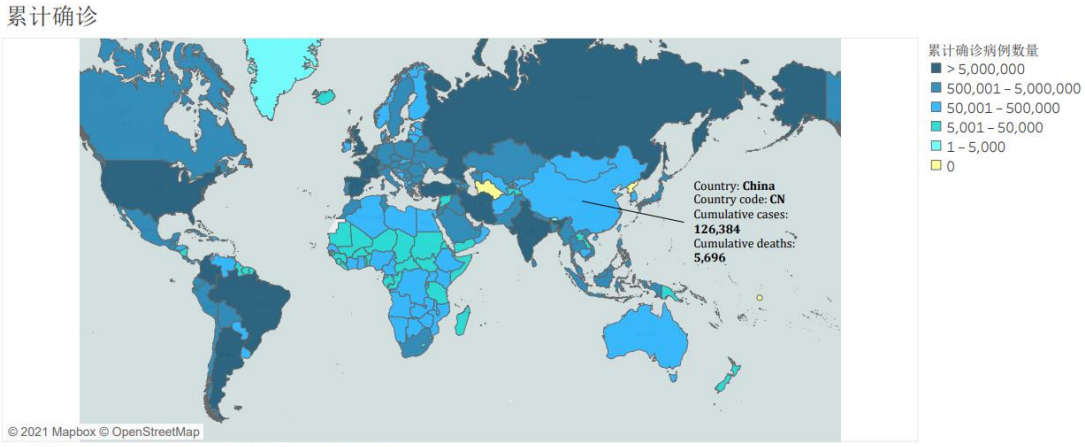
①全球累计确诊/死亡病例的分布情况。
②全球疫苗接种情况
③自疫情开始以来，疫情的发展趋势
④我们关注的某几个国家疫情发展趋势

第五步：

针对上述问题，利用可视化解释数据集。

①全球累计确诊病例的分布情况。

第一组图，是利用表 **WHO-COVID-19-global-data.csv** 进行制作的，问题是我们只有每一天的新增确诊/死亡病例人数，我们想得到总的确诊/死亡病例人数，只需要建立一个新的变量，以国家为单位分组，对每天新增确诊/死亡病例人数进行求和，即可得到累计的人数。



截止**2021**年**11**月**4**日全球累计新冠肺炎确诊人数
累计确诊病例：**247,968227**
累计死亡病例：**5,020,204**

累计死亡

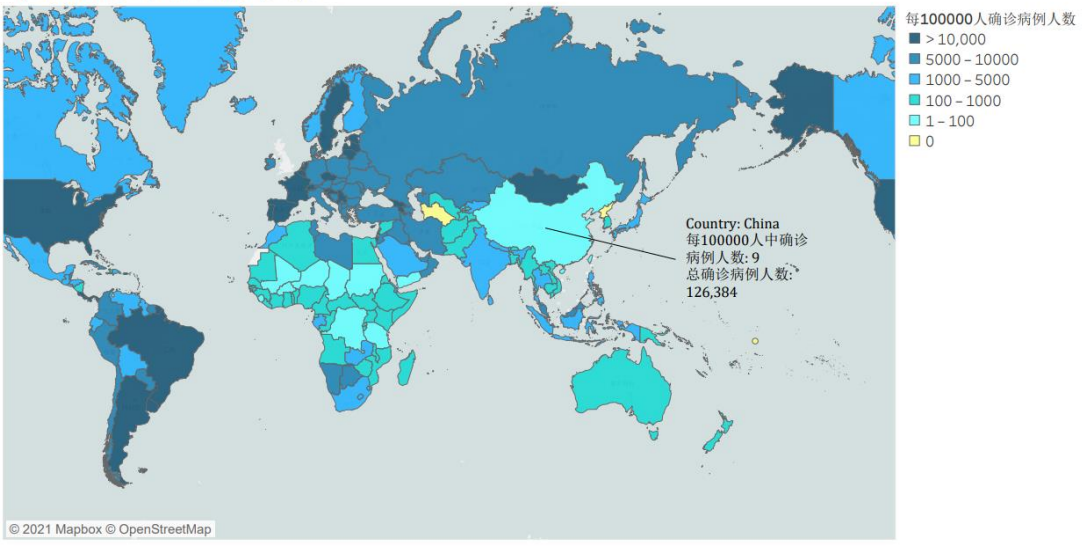
截止**2021**年**11**月**4**日全球累计新冠肺炎确诊人数
累计确诊病例：**247,968227**
累计死亡病例：**5,020,204**

