Data Science and Data Analytics

The Data Science Workflow IV – Communicate

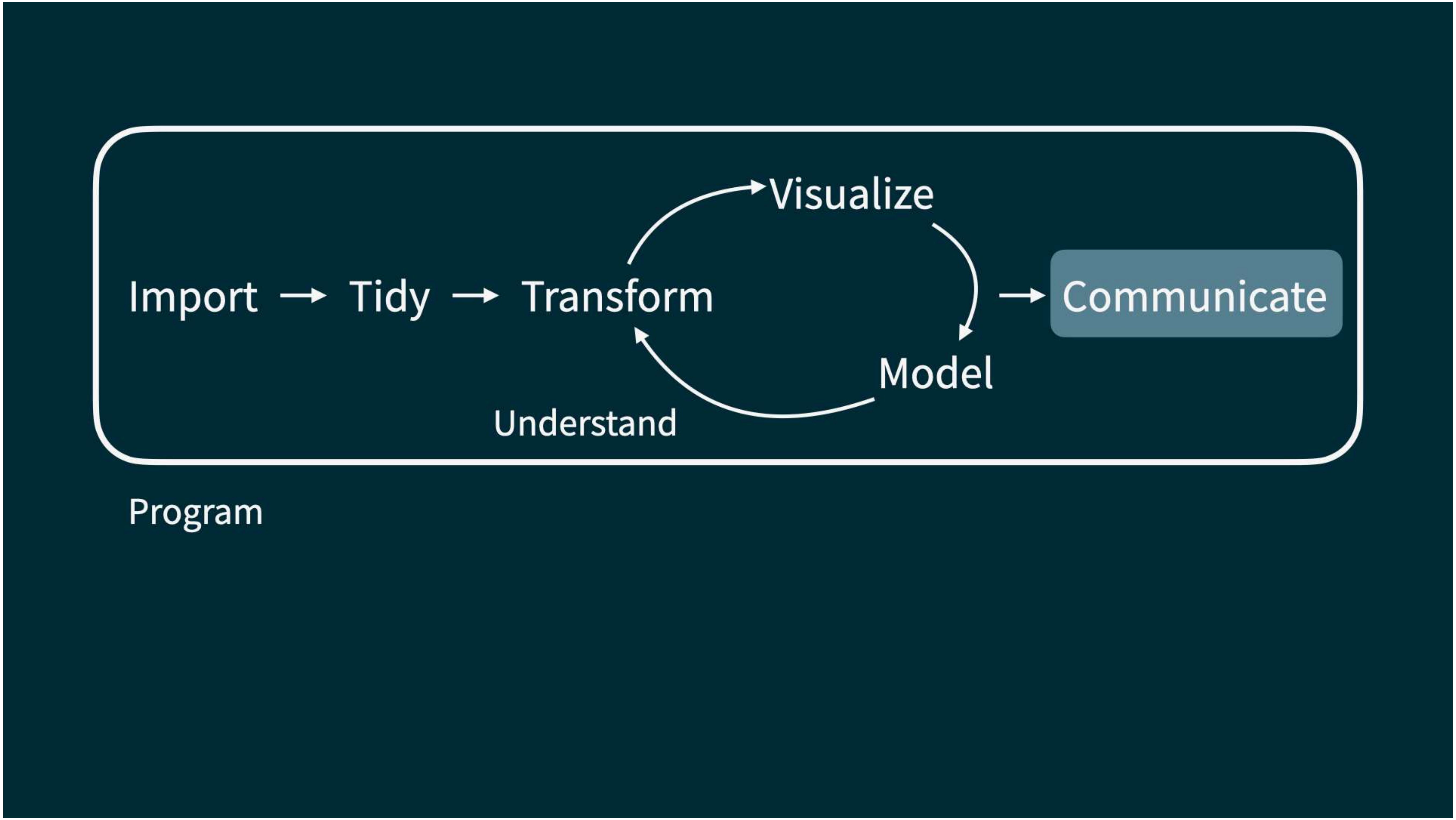
Julian Amon, PhD

Charlotte Fresenius Privatuniversität

May 28, 2025

Introduction to Quarto

The Data Science workflow – Communicate



Putting it all together…

- The final product of a data science project is often a **report**, which can take multiple different forms:
  - a scientific **publication** (typically PDF)
  - a **technical report** for your company (typically Word, Pages or PDF)
  - a **presentation** (typically PowerPoint, Keynote or PDF)
  - a **blog post** (typically HTML)
  - …
- Depending on the audience, such reports may include
  - **text** to describe project, methods and findings
  - **figures and tables** to describe and visualize data and results
  - **code** you ran to get to these results

Putting it all together…

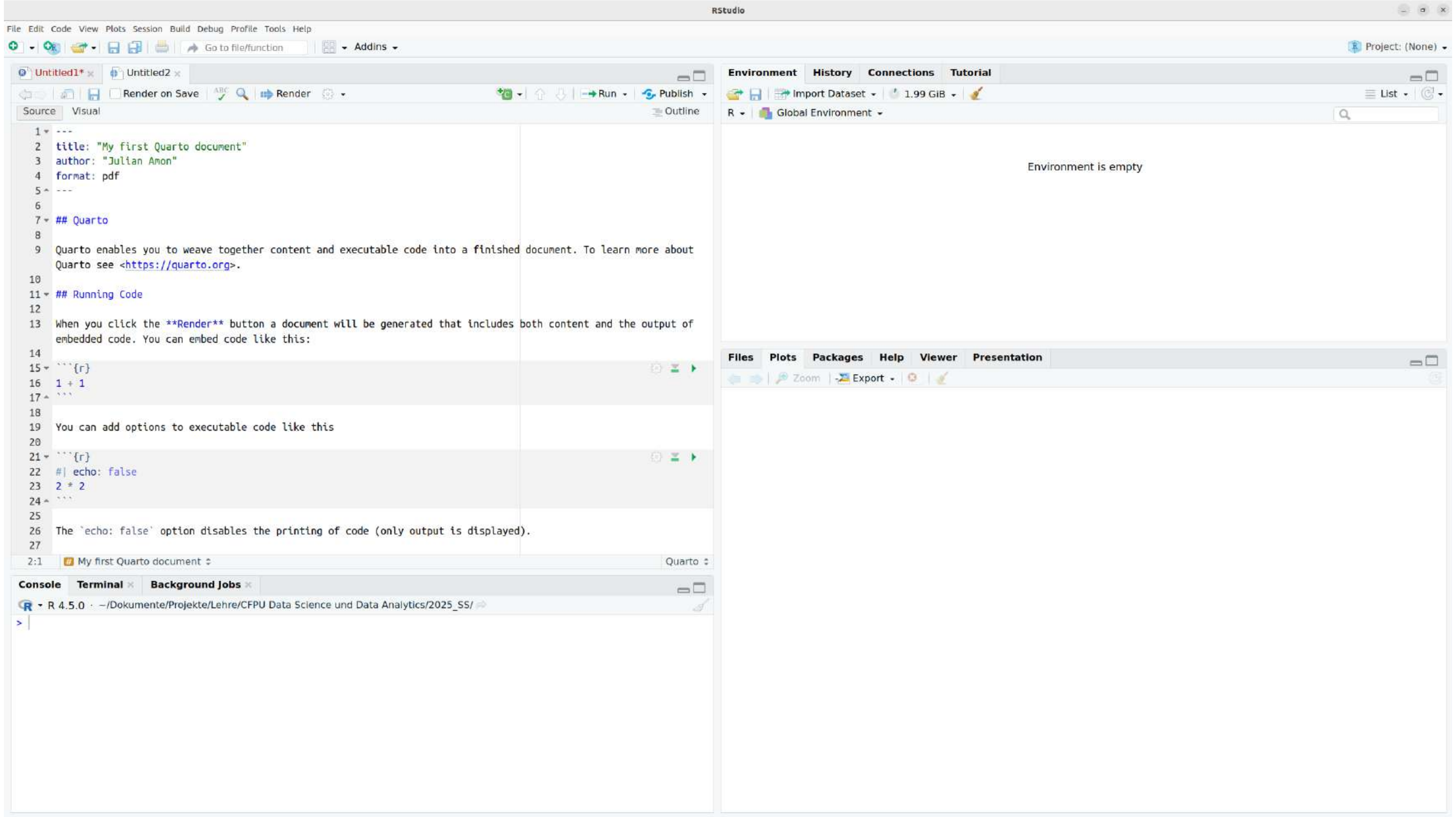
- Now, imagine…
  - you have to **re-run your entire analysis** with a new data set because the initial one was wrong in some way.
  - you realize that a mistake was made and the code needs to be fixed and re-run entirely.
  - a colleague from another department wants to **reproduce** the results from your report using your data and code.
- Situations like these are very commonplace for a data scientist. It is for this reason that we need a framework that integrates text, figures, tables, code and data to create **reproducible reports**.
- Such a framework is given by **Quarto**.

What is Quarto?

