p8105_hw1_dl3757

Dang Lin dl3757

2024-09-21

```
# Import the "tidyverse" library
library(tidyverse)

# Import the "ggplot2" library
library(ggplot2)
```

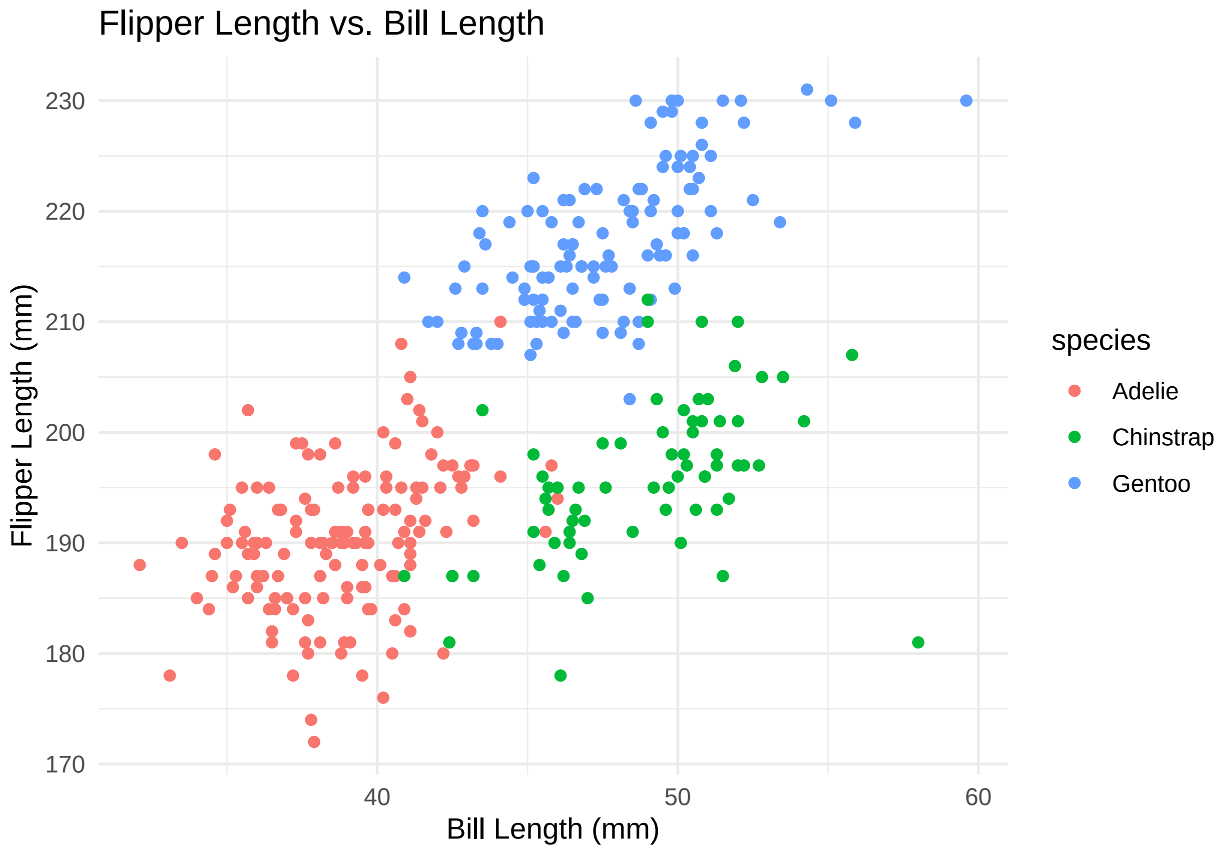
Problem 1

```
# Load the "penguins" data
data("penguins", package = "palmerpenguins")

# Remove the missing values in the dataset
penguins_clean <- penguins %>% na.omit()
```

The "penguins" dataset incorporates eight variables. Firstly, the dataset contains three species of penguins: Adelie, Gentoo, and Chinstrap, which are found on three different islands: Torgersen, Dream, and Biscoe. Besides, four numeric variables: bill length, bill depth, flipper length, and body mass of each penguin have been included in the dataset. Furthermore, this dataset also contains an ordinal variable "year", and a nominal data "sex" of the penguin. The "penguins" dataset has 344 rows and 8 columns, and the mean of flipper length is 200.966967 mm.

```
# Make a scatterplot of Flipper Length vs. Bill Length
ggplot(penguins_clean, aes(y = flipper_length_mm,
                           x = bill_length_mm,
                           color = species)) +
  geom_point() +
  labs(x = 'Bill Length (mm)',
       y = 'Flipper Length (mm)',
       title = 'Flipper Length vs. Bill Length') +
  theme_minimal()
```



```
# Export the scatterplot in PNG form
ggsave("flipper_length_vs_bill_length.png")
```

```
## Saving 6.5 x 4.5 in image
```

Problem 2

```
# Set seed for reproducibility
set.seed(3757)

# Create a data frame
df <- tibble(
  numeric_vector = rnorm(n = 10),
  logical_vector = numeric_vector > 0,
  character_vector = c("A", "B", "C", "D", "E",
                       "F", "G", "H", "I", "J"),

  factor_vector = factor(rep(c("Level_1", "Level_2", "Level_3"),
                     length.out = 10))
)
```

```
# Take the mean of each variable in the data frame
numeric_mean <- df %>% pull(numeric_vector) %>% mean()
logical_mean <- df %>% pull(logical_vector) %>% mean()
character_mean <- df %>% pull(character_vector) %>% mean()
factor_mean <- df %>% pull(factor_vector) %>% mean()
```

```
numeric_mean
```

```
## [1] 0.7582547
```

```
logical_mean
```

```
## [1] 0.8
```

```
character_mean
```

```
## [1] NA
```

```
factor_mean
```

```
## [1] NA
```

The mean of the numeric vector is 0.7582547, and the mean of the logical vector is 0.8. However, we cannot calculate the mean of the character and factor vectors, as they are not numeric.

```
# Convert the logical, character, and factor variables to the numeric variables
logical_numeric <- as.numeric(df[["logical_vector"]])
character_numeric <- as.numeric(df[["character_vector"]])
factor_numeric <- as.numeric(df[["factor_vector"]])
```

```
# Take the mean of converted variables
mean(logical_numeric)
```

```
## [1] 0.8
```

```
mean(character_numeric)
```

```
## [1] NA
```

```
mean(factor_numeric)
```

```
## [1] 1.9
```

After the conversion, values in the logical variable are converted into 0 or 1, resulting in a mean of 0.8. However, we still cannot calculate the mean of the character variable after the conversion because characters cannot be converted into any numeric values. Surprisingly, after we applied the as.numeric() function to the factor variable, it assigned numeric values to different factor variable levels, with each level represented by a corresponding number. Therefore, we are able to calculate the mean of the factor variable after conversion, which is 1.9.