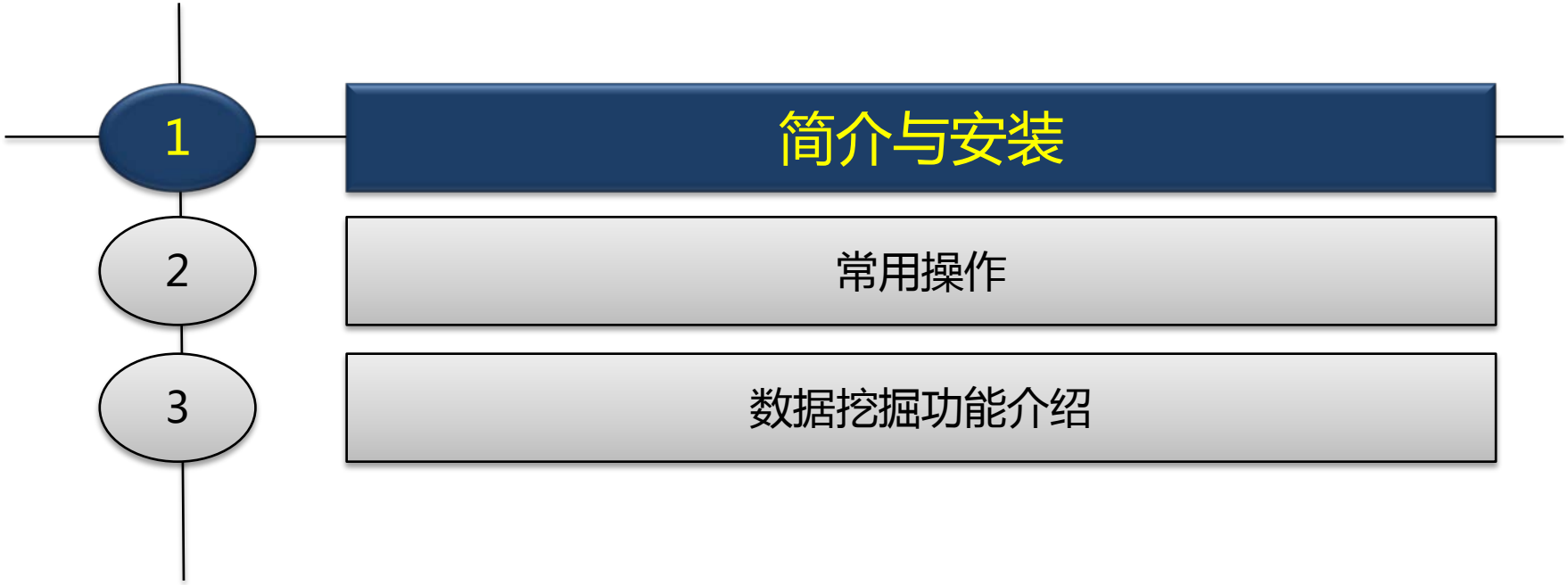
信息分析与预测实验

第1章 R软件简介

经济与管理学院 孙蕾

目录



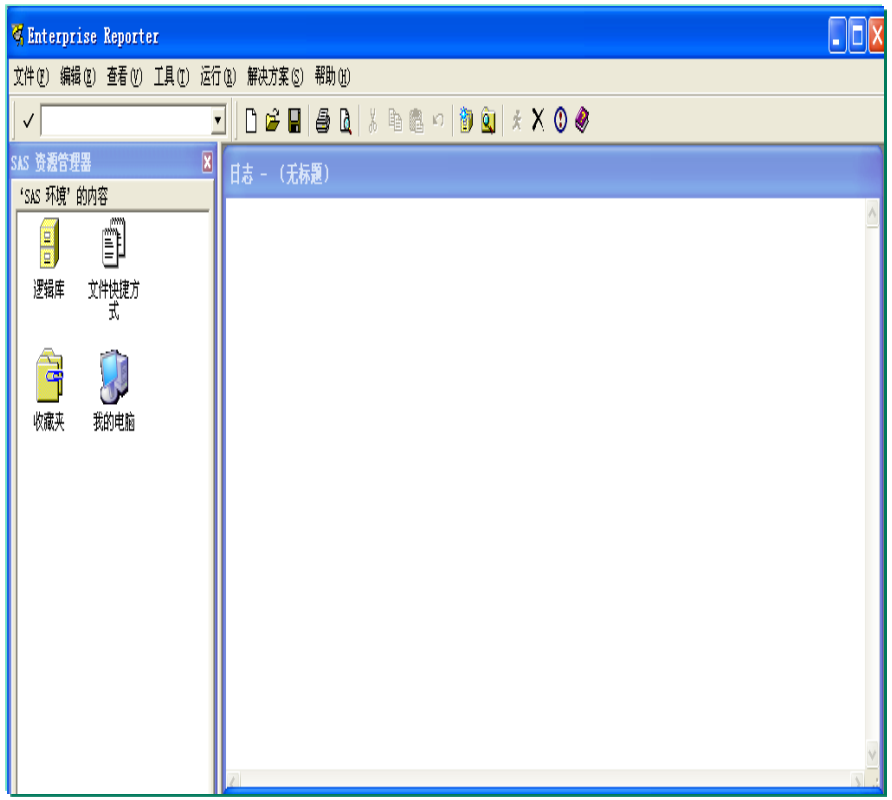
什么是R？

R的源起

- R是S语言的一种实现。S语言是由 AT&T贝尔实验室开发的一种用来进行数据探索、统计分析、作图的解释型语言。最初S语言的实现版本主要是S-PLUS。
- S-PLUS是一个商业软件，它基于S语言，并由MathSoft公司的统计科学部进一步完善。后来Auckland大学的Robert Gentleman 和 Ross Ihaka 及其他志愿人员开发了一个R系统。R的使用与S-PLUS有很多类似之处，两个软件有一定的兼容性。

