# 报告目录

# 一 分析背景与目的

## 1.1 分析背景

新型冠状病毒（简称新冠）会通过一段极短的时间在全球范围内大规模流行，正是由于它的高传播速度以及高致死率，社会迫切需要疫情相关的信息。自疫情爆发以来，各级政府部门在第一时间通过各种渠道及时发布第一手数据，众多组织及个人也采取迅速行动，利用多种分析手段为公众提供疫情分析和解读，从而更好的抗击疫情，推动我国打赢疫情防控狙击战。

## 1.2 分析目标

- 统计疫情数据
- 设计可视化数字大屏，展示新冠疫情的时空变化情况
- 利用可视化工具，绘制城市疫情风险图
- 分析国内和国际的疫情变化情况

## 1.3 分析工具

本文主要使用的是 Python 工具，版本号为 3.7.0，其中主要使用的第三方库如下：

- Pandas（数据读取与处理）
- Numpy（科学计算）
- PowerPoint、Matplotlib、Pyecharts（可视化）
- DataV（可视化数字大屏）

# 二 分析思路

## 2.1 数据的基本处理分析思路

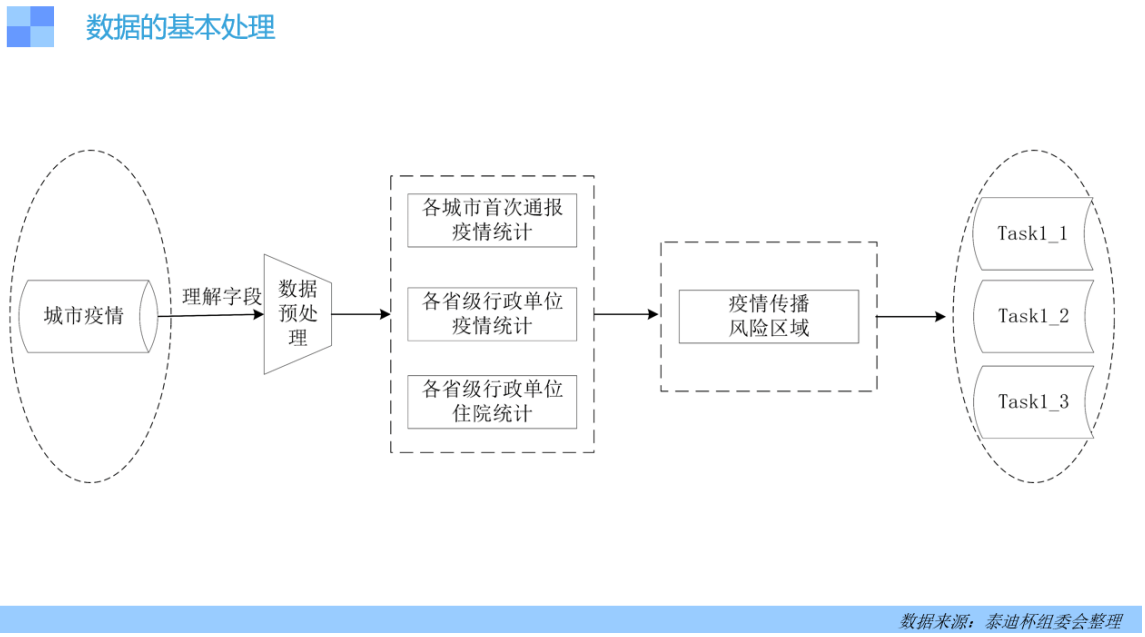



**图 1 数据的基本处理流程图**

## 2.2 数字大屏设计分析思路

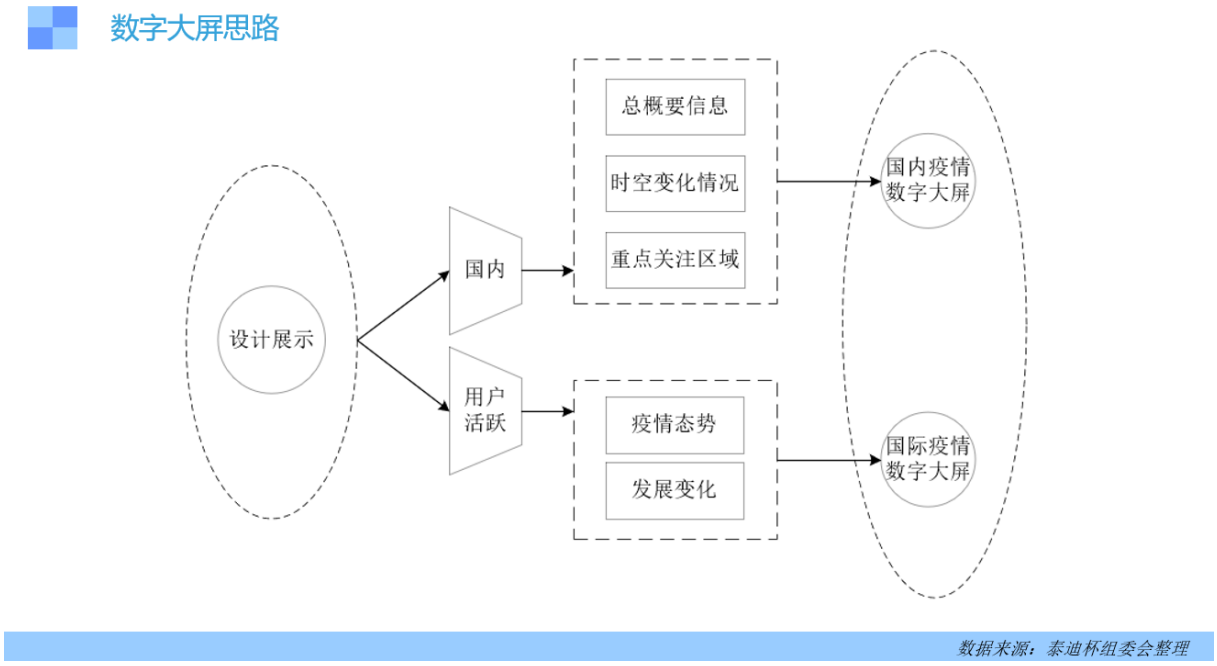



**图 2 数字大屏设计流程图**

## 2.3 国际疫情发展分析分析思路

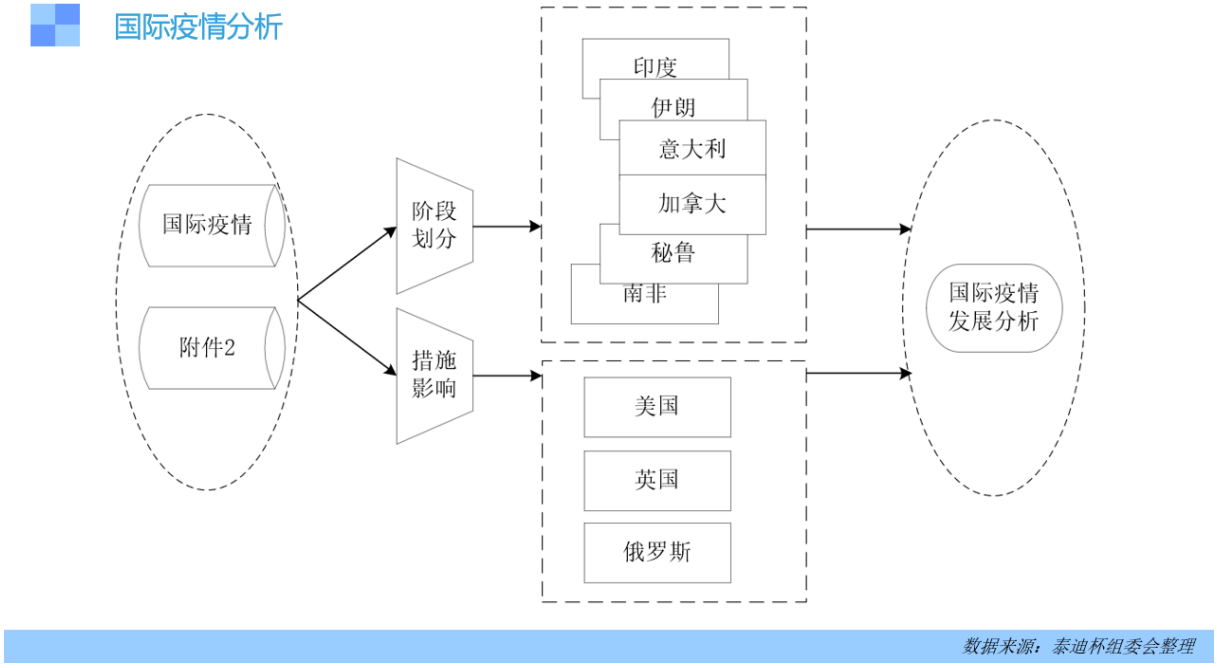



**图 3 国际疫情发展分析流程图**

# 三 数据说明

## 3.1 数据集描述

- 数据集名称：疫情分析数据
- 数据集来源：泰迪杯组委会整理
- 数据时间范围：2020 年 1 月 1 日至 2020 年 6 月 30 日

## 3.2 数据集信息

本次数据主要是新冠疫情相关的数据，包括城市疫情、城市省份对照、国际疫情以及 A 市涉疫场所分布数据。

## 3.3 数据量

- 城市疫情：10245 条记录
- 城市省份对照：479 条记录
- 国际疫情：25451 条记录
- A 市涉疫场所分布：152 条记录

# 四 分析内容

## 4.1 数据的基本处理

### 4.1.1 各城市首次通报疫情统计

**1）理解字段的含义**

原数据表的字段的列名非常规范，无需对列名称进行重命名，数据分析之前需要重复理解每个数据表所代表的信息。

各个表字段主要由三方面信息构成，分别是时间、地点、人数，最重要的是理解新增确诊、新增治愈、新增死亡。新增确诊是当天新检测出的人数，新增治愈是当天治愈人数，新增死亡则是当天死亡的人数，需要假设的是治愈后的人群不再能够被感染。

**2）数据预处理**

通过观察数据发现，本次数据集较为干净，不存在缺失值和重复值，同时数值型变量也不存在异常值，不需要进行数据清洗操作。

我们为了方便后续工作开展，考虑对数据进行规范化处理，主要是进行了格式的转换。

**表 1 数据规范化**

| 表面-列名 | 规范化操作 |
|---|---|
