Data manipulation - solutions

Gbadamassi G.O. Dossa

Updated on 2023-11-12 (created on 2021-09-13)

Acknowledgements

The content of this module are based on materials from:

olivier gimenez's materials

Question 1a

```r
#read libraries
library(palmerpenguins)
library(tidyverse)
penguins # display data
```

```
## # A tibble: 344 × 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
##  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
##  7 Adelie  Torgersen           38.9          17.8               181        3625
##  8 Adelie  Torgersen           39.2          19.6               195        4675
##  9 Adelie  Torgersen           34.1          18.1               193        3475
## 10 Adelie  Torgersen           42            20.2               190        4250
## # ℹ 334 more rows
## # ℹ 2 more variables: sex <fct>, year <int>
```

```r
glimpse(penguins)
```

Question 1a

a. Display the data penguins.

```
penguins %>% glimpse() # display data
```

```
## Rows: 344
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel…
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse…
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, …
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, …
## $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186…
## $ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, …
## $ sex               <fct> male, female, female, NA, female, male, female, male…
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007…
```

Question 1b

b. Make sure you understand the columns we have in this dataset. c. Filter out penguins for which sex is missing. d. Select variables species, island, bill_length_mm and body_mass_g. e. Store the new dataset in a dat object.

```
dat <- penguins %>%
# filter out missing sex
  filter(!is.na(sex)) %>%
# select variables
  select(species, island, bill_length_mm, body_mass_g)

glimpse(dat)
```

```
## Rows: 333
## Columns: 4
## $ species        <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adeli
## $ island         <fct> Torgersen, Torgersen, Torgersen, Torgersen, T
```

Question 2a

a. How many penguins do we have in the dataset?

```
dat # nb of penguins
```

```
## # A tibble: 333 × 4
##    species island    bill_length_mm body_mass_g
##    <fct>   <fct>              <dbl>       <int>
##  1 Adelie  Torgersen           39.1        3750
##  2 Adelie  Torgersen           39.5        3800
##  3 Adelie  Torgersen           40.3        3250
##  4 Adelie  Torgersen           36.7        3450
##  5 Adelie  Torgersen           39.3        3650
##  6 Adelie  Torgersen           38.9        3625
##  7 Adelie  Torgersen           39.2        4675
##  8 Adelie  Torgersen           41.1        3200
##  9 Adelie  Torgersen           38.6        3800
## 10 Adelie  Torgersen           34.6        4400
## # ℹ 323 more rows
```

Questions 2b and 2d

b. How many species? d. Count the number of penguins per species.

```
# nb of species, and penguins per species
glimpse(dat)
```

```
## Rows: 333
## Columns: 4
## $ species        <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adeli
## $ island         <fct> Torgersen, Torgersen, Torgersen, Torgersen, T
## $ bill_length_mm <dbl> 39.1, 39.5, 40.3, 36.7, 39.3, 38.9, 39.2, 41.
## $ body_mass_g    <int> 3750, 3800, 3250, 3450, 3650, 3625, 4675, 320
```

```
levels(dat$species)
```

```
## [1] "Adelie"    "Chinstrap" "Gentoo"
```

Questions 2b and 2d

b. How many species? d. Count the number of penguins per species.

```
dat %>% count(species, sort = TRUE) # idem, arranged by n
```

```
## # A tibble: 3 × 2
##   species       n
##   <fct>     <int>
## 1 Adelie      146
## 2 Gentoo      119
## 3 Chinstrap    68
```

Question 2c

c. How many islands?

```r
dat %>% count(island) # nb of island, and penguins per island
```

```
## # A tibble: 3 × 2
##   island        n
##   <fct>     <int>
## 1 Biscoe      163
## 2 Dream       123
## 3 Torgersen    47
```

