R入门

初识R

魏振军

2015年

R是什么？

R是一个开放的交互式数据分析和可视化平台，是一个统计编程环境，是一种用于探索、展示和理解数据的语言。

它已成为数据科学领域的的全球通用语言。它提供各种用于分析和理解数据的方法，从最基础的到最前沿的，无所不包。

R语言的发展

R是受S语言和Scheme语言影响发展而来

最初由新西兰奥克兰大学统计系的Ross Ihaka和Robert Gentleman合作编写

Ross Ihaka

Robert Gentleman

R语言的发展

1995年6月R语言的源代码正式发布到自由软件协会的FTP上.

随着R语言的进一步开发，程序版本的归档又成了一个问题. 维也纳工业大学的Kurt Hornik承担了这个任务，在维也纳建立了R程序的归档，使程序版本的发布更加规范. 同时世界各地也出现了R程序的镜像.

R语言的发展

1997年中期R核心团队正式成立， 包含11个早期成员. 现在R语言版本依然还是由“R开发核心团队”负责开发. 截止到2013年，R核心团队已经达到20人.

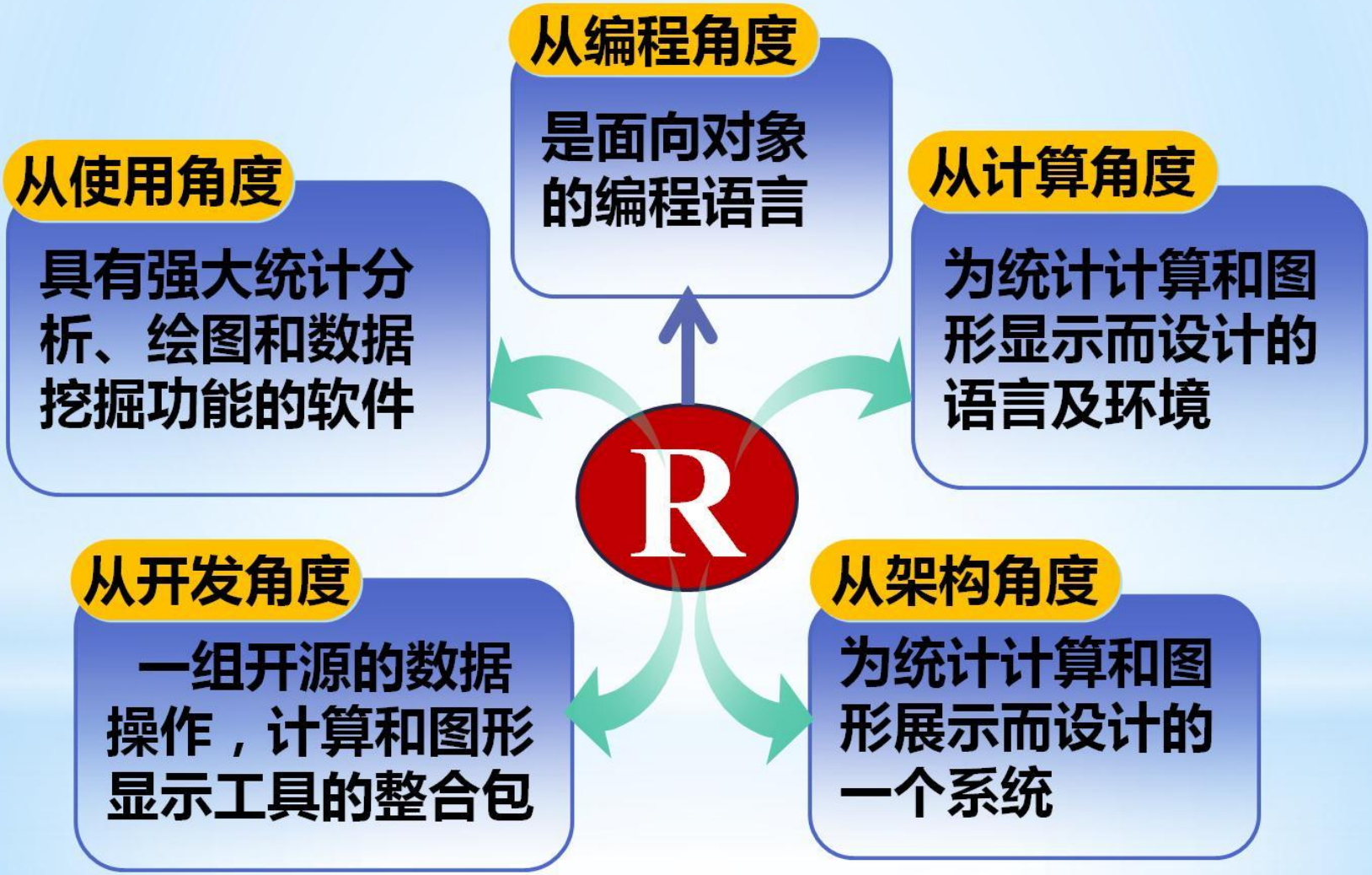
由于R语言自身扩展性非常强, 随着发展和使用人数增多,也吸引了大量用户编写的自定义的程序包供更多人使用 .这些包可以从世界各地的CRAM镜像上下载.

R语言的发展

2010年，美国统计协会将第一届“统计计算及图形奖”授予R语言，用于表彰其在统计应用和统计研究广泛的影响.

目前R最新版本是3.2.1(2015年6月18日发布）.

R是什么？


从编程角度
是面向对象的编程语言
从使用角度
具有强大统计分析、绘图和数据挖掘功能的软件
从计算角度
为统计计算和图形显示而设计的语言及环境
R
从开发角度
一组开源的数据操作，计算和图形显示工具的整合包
从架构角度
为统计计算和图形展示而设计的一个系统

为什么要使用R?