- **Quarto** is a format for so-called **literate programming documents**. Literate programming weaves textual documentation and comments together with code, producing a document that is optimized for human understanding.
- The basic workflow is very simple:
    - Create a **Quarto file** (ends in `.qmd`)
    - In this file, specify the output format (PDF, HTML, Word, …) as well as metadata about the document, such as a title, the name of the author, date, etc.
    - Add your code and text.
    - **Compile** the document into the final report. In this process, Quarto runs all your code, creates all the plots and tables and integrates everything with your text into the specified output format.
- Quarto is **very powerful**. The rest of this course gives a basic introduction.

Our first Quarto document

In RStudio, we can create a new Quarto file by clicking on `File > New File > Quarto Document`. Give the document a fitting title and author and select `PDF` as the output format (for now). RStudio then creates a template `.qmd` file:



~~~
---
title: "My first Quarto document"
author: "Julian Amon"
format: pdf
---

## Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see <https://quarto.org>.

## Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```{r}
1 + 1
```

You can add options to executable code like this

```{r}
#| echo: false
2 * 2
```

The `echo: false` option disables the printing of code (only output is displayed).

~~~

```
R 4.5.0 · ~/Dokumente/Projekte/Lehre/CFPU Data Science und Data Analytics/2025_SS/
>
```

Environment is empty

Elements of a Quarto document – YAML header

- At the very top of the document – demarcated by three dashes (`---`) on either end – is the so-called **YAML header**:

```yaml
title: "My first Quarto document"
author: "Julian Amon"
format: pdf
```

- YAML is a widely used language mainly used for providing configuration data.
- With Quarto, it is used to define **metadata** about the document, such as (in this case) `title`, `author` and (output) `format`.
- We can define many other things in the header than what is included in the template, e.g. `theme`, `fontcolor` or `fig-width`. A complete list of available YAML fields for PDF documents can be found here.

Elements of a Quarto document – Code chunks

- To run code inside a Quarto document, you need to put it inside a so-called **code chunk**. Such a code consists of four elements:
  - **Three back ticks** and the programming language in curly braces to start the chunk (since we only use R, for us, this is always `` ```{r} ``).
  - **Chunk options** (optional): these are key-value pairs that can be used for customizing code chunks, each on their own line starting with `#|`.
  - The actual code.
  - **Three back ticks** again to signify the end of the code chunk.
- The template contains an example of a code chunks with options:

~~~
```{r}
#| echo: false

2 * 2
```
~~~

Elements of a Quarto document – Code chunks

- The full list of code options is available here. The most important ones are:
  - `echo: false` prevents code, but not the results from appearing in the finished file. Use this when writing reports aimed at people who do not want to see the underlying R code.
  - `eval: false` prevents code from being run. This is useful for displaying example code, or for disabling a large block of code without commenting each line.
  - `include: false` runs the code, but doesn't show the code or results in the final document. Use this for setup code (like loading libraries) that you do not want cluttering your report.
  - `message: false` or `warning: false` prevents messages or warnings from appearing in the finished file.
- By default, all of these code chunk options are `true`.

Elements of a Quarto document – Code chunks

- As you work more with Quarto, you will discover that some of the default chunk options do not fit your needs and you want to change them.
- For example, consider a situation where you are writing a report for a non-technical audience. To hide the underlying R code, we would have to set `echo: false` in each and every chunk.
- To avoid this, we can set the preferred options **globally** in the YAML header under `execute`. So, to achieve global code suppression, we would set:

```yaml
title: "My first Quarto document"
author: "Julian Amon"
execute:
  echo: false
format: pdf
```

- These **global** code chunk options can be overridden locally in any chunk.

Elements of a Quarto document – Markdown text

- Finally, besides the YAML header and code chunks, the last (main) type of content we see in a Quarto file is **text** mixed with simple formatting instructions. This text is written in a so-called **Markdown format**.
- Instead of using buttons like in Word or Pages to format our text, **Markdown** lets us add **headings, bold or italic text, lists, links and more**, just by typing special characters. Consider the following examples:
  - `# Heading`, `## Subheading` and `### Subsubheading` for section titles.
  - `*italic*` for text *in italics*.
  - `**bold**` for text **in boldface**.
  - `- Item 1`

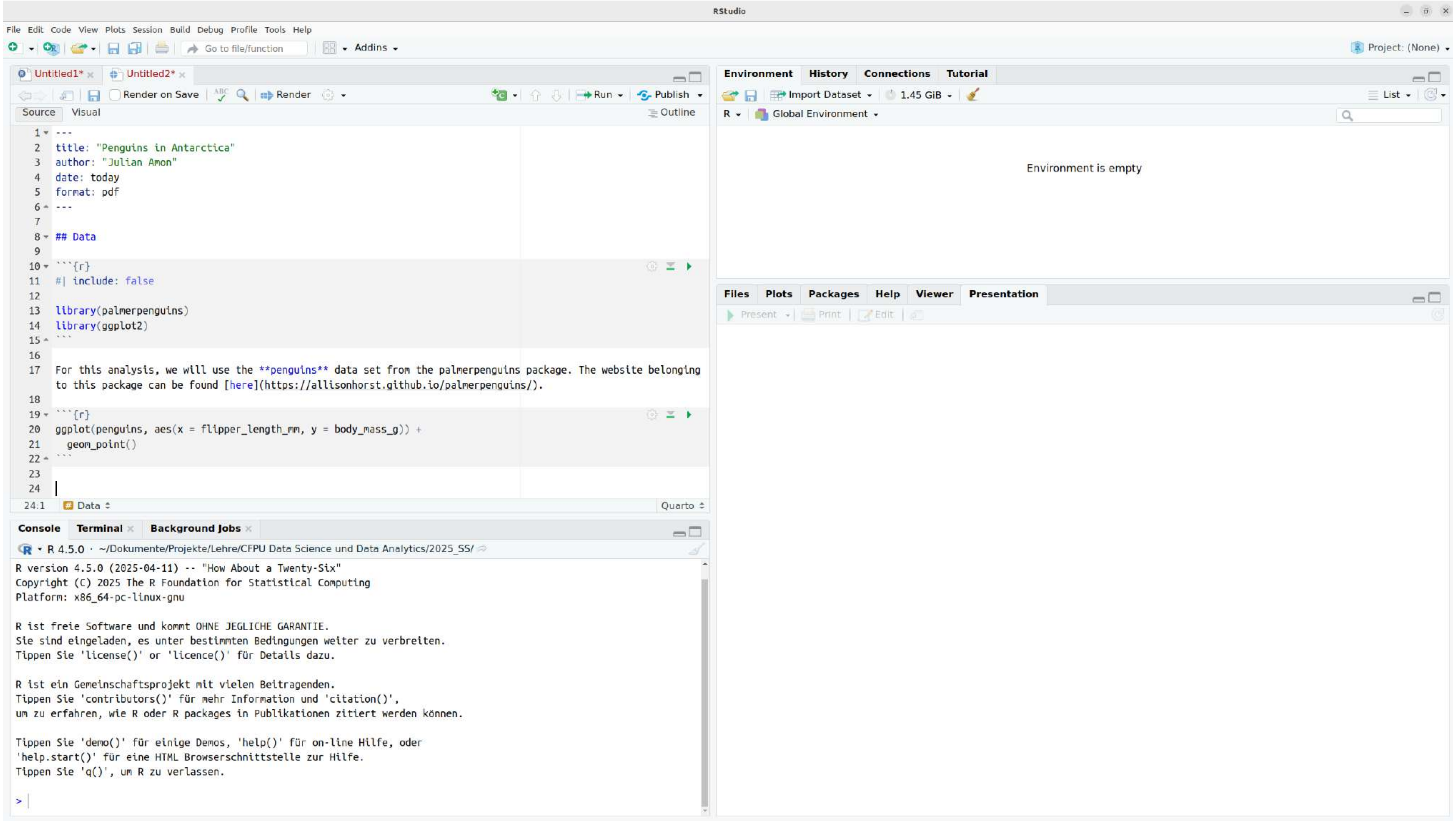
    `- Item 2` for bullet point lists.
  - `[Link text](https://quarto.org)` for clickable links.

Combining all three elements in Quarto

- Let’s use our knowledge about the three elements of a Quarto document (YAML header, code chunks and Markdown text) to adapt the template that RStudio created for us in the following ways:
  - Add a `date` field to the header and set it (literally) to `today`.
  - Change the title of the document to `Penguins in Antarctica`.
  - Change the level 2 header `## Quarto` to `## Data`.
  - Immediately after this header, add a code chunk that only loads the libraries `palmerpenguins` and `ggplot2`. For this chunk, set `include: false`.
  - After that chunk, write a simple sentence in Markdown that points out that we use data from the `palmerpenguins` package. Make sure to include a link to the website of the package, which can be found here.
  - Finally, in another chunk, create a scatter plot of `flipper_length_mm` length against `body_mass_g`. Remove all other content from the template.

Combining all three elements in Quarto

After you did all that, your Quarto document should look something like this:



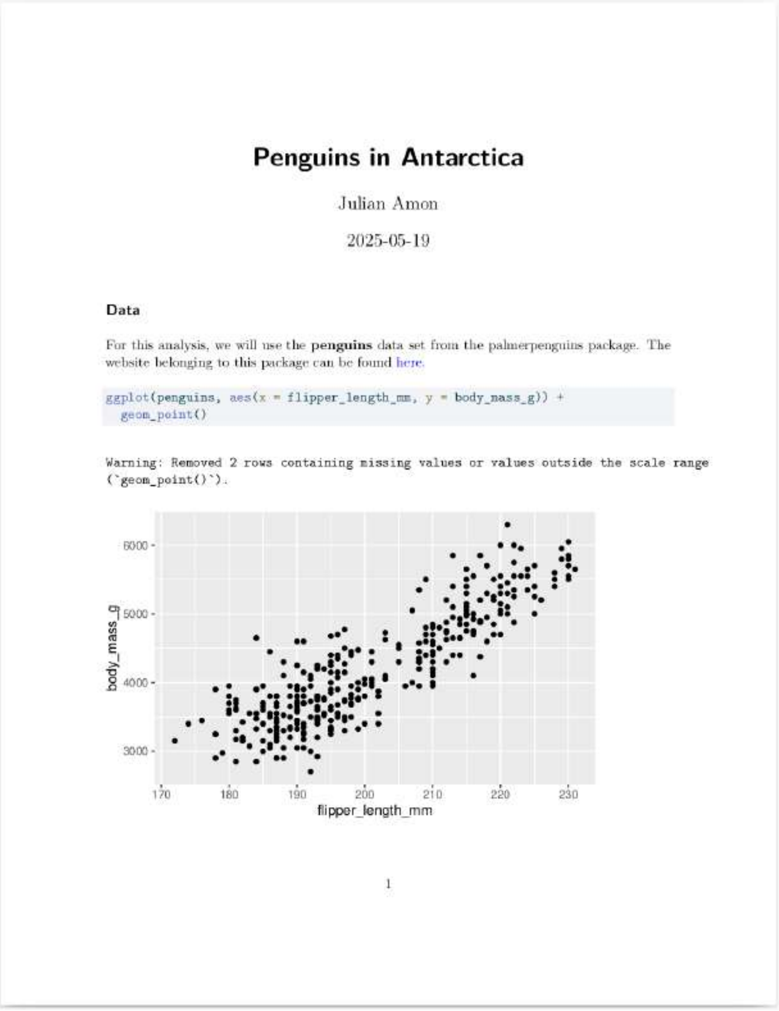
~~~
---
title: "Penguins in Antarctica"
author: "Julian Amon"
date: today
format: pdf
---

