**新冠肺炎的可视化和预测分析**

管飞宇 U201715262

1. 项目背景

自从今年1月以来，新冠肺炎肆虐，各国感染人数急速攀升，目前总人数已超过400万，即将达到500万乃至更多。本项目基于python语言及机器学习技术，尝试对已有数据进行可视化分析并给出对于未来感染人数的预测。

1. 项目数据

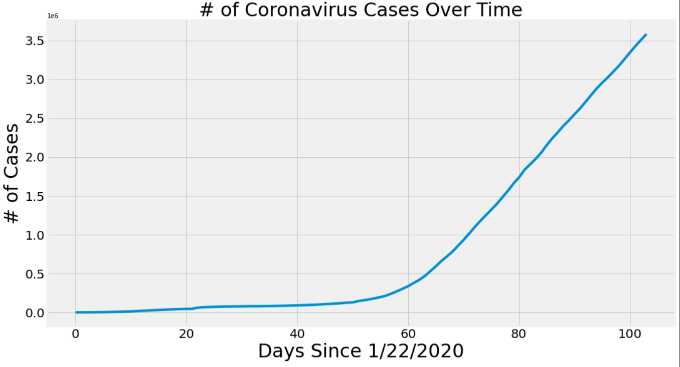
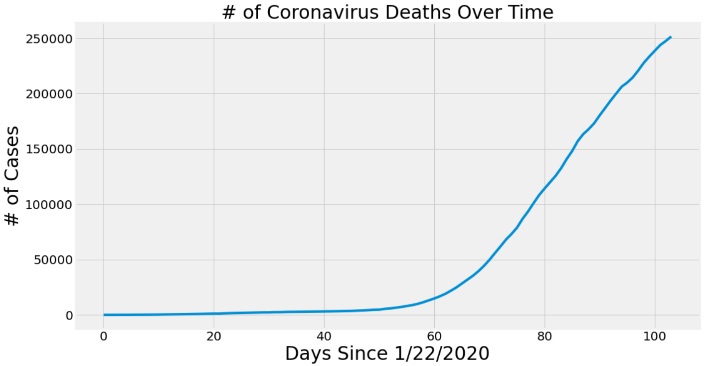
使用kaggle公开数据集<https://www.kaggle.com/corochann/covid-19-eda-with-recent-update-on-april/data?scriptVersionId=32149572>（不翻墙会很慢），日期截止至4月14日，感染人数预测部分使用数据截止至4月25日。

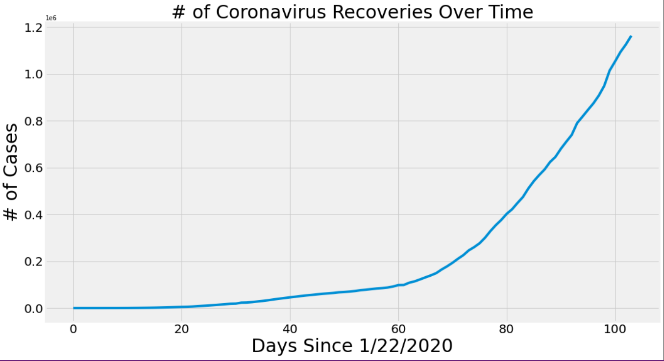
1. 项目使用技术分析

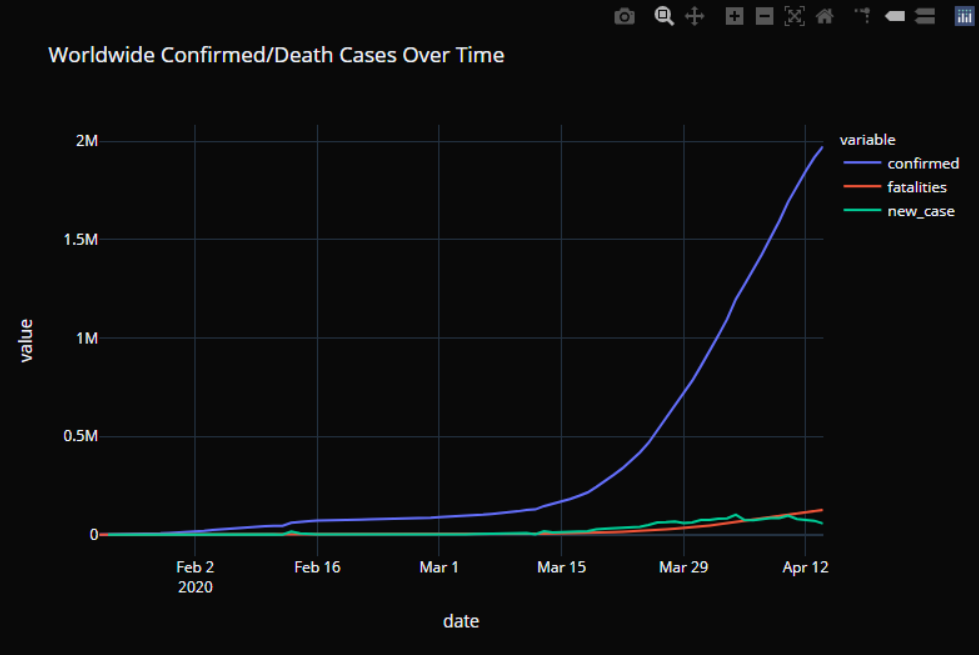
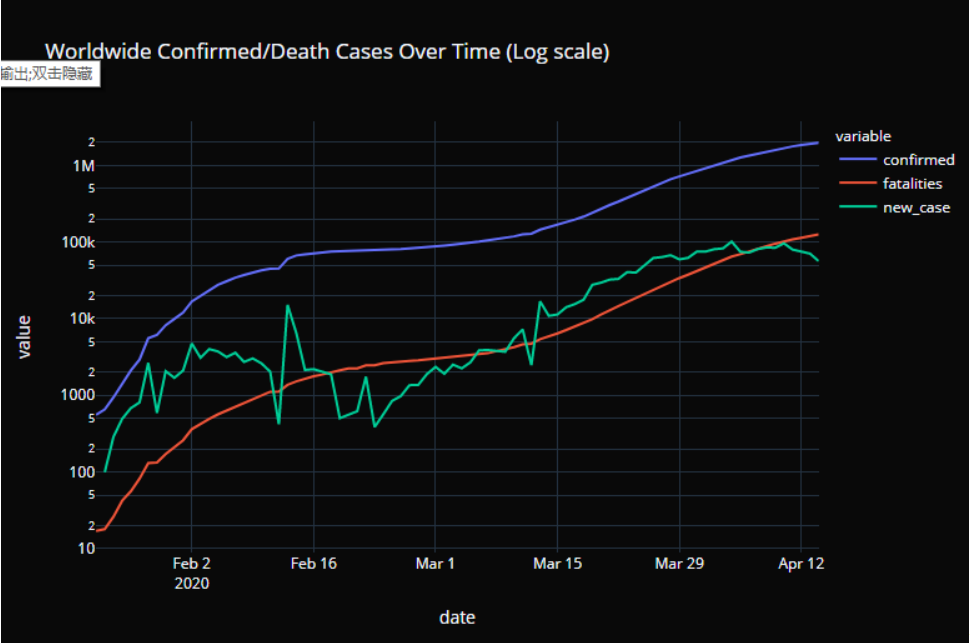
对于第一部分新冠感染人数的可视化，主要采用python的pandas包读取数据，采用numpy和scipy包进行一些数据处理，最后使用plotly包实现作图输出，达到可视化的目的。

对于第二部分，未来感染人数的预测，主要采取两种方法进行拟合，第一种是使用传统方法，基于sigmod函数的s型曲线拟合，第二种是使用机器学习技术，采用Facebook开源的prophet框架进行预测。

1. 结果分析
2. **数据可视化分析**
3. **全球疫情总体情况**

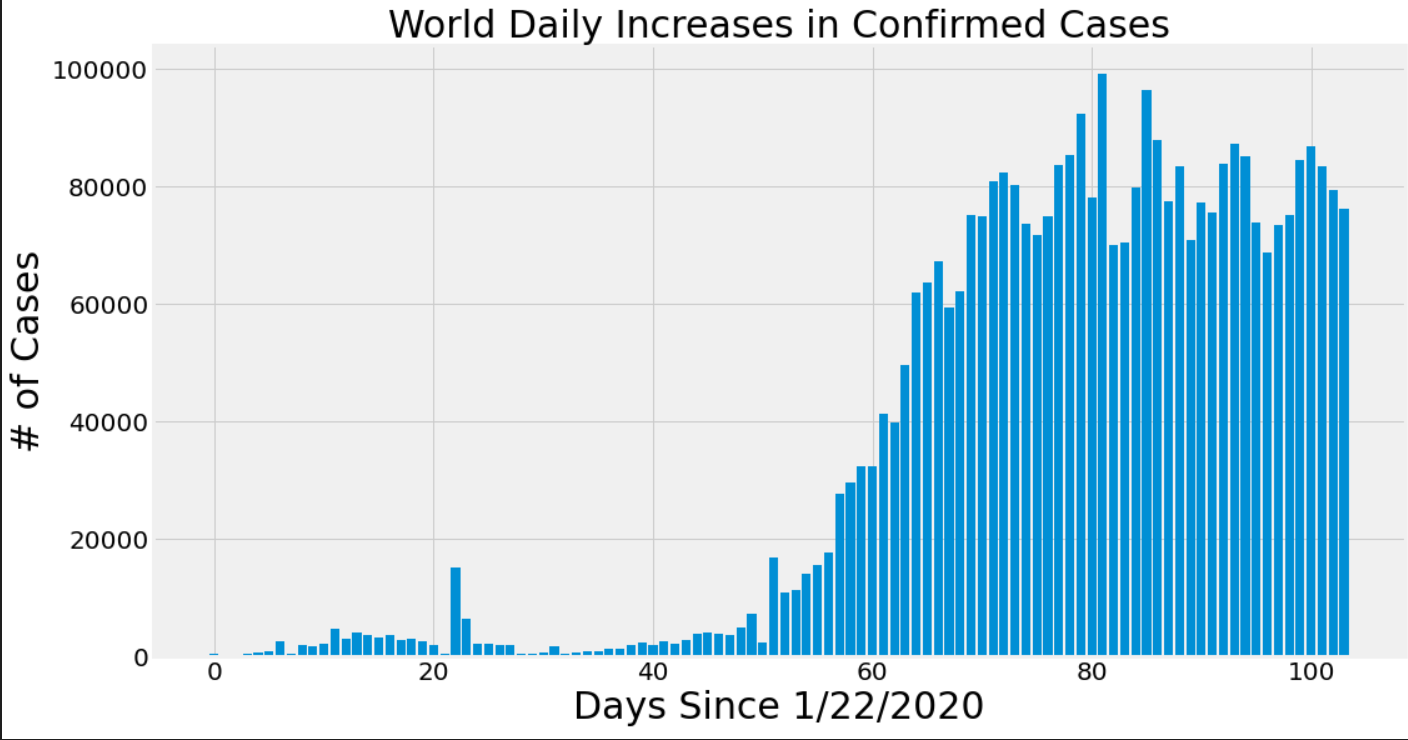
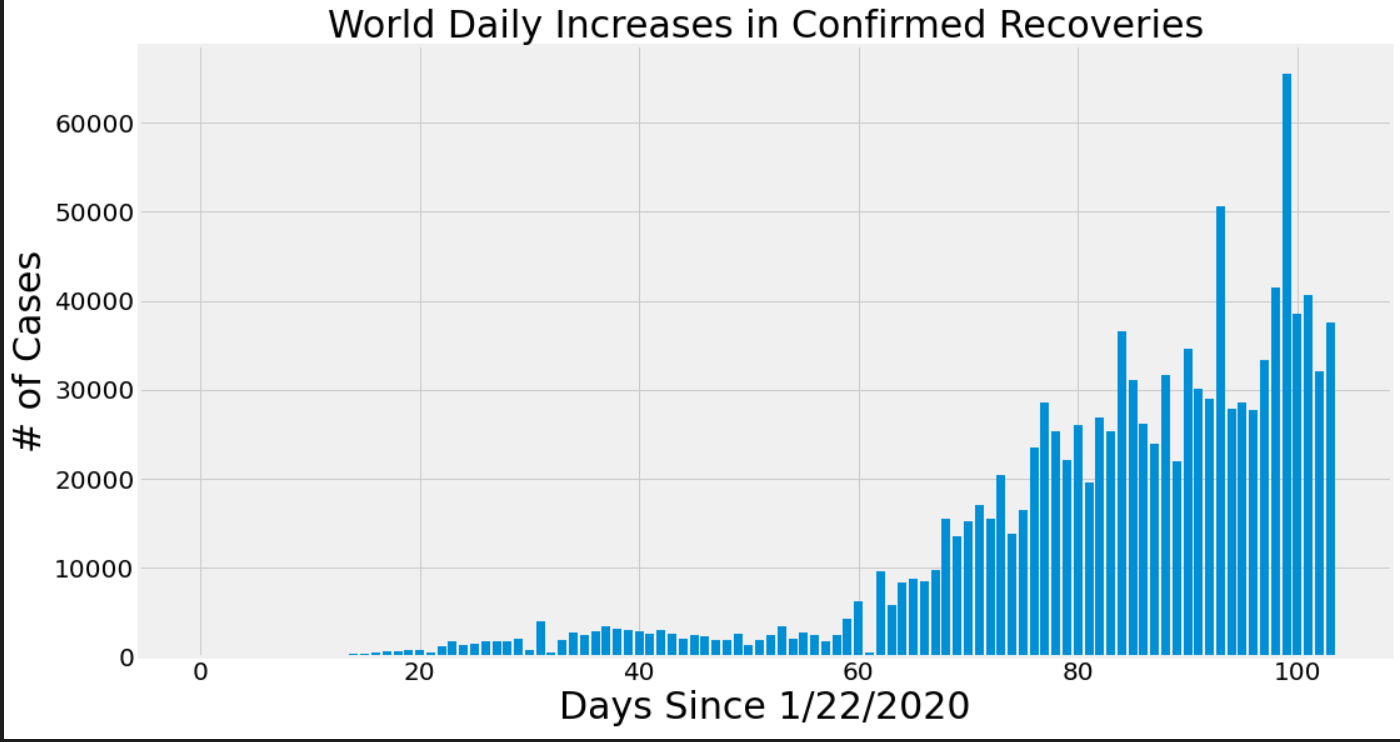
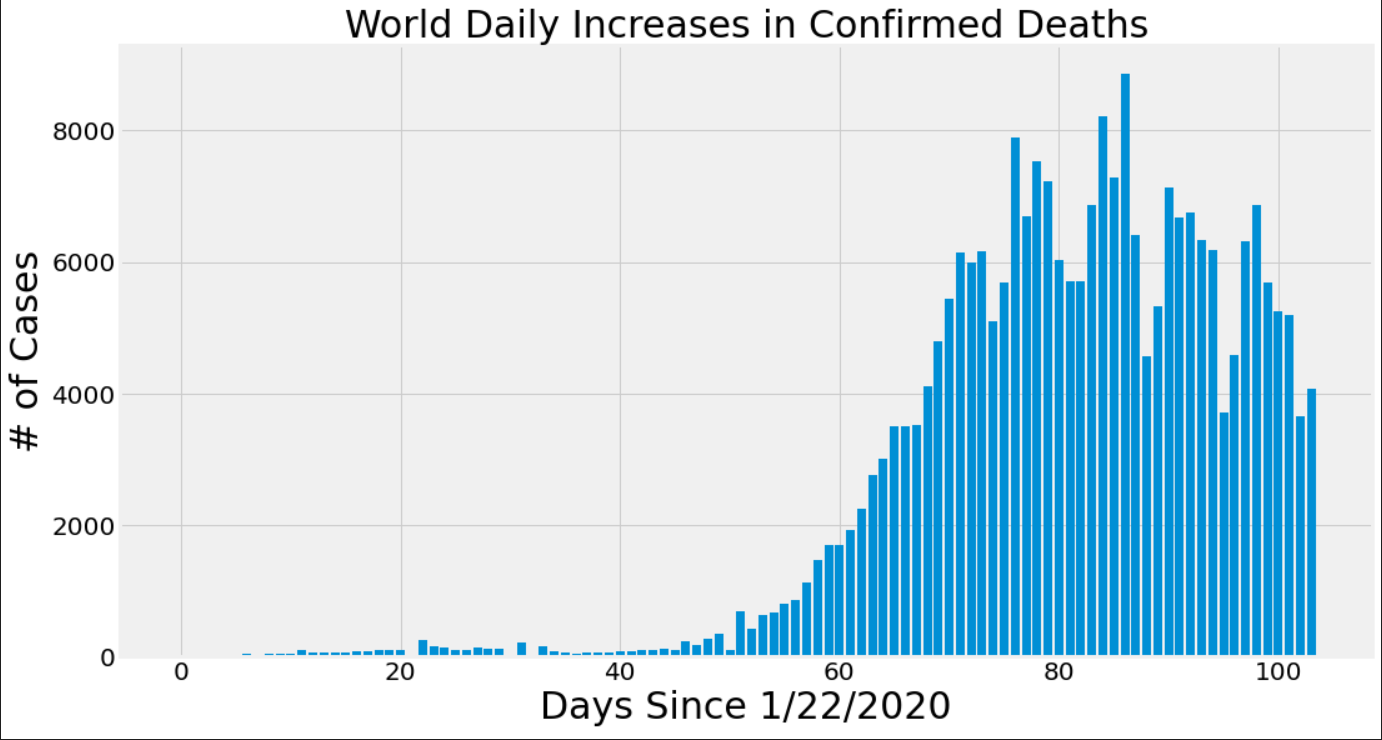




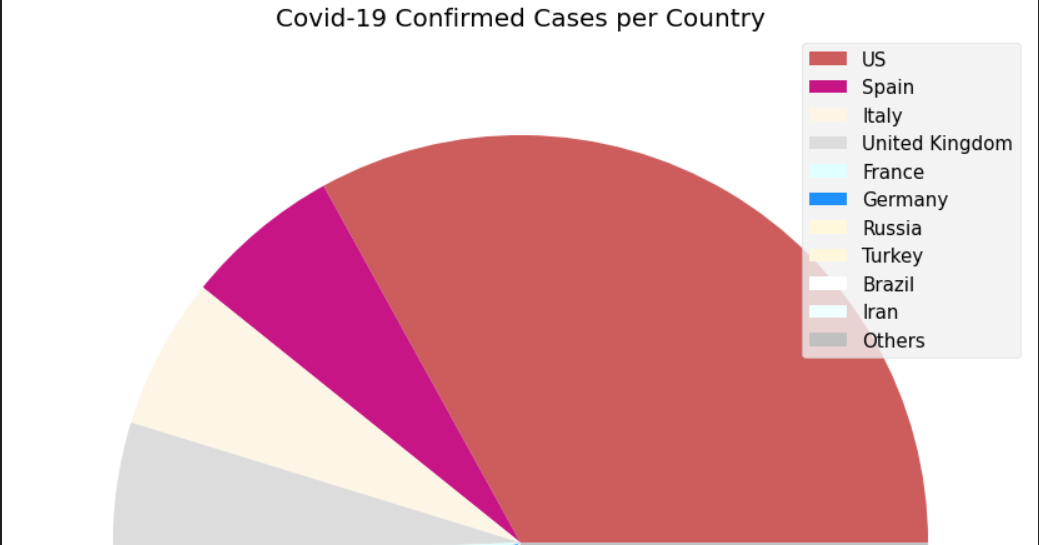
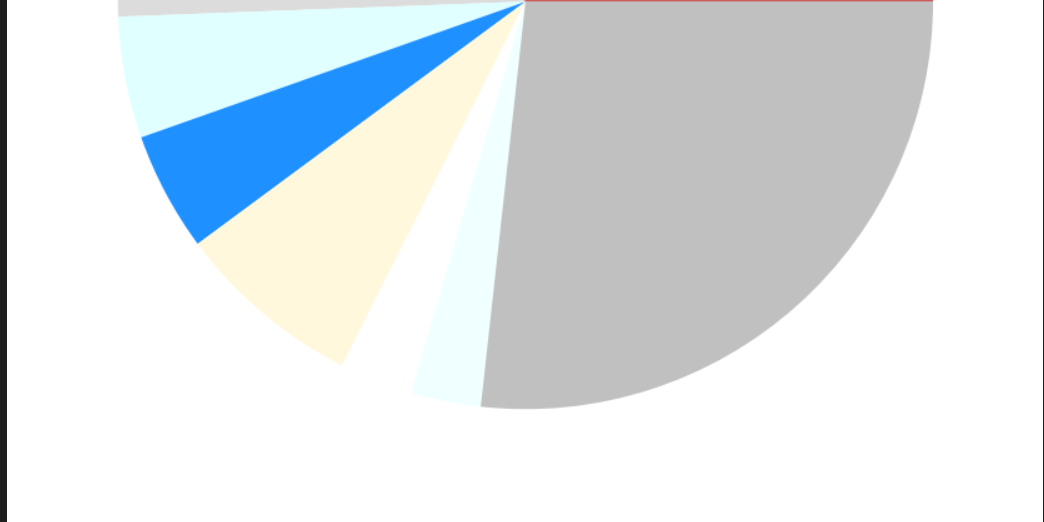
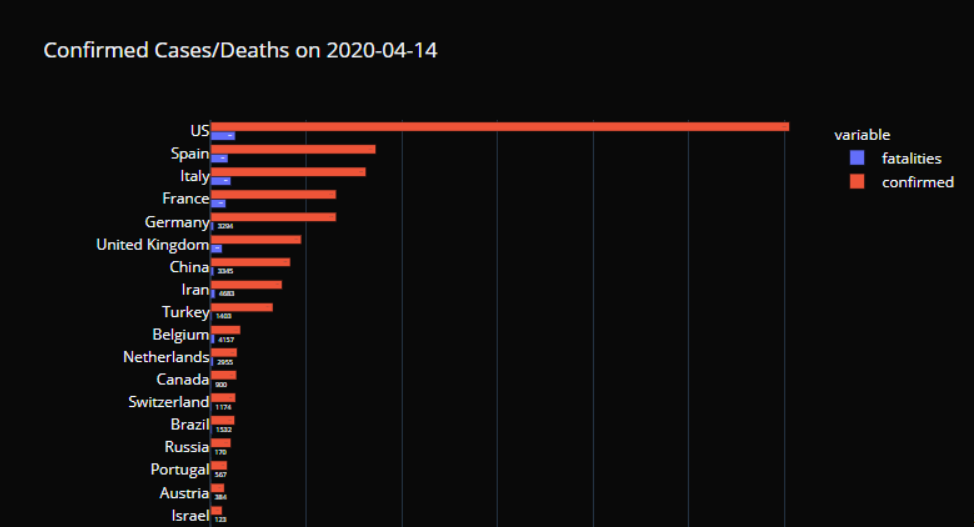
采用另一种方式呈现