功能强大

开源、免费、跨平台

可编程　　不断更新

良好的扩展性　完备的帮助系统

丰富的网络资源

强大的社区支持

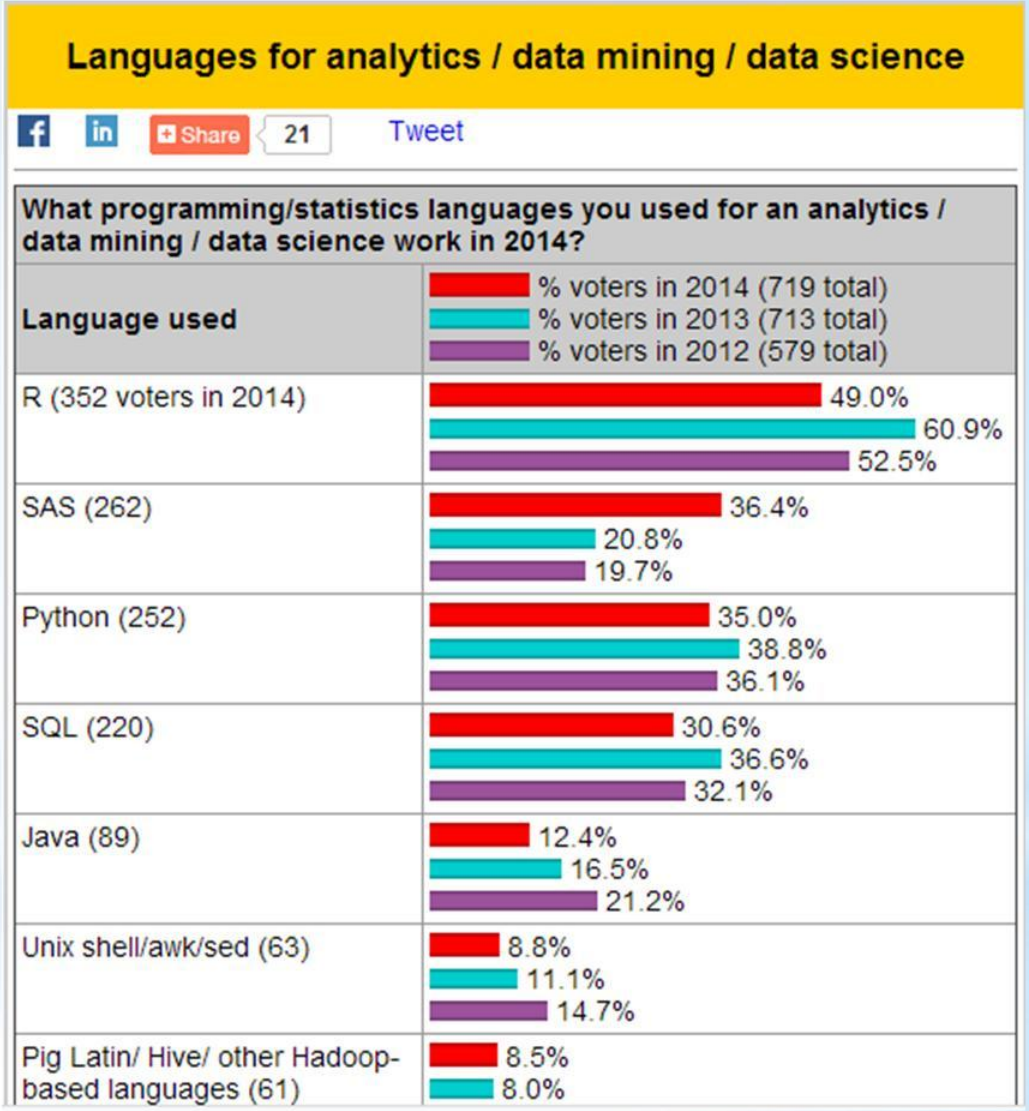
R的使用者

R语言最早是学术界统计学家在用。许多志愿者坚持不懈推动着R语言的发展. 众多统计学者或相关领域的程序员也纷纷贡献自己的力量，将大量统计方法以附加包的形式发布出来，使用户可以以最快的速度用上最新的统计方法.

近些年来，由互联网引发的大数据革命，也让工业界的人，开始认识R，加入R

R语言是数据分析领域的通用语言

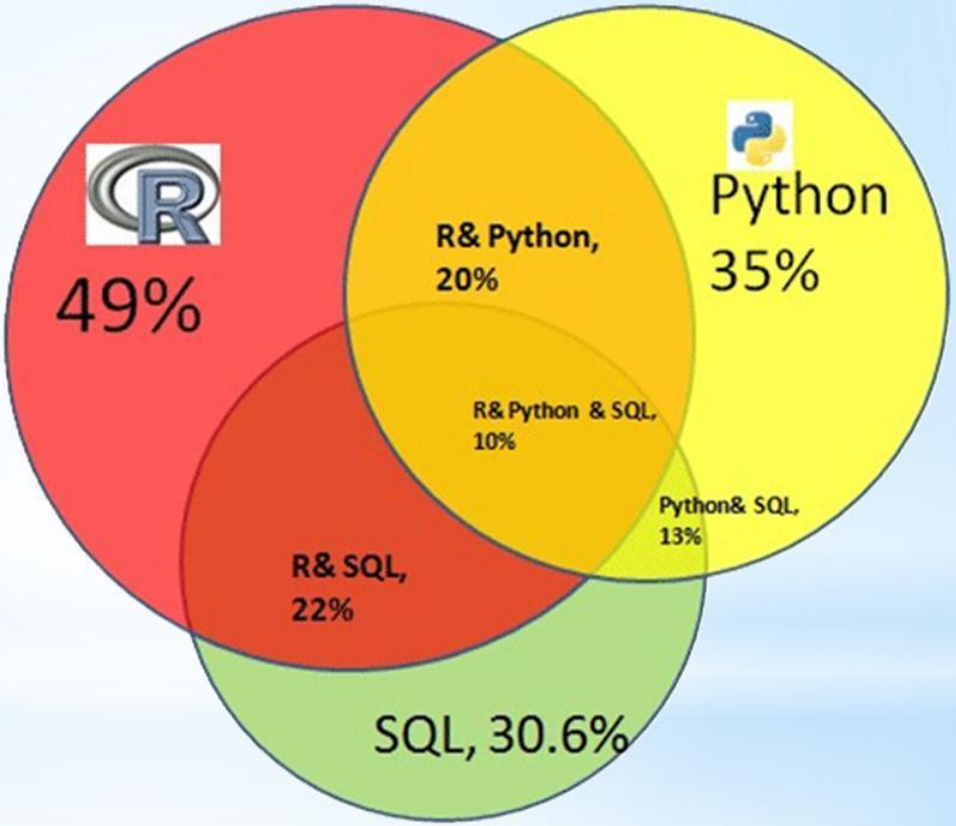
根据kdnuggets的调查显示，在2012~2014连续三年，R语言都是分析/数据挖掘/数据科学领域排名第一的主流语言和工具