## Data

```{r}
#| include: false

library(palmerpenguins)
library(ggplot2)
```

For this analysis, we will use the **penguins** data set from the palmerpenguins package. The website belonging
to this package can be found [here](https://allisonhorst.github.io/palmerpenguins/).

```{r}
ggplot(penguins, aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point()
```

~~~

```
R version 4.5.0 (2025-04-11) -- "How About a Twenty-Six"
Copyright (C) 2025 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu

R ist freie Software und kommt OHNE JEGLICHE GARANTIE.
Sie sind eingeladen, es unter bestimmten Bedingungen weiter zu verbreiten.
Tippen Sie 'license()' or 'licence()' für Details dazu.

R ist ein Gemeinschaftsprojekt mit vielen Beitragenden.
Tippen Sie 'contributors()' für mehr Information und 'citation()',
um zu erfahren, wie R oder R packages in Publikationen zitiert werden können.

Tippen Sie 'demo()' für einige Demos, 'help()' für on-line Hilfe, oder
'help.start()' für eine HTML Browserschnittstelle zur Hilfe.
Tippen Sie 'q()', um R zu verlassen.

>
```

Compiling our first Quarto report

- To **compile** (or **render**) your final report from this Quarto file, first **save the file** into an appropriate location on your computer.

- Then, click on the Render button at the top of the screen. This will run all the code, and **combine text, code and results** in the desired output format:



Penguins in Antarctica

Julian Amon

2025-05-19

Data

For this analysis, we will use the **penguins** data set from the palmerpenguins package. The website belonging to this package can be found here.

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

```
Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_point()`).
```

1

The visual editor

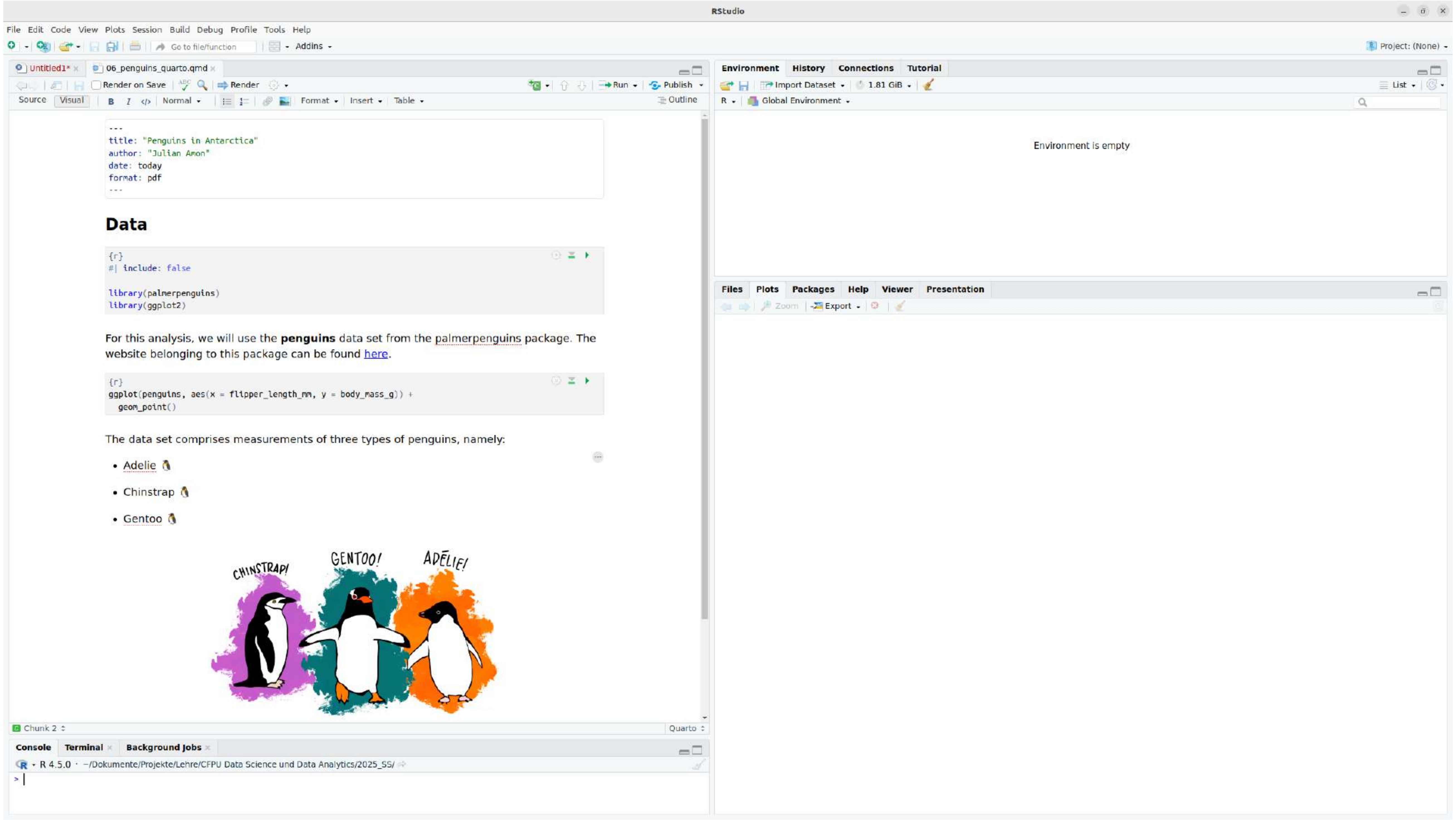
- We have successfully created our first **Quarto report**! 🎉
- While writing text in Markdown format is in principle easy to read and write, it still requires learning new syntax. For this reason, RStudio offers the **visual editor** to assist you in editing your Quarto document:
  - The idea of the visual editor is to provide a **graphical user interface (GUI)** that is more similar to working with, say, Microsoft Word.
  - For example, you can use the buttons on the menu bar for **text formatting** or to insert lists, images, tables, cross-references, etc. You can also **paste an image** from your clipboard directly into the visual editor and much more…
  - While the visual editor displays your content with formatting, under the hood, it saves your content in plain Markdown. You can switch back and forth between **visual and source editor** to view and edit your content.

The visual editor

- Let's use the visual editor to make the following **changes** to our Quarto document on the penguins from the `palmerpenguins` package:
  - Write a sentence that explains that the data comprises three penguin species and list them with **bullet points** (Adelie, Chinstrap and Gentoo).
  - At the end of each bullet point, insert a penguin emoji (`Insert > Special Characters > Insert Emoji...`).
  - Below these bullet points, insert the cartoon displaying all three species from the package website. Make sure that it is aligned in the centre.
- At the end, switch back to the source editor and see how your changes were implemented in Markdown. Finally, render your report again to see how the changes affect the final output.

Implementing changes in the visual editor

With all these changes, the visual editor should show something like this:



```
---
title: "Penguins in Antarctica"
author: "Julian Amon"
date: today
format: pdf
---
```

Data

```
{r}
#| include: false

library(palmerpenguins)
library(ggplot2)
```

For this analysis, we will use the **penguins** data set from the palmerpenguins package. The website belonging to this package can be found here.

