Introduction to R - Young Researchers Fellowship Program

Lecture 2 - Introduction to the tidyverse and data importing

Daniel Sánchez Pazmiño

Laboratorio de Investigación para el Desarrollo del Ecuador

September 2024

The tidyverse or how modern R code is being written

Tidy data

Tidying your data means storing it in a consistent form that matches the semantics of the dataset with how it is stored (Wickham et al, 2023)

- Tidy data is a standard way of mapping the meaning of a dataset to its structure.
- A dataset is messy or tidy depending on how rows, columns, and tables are matched up with observations, variables, and types.
- In tidy data:
    - Each variable forms a column.
    - Each observation forms a row.
    - Each type of observational unit forms a table.

Who came up with this?

- Hadley Wickham introduced the concept of tidy data in his paper “Tidy Data” published in the Journal of Statistical Software in 2014.
- In the R for Data Science book (R4DS), the tidyverse is introduced as a collection of R packages designed to tidy data and work with it in a data science context.
- The tidyverse philosophy revolutionized the way R code is written and data is handled, making it more efficient and easier to understand.

The data science vs. the research perspective

- According to Hadley Wickham, *data science is an exciting discipline that allows you to transform raw data into understanding, insight, and knowledge*.
- This means we need not be afraid that the tidyverse will make us lose the ability to do research.
    - In this view, data science is not only predictive modeling.

The tidyverse steps in a data science project

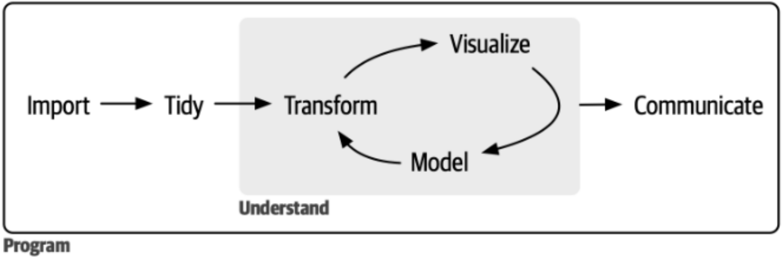

Figure 1: Tidy steps in Data

Source: R for Data Science by Wickham, Cetinkanya-Rundel & Grolemund (2023)

The tidyverse steps in a research project

1. **Import**: read data from a file or database.
2. **Tidy**: transform the data into a format that makes it easy to work with.
3. **Transform**: perform operations on the data to create new variables or summaries.

(together, transform and tidy are often referred to as wrangling - it feels like a wrestling match sometimes!)

4. **Visualize**: generate static graphics for exploratory data analysis.
5. **Model**: fit quantitative models to understand relationships between variables, complementary to visualization.
6. **Communicate**: generate reports or dashboards, or create a Shiny app. The most important step!

Where does programming fit in?

- Programming is an outer step in the process as it will be used all along the way.
- We use programming to automate the steps in the process and solve problems effectively.

The tidyverse packages

- The tidyverse is a collection of R packages that share an underlying design philosophy, grammar, and data structures.
- The packages in the tidyverse are designed to work together, and it is easier to learn them together.

The core tidyverse packages

- **ggplot2**: for data visualization.
- **dplyr**: for data manipulation.
- **tidyr**: for data tidying.
- **readr**: for data import.
- **purrr**: for functional programming.
- **tibble**: for tibbles, a modern reimagining of data frames.
- **stringr**: for strings.
- **forcats**: for factors.

Installing the tidyverse

- We install them all at once through the `tidyverse` package, which is a meta-package that installs the core tidyverse packages.

```
install.packages("tidyverse")
```

Importing data with the tidyverse

- The `readr` package is part of the tidyverse and provides a fast and friendly way to read rectangular data.

Tidyverse vs. base R - some brief comments

- The tidyverse often follows a certain style of programming that is different from base R.
- The tidyverse is typically easier to learn, but it is not the only way to write R code.