R 入门指引 (The learning path on R)

https://github.com/jun3970/R-Intro

胡弘宇

12 June 2020

> R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.
>
> — The Comprehensive R Archive Network - R FAQ - What is R?

前言

又临毕业季，导师（暨南大学经济学院谢子雄教授）提议，让我为师门即将入学的新生撰写一份 R 语言入门指引。[1] 得谢老师信任，诚惶诚恐，我尽力为之。需要强调的是，本文基于个人学习经验写就，少不了有一些弊病。读者当审慎参考文内的粗浅看法与不成熟建议。

R 是一门开源的专业统计计算语言，脱胎于贝尔实验室的 *S* 语言。[2] 依托于其灵活强大的统计计算与数据绘图功能，加之程式语言较易理解（核心开发团队成员大多

[1] There are two major types of programming languages: low-level languages and high-level languages. Low-level languages are referred to as 'low' because they are very close to how different hardware elements of a computer actually communicate with each other. A high-level language is a programming language that uses English and mathematical symbols, like +, -, % and many others, in its instructions. When using the term 'programming languages,' most people are actually referring to high-level languages, like as R. https://study.com/academy/lesson/machine-code-and-high-level-languages-using-interpreters-and-compilers.html

[2] R 的语法來自 Scheme。Scheme 是一种函数式编程语言，遵循极简主义哲学，以一个小型语言核心作为标准，加上各种强力语言工具（语法糖）来扩展语言本身。著名计算机科学入门教材《计算机程序的构造和解释》（SICP）利用 Scheme 來解释程序设计。Wikipedia:Scheme

图 1: R logo

数为统计学家），拓展包（package）丰富，R 成为当前学术界（尤其是统计、生物、经济等学科）最青睐的量化工具之一。[3] 2009 年 1 月 6 日被纽约时报科技版专文报导 Data Analysts Captivated by R's Power. 我们可以在 CRAN 免费获取 R 的安装文件与官方文档。[4] RStudio 公司为我们打造了一款非常棒的开源 R 语言集成开发环境（IDE）— RStudio。[5]

好的教材可以使读者更快地入门某领域，降低读者的学习成本。网络上关于 R 的学习资料琳琅满目，新手难免陷入选择的困惑。本文以推荐 R 语言优秀教材为主体内容，附带的说明性文字和补充材料读者可根据自身的需求选择性撷取。书籍的推荐顺序，在学习理解难度上，大体上保持了由易到难的顺序。具有一定 R 语言基础的读者，可以直接点击文内链接到资料相应页面做有针对性的学习。本文所推荐书籍的标准是，

[3]我到底该学 R 还是学 Python? 首先看导师给的项目（要求）和个人规划；其次二者并不冲突，对于非计算机专业的同学来讲，无论熟练掌握哪一门都是极好的。具体来说，R 的战场只有两个：统计计算和数据可视化。而 Python 的应用领域有：web 开发、服务器运维、自动化测试、人工智能（机器学习）、网络爬虫。可以看出 Python 并不是为了某一类业务岗而专门设计的程序语言（“胶水语言”），由于其本身并不专精，故大家多是将其作为次语言使用（其在国内声名这么火爆，和培训机构的狂轰乱炸，割大学生韭菜有很大关系）。

[4]CRAN 全称 Comprehensive R Archive Network. 是 R 官方组织维护的 R 资料讯息聚集地。网页样式虽停留在上一个世纪，但该有的干货一点儿也不少！为提高数据的网络传输速度，我们选择一个距离我们较近的镜像。大陆地区我推荐清华或北外的镜像。此外，浏览 R 的官方主页 https://www.r-project.org/ 我们可以实时了解 R 的最新动态，如所有被 CRAN 收录的包的清单：Table of available packages, sorted by date of publication.

