Programming Tools for Data Science

Exploratory Data Analysis using R

Lazaros Athanasiadis
MSc student – Data Science and Machine Learning
Student Number: 03400201
email: lazarosathanassiadis@mail.ntua.gr

School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece

January 24, 2024

1 Introduction

The Programme for International Student Assessment (PISA) is a global study conducted every three years. It is designed to evaluate the performance of 15-year-olds worldwide in mathematics, reading and science, focusing on skills that will aid them in lifelong learning and their full participation in society [3]. For each participating country, the published results dataset contains the students' mean score in each discipline, as well as the mean score for the male and female students, resulting in a total of nine data points for each country. The objective of this Project is to perform exploratory data analysis on the 2015 PISA study results using the R programming language, with an emphasis on the `ggplot2` and `data.table` libraries. The investigation aims to determine whether three main factors (gender, country and region) affect student performance. The following Chapters are organized as follows:

- In Chapter 2, I outline the dataset preprocessing steps that facilitated my analysis.
- In Chapter 3, I examine the student performance in each country.
- In Chapter 4, I investigate whether countries in the same region perform similarly.
- In Chapter 5, I analyze the results of male and female students separately and check if one gender outperforms the other.
- Finally, in Chapter 6, I provide concluding remarks.

2 Data Preprocessing

As with most data analysis tasks, the dataset must be cleaned and transformed so that working with it is easy. I start by reading the dataset with `fread`, adding better header names, dropping the column that explains the test code and looking into the NA values.

```r
library(data.table)
library(ggplot2)
library(countrycode)
library(maps)

DT <- fread("pisa2015.csv", header = TRUE, na.strings = "..")
names(DT) <- c("Country", "CountryCode", "Test", "TestCode", "Score")
DT[, Test:=NULL]

has_na <- subset(DT, is.na(Score) == TRUE)
has_na <- has_na[, .N != 9, CountryCode]
has_na[V1 != FALSE]
```

Listing 1: Reading the dataset plus some basic preprocessing

The last expression, which shows how many countries have NA values in some (but not all) of the rows corresponding to them, returns an empty data table. This means that there is no need to specifically examine countries with partial data, because no such countries exist in the dataset. Moving on, I remove every row that has a NA value in the "Score" column (called 2015 in the original dataset).

```r
DT <- na.omit(DT)
```

Listing 2: Dropping rows with NA values

2.1 Adding Region to the Dataset

One of the tasks of this Project is to look into whether a country's region plays a role in its students' performance in the PISA tests. However, the provided dataset does not include any columns related to that purpose. For that reason, I used the `countrycode` library [1], which can convert country codes to names and vice-versa. It can also provide each country's region and EU membership status, functionalities I used in order to augment the dataset with two extra columns, `Region` and `EU`.

```r
DT$Region <- countrycode(DT$CountryCode, "iso3c", "region")
# following hack is needed because countrycode returns "EU" or NA instead of true or false
DT$EU <- FALSE
DT[!is.na(countrycode(CountryCode, "iso3c", "eu28"))]$EU <- TRUE
```

Listing 3: Adding Region and EU columns to the dataset

Region, in the context of `countrycode`, refers to the seven world regions as defined by the World Bank [2]. These are:

- North America
- Latin America & Caribbean
- Europe & Central Asia
- Middle East & North Africa
- East Asia & Pacific
- Sub-Saharan Africa
- South Asia

With the Region column added to the data table, it is easy to list how many countries participated in the study from each region. Then, using a list of every country from the `countrycode` package, I calculated the participation ratio for each one [1].

```r
# Create new col with how many countries participated from each region
region_cnt <- DT[, .(Participating = .N/9), Region]
# Get a count of how many countries are in each region
# using codelist, a dataframe provided by the countrycode package
all_regions <- data.table(Region = na.omit(codelist$region))
all_regions_cnt <- all_regions[, .(All = .N), Region]
# Join
region_cnt <- region_cnt[all_regions_cnt, on=.(Region)]
# replace NAs with 0
region_cnt[is.na(Participating)]$Participating <- 0
region_cnt <- region_cnt[, .(ParticipatingRatio = round(Participating/All,2)), Region]
region_cnt[order(-ParticipatingRatio)]
```

Listing 4: Calculating the perticipation percentage for each region

[1] `countrycode` is not limited to sovereign countries, but also contains territories that have some autonomy. For example, the French islands of Saint Pierre and Miquelon, near the coast of Canada, are listed as a territory in the North American region. This means that the computed ratio is the participating territories ratio, not the participating countries one. However, it will be a useful approximate measure of which regions are under-represented.

