

The background features a dark blue gradient with several orange vertical bars of varying heights. Overlaid on these bars are white lines with circular markers at various points, creating a complex, abstract pattern that suggests data analysis or statistical trends.

# Statistiques Descriptives en R

Alison PATOU

Patou.alison@gmail.com



# Programme

- Définitions
- Mesures du centre, mesure de la variation
- Rmarkdown

1

# Définitions

# Variables qualitatives

- Les caractères qualitatifs sont ceux dont les modalités ne peuvent pas être ordonnées, c'est-à-dire que si l'on considère deux caractères pris au hasard, on ne peut pas dire de l'un des caractères qu'il est inférieur ou égal à l'autre.
- Exemple : La région, le pays, couleurs sont des variables qualitatives

# Variables quantitatives

- Les caractères quantitatifs sont des caractères dont les modalités peuvent être ordonnées.
- Exemple : l'âge, la taille de vie ou le salaire d'un individu sont des caractères quantitatifs

# Effectif

L'effectif de la valeur  $x_i$  est le nombre d'individus de la population ayant cette valeur ou appartenant à cette classe : on le note  $n_i$ .

L'effectif total  $N$  est la somme de tous les effectifs :  $N = n_1 + n_2 + \dots + n_k$ .

En rangeant les valeurs du caractère dans l'ordre croissant, on peut calculer l'effectif cumulé croissant en faisant la somme des effectifs de cette valeur et de tous ceux qui la précèdent.

## Exemple

Note	19	11	8	12	10	17	8	10	12
Note	8	10	11	12	17	19	Effectif		
Effectif	2	2	1	1	1	1			
Note	8	10	11	12	17	19			
Effectif	2	2	1	1	1	1			
Eff. cumulé croissant	2	4	5	6	7	8			

# Fréquence

La fréquence d'une valeur est le quotient de l'effectif de la valeur par l'effectif total.

En rangeant les valeurs du caractère dans l'ordre croissant, on peut calculer les fréquences cumulées croissantes en faisant la somme des fréquences de cette valeur et de tous ceux qui la précèdent.

**La fréquence est comprise entre 0 et 1**

## *Exemple*

Note	8	10	11	12	17	19
Effectif	2	2	1	1	1	1
Eff. cumulé croissant	2	4	5	6	7	8
Fréquence	0,25	0,25	0,125	0,125	0,125	0,125
Fréq. Cumulée croissant	0,25	0,5	0,625	0,75	0,875	1

# Manipulation de données

```
> iris
  Sepal.Length Sepal.width Petal.Length Petal.width  Species
1          5.1         3.5         1.4         0.2    setosa
2          4.9         3.0         1.4         0.2    setosa
3          4.7         3.2         1.3         0.2    setosa
4          4.6         3.1         1.5         0.2    setosa
5          5.0         3.6         1.4         0.2    setosa
-          -         -         -         -         -
```

On peut manipuler les matrices pour en extraire des vecteurs ou des sous-matrices.

```
> iris[,1:4]
  Sepal.Length Sepal.width Petal.Length Petal.width
1          5.1         3.5         1.4         0.2
2          4.9         3.0         1.4         0.2
3          4.7         3.2         1.3         0.2
4          4.6         3.1         1.5         0.2
5          5.0         3.6         1.4         0.2
```

**Matrices[lignes, colonnes]**

Pour la selection d'une colonne unique :

```
> iris[,c(-1,-4)]
  Sepal.width Petal.Length  Species
1          3.5         1.4    setosa
2          3.0         1.4    setosa
3          3.2         1.3    setosa
4          3.1         1.5    setosa
5          3.6         1.4    setosa
```

**Matrice\$colonne**



## Fonctions utiles

- **tapply()** permet de réaliser des calculs sur un vecteur, conditionnellement aux valeurs prises par un ou plusieurs facteurs

```
tapply(iris$Sepal.Length, iris$Species, mean)
      setosa versicolor virginica
      5.006      5.936      6.588
```

## Fonctions utiles

- **table()** permet de calculer le nombre d'effectif associé à chaque valeur de la colonne.

```
table(iris$Species)
```

```
      setosa versicolor virginica  
      50         50         50
```

```
.
```

2

Mesures du centre

# Mesures du centre

Mesurer le centre permet de comprendre quelles sont les caractéristiques d'une observation qui serait le centre de l'échantillon, autrement dit une observation qui représente un exemple typique tiré de la population.

