开放案例研究：通过现实世界应用开展统计学与数据科学教育

Carrie Wright[1], Qier Meng[1], Michael R. Breshock[2], Lyla Atta[2], Margaret A. Taub [1], Leah R. Jager [1], John Muschelli [1,3], 和 Stephanie C. Hicks[1], *

[1]约翰斯·霍普金斯彭博公共卫生学院生物统计学系 [2]约翰斯·霍普金斯大学生物医学工程系 [3]Streamline Data Science*通信请联系 shicks19@jhu.edu

arXiv:2301.05298v1 [stat.AP] 12 Jan 2023

摘要

随着人们对数据科学教育空前且日益增长的兴趣，教育工作者可用的材料却十分有限，无法为学习者提供有意义的机会，让他们运用统计思维，即 Wild和Pfannkuch 所定义的统计思维 [1], , 处理应对现实世界挑战的杂乱数据。为此，Nolan 和 Speed [2] 主张在本科统计学课程中，借助深入的案例研究，让应用成为教学重点，以鼓励和培养课堂中的统计思维。这种方法的局限性包括开发案例研究所需的大量时间投入——具体来说，要选定一个有启发性的问题并制作示例性数据分析——以及所需的领域专业知识。因此，基于现实挑战而非玩具示例的案例研究十分稀少。为解决这一问题，我们开发了开放案例研究项目（opencasestudies.org），提出了一种新的统计学与数据科学教育案例研究模式。这项教育资源提供自成一体、多模态、经同行评审且开源的指南（或案例研究），以现实世界的实例为基础，帮助学习者主动完成完整的数据分析。我们还编写了教育者指南，说明如何最有效地使用这些案例研究，如何在课堂中修改和调整案例研究的组成部分，以及如何贡献新的案例研究。（opencasestudies.org/OCS Guide）。

关键词：应用统计学、数据科学、统计思维、案例研究、教育、计算

1 引言

数据科学和统计学教学实践中的一大挑战，是能够提供有意义的机会，让学生使用杂乱数据应对现实世界挑战，并在不同情境领域中练习和应用统计思维的课程和课程材料数量有限。思维，其定义见Wild和Pfannkuch [1], 。为解决这一问题，Nolan 和 Speed [2] 提出了一种开发案例研究（也称为“实验”）的模式，供本科统计学课程使用，具体目标是“鼓励并培养统计思维”。具体而言，该模式要求每个案例研究都应：

> “一项内容充实、解法并非显而易见的练习，既允许进行不同的分析，又能成为课程的核心组成部分。实验课应为学习理论统计学提供动机和框架，并让学生体验如何运用统计学来回答科学问题。这种方法的一个重要目标是，在传授数理统计学知识的同时，鼓励并培养统计思维。” [2]

2018年，Hicks 和 Irizarry [3] 指出，他们提出的数据科学教学五项原则之一，是按照 Nolan 和 Speed [2], 的模式，“围绕一系列不同的案例研究来组织课程”，以培养统计思维，并将现实世界应用带入课堂。课堂上也在众多不同领域中使用案例研究，包括统计学 [4–8], 进化生物学 [9],工程学 [10],和环境科学 [11]。

然而，扩大案例研究的应用仍面临若干限制因素。首先，选择驱动性问题 [12], 寻找真实世界中引人入胜的数据 [13, 14], 描述数据所处的背景 [15, 16], 并准备多样化的教学数据分析，需要在初期投入大量时间和精力 [3]。其次，最有条件开发出有效且富有洞见的案例研究的人，是实践型教师 [17], , 也就是在特定领域背景下教学和实践的应用统计学家与数据科学家。对这些人而言，成功——

要在广泛的背景主题中全面构建一套多样化的案例研究，可能需要与其他学科的人员合作；如果其学术机构不提供受保护的时间和精力，这种合作可能很难实现 [18]。第三，尽管有丰富的数据集存储库 [7]，但展示如何利用数据来说明数据科学和统计学基本概念（并可能产生意外结果）的配套数据分析合集却很少[19]。对于复杂且杂乱的数据而言，这一点尤为突出，因为分析决策必须超越数据简要总结（例如 README 文件）所能概括的内容 [20, 21]。这些挑战导致基于真实世界挑战而非简单玩具示例的案例研究十分匮乏。此外，许多数据存储库会针对数据子集推荐不同的处理和分析方法，而这些方法通常被当作"标准"分析，却没有对研究过程中其他可选方案进行充分讨论。

为应对这些挑战，我们开发了一项开放式教育资源——开放案例研究（OCS）项目（opencasestudies.org）。该资源包含深入、独立完整、多模态且经过同行评审的体验式指南（即案例研究），通过展示涵盖统计学和数据科学各类主题的示例性数据分析，教会学习者如何有效地从数据中获取知识。授课教师可以利用这些指南，在课堂上突出应用；独立学习者也可以在课堂之外使用这些指南。最后，我们编写了一份教育者指南，介绍如何最有效地使用这些案例研究、如何在课堂上修改和调整案例研究的组成部分，以及如何贡献新的案例研究。（opencasestudies.org/OCS Guide）。

本文其余部分安排如下。首先，我们概述并讨论开放案例研究模型的各个组成部分（**第2**节），这是一个对 [2] 案例研究模型加以扩展的新模型。其次，我们介绍开放案例研究教育资源（**第3**节）。第三，我们根据自身经验，为其他人如何创建自己的案例研究提供指导（**第4**节），其中包括如何创建交互式案例研究。最后，我们总结此类案例研究在课堂内外的用途（**第5**节）。

2 将 OCS 模型付诸实践

2.1 OpenCase 案例研究模型概述

Nolan 和 Speed[2] 所描述的案例研究模型将每个案例研究分为五个主要组成部分：(i) 引言，(ii) 数据描述，(iii) 背景，(iv) 调查，以及(v) 理论，并可选择加入一个关于高级分析或相关理论材——的章节

料。在我们的开放案例研究（OCS）模型中，我们将这些组成部分扩展为十三个组成部分。表1介绍了 OCS 模型的组成部分，以及我们的模型与 Nolan 和 Speed 的原始模型之间的对应关系 [2]。

我们强调，尽管这两种案例研究模型的结构相似，但我们的 OCS 模型与 Nolan 和 Speed 提出的模型目的不同 [2]。简而言之，Nolan 和 Speed [2] 设计的案例研究有两种用途：（i）用于讲座中的开放式讨论，或（ii）作为开放式实验练习，学生在课外进行大量分析，并撰写报告，记录观察结果和解答。在这两种用途中，案例研究都是开放式的；背景内容可以先在课堂上讨论，也可以作为作业的一部分，但学生会独立或以小组形式开展工作，提出自己的解答，并在完整篇幅的报告中总结自己的发现，以回答最初的问题。相比之下，我们选择设计完整、深入的体验式指南，带领学习者走过数据分析的整个过程，并着重讲解计算 [22]，：从一个引人入胜的问题出发，最终总结结果。我们的目标是让教育工作者能够在课堂上直接使用完整的案例研究，或选取部分材料加以改编。例如，教育工作者可以选择展示案例研究中提供的解答、展示不同的解答，或让讨论保持开放式。我们提供完整指南的理由是，对教育工作者而言，删减或修改材料通常比从头创建材料更容易。通过这种方式，我们旨在服务比课堂教育工作者更广泛的受众，因为任何对某个主题感兴趣的学习者都可以通过案例研究，了解完整数据分析的实例。此外，这种方法对那些可能不太有信心从头开展分析的授课教师尤其有帮助，特别是在分析内容超出其主要专业领域的情况下；我们的案例研究由领域专家参与编写，并经过同行评审。

2.2 开放案例研究模型的组成部分

我们将以一个案例研究为例，介绍开放案例研究模型的十三个组成部分（**表 1**）。目前，我们的所有案例研究都展示了如何使用 R 统计编程语言 [23] 进行数据分析，不过我们的模型也可以使用其他编程语言。在这里，我们使用"探索二氧化碳$_2$ 排放随时间的变化"案例研究（opencasestudies.org/ocs-bp-co2-emissions），该研究探讨了从 18 世纪到 2014 年全球及各国的二氧化碳（$_2$）排放情况（**图 1**）。该案例研究还调查了二氧化碳$_2$ 排放率与美国（US）不断升高的温度及自然灾害发生率之间可能存在的关系。我们还介绍另外四个案例研究（**表 2**），并给出

两个案例研究模型组成部分的对应关系			
开放案例研究模型		Nolan 和 Speed 的案例研究模型 [2]	
组成部分	**描述**	**组成部分**	**描述**
1. 动机	案例研究开篇的引导性图示和文本	引言	介绍科学问题的背景 和动机
2. 主要问题	案例研究 科学问题		
3. 学习目标	数据科学和统计学的学习 目标	背景	用于提出问题的信息 使用非专业语言的背景 语言
4. 背景	问题或数据的背景		
5. 局限性	案例研究或所用 使用的数据		
6. 数据是什么?	数据来源总结 以及数据包含什么	数据描述	收集数据的文档 以回答问题
7. 数据导入	数据导入分析	调查	回答方面的建议 问题的建议（难度各异）
8. 数据整理与 ex- 探索	用于数据整理和探索的分析 数据		
9. 数据可视化	用于数据可视化的分析		
10. 数据分析	包含统计概念的分析 和用于回答问题的方法	理论	描述相关统计 概念和方法论，以回答 问题
11. 总结	结果总结	扩展材料 （可选）	介绍高级分析或 相关理论材料
12. 建议作业	可进一步探讨的问题		
13. 补充信息	有用的链接或使用的软件包		

表 1. 开放案例研究的组成部分 我们开放案例研究模型各组成部分的描述（左），以及这些组成部分与 Nolan 和 Speed 提出的案例研究模型中的组成部分之间的对应关系 [2]（右）。我们注意到，Nolan 和 Speed 的模型 [2] 将“数据描述”排在“背景”之前。不过，为了更容易与我们的开放案例研究模型对应，这里将背景列在前面。

所有案例研究涵盖的主题示例（**表 S1**）。

1. 动机。每个案例研究都以一个用于激发兴趣的数据可视化开篇。这个想法源自 Mine Cetinkaya-Rundel 博士题为《先让他们吃蛋糕！》的演讲，该演讲于 2018年哈佛大学统计系的 David K. Pickard 纪念讲座 [24] 上发表。她认为，就像一本介绍如何烘焙蛋糕的食谱，先向学习者展示可视化图表，能够激发他们的兴趣，并让学习者了解接下来要做什么。在数据分析开始时先展示可视化图表，然后向学习者展示生成该数据可视化的代码，这种做法能让学习者更清楚地了解最终成果，也能激励他们学习制作可视化图表所需的、更具挑战性的概念。

CO_2 排放案例研究中的**图 1**展示了用于说明动机的图。本案例研究中，我们还加入了阐述案例研究动机的文本。我们的案例研究通常由近期发布的报告或出版物所引发，这些报告或出版物探讨了特定的科学问题。在本节中，我们将解释该主题为何值得关注，并定义理解主要问题（将在下一节介绍）所需的术语。

