Contents

List of Figures

List of Tables

1 Introduction

The assignment focuses on the dataset of the Program for International Student Assessment (PISA) for the year 2015, comprising 1161 data entries from various countries. Our goal is to explore the impact of gender and country on the academic performance of 15-year-old students in reading, mathematics, and science. Through the use of R, specifically the ggplot2 and data.table libraries, we will employ exploratory data analysis techniques to uncover patterns and trends in the data. Ultimately, we aim to draw conclusions based on our findings.

2 Data Preprocessing

At first, we need to preprocess the given dataset in a way that is more useful for our analysis. For this reason, we remove the records that have missing Score values, since we cannot derive any conclusions from them. We also remove redundant columns, and from the remaining columns, we factorize the categorical ones, to make sure that they have the correct type.
The R code associated with the before mentioned, is the following:

```r
# load the necessary libraries
library(ggplot2)
library(data.table)
library(maps)

# create initial dataframe
PATH = 'C:/Users/milio/Desktop/edemm/programming_tools/R/ergasia/'
FILE_NAME = "Pisa mean perfomance scores 2015 Data.csv"
FILE_PATH = file.path(PATH,FILE_NAME)
df = fread(FILE_PATH, na.strings="..", header=TRUE)

# drop rows with NA values
df <- na.omit(df)

# delete unnecessary columns
col_to_delete = c('Country Name','Series Name')
df[, (col_to_delete) := NULL]

# rename columns
rename_col <- function(df,oldname,newname) {
  colnames(df)[colnames(df) == oldname] <- newname
  return (df)
}
df = rename_col(df,'2015','Score')
df = rename_col(df,'Country Code','Country')
df = rename_col(df,'Series Code','Code')

# create new columns
df[, 'Code' := substr(Code, 9, nchar(Code))]
df[, 'Subject' := substr(Code, 1, 1)]
df[, 'Gender' := substr(Code, 5, 5)]
df[, 'Code' := NULL]

# rearrange columns order, to get 'Score' at last place
df <- df[, c(setdiff(names(df), 'Score'), 'Score'), with = FALSE]

# convert columns to factors
df[, Gender := factor(ifelse(Gender == "", "B", Gender))] # B for both genders
df[, Subject := factor(Subject)]
df[, Country := factor(Country)]
```

We end up with a dataframe including 3 categorical columns: Country, Subject, Gender and 1 continuous variable (Score). The rows have been reduced from 1161 to 612, due to missing values in the Score column.

3 Data Analysis

In our analysis, we will split our focus into two primary sections. The first section will be dedicated to a comparative analysis of performance scores between genders. We will explore how male and female students' scores differ across various subjects and seek to identify any significant trends or patterns. In the second section, we will compare the performance scores across different countries. This will allow us to compare educational outcomes on a global scale and highlight any notable disparities between nations.

3.1 Gender Analysis

For the first part of our analysis, we are going to create boxplots, which will help us visualize potential differences between genders at the 3 subjects, followed by their stats. Finally, we will comment on how gender affects the performance score.

The code for plotting the boxplots and creating their statistics is the following:

```
both    <- subset(df, Gender == "B")
male    <- subset(df, Gender == "M")
female  <- subset(df, Gender == "F")
mf      <- subset(df, Gender == "M" | Gender == "F")

# Box plot for Gender Comparisons
boxplot = ggplot(mf)+
  aes(x=Gender, y=Score)+
  geom_boxplot(alpha=0.75) +
  aes(fill=Gender) +
  scale_fill_manual(values = c("pink", "lightblue")) +
  scale_x_discrete(labels = c("F"="Female", "M"="Male")) +
  theme_linedraw() +
  labs(x="Gender",y="Score") +
  facet_wrap(.~Subject, scales = "free_x", nrow=1, ncol=3 ,
             labeller = labeller(Subject = c("M"="Math", "R"="Reading", "S"="Science")))

# Function to calculate boxplot stats for each subject and gender and return as a data frame
calculate_stats <- function(data, gender, subject) {
  scores <- data$Score[data$Gender == gender & data$Subject == subject]
  stats <- fivenum(scores)
  mean_score <- mean(scores, na.rm = TRUE)
  stats_df <- data.table(
    Subject = subject,
    Gender = gender,
    Min = round(stats[1], 2),
    Q1 = round(stats[2], 2),
    Median = round(stats[3], 2),
    Mean = round(mean_score, 2),
    Q3 = round(stats[4], 2),
    Max = round(stats[5], 2)
  )
  return(stats_df)
}

all_stats <- data.table()
for (subject in c("M", "R", "S")){
  for (gender in c("F", "M")) {
    all_stats <- rbind(all_stats, calculate_stats(df, gender, subject))
  }
}
# Convert subject codes to full names
all_stats$Subject <- factor(all_stats$Subject, levels = c("M", "R", "S"),
                            labels = c("Math", "Reading", "Science"))
print(all_stats)
```

