Class 02

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-08-21

Load Packages

```r
library(Hmisc)
library(GGally)
library(palmerpenguins)
library(tidyverse)

theme_set(theme_bw())
```

Welcome Back

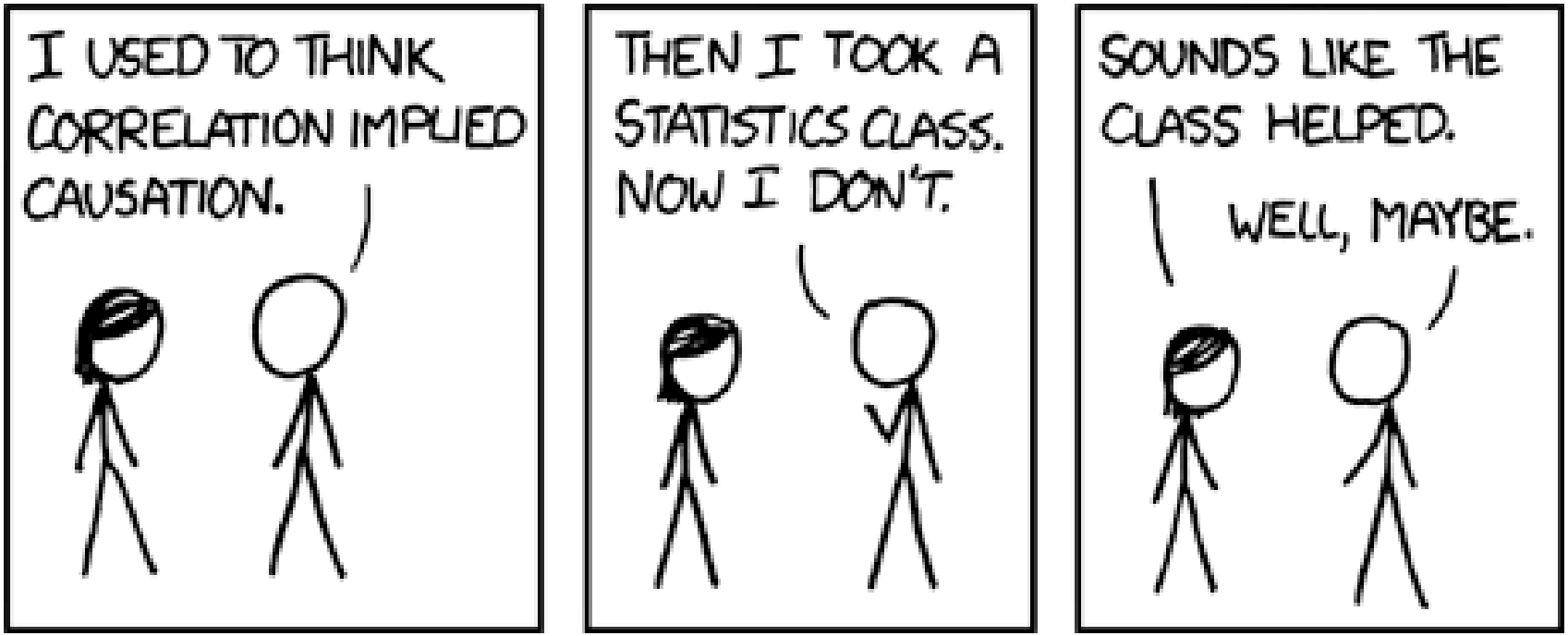
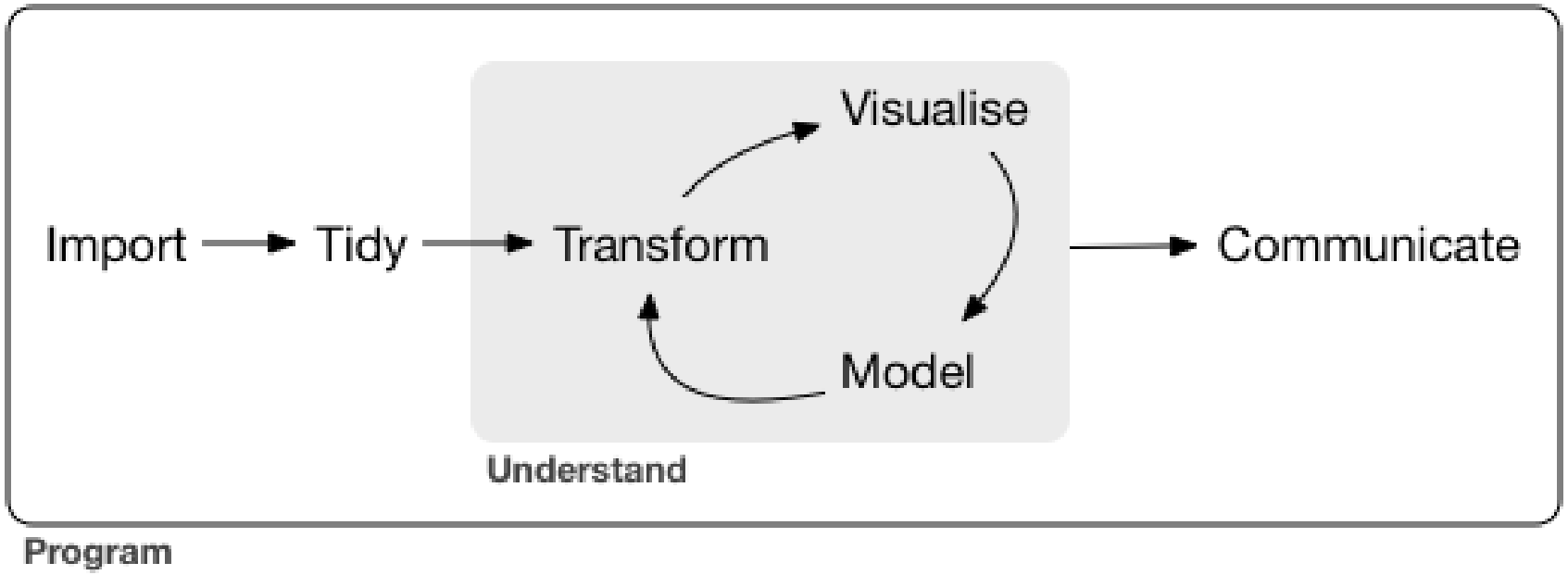


Figure 1: *Correlation*

- *Source: XKCD*

What is data science?



Source: Figure 1.1 in https://r4ds.hadley.nz/intro.html

Chatfield's Six Rules for Data Analysis

1. *Do not attempt to analyze the data until you understand what is being measured and why.*
2. *Find out how the data were collected.*
3. *Look at the structure of the data.*
4. *Carefully examine the data in an exploratory way, before attempting a more sophisticated analysis.*
5. *Use your common sense at all times.*
6. *Report the results in a clear, self-explanatory way.*

Chatfield, Chris (1996) Problem Solving: A Statistician's Guide, 2nd ed.

Introduction to R

- *R is an open source statistical programming language managed by a core team of 15 people and thousands of code writers/statisticians who contribute their work*
- *Most of R is written in R*
- *Community available for fixing bugs/software*
- *Promotes reproducible research through open and accessible tools*

R vs RStudio

- *R is the programming language itself*
- *RStudio is an interface for working with R*

