# COVID-19疫情扩散大数据研究报告

李牧元

(河北金融学院 保定 071000)

**摘要**：2020年COVID-19袭击了我国，获取疫情大数据分析疫情的传播并做好数据可视化帮助公众了解疫情发展情况是数据工作者的职责。本文通过爬虫获取网易疫情数据，随后利用Pandas进行数据清洗，接着利用Matplotlib和百度开源数据可视化工具Pyecharts进行数据可视化，最后利用疫情传播SIR算法模型预测疫情未来发展。

**Abstract**：In 2020, covid-19 attacked China. It is the responsibility of data workers to obtain big data to analyze the spread of the epidemic and help the public understand the development of the epidemic through data visualization.In this paper, the epidemic data of netease was obtained through the crawls, and then the data was clean with Pandas, and then the data visualization was carried out with Matplotlib and baidu open source data visualization tool Pyecharts. Finally, the SIR algorithm model of epidemic transmission was used to predict the future development of the epidemic.

关键词：COVID-19、spider、pandas、matplotlib、pyecharts、SIR模型、数据可视化

===========================================================

## 引言：

**COVID-19:**

新型冠状病毒肺炎（Corona Virus Disease 2019，COVID-19），简称“新冠肺炎”，世界卫生组织命名为“2019冠状病毒病” [1-2] ，是指2019新型冠状病毒感染导致的肺炎。2019年12月以来，湖北省武汉市部分医院陆续发现了多例有华南海鲜市场暴露史的不明原因肺炎病例，现已证实为2019新型冠状病毒感染引起的急性呼吸道传染病。

**背景：**

2019年12月，湖北省武汉市发现多起“不明原因肺炎”，随后的患者多有华南海鲜市场暴露史或家庭聚集现象，2020年1月7日21时，研究人员在患者标本中检出一种新型冠状病毒；1月10日24时，完成病原核酸检测；1月12日，世界卫生组织（WHO）将造成武汉肺炎疫情的新型冠状病毒暂命名为2019新型冠状病毒（2019-nCoV）；1月30日，WHO宣布将新型冠状病毒疫情列为国际关注的突发公共卫生事件。2月7日，国家卫健委决定将“新型冠状病毒感染的肺炎”暂命名为“新型冠状病毒肺炎”，简称“新冠肺炎”。英文名称为“Novel Coronavirus Pneumonia”，简称“NCP”； 2月11日，WHO将由新型冠状病毒引发的疾病命名为2019冠状病毒病，英文名称为“Corona Virus Disease 2019”，简称“COVID-19”；

随后COVID-19在我国全面爆发，经过全国人民的艰苦努力，截至4月30日24时，据31个省（自治区、直辖市）和新疆生产建设兵团报告，现有确诊病例599例（其中重症病例38例），累计治愈出院病例77642例，累计死亡病例4633例，累计报告确诊病例82874例，现有疑似病例9例。累计追踪到密切接触者732966人，尚在医学观察的密切接触者7761人。累计收到港澳台地区通报确诊病例1511例。其中，香港特别行政区1037例（出院846例，死亡4例），澳门特别行政区45例（出院35例），台湾地区429例（出院322例，死亡6例）。我国的疫情暂时得到了有效控制，但是面对境外输入的病例，中国人民的抗疫任务依然严峻。

