

R简介

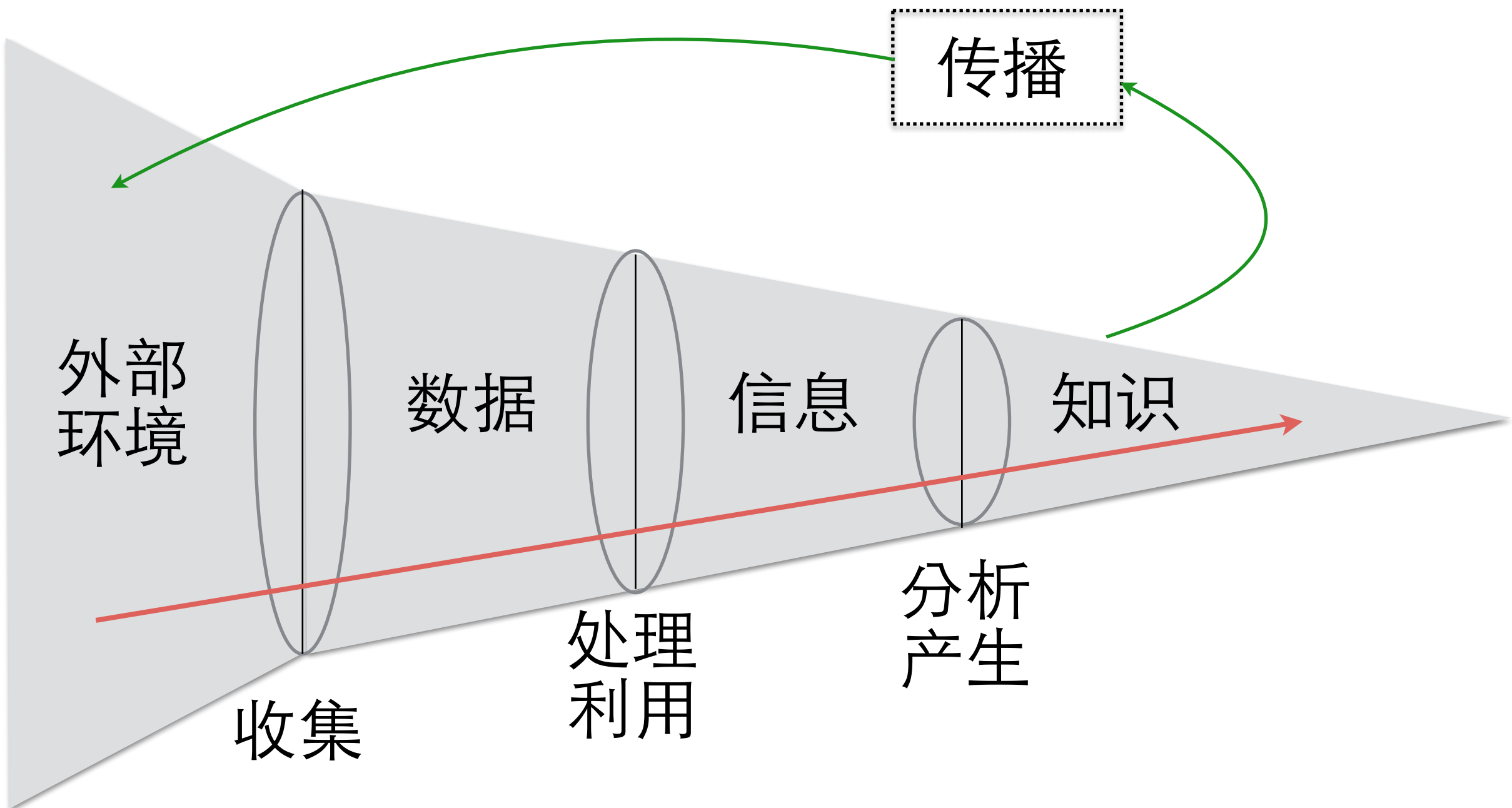


- 数据分析简介
- R软件简介
- 练习

数据分析简介

- 反应原始事物的一组定性或定量变量的集合
 - ✳ 可以测量、收集、报告和分析
 - ✳ 原始数据 vs 处理过数据
- 计算机发明之前，数据的产生和共享仅限于很少的形式
 - ✳ 手写书信 vs Email
 - ✳ 胶卷 vs 数码相机
 - ✳ 打印书籍 vs 电子书

