

Introduction à la Data Science avec R

Fousseynou Bah

Faculté des Sciences Economiques et de Gestion (FSEG)
Université des Sciences Sociales et de Gestion de Bamako (USSGB)

30-Jan-2019

- 1 Définir la *data science*

- 2 R

- 3 RStudio

Définir la *data science*

Commençons par quelques exemples

Fait de la *data science*:

- l'économiste qui examine le niveau du PIB sur 30 ans et cherche à dégager des scénarii pour des futures évolutions
- le sociologue qui s'appuie le taux de natalité et le taux de participation des femmes au marché du travail pour comprendre l'évolution de la place de la femme dans la société
- le météorologue qui cherche à prédire la pluviométrie de la semaine à venir en modélisant les données historiques
- l'épidémiologue qui cartographie le taux de prévalence du paludisme pour appuyer un programme stratégique
- ...

Tout ça, pour dire que la *data science* est une discipline transversale. Elle se pratique dans plusieurs domaines.

Une discipline carrefour

Selon [Wikipédia](#), la *data science* est un champ interdisciplinaire qui utilise les méthode, processus, algorithmes et systèmes scientifiques pour extraire des données - tant structurées que non structurées - des informations utiles à la compréhension et à la prise de décision. De ce fait, elle s'appuie sur diverses méthodes (mathématiques, statistiques, informatiques, etc.) pour tirer des données une compréhension meilleure de phénomènes d'intérêt.

Le *data scientist*, le métier du 21eme siècle

Face à la génération massive des données, dans tous les secteurs, le besoin de *data scientist* se fait pressant.

Qu'est-ce qu'un *data scientist*? Quelqu'un qui pratique la *data science*, évidemment! Mais plus que ça, c'est un détective des données, quelqu'un qui investigue les données en vue de découvrir des récurrences, de reveler des éléments surprenants ou tout simplement de prendre la mesure des faits déjà connus...et tout ça dans le dessein de guider la prise de décision.

Comme tout bon détective, il se doit de posséder ou de cultiver un certains nombres de compétences. Quelles sont celles-ci?

Ce qu'il faut pour être *data scientist*

- Pas nécessairement un diplôme avancé en mathématiques ou en statistiques...quoiqu'il est utile de maîtriser des concepts de bases (vecteur, matrice, moyenne, écart-type, etc.)
- Pas forcément un diplôme en informatique ou en programmation...quoiqu'il est utile de connaître les notions de bases (qu'est-ce qu'un objet, un environnement? quels types d'objets peut-on manipuler dans un environnement donnée...)
- Une connaissance avérée dans un domaine spécifique dans lequel l'on peut soulever des questions, mobiliser des outils théoriques auxquels on confronte les résultats de l'analyse conduite sur les données
- Un esprit curieux, quelle que soit l'avenue que l'on emprunte.

Vous pourrez avoir une meilleure idée en surfant sur le net. Pour commencer, regardez [ici](#).

R

Qu'est-ce que c'est que R?

R est un langage de programmation et un logiciel gratuit et libre. Il est surtout utilisé pour le développement de programmes statistiques et des analyses de données. Il gagne en popularité depuis quelques années avec l'émergence de la *data science* et du fait qu'il est gratuit et ouvert (*open-source*).

R est née d'un projet de recherche mené par deux chercheurs, Ross Ihaka et Robert Gentleman à l'université d'Auckland (Nouvelle-Zélande) en 1993. En 1997 est mis en place le *Comprehension R Archive Network (CRAN)* qui centralise les contributions au projet. Depuis le projet connaît une croissance soutenue, grâce à des contributions de la part de milliers de personnes à travers le monde.

Pourquoi R?

