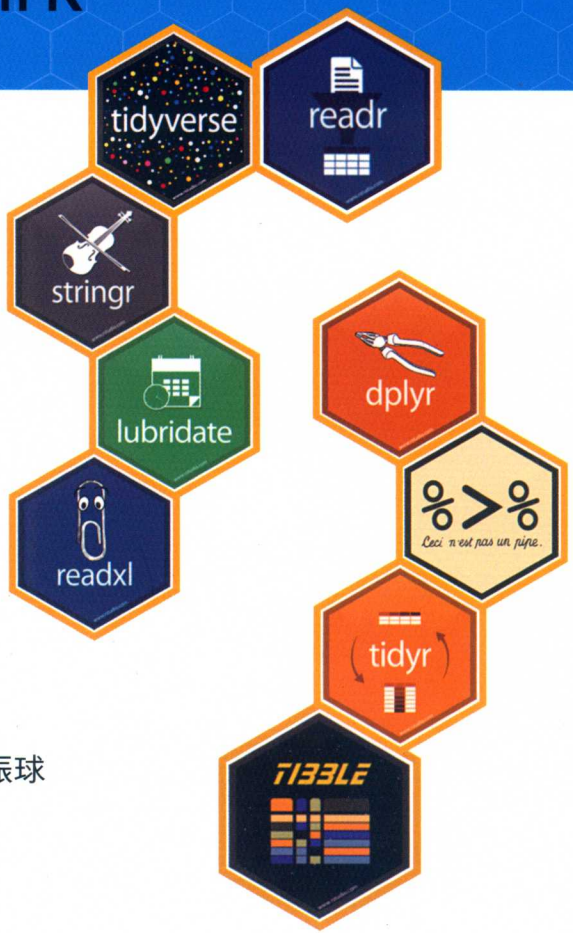
- 深入浅出，丝丝入扣，快速掌握R语言数据清洗的方法与技巧
- 玩转R语言，开心做科研

R语言与数据清洗

Data Cleaning with R



主　审　何　纳
主　编　陈兴栋　张铁军　刘振球

人民卫生出版社

语言与数据清洗

Data Cleaning with R

主　审　何　纳

主　编　陈兴栋　张铁军　刘振球

编　者（以姓氏笔画为序）

文育锋（皖南医学院）
邓文江（瑞典卡罗林斯卡医学院）
吕　明（山东大学齐鲁医院）
刘思寒（中南大学）
刘振球（复旦大学）
严　琼（复旦大学）
苏　燕（中国科学院上海营养与健康研究所）
杜　雨（北京三快在线科技有限公司）
张铁军（复旦大学）
陈　超（中南大学）
陈兴栋（复旦大学）
陈晓晨（复旦大学）
周亭攸（浙江财经大学）
袁子宇（复旦大学泰州健康科学研究院）
袁黄波（复旦大学）
索　晨（复旦大学）
徐　梦（中南大学）
徐珂琳（复旦大学）
蒋艳峰（复[illegible]）

秘　书　方绮雯（复[illegible]）

人民卫生出版社

图书在版编目(CIP)数据

R语言与数据清洗/陈兴栋，张铁军，刘振球主编
. —北京：人民卫生出版社，2019
ISBN 978-7-117-28059-4

Ⅰ. ①R… Ⅱ. ①陈… ②张… ③刘… Ⅲ. ①程序语言—应用—数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2019)第024235号

R语言与数据清洗

主　　编：陈兴栋　张铁军　刘振球
出版发行：人民卫生出版社（中继线 010-59780011）
地　　址：北京市朝阳区潘家园南里19号
邮　　编：100021
E - mail：pmph @ pmph.com
购书热线：010-59787592　010-59787584　010-65264830
印　　刷：河北新华第一印刷有限责任公司
经　　销：新华书店
开　　本：710×1000　1/16　　印张：20　　插页：1
字　　数：370千字
版　　次：2019年4月第1版　2019年4月第1版第1次印刷
标准书号：ISBN 978-7-117-28059-4
定　　价：52.00元
打击盗版举报电话：010-59787491　E-mail：WQ @ pmph.com
（凡属印装质量问题请与本社市场营销中心联系退换）

