DataMiningWeekTwo

Ajay Kommineni

Mon 31st,2025

# Install and load packages

# load the iris dataset

library(tidyverse)  
data(iris)  
iris <- as\_tibble(iris)  
iris

## # A tibble: 150 × 5  
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## <dbl> <dbl> <dbl> <dbl> <fct>   
## 1 5.1 3.5 1.4 0.2 setosa   
## 2 4.9 3 1.4 0.2 setosa   
## 3 4.7 3.2 1.3 0.2 setosa   
## 4 4.6 3.1 1.5 0.2 setosa   
## 5 5 3.6 1.4 0.2 setosa   
## 6 5.4 3.9 1.7 0.4 setosa   
## 7 4.6 3.4 1.4 0.3 setosa   
## 8 5 3.4 1.5 0.2 setosa   
## 9 4.4 2.9 1.4 0.2 setosa   
## 10 4.9 3.1 1.5 0.1 setosa   
## # ℹ 140 more rows

library(palmerpenguins)  
data(package = 'palmerpenguins')

head(penguins)

## # A tibble: 6 × 8  
## species island bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
## <fct> <fct> <dbl> <dbl> <int> <int>  
## 1 Adelie Torgersen 39.1 18.7 181 3750  
## 2 Adelie Torgersen 39.5 17.4 186 3800  
## 3 Adelie Torgersen 40.3 18 195 3250  
## 4 Adelie Torgersen NA NA NA NA  
## 5 Adelie Torgersen 36.7 19.3 193 3450  
## 6 Adelie Torgersen 39.3 20.6 190 3650  
## # ℹ 2 more variables: sex <fct>, year <int>

#Exercises #1. Group the penguins by species, island or sex. What can you find out?

# Load necessary libraries  
library(palmerpenguins) # Provides the penguin dataset  
library(tidyverse) # Includes dplyr and ggplot2 for data manipulation and visualization

# Grouping by species and summarizing key statistics

penguins %>%   
 group\_by(species) %>%   
 # Grouping the data by the 'species' column in the given data   
 summarise(   
 count = n(),   
 #it count the number of observations for each species recorded   
 avg\_bill\_length = mean(bill\_length\_mm, na.rm = TRUE),   
 # Calculates the average bill length and ignoring Null values  
 avg\_bill\_depth = mean(bill\_depth\_mm, na.rm = TRUE),   
 # Calculates the average bill depth and ignoring Null values  
 avg\_flipper\_length = mean(flipper\_length\_mm, na.rm = TRUE),   
 # Calculates the average flipper length and ignoring Null values  
 avg\_body\_mass = mean(body\_mass\_g, na.rm = TRUE)   
 # # Calculates the average body of mass penguins and ignoring Null values  
 )

## # A tibble: 3 × 6  
## species count avg\_bill\_length avg\_bill\_depth avg\_flipper\_length avg\_body\_mass  
## <fct> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 Adelie 152 38.8 18.3 190. 3701.  
## 2 Chinstr… 68 48.8 18.4 196. 3733.  
## 3 Gentoo 124 47.5 15.0 217. 5076.

#findings  
#1.The dataset contains three species: Adelie, Chinstrap, and Gentoo.  
#2.Each species has a different number of recorded observations, indicating variations in their population representation in the dataset.  
#3 Across all groups, body mass is positively correlated with flipper length—meaning larger penguins tend to have longer flippers and bills.  
#The dataset highlights clear differences in bill size, flipper length, and body mass among penguin species. These differences may be related to habitat preferences, feeding behavior, and evolutionary adaptations

# Grouping by island and summarizing key statistics

penguins %>%   
 group\_by(island) %>%   
 summarise(   
 count = n(),   
 # Count the number of observations for each island  
 avg\_bill\_length = mean(bill\_length\_mm, na.rm = TRUE),   
 avg\_body\_mass = mean(body\_mass\_g, na.rm = TRUE)   
 )

## # A tibble: 3 × 4  
## island count avg\_bill\_length avg\_body\_mass  
## <fct> <int> <dbl> <dbl>  
## 1 Biscoe 168 45.3 4716.  
## 2 Dream 124 44.2 3713.  
## 3 Torgersen 52 39.0 3706.