Pour un apprenti *data scientist*, le choix du langage et/ou du programme est une décision critique. Considérant le temps qu'il investira en apprentissage et le retour qu'il espéra à travers l'utilisation de ses nouvelles connaissances dans sa profession, il est utile de considerer divers critères dont:

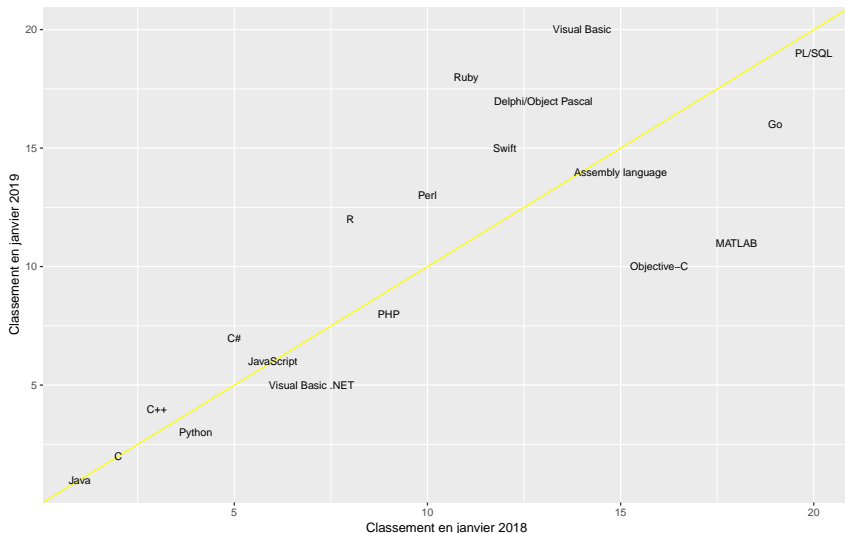
- l'accessibilité de l'outil en termes de coûts (tout n'est pas gratuit comme R!);
- l'accessibilité du langage en termes de syntaxe;
- la popularité du langage parmi les paires (en vue de faciliter des collaborations);
- la dynamique de développement du langage.

Et sur tous ces points, R présente des avantages.

R dans l'écosystème des langages: TIOBE Index (1)

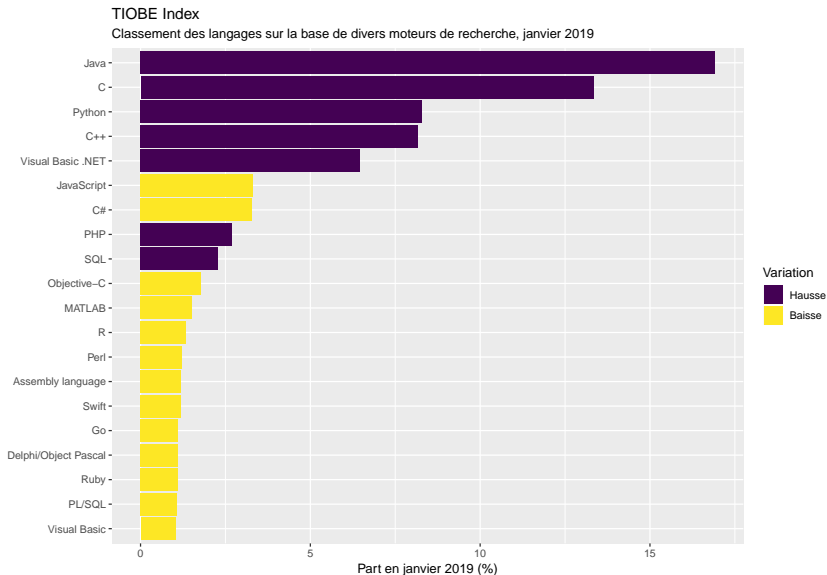
TIOBE Index

Classement des langages sur la base de divers moteurs, janvier 2019



Source: Données tirées de <https://www.tiobe.com/tiobe-index/>

R dans l'écosystème des langages: TIOBE Index (2)

