Part 1. R 프로그래밍 (데이터 분석 전문가 양성과정)

09

탐색적 데이터 분석

경북대학교 배준현 교수
(joonion@knu.ac.kr)

09. 탐색적 데이터 분석

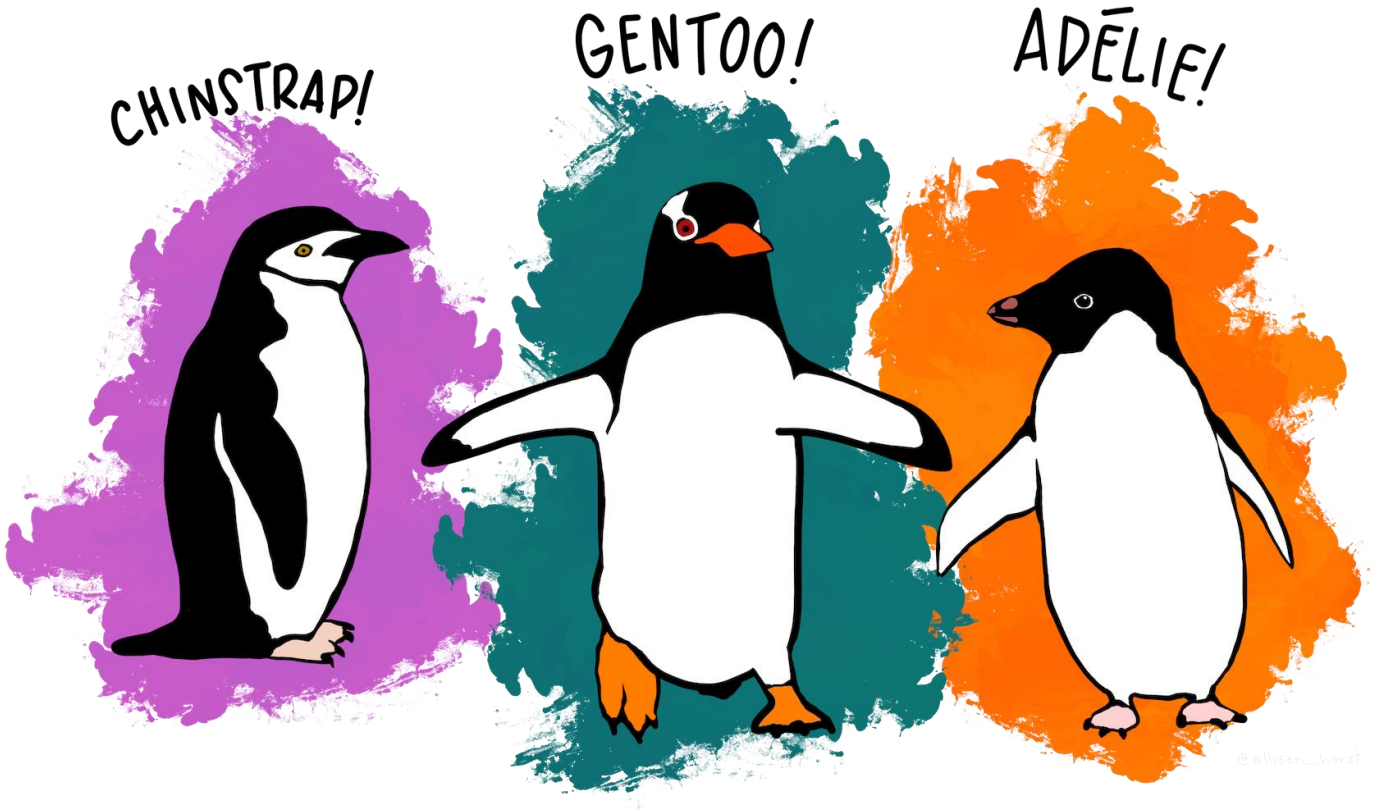
- **데이터에 대한 두 가지 접근법**: CDA *.vs.* EDA
  - **확증적** 데이터 분석: CDA, *confirmatory* data analysis
    - 가설을 수립하고 데이터를 통해 통계적 유의성을 **검정**하는 전통적 분석 기법
    - *Ronald Fisher*: 가설검정, 신뢰구간, 유의확률, 유의수준(*p-value*)
  - **탐색적** 데이터 분석: EDA, *exploratory* data analysis
    - 정해진 가설과 모형없이 데이터의 구조와 특성을 통해 **통찰**을 얻는 분석 기법
    - *John Tukey*: EDA는 우리가 존재한다고 믿는 것들은 물론이고, 존재하지 않는다고 믿는 것들을 발견하려는 태도, 유연성, 그리고 자발성이다.

