零基础学数据可视化

一、导论

毛益民　副教授

浙江工商大学公共管理学院

2020-02-21

为何学数据可视化?


2018年09月28日 星期五 14:50
龙城街道智慧食安大数据智慧中心
龙城街道办
2018年09月份数据
主体巡查不合格数据统计
隐患项目
次数
不合格率
未超出许可事项从事冷食类食品（如沙拉类、焖烤等）、生食类食品（如刺身）等
174
12.37%
按从业人员持有效健康证上岗
131
9.32%
按照“明厨亮灶”建设规范指引要求，公示加工操作关键过
112
8.45%
按照“明厨亮灶”建设规范指引要求，公示加工操作关键过
101
7.37%
一般餐饮单位
集体餐饮单位
农贸商超
食安风险分布图
问题单位分布图
巡查员巡查覆盖分布图
检测员抽检覆盖分布图
被执法单位分布图
低风险
中风险
高风险
主体快检样品不合格公示
抽查时间
样品
主体名称
不合格率
2018-10-17
猪肉
潮汕三及第汤饭
监督销毁
2018-10-16
苹果
轻时光沙拉
执法抽检
2018-10-13
芝麻酱
福建沙田小吃 龙岗街道
监督销毁
2018-10-11
草鱼
五谷鱼粉
监督销毁
主体巡查整改统计
已整改
改整中
未改正
明亮厨房视频监控
主体巡查数量趋势
主体总数量
1569家
当前巡查员
1569人
当前快检员
156人
主体快检合格率统计
已整改
改整中
未改正
日
周
月
巡查总量：2610 家
不合格率：26.63%
快检总量：1610 批次
合格率：76.63%
整改总量：1110 批次
未整改率：46.63%

图表的作用

- 真实、准确、全面地展示数据；
- 以较小的空间承载较多的信息；
- 揭示数据的本质、关系、规律。

- -

可视化的终极目标是洞悉蕴含在数据中的现象和规律。

这包括多重含义：发现、决策、解释、分析、探索和学习.

正所谓：“一图胜千言”

Data Scientist The Sexy Job

October 2012 Issue

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

- See also an old article by NYT (2009): For Today's Graduate, Just One Word: Statistics
- And another famous McKinseay 2011 Report: Big data: The next frontier for innovation, competition, and productivity

为何采用R做数据分析?

R是什么?

R能做什么?

小调查

- 1.有处理过数据吗? 量比较大，内容比较复杂那种?
- 2.你们平时用什么软件处理数据?

绘图软件比较

LOGO	名称	开源	付费	技能要求	官方网站
	Excel	否	是	界面操作	https://support.office.com/en-GB/Excel
	Origin	否	是	界面操作	http://originlab.com/
	SigmPlot	否	是	界面操作	https://systatsoftware.com/products/sigmaplot/
	GraphPad Prism	否	是	界面操作	http://www.graphpad.com/
	MATLAB	否	是	编程	https://www.mathworks.com/products/matlab.html
	Python	是	否	编程	https://www.python.org/
	R	是	否	编程	https://www.r-project.org/