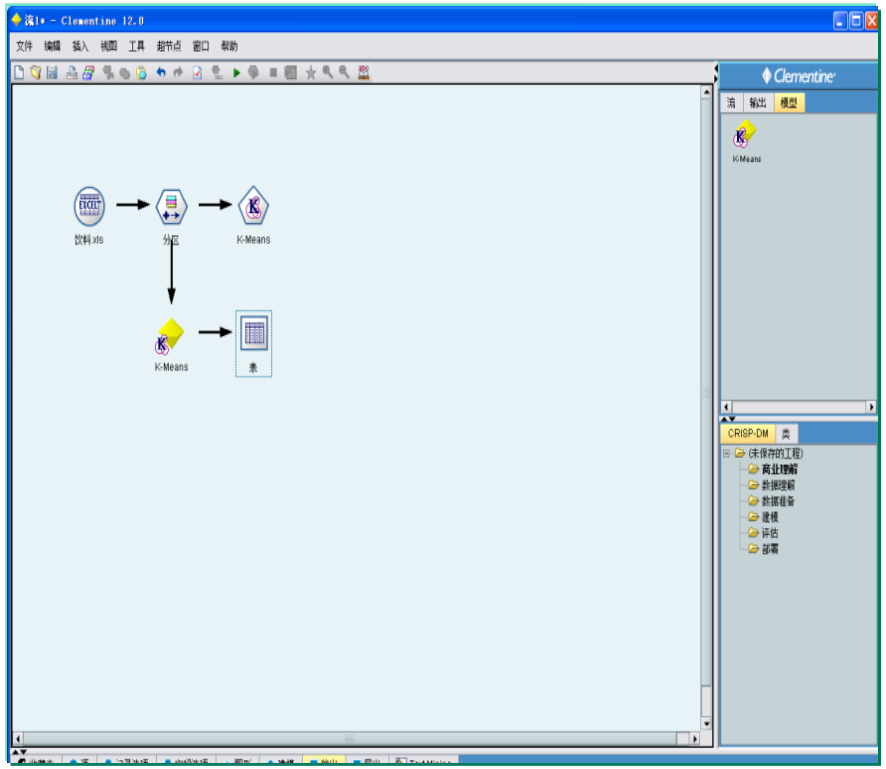

常用数据挖掘工具-----商用

SAS：商业软件，模块固定不可修改，提供菜单操作和编程。



收费

SPSS modeler：商业软件，流操作的图形界面模式，模块固化。



R语言特点

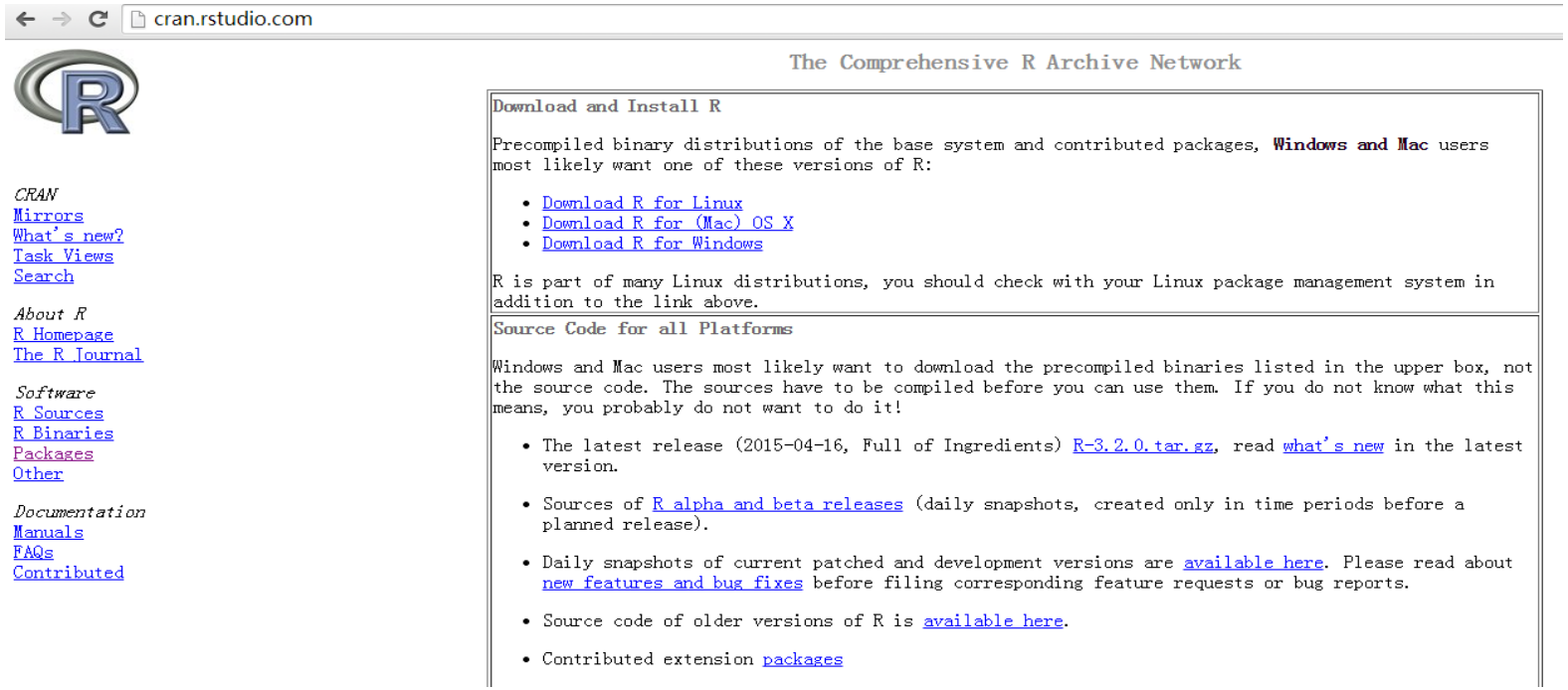
- R是一种为统计计算和绘图而生的语言和环境，它是一套开源的数据分析解决方案，由一个庞大且活跃的全球性研究型社区维护。
- 多数商业统计软件价格不菲，而R是免费的！
- R语言具备可扩展能力且拥有丰富的功能选项，帮助开发人员构建自己的工具及方法，从而顺利实现数据分析。
- R可运行与多种平台之上，包括Windows、Unix和Mac OS X。这基本上意味着它可以运行于你所能拥有的任何计算机上。
- 国际上R语言已然是专业数据分析领域的标准