- **팔머펭귄 데이터셋: *palmerpenguins* dataset**
  - 남극의 팔머 군도에 서식하는 3종의 펭귄에 대한 데이터셋
  - 데이터 분석과 시각화 교육용으로 적절한 특성을 가지고 있음

https://bit.ly/36rDgFx

Gorman, Kristen B., Tony D. Williams, and William R. Fraser. "Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis)." PloS one 9.3 (2014): e90081.

- **팔머펭귄의 종류:**
  - 턱끈: *chinstrap*
  - 젠투: *gentoo*
  - 아델리: *adelie*

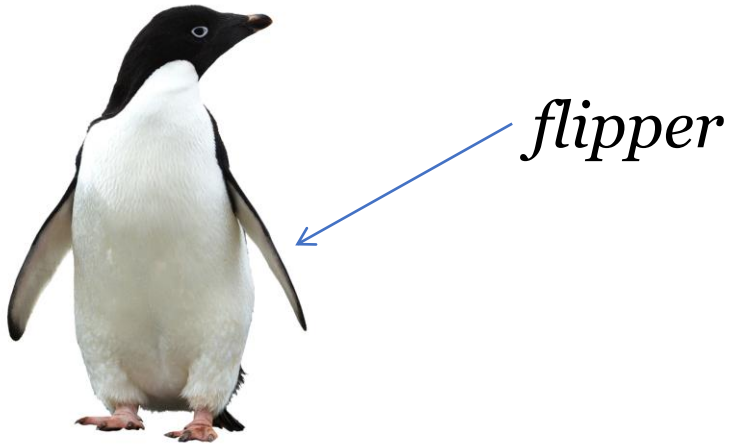


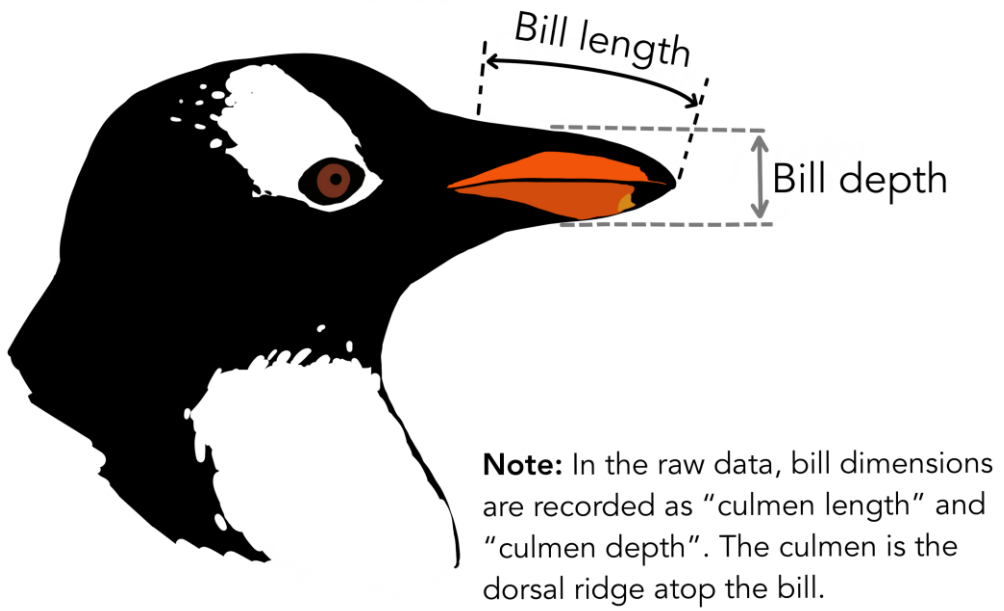
Artwork by @allison_horst

09. 탐색적 데이터 분석

- **데이터셋 정보:**
  - 관측값: 344개
  - 특징변수: 8개
    - 수치형 변수: 5개
    - 범주형 변수: 3개
    - 종속변수: *species*
    - 독립변수: 7개

- **수치형 변수**: *numeric* variables
  - *bill_length_mm*: **부리의 길이**
  - *bill_depth_mm*: **부리의 깊이**
  - *flipper_length_mm*: **팔(?)의 길이** (**날개? 지느러미?**)
  - *body_mass_g*: **체중**
  - *year*: **연구년도**(2007, 2008, 2009)

