Josep Carreras LEUKAEMIA Research Institute

Data visualization in R with the ggplot2 package

Angelika Merkel (Head of Bioinformatics Unit IJC)
17/11/2023

http://www.carrerasresearch.org/en/units/bioinformatics

IJC Bioinformatics

Materials

Course book:

R for Data Science, 2nd edition (Wickham, Cetinkaya-Rundel and Grolemund, 2023)

RStudio course server:

https://rstudio1.services.carrerasresearch.org/

BIT course webpage:

https://ijcbit.github.io/Workshops/

Who we are

Bioinformatics Unit IJC

Angelika Merkel

(Head of Unit)

Izar de Villasante

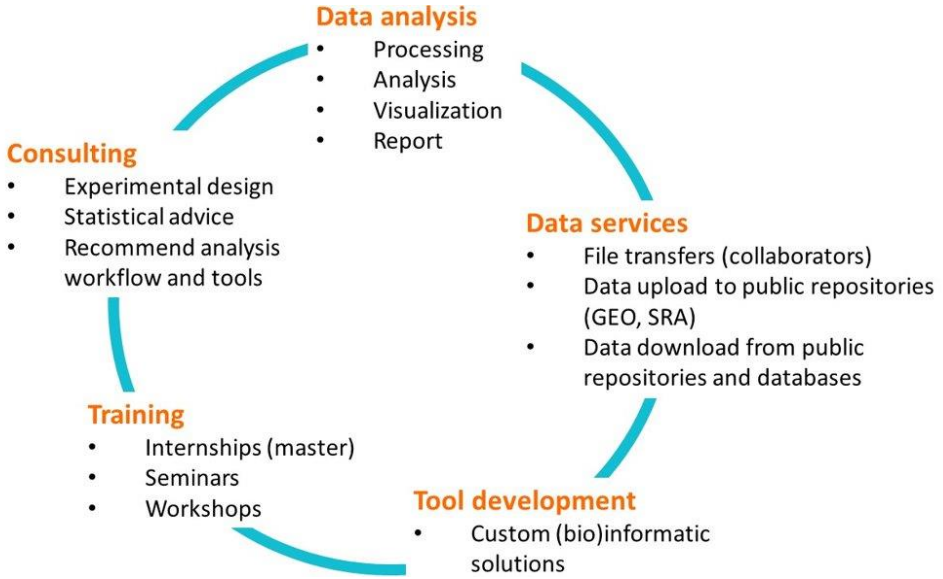
(Bioinformatician)

Emilio Lario

(Software engineer)

Marta Meroño

(Master student)



Office: Sala Prof. Alber Grañena (1st floor); phone: 4300

https://carrerasresearch.sharepoint.com/sites/BIT

https://www.carrerasresearch.org/en/bioinformatics-unit

Data visualization in R

Multiple packages exist:

- {graphics} for basic graphic
- {lattice}, for high level data visualizations for multivariate data
- {ComplexHeatmap}, {pheatmap} for specialized graphic such as heatmaps
- {ggplot2} coherent system for data visualizations based on 'the grammar of graphics'

The Tidyverse

Tidy verse = collection of approx. 25 packages for manipulation, visualization, transformation of "tidy data" (incl ggplot2)

Tidy data (and data frames aka 'tibbles'):
= each value is placed in its own “cell”,
each variable in its own column,
and each observation in its own row.


```
table1
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
```

```
table2
#> # A tibble: 12 × 4
#>   country      year type           count
#>   <chr>       <dbl> <chr>          <dbl>
#> 1 Afghanistan  1999 cases            745
#> 2 Afghanistan  1999 population  19987071
#> 3 Afghanistan  2000 cases           2666
#> 4 Afghanistan  2000 population  20595360
#> 5 Brazil       1999 cases           37737
#> 6 Brazil       1999 population 172006362
#> # ℹ 6 more rows
```

```
table3
#> # A tibble: 6 × 3
#>   country      year rate
#>   <chr>       <dbl> <chr>
#> 1 Afghanistan  1999 745/19987071
#> 2 Afghanistan  2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> 4 Brazil       2000 80488/174504898
#> 5 China        1999 212258/1272915272
#> 6 China        2000 213766/1280428583
```

Base R and the tidyvers

BaseR

- better for software development
- better for running quick simulations
- generally faster performance
- more appealing to users with previous programming experience

Use if:

- Most of your work involves software or package development, advanced statistical procedures, or computationally expensive operations
- You're used to other languages that have more in common with Base-R
- Most of your collaborators and online network use it too

Tidyverse

- ease of use, functions have the same structure and easier names, enables reading functions as instructions
- quick and easy data manipulation
- grouping datasets with many variable for summary statistics with dplyr
- over 25 packages in the tidyverse, each requiring its own updates to stay current
  -> adds overhead, difficult to reproduce, limits submission to code repros as R cran or bioconductor

Use if:

- Most of your work involves data cleaning, visualization, and common statistics
- You're newer to R and find it easier to read and understand than base-R
- Most of your collaborators and online network use it too

Practical session

R for Data Science, 2nd edition (Wickham) Chapter: Data Visualization

Questions?

Thank you!

Further resources

Tutorials:

- Datanovia

Inspirations with code examples:

- R gallery