

大数据高端人才专项计划



東北大學
Northeastern University



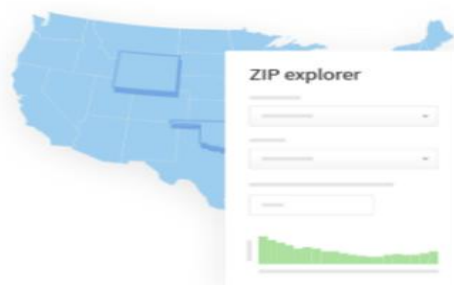
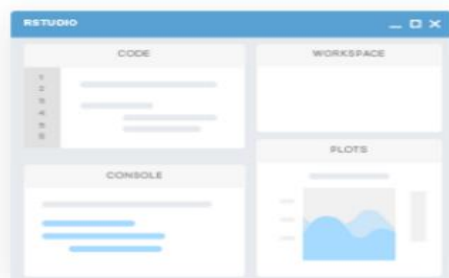
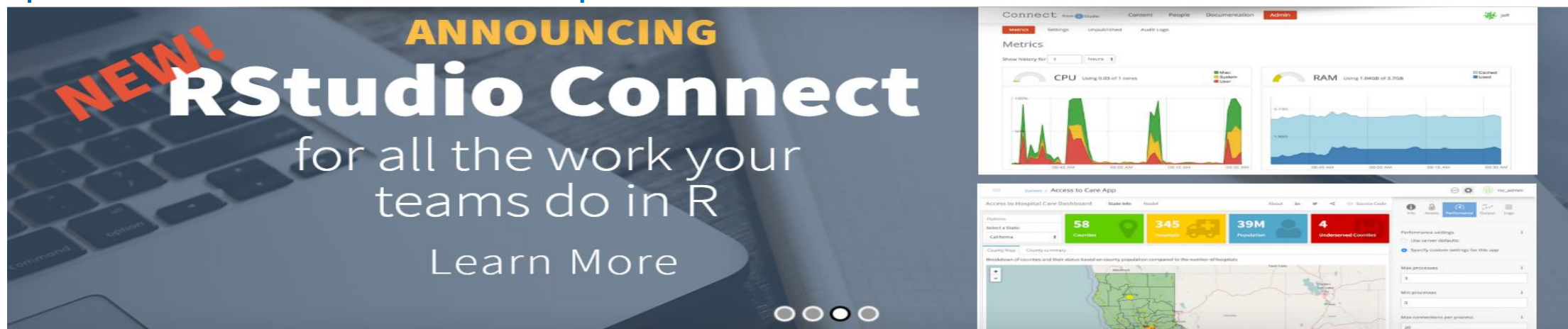
R环境准备

➤ 环境准备:

准备 RStudio-1.1.456 、 R 3.3.2以上 、 R NOTEBOOK 。

➤ 获取网址:

<https://www.rstudio.com/> <https://cran.rstudio.com/>



R包

➤ R包介绍:

dplyr <http://www.xueqing.tv/course/31>

data.table

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

leaflet <http://rstudio.github.io/leaflet/shiny.html>

ggplot2

plotly https://cpsievert.github.io/plotly_book/

h2o <https://www.h2o.ai/>

熟悉R环境

➤ R环境介绍:

选中R源文件中部分代码，点击Run即可执行

数据查看

The screenshot displays the RStudio environment with the following components:

- Script Editor:** Contains R code for creating vectors and matrices. A red circle highlights the first line of code: `a <- c(1, 2, 5, 3, 6, -2, 4)`.
- Console:** Shows the execution of the first line of code: `> a <- c(1, 2, 5, 3, 6, -2, 4)`.
- Environment Pane:** Displays the current data environment. It shows a global environment with a data frame named `patientdata` containing 4 observations and 4 variables. The variables are `x` (integer), `y` (integer), `a` (numeric), and `age` (numeric). The values for `a` are 1, 2, 5, 3, 6, -2, 4.
- Buttons:** A red circle highlights the `Run` button in the toolbar.

