R Without Statistics

David Keyes

Contents

About the Book

This is the in-progress version of *R Without Statistics*, a forthcoming book from No Starch Press.

Since R was invented in 1993, it has become a widely used programming language for statistical analysis. From academia to the tech world and beyond, R is used for a wide range of statistical analysis.

R's ubiquity in the world of statistics leads many to assume that it is only useful to those who do complex statistical work. But as R has grown in popularity, the number of ways it can be used has grown as well. Today, R is used for:

- Data visualization
- Map making
- Sharing results through reports, slides, and websites
- Automating processes
- And much more!

The idea that R is only for statistical analysis is outdated and inaccurate. But, without a single book that demonstrates the power of R for non-statistical purposes, this perception persists.

Enter R Without Statistics.

R Without Statistics will show ways that R can be used beyond complex statistical analysis. Readers will learn about a range of uses for R, many of which they have likely never even considered.

Each chapter will, using a consistent format, cover one novel way of using R.

1. Readers will first be introduced to an R user who has done something novel and learn how using R in this way transformed their work.
2. Following this, there will be code samples that demonstrate exactly how the R user did the thing they are being profiled for.

3. Finally, there will be a summary, with lessons learned from this novel way of using R.

Written by David Keyes, Founder and CEO of R for the Rest of Us, R Without Statistics will be published by No Starch Press.

Introduction

Why R Without Statistics?

How New Zealand Used R to Fought COVID with R

TODO

How I Came to Use R

My own relationship with R goes back to 2016. At the time, I was a consultant, helping non-profits, government agencies, and educational institutions to measure how effective their work is (a field known as program evaluation). A lot of my work involved conducting surveys, analyzing the resulting the data, and sharing these results with clients.

The work itself was fine, but the tools I was using to do it were getting on my nerves. Well, one tool really: Excel.

Now look, this is not a place for an anti-Excel rant. Excel is a fine tool that has empowered millions to work with data in ways they would never have been able to without this tool.

But, for me, Excel was tedious. The amount of pointing and clicking I had to do when working with the amount of data I had got old fast. Each time I would conduct a survey, I'd know that it would yield an avalanche of data and that my wrists would end up exhausted after hours of pointing and clicking on Excel buttons.

No matter what I did, analyzing data and creating charts in Excel just involved a lot of repetitive pointing and clicking.

TODO: Add graph of clicks going up

Endless pointing and clicking was just one problem I faced using Excel. Annoying though it was, it didn't affect the quality of my work. Or so I thought until I recalled a project I had worked on a few years earlier.

In this project, I was looking at which school districts in the state of Oregon have outdoor education programs known as Outdoor School. As part of this project, I had to download data on all school districts throughout Oregon, filter to only include relevant districts with fifth or sixth graders (the ages Outdoor School takes place), and then merge this with data that I collected as part of a survey I conducted.

I did the work in Excel, using a lot of (you guessed it!) pointing and clicking. The problem came when I was almost done with the project. I've blocked the details from my memory (as I've done with most things Excel-related), but what I do recall is that I wasn't 100% certain I had done my filtering and joining correctly. And, to make it worse, I had no way to check. Why? Because all my pointing and clicking was ephemeral, gone in the ether as soon as I had completed it.

I finished the Outdoor School project and submitted my report. I think the work I did was probably accurate, but maybe it wasn't?

Now, you may be reading this thinking: why didn't you write down the steps you used in Excel so you could retrace them later? Sure, I could (and should) have done that. But let's be honest: most of us don't.

And the reality is, we're human. We all make mistakes. And without a straightforward way to audit your work (and keeping a list of all of your Excel points and clicks in a separate document is not, in my view, straightforward), mistakes will happen. If you've used Excel to work with data, I guarantee you've made a mistake, just like me.

The good news is that it's ok. There's a solution. And that solution is called R.

If I were to redo that project on Outdoor School with R, here's what would be different. Rather than watching points and clicks disappear into the ether, I'd write code that would serve as a record of everything I did. This code would:

1. Download data on all school districts
2. Filter to only include districts with fifth or sixth graders
3. Join the filtered data on school districts with my survey data

TODO: Add pseudocode to show this in R

Code can be scary. Having to write code is one of the reasons many people never learn R. But code is just a list of things you want to do to your data. It may be written in a hard-to-parse syntax (though it quickly gets easier to make sense of it), but it's just a set of steps. The same steps that we should write out when we're working in Excel, but never do.

If I had done things this way when working on the Outdoor School project, I could have looked back at any point to make sure what I thought was happening to my data was in fact happening. That nagging sensation I had near the end of the project that I may have made a mistake in one of my early points or clicks? It never would come up because I could just review my code to make sure it did what I thought it did.