[5]RStudio::conf 是北美地区的数据科学盛会，会议报告人大体为热门 R 包开发者，数据科学从业者，以及具有使用 R 丰富经验的技术爱好者。会议材料与现场视频（2017 年至今）被收录在 RStudio 官网下的 Webinars & Viedos 栏目（访问需要外网环境，国内用户可到该网站的项目仓库（repository）rstudio-conf 获取会议报告幻灯片）。统计之都有一篇 RStudio 公司创始人 J.J. Allaire 接受记者采访的翻译稿。

1. 作者本身为 R 核心开发团队成员、统计学家，又或者 R 社区的领军人物，有代表性的 R 包作品，即书籍内容的质量有所保证；

2. 优先推荐开源书籍（获取方便），书内示例代码和数据可以在 GitHub 找到。[6]

3. 书籍之间在内容上具有一定的互补性，或者说各具代表性。这是因为有层次、互补性地学习可以使我们将 R 掌握地更加牢固与精深。

若读者手头没有本地的数据项目需要处理，则我建议读者在学习过程中~~有选择地~~**敲键盘复盘书中例子**（尤其是理解起来吃力的部分）。若读者在学习过程中能做到时常尝试换一种写法，实现相同的程序目的，那么读者很快就能入门 R。因为 R 语言不难于理解，单难于理解其语法逻辑（向量化）后的持续性使用，典型情境就是学了就忘（手头有感兴趣的数据项目，还有人 push 你思考，则此问题迎刃而解）。此外，微小的写法差异累积起来就是自己的 R 代码风格。

入门量化分析（数据科学）

A. The R for data science: Import, Tidy, Transform, Visualize, and Model Data by Garrett Grolemund, Hadley Wickham - O'Reilly Media

本书不适合具有其它统计语言编程经验（SAS, Stata 等），想直接上手 R 做统计计算的应用统计者。本书适合两类读者：

1. 对数据的认知与实践大体为做描述性统计，当下想入门 R，在论文中做一些稍微复杂点的量化分析工作。典型如社会科学专业的本科生和低年级研究生。[7]

2. 了解 R 基本语句的大致用法，对 R 中当前热火朝天的 tidyverse 感兴趣的数据科学爱好者。

[6]Github 是世界上最大的代码存放网站和开源社区，绝大多数的 R 包源码都托管于 Github, 当 *CRAN* 上某包未及时更新（稳定版），我们可通过 devtools 包中的 `install_github()` 函数直接从 Github 仓库安装最新开发版（evelopment version）。

[7]与此观点不同，本文所推荐的第二本书 *The Art of R Programming: A Tour of Statistical Software Design* 的作者 Norman Matloff 写有一篇反对没有计算机基础的新人以 tidyverse 入门 R 语言的长文 Tidyverse Skeptic: An alternate view of the Tidyverse "dialect" of the R language, and its promotion by RStudio. 根据个人经验，若读者的自学基本是单兵作战，只能借助教材和网路信息进行自主学习，本人还是建议以 tidyverse 入门 R 语言，原因为 tidyverse 内容虽然多，但理解和执行（调用）难度并不大。

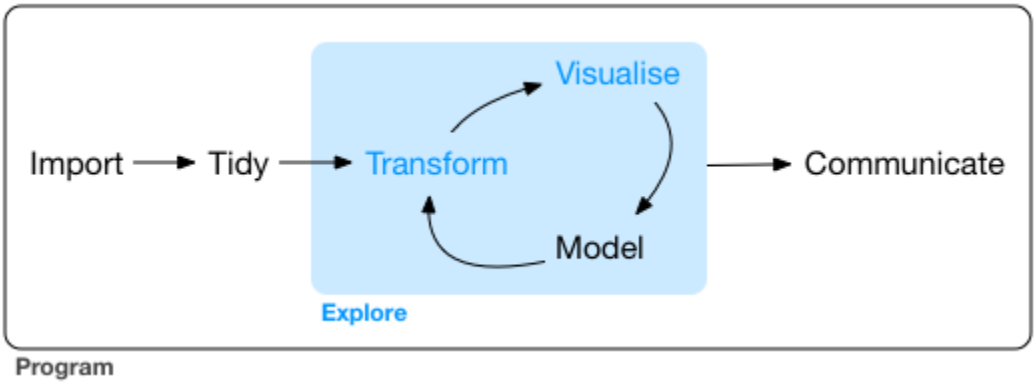

图 2: 数据分析工作流程

> The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.
>
> — Hadley Wickham [8]

