



# The Data Scientist's Toolbox

Johns Hopkins Bloomberg School of Public Health

```
## Error in eval(expr, envir, enclos): object 'opts_chunk' not found
```

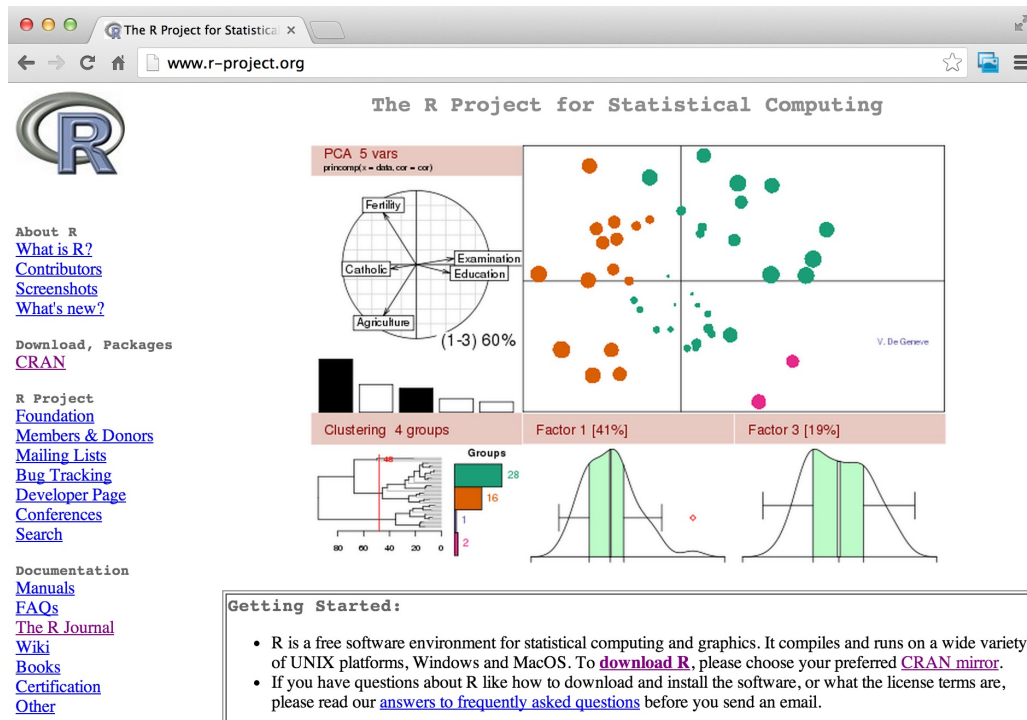
```
## Error in eval(expr, envir, enclos): object 'knit_hooks' not found
```

```
## Error in eval(expr, envir, enclos): object 'knit_hooks' not found
```

# 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

# The main workhorse of data science



The screenshot shows the R Project for Statistical Computing website. The browser address bar displays 'www.r-project.org'. The page features the R logo on the left, a navigation menu with links like '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', and 'Documentation' (with sub-links for 'Manuals', 'FAQs', 'The R Journal', 'Wiki', 'Books', 'Certification', and 'Other'). The main content area is titled 'The R Project for Statistical Computing' and contains several data visualization examples: a PCA plot of 5 variables (Fertility, Examination, Education, Catholic, Agriculture) with a loading plot and a scatter plot; a clustering dendrogram with 4 groups; and two histograms showing the distribution of Factor 1 (41%) and Factor 3 (19%).

