|  |  |
| --- | --- |
| 学号 | 10212818103 |



课 程 设 计

课程名称 python程序设计课程设计

|  |  |
| --- | --- |
| 题 目 | 数据分析的学习与实践—地震数据分析 |
| 专 业 | 软件工程 |
| 班 级 | 软件1181 |
| 姓 名 | 饶天成 |
| 成 绩 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 指 导 老 师 | 徐勇 |

2021 年 1 月 11 日至 2021 年 1 月 22 日

武汉华夏理工学院信息工程学院

**课 程 设 计 任 务 书**

课程名称：python程序设计课程设计 指导教师：徐勇

班级名称： 软件1181-1182 开课院、系：计算机与网络工程系

**一、课程设计目的与任务**

“python程序设计课程设计”是一个综合性的学习实践型实验教学环节，将在“python程序设计”课程的授课基础上，对python的基础语法、python 对文件的操作、python 对数据的操作、数据处理基础、数据可视化等若干个知识点进行综合运用。

python是一种具有天然开源基因的编程语言，了解开源社区和广泛的使用开源工具，也是Python学习实践的重要环节。因此在本次课程设计中，要求掌握jupyter notebook、Git等常用工具，以及github等重要开源社区的使用。

**二、课程设计的内容与基本要求**

数据分析的基本技术和方法在“python程序设计”课程中已经进行完整的讲授，由于数据分析本身是基于业务场景的，因此本次课程设计环节更加偏重于实际的业务场景的实践。通过对近期互联网热点的调查，准备了15个具体的业务场景，用于本次课程设计的具体任务场景。具体包括如下：

1. 北上广深租房状况分析；
2. 蔡某坤粉丝数及转发数据真假状况分析；
3. 地震的数据分析；
4. 英文名字的数据分析；
5. 外籍英文老师收入虚高情况数据分析；
6. 我国城市空气污染和烟花燃放的关系分析；
7. 针对996工作，程序员群体的看法的分析；
8. 吴某凡微博热点的分析；
9. 节假日长假景点人满为患的数据分析；
10. 针对荔枝的品种、销售地等维度，进行价格数据分析；
11. 分析芒果TV《我是大侦探》的观众评论数据；
12. 针对当前儿科医生的缺乏，对相关数据进行分析；
13. 著名网游《绝地求生》的数据分析；
14. 实习岗位状况的数据分析；
15. 电影《流浪地球》的观众评价的数据分析。

每三个学生组建一个课程设计小组，最后的任务输出包括代码每小组一份、课程设计报告每人一份、答辩ppt每小组一份，并进行课程设计成果答辩。小组成员均参与前述工作，但是每个人的侧重点不同。

每个课程设计小组可以从上述15个场景中选取一个作为课程设计的选题，选题中提供了待分析的数据，和现有的分析方法。各小组，通过学习和实践现有的分析方法，理解实战分析的思维过程并锻炼实际动手能力，再此基础上可以扩展更多维度的分析和数据展现形式。

每个班每个选题最多只能被两个小组选中，先选先得。同时，如果各小组发现更有意思的场景，并能够获取到相关待分析的数据，也可以申请作为选题方向。

本次课程设计的目标是培养学生的团队协作能力、对python知识点的综合运用、对实际场景的理解和适应能力、针对答辩的表达能力等。注重过程，期待成果，但不强求结果的尽善尽美。

**三、学时分配进度安排**

|  |  |  |
| --- | --- | --- |
| 序号 | 设计内容 | 所用时间 |
| 1 | 下发任务书，学生查阅资料 | 1天 |
| 2 | 组建团队，并进行选题和团队匹配 | 1天 |
| 3 | 各小组对自己的选题开始进行研究 | 2天 |
| 4 | 各小组开始准备数据并形成初步处理意见 | 1天 |
| 5 | 代码和实现分析的动手实践 | 2天 |
| 6 | 形成初步的报告书和ppt | 1天 |
| 7 | 答辩并完成报告书 | 2天 |
| 合 计 | | 2周 |

**四、课程设计考核及评分标准**

**1.设计报告要求**

课程设计报告要求逻辑清晰、层次分明、书写整洁。课程设计报告为每人一份，同一个小组的各成员的整体报告内容结构一致，但个人侧重点不同，个人着重撰写自己的工作内容，其他人的内容只要体现文档结构的完整性即可。

课程设计考核将综合考虑学生考勤和参与度、团队协作能力，过程管理能力、成果达成情况等。

**2.过程要求**

整个过程要求通过github来进行过程化发布，即阶段性地在github上提交过程结果。

**3.评分标准**

|  |  |
| --- | --- |
| **评分依据** | **评分成绩** |
| 1．团队协作能力 | 25分 |
| 2．python综合运用能力 | 25分 |
| 3．态度认真、刻苦钻研、遵守纪律 | 10分 |
| 4．过程完成、对工具的使用、对github的运用 | 20分 |
| 5．课程设计答辩逻辑清晰，内容正确 | 10分 |
| 6. 课程设计期间的课堂考勤、创新能力 | 10分 |
| 总分 | 100分 |

注：按上述六项分别记分后求和，根据小组成员贡献率综合评定，记载个人最后成绩。

成绩等级：优（90分—100分）、良（80分—89分）、中（70分—79分）、及格（60分—69分）、60分以下为不及格。

**五、指导地点与时间**

本课程设计将安排在第20-21周，采用腾讯会议和QQ群的方式，以在线形式进行。具体安排如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 周次 | 星期一 | 星期二 | 星期三 | 星期四 | 星期五 |
| 第20周 | 第5-8节 | 第5-8节 |  | 第5-8节 |  |
| 第21周 |  | 第5-8节 |  | 第5-8节 |  |

执笔:徐 勇 日期：2021-1-8

审阅:钱小红 日期：2021-1-8

**目 录**

[前言 1](#_Toc20876)

[1编程环境与相关技术 2](#_Toc24182)

[1.1硬件平台 2](#_Toc15773)

[1.2项目托管平台-GitHub 2](#_Toc11176)

[1.3开发平台-jupyter 4](#_Toc10104)

[1.4编程语言-python 4](#_Toc26849)

[2程序设计（仅展示关键代码） 6](#_Toc12809)

[2.1.1数据分析库导入 6](#_Toc27293)

[2.1.2数据分析库注解 6](#_Toc4140)

[2.2数据表 7](#_Toc20299)

[2.2.1数据表导入 7](#_Toc26998)

[2.2.2数据表说明 8](#_Toc8074)

[2.3数据清洗 9](#_Toc3584)

[2.3.1数据清洗代码 9](#_Toc30010)

[2.4震级分析图 9](#_Toc14762)

[2.4.1数据提取代码 9](#_Toc2098)

[2.4.2生成震级分析图代码 10](#_Toc27262)

[2.5 20年以来全球4.5级以上地震统计图 10](#_Toc13730)

[2.5.1统计图生成代码 10](#_Toc11026)

[2.6 20年以来全球7级以上地震统计图 11](#_Toc32604)

[2.7 近5年以来全球每月4.5级以上地震统计图 12](#_Toc31008)

[2.7.1统计图生成代码 12](#_Toc10912)

[2.8地震类型分析 13](#_Toc29097)

[2.8.1地震类型分析生成图代码 13](#_Toc8741)

[2.8.2地震类型分析图 14](#_Toc30554)

[2.9地震高频发生地区分析 15](#_Toc7530)

[2.9.1地震类型分析生成图代码 15](#_Toc27930)

[3数据分析 15](#_Toc24474)

[3.1问题引入 15](#_Toc11739)

[3.2数图分析 16](#_Toc12935)

[3.3引发地震的因素有哪些？ 18](#_Toc29737)

[3.4我国哪些地方比较经常发生地震？ 20](#_Toc9862)

[3.5为什么好像2008年之后四川经常发生地震？ 21](#_Toc14097)

[3.6全世界地震频发的地区有哪些？ 24](#_Toc9787)

