www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917

Le Bulletin de la Dialyse à Domicile

Introduction to data analysis with R software

Introduction to data visualization under R, with the add in Esquisse

Note : le texte original en version Française est disponible à la même adresse url : https://doi.org/10.25796/bdd.v2i3.21313

Claire Della Vedova

1087 chemin de Sainte Roustagne, 04100 MANOSQUE, France

Editor's note: The main purpose of the RDPLF is to help dialysis teams involved in home dialysis to evaluate their clinical practices and also to conduct studies based on anonymous exports of the registry data. To this end, since June 2019, a training series on the use of the R free-software is published quarterly with each issue of the Bulletin de la Dialyse à Domicile. The goal is to allow all teams to perform basic statistics and quickly visualize their data.

The first article in this tutorial was devoted to downloading and installing R software on Macintosh and PC computers: https://doi.org/10.25796/bdd.v2i2.20513.

This second article is devoted to graphic visualization of statistical data. The basic package used is usually ggplot2. But the latter requires a small learning phase that is not necessarily available time when the need to create a graphic for a presentation is urgent.

It may be interesting then to use a simple and fast learning graphical software to quickly showcase its digital data in an attractive form.

The Esquisse package meets this need and will be the subject of this article.
As in the first article, an example file from the RDPLF database will be used.

Claire Della Vedova is an engineer in biostatistics / data analyst, she uses the R software daily to analyze data. She has worked for more than 15 years in the fields of environment and health, and has trained many students and researchers in the use of R.

She runs since November 2017 the Franche Statistiques and Logiciel R whose goal is help beginners to better understand the standard statistical methods and to use the R software more effectively, especially through tutorials: https://statistique-et-logiciel-r.com/.

The total training is done over 15 months, at the rate of one article per quarter each publication of the BDD. This will leave ample time to assimilate and test the knowledge gained between each article. For those who wish to go faster, they can go to the blog(https://statistique-et-logiciel-r.com/).

Dates des prochaines parutions :

- article 3 (Dcember 2019) : initiation to ggplot2
- article 4 (April 2020) : automated statistical analysis reports with Rmarkdown
- article 5 (June 2020) : data manipulation (with dplyr, including the group_by and summarize functions)
- article 6 (Septembre 2020) : Descriptive analysis (statistical parameters and graphs) in the form of a dashboard with the flexboard package

Key words : biostatistic, epidemiology, R Software, RDPLF

journal officiel du Registre de Dialyse Péritonéale de Langue Française RDPLF www.rdplf.org

www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917

The use of this package is not complicated in itself, but it nevertheless requires investing a little time in its learning.

And sometimes, it's a bit frustrating, because when you're interested in making graphics under R, it's usually because you need it right now!

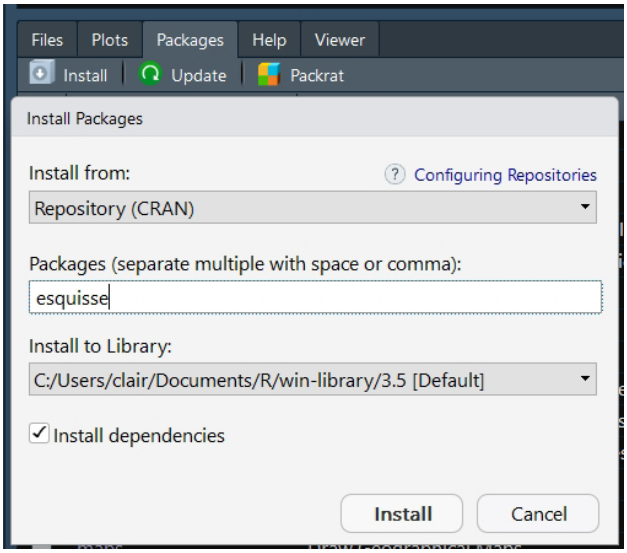
So in this article, I'll give you a tip to start using ggplot2 immediately. Without going through the learning box!