R语言的缺点

- R是一种解释性语言，和编译性语言相比，速度显得略慢一点。
- R所有计算都是在内存中进行的。
- 由于R语言的自由，各种包的编写者来自不同的领域，所以在一定程度上是比较混乱的。

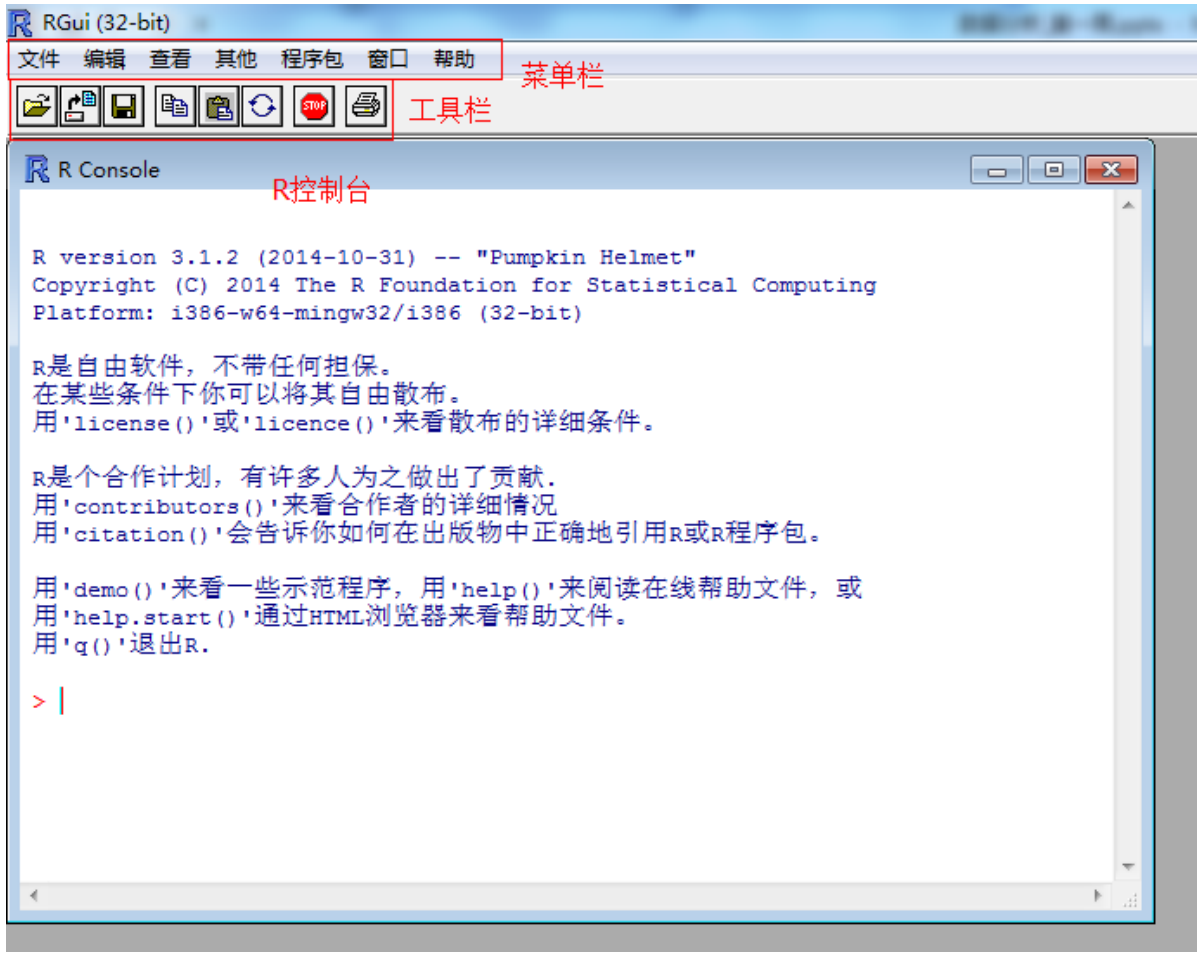
R语言的获取和安装

- R可以在CRAN(Comprehensive R Archive Network) http://cran.r-project.org/mirrors.html 上免费下载。
- Linux、Mac OS X和Windows都有相应编译好的二进制版本。
- 可以通过安装成为包(package)的可选模块(同样可从CRAN下载)来增强R的功能。