On dispose de trois manières usuelles de mesurer le centre de la distribution d'une variable :

- Moyenne
- Médiane
- Mode

# Moyenne

La moyenne correspond à la somme de toutes les mesures de votre échantillon divisé par le nombre total d'individus.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

n = effectif total  
xi = i-ème valeur de la variable

## Exemple

Elève	Note
Matthieu	12
Lisa	17
Jean	10
Gaspard	15



$$\text{Moyenne} = \frac{12 + 17 + 10 + 15}{4} = 13,5$$

# Moyenne (formule généralisée)

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \dots + n_kx_k}{n_1 + n_2 + \dots + n_k} = \frac{1}{N} \times \sum_{i=1}^{i=k} n_i x_i$$

N = effectif total

n<sub>i</sub> = l'effectif de la valeur x<sub>i</sub>

x<sub>i</sub> = i-ème valeur de la variable

## Exemple

Age	Effectif
20	3
21	2
22	2
23	6



$$\text{Moyenne} = \frac{20 \cdot 3 + 21 \cdot 2 + 22 \cdot 2 + 23 \cdot 6}{3 + 2 + 2 + 6} = 21,8$$

# Sous R

*Tous les exemples de ce cours seront avec le dataset Iris, présent nativement sous R.*

```
# Affichage de la structure du dataset
```

```
str(iris)
```

```
'data.frame':  150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species      : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Affichage de la moyenne de la colonne Sepal.Length
```

```
> mean(iris$Sepal.Length)
```

```
[1] 5.843333
```

# Médiane

La médiane est littéralement le milieu de votre échantillon. Il faut cependant que celui-ci soit ordonné de la plus petite valeur à la plus grande.

## *Exemple*

Personne	Salaire
Matthieu	19 000\$
Lisa	27 000\$
Jean	35 000\$
Gaspard	47 000\$
Sarah	56 000\$



La médiane est de 35 000\$

Personne	Salaire
Matthieu	19 000\$
Lisa	27 000\$
Jean	35 000\$
Gaspard	47 000\$



La médiane est de 31 000\$



# Sous R

*Tous les exemples de ce cours seront avec le dataset Iris, présent nativement sous R.*

```
# Affichage de la structure du dataset
```

```
str(iris)
```

```
'data.frame':  150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species      : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Affichage de la médiane de la colonne Sepal.Length
```

```
> median(iris$Sepal.Length)
```

```
[1] 5.84
```

# Médiane

## Généralisation aux quantiles

Quantile d'ordre  $\frac{1}{4}$  (1<sup>er</sup> quartile) : C'est la valeur  $Q1$  tel que  $F(Q1) = 0.25$ .

Quantile d'ordre  $\frac{3}{4}$  (3<sup>ème</sup> quartile) : C'est la valeur  $Q3$  tel que  $F(Q3) = 0.75$  (on a  $Me = Q2$ ).

Déciles d'ordre  $1/10, 2/10, \dots$  :  $F(D1)=0.1, F(D2)=0.2$ .

# Sous R

*Tous les exemples de ce cours seront avec le dataset Iris, présent nativement sous R.*

```
# Affichage de la structure du dataset
```

```
str(iris)
```

```
'data.frame':  150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species      : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Affichage des quantiles de la colonne Sepal.Length
```

```
> quantile(iris$Sepal.Length)
```

```
0% 25% 50% 75% 100%
4.3 5.1 5.8 6.4 7.9
```

# Mode

Le mode est le nombre qui apparaît le plus fréquemment dans votre échantillon. S'il y n'a aucune valeur qui se répète, alors le mode ne peut pas être calculé. A l'inverse si vous avez plusieurs valeurs qui se répètent, on prend la fréquence la plus grande

# Sous R

```
# Affichage du summary de la colonne Sepal.Length  
summary(iris$Sepal.Length)
```

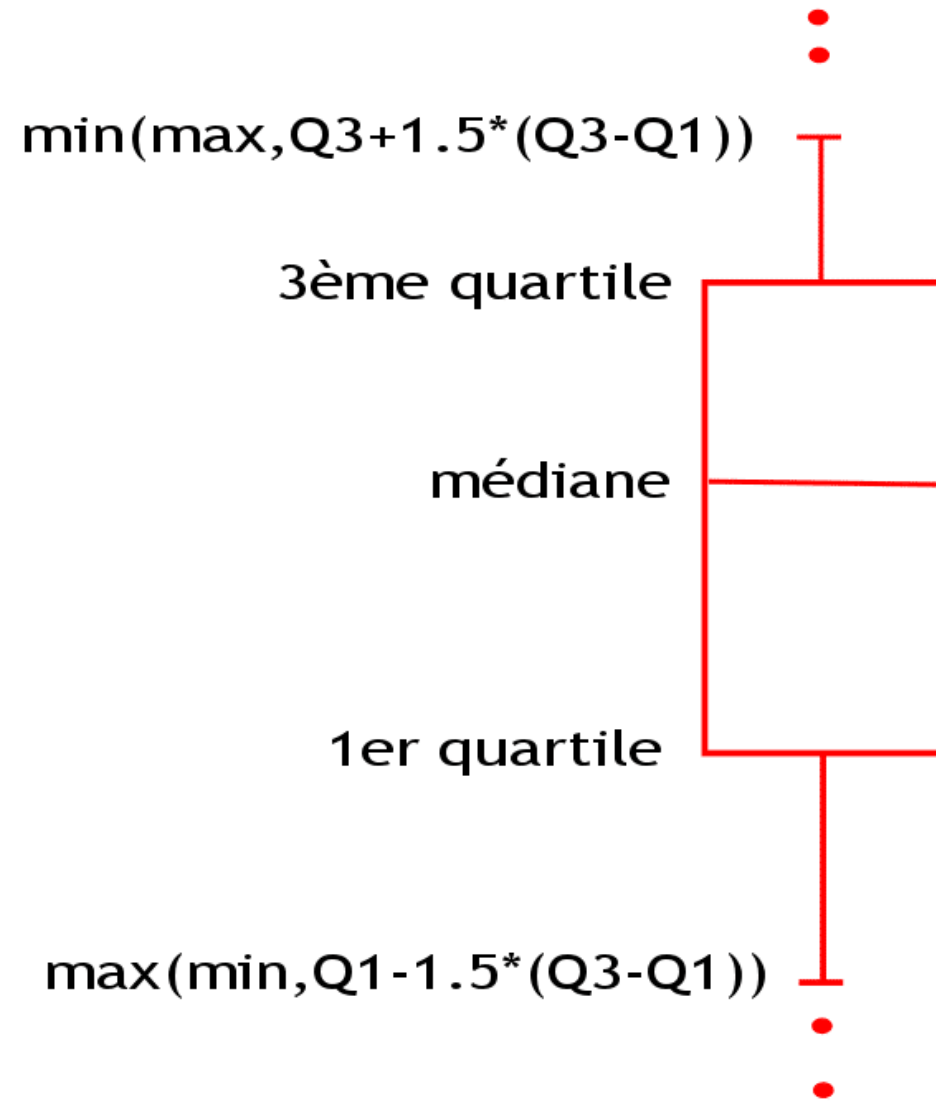
```
Min. 1st Qu. Median Mean 3rd Qu. Max.  
4.300 5.100 5.800 5.843 6.400 7.900
```



