

# Data transformation with dplyr

## Practice with penguins

HARON OTIM

### Table of contents

Task 1: Create a subset of the data using filter() . . . . .	2
Task 2: Add new columns with mutate() . . . . .	6
Task 3: Summarize data with group_by() and summarize() & count() . . . . .	9
Task 7: Data communication . . . . .	11
Task 8: Stage, Commit & Push to GitHub . . . . .	11

```
library(palmerpenguins)
library(dplyr)
```

All exercises in this assignment use the `penguins` data as a starting point.

1. Run all code chunks above.
2. Run the code chunk that contains `glimpse(penguins)`.
3. How many variables are in the data set? #Haron: There are 8 variables in the data set evident in the number of columns.
4. How many observations are in the data set? #Haron: There are 344 observations in the data set evident by 344 number of rows.
5. What data types are contained in the variables? (Reminder: <https://ds4owd-001.github.io/website/slides/lec-02-visualisation.html#/types-of-variables>) #Haron: species (categorical variable) island (categorical variable) bill\_length\_mm (continuous variable) bill\_depth\_mm (continuous variable) flipper\_length\_mm (discrete variable) body\_mass\_g (discrete variable) sex (categorical variables) year (continuous variable)

```
glimpse(penguins)
```

```
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~
```

## Task 1: Create a subset of the data using filter()

Use `filter()` to create a subset from `penguins` that only contains observations for Adelie penguins.

```
penguins |>
  filter(species == "Adelie")
```

```
# A tibble: 152 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
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 142 more rows
# i 2 more variables: sex <fct>, year <int>
```

Use `filter()` to create a subset from `penguins` that only contains observations where body mass is less than or equal to 2900 g.

```
penguins |>
  filter(body_mass_g <= 2900)
```

```
# A tibble: 7 x 8
  species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>         <dbl>         <dbl>         <int>         <int>
1 Adelie  Biscoe           34.5           18.1           187           2900
2 Adelie  Biscoe           36.5           16.6           181           2850
3 Adelie  Biscoe           36.4           17.1           184           2850
4 Adelie  Dream            33.1           16.1           178           2900
5 Adelie  Torgersen          38.6            17           188           2900
6 Chinstrap Dream          43.2           16.6           187           2900
7 Chinstrap Dream          46.9           16.6           192           2700
# i 2 more variables: sex <fct>, year <int>
```

Use `filter()` to create a subset from `penguins` that only contains observations for Adelie penguins with a bill length greater than 40 mm.

```
penguins |>
  filter(species == "Adelie" & bill_length_mm > 40)
```

```
# A tibble: 51 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          40.3            18           195           3250
2 Adelie  Torgersen          42            20.2           190           4250
3 Adelie  Torgersen          41.1           17.6           182           3200
4 Adelie  Torgersen          42.5           20.7           197           4500
5 Adelie  Torgersen          46            21.5           194           4200
6 Adelie  Biscoe          40.6           18.6           183           3550
7 Adelie  Biscoe          40.5           17.9           187           3200
8 Adelie  Biscoe          40.5           18.9           180           3950
9 Adelie  Dream           40.9           18.9           184           3900
10 Adelie Dream           42.2           18.5           180           3550
# i 41 more rows
# i 2 more variables: sex <fct>, year <int>
```

Use `filter()` to create a subset from `penguins` that excludes observations for chinstraps.

```
penguins |>
  filter(species != "Chinstrap")
```

```
# A tibble: 276 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
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 266 more rows
# i 2 more variables: sex <fct>, year <int>
```

Use `filter()` to create a subset from `penguins` that only contains gentoo penguins with a bill depth greater than or equal to 15.5 millimeters.

```
penguins |>
  filter(species == "gentoo", bill_depth_mm >= 15.5)
```

```
# A tibble: 0 x 8
# i 8 variables: species <fct>, island <fct>, bill_length_mm <dbl>,
#   bill_depth_mm <dbl>, flipper_length_mm <int>, body_mass_g <int>, sex <fct>,
#   year <int>
```

Use `filter()` to create a subset from `penguins` that contains observations for male penguins recorded at Dream and Biscoe Islands.

```
penguins |>
  filter(sex == "male", island %in% c("Dream","Biscoe"))
```

```
# A tibble: 145 x 8
  species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>         <dbl>         <dbl>         <int>         <int>
1 Adelie  Biscoe          37.7          18.7          180          3600
2 Adelie  Biscoe          38.2          18.1          185          3950
3 Adelie  Biscoe          38.8          17.2          180          3800
4 Adelie  Biscoe          40.6          18.6          183          3550
```

```

5 Adelie Biscoe          40.5          18.9          180          3950
6 Adelie Dream           37.2          18.1          178          3900
7 Adelie Dream           40.9          18.9          184          3900
8 Adelie Dream           39.2          21.1          196          4150
9 Adelie Dream           38.8          20           190          3950
10 Adelie Dream          39.8          19.1          184          4650
# i 135 more rows
# i 2 more variables: sex <fct>, year <int>

