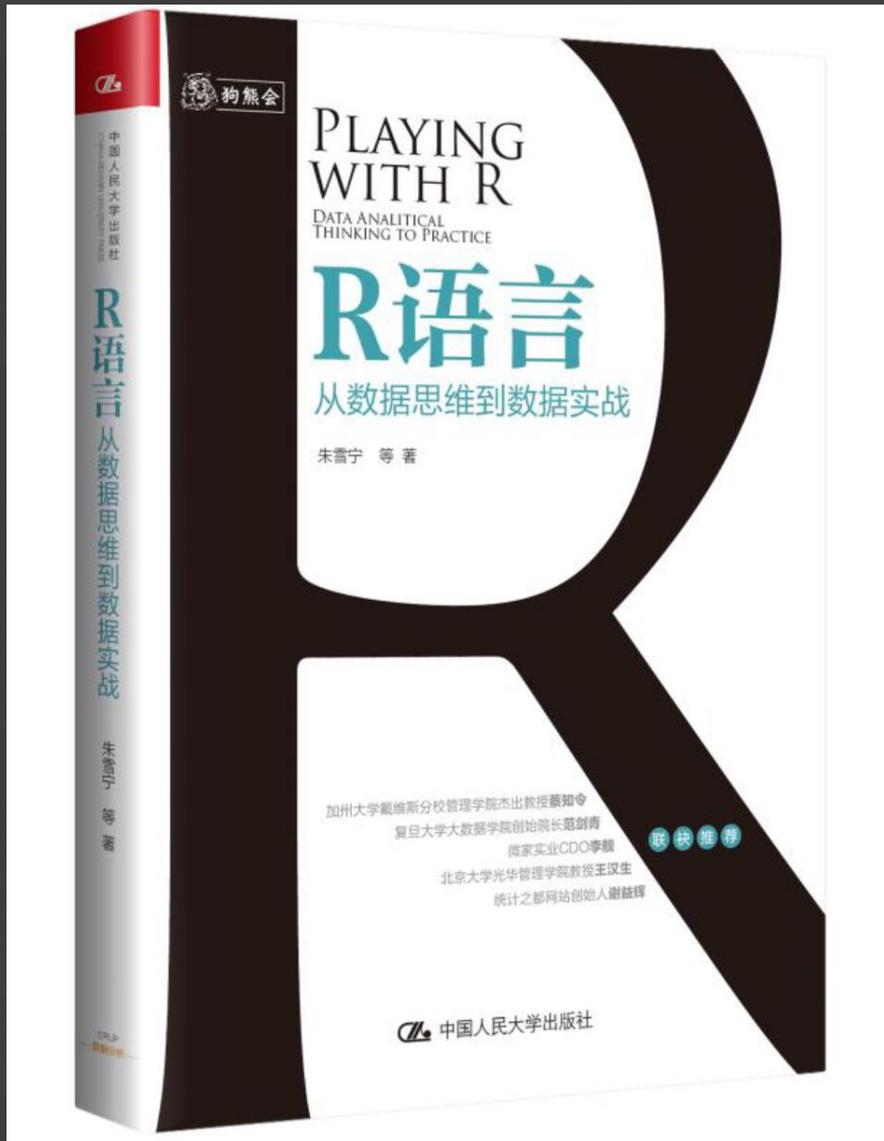


动手学习R语言：R you Ready?