**研究目的**：

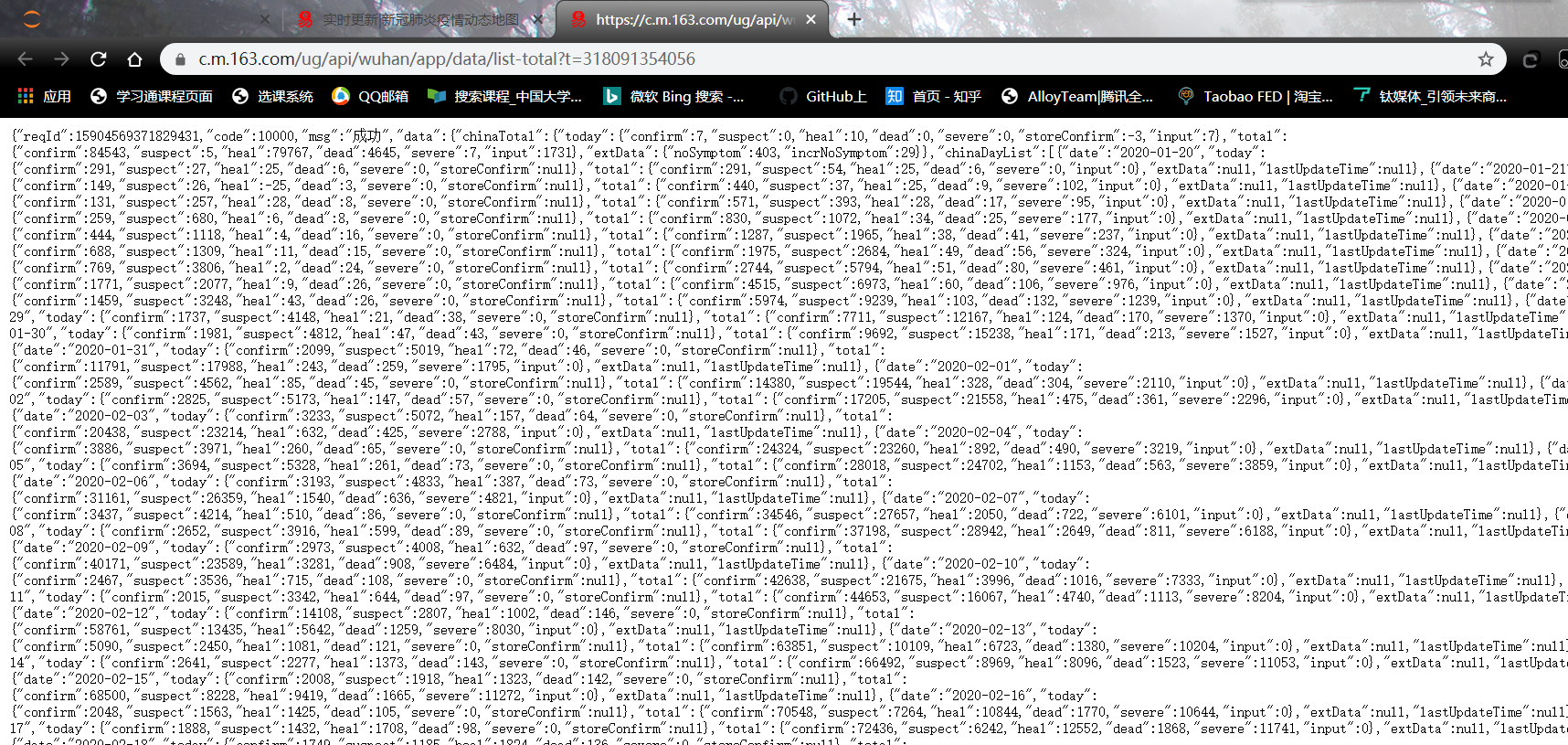
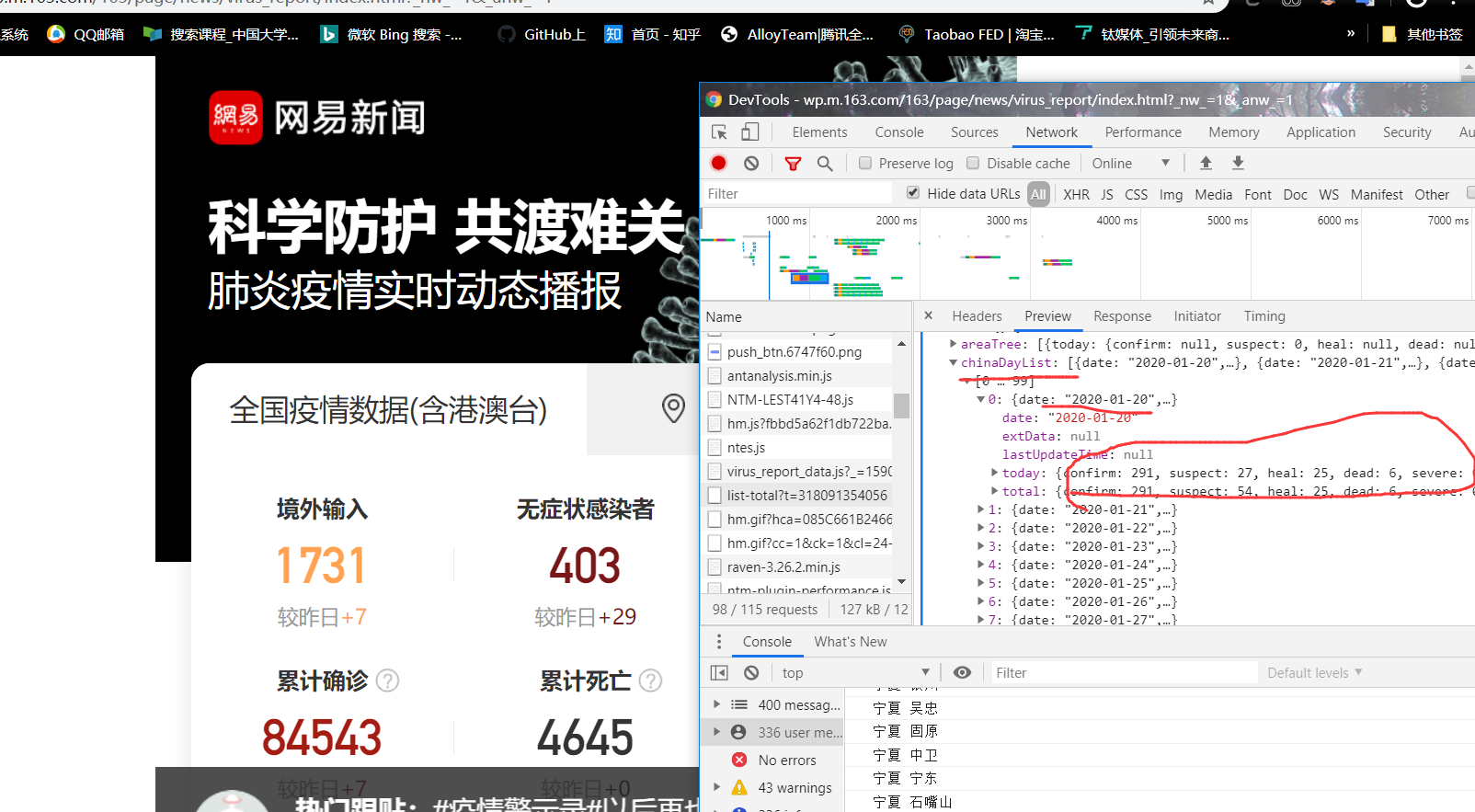
面对突如其来的疫情，勇敢的中国人民，众志成城，团结一致积极应对取得了显著的成果，作为数据工作者，对疫情数据进行有效分析帮助人民了解疫情发展情况，缓解恐慌是我们义不容辞的义务。