Region	Ratio of Participating Countries
North America	0.50
Europe & Central Asia	0.42
Middle East & North Africa	0.33
East Asia & Pacific	0.23
Latin America & Caribbean	0.19
South Asia	0
Sub-Saharan Afric	0

Table 1: Output from Listing 4

Looking at Table 1, it is apparent that the study is dominated by western countries, while there are regions from which no country participated. This bias should be kept in mind while arriving at conclusions using this data. Also, we should be careful when attempting to generalize any conclusion for the entire world.

2.2 Test Code Variables

For each test code, I defined a literal string variable in order for the code to be more readable and reusable, if the test codes for a later PISA study change.

```
MATH <- "LO.PISA.MAT"
MATH_M <- "LO.PISA.MAT.MA"
MATH_F <- "LO.PISA.MAT.FE"
READ <- "LO.PISA.REA"
READ_M <- "LO.PISA.REA.MA"
READ_F <- "LO.PISA.REA.FE"
SCI <- "LO.PISA.SCI"
SCI_M <- "LO.PISA.SCI.MA"
SCI_F <- "LO.PISA.SCI.FE"
```

Listing 5: String literals for each test code

2.3 Filtered Datasets

The provided form of the dataset is not the optimal one, because for each country, the information of some data points (the mean scores of each gender at each discipline) is also contained in the rest of the data points (the mean score overall at each discipline). For this reason, I created two new data tables: `mean_scores`, which contains the mean scores for each country at each discipline, and `dt_gendered`, which contains the mean scores for each country and gender at each discipline. In order to achieve the best synergy with `ggplot`'s facet commands, the second dataset contains a `Gender` column, instead of relying on a different test code for each gender.

```
mean_scores <- DT[TestCode == MATH | TestCode == READ | TestCode == SCI]

dt_gendered <- DT[TestCode != SCI & TestCode != MATH & TestCode != READ]
dt_gendered$Gender <- "Male"
dt_gendered[TestCode == MATH_F
            | TestCode == READ_F
            | TestCode == SCI_F]$Gender <- "Female"
# replace gender specific test codes with the generic ones
```

```
dt_gendered[startsWith(TestCode, MATH)]$TestCode <- MATH
dt_gendered[startsWith(TestCode, READ)]$TestCode <- READ
dt_gendered[startsWith(TestCode, SCI)]$TestCode <- SCI
```

Listing 6: Adding Region and EU columns to the dataset

3 Variation Between Countries

I start my analysis by examining the variation between countries. I settled on a world map plot depicting the mean score across all disciplines for each country as a quick way to do that. It will also highlight which regions are not well represented in the study.

First, I created the `one_mean` data table, containing the mean performance of each country. Afterwards, I loaded the world map data using the `map_data` function from `ggplot2`, extracted a data table of all countries, then joined it with `one_mean`, after changing some countries' names so that they would match.

```
one_mean <- mean_scores[, .(Score = mean(Score)), Country]

# Change some country names to match with the ones returned by map_data
change_name <- function(DT, old, new){
    DT[Country == old, "Country" := new]
}
updates <- list(c("Russian Federation", "Russia"),
                c("United States", "USA"),
                c("United Kingdom", "UK"),
                c("Macedonia, FYR", "North Macedonia"),
                c("Korea, Rep.", "South Korea"))
for (pair in updates){
    change_name(one_mean, pair[1], pair[2])
}

# Load map data, get a data table of all countries and join it with one_mean
wmap <- setDT(map_data("world", wrap = c(-180, 180)))
all_countries <- data.table(Country = unique(wmap$region))
one_mean <- one_mean[all_countries, on = "Country"]
```

Listing 7: Creating a data table with NA values for countries that did not participate in the study

