Ladies

Amsterdam

Introduction to (Data Wrangling with) Tidyverse

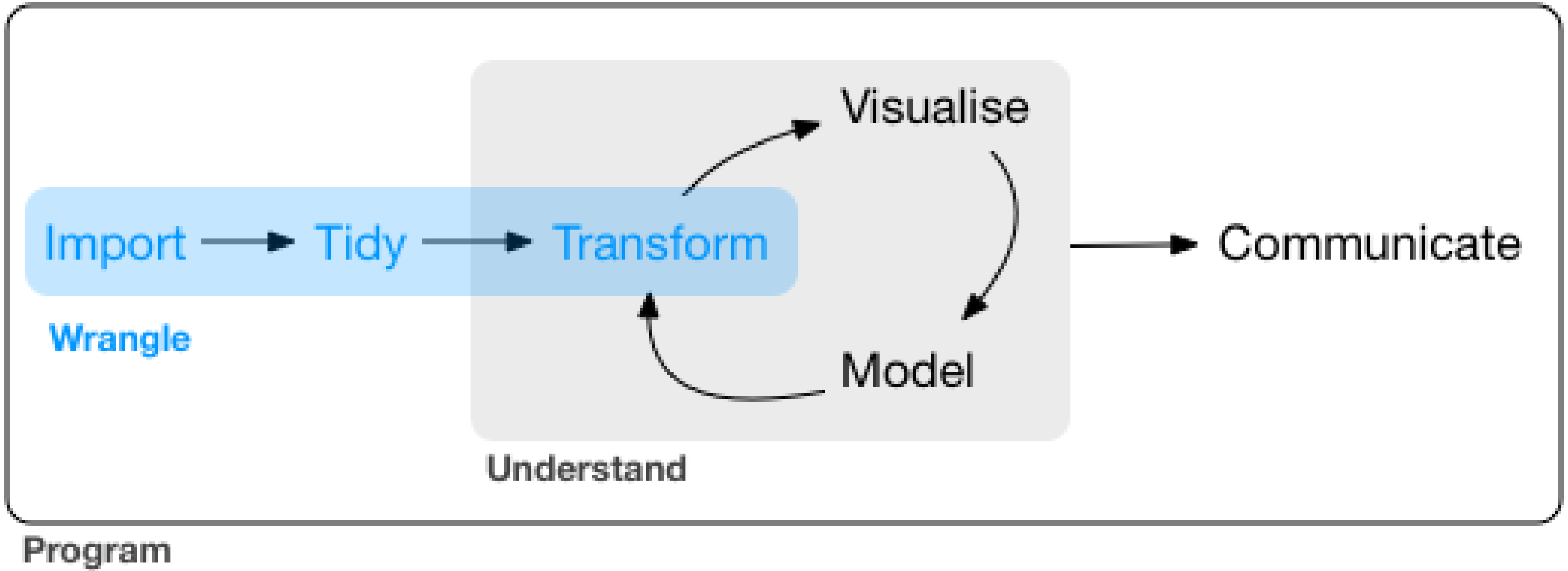
By Nutsa Nanuashvili

24-03-2021

Tidyverse

tidyverse
www.rstudio.com
readr
www.rstudio.com
tidyr
tidyr.tidyverse.org
dplyr
dplyr.tidyverse.org
TIBBLE
www.rstudio.com
ggplot2
www.rstudio.com
stringr
www.rstudio.com
purrr
www.rstudio.com
forcats
www.rstudio.com

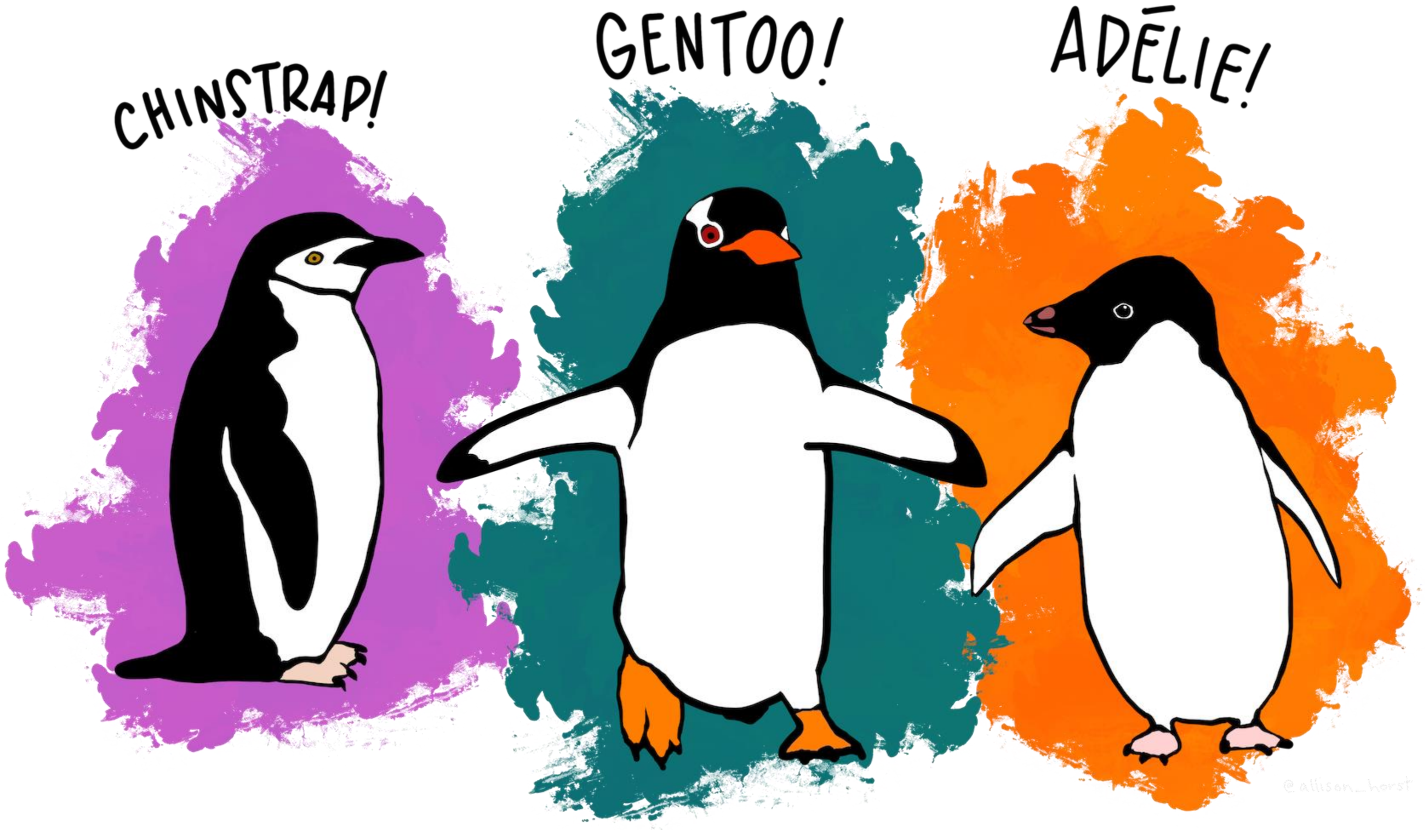
Main Workflow



Packages we’re going to use today

Introducing Data - Palmer Penguins

```r
library(palmerpenguins)
```



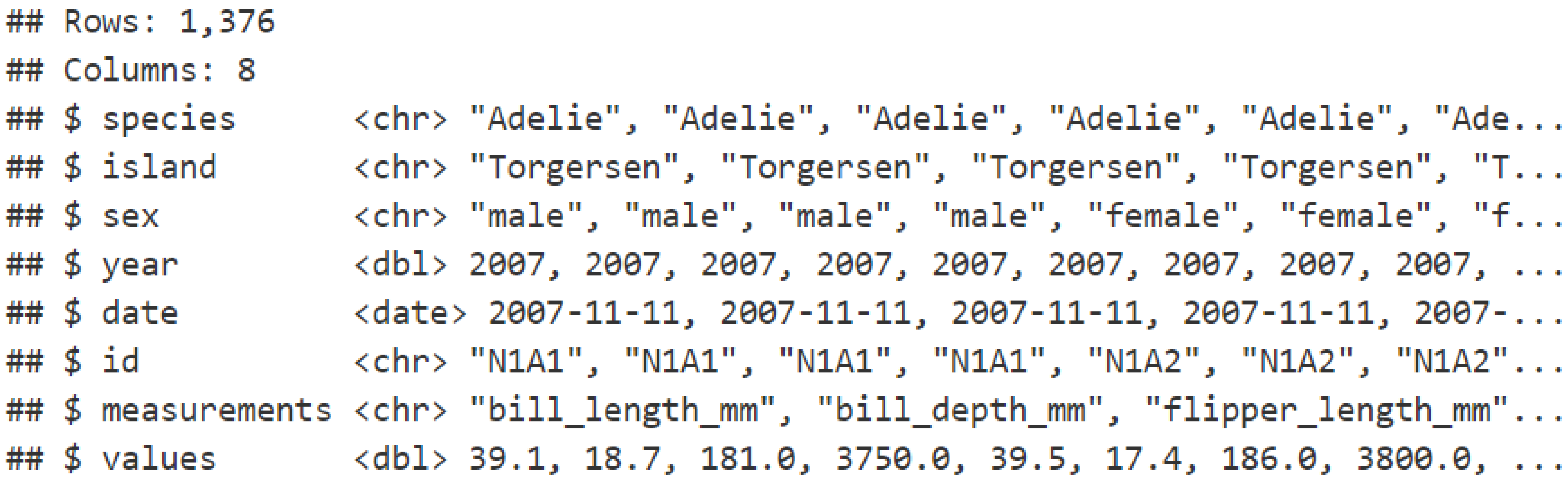
Import Data

```
library(tidyverse) OR library(readr)
read_csv(file, col_names = TRUE, col_types = NULL,
         na = c("", "NA“))


penguin_data = read_csv("dataset/penguins_data.csv")
```

