learningtower: an R package for Exploring Standardised Test Scores Across the Globe

by Priya Ravindra Dingorkar, Kevin Y.X. Wang, and Dianne Cook

Abstract An abstract of less than 150 words - Discuss what the paper talks about with a little introduction.

Introduction

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. 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. The PISA study, conducted every three years, provides comparative statistics on 15-year-olds' performance in reading, maths, and science. This paper describes how to utilize the `learningtower` package, which offers OECD PISA datasets from 2000 to 2018 in an easy-to-use format. This dataset comprises information on their test results and other socioeconomic factors, as well as information on their schools, infrastructure and the countries participating in the program.

What is PISA?

PISA assesses the extent to which children approaching the end of compulsory school have learned some of the information and abilities required for full participation in modern society, notably in maths, reading, and science. The examination focuses on reading, mathematics, science, and problem solving. It also assesses students capacity to replicate information and extrapolate from what they have learned and apply that knowledge in unexpected circumstances, both inside and outside of school. This approach reflects the fact that individuals are rewarded in modern economies not for what they know, but for what they can accomplish with what they know.

This evaluation which is carried out every three years, assists in identifying students' development of knowledge and skills throughout the world, which can provide actionable insights and therefore assist education policymakers. PISA is well known for its distinctive testing characteristics, which include policy orientation, an innovative notion of literacy, relevance to lifelong learning, regularity, and breadth of coverage. PISA is now used as an assessment tool in many regions around the world. In addition to OECD member countries, the survey has been or is being conducted in East, South and Southeast Asia, Central, Mediterranean and Eastern Europe, and Central Asia, The Middle East, Central and South America and Africa.

For each year of the PISA study, one domain subject is thoroughly examined. In 2018, for example, reading was assessed alongside mathematics and science as minor areas of assessment. The 2012 survey concentrates on mathematics, with reading, science, and problem solving serving as minor evaluation topics. PISA targets a certain age group of students in order to properly compare their performance worldwide. PISA students are aged between 15 years 3 months and 16 years 2 months at the time of the assessment, and have completed at least 6 years of formal schooling. They can enroll in any sort of institution, participate in full-time or part-time education, academic or vocational programs, and attend public, private, or international schools inside the country. Using this age across nations and throughout time allows PISA to compare the knowledge and abilities of people born in the same year who are still in school at the age of 15, irrespective of their diverse schooling.

The PISA test is primarily computer-based and lasts around 2 hours. The examination comprises both multiple choice and free entry questions. Some countries that were not ready for computer-based delivery carried out the testing on paper. Each student may have a unique set of questions. An example of the test may be seen here. PISA assessment areas seek to measure the following aspects of students' literacy in math, reading, and science. The goal of mathematical literacy is to assess students ability to grasp and interpret mathematics in a variety of settings. Reading literacy assesses students' capacity to absorb, apply, analyze, and reflect on texts in order to attain required goals and participate in society. Science literacy is described as the ability to engage with science-related issues and scientific concepts as a reflective citizen.

PISA data is publicly accessible for download. Furthermore, reading the data documentation reveals that the disclosed PISA scores are generated using a sophisticated linear model applied to the data. For each student, several values are simulated. This is known as synthetic data, and it is a popular technique to ensuring data privacy. The data can still be deemed accurate within the mean, variance, and stratum used in the original data's modelling. In addition, the PISA website provides the data in SPSS and SAS format, which can limit accessibility due to the commercial nature of these software. Furthermore, all questions are assigned with unique IDs within each year of the PISA study, but do not always agree across the different years. This data has now been curated and simplified into a single R package called `learningtower`, which contains all of the PISA scores from the years 2000 to 2018.

