www.bdd.rdplf.org Volume 2, n° 2, Juin 2019
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917

Le Bulletin de la Dialyse à Domicile

Introduction to data analysis with R software

(Introduction à l'analyse de données avec le logiciel R)

Claire Della Vedova

1087 chemin de Sainte Roustagne, 04100 MANOSQUE, France

Note : le texte original en version Française est disponible à la même adresse url : https://doi.org/10.25796/bdd.v2i2.20513

Editor's note: The main purpose of the RDPLF is to help nurses etn nephrologists involved in home dialysis treatment to evaluate their clinical practices and also to conduct studies based the anonymous content of the Registry data base. Any evaluation or study based on biological or clinical data requires the use of statistical calculations. If the complex models are the domain of specialists, many basic calculations, such as average, median, etc .. are available to all with a minimum effort. There is a free software, «R», extremely powerful, that is easy to install on any computer and that allows the simplest calculations as the most complicated one according to the user's skills.We thought that BDD readers, nurses and doctors would be interested in taking basic training that would give them autonomy for simple or complicated studies that they would like to lead. The RDPLF is also at their disposal to send them any anonymized export file.We start with this issue a series of articles written by Claire Della Vedova that we thank for the help provided.Claire is an engineer in biostatistics / data analyst, she uses daily R software 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 to the use of R. Since November 2017 she animates the blog Statistics and R Software whose goal is to 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/.This first article is reserved for the installation of the free software R.

This course is initially writtent in French, but we thought that some english readers might be interested and we translated it for them. We apologize for any english non conventional writing.

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/).

Next issues :
- article 2 (september 2019) : la réalisation de représentations visuelles avec l'add on esquisse
- 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

www.bdd.rdplf.org Volume 2, n° 2, Juin 2019
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917

1. What is R ?

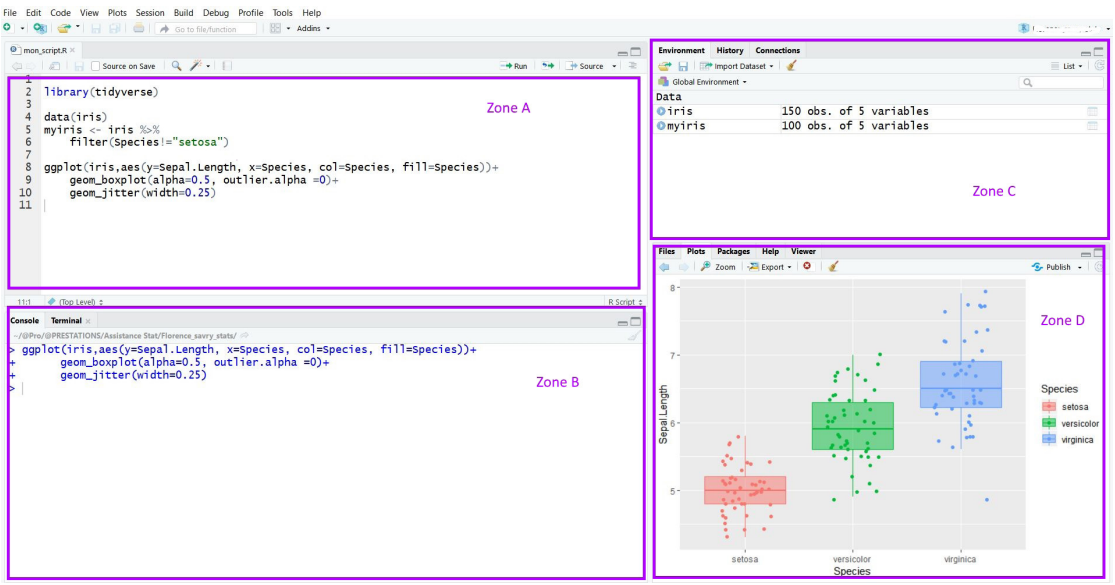
Originally, R was a statistical software. Today, given its evolution, it is more qualified as a data science software because it can also be used to make data visualization, machine learning, cartography, automated analysis reports, dashboards, or web applications (with shiny).

R is very widely used in the medical field, because it allows performing descriptive analyzes, to use multiple hypothesis tests (to compare averages or percentages for example), or to use many regression models (multiple linear regression, logistic regression, survival models etc ...).

But R is also a programming language! In practice it means that it works, not with buttons, or drop-down menus (like Excel for example), but mainly with command lines. These command lines have kinds of «keywords» that are implemented (ie, ready to use) functions. For example, the mean () function calculates an average. It is, of course, also possible to code its own functions.
This use of the command lines can be a bit scary at first. But on the one hand, there is an environment for R, called R Studio, that greatly facilitates its use. And secondly, my experience in coaching beginners, shows me that by investing a little time in language learning, it is quite possible to become autonomous quite quickly.
Finally, R is a free software, which is available for Windows, MacOS and Linux operating systems.

1.2. How does it work ?

To use R, you first need to download and install R, then do the same with R Studio (this is explained in detail below). Then, R is used through R Studio. This is a graphical interface that consists of four windows (or zones):
Zone A is the zone dedicated to the edition of R codes. It is at its level that the opening, the creation or the modification of scripts of R commands takes place. These scripts are in .R or .Rmd.
Zone B is the R software console; it allows the execution of codes. The command lines can be directly entered into the console, or transferred from zone A to zone B by a copy-and-paste, or by the shortcut Ctrl + Enter after being positioned on the line.
The zone C makes it possible in particular to have access to the objects present in the memory of R, as well as the datasets imported or created. This area also contains a data import tool via the Import Dataset drop-down menu.
Zone D allows, among other things, to have access to the tool for downloading packages, a window for viewing graphics, a file browser (such as under Windows), or to access the page of help functions.



www.bdd.rdplf.org Volume 2, n° 2, Juin 2019
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917

The functions used by R are grouped into packages, called packages. These packages are developed individually by specialists in the field they are interested in. They are made available to the R community, usually on the CRAN site (in the package tab of the left menu), but sometimes also on the GitHub account of the developer. Packages containing the basic functions are downloaded and installed automatically with the R software (this is the case of the packages stats, graphics, grDevices, datasets, methods, base). Other packages must be downloaded and installed voluntarily. We will install in the rest of this article, the funModelling package to use some of the functions it contains to perform a descriptive analysis of qualitative variables (or categorical).