R的图形用户界面

- 在R的GUI窗口里，有菜单栏、工具栏和R的控制台。



Rstudio:一个友好的编辑器

- R自身带的编辑器很不好用，这里推荐Rstudio，它是专门用于R语言环境的IDE。
- Rstudio可以从其官网 http://www.rstudio.com/ 上免费下载安装。请根据本机操作系统选择系统支持版本自行下载安装。

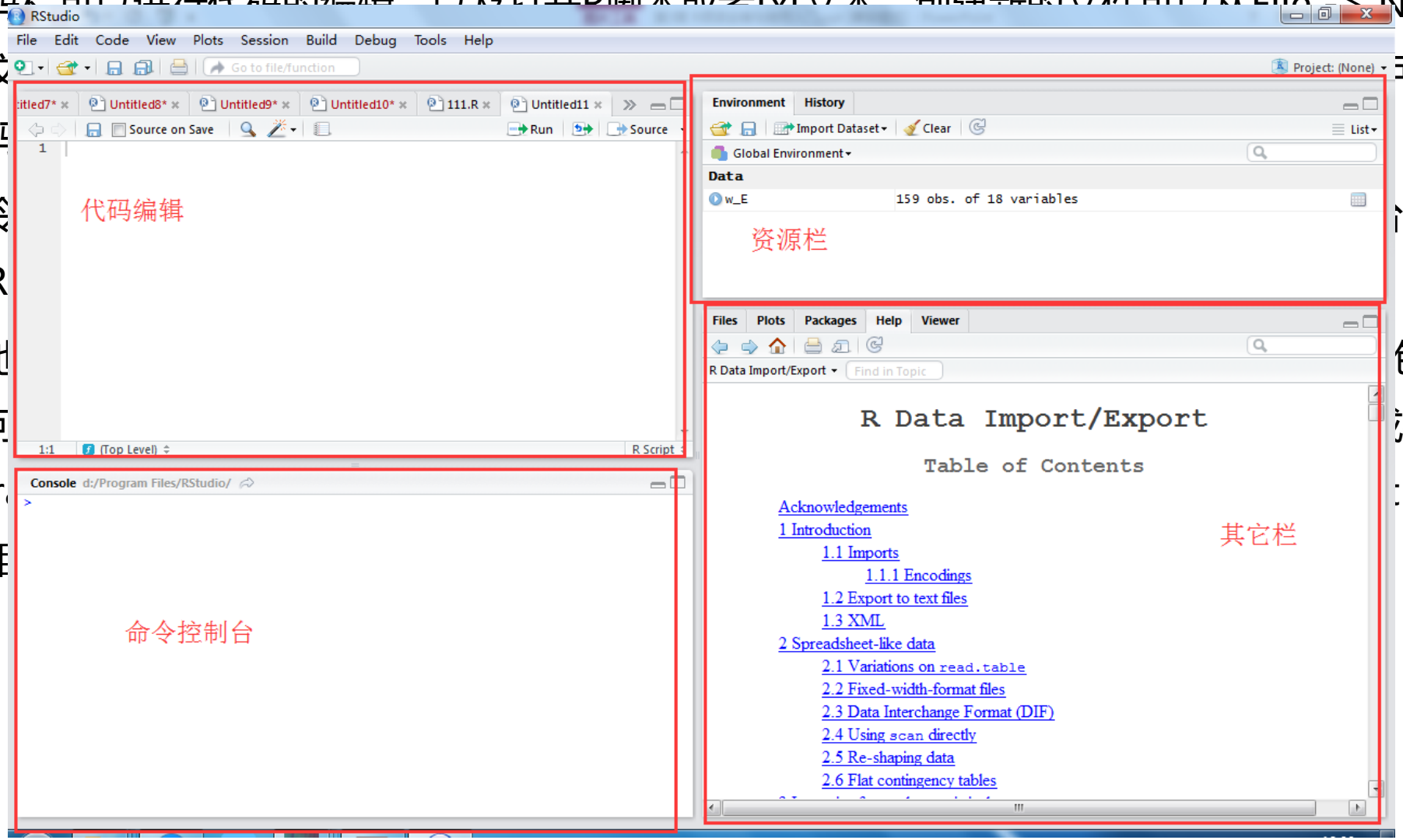
R的更新

- R的升级同城是通过从CRAN（http://cran.r-project.org/bin/）上下载和安装最新版的R，这种方式需要重新设置各种自定义选项，包括之前安装的扩展包。
- 可以将R目录下etc文件夹中的Rprofile.site 文件及R目录下的library文件夹保存到其它地方，带安装新版本后，再移动到相应的位置即可。
- 单击安装目录bin下的Rgui.exe，然后运行以下代码

```
install.packages(“installr”)
require(installr)
updateR()
```