First Overview of the Data

```r
glimpse(penguin_data)
```

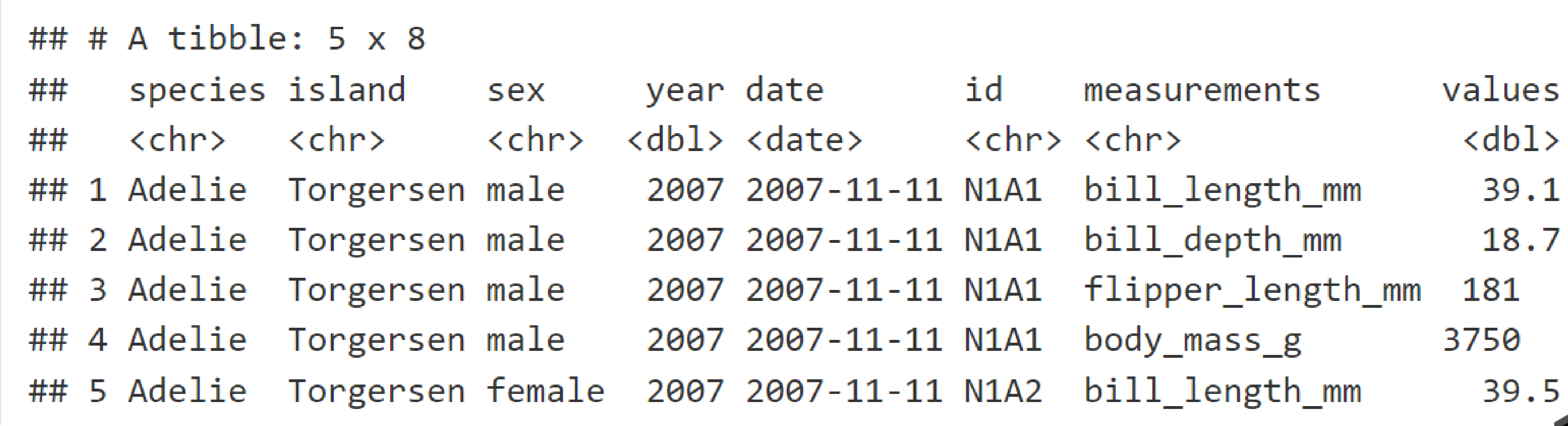


```
## Rows: 1,376
## Columns: 8
## $ species      <chr> "Adelie", "Adelie", "Adelie", "Adelie", "Adelie", "Ade...
## $ island       <chr> "Torgersen", "Torgersen", "Torgersen", "Torgersen", "T...
## $ sex          <chr> "male", "male", "male", "male", "female", "female", "f...
## $ year         <dbl> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, ...
## $ date         <date> 2007-11-11, 2007-11-11, 2007-11-11, 2007-11-11, 2007-...
## $ id           <chr> "N1A1", "N1A1", "N1A1", "N1A1", "N1A2", "N1A2", "N1A2"...
## $ measurements <chr> "bill_length_mm", "bill_depth_mm", "flipper_length_mm"...
## $ values       <dbl> 39.1, 18.7, 181.0, 3750.0, 39.5, 17.4, 186.0, 3800.0, ...
```

@RLadiesAMS
#RLadies #rstats

First Overview of the Data

```
slice_head (penguin_data, n = 5)
```



```
## # A tibble: 5 x 8
##   species island    sex     year date       id    measurements      values
##   <chr>   <chr>     <chr>  <dbl> <date>     <chr> <chr>              <dbl>
## 1 Adelie  Torgersen male    2007 2007-11-11 N1A1  bill_length_mm      39.1
## 2 Adelie  Torgersen male    2007 2007-11-11 N1A1  bill_depth_mm       18.7
## 3 Adelie  Torgersen male    2007 2007-11-11 N1A1  flipper_length_mm  181
## 4 Adelie  Torgersen male    2007 2007-11-11 N1A1  body_mass_g       3750
## 5 Adelie  Torgersen female  2007 2007-11-11 N1A2  bill_length_mm      39.5
```

slice_tail and slice_sample

Exercises Part 1

Converting Variable Types

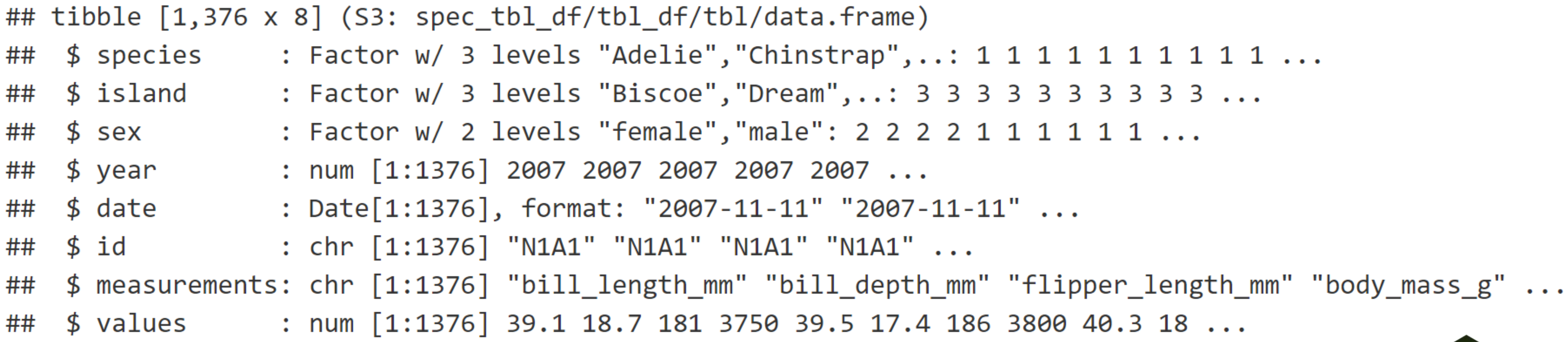
`map()` from `purrr`

```
penguin_data[, c( "species", "island", "sex“ )] =
purrr::map( penguin_data[, c( "species", "island",
"sex“)], factor)


# display data structure
str(penguin_data, give.attr = F)
```

Converting Variable Types

`map()` from `purrr`



```
## tibble [1,376 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ species     : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ island      : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ sex         : Factor w/ 2 levels "female","male": 2 2 2 2 1 1 1 1 1 1 ...
##  $ year        : num [1:1376] 2007 2007 2007 2007 2007 ...
##  $ date        : Date[1:1376], format: "2007-11-11" "2007-11-11" ...
##  $ id          : chr [1:1376] "N1A1" "N1A1" "N1A1" "N1A1" ...
##  $ measurements: chr [1:1376] "bill_length_mm" "bill_depth_mm" "flipper_length_mm" "body_mass_g" ...
##  $ values      : num [1:1376] 39.1 18.7 181 3750 39.5 17.4 186 3800 40.3 18 ...
```