以本书入门 R 语言的优势

1. 本书叙述逻辑框架是数据分析的工作流程。阅读本书，能使读者对寻常数据项目的内容与流程有一个系统性的认知。

2. 书中对 tidyverse （极乐净土）做了系统且全面的介绍。tidyverse 系列包是一套以数据项目为内容对象的工具包。其一定程度上重新包装了 R 基础语句（尤其是管道符的加入），使得未有编程经验的新手能跳过细碎的语法细节，进入以函数为核心的数据操纵环节。

3. tidyverse 用户众多，容易寻求帮助。编写 R 代码过程中遇到具体问题，可以到 stack overflow 向全世界用户乃至开发者本人寻求帮助。

4. 以 tidyverse 中某个包为依赖环境的拓展包不在少数，部分亦十分优秀（如 dtplyr, dbplyr），学海无涯…… [9]

[8]此处必须介绍一下大佬 Hadley Wickham。Hadley 供职于开源软件公司 RStudio，在 2019 年其凭借自身在数据科学领域的卓出贡献获得统计学 “诺贝尔” 奖—考普斯总统奖 (COPSS Presidents' Award)。获奖词为：Wickham was awarded the international COPSS Presidents' Award in 2019 for “influential work in statistical computing, visualisation, graphics, and data analysis” including “making

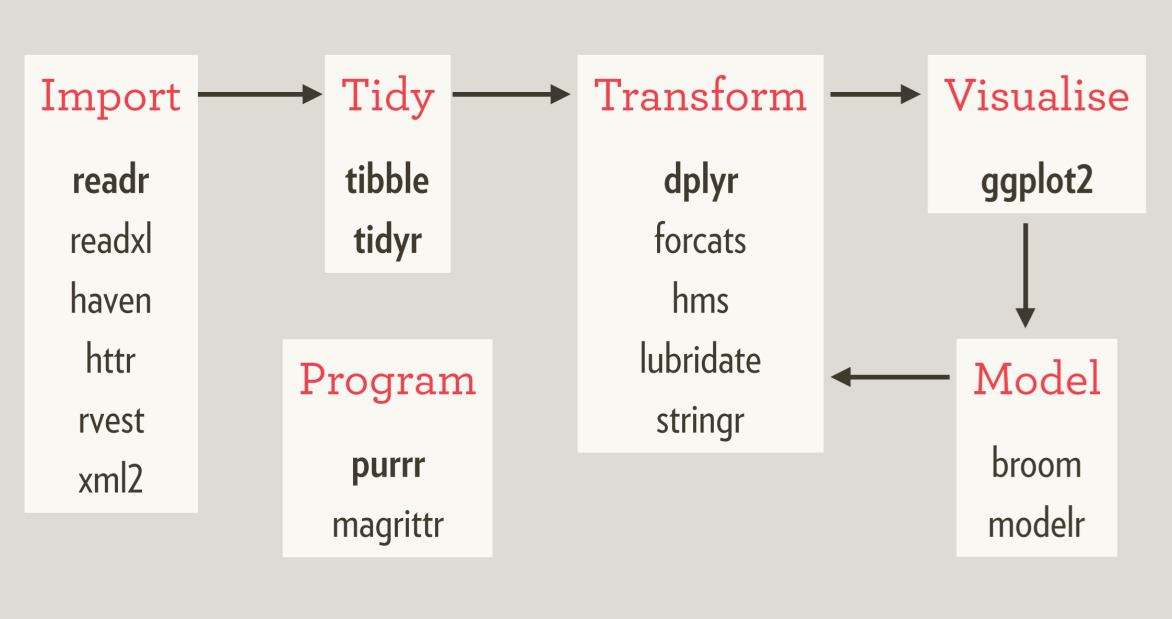

图 3: tidyverse 工作流

本书不友善之处

1. R 语言本身语法内容较少，且在书中分布零散（例如对数据类型说明被排在第二十章）。这将致使读者难以对 *R* 语言有具体入微的理解与掌握。因此当我们开始处理本地个人项目时，比较容易陷入茫然无措的境地。

2. 大概率成为“净土宗”的狂热信徒，表现为密集（强迫症）使用管道符。[10]

3. tidyverse“黑魔法”较多，写时优雅，但排除 bug 费时费力。此外，近两年 tidyverse 版本更新较快（以 Hadley 为首的开发团队庞大）。粗听起来这似乎是优点，但当你追逐潮流地更新包版本，重新跑之前写好的脚本，遇到奇怪的 bug 时……那滋味，很酸爽！

statistical thinking and computing accessible to a large audience”.

