R4DS(2e) 1. 조망하기 - v1.1

한상곤(sangkon@pusan.ac.kr / sigmadream@gmail.com)

2023.06.13, updated: 2023.12.01

Contents

데이터 시각화

R을 사용해서 데이터분석의 전반적인 과정을 진행하기 위해서 필요한 것은 tidyverse라는 패키지 입니다. 해당 패키지를 설치하신 후 관련 라이브러리를 불러와주세요.

```
library(tidyverse)
```

R4DS(2e)에서 사용하게 될 팔머펭귄 데이터를 제공하는 palmerpenguins 패키지와 색맹에 안전한 시각적 표현법을 제공하는 ggthemes 패키지도 설치 후 불러와주세요.

```
library(palmerpenguins)
library(ggthemes)
```

첫번째. 데이터 확인

우리가 사용한 팔머펭귄 데이터를 확인해보도록 하겠습니다.

```
penguins
```

```
## # A tibble: 344 x 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
...
```

tidyverse를 사용하기 때문에 일반적인 형태의 data.frame이 아니라 tibble을 사용하는 것을 확인할 수 있습니다. tibble에 대한 세부적인 사항은 스터디를 진행하면서 알아보도록 하겠습니다. 이후, 데이터 셋의 각 변수(컬럼)에 대한 정보를 확인하도록 하겠습니다.

```
glimpse(penguins)
```

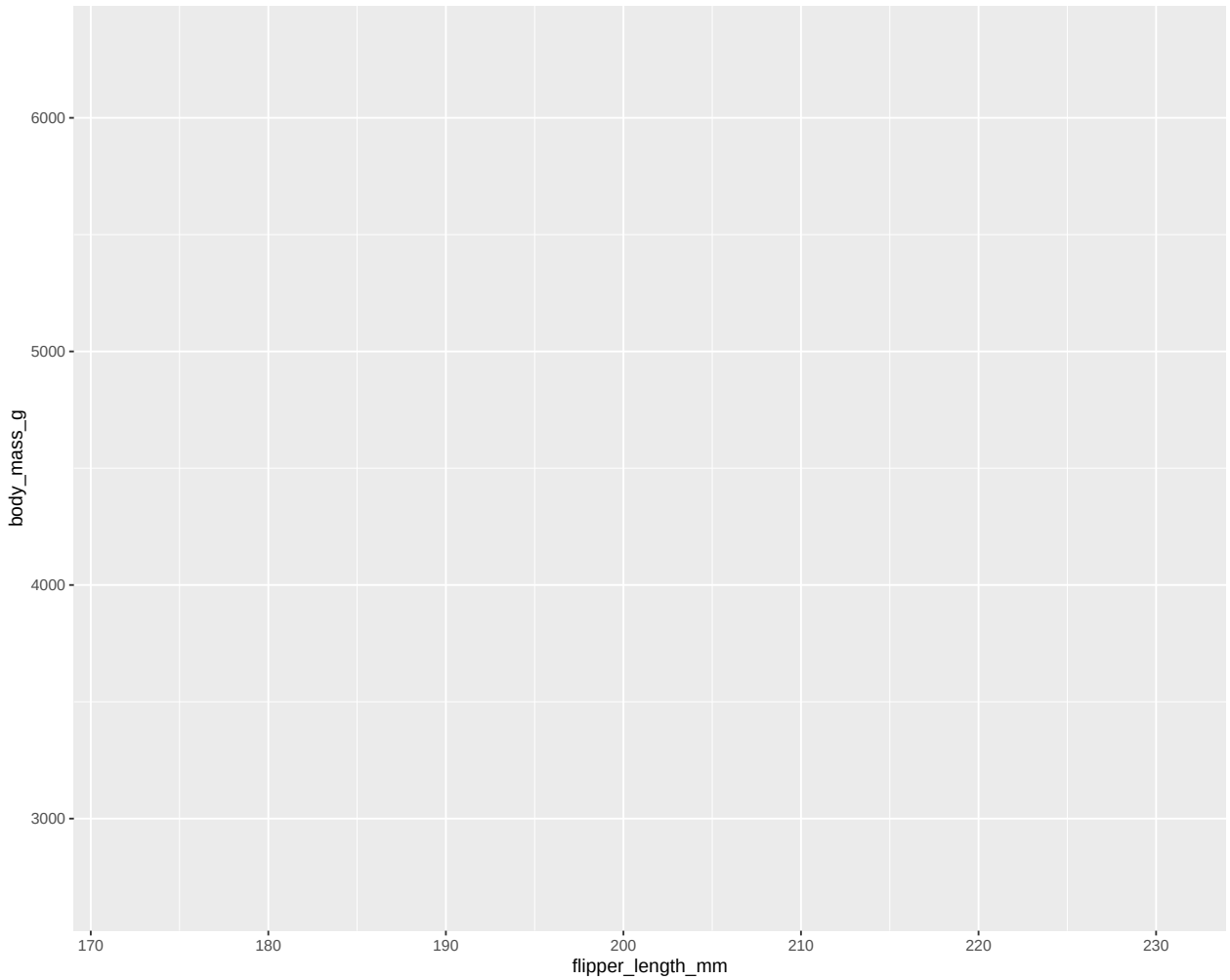
```
## Rows: 344
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
...
```

첫번째에서 우리가 원하는 것은 'flipper lengthsandbody masses의 상관관계를 시각적으로 표현하는 것 입니다. 일단 간단하게ggplot2'를 사용하도록 하겠습니다. 일단 ggplot2를 사용해서 그래프를 그릴 수 있는 객체를 생성합니다. 해당 그래프에 추가될 데이터에 대한 정보를 매개변수로 전달합니다.

```
ggplot(data = penguins)
```