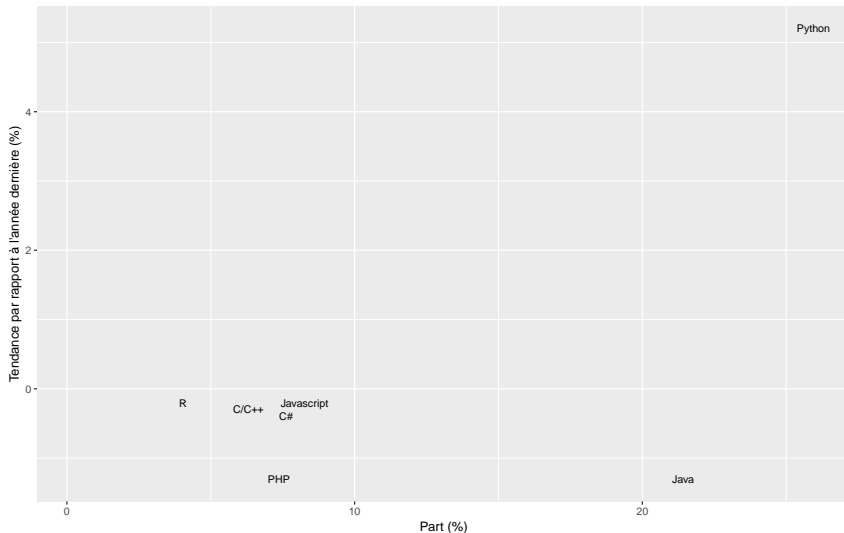


Source: Données tirées de <https://www.tiobe.com/tiobe-index/>

R dans l'écosystème des langages: PYPL Index (1)

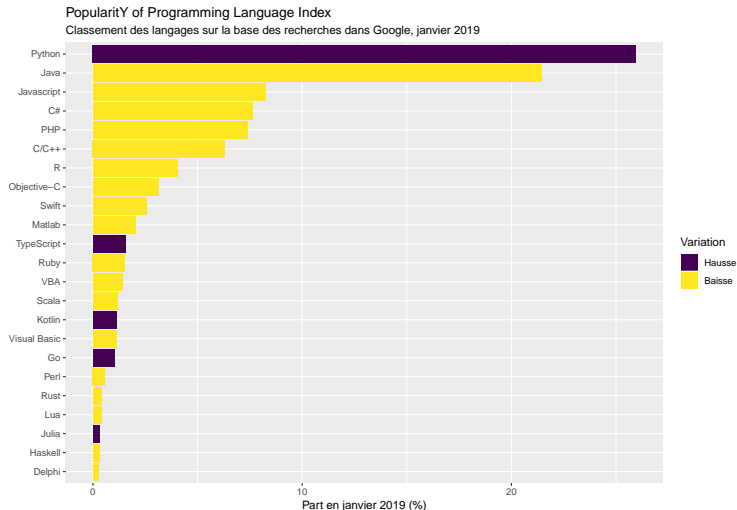
Popularity of Programming Language Index

Classement des langages sur la base des recherches dans Google, janvier 2019



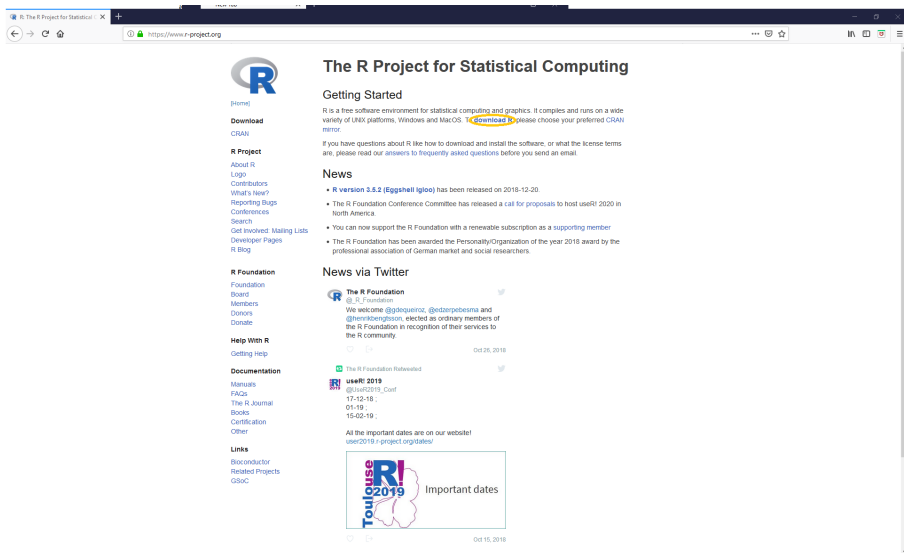
Source: Données tirées de <http://pypl.github.io/PYPL.html>