讲师：布丁（朱雪宁） | 狗熊会



宣传一下新书 😊



工欲善其事，必先利其器。

01 Why R?

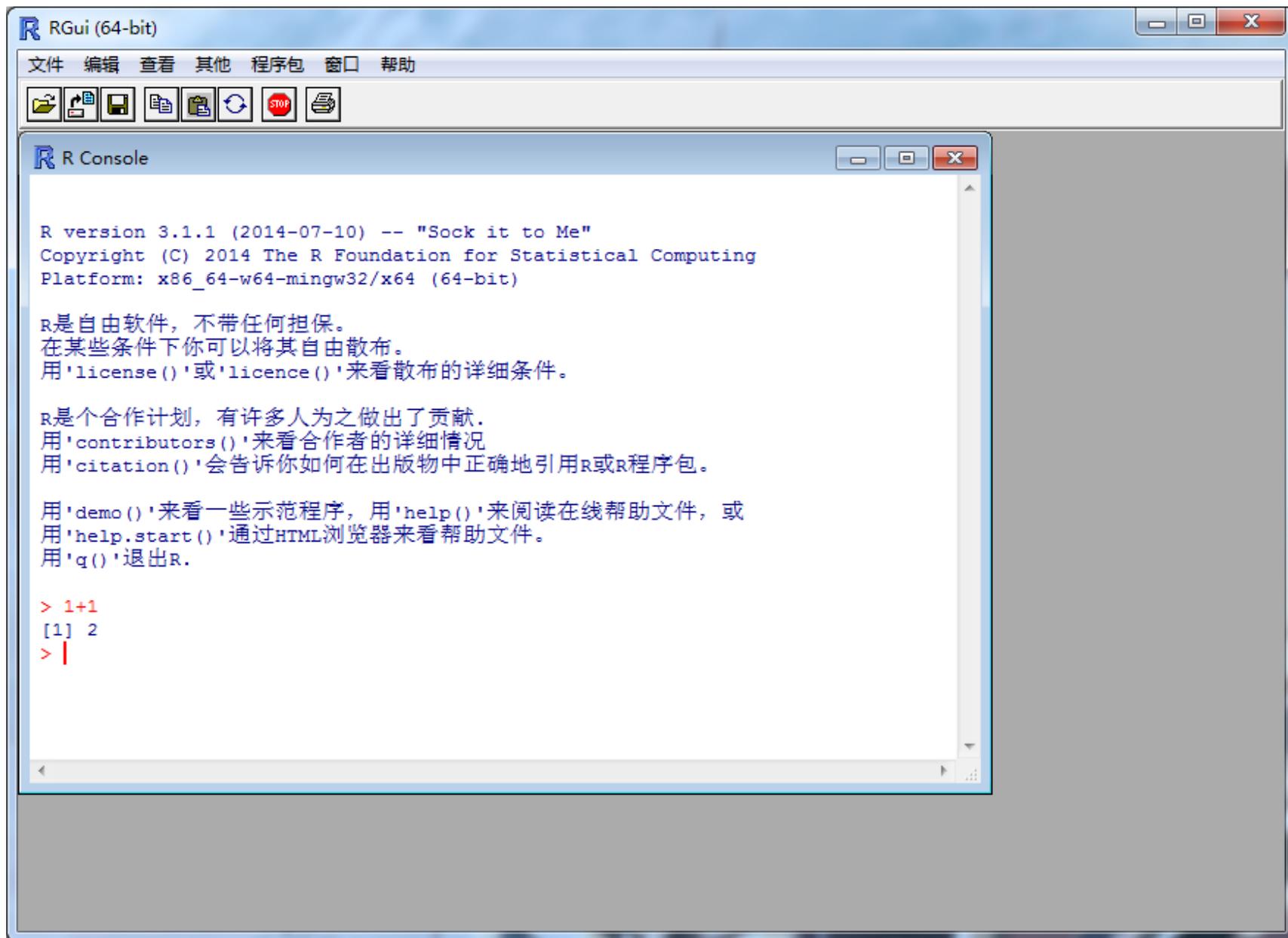
R 的起源

Ross Ihaka



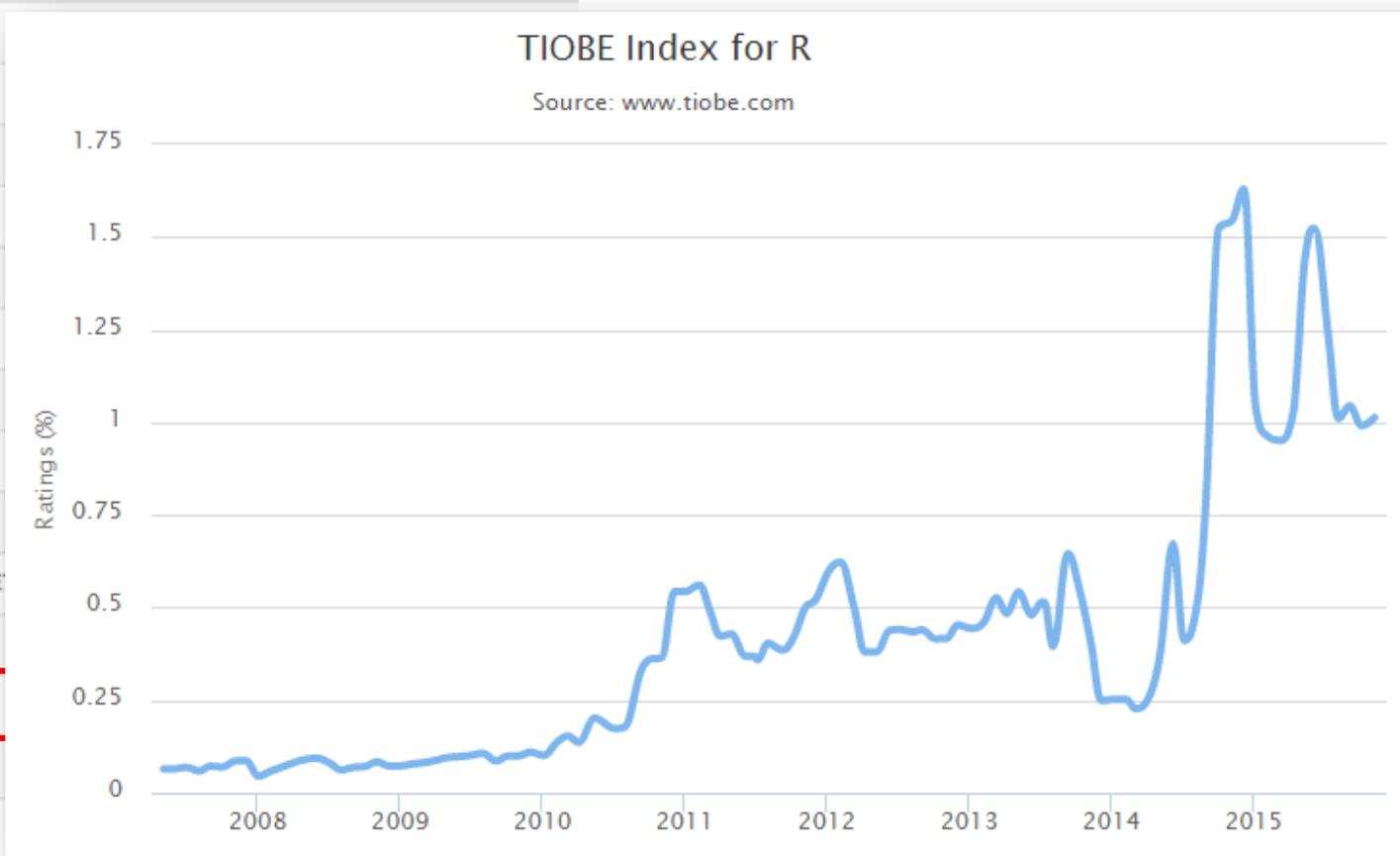
Robert Gentleman





Why R?

Nov 2014	Nov 2013	Change	Programming Language	Ratings	Change
1	1		C		
2	2		Java		
3	3		Objective-C		
4	4		C++		
5	5		C#		
6	6		PHP		
7	8	▲	Python		
8	10	▲	JavaScript		
9	12	▲	Perl		
10	11	▲	Visual Basic .NET		
11	-	▲▲	Visual Basic		
12	31	▲▲▲	R		
13	9	▼▼	Transact-SQL		
14	13	▼	Ruby		
15	17	▲	Delphi/Object Pascal	0.957%	+0.31%



R语言在业界

REVOLUTION
ANALYTICS



Microsoft



R能干啥: 优势篇

- 几乎覆盖整个统计领域的前沿算法 (3700+ 扩展包)
- 开放的源码、强大的社区
- 数据挖掘平台
- 方便的扩展性
 - 连接数据库 (Oracle, MySQL...)
 - 同其他语言互调 (Python、C..)