RStudio

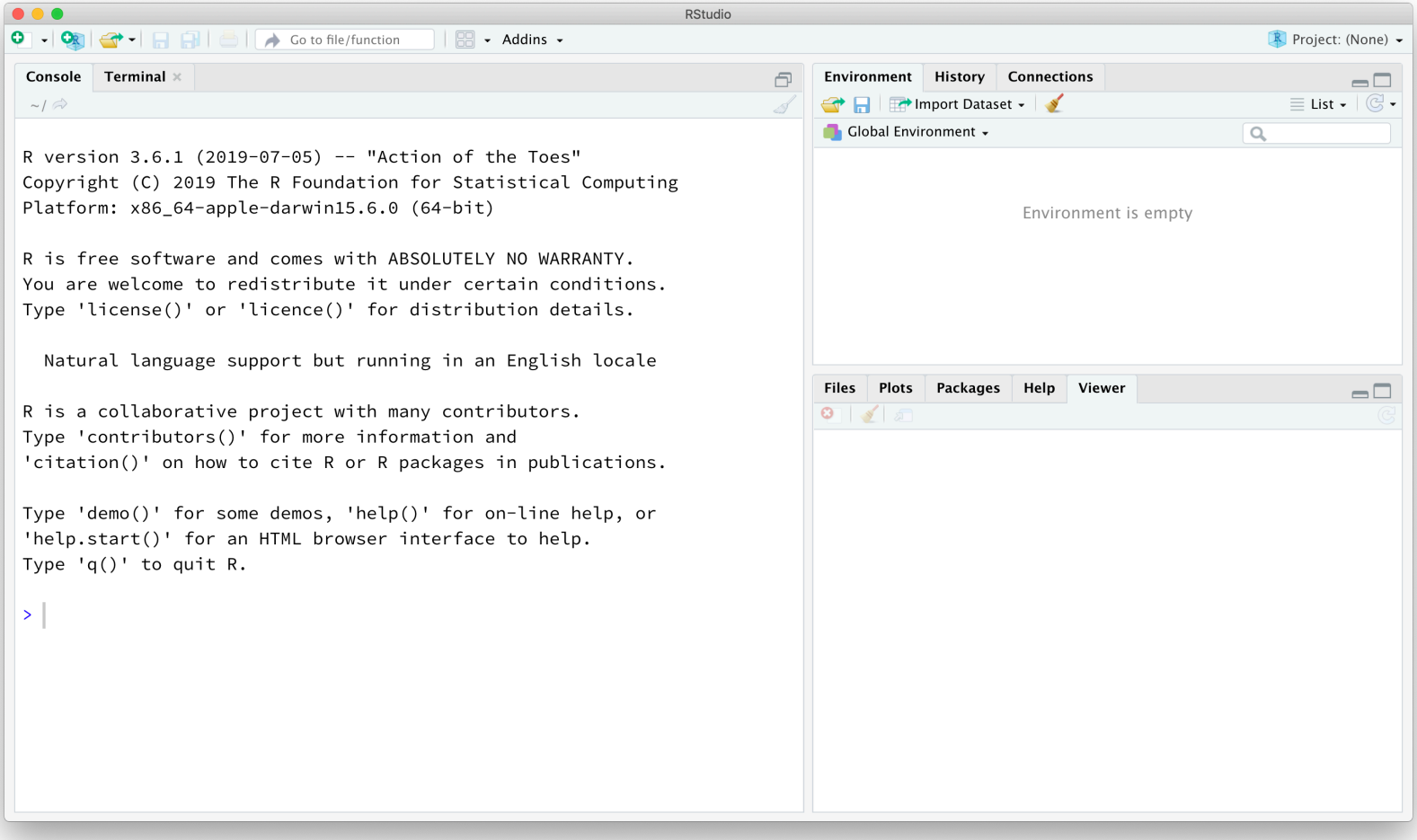


Figure 2: *RStudio Default Screen*

Projects

Projects are a convenient way to keep all your files for an analysis in one place.

Go to File > New Project to begin one now. Call the project "homework1" and save it to your computer in a folder for this class.

Types of Document

- *R script*
  - *End with .R and are pure code. If you run them, output will appear in the bottom left corner called the console.*
- *Quarto documents*
  - *End in .qmd and use markdown language to turn characters into formatted text.*
  - *Processes code in code chunks, and output appears directly in the document*
- *Begin a new markdown script now*

Environment

Your project now has it's own "environment" in which you can store your data, variables and results.

Add a code chunk to your document, copy the code below, and run it.

Example:

```
x <- c(1, 2, 3, 4, 5)
```

Environment (cont.)

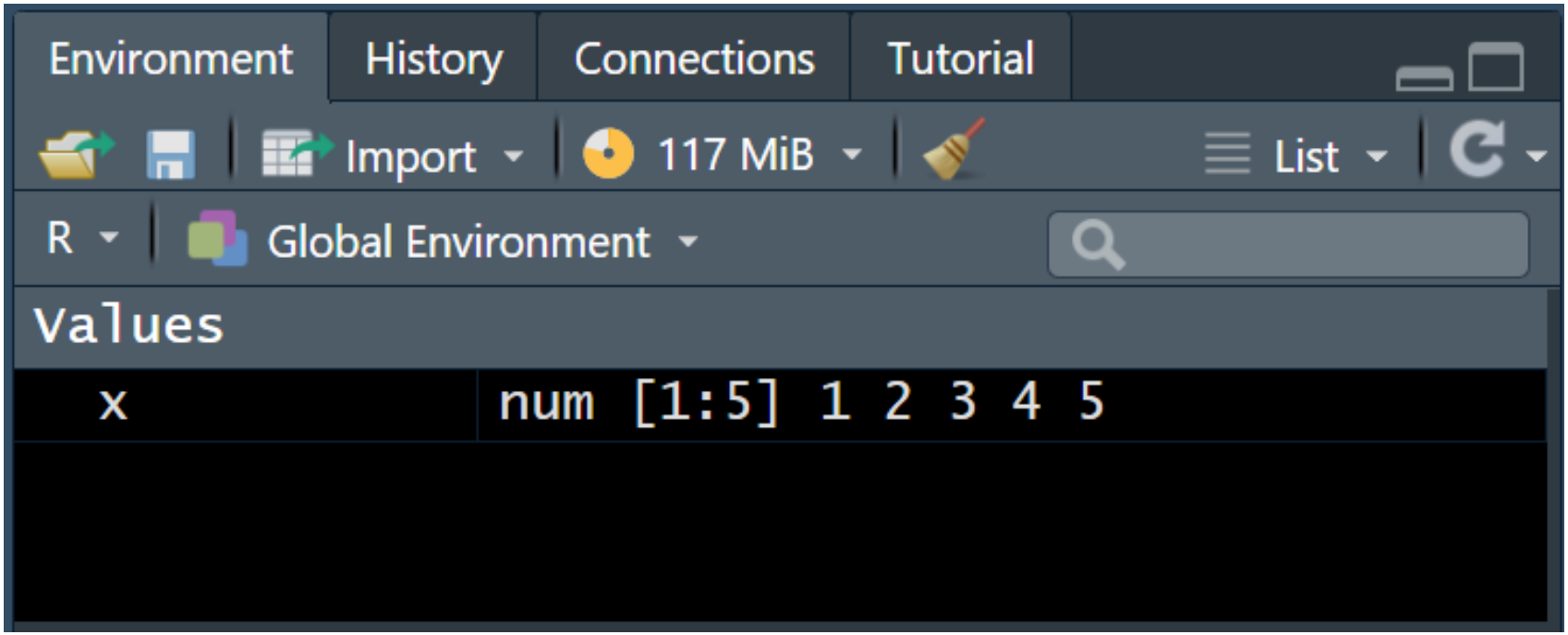
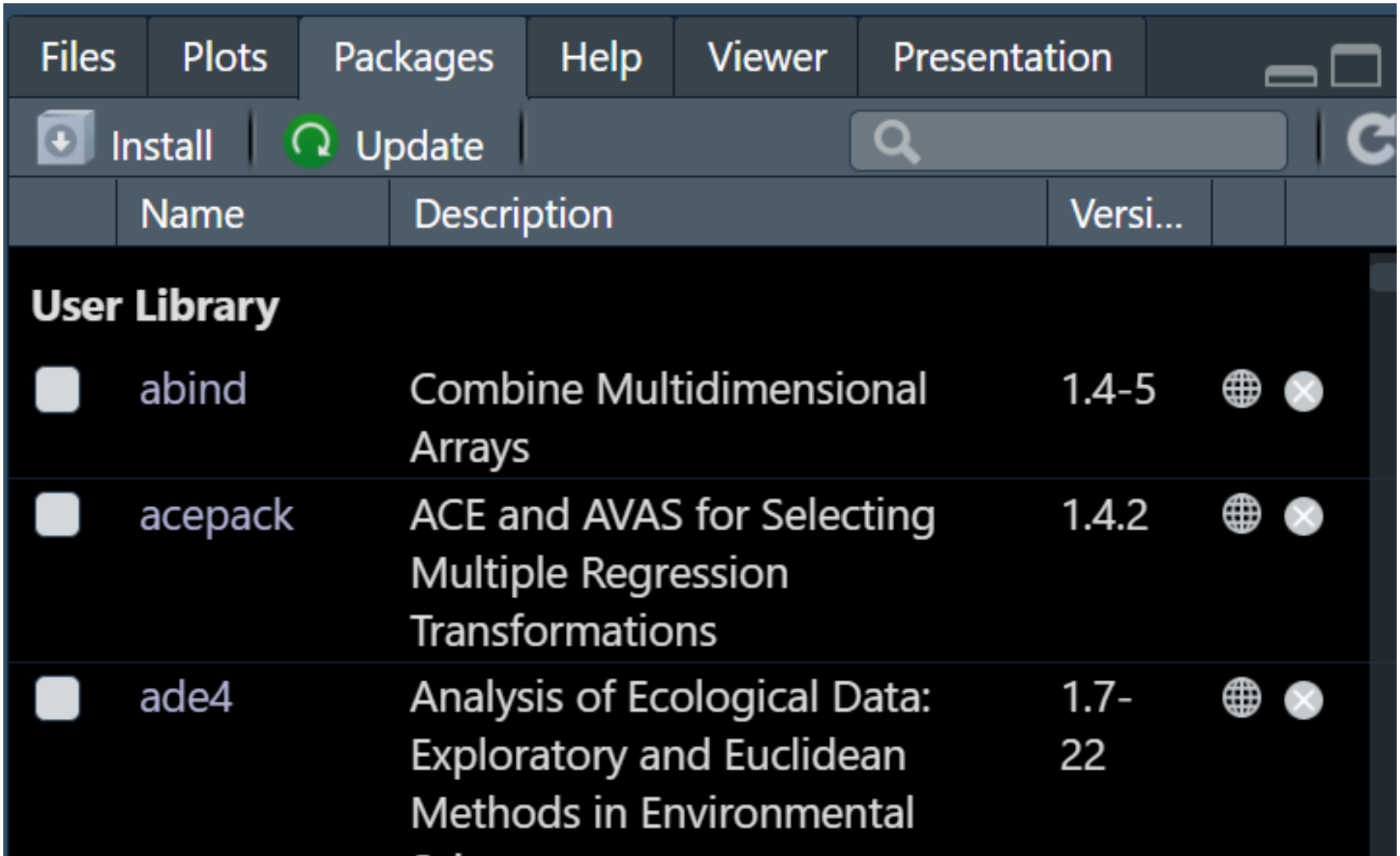