在第一组图中，我们以颜色的深浅表示每个国家确诊/死亡人数的多少，颜色越深，表示越严重，右边有图例标注，最下方也有总的数据的说明。

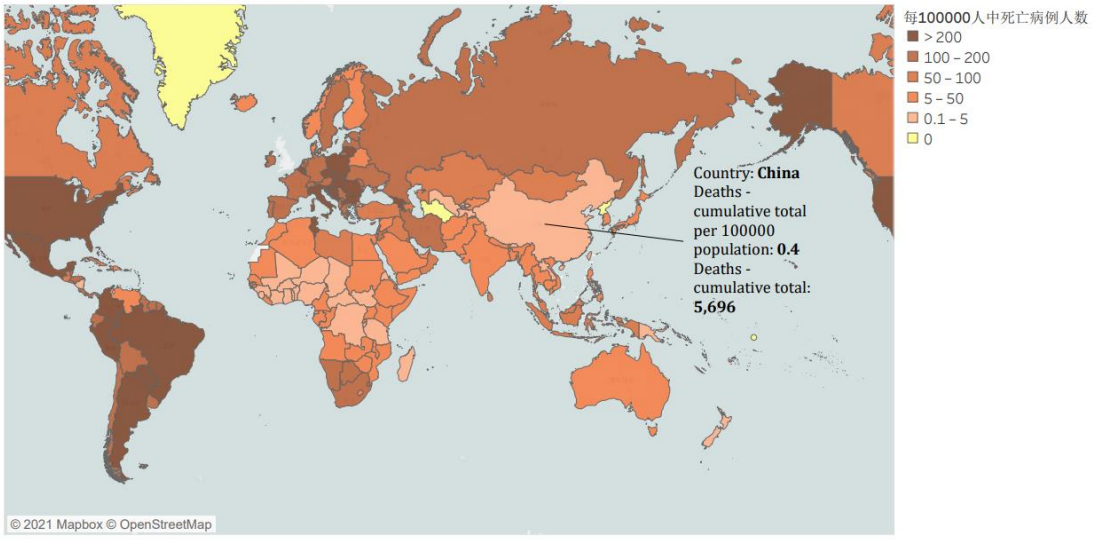
我是利用 Tableau 软件进行制作的，在软件中可以利用鼠标交互查看每个国家的具体信息，但是在图片中无法展示出来，所以我就示例性地标注出中国的详细信息。

不过，仅仅通过人数的信息，不能全面客观的描述国家的疫情情况，因为不同国家人口差异很大，所以我们可以用一个国家中每 100000 人中确诊病例人数来描述国家疫情的严重程度。