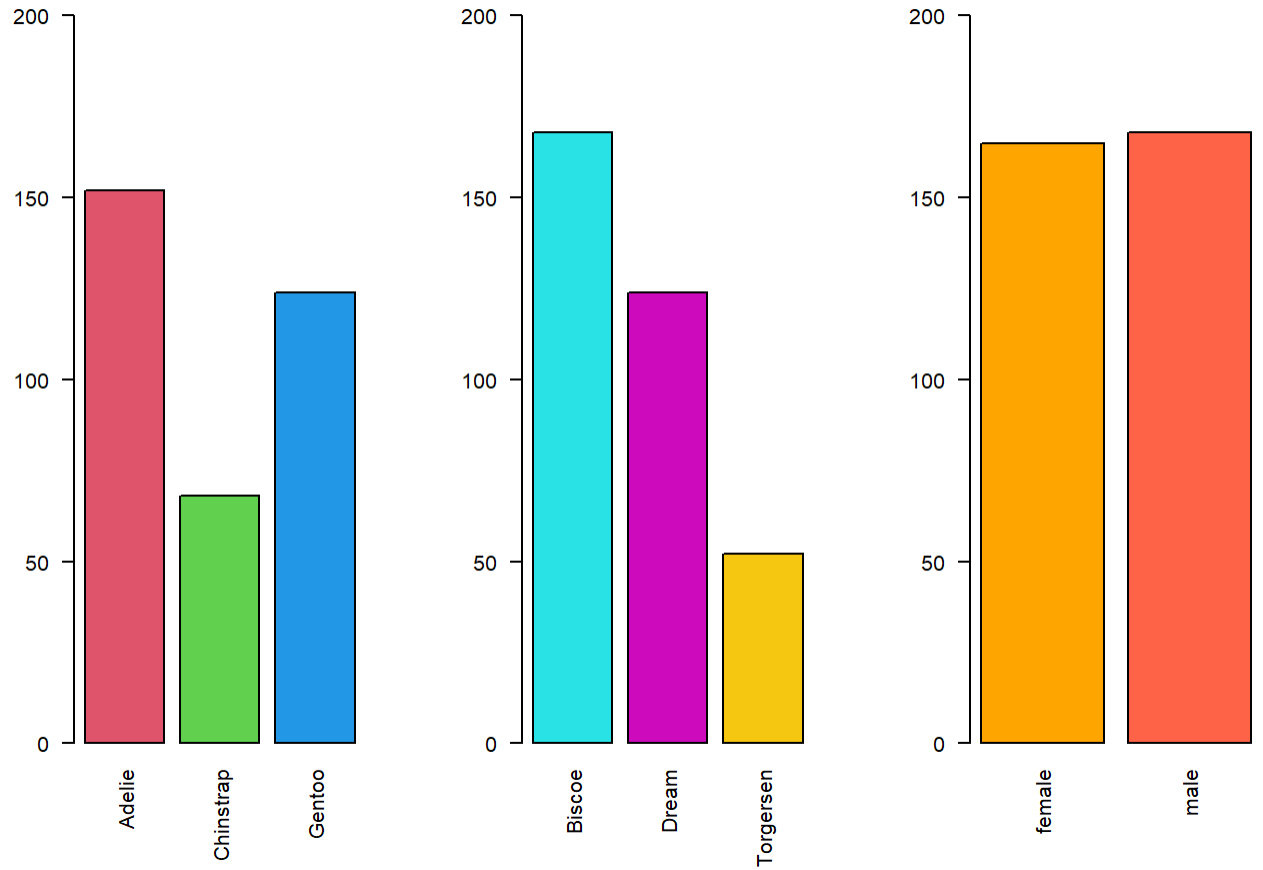




- **범주형 변수**: *categorical* variables
  - *species*: **종**
    - Adelie, Chinstrap, Gentoo
  - *island*: **섬(서식지)**
    - Biscoe, Dream, Torgersen
  - *sex*: **성별**
    - female, male

```
> summary(penguins[, c(1, 2, 7)])
      species          island         sex
 Adelie   :152   Biscoe   :168   female:165
 Chinstrap: 68   Dream    :124   male  :168
 Gentoo   :124   Torgersen: 52   NA's  : 11
```

- 3가지 범주형 변수에 대한 막대그래프를 그려보자.
- par() 함수로 3개의 파티션을 구분하여 하나의 플롯으로 그린다.



- 3가지 범주형 변수에 대해서 각 범주의 비율을 표로 만들어보자.

