Project01_PISA

Introduction and background

Dataset description

In this dataset we have the PISA score and country characteristics for all OECD country for every 3 years (2006-2022). Due to COVID, the PISA test that supposed to help in 2021 was moved to 2022.

Data overview

Load the data from selected years.

```
library(readxl)
library(dplyr)
```

```
Warning: package 'dplyr' was built under R version 4.3.2
```

```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
# Load the data for 2018 and 2022
data_2009 = read_excel("PISA_project.xlsx", sheet = "2009")
data_2018 = read_excel("PISA_project.xlsx", sheet = "2018")
data_2021 = read_excel("PISA_project.xlsx", sheet = "2021")
data_2022 = read_excel("PISA_project.xlsx", sheet = "2022")

# Preview the data
# head(data_2009)
# head(data_2018)
# head(data_2021)
# head(data_2022)
```

Data cleaning

Replace missing value in 2022 by 2021 data:

Due to the fact that the PISA was cancelled in 2021, some economic indicators are not provided in 2022. We fill the unknown value in 2022 with 2021 data.

```r
# Prepare a subset of 2021 data with relevant columns
drop = c("Math_Female2021", "Math_Male2021", "Science_Female2021", "Science_Male2021")
data_2021_subset = data_2021[,!(names(data_2021) %in% drop)]

# Merge 2022 data with 2021 subset based on 'Country'
merged_data <- merge(data_2022, data_2021_subset, by="Country", suffixes = c("", "_from_2021

# Fill missing values in 2022 using 2021 data
for(column in names(merged_data)) {
  if(grepl("_from_2021$", column)) {
    original_column <- gsub("_from_2021$", "", column)
    # Check for NA in original column and fill
    is_na <- is.na(merged_data[, original_column])
    merged_data[is_na, original_column] <- merged_data[is_na, column]
  }
}

# Clean up - remove the extra columns from 2021
columns_to_remove = grep("_from_2021$", names(merged_data), value = TRUE)
merged_data = merged_data[, !(names(merged_data) %in% columns_to_remove)]
```

Overall score for math and science:

We assume the proportion of each gender are the same in each country, here the overall score is estimated by averaging male and female scores.

```
# Calculating overall Math and Science scores for 2009
data_2009$Overall_Math_Score_2009 = rowMeans(data_2009[,c('Math_Female2009', 'Math_Male2009'
data_2009$Overall_Science_Score_2009 = rowMeans(data_2009[,c('Science_Female2009', 'Science_
gender_2009 = c('Math_Female2009', 'Math_Male2009','Science_Female2009', 'Science_Male2009')
data_2009 = data_2009[, !(names(data_2009) %in% gender_2009)]

# Calculating overall Math and Science scores for 2018
data_2018$Overall_Math_Score_2018 = rowMeans(data_2018[,c('Math_Female2018', 'Math_Male2018'
data_2018$Overall_Science_Score_2018 = rowMeans(data_2018[,c('Science_Female2018', 'Science_
gender_2018 = c('Math_Female2018', 'Math_Male2018','Science_Female2018', 'Science_Male2018')
data_2018 = data_2018[, !(names(data_2018) %in% gender_2018)]

# Calculating overall Math and Science scores for 2022
merged_data$Overall_Math_Score_2022 = rowMeans(merged_data[,c('Math_Female2022', 'Math_Male2
merged_data$Overall_Science_Score_2022 = rowMeans(merged_data[,c('Science_Female2022', 'Scie
gender_2022 = c('Math_Female2022', 'Math_Male2022','Science_Female2022', 'Science_Male2022')
data_2022 = merged_data[, !(names(merged_data) %in% gender_2022)]
```

Initial Exploration

[Due to the page limit, the code here will be commented and only show the written description instead.]

```
# Summary statistics
# summary(data_2009)
# summary(data_2018)
# summary(data_2022)

# Missing values of selected years
# missing_values_2009 = colSums(is.na(data_2009))
# missing_values_2009 = missing_values_2009[missing_values_2009 > 0]
# missing_values_2018 = colSums(is.na(data_2018))
# missing_values_2018 = missing_values_2018[missing_values_2018 > 0]
# missing_values_2022 = colSums(is.na(data_2022))
# missing_values_2022 = missing_values_2022[missing_values_2022 > 0]

# Print out missing values information
```

```
# print(missing_values_2009)
# print(missing_values_2018)
# print(missing_values_2022)
```

For the dataset given, here's a summary of our findings:

- **Variables and Data Types:** All the selected spreadsheet includes a mix of numerical (e.g., Math and Science scores, GDP, urban population percentage) and categorical data (e.g., Country names).
- **Central Tendencies and Dispersion**: Scores and several economic indicators show a reasonable spread around their means, indicating variability across OECD countries.
- **Missing value:**
  - Year 2009
    - `Math_Male_2009`, `Math_Female_2009` , `Science_Male_2009` and `Math_Female_2009` each have 1 missing value.
    - **`Tax revenue (% of GDP)`** has 2 missing values.
    - **`Population in the largest city (% of urban population)`** has 3 missing values.
    - **`Gini index`** , **`School enrollment, secondary (% gross)`** and **`Researchers in R&D (per million people)`** each have 5 missing value.
    - **`Pupil-teacher ratio, secondary`** has 15 missing values.
    - **`Secure Internet servers (per 1 million people)`** have significant missing data (100% missing).
  - Year 2018
    - **`Tax revenue (% of GDP)`** has 1 missing values.
    - **`Population in the largest city (% of urban population)`** has 3 missing values.
    - **`Gini index`** has 5 missing values.
    - **`Researchers in R&D (per million people)`** has 6 missing value.
    - **`Pupil-teacher ratio, secondary`** has 36 missing values.
  - Year 2022 (Merged with 2021 data)

* **Math_Male_2022**, **Math_Female_2022** , **Science_Male_2022** , **Math_Female_2022** , **Tax revenue (% of GDP)** and **School enrollment, secondary (% gross)** each have 1 missing value.
* **Population in the largest city (% of urban population)** has 3 missing values.
* **Researchers in R&D (per million people)** has 8 missing value.
* **Gini index** has 11 missing value.
* **CO2 emissions (metric tons per capita)** , **Pupil-teacher ratio, secondary** , **Secure Internet servers (per 1 million people)** have significant missing data (100% missing).

Handling missing value

Fill in known data from external database

Gini index

We fill in the missing value of Gini index according to OECD and UN database.

```
# They only have 2020 value for Australia, Germany, Israel, Mexico, New Zealand, Switzerland
# Chile only have Gini index of 2022, so the 2021 Gini index of Chile will be filled by 2022

Gini_2021 <- c("Australia" = 31.8 , "Canada" = 29.2 , "Chile" = 44.8, "Germany" = 30.3, "Isra

for (country in names(Gini_2021)) {
  data_2022$`Gini index`[data_2022$Country == country] <- Gini_2021[country]
}

# They only have 2017 value for Chile and Iceland so the 2018 Gini index of them will be fil

Gini_2018 <- c("Chile" = 44.4, "Iceland" = 26.1)

for (country in names(Gini_2018)) {
  data_2018$`Gini index`[data_2018$Country == country] <- Gini_2018[country]
}
```

Due to the fact that we can't find the values for other variables, we proceed to removal of missing value.