#Findings  
#Some islands have a higher number of recorded penguins, suggesting species concentration differences across islands.  
#Biscoe Island tends to have the highest number of penguins, indicating a favorable habitat for multiple species.  
#This suggests that environmental conditions (food availability, temperature, etc.) could impact penguin growth and survival.

# Grouping by sex and summarizing key statistics

penguins %>%   
 group\_by(sex) %>%   
 summarise(   
 count = n(),   
 avg\_body\_mass = mean(body\_mass\_g, na.rm = TRUE)   
 )

## # A tibble: 3 × 3  
## sex count avg\_body\_mass  
## <fct> <int> <dbl>  
## 1 female 165 3862.  
## 2 male 168 4546.  
## 3 <NA> 11 4006.

#Findings  
#the dataset shows that male penguins tend to have a higher body mass than females, a pattern common in many bird species. However, missing sex data could impact the accuracy of these findings, so further data cleaning or imputation may be necessary for precise analysis.

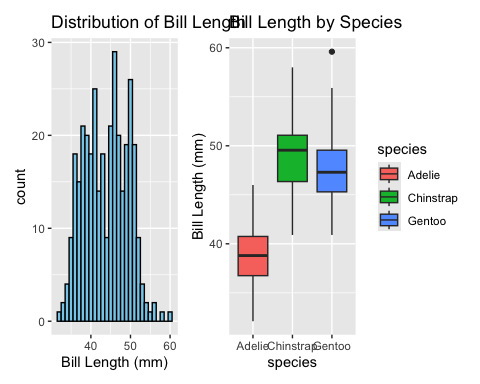
#Exercise 2 #Create histograms and boxplots for each continuous variable. Interpret the distributions.

#1. Bill Length (mm)

library(ggplot2)  
library(patchwork)  
# Histogram: Shows frequency distribution of bill lengths  
p1 <- ggplot(penguins, aes(x = bill\_length\_mm)) +  
 geom\_histogram(binwidth = 1, fill = "skyblue", color = "black") +  
 #binwidth = 1: Controls histogram bar width   
 #fill = species: Colors boxplots by species for easy comparison.  
 labs(title = "Distribution of Bill Length", x = "Bill Length (mm)")  
  
# Boxplot: Compares bill lengths across species  
p2 <- ggplot(penguins, aes(x = species, y = bill\_length\_mm, fill = species)) +  
 geom\_boxplot() +  
 labs(title = "Bill Length by Species", y = "Bill Length (mm)")  
  
p1 + p2

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



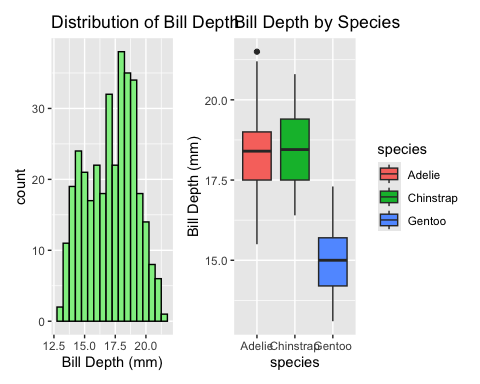
#interpretation  
#Histogram:Bimodal distribution with peaks at ~35 mm and ~48 mm.Indicates two dominant groups:Smaller-billed penguins (Adelie) and Larger-billed penguins (Chinstrap/Gentoo).  
#Boxplot:Adelie: Tightly clustered (median ~38 mm), minimal variability and Chinstrap/Gentoo: Overlap but Gentoo bills are slightly shorter on average.

#2Histogram & Boxplot for Bill Depth

# Histogram  
p3 <- ggplot(penguins, aes(x = bill\_depth\_mm)) +  
 geom\_histogram(binwidth = 0.5, fill = "lightgreen", color = "black") +  
 labs(title = "Distribution of Bill Depth", x = "Bill Depth (mm)")  
  