Figure 3: *Stored variable now appears in the environment*

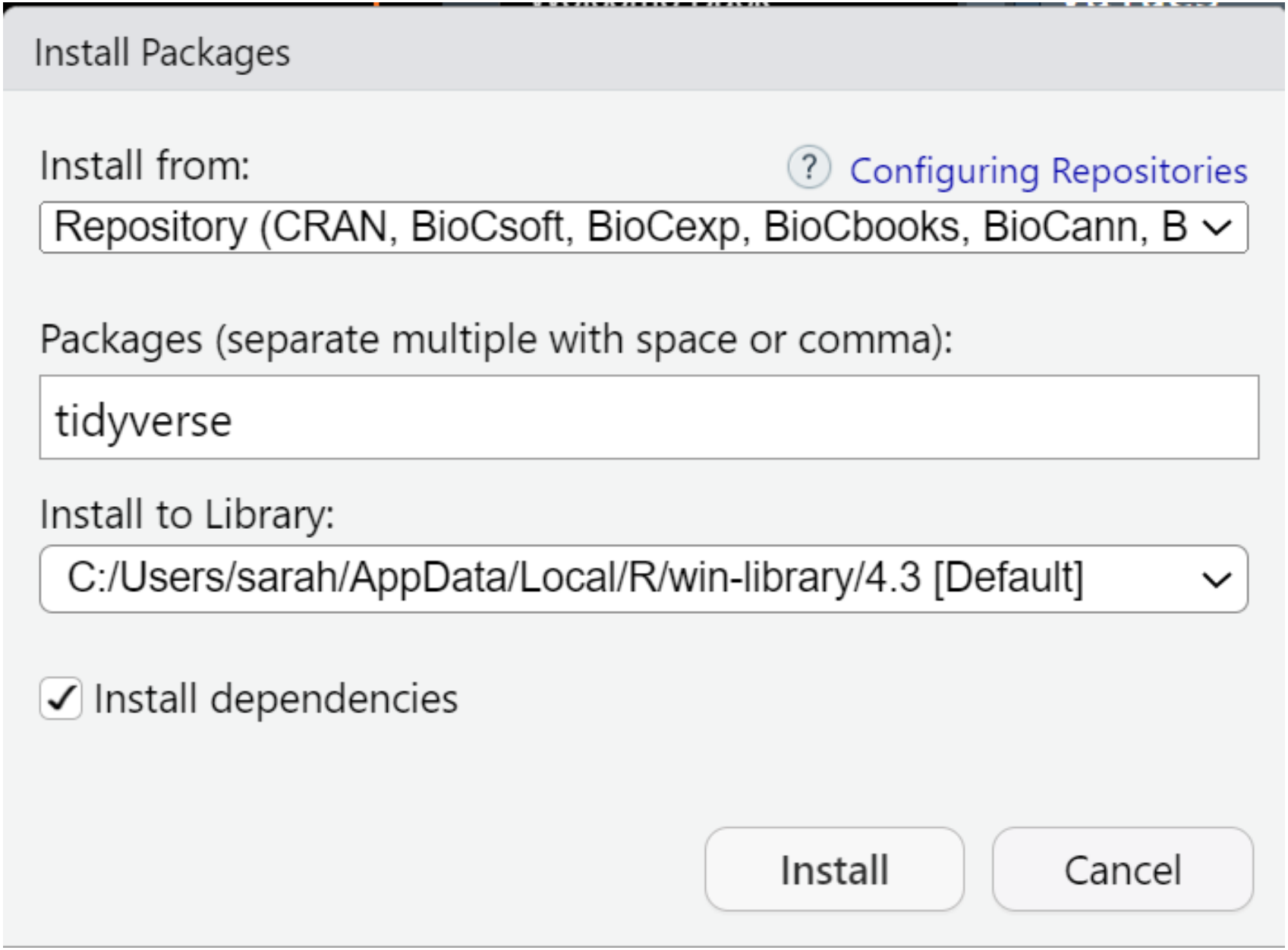
Packages

Packages are collections of functions to use for statistical analyses. Some are loaded automatically, and some need to be separately installed. Let's install the `tidyverse` *package.*



Install Packages

Either...



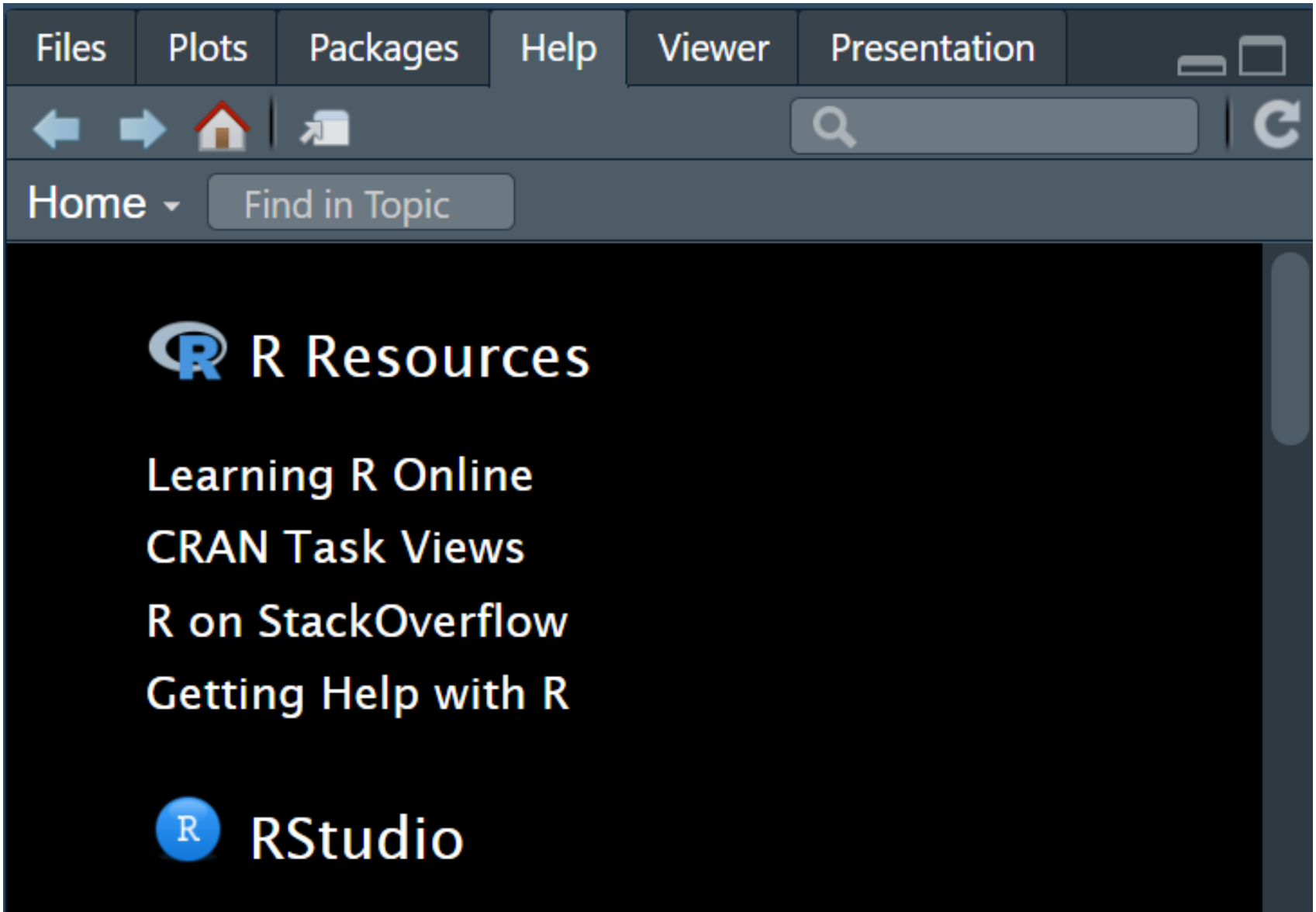
Install packages

or...

```
install.packages('tidyverse',
                 dependencies = T)
install.packages('openintro',
                 dependencies = T)
```

Getting Help

Search for functions, packages, vignettes, and more directly in RStudio in the "Help" panel.



Exercise: Palmer Penguins

Figure 6: *The Palmer Penguins*

Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi: 10.5281/zenodo.3960218.