[3.7 2000年以来引发全世界关注的8大地震 25](#_Toc28495)

[4数据挖掘 27](#_Toc9609)

[4.1数据挖掘基本信息 27](#_Toc3819)

[4.2数据挖掘的基本步骤 27](#_Toc11590)

[4.3数据挖掘的常用方法 28](#_Toc1700)

**前言**

地震是一种及其普通和常见的一种自然现象，但由于地壳构造的复杂性和震源区的不可直观性，关于地震特别构造地震，它是怎样孕育和发生的，其成因和机制是什么的问题，至今尚无完满的解答，但目前科学家比较公认的解释是构造地震是由地壳板块运动造成的。

震级是地震大小的一种度量，根据地震释放能量的多少来划分，用“级”来表示。同样大小的地震，造成的破坏不一定相同；同一次地震，在不同的地方造成的破坏也不同。为衡量地震破坏程度，科学家又“制作”了另一把“尺子”一一[地震烈度](https://baike.baidu.com/item/%E5%9C%B0%E9%9C%87%E7%83%88%E5%BA%A6" \t "https://baike.baidu.com/item/%E5%9C%B0%E9%9C%87/_blank)。在中国地震烈度表上，对人的感觉、一般房屋震害程度和其他现象作了描述，可以作为确定烈度的基本依据。影响烈度的因素有震级、震源深度、距震源的远近、地面状况和地层构造等。

对于地震，我们更应该做的是提高建筑抗震等级、做好防御，而不是预测地震。但我还是不禁好奇地震的最近地震真的比以往频繁吗？引发地震的因素有哪些？我国哪些地方比较经常发生地震？我不经有了更多的疑问。

# 1编程环境与相关技术

## 1.1硬件平台

CPU：Intel(R) Core(TM) i5-8300H CPU @ 2.30GHz (8 CPUs), ~2.3GHz

最低要求：64位7代i3处理器及以上

内存：16G

最低要求：4G

硬盘空间：SSD-120G

最低要求：机械硬盘-20G

**1.2项目托管平台-GitHub**

GitHub 是一个面向[开源](https://baike.baidu.com/item/%E5%BC%80%E6%BA%90/20720669" \t "https://baike.baidu.com/item/_blank)及私有[软件](https://baike.baidu.com/item/%E8%BD%AF%E4%BB%B6/12053" \t "https://baike.baidu.com/item/_blank)项目的托管平台，因为只支持 Git 作为唯一的版本库格式进行托管，故名 GitHub。

GitHub 于 2008 年 4 月 10 日正式上线，除了 [Git](https://baike.baidu.com/item/Git/12647237" \t "https://baike.baidu.com/item/_blank) 代码仓库托管及基本的 Web 管理界面以外，还提供了订阅、讨论组、文本渲染、在线文件编辑器、协作图谱（报表）、代码片段分享（Gist）等功能。目前，其注册用户已经超过 350 万，托管版本数量也是非常之多，其中不乏知名开源项目 [Ruby](https://baike.baidu.com/item/Ruby/11419" \t "https://baike.baidu.com/item/_blank) on Rails、[jQuery](https://baike.baidu.com/item/jQuery/5385065" \t "https://baike.baidu.com/item/_blank)、[python](https://baike.baidu.com/item/python/407313" \t "https://baike.baidu.com/item/_blank) 等。

2018 年 6 月 4 日，微软宣布，通过 75 亿美元的股票交易收购代码托管平台 GitHub。

作为开源代码库以及版本控制系统，Github拥有超过900万开发者用户。随着越来越多的应用程序转移到了云上，Github已经成为了管理软件开发以及发现已有代码的首选方法。

如前所述，作为一个分布式的版本控制系统，在Git中并不存在主库这样的概念，每一份[复制](https://baike.baidu.com/item/%E5%A4%8D%E5%88%B6" \t "https://baike.baidu.com/item/_blank)出的库都可以独立使用，任何两个库之间的不一致之处都可以进行合并。

GitHub可以托管各种git库，并提供一个web界面，但它与外国的[SourceForge](https://baike.baidu.com/item/SourceForge/6562141" \t "https://baike.baidu.com/item/_blank)、[Google Code](https://baike.baidu.com/item/Google Code" \t "https://baike.baidu.com/item/_blank)或中国的[coding](https://baike.baidu.com/item/coding/8921246" \t "https://baike.baidu.com/item/_blank)的服务不同，GitHub的独特卖点在于从另外一个项目进行分支的简易性。为一个项目贡献代码非常简单：首先点击项目站点的“fork”的按钮，然后将代码检出并将修改加入到刚才分出的代码库中，最后通过内建的“pull request”机制向项目负责人申请代码合并。已经有人将GitHub称为代码玩家的MySpace。

在GitHub进行分支就像在[Myspace](https://baike.baidu.com/item/Myspace" \t "https://baike.baidu.com/item/_blank)（或[Facebook](https://baike.baidu.com/item/Facebook" \t "https://baike.baidu.com/item/_blank)…）进行交友一样，在社会关系图的节点中不断的连线。

GitHub项目本身自然而然的也在GitHub上进行托管，只不过在一个私有的，公共视图不可见的库中。[开源项目](https://baike.baidu.com/item/%E5%BC%80%E6%BA%90%E9%A1%B9%E7%9B%AE" \t "https://baike.baidu.com/item/_blank)可以免费托管，但私有库则并不如此。Chris Wanstrath，GitHub的开发者之一，肯定了通过付费的私有库来在财务上支持免费库的托管这一计划。

通过与客户的接洽，开发FamSpam，甚至是开发GitHub本身，GitHub的私有库已经被证明了物有所值。任何希望节省时间并希望和团队其它成员一样远离页面频繁转换之苦的人士都会从GitHub中获得他们真正想要的价值。

在GitHub，用户可以十分轻易地找到海量的[开源](https://baike.baidu.com/item/%E5%BC%80%E6%BA%90/20720669" \t "https://baike.baidu.com/item/_blank)代码。

## 1.3开发平台-jupyter

Jupyter Notebook（此前被称为 IPython notebook）是一个交互式笔记本，支持运行 40 多种编程语言。

Jupyter Notebook 的本质是一个 Web[应用程序](https://baike.baidu.com/item/ %E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F/5985445" \t "https://baike.baidu.com/item/Jupyter/_blank)，便于创建和共享文学化程序文档，支持实时代码，数学方程，可视化和 [markdown](https://baike.baidu.com/item/markdown/3245829" \t "https://baike.baidu.com/item/Jupyter/_blank)。 用途包括：数据清理和转换，数值模拟，统计建模，机器学习等等

　用户可以通过电子邮件，Dropbox，GitHub 和 Jupyter Notebook Viewer，将 Jupyter Notebook 分享给其他人。在Jupyter Notebook 中，代码可以实时的生成图像，视频，LaTeX和JavaScript。

当你还处于原型开发阶段时，Jupyter Notebooks 的作用更是引人注目。这是因为你的代码是按独立单元的形式编写的，而且这些单元是独立执行的。这让用户可以测试一个项目中的特定代码块，而无需从项目开始处执行代码。很多其它 IDE 环境（比如 RStudio）也有其它几种方式能做到这一点，但我个人觉得 Jupyter 的单个单元结构是最好的。它们甚至允许你运行 Python 之外的其它语言，比如 R、SQL 等。因为它们比单纯的 IDE 平台更具交互性，所以它们被广泛用于以更具教学性的方式展示代码。

**1.4编程语言-python**

自从20世纪90年代初Python语言诞生至今，它已被逐渐广泛应用于系统管理任务的处理和[Web](https://baike.baidu.com/item/Web/150564" \t "https://baike.baidu.com/item/Python/_blank)编程。

