Class 04

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-09-06

Load Packages for Today's Slides

```
# Contains the describe() function for comprehensive data s
library(Hmisc)
# Contains the palmer penguins_df dataset
library(palmerpenguins)
# For scatterplot matrices
library(GGally)
# Always load the tidyverse last
library(tidyverse)

# Set favorite ggplot2 theme for visualizations
theme_set(theme_bw())
```

Types of Data

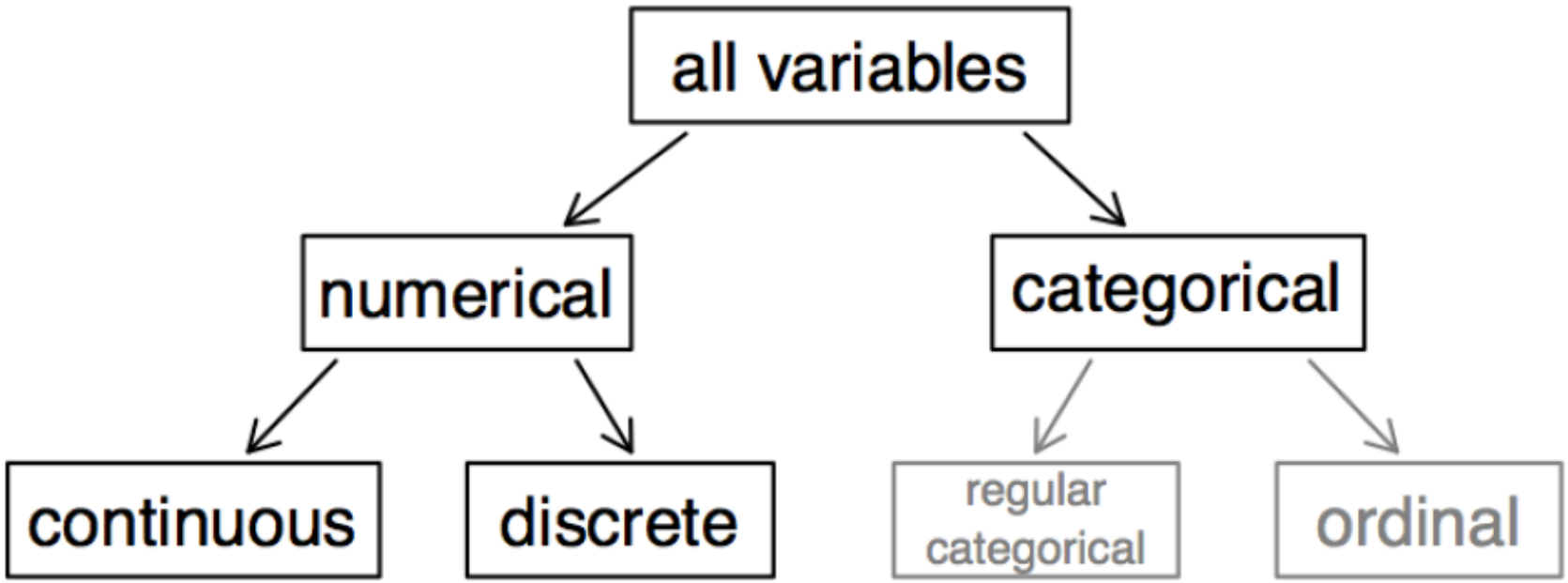


Figure 1: The two primary types of data we'll analyze are numerical and categorical variables.

How can I tell what kind of variable I have?

- Inspect the data in the global environment
- Print the data to the console or in a code chunk
- Use the `glimpse()` function for a quick summary
- Use the `describe()` function from the `Hmisc` package for a detailed summary

Let's work with palmerpenguins!

