

数据科学中的 R 语言

王敏杰

2020 年 11 月 10 日

四川师范大学

开场白

谢谢主持人，能听清我说话吗？

各位老师，各位同学，大家晚上好，我叫王敏杰，

- 是四川师范大学的一名老师，
- 是 R 语言 tidyverse 的死忠粉。

今天很高兴、也很荣幸（借百智享和财经联盟平台）为大家一起交流、分享 R 语言。

开场白

开场白

- 非常感谢百智享公司唐总的热情邀请，以及林晗川工程师的细致安排

开场白

- 非常感谢百智享公司唐总的热情邀请，以及林晗川工程师的细致安排

本节课的目的

前面几周，大家已经向很多教授和专家学习了 python 语言。python 语言是一个强大的语言，是数据科学的重要工具，在学术研究和工业领域，都有非常广泛的应用。

经过一段时间的学习，我相信大家应该收获很多，逐步上手了。但也有同学，可能觉得比较难，如果，你觉得 python 语法很难学，IDE 不好用，画图不太好看的时候，可以考虑学习我们的 R 语言，哈哈，有点不怀好意，像来砸 python 场子的感觉。

事实上，任何一门语言的定位都是不一样的，都有它优势和缺点，我们可以根据自己的需求，选择自己喜欢的语言。（人有高矮胖瘦，可以根据自己的身材，选择合适的衣服，没有绝对正确的工具，掌握的最好的，就在好工具）。我想这就是财经联盟平台包容、强大的地方吧，能够为用户提供更多的选择，尤其在多元的时代。

本节课的目的

本节课的目的

本节课的目的，我这里假定，大家都是小白用户，但是今天，我不打算讲具体的语法，毕竟只有一个小时的时间。

我是想，**利用这一个小时的时间，给大家留下两个印象：**

- 第一，R 能给我们生活带来什么？（介绍什么是 R？R 能做什么？我们为什么选择 R）（明确目标）
- 第二，以最快的速度让大家获得一次成功的经验（有个切身的感受，这个对新用户来说，非常关键）

大家在听我胡说八道的过程中，可以先注册一个
<https://rstudio.cloud> 账号，然后登录
<https://rstudio.cloud/project/1847233>

本节课的目的

- R 能给我们生活带来什么 ?
 - R 是什么 ?
 - R 能干什么 ?
 - 为什么是 R ?
- 五分钟上手 R 语言
 - 需要一台电脑
 - 注册一个 <https://rstudio.cloud> 账号
 - 登录 <https://rstudio.cloud/project/1847233>

R 是什么

好，我们开始。

- 首先说说，R 的那些事

R 是什么

R 那些事

也就是说，R 为统计而生的一门编程语言，

- 所以这个属性，决定了它是**统计编程**语言，不是通用性语言

R 那些事

- 1992 年，新西兰奥克兰大学统计学教授 Ross Ihaka 和 Robert Gentleman，为了方便地给学生教授统计学课程，他们设计开发了 R 语言（他们名字的首字母都是 R）。



Ross Ihaka



Robert Gentleman

R 是什么

R 是什么

R 语言是用于统计分析，图形表示和报告的编程语言：

- R 是一个**统计编程语言** (statistical programming)
- R 可运行于多种平台之上，包括 Windows、UNIX 和 Mac OS X
- R 拥有顶尖水准的**制图**功能
- R 是免费的
- R 应用广泛，拥有丰富的**库包**
- 活跃的**社区**(#rstats)

R 的前世今生

- 2000 年, R1.0.0 发布
- 2004 年, 第一届国际 useR! 会议 (随后每年举办一次)
- 2005 年, ggplot2 宏包 (2018.8 - 2019.8 下载量超过 1.3 亿次)
- 2012 年, R2.15.2 发布
- 2013 年, R3.0.2 发布, CRAN 上的宏包数量 5026 个
- 2016 年, Rstudio 公司推出 tidyverse 宏包数据科学当
前最流行的 R 宏包
- 2017 年, R3.4.1 发布, CRAN 上的宏包数量 10875 个
- 2019 年, R3.6.1 发布, CRAN 上的宏包数量 15102 个
- 2020 年, R4.0.0 发布, CRAN 上的宏包数量 16054 个

R 的前世今生

- 2000 年, R1.0.0 发布
- 2004 年, 第一届国际 useR! 会议 (随后每年举办一次)
- 2005 年, ggplot2 宏包 (**2018.8 - 2019.8 下载量超过 1.3 亿次**)
- 2012 年, R2.15.2 发布
- 2013 年, R3.0.2 发布, CRAN 上的宏包数量 5026 个
- 2016 年, Rstudio 公司推出 tidyverse 宏包 (**数据科学当前最流行的 R 宏包**)
- 2017 年, R3.4.1 发布, CRAN 上的宏包数量 10875 个
- 2019 年, R3.6.1 发布, CRAN 上的宏包数量 15102 个
- 2020 年, R4.0.0 发布, CRAN 上的宏包数量 16054 个

The History of R

R 语言发展趋势

从 2019 年 20 名上升到 2020 年 8 名，

- 可能与新冠肺炎大流行有关，可能大家突然发现，R 很原来这么好用

R 语言发展趋势

Jul 2020	Jul 2019	Change	Programming Language	Ratings	Change
1	2	▲	C	16.45%	+2.24%
2	1	▼	Java	15.10%	+0.04%
3	3		Python	9.09%	-0.17%
4	4		C++	6.21%	-0.49%
5	5		C#	5.25%	+0.88%
6	6		Visual Basic	5.23%	+1.03%
7	7		JavaScript	2.48%	+0.18%
8	20	▲	R	2.41%	+1.57%
9	8	▼	PHP	1.90%	-0.27%
10	13	▲	Swift	1.43%	+0.31%
11	9	▼	SQL	1.40%	-0.58%
12	16	▲	Go	1.21%	+0.19%
13	12	▼	Assembly language	0.94%	-0.45%
14	19	▲	Perl	0.87%	-0.04%