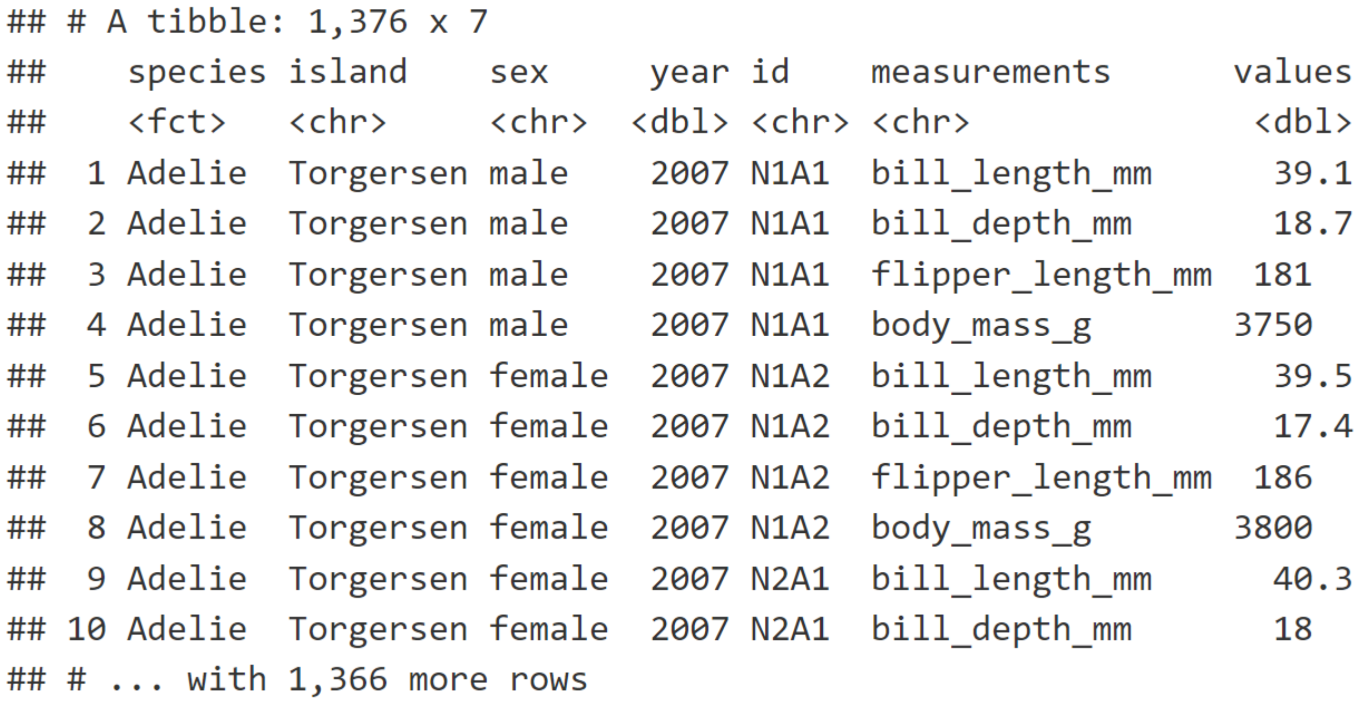
Converting Variable Types

col_types

```
penguin_data02 = read_csv("dataset/penguins_data.csv",
             col_types = cols(species = col_factor( c ( "Adelie",
"Gentoo",  "Chinstrap‶ )),
                    #skip the date column while reading the file
                    date =  col_skip()))
```

Converting Variable Types

col_types



```
## # A tibble: 1,376 x 7
##    species island    sex     year id    measurements      values
##    <fct>   <chr>     <chr>  <dbl> <chr> <chr>              <dbl>
##  1 Adelie  Torgersen male    2007 N1A1  bill_length_mm      39.1
##  2 Adelie  Torgersen male    2007 N1A1  bill_depth_mm       18.7
##  3 Adelie  Torgersen male    2007 N1A1  flipper_length_mm  181
##  4 Adelie  Torgersen male    2007 N1A1  body_mass_g       3750
##  5 Adelie  Torgersen female  2007 N1A2  bill_length_mm      39.5
##  6 Adelie  Torgersen female  2007 N1A2  bill_depth_mm       17.4
##  7 Adelie  Torgersen female  2007 N1A2  flipper_length_mm  186
##  8 Adelie  Torgersen female  2007 N1A2  body_mass_g       3800
##  9 Adelie  Torgersen female  2007 N2A1  bill_length_mm      40.3
## 10 Adelie  Torgersen female  2007 N2A1  bill_depth_mm       18
## # ... with 1,366 more rows
```

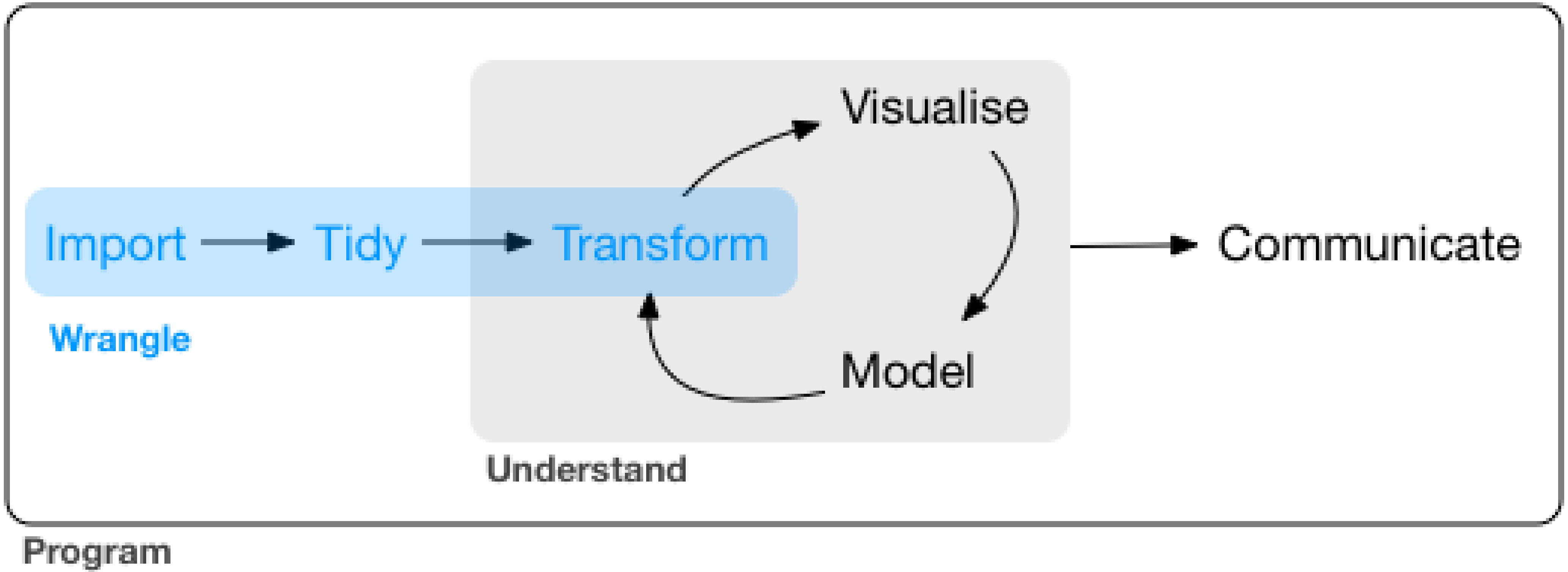
The functions for converting variables

```
col_double()
col_character()
col_date (format = "")
```

```
col_factor()
col_logical()
col_numeric()
```

Exercises Part 2

Main Workflow



Pipe Operator

```
function(data, arguments)
```

Same as

```
data %>% function(arguments)
```

And

```
function_2( function_1 (A) )
```

Is equivalent of

```
A %>%
 function_1() %>%
 function_2()
```

Pipe Operator

Can be read as ***“then”***

```
pasta %>%
 boil_water() %>%
 put_pasta(type = “penne”)
 add_souce (type = “marinara”)
```

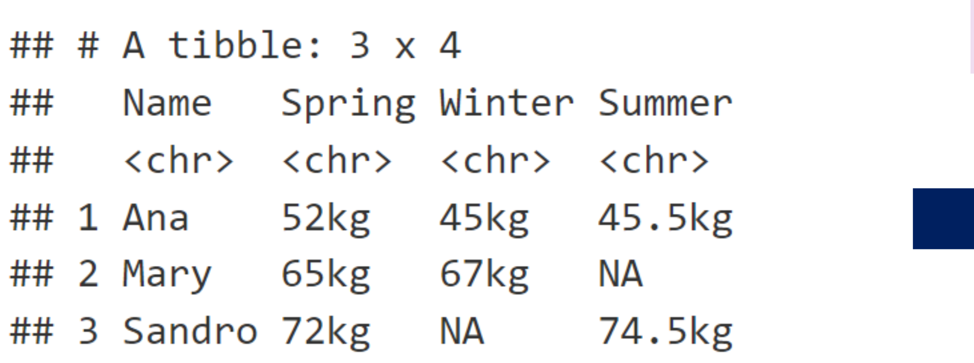
Tidy Data

Name	Spring	Winter	Summer
Ana	52kg	45kg	45.5kg
Mary	65kg	67kg	NA
Sandro	72kg	NA	74.5kg

