Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

bot²

Introduction to the tidyverse

Roman M. Link

University of Würzburg

October 30, 2019

What is the tidyverse?

- Suite of interrelated R packages that share a common design philosophy, grammar, and data structures

What is the tidyverse?

- Suite of interrelated R packages that share a common design philosophy, grammar, and data structures
  - Core `tidyverse`: `ggplot2`, `dplyr`, `tidyr`, `readr`, `purrr`, `tibble`, `stringr`, and `forcats`

What is the tidyverse?

- Suite of interrelated R packages that share a common design philosophy, grammar, and data structures
  - Core `tidyverse`: `ggplot2`, `dplyr`, `tidyr`, `readr`, `purrr`, `tibble`, `stringr`, and `forcats`
  - hundreds of additional packages that follow the same principles

What is the tidyverse?

- Suite of interrelated R packages that share a common design philosophy, grammar, and data structures
  - Core `tidyverse`: `ggplot2`, `dplyr`, `tidyr`, `readr`, `purrr`, `tibble`, `stringr`, and `forcats`
  - hundreds of additional packages that follow the same principles
- Core packages are aimed at data handling, processing and visualisation, **not statistical analysis**

Philosophy

- Code that mimics human language is easier to understand

Philosophy

- Code that mimics human language is easier to understand
  - the `tidyverse` is centered around functions

Philosophy

- Code that mimics human language is easier to understand
  - the `tidyverse` is centered around functions
  - functions perform the role of verbs
    $\rightarrow$ describe what is done with an object

Philosophy

- Code that mimics human language is easier to understand
  - the `tidyverse` is centered around functions
  - functions perform the role of verbs
    $\rightarrow$ describe what is done with an object
  - code follows a simple grammatical structure

Philosophy

- Code that mimics human language is easier to understand
  - the `tidyverse` is centered around functions
  - functions perform the role of verbs
    $\rightarrow$ describe what is done with an object
  - code follows a simple grammatical structure
  - in a tidy workflow, multiple interactions with the same object can be read as sentences

Design

- Consistent style around all packages

Design

- Consistent style around all packages
  - the names of functions that perform one class of operations share a common beginning (eg. `summarize_at()`, `summarize_if()`, `summarize_all()`)

Design

- Consistent style around all packages
  - the names of functions that perform one class of operations share a common beginning (eg. `summarize_at()`, `summarize_if()`, `summarize_all()`)
  - the `data` argument comes always first

Design

- Consistent style around all packages
  - the names of functions that perform one class of operations share a common beginning (eg. `summarize_at()`, `summarize_if()`, `summarize_all()`)
  - the `data` argument comes always first
  - all functions designed to work with *tidy* data structures

The concept of "tidy data"

- General structure of data: collection of *values* that belong to a *variable* and an *observation*

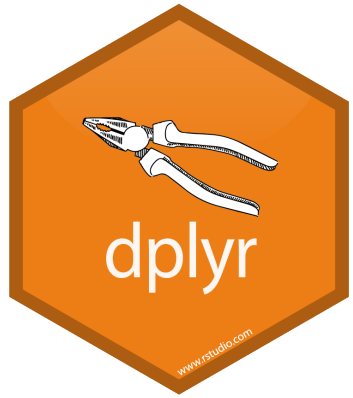
The concept of "tidy data"

- General structure of data: collection of *values* that belong to a *variable* and an *observation*
- Datasets are tidy when
  1. Each *variable* forms a column
  2. Each *observation* forms a row
  3. Each type of *observational unit* forms a table

The concept of "tidy data"

- General structure of data: collection of *values* that belong to a *variable* and an *observation*
- Datasets are tidy when
  1. Each *variable* forms a column
  2. Each *observation* forms a row
  3. Each type of *observational unit* forms a table
- Leading paradigm for data handling in the `tidyverse`

dplyr: a toolbox for data wrangling



- Large number of important `tidyverse` verbs
- Focus on data processing
- Clean and easy-to-follow syntax