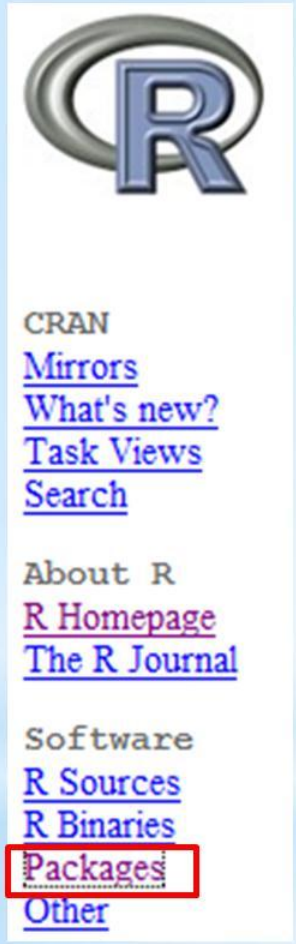
2014年在数据科学/数据挖掘使用排名前4的语言

KDnuggets 2014 Poll:Languages used for Analytics/Data Mining

KDnuggets调查：2014年在分析/数据挖掘中使用的四个主要语言的优势分析：R,SAS,Python,和使用SQL



R不断增加的程序包



R语言最强大的是它的统计分析功能，专业的R语言代码。世界各国统计学家、统计工作者不少都把R作为展示统计方法的首选平台。

截止到2015年8月7日

Contributed Packages

Available Packages

贡献的包多达6973个

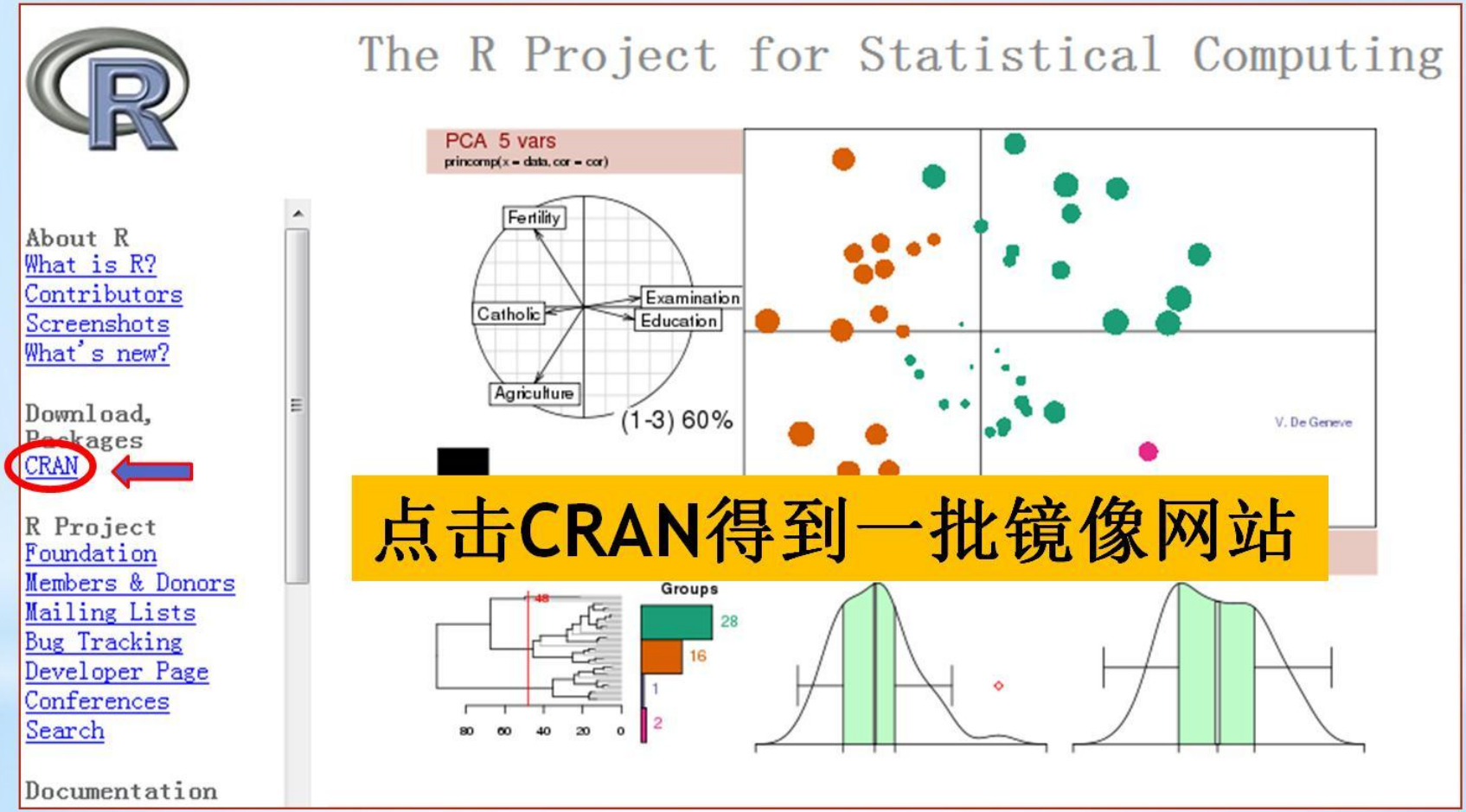
Currently, the CRAN package repository features 6973 available packages.

Table of available packages, sorted by date of publication

Table of available packages, sorted by name

可用的包列表

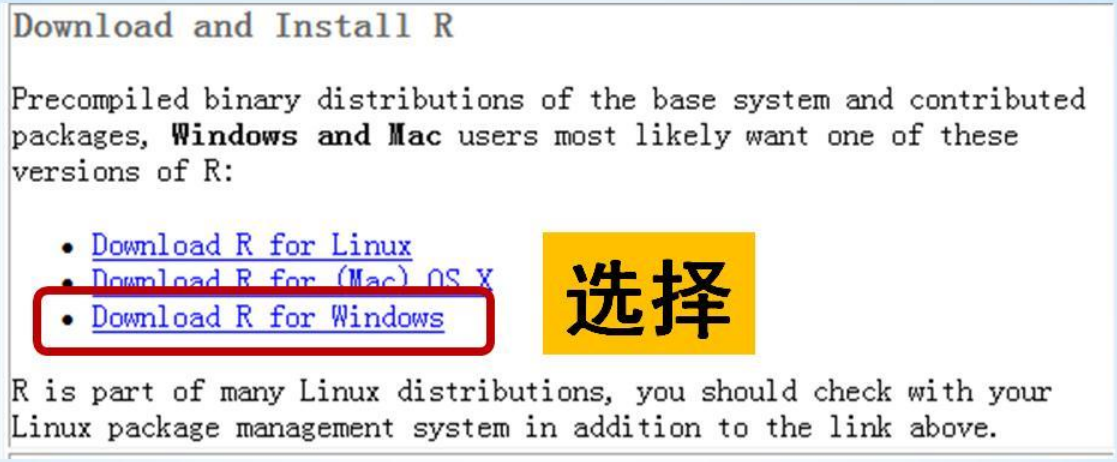
R的下载和安装

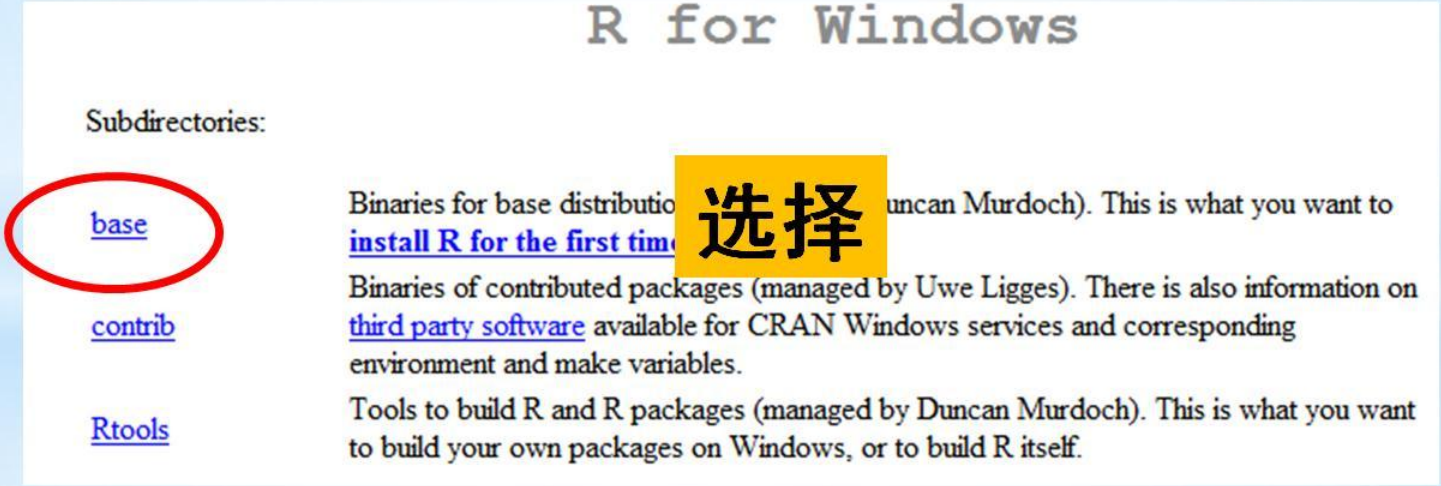

The R Project for Statistical Computing
About R
What is R?
Contributors
Screenshots
What's new?
Download,
Packages
CRAN
R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search
Documentation
PCA 5 vars
princomp(x = data, cor = cor)
Fertility
Catholic
Examination
Education
Agriculture
(1-3) 60%
V. De Geneve
点击CRAN得到一批镜像网站
Groups
28
16
1
2
80
60
40
20
0