This trick is the addin «Esquisse»! This is a graphical interface that uses many features of ggplot2 but without code, only by manipulating labels (in drag and drop).

And icing on the cake, Sketch shows you the code ggplot2 corresponding to the graph! It's a little magic! We owe this tool to Victor Perrier of DreamRs, a data science consulting and expertise company specializing in R.3. Installation et accès de l'addin Esquisse

3.1 Installing «Esquisse» package

To use the Esquisse addin, it is necessary to install the package of the same name, for example using the R Studio installation tool:



It is also necessary to install the ggplot2 package, in the same way.

3.2 Import your data

To use the addin, it is also necessary that the data set on which the graphical representation (s) will be carried is imported into R. For example, we will, here, use the same data as in the first article, they can be downloaded in csv format at this address:

https://www.rdplf.org/exempleR/FichierExempleStat.csv

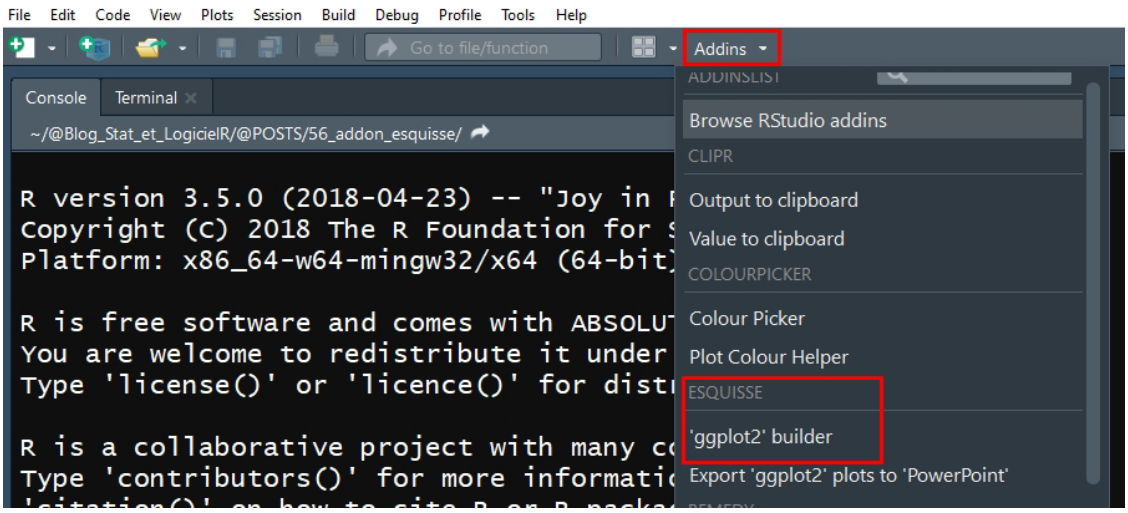
Once downloaded, place the csv file in the «data» folder of your R project, then use the following command:
mydata <- read.csv2("data/FichierExempleStat.csv")