Data Compilation

Each developer at the ROpenSci OzUnconf was assigned to curate a specific year of the PISA study. 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 with the exception of the year 2000 and 2003. Due to formatting issues, the data for these two particular years were first read using SPSS and then exported into compatible `.sav` files. 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 developers was to ensure the consistency of variables over the years. For example, a student's mother's highest level of education was never recorded in 2000, but it was categorised as "ST11R01" between 2003 and 2012 and "ST005Q01TA" between 2015 and 2018. Such a problem was tackled manually by curating these values as an integer variable named "mother_educ" in the output data. These final RDS file for each PISA year were then thoroughly vetted and made available in a separate GitHub repository.

What is`learningtower`?

'learningtower' is an easy-to-use R package that provides quick access to a variety of variables using OECD PISA data collected over a three-year period from 2000 to 2018. 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 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 maths 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.

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.

`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.

Example Analysis

In this section we will illustrate how the `learningtower` package can be utilized to answer some research questions by applying various methodologies and statistical computations on the `learningtower` datasets.

We will use only the 2018 data here for illustrative purpose. We will begin first by constructing a `data.frame` that records the (weighted) average maths score for each country/region, grouped by gender. We will also calculate the difference between the two averages by gender. In order to show the variability in the mean estimate, we will perform sampling with replace on the data and calculate the same mean difference estimate. The empirical 90% confidence intervals are shown for each estimate for each country. We will repeat this process for reading and science test scores.