研究思路：

首先，确定数据源，由于疫情的传播具有实时性，本文选择网易疫情实时动态播报网作为数据源[1]。第二，进行数据清洗，获得数据中存在英文表头并且没有当日确诊数，所以要进行更改表头并从直接数据中获得间接数据。第三，通过matplotlib和echarts进行数据可视化。最后，利用SIR（疫情传染模型）对疫情未来传播进行预测。

**研究内容**：

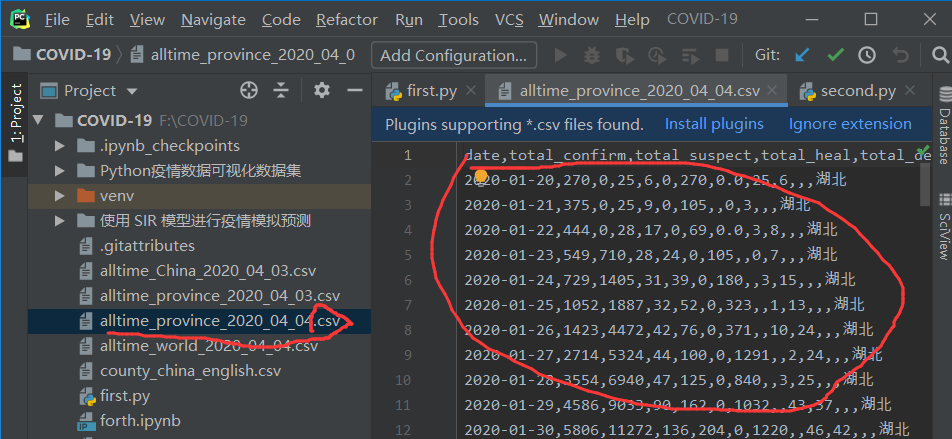
1. *爬虫：*

分析如何获得数据。

本页面的数据是动态来源于服务器的json数据文件，所以只要通过requsets请求url便可以获得数据。

1. *数据的存储：*

通过字典的键值查询，找到所需的数据并转存csv。

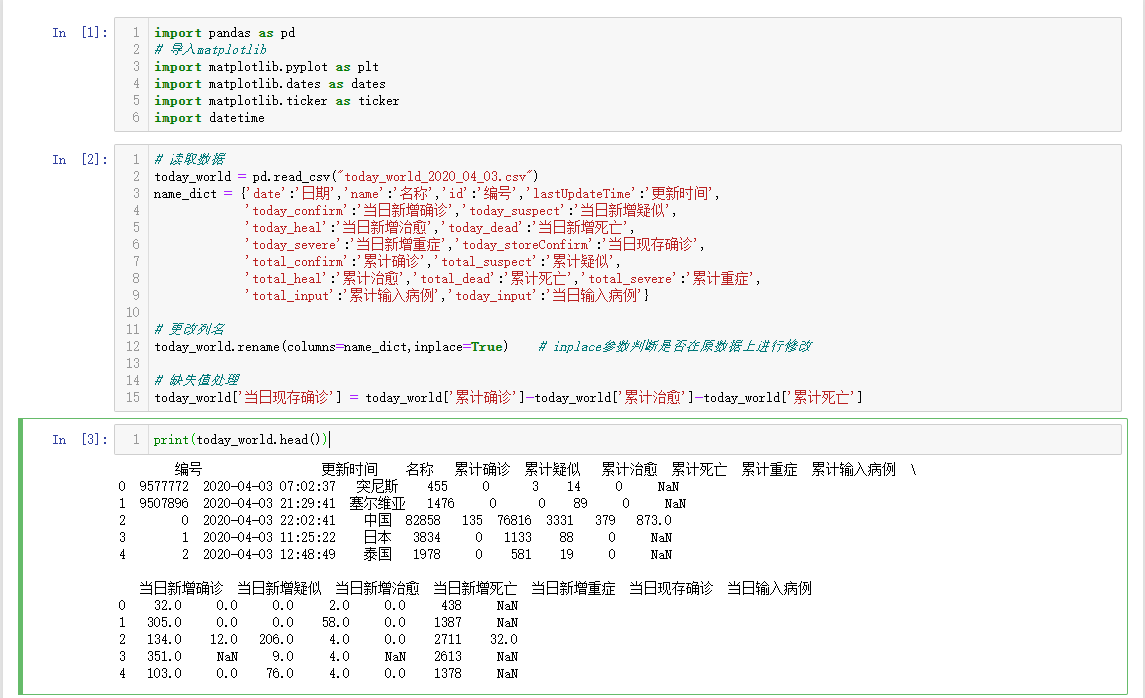


1. *利用pandas进行简单的数据清洗：*

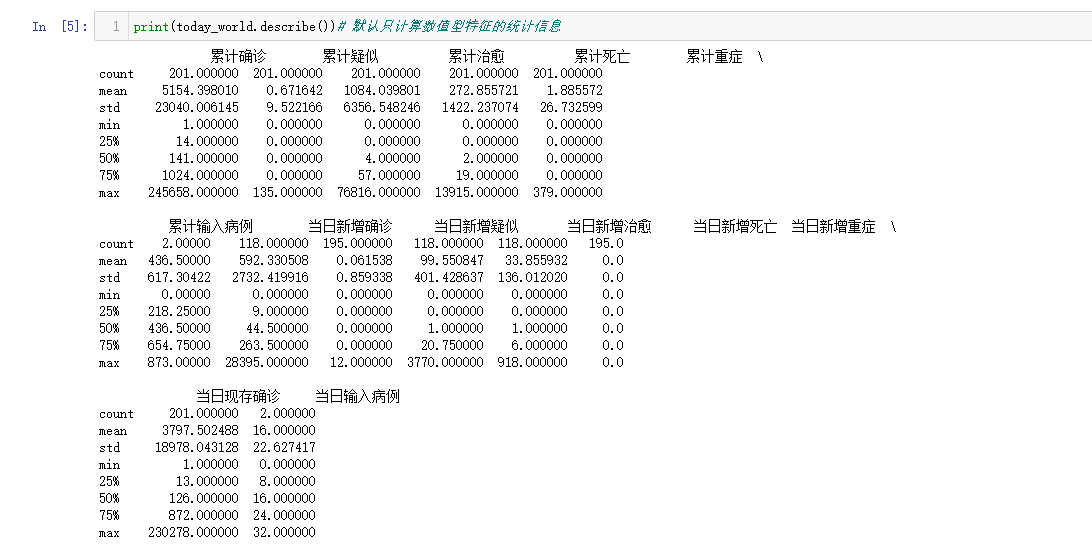
由于获得的表头全部是英文并且当日确诊数量没有明确给出，所以需要更换表头并获得当日确诊人员数量，公式为：