Visualisation

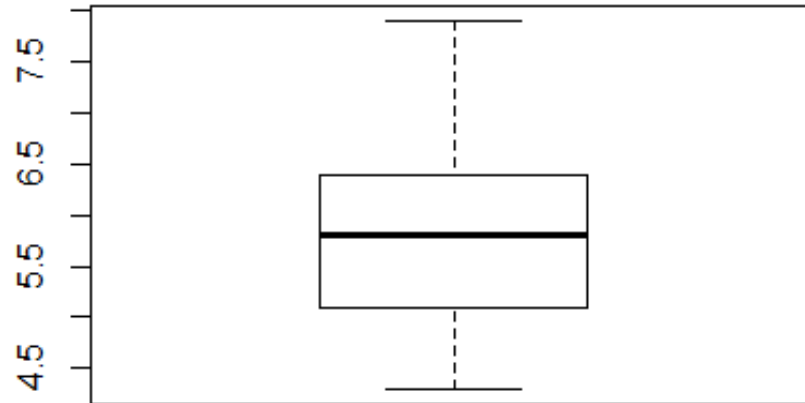
**La boîte à moustaches / Le boxplot**

## Boxplot – Boite à moustaches



# Sous R

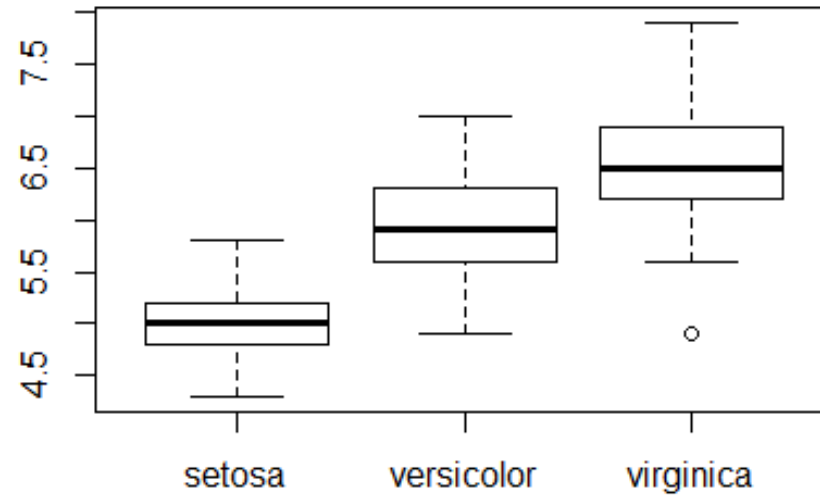
```
# Création d'un boxplot sur la colonne Sepal.Length  
boxplot(iris$Sepal.Length)
```



Quelle est la médiane ?  
Le 1<sup>er</sup> et 3<sup>ème</sup> quartile ?

# Sous R

```
# Création d'un boxplot sur la colonne Sepal.Length en fonction de l'espèce  
boxplot(iris$Sepal.Length ~ iris$Species)
```





# R Markdown

Comment optimiser et  
présenter son code ?

---

# Notebook

**Il existe sous Python, différents notebooks afin de centraliser code et résultats : Jupyter est souvent cité en référence.**

***Quid de R ?***

# Notebook

## Qu'est ce que c'est ?

Un notebook contient le code du développeur mais aussi les différentes étapes d'analyses, les visualisations, les commentaires et des découpages grâce à des titres et sous-titres pour une meilleure lisibilité.

