NATIONAL TECHNICAL UNIVERSITY OF ATHENS

MSc in Data Science and Machine Learning

Exploratory Data Analysis using R

PISA 2015: Gender and Nationality Impact

Theofili Stavroula
Student ID: 03400166
Email: stheofil98@gmail.com

January 26, 2024

Contents

1 Introduction

The Program for International Student Assessment (PISA) is a globally recognized initiative designed to evaluate the academic performance students in areas like Mathematics, Reading, and Science. This comprehensive assessment, administered across various countries, serves as a valuable tool for understanding the educational landscape on an international scale. Our project focuses on conducting an in-depth exploratory data analysis on the PISA dataset, aiming to shed light on the impact of factors such as gender, country of origin, and regional variations on students' academic achievements. Our investigation unfolds in two distinct parts. The first part of our analysis delves into the influence of countries and regions on students' academic performance, while transitioning to the second part of our project, we direct our focus towards the gender of students. By orchestrating this comprehensive analysis, we seek to contribute valuable insights to the broader discourse on education, offering a holistic understanding of the intricate factors shaping students' performance on a global scale.

2 Data Preprocessing

The cleaning and preprocessing procedures were a prerequisite for the analysis of the available data. To begin with, we install and load the `data.table` package into the R environment. Then, we read these data from a csv file, indicating that the first row of the file contains column names. Before checking for missing values, we decide to rename the column named "2015" to "Performance Score in 2015" for better understanding. During the inspection of our data, we realise that the only column that includes missing values is the last one ("Performance Score in 2015"). To be more specific, there are 61 countries for whom we don't have the students' scores. These countries cannot be included in our analysis because they don't provide us with the information needed. Therefore, we create a new dataset, excluding rows (countries) where the "Performance Score in 2015" is missing.To continue, we check and print the structure of our `ndata` data frame. As we can see, the "Performance Score in 2015" column contains non-numeric characters which need to be converted in order to be able to perform our analysis and conduct some comparisons. To achieve that, we use the `as.numeric()` function to convert these values from character (char) data type to numeric. Our last step in data preprocessing is to drop 2 columns: "Country Code", "Series Name". In my estimation, these columns are unnecessary since we already have "Country Name" and "Series Code" as identifiers and descriptors that capture the same information.
The R Code developed for the aforementioned steps is presented below (R Code Snippet 1).

```
install.packages("data.table") #Install the data.table Package
library(data.table) #Load the data.table Package

#Read Data from a CSV File
data <- fread("C:\\Users\\Stavroula Theofili\\Downloads\\Pisa
   mean perfomance scores 2015 Data.csv" , header=TRUE)
   print(data)

#Change the Name of the Column '2015'
```

```r
setnames(data, old = "2015", new = "Performance Score in 2015")

#Missing Values
total_missing_values <- sum(data$`Performance Score in 2015` == "..", na.rm = TRUE) #Counting Total Missing Values
countries_with_no_info <- unique(subset(data, `Performance Score in 2015` == "..")[, "Country Name"]) #Listing Countries with Missing Values
print(countries_with_no_info)

#Creating a Dataset without Missing Values
data_without_missing_values <- subset(data, `Performance Score in 2015` != "..")
print(data_without_missing_values)
ndata <- data_without_missing_values

#Check datatype - Make last column numeric
summary(ndata)
numeric_columns <- sapply(ndata, is.numeric)
print(numeric_columns)
ndata$`Performance Score in 2015` <- as.numeric(ndata$`Performance Score in 2015`)

#Drop Columns "Country Code", "Series Name"
columns_to_remove <- c("Country Code", "Series Name")
ndata <- ndata[, !(names(ndata) %in% columns_to_remove), with = FALSE]
```

Snippet 1: Data preprocessing.

3 Data Analysis

With the dataset cleaned and organized into a data table including the columns 'Country Name', 'Series Code' and 'Performance Score in 2015', the process of data analysis shall begin. As mentioned in the Introduction, this process was divided into two parts. The first part focuses on how countries and regions affect the students' performance in the three disciplines, while in the second part we emphasize on the gender of each student, making comparisons between males and females.

3.1 Impact of Geographic Factors on Student Performance

In order to start analyzing the influence of countries and regions on student achievements, we first need to create a new data frame named `mean_scores`. This data frame is constructed by filtering the original dataset, `ndata`, to include only relevant entries related to performance scores in the three disciplines (Mathematics, Reading, and Science) using the specified series codes ('LO.PISA.MAT', 'LO.PISA.REA', 'LO.PISA.SCI'). Subsequently, we group the filtered data by country using the `group_by` function from the `dplyr` library. For each country, we calculate the mean score in 2015 across the three disciplines, and this information is stored in the new data frame. Finally, we arrange the data frame in descending order based on the mean scores, providing a clear representation of the average performance in 2015 across countries. The R Code developed for the creation of our new data frame is presented in Snippet 2:

```
library(dplyr)

#Keep the rows where 'Series Code' matches 'LO.PISA.MAT','LO.PISA.REA', 'LO.PISA.SCI'
filtered_data <- ndata %>%
     filter(`Series Code` %in% c('LO.PISA.MAT', 'LO.PISA.REA', 'LO.PISA.SCI'))

#Calculating Mean Scores
mean_scores <- filtered_data %>%
      group_by(`Country Name`) %>%
      summarise(
           `Series Code` = 'LO.PISA.MEAN',
           `Mean Score in 2015` = mean(as.numeric(`Performance Score in 2015`), na.rm = TRUE)
       )

#Arranging in Descending Order
mean_scores <- mean_scores %>%
      arrange(desc(`Mean Score in 2015`))
```

Snippet 2: Creating the new dataframe of Mean Scores

After organizing the mean_scores dataframe, which contains the mean PISA scores for each country in 2015, we can proceed to the visualization of our data. Using the ggplot2 library we create a bar plot representing the mean scores of different countries. The bars are colored differently, with a distinctive shade for Greece (red color) which helps us emphasize our country's position within the dataset. The barplot described is presented in Figure 1.