序

我们生活在一个信息爆炸的时代。正如狄更斯所言:“这是最好的时代,这是最坏的时代。”“好”在于海量的信息让我们的生活变得更加的便捷,让世界进一步变小,而“坏”在于当我们面对海量的信息时,就如同面对浩瀚的宇宙一般,显得渺小而无助。诚然,以现阶段的技术,人类还难以解析当下产生的所有数据的密码。比如,随着测序技术的快速发展,我们现在能够轻松获取某个个体的全部基因组、转录组、蛋白组、代谢组、宏基因组等一系列生物学信息。就单个个体而言,这些数据加在一起就是 GB 甚至 TB 级别的。数据的“大”,一方面为我们提供了更广阔的探索空间,另一方面,也为数据分析和挖掘带来了极大的挑战。如何从庞杂的数据中,挖掘出真正有价值的信息,这是许多科研工作者一直孜孜以求的事情。

当前,以预防为主、关口前移的健康维护理念已在全球得到普及,现代生物医学模式向预防性(Preemptive)、预测性(Predictive)、个体化(Personalized)和参与性(Participatory)、精准化(Precise)的“5P 模式”发展,也对人类健康的精准维护提出了更高的要求。要真正实现预防及解密疾病,就要找到患病因素和疾病发生机制,提出更具针对性的健康维护方案。基于这一思想,生物医学产生了多个研究热点和创新方向,包括转化医学、精准医学和人类表型组学等。随之而来的是全面、系统、多维度组学和表型数据的产生。而由于缺乏统一的、完备的数据收集和管理标准,我们总是难以对各种来源的原始数据进行有效约束,以至于呈现出“千奇百怪”的数据姿态。这无疑给后续的整合、分析、理解带来了很大的挑战。科研人员在进行真正的数据分析之前,总要花费大量的时间进行原始数据清洗,才能使其满足后续分析的要求。相比于数据分析与挖掘,数据的清洗更费时间、更具挑战,也是极为重要的一个环节。幸运的是,随着计算机科学的发展,涌现出了一批简单而强大的工具,能够帮助我们完成这些繁琐的工作。R 语言就是其中的佼佼者。

来自复旦大学等多个单位的一线科研工作者,结合泰州自然人群队列建

设和研究过程中的实际经验及自身经历，系统全面地总结了如何利用 R 语言这一有力工具进行数据清洗。目前国内尚缺乏此类工具书籍，相信此书的问世会给科研一线工作者带来直接有效的帮助，助力其更好地完成科研工作。

复旦大学副校长
中国科学院院士
2019 年 1 月

前　言

随着信息技术的快速发展，医学相关数据在呈现指数级的增长，以基因数据为例，曾有统计表明，NCBI 网站上 GenBank 数据库中所包含的基因数据，每隔两个月就会翻一番，而这一速度还在继续加快。此外，随着电子化病历的普及，医院临床上一线的患者数据也在快速增加，而且随着时间的推移，这些数据在不断的堆积；流行病学研究中，假设开展一个基数为 5 万人的队列，每个对象收集大约 100 个指标的信息，那么一次随访就可产生 500 万个具体的数字。以上事例说明了当前医学研究中的数据也在不断变大。其实，相比于数据的“大”，从一线收集的数据，另一个更加明显的特点是“杂”。一线数据的杂乱是由多种原因造成的，比如收集时，指标定义不够明确；一线工作人员出现差错；数据因种种原因出现缺失等。这些杂乱的数据为我们后续的进一步分析带来了很大的困难，因此，在进行数据分析和统计建模之前，必须对数据进行必要的“清洗”。举个通俗的例子，你想要炒一盘土豆丝，不可能直接把刚从地里挖出来的土豆放进锅里炒，而是首先要将土豆洗净、去皮、切丝，等这些准备工作完成了，再来生火、加油，而这土豆洗得干不干净，切得是否匀称，都会影响出锅时的口感。数据分析其实也是这个流程，我们对原始数据进行必要的“清洗”，比如数据格式的转换、变量的新增与删除、日期数据的处理等，这些工作的目的就是为了使数据变得整齐，满足统计分析软件所需要的数据格式。有过数据清洗和分析经验的同学应该知道，在整个的数据分析流程中，数据清洗这一步最为繁琐，也是耗时最多的。

如今流行的几种统计分析软件中，SPSS 和 Stata 虽然简单，但是灵活性不够，因此难以胜任数据清洗的任务；SAS 和 R 语言足够灵活，因此一直更受青睐。但是 SAS 属于商业软件，价格昂贵，而 R 语言属于免费的开源软件，世界范围内，有许许多多的 R 语言爱好者一直在为 R 语言的快速发展贡献自己的力量，因此，近十年来，R 语言发展势头十分迅猛，已然成为数据科学领域必不可少的一门工具了。

