What is the tidyverse?

2017-06-08

by Joseph Rickert

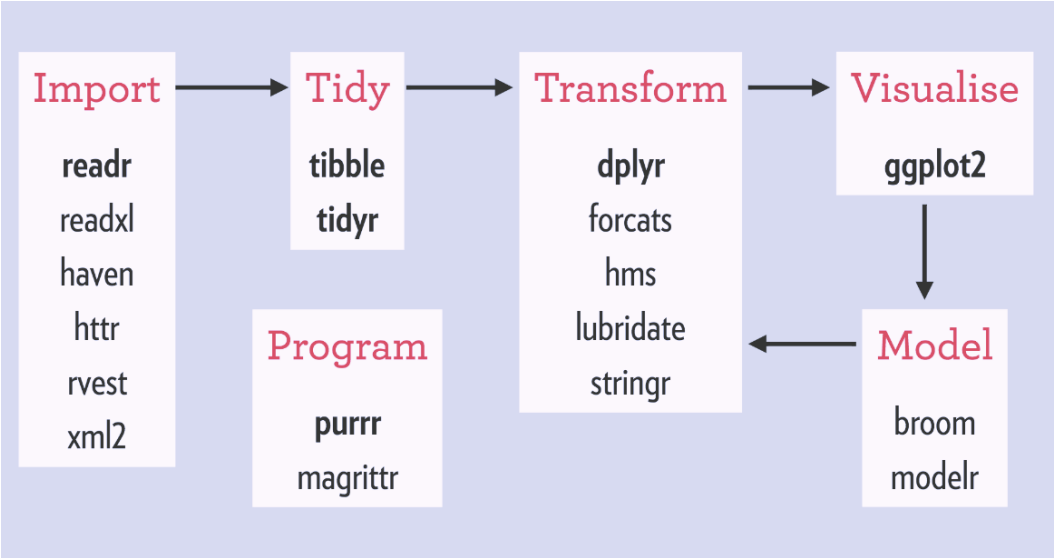
Last week, I had the opportunity to talk to a group of Master's level Statistics and Business Analytics students at Cal State East Bay about R and Data Science. Many in my audience were adult students coming back to school with job experience writing code in Java, Python and SAS. It was a pretty sophisticated crowd, but not surprisingly, their R skills were stitched together in a way that left some big gaps. Many for example, didn't fully understand the importance of CRAN Task Views as curated source for the best packages to support their work in machine learning, time series and the other areas of Statistics they were studying. So, it made sense that even though `ggplot2` and `dplyr` were mentioned in some of the student's questions, a faculty member present asked: "What is the tidyverse?" in an attempt to cover an area that he knew was one of those gaps.

There is an incredible amount of good material available online about the tidyverse, and I will point to some of that below. But here, I'll elaborate on the answer I gave during the Q&A.

The Basics

The tidyverse is a coherent system of packages for data manipulation, exploration and visualization that share a common design philosophy. These were mostly developed by Hadley Wickham himself, but they are now being expanded by several contributors. Tidyverse packages are intended to make statisticians and data scientists more productive by guiding them through workflows that facilitate communication, and result in reproducible work products. Fundamentally, the tidyverse is about the connections between the tools that make the workflow possible.

It is also the case that the tidyverse is work in progress. You can find the current state of development at tidyverse.org. Clicking on the icon for each package on this website will bring you to detailed documentation for each package. The following figure illustrates a canonical data science workflow, and shows how the individual packages fit in.



If you have some experience with R, you ought to be able to jump right into the online documentation and find your way around. If you are new to R, and maybe new to data science as well, you can't do any better than work through the book R for Data Science by Hadley Wickham and Garrett Grolemund.

Advantages

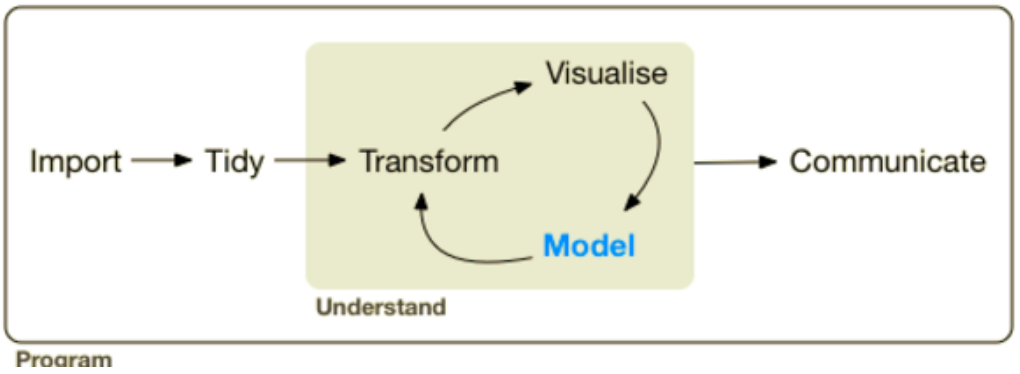
The advantages of the tidyverse include consistent functions, workflow coverage, a path to data science education, a parsimonious approach to the development of data science tools, and the possibility of greater productivity.