TIOBE index

安装很方便

R 语言官方网站一直很朴素。所以，很容易找到下载链接，找到系统对应的版本，就可以了



安装很方便

官网地址: <https://www.r-project.org/>



[Home]

[Download](#)

[CRAN](#)

[R Project](#)

[About R](#)

[Logo](#)

[Contributors](#)

[What's New?](#)

[Reporting Bugs](#)

[Conferences](#)

[Search](#)

[Get Involved: Mailing Lists](#)

[Developer Pages](#)

[R Blog](#)

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

- [R version 4.0.3 \(Bunny-Wunnies Freak Out\)](#) has been released on 2020-10-10.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the [R Consortium YouTube channel](#).
- [R version 3.6.3 \(Holding the Windsock\)](#) was released on 2020-02-29.
- You can support the R Foundation with a renewable subscription as a [supporting member](#)

IDE 很舒服

我喜欢 R，还有一个很重要的原因，就是 R 的 IDE，

- Rstudio, 用起来很贴心
- 同时，Rstudio 是一款可视化编辑器。基于 Pandoc 的 markdown 标准的，良心可视化编辑器只有两款（Typora 和 Rstudio），其中一款就是 Rstudio

现场演示：

- 它有四个窗口：
- 如果用最新的版本 1.4 版本，你会发现很多又好玩、又适用的功能，比如，这里有个 A 图标的，切换到可视化模式，**直接修改表格**，有点像 word 所见即所得

IDE 很舒服

官网地址: <https://rstudio.com/>

The screenshot displays the RStudio interface with several panes:

- Editor**: Shows an R script named "mpg-plot.R" with the following code:

```
1 library(ggplot2)
2
3 ggplot(mpg, aes(x = displ, y = hwy)) +
4   geom_point(aes(colour = class))
5
```
- Console**: Shows the R command history:

```
> library(ggplot2)
> ggplot(mpg, aes(x = displ, y = hwy)) +
+   geom_point(aes(colour = class))
>
```
- Environment**: Shows the Global Environment pane with the message "Environment is empty".
- Output**: Shows a scatter plot of "hwy" vs "displ" with points colored by "class". The legend defines the colors for different vehicle classes:

class	Color
2seater	Red
compact	Yellow
midsized	Green
minivan	Cyan
pickup	Blue
subcompact	Purple
suv	Magenta