I also binned the mean scores, in order to have a clearer plot. Then, I plotted the world map, coloring each country according to the bin their mean performance is in, adding thin, black borders to each country and removing the background grid.

```
# Binning
discrete_values <- seq(300, 550, 50)
one_mean$Discrete = cut(one_mean$Score, breaks = discrete_values)
# Preparing the legend
legend <- c()
for (i in 1:(length(discrete_values) - 2)){
    legend <- append(legend, paste(discrete_values[i], "-",
                     discrete_values[i+1]))
}
legend <- append(legend, c("> 500", "No Data"))
# Plot
ggplot(one_mean, aes(map_id = Country)) +
    geom_map(aes(fill = Discrete), map = wmap, color = "black", linewidth = 0.05) +
    expand_limits(x = wmap$long, y = wmap$lat) +
    scale_fill_brewer(palette = "RdYlBu", na.value = "lightgrey",
                      name = "Mean Exam Score", labels = legend) +
    coord_quickmap() +
    theme_void() +
    theme(legend.position = "bottom")
```

Listing 8: Generating the world map plot of Figure 1

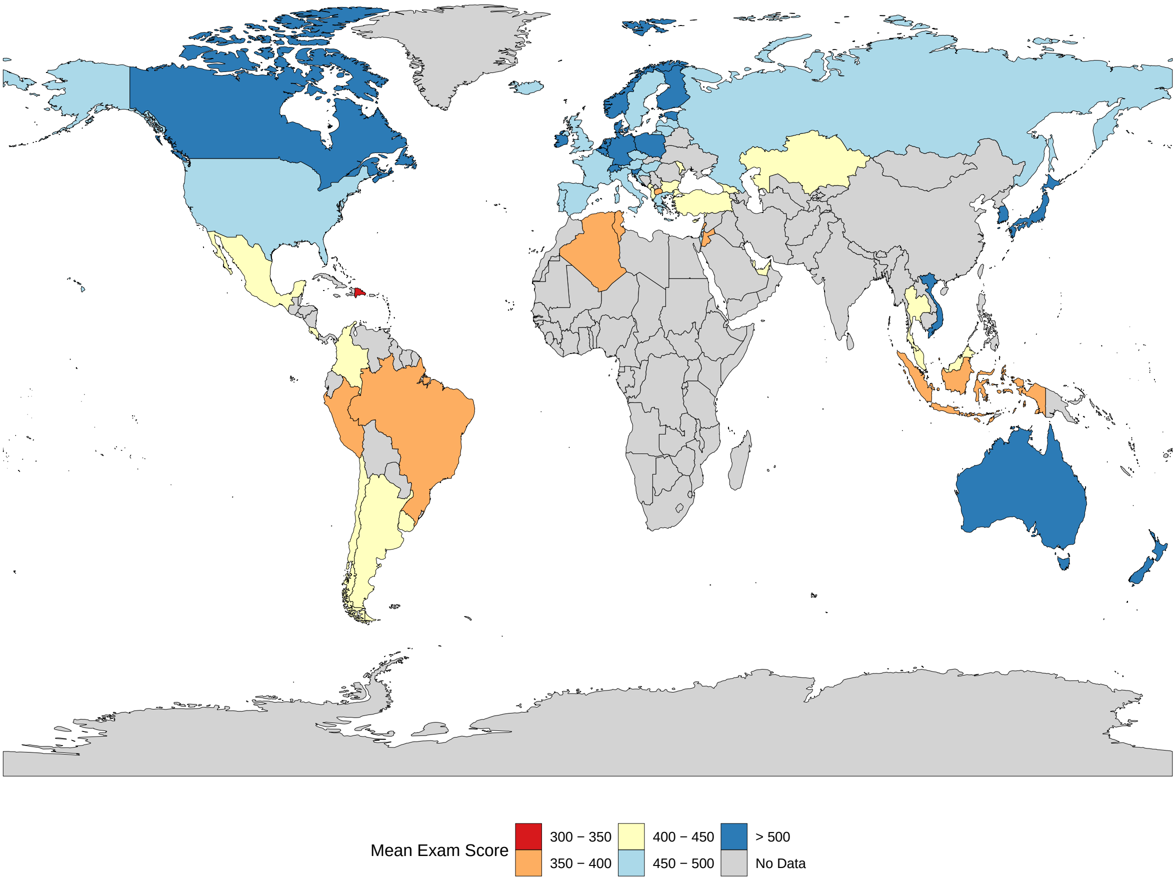

Figure 1: Mean performance of each country in the 2015 PISA study

From Figure 1, we see that the performance of each country varies greatly. It is also easy to notice the regions that are not represented at all (Sub-Saharan Africa and South Asia) and the ones that are underrepresented, such as Eastern Europe and the Middle East. Furthermore, a correlation between the region of a country emerges. For example, countries in Latin America performed worse than most European countries.