Cela permet notamment une meilleure :

- Efficacité
- Interactivité
- Collaborativité
- Reproductibilité
- Automatisation
- ...

# Notebook

## Quand l'utiliser ?

Le notebook s'utilise durant les différentes étapes d'analyse :

- **Nettoyage des données** : faire le tri entre les données importantes et celles qui ne le sont pas dans l'analyse des ensembles de mégadonnées
- **Modélisation statistique** : méthode mathématique permettant d'établir la probabilité de répartition d'une caractéristique particulière
- **Création et mise en œuvre de modèles d'apprentissage automatique** : étude, programmation et apprentissage de modèles
- **Visualisation de données** : représentation graphique de données pour faire apparaître des structures, des tendances, des relations, etc.

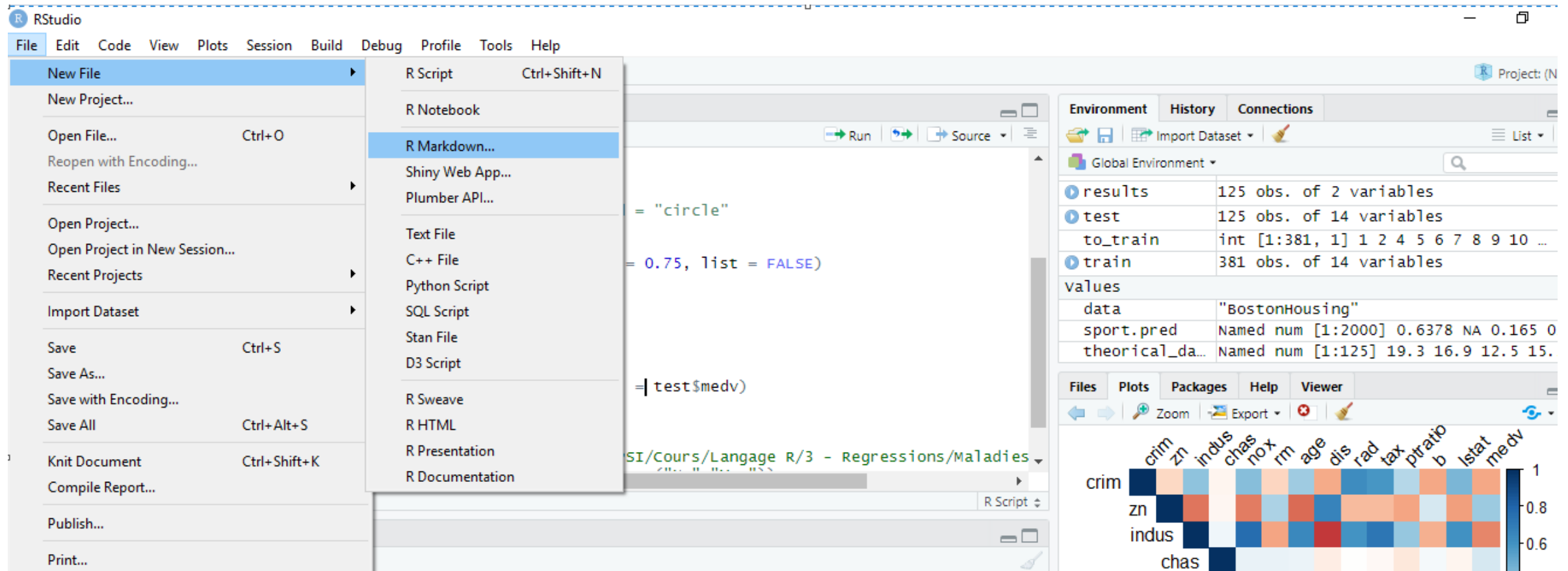
# R Markdown

- R Markdown est un package qui, combiné à R Studio, permet la création rapidement d'un notebook de travail et ainsi créer une page Web.
- Le package Markdown fonctionne avec le package ***knitr***.

Le package ***knitr*** permet de créer des tableaux mais aussi des graphiques et ainsi permettre une reproductibilité des travaux de manière efficace et simple.

# R Markdown

## Ouvrir R Markdown :



The screenshot shows the RStudio interface. The 'File' menu is open, and 'R Markdown...' is selected. The background shows a code editor with R code and a console window.

**File Menu Options:**

- New File
  - R Script (Ctrl+Shift+N)
  - R Notebook
  - R Markdown...**
  - Shiny Web App...
  - Plumber API...
- New Project...
- Open File... (Ctrl+O)
- Reopen with Encoding...
- Recent Files
- Open Project...
- Open Project in New Session...
- Recent Projects
- Import Dataset
- Save (Ctrl+S)
- Save As...
- Save with Encoding...
- Save All (Ctrl+Alt+S)
- Knit Document (Ctrl+Shift+K)
- Compile Report...
- Publish...
- Print...

**Code Editor Content:**

```
= "circle"  
= 0.75, list = FALSE)  
= test$medv)  
SI/Cours/Langage R/3 - Regressions/Maladies
```

**Environment Panel:**

Object	Details
results	125 obs. of 2 variables
test	125 obs. of 14 variables
to_train	int [1:381, 1] 1 2 4 5 6 7 8 9 10 ...
train	381 obs. of 14 variables
data	"BostonHousing"
sport.pred	Named num [1:2000] 0.6378 NA 0.165 0
theoretical_da...	Named num [1:125] 19.3 16.9 12.5 15.