2. 主要问题。在本节中，我们会在开始分析之前，突出并明确提出一组精确的科学问题或问题 [25]。当案例研究由先前的出版物引出时，这些问题可能与原论文或报告中最初探讨的问题不完全相同。例如，案例研究可能只探讨报告或出版物中所呈现结果的一小部分。或者，案例研究可能根本不探讨相同的问题，而是利用报告或出版物中的数据来展示特定的数据科学或统计学学习目标。该框架还重申，许多问题在分析开始前就已有一组待解答的问题；不建议先寻找答案，再事后设计问题。数据探索是分析框架的重要组成部分，也会在案例研究中体现，但 OCS 强调，应在分析之前审慎确定问题。

在 CO_2 排放案例研究中，科学问题如下：

1. 全球 CO_2 排放率随时间发生了怎样的变化？尤其是在美国，这些变化如何？美国与其他国家相比如何？
2. 美国的 CO_2 排放、全球气温和美国的自然灾害发生率之间是否存在关联？

3. 学习目标。每个案例研究都包含一组教学学习目标。我们将每个目标归类为（i）数据科学或（ii）统计学相关目标；后者涉及统计学课程中通常教授的概念，例如线性回归、多重检验校正和显著性，而前者则是

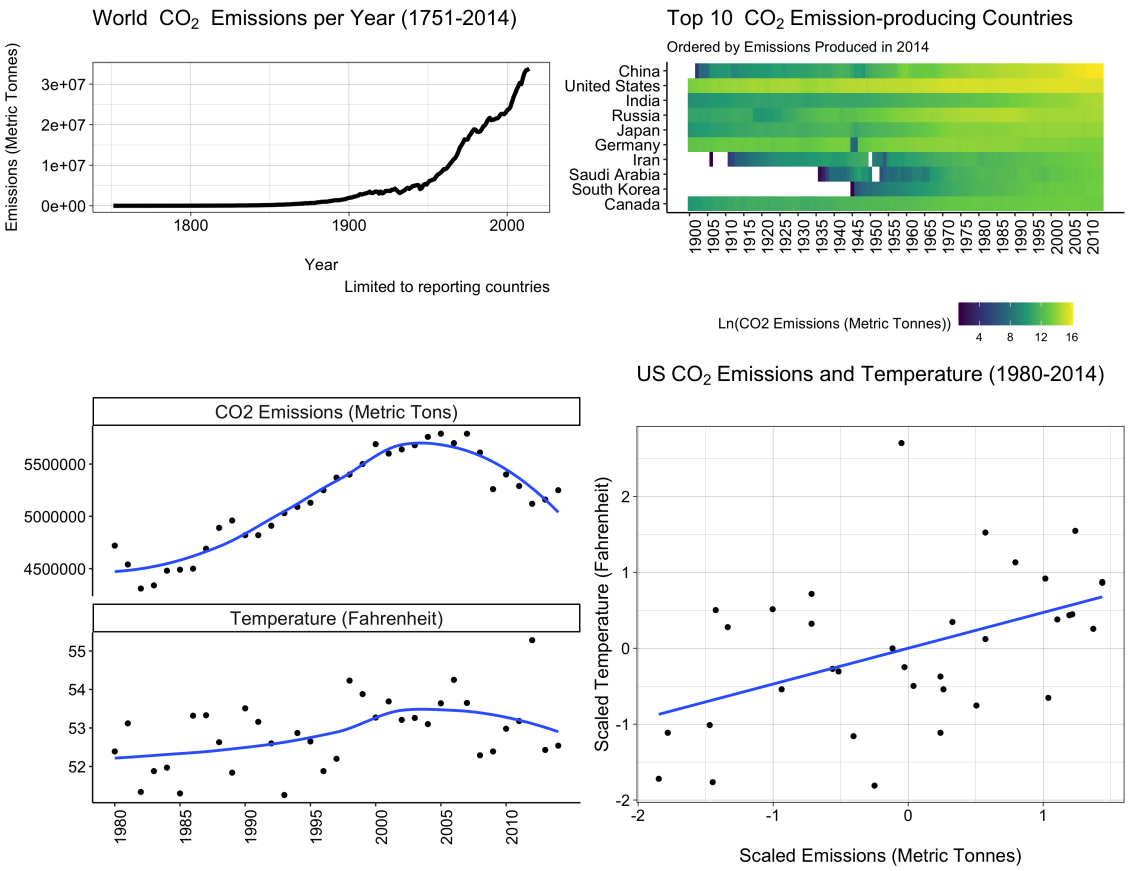


图1. “探索不同时间段的二氧化碳$_2$ 排放”案例研究中的动机图示例 完整案例研究请见（opencasestudies.org/ocs-bp-co2-emissions）。 **上排**：折线图展示二氧化碳$_2$排放量随时间的增加情况（左）。纵向热图突出显示美国历史上和目前一直是二氧化碳排放量最高的国家之一（右）。 **下排**：散点图展示二氧化碳$_2$排放量与温度随时间变化的趋势。

经常出现在传统统计学课程之外的概念，例如重新编码数据值、从网站抓取数据，或为数据集创建仪表板。根据案例研究的目的，还可以考虑其他类别。这种区分也使教育工作者能够将材料调整到除 R 之外的其他计算框架和语言，例如 Python。

我们列出这些学习目标有三个原因：(i) 帮助教育工作者选择符合其教学目标的案例研究；(ii) 帮助学习者选择能够展示其所希望学习内容的案例研究；(iii) 让教育工作者和学习者都能清楚了解特定案例研究的目标。例如，一项关于在本科科学课程中使用学习目标的研究发现，学生认为学习目标有助于缩小学习范围并安排学习内容 [26]。

对于 CO_2 排放案例研究，我们围绕以下学习目标设计了该案例研究：

(i) 数据科学学习目标：

- 从各种类型的 Excel 文件和 CSV 文件导入数据
- 运用 dplyr 中的动作动词 [27] 进行数据整理
- 如何在“长格式”和“宽格式”数据集之间转换
- 使用 dplyr 合并多个数据集
- 如何使用 ggplot 创建有效的纵向数据可视化2 [28]
- 如何为 ggplot2 图表添加文本、颜色和标签
- 如何创建分面 ggplot2 图

(ii) 统计学习目标：

- 相关系数作为汇总统计量

OCS资源中的案例研究示例

主题	问题	数据源	原始数据	数据科学技能	统计概念
空气污染 [超文本标记语言]	我们能否预测年度细颗粒物空气污染浓度，使用预测因子，例如人口密度，城市化，和卫星数据?	重量法测定的美国环境保护署空气污染数据（来自2008年）和预测变量数据来自 NASA、美国人口普查局，和美国国家卫生统计中心	单一整理后的 CSV file	tidymodels、相关性可视化，地理空间可视化	机器学习、线性回归、随机森林
电子烟使用 [超文本标记语言]	烟草／尼古丁产品在美国青少年中的使用情况自 2015 年以来发生了哪些变化?是否存在某种关系电子烟使用与电子烟使用及其他烟草／尼古丁产品 use?	2015—2019年全国青少年烟草调查 data	Excel 文件 and 代码本每年	导入Excel文件，导入多个文件高效地合并数据，编写函数，函数性的编程，纵向可视化	调查加权，调查加权逻辑回归调查加权，纵向数据，代码本
CO_2 排放 [超文本标记语言]	全球 CO_2 如何排放率发生了怎样的变化随时间推移? 尤其是对于美国，以及如何做到、】【美国与其他国家相比如何? // CO_2 美国的排放量、全球温度，以及自然美国的自然灾害发生率是否相关?	CO_2 排放（来自1751—2019年，国内生产总值和来自能源使用数据 Gapminder。美国温度和灾害来自美国国家海洋和大气管理局的数据	XLSX文件和 CSV 文件	从以下来源导入数据 Excel 文件和 CSV 文件，数据连接，纵向数据可视化、图表带有文本和标签	相关系数，.....之间的关系相关性与线性回归、相关性与因果关系
美国学校枪击事件 [超文本标记语言]	每年的校园枪击事件发生率以及在美国哪些地区曾发生过这些事件过去50年（从1970年1月到2020年6月）?	开源K-12校园枪击事件数据库（1970—2020年）	单个CSV文件文件，Google 表格	导入 Google 表格，日期格式、地理编码，交互式表格、R Markdown、地图，交互式仪表板	计算百分比针对存在缺失值的数据

表 2. OCS资源中四个案例研究示例的描述此表展示了四个独立案例研究涵盖的主题，以及原始数据的相关信息。EPA = 美国环境保护署，NASA = 美国国家航空航天局，NCHS = 美国国家卫生统计中心，NYTS = 全国青年烟草调查，NOAA = 美国国家海洋和大气管理局。CO_2 排放数据取自Gapminder，最初来自世界银行。校园枪击数据取自海军研究生院（NPS）的国土防卫与安全中心。

- 相关性与线性回归之间的关系
- 相关性不等于因果关系

此外，在案例研究中阐明这些学习目标，可以帮助学生开始了解如何在未来的分析中应用这些概念。最后，我们在开放案例研究网站（opencasestudies.org）上提供了一个可交互的学习目标搜索表，方便查找能够展示教师或学习者可能感兴趣的特定技术、方法或概念的案例研究。

4. 背景。背景部分提供必要的背景信息，以帮助读者理解所关注问题的背景，以及用于回答这些问题的数据 [15, 16]。其中可能包括案例研究所依据的出版物中的信息，也可能包括其他背景文献。以公共卫生为例，案例研究可以描述目前对该主题健康影响的认知（或未知之处）。这有助于说明具体问题如何融入更广泛的科学背景。

对于 CO_2 案例研究，背景部分讨论了气候变化可能造成的影响对人类健康的影响，概述了未来几年变暖可能的发展趋势，以及对环境其他组成部分的潜在影响，例如海洋酸度和降雨量。

5. 局限性。除了阐述每个案例研究的动机和背景外，还应正式说明所呈现分析的局限性，因为这能为教育工作者或学习者提供重要背景 [7]。局限性的示例包括：(i) 可用数据带来的局限性，例如使用替代变量或指标；(ii) 所用方法的局限性，例如对可能每天或每月变化的数量采用年度平均估计值；(iii) 对观测数据进行抽样所造成的选择偏差。数据科学的一个关键概念是，分析得出的结论好坏取决于所用数据和分析方法，因此呈现这些局限性是一次宝贵的学习机会。

在 CO_2 案例研究中，我们描述了数据存在局限性：由于只有部分国家报告了CO_2 排放量，且仅涵盖某些年份，因此数据并不完整。我们还说明了其他排放量的来源。

这些排放量来自未纳入数据的国家。这有助于学习者理解，尽管数据能帮助我们了解 CO_2 排放量，但无法呈现全貌。

6. 什么是数据? 为使数据来源透明，我们会说明原始数据的获取方式和使用方法。如果数据来自网站、调查或报告，并且条件允许，我们还会说明这些数据最初是如何收集的。通常，我们会在案例研究后面的部分介绍每个数据集中的变量，以便学习者在导入并探索数据后，能更好地了解这些变量。