- Examples
  - `filter()`: filter rows of a dataframe
  - `select()`: select columns of a dataframe
  - `mutate()`: change columns in a dataframe
  - `summarize()`: compute summary statistics of columns in a dataframe
  - `group_by()`: group a data.frame by one or several variables
  - `full_join()` / `left_join()` / `right_join()` etc.: join dataframes using database joins

magrittr: something in the pipeline

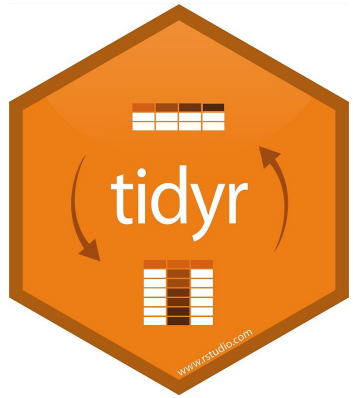
- Defines pipeline operator %>% as used in `dplyr`
- Allows to concatenate series of operations by using the output of one function as the input of the next function
- By default, piped objects go to the first argument of a function (which in tidyverse functions always is a dataset)
  - **without pipe:**

```
summarize(
  mutate(data, z = x + y),
  mean = mean(z), sd = sd(z))
```

  - **with pipe:**

```
data %>%
  mutate(z = x + y) %>%
  summarize(mean = mean(z), sd = sd(z))
```

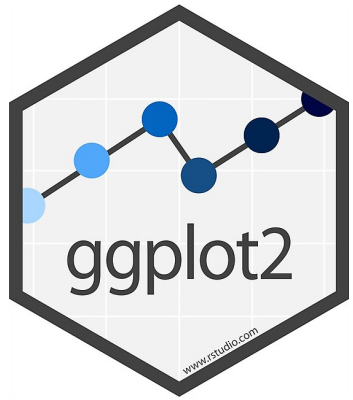
tidyr: making your data tidier



- Allows to switch from long table to wide table format and vice versa

- `gather()`/`pivot_longer()`: conversion to long format
- `spread()`/`pivot_wider()`: conversion to wide format

ggplot2: a grammar of graphics

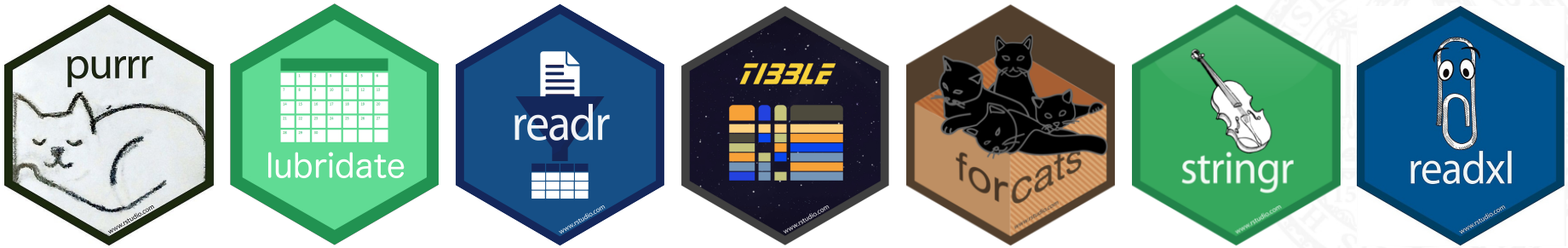


- The most powerful graphics package in R
- Very complex graphics with few lines of code

- Bit tricky to use, but absolutely worth the effort
- One of the earliest `tidyverse packages` - I use it since 2010

Other notable packages

- `purrr` - tools for functional programming
- `lubridate` - handling dates and times
- `tibble` - improved `data.frames`
- `readr` - improved data reading functions
- `forcats` - factor handling
- `stringr` - handling of character strings
- `readxl` - reading Excel files directly
- ...



