Center for Analytics and Data Science

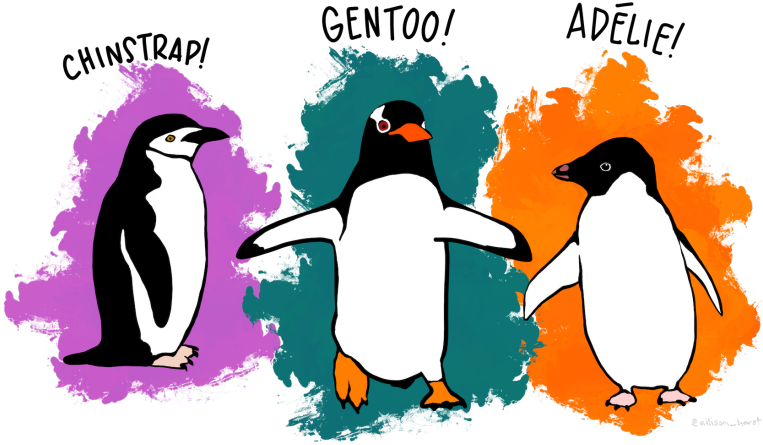
R Review Workshop

Exploring the Palmer Penguins Data

Bentley University

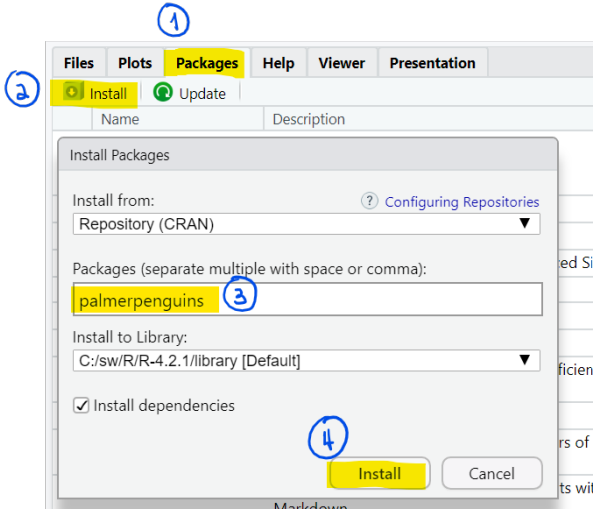
Fall 2022

Meet the Penguins



Source: Artwork by @allison_horst
https://allisonhorst.github.io/palmerpenguins/

Install the Package



Load some Packages and Access the Data

Load the `tidyverse` package into your environment so we have some functions to work and visualize the data.

```
> library(tidyverse)
```

Load the `palmerpenguins` package into your environment and store the penguins data into an object. I like to call it `dta`; a short version of 'data' to save some typing.

```
> library(palmerpenguins)
```

```
> dta <- penguins
```

In your environment you should see an object named `dta` with 344 observations and 8 variables.

The rest of the document asks some questions about the data and suggests some functions you can use to answer those questions. Do not feel obliged to use the suggested functions. There are many ways to implement the computations needed to explore the data.

What are we working with?

1. How many observations and how many features (variables) do we have in the data set? **dim()** or **nrow()** and **ncol()**
2. What are the names of the variables in the data set? **names()**
3. Are all the variables in the data set numeric or character? The function **str()** will give you the **structure** of the data set and a few observations for each variable.
4. Print out the first few rows of the data set. **head()**
5. Print the last 15 rows of the data set. **tail()**

What kinds of penguins do we have?

1. How many different species of penguins do we have? **unique()** and **$**
2. How many observations do we have of each type of penguin? **table()** or **xtabs()**
3. I have heard that some penguins do not like some of the islands, but other penguins like all of the islands. Is that true? Create a table where the rows are the penguin species and the columns are the islands. The entries in the table should be the number of penguins. **table()** or **xtabs()**
4. Do the sum of all your table entries match the total number of observations in your data set? **sum()**
5. Are there any missing values in the body mass variable? **any()**, **is.na()**
6. Which observations are those? **which()**
7. Display those records with missing body mass. **[]**
8. Remove the rows of data with a missing weight and store the data in a new variable.

Let's look at penguins weight

1. What is the average weight (in kilograms) across all penguins? **mean()**
2. Which penguins weigh between 6,000 and 6,500 grams? Create an index variable, called `idx`, that is true if the weight is in this range and false otherwise. How many penguins fall in this weight range?
3. Display the penguins in this range.
4. I have heard that all three types of penguins are about the same size. Is it true that average weight for all three penguins is the same? **group_by()**, **summarize()**, **%>%**
5. What is the one standard deviation interval around the mean weight for these penguins? **group_by()**, **summarize()**, **mean()**, and **sd()**

Body Mass Index for penguins?

Humans use the body mass index (BMI) to determine if someone is under-weight, normal weight, or over-weight. Let's **invent** a penguin body mass index. Maybe we can call it PMI for penguin mass index.
Humans use the formula

$$\text{BMI} = \frac{\text{Weight}}{\text{Height}^2}$$

where weight is in kilograms and height is in meters.
Let's try the following formula for the penguins

$$\text{PMI} = \frac{\text{Weight}}{(2.3 \times \text{Flipper})^2}$$

where weight is in grams and flipper is the flipper length in centimeters. (There are 10 millimeters in each centimeter.)

Penguin Mass Index

1. Add a new variable to the data set, say `pmi`, that computes the penguin mass index as given by the above formula.
2. Give a summary of the penguin mass index. **summary()**
3. Plot a histogram of the penguin mass index. **ggplot()**, **geom_histogram()**, **labs()**

Penguin Bills (not the utility kind)

1. Is the bill length and depth related to each other? Create a scatterplot with bill length on the x-axis and bill depth on the y-axis. What kind of patterns do you see in the scatterplot? **ggplot()**, **geom_point()**, **labs()**
2. Can we color the points differently by the penguin's species?
3. Maybe we should remove the rows where we have missing information. Find out, for each variable in the data set, which rows have missing information. **lapply()**, **function()**
4. Remove the rows with missing measurements.
5. Fit three linear regressions where the response variable is bill depth and the predictor variable is bill length. **lm()**, **summary()**
6. Are the slopes of the lines comparable across species?
7. Compute 96% confidence intervals for the coefficients of the the three regression lines and compare them. **confint()**, **rbind()**.
8. Can we add least squares regression lines to the previous scatterplot? **geom_smooth()**

Other questions

Human females tend (on average) to be smaller than human males. Does the same thing happen for penguins in each species?
Since we have lots of rows with missing sex information, let's remove them from the data set. **na.omit()**

1. Compute average weight, flipper, bill length and depth, by species and sex. **group_by()**, **summarize()**, **mean()**
2. Are the differences significant? Compute standard deviations and use them to see if there are big differences? **group_by()**, **summarize()**, **sd()**