CO_2 案例研究的数据来源包括 Gapminder（gapminder.org）（最初来自世界银行）以及美国国家海洋和大气管理局。在案例研究中，我们会用表格列出不同的数据来源，并简要描述每个来源，包括应予引用的来源。

7. 数据导入。接下来，我们会介绍将原始数据读入分析环境所需的步骤并提供相应代码。目前，我们的所有案例研究都使用 R 编程语言进行分析。有些情况下，导入原始数据相当简单，因此本节篇幅较短。其他案例研究的数据导入部分则更长、涉及的步骤也更多，例如从 PDF 抓取数据、通过应用程序编程接口（API）访问数据，或编写函数以高效访问多个格式相同的文件。重要的是，我们以读写式编程的方式介绍案例研究中所有代码的用法 [29]，，也就是说，我们会逐步进行说明，确保学习者能够理解。

由于 CO_2 案例研究的数据存储在 Excel 和逗号分隔值 (CSV) 文件中，我们使用标准数据导入函数：readxl 包中的 read excel() 和 R 的 readr 包中的 read csv()，来导入数据 [30] 。

8. 数据整理与探索。在我们的许多案例研究中，数据整理部分通常是篇幅最长的部分之一，其中介绍了将已导入的原始数据处理成可用于分析和创建可视化的状态所需的所有步骤。我们还演示了如何进行探索性数据分析 [31]。

例如，在 CO_2 案例研究中，需要将原始数据从“宽格式”转换为“长格式”，使每个国家—年份观测都占据单独一行。对来自各个来源的数据进行整理后，我们将演示如何通过匹配国家—年份组合，将不同来源的数据集连接起来。最终，我们会创建一个大型数据集，其中包含分析所需的所有变量（列），并为每个国家—年份组合（行）保留一条记录。

9. 数据可视化。我们展示了简单和复杂的数据可视化，以探索并展示各种图形设计选择，包括图表类型及其他美学选择，从而最佳地呈现我们感兴趣的变量类型。此外，大多数案例研究还会介绍如何分面或组合图表，以便在一张数据可视化中呈现该案例研究的所有主要发现。

在 CO_2 案例研究中，我们针对部分变量创建数据可视化。例如，我们使用折线图来展示全球 CO_2 排放量（以公吨计）随时间的变化（**图 1**）。我们详细介绍如何为折线着色和添加标签、如何在时间尺度轴上缩放，以及如何添加信息丰富的图表标题和坐标轴标签。我们展示了在查看不同国家的 CO_2 排放量随时间的变化时，为标签做特别处理的必要性。我们还展示了热力图或瓦片图如何以不那么繁杂的方式，出色地呈现各国之间的主要差异（**图 1**）。我们还展示了分面图如何同时可视化更多变量随时间的变化。此外，我们还展示了如何通过添加平滑线或线性回归线的散点图，开始寻找数据中的关联或趋势。

10. 数据分析。我们的案例研究旨在介绍如何实施和解读特定的统计检验或数据科学技术，以回答感兴趣的科学问题。不过，我们会引导学习者经历一个意料之外的结果，并说明我们如何诊断这一结果 [19]。我们提供有关统计概念的背景信息，并介绍这些概念如何应用于示例分析。

我们的 CO_2排放 Open Case Study 分析部分的主题是相关性，以及相关性与线性回归的关系。我们讨论相关背景信息，例如汇总统计量的定义、相关系数的含义，以及相关系数的数学计算方法。我们还介绍相关性分析的局限性，以及相关性并不意味着因果关系。我们展示如何实施相关性评估，以及如何解读结果。

11. 总结。本节提供一幅总结图，以直观呈现该案例研究的一些主要发现。这幅可视化旨在展示如何向更广泛的受众传达分析结果 [6]。这通常需要组合图表并添加注释。这幅总结图是在案例研究开篇使用的引导图。除了这幅图，我们还提供了该案例研究的概要，其中重述并总结了研究的动机、背景和

科学问题，同时介绍数据整理、数据探索和分析的主要步骤。我们讨论了分析的主要发现，并着重说明这些发现对于科学问题的更广泛背景可能意味着什么，以及仍有哪些未知之处。

在 CO_2 排放 Open Case Study 中，总结图（**图1**）结合了案例研究中的若干图表，以展示主要发现。概要回顾了我们使用的数据（1751—2014 年部分国家的 CO_2排放数据）以及分析中得出的结论，其中也涉及开篇列出的学习目标。我们对分析部分讨论的统计概念作了更简单的解释，在本案例中，这些概念是相关性和回归。我们进一步讨论了自己能够或无法回答哪些感兴趣的问题。我们介绍了如何发现全球 CO_2 排放量随时间大幅增加，以及一些国家似乎对此负有特别大的责任。我们还指出，尽管数据表明美国的温度与 CO_2排放之间存在关联，但根据我们对气候变化的了解，还有许多其他重要因素需要考虑，包括：大气层中来自其他国家的 CO_2 排放的影响、其他温室气体的影响、大气层中已有的 CO_2 会持续多年蓄热这一事实，以及以往排放造成的海洋蓄热会导致地表温度发生滞后变化这一事实。我们还指出，分析结果可能会对我们应如何考虑减缓气候变化影响，以及气温变暖未来可能如何影响社会产生何种启示。

12. 建议作业。每个案例研究都会为学生提供一项建议作业，供他们独立尝试
这些活动通常要求学生运用已学技能处理新的数据集，或扩展分析以评估数据的另一个子集。也可能会要求学生根据这些分析制作可视化。

针对二氧化碳排放 Open Case Study，建议的作业如下：

- 制作一张带标签的图表，展示二氧化碳排放量最低的国家。
- 在散点图上绘制二氧化碳排放量和其他变量（如能源使用），计算皮尔逊相关系数，并讨论结果。

这些建议能让学习者练习可视化和分析技能，在较少指导的情况下进一步探索数据。

13. 补充信息。本节介绍有关该案例研究所涉及的更广泛科学主题、用于分析数据的方法，以及分析中使用的具体数据集的补充信息。相关信息以外部在线资源链接的形式提供，包括博客文章、科学论文、科学报告和教育网站。我们还提供所用 R 包的文档链接，以及所用软件包的具体版本信息。此外，我们也链接到参与案例研究开发的相关领域专家的信息。

CO_2 排放 Open Case Study 提供了相关资源链接，帮助读者进一步了解案例研究中使用的各种 R 包（例如 here [32], readxl[33], readr [30], dplyr [27], magrittr [34], stringr[35], purrr [36], tidyr [37], tibble [38], forcats[39], ggplot2 [28], directlabels [40], ggrepel [41], broom [42], patchwork [43]），以及这些软件包的使用方法；同时还介绍了案例研究涉及的统计主题，包括相关性、回归和时间序列分析。这些内容超出了案例研究中呈现的部分材料，旨在为授课教师或学习者提供更多感兴趣主题的相关资源。

3 OCS教育资源

OCS资源可在线查阅（opencasestud-ies.org）。此外，我们还编写了教育者指南，介绍如何最有效地使用案例研究、如何在课堂上修改和调整案例研究的组成部分，以及如何贡献新的案例研究。（opencasestudies.org/OCS Guide）。

3.1 开放案例研究网站与搜索工具

我们的案例研究资源托管在开放案例研究（OCS）网站上（**图 2**）。为了方便浏览案例研究，我们提供了一个交互式搜索表，该表使用 `DT` 软件包 [44], 构建，让感兴趣的读者可以按主题、统计学习目标、数据科学学习目标以及所展示的 R 软件包来搜索案例研究。该表包含每个案例研究的代码和数据链接，以及案例研究网页版本的链接，读者可以在这些网页上完整阅读整个分析过程。

3.2 GitHub上的开放案例研究

每个案例研究的代码和数据都托管在 GitHub代码仓库中（**图 2**）。我们的案例研究使用 R Markdown 构建，文本、图像和描述背景及数据分析过程的动图可以与展示分析中实际所用代码的代码块交错呈现 [29]。这些案例研究是在 Quarto发布系统发布之前开发的（quarto.org/quarto）。随后，我们使用 GitHub Actions 将这些案例研究“编织”成渲染后的 HTML 格式文件[45]，以实现持续集成和部署。持续集成是指对所做的更改进行跟踪

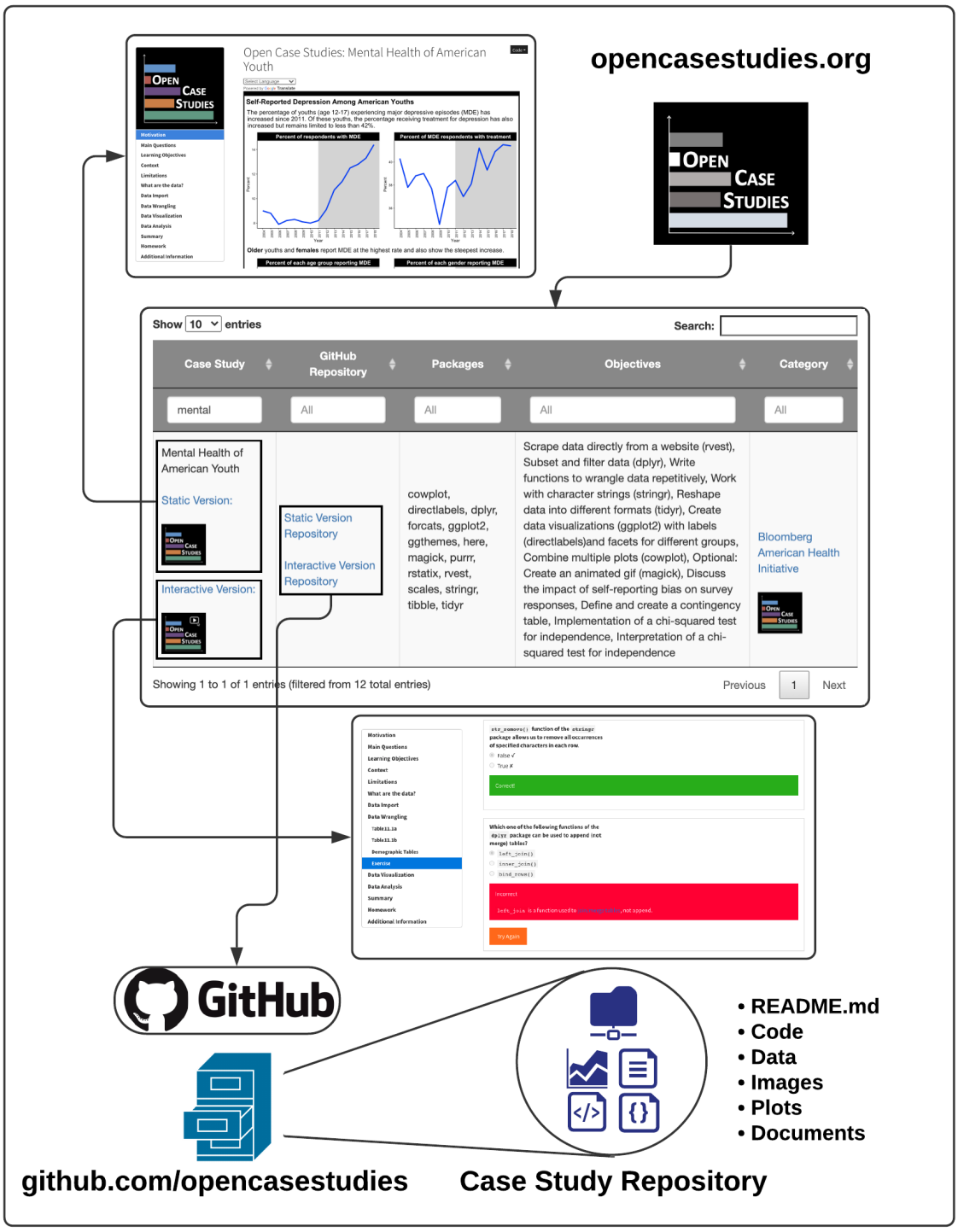


图 2. OCS 教育资源概览 开放案例研究网站包含一个可搜索的数据库，收录了所有可用的案例研究。用户可以按案例研究名称、所用 R 包、学习目标和类别进行搜索。每个案例研究都链接到一个网站，其中提供完整分析的呈现版本，并链接到 GitHub 代码仓库。GitHub 代码仓库存放在线课程，以及跟随学习或开展新分析所需的所有相关代码、数据、图像、图表和文档文件。部分案例研究现已提供交互式版本，其中包含实时测验和编程教程。