```
{r}
ggplot(penguins, aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point()
```

The data set comprises measurements of three types of penguins, namely:

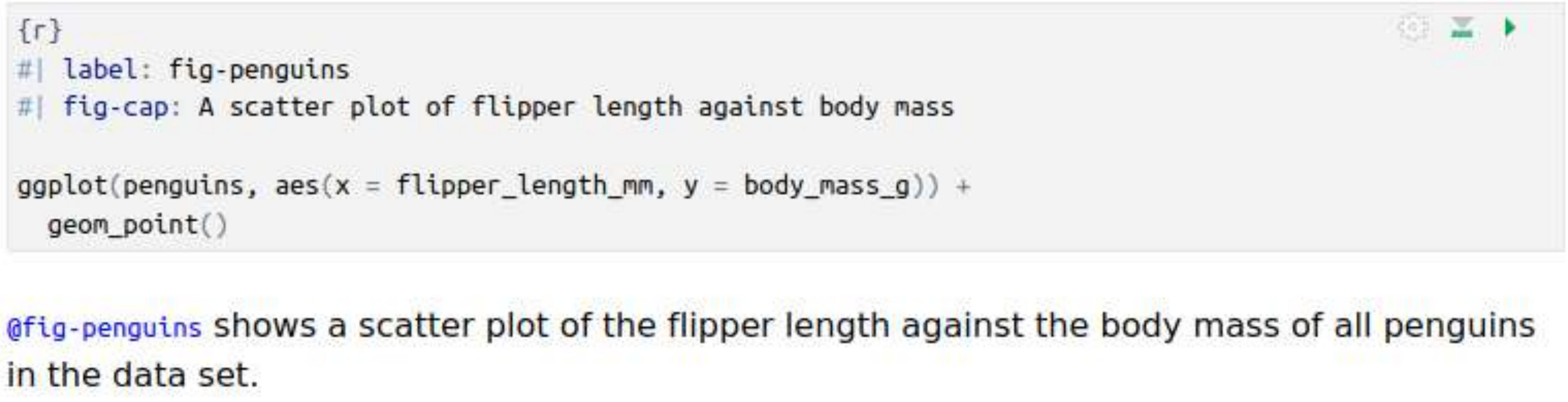
- Adelie 🐧
- Chinstrap 🐧
- Gentoo 🐧

```
R 4.5.0 · ~/Dokumente/Projekte/Lehre/CFPU Data Science und Data Analytics/2025_SS/
>
```

Cross-references and citations in Quarto

- To write a scientific report or paper, two features are almost always required:
  - **Cross references**: referring to a figure or a table by a `label`, so that you can reference it in your text using its number (e.g., “see Figure 1”).
  - **Citations**: citing academic literature using a `.bib`-file, so that references are properly formatted and listed automatically at the end of the document.
- Fortunately, the visual editor makes both of these things very easy. To create a cross reference for a plot produced by a code chunk, we need to do two things:
  - Give the chunk a `label` starting with `fig-`, e.g. `fig-penguins1`.
  - Provide a caption for the plot in the `fig-cap` code chunk option, e.g. “*A scatter plot of flipper length against body mass*”.
- Once, we have done this, we can then refer to this figure in the text with `@fig-penguins1` or insert the cross reference via `Insert > Cross Reference`.

Cross-references in Quarto

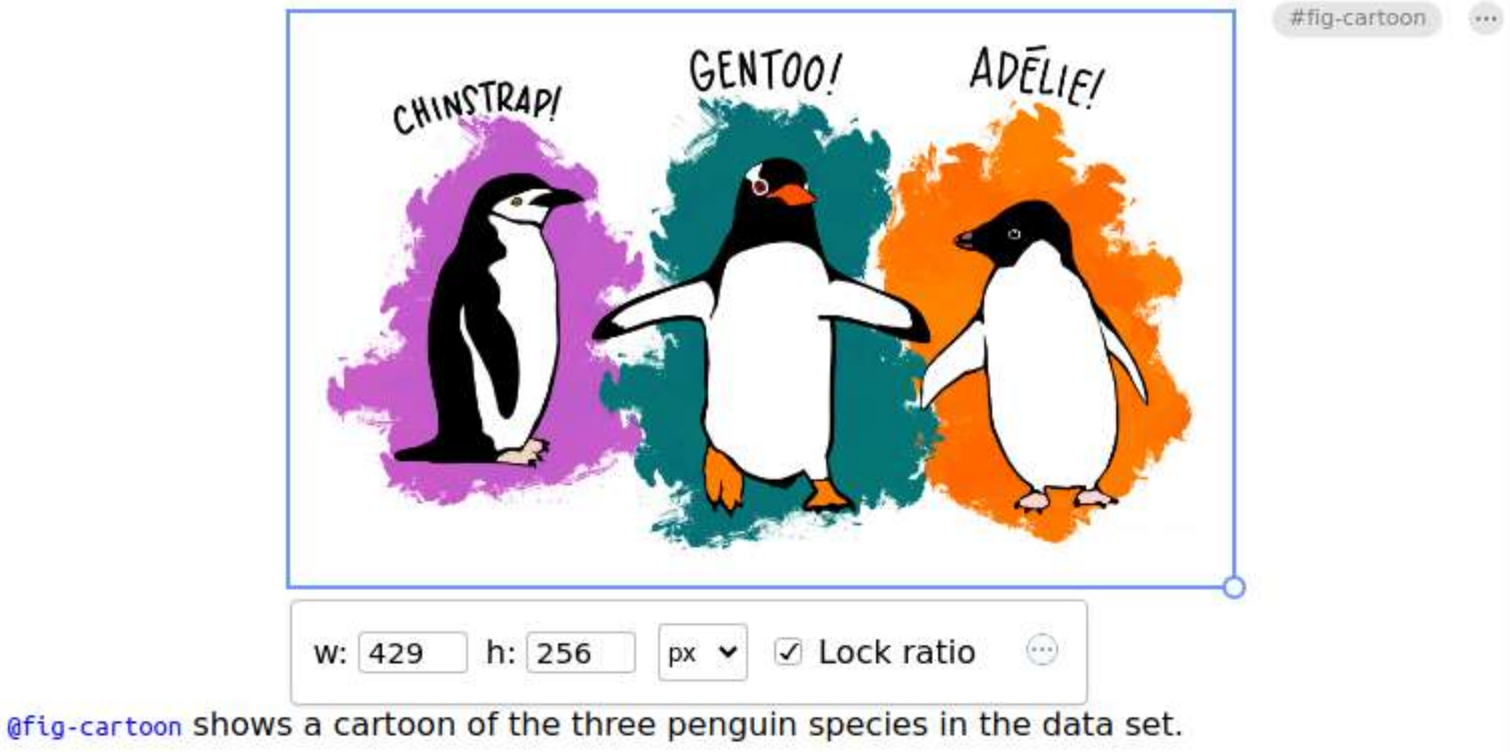


```
{r}
#| label: fig-penguins
#| fig-cap: A scatter plot of flipper length against body mass

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

`@fig-penguins` shows a scatter plot of the flipper length against the body mass of all penguins in the data set.

- In this way, it is ensured that wherever we move the chunk or its reference in the report, the reference will always go the correct plot.
- We can create such references also for **inserted images**, of course:
  - Simply click on the three dots next to the inserted image and go to the `Attributes` tab.
  - Give the image a label starting with `#fig-`, e.g. `#fig-cartoon`
  - You can then refer to the image by `@fig-cartoon` (see next slide).