**PCA 5 vars**  
`princomp(x = data, cor = cor)`

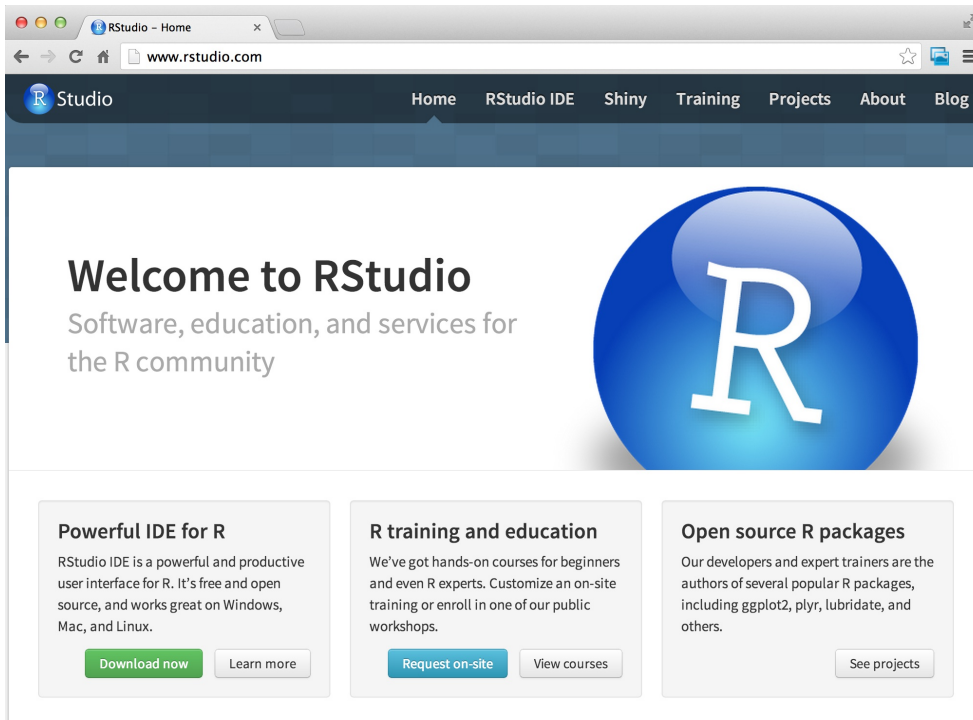
Factor 1 [41%]      Factor 3 [19%]

