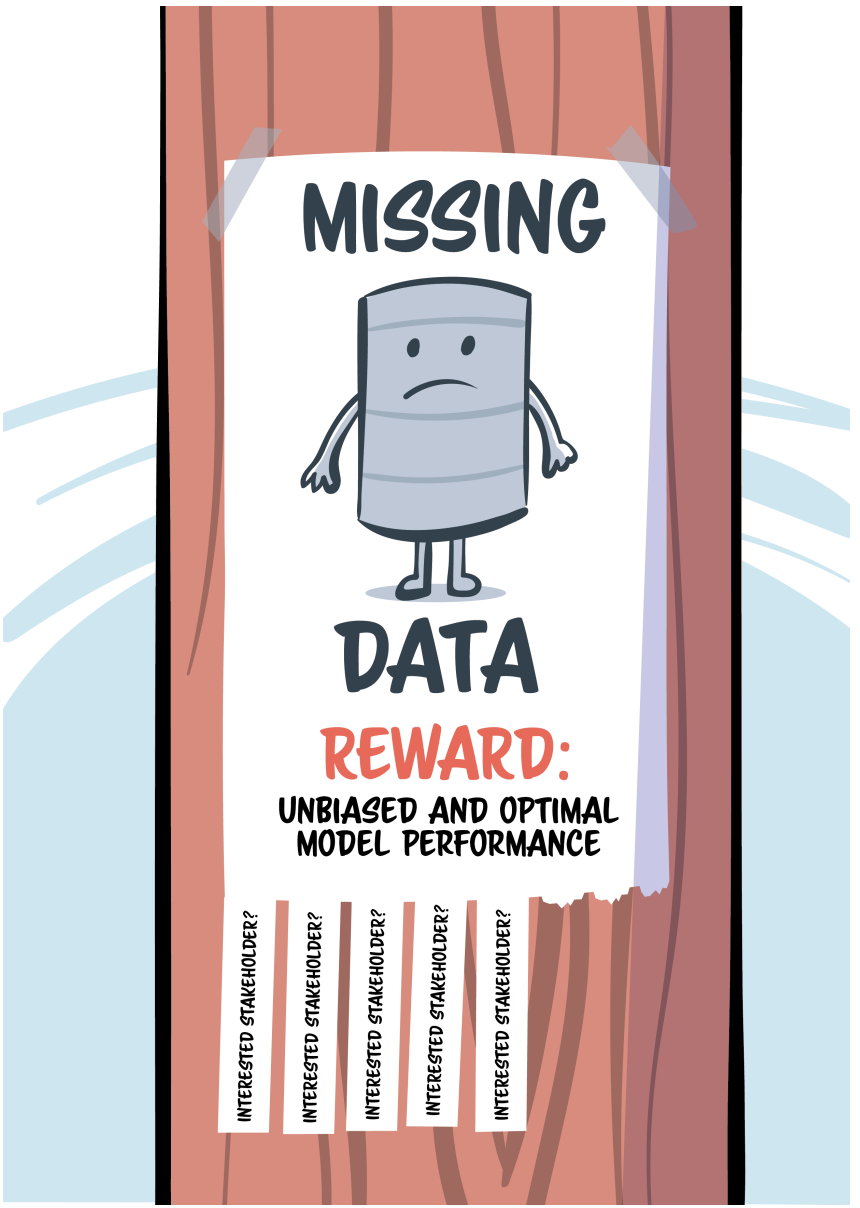
Data visualization

for incomplete datasets in R

Hanne Oberman

PhD candidate at Utrecht University

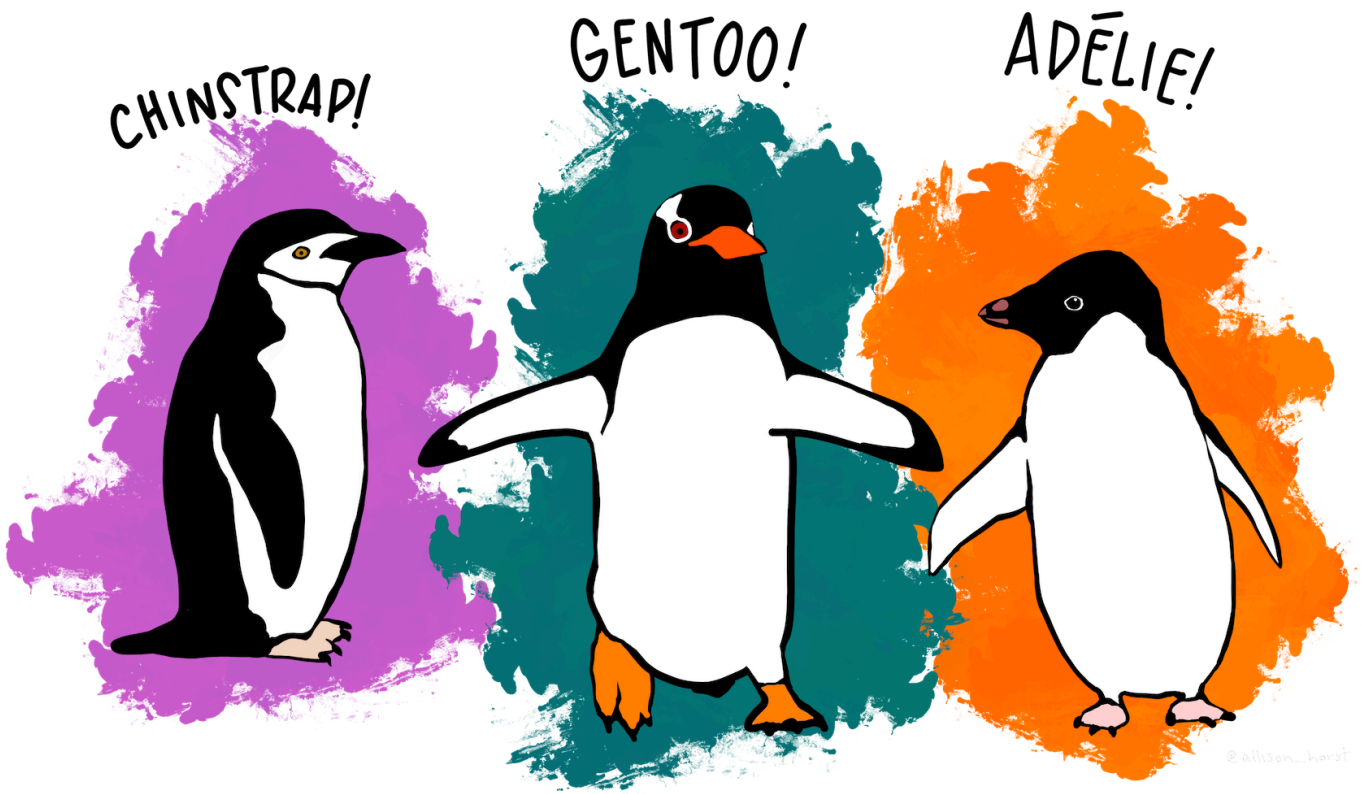
Missingness



Hanne Oberman, hanneoberman.github.io/presentations

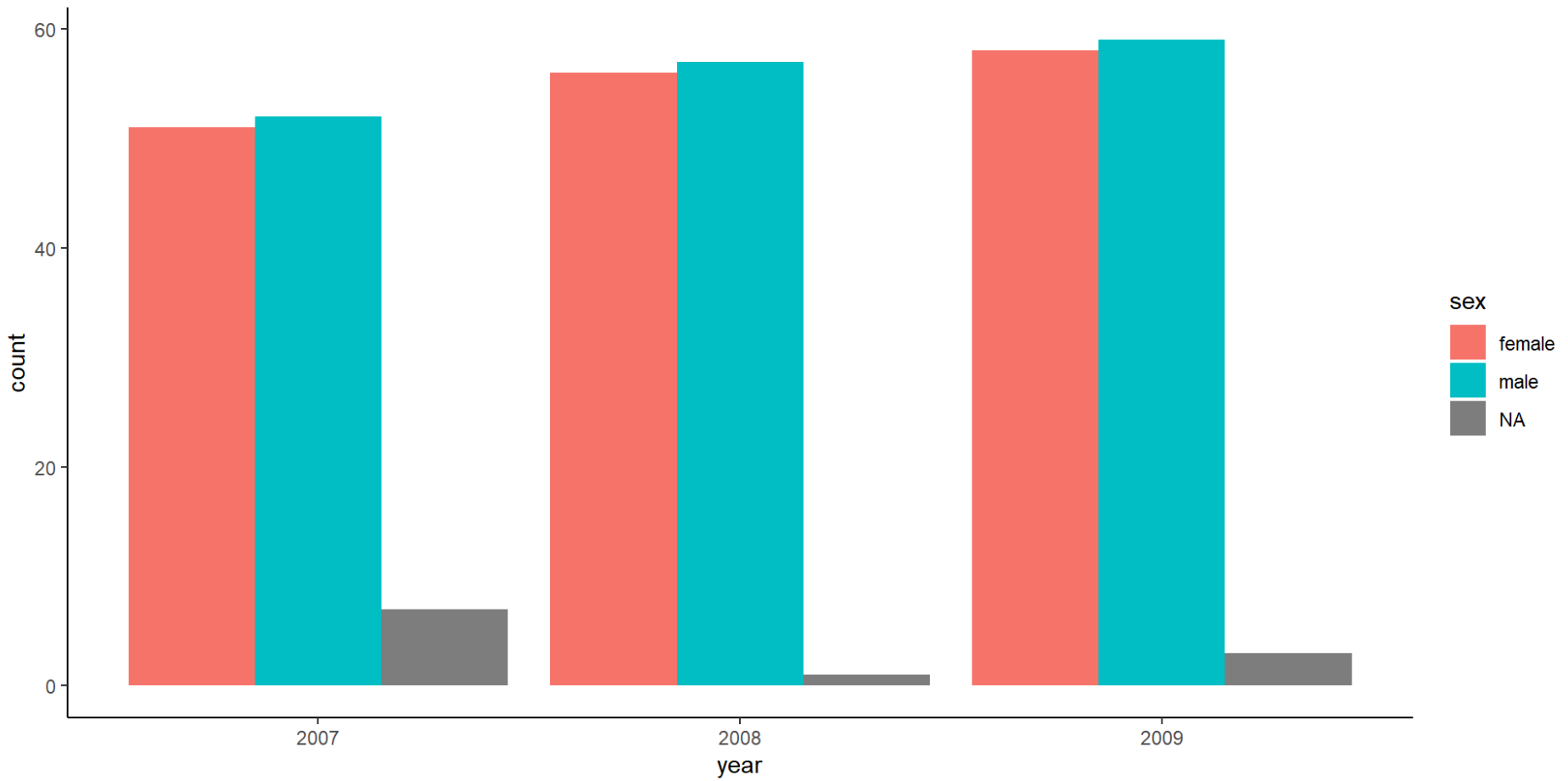
Case study

```
set.seed(123)
library(palmerpenguins)
library(mice)
library(ggmice)
library(ggplot2)
```