Python的创始人为荷兰人吉多·范罗苏姆 （Guido van Rossum）。1989年圣诞节期间，在[阿姆斯特丹](https://baike.baidu.com/item/%E9%98%BF%E5%A7%86%E6%96%AF%E7%89%B9%E4%B8%B9/2259975" \t "https://baike.baidu.com/item/Python/_blank)，Guido为了打发[圣诞节](https://baike.baidu.com/item/%E5%9C%A3%E8%AF%9E%E8%8A%82/127881" \t "https://baike.baidu.com/item/Python/_blank)的无趣，决心开发一个新的脚本解释程序，作为ABC语言的一种继承。之所以选中Python（大蟒蛇的意思）作为该编程语言的名字，是取自英国20世纪70年代首播的电视喜剧《蒙提·派森的飞行马戏团》（Monty Python's Flying Circus）。

ABC是由Guido参加设计的一种[教学](https://baike.baidu.com/item/%E6%95%99%E5%AD%A6" \t "https://baike.baidu.com/item/Python/_blank)语言。就Guido本人看来，ABC这种语言非常优美和强大，是专门为非专业程序员设计的。但是ABC语言并没有成功，究其原因，Guido认为是其非开放造成的。Guido决心在Python中避免这一错误。同时，他还想实现在ABC中闪现过但未曾实现的东西。

就这样，Python在Guido手中诞生了。可以说，Python是从ABC发展起来，主要受到了[Modula-3](https://baike.baidu.com/item/Modula-3/17009923" \t "https://baike.baidu.com/item/Python/_blank)（另一种相当优美且强大的语言，为小型团体所设计的）的影响。并且结合了[Unix shell](https://baike.baidu.com/item/Unix shell" \t "https://baike.baidu.com/item/Python/_blank)和[C](https://baike.baidu.com/item/C/7252092" \t "https://baike.baidu.com/item/Python/_blank)的习惯。

Python  已经成为最受欢迎的[程序设计语言](https://baike.baidu.com/item/%E7%A8%8B%E5%BA%8F%E8%AE%BE%E8%AE%A1%E8%AF%AD%E8%A8%80/2317999" \t "https://baike.baidu.com/item/Python/_blank)之一。自从2004年以后，python的使用率呈线性增长。Python 2于2000年10月16日发布，稳定版本是Python 2.7。Python 3于2008年12月3日发布，不完全兼容Python 2。 2011年1月，它被[TIOBE](https://baike.baidu.com/item/TIOBE" \t "https://baike.baidu.com/item/Python/_blank)编程语言[排行榜](https://baike.baidu.com/item/%E6%8E%92%E8%A1%8C%E6%A6%9C/4895" \t "https://baike.baidu.com/item/Python/_blank)评为2010年度语言。

由于Python语言的[简洁](https://baike.baidu.com/item/%E7%AE%80%E6%B4%81" \t "https://baike.baidu.com/item/Python/_blank)性、易读性以及可扩展性，在国外用Python做科学计算的研究机构日益增多，一些知名大学已经采用Python来教授程序设计[课程](https://baike.baidu.com/item/%E8%AF%BE%E7%A8%8B" \t "https://baike.baidu.com/item/Python/_blank)。例如[卡耐基梅隆大学](https://baike.baidu.com/item/%E5%8D%A1%E8%80%90%E5%9F%BA%E6%A2%85%E9%9A%86%E5%A4%A7%E5%AD%A6" \t "https://baike.baidu.com/item/Python/_blank)的编程基础、[麻省理工学院](https://baike.baidu.com/item/%E9%BA%BB%E7%9C%81%E7%90%86%E5%B7%A5%E5%AD%A6%E9%99%A2/117999" \t "https://baike.baidu.com/item/Python/_blank)的计算机科学及编程导论就使用Python语言讲授。众多开源的科学计算软件包都提供了Python的调用[接口](https://baike.baidu.com/item/%E6%8E%A5%E5%8F%A3" \t "https://baike.baidu.com/item/Python/_blank)，例如著名的计算机视觉库[OpenCV](https://baike.baidu.com/item/OpenCV" \t "https://baike.baidu.com/item/Python/_blank)、三维可视化库VTK、医学图像处理库ITK。而Python专用的科学计算扩展库就更多了，例如如下3个十分经典的科学计算扩展库：NumPy、SciPy和matplotlib，它们分别为Python提供了快速数组处理、数值运算以及绘图功能。因此Python语言及其众多的扩展库所构成的开发环境十分适合[工程](https://baike.baidu.com/item/%E5%B7%A5%E7%A8%8B" \t "https://baike.baidu.com/item/Python/_blank)技术、科研人员处理实验数据、制作图表，甚至开发科学计算[应用程序](https://baike.baidu.com/item/%E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F" \t "https://baike.baidu.com/item/Python/_blank)。

**2程序设计（仅展示关键代码）**

**2.1数据库分析**

### 2.1.1数据分析库导入见下图2.1-1

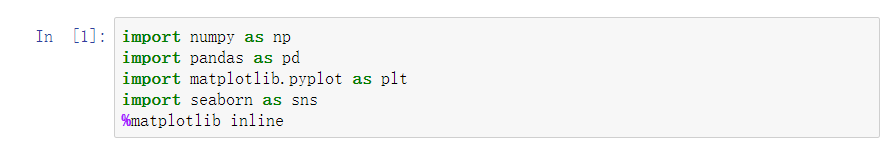


图2.1-1

### 2.1.2数据分析库注解

pandas 是 python 的数据分析处理库 NumPy是Python语言的一个扩充程序库。支持高级大量的维度数组与bai矩阵运算，此外也针对数组运算提供大量的数学函数库。

Numpy内部解除了CPython的GIL（全局解释器锁），运行效率极好，是大量机器学习框架的基础库！

使用%matplotlib命令可以将matplotlib的图表直接嵌入到Notebook之中，或者使用指定的界面库显示图表，它有一个参数指定matplotlib图表的显示方式。inline表示将图表嵌入到Notebook中。

%matplotlib inline Seaborn是一种基于matplotlib的图形可视化python libraty。它提供了一种高度交互式界面，便于用户能够做出各种有吸引力的统计图表。

Seaborn其实是在matplotlib的基础上进行了更高级的API封装，从而使得作图更加容易，在大多数情况下使用seaborn就能做出很具有吸引力的图，而使matplotlib就能制作具有更多特色的图。

Matplotlib是一个Python的2D绘图库，它以各种硬拷贝格式和跨平台的交互式环境生成出版质量级别的图形。通过Matplotlib，开发者可以仅需要几行代码，便可以生成绘图。一般可绘制折线图、散点图、柱状图、饼图、直方图、子图等等。Matplot使用Numpy进行数组运算，并调用一系列其他的Python库来实现硬件交互。