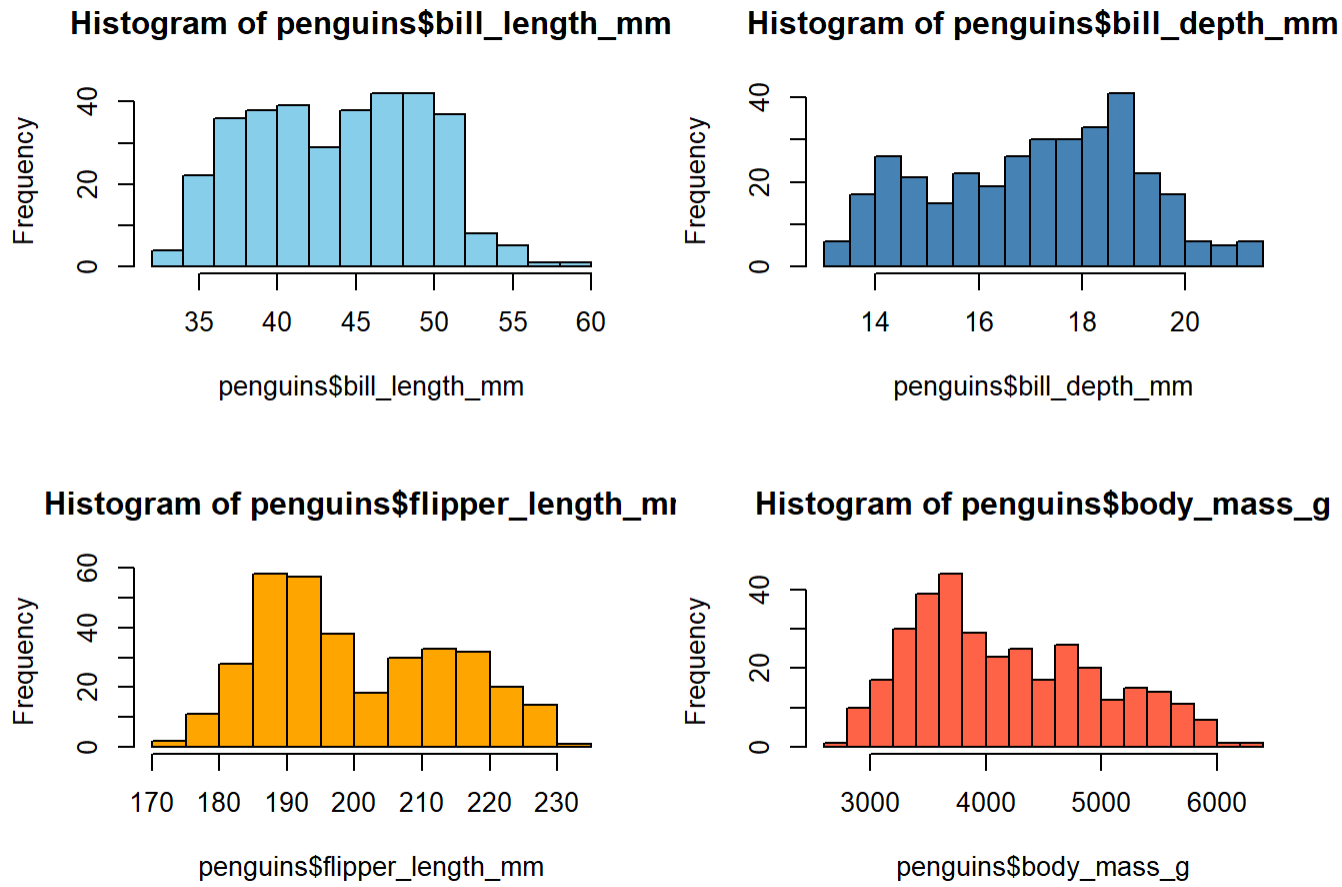
```
>
   Adelie Chinstrap    Gentoo
0.4418605 0.1976744 0.3604651

>
   Biscoe     Dream Torgersen
0.4883721 0.3604651 0.1511628

>
   female      male
0.4954955 0.5045045
```

- 각 종별로 어떤 섬에 서식하고 있는지 교차표를 만들어보자.
- gmodels 패키지의 CrossTable() 함수를 이용하여 교차표를 만들어보자.

```
>
           island
species     Biscoe Dream Torgersen
  Adelie        44    56         52
  Chinstrap      0    68          0
  Gentoo       124     0          0
```

09. 탐색적 데이터 분석

- 4가지 수치형 변수에 대한 히스토그램을 그려보자.
- par() 함수로 2*2 개의 파티션을 나누어 그려본다.