R的下载和安装


About R
What is R?
Contributors
Screenshots
What's new?
Download,
Packages
CRAN
Chile
http://dirichlet.mat.puc.cl/
Pontificia Universidad Cato
Santiago
China
http://ftp.ctex.org/mirrors/CRAN/
CTEX.ORG
http://mirror.bjtu.edu.cn/cran
Beijing Jiaotong University,
http://mirrors.ustc.edu.cn/CRAN/
University of Science and T
China
http://mirrors.xmu.edu.cn/CRAN/
Xiamen University
Colombia
http://www.laqee.unal.edu.co/CRAN/
National University of Colo
http://www.icesi.edu.co/CRAN/
Icesi University
Denmark
dotsrc.org, Aalborg
点击所选镜像网站

R的下载和安装


Download and Install R
Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:
Download R for Linux
Download R for (Mac) OS X
Download R for Windows
选择
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

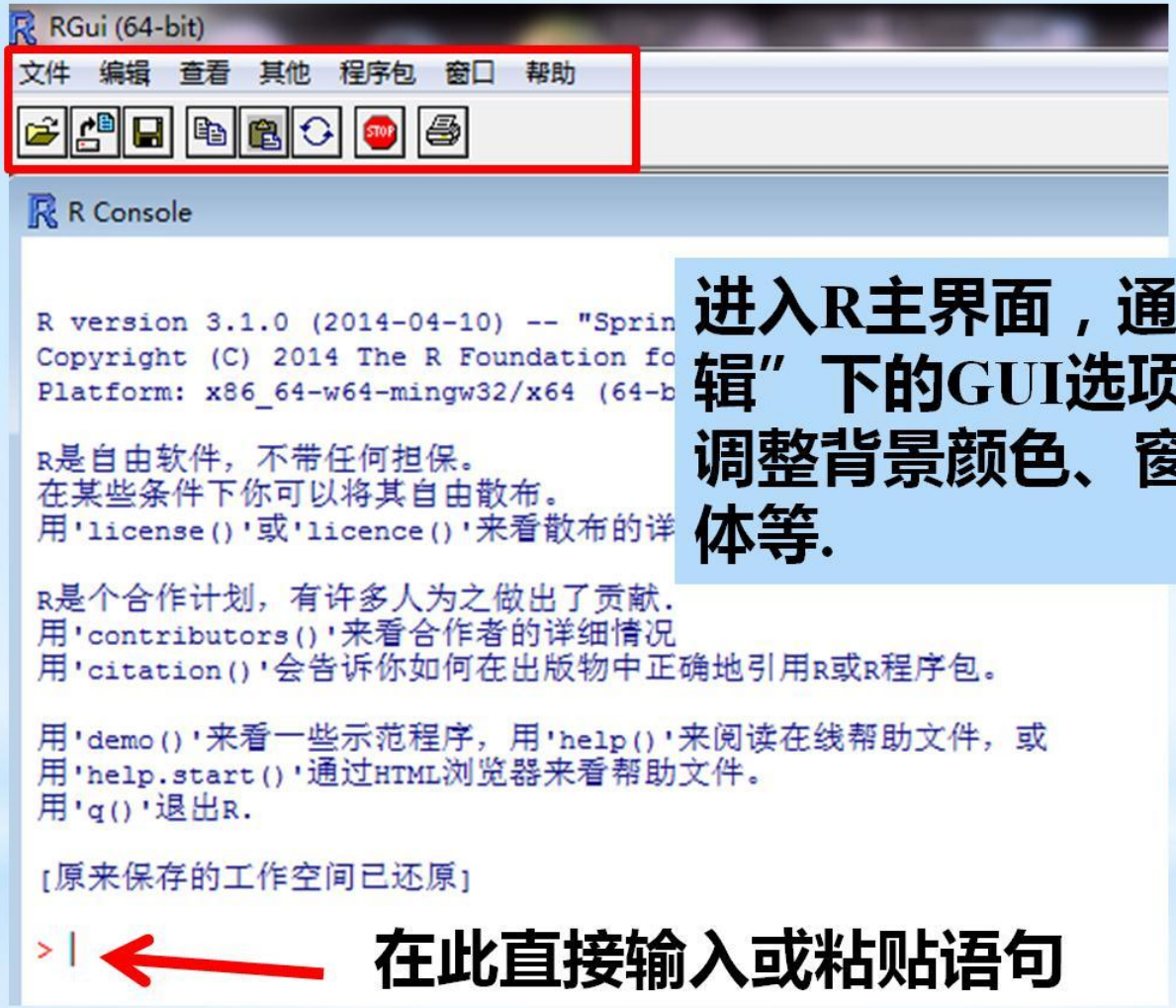

R for Windows
Subdirectories:
base
选择
Binaries for base distributio
uncan Murdoch). This is what you want to
install R for the first tim
contrib
Binaries of contributed packages (managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
Rtools
Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows, or to build R itself.

R的下载和安装

选择这个,下载安装文件

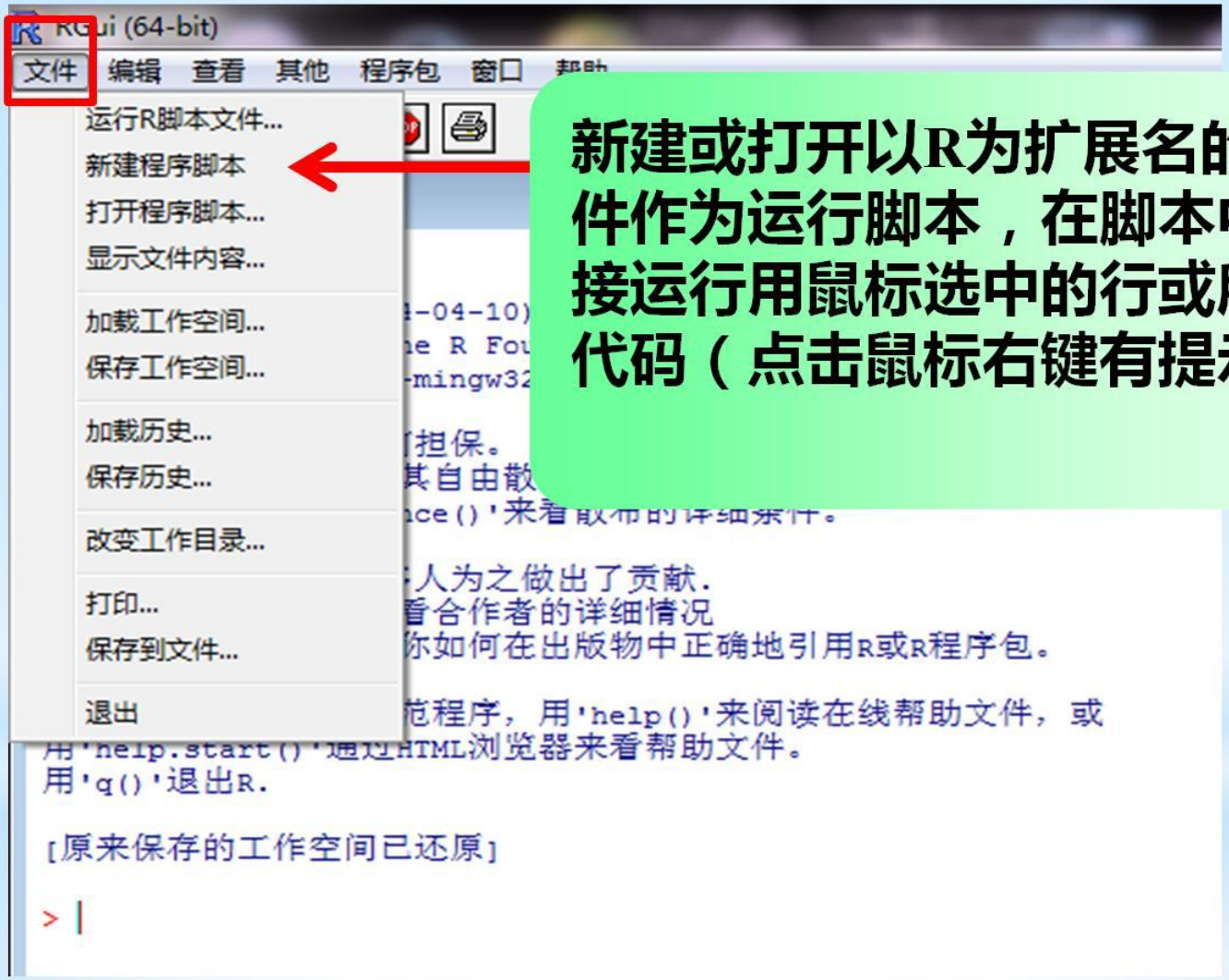
下载的R软件中包含几个标准（基本）R包，其它R包（扩展包）可以在需要时下载.