世界范围内确诊与死亡数对比 确诊与死亡数对比（对数坐标）

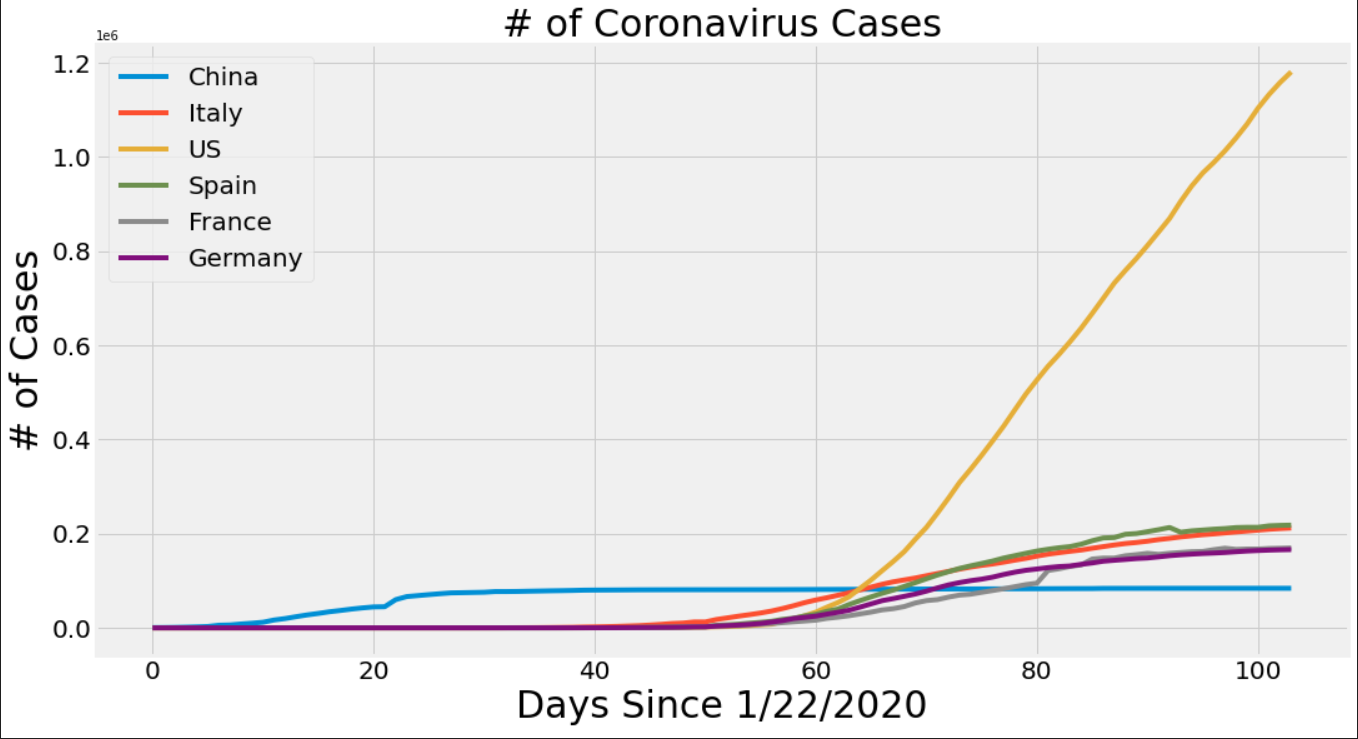
可以看到，在世界范围内，新冠肺炎的确诊人数和死亡人数仍在高速增长，虽然痊愈人数也在增加，但是远赶不上确诊病例增加的速度，说明接下来一段时间内疫情蔓延的势头还不会停止。另一组图能更加直观的反映这一点：

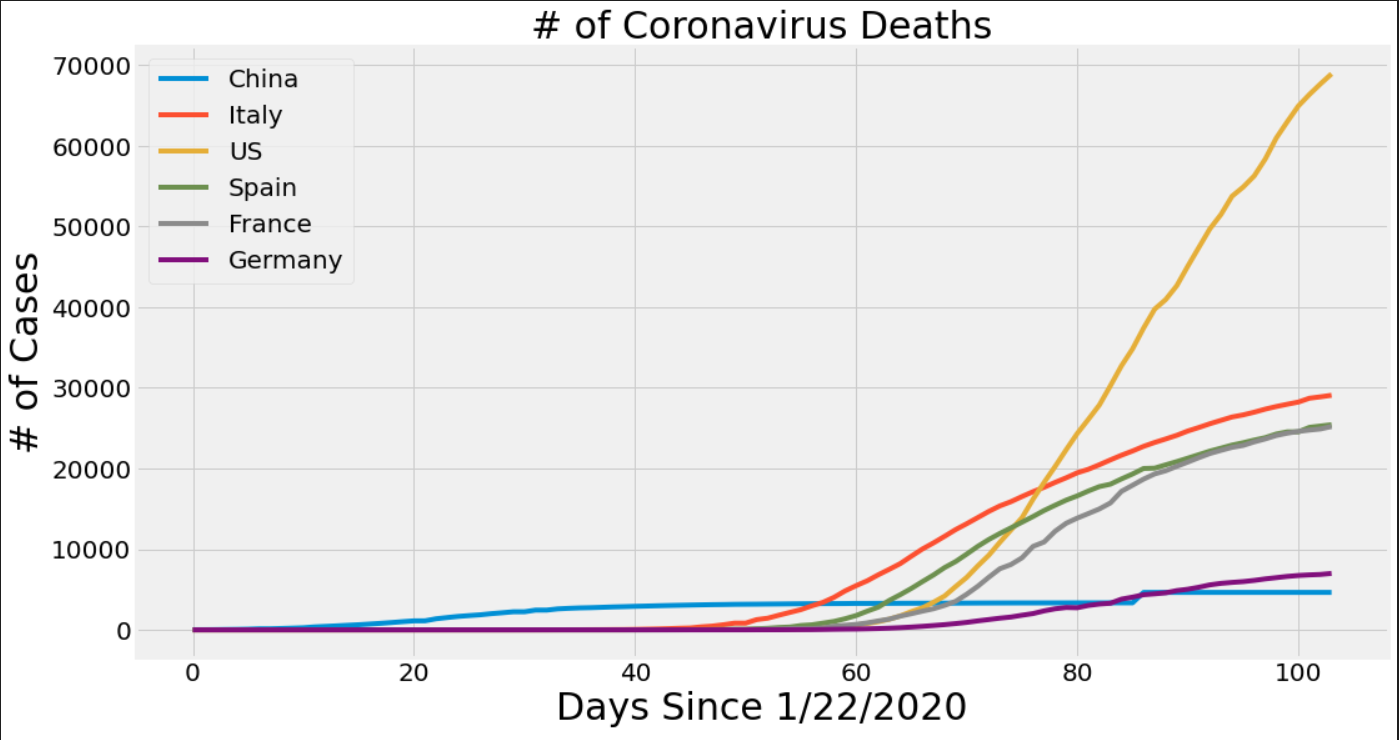
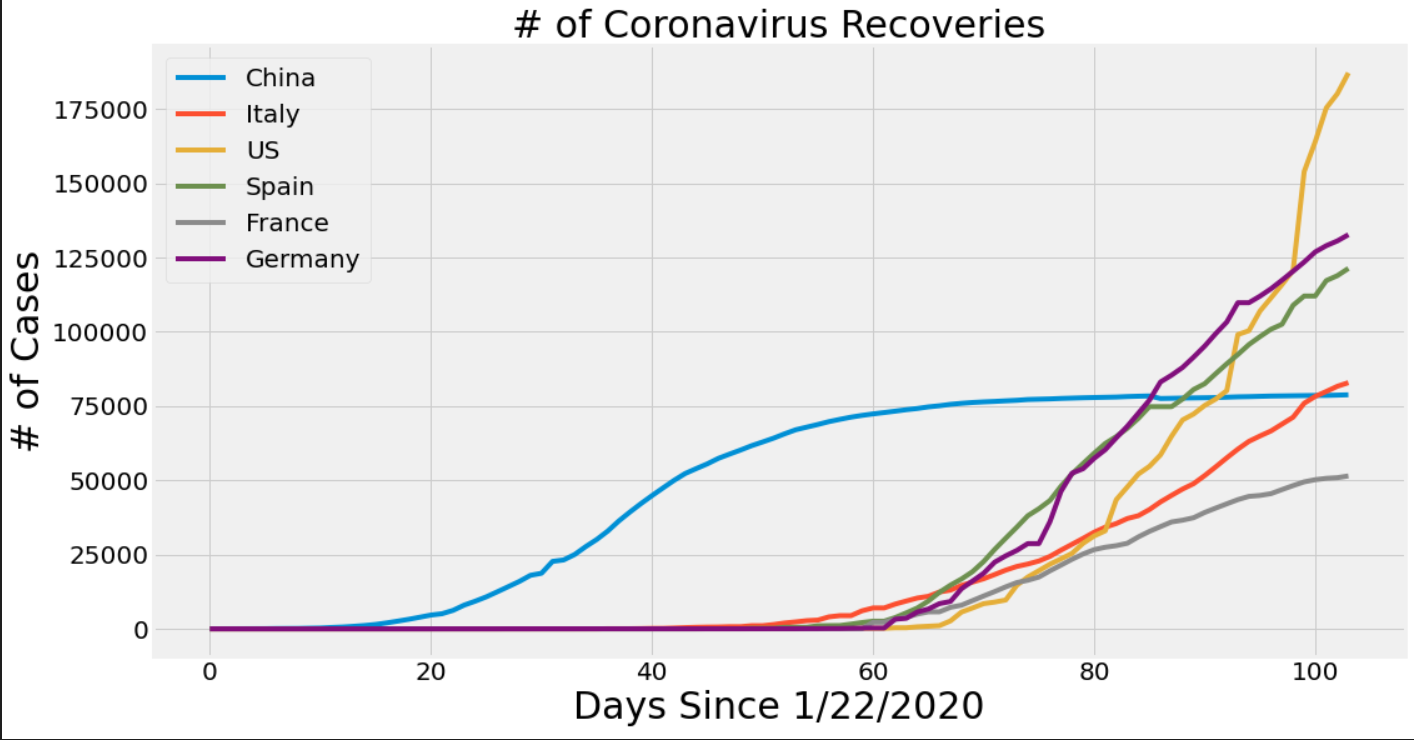


可以看到，每日新增确诊一直稳定在7万附近，而痊愈人数却基本停留在3-4万人左右，同时还伴随着每天近4000人的死亡，疫情形势不容乐观

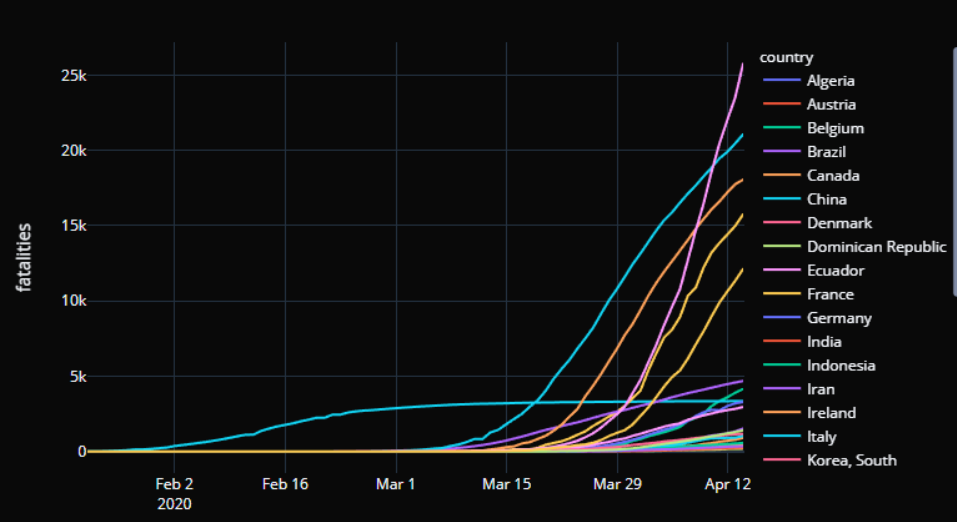
1. **各国总体情况**

对比世界各个国家的确诊病例和死亡病例，可以看到，美国的确诊数遥遥领先于其他国家，同时欧洲最早爆发疫情的西班牙和意大利也是重灾区，此外还有法国，德国，英国也都有一定数量的确诊病例。

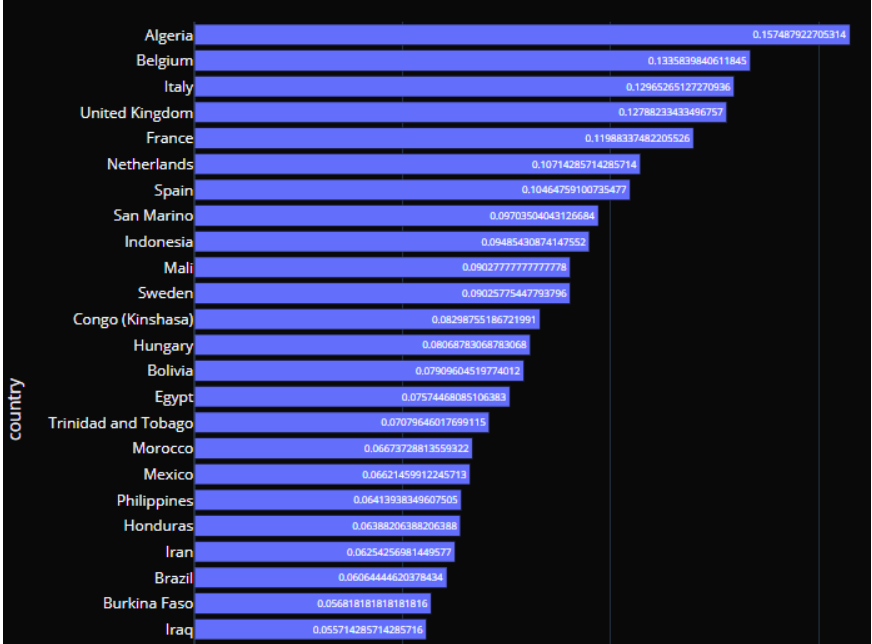
选取排名靠前的五个国家，美国、西班牙、意大利、法国、德国在加上中国，得到下图：



可以看到，对于最早采取最严重的管控措施的中国，疫情已经基本控制住，确诊数和死亡数基本不再上升，对于采取了一定管控措施的欧洲各国，确诊数和死亡数还在上升，但是数据远远好于采取自由放任态度的美国，这证明了严格的管控措施和隔离措施是十分有效的。

