MONASH University

Report of developing Learningtower R package

Shabarish Sai Subramanian
Master of Business Analytics

Guan Ru, Chen
Master of Business Analytics

Report for
ETC5543 Business analytics creative activity, 2024

29 October 2024

MONASH BUSINESS SCHOOL

Department of Econometrics & Business Statistics

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

ABN: 12 377 614 012

1 Abstract

A powerful tool for exploring and analysing data from the 2000–2022 Programme for International Student Assessment (PISA), which is led by the Organisation for Economic Co-operation and Development (OECD), is the learningtower R package. Reading, math, and science ability among 15-year-old pupils is measured as part of the triennial PISA international research, which assesses educational systems around the world. This thorough evaluation shows the efficacy of various educational policies and methods across countries in addition to offering insights on student accomplishment.

With a suite of tools that improve data curation, visualisation, and statistical modelling, the learningtower package focusses on important areas of educational research, including gender inequities, socioeconomic impacts, and temporal trends in student performance. Learningtower makes it easier for academics, decision-makers, and educators to access and manipulate PISA data, allowing them to perform comprehensive analyses that support evidence-based decision-making.

The learningtower package's intuitive design and features enable users to find important information about the performance of education around the world, emphasising injustices and differences between nations. Learningtower plays a critical role in promoting educated conversations on educational results and equity by making PISA data easier to access and handle. This, in turn, helps to improve educational practices and policies globally.

2 Background

With its emphasis on school and student-level data, the learningtower R package is an extensive resource that offers crucial insights into a range of educational measures, such as socioeconomic backgrounds, performance indicators, and the accessibility of educational resources. A multitude of datasets from worldwide examinations, including the Programme for worldwide Student Assessment (PISA) before 2022, which assess students' proficiency in a variety of disciplines globally, are included in this bundle.

To guarantee correctness and dependability, learningtower's datasets go through a rigorous cleaning and standardisation procedure. This entails a number of adjustments, including computing important metrics like the student-teacher ratio and aligning variables for uniformity. Making these changes is essential to improving the dataset's usability and preparing it for in-depth study.

Student accomplishments in foundational subjects like science, math, and reading are among the main performance indicators that the learningtower package records. Furthermore, the databases include

crucial data on national and school identities as well as the accessibility of educational resources, which may include elements like staffing levels and school size. These indicators offer a framework for assessing how many factors affect academic performance and resource allocation, and they are crucial for comprehending the educational landscape.

The learningtower package helps academics and policymakers identify inequalities in educational systems by making it easier to analyse regional and worldwide trends in educational results. These analyses provide insight into the different elements affecting the allocation of resources, the effectiveness of instruction, and eventually, student achievement. Learningtower's all-encompassing strategy seeks to advance evidence-based tactics for raising educational quality and equity while deepening our understanding of educational dynamics.

3 Introduction

With a focus on the Programme for International Student Assessment (PISA) data from 2000 to 2022, the learningtower package is a strong and adaptable tool for evaluating educational data globally. The Organisation for Economic Co-operation and Development (OECD) administers PISA every three years to assess 15-year-old pupils' competency in important academic subjects like science, arithmetic, and reading. This test is intended to gauge how well students can apply their knowledge and abilities to practical issues, offering important information about how ready they are for adulthood and further education. Countries can track changes in their educational systems over time and evaluate the efficacy of their policies and practices by assessing students' proficiency in these fundamental disciplines.

Although PISA data is accessible to the general public via the OECD website, researchers and analysts may encounter considerable difficulties due to the datasets' intricate structure and variety of forms. Effective analysis may be hampered by these intricacies, making it challenging to glean significant insights from the abundance of available data. By combining extensive datasets covering teacher-student ratios, school resources, student performance measures, and important socioeconomic aspects, the learningtower package tackles these issues head-on. Learningtower enables comprehensive cross-sectional and longitudinal research through a methodical process of data cleansing, standardisation, and preparation, making it easy for users to traverse the complexities of educational data.

Moreover, the recent addition of the 2022 dataset enhances the package's relevance and utility, providing the most current information on global educational trends. This update allows for thorough cross-national comparisons and assessments, enabling researchers and policymakers to identify patterns and correlations that impact student outcomes. By exploring the relationships between

educational resources, teaching methodologies, and student performance, the learningtower package empowers users to draw informed conclusions and develop evidence-based strategies aimed at improving educational equity and effectiveness across diverse contexts.

In conclusion, the learningtower package is an invaluable tool for educators and policymakers who want to comprehend and address the global variables driving student accomplishment, in addition to being a vital resource for education researchers.

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, since the year 2000, it has involved more than 100 countries and economies and has conducted tests of more than 3.7 million students worldwide.(https://www.oecd.org/pisa/aboutpisa/pisa-participants.html).(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.

Specific Goals related to the Learningtower package

1. Simplified Access to Complete PISA Data: Make a user-friendly R package available that makes it easier to access a well selected, reliable, and complete subset of PISA data from 2000 to the most recent year. By doing this, users will need to spend less time and technical effort downloading, cleaning, and standardising the data.

2. Improved Data for Comparative Analysis: Make it possible to conduct longitudinal and cross-sectional studies of educational performance across nations, giving researchers the ability to monitor trends over time and evaluate how different socioeconomic factors affect student outcomes.

3. Emphasis on Educational Equity: Encourage the investigation of performance gaps, paying special attention to variables including gender, socioeconomic position, and school resources. This goal is to enable policymakers and educational academics to recognise and resolve disparities both within and across nations.

4. Promoting Policy-Relevant Insights: Make it possible for academics and policymakers to obtain practical insights that guide instructional techniques, particularly with regard to achievement

gaps. The package aids in the development of policy interventions aimed at underperforming populations by examining the variables that affect student success.

5. Simple Usability for the R Community: Provide thorough documentation, effective data structures in R, and sample analytic code so that people with different levels of experience can interact with PISA data for educational research.

6. Continuous upgrades and Data Consistency: Make sure that learningtower stays pertinent for the demands of contemporary research by committing to frequent package upgrades that incorporate fresh PISA data as it becomes available (next in 2025). Furthermore, in order to accommodate modifications in OECD schedules, the program guarantees consistency of variables across several assessment years in order to handle post-COVID-19 education dynamics.

7. Encouraging Reproducibility and Transparency: Make the carefully selected PISA data and analytic techniques publicly accessible via GitHub to encourage transparency and allow other researchers to duplicate and expand on the results.