R控制台

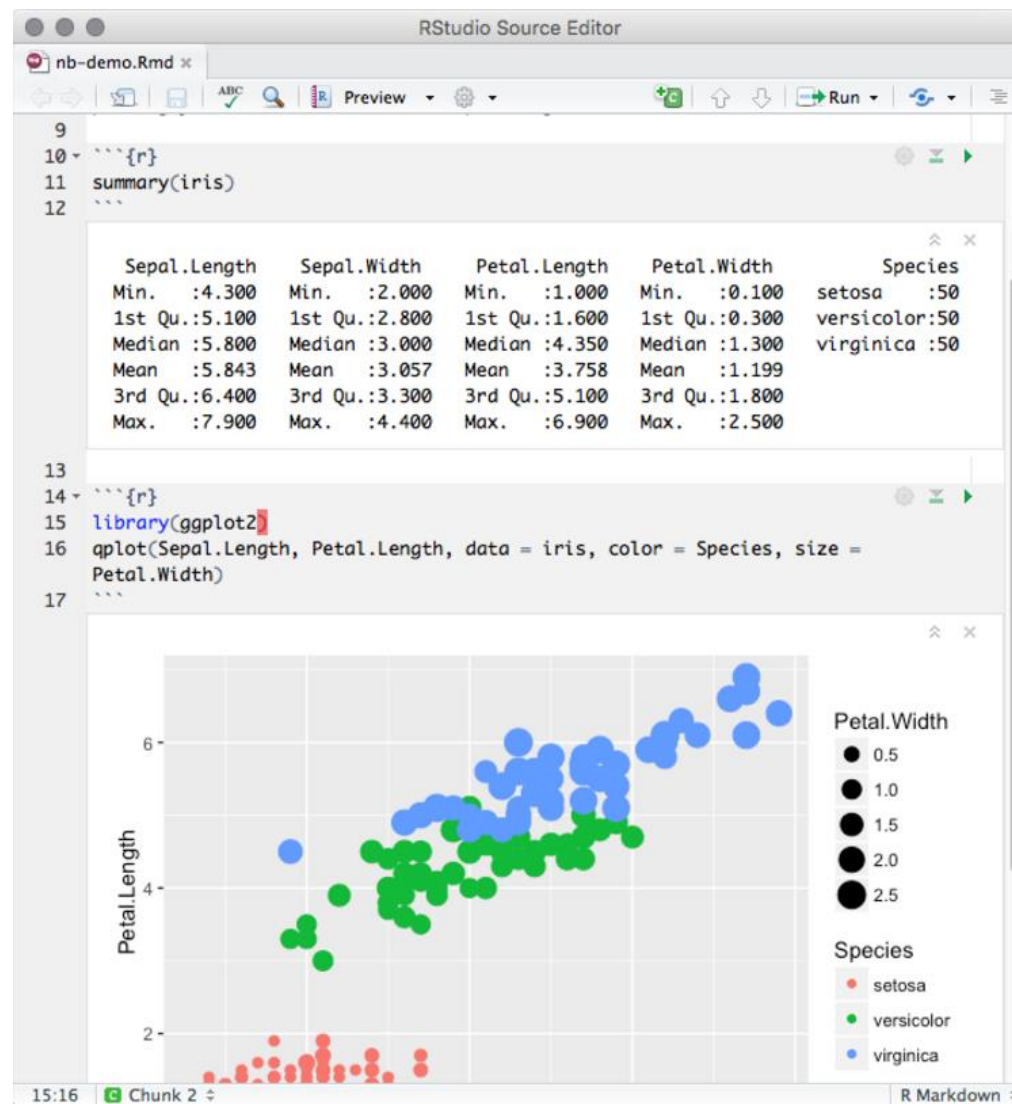
图形绘制区域等

熟悉R环境

➤ R NOTEBOOK 介绍:

