Homework Assignment 1

Brooklynn McNeil

2024-09-16

This is an R Markdown document for the first homework assignment in p8105 class: Data Science 1.

Problem 1

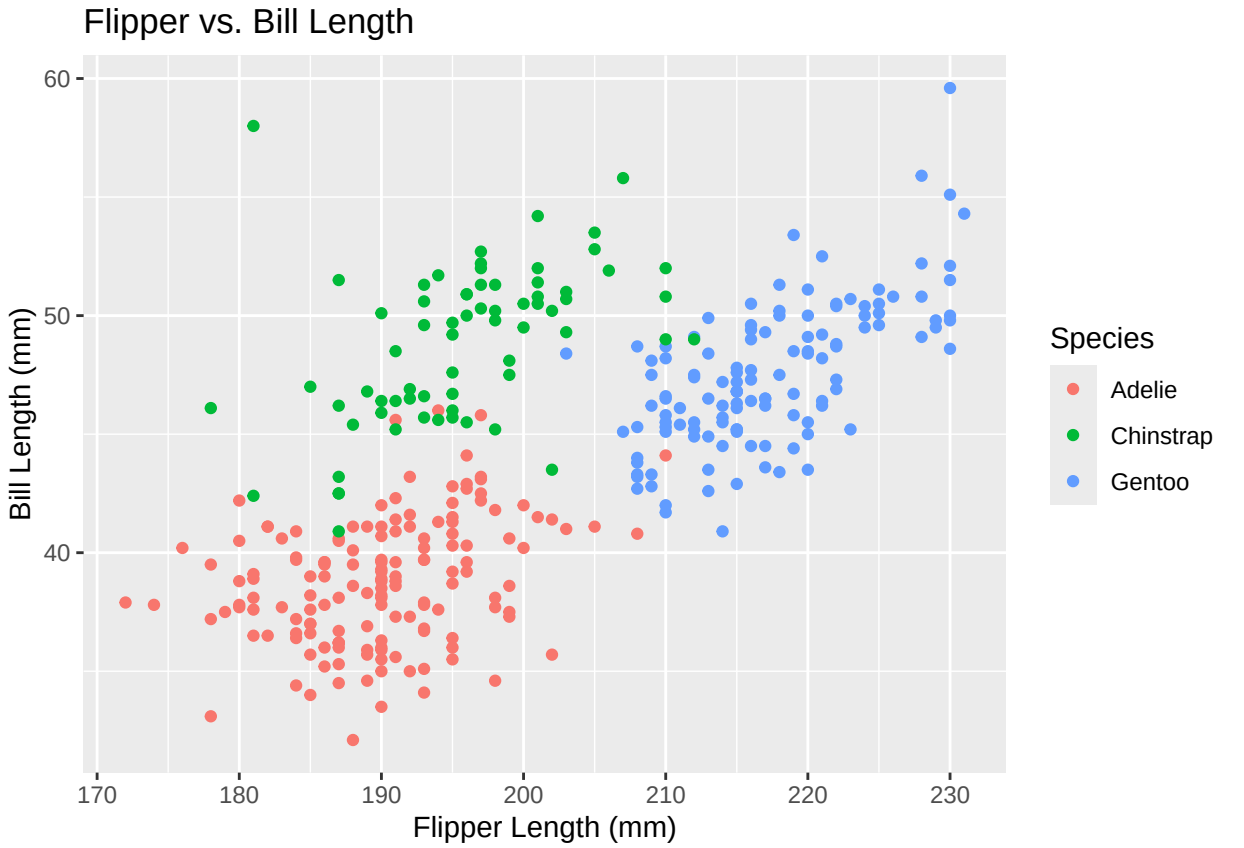
Description of the Penguins Dataset

The *penguins* data set is a subset of measurments of penguin species from the *Torgersen, Biscoe, Dream* islands in Palmer Archipelgo. The *penguins* data set includes the variables: *species, island, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, sex, year*. Penguins is a data frame with 344 rows and 8 columns. The mean flipper length is 200.9152047

Plots of Penguin Dataset

```
# create data frame
penguin_df = tibble(
  x = penguins$flipper_length_mm,
  y = penguins$bill_length_mm)

# create plot
ggplot(penguin_df) +
  geom_point(mapping = aes(x = x, y = y, color = penguins$species)) +
  labs(title = "Flipper vs. Bill Length",
       x = "Flipper Length (mm)",
       y = "Bill Length (mm)",
       color = "Species")
```




```
# save plot
ggsave(filename = "Flipper_vs_Bill_Length.png")
```

Problem 2

Create a data frame with 4 types of variables (numerical, logical, character, and factor) with a length of 10 for each.

```
my_df = tibble(
  vec_num = rnorm(n = 10),
  vec_lgl = vec_num > 0,
  vec_chr = c("hi", "my", "name", "is", "brooklynn", "and", "i", "like", "yummy","pasta"),
  vec_fac = factor(x = c(1,2,3,1,2,3,1,2,3,1))
)
my_df
```

```
## # A tibble: 10 x 4
##    vec_num vec_lgl vec_chr   vec_fac
##      <dbl> <lgl>   <chr>     <fct>
##  1  -0.787 FALSE   hi        1
##  2   0.206 TRUE    my        2
##  3  -1.39  FALSE   name      3
##  4  -1.73  FALSE   is        1
##  5   1.20  TRUE    brooklynn 2
##  6  -0.481 FALSE   and       3
##  7   0.883 TRUE    i         1
##  8  -0.851 FALSE   like      2
##  9  -0.488 FALSE   yummy     3
```

```
## 10   0.410 TRUE    pasta     1
```

```
## We can take the mean of numerical and logical vectors
mean(pull(my_df, var = vec_num))
```

```
## [1] -0.3023064
```

```
mean(pull(my_df, var = vec_lgl))
```

```
## [1] 0.4
```

```
## We cannot take the mean of character of factor vectors
mean(pull(my_df, var = vec_chr))
```

```
## Warning in mean.default(pull(my_df, var = vec_chr)): argument is not numeric or
## logical: returning NA
```

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

```
mean(pull(my_df, var = vec_fac))
```

```
## Warning in mean.default(pull(my_df, var = vec_fac)): argument is not numeric or
## logical: returning NA
```

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

This chunk shows an attempt at converting all the vectors in **my__df** data frame to a numerical format. This only works for the logical variable and not for the character or factor variables. This explains why taking a mean was impossible to calculate for these variables.

```
as.numeric(my_df$vec_log)
```

```
## Warning: Unknown or uninitialised column: `vec_log`.
```

```
## numeric(0)
```

```
as.numeric(my_df$vec_chr)
```

```
## Warning: NAs introduced by coercion
```

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

```
as.numeric(my_df$vec_fac)
```

```
##  [1] 1 2 3 1 2 3 1 2 3 1
```