By achieving these objectives, the learningtower package hopes to become a vital tool for educational research, advancing evidence-based perspectives on global trends in education and aiding initiatives to promote fair educational opportunities across the globe.

5 Methodology

The learningtower R program was developed through a series of methodical procedures, including data collection, processing, variable consistency checks, and analysis, for the study of PISA data from 2000 to 2022. Each phase was meant to ensure that the data remains robust, consistent, and available for educational research and comparative studies.

1. Data Acquisition and Import

a. Data Download: The most recent PISA datasets, which included student and school data from several nations, were downloaded in SAS or SPSS formats from the OECD website

b. Data Loading in R: Following the download, the data was loaded into the R environment using scripts that made managing big SPSS/SAS files easier and guaranteed R compatibility.

2. Data Cleaning and Transformation

a. Initial Cleaning: To eliminate discrepancies, including missing or incorrect entries, and to reformat categorical variables for standardisation, data wrangling programs were used. The data was prepared for additional processing thanks to this first cleaning.

b. Variable Recategorisation: To ensure uniformity between PISA years, variables of interest were reclassified where needed.

c. Taking care of Missing Variables: The 2022 dataset lacked some variables, including "possession of desk." These variables were either explicitly selected as character variables or, when appropriate, substituted with comparable indications in order to address this.

3. Ensuring Variable Consistency

a. Cross-Year Alignment: Maintaining the consistency of important variables between PISA assessment years was a significant methodological difficulty. Some variables, such as the WEALTH index, were absent from newer datasets due to changes in questionnaire design. To retain a measure of socioeconomic level in these situations, other variables such as ESCS (Economic, Social, and Cultural level) were employed.

b. Categorisation Adjustments: To guarantee consistency in analysis, the classification for parental education levels, household belongings, and technological access was standardised across datasets.

4. Data Transformation and Storage

a. Data Transformation: The processed data was transformed and saved as.rds files after being cleaned and having its consistency checked. The cleaned and categorised datasets are stored in these RDS files, which are small and easy to load quickly in R.

b. File Organisation by Year: To assist cross-sectional research within each year and to enable longitudinal analysis across years, each PISA dataset was saved independently. This method guarantees that the data by year may be readily accessed by researchers without the need for extra processing.

5. Validation and Quality Checks Each dataset was examined for accuracy and completeness following processing. Particularly for new or altered variables, the integrity of socioeconomic indicators and the coherence of variable definitions were verified twice.

6. Data Analysis Techniques

a. Bootstrap Sampling: For confidence intervals in analyses like gender gaps, socioeconomic effect, and longitudinal trends, bootstrap sampling was used to produce reliable results.

b. The dataset is appropriate for both cross-sectional and longitudinal educational research since methods were used to investigate trends in gender disparities, socioeconomic factors, and the influence of variables across time.

Learningtower's high-quality, consistent, and easily accessible PISA data is guaranteed by this methodical approach, which makes thorough analysis and policy-oriented educational research possible.

6 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.

7 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.

8 Overview of the data

8.1 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.

8.2 School Dataset

The school dataset covers a variety of educational topics and offers data about schools across multiple nations.

year: Year represents the year during which the data was collected.

country: Country refers to the countries from which the data was collected.

School's information:

1. **school_id**: A unique identifier for each school, allowing for consistent tracking of school-related data across different records and years.
2. **public_private**: Designates the type of school administration, distinguishing between public (government-funded) and private (independently funded) institutions.
3. **stratio**: Represents the student-teacher ratio, indicating the average number of students per teacher in the school. Lower ratios typically suggest smaller class sizes and potentially more individualized attention.
4. **staff_shortage**: A measure indicating the level of staffing challenges faced by the school. Positive values may suggest severe shortages, while negative values or zeros may indicate minimal issues.
5. **school_size**: The total number of students enrolled, which can reflect the school's capacity and influence on the learning environment. Larger schools may offer more diverse programs, while smaller ones might provide a more personalized setting.
6. **fund_gov**: the percentage of funding that the government provides to the school. Understanding the resources available for staffing, programs, and infrastructure can be greatly aided by this.
7. **fund_fees**: The proportion of money that came from student fees demonstrates the school's reliance on fee-based revenue, which is more usual in private institutions.
8. **fund_donation**: represents the proportion of funds raised through donations, which could be a sign of extra support systems or community involvement.
9. **sch_wgt**: In order to guarantee that the sample appropriately reflects the school population, school weight is utilised in statistical analysis, especially in weighted averages or survey data.

8.3 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.

9 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 utilise 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.

9.1 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) R package. 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.

