DaSEH Instructor Guide

October, 2025

Table of Contents

# About this guide

## 0.1 Formats

You can find an online version of this guide at <https://hutchdatascience.org/daseh_instructor_guide/>.

## 0.2 Summary

The [Data Science for Environmental Health Short Course (DaSEH)](https://daseh.org/), developed at the [Fred Hutch Data Science Lab](https://hutchdatascience.org/), combines online learning and an in-person project-focused intensive. DaSEH is tailored for beginners and novices in R programming, offering instruction on importing, wrangling, visualizing, and analyzing data. It provides hands-on training in using R for statistical computing, a widely-used open-source tool for data analysis and visualization.

This training initiative is funded by National Institute of Environmental Health Sciences 1R25ES035590-01.

DaSEH guides can be used:

1. As a course or to add to a curriculum (either onsite or online) by engaging students to actively participate in data science education for environmental health.
2. Outside of the classroom by providing an archive of examples code for data science best practices for self-learning.

To help guide educators on how to most effectively use the DaSEH resources (either in or outside of the classroom), this guide documents various entry points to using the materials, examples of how to use the materials, how to modify and adapt components of the resources for the classroom, and how to contribute our resources.

# 1 Introduction

## 1.1 Learning Objectives

In this chapter we will:

* Describe the audience for this guide
* Discuss what you can expect in this guide
* Introduce the [Data Science for Environmental Health Short Course (DaSEH)](https://daseh.org/)
* Introduce the resources that make up this project

## 1.2 Motivation

The [Data Science for Environmental Health Short Course (DaSEH)](https://daseh.org/), developed at the [Fred Hutch Data Science Lab](https://hutchdatascience.org/), is an education platform that provides open-source teaching materials using real-world examples for active experiences of introductory concepts in environmental health data analyses. The intention of this guide is to provide instructors more information about how to make the most of our resources.

## 1.3 Target Audience

This guide is intended for educators who are interested in using DaSEH materials for instruction.

Elements of our materials can be helpful for instructors who teach:

* High school students
* Undergraduate students
* Graduate students
* Professionals (researchers and public health practitioners)

Elements of the case studies can assist with teaching courses about the following topics:

* Environmental Health
* Data science
* Statistics
* Public health
* Programming
* Technical writing

… and more

## 1.4 Curriculum

This guide documents:

* The DaSEH philosophy
* The general structure of the short course
* Various entry points to using the short course materials (including our hands-on lab exercises, homework ideas and more)
* Examples of how to use the materials in the classroom
* Instructions on how to modify and adapt components of our materials for the classroom
* Guidelines for contributing to our materials or creating your own short course

## 1.5 DaSEH Philosophy

The DaSEH short course is an educational resource that educators can use in the classroom to teach students how to effectively derive knowledge from data in real-world challenges about environmental health.

We provide independent learners and educators with resources for environmental health data science education. Most of the examples use real data from sources that are used for actual research or public health initiatives in an effort to make the material more engaging and relevant.

Our resources are not intended to show all aspects of the research process, however they are intended to introduce beginners, demonstrate the decision making process of real data analyses, the basic challenges of analyses, to show examples of applications of data methods, and to provide context for when such methods could be applied to glean information from data. We hope to demonstrate best practices for mindful data cleaning, reproducibility, and effective data science communication.

We emphasize learning the concepts of best practices as well as how to troubleshoot and look up documentation to support continued learning, rather than memorization of small coding details. For more information on our teaching philosophy see our [paper about teaching R](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012018).

All DaSEH materials are open source and free and available at our [GitHub Repository](https://github.com/fhdsl/DaSEH). We provide transparency about where the data came from and where possible how it was produced. We also cover data method limitations and ethical considerations.

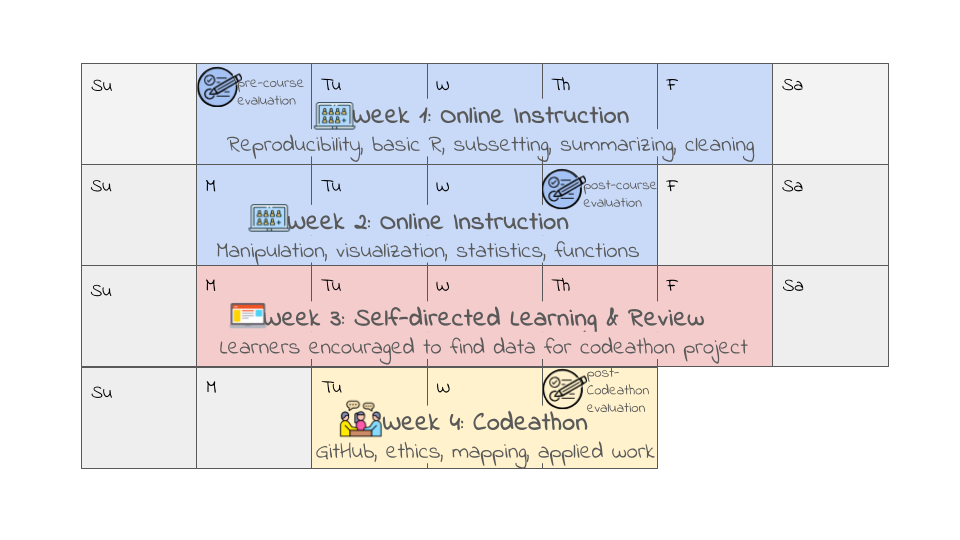
Our materials can be used for small additional assignments in a classroom setting or as the main content that educators can slowly work through with students.

We hope you find our resources useful!

### 1.5.1 What problem are we addressing?

Despite a growing interest in data science and environmental health on campuses, instructors do not always have time to create course content to support new courses. These resources are intended to help others support more environmental health data science education to promote more research and advancement in this important field of public health.

## 1.6 DaSEH Short Course Resources



The DaSEH Short Course has 3 main stages:

### 1.6.1 Stage 1: Online Instructions

In the first stage, learners participate in a synchronous online course where they watch didactic lectures on a particular topic, followed by hands-on lab activities. This last for two weeks. We cover the following topics.

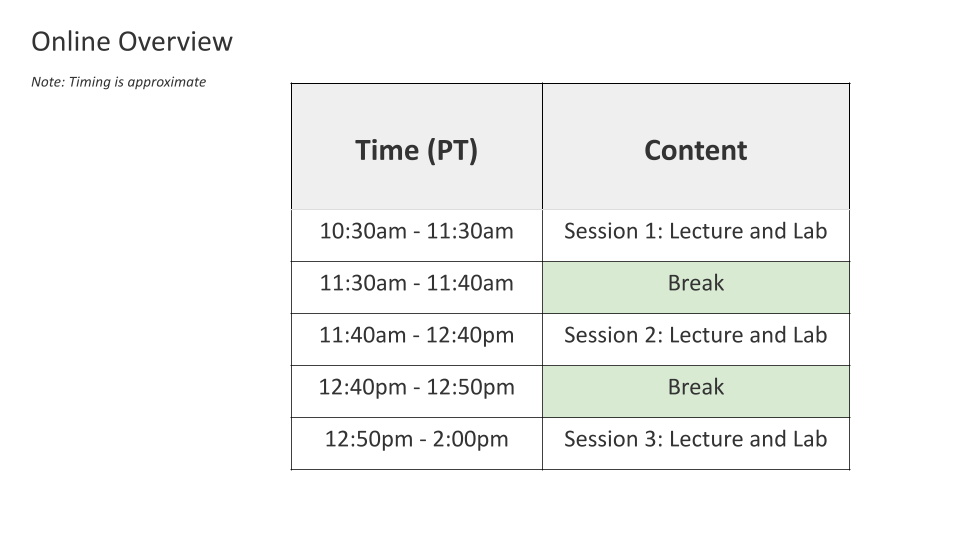
Week one:

* Reproducibility
* Basic R
* Subsetting Data
* Summarizing Data
* Cleaning Data

Week two:

* Manipulation Data
* Visualization of Data
* Statistics
* Functions

Course instruction lasts for 4 hours each day with breaks.



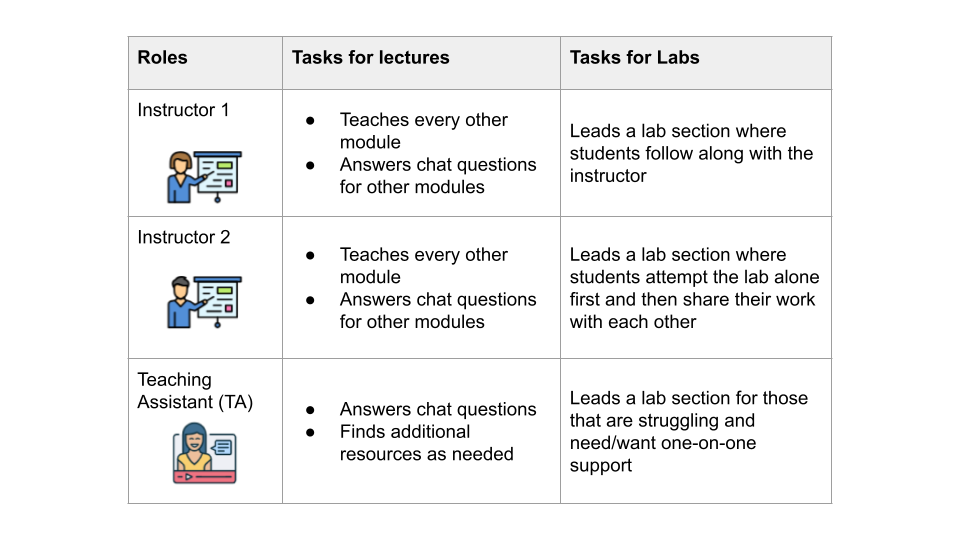
During each session we cover one or two modules, which include didactic lectures led by an instructor and followed by hands-on lab exercises.

