MONASH University

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

24 October 2024

MONASH
BUSINESS
SCHOOL

Department of Econometrics & Business Statistics

(03) 9905 2478
BusEco-Econometrics@monash.edu

ABN: 12 377 614 012

1 Abstract

2 Background

The learningtower package, which focusses on school and student-level data like performance indicators, socioeconomic backgrounds, and educational resources, contains educational datasets from international exams like PISA before 2022. After being cleaned and standardised, these datasets go through modifications including variable alignment for consistency and student-teacher ratio calculations. Key performance indicators include student performance in disciplines including reading, mathematics, and science; school and country identities; and educational resources (e.g., staff shortages, school size). The information makes it possible to analyse regional and worldwide trends as well as differences in educational systems, which sheds light on the variables influencing resource allocation and academic performance.

3 Introduction

A potent tool for analysing global educational data, the learningtower program focusses on Programme for International Student Assessment (PISA) statistics gathered over a number of years. These databases include precise school and student-level statistics from several nations, including student achievement in areas like reading, mathematics, and science, as well as contextual aspects like school resources, teacher-student ratios, and socioeconomic backgrounds. In order to make data from prior years (before 2022) appropriate for comprehensive cross-sectional and longitudinal investigations, the package makes sure that the data is cleaned, standardised, and converted to maintain consistency. With the help of this preparation, users can investigate worldwide trends in education, spot inequalities, and evaluate how various factors affect student achievements. We are now integrating the 2022 dataset, which has been effectively configured inside the package, guaranteeing that it is prepared for examination. With this update, the package can now offer the most accurate and up-to-date insights into global educational trends, allowing for greater cross-country comparison and analysis.

3.1 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.

3.2 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.

4 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.

5 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). These final RDS file for each PISA year were then thoroughly vetted and made available in a separate GitHub repository.

6 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.

7 Overview of the data

7.1 Student Dataset

The dataset offers student-level information from a number of nations and captures variables that affect academic performance. it contains the 23 variables, which could be categorized into groups:

Year*: represents the year of data collection, which is manually constructed by the contributor.

Country: specifies the country from which the student data has been collected, using country codes.

School's inforamtion: represents the unique identifier of each student's school.

Student's inforamtion: This group provides some information about each student in the dataset.

1. **Parent's education**: record the parent's highest level of education based on the International Standard Classification of Education (ISCED) levels, ranging from "less than ISCED1" to "ISCED 3B, C".
2. **Gender**: categorizes the gender of each student as "male" or "female".
3. **Household possession**: record several variables related to students' household resources. Including whether the student has access to a computer and internet at home, both marked as "yes" or "no." Additional household resources are indicated by variables for a desk, separate room, dishwasher, television, and car. The number of computers and laptops is also available. Finally, the number of books in the student's home is categorized into ranges, such as "0-10" or "101-200.
4. **Math, Read, Science**: These columns provide the scores in mathematics, reading, and science subjects, respectively.
5. **Stu_Wgt**: Represents the student weight, used for calculating weighted averages in the analysis to ensure representative data.

6. **Wealth**: This column provides a measure of the student's economic wealth, where higher values indicate greater wealth. However this variable is not recorded in 2022 dataset.

7. **ESCS**: Represents the Economic, Social, and Cultural Status index, which is a composite measure of a student's socio-economic background.

7.2 School Dataset

The 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.

```
# A tibble: 6 x 13
  year  country school_id fund_gov fund_fees fund_donation enrol_boys
  <fct> <fct>   <fct>        <dbl>     <dbl>         <dbl>      <dbl>
1 2000  ALB     01001          100         0             0       1191
2 2000  ALB     01004           98         1             1        334
3 2000  ALB     01005           91         5             2        403
4 2000  ALB     01010          100         0             0        114
5 2000  ALB     01013            0        50            30        250
6 2000  ALB     01017           95         2             3        771
# i 6 more variables: enrol_girls <dbl>, stratio <dbl>, public_private <fct>,
#   staff_shortage <dbl>, sch_wgt <dbl>, school_size <dbl>
```

Year*: represents the year of data collection, which is manually constructed by the contributor.

Country: specifies the country from which the student data has been collected, using country codes.

School's information: This group provides some information about each school in the dataset.

7.3 Countrycode Dataset

This dataset includes a mapping of a country/region's ISO code to its full name. More information on the participating countries can be found here.

8 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 solely utilize the 2022 PISA data and scores for illustrative purposes throughout the analysis section. Some of these questions include if there is any significant gender difference between girls and boys and explore their performance in the areas of mathematics, reading, and science. Furthermore, we will inspect the various socioeconomic characteristics reflected in the student data and investigate if they have any substantial impact on the scores of these students.

