



The Data Scientist's Toolbox

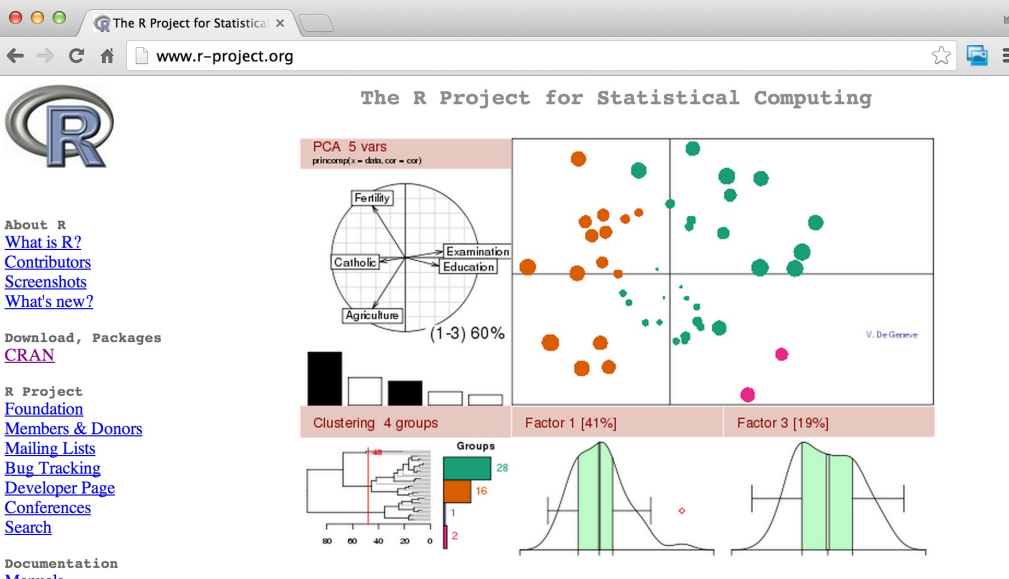
Johns Hopkins Bloomberg School of Public Health

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 R Project for Statistical Computing



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](#)
[Other](#)

PCA 5 vars
`prcomp(x = data, cor = cor)`

Clustering 4 groups

Factor 1 [41%]

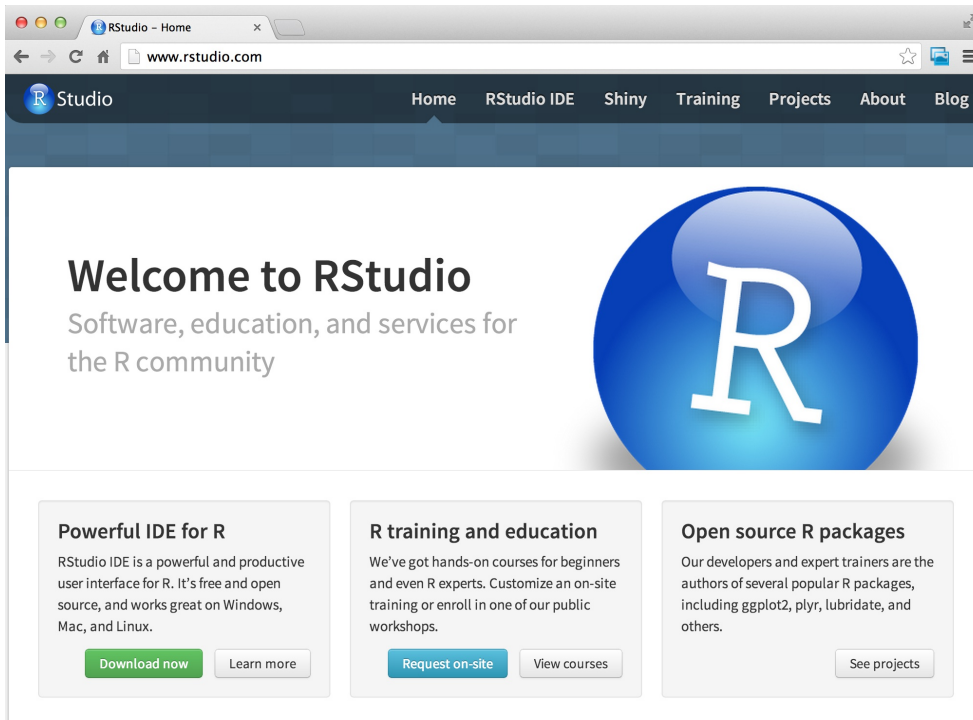
Factor 3 [19%]

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.