此外，我还绘制了各国死亡人数和死亡率排名的图：

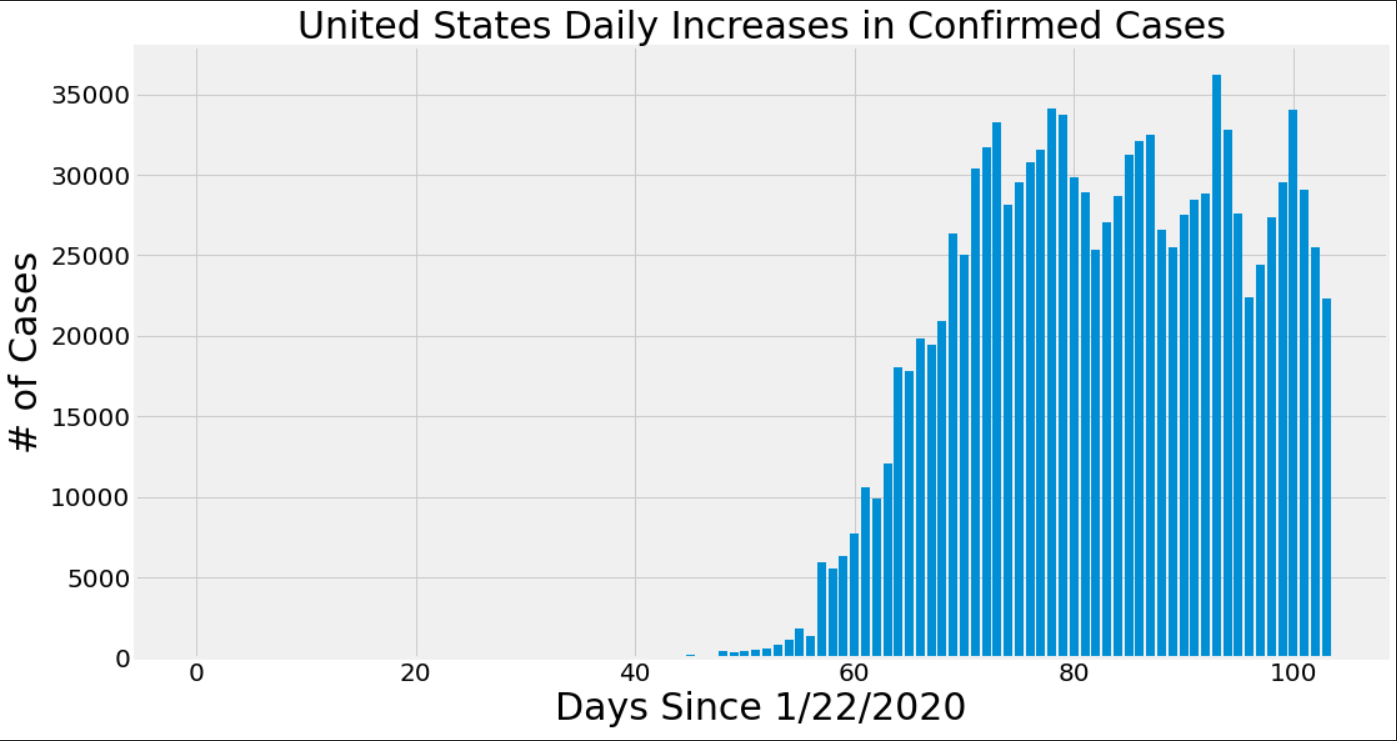
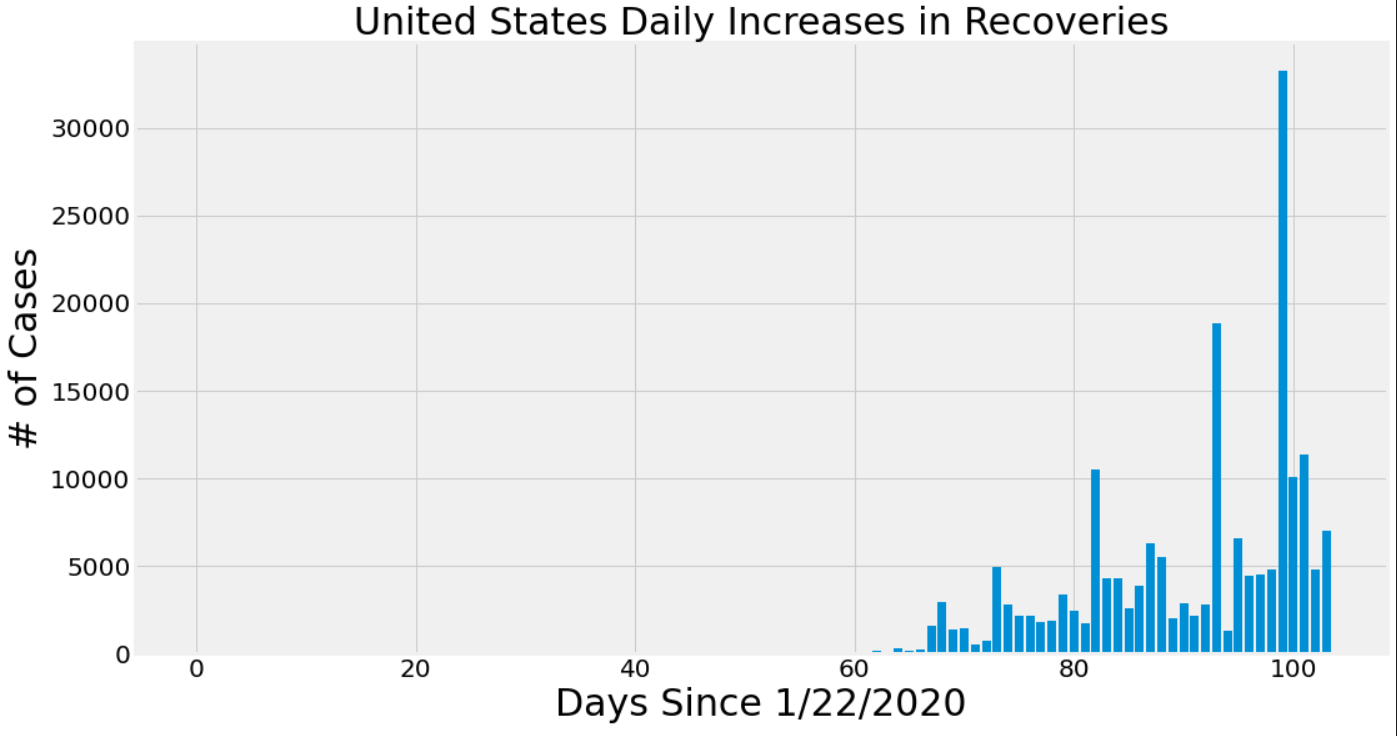
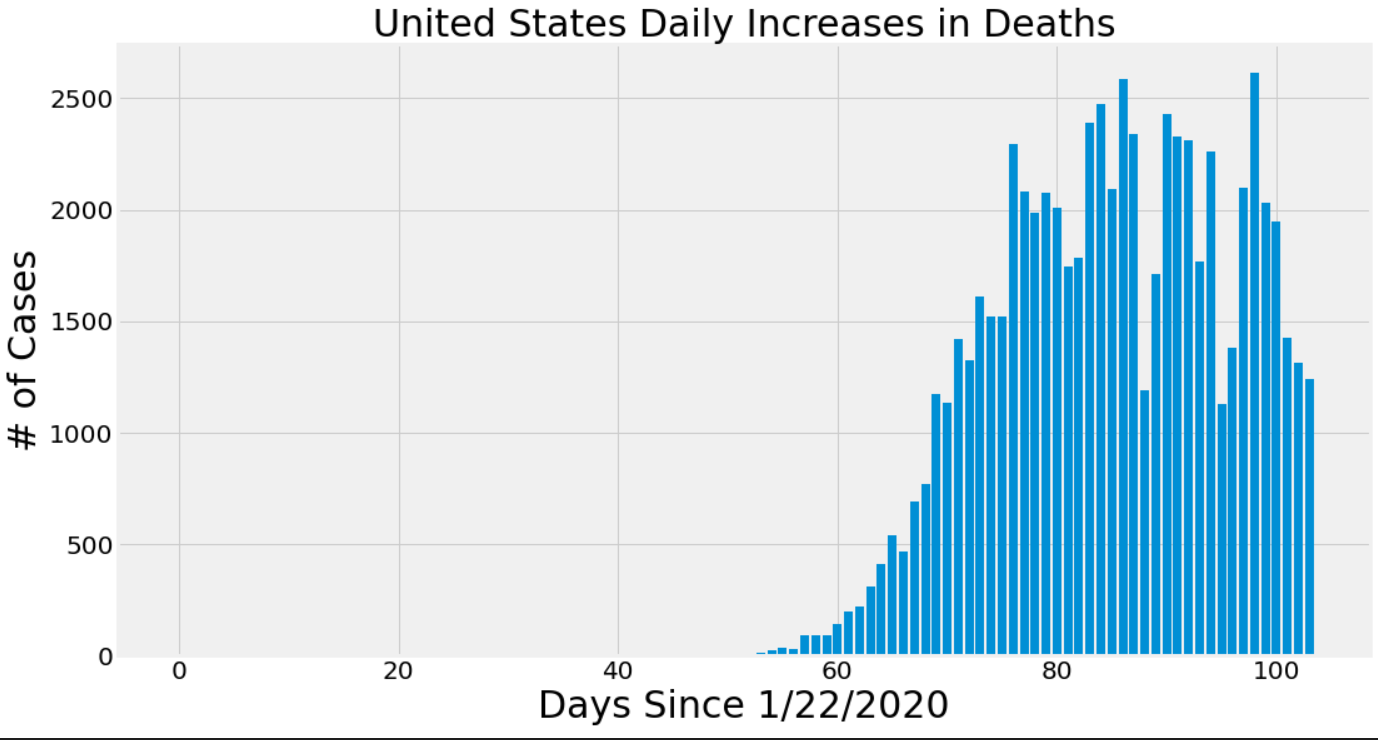
死亡人数排名靠前的国家

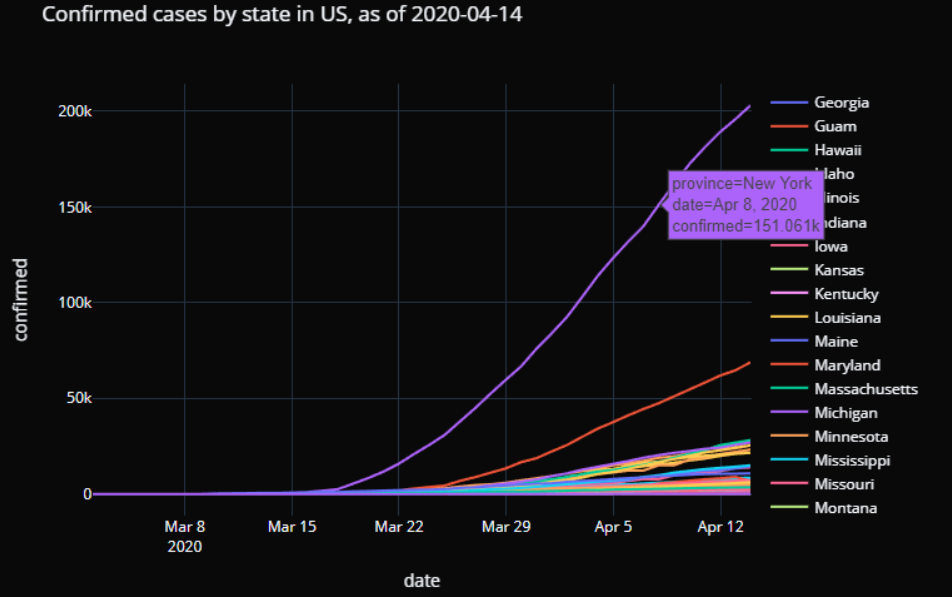
死亡率排名靠前国家

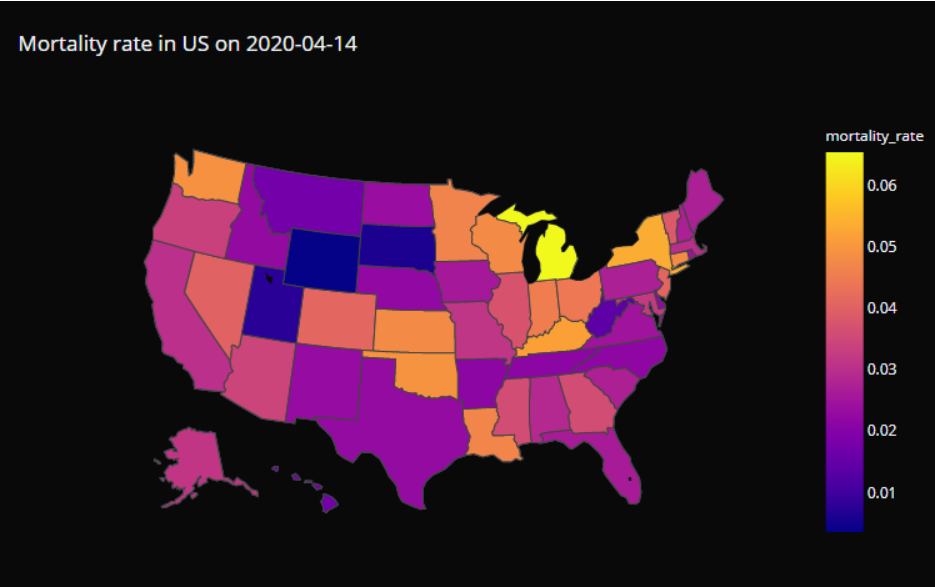
1. **各国具体分析（以美国、欧洲、亚洲为例）**

这里我做了很多国家的图，除此之外还包括了加拿大、澳大利亚等，这里只对有代表性的美国、欧洲、亚洲（由于中国境内疫情已基本停息，数据不再增长，所以作为一个整体进行分析比较合适）进行分析。

1. 美国

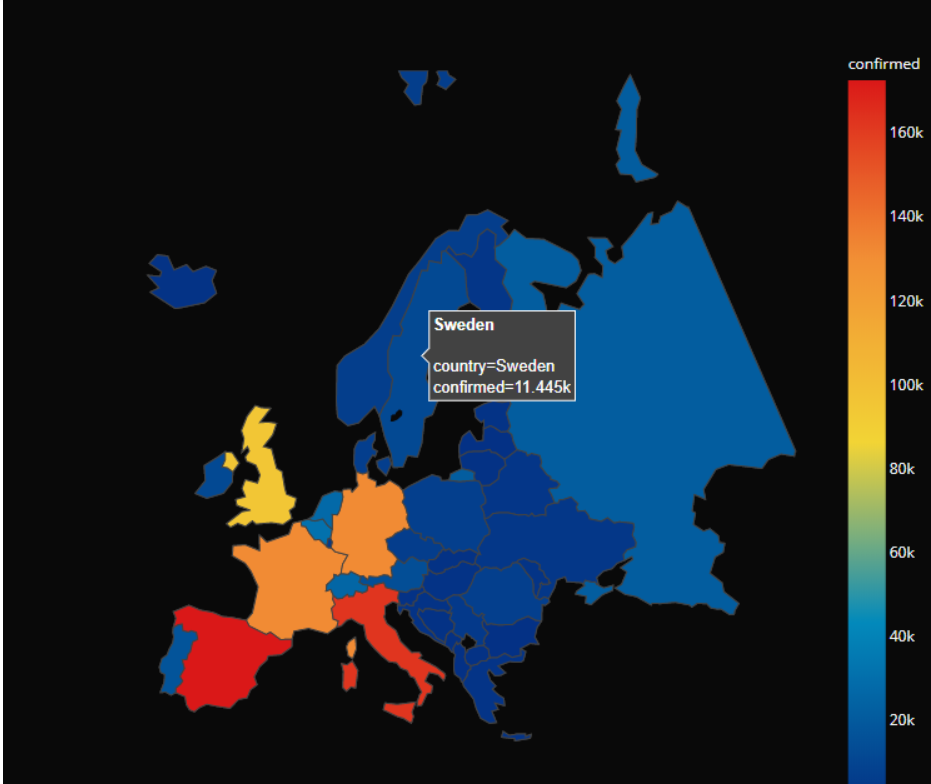


可以看到，美国作为当前疫情最严重的国家，每日新增病例仍远远大于痊愈病例，累计患者数目仍在快速上升，死亡人数也居高不下。下面我们看看各州的情况：

纽约州的情况最为严重，远远超过其他49个州，此外，新泽西州的疫情也比较严重。下面我们用更直观的图看一下各州的死亡率情况：

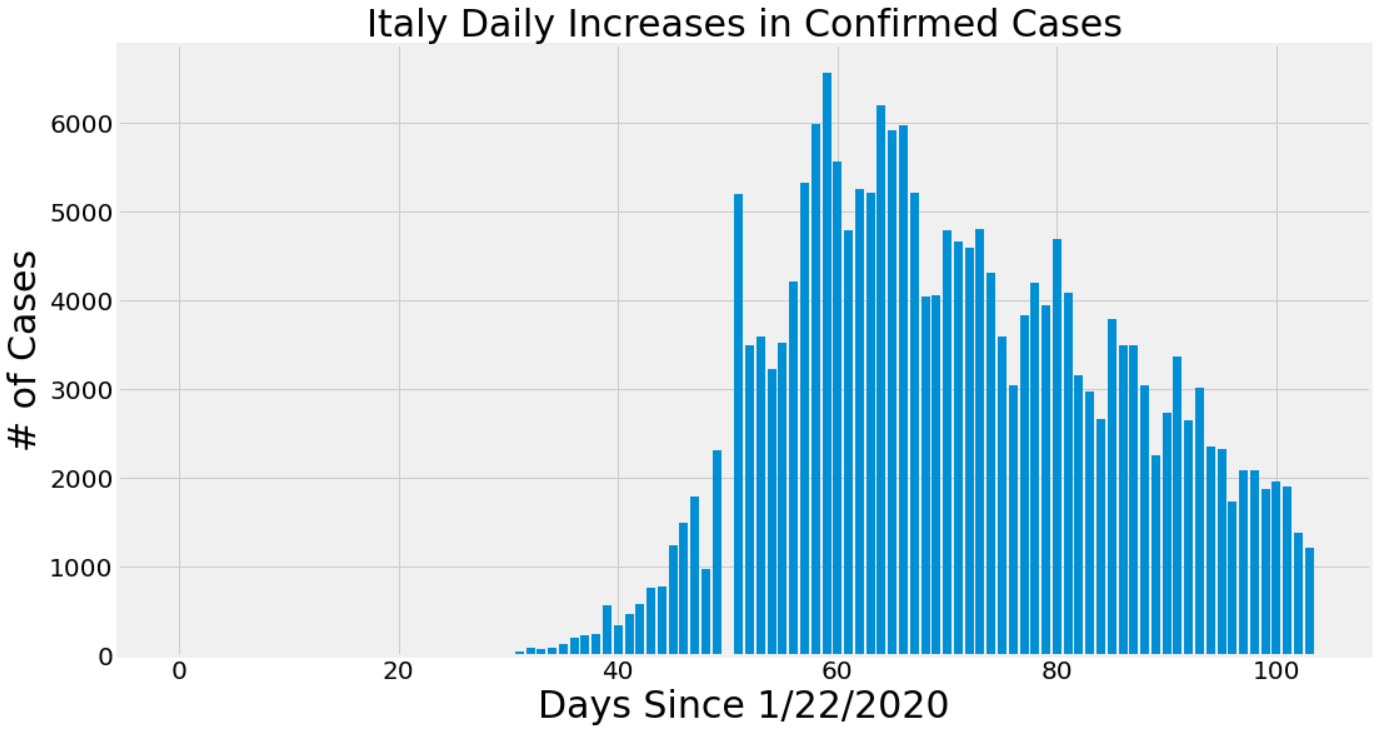
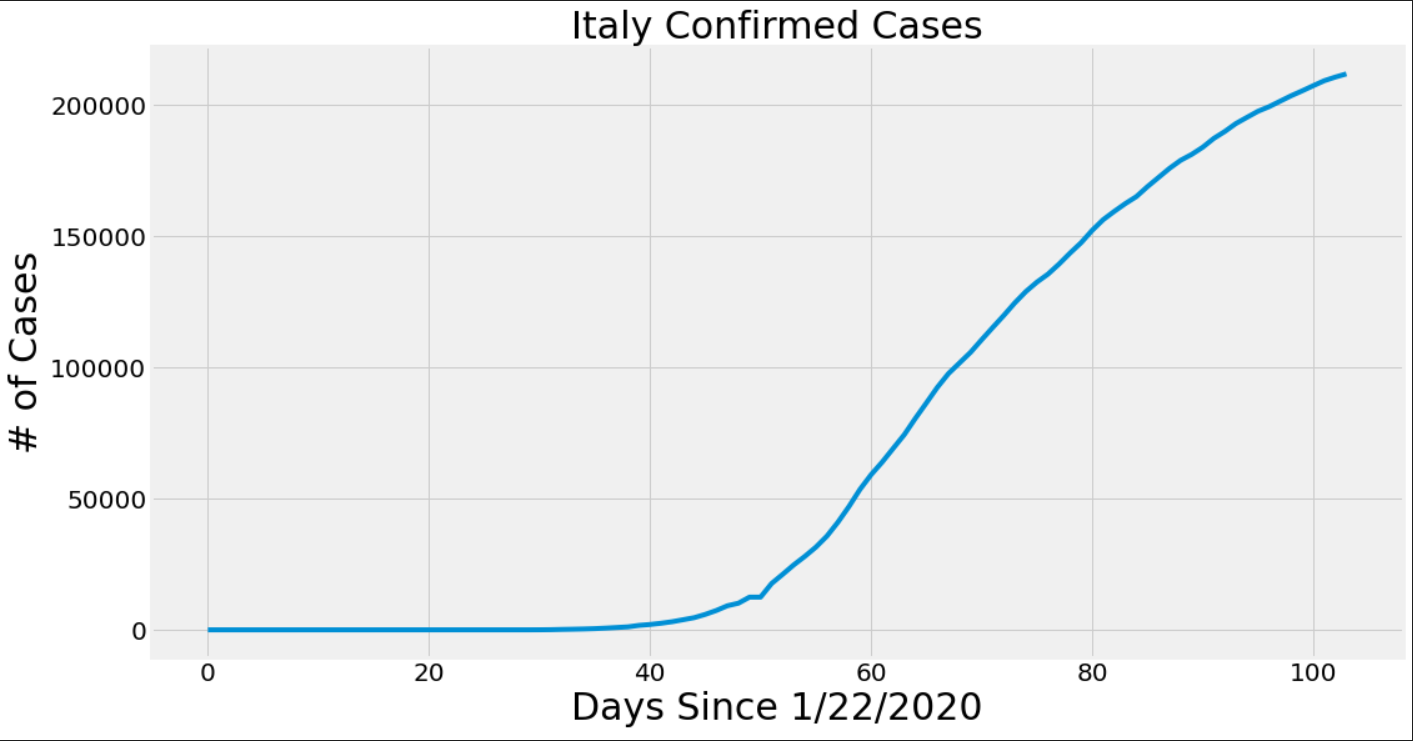
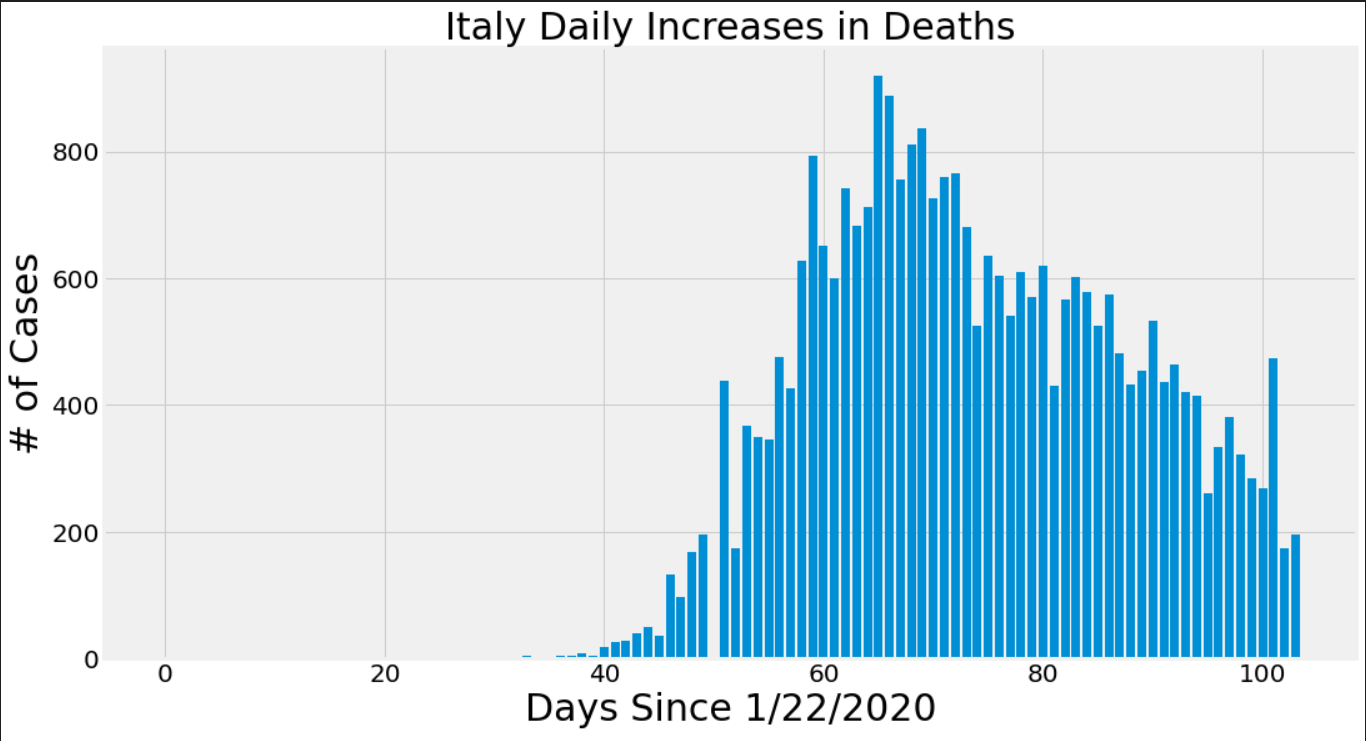
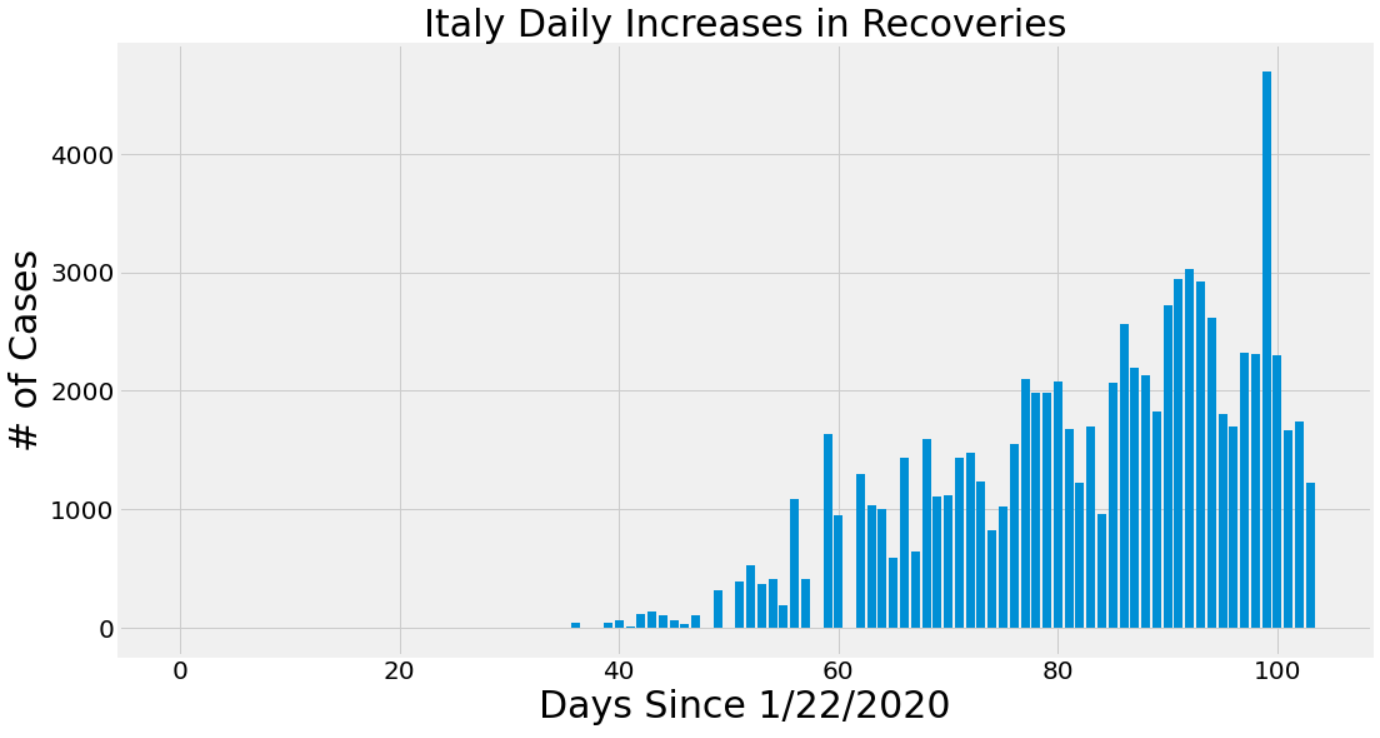
可以看到，密歇根州的情况最为严重，病死率超过了6%，此外，纽约州、华盛顿州、肯塔基州和路易斯安那州的情况也不容乐观。

