

Effective Data Science

Zak Varty

2022-07-12

Contents

Preface	5
Acknowledgements	5
0.1 Usage	5
0.2 Render book	6
0.3 Preview book	6
1 Introduction	7
1.1 Module Description	7
1.2 Allocation of Study Hours	7
2 Data Science Workflows	9
2.1 Chapters and sub-chapters	9
2.2 Captioned figures and tables	9
3 Aquiring and Sharing Data	13
4 Cleaning, Exploring and Visualising	15
4.1 Footnotes	15
4.2 Citations	15
5 Getting Your Work into Production	17
5.1 Equations	17
5.2 Theorems and proofs	17
5.3 Callout blocks	17
6 Wider Context and Ethics	19
6.1 Publishing	19
6.2 404 pages	19
6.3 Metadata for sharing	19

Preface

These notes are intended for students on the course **MATH70076: Data Science** in the academic year 2022/23.

As the course is scheduled to take place over five weeks, the suggested schedule is

- 1st week: Chapters 1 and 2
- 2nd week: Chapter 3
- 3rd week: Chapter 4
- 4th week: Chapters 5
- 5th week: Chapter 6

Acknowledgements

These notes were created by Dr Zak Varty based on a lecture series at Imperial College London that was developed by Dr Purvasha Chakravarti and others.

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports; for example, a math equation $a^2 + b^2 = c^2$.

0.1 Usage

Each **bookdown** chapter is an .Rmd file, and each .Rmd file can contain one (and only one) chapter. A chapter *must* start with a first-level heading: **# A good chapter**, and can contain one (and only one) first-level heading.

Use second-level and higher headings within chapters like: **## A short section** or **### An even shorter section**.

The `index.Rmd` file is required, and is also your first book chapter. It will be the homepage when you render the book.

0.2 Render book

You can render the HTML version of this example book without changing anything:

1. Find the **Build** pane in the RStudio IDE, and
2. Click on **Build Book**, then select your output format, or select “All formats” if you’d like to use multiple formats from the same book source files.

Or build the book from the R console:

To render this example to PDF as a `bookdown::pdf_book`, you’ll need to install XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): <https://yihui.org/tinytex/>.

0.3 Preview book

As you work, you may start a local server to live preview this HTML book. This preview will update as you edit the book when you save individual `.Rmd` files. You can start the server in a work session by using the RStudio add-in “Preview book”, or from the R console:

Chapter 1

Introduction

1.1 Module Description

Model building and evaluation are necessary but not sufficient skills for the effective practice of data science. In this module you will develop the technical and personal skills that are required to work successfully as a data scientist within an organisation.

During this module you will critically explore how to:

- effectively scope and manage a data science project;
- efficiently acquire, manipulate, and present data;
- interpret and explain your work for a variety of stakeholders;
- ensure that your work can be put into production;
- assess the ethical implications of your work as a data scientist.

This interdisciplinary course will draw from fields including statistics, computing, management science and data ethics. Each topic will be investigated through a selection of lecture videos, conference presentations and academic papers, hands-on lab exercises, and readings on industry best-practices from recognised professional bodies.

This course will be assessed entirely by coursework, reflecting the practical and pragmatic nature of the course material.

1.2 Allocation of Study Hours

Lectures: 10 Hours (2 hours per week)

Group Teaching: 5 Hours (1 hour per week)

Lab / Practical: 5 hours (1 hour per week)

Independent Study: 105 hours (17 hours per week + 20 hours coursework)

1.2.1 Learning outcomes

On successful completion of this module students should be able to:

1. Independently scope and manage a data science project;
2. Source data from the internet through web scraping and APIs;
3. Clean, explore and visualise data, justifying and documenting the decisions made;
4. Evaluate the need for (and implement) approaches that are explainable, reproducible and scalable;
5. Appraise the ethical implications of a data science projects, particularly the risks of compromising privacy or fairness and the potential to cause harm.

1.2.2 Module Content

This module will cover:

- effective management of a data science project;
- open and reproducible work flows;
- sourcing and preparing data for analysis;
- exploratory and expository data visualisation;
- minimal requirements for models to go into production;
- ethical implications of modern data science.

An unnumbered section

Chapters and sections are numbered by default. To un-number a heading, add a `{.unnumbered}` or the shorter `{-}` at the end of the heading, like in this section.

Chapter 2

Data Science Workflows

Cross-references make it easier for your readers to find and link to elements in your book.

