2020 年"泰迪杯"数据分析职业技能大赛

B 题

新冠疫情数据分析

目录

一、 背景

1、行业背景

2020 年 1 月新型冠状病毒（以下简称新冠）肺炎在极短时间内就在全球范围内大规模流行，据美国约翰斯·霍普金斯大学 11 月 8 日发布的新冠疫情最新统计数据显示，截至美国东部时间 11 月 8 日 11 时 24 分全球累计确诊人数超过 5000 万，死亡人数超过 125 万。由于新冠病毒的传播速度快、致死率较高，世界卫生组织称新冠是百年一遇的人类公敌。自新冠肺炎爆发以来，面对社会对疫情信息的迫切需求，各级政府部门通过多种渠道及时发布第一手相关数据，许多组织和个人也迅速行动，利用多种分析手段为公众提供疫情数据的解读分析，以消除公众的恐慌情绪，提高人们自我防护意识，配合政府防疫措施，为我国最终打赢疫情防控阻击战发挥了巨大的推动作用。

2、分析目标

（1）对疫情数据进行简单的统计。

（2）设计可视化数字大屏展示新冠疫情的时空变化情况。

（3）使用可视化工具绘制城市疫情风险图。

（4）对国内和国际的疫情变化情况进行分析。

二、新冠疫情数据分析

1、数据的基本处理

1.1 任务

根据附件 1“城市疫情”中的数据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，将结果保存为“task1_1.csv”，并列表给出武汉、深圳、保定每月 10、25 日的统计结果。

解题思路：

对各地区进行分组匹配出相对应的省份，按日期累加每日的新增确诊的人数、新增治愈人数、新增死亡人数作为累计确诊人数、累计治愈人数和累计死亡人数，然后保存为 CSV 文件，取出三个城市每月 10、25 日的数据进行统计制作成列表。

解题过程：

用 to_datetime 进行转换表中的日期为标准化日期，使用 groupby 对城市进行分类，通过 cumsum 函数实现累加，然后对数据进行保存，建立透视表；取出武汉、深圳、保定信息后定义一个函数返回 10、25 号的累计数据。

结果展示：

	城市	日期	累计确诊人数	累计治愈人数	累计死亡人数	月	日
0	武汉	2020-01-10	41	2	1	1	10
4045	武汉	2020-02-10	18454	1173	748	2	10
8532	武汉	2020-03-10	49978	33264	2423	3	10
9407	武汉	2020-04-10	50008	46154	2577	4	10
9851	武汉	2020-05-10	50339	46464	3869	5	10,
	城市	日期	累计确诊人数	累计治愈人数	累计死亡人数	月	日
492	武汉	2020-01-25	618	40	45	1	25
6931	武汉	2020-02-25	47441	12026	2085	2	25
9088	武汉	2020-03-25	50006	44020	2531	3	25
9662	武汉	2020-04-25	50333	46452	3869	4	25)

图 1-1-1　武汉每月 10 号 25 号的统计结果

	城市	日期	累计确诊人数	累计治愈人数	累计死亡人数	月	日
4013	深圳	2020-02-10	375	56	0	2	10
8520	深圳	2020-03-10	417	387	3	3	10,
	城市	日期	累计确诊人数	累计治愈人数	累计死亡人数	月	日
469	深圳	2020-01-25	27	2	0	1	25
6909	深圳	2020-02-25	417	262	3	2	25)

图 1-1-2　深圳每月 10 号 25 号的统计结果

	城市	日期	累计确诊人数	累计治愈人数	累计死亡人数	月	日
3886	保定	2020-02-10	30	9	0	2	10,
	城市	日期	累计确诊人数	累计治愈人数	累计死亡人数	月	日
370	保定	2020-01-25	3	0	0	1	25
10202	保定	2020-06-25	45	33	0	6	25)

图 1-1-3 保定每月 10 号 25 号的统计结果

任务 1.2

根据任务 1.1 的结果，并结合附件 1“城市省份对照表”统计各省级行政单位按日新增和累计数据，将结果保存为“task1_2.csv”。在论文中给出实现方法的相关描述，并列表给出湖北、广东、河北每月 15 日的统计结果。