8.1 Gender Gap

Gender gaps have always been a topic of interest among researchers, and when it comes to PISA data and scores of 15-year-old students around the world, uncovering patterns based on their gender would help gain meaningful insights in the field of education for various education policymakers around the world. Based on the 2022 PISA results, let us see if there is a major gender disparity between girls and boys throughout the world in mathematics, reading, and science. To begin, we will create a 'data.frame' that stores the weighted average math score for each nation as well as the various regions of the countries grouped by country and gender, in order to create this `data.frame` and represent data in the tidy format we use the `tidyverse` (Wickham et al. 2019) and `dplyr` (Wickham et al. 2021) R packages. Survey weights are critical and must be used in the analysis to guarantee that each sampled student accurately represents the total number of pupils in the PISA population. In addition, we compute the gender difference between the two averages. To demonstrate the variability in the mean estimate, we use bootstrap sampling with replacement using the `map_dfr` function on the data and compute the same mean difference estimate. For each country, the empirical 90 percent confidence intervals are presented. The same process is used for reading and science test scores.

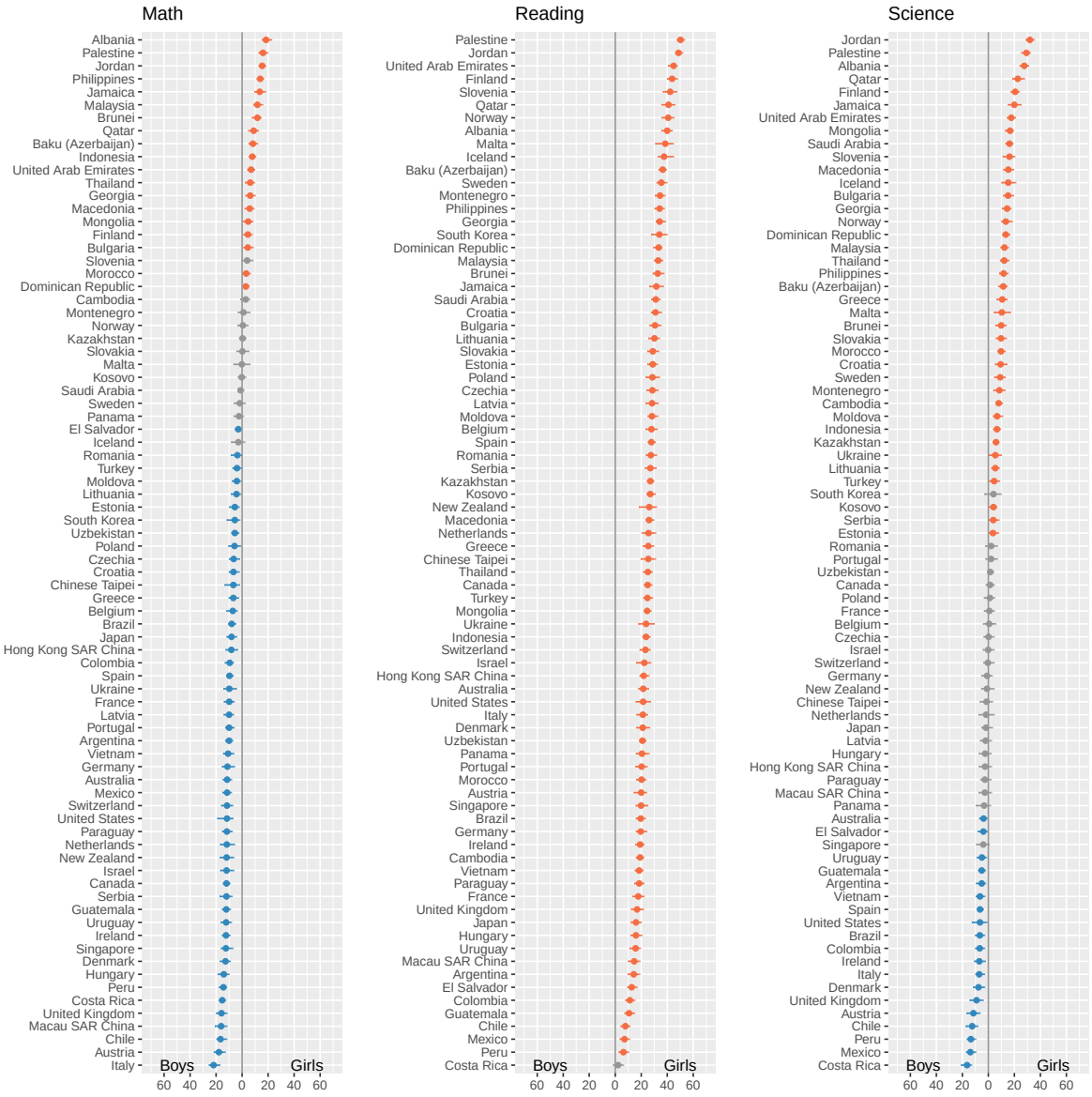


Figure 1: *The chart above depicts the gender gap difference in 15-year-olds' in math, reading, and science results in 2022. The scores to the right of the grey line represent the performances of the girls, while the scores to the left of the grey line represent the performances of the boys. One of the most intriguing conclusions we can get from this chart is that in the PISA experiment in 2022, girls from all countries outperformed boys in reading.The chart above depicts the gender gap difference in 15-year-olds' in math, reading, and science results in 2022. The scores to the right of the grey line represent the performances of the girls, while the scores to the left of the grey line represent the performances of the boys. One of the most intriguing conclusions we can get from this chart is that in the PISA experiment in 2022, girls from all countries outperformed boys in reading.*

Figure 1 illustrates the global disparities in mean math, reading, and science outcomes, before we get to the plot conclusion, let's have a look at the variables that have been plotted. The grey line here indicates a reference point, and all of the scores to the right of the grey line show the scores of girls in math, reading, and science. Similarly, the scores on the left side of this grey line indicate the scores

of boys in the three disciplines. Based on Figure 1, because most math estimates and confidence intervals lie to the left of the grey line, we may conclude that most boys outperformed girls in math.