| 城市疫情-日期 | 转化为 datetime 格式 |
| 城市疫情-新增确证，新增治愈，新增死亡 | 转化为 int32 格式 |
| 国际疫情-日期 | 转化为 datetime 格式 |
| 国际疫情-新增确证，新增治愈，新增死亡 | 转换为 int32 格式 |
| A 市涉疫场所分布-通过日期 | 转化为 int32 格式 |
| A 市涉疫场所分布-横坐标（公里），纵坐标（公里） | 转化为 float32 格式 |

**3）分组聚合统计结果**

由于需要统计各城市自首次通报确诊病例后至 6 月 30 日的每日累计确诊人数、累计治愈人数和累计死亡人数，我们采用数据透视表的方式，将 index 设置为“城市”与“日期”，将 values 设置为新增确证、新增治愈以及新增死亡，而 aggfunc 采用 numpy.sum 的方式，创建的透视表可以基本满足需求。

同时为了使得时间连续，以各个城市首次确证时间作为起始点，6 月 30 日作为终点进行填充缺失的数据，最后用循环的方式可以得到想要的结果，保存至 task1_1 文件。

**4）展示三城统计结果**

生成的武汉、深圳、保定每月 10、25 日统计结果如下表所示：

**表 2 三城统计结果**

| 城市 | 日期 | 累积死亡 | 累积治愈 | 累积确诊 |
|---|---|---|---|---|
| 武汉 | 2020/1/10 | 1 | 2 | 41 |
| 武汉 | 2020/1/25 | 45 | 40 | 618 |
| 武汉 | 2020/2/10 | 748 | 1173 | 18454 |
| 武汉 | 2020/2/25 | 2085 | 12026 | 47441 |
| 武汉 | 2020/3/10 | 2423 | 33264 | 49978 |
| 武汉 | 2020/3/25 | 2531 | 44020 | 50006 |
| 武汉 | 2020/4/10 | 2577 | 46154 | 50008 |
| 武汉 | 2020/4/25 | 3869 | 46452 | 50333 |
| 武汉 | 2020/5/10 | 3869 | 46464 | 50339 |
| 武汉 | 2020/5/25 | 3869 | 46465 | 50340 |
| 武汉 | 2020/6/10 | 3869 | 46471 | 50340 |
| 武汉 | 2020/6/25 | 3869 | 46471 | 50340 |
| 深圳 | 2020/1/25 | 0 | 2 | 27 |
| 深圳 | 2020/2/10 | 0 | 56 | 375 |
| 深圳 | 2020/2/25 | 3 | 262 | 417 |
| 深圳 | 2020/3/10 | 3 | 387 | 417 |
| 深圳 | 2020/3/25 | 3 | 414 | 417 |
| 深圳 | 2020/4/10 | 3 | 414 | 419 |
| 深圳 | 2020/4/25 | 3 | 414 | 422 |

