INTRODUCTION TO DATA SCIENCE WITH R

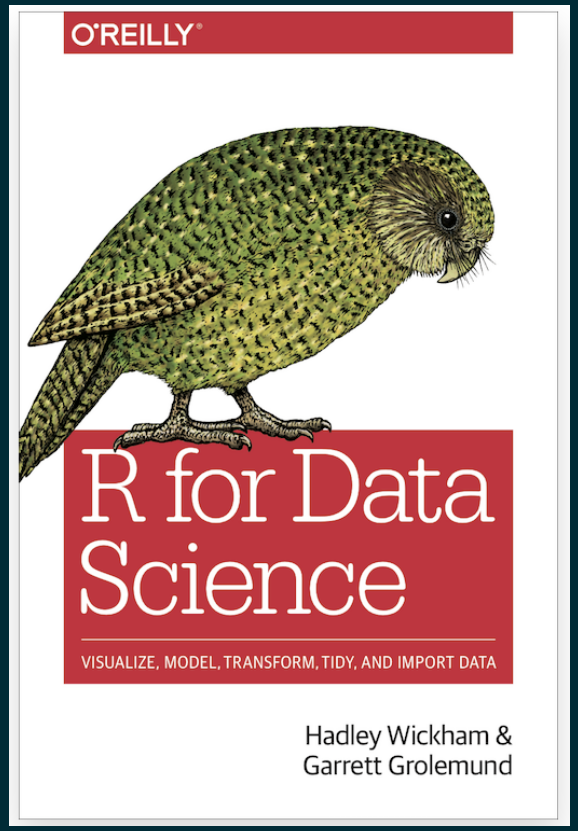
Session 1: Welcome!

Ina Bornkessel-Schlesewsky

January 18, 2023

WHAT IS DATA SCIENCE?

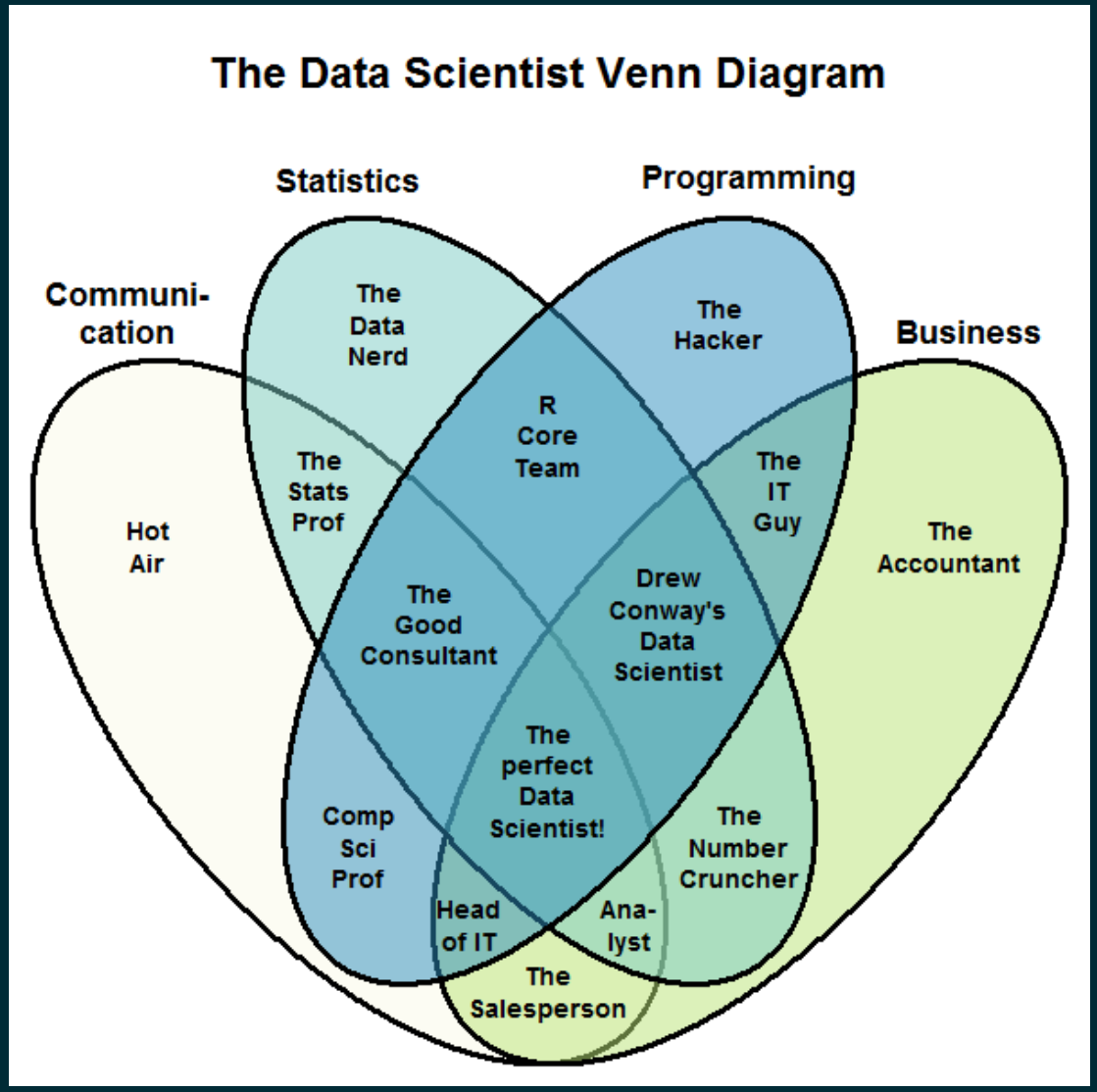
> Data science is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge.



