

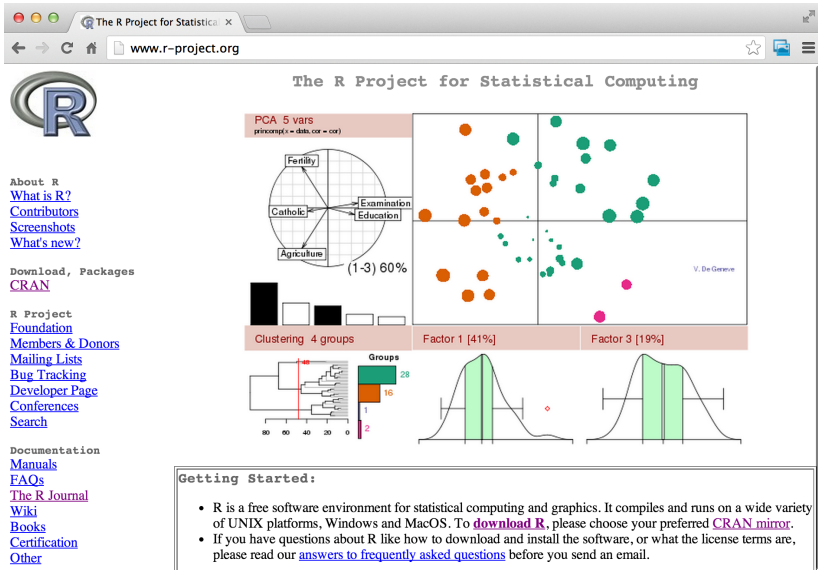
# The Data Scientist's Toolbox

May 17, 2016

# What do data scientists do?

- ▶ Define the question
- ▶ Define the ideal data set
- ▶ Determine what data you can access
- ▶ Obtain the data
- ▶ Clean the data
- ▶ Exploratory data analysis
- ▶ Statistical prediction/modeling
- ▶ Interpret results
- ▶ Challenge results
- ▶ Synthesize/write up results
- ▶ Create reproducible code
- ▶ Distribute results to other people

# The main workhorse of data science



The screenshot shows the homepage of the R Project for Statistical Computing. The browser address bar displays 'www.r-project.org'. The page features the R logo on the left and a navigation menu with links to 'About R', 'What is R?', 'Contributors', 'Screenshots', 'What's new?', 'Download, Packages', 'CRAN', 'R Project', 'Foundation', 'Members & Donors', 'Mailing Lists', 'Bug Tracking', 'Developer Page', 'Conferences', 'Search', 'Documentation', 'Manuals', 'FAQs', 'The R Journal', 'Wiki', 'Books', 'Certification', and 'Other'.

The main content area is titled 'The R Project for Statistical Computing' and displays several statistical plots:

- PCA 5 vars**: A biplot showing the relationship between five variables (Fertility, Examination, Education, Catholic, Agriculture) and the first three principal components. The first three components account for 60% of the variance.
- Clustering 4 groups**: A dendrogram showing the hierarchical clustering of data points into four groups.
- Factor 1 [41%]** and **Factor 3 [19%]**: Two histograms showing the distribution of data points for the first and third principal components, respectively.

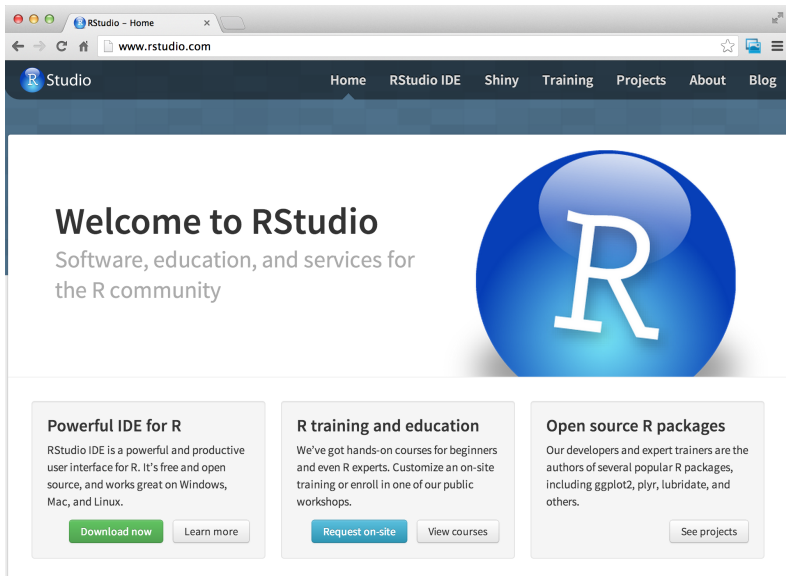
A 'Getting Started:' section is located at the bottom right, providing information about the R software environment and how to download it.

**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. To [download R](#), 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.