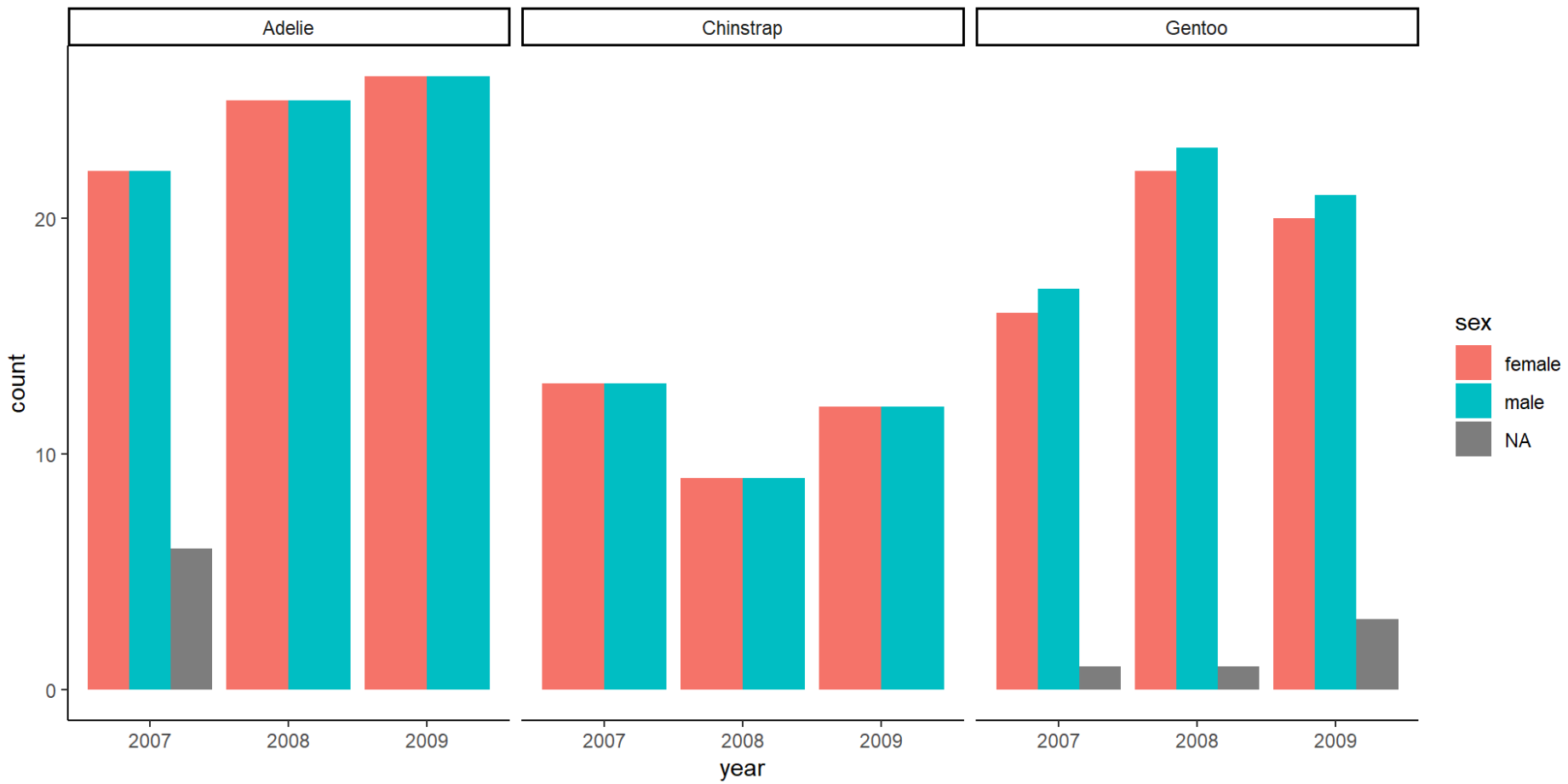
Penguin populations

▶ Code



Penguin populations

- Code



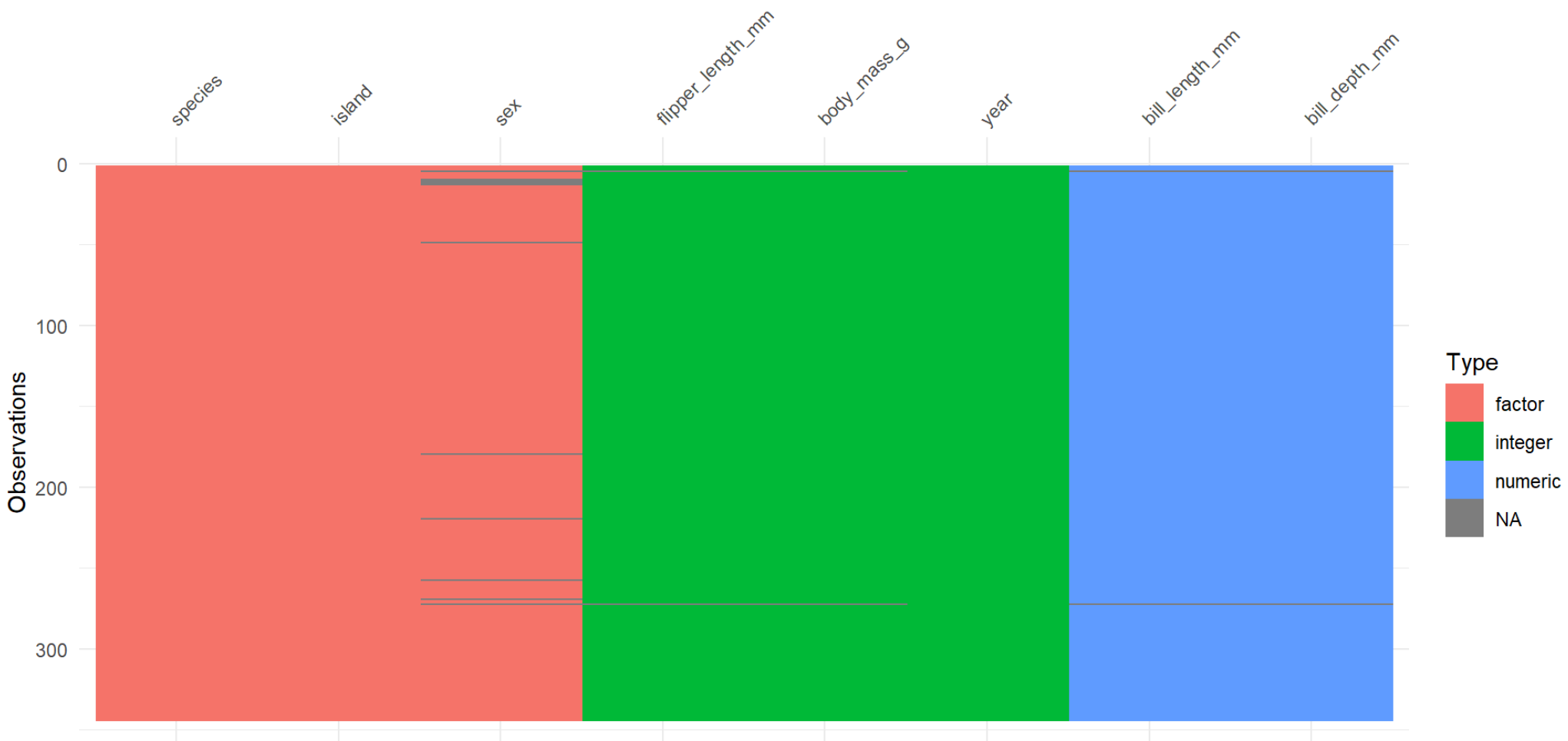
Incomplete data

```
1 str(penguins)
```

```
tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1
1 1 1 1 ...
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3
3 3 ...
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1
42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1
20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475
4250 ...
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA
NA ...
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007
2007
```


Incomplete data

```
1  visdat::vis_dat(penguins)
```



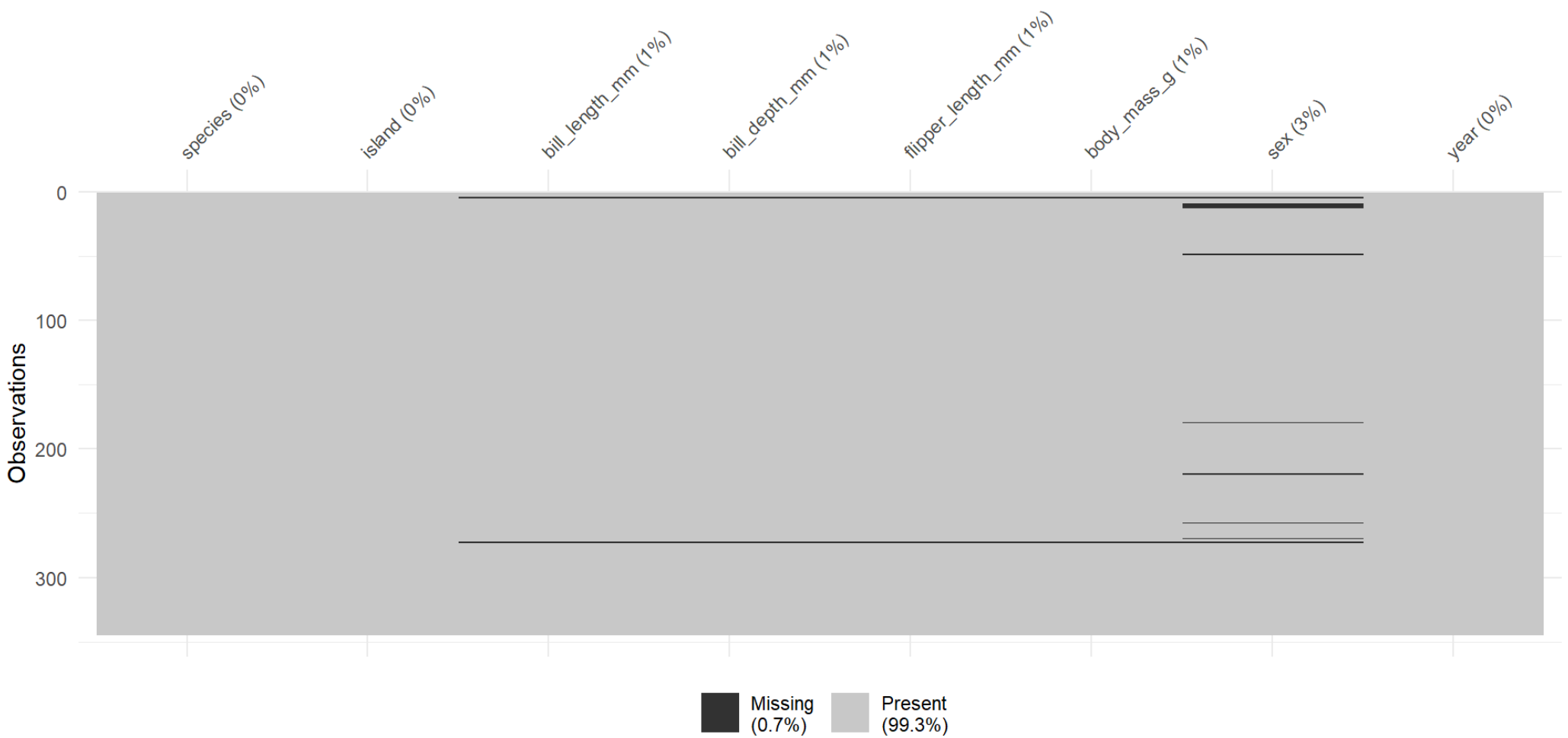
Response indicator

```
1 is.na(penguins)
```

```
      species island bill_length_mm bill_depth_mm flipper_length_mm
 [1,]   FALSE  FALSE          FALSE         FALSE             FALSE
 [2,]   FALSE  FALSE          FALSE         FALSE             FALSE
 [3,]   FALSE  FALSE          FALSE         FALSE             FALSE
 [4,]   FALSE  FALSE           TRUE          TRUE              TRUE
 [5,]   FALSE  FALSE          FALSE         FALSE             FALSE
 [6,]   FALSE  FALSE          FALSE         FALSE             FALSE
 [7,]   FALSE  FALSE          FALSE         FALSE             FALSE
 [8,]   FALSE  FALSE          FALSE         FALSE             FALSE
 [9,]   FALSE  FALSE          FALSE         FALSE             FALSE
[10,]   FALSE  FALSE          FALSE         FALSE             FALSE
[11,]   FALSE  FALSE          FALSE         FALSE             FALSE
[12,]   FALSE  FALSE          FALSE         FALSE             FALSE
[13,]   FALSE  FALSE          FALSE         FALSE             FALSE
[14,]   FALSE  FALSE          FALSE         FALSE             FALSE
```