1. 欧洲

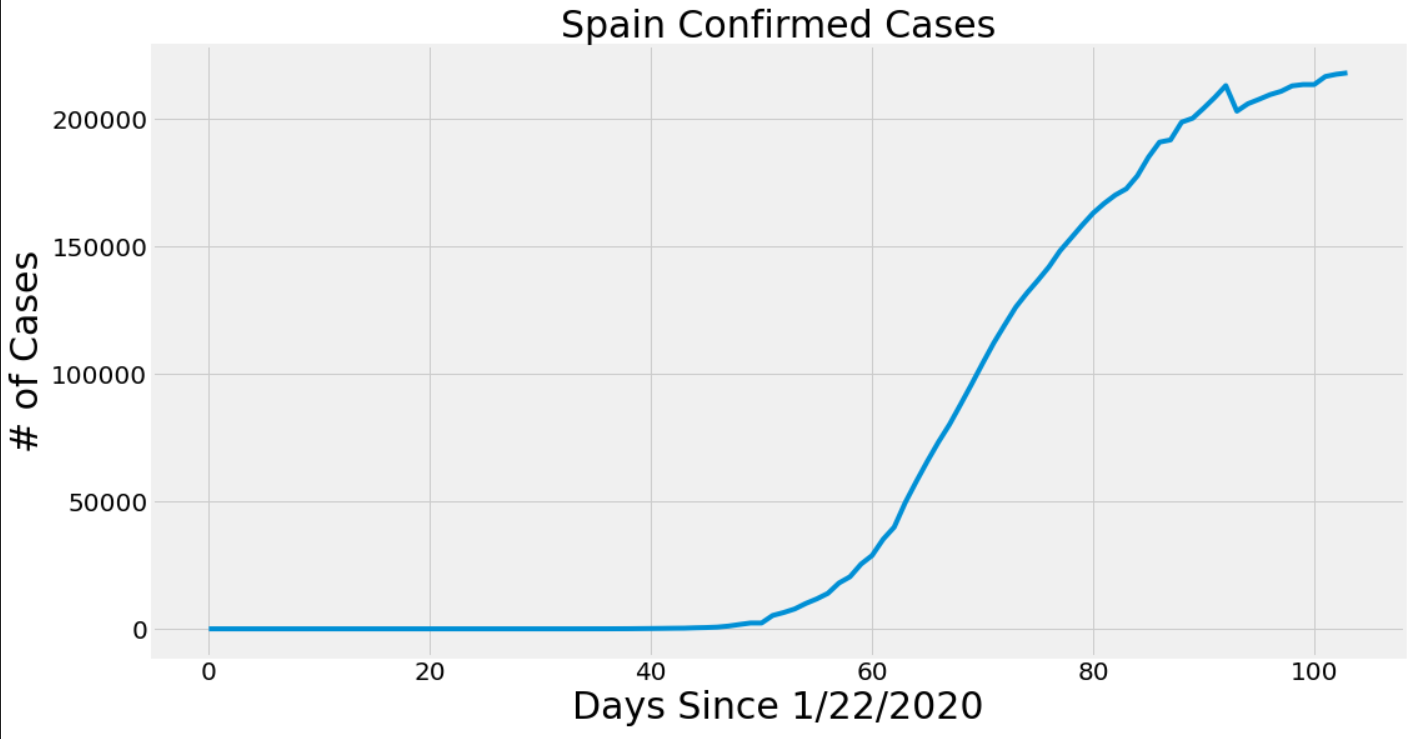
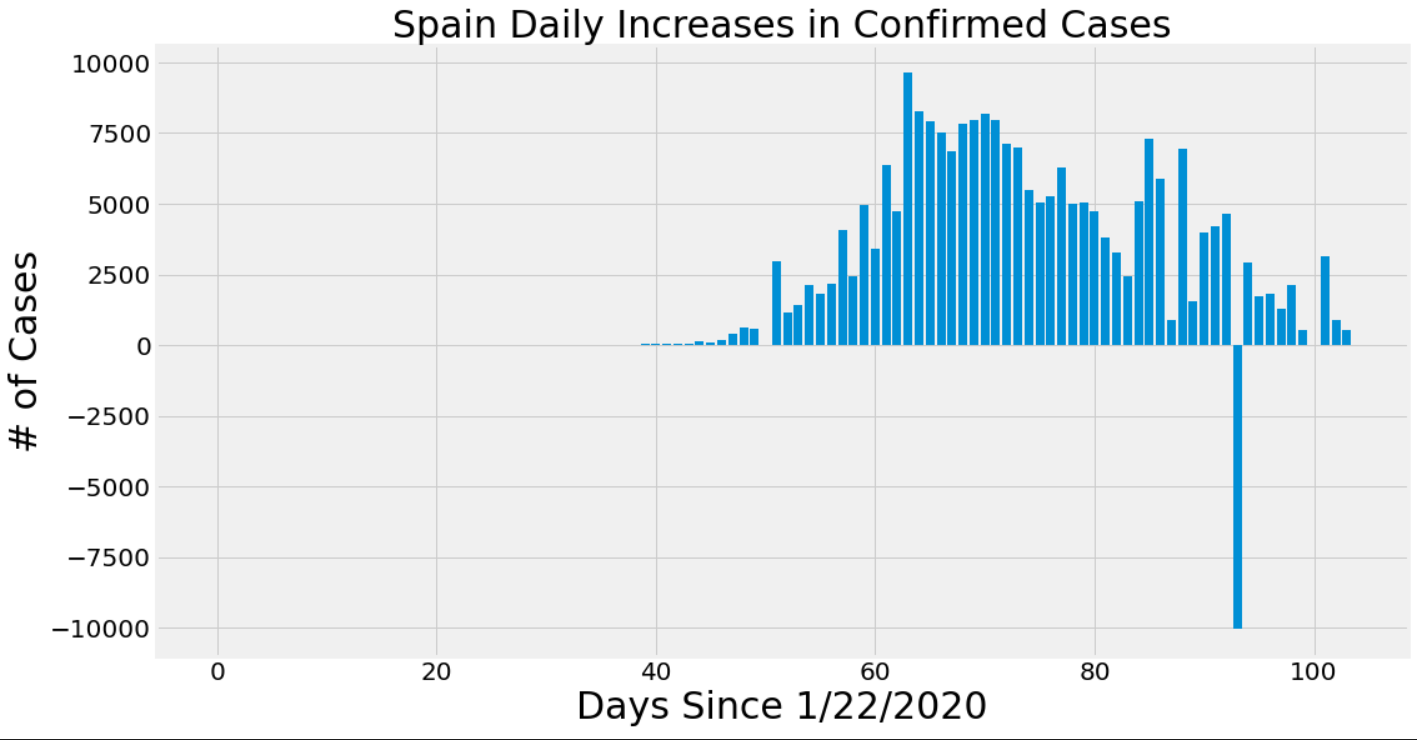
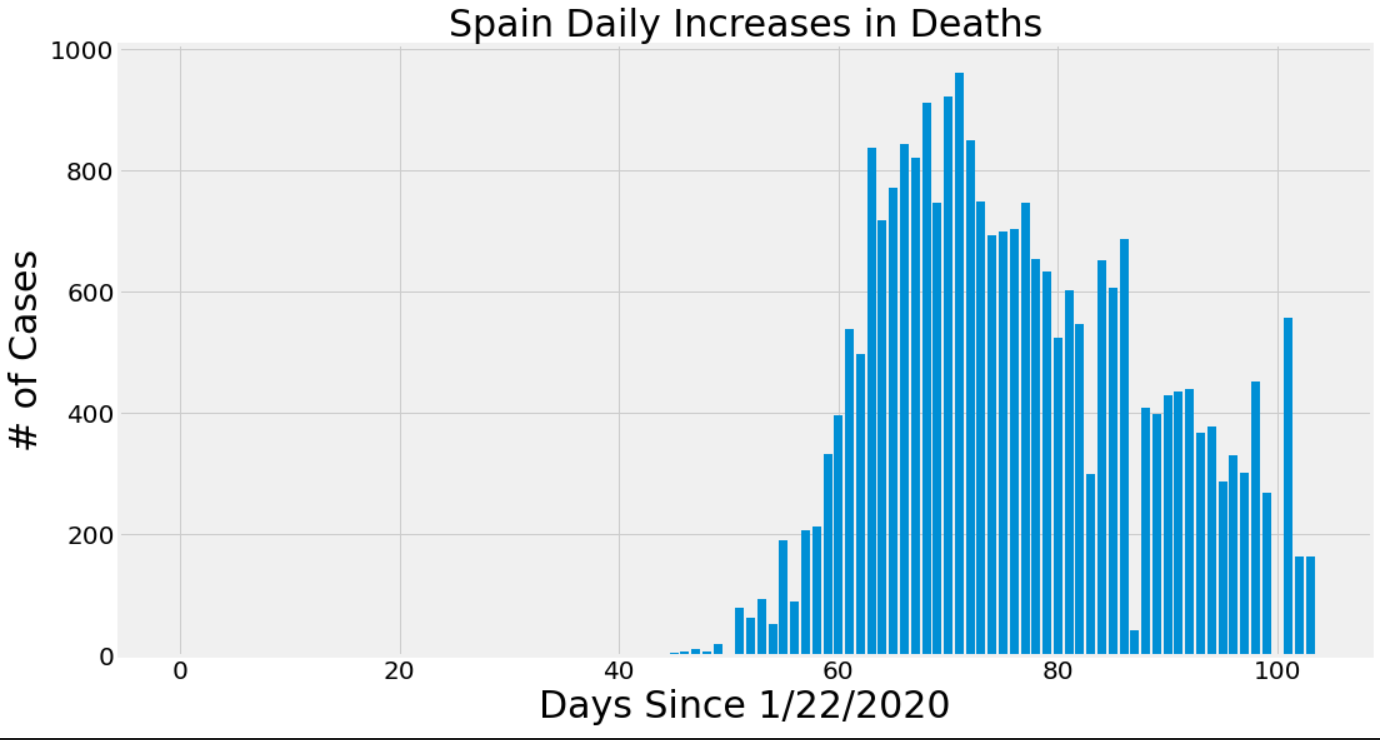
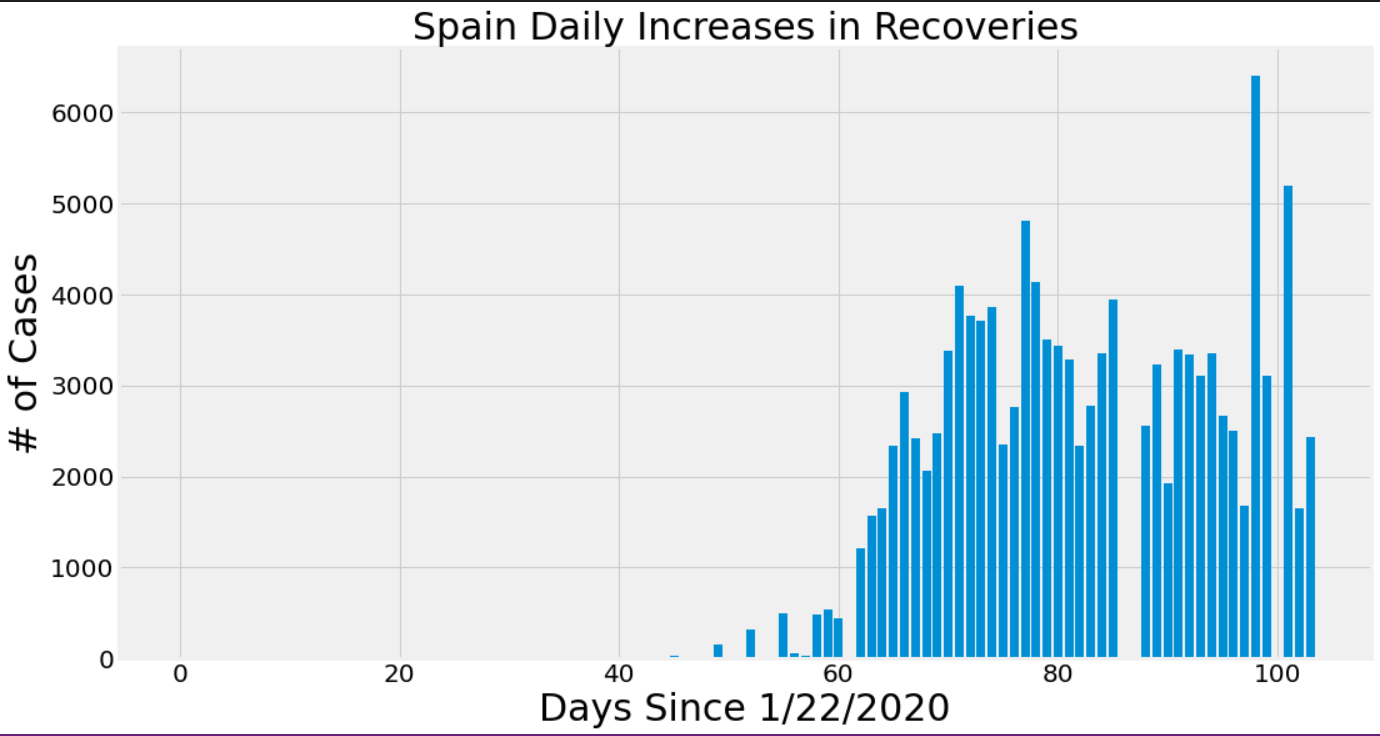
通过这张图能直观的反应欧洲各国的疫情情况：

