

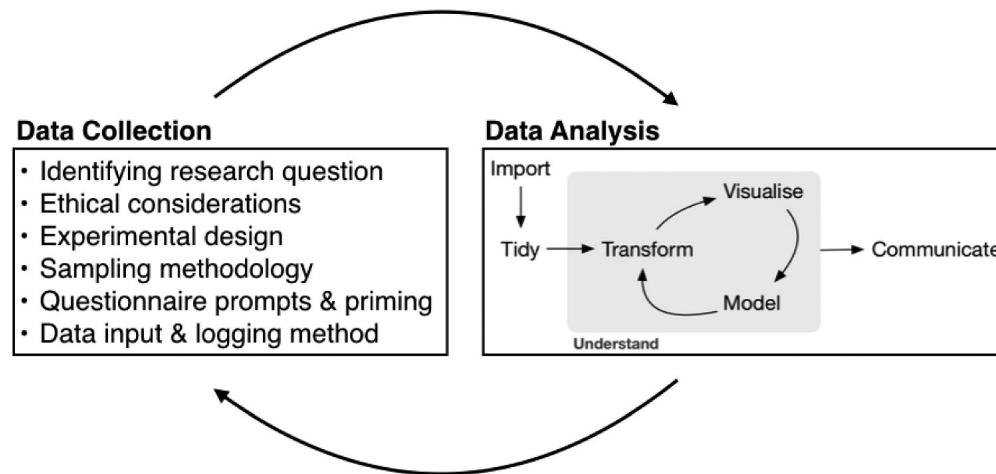
Exploratory Data Analysis in R

Ciaran Evans

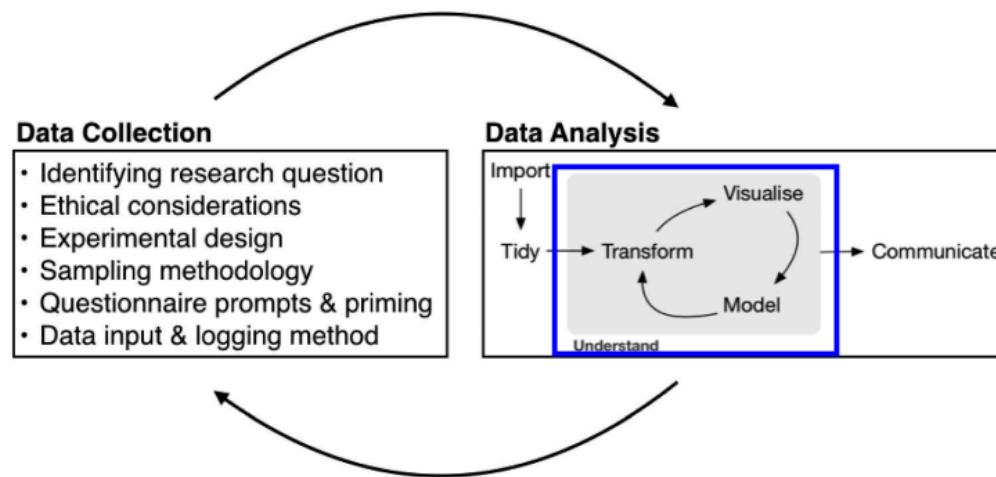
Agenda

- + Overview of exploratory data analysis
- + Introduction to R and RStudio
- + Class activity: penguins!