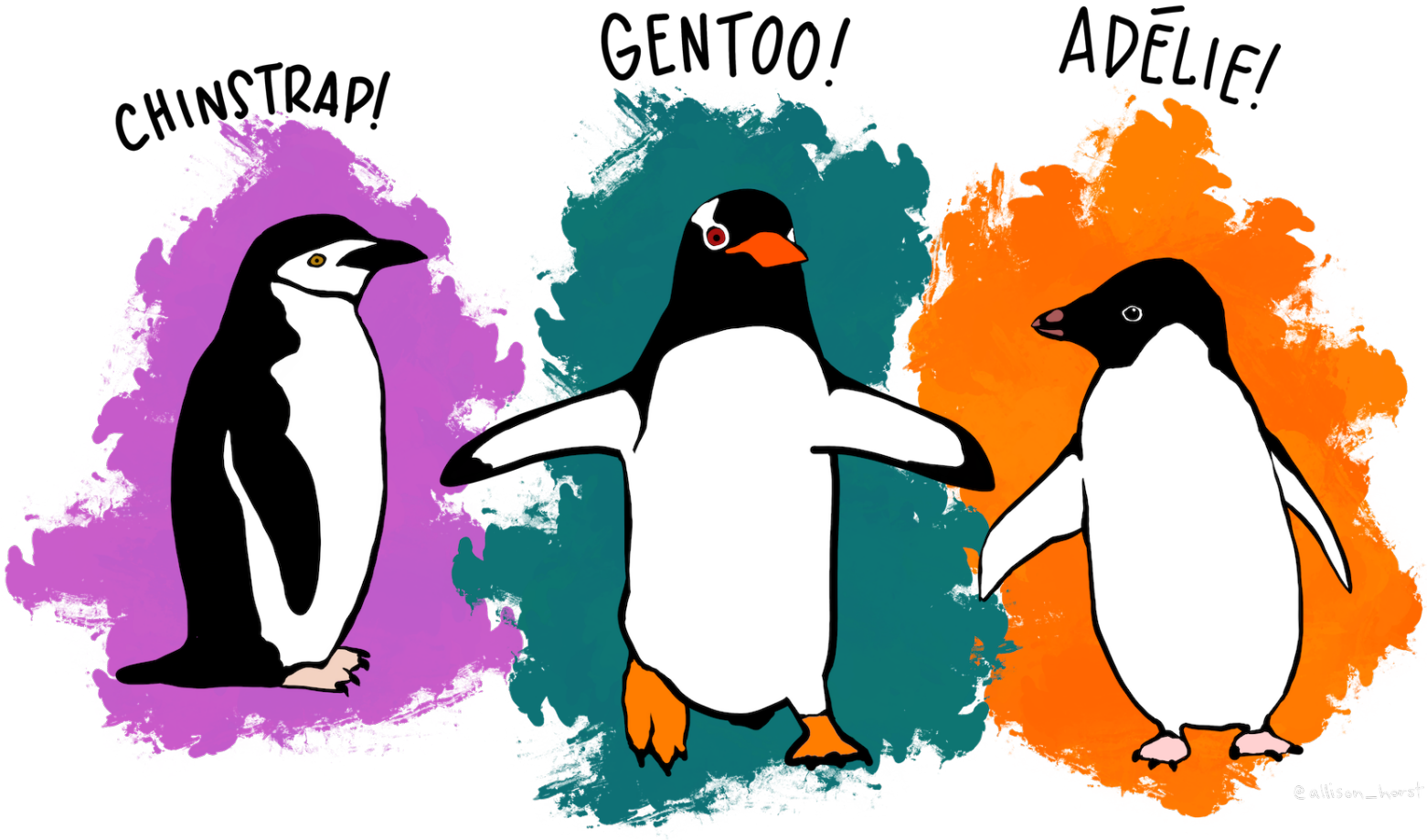


Figure 2: Meet the Palmer Penguins

Load the palmerpenguins::penguins_dfdata set

```r
penguins_df <- palmerpenguins::penguins

head(penguins_df)
```

```
# 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>
```

Data Dictionary

name	description	variable type
`species`	Penguin species: chinstrap, gentoo, adelie	categorical
`bill_length_mm`	how long is the bill from base to tip	numerical, continuous
`bill_depth_mm`	how wide is the bill from bottom to top	numerical

Use Hmisc::describe() for a detailed summary

```
# describe is a common function name, so
# it is a good habit to call this version
# directly from the package using package_name::
# to prevent conflicts and errors
penguins_df |>
  select(species, bill_length_mm, bill_depth_mm) |>
  Hmisc::describe()
```

Use Hmisc::describe() for a detailed summary

```
select(penguins_df, species, bill_length_mm, bill_depth_mm)

 3  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
--------------------------------------------------------------------
bill_length_mm
       n  missing distinct     Info     Mean      Gmd
     342        2      164        1    43.92    6.274     35
     .25      .50      .75      .90      .95
   39.23    44.45    48.50    50.80    51.99
```

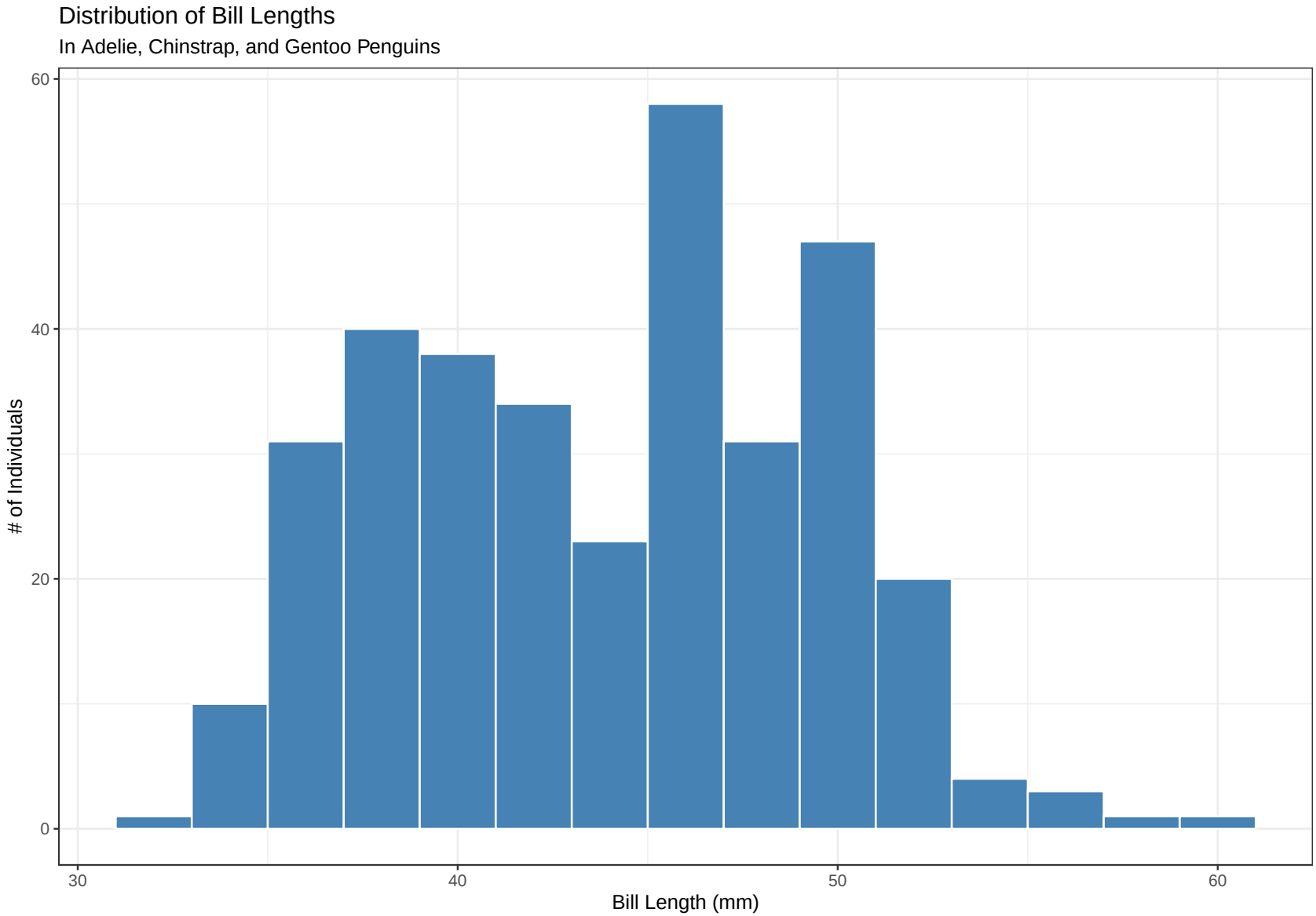
Visualizing Data

- Distributions and numerical summaries of both explanatory and response variables
  - Histogram, bar plot
  - Density or violin plots
  - Boxplots
- Associations, relationships, correlations between explanatory and response variables
  - Scatter plots, regression
  - Scatterplot matrices