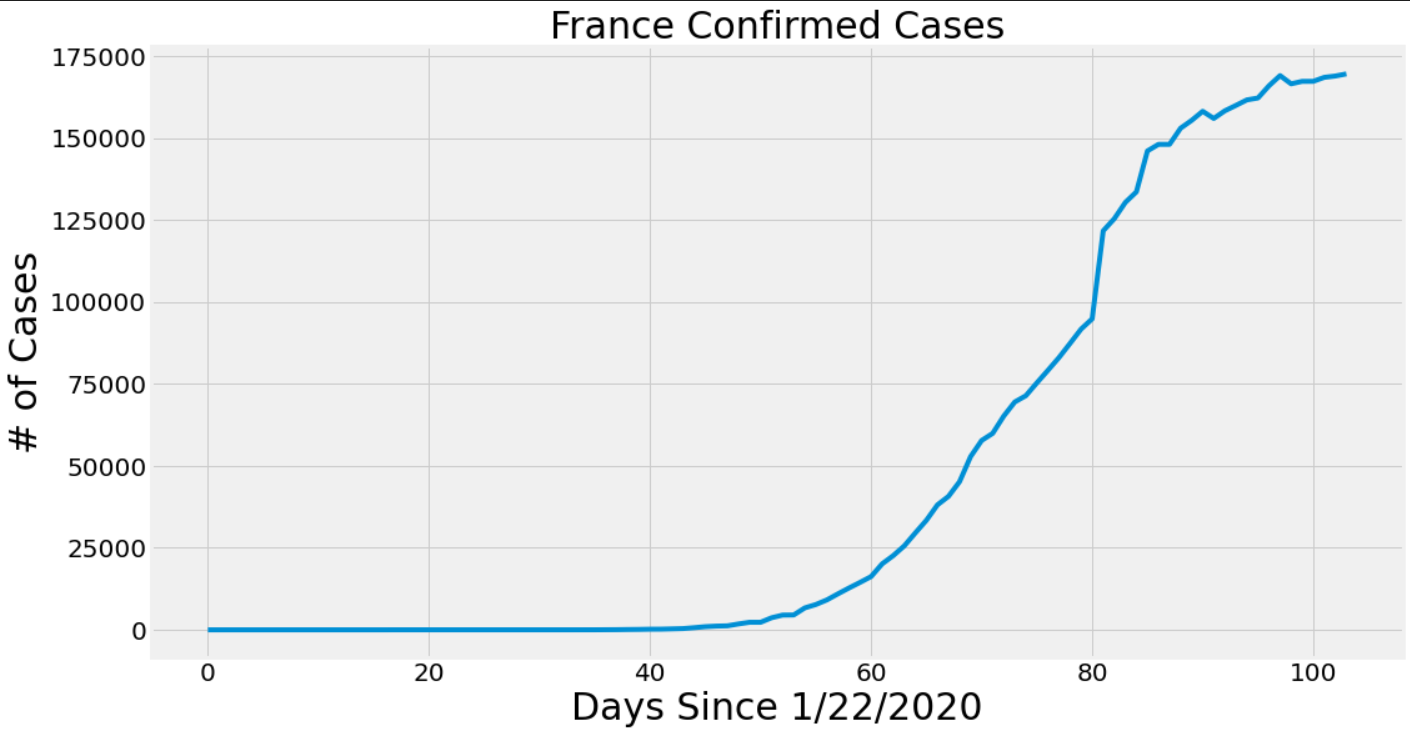
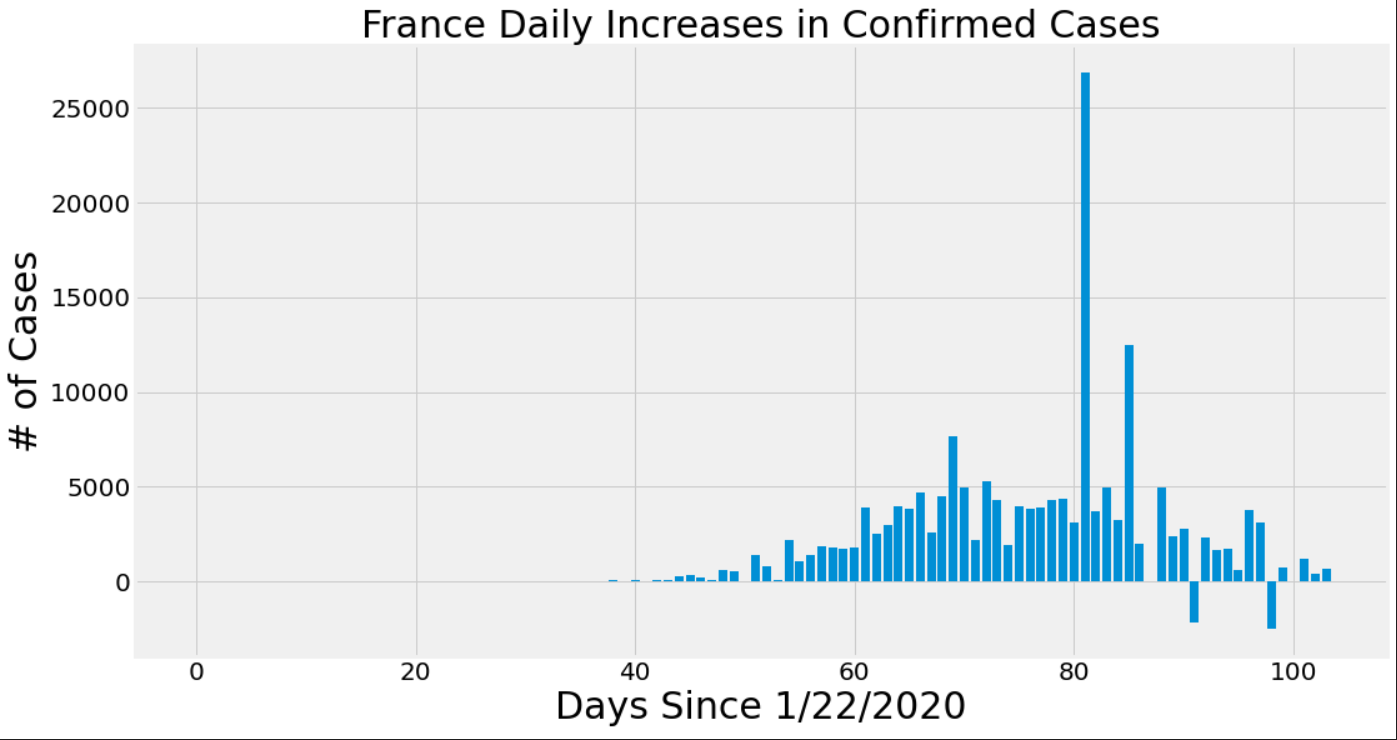
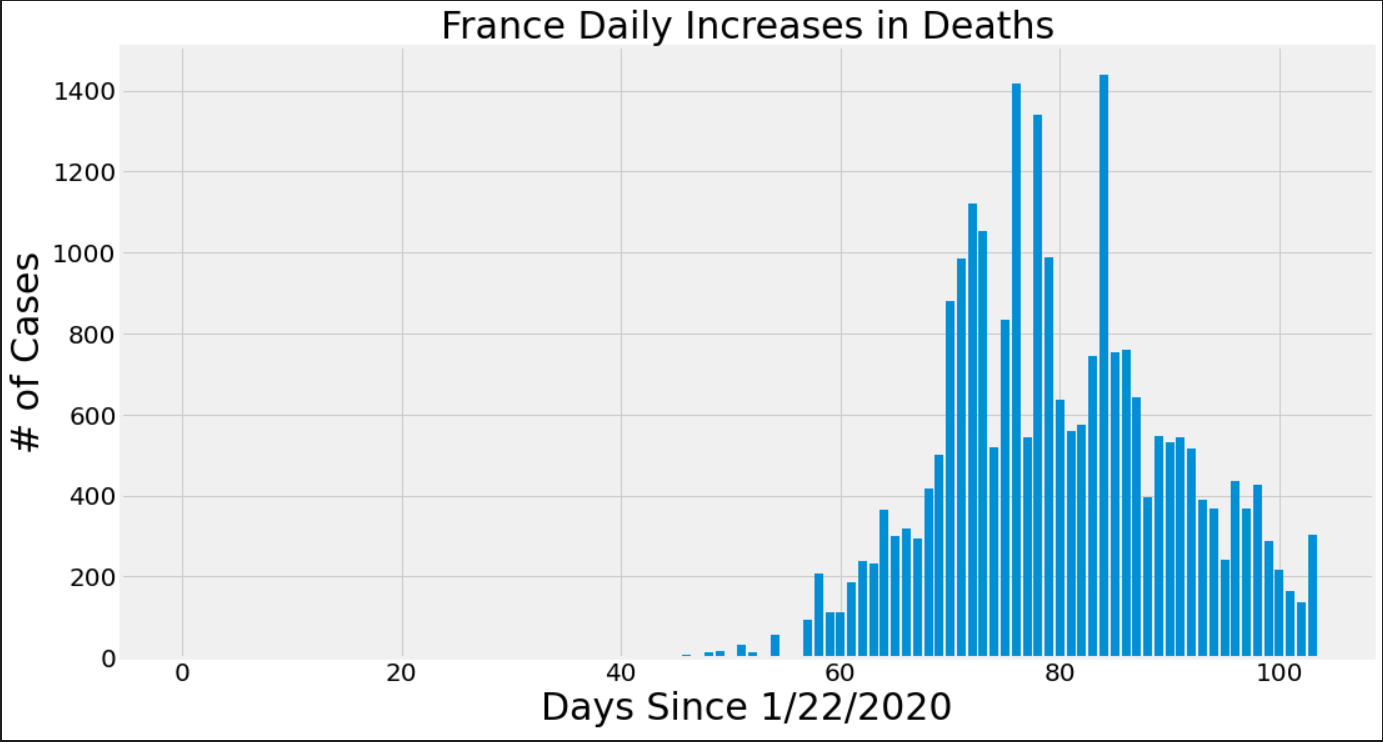
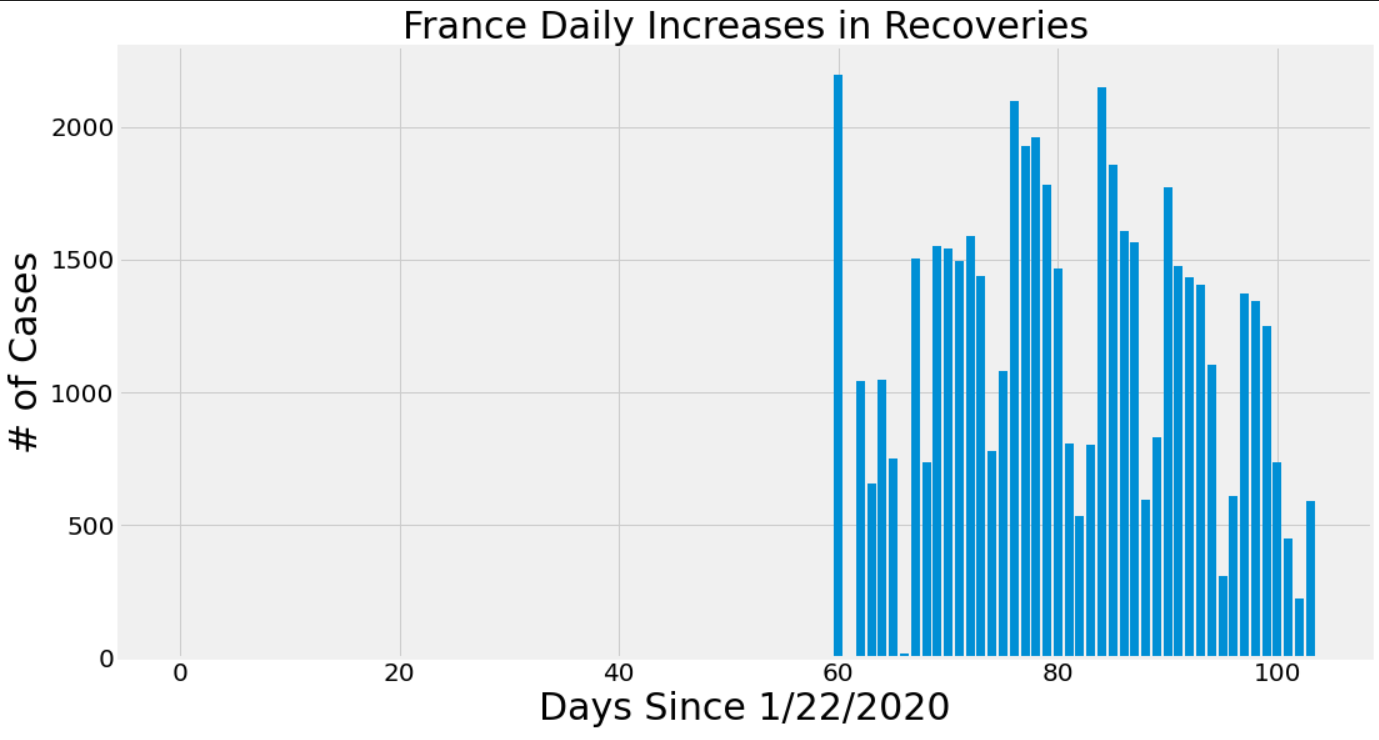
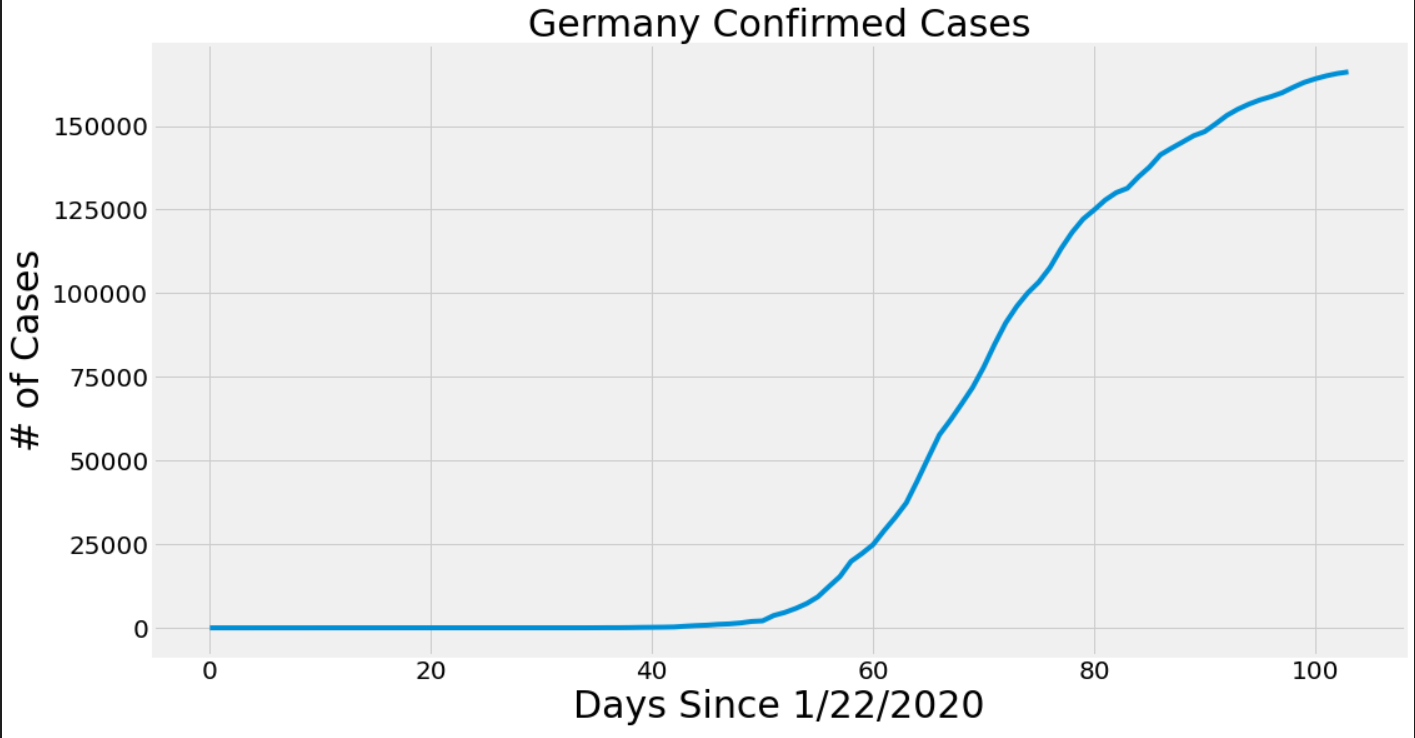
可以看到，作为欧洲最早爆发地的意大利和西班牙目前情况最为严重，感染人数均超过了16万人，紧随其后的是德国、法国以及首相王储身先士卒践行“群体免疫”的大英帝国，其他国家的情况相对来说还处于可控的状态中。

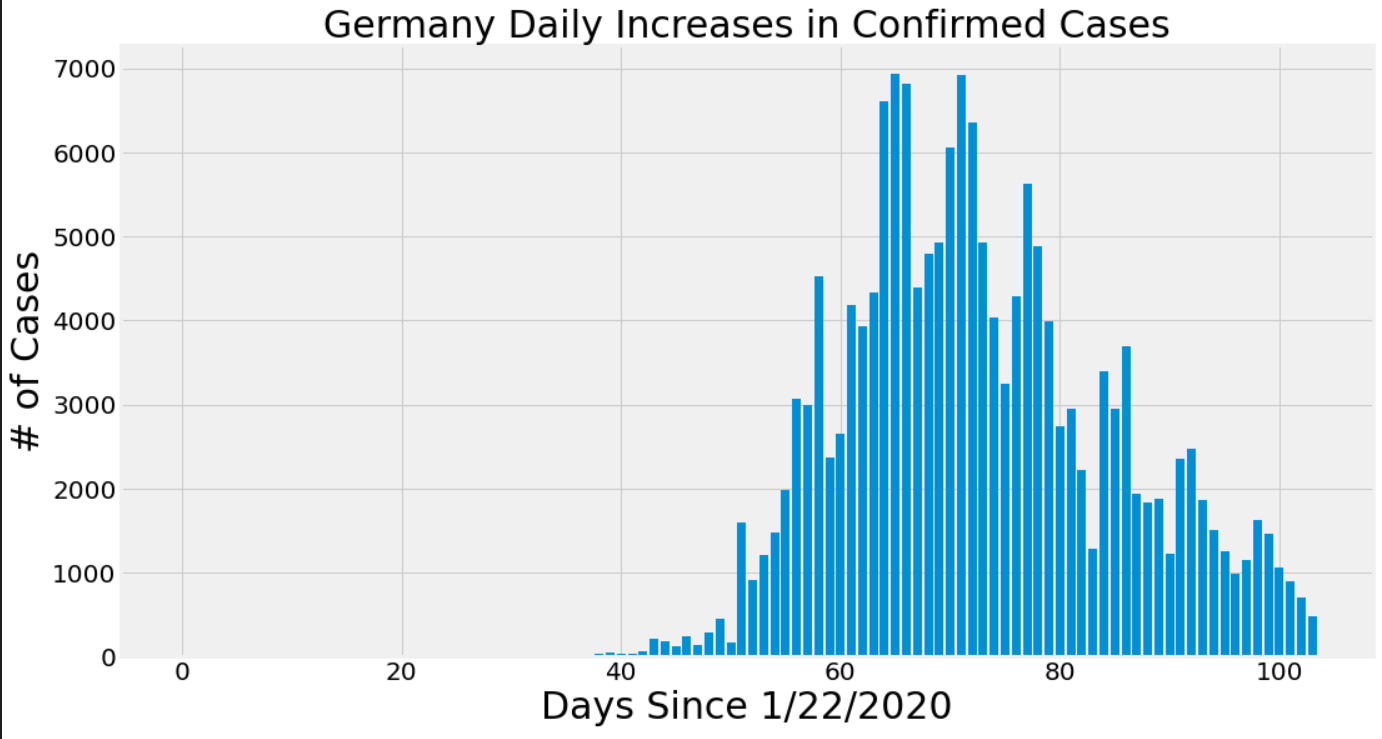
下边我们具体看一下，意大利、西班牙、法国、德国、英国以及最近增长较快的俄罗斯的每日新增感染、死亡、痊愈人数：