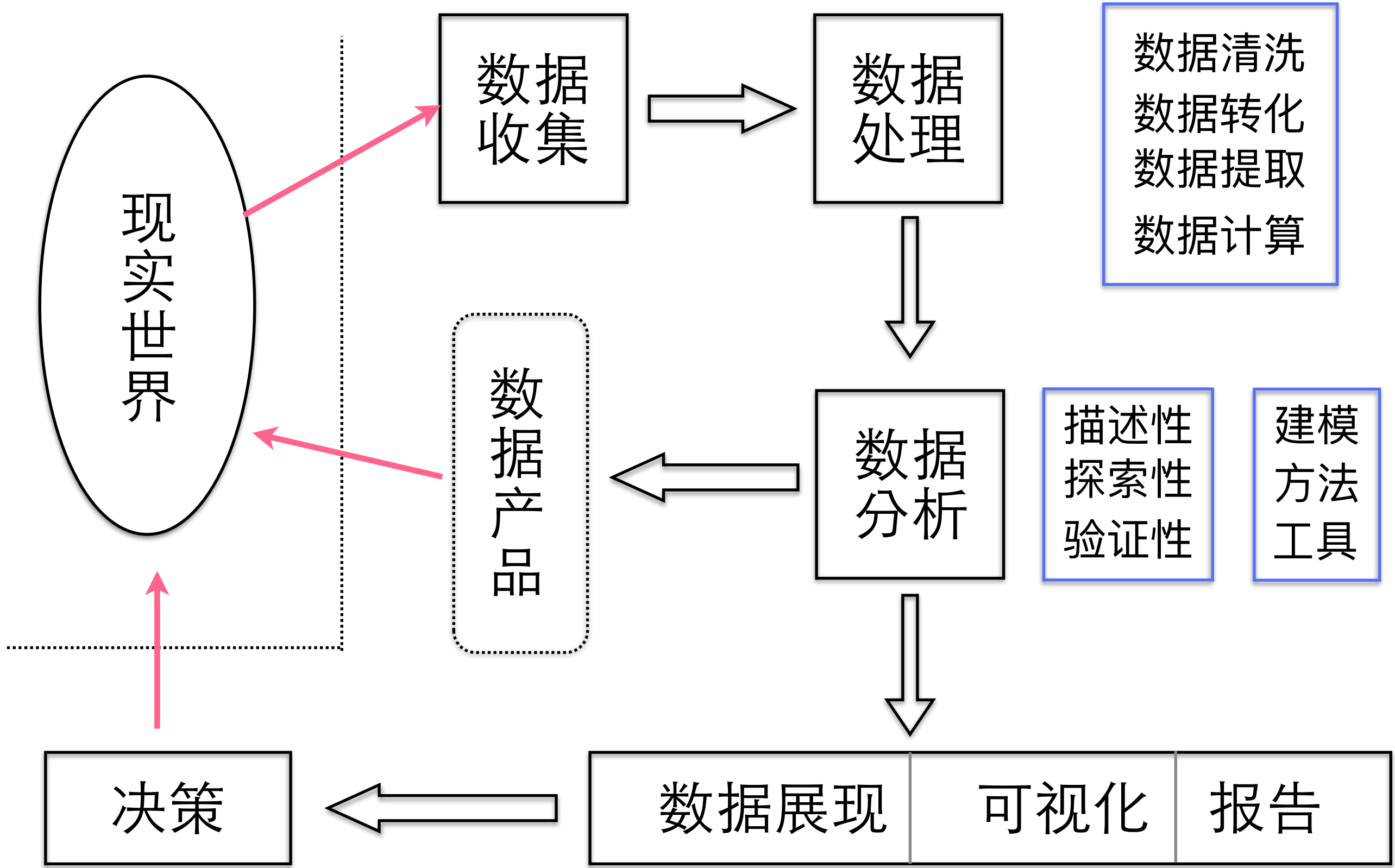
数据、信息、知识



- 对数据进行分析
- 用适当的统计分析方法对收集来的大量数据进行分析，将他们加以汇总和理解消化，以求最大化的开发数据的功能，发挥数据的作用。
 - ✱ 是为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。
 - ✱ 数据也称观测值，是通过实验、测量、观察、调查等方式获取的结果，常常以数量的形式展现出来。

- 目的
 - ✱ 把隐藏在一大批看似杂乱无章的数据背后的信息集中和提炼出来，总结出研究对象的内在规律。
 - ✱ 实际工作中，数据分析能够帮助管理者进行判断和决策，以便采取适当的策略与行动。
 - ✱ 统计学：描述性数据分析、探索性数据分析（发现新特征）、验证性数据分析（验证已有假设）。
- 作用：现状分析、原因分析、预测分析