**当日确诊=累计确诊-累计治愈-累计死亡**

结果如下：



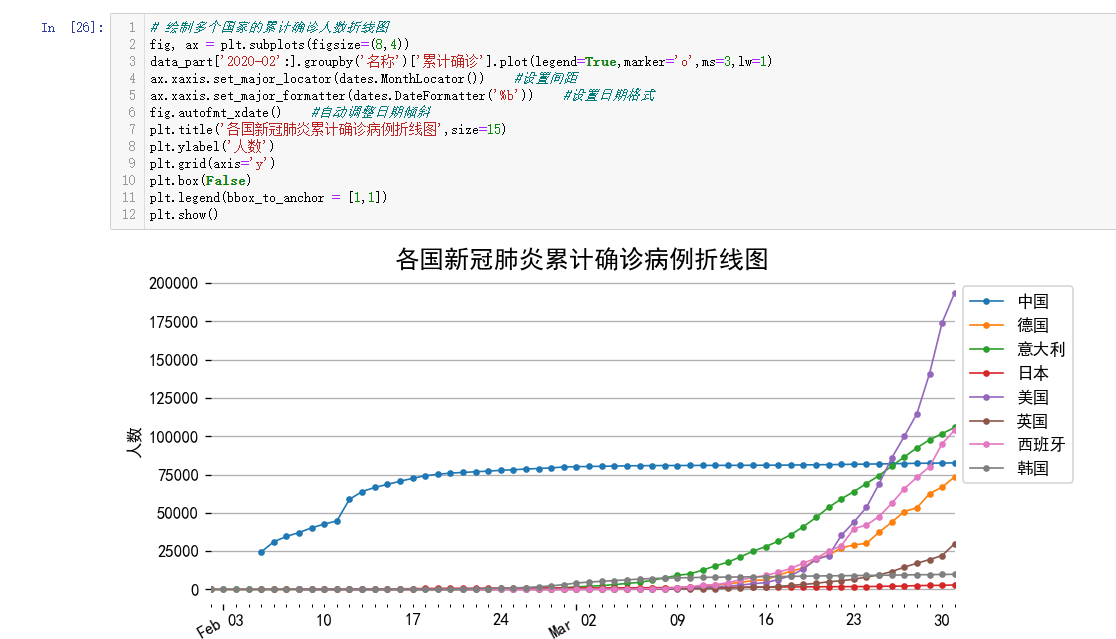
查看4月3日的世界疫情基本统计信息：



可以看到自疫情爆发以来，累计确诊人数惊人。

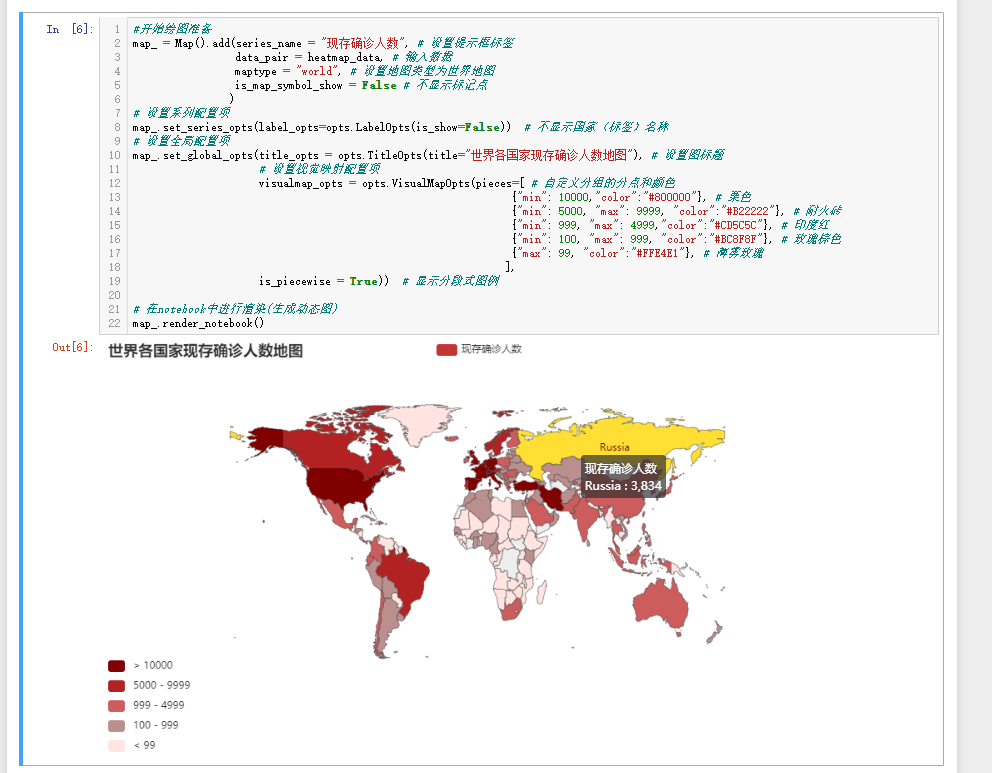