Install the package and load the library

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

Find the data

```r
data(package = 'palmerpenguins')

# Add the data to your environment
penguins <- penguins
```

Inspect the data

```r
head(penguins)
```

```
# A tibble: 6 x 8
  species island    bill_length_mm bill_depth_mm flipper_le
  <fct>   <fct>              <dbl>         <dbl>
1 Adelie  Torgersen           39.1          18.7
2 Adelie  Torgersen           39.5          17.4
3 Adelie  Torgersen           40.3          18
4 Adelie  Torgersen           NA            NA
5 Adelie  Torgersen           36.7          19.3
6 Adelie  Torgersen           39.3          20.6
# i 2 more variables: sex <fct>, year <int>
```

Dataframes

- *Data structure in rows and columns like a spreadsheet*
- *Rows: (ideally) uniquely identified observations*
- *Columns: parameters which describe the observations*

How many rows does `penguins` have?

```
nrow(penguins)
```

```
[1] 344
```

How many variables does `penguins` have?

```r
ncol(penguins)
```

```
[1] 8
```

```r
colnames(penguins)
```

```
[1] "species"           "island"            "bill_length_mm
[4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
[7] "sex"               "year"
```

Can I find this out more quickly?

```r
glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, A
$ island            <fct> Torgersen, Torgersen, Torgersen,
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3,
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6,
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181,
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650,
$ sex               <fct> male, female, female, NA, female,
$ year              <int> 2007, 2007, 2007, 2007, 2007, 200
```

How else can I get a description of the data?

Use the `Hmisc::describe()` function to quickly summarize data.

```
Hmisc::describe(penguins)
```

How else can I get a description of the data?

```
penguins

 8  Variables      344  Observations
---------------------------------------------------------------
species
       n  missing distinct
     344        0        3

Value          Adelie Chinstrap    Gentoo
Frequency         152        68       124
Proportion      0.442     0.198     0.360
---------------------------------------------------------------
island
       n  missing distinct
     344        0        3

Value          Biscoe     Dream Torgersen
Frequency         168       124        52
```

Meet the penguins!

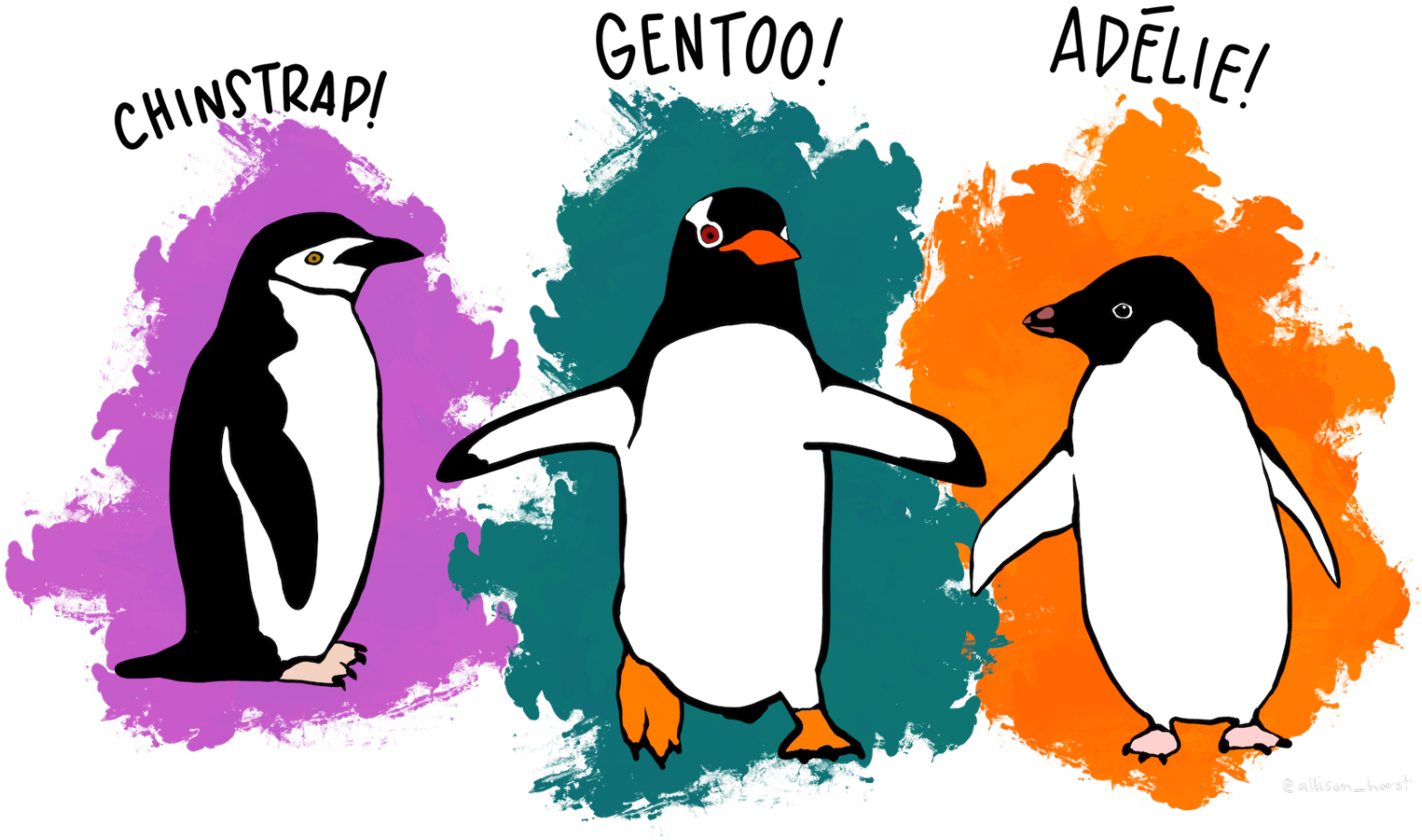


Figure 7: *Meet the Palmer Penguins*

Types of data

The key distinction we'll make is between

- ***quantitative*** *(numerical) and*
- ***categorical*** *(qualitative) information.*

Information that is quantitative describes a ***quantity****.*

- *All quantitative variables have units of measurement.*
- *Quantitative variables are recorded in numbers, and we use them as numbers (for instance, taking a mean of the variable makes some sense.)*

Continuous vs. Discrete Quantities

Continuous *variables (can take any value in a range) vs.* ***Discrete*** *variables (limited set of potential values)*

- *Is time a continuous or a discrete variable?*
- *Time is certainly continuous as a concept, but how precise is our unit (e.g. hour, year, decade)?*

Qualitative (Categorical) Data

Qualitative variables consist of names of categories.

- *Each possible value is a code for a category (could use numerical or non-numerical codes.)*
  - ***Binary*** *categorical variables (two categories, often labeled 1 or 0)*
  - ***Multi-categorical*** *variables (three or more categories)*
- *Can distinguish nominal (no underlying order) vs. ordinal (categories are ordered.)*

Some Categorical Variables

- *How is your overall health? (Excellent, Very Good, Good, Fair, Poor)*
- *Which candidate would you vote for if the election were held today?*
- *Did this patient receive this procedure?*
- *If you needed to analyze a small data set right away, which of the following software tools would you be comfortable using to accomplish that task?*

Are these quantitative or categorical?

1. *Do you **smoke**? (1 = Non-, 2 = Former, 3 = Smoker)*
2. *How much did you pay for your most recent **haircut**? (in $)*
3. *What is your favorite **color**?*
4. *How many hours did you **sleep** last night?*
5. *Statistical thinking in your future **career**? (1 = Not at all important to 7 = Extremely important)*

- *If quantitative, are they discrete or continuous? Do they have a meaningful zero point?*
- *If categorical, how many categories? Nominal or ordinal?*

Data Dictionary

name	description
`species`	Penguin species: chinstrap, gentoo, adelie
`island`	Island where penguin was observed
`bill_length_mm`	how long is the bill from base to tip
`bill_depth_mm`	how wide is the bill from bottom to top
`flipper_length_mm`	length of flipper
`body_mass_g`	body mass
`sex`	male or female
`year`	2007, 2008, 2009