1.3. Why use R rather than Excel to analyze data?

First, because Excel is not particularly intuitive to do statistical analysis, whether descriptive or comparative. Second, because the statistical analyzes that can be done in Excel are extremely limited. The same is true of graphic possibilities.
On the side of R, the possibilities are almost limitless! Not only in terms of statistical analysis, but also in terms of data visualization, reporting (dashboard, automated analysis reports), and open access resources.

And then R is very much used in the medical field; knowing how to use it has become an important and sought-after skill.

And finally, as said before R is free!

2. R et RStudio installation

The installation of R & R studio takes place in 5 steps:
1) R Software Download
2) Installing the R software
3) Downloading from R studio
4) Installing R studio
5) Opening of R studio

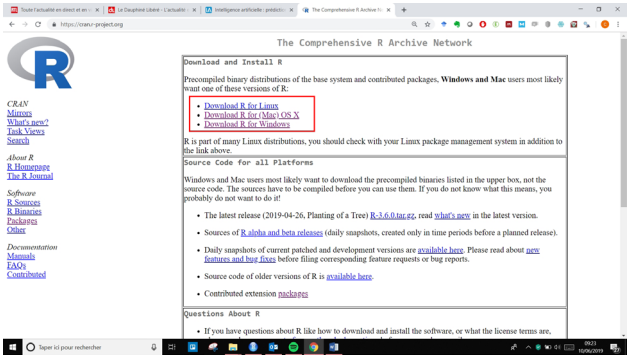
2.1 R software download:

R is available on the CRAN website CRAN : The Comprehensive R Archive Network (https://cran.r-project.org/).To download it, just go to this site, and follow the steps presented here step by step for Windows. The procedure for other operating systems is substantially the same:

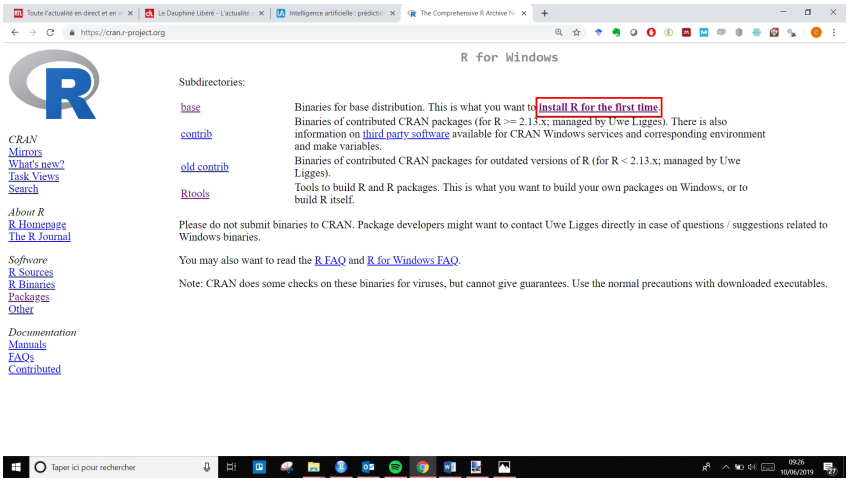
2.2 Installation du logiciel R

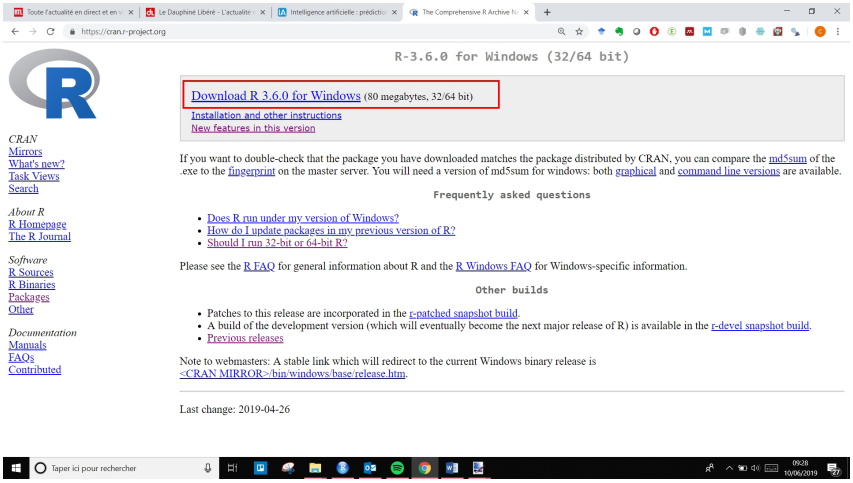
Once the file is downloaded, it is necessary to go to the download folder, and double-click on it to execute it.

Normally, the language selected by default is yours. If this is not the case, you can choose «French» or «English» from the pop-up menu.
For the continuation, it is enough to always click on «following», accepting the options by default, until the end of the installation.



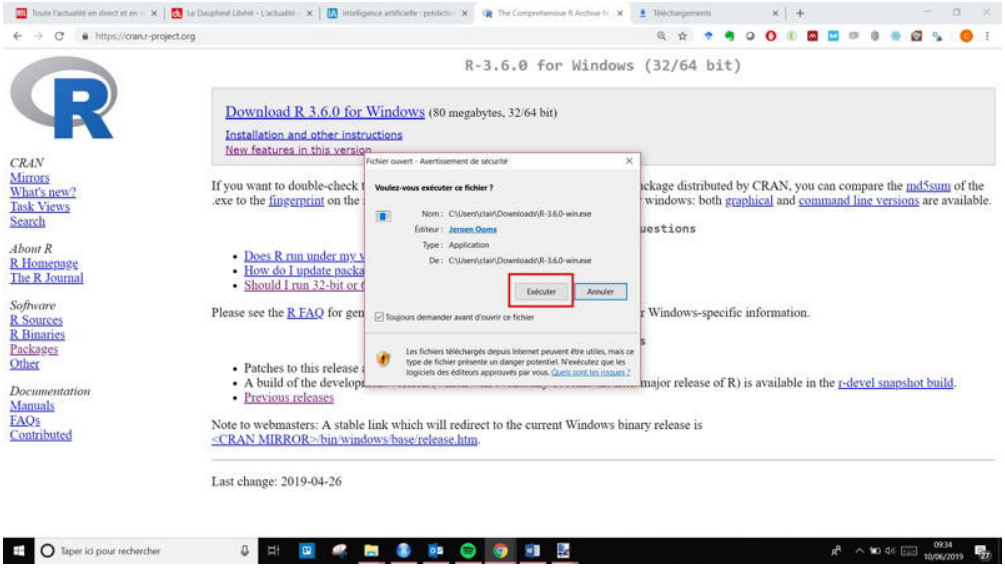
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917