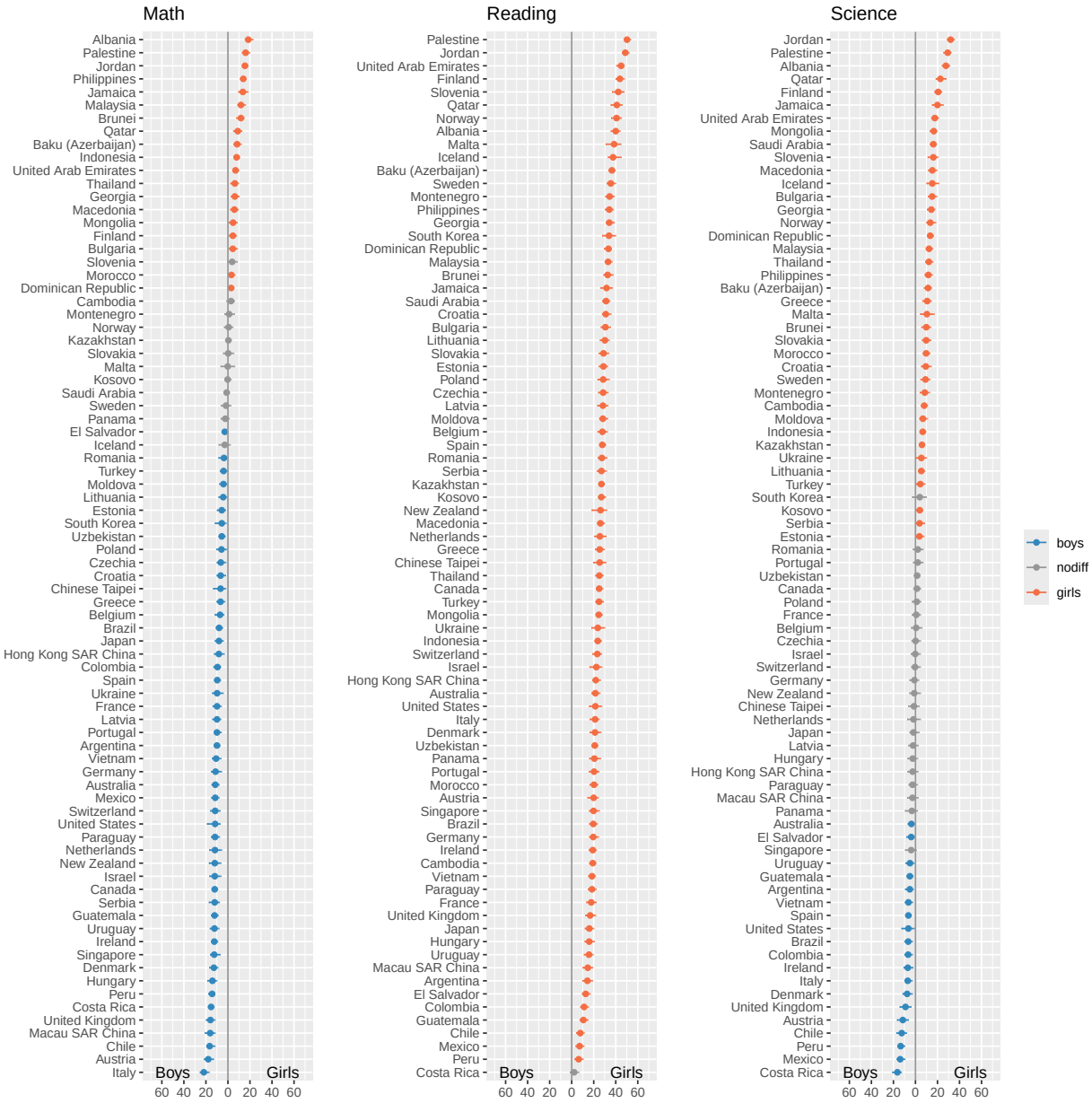


Figure 1: *The chart above depicts the gender gap difference in 15-year-olds' in math, reading, and science results in 2022. The scores to the right of the grey line represent the performances of the girls, while the scores to the left of the grey line represent the performances of the boys. One of the 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.

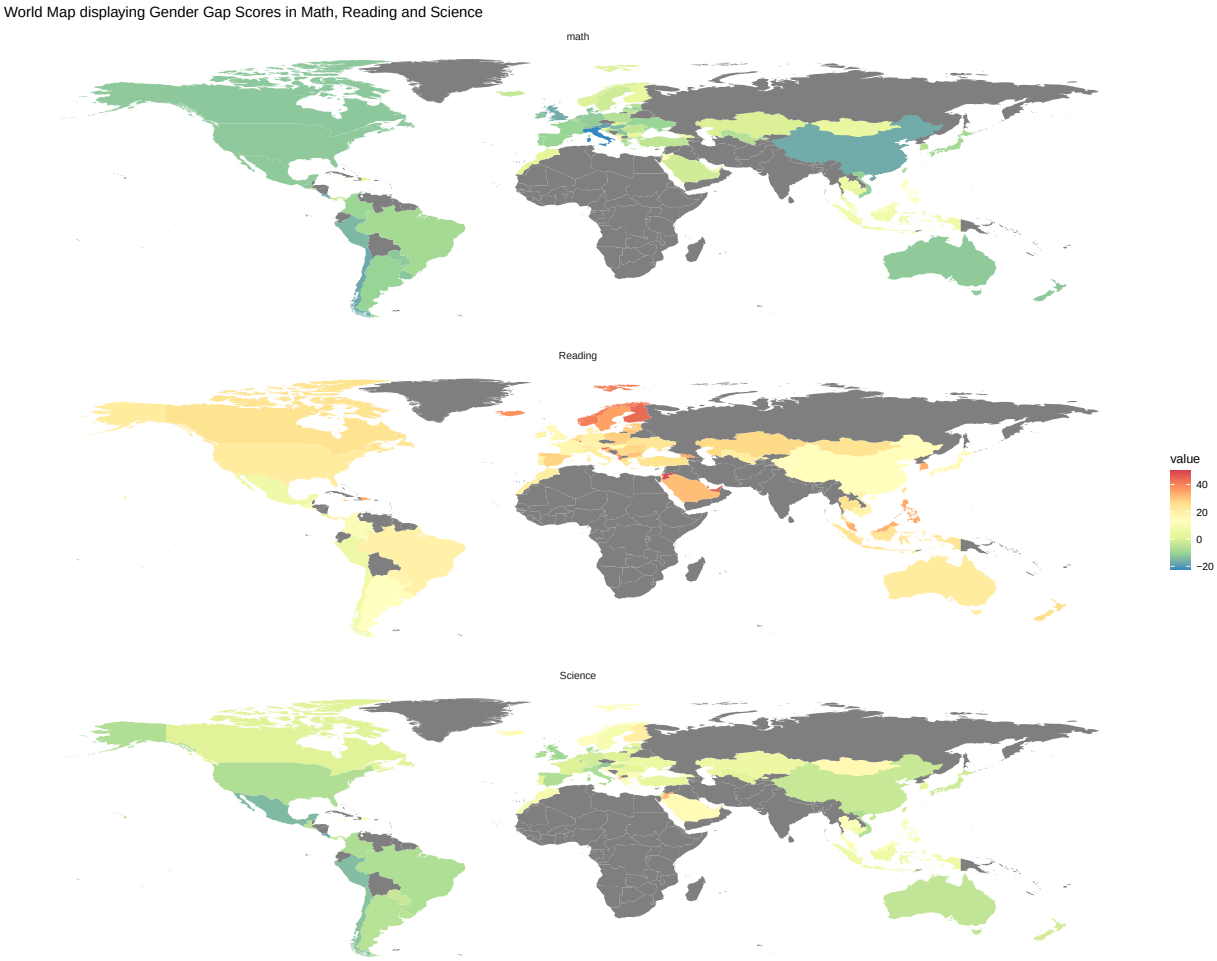


Figure 2: *Maps showing the gender gap in math, reading, and science results between girls and boys across the world. A positive score for a country indicates that girls outperformed boys in that country, whereas a negative score for a country difference indicates that boys outperformed girls in that 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.