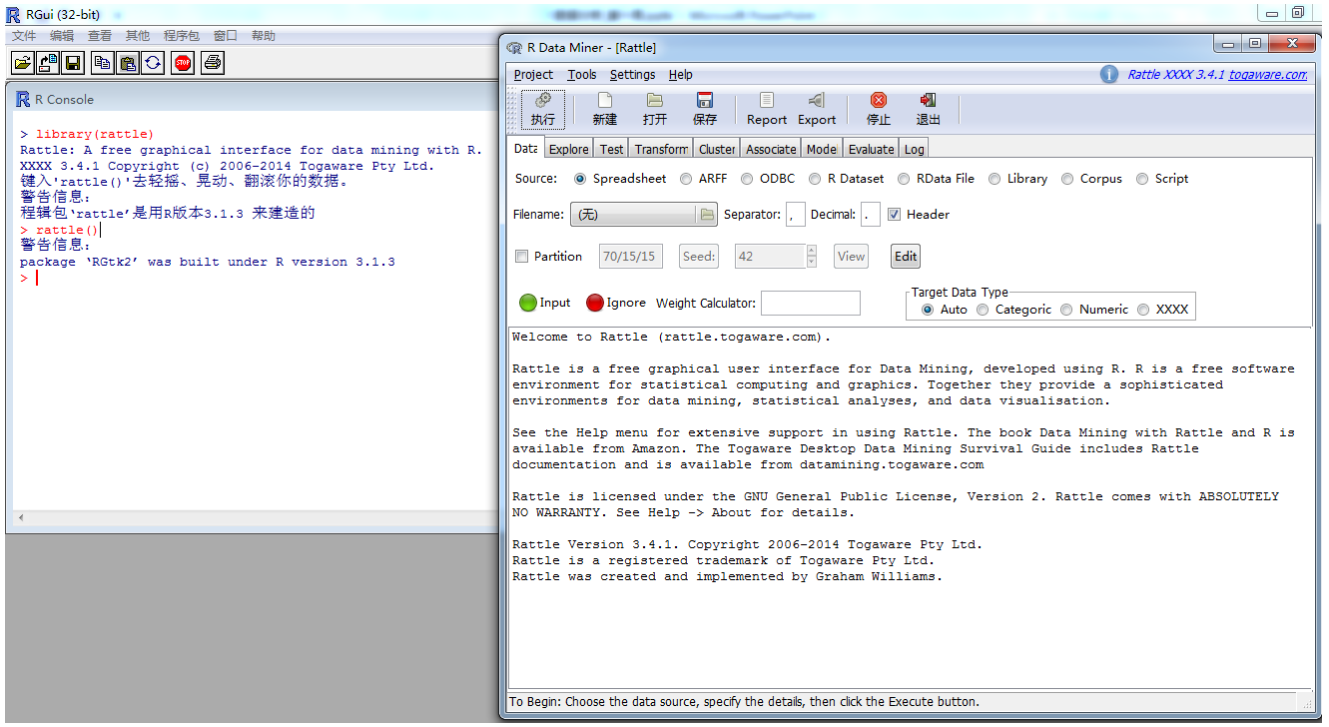
RStudio窗口介绍

- 代码栏可以进行代码的编辑，以及打开R脚本或者txt文本，创建新的文件可以从File -> New里选择，打开文
- 命令
- 其他



rattle:可视化数据挖掘工具

- 数据可视化旨在借助图形化手段，清晰有效地传达与沟通信息。
- R语言有众多的绘图工具包，例如ggplot2、lattice等。而在动态绘图方面，则可以利用rggobi和ggobi软件进行协同工作。
- 对懒得敲命令的读者来说，还可以利用rattle工具的图形界面进行数据挖掘和可视化工作。



目录

常用操作

- R是一种区分大小写的解释性语言。
- 可以在命令提示符(>)后每次输入并执行一条命令，或者一次性执行写在脚本文件中的一组命令。
- R中有多种数据类型，包括向量、矩阵、数据框以及列表（各种对象的集合）。我们将在后面中讨论这些数据类型。
- R中的多数功能是由程序内置函数和用户自编函数提供的，一次交互会话期间的所有数据对象都被保存在内存中。
- 一些基本函数是默认直接用的，而其他高级函数则包含于按需要加载的程序包中。