1. *用matplotlib进行简单的数据可视化分析：*

实验结果：



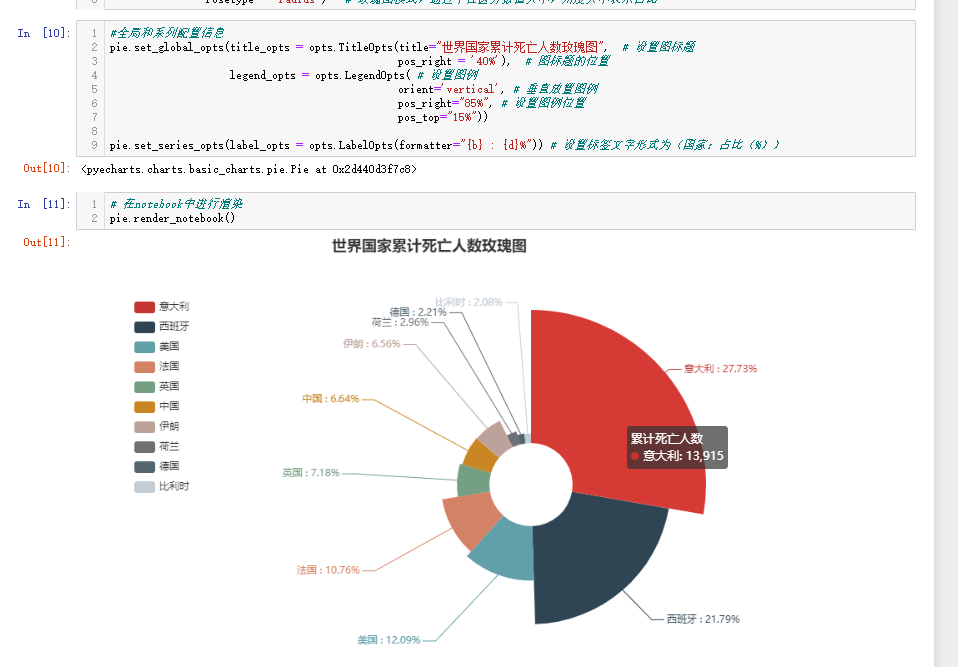
根据上表，可以看到我国的疫情以基本得到控制，而美国和一些西方国家的疫情开始迅速蔓延。