I conclude this section by plotting the mean performance in a histogram, for a more succinct visualization of how many countries belong to each bin, as defined in the previous figure.

```
ggplot(na.omit(one_mean)) +
    geom_histogram(aes(x = Score, fill=Discrete), binwidth = 10, color = "black") +
    scale_y_continuous(breaks = seq(0,20,2), labels=seq(0,20,2)) +
    scale_fill_brewer(palette = "RdYlBu") +
    theme(legend.position = "none") +
    labs(x = "Mean Exam Score", y = "Count")
```

Listing 9: Generating the histogram of Figure 2

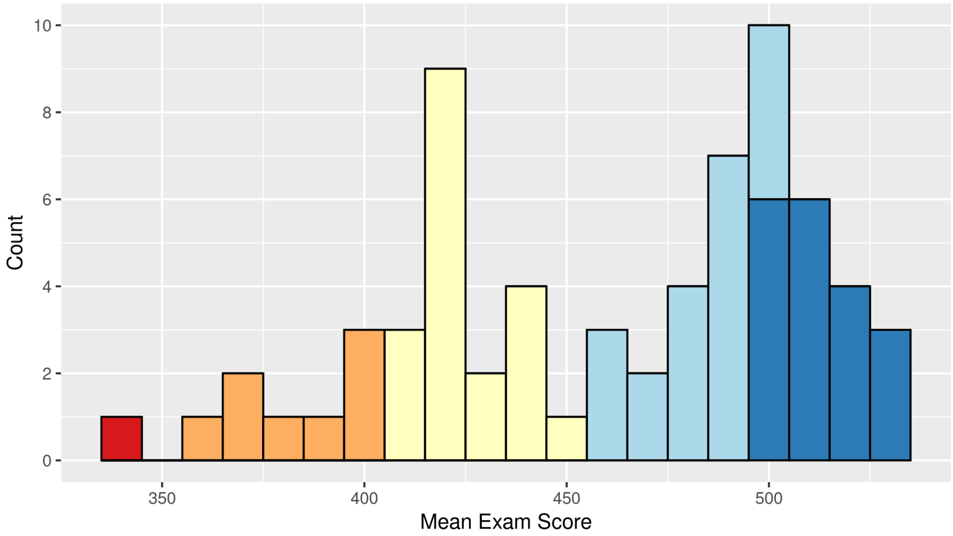

Figure 2: Histogram of mean performance of each country, using bin colors from Figure 1

Figure 2 brings to our attention the fact that most countries in the $450 - 500$ bin have mean performance closer to 500 than to 450. It also hints towards a bimodal underlying density, so I plotted a density histogram to get a better look.

```
ggplot(na.omit(plot_dt)) +
    geom_histogram(aes(x = Score, y = after_stat(density)), bins = 20,
                   color = "deepskyblue", fill = "white") +
    geom_density(aes(x = Score), color = "mediumblue") +
    labs(x = "Mean Exam Score", y = "Density")
```

Listing 10: Generating Figure 3

In Figure 3, the density of the mean exam score resembles a mixture of two normal distributions. This finding can be interpreted in a number of ways. If we split countries in two categories ("developed" and "developing"), then we would expect countries in each category to perform similarly. Then, each normal distribution would represent the mean performance of each class of countries. However, if we treat the development of a country in a more continuous manner, then the small amount of mean scores between the two peaks could mean that once some initial hurdles are overpassed, a country's education system will quickly progress. It could also be a methodological error of the study. If a number of exercises were to be solved in a related way, then most students would either solve either all of them or none. This would result in scores that are not equally distributed, which would explain this finding.