## 2.2数据表

### 2.2.1数据表导入见下图2.2-1



图2.2-1

### 2.2.2数据表说明

数据来源：美国地质勘探局（USGS）2000年以来全球发生的所有4.5级以上地震数据131865条

数据表注释：

1999-2019.csv数据文件的数据代表的是1999年-2019年全球的地震信息数据

eartquake.csv数据文件的数据代表的是全球热点地区的地震信息数据

sichuan.csv数据文件的数据代表的是中国四川近年来的地震信息数据

## 2.3数据清洗

### 2.3.1数据清洗代码见下图2.3-1



图2.3-1

## 2.4震级分析图

### 2.4.1数据提取代码见下图2.4-1



图2.4-1

### 2.4.2生成震级分析图代码见下图2.4-2

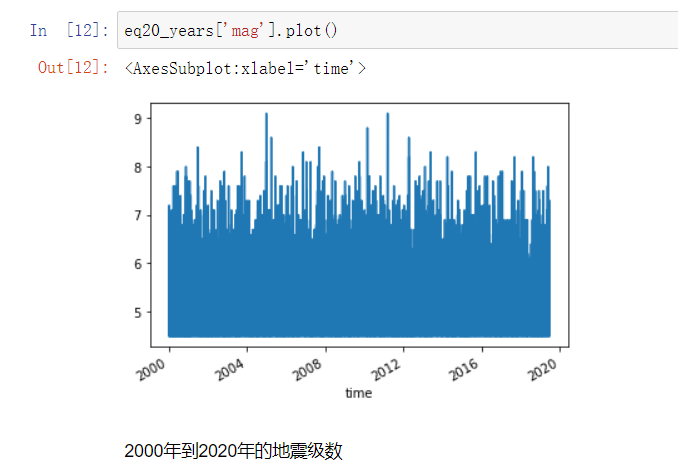


图2.4-2

## 2.5 20年以来全球4.5级以上地震统计图

### 2.5.1统计图生成代码见下图2.5-1和2.5-2



图2.5-1

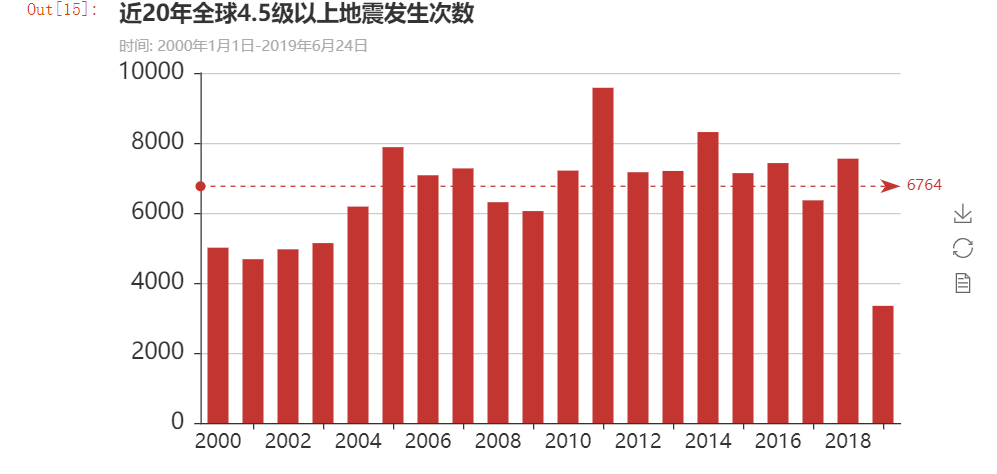


图2.5-2

**2.6 20年以来全球7级以上地震统计图**

**2.6.1统计图生成代码见下图2.6-1**

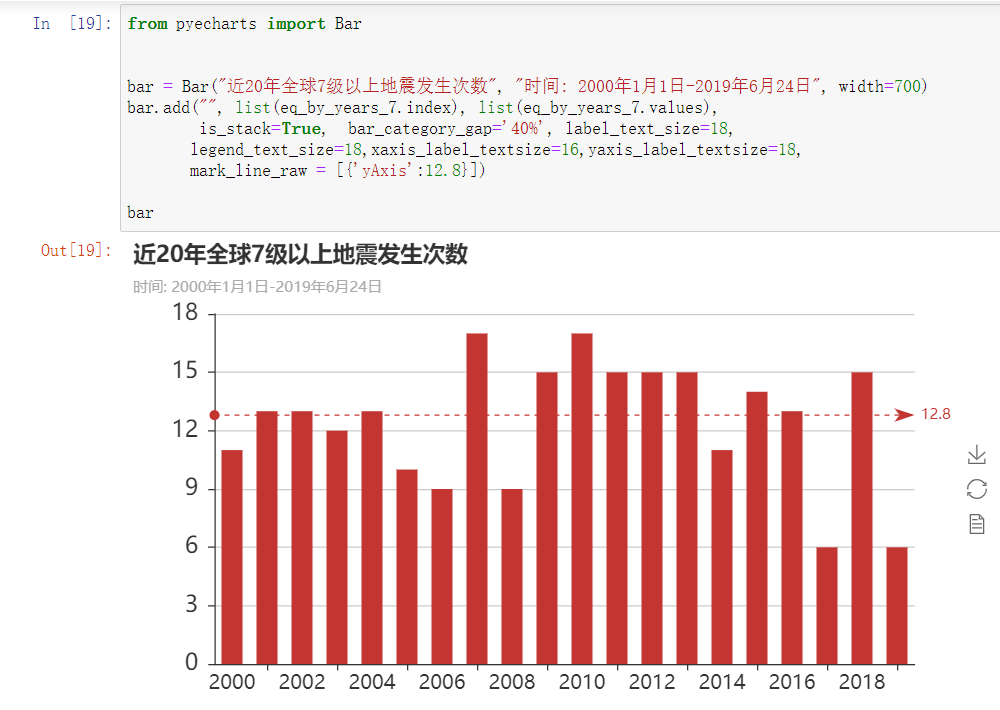
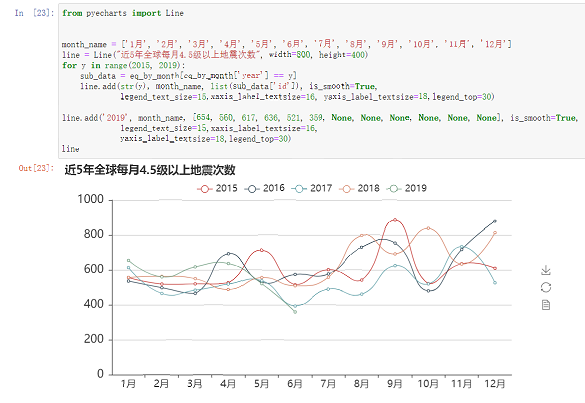