2016年10月5日发布RStudio1.0版本以上的新功能。

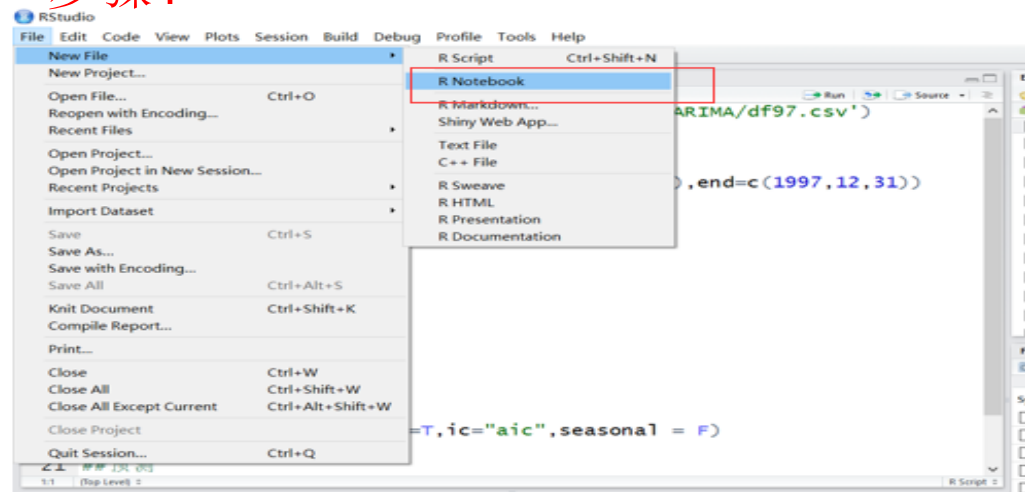
特色：交互式编程界面。



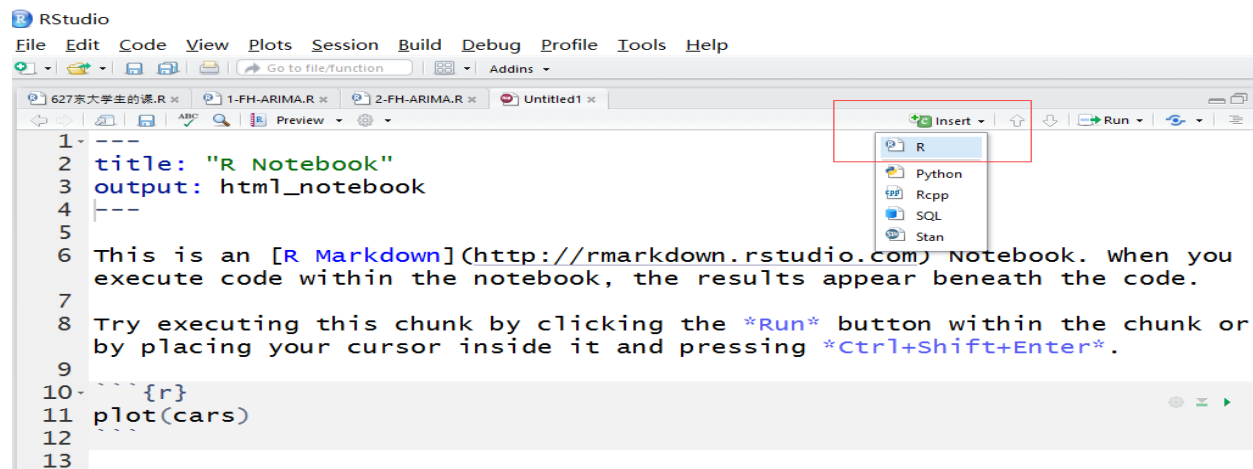
熟悉R环境

➤ R NOTEBOOK 应用介绍:

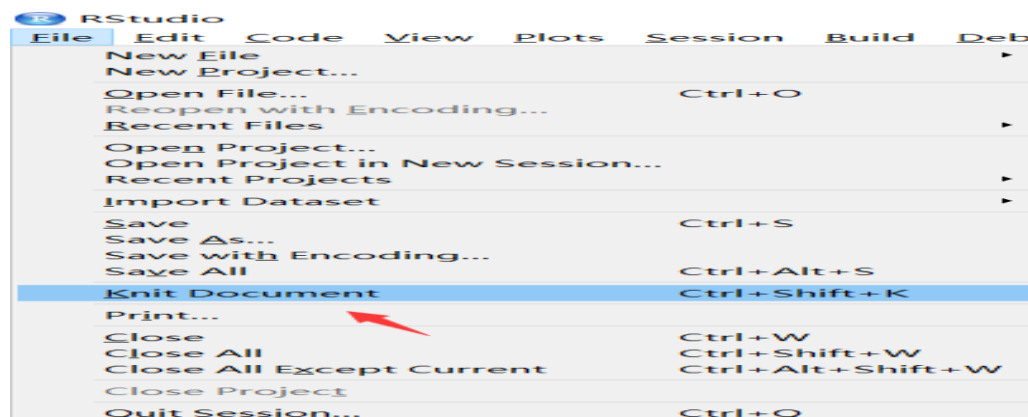
步骤1



步骤2



步骤3

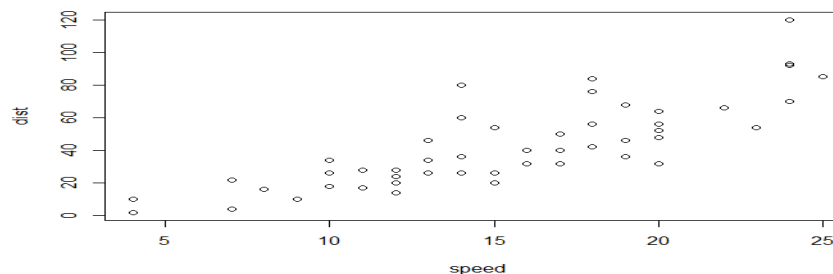


步骤4

R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.
Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
plot(cars)
```



知识点回顾

➤ 数据集的创建:

数据集 (data set) 是一个数据的集合, 通常以数据库表格的形式出现。

➤ 基本的数据结构:

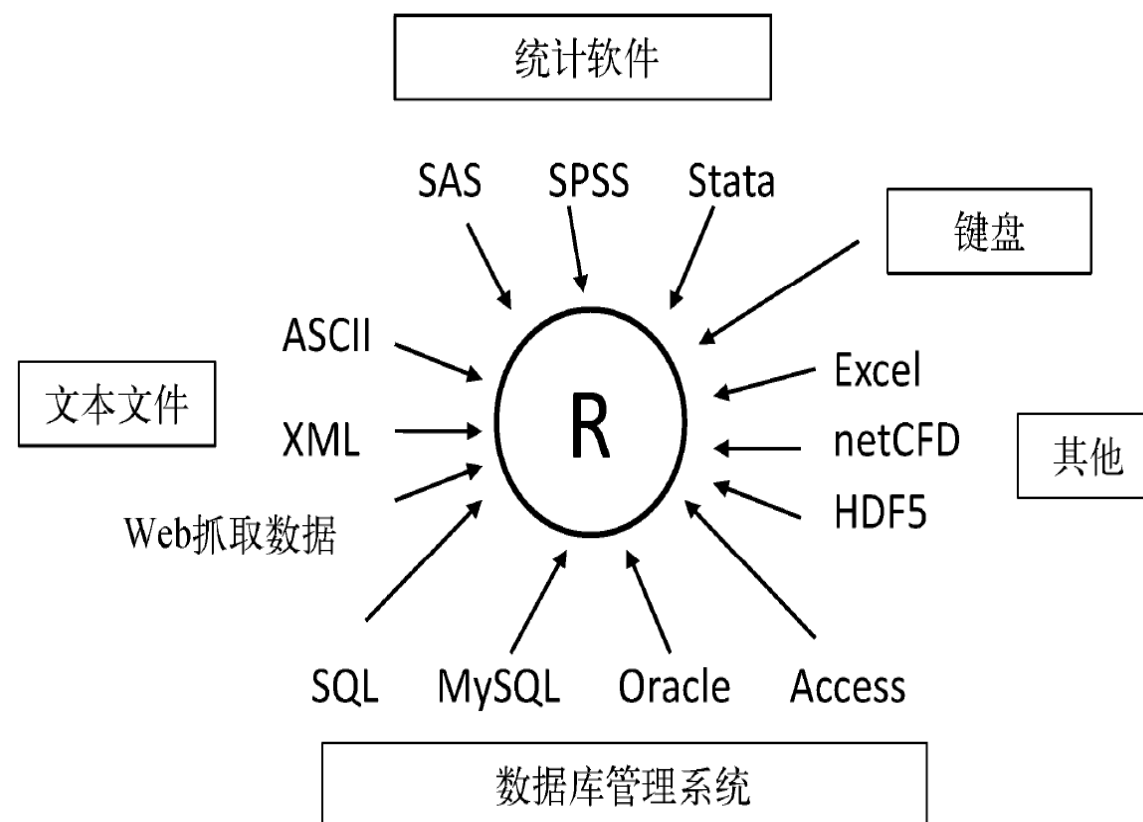
向量、矩阵、数组、数据框、因子、列表

➤ 数据的输入和导入:

R支持很多种输入和导入, 如右图

➤ 处理数据对象的常用函数:

如: `dim()`查看某对象的维度
`class()`查看某对象的类型
`head()`查看某对象的开始部分
`tail()`查看某对象的结尾部分
`cbind()`按列合并对象
`rbind()`按行合并对象



实训项目

题目：

1.安装R环境，熟悉环境，熟悉R NOTEBOOK的用法。

2.基本数据结构练习：

- (1) 创建数值从1到10，间隔为2，名称为A的向量
- (2) 将数字1到12每个重复3次写入向量B中：(1,1,1, ...,12,12,12) 提示rep函数
- (3) 输出B向量的长度和去重后的向量数值
- (4) 输出A、B的交集，并集C，差集
- (5) 将集合C按降序排序