<http://www.r-project.org/>

# Where we will work on coding



The image is a screenshot of a web browser displaying the RStudio homepage. The browser's address bar shows 'www.rstudio.com'. The website has a dark blue header with the 'RStudio' logo and navigation links: 'Home', 'RStudio IDE', 'Shiny', 'Training', 'Projects', 'About', and 'Blog'. The main content area features a large blue circle with a white 'R' logo. Below this, there are three columns of text and buttons. The first column, 'Powerful IDE for R', describes the RStudio IDE as a powerful and productive user interface for R, free and open source, and works on Windows, Mac, and Linux. It includes a green 'Download now' button and a grey 'Learn more' button. The second column, 'R training and education', mentions hands-on courses for beginners and experts, and offers on-site training or public workshops. It includes a blue 'Request on-site' button and a grey 'View courses' button. The third column, 'Open source R packages', states that developers and expert trainers are the authors of several popular R packages, including ggplot2, plyr, lubridate, and others. It includes a grey 'See projects' button.

RStudio – Home  
www.rstudio.com

RStudio Home RStudio IDE Shiny Training Projects About Blog

## Welcome to RStudio

Software, education, and services for the R community

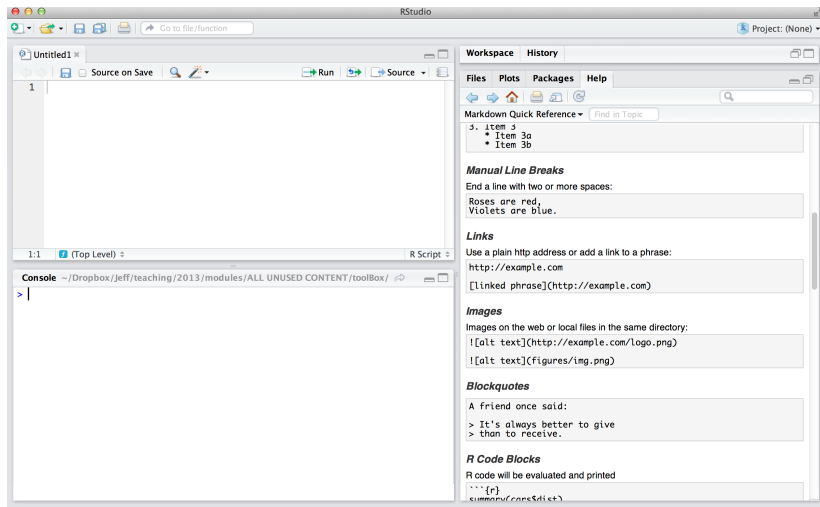
**Powerful IDE for R**  
RStudio IDE is a powerful and productive user interface for R. It's free and open source, and works great on Windows, Mac, and Linux.  
[Download now](#) [Learn more](#)

**R training and education**  
We've got hands-on courses for beginners and even R experts. Customize an on-site training or enroll in one of our public workshops.  
[Request on-site](#) [View courses](#)

**Open source R packages**  
Our developers and expert trainers are the authors of several popular R packages, including ggplot2, plyr, lubridate, and others.  
[See projects](#)