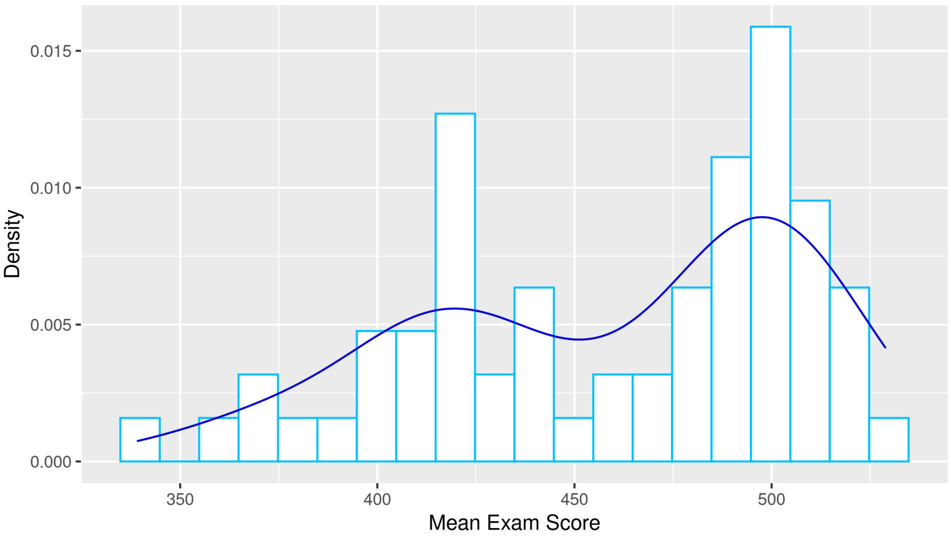

Figure 3: Density histogram for the mean exam score, with added density curve

4 Performance of Regions

In this Chapter, I examine the correlation between the region of a country and its performance in greater detail. To accomplish this, I graphed box plots for each discipline, faceted by region. This way, I will find out the best and worst performing regions, and I will investigate whether the mean scores for each discipline are similar, or if there exist some regions that perform noticeably better in one discipline compared to the rest.

```r
test_labels <- c("Math", "Reading", "Science")
names(test_labels) <- c(MATH, READ, SCI)

ggplot(mean_scores) +
    geom_boxplot(aes(x = TestCode, y = Score)) +
    facet_wrap(facets = vars(Region)) +
    scale_x_discrete(labels = test_labels) +
    labs(x = "Tested Discipline", y = "Score")
```

Listing 11: Generating Figure 4

By observing Figure 4, it is evident that the mean score for every discipline is similar in each region. In some cases (e.g. North America, Latin America & Caribbean), students performed slightly worse in Mathematics. Also, we can quickly observe that North America is the best performing region, while Latin America & Caribbean and Middle East & North Africa are the worst performing ones.