Remove variable that have too many (>30%) missing values

```
# Remove variables that have more than 1/3 of missing values
threshold <- 0.3

# For 2009 data
missing_percentage_2009 = colMeans(is.na(data_2009))
columns_to_drop_2009 = names(missing_percentage_2009[missing_percentage_2009 > threshold])
data_2009_clean = data_2009[, !(names(data_2009) %in% columns_to_drop_2009)]

# For 2018 data
missing_percentage_2018 = colMeans(is.na(data_2018))
columns_to_drop_2018 = names(missing_percentage_2018[missing_percentage_2018 > threshold])
data_2018_clean = data_2018[, !(names(data_2018) %in% columns_to_drop_2018)]

# For 2022 data
missing_percentage_2022 = colMeans(is.na(data_2022))
columns_to_drop_2022 = names(missing_percentage_2022[missing_percentage_2022 > threshold])
data_2022_clean = data_2022[, !(names(data_2022) %in% columns_to_drop_2022)]
```

- **Removed variables:**
  - Year 2009
    - **Pupil-teacher ratio, secondary**
    - **Secure Internet servers (per 1 million people)**
  - Year 2018
    - **Pupil-teacher ratio, secondary**
  - Year 2022 (Merged with 2021 data)
    - **CO2 emissions (metric tons per capita)**
    - **Pupil-teacher ratio, secondary**
    - **Secure Internet servers (per 1 million people)**

Exploratory Data Analysis (EDA)

Overall Math Score

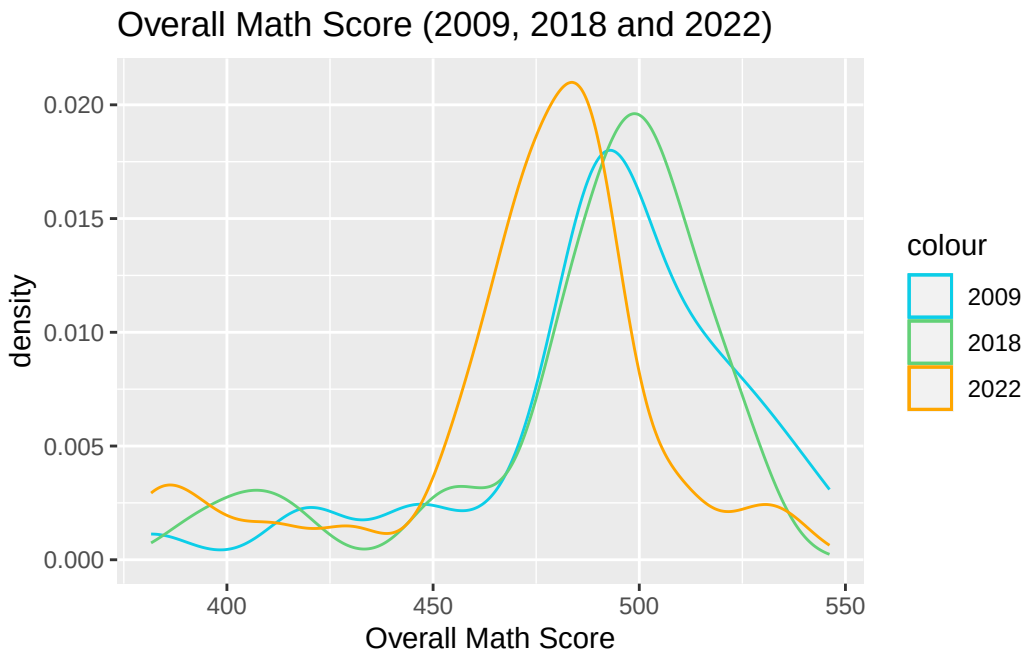
```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.3.2
```

```
suppressWarnings({
ggplot() +
  geom_density( data_2009_clean, mapping = aes(x = Overall_Math_Score_2009, colour = "2009")) +
  geom_density(data_2018_clean, mapping = aes(x=Overall_Math_Score_2018, colour = "2018")) +
  geom_density(data_2022_clean, mapping = aes(x=Overall_Math_Score_2022, colour = "2022")) +
  scale_color_manual(values = c("#0ccee8", "#62d177", "#ffa600"))+
  xlab("Overall Math Score") +
   ggtitle("Overall Math Score (2009, 2018 and 2022)")
})
```

```
Warning: Removed 1 rows containing non-finite values (`stat_density()`).
```

```
Warning: Removed 1 rows containing non-finite values (`stat_density()`).
```

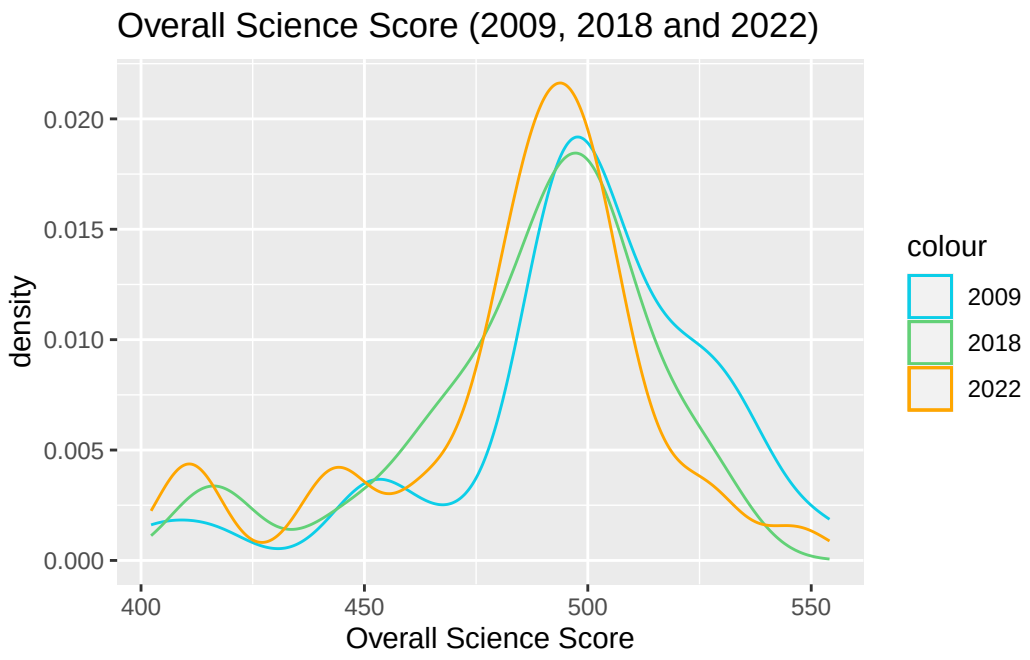


The overall math scores increased from 2009 to 2018, but decreased from 2018 to 2022.