```

Use `filter()` to create a subset from `penguins` that contains observations for female Adelie penguins with bill lengths less than 35 mm.

```

penguins |>
  filter(sex == "female" & species == "Adelie" & bill_length_mm < 35)

# A tibble: 7 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     34.4          18.4          184          3325
2 Adelie Biscoe       34.5          18.1          187          2900
3 Adelie Torgersen     33.5          19           190          3600
4 Adelie Torgersen     34.6          17.2          189          3200
5 Adelie Dream        34           17.1          185          3400
6 Adelie Dream        33.1          16.1          178          2900
7 Adelie Dream        32.1          15.5          188          3050
# i 2 more variables: sex <fct>, year <int>

```

Use `filter()` to create a subset from `penguins` containing observations for female chinstrap penguins on Dream and Torgersen Islands.

```

penguins |>
  filter(sex == "female", species == "Chinstrap", island %in% c("Dream","Torgersen"))

# A tibble: 34 x 8
  species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>         <dbl>         <dbl>         <int>         <int>
1 Chinstrap Dream     46.5          17.9          192          3500
2 Chinstrap Dream     45.4          18.7          188          3525
3 Chinstrap Dream     45.2          17.8          198          3950
4 Chinstrap Dream     46.1          18.2          178          3250

```

```

5 Chinstrap Dream      46      18.9      195      4150
6 Chinstrap Dream      46.6     17.8      193      3800
7 Chinstrap Dream      47      17.3      185      3700
8 Chinstrap Dream      45.9     17.1      190      3575
9 Chinstrap Dream      58      17.8      181      3700
10 Chinstrap Dream     46.4     18.6      190      3450
# i 24 more rows
# i 2 more variables: sex <fct>, year <int>

```

Use `filter()` to create a subset from `penguins` that contains penguins that are either gentoos OR have a body mass greater than 4500 g.

```

penguins |>
  filter(species == "Gentoo" | body_mass_g > 4500)

# A tibble: 133 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        39.2          19.6           195         4675
2 Adelie  Dream           39.8          19.1           184         4650
3 Adelie  Dream           39.6          18.8           190         4600
4 Adelie  Torgersen        42.9          17.6           196         4700
5 Adelie  Biscoe          41           20            203         4725
6 Adelie  Biscoe          43.2          19            197         4775
7 Adelie  Biscoe          45.6          20.3           191         4600
8 Gentoo  Biscoe          46.1          13.2           211         4500
9 Gentoo  Biscoe          50           16.3           230         5700
10 Gentoo Biscoe          48.7          14.1           210         4450
# i 123 more rows
# i 2 more variables: sex <fct>, year <int>

```

## Task 2: Add new columns with `mutate()`

Add a column to `penguins` that contains a new column `flipper_m`, which is the `flipper_length_mm` (flipper length in millimeters) converted to units of meters.

```

penguins |>
  mutate(flipper_m = flipper_length_mm / 1000)

```

```
# A tibble: 344 x 9
  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 3 more variables: sex <fct>, year <int>, flipper_m <dbl>
```

Add a new column to `penguins` that contains a new column `body_mass_kg`, which is the `body_mass_g` (body mass in grams) converted to units of kilograms.

```
penguins |>
  mutate(body_mass_kg = body_mass_g / 1000)
```

```
# A tibble: 344 x 9
  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 3 more variables: sex <fct>, year <int>, body_mass_kg <dbl>
```

Add a new column to `penguins` that contains a new column `bill_ratio`, which is the ratio of bill length to bill depth.

```
penguins |>
  mutate(bill_ratio = bill_length_mm / bill_depth_mm)
```

```
# A tibble: 344 x 9
  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 3 more variables: sex <fct>, year <int>, bill_ratio <dbl>
```

Add a new column called `id` to `penguins` with a sequence of values from 1 to the length of the data frame. Use `relocate()` to move the column to the first position in the data frame.

```
penguins |>
  mutate(id = 1:n()) |>
  relocate(id)
```

```
# A tibble: 344 x 9
   id species island bill_length_mm bill_depth_mm flipper_length_mm
  <int> <fct>   <fct>         <dbl>         <dbl>         <int>
1     1 Adelie Torgersen      39.1          18.7          181
2     2 Adelie Torgersen      39.5          17.4          186
3     3 Adelie Torgersen      40.3           18          195
4     4 Adelie Torgersen      NA           NA           NA
5     5 Adelie Torgersen      36.7          19.3          193
6     6 Adelie Torgersen      39.3          20.6          190
7     7 Adelie Torgersen      38.9          17.8          181
8     8 Adelie Torgersen      39.2          19.6          195
9     9 Adelie Torgersen      34.1          18.1          193
10    10 Adelie Torgersen      42           20.2          190
# i 334 more rows
# i 3 more variables: body_mass_g <int>, sex <fct>, year <int>
```