In nations such as Panama, Malta, Saudi Arabia, Sweden, Kazakhstan, Norway, Slovenia, Iceland, Kosovo, Cambodia, Montenegro and Slovakia, there is almost no gender difference in average math scores. When we look at the reading scores, we notice a remarkable trend in that all girls outpaced boys in reading in all countries in 2022. The highest reading scores were achieved by girls from Palestine, Jordan and United Arab Emirates. Looking further into the science plot, we see an unexpected pattern here where most countries have very little gender difference in science scores, implying that most boys and girls perform equally well in science. Boys from Costa Rica, Mexico and and Peru perform well in science and girls from Jordan, Palestine, and Albania are the top scores for science. Figure 1 helps us to depict the gender gap in math, reading, and science for all nations and regions that took part in the 2022 PISA experiment.

We gathered meaningful insights about the gender gap between girls and boys across the world from the above Figure 1 because this is a geographical research communication topic, the findings will help us better comprehend the score differences in the three educational disciplines using world maps. Let us continue to investigate and discover patterns and correlations using map visualization. To illustrate the gender gap difference between girls and boys throughout the world, we summarize regions on a country level and utilize the `map_data` function to get the latitude and longitude coordinates needed to construct a map for our data. We connect these latitude and longitude coordinates to our PISA data and render the world map using the `geom_polygon` function wrapped within `ggplot2` (Wickham 2016), the interactive features and placement of the plots are made using `plotly` (Sievert 2020) and `patchwork` (Pedersen 2020) packages in R.

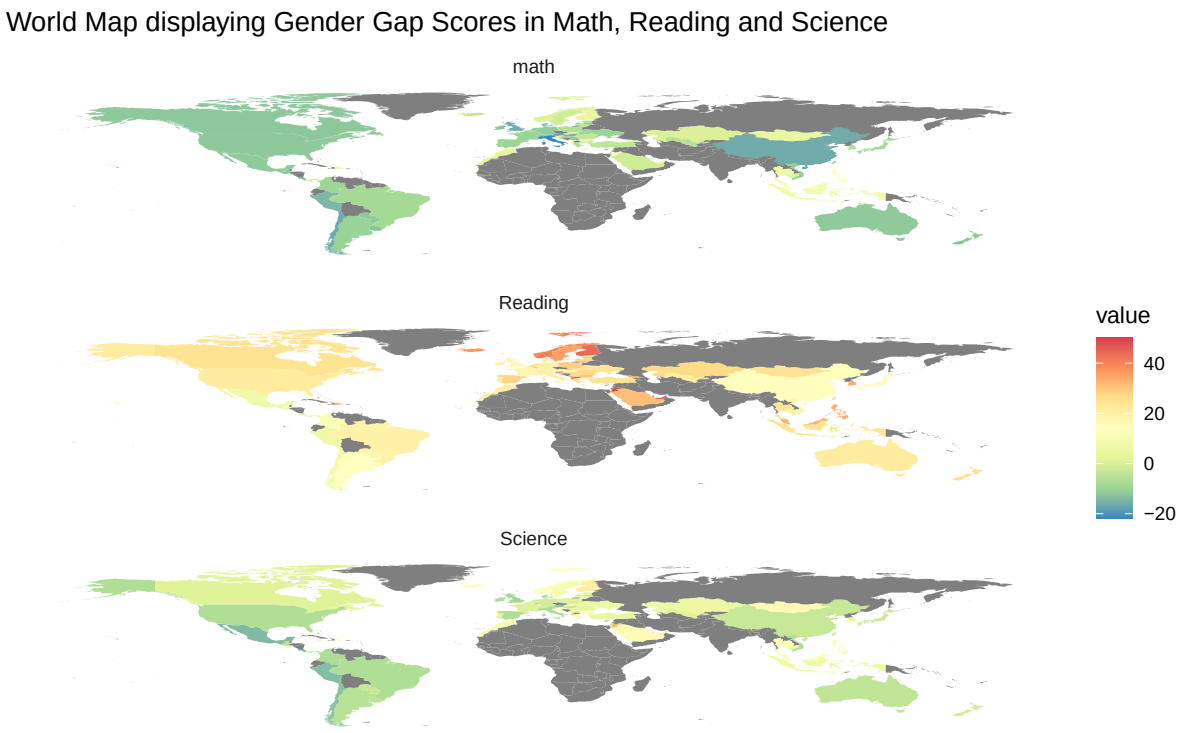


Figure 2: *Maps showing the gender gap in math, reading, and science results between girls and boys across the world. A positive score for a country indicates that girls outperformed boys in that country, whereas a negative score for a country difference indicates that boys outperformed girls in that country. The diverging colour scale makes it possible to interpret the range of scores and the also helps us intrepret the gender gap difference among these students across the globe.*

