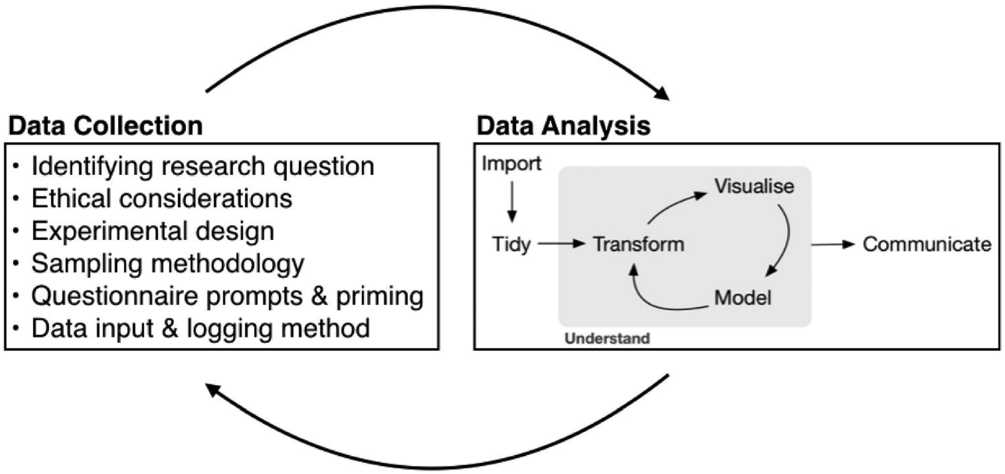
Exploratory Data Analysis in R

Ciaran Evans

Agenda

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

The data analysis process



Data Collection

- Identifying research question
- Ethical considerations
- Experimental design
- Sampling methodology
- Questionnaire prompts & priming
- Data input & logging method

Data Analysis

Import

Tidy

Transform

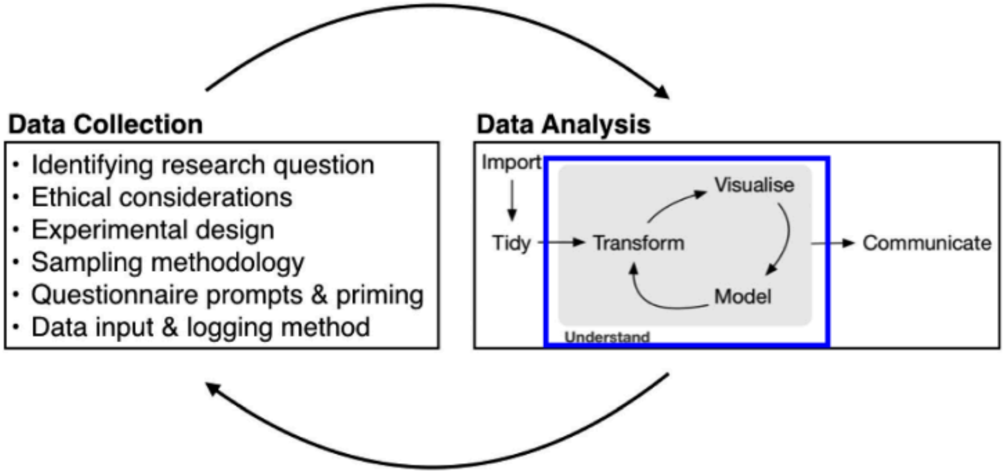
Visualise

Model

Understand

Communicate

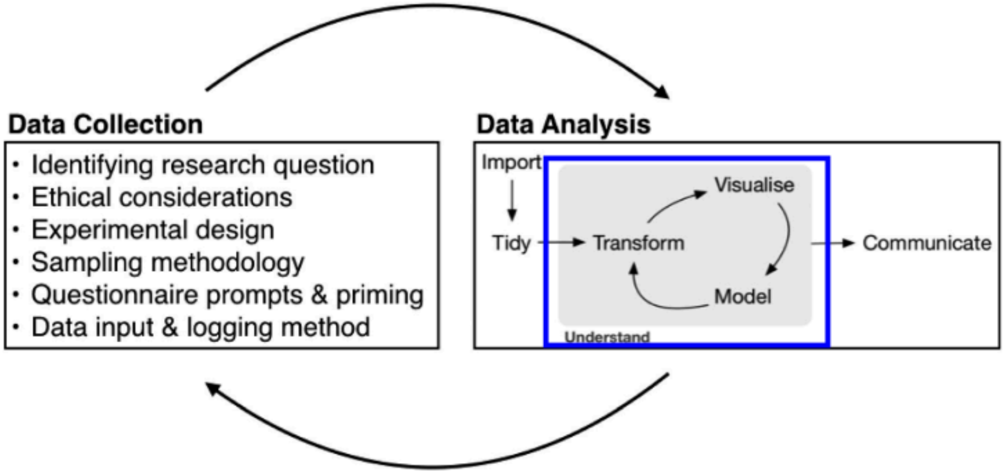
The data analysis process



Understanding:

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

Exploratory data analysis (EDA)