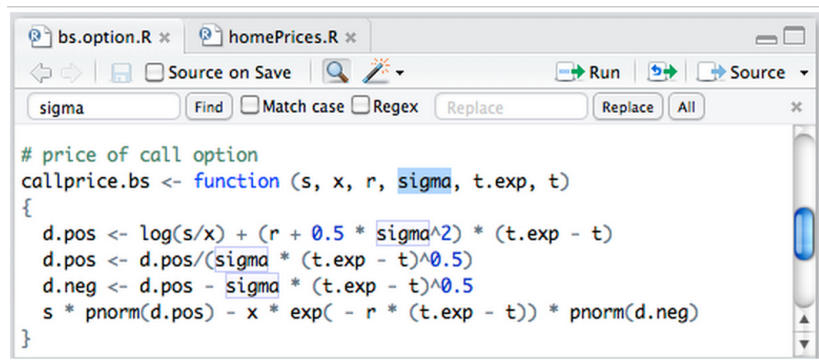
`http://www.rstudio.com/`

# Rstudio's interface



<http://www.rstudio.com/>

## Primary file types - R script

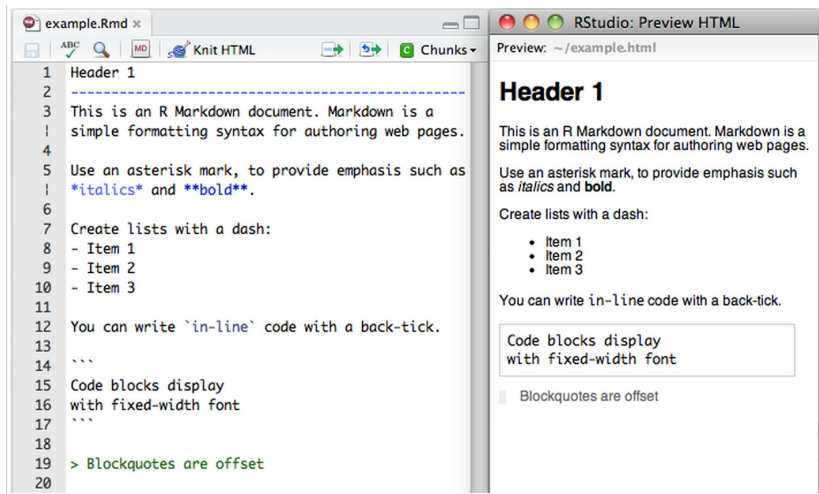


The screenshot shows an R script editor window with two tabs: 'bs.option.R' and 'homePrices.R'. The 'homePrices.R' tab is active. The editor has a toolbar with icons for navigation, saving, and running. Below the toolbar is a search bar with the text 'sigma' and buttons for 'Find', 'Match case', 'Regex', 'Replace', and 'All'. The main text area contains the following R code:

```
# price of call option
callprice.bs <- function(s, x, r, sigma, t.exp, t)
{
  d.pos <- log(s/x) + (r + 0.5 * sigma^2) * (t.exp - t)
  d.pos <- d.pos/(sigma * (t.exp - t)^0.5)
  d.neg <- d.pos - sigma * (t.exp - t)^0.5
  s * pnorm(d.pos) - x * exp(- r * (t.exp - t)) * pnorm(d.neg)
}
```

<http://www.rstudio.com/ide/docs/using/source>

# Primary file types - R markdown document



The screenshot displays the RStudio interface with two main panes. The left pane shows the source R Markdown file 'example.Rmd' with line numbers 1 through 20. The right pane shows the 'Preview HTML' of the document.

**Source File (example.Rmd):**

```
1 Header 1
2 -----
3 This is an R Markdown document. Markdown is a
4 simple formatting syntax for authoring web pages.
5 Use an asterisk mark, to provide emphasis such as
6 italics and bold.
7 Create lists with a dash:
8 - Item 1
9 - Item 2
10 - Item 3
11
12 You can write `in-line` code with a back-tick.
13
14 ```
15 Code blocks display
16 with fixed-width font
17 ```
18
19 > Blockquotes are offset
20
```

**Preview HTML:**

## Header 1

This is an R Markdown document. Markdown is a simple formatting syntax for authoring web pages.

Use an asterisk mark, to provide emphasis such as *italics* and **bold**.

Create lists with a dash:

- Item 1
- Item 2
- Item 3

You can write in-line code with a back-tick.

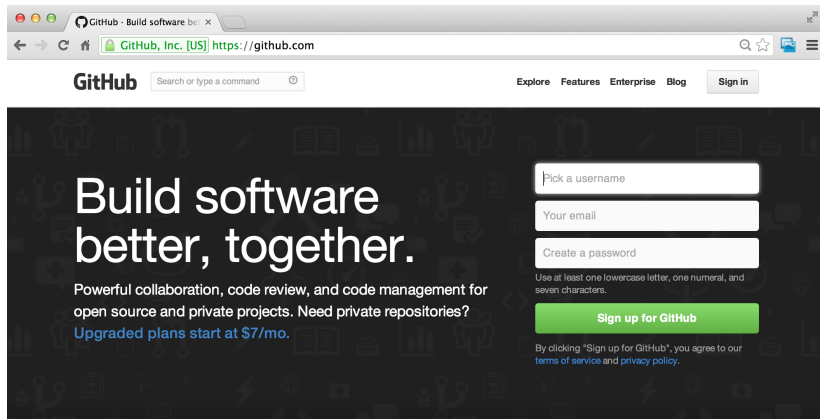
```
Code blocks display
with fixed-width font
```

Blockquotes are offset

http:

[//www.rstudio.com/ide/docs/authoring/using\\_markdown](http://www.rstudio.com/ide/docs/authoring/using_markdown)

# Sharing your results - Github & Git



The screenshot shows the GitHub homepage in a web browser. The browser's address bar displays "https://github.com". The GitHub logo is in the top left, followed by a search bar with the placeholder text "Search or type a command". To the right of the search bar are links for "Explore", "Features", "Enterprise", and "Blog", and a "Sign in" button. The main content area has a dark background with the text "Build software better, together." in large white letters. Below this, it says "Powerful collaboration, code review, and code management for open source and private projects. Need private repositories? Upgraded plans start at \$7/mo." To the right of this text is a sign-up form with three input fields: "Pick a username", "Your email", and "Create a password". Below the "Create a password" field is a note: "Use at least one lowercase letter, one numeral, and seven characters." Below the form is a green "Sign up for GitHub" button. At the bottom of the sign-up section, it says "By clicking 'Sign up for GitHub', you agree to our terms of service and privacy policy."

Build software better, together.

Powerful collaboration, code review, and code management for open source and private projects. Need private repositories? Upgraded plans start at \$7/mo.

Pick a username

Your email

Create a password

Use at least one lowercase letter, one numeral, and seven characters.

Sign up for GitHub

By clicking "Sign up for GitHub", you agree to our terms of service and privacy policy.

## Why you'll love GitHub.

Powerful features to make software development more collaborative.



# Where to run Github commands - the shell