图2.6-1

## 2.7 近5年以来全球每月4.5级以上地震统计图

### 2.7.1统计图生成代码见下图2.7-1

图2.7-1

## 2.8地震类型分析

### 2.8.1地震类型分析生成图代码见下图2.8-1

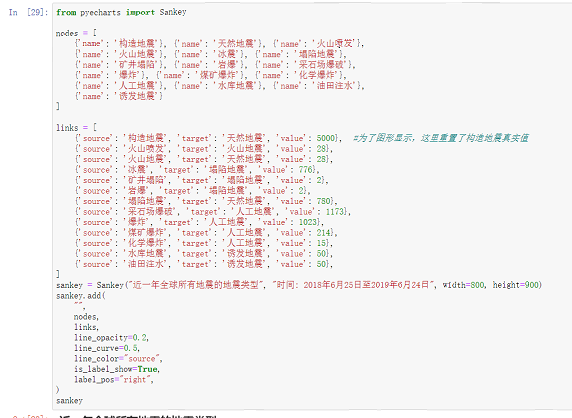


图2.8-1

### 2.8.2地震类型分析图见下图2.8.2

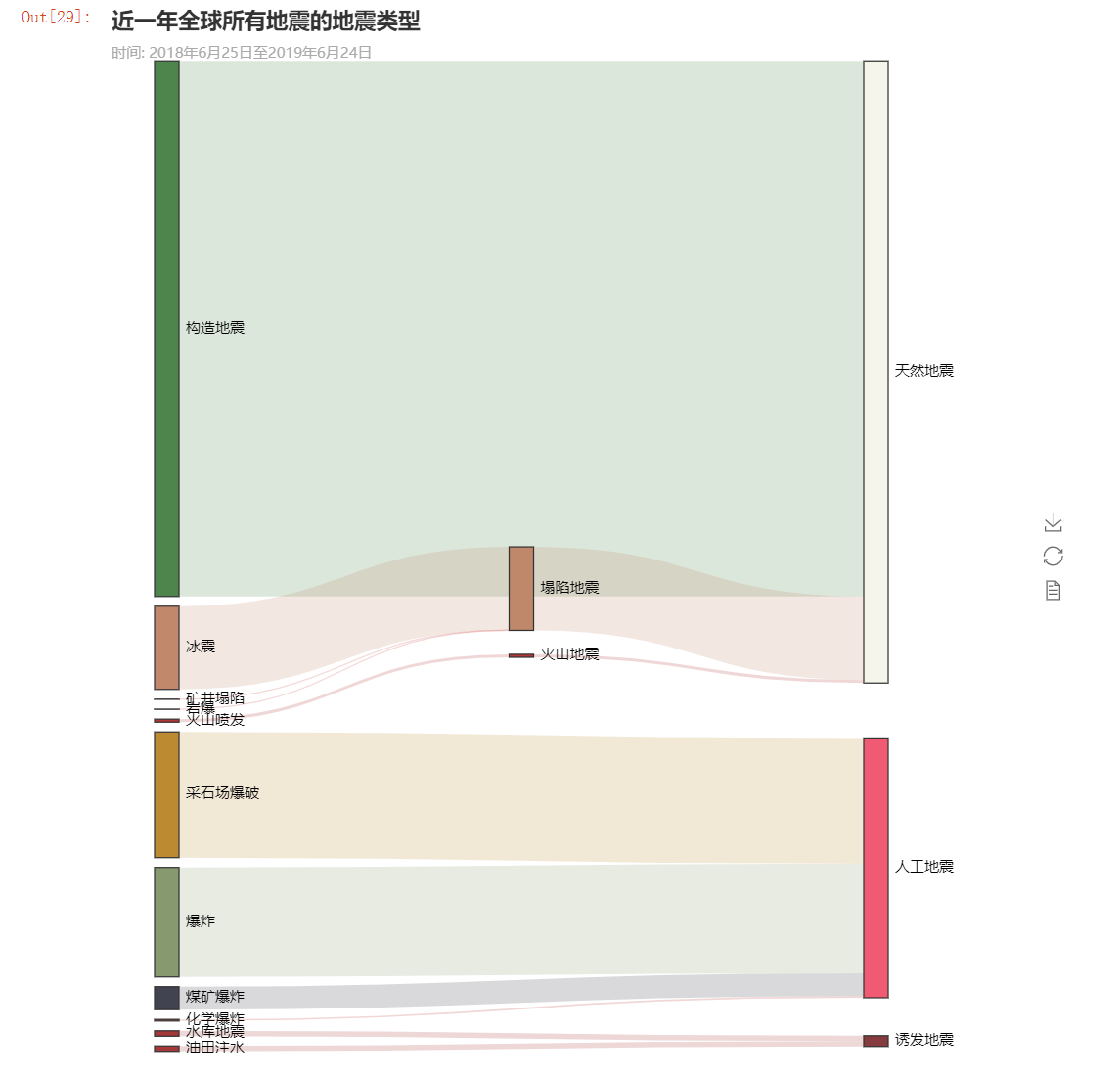


图2.8-2

## 2.9地震高频发生地区分析

### 2.9.1地震类型分析生成图代码见下图2.9-1



图2.9-1

**3数据分析**

## 3.1问题引入

最近宜宾两次震感明显的地震，楚雄一次震感明显的地震，大家都在说最近地震怎么那么频繁呢？是不是最近一年的地震真的比以往多呢？相关网络言论见图3.1-1



图3.1-1

## 3.2数图分析

我们先看全球近20年以来4.5级以上的地震发生次数，见图3.2-1

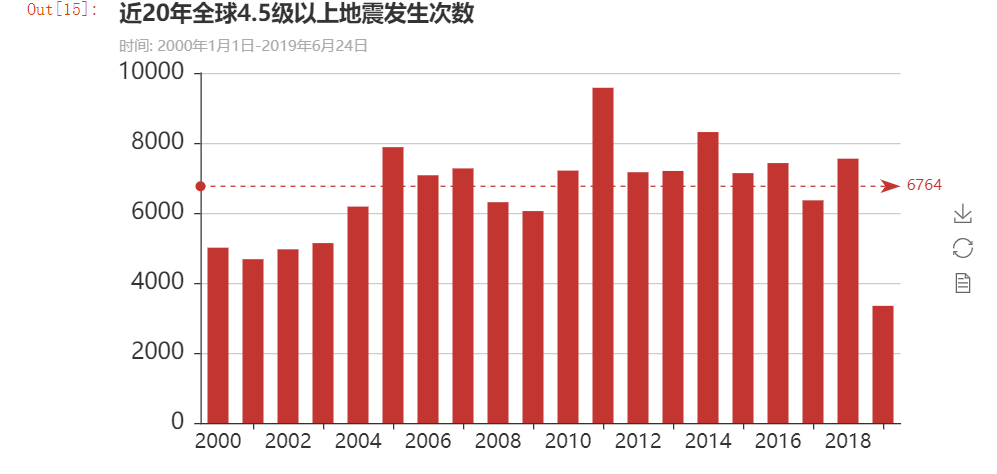


图3.2-1

这是2000年以来全球4.5级以上地震每年发生次数图。可以看到，2000年以来全球平均每年发生6764次4.5级以上地震（是的，其实地球一直开启的都是震动模式，地壳总是在运动的），2019年到6月24日止已发生3347次，对比其它年度并没有明显的提升。

还有人说，地震和太阳黑子活动周期都是11年，上次是2008年，这次是2019年。相关网络言论见图3.2-2

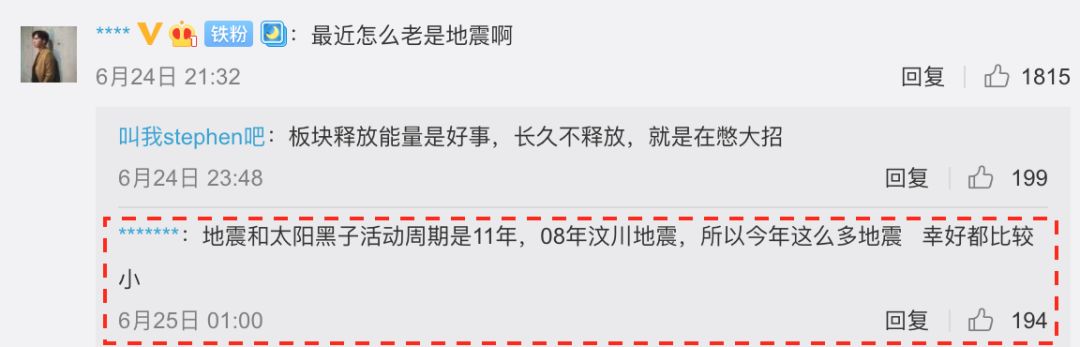


图3.2-2

可见， 2008年的4.5级以上地震发生次数为6309次，比平均次数少455次，并不存在所谓的地震受太阳黑子爆发的周期性（11年）影响。当然可能有人要说，这里说的周期性指的是大地震，并不包括4.5级这些中小地震。我们再来看图3.2-3。

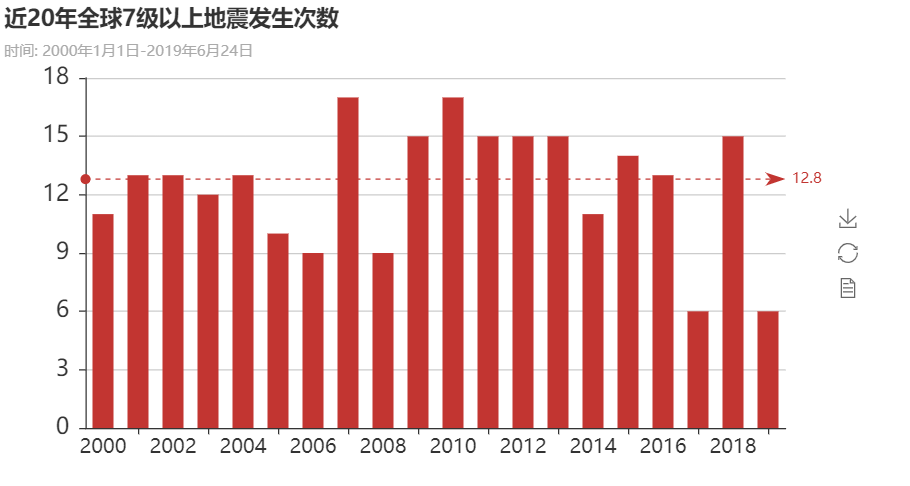


图3.2-3

这是近20年全球7级以上地震发生次数图。全球近19年平均每年发生了12.8次7级以上的大地震和特大地震。今年已经发生了6次，对比其它年份并没有明显提升，而11年前的2008年发生了9次，比平均次数少3.8次。所以说，把地震跟太阳黑子活动周期联系起来的说法，一定程度上是试图把2019年地震和2008年汶川地震联系起来，会造成大家的恐慌，是不可取的。