Following the bar plot, we generated a heatmap (Figure 2) to provide a nuanced visualization of the mean PISA results. The color gradient, ranging from the lightest to darker blue hues, effectively conveys the variations in student performance across different countries. Notably, countries without available information are excluded from the map, allowing for a focused analysis of the presented data.
The figures, along with their corresponding comments and the code utilized to generate them, are presented below.

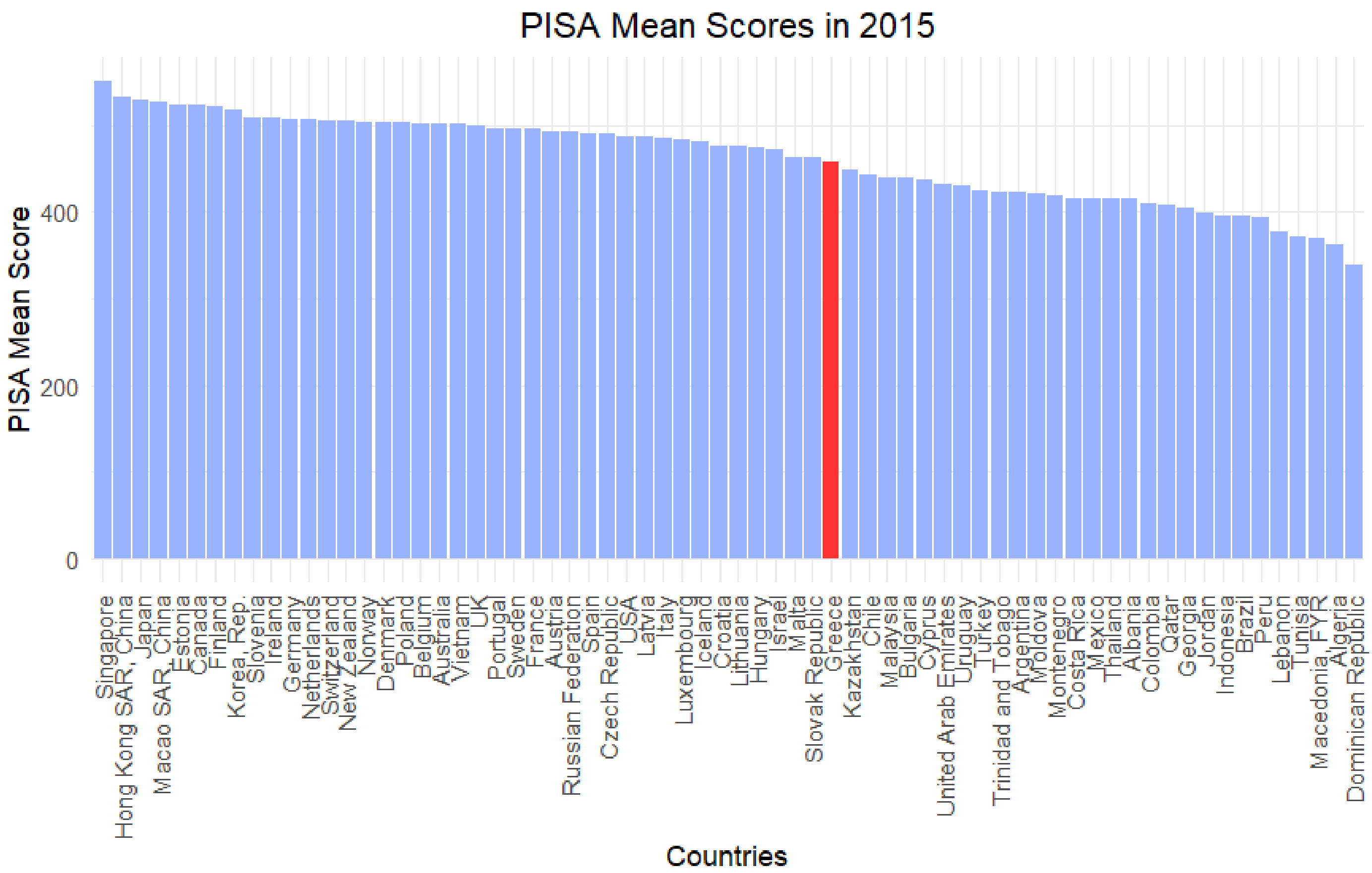

Figure 1: Mean Scores of Different Countries: Highlighting Greece in Red

As we observe, the top-performing countries based on the mean performance score of the three subjects, are Singapore, Hong Kong, Japan, Macao, and Estonia. The prominence of these countries in the top 5 positions is not surprising given the distinct educational strategies and societal priorities embraced by these nations. To be more specific, Singapore's systematic and forward-thinking education policies have consistently placed it at the forefront of global educational rankings. The country's emphasis on teacher quality, rigorous curriculum, and a supportive learning environment fosters a culture of excellence. Similarly, Hong Kong and Japan's historical commitment to education, coupled with disciplined study habits and a culture valuing academic achievement, positions them consistently among the top performers. Finally, Macao's multicultural educational environment and Estonia's innovative approach to integrating technology, contribute to their exceptional performance.

In contrast, Greece's position at No. 40 among 68 countries is quite disappointing. This specific ranking may be influenced by several factors. Economic challenges in recent years could have impacted education funding, leading to resource constraints and potentially affecting the quality of education. Additionally, systemic issues, such as outdated curriculum or teaching methods, could also contribute to our country's placement in the middle of the bar plot.