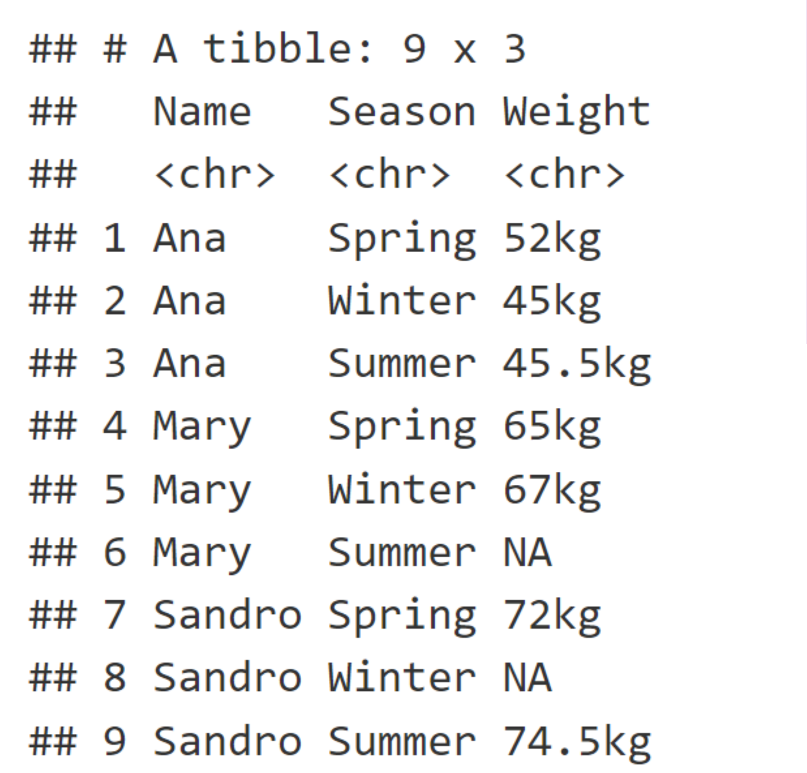
Name	Season	Weight
Ana	Spring	52kg
Ana	Winter	45kg
Ana	Summer	45.5kg
Mary	Spring	65kg
Mary	Winter	67kg
Mary	Summer	NA
Sandro	Spring	72kg
Sandro	Winter	NA
Sandro	Summer	74.5kg

Tidy Data

```
#reshape into longer format
weight_df_long = weight_df %>% pivot_longer(cols = c( "Spring“ ,
"Winter", "Summer“ ),
             names_to = "Season“,
             values_to = "Weight“,
             values_drop_na = FALSE )
```



```
## # A tibble: 3 x 4
##   Name   Spring Winter Summer
##   <chr>  <chr>  <chr>  <chr>
## 1 Ana    52kg   45kg   45.5kg
## 2 Mary   65kg   67kg   NA
## 3 Sandro 72kg   NA     74.5kg
```



```
## # A tibble: 9 x 3
##   Name   Season Weight
##   <chr>  <chr>  <chr>
## 1 Ana    Spring 52kg
## 2 Ana    Winter 45kg
## 3 Ana    Summer 45.5kg
## 4 Mary   Spring 65kg
## 5 Mary   Winter 67kg
## 6 Mary   Summer NA
## 7 Sandro Spring 72kg
## 8 Sandro Winter NA
## 9 Sandro Summer 74.5kg
```

Tidy Data

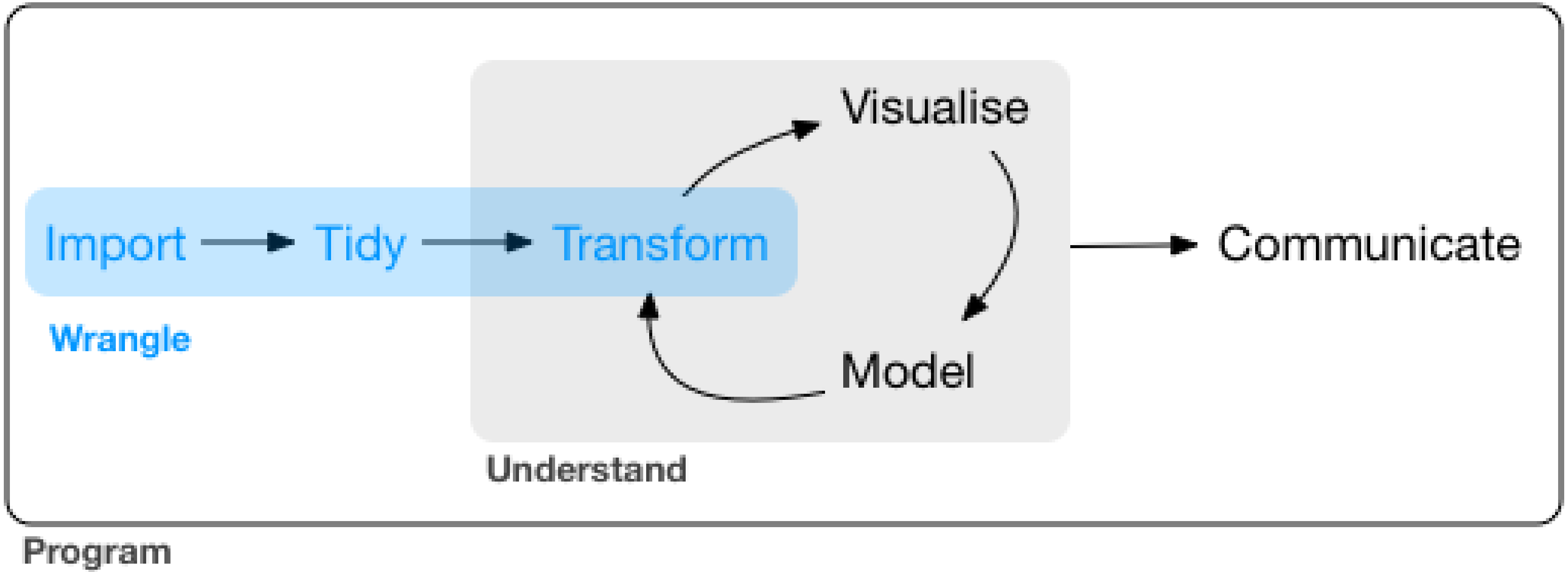
```
pivot_wider (weight_df_long,
                    names_from = Season,
                    values_from = Weight )
```

```
## # A tibble: 3 x 4
##   Name   Spring Winter Summer
##   <chr>  <chr>  <chr>  <chr>
## 1 Ana    52kg   45kg   45.5kg
## 2 Mary   65kg   67kg   NA
## 3 Sandro 72kg   NA     74.5kg
```

```
## # A tibble: 9 x 3
##   Name   Season Weight
##   <chr>  <chr>  <chr>
## 1 Ana    Spring 52kg
## 2 Ana    Winter 45kg
## 3 Ana    Summer 45.5kg
## 4 Mary   Spring 65kg
## 5 Mary   Winter 67kg
## 6 Mary   Summer NA
## 7 Sandro Spring 72kg
## 8 Sandro Winter NA
## 9 Sandro Summer 74.5kg
```

Exercises Part 3

Main Workflow

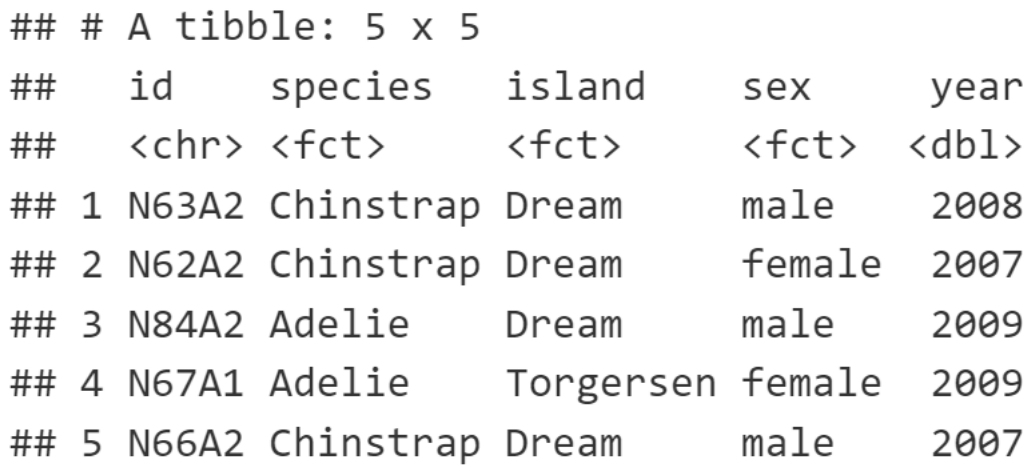


Data Transformation

Select ()