数据分析过程

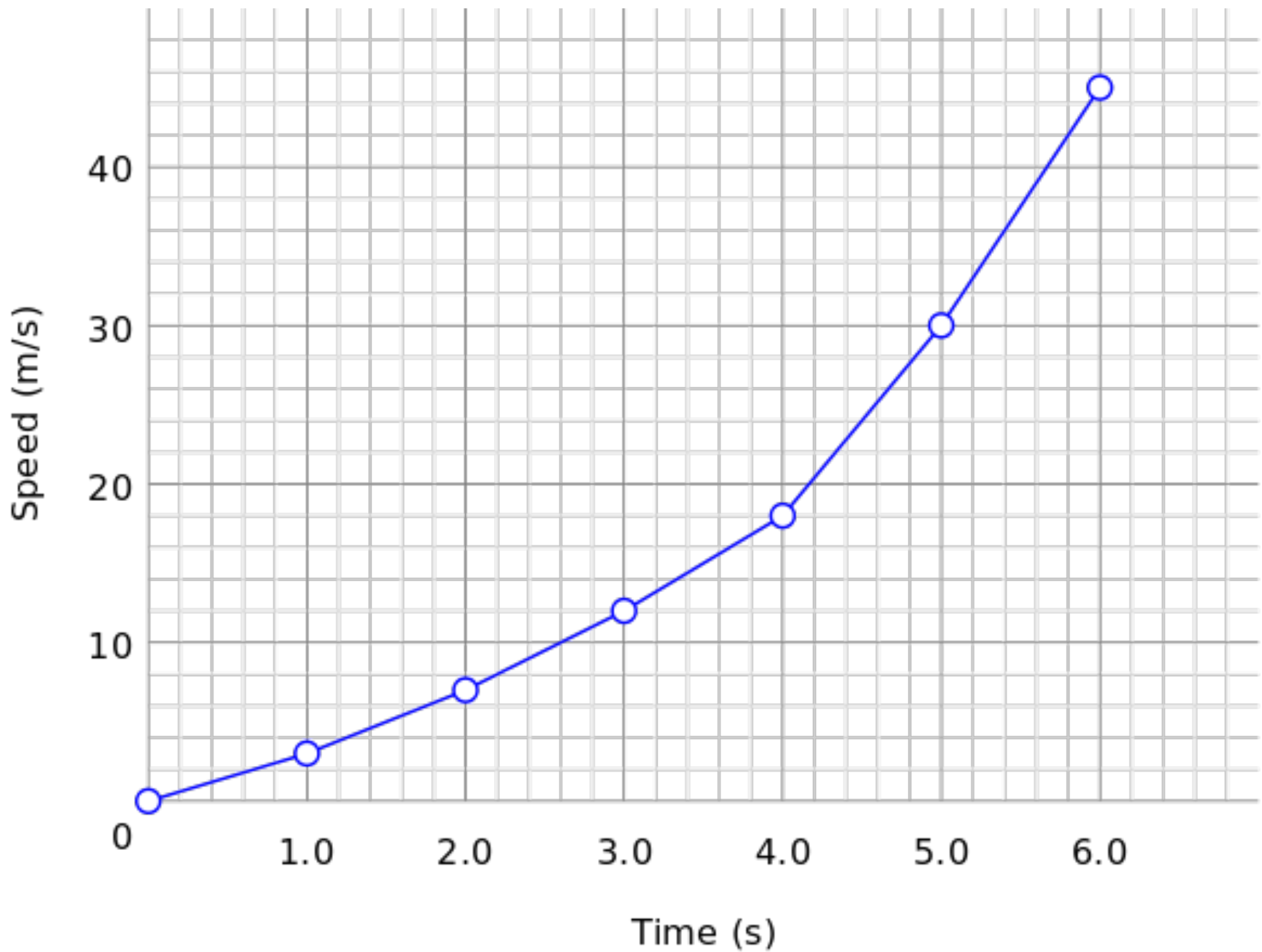


- 检索一个值
- 过滤
- 计算派生值
- 发现极值
- 排序
- 确定边界
- 刻画分布
- 发现异常

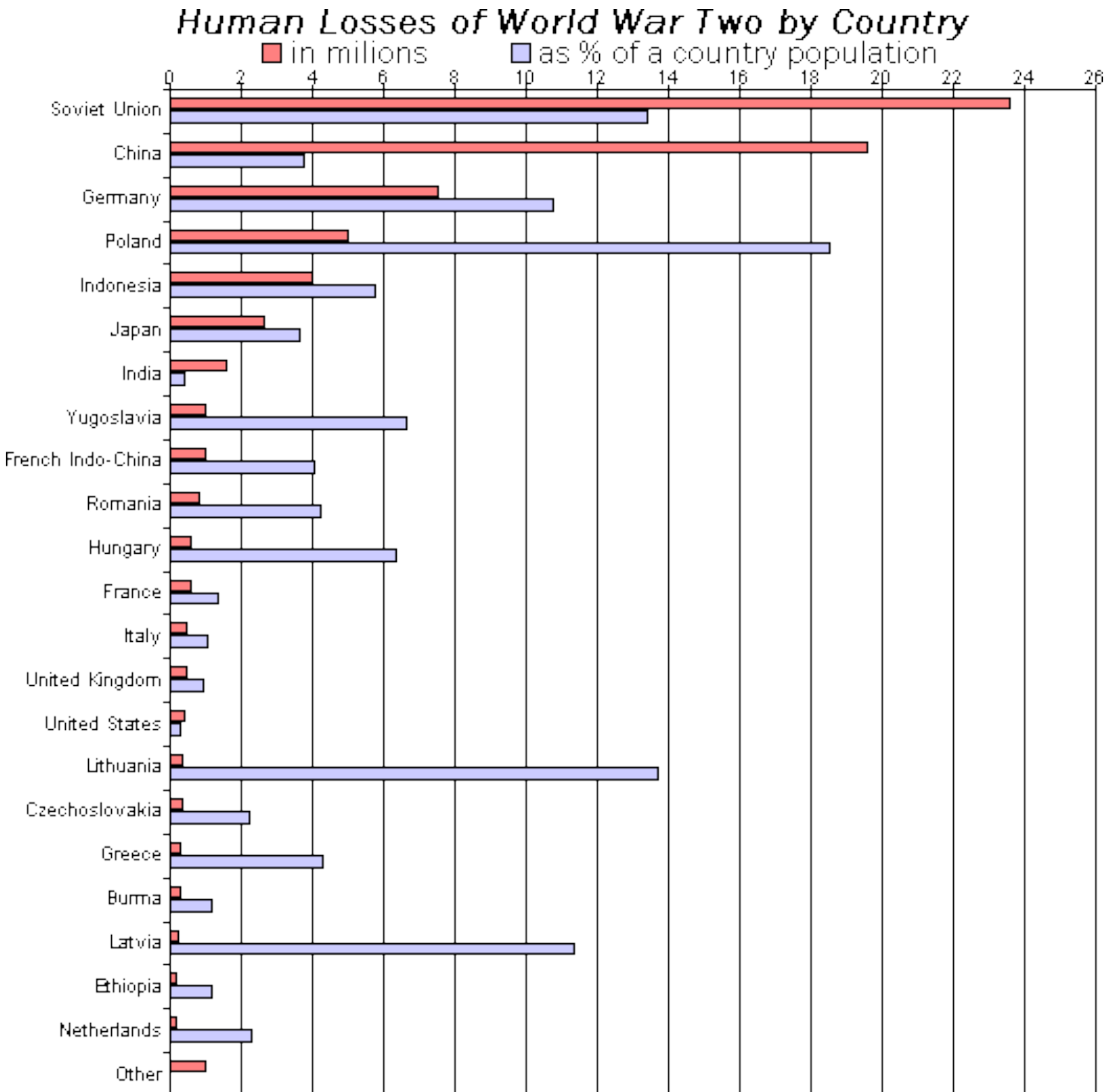
- 平均值、最大值、最小值、众数