Data Collection

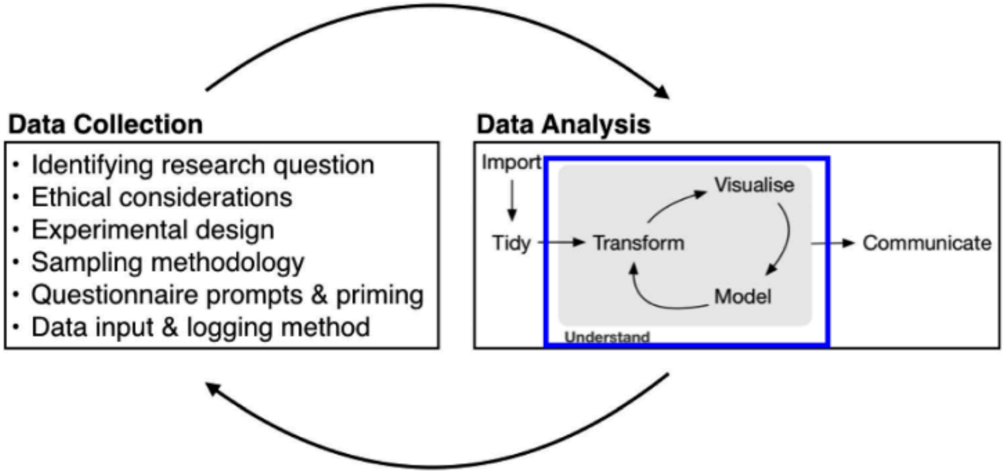
- Identifying research question
- Ethical considerations
- Experimental design
- Sampling methodology
- Questionnaire prompts & priming
- Data input & logging method

Data Analysis

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.

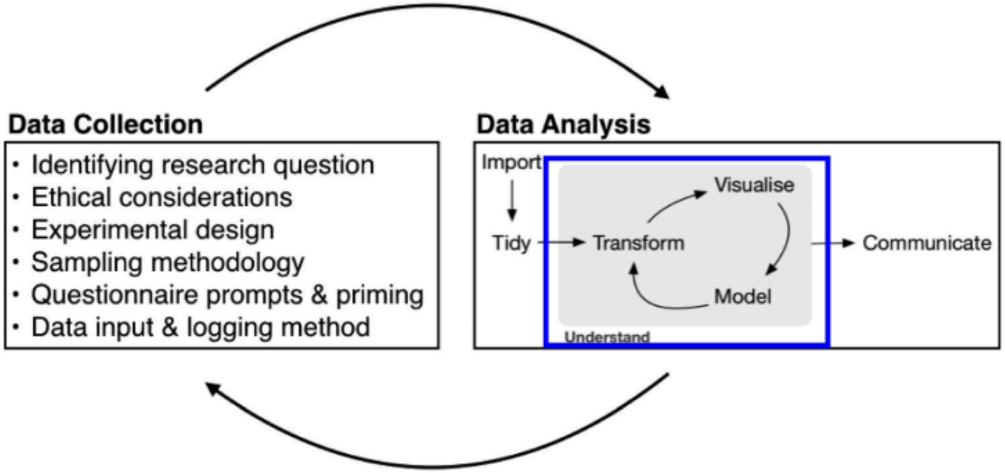
Exploratory data analysis (EDA)



Goal: get familiar with the data

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

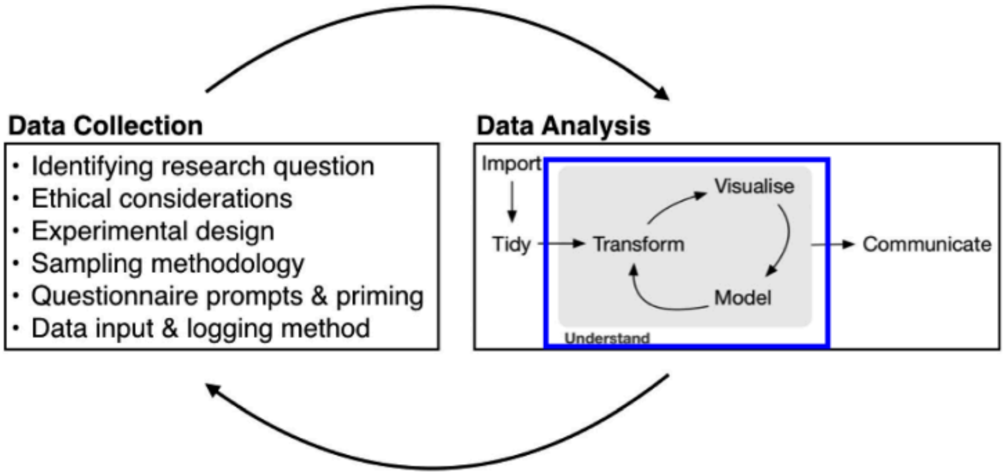
Exploratory data analysis (EDA)



Goal: get familiar with the data

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

Exploratory data analysis (EDA)



Data Collection

- Identifying research question
- Ethical considerations
- Experimental design
- Sampling methodology
- Questionnaire prompts & priming
- Data input & logging method