1. *用echarts进行更详细的可视化分析：*



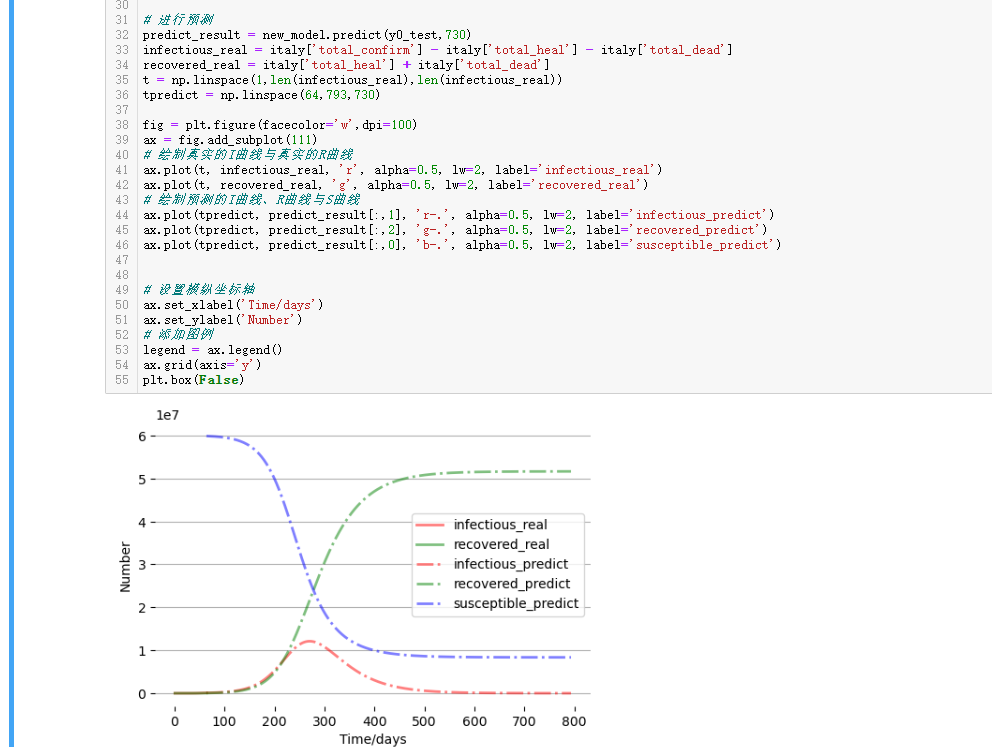
如同，可以看到4月3日美洲中部、西欧和中东的疫情情况严峻，而非洲地区和极低地区的格陵兰岛疫情相较轻，推测病毒可能耐高温和耐寒能力差。

玫瑰图：



可以看到意大利的死亡人数最高，上午查询相关资料得知，根本原因在于意大利老龄化严重并且该病毒在老年人中致死率极高，所以出现了如上图中的结果。

1. *利用SIR传染模型对意大利未来疫情进行预测：*



根据曲线结果（红色的点线），预测距离意大利真正出现疫情扩散的拐点，至少还有接近半年的时间。

## 总结：

根据以上研究结果，距离疫情的结束还有相当长的距离，相信在世界人民的共同努力下，我们能共同抵过难关，向在抗疫斗争中牺牲的人民英雄们致敬！

===========================================================

参考资料：

[1] https://wp.m.163.com/163/page/news/virus\_report/index.html?\_nw\_=1&\_anw\_=1