"there is art to data science"

NBME useRs

VOL.II. . . No.1 *FEBRUARY 23, 2021* PRICELESS

Happy 2021!

Welcome back! I hope you all had some enjoyable time off and are having a great start to the new year - especially now that we're almost two months in! One of my goals for the year is to publish at least 6 issues of the NBME useRs newsletter. **I won't be able to make this goal without your help**, so please continue to send me ideas about things that you want to learn and know about R. For this issue I'm going to discuss connecting R to ODBC, as well as random things of interest I learned over the last couple of months at an R workshop and conference.

Using R with Open Database Connectivity (ODBC)[1]

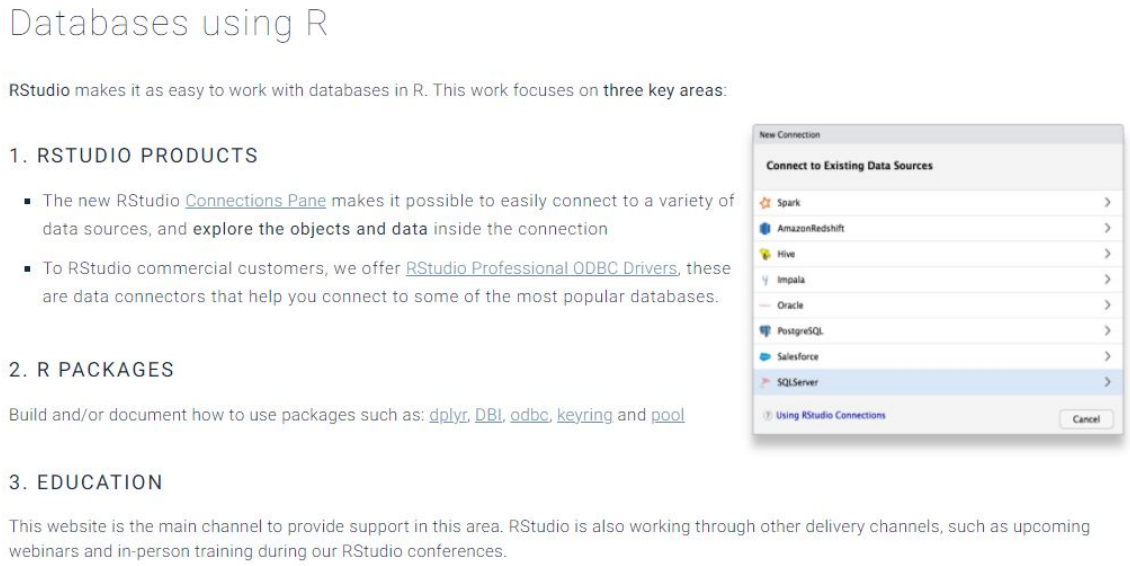

Figure 1: **Databases using R**

Since the NBME announced that they will be moving toward using R and Python for operational work, I've started paying more attention to how can R can be deployed at the level of an enterprise and not just at the individual (or team) level. Fortunately the folks at R Studio (and similar organizations) long ago recognized that supporting such work would be beneficial to their business. As such, there are numerous resources online at RStudio and other places that discuss such endeavors, including using R with ODBC and working in different environments.

[1] Thanks to Laurel Smith for the idea to cover this topic!

An short list of some related helpful resources:

- **Databases using R** This is the main landing page for RStudio resources for connecting to databases with R (pictured on the first page). The left-hand side of this page (not pictured) has information on the R Packages that have been developed to facilitate database connections. This is followed by information on Professional Drivers that can be used (these are *not* free services), and then a section on best practices when using database connections.
- **Connect to a Database** This RStudio page is a basic introduction to connecting to a database with R, complete with R package downloads and examples along the way.
- **Databases** This page provides a list of database types that can be connected via R. Clicking on the name of the database (Apache Hive, MySQL, SQLite, Salesforce, etc) takes you to an example page of how to establish a connection with that database.
- **SQL server connections in R** This is a blog post from NHS-R, a community of R users that work with or analyze data from the National Health Service in England.
- **Database Connections with R Webinar** This is a 60-minute webinar that was presented by NHS-R in 2020 that is a nice overview on connecting databases with R. Jump ahead to the 5:30 mark to get to the start of the actual webinar presentation. I am a novice with connecting to databases, and I learned a lot watching this webinar. The materials and scripts that the presenter uses in the webinar are available at their github page.

In general, I would encourage you to work with your teammates to help navigate the best setup for your team / use case to use when connecting R to different databases, as I would imagine this will vary across teams at NBME.