Response indicator

```
1 naniar::vis_miss(penguins)
```



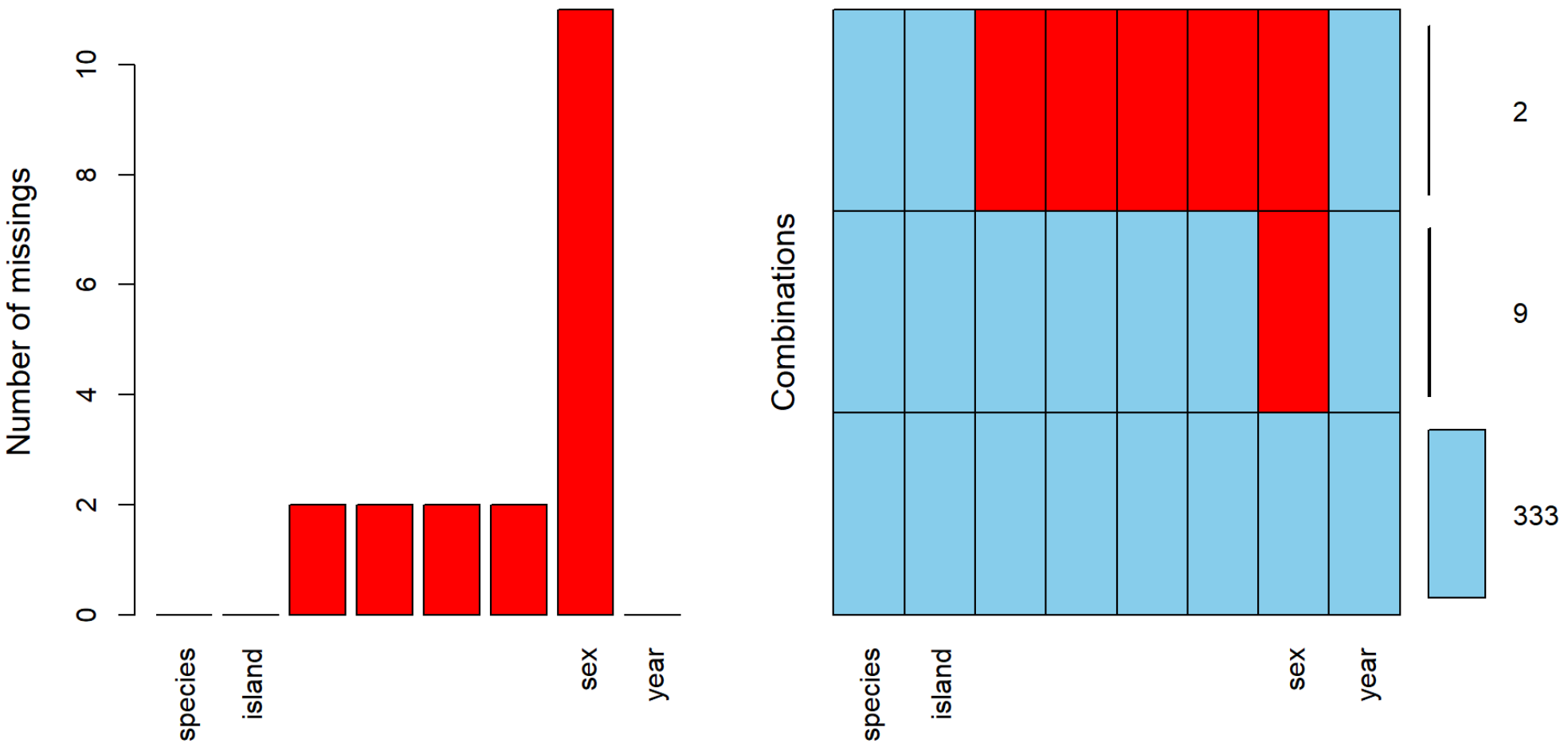
Missingness rate

```
1 colSums(is.na(penguins))
```

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


Missingness rate

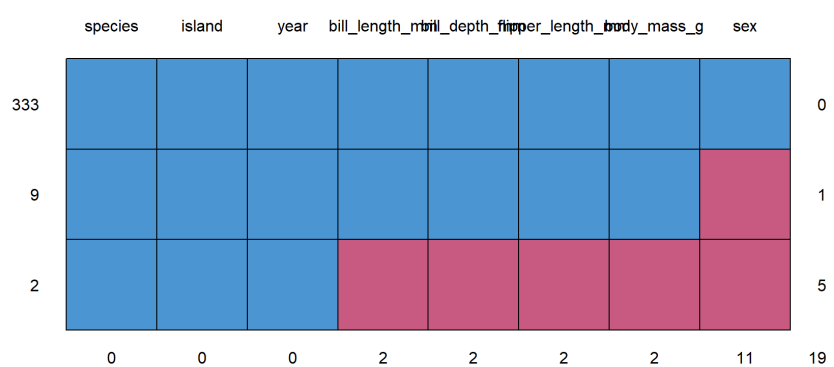
```
1  VIM::aggr(penguins, numbers = TRUE, prop = FALSE)
```



Missing data pattern

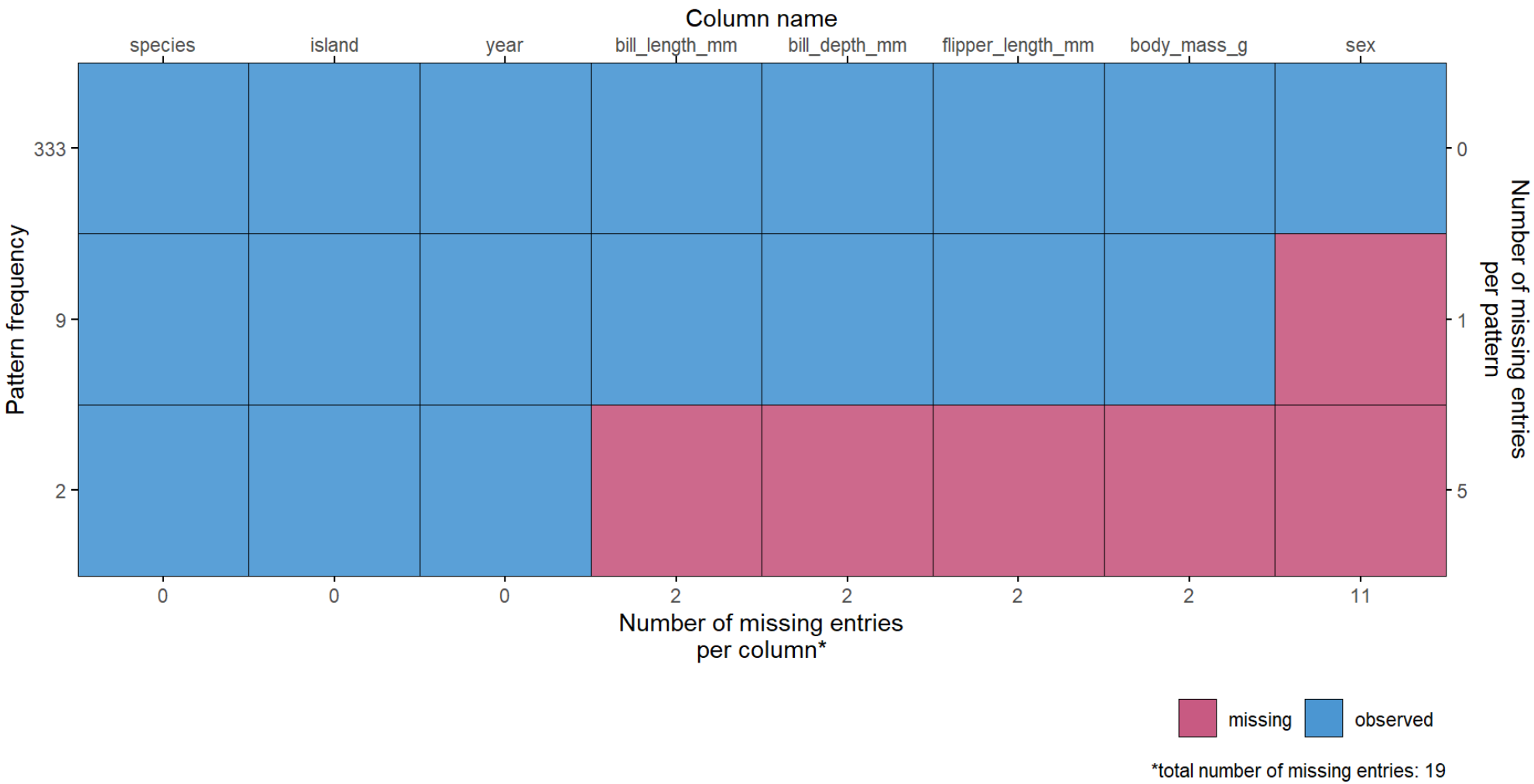
```
1 md.pattern(penguins)
```

```
    species island year bill_length_mm bill_depth_mm flipper_length_mm
333       1      1    1              1             1                  1
9         1      1    1              1             1                  1
2         1      1    1              0             0                  0
          0      0    0              2             2                  2
    body_mass_g sex
333           1   1   0
9             1   0   1
2             0   0   5
              2  11  19
```