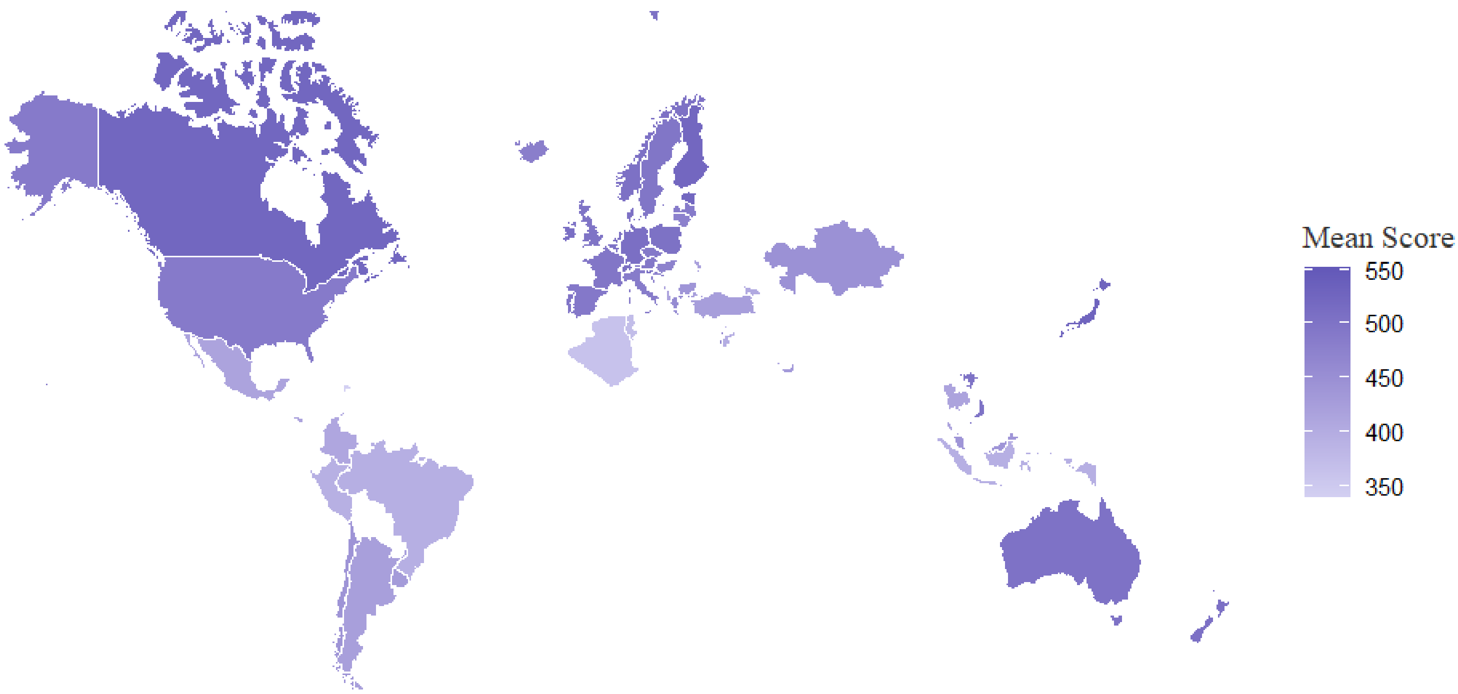

Figure 2: World Heatmap about Mean PISA Scores

Upon careful examination of the heatmap illustrating PISA scores in 2015, a discernible pattern emerges, indicating that countries with the highest-performing students are predominantly situated in Australia and nearby regions, as well as in Northern Europe and North America (Canada). The observed concentration of top-performing nations in these areas may be attributed to a combination of various factors. First and foremost, these regions often prioritize and invest significantly in education, fostering an environment that values academic achievement. Moreover, educational policies in these countries frequently emphasize innovation, critical thinking, and a holistic approach to learning, contributing to the development of well-rounded and highly skilled students. Additionally, cultural attitudes towards education, parental involvement, and robust support systems for teachers may also play pivotal roles in shaping the academic success of students in these regions.

On the other hand, it becomes evident that some countries in North Africa, for which data is available, exhibit lower academic performance compared to other regions. The observed lower scores in these North African nations may be attributed to a combination of socioeconomic challenges and disparities in educational resources. Economic factors, such as poverty and limited access to quality education, can significantly impact the learning environment in these regions.