[9]dbplyr 可以实现与 dplyr 语句结果等价的 SQL 数据库操作，利用 dtplyr 可以实现间接调用 data.table 包提高数据处理速度。

[10]在由管道符构建的当前局部环境下，我们以变量的形式索引数据（dplyr verbs use tidy evaluation；在全局 base-R 环境下，若按变量名称索引数据，我们必须使用该变量的名称字符）。这种设计固然在一定程度上为我们带来了书写代码的便捷，但与全局环境下 base-R 语法的不一致，极可能会使新人混淆 R 中索引数据的底层逻辑。尤其是在管道符环境下调用本地编写的函数，稍有不慎，全局变量参数被按照字符进行传参，bug 便如影随形……

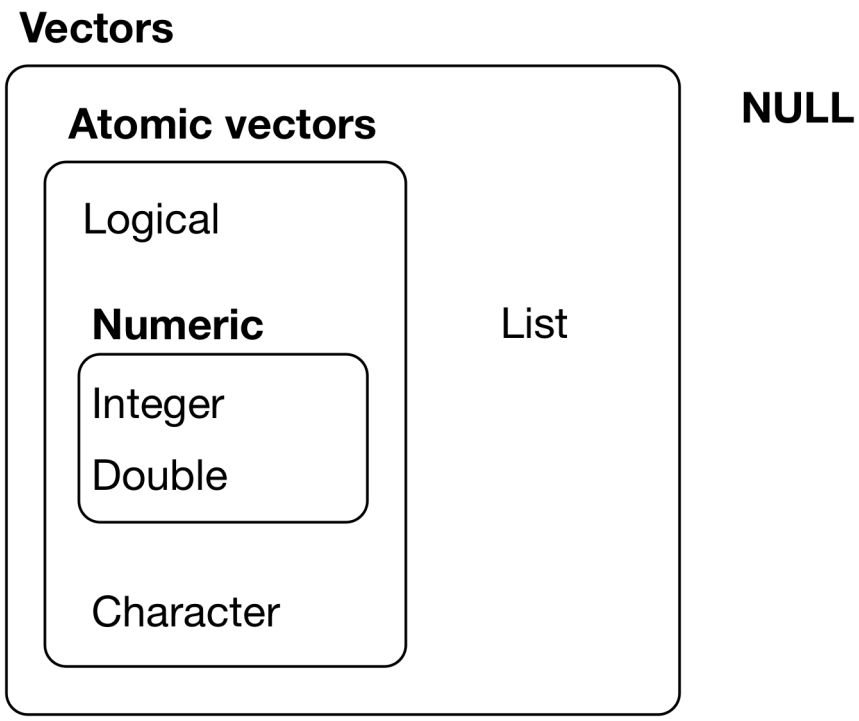

图 4: R 数据结构

B. *The Art of R Programming: A Tour of Statistical Software Design* by Norman Matloff - No Starch Press

本书面向的是对 R 有了一定的了解，想深入理解 R 的语法特性与数据对象，以及提高 R 编程水平的入门读者。尤其是适合拥有其它统计编程语言使用经验的学习者。

面对五花八门的数值，选择合适的数据类型（type）以及存储形态，能极大地简化后续统计计算环节的语句。作者在书中花费五章（共 110 页）详细叙述了 R 的数据结构（`vector`, `array`, `list`, `data.frame`, `factor`），这能为读者打下坚实的数据对象基础（尤其要牢牢掌握 vector, list, data.frame）。建议认真阅读本书**第二至第七章**内容。

本书的令一大优点是同一个例子，作者会以渐进优化的逻辑实现两到三次（语句更简洁高效与通用）。部分章节还附有一个综合性的例子，仔细琢磨这些例子能较好地提高我们应用 R 语言的水平，写出高质量的 R 代码。

