Report of developing Learningtower R package

Shabarish Sai Subramanian
Master of Businss Analytics

Guan Ru, Chen
Master of Businss Analytics

Report for
Department of Econometrics & Business Statistics, Monash Un

8 October 2024

MONASH
BUSINESS
SCHOOL

**Department of
Econometrics &
Business Statistics**

1 Background

1.1 Introduction

PISA

The Organization for Economic Cooperation and Development OECD is a global organization that aims to create better policies for better lives. Its mission is to create policies that promote prosperity, equality, opportunity, and well-being for all. (Organization for Economic Cooperation and Development 2021a) PISA is one of OECD's Programme for International Student Assessment. PISA assesses 15-year-old students' potential to apply their knowledge and abilities in reading, mathematics, and science to real-world challenges. OECD launched this in 1997, it was initially administered in 2000, and it currently includes over 80 nations. (Organization for Economic Cooperation and Development 2021b) The PISA study, conducted every three years, provides comparative statistics on 15-year-old students' performance in reading, math, and science. This report describes how to utilize the `learningtower` package, which offers OECD PISA datasets from 2000 to 2022 in an easy-to-use format. The datasets comprise information on students' test results and other socioeconomic factors, as well as information on their schools, infrastructure and the countries participating in the program.

Learningtower Package

'learningtower' The R package (Wang et al. 2021) provides quick access to a variety of variables in the OECD PISA data collected over a three-year period from 2000 to 2022. This dataset includes information on the PISA test scores in mathematics, reading, and science. Furthermore, these datasets include information on other socioeconomic aspects, as well as information on their school and its facilities, as well as the nations participating in the program.

The `learningtower` package primarily comprised of three datasets: `student`, `school`, and `countrycode`. The `student` dataset includes results from triennial testing of 15-year-old students throughout the world. This dataset also includes information about their parents' education, family wealth, gender, and presence of computers, internet, vehicles, books, rooms, desks, and other comparable factors. Due to the size limitation on CRAN packages, only a subset of the student data can be made available in the downloaded package. These subsets of the student data, known as the `student_subset_yyyy` (yyyy being the specific year of the study) allow uses to quickly load, visualise the trends in the full data. The full student dataset can be downloaded using the `load_student()` function included in this package. The `school` dataset includes school weight as well as other information such as school funding distribution, whether the school is private or public, enrollment of boys and girls, school size, and similar other characteristics of interest of different schools these 15-year-olds attend around the world. The `countrycode` dataset includes a mapping

of a country/region's ISO code to its full name.

1.2   Goals

The motivation for developing the `learningtower` package was sparked by the announcement of the PISA 2018 results, which caused a collective wringing of hands in the Australian press, with headlines such as "Vital Signs: Australia's slipping student scores will lead to greater income inequality" and "In China, Nicholas studied math 20 hours a week. In Australia, it's three". That's when several academics from Australia, New Zealand, and Indonesia decided to make things easier by providing easy access to PISA scores as part of the ROpenSci OzUnconf, which was held in Sydney from December 11 to 13, 2019.

The data from this survey, as well as all other surveys performed since the initial collection in 2000, is freely accessible to the public. However, downloading and curating data across multiple years of the PISA study could be a time consuming task. As a result, we have made a more convenient subset of the data freely available in a new R package called `learningtower`, along with sample code for analysis.

`learningtower` developers are committed to providing R users with data to analyse PISA results every three years. Our package's future enhancements include updating the package every time additional PISA scores are announced. Note that, in order to account for post COVID-19 problems, OECD member nations and associates decided to postpone the PISA 2021 evaluation to 2022 and the PISA 2024 assessment to 2025.

1.3   Compiling the Data(more details about the process and problems faced)

We are responsible for the curation of the newest PISA study, year 2022. Data on the participating students and schools were first downloaded from the PISA website, in either SPSS or SAS format. The data were read into an R environment. After some data cleaning and wrangling with the appropriate script, the variables of interest were re-categorised and saved as RDS files. One major challenge faced by the us was to ensure the consistency of variables over the years. However, several variables may be missing due to the reconstruction of questionnaires. For instance, a question regarding student's possession of desk is not recorded in 2022, but it was included in previous questionnaires, hence these variables were manually curated as an character variable in the output data. Another important issue we faced is a missing variable `WEALTH`, this variable used to be a good measurement of a student's socioeconomic status. But we also discovered a variable called `ESCS` (economic, social and cultural status). Also for 2022 PISA test, the student's curiosity was the first time included in the questionnaire. As "the finding show a clear link between student curiosity and student performance in mathematics.",

we decide to includ this attribute in the student dataset as well. These final RDS file for each PISA year were then thoroughly vetted and made available in a separate GitHub repository.

1.4 Communication and Documentation Tools

Slack and Notion can be effectively utilized together to enhance team communication and documentation management. Slack serves as a real-time communication tool, allowing teams to quickly exchange information, discuss projects, and stay updated on tasks, making it ideal for team collaboration.

Notion, on the other hand, excels as a centralized workspace for recording and organizing important documents, such as meeting journals, project notes, and other key materials, ensuring that important information is organized and easily accessible.

By using Slack for dynamic conversations and Notion for structured documentation, teams can ensure seamless communication while maintaining an organized record of all important documents, meeting notes, and long-term planning.

2 Overview of the data

3 Analysis

Gender

Pandemic effects

EcoSocio factors

4 Comparison with previous years' datasets

5 Reference

5.1 Git respository of the report

https://github.com/Shabarish161/Learningtower_Rpackage

References

Organization for Economic Cooperation and Development (2021a). *About OECD*. Accessed: 2021-11-03. https://www.oecd.org/about/.

Organization for Economic Cooperation and Development (2021b). *About PISA*. Accessed: 2021-11-03. https://www.oecd.org/pisa/.

Wang, K, P Yacobellis, E Siregar, S Romanes, K Fitter, G Valentino Dalla Riva, D Cook, N Tierney & P Dingorkar (2021). *learningtower: OECD PISA datasets from 2000-2018 in an easy-to-use format*. https://kevinwang09.github.io/learningtower/, https://github.com/kevinwang09/learningtower.