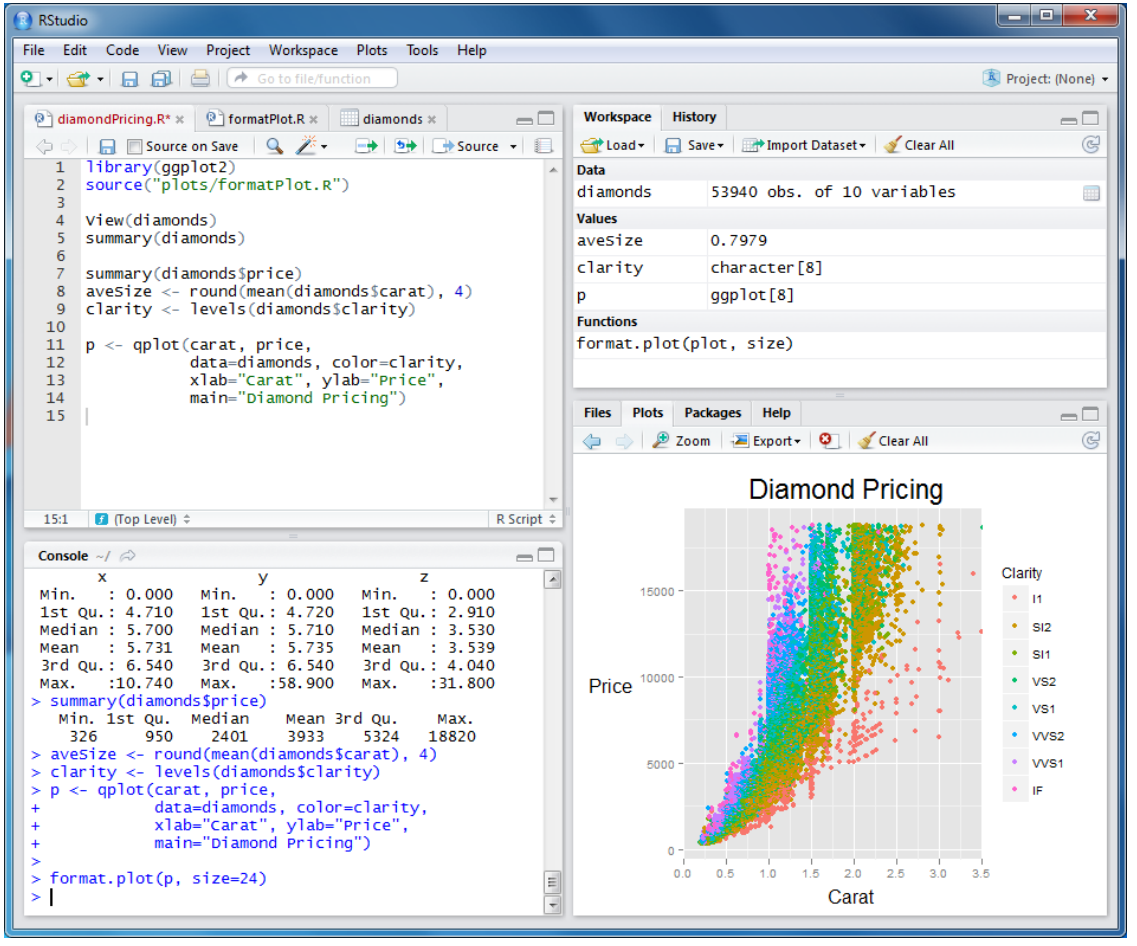
R是什么?

R is a language and environment for statistical computing and graphics.It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

> The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

Rstudio IDE


RStudio
File Edit Code View Project Workspace Plots Tools Help
Project: (None)
diamondPricing.R* formatPlot.R diamonds
Source on Save Source
1 library(ggplot2)
2 source("plots/formatPlot.R")
3
4 View(diamonds)
5 summary(diamonds)
6
7 summary(diamonds$price)
8 aveSize <- round(mean(diamonds$carat), 4)
9 clarity <- levels(diamonds$clarity)
10
11 p <- qplot(carat, price,
12 data=diamonds, color=clarity,
13 xlab="Carat", ylab="Price",
14 main="Diamond Pricing")
15
15:1 (Top Level) R Script
Console ~/
x y z
Min. : 0.000 Min. : 0.000 Min. : 0.000
1st Qu.: 4.710 1st Qu.: 4.720 1st Qu.: 2.910
Median : 5.700 Median : 5.710 Median : 3.530
Mean : 5.731 Mean : 5.735 Mean : 3.539
3rd Qu.: 6.540 3rd Qu.: 6.540 3rd Qu.: 4.040
Max. :10.740 Max. :58.900 Max. :31.800
> summary(diamonds$price)
Min. 1st Qu. Median Mean 3rd Qu. Max.
326 950 2401 3933 5324 18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+ data=diamonds, color=clarity,
+ xlab="Carat", ylab="Price",
+ main="Diamond Pricing")
>
> format.plot(p, size=24)
>
Workspace History
Load Save Import Dataset Clear All
Data
diamonds 53940 obs. of 10 variables
Values
aveSize 0.7979
clarity character[8]
p ggplot[8]
Functions
format.plot(plot, size)
Files Plots Packages Help
Zoom Export Clear All
Diamond Pricing
Price
15000
10000
5000
0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Carat
Clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF

R能做什么?