- 全距、平均偏差、标准差、方差
- 矩、偏度、峰度
- 二项式、正态、泊松、柏努力
- 时间、地域
-

目的明确!

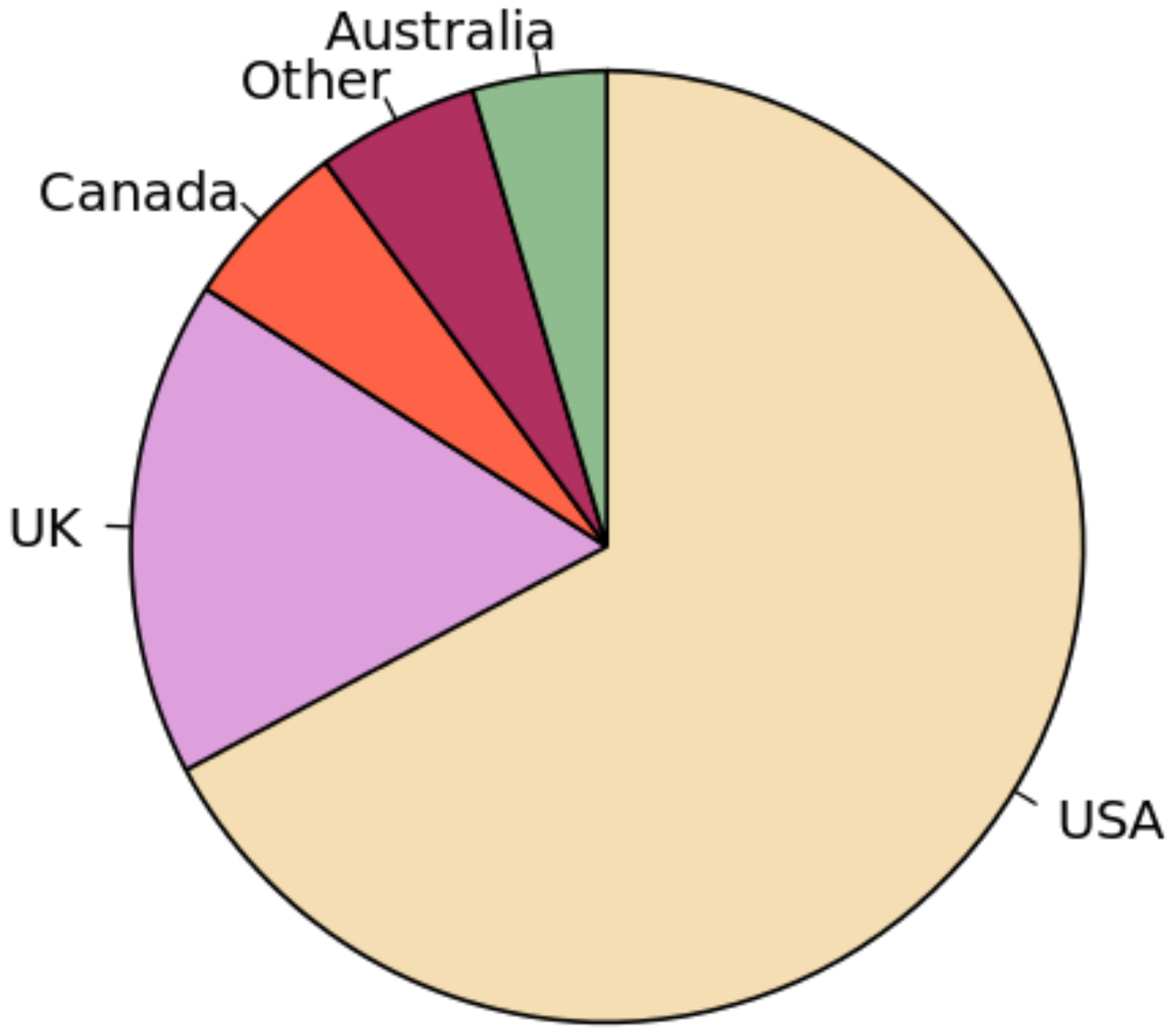
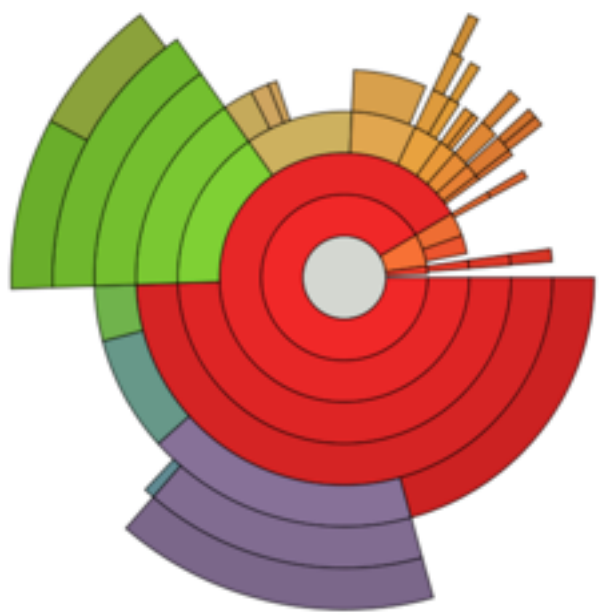
折线图