常用操作——help

- help：打开帮助文档，可以采用“？”命令的模式，或者help(命令)显示函数的说明。

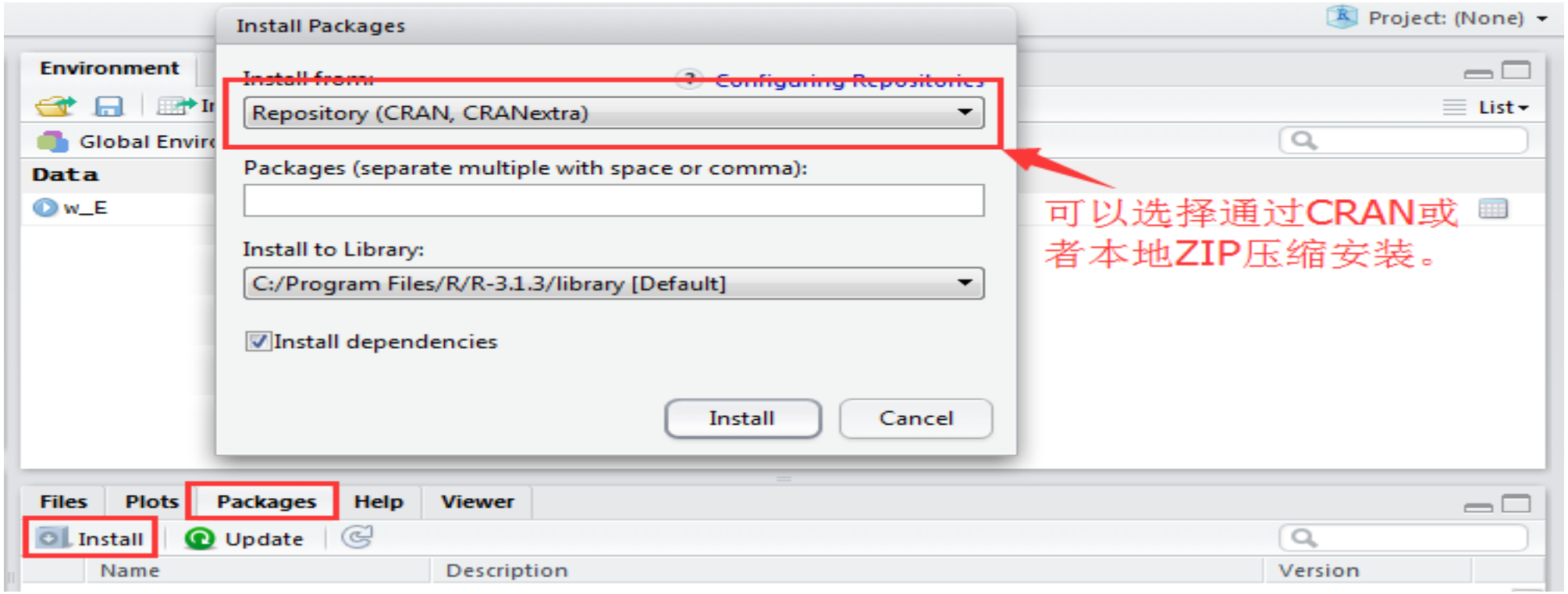
函数	功能
help.start()	打开帮助文档
help("plot")或者 ?plot	查看函数plot的帮助（引号可以省略）
help.search("plot")或者 ??plot	以plot为关键词搜索本地帮助文档
example(“plot”)	函数plot的使用示例（引号可以省略）
RSiteSearch("plot")	以plot为关键词搜索在线文档个邮件列表存档
apropos("plot",mode="function")	列出名称中含有plot的所有可用函数
data()	列出当前以加载包中所含的所有可用示例数据集
vignette()	列出当前已经安装包中所有可能的vignette文档
vignette(“plot”)	为主题plot显示指定的vignette文档

常用操作——package

- 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。
- 计算机上存储包的目录称为库(library)。
- 函数.libPaths()能够显示库所在的位置。
- 函数library()则可以显示库中有哪些包。
- R自带了一系列默认包(包括base、datasets、utils、grDevices、graphics、stats以及methods),它们提供了种类繁多的默认函数和数据集。其他包可通过下载来进行安装。

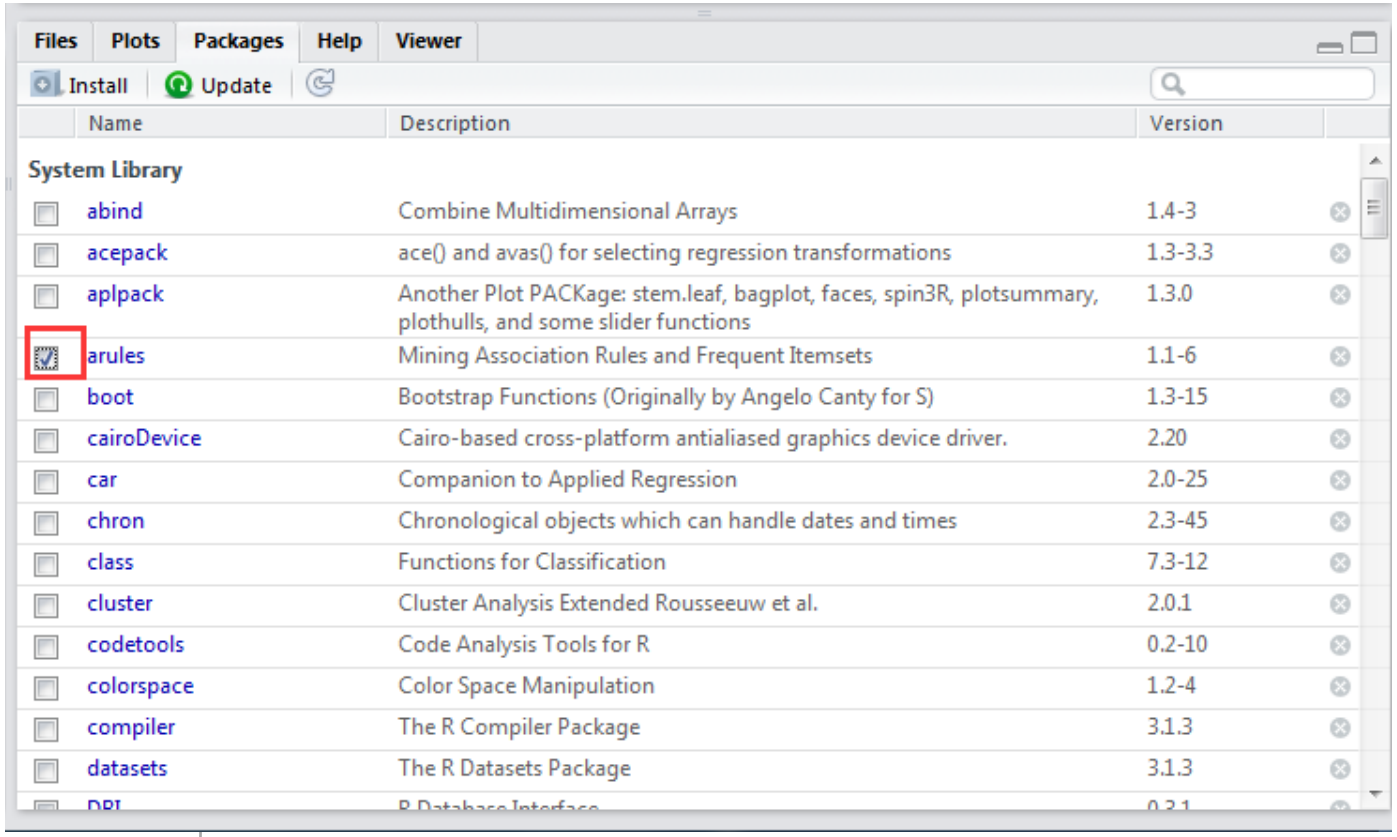
常用操作——package

- install.packages(“”)：安装R包命令
- 或者通过RStudio的Packages目录下的install进行安装可以选择有网和无网安装。



常用操作

- library(package_name) ：加载已经安装的R包的命令，或者通过RStudio的Packages目录下选择相应的R包就行。
- detach(“package:bmp”)：分离加载的bmp包。