09. 탐색적 데이터 분석

- psych 패키지의 describe() 함수를 통해서 기술통계량을 산출해보자.

```
>
                  vars   n    mean      sd  median trimmed     mad     min     max
species*             1 344    1.92    0.89    2.00    1.90    1.48     1.0     3.0
island*              2 344    1.66    0.73    2.00    1.58    1.48     1.0     3.0
bill_length_mm       3 342   43.92    5.46   44.45   43.91    7.04    32.1    59.6
bill_depth_mm        4 342   17.15    1.97   17.30   17.17    2.22    13.1    21.5
flipper_length_mm    5 342  200.92   14.06  197.00  200.34   16.31   172.0   231.0
body_mass_g          6 342 4201.75 801.95 4050.00 4154.01 889.56 2700.0 6300.0
sex*                 7 333    1.50    0.50    2.00    1.51    0.00     1.0     2.0
year                 8 344 2008.03    0.82 2008.00 2008.04    1.48 2007.0
```

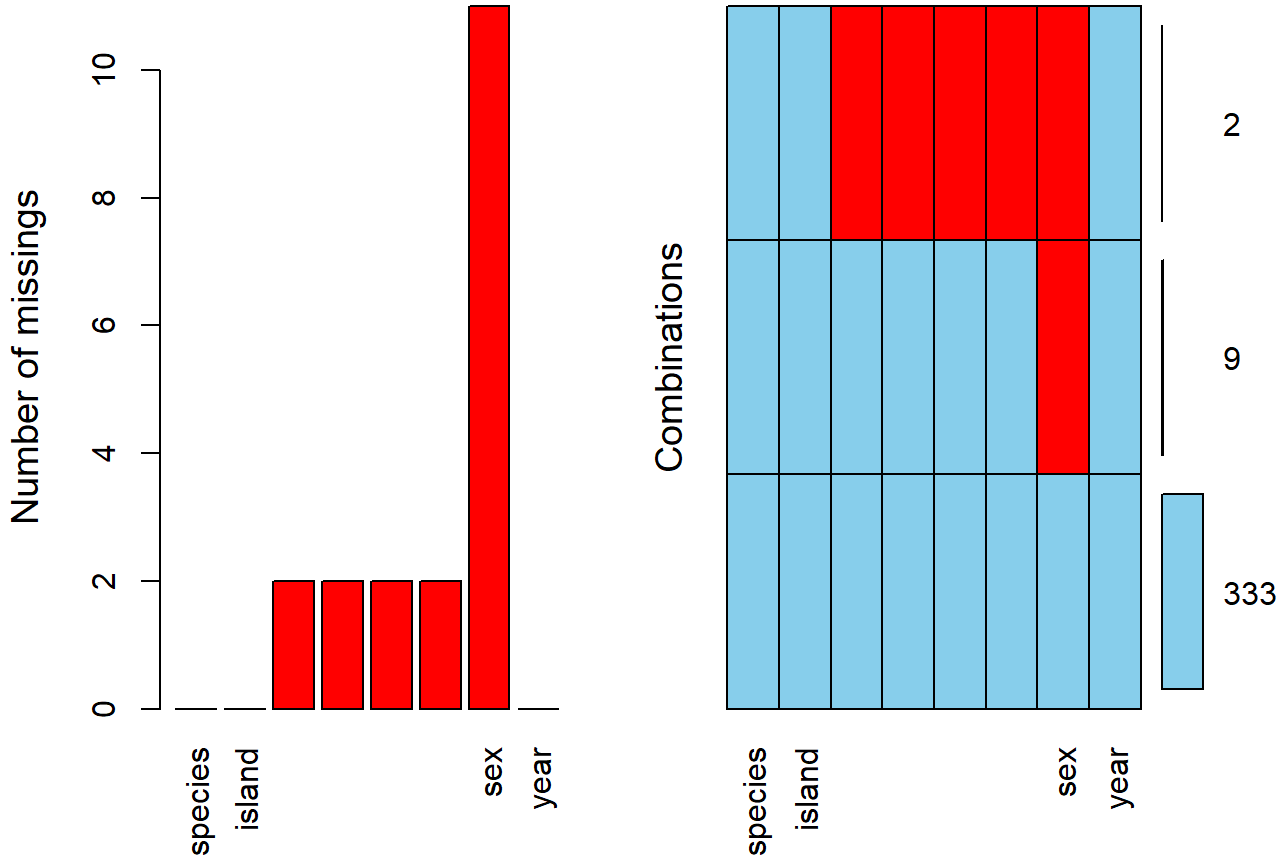
- aggregate() 함수를 이용하여 범주별 기술통계량을 산출할 수 있다.
- 펭귄의 종별로 부리의 길이와 깊이, 날개의 길이, 체질량의 평균을 확인해보자.

```
>
    species bill_length_mm
1    Adelie       38.79139
2 Chinstrap       48.83382
3  Gentoo        47.50488

>
    species bill_depth_mm
1    Adelie      18.34636
2 Chinstrap      18.42059
3    Gentoo      14.98211
```