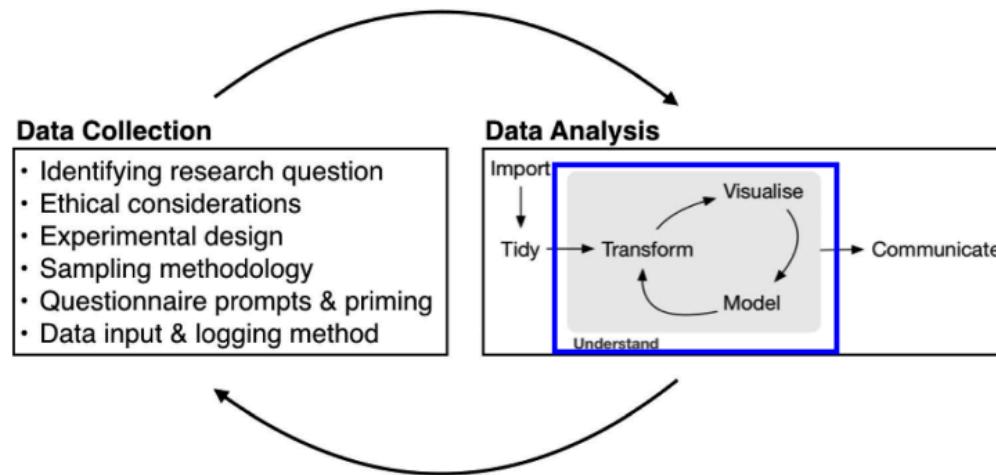
The data analysis process



The data analysis process



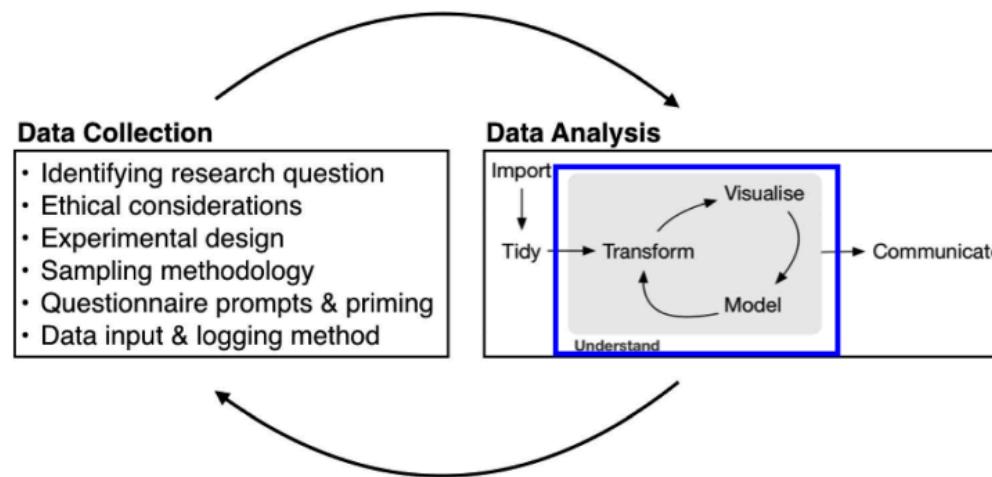
The data analysis process



Understanding:

- + Not a linear process
- + Begins with *exploratory data analysis*

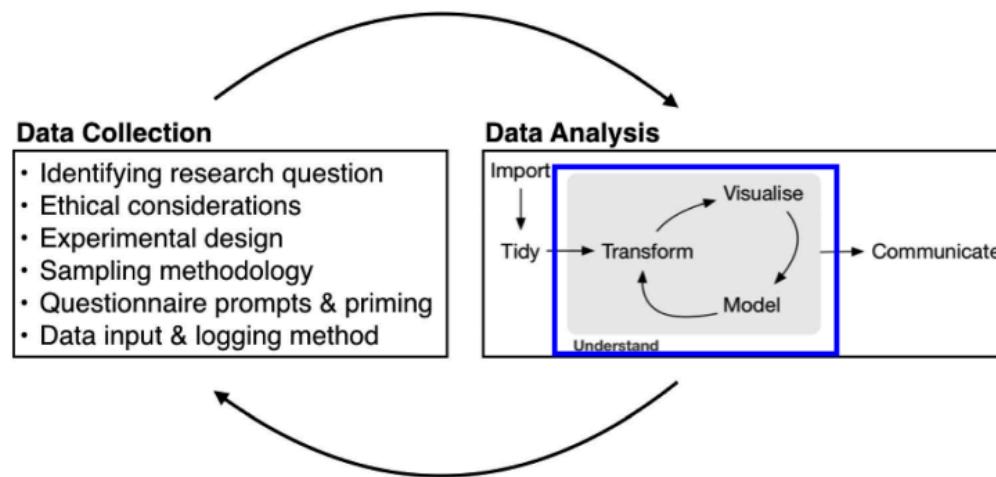
Exploratory data analysis (EDA)



Goal: get familiar with the data

- + What does the data represent?
- + How big is the data?
- + What are the rows and columns?
- + Where and when was it collected?
- + Who collected it, and what choices did they make?
- + Etc.

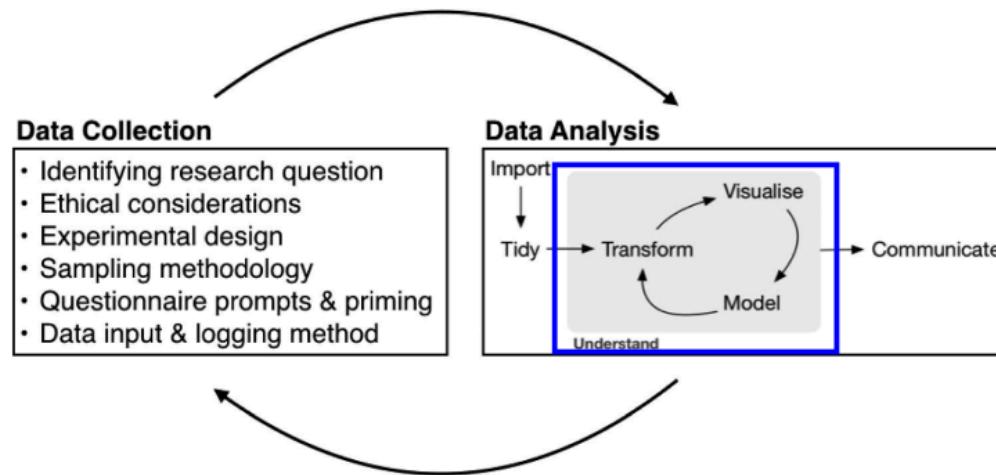
Exploratory data analysis (EDA)



Goal: get familiar with the data

- + What do the variables look like? (univariate EDA)
 - + histograms, frequency tables, summary statistics, etc.
 - + any outliers?