10 Socioeconomic factors

Socioeconomic status is an economic and sociological complete measure of a person's work experience, economic access to resources, and social standing in relation to others. Do these socioeconomic factors influence students' academic performance? In this section, we will investigate if different socioeconomic factors owned by a family have a significant impact on a student's academic performance. The student dataset in the `learningtower` package contains scores of 15-year-olds from triennial testing across the world. This dataset also includes information about their parents' education, family wealth, gender, and ownership of computers, internet, cars, books, rooms, desks, and dishwashers. Next, we will mainly explore some fascinating aspects of the influence of a few socioeconomic factors on student performance in math, reading, and science. Let us further explore the impact of a selection of socioeconomic factors on the students' score.

Parents qualification is a vital element of childhood development. As previously stated, the student dataset in the package includes information regarding the parents qualification. In this section, we will investigate if both the mother's and father's qualifications have a significant impact on their child's performance. The mother's education and father's education variables are originally recorded in the student dataset in the `learningtower` package at distinct International Standard Classification of Education (ISCED) levels which are less than ISCED1 equivalent to ISCED 0, ISCED 1, ISCED 2, ISCED 3A and ISCED 3B, C, where:

- level 0 indicates pre-primary education or no education at all
- level 1 indicates primary education or the first stage of basic education
- level 2 indicates lower secondary education or the second stage of basic education, and
- level 3 indicates upper secondary education. ISCED level 3 have been further classified into three distinct levels, with ideally very little difference in their classification. This may also be found in the publication OECD Handbook for Internationally Comparative Education Statistics (Economic Cooperation & Development 1999) published by OECD.

To determine the impact of the parents' qualification we first create data frames that are categorized by the various countries and regions and grouped by the father's and mother's qualification. We next compute the weighted average of math scores while accounting for student survey weights. Furthermore, we re-factored the parents qualification variable based on the multiple levels of classification, dividing it into four unique levels of education, namely early childhood, primary, lower, and secondary education. Furthermore, we display the weighted math average versus qualification colored by the re factored qualifications levels for both the mother and father using the `geom_quasirandom` function wrapped within `ggplot2` (Wickham 2016), we further plot this with the help of `viridis` (Garnier et al. 2021) and `patchwork` (Pedersen 2020) packages in R.

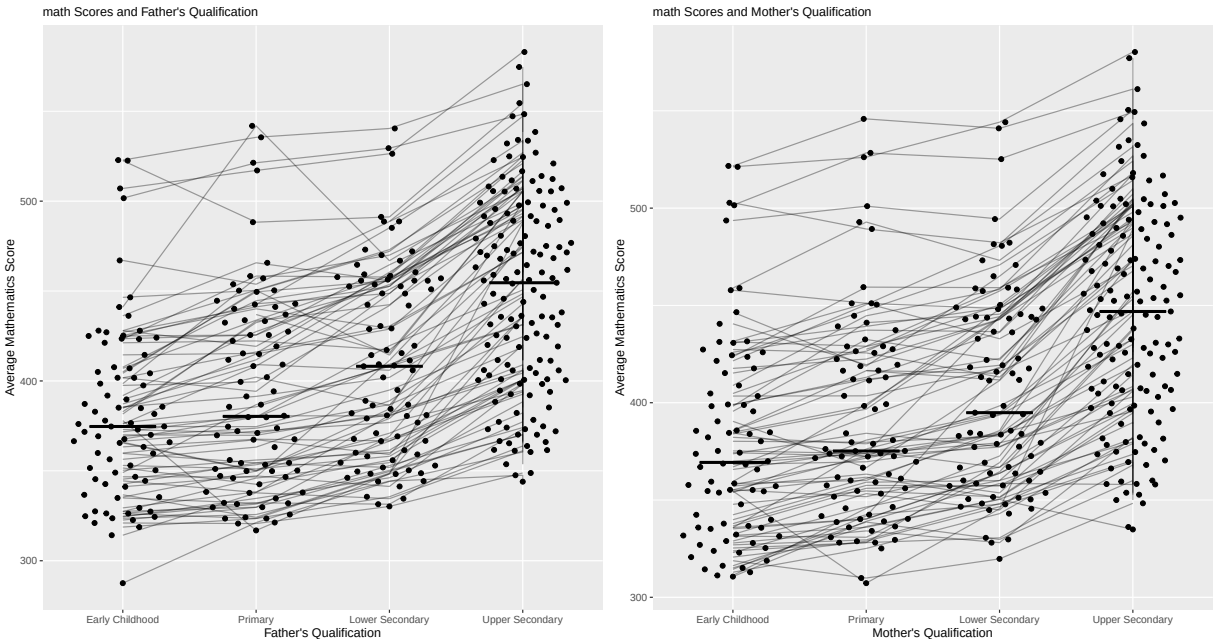


Figure 3: *The impact of parents' education on their children's academic progress is depicted in this graph. When the parents have greater levels of education, we see a considerable rise in scores and an increase in the median of scores for each category, as shown in the figure. In comparison to parents 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.*

The Figure 3 depicts the impact of mothers' and fathers' qualifications on students academic performance. The Figure 3 allows us to deduce a very important and remarkable insight in which we see a constant increase in the students' academic performance when both mother and father qualifications shift towards higher levels of education. The bold horizontal black lines that we see in each category for mother's and father's qualification here represent median score for that qualification category across countries. As the parent attains higher qualifications, we notice an increasing trend in these medians for each category. Taking a closer look at the Figure 3, we can see that there is a considerable boost in scores when both the mother and father have upper secondary education. Furthermore, the

`geom_quasirandom()` function in the `ggbeeswarm` (Clarke & Sherrill-Mix 2017) package makes this plot more accessible and understandable by providing a way to offset points inside categories to prevent overplotting. Thus, we can clearly see that both the mother's and father's qualifications has a significant influence on the student's academic performance, with the more educated the parent more likely to have their child academically performing better.

Students are becoming more active and adept learners as technologies like computers and internet mature and becoming more commonplace over the past twenty years. We will investigate if having a computer with internet access at the age of 15 has a positive or negative impact on student academic achievement. We will plot the average math results of the several nations that participated in the PISA experiment in 2022 to determine the effect of owning a computer and having access to the internet. We first create `data.frame` that is grouped by the nations and the frequency of whether the student possesses a computer or not, as well as a students' access to the internet or not. We will plot this result against the weighted average mathematical score to determine the influence of various of television and internet on the student academic performance using the several functions available in the `ggplot2`(Wickham 2016) package.

