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 à l'analyse de données avec le logiciel R

Introduction à la data visualisation sous R, avec l'add in Esquisse

Claire Della Vedova

1087 chemin de Sainte Roustagne, 04100 MANOSQUE, France

NDLR : Le RDPLF a pour but principal d'être une aide pour permettre aux équipes de dialyse à domicile d'évaluer leurs pratiques cliniques et également conduire des études à partir d'exports anonymisés des données qu'elles saisissent. A cette fin, depuis juin 2019, un article de formation à l'utilisation du logiciel Libre R est publié trimestriellement à chaque parution du Bulletin de la Dialyse à Domicile. Le but est de permettre à toutes les équipes de réaliser des statistiques de bases et visualiser rapidement leur données.

Le premier article de cette série d'initiation était consacré au téléchargment et à l'installation du logiciel R sur les ordinateurs Macintosh et PC : https://doi.org/10.25796/bdd.v2i2.20513 .

Ce second article est consacré à la visualisation graphiques des données statistiques. Le package de base utilisé est habituellement ggplot2. Mais ce dernier demande une petite phase d'apprentissage dont on ne dispose pas forcement le temps lorsque le besoin de creer un graphique pour une présentation est urgent.

Il peut être intéressant alors de recourir à un logiciel graphique d'apprentissage simple et rapide permettant rapidement de mettre en valeur ses données numériques sous une forme attrayante.

Le package Esquisse répond à ce besoin et sera le sujet de cet article.

Comme dans le premier article un fichier exemple tiré de la base de données du RDPLF sera utlisé.

Claire Della Vedova est Ingénieure en biostatistique / data analyste, Elle utilise quotidiennement le logiciel R pour analyser des données. Elle a travaillé pendant plus de 15 ans dans les domaines de l'environnement et de la santé, et a formé de nombreux étudiants et chercheurs à l'utilisation de R. Elle anime depuis novembre 2017 le blog Statistique et Logiciel R dont le but est d'aider les débutants à mieux appréhender les méthodes statistiques classiques et à utiliser le logiciel R plus efficacement, notamment au travers de tutoriels : https://statistique-et-logiciel-r.com/.

La formation totale se fait sur 15 mois, au rythme d'un article par trimestre à chaque parution du BDD. Cela laissera largement le temps d'assimiler et tester les connaissances acquises entre chaque article. Pour ceux qui souhaiteront aller plus vite, ils pourront aller sur le blog (https://statistique-et-logiciel-r.com/).

Dates des prochaines parutions :

- article 3 (Décembre 2019) : une initiation à ggplot2
- article 4 (Avril 2020) : la réalisation de rapports d'analyses statistiques automatisées avec Rmarkdown
- article 5 (Juin 2020) : la manipulation de données (avec dplyr, notamment les fonction group_by et summarise)
- article 6 (Septembre 2020) : la réalisation d'analyses descriptives (paramètres statistiques et graphs) sous forme de dashboard avec le package flexboard

Mots clés : biostatistique, épidémiologie, logiciel R, 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

Introduction à la data visualisation sous R, avec l'add in Esquisse

1. Introduction :

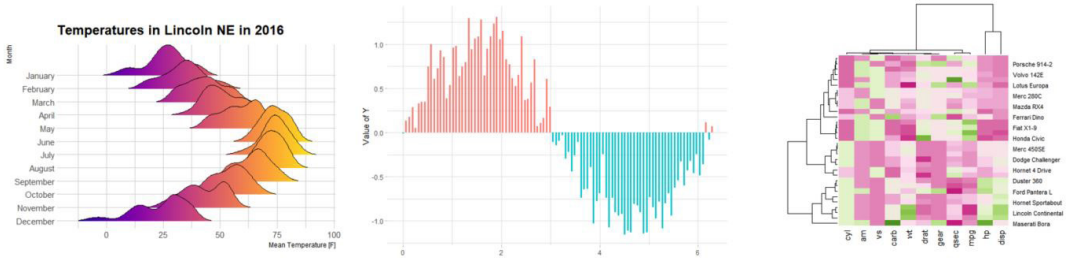
Dans le premier article (*https://www.bdd.rdplf.org/index.php/bdd/article/view/20513*), nous avons vu comment installer le logiciel R et son interface RStudio, comment préparer l'organisation de son travail sous la forme d'un projet R, puis comment importer des données, les vérifier et enfin les valider en utilisant R.

Dans ce second article, nous allons parler de data visualisation, autrement dit, de la réalisation de graphiques avec R !