2.1 Chapters and sub-chapters

There are two steps to cross-reference any heading:

1. Label the heading: `# Hello world {#nice-label}`.
 - Leave the label off if you like the automated heading generated based on your heading title: for example, `# Hello world = # Hello world {#hello-world}`.
 - To label an un-numbered heading, use: `# Hello world {-#nice-label}` or `{# Hello world .unnumbered}`.
2. Next, reference the labeled heading anywhere in the text using `\@ref(nice-label)`; for example, please see Chapter ??.
 - If you prefer text as the link instead of a numbered reference use: any text you want can go here.

2.2 Captioned figures and tables

Figures and tables *with captions* can also be cross-referenced from elsewhere in your book using `\@ref(fig:chunk-label)` and `\@ref(tab:chunk-label)`, respectively.

See Figure 2.1.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Don't miss Table 2.1.



Figure 2.1: Here is a nice figure!

```
knitr::kable(  
  head(pressure, 10), caption = 'Here is a nice table!',  
  booktabs = TRUE  
)
```

Table 2.1: Here is a nice table!

temperature	pressure
0	0.0002
20	0.0012
40	0.0060
60	0.0300
80	0.0900
100	0.2700
120	0.7500
140	1.8500
160	4.2000
180	8.8000

Chapter 3

Aquiring and Sharing Data

You can add parts to organize one or more book chapters together. Parts can be inserted at the top of an .Rmd file, before the first-level chapter heading in that same file.

Add a numbered part: `# (PART) Act one {-}` (followed by `# A chapter`)

Add an unnumbered part: `# (PART*) Act one {-}` (followed by `# A chapter`)

Add an appendix as a special kind of un-numbered part: `# (APPENDIX) Other stuff {-}` (followed by `# A chapter`). Chapters in an appendix are prepended with letters instead of numbers.

Chapter 4

Cleaning, Exploring and Visualising

4.1 Footnotes

Footnotes are put inside the square brackets after a caret `^[]`. Like this one ¹.

4.2 Citations

Reference items in your bibliography file(s) using `@key`.

For example, we are using the **bookdown** package (Xie, 2022) (check out the last code chunk in `index.Rmd` to see how this citation key was added) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015) (this citation was added manually in an external file `book.bib`). Note that the `.bib` files need to be listed in the `index.Rmd` with the YAML `bibliography` key.

The `bs4_book` theme makes footnotes appear inline when you click on them. In this example book, we added `csl: chicago-fullnote-bibliography.csl` to the `index.Rmd` YAML, and include the `.csl` file. To download a new style, we recommend: <https://www.zotero.org/styles/>

The RStudio Visual Markdown Editor can also make it easier to insert citations: <https://rstudio.github.io/visual-markdown-editing/#/citations>

¹This is a footnote.

Chapter 5

Getting Your Work into Production

5.1 Equations

Here is an equation.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (5.1)$$

You may refer to using `\@ref{eq:binom}`, like see Equation (5.1).

5.2 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref{thm:tri}`, for example, check out this smart theorem 5.1.

Theorem 5.1. *For a right triangle, if c denotes the length of the hypotenuse and a and b denote the lengths of the **other** two sides, we have*

$$a^2 + b^2 = c^2$$

Read more here <https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html>.

5.3 Callout blocks

The `bs4_book` theme also includes special callout blocks, like this `.rmdnote`.

You can use **markdown** inside a block.

```
head(beaver1, n = 5)
#>   day time  temp activ
#> 1 346  840 36.33     0
#> 2 346  850 36.34     0
#> 3 346  900 36.35     0
#> 4 346  910 36.42     0
#> 5 346  920 36.55     0
```

It is up to the user to define the appearance of these blocks for LaTeX output.

You may also use: `.rmdcaution`, `.rmdimportant`, `.rmdtip`, or `.rmdwarning` as the block name.

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: <https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html>

Chapter 6

Wider Context and Ethics

6.1 Publishing

HTML books can be published online, see: <https://bookdown.org/yihui/bookdown/publishing.html>

6.2 404 pages

By default, users will be directed to a 404 page if they try to access a webpage that cannot be found. If you'd like to customize your 404 page instead of using the default, you may add either a `_404.Rmd` or `_404.md` file to your project root and use code and/or Markdown syntax.

6.3 Metadata for sharing

Bookdown HTML books will provide HTML metadata for social sharing on platforms like Twitter, Facebook, and LinkedIn, using information you provide in the `index.Rmd` YAML. To setup, set the `url` for your book and the path to your `cover-image` file. Your book's `title` and `description` are also used.

This `bs4_book` provides enhanced metadata for social sharing, so that each chapter shared will have a unique description, auto-generated based on the content.

Specify your book's source repository on GitHub as the `repo` in the `_output.yml` file, which allows users to view each chapter's source file or suggest an edit. Read more about the features of this output format here:

https://pkgs.rstudio.com/bookdown/reference/bs4_book.html

Or use:

?bookdown::bs4_book

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2022). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.26.