每**100000**人中确诊病例人数

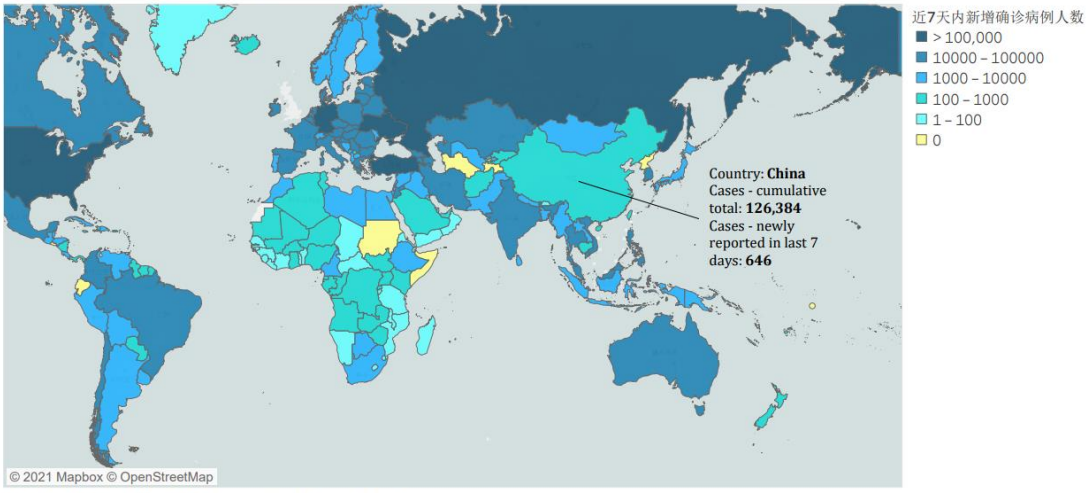


每**100000**人中死亡病例人数



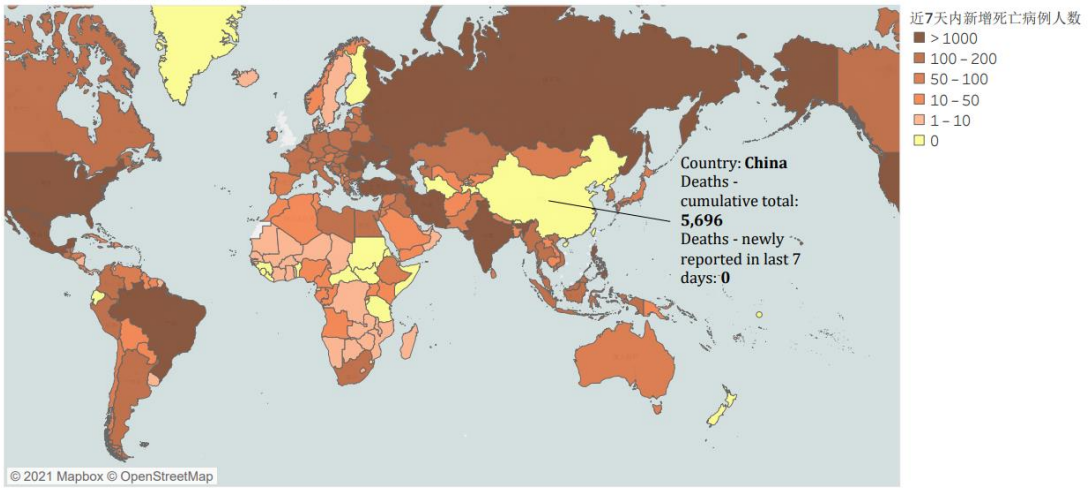
上面两组图没有考虑到时间的变量，所以无法表示近期（2021.11.4）的疫情情况，于是可以用第二张表 **WHO-COVID-19-global-table-data** 中的信息，绘制出近七天新增确诊/死亡病例人数的可视化图。

近**7**天内新增确诊病例人数



截止**2021**年**11**月**4**日近**7**天内新增确诊病例人数

近7天内新增死亡病例人数



截止**2021**年**11**月**4**日近**7**天内新增死亡病例人数

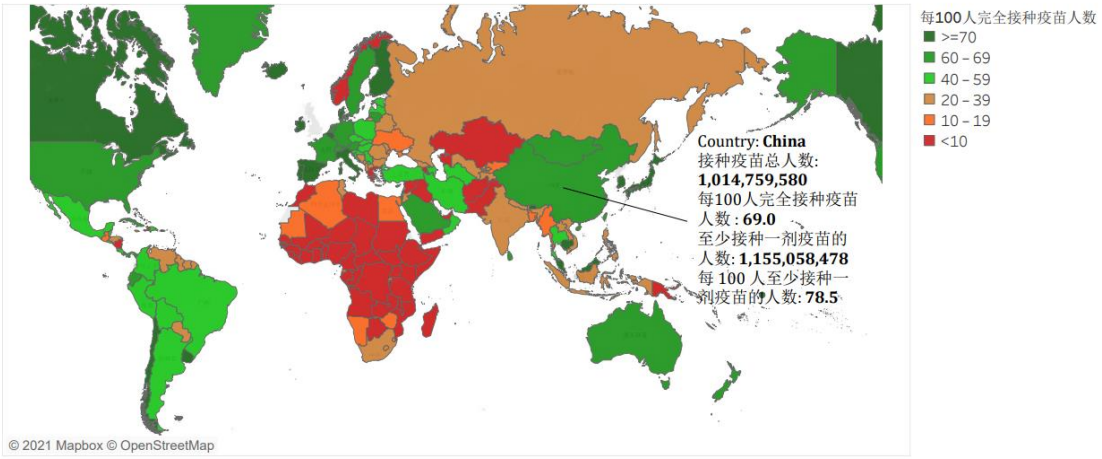
在上面的图像中，我们可以看到美国、印度及非洲地区的确诊人数、死亡人数都很高，说明疫情在这些国家地区比较严重。