We recommend having two instructors and one TA. We have the instructors split up the module lectures for the following reasons:

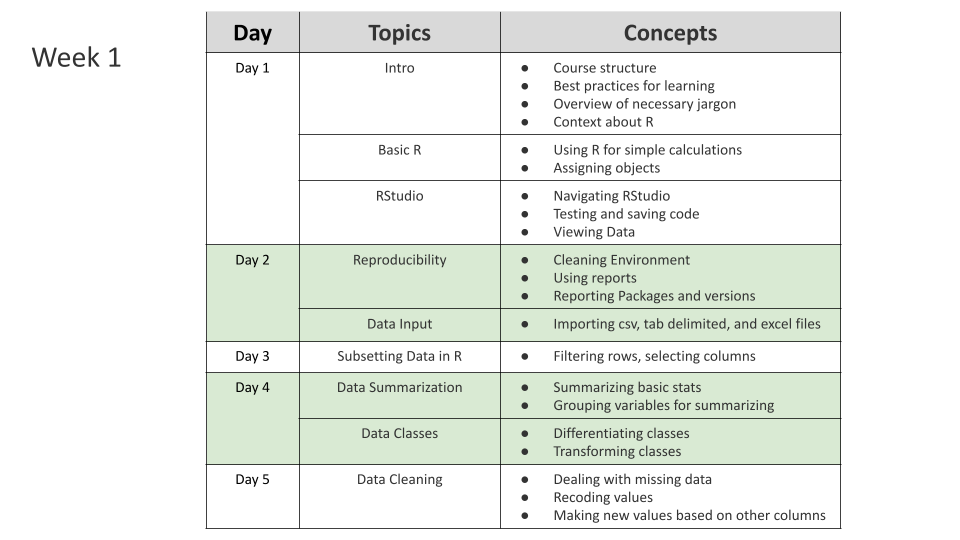
* to reduce instruction burden
* to allow learners the opportunity to ask chat questions during the course without slowing the rest of the course
* to allow the learners to have instructors with slightly different teaching styles/explanations

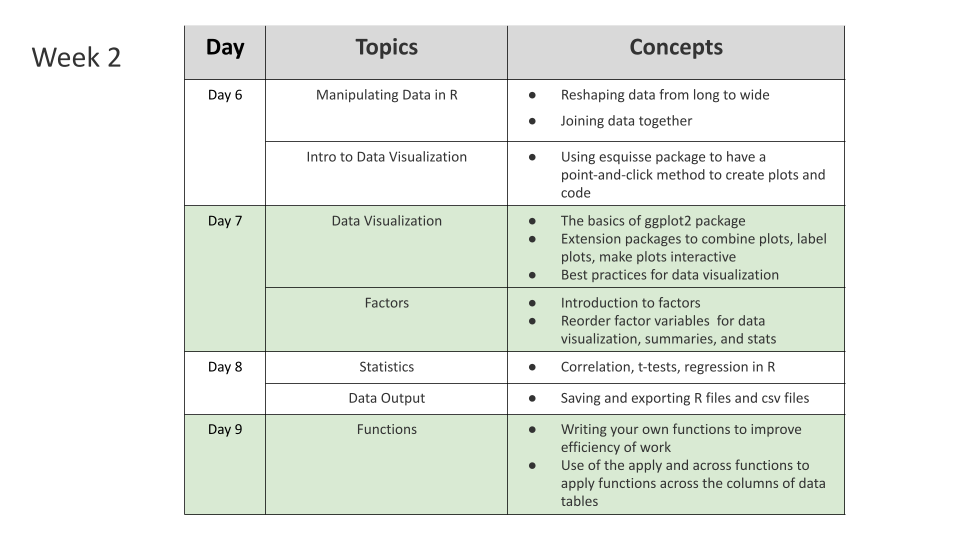
For labs we provide learners with 3 options:

* A room led by one instructor who waits briefly for each question and then walks through the exercises
* A room led by one instructor who allows students time to work through all exercises and then students share their work and challenges with guidance and troubleshooting from the instructor
* A room where the teaching assistant helps those who need one-on-one support



Topics covered in each week of the online course:





Students have access to [our website](https://daseh.org) for course resources which includes lecture slides, lab exercises, cheatsheets to remind them about important functions, homework assignments, and additional resources.

### 1.6.2 Stage 2: Self-Directed Review and Project Planning

In this stage, learners can complete homework problems for extra practice or watch recordings of the lectures and labs from the previous week. This stage also lasts at least a week. It also gives students time to digest what they learned and catch up on other work and personal responsibilities before the code-a-thon.

Learners are also encouraged to find data and are given resources on where one can [find public data](https://daseh.org/data.html), so that they can identify a research question for their project in the code-a-thon stage. Learners are provided with an [example project](https://daseh.org/modules/Project_Example/Project_Example.html) and a [project template R Markdown file](https://daseh.org/modules/Project_Template/Project_Template.Rmd).

### 1.6.3 Stage 3: Code-a-thon

In the final stage of the short course, learners participate in a 3 day code-a-thon, which allows them to apply what they learned to a topic of interest, get peer and instructor in person assistance, ask more extensive questions, and learn about some more advanced topics in reproducibility and data ethics.

See this [code-a-thon schedule](https://docs.google.com/document/d/1ZD-w0vc3Vtv1vf95h6323zaaxORg9vC7Hp4I_IUeW0I/edit?tab=t.0) for more details.

## 1.7 Getting Started

DaSEH materials are designed to be beginner friendly. Users can work through our materials with no prerequisite knowledge in programming, environmental health, public health, or statistics. However, the appropriate technology, software, and a basic familiarity with R Studio is helpful. This section will detail what learners and instructors will need to be able to jump into our resources.

### 1.7.1 Technical Requirements

Our materials the R statistical programming language for data analysis. R is available for Windows, Mac, and Linux and can be downloaded from the [The Comprehensive R Archive Network (CRAN)](https://cran.r-project.org/). While there is no specific R version requirement, we recommend that instructors have all students use the same version (typically the most recent version of R).

Additional R packages are also required including:

* tidyverse
* naniar
* janitor
* esquisse

Cheat sheets detailing how to use common packages and functions are available at the [RStudio website](https://www.rstudio.com/resources/cheatsheets/).

Please note that to install R version 3.5 and higher you will need to have a [web browser](https://en.wikipedia.org/wiki/Web_browser) and a compatible [operating system](https://en.wikipedia.org/wiki/Operating_system). See this [link](https://cran.r-project.org/) for more information.

### 1.7.2 RStudio

To work with R, we recommend the RStudio Integrated Development Environment. RStudio includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server. More information and program installation instruction are available at the [RStudio website](https://www.rstudio.com/products/rstudio/).

For reproducible data analyses and easy publishing of reports and presentations, we recommend using RMarkdown. More information about RMarkdown is also available at the [RStudio website](https://rmarkdown.rstudio.com/index.html).

### 1.7.3 GitHub

DaSEH is hosted on GitHub. GitHub is a website and cloud service that enables developers to store, manage, and track changes to their code. DaSEH uses GitHub for both development and distribution purposes. Users have complete access to all materials through our [DaSEH GitHub repository](https://github.com/fhdsl/daseh). The repository contains all the materials needed for the case study.

This includes the slide files to be distributed to learners, data used in the examples (that aren’t installed in R otherwise), cheatsheets to remind learners of packages, homework and lab files, including both keys and raw files for learners. While experience with GitHub is not needed for users to access the materials, users can use GitHub to streamline download and modification of our resources. An introduction on how to get started with GitHub is available at [Happy Git](https://happygitwithr.com/index.html).

## 1.8 Website

Our website has the following pages:

1. An about page explaining what the course is for: <https://daseh.org/index.html>
2. A page that links to a form for people to apply to the course: <https://docs.google.com/forms/d/e/1FAIpQLScpzjYQBuM2rpo4SMZL_H65PI6U0sR1lPDjMI-DG2MtSld-1A/viewform>
3. A page with information about how people can participate and what is required: <https://daseh.org/logistics.html>
4. A schedule with all associate materials: <https://daseh.org/materials_schedule.html>
5. Resources about where to find environmental health data: <https://daseh.org/data.html>
6. Resources for extra learning and help: <https://daseh.org/resources.html>
7. A help page for common issues, common errors, and how to change preferences in RStudio: <https://daseh.org/help.html>

## 1.9 Feedback

We are continually working to improve the DaSEH resources to address learner and educator needs. Feedback is essential for this goal. If you use DaSEH materials, we would love to hear from you!