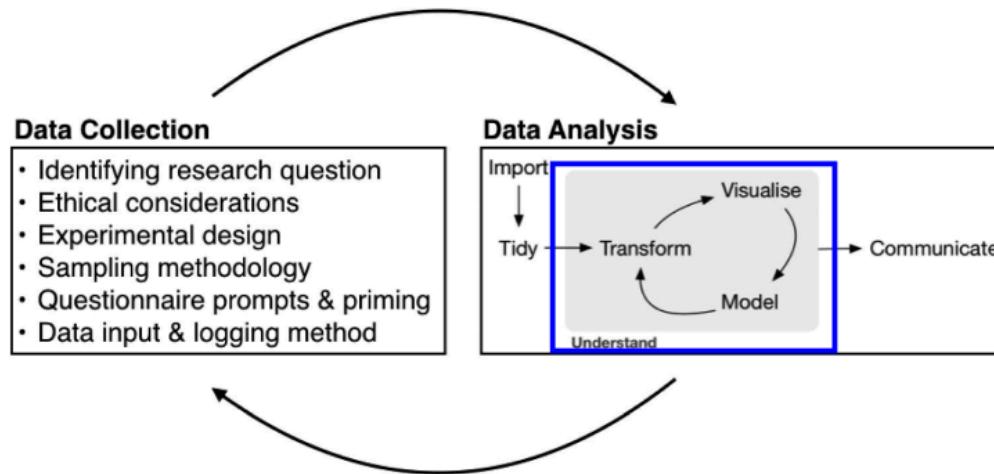
Exploratory data analysis (EDA)



Goal: get familiar with the data

- + How are the variables related? (multivariate EDA)
 - + two-way tables, scatterplots, boxplots, etc.

Exploratory data analysis (EDA)



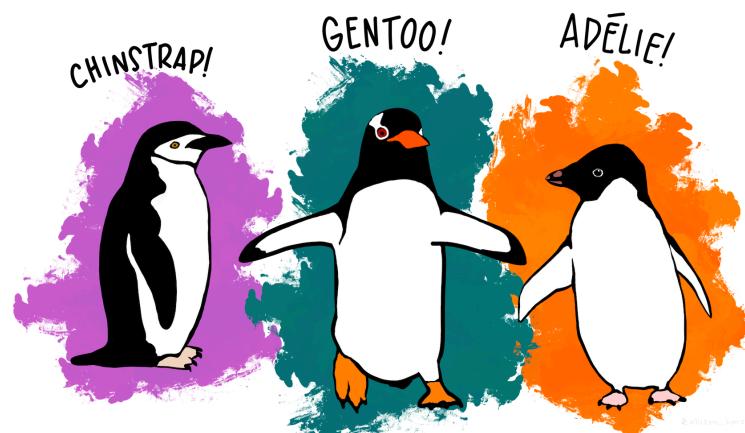
Goal: get familiar with the data

- + What relationships might we want to model?
- + generally informed by *why* we're looking at the data

Data: Penguins!

Data on 344 penguins from 3 species (Adelie, Chinstrap, Gentoo).
Variables include

- + Species
- + Bill length
- + Bill depth
- + ...



Artwork by @allison_horst

Visualizations

Bill length is a quantitative variable. What plot could we use to visualize the distribution of bill length in the penguins dataset?

(A) Scatterplot

(B) Histogram

(C) Bar chart

(D) Pie chart

Visualizations

Bill length is a quantitative variable. What plot could we use to visualize the distribution of bill length in the penguins dataset?

(A) Scatterplot

(B) Histogram

(C) Bar chart

(D) Pie chart

Answer: A histogram is a good choice for visualizing the distribution of a single quantitative variable.

Visualizations

Species is a categorical variable. What plot could we use to visualize the distribution of species in the penguins dataset?

(A) Scatterplot

(B) Histogram

(C) Bar chart

(D) Pie chart

Visualizations

Species is a categorical variable. What plot could we use to visualize the distribution of species in the penguins dataset?

(A) Scatterplot

(B) Histogram

(C) Bar chart

(D) Pie chart

Answer: A bar chart is a good choice for visualizing the distribution of a single categorical variable. Pie charts also work, but I find them harder to read.

Visualizations

Bill length and *bill depth* are both quantitative variables. What plot could we use to visualize the relationship between these two variables?

(A) Scatterplot

(B) Histogram

(C) Bar chart

(D) Pie chart

Visualizations

Bill length and *bill depth* are both quantitative variables. What plot could we use to visualize the relationship between these two variables?

(A) Scatterplot

(B) Histogram

(C) Bar chart

(D) Pie chart

Answer: A scatterplot shows the relationship between two quantitative variables.