**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.

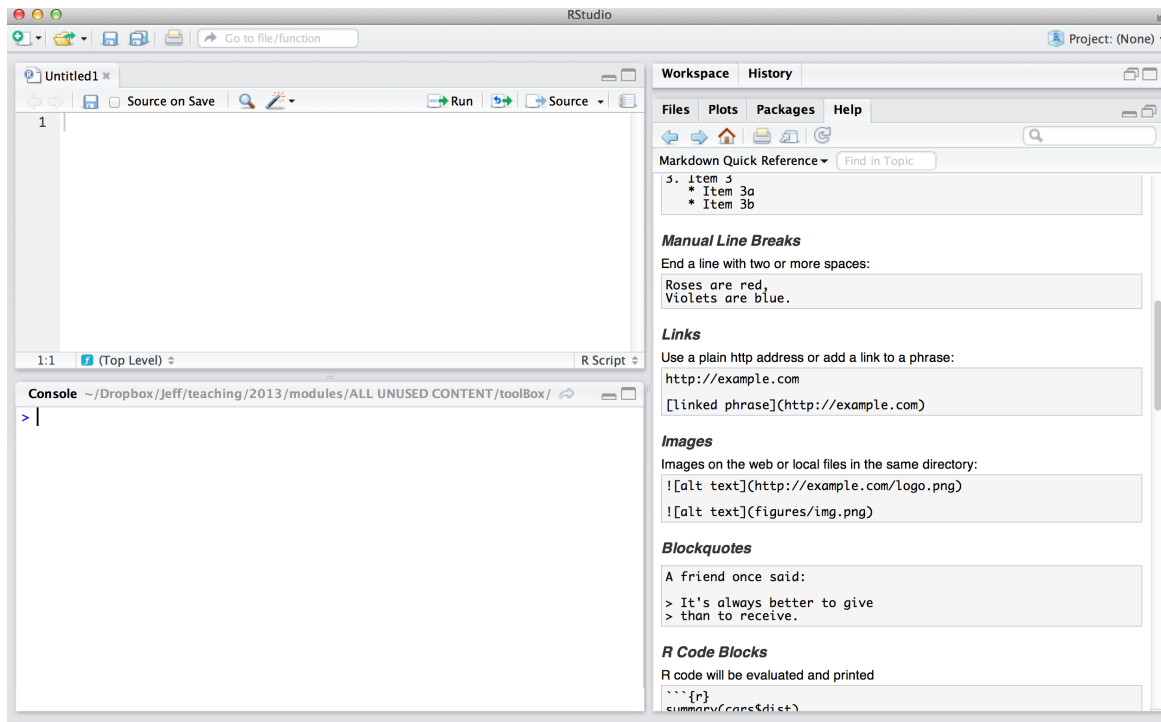
<http://www.r-project.org/>

# Where we will work on coding



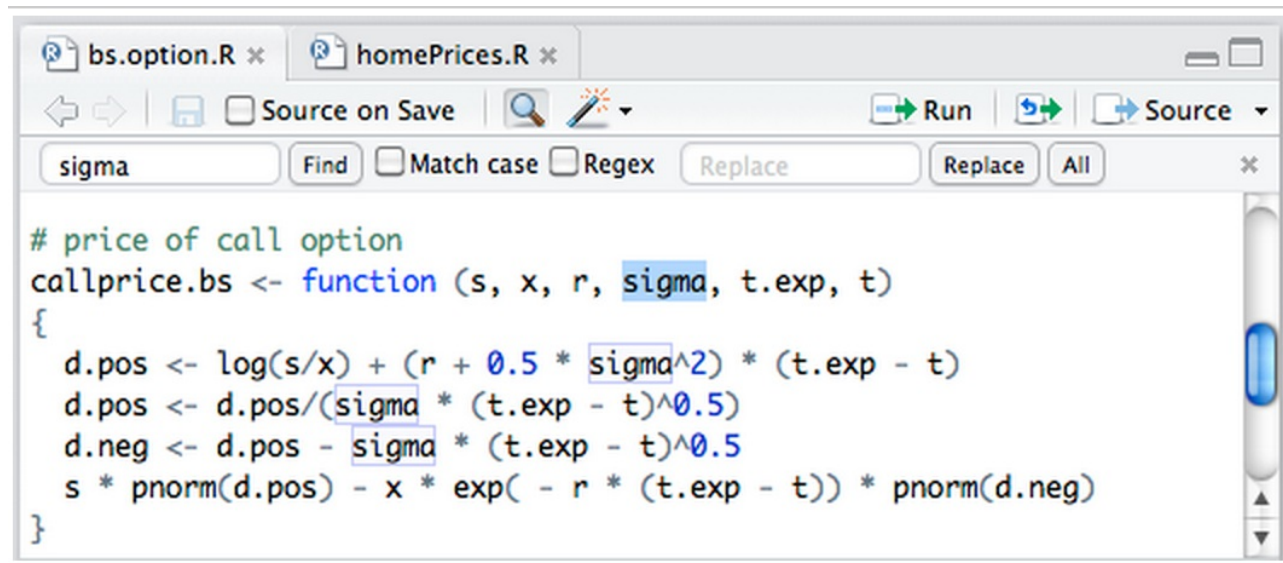
<http://www.rstudio.com/>

# Rstudio's interface



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

# Primary file types - R script



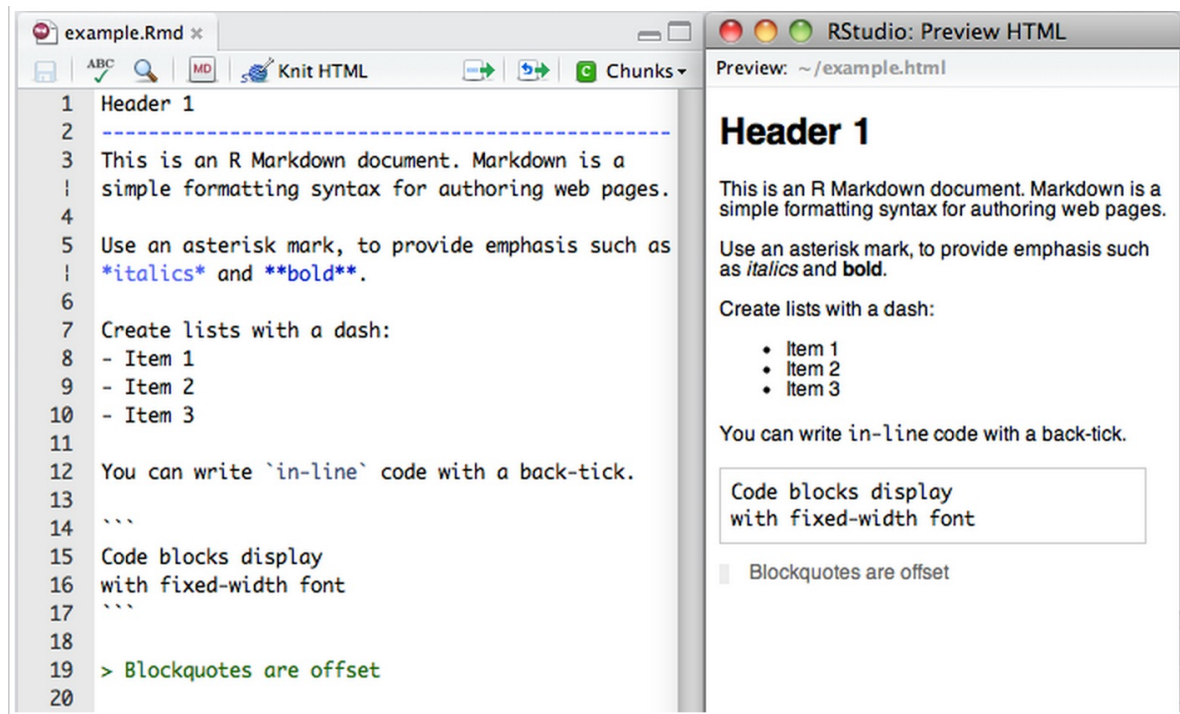
The screenshot shows the RStudio Source editor with two tabs: 'bs.option.R' and 'homePrices.R'. The 'bs.option.R' tab is active. The editor contains the following R code:

```
# 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)
}
```