Selects columns by their name and returns a tibble

```
penguin_df_wide %>%
  select(id,
         species:year) %>%
  slice_sample(n = 5)
```



```
## # A tibble: 5 x 5
##   id    species   island    sex     year
##   <chr> <fct>     <fct>     <fct>  <dbl>
## 1 N63A2 Chinstrap Dream     male    2008
## 2 N62A2 Chinstrap Dream     female  2007
## 3 N84A2 Adelie    Dream     male    2009
## 4 N67A1 Adelie    Torgersen female  2009
## 5 N66A2 Chinstrap Dream     male    2007
```

@RLadiesAMS
#RLadies #rstats

Deleting columns using `select`

```
penguin_df_wide %>%
      select( -(year:id) ) %>%
      slice_sample(n = 5)
```

```
## # A tibble: 5 x 7
##   species island sex   bill_length_mm bill_depth_mm flipper_length_~ body_mass_g
##   <fct>   <fct>  <fct>          <dbl>         <dbl>            <dbl>       <dbl>
## 1 Chinst~ Dream  fema~           40.9          16.6              187        3200
## 2 Adelie  Dream  fema~           36.6          18.4              184        3475
## 3 Adelie  Torge~ fema~           40.2          17                176        3450
## 4 Gentoo  Biscoe fema~           46.2          14.5              209        4800
## 5 Chinst~ Dream  male            51.3          19.2              193        3650
```

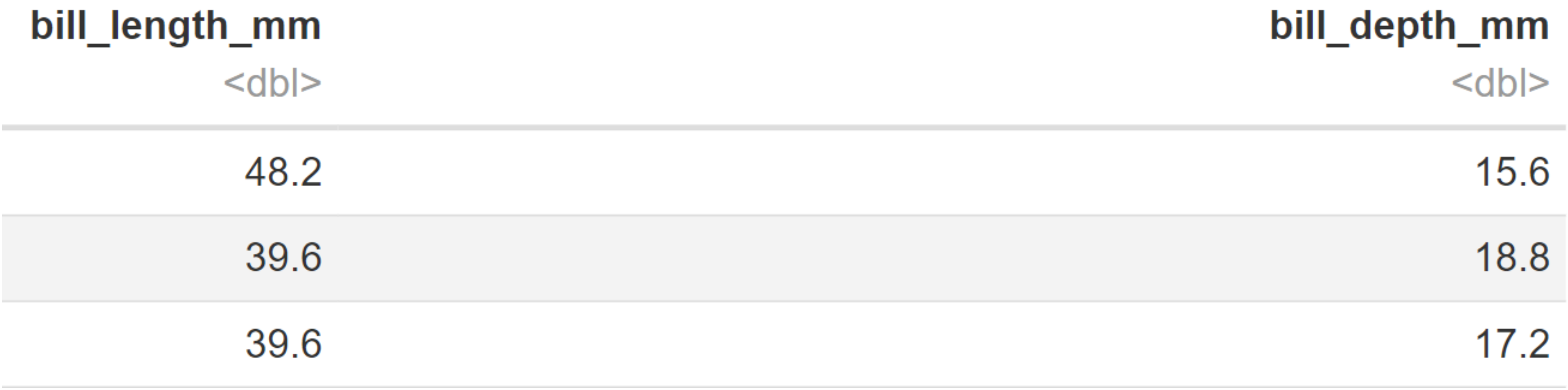
“Helper” verbs for `select`

- `starts_with()`
- `ends_with()`

- `contains()`
- `everything()`
- `where()`

“Helper” verbs

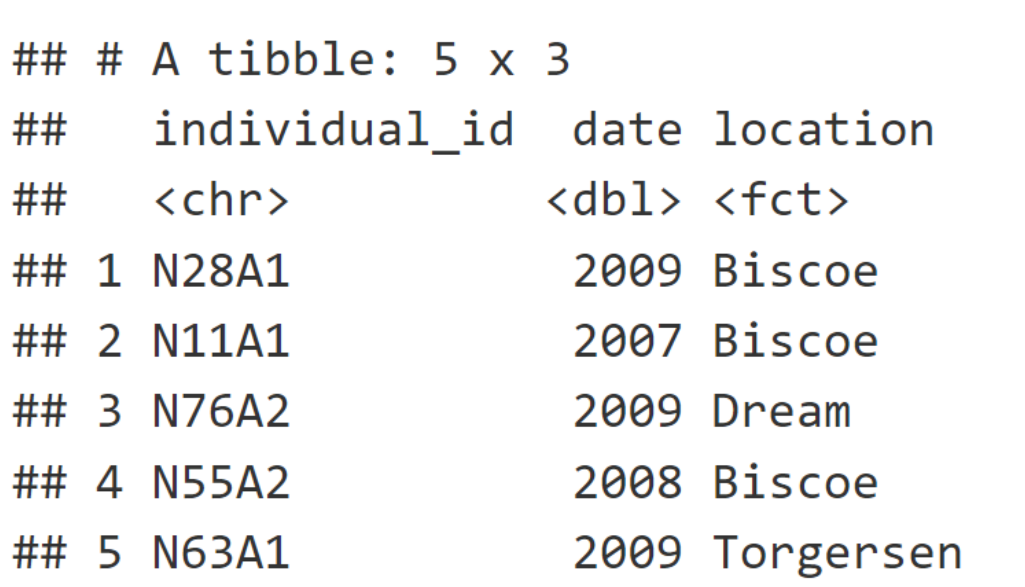
```
penguin_df_wide %>%
  select( starts_with ("bill“) ) %>%
  slice_sample(n = 3)
```



bill_length_mm <dbl>	bill_depth_mm <dbl>
48.2	15.6
39.6	18.8
39.6	17.2

Renaming

```
penguin_df_wide %>%
  select(individual_id = id,
         date = year,
         location = island ) %>%
  slice_sample (n = 5)
```

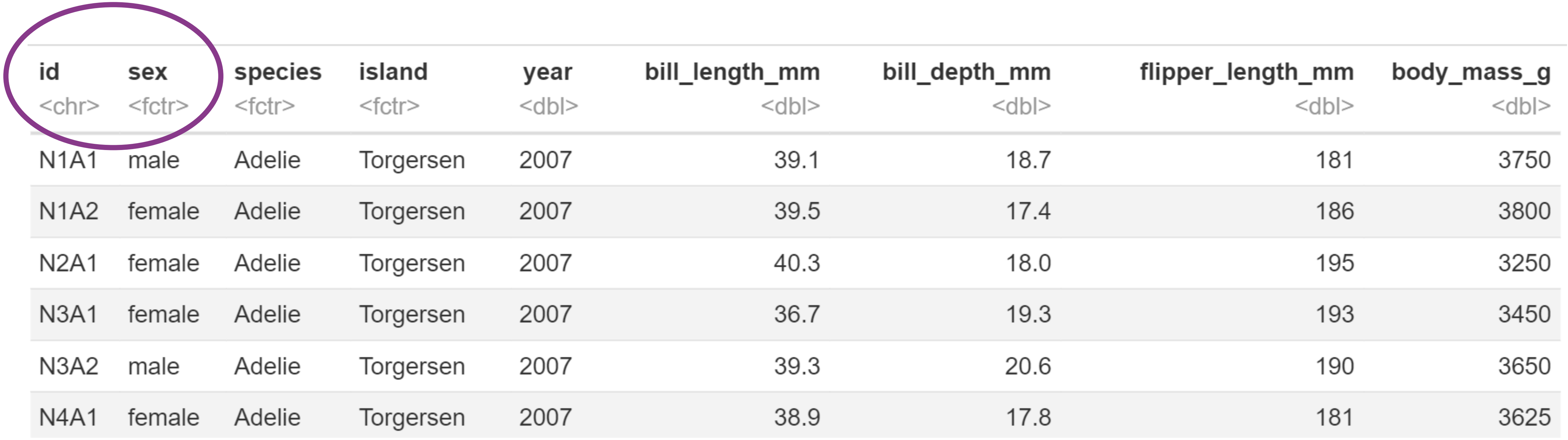


```
## # A tibble: 5 x 3
##   individual_id  date location
##   <chr>         <dbl> <fct>
## 1 N28A1          2009 Biscoe
## 2 N11A1          2007 Biscoe
## 3 N76A2          2009 Dream
## 4 N55A2          2008 Biscoe
## 5 N63A1          2009 Torgersen
```

Rearranging columns

```
penguin_df_wide %>%
   select(id, sex,
         everything()) %>%
# drop the missing values from every row
  drop_na()
```