R的基本操作



进入R主界面，通过“编辑”下的GUI选项，可以调整背景颜色、窗口字体等.

R的基本操作



新建或打开以R为扩展名的文件作为运行脚本，在脚本中直接运行用鼠标选中的行或所选代码（点击鼠标右键有提示）.

R的基本操作

几点提醒

- 所有代码中的标点符号都必须使用半角格式
- R代码对大小写敏感
- 每一行可输入多个语句，之间用“;”分割
- 一行中，从”#”开始到句子收尾之间的语句是注释
- 可以用向上光标键来找回以前运行的命令再次运行或修改后再运行

R 当前的工作目录

在用R工作时，首先要设定工作目录.

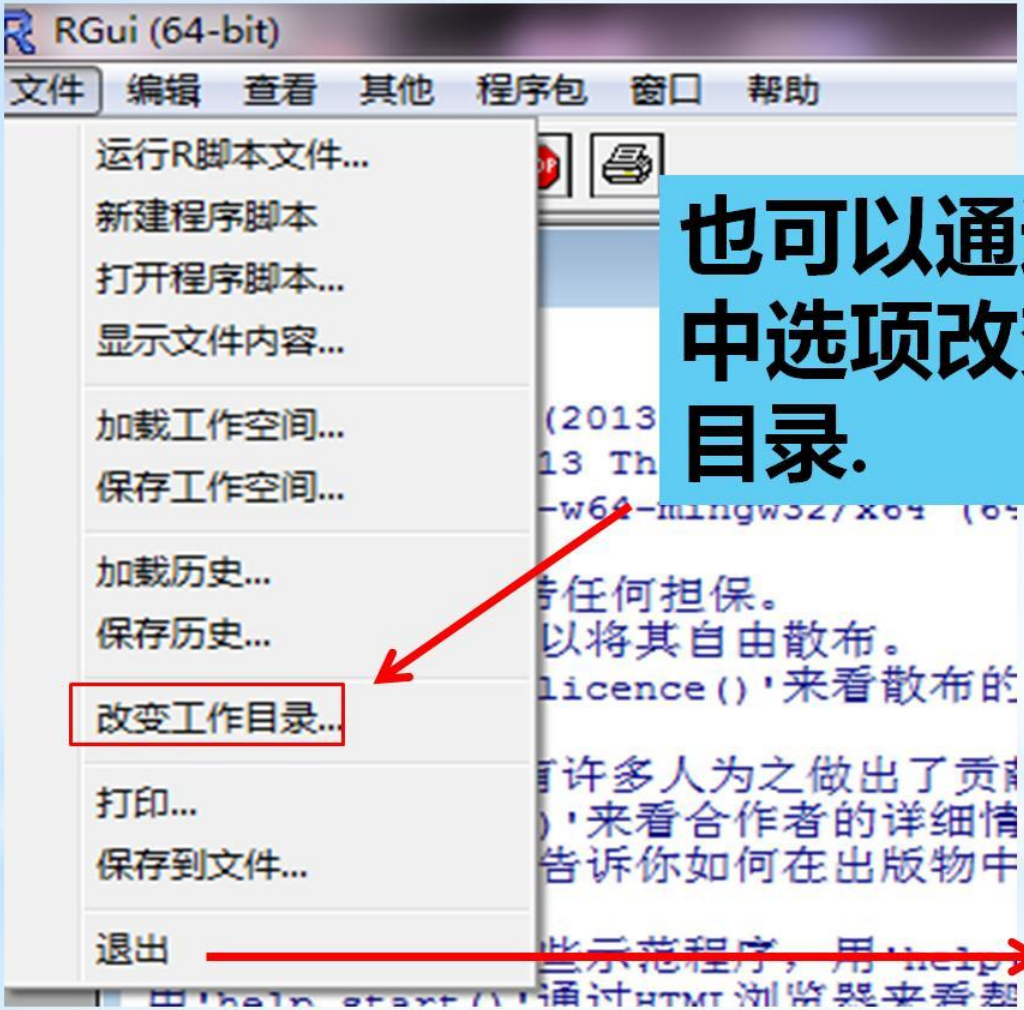
调用工作目录中的数据不必写路径.

使用getwd( ) 命令获得R 的工作目录
使用setwd( ) 设置当前工作目录位置

如：setwd(“d:/2015stat”)

或　setwd(“d:\\2015stat”)

R 当前的工作目录



也可以通过菜单中选项改变工作目录.

q()

查看R自带的图形演示

使用demo( ) 列出R中可用演示

我们来看其中的三个图形演示

使用指令：

```
demo(graphics)
```

```
demo(persp)
```

```
demo(image)
```

R包(Packages)

R包是R函数、数据、预编译代码以一种定义完善的格式组成的集合. 有详细的说明和示例. Window下的R包是经过编译的zip包.

R包的安装和使用

R语言的使用，很大程度上是借助各种各样的包(Packages)的辅助。这些包提供了横跨多个领域、数量惊人的新功能.

R包使R的功能不断扩展，特定的分析功能，需要用相应的包实现.

计算机上存储包的目录称为库（library）

R包的安装和使用

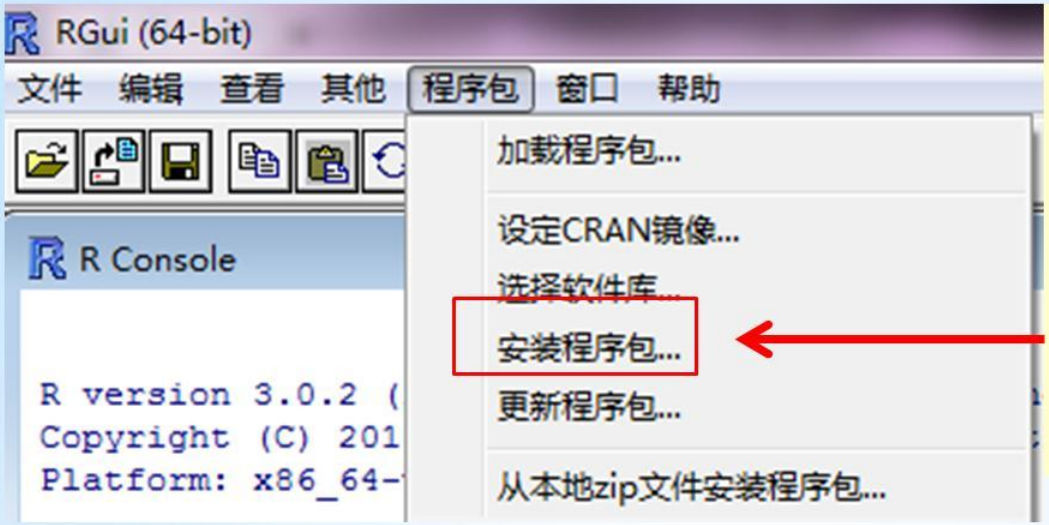
使用 .libPaths( ) 显示库所在的位置

使用 library( ) 显示库中有哪些包

对每个包有一个简要说明

除R自带的包外，需要使用的其他包可通过下载来进行安装.