Tools for working with data

R: Statistical software for data manipulation, visualization, computing, modeling

RStudio: Integrated development environment (IDE) that makes it easy to use R

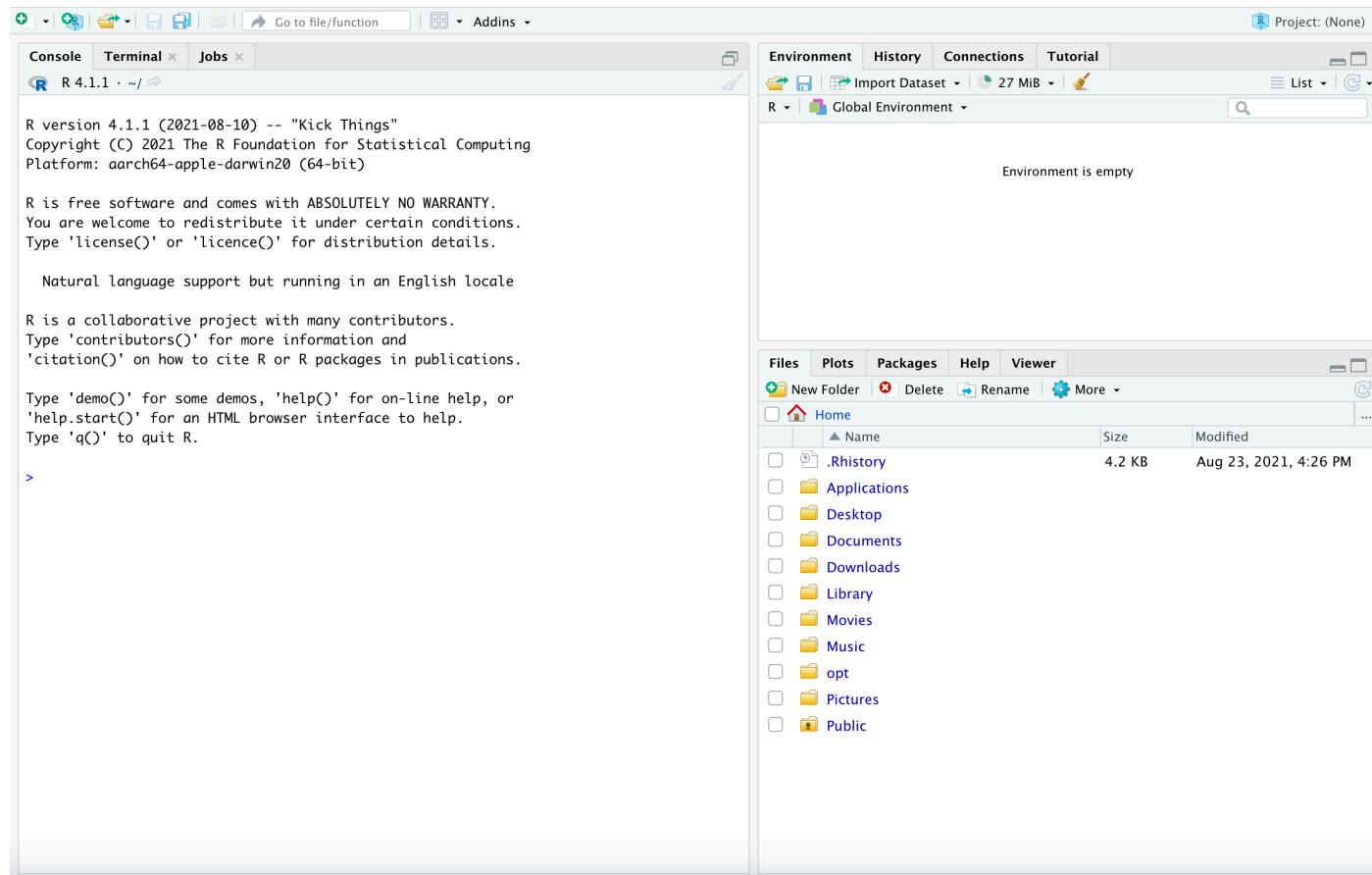
R: Engine



RStudio: Dashboard



Overview of RStudio



Panes

The screenshot shows the RStudio interface with four main panes:

- Console** (leftmost pane): Displays the R startup message and a single command prompt (>).
- Environment** (top right pane): Shows the Global Environment with a note that it is empty.
- Files** (bottom right pane): A file browser showing the contents of the Home directory. The contents are:

Name	Size	Modified
.Rhistory	4.2 KB	Aug 23, 2021, 4:26 PM
Applications		
Desktop		
Documents		
Downloads		
Library		
Movies		
Music		
opt		
Pictures		
Public		

- Plots** (not explicitly visible in the screenshot but implied by the title)

Console (Caption for the left pane)

Environment (Caption for the top right pane)

Files, plots, help (Caption for the bottom right pane)

Panes

Create a new file

The screenshot shows the RStudio interface with four main panes:

- Code Pane:** Displays an R Markdown document titled "Untitled". The code includes setup code for knitr, a brief introduction to R Markdown, and a note about the Knit button.
- Environment Pane:** Shows the R environment as empty.
- Files Pane:** Lists files and folders in the current directory, including .Rhistory, Applications, Desktop, Documents, Downloads, Library, Movies, Music, opt, Pictures, and Public.
- Console Pane:** Displays the R startup message and basic information about the R version and platform.