并且，借助 Git 和 GitHub 保存来自不同作者的代码历史记录，并将其汇总到一个主版本中。 [46] 通过持续部署，我们指的是：一旦在 GitHub 上建立新版本，案例研究的网站版本就会自动生成并向公众开放。这些案例研究的网站版本也托管在 GitHub 上。目前，我们的所有案例研究都使用 R 编程语言编写，不过，当前格式也可以扩展，以支持使用其他编程语言编写的教程。我们的案例研究设有目录，方便教师和学习者轻松浏览各个章节，从而专注于最符合自身需求的材料。此外，每个案例研究都以一幅图形或图表开篇，用于展示该案例研究的主要发现。所有案例研究都采用相同的基本结构，方便学习者更轻松地浏览案例研究，并了解不同案例研究中分析方法的规律（**图S1**）。

3.3 OpenCase 案例研究文件结构

每个案例研究代码库都具有相似的文件结构，其中的数据目录包含原始数据，以及经过各种处理的数据版本，以便授课教师/学习者根据自身需求，将案例研究模块化（**图 3**）。例如，授课教师可以跳过案例研究中的数据导入和数据整理部分，专注于使用完全清理好的数据集进行可视化和分析。为支持这种模块化教学方式，每个案例研究都会在各部分开头提供命令，以导入上一部分结束时的数据状态。数据的这些不同阶段按五类整理在一个数据文件夹中：原始、已导入、已整理、简易导入和额外数据。原始数据目录包含原始数据源提供的原始文件，保持其原始状态和文件格式（某些情况下，原始文件可能是 CSV 文件、Excel 文件、PDF 文件或其他格式）。导入数据目录包含以与 R 直接兼容的格式存储数据的文件，例如 RData 文件（通常简称为 Rda）。整理后数据目录也包含一个 RData 文件，其中存储了经过预处理、整洁干净且可用于分析的数据版本；此外，还包含 CSV 文件，供希望演示更简易数据导入方式的授课教师使用。简易导入文件夹包含已转换为 CSV 文件格式或其他更易导入 R 的格式的原始文件。额外数据文件夹包含数据文件，供个人开展超出案例研究范围的分析（这些额外文件的格式各不相同）。每个代码库还包含一个 README 文件 [20, 21] ，用于说明案例研究的模块化特点，以及如何将案例研究用于教学等其他信息（**表 S2**）。

3.4 OpenCase 案例研究中的交互元素。

为了让案例研究更具体验性，我们使用 learnr [47] andgradethis[48] 软件包引入了测验和编程练习等交互元素。

每个案例研究中都包含选择题和编程练习。编程练习穿插在案例研究内容中，让学生有机会为分析中的特定步骤编写代码。练习的答案（即案例研究中用于完成这些步骤的代码/输出）会隐藏在练习窗口后方可点击展开的部分中。学生可以将自己的代码和输出与这些答案进行比较。我们还会在案例研究各主要部分的末尾设置练习小节，其中既有选择题，也有编程练习。学生可以通过这些练习检验自己对各节内容的理解。所有选择题都会提供实时反馈，在答错后给出提示，并允许学生重试。对于大多数编程练习，都提供提示和/或解答。借助 gradethis 程序包，其中一些编程题还会在学生提交代码后提供实时反馈。

4 构建自己的案例研究

对于有意构建自己的案例研究的教育工作者，我们将在下一节根据本项目中的经验和挑战，介绍我们对这一过程的建议。我们也在《教育者指南》（open-casestudies.org/OCS Guide）中介绍了这些建议。

4.1 为案例研究确定问题和数据

选择数据来源和感兴趣的问题，可以说是构建案例研究过程中最重要的部分。我们可以先找到一个有趣且公开可用的数据集，再围绕与该数据相关的主题提出一个切合时宜且引人入胜的问题；也可以先确定一个有趣的问题，然后寻找公开可用的数据来回答这个问题。将问题与公开可用数据联系起来的过程，往往需要一些反复试错，并在牢记案例研究旨在展示什么的同时，对问题进行调整，必要时也调整案例研究的目标。

在开发案例研究的过程中，我们发现，先确定一个数据集，往往比依靠寻找数据集来回答某个特定问题更容易。虽然我们的许多案例研究都是专门为应对公共卫生挑战而设计的，但有时我们很难找到适合所关注问题或一组问题的公开可用数据

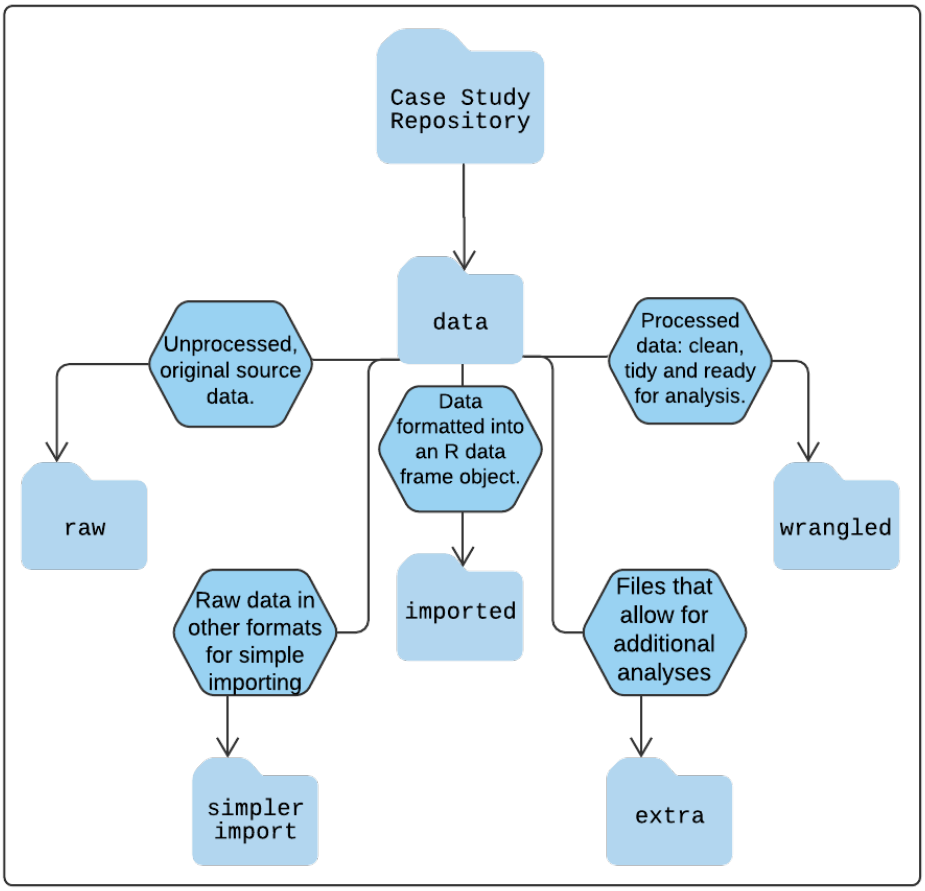


图 3。 GitHub 上数据 文件结构概览 展示数据目录结构的仓库树状图。每个气泡说明子文件夹中可找到的数据文件类型。

。与领域专家合作，对解决这一挑战尤其有帮助。在开展案例研究时，我们与公共卫生专家合作，既确定有趣、及时且可检验的问题，也寻找能够回答这些问题的公开数据来源。

我们发现，可以将获取标准格式（例如 Excel、CSV）的数据这一难题作为教学机会；同时，对数据来源持开放态度，也让我们能够展示非常规技能。例如，当我们无法轻易访问已发表报告中表格所存储的数据时，我们演示了直接从 PDF 中提取数据这一数据科学技能。作为未来的数据科学家，学生需要具备灵活获取数据的技能，因为这些数据未必能够直接读入 R，或以原样导入 R。

虽然我们通常会先确定一系列数据科学和统计学习目标，再着手开发每个案例研究，但有时我们会陷入两难：一方面要找到能够实现这些特定目标的数据集，另一方面又要让数据引导案例研究的方向。我们发现，顺应数据本身所提供的机遇，会让我们举出的例子更贴近现实世界中的数据分析情境。我们在案例研究中记录了其中一些挑战，以便学生更好地理解寻找合适数据来回答感兴趣问题的过程（以及可能需要重新聚焦问题的情况）。尤其是局限性一节，为课堂讨论提供了最有用的材料之一，可用于探讨数据能够和不能回答的问题类型，以及我们有时必须简化分析，以反映现有数据的局限性。

作为在研究领域围绕性别和种族议题进行反思与社会变革的时期工作的教育工作者，我们也特别注意指出数据集中一些历来遭到忽视的方面。例如，使用调查收集数据时，如果调查在族裔、种族或种族与性别交叉方面提供的选项有限，就会限制我们准确呈现受研究人群多样性的能力。关于青年脱节的案例研究（opencasestudies.org/ocs-bp-youth-disconnection）可供读者参考。

对于一些案例研究，我们着重寻找相对干净且完整的数据，以便展示某些概念，例如机器学习或如何创建仪表板。在这些案例研究中，我们知道分析内容会相当深入且篇幅较长，因此特意寻找能够让我们迅速着手的数据集，避免在收集、清理和导入数据方面遇到太多困难。

O整体 ll 建议 tions for starti一个案例研 tud究的要点是:

- **对数据来源保持开放和灵活：**与进行真实分析时不同，分析人员可能会选择避开访问数据时遇到的复杂问题（如果可以选择更容易获取的数据集）。但这类复杂问题可以带来教学机会，让学生为将来遇到无法选择更简单方案的情况做好准备。