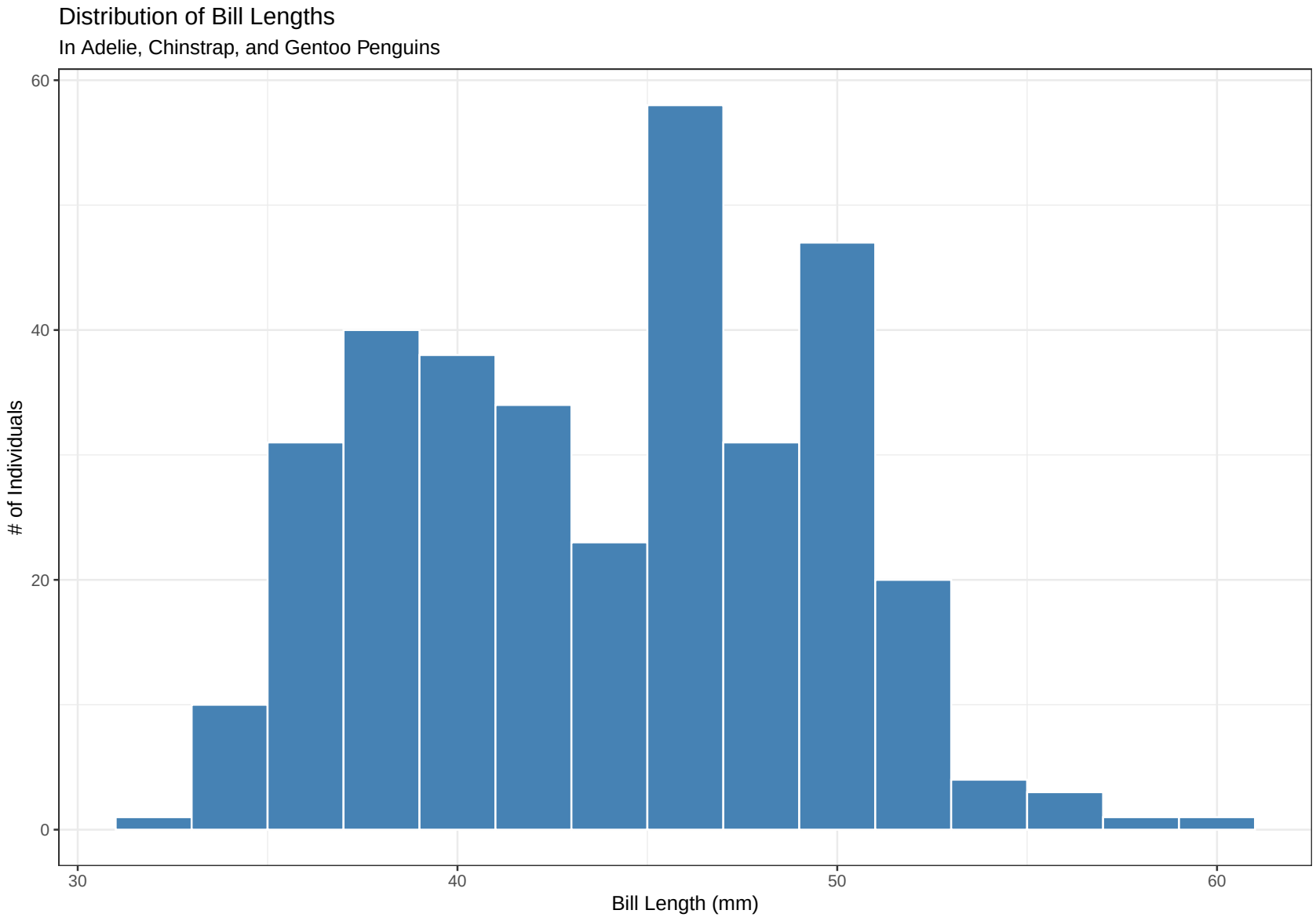
How do you visualize variables?

- *Histogram (bar plot)*
- *Density, violin plot*
- *Boxplot*

Histogram

```
penguins |>
  ggplot(aes(x = bill_length_mm)) +
  geom_histogram(binwidth = 2,
                 fill = 'steelblue',
                 col = 'white') +
  labs(title = 'Distribution of Bill Lengths',
       subtitle = 'In Adelie, Chinstrap, and Gentoo Penguin
       x = 'Bill Length (mm)',
       y = '# of Individuals')
```

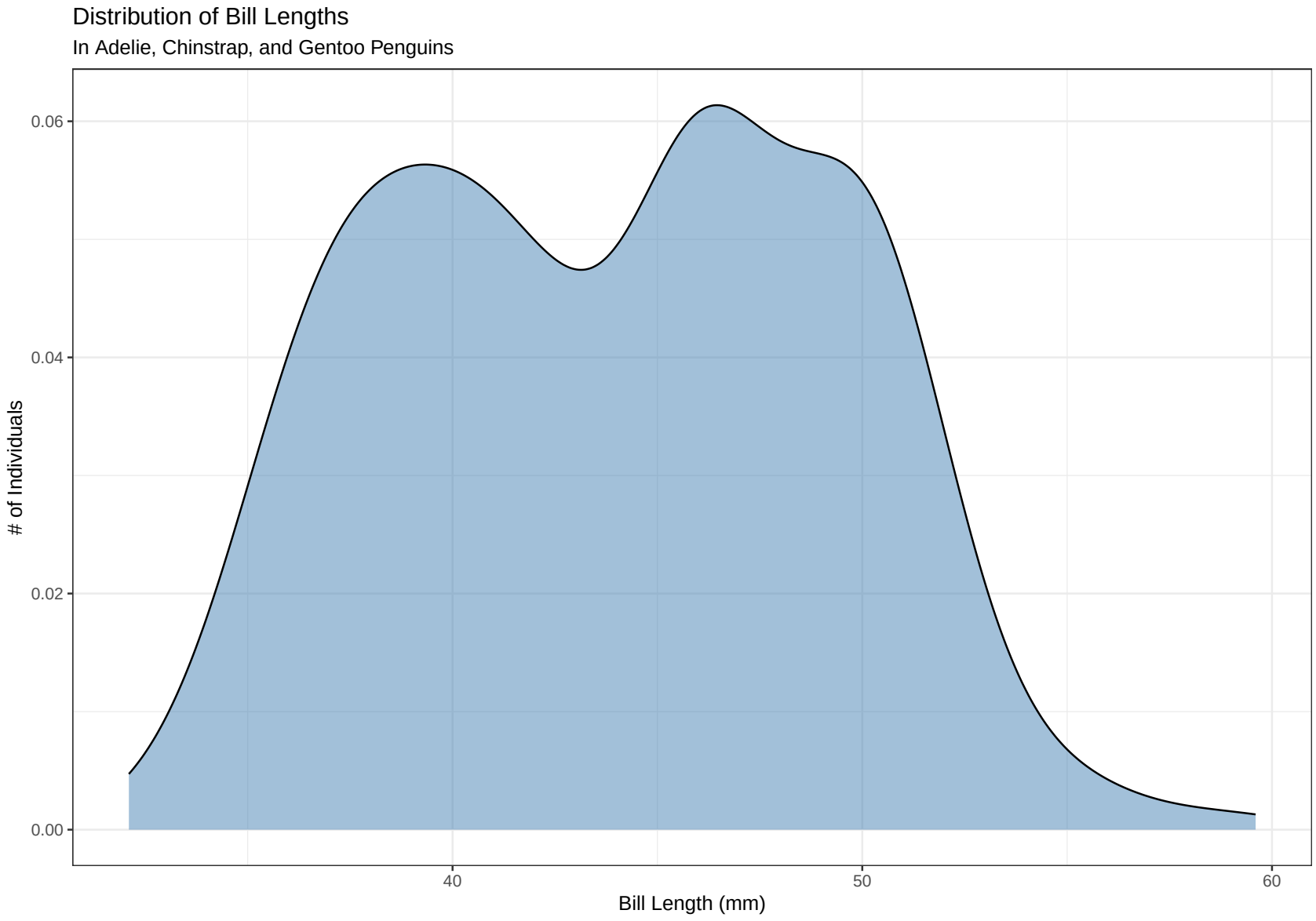
Histogram


Distribution of Bill Lengths
In Adelie, Chinstrap, and Gentoo Penguins
60
40
20
0
# of Individuals
30
40
50
60
Bill Length (mm)

Density Plot

```
penguins |>
  ggplot(aes(x = bill_length_mm)) +
  geom_density(fill = 'steelblue', alpha = 0.5) +
  labs(title = 'Distribution of Bill Lengths',
       subtitle = 'In Adelie, Chinstrap, and Gentoo Penguin
       x = 'Bill Length (mm)',
       y = '# of Individuals')
```