两者的关系

R 和 Rstudio 是什么关系呢？

R 好比汽车的发动机，在里面完成计算，最后通过 Rstudio 把（运行的结果，生成的图片）呈现出来，所以 Rstudio 就好比汽车的仪表盘

有时候，我这样比喻，Rstudio 是好看的皮囊，R 是有趣的灵魂

两者的关系

R: Engine



RStudio: Dashboard



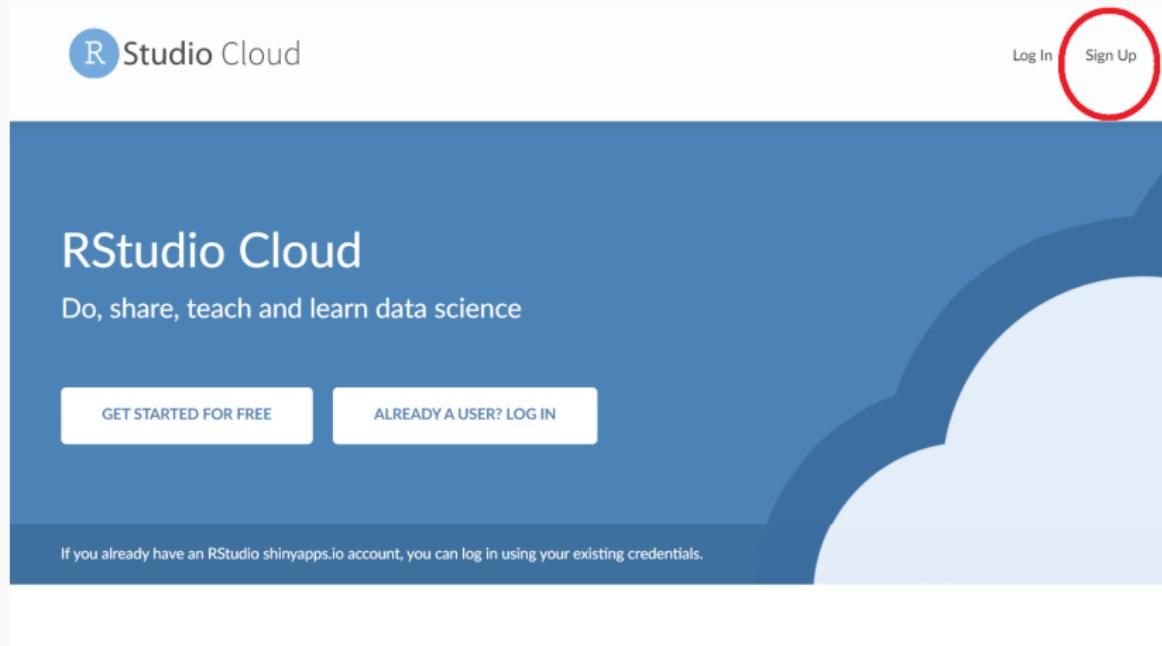
也可以偷懒

安装 R + Rstudio 很简单，就像电脑安装 QQ 一样。

尽管如此，我们还是可以偷懒的，那就是云平台，就是开始，让大家注册的那个网站

云平台有个好处：省去了安装 R + Rstudio 的时间，直接写代码就可以了

也可以偷懶



注册就可使用：<https://rstudio.cloud/>

平台很友好

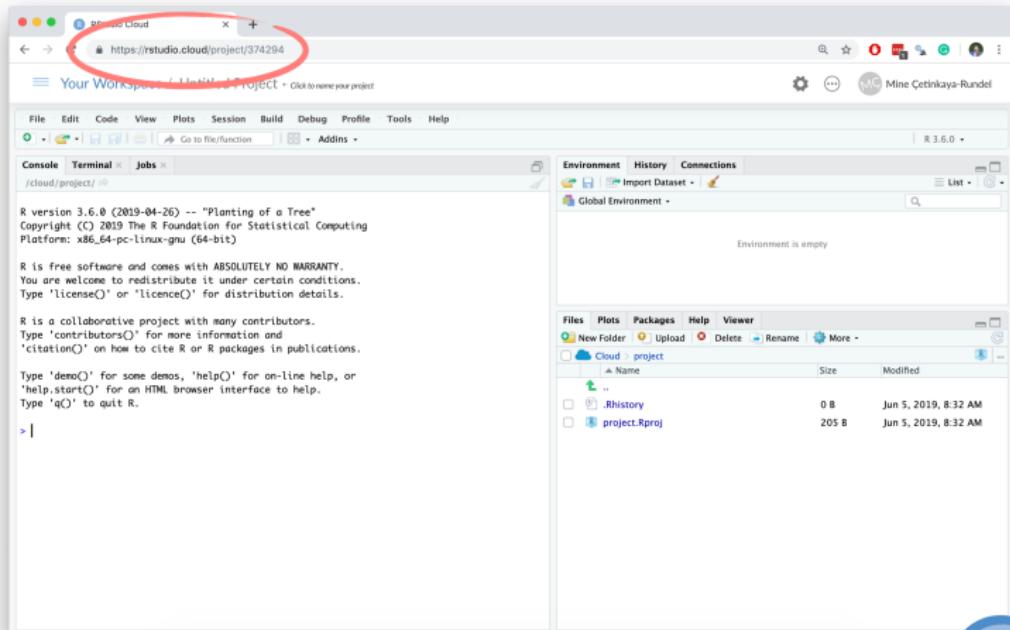
云平台的界面和本地的 Rstudio 的界面是一样的，只不过是在浏览器里，所以最上面多了一行网址

好!

我们接着讲其它的内容，留给一点时间

- 还没有注册的同学，接着注册
- 注册并登录成功的同学，可以打个 +1

平台很友好



R 路上的大神

- 现在，R 语言做（数据探索 + 可视化 + 数据分析），大都喜欢用 tidyverse 包（流行）。
- 个人觉得，数据科学领域里，**R 之所以能和 python 平分天下，主要得益于 tidyverse 的存在**
- tidyverse 是美国 Rstudio 公司（就是我们之前提到的 IDE）首席科学家 Hadley Wickham 和他的团队开发的
- tidyverse 深受用户喜欢，Hadley Wickham 也因此被称为 R 路上的大神，一个改变了 R 语言的人
- 2019 年 8 月，国际统计学年会将考普斯总统奖（被誉为统计学的诺贝尔奖）奖颁给 Hadley Wickham
- 说明 R 语言得到了学术界的充分认可。