**Plots Panel:**

Heatmap showing the relationship between variables: crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, b, lstat, medv. The color scale ranges from 0.6 (dark blue) to 1.0 (dark red).

**Document**

Presentation

Shiny

From Template

**Title:**

**Author:**

**Default Output Format:**

☒ **HTML**  
Recommended format for authoring (you can switch to PDF or Word output anytime).

☐ **PDF**  
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

☐ **Word**  
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).

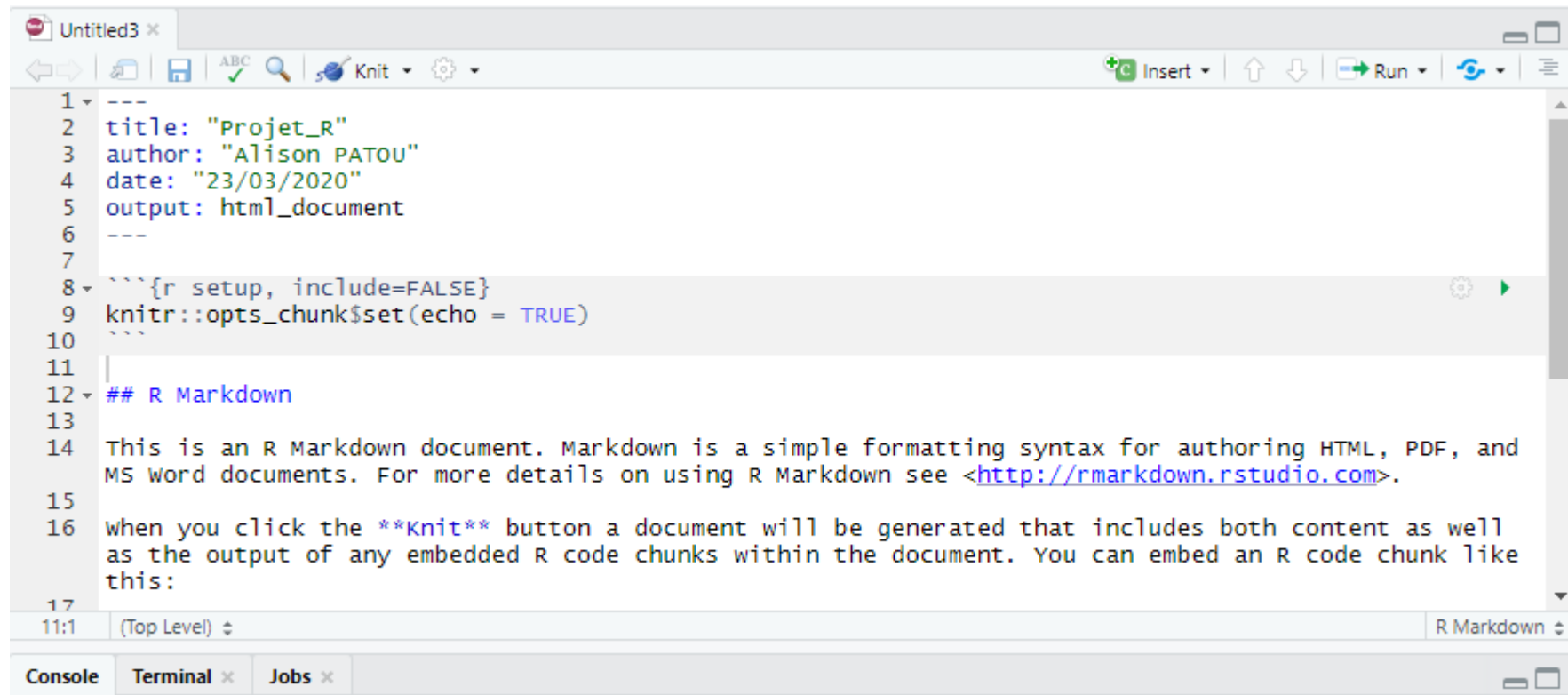
OK Cancel

## R Markdown

- **3 fichiers en « sortie » possibles :**
- HTML : le plus commun, sera lu via un navigateur
- PDF : pour des rapports mais requiere tex
- Word



# Structure R Markdown



The screenshot shows a text editor window titled 'Untitled3' with a toolbar at the top containing icons for navigation, saving, and running. The document content is as follows:

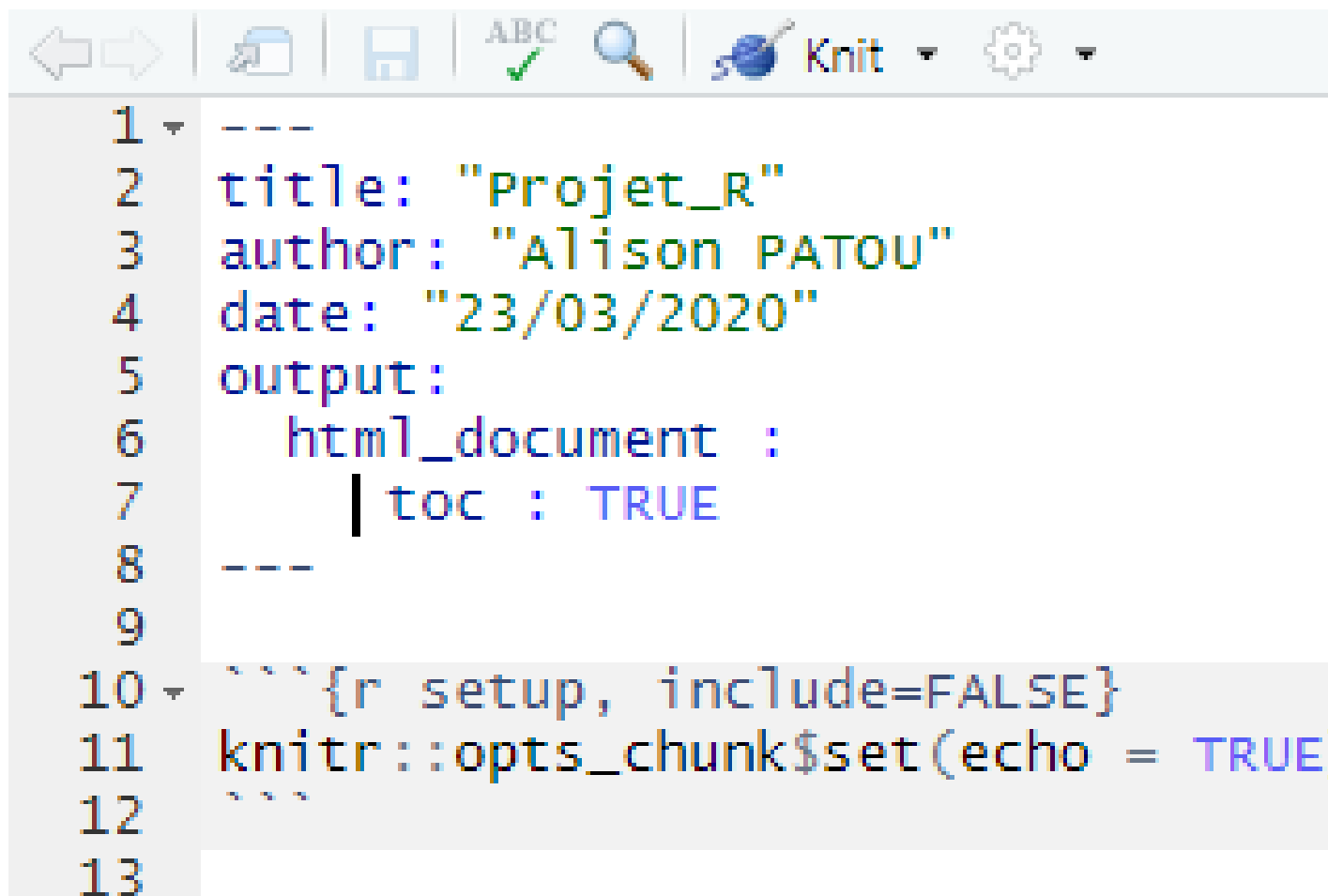
```
1 ---
2 title: "Projet_R"
3 author: "Alison PATOU"
4 date: "23/03/2020"
5 output: html_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and
15 MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
16
17 when you click the **knit** button a document will be generated that includes both content as well
18 as the output of any embedded R code chunks within the document. You can embed an R code chunk like
19 this:
```

The editor interface includes a status bar at the bottom with tabs for 'Console', 'Terminal', and 'Jobs', and a dropdown menu showing 'R Markdown'.