- **根据案例研究的目标确定灵活程度：**如果案例研究旨在展示某种特定的统计方法或数据导入方法，那么可能需要投入更多精力来寻找符合这一特定教学预期的数据。在我们的案例中，我们知道自己计划编写多个案例研究，因此允许其中一些案例研究自然地朝着我们最初未曾设想的方向发展。这最终带来了一些我们未曾预料到的教学机会。不过，对于其中一些案例研究，我们对数据需求的要求更为严格。

- **考虑案例研究的范围：** 请牢记：1）案例研究面向的学习者类型；2）案例研究旨在展示的数据分析方法目标。尽量避免选择一个既需要大量数据导入/数据整理、又需要大量数据分析的案例研究。在后续阶段，可能需要重新评估案例研究的整体方向和范围。如果案例研究过长，可以考虑将其拆分为多个案例研究。

- **保持简单：** 以初学者水平解释一个过程时，通常需要在案例研究中占用比预期更多的篇幅。使案例研究计划保持简单，有助于应对可能出现的意外教学机会，因为这些机会可能需要更多讲解。

4.2 先进行分析，但要心系学习者

要呈现分析叙述，必须先进行分析，再着手撰写叙述性描述。然而，案例研究本身不应只是对数据分析过程的复现，而应通过简化和调整，为学生呈现清晰连贯的内容。为此，务必妥善记录初步分析期间采取的所有步骤，包括解释和注释，以说明分析过程中所做选择的理由。尤其应记录原始数据的确切获取方式。

我们自己通常进行分析的方式，往往并不总是最适合教学。例如，一位经验丰富的数据分析师可能会先编写一个函数，用于对多个相似的数据文件进行数据整理。然而，这并不适合用作面向初学者的案例研究的起点。相反，可以先选择一个特定文件，对其进行非常详细的数据整理，然后再尝试将代码概括为一个函数。因此，在真正生成描述这一过程的对话之前，我们会先根据目标受众的水平，确定整体的数据导入和数据整理流程。

我们还发现，案例研究中的数据探索步骤，以及如何整理数据这一决策过程所涉及的步骤，往往需要简化。例如，我们最终可能会决定从分析中移除某个数据源，因为我们发现数据中存在错误，而处理这些错误超出了目标受众的范围。虽然向学生介绍这些数据错误[19] 以及如何处理这些错误可能很有用，但我们也需要保持适当的细节程度，以免让他们不堪重负。

另一种可能需要修改分析的情况是，某个过程需要大量反复试错。与其向学生展示反复试错的所有迭代过程以及其中的所有决策，我们可能只演示其中一小部分，以免案例研究篇幅过长。例如，在一个关于机器学习的案例研究中，我们希望达到一定的性能水平，因此花了相当多的时间演示如何优化和调优参数。虽然我们简要介绍了调优策略，但没有展示所有中间模型，而是最终选取了两个有意思且有助于说明参数调优的模型。

总而言之，我们可能在自己的分析中经历了一个学习过程，最终得出了更完善的方法。有时，我们不会描述如何一步步走到这一步，而是直接呈现最终方法，同时在叙述中说明，实际操作中还需要投入更多精力。我们既希望如实呈现数据分析过程，也需要做到清晰明了、重点突出。

4.3 撰写案例研究叙述

分析人员完成分析并解答相关问题后，就可以开始撰写叙述了。首先，我们通过介绍与案例研究所探讨的具体问题相关的研究，来引出并阐明主题的重要性。

首先，我们介绍数据导入、数据整理和分析流程。如上所述，这可能无法忠实再现我们自己的分析过程，而是会重新编排，以尽可能满足案例研究的教学目标。至于补充的叙述，我们

会尽力引导学生理解我们所呈现的新信息。首次使用某个函数时，我们会说明它的作用、主要参数以及所属的软件包。我们会阐述从一个步骤到下一个步骤的决策过程，有时也会举例说明尝试某种方法却未能奏效的情形，以反映现实世界数据分析的实际情况。

我们还会尽可能通过点击展开的部分来说明术语和背景信息，以免打断案例研究的整体行文。例如，展开的部分会解释“管道操作”如何运作，即如何让对象依次经过一系列步骤。这样既不会拖慢已经熟悉管道操作的学生，也能让从未接触过管道操作的学生跟上。其他适合放在此类可展开部分中的内容包括：介绍 `ggplot2` 软件包的“图形语法”，或在进行统计检验前提供统计学背景信息。有时，我们会在另一项案例研究中详细阐述某个概念，因此会链接到那里的描述；但总体而言，我们至少会在每项案例研究中简要介绍大多数概念和方法，使其尽可能自成一体。同样，我们发现，加入部分 RStudio 速查表 [49] 对某些主题很有帮助，例如介绍正则表达式或连接函数。有时，我们发现，先用一个较小的数据集（已导入 R 或由我们在 R 中创建）来举例，更容易解释某个概念或难点。对于可能已经熟悉这些概念的学生，这些内容也会放在点击展开的部分中。

在构建叙述时，我们会考虑在哪里加入提问环节。这些环节可以让教师发起讨论，探讨分析决策过程，例如为什么某种图形选择并非总是有效，或者为什么某种数据整理方法可能无法复现。我们也可以提示学生回忆如何完成案例研究中先前演示过的任务。如 3.4 节所述，我们的交互式案例研究还包括测验题和编程练习。

最后，我们通过总结如何传达分析的主要发现 [6] 来结束叙述。我们还会说明这些结果如何融入该领域更广泛的背景、其影响、研究的局限性以及仍有哪些未知之处。最后，我们会通读案例研究，列出其中展示过的所有资源。

在创建此资源的过程中，我们发现了各种挑战，也总结了克服这些挑战的策略，详见**表S3** 以及创建新案例研究的指南（补充说明 S1）。

4.4 创建交互式案例研究

我们还在部分案例研究中使用软件包（learnr [47] 和 gradethis [48]）加入了交互元素。这些软件包基于 shiny [50] 软件包构建，后者使 R 用户能够更轻松地创建 Web 应用程序。learnr 软件包允许用户创建选择题和编程练习，而 gradethis 则允许自定义学习者回答问题或完成练习时收到的反馈。

有两种方法可以实现这一点。一种方法是将每项练习分别托管为一个 Shiny 应用程序，然后使用内嵌框架（超文本标记语言中的“iframe”）将这些应用程序嵌入案例研究中。第二种方法是创建一个单独的应用程序，将练习整合到案例研究中（补充说明 S2 和补充说明 S3）。

5 讨论与结论

本文介绍了一种创建完全开源、经过同行评审且内容完整的案例研究的方法，旨在建立一个最佳实践示例档案，引导学生使用真实、复杂、杂乱且有趣的数据开展数据分析。由于案例研究便于浏览，并采用通用的模块化架构进行组织，授课教师可以在课堂上使用我们的档案，引导学生学习案例研究的任意部分。独立学习者也可以利用其中详尽的叙述、交互元素和完整的分析来学习。学生和学习者可以了解新主题，也可以重温某个案例研究，复习特定方法或技术的细节。我们的案例研究中的数据，以及其中讲解的数据分析和数据科学方法，可供授课教师用于本科生、研究生以及高中生的教学，涵盖统计学、公共卫生、编程和数据科学等多种主题。这使授课教师能够使用与当前公共卫生问题相关、因而能引起各类学生兴趣的数据，而无需耗费精力寻找此类数据或确定可以对其进行哪些分析。这样，授课教师便能专注于通过更具互动性的课堂讨论向学生提出挑战，并帮助学生了解数据分析所需的决策过程。

总而言之，OCS 提供了一个一致的框架，该框架以 [2], 开放为基础，同时还就如何教授相关材料提供建议。借助 OCS资源，教育工作者也可以创建自己的资源，并在回馈社区的同时扩展 OCS。我们相信，这些补充内容旨在弥合分析教育最后一公里中的鸿沟。

文后部分

作者贡献

根据 CRediT 系统列出贡献。

- 概念化：CW、MAT、LRJ、SCH
- 软件：所有共同作者
- 形式分析：所有共同作者
- 调查：CW、SCH
- 数据整理：CW, SCH
- 撰写——初稿：CW, SCH
- 撰写——审阅与编辑：所有合著者
- 可视化：CW, MB
- 监督：CW, MAT, LRJ, SCH
- 经费获取：CW, MAT, LRJ, SCH

致谢

我们谨向约翰斯·霍普金斯数据科学实验室（jhudatascience.org）表示感谢，特别感谢 Roger Peng、Jeff Leek、Brian Caffo 和 Jessica Crowell 对开放案例研究项目的支持和宝贵反馈。感谢 Ira Gooding 就如何将案例研究纳入 Coursera 平台提供反馈。此外，我们还要感谢所有评审案例研究的数据科学和统计学专家，包括：Shannon Ellis、Nicholas Horton、Leslie Myint、Mine C,etinkaya-Rundel、Michael Love 和 Christina P. Knudson，以及以下学生评审：Jensen Stanton、Tina Trinh 和 Ruby Ho。我们也要感谢以下主题评审：Roger Peng、Tamar Mendelson、Brendan Saloner、Renee Johnson、Jessica Fanzo、Daniel Webster、Elizabeth Stuart、Aboozar Hadavand、Megan Latshaw、Kirsten Koehler 和 Alexander McCourt。我们还要感谢 Ashkan Afshin 和 Erin Mullany，感谢他们向我们提供题为“探索与健康风险相关的饮食行为的全球模式”的案例研究所用数据。最后，感谢约翰斯·霍普金斯大学布隆伯格公共卫生学院生物统计学系最初为本项目提供资金。

资助

本出版物中报告的开放案例研究项目获得了彭博美国健康倡议于2019—2020年提供的高影响力项目资助，用于创建目前纳入该项目的大部分案例研究。约翰斯·霍普金斯大学教务长办公室于2020年提供的数字教育与学习技术加速（DELTA）资助，支持了交互式案例研究以及许多支持其使用的工具（例如搜索工具）的开发。开放案例研究指南的经费作为基因组数据科学社区网络（GDSCN）资助的延续部分提供。GDSCN通过美国国家人类基因组研究所向约翰斯·霍普金斯大学提供的一项合同（75N92020P00235）获得支持。JM获得的支持来自 Streamline Data Science、U24HG010263-01（ANVIL）、UL1TR003098（美国国立卫生研究院／国家转化科学促进中心）：机构临床与转化科学奖，以及 UE5CA254170。

利益冲突

合著者 Carrie Wright 和 Stephanie Hicks 从 Leanpub 上的一本书以及 Coursera 上的一门课程中获得版税。这本书和课程均名为“Tidyverse Skills for Data Science”，并且都采用了开放案例研究项目中的三个案例研究。

参考文献

[1] C. J. Wild 和 M. Pfannkuch。实证探究中的统计思维。《国际统计评论》，67(3):223–248，1999年。

[2] Deborah Nolan 和 Terry P. Speed。通过应用教授统计理论。《美国统计学家》，53:370，1999年。

[3] Stephanie C. Hicks 和 Rafael A. Irizarry。数据科学教学指南。《美国统计学家》，72(4):382–391，2018年。PMID：31105314。