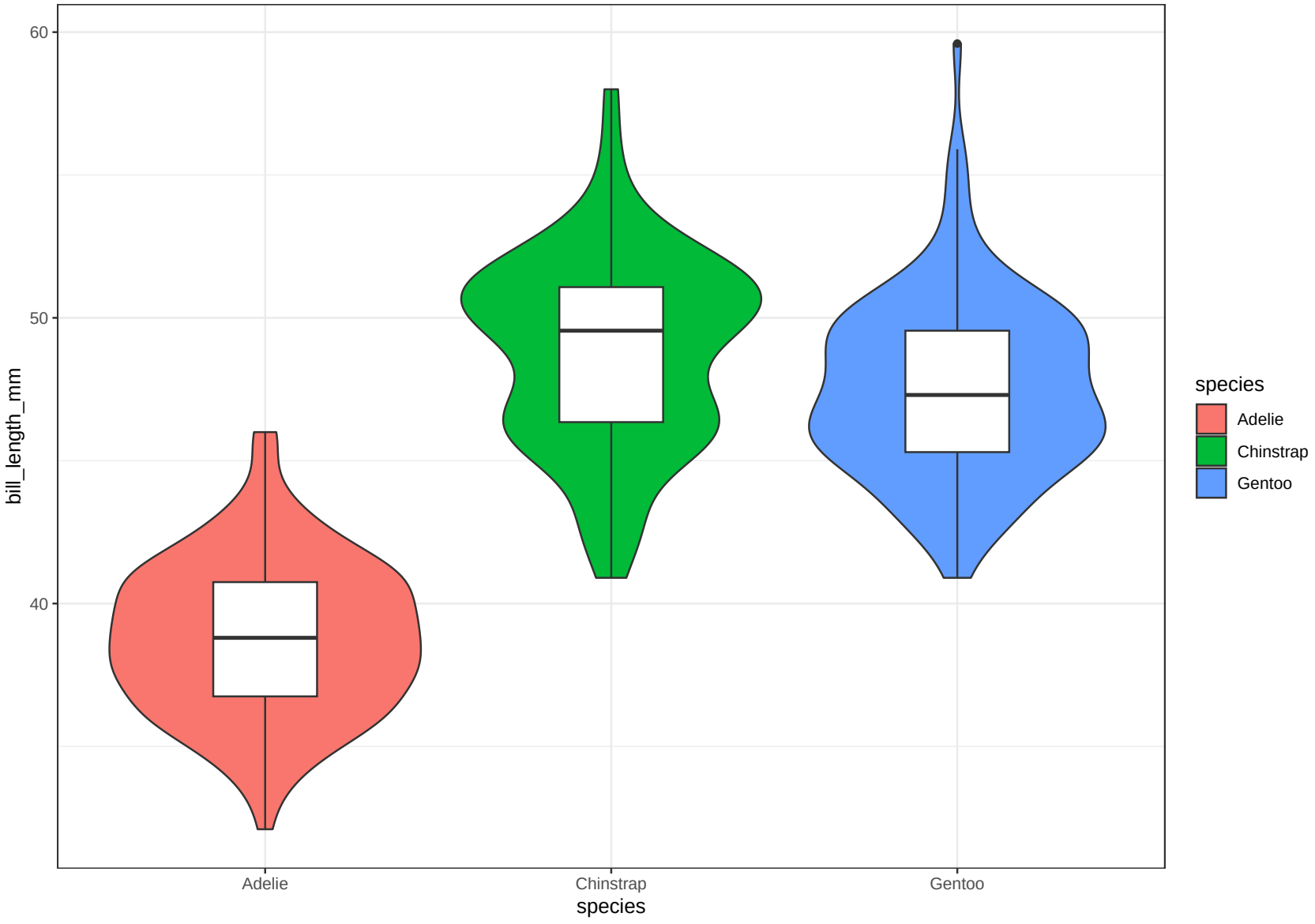
Density Plot



Boxplot + Violin

```
penguins |>
  ggplot(aes(x = species, y = bill_length_mm)) +
  geom_violin(aes(fill = species)) +
  geom_boxplot(width = 0.3)
```

Boxplot + Violin



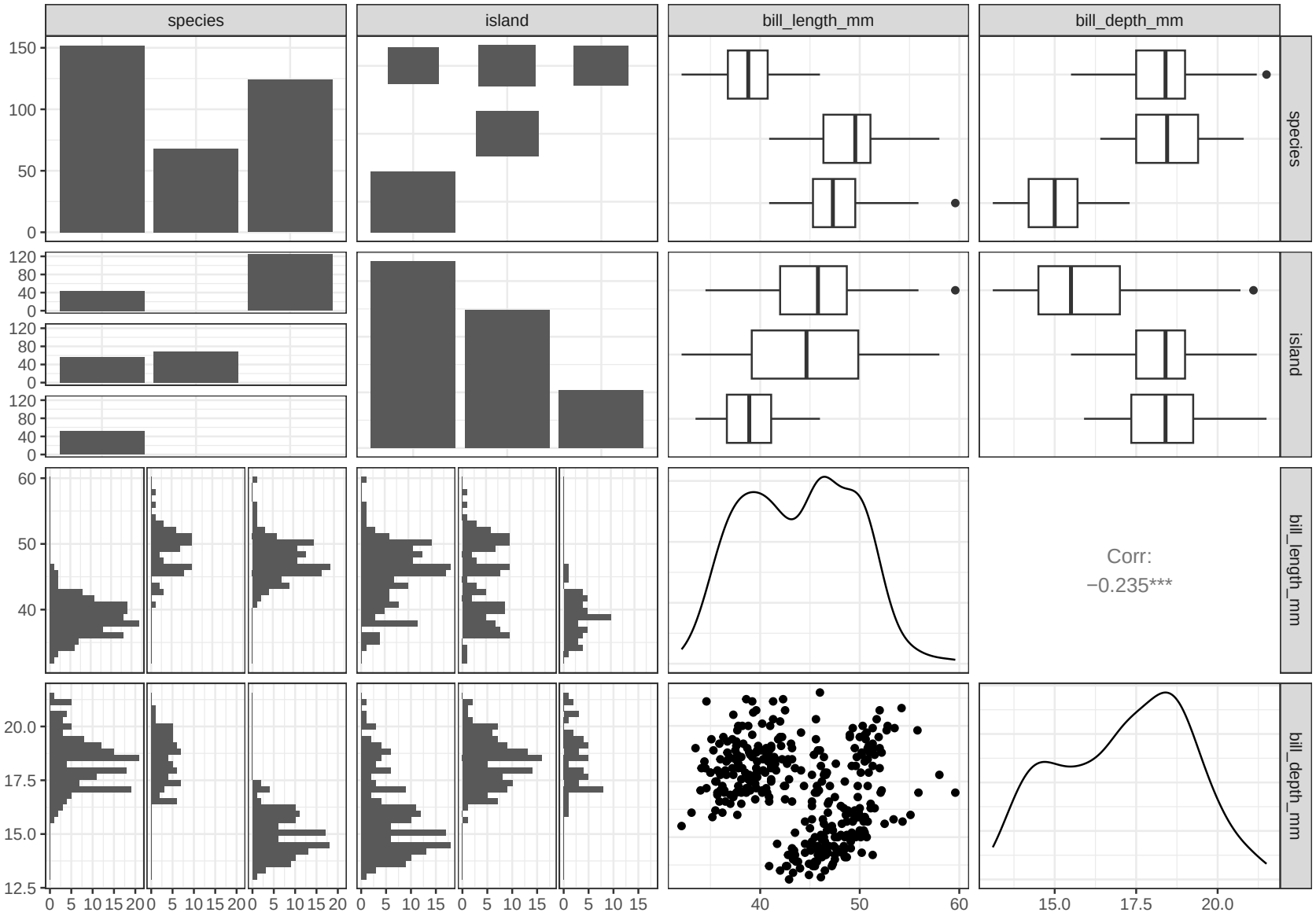
How do you find relationships between variables?

- *Develop a research question*
- *Examine summary statistics*
- *Data exploration*

Visualizing the Data - Scatterplot Matrices

```
penguins |>
  select(species, island, bill_length_mm,
         bill_depth_mm) |>
  ggpairs()
```

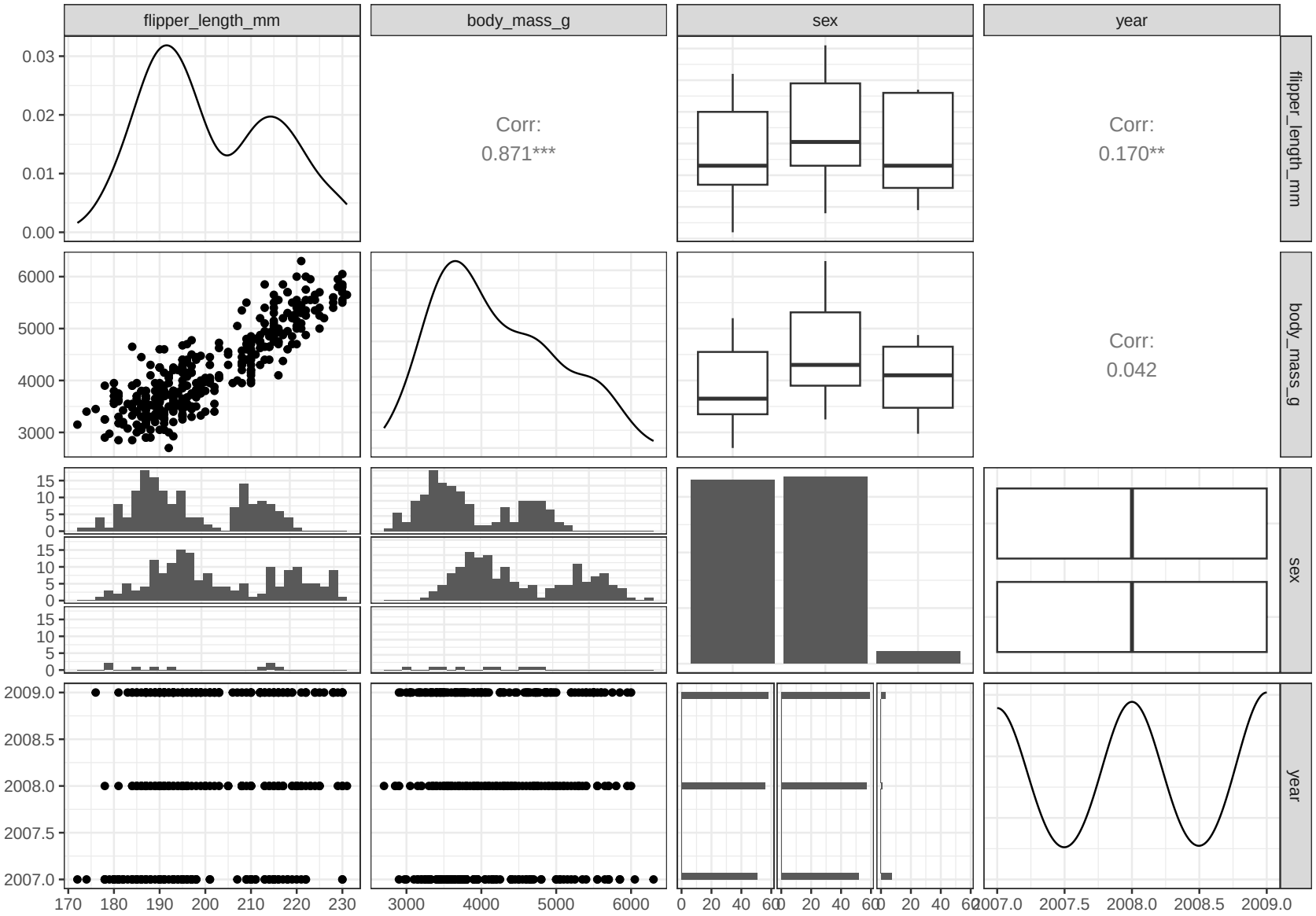
Visualizing the Data - Scatterplot Matrices



Visualizing the Data - Scatterplot Matrices

```
penguins |>
  select(flipper_length_mm, body_mass_g,
         sex, year) |>
  ggpairs()
```

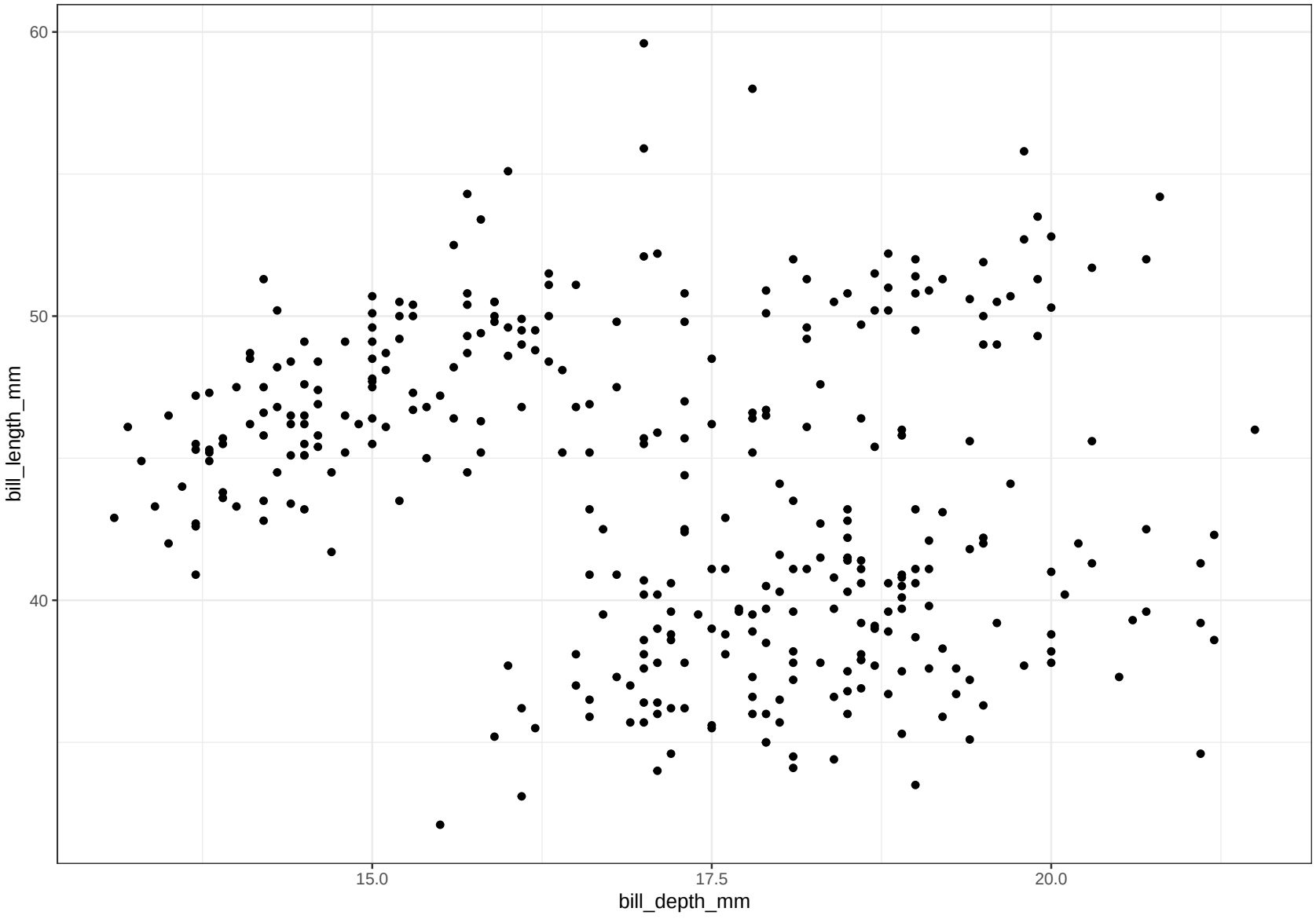
Visualizing the Data - Scatterplot Matrices



Plotting Relationships

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

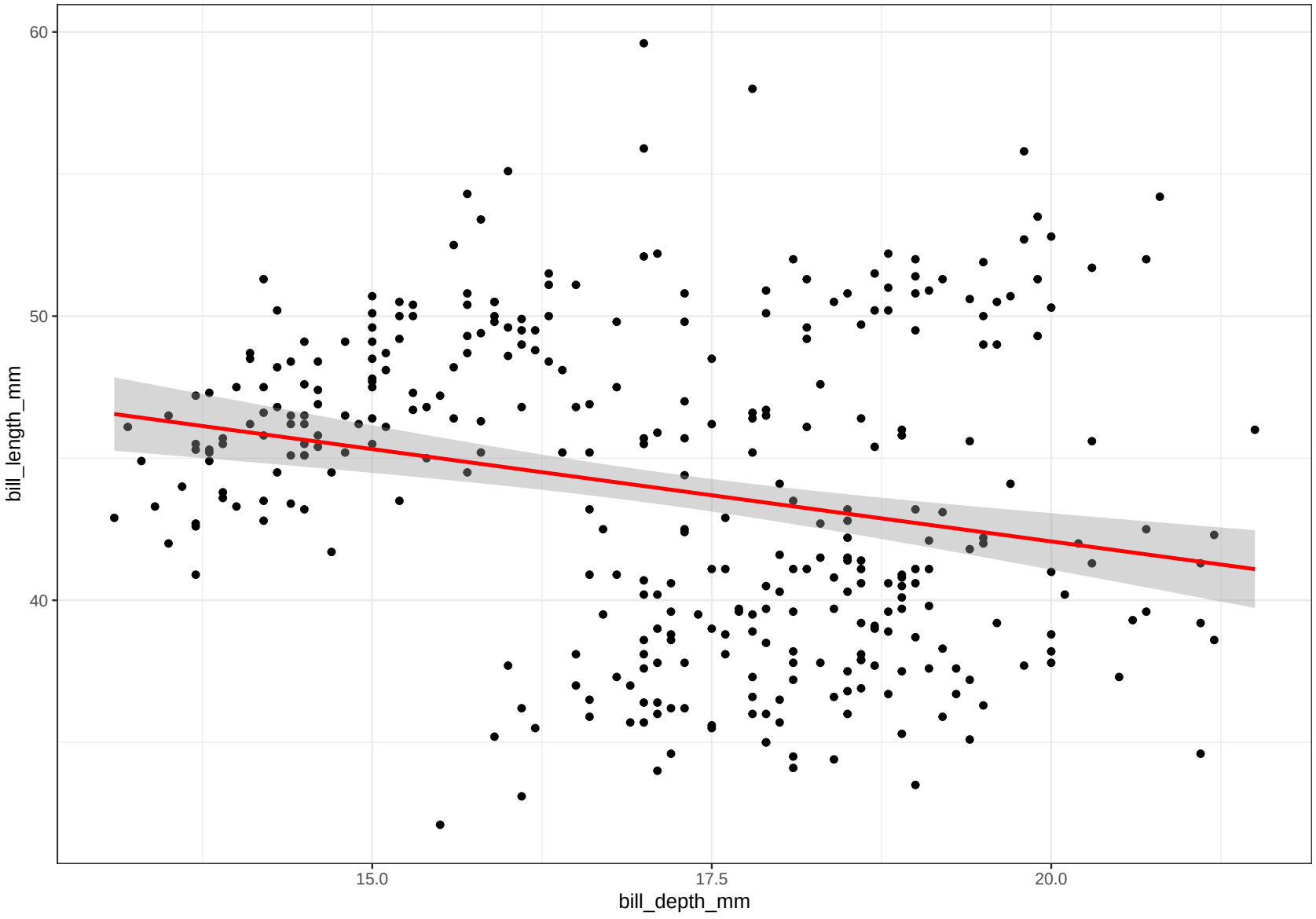
Plotting Relationships



Adding Regression Lines - LM

```
penguins |>
  ggplot(aes(x = bill_depth_mm,
             y = bill_length_mm)) +
  geom_point() +
  geom_smooth(method = 'lm', col = 'red')
```