R 路上的大神

2019 年 8 月，国际统计学年会将考普斯总统奖（被誉为统计学的诺贝尔奖）奖颁给 tidyverse 的作者



- Hadley Wickham
- R 路上的大神
- 一个改变了 R 语言的人

R 能干什么

那么 R 或者 tidyverse 能干什么事情呢？

听我慢慢道来

R 能干什么

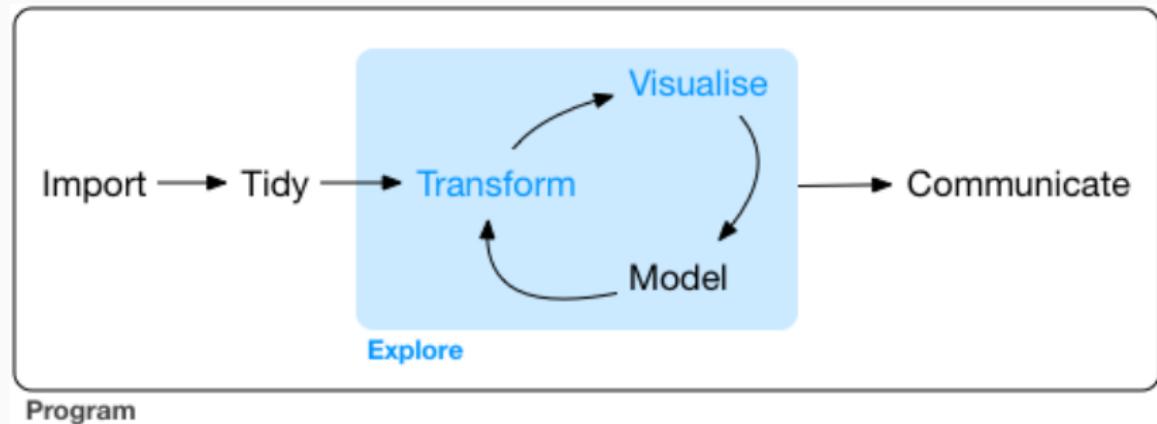
数据科学的流程

刚才讲到了大神 Hadley Wickham，他将数据科学分成了 6 个环节

- 每一个环节，都在代码里完成。
- 代码里完成，有什么好处？最大的好处：可重复性，等会演示给大家看

数据科学的流程

Hadley Wickham 将数据科学流程分解成 6 个环节



tidyverse 套餐

光定义一个流程是不够的，

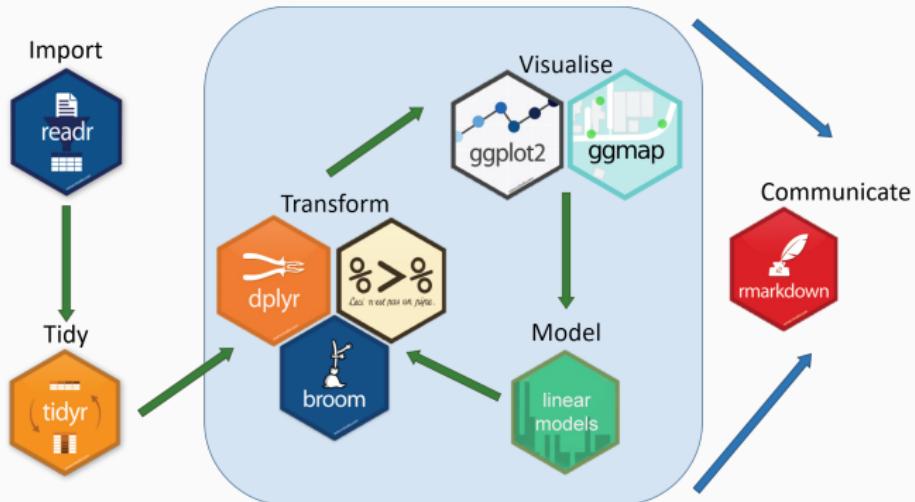
Hadley Wickham 更重要的贡献，就是，**为每个环节都开发设计了相应的宏包**（工具箱）。这样用起来，就很超级方便了。比如读取数据用 `readr...`

这些宏包的集合（这些宏包打个包），就是 tidyverse，即 tidyverse 套餐

因为 tidyverse 套餐是同一个团队开发的，因此 tidyverse 相比其它宏包而言，具有明显的**优势**

- 语法一致性（学习一个宏包，可以帮助理解其他宏包）
- 代码可读性，接近人类语言（`%>%` 太酷了），写代码和说话一样自然

tidyverse 套餐



<https://www.tidyverse.org/>

以最快的速度获得一次成功的经验

打了很长时间的广告，到底好不好，还要看疗效。

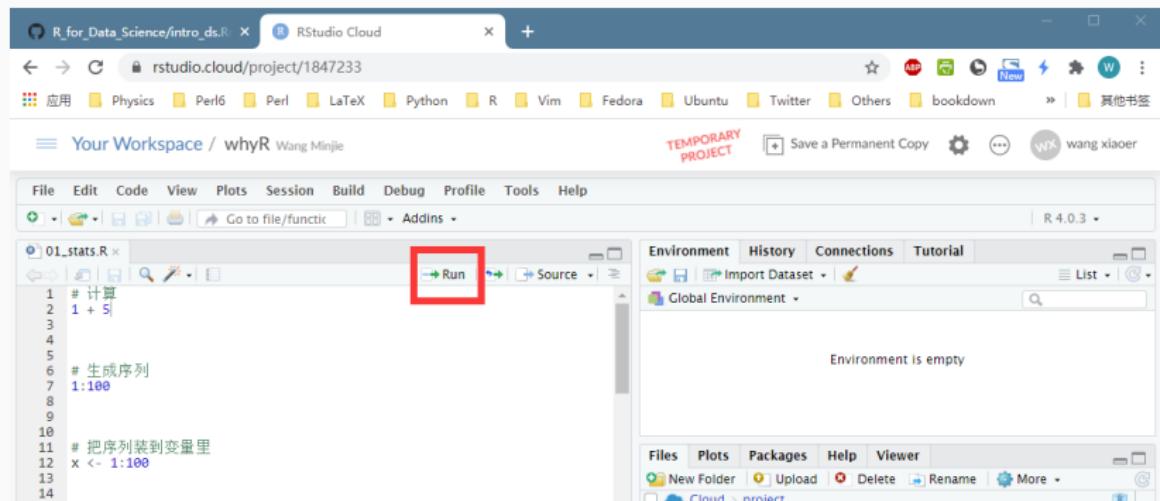
下面我们就通过代码，演示下 R 能干什么。

我们现在一起，大家登录云平台，然后打开链接
就出现这个界面...

这里要演示 4 份代码

以最快的速度获得一次成功的经验

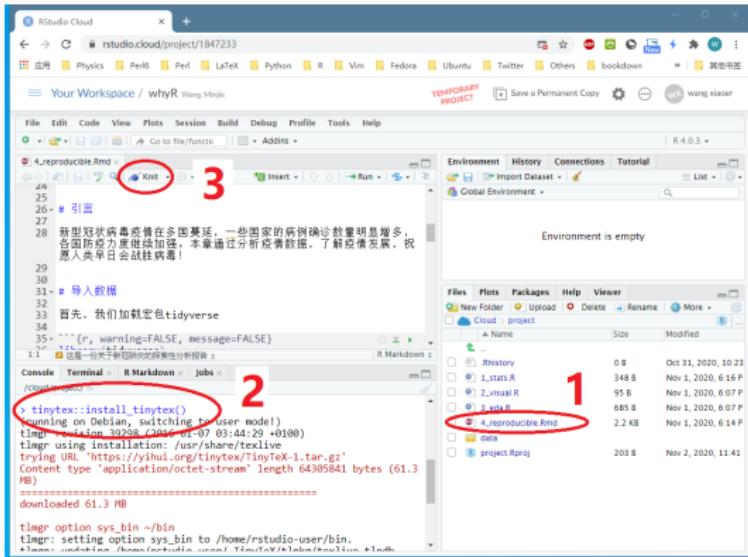
- 登录 <https://rstudio.cloud>
- 打开链接 <https://rstudio.cloud/project/1847233>
- 运行代码（点击右上角的 Run）