2. Les graphiques sous R

Sous R, la représentation graphique des données passe, quasiment nécessairement, par l'utilisation du package ggplot2. Ce package permet de réaliser de nombreux types de graphiques, et d'obtenir des rendus très élégants, notamment pour des publications.

En voici quelques exemples :
D'après *https://www.r-graph-gallery.com*



Le package ggplot2 repose sur le principe de couches successives, qui sont appliquées au graphique, pour le construire en pas à pas.

Ainsi, on commence par déclarer un canevas, puis on ajoute une couche pour définir le type de graphique, puis une autre pour ajouter une courbe de tendance par exemple, puis une autre pour gérer la couleur, puis une autre pour gérer les échelles des axes etc…

L'utilisation de ce package n'est pas compliquée en elle-même, mais elle nécessite néanmoins d'investir un peu de temps dans son apprentissage.

Et parfois, c'est un peu frustrant, parce que lorsqu'on s'intéresse à la réalisation de graphiques sous R, c'est généralement parce qu'on en a besoin tout de suite !

Alors dans cet article, je vais vous donner une astuce pour commencer à utiliser ggplot2 immédiatement. Sans passer par la case apprentissage !

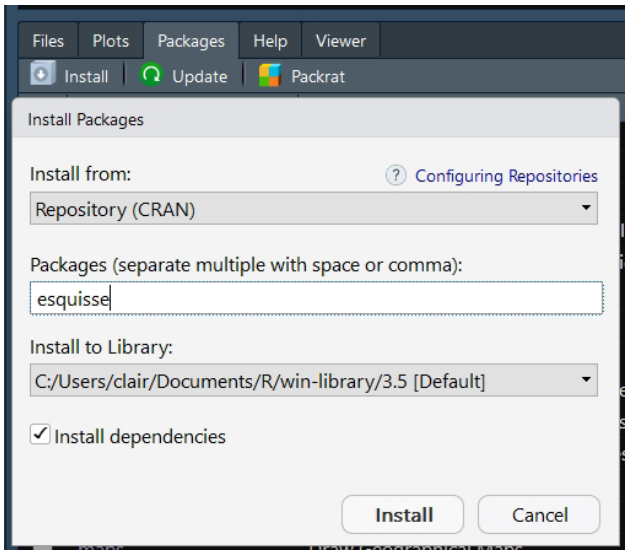
Cette astuce, c'est l'**addin "Esquisse"** ! Il s'agit d'une interface graphique qui utilise de nombreuses fonctionnalités de ggplot2 mais sans code, uniquement en manipulant des étiquettes (en drag and drop).
Et cerise sur le gâteau, Esquisse vous montre le code ggplot2 correspondant au graphique réalisé ! C'est un peu magique ! Nous devons cet outil à Victor Perrier de DreamRs, une entreprise de conseils et d'expertises en data sciences spécialisée en R.

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

3. Installation et accès de l'addin Esquisse

3.1 Installer le package Esquisse

Pour utiliser l'addin Esquisse, il est nécessaire d'installer le package du même nom, par exemple en utilisant l'outil d'installation de R Studio :



Il est également nécessaire d'installer le package ggplot2, de la même façon.

3.2 Importer vos données

Pour utiliser l'addin, il est également nécessaire que le jeu de données sur lequel va porter la, ou les, représentations graphiques, soit importé dans R. A titre d'exemple, nous allons, ici, utiliser les mêmes données que dans le premier article, elles sont téléchargeables au format csv, à cette adresse :

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

Une fois téléchargé, placez le fichier csv dans le dossier "data" de votre projet R , puis utilisez la commande suivante :