②全球疫苗接种情况

针对第二个问题，我们可以用第三表 **vaccination-data** 中的数据来回答。

我们看到 **vaccination-data** 表中，有很多数据，但如果我们想看到每个国家疫苗的“普及率”，不应该用总的接种人数，而是用接种比例来进行可视化，所以我们可以根据每个国家每 100 人完全接种疫苗人数来为地图上色。

疫苗接种



截止**2021**年**11**月**4**日全球疫苗接种情况
完全接种疫苗总人数：**2,994,198,155**

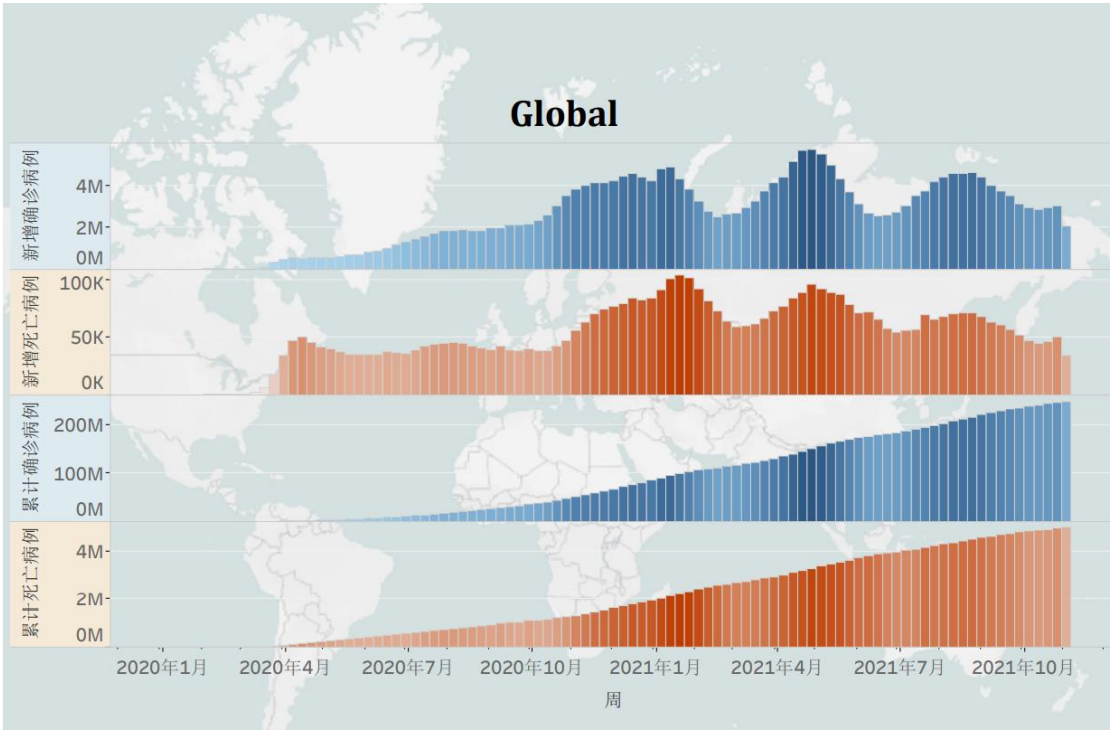
颜色越绿表明这个国家疫苗的“普及率”越高，同时接种人数等信息可以通过标注显示出来，因为图片没法显示鼠标交互效果，所以也是示例性地表示出中国的交互标记信息。把鼠标放到对应的国家上就可以显示出接种疫苗总人数、每 100 人完全接种疫苗人数、至少接种一剂疫苗的人数、每 100 人至少接种一剂疫苗的人数的信息。

