****

**新冠肺炎数据可视化及预测分析**

|  |  |
| --- | --- |
| **学生姓名** | **叶彤彤** |
| **学 号** | **091701113** |
| **专 业** | **信息管理与信息系统** |

**摘 要**

在新型冠状病毒感染的肺炎疫情牵动社会人心的关键时刻，本文将利用数据分析、数据挖掘、机器学习相关方法，围绕疫情态势展示、疫情走势预测进行分析，并以生动直观的形式呈现；

论文首先从全球的疫情发展现况进行分析，客观地描绘了当前整个世界的发展趋势，体现了在灾难面前，没有一个国家能够独善其身。在此基础上，再对各大洲（美国、亚洲、欧洲）进行分析，直观的展现各大洲目前的发展形态，最后使用Logistic增长模型预测美国两个月内（90天）转折点以及确诊病人数目；

数据来源：https://github.com/CSSEGISandData/COVID-19

数据范围：2020/1/22-2020/5/5

工具：anaconda（jupyter）、matplotlib、plotly等；

语言：python

注：由于数据获取方面问题，存在某一部分数据可能为空，因此可视化方面仅关注大体趋势；

**关键词:** 新型冠状病毒；数据分析；Logistic增长模型；预测分析；

**目录**

[**1. 绪论 1**](#_Toc19427)

[**1.1. 研究背景和意义 1**](#_Toc14699)

[**2. 全球新冠疫情发展趋势 1**](#_Toc11798)

[**3. 各大洲及国家新冠疫情发展趋势 3**](#_Toc18968)

[**3.1. 亚洲 3**](#_Toc29800)

[**3.2. 欧洲 4**](#_Toc6388)

[**3.3. 各国家疫情对比分析 6**](#_Toc6388)

[**4. 基于Logistic增长模型预测美国确诊数目 9**](#_Toc13423)

[**4.1. 美国疫情发展趋势 9**](#_Toc29800)

[**4.2. 美国确诊数目预测 10**](#_Toc6388)

[**5. 总结与展望 13**](#_Toc1444)

[**6. 参考文献 13**](#_Toc16195)

# 绪论

## 研究背景和意义

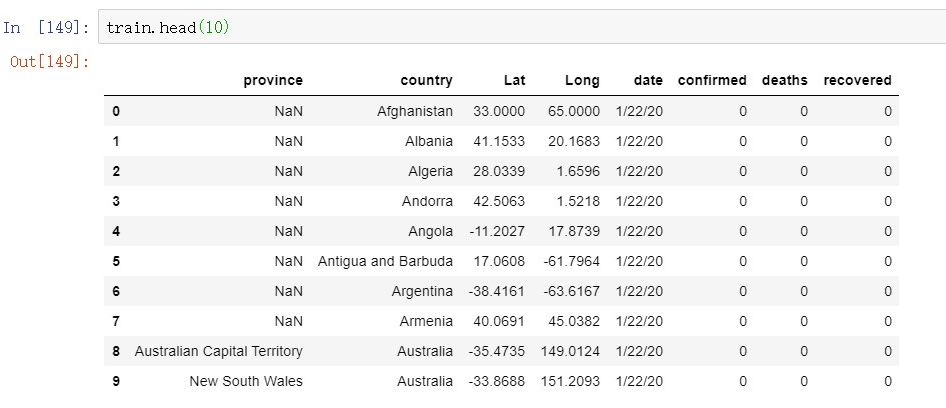
2019年12月下旬，中国武汉发生了不明原因的群发肺炎病例，引起卫生主管部门关注。12月31日，中国疾病预防控制中心派快速反应小组前往武汉。逐个排除了可能的原因包括流感、禽流感、腺病毒、严重急性呼吸系统综合症冠状病毒（SARS- CoV）和中东呼吸综合征冠状病毒（MERS-CoV）。流行病学调查指向病人感染可能与武汉市华南海鲜市场有关，1月1日，当地政府关闭武汉华南海鲜市场，并对市场进行了消毒，同时要求进行病例的主动搜索和应急监测。1月3日，中国政府向世界卫生组织（WHO）通报了疫情。