R能干啥：清洗数据

理想中的数据分析师：

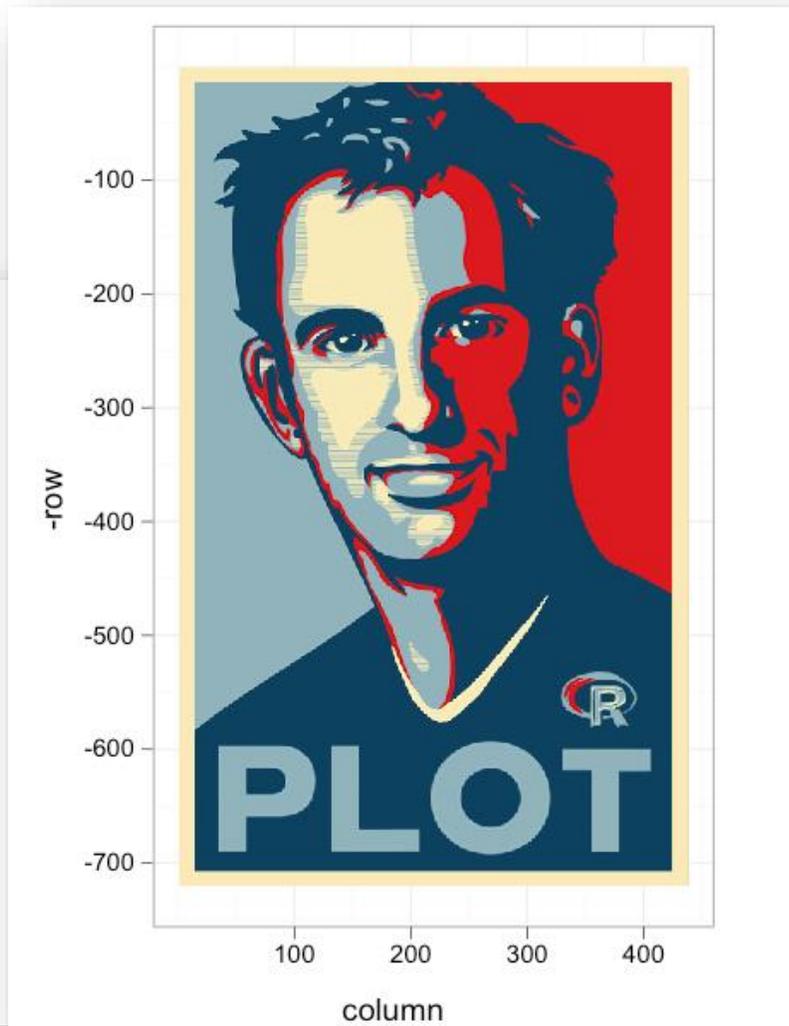
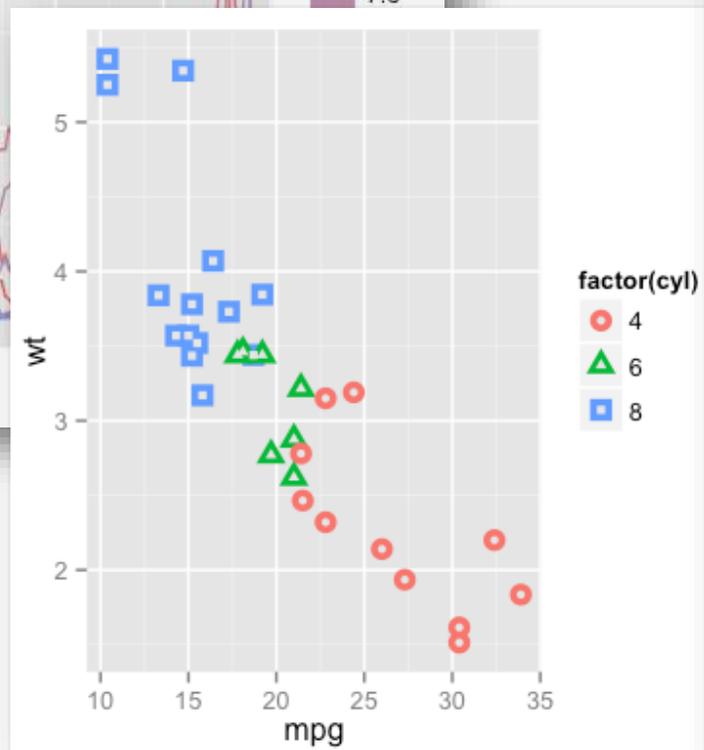