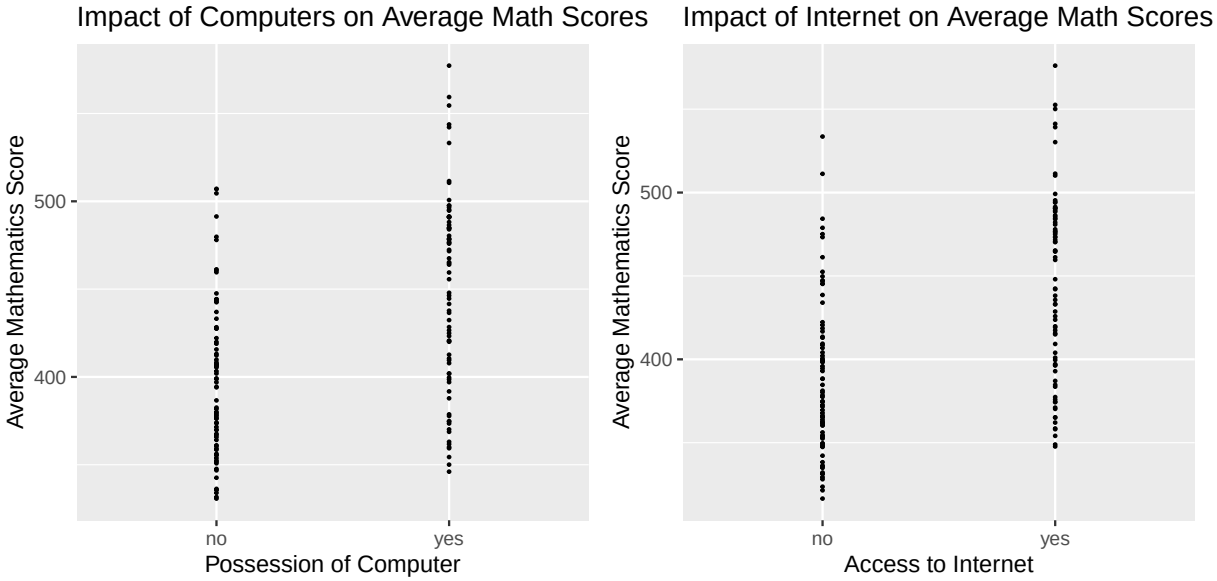


Figure 4: *Computers and the Internet are two of the most important inventions in the history of technology. In this figure, we observe the impact of owning a computer and having access to the internet on 15-year-old students all over the world. A remarkable finding from the plot is that all nations have higher scores in student performance when they own a computer and have access to the internet.*

In the Figure 4, we see that students who own a computer and have access to the internet consistently outperform students who do not own a computer or have access to the internet at all. No country has exempted from this finding. Thus, on average, a 15-year-old student's access to a computer and the internet is unquestionably has significant positive influence on their academic performance. While the

increase in academic performance is expected, but the magnitude of this increase and the associated educational benefits may be used by policymakers to improve their own domestic access to these technologies.

Followed on previous discussion, computer has became an essential in learning experience. In this segment of the article, we investigate the influence of number of computers by countries/regions, as well as whether this technology has a significant impact on students' academic performance. The computer variables that are recorded in the student dataset is a factor variable that records whether or not the students participating in this study have a computer and, if they do, the quantity of computers per family is recorded via the PISA survey. Especially, in 2022 dataset, the survey also included the laptop variable. The computer and laptop variables are initially recorded has four levels: "No computer/laptop", "1 computer/laptop", "2 computers/laptops", or "3+ computer/laptops", respectively. Because we are interested in researching the impact of these computers and laptops on the students' scores, we visualized the confidence intervals for each of these levels in order to determine the uncertainty of the results at each level. We begin with initially creating a `data.frame` that is grouped by country and the number of computers and laptops per household for each country. Next we fit a linear model between the math average with computer and laptops in the 2022 PISA data and finally plot the impact for all countries sorted as per the slope with the help of the functions available in the `ggplot2` (Wickham 2016) package.

