'Master the Tidyverse' - conversion of course to use pharma industry data

Mike K Smith

2024-04-26

Signatories

Project team

Mike K Smith (Pfizer) - Lead;
Natalia Andriychuk (Pfizer) - Co-lead;
Rajni Bhaya (Pfizer) - Contributor

Contributors

Mine Cetinkaya-Rundel (Posit);
Garrett Grolemund (Posit)

Consulted

Mehar Pretap Singh (Procogia; Chair, R Consortium Board);
Joseph Rickert (Procogia; Executive Director, R Consortium Board);
Henrik Bengtsson (R Foundation; ISC Chair);
Sumesh Kalappurakal (Janssen; R Consortium Board);
Jared Lander (Lander Analytics; R Consortium Board);
Michael Lawrence (Genentech; R Foundation; R Consortium Board);
Uday Preetham Palukuru (Merck; R Consortium Board);
Vincent Shen (Roche; R Consortium Board);
David Smith (Microsoft; R Consortium Board);
Hadley Wickham (Posit; R Consortium Board)

The Problem

Posit's "Master the Tidyverse" training is a tried-and-tested course to introduce data scientists to the {tidyverse} workflows and associated packages. https://www.tidyverse.org/learn/

The course has been presented to thousands of learners worldwide by Posit trainers and certified trainers https://education.rstudio.com/trainers/. The course materials are available at a Github repository and are made available to modify under a Creative Commons CC-BY-SA-4.0 license. https://github.com/rstudio-education/remaster-the-tidyverse

Datasets used in the training are general and typically served via R packages for ease of sharing and availability. With R adoption growing within the pharmaceutical industry it may be advantageous to swap out the datasets used for example datasets specific to the pharmaceutical industry - clinical trial data or data associated with pharma industry processes. However, data sharing outside of individual companies is very difficult for reasons relating to intellectual property and patient confidentiality.

It would be advantageous to work on the conversion of the datasets used in this course in an open-source environment in order to avoid duplication of effort across companies and organisations.

The proposal

Conversion of 'Master the Tidyverse' course material to use pharmaceutical industry data examples.

Overview

We request a Github repository to be set up under the R Consortium Github organisation (https://github.com/RConsortium). Having a central location for material will facilitate contributions across the community via pull-request and a way to centrally track Issues and Discussion.

Detail

PhUSE have provided a Github repository of synthetic data which can be used as the basis for datasets used in the course. Because this data is synthetic and in the public domain (under MIT license), we can craft examples which can be shared publically and across companies and organisations without concern of intellectual property concerns or patient confidentiality.

We propose:

- To create example datasets for use in the "Master the Tidyverse" training material.
- To create an R package containing these example datasets (for documentation and ease of distribution). The R package will contain data, but also test cases based on the training material to ensure that the data provided works with the example code in the training materials. These test cases should be updated as the training material evolves.
- To convert the training material (slide-ware and example code) in the "Remaster the Tidyverse" Github repository (which stores the master versions of trainer material used in the "Master the Tidyverse" training) to use these example datasets.
- To convert the training material above to use Quarto, rather than Keynote. This will facilitate updating the material and testing code within the slides against latest {tidyverse} versions. Ultimately it might be possible to use {webR} to allow live coding by trainers within the slide presentation.

Project plan

The proposed initiative should aim to complete within 12 months of initiation.

Start-up phase

- Fork the "Remaster the Tidyverse" repository to the RConsortium Github repository.
- Identify datasets used in the training and their context within the training.
  - Set up discussion for each dataset to identify features and attributes of each dataset used in the training contexts
- Identify SDTM / ADaM data that could be used in place of the existing example datasets.
  - Identify any pre-processing / summarisation that might be required to convert the original data to a minimal example dataset that could be used in training.
- Define Issues in Github repo to capture tasks required.
  - Use tags to identify the nature of the work involved. This will facilitate engagement with the broader community to contribute, as they will know what pre-requisite skillset is required to contribute.
- Define process of review and feedback for Pull Requests
- Craft Code of Conduct for the Repository.

- Define LICENSE file for the Repository - Creative Commons Attribution Share Alike (CC-BY-SA 4.0) license, to match originating repository.
- Set up a governance team who can review progress and feedback to ISC, schedule meetings.
- Draft communication to the community to promote the proposed initiative and disseminate - blog posts, slides, etc.