www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917

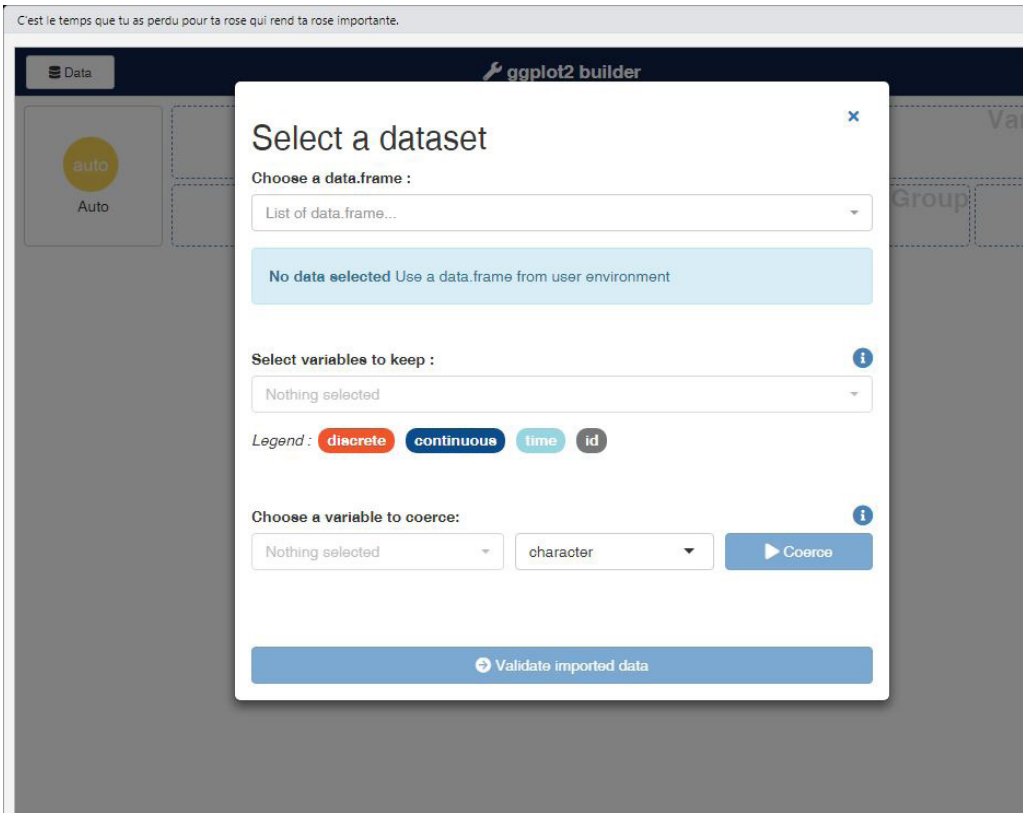
Notes: For more information on R projects and its «data» folder, see the article «Introduction to data analysis with R software» (https://www.bdd.rdplf.org/index.php/bdd/article/view/20513/19163)).

4. Open Esquisse

Now that everything is ready, you can open the graphic interfaceI, by going to the Addins menu, then choosing «ggplot2 builder» in the «ESQUISSE» section.



Here is the window that you should see appear:



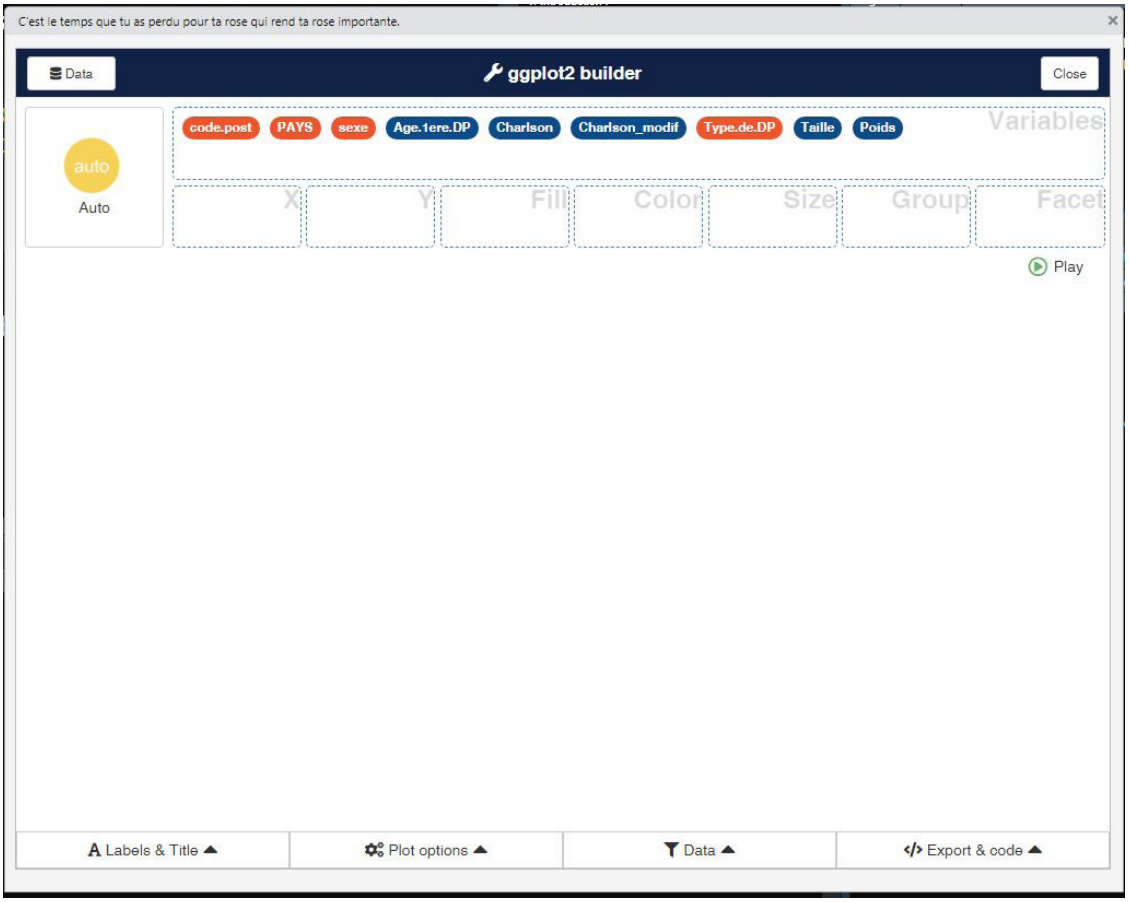
Choose your dataset, and click on «Validate imported data»!

www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917

5. Interface

5.1 Les éléments

L'interface apparaît alors, comme ceci:



The variables in the dataset are represented in the form of labels at the top.
In the part directly below, there are boxes, named:

X: used to define the variable represented on the X axis,
Y: used to define the variable represented on the Y axis,
Fill: used to define the variable which will control the color of the graphs containing «boxes», like the boxplots or the barplots for example,
Color: used to define the variable that will control the color of graphics containing points and / or lines etc ...
Size: used to define the variable that will control the size of the points,
Group: used to define a variable that will control the grouping of data in some graphics,
Facet: used to define a variable that will control the division of the graph into multiple entities (eg a scatterplot by Country).
Do not panic if you are lost, we will see this in examples.

5.2 option menus

Dans la partie inférieure de l'interface, se trouvent différents menus, qui permettent de personnaliser les