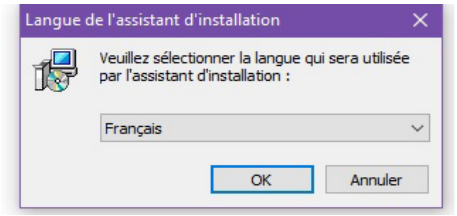


2.3 R Studio download

R studio downloads from the R-studio website (https://www.rstudio.com/).
Again, the approach is presented step by step, for Windows.



www.bdd.rdplf.org Volume 2, n° 2, Juin 2019
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917

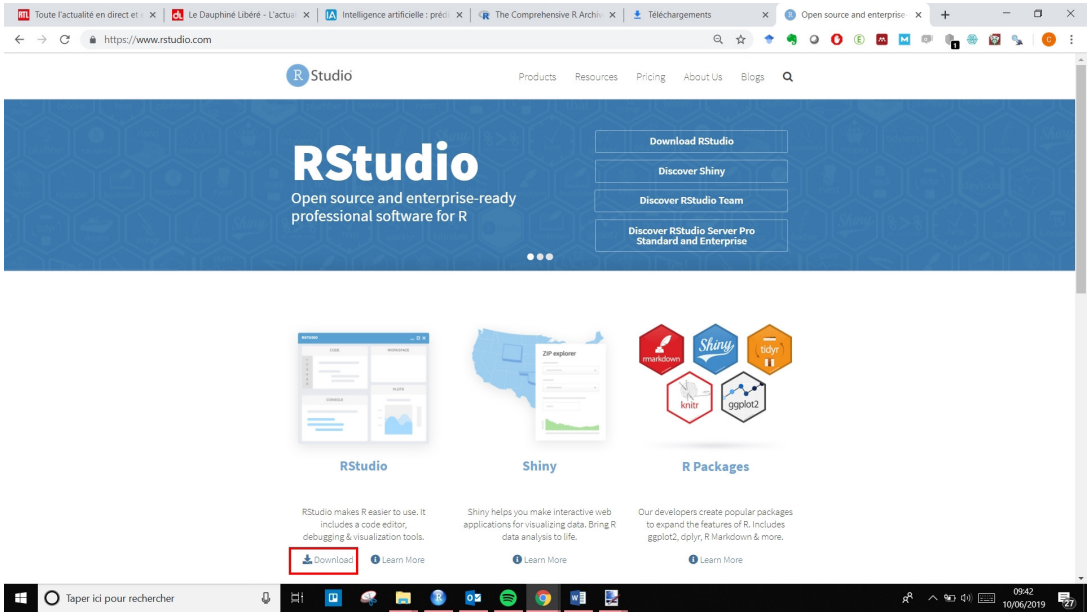


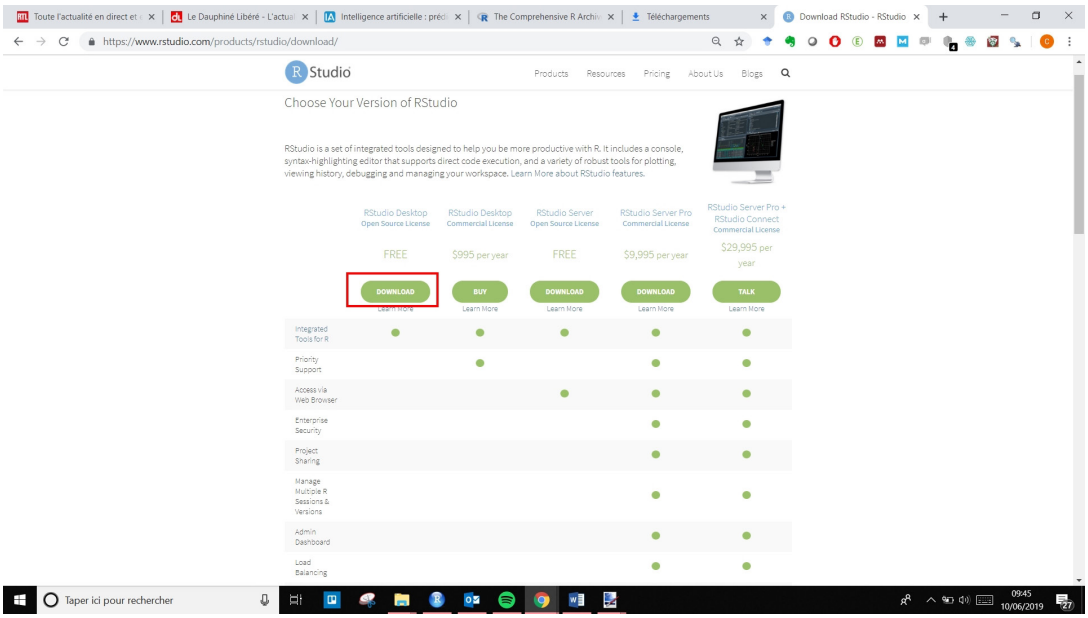
2.4 R-Studio installlation

A ce stade, il faut aller chercher, dans le dossier de Téléchargement, le fichier de R studio que l'on vient de télécharger, puis de double cliquer dessus pour commencer son installation.