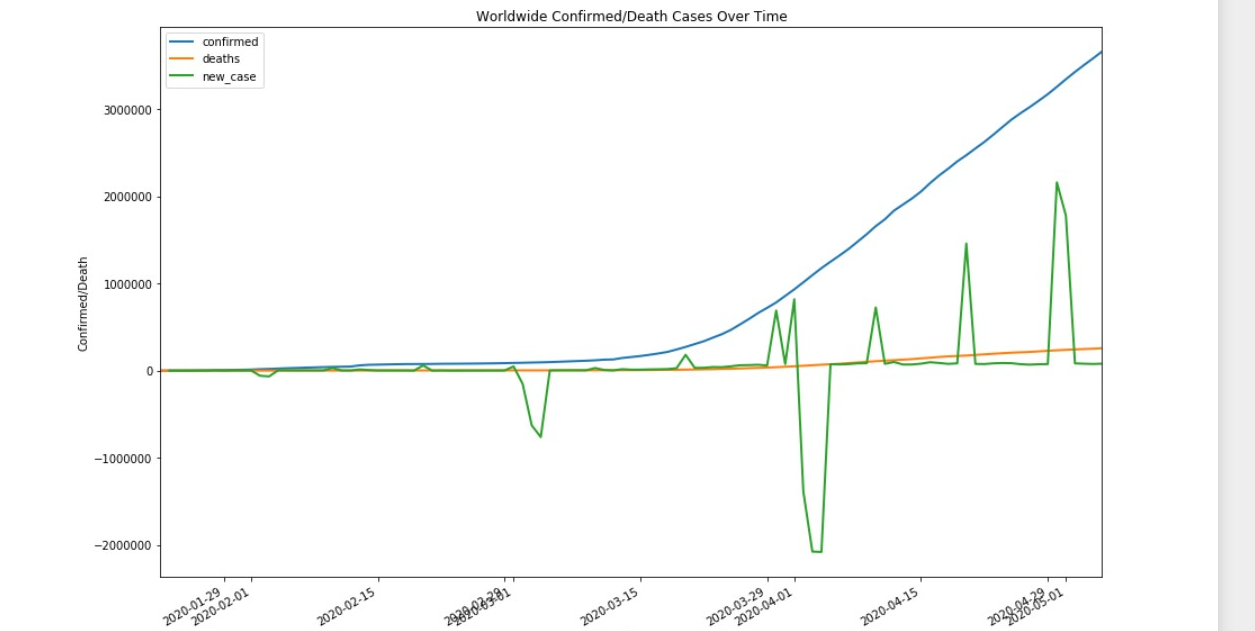
2020年1月7日，致病病原被鉴定为新型冠状病毒(2019-nCoV)，随后进行了基因序列分析，开发了检测方法。现在世界卫生组织将该病命名为COVID-19。

至今，这场新型冠状肺炎仍在全球迅速蔓延，不断地威胁着人们的生命健康，但幸运的是我们是中国人，我们有国家、有政府、有很多可爱的“逆行者”，是他们不断的给我们信心，减少我们的恐慌，在此真诚的感谢无私奉献的“逆行者”，感谢政府，感谢国家。

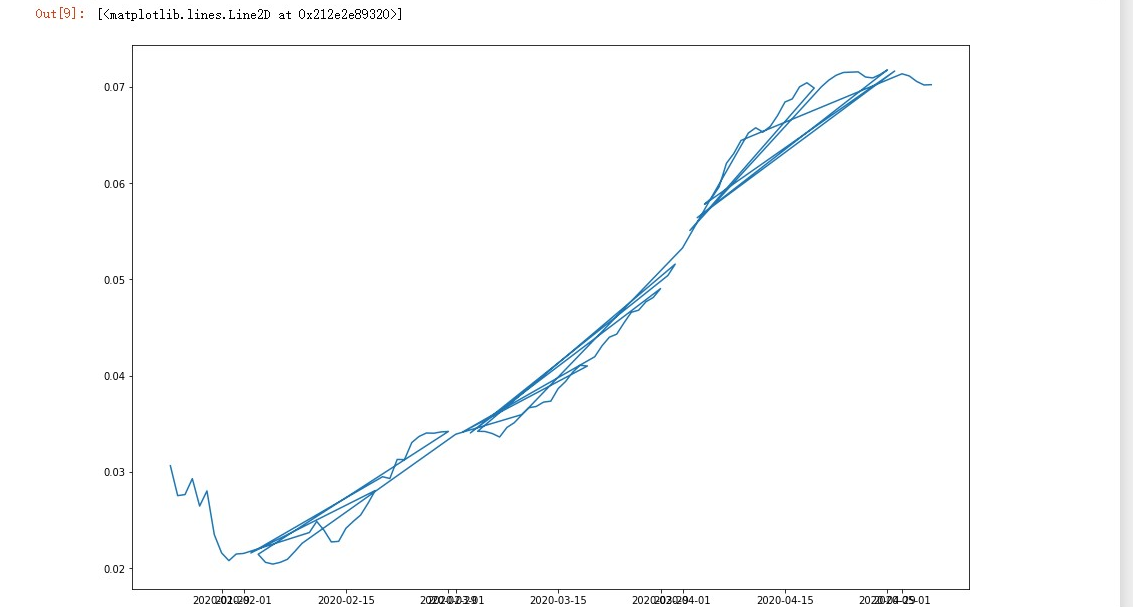
本文致力于展现最真实的全球疫情趋势，帮助自己以及阅读者对当前新冠病毒发展有更直观的认识。

# 全球新冠疫情发展趋势





（图1：全球确诊/死亡/新增病例）



（图2：全球致死率）

分析：

1. 从图1我们可以看到，世界疫情从三月中旬开始，便呈现指数级增长。在半个月内，全球确诊病例翻了一倍，死亡率也随之指数级上升。而三月中旬正好国内疫情进入攻坚后半段，形势好转，确诊人数各省逐步清零，但这时仅仅是国外疫情爆发的开始…
2. 从图2可以表明，但疫情仅在国内大规模爆发时，死亡率在3%附近；而截至于2020/5/5，该病毒的死亡率率已经接近7%，而这也由于国外对于疫情的不重视以及采取错误的决策导致；