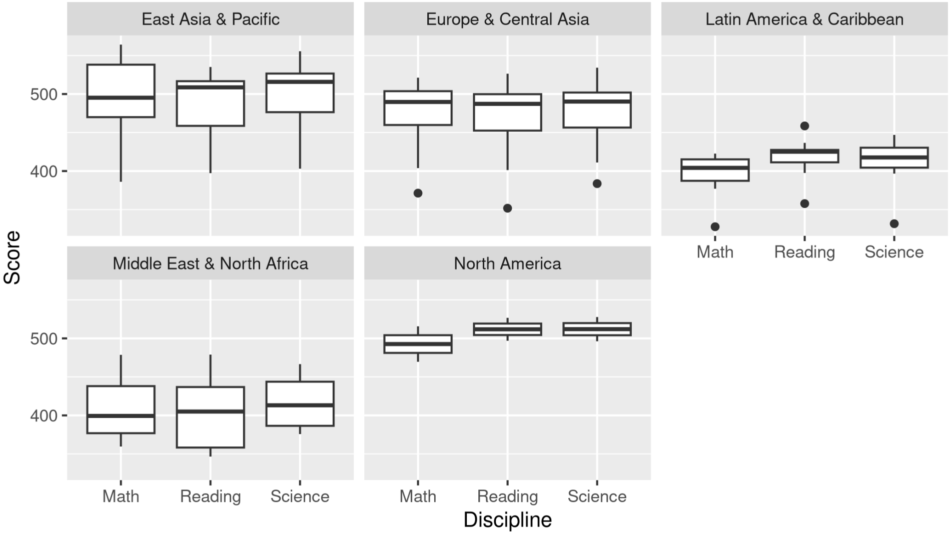

Figure 4: Performance of each region at every discipline

4.1 Performance of the European Union

Using the `EU` column that I have added in the dataset, it is also possible to plot the performance of countries in the European Union compared to the rest of the world. In Figure 5 we see that students in EU countries performed better than the ones in the rest of the world, as one could have noticed from Figure 1.

```r
# custom labels instead of true or false
facet_labels <- c(`TRUE` = "EU Members", `FALSE` = "Rest of the World")

ggplot(mean_scores) +
    geom_boxplot(aes(x = TestCode, y = Score)) +
    # change to factor in order to fix the order of the facets
    facet_wrap(~factor(EU, levels = c(TRUE, FALSE)),
               labeller = as_labeller(facet_labels)) +
    scale_x_discrete(labels = test_labels) +
    labs(x = "Discipline", y = "Score")
```

Listing 12: Generating Figure 5

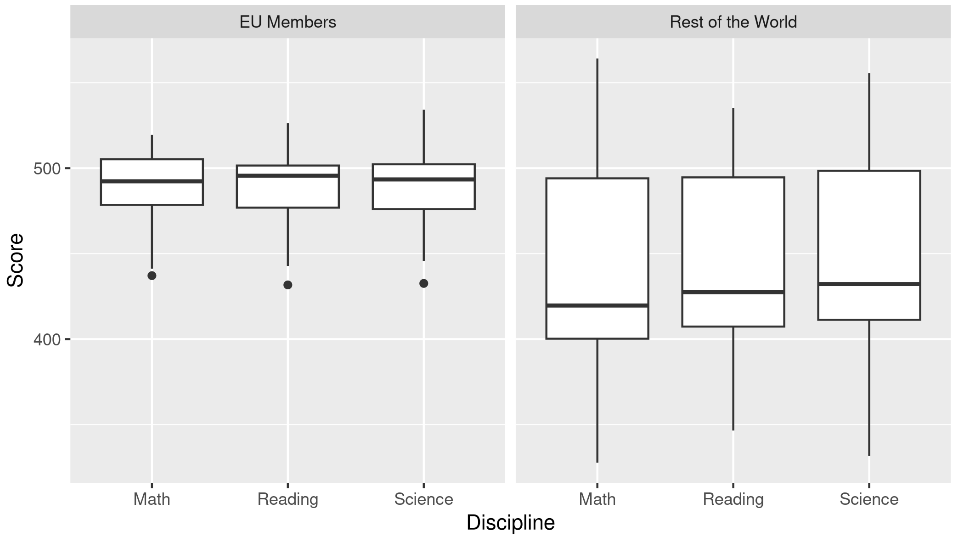

Figure 5: Comparing the performance of the EU vs the rest of the world

5 The Effect of Gender

After examining how their country and region affects students' performance, I will compare the scores of male and female ones separately, in order to examine whether one gender performs consistently better in some disciplines. I do this with a manner similar to the one in the previous Chapter, creating a box plot for each discipline-gender pair, coloring them according to gender. After examining Figure 6 it appears that girls perform better, on average, than boys on Reading, while their scores for Mathematics and Science are more or less comparable.

```
ggplot(dt_gendered) +
    geom_boxplot(aes(x = TestCode, y = Score, color = Gender)) +
    scale_x_discrete(labels = test_labels) +
    labs(x = "Discipline", y = "Score")
```