通过阅读本书你将学到：

1. R 语言的底层逻辑，常用数据类型以及存储结构。
2. 习惯于编写自己的函数，开启函数式编程之旅。[11]
3. 程序循环控制流与 R 语句优化逻辑与思维。

C. R Programming for Data Science by Roger D. Peng - Leanpub

学完上面两本稍偏技术层面的书，读者应该就能处理百分之九十以上的数据项目了。接下来，学习 Peng 的这本 *R Programming for Data Science* 一定程度上能为读者揭开 R 语言的面纱（以一种大巧若拙的高度应用 R）。

该书特点是，短则两三行，长则五六行，Peng 即能叙述清一个重要函数或某类操作的概念及用法。且语言极其简洁精炼，每节内容很少超过一页纸。Peng 为我们高屋建瓴地概述了 R 中常用数据操作的逻辑，并辅以简单的示例（喜欢通过观看视频学习的同学（有外网环境），可直接经由每章开头的视频链接（Youtube）观看教学视频）。阅读本书，能帮助我们从 *The Art of R Programming* 的程序语言框架中跳出来，复以处理数据的视角审视 R。

通过学习本书你将掌握：

1. R 中基础函数（base-R）的灵活使用。
2. 码其然，知其所以然。如 R 为什么会有内存瓶颈（Scoping Rules of R）?
3. 了解如何有效地优化某段 R 代码（从其实际占用 CPU 时间出发）以及 Debug 的思路。
4. 代码也要排版（8 个 space 为一个 Tab，每行内容不超过 80 个字符）!

D. Advanced R by Hadley Wickham - Chapman & Hall's R Series

本书是第一本书 *The R for data science* 的进阶版，更多地站在开发者的角度讲解 R 的语法设计。书中四分之三的内容由 programming 组成, 甚至包含 R & C++

[11] 除了 if / for 等基本程序循环控制流，`base-R` 环境中的 `apply` 族函数（我自己最常用的是 `lapply` 与 `sapply`），`purrr` 包中的 `map` 族函数（我自己最常用的是 `map` 与 `map2`，`map_dbl` 与 `map_dfr`）亦十分有必要掌握。值得强调的是，`purrr` 中的 `map` 族函数的威力在于与 list-colum 式数据储存结构的结合使用。此外，与 base-R 的 `apply` 族函数相比，`map` 族函数语法更加协调统一，易于记忆。若实在记不清，可在 `help()` 之外，配合 `args()` 直接查看函数参数设定情况（形参与实参，参数位置顺序等）。

内容。对于只是拿 R 做统计计算的应用者而言，学习收益与时间成本不太划算。更推荐熟练掌握 RStudio 公司制作的 cheat sheet. [12]

本书是本文所荐的四本技术书中我唯一没有读过的一本。依然将其摆在这里，是我考虑到可能会有学弟妹已有一定的计算机科学基础，想无缝衔接到 R。如是，则可直接上手 Hadley 大佬的这本 *Advanced R*.

E. 干货视频（Youtube）

R Programming Tutorial - Learn the Basics of Statistical Computing
Hadley Wickham: Managing many models with R
Intro to Data Visualization with R & ggplot2
RStudio Essentials programming 2: Debugging Code
RStudio Tips and Tricks
Code smells and feels

计量经济 & R

> I've written thousands of lines of R, but never actually used it for statistical analysis as opposed to using it as a functional programming language.
> — Tom Lawton

本人计量经济的水平十分有限，等同于大一新生（完全没学），因此本节内容粗浅异常，读者可自寻通过别的渠道规划学习方案。

A. Using R for Introductory Econometrics by Florian Heiss - Independently published

我真正阅读过的应用 R 分析经济问题的只有这本 *Using R for Introductory Econometrics*. 在本书中，Florian Heiss 用 R 重现了 *Introductory Econometrics: A Modern Approach* (by Jeffrey M. Wooldridge) 书中所有例子的计量结果。通过阅读本书，我们可以很快地明白如何通过调包调函数得到我们想要的计量结果。本书缺点亦

[12] cheat sheet 是以包为单位总结了（辅以可视化）包内部主体函数简单用法的一页 PDF。一份好的 cheat sheet 能极大地帮助我们理解函数的用法（操作与目的），解放我们的记忆力。我们也可根据自己的需求到该项目 GitHub 仓库 下载更多网友制作的 cheat sheet。一般来说，带有 cheat sheet 的包不会是差包。