Data Analysis

Import

Tidy

Transform

Visualise

Model

Understand

Communicate

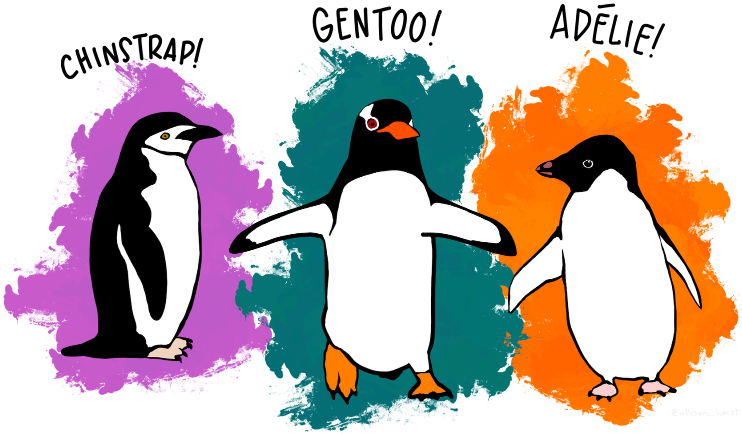
Goal: get familiar with the data

- 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?

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?

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?

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

Tools for working with data

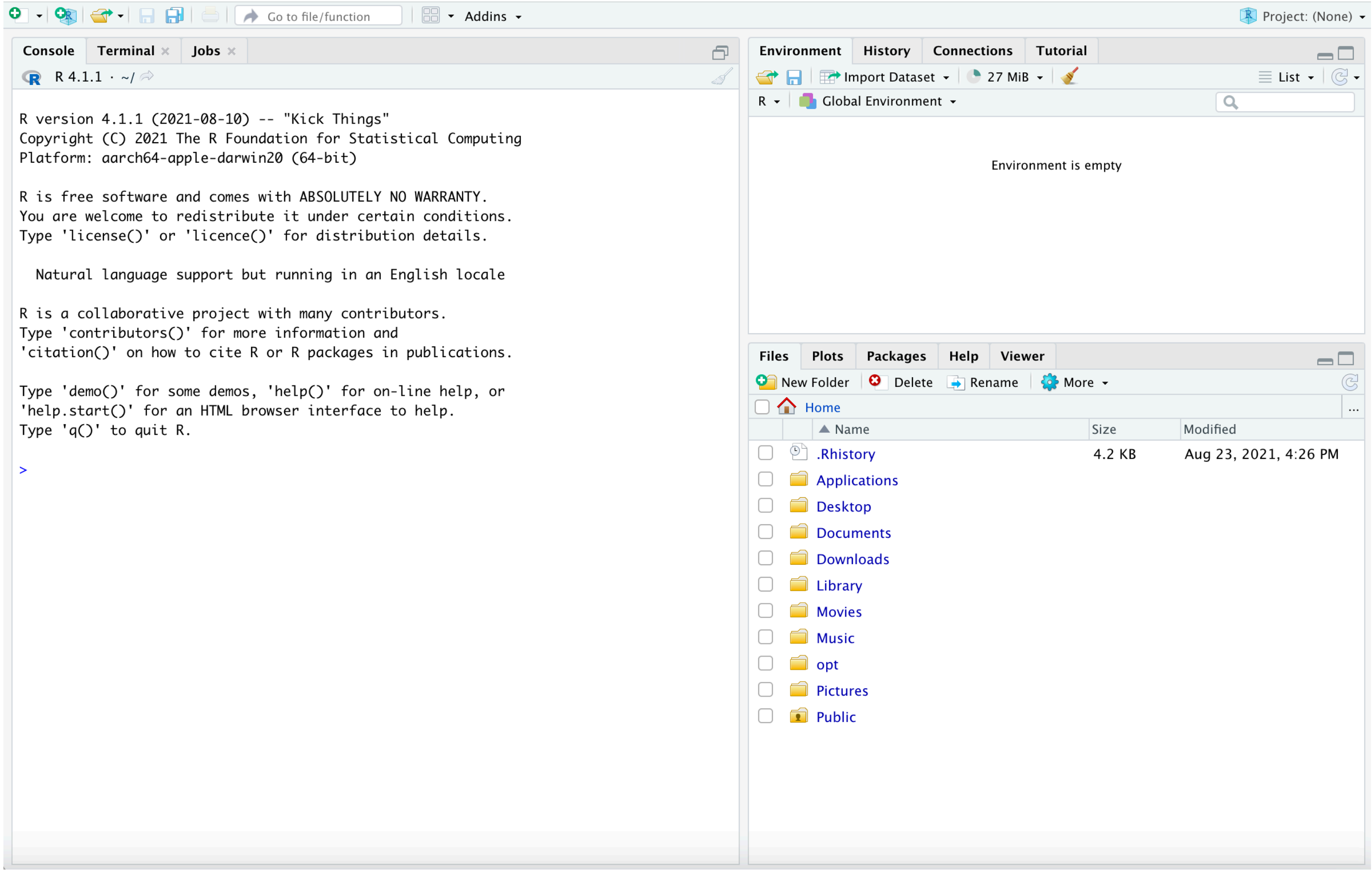
R: Statistical software for data manipulation, visualization, computing, modeling