# 各大洲（及国家）新冠疫情发展趋势

## 亚洲



（图3：截至到2020/5/5亚洲确诊情况）

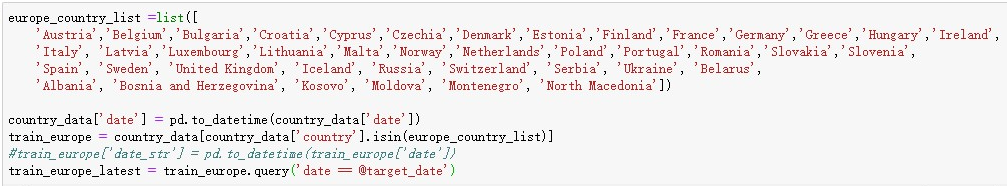


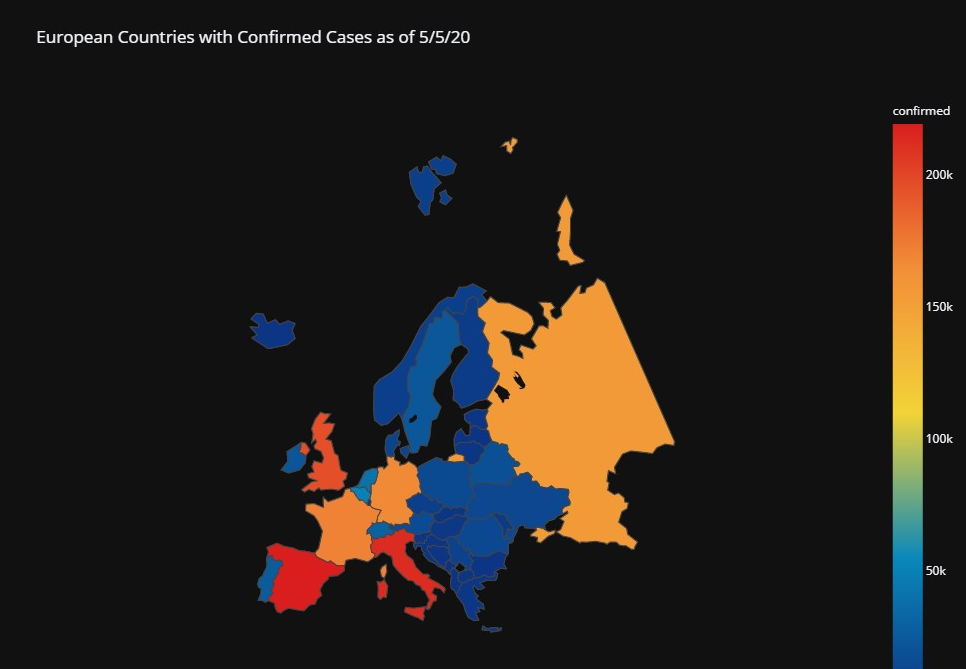
（图4：截至到2020/5/5部分国家确诊情况）

分析：

1. 从图3可以看出，在亚洲大陆上中国、印度、伊朗、土耳其累计确诊人数居高，均超过了5万人。从图4中，中国的确诊人数早已趋于稳定，而土耳其、伊朗等其他国家仍处于上升趋势。
2. 结合近期新闻，土耳其和伊朗在疫情前都没有做好充分的政策和认识准备，在国内疫情日益严重时，才逐渐加码政策，最近也看到了胜利的曙光。而印度正处在疫情扩大阶段，由于检测量不足，其数据不足以反映疫情的规模和进展态势，而印度目前采取的社交隔离政策较日本更严厉。