In the Figure 2, we have shown the gender gap difference between girls and boys in math, reading, and science in 2022. Map visualization aids in the comprehension of large volumes of data in a more efficient manner and increases the ability to compare outcomes across many geographical locations at a glance. In this figure, we see both positive and negative score difference scale ranges in all three maps. A positive country score indicates that girls outperformed boys in that country, whereas a negative country score shows that boys outscored girls in that country. The diverging spectral color scale and the legend of these maps makes it possible for us to deduce and identify regions across the globe showing large gender discrepancy between girls and boys. The grey colour for different geographic locations across the maps in Figure 2 indicates that these regions were not a part of the PISA experiment in year 2022.

Even though the map visualization embeds the same scores as Figure 2, one of the most striking thing on this map is the lack of data for the Africa continent. We see that there is less of a gender disparity seen in the science scores compared to maths and reading. In addition, the color scale for scores of each subject aids in identifying the countries that took part in the PISA experiment. As a result, in this section, we have seen the gender gap scores and striking trends between 15-year-old girls and

boys in math, reading, and science. Our main conclusion from this gender study is the performance of girls in reading. The fewer gender disparity is evident in the science scores, and the majority of boys perform better than girls in mathematics.

8.2 EcoSocio factors

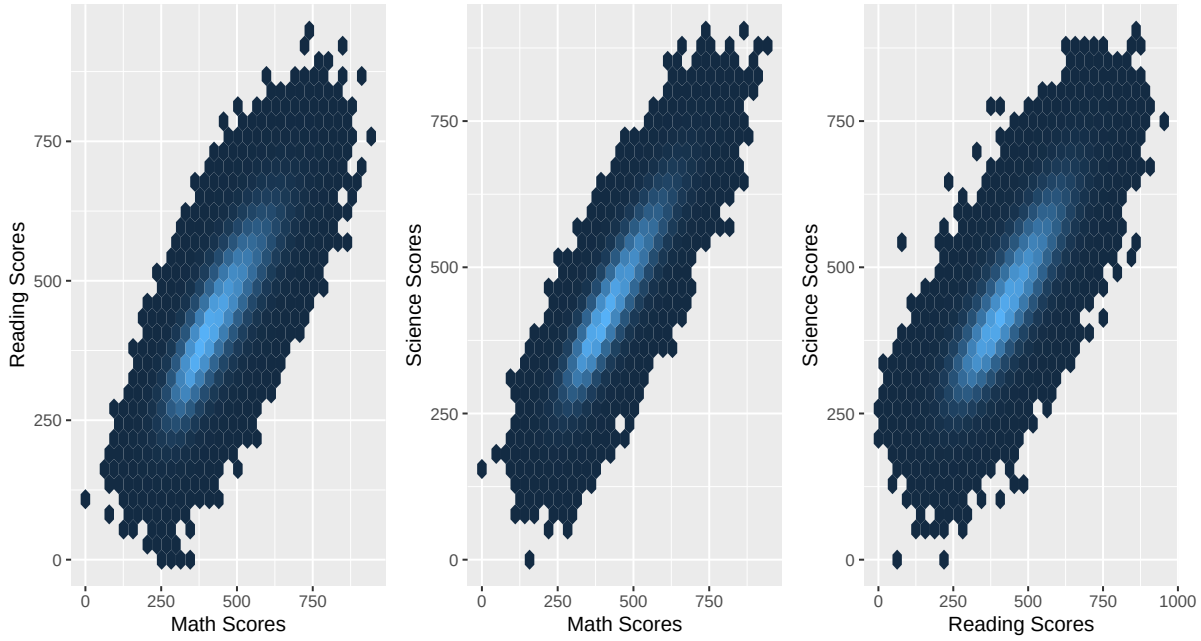


Figure 3: *The scatterplot displays the relationship between math, reading, and science scores for all PISA countries that participated in the experiment in 2022. This scatterplot shows that all three subjects have a significant and positive correlation with one another.*