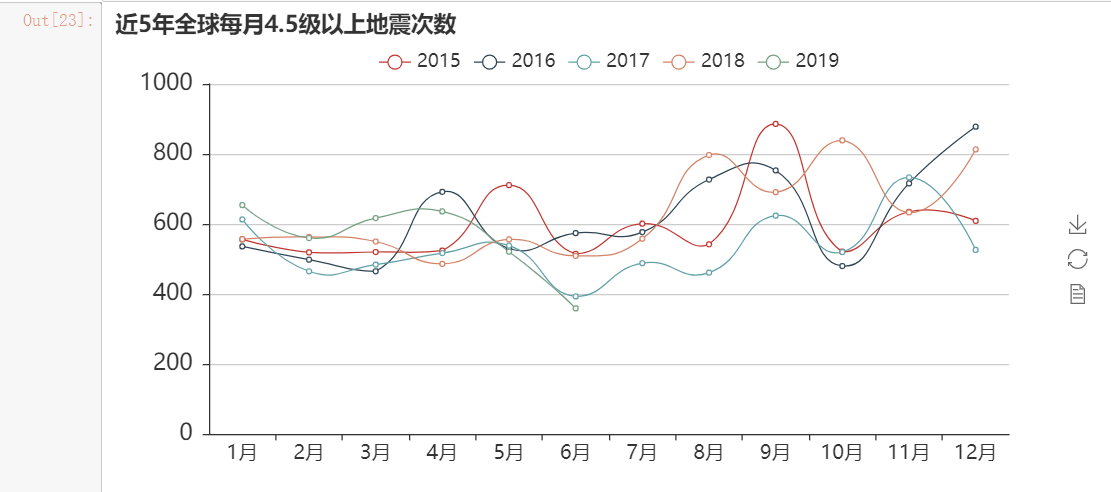


图3.2-4

这是2015年来每月发生地震的次数图，可以看到2019年6月份发生地震的次数也跟其它年份没有什么太大的差异。

总结：所以，我们能做的，不是以讹传讹，而是普及地震常识，多参加地震逃生演练，做一些必要的准备，知道一旦发生地震要怎么正确应对，就可以了。

## 3.3引发地震的因素有哪些？

地震主要包括天然地震和人工地震，天然地震中，绝大部份是由于板块碰撞挤压、地下岩石破裂、错动导致的构造地震。我们获取了2018年6月25日至2019年6月24日一年间，全球发生的所有0级以上地震数据161565条，发现97.9%的地震都属于构造地震。以下是近一年全球所有地震的各种地震类型发生数量图3.3-1。

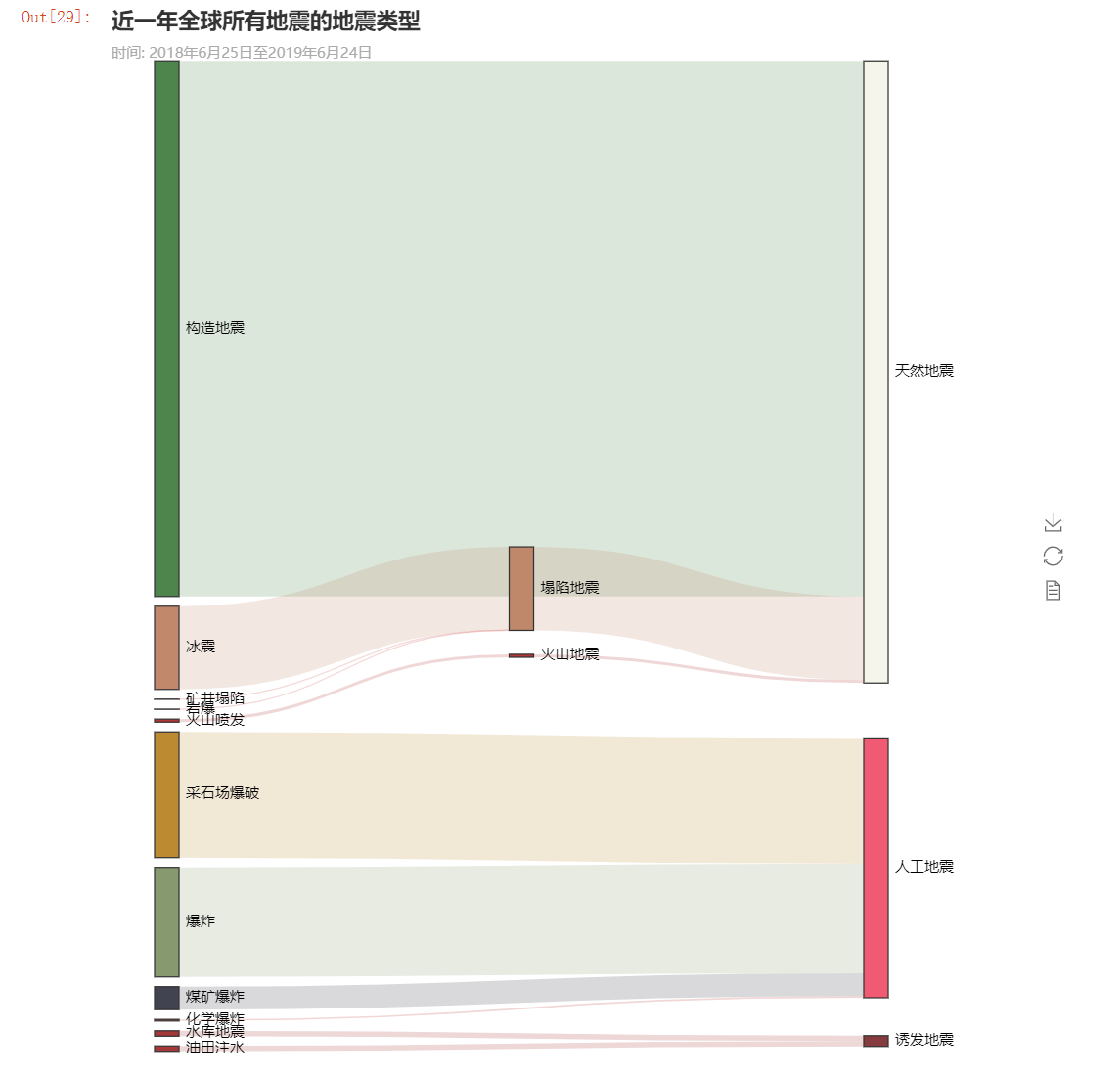


图3.3-1

(注：为了更好地显示，这里重置了构造地震的真实值)

## 3.4我国哪些地方比较经常发生地震？

我们把我国及其周边地区近20年4.5级以上的地震都以圆点的形式投射在了地图上，圆点越大，表明地震震级越大。见图3.4-1



图3.4-1

由图可见，我国的地震主要集中在西部的四川、云南、西藏、青海、甘肃、新疆以及东部的台湾，这些地方也就是板块活动频繁的地方。所以生活在这些地区的人们，需要更加注重了解地震的相关知识以及逃生技能。其它省份的人们不需要太过于担心地震，当然也不能掉以轻心。另外，我国周边的尼泊尔、缅甸、菲律宾、日本等国家，也是地震经常光顾的主儿。

3.5为什么好像2008年之后四川经常发生地震？

有很多朋友可能有这样一些疑问：怎么四川省自从2008年汶川地震之后就好像经常发生地震啊？2008年之前好像都没有听说过四川省有地震呢？

真的是这样吗？我们获取了1979年以来近40年四川省及其周围4.5级以上地震的所有数据，并且绘制出每年四川省及其周围地区发生4.5级以上地震次数图。见图3.5.1