## 欧洲





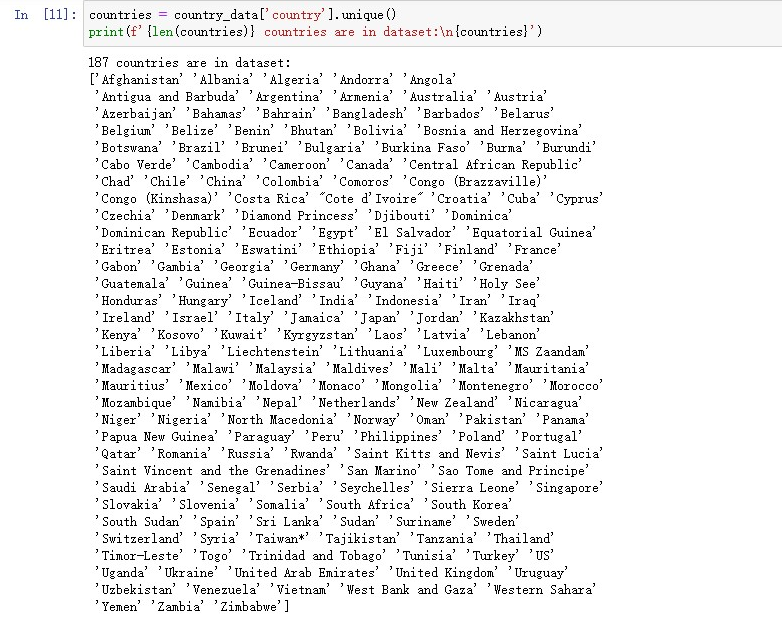
（图5：截至到2020/5/5欧洲确诊情况）

分析：

图5可以看出，意大利、西班牙以及英国是欧洲内部较严重的国家，均达到了20万左右的确诊人数。与亚洲疫情严重的国家相比，都不能处在同一个向量级了..我认为这与国家面对疫情时所采取的政策、本身医疗系统以及民众对待病情的态度有着密不可分的关系；

1. 比如英国在面对疫情时采取的“群体免疫政策”，企图通过大部分人在隐匿性感染后无症状或仅有轻微症状，从而在人群中获得普遍免疫，以控制疫情。事实证明这种政策造成了英国疫情的大爆发；
2. 由于不同国家的信仰习惯不同，社交习惯不同，导致在流行传染病面前，许许多多的外国人仍在坚持不戴口罩以及聚众玩闹，这是让我们经历过疫情大爆发的人最最揪心；
3. 截至目前，德国新冠病毒感染人数为17万人，死亡率远低于英国、意大利等疫情重灾区。这与德国有着欧洲最庞大的医疗体系是密切相关的，与此同时，他们的检测数量也是早早领先于全欧洲的。