生成 pdf

出 pdf 的时候，应该有掌声，尽管我听不到

生成 pdf



- Console 中输入 `tinytex::install_tinytex()` 回车
- 等待 2 分钟，然后重新点击 Knit 就可以看到 pdf 了

难吗？

难吗？

感觉很难吗？

如果是，那说明你认真听了

看了这些代码，可能第一眼感觉是这样的

这是一部科幻电影，讲的是外星人突然降临到地球，并且向人类发出了讯号，这个奇怪的圈圈。

- 我想第一次看 R 代码，和看到外星人的文字，感觉是一样的吧

看了这些代码，可能第一眼感觉是这样的



图 1：图片来自电影《降临》

但我更希望学完后

但我更希望学完后



图 2：图片来自美剧《权利的游戏》

为什么是 R

好，我们接着讲。

- 既然很难，但我们为什么还要选择 R ?
- 因为，**不做难一点的事情，我们怎么超越别人呢？**
- 事实上，我列出了几点理由

为什么是 R

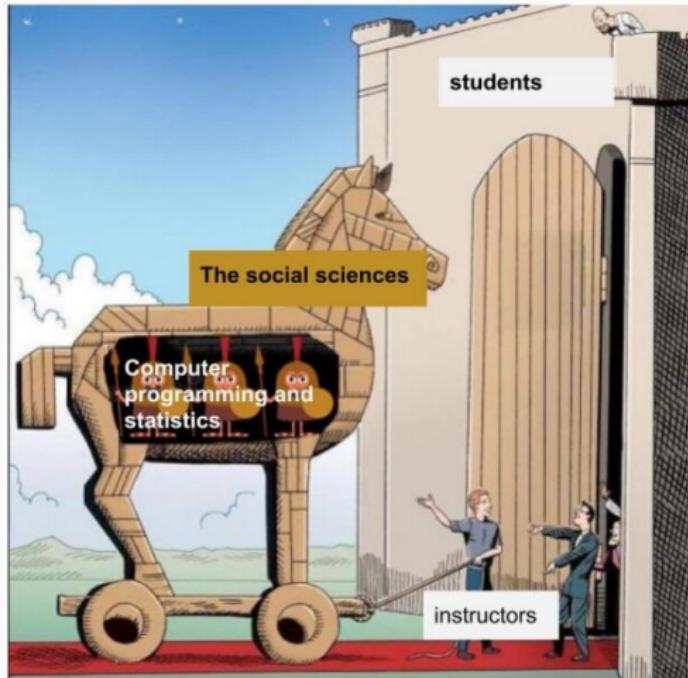
社会科学需要统计

一个学科之所以成为一门科学，必须要有数学作为基础。我说这话，相信很多人会反驳我。我接受反驳。但我还是会坚持我的观点。

很多同学在选专业的时候，导师会说，这个专业不会用太多数学，事实上被忽悠了，尤其在（新文科建设、跨学科研）背景下，社会科学（包括心理学、语言学）都在交叉融合，都需要用数学和计算机。

所以，我们不是学统计的，但需要用统计。**一个更残酷的现实，往往用的统计的，都不是学统计的。**

社会科学需要统计



往往用的统计的，都不是学统计的

社会科学需要可视化

可视化，这个理由很显然了。我们人，都是视觉动物，都喜欢看漂亮美好的东西。如果文章或者报告太多表格，不会给人留下深刻影响；相反，用图片，重点突出，一图胜千言，很容易传递信息。当然，前提是，画图要画的好。

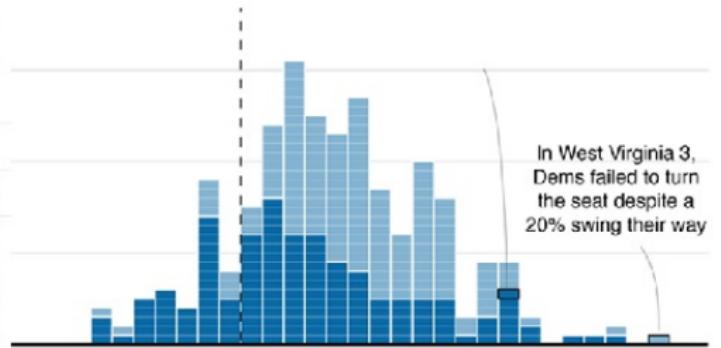
又一个残酷的现实

- 在这个看脸的时代，没有好看的皮囊，没人愿意了解你的灵魂。

社会科学需要可视化



	table.Q1b.- Awareness - %	table.Q4.Brand-Attitude.T2B - %	table.Q3.Preferred.cola - %
Coca-Cola	100.0	73.8	42.8
Diet Coke	100.0	34.5	10.3
Coke Zero	92.7	42.7	17.3
Pepsi	100.0	45.2	9.2
Diet Pepsi	82.3	17.5	2.7
Pepsi Max	91.0	40.7	15.7
NET	100.0	96.0	100.0