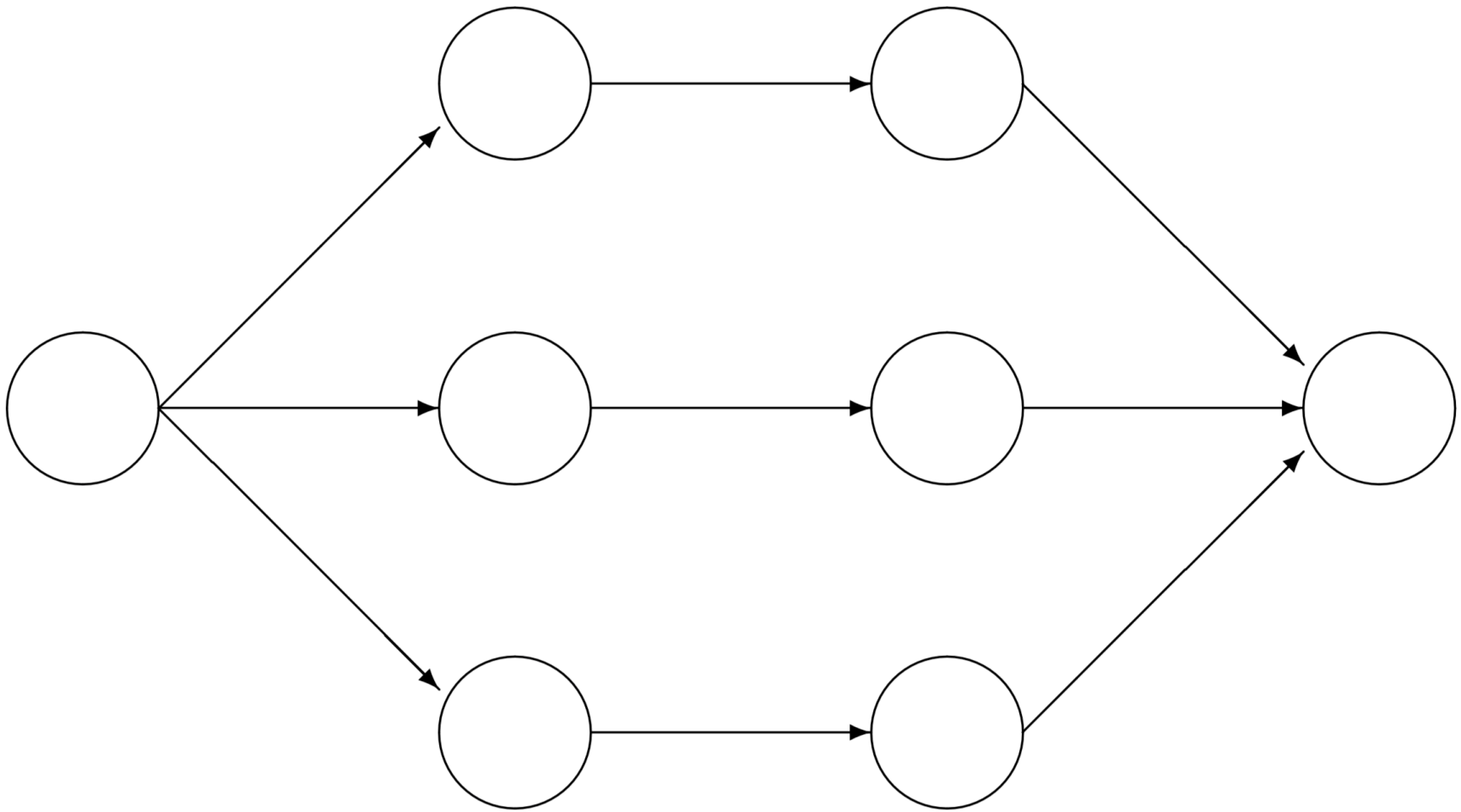


Missing data pattern

```
1 plot_pattern(penguins)
```



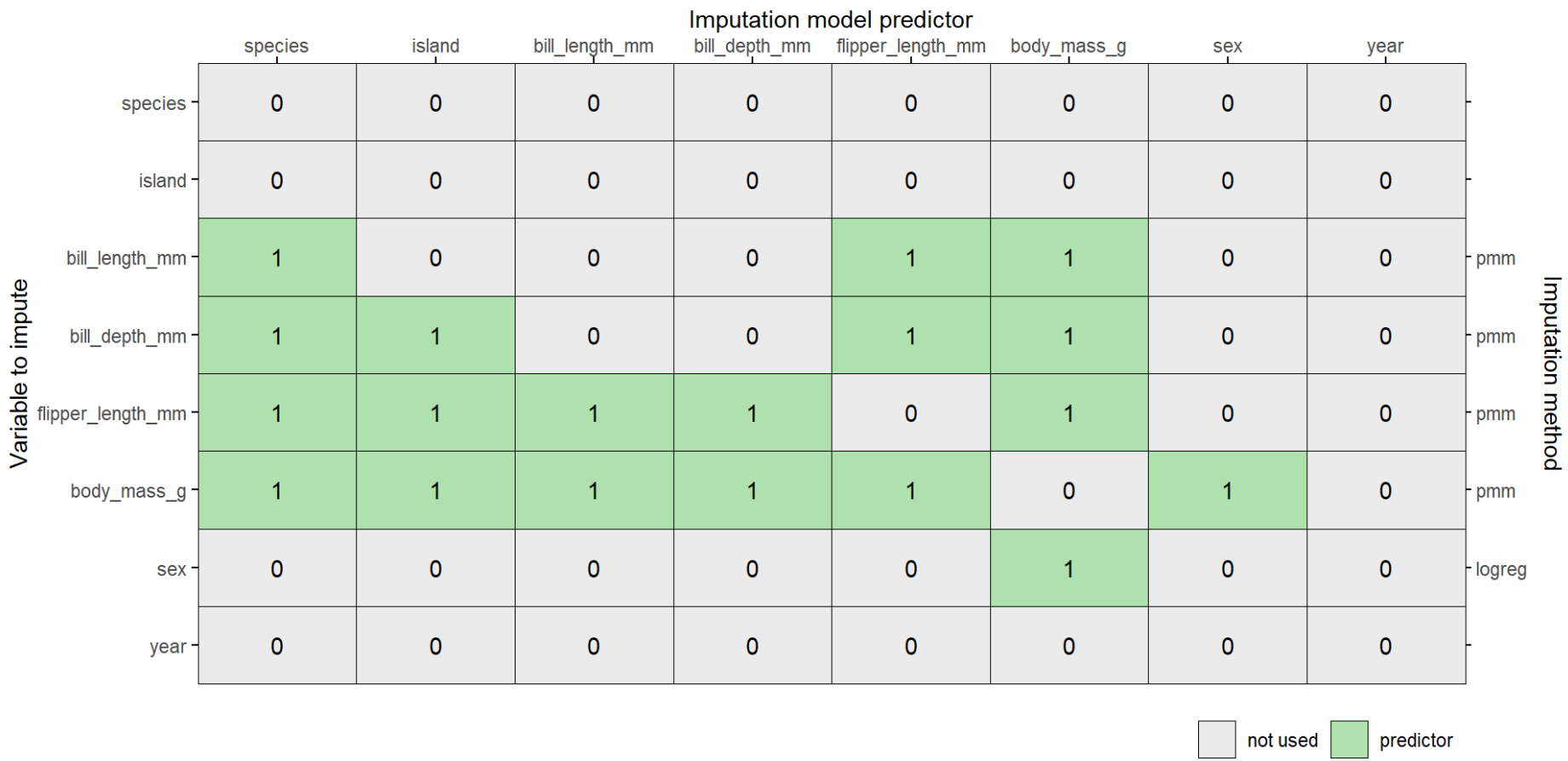
Imputation workflow



Incomplete data Imputed data Analysis results Pooled result

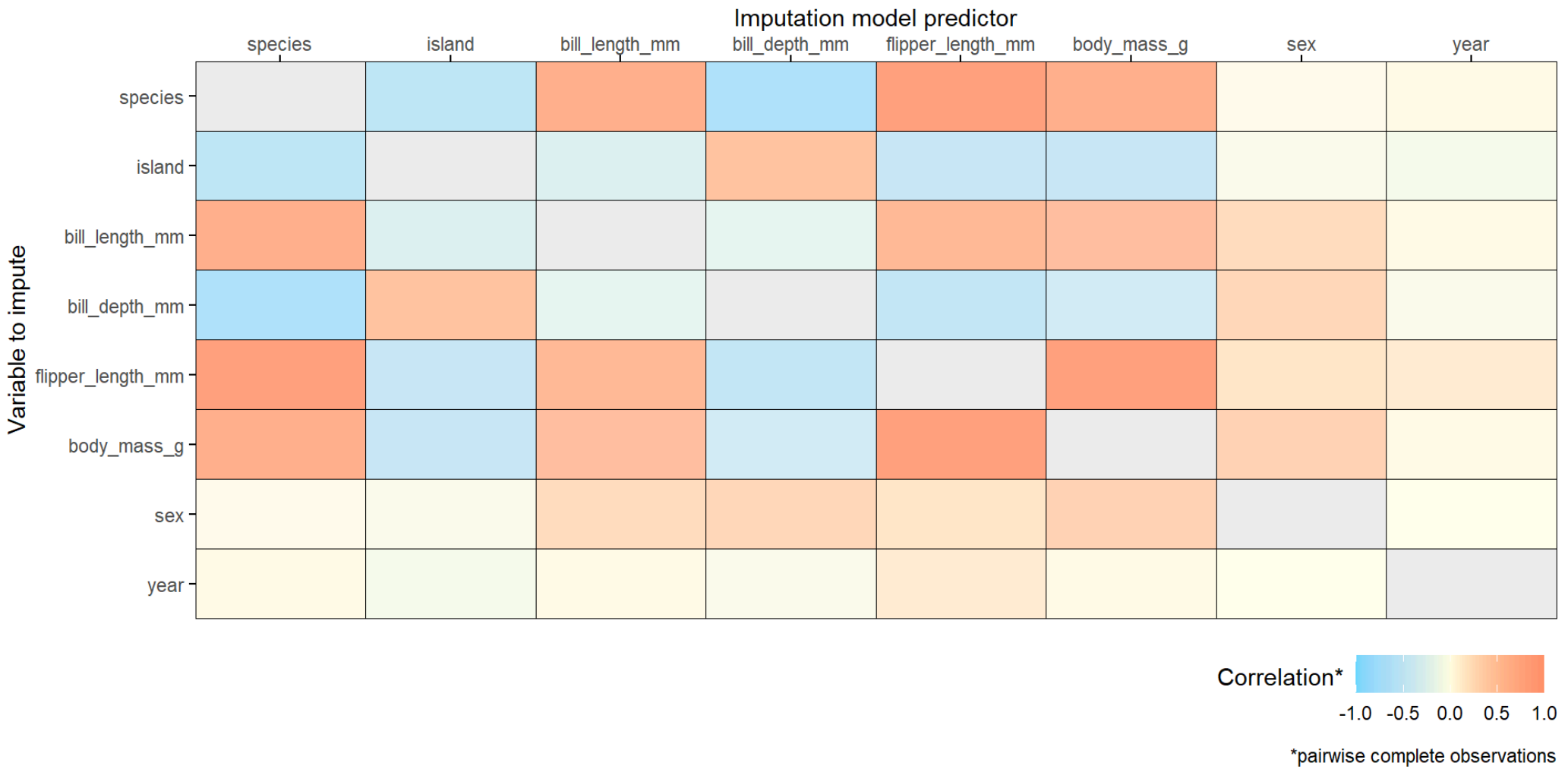
Imputation models

```
1 pred <- quickpred(penguins, mincor = 0.4)
2 meth <- make.method(penguins)
3 plot_pred(pred, method = meth, square = FALSE)
```