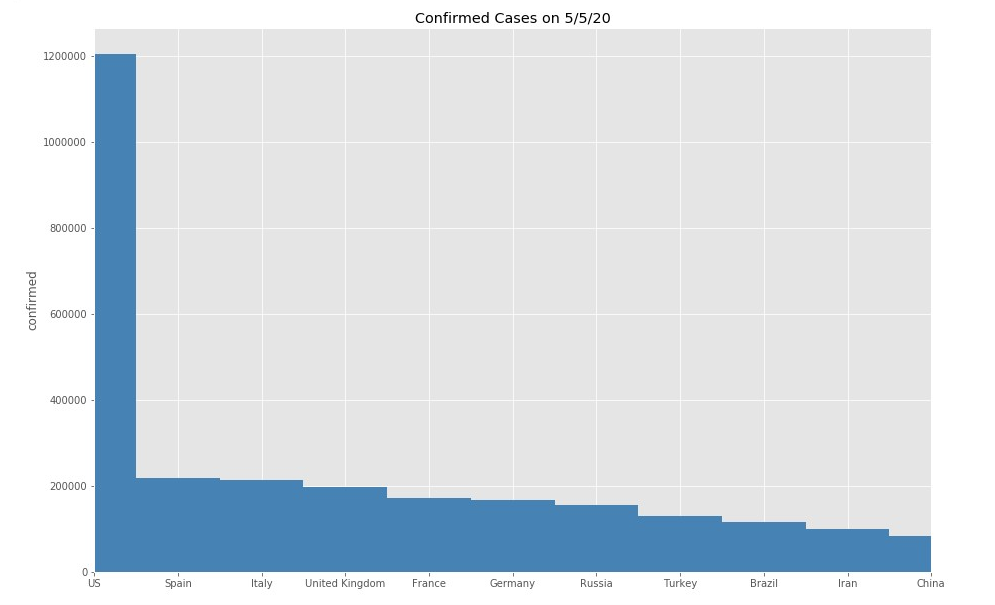
## 各国家疫情对比分析



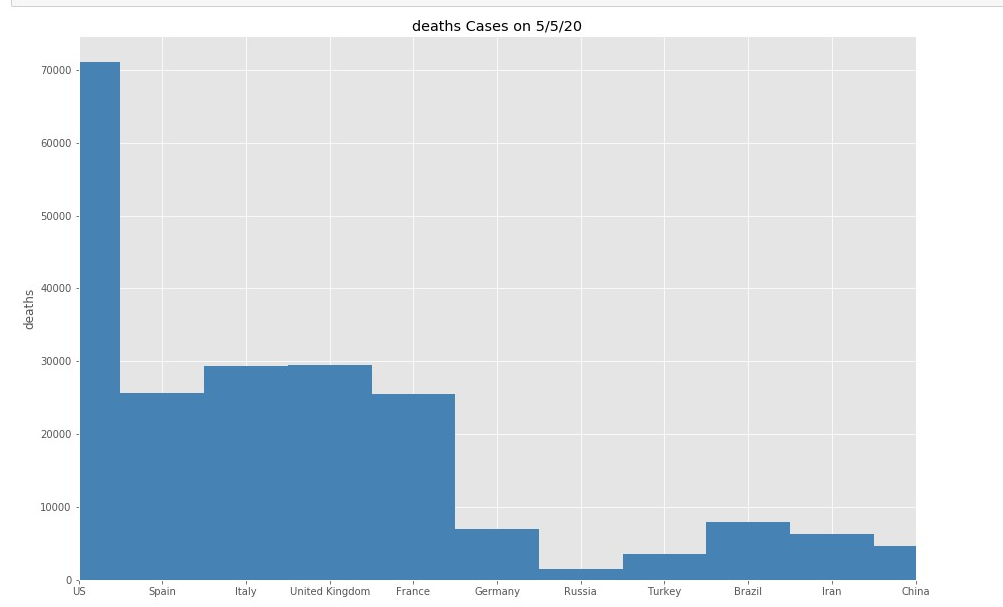
（图6：数据集中所包含的国家共187个）



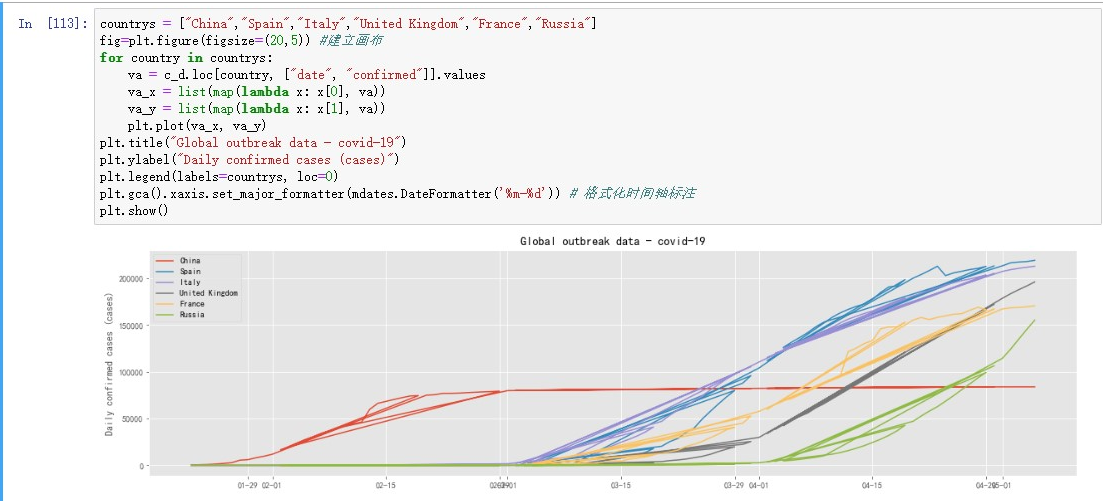
（图7：截至2020/5/5疫情确诊人数排名前十一的国家）



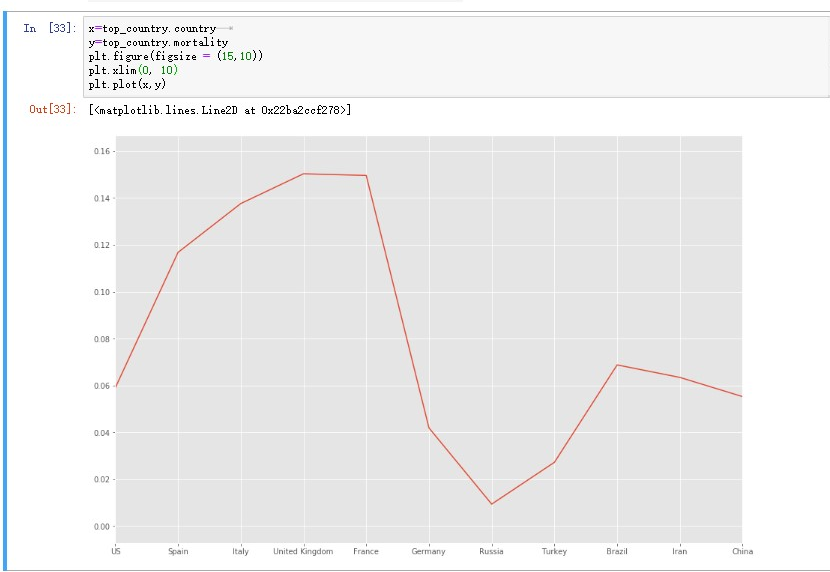
（图8：全球范围确诊前十一国家）



（图9：图8国家对应的死亡人数）



（图10：代表性国家（除美国）当前疫情发展趋势）



（图11：截至2020/5/5各国家死亡率）

分析：结合图8、9、10、11可以得出

1. 美国虽然是目前确诊人数最多的国家，但是其死亡率远低于英国、法国、意大利；截至到目前为止，俄罗斯、德国、土耳其国内都保持着较低死亡率；
2. 对于死亡率排名前三的国家，深究其因，除了前面提到的医疗体系不够健全，普通百姓的防范意识不强。还存在另一个重要的原因，这几个国家的人口老龄化十分严重。众所周知，在疫情爆发之初，我国的专家就说过，此次的新冠病毒对年轻人还是很友好的，治愈率很高。但是对于年老体迈的老人来说，是不太友好的。特别是那些身体有其他隐疾的老人，是十分危险的，这就不难理解这三个国家确诊人数比不上美国，但死亡率为什么突然之间成为了世界前三；
3. 在本次数据中，更值得注意的是俄罗斯。它确诊人数高，但是其死亡率却明显的低于其他十个国家；