R dans l'écosystème des langages: PYPL Index (2)



Source: Données tirées de <http://pypl.github.io/PYPL.html>

Télécharger R (1)



The screenshot shows the official website of the R Project for Statistical Computing. The browser's address bar displays 'https://www.r-project.org'. The website layout includes a navigation menu on the left with links to Home, Download, CRAN, R Project, About R, Logo, Contributors, What's New?, Reporting Bugs, Conferences, Search, Get Involved, Mailing Lists, Developer Pages, and R Blog. The main content area features the title 'The R Project for Statistical Computing' and a 'Getting Started' section. This section explains that R is a free software environment for statistical computing and graphics, available on various platforms. It provides a 'Download' link and advises users to read the 'answers to frequently asked questions' before sending an email. Below this is a 'News' section with three bullet points: the release of R version 3.5.2 (Eggshell Igloo) on 2018-12-20, the R Foundation Conference Committee's call for proposals for the 2020 userR conference in North America, and the R Foundation's award of the Personality/Organization of the year 2018 award to the professional association of German market and social researchers. A 'News via Twitter' section follows, showing two tweets. The first tweet from @R_Foundation welcomes @gdequeiroz, @bedzperbesma, and @henrikbengtsson as new members of the R Foundation. The second tweet from @userR2019_Conf announces the userR 2019 conference dates: 17-12-18, 01-19, and 15-02-19. A graphic for 'Toulouse R! 2019' with the text 'Important dates' is also shown. The left sidebar contains links to the R Foundation (Foundation, Board, Members, Donors, Donate), Help With R (Getting Help), Documentation (Manuals, FAQs, The R Journal, Books, Certification, Other), and Links (Bioconductor, Related Projects, GSOC).

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. [Download](#) to please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

- **R version 3.5.2 (Eggshell Igloo)** has been released on 2018-12-20.
- The R Foundation Conference Committee has released a [call for proposals](#) to host userR 2020 in North America.
- You can now support the R Foundation with a renewable subscription as a [supporting member](#)
- The R Foundation has been awarded the Personality/Organization of the year 2018 award by the professional association of German market and social researchers.

News via Twitter

The R Foundation @R_Foundation
We welcome @gdequeiroz, @bedzperbesma and @henrikbengtsson, elected as ordinary members of the R Foundation in recognition of their services to the R community.
Oct 26, 2018

userR 2019 @userR2019_Conf
17-12-18 ;
01-19 ;
15-02-19 ;
All the important dates are on our website!
user2019.r-project.org/dates/
Oct 16, 2018

Toulouse R! 2019 Important dates

Télécharger R (2)