Using R won't mean you'll never make mistakes again (trust me, you will). But it will mean that you can easily spot your mistakes, make changes, and rerun your code to fix any issues.

I started learning R to avoid tedious pointing and clicking. But what I found was that R improved my work in ways I never expected. It's not just that my wrists are less tired. I now have more confidence that my work is accurate.

My own uncertainty about the way I use R

I used to feel ashamed about the way I use R.

I use R, a tool for statistical analysis, but I don't use it for complex statistical analysis. I don't do machine learning. I don't know what a random forest is. I've never even run a regression in R.

The only statistics in R are descriptive statistics. Counts, sums, averages: they are the statistics that I do in R.

For a long time, I felt like I wasn't a "real" R user. Real R users, in my mind, used R for hardcore stats. I "only" used R for descriptive stats.

TODO: Does this fit? "It sometimes feels like I'm using a souped up Ferrari sports car to drive 20 miles an hour to the grocery store."

But eventually I realized that this framing misses the point. R started out as a tool created by statisticians for other statisticians (TODO: Add link to R history article). But, over a quarter century since its creation, R can do much more than statistical analysis.

TODO: Remove next two sections?

My background as an anthropologist

Never used R in grad school

I use R for three main things:

Here's what I use R for today:

1. **Data visualization**: making graphs, maps, and tables that look good and communicate effectively.
2. **Reporting with RMarkdown**: moving away from the inefficiency and error-prone workflow of using multiple tools to create reports by instead doing it all in the one tool that I think of as R's killer feature.
3. **Automating tedious practices**: Remember my Excel-burdened wrists? Since I moved to R I've found so many ways to automate tedious practices, from gathering data directly from the U.S. Census Bureau to pulling survey results in from Qualtrics and more.

I think of my work in three buckets:

1. Illuminate
2. Communicate
3. Automate

But then I realized what people get most excited about is:

The main reason I've come to accept that my way of using R is as valid as anyone else's has come through realizing that more "sophisticated" R users are doing many of the same things I am. Sure, they may also be doing statistical analyses that I am not, but everyone who uses R needs to illuminate, communicate, and automate.

Canadian statistician Sharla Gelfand has talked about how they used R to automate an annual report on nursing registration exams in Ontario. Sharla told me in 2019 that, despite being a statistician, the most statistical thing they did was calculating a median.

Take a look at the R community on Twitter (where users congregate under the #rstats hashtag). What gets people most excited is not the latest complex statistical analysis. It's tips and tricks on data wrangling.

TODO: Switch examples below to focus on illuminate, communicate, and automate?

Like how to rename 192 variables without writing 192 lines of code.

Or love letters to the `clean_names()` function from the `janitor` package, which takes messy variable names and makes them easy to work with in R.

Or oohing and aahing over the latest gorgeous data visualization made as part of the Tidy Tuesday project.

TODO: Add Cedric's Spotify viz

No matter what else you do in R, you have to **illuminate** your findings and **communicate** your results. And, the more you use R, the more you'll find yourself wanting to **automate** things you used to do manually (your wrists will thank you).

I realize now that the things that I use R for *are* the things that everyone uses R for. R was created for statistics. But today you can use R without statistics.

Book Overview

How each chapter works

Broad scope of book

- I've tried to choose topics that are relevant to everyone, no matter what you do (e.g. art with R is cool but not everyone wants to do it)

Why didn't you cover X topic?

- That's a great idea, but I can't cover everything!
- The fact that R can do X is a great example of its versatility (please write your own book!)

Book title is not to be taken literally

Pedants of the world (as one of you, I come in peace), I have a favor to ask.

This book is called R Without Statistics. But it's not meant to be taken literally.

Of course it's true that if you're making a graph you're using statistics. So, before you start typing an angry email to me, please know that R Without Statistics is a mindset rather than a literal statement.

We're all using R with statistics already. Let's also learn to use R without statistics.

Illuminate

Use General Principles of High-Quality Data Viz in R

Develop a Custom Theme to Keep Your Data Viz Consistent

R is a Full-Fledged Map-Making Tool

Make Tables That Look Good and Share Results Effectively

Communicate

Use RMarkdown to Communicate Accurately and Efficiently

Use RMarkdown to Instantly Generate Hundreds of Reports

https://urban-institute.medium.com/using-r-markdown-to-track-and-publish-state-data-d1291bfa1ec0

Create Beautiful Presentations with RMarkdown

Make Websites to Share Results Online

Automate

Access Up to Date Census Data with the `tidycensus` Package

Pull in Survey Results as Soon as They Come In

Stop Copying and Pasting Code by Creating Your Own Functions

Bundle Your Functions Together in Your Own R Package

Conclusion

Come for the Data, Stay for the Community