The boxplots of the above mentioned code are the following:

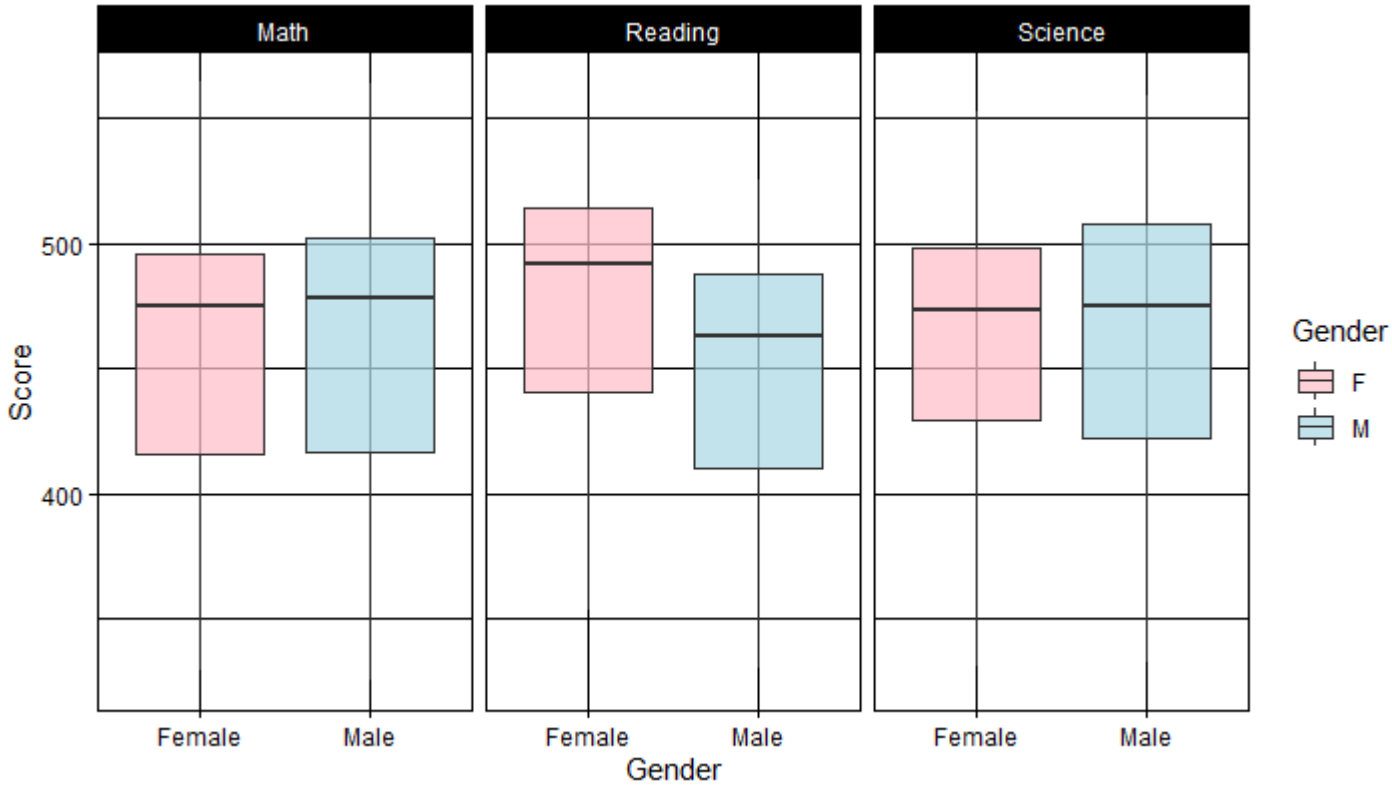

Figure 1: Boxplot of Scores by Gender for each Subject

The boxplot statistics are listed below:

Subject	Gender	Min	Q1	Median	Mean	Q3	Max
Math	F	329.75	415.09	475.30	458.13	495.73	564.25
Math	M	325.59	415.97	477.89	462.29	504.03	564.13
Reading	F	353.28	438.78	492.14	476.11	514.23	550.51
Reading	M	330.14	409.90	463.43	445.86	487.99	525.32
Science	F	330.83	429.05	473.62	465.62	498.60	552.27
Science	M	332.48	421.65	475.00	464.68	507.74	558.66

Table 1: Gender statistics

Based on figure (1) and table (1), we can make several observations about the performance scores by gender across three subjects: Math, Reading, and Science.