在上面的可视化结果中，可以看出，中国、澳大利亚、加拿大、美国和欧洲部分地区的疫苗普及率较高。

③自疫情开始以来，疫情的发展趋势

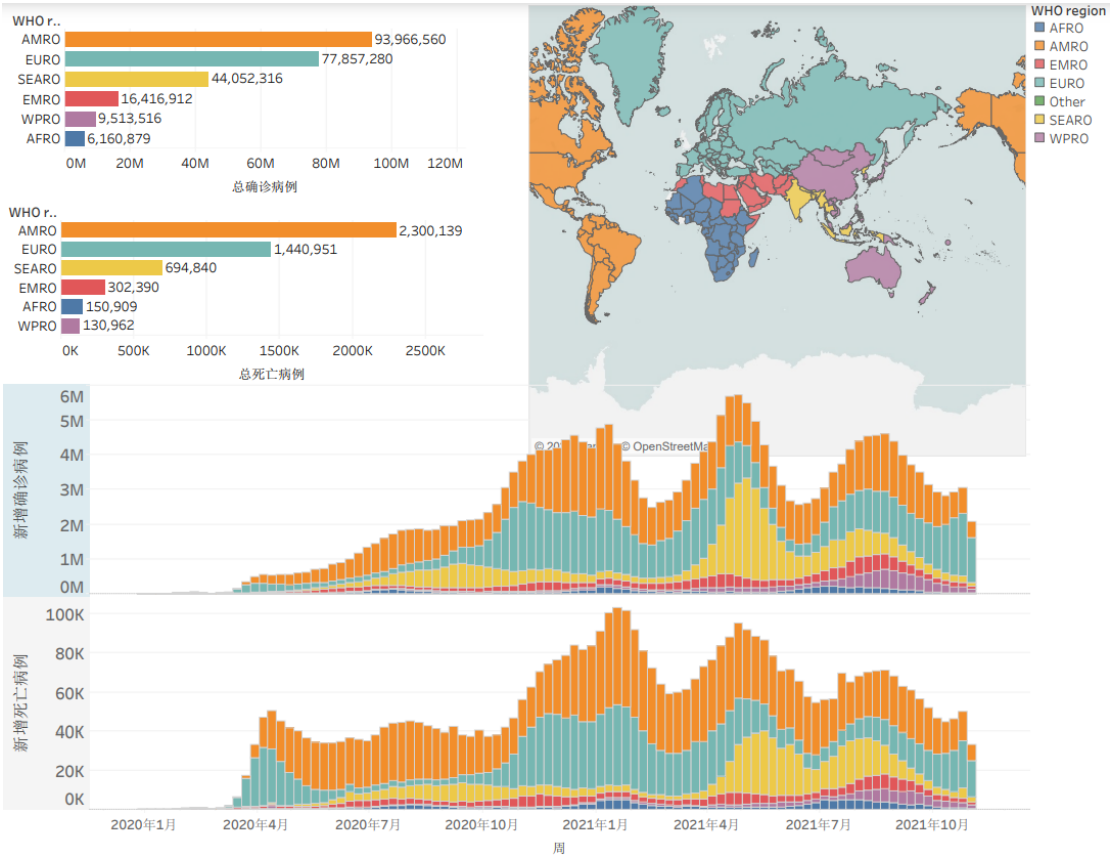
利用柱状图实现，每日新增的确诊/死亡病例人数是已有的数据，累计的确诊/死亡病例人数在 Tableau 软件中可以很方便的利用累加算出来，同时我们加上颜色信息，单日病例人数增长越大颜色越深，蓝色表示确诊病例，红色代表死亡病例。