| | | | | |
|---|---|---|---|---|
| 深圳 | 2020/5/10 | 3 | 414 | 423 |
| 深圳 | 2020/5/25 | 3 | 414 | 423 |
| 深圳 | 2020/6/10 | 3 | 414 | 423 |
| 深圳 | 2020/6/25 | 3 | 414 | 423 |
| 保定 | 2020/1/25 | 0 | 0 | 3 |
| 保定 | 2020/2/10 | 0 | 9 | 30 |
| 保定 | 2020/2/25 | 0 | 32 | 32 |
| 保定 | 2020/3/10 | 0 | 32 | 32 |
| 保定 | 2020/3/25 | 0 | 32 | 32 |
| 保定 | 2020/4/10 | 0 | 32 | 32 |
| 保定 | 2020/4/25 | 0 | 32 | 32 |
| 保定 | 2020/5/10 | 0 | 32 | 32 |
| 保定 | 2020/5/25 | 0 | 32 | 32 |
| 保定 | 2020/6/10 | 0 | 32 | 32 |
| 保定 | 2020/6/25 | 0 | 33 | 45 |

从生成的结果可以看出，武汉累计人数明显大于深圳、保定，并且对武汉单独进行分析，结果如下图所示，从图中可以看到：

- 10 号于 15 号没有明显差别，总体分布类似
- 1 月至 3 月处于急剧上升阶段，病毒爆发；
- 3 月后进入抗疫阶段，趋于稳定。

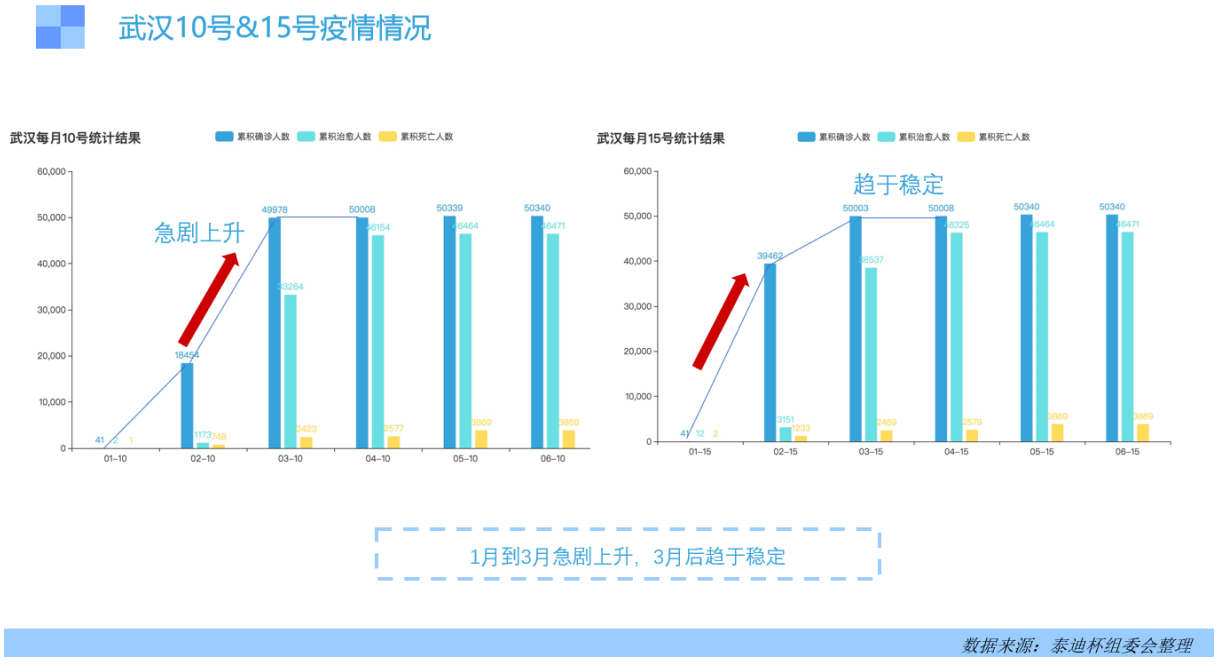



**图 4 10 号&15 号疫情情况**

#### 4.1.2 各省级行政单位疫情统计

根据任务 1.1 结果，以及有各个城市及其时间所对应的每日累计确诊人数、累计治愈人数和累计死亡人数，结合附件 1“城市省份对照表”，进行键值对匹配，利用数据透视表 pivot_table，将原有数据体现在省份上，最后生成各省级行政单位按日新增和累计数据，结果保存为 task1_2.csv。