A black arrow points from the text "Create a new file" towards the top-left corner of the RStudio window, specifically the area where the "New File" icon is located in the toolbar.

Open and edit files

Environment

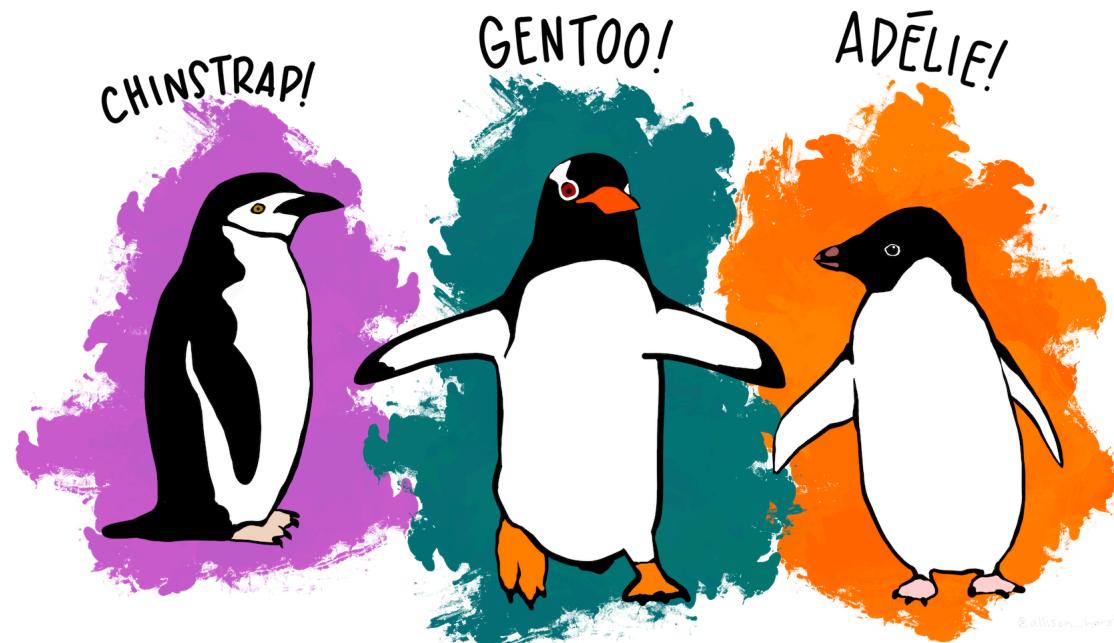
Environment is empty

Files, plots, help

Console

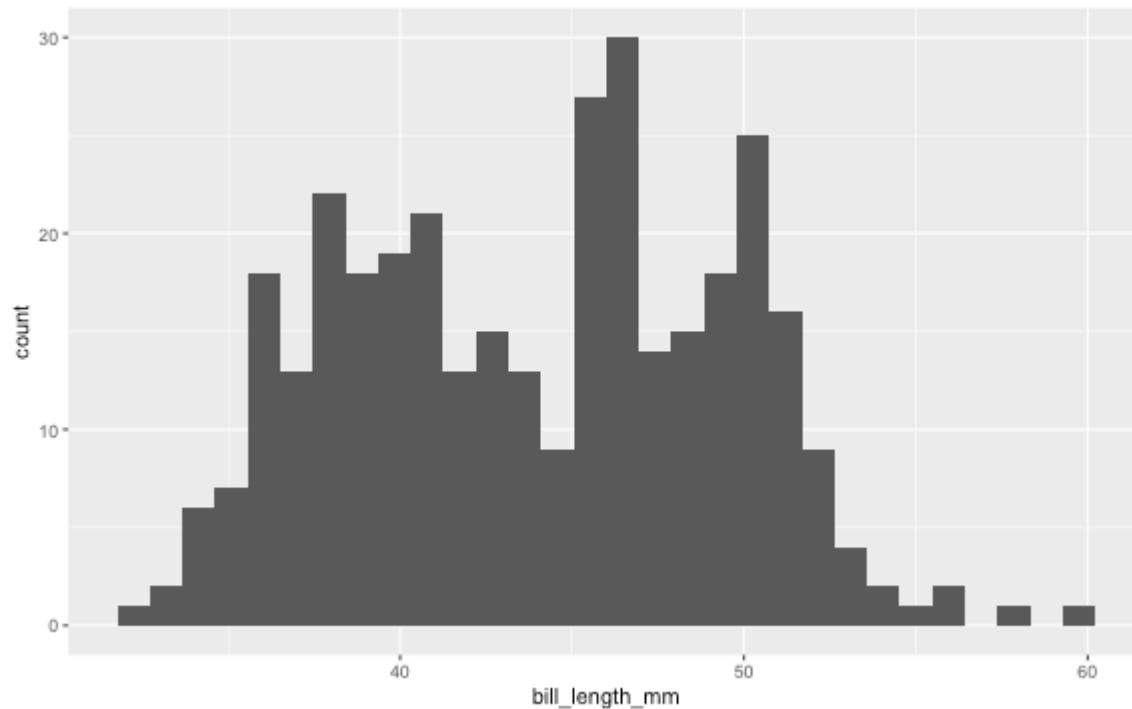
Class activity: EDA with penguins

https://sta112-s26.github.io/class_activities/ca_02.html

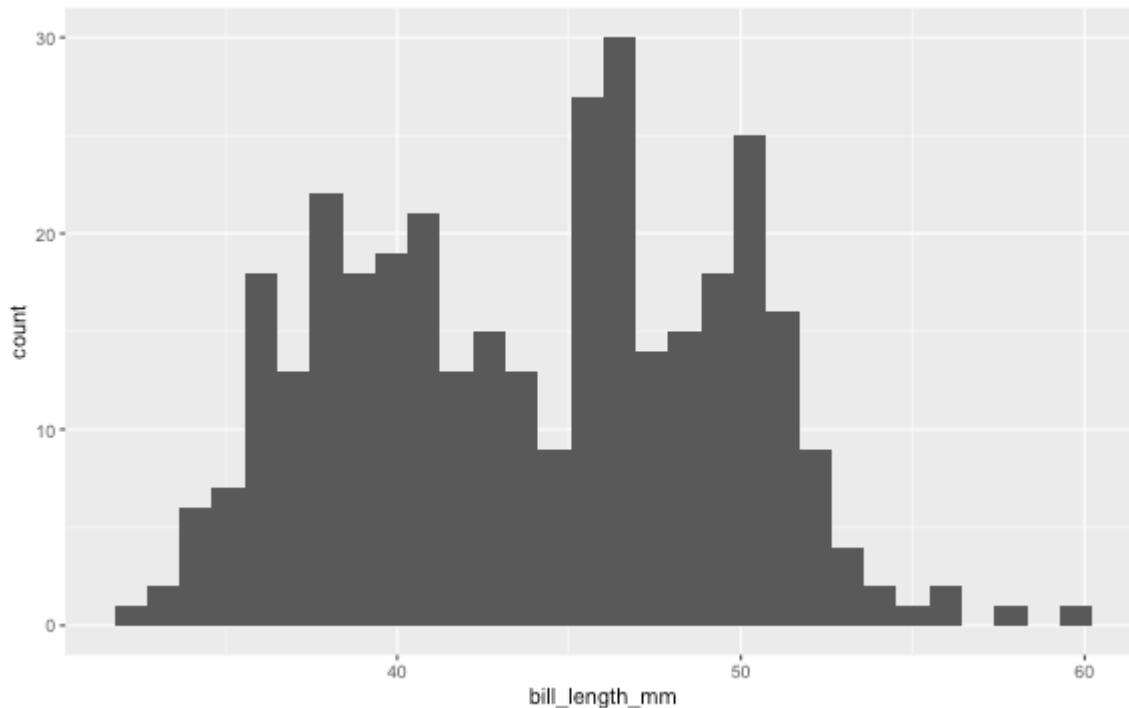