[4]Sharon Lawner Weinberg 和 Sarah Knapp Abramowitz。运用主动案例研究方法，使一般性原则在课堂中生动呈现。《统计学教育期刊》，8(2):null，2000年。doi：10.1080/10691898.2000。12131290。
U
RLhttps://doi.org/10.1080/10691898.2000.12131290。

[5] Daniel W. Schafer 和 Fred L. Ramsey。数据分析的实践教学。《统计学教育期刊》，11(1):null，2003年。doi：10.1080/10691898.2003。11910692。

[6]Davit Khachatryan 和 Nathaniel Karst。口头表达的 V：提升统计学沟通技能的策略。《统计学教育期刊》，25(2):68–78，2017年。ISSN 1069-1898。doi：10.1080/10691898.2017.1305261。
U
R
L
h
ttps://www.tandfonline.com/doi/full/10.1080/10691898.2017.1305261。

[7]Roberto Rivera、Mario Marazzi 和 Pedro A. Torres-Saavedra。将开放数据纳入统计学入门课程。《统计学教育期刊》，27(3):198–207，2019 年。ISSN 1069-1898。doi：10.1080/10691898.2019.1669506。网址：
h
ttps://www.tandfonline.com/doi/full/10.1080/10691898.2019.1669506。

[8] Thomas Donoghue、Bradley Voytek 和 Shannon E. Ellis。大规模教授创意性与实践数据科学。《统计学与数据科学教育期刊》，29(sup1):S27–S39，2021年。doi: 10.1080/10691898.2020.1860725。

[9] Luna L. Sanchez Reyes 和 Emily Jane McTavish。易于理解的案例研究有助于数据科学中的学习与可重复性：以进化生物学为例。《统计学与数据科学教育期刊》，30(3):304–310，2022年。doi: 10.1080/ 26939169.2022.2099487。

[10] R. Romero、A. Ferrer、C. Capilla、L. Zunica、S. Bal- asch、V. Serra 和 R. Alcover。面向工程师的统计学教学：一种创新的教学经验。《统计学教育期刊》，3(1):5，1995年。ISSN 1069-1898。doi: 10.1080/10691898.1995.11910481。URL https://www.tandfonline.com/doi/full/10.1080/10691898.1995.11910481。

[11] Allison S. Theobold、Stacey A. Hancock 和 Sara Mannheimer。为数据密集型环境科学研究设计数据科学研讨会。《统计学与数据科学教育期刊》，29 (sup1):S83–S94，2021年。doi: 10.1080/10691898.2020. 1854636。

[12] Pip Arnold 和 Christine Franklin。什么样的统计学问题才是好问题？《统计学与数据科学教育期刊》，29(1):122–130，2021年。ISSN 2693-9169。doi: 10.1080/26939169.2021.1877582。URL https://www.tandfonline.com/doi/full/10.1080/26939169.2021.1877582。

[13]David L. Neumann、Michelle Hood 和 Michelle M. Neumann。《统计学教学中使用现实数据：学生对这项策略在统计学导论课程中的看法》。统计学教育研究期刊，12(2):59–70，2013。ISSN 1570-1824。doi: 10.52041/serj.v12i2. 304。网址 https://iase-web.org/ojs/SERJ/article/view/304。

[14] David Donoho。《数据科学的50年》。 计算与图形统计学杂志，26(4)：745–766，2017年。doi：10.1080/10618600.2017. 1384734。网址https://doi.org/10.1080/10618600.2017.1384734。[15] Beverly L. Wood、Megan Mocko、Michelle Everson、Nicholas J. Horton 和 Paul Velleman。《更新指南，更新课程：GAISE大学报告与面向现代学生的统计学导论》。 *CHANCE*，31(2)：53–59，2018年4月。doi：10.1080/09332480.2018.1467642。网址 https://www.tandfonline.com/doi/full/10.1080/09332480.2018.1467642。[16] 数据科学学科构想委员会：本科生视角、计算机科学与电信委员会、数学科学与分析委员会、应用与理论统计学委员会、工程与物理科学部、科学教育委员会、行为与社会科学及教育部，以及美国国家科学院、工程院和医学院。 本科生数据科学：机遇与选择。美国国家科学院出版社，2018年。ISBN 9780309475594。doi：10.17226/25104。网址 https://www.nap.edu/catalog/25104。

[17] Sean Kross 和 Philip J. Guo。在工业界和学术界教授数据科学：期望、工作流程与挑战。载于《*2019* 年 *CHI* 人机交互系统因素会议论文集》，CHI '19，第 1–14 页，纽约，美国纽约州，2019 年。计算机协会。ISBN 9781450359702。

[18] Lance A. Waller. 统计学与生物统计学系学术晋升中数据科学贡献的记录与评估。《美国统计学家》，72(1):11–19，2018年。doi: 10.1080/00031305.2017.1375988。

[19]Roger D. Peng、Athena Chen、Eric Bridgeford、Jeffrey T. Leek 和 Stephanie C. Hicks。诊断课堂中的数据分析问题。《统计学与数据科学教育期刊》，29(3):267–276，2021年。ISSN 2693-9169。doi: 10.1080/26939169.2021.1971586。URL https://www.tandfonline.com/doi/full/10.1080/26939169.2021.1971586。

[20]Lars Vilhuber、Hyuk Harry Son、Meredith Welch、David N. Wasser 和 Michael Darisse。面向大规模可复现性验证的教学。《统计学与数据科学教育期刊》，30(3):274–281，2022年。ISSN 2693-9169。doi: 10.1080/26939169.2022.2074582。URL https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2074582。

[21] Mine Dogucu 和 Mine C.etinkaya Rundel。可复现教学的工具与建议。《统计学与数据科学教育期刊》，30(3):251–260，2022年。ISSN 2693-9169。doi: 10.1080/26939169.2022.2138645。URL https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2138645。[22] Deborah Nolan 和 Duncan Temple Lang。统计学课程体系中的计算。《美国统计学家》，64(2):97–107，2010年。doi: 10.1198/tast. 2010.09132。[23] R 核心团队。*R*：用于统计计算的语言与环境。R 统计计算基金会，奥地利维也纳，2021年。URLhttps://www.R-project.org/。

[24] Mine Cetinkaya Rundel。先让他们吃蛋糕！[视频]。*YouTube*。（*2018*年*11*月*9*日），2018年。URLhttps://www.youtube.com/watch?v=RsVOrpXAPXo。

[25]Simmi K. Ratan、Tanu Anand 和 John Ratan。《研究问题的提出——循序渐进的方法》。《印度儿科外科医师协会杂志》，24(1):15，2019年3月。网址：https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6322175/。出版商：Wolters Kluwer – Medknow Publications。

[26] Bethany Osueke、Birook Mekonnen 和 Julie Dangremond Stanton。《本科理科学生如何利用学习目标进行学习》。《微生物学与生物学教育杂志》，19 (2)，2018年1月。ISSN 1935-7877，1935-7885。网址：https://journals.asm.org/doi/10.1128/jmbe.v19i2.1510。

[27] Hadley Wickham、Romain Francois、Henry Lionel 和 Kirill Müller。*dplyr*：数据操作语法，2022年。网址：https://CRAN.R-project.org/package=dplyr。

[28] Hadley Wickham。《ggplot2：数据分析的优雅图形》。Springer-Verlag New York，2016年。ISBN 978-3-319-24277-4。网址 https://ggplot2.tidyverse.org。[29] D. E. Knuth。《文学化编程》。《计算机杂志》，27(2):97–111，1984年1月。ISSN 0010-4620。doi: 10.1093/comjnl/27.2.97。网址 https://doi. org/10.1093/comjnl/27.2.97。[30] Hadley Wickham 和 Jim Hester。readr：读取矩形文本数据，2020年。网址 https://CRAN.R-project.org/package=readr。[31] John W. Tukey。数据分析的未来。《数学统计年刊》，33(1):1 – 67，1962年。[32] Kirill Müller。here：更简便地查找文件的方法，2020年。网址 https://CRAN.R-project.org/package=here。R 软件包版本 1.0.1。[33] Hadley Wickham 和 Jennifer Bryan。readxl：读取 Excel 文件，2022年。网址 https://CRAN.R-project.org/package=readxl。R 软件包版本 1.4.0。[34]Stefan Milton Bache 和 Hadley Wickham。magrittr：R 的前向管道运算符，2022年。网址 https://CRAN.R-project.org/package=magrittr。R 软件包版本 2.0.3。[35]Hadley Wickham。stringr：常用字符串操作的简洁一致封装，2022年。网址 https://CRAN.R-project.org/ package=stringr。R 软件包版本 1.4.1。

[36] Lionel Henry 和 Hadley Wickham。purrr：函数式编程工具，2020年。网址 https: //CRAN.R-project.org/package=purrr。R 软件包版本 0.3.4。[37] Hadley Wickham 和 Maximilian Girlich。tidyr：整理杂乱数据，2022年。网址 https://CRAN. R-project.org/package=tidyr。R 软件包版本 1.2.0。[38] Kirill Müller 和 Hadley Wickham。tibble：简易数据框，2022年。网址 https://CRAN. R-project.org/package=tibble。R 软件包版本 3.1.8。[39] Hadley Wickham。forcats：处理分类变量（因子）的工具，2021年。网址 https://CRAN.R-project.org/package=forcats。R 软件包版本 0.5.1。[40] Toby Dylan Hocking。directlabels：多色图的直接标签，2021年。网址 https://CRAN.R-project.org/package=directlabels。R 软件包版本 2021.1.13。[41] Kamil Slowikowski。ggrepel：使用 'ggplot2' 自动放置互不重叠的文本标签，2021年。网址 https://CRAN.R-project.org/ package=ggrepel。R 软件包版本 0.9.1。[42] David Robinson、Alex Hayes 和 Simon Couch。broom：将统计对象转换为整洁小表，2022年。网址 https://CRAN.R-project.org/ package=broom。R 软件包版本 1.0.0。[43] Thomas Lin Pedersen。patchwork：图表编排器，2020年。网址 https://CRAN.R-project. org/package=patchwork。R 软件包版本 1.1.1。[44] Yihui Xie、Joe Cheng 和 Xianying Tan。DT：JavaScript 库 'DataTables' 的封装，2021年。网址 https://CRAN.R-project.org/package=DT。R 软件包版本 0.17。[45] GitHub Actions，2021年。网址 https://github.com/features/actions。[46]Mojtaba Shahin、Muhammad Ali Babar 和 Liming Zhu。持续集成、交付与部署：方法、工具、挑战与实践的系统综述。IEEE Access，5:3909–3943，2017年。ISSN 2169-3536。网址 http://ieeexplore.ieee.org/document/7884954/。[47] Barret Schloerke、JJ Allaire 和 Barbara Borges。learnr：R 的交互式教程，2020年。网址 https: //CRAN.R-project.org/package=learnr。R 软件包版本 0.10.1。