www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917

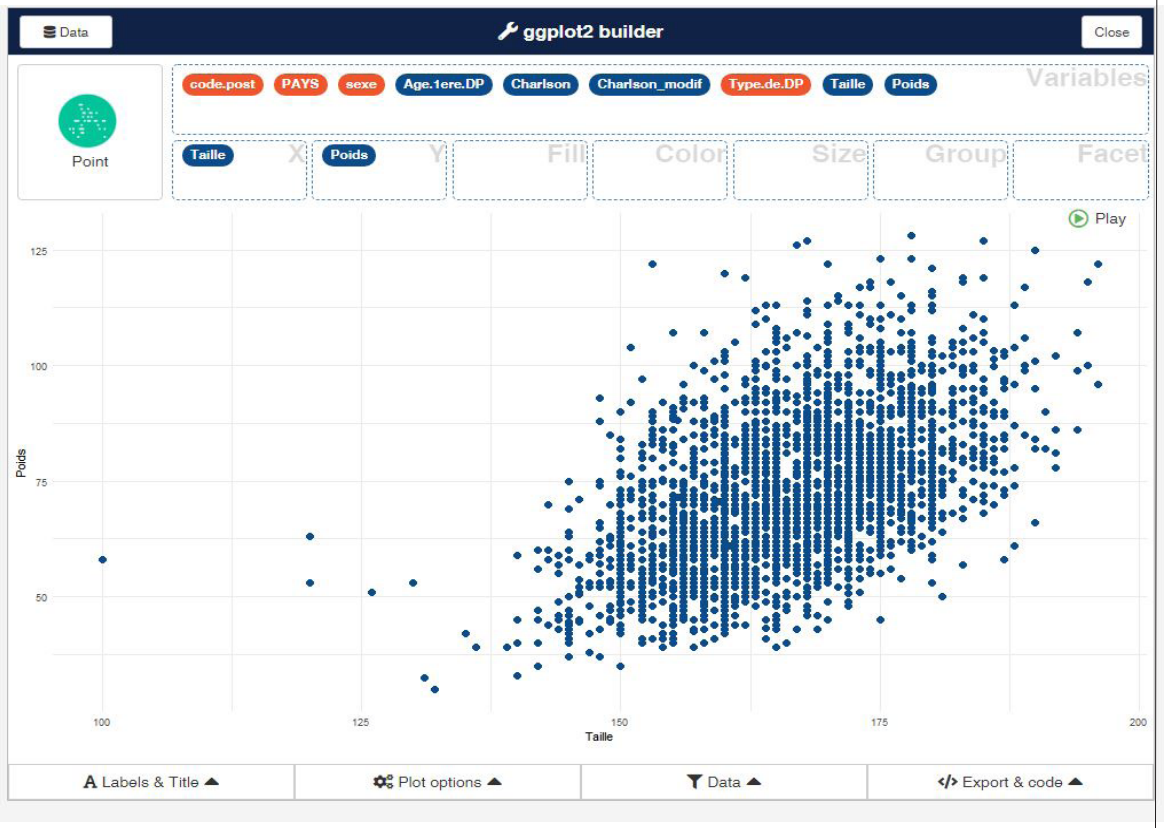
visualisations :.

- «Labels and Title», which allows to manage the titles and the names of the axes,
- «Plot options», which allows, for example, to use a log scale on the axes, or to modify the colors used, or the plot theme,
- «Data», which allows, for example, to select a subgroup of data,
- «Export & Code», in which you will find buttons to export the realized graph, and the corresponding ggplot2 code.