Technical delivery

- Fork the repository (**within 1 week of start date**)
- Identify datasets used in the Master the Tidyverse training course and their context. (**within 1 month of start date**)
  - Document in Discussion threads
  - Include code snippets / reprex for context
  - Identify features and attributes of the data that motivates their use in the example code.
  - Encourage discussion of the examples and identification of SDTM / ADaM / pharma data that could be used to substitute for the original data.
- Identify proposed data examples (based on SDTM / ADaM) that will be used to replace the existing datasets. (**within 3 months of start date**)
  - Features and attributes of the updated pharma example datasets
  - Code to produce these example datasets from the starting point of the PhUSE synthetic data.
- Update training materials using proposed data examples (**within 9 months of start date**)
  - Update slideware - text, screenshots, visualizations, tables
  - Update example code
- Prepare R package of example datasets used (**within 12 months of start date**)
  - Create R package
  - Create documentation for each dataset - context, attributes.
  - Create {testthat} test scripts for each example code snippet to ensure that code works with the associated data.
- Convert existing "Master the Tidyverse" slideware to Quarto (**within 6 months of start date**)
  - Prepare for conversion of the slideware to Quarto by starting with the existing training material.
- Delivery of "Master the Tidyverse" pilot course using new pharma dataset examples (**within 12 months of start date**)

Other aspects

- Announcement of proposed initiative via R Consortium blog post / LinkedIn / X (Twitter) / Mastodon
  - Invitation to contribute through Github repository
- Announcement of progress of proposed initiative at UseR! (July 2024), R in Pharma (Oct 2024), PhUSE EU Connect (Nov 2024)
- Feedback to ISC as required.

Requirements

People

We need some oversight of the Github repository to review any Pull Requests and administer the repository. So one or two individuals will need to be administrators of the Github repository.

We need individuals familiar with the training material of the "Master the Tidyverse" course to identify the features and attributes of the data required to illustrate the technique or functions used in the training. This would ideally be a certified tidyverse trainer.

We need individuals familiar with CDISC SDTM and ADaM dataset structures to help craft the example data using the PhUSE synthetic datasets as the starting material. These individuals will also help craft

documentation for the resulting datasets to help learners understand the context of the data they are working with.

We need individuals familiar with crafting R packages (preferably packages with data content) to help build an R package for disseminating the example data. These individuals should also craft test scripts to test the functionality described in the training material with the proposed example data to ensure that the example code works against the datasets provided.

We need individuals to update the training materials in the "Remaster the Tidyverse" Github repository to reflect the changed example datasets.

We need individuals who could convert the training materials from Keynote presentation format to Quarto.

We propose to engage the community and find individuals to assist through R Consortium blog post, R in Pharma announcement, engagement via Posit Data Science Hangout.

Mike K Smith (Pfizer Ltd) and Natalia Andriychuk (Pfizer Ltd) will lead and drive the initiative forward, but we rely on contributions and input from individuals across the industry, from certified tidyverse trainers within the industry and from Posit Education team members.

Garrett Grolemund and Mine Cetinkaya-Rundel have expressed their willingness to provide advice and input to the proposed work. Their contribution is advisory and on an as-needed basis, but as principal authors of the original material their guidance is very welcome.

Processes

We will use the Github Discussion and Issues features to plan and manage tasks. Contributors will fork the Github repository, craft their contributions then submit back to Github via Pull Request. Their contributions will be noted in the Package DESCRIPTION file and in training material acknowledgements.

A Code of Conduct for the repository will be prepared.

We will provide updates to the ISC as needed.

Work on the project will be open-sourced from the outset under a Creative Commons CC-BY-SA 4.0 license, to match the "Share Alike" requirement of the original material.

Tools & Tech

A Github repository set up under the R Consortium organisation, with Mike K Smith and Natalia Andriychuk as administrators.

Funding

Funding of this initiative is not strictly required. However we would welcome the ISC's suggestions on whether having funding might expedite creation of this repository since we would be able to fund resources from SME CRO organisations to contribute time and effort towards the effort.

Summary

The requirements for this proposal are largely expertise and programming resources to prepare the datasets, craft the R package, adapt the slides to Quarto and update with the new example material. A Github repository is required to collate and coordinate work across the proposed initiative.

Success

Definition of done

- Updated training material using pharma industry datasets to motivate {tidyverse} concepts and examples.
- Datasets used are contained in an R package with associated documentation to facilitate easy dissemination and distribution.
- Training material using Quarto templates rather than Keynote.
- Updated material made available via Github
  - Easy setup for courses via setup.R script
  - Installation of R package containing data with minimal dependencies (over and above those packages required for the course)
  - Deployment to Posit Cloud

Measuring success

- "Master the Tidyverse" course using the updated examples and training material held.
- Positive feedback from course delegates about the relevance of example data to their day-to-day job.

Future work

- Could enhance the updated Slideware using Quarto to HTML / revealjs and extend to use {webR} for live-coding examples.
- Could use similar approach to convert {tidymodels} training to use pharma industry example data.
- Could extend to {pharmaverse} training material providing a common template across both training course materials.

Key risks

- PhUSE synthetic data does not adequately illustrate concepts needed by the training.
- Low level of engagement and input from the community to move this proposed initiative forward.
  - Takes longer than proposed