The code developed for the analysis of mean PISA scores in 2015, including the creation of the heatmap, is presented in Snippet 3.

```
# Mean Scores Barblot - Red GREECE
```

```
library(ggplot2)
mean_scores <- mean_scores[order(-mean_scores$‘Mean Score in 2015‘), ]

# Data Preparation
mean_scores <- mean_scores %>%
mutate(color = ifelse(‘Country Name‘ == "Greece", "#99B3FF", "#FF3333"))

# Plotting
ggplot(mean_scores, aes(x = reorder(‘Country Name‘, -‘Mean Score in 2015‘), y = ‘Mean Score in 2015‘, fill = color)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
      legend.position = "none") +
  labs(x = "Countries", y = "PISA Mean Score", title = "PISA Mean Scores in 2015") +
  scale_fill_manual(values = c("#FF3333", "#99B3FF")) +
  theme(plot.title = element_text(hjust = 0.5))

# Heatmap
mean_scores$‘Country Name‘[mean_scores$‘Country Name‘ == "United States"] <- "USA"
mean_scores$‘Country Name‘[mean_scores$‘Country Name‘ == "United Kingdom"] <- "UK"

# Load world map data
world <- map_data(’world’)

# Create the input for the world map
wdt <- world[world$region %in% mean_scores$‘Country Name‘, ]
wdt$value <- mean_scores$‘Mean Score in 2015‘[match(wdt$region, mean_scores$‘Country Name‘)]

# Plot the world map
plot_mean_scores <- ggplot(wdt, aes(x = long, y = lat, group = group, fill = value)) +
  geom_polygon(colour = "white", size = 0.05) +
  theme_bw() +
  scale_fill_gradient(low = "#d3d0f2", high = "#6258b8", name = "Mean Score") +
  labs(fill = "Mean Score", x = "", y = "") +
  scale_y_continuous(breaks = NULL) +
  scale_x_continuous(breaks = NULL) +
  coord_map(xlim = c(-170, 170), ylim = c(-50, 100)) +
  theme(panel.border = element_blank(),
      plot.title = element_text(hjust = 0.5, size = 15, face = "plain", color = "#333333", family = "serif", margin = margin(b = 10)),
```

```
        legend.title = element_text(size = 12, face = "plain",
           color = "#333333", family = "serif"))

# Display the plot
print(plot_mean_scores)
```

Snippet 3: Barplot of Mean Scores - Heatmap

To delve into more specific details after scrutinizing the global heatmap, we narrow our focus to a European context. The European heatmap, presented in Figure 3, allows us to concentrate our analysis on the intricacies of our continent and draw meaningful comparisons.

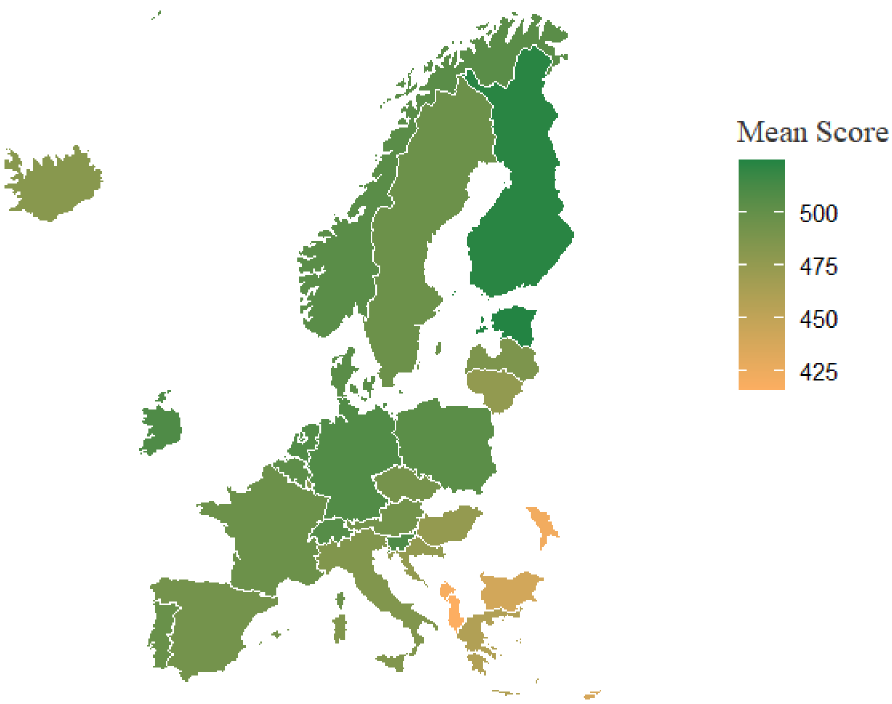

Figure 3: European Variations: Heatmap of Educational Performance

Highlighting the diverse educational landscape within the continent, the European heatmap in Figure 3 reaffirms our earlier findings from the barplot in Figure 1. Estonia and Finland stand out as high-performing nations, showcasing notable PISA scores. A closer inspection of the map reveals Greece positioned among countries with comparatively lower scores in all three subjects, underscoring the nuanced educational disparities present across Europe.

```
#European HeatMap
european_countries <- c("Albania", "Andorra", "Austria",
   "Belarus", "Belgium", "Bosnia and Herzegovina", "Bulgaria",
   "Croatia", "Cyprus", "Czech Republic", "Denmark", "Estonia",
   "Finland", "France", "Germany", "Greece", "Hungary",
```

```
    "Iceland", "Ireland", "Italy", "Kosovo", "Latvia",
    "Liechtenstein", "Lithuania", "Luxembourg", "Malta",
    "Moldova", "Monaco", "Montenegro", "Netherlands", "North
    Macedonia", "Norway", "Poland", "Portugal", "Romania",
    "Russia", "San Marino", "Serbia", "Slovakia", "Slovenia",
    "Spain", "Sweden", "Switzerland", "Ukraine", "United
    Kingdom", "Vatican City")

# Filter world map data to include only European countries
wdt_europe <- wdt[wdt$region %in% european_countries, ]

# Plot the heatmap for Europe
plot_mean_scores_europe <- ggplot(wdt_europe, aes(x = long, y =
    lat, group = group, fill = value)) +
  geom_polygon(colour = "white", size = 0.05) +
  theme_bw() +
  scale_fill_gradient(low = "#fdae61", high = "#238443", name =
     "Mean Score") +  # Orange to green color palette
  labs(fill = "Mean Score", x = "", y = "") +
  scale_y_continuous(breaks = NULL) + scale_x_continuous(breaks
     = NULL) + coord_map(xlim = c(-20, 40), ylim = c(35, 75)) +
     theme(panel.border = element_blank(), plot.title =
     element_text(hjust = 0.5, size = 15, face = "plain", color
     = "#333333", family = "serif", margin = margin(b = 10)),
     legend.title = element_text(size = 12, face = "plain",
     color = "#333333", family = "serif"))

# Display the plot for Europe
print(plot_mean_scores_europe)
```

Snippet 4: Creating the European HeatMap

3.2 Impact of Gender on Student Performance

In this subsection, as we navigate through the data, our focus shifts towards the gender dimension, unraveling the intricacies of student performance. By distinguishing between male and female students, we aim to shed light on any existing disparities in the subjects of Mathematics, Reading, and Science. Are there discernible trends in performance that align with traditional gender stereotypes, or do the data reveal a more nuanced picture of academic achievement?