Cross-references in Quarto

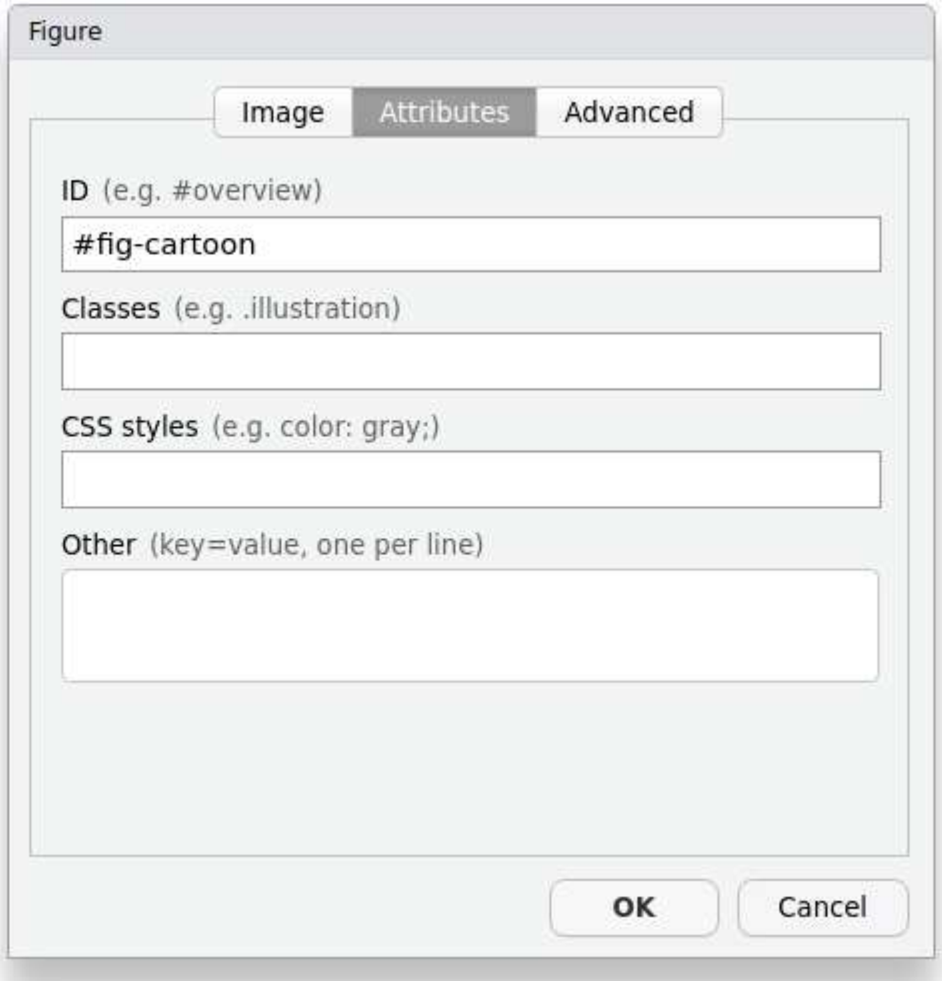


#fig-cartoon

w: 429 h: 256 px

☑ Lock ratio

@fig-cartoon shows a cartoon of the three penguin species in the data set.



ID (e.g. #overview): #fig-cartoon

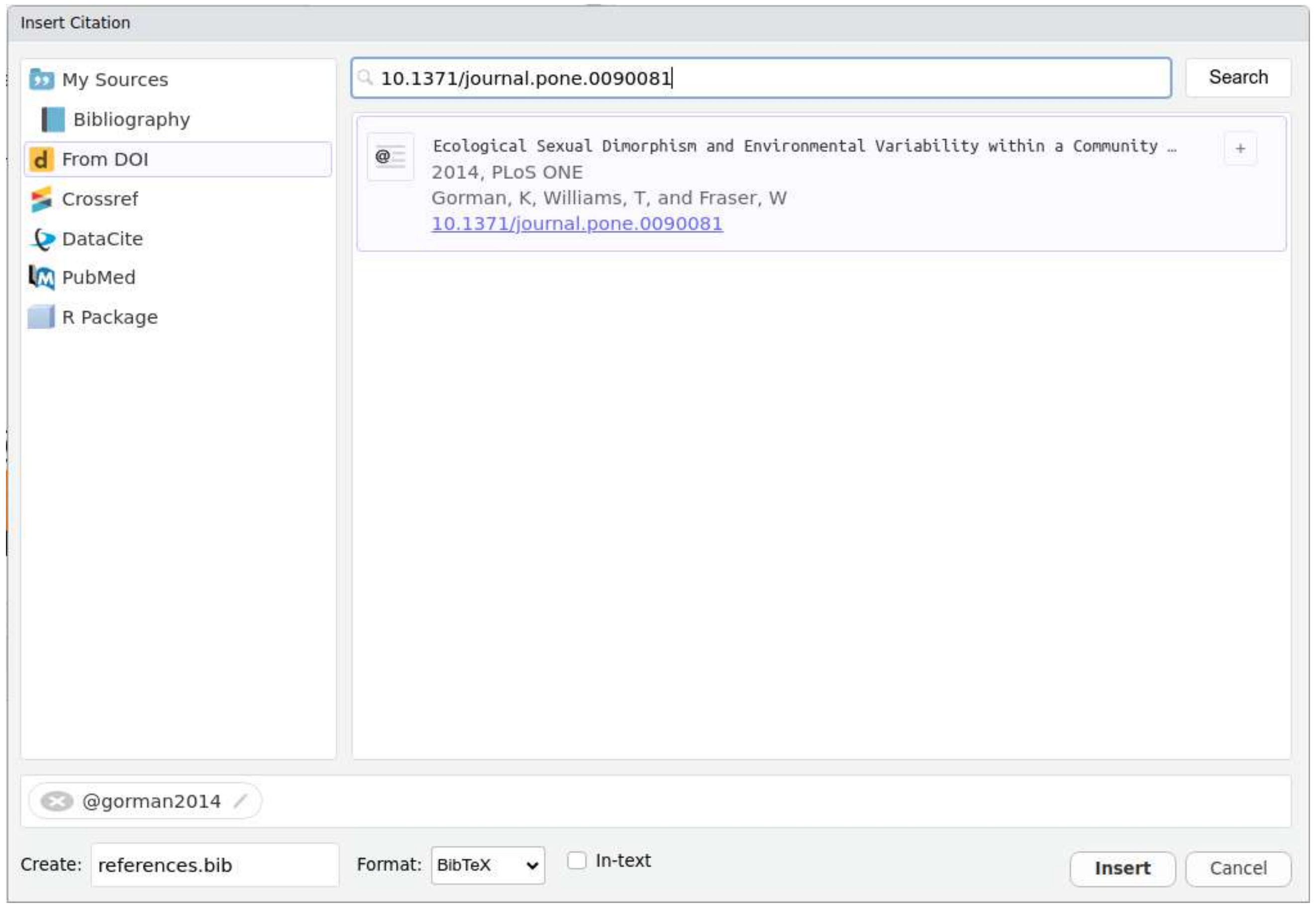
Classes (e.g. .illustration):

CSS styles (e.g. color: gray;):

Other (key=value, one per line):

Citations in Quarto

- Let's say we wanted to cite the paper in which the penguin data was originally published. It can be found here.
- To insert a citation in the visual editor, simply select `Insert > Citation` at the place where you want the citation to be. RStudio then gives you several options. The easiest way is to paste the **DOI** of the paper:



10.1371/journal.pone.0090081

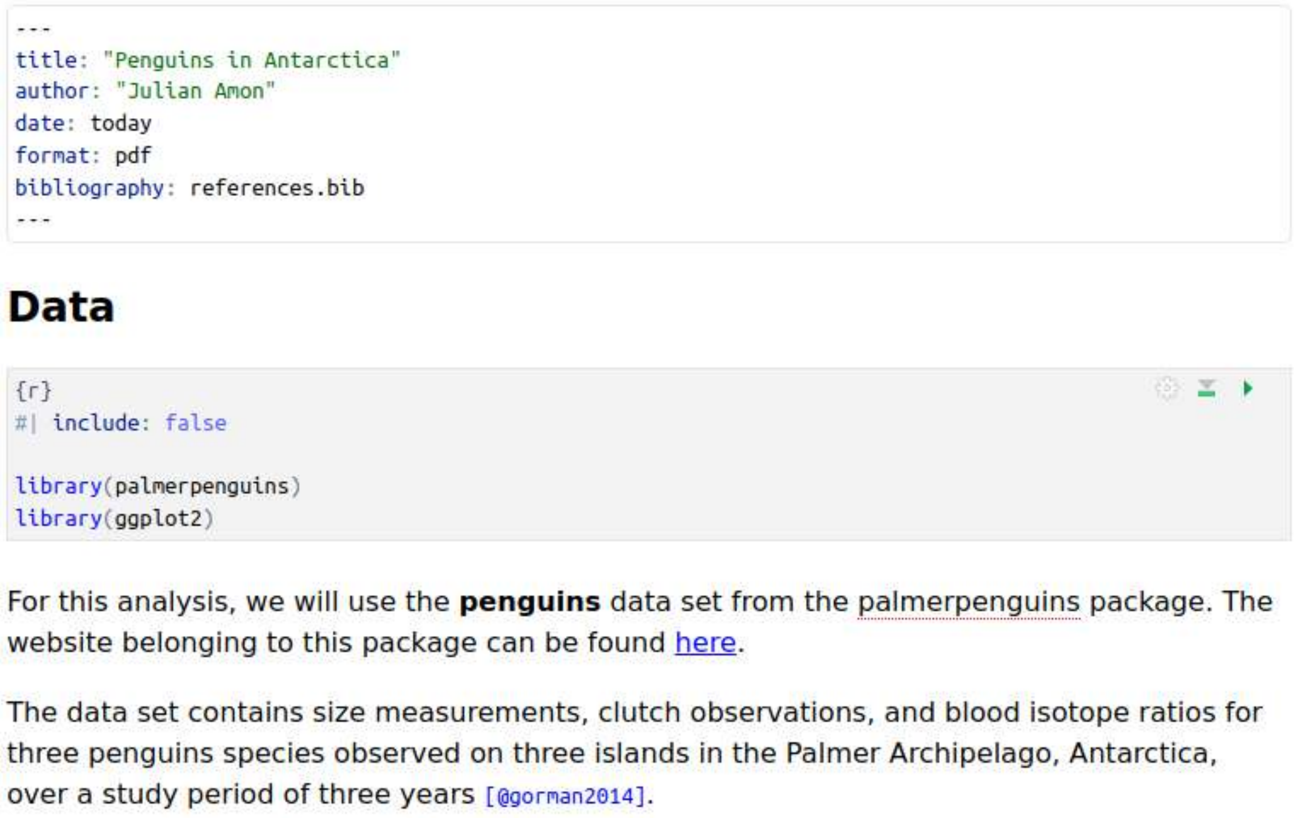
Ecological Sexual Dimorphism and Environmental Variability within a Community …
2014, PLoS ONE
Gorman, K, Williams, T, and Fraser, W
10.1371/journal.pone.0090081