解题思路：

通过任务 1.1 的 task1_1 表与附件 1 中的城市省份对照表进行联合，合成省份疫情表。

解题过程：

先对 task1_1 表与附件 1 中城市省份对照表进行联合，得出省份疫情表，后读取省份疫情表，先对省份疫情表中的日期数据进行标准化处理，然后提取出日期中的‘月’与‘日’，插入表中，方便提取；

然后提取出省份为‘湖北’，‘广东’，‘河北’中日为 15 的数据表；

最后进行以省份与日期为键的聚类分组处理。

结果展示：

	省份	日期	新增确诊	新增治愈	新增死亡	累计确诊人数	累计治愈人数	累计死亡人数	月	日
0	湖北	2020-01-10	41	2	1	41	2	1	1	10
1	湖北	2020-01-11	0	4	0	41	6	1	1	11
2	湖北	2020-01-12	0	1	0	41	7	1	1	12
3	湖北	2020-01-15	0	5	1	41	12	2	1	15
4	湖北	2020-01-16	4	3	0	45	15	2	1	16
...	...	...	...	...	...	...	...	...	...	...
10240	湖南	2020-04-17	0	1	0	1	1	0	4	17
10241	安徽	2020-04-08	1	0	0	1	0	0	4	8
10242	安徽	2020-04-24	0	1	0	1	1	0	4	24
10243	辽宁	2020-04-16	1	0	0	1	0	0	4	16
10244	辽宁	2020-05-01	0	1	0	1	1	0	5	1

10245 rows × 10 columns

图 1-2-1 处理后的省份疫情表

		累计确诊人数	累计治愈人数	累计死亡人数
省份	日期			
广东	2020-02-15	1253	408	2
	2020-03-15	428	401	3
	2020-04-15	458	410	1
	2020-06-15	233	88	0
河北	2020-02-15	254	85	3
	2020-04-15	10	4	0
	2020-06-15	36	32	0
湖北	2020-01-15	41	12	2
	2020-02-15	56187	5852	1596
	2020-03-15	66455	53940	3070
	2020-04-15	50008	46325	2579

图 1-2-2 湖北、广东、河北的统计结果

任务 1.3

根据任务 1.2 的结果，统计各省级行政单位每天新冠病人的住院人数，将结果保存为“task1_3.csv”。在论文中给出实现方法的相关描述，并列表给出湖北、广东、上海每月 20 日的统计结果。

解题思路：

先读取任务 1.2 所得的表，然后通过计算公式：每天新管病人住院人数=累计确诊人数-累计治愈人数-累计死亡人数，得出每天新冠病人住院人数。

解题过程：

（1）读取 task1_2 表。

（2）提取出 task1_2 表中的累计确诊人数，累计治愈人数，累计死亡人数，然后使用公式计算出住院人数，再插入到表中。

（3）然后提取出省份为‘湖北’，‘广东’，‘上海’中日为 20 的数据表。

（4）进行以省份与日期为键的聚类分组处理。

结果展示：

	省份	日期	新增确诊	新增治愈	新增死亡	累计确诊人数	累计治愈人数	累计死亡人数	住院人数
0	湖北	2020-01-10	41	2	1	41	2	1	38
1	湖北	2020-01-11	0	4	0	41	6	1	34
2	湖北	2020-01-12	0	1	0	41	7	1	33
3	湖北	2020-01-15	0	5	1	41	12	2	27
4	湖北	2020-01-16	4	3	0	45	15	2	28
...	...	...	...	...	...	...	...	...	...
10240	湖南	2020-04-17	0	1	0	1	1	0	0
10241	安徽	2020-04-08	1	0	0	1	0	0	1
10242	安徽	2020-04-24	0	1	0	1	1	0	0
10243	辽宁	2020-04-16	1	0	0	1	0	0	1
10244	辽宁	2020-05-01	0	1	0	1	1	0	0

图 1-3-1 整合后的数据表

3]:

省份	日期	住院人数
上海	2020-01-20	1
	2020-02-20	105
	2020-03-20	41
	2020-04-20	100
	2020-05-20	20
广东	2020-01-20	14
	2020-02-20	557
	2020-03-20	52
	2020-04-20	8
	2020-05-20	110
湖北	2020-01-20	239
	2020-02-20	48725
	2020-03-20	5595
	2020-04-20	102