Histogram - How common are certain ranges of values? (Discrete)

```
# Pipe data into ggplot2
penguins_df |>
  # Initialize the plot parameters with aes
  ggplot(aes(x = bill_length_mm)) + # ggplot2 only uses +!
  # add a histogram to the plot
  geom_histogram(binwidth = 2, # each bin spans 2 mm
                 fill = 'steelblue', # some color for fun
                 color = 'white') +
  # Add titles and axis labels
  labs(title = 'Distribution of Bill Lengths',
       subtitle = 'In Adelie, Chinstrap, and Gentoo Penguin
       x = 'Bill Length (mm)',
       y = '# of Individuals')
```

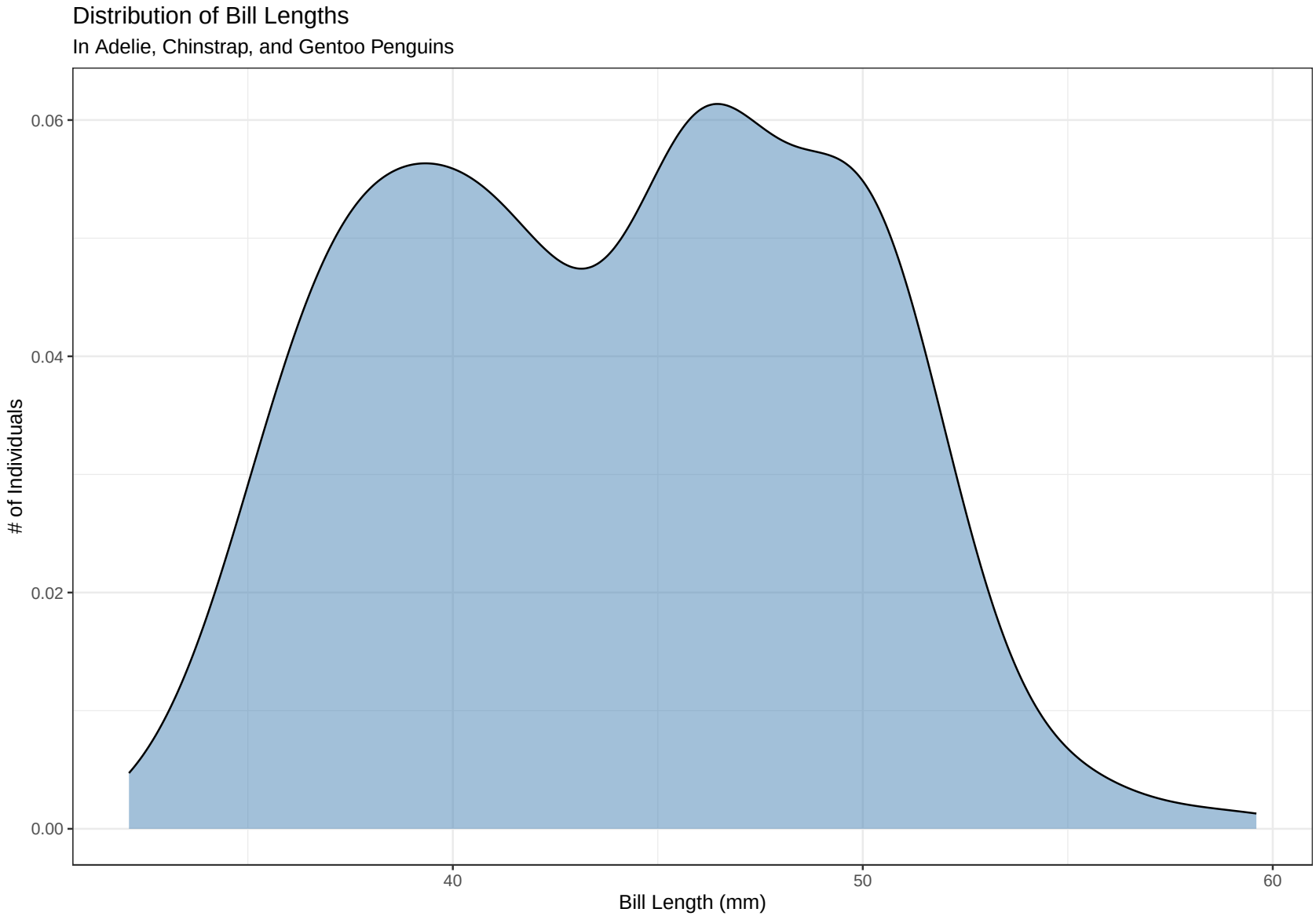
Histogram - How common are certain ranges of values? (Discrete)



Density Plot - How common are certain ranges of values? (Continuous)

```
# Pipe data into ggplot2
penguins_df |>
  # Initialize the plot parameters with aes
  ggplot(aes(x = bill_length_mm)) +
  # add a density curve to the plot
  geom_density(fill = 'steelblue', # add some color and mak
               alpha = 0.5) +
  # Add titles and axis labels
  labs(title = 'Distribution of Bill Lengths',
       subtitle = 'In Adelie, Chinstrap, and Gentoo Penguin
       x = 'Bill Length (mm)',
       y = '# of Individuals')
```