6 Demonstration

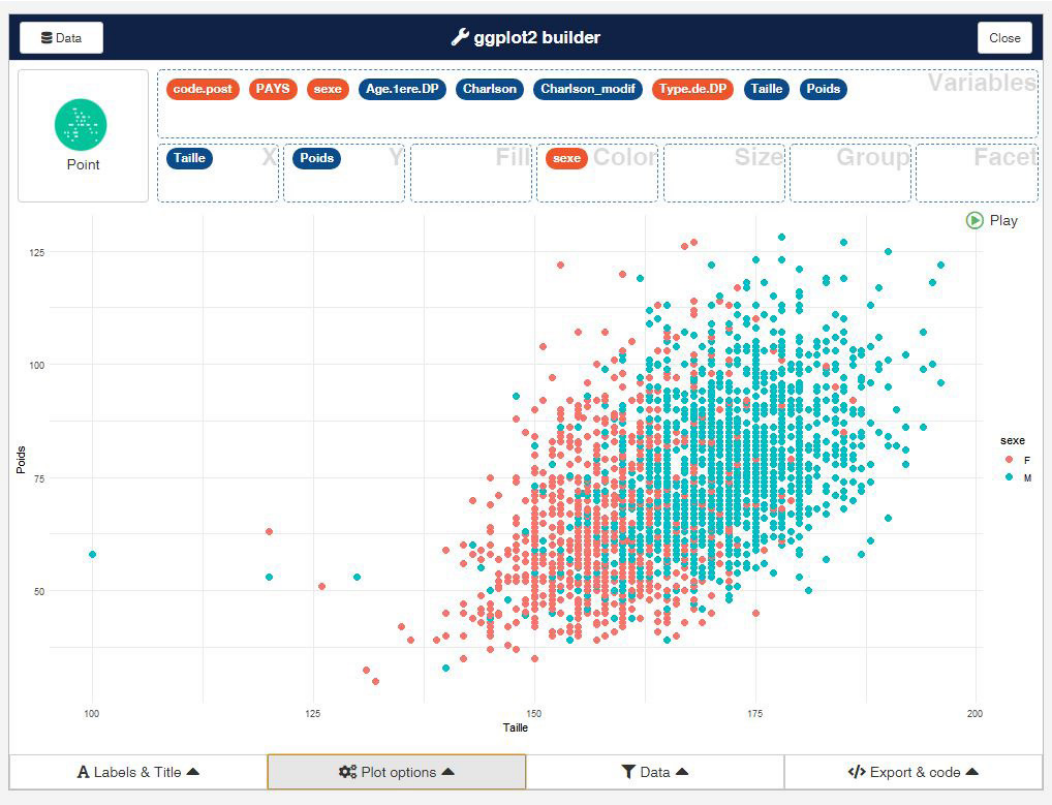
6.1 Exemples with scatter plot

Here is a series of examples based on the realization of a scatter plot of the variables Size (in X) and Weight (in Y):



By adding the Sex tag in the Color box, we can distinguish men from women:

www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917



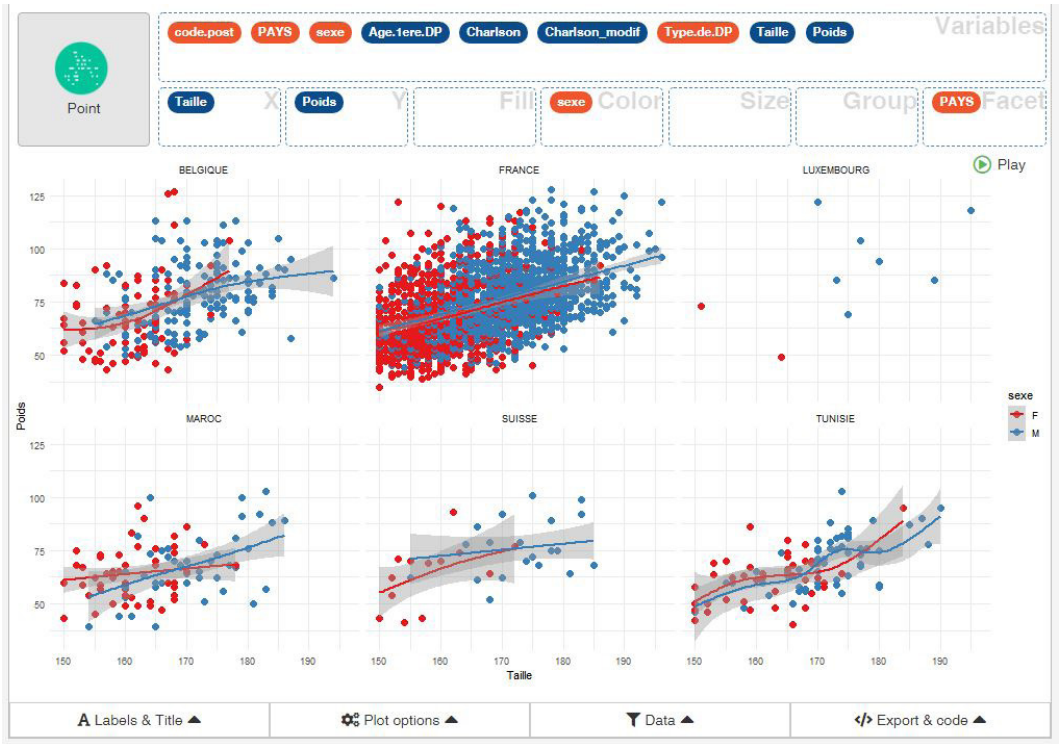
En ajoutant l'étiquette Pays dans la boite "Facet", un graphique par pays est réalisé :

Il est possible d'ajouter une courbe de tendance, en allant dans le menu "Plot Option", en activant l'option

www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917

Smooth.