- penguins 데이터셋에는 결측치(NA)가 포함되어 있다.
- is.na() 함수를 이용하여 각 변수별로 결측치가 몇 개인지 확인해보자.

```
>
species 0
island 0
bill_length_mm 2
bill_depth_mm 2
flipper_length_mm 2
body_mass_g 2
sex 11
year 0
```

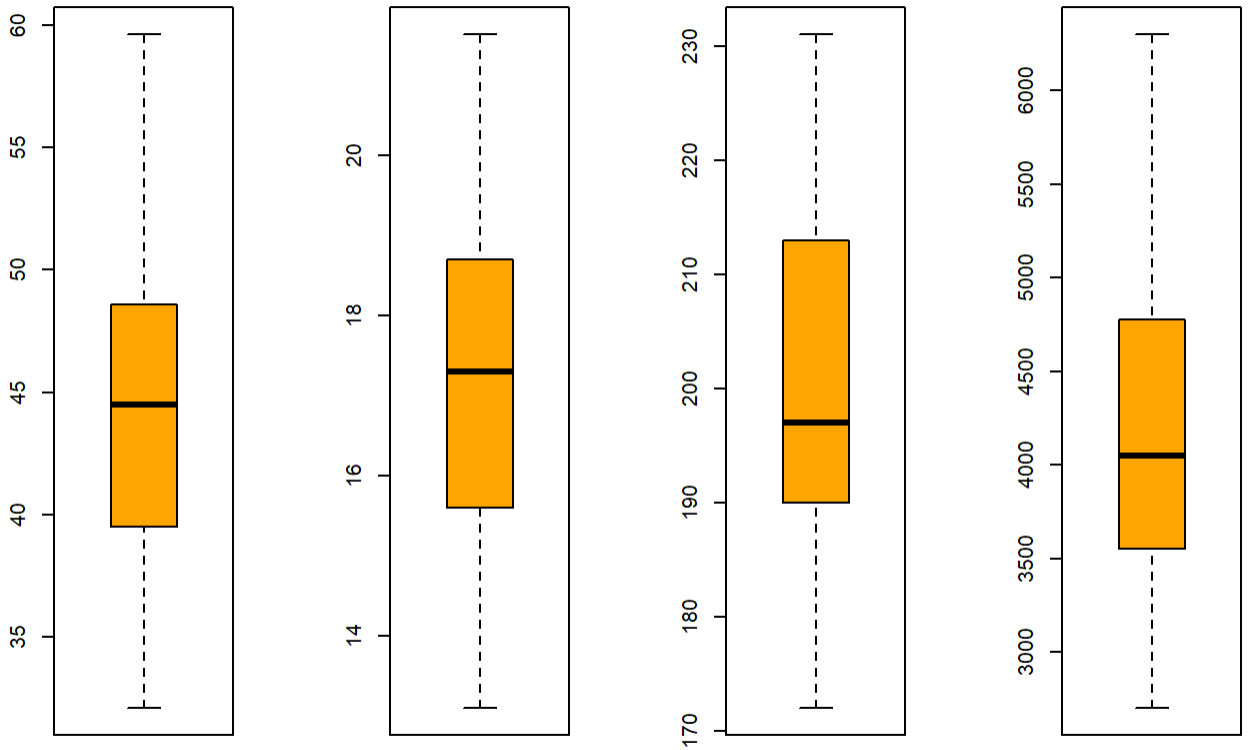
- VIM 패키지의 aggr() 함수를 이용하여 결측값의 패턴을 확인해보자.
- aggr() 함수는 변수별로 결측치의 분포와 변수별 조합에 따른 결측치의 패턴을 그림으로 보여준다.



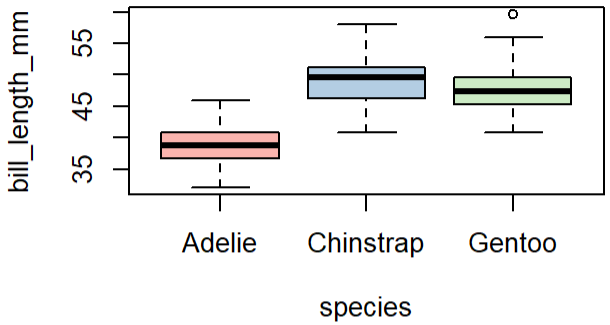
- complete.cases() 함수를 이용하여 결측치가 포함된 행을 확인할 수 있다.
- 결측치를 포함한 데이터를 삭제하여 pg 라는 이름의 데이터프레임으로 저장하자.