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

R: Engine	RStudio: Dashboard

Source: Modern Dive

Overview of RStudio

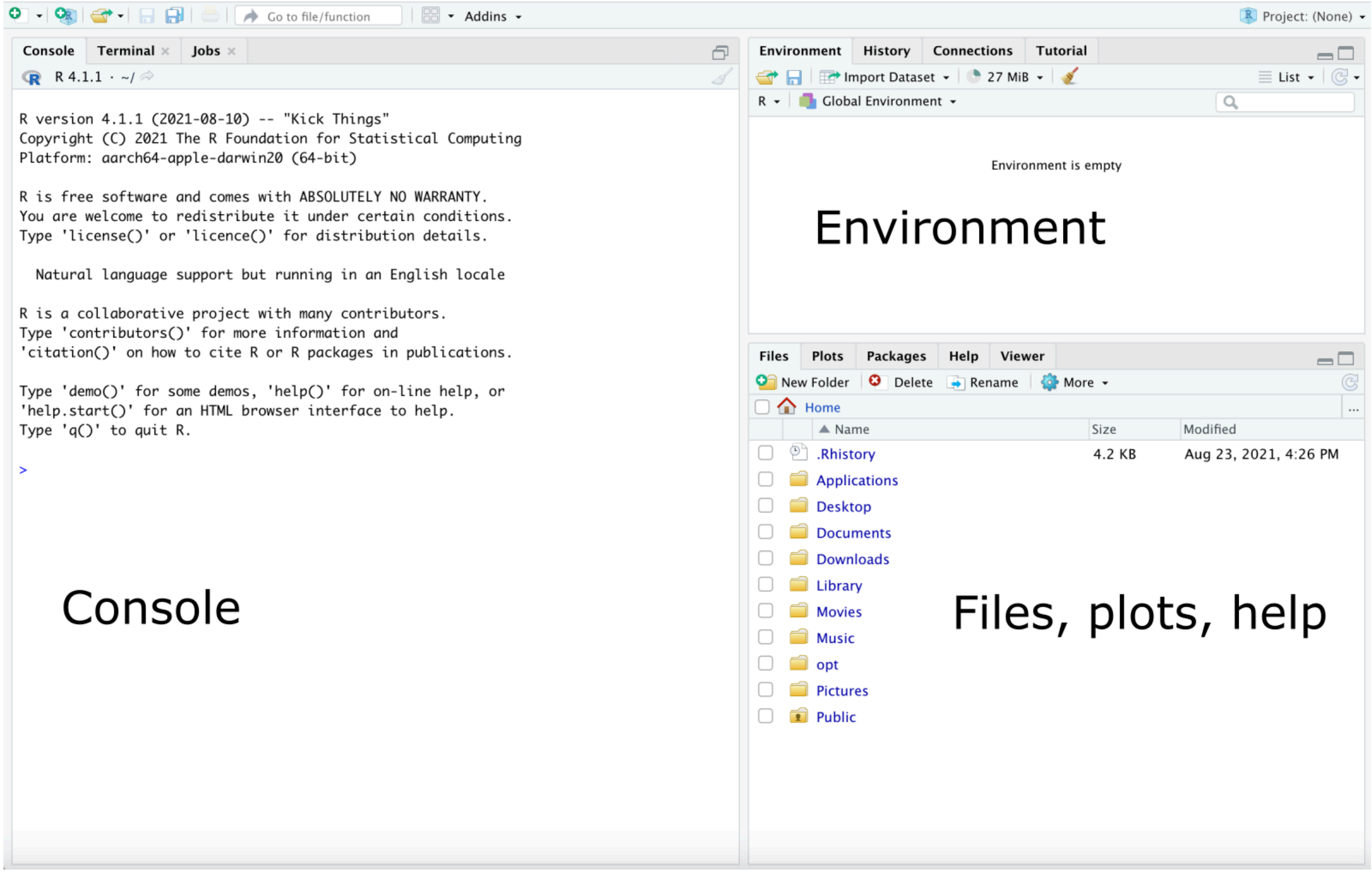


```
R version 4.1.1 (2021-08-10) -- "Kick Things"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Environment is empty

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

Panes



```
R version 4.1.1 (2021-08-10) -- "Kick Things"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Environment is empty