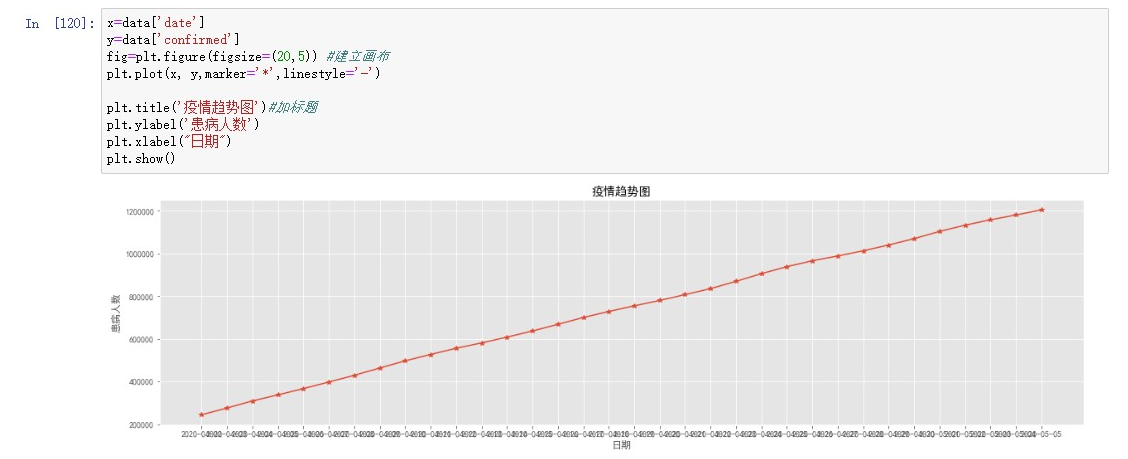
从医疗系统条件分析，依据知名医学期刊《柳叶刀》(Lancet)发布了最新一份《全球195个国家和地区的医疗质量与可及性排名》，俄罗斯排名第58位，排名并不靠前（其中意大利排名第9，德国排名第18，西班牙排名第19，法国排名第20，美国排名第29），由此可以看出俄罗斯在这场全球新冠肺炎战役中，医疗条件并不占优势；

从俄罗斯新冠肺炎应对政策，尽管俄罗斯宣布在全国范围内实施“封锁”措施，但其力度比起中国不只是弱了一点，可中国的新冠肺炎患者死亡率是5.5%，是俄罗斯死亡率的6倍。很多国家和地区也采取了与俄罗斯差不多的防控措施，可见俄罗斯新冠肺炎应对政策也不见优势。

排除这些可能后，可以注意到，俄罗斯向来被称为“战斗民族”，在俄罗斯特别流行的户外健身，男女老少在家里或者街上只要找到空地就能动起来，利用随手可得的器械锻炼肌肉力量。在这样的氛围下，他们强壮的不仅是体魄，还有精神力量。有着如此的高身体整体素质，强大的免疫力，也能理解低死亡率的原因；