In order to start analyzing the influence of the gender on student achievements, we firstly need to transform our data and calculate the differences, total score, and average difference for each country across the three different subjects. To be more specific, we use the `reshape` function to transform our data from long format to wide format, making each unique "Series Code" a separate column, specifying "Country Name" as the identifier variable (`idvar`), "Series Code" as the time variable (`timevar`), and finally dropping the columns containing "Performance Score in 2015". Then we use `gsub` to remove the common prefix "Performance Score in 2015." from the column names and we compute the differences for each subject (Maths, Reading, Science) and the average difference across genders. The R Code developed for these alterationsis presented in Snippet 4:

```
#Reshape Data
de_data <- reshape(
     ndata,
     idvar = "Country Name",
     timevar = "Series Code",
     direction = "wide",
     drop = "Performance Score in 2015." )

#Rename Columns
colnames(wide_data) <- gsub("Performance Score in 2015.", "",
   colnames(wide_data))

#Calculate Differences, Total Score, and Average Difference:
wide_data$Maths.Diff <- ((wide_data$LO.PISA.MAT.FE -
   wide_data$LO.PISA.MAT.MA) / wide_data$LO.PISA.MAT.MA) * 100
wide_data$Reading.Diff <- ((wide_data$LO.PISA.REA.FE -
   wide_data$LO.PISA.REA.MA) / wide_data$LO.PISA.REA.MA) * 100
wide_data$Science.Diff <- ((wide_data$LO.PISA.SCI.FE -
   wide_data$LO.PISA.SCI.MA) / wide_data$LO.PISA.SCI.MA) * 100
wide_data$Total.Score <- rowSums(wide_data[, c("LO.PISA.MAT",
   "LO.PISA.REA", "LO.PISA.SCI")], na.rm = TRUE)
wide_data$Avg.Diff <- rowMeans(wide_data[, c("Maths.Diff",
   "Reading.Diff", "Science.Diff")], na.rm = TRUE)
```

Snippet 5: Preparing Gender-Based Analysis of Student Performance

After the preparation of our data, it is time to create some plots so that we can visually analyze and compare the performance differences between male and female students across different countries. More precisely, these plots represent the % difference in scores using males as reference. A positive difference means females scored higher, while a negative difference means females scored lower. The dark red line represents the mean difference across all countries. We can observe these graphical representations in Figure 3 and 4.