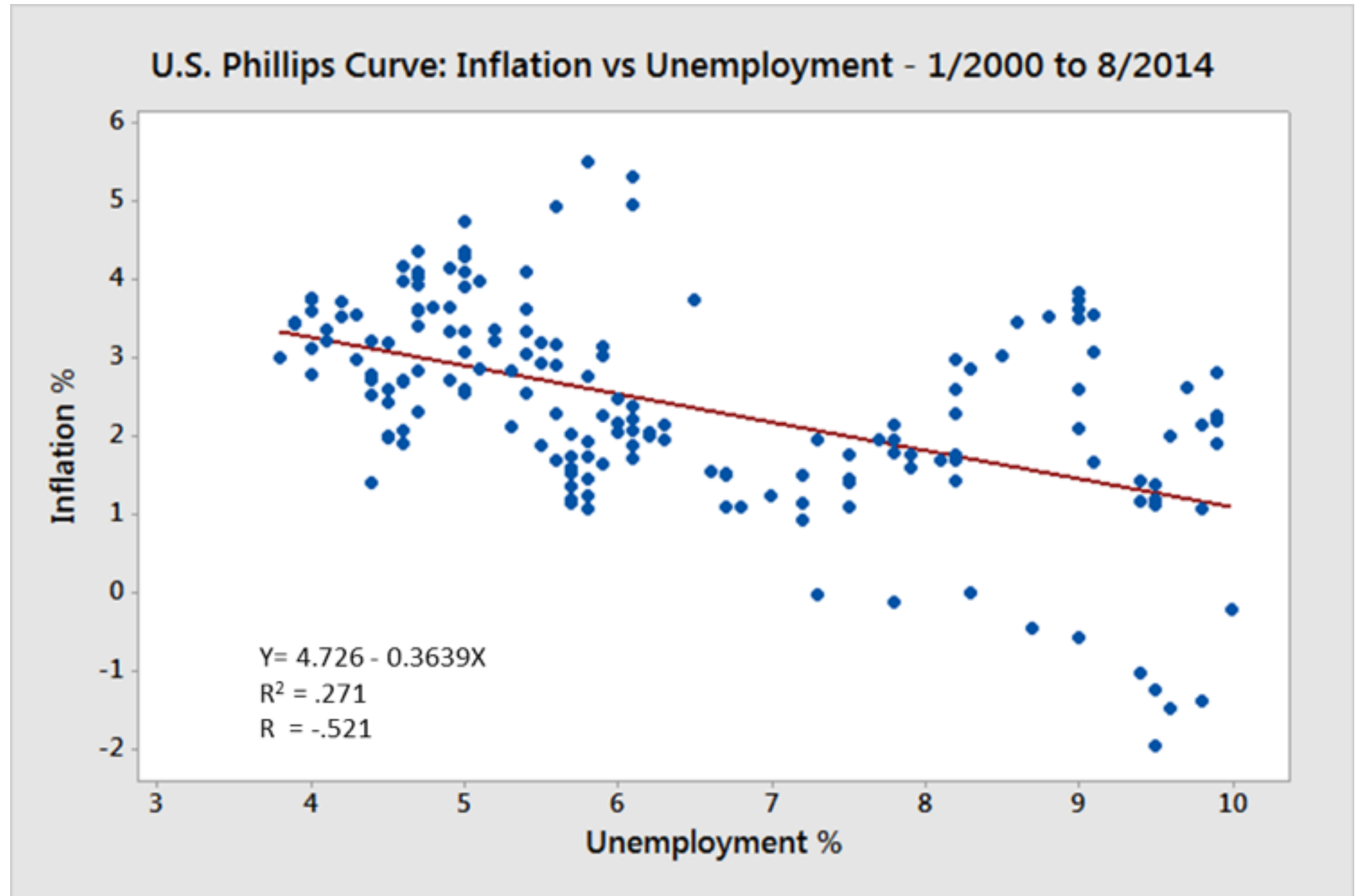
条形图



饼图

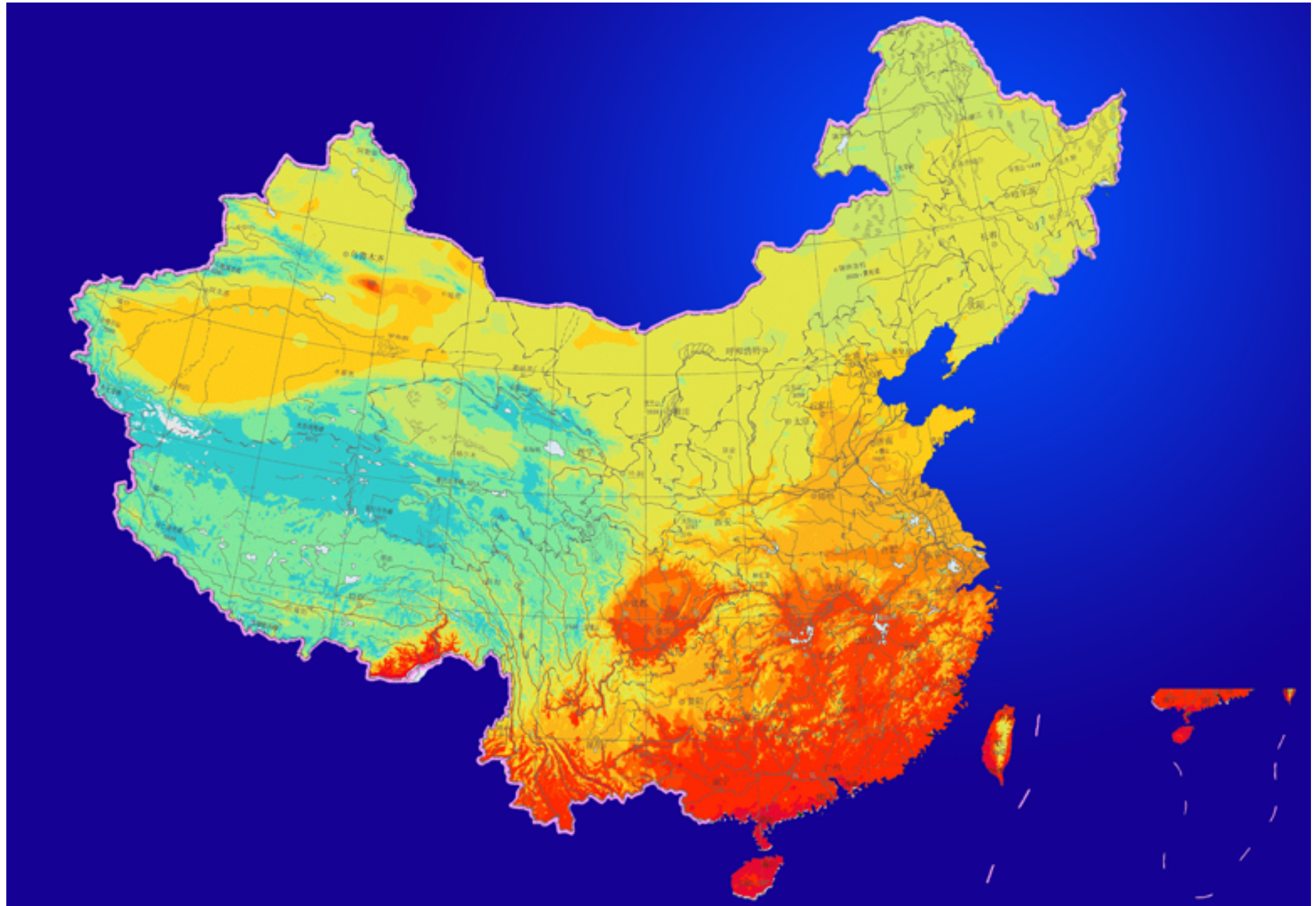


散点图



Source Data: FRED Database
Inflation: CPI for All Urban Consumers

统计地图



热图



Java



C



C++



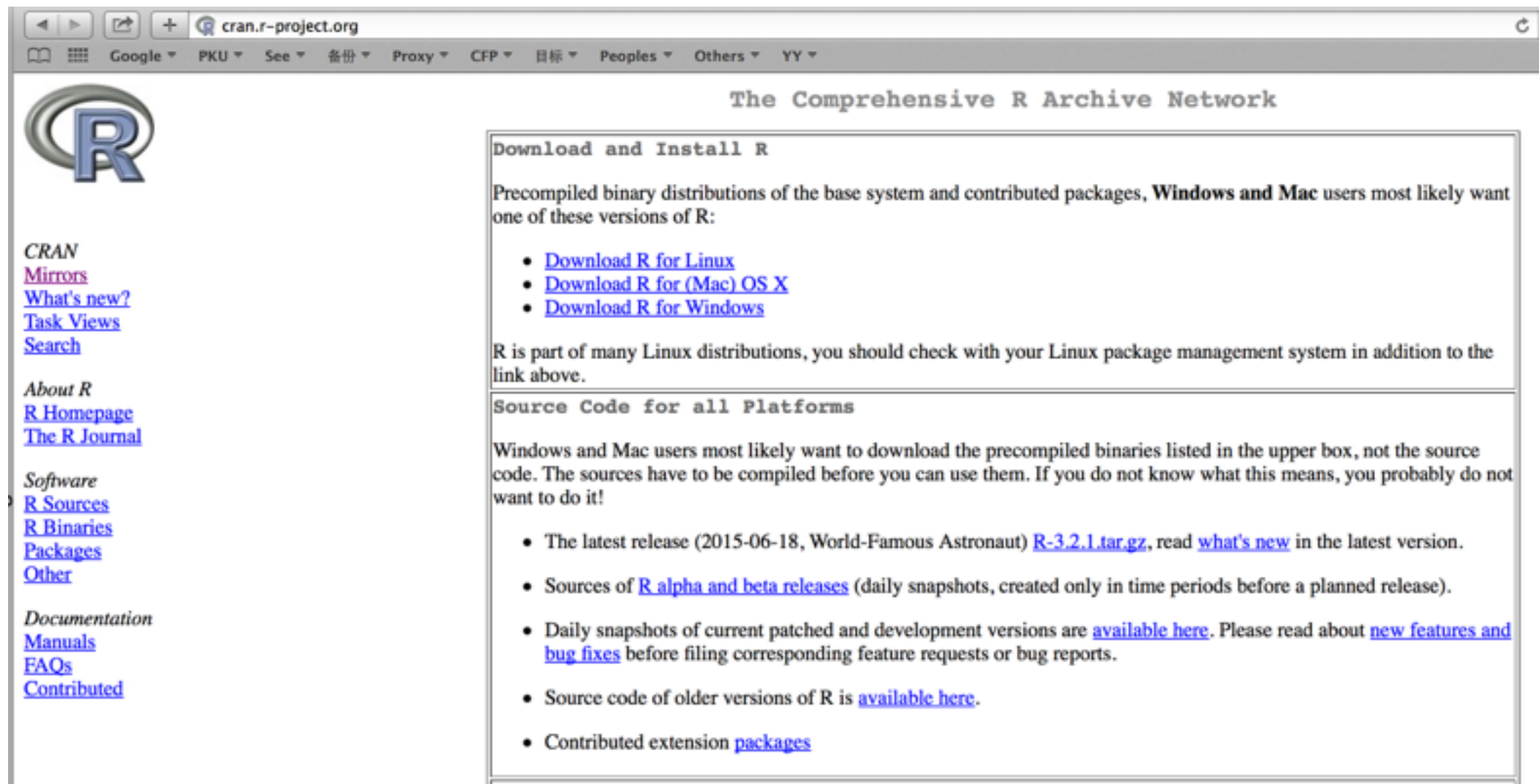
- 对整个数据分析过程的一个总结与呈现
- 把数据分析的起因、过程、结果和建议完整呈现出来
- 好的分析框架，图文并茂、层次明晰
- 明确的结论，建议和解决方案
- 分析背景和目的、分析思路、分析正文、总结和建议