### 1.9.1 Survey

Please consider participating in our user experience survey.

avocado need to update for our own survey!

We appreciate feedback about all aspects of the DaSEH user experience including but not limited to technical challenges, potential improvements, or new ideas on how to use the our materials in the classroom.

## `google-chrome`, `chromium-browser` and `chrome` were not found. Try setting the `CHROMOTE\_CHROME` environment variable to the executable of a Chromium-based browser, such as Google Chrome, Chromium or Brave or adding one of these executables to your PATH.

### 1.9.2 Feedback for this Guide

If you have feedback for this guide about how we can improve the content, please see [here](https://github.com/fhdsl/daseh_instructor_guide/issues/new/choose).

## 1.10 Contact Us

If you have any questions, suggestions for improvement for individual resources, ideas, or thoughts, you can contact us through email at. avocado add info

You can also find us XYZ avocado add

# 2 Open Case Study Infrastructure

## 2.1 Learning Objectives

In this chapter we will discuss the overall infrastructure of the Open Case Studies platform, which includes:

* The Open Case Studies (OCS) website
* Methods to provide feedback
* A search tool to find case studies
* The OCS GitHub organization
* Our R package called OCSdata

## 2.2 OCS Website

The [OpenCaseStudies website](https://www.opencasestudies.org/) describes the mission of the Open Case Studies project, the history of its inception, current and previous members of the OCS team, an archive of talks and blog posts and other information.

Links to all of our case studies can be found on the Open Case Studies website. The case studies are listed in a searchable table that will be detailed further in the following section.

## 2.3 Feedback

We are continually striving to make our case studies better. Please contact us if you have ideas for suggestions for the project or new ideas for how case studies can be used in the classroom.

Also please let us know if you notice typos or errors, or if you are interested in getting involved.

### 2.3.1 Email Form

The website contains a contact [email form](https://www.opencasestudies.org/OCS_Guide/introduction.html#contact-us) that may be used to send a message to Open Case Studies to ask a question or provide suggestions.

### 2.3.2 Survey

There is also a [survey](https://docs.google.com/forms/d/e/1FAIpQLSfpN4FN3KELqBNEgf2Atpi7Wy7Nqy2beSkFQINL7Y5sAMV5_w/viewform?usp=sf_link) available on our website and also within the case studies themselves that allows us to do research on case study use.

The survey should take no more than 10 minutes to complete. Your feedback helps us learn more about how to improve the data science education experience. Part of this includes getting a better understanding of who is using our case studies and how so that we can better design our case studies. We would greatly appreciate you filling it out if you have the time!

## 2.4 OCS Case Study Search Tool

The website also includes the case study search tool to aid instructors in finding appropriate case studies for their learning objectives. Accessing the search tool and how to use it is described in more detail below.

This diagram illustrates the workflow of accessing a case study from the OCS website through the case study search table. From the table, users can use the provided links to view the original static case studies, interactive case studies, and the GitHub repositories for each. Users may find all case study source files in the case study repository, as well as instructions on how to use the case study.

### 2.4.1 Interactive Case Studies

The interactive versions of the case studies are a recent development. These versions include live tutorials through quizzes and interactive coding exercises with real-time feedback. The interactive case studies were made using the [learnr](https://rstudio.github.io/learnr/) and [gradethis](https://pkgs.rstudio.com/gradethis/) packages.

If you’d like to learn more about the interactive case studies, graduate student [Qier Meng](https://www.opencasestudies.org/authors/qmeng/) discusses the interactive versions in further detail in the video below:

If you’d like to learn more about the interactive case studies you can read this [thesis](https://jscholarship.library.jhu.edu/handle/1774.2/66820) by former graduate student [Michael Breshock](https://mbreshock.github.io/) ([**breshock\_expanding\_2021?**](#ref-breshock_expanding_2021)).

The [Open Case Study Search](https://www.opencasestudies.org/#searchtab) tool can be found at the bottom of the [OCS Website](https://www.opencasestudies.org/). The tool consists of a table with searchable columns and each row describing an individual case study. This searchable table is designed to aid instructors in identifying appropriate case studies for their learning objectives. The columns are organized as such:

* The “Case Study” column contains the case study name and a link to the static and interactive versions of the case study (if available)
* The “GitHub Repository” column provides links to the case study’s associated GitHub repository that contains all case study source files, data, code, and more
* The “Packages” column details all the R packages used in the case study, and can help identify if a case study teaches a specific data import, wrangling, analysis, or visualization skill
* The “Objectives” column details the learning objectives of each case study (e.g. importing data from PDF files, reshaping data, specific statistical analysis, etc.)
* The “Category” column lists the source of funding or project that the case study is associated with

The main two columns likely to be helpful in identifying appropriate case studies are the “Packages” and “Objectives” columns. Users may search for keywords across all columns using the overall search bar, otherwise users can search individual columns of interest.

This table can be used to access all case study resources:

This video provides a live demonstration on how to use the search tool:

## 2.5 Open Case Studies GitHub Organization

GitHub is a website and cloud service that enables developers to store, manage, and track changes to their code. OCS uses GitHub for both development and distribution purposes. Users have complete access to all case study material through our [OCS GitHub page](https://github.com/opencasestudies) where each case study is hosted in an individual repository. The repository contains all the materials needed for the case study. This includes the case study text to be distributed to students, the data used in the case study (discussed below), additional documents and references, and brief guidelines on case study use.

Data included in the GitHub repository is available in multiple formats to enable modular use of the case studies. This diagram explains the case study data folder structure and how data is categorized into different sub-folders:

Data included in the GitHub repository is available in multiple formats to facilitate modularization of the case studies as described below. To use the case study data, you can download the GitHub repository directly or use the OCSdata R package described below.

## 2.6 OCSdata

To simplify the process of accessing the data required for each case study, we have created the OCSdata R package. Briefly, the OCSdata package creates a new folder called “OCSdata” where it downloads the data needed for a specific case study. Users can download the data in its original raw format or in various processed formats that correspond to different stages of data wrangling and cleaning. This allows users to perform the data exploration and wrangling or the data visualization and analysis sections of the case study without having to process the data from the raw files. For some of the case studies, the OCSdata package also downloads extra source data that is not used in the case study.

The following are the main functions to import data in various formats using the OCSdata package. Each function is described in more detail in the OCSdata [package documentation](https://cran.r-project.org/web/packages/OCSdata/index.html).

| Data Folder | Case Study Section | OCSdata Function |
| --- | --- | --- |
| raw | Data Import | raw\_data |
| imported | Data Exploration, Data Wrangling | imported\_data |
| wrangled | Data Visualization, Data Analysis | wrangled\_csv, wrangled\_rda |
| simpler\_import | Data Import | simpler\_import\_data |
| extra | Suggested Homework (?) | extra\_data |

The package source files and documentation are also available on [GitHub](https://github.com/opencasestudies/OCSdata).

### 2.6.1 Getting Started with OCSdata

The OCSdata package is available on the package repository [CRAN](https://cran.r-project.org/web/packages/OCSdata/index.html). It requires R 3.5 or higher and can be installed in R as follows:

### 2.6.2 Downloading raw data

The raw\_data function will download the raw data files that can be imported into R.

The first argument is the name of the case study. A list of case study names can be found in the package documentation [online](https://cran.r-project.org/web/packages/OCSdata/vignettes/instructions.html#casestudy) or by typing ?raw\_data in R.

The outpath argument is a string specifying the folder where the data should be downloaded. To download the data to a folder named “OCS\_data” in the current working directory, you can supply getwd() to the output argument. If nothing is provided for the argument, you will be prompted to enter 1, 2, or 3 to download the data into the current director, to specify the download path, or to cancel, respectively.

In the following example, we download the raw data for the “Opioids in the United States” case study to the current directory.

### 2.6.3 Downloading data in other formats

The OCSdata package can be used to download the data in various processed formats that may be helpful in skipping certain case study sections and focusing on data wrangling and/or analysis and visualization. All of the functions take the same arguments described above.

#### 2.6.3.1 Simpler import

The simpler\_import\_data() function will download raw data files that have been converted to file formats that are easier to import into R, typically .csv. Some case studies offer this option when the original raw files require a more complicated import step.

#### 2.6.3.2 Importing data as R objects

The imported\_data() function will download raw data files in .rda format. This means the data have already been imported into R objects. This can be used to skip the data import section and start directly with data wrangling. The R objects files can be imported into R by either double clicking on the files in RStudio or using the load() function as follows.

#### 2.6.3.3 Importing wrangled data

The following functions will download the data files that have already been wrangled and are ready to be analyzed. These come in both .csv and .rda formats.

Download as csv files:

Downloading as R objects:

### 2.6.4 Downloading extra data

Some case studies have extra data are not used in the case study but can be used to explore the case study subject from different perspectives. These data but can This data can be downloaded using the extra\_data() function.

### 2.6.5 Downloading all case study data

The zip\_ocs() function will download the all of the repository files in a .zip folder and unzip them into a specified directory. This includes the case study data in all the formats detailed above (raw, simpler\_import, imported, wrangled, and extra). It also includes the case study .Rmd file, which can be modified by instructors as needed. We recommend using this method over cloning or forking (terms that you may be familiar with if you are familiar with Git and GitHub), as this will not result in the user getting all of our git history.

