P8105 Homework I

2022-09-27

```
# read and name data
data("penguins", package = "palmerpenguins")

# load the packages necessary for submission to knit.
library(tidyverse)
```

Problem 1

The database *penguins* has 8 variables and 344 observations. Here's a code chunk that creates an overall summary table, including the names of all the variables. As we can see in the table, the mean **flipper length** is 200.9152047.

```
## create overall summary table
library(gtsummary)
penguins %>%
  tbl_summary(missing_text = "(Missing)", # counts missing values
              statistic = list(all_continuous() ~ "{mean} ({sd})"), #continuous variables
              label = list(sex ~ "Sex", # label variables
                           species ~"Species",
                           island ~ "Island",
                           year ~ "Year",
                           body_mass_g ~ "Body Mass",
                           flipper_length_mm ~ "Flipper Length",
                           bill_depth_mm ~ "Bill Depth",
                           bill_length_mm ~ "Bill Length" )) %>%
  bold_labels() %>%
  italicize_levels()
```

Characteristic	**N = 344**
Species	
Adelie	152 (44%)
Chinstrap	68 (20%)
Gentoo	124 (36%)
Island	
Biscoe	168 (49%)
Dream	124 (36%)
Torgersen	52 (15%)
Bill Length	43.9 (5.5)
(Missing)	2
Bill Depth	17.15 (1.97)
(Missing)	2
Flipper Length	201 (14)
(Missing)	2
Body Mass	4,202 (802)
(Missing)	2

Characteristic	**N = 344**
Sex	
female	165 (50%)
male	168 (50%)
(Missing)	11
Year	
2007	110 (32%)
2008	114 (33%)
2009	120 (35%)

We now want to create a summary table by species. As we can see in the table, birds from the Gentoo species have, on average, lager flippers:

```
# create summary table by species
penguins %>%
  tbl_summary(by = "species",   # stratify by species
              missing_text = "(Missing)",
    statistic = list(all_continuous() ~ "{mean} ({sd})"),
    label = list(sex ~ "Sex", # label variables
                 species ~"Species",
                 island ~ "Island",
                 body_mass_g ~ "Body Mass",
                 flipper_length_mm ~ "Flipper Length",
                 bill_depth_mm ~ "Bill Depth",
                 bill_length_mm ~ "Bill Length" )) %>%
  bold_labels() %>%
  italicize_levels()
```

Characteristic	**Adelie**, N = 152	**Chinstrap**, N = 68	**Gentoo**, N = 124
Island			
Biscoe	44 (29%)	0 (0%)	124 (100%)
Dream	56 (37%)	68 (100%)	0 (0%)
Torgersen	52 (34%)	0 (0%)	0 (0%)
Bill Length	38.8 (2.7)	48.8 (3.3)	47.5 (3.1)
(Missing)	1	0	1
Bill Depth	18.35 (1.22)	18.42 (1.14)	14.98 (0.98)
(Missing)	1	0	1
Flipper Length	190 (7)	196 (7)	217 (6)
(Missing)	1	0	1
Body Mass	3,701 (459)	3,733 (384)	5,076 (504)
(Missing)	1	0	1
Sex			
female	73 (50%)	34 (50%)	58 (49%)
male	73 (50%)	34 (50%)	61 (51%)
(Missing)	6	0	5
year			
2007	50 (33%)	26 (38%)	34 (27%)
2008	50 (33%)	18 (26%)	46 (37%)
2009	52 (34%)	24 (35%)	44 (35%)

We can also describe this data with plots.

