Data Mining et Data visualisation

Cours 1 : Introduction au langage R
et graphiques de base

Laurent Politis
politis.laurent@gmail.com
ESSCA 2022-2023

INTRODUCTION

Le langage R

- Vocation aux études statistiques et à la visualisation de données
- Logiciel de compilation gratuit et en « open source » (Rstudio est l'interface)
- Nouveaux développements en permanence (packages)
- Très flexible et grande communauté d'utilisateurs
- Compatible avec Windows, Linux, Mac

https://openclassrooms.com/fr/courses/1393696-effectuez-vos-etudes-statistiques-avec-r/1393758-introduction-a-r

Les principaux outils dérivés

- Création de packages pour intégrer d'autres langages de programmation ou vos propres fonctions

- Langage R Shiny Dashboard pour application web

  Quelques exemples : https://shiny.rstudio.com/gallery/

- Fonctionnalité R Markdown pour des rapports écrits (équivalent à LaTeX)

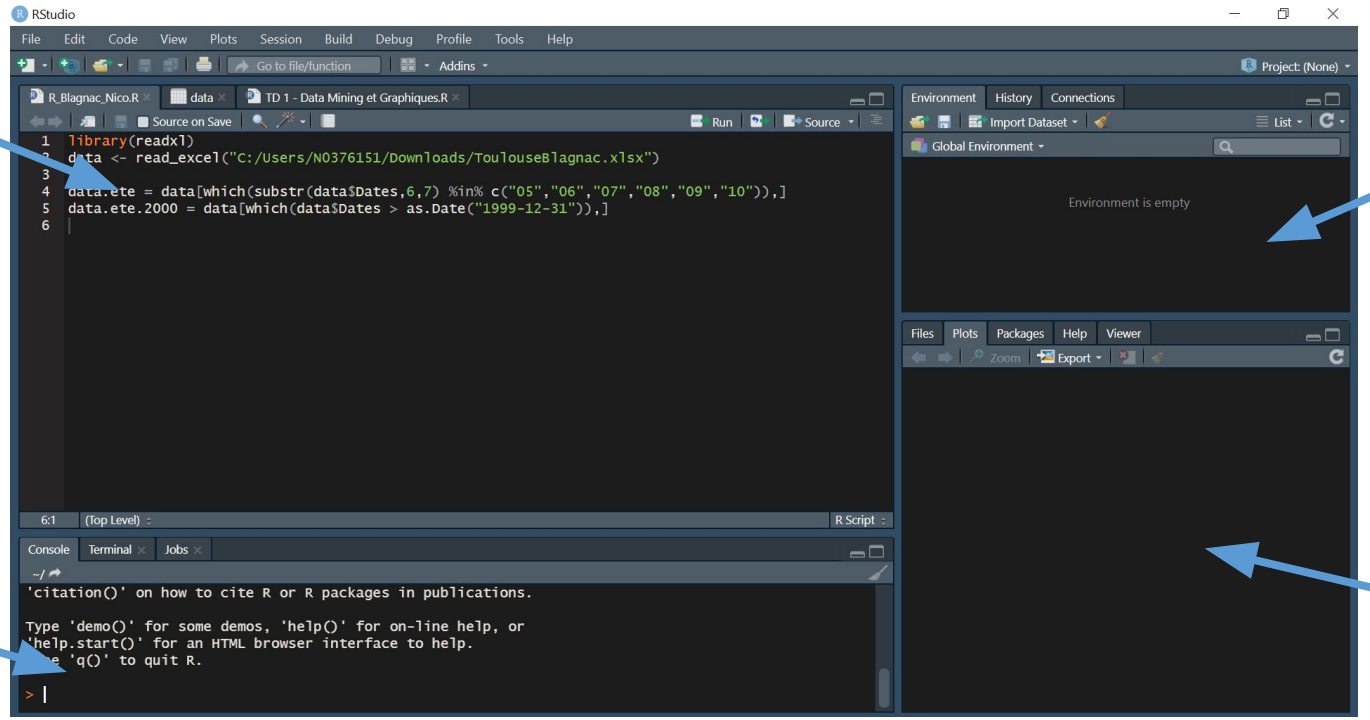
  Quelques exemples :
  https://rmarkdown.rstudio.com/gallery.html