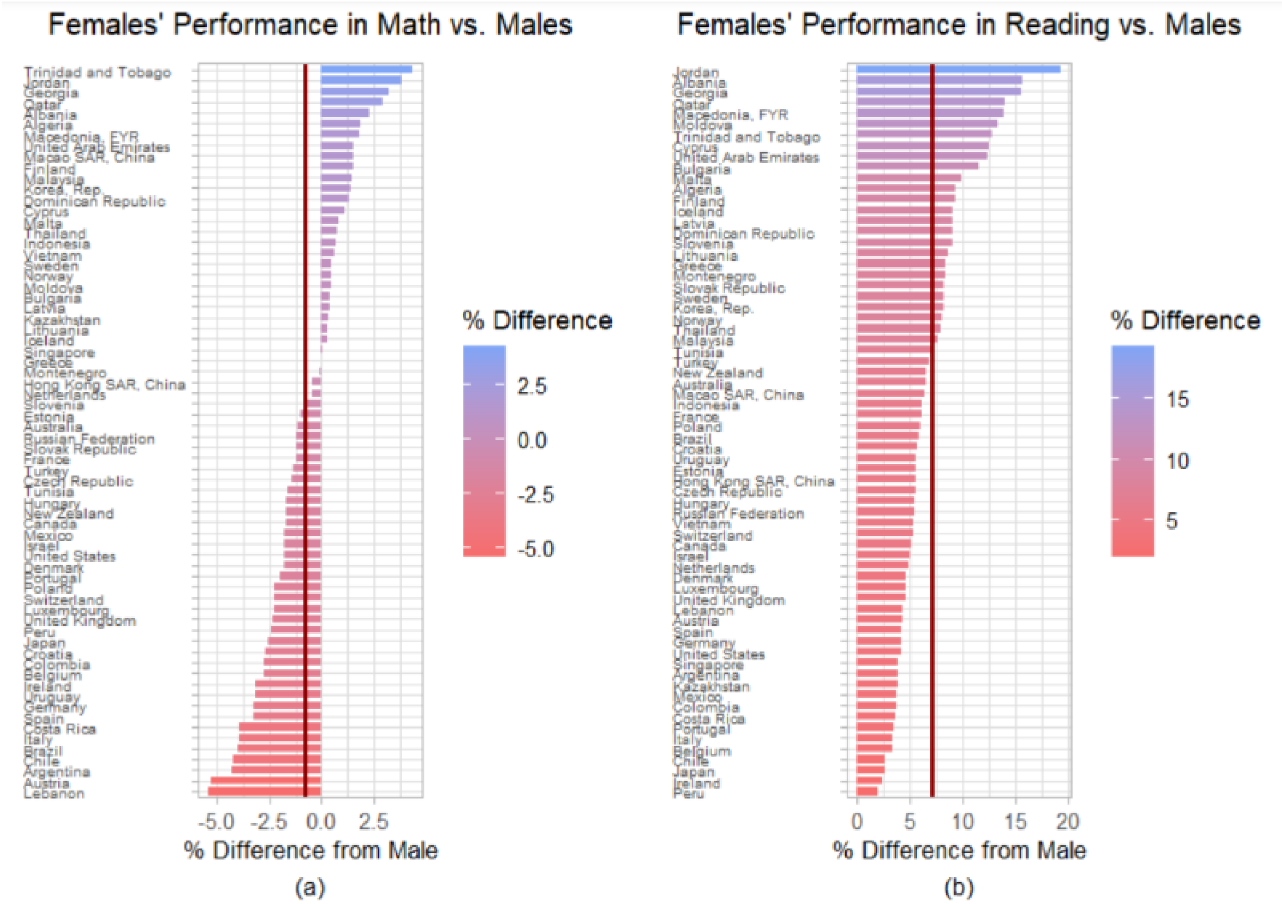

Figure 4: Males vs Females - Math / Reading

As we observe in barplot (a), approximately half of the countries assessed indicate that girls exhibit superior mathematical skills, while the remaining half demonstrates a tendency for boys to outperform their female counterparts. This balanced distribution of gender-specific strengths suggests a nuanced global landscape in mathematical achievement. Strikingly, in Trinidad and Tobago, female students exhibit a pronounced advantage in mathematical proficiency while in countries such as Austria and Lebanon, male students demonstrate a significant edge in math abilities. Notably, Greece emerges with a more balanced scenario, as the observed gender difference in this subject is deemed insignificant.

Diagram (b) reveals a consistent trend across all countries, indicating a notable proficiency in reading skills among females. In comparison to their male counterparts, females consistently achieved higher reading scores. This observation underscores a significant gender-related disparity in reading abilities, suggesting that females, on average, outperformed males in the assessed reading skills across diverse national contexts.

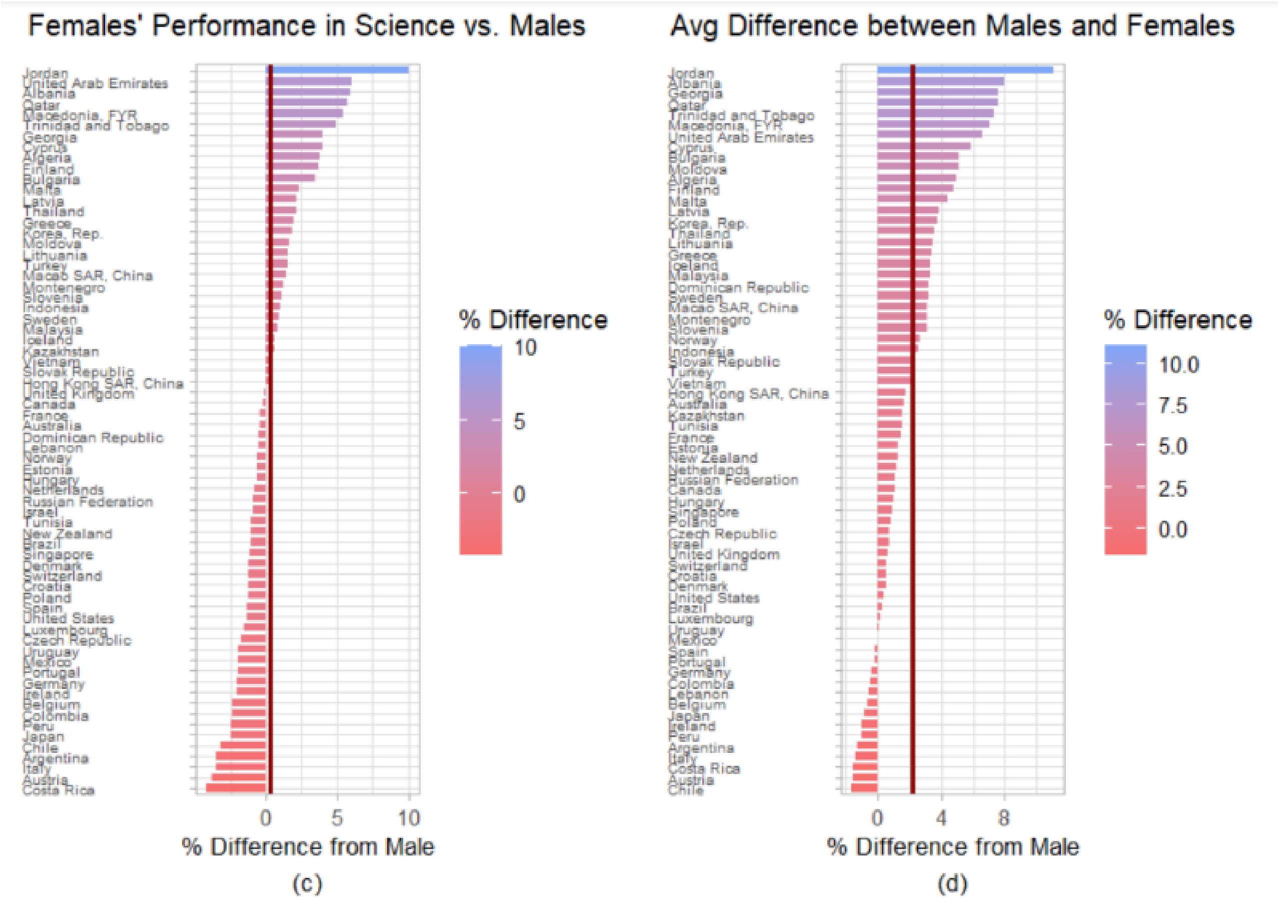

Figure 5: Males vs Females - Science / Average

The findings depicted in the third diagram, specifically focusing on science, mirror those observed in the initial analysis of mathematical performance. However, a distinctive feature is evident in the case of Greece, where female students exhibit a considerable advantage in science-related skills.