Example: Costa-Rican tree diameters

- Tree diameters and corresponding comments in a wide table

```
> library(tidyverse)
> library(lubridate)
> data <- read_csv("cangreja.csv", col_types = cols(.default = col_character()))
> data
# A tibble: 40 x 30
   site  ID    family species dbh_20150803 comment_20150803 dbh_20151114 comment_20151114 dbh_20160219 comment_20160219 dbh_20160513
   <chr> <chr> <chr>  <chr>   <chr>         <chr>           <chr>         <chr>           <chr>         <chr>           <chr>
 1 Cang… CA_0… Myris… Virola… 334.7         NA              335.4         NA              335.5         NA              336.3
 2 Cang… CA_0… Vochy… Vochys… 801.8         NO cambio de pu… 805.4        NA              805.5         NA              805.6
 3 Cang… CA_0… Lecyt… Lecyth… 322.4         NA              324.6         error de transc… 326.6        NA              329.4
 4 Cang… CA_0… Lecyt… Gustav… 282.3         NA              282.4         NA              282.6         NA              282.4
 5 Cang… CA_0… Melia… Carapa… 366.4         Ajuste dendróme… 368.6        NA              369.9         NA              372.1
 6 Cang… CA_0… Vochy… Vochys… 662.7         NA              664.9         NA              665.8         NA              668.4
 7 Cang… CA_0… Vochy… Vochys… 809.2         Ajuste dendróme… 809.9        NA              809.8         Ajuste dendróme… 810.5
 8 Cang… CA_0… Melia… Carapa… 321.2         NA              321.4         NA              321.4         NA              321.5
 9 Cang… CA_0… Lecyt… Lecyth… 374.3         Ajuste dendróme… 374.6        NA              374.4         NA              375
10 Cang… CA_0… Melia… Carapa… 287.6         NA              287.4         NA              287.3         NA              287.3
# … with 30 more rows, and 19 more variables: comment_20160513 <chr>, dbh_20160813 <chr>, comment_20160813 <chr>,
#   dbh_20161026 <chr>, comment_20161026 <chr>, dbh_20170125 <chr>, comment_20170125 <chr>, dbh_20170327 <chr>,
#   comment_20170327 <chr>, dbh_20170528 <chr>, comment_20170528 <chr>, dbh_20170731 <chr>, comment_20170731 <chr>,
#   dbh_20171018 <chr>, comment_20171018 <chr>, dbh_20180122 <chr>, comment_20180121 <chr>, dbh_20180328 <chr>,
#   comment_20180328 <chr>
```

Example: Costa-Rican tree diameters

- Tree diameters and corresponding comments in a wide table
- Dates in header

```
> library(tidyverse)
> library(lubridate)
> data <- read_csv("cangreja.csv", col_types = cols(.default = col_character()))
> data
# A tibble: 40 x 30
   site  ID    family species dbh_20150803 comment_20150803 dbh_20151114 comment_20151114 dbh_20160219 comment_20160219 dbh_20160513
   <chr> <chr> <chr>  <chr>   <chr>         <chr>           <chr>         <chr>           <chr>         <chr>           <chr>
 1 Cang… CA_0… Myris… Virola… 334.7         NA              335.4         NA              335.5         NA              336.3
 2 Cang… CA_0… Vochy… Vochys… 801.8         NO cambio de pu… 805.4        NA              805.5         NA              805.6
 3 Cang… CA_0… Lecyt… Lecyth… 322.4         NA              324.6         error de transc… 326.6        NA              329.4
 4 Cang… CA_0… Lecyt… Gustav… 282.3         NA              282.4         NA              282.6         NA              282.4
 5 Cang… CA_0… Melia… Carapa… 366.4         Ajuste dendróme… 368.6        NA              369.9         NA              372.1
 6 Cang… CA_0… Vochy… Vochys… 662.7         NA              664.9         NA              665.8         NA              668.4
 7 Cang… CA_0… Vochy… Vochys… 809.2         Ajuste dendróme… 809.9        NA              809.8         Ajuste dendróme… 810.5
 8 Cang… CA_0… Melia… Carapa… 321.2         NA              321.4         NA              321.4         NA              321.5
 9 Cang… CA_0… Lecyt… Lecyth… 374.3         Ajuste dendróme… 374.6        NA              374.4         NA              375
10 Cang… CA_0… Melia… Carapa… 287.6         NA              287.4         NA              287.3         NA              287.3
# … with 30 more rows, and 19 more variables: comment_20160513 <chr>, dbh_20160813 <chr>, comment_20160813 <chr>,
#   dbh_20161026 <chr>, comment_20161026 <chr>, dbh_20170125 <chr>, comment_20170125 <chr>, dbh_20170327 <chr>,
#   comment_20170327 <chr>, dbh_20170528 <chr>, comment_20170528 <chr>, dbh_20170731 <chr>, comment_20170731 <chr>,
#   dbh_20171018 <chr>, comment_20171018 <chr>, dbh_20180122 <chr>, comment_20180121 <chr>, dbh_20180328 <chr>,
#   comment_20180328 <chr>
```