Overall Science Score

```
library(ggplot2)

suppressWarnings({
ggplot() +
  geom_density(data_2009_clean, mapping = aes(x = Overall_Science_Score_2009, colour = "2009"
  geom_density(data_2018_clean, mapping = aes(x=Overall_Science_Score_2018, colour = "2018")
  geom_density(data_2022_clean, mapping = aes(x=Overall_Science_Score_2022, colour = "2022")
  scale_color_manual(values = c("#0ccee8", "#62d177", "#ffa600"))+
  xlab("Overall Science Score") +
   ggtitle("Overall Science Score (2009, 2018 and 2022)")
})
```

```
Warning: Removed 1 rows containing non-finite values (`stat_density()`).
Removed 1 rows containing non-finite values (`stat_density()`).
```



The overall science scores were about the same in both 2009 and 2018, but decreased from 2018 to 2022.

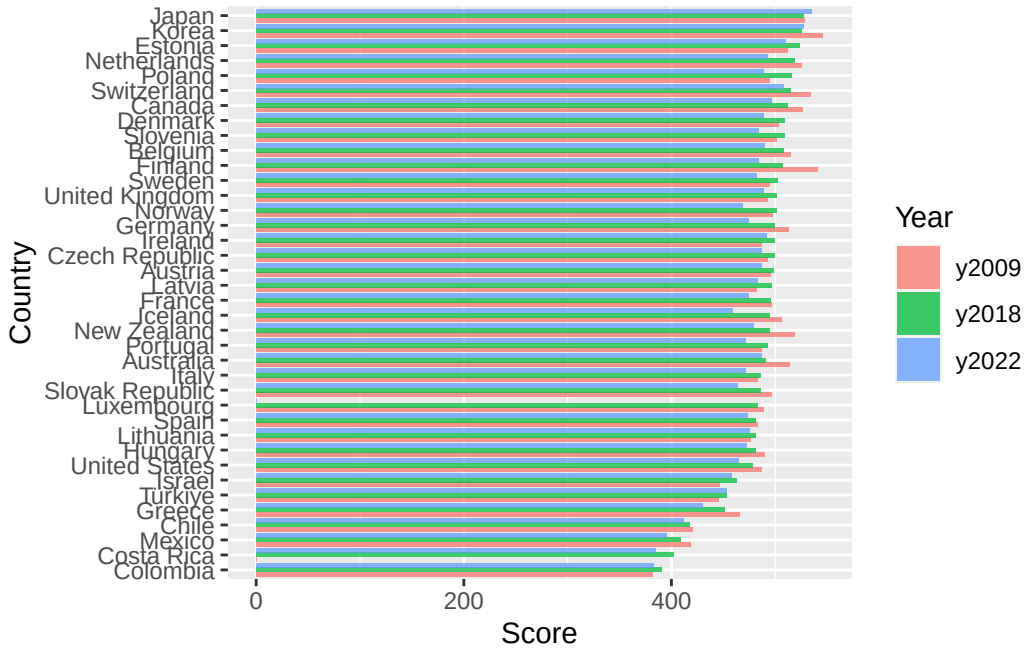
Overall Math Score by country

In this plot we can see the math score annual changes in the OECD countries.

```
# Fill missing value with 0
M2009 = replace(data_2009_clean, is.na(data_2009_clean), 0)
M2018 = replace(data_2018_clean, is.na(data_2018_clean), 0)
M2022 = replace(data_2022_clean, is.na(data_2022_clean), 0)
Math = data.frame(M2009$Country, M2009$Overall_Math_Score_2009, M2018$Overall_Math_Score_2018
colnames(Math) <- c('Country','y2009','y2018','y2022')
Math_sort <- Math[order(Math$y2018),]

# Wide to long
library(tidyr)
```

```
Warning: package 'tidyr' was built under R version 4.3.2
```

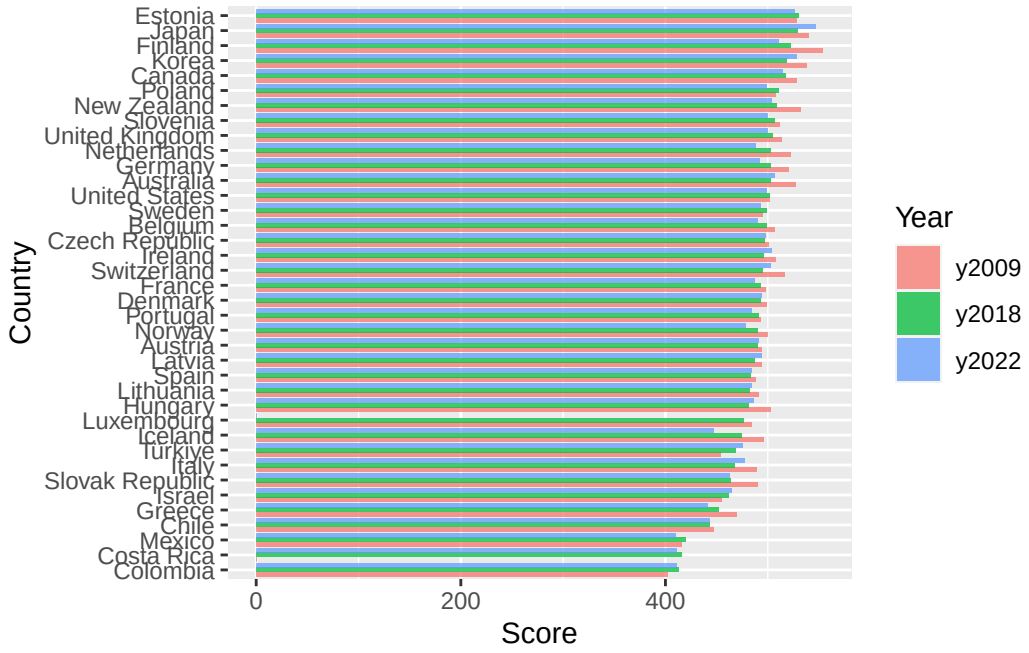
```
Math_long <- gather(Math_sort, Year, Score, "y2009", "y2018", "y2022")

# Plot the bar graph
Math_long$Country <- factor(unique(Math_long$Country), levels = unique(Math_long$Country))
ggplot(data = Math_long, aes(x = Country, y = Score, fill = Year)) +
  geom_bar(stat = "identity", position = position_dodge(), alpha = 0.75) +
  coord_flip()
```



Overall Science Score by country

In this plot we can see the science score annual changes in the OECD countries.

```
# Fill missing value with 0
D2009 = replace(data_2009_clean, is.na(data_2009_clean), 0)
D2018 = replace(data_2018_clean, is.na(data_2018_clean), 0)
D2022 = replace(data_2022_clean, is.na(data_2022_clean), 0)
Science = data.frame(D2009$Country, D2009$Overall_Science_Score_2009, D2018$Overall_Science_S
colnames(Science) <- c('Country','y2009','y2018','y2022')
Science_sort <- Science[order(Science$y2018),]

# Wide to long
library(tidyr)
Science_long <- gather(Science_sort, Year, Score, "y2009", "y2018", "y2022")

# Plot the bar graph
Science_long$Country <- factor(unique(Science_long$Country), levels = unique(Science_long$Cou
ggplot(data = Science_long, aes(x = Country, y = Score, fill = Year)) +
  geom_bar(stat = "identity", position = position_dodge(), alpha = 0.75) +
  coord_flip()
```



From the plots above, we can see that Japan and Colombia are the two extremes among OECD countries in PISA math and science score.