Wickham, H. & Grolemund, G. (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media. Henceforth: *R4DS*

freely available online (Note: we will be using the in-progress 2nd edition)

WHAT IS DATA SCIENCE?



StackExchange Data Science user Stephan Kolassa CC BY-SA 4.0 via Wikimedia Commons

WHAT IS DATA SCIENCE?

> ‘Hal Varian, the chief economist at Google, is known to have said, “The sexy job in the next 10 years will be statisticians. People think I’m joking, but who would’ve guessed that computer engineers would’ve been the sexy job of the 1990s?” If “sexy” means having rare qualities that are much in demand, data scientists are already there. They are difficult and expensive to hire and, given the very competitive market for their services, difficult to retain. There simply aren’t a lot of people with their combination of scientific background and computational and analytical skills.’

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

From the October 2012 Issue

I'M NOT LOOKING FOR A SEXY NEW JOB. WHY IS THIS RELEVANT FOR ME?

In academic research, across a wide range of disciplines, we're also interested in turning "raw data into understanding, insight, and knowledge", as well as in communicating our results!

WHAT YOU WILL LEARN HERE

- how to gain insights from data using contemporary computational tools
- basic programming skills in an open source programming language (i.e. R)
- how to produce reproducible reports (good for science and good for you!)
- how to use online repositories such as GitHub or the Open Science Framework to share data and code

You will also develop an understanding of how these tools help to foster open science, reproducible research and thus the ethical treatment of data.

These skills are readily generalisable across a wide range of domains.

SO IS THIS JUST ANOTHER STATS COURSE?

Apart from the fact that we're using R?

"Oh no you didn't" gif by *happydog* from https://giphy.com

NOT JUST ANOTHER STATS COURSE ...

OUR FOCUS WILL BE ON

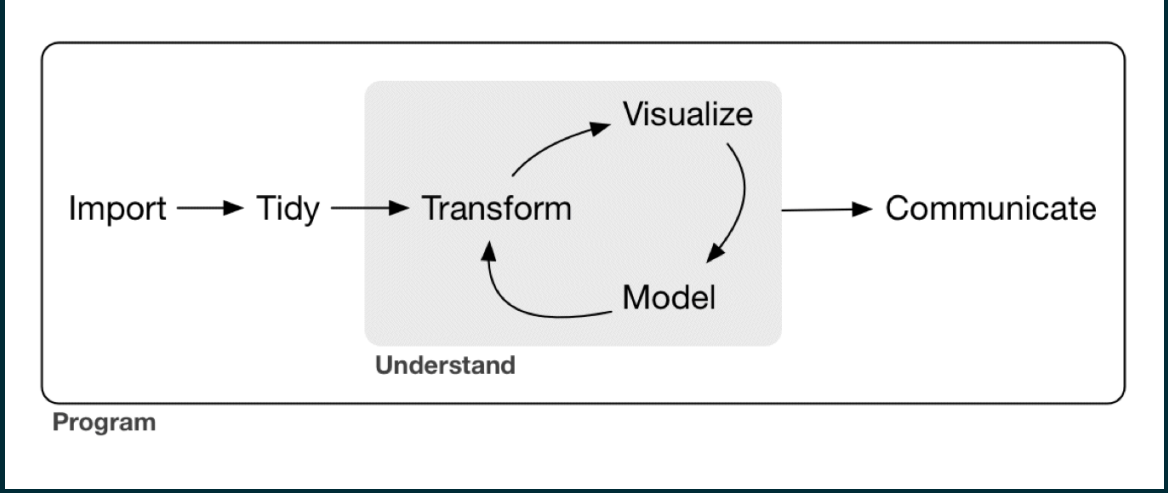
- understanding data rather than statistical tests per se (though they may come up in passing)
- philosophy / workflow rather than “results”
- (moral of the story: it’s not just about statistical significance!)