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

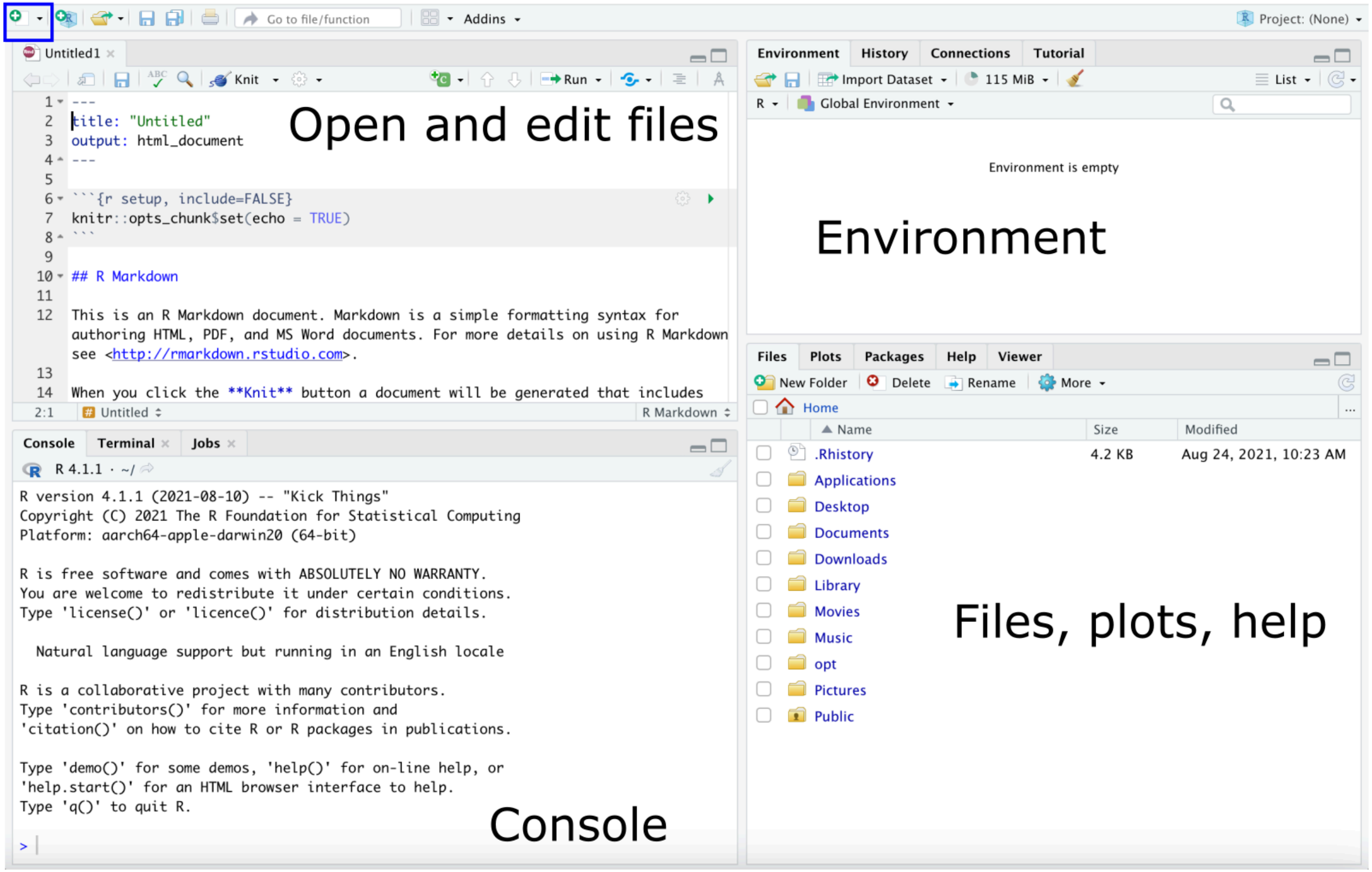
Environment

Console

Files, plots, help

Panes

Create a new file



Open and edit files

~~~
---
title: "Untitled"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for
authoring HTML, PDF, and MS Word documents. For more details on using R Markdown
see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes
~~~