Correlation matrix

2009 Correlation matrix

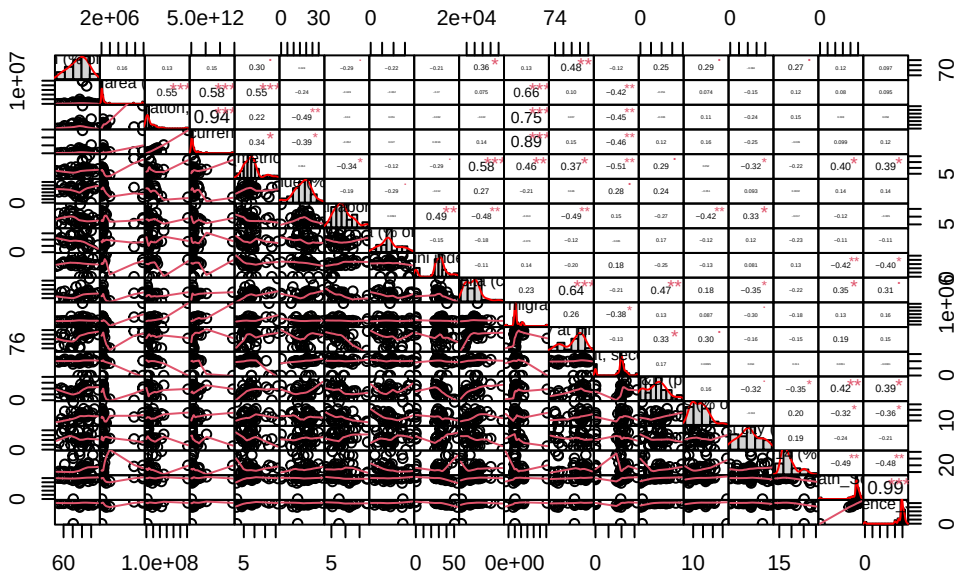
```
suppressWarnings({
library("PerformanceAnalytics")
Y2009 = as.matrix(D2009[-1])
chart.Correlation(Y2009, histogram=TRUE)
})
```

```
Loading required package: xts
```

```
Loading required package: zoo
```

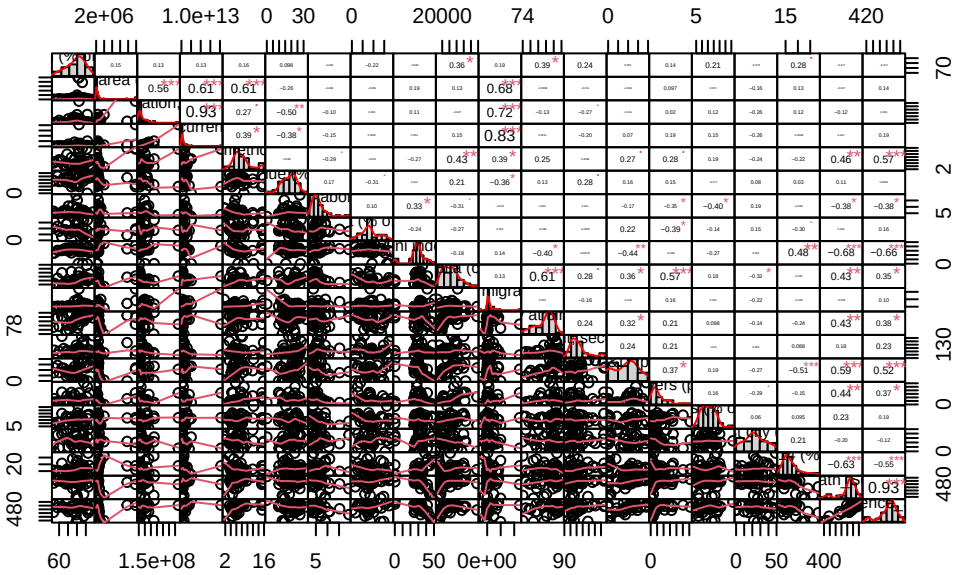
```


Attaching package: 'zoo'
```

```
The following objects are masked from 'package:base':

    as.Date, as.Date.numeric


######################### Warning from 'xts' package #########################
#                                                                            #
# The dplyr lag() function breaks how base R's lag() function is supposed to #
# work, which breaks lag(my_xts). Calls to lag(my_xts) that you type or      #
# source() into this session won't work correctly.                           #
#                                                                            #
# Use stats::lag() to make sure you're not using dplyr::lag(), or you can add #
# conflictRules('dplyr', exclude = 'lag') to your .Rprofile to stop          #
# dplyr from breaking base R's lag() function.                               #
#                                                                            #
# Code in packages is not affected. It's protected by R's namespace mechanism #
# Set `options(xts.warn_dplyr_breaks_lag = FALSE)` to suppress this warning. #
#                                                                            #
##############################################################################


Attaching package: 'xts'

The following objects are masked from 'package:dplyr':

    first, last


Attaching package: 'PerformanceAnalytics'

The following object is masked from 'package:graphics':

    legend
```



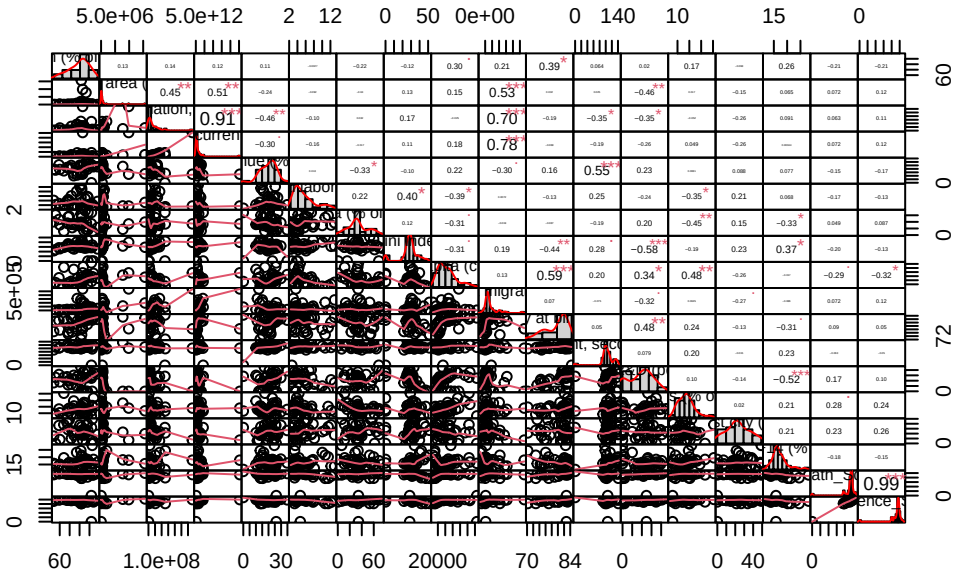
2018 Correlation matrix

```
suppressWarnings({
library("PerformanceAnalytics")
Y2018 = as.matrix(D2018[-1])
chart.Correlation(Y2018, histogram=TRUE)
})
```