NOT JUST ANOTHER STATS COURSE ...

YOU WILL BE INTRODUCED TO A SET OF TOOLS AND WORKFLOW THAT

- foster good practices in dealing with data (i.e. we try to draw the best insights we can from a dataset)
- foster open science (i.e. we share our data and “show our work”, which is good for science and for sharing knowledge)
- are economical and reproducible (i.e. we avoid doing stuff by hand and can repeat what we did)

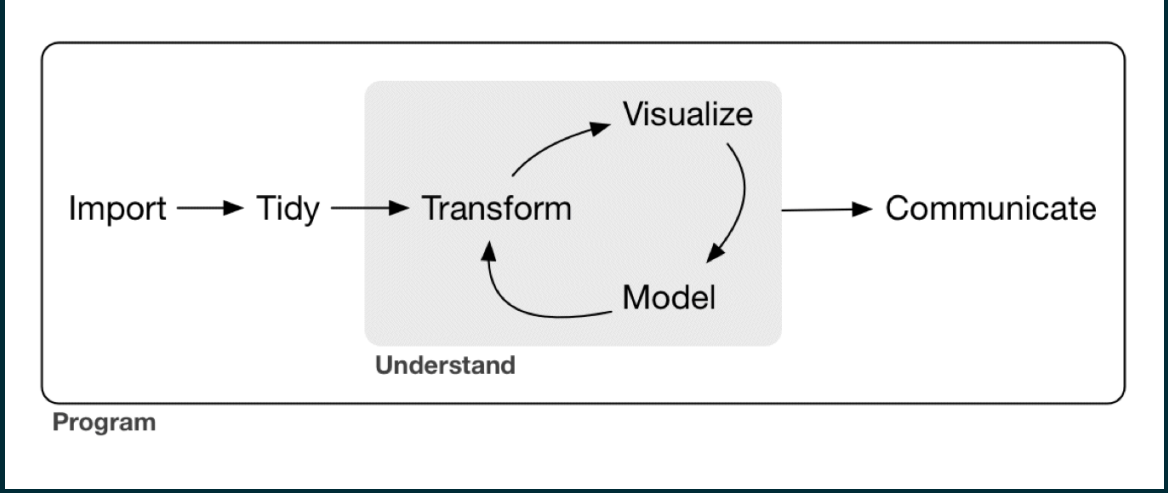
SPEAKING OF WORKFLOW



from R4DS

- **import** data (into R)
- **tidy** data: bring it into a consistent format that can be used for multiple purposes (each column = variable; each row = observation)
  - lets you focus on understanding the data rather than which format you need