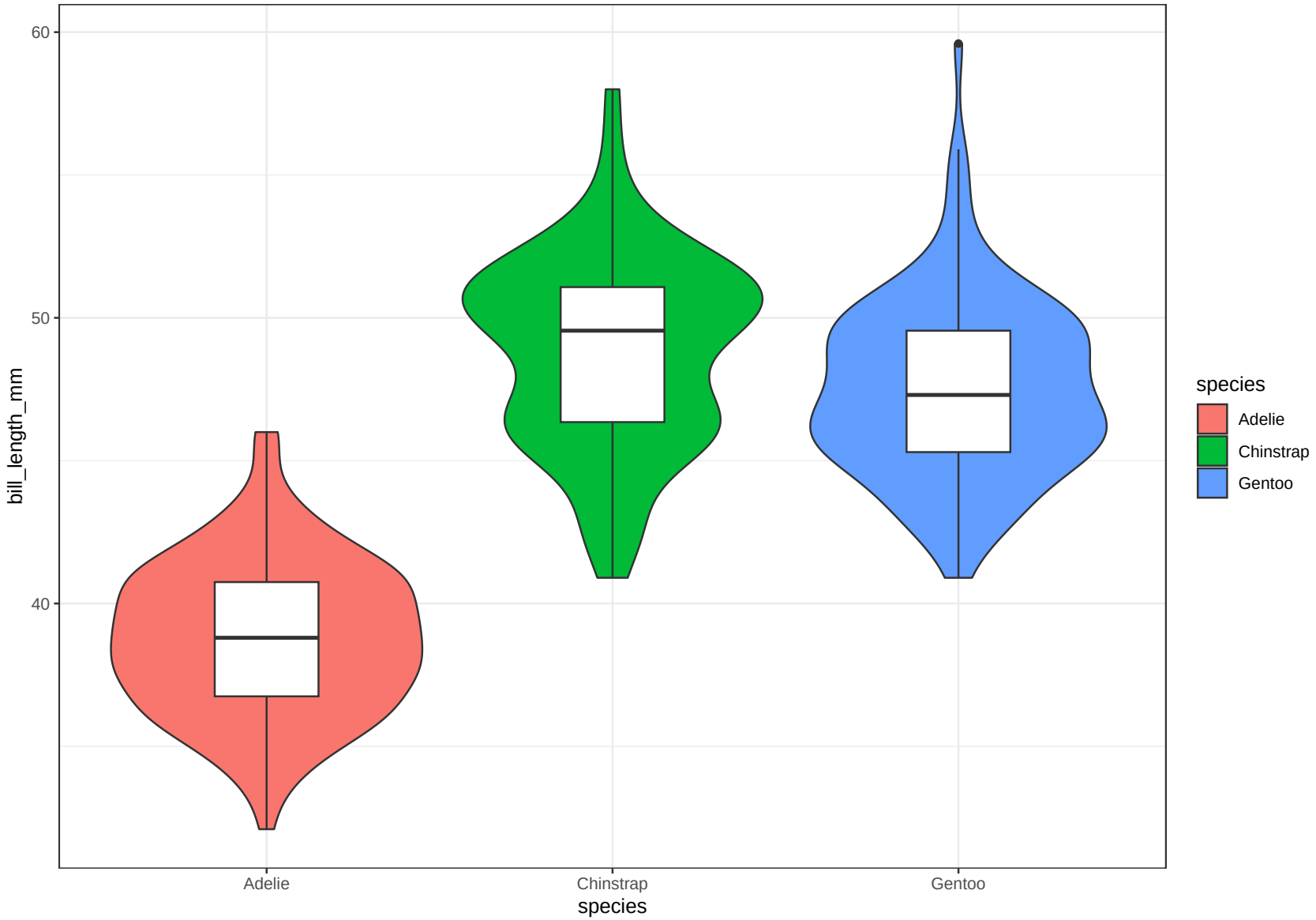
Density Plot - How common are certain ranges of values? (Continuous)



Boxplot + Violin - Numerical summary + density curves

```
# Pipe data into ggplot2
penguins_df |>
  # Initialize the plot parameters with aes
  ggplot(aes(x = species, y = bill_length_mm)) +
  # Add a violin plot as the base layer
  geom_violin(aes(fill = species)) +
  # Add a boxplot on top of the violin plot
  geom_boxplot(width = 0.3)
```

Boxplot + Violin - Numerical summary + density curves



Variable Terms

- ***Independent*** or ***explanatory*** variable
  - Typically on the x-axis
  - "Cause" variable
- ***Dependent*** or ***response*** variable
  - Typically on the y-axis
  - "Effect" variable

Association vs. Independence

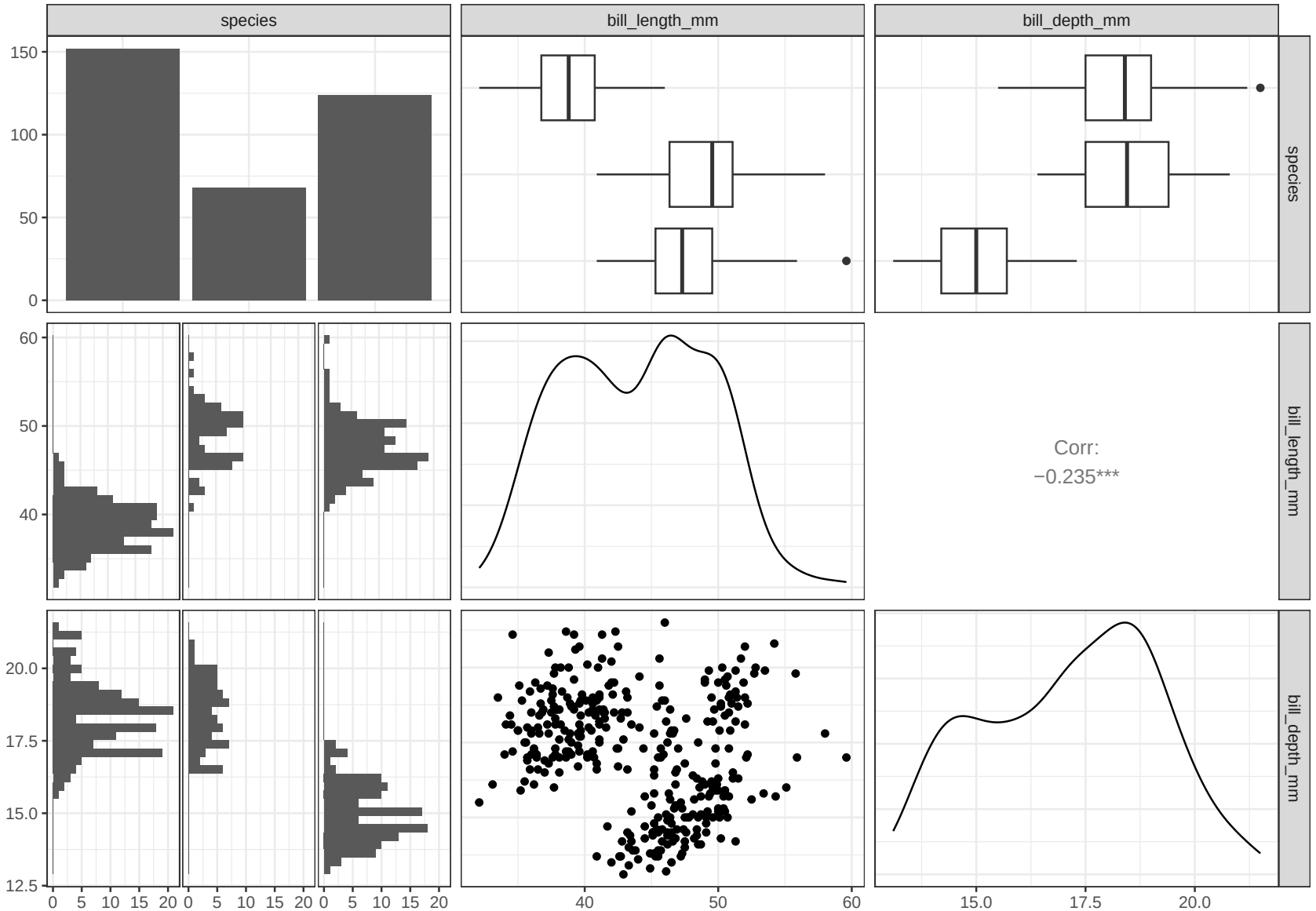
- When two variables show some connection with one another, they are called ***associated*** variables.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be ***independent***.