2022 Correlation matrix

```
suppressWarnings({
library("PerformanceAnalytics")
Y2022 = as.matrix(D2022[-1])
chart.Correlation(Y2022, histogram=TRUE)
})
```



By the correlation matrix we can see some pattern of each variable.

Across the years, we see found:

- PISA math and PISA science are highly positively correlated.
- GDP per capita (current US$) and PISA scores are correlated. (Direct changed since 2022)

The pattern of PISA performance with country characteristics changed since 2022. Some highly correlated features no longer have relationship with PISA performance. It can potentially due to some global event (e.g., COVID).

Statistical Analysis

Q1: Effect of Characteristics on Scores (2018, 2022)

In this question, we aimed to understand what country characteristic affect the PISA score (math and science) for that designated year (2018 and 2022) for the OECD countries.

Data preperation

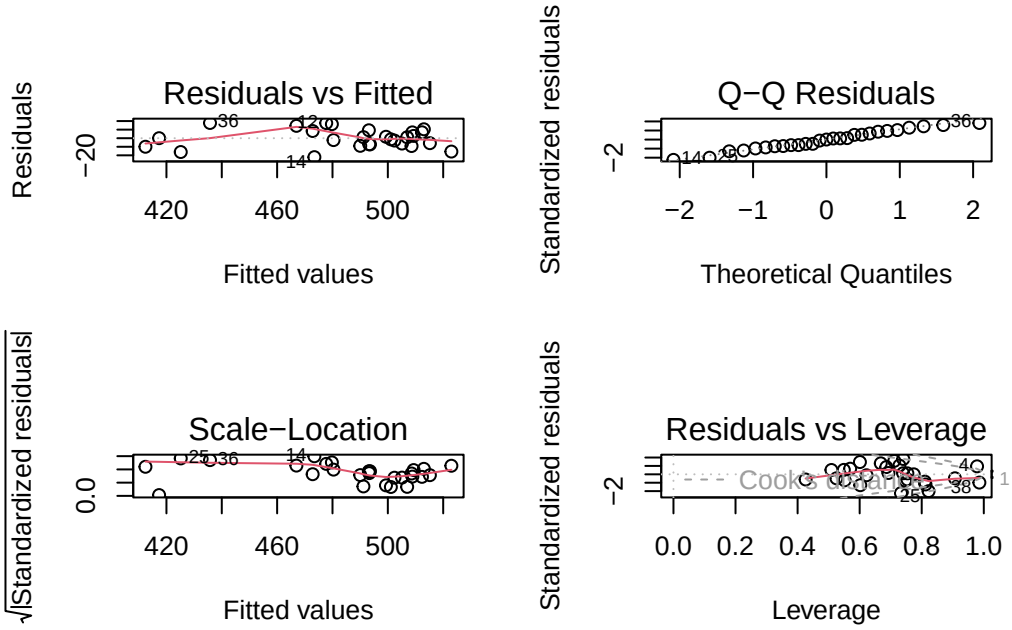
```
# Exclude data that are irrelevant to the following analysis.
S18 = c("Overall_Science_Score_2018", "Country")
math18 = data_2018_clean[, !(names(data_2018_clean) %in% S18)]
S22 = c("Overall_Science_Score_2022", "Country")
math22 = data_2022_clean[, !(names(data_2022_clean) %in% S22)]
M18 = c("Overall_Math_Score_2018", "Country")
science18 = data_2018_clean[, !(names(data_2018_clean) %in% M18)]
M22 = c("Overall_Math_Score_2022", "Country")
science22 = data_2022_clean[, !(names(data_2022_clean) %in% M22)]
```

Assumption check

PISA Math 2018

```
full_math18 <- lm(Overall_Math_Score_2018 ~ ., data = math18)
par(mfrow = c(2,2))
plot(full_math18)
```

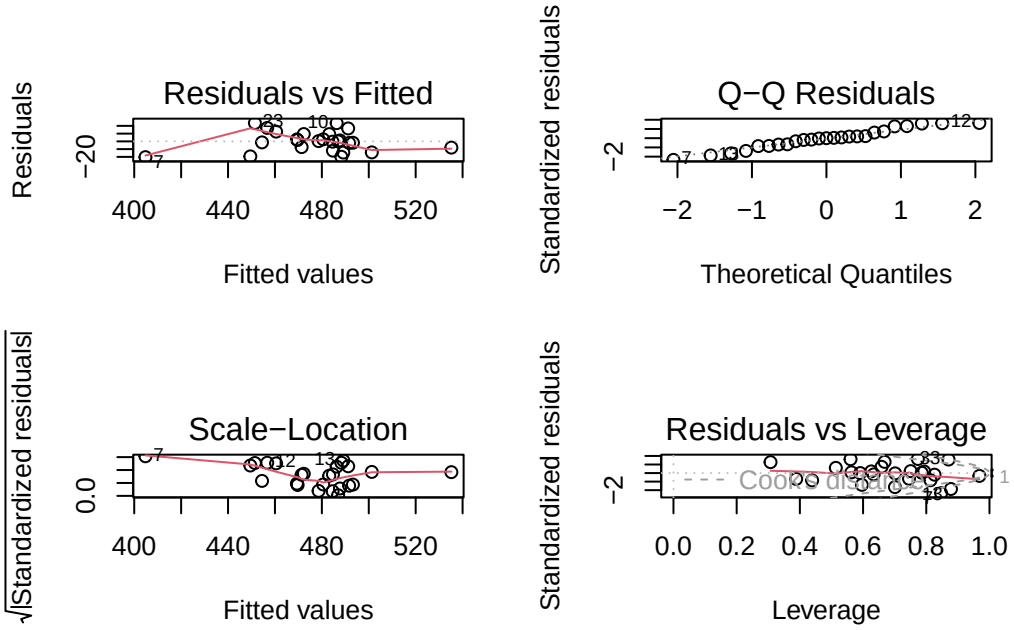
```
Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



PISA Math 2022