CRAN - Mirrors

https://cran.r-project.org/mirrors.html

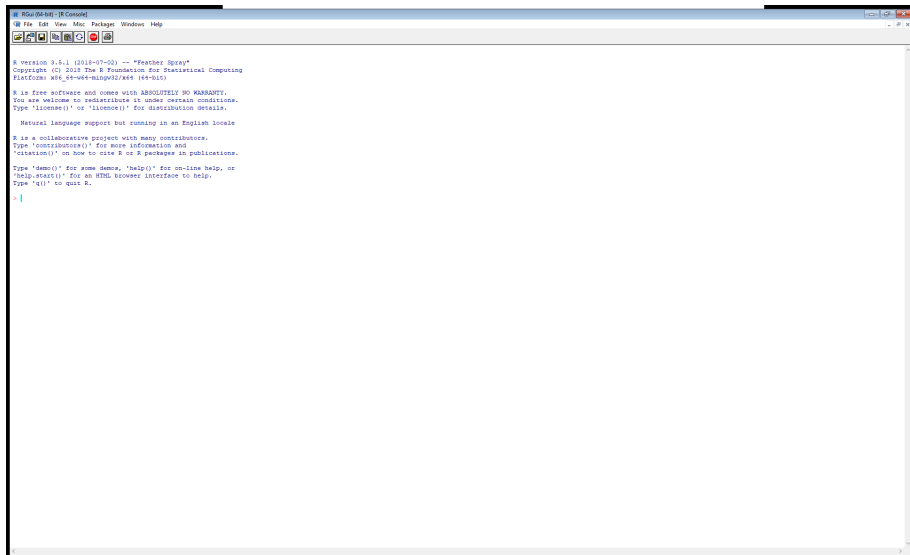
CRAN MIRRORS

The Comprehensive R Archive Network is available at the following URL's, please choose a location close to you. Some statistics on the status of the mirrors can be found here: [main page](#), [windows release](#), [windows old release](#).

If you want to host a new mirror at your institution, please have a look at the [CRAN Mirror HOWTO](#).

0-Cloud	https://cloud.r-project.org/ http://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio Automatic redirection to servers worldwide, currently sponsored by Rstudio
Algeria	https://cran.usbhb.dz/ http://cran.usbhb.dz/	University of Science and Technology Houari Boumediene University of Science and Technology Houari Boumediene
Argentina	http://mirror.fraulp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata
Australia	https://cran.csiro.au/ http://cran.csiro.au/ https://mirror.aarnet.edu.au/pub/CRAN/ https://cran.mus.unimelb.edu.au/ https://cran.curtin.edu.au/	CSIRO CSIRO AARNET School of Mathematics and Statistics, University of Melbourne Curtin University of Technology
Austria	https://cran.wu.ac.at/ http://cran.wu.ac.at/	Wirtschaftsuniversität Wien Wirtschaftsuniversität Wien
Belgium	http://www.freestatsci.org/cran/ https://lib.ugent.be/CRAN/ http://lib.ugent.be/CRAN/	K.U. Leuven Association Ghent University Library Ghent University Library
Brazil	http://lincapb.usp.br/mirrors/cran/ https://cran-e.cbl.ufpe.br/ http://cran-x.cbl.ufpe.br/ https://cran.fiocruz.br/ http://cran.fiocruz.br/ https://cps.fiocruz.usp.br/CRAN/ http://cps.fiocruz.usp.br/CRAN/ https://briezer.esalq.usp.br/CRAN/ http://briezer.esalq.usp.br/CRAN/	Computational Biology Center at Universidade Estadual de Santa Cruz Universidade Federal do Paraná Universidade Federal do Paraná Oswaldo Cruz Foundation, Rio de Janeiro Oswaldo Cruz Foundation, Rio de Janeiro University of Sao Paulo, Sao Paulo University of Sao Paulo, Sao Paulo University of Sao Paulo, Piracicaba University of Sao Paulo, Piracicaba
Bulgaria	https://ftp.uni-sofia.bg/CRAN/ http://ftp.uni-sofia.bg/CRAN/	Sofia University Sofia University
Canada	https://mirror.its.sfu.ca/mirror/CRAN/ http://cran.stat.sfu.ca/ https://moss.ca/mirror/cran/ http://moss.ca/mirror/cran/ https://mirror.its.dal.ca/cran/ http://mirror.its.dal.ca/cran/ http://cran.utoronto.ca/	Simon Fraser University, Burnaby Simon Fraser University, Burnaby Manitoba Unix User Group Manitoba Unix User Group Dalhousie University, Halifax Dalhousie University, Halifax University of Toronto
Chile	http://dirichlet.mat.puc.cl/ http://dirichlet.mat.puc.cl/	Pontificia Universidad Católica de Chile, Santiago Pontificia Universidad Católica de Chile, Santiago