If you choose to fork the repository you will automatically generate a repository on GitHub and your repository will have connections to the original case study. This can be helpful for pulling any changes to the original case study. It can also be helpful if you want to send edits to the original case study in the form of what is called a pull request.

If you clone the case study repository, you can set it up on GitHub as well with a few more steps and you will not preserve any connection to the original case study repository.

Again, don’t worry if all these terms are new to you. You can just use the zip\_ocs() function instead. Otherwise take a look at ([**happygitwithr?**](#ref-happygitwithr)) to learn more.

### 2.6.6 Fork or clone the case study repository

If instead users are familiar with Git and GitHub and want to fork or clone the case study repository, this can also easily be done using the OCSdata package. The clone\_ocs() function of the OCSdata package can be used to do either. If the fork\_repo function is set to TRUE it will fork the repo, otherwise, by default, it will clone the repository. These functions will result in the same outcome as using GitHub to clone or fork the repo.

Again you can also specify the outpath location as in the previous description about the zip\_ocs() function.

However, using these functions will involve the users getting all of our git history so we suggest that users use the zip\_ocs() function (described in the above section) of OCSdata instead.

If you’d like to learn more about the OCSdata package or the OCS GitHub organization page, you can read this [thesis](https://jscholarship.library.jhu.edu/handle/1774.2/66820) by former graduate student [Michael Breshock](https://mbreshock.github.io/) ([**breshock\_expanding\_2021?**](#ref-breshock_expanding_2021)).

## 2.7 Session info

## R version 4.3.2 (2023-10-31)  
## Platform: x86\_64-pc-linux-gnu (64-bit)  
## Running under: Ubuntu 22.04.4 LTS  
##   
## Matrix products: default  
## BLAS: /usr/lib/x86\_64-linux-gnu/openblas-pthread/libblas.so.3   
## LAPACK: /usr/lib/x86\_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so; LAPACK version 3.10.0  
##   
## locale:  
## [1] LC\_CTYPE=en\_US.UTF-8 LC\_NUMERIC=C   
## [3] LC\_TIME=en\_US.UTF-8 LC\_COLLATE=en\_US.UTF-8   
## [5] LC\_MONETARY=en\_US.UTF-8 LC\_MESSAGES=en\_US.UTF-8   
## [7] LC\_PAPER=en\_US.UTF-8 LC\_NAME=C   
## [9] LC\_ADDRESS=C LC\_TELEPHONE=C   
## [11] LC\_MEASUREMENT=en\_US.UTF-8 LC\_IDENTIFICATION=C   
##   
## time zone: Etc/UTC  
## tzcode source: system (glibc)  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## loaded via a namespace (and not attached):  
## [1] compiler\_4.3.2 fastmap\_1.1.1 bookdown\_0.43 cli\_3.6.2   
## [5] htmltools\_0.5.7 tools\_4.3.2 yaml\_2.3.10 rmarkdown\_2.25   
## [9] knitr\_1.50 digest\_0.6.34 xfun\_0.52 rlang\_1.1.6   
## [13] evaluate\_1.0.4

# 3 Use of Open Case Studies

## 3.1 Learning Objectives

This chapter will provide guidance on how to use Open Case Studies for instruction.

* We will give a coarse guide on which case studies include material appropriate for beginner, intermediate, or advanced learners.
* We will describe ways that instructors can use full case studies, part of a case study, or just the data.
* We will present ways in which others have used the case studies.
* We will present some examples of extensions that can accompany the case studies and can serve as a template for homework assignments or independent student exploration.

The examples presented in this chapter are merely suggestions - modifications to the case studies to fit student needs are expected and encouraged! If you come up with a different way to use the case studies, please [let us know](https://www.opencasestudies.org/#contact) what you come up with so that other educators may be inspired by your creativity.

### 3.1.1 Prerequisites

#### 3.1.1.1 Public Health Subject Matter

The case studies in Open Case Studies generally deal with topics in public health. We do not require any prior knowledge on the public health subjects examined in the case studies. The Getting Started section in each case study (specifically, the case study context) will present the subject material relevant to understanding the data and the case study implications.

#### 3.1.1.2 Statistics

The case studies cover a range of statistical approaches and skillsets. Some case studies will build on more foundational statistical concepts. The Experience Level categorization for each case study (see the [case study recommendation section](https://www.opencasestudies.org/OCS_Guide/use-of-open-case-studies.html#case-study-recommendations)) will indicate the expected prior skills that the case study will expect. Furthermore, the exact skills that the case study will use are listed in the Case Study Search Tool under the Objectives column.

#### 3.1.1.3 Coding/Data Science

All case studies use the R statistical programming language for data analysis. Some familiarity with R basics is expected for effective use of the case studies. However, depending on the data used and the extent to which it needs to be cleaned and processed before analysis, each case study may require experience with additional programming and data wrangling skills. This will be indicated by the Experience Level designation for the case study ([case study recommendation section](https://www.opencasestudies.org/OCS_Guide/use-of-open-case-studies.html#case-study-recommendations)).

#### 3.1.1.4 Software

All case studies use the R statistical programming language for data analysis. While there is no specific R version requirement for the case studies, the OCSdata package, which can be used to get and load the data, does require R 3.5. Furthermore, R packages used to run specific analyses in each case study may have their own R version requirements. R version requirements may be checked in the sessionInfo() section in each case study.

### 3.1.2 Experience Level Descriptions

The table below explains how we define the beginner, intermediate, and advanced experience levels.

| Experience Level | Description |
| --- | --- |
| Beginner | Little to no previous experience with coding and/or statistical analysis. |
| Intermediate | Familiar with at least one programming language and has experience working with data and statistics. |
| Advanced | Fluent in at least one programming language and likely familiar with more. Has a depth of experience working with data science projects. |

Typically, most middle/high school and first year undergraduate students will fit in the beginner category. Upperclassmen undergraduates and some graduate students are often at the intermediate level. Most advanced level students will be at the graduate level. However, this is a generalization, and a student may be considered beginner, intermediate, or advanced at any academic level depending on their independent studies and experiences.

## 3.2 Open Case Studies in the Classroom

The case studies are structured to support both partial and full use of a case study. Educators are also free to use case study data by itself.

### 3.2.1 Teaching a Full Case Study

The case studies are written to provide a cohesive story that simulates data science in the real world. Reading through an entire case study is an excellent exercise for students to experience a standard data science workflow and learn the best practices of today from start to finish. The following list provides a few examples of how educators have used a full case study in the past:

* Converted the case study into a [slide deck](https://cogs137.github.io/website/lecslides/16-cs2-data.html#1) ([Practical Data Science in R](https://cogs137.github.io/website/) taught by [Dr. Shannon Ellis](https://www.shanellis.com/) at University of California Santa Barbara)
* Assigned students to read the case study and write a report as homework ([Advanced Data Science course at Johns Hopkins](https://github.com/advdatasci/homework9))
* Assigned students to extend analysis beyond case study ([Advanced Data Science course at Johns Hopkins](https://github.com/advdatasci/homework11))

**Case Study Reading Time and Readability Index**

The reading time and readability index were calculated for each case study with [koRpus](https://github.com/unDocUMeantIt/koRpus). A readability index estimates the reading difficulty level of a particular text. The following table lists the reading time and the Flesch-Kincaid readability index for each case study. This information may be useful for deciding which case study to use in your curriculum. Based on the course taught by Dr. Ellis, in which the case studies were used for lecture material, it appears that for intermediate level data science students, a 10 week course could cover 2 longer case studies or 3 shorter case studies. How long a case study will take however, will depend on the experience level of the students.

| **Case Study** | **Reading Time** (minutes) | Readability Index (Flesch-Kincaid) |
| --- | --- | --- |
| [School Shootings in the United States](https://www.opencasestudies.org/ocs-bp-school-shootings-dashboard) | 110 | Grade 9, Age 14 |
| [Disparities in Youth Disconnection](https://www.opencasestudies.org/ocs-bp-youth-disconnection) | 85 | Grade 8, Age 13 |
| [Opioids in United States](https://www.opencasestudies.org/ocs-bp-opioid-rural-urban) | 90 | Grade 9, Age 14 |
| [Vaping Behaviors in American Youth](https://www.opencasestudies.org/ocs-bp-vaping-case-study) | 75 | Grade 10, Age 15 |
| [Mental Health of American Youth](https://www.opencasestudies.org/ocs-bp-youth-mental-health) | 90 | Grade 8, Age 13 |
| [Exploring global patterns of obesity across rural and urban regions](https://www.opencasestudies.org/ocs-bp-rural-and-urban-obesity) | 70 | Grade 9, Age 14 |
| [Influence of Multicollinearity on Measured Impact of Right-to-Carry Gun Laws Part 1](https://www.opencasestudies.org/ocs-bp-RTC-wrangling) | 55 | Grade 9, Age 14 |
| [Influence of Multicollinearity on Measured Impact of Right-to-Carry Gun Laws Part 2](https://www.opencasestudies.org/ocs-bp-RTC-analysis) | 60 | Grade 11, Age 16 |
| [Exploring CO2 emissions across time](https://www.opencasestudies.org/ocs-bp-co2-emissions) | 70 | Grade 9, Age 14 |
| [Exploring global patterns of dietary behaviors associated with health risk](https://www.opencasestudies.org/ocs-bp-diet) | 100 | Grade 10, Age 15 |
| [Predicting Annual Air Pollution](https://www.opencasestudies.org/ocs-bp-air-pollution) | 100 | Grade 10, Age 15 |