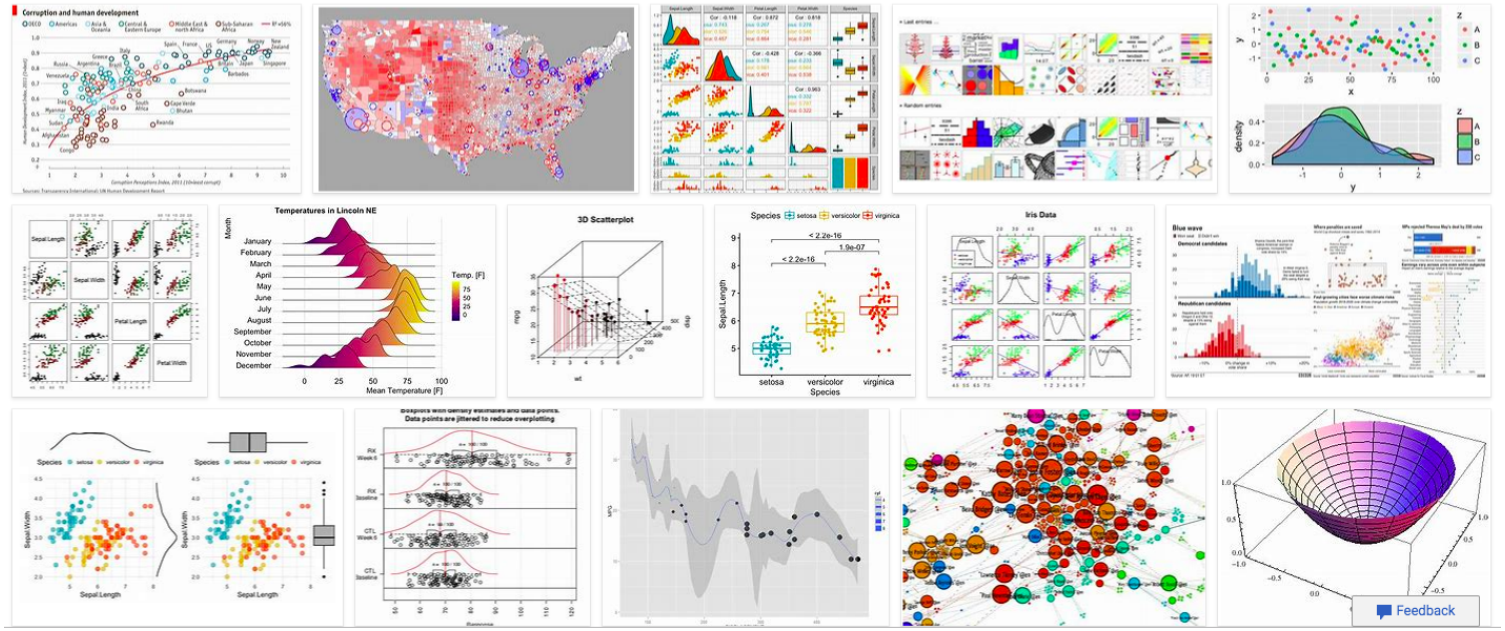
这个问题问得好? 

答案是:

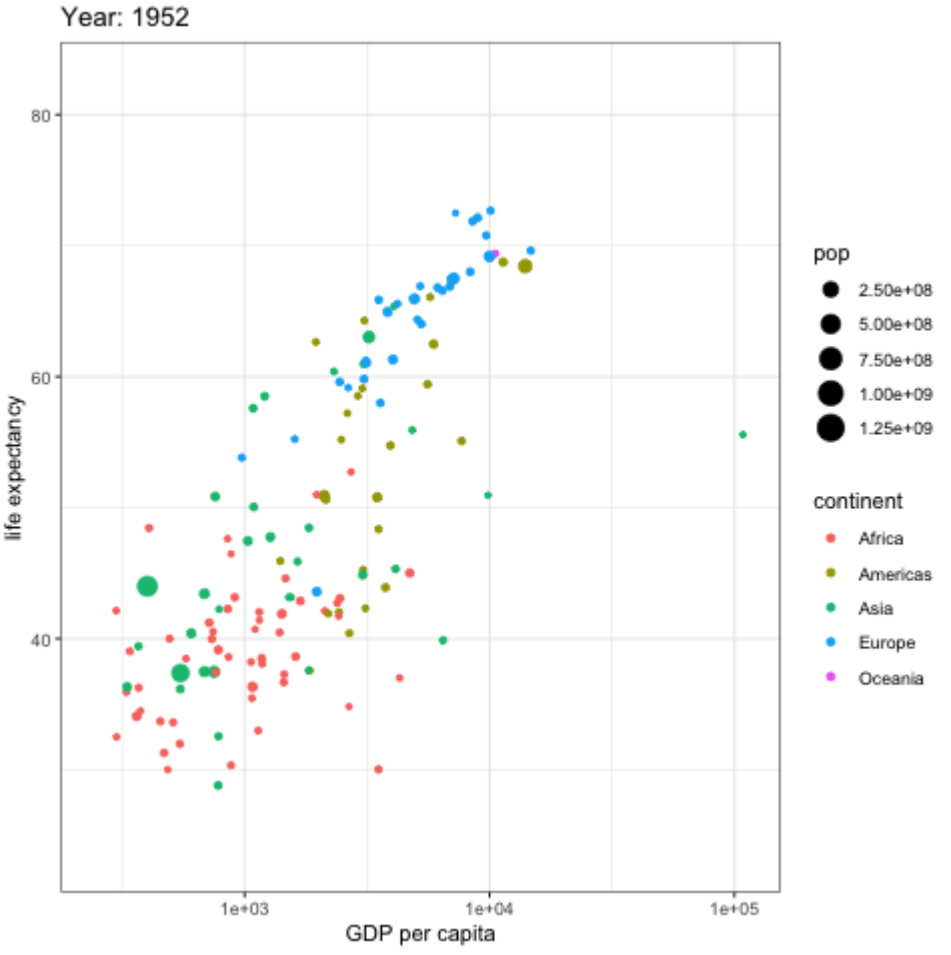
在数据分析领域，无所不能

就算现在不能，以后也一定能

用R绘制的图形


Corruption and human development
Temperatures in Lincoln NE
Mean Temperature [F]
3D Scatterplot
Iris Data
Species
setosa
versicolor
virginica
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Data points are jittered to reduce overplotting
Feedback