1. Math

   - Female students have a median score of 475.30, which is slightly lower than male students' median of 477.89. This suggests that male students are slightly outperforming female students at the median level.
   - The mean score for males is higher (462.29) compared to females (458.13), reinforcing that on average, males perform slightly better in Math.
   - The interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3), is slightly broader for males than females, indicating more variability in male students' Math scores.

2. Reading:

   - Female students show a higher median score of 492.14 compared to 463.43 for male students. This suggests that female students generally outperform male students in Reading.
   - The mean scores reflect this as well, with females averaging 476.11 and males 445.86, showing a bigger difference in Reading than in Math.

3. Science

   - The median scores for Science are quite close, with females at 473.62 and males at 475.00, suggesting a very similar performance.
   - The mean scores are also similar, with females at 465.62 and males at 464.68, indicating almost equal average performance in Science across genders.
   - Males have a slightly wider IQR, which could indicate a greater variability in performance among male students.

Based on our observations from the statistical table and boxplot, we conclude that:

- The distribution of scores in Math and Science is relatively similar between genders, with a slight edge for males in Math.
- In Reading, there is a big difference, with females having both higher median and mean scores and also a tighter IQR, suggesting that female performance is consistently higher in this subject.
- There are no obvious outliers in the boxplot, suggesting that most of the scores fall within a typical range.
- There are gender disparities in student performance across different subjects. While males slightly outperform females in Math, females show a significant lead in Reading, and both genders have nearly equal performance in Science.

3.2 Country Analysis

For the second part of our analysis, we focus on the impact that each country has on the students' scores. At first, we create boxplots for all 3 subjects, to detect possible outliers.

3.2.1 Part A

The used R code is the following:

```
# Create the boxplot
ggplot(both, aes(x = Subject, y = Score, fill = Subject)) +
  geom_boxplot() +  # Boxplot layer
  geom_jitter(width = 0.2, alpha = 0.5, color = "black") +
  scale_fill_brewer(palette="Pastel1") +
  labs(x = "Subject", y = "Score") +
  theme_minimal()
```

The above mentioned code creates this boxplot graph:

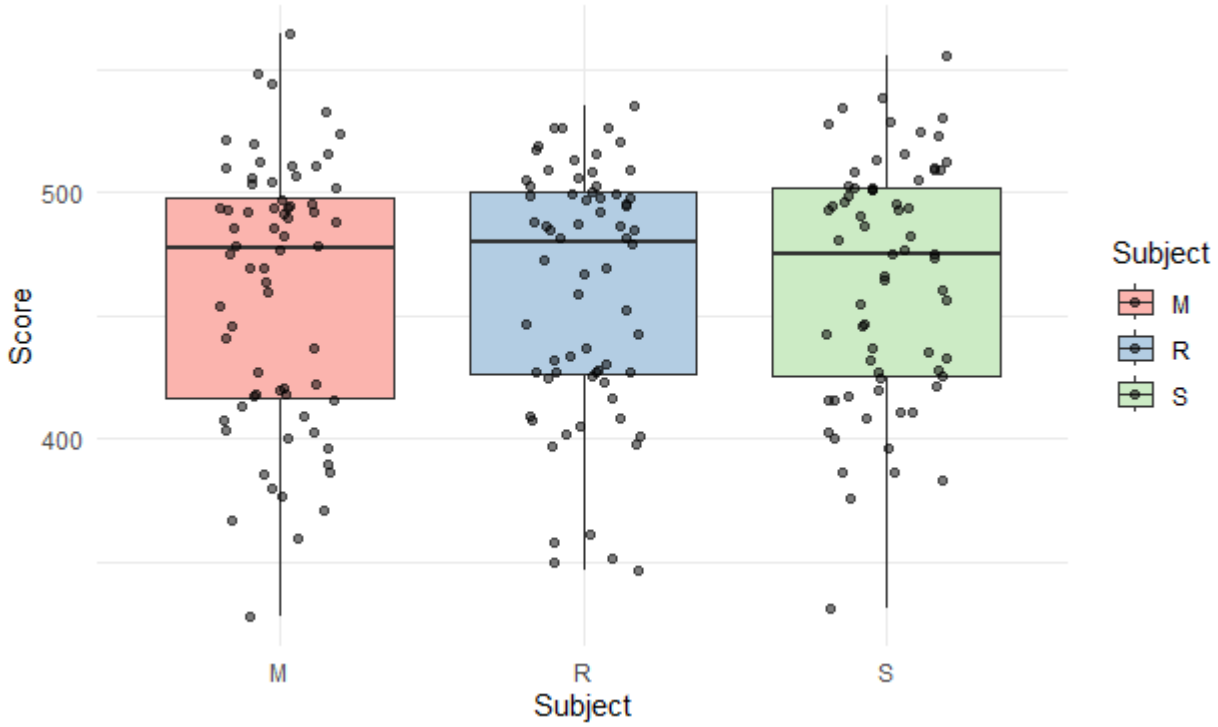

Figure 2: Boxplot of Scores by Subject

From figure (2), we can observe that:

- The range of scores within each subject is relatively broad, indicating variability in performance across the countries.
- The median scores for each subject, indicated by the line within each box, appear to be relatively close to each other.
- The IQR, represented by the height of each box, differs between subjects. Reading seems to have the smallest IQR, suggesting that scores are more clustered around the median, while Math and Science have a wider spread.
- There are multiple potential outliers for each subject, both above and below the boxes, depicted as individual points.