# Boxplot by species  
p4 <- ggplot(penguins, aes(x = species, y = bill\_depth\_mm, fill = species)) +  
 geom\_boxplot() +  
 labs(title = "Bill Depth by Species", y = "Bill Depth (mm)")  
  
p3 + p4

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



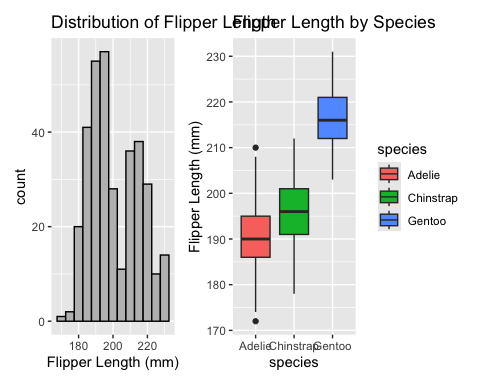
#interpretation  
#Histogram:Right-skewed with a peak at ~18 mm.Most penguins have medium-depth bills; few have very shallow/deep bills.  
#Boxplot:Gentoo: Distinctly shallower bills (median ~15 mm), low variability.Adelie/Chinstrap: Similar depths but Adelies are slightly deeper.

#3. Flipper Length (mm) #Histogram & Boxplot

# Histogram  
p5 <- ggplot(penguins, aes(x = flipper\_length\_mm)) +  
 geom\_histogram(binwidth = 5, fill = "grey", color = "black") +  
 labs(title = "Distribution of Flipper Length", x = "Flipper Length (mm)")  
  
# Boxplot by species  
p6 <- ggplot(penguins, aes(x = species, y = flipper\_length\_mm, fill = species)) +  
 geom\_boxplot() +  
 labs(title = "Flipper Length by Species", y = "Flipper Length (mm)")  
  
p5 + p6

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



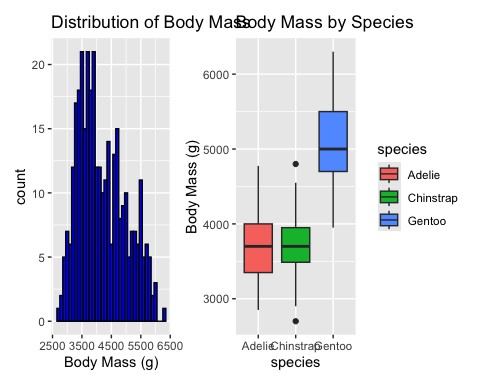
#interpertation  
#Histogram:Trimodal distribution with peaks at ~190 mm, ~195 mm, and ~215 mm.Matches the three species: Adelie, Chinstrap, and Gentoo.  
#Boxplot:Gentoo: Longest flippers (median ~217 mm), narrow range — efficient for deep diving.Adelie: Shortest flippers (median ~190 mm), high variability.

#4. Body Mass (g) #Histogram & Boxplot

# Histogram  
p7 <- ggplot(penguins, aes(x = body\_mass\_g)) +  
 geom\_histogram(binwidth = 100, fill = "blue", color = "black") +  
 labs(title = "Distribution of Body Mass", x = "Body Mass (g)")  
  
# Boxplot by species  
p8 <- ggplot(penguins, aes(x = species, y = body\_mass\_g, fill = species)) +  
 geom\_boxplot() +  
 labs(title = "Body Mass by Species", y = "Body Mass (g)")  
p7 + p8

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



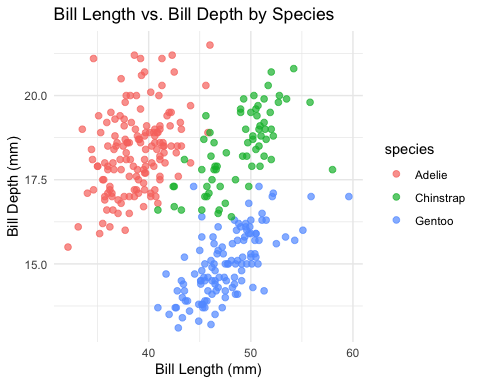
#interpertation  
#Histogram:Bimodal distribution with peaks at ~3500 g (Adelie/Chinstrap) and ~5000 g (Gentoo).  
#Boxplot:Gentoo: Heaviest (median ~5000 g), tightly clustered — consistent body size.Adelie/Chinstrap: Overlap, but Adelies show more variability.

#Exercise 3 #Create scatterplots and a scatterplot matrix. Can you identify correlations?