The code defines a function `callprice.bs` that calculates the price of a call option. The parameters are `s` (stock price), `x` (strike price), `r` (risk-free rate), `sigma` (volatility), `t.exp` (expiration time), and `t` (current time). The function uses the Black-Scholes formula to calculate the option price.

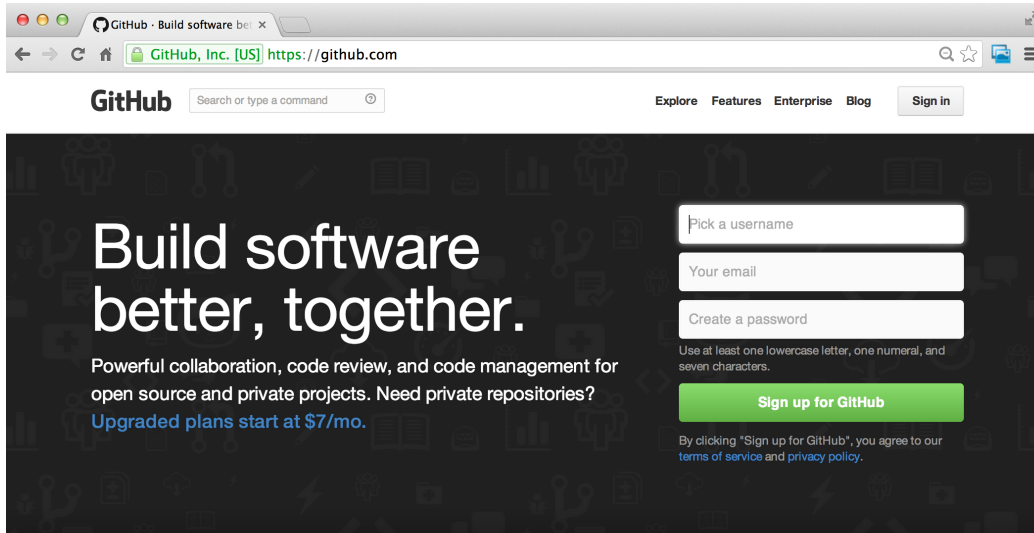
<http://www.rstudio.com/ide/docs/using/source>

# Primary file types - R markdown document



[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 on the left, and a search bar with the placeholder "Search or type a command" is next to it. On the right, there are links for "Explore", "Features", "Enterprise", "Blog", and a "Sign in" button. The main content area has a dark background with the text "Build software better, together." and a description of GitHub's features. 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 these fields is a green "Sign up for GitHub" button. At the bottom of the form, there is a line of text stating that by clicking "Sign up for GitHub", the user agrees to the terms of service and privacy policy.

GitHub

Search or type a command

Explore Features Enterprise Blog Sign in

## 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