De même, vous pouvez changer les couleurs des points et limiter les données (ci-dessous, seules les données des patients mesurant au moins 1 mètre 50 sont considérées).

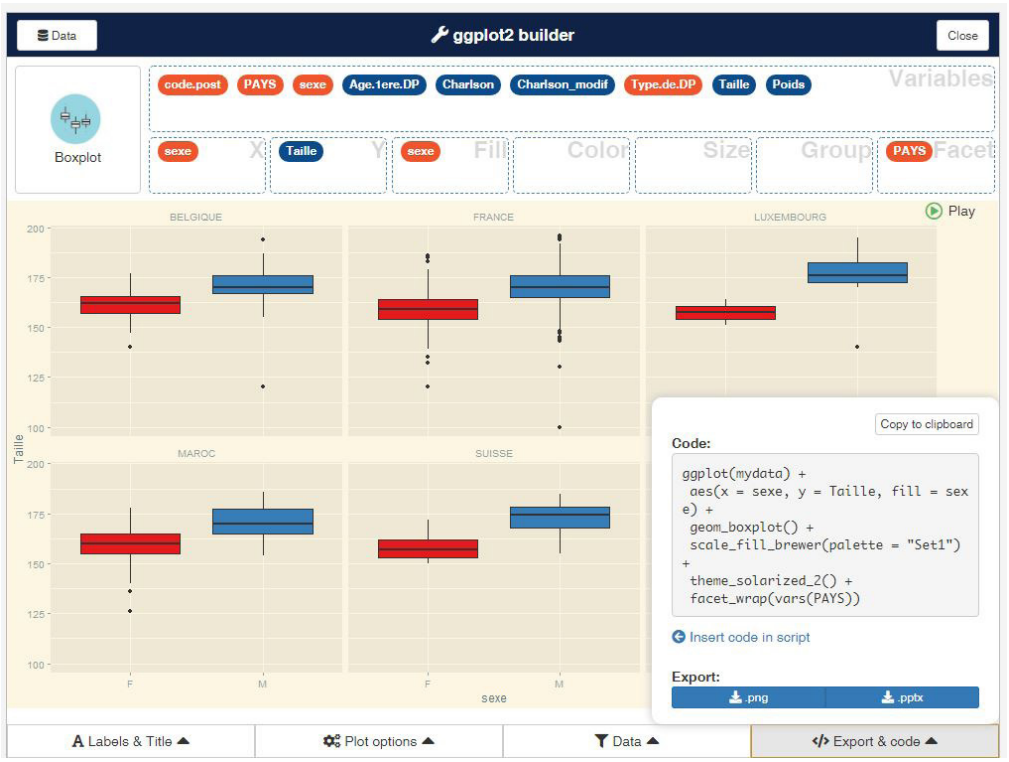


www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917

6.2 Other exemples

Here, are the distribution densities of the weight variable for men and women.