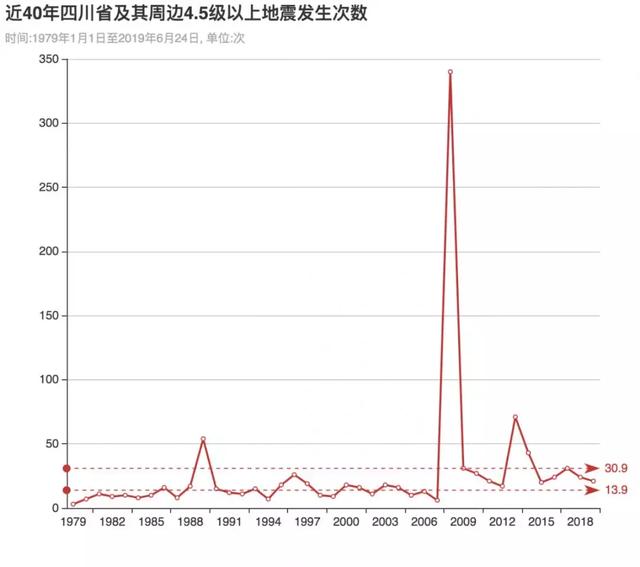


图3.5-1

我们可以把2008年作为一个分界线。2009年到2018年这10年间，每年4.5级地震平均发生次数是30.9次，而1979年到2007年，每年4.5级地震发生次数是13.9次，确实2008年之后四川及其周边地区地震发生次数比2008年之前频繁了一些。

但是也可以看到，2008年之前四川及其周边地区也是有地震的，1979年到2008年，四川及其周边共发生了9次6级以上的地震。比如说1981年1月23日，四川甘孜藏族自治州道孚县发生了6.9级地震，造成各类房屋倒塌2992幢，死123人，伤489人，成都市有震感。

为什么我们觉得2008年以前好像没有听说过四川有什么地震，2008年以后四川地震很频繁呢？我们总结了以下一些原因：

1. 2008年之后四川及其周边地震确实比以前频繁了一些；

2. 信息流通变得越来越通畅了。现在某个地方发生一个4.5级的地震，朋友圈里面的震感可能比震中还要强烈。而以前某个地方发生地震，离得稍远的地方有人可以感受到地震，但不会产生大的讨论；

3. 由于2008年汶川地震的影响，现在地震越来越被大家所关注。

同样，我们把这份数据以圆点的形式投射在了地图上。可以看到，位于雅安、成都、德阳、绵阳、广元北边有一条非常明显的东北-西南走向的地震带。见下图3.5-2



图3.5-2

3.6全世界地震频发的地区有哪些？

我们经常听说印尼地震、日本地震等，那么全世界地震频发的国家或地区有哪些呢？见下图3.6-1

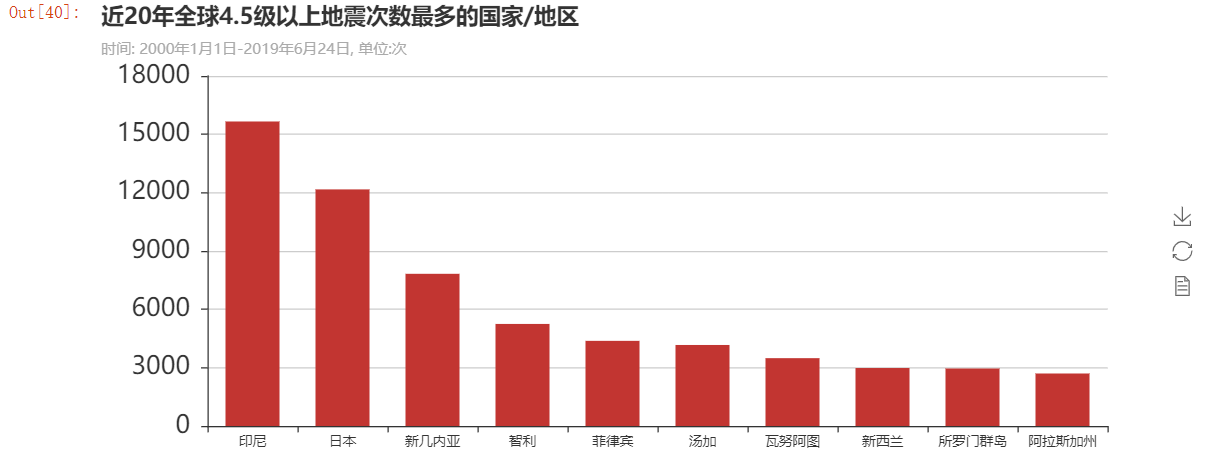


图3.6-1

我们选取了近20年4.5级以上地震发生次数最多的前10个国家或地区。可以看到，印尼和日本果然是全球地震次数最多的国家。特别是印尼，2000年以来发生了15661场4.5级以上的地震，平均每天有2.2场。除此之外，还有亚洲的菲律宾、大洋洲的巴布亚新几内亚、汤加、瓦努阿图、新西兰、所罗门群岛，南美的智利，北美的阿拉斯加州。

为什么地震经常发生在这些国家或者地区呢？我们把近20年全球4.5级以上地震以圆点形式投射在地图上。见图3.6-2

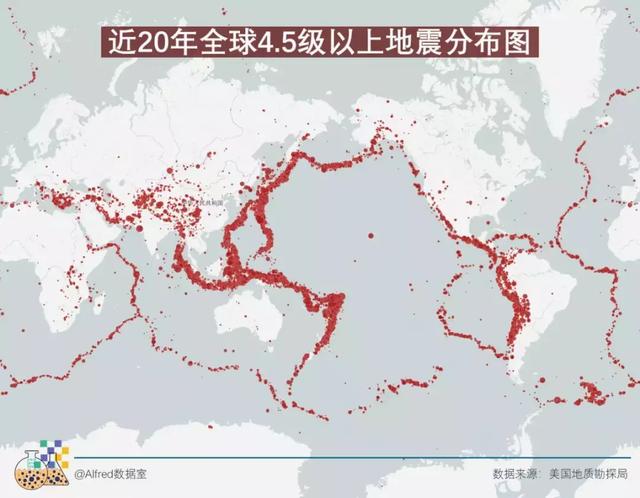


图3.6-2

这幅地图，便是全球地震带的分布图。不同的地震带把地球分成了很明显的好几个板块。这些地震频发的国家，几乎都位于环太平洋火山地震带上。

3.7 2000年以来引发全世界关注的8大地震

地震是一种无法干预、目前还不能完全预测的自然现象和灾害。它告诉我们要敬畏自然、理解自然，并且要通过学习掌握自然规律，来减少灾害造成的影响。我们来看一看2000年以来引发全世界关注的8大地震。见图3.7-1



图3.7-1

**4数据挖掘**

**4.1数据挖掘基本信息**

数据挖掘(Data Mining)是通过分析每个数据，从大量数据中寻找其规律的技术，主要有数据准备、规律寻找和规律表示3个步骤。数据挖掘的任务有关联分析、[聚类分析](https://baike.so.com/doc/4551628-4762184.html" \t "https://baike.so.com/doc/_blank)、分类分析、异常分析、特异群组分析和演变分析等。

**4.2数据挖掘的基本步骤**

数据挖掘的步骤会随不同领域的应用而有所变化，每一种数据挖掘技术也会有各自的特性和使用步骤，针对不同问题和需求所制定的数据挖掘过程也会存在差异。此外，数据的完整程度、专业人员支持的程度等都会对建立数据挖掘过程有所影响。这些因素造成了数据挖掘在各不同领域中的运用、规划，以及流程的差异性，即使同一产业，也会因为分析技术和专业知识的涉入程度不同而不同，因此对于数据挖掘过程的系统化、标准化就显得格外重要。如此一来，不仅可以较容易地跨领域应用，也可以结合不同的专业知识，发挥数据挖掘的真正精神。