@gorman2014

Create: references.bib
Format: BibTeX
☐ In-text

Citations in Quarto

At the next rendering, Quarto will create a `.bib` file holding the added references, **add the citation** at the place where you have inserted it and **append a list of references** to the end of your document.



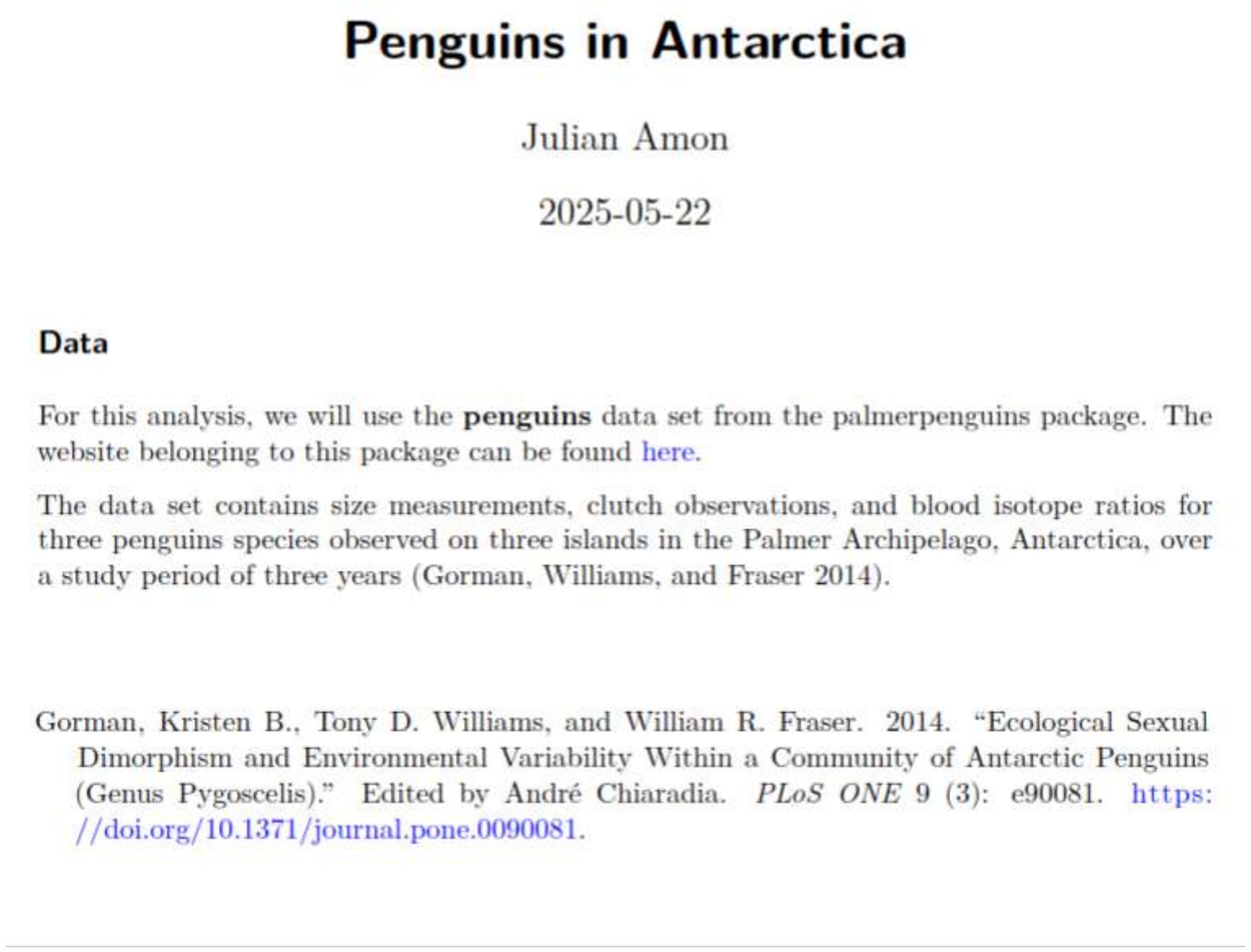
```
---
title: "Penguins in Antarctica"
author: "Julian Amon"
date: today
format: pdf
bibliography: references.bib
---
```

Data

```
{r}
#| include: false

library(palmerpenguins)
library(ggplot2)
```

For this analysis, we will use the **penguins** data set from the palmerpenguins package. The website belonging to this package can be found here.

The data set contains size measurements, clutch observations, and blood isotope ratios for three penguins species observed on three islands in the Palmer Archipelago, Antarctica, over a study period of three years `[@gorman2014]`.



Penguins in Antarctica

Julian Amon

2025-05-22

Data

For this analysis, we will use the **penguins** data set from the palmerpenguins package. The website belonging to this package can be found here.

The data set contains size measurements, clutch observations, and blood isotope ratios for three penguins species observed on three islands in the Palmer Archipelago, Antarctica, over a study period of three years (Gorman, Williams, and Fraser 2014).

Gorman, Kristen B., Tony D. Williams, and William R. Fraser. 2014. "Ecological Sexual Dimorphism and Environmental Variability Within a Community of Antarctic Penguins (Genus Pygoscelis)." Edited by André Chiaradia. *PLoS ONE* 9 (3): e90081. https://doi.org/10.1371/journal.pone.0090081.

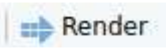
Quarto formats

- One of the many great things about Quarto is that it offers many different **output formats** in which to render our report.
- Switching between output formats is very easy: simply change the `format` in the YAML header.
- For **reports**, the most common formats are:
  - `pdf`: what we were using for our penguin report so far.
  - `html`: an HTML document with more flexibility than a PDF (e.g. animations).
  - `docx`: a Microsoft Word document.
- For **presentations**, the most common formats are:
  - `revealjs`: an HTML-based presentation format allowing for interactivity, animations, dynamic content and much more.
  - `pptx`: a Microsoft PowerPoint presentation.