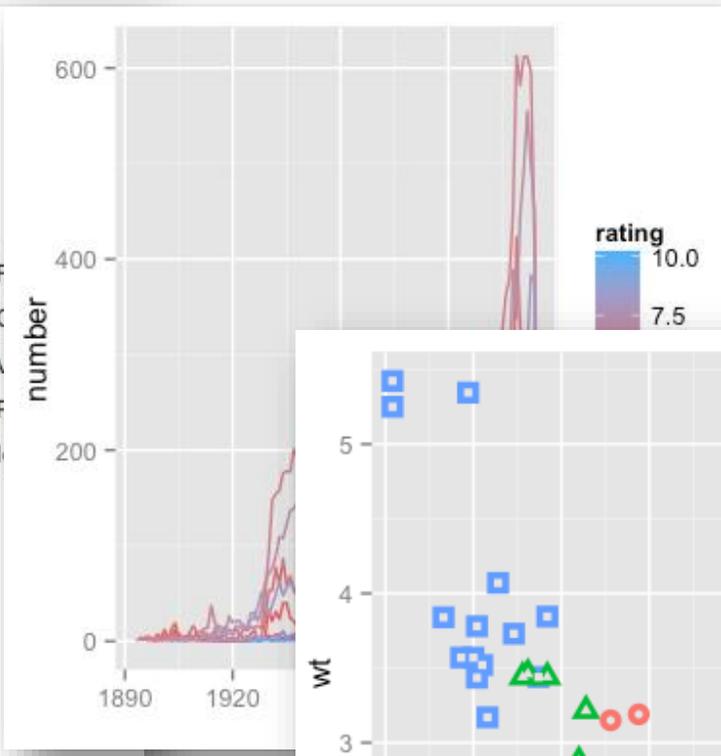
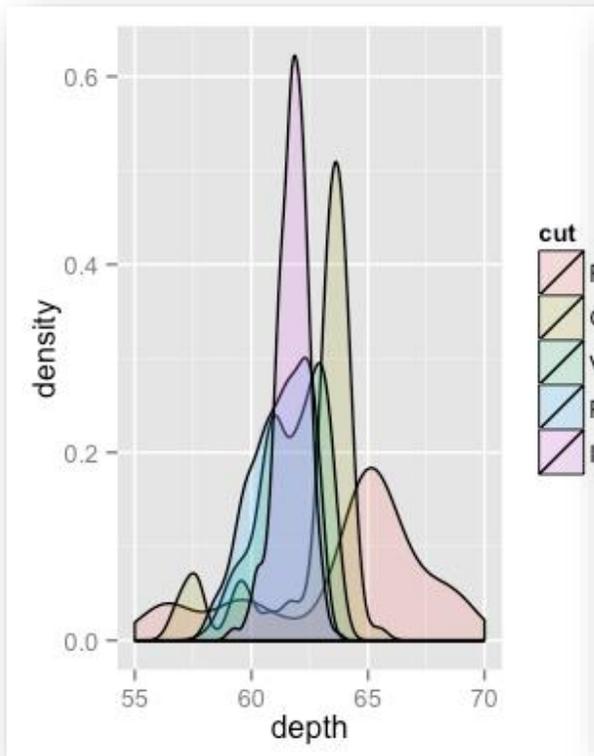


现实中的你：



实际项目中**80%**的时间
用于清洗数据!!!