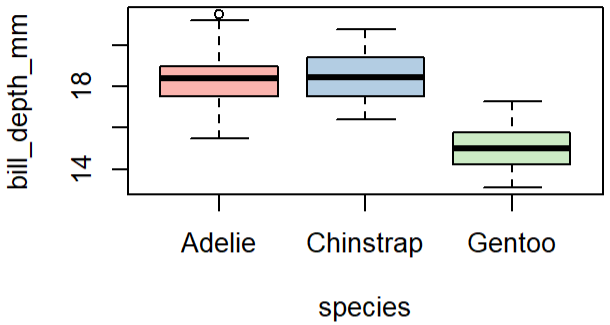
```
> sum(!complete.cases(penguins))
[1] 11
> penguins[!complete.cases(penguins), ]
# A tibble: 11 × 8
   species island    bill_length_mm bill_depth_mm flipper_…¹ body_…² sex    year
   <fct>   <fct>              <dbl>         <dbl>      <int>   <int> <fct> <int>
 1 Adelie  Torgersen             NA            NA         NA      NA <NA>   2007
 2 Adelie  Torgersen           34.1          18.1        193    3475 <NA>   2007
 3 Adelie  Torgersen           42            20.2        190    4250 <NA>
......

> pg <- penguins[complete.cases(penguins), ]
> dim(pg)
[1] 333   8
```

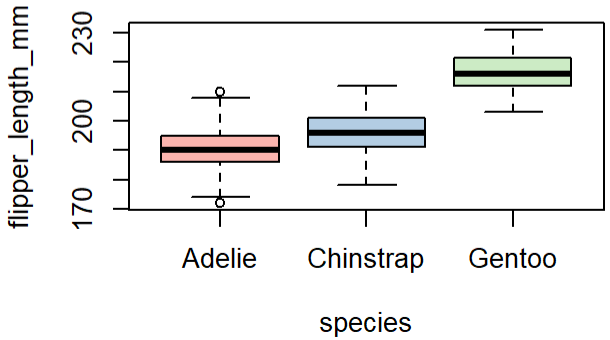
- 결측치를 제거한 penguins 데이터셋에서 4가지 수치형 변수에 대해 박스플롯을 그려보자.

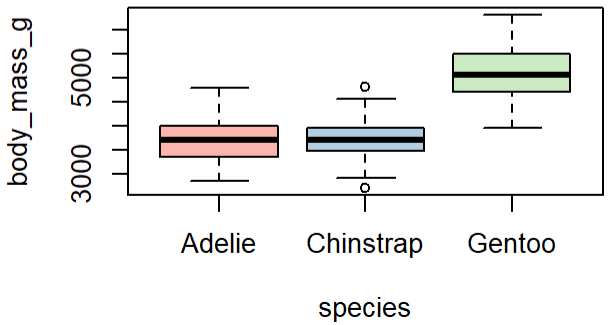


- 펭귄의 종별로 박스플롯을 그려보자.









- boxplot.stats() 함수를 이용하여 Adelie 펭귄의 날개 길이에서 나타난 이상치의 값과 Chinstrap 펭귄의 체질량에서 나타난 이상치의 값을 확인해보자.

```
>
$stats
[1] 174 186 190 195 208

$n
[1] 146

$conf
[1] 188.8231 191.1769

$out
[1] 172 210
```

- order() 함수를 이용하여 penguins 데이터셋을 날개의 길이를 기준으로 오름차순으로 정렬해보자.

```
# A tibble: 333 × 8
   species   island    bill_length_mm bill_depth_mm flippe…¹ body_…² sex    year
   <fct>     <fct>              <dbl>         <dbl>    <int>   <int> <fct> <int>
 1 Adelie    Biscoe              37.9          18.6      172    3150 fema…  2007
 2 Adelie    Biscoe              37.8          18.3      174    3400 fema…  2007
 3 Adelie    Torgersen           40.2          17        176    3450 fema…  2009
 4 Adelie    Dream               39.5          16.7      178    3250 fema…  2007
 5 Adelie    Dream               37.2          18.1      178    3900 male   2007
 6 Adelie    Dream               33.1          16.1      178    2900 fema…  2008
 7 Chinstrap Dream               46.1          18.2      178    3250 fema…  2007
 8 Adelie    Biscoe              37.7          18.7      180    3600 male   2007
 9 Adelie    Biscoe              38.8          17.2      180    3800 male   2007
10 Adelie    Biscoe              40.5          18.9      180    3950 male   2007
```