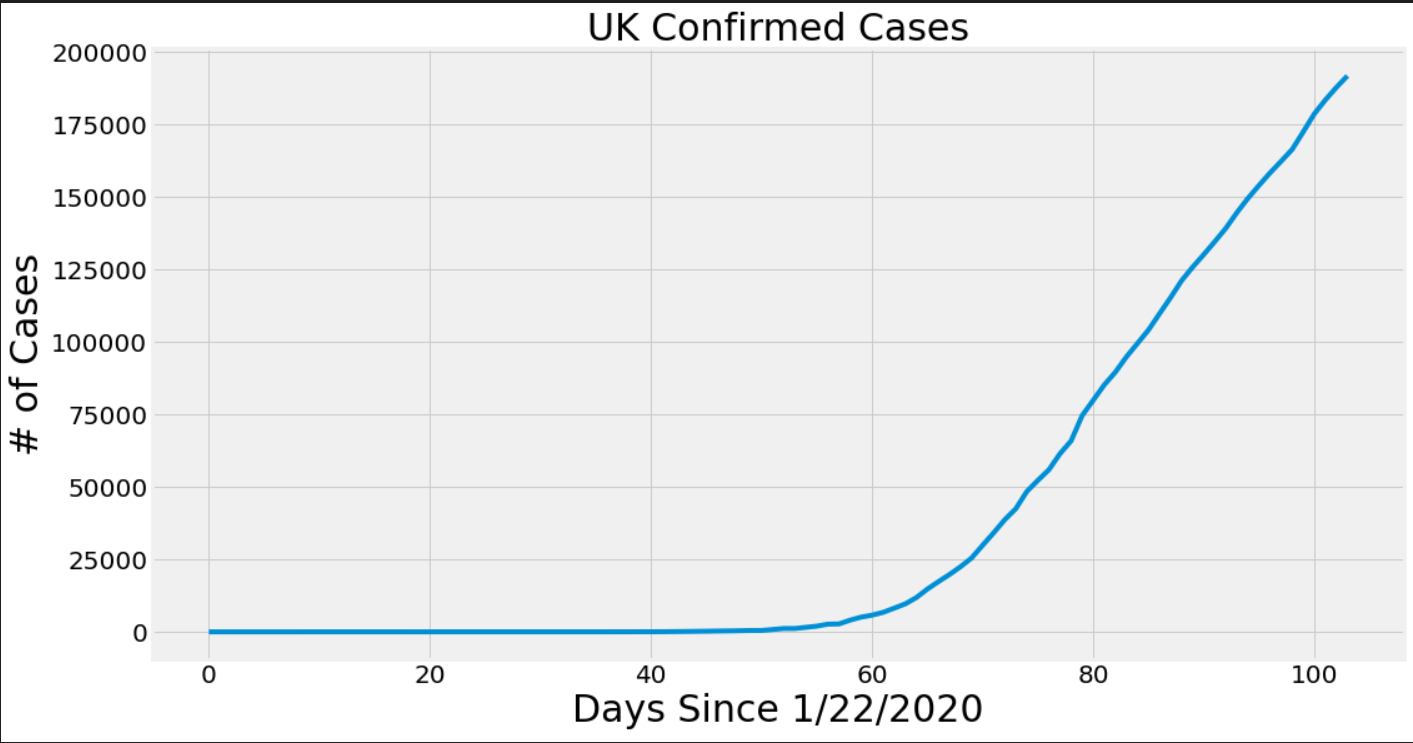
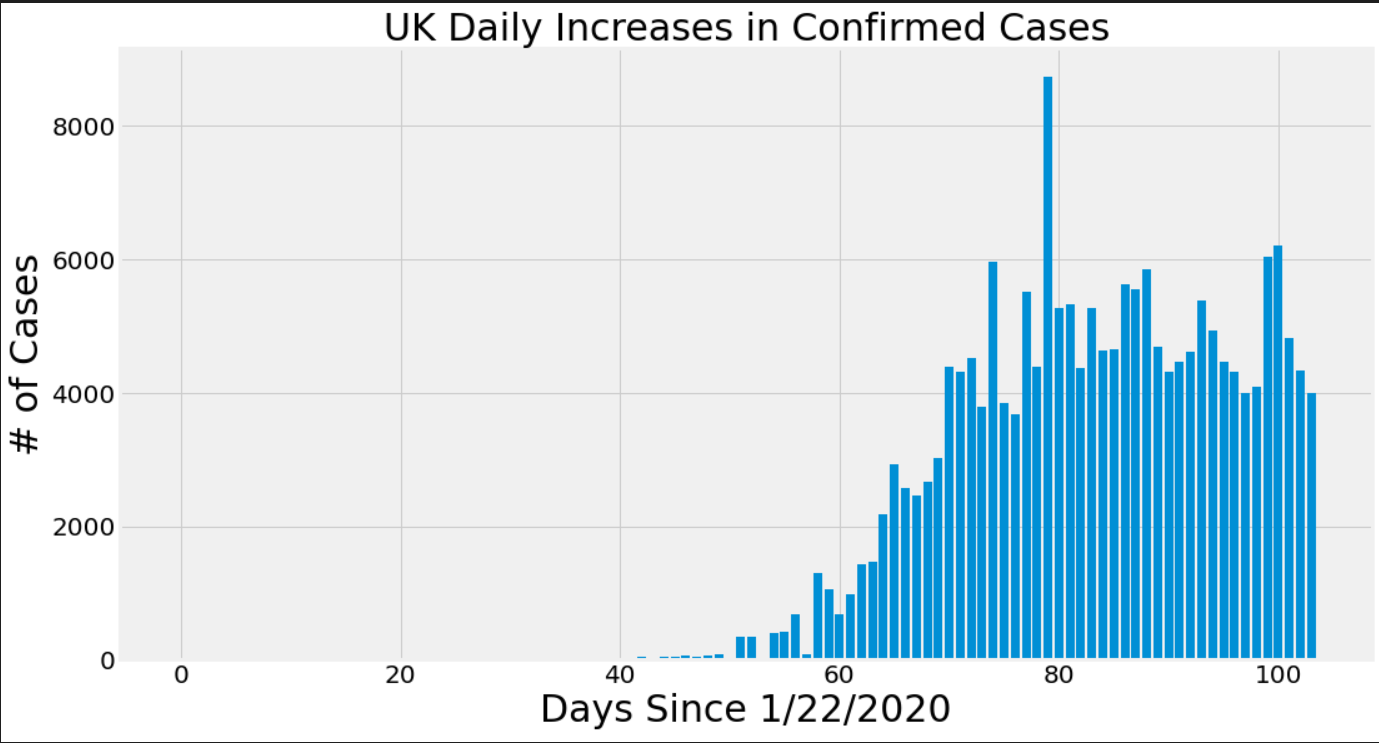
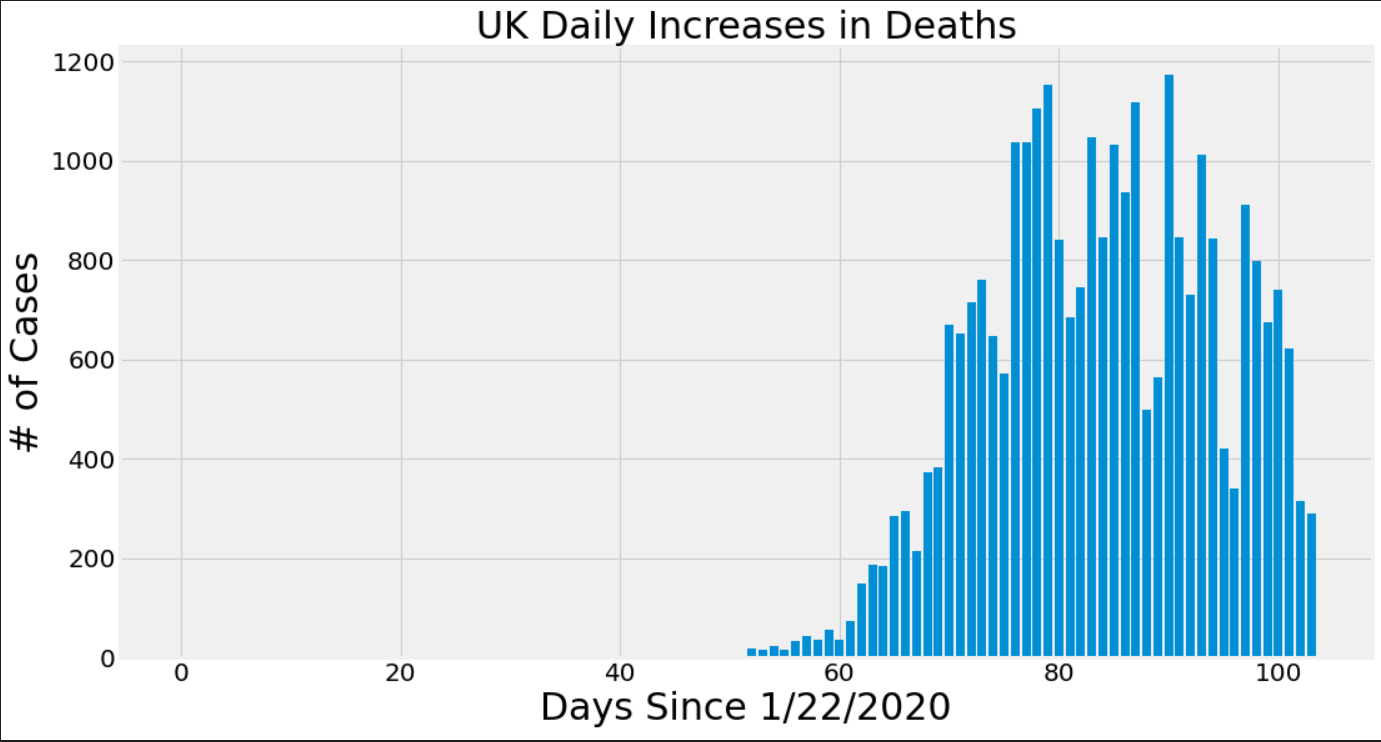
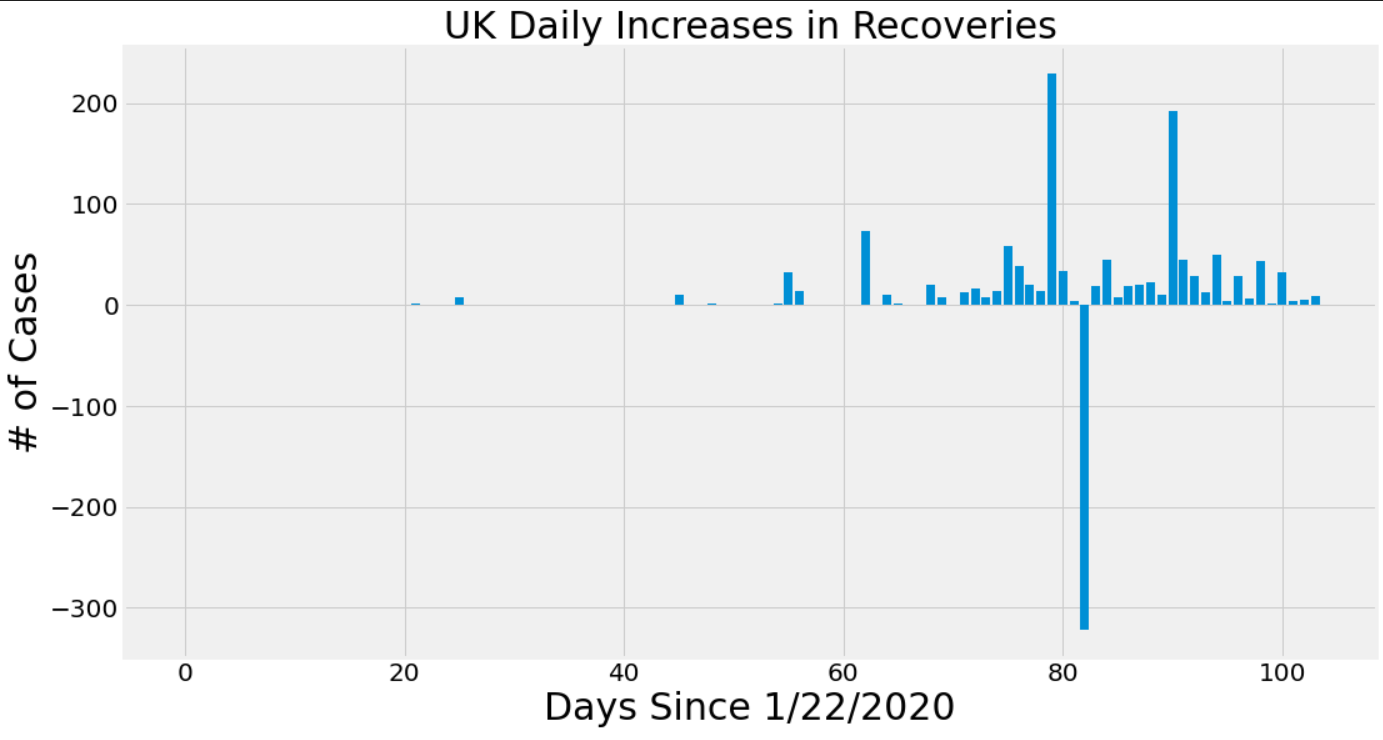
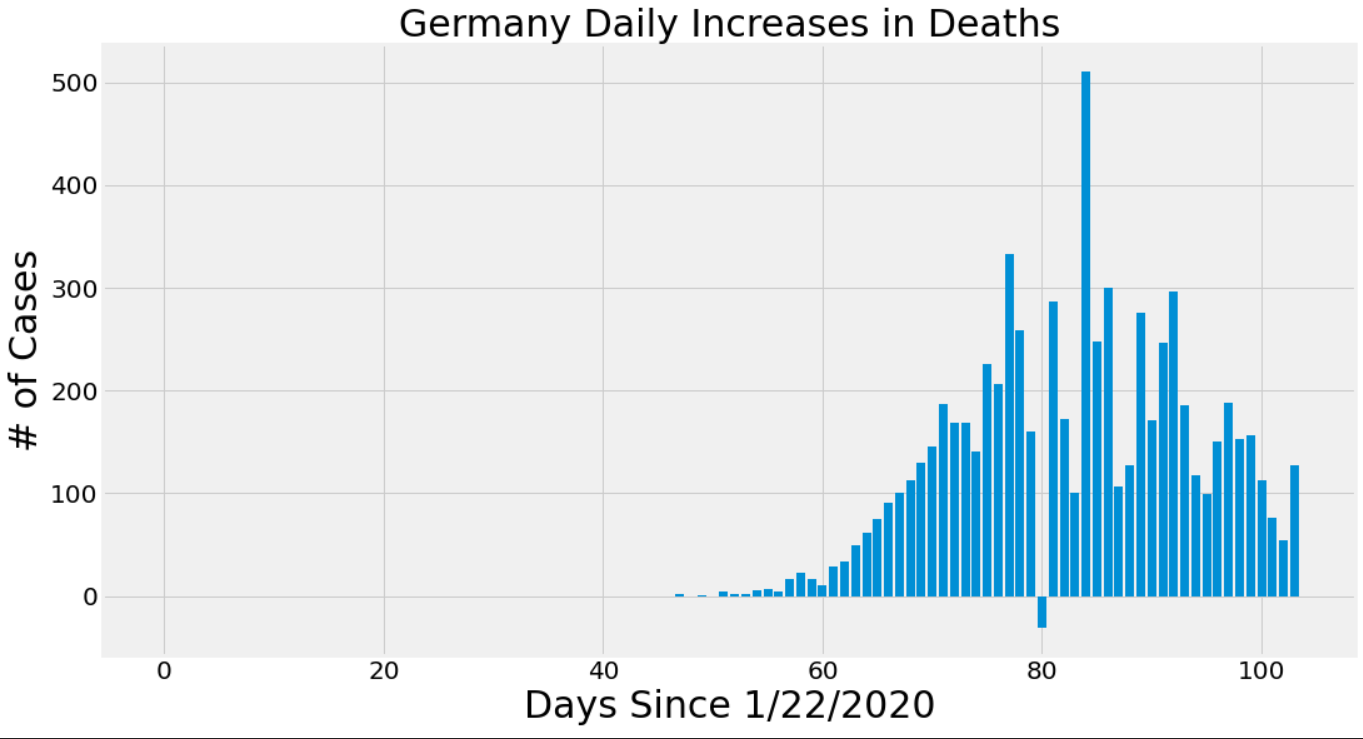
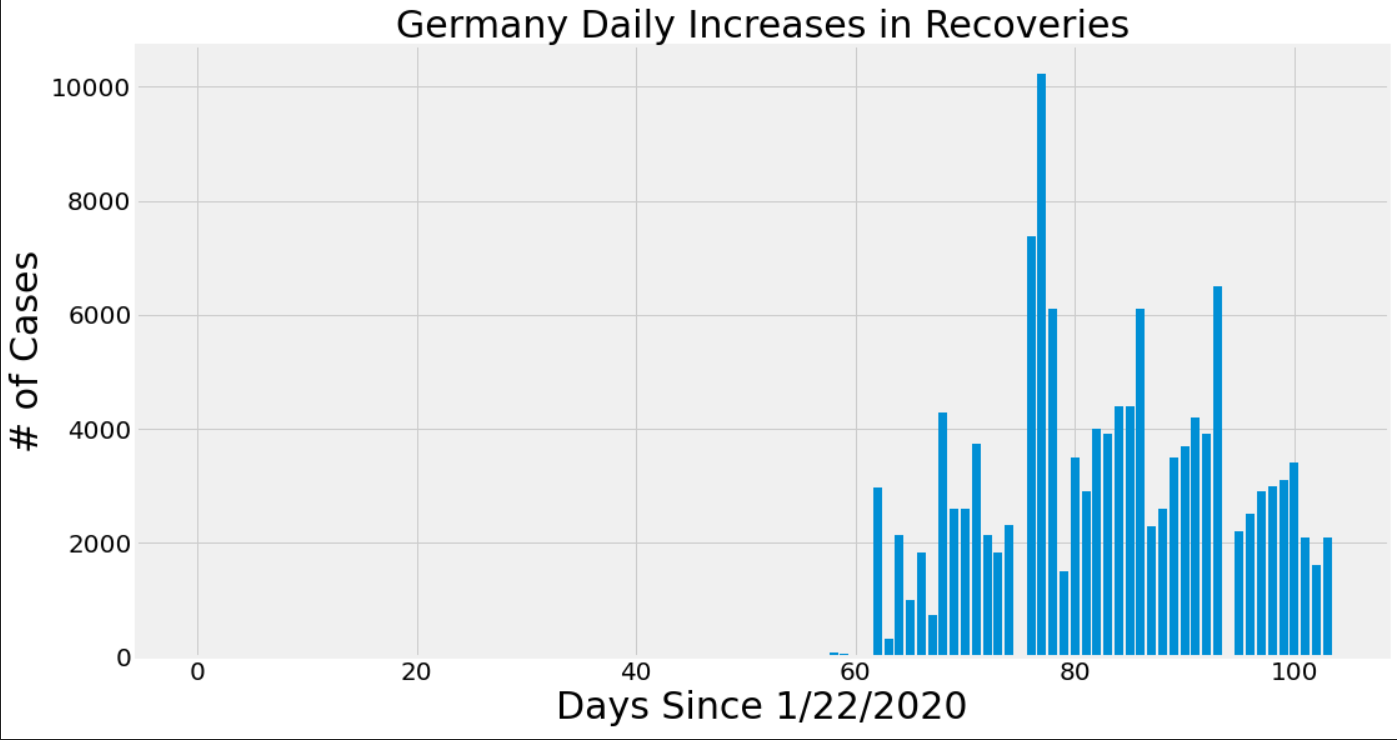
意大利：

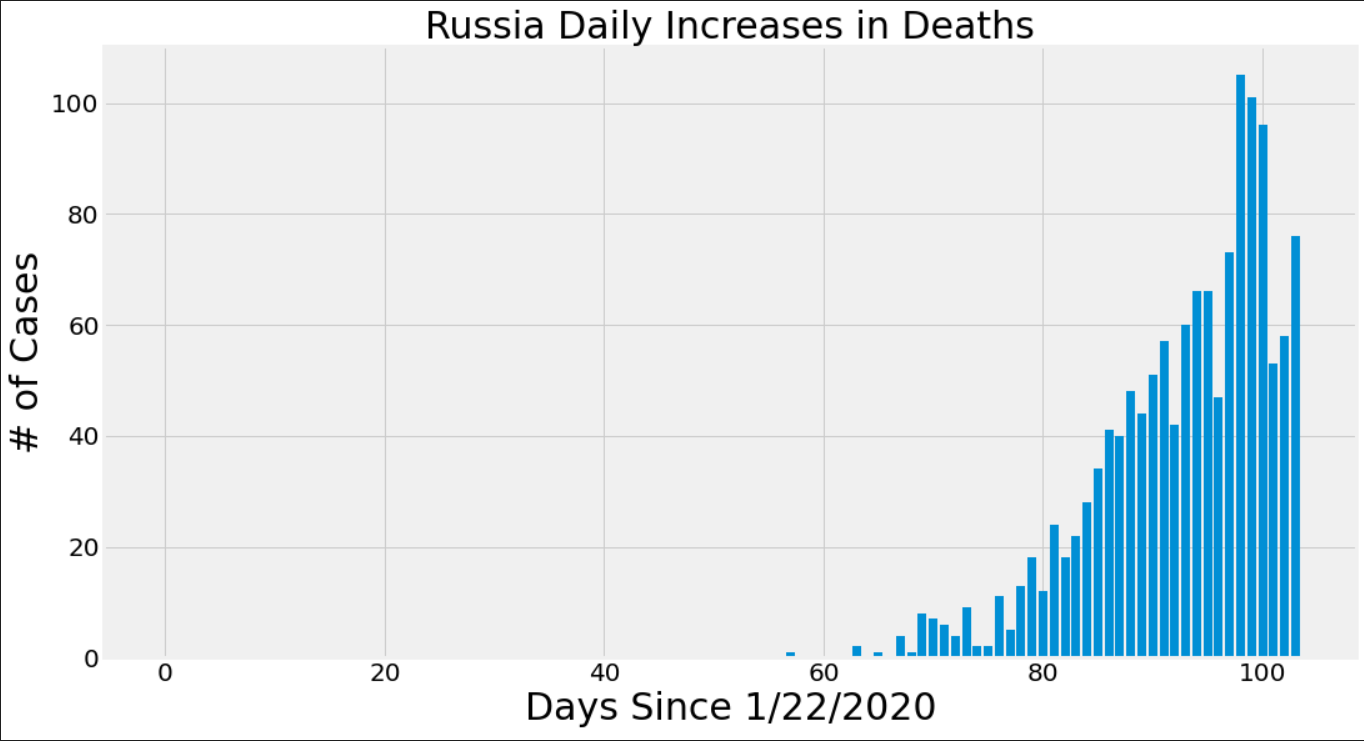
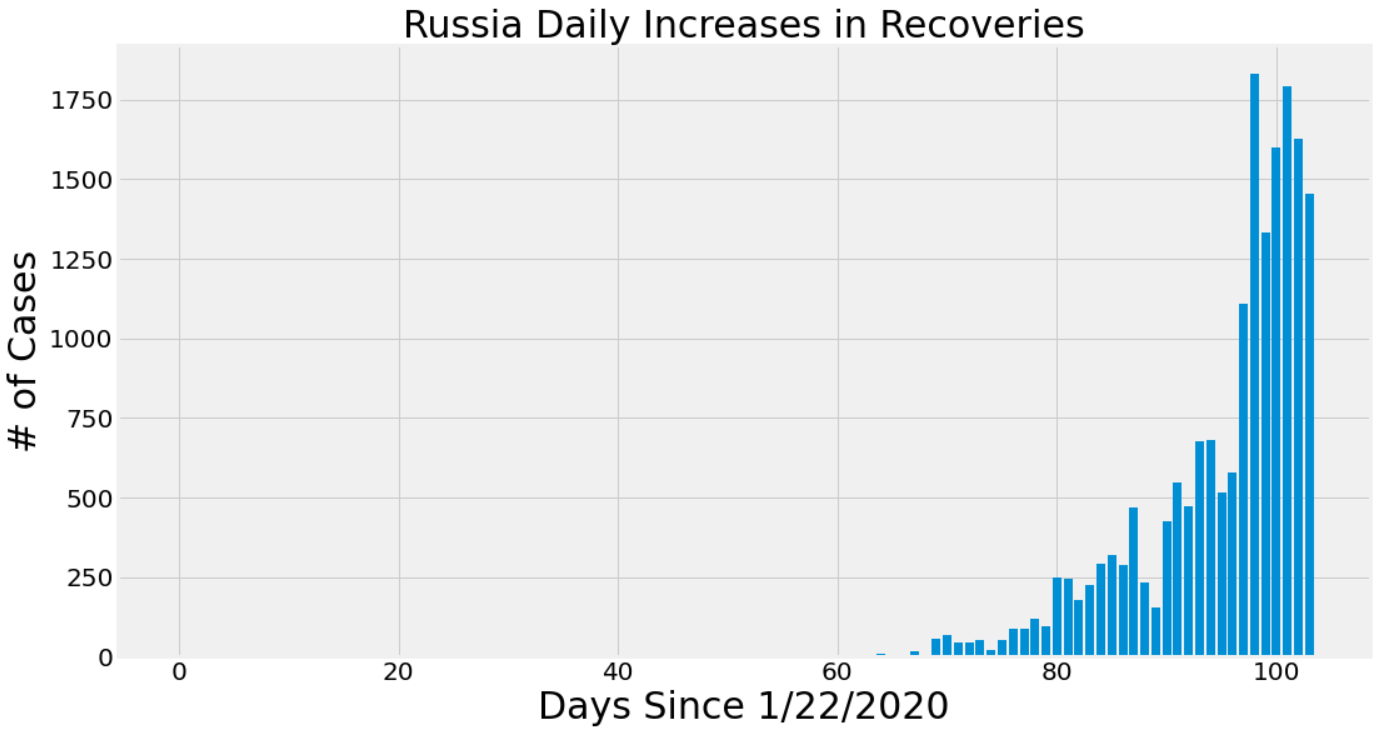
西班牙：