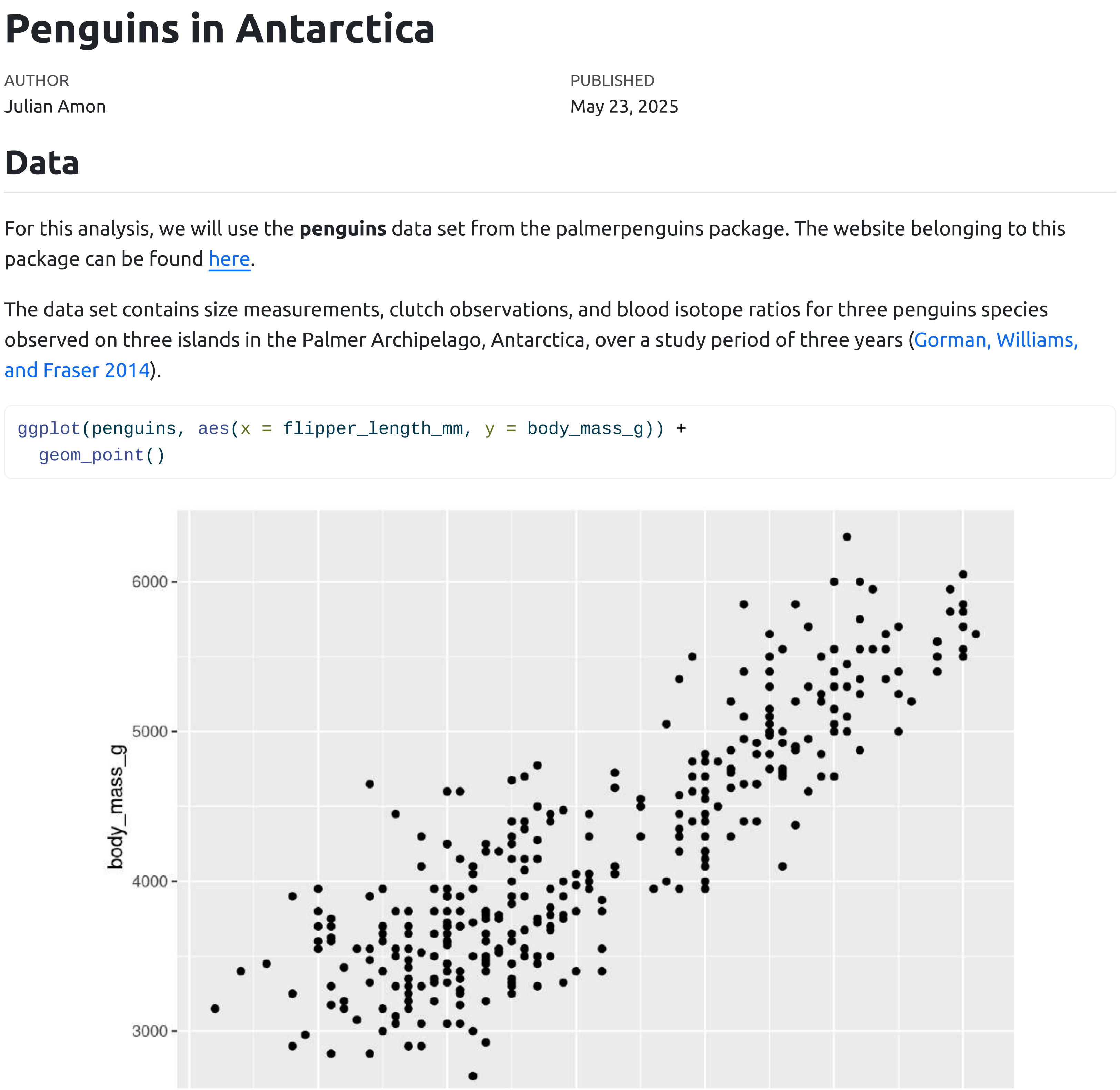
Quarto formats – Switching to HTML

- As you may have noticed, the penguin emojis we included into our Quarto document were not rendered into the output PDF. This is because the PDF output format cannot handle emojis.
- To remedy that, let's switch to `html` in the YAML header:

```yaml
title: "Penguins in Antarctica"
author: "Julian Amon"
date: today
format: html
bibliography: references.bib
```

- When we now click on , the report will be generated as an **HTML** and not as a **PDF**. Note how it changes in appearance slightly and now correctly displays the emojis we added earlier.

Quarto formats – Switching to HTML



Penguins in Antarctica

AUTHOR
Julian Amon

PUBLISHED
May 23, 2025

Data

For this analysis, we will use the **penguins** data set from the palmerpenguins package. The website belonging to this package can be found here.

The data set contains size measurements, clutch observations, and blood isotope ratios for three penguins species observed on three islands in the Palmer Archipelago, Antarctica, over a study period of three years (Gorman, Williams, and Fraser 2014).

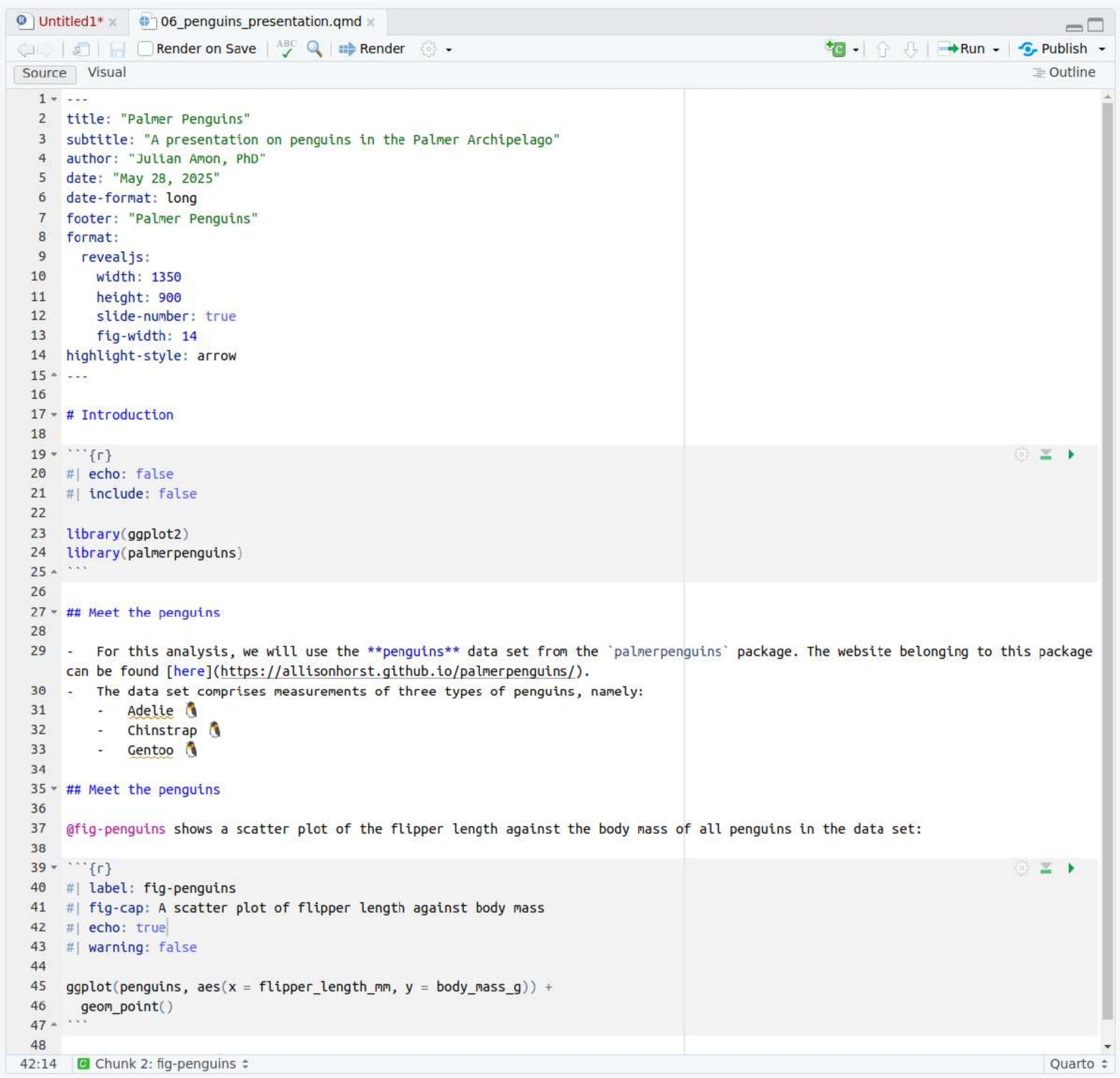
```
ggplot(penguins, aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point()
```

Quarto formats – Presentation with revealjs

- As indicated before, Quarto allows us not only to generate **reports**, but also **presentations**.
- While in principle, we could also simply switch the `format` to one of the presentation formats (e.g. `revealjs` or `pptx`) to generate a presentation, there are some special things to consider when doing that.
- Presentations work by dividing your content into **slides**. To do this with Quarto, the different level **headings** receive a special meaning when the output format is `revealjs` or `pptx`:
  - Second level headings (starting with `##`) demarcate slides, i.e. a new second level heading will create a new slide.
  - First level headings (starting with `#`) indicate the beginning of a new section with a **section title slide**.

Quarto formats – Presentation with revealjs

A really basic presentation for the penguin example could look like this:



~~~
---
title: "Palmer Penguins"
subtitle: "A presentation on penguins in the Palmer Archipelago"
author: "Julian Amon, PhD"
date: "May 28, 2025"
date-format: long
footer: "Palmer Penguins"
format:
  revealjs:
    width: 1350
    height: 900
    slide-number: true
    fig-width: 14
highlight-style: arrow
---

# Introduction