图 1-3-2 湖北、广东、上海每月 20 日的统计结果

任务 1.4

假设新冠病人的传播半径为 1 km，根据附件 1“A 市涉疫场所”在平面图中分别绘制该市第 6 天和第 10 天的疫情传播风险区域。

解题思路：

先算出疫情传播的面积，再算新冠病人的传播范围。

解题过程：

使用 Bi 平台，根据表 A 市涉疫场所中的数据，分别描绘出第 6 天和第 10 天的传播横坐标和纵坐标，算出所对应的疫情场所的传播面积。

结果展示：

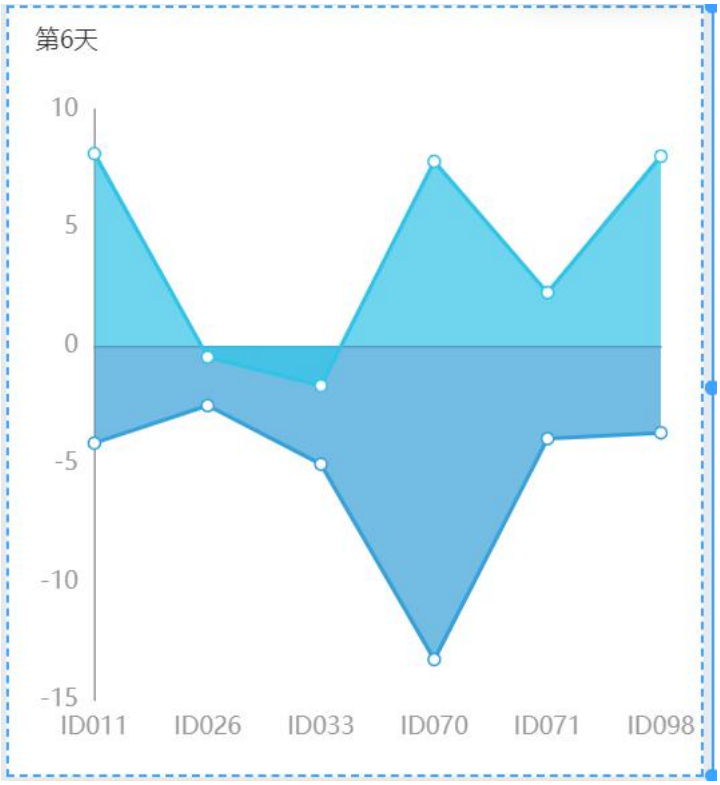


图 1-4-1 第 6 天的疫情传播风险区域

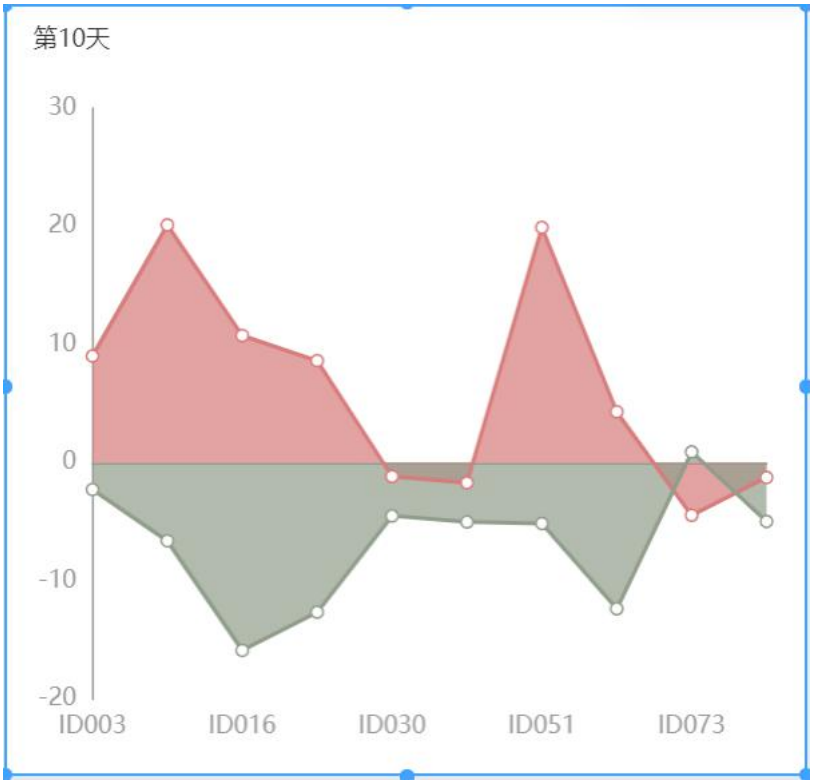


图 1-4-2 第 10 天的疫情传播风险区域

结果分析：

新冠病人的传播的半径不超出这个范围疫情则不会继续扩大。

2、数字大屏设计

任务 2.1

设计数字大屏,展示国内新冠疫情汇总概要信息、时空变化情况、重点关注区域等。

解题过程：

国内新冠疫情汇总概要可以统计国内各城市的新增总人数、确诊总人数、死亡总人数用于中共地图中的城市数据分布，时空变化情况通过雷达图进行同一地点不同时间的数据对比，重点关注区域通过词云图对各城市的数据统计展现出受疫情明显的城市。

结果展示：