常用操作——RStudio使用小技巧:

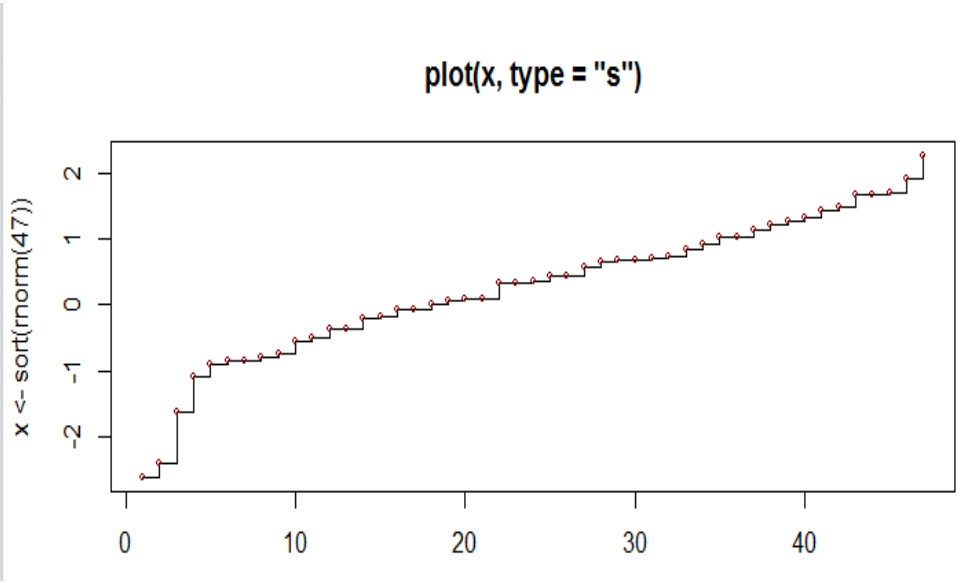
- rm() ：清除单个变量使用或者内存中所有的变量（ rm(list = ls(all = TRUE)) ）；
- Ctrl+L：清除console中的所有显示内容；
- data ( ) ：查看R包内置的数据集；
- getwd()/setwd()：获取或者设置当前工作目录的位置；
- file.choose() ：打开一个Windows 标准文件选择对话框，手动选择文件，choose.dir()的功能相同；
- R 里面使用必须使用双反斜杠或单斜杠表示文件路径，如

  d : \ \scripts\\ xgobi.csv或者d : /scripts/xgobi.csv

常用操作——plot

- plot： 画图函数，可以设置参数进行定制化。

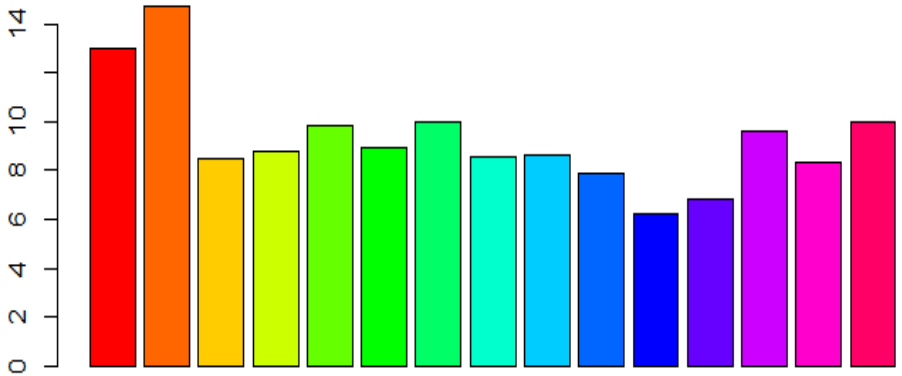
```
> plot(x <- sort(rnorm(47)), type = "s", main
= "plot(x, type = \"s\")")
> points(x, cex = .5, col = "dark red")
```



常用操作——plot

- barplot：画条形图，可以设置参数进行定制的图像制作。R默认情况下提供8种颜色，还有一些专门处理“色谱”的包，比如RColorBrewer 和colorRamps，尤其是RColorBrewer提供了连续型(sequential)、离散型(diverging)、定性型(qualitative)三种配色方案。R内置的颜色可以通过colors()得到，R中的颜色通过col2rgb()函数与RGB的颜色对应和相互转换。

```
>palette(rainbow(15))    ##R默认8种颜色，可以使用palette函数进行修改
> barplot(rnorm(15,10,3),col=1:15)    ##col参数为颜色参数
```