位于此：规范的示例数据我们可以从 `woodlridge` 包 (by Justin M. Shea and Kennth H. Brown) 中直接获取，跑回归则直接一个函数命令结束。当我们处理自己的数据项目时，哪有这么爽歪歪的 “老板” 的待遇！因此本书搭配 *Introductory Econometrics* 可供高年级本科生学习，不太适合研究生。

值得强调的是，本书内容无缝衔接 Wooldridge 的计量教材 *Introductory Econometrics*, 能帮助我们更深入地理解 *Introductory Econometrics* 书中的经典例子，而不是跑旁的乱七八糟的回归。

B. Applied Econometrics with R by Christian Kleiber, Achim Zeileis - Springer

触之所及，另一本 *Applied Econometrics with R* 十分值得推荐。此书作者不仅花费大段文字提点我们在应用某包函数跑某类计量模型时的逻辑与注意事项（难得的经验），还给出了核心统计量的数学表达式（矩阵形式）。此书还有一个高大上的简称 *AER*，本书的相应包 `AER` 亦被 CRAN 收录。[13]

C. Econometrics in R by Grant V. Farnsworth

此处我还想题一题另一份只有五十页的 *Econometrics in R* . 该文档中 Cross Sectional Regression, Special Regression Time Series Regression 三节内容只占十页，半天即可读完。读者可凭此快速了解跑某类计量经济模型有哪些包可供选择（缺点是本文写于 2008 年年底，比较旧了）。

D. CRAN - Task Views

> CRAN task views aim to provide some guidance which packages on CRAN are relevant for tasks related to a certain topic.

我们可到 CRAN 的 Tasks Views 页面扫览有哪些代表性的包可被应用于本学科统计计算任务。CRAN 上的 tasks 大致是基于统计学科的子领域和学科应用范围划分的（相当于可供学院开设一至两个学期的专业选修课）。与经济相关的有 Econometrics, Finance, TimeSeries, Robust. 各 task 下的内容按模型方法划分，对我们会遇到的统计计算任务，两三句话告知我们有哪些相关的包可供调用。

[13] 写到这里，很后悔为什么我没有在接触到这本书的当时就啃它（文档下载时间显示我是 2018 年 12 月 12 日知晓本书存在的），希望以学术为志的学弟妹引以为戒，不要花太多时间在学习数据处理技术上，核心的应该是掌握回归模型的计算推导以及模型背后的经济学阐释。

附录

遇到问题如何解决

1. 通过 *R* 内置的帮助系统（`?`, `help()`）以及参数查询函数 `args()` 深入了解某一函数用法；

2. 善用 搜索引擎 与程序语言问答网站 stack overflow；

我们遇到的大多数问题网友都曾遇到过，因此网络上基本都有答案，顶多需要我们对多个解决方案进行比较与重组。此外，阅读网友的解决方案或示例代码，不光能解决我们遇到的某类问题，也能加深我们对语法的理解；

3. 查阅 R 官方给出的常见 FAQs 与 The R Manuals

需要注意的是，R Manuals 更多面向的是开发者与服务器平台用户（在教读者如何安装以及管理 R 的子文档中，Installing R under Unix-alikes 排在了 Installing R under Windows 与 Installing R under macOS 前面）。站在应用的层面，只建议有针对性地查阅 An Introduction to R, R Data Import/Export 两份主文档（利用浏览器内置的文本搜索功能，在 HTML 页面检索答案）。学习到这里，你大概已经是两个 σ 的 R 中毒者了！

包的食用方法

R 中的包由三部分内容组成：函数（包的主体内容），帮助文档（用法说明），以及示例数据。[14] 如何系统地了解包内附有哪些函数，以及这些函数的具体用法？

最直接高效的方式就是到该包主页看有无语法说明，被 CRAN 收录的包的网络地址格式为 https://cran.r-project.org/package=dplyr。其中 `dplyr` 是我输入的示例包名，当我们查询欲浏览某包主页，我们只需将其替换为我们相应包名。如我们现在想看 `ggplot2` 主页，我们在浏览器的网址栏输入 https://cran.r-project.org/package=ggplot2, 回车即可开始冲浪。我们主要看页面中哪些内容呢？

[14]这是一份网友站在统计学者和数据科学的角度给的一份 R 包清单: Essential list of useful R packages for data scientists. Many useful functions are available in many different R packages, many of the same functionalities also in different packages, so it all boils down to user preferences and work, that one decides to use particular package.