Header (présentation) du document

Corps du document

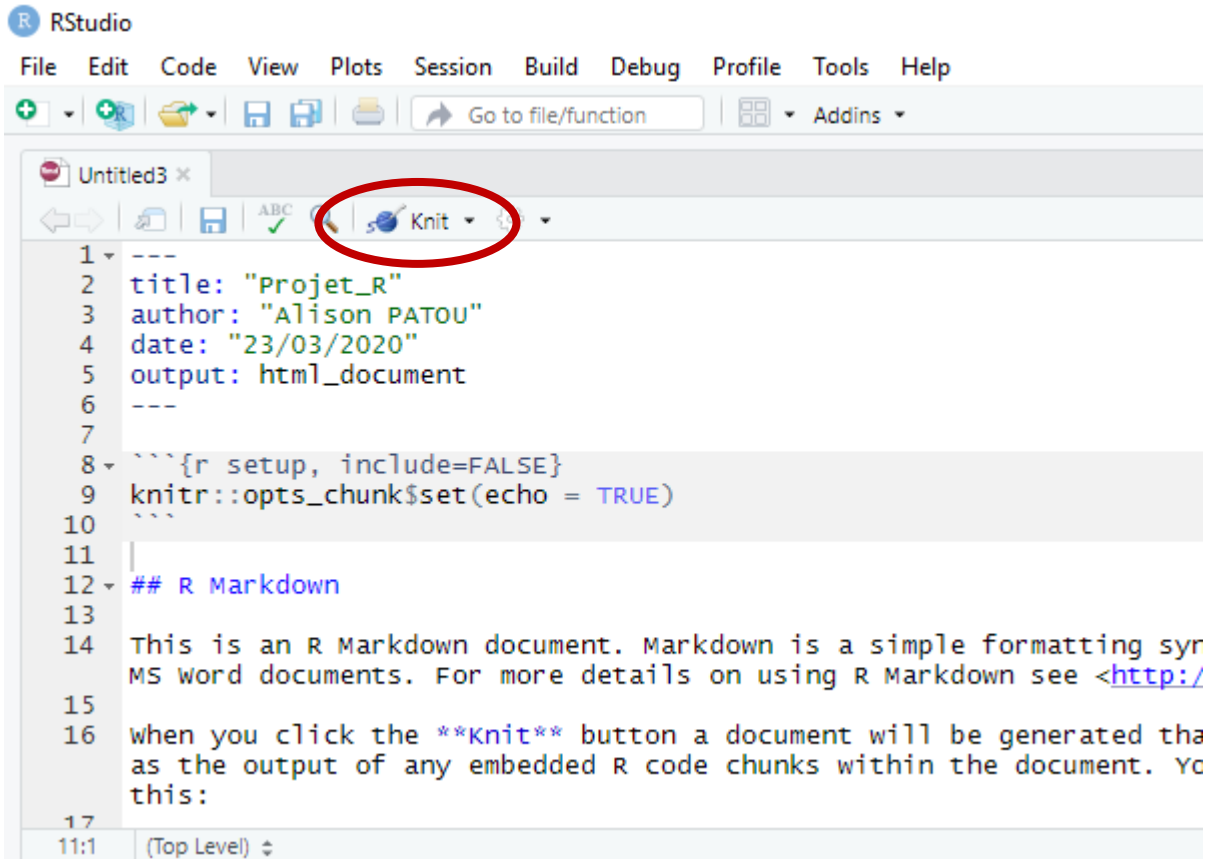
# Header



```
1 ---
2 title: "Projet_R"
3 author: "Alison PATOU"
4 date: "23/03/2020"
5 output:
6   html_document :
7     | toc : TRUE
8 ---
9
10 ```{r setup, include=FALSE}
11 knitr::opts_chunk$set(echo = TRUE
12 ```
13
```

- Plusieurs customisations sont possible concernant le Header (numérotation des titres, insertion logo, ...)
- Toc : TRUE
  - > permet l'affichage de votre structure de document (titre, sous-titre, ...)

# Exécution

A screenshot of the RStudio application window. The title bar says 'RStudio'. The menu bar includes 'File', 'Edit', 'Code', 'View', 'Plots', 'Session', 'Build', 'Debug', 'Profile', 'Tools', and 'Help'. Below the menu bar is a toolbar with various icons. The 'Knit' button, which has a blue ball of yarn icon, is circled in red. Below the toolbar is a text editor window titled 'Untitled3'. It contains R Markdown code. The code starts with a YAML header: 'title: "Projet\_R"', 'author: "Alison PATOU"', 'date: "23/03/2020"', and 'output: html\_document'. This is followed by an R code chunk: '{r setup, include=FALSE}' and 'knitr::opts\_chunk\$set(echo = TRUE)'. Then there is a section header '## R Markdown' and a paragraph of text explaining that clicking the 'knit' button will generate a document with the output of the R code chunks. The status bar at the bottom shows '11:1' and '(Top Level)'.

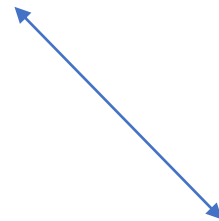
```
1 ---
2 title: "Projet_R"
3 author: "Alison PATOU"
4 date: "23/03/2020"
5 output: html_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syr
15 MS word documents. For more details on using R Markdown see <http:/
16 when you click the knit button a document will be generated tha
17 as the output of any embedded R code chunks within the document. Yc
18 this:
19
20 11:1 (Top Level)
```

Le bouton knit va tricoter (exécuter)  
votre code pour créer votre notebook

# Exécution

Vos deux fichiers ont été créés (le .Rmd et le .html)

Markdown_cours	23/03/2020 11:15	Document HTML	629 Ko
Markdown_cours.Rmd	23/03/2020 11:15	Fichier RMD	1 Ko