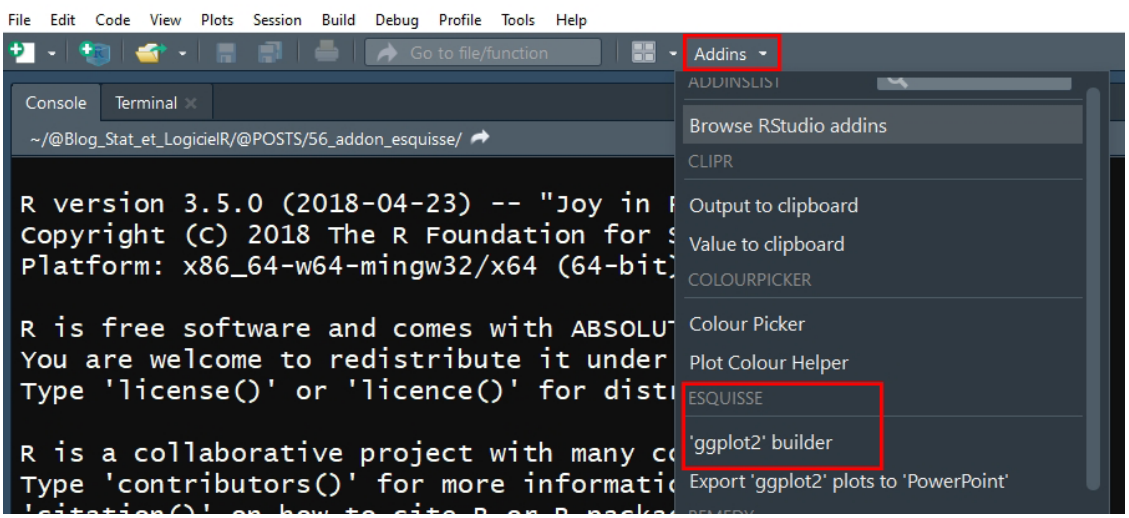
```
mydata <- read.csv2("data/FichierExempleStat.csv")
```

Remarques : pour plus d'informations sur les projets R et son dossier "data, consultez l'article "Introduction à l'analyse de données avec le logiciel R» (https://www.bdd.rdplf.org/index.php/bdd/article/view/20513/19163)).

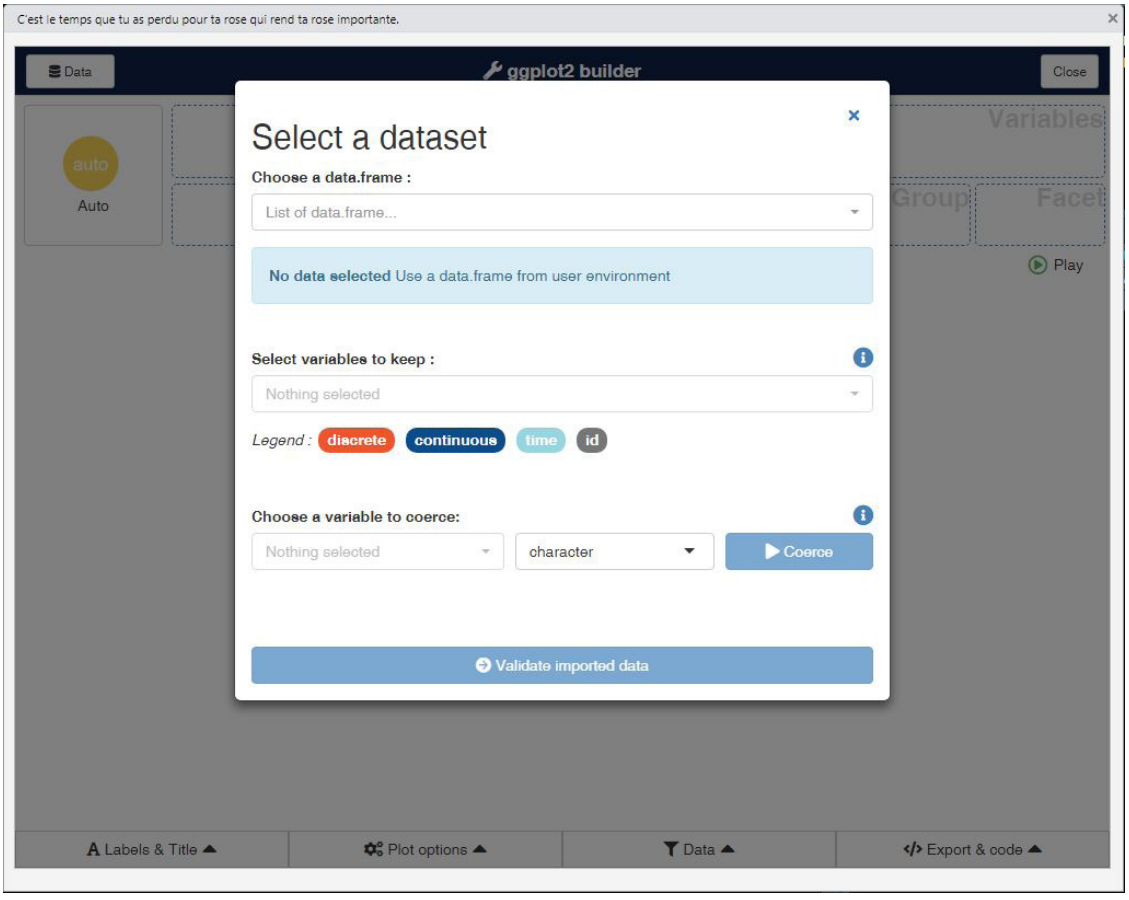
4. Ouverture d'Esquisse

Maintenant que tout est prêt, vous pouvez ouvrir l'interface graphique, en allant dans le menu Addins, puis en choisissant "ggplot2 builder" dans la partie «ESQUISSE :