根据生成结果，统计湖北、广东、河北每月 15 日的数据如下表所示：

**表 3 三省每日统计表**

| 省份 | 日期 | 新增确诊人数 | 新增治愈人数 | 新增死亡人数 | 累积确诊人数 | 累积治愈人数 | 累积死亡人数 |
|---|---|---|---|---|---|---|---|
| 广东 | 2020/2/15 | 22 | 50 | 0 | 1316 | 434 | 2 |
| 广东 | 2020/3/15 | 4 | 5 | 0 | 1361 | 1292 | 8 |
| 广东 | 2020/4/15 | 5 | 0 | 0 | 1571 | 1392 | 8 |
| 广东 | 2020/5/15 | 0 | 0 | 0 | 1589 | 1392 | 8 |
| 广东 | 2020/6/15 | 3 | 0 | 0 | 1628 | 1394 | 8 |
| 河北 | 2020/2/15 | 9 | 14 | 0 | 300 | 100 | 3 |
| 河北 | 2020/3/15 | 0 | 0 | 0 | 318 | 310 | 6 |
| 河北 | 2020/4/15 | 1 | 1 | 0 | 328 | 314 | 6 |
| 河北 | 2020/5/15 | 0 | 0 | 0 | 328 | 319 | 6 |
| 河北 | 2020/6/15 | 4 | 0 | 0 | 332 | 320 | 6 |
| 湖北 | 2020/1/15 | 0 | 5 | 1 | 41 | 12 | 2 |
| 湖北 | 2020/2/15 | 1839 | 849 | 139 | 56197 | 5860 | 1596 |
| 湖北 | 2020/3/15 | 4 | 816 | 14 | 67798 | 55247 | 3099 |
| 湖北 | 2020/4/15 | 0 | 33 | 0 | 67803 | 63477 | 3222 |
| 湖北 | 2020/5/15 | 0 | 0 | 0 | 68134 | 63616 | 4512 |
| 湖北 | 2020/6/15 | 0 | 0 | 0 | 68135 | 63623 | 4512 |

同时由于湖北在疫情中的特殊性，在要求外另外单独绘制湖北统计结果如图 5 所示。从图中可以看出：

- 1 月至 2 月是湖北省快速上升确诊人数的阶段，这一段时间是由于没有采取强力措施导致的；
- 2 月至 3 月的区间内，政府迅速采取有力措施，使得累计确诊人数区域稳定，正式进入抗疫阶段；
- 新增人数在 3 月后处于非常低的水平，可以看出措施的有效性以及稳定性；
- 3 月到达顶峰后趋于稳定值，这也为全国打赢攻坚战创下“定心丸”。

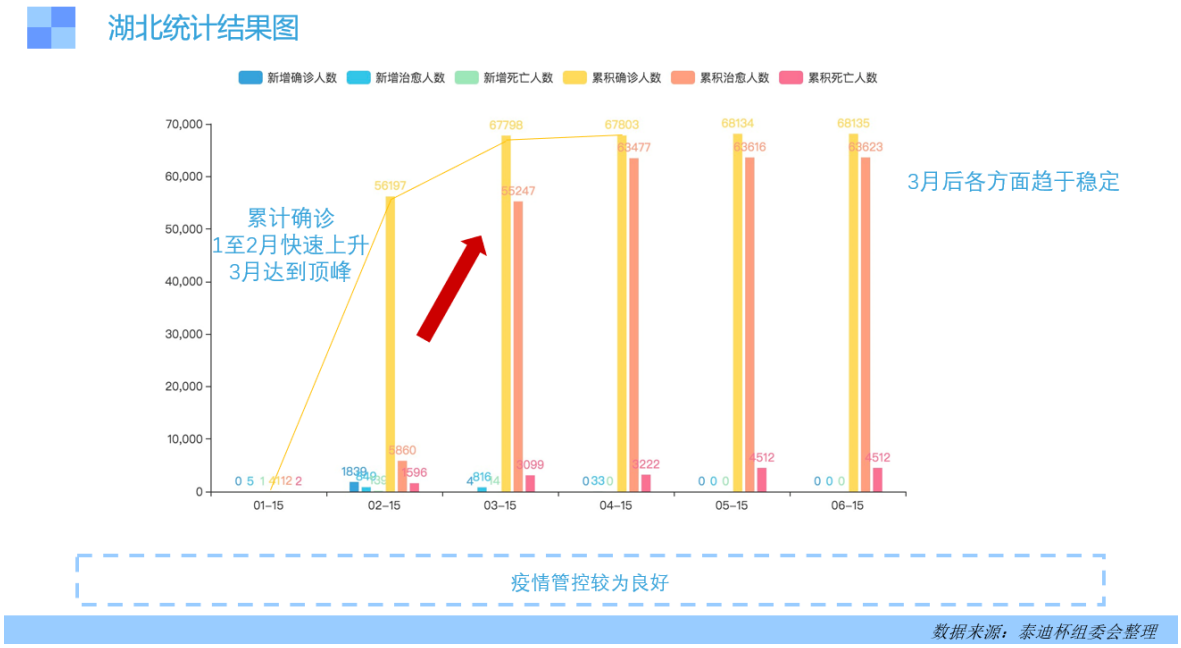



**图 5 湖北统计结果图**

### 4.1.3 各省级行政单位每日住院统计

根据任务 1.2 的结果，已经生成 task1_2.csv 关于各省级行政单位每日的新增确诊人数、新增治愈人数、新增死亡人数、累计确诊人数、累计治愈人数、累计死亡人数的数据。由于国内新冠确诊病人都入院收治，定义住院人数计算公式如下

*住院人数=累计确诊人数-累计治愈人数-累计死亡人数*

利用公式计算结果保存为 task1_3.csv，并且筛选出湖北、广东、上海城市的数据如下表所示：

**表 4 上海、广东、湖北统计结果表**

| 日期 | 上海住院人数 | 广东住院人数 | 湖北住院人数 |
|---|---|---|---|
| 2020/1/20 | 1 | 14 | 239 |
| 2020/2/20 | 135 | 665 | 48725 |
| 2020/3/20 | 55 | 89 | 5602 |
| 2020/4/20 | 115 | 182 | 102 |
| 2020/5/20 | 30 | 191 | 7 |
| 2020/6/20 | 38 | 232 | 0 |