没有好看的皮囊，没人愿意了解你的灵魂

社会科学需要编程

为什么要编程，回答这个问题，相当于回答，为什么不能用 excel 做数据分析？

我们这里用这个图，说明这个问题

- 横坐标 = 数据量的大小或者问题复杂程度，
- 纵坐标 = 解决问题的困难程度。

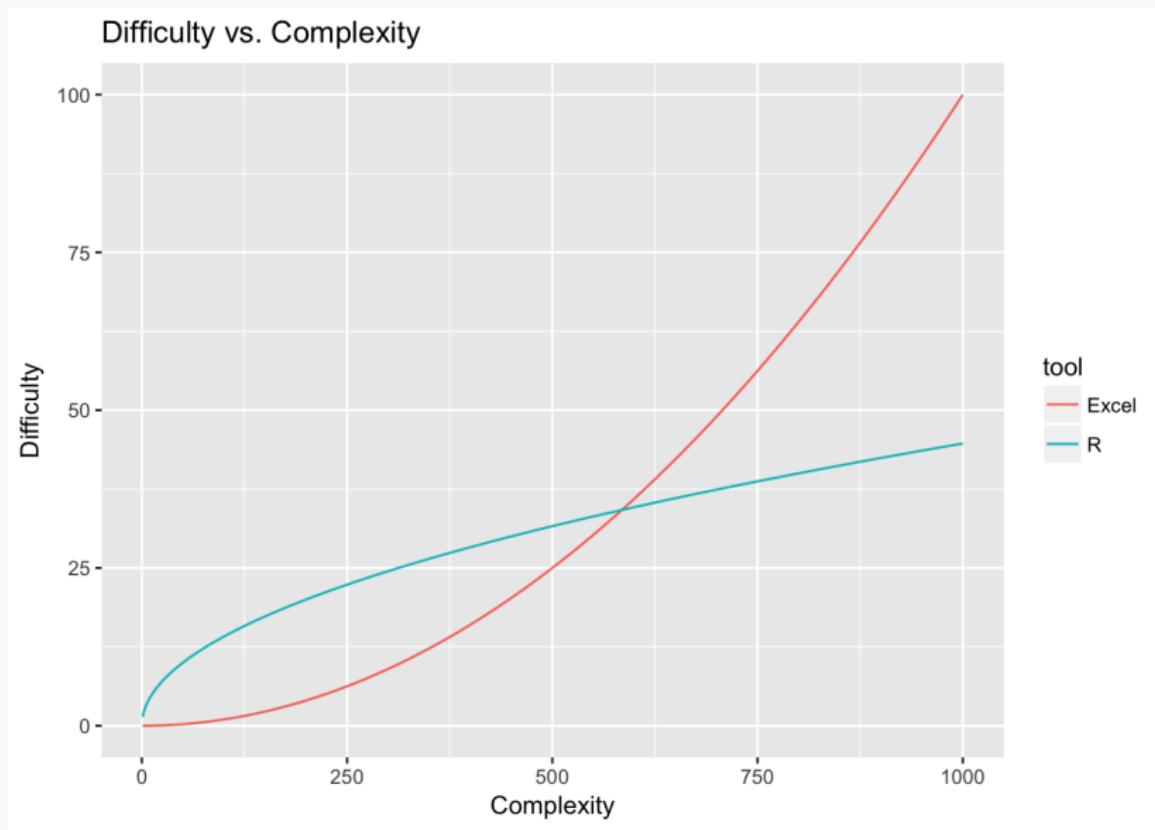
大家看这条 excel 红线，

- 对于数据量不大，或者复杂程度不高的需求来说，比如 10 行 10 列，excel 很方便也很直观，很容易搞定。
- 但，随着数据量或复杂程度不断增大，用 excel 解决起来，难度系数就会陡增，或者无法搞定，这就需要借助编程完成。

从另外一个角度看，掌握了编程技能，比如这里的 R，对于简单的问题和复杂的问题，难度系数是差不多了。

所以，**第三残酷的现实：现在小学生都开始学编程了**

社会科学需要编程



社会科学需要可重复性

科学的可重复性危机，已经成为举世瞩目的热点议题。

- 科研结果**可重复性低**的原因很多很多。
- **不可重复，说明事情没那么简单。**

或许，科学固有不确定性，但需要从研究方法、实验设计和统计方法方面改进

第四个残酷的现实：科学研究的方向是（开放科学框架（Open Science Framework, OSF）），正如 Nature 期刊要求的一样，需要公布原始数据和如何分析的代码

社会科学需要可重复性



EDITORIAL

OPEN

Reproducibility: let's get it right from the start

From September 12th 2018, *Nature Communications* will be setting a higher standard of data reporting for papers under peer review. We believe that sharing raw data at an early stage with editors and reviewers is the best way to build confidence in the reproducibility of your findings. Learn here how to ensure that your paper makes the grade.

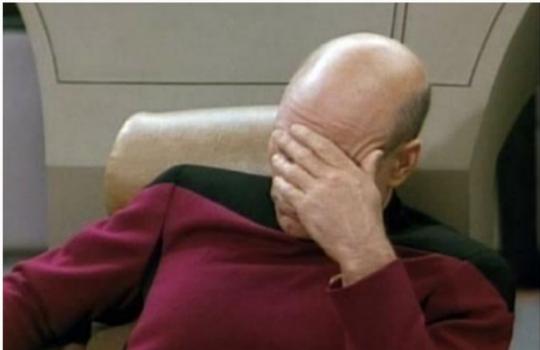
Whether you prefer to call it a crisis, a challenge or a revolution, the growing awareness of reproducibility as an important issue in science is surely a cause for optimism. In biomedical research in particular, journals now compete with each other to

this can occur at a relatively late stage in review, resulting in a waste of time for reviewers and authors alike. Ensuring that reporting is transparent, right from the start of the peer review process, would allow our reviewers to scrutinize the level of support for the findings with greater confidence and at a point where any pro-