图 2-1-1 国内疫情汇总

结果分析：

从仪表盘中可以看到疫情影响的地区多半集聚于中国东部沿海地区较为发达的地区。从词云图中可以看出应该重点关注的地区，由于新增人数、死亡人数的增多和治愈人数的降低，我们可以把这个区域列为严重疫情灾区，从而可以对该区域进行防疫物资及防疫宣传工作的展开。

任务 2.2

设计数字大屏，展现并分析国际疫情态势和发展变化。

解题过程：

用一个雷达图统计各国数据，制作一个世界地图统计各国的疫情情况是否严重的分布，用折线图体现出世界各国随着时间的变化疫情累计的确诊人数、累计的治愈人数和累计的死亡人数，用于分析疫情的走势。

结果展示：

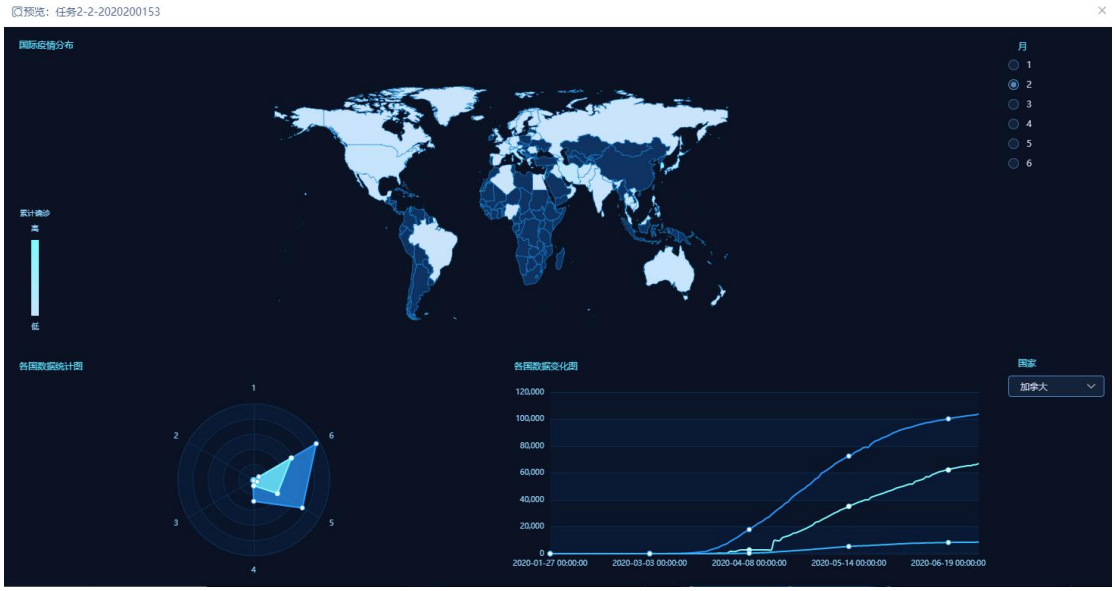


图 2-1-2 国际疫情情况

结果分析：

(1)1 月-3 月，缓慢增长期：该阶段的主要特征是累计死亡病例数不超过 100 例，而且全球累计死亡病例以亚洲为主导，亚洲累计死亡病例占全球总数的 93%，欧洲约占 7%;

(2)3 月-5 月，快速增长期：该阶段的主要特征是单日新增死亡病例数大于 200 例但不超过 2000 例，全球累计死亡病例分布在除南极洲之外的其他六大洲，且以欧洲为主导，欧洲的累计死亡病例占全球总数的 63%，亚洲占 34%，美洲(主要是北美洲)占 3%;