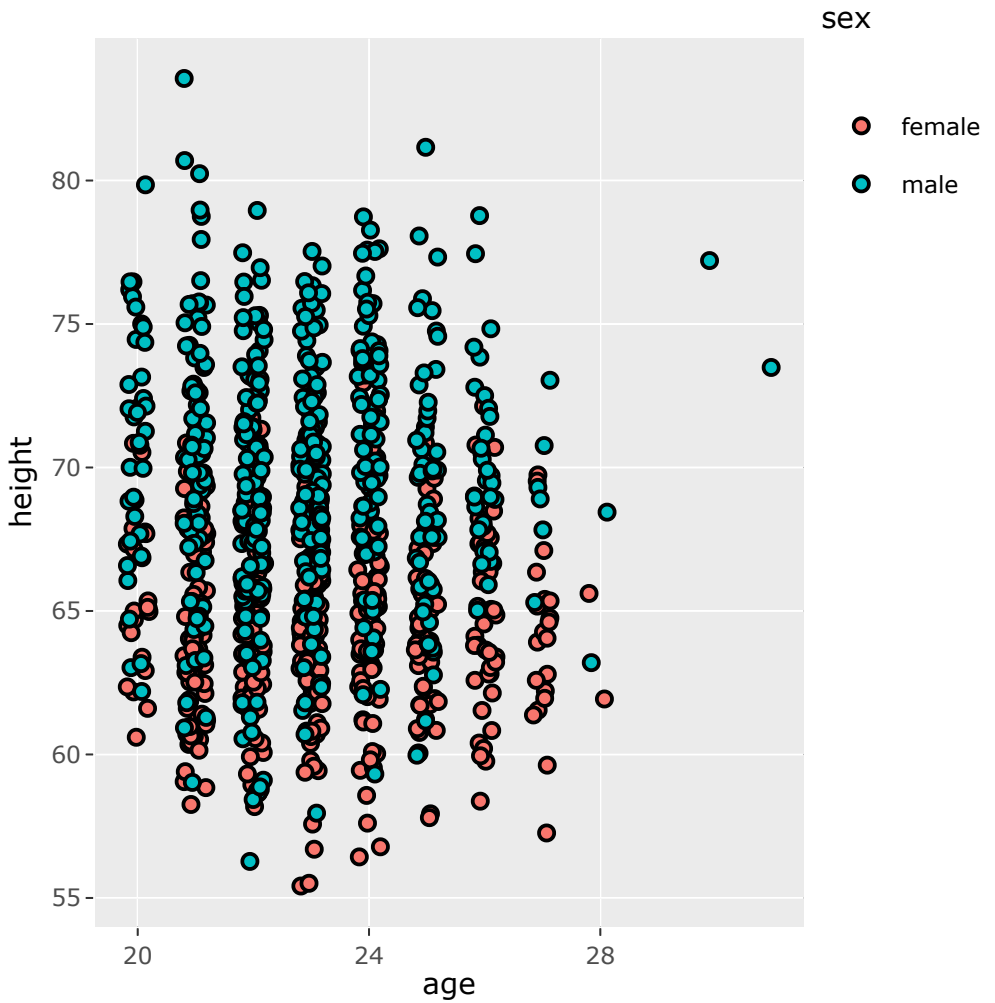
用R绘制的图形


Year: 1952
80
60
40
life expectancy
1e+03
1e+04
1e+05
GDP per capita
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09
continent
Africa
Americas
Asia
Europe
Oceania