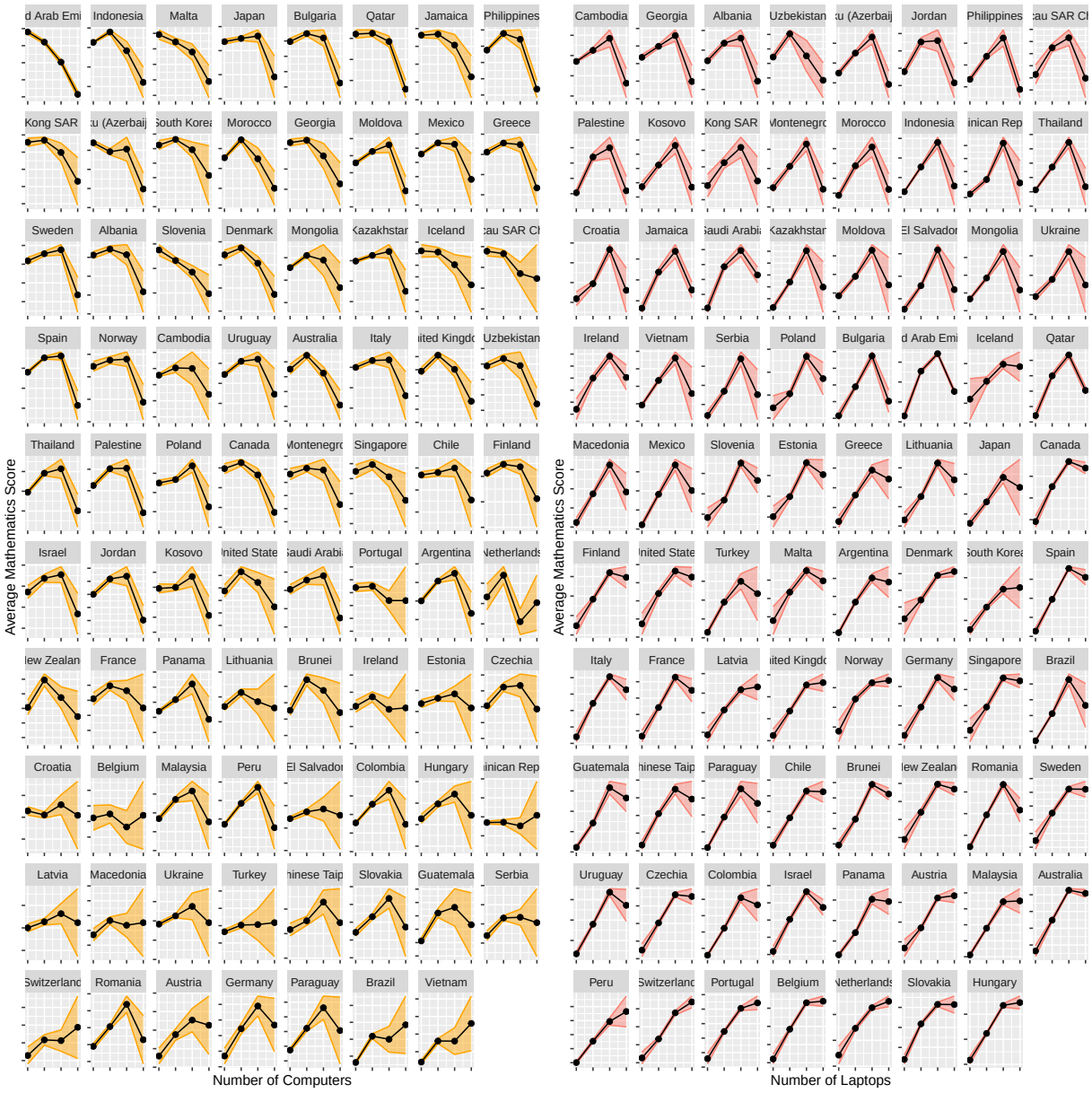


Figure 5: *Relationship between number of computers and laptops in a household and average math scores across countries. Number of computers/laptops ranges from 0 to 3 or more. The orange bands indicate 95 percent standard confidence intervals. The impact of computer/laptop on student performance is a contentious issue. It is interesting that the effect of computers and laptops are different across countries*

In the Figure 5, we can see highly striking patterns as well as a significant influence of television on students' academic performance. We have arranged the nations in the Figure 5 according to the slope of math average scores fitted against the different levels of computer/laptop described previously. United Arab Emirates and Indonesia have a lower influence of computer on student performance, whereas Brazil and Vietnam have a rising tendency and therefore a larger impact of computer on students' performance. Furthermore, the confidence interval plotted in the figure Figure 5 show that there is a lot of uncertainty in the level of scores when a household have more computers in

the majority of the countries. On the other hand, Cambodia and Georgia have a lower influence of laptop on student performance, whereas Slovakia and Hungary have a rising tendency and therefore a larger impact of laptops on students' performance. Taking a closer look we observe that when the slope of laptop increases in countries, compared to the slope of computer, the confidence interval of such countries becomes narrower. Hence laptop could have more potential educational benefit compared to computer. It's quite interesting because the similar nature of computer and laptop, we would assume the effect of these two assets should be similar, yet the Figure 5 has not only shown a different result. Nevertheless, having multiple computers and laptops may turn out to be having negative influence on student's performance in most of the countries, excessive access could lead to distractions or excessive screen time, which can negatively impact performance.

10.1 Temporal Analysis

The 2022 PISA results offer a unique opportunity to analyze the impact of the COVID-19 pandemic on student's academic performance. As countries faced prolonged school closures, remote learning challenges, and varying socio-economic pressures, concerns grew over potential learning losses and their effects on core academic skills. In this section, our temporal analysis will examine how student performance in each subject may have shifted compared to pre-pandemic data, providing insights into the pandemic's educational impact. Additionally, the study will revisit the gender gap, a persistent trend in PISA assessments. By investigating whether this gap has narrowed or persisted in 2022, we can better understand how gender disparities in education might have evolved, offering valuable guidance for future efforts to support equity and learning recovery worldwide.

Pandemic effects

To better comprehend pandemic effect on student's academic performance, We illustrate the temporal trend of Australia in comparison to a few other nations. We evaluate these countries performance using a statistical procedure bootstrapping using the `map_dfr` function which re-samples a single dataset to generate a large number of simulated samples. We will compare the results of these bootstrap samples across all the years they participated in PISA.