Distribution of bill length

```
penguins |>  
  ggplot(aes(x = bill_length_mm)) +  
  geom_histogram()
```



Distribution of bill length



- + Most bill lengths between 35mm and 55mm
- + Multimodal, with peaks around 40mm, 45mm, and 50mm
- + Fairly symmetric, no clear outliers

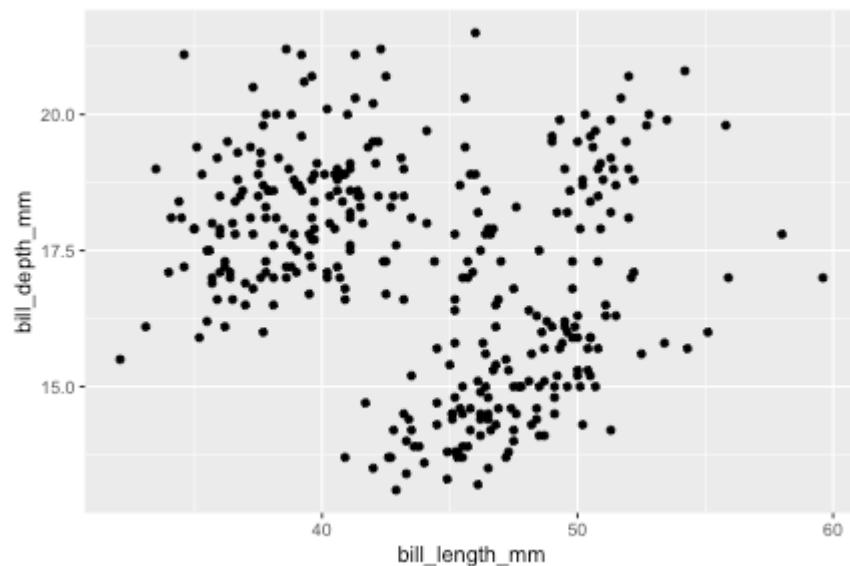
Aside: changing the number of bins

```
penguins |>  
  ggplot(aes(x = bill_length_mm)) +  
    geom_histogram(bins = 20)
```