Un aperçu de R: *GUI*



```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

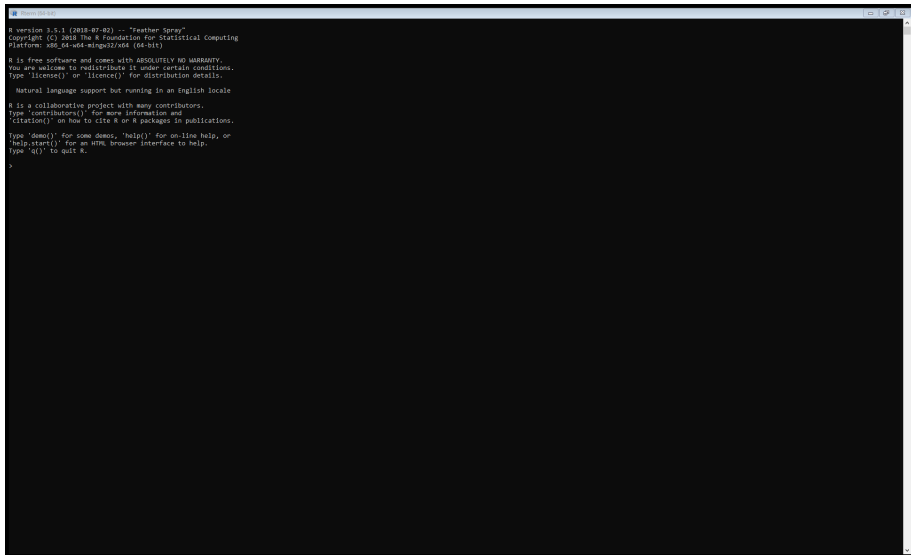
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Un aperçu de R: Terminal



```
R version 3.6.1 (2018-07-02) -- "Teardrop Spring"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