因为我们的日期是从 2020 年 1 月开始，直到 2021 年 11 月结束，其中有六百多天，如果我们以天为单位来绘制柱状图，x 坐标会很密集，效果不好。于是我用周（每 7 天）为单位进行绘制，每个“柱”表示这一周的 7 天的新增病例加起来的大小，这样效果会好很多。



我们注意到数据集中有一个属性为 WHO_Region 表示世卫组织划分的地区，可以利用更加丰富的颜色变量，描述出不同地区之间的对比情况。

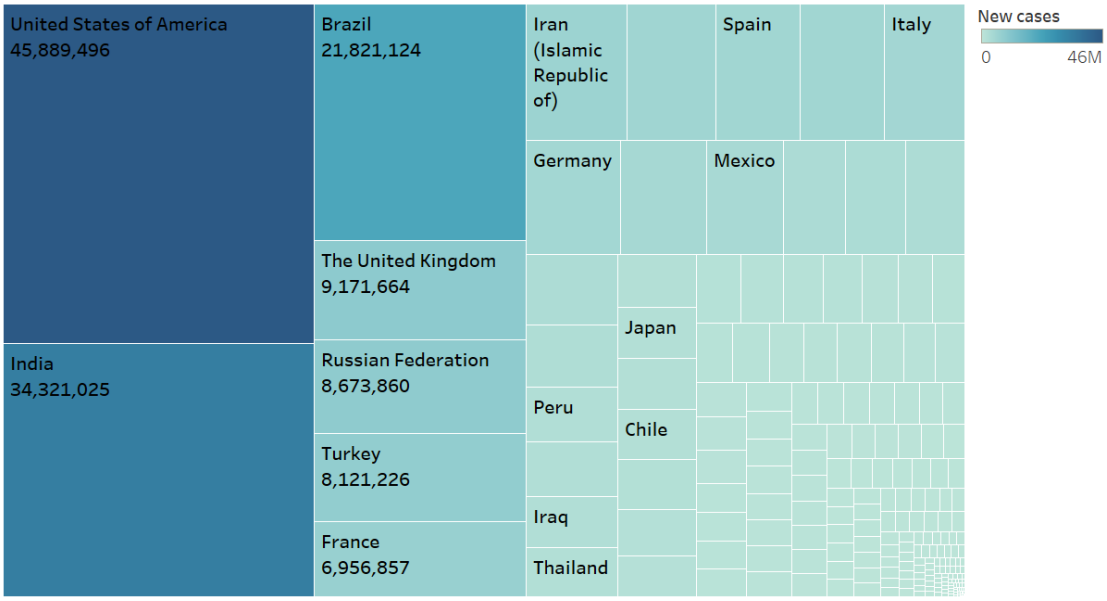
在上面的图像中，可以看到全球疫情在 2021 年 1 月和 4 月是疫情增长的高峰期。