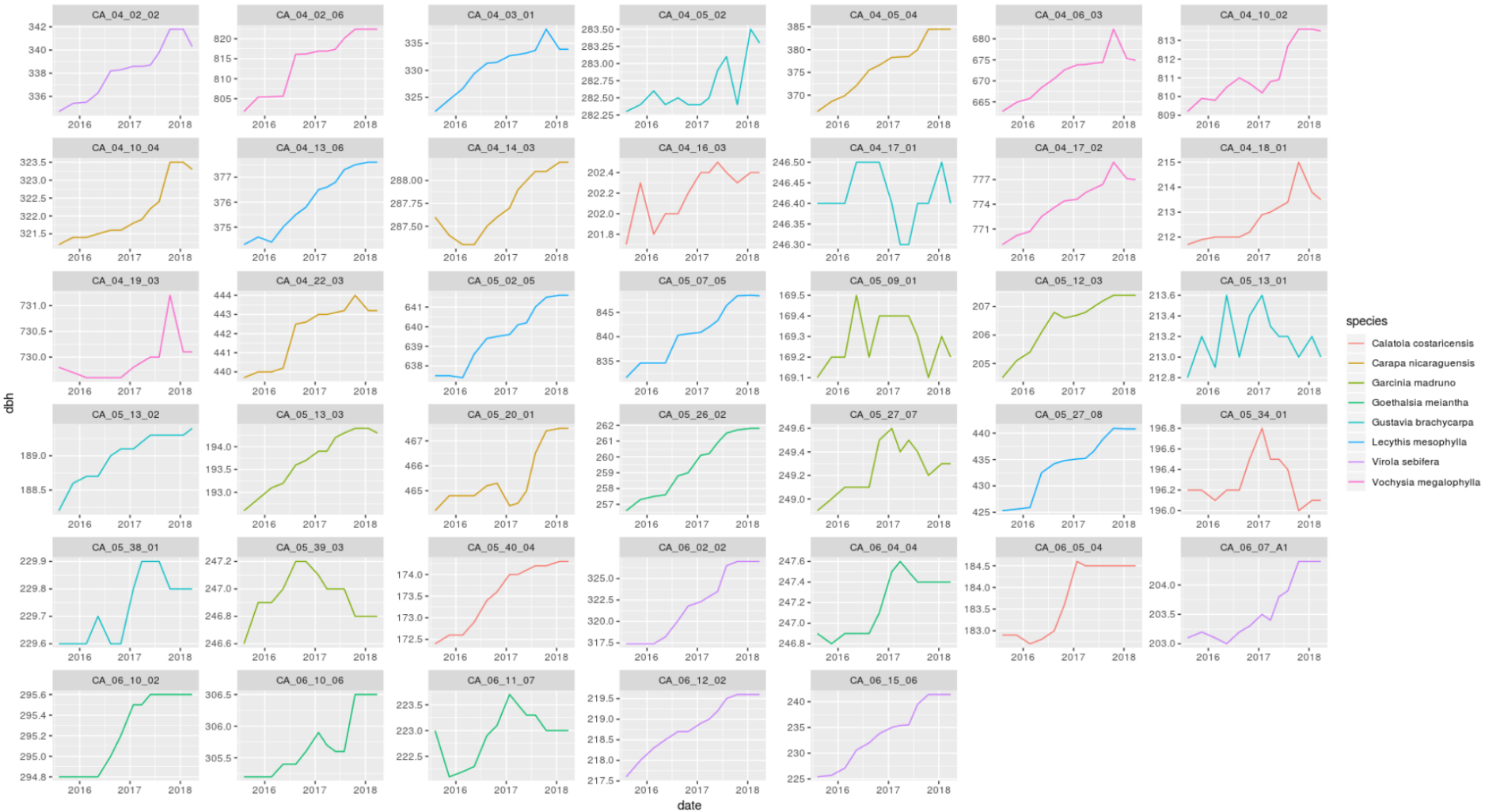
Example: Costa-Rican tree diameters

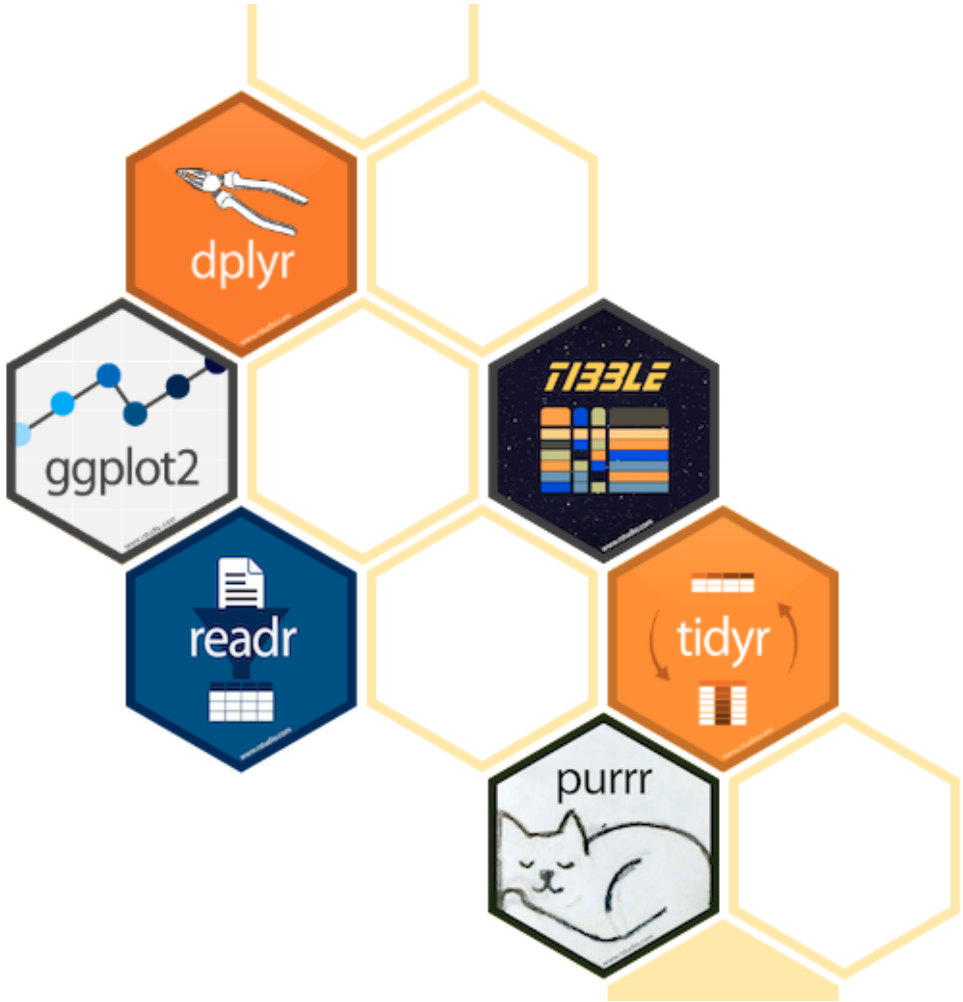
- Tree diameters and corresponding comments in a wide table
- Dates in header
- How to tidy this mess up and plot diameter vs. time?

```
> library(tidyverse)
> library(lubridate)
> data <- read_csv("cangreja.csv", col_types = cols(.default = col_character()))
> data
# A tibble: 40 x 30
   site  ID    family species dbh_20150803 comment_20150803 dbh_20151114 comment_20151114 dbh_20160219 comment_20160219 dbh_20160513
   <chr> <chr> <chr>  <chr>   <chr>         <chr>            <chr>         <chr>            <chr>         <chr>            <chr>
 1 Cang… CA_0… Myris… Virola… 334.7         NA               335.4         NA               335.5         NA               336.3
 2 Cang… CA_0… Vochy… Vochys… 801.8         NO cambio de pu… 805.4         NA               805.5         NA               805.6
 3 Cang… CA_0… Lecyt… Lecyth… 322.4         NA               324.6         error de transc… 326.6         NA               329.4
 4 Cang… CA_0… Lecyt… Gustav… 282.3         NA               282.4         NA               282.6         NA               282.4
 5 Cang… CA_0… Melia… Carapa… 366.4         Ajuste dendróme… 368.6         NA               369.9         NA               372.1