```
full_math22 <- lm(Overall_Math_Score_2022 ~ . ,data = math22)
par(mfrow = c(2,2))
plot(full_math22)
```

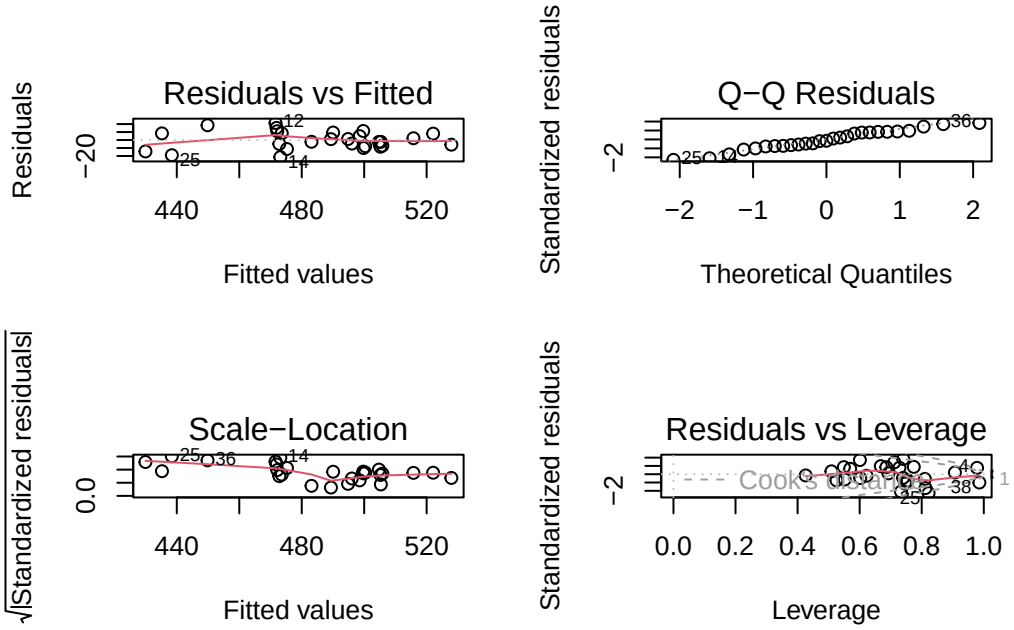
```
Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



PISA Science 2018

```
full_science18 <- lm(Overall_Science_Score_2018 ~ . ,data = science18)
par(mfrow = c(2,2))
plot(full_science18)
```

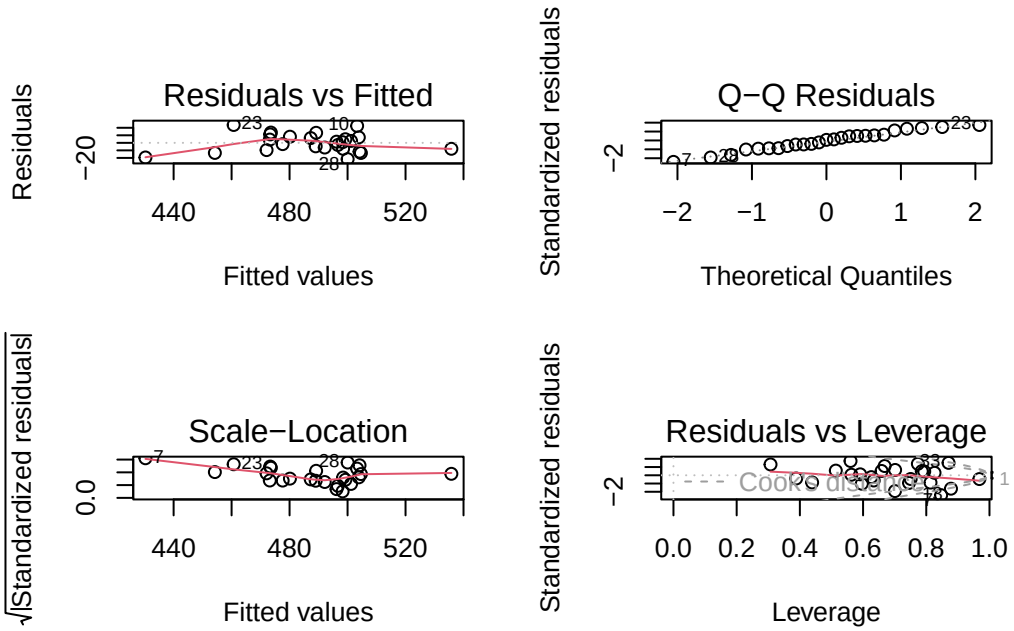
```
Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



PISA Science 2022

```
full_science22 <- lm(Overall_Science_Score_2022 ~ . ,data = science22)
par(mfrow = c(2,2))
plot(full_science22)
```

```
Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Assumption check: Summary

- All the 4 models have the similar **assumption results**:
  - **Residuals vs Fitted**. Used to check the linear relationship assumptions. We found the data may be curve-linear.
  - **Normal Q-Q**. Used to examine whether the residuals are normally distributed. The residuals points follow the straight dashed line, which showed our data follows a normal distribution.
  - **Scale-Location.** Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity. This is not the case in our example, where we have a heteroscedasticity problem.
  - **Residuals vs Leverage**. Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. We can see there are some extreme cases in our data. Due to the fact that we want to have a comprehensive understanding of PISA and OECD countries' characteristics, we are not going to exclude any data point from the data for the following analysis even if they are outliers by definition.

In order to answer the given question, we utilize linear regression analysis despite some potential non-linearity. Because this is the easiest way to interpret and compare the results, given we have so many variables in the dataset.

Note: I did consider to add interaction term and transform the data to resolve the non-linearity. But adding the interaction term will make the result extreme difficult to interpret since we have too many variables. Plus, they are not comprehensively selected we may not effectively find out the proper interaction effect. Similarly, transform the data will significantly decrease the interpretability.

Linear regression

PISA Math 2018

```
summary(full_math18)
```

```
Call:
lm(formula = Overall_Math_Score_2018 ~ ., data = math18)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0915  -7.1642   0.0085   7.6420  17.7431

Coefficients:
                                                                                Estimate
(Intercept)                                                                    5.609e+02
`Urban population (% of total population)`                                    -4.112e-01
`Surface area (sq. km)`                                                        1.105e-06
`Population, total`                                                            3.205e-07
`GDP (current US$)`                                                           -3.609e-12
`CO2 emissions (metric tons per capita)`                                       3.012e+00
`Tax revenue (% of GDP)`                                                       8.804e-01
`Unemployment, total (% of total labor force) (modeled ILO estimate)` -9.148e-01
`Forest area (% of land area)`                                                 2.197e-01
`Gini index`                                                                  -1.801e+00
`GDP per capita (current US$)`                                                 1.886e-04
`Net migration`                                                               -3.228e-05
`Life expectancy at birth, total (years)`                                     -1.070e-01
`School enrollment, secondary (% gross)`                                       5.069e-01
`Researchers in R&D (per million people)`                                     -9.647e-04
`Secure Internet servers (per 1 million people)`                               1.699e-04
```

```
`High-technology exports (% of manufactured exports)`                 3.180e-01
`Population in the largest city (% of urban population)`             -9.453e-02
`Population ages 0-14 (% of total population)`                       -5.259e+00
                                                                     Std. Error
(Intercept)                                                           2.258e+02
`Urban population (% of total population)`                            6.403e-01
`Surface area (sq. km)`                                               4.350e-06
`Population, total`                                                   3.358e-07
`GDP (current US$)`                                                   6.087e-12
`CO2 emissions (metric tons per capita)`                              3.607e+00
`Tax revenue (% of GDP)`                                              1.623e+00
`Unemployment, total (% of total labor force) (modeled ILO estimate)` 1.726e+00
`Forest area (% of land area)`                                        3.519e-01
`Gini index`                                                          2.409e+00
`GDP per capita (current US$)`                                        5.558e-04
`Net migration`                                                       6.659e-05
`Life expectancy at birth, total (years)`                             3.302e+00
`School enrollment, secondary (% gross)`                              3.644e-01
`Researchers in R&D (per million people)`                             7.517e-03
`Secure Internet servers (per 1 million people)`                      2.409e-04
`High-technology exports (% of manufactured exports)`                 9.027e-01
`Population in the largest city (% of urban population)`              6.439e-01
`Population ages 0-14 (% of total population)`                        3.131e+00
                                                                      t value
(Intercept)                                                             2.484
`Urban population (% of total population)`                             -0.642
`Surface area (sq. km)`                                                 0.254
`Population, total`                                                     0.954
`GDP (current US$)`                                                    -0.593
`CO2 emissions (metric tons per capita)`                                0.835
`Tax revenue (% of GDP)`                                                0.542
`Unemployment, total (% of total labor force) (modeled ILO estimate)`  -0.530
`Forest area (% of land area)`                                          0.624
`Gini index`                                                           -0.748
`GDP per capita (current US$)`                                          0.339
`Net migration`                                                        -0.485
`Life expectancy at birth, total (years)`                              -0.032
`School enrollment, secondary (% gross)`                                1.391
`Researchers in R&D (per million people)`                              -0.128
`Secure Internet servers (per 1 million people)`                        0.705
`High-technology exports (% of manufactured exports)`                   0.352
`Population in the largest city (% of urban population)`               -0.147
`Population ages 0-14 (% of total population)`                         -1.680
```

```
                                                                  Pr(>|t|)
(Intercept)                                                         0.0378
`Urban population (% of total population)`                          0.5387
`Surface area (sq. km)`                                             0.8060
`Population, total`                                                 0.3678
`GDP (current US$)`                                                 0.5696
`CO2 emissions (metric tons per capita)`                            0.4280
`Tax revenue (% of GDP)`                                            0.6024
`Unemployment, total (% of total labor force) (modeled ILO estimate)`  0.6106
`Forest area (% of land area)`                                      0.5497
`Gini index`                                                        0.4760
`GDP per capita (current US$)`                                      0.7431
`Net migration`                                                     0.6408
`Life expectancy at birth, total (years)`                           0.9749
`School enrollment, secondary (% gross)`                            0.2017
`Researchers in R&D (per million people)`                           0.9011
`Secure Internet servers (per 1 million people)`                    0.5007
`High-technology exports (% of manufactured exports)`               0.7338
`Population in the largest city (% of urban population)`            0.8869
`Population ages 0-14 (% of total population)`                      0.1315

(Intercept)                                                         *
`Urban population (% of total population)`
`Surface area (sq. km)`
`Population, total`
`GDP (current US$)`
`CO2 emissions (metric tons per capita)`
`Tax revenue (% of GDP)`
`Unemployment, total (% of total labor force) (modeled ILO estimate)`
`Forest area (% of land area)`
`Gini index`
`GDP per capita (current US$)`
`Net migration`
`Life expectancy at birth, total (years)`
`School enrollment, secondary (% gross)`
`Researchers in R&D (per million people)`
`Secure Internet servers (per 1 million people)`
`High-technology exports (% of manufactured exports)`
`Population in the largest city (% of urban population)`
`Population ages 0-14 (% of total population)`
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 19.09 on 8 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.892, Adjusted R-squared:  0.649
F-statistic: 3.671 on 18 and 8 DF,  p-value: 0.03304
```

Coefficients

- **Intercept (560.9)**: The expected value of the Overall Math Score in 2018 when all other variables are zero. The significant p-value (0.0378) indicates that the intercept is significantly different from zero.
- **Variables**: Most predictors are not statistically significant at the 0.05 level (no stars next to their p-values), meaning there's not enough evidence to assert they have a strong linear relationship with Math scores under this model setup.

Model Fit

- **Adjusted R-squared (0.649)**: A decrease from 0.892 (original) to 0.649 in R-squared suggests that some predictors may not be contributing much to the model. About 64.9% of the variability in the Overall Math Scores in 2018 is explained by the model's predictors.
- **F-statistic (3.671) and p-value (0.03304)**: This tests the null hypothesis that all regression coefficients are equal to zero (no linear relationship). A p-value of 0.03304 indicates the model is statistically significant at the 0.05 level, suggesting at least one of the predictors has a linear relationship with the Overall Math Score in 2018.

Overall

- **Non-significance of Many Predictors**: Most of the predictors are not statistically significant, which might suggest that either these factors do not have a strong linear relationship with Math scores or that the model could be improved. Especially, we may not actually select the most relevant variables to be the predictors.