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



Voici la fenêtre que vous devriez voir apparaître :



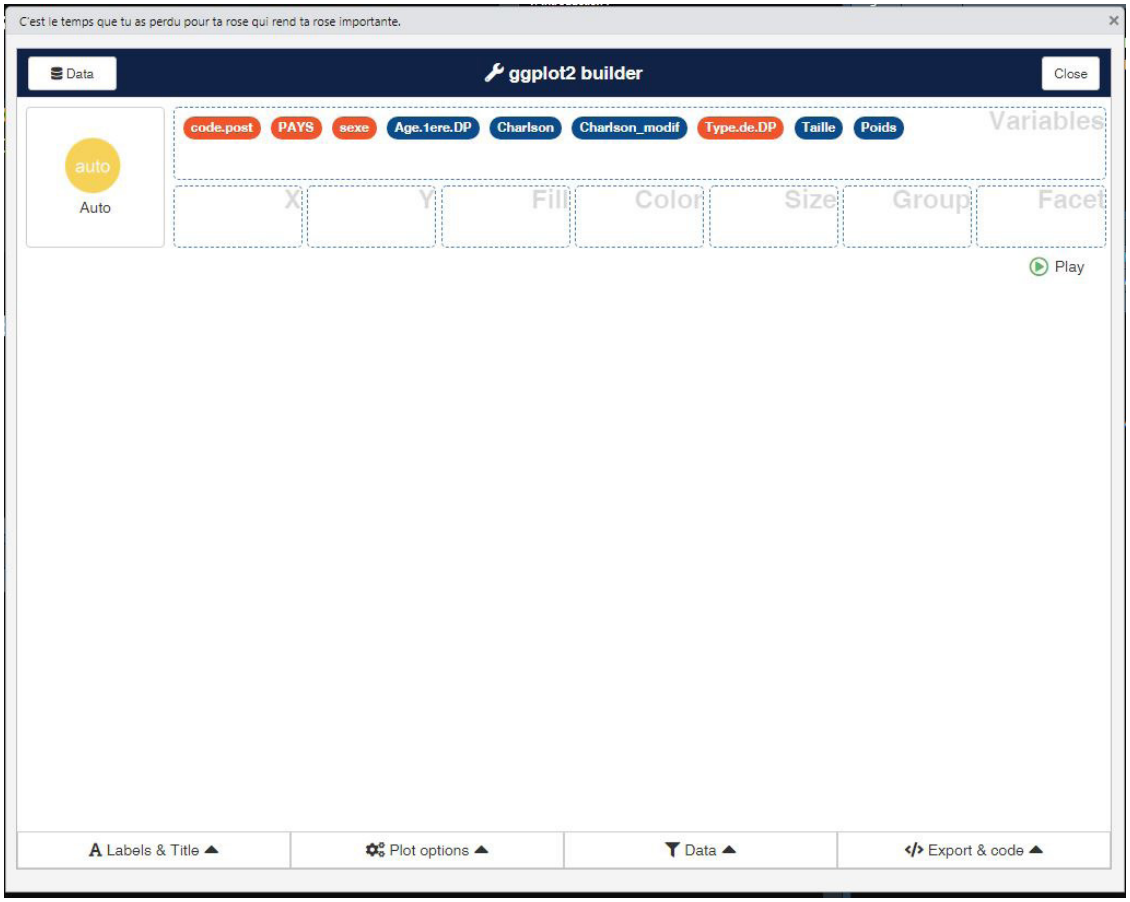
Choisissez votre jeu de données, et cliquez sur “Validate imported data” !