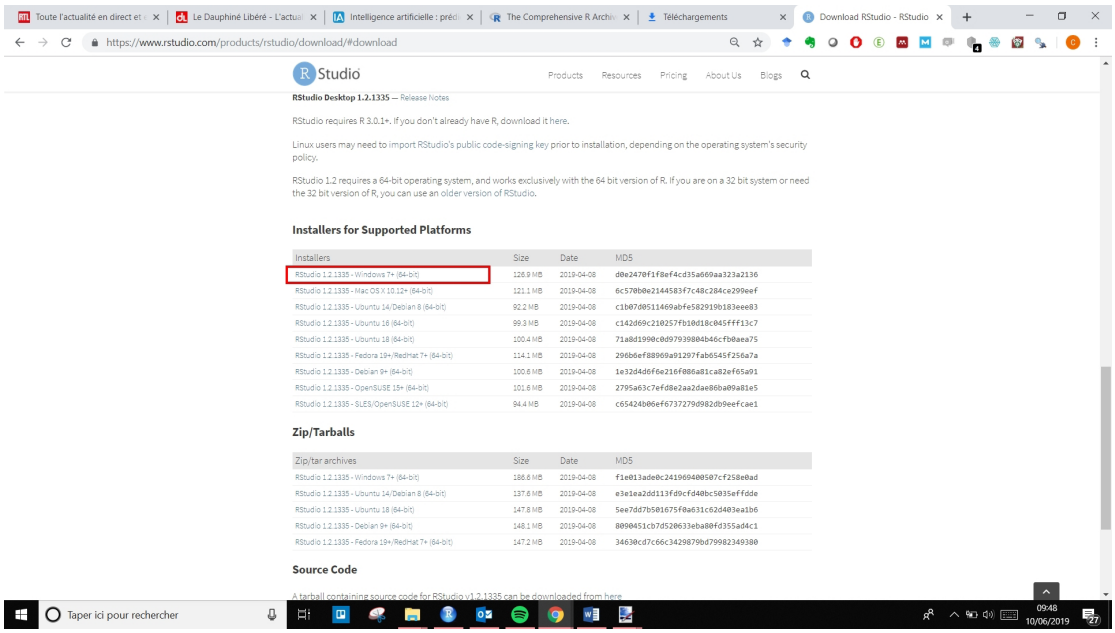
2.5 R and RStudio opening

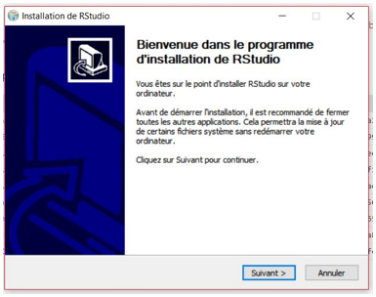
To use R, from R studio, you just have to open R studio. If the installation was successful, an R studio icon is now visible in the Windows drop-down menu, or in the list of installed applications. Open R studio by clicking on it.





www.bdd.rdplf.org Volume 2, n° 2, Juin 2019
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917





You should get a screen similar to this one:

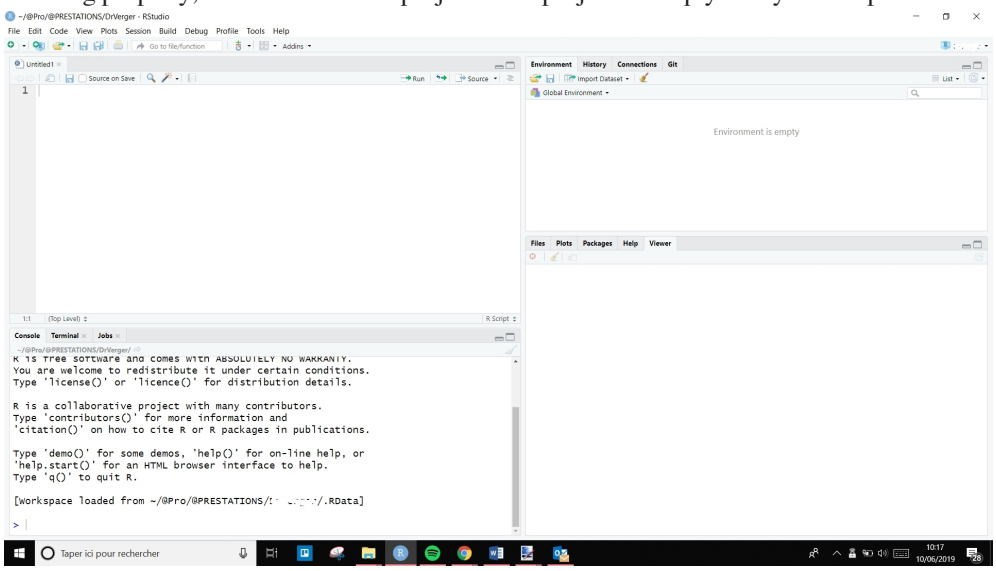
Do not panic if you do not have four windows, just click on the icon at the top right (of the window) to open it. Do not worry either if you miss the Git tab in area C (top left), it's normal!

Rest assured, everything is ready to use R and RStudio.

3.First steps with R :

As a first step, we will import a dataset and make a first descriptive analysis of it.

But before starting properly, we will create a R-project. A R -project is simply a way to encapsulate his work.



www.bdd.rdplf.org Volume 2, n° 2, Juin 2019
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917

It consists of associating a working folder of the computer with R Studio. This is very convenient, because by default, the working directory of the R software will be at the root of this folder. So, every time you save a script or a graph, by default it will be done in this folder.

3.1 Creating a R-project

- 3.1.1 Creating a working directory

To do this, create a working folder wherever you want on your computer.
For example, here I created (with Windows Explorer) a folder «Analyzes_stats» in «Document». And in this folder «Analyzes_stats», I created a folder «Projet1». This is the folder that I will associate with R.