Environment is empty

Environment

Console

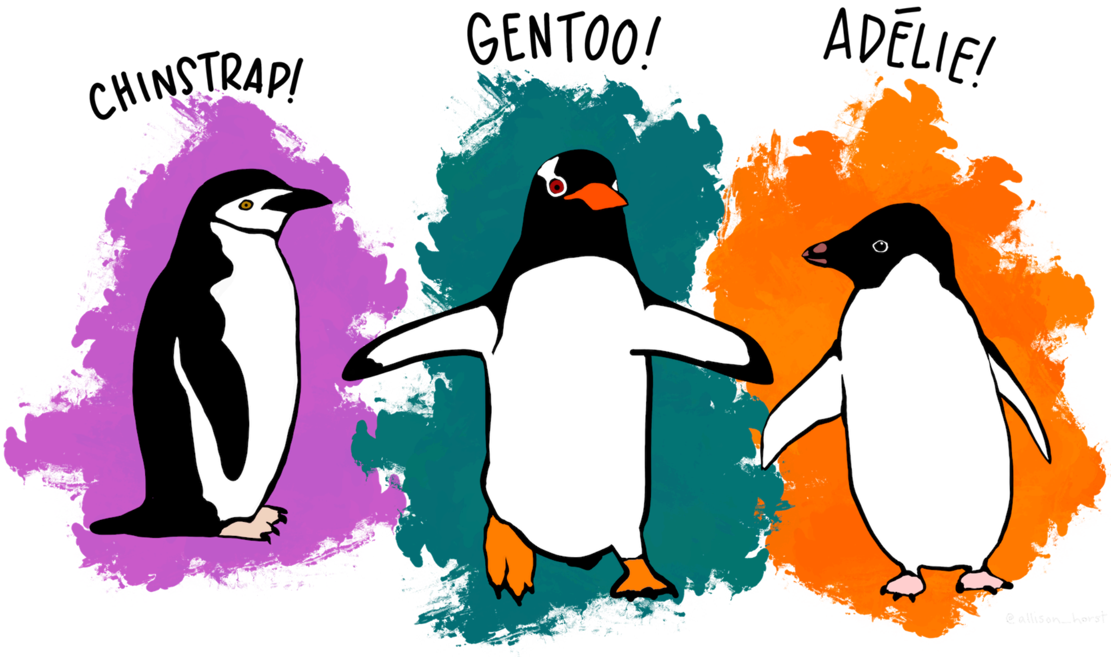
```
R version 4.1.1 (2021-08-10) -- "Kick Things"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Files, plots, help

	Name	Size	Modified
□	.Rhistory	4.2 KB	Aug 24, 2021, 10:23 AM
□	Applications		
□	Desktop		
□	Documents		
□	Downloads		
□	Library		
□	Movies		
□	Music		
□	opt		
□	Pictures		
□	Public		

Class activity: EDA with penguins

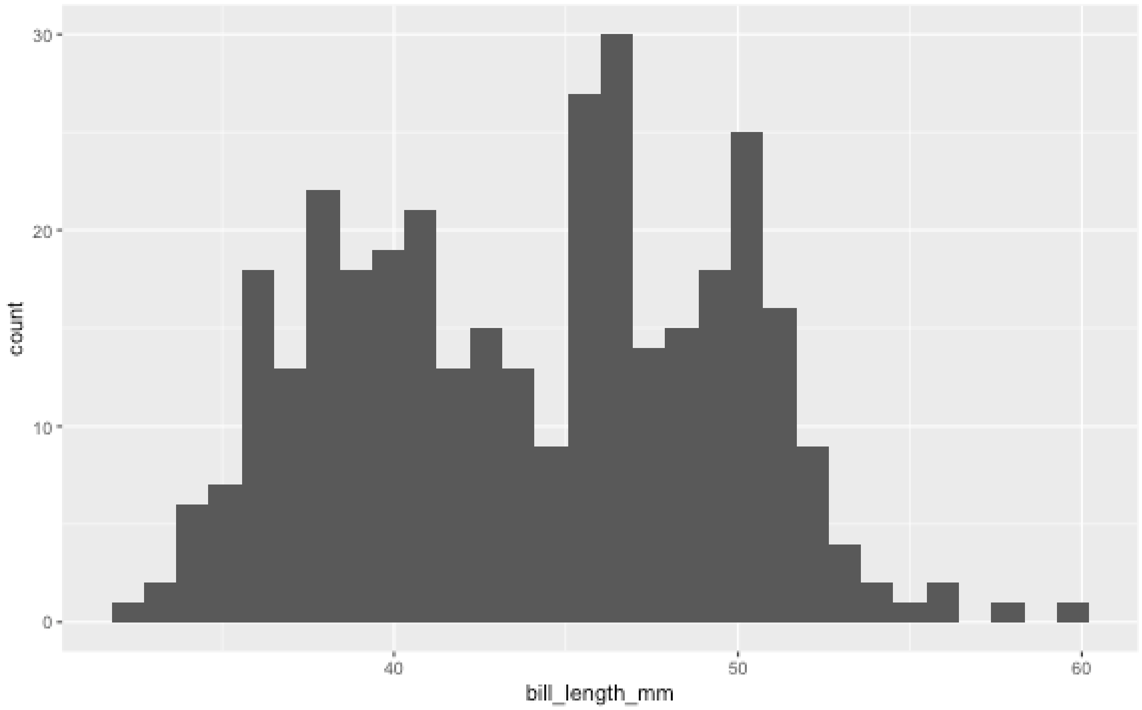