为了更好的对三地进行比较，进行可视化呈现，从图中可以看到：

- 湖北的住院人数最多，其中在 2 月达到峰值；
- 广东、上海住院人数处于较低水平；
- 1 月到 3 月变化比较剧烈；
- 2 月可以明显看到疫情管控效果。

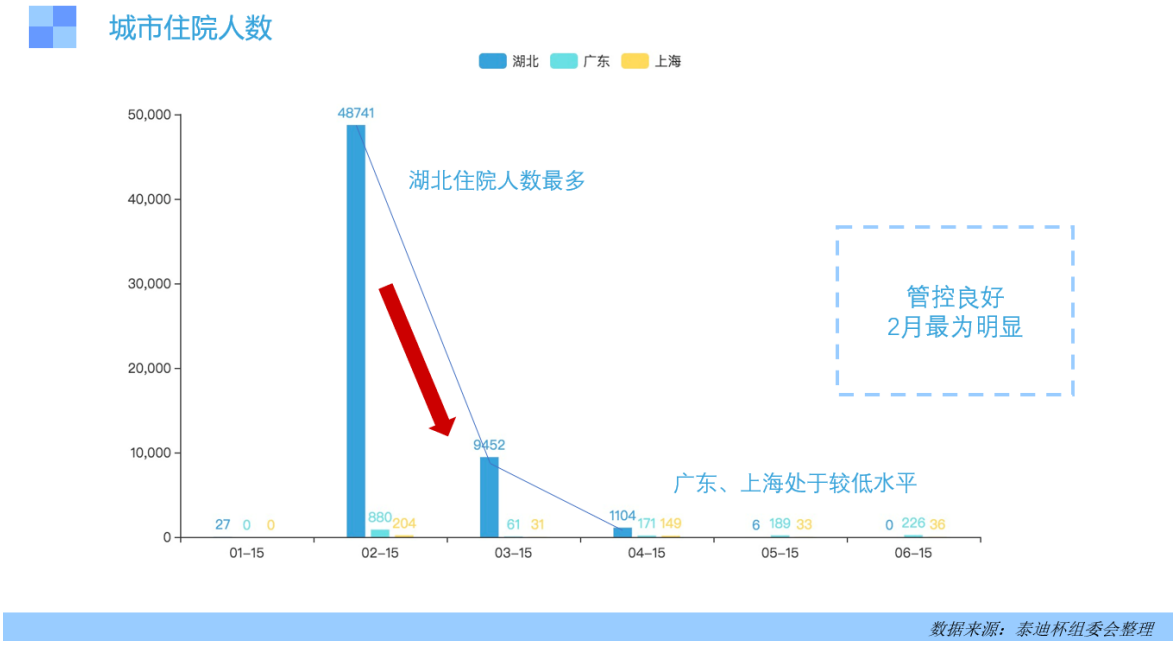



**图 6 湖北、广东、上海对比图**

### 4.1.4 A 市疫情传播风险区域

假设所有新冠患者在统计日期内均具有传染性，且未痊愈。则直到第六天，患者疫情场所是前六天的总和。

由于数据进行脱敏操作，为了绘制传播情况，无法在真实地图中展示，为此利用 pyecharts 进行散点图的绘制。

首先利用 K-Means 对这些坐标进行聚类，按照两类进行划分，分别是高危险取悦和低危险区域，聚类结果如下图所示：

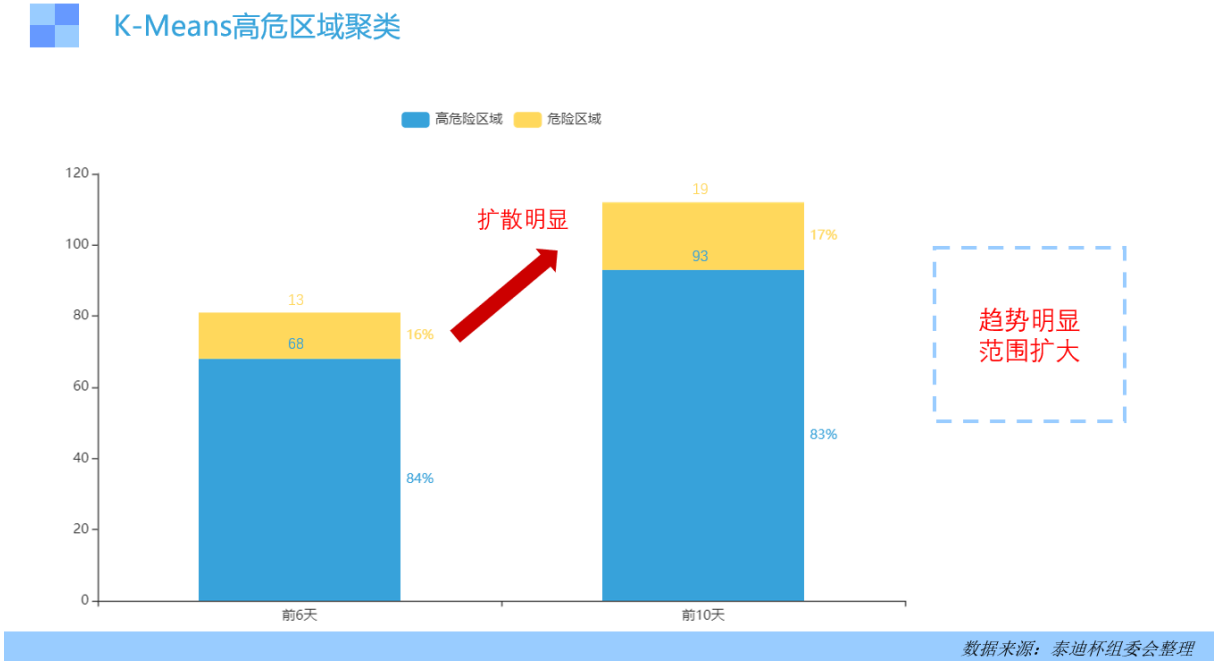



图 7 K-Means 区域聚类图

从 A 市涉疫场所分布该表中提取出前六天的横坐标与纵坐标，因每个新冠病人的传播半径为 $1km$，所以设置每个点的辐射区域为 $\pi km^2$，以此绘制散点图；同理从 A 市涉疫场所分布该表中提取出前十天的横坐标与纵坐标绘制散点图