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

5. L'interface

5.1 Les éléments

L'interface apparaît alors, comme ceci:



Les variables du jeu de données sont représentées sous la forme d'étiquettes, dans la partie supérieure. Dans la partie directement en dessous, se trouve des boites, nommées :

- X : sert à définir la variable représentée sur l'axe des X,
- Y : sert à définir la variable représentée sur l'axe des Y,
- Fill : sert à définir la variable qui contrôlera la couleur des graphiques contenant des "boites", comme les boxplots ou les barplots par exemple,
- Color : sert à définir la variable qui contrôlera la couleur des graphiques contenant des points et/ou des lignes etc…
- Size : sert à définir la variable qui contrôlera la taille des points,
- Group : sert à définir une variable qui contrôlera le regroupement des données dans certains graphiques,
- Facet : sert à définir une variable qui contrôlera la division du graphique en multiples entités (par exemple un scatterplot par Pays).

Pas de panique si vous êtes perdu, nous allons voir cela dans des exemples.

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

5.2 Les menus d'option

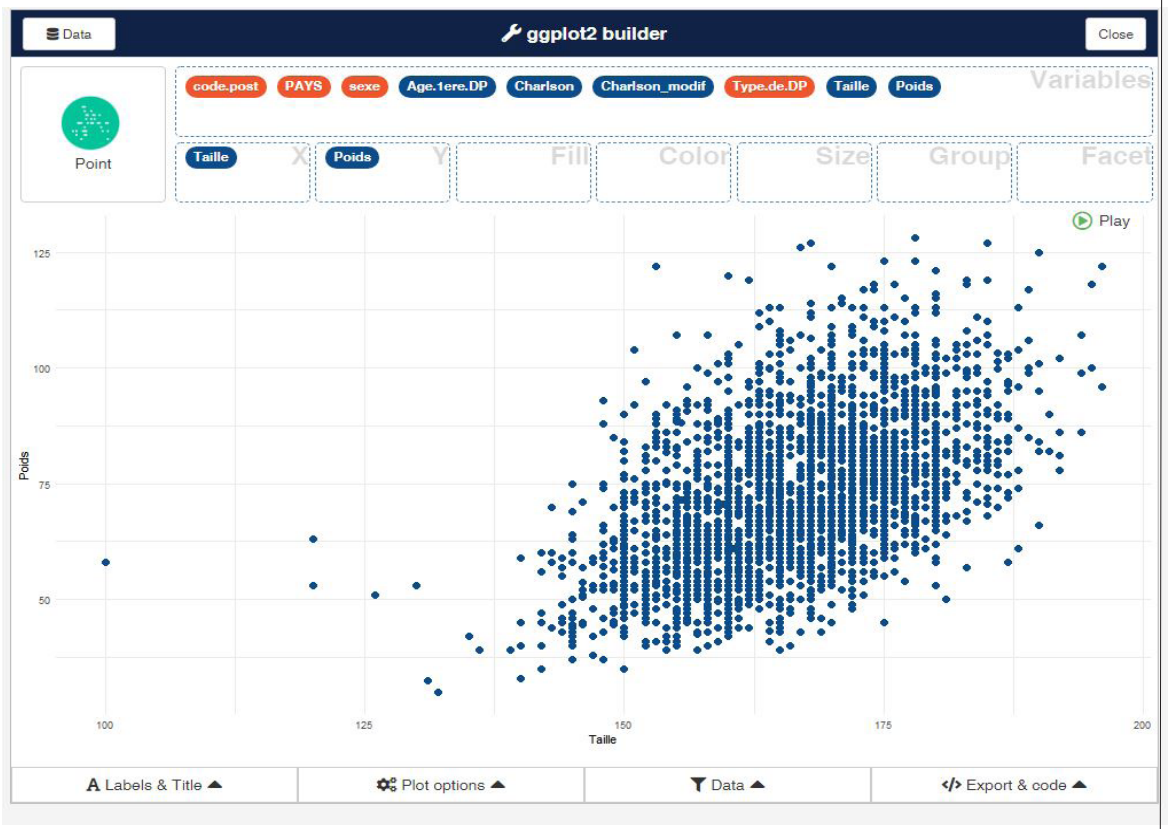
Dans la partie inférieure de l'interface, se trouvent différents menus, qui permettent de personnaliser les visualisations :

- "Labels et Title", qui permet de gérer les titres et les noms des axes,
- "Plot options", qui permet, par exemple, d'employer une échelle log sur les axes, ou encore de modifier les couleurs employées, ou le thème du plot,
- "Data", qui permet, par exemple, de sélectionner un sous-groupe de données,
- "Export & Code", dans lequel vous retrouverez des boutons pour exporter le graphique réalisé, et le code ggplot2 correspondant.