论文，要公布原始数据和如何分析的代码

4 个残酷的现实

刚才说了 4 个残酷的现实，很沮丧，何以解忧？

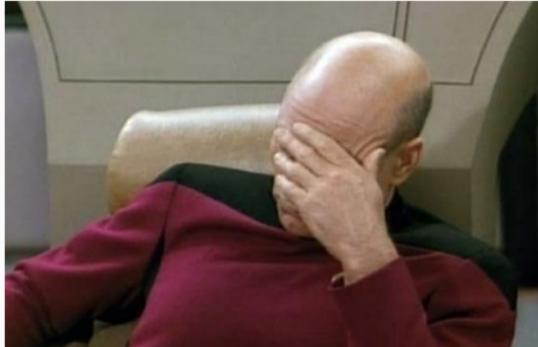


残酷的现实

何以解忧

何以解忧，唯有撸 R

我想，R 语言之美，可以缓解你的压力



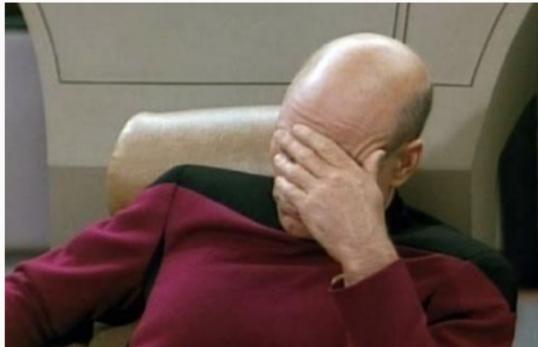
残酷的现实



何以解忧，唯有撸 R

何以解忧

怕怕



残酷的现实



何以解忧，唯有可以撸 R

R 语言之美，你值得拥有

我想，R 语言之美，可以缓解你的压力

- 首先，R 语言做统计分析，是它的看家本领，非常好用（可以缓解第一个残酷）
- 其次，ggplot2 画图，是颜值担当，非常好看，一直被模仿，从未被超越（可以解决第一个残酷）
- tidyverse 来编程，代码可读性强，用的是人类语言，非常好学（在解决残酷现实的同时，还让你感受到乐趣）
- 关于第四点，需要特别说明下，Rmarkdown 并不能保证研究结果可重复性，因为影响结果可重复性的原因很多很多，这不是程序语言能解决的事。但是，R 语言能帮你的，就是减少低级的计算错误和复制粘贴等繁琐工作，可以生成 html、word 或者 pdf 格式的可重复性报告文档，可以方便快捷做幻灯片、海报、论文、书籍、网页。所以还是挺好玩的/

所以，R 语言之美，你值得拥有

R 语言之美，你值得拥有

序号	内容	特性	评价
1	统计分析	看家本领	好用
2	ggplot2 画图	颜值担当	好看
3	tidyverse 语法	人类语言	好学
4	可重复性报告	方便快捷	好玩

当今最值得学习的数据科学语言

所以，这篇文章，旗帜鲜明的指出，**R 语言，是当今最值得学习的数据科学语言**。罗列了很多理由，其中的 3 点理由，我圈出来了（传统的统计学，贝叶斯新统计、数据可视化），我看完这篇文章的感受是：

- 第一、在数据科学领域，python 能做的，R 也能做，甚至更好，比如可视化。
- 第二、有一定 R 基础后，对统计学的学习帮助很大，这是 python 语言不具备的
- 第三、我觉得 R 的语法更符合人的思维方式。

说到思维方式，忍不住想吐槽 python 语言了（云平台）
不能再说 python 的坏话，再说，林总可能要把我提出群了

当今最值得学习的数据科学语言



HOME ABOUT RSS ADD YOUR BLOG! LEARN R R JOBS CONTACT US

Why R is the best data science language to learn today

Posted on January 3, 2017 by Sharp Sight Labs in R bloggers | 0 Comments

[This article was first published on r-bloggers – SHARP SIGHT LABS, and kindly contributed to R-bloggers]. (You can report issue about the content on this page here)

Want to share your content on R-bloggers? click here if you have a blog, or here if you don't.

f Share

t Tweet

In last week's blog, I explained why you should Master R (even if it may eventually become obsolete).

I wrote that article to address people who claim mastering R is a bit of a waste of time (because it will eventually become obsolete).

But when I suggested that R may eventually become obsolete, this seemed to provoke fear that R is becoming obsolete right now.

I want to allay your fears: R is still very popular.

R has been one of the fastest growing programming languages of the last decade.

In fact, if you're getting started with data science, it's still the language that I recommend.

So, I want to reassure you. R is definitely not obsolete. In fact, R is extremely popular and a best-in-class data language.

To that end, I want to explain all of the reasons why I'm very optimistic about R's long term prospects, and why I think it's perhaps the best data science language to learn today.

Learn frequentist statistics with R

The same can be said for statistics books.

Because R has statistics "built into its DNA," many statistics textbooks use R as a learning tool.

For an introductory look at frequentist statistics, here's one excellent book:

- Statistics: an Introduction using R

Again, if you do a quick search on Amazon, and look at many intro stats books, you'll find that if they use any programming language as a teaching tool, they are more likely to use R than almost any other language.

Learn Bayesian statistics with R

This becomes even more pronounced if you want a hands-on book for learning Bayesian statistics.

If you want to learn Bayesian stats and Bayesian analysis, nearly all of the books use R. There are some exceptions, like a few books that teach Bayesian analysis in C or Python, but overwhelmingly the best books that teach Bayesian statistics use R.

If you're interested in Bayesian stats, check out these:

- Introduction to Bayesian Statistics
- Statistical Rethinking
- Doing Bayesian Data Analysis

If you're interested in Bayesian methods, these books are "best in class," and they all use R.