Scatterplot Matrices - Quick Look at Many Relationships

This code will sometimes run slowly and generate lots of warning messages.

```
penguins_df |>
# Select variables of interest
  select(species, bill_length_mm,
         bill_depth_mm) |>
  # send to ggpairs to create the matrix
  ggpairs()
```

Scatterplot Matrices - Quick Look at Many Relationships

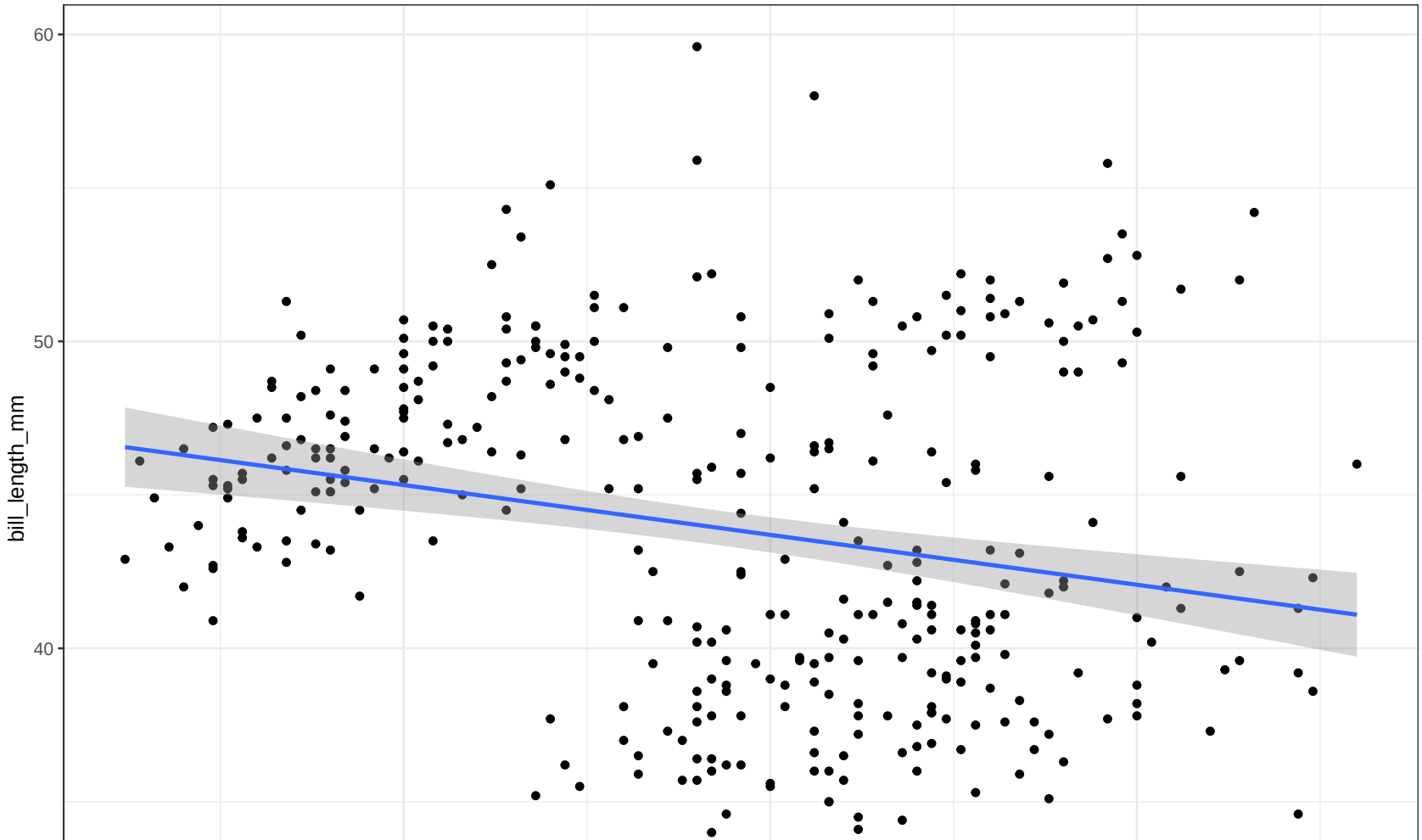


Scatter plot + Linear Regression - Detailed Look at 1 Relationship

```
# Pipe data into ggplot2
penguins_df |>
  # Set x and y variables with aes
  ggplot(aes(x = bill_depth_mm,
             y = bill_length_mm)) +
  # add a scatterplot
  geom_point() +
  # add a linear model regression line
  geom_smooth(formula = y ~ x,
              # set method to lm
              method = 'lm',
              # keep standard error shading
              se = T)
```