 6 Cang… CA_0… Vochy… Vochys… 662.7         NA               664.9         NA               665.8         NA               668.4
 7 Cang… CA_0… Vochy… Vochys… 809.2         Ajuste dendróme… 809.9         NA               809.8         Ajuste dendróme… 810.5
 8 Cang… CA_0… Melia… Carapa… 321.2         NA               321.4         NA               321.4         NA               321.5
 9 Cang… CA_0… Lecyt… Lecyth… 374.3         Ajuste dendróme… 374.6         NA               374.4         NA               375
10 Cang… CA_0… Melia… Carapa… 287.6         NA               287.4         NA               287.3         NA               287.3
# … with 30 more rows, and 19 more variables: comment_20160513 <chr>, dbh_20160813 <chr>, comment_20160813 <chr>,
#   dbh_20161026 <chr>, comment_20161026 <chr>, dbh_20170125 <chr>, comment_20170125 <chr>, dbh_20170327 <chr>,
#   comment_20170327 <chr>, dbh_20170528 <chr>, comment_20170528 <chr>, dbh_20170731 <chr>, comment_20170731 <chr>,
#   dbh_20171018 <chr>, comment_20171018 <chr>, dbh_20180122 <chr>, comment_20180121 <chr>, dbh_20180328 <chr>,
#   comment_20180328 <chr>
```

```
> datalong <- data %>%
+   pivot_longer(cols = -(site:species),
+                names_to = c("var", "date"),