Gender Analysis

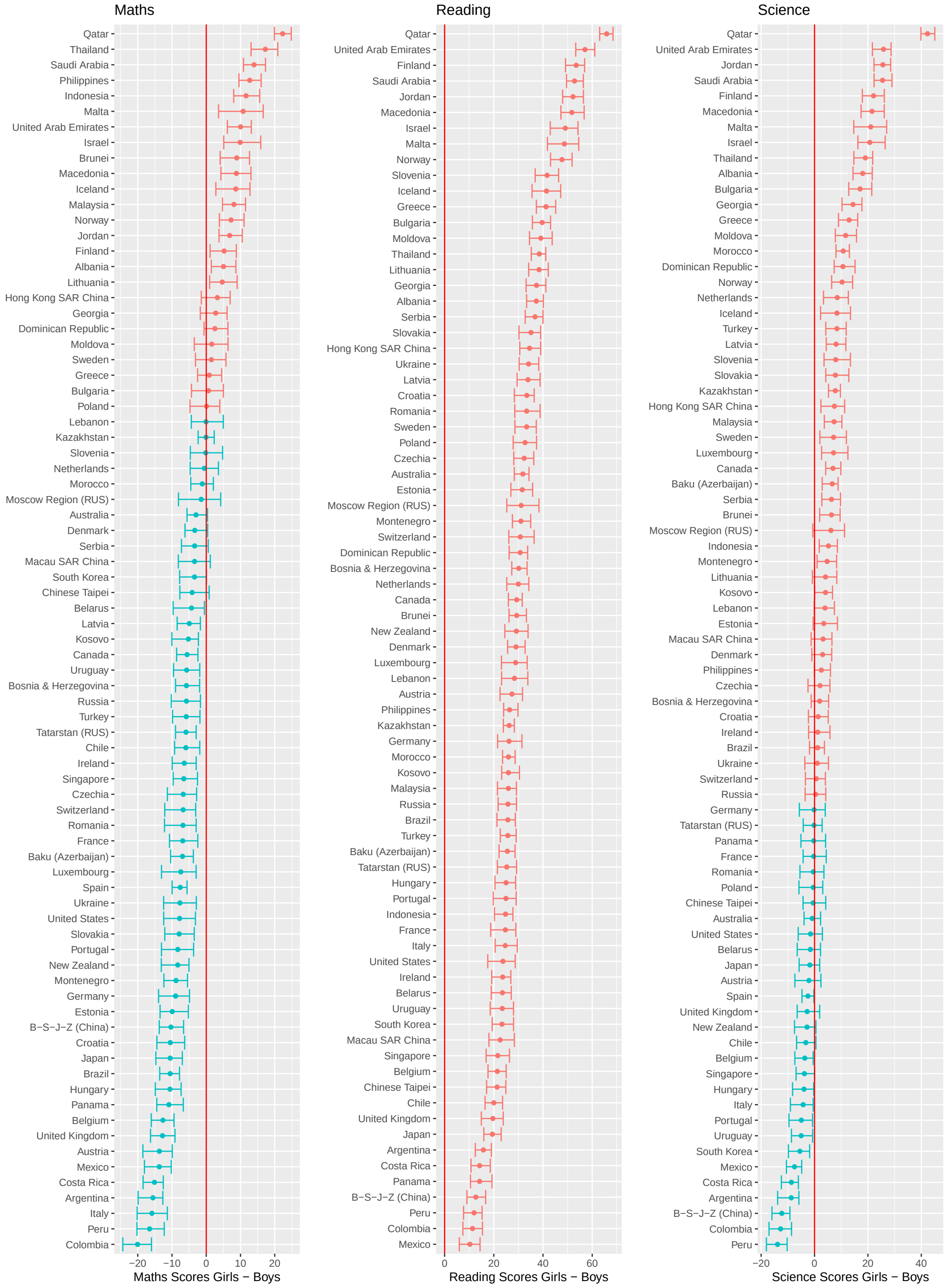

Figure 1: Gender Analysis

Figure 1 shows differences between mean scores for the three topics.

Maps

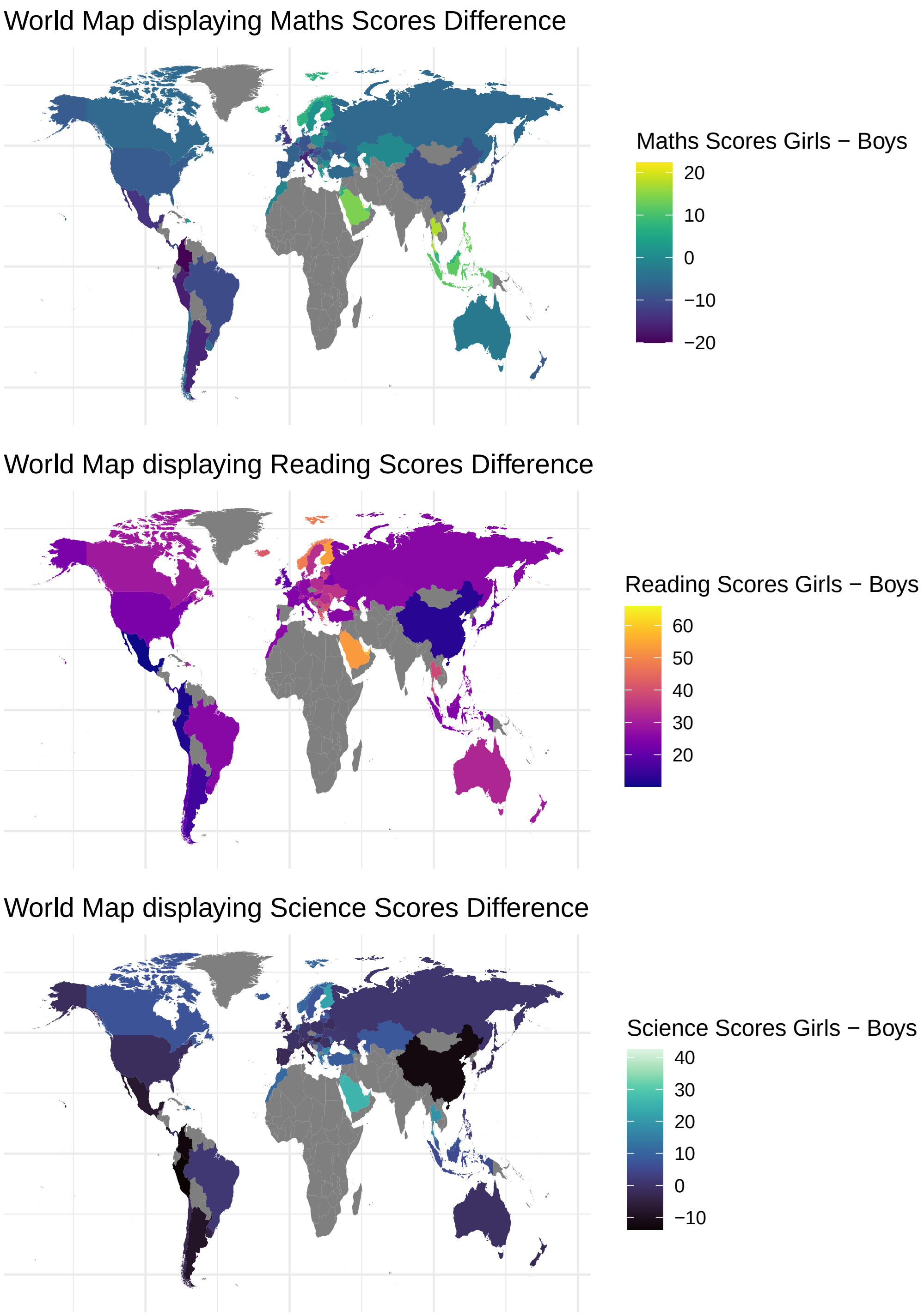
World Map displaying Maths Scores Difference
Maths Scores Girls − Boys
20
10
0
−10
−20
World Map displaying Reading Scores Difference
Reading Scores Girls − Boys
60
50
40
30
20
World Map displaying Science Scores Difference
Science Scores Girls − Boys
40
30
20
10
0
−10