用R绘制的图形

用R绘制的图形

用R绘制的图形



各种图形

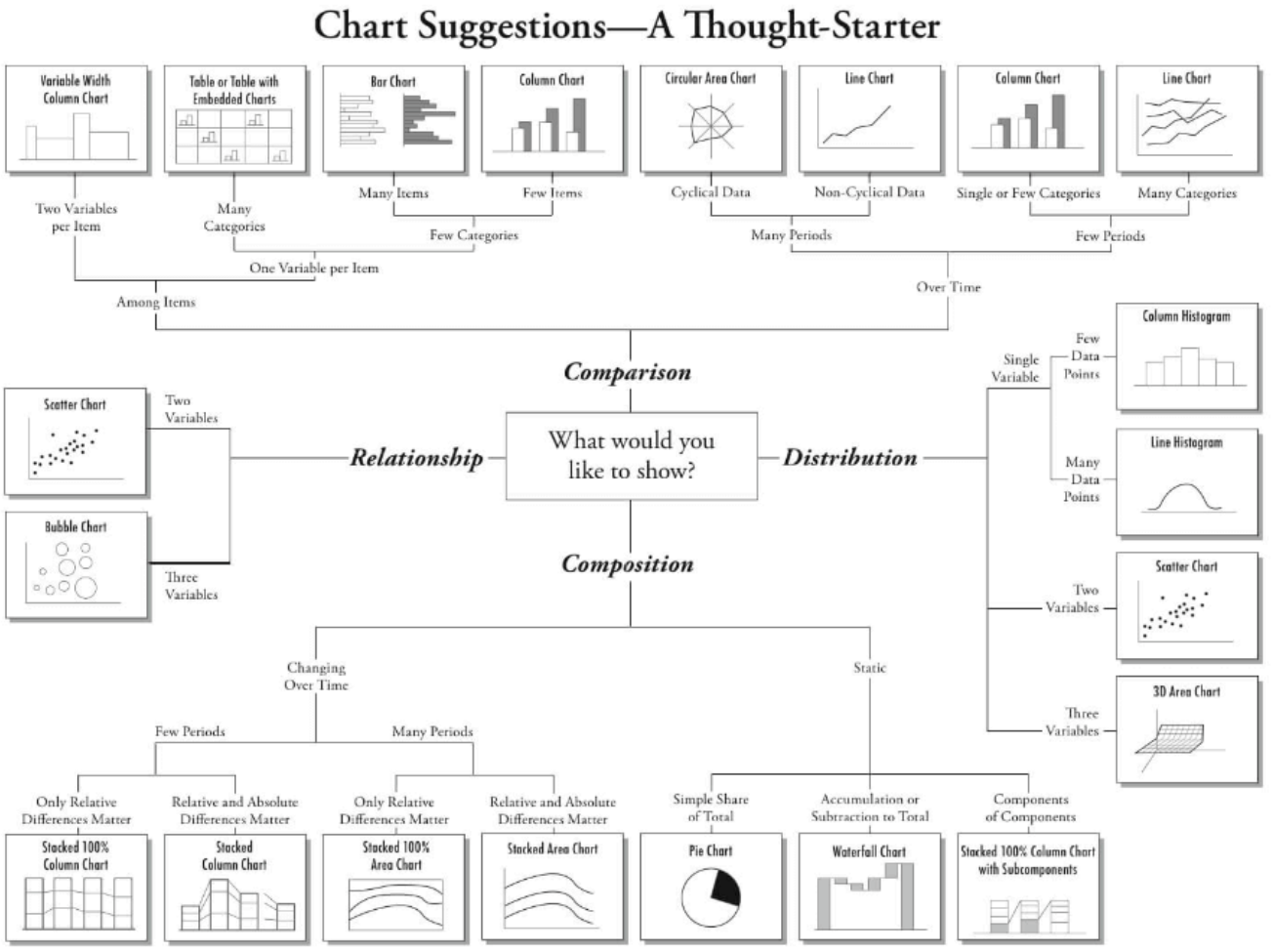


Chart Suggestions—A Thought-Starter
Variable Width Column Chart
Table or Table with Embedded Charts
Bar Chart
Column Chart
Circular Area Chart
Line Chart
Column Chart
Line Chart
Two Variables per Item
Many Categories
Many Items
Few Items
Few Categories
Cyclical Data
Non-Cyclical Data
Many Periods
Single or Few Categories
Many Categories
Few Periods
One Variable per Item
Among Items
Over Time
Comparison
Scatter Chart
Two Variables
Bubble Chart
Three Variables
Relationship
What would you like to show?
Distribution
Composition
Single Variable
Few Data Points
Many Data Points
Column Histogram
Line Histogram
Two Variables
Scatter Chart
Three Variables
3D Area Chart
Changing Over Time
Static
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Stacked 100% Column Chart
Stacked Column Chart
Stacked 100% Area Chart
Stacked Area Chart
Pie Chart
Waterfall Chart
Stacked 100% Column Chart with Subcomponents

R 代码

```r
#做个计算题
5+5+7+9+10
```

```
# [1] 36
```

```r
# 生成十个数字
rnorm(10,mean = 0,sd = 1)
```

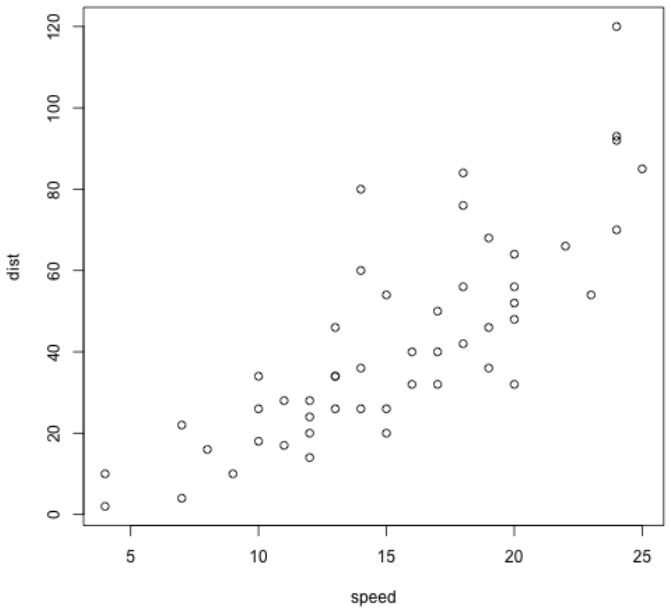
```
#  [1]  1.30709064  0.03676991 -0.85136309  0.19612334 -0.02014146
#  [6]  1.42881203  0.86934421 -1.42369295  0.11197427  0.55488916
```

```r
# 处理文字
dojutsu = c('地爆天星', '天照', '加具土命', '神威', '須佐能乎', '無限月読')
grep('天', dojutsu, value = TRUE)
```

```
# [1] "地爆天星" "天照"
```