根据 K-Means 结果进行调整，得到风险区域如下图所示：

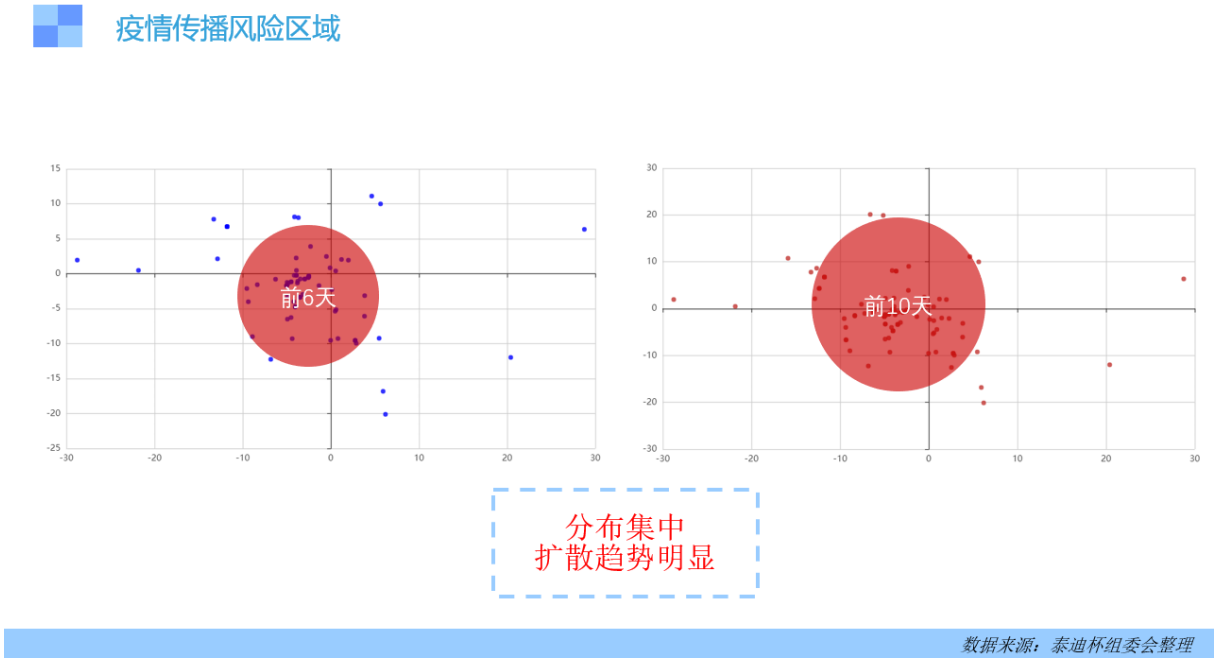



**图 8 疫情传播风险区域示意图**

由图可知，前六天的风险疫情场所主要集中在以（-3，-2）的半圆内，而 10 天后范围得到扩大，向中心区域蔓延。

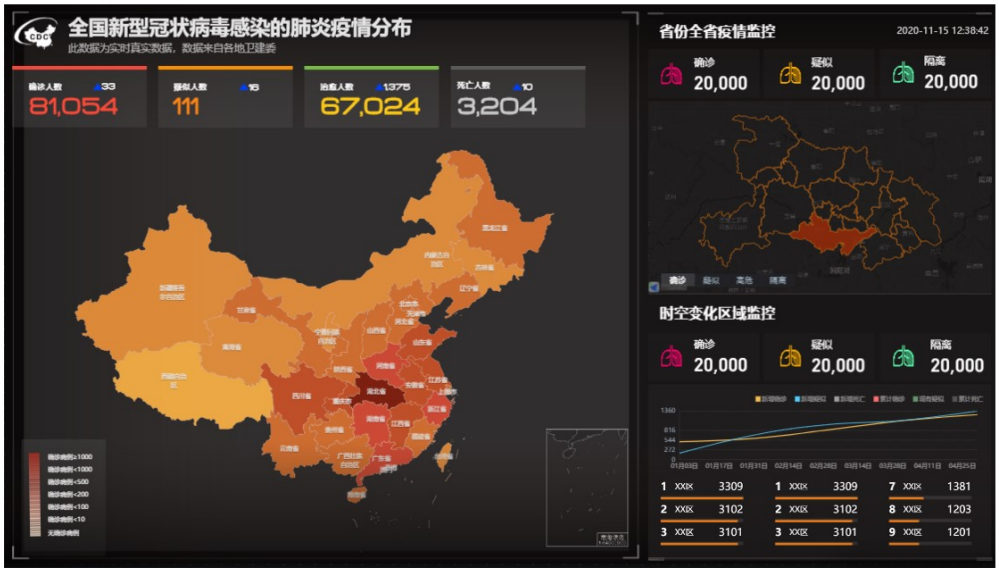
因此我们判断，以（0，0）为中心区域的范围内人流量较大，容易使得病毒得到蔓延，同时还可以清楚的看出病毒具有明显向边缘扩散的趋势以及中心疫情加剧的情况。

