# dplyr - Part I

Author

Date

# Outline

- Introducing `dplyr` and its verbs
- `select()`: extracts columns from your dataset
- `mutate()`: adds new variables that are functions of existing variables
- `rename()` & `relocate()`
- Exercises

# Dplyr - A Grammar of Data Manipulation

- Grammar in this context means a framework which follows a (layered) approach to manipulate and transform your data in a structured manner (adapted from this post)
- In the words of the authors: "Set of verbs that help you solve the most common data manipulation challenges."
- `select()`: extracts columns from your dataset
- `mutate()`: adds new variables that are functions of existing variables
- `filter()`: picks cases based on their values
- ...

# Select(): Subsetting columns

```
penguins %>%
  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 |

```
penguins %>%
  select(species, island, bill_length_mm) %>%
  head(n = 3) %>%
  kable() %>% kable_styling(font_size = 14)
```

| species | island | bill_length_mm |
|---|---|---|
| Adelie | Torgersen | 39.1 |
| Adelie | Torgersen | 39.5 |
| Adelie | Torgersen | 40.3 |

## Subsetting by column names with :

```
penguins %>%
  select(species:bill_length_mm) %>%
  head(n = 3) %>%
  kable() %>% kable_styling(font_size = 10)
```

| species | island | bill_length_mm |
|---|---|---|
| Adelie | Torgersen | 39.1 |
| Adelie | Torgersen | 39.5 |
| Adelie | Torgersen | 40.3 |

## Subsetting by column positions

```
penguins %>%
  select(1:3) %>%
  head(n = 3) %>%
  kable() %>% kable_styling(font_size = 10)
```

| species | island | bill_length_mm |
|---|---|---|
| Adelie | Torgersen | 39.1 |
| Adelie | Torgersen | 39.5 |
| Adelie | Torgersen | 40.3 |

## Subsetting by column positions with `!`

```
penguins %>%
  select(!1:3) %>%
  head(n = 3) %>%
  kable()
```

| bill_depth_mm | flipper_length_mm | body_mass_g | sex | year |
|---|---|---|---|---|
| 18.7 | 181 | 3750 | male | 2007 |
| 17.4 | 186 | 3800 | female | 2007 |
| 18.0 | 195 | 3250 | female | 2007 |

## Dropping variables

```
penguins %>%
  select(-body_mass_g) %>%
  head(n = 1) %>%
  kable() %>% kable_styling(font_size = 10)
```

| species | island | bill_length_mm | bill_depth_mm | flipper_length_mm | sex | year |
|---|---|---|---|---|---|---|
| Adelie | Torgersen | 39.1 | 18.7 | 181 | male | 2007 |

```
penguins %>%
select(-c(year, sex, species)) %>%
  head(n = 1) %>%
  kable() %>% kable_styling(font_size = 10)
```

| island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g |
|---|---|---|---|---|
| Torgersen | 39.1 | 18.7 | 181 | 3750 |

# Questions?

# Useful helper functions

- `starts_with()`

```
penguins %>%
  select(starts_with("bill")) %>%
  head(n = 2) %>%
  kable()
```

| bill_length_mm | bill_depth_mm |
|---|---|
| 39.1 | 18.7 |
| 39.5 | 17.4 |

- ends_with()

```
penguins %>%
  select(ends_with("mm")) %>%
  head(n = 2) %>%
  kable() %>% kable_styling(font_size = 12)
```

| bill_length_mm | bill_depth_mm | flipper_length_mm |
|---|---|---|
| 39.1 | 18.7 | 181 |
| 39.5 | 17.4 | 186 |

- contains()

```
penguins %>%
  select(contains("length")) %>%
  head(n = 2) %>%
  kable() %>% kable_styling(font_size = 12)
```

| bill_length_mm | flipper_length_mm |
|---|---|
| 39.1 | 181 |
| 39.5 | 186 |

- `where()`: Selects the variables for which a function returns `TRUE`