R 基本绘图

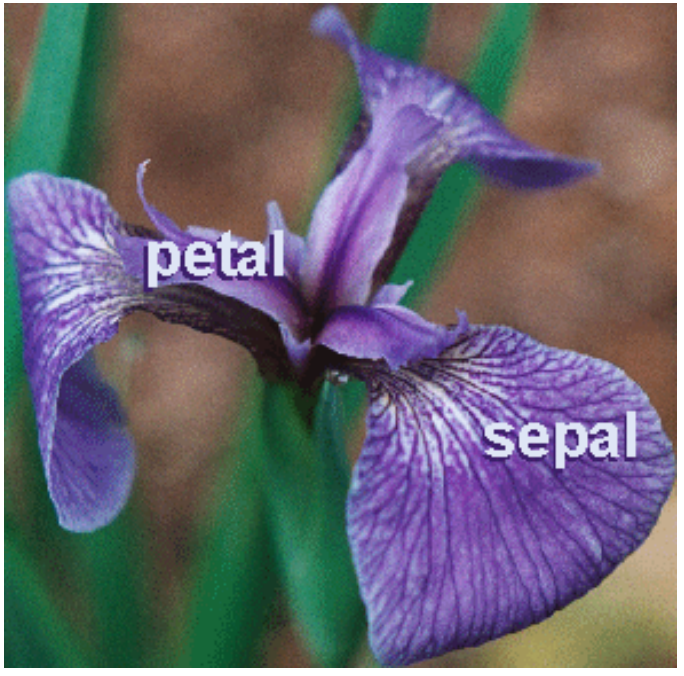
```
plot(cars)
```



R 绘图

Iris Data Set（鸢尾属植物数据集）

Iris Data Set（鸢尾属植物数据集）首次出现在著名的英国统计学家和生物学家Ronald Fisher 1936年的论文《The use of multiple measurements in taxonomic problems》。



Iris Data Set（鸢尾属植物数据集）

在这个数据集中，包括了三类不同的鸢尾属植物：Iris Setosa，Iris Versicolour，Iris Virginica。每类收集了50个样本，因此这个数据集一共包含了150个样本。

该数据集测量了所有150个样本的4个特征，分别是：

- sepal length（花萼长度）
- sepal width（花萼宽度）
- petal length（花瓣长度）
- petal width（花瓣宽度）

以上四个特征的单位都是厘米（cm）。 通常使用m表示样本量的大小，n表示每个样本所具有的特征数。因此在该数据集中，m=150,n=4

表格显示

Show 6 entries Search:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Showing 1 to 6 of 40 entries

Previous 1 2 3 4 5 6 7 Next

做个练习

三类不同的鸢尾属植物：

- Iris Setosa
- Iris Versicolour
- Iris Virginica

4个特征：

- sepal length（花萼长度）
- sepal width（花萼宽度）
- petal length（花瓣长度）
- petal width（花瓣宽度）

> 问题：请按种类分别求各特征的均值

让R来帮你完成

```
library(dplyr)
datatable(iris %>%
  group_by(Species) %>%
  summarise_all(mean))
```

Show 10 entries Search:

	Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	setosa	5.006	3.428	1.462	0.246
2	versicolor	5.936	2.77	4.26	1.326
3	virginica	6.588	2.974	5.552	2.026

Showing 1 to 3 of 3 entries Previous 1 Next

数据分析的基本流程

谢谢聆听!

下次请带上电脑，我们一起感受R的乐趣！

毛益民

浙江工商大学公共管理学院