Rearranging columns



id	sex	species	island	year	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
<chr>	<fctr>	<fctr>	<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
N1A1	male	Adelie	Torgersen	2007	39.1	18.7	181	3750
N1A2	female	Adelie	Torgersen	2007	39.5	17.4	186	3800
N2A1	female	Adelie	Torgersen	2007	40.3	18.0	195	3250
N3A1	female	Adelie	Torgersen	2007	36.7	19.3	193	3450
N3A2	male	Adelie	Torgersen	2007	39.3	20.6	190	3650
N4A1	female	Adelie	Torgersen	2007	38.9	17.8	181	3625

Relocate

```
relocate(.data, ..., .before = NULL, .after = NULL)
```

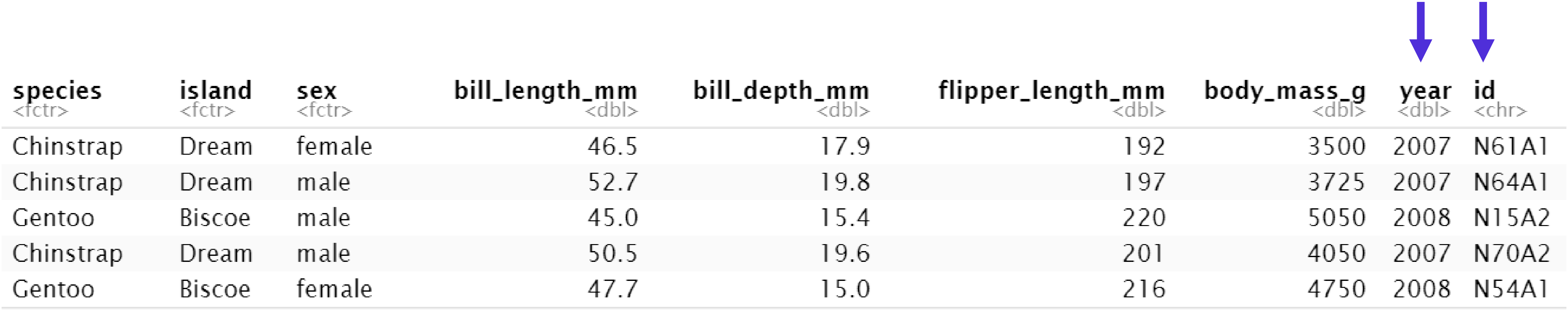
```
penguin_df_wide %>%

  relocate( year:id, .after = last_col()) %>%

  slice_sample(n = 5)
```

Relocate

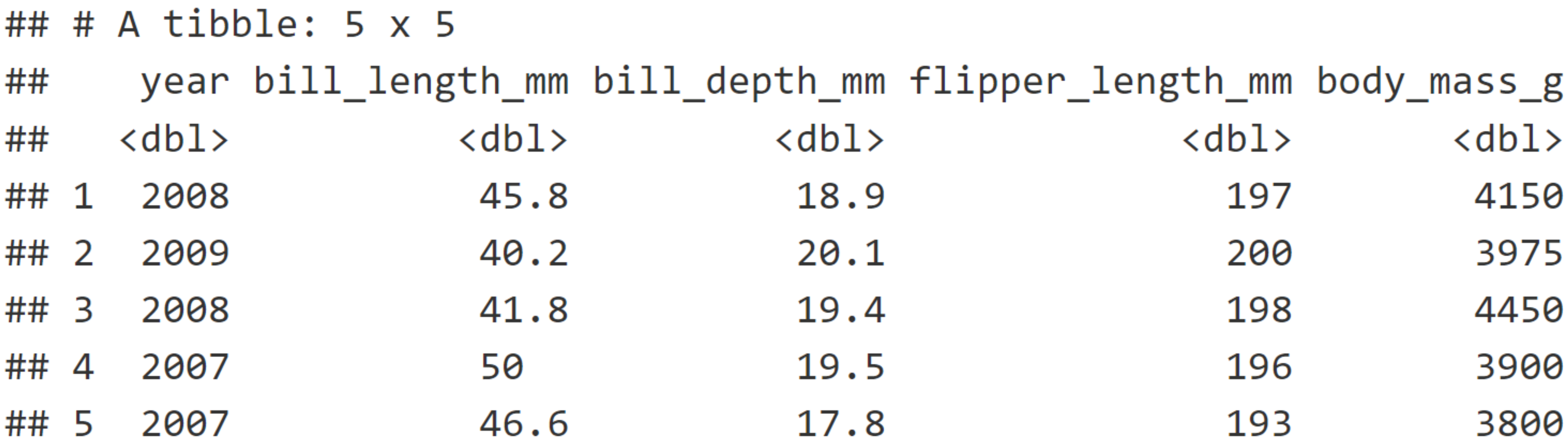
```
relocate(.data, ..., .before = NULL, .after = NULL)
```



species	island	sex	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	year	id
<fctr>	<fctr>	<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
Chinstrap	Dream	female	46.5	17.9	192	3500	2007	N61A1
Chinstrap	Dream	male	52.7	19.8	197	3725	2007	N64A1
Gentoo	Biscoe	male	45.0	15.4	220	5050	2008	N15A2
Chinstrap	Dream	male	50.5	19.6	201	4050	2007	N70A2
Gentoo	Biscoe	female	47.7	15.0	216	4750	2008	N54A1

Select based on a condition

`where()` selects a column *where* the condition is **`TRUE`**



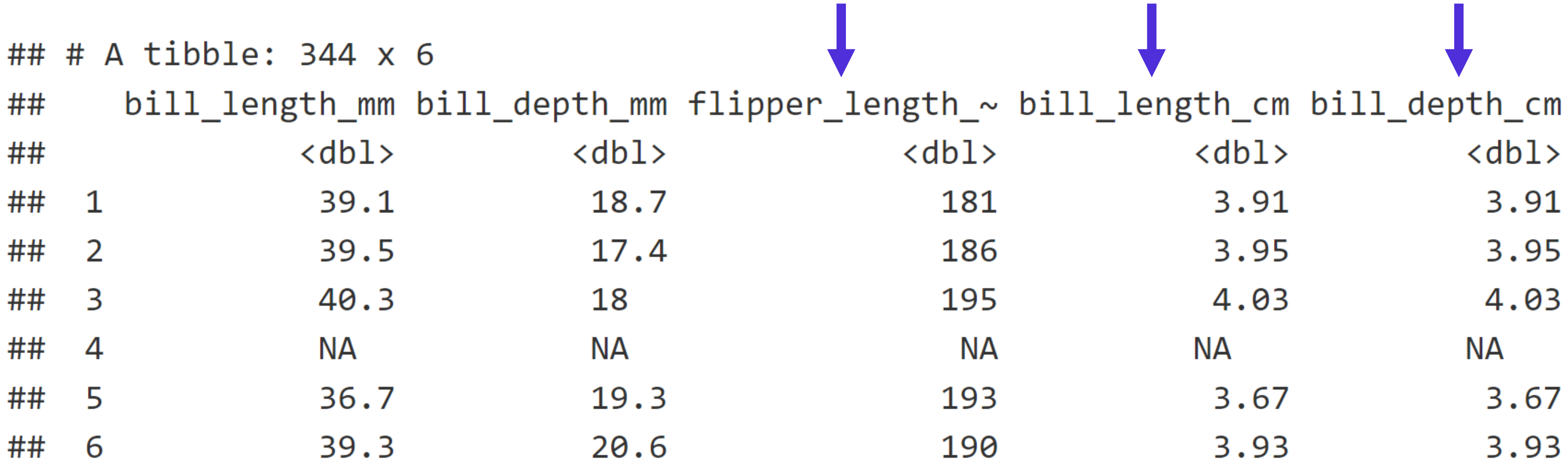
```
## # A tibble: 5 x 5
##    year bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <dbl>          <dbl>         <dbl>             <dbl>       <dbl>
## 1  2008           45.8          18.9               197        4150
## 2  2009           40.2          20.1               200        3975
## 3  2008           41.8          19.4               198        4450
## 4  2007           50            19.5               196        3900
## 5  2007           46.6          17.8               193        3800
```

Exercises Part 4

Forming new columns with `mutate`

```
penguin_df_wide %>%
   select(contains("mm“)) %>%
   mutate(bill_length_cm = bill_length_mm / 10,
         bill_depth_cm =  bill_length_mm / 10,
         flipper_length_cm = flipper_length_mm / 10)
```