R包的安装和使用



如果计算机联网，通过选择下载镜像，下载安装各种R包

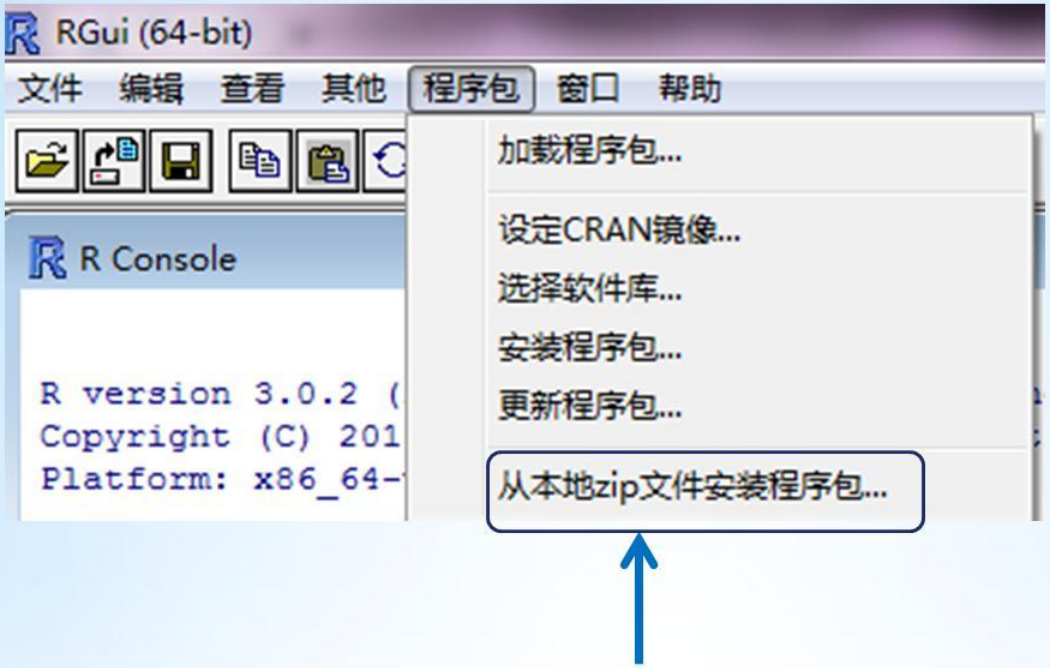
或使用指令：install.packages(“包名")

联网直接安装需要的程序包.

例如：安装随机森林包

install.packages(“randomForest”)

R包的安装和使用



对于不联网的机器，可以事先从镜像网站下载可用的R包，存放在本机上，使用时再安装.

R包的安装和使用

使用 installed.packages( )

列出安装的包，以及它们的版本号、依赖关系等信息

使用 .packages (all.available=TRUE)

获得本地安装的包列表

R包的安装和使用

加载包

包安装后，要使用包中的函数。必须先把包加载到内存中（R启动后默认加载基础包）

加载包命令：　library(包名)

require(包名)

例如：把随机森林包加载到内存中

```
library(randomForest)
```

查看当前环境哪些R包已加载

search( )　　仅告知包名

find.package( )

path.package( )　　同时告知包所在库路径

注意与 library( ) 的区别

查看库中有哪些包
不论是否加载

R启动后内存中自动加载的包

R 初始状态载入包（基础包）列表

包	描述
stats	常用统计函数
graphics	基础绘图函数
grDevices	基础或grid 图形设备
utils R	工具函数
datasets	基础数据集
methods	用于R对象和编程工具的方法和类的定义
base	基础函数

查看某个包的主要内容

help(package="包名")

例如： help(package="MASS")

Documentation for package 'MASS' version 7.3-26

- DESCRIPTION file.
- Package NEWS.

数据包的简介

Help Pages

A B C D E F G H I K L M N O P Q R S T U V W misc

-- A --

abbey	Determinations of Nickel Content
accdeaths	Accidental Deaths in the US 1973-1978

函数和数据集的索引

查看某个包的主要内容

```
library(help=包名)
```

主要内容包括：例如：包名、作者、版本、更新时间、功能描述、开源协议、存储位置、主要的函数和数据集

例如：查看“随机森林包”的帮助

```
library(help=randomForest)
```

注意：这里包名无引号

某个包一经载入，我们就可以使用该包中一系列新的函数和数据集

更新已安装的包

有的包出了新版本，增加了新功能
在计算机联网的情况下，使用

```
update.packages( )
```

通过选择下载镜像，下载安装该包的最新版本。

查看某个包中的数据文件

data(package="包名")

例如：查看MASS包中的数据文件

data(package="MASS")

```
Data sets in package 'MASS':

Aids2                   Australian AIDS Survival Data
Animals                 Brain and Body Weights for 28 Species
Boston                  Housing Values in Suburbs of Boston
Cars93                  Data from 93 Cars on Sale in the USA in 1993
Cushings                Diagnostic Tests on Patients with Cushing's
                        Syndrome
DDT                     DDT in Kale
GAGurine                Level of GAG in Urine of Children
Insurance               Numbers of Car Insurance claims
Melanoma                Survival from Malignant Melanoma
OME                     Tests of Auditory Perception in Children with
                        OME
Pima.te                 Diabetes in Pima Indian Women
                              :
                              :
```

查看当前所有已加载包中的数据集

data( )

在命令行键入列出的任一数据集名，就可查看该数据集

```
> Insurance
   District   Group    Age Holders Claims
1         1     <1l    <25     197     38
2         1     <1l  25-29     264     35
3         1     <1l  30-35     246     20
4         1     <1l    >35    1680    156
5         1 1-1.5l    <25     284     63
6         1 1-1.5l  25-29     536     84
7         1 1-1.5l  30-35     696     89
8         1 1-1.5l    >35    3582    400
9         1 1.5-2l    <25     133     19
10        1 1.5-2l  25-29     286     52
11        1 1.5-2l  30-35     355     74
12        1 1.5-2l    >35    1640    233
```