SPEAKING OF WORKFLOW



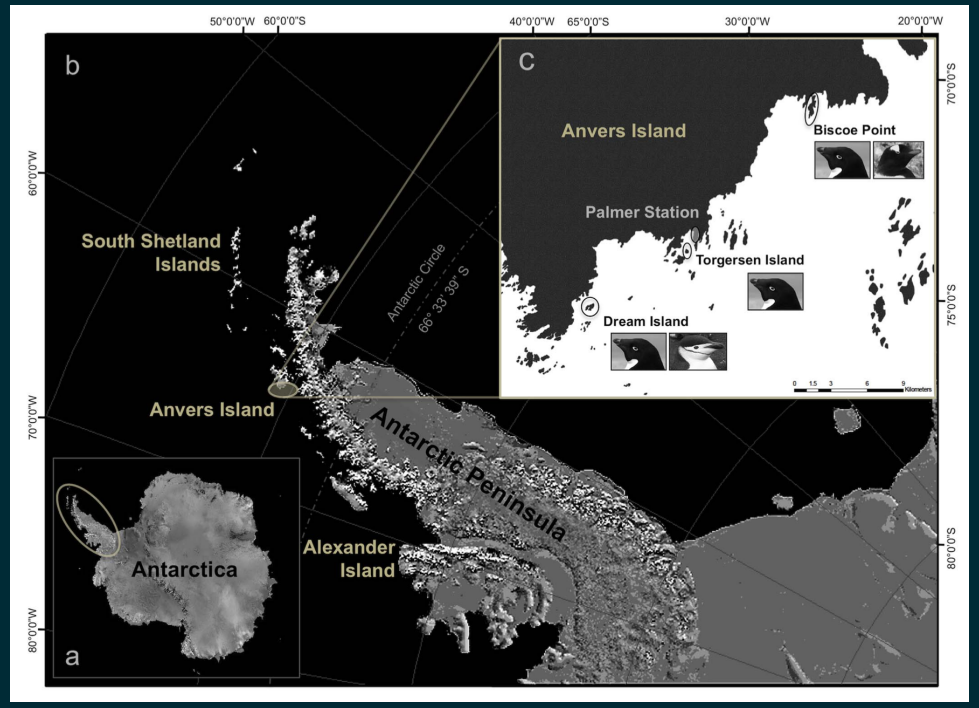
from R4DS

- **transform** data
  - e.g. focus on observations of interest, create new variables, compute summary statistics
- **visualise** data
  - essential for understanding
- **model** data
  - use (statistical) models to answer your questions about the data
- **communicate** insights

LET'S GIVE IT A GO!

PENGUINS!

Artwork by @allison_horst



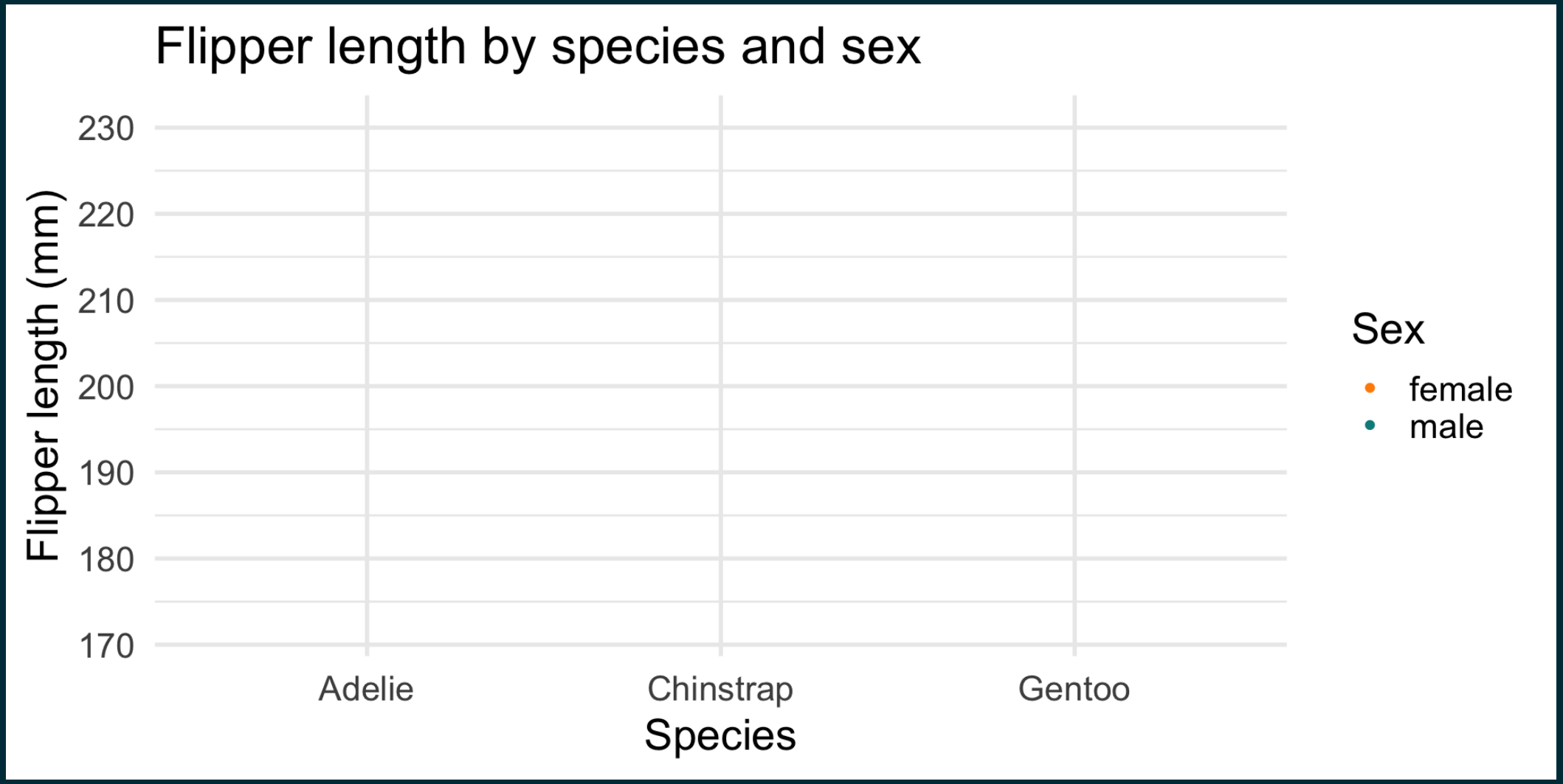
- data on penguins from the Palmer Archipelago in Antarctica
- original data (and image above) from Gorman et al. (2014, Plos One)