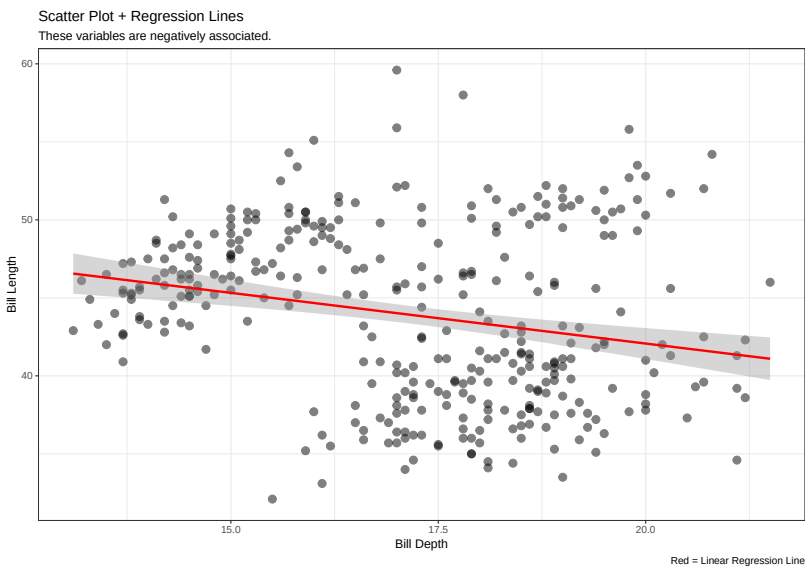
Scatter plot + Linear Regression - Detailed Look at 1 Relationship

Does this look like this regression line accurately describes the relationship between bill depth and bill length? Do you see any patterns in the points?



Negatively Correlated Variables

The regression line slopes downwards from the upper left-hand corner towards the lower right-hand corner.



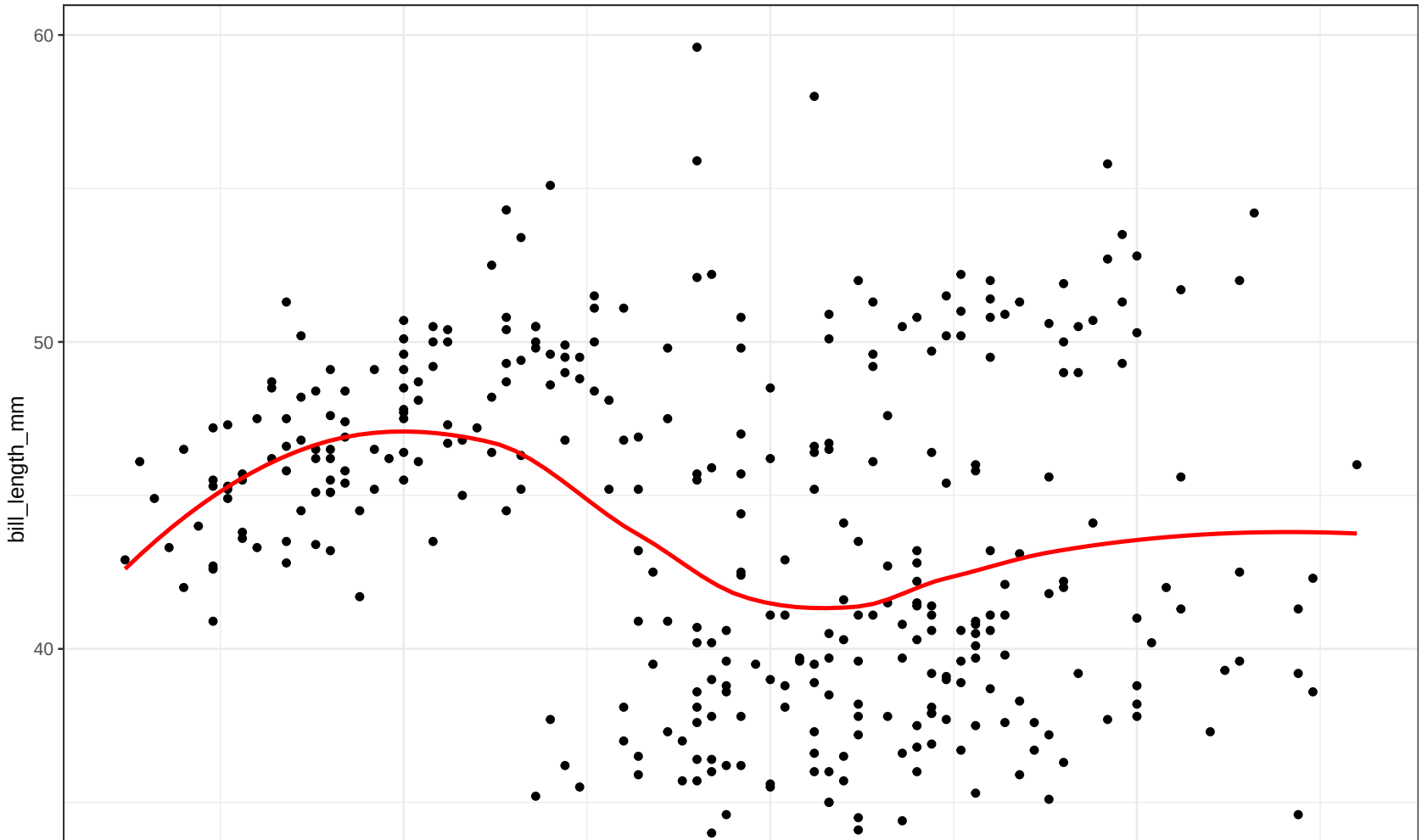
- Bill length is negatively associated with bill depth for all penguins_df sampled.
- Bill length is negatively correlated to bill depth.
- As bill depth increases, bill length decreases.

Scatter plot + LOESS Regression - Detailed Look at 1 Relationship

```
# Pipe data into ggplot2
penguins_df |>
  # Set x and y variables with aes
  ggplot(aes(x = bill_depth_mm,
             y = bill_length_mm)) +
  # add a scatterplot
  geom_point() +
  # add a LOESS regression line
  geom_smooth(formula = y ~ x,
              # set method to loess
              method = 'loess',
              # change the color
              color = 'red',
              # remove standard error shading
              se = F)
```

