A Little Book of Data Analysis

Leighton Pritchard

2024-01-05

Table of contents

Preface to *A Little Book of Bioinformatics*

Welcome to A Little Book of Bioinformatics. This is an online book, under continual development, which I am building as and when topics come to mind or prominence.

My goal is that this online book will come to be a fairly transparent and honest reference for students in bioinformatics, and maybe for some researchers, too.

I would be very grateful for feedback by email or through the GitHub repository Issues page

1 Introduction

The Introduction page is intended as a short introduction to the book.

Like most Quarto books, this is a book created from markdown and executable code.

This kind of book is an example of literate programming - the intertwining of nicely-formatted text and images, and executable code. For example, the `R` code cell below executes and produces output when the book is compiled:

```
1 + 1
```

```
[1] 2
```

But the `R` code cell below does not:

```
summary(cars)
```

See Knuth (1984) for additional discussion of literate programming.

Part I

Early Section

This `.qmd` file introduces a `Part` of the Quarto book. We use the `{#sec-REFERENCE}` option to make it cross-referenceable elsewhere in the text, and the `{.unnumbered}` option to avoid giving it a section number.

2 Early Section Topic

This `.qmd` file represents some topic-related text. We use the `{#sec-REFERENCE}` option to make it cross-referenceable elsewhere in the text.

Part II

Late Section

This `.qmd` file introduces a `Part` of the Quarto book. We use the `{#sec-REFERENCE}` option to make it cross-referenceable elsewhere in the text, and the `{.unnumbered}` option to avoid giving it a section number.

3 R Playground

```
#| context: setup

library(ggplot2)
library(palmerpenguins)
library(tidyverse)
```

3.1 Introduction

This page provides a WebR cell for you to use as a playground to experiment with some example datasets. You can use this page to explore data management and visualisation in R.

3.2 Playground

```
# Use this WebR cell to experiment with some practice biological datasets
```

3.3 Things you can do

This WebR instance has three packages installed:

- ggplot2
- GGally
- tidyverse
- palmerpenguins

Open the callout boxes below to see some examples you can try in the code cell above.

Investigate Palmer's Penguins

The `penguins` dataset contains data about three different species of penguins. Take a look at the format of the dataset:

```
glimpse(penguins)
```

You'll see there are eight variables, including `species`, `weight`, `sex`, etc. - some of these variables are *categorical* (i.e. a category, like `species`), and others are *continuous* (i.e. numerical). You can see a visual overview of how the data is related using the `plot()` function:

```
plot(penguins)
```

We can visualise the number of penguins of each species in a bar chart:

```
fig <- ggplot(penguins, aes(species, fill=species)) +
         geom_bar()
fig
```

And break this down in a facet plot, by sex:

```
fig <- ggplot(penguins, aes(species, fill=species)) +
         geom_bar() +
         facet_wrap(~sex)
fig
```

We can make a box and whisker plot of penguin body mass by species:

```
fig <- ggplot(penguins, aes(x=species, y=body_mass_g, fill=species)) +
         geom_boxplot()
fig
```

And plot the body mass for each sex side-by-side

```
fig <- ggplot(penguins, aes(x=species, y=body_mass_g, fill=sex)) +
         geom_boxplot()
fig
```

We can investigate correlations, such as between body mass and flipper length:

```
fig <- ggplot(penguins, aes(x=body_mass_g, y=flipper_length_mm)) +
         geom_point()
fig
```

We can colour datapoints by species:

```
fig <- ggplot(penguins, aes(x=body_mass_g, y=flipper_length_mm, colour=species)) +
         geom_point()
fig
```

And fit a linear regression to each species separately:

```
fig <- ggplot(penguins, aes(x=body_mass_g, y=flipper_length_mm, colour=species)) +
         geom_point() +
         geom_smooth(method="lm")
fig
```

Note

`R` comes with a number of example datasets you can practice with, including:

- `mtcars`: fuel consumption and other statistic for 32 automobiles
- `Titanic`: information on the fate of passengers on the fatal maiden voyage of the ocean liner *Titanic*

You can see a full list by running the command

```
library(help = "datasets")
```

References

Knuth, Donald E. 1984. “Literate Programming.” *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.