RStudio Table Competition

In late December RStudio announced the winners of their very first table contest. Like all RStudio competitions, I'm amazed at the wide variety of ways that people think about using R. Almost every entry has some feature that's inspiring and novel, and makes me think about the ways that I could construct more informative tables.

One table of note is this `tidyverse` table about the different packages and functions that can be accomplished in R by using the `tidyverse` package. The table is interesting both because of its content and the beauty of its construction.

Figure 2: A beautiful, but uninformative table.

Featured R packages

R package: `renv`

renv pdf vignette

R dependency management.

Suggested audience: Teams that need to ensure package updates and related software changes don't disrupt normal operational workflows.

One topic that was often mentioned during conference talks was R package dependency management. As you may know, R packages are often updated by their developers for a variety of reasons, most common of which are extending package functionality and bug fixes (one requirement of having your package hosted on the official CRAN network is promising to fix bugs that are discovered by users). A common problem that can occur when you are using multiple R packages for a project is that when you update your packages, somewhere the packages no longer "talk" to one another and your program no longer functions as intended. And the problem may not be caused by the package you think is causing the problem; many packages have *other* packages on which they depend, and updating the main package also updates the packages on which the main package depends (see below). When programs fail to run in research or other low-stakes situations, it can be annoying and frustrating, especially when the fix isn't quickly evident. When operational work is interrupted by errors caused by package updates, it can be even more problematic.

AER: Applied Econometrics with R

Functions, data sets, examples, demos, and vignettes for the book Christian Kleiber and Achim Zeileis (2008),

Version:	1.2-9
Depends:	R (≥ 3.0.0), car (≥ 2.0-19), lmtest, sandwich (≥ 2.4-0), survival (≥ 2.37-5), zoo
Imports:	stats, Formula (≥ 0.2-0)

Figure 3: Example of package dependencies with the AER package.

Fortunately, this problem is not a new problem and there have long existed *dependency management* tools as one possible solution. Such tools help you reproduce the environment where you initially built your program to ensure that package updates don't break the functionality of an important process. The `renv` package is the newest package developed by RStudio to help accomplish this task (`packrat` was the previous package developed for this purpose). By integrating `renv` into your programming, you can have more confidence that your program will consistently work as intended over time, as well as on other machines! I highly recommend looking at the Introduction to renv page to learn more.

For those interested in more information about R environments, namespaces, exports, imports, etc., I found this blog post to be very informative and well worth the time to read.

Conference Highlights

I was fortunate enough to virtually attend a few conferences in January and February. What I enjoy about virtual conferences: (1) they're (mostly) free; (2) because they're free, I don't feel bad if I can only make 1 or 2 talks because of my schedule and other work demands; and (3) most conferences have been posting the talks online afterward. Here are a few of the talks that I thought to be most helpful and/or interesting.

ESMAR (Evidence Synthesis & Meta-Analysis in R)

Workshop: R functions and packages

This workshop talks about R functions and packages differently than other presentations I've seen on the same topic. Diagramming the work flow of functions and packages was new to me, and seems to have some utility. All of the other ESMAR talks can be found on this YouTube page. Topics for other talks include automation of tasks (such as text analysis), data visualization, an introduction to Github, etc.

restudio::global 2021

▷ Maintaining the house that tidyverse built

This is a talk on code maintenance by Hadley Wickham, the Chief Scientist at RStudio and someone whom I admire greatly. Admittedly the topic isn't "sexy" (by Hadley's own admission), but I always enjoy hearing Hadley talk about anything. The logic and clarity in his thought is remarkable, and this shows up in the design of his packages and in the elegance of his coding. If you haven't ever heard Hadley give a talk, are curious, and have 30 minutes to spare, I'd suggest checking it out. Here is a YouTube playlist of some of Hadley's talks.

▷ The rest of the rstudio::global 2021 talks can be found here

The rest[2] of the rstudio::global 2021 talks short (5-20 minutes) and cover a large breadth of topics. RStudio has categorized the talks into the following broad categories: Data for Good, Language Interop (R talking to other languages, such as Python and SQL), Learning, Modeling, Organizational Tooling, Package Development, Programming, Teaching, and Visualization. **If anyone sees a talk that they particularly like and think others would find interesting, please let me know!**

Google Fonts

This last bit of information isn't related to R, but is too cool not to share. I've always found fonts to be pretty interesting (check out the Helvetica documentary), but haven't recently taken the time to look for free fonts online that might be useful for documents / projects. Over the holiday season I learned about Google fonts, an incredible collection of beautiful fonts that are freely available to download and install. Go and explore and find your favorite (free!) new font!

[2]Those talks that are *not* Keynote talks.