(3)5 月-6 月，持续爆发期：该阶段的主要特征是单日新增死亡病例数超过 2200 例，目前单日新增死亡最高达 6662 例;全球累计死亡病例分布仍以欧洲为主导，欧洲的累计死亡病例占全球总数 71.6%，且美洲(主要是北美洲)的累计死亡病例数呈持续快速增长态势。

（4）折线图可以看出随着时间的推移，在 3 月至 5 月上升坡度较大，累计确诊人数数量急剧上升，累计治愈也越来越多，累计死亡人数在疫情初期也是潜伏期没有对生命造成很大的伤害，在 5 月后，人们对疫情的认识和防控也有了进一步的了解，所以疫情的发展趋势慢慢得到缓解。

3、国际疫情的发展分析

任务 3.1

根据附件 1“国际疫情”中的数据，对印度、伊朗、意大利、加拿大、秘鲁、南非在各个时间段中所处的疫情发展阶段进行划分，并在论文中给出划分的依据和结果。

解题思路：

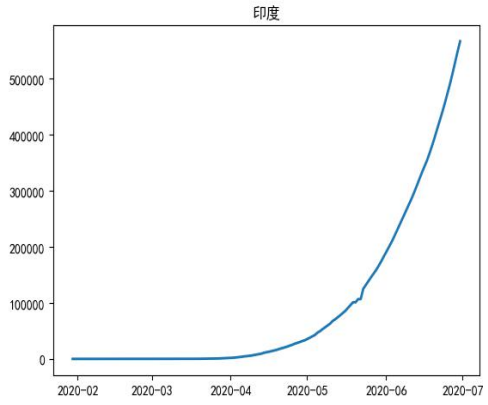
首先读取附件 1 中的国际疫情表，然后对国际疫情表中的国家进行提取，提取后建立‘日期与累计确诊折线图’与‘确诊月环比率折线图’通过两图对国家的疫情进行时间段划分。

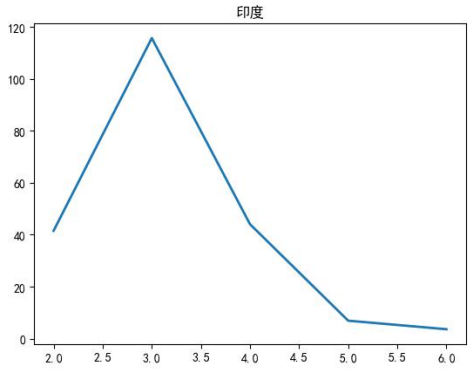
解题分析：

读取表后，开始对各国家进行提取，提取后进行绘制折线图，与月环比率折线图。通过月环比率折线图得出分区数。

结果展示与分析：

(1) 印度时间段疫情发展阶段划分：





目录

图 3-1-1 印度疫情增长折线图　　**图 3-1-2 印度疫情月环比率折线图**

通过月环比率折线图可得知，在印度 1，2 月份属于快速增长趋势，因此为疫情增长区，3、4、5、6 月份月环比率开始下降，因此 3、4、5、6 月份属于疫情减缓区，但月环比率还是处于正增长趋势，因此疫情还未得以全部控制。