RStudio

- Interface privilégiée pour le développement de langage R, gratuite

- Composée en plusieurs parties :

Votre script



Vos variables

La console

La visualisation de données

LES BASES DE R

Le cœur de la programmation sur R

- Vous pouvez créer des **variables** avec une ligne de code, importer des tableaux à partir de fichier txt, csv, excel…
- Avec des **fonctions** et des **boucles**, vous pouvez modifier ces variables, les refaçonner selon vos intérêts
- Plusieurs outils sont ensuite à votre disposition pour visualiser les résultats et les transmettre : **graphiques, cartes, rapports dynamiques…**

La gestion des erreurs : vous serez confrontés forcément à des messages d'erreur lors de vos travaux, jeter un œil à la documentation des fonctions, vérifier les classes des variables et fouiller sur Internet sont les meilleurs moyens de les régler

Quelques classes de variables

- Classes à 1 dimension : vector (numeric, character, date, logical, …)

  data[1] désignera le 1er élément du vecteur

- Classes à 2 dimensions : matrix, data.frame, table

  data[1,1] désignera l'élément de la 1ère ligne / 1ère colonne du tableau

- Classes à n dimensions : array

  data[,,1] désignera l'ensemble du 1er tableau de l'array

- Autre : list (un peu « fourre-tout »)

  data[[1]] désignera le premier élément de la liste

La fonction ***class()*** permettra de déterminer quelle est la classe de la variable
Exemple : ***class(x=data)***

Les boucles

- La boucle ***for*** s'utilise pour effectuer un nombre déterminé de fois une commande ou une série de commandes (ici 5 fois) :

```
for(i in 1:5){
  print(i)
}
```

Le bout de code ci-dessus permet d'afficher (grâce à la fonction ***print()***) les valeurs que prend la variable ***i***. La première ligne détermine que ***i*** prendra tour à tour les valeurs entières allant de 1 à 5 (« : »)

- La boucle ***if*** s'utilise pour effectuer une commande ou une série de commandes si une condition est vérifiée :

```
a = 5
if(a > 2){
  print(a)
}
```

Ici, on définit dans un premier temps la variable ***a*** (qui vaut 5). La condition ***a > 2*** étant vérifiée, on affiche la valeur de ***a***. Si ***a*** valait 1, la condition ne serait pas remplie et la commande ne serait donc pas exécutée

Le signe **=** assigne une valeur à une variable
Le signe **==** est un test qui compare les deux éléments de l'égalité (comparable à la condition vue ci-dessus)

Les fonctions

- Les fonctions dans R s'utilisent avec des arguments. Ex : f(x) x est l'argument Comme vu dans l'exemple ***class(x=data)***, il faut attribuer une valeur ou une variable à l'argument

- La plupart des fonctions ont plusieurs arguments, il faut donc préciser à quel argument on associe une variable

- Dans R, les fonctions font généralement partie d'un *package* qu'il est nécessaire d'installer (certains packages sont installés par défaut)

- Chaque fonction a une documentation (disponible sur internet) qui nous informe comment et pourquoi l'utiliser

Pour installer un package :
install.packages()
library()

DATA MINING
et
DATA VISUALISATION

Introduction au Data Mining

- Lorsque l'on travaille avec un nouveau set de données, il est important d'en prendre connaissance et de le retravailler pour l'utilisation que l'on va en avoir

- Cela consiste à savoir sous quel format la donnée est présentée, ce qu'elle contient et toute autre information la concernant (metadata)

- Il peut être nécessaire de convertir les unités, le format des dates, de retirer ou ajouter des colonnes et même de remplacer les valeurs manquantes

Un Data Mining approfondi fera appel aux outils de visualisation de la donnée : les graphiques

Introduction à la visualisation

- Les supports graphiques permettront de comprendre rapidement la dynamique de chaque paramètre du set de données

- On pourra mettre visuellement en évidence les valeurs manquantes ou aberrantes, les tendances et saisonnalité de la donnée ainsi que les éventuelles relations entre les paramètres du set de données

- De manière plus générale, les graphiques (s'ils sont bien construits) sont le meilleur moyen de présenter les résultats d'une étude, plus parlants que les phrases

Nous étudierons lors de ce cours les outils graphiques R de base
D'autres, plus esthétiques mais plus complexes, existent