6 Démonstration

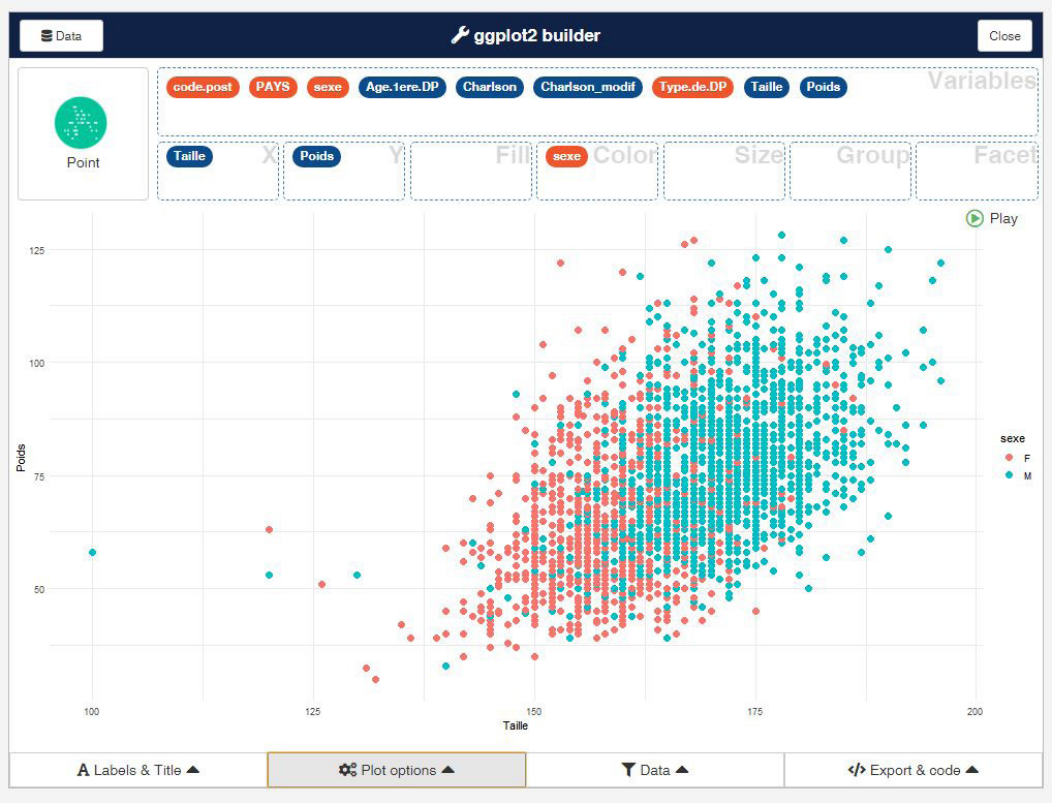
6.1 Exemples autour d'un scatter plot

Voici une série d'exemples basés sur la réalisation d'un scatter plot des variables Taille (en X) et Poids (en Y):



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 Sexe dans la boite Color, nous pouvons distinguer les hommes des femmes :