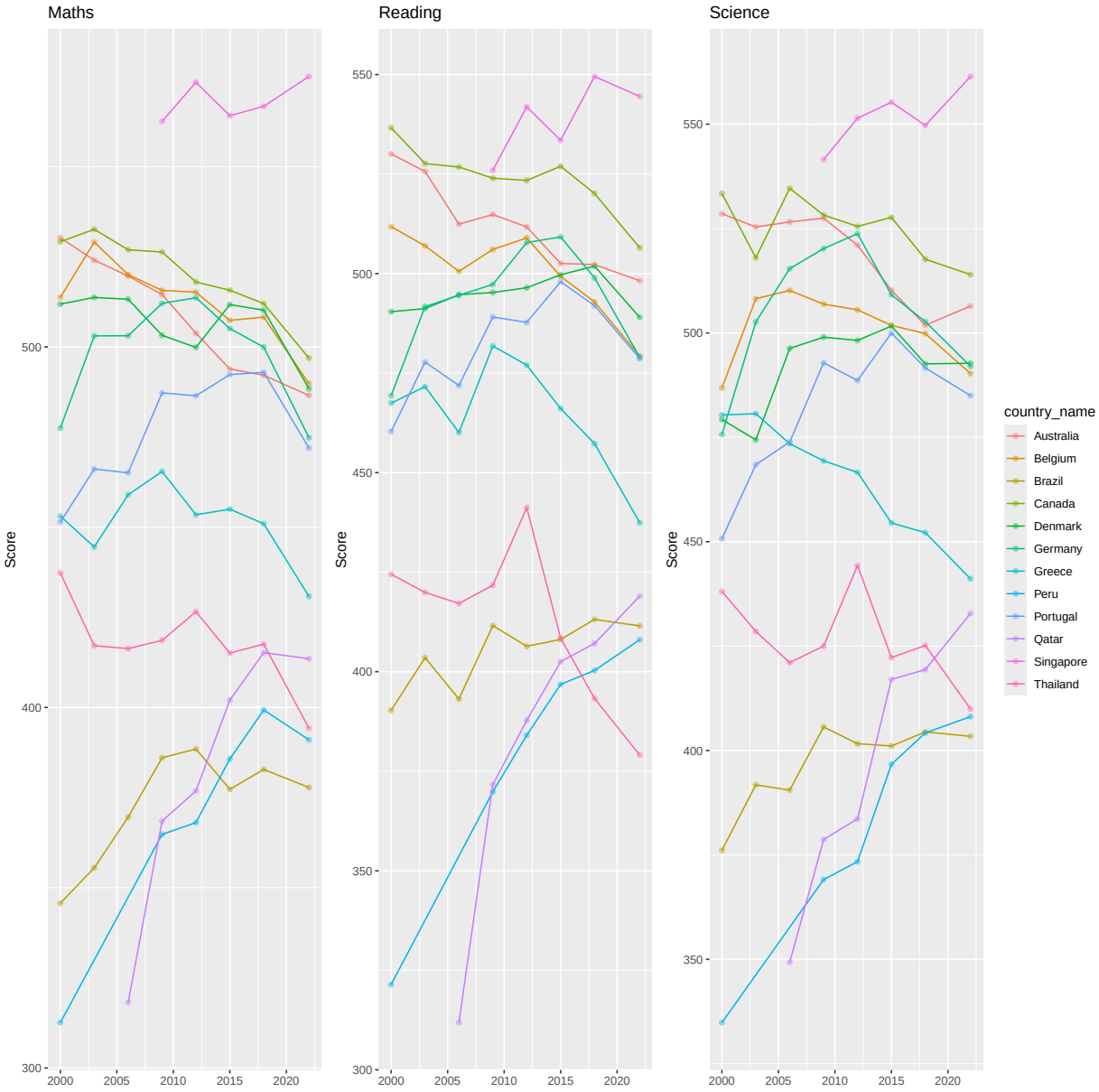
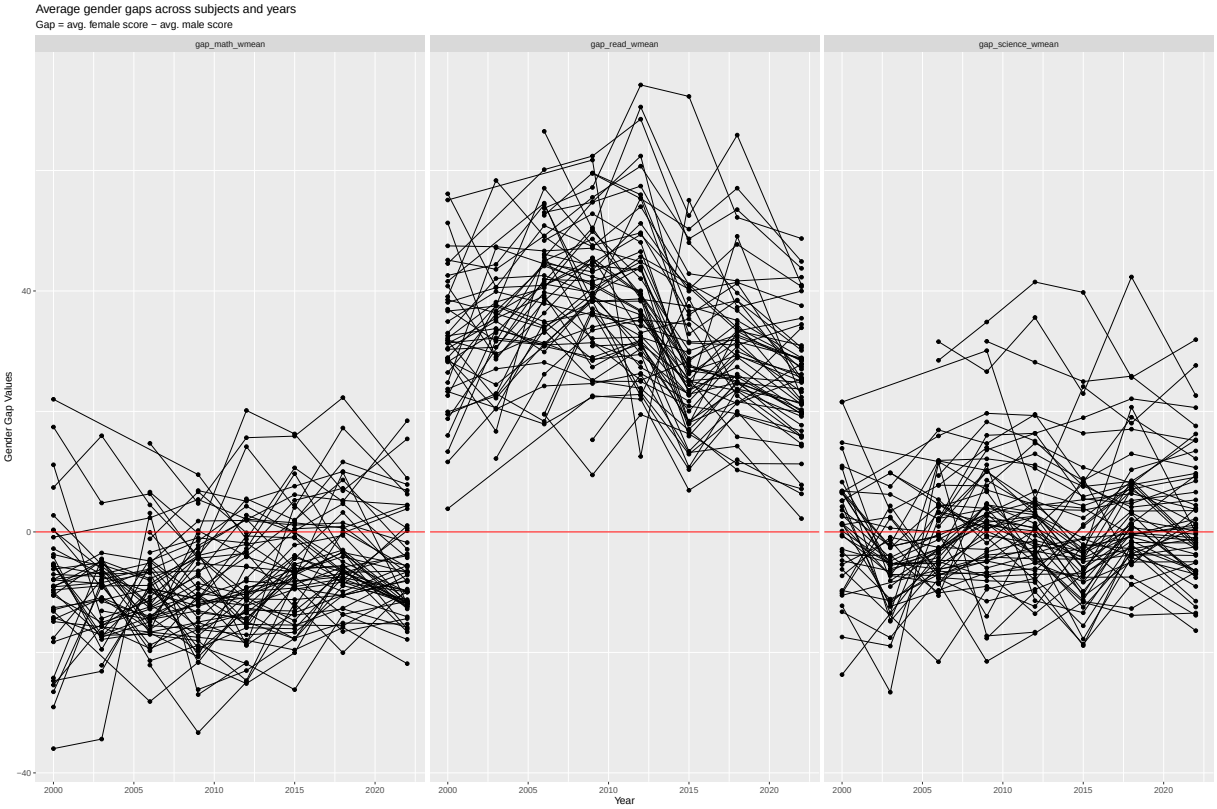


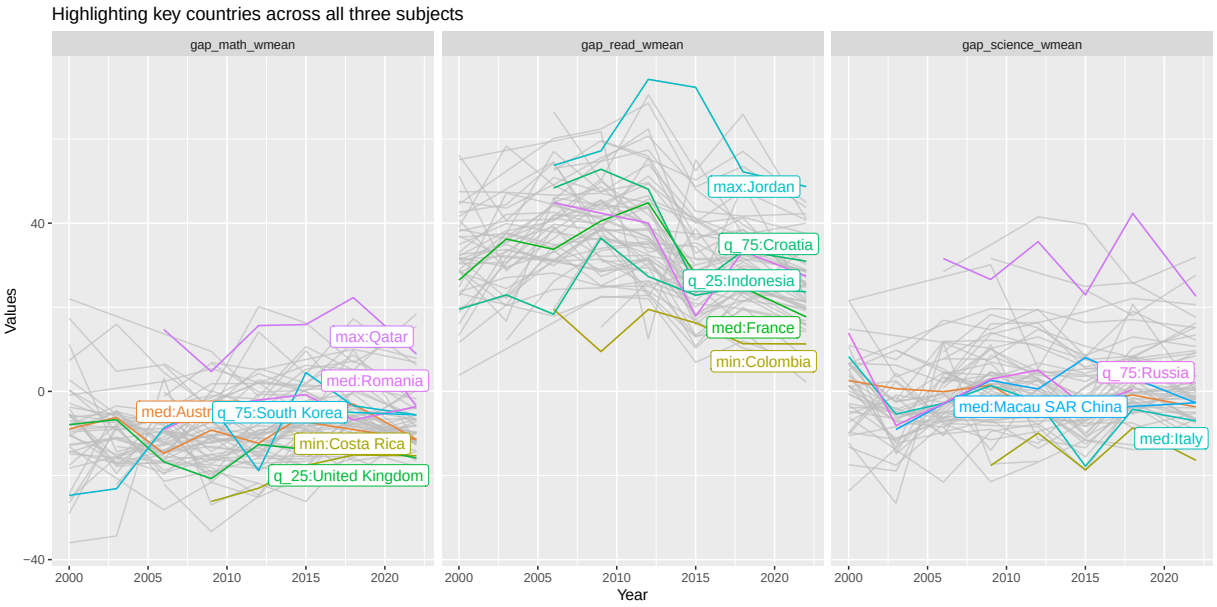
Figure 6: *Temporal patterns in math, reading, and science in a variety of countries. The highlighted countries in the chart help us infer Australia's performance in contrast to the other countries; we can see that Australia's scores have always been among the highest in the PISA survey throughout all years.*