Forming new columns with `mutate`



```
## # A tibble: 344 x 6
##    bill_length_mm bill_depth_mm flipper_length_~ bill_length_cm bill_depth_cm
##             <dbl>         <dbl>            <dbl>          <dbl>         <dbl>
##  1           39.1          18.7              181           3.91          3.91
##  2           39.5          17.4              186           3.95          3.95
##  3           40.3          18                195           4.03          4.03
##  4           NA            NA                 NA          NA            NA
##  5           36.7          19.3              193           3.67          3.67
##  6           39.3          20.6              190           3.93          3.93
```

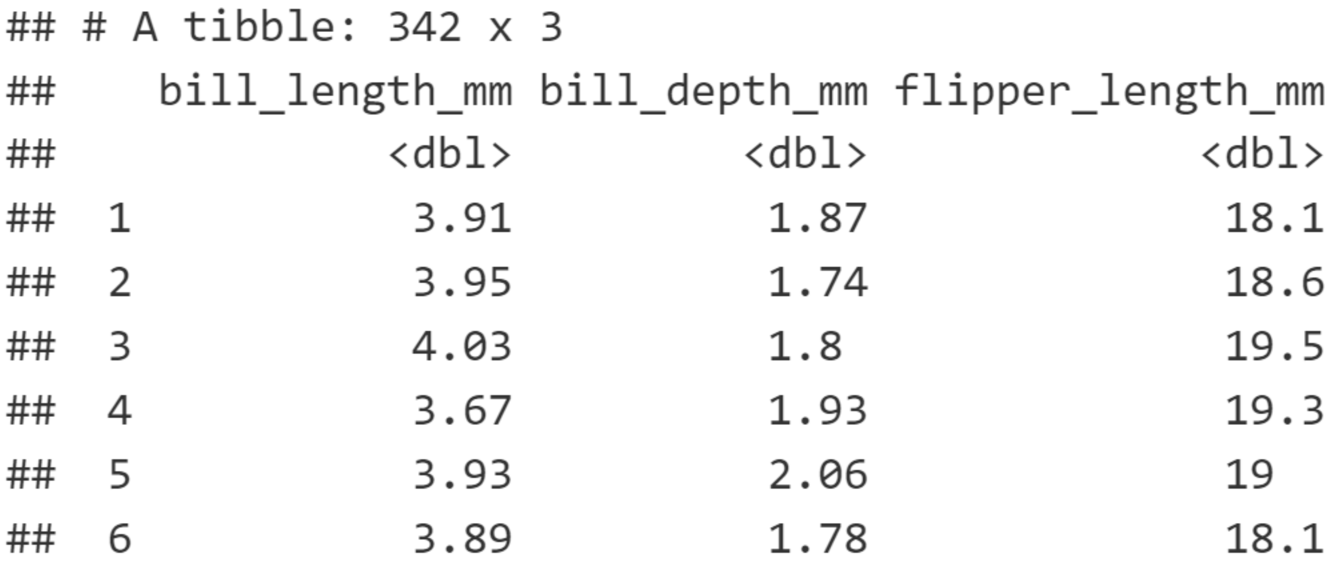
across

Takes 2 arguments - columns to transform & a function to apply

```
penguin_df_wide %>%
  # select every column that contains "mm"in name
  select(contains("mm“)) %>%
  # remove missing values
  drop_na() %>%
  # divide every column by 10
  mutate(across (everything(), ~.x / 10 ) )
```

across

Takes 2 arguments - columns to transform & a function to apply



```
## # A tibble: 342 x 3
##    bill_length_mm bill_depth_mm flipper_length_mm
##             <dbl>         <dbl>             <dbl>
##  1           3.91          1.87              18.1
##  2           3.95          1.74              18.6
##  3           4.03          1.8               19.5
##  4           3.67          1.93              19.3
##  5           3.93          2.06              19  
##  6           3.89          1.78              18.1
```

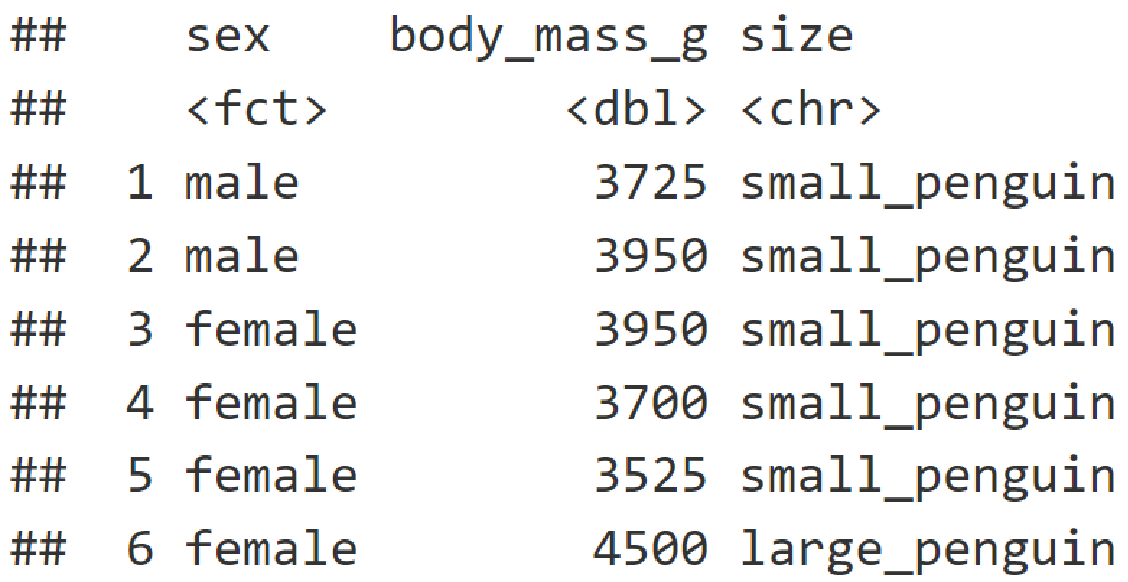
if_else(condition, true, false)

Let’s divide penguins into small and large

```r
# calculate median body mass of all penguins
median_mass = median(penguin_df_wide$body_mass_g, na.rm = T)


penguin_df_wide %>%
  select(sex, body_mass_g) %>%
#create a new column to categorize penguins based on their mass
  mutate(size = if_else(body_mass_g  >= median_mass,
"large_penguin", "small_penguin"))
```

if_else(condition, true, false)



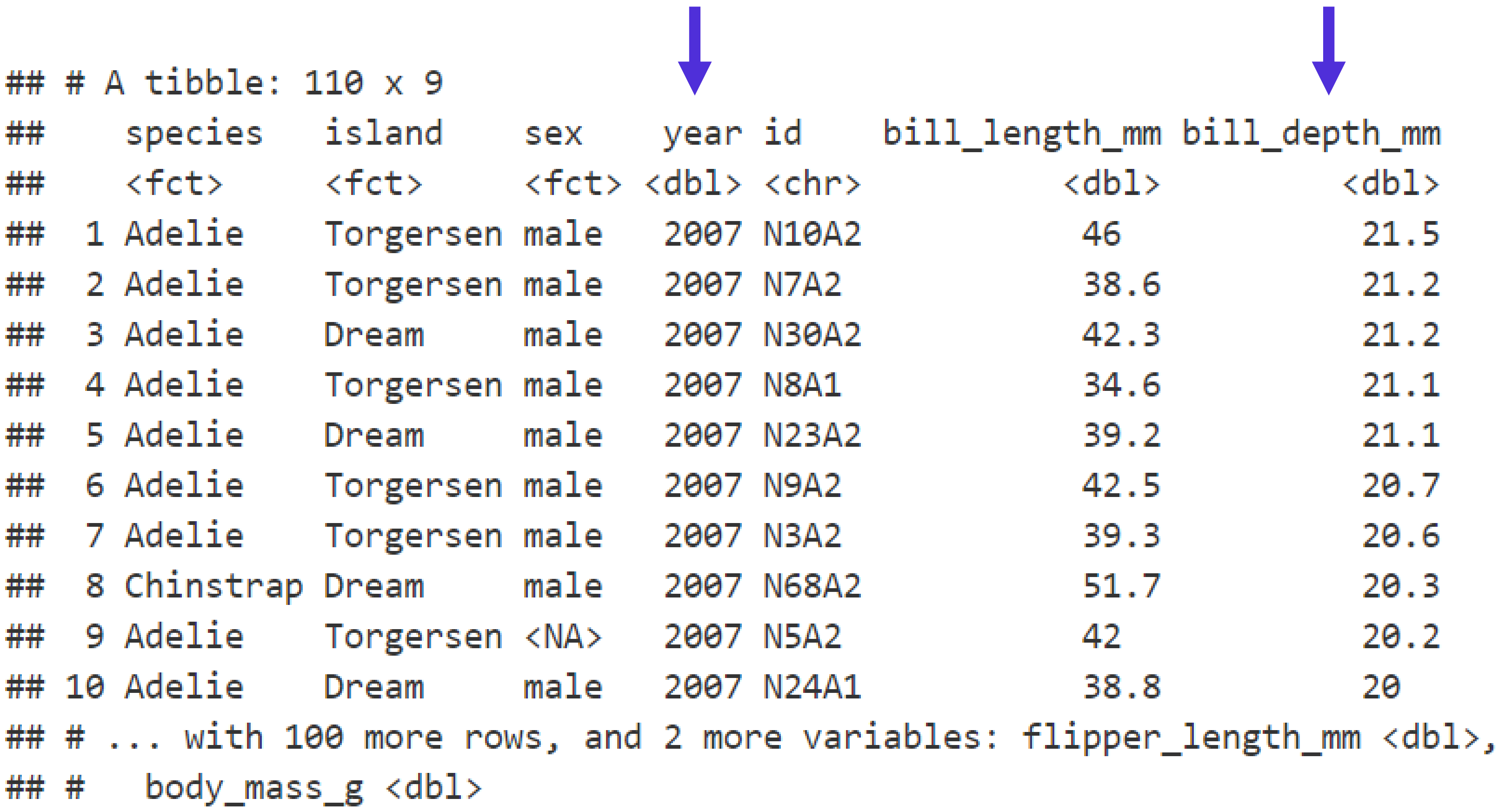
```
##     sex    body_mass_g size
##     <fct>        <dbl> <chr>
##   1 male          3725 small_penguin
##   2 male          3950 small_penguin
##   3 female        3950 small_penguin
##   4 female        3700 small_penguin
##   5 female        3525 small_penguin
##   6 female        4500 large_penguin
```