PENGUINS DATA

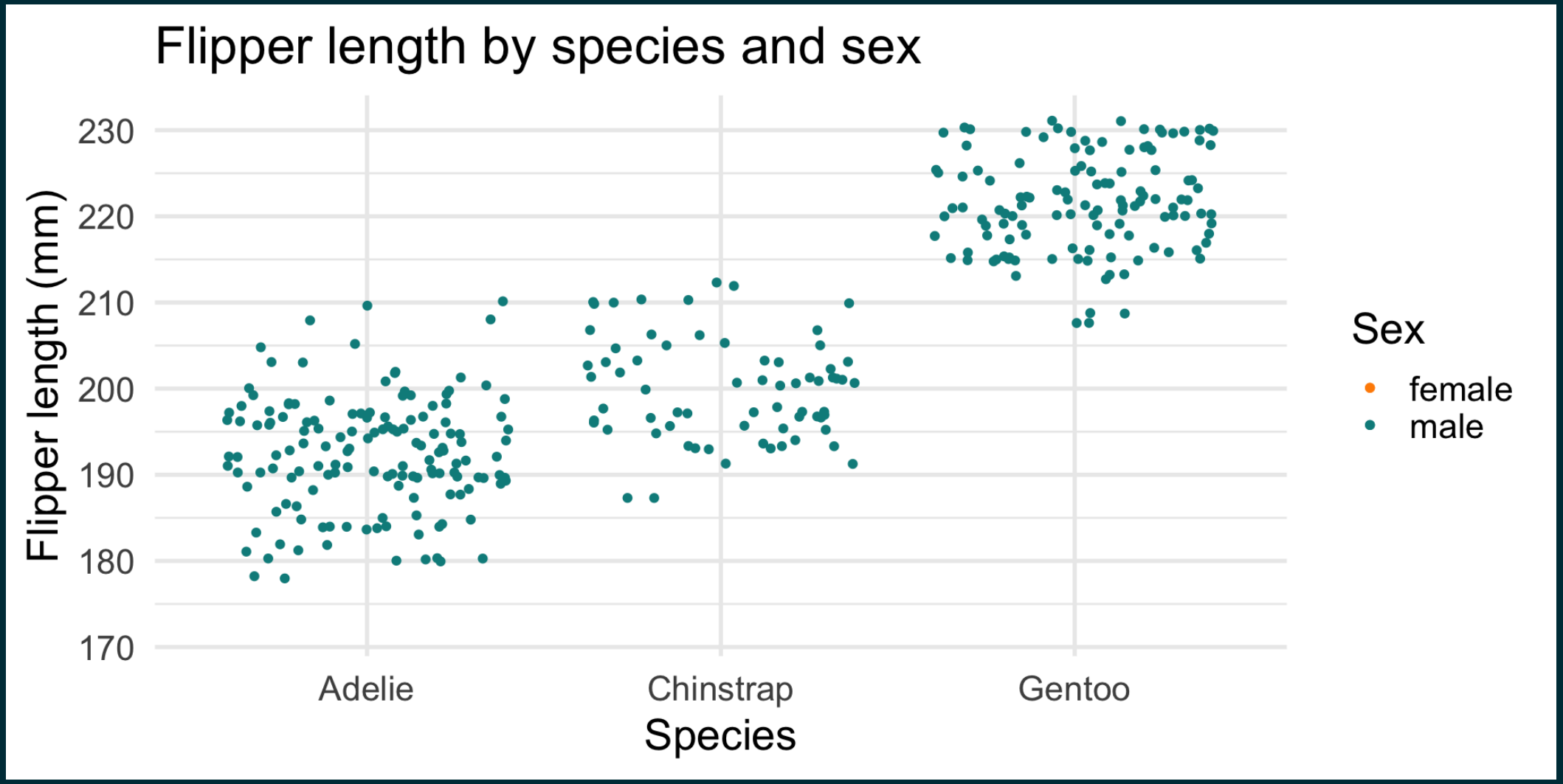
- from the `palmerpenguins` R package (Horst, 2020) - don't worry, you will learn more about what an R package is later