https://sta112-f25.github.io/class_activities/ca_02.html



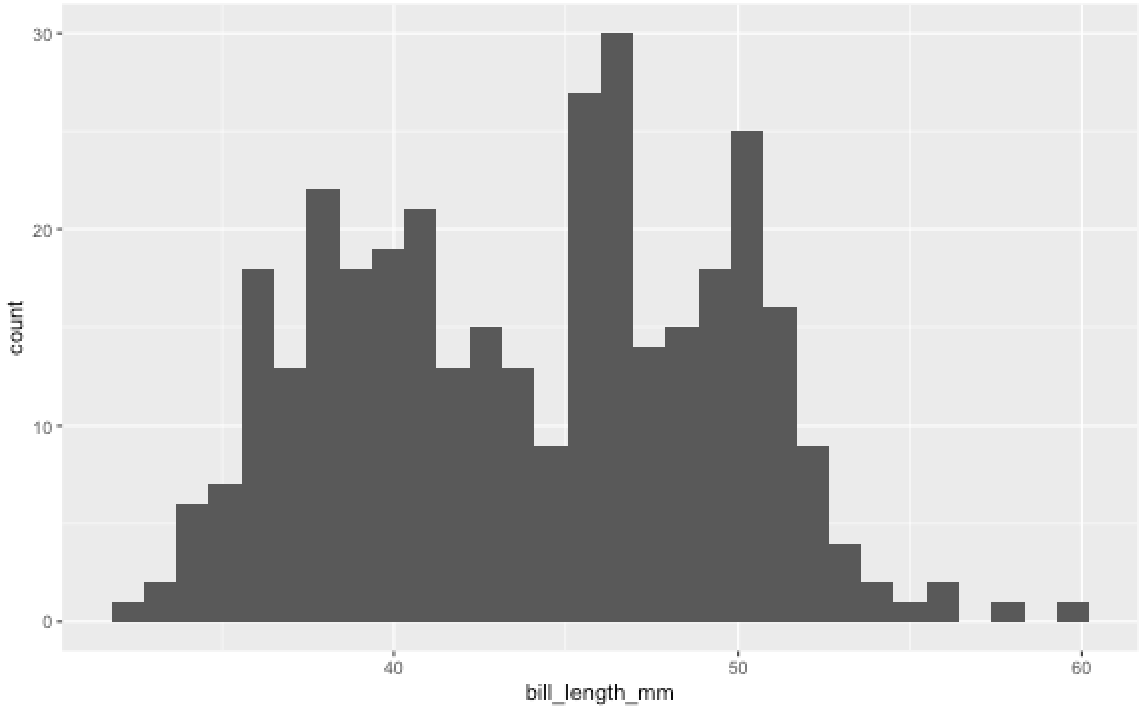
Artwork by @allison_horst

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

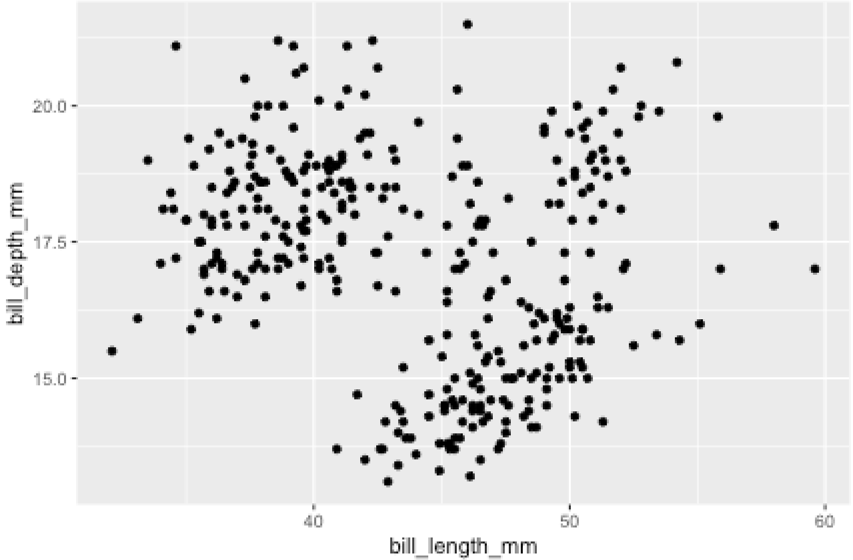
Aside: changing the number of bins

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

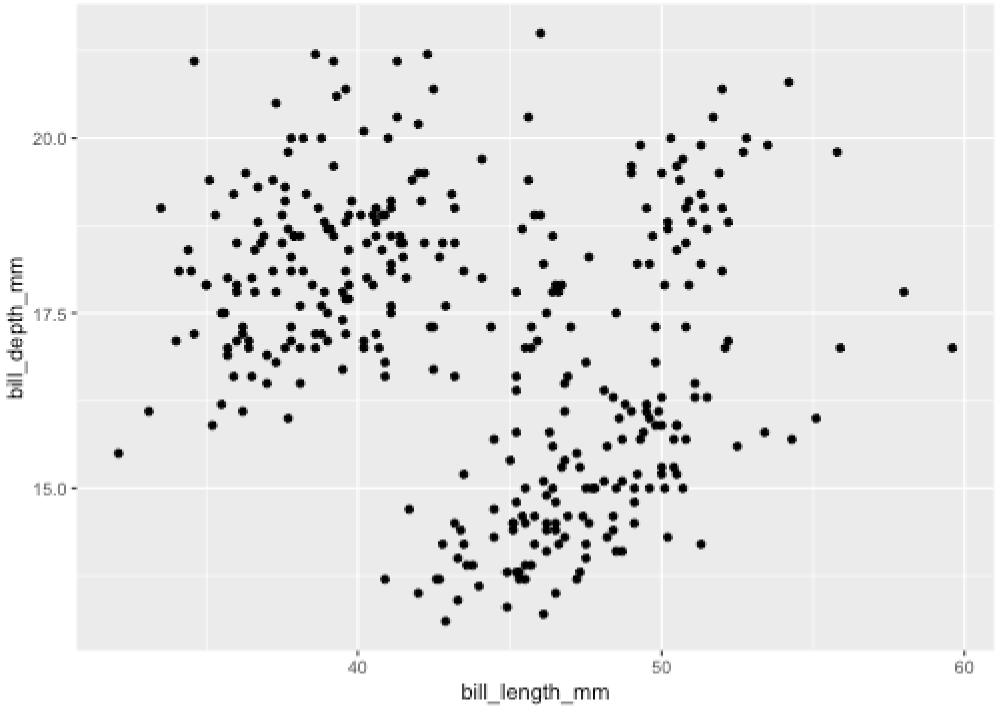
When making histograms, it is good to try different numbers of bins. The default in `geom_histogram` is 30, but can be changed with `bins = ...`