# 基于Logistic模型预测美国确诊数目

## 美国疫情发展趋势



该数据日期从2020/4/01-2020/5/5

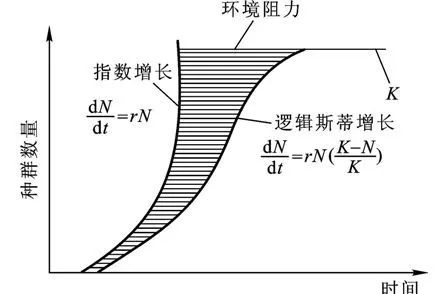
（图12：截至2020/5/5美国疫情发展趋势）

分析：

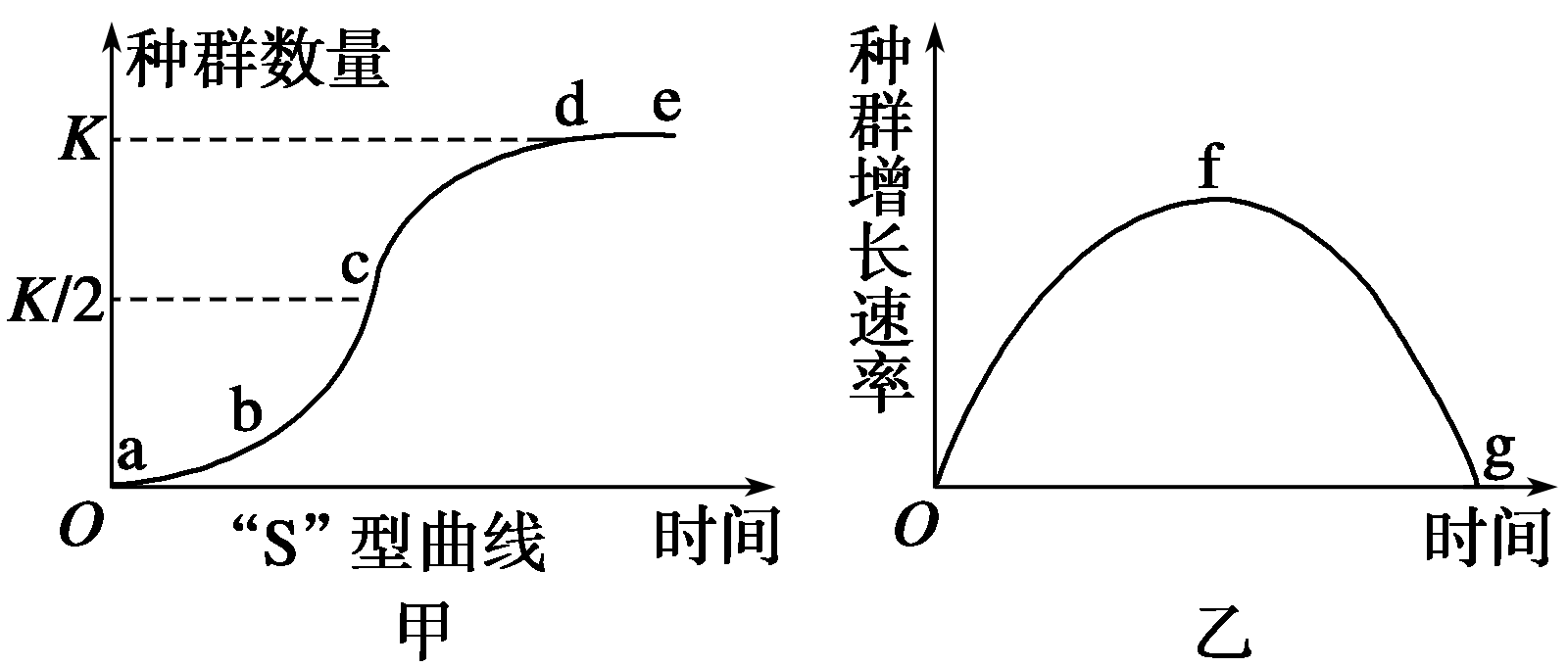
1. 图12可以看出，截至到2020/5/5，美国疫情仍在上升趋势，疫情拐点还未到来；