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007

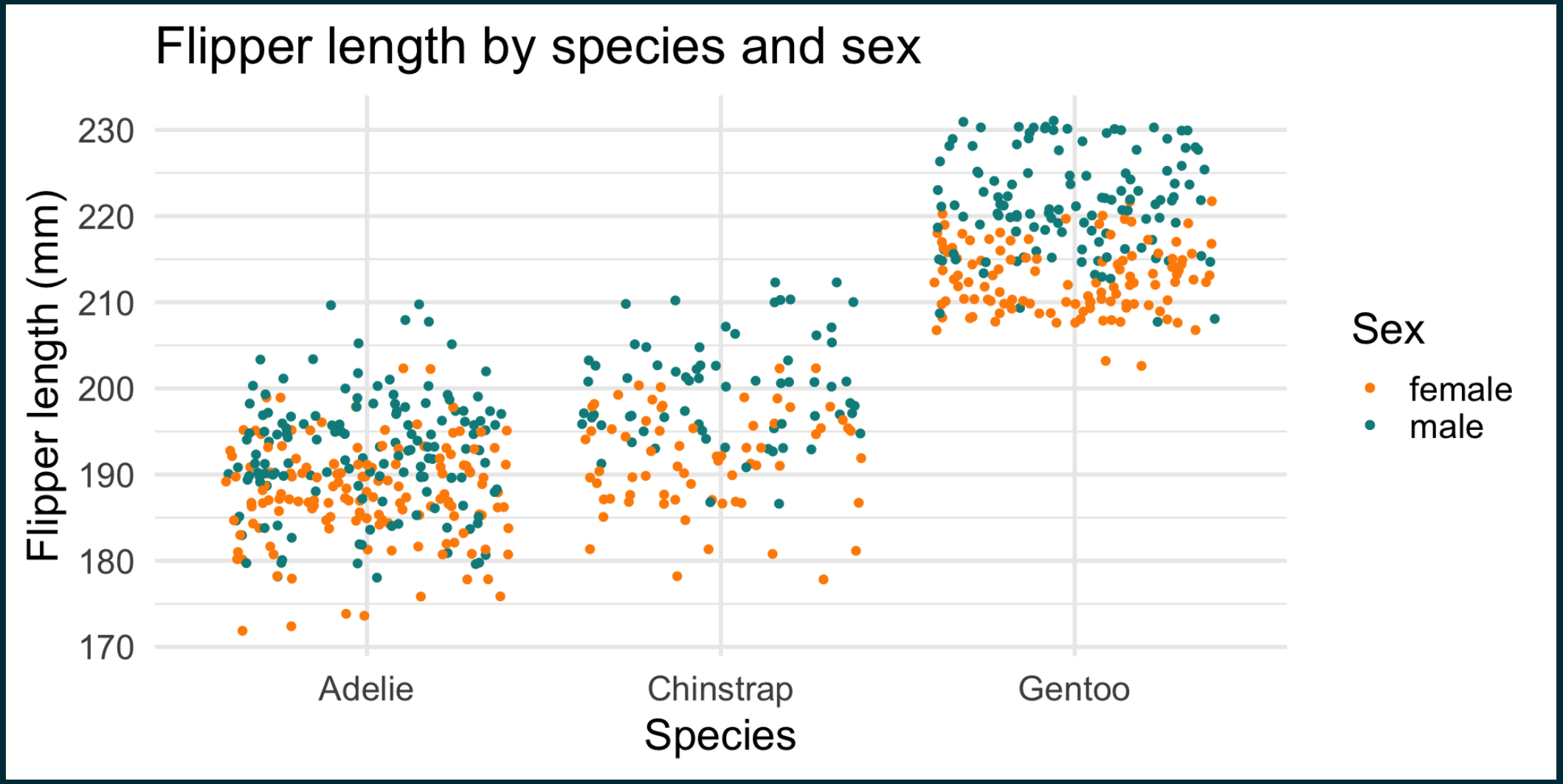
WHICH PENGUINS ARE LARGEST?