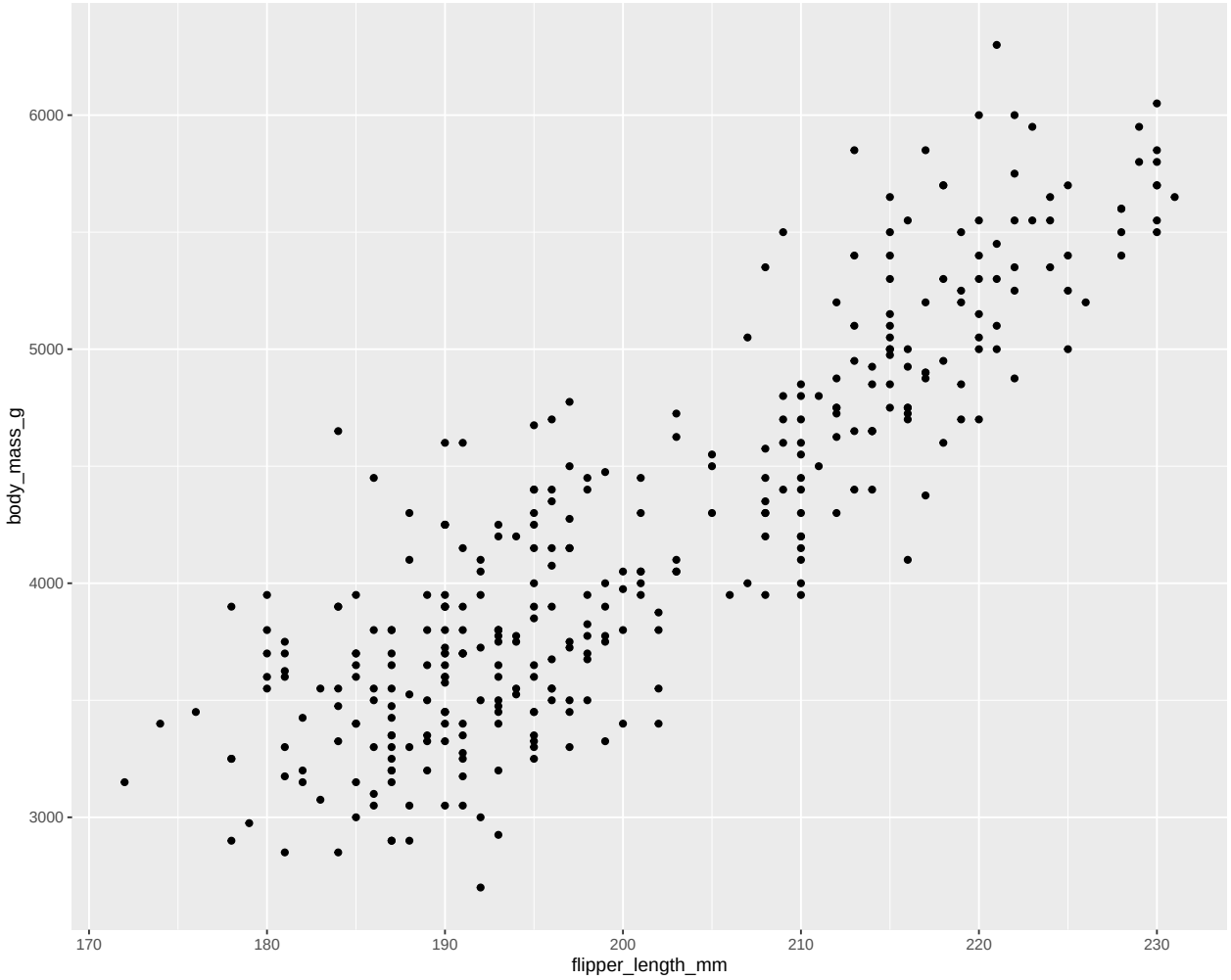

다음으로, 데이터의 정보를 시각적으로 표현하는 방법을 ggplot()에 전달해야 합니다. ggplot() 함수의 `mapping` 매개변수는 데이터 집합의 변수가 플롯의 시각적 속성(aesthetics)을 정의합니다. 시각적 속성은 aes() 함수를 사용해서 정의하며, aes()의 x 및 y 인수는 x축과 y축의 값을 정의합니다. 여기서는 x는 'flipper lengths이고 `y` 는body masses'입니다.

```
ggplot(data = penguins,
       mapping = aes(x = flipper_length_mm, y = body_mass_g))
```



기존에 존재하지 않던 격자가 생성되었다는 것을 통해서 x,y의 데이터가 정의되었다는 것을 확인할 수 있습니다. 하지만 시각적으로 표현하기 위한 방법을 아직 정의하지 않았기 때문에 관련 정보가 출력되지 않습니다. 데이터를 표현하는 데 사용하는 기하학적 개체인 geom을 정의해야 합니다. 이러한 기하학적 개체는 geom_로 시작하는 함수를 통해 ggplot2에서 사용할 수 있습니다. 예를 들어 막대형 차트에는 막대 도형(geom_bar()), 꺾은선형 차트에는 선 도형(geom_line()), 박스 플롯에는 박스 플롯 도형(geom_boxplot()), 스캐터 플롯에는 점 도형(geom_point()) 등을 사용합니다.

```
ggplot(data = penguins,
       mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point()
```



산점도의 특성상 변수에 대한 내용을 명확하게 이해가 쉽지 않기 때문에 색을 활용해서 추가 정보를 표현하도록 하겠습니다.

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g, color = species)
) +
  geom_point()
```


6000
5000
4000
3000
body_mass_g
170
180
190
200
210
220
230
flipper_length_mm
species
Adelie
Chinstrap
Gentoo