```{r}
#| echo: false
#| include: false

library(ggplot2)
library(palmerpenguins)
~~~

Meet the penguins

-   For this analysis, we will use the **penguins** data set from the `palmerpenguins` package. The website belonging to this package can be found [here](https://allisonhorst.github.io/palmerpenguins/).
-   The data set comprises measurements of three types of penguins, namely:
    -   Adelie 🐧
    -   Chinstrap 🐧
    -   Gentoo 🐧

Meet the penguins

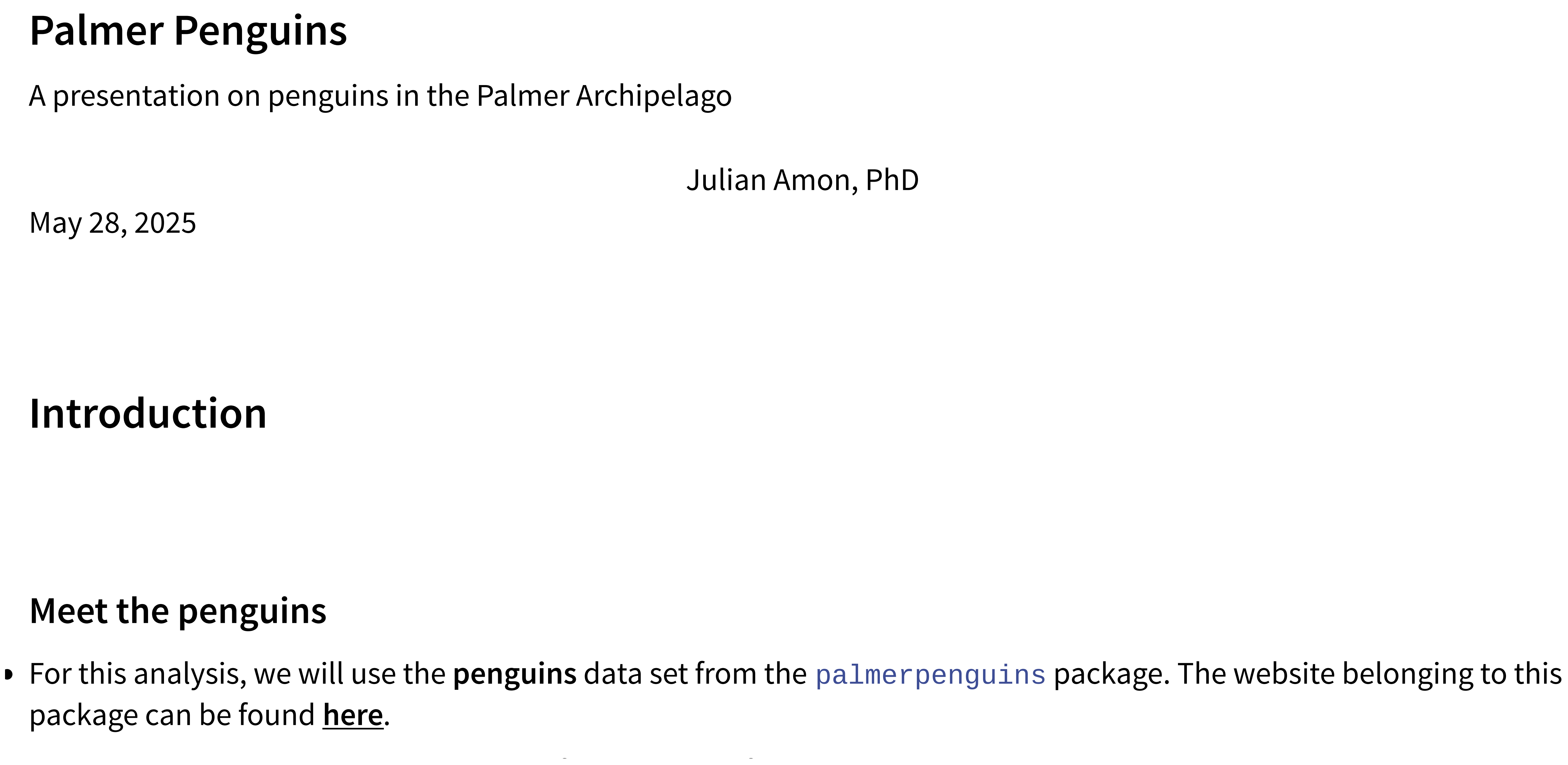
@fig-penguins shows a scatter plot of the flipper length against the body mass of all penguins in the data set:

```{r}
#| label: fig-penguins
#| fig-cap: A scatter plot of flipper length against body mass
#| echo: true
#| warning: false

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

```
```

Quarto formats – Presentation with revealjs

When rendered, this looks like this:



Palmer Penguins

A presentation on penguins in the Palmer Archipelago

Julian Amon, PhD

May 28, 2025

Introduction

Meet the penguins

- For this analysis, we will use the **penguins** data set from the `palmerpenguins` package. The website belonging to this package can be found **here**.

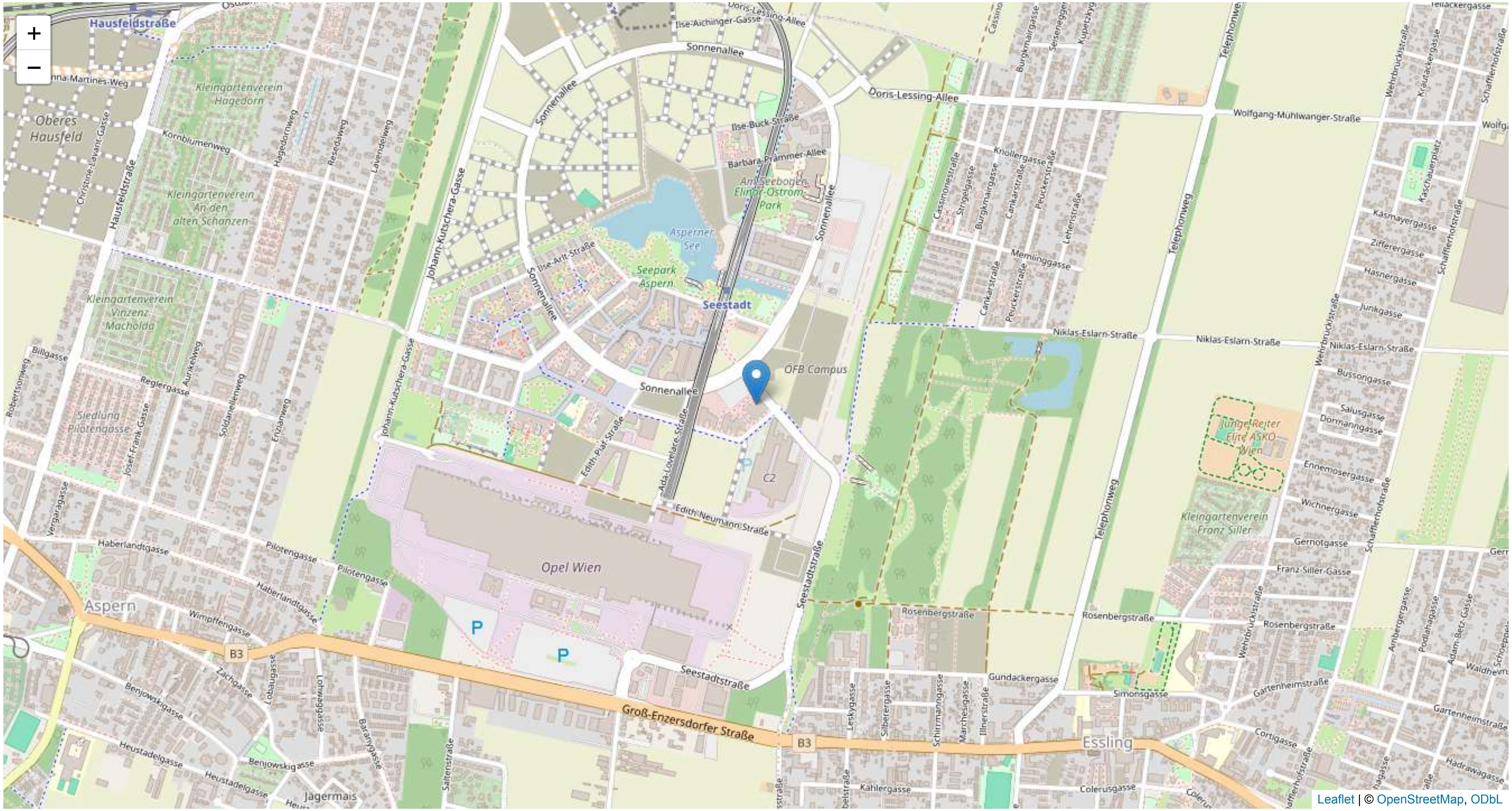
Presentations with Quarto

- Quarto has an incredible set of features for presentations. In fact, all of the slides for this course were created with Quarto!
- Some of these features include:
  - Code highlighting and animations:

```r
ggplot(penguins, aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point()
```

  - Slide transitions
  - Tabsets (as in tabs splitting plot and code, for example)
  - Interactive slides (see example on the next slide)
  - and countless more…
- Basically, whatever you want to do, **Quarto makes it possible**! Thus, we cannot even remotely cover everything, features are best explored by yourself!

Presentations with Quarto



Further references for Quarto

- The main resource is the Quarto website. There you can find:
  - Basic tutorials for getting started.
  - Specific guides on topics like presentations, computations, books, tools, authoring, websites, publishing, etc.
  - A reference that explains all options for the possible output formats.
  - A demo presentation made with `revealjs` that showcases some more of the cool features that Quarto offers for presentations.
- A Markdown tutorial for practice with elementary Markdown.
- A Markdown table generator that gives you the option to create Markdown tables in a very convenient way (including importing from Excel or similar).
- The R for Data Science book, which is freely available online. In its last two chapters, it also contains a great introduction to Quarto.