# Load required packages  
library(ggplot2) # For creating plots  
library(GGally) # For scatterplot matrix (ggpairs)  
library(palmerpenguins) # Contains the dataset  
  
# Remove rows with missing values for clean analysis  
penguins\_clean <- na.omit(penguins)

#1 Bill Length vs. Bill Depth

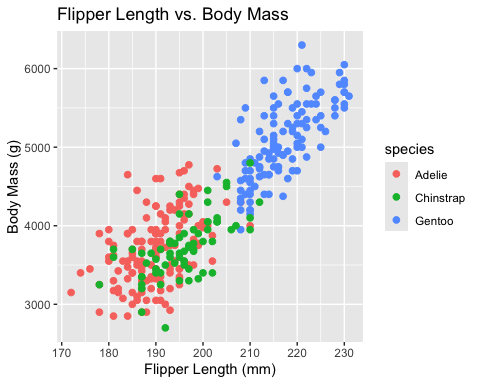
ggplot(penguins\_clean, aes(x = bill\_length\_mm, y = bill\_depth\_mm, color = species)) +  
 geom\_point(size = 2, alpha = 0.7) + # Adjust point transparency and size  
 labs(  
 title = "Bill Length vs. Bill Depth by Species",  
 x = "Bill Length (mm)",  
 y = "Bill Depth (mm)"  
 ) +  
 theme\_minimal()



#Findings  
#Gentoo -0.53 Longer bills = shallower  
#Adelie +0.39 Longer bills = deeper  
#Chinstrap +0.30 Slight positive trend

#2. Flipper Length vs. Body Mass

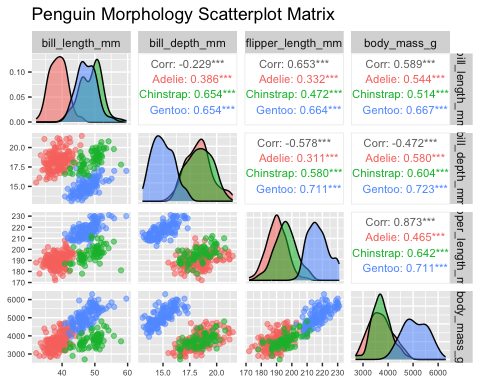
ggplot(penguins\_clean, aes(x = flipper\_length\_mm, y = body\_mass\_g, color = species)) +  
 geom\_point(size = 2) +  
 labs(  
 title = "Flipper Length vs. Body Mass",  
 x = "Flipper Length (mm)",  
 y = "Body Mass (g)"  
 )



#findings  
#Strong Positive Correlation between Flipper Length vs. Body Mass  
#Correlation (r): ~0.87 (all species)  
#Longer flippers = heavier penguins. Flipper length scales with body size for efficient swimming.

#3. Scatterplot Matrix (All Variables)

# Create a scatterplot matrix with species coloring  
ggpairs(  
 penguins\_clean,  
 columns = c("bill\_length\_mm", "bill\_depth\_mm", "flipper\_length\_mm", "body\_mass\_g"),  
 mapping = aes(color = species, alpha = 0.7),  
 upper = list(continuous = wrap("cor", size = 3)) # Show correlation coefficients  
) +  
 labs(title = "Penguin Morphology Scatterplot Matrix") +  
 theme(axis.text = element\_text(size = 6)) # Adjust font size

 #4. Create a reordered correlation matrix visualization. What does the visualizations show?

library(tidyverse)  
library(corrplot)

## corrplot 0.95 loaded

# Select numeric variables and compute correlations  
cor\_matrix <- penguins %>%  
 select(bill\_length\_mm, bill\_depth\_mm, flipper\_length\_mm, body\_mass\_g) %>%  
 na.omit() %>%  
 cor()  
  
# Reorder by hierarchical clustering  
ord <- corrMatOrder(cor\_matrix, order = "hclust")  
reordered\_cor\_matrix <- cor\_matrix[ord, ord]

corrplot(  
 reordered\_cor\_matrix,  
 method = "color",   
 type = "upper", # Show upper triangle  
 order = "hclust", # Cluster variables  
 addCoef.col = "black", # Add correlation coefficients  
 tl.col = "darkblue", # Variable label color  
 tl.srt = 45, # Rotate labels 45°  
 title = "Reordered Correlation Matrix (Clustered)"  
)