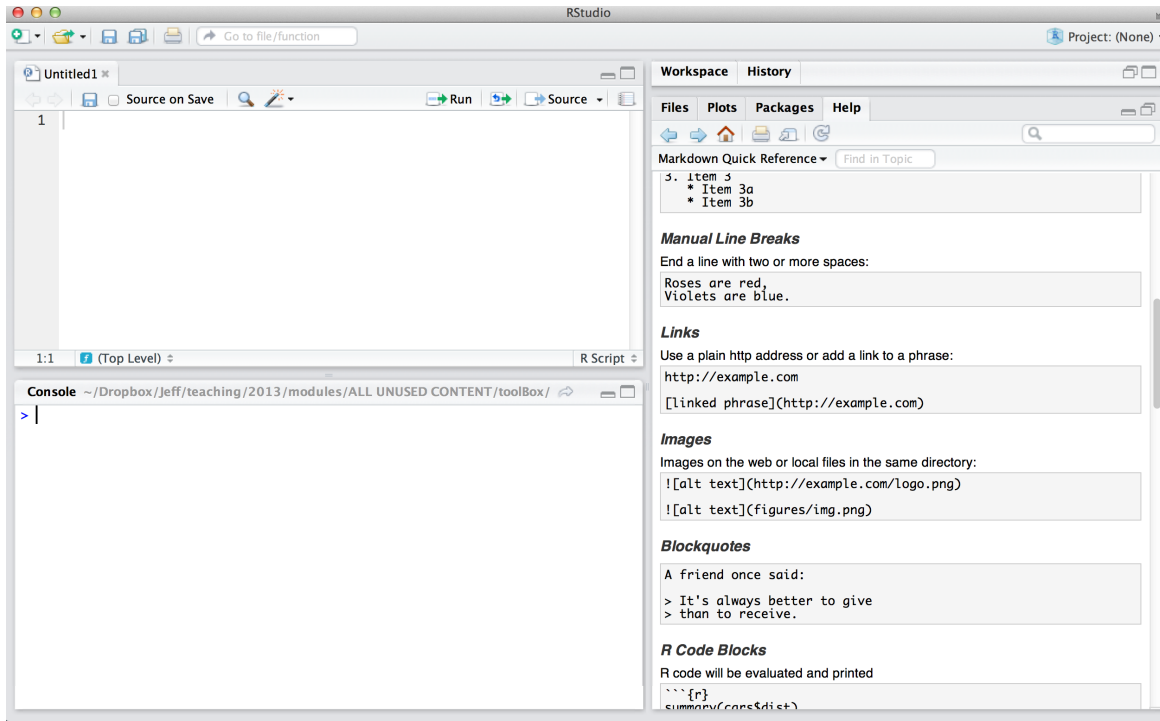
<http://www.r-project.org/>

Where we will work on coding



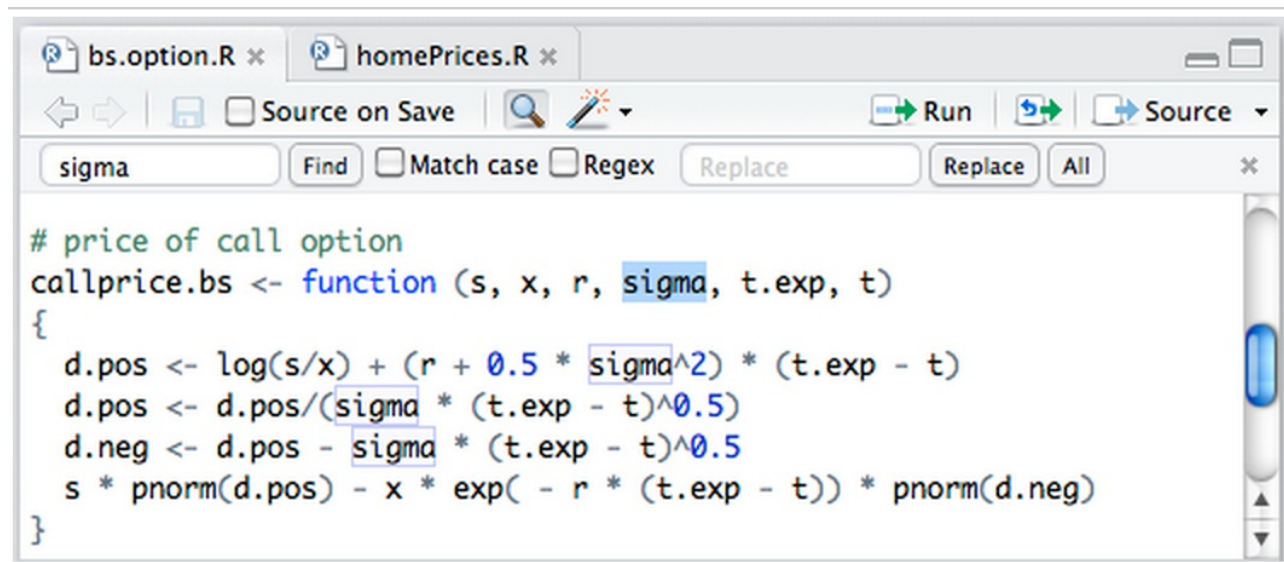
<http://www.rstudio.com/>

Rstudio's interface



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

Primary file types - R script



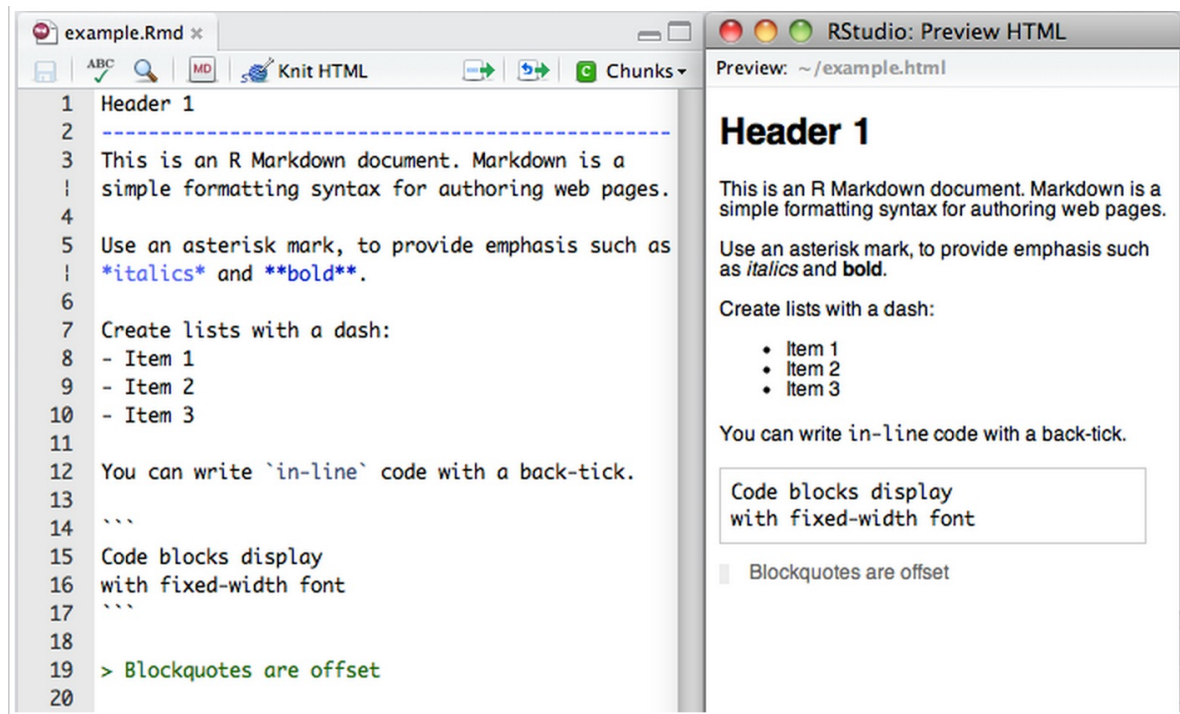
The screenshot shows the RStudio Source editor with two tabs: 'bs.option.R' and 'homePrices.R'. The 'bs.option.R' tab is active. The editor contains the following R code:

```
# 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.

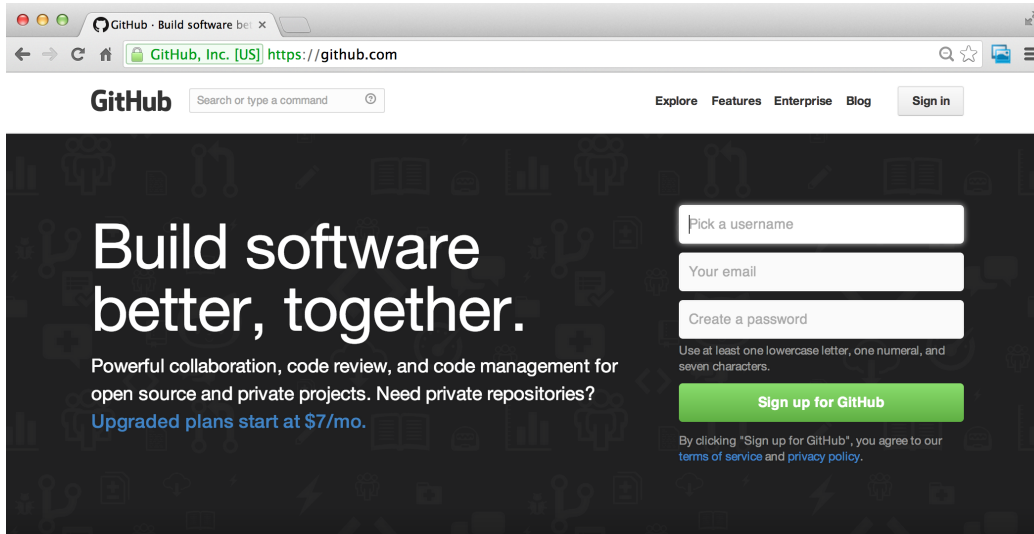
<http://www.rstudio.com/ide/docs/using/source>

Primary file types - R markdown document



http://www.rstudio.com/ide/docs/authoring/using_markdown

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 navigation links for "Explore", "Features", "Enterprise", "Blog", and "Sign in" are on the right. The main content area has a dark background with the text "Build software better, together." and a description of GitHub's features. A sign-up form is on the right, including fields for "Pick a username", "Your email", and "Create a password", followed by a green "Sign up for GitHub" button. Below the button is a link to the "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