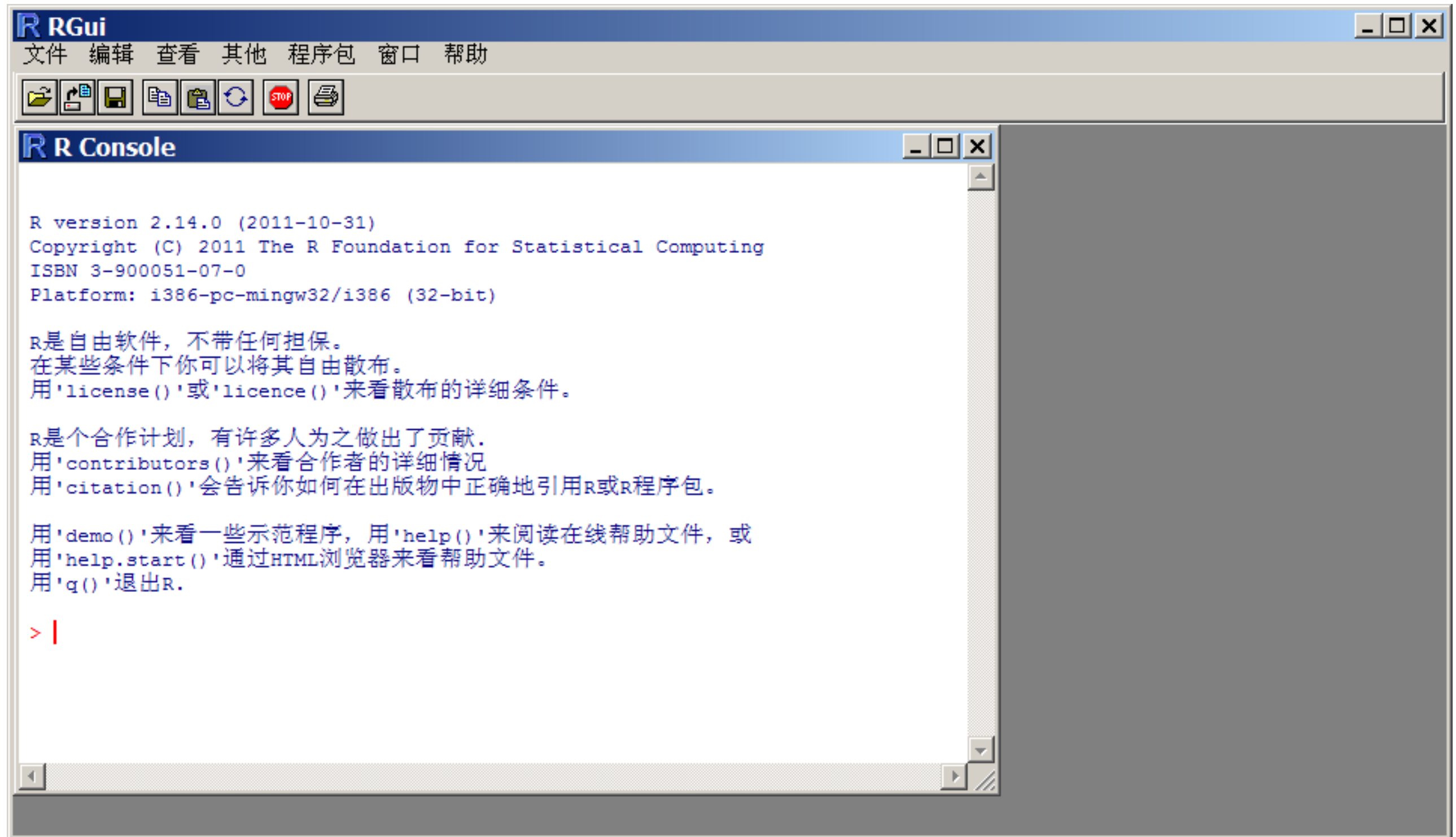
R软件简介

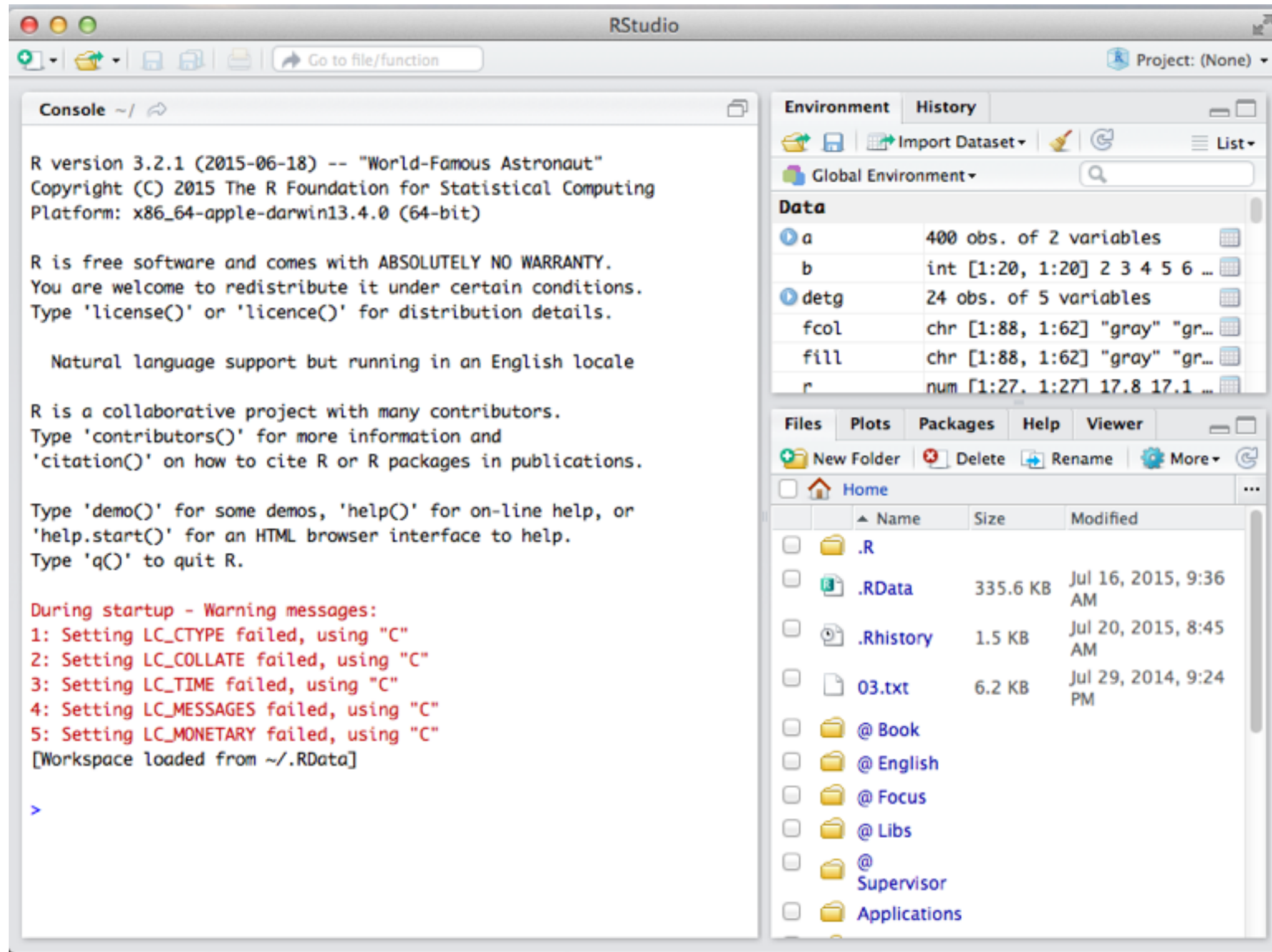
- R是S语言的一种实现。S语言是由 AT&T 贝尔实验室开发的一种用来进行数据探索、统计分析、画图的解释型语言。最初S语言的实现版本主要是S-PLUS。S-PLUS是一个商业软件，它基于S语言，并由MathSoft公司的统计科学部进一步完善。
- Auckland大学的Robert Gentleman 和 Ross Ihaka 及其他志愿人员开发了一个R系统。R的使用与S-PLUS有很多类似之处，两个软件有一定的兼容性。
- R是用于统计分析、绘图的语言和操作环境。R是属于GNU系统的一个自由、免费、源代码开放的软件，它是一个用于统计计算和统计制图的优秀工具。

- 自由软件，免费
- 功能强大，不弱于同类软件
- 和其他语言、数据库有非常好的接口
- 容易扩展，新功能更新更快，网上资源丰富
- 多平台支持，多种GUI支持
- 面向对象编程语言，语言简洁高效，顶尖水平的制图能力
- 连贯完整的数据分析中间工具，有效的数据处理和保存机制，可进行交互式数据分析
- <<<<<<<<<<< 学习曲线较为陡峭

- 官方网站: <http://cran.r-project.org> 免费获取
- Windows、Mac、Linux
- 包 (package) 增强R功能







- 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个命令，也可以一次性执行写在脚本文件的一组命令
- 输入的指令成为**表达式**，R的解析器读入解释这些指令
- R的功能多数由程序内置函数和用户自编函数提供，基本函数默认可以直接使用，其余函数可以按需加载
- R语句由函数和赋值构成，R使用“`<-`”，而不是传统的“`=`”作为赋值符号
- 注释由符号“`#`”开头，但不支持多行