We can address the last observation using the known formula for boxplot elements x_i that:
$x_i \notin [Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR] \Rightarrow \ x_i$ is outlier
Calculating the intervals for every subject separately, we get:

Subject	Lower	Upper
Math	295.01	619.78
Reading	316.37	610.02
Science	309.95	617.46

Table 2: Outlier boundaries for the Subjects

After inspecting the country scores for values outside of the subject intervals, we conclude that there are no outliers in the data.

3.2.2 Part B

Following the outlier detection, we plot the performance scores of each subject for all countries, in ranked order, using barcharts that include a red line at the mean subject Score, with the following code:

```
# Barplots for Country comparisons
math    <- subset(df, Subject == "M" & Gender == "B")
reading <- subset(df, Subject == "R" & Gender == "B")
science <- subset(df, Subject == "S" & Gender == "B")

create_barplot = function(df, subj) {
  ggplot(df) +
  aes(x = reorder(Country, Score), y = Score, fill = Score) +
  geom_bar(stat = 'identity') +
  coord_flip() +
  scale_fill_gradient(name = "Score Level") +
  geom_hline(yintercept = mean(df$Score), size = 1, col = "red") +
  theme_grey() +
  labs(x = "Country Name", y = paste(subj, " Score"))
}

create_barplot(math, 'Math')
create_barplot(reading, 'Reading')
create_barplot(science, 'Science')

all_stats <- data.table()
all_stats <- rbind(all_stats, calculate_stats(math, 'B', 'M'))
all_stats <- rbind(all_stats, calculate_stats(reading, 'B', 'R'))
all_stats <- rbind(all_stats, calculate_stats(science, 'B', 'S'))
print(all_stats)
```

Below is a bar plot that illustrates the performance scores for the math subject:

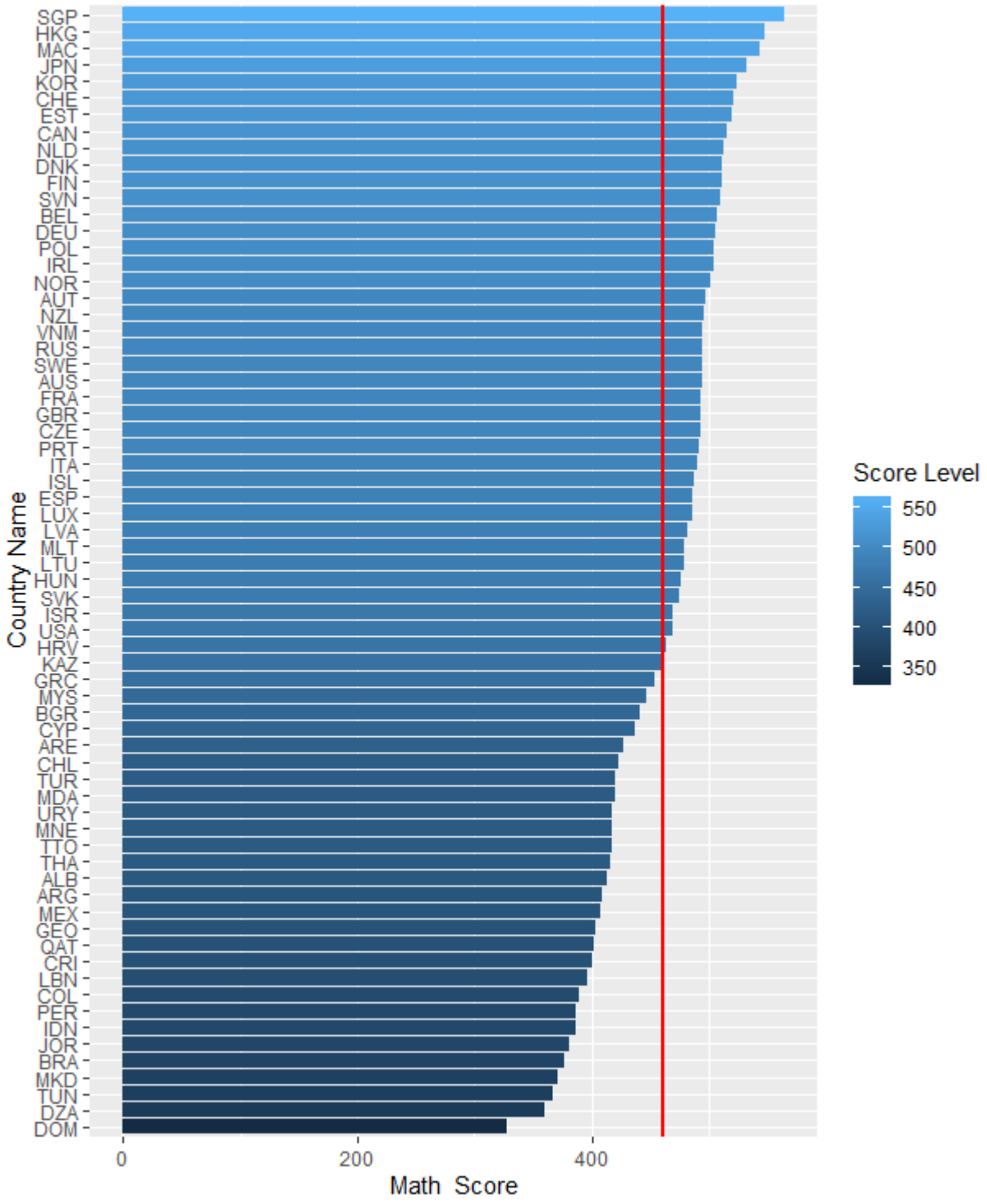

Figure 3: Barplot of Math Scores by Country

The math scores statistics are listed below:

Min	Q1	Median	Mean	Q3	Max
327.7	416.35	477.61	460.19	499.24	564.19

Table 3: Math scores statistics

From figure (3), table (3) we can observe that:

- Mean score for math performance as shown by the red line is 460.19, while median score is 477.61, indicating that countries with low scores pull the average down.

- The top performers are: Singapore (SGP) leads with the highest score (564.19), followed by Hong Kong (HKG) and Japan (JPN).
- There is considerable variability in scores, with the Dominican Republic (DOM) having the lowest score (327.70) listed in the plot, which is significantly below the scores of the top-performing countries.

Conclusions

- There is a wide gap in math performance between the best and the worst performance countries in the graph.
- The data suggests geographic patterns in scores, where East Asian countries are generally high performers, and certain countries in the Middle East, Latin America, and North Africa are among the lower scorers.
- While some European countries are high performers, others like Albania (ALB) and Macedonia (MKD) have scores on the lower end, highlighting educational disparities within the continent.
- The math score of a country may be correlated with it's economic situation, as many of the higher-scoring countries also have higher GDP per capita, which could suggest a link between economic resources and educational outcomes.

Below is a bar plot that illustrates the performance scores for the reading subject:

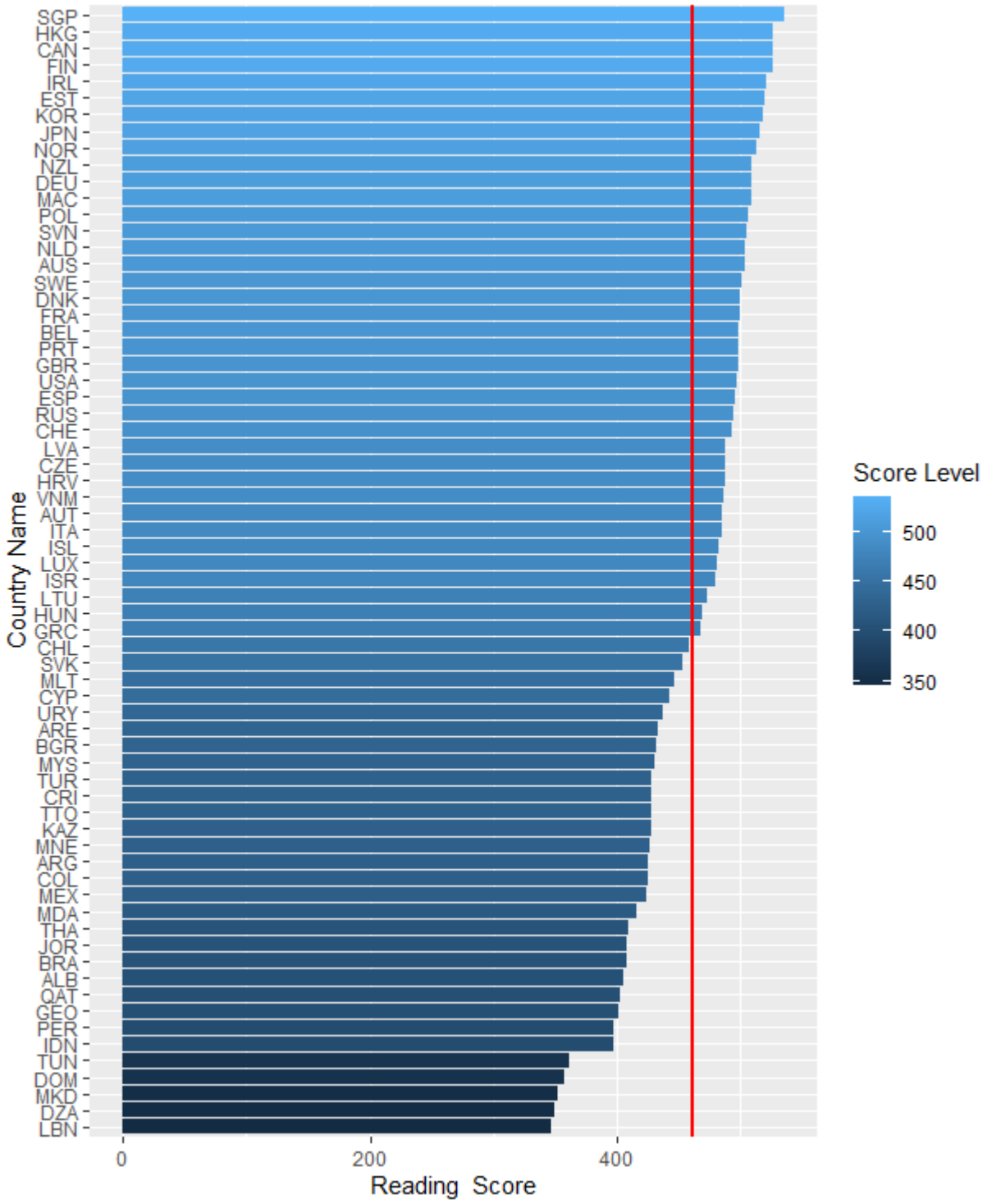

Figure 4: Barplot of Reading Scores by Country

The reading scores statistics are listed below:

Min	Q1	Median	Mean	Q3	Max
346.55	426.09	480.2	460.92	499.99	535.10

Table 4: Reading scores statistics

From figure (4), table (4) we can observe that:

- The scores span from a minimum of 346.55 to a maximum of 535.10, indicating a broad range of reading abilities among the countries. There's a substantial gap of nearly 190 points between the highest and

lowest scores.

- Singapore (SGP), with the highest score of 535.10, Hong Kong (HKG), and Canada (CAN) are the top performers, suggesting that students in these countries have a higher proficiency in reading compared to their fellow-students globally.
- Mean score for math performance as shown by the red line is 460.92, while median score is 480.2, a small number of countries with low reading scores are affecting the average more than those at the high end.

Conclusions

- The top performing region is, again, East Asia, including countries, like Singapore (SGP), Hong Kong (HKG), South Korea (KOR), Japan (JPN) and Macau (MAC).
- There's a group of European countries significantly above the global average, including countries like Finland (FIN), Estonia (EST), and Germany (DEU), which are known for their strong educational systems.
- Many of the higher-scoring countries have strong economies, which might afford them better educational resources.

Below is a bar plot that illustrates the performance scores for the science subject:

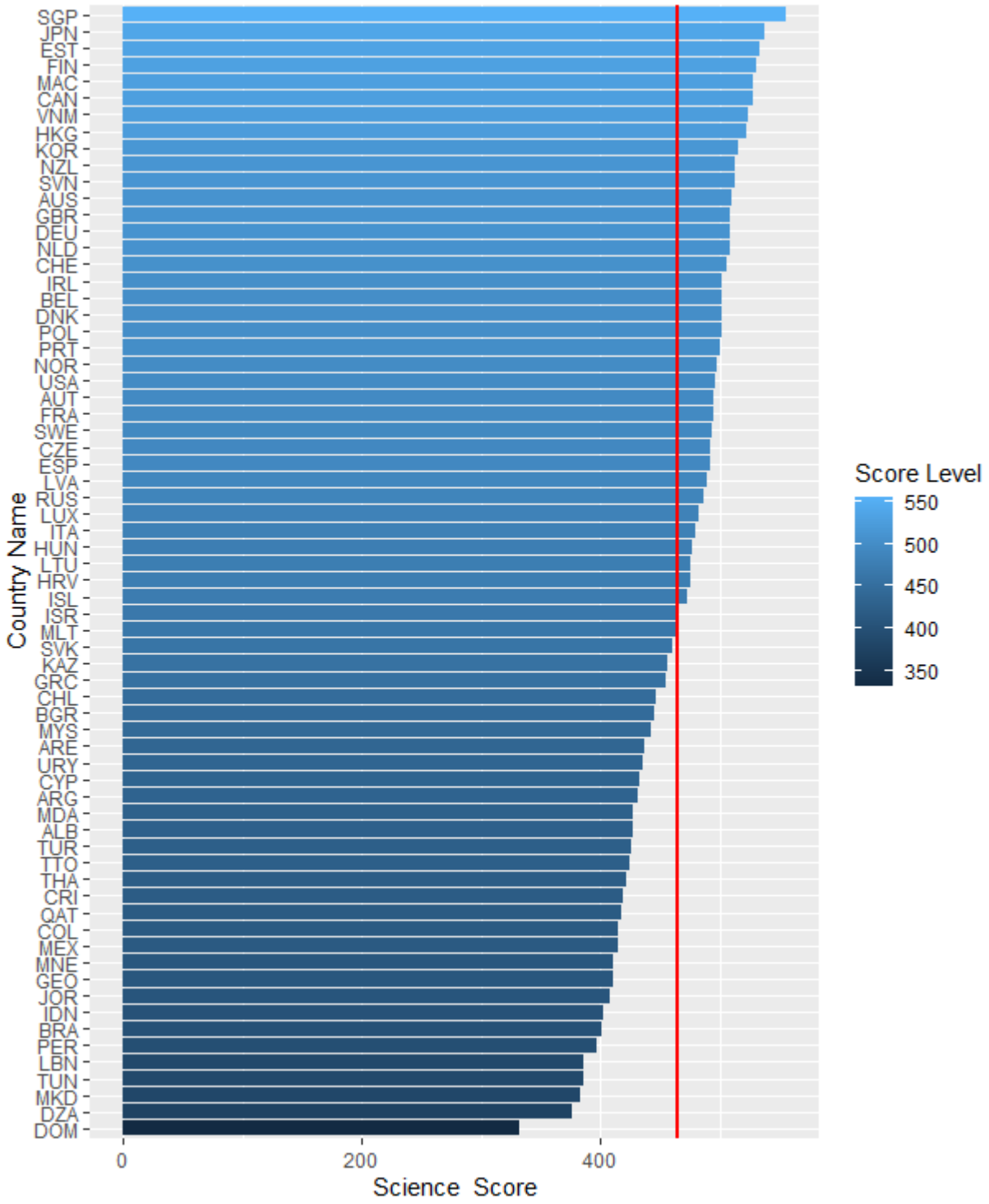

Figure 5: Barplot of Science Scores by Country

The science scores statistics are listed below:

Min	Q1	Median	Mean	Q3	Max
331.64	425.04	475.40	465.12	502.29	555.57

Table 5: Science scores statistics

From figure (5), table (5), we can observe that:

- Singapore (SGP), with the highest score of 555.57, leads the chart, followed by countries like Japan (JPN) and Finland (FIN), which also have high scores.

- The range of scores from 331.64 to 555.57 indicates a significant disparity in science proficiency across the countries.
- The distribution of scores is skewed towards the lower end, as indicated by the mean score (465.12) being lower than the median (475.40).

Conclusions

- Generally, countries with higher GDP per capita tend to have higher scores, although there are exceptions.
- The countries leading the scores likely invest significantly in education, with a particular focus on science.
- Countries from East Asia, such as Singapore (SGP), Japan (JPN), and Hong Kong (HKG), are at the top of the science score rankings. This reflects a consistent trend where East Asian education systems achieve high performance.

Furthermore, we will showcase a bar chart that aggregates the scores across the three subjects. This visualization will be used for drawing conclusions about the overall patterns in the performance data.
The code in R used is the following:

```
# Aggregated bar plot
aggregate_scores <- aggregate(Score ~ Country, data = df, FUN = sum)
aggregate_scores$Gender <- 'B'
aggregate_scores$Subject <- 'All'
create_barplot(aggregate_scores, '')
all_stats <- calculate_stats(aggregate_scores,'B','All')
print(all_stats)
```

The aggregated scores statistics are listed below:

Min	Q1	Median	Mean	Q3	Max
3050.67	3780.59	4278.95	4158.92	4528.69	4965.06

Table 6: Aggregated scores statistics

Below is a bar plot that illustrates the total performance scores for the 3 subjects:

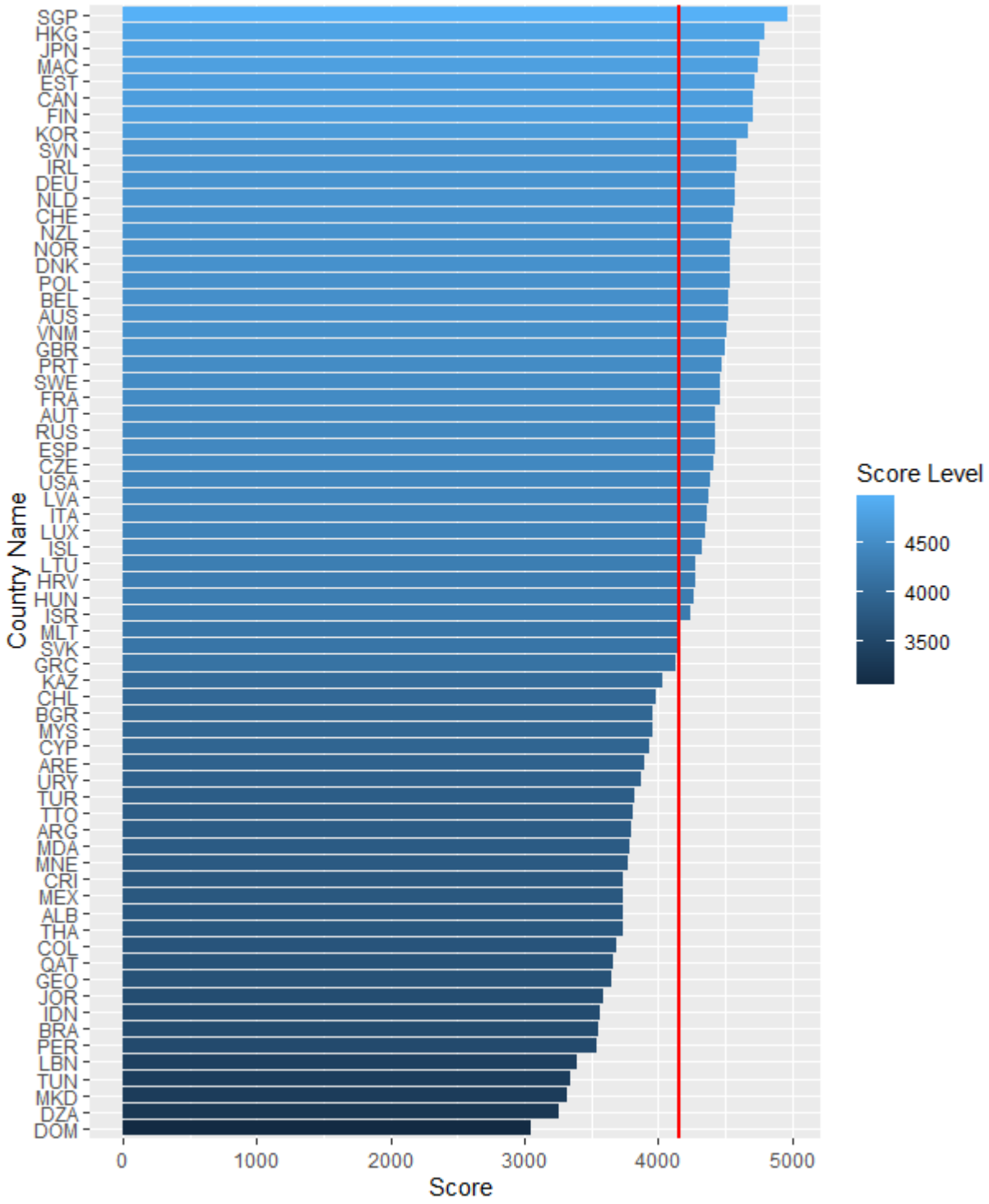

Figure 6: Aggregated barplot of all subjects

From figure (6), table (6) we draw the following conclusions:

- East Asia is the most dominant region overall. This suggests that there may be underlying educational philosophies or cultural attitudes that prioritize academic excellence.
- Central and Northern Europe showcanse strong educational outcomes, which is likely due to strong educational systems or cultural values that emphasize the importance of academic achievement.
- There appears to be a correlation between the economic development of a country and its educational performance. Developed countries, which often have more resources to invest in education, tend to score

higher on the PISA assessments. For instance, countries with higher GDPs like Canada (CAN) and Australia (AUS) show higher scores, indicating that economic growth may be linked to better education.

- Developing countries, particularly those in regions like Latin America, North Africa, and some parts of Asia, are represented in the lower end of the score spectrum. Countries like the Dominican Republic (DOM), Algeria (DZA), and Peru (PER) have aggregated scores on the lower side.

References

[1] Fouskakis Dimitris: *Data Tables & ggplot2 visualizations*, Programming Tools & Technologies for Data Science, NTUA, 2023-24, available at: `http://www.math.ntua.gr/~fouskakis/Programming_R/progr_slides.html`.

[2] LaTeX Documentation, available at: `https://www.overleaf.com/learn`.