### 3.2.2 Teaching Part of a Case Study

Some educators may find that only certain sections of a case study are relevant to their course learning objectives. For these educators the case studies are structured to allow for modular use. The case study [GitHub repositories](https://github.com/opencasestudies) provide the appropriate data files to be used at the start of each section. These data files can be downloaded directly from GitHub or with the [OCSdata](https://github.com/opencasestudies/OCSdata) package (see the [OCSdata section](https://www.opencasestudies.org/OCS_Guide/open-case-study-infrastructure.html#ocsdata) in chapter 2). The following table describes which data folder contains the corresponding data for each case study section. An example use for each data type is provided as well.

| Data Folder | Case Study Section | Example Use |
| --- | --- | --- |
| raw | Data Import | Assigning public health students to read through a case study starting from the beginning |
| imported | Data Exploration, Data Wrangling | Data science students practicing wrangling methods with in class exercises |
| wrangled | Data Visualization, Data Analysis | Statistics course practicing analysis methods with real data as a live lesson |
| simpler\_import | Data Import | Introductory data science instructor wants students to practice data import without over-complication |
| extra | Not Used in Case Study | Public health course assigns homework to practice case study methods with similar but new data |

### 3.2.3 Teaching With Case Study Data Only

Educators can use the data available with the case studies without using the case study as a whole. The data is available on GitHub and can be accessed using the OCSdata R package.

## 3.3 Case Study Recommendations

Each case study demonstrates an entire data analysis starting from data import and wrangling and continuing to statistical analysis and data visualization. This means that individual case studies can often incorporate different skill levels for different stages of the analyses. For example, a case study can use advanced data wrangling approaches but only need beginner level statistical analysis methods. Below we provide broad categorizations of each of the case studies in terms of the skill levels required (beginner, intermediate, or advanced) for each of the general stages of the case studies (data wrangling, data visualization, and statistics). More details on the specific skills taught in each case study can be found in the [Case Study Search Tool](https://www.opencasestudies.org/#searchtab) or in the Learning Objectives section in each case study.

Here, we are using the following interpretations of “beginner”, “intermediate”, and “advanced”:

| Skill Level | Data Import | Data Wrangling | Data Visualization | Statistics |
| --- | --- | --- | --- | --- |
| Beginner | No experience with importing data into any programming language | No experience wrangling and cleaning raw data in any programming language | No experience visualizing data in any programming language | No experience with statistical concepts |
| Intermediate | Some experience with importing common data formats (e.g. CSVs) into R or significant experience in another programming language | Some experience wrangling or cleaning raw data in common formats (e.g. numerical data) in R or significant experience in another programming language | Some experience with common visualization packages in R (e.g. ggplot) or significant experience in another programming language | Some familiarity with common statistical concepts (e.g. summary statistics, hypothesis testing) and techniques (e.g. t-test) |
| Advanced | Experience with importing uncommon data types (e.g. PDFs or web-scraping) and comfort with troubleshooting import challenges | Experience cleaning and wrangling raw data in uncommon formats (e.g. regular expressions) in R and comfort with troubleshooting wrangling challenges | Experience with creating complex data visualizations in R and comfort with visualization challenges | Good understanding of foundational statistical concepts and comfort with applying foundational statistical techniques |

The following table lists a few example case studies that would be suitable for each experience level.

| Case Study | Data Import | Data Wrangling | Data Visualization | Statistics |
| --- | --- | --- | --- | --- |
| School Shootings in the United States | Intermediate | Beginner | Advanced | Beginner |
| Disparities in Youth Disconnection | Intermediate | Beginner | Beginner | Beginner |
| Opioids in United States | Intermediate | Beginner | Beginner | Intermediate |
| Vaping Behaviors in American Youth | Beginner | Beginner | Beginner | Intermediate |
| Mental Health of American Youth | Advanced | Beginner | Beginner | Beginner |
| Exploring global patterns of obesity across rural and urban regions | Intermediate | Beginner | Beginner | Beginner |
| Influence of Multicollinearity on Measured Impact of Right-to-Carry Gun Laws - Part 1 | Beginner | Advanced | NA | NA |
| Influence of Multicollinearity on Measured Impact of Right-to-Carry Gun Laws - Part 2 | NA | NA | Intermediate | Advanced |
| Exploring CO2 emissions across time | Beginner | Beginner | Beginner | Beginner |
| Exploring global patterns of dietary behaviors associated with health risk | Intermediate | Intermediate | Intermediate | Intermediate |
| Predicting Annual Air Pollution | Beginner | Beginner | Intermediate | Advanced |
| Exploring health expenditure using state-level data in the United States | Beginner | Beginner | Beginner | Beginner |

## 3.4 Troubleshooting

You may encounter errors trying to render our case studies.

In which case, we suggest that you check the “Session Info” section under the “Additional Information”section of the case study to see what versions of packages we used.

R packages versions can have updates to arguments and function names that can cause code to work differently or can break the code.

If you encounter an error, this is likely the reason. We try to update our case studies when we can, but we can’t always update the information in a timely manner as this is currently a passion project. You can either use the error message from trying to knit the case study to determine what function may have been updated or deprecated (we recommend this option to help you or your students learn the most up-to-date information), or you can use the versions of the packages that are shown in our Session Info section and load the versions that we used, following the directions [here](https://search.r-project.org/CRAN/refmans/remotes/html/install_version.html).

## 3.5 Example Use Cases

Because the case studies were developed to be modular and stand-alone, they can be used in a variety of ways that cater to the learner’s goals, experience, and interests. Below, we provide a few examples of how case studies have been used previously. If you use Open Case Studies in a new way, we would love to [hear](https://www.opencasestudies.org/#contact) about it!

### 3.5.1 Using Case Studies as Lecture Content

[Practical Data Science in R](https://cogs137.github.io/website/) is a 10-week intermediate undergraduate course taught by [Dr. Shannon Ellis](https://www.shanellis.com/) at University of California Santa Barbara. In 2021, Dr. Ellis taught the course using three Open Case Studies and used them to illustrate how foundational data science skills and statistical concepts taught throughout the course can be applied to real data.

Here you can see how the course used the OCSdata package:

[Dr. Ellis](https://www.shanellis.com/) incorporated labs and homework assignments into the course, which had guided coding and analysis exercises related to the concepts discussed in lecture which used content from the case studies in a slide format. She also assigned written reports where students presented the analysis they conducted related to the case study in the format of a scientific article (see example assignment below).

### 3.5.2 Using Case Studies for Assignments

[Advanced Data Science](http://jtleek.com/ads2020/) was a semester-long graduate data science course taught by [Dr. Jeff Leek](https://jtleek.com/) and [Dr. Roger Peng](https://rdpeng.org/) in 2020 at Johns Hopkins Bloomberg School of Public Health, primarily for PhD students. This course is designed for students to gain experience in designing and communicating data analyses effectively and critically analyzing analyses. Assignments included [writing scientific journal sections](https://github.com/advdatasci/homework9) (e.g. Introduction, Methods, Results, Discussion) based on the case studies and [extending analyses](https://github.com/advdatasci/homework11) based on results presented in the case study.

### 3.5.3 Independent Study

Case studies can be used for learners to gain experience in statistics and data science independently. We strongly recommend that independent learners aim to actively engage with the case study by running the analyses independently, exploring the data beyond what is presented in the case study, and extending the analyses by to investigate their own hypotheses. Furthermore, creating a finished product, such as a blog post or a presentation, can be an excellent demonstration of the skills learned.

### 3.5.4 Interactive Case Studies

Some of the case studies also have interactive versions. These versions allow students to write and run code in the browser interactively, with hints and answers available for students to check their progress as they go through the case study. Interactive case studies could be appropriate for independent learning or for in class labs, as they provide real time feedback and can reduce demands on the educator to provide intensive personalized feedback. Please see the following video for a demonstration on how to use the interactive case studies:

## 3.6 Examples of assignments

Educators are not limited to having the students go through the case study in their current format. Case studies can be a spring board for further exploration and additional assignments. For example, in addition to helping students develop data analysis skills, we hope that the case studies can also help students develop their curiosity, technical writing and communication skills. Additional assignments can include but is not limited to data visualization and presentation, written reports, and oral presentations. Below we provide a few examples of potential assignments that educators can use to tailor instruction to the desired learning objectives. See [here](http://jtleek.com/ads2020/week-5.html) guidelines about considerations for effective and ethical data visualizations from the [Advanced Data Science Course](http://jtleek.com/ads2020/) taught by [Jeff leek](https://jtleek.com/) and [Roger Peng](https://rdpeng.org/) at [Johns Hopkins Bloomberg School of Public Health](https://publichealth.jhu.edu/).