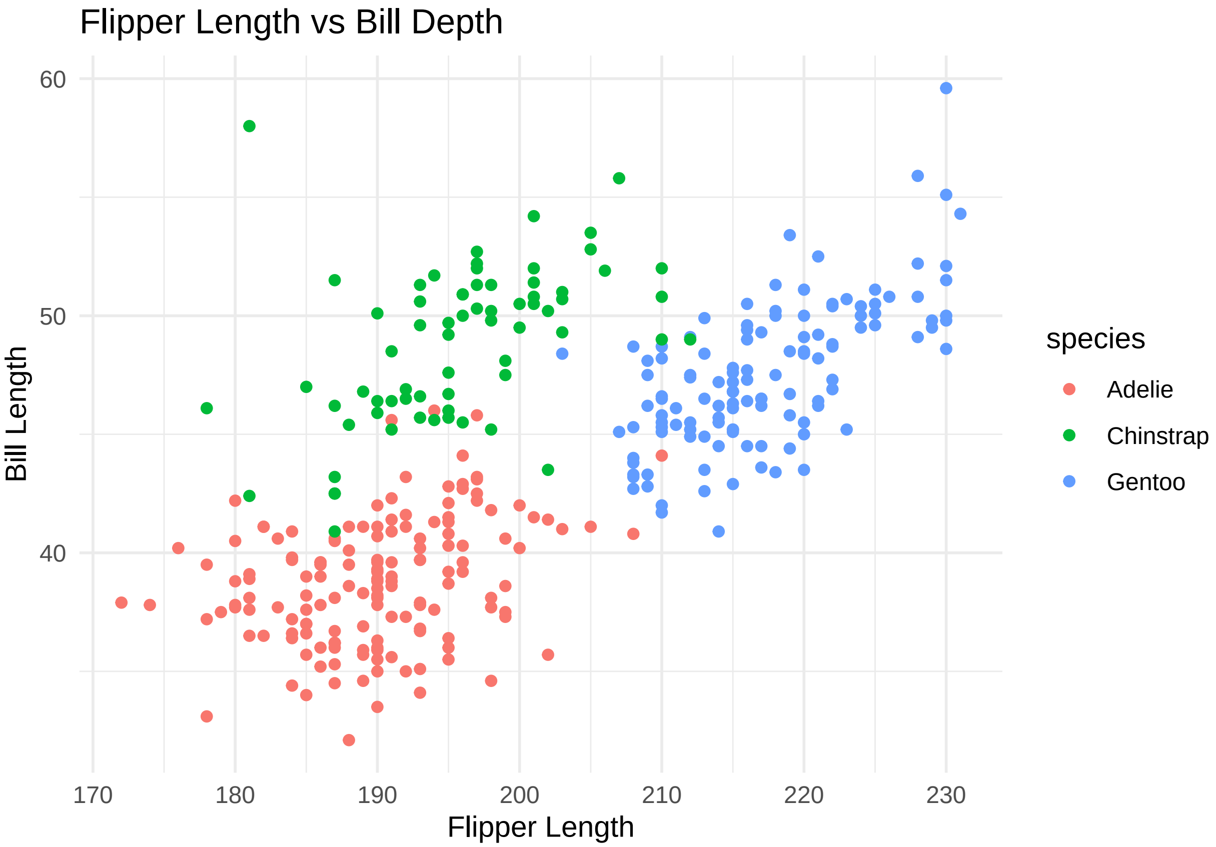
```
library(ggplot2)

#create scatterplot

f1 <- ggplot(penguins) +
      aes( x= flipper_length_mm, y= bill_length_mm,color=species) +
      geom_point () +
      labs(title="Flipper Length vs Bill Depth", x = "Flipper Length", y = "Bill Length") +
      theme_minimal()

f1
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



```
    png("f1.png")
```

Problem 2

Step 1 - Create a data frame:

```
# Create data frame

df <- tibble(
        random_sample = rnorm(10),
        logical_vector = random_sample > 0,
        character_vector = c("a", "b","c","d","e", "f", "g", "h","i", "j"),
        factor_vector = factor(c("low","medium","high","low","medium","high","low","medium","high","low
        )
```

Step 2 - Take the mean of each variable in your dataframe:

```
mean(df %>% pull(random_sample)) # Works
mean(df %>% pull(logical_vector)) # Works
mean(df %>% pull(character_vector)) # Does not work, argument is not numeric or logical
mean(df %>% pull(factor_vector)) # Does not work, argument is not numeric or logical
```

Step 3 - Convert variables from one type to another and calculate the mean:

```
# Code chunk that applies the as.numeric function to the logical, character, and factor variables

new_1 <-as.numeric(df %>% pull(logical_vectir))
mean(numeric_log_var) #True is converted to 1 and False is converted to 0

new_2 <- as.numeric(df %>% pull(character_vector))
mean(new_2) #The vector is now numeric. We can now calculate the mean.

new_3 <- as.numeric(df %>% pull(factor_vector))
mean(new_3) ##The vector is now numeric. We can now calculate the mean.
```