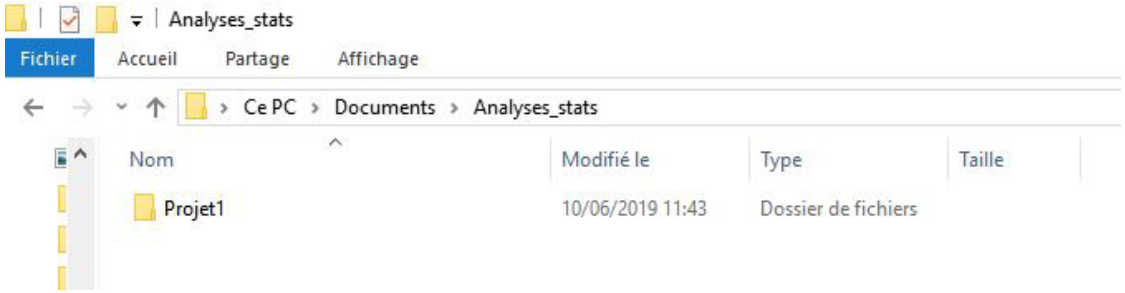
- 3.1.2 Association du dossier à R Studio

Pour cela, dans R Studio, allez dans le menu File (en haut à gauche), puis choisissez “New Project”,et indiquez votre dossier de travail :

You should get a screen identical to the one below :

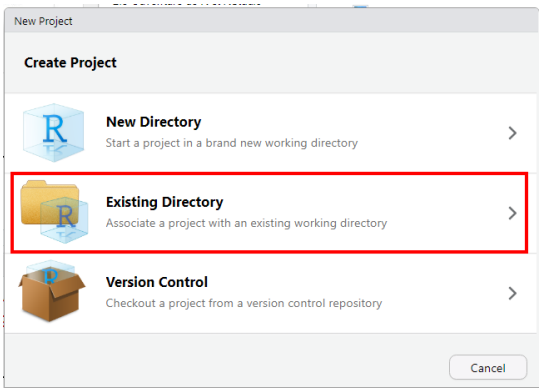
- 3.1.3 Creating a data folder

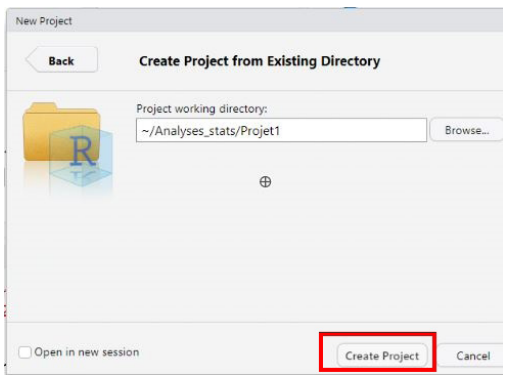
The last step is to create a folder «data», inside folder «Project1». As its name suggests, it is in this folder «data», that will be placed the data sets to import into R to analyze.
You can create this «data» folder in the «project1» folder associated with R Studio, using Windows Explo-



rer (in the same way that you created the Project1 folder), or by using the explorer. RStudio, in zone D.3.2 Quelques éléments à connaître pour bien débuter

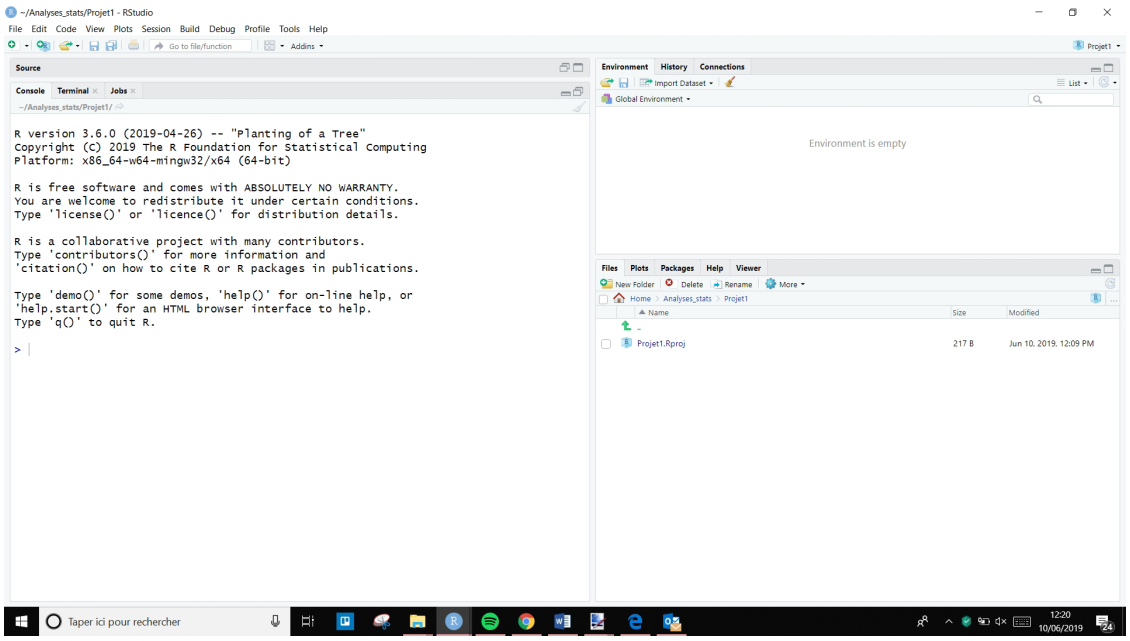
Before starting to speak properly, some elements of the use of R are important to know:





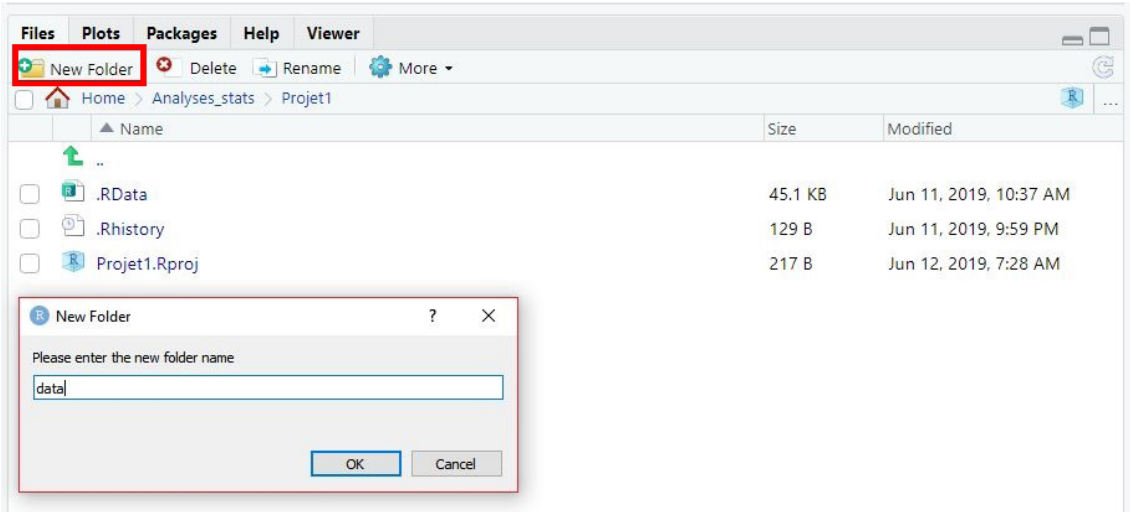
www.bdd.rdplf.org Volume 2, n° 2, Juin 2019
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917

1. A command can be entered in the R console when the prompt, ie the «>» sign is present. When this sign is absent, it means that a command is not completed. To find the prompt, the Esc key can be used.



2. R is case sensitive. This means that a capital letter is not equivalent to a lowercase letter. We must be attentive on this point, when writing code. In the example below when the Mean function is written with a capital letter, R returns an error because mean must be written in lowercase.
3. It is possible to reuse a previously executed R command using the up arrow on the keyboard.

4. Under R, decimals are points.



5. The assignment arrow is used to create an object. Here, the object «a» takes the value 5.
a <- 5

6. For help on a function, we use the syntax? Function, for example with the mean () function that calculates the mean. The help page opens in area D (bottom right) ?mean()
7. Pour connaître les fonctions disponibles, on utilise la fonction ***apropos()***. Par exemple pour connaître toutes les fonctions disponibles qui contiennent le mot “mean”.

https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917


```
apropos(«mean»)
##  [1] «.colMeans»      «.rowMeans»      «colMeans»       «cummean»
##  [5] «kmeans»         «mean»           «mean.Date»      «mean.default»
##  [9] «mean.difftime»  «mean.POSIXct»   «mean.POSIXlt»   «mean_cl_boot»
```


```
## [13] «mean_cl_normal» «mean_sdl»      «mean_se»       «rowMeans»
## [17] «weighted.mean»
```

3.3 Data importation

The data to be imported is contained in a csv file, named «SampleStateFile» extracted from the RDPLF database. It can be downloaded here: https://www.rdplf.org/exempleR/FichierExempleStat.csv

Put them in the «data» folder created previously.
To import the csv data file into R, we will use the read.csv2 () function here, and name the imported file «mydata». In practice, the result of the read.csv2 () function is assigned to mydata, which will contain the data.
mydata <- **read.csv2**(«**data/FichierExempleStat.csv**»)

By default, the *read.csv2()* function considers:

• that the dataset contains variable names (this is called a header),
• that the data separator is a semicolon (this is what is used in a European csv file),
• that the decimal point separator is a comma (this is also what is used by default in a European csv file).

If the file to import does not contain these characteristics, then it will be necessary to specify it in the function. For more information, you can read this article:

https://statistique-et-logiciel-r.com/nettoyer-et-valider-les-donnees-avec-r/

Once the import is done, it is important to check that everything went smoothly. The head () function can then be used to display the first six lines of the dataset:head(mydata)

	A	B	C	D	E	F	G	H	I
1	code post	PAYS	sexe	Age 1ere DP	Charlson	Charlson_mo	Type de DP	Taille	Poids
2	56100	FRANCE	F	20,13	3	3	DPCA	132	30
3	1006	TUNISIE	F	50,74	3	2	DPCA	131	32,5
4	2540	LUXEMBOUR	M	18,86	5	5	DPA quotidie	140	33
5	3200	FRANCE	F	50,86	6	5	DPA quotidie	142	35
6	31603	FRANCE	F	54,07	6	5	DPCA	150	35
7	74374	FRANCE	F	53,89	4	3	DPCA	145	37
8	20420	MAROC	F	22,25	2	2	DPCA	148	37
9	25000	FRANCE	F	86,5	7	3	DPCA	147	38
10	10000	MAROC	F	18,22	2	2	DPA quotidie	136	39
11	83100	FRANCE	F	83,75	7	3	DPCA	139	39

www.bdd.rdplf.org Volume 2, n° 2, Juin 2019
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917

Here, everything seems consistent!
RStudio also has an integrated table that allows you to view, but also to sort or filter the data (as in Excel). To open this spreadsheet, just click on the data in the Environment tab of area C:
The table opens in zone A:

4 Verification and description of the data

Now that the data is correctly imported, it is necessary to check it. In practice this means checking that all the rows and columns are present, that the data has the correct format (for example a quantitative variable is well regarded as a numerical variable, and not as text), and again to look for the presence of possible outliers.