```
penguins %>%
  select(where(is.factor)) %>%
  head(n = 2) %>%
  kable() %>% kable_styling(font_size = 12)
```

| species | island | sex |
|---|---|---|
| Adelie | Torgersen | male |
| Adelie | Torgersen | female |

```
penguins %>%
  select(where(is.numeric)) %>%
  head(n = 2) %>%
  kable() %>% kable_styling(font_size = 12)
```

| bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | year |
|---|---|---|---|---|
| 39.1 | 18.7 | 181 | 3750 | 2007 |
| 39.5 | 17.4 | 186 | 3800 | 2007 |

# Combinations are possible

```
penguins %>%
  select(year,
         contains("length"),
         island) %>%
  head(n = 2) %>%
  kable()
```

| year | bill_length_mm | flipper_length_mm | island |
|---|---|---|---|
| 2007 | 39.1 | 181 | Torgersen |
| 2007 | 39.5 | 186 | Torgersen |

# Mutate(): Create and modify columns

```
penguins %>%
  mutate(body_mass_kg = body_mass_g/1000, .keep = "used") %>%
  head(n = 3) %>%
  kable() %>% kable_styling(font_size = 12)
```

| body_mass_g | body_mass_kg |
|---|---|
| 3750 | 3.75 |
| 3800 | 3.80 |
| 3250 | 3.25 |

```
penguins %>%
  mutate(bill_ratio = bill_length_mm/bill_depth_mm, .keep = "used") %>%
  head(n = 1) %>%
  kable() %>% kable_styling(font_size = 12)
```

| bill_length_mm | bill_depth_mm | bill_ratio |
|---|---|---|
| 39.1 | 18.7 | 2.090909 |

```r
penguins %>%
  mutate(body_mass_kg = body_mass_g/1000,
         bill_square = bill_length_mm*bill_length_mm, .keep = "used") %>%
  head(n = 3) %>%
  kable()
```

| bill_length_mm | body_mass_g | body_mass_kg | bill_square |
|---|---|---|---|
| 39.1 | 3750 | 3.75 | 1528.81 |
| 39.5 | 3800 | 3.80 | 1560.25 |
| 40.3 | 3250 | 3.25 | 1624.09 |

# Rename(): Change variable names

```
penguins %>%
  rename(weight = body_mass_g,
         island_name = island) %>%
  head(n = 3) %>%
  kable() %>% kable_styling(font_size = 12)
```

| species | island_name | bill_length_mm | bill_depth_mm | flipper_length_mm | weight | 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 |

# Pro tip

- `select()` can also rename variables!

```
penguins %>%
  select(island,
         weight = body_mass_g) %>%
  head(n = 3) %>%
  kable()
```

| island | weight |
|---|---|
| Torgersen | 3750 |
| Torgersen | 3800 |
| Torgersen | 3250 |

# Relocate(): Change column order

- `year` as first column

```
penguins %>%
  relocate(year) %>%
  head(n = 1) %>%
  kable() %>% kable_styling(font_size = 12)
```

| year | species | island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | sex |
|---|---|---|---|---|---|---|---|
| 2007 | Adelie | Torgersen | 39.1 | 18.7 | 181 | 3750 | male |

- `sex` after `species`

```
penguins %>%
  relocate(sex, .after = species) %>%
  head(n = 1) %>%
  kable() %>% kable_styling(font_size = 12)
```

| species | sex | island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | year |
|---|---|---|---|---|---|---|---|
| Adelie | male | Torgersen | 39.1 | 18.7 | 181 | 3750 | 2007 |

- `island` before `species`

```
penguins %>%
  relocate(island, .before = species) %>%
  head(n = 1) %>%
  kable() %>% kable_styling(font_size = 12)
```

| island | species | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | sex | year |
|---|---|---|---|---|---|---|---|
| Torgersen | Adelie | 39.1 | 18.7 | 181 | 3750 | male | 2007 |

# Questions?

# Recap & outlook

- Introduced to `dplyr`
- `select()`: picks variables based on their names
- `mutate()`: adds new variables that are functions of existing variables
- `rename()` & `relocate()`
- Next up: `filter()`, `group_by()` and `summarise()`

# Time to exercise!