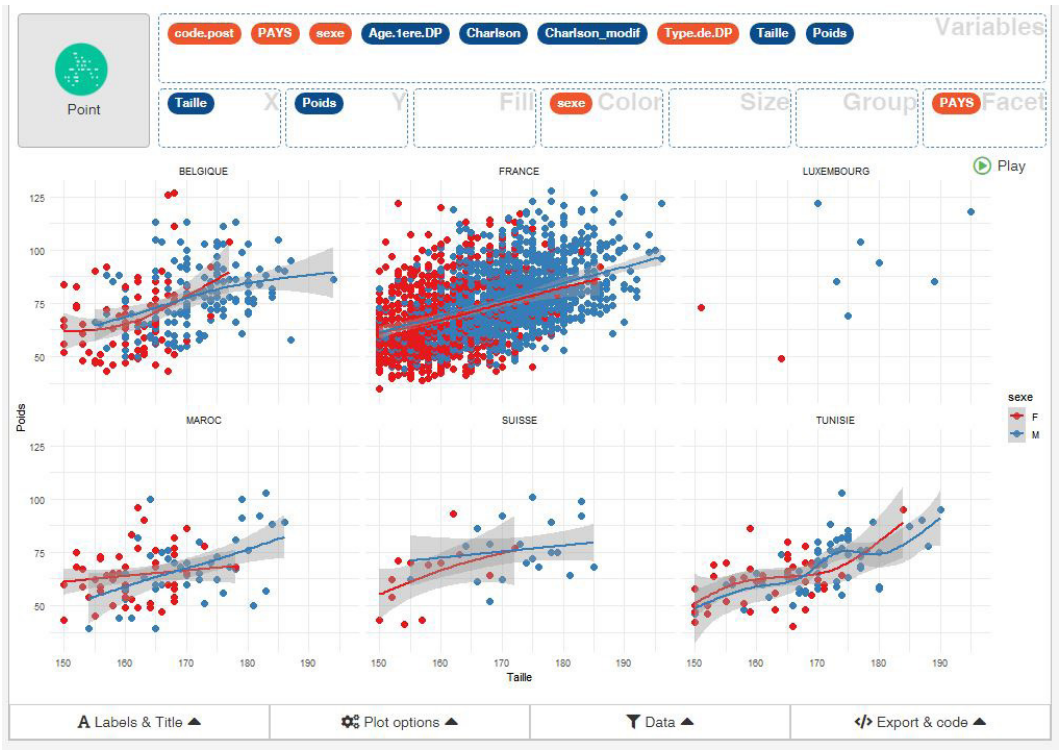


En ajoutant l'étiquette Pays dans la boite "Facet", un graphique par pays est réalisé :

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

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

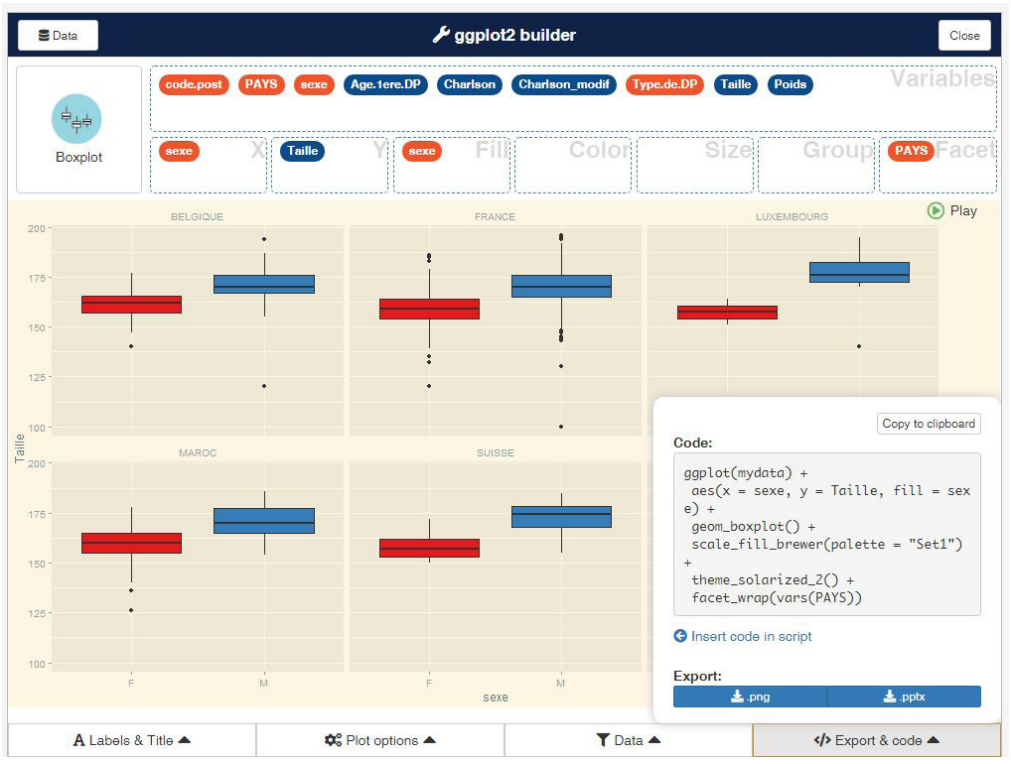


www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917

6.2 Autres exemples

Voici, ici, les densités de distribution de la variable poids, chez les hommes et les femmes.