(2) 意大利时间段疫情发展阶段划分：

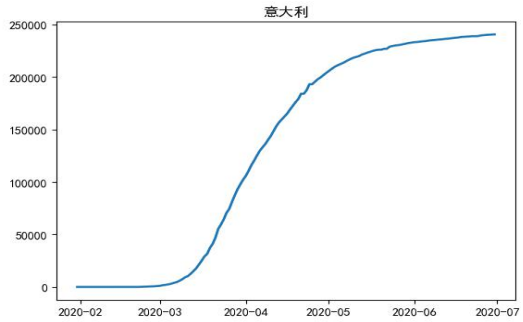


图 3-1-3 意大利疫情增长折线图

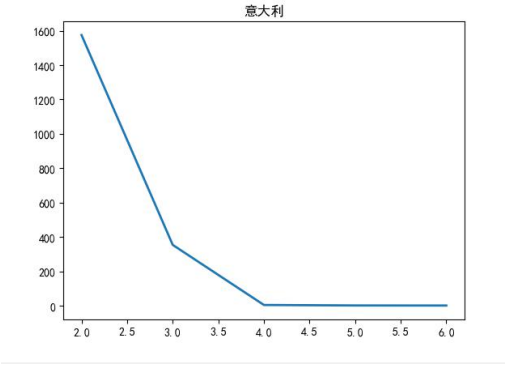


图 3-1-4 意大利疫情月环比率折线图

通过月环比率折线图可得知，意大利从疫情开始时，环比增长是非常大的，2 月，3 月属于潜伏增长区，新冠疫情人数基数开始增长，3 月份环比率比 2 月份环比率低。从 4 月份开始月环比率趋近于 0，说明疫情以被控制或者已稳定。

(3)加拿大时间段疫情发展阶段划分：

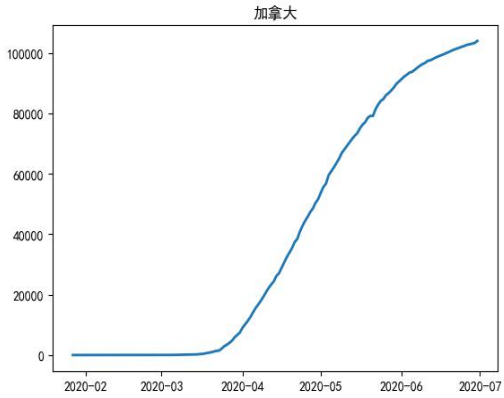


图 3-1-4 加拿大疫情增长折线图

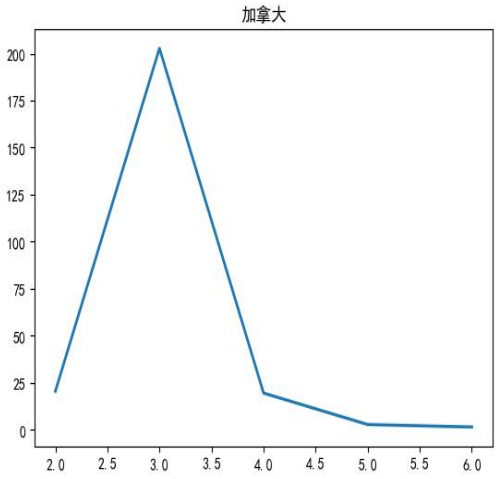


图 3-1-5 加拿大疫情月环比率折线图

通过加拿大疫情月环比例率，可得知，加拿大 1，2 月份属于疫情月环比率增加，并且增加幅度大，因此属于潜伏发展区，3 月份，月环比率降低，但比例数大，因此属于发展疫情区，4，5 月份，月环比率降低到 10 点以下，以此属于疫情稳定区。