RStudio

File Edit Code View Plots Console Environment

~/Personnel/EPIS/Cours/Langage R/4 - Rmarkdown/Markdown\_cours.html

Markdown\_cours.html Open in Browser Find

## Projet\_R

Alison PATOU  
23/03/2020

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:


```
summary(cars)
```

```
##      speed      dist
##  Min.   : 4.0   Min.   : 2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

## Including Plots

You can also embed plots, for example:

# Mise en forme

Syntax	Becomes
Plain text	Plain text
End a line with two spaces to start a new paragraph.	End a line with two spaces to start a new paragraph.
<code>*italics*</code> and <code>_italics_</code>	<i>italics and italics</i>
<code>**bold**</code> and <code>__bold__</code>	<b>bold and bold</b>
<code>superscript^2^</code>	<sup>superscript</sup> 2
<code>--strikethrough--</code>	<del>strikethrough</del>
<code>[link](www.rstudio.com)</code>	<a href="http://www.rstudio.com">link</a>
<code># Header 1</code>	<h2>Header 1</h2>
<code>## Header 2</code>	<h3>Header 2</h3>
<code>### Header 3</code>	<h4>Header 3</h4>
<code>#### Header 4</code>	<h5>Header 4</h5>
<code>##### Header 5</code>	<h6>Header 5</h6>
<code>##### Header 6</code>	<h7>Header 6</h7>
<code>endash: --</code>	endash: —
<code>emdash: ---</code>	emdash: —
<code>ellipsis: ...</code>	ellipsis: ...
<code>inline equation: <math>A = \pi r^2</math></code>	inline equation: $A = \pi r^2$
<code>image: </code>	image: 

**NB : La Cheat Sheet R Markdown est sur mon github (<https://github.com/apatou>)**

# Code

L'avantage de Rmarkdown c'est que vous pouvez continuer d'écrire votre code comme dans un script R classique

Votre Code Chunk (corps de texte) va se trouver sous cette zone :

```
7  
8 {r setup, include=FALSE}  
9 knitr::opts_chunk$set(echo = TRUE)  
10  
11
```

Chunk options		
option	default value	description
Code evaluation		
child	NULL	A character vector of filenames. Knitr will knit the files and place them into the main document.
code	NULL	Set to R code. Knitr will replace the code in the chunk with the code in the code option.
engine	'R'	Knitr will evaluate the chunk in the named language, e.g. <code>engine = 'python'</code> . Run <code>names(knitr::knit_engines\$get())</code> to see supported languages.
eval	TRUE	If FALSE, knitr will not run the code in the code chunk.
include	TRUE	If FALSE, knitr will run the chunk but not include the chunk in the final document.
purL	TRUE	If FALSE, knitr will not include the chunk when running <code>purL()</code> to extract the source code.
Results		
collapse	FALSE	If TRUE, knitr will collapse all the source and output blocks created by the chunk into a single block.
echo	TRUE	If FALSE, knitr will not display the code in the code chunk above it's results in the final document.
results	'markup'	If 'hide', knitr will not display the code's results in the final document. If 'hold', knitr will delay displaying all output pieces until the end of the chunk. If 'asis', knitr will pass through results without reformatting them (useful if results return raw HTML, etc.)
error	TRUE	If FALSE, knitr will not display any error messages generated by the code.
message	TRUE	If FALSE, knitr will not display any messages generated by the code.
warning	TRUE	If FALSE, knitr will not display any warning messages generated by the code.
Code Decoration		
comment	"##"	A character string. Knitr will append the string to the start of each line of results in the final document.
highlight	TRUE	If TRUE, knitr will highlight the source code in the final output.
prompt	FALSE	If TRUE, knitr will add > to the start of each line of code displayed in the final document.
strip.white	TRUE	If TRUE, knitr will remove white spaces that appear at the beginning or end of a code chunk.
tidy	FALSE	If TRUE, knitr will tidy code chunks for display with the <code>tidy_source()</code> function in the <code>formatR</code> package.

# Code

The screenshot shows the RStudio interface with an R Markdown document open. The document is titled 'Projet\_R' and includes the following content:

```
1 ---
2 title: "Projet_R"
3 author: "Alison PATOU"
4 date: "23/03/2020"
5 output: html_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 ### Import de la librairie
15 ```{r}
16 #Charger la librairie data.table
17 library(data.table)
18 ```
19
20 ### Import dataset
21 ```{r}
22 datasetfleur = fread("C:/Users/patou/Documents/Personnel/EPST/Cours/Langage R/2 - 3
23 ce/datasetfleur.csv")
24 ```
25
26 ### Visualisation
27 ```{r}
28 plot(datasetfleur$Valeurs)
29 ```
30 Globalement les valeurs semblent assez hom
```

The right pane shows the rendered HTML output of the document, which includes the title 'Projet\_R', the author 'Alison PATOU', the date '23/03/2020', and the section 'R Markdown'. Below this, the 'Import de la librairie' section is rendered with a code block for loading the 'data.table' library. The 'Import dataset' section is rendered with a code block for loading the 'datasetfleur' dataset. The 'Visualisation' section is rendered with a code block for plotting the 'Valeurs' column of the 'datasetfleur' dataset. The bottom pane shows a console window with the output of the code execution.

- ## Titre
- ### Sous titre
- Le code R qui va être exécuté doit se trouver entre :
  - ```{r}
  - VOTRE CODE
  - ```

# Pour aller plus loin

- Possibilité de customiser le rendu HTML avec du CSS (markdown.css)
- Ne pas hésiter à mettre des visualisations/graphiques dynamiques. Plusieurs packages en proposent :
  - Ggplot/ggplot2
  - Plotly
  - rAmChart