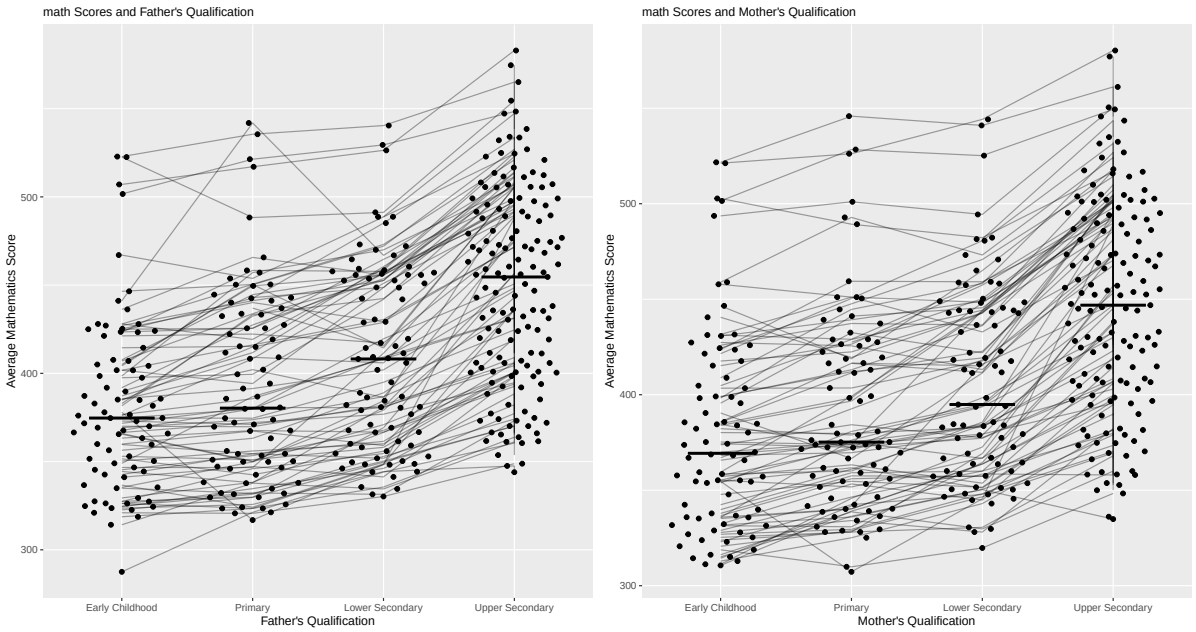


Figure 4: *The impact of parents' education on their children's academic progress is depicted in this graph. When the parents have greater levels of education, we see a considerable rise in scores and an increase in the median of scores for each category, as shown in the figure. In comparison to parents with lower levels of education qualifications. Parents who have tend to have upper secondary qualification or equivalent credentials their children are more likely to perform better in academics when compared with parent having lesser levels of qualifications.*