本书将继续秉持《R 语言与医学统计图形》的思想，从实际出发，以解决实际问题为目标，力求全面地介绍如何利用 R 语言进行数据的清洗工作，目的在于让大家以最快的速度掌握 R 语言的基础，在此基础上，掌握数据清洗的方法，并习得一定的技巧，助力自己的科研工作。

从结构上来说，全书分为十六章，前六章着重介绍了 R 语言的基本概念和基本操作，帮助不熟悉 R 语言的读者快速入门。从第七章开始，正式进入数据清洗的世界。第七章介绍了各种数据的读取与导出；第八章介绍了数据框的预处理，包括子集提取，数据结构转换等；第九章介绍了 R 语言中常用的数据汇总操作函数；第十章介绍的是异常值与缺失值的处理方法；第十一章介绍了字符串的处理方法，包含正则表达式的内容；第十二章介绍了分类变量，即因子变量的处理方法；第十三章介绍了时间日期以及时间序列数据的处理方式；第十四章介绍了 DNA 等基因序列数据的处理方式；第十五章介绍了 R 语言与 MySQL，SQL Server 等主流数据库的连接方式；第十六章介绍了利用 R 语言进行网络数据抓取的方式。

同样的，本书的编写得到了国家重点研发计划、国家自然科学基金、教育部博士点基金、上海市自然科学基金的资助。我们特别邀请了我国著名遗传学家金力教授和流行病学家何纳教授对本书进行审阅，两位老师丰富的学识和严谨的科学态度使得本书增色不少。感谢复旦大学蔡宁同学、毛宪化同学在书稿校对上的帮助。同时感谢“医学方”微信公众平台的支持。我们求学于复旦、成长于复旦，衷心感谢复旦大学对于本书的大力支持；感谢每一位编委的辛劳付出；感谢 Hadley Wickham 的无私支持；感谢 R 语言道路上的先驱与前辈，是您们的智慧成就了 R 语言的今天，才让我们能够站在巨人的肩膀上继续前行。

R 语言是一门开源共享的语言，正是有了大家的共同努力与促进，R 语言才能发展的如此之快。由于笔者水平有限，书中难免有疏漏错误之处，恳请各位专家、老师、同学批评指正（邮件可发送至 zhenqiuliu@outlook.com）。

陈兴栋　张铁军　刘振球

于复旦大学生命科学学院

2019 年 1 月

目　录

第一章

整洁数据的原则

作为本书的开篇，第一章将为各位读者介绍整洁数据（tidy data）的一些基本特征和原则。整洁数据，有时也称作“干净数据”，与之对应的就是“脏数据”（messy data）。顾名思义，整洁数据就是已经清洗好的数据，可以直接拿来进行统计分析或者数据可视化。关于整洁数据的定义和原则，业内有各种不同的版本，而笔者认为最全面的莫过于 Hadley Wickham 博士的总结。因此，本章笔者特意将 Wickham 博士于 2014 年发表在 *Journal of Statistical Software* 上的一篇题为“Tidy Data”的文章进行了翻译，以飨读者。考虑到中英文语言差异，我们对部分内容进行了略微改动，力求更好地传达 Hadley 的意思。由于笔者水平有限，可能难以还原原文全部之精髓，因此推荐有能力的读者阅读原文。原文可在 https://www.jstatsoft.org/article/view/v059i10 上下载。本文的中译版本也得到了 Hadley 本人的授权。

整洁数据

Hadley Wickham

Rstudio

摘要

为了将数据清洗成可以分析的形式，研究者们已经做了大量的努力。但是，目前还很少有研究告诉大家如何以既快又高效的方式进行数据清洗。本文涉及了数据清洗过程中一个很小但是却很重要的部分：数据整理。整洁的数据容易操作、建模以及可视化，而且具备一些特有的结构：每一列代表一个变量，每一行代表一个观测，每种类型的观察单元构成一个表格。这种框架使得整理杂乱的数据集变得很容易，因为处理大量不整洁的数据集只需要一小部分工具。这种结构还使开发用于数据分析的工具变得更容易，这些工具既可以输入数据，也可以输出整洁的数据集。此外，本文通过一个简单的案例研究证实了一致的数据结构和匹配工具的优点。