+                names_sep = "_") %>%
+   pivot_wider(names_from = var,
+               values_from = value) %>%
+   mutate(dbh  = as.numeric(dbh),
+          date = ymd(date))
> datalong
# A tibble: 560 x 7
   site     ID          family        species          date         dbh comment
   <chr>    <chr>       <chr>         <chr>            <date>     <dbl> <chr>
 1 Cangreja CA_04_02_02 Myristicaceae Virola sebifera 2015-08-03  335. NA
 2 Cangreja CA_04_02_02 Myristicaceae Virola sebifera 2015-11-14  335. NA
 3 Cangreja CA_04_02_02 Myristicaceae Virola sebifera 2016-02-19  336. NA
 4 Cangreja CA_04_02_02 Myristicaceae Virola sebifera 2016-05-13  336. NA
 5 Cangreja CA_04_02_02 Myristicaceae Virola sebifera 2016-08-13  338. NA
 6 Cangreja CA_04_02_02 Myristicaceae Virola sebifera 2016-10-26  338. NA
 7 Cangreja CA_04_02_02 Myristicaceae Virola sebifera 2017-01-25  339. NA
 8 Cangreja CA_04_02_02 Myristicaceae Virola sebifera 2017-03-27  339. NA
 9 Cangreja CA_04_02_02 Myristicaceae Virola sebifera 2017-05-28  339. NA
10 Cangreja CA_04_02_02 Myristicaceae Virola sebifera 2017-07-31  340. NA
# … with 550 more rows
```

```
> ggplot(datalong, aes(x = date, y = dbh, col = species, group = ID)) +
+   geom_line() +
+   facet_wrap(~ID, scales = "free")
```





Room for questions