数据挖掘完整的步骤如下：

① 理解数据和数据的来源（understanding）。

② 获取相关知识与技术（acquisition）。

③ 整合与检查数据（integration and checking）。

④ 去除错误或不一致的数据（data cleaning）。

⑤ 建立模型和假设（model and hypothesis development）。

⑥ 实际数据挖掘工作（data mining）。

⑦ 测试和验证挖掘结果（testing and verification）。

⑧ 解释和应用（interpretation and use）。

由上述步骤可看出，数据挖掘牵涉了大量的准备工作与规划工作，事实上许多专家都认为整套数据挖掘的过程中，有80%的时间和精力是花费在数据预处理阶段，其中包括数据的净化、数据格式转换、变量整合，以及数据表的链接。可见，在进行数据挖掘技术的分析之前，还有许多准备工作要完成。

**4.3数据挖掘的常用方法**

在大数据时代，数据挖掘是最关键的工作。大数据的挖掘是从海量、不完全的、有噪声的、模糊的、随机的大型数据库中发现隐含在其中有价值的、潜在有用的信息和知识的过程，也是一种决策支持过程。其主要基于人工智能，机器学习，模式学习，统计学等。通过对大数据高度自动化地分析，做出归纳性的推理，从中挖掘出潜在的模式，可以帮助企业、商家、用户调整市场政策、减少风险、理性面对市场，并做出正确的决策。目前，在很多领域尤其是在商业领域如银行、电信、电商等，数据挖掘可以解决很多问题，包括市场营销策略制定、背景分析、企业管理危机等。大数据的挖掘常用的方法有分类、回归分析、聚类、关联规则、神经网络方法、Web 数据挖掘等。这些方法从不同的角度对数据进行挖掘。

(1)分类：分类是找出数据库中的一组数据对象的共同特点并按照分类模式将其划分为不同的类，其目的是通过分类模型，将数据库中的数据项映射到摸个给定的类别中。可以应用到涉及到应用分类、趋势预测中，如淘宝商铺将用户在一段时间内的购买情况划分成不同的类，根据情况向用户推荐关联类的商品，从而增加商铺的销售量。

(2)回归分析：回归分析反映了数据库中数据的属性值的特性，通过函数表达数据映射的关系来发现属性值之间的依赖关系。它可以应用到对数据序列的预测及相关关系的研究中去。在市场营销中，回归分析可以被应用到各个方面。如通过对本季度销售的回归分析，对下一季度的销售趋势作出预测并做出针对性的营销改变。

(3)聚类：聚类类似于分类，但与分类的目的不同，是针对数据的相似性和差异性将一组数据分为几个类别。属于同一类别的数据间的相似性很大，但不同类别之间数据的相似性很小，跨类的数据关联性很低。

(4)关联规则：关联规则是隐藏在数据项之间的关联或相互关系，即可以根据一个数据项的出现推导出其他数据项的出现。关联规则的挖掘过程主要包括两个阶段：第一阶段为从海量原始数据中找出所有的高频项目组;第二阶段为从这些高频项目组产生关联规则。关联规则挖掘技术已经被广泛应用于金融行业企业中用以预测客户的需求，各银行在自己的ATM 机上通过捆绑客户可能感兴趣的信息供用户了解并获取相应信息来改善自身的营销。

(5)神经网络方法：神经网络作为一种先进的人工智能技术，因其自身自行处理、分布存储和高度容错等特性非常适合处理非线性的以及那些以模糊、不完整、不严密的知识或数据为特征的处理问题，它的这一特点十分适合解决数据挖掘的问题。典型的神经网络模型主要分为三大类：第一类是以用于分类预测和模式识别的前馈式神经网络模型，其主要代表为函数型网络、感知机;第二类是用于联想记忆和优化算法的反馈式神经网络模型，以Hopfield 的离散模型和连续模型为代表。第三类是用于聚类的自组织映射方法，以ART 模型为代表。虽然神经网络有多种模型及算法，但在特定领域的数据挖掘中使用何种模型及算法并没有统一的规则，而且人们很难理解网络的学习及决策过程。

(6)Web数据挖掘：Web数据挖掘是一项综合性技术，指Web 从文档结构和使用的集合C 中发现隐含的模式P，如果将C看做是输入，P 看做是输出，那么Web 挖掘过程就可以看做是从输入到输出的一个映射过程。

当前越来越多的Web 数据都是以数据流的形式出现的，因此对Web 数据流挖掘就具有很重要的意义。目前常用的Web数据挖掘算法有：PageRank算法，HITS算法以及LOGSOM 算法。这三种算法提到的用户都是笼统的用户，并没有区分用户的个体。目前Web 数据挖掘面临着一些问题，包括：用户的分类问题、网站内容时效性问题，用户在页面停留时间问题，页面的链入与链出数问题等。在Web 技术高速发展的今天，这些问题仍旧值得研究并加以解决。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **小组分工明细表** | | | | | |
| **项目名称** | **地震数据分析** | | | | |
| **小组成员** | **林寒** | **周莹** | | **饶天成** | |
| **任务名称** | **主负责人** | | **协助人** | | **协助人** |
| **选题** | **林寒** | | **周莹** | | **饶天成** |
| **小组讨论分析** | **林寒** | | **周莹** | | **饶天成** |
| **建立github** | **周莹** | | **无** | | **无** |
| **数据采集** | **饶天成** | | **林寒** | | **无** |
| **环境搭建** | **林寒** | | **周莹** | | **饶天成** |
| **编码** | **林寒** | | **周莹** | | **饶天成** |
| **数据图形化** | **林寒** | | **周莹** | | **无** |
| **数据分析** | **饶天成** | | **林寒** | | **无** |
| **代码检查** | **周莹** | | **林寒** | | **无** |
| **注释文档编写** | **周莹** | | **林寒** | | **无** |
| **PPT报告编写** | **饶天成** | | **无** | | **无** |
| **Word报告编写** | **林寒** | | **周莹** | | **饶天成** |
| **上传项目文件** | **周莹** | | **无** | | **无** |
| **Github管理** | **周莹** | | **无** | | **无** |
| **项目汇报** | **林寒** | | **周莹** | | **饶天成** |

**课程设计成绩评定表**

|  |  |  |  |
| --- | --- | --- | --- |
| 课程设计题目 | 数据分析的学习与实践-地震数据分析 | | |
| 课程设计学生答辩或质疑记录：  1.数据分析的目的？  答：数据分析的目的是把隐没在一大批看来杂乱无章的数据中的信息集中、萃取和提炼出来，以找出所研究对象的内在规律。  2.数据清洗的目的何在？  简单的来说不干净的数据会导致分析过程中的错误以及结果的错误。举个例子，做柱形图这种类型的图时，如果大部分数据集中在某个区间而一两个数据离得很远，如果不去除这一两个有问题的数据，那整体的图画出来就会有问题，不能反映数据的情况。  3.[%matplotlib inline的含义](https://www.cnblogs.com/chester-cs/p/11825282.html)  答：用在Jupyter notebook中具体作用是当你调用matplotlib.pyplot的绘图函数plot()进行绘图的时候，或者生成一个figure画布的时候，可以直接在你的python console里面生成图像。 | | | |
| **评 分 依 据** | | **分 值** | **评分成绩** |
| 1．团队协作能力 | | 25分 |  |
| 2．python综合运用能力 | | 25分 |  |
| 3．态度认真、刻苦钻研、遵守纪律 | | 10分 |  |
| 4．过程完成、对工具的使用、对github的运用 | | 20分 |  |
| 5．课程设计答辩逻辑清晰，内容正确 | | 10分 |  |
| 6. 课程设计期间的课堂考勤、创新能力 | | 10分 |  |
| 总 分 | | 100分 |  |
| 最终评定等级为：  指导老师签字：  2021 年 1 月 15日 | | | |