1. 前言

通常认为，数据分析过程中 80% 的时间都是花在了数据的清洗和准备阶段（Dasu and Johnson 2003）。数据准备不仅仅是第一步的工作，而是在分析过程中，每当发现新问题或者收集了新数据时，都必须多次重复的一个过程。尽管这一过程很费时间，但遗憾的是目前并没有什么研究讲述如何更好地清洗数据。我想其中部分挑战就来源于数据清洗涉及的工作太广了：从异常值的检查，到日期数据解析，再到缺失值插补。为了解决这些问题，本文着重关注数据清洗过程中一个虽小但很重要的方面，即将数据集进行结构化，以更利于后续的分析。我把这一过程称为数据整理（data tidying）。

整洁数据的原则可以看作是在数据集中处理数据的一种标准方式。标准化的方式使数据清洗变得更加简单，因为你不用从零开始，也不必每一次都做重复的工作。而且这一方式已被我们用来进行最初的数据探索及分析，并且简化了数据分析工具的开发工作。麻烦的是，当前我们往往需要在不同的工具间进行不断地切换。也就是说，你需要花一些时间处理一个工具的输出（output），以便作为另外一个工具的输入（input）。如果能将一个整洁的数据集和整理数据的工具紧密结合起来，数据分析会变得更加简单。这样，你就可以只关注感兴趣的问题，而忽略那些不好玩的数据逻辑。

整洁数据的原则与关系数据库和 Codd 的关系代数密切相关（Codd 1990），不过可被搭建成统计学家所熟悉的计算机语言。计算机领域的科学家在数据清洗方面也做了很多的贡献。比如，Lakshmanan，Sadri 以及 Subramanian（1996）定义了一种 SQL 的扩展，可以借此处理脏数据；Raman 和 Hellerstein（2001）提供了一个数据清洗的框架；Kandel，Paepcke，Hellerstein 和 Heer（2011）开发了一款界面友好、可自动生成数据清洗代码的交互式工具。这些工具虽然都很有用，但它们往往以统计学家不熟悉的计算机语言来呈现，并没有给出更多关于数据集如何结构化的建议，且缺乏与数据分析工具之间的有效连接。

我日常处理现实数据的经验促使了整洁数据的开发。由于缺乏对数据结构的限制（即使有也不多），现实数据通常呈现出一些很奇怪的构造。即便是为了让这些数据以统一的方式进行构造以便后续的分析，就花费了我大量的时间，更别说让这一工作变得简单了。同样，我也力求将这些技巧都传授给我的学生，让他们可以独自处理来自真实世界的数据。在这过程中，我开发了 *reshape* 包和 *reshape2* 包（Wickham 2007）。尽管我有意识地使用这些工具并通过一些示例来讲授它们的用法，但依然缺乏一个把我的想法变得更加直观的框架。这篇文章提供了这一框架，它传达了一个全面的“数据哲学”，而这一哲学奠定了我在 *plyr* 包（Wickham 2011）和 *ggplot2* 包（Wickham 2009）

中工作的基础。

这篇论文整体结构如下：第二部分介绍了整洁数据的三个构成特点。考虑到大部分真实数据是不整洁的，第三部分介绍了将脏数据变得整洁的方法，并以大量的实例来解释这些方法。第四部分定义了整洁工具（tidy tools），这些工具可以输入和输出整洁的数据集；此外，第四部分还讨论了如何将整洁数据和整洁工具结合在一起，从而让数据分析变得更简单。第五部分提供了一个小小的案例研究，以阐明上述的所有原则。第六部分讨论了这一框架的缺陷，并提出了可以采取哪些其他行之有效的方法。

2. 整洁数据的定义

幸福的家庭大多一个样，而不幸的家庭却各有各的不幸。

——列夫•托尔斯泰

如同托尔斯泰形容的家庭，整洁的数据大多一个样，而脏数据却各有各的脏法。整洁的数据集提供了一个标准化的方式，能够将数据集的结构（其实体布局）与其语义（其实际含义）连接起来。这部分，我会给出几个标准术语来描述数据集的结构和语义，再用这些术语来定义整洁数据。

2.1 数据的结构

大部分的统计数据集都是由行和列组成的长方形表格。列通常情况下是有标签的，而行有时也会有一个行标签。表 1 展示了一个现实中常见数据格式的假想实验数据，它有 3 行 2 列，行和列都有对应的标签。