The bar plot (d), depicting the average difference between males and females across subjects, provides a compelling insight. As observed, girls consistently outperform boys in mean scores across the majority of countries. Although boys exhibit higher scores in mathematics and science, the overall trend suggests that girls, on average, surpass boys in academic achievements. In Greece, an intriguing observation emerges, indicating that girls outperform boys by approximately 3%. Nevertheless, the PISA scores we've examined are specific to the year 2015 and certain subjects, so drawing general conclusions about girls' or boys' academic superiority is limited. Educational assessments involve many factors, and these results provide only a snapshot of a larger and more complex picture.

In snippet 6, we can find the code for generating the four diagrams.

```
# How much better are females at math?
ggplot(data = wide_data, aes(x = reorder(`Country Name`,
   Maths.Diff), y = Maths.Diff)) + geom_bar(stat = "identity",
```

```
    aes(fill = Maths.Diff), width = 0.7) +
    scale_fill_gradient(name = "% Difference", low = "#F76C6C",
    high = "#7FA5F7") + coord_flip() + theme_light() +
    geom_hline(yintercept = mean(wide_data$Maths.Diff), size = 1,
    color = "red4") + labs(title = "Females’ Performance in Math
    vs. Males", x = "", y = "% Difference from Male", caption =
    "(a)") + theme(axis.text.y = element_text(size = 6, hjust =
    0), axis.title = element_text(size = 10), plot.title =
    element_text(size = 12, hjust = 0.5, margin = margin(b =
    10)), plot.caption = element_text(size = 10, hjust = 0.5))

# How much better are females at reading?
ggplot(data = wide_data, aes(x = reorder(‘Country Name‘,
    Reading.Diff), y = Reading.Diff)) + geom_bar(stat =
    "identity", aes(fill = Reading.Diff), width = 0.7) +
    scale_fill_gradient(name = "% Difference", low = "#F76C6C",
    high = "#7FA5F7") + coord_flip() + theme_light() +
    geom_hline(yintercept = mean(wide_data$Reading.Diff), size =
    1, color = "red4") + labs(title = "Females’ Performance in
    Reading vs. Males", x = "", y = "% Difference from Male",
    caption = "(b)") + theme(axis.text.y = element_text(size = 6,
    hjust = 0), axis.title = element_text(size = 10), plot.title
    = element_text(size = 12, hjust = 0.5, margin = margin(b =
    10)), plot.caption = element_text(size = 10, hjust = 0.5))

# How much better are females at science?
ggplot(data = wide_data, aes(x = reorder(‘Country Name‘,
    Science.Diff), y = Science.Diff)) + geom_bar(stat =
    "identity", aes(fill = Science.Diff), width = 0.7) +
    scale_fill_gradient(name = "% Difference", low = "#F76C6C",
    high = "#7FA5F7") + coord_flip() + theme_light() +
    geom_hline(yintercept = mean(wide_data$Science.Diff), size =
    1, color = "red4") + labs(title = "Females’ Performance in
    Science vs. Males", x = "", y = "% Difference from Male",
    caption = "(c)") + theme(axis.text.y = element_text(size = 6,
    hjust = 0), axis.title = element_text(size = 10), plot.title
    = element_text(size = 12, hjust = 0.5, margin = margin(b =
    10)), plot.caption = element_text(size = 10, hjust = 0.5))

# Average difference across all three subjects
ggplot(data = wide_data, aes(x = reorder(‘Country Name‘,
    Avg.Diff), y = Avg.Diff)) + geom_bar(stat = "identity",
    aes(fill = Avg.Diff), width = 0.7) + scale_fill_gradient(name
    = "% Difference", low = "#F76C6C", high = "#7FA5F7") +
    coord_flip() + theme_light() + geom_hline(yintercept =
    mean(wide_data$Avg.Diff), size = 1, color = "red4") +
    labs(title = "Average Difference between Males and Females
    across subjects", x = "", y = "% Difference from Male",
    caption = "(d)") + theme(axis.text.y = element_text(size = 6,
    hjust = 0), axis.title = element_text(size = 10), plot.title
    = element_text(size = 12, hjust = 0.5, margin = margin(b =
```

```
    10)), plot.caption = element_text(size = 10, hjust = 0.5))
```

Snippet 6: Males vs Females Diagrams

Following the analysis comparing the academic performance of male and female students, we will proceed to construct a scatter plot illustrating the relationship between the difference in scores among males and females and the total combined scores across mathematics, reading, and science disciplines.