Bill depth vs. bill length

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



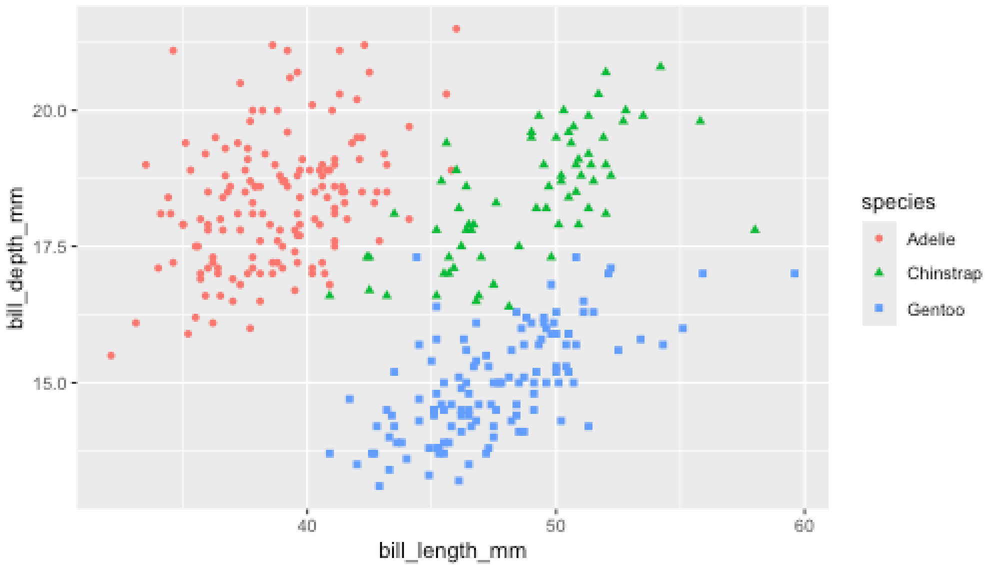
Bill depth vs. bill length



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

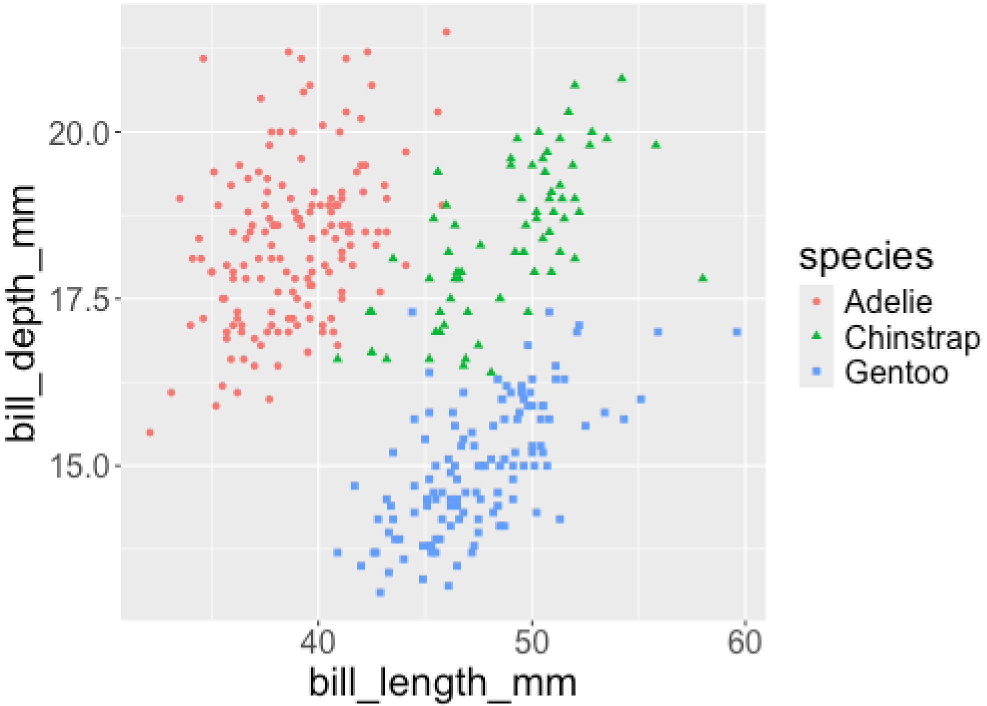
Coloring by species

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



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 = ?