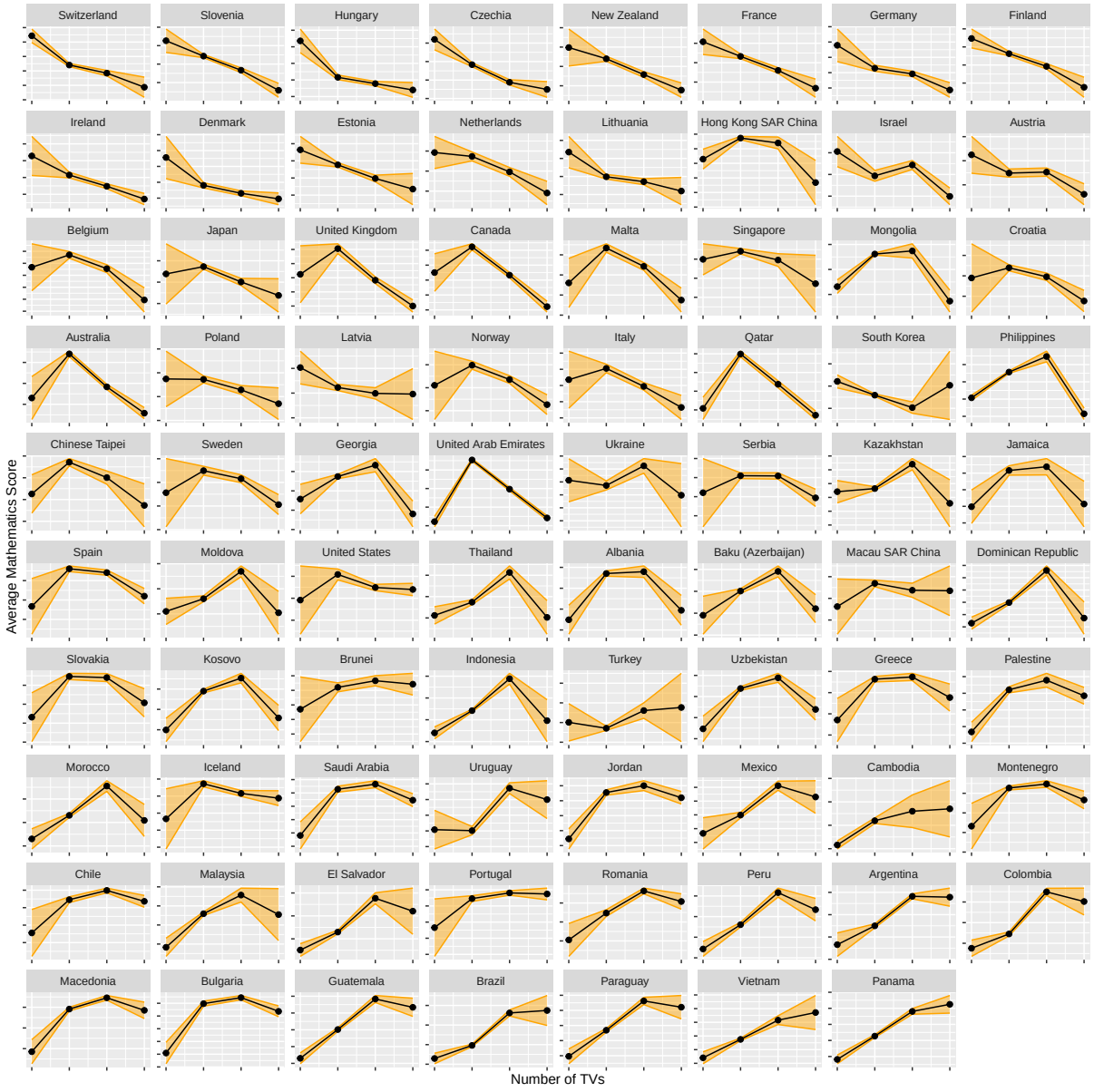


Figure 5: *Relationship between number of TVs in a household and average math scores across countries. Number of TVs ranges from 0 to 3 or more. The orange bands indicate 95 percent standard confidence intervals. The impact of television on student performance is a contentious issue. It is interesting that in some countries for example in South Korea's effect appears to be positive, but in other countries like Poland and Germany there is a decline in average math scores.*

Figure 6: *Impact of the number of books on average math score. Number of books ranges from 0 to 500 and more. 95 percent standard confidence bands shown in orange. Math scores generally increase as the number of books increases. Averages for some countries at the higher number of books are less reliable, and hence the decline reflects more that there are few households with this many books than a true decline.*

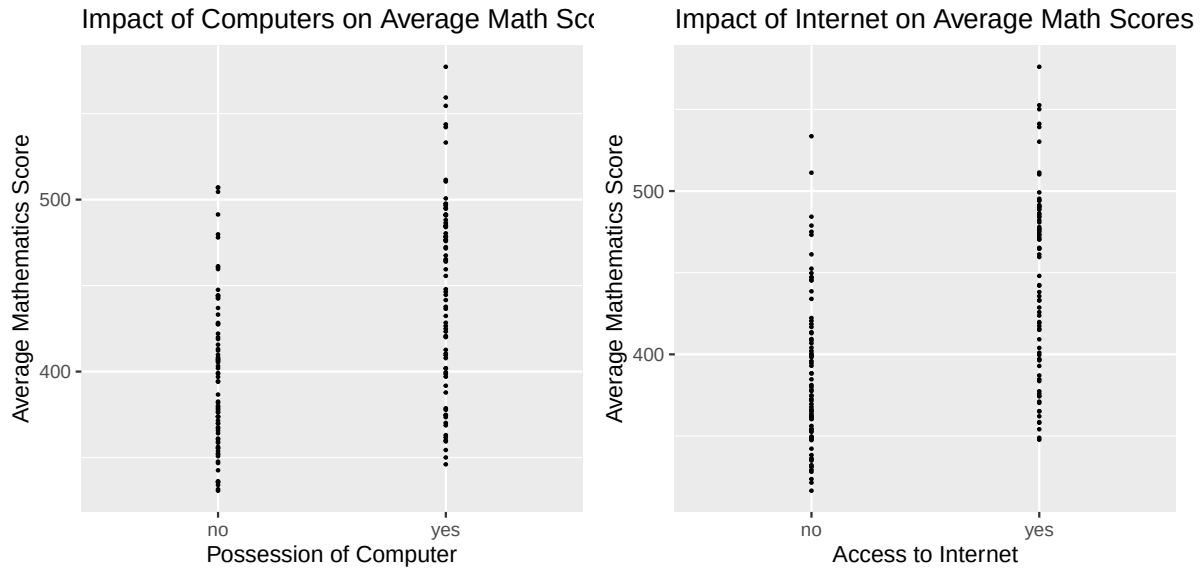


Figure 7: *Computers and the Internet are two of the most important inventions in the history of technology. In this figure, we observe the impact of owning a computer and having access to the internet on 15-year-old students all over the world. A remarkable finding from the plot is that all nations have higher scores in student performance when they own a computer and have access to the internet.*