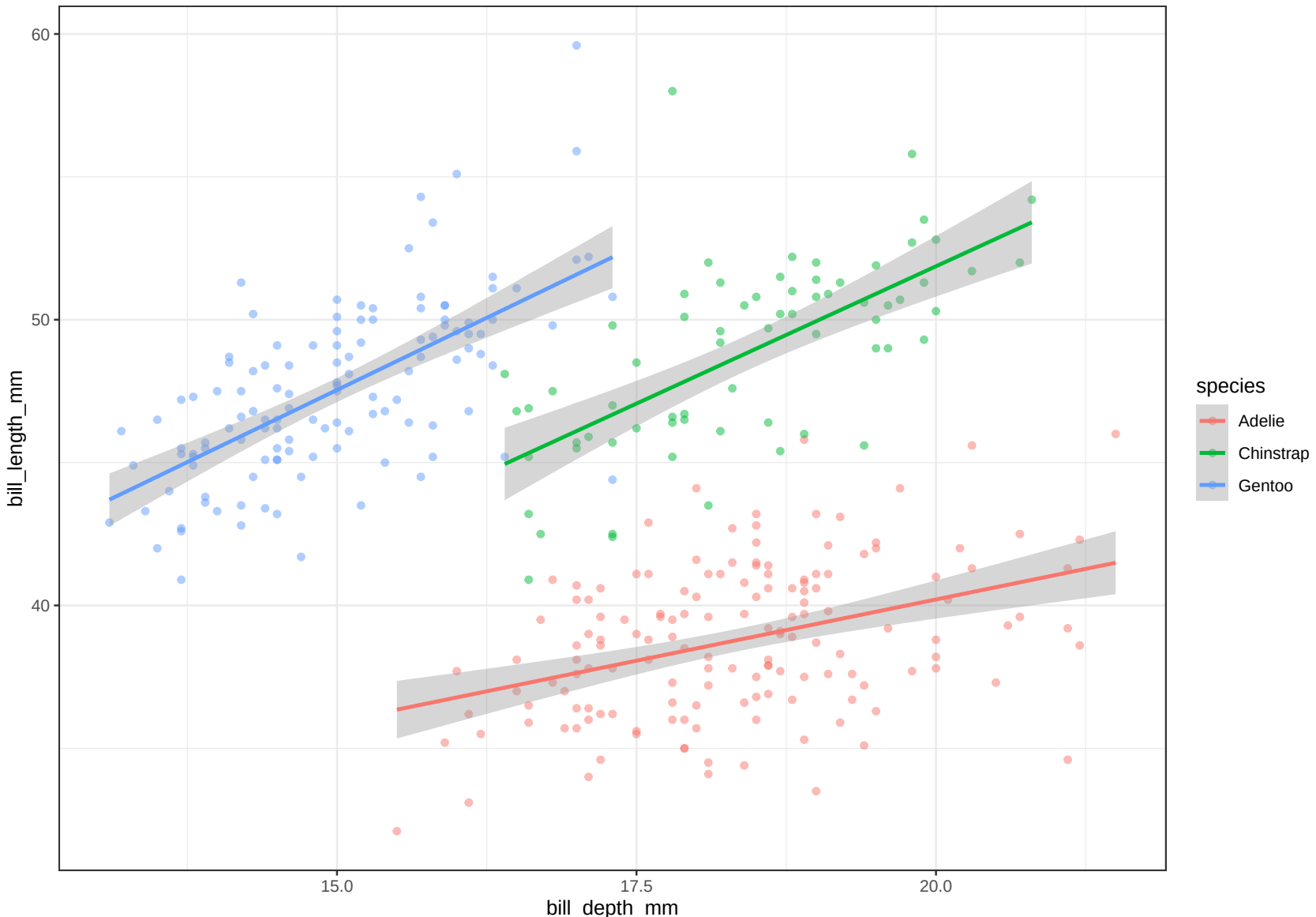
Scatter plot + LOESS Regression - Detailed Look at 1 Relationship

Using the localized regression technique LOESS can help you identify trends in your data that traditional linear models can miss. What do you see here?



What happens when we consider penguin species?

How does this plot differ from the first linear regression analysis on data from `penguins_df` not grouped by species?

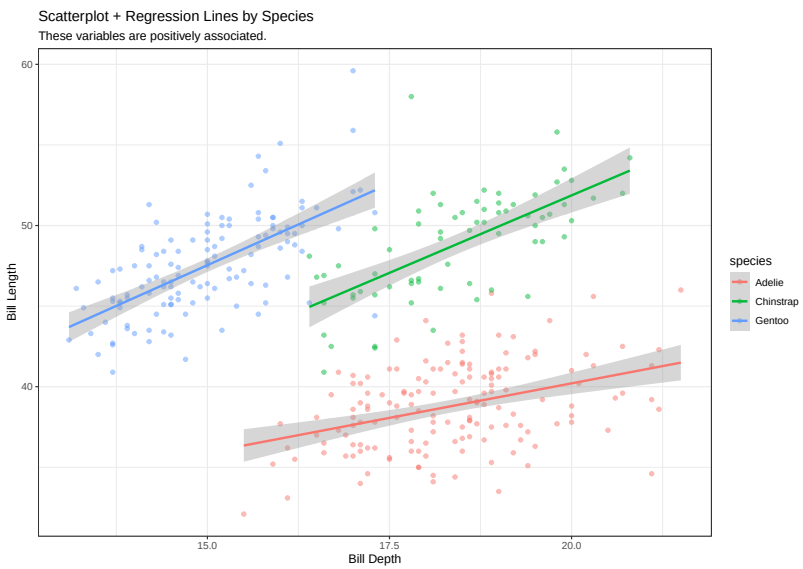


What happens when we consider penguin species?

```
# Pipe data into ggplot2
penguins_df |>
  # Set x and y variables with aes
  ggplot(aes(x = bill_depth_mm,
             y = bill_length_mm,
             # group the data by species
             group = species,
             # color the points/lines by species
             color = species)) +
  # add a scatterplot
  geom_point(alpha = 0.5) +
  # add a linear model regression line
  geom_smooth(formula = y ~ x,
              # set method to lm
              method = 'lm',
              # keep standard error shading
              se = T)
```

Positively Correlated Variables

The regression line slopes upwards from the bottom left-hand corner towards the upper right-hand corner.



- Bill length is positively associated with bill depth within each of the 3 penguin species.
- Bill length is negatively correlated to bill depth
- As bill depth increases, bill length decreases.