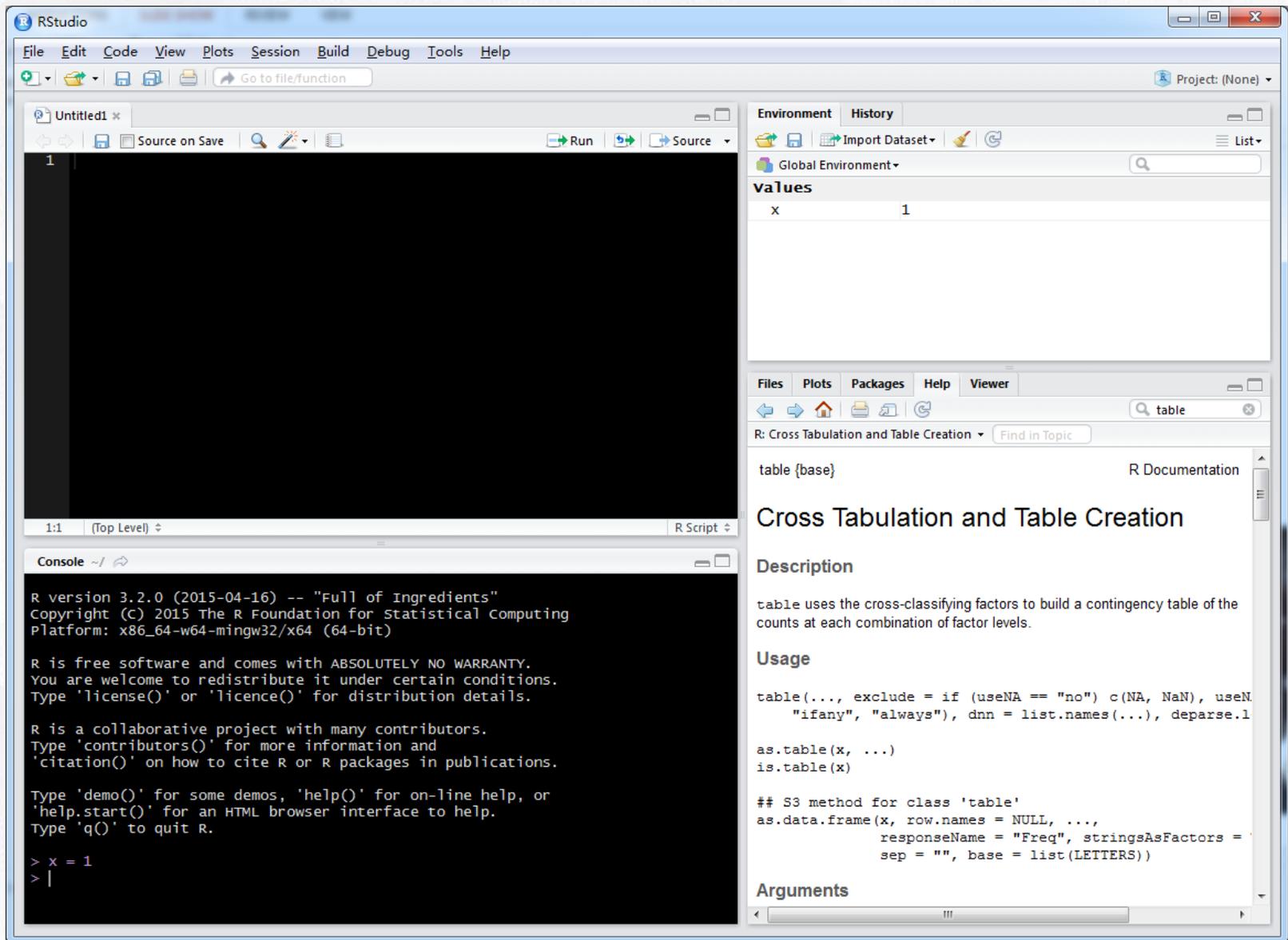
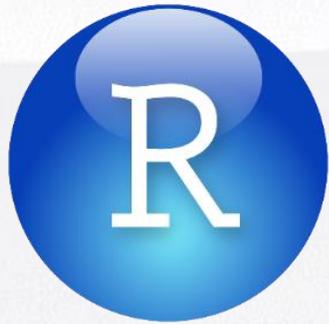
R能干啥：描述分析可视化