法国：

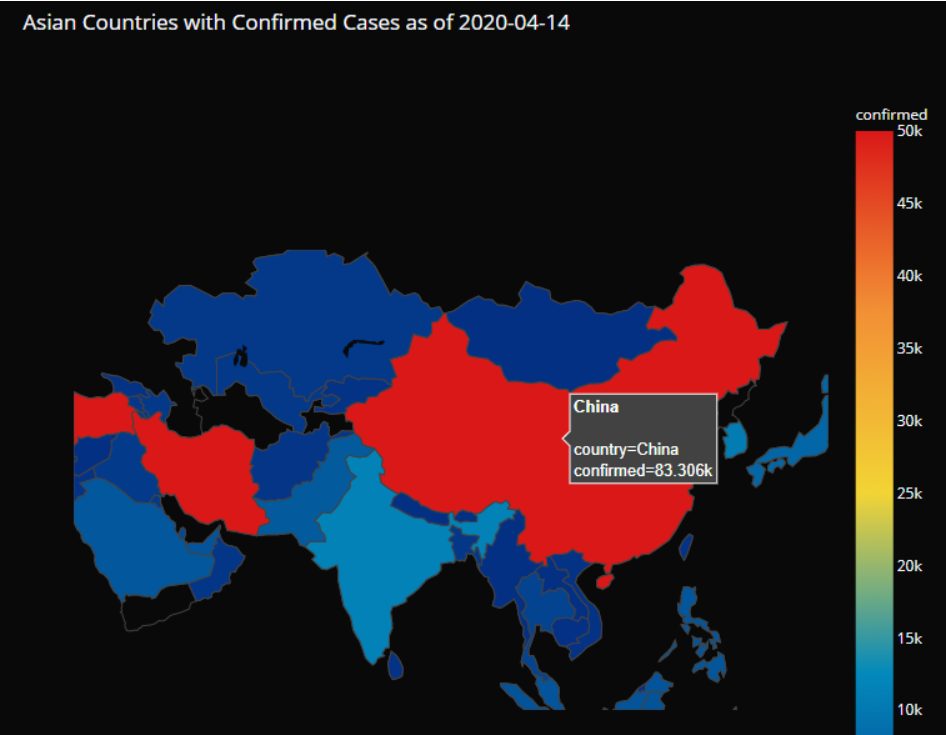
德国：

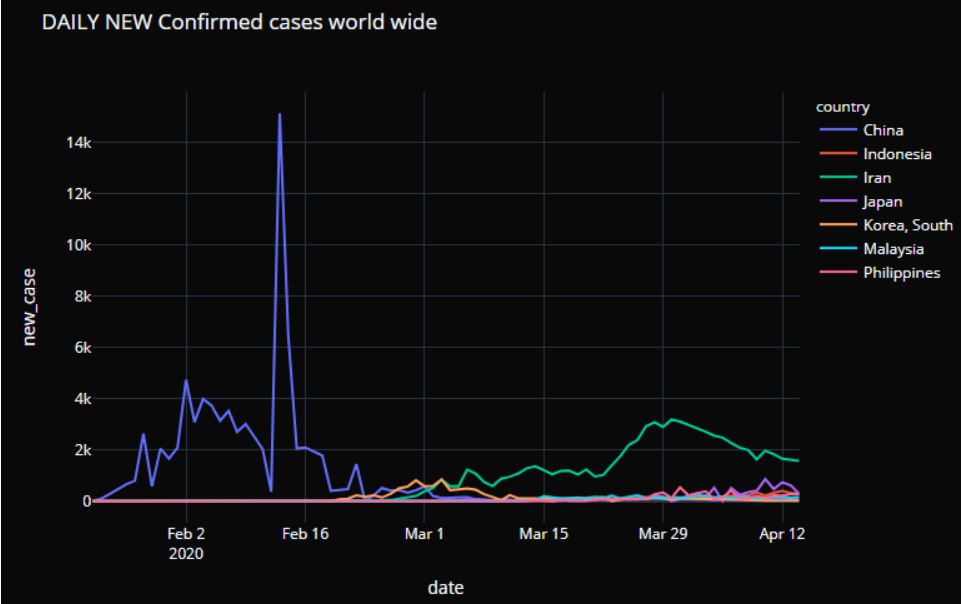
英国：

俄罗斯：



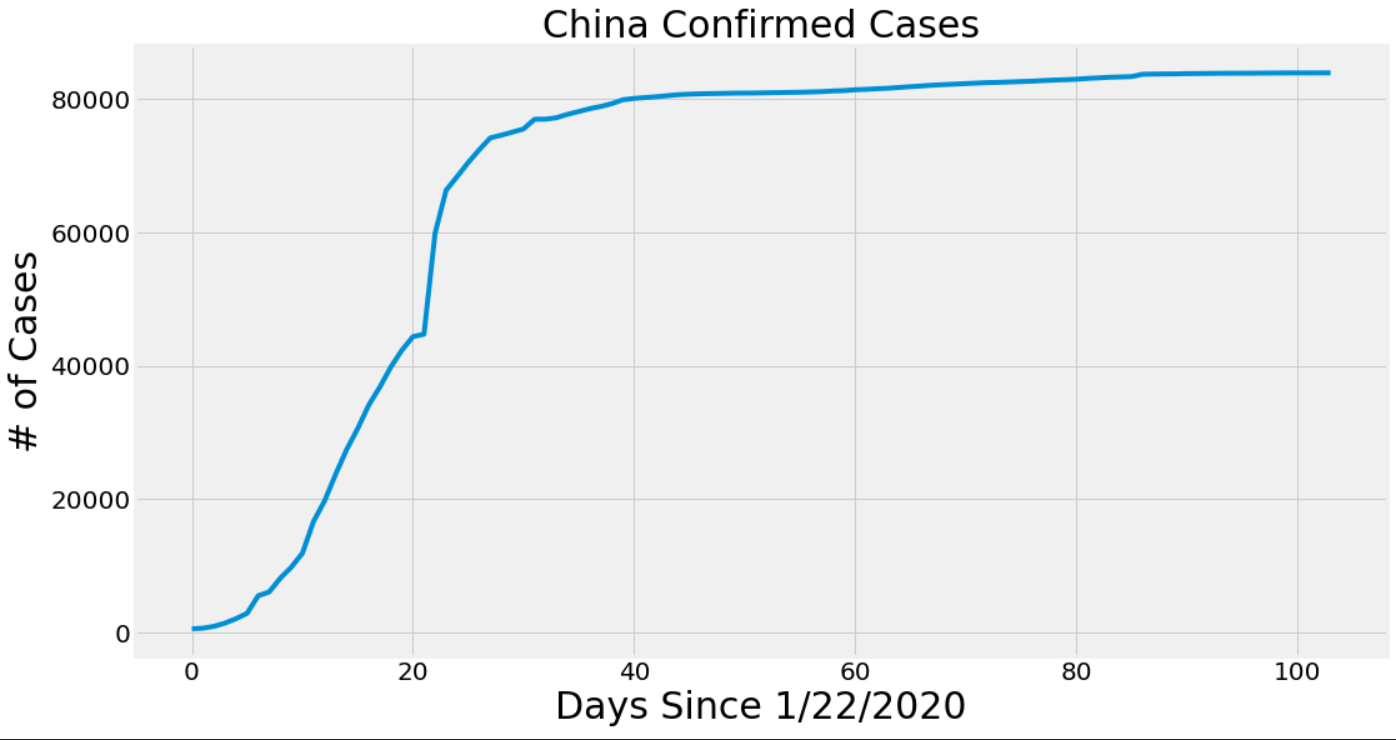
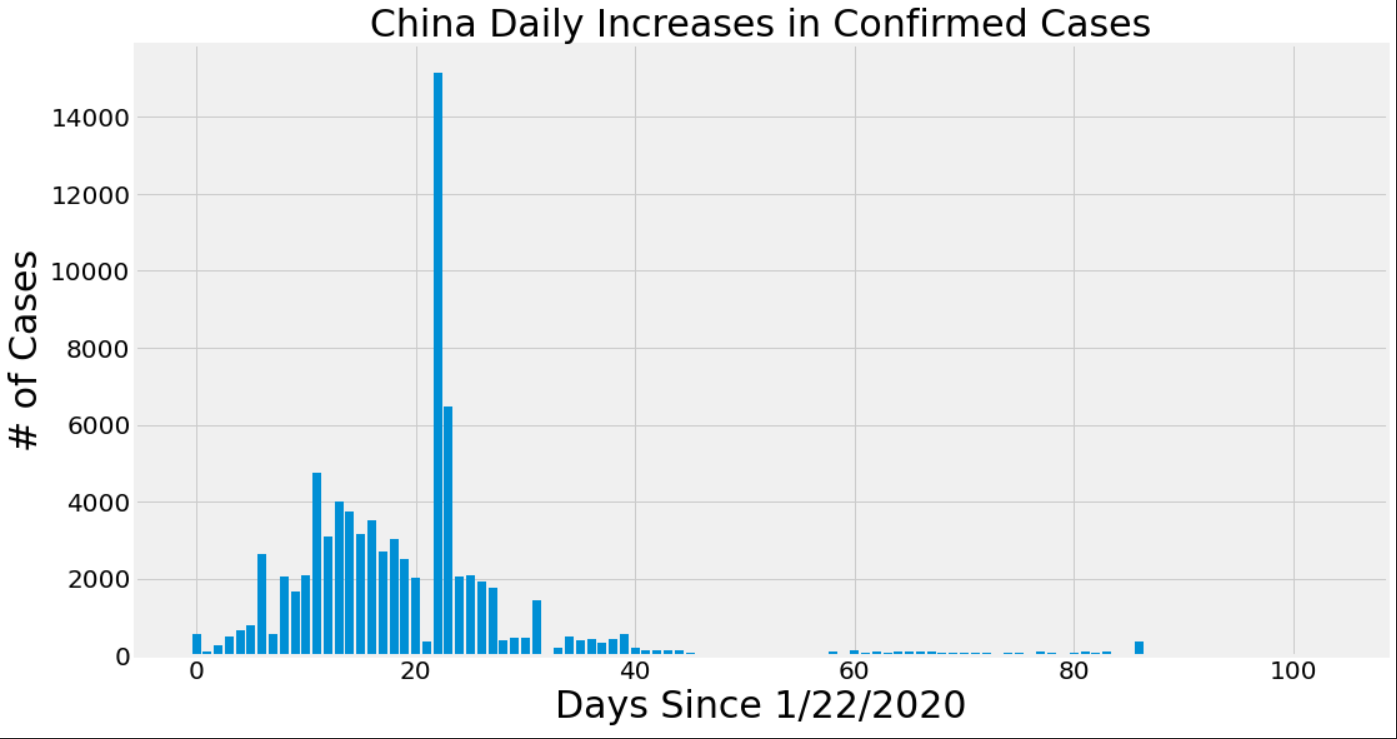
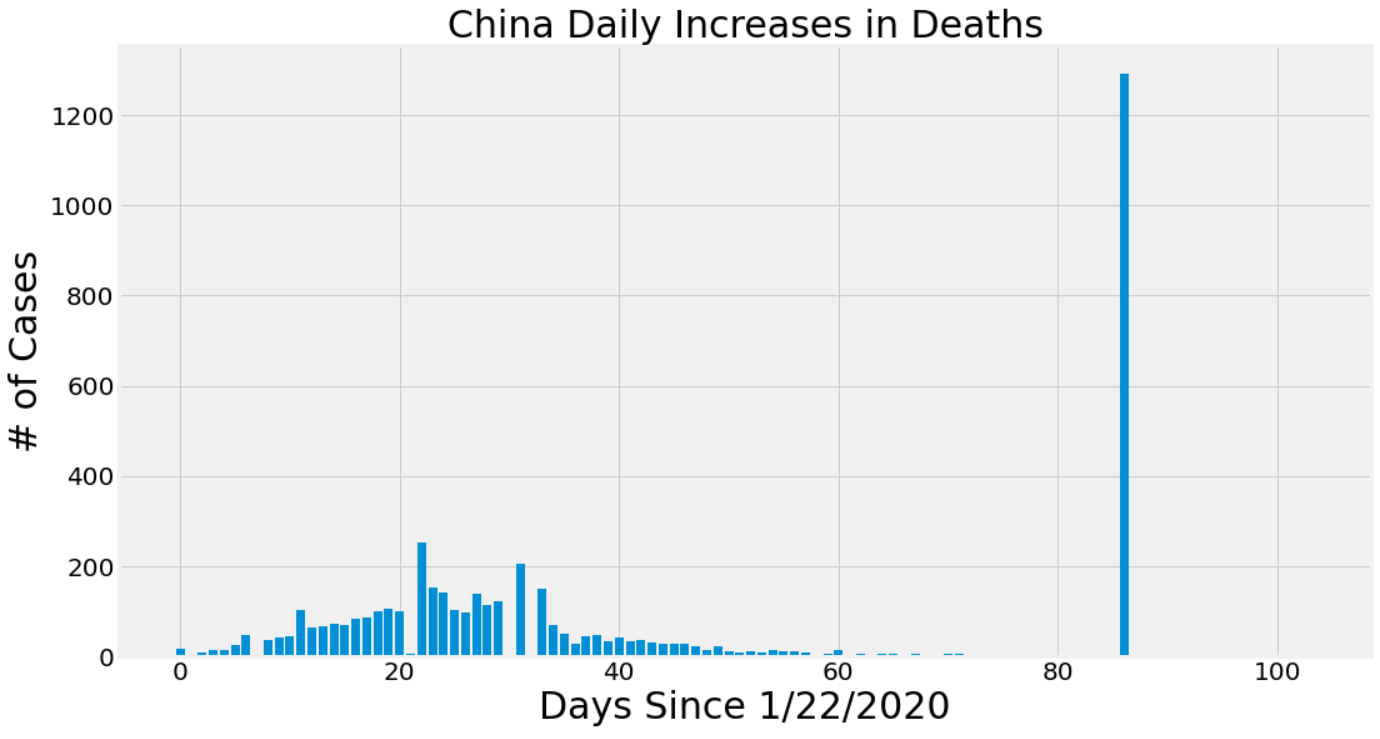
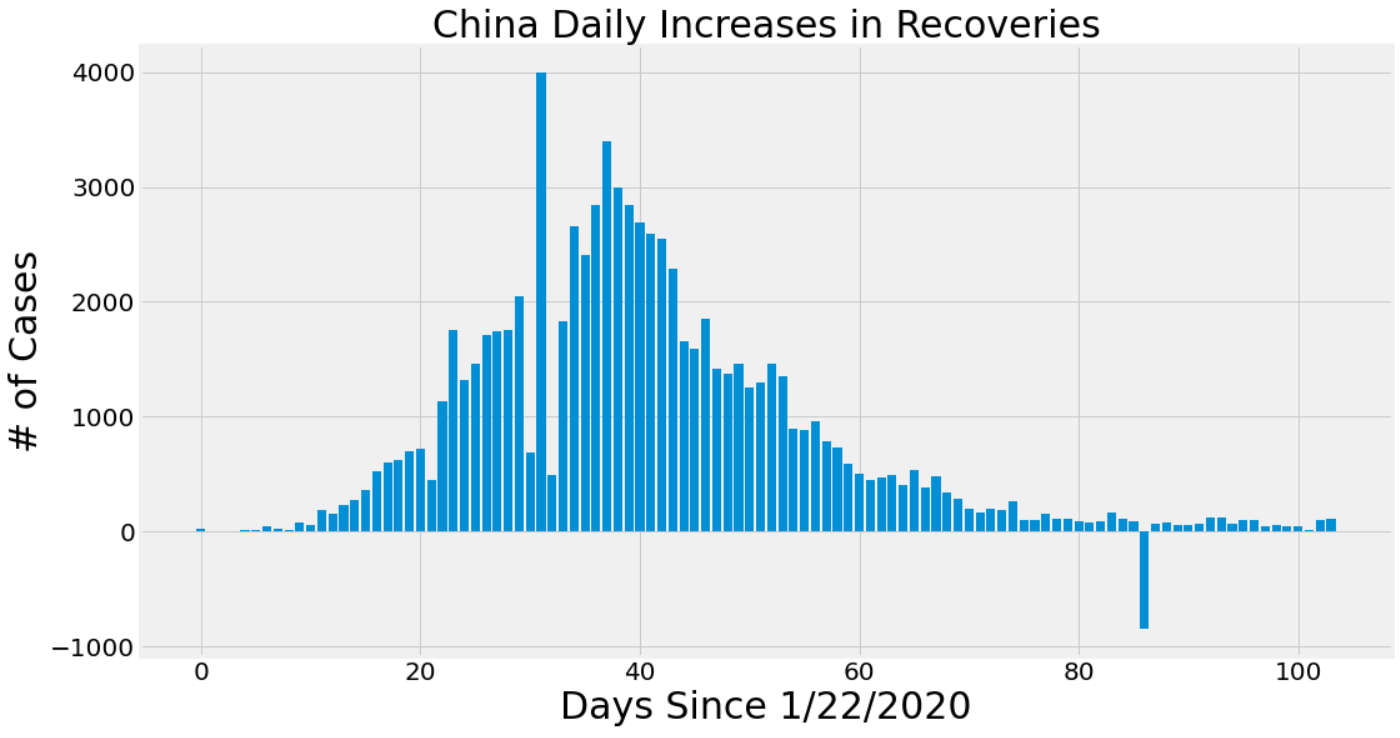
1. 亚洲

