
# An Introduction to the Tidyverse

If you've used R you've probably stumbled into the name ggplot2, dplyr or tidyverse. These names are R packages created by RStudio's Chief Scientist Hadley Wickham. The tidyverse is a series of packages that work extremely well with each other and have been created with the aim of easing the data analysis process. As stated by its main author, the philosophy of the tidyverse is to allow the analyst to concentrate on the substantive questions rather than on technicalities of data analysis.

The tidyverse is comprised of over 15 packages, each one tackling a specific process within the data exploration process. In these series of seminars we will introduce you to the core packages and concepts that will allow you to start conducting analysis right away.

The outline of the seminar series is as follows:

In the first seminar we will start with the philosophy of the tidyverse. We will discuss why these packages work well together and how they complement each other. Next off we'll jump right into an analysis using the dplyr and ggplot2 packages in order to give you the tools to start using it. We will learn how to read and save data from Excel, Stata, SPSS, among other software. We'll finish off with a series of exercises that will test your understanding of the concepts.

The second seminar will concentrate on the two things non-R users despise from R: data cleaning and data wrangling. The tidyr and the dplyr packages have turned the tables around by making data manipulation extremely easy and intuitive. We'll touch upon the basic 'verbs'

contained in dplyr and the basic transformation techniques available in tidyr. We'll also introduce the tidyverse's version of data frames called tibbles. As in every other session, we will finish with a series of exercises, this time related to data cleaning.

The third seminar will introduce ggplot2, one of R's most used package. This session will show you how to create graphs following the notion of 'the grammar of graphics'. The grammar of graphics consists of a series of verbs that allow you to construct graphs just as if you were constructing a sentence. This package allows you to create simple as well as complicated graphs following a very intuitive syntax and a couple of simple rules.

The fourth session will show you the power of functional programming. This means that we'll be able to automate things and concentrate on analyzing rather than writing code. This session will show you how the tidyverse can be incorporated with the statistical side of R by conducting some simple linear modelling and visualizing the results.

Finally, the fifth session will wrap everything by showing you how to communicate your results and make your analysis reproducible. We'll introduce you to Rmarkdown, a simple feature that allows you to write text and code in the same document. Yes! That means you can truly have your analysis together with your writeups all in one reproducible document. We will finish off by exploring a dataset and documenting everything into one nicely formatted document.

These sessions will be a simplification of what the tidyverse is capable of. For those interested in diving deep into what you can do with the it, we point you towards the book R for Data Science, written by Garrett Grolemund and Hadley Wickham (the main author of the tidyverse). These seminars are based extensively on this book which can be read for free here or bought here.

We intend to make all seminars as interactive as possible. For that, we plan to learn the tidyverse by using it. We will develop the exercises for each class and give you the tools to start your analysis on the spot.

**Requirements:**

1. You should be familiar with R to the point that you understand vectors, data frames, lists, packages and the core functions in R. Ideally, you should have conducted some type of analysis in R in the past.
2. Some familiarity with statistics as we will touch upon linear modelling and descriptive statistics.
3. You are motivated enough to put up with some initial (and continuing) frustration :)

If you don't comply with any of the first two requirements, you're still more than welcome to come but be aware that we will not teach these concepts in class.

## Seminars Program:

*Please, bring your laptop with R and Rstudio already installed (and possibly with the battery fully charged although there will be charging stations in the classroom). All seminars are free but if you're interested in participating, please email Jorge Cimentada at jorge.cimentada@upf.edu to be included in our mailing list.*

**Seminar 1. Data Exploration –**

**Thursday 16th of February - 15.00 to 17.00;**

**Room: 24.019.**

**Instructor: Jorge Cimentada**

- Introduction to the tidyverse
- Data visualization with ggplot2
- A brief introduction to the pipe
- Data transformation with dplyr
- Data import with haven and readr
- Create your own Rstudio project
- Exploratory Data Analysis (EDA)

**Seminar 2. Data cleaning made easy**

**Thursday 23rd of February - 15.00 to 17.00;**

**Room: 24.019.**

**Instructor: Jorge Cimentada**

- Tibbles - the new data frame
- What is a ‘tidy’ dataset?
- An introduction to tidyr and its verbs
- Why is R so difficult at data cleaning?
- Data manipulation the dplyr way
- What else should you learn? Factors, Dates and Regular expressions.

**Seminar 3. ggplot2: The grammar of graphics**

**Thursday 2nd of March - 15.00 to 17.00;**

**Room: 24.019.**

**Instructor: Robert T. Lange**

- Philosophy - Plots as adaptable objects
- An Introduction to the multilayer system of ggplot2
    - Data, geom-layers, themes, legend, labels, facets
- When to use? - Comparison to internal plot function
- Different types of plots in ggplot2
    - Univariate, Bivariate, Multivariate
- Fine-tuning plots (adding LaTex in Labels, etc.)
- One complete ggplot2 workflow example
- (Dynamic visualization with gganimate, 3d-plotting with plotly)

**Seminar 4. Automating everything: we should always strive for it!**

**Thursday 9th of March - 15.00 to 17.00;**

**Room: 24.019.**

**Instructor: Jorge Cimentada**

- A subtle introduction to loops and the apply family.
- Our first function
- The purrr package and how it replaces loops
- Integrating statistical modelling into the tidyverse
- Visualizing regression models with broom and ggplot2

**Seminar 5. Making your analysis reproducible**

**Thursday 16th of March - 15.00 to 17.00;**

**Room: 24.019.**

**Instructor: Jorge Cimentada**

- Philosophy of reproducibility
- An introduction to Rmarkdown
- Creating your first report
- A subtle introduction to Sweave.
- At this point we will provide everyone with a dataset and set a research question for you to explore. You'll create a reproducible document where you'll include both your comments and your results as we go through everyone's desks to help.

**The organizers of the tidyverse seminar**

**Jorge Cimentada, Robert T Lange and Bruno Arpino**