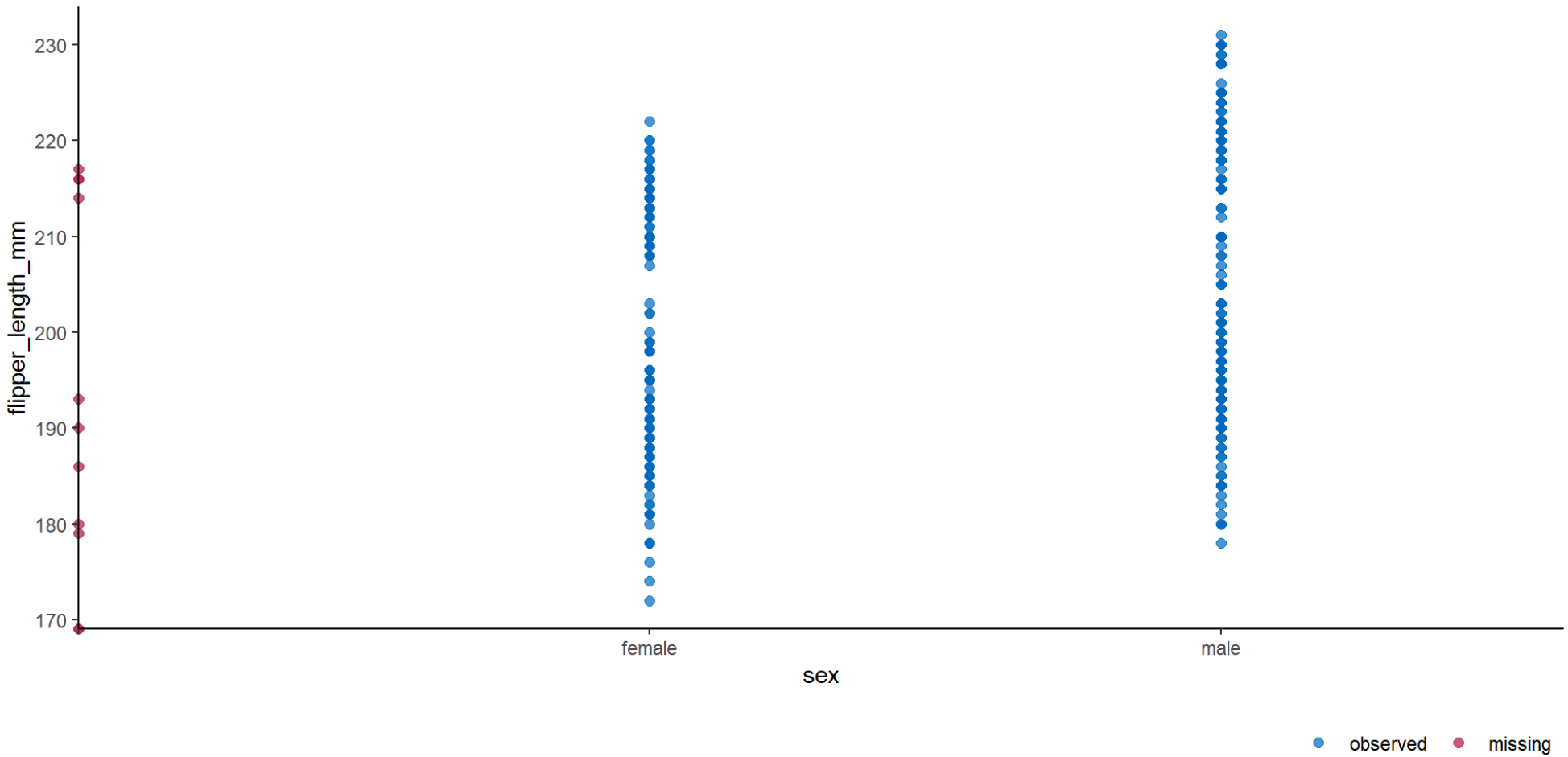
Correlation

```
1 plot_corr(penguins, square = FALSE)
```



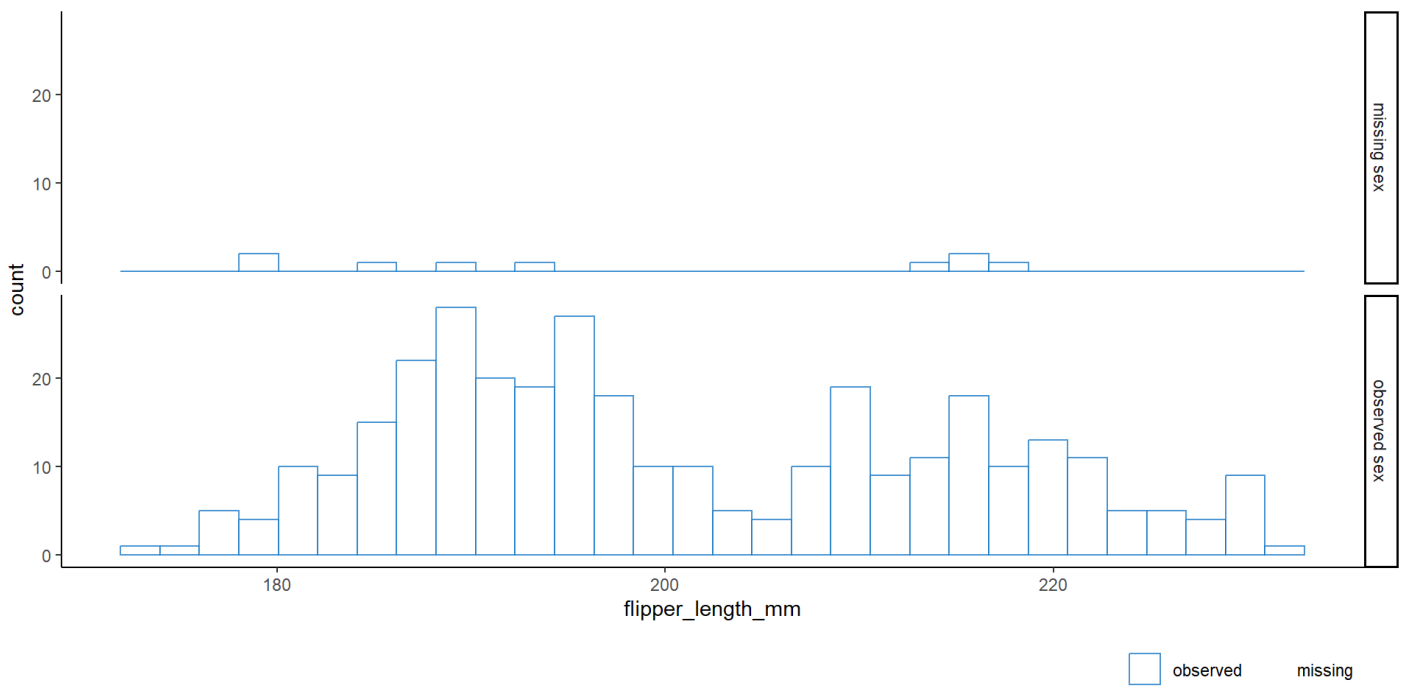
Scatter plot

```
1 ggmice(penguins, aes(sex, flipper_length_mm)) +
2   geom_point(size = 2)
```



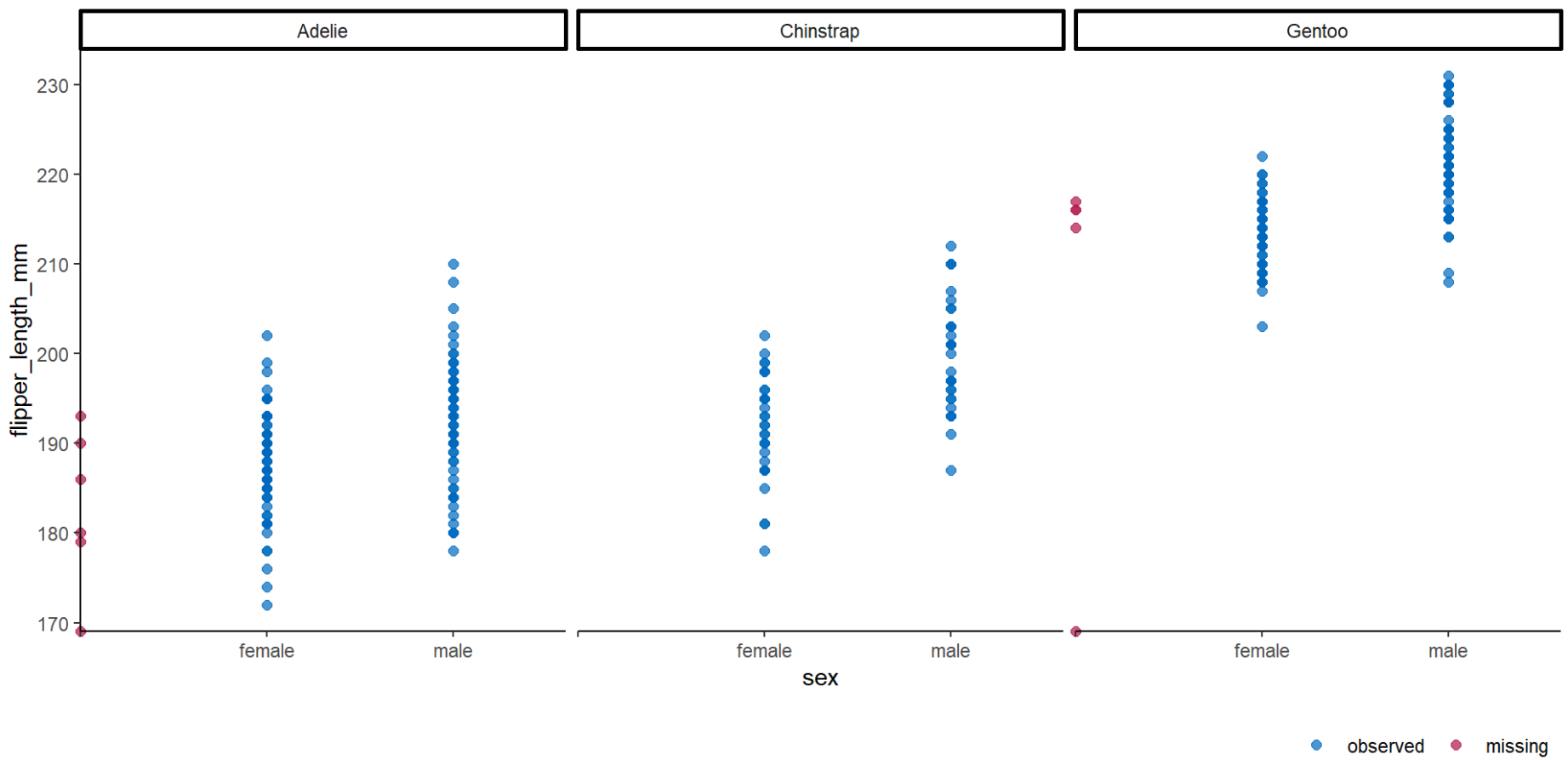
Faceted distribution

```
ggmice(penguins, aes(flipper_length_mm)) +
  geom_histogram(fill = "white") +
  facet_grid(factor(
    is.na(sex),
    levels = c(TRUE, FALSE),
    labels = c("missing sex", "observed sex")
    ) ~ .)
```



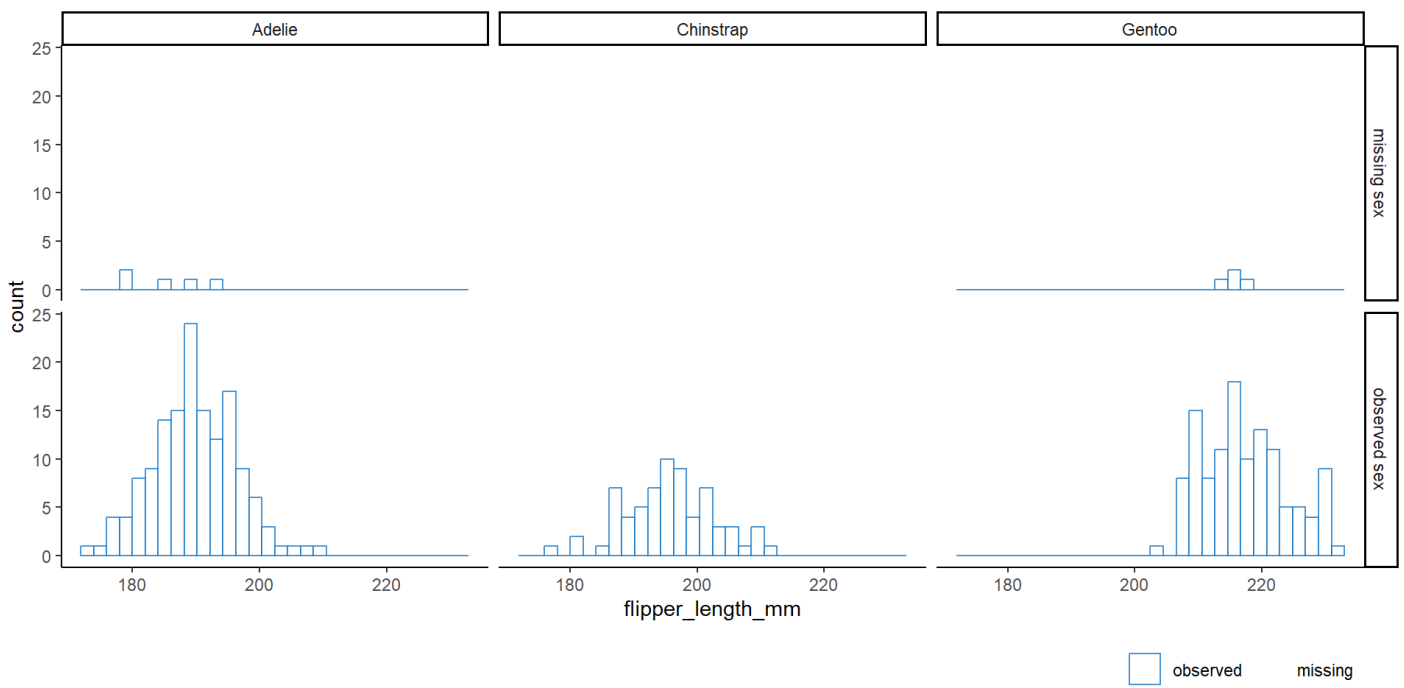
Faceted scatter plot

```
1  ggmice(penguins, aes(sex, flipper_length_mm)) +
2    geom_point(size = 2) +
3    facet_wrap(~species)
```



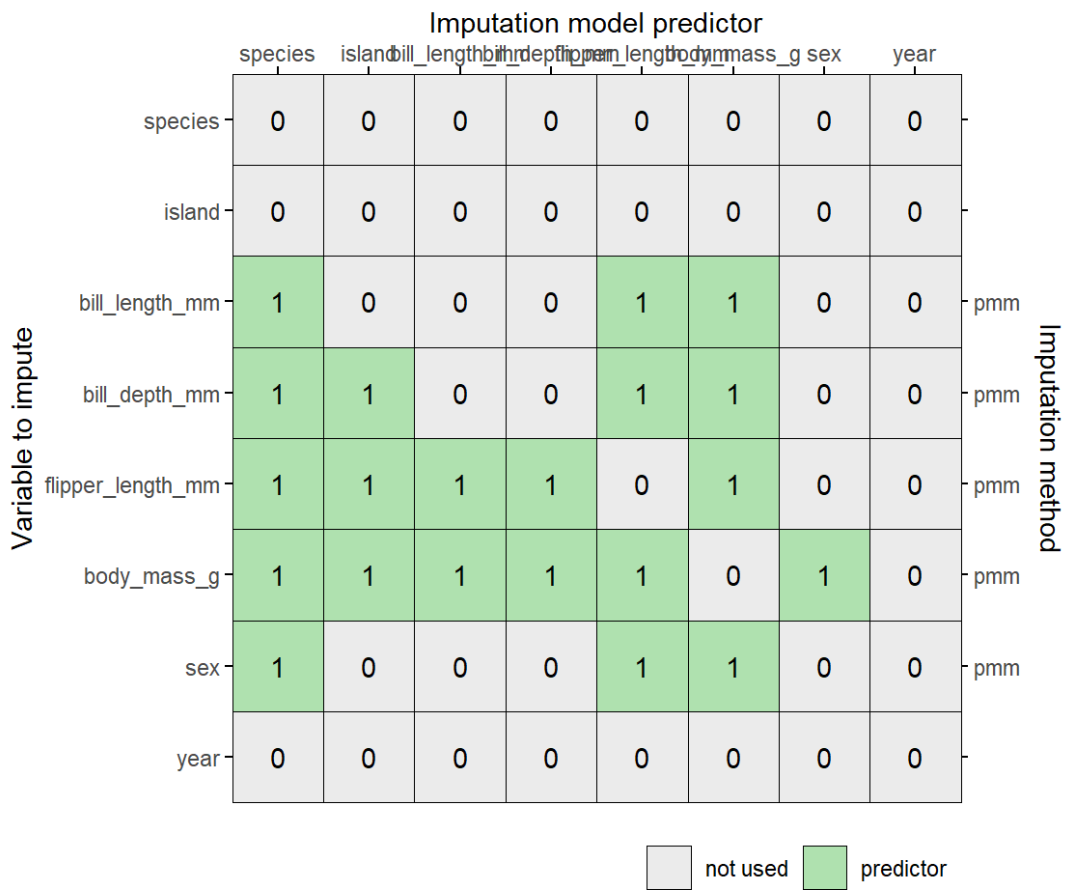
Faceted distribution

```
ggmice(penguins, aes(flipper_length_mm)) +
  geom_histogram(fill = "white") +
  facet_grid(factor(
    is.na(sex),
    levels = c(TRUE, FALSE),
    labels = c("missing sex", "observed sex")
    ) ~ species)
```



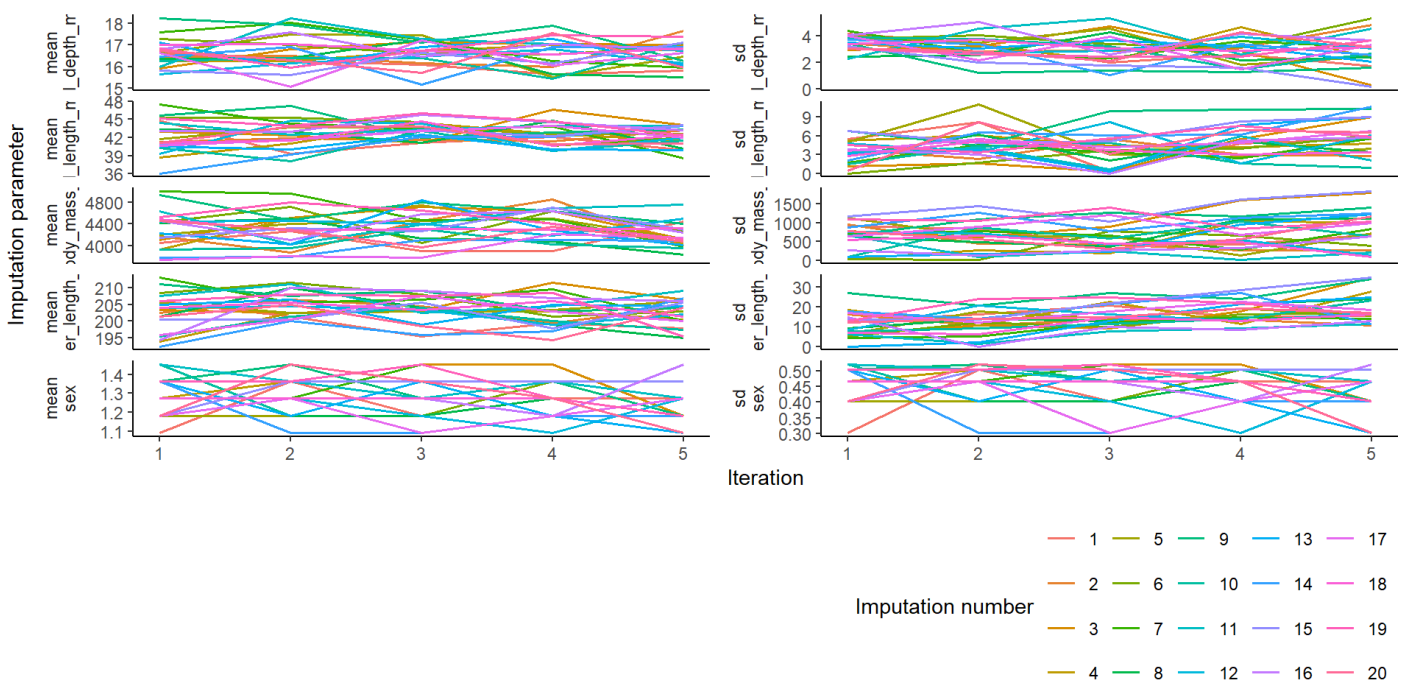
Adjust imputation models

```
1 pred["sex", c("species", "flipper_length_mm")] <- 1
2 meth["sex"] <- "pmm"
3 plot_pred(pred, method = meth)
```



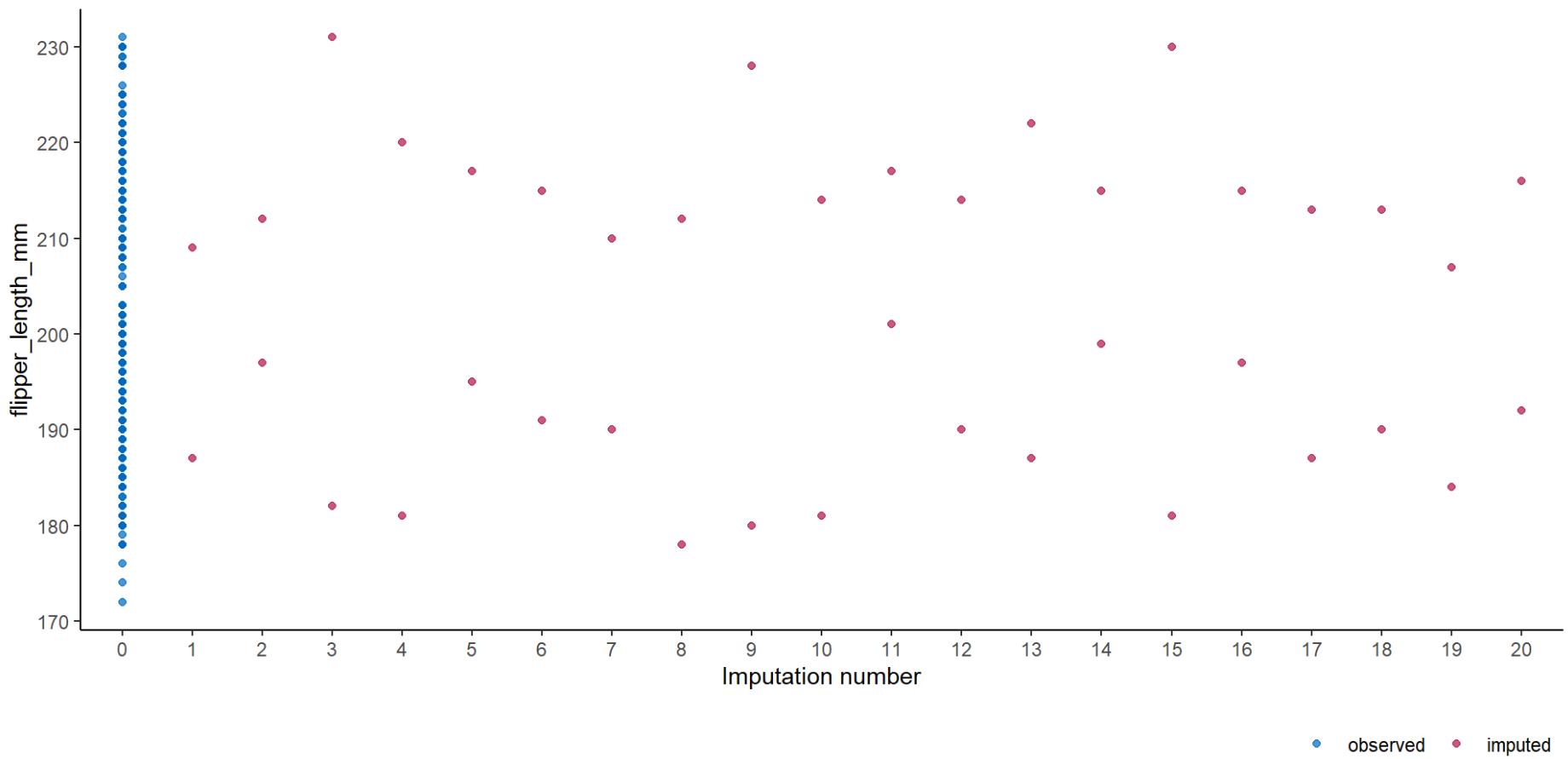
Impute

```
imp <- mice(
  penguins,
  pred = pred,
  method = meth,
  m = 20,
  print = FALSE)
plot_trace(imp)
```



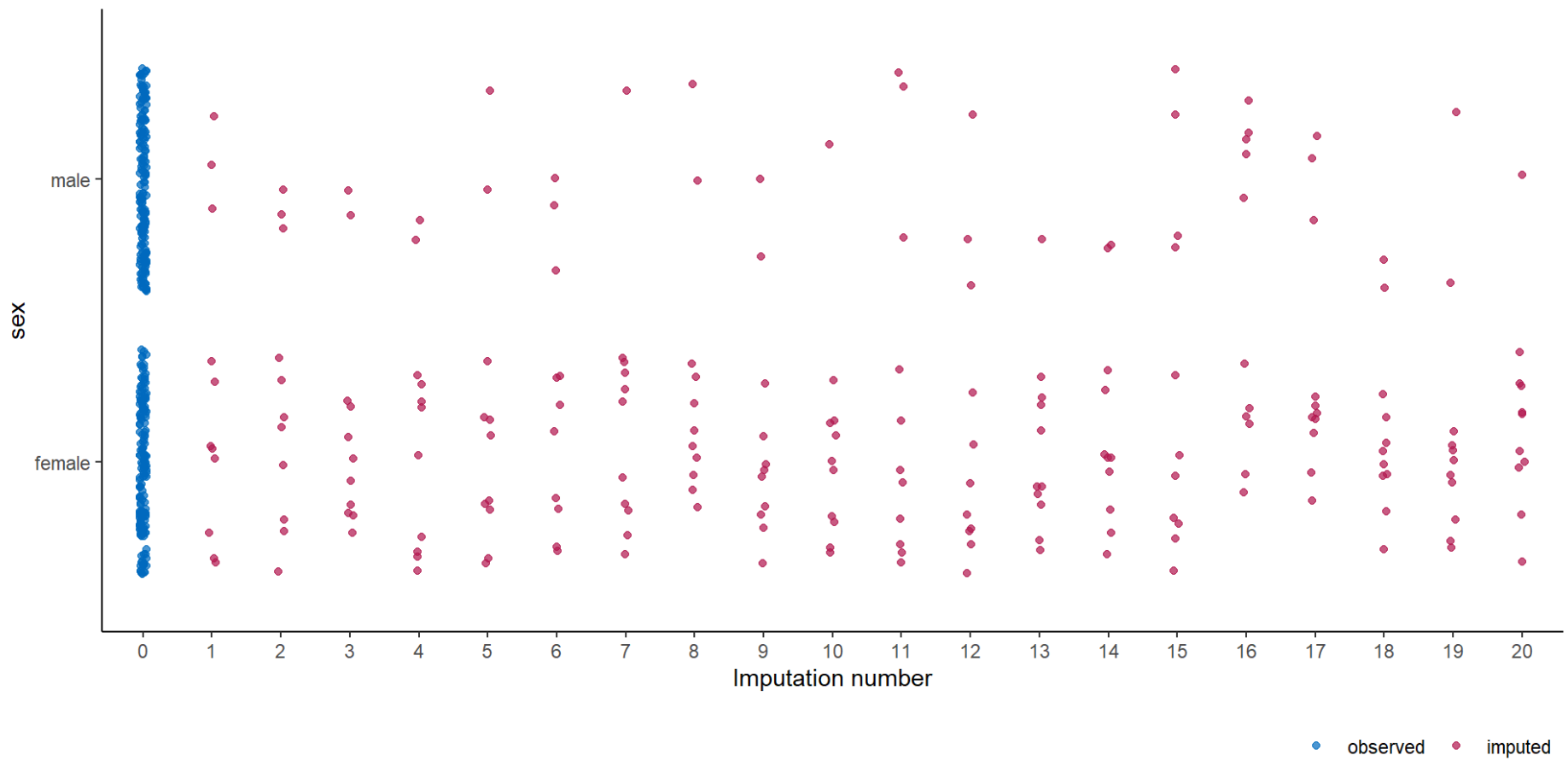
Stripplot

```
ggmice(imp, aes(x = .imp, y = flipper_length_mm)) +
  geom_point() +
  labs(x = "Imputation number")
```



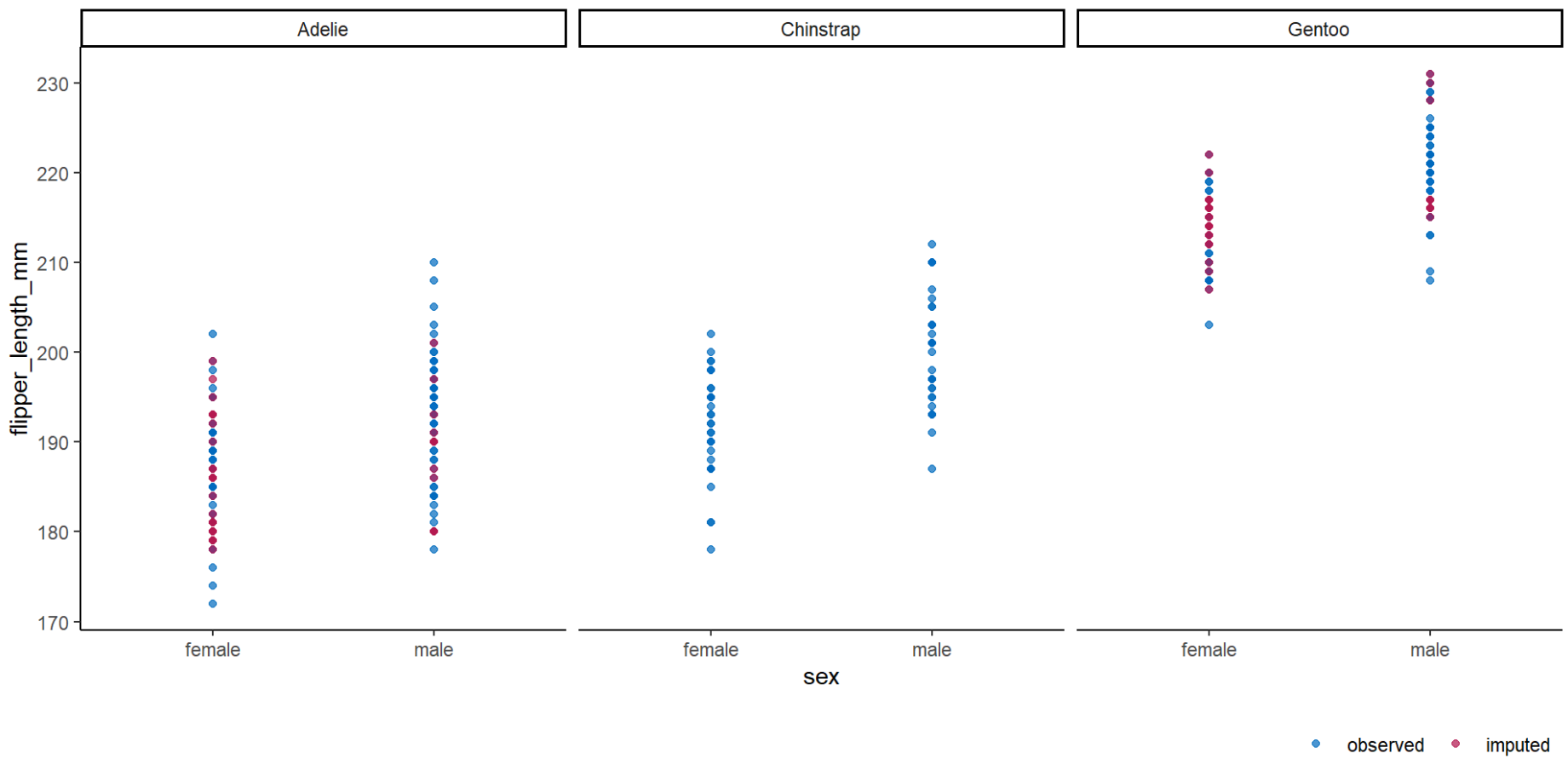
Stripplot

```
ggmice(imp, aes(x = .imp, y = sex)) +
  geom_jitter(width = 0.05) +
  labs(x = "Imputation number")
```



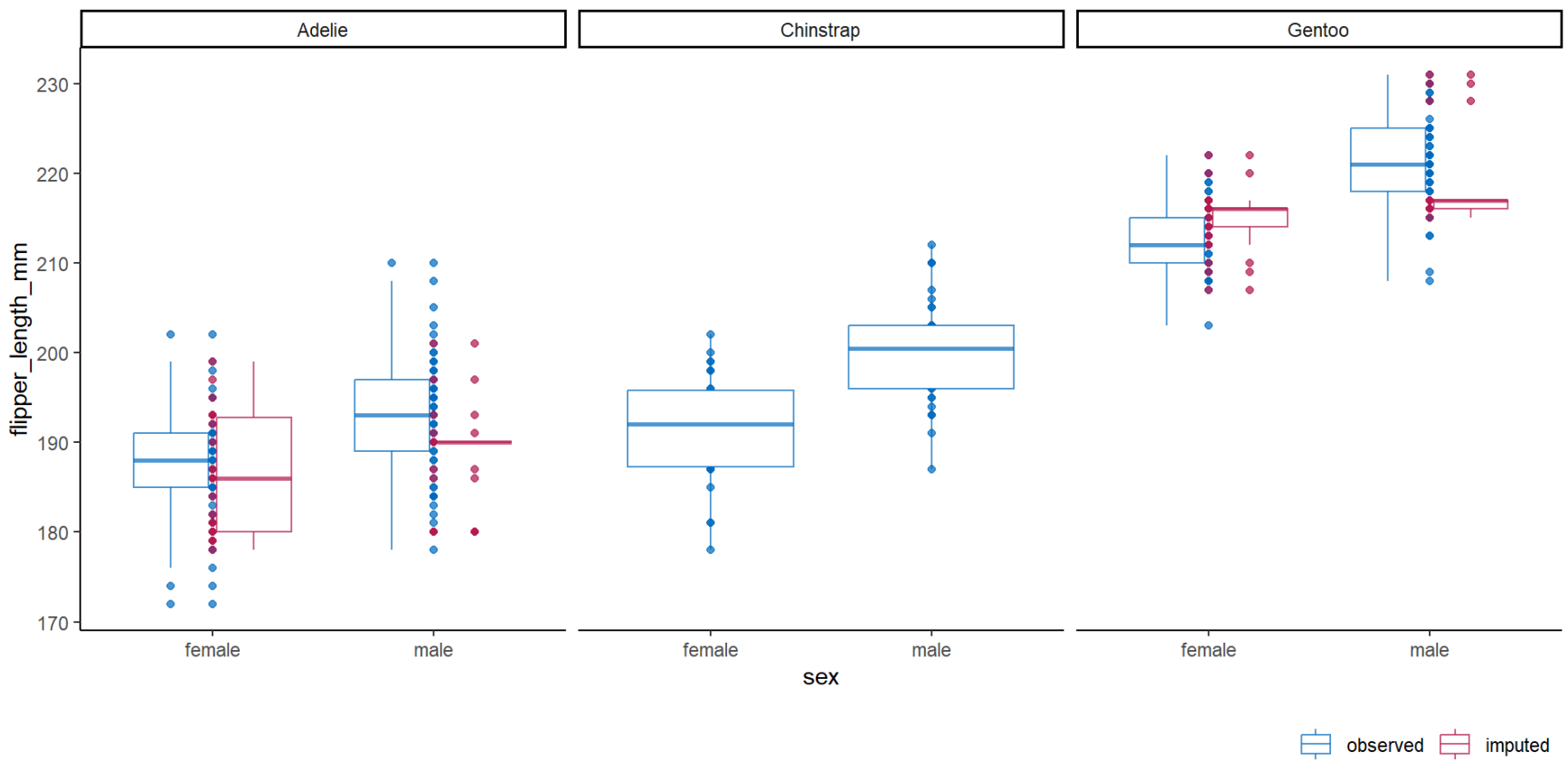
Scatter plot

```
1  ggmice(imp, aes(sex, flipper_length_mm)) +
2    geom_point() +
3    facet_grid(~species)
```



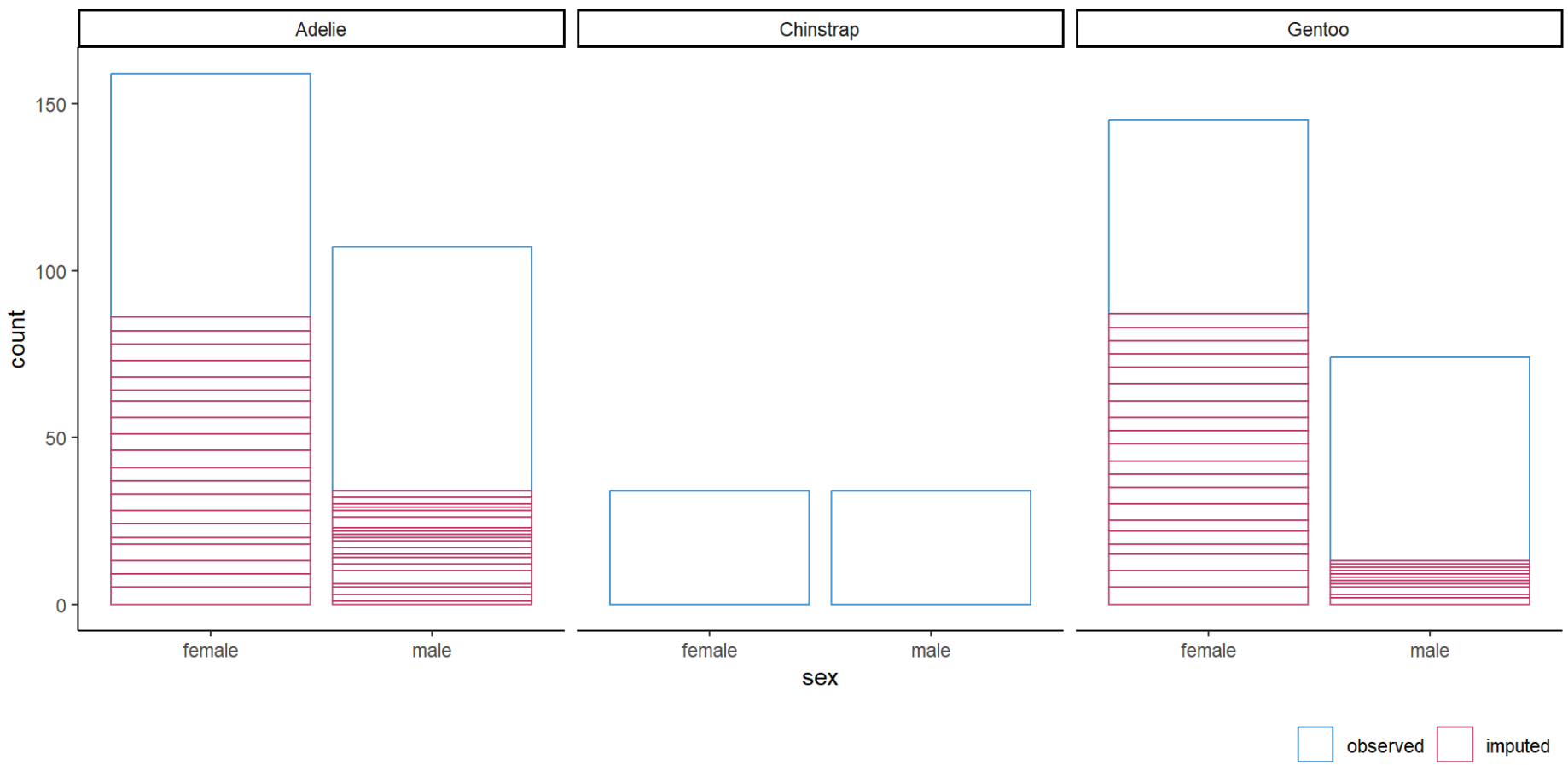
Scatter plot with boxplot

```
ggmice(imp, aes(sex, flipper_length_mm)) +
  geom_point() +
  geom_boxplot() +
  facet_grid(~species)
```



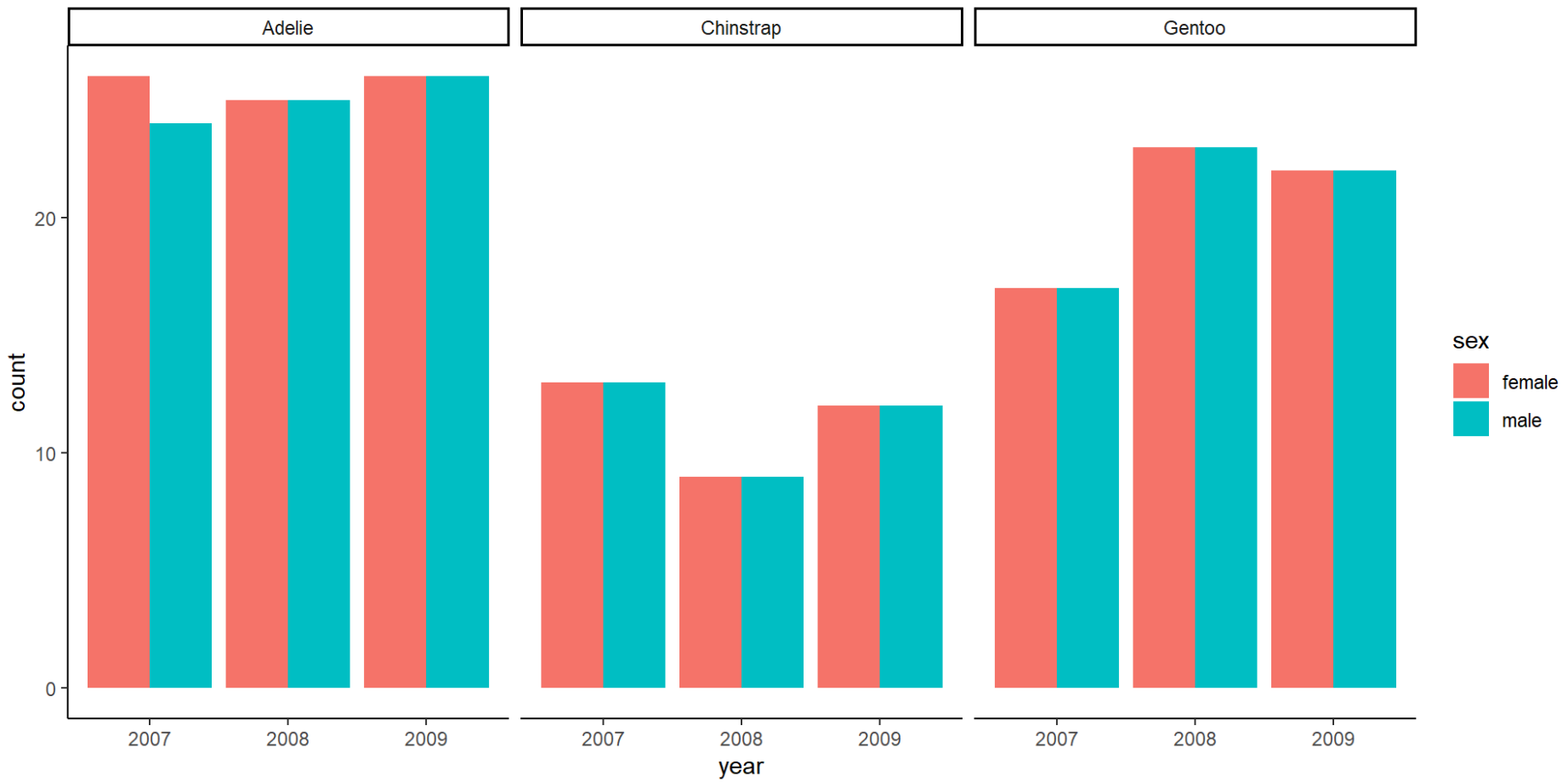
Faceted distribution

```
ggmice(imp, aes(sex, group = .imp)) +
  geom_histogram(fill = "white", stat="count") +
  facet_grid(~ species)
```



Populations after imputation

- Code



Thank you!