表1-1 10名婴儿的月龄和体重

年龄 (月)	体重 (kg)	年龄 (月)	体重 (kg)
01	4.4	09	7.3
03	5.3	03	6.0
05	7.2	09	10.4
02	5.2	12	10.2
11	8.5	03	6.1

```
> age <- c(1, 3, 5, 2, 11, 9, 3, 9, 12, 3)
```

```
> weight <- c(4.4, 5.3, 7.2, 5.2, 8.5,  
             7.3, 6, 10.4, 10.2, 6.1)
```

```
> mean(weight)
```

```
> sd(weight)
```

```
> cor(age, weight)
```

```
> plot(age, weight)
```

Example 0001

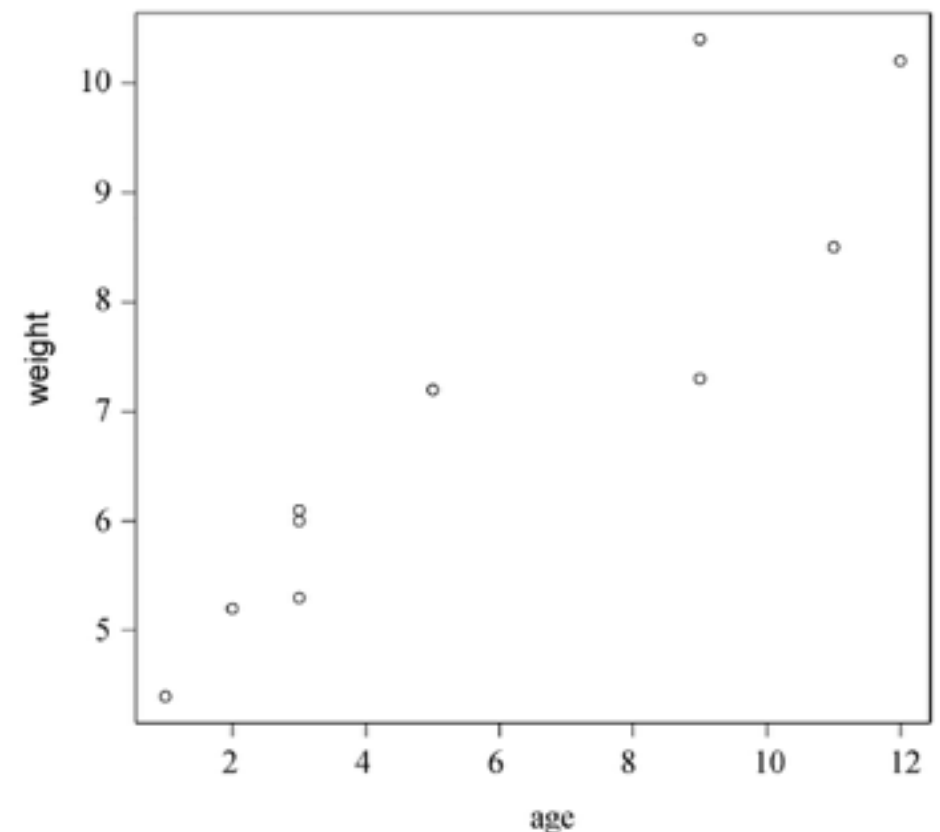
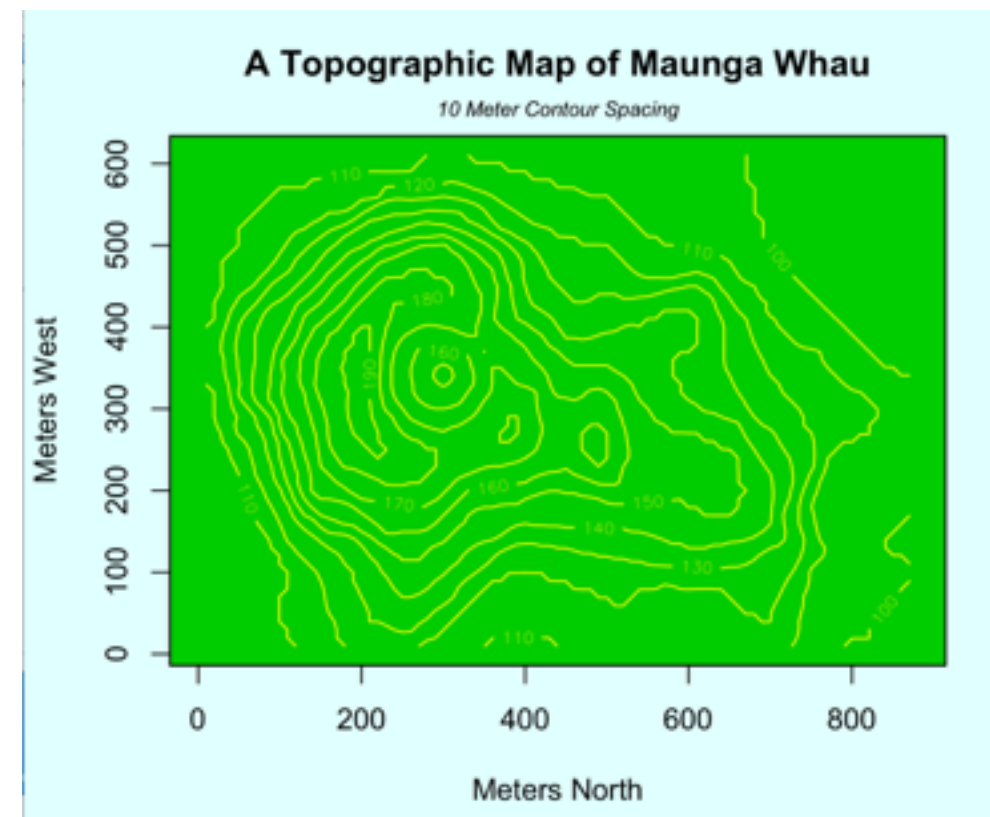
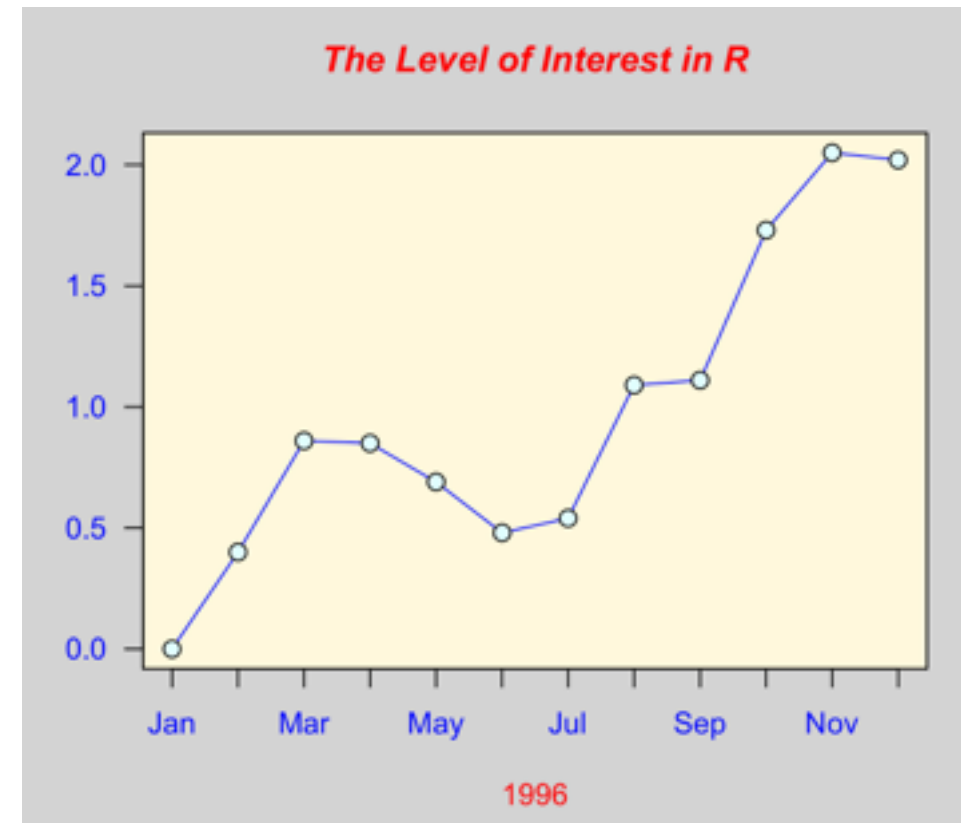
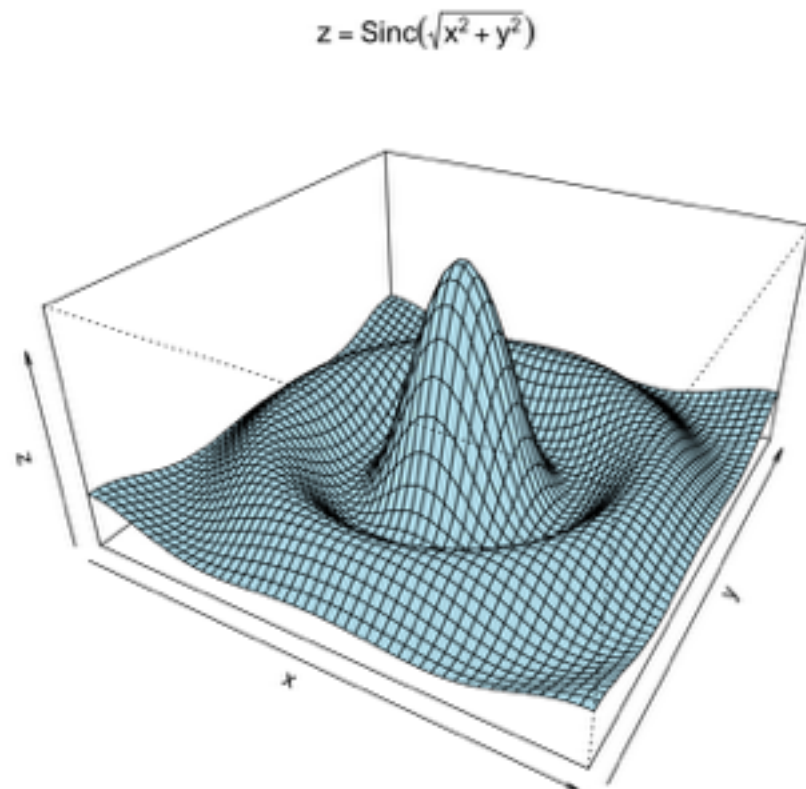