### 3.6.1 Written Report

Below is an example of a scientific-style paper written based on the [Opioid Use Case Study](https://www.opencasestudies.org/ocs-bp-opioid-rural-urban/#Main_Question). We also include an example rubric by which this paper can be evaluated adapted from [here](https://ocw.mit.edu/courses/biological-engineering/20-109-laboratory-fundamentals-in-biological-engineering-spring-2010/assignments/guidelines-for-writing-up-your-research/#Evaluation).

* [OCS Opioid Use in the US Example Report](https://raw.githubusercontent.com/opencasestudies/OCS_Guide/main/assets/OCS_Opioids_Example_Report.pdf)
* [OCS Opioid Use in the US Example Report Rubric](https://raw.githubusercontent.com/opencasestudies/OCS_Guide/main/assets/OCS_Opioids_Example_Report_Rubric.pdf)

### 3.6.2 Oral Presentation

Below is an example of a presentation based on the [Vaping Behaviors in American Youth Case Study](https://www.opencasestudies.org/ocs-bp-vaping-case-study/). This presentation focuses on the context of the study and the methods used in the analysis. Assignments for presentations can be modified to emphasize presenting results and conclusions or to emphasize communication to different audiences (e.g. policy makers, other researchers, the public, etc. ).

* [OCS Vaping Case Study Oral Presentation Assignment Example](https://github.com/advdatasci/homework12)
* [OCS Vaping Case Study Oral Presentation Example](https://www.youtube.com/watch?v=noWLCSipKEU)

### 3.6.3 Data Visualization

The data visualizations included in the case studies are not the only way to present the data used within the analyses. While the principles of effective data visualization are a focus of the case studies, the data included as well as the study questions can be used to guide students through the design choices that are commonly considered when determining how to best present data. Students can be assigned to create a new visualization beyond what is included in the case studies that emphasizes different aspects in the data.

### 3.6.4 Further Exploration

Several case studies have additional data that is not discussed. This can be used for further exploration of the subject area that was discussed in the case study. This can be guided by the questions included in the Homework section of the case studies. Case studies that have additional data include the [Opioid Use Case Study](https://www.opencasestudies.org/ocs-bp-opioid-rural-urban/#Main_Question), the [Right to Carry Case Study](https://www.opencasestudies.org/ocs-bp-RTC-analysis/), and the [CO2 Emissions Case Study](https://www.opencasestudies.org/ocs-bp-co2-emissions/).

Below is an example of an assignment assigned to students in the [Advanced Data Science Course at Johns Hopkins](http://jtleek.com/ads2020/). In this assignment students were asked to create a simpler machine learning analysis based on the case study. They were also asked to share their analysis with another student and write a summary and critique of the other student’s analysis. This provides an opportunity for the student to try out analysis skills with their own analysis work, and also work on their comprehension skills of reading other people’s work. It also shows how we can get different results with minor changes in our analyses.:

* [OCS Air Pollution Extension Assignment Example](https://github.com/advdatasci/homework12)

*Acknowledgments*

We would like to thank the following people for generously sharing the ways in which they used Open Case Studies materials in their teaching:

* [Dr. Shannon Ellis](https://www.shanellis.com/)
* [Dr. Jeff Leek](https://jtleek.com/)
* [Dr. Roger Peng](https://rdpeng.org/)

## 3.7 Session info

## R version 4.3.2 (2023-10-31)  
## Platform: x86\_64-pc-linux-gnu (64-bit)  
## Running under: Ubuntu 22.04.4 LTS  
##   
## Matrix products: default  
## BLAS: /usr/lib/x86\_64-linux-gnu/openblas-pthread/libblas.so.3   
## LAPACK: /usr/lib/x86\_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so; LAPACK version 3.10.0  
##   
## locale:  
## [1] LC\_CTYPE=en\_US.UTF-8 LC\_NUMERIC=C   
## [3] LC\_TIME=en\_US.UTF-8 LC\_COLLATE=en\_US.UTF-8   
## [5] LC\_MONETARY=en\_US.UTF-8 LC\_MESSAGES=en\_US.UTF-8   
## [7] LC\_PAPER=en\_US.UTF-8 LC\_NAME=C   
## [9] LC\_ADDRESS=C LC\_TELEPHONE=C   
## [11] LC\_MEASUREMENT=en\_US.UTF-8 LC\_IDENTIFICATION=C   
##   
## time zone: Etc/UTC  
## tzcode source: system (glibc)  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## loaded via a namespace (and not attached):  
## [1] compiler\_4.3.2 fastmap\_1.1.1 bookdown\_0.43 cli\_3.6.2   
## [5] htmltools\_0.5.7 tools\_4.3.2 yaml\_2.3.10 rmarkdown\_2.25   
## [9] knitr\_1.50 digest\_0.6.34 xfun\_0.52 rlang\_1.1.6   
## [13] evaluate\_1.0.4

# 4 Modifying open case studies

## 4.1 Learning Objectives

This chapter will cover how to modify the case studies to your own needs using the following methods:

* Modular case study use with the help of the [OCSdata package](https://github.com/opencasestudies/OCSdata)
* Modifying a case study with [GitHub](https://github.com/opencasestudies) and [RStudio](https://www.rstudio.com/)

## 4.2 Modular use

Some educators may find that only certain sections of a case study are relevant to their specific needs. For example, a statistics teacher may want students to practice the skills covered in the data analysis section, but doesn’t have time to go through the whole case study. The case studies are designed to allow for such use. This educator and their students may jump right in to any case study section without working through any previous sections. This is made possible because the data files are saved at the end of each section. These data files are made available on the case study’s GitHub repository and may also be downloaded with the help of the OCSdata package. See Chapter 2 for more details on the structure and organization of a case study data folder. The table below explains which data sub-folder and package function to use for each case study section.

| Data Folder | Case Study Section | OCSdata Function |
| --- | --- | --- |
| raw | Data Import | raw\_data |
| imported | Data Exploration, Data Wrangling | imported\_data |
| wrangled | Data Visualization, Data Analysis | wrangled\_csv, wrangled\_rda |
| simpler\_import | Data Import | simpler\_import\_data |
| extra | Suggested Homework (?) | extra\_data |

If you’d like to learn more about modular use of Open Case Studies you can read this [thesis](https://jscholarship.library.jhu.edu/handle/1774.2/66820) by former graduate student [Michael Breshock](https://mbreshock.github.io/) ([**breshock\_expanding\_2021?**](#ref-breshock_expanding_2021)).

### 4.2.1 Example of Modular Use

The following steps illustrate how one would skip to a specific case study section. The data analysis section from the “Opioids in United States” case study is used for this example, but these directions apply for any section in any case study.

#### 4.2.1.1 Steps for modular use

1. Use the table of contents to navigate to the section of interest. Click on the arrow that reads “If you skipped the previous sections click here.”
2. Follow the instructions provided to download the data files from the previous section either with OCSdata or manually through GitHub. We will now demonstrate each option.

#### 4.2.1.2 Downloading data with OCSdata:

1. First install and load the OCSdata package:
2. Now download the wrangled data into your R environment using the following function command:

*This command will download the wrangled data in .RDA format. You may also be downloaded wrangled data in .CSV format by replacing ‘wrangled\_rda’ with ‘wrangled\_csv’*

1. Load the RDA files with the following commands:

#### 4.2.1.3 Manually Download Data from GitHub:

1. Download the .RDA files available on the case study GitHub repository from [here](https://github.com/opencasestudies/ocs-bp-opioid-rural-urban/tree/master/data/wrangled).

*The CSV versions of the files may also be downloaded here, if preferred*

1. Move the data files from your ‘Downloads’ folder to your R session’s current working directory (you can see what this is with getwd()).
2. Load the RDA files with the following commands:
3. All the data you need to work through the current section is now loaded into your environment. You are ready to work through the section of interest, without needing to work through any of the previous sections.

## 4.3 Modify a case study

The case studies are written in [R Markdown](https://rmarkdown.rstudio.com/) documents and developed within an RStudio project. R Markdown documents are denoted with the file extension “.Rmd” and allow for the inclusion of code chunks and outputs in a written report. They are written using [Markdown syntax](https://raw.githubusercontent.com/rstudio/cheatsheets/main/rmarkdown.pdf). RStudio projects are used to organize the case studies. The [knitr](https://yihui.org/knitr/) package “knits” the case study written in R Markdown and outputs the document as an HTML file. Open Case Studies uses these HTML files to post the case studies online. The entire case study project is contained in a GitHub repository which allows for easy distribution and version control. [GitHub Pages](https://pages.github.com/) is used to host the case study webpage from the case study repository.

Modifying a case study requires the following simple steps:

1. Use the OCSdata package zip\_ocs() function to download the case study files without our git history. Alternatively, if you are familiar with GitHub and wish to you can clone or fork the case study repository from GitHub. See [this section](https://www.opencasestudies.org/OCS_Guide/open-case-study-infrastructure.html#fork-or-clone-the-case-study-repository) of chapter 2 for more information.
2. In the repository folder, open the case study .Rproj file to open the project up in RStudio.
3. Edit the sections to be modified in the index.Rmd file.
4. Save your changes, then click on the “Knit” drop down menu in the top left corner of RStudio. Choose which file format you’d like to knit to.
5. Distribute your modified case study as you please!

You can share HTML, PDF and Word files directly with your students or you can host your case study for free on GitHub using [GitHub pages](https://pages.github.com/) ([**pages\_github?**](#ref-pages_github)).