WHICH PENGUINS ARE LARGEST?



WHICH PENGUINS ARE LARGEST?



THIS MUST BE TRICKY, RIGHT?

- You will be able to generate (simple) publication-quality figures not unlike these by the end of our first session tomorrow

DATA EXPLORATION EXERCISE

EXPLORE THE PALMER PENGUINS DATA

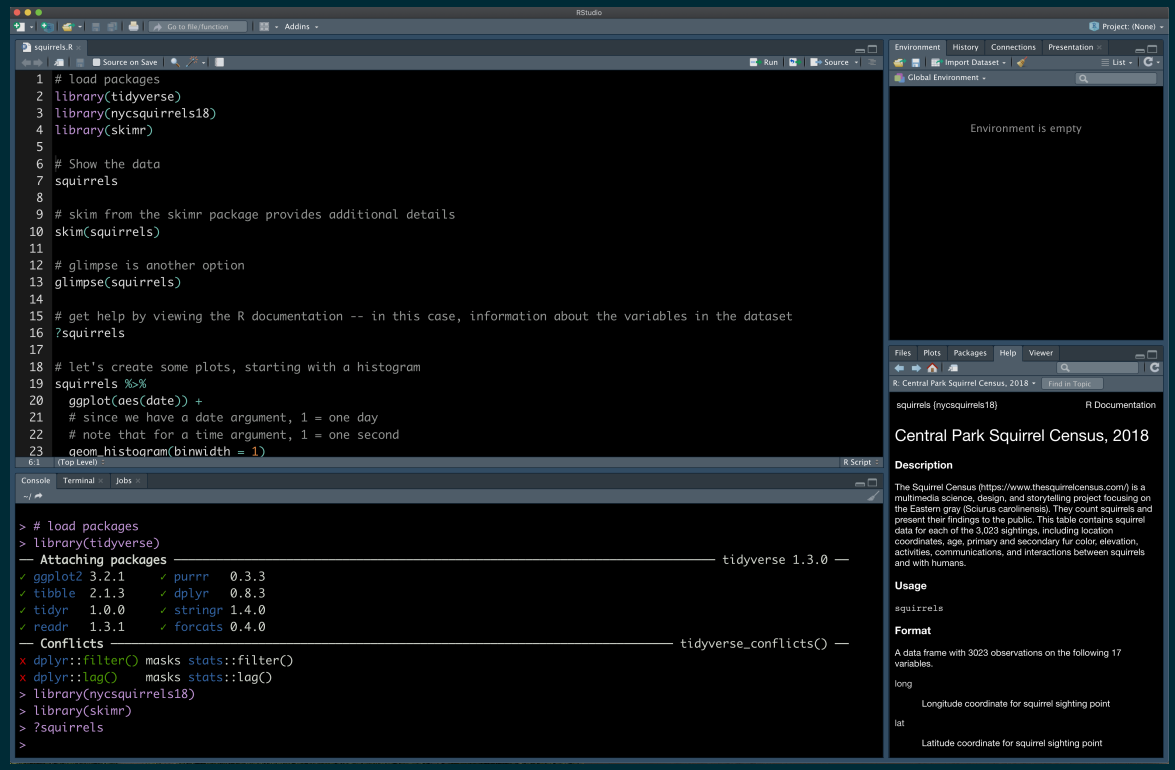
- Open this web app: https://ibsneuro.shinyapps.io/palmer_penguins/
- Tab 1 contains information about the data set and lets you inspect the data frame
- Tab 2 allows you to generate plots by selecting the type of graph, which variables to put on the x and y axes and which variable to group by (using different colours)
- In your exploration, consider the questions on the following slide
- For each question, note down not only your answer but also the strategy you chose to get to it: how did you choose to construct your graph for the question and why?