4.1 Contrôle de la structure avec la fonction **str()**

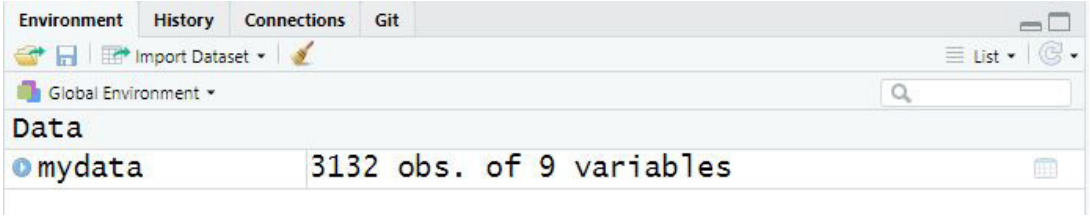
str (mydata)
he function «str» makes it possible to check:
• the data are saved in a data table (data.frame en R)
• the numbers of lines (here 3132) and columns (here 9)
• the nature of the variables:
- «Factor» for a categorical variable,
- «num» or «int» for numeric variabless.

```
Head (mydata)

##   code.post       PAYS sexe Age.1ere.DP Charlson Charlson_modif
## 1     56100     FRANCE    F       20.13        3              3
## 2      1006    TUNISIE    F       50.74        3              2
## 3      2540 LUXEMBOURG    M       18.86        5              5
## 4      3200     FRANCE    F       50.86        6              5
## 5     31603     FRANCE    F       54.07        6              5
## 6     74374     FRANCE    F       53.89        4              3
##         Type.de.DP Taille Poids
## 1             DPCA    132  30.0
## 2             DPCA    131  32.5
## 3 DPA quotidienne    140  33.0
## 4 DPA quotidienne    142  35.0
## 5             DPCA    150  35.0
## 6             DPCA    145  37.0
```

4.2 Description of the quantitative data

The summary () function is very useful for describing data and highlighting any outliers. Indeed, this function returns, for the numeric variables, the smallest value (min), the highest value (max), the mean (mean),



www.bdd.rdplf.org Volume 2, n° 2, Juin 2019
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917

mydata

Filter

	code.post	PAYS	sexe	Age.1ere.DP	Charlson	Charlson_modif	Type.de.DP	Taille	Poids
1	56100	FRANCE	F	20.13	3	3	DPCA	132	30.0
2	1006	TUNISIE	F	50.74	3	2	DPCA	131	32.5
3	2540	LUXEMBOURG	M	18.86	5	5	DPA quotidienne	140	33.0
4	3200	FRANCE	F	50.86	6	5	DPA quotidienne	142	35.0
5	31603	FRANCE	F	54.07	6	5	DPCA	150	35.0
6	74374	FRANCE	F	53.89	4	3	DPCA	145	37.0
7	20420	MAROC	F	22.25	2	2	DPCA	148	37.0
8	25000	FRANCE	F	86.50	7	3	DPCA	147	38.0
9	10000	MAROC	F	18.22	2	2	DPA quotidienne	136	39.0

the median (med), as well as the first and the third quartile.

For more information on descriptive analyzes of quantitative variables, you can read this article: http://bit.ly/2XPUGC9 ainsi que celui-là : http://bit.ly/2wTJQPC

summary(mydata)

4.3 Description of qualitative data

The funModeling package offers particularly interesting functions for describing (and even visualizing) categorical (or qualitative) variables, here the variables «code.post», «PAYS», «sexe», and «Type.DP»
We will start by downloading and installing this package. To download it, we can use the tool present in the

```
> str(mydata)
'data.frame':   3132 obs. of  9 variables:
 $ code.post     : Factor w/ 177 levels "10000","1006",..: 84 2 35 49 48 131 28 34 1 151 ...
 $ PAYS          : Factor w/ 6 levels "BELGIQUE","FRANCE",..: 2 6 3 2 2 2 4 2 4 2 ...
 $ sexe          : Factor w/ 2 levels "F","M": 1 1 2 1 1 1 1 1 1 1 ...
 $ Age.1ere.DP   : num  20.1 50.7 18.9 50.9 54.1 ...
 $ Charlson      : int  3 3 5 6 6 4 2 7 2 7 ...
 $ Charlson_modif: int  3 2 5 5 5 3 2 3 2 3 ...
 $ Type.de.DP    : Factor w/ 4 levels "DPA quotidienne",..: 2 2 1 1 2 2 2 2 1 2 ...
 $ Taille        : num  132 131 140 142 150 145 148 147 136 139 ...
 $ Poids         : num  30 32.5 33 35 35 37 37 38 39 39 ...
```

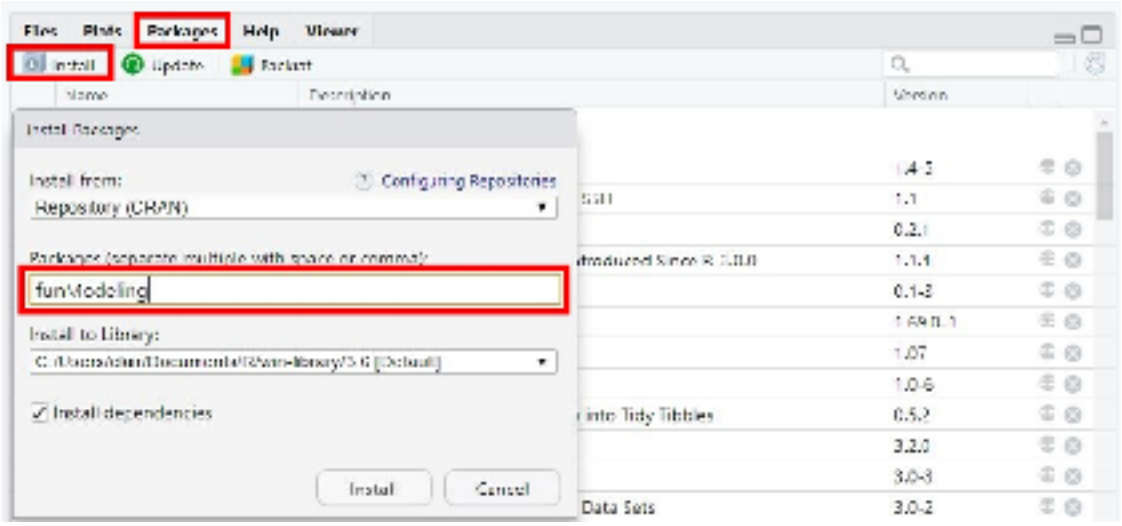
D zone of R Studio:

To be able to use the functions present in this package, it is necessary, beforehand, to load it into R, using the following command:library(funModeling)
Then we use the freq () function to obtain, for each modality (or level of a given variable):
• the number of data
• the percentage represented
• the cumulative percentage
In the command line below, the syntax [«Type.DP»] allows you to specify the variable you wish to study.
freq(mydata[«Type.de.DP»])