To do so, you will need to set up your version of the case study on GitHub. If you are new to GitHub, also be sure to check out ([**happygitwithr?**](#ref-happygitwithr)).

These steps are demonstrated in the following video about modifying case studies:

## 4.4 Session info

## R version 4.3.2 (2023-10-31)  
## Platform: x86\_64-pc-linux-gnu (64-bit)  
## Running under: Ubuntu 22.04.4 LTS  
##   
## Matrix products: default  
## BLAS: /usr/lib/x86\_64-linux-gnu/openblas-pthread/libblas.so.3   
## LAPACK: /usr/lib/x86\_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so; LAPACK version 3.10.0  
##   
## locale:  
## [1] LC\_CTYPE=en\_US.UTF-8 LC\_NUMERIC=C   
## [3] LC\_TIME=en\_US.UTF-8 LC\_COLLATE=en\_US.UTF-8   
## [5] LC\_MONETARY=en\_US.UTF-8 LC\_MESSAGES=en\_US.UTF-8   
## [7] LC\_PAPER=en\_US.UTF-8 LC\_NAME=C   
## [9] LC\_ADDRESS=C LC\_TELEPHONE=C   
## [11] LC\_MEASUREMENT=en\_US.UTF-8 LC\_IDENTIFICATION=C   
##   
## time zone: Etc/UTC  
## tzcode source: system (glibc)  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## loaded via a namespace (and not attached):  
## [1] compiler\_4.3.2 fastmap\_1.1.1 bookdown\_0.43 cli\_3.6.2   
## [5] htmltools\_0.5.7 tools\_4.3.2 yaml\_2.3.10 rmarkdown\_2.25   
## [9] knitr\_1.50 digest\_0.6.34 xfun\_0.52 rlang\_1.1.6   
## [13] evaluate\_1.0.4

# 5 New Case Studies - Building and Contributing

## 5.1 Learning Objectives

In this chapter we will discuss: - Creating your own case study with our [template](https://github.com/opencasestudies/ocs-bp-template) - The guidelines for how to publish your own case studies as part of our project

## 5.2 Create a Case Study

### 5.2.1 Template Case Study

A template case study is available in a repository on our GitHub page at [github.com/opencasestudies/ocs-bp-template](https://github.com/opencasestudies/ocs-bp-template). This template contains the basic skeletal structure used for our case studies. Creating a new case study with the template is very similar to modifying an existing case study:

1. Click on the “use this template” button at [opencasestudies/ocs-bp-template](https://github.com/opencasestudies/ocs-bp-template).
2. Open the index.Rmd file in RStudio.
3. Add the case study content to the index.Rmd file. Use the instructions provided in this file to add different elements such as images and videos.
4. Save your changes and knit the case study to the preferred file format.
5. Distribute the knitted case study as you please!

All of these steps are demonstrated in the following video:

If you’d like to learn more about MakeCaseStudies, you can read this [thesis](https://jscholarship.library.jhu.edu/handle/1774.2/66820) by former graduate student [Michael Breshock](https://mbreshock.github.io/) ([**breshock\_expanding\_2021?**](#ref-breshock_expanding_2021)).

## 5.3 Case Study Libraries

To help curate case studies for educators and learners, we would like to encourage others to help us expand our library of case studies.

We recognize that other educators may have case studies in various stages of development that could be helpful to others.

We hope to facilitate sharing of these case studies on the Open Case Studies project.

However, we unfortunately do not have the resources or time to help integrate all case studies we receive into the official Open Case Studies library

For this reason, we have outlined the following two modes of submission:

1. **Submissions for integration into the official library:**

With this submission type, the Open Case Studies team will work with you starting from a case study idea through development and peer review stages, with the ultimate goal of adding your case study to the Open Case Studies project. Please note that we will only be able to accept one or two case studies a year to the official repository, as we are currently doing this as a passion project and have limited time to review and collaborate on the case studies. Please note that it is not required that case studies explore a public health topic, instead the just need to explore an interesting and timely problem.

1. **Submissions for a publicly available community repository:**

We hope to also create a more casual community library that allows others to share their work more easily and quickly. This repository will contain case studies submitted by educators but not included in the official Open Case Studies project library. The submission process is much simpler.

Please note that case studies submitted to either collection need to be open source and with a license that permits sharing of derivatives to allow other educators to make use of our case studies as needed.

The original others will retain the rights of their case studies, but need to provide citation information for others to attribute their case studies. For official case studies, the OCS team may help modify the case study and may therefore be included as authors.

## 5.4 Submission Process

The process of submitting community case studies, involves filling out a simple form. The process of submitting to our official library involves ensuring that your case study meets our more lengthy requirements and a review process. Now we will describe the submission process for each library

### 5.4.1 Community Library Submissions

These case studies will be shared publicly on the Open Case Studies Community Repository for the benefit of other educators and learners, with minimal review from the Open Case Studies team. To submit a case study to the Community Repository, please fill out the form below:

[**Open Case Studies Community Repository Submission Form**](https://forms.gle/BgkQMbb13wtaYHMo6)

### 5.4.2 Official Library Submissions

If you have not yet created a case study, but want to create one for our official library, please complete the form below to contact the Open Case Studies team regarding a new case study idea. The team will get back to do as soon as we can.

[**Open Case Studies New Idea Offical Repository Submission Form**](https://forms.gle/2EEYfnTJkNB7Nn9dA)

Ultimately your case study will be submitted using an issue template to the GitHub repository for this guide, located at [this link](https://github.com/opencasestudies/OCS_Guide/issues/new?assignees=carriewright11%2Cstephaniehicks&labels=&template=new-official-case-study-submission.md&title=). The form indicates what is required for case studies to be considered for inclusion in our official collection, including peer review. We will now also describe some of these requirements in more depth in the next section.

## 5.5 Offical Case Study Guidelines

To ensure that the submitted case studies provide the most benefit to the community, we hope that they adhere as much as possible to the core structure and ideas of the Open Case Studies project. We provide the guidelines detailed below for submission to the official library. Furthermore, because the Open Case Studies team does not have the infrastructure to perform a full content review of submitted case studies, we ask that you find at least two independent reviewers to comment about the validity of the content and conclusions presented in your case study.

**Programming language:** Case studies should be written in open source programming languages (such as R or python).

**Data:**

* Case studies should use data that is publicly available or can be made publicly available. Please ensure that you are allowed to make the data public if it is not already.
* Transparent descriptions of data sources and how data was generated should be included when possible.

**Core sections**: all case studies should include the following sections consistent with the published case studies. See [Chapter 1, Case Study Anatomy](https://www.opencasestudies.org/OCS_Guide/introduction.html#open-case-studies-anatomy) for a detailed description on what each section should include:

* **Getting Started:** *Case study context, Study motivation, Main question, Learning objectives, Study limitations.* This section will outline previous literature and overarching questions in the field that make this case study question important, providing motivation for conducting the analysis in the case study. It will specify the exact question that the case study endeavors to address as well as the defining the statistical and data analysis learning objectives for the case study. It will also discuss the limitations of the investigation that may prevent the analysis from answering the main question of interest.
* **Analyzing the Data:** *Data description, Import and exploration, Wrangling, Visualization, Analysis.* This section includes the bulk of the data analysis, beginning with a detailed description of the data used in the case study, how it was generated, and where it was obtained. Then it walks through, step-by-step, the data import, exploration, wrangling, procedures. Finally, it demonstrates the visualization and statistical analysis steps used to address the case study question.
* **Wrapping-up:** *Analysis conclusions, Case study summary, Next steps, Homework, Additional information.* This section synthesized conclusions to the main study question based on the data analysis results demonstrated in the previous sections. It summarizes the analysis steps and potential next steps for further exploration, which can be suggested as homework. If applicable, additional information relevant to further study can be provided here.

**Content:** Despite often being motivated by articles, case studies are not intended to demonstrate the methods of a paper - they are intended as an educational resource where users are guided through the data science process.

* Links to literature or other sources to motivate the scientific topic of the case study should be included where possible.
* Case studies should aim to describe the decision making process involved in performing data science related tasks.

**Attribution:** all outside resources used in the case study should be referenced appropriately.

* Case studies should include disclaimers and appropriate license agreements.
* All included images (that are not original to the case study) should include relevant sources.

**Data:**

* All data files should be saved in a folder named “data” within the project directory. All files should be contained in a sub-folder using the design scheme outlined in this diagram:

This diagram is from former graduate student [Michael Breshock’s](https://mbreshock.github.io/) [thesis](https://jscholarship.library.jhu.edu/handle/1774.2/66820) on the Open Case Studies. In this thesis you can find a detailed description on the organization and structure of the case study data files ([**breshock\_expanding\_2021?**](#ref-breshock_expanding_2021)).