EXPLORE THE PALMER PENGUINS DATA

- If you wanted to predict a penguin's body mass, which other attributes could you look at (e.g. flipper length, bill length, sex etc.)? In other words, which of the other attributes appear to be most predictive of body mass?
- Is there a close relationship between bill length and bill depth?
- Is it possible to look at effects of island (i.e. the environment in which the penguins live) independently of other factors such as species or sex? If not, why not?
- Explore another 2 or 3 questions that interest you
- Finally, reflect on what this exercise has shown you regarding the use of different graph types to address different questions

SO HOW DOES THIS WORK?

ENTER R AND RSTUDIO (POSIT)



- R is a programming language for statistical computing (but it can also be used for other things)
- RStudio (Posit) is an integrated development environment (IDE) for R, which is a fancy way of saying that it provides a convenient platform within which we can use R

POSIT CLOUD

- For (the first part of) this workshop, we will be using Posit Cloud, which provides a web-based version of RStudio
- This means that you won't have to install anything on your computer and that you will have direct access to all of the materials that I have prepared
- Go to https://posit.cloud and create a login
- Once you have done this, use the link that you were emailed to access the **data_science_2023** workspace on Posit Cloud and select the project *summer_2023*
- When you access the project, you will receive your own permanent copy of it to work on

EXERCISE: PALMER PENGUINS

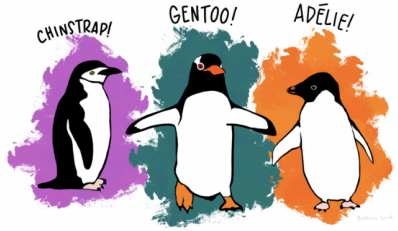
- go to the Files tab in the lower right pane of your RStudio Cloud project
- go to exercises > 01a_penguins
- click on the document *01a_penguins.qmd*
- this is a *Quarto* document which mixes text and code and is an excellent format for reproducible research reports, as we will see later - don't worry too much about the code for now; this is what the document looks like when rendered

Palmer Penguins

Introduction

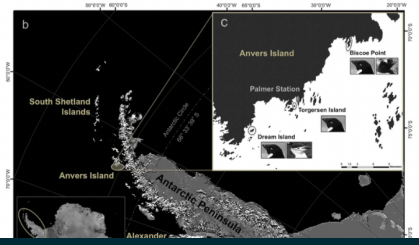
Here, we explore data on penguins from the Palmer Archipelago in Antarctica. The data are from a paper by Gorman et al. (2014, Plos One) entitled "Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis)".

The term sexual dimorphism refers to the sexes of a species differing in physical appearance (e.g. body size). In their study, Gorman and colleagues examined sexual dimorphism in three species of Pygoscelis ("brush tailed") penguins: Adélie, Chinstrap, and Gentoo.



Artwork by @allison_horst

Data were collected during field research on several islands of the Palmer Archipelago, west of the Antarctic Peninsula:



EXERCISE: PALMER PENGUINS

- click on the *Render* button at the top of the document to produce the rendered version
- have a read-through and look at the figures
- there is also an interactive table to remind you of what the data look like (they're the same as in the web app that you interacted with earlier)

YOUR TURN!

For each of the following challenges, go back to the raw document (i.e. the one that doesn't look pretty 😁), try to figure out how to make the relevant change and then render the document using *Knit* to see whether you were correct!

- For the relationship between bill length and depth, change *Gentoo* to *Adelie*; check the figure to see if it worked
- Change the outcome variable in the histogram from body mass to something else and observe what happens. Remember that you can go back to the table at the top of the rendered document to have a look at the available variables. Note that the figure title will only change if you also adapt the text in "title". (Hint: you may need to change the bin width! What would make sense for the outcome variable you have chosen?)
- Look at the effect of body mass by island rather than species. What do you see?
- Look at flipper length by sex rather than species

YOUR TURN!

- Look at the effect of body mass by island rather than species. What do you see?
- Look at flipper length by sex rather than species