Taking a deeper look at the figure Figure 6, we notice the changing scales of their scores in all three plots of math, reading, and science and we could observe that there's a obvious decline in mathematics score across countries, only Singapore still demonstrate increase in scores; same in reading most countries have decline, yet only Qatar and Peru have shown improvement. As for science, there's no clear drop in score, most countries have maintained in the same level as previous years.

Gender Gaps Across Subjects and Years



Highlighting Key Countries



11 Discussion

11.1 Results

The learningtower package's study of PISA data reveals important educational trends that affect student performance globally, particularly in the areas of gender, socioeconomic status, and technology access. Girls typically outperform boys in reading across all countries, according to the gender-based data, with particularly large gaps observed in Jordan, Palestine, and the United Arab Emirates. This pattern implies that although reading is typically thought of as a subject in which girls do best, this disparity may potentially be exacerbated by cultural and educational variables. Though some nations, like Sweden and Kazakhstan, exhibit near parity, suggesting potential success in implementing gender-neutral or inclusive educational approaches in maths, boys still often perform better than girls in maths. In contrast, gender variations in science scores are negligible in many regions, indicating that science instruction may inherently support more balanced performance between boys and girls.

In every subject, socioeconomic considerations show up as important predictors of academic performance. Higher scores are typically attained by students with more educated parents and those from higher ESCS (Economic, Social, and Cultural Status) backgrounds. Learning is impacted by parental support, family wealth, and access to educational resources, as seen by the positive link found between socioeconomic resources and student performance. Additionally, having access to technology—such as home computers and internet connectivity—is substantially linked to better academic achievement; kids who have these resources often outperform those who do not. Performance differences are less in nations with higher levels of digital penetration, indicating that equal access to technology may be crucial in closing the achievement gap between kids from different socioeconomic backgrounds.

Student performance trends throughout time demonstrate the COVID-19 pandemic's effects, especially in mathematics, where 2022 scores fell in many nations. This pattern might be a reflection of the difficulties associated with distance learning and the scarcity of resources while schools are closed. On the other hand, science and reading scores were comparatively stable, with several nations—like Peru and Qatar—even seeing improvements in their reading proficiency. This resiliency in science and reading might indicate that these courses were better served by online resources or that students were better able to adjust to distance learning for these skills. The significance of flexible teaching strategies and infrastructure is emphasised by the temporal analysis, which can protect educational systems from interruptions and support the maintenance of uniform learning outcomes across disciplines.

The learningtower package's mapping and visualisation capabilities offer a thorough understanding of these findings, facilitating more understandable cross-national comparisons and emphasising distinctive regional insights. The significance of investments in digital and educational resources is further supported by the fact that nations with strong digital infrastructure and fair socioeconomic

conditions typically exhibit narrower achievement inequalities. Together, these results highlight the necessity of evidence-based approaches that tackle regional and national educational issues in addition to global trends. The learningtower package supports efforts to build more inclusive, equitable, and resilient educational systems around the world by educating educators and policymakers about these important discoveries.

11.2 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.
- Handling of Missing Data: Some variables are missing or incomplete in certain years or countries. These missing data points can limit the analysis and reduce comparability accross regions or over time.
- Dependency on External data formats: The package relies on data orginally available in SPSS or SAS formats from the OECD. Any changes in these formats, structures or access protocols on the OECD website could impact the package's ability to import and process the data accurately.
- Potential Sampling Bias: Although PISA aims for representative samples, certain countries or regions may face challenges in collecting the data that fully reflects their population. This could introduce sampling bias.

12 Conclusion

The learningtower R package has been significantly improved with the addition of the 2022 PISA data, providing researchers, educators, and policymakers with a useful tool for examining and comprehending student performance and school characteristics worldwide. This most recent version guarantees a consistent and thorough dataset covering the years 2000–2022, in addition to providing up-to-date insights concerning educational outcomes. The software enables precise cross-country comparisons and thorough longitudinal studies by preserving consistency in variables and data structures, both of which are essential for spotting and solving educational trends.

The learningtower package offers a strong framework for investigating important topics in education, including socioeconomic effects, gender inequality, and the function of technology in the classroom. Users can explore the ways in which various factors, such as parental education levels and digital access, impact student accomplishment with this degree of exposure to comprehensive, curated data. This makes it possible for stakeholders to make data-driven choices that have a direct impact on educational policies and initiatives meant to raise educational quality and equity.

The learningtower package stresses the value of open, repeatable research in addition to its analytical capabilities. The package facilitates a collaborative environment where discoveries may be shared and validated across research groups by providing the data and tools in an easily accessible format. This openness promotes future developments in the discipline when new PISA data becomes available and enhances the validity of educational analyses.

In the future, the learningtower package will continue to be updated with new PISA cycles, maintaining its usefulness and relevance as a vital resource for researching worldwide trends in education. The package's capacity to include new data will guarantee that it continues to be an essential tool for comprehending and resolving educational inequities as the educational landscape changes, especially in reaction to crises like the COVID-19 pandemic. Learningtower will promote evidence-based educational advancements with continued contributions and enhancements, eventually leading to a more effective and equitable global education system.

13 Reference

13.1 Git respository of the report

https://github.com/Shabarish161/Learningtower_Rpackage

References

Clarke, E & S Sherrill-Mix (2017). *ggbeeswarm: Categorical Scatter (Violin Point) Plots*. R package version 0.6.0. https://CRAN.R-project.org/package=ggbeeswarm.

Economic Cooperation, O for & Development (1999). *ISCED-97 Implementation in OECD Countries*. Accessed: 2021-11-03. https://www.oecd.org/education/1841854.pdf.

Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, A Pedro, Sciaini, Marco, Scherer & Cédric (2021). *viridis - Colorblind-Friendly Color Maps for R*. R package version 0.6.2. https://sjmgarnier.github.io/viridis/.

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.