图 5: 布满 R 核心开发团队成员签名的 R 1.0.0 版 CD

- In views

  通过页面中的 In views 条目，我们可以知晓该包被哪些 tasks 索引（其定位）。

- Materials - README

  README 一般都包含有 Overview, Installation, Usage 三部分内容。友好的开发者会在文内对包的核心函数做简略介绍，对大多数用户而言，熟练掌握这些就够了。

- Vignettes

  包开发者对包内主要函数及语法更详细的说明，经常附带些“黑魔法”。建议耐心阅读。

- Reference manual

  Reference manual 是作者遵照 R 社区包开发标准，对此包内部所有函数的完整介绍（包括描述介绍、函数参数、示例代码等）。文档常长达百页（为兼顾阅读舒适性，排版较为稀疏）。[15]

[15]此处为 pdf 文档，pdf 格式的帮助手册的优势是可集中学习，不用在 HTML 页面来回跳转。其实，该份 manual 手册已被 R 内置了超链接索引。还是以包 `dplyr` 为例，我们只要在 R 终端（console）输入 `help(package = "dplyr")` 并回车，就能在默认浏览器打开该包的 Reference manual 手册（HTML）。

以 R 包为卖点的两本期刊

1. The R Journal

The R Journal 是 R 官方编辑的期刊，拥有 ISSN 编号，一年两期。每篇论文以一个新包为叙述主体（作者为实现某类统计计算方法所开发），论文可供随意下载，并附有示例 R 代码。文内的包大多遵循 Creative Commons Attribution 4.0 International license, 即只要非商业用途，署上作者名便可随意使用。

2. Journal of Statistical Software

JSS 期刊是我在搜索引擎检索某个 R 语言问题时发现的，同 The R Journal 一样，志在曝光新包，论文与示例代码可通过杂志官网超链接直接获取。并附有包的源码（避免未被 CRAN 收录的可能）。The R Journal 与 JSS 双管齐下，保准你站在 R 浪头之巅。

对 tidyverse 系列包的补充说明

- readr & tidyr & tibble

  readr 负责定制化读取数据，tidyr 负责将我们的数据整理成列（column）为变量（特征），行（row）为观测对象（单个样本）的 tabular data，tibble 增强了 R 中 data.frame 数据在终端中的阅读性，统一了对 data.frame 中数据做索引时返回值的数据属性。[16]

- dplyr

  dplyr 是 tidyverse 的核心，主要有三个方面的应用，[17]

若是在集成编辑器 RStudio 中敲相应命令，手册将直接显示在 Help 窗口（不能白白占用四五百 MB 的内存啊）。当然，我们也记不住那么多的函数（哪些是实参，哪些是形参）。我们更多的还是以查询工具书的方式浏览它。

[16]一定程度上能避免产生自己难以意识到的数据结构层面的 bug。举例来说，从 data.frame 中单取出一列数据，返回值依然是 data.frame 还是一维的向量，视索引函数而定，而不依赖于数据的维度。

[17]dplyr vs. data.table. data.table 是 R 中另一个十分优秀的数据操作包。data.table 会隐式地根据目标任务展开并行计算，优化内存占用量，计算速度表现十分出众。data.table 的逻辑与 tidyverse 的逻辑可以说是完全相反。tidyverse 强调语句的直观性，辅以大量各具特色的函数，企图让数据操作语句成为一套优雅的拳法；而 data.ble 语法就像是陈咬金的三板斧，虽然就一招三参量，但依凭复杂的 R 基础语句（绝对的抽象），依然可实现复杂的数据操作。