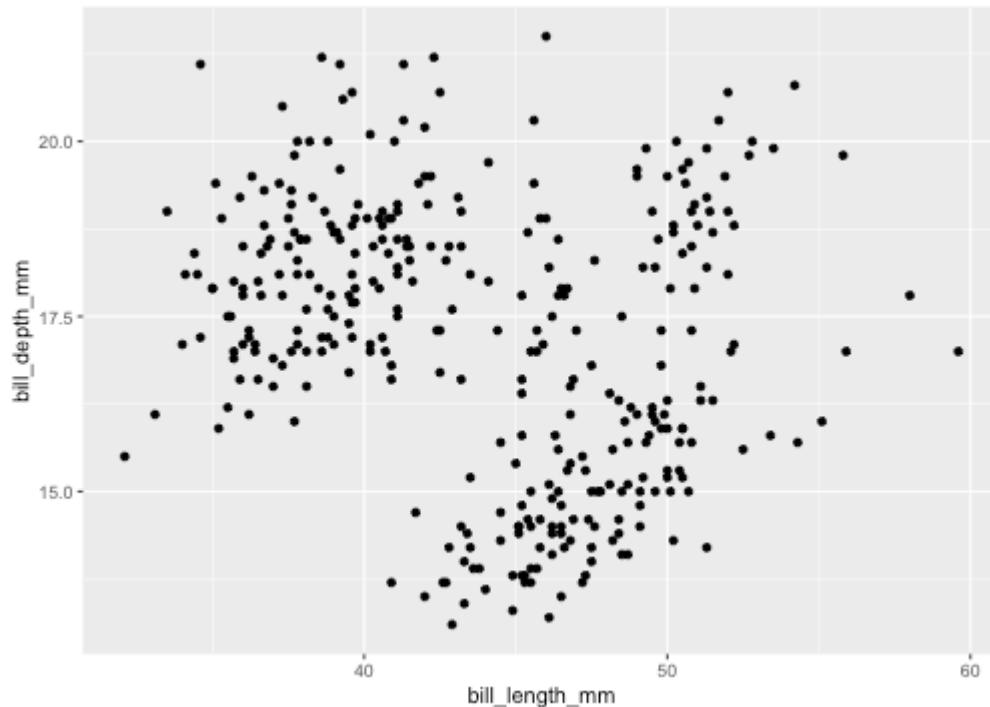
When making histograms, it is good to try different numbers of bins. The default in `geom_histogram` is 30, but can be changed with `bins = ...`

Bill depth vs. bill length

```
penguins |>  
  ggplot(aes(x = bill_length_mm,  
             y = bill_depth_mm)) +  
  geom_point()
```



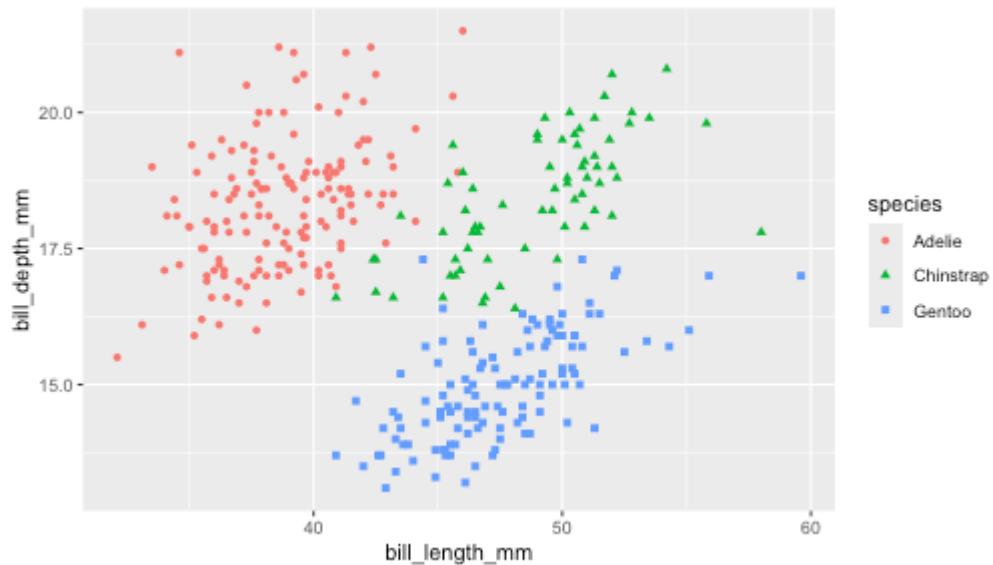
Bill depth vs. bill length



- + There does not appear to be a relationship between bill length and bill depth