Figure 2: Maps

Figure 2

Socioeconomic Factors Analysis

Corr plot

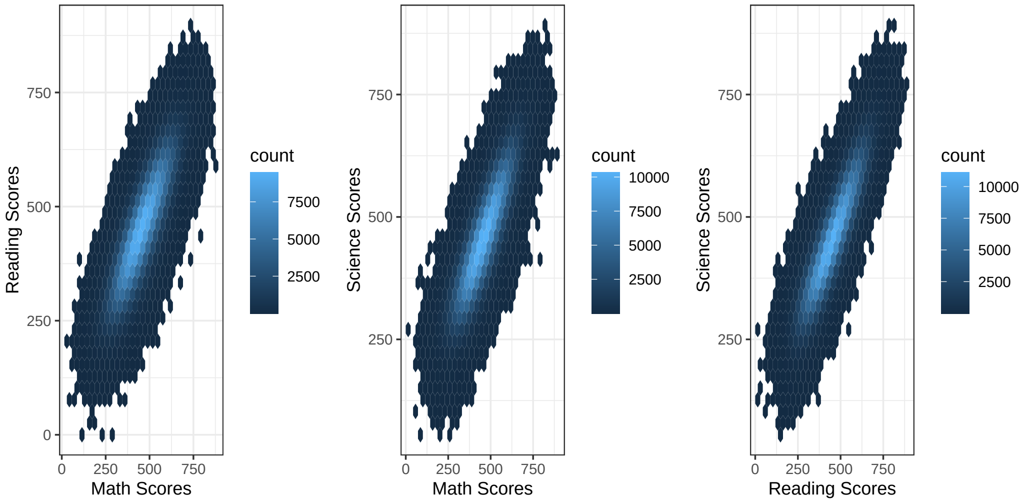

Figure 3: Correlation Plot

Figure 3

Qual Plot

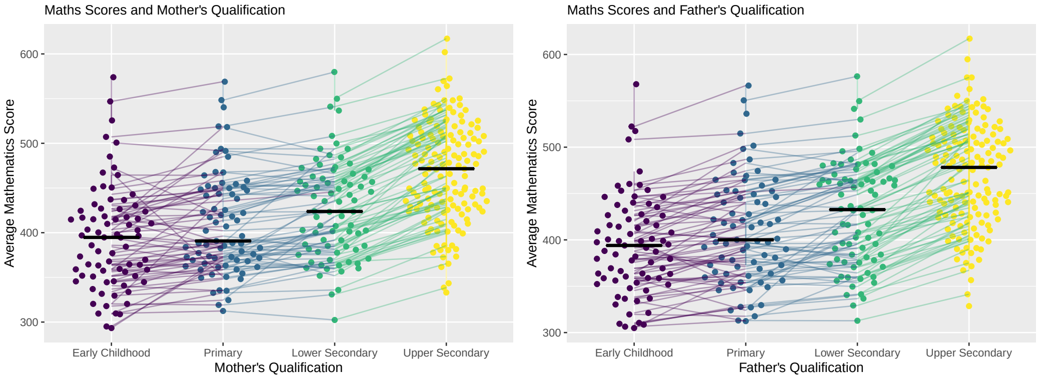

Figure 4: Qualification Plots

Figure 4

TV Plots

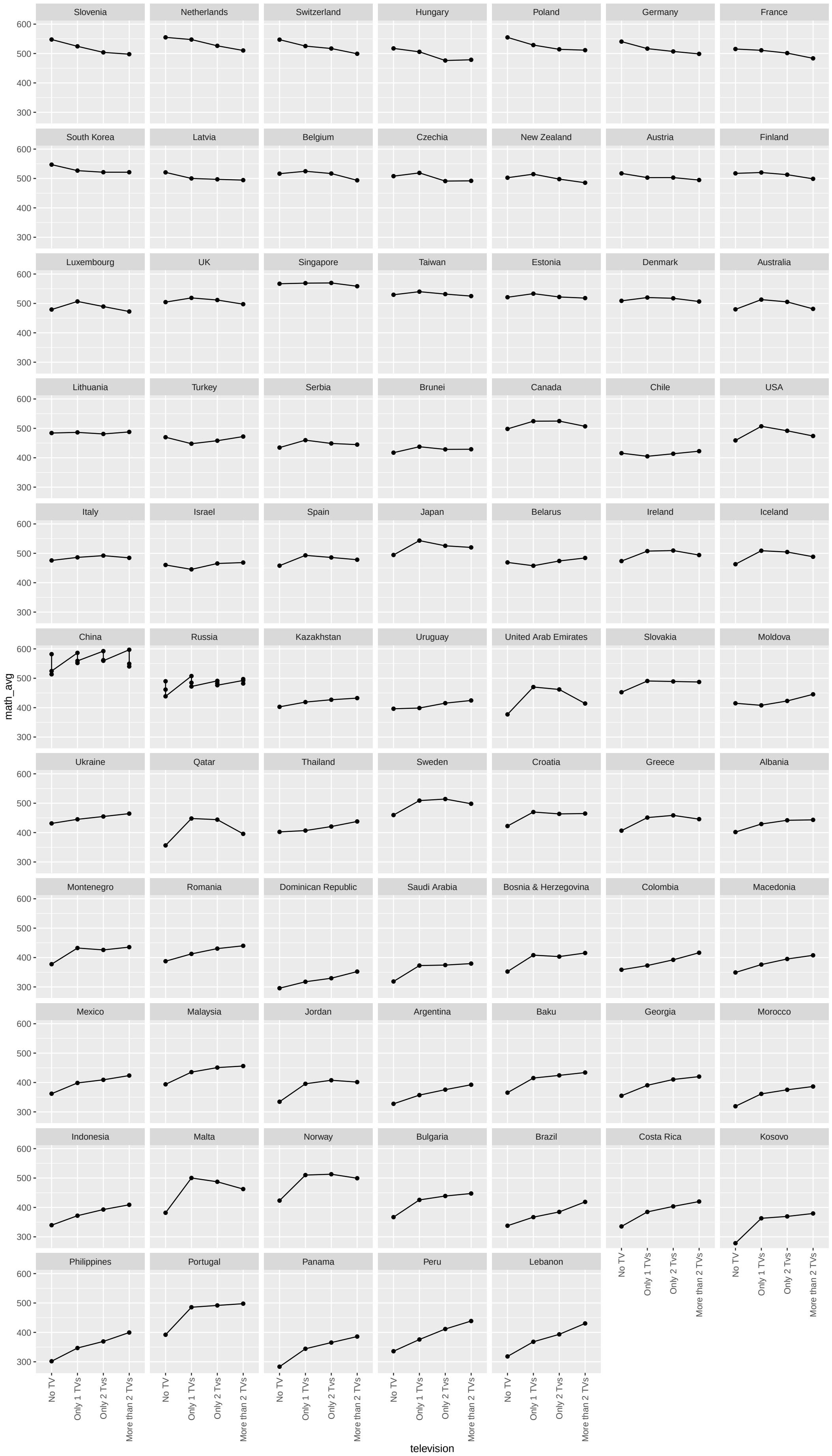

Figure 5: TV Plot

Figure 5

Book Plot

