HW1_jw4693

Jianming Wang

2024-09-13

This is a file for my homework 1 in Data Science I.

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

problem 0.1

I have committed to git for the first time for problem 0.1, which can be viewed in the commit history.

problem 0.2

This “problem” focuses on correct styling for your solutions to Problems 1 and 2.

problem 1

```
data("penguins", package = "palmerpenguins")
summary(penguins)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                   NA’s   :2       NA’s   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA’s  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA’s   :2         NA’s   :2
```

```r
nrow(penguins)
```

```
## [1] 344
```

```r
ncol(penguins)
```

```
## [1] 8
```

```r
mean(penguins$flipper_length_mm,na.rm = T)
```

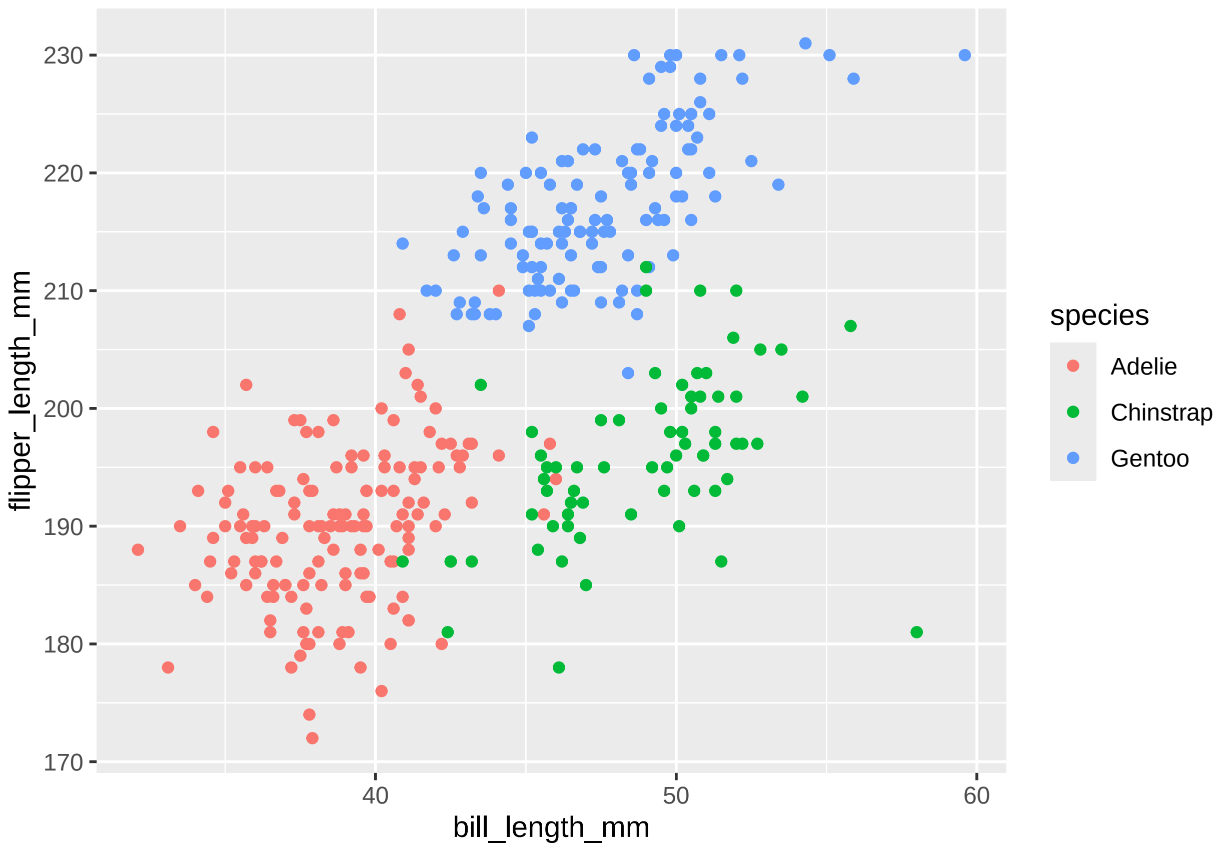
```
## [1] 200.9152
```

The 344 penguins counted in the data are from 3 different species, Adelie(152), Chinstrap(68) and Gentoo(124). They are from 3 island, Biscoe(168), Dream(124) and Torgersen(52). Except the 2 variables above, the data also contain 6 important variables for penguins, including bill_length_mm(mean 43.92, min 32.10, max 59.60 and median 44.45), bill_depth_mm(mean 17.15, min 13.10, max 21.50 and median 17.30), flipper_length_mm(mean 200.9, min 172.0, max 231.0 and median 197.0), body_mass_g(mean 4202, min 2700, max 6300 and median 4050),sex(male 168, female 165) and year(mean 2008, min 2007, max 2009 and median 2008). The variable bill_length_mm, bill_depth_mm, flipper_length_mm and body_mass_g all have 2 NAs(missing data), the variable sex has 11 NAs(missing data).

The data have 344 rows and 8 columns, and the mean flipper length is 200.9152047mm.

```r
ggplot(data = penguins, aes(x = bill_length_mm, y = flipper_length_mm, colour = species))+
  geom_point()
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



```
ggsave('scatterplot_for_problem_1.pdf')
```

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

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```

Tip: Missing data are removed automatically when drawing plot.

problem 2

```
set.seed(111)
test_data <- tibble(random_number = rnorm(10),
                    logical_vector = if_else(random_number>0, TRUE, FALSE),
                    character_vector = c('this','is','a','test','in','hw1','problem2','for','mean','valu
                    factor_vector = factor(c('left','right','right','left','middle','middle','middle','l
randomnumber <- pull(test_data,1)
mean(randomnumber)
```

```
## [1] -0.6690135
```

```r
logicalvector <- pull(test_data,2)
mean(logicalvector)
```

```
## [1] 0.2
```

```r
charactervector <- pull(test_data,3)
mean(charactervector)
```

```
## Warning in mean.default(charactervector): argument is not numeric or logical:
## returning NA
```

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

```r
factorvector <- pull(test_data,4)
mean(factorvector)
```

```
## Warning in mean.default(factorvector): argument is not numeric or logical:
## returning NA
```

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

The mean of random sample and logical vector can be calculated, the ‘TRUE’ and ‘FALSE’ respectively equal 1 and 0. The mean of character vector and factor vector cannot be calculated.

```r
charactervector <- as.numeric(pull(test_data,3))
```

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

```r
charactervector
```

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

```r
mean(charactervector)
```

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

```r
factorvector <- as.numeric(pull(test_data,4))
factorvector
```

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

```r
mean(factorvector)
```

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

After converting character vector and factor vector to numeric, the factor vector can be calculated and levels ‘left’, ‘middle’,‘right’ are respectively converted to 1,2 and 3. However, the character vector cannot be converted to numeric and calculated, and they will finally become NAs.