图1-4 婴儿体重 (千克) 和年龄 (月) 的散点图

- `demo(graphics)`
- `demo(image)`
- `demo(Hershey)`
- `demo(persp)`



- 包是R函数、数据、预编译代码组成的集合
- R一般自带8个默认包
- `library()`
- `.libPaths()`
- `search()`
- `install.packages()`
- `update.packages()`
- `help(package="packagename")`

函 数	功 能
<code>getwd()</code>	显示当前的工作目录
<code>setwd("mydirectory")</code>	修改当前的工作目录为 <code>mydirectory</code>
<code>ls()</code>	列出当前工作空间中的对象
<code>rm(objectlist)</code>	移除（删除）一个或多个对象
<code>help(options)</code>	显示可用选项的说明
<code>options()</code>	显示或设置当前选项
<code>history(#)</code>	显示最近使用过的#个命令（默认值为25）
<code>savehistory("myfile")</code>	保存命令历史到文件 <code>myfile</code> 中（默认值为 <code>.Rhistory</code> ）
<code>loadhistory("myfile")</code>	载入一个命令历史文件（默认值为 <code>.Rhistory</code> ）
<code>save.image("myfile")</code>	保存工作空间到文件 <code>myfile</code> 中（默认值为 <code>.RData</code> ）
<code>save(objectlist, file="myfile")</code>	保存指定对象到一个文件中
<code>load("myfile")</code>	读取一个工作空间到当前会话中（默认值为 <code>.RData</code> ）
<code>q()</code>	退出R。将会询问你是否保存工作空间

表1-4 用于保存图形输出的函数

函 数	输 出
<code>pdf("filename.pdf")</code>	PDF文件
<code>win.metafile("filename.wmf")</code>	Windows图元文件
<code>png("filename.png")</code>	PBG文件
<code>jpeg("filename.jpg")</code>	JPEG文件
<code>bmp("filename.bmp")</code>	BMP文件
<code>postscript("filename.ps")</code>	PostScript文件

- `source("filename")`
- `sink("filename")`
 - `append=TRUE`, 将文本添加到文件后
 - `split=TRUE`, 同时输入到屏幕和输出文件
- `dev.off()`

- 大小写错误
- 忘记引号
- 函数调用忘记括号
- Windows环境下路径名使用了“\”
- 使用了一个没有加载包的函数
- R的报错信息很模糊

<code><i>help.start()</i></code>	打开帮护文档首页
<code><i>help("fun"), ?fun</i></code>	查看函数fun的帮助
<code><i>example("fun")</i></code>	函数fun的使用示例
<code><i>data()</i></code>	当前已经加载包中的所有可用示例数据集

练习

- 安装R
 - 安装RStudio
-
- 熟悉R和RStudio的界面和菜单功能
-
- 安装swirl, 完成三个课堂练习 (估计30-60分钟)

`install_course_github("pkusdatanalysis","C1_R_Basis")`

- 输入并执行例子0001
-

- 执行一下前面讲的几个Demo
-

- 熟悉包的发现和安装
-

- 熟悉工作空间
-

- 熟悉输入输出（例子0001执行结果分别或同时送到文件和屏幕）

- 打开帮助文档首页，并查阅其中的“Introduction to R”的第1章和13章。
-
- 安装vcd包。
 - 列出此包中可用的函数和数据集。
 - 载入这个包并阅读数据集Arthritis的描述。
 - 显示数据集Arthritis的内容 (直接输入一个对象的名称将列出它的内容)。
 - 运行数据集Arthritis自带的示例。它基本上显示了接受治疗的关节炎患者较接受安慰剂的患者在病情上有了更多改善。
 - 退出。

- 看RIA第一章
 - 使用帮助看所有出现的函数说明，了解所有函数功能
 - 把第8页、第11页和第12页的三个例子输入到R文件，看执行效果
-
- 看A Introduction to R的第二章和附录A（可选）
-
- 实现一个简单的hello world程序

An Integrated Development Environment for R

Getting Started with

RStudio



O'REILLY®

John Verzani

- 第一章
- 第三章
- 第五章



谢谢!

孙惠平

sunhp@ss.pku.edu.cn