The *freq()* function also automatically creates a graphical representation of the studied da

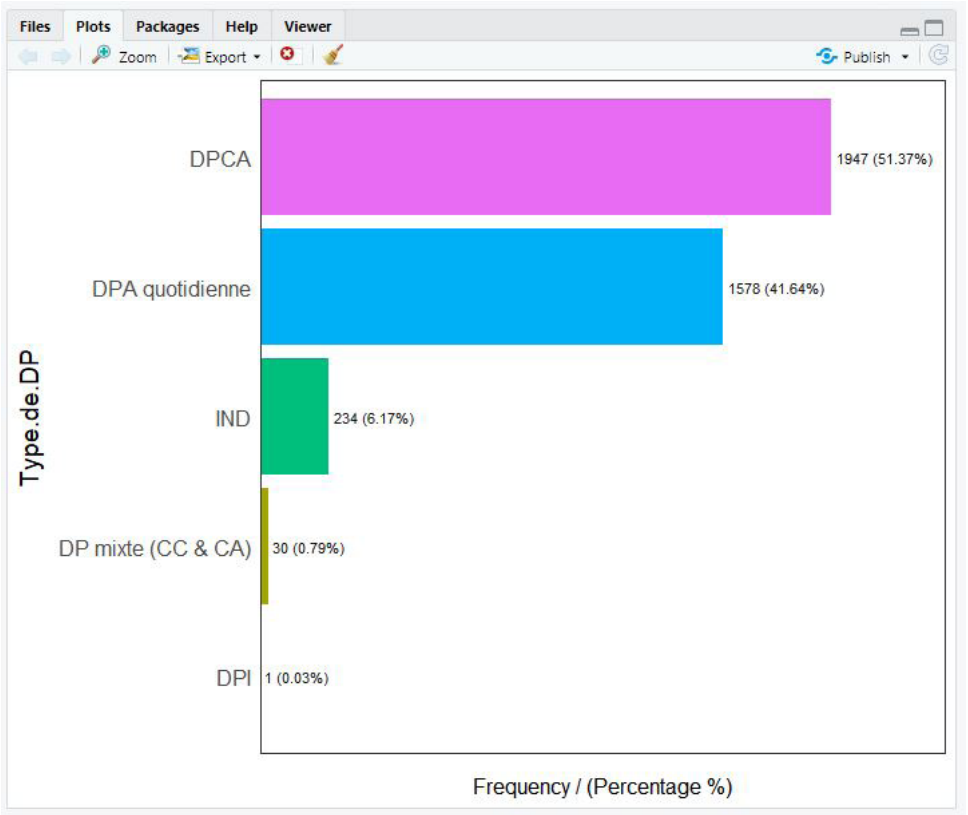
www.bdd.rdplf.org Volume 2, n° 2, Juin 2019
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917

```
> summary(mydata)
   code.post             PAYS        sexe       Age.1ere.DP       Charlson
 10000  :  85   BELGIQUE  : 244   F:1265   Min.   :16.01   Min.   : 2.000
 4000   :  70   FRANCE    :2639   M:1867   1st Qu.:54.77   1st Qu.: 4.000
 98849  :  60   LUXEMBOURG:  10            Median :68.10   Median : 6.000
 14033  :  54   MAROC     :  99            Mean   :65.07   Mean   : 5.843
 75877  :  52   SUISSE    :  33            3rd Qu.:77.69   3rd Qu.: 8.000
 63400  :  51   TUNISIE   : 107            Max.   :98.39   Max.   :16.000
 (Other):2760
 Charlson_modif            Type.de.DP       Taille          Poids
 Min.   :-2.000   DPA quotidienne:1348   Min.   :100.0   Min.   : 30.00
 1st Qu.: 2.000   DPCA           :1782   1st Qu.:160.0   1st Qu.: 61.00
 Median : 3.000   DPI            :   1   Median :166.0   Median : 72.00
 Mean   : 3.708   IND            :   1   Mean   :166.1   Mean   : 72.63
 3rd Qu.: 5.000                          3rd Qu.:173.0   3rd Qu.: 83.00
 Max.   :14.000                          Max.   :196.0   Max.   :128.00
```



www.bdd.rdplf.org Volume 2, n° 2, Juin 2019
https://doi.org/10.25796/bdd.v2i2.20513
ISSN 2607-9917

```
##         Type.de.DP frequency percentage cumulative_perc
## 1             DPCA      1782      56.90           56.90
## 2 DPA quotidienne      1348      43.04           99.94
## 3              DPI         1       0.03           99.97
## 4              IND         1       0.03          100.00
```



If no variable is defined as an argument to the freq () function, it is automatically applied to all categorical variables.

5. To conclude

I hope that this first introductory article to R have convinced you to take your first steps with this software, and that it is pushing you wanting to know more. The next article will be dedicated to the data visualization , that is to say to the realization of graphics. Later, we'll see how to do reporting with R. Because analyzing your data is good, but releasing an analysis report is even better! Especially if it is done in a completely automated way.

In the meantime, if you wish to acquire new knowledge, here is a list of particularly interesting French resources:

- the training (in French)"Introduction à la statistique avec R"(http://bit.ly/2XPyOXF) on the FUN platform, provided by Bruno Falissard and Christophe Lalanne. This is a very complete training, over 5 weeks. Unfortunately, it is only open at certain times of the year.

- The online book Contes et Stats R by Lise Vaudor. (http://bit.ly/2RmC8H5)
- The online book «Introduction à R» and tidyverse by Julien Barnier. (http://bit.ly/2XpIwTH)
- The blog R-atique by Lise Vaudor. (http://perso.ens-lyon.fr/lise.vaudor/)
- The STHDA site by Alboukadel Kassambara. (http://www.sthda.com/french/wiki/wiki.php)
- The page »Débutants- commencez ici de mon blog Statistique et logiciel R».(http://bit.ly/2Km-c8ew)

You can find a more complete list of French resources here : https://wp.me/p93iR1-m6