Take-Home Lessons

- Conclusions are shaped by the assumptions we make during the analysis
- Context is important!
- A picture is worth a thousand words

Study Approaches

- **Case Study (anecdotal evidence)**: very few observations, often $n = 1$
- **Sampling**: a subset of all possible observations
  - **Random**: a sampling technique used to obtain data from the subset that is representative of all possible observations
  - **Voluntary response**: data is volunteered by study subjects
  - **Convenience**: data is obtained from the subset of all possible observations that is most easily accessible
- **Census**: all possible observations

Study Methods

- **Observational Study**: researchers do not affect the data that is collected from subjects (ex: a political poll)
- **Interventional Study**: researchers do *something* which modifies the data collected from subjects (ex: a clinical trial that assigns subjects to control and treatment groups)
- **Prospective Study**: subjects are recruited into a study prior to future data collections (ex: a 5-year study of crime rates in juveniles following expulsion from school)
- **Retrospective Study**: data is collected from subjects on events that have already taken place (ex: a study of the past medical records of hospital patients who were diagnosed with a disease)

Study Population Definitions

- **Study Population**: all possible subjects who could have been observed in the study
- **Sample Population**: the subjects who were actually observed in the study
- **Target Population**: the subjects who the research conclusions should be applied to

Evaluating a Study's Data

- **Reliability**: how much can we trust the data? is it accurate?
- **Validity**: how well does the sample population represent the study population?
- **Generalizability**: would the conclusions from the study population be applicable to the target population?

Populations & Sampling: Example 1

- *Grambeau, K., Osborne, B. Weight, E.A. (2020). Student-Athlete Perceptions of Name, Image, and Likeness Compensation. Chapel Hill, NC: Center for Research in Intercollegiate Athletics.*
- **Who:** The Center for Research in Intercollegiate Athletics at the University of North Carolina at Chapel Hill
- **Research Question**: In light of impending NCAA rule changes, are NCAA student-athletes in the Power Four conferences in favor of receiving compensation in exchange for the use of their name, image, and likeness (NIL)?

Populations & Sampling: Example 1

- **Research Question**: In light of impending NCAA rule changes, are NCAA Division I student-athletes in favor of receiving compensation in exchange for the use of their name, image, and likeness (NIL)?
- **Methods**: Surveys from $n = 1201$ current student-athletes at an NCAA Division I Power Four conference school were analyzed.

1. What is the target population?
   all NCAA Division I student-athletes
2. What is the study population?
   Current student-athletes at an NCAA Division I Power Four conference school
3. What is the sample population?
   1,201 current student-athletes at an NCAA Division I Power Four conference school who responded to the survey

Populations & Sampling: Example 1

- **Research Question**: In light of impending NCAA rule changes, are NCAA Division I student-athletes in favor of receiving compensation in exchange for the use of their name, image, and likeness (NIL)?
- **Methods**: Surveys from $n = 1201$ current student-athletes at an NCAA Division I Power Four conference school were analyzed.

1. Is the data reliable?
   self-reported data may not be honest
2. Is the data valid?
   sampling technique not reported (random? convenience?)
   response rate not reported (non-responders excluded)
3. Is the data generalizable?
   are Power Four student-athletes representative of all Division I student-athletes?

Populations and Sampling: Example 2

- *Harvey, S. B., Øverland, S., Hatch, S. L., Wessely, S., Mykletun, A., & Hotopf, M. (2018). Exercise and the Prevention of Depression: Results of the HUNT Cohort Study. American Journal of Psychiatry, 175(1), 28–36. https://doi.org/10.1176/appi.ajp.2017.16111223*
- **Research Question**: Does exercise provide protection against new-onset depression and anxiety?
- **Methods**: Baseline and follow-up surveys about lifestyle and medical history were sent to all inhabitants aged 20+ years of Nord-Trondelag County in Norway ($n = 85100$) in 1984-1986 (baseline) and 1995-1997 (follow-up). 60,980 subjects responded, from which 33,908 subjects were selected due to having no pre-existing physical or mental health conditions.

Populations and Sampling: Example 2

- **Research Question**: Does exercise provide protection against new-onset depression and anxiety?
- **Methods**: Baseline and follow-up surveys about lifestyle and medical history were sent to all inhabitants aged 20+ years in rural Nord-Trondelag County in Norway ($n = 85100$) in 1984-1986 (baseline) and 1995-1997 (follow-up). 60,980 subjects responded, from which 33,908 subjects were selected due to having no pre-existing physical or mental health conditions.

1. What is the target population?
   all Norwegian adults? all white adults? all adults?
2. What is the study population?
   Norwegian adults aged 20+ years from Nord-Trondelag County from 1984-1997
3. What is the sample population?
   33,908 "healthy" Norwegian adults aged 20+ years from Nord-Trondelag County who responded to both surveys

Populations & Sampling: Example 2

- **Research Question**: Does exercise provide protection against new-onset depression and anxiety?
- **Methods**: Baseline and follow-up surveys about lifestyle and medical history were sent to all inhabitants aged 20+ years in rural Nord-Trondelag County in Norway ($n = 85100$) in 1984-1986 (baseline) and 1995-1997 (follow-up). 60,980 subjects (72%) responded, from which 33,908 subjects were selected due to having no pre-existing physical or

1. Is the data reliable?
   self-reported data may not be honest
2. Is the data valid?
   high response rate
   "unhealthy" subjects excluded
3. Is the data generalizable?
   Are residents of rural Nord-Trondelag county representative of all Norwegians? adults who live in the US? adults who live in China?

Sampling with Skittles

You're a researcher with the FDA investigating the safety of food additives. Some skittles are made with Red 40, a food dye. This additive can harm the body in large doses, so you want to make sure people who eat Skittles don't get too much. You need to know on average how many red skittles come in a package to determine if levels are below the legal limit.

Research Question

On average, how many red Skittles are in the a package?

What is being measured?

Count of red skittles in a package

Sampling Strategy

- Where will we get Skittles to study from?
- How many packages should we count?

Target Population

- To what or whom will the conclusions of our study apply?
- How does that affect data collection?

Population Definitions

- What is the study population?
- What is the sample population?

Evaluating Inferences

- How reliable would data collected this way be?
- How valid would data collected this way be?
- To whom/what is the data generalizable?

Session Info

```
xfun::session_info()
```

```
R version 4.4.1 (2024-06-14 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 22631)

Locale:
  LC_COLLATE=English_United States.utf8
  LC_CTYPE=English_United States.utf8
  LC_MONETARY=English_United States.utf8
  LC_NUMERIC=C
  LC_TIME=English_United States.utf8

Package version:
  askpass_1.2.0          backports_1.5.0       base64enc_0.1-3
  bit_4.0.5              bit64_4.0.5           blob_1.2.4
  broom_1.0.6            broom.helpers_1.17.0  bslib_0.8.0
  cachem_1.1.0           callr_3.7.6           cards_0.2.2
```