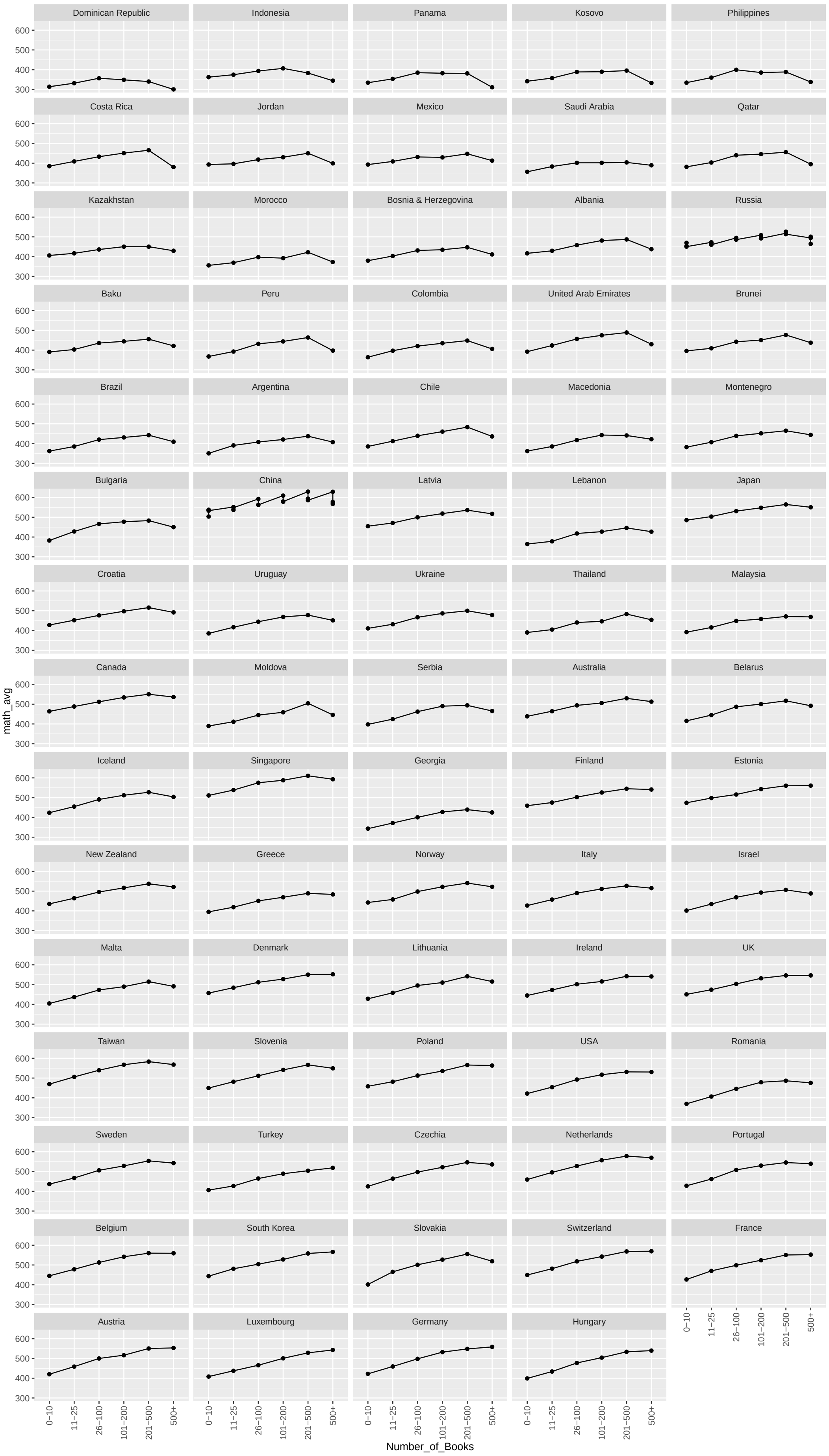

Figure 6: Book Plot

Figure 6

Internet and Computer Plot

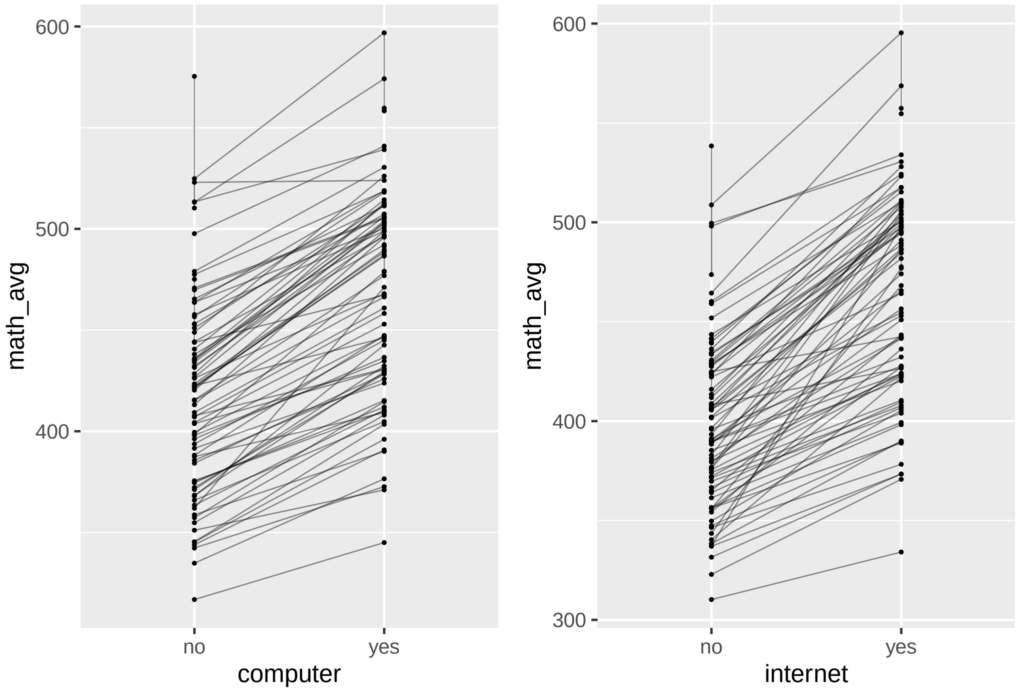

Figure 7: Qualification Plots

Figure 7

Temoral Trend Australia

Priya Ravindra Dingorkar
Monash University
Department Econometrics and Business Statistics
Clayton, Australia
https://www.linkedin.com/in/priya-dingorkar/
priyadingorkar@gmail.com

Kevin Y.X. Wang
University of Sydney
Data scientist
Illumina, Inc.
School of Mathematics and Statistics
Sydney, Australia
https://kevinwang09.github.io/
kevinwangstats@gmail.com

Dianne Cook
Monash University
Department Econometrics and Business Statistics
Clayton, Australia
http://dicook.org/
dicook@monash.edu