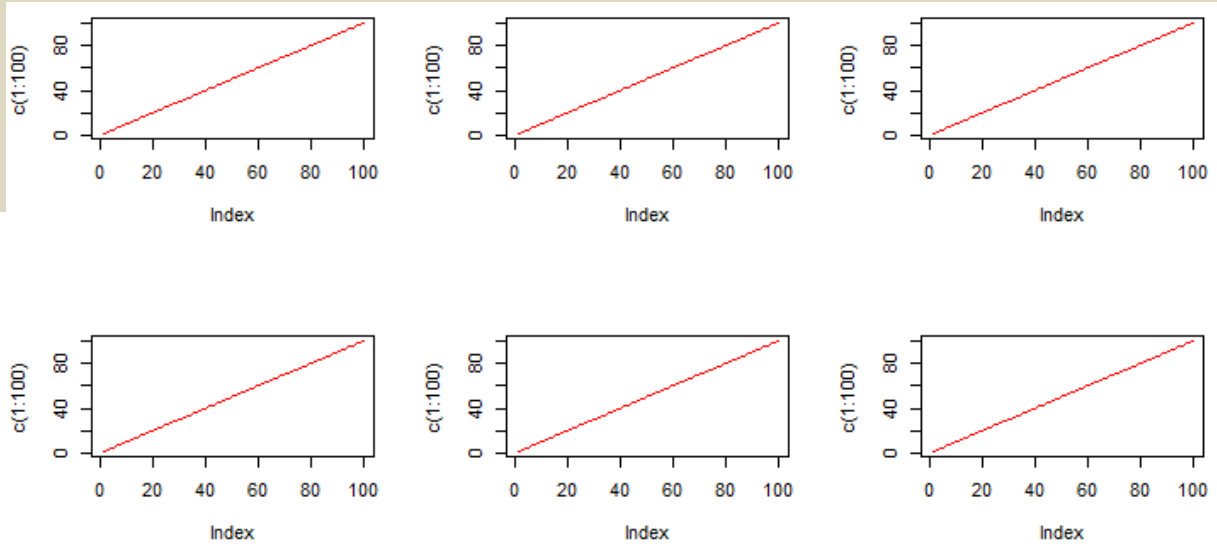


常用操作——plot

- par/layout：提供在同一画面画出多张图。par通过mfrow或者mfcol参数进行修改。

➢par(mfrow=c(2,3))　##设置图形排列方式：2行3列一共6个图，其中mfrow是按照行的顺序排列，##mfcol参数是按照列的顺序排列。

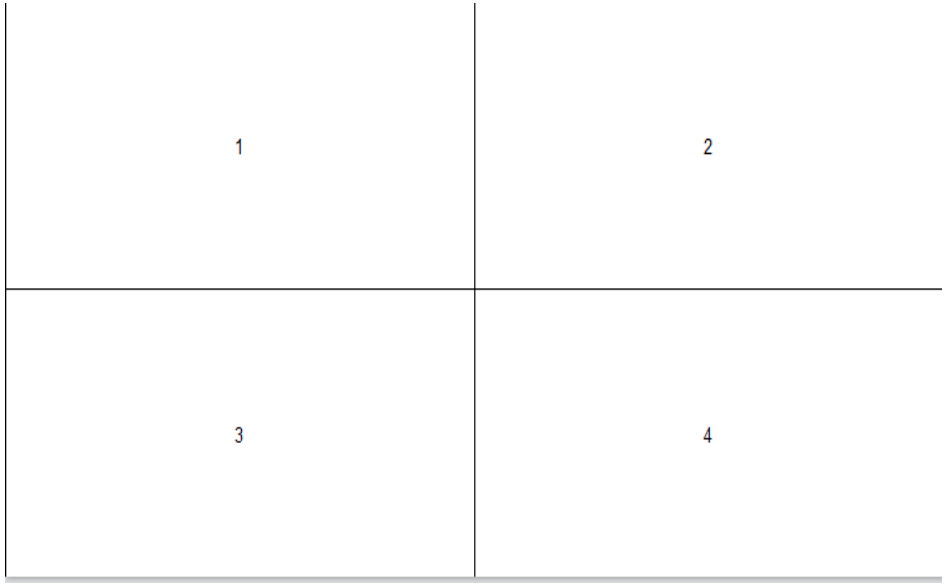
```
> plot(c(1:100),type="l",col="red")
> plot(c(1:100),type="l",col="red")
> plot(c(1:100),type="l",col="red")
> plot(c(1:100),type="l",col="red")
> plot(c(1:100),type="l",col="red")
> plot(c(1:100),type="l",col="red")
```



常用操作——layout

- layout：layout可以设置图形绘制顺序和图形大小。其输入参数至少要对{1 … n}里每一个值都有参考值，其中0代表没有。可以通layout.show(n)命令查看图形的布局。

```
> layout(matrix(c(1,2,3,4), 2, 2, byrow = TRUE))     ##生成可以放4张图的窗口，为2行2列。
> layout.show(4)                ##查看窗口布局
```




```
> layout(matrix(c(1,2,0,4), 2, 2, byrow = TRUE))
##这样就会出错，提示格式矩阵至少要对{1 … 4}里每一个值都有参考值
```

目录

R数据挖掘相关包

- R在数据挖掘领域也提供了足够的支持。比如关联规则挖掘、聚类、分类等，通过加载不同的R包就能够使用数据挖掘的功能。

功能	函数及加载包
分类与预测	nnet()需要加载BP神经网络nnet包; randomForest()需要加载随机森林randomForest ; svm()需要加载e1071包; tree()需要加载CRAT决策树tree包等;
聚类分析	hclust()函数、kmeans()函数在stats包中
关联规则	apriori()需要加载arules包
时间序列	arima()需要加载forecast、tseries包