[48] Garrick Aden-Buie、Daniel Chen、Garrett Grole- mund 和 Barret Schloerke。*gradethis*：为 *'learnr'* 教程中的学生练习提供自动反馈，2021年。网址https://pkgs.rstudio.com/gradethis/。

[49]*RStudio* 速查表，2022年。网址：https://rstudio.com/resources/cheatsheets/。

[50]Winston Chang、Joe Cheng、JJ Allaire、Carson Sievert、Barret Schloerke、谢益辉、Jeff Allen、Jonathan McPherson、Alan Dipert 和 Barbara Borges。*shiny*：*R* 的 *Web* 应用程序框架，2021年。网址：https://CRAN.R-project.org/ 软件包=shiny。R 软件包版本 1.6.0。[51]Desir´ee De Leon。适用于 *R Markdown* 网站的交互式教程窗口，2022年。网址：https://desiree.rbind.io/blog/learnr-iframes/。

补充材料

开放案例研究：通过现实世界应用开展统计学与数据科学教育

CarrieWright、Qier Meng、Michael Breshock、Margaret A. Taub、Leah R. Jager、John Muschelli、Stephanie C. Hicks*

*通信请联系shicks19@jhu.edu

目录

1. **补充表 S1-S3**。2. **补充图 S1-S5**。3. **补充说明 S1-S3**。

补充表格

Open Case Study 资源中主题的多样性		
问题类型	某件事如何随时间变化? 群体随时间如何变化? 某些群体或亚群体是否更脆弱? 如何预测新数据的结果? 变量之间是否相互关联?	群体之间如何比较? 配对群体之间如何比较? 某个事物在不同地区之间有何差异? 这会影响我的数据吗? 如何展示这些数据?
Data 类型	多个文件 CSV 网站 API 调查数据 / 代码本	PDF Excel 图像文本 Google 表格
数据整理方法	从 PDF 中提取数据 重新编码数据 数据修改 数据重塑	地理编码数据 数据合并 处理文本 重复性过程
Data 可视化	格式化表格 折线图 箱线图 热力图 缺失数据图	散点图 条形图 饼图 / 华夫图 相关性图 Maps
高级可视化	匹配绘图风格 直接组标签 图表注释 组合图表 交互式地图 向图表添加图像	分面图 突出显示组 图表误差线 交互式图表 交互式表格 交互式仪表板
分析概念 and 方法	t 检验 线性回归 曼-肯德尔趋势检验 卡方独立性检验 缺失数据百分比计算	方差分析 逻辑回归 机器学习 威尔科克森符号秩检验 威尔科克森秩和检验

补充表 S1。OCS资源中的示例主题 这些主题如何应用于案例研究示例，请参见**表 2**。

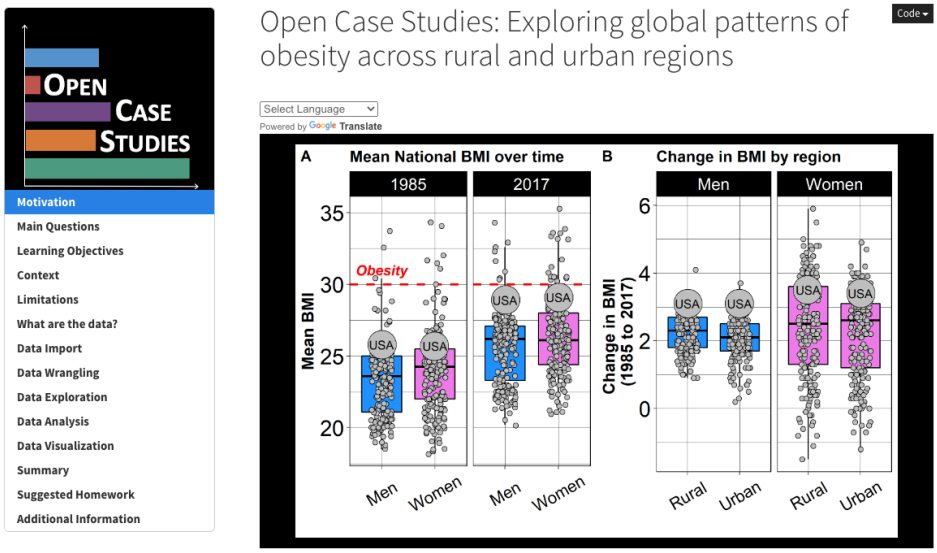
模块化案例研究应用示例		
User	使用示例	数据文件夹
学生	正在寻找的学生 开放数据以用于课程项目的数据科学学生	raw
学生	练习数据整理的公共卫生专业学生和可视化	已导入
教育工作者	课程讲师布置作业 使用相关但全新的数据，这些数据 超出了案例研究的范围	额外
教育工作者	希望从事数据分析教学的教师 学生练习一些简单的 数据导入，但时间有限	更简单的导入
自学者	寻找分析示例的研究人员 示例	整理过的数据

补充表 S2. 模块化案例研究应用示例案例研究的潜在用途示例，不仅限于在课堂上展示完整案例研究，或供学习者跟随学习。

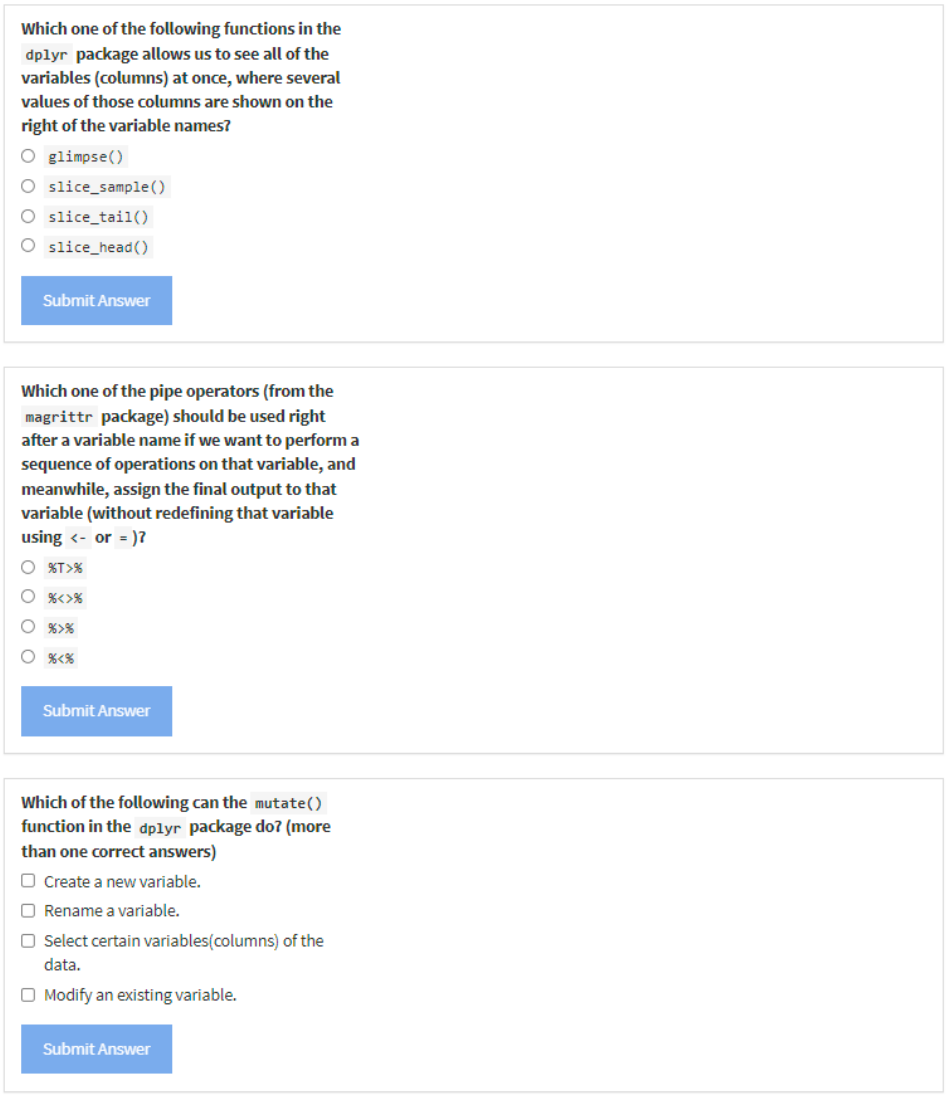
创建 OCS资源中的案例研究所面临的挑战	
挑战	建议的解决方案
为.....找到合适的范围、数据和问题 目标受众	对数据来源保持开放态度，灵活地进行修订或 移除数据
平衡教学目标与教学机会 由数据呈现	保持计划简单，为意料之外的教学留出空间 机会
展示恰到好处的数据科学流程	先进行分析，再筛选要纳入的内容
在假定读者具备一定先验知识与确保 案例研究内容完整自足之间取得平衡	点击展开的部分，提供补充信息以及指向 其他案例研究和资源的链接
满足多种受众类型的需求	将案例研究模块化，使不同用户能够只使用 案例研究中的某些部分，并加入点击展开功能 为需要更多背景信息的学生提供章节
案例研究模块化	在每个章节结束后保存数据，并在每个章节开始时加载 每个章节

补充表 S3。创建 OCS 资源中案例研究的挑战我们遇到的主要挑战及建议的解答列表。

补充图



补充图 S1. 案例研究结构 案例研究顶部的图片，左侧显示交互式目录，另一幅图片总结了该案例研究的主要发现。

```r
```{r dw_quiz, echo = FALSE}
quiz(caption = "",
  question("Which one of the following functions in the `dplyr` package allows us to see all of the variables (columns) at once,
where several values of those columns are shown on the right of the variable names?",
    answer("`slice_head()`", message = "This function allows us to see just the first rows of the data."),
    answer("`slice_sample()`", message = "This function allows us to see a selection of random rows of the data."),
    answer("`slice_tail()`", message = "This function allows us to see the rows at the end of the data."),
    answer("`glimpse()`", correct = TRUE),
    allow_retry = TRUE,
    random_answer_order = TRUE
  ),
  question("Which one of the pipe operators (from the `magrittr` package) should be used right after a variable name if we want to
perform a sequence of operations on that variable, and meanwhile, assign the final output to that variable (without redefining that
variable using `<-` or `=`)?",
    answer("`%>%`", message = "This operator cannot assign the final output to that variable."),
    answer("`%<%`", message = "This is not a valie pipe operator."),
    answer("`%T>%`", message = "This is the side-effects operator."),
    answer("`%<>%`", correct = TRUE),
    allow_retry = TRUE,
    random_answer_order = TRUE
  ),
  question("Which of the following can the `mutate()` function in the `dplyr` package do? (more than one correct answers)",
    answer("Select certain variables(columns) of the data.", message = "Selecting certain variables(columns) of the data is the
function of the `select()` function."),
    answer("Rename a variable.", message = "Renaming a variable is the function of the `rename()` function."),
    answer("Create a new variable.", correct = TRUE),
    answer("Modify an existing variable.", correct = TRUE),
    allow_retry = TRUE,
    random_answer_order = TRUE
  )
)
```
```

补充图 S2. 使用`learnr` 创建选择题测验的代码块示例，摘自二氧化碳排放案例研究。



补充图 S3。 案例研究中呈现的选择题测验 该选择题测验位于 二氧化碳排放案例研究使用图中代码创建 S2。

````
However, we can also do the same thing in a single step, where we replace the `women` dataset
with the code we used to create that dataset.  So first, we select the "Women" and then we split
the data by " Women" to create a new dataset all within the same chunk of code. Write this code
yourself!

