Packages

Dr. Aline Coutinho

What is a package?

A *package* is a collection of functions and datasets. Imagine a set of tools that you need to execute a specific analytical computing or command. You may want to run a regression, create a results table, visualize a model, or even write a full report or an article. Using a package is analogous to going to a *library* to rent certain tools in order to perform a desired activity.

How to download and run a package?

Packages can be downloaded from repositories. The most common repository is `CRAN` (the *comprehensive R archive network*). Another widely used is `Github` (a famous repository for open source projects, not only R-related).

To download a package you must write the command:

```
install.packages("name of package")
or
install.packages("devtools") followed by install_github()
```

To use a package, you have to call it from your *library*, the place where your computer keeps your packages (most likely a folder on your computer). You can load a package by typing the following function:

```
library(name of package).
```

What packages will you most likely need for this course?

There are a few packages I would like you to download for this course. I will talk in-depth about them in class. For now, let's just do a brief overview.

The tidyverse package (click here for more information).

The tidyverse is a collection of R packages that we will frequently use during data analysis and visualization. It includes the famous ggplot2.

ggplot 2 is an R package designed by Hadley Wickham for producing statistical and data graphics. It has a simple set of core principles based on Wilkinson's Grammar of Graphics (Wickham, 2016; Wilkinson, 2005). It is designed to work iteratively, meaning that one can start with a layer showing the raw data, and then incrementally adding other layers of annotations and statistical summaries. The commands used to build a new graphic can easily incorporate new datasets (this is great for datasets that are regularly updated).

But what is the grammar of graphics? As Hadley Wickham puts it:

The tidyverse

Components

Figure 1: The tidyverse package, figure by Steven M. Mortimer

> the grammar tells us that a statistical graphic is a mapping from data to aesthetic attributes (colour, shape, size) of geometric objects (points, lines, bars). The plot may also contain statistical transformations of the data and is drawn on a specific coordinate system. Facetting can be used to generate the same plot for different subsets of the dataset. It is the combination of these independent components that make up a graphic.
>
> — Wickham, 2016: 4.

All plots include:

- the **data** that one wants to visualize;
- the **mappings** describe how variables are mapped to *aesthetic* attributes;
- the **geoms**, short for the geometric representations of the data: points, lines, bars, pies, polygons;
- the **stats** or stattistical transformations of the data: the summarization of the data;
- the **scales** that draw legends and axes;
- the **coord** that describes the data coordinates;
- the **facets** that break and display the data into subsets;
- the **theme** which controls the display of the graphic, such as background colours and grids, font size, and so on.

The following is a typical set of commands to create a scatterplot with ggplot:

```
ggplot(dataset, aes(x = name of variable, y = name of variable)) + geom_point()
```

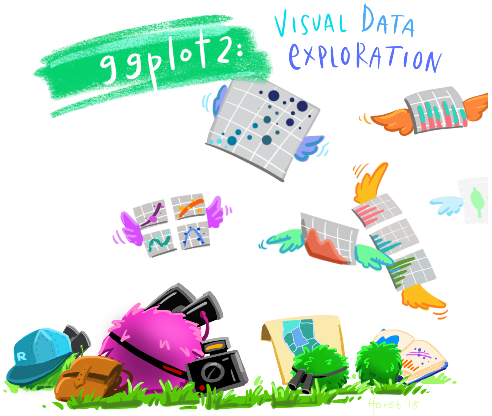

Figure 2: The ggplot2 package, figure by Allison Horst

To download and load `tidyverse` you must type the following commands:

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

The patchwork package (click here for more information).

This package was developed by Thomas Lin Pedersen. It allows you to seamlessly combine multiple graphs into one display.

```
install.packages("patchwork")
library(patchwork)
```

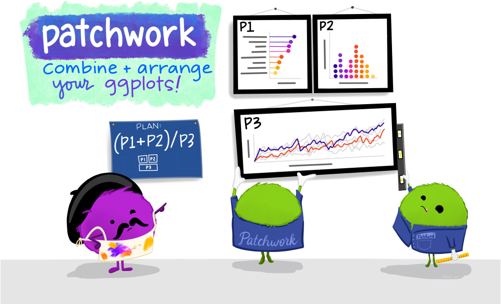

Figure 3: The patchwork package, figure by Allison Horst

The cansim package (click here for more information).

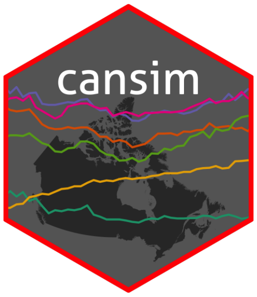

Figure 4: The cansim package, figure by Jens von Bergmann

This package was developed by Jens von Bergmann. It allows you to retrieve data from Statistics Canada.

I highly recommended that you watch two videos from Statistics Canada in order to better understand their data tables:

- The first video is about some changes in their website: https://www.statcan.gc.ca/eng/sc/video/new
- The second is a brief tutorial about their tables: https://www.statcan.gc.ca/eng/sc/video/howto

CANSIM tables are now replace by data tables that are dynamically updated as new data is collected. Does it render the `cansim` package useless? No. The reason is because there is a certain concordance of DOI (Digital Object Identifier) between the old cansim and the new tables used by Statistic Canada.

To download and open the package, write the following commands:

```
install.packages("cansim")
library(cansim)
```

The stargazer package (click here for more information)

The `stargazer` package creates high-quality regression and summary statistics tables. It saves an immense amout of time as its users don't need to create new tables everytime they tweak their dataset.

```
install.packages("stargazer")
library(stargazer)
```

The gt package (click here for more information)

The `gt` package creates a wide variety of tables. It is a fairly new package, so it currently only supports HTML output, but LaTeX and RTF are planned for the near future.

```
install.packages("gt")
library(gt)
```

Links to additional resources

A Modern Dive into R and the Tidyverse

FAQ about Statistics Canada's Tables

ggplot2 cheat sheet

More info on Stargazer

References

Grolemund, G. & Wickham, H. (2017). *R for Data Science.* O'Reilly Media.

Wickham, H. (2016). *ggplot 2: Elegant Graphics for Data Analysis.* (2nd edn). Springer.

Wilkinson, L. (2005). *The Grammar of Graphics: Statistics and Computing.* (2nd edn.). Springer