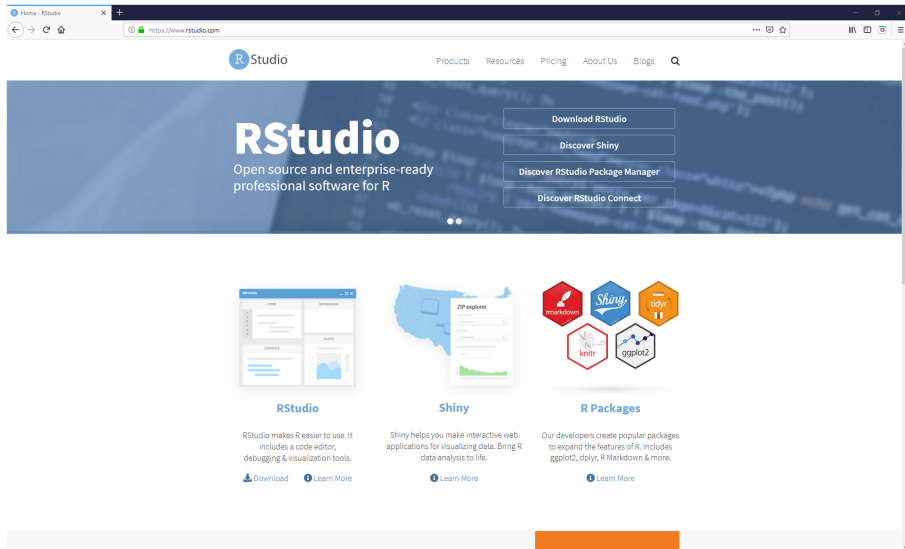
>
```

RStudio

Qu'est-ce que c'est que RStudio

- C'est une IDE (*Integrated Development Environment*) ou Environnement Intégré de Développement
- Il sert d'interface entre R et l'utilisateur, offre à celui diverses commodités d'utilisation

Télécharger RStudio (1)



The screenshot shows the RStudio website homepage in a web browser. The browser's address bar displays "https://www.rstudio.com". The website features a dark blue header with the RStudio logo and navigation links: Products, Resources, Pricing, About Us, Blogs, and a search icon. The main content area has a large blue banner with the text "RStudio" and "Open source and enterprise-ready professional software for R". Below this banner are four buttons: "Download RStudio", "Discover Shiny", "Discover RStudio Package Manager", and "Discover RStudio Connect". Further down, there are three sections: "RStudio" with a screenshot of the IDE, "Shiny" with a map of the United States and a "ZIP explorer" interface, and "R Packages" with icons for markdown, Shiny, tidy, knitr, and ggplot2. Each section includes a brief description and a "Learn More" link.

Home - RStudio
https://www.rstudio.com

RStudio

Products Resources Pricing About Us Blogs

RStudio
Open source and enterprise-ready professional software for R

Download RStudio
Discover Shiny
Discover RStudio Package Manager
Discover RStudio Connect

RStudio
RStudio makes R easier to use. It includes a code editor, debugging & visualization tools.
[Download](#) [Learn More](#)

Shiny
Shiny helps you make interactive web applications for visualizing data. Bring R data analysis to life.
[Learn More](#)

R Packages
Our developers create popular packages to expand the features of R. Includes ggplot2, dplyr, R Markdown & more.
[Learn More](#)

Télécharger RStudio (2)

Download RStudio - RStudio







https://www.rstudio.com/products/rstudio/download/

RStudio

Products Resources Pricing About Us Blogs

Choose Your Version of RStudio

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace. [Learn More about RStudio features.](#)

FREE \$995 per year FREE \$9,995 per year \$29,995 per year

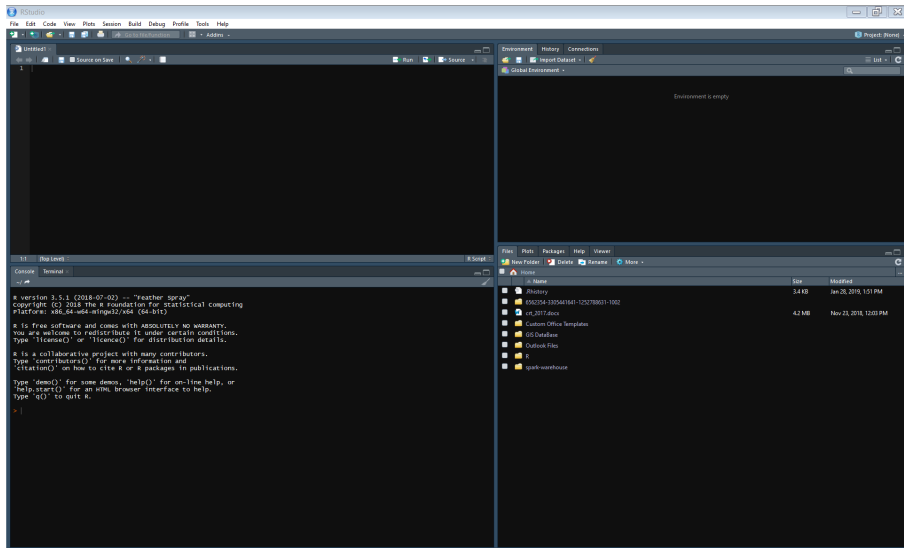
DOWNLOAD BUY DOWNLOAD DOWNLOAD TALK

Learn More Learn More Learn More Learn More Learn More

Integrated Tools for R	●	●	●	●	●
Priority Support		●		●	●
Access via Web Browser			●	●	●
Enterprise Security				●	●
Project Sharing				●	●
Manage Multiple R Sessions & Versions				●	●
Admin Dashboard				●	●
Load Balancing				●	●

Try RStudio Server Pro for free!

Un aperçu de RStudio



Et on démarre

Maintenant, vous avez les outils nécessaires pour commencer la formidable aventure!