```{r, echo=FALSE}
save(rural_urban, file = here::here("www", "exercise", "dw_code2.rda"))
```

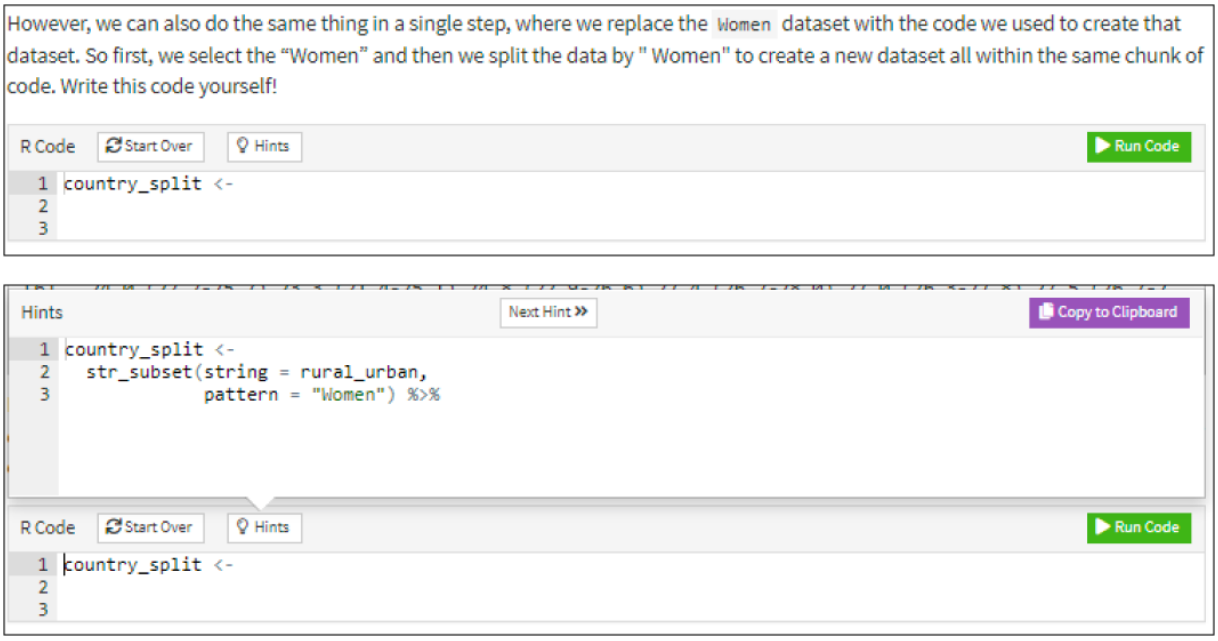
```{r dw_code2-setup}
library(tidyverse)
library(magrittr)

load(here::here("www", "exercise", "dw_code2.rda"))
```

```{r dw_code2, exercise=TRUE}
country_split <-
```

```{r dw_code2-hint-1}
country_split <-
  str_subset(string = rural_urban,
             pattern = "Women") %>%
```

```{r dw_code2-solution}
country_split <-
  str_subset(string = rural_urban,
             pattern = "Women") %>%
  stringr::str_split(pattern= " Women") %>%
  unlist()
# View the first rows of the data
head(country_split)
```
````

补充图 S4. 使用 `learnr` 创建编程练习的代码块示例，来自 肥胖症案例研究。



补充图 S5。 案例研究中的编程练习呈现 在肥胖症案例研究 中呈现的编程练习，S4图中的代码。上：练习题目。下：显示卡住时可以使用的提示函数。这些提示也使用`learnr`制作。

补充说明

补充 附注 S1：在 OCS 案例研究集中创建案例研究的指南 n

为维护本资源核心理念的完整性，我们建议纳入开放案例研究集的案例研究遵循以下指南：

- 案例研究应使用开源编程语言编写。
- 案例研究应使用公开可获取或能够公开的数据。如果数据目前尚未公开，请确保其可以公开。

- 案例研究应包含免责声明和适当的许可协议。
- 应尽可能透明地说明数据的原始来源。
- 所有包含的图像（非案例研究原创）都应注明来源。
- 案例研究必须包含以下核心部分：动机、主要问题、学习目标、局限性（概述数据的局限性）、数据是什么？以及总结（包括对所呈现分析的局限性说明）。
- 案例研究应力求描述执行与数据科学相关任务时所涉及的决策过程。
- 在可能的情况下，应加入指向文献或其他来源的链接，以阐明案例研究的科学主题。
- 尽管案例研究通常由文章激发，但其目的并不是展示论文中的方法，而是作为一种教育资源，引导用户完成数据科学流程。

SupplementalNoteS2 创建交互式案例研究的两种方法的技术方面

在本说明中，我们将介绍如何实现创建交互式案例研究的两种方案：1）将每个练习托管为单独的 Shiny 应用程序，然后将这些应用程序嵌入案例研究中；或 2）通过将案例研究创建为 learnr 教程，构建一个包含案例研究内各项练习的单一应用程序。

方法 1——嵌入交互元素。我们通过调整 De Leon 的方法实现了这一点 [51]。De Leon 基于 iFrame Resizer JavaScript 库，引入了使用超文本标记语言文件实时调整 `learnr`教程窗口大小的方法。该超文本标记语言文件被包含在练习的 R Markdown 文件的 YAML 标头中，同时在最后一行加入 div 标签，以标示 iFrame Resizer 内容的结束位置。我们还通过创建另一个包含 iFrame Resizer JavaScript 的超文本标记语言文件，并将其包含在 YAML 标头中，加入了这一功能。随后，每个练习的 Shiny 应用程序都会作为类名为“interactive”的‘iframe’添加到案例研究中。这是通过将 `learnr` 练习创建在单独的 R Markdown 文件中实现的。然后，每个练习都会作为单独的 Shiny 应用程序在线发布。发布后，每个练习都会拥有一个用于托管它的唯一 URL 地址。接着，可以在案例研究自身的可调整大小的窗口中呈现这些练习：在案例研究的 R Markdown 文件中嵌套 iFrame 超文本标记语言代码块，并将练习 URL 作为输入。每个案例研究的 R Markdown 文件末尾都会添加一个超文本标记语言标签，指定仅对类名为“interactive”的元素应用调整器。R `knitr` 提供了一个 `include app()` 函数来嵌入 Shiny 应用程序，但与之相比，De Leon 的方法更能适应网站的美观设计，以及教程中的弹出窗口（如提示、解答和反馈）。

方法 2 - 单个交互式案例研究应用程序。 第二种选择是将整个案例研究制作成一个`learnr` 教程，在案例研究 R Markdown 文档中直接创建练习。为此，我们将 `runtime: shiny prerendered` 添加到本文档的 YAML 标头中。为了发布单个案例研究应用程序，需要将渲染案例研究所需的所有文件（数据、图像、CSS 样式代码等）放入 `www/` 子目录中。该文件夹将与 `index.Rmd` 文件一起发布到 Shiny。可以使用 `quiz()`、 `question()` 和 `answer()` 函数，在单个 R 代码块中创建测验，这些函数来自 `learnr`（**图 S2-S3**）。必须使用多个 R 代码块创建编程练习。以下每一项分别使用一个代码块：练习设置、提示、解答和练习题目（**图 S4**）。在渲染后的案例研究中，这些独立的代码块会合并为一个编程练习元素（**图 S5**）。最终，该文档会作为单个 Shiny 应用程序发布（opencasestudies.org/ocs-make-interactive-tutorial）。

不同方法之间的差异。 第一种创建交互元素的方法具有优势，因为通常更易于维护。如果某个练习出现问题，我们可以很容易地确定是哪个练习出了问题。但是，这种方法的效率可能较低，因为单个案例研究需要多个 Shiny 应用程序；而且由于每个应用程序都是独立运行的，案例研究中的数据无法直接传递到练习中。因此，当练习需要在分析的特定阶段使用案例研究中的数据时，设置过程可能会变得更加复杂。

第二种方法效率更高，因为整个案例研究都写在同一个文件中。不过，Shiny 应用程序的启动时间必须少于 60 秒。对于渲染时间较长的案例研究，需要将部分数据分析结果单独存储，以缩短启动时间。此外，由于单个文件更难维护，因此务必为案例研究留下详细的维护记录。

我们决定采用第二种方法，因为我们创建了许多练习，若采用第一种方法，就会产生数十个 Shiny 应用程序。第一种方法是将每道练习题制作成一个 learnr 教程，然后将其作为 Shiny 应用程序发布，或通过 Posit Connect 发布。采用这种方法时，交互式案例研究的实际渲染版本仍托管在 GitHub 上，但每道练习都会使用超文本标记语言的 ‘ifra me‘，作为独立窗口嵌入网站中。我们最终采用的第二种方法，是将交互式案例研究制作成单个 learnr 教程。在这种情况下，发布的不是托管在 GitHub 上的渲染后网站版本，而是 Shiny 应用程序。我们选择使用 Posit Connect 来实现这一点，这需要支付一定费用。目前，Shiny 允许个人用户免费发布最多 5 个应用程序。

Supp补充说明 S3 案例研究中的交互式问题示例 es

我们加入了几种交互元素。第一种是在案例研究的分析过程中设置编程练习。这些练习安排在分析以另一个数据集重复已有操作的地方。学生可以运用所学知识，同时始终置身于案例研究的背景中。第二种是在案例研究各部分末尾设置选择题测验和编程练习。

作为第一种类型的示例，如果有两个相似的数据集需要更改变量名，我们可以演示如何更改其中一个数据集的变量名，并要求学生更改另一个数据集的变量名。这类练习也适用于其他情况，例如介绍函数的某个参数时。练习窗口可用于探索该参数的作用：通过编写带有或不带该参数的代码来进行尝试。例如，介绍 set.seed() 函数时，练习窗口允许学生试验使用不同的种子会如何改变输出。

第二类交互元素旨在帮助学习者在案例研究各章节的末尾通过测验题和编程练习评估自己的知识掌握情况。我们将这些练习单独设为一个小节，以便读者轻松找到。与侧重于分析某个具体步骤不同，这些题目综合考查整个章节的内容。选择题用于测试学生对所介绍概念的理解。例如，题目可能会要求学生说出线性回归模型的关键假设、R 函数能够实现什么功能等。编程题则考查学生在数据分析中使用 R 函数的能力。例如，题目可能会要求学生将一个虚构的数据集从宽格式转换为长格式，或使用 R 内置数据集构建机器学习模型等。