## 美国确诊数目预测

### Logistic模型介绍（逻辑斯蒂增长模型）



logistic增长的曲线也称为s型曲线。下图左图为曲线数量，右图为增长速率。



逻辑斯蒂模型，又叫阻滞增长模型

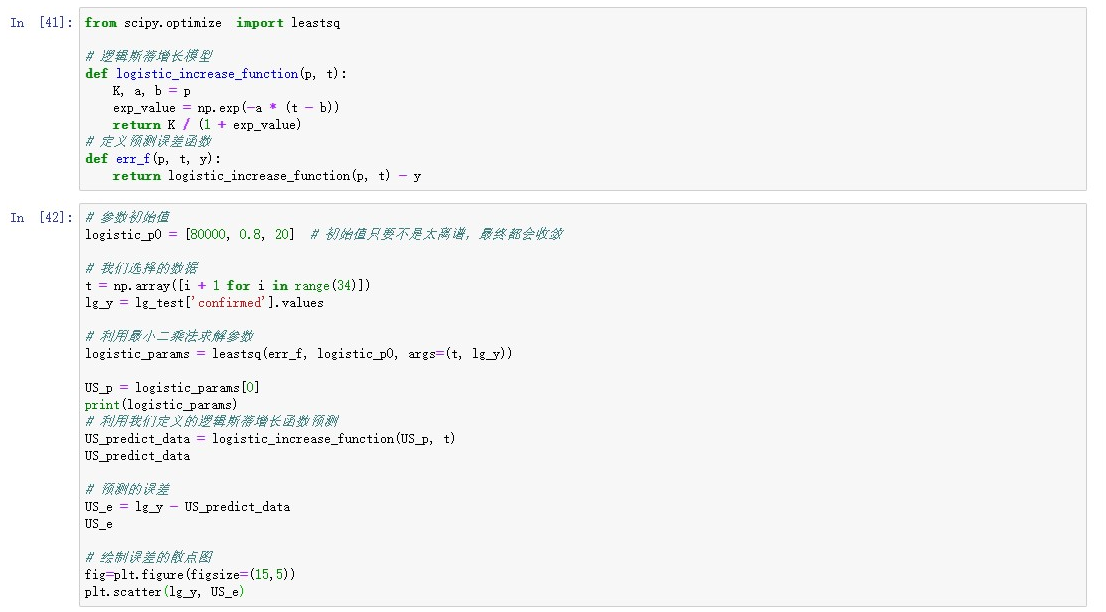
逻辑斯蒂曲线通常分为5个时期：

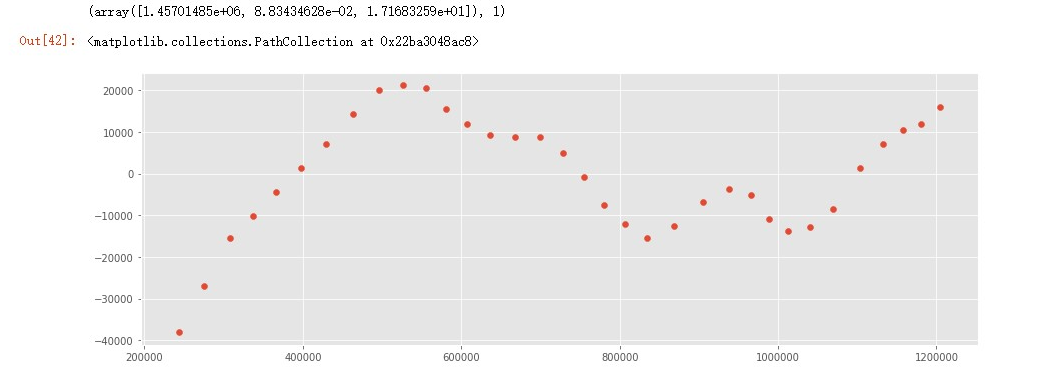
* 1. 开始期，由于种群个体数很少，密度增长缓慢，又称潜伏期。
  2. 加速期，随个体数增加，密度增长加快。
  3. 转折期，当个体数达到饱和密度一半（K/2),密度增长最快。
  4. 减速期，个体数超过密度一半（K/2)后，增长变慢。
  5. 饱和期，种群个体数达到K值而饱和。

根据历史经验，2003年非典患者预测，部分学者利用逻辑斯蒂增长模型进行预测，并且准确率很高，所以我也尝试利用逻辑斯蒂增长模型进行新型冠状病毒患者数量，逻辑斯蒂增长模型具体为：

（其中K为环境最大容量，P0为初始容量，r为增长速率，r越大则增长越快（即更快的逼近上限）。）

### 预测美国确诊数目

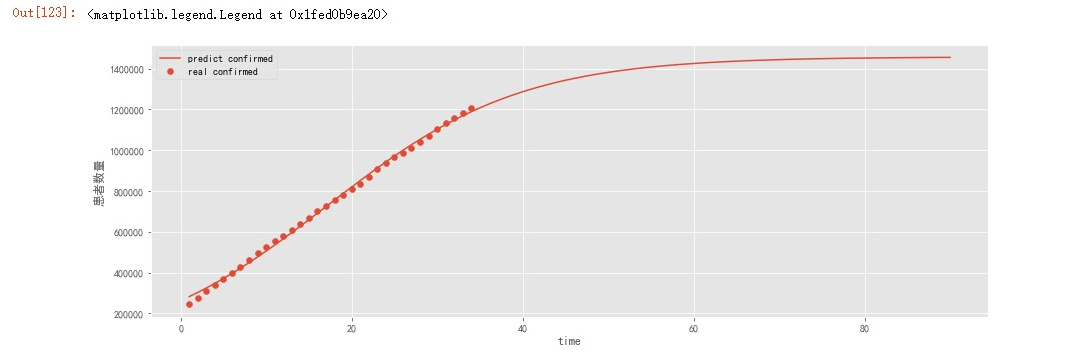




（图13：绘制散点误差图）







（图14：预测近未来90天内的疫情人数）

分析：

1. 模型预测的结果是近90天内美国新冠肺炎最终的患者数量是145万人左右，拐点出现在第80天，即6月19日左右。
2. 模型拟合度较高，但结合实际仍存在偏差；模型预测5.6号-5.12号确诊人数在120万-130万之间浮动，但实际数据中美国在此期间早已突破130万确诊人数，在5.13号已经突破140万确诊人数（模型预测为131万）。因此后续还需要对模型进行进一步的调参以完善预测数据；

# 总结与展望

本文仅通过个人的资料收集，结合决策支持系统课程所学的数据分析、数据挖掘以及机器学习等方面内容，对目前牵动全球人心的新冠肺炎疫情进行了基础性的分析和预测；总体上存在许多不足，可视化分析以及预测拟合程度仍需提高；

但通过这次的分析，不仅增加了我对数据分析及机器学习方面的知识，也让我对目前的疫情发展有了更清晰的认识。虽然我们国内的疫情趋势正在转好，但也不能掉以轻心，这场疫情攻坚战还有很长的一段路要走，但相信只要我们齐心协力，一定能迎来属于全世界人民的春天！

# 参考文献

[1] 《柳叶刀》(Lancet).《全球195个国家和地区的医疗质量与可及性排名》2019

[2] 王旭艳，喻勇，胡樱，宇传华 . 《基于指数平滑模型的湖北省新冠肺炎疫情预测分析》.公共卫生与预防医学,2020

[3] 伊赫亚,李川,王芃,崔增伟,夏建国.《全球新型冠状病毒肺炎疫情与早期防控对策》. 预防医学,2020

[4] 博主「风度78」.新冠肺炎的可视化和预测分析