### Task 3: Summarize data with `group_by()` and `summarize()` & `count()`

Starting with `penguins`, group the data by species, then create a summary table containing the maximum and minimum length of flippers (call the columns `flip_max` and `flip_min`). How will you handle NA values?

```
penguins |>
  group_by(species) |>
  summarize(flip_max = max(flipper_length_mm, na.rm = TRUE),
            flip_min = min(flipper_length_mm, na.rm = TRUE))
```

```
# A tibble: 3 x 3
  species flip_max flip_min
  <fct>    <int>    <int>
1 Adelie     210      172
2 Chinstrap  212      178
3 Gentoo    231      203
```

Starting with `penguins`, group the data by species and year, then create a summary table containing the mean bill depth (call this `bill_depth_mean`), the mean bill length (call this `bill_length_mean`), and the count for each group. How will you handle NA values?

```
penguins |>
  group_by(species, year) |>
  summarize(bill_depth_mean = mean(bill_depth_mm, na.rm = TRUE),
            bill_length_mean = mean(bill_length_mm, na.rm = TRUE))
```

```
# A tibble: 9 x 4
# Groups:   species [3]
  species year bill_depth_mean bill_length_mean
  <fct>   <int>         <dbl>         <dbl>
1 Adelie  2007          18.8          38.8
2 Adelie  2008          18.2          38.6
3 Adelie  2009          18.1          39.0
4 Chinstrap 2007          18.5          48.7
5 Chinstrap 2008          18.4          48.7
6 Chinstrap 2009          18.3          49.1
7 Gentoo   2007          14.7          47.0
8 Gentoo   2008          14.9          46.9
9 Gentoo   2009          15.3          48.5
```

Use the `count()` function to count the number of observations for each species in `penguins`.

```
penguins |>
  filter(!is.na(bill_depth_mm), !is.na(bill_length_mm)) |>
  group_by(species, year) |>
  summarize(bill_depth_mean = mean(bill_depth_mm),
            bill_length_mean = mean(bill_length_mm),
            count = n())
```

```
# A tibble: 9 x 5
# Groups:   species [3]
  species    year bill_depth_mean bill_length_mean count
  <fct>    <int>         <dbl>         <dbl> <int>
1 Adelie   2007           18.8           38.8    49
2 Adelie   2008           18.2           38.6    50
3 Adelie   2009           18.1           39.0    52
4 Chinstrap 2007           18.5           48.7    26
5 Chinstrap 2008           18.4           48.7    18
6 Chinstrap 2009           18.3           49.1    24
7 Gentoo   2007           14.7           47.0    34
8 Gentoo   2008           14.9           46.9    46
9 Gentoo   2009           15.3           48.5    43
```

Use the `count()` function to count the number of observations for each species and island in `penguins`.

```
penguins |>
  count(species, island)
```

```
# A tibble: 5 x 3
  species island      n
  <fct>    <fct>    <int>
1 Adelie  Biscoe      44
2 Adelie  Dream       56
3 Adelie  Torgersen   52
4 Chinstrap Dream      68
5 Gentoo  Biscoe     124
```

Use `filter()` to create a subset from `penguins` that contains observations for female penguins recorded at Torgersen and Biscoe Islands. Then use add the pipe `|>` and `count()` to verify that you written the correct code.

```
penguins |>
  filter(sex == "female", island %in% c("Torgersen","Biscoe")) |>
  count(sex,island)
```

```
# A tibble: 2 x 3
  sex    island      n
  <fct> <fct>    <int>
1 female Biscoe      80
2 female Torgersen   24
```

## Task 7: Data communication

In the YAML header (between the three dashes at the top of the document)

1. Add your name as the author of this document
2. Render the document and fix any errors

## Task 8: Stage, Commit & Push to GitHub

1. Open the Git pane in RStudio. It's in the top right corner in a separate tab.
2. **Stage** your changes by checking appropriate box next to all files (if you select one file with your mouse, you can then highlight them all with Ctrl + A on your keyboard and check all boxes).
3. Write a meaningful commit message (e.g. "Completed part a of homework assignment 03.") in the **Commit message** box.
4. Click **Commit**. Note that every commit needs to have a commit message associated with it.