## 4.2 数字大屏设计

### 4.2.1 设计国内新冠疫情汇总大屏

大屏展示（数据仅供参考，并无实际意义）：

国内疫情数字大屏




*数据来源：泰迪杯组委会整理*

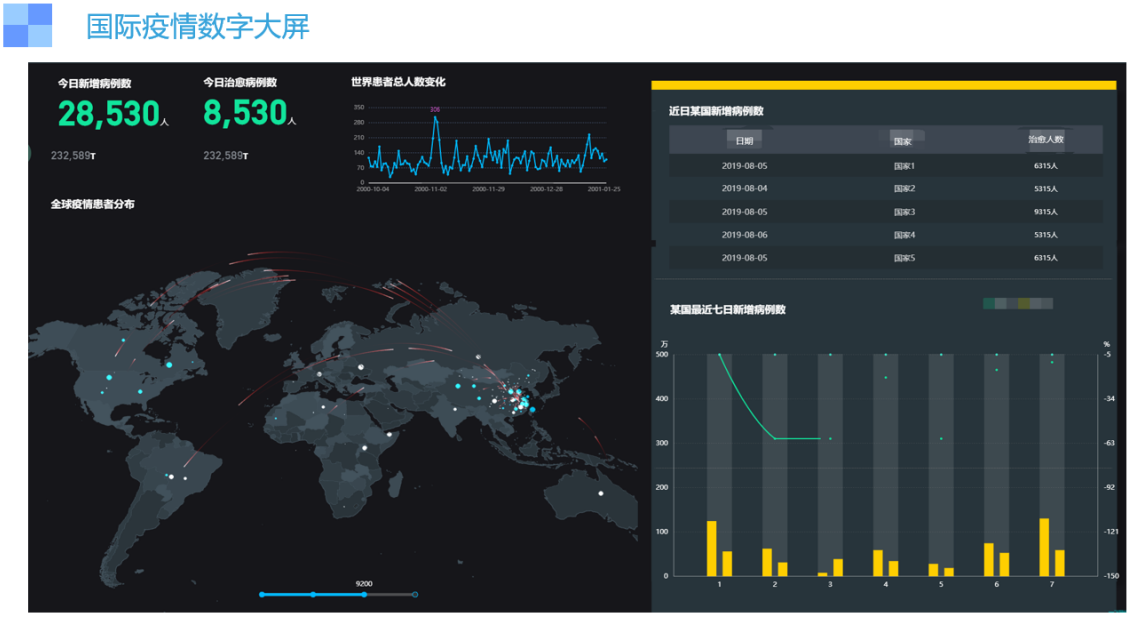
**图 9 国内疫情数字大屏展示**

设计思路如下——

- 设计国内新冠疫情数据大屏需要反映信息：国内各省份的病例密度、省内各区域病例密度、某区域病例随时间的变化。
- 画布左侧展示国内新冠疫情汇总概要信息，包括总确诊人数、疑似人数、治愈人数和死亡人数；地图上各个省份的深度表示确诊病例的人数。
- 右上角展示重点关注区域（如湖北），可以展示出省份内各区域的疫情情况，如有大幅新增情况则表现出红色；上面可以统计该省内确诊、疑似和隔离病例个数。
- 右下角的时空变化区域监控用来反馈各区域患者情况，同样统计确诊、疑似和隔离病例个数，折线图可以轮播反映时间变化时新增病例、新增确诊、新增死亡、累计确诊、现有疑似、累计死亡的增减，下面可以轮播各个区域的新增确诊来反映空间的变化。

### 4.2.2 设计国际疫情态势和发展变化大屏

大屏展示（数据仅供参考，并无实际意义）：




*数据来源：泰迪杯组委会整理*

图 10 国际疫情数字大屏展示

设计思路如下——

- 将画布分为左右两侧，左侧展示全球范围内新冠患者分布情况上方统计出今日新增的病例数和今日治愈病例数，旁边折线图反映近三个月内世界患者总人数变化。
- 画布右侧上方轮播某日某国的治愈人数，下方条形图和折线图可以反映新增病例数和新增病例数占总病例数的比重。

## 4.3 国际疫情发展分析

### 4.3.1 划分六国疫情发展阶段

由于六个国家国情存在不同，考虑实际情况，无法进行定性的划分。因此确定以下划分依据——

先对一个国家求出存活病例数，利用 K-Means 算法对存活病例数进行聚类，得到两个阈值与时间和存活病例关联的折线图进行匹配对时间进行划分，得出不同的各个时间段，即疫情的发展阶段。