表 1 典型的示例数据集

	treatmenta	treatmentb
John Smith	-	2
Jane Doe	16	11
Mary Johnson	3	1

表 2 以不同的结构展示表 1 相同的数据

	John Smith	Jane Doe	Mary Johnson
treatmenta	-	16	3
treatmentb	2	11	1

有很多方式可以构造相同的数据，如表 2 展示的是与表 1 中相同的数据，但是行和列进行了转置，数据是相同的，不过结构不同。我们的行和列词汇表不够丰富，无法描述为什么这两个表可以展示相同的数据。除了数据的外观，我们也需要一种方式去描述表格中每一个数值的潜在语义或者含义。

2.2 数据的语义

数据集是数据的集合，而数据通常是数字（定量数据）或字符串（定性数据）。一个数据由两个维度构成，它既属于一个变量（variable），也属于一个观测（observation）。一个变量包含了衡量同一特征，比如身高、温度、持续时间等的所有观测单位的数值。一个观测包含了同一个观测单位（比如人、天或者种族）所有特征的数值。

表 3 中重新组织了表 1 的数据，使得值、变量和观测更加清楚。该数据集包含了 18 个值，代表 3 个变量和 6 个观测。

表 3　数据与表 1 相同，但是变量和观测分别以列和行展示

person	treatment	result
John Smith	a	-
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

变量如下：

1. person：包含三个可能的取值（John Smith，Mary Johnson 和 Jane Doe）。

2. treatment：包含两个可能的取值（a 和 b）。

3. result：包含五个或者六个可能的取值，这取决于你如何看待缺失值（-，16，3，2，11，1）。

实验设计告诉了我们更多观测的结构信息。在这个完全交叉设计的实验中，对每一对 person 和 treatment 的组合进行了测量。实验设计同样决定了缺失值是否应该被安全舍弃。在本实验中，缺失值表示某个观测应当被测量，但是实际上并没有，因此保留它是很重要的。而结构缺失值，即难以进行测量的数值，比如怀孕的男性的数量，就可以放心地删掉。

对于一个给定的数据集，很容易就能分辨出哪些是观测，哪些是变量，但奇怪的是，通常很难准确地定义变量和观测。比如，如果表 1 中的列是身高和体重，那么我们可以很轻松地将它们称作变量；如果列是高度和宽度，那么它就不那么清晰了，因为我们可能认为高度和宽度是维度变量的值。如果列是家庭电话和工作电话，我们可以将它们视为两个变量。但是在分辨诈骗电话时，我们可能需要两个变量：电话号码和号码类型，因为对多人使用一个电话号码可能意味着诈骗号码。一个经验法则是，描述变量间的函数关系比描

述行之间的关系更加容易，比如 z 是 x 和 y 的线性组合，密度是重量和体积的比值。不过，比较不同组别的观测（比如 a 组的平均值和 b 组的平均值），相对于比较不同组别的列更加容易。

在一个特定的分析中，很有可能存在多个维度的观测。比如说，在一项新型的过敏药物实验中，可能存在三个观测类型：每个研究对象的人口学数据（年龄、性别、种族），每个研究对象每天的医疗数据（打喷嚏的次数、眼睛是否红肿），以及每天的气象资料（温度、花粉量）。

2.3 整洁数据

整洁数据是一种将数据集的含义映射到其结构上的标准方式。一个数据集是否整洁，取决于它的行、列和表格同观测、变量以及类型的匹配程度。在整洁数据中：

1. 一个变量构成一列。
2. 一个观测构成一行。
3. 每一个类型的观测单元构成一个表格。

这是 Codd 的第三种标准形式（Codd 1990），但在统计语言上有框架的约束，而且其重点是放在单个数据集上，而非关系数据库中常见的许多连接数据集。脏数据是指任何其他组织形式的数据。

表 3 是表 1 的整洁形式，每一行代表一个观测，即每一种疗法（treatment）在每一个研究对象（person）上的疗效（result），每一列代表一个变量。