- penguins 데이터셋을 날개의 길이를 기준으로 오름차순으로 정렬하되,
- 날개의 길이가 같으면 체질량을 기준으로 내림차순으로 정렬해보자.

```
# A tibble: 333 × 8
   species   island    bill_length_mm bill_depth_mm flippe…¹ body_…² sex    year
   <fct>     <fct>              <dbl>         <dbl>    <int>   <int> <fct> <int>
 1 Adelie    Biscoe              37.9          18.6      172    3150 fema…  2007
 2 Adelie    Biscoe              37.8          18.3      174    3400 fema…  2007
 3 Adelie    Torgersen           40.2          17        176    3450 fema…  2009
 4 Adelie    Dream               37.2          18.1      178    3900 male   2007
 5 Adelie    Dream               39.5          16.7      178    3250 fema…  2007
 6 Chinstrap Dream               46.1          18.2      178    3250 fema…  2007
 7 Adelie    Dream               33.1          16.1      178    2900 fema…  2008
 8 Adelie    Biscoe              40.5          18.9      180    3950 male   2007
 9 Adelie    Biscoe              38.8          17.2      180    3800 male   2007
10 Adelie    Biscoe              37.7          18.7      180    3600 male   2007
```

- penguins 데이터셋에서 부리의 길이가 가장 긴 10개의 데이터를 출력해보자.

```
# A tibble: 10 × 8
   species   island bill_length_mm bill_depth_mm flipper_l…¹ body_…² sex    year
   <fct>     <fct>           <dbl>         <dbl>       <int>   <int> <fct> <int>
 1 Gentoo    Biscoe           59.6          17           230    6050 male   2007
 2 Chinstrap Dream            58            17.8         181    3700 fema…  2007
 3 Gentoo    Biscoe           55.9          17           228    5600 male   2009
 4 Chinstrap Dream            55.8          19.8         207    4000 male   2009
 5 Gentoo    Biscoe           55.1          16           230    5850 male   2009
 6 Gentoo    Biscoe           54.3          15.7         231    5650 male   2008
 7 Chinstrap Dream            54.2          20.8         201    4300 male   2008
 8 Chinstrap Dream            53.5          19.9         205    4500 male   2008
 9 Gentoo    Biscoe           53.4          15.8         219    5500 male   2009
10 Chinstrap Dream            52.8          20           205    4550 male   2008
```

- split() 함수를 이용하여 penguins 데이터셋을 종류별로 구분해보자.

```
> sp <- split(pg, pg$species)
> class(sp)
> summary(sp)

          Length Class  Mode
Adelie    8      tbl_df list
Chinstrap 8      tbl_df list
Gentoo    8      tbl_df list
```

- rbind() 함수를 이용하여 앞에서 분리한 sp 리스트에서 Adelie와 Gentoo를 합쳐보자.
- split()으로 분리할 때 species의 범주 정보가 남아 있으므로, rbind()를 한 이후에 factor() 생성자로 형 변환을 하면 범주가 두 개만 남게 된다

```
> AG <- rbind(sp$Adelie, sp$Gentoo)
> levels(AG$species)
> AG$species <- factor(AG$species)
> levels(AG$species)
```

- penguins 데이터셋에서 sample() 함수를 이용하여 20개의 데이터를 임의로 추출해 보자.
- 단, 추출 후에도 원래 데이터를 구분할 수 있도록 먼저 id 변수를 추가하도록 한다.

```
> set.seed(2022)
> df.origin <- data.frame(id = 1:nrow(pg), species = pg$species,
                          bill_length = pg$bill_length_mm,
                          bill_depth = pg$bill_depth_mm)
> idx <- sample(1:nrow(df.origin), size = 20)
> df.sample <- df.origin[idx, ]
> df.sample
```

- df.sample 데이터프레임을 부리의 길이와 깊이를 따로 분리한 후에, 두 개의 데이터프레임에서 각각 10개의 데이터를 샘플링한다.
- 각각 10개의 데이터를 가진 두 개의 데이터프레임을 merge() 함수를 이용하여 다시 합쳐보자.
- 기준 변수는 원래 데이터셋의 id 변수를 사용한다.

```
> set.seed(2022)
> pg.x <- df.sample[sample(1:nrow(df.sample), size = 10), ][, c(1, 2, 3)]
> pg.x[order(pg.x$id), ]
> pg.y <- df.sample[sample(1:nrow(df.sample), size = 10), ][, c(1, 2, 4)]
> pg.y[order(pg.y$id), ]
> merge(x = pg.x, y = pg.y, by = c("id", "species"), all = T)
```

Any Questions?