02 如何学会R语言?

下面是本次直播最重要的**3**件事情!

Step 1: 选好编辑器 (IDE)

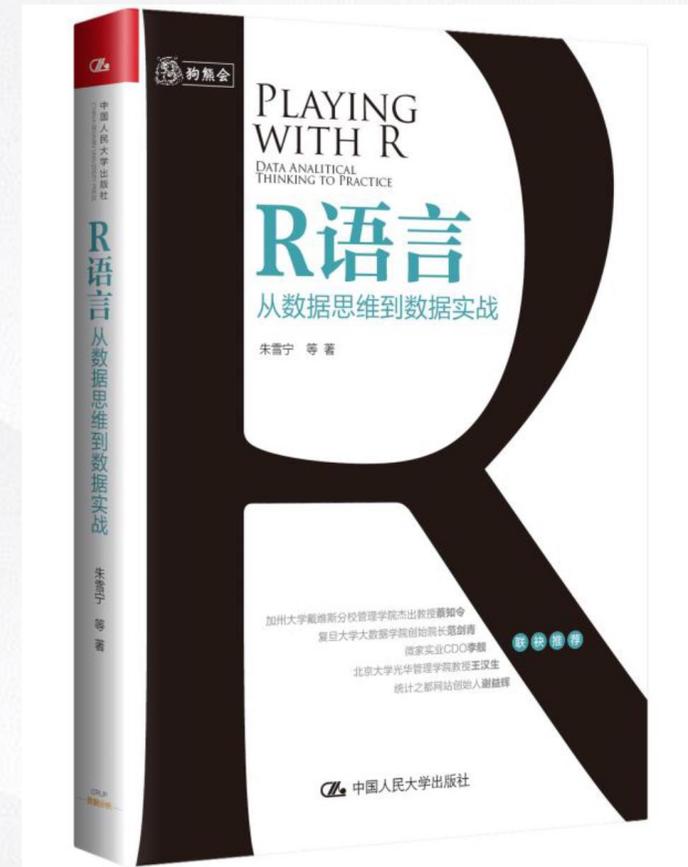


Step 2: 代码规范、撰写文档

- **简洁、清晰、高效**
- **代码分块，逐行注释!**
 - ✓ 增加功能性、可读性
 - ✓ 对于函数：写好输入输出、函数功能
 - ✓ 对于文件：写好文件用途等
- 函数、变量命名：有一定的可读性
- 不要太拥挤!
 - ✓ 不要吝惜空格：逗号前空格，等号前后空格

Step 3: 勤动手、会求助

- 狗熊会：
 - R语言：从数据思维到数据实战
 - R语千寻 专栏（在公众号输入R返回历史文章）
- 勤练习：
 - www.kaggle.com
- 会求助
 - www.google.com
 - StackOverFlow
 - Github
 -



练习一下写代码吧!

能动手，别BB!

求助开源社区

- StackOverFlow
- Github
- 统计之都

友情提示：

求助之前：

Do Your Homework!!!