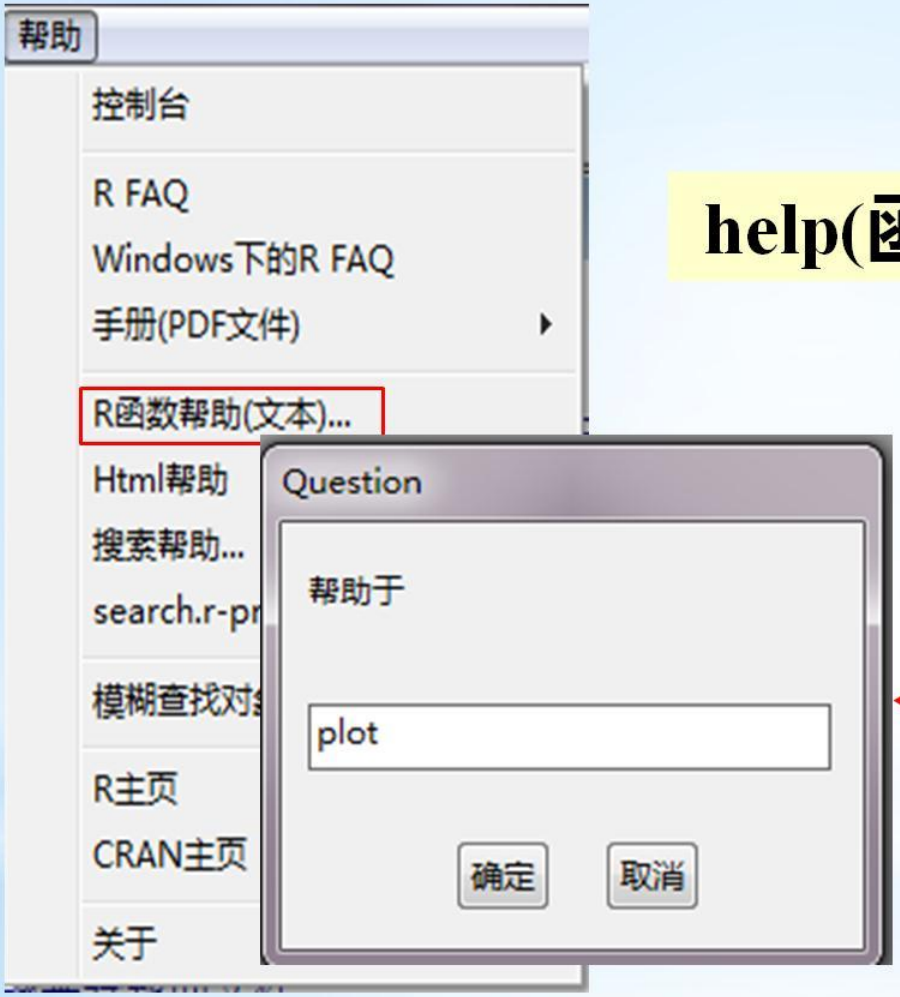
example(数据集名)

help(数据集名)

显示某数据集的内容及示例

运行数据集自带的示例

获取某函数的帮助



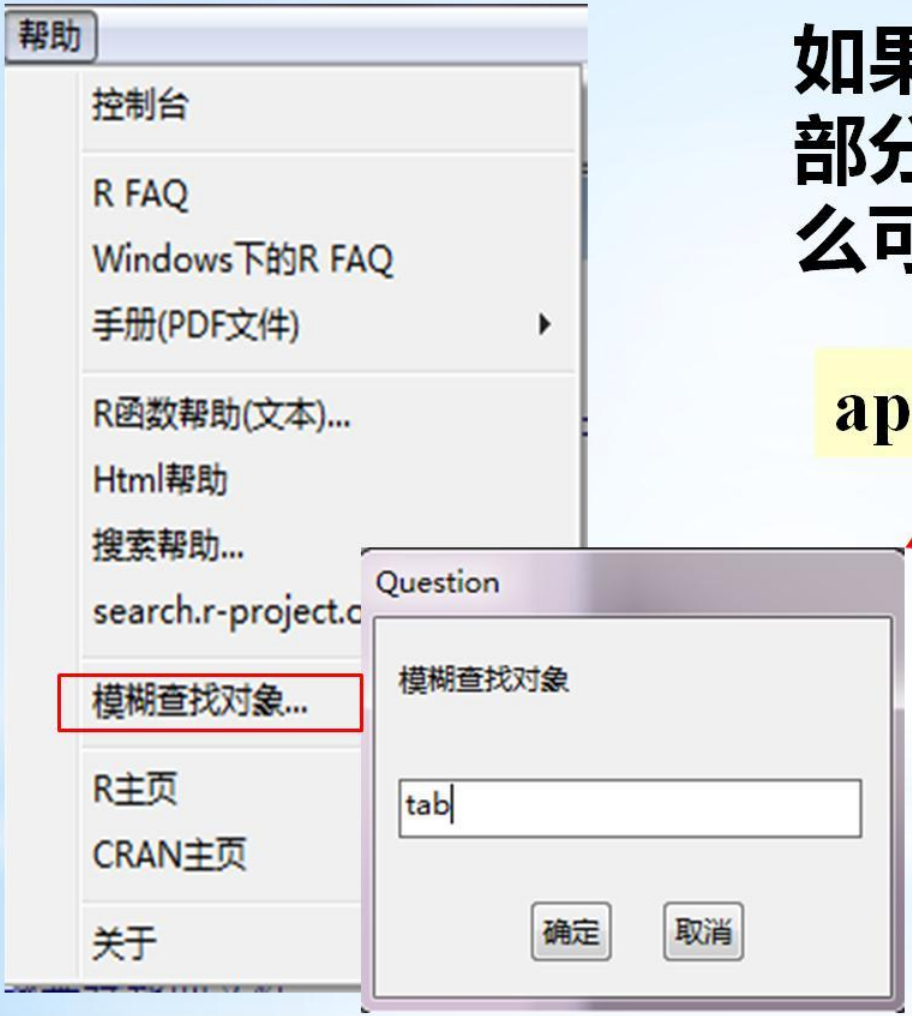
help(函数名) 或 ?函数名

help(plot)

或

?plot

获取某函数的帮助



如果你只知道函数的部分名称，如tab,那么可以使用

```
apropos("tab")
```

搜索得到载入内存的所有包含tab字段的函数

获取某函数的帮助

在工作窗口直接运行函数帮助自带的示例:

```
example(函数名)
```

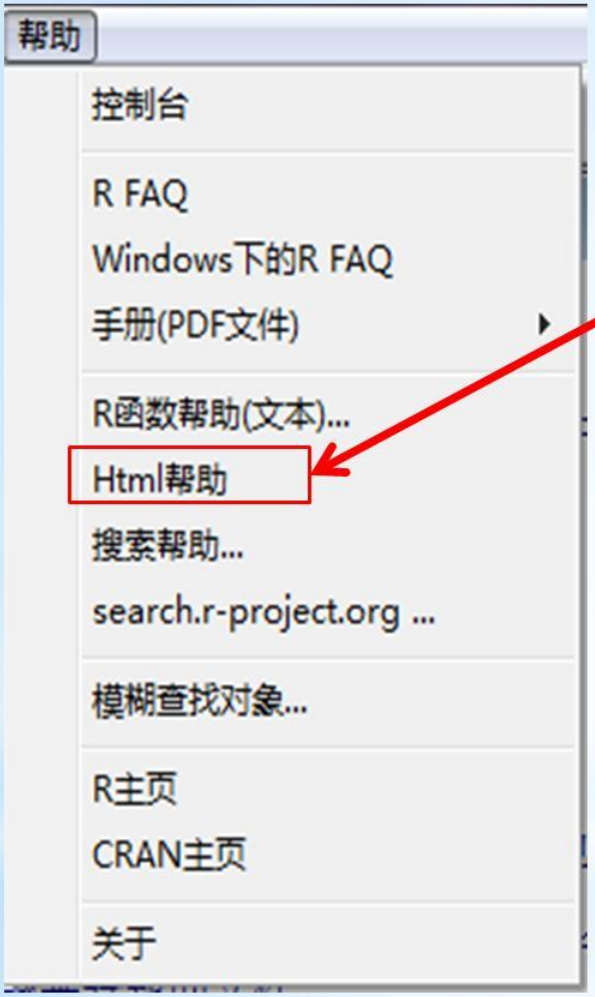
例如：

```
example(plot)
```

```
example(mean)
```

不需进入帮助文档，R会负责运行文档中的例子，并显示结果.