用不同的颜色表示不同的地区，可以非常直观的看出每个地区之间确诊病例死亡病例的比较。

根据这一张图可以看出，美洲地区和欧洲地区是疫情最严重的两个地区。

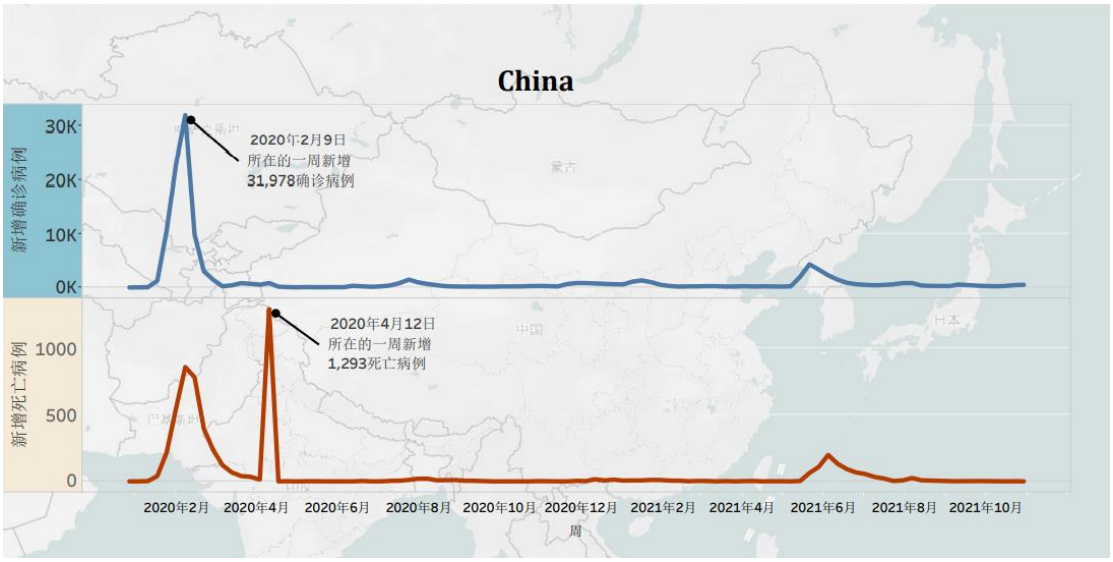
tree_map



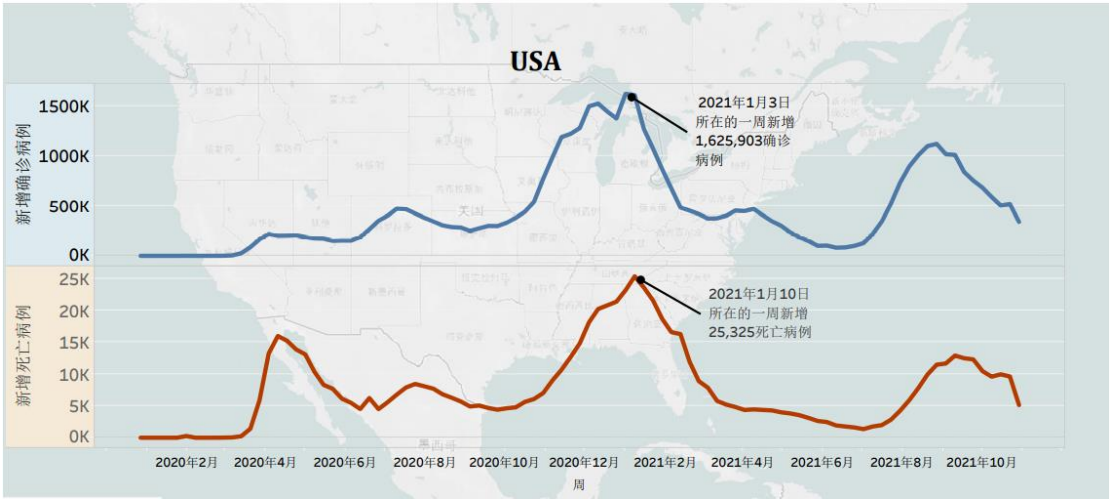
利用这个 tree_map 可以看出，美国、印度、巴西是累计确诊病例最多的三个国家