03 如何使用我们的书？

定位

- 特色:

- 代码详尽
- 数据完善
- 语言轻松

- 适用于:

- 初学者: 没有数据、无从下手
- 业界: 数据分析师
- 教师: 可选作讲课素材

这里要特别提醒的是: 这些函数看似简单, 但在实操过程中非常重要。当对不同数据源做各种合并、匹配、排序、删除等操作时, 时刻要有一种“看到新数据先 `summary()`, 看看是不是符合自己的预期, 有没有新异常”的习惯, 也就是要有“时刻把握你的数据状态、进度”的意识, 否则, 待到后面报错时再重新回来查看就费时费力, 万一本来匹配有误最后程序还没报错, 后果会很严重。所以我们要记得不断查看处理的每一步到底对数据做了什么。

下面是 `str()` 和 `summary()` 函数的演示结果。可以清晰看到 `movie` 数据集中哪些变量是数值型 `num`, 哪些变量是因子型 `factor`, 以及它们各自的取值范围。

```
str(movie)
## 'data.frame': 19 obs. of 11 variables:
## $ name : chr "叶问3" "美人鱼" "女汉子真爱公式" "西游记之孙悟空三打白骨精" ...
## $ boxoffice : num 77060 338583 6184 119957 111694 ...
## $ doubanscore: num 6.4 6.9 4.5 5.7 4 7.7 6.5 6.4 5 5.6 ...
## $ type : chr "动作" "喜剧" "喜剧" "喜剧" ...
## $ duration : int 105 93 93 120 112 95 131 108 95 102 ...
## $ showtime : chr "2016/3/4" "2016/2/8" "2016/3/18" "2016/2/8" ...
## $ director : chr "叶伟信" "周星驰" "郭大雷" "郑保瑞" ...
## $ star1 : chr "甄子丹" "邓超" "赵丽颖" "郭富城" ...
## $ index1 : int 11385 41310 181979 12227 16731 178 13499 14759 13251 6911 ...
## $ star2 : chr "张晋" "林允" "张翰" "巩俐" ...
## $ index2 : int 4105 9292 44277 8546 30277 1540 77260 755 9549 5614 ...
```

目录

目 录

CONTENTS

第 1 章 初识 R 语言	1
1.1 初识 R 语言	1
1.2 安装 R 语言	10
1.3 获取 R 帮助文档	24
第 2 章 R 语言数据操作	39
2.1 R 中的数据类型	39
2.2 数据读入	90
第 3 章 R 语言与统计分析	109
3.1 描述分析及可视化	109
3.2 统计检验	162
3.3 回归分析	171
3.4 代码规范与文档撰写	207
第 4 章 R 语言与非结构化数据分析	223
4.1 文本分析	223
4.2 图像分析	238
第 5 章 R 语言与机器学习	265
5.1 机器学习概述	265

2 R 语言：从数据思维到数据实战

5.2 数据预处理	273
5.3 模型训练与调参	283
5.4 模型训练与集成	289
第 6 章 R 语言爬虫初介	298
6.1 HTML 基础与 R 语言解析	298
6.2 XML 与 XPath 表达式以及 R 爬虫应用	305
6.3 HTTP 协议	311
6.4 AJAX 与网页动态加载	319
6.5 正则表达式与字符串处理函数	324
6.6 R 语言爬虫实战	331

课时安排

章节	内容	上课周 (每周3课时)
初识R语言	<ol style="list-style-type: none">1. R的背景、功能2. 安装R语言, 编辑器介绍3. 获取帮助	1
R的数据操作	<ol style="list-style-type: none">1. 数据类型2. 数据读入	2—4
R语言与统计分析	<ol style="list-style-type: none">1. 描述分析2. 可视化3. 统计检验4. 回归分析	5—8 (小组汇报)

课时安排

章节	内容	上课周（每周3课时）
R语言与非结构化数据	<ol style="list-style-type: none">1. 文本数据分析2. 图像数据分析	9—10（小组汇报）
R语言与机器学习	<ol style="list-style-type: none">1. 机器学习概述2. 数据预处理3. 模型训练与调参4. 集成学习	11—13（小组汇报）
R语言爬虫	<ol style="list-style-type: none">1. 爬虫基础（HTML,XML,Xpath等介绍）2. 正则表达式3. 爬虫实战	14—15（可按程度布置自学内容）
课程汇报	用R语言自爬数据、自命题，进行数据分析、建模及模型测试与解读	16



谢谢!