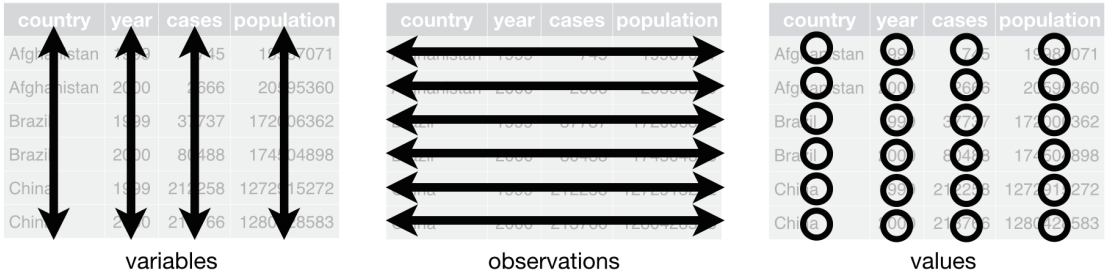

图 6: tabular data

1. 简洁且自动化程度高（如 select_* 族函数）地对行列做筛选；添加列变量，包括对 data.fram 中已有变量做数据变化（取对数等），对多列变量做线性组合等。
2. 多对个 data.frame 做各式各样的合并操作（*_join 族函数，交、并、补）。
3. `group_by()` 函数 + `summarise()` 函数 + 科学计算表达式或内置计算函数。熟练掌握此套招数，计算以及比较不同组别间的数据统计量变得非常简单，代码也十分简洁优雅！

- purrr

  purrr 构建的 list-column workflow (column-wise programming) 优雅非常！我们借助 `map` 族函数在行（row）之间实现隐式循环操作，在列（column）之间存储阶段性编程结果。分散的数据结果（以 list 为基本数据单位，包括 tabualr data、模型结果乃至可视化图表等等）集合到了一张表，能极大地简化中间变量和减少代码行数！

补充网站 & 论坛

- https://bookdown.org/

谢益辉谢大开发的一系列 `*down` 包为我们撰写各类文档文件提供了极大的助益，本站是 bookdown 的官网，一系列基于 bookdown 攥写的书籍开源于此。上文我推荐的书籍 *R for Data Science* 和 *R Programming for Data Science* 皆在列表之中。大家可根据自己的数据项目需求，选择某一主题的书进行强化阅读。[18]

[18]此处提一嘴由谢大个人单枪匹马开发的 `R markdown`。`R markdown` 在 R 的基础之上，利用 knitr

- 统计之都

由谢大发起并创办的统计之都是国内最专业的统计知识交流社区。主页的文章干货满满，三十几期的专访（统计学者为主）是我等菜鸡不竭的鸡汤来源。其下论坛也有 R 子板，坛友之间纯交流技术和统计学问题。无娱乐无政治无新闻，也不是乱七八糟的资源分享站，清净而能学到东西。

- The R Graph Gallery

一个展示如何利用 tidyverse 和 ggplot2 做数据可视化的网站，每张图片的下方便是相应的 R 代码。该网站除却数据图形非常齐全，更难得的是附带的绘图代码写的非常规范，且基本做到了每行命令皆有注释说明，十分适合对数据可视化有兴趣的同学浏览学习（最好是读过一两本讲述数据可视化的书籍，有一定的数据科学常识和图形品味之后）。这里推荐一本书 Fundamentals of Data Visualization,

> The book is meant as a guide to making visualizations that accurately reflect the data, tell a story, and look professional. In my experience, scientists frequently (though not always!) know how to visualize data without being grossly misleading. However, they may not have a well developed sense of visual aesthetics, and they may inadvertantly make visual choices that detract from their desired message. Designers, on the other hand, may prepare visualizations that look beautiful but play fast and loose with the data. It is my goal to provide useful information to both groups.
> — Claus O. Wilke

- https://www.r-bloggers.com/

该网站的定位是: *R news and tutorials contributed by hundreds of R bloggers.* 一天的博客量在三到十条左右。阅读这些博客可以极大拓展地我们的数据视野，避免我们自身陷入技术孤岛。

包让生成可重复性研究报告变得轻而易举，并且借助 pandoc 让我们可以随意选择文档输出格式（本文即是用 R markdown 写的）。站在对受众的影响和受众的范围来讲，R markdown 的产品杰出程度甚至要超过 tidyverse。感兴趣的请移步 How Rmarkdown changed my life. 此外，谢大的个人主页上 - 演讲和讲座 - 栏目亦是干货满满，塞满了其四处演讲的幻灯片！我研一上学期非常喜欢读谢大的博客（最早可追溯到 2005 年年初），文字纯真而富有启迪。