Ici, des boxplots permettant de visualiser la distribution des tailles des patients par pays, en distinguant les hommes et les femmes, avec un thème «solarized» :



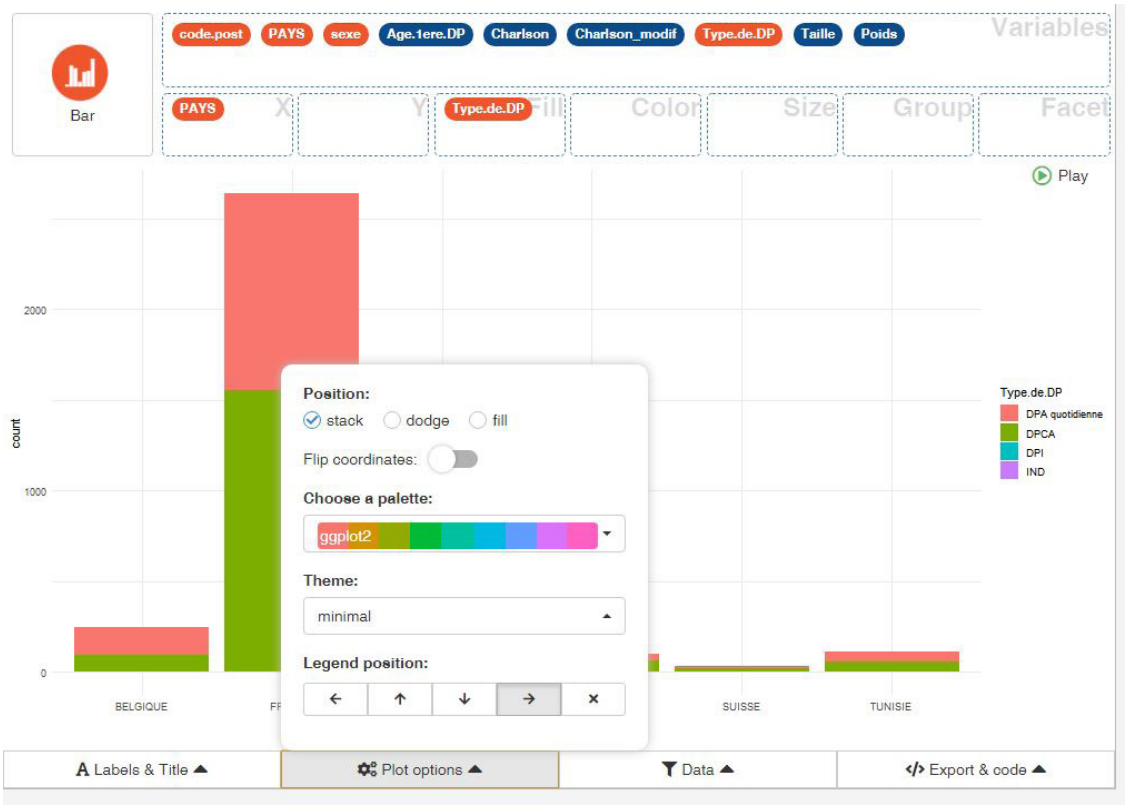
www.bdd.rdplf.org Volume 2, n° 3,Sept 2019
https://doi.org/10.25796/bdd.v2i3.21313
ISSN 2607-9917

Notez la retranscription du graphique réalisé en code ggplot2, directement utilisable dans la console ou dans un script. Pour cela, n'oubliez, au préalable, pas de charger le package ggplot2 avec la commande suivante : library(ggplot2)

Vous trouverez également d'autres exemples d'utilisation de l'addin Esquisse, ici : (https://statistique-et-logiciel-r.com/decouvrez-laddin-esquisse/)

7 Pour conclure

Il ne vous reste plus qu'à explorer, par vous-même, avec d'autres données peut être, les capacités de cet outil. Je vous recommande, en particulier, de tester les différentes possibilités de barplots, en utilisant les options "stack" "dodge" et "fill".



Si vous avez besoin d'inspiration, je vous recommande d'aller faire un tour sur la "R Graph Gallery" (https://www.r-graph-gallery.com/)
Et pour avoir plus d'informations sur l'utilisation de l'argument "Facet", ou "Group", ou plus généralement sur ggplot2, je vous recommande le livre R graphics cookbook (consultable en ligne) de Winston Chang (https://r-graphics.org/).
Vous pouvez également consulter cette page (http://www.cookbook-r.com/Graphs/) et cet article d'introduction au package ggplot2. (https://statistique-et-logiciel-r.com/introduction-a-la-visualisation-sous-r-avec-le-package-ggplot2/)
Dans le prochain article, nous apprendrons à utiliser ggplot2, directement avec les lignes de commandes (la case apprentissage en somme;-)).

CONFLITS D'INTERET

l'auteur déclare ne pas avoir de conflit d'intérêt pour cet article.

LICENCE ET DROIT D'AUTEUR

Reçu le 07/07/19, accepté après révision le 09/08/19, publié le 15/09/19