Listing 13: Generating Figure 6

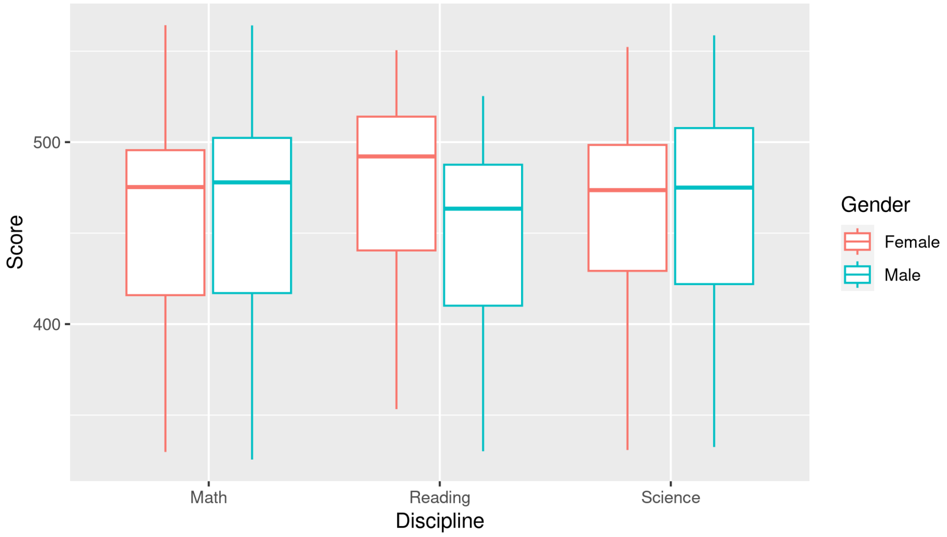

Figure 6: Comparing the performance of boys and girls in each discipline

5.1 In Each Region

It would be possible for some regions to have different gender trends than the rest of the world. For this reason, I created the same plot, but with added facets for region. Looking at Figure 7, it appears that the pattern of girls performing better than boys at Reading holds for every region, albeit with varying scale. In the Middle East & North Africa we also see them performing better than boys in Science. On the other hand, male students appear to score slightly higher than their female counterparts in Math in the Americas and the Caribbean.

```r
ggplot(dt_gendered) +
    geom_boxplot(aes(x = TestCode, y = Score, color = Gender)) +
    facet_wrap(~Region) +
    scale_x_discrete(labels = test_labels) +
    labs(x = "Discipline", y = "Score")
```

Listing 14: Generating Figure 7

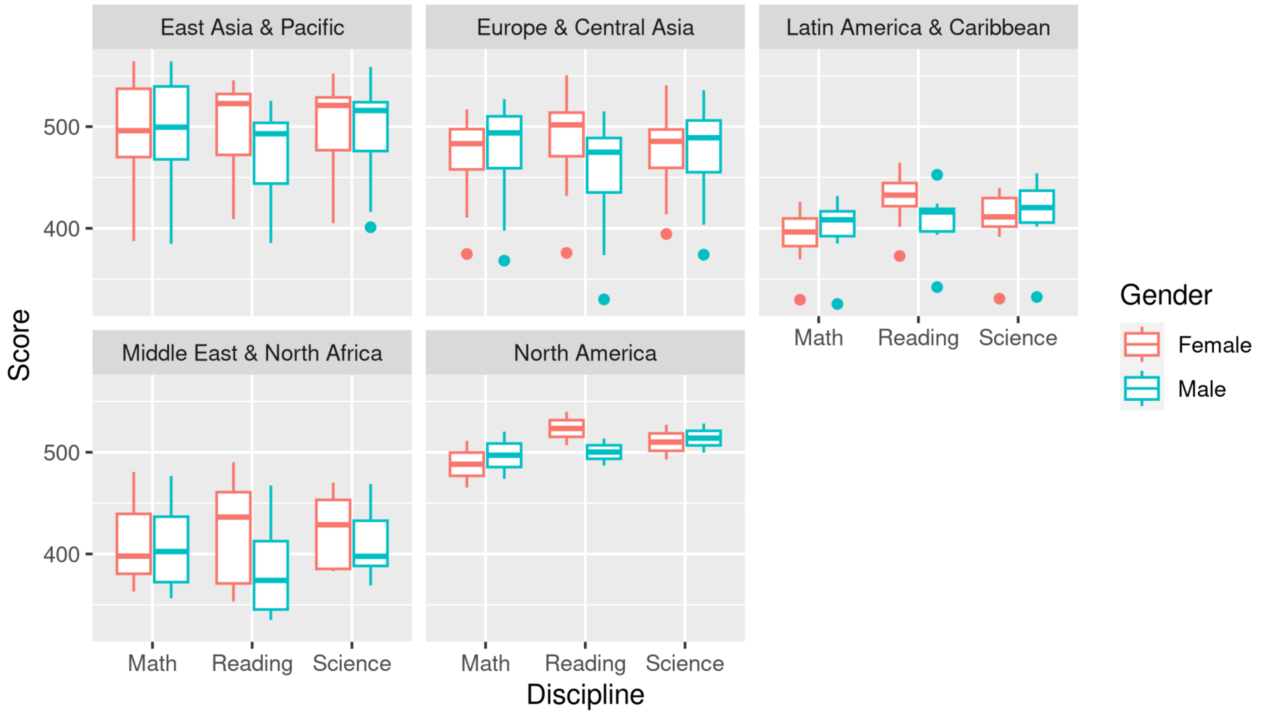

Figure 7: Comparing the performance of boys and girls in each discipline and region