对于分析师和计算机来说，很容易从整洁数据中提取所需的变量，因为它提供了一个数据结构化的标准方式。对比表 3 和表 1，你会发现：对于表 1，你需要采用不同的策略去提取不同的变量，这不利于数据分析，且容易产生错误。考虑一下总共有多少数据分析操作会涉及某个变量中的所有数值（每一个集合函数），你就会明白用一种简单而标准的方式来提取这些数值是多么的重要了。整洁数据尤其适用于向量化编程语言，比如 R（R Core Team 2014），因为数据的结构保证了同一观测中不同变量的取值总是成对出现的。

虽然变量和观测的顺序对于数据分析没什么影响，但良好的排序会使查看原始数据变得更简单。我们可以根据数据分析中变量的作用来组织它们：变量的取值是在数据收集的设计时就确定的？还是在实验的过程中测得的？固定的变量描述的是实验的设计，这是我们事先就可以知道的。计算机科学家通常称固定变量为维数，而统计学家习惯用随机变量的下标来表示它们。被测变量即我们在研究中实际测得的。固定变量应该排在前面，然后是被测变量，每个变量都是有序的，这样相关的变量就是连续的。然后，可以通过第一个变量对行进行排序，从而断开与第二个及后续（固定）变量的关联。这是本篇论文中所有表格展示时所采用的约定。

3. 清理脏数据

真实数据几乎常以各种可以想象的方式违背整洁数据的三大原则。虽然有时候你拿到一个数据就可以马上进行分析，但这仅是个例，而非常态。本部分讲述了脏数据的五种最常见问题及它们对应的处理方式：

- 列头是值，而非变量名。
- 多个变量被堆在一列里。
- 变量既以列的形式存储，又以行的形式存储。
- 各种类型的观测单元被存储在同一个表格中。
- 单个的观测单元被存储在多个表格中。

奇怪的是，大部分的脏数据，包括那些没有在上述提及的类型，都可以用少量的工具进行处理：拆分（melting），字符串切割（string splitting），以及整合（casting）。接下来，我会用用真实的数据集来讲解我碰到的每个实际问题，并且告诉大家如何处理它们。这里所用到的完整数据集及相应的清洗代码可以从 https://github.com/hadley/tidy-data 网站上获取，也可以在本文的在线补充材料中找到。

3.1 列头是值，而非变量名

脏数据的一种常见类型是以表格形式展示数据，但是在表格中，行和列里面都有变量，而且列头通常是数值，而非变量名。不过，不得不说这种“脏乱”的排列在某些情况下是很有用的。它为完全交叉设计提供了有效的存储，且如果我们的操作可以以矩阵运算的形式进行的话，采用这种表格，在计算上是非常高效的。这一问题将在第六部分深入讨论。

表 4 展示了部分此种形式的表格，这个数据集研究的是美国宗教信仰和收入之间的关系。数据来自于皮尤研究中心的一份报告，该研究中心是美国的一个智库，专门收集民众对各种话题的态度，从宗教到互联网都有涉及，而且会制作出包含类似于表 4 的数据集的报告。

表 4　收入和宗教数据的前十行（数据来自于皮尤论坛，此处省略其中三列：$75 ~ 100k, $100 ~ 150k, >150k）

religion	<$10k	$10 ~ 20k	$20 ~ 30k	$30 ~ 40k	$40 ~ 50k	$50 ~ 75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486

续表

religion	<$10k	$10 ~ 20k	$20 ~ 30k	$30 ~ 40k	$40 ~ 50k	$50 ~ 75k
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

该数据集包含三个变量，宗教、收入和频数。为了清洗它，我们需要首先“熔化”(melt)或者对它进行堆栈操作(stack)。换句话说，我们需要将列变成行。然而这种操作通常被称作将宽型数据变成长型数据，因为它们不够准确，我会避免使用这些术语。熔化是由已经是变量的列[简称列变量(colvars)]来参数化的。其他列则转化成两个变量，其中一个称为 column 的新变量包含的是重复的列头，另一个称为 value 的新变量包含的是先前分离的列的串联数据值。表 5 的玩具数据集展示的就是这样一个过程。数据集被熔化后就是熔化的数据集。

表 5　一个简单的数据熔化示意(a)中熔化了一个列变量(row)，得到了熔化后的数据集(b)两个表格中的信息是完全相同的，只不过存储方式不同

row	a	b	c
A	1	4	7
B	2	5	8
C	3	6	9

(a)原始数据

row	column	value
A	a	1
B	a	2
C	a	3
A	b	4
B	b	5
C	b	6
A	c	7
B	c	8
C	c	9

(b)熔化后的数据