Question 2e

e. Count the number of penguins per species and per island.

```
# penguins per species and island
glimpse(dat)
```

```
## Rows: 333
## Columns: 4
## $ species        <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adeli
## $ island         <fct> Torgersen, Torgersen, Torgersen, Torgersen, T
## $ bill_length_mm <dbl> 39.1, 39.5, 40.3, 36.7, 39.3, 38.9, 39.2, 41.
## $ body_mass_g    <int> 3750, 3800, 3250, 3450, 3650, 3625, 4675, 320
```

```
dat %>% count(species, island)
```

```
## # A tibble: 5 × 3
```

Question 3a: mean body mass

a. Calculate the overall mean body mass.

```
# option 1
mean(penguins$body_mass_g) # Gives NA because the original data has missing values
```

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

```
mean(dat$body_mass_g)# here no more NA becuse we filtered out NA
```

```
## [1] 4207.057
```

```
a<-dat %>%
  mutate(mean_bm = mean(body_mass_g))
glimpse(dat)
```

```
## Rows: 333
## Columns: 4
## $ species        <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adelie,…
## $ island         <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, …
## $ bill_length_mm <dbl> 39.1, 39.5, 40.3, 36.7, 39.3, 38.9, 39.2, 41.1, 38.6, 3…
## $ body_mass_g    <int> 3750, 3800, 3250, 3450, 3650, 3625, 4675, 3200, 3800, 4…
```

Question 3a: mean body mass

```
# option 2
dat %>%
  summarise(mean_bm = mean(body_mass_g))
```

```
## # A tibble: 1 × 1
##   mean_bm
##     <dbl>
## 1   4207.
```

Question 3b: mean body mass per species

b. Calculate the mean body mass for each species.

```
# option 1
dat %>%
  group_by(species) %>%
  mutate(mean_bm = mean(body_mass_g))
```

```
## # A tibble: 333 × 5
## # Groups:   species [3]
##    species island    bill_length_mm body_mass_g mean_bm
##    <fct>   <fct>              <dbl>       <int>   <dbl>
##  1 Adelie  Torgersen           39.1        3750   3706.
##  2 Adelie  Torgersen           39.5        3800   3706.
##  3 Adelie  Torgersen           40.3        3250   3706.
##  4 Adelie  Torgersen           36.7        3450   3706.
##  5 Adelie  Torgersen           39.3        3650   3706.
##  6 Adelie  Torgersen           38.9        3625   3706.
##  7 Adelie  Torgersen           39.2        4675   3706.
##  8 Adelie  Torgersen           41.1        3200   3706.
##  9 Adelie  Torgersen           38.6        3800   3706.
## 10 Adelie  Torgersen           34.6        4400   3706.
## # ℹ 323 more rows
```

Question 3b: mean body mass per species

b. Calculate the mean body mass for each species.

```
# option 2
dat %>%
  group_by(species) %>%
  summarize(mean_bm = mean(body_mass_g))
```

```
## # A tibble: 3 × 2
##   species   mean_bm
##   <fct>       <dbl>
## 1 Adelie      3706.
## 2 Chinstrap   3733.
## 3 Gentoo      5092.
```

Question 3c: mean traits

c. Calculate the mean of both traits bill length and body mass measured for each species.

```r
# all at once, through column selection
dat %>%
  group_by(species) %>%
  summarize(across(bill_length_mm:body_mass_g, mean))
```

```
## # A tibble: 3 × 3
##   species   bill_length_mm body_mass_g
##   <fct>              <dbl>       <dbl>
## 1 Adelie              38.8       3706.
## 2 Chinstrap           48.8       3733.
## 3 Gentoo              47.6       5092.
```

Question 3c: mean traits

```
# all at once, through column format selection
glimpse(dat)
```

```
## Rows: 333
## Columns: 4
## $ species        <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adeli
## $ island         <fct> Torgersen, Torgersen, Torgersen, Torgersen, T
## $ bill_length_mm <dbl> 39.1, 39.5, 40.3, 36.7, 39.3, 38.9, 39.2, 41.
## $ body_mass_g    <int> 3750, 3800, 3250, 3450, 3650, 3625, 4675, 320
```

```
dat3<-penguins%>%
  filter(!is.na(sex))%>%
  select(species, island, sex, bill_length_mm, body_mass_g)
glimpse(dat3)
```