Consistency

The tidyverse aspires to consistency on multiple levels. Examples of “micro”-level consistency include the convention of having variable names glide along in `snake_case`, and the signatures of tidyverse functions follow a regular pattern. (The first formal argument is always a data frame that provides the function’s input.) Higher-level consistency includes the idea of tidy data - a data frame where each row is an observation and each column contains the value of a single variable - and the way in which the pipe operator, `%>%`, channels the flow of tidy operations. Under the covers, there are even more levels of structure that aid the pursuit of consistency, including uniform standards for package organization, testing procedures, coding style, etc.

Coverage

The workflow shown above, with tidyverse packages associated with the various steps, or more usually rendered with the following iconic tidyverse diagram, preceded and motivated the development of the tidyverse.



It is an abstraction of the canonical data analysis workflow that has always guided statisticians, but now informs data science as

a map to organize, streamline, automate and optimize the various processes involved. The fact that tidyverse packages are associate with all of the processes indicates that it comprises enough fundamental building blocks to support the entire end-to-end workflow for a variety of data sources and analysis goals. Moreover, the relatively recent addition of the `purrr` package extends the reach of the tidyverse to support the creation of new data science tools.

Critical Mass

A great strength of the R language is that with over ten thousand user contributed packages on CRAN alone, it has a lot to offer. This kind of organic growth makes it inevitable that packages will offer overlapping features. Users have to make decisions about which package, or suite of packages, they will make the effort to learn. For many users, the decision hinges on whether a collection of packages visibly supports important work. Does it have a large community of users and is it backed by committed developers and maintainers? All of the signals indicate that (at least, among R-using data scientists) the tidyverse has reached critical mass. For example, the tidyverse package has been downloaded 50,000 times in the last month. Moreover, it appears that tidyverse principles are propagating into other application areas. The tidyquant package, for example, is a serious attempt to bring tidy principles to Finance.

Education

A typical R user gets involved with R in the first place through a desire to compute in some quantitative field. The path to R competency frequently begins with mastering a small number of relevant functions. Statisticians, for example, may learn to read in data from a `.csv` file and build a linear regression model with `lm()`. Financial analysts may be introduced to R through a package like `quantmod`, which enables a new user to do quite a bit of real work. The tidyverse provides the path of least resistance, or "pit of success", for data scientists interested

in R. For example, the small number of compatible building blocks provided by dplyr enable even a relatively inexperienced user to tidy up a messy data set quickly and easily.

Parsimony

The packages and functions of the tidyverse are the result of trial-and-error experimentation carried out over several years, to find a minimum set of functions that are sufficient to enable the canonical data science workflow. Those of you who have been following Hadley's work will remember `cast()` and `melt()` from the `reshape` and `reshape2` packages, and `ddply()` from the `plyr` package, which were early attempts to find a vocabulary for wrangling data frames. After several attempts to identify and construct the most advantages set of primitive building blocks, the tidyverse has matured into its present form.

Productivity

Hadley has always been clear that a major goal for the tidyverse - and indeed much of his work over the years - has been to help anyone who needs to analyze data work productively, and he is fond of quoting Hal Abelson: "Programs must be written for people to read and only incidentally for machines to execute". My take is that a major reason for the popularity of tidyverse packages is that they help people achieve and maintain flow in their daily data analysis work.

Some Limitations

The tidyverse, of course, is not without limitations. Some of these are due to factors that are beyond the designer's control, and others may be by design. Limitations of the first kind may arise from a lack of agreement as to whether some data can be, or should be, forced into a "rectangular" data structure. For example, although there are scientists and data scientists working in genomics that are fans of `dplyr` and `ggplot2`,

much of the work done in the Bioconductor Project remains outside of the tidyverse workflow.

The need for the close coordination of tidyverse packages produces some limitations of the second sort. There are many high-quality R packages that are of great use to data scientists, but based on design goals that differ from those of the tidyverse. There will always be more than the tidyverse.

A Bigger Picture

A powerful, but perhaps under-appreciated, capability of the R language is its ability to support the design and programming of Domain Specific Languages. Joe Cheng highlighted this feature in an interview he gave to R Views last year. He described R as being "shockingly close to LISP", of which Joe says: "it's almost like you change the language itself to be a DSL for whatever problem you're trying to solve ... the elegant, terse syntax of dplyr and the pipe operator are possible because of how malleable a language R is, and how great it is for writing DSLs in it."

So, from a wider perspective, the tidyverse can be seen as sub-dialect of the R language that is evolving to express ideas and tasks inherent in Data Science workflows and software development. This dialect may not be for everyone, but it does seem to be helping many R fluent data scientists frame their conversations.

Some Resources

The following are some resources that you may find helpful in learning and mastering the tidyverse.

- The video of Hadley Wickham's Keynote address at rstudio::conf 2017
- The slides corresponding to the above video

- R for Data Science by Hadley Wickham and Garrett Grolemund
- Text Mining with R
- Getting Started with the Tidyverse in R

R Language · R Packages · RStudio · Data Science · Opinion Comments Share
R · tidyverse · dplyr · ggplot2

OLDER

A Shiny App for Exploring Commodities Prices and Economic Indicators, via Quandl

NEWER

Mapping Quandl Data with Shiny

You may leave a comment below or discuss the post in the forum community.rstudio.com.

Comments Community 1 Login

Recommend 11 Tweet Share

Sort by Best

Start the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name