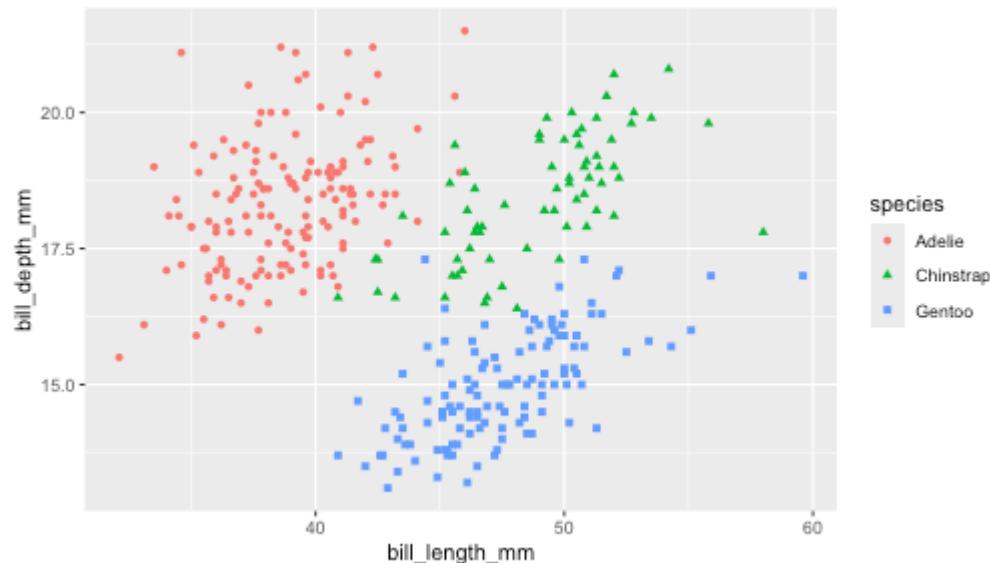
Coloring by species

```
penguins |>  
  ggplot(aes(x = bill_length_mm,  
             y = bill_depth_mm,  
             color = species, shape=species)) +  
  geom_point()
```



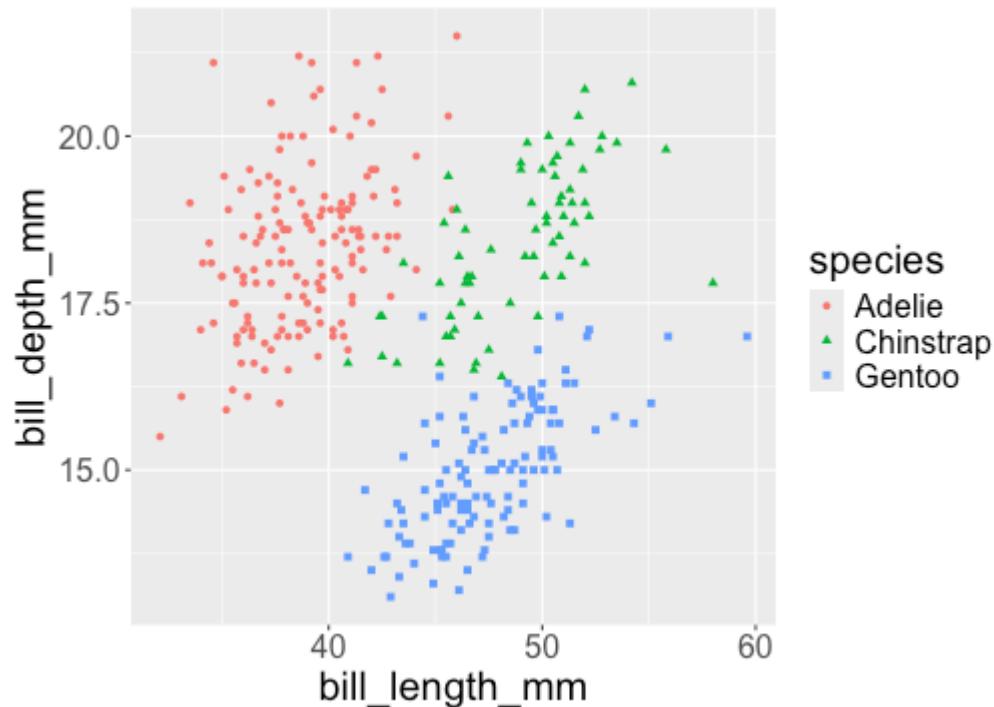
Coloring by species

```
penguins |>  
  ggplot(aes(x = bill_length_mm,  
             y = bill_depth_mm,  
             color = species, shape=species)) +  
  geom_point()
```



Within each species, there appears to be a positive, linear relationship between bill length and bill depth.

Predicting species



New penguin 🐧:

- + Bill length = 50mm, bill depth = 15mm
- + Predicted species = ?