④我们关注的某几个国家疫情发展趋势

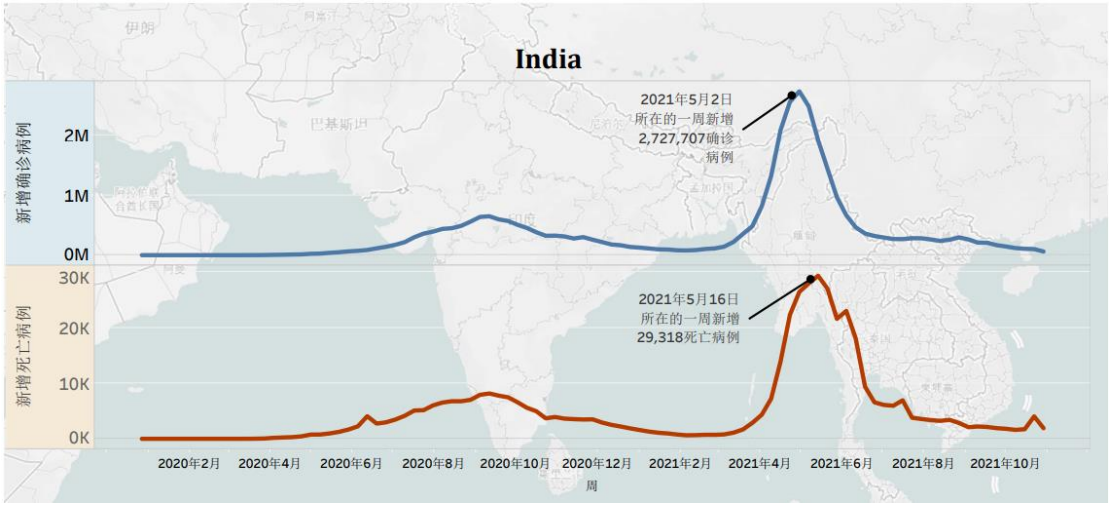
我们还可以单独绘制每个国家的疫情发展情况，背景用对应国家的地图，更加美观。与之前一样，为了使折线图更加平滑美观，以周而不是天做单位。单独标注出最大值点。这里我分别绘制了中国、美国、印度。



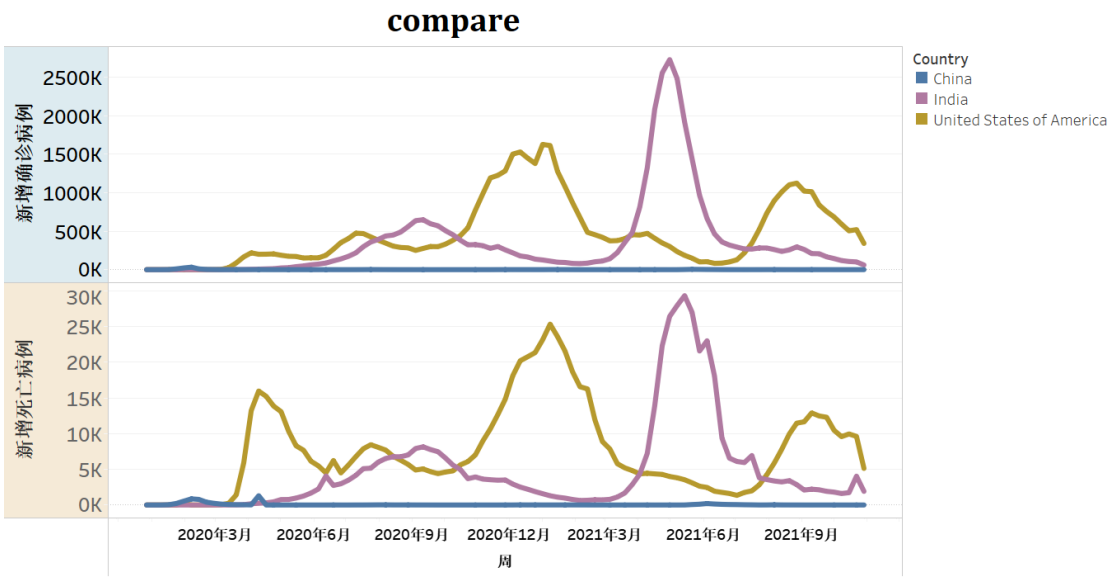
中国在 2020 年 2 月迎来了疫情的爆发，2020 年 4 月死亡病例激增。但在后面除了 2021 年 6 月经历了较小的波动外，基本控制住了疫情。



美国整体病例增长率一直很高，尤其是在 2020 年末和 2021 年初，并且反复现象明显，每日新增病例居高不下。



印度的疫情在 2021 年 5 月左右大爆发，一周内新增确诊病例将近三百万。



对比中国、美国、印度确诊和死亡病例可以发现美国和印度的疫情比中国严重得多

END

全部可视化图像可见另一个 pdf 文件，或 Tableau 工程文件。