$$存活病例=累计确诊-累计治愈-累计死亡$$

根据聚类结果，绘制发展阶段图：

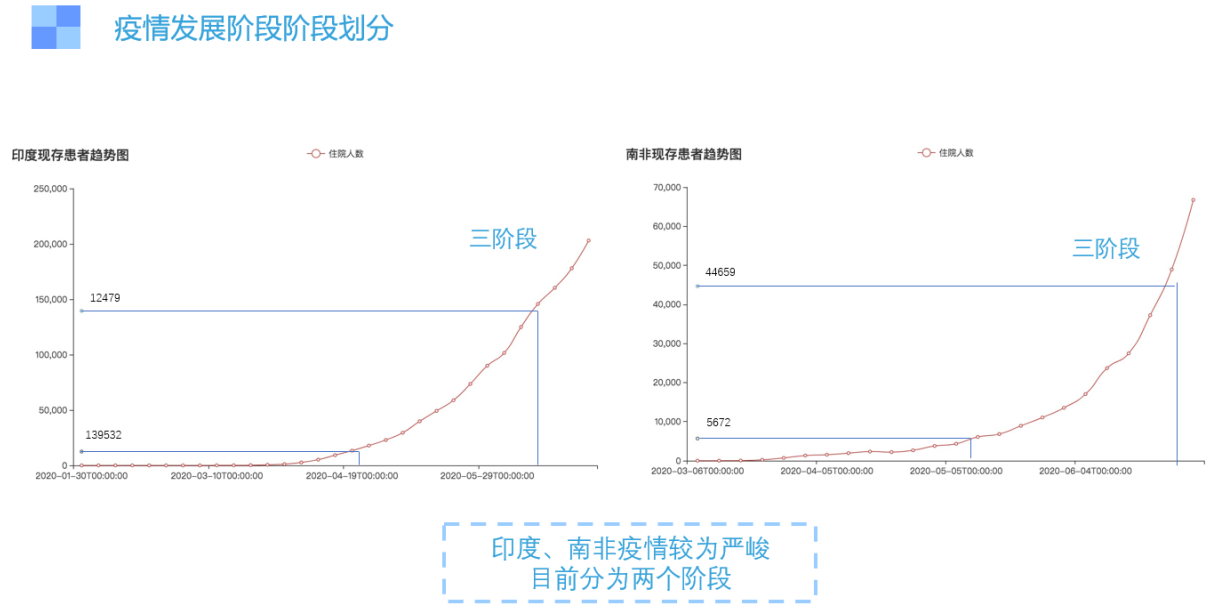



**图 11 印度、南非划分图**

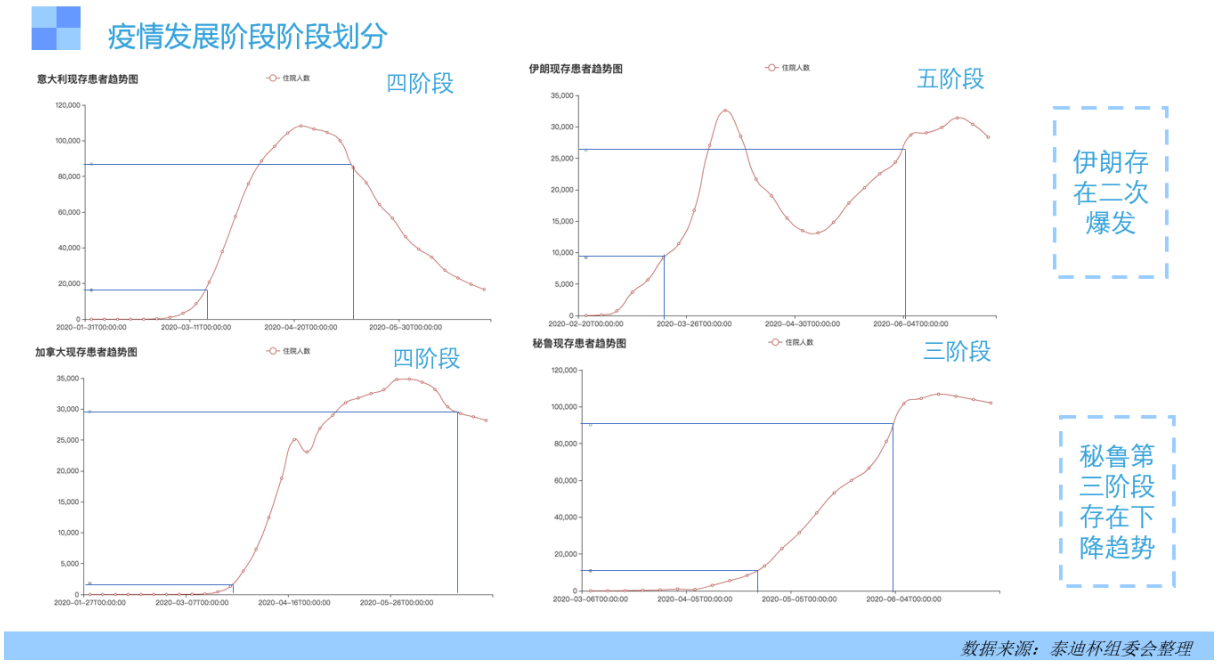



**图 12 意大利、伊朗、加拿大、秘鲁阶段划分图**

我们将这个国家划分为三阶段、四阶段以及五阶段。

- 三阶段：潜伏阶段、爆发阶段、持续阶段
- 四阶段：潜伏阶段、爆发阶段、持续阶段、治愈阶段
- 五阶段：潜伏阶段、爆发阶段、持续阶段、治愈阶段、二次爆发阶段

**从图 11 可以看出：**

印度和南非划分为三阶段，病例数呈现出上升趋势，并没有得到相应的控制。

**从图 12 可以看出：**

意大利划分为四阶段，并且第四阶段处于一个良好的变化趋势，政府及公民的努力得到了体现。

伊朗划分为五阶段，新冠病毒得到二次爆发，需要继续防控策略工作的展开。

加拿大划分为四阶段，虽然病毒得到控制，但是第四阶段效果并不显著，

秘鲁划分为三阶段，虽然第三阶段处于保持状态，但有微弱迹象显示有疫情控制的趋向。

### 4.3.2 分析三国疫情防控策略

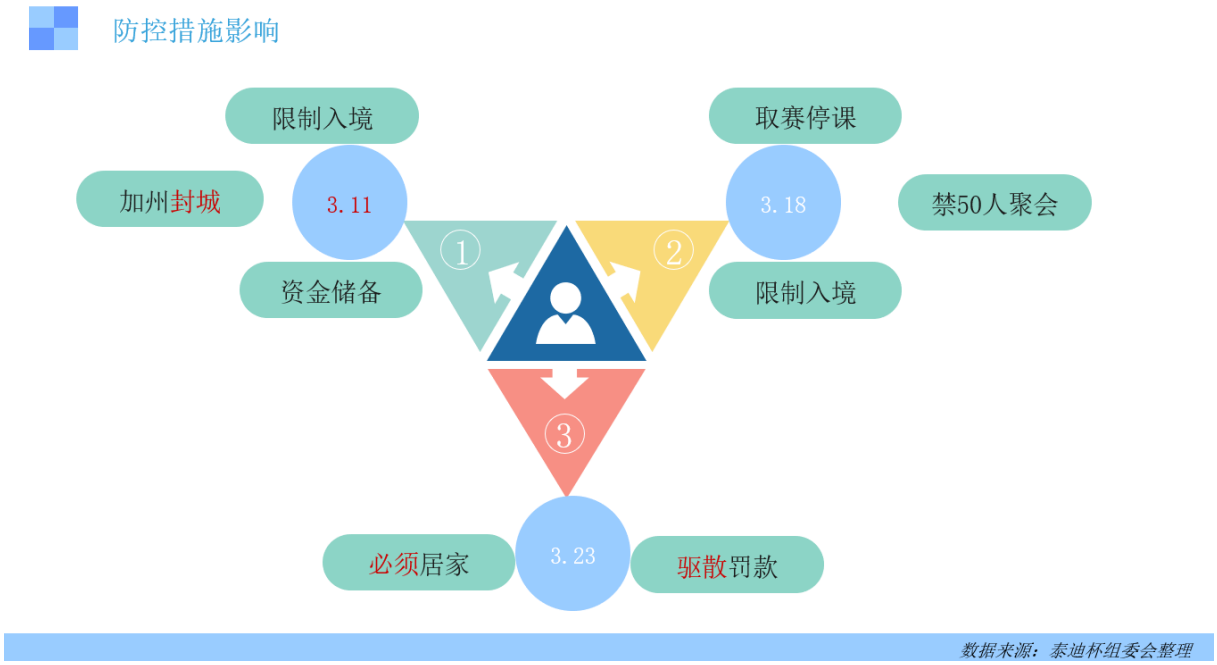



**图 13 防控策略**

**1）美国**

策略：从 13 日起禁止所有欧洲居民入境美国，启动紧急资金用于应对新冠疫情。19 日，加州封城。

分析：美国执行措施比其他两国都要早，对新冠疫情投入了大量的人力物力财力，管理执行较为严格（加州采取封城），但可能因为国民对此并不重视或者政府实际工作效率低下，所以直到 6 月 30 日，累计确诊人数还保持增长趋势。

**2）英国：**

策略：3 月 23 日起非必需禁止外出，禁止 50 人以上聚集。

分析：即使从 3 月 23 日开始禁止外出，但是非强制，聚集的处罚太低，没有起到督促作用。所以直到 6 月 30 日，英国新增确诊数并没有放缓的趋势。

**3）俄罗斯**

策略：3 月 18 日至 5 月 1 日，限制入境，取消所有体育赛事，莫斯科停课，禁止 50 人以上集会。

分析：俄罗斯从 18 日开始颁布限制措施，此时俄罗斯只有不到 200 人的累计病例，且限制措施较为严格，所以到 6 月 30 日时，新增病例大幅减少，疫情逐渐被控制。

**4）总体分析**

三个国家都有采取相应的措施，但是由于行动速度慢、行动效率低、国民配合度等多方面因素影响没有使得疫情防控策略切实作用于疫情防控。