8.3 Temporal Analysis

Pandemic effects

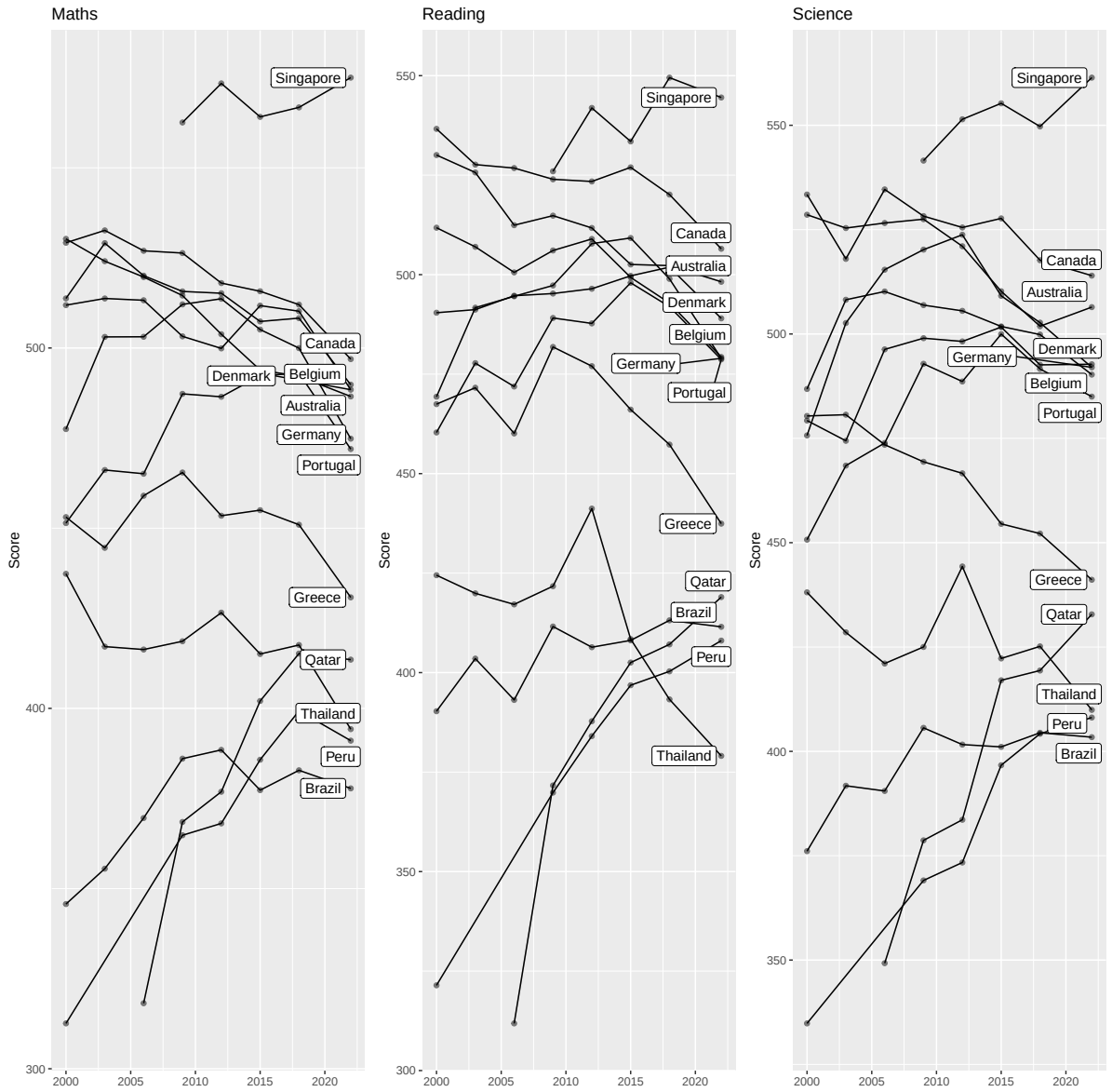
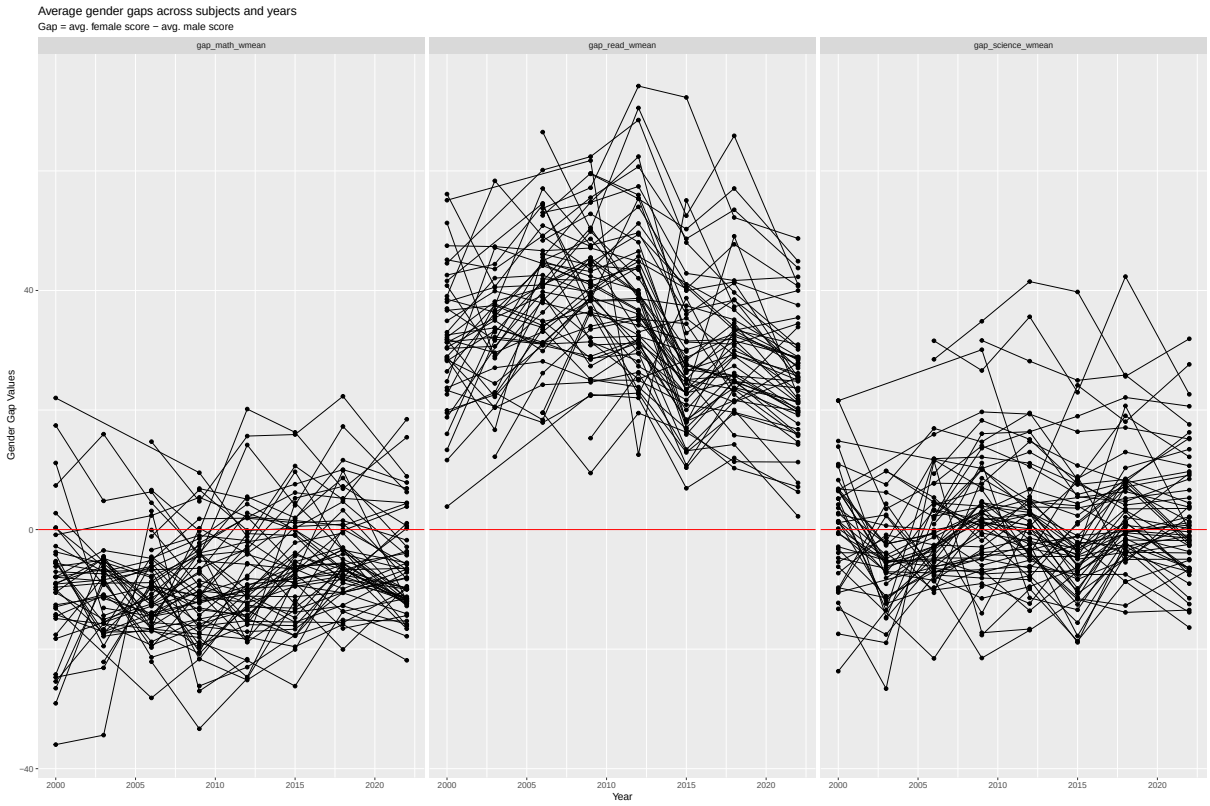
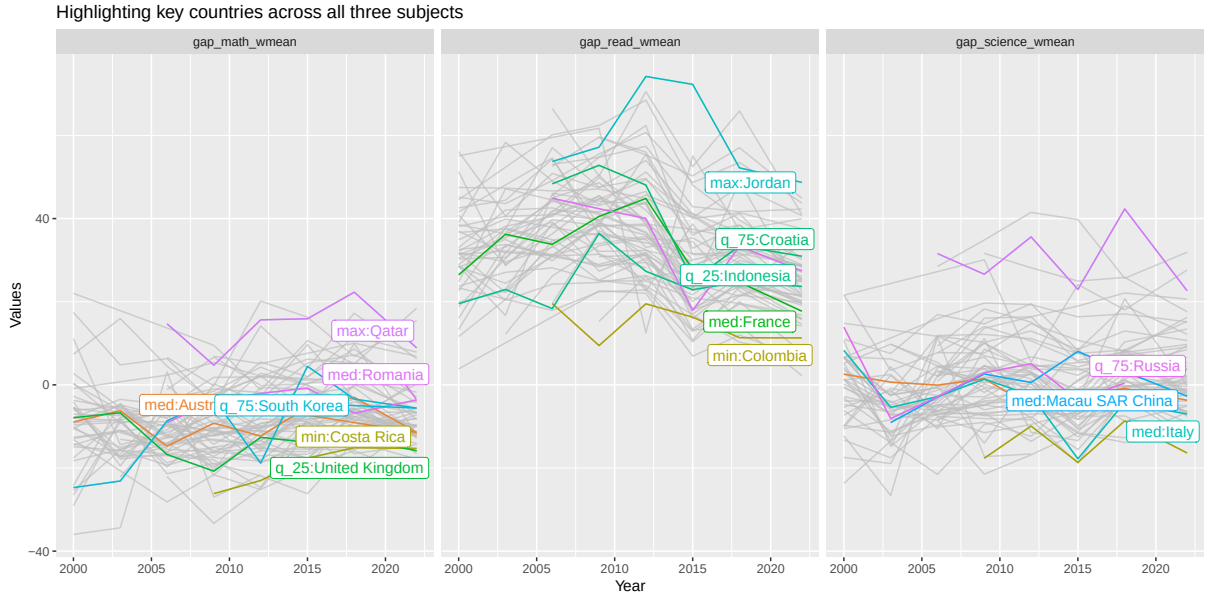