5.2 In the European Union

Finally, I check whether the gender trends are different in the EU, using an identical approach as with the previous Section, except with facets for the EU column, instead of region. By examining Figure 8, this does not seem to be the case. In the EU, boys perform slightly better in Math than girls, but it is a small increase.

```r
ggplot(dt_gendered) +
    geom_boxplot(aes(x = TestCode, y = Score, color = Gender)) +
    # change to factor in order to fix the order of the facets
    facet_wrap(~factor(EU, levels = c(TRUE, FALSE)),
               labeller = as_labeller(facet_labels)) +
    scale_x_discrete(labels = test_labels) +
    labs(x = "Discipline", y = "Score")
```

Listing 15: Generating Figure 8

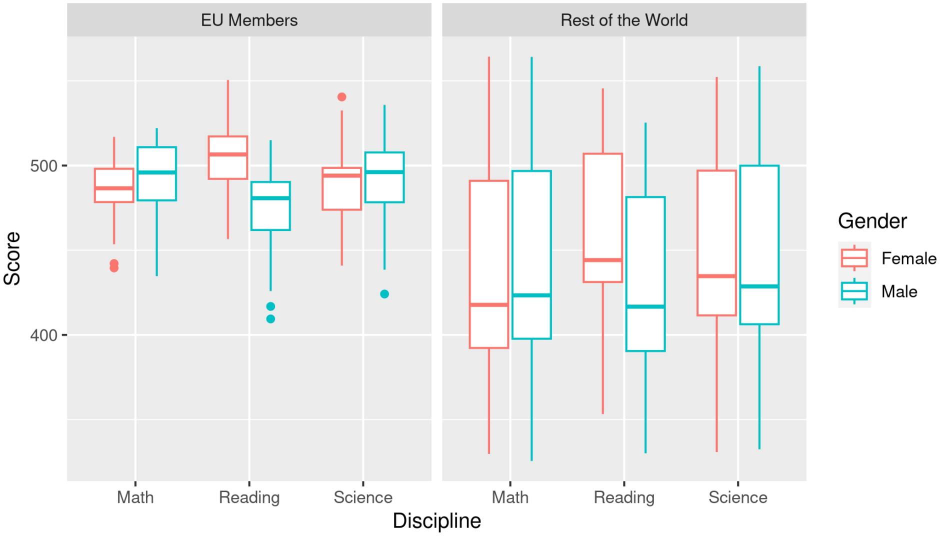

Figure 8: Comparing the effect of gender in the EU vs the rest of the world

6 Conclusion

I presented an introductory analysis of the 2015 PISA study, exploring whether three factors (country, region, gender) affected student performance in three disciplines – Math, Reading and Science. The performance of each country is varied, with some regions outperforming the others. Male and female students had comparable scores, with girls having a lead in Reading.

Every conclusion reached was done purely through graphical means and should not be taken at face value. Statistical testing is required, if we want to be more confident in them. However, they do paint a general outline and can point us in the right direction. Care should also be taken, when attempting to generalize any statement for a whole region or even the world, because most countries did not participate in the study.

References

[1] Vincent Arel-Bundock, Nils Enevoldsen, and Cj Yetman. “countrycode: An R package to convert country names and country codes”. In: *Journal of Open Source Software* 3.28 (2018), p. 848. DOI: 10.21105/joss.00848. URL: https://doi.org/10.21105/joss.00848.

[2] World Bank. *The World by Income and Region*. URL: https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html. (accessed: 3.1.2024).

[3] OECD. *PISA Frequently Asked Questions*. URL: https://www.oecd.org/pisa/pisafaq/. (accessed: 3.1.2024).