dplyr - Part II

Author

Date

Outline

- Continue working with `dplyr`
- `filter()`: subset rows using column values
- `group_by()`: group by one or more variables
- `summarise()`: summarise each group to fewer rows
- Exercises

Filter(): Subset rows using column values

- The following expressions are mostly used:
  - `==`, `>`, etc.
  - `&` *(and)*
  - `|` *(or)*
  - `!` *(is not)*
  - `xor()` *(elementwise exclusive OR)*
  - `is.na()`
  - `between()` *(shortcut for `x >= left` & `x <= right`)*

Quantitative variables

```
penguins %>%
  filter(bill_length_mm > 40) %>%
  head(n = 2) %>%
  kable() %>%  kable_styling(font_size = 14)
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	42.0	20.2	190	4250	NA	2007

Qualitative variables

```
penguins %>%
  filter(species == "Adelie") %>%
  head(n = 3) %>%
  kable() %>%  kable_styling(font_size = 14)
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007

Boolean operators

```
penguins %>%
  filter(species == "Adelie" & island == "Dream") %>%
  head(n = 1) %>%
  kable() %>%  kable_styling(font_size = 14)
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Dream	39.5	16.7	178	3250	female	2007

```
penguins %>%
  filter(species == "Adelie" & year != 2008) %>%
  head(n = 2) %>%
  kable() %>%  kable_styling(font_size = 14)
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007

- `xor()`

```
penguins %>%
  filter(xor(species == "Adelie", flipper_length_mm > 200)) %>%
  head(n = 2) %>%
  kable() %>%  kable_styling(font_size = 14)
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007

- `is.na()`

```
penguins %>%
  filter(is.na(body_mass_g)) %>%
  head(n = 2) %>%
  kable() %>% kable_styling(font_size = 14)
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Gentoo	Biscoe	NA	NA	NA	NA	NA	2009

- `between()`

```
penguins %>%
  filter(between(body_mass_g, 5500, 6000)) %>%
  head(n = 2) %>%
  kable() %>%  kable_styling(font_size = 14)
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Gentoo	Biscoe	50	16.3	230	5700	male	2007
Gentoo	Biscoe	50	15.2	218	5700	male	2007

Quick recap

- So far, we know how to...
- subset columns with `select()`
- modify existing and create new columns with `mutate()`
- rename columns with `rename()`
- change the order of columns with `relocate()`
- subset rows with `filter()`
- **Questions?**

How do we work with different groups in our datasets?

Group_by: Group by one or more variables

- Examples of groups in the `Palmerpenguins` dataset:

```
distinct(penguins, species)
```

```
## # A tibble: 3 x 1
##   species
##   <fct>
## 1 Adelie
## 2 Gentoo
## 3 Chinstrap
```

```
distinct(penguins, sex)
```

```
## # A tibble: 3 x 1
##   sex
##   <fct>
## 1 male
## 2 female
## 3 <NA>
```

- Group_by is a powerful function allowing operations per group (which is usually more interesting)
- What is the average weigth of the penguins per year?

```
penguins_weigth <- penguins %>%
  group_by(year) %>%
  mutate(weigth_avg = mean(body_mass_g, na.rm = TRUE))

distinct(penguins_weigth, weigth_avg) %>%
  kable()
```

year	weigth_avg
2007	4124.541
2008	4266.667
2009	4210.294

- What is the maximum bill length of the three different penguin species?

```
penguins_bill <- penguins %>%
  group_by(species) %>%
  mutate(bill_length = max(bill_length_mm, na.rm = TRUE))

distinct(penguins_bill, bill_length) %>%
  kable()
```

species	bill_length
Adelie	46.0
Gentoo	59.6
Chinstrap	58.0

- You can also group by several groups
- What is the maximum bill length of the three different penguin species on the three different islands?

```
penguins_bill_island <- penguins %>%
  group_by(species, island) %>%
  mutate(bill_length_max = max(bill_length_mm, na.rm = TRUE))

distinct(penguins_bill_island, bill_length_max) %>%
  kable()
```

species	island	bill_length_max
Adelie	Torgersen	46.0
Adelie	Biscoe	45.6
Adelie	Dream	44.1
Gentoo	Biscoe	59.6
Chinstrap	Dream	58.0

- There's a tidier way to obtain the same result

Summarise(): Summarise each group to fewer rows

- Let's look at the same examples again
- What is the average weigth of the penguins per year?

```
penguins %>%
  group_by(year) %>%
  summarise(weigth_avg = mean(body_mass_g, na.rm = TRUE)) %>%
  kable()
```

year	weigth_avg
2007	4124.541
2008	4266.667
2009	4210.294

- What is the maximum bill length of the three different penguin species on the three different islands?

```
penguins %>%
  group_by(species, island) %>%
  summarise(bill_length_max = max(bill_length_mm, na.rm = TRUE)) %>%
  kable()
```

species	island	bill_length_max
Adelie	Biscoe	45.6
Adelie	Dream	44.1
Adelie	Torgersen	46.0
Chinstrap	Dream	58.0
Gentoo	Biscoe	59.6

- Again, combining different operations is possible

```
penguins %>%
  group_by(species) %>%
  summarise(bill_length_max = max(bill_length_mm, na.rm = TRUE),
            bill_length_min = min(bill_length_mm, na.rm = TRUE),
            average_weigth = mean(body_mass_g, na.rm = TRUE)) %>%
  kable()
```

species	bill_length_max	bill_length_min	average_weigth
Adelie	46.0	32.1	3700.662
Chinstrap	58.0	40.9	3733.088
Gentoo	59.6	40.9	5076.016

Questions?

Recap & outlook

- Introduced to `dplyr`
- `select()`: picks variables based on their names.
- `mutate()`: adds new variables that are functions of existing variables
- `filter()`: picks cases based on their values.
- `group_by`: allows operations across groups
- `summarise()`: reduces multiple values down to a single summary
- Next up: Data visualisation with `ggplot2` and applying everything in a final exercise

Time to exercise!