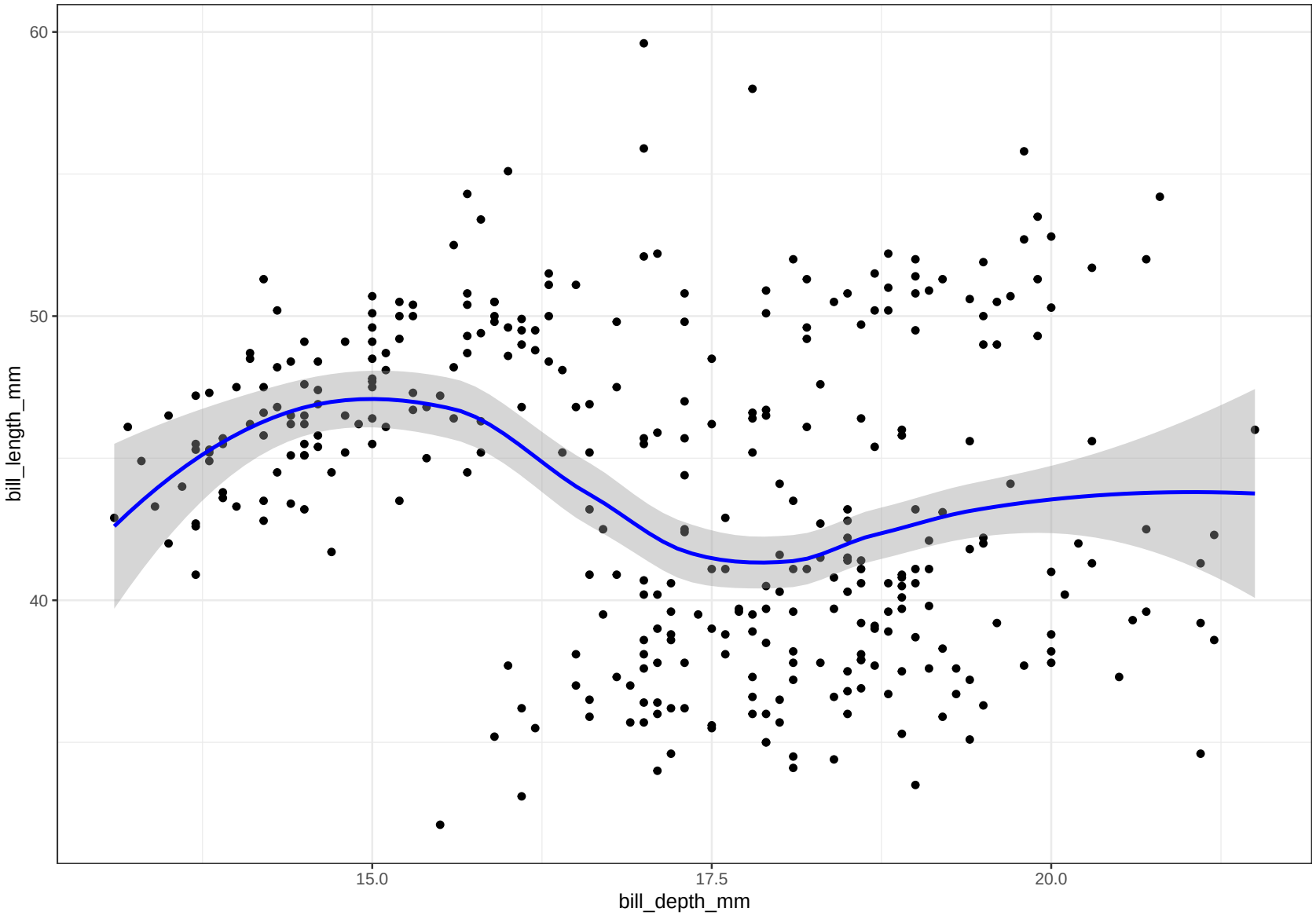
Adding Regression Lines - LM



Plotting Regression Lines - LOESS

```
penguins |>
  ggplot(aes(x = bill_depth_mm,
             y = bill_length_mm)) +
  geom_point() +
  geom_smooth(method = 'loess', col = 'blue')
```

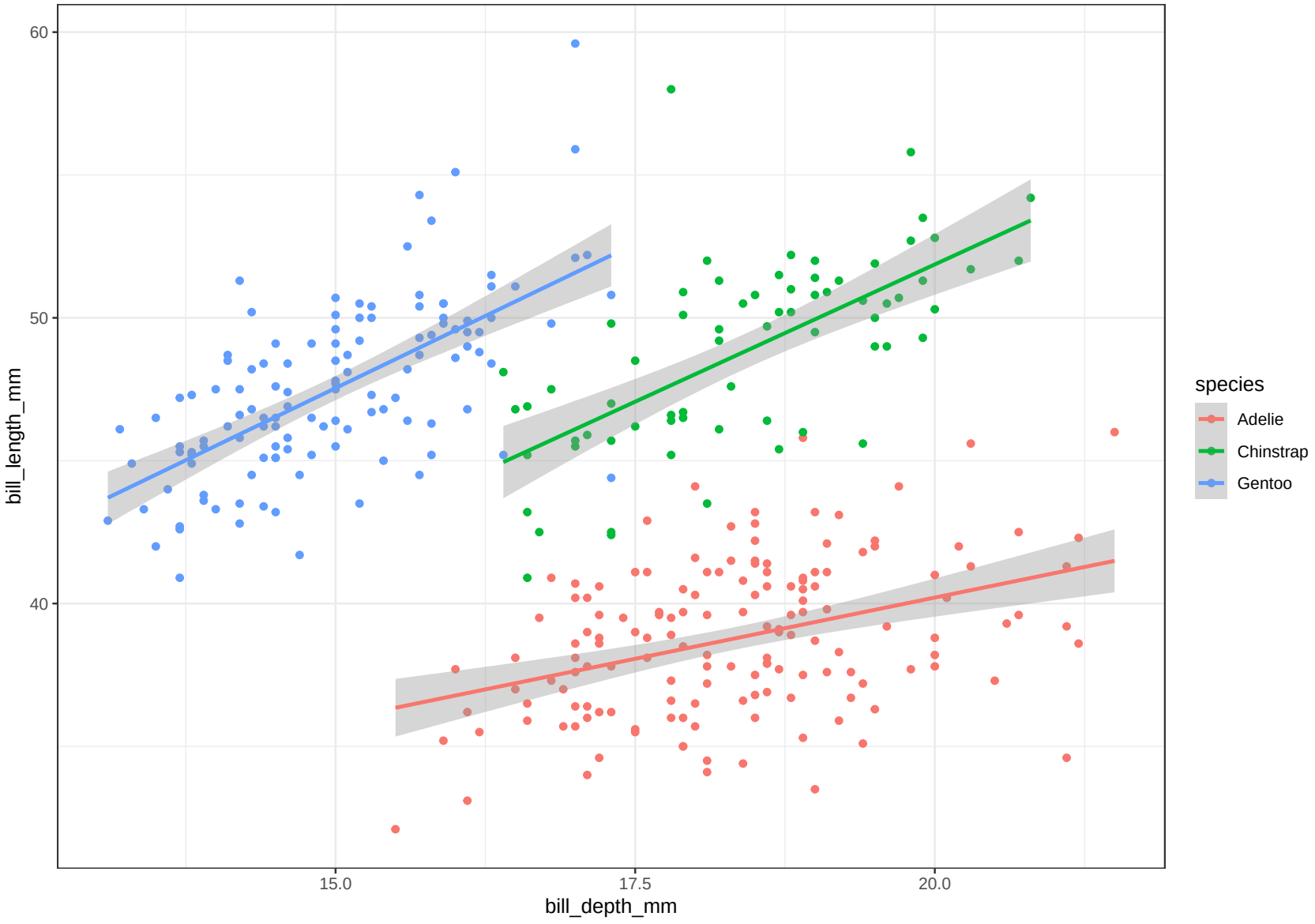
Plotting Regression Lines - LOESS



Plotting Regression Lines - By Group

```
penguins |>
  ggplot(aes(x = bill_depth_mm,
             y = bill_length_mm,
             col = species, group = species)) +
  geom_point() +
  geom_smooth(method = 'lm')
```

Plotting Regression Lines - By Group



Quarto Resources

- How to use Quarto in RStudio:
  https://quarto.org/docs/get-started/hello/rstudio.html
- Markdown language basics:
  https://quarto.org/docs/authoring/markdown-basics.html
- Themes for projects:
  https://quarto.org/docs/output-formats/html-themes.html