* The data folder is required to have at least raw, imported, and wrangled data files. This allows users to skip sections when convenient.
* The raw data folder consists of data files as they came from the source. These are used at the beginning of the case study analysis at the Data Import section.
* Imported data is the version of the data after it has already been imported into R. These should be in the form of R Data files (extensions include .RData, .Rda, and .Rds). To create these files, use the save() function in R after importing the raw data.
* Wrangled data is the version of the data after it has been cleaned and is ready for analysis. Use the save() function in R after wrangling the data to create these files. This data should be provided in both RDA and CSV format.
* The next two data sub-folders are optional and used as needed:
  + If the raw data files for your case study come in a format that requires a complicated data import process (such as web scraping) you may consider providing the raw data in an import friendly format such as CSV or XLS(X). Place these files in the “simpler\_import” sub-folder.
  + If there are any raw data files from your source that are not used in the case study analysis, but could be analyzed in a similar fashion, please include these in the “extra” sub-folder.

## 5.6 Session info

## R version 4.3.2 (2023-10-31)  
## Platform: x86\_64-pc-linux-gnu (64-bit)  
## Running under: Ubuntu 22.04.4 LTS  
##   
## Matrix products: default  
## BLAS: /usr/lib/x86\_64-linux-gnu/openblas-pthread/libblas.so.3   
## LAPACK: /usr/lib/x86\_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so; LAPACK version 3.10.0  
##   
## locale:  
## [1] LC\_CTYPE=en\_US.UTF-8 LC\_NUMERIC=C   
## [3] LC\_TIME=en\_US.UTF-8 LC\_COLLATE=en\_US.UTF-8   
## [5] LC\_MONETARY=en\_US.UTF-8 LC\_MESSAGES=en\_US.UTF-8   
## [7] LC\_PAPER=en\_US.UTF-8 LC\_NAME=C   
## [9] LC\_ADDRESS=C LC\_TELEPHONE=C   
## [11] LC\_MEASUREMENT=en\_US.UTF-8 LC\_IDENTIFICATION=C   
##   
## time zone: Etc/UTC  
## tzcode source: system (glibc)  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## loaded via a namespace (and not attached):  
## [1] compiler\_4.3.2 fastmap\_1.1.1 bookdown\_0.43 cli\_3.6.2   
## [5] htmltools\_0.5.7 tools\_4.3.2 yaml\_2.3.10 rmarkdown\_2.25   
## [9] knitr\_1.50 digest\_0.6.34 xfun\_0.52 rlang\_1.1.6   
## [13] evaluate\_1.0.4

# About the Authors

These credits are based on our [course contributors table guidelines](https://www.ottrproject.org/more_features.html#giving-credits-to-contributors).

| Credits | Names |
| --- | --- |
| **Pedagogy** |  |
| Lead Content Instructor(s) | [FirstName LastName](link%20to%20personal%20website) |
| Lecturer(s) (include chapter name/link in parentheses if only for specific chapters) - make new line if more than one chapter involved | Delivered the course in some way - video or audio |
| Content Author(s) (include chapter name/link in parentheses if only for specific chapters) - make new line if more than one chapter involved | If any other authors besides lead instructor |
| Content Contributor(s) (include section name/link in parentheses) - make new line if more than one section involved | Wrote less than a chapter |
| Content Editor(s)/Reviewer(s) | Checked your content |
| Content Director(s) | Helped guide the content direction |
| Content Consultants (include chapter name/link in parentheses or word “General”) - make new line if more than one chapter involved | Gave high level advice on content |
| Acknowledgments | Gave small assistance to content but not to the level of consulting |
| **Production** |  |
| Content Publisher(s) | Helped with publishing platform |
| Content Publishing Reviewer(s) | Reviewed overall content and aesthetics on publishing platform |
| **Technical** |  |
| Course Publishing Engineer(s) | Helped with the code for the technical aspects related to the specific course generation |
| Template Publishing Engineers | [Candace Savonen](https://www.cansavvy.com/), [Carrie Wright](https://carriewright11.github.io/), [Ava Hoffman](https://www.avahoffman.com/) |
| Publishing Maintenance Engineer | [Candace Savonen](https://www.cansavvy.com/) |
| Technical Publishing Stylists | [Carrie Wright](https://carriewright11.github.io/), [Ava Hoffman](https://www.avahoffman.com/), [Candace Savonen](https://www.cansavvy.com/), [Katherine Cox](https://katherinecox.github.io/) |
| Package Developers ([ottrpal](https://github.com/ottrproject/ottrpal)) | [Candace Savonen](https://www.cansavvy.com/), [Ava Hoffman](https://www.avahoffman.com/), [Howard Baek](https://www.linkedin.com/in/howard-baik/), [Kate Isaac](https://kweav.github.io/), [Carrie Wright](https://carriewright11.github.io/), [John Muschelli](https://johnmuschelli.com/) |
| **Art and Design** |  |
| Illustrator(s) | Created graphics for the course |
| Figure Artist(s) | Created figures/plots for course |
| Videographer(s) | Filmed videos |
| Videography Editor(s) | Edited film |
| Audiographer(s) | Recorded audio |
| Audiography Editor(s) | Edited audio recordings |
| **Funding** |  |
| Funder(s) | Institution/individual who funded course including grant number |
| Funding Staff | Staff members who help with funding |

## ─ Session info ───────────────────────────────────────────────────────────────  
## setting value  
## version R version 4.3.2 (2023-10-31)  
## os Ubuntu 22.04.4 LTS  
## system x86\_64, linux-gnu  
## ui X11  
## language (EN)  
## collate en\_US.UTF-8  
## ctype en\_US.UTF-8  
## tz Etc/UTC  
## date 2025-10-16  
## pandoc 3.1.1 @ /usr/local/bin/ (via rmarkdown)  
##   
## ─ Packages ───────────────────────────────────────────────────────────────────  
## package \* version date (UTC) lib source  
## bookdown 0.43 2025-04-15 [1] CRAN (R 4.3.2)  
## cachem 1.0.8 2023-05-01 [1] RSPM (R 4.3.0)  
## cli 3.6.2 2023-12-11 [1] RSPM (R 4.3.0)  
## devtools 2.4.5 2022-10-11 [1] RSPM (R 4.3.0)  
## digest 0.6.34 2024-01-11 [1] RSPM (R 4.3.0)  
## ellipsis 0.3.2 2021-04-29 [1] RSPM (R 4.3.0)  
## evaluate 1.0.4 2025-06-18 [1] CRAN (R 4.3.2)  
## fastmap 1.1.1 2023-02-24 [1] RSPM (R 4.3.0)  
## fs 1.6.3 2023-07-20 [1] RSPM (R 4.3.0)  
## glue 1.7.0 2024-01-09 [1] RSPM (R 4.3.0)  
## htmltools 0.5.7 2023-11-03 [1] RSPM (R 4.3.0)  
## htmlwidgets 1.6.4 2023-12-06 [1] RSPM (R 4.3.0)  
## httpuv 1.6.14 2024-01-26 [1] RSPM (R 4.3.0)  
## knitr 1.50 2025-03-16 [1] CRAN (R 4.3.2)  
## later 1.3.2 2023-12-06 [1] RSPM (R 4.3.0)  
## lifecycle 1.0.4 2023-11-07 [1] RSPM (R 4.3.0)  
## magrittr 2.0.3 2022-03-30 [1] RSPM (R 4.3.0)  
## memoise 2.0.1 2021-11-26 [1] RSPM (R 4.3.0)  
## mime 0.12 2021-09-28 [1] RSPM (R 4.3.0)  
## miniUI 0.1.1.1 2018-05-18 [1] RSPM (R 4.3.0)  
## pkgbuild 1.4.3 2023-12-10 [1] RSPM (R 4.3.0)  
## pkgload 1.3.4 2024-01-16 [1] RSPM (R 4.3.0)  
## profvis 0.3.8 2023-05-02 [1] RSPM (R 4.3.0)  
## promises 1.2.1 2023-08-10 [1] RSPM (R 4.3.0)  
## purrr 1.0.2 2023-08-10 [1] RSPM (R 4.3.0)  
## R6 2.5.1 2021-08-19 [1] RSPM (R 4.3.0)  
## Rcpp 1.0.12 2024-01-09 [1] RSPM (R 4.3.0)  
## remotes 2.4.2.1 2023-07-18 [1] RSPM (R 4.3.0)  
## rlang 1.1.6 2025-04-11 [1] CRAN (R 4.3.2)  
## rmarkdown 2.25 2023-09-18 [1] RSPM (R 4.3.0)  
## sessioninfo 1.2.2 2021-12-06 [1] RSPM (R 4.3.0)  
## shiny 1.8.0 2023-11-17 [1] RSPM (R 4.3.0)  
## stringi 1.8.3 2023-12-11 [1] RSPM (R 4.3.0)  
## stringr 1.5.1 2023-11-14 [1] RSPM (R 4.3.0)  
## urlchecker 1.0.1 2021-11-30 [1] RSPM (R 4.3.0)  
## usethis 2.2.3 2024-02-19 [1] RSPM (R 4.3.0)  
## vctrs 0.6.5 2023-12-01 [1] RSPM (R 4.3.0)  
## xfun 0.52 2025-04-02 [1] CRAN (R 4.3.2)  
## xtable 1.8-4 2019-04-21 [1] RSPM (R 4.3.0)  
## yaml 2.3.10 2024-07-26 [1] CRAN (R 4.3.2)  
##   
## [1] /usr/local/lib/R/site-library  
## [2] /usr/local/lib/R/library  
##   
## ──────────────────────────────────────────────────────────────────────────────

# 6 References