Learn Data Visualization in R

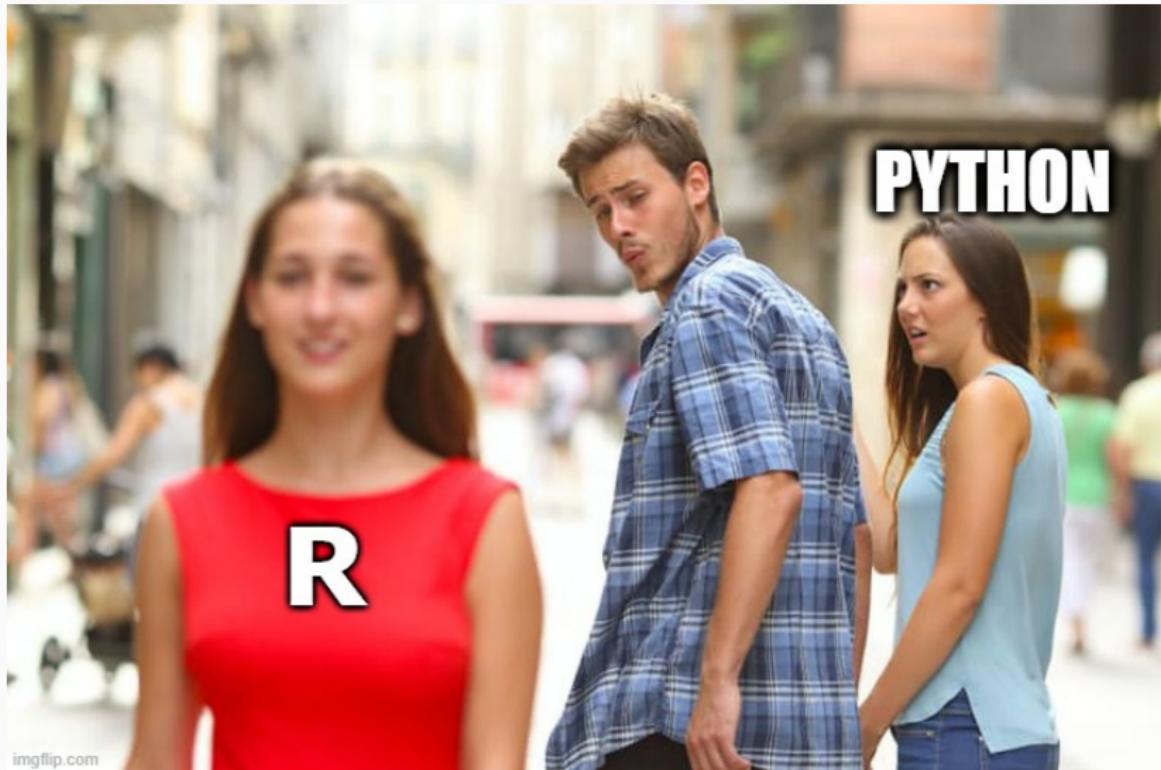
When you're learning data visualization, there's a slightly larger range of programming languages to choose from, but I still maintain that most of the best learning materials use R.

If you're learning data visualization, I highly recommend the work of Nathan Yau. His

一见钟情，还是相见恨晚？

所以，是一见钟情，还是相见恨晚？

一见钟情，还是相见恨晚？



剧情，是这样发展的....



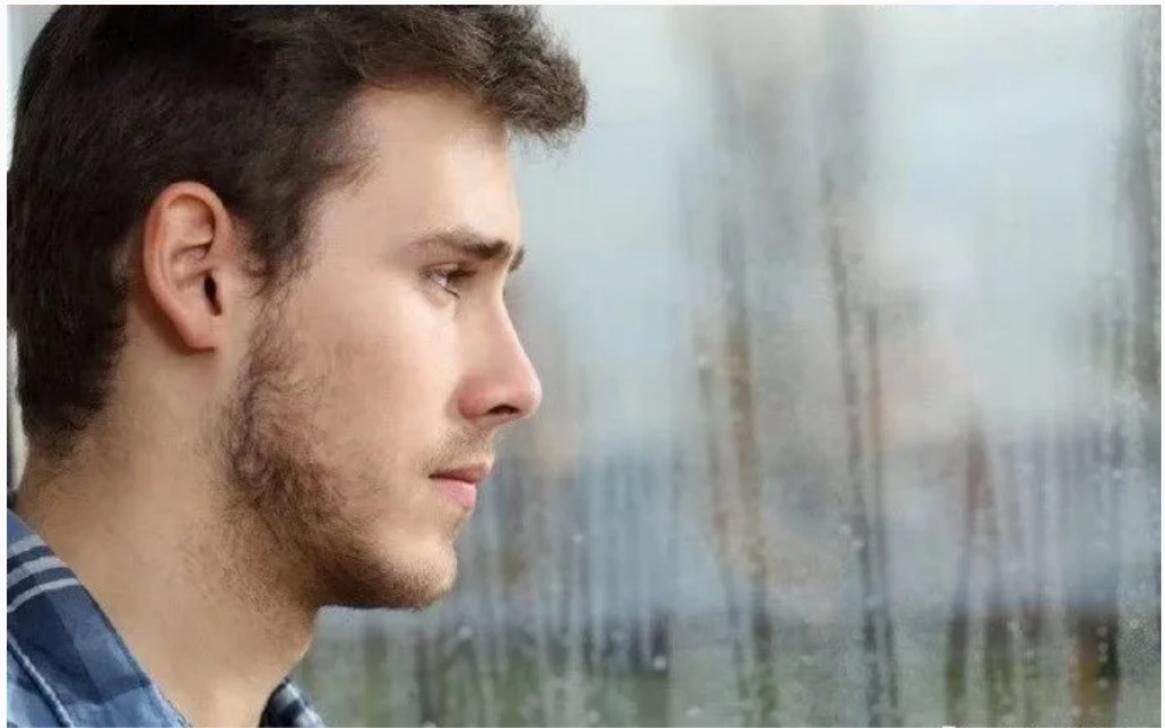
然后，...



但结局，总是出乎人的意料…



我想说的是，R 和 python 实际上是好朋友，他们相互吐槽，也相互学习了很多，比如 python 向 R 学习了数据框的概念，R 也 python 借鉴了爬虫的思想



最后，欢迎大家来到 R 语言的社区。

希望，**大家多用 R**，把 R 当作知识学习的**脚手架**！它就像修房子时候搭的架子一样，你的学科就是这个房子，R 就是好用的工具。我一直比较满意，自己找到的这个比喻词。

最后，再次感谢唐总和林总，感谢同学聆听 /

我今天，在这里给大家讲座，很开心，让我获得了很多激励和灵感，我期待下次，还有很多机会再和大家相聚。

谢谢大家。

Welcome to R, RStudio, and
the tidyverse!