Exercises Part 5

Filtering and changing the row order

```
# choose only the rows corresponding to year 2007
penguin_df_wide %>%
  filter(year == '2007` )%>%
# sort bill depth in descending order
  arrange (desc(bill_depth_mm))
```

Filtering and changing the row order



```
## # A tibble: 110 x 9
##    species   island    sex    year id     bill_length_mm bill_depth_mm
##    <fct>     <fct>     <fct> <dbl> <chr>           <dbl>         <dbl>
##  1 Adelie    Torgersen male   2007 N10A2            46            21.5
##  2 Adelie    Torgersen male   2007 N7A2             38.6          21.2
##  3 Adelie    Dream     male   2007 N30A2            42.3          21.2
##  4 Adelie    Torgersen male   2007 N8A1             34.6          21.1
##  5 Adelie    Dream     male   2007 N23A2            39.2          21.1
##  6 Adelie    Torgersen male   2007 N9A2             42.5          20.7
##  7 Adelie    Torgersen male   2007 N3A2             39.3          20.6
##  8 Chinstrap Dream     male   2007 N68A2            51.7          20.3
##  9 Adelie    Torgersen <NA>   2007 N5A2             42            20.2
## 10 Adelie    Dream     male   2007 N24A1            38.8          20
## # ... with 100 more rows, and 2 more variables: flipper_length_mm <dbl>,
## #   body_mass_g <dbl>
```

Filtering

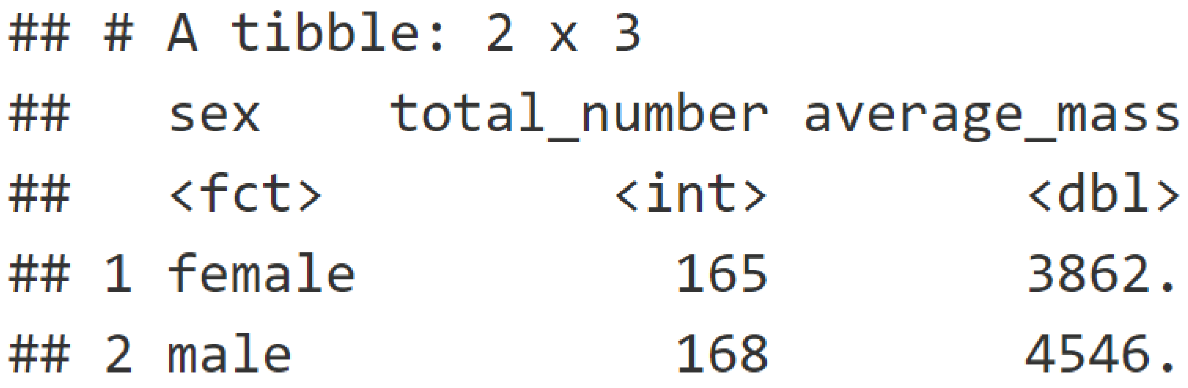
```r
# minimum body mass (kg) of female penguins from the Dream
# island in 2007
penguin_df_wide %>%
  filter(island == 'Dream', sex == 'female') %>%
  # calculate body mass in kg
  transmute(body_mass_kg = body_mass_g / 1000) %>%
  slice_min(body_mass_kg)
```

```
## # A tibble: 1 x 1
##   body_mass_kg
##          <dbl>
## 1          2.7
```

Group and Summarize Data

```
penguin_df_wide %>%
   # group for females and males
  group_by(sex) %>%
  # summarize number of penguins and average mass for each
group
summarise (total_number = n(),
           average_mass = mean(body_mass_g, na.rm = T))
```

Group and Summarize Data



```
## # A tibble: 2 x 3
##   sex    total_number average_mass
##   <fct>         <int>        <dbl>
## 1 female          165        3862.
## 2 male            168        4546.
```

Exercises Part 6

The rest of tidyverse

Visualization

Working on stings

Manipulating dates

Newer better data.frame

All about factors

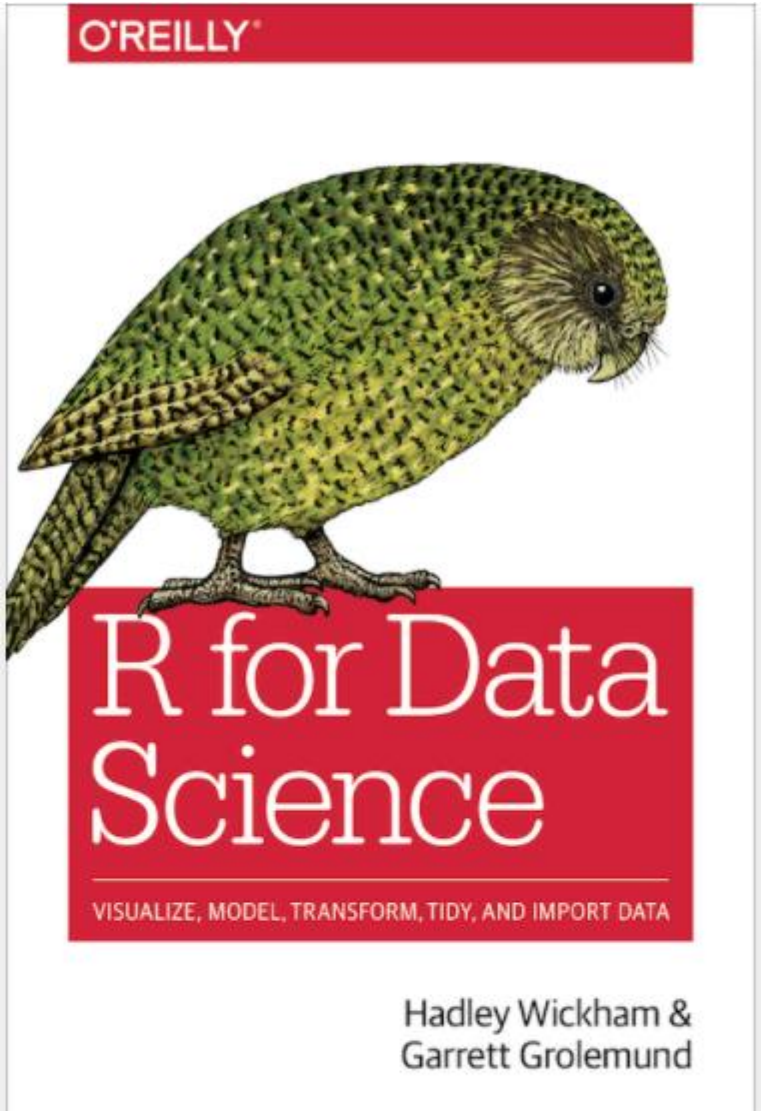
Advanced programming

The rest of tidyverse

From beginner to advanced guide into tidyverse

https://r4ds.had.co.nz/

https://github.com/tidyverse/tidyverse



@RLadiesAMS
#RLadies #rstats

Thank you!

✉ nutsa.nanuashvili@gmail.com

@Nutsa_Nanuash

@Nutsa-N

https://www.linkedin.com/in/nutsa-nanuashvili/