Palmer Penguins Dataset Analysis in R - FSML S24

Martin Van Waerebeke

19/07/2024

1 Data Loading and Inspection

- Load the Palmer Penguins dataset into R using `data("penguins", package = "palmerpenguins")`.
- Use `head(dataset)` to view the first few rows of the dataset.
- Display a summary using the `summary(.)` function, showing types of variables and missing values.

2 Data Exploration and Manipulation

- List all variables using `names(.)`.
- Count observations using `nrow(.)` or `dim(.)[1]`. What is `dim(.)[2]` ?
- Calculate mean and median of a variable like flipper length using `mean(dataset$YOUR_VARIABLE)` and `median(dataset$YOUR_VARIABLE)`.
- Filter dataset with `subset(dataset, YOUR_VARIABLE == "VALUE_FOR_THIS_VARIABLE")` or `dataset[dataset$YOUR_VARIABLE > threshold,]`.
- Create a new variable using `dataset$new_variable <- expression`, like BMI calculated from mass and flipper length.

3 Data Visualization

- Use `boxplot(.)` to display a boxplot of the penguins's flipper length.
- Create a histogram with `hist(.)` (histograms' input are only one column, adapt accordingly !).
- Generate a scatter plot using `plot(., .)`. Again, scatter plot draws a 1 dimensional data as a function of another 1D data, be careful with the input !

- Produce a bar plot with `barplot(table(.))`. (Still 1D)

4 Data Grouping and Summarization

- Group by species and calculate average flipper length using `aggregate(attribute_to_measure˜attribute_by_which_to_group, data=dataset_name, mean`.
- Count penguins in each species-sex combination with `table(dataset$species, dataset$sex)`.
- (Advanced) Summarize multiple variables - ask ChatGPT for explanations: Summarize multiple variables using `aggregate(cbind(flipper_length, body_mass)   species, data = dataset, FUN = function(x) c(mean = mean(x), sd = sd(x)))`.

5 Data Analysis

Questions:

- **Variance (using `var()` function)**
  1. Calculate the variance of the flipper length (in mm) among the penguins in the dataset using the `var()` function. What does this variance indicate about the spread of flipper length measurements?
  2. Use the `var()` function to compute the variance of the body mass (in g) of the penguins. How can this variance help you understand the distribution of body mass values in the dataset?
- **Covariance (using `cov()` function)**
  1. Compute the covariance between the flipper length and body mass of the penguins using the `cov()` function. Interpret the sign (positive or negative) of the covariance value. What does it suggest about the relationship between these two variables?
- **Correlation Coefficient (using `cor()` function)**
  1. Determine the correlation coefficient between the flipper length and body mass using the `cor()` function. Explain how the correlation coefficient differs from covariance and what it reveals about the strength and direction of the relationship between flipper length and body mass.