查看R的文档

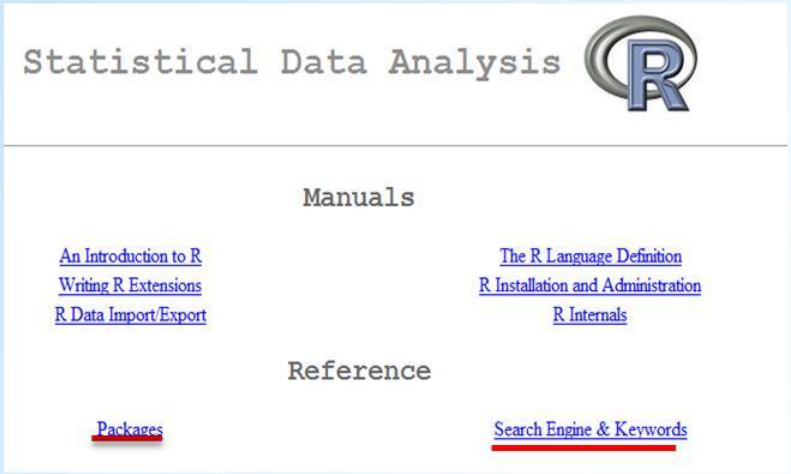


help.start()

该函数会打开一个浏览器窗口，显示出最顶层的目录

见下页图

查看R的文档



Packages

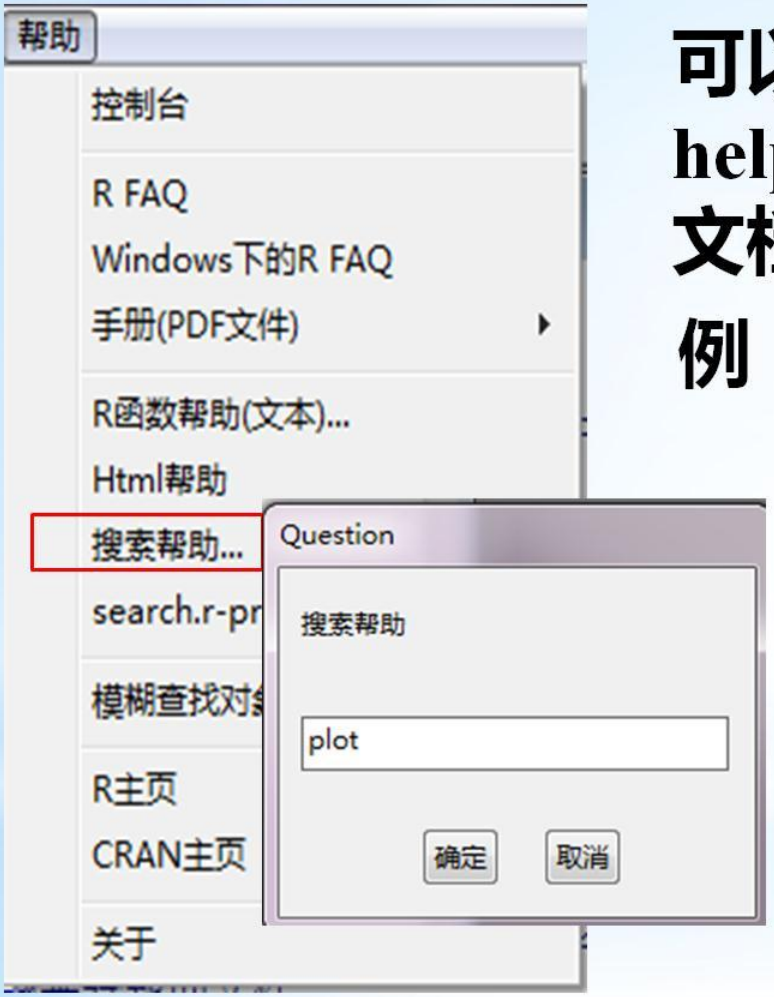
点击可查看所有已经安装的包。点击各个包的名称，就能看到其中的函数.

两个链接特别有用

Search Engine \& Keyword

一个简单的搜索引擎，可以用关键字搜索文档.

查看R的文档



可以直接使用 ?topic 或 help(topic) 来搜索本地帮助文档获取 topic 的帮助信息

例：help(plot) 或 ??plot

搜索帮助中也可键入包名，查看数据包的简介，进一步按照函数和数据集的索引查找相关信息

在网上搜索帮助信息

在网络上搜索与R有关的信息和问题解答

在R中使用 RSiteSearch("关键字或短语")

如 RSiteSearch("canonical correlation")

进入R社区网站，得到专家的帮助

如 R-help mailing list

stackoverflow

或用浏览器直接对有关R的一些问题进行搜索

内存分配和使用

查看当前设置下操作系统能分配给R的最大内存

使用指令 memory.limit( )

查看当前R已使用的内存，可使用

memory.size(F)

查看已分配的内存 memory.size(T)

查看对象x的存储模式：

storage.mode(x)

内存分配和使用

当对象占用内存过多时，会出现无法分配内存的错误，影响程序的运行。

对不再需要的对象，可使用

rm(对象名）

删除它来释放一些内存空间

删除内存中所有对象： rm(list=ls( ))

通过rm()删除对象后R进程占用的内存并没有被立即释放，而是过一段时间后才会清理

内存分配和使用

如果想要删除的对象立刻被清理

使用垃圾处理函数 gc( )

立刻释放内存空间

通常当内存不够时系统会自动清理垃圾。我们要做的只是随时关注内存的使用情况，将不再使用的对象用rm( )删除。还可以通过处理对象等方法来获取更大的可用内存。

内存分配和使用

从R 3.0.0发布起，R 正式进入了 3.x 时代

R 3.0.0 带来了约 100 项的新特性，长向量的全面支持和若干项性能提升。

在官方公告中，有对新特征的说明

对2^31 以上向量的支持仅限于 64-bit 系统

64-bit 版本的 R 可分配内存的大小仅受系统的限制。可通过系统工具（如 bash shell 下的 ulimit ）来设置单个 R 进程的整体内存占用，尤其是在多用户环境下。有若干包需要 4GB 以上的虚拟内存来加载。

内存分配和使用

64-bit Windows 版本的 R 可用内存大小默认限制为已安装内存的大小。该值可通过启动参数 --max-mem-size 或环境变量 R_MAX_MEM_SIZE 设置。

最好在64-bit Windows 版本安装R

实践1

1.安装一个用于可视化类别数据的包(vcd包)
2.查看该包的主要内容
3.加载该包
4.查看包中的数据集
5.显示某数据集的内容(如Arthritis)
6.查看该数据集
7.运行该数据集的相关示例
8.查看对其中某函数的帮助(如cd_plot)
9.运行该函数的示例
10.退出

实践1代码

```
install.packages(“vcd”)#联网时可用
#不联网时从本地安装
help(package="vcd")
library(vcd)
data(package="vcd")
help(Arthritis)
Arthritis
example(Arthritis)
help(cd_plot)
example(cd_plot)
q()
```