- (6) 选取C的第3个元素；第4到最后的元素；数值在区间大于等于3小于7的元素
- (7) 将数值型向量C转化为字符型
- (8) 向量C的长度，最大值及其索引
- (9) 将A转化成数组类型变量名为a,查看a的类型
- (10) 用1~20的数字构成两个4*5的矩阵，其中M1为按列输入，M2为按行输入，计算M3为M1+M2；并构建M4，它由M3各列构成，但不包含第3列。
- (11) 用1~9的数字构成一个3*3的按列输入的矩阵M5；求M5的对角阵M6；计算M5与M6的矩阵乘法得到M7；求M7的转置矩阵M8
- (12) 用1~12的数字构成一个4*3的按列输入的矩阵M9，求M9的列加和；求M9的行平均

实训项目

题目：

3.数据的导入和处理对象常用函数练习：

- (1) 从csv文件中读取algae数据集赋值给algae1
- (2) 查看algae1的前10条数据
- (3) 输出algae1的基本统计信息，数据的维度，特征名称，查看season、size和NO3各列的数据类型
- (4) 选出季节为夏天的样本存在newalgae1，输出newalgae1样本行数
- (5) 将newalgae1的列cl中的缺失值用本列得中位数填充
- (6) 将algae1删除含有缺失值的样本，输出原始样本的行数和剩余样本行数
- (7) 编辑algae1并另存为algae2（任意修改某个点的值）
- (8) 将algae1和algae2，按行合并得到algae3，输出algae1、algae2 和algae3的样本行数

提交内容：将题目2和题目3的代码和结果保存并提交，格式：R第一次实训-姓名+学号。

THANKS