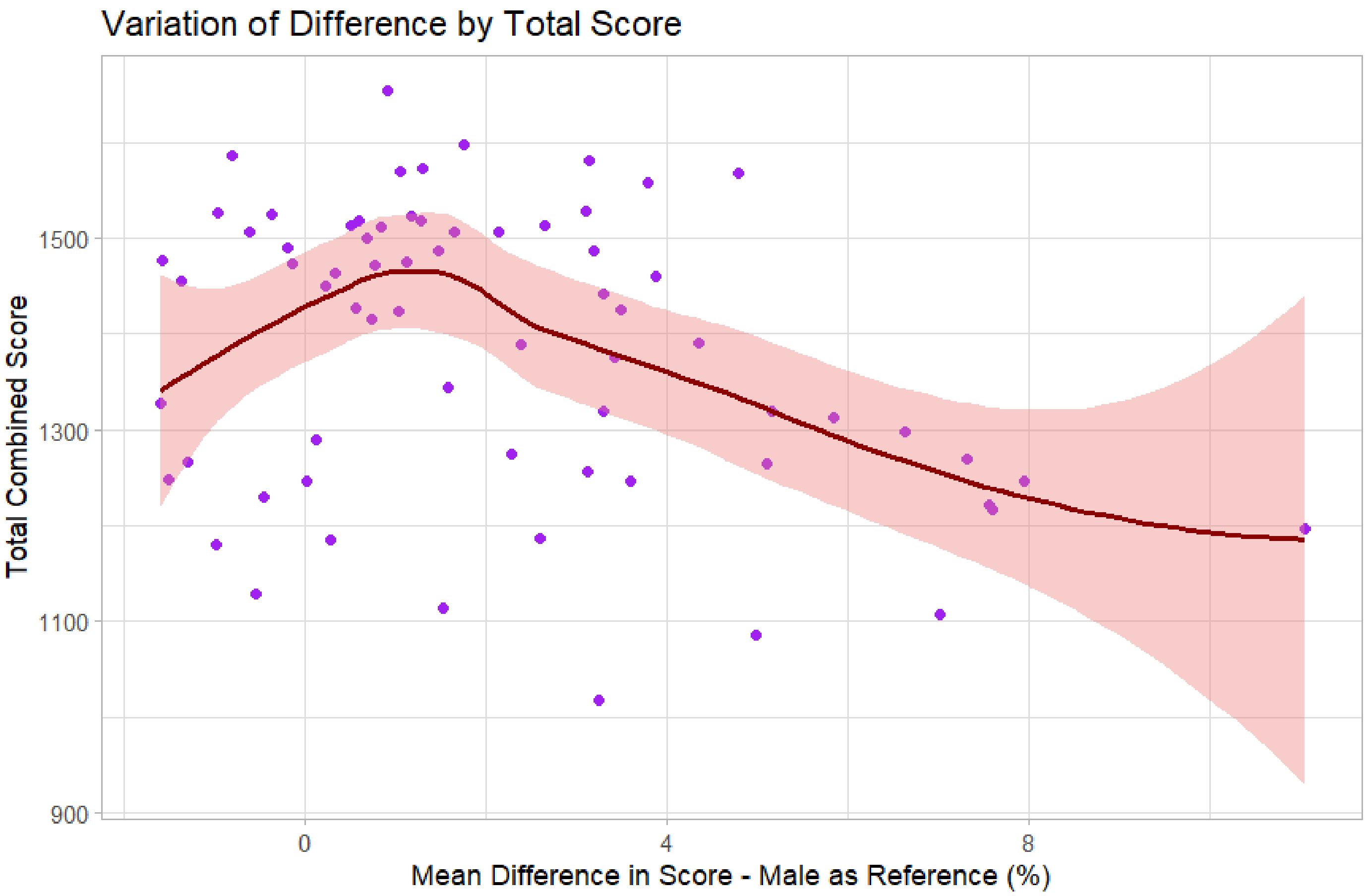

Figure 6: Gender Disparities vs. Total Scores Scatter Plot

The presented scatter plot visually illustrates the correlation between the difference in scores among males and females and the total combined scores across the three key subjects. Upon scrutinizing the plot, a discernible trend emerges, indicating that the disparity in scores tends to be smaller, approaching zero, in countries characterized by higher academic achievement. This intriguing pattern suggests that as countries exhibit superior performance across the three disciplines, the magnitude of score differences between males and females diminishes. While causation cannot be directly inferred from this observation, it prompts us to consider potential factors influencing this phenomenon. It is plausible that in countries with robust educational systems and higher overall performance, the emphasis on equal opportunities and inclusivity may contribute to a more balanced distribution of academic achievements between genders.

The R code used to generate the scatter plot in Figure 5 is provided in snippet 6.

```
plot(data=wide_data, aes(x=Avg.Diff, y=Total.Score)) +
   geom_point(color="purple") +
   geom_smooth(fill="lightcoral", colour="darkred", linetype=1) +
   theme_light() +
   labs(title="Variation of Difference by Total Score",
        x="Mean Difference in Score - Male as Reference (%)",
           y="Total Combined Score")
```

Snippet 7: Scatter Plot for Gender Disparities and Total Scores

4 Outlook

Taking all the above into serious consideration, we come to the conclusion that the comprehensive analysis of the PISA 2015 data has provided a multifaceted understanding of global educational landscapes, encompassing both national and gender dimensions. The top-performing countries, such as Singapore, Hong Kong, Japan, Macao, Estonia, underscore the impact of diverse educational strategies and societal priorities, with a focus on excellence, innovation, and supportive learning environments. In contrast, Greece's mid-ranking position raises questions about potential influences from economic challenges, resource constraints, or systemic issues within its educational framework.

The examination of gender-specific performance across mathematical, reading, and science skills revealed intriguing patterns. While gender differences in mathematical proficiency displayed a balanced global distribution, females consistently outperformed males in reading skills. Greece, notably, exhibited a balanced scenario in mathematics, while females showcased a considerable advantage in science-related skills. The average difference between genders across subjects indicated that, on average, girls tend to surpass boys in academic achievements.

Finally, the scatter plot we created revealed an interesting trend – in countries where students perform better overall, the gap in scores between boys and girls tends to be smaller. This suggests that strong education systems, emphasizing equal opportunities, may contribute to a more balanced performance between genders.

In conclusion, we believe that in the ever-evolving landscape of education, these insights open doors for further exploration, encouraging ongoing investigation into the multifaceted factors shaping academic achievement globally.