Here, boxplots to visualize the distribution of patient sizes by country, distinguishing between men and women, with a «solarized» theme:



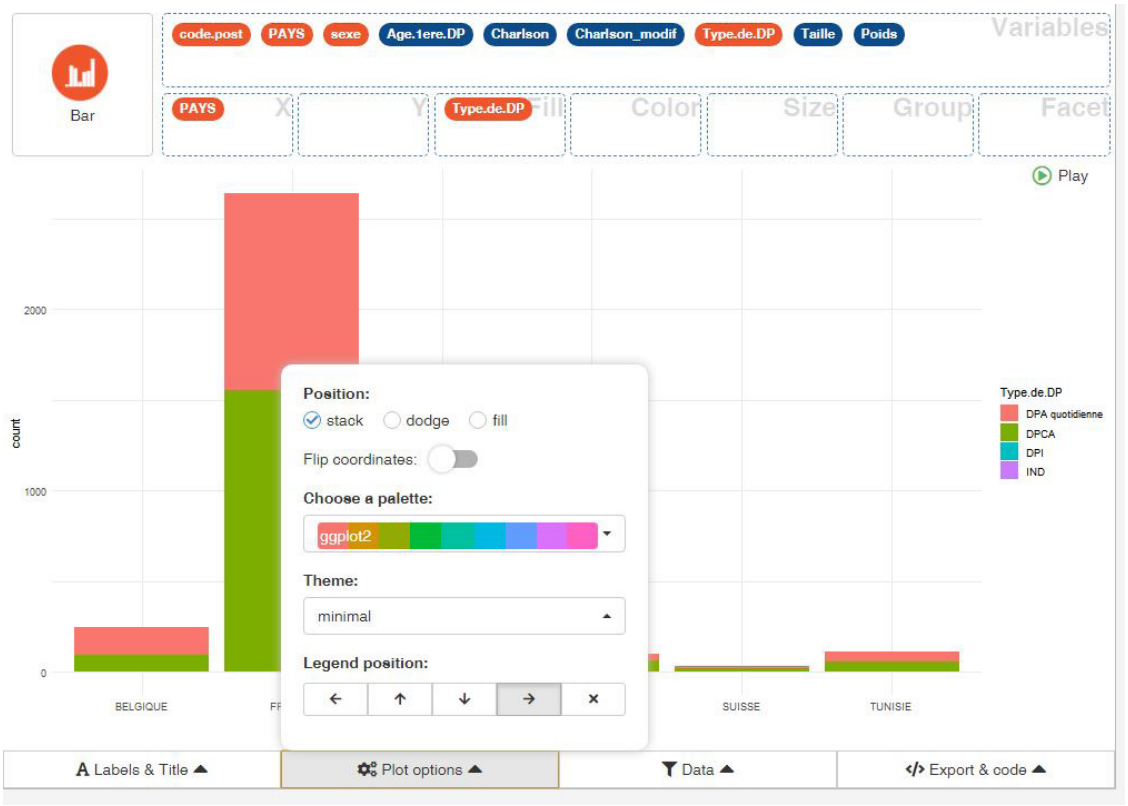
www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917

Note the transcript of the graph made in ggplot2 code, directly usable in the console or in a script. For that, do not forget to load the ggplot2 package with the following command: library(ggplot2)

You will also find other examples of using the Sketch addin here : (https://statistique-et-logiciel-r.com/decouvrez-laddin-esquisse/)

7 To conclude

All you have to do is explore, by yourself, with other data, the capabilities of this tool. I recommend, in particular, to test the different possibilities of barplots, using the options «stack» «dodge» and «fill».



If you need inspiration, I recommend going for a ride on the «R Graph Gallery» (https://www.r-graph-gallery.com/)

And for more information on the use of the argument «Facet», or «Group», or more generally on ggplot2, I recommend the book R graphics cookbook (available online) of Winston Chang (https://r-graphics.org/).

You can also consult this page (http://www.cookbook-r.com/Graphs/) and this introductory article to the ggplot2 package. (https://statistique-et-logiciel-r.com/introduction-a-la-visualisation-sous-r-avec-le-package-ggplot2/)

In the next article, we will learn how to use ggplot2, directly with the command lines (the learning box in sum ;-)).

www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917

DISCLOSURES

the author declares that she has no conflict of interest in this article.

LICENCE AND COPYRIGHT

Received 07/07/19, accepted after revision 09/08/19, published 15/09/19