下边这幅图能比较直观的反应亚洲各国的疫情情况：

亚洲各国累计确诊数量

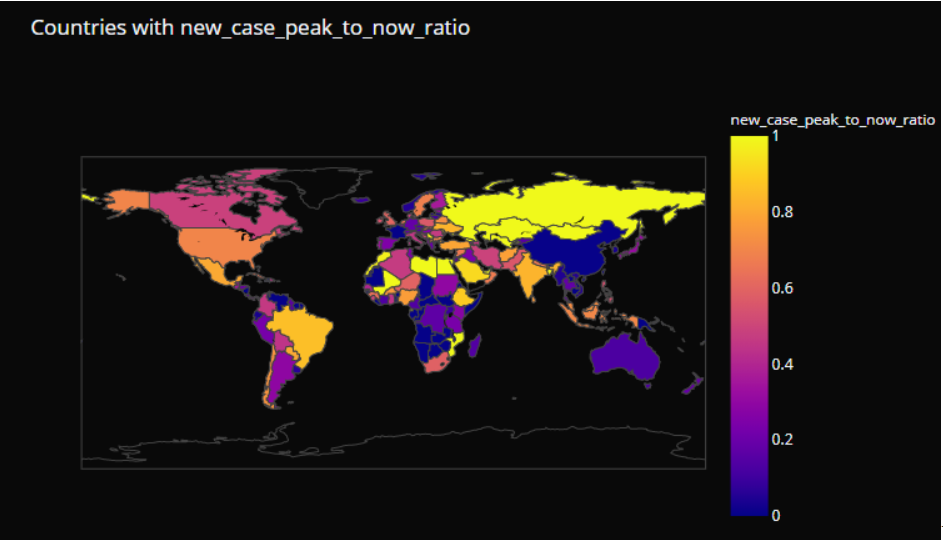
亚洲各国每日新增确诊数量

可以看出，亚洲范围内，除了最早爆发疫情但目前已基本控制住的中国外，情况

较危急的国家有伊朗和土耳其，其他国家目前局势总体还比较平稳。由于伊朗以及土耳其其他数据缺失比较严重，所以主要分析下中国的情况：

可以看到，中国虽然疫情初期爆发较快，但在党中央的正确领导以及全国人民的积极配合下，疫情已经基本控制住，新增病例和新增死亡都趋近于0，现存病例也基本清零，是目前世界各国中表现最好的国家。

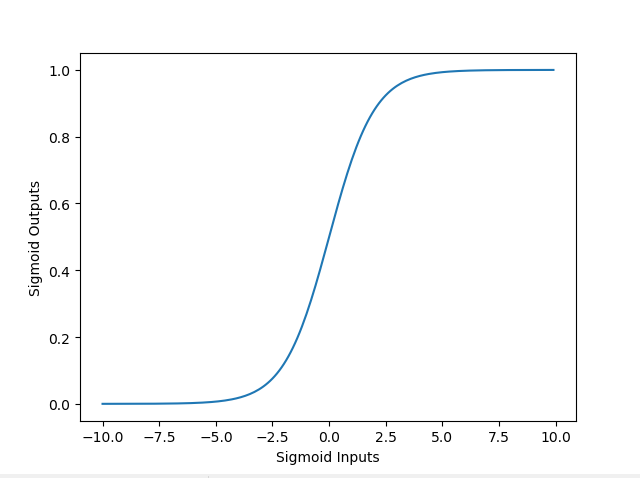
4.**哪里正在复苏**

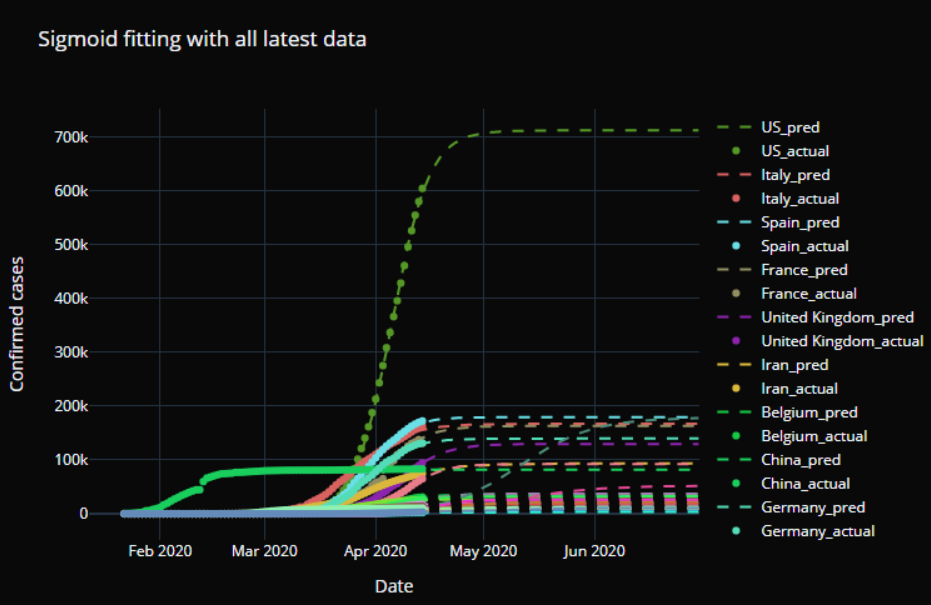
将历史单日新增确诊数的最大值与当前的每日新增确诊数取比值，越趋近于1说明仍在每日新增确诊数仍在高速增长，越趋近于0说明增长已基本停止，绘图：

可以看到，中国目前已成为世界上最安全的国家，危险系数最低，也最有可能开始复苏，而俄罗斯、蒙古、哈萨克斯坦以及部分非洲国家仍处于高速增长阶段，令人担忧。

1. **未来疫情形势预测**
2. **使用sigmoid函数拟合预测**

sigmoid函数是机器学习领域最常用的激活函数之一，其表达式为：

其中，e为自然对数的底数，约为2.718，x为自变量，其图像为

可以看到，其特点在于，当x趋近于时，p趋近于0；x趋近于时，p则趋近于1，即可以讲任何范围内的输入转化为0-1之间的输出，并且曲线过度平滑，即任何地方的梯度均不等于0或，因而有着非常好的性质，因此被用于机器学习和拟合预测的各个领域，下面我们用sigmoid函数进行拟合预测：

根据预测，美国的新冠肺炎疫情最终将感染约70万人左右（当然，这是使用4月14日的数据做出的预测，到5月19日本文写作时，目前感染人数已经达到154万人以上，远远超过了模型预测结果，这也在一定程度上证明了美国防控措施的不力），此外，根据模型预测，意大利。西班牙、伊朗等国的疫情规模将处于20万人左右（根据目前数据，西欧各国感染人数已达22-24万人，若接下来疫情规模不再扩大则预测比较准确，不过从目前形势来看这一条件成为现实的可能性不大）。

1. **基于prophet的机器学习预测**

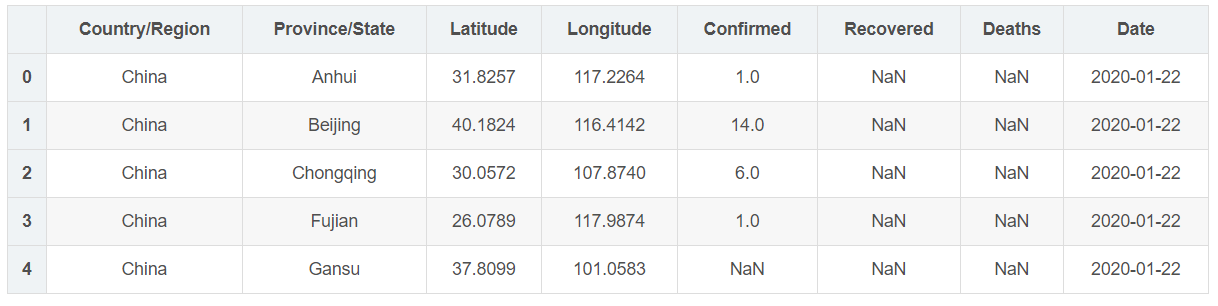
机器学习是研究怎样使用计算机模拟或实现人类学习活动的科学，是人工智能中最具

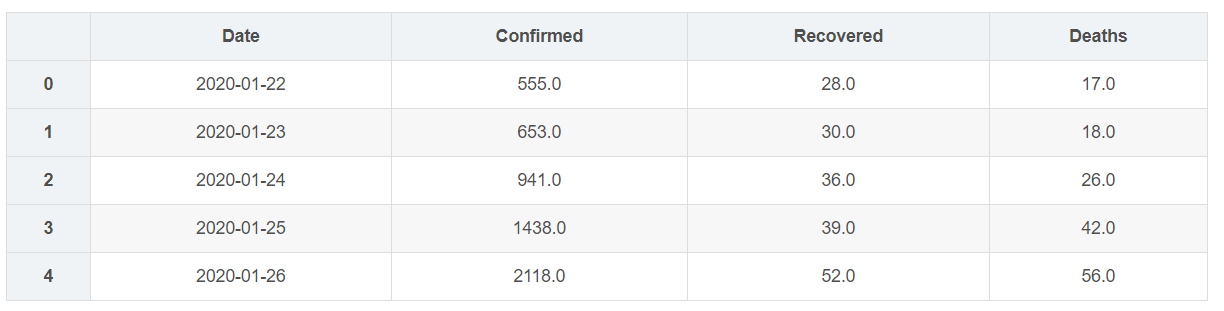
智能特征，最前沿的研究领域之一。自 20世纪80年代以来，机器学习作为实现人工智能的途径，在人工智能界引起了广泛的兴趣，特别是近十几年来，机器学习领域的研究工作发展很快，它已成为人工智能的重要课题之一。其主要算法有支持向量机svm、决策树、随机森林、贝叶斯概率以及当下最火热的神经网络（深度神经网络即狭义的人工智能）等。

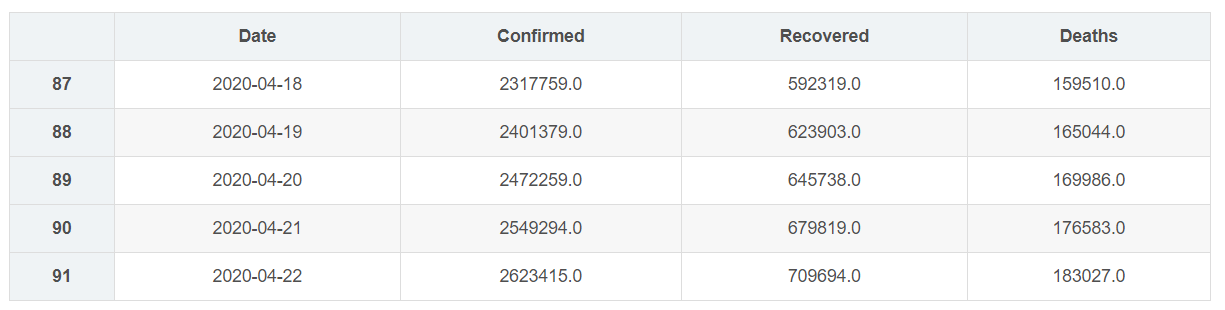
Prophet框架是Facebook于2018年3月开源的一款针对时间序列的，基于Python和R语言的数据预测工具。prophet是一个工业级应用，具有非常良好的普适性与易用性。其针对时间序列预测的主要特点为：

1. 对于历史在至少几个月（最好是一年）的每小时、每天或每周的观察
2. 强大的多次的「人类规模级」的季节性：每周的一些天和每年的一些时候
3. 事先知道的以不定期的间隔发生的重要节假日（如，双十一）
4. 合理数量的缺失的观察或大量异常
5. 历史趋势改变，比如因为产品发布或记录变化
6. 非线性增长曲线的趋势，其中有的趋势达到了自然极限或饱和

基于这些特点，prophet是针对本次疫情分析预测比较合适的框架之一，因此选用prophet框架尝试对疫情感染人数变化情况进行预测。（此部分完成于4月25日，因此采用数据截止至4月25日）

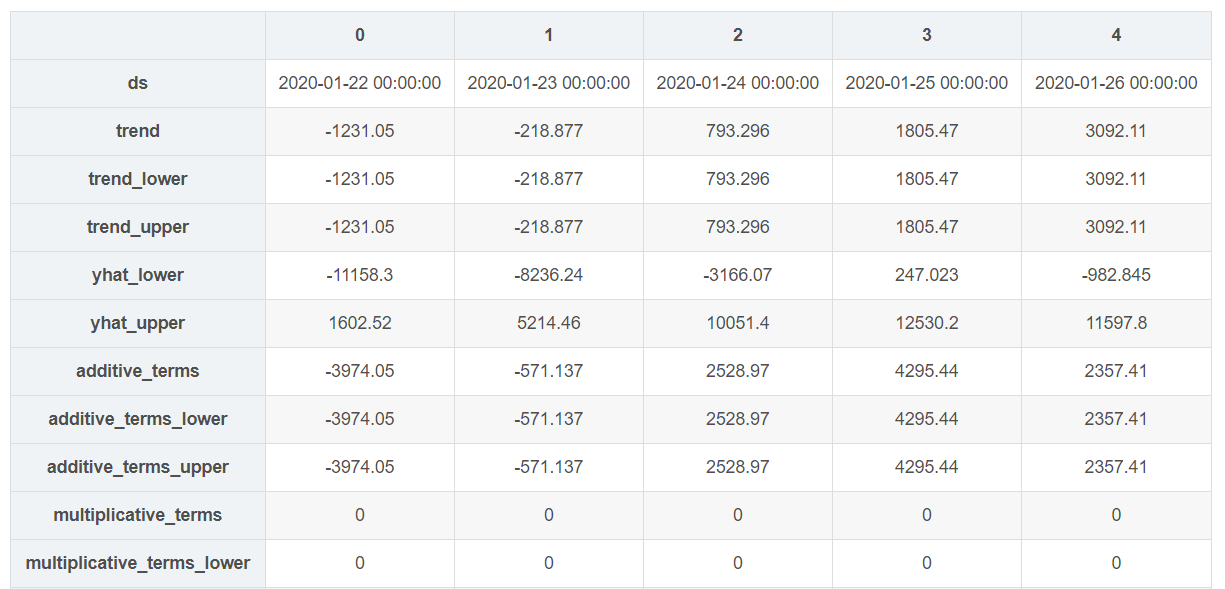
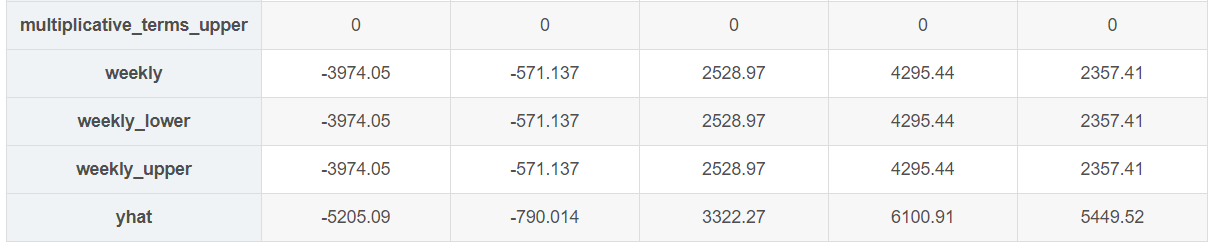
**数据读取：**

**部分数据展示：**

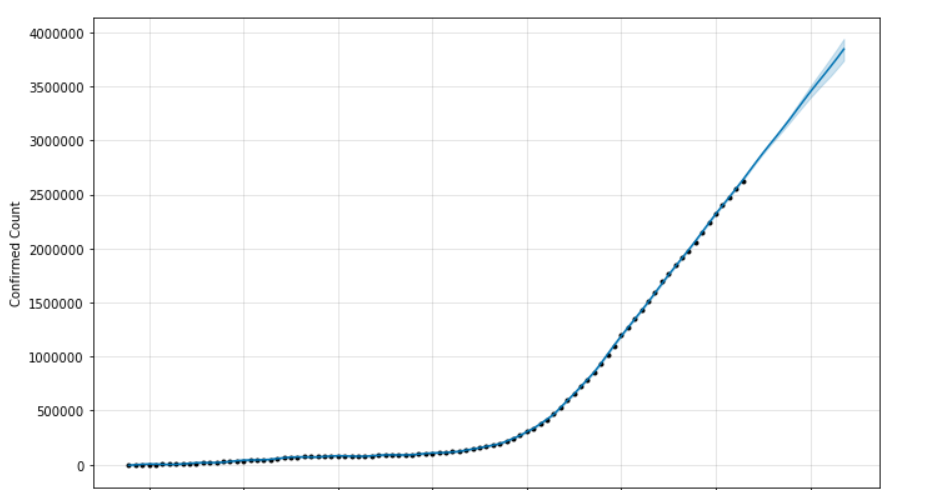


观察前边几行，在1月份时，疫情主要发生在中国，单日新增病例只有几百人左右，而到了4月底左右，每日新增确诊数均超过了10万，令人触目惊心。

**数据处理部分展示：**

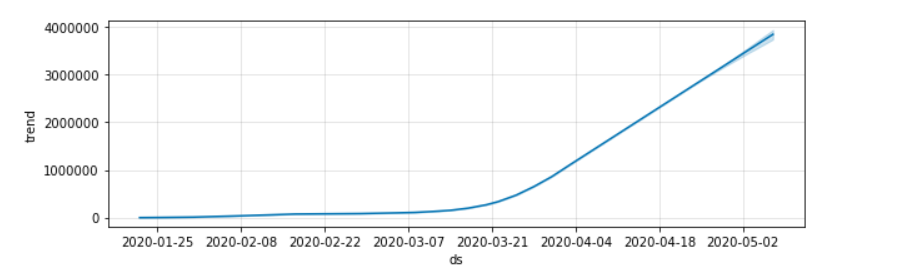




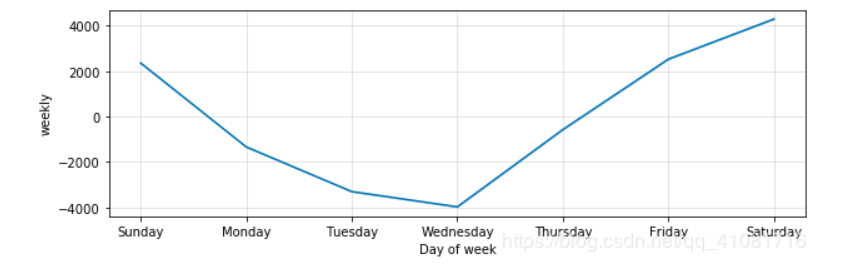
**预测结果：**

黑点是输入数据，蓝线是预测结果，可以看到，由于数据的绝对值比较大，点线近似重合了，即误差十分的小，结果比较可信。

对结果进行成分分析：

按日期序列展示：

可以看到，根据模型预测结果，到5月10日左右，新冠肺炎感染人数就将达到400万人左右，站在今天的结果看5月19日，5月19日全球新冠肺炎感染人数为4801923人，可以说，这一预测结果是比较接近的。

按星期序列展示：

这张图给出的是模型预测的未来每周周一到周日的感染人数，即按星期几的序列给出预测。**对于这个结果，我个人的观点是这个结果未必可信**，prophet之所以有这一功能，最初是基于商业数据对时间序列进行预测，比如超市营业额或者电影院客流量等，星期几不同对预测结果的影响有比较清晰的内在联系（工作日的人流量显然会少于周末）。而我个人认为在当前情况下疫情感染人数和星期几关系不大，很难说有比较靠谱的内在联系，**所以我倾向于认为这一预测结果只是模型对于输入数据机械性的拟合预测，强行建立相关性，因此可信度存疑。**从预测结果来看，周二周三的新增病例竟然是负的，似乎有点不太合常理，因此我更加倾向于模型对输入数据强行建立了相关性，预测结果可信度不高。

**误差分析：**

用sklearn库计算误差

MSE（Mean Squared Error）均方误差：



MAE （Mean absolute Error）平均绝对误差：

用numpy库计算误差

MSE（Mean Squared Error）均方误差：



MSE（Mean Squared Error）均方误差：

RMSE（Root Mean Squared Error）均方根误差：

可以看到，两个库不同的内置函数计算出的结果一致，检验了计算的准确性

1. 感想心得

这个项目既是多相流课程的大作业，同时也是一次难得的社会实践的项目经历。在这个项目过程中，一方面让我学到了许多知识，我一点点摸索plotly库的应用，逐渐画出了我想要的可视化图像效果，尤其是将数据展示在地图上，真的是花了我好大好大的精力，遇到了各种问题；从头开始学习prophet框架，从安装到使用，一点点地摸索尝试，也遇到了很多困难，有的时候一卡壳就是好几天，找了各种资料也不一定有用，总之，就和之前无数次涉足自己没接触过领域一样，遇到各种困难，然后只能自己尝试着一点点地去解决，这个过程虽然很坎坷，也很痛苦，但是不得不说，确实锻炼了我解决问题的能力，教会了我怎么去面对各种挫折。

另一方面，在新冠肺炎疫情肆虐全球的当下，这一项目也确实很具有现实意义和社会意义，习大大之前也一直强调，大学生不仅要学会在课堂里学习，更要学会扎根在祖国大地上。因此这个项目也让我更加懂得了要尽可能为用自己所学的知识为国家做力所能及的事，将自己的命运紧紧地同祖国联系在一起，因此这个项目也是一次很好的社会实践和爱国主义教育。

最后，希望疫情能够尽快平息，希望世界各地的人们都能尽快过上安全稳定而生活，希望在不远的未来（就本世纪吧，毕竟21世纪是生物的世纪），生命科学和医学能够有长足进步，助力人类战胜各种病魔，健康长寿。

1. 附源代码

已上传至github：

<https://github.com/yasuki666/covid19-analysis-and-forecast-multiple-phase-flows-project>