Figure 8: *Temporal patterns in math, reading, and science in a variety of countries. The highlighted countries in the chart help us infer Australia's performance in contrast to the other countries; we can see that Australia's scores have always been among the highest in the PISA survey throughout all years.*

Gender Gaps Across Subjects and Years



Highlighting Key Countries



9 Discussion

9.1 Limitations

- Size limitation on CRAN packages: The data size would be bigger if keep uploading the newest data, so further curation process of data should be considered, or explore alternative data

compression for the datasets.

- Variables Consistency: The construction of questionnaire would be different every survey, as well as the coding mechanism of the original dataset, so curation process must be examined everytime to ensure the consistency of variables.

10 Conclusion

11 Reference

11.1 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/.

Pedersen, TL (2020). *patchwork: The Composer of Plots*. R package version 1.1.1. https://CRAN.R-project.org/package=patchwork.

Sievert, C (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. https://plotly-r.com.

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.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, H, M Averick, J Bryan, W Chang, LD McGowan, R François, G Grolemund, A Hayes, L Henry, J Hester, M Kuhn, TL Pedersen, E Miller, SM Bache, K Müller, J Ooms, D Robinson, DP Seidel, V Spinu, K Takahashi, D Vaughan, C Wilke, K Woo & H Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**(43), 1686.

Wickham, H, R François, L Henry & K Müller (2021). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr.