R-markdown learning

Hiep T.Bui

20 November 2023

Contents

1 Preprocessing

1.1 Choosing the data

This task will use the penguins dataset. Let's working with it

The source is taking from OpenML To catching anything, you need to

> quote out like this

to make an empty line, simply add ‘’ like this Penguins is cute!

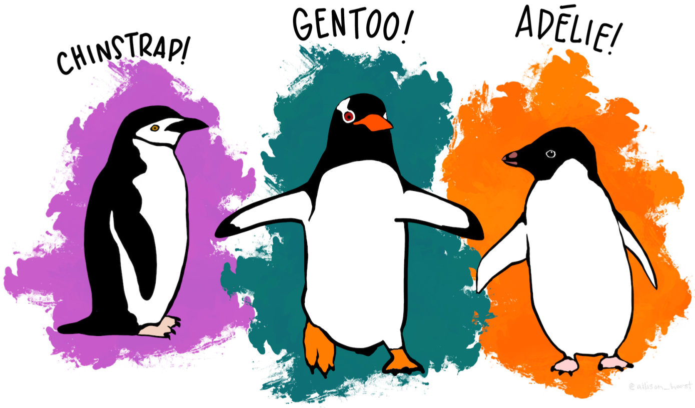
1.2 Loading the dataset, reading data

First, I would like to:

1. Load *packages*

2. Read the data

3. **Remove missing values**



To add an images, you use:

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'forcats' was built under R version 4.3.2
```

```
## Warning: package 'lubridate' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- 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 error
```

```
# load the dataset from OpenML Library
d <- OpenML::getOMLDataSet(data.id = 43347)
```

```
## Downloading from 'http://www.openml.org/api/v1/data/43347' to 'C:\Users\hiepa\AppData\Local\Temp\Rtm
## Downloading from 'https://api.openml.org/data/v1/download/22102172/Palmer-Penguins-Dataset-Alternati
```

```
# convert the OpenML object to a tibble (enhanced data.frame)
penguins <- d %>% dplyr::as_tibble()
#skimmed_penguins <- skimr::skim(penguins)
head(penguins)
```

```
## # A tibble: 6 x 7
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <chr>   <chr>              <dbl>         <dbl>             <dbl>       <dbl>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 1 more variable: sex <chr>
```

```
# Run again the code abow
library(tidyverse)
head(penguins)
```

```
## # A tibble: 6 x 7
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <chr>   <chr>              <dbl>         <dbl>             <dbl>       <dbl>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 1 more variable: sex <chr>
```

How the output and the code itself shows up in the document also be modified:

- ‘echo = FALSE’: only output is visible
- ‘include = FALSE’ hides both output and code
- ‘warning = FALSE’ to suppress warning message (also works with errors and message)
- ‘eval = FALSE’: code is not run

1.3 Removing missing data

```
# Before removing na_values
nrow(penguins)
```

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

```
# After removing
penguins <- penguins %>% drop_na()
nrow(penguins)
```

```
## [1] 333
```

Or we can write like this. I had removed missing values, so the data now only has 333 rows.

1.4 Descriptive statistics

The mean bill length is 43.9927928 mm. The bill depth is between 13.1 and 21.5

2 Graphs

Some ideas for the graphs:

- weight by flipper length
  - for the entire data
  - separately for each species
  - additionally by sex
- flipper and bill length

2.1