(4)伊朗时间段疫情发展阶段划分：

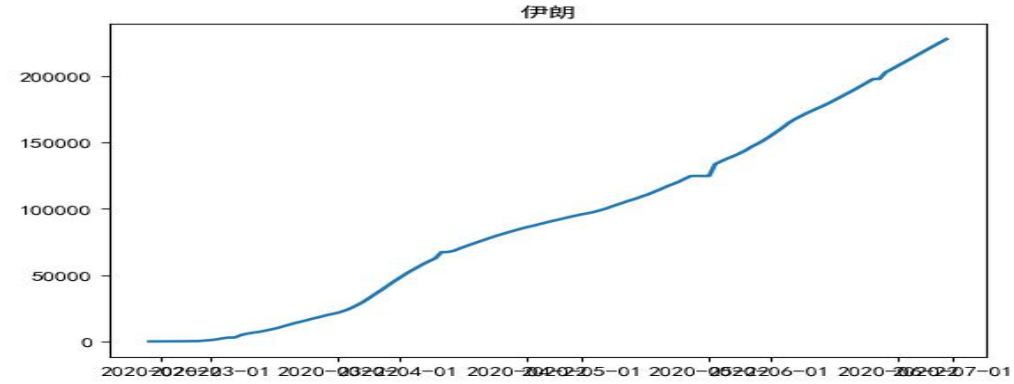


图 3-1-6 伊朗疫情增长折线图

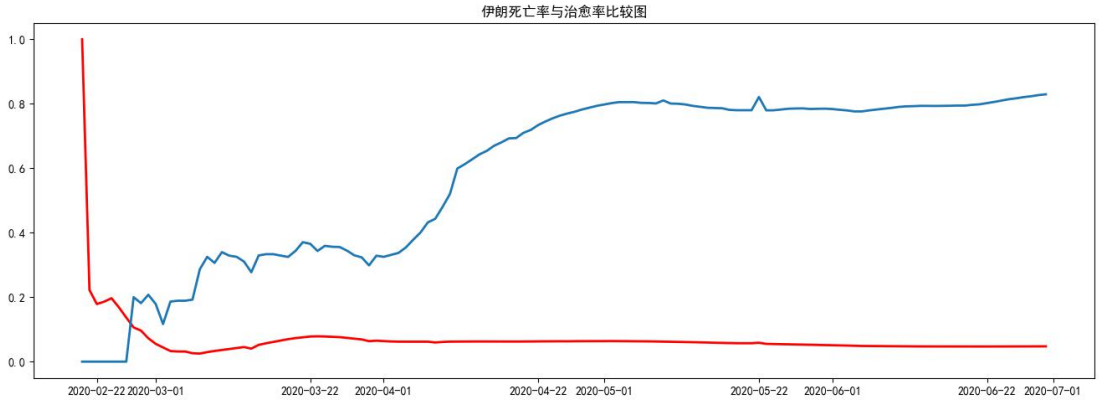


图 3-1-7 伊朗疫情死亡率与治愈率比较图（红线为死亡率，蓝线为治愈率）

通过伊朗疫情死亡率与治愈率比较图，可得知，2 月份，死亡率比治愈率要高，并且疫情还在增长，因此属于疫情增长区；3 月份，死亡率与治愈率较为接近，因此为疫情缓和区，从 4 月份开始，治愈率在上升而死亡率正在下降，因此为疫情平稳区。

结果总结：

观察各国家的疫情发展情况，可普遍划分为 1，2 月疫情潜伏增长区，1，2 月份增长率正在不断升高；3，4 月份为疫情平缓区，疫情增长率逐渐降低；5，6 月份为疫情缓和区，5，6 月份大部分国家疫情增长率为 0，以属于稳定疫情。

任务 3.2

根据附件 2 中的信息，分析美国、英国、俄罗斯 3 个国家推出的疫情防控措施对本国疫情变化情况的影响。

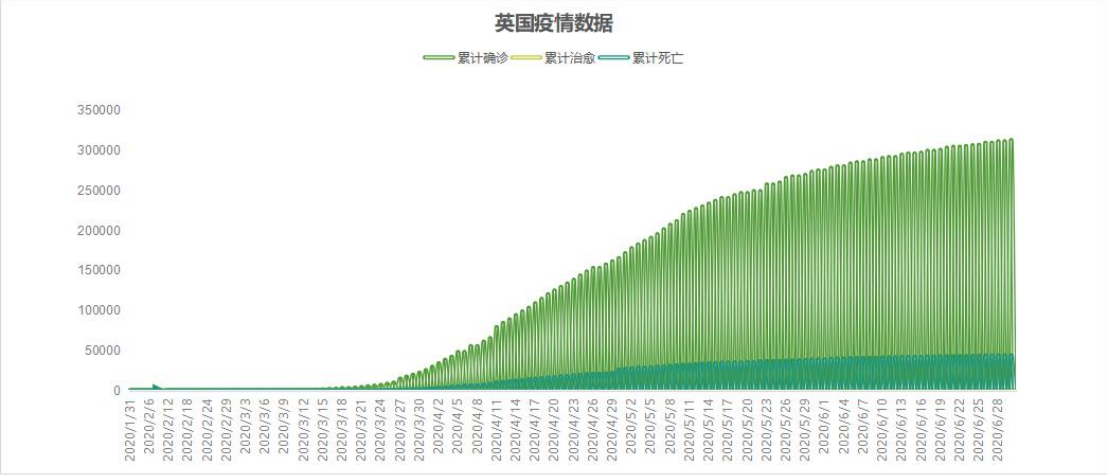


图 3-2-1 英国疫情数据

结果分析：

英国虽然采取了强硬的措施，可是没有将疫情的严重性让国民清楚的了解，也没有宣传防控的注意事项。为何现在疫情还迟迟得不到很好的缓解原因有以下几点

（1）政府前期疫情过于佛系，人口感染基数不断扩大。尽管如今，英国政府深感到病毒的危害，英国在各方面都加强了措施，甚至对部分地区实行封禁，但为时已晚。

（2）政府呼吁“群体免疫”的政策。他们想要所有人不感染不可能，当大部分人感染后可获得“群体免疫力“。

（3）国民的不重视，信任政府的，对日常的防控没有意识没有加强，不注重日常的卫生和室内通风，不戴口罩，人口密集。

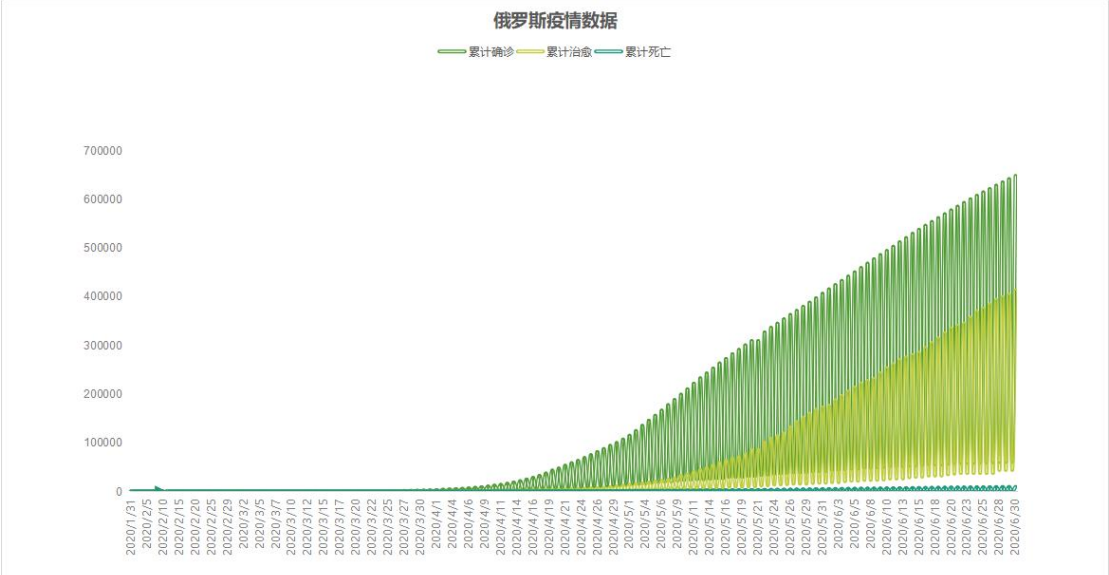


图 3-2-2 俄罗斯疫情数据

结果分析：

虽然，俄罗斯 3 月 18 号-5 月 1 号，进行封国，推出了强有力的政策，在国内不断普及新冠病毒的危害，而且禁止外籍人士进入俄罗斯，一开始推出政策是有效控制了疫情的增长。但是疫情有一定的潜伏期，大量的俄罗斯人回国，而且俄罗斯下层政府不够重视，没有落实好政策，民众仍然不够重视，导致确诊人数不断攀升，幸好国家医疗工作完善还有邻国的帮助，致使疫情相对控制。

（1）对于俄罗斯的下层政府很多官员来说，他们并不是十分重视这次的新冠肺炎疫情，因此在宣传病毒危害时，不戴口罩，甚至自己在出行时也不戴口罩。

（2）大量境外欧洲俄罗斯人回国，境外输入疫情防控不力，导致新冠肺炎疫情。

（3）检测手段和试剂盒的不断增加，检测人群的扩大，确诊患者的数量不断增多。

图 3-2-3 美国疫情数据

结果分析：

虽然美国，疫情一开始的确得到了很好的抑制，可是之后的一系列暴乱和政策引起了疫情的加重，而且有部分美国人对特朗普的群体感染政策抱有希望，这些举动对美国的疫情防控工作严重的冲击。

（1）美国大家都知道了，特朗普政府不作为，不抗击疫情，还甩锅中国，抹黑中国。

（2）美国民众也轻视，甚至不戴口罩聚会依旧。加上昨天特朗普甚至决定要取消美国的疫情应对小组，迫于舆论不得不改口，但可以看出未来美国疫情应对小组必将不会受到特朗普的多大重视。

（3）美国已经在逐步进行复工复产，一些封禁的州也在不断解封。

（4）黑人事件引发的骚乱。