PISA Math 2022

```
summary(full_math22)
```

```
Call:
lm(formula = Overall_Math_Score_2022 ~ ., data = math22)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-20.3326  -8.0425   0.0091   9.3866  23.7259

Coefficients:
                                                                                Estimate
(Intercept)                                                                    6.241e+02
`Urban population (% of total population)`                                     6.088e-01
`Surface area (sq. km)`                                                       -7.058e-05
`Population, total`                                                            2.100e-06
`GDP (current US$)`                                                           -4.166e-11
`Tax revenue (% of GDP)`                                                       1.105e+00
`Unemployment, total (% of total labor force) (modeled ILO estimate)`         -3.697e+00
`Forest area (% of land area)`                                                 7.299e-01
`Gini index`                                                                  -2.281e+00
`GDP per capita (current US$)`                                                 8.292e-04
`Net migration`                                                                1.069e-04
`Life expectancy at birth, total (years)`                                     -1.760e+00
`School enrollment, secondary (% gross)`                                      -1.834e-01
`Researchers in R&D (per million people)`                                     -2.951e-04
`High-technology exports (% of manufactured exports)`                          6.482e-01
`Population in the largest city (% of urban population)`                       4.982e-01
`Population ages 0-14 (% of total population)`                                -3.102e+00
                                                                              Std. Error
(Intercept)                                                                    2.277e+02
`Urban population (% of total population)`                                     8.068e-01
`Surface area (sq. km)`                                                        5.232e-05
`Population, total`                                                            8.718e-07
`GDP (current US$)`                                                            1.975e-11
`Tax revenue (% of GDP)`                                                       1.492e+00
`Unemployment, total (% of total labor force) (modeled ILO estimate)`          2.781e+00
`Forest area (% of land area)`                                                 4.916e-01
`Gini index`                                                                   1.633e+00
`GDP per capita (current US$)`                                                 5.973e-04
`Net migration`                                                                1.028e-04
`Life expectancy at birth, total (years)`                                      3.337e+00
`School enrollment, secondary (% gross)`                                       5.281e-01
`Researchers in R&D (per million people)`                                      5.915e-03
`High-technology exports (% of manufactured exports)`                          1.230e+00
`Population in the largest city (% of urban population)`                       6.090e-01
`Population ages 0-14 (% of total population)`                                 3.597e+00
                                                                              t value
(Intercept)                                                                      2.741
`Urban population (% of total population)`                                       0.755
```

```
`Surface area (sq. km)`                                                  -1.349
`Population, total`                                                       2.409
`GDP (current US$)`                                                      -2.109
`Tax revenue (% of GDP)`                                                  0.741
`Unemployment, total (% of total labor force) (modeled ILO estimate)`    -1.330
`Forest area (% of land area)`                                            1.485
`Gini index`                                                             -1.397
`GDP per capita (current US$)`                                            1.388
`Net migration`                                                           1.040
`Life expectancy at birth, total (years)`                                -0.527
`School enrollment, secondary (% gross)`                                 -0.347
`Researchers in R&D (per million people)`                                -0.050
`High-technology exports (% of manufactured exports)`                     0.527
`Population in the largest city (% of urban population)`                  0.818
`Population ages 0-14 (% of total population)`                           -0.862
                                                                        Pr(>|t|)
(Intercept)                                                               0.0254
`Urban population (% of total population)`                                0.4722
`Surface area (sq. km)`                                                   0.2143
`Population, total`                                                       0.0426
`GDP (current US$)`                                                       0.0680
`Tax revenue (% of GDP)`                                                  0.4800
`Unemployment, total (% of total labor force) (modeled ILO estimate)`     0.2203
`Forest area (% of land area)`                                            0.1759
`Gini index`                                                              0.2000
`GDP per capita (current US$)`                                            0.2025
`Net migration`                                                           0.3287
`Life expectancy at birth, total (years)`                                 0.6122
`School enrollment, secondary (% gross)`                                  0.7373
`Researchers in R&D (per million people)`                                 0.9614
`High-technology exports (% of manufactured exports)`                     0.6125
`Population in the largest city (% of urban population)`                  0.4370
`Population ages 0-14 (% of total population)`                            0.4137

(Intercept)                                                              *
`Urban population (% of total population)`
`Surface area (sq. km)`
`Population, total`                                                      *
`GDP (current US$)`                                                      .
`Tax revenue (% of GDP)`
`Unemployment, total (% of total labor force) (modeled ILO estimate)`
`Forest area (% of land area)`
`Gini index`
```

```
`GDP per capita (current US$)`
`Net migration`
`Life expectancy at birth, total (years)`
`School enrollment, secondary (% gross)`
`Researchers in R&D (per million people)`
`High-technology exports (% of manufactured exports)`
`Population in the largest city (% of urban population)`
`Population ages 0-14 (% of total population)`
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.13 on 8 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.7737,    Adjusted R-squared:  0.321
F-statistic: 1.709 on 16 and 8 DF,  p-value: 0.2239
```

Coefficients

- **Intercept (624.1)**: This is the predicted Overall Math Score in 2022 when all predictor variables are held at zero. It is statistically significant at the 0.05 level (p-value: 0.0254), suggesting a baseline level of math proficiency across the sample when no other factors are considered.
- **Variables**:
  - `Population, total` has a positive coefficient (2.100e-06) with a p-value of 0.0426, indicating that a larger population is associated with a higher math score, and this relationship is statistically significant at the 0.05 level.
  - `GDP (current US$)` shows a negative coefficient (-4.166e-11), with a p-value close to the significance threshold (0.0680), suggesting that higher GDP might be associated with lower math scores, although this result is not statistically significant at the conventional 0.05 level.

Model Fit

- **Adjusted R-squared (0.321)**: After adjusting for the number of predictors, the adjusted R-squared significantly drops (from 0.7737), suggesting that some of the model's explanatory power may be due to the number of predictors rather than their individual explanatory value. About 32.1% of the variability in the Overall Math Scores in 2022 is actually explained by the model's predictors.
- **F-statistic and p-value**: The F-statistic tests whether at least one of the predictors is significantly related to the response variable. With a p-value of 0.2239, the test suggests that the model, as a whole, might not be statistically significant, indicating that the

predictors collectively may not explain the variation in math scores as well as suggested by the R-squared value.

Overall

- **Significance**: Only a few predictors are statistically significant, indicating that many variables included in the model may not have a strong linear relationship with the Overall Math Score in 2022.

PISA Science 2018

```
summary(full_science18)
```

```
Call:
lm(formula = Overall_Science_Score_2018 ~ ., data = science18)

Residuals:
    Min      1Q  Median      3Q     Max
-21.440  -7.375  -2.194   8.416  21.362

Coefficients:
                                                                      Estimate
(Intercept)                                                          4.678e+02
`Urban population (% of total population)`                          -1.552e-01
`Surface area (sq. km)`                                             -9.701e-07
`Population, total`                                                  2.585e-07
`GDP (current US$)`                                                 -2.350e-12
`CO2 emissions (metric tons per capita)`                             4.674e+00
`Tax revenue (% of GDP)`                                            -2.219e-01
`Unemployment, total (% of total labor force) (modeled ILO estimate)` -7.973e-01
`Forest area (% of land area)`                                       1.568e-01
`Gini index`                                                        -5.279e-01
`GDP per capita (current US$)`                                      -1.759e-04
`Net migration`                                                     -1.800e-05
`Life expectancy at birth, total (years)`                           -2.236e-02
`School enrollment, secondary (% gross)`                             4.730e-01
`Researchers in R&D (per million people)`                            5.058e-03
`Secure Internet servers (per 1 million people)`                     5.488e-05
`High-technology exports (% of manufactured exports)`                1.600e-01
`Population in the largest city (% of urban population)`             2.651e-01
`Population ages 0-14 (% of total population)`                      -3.642e+00
```

	Std. Error
(Intercept)	2.340e+02
`Urban population (% of total population)`	6.636e-01
`Surface area (sq. km)`	4.508e-06
`Population, total`	3.480e-07
`GDP (current US$)`	6.309e-12
`CO2 emissions (metric tons per capita)`	3.738e+00
`Tax revenue (% of GDP)`	1.683e+00
`Unemployment, total (% of total labor force) (modeled ILO estimate)`	1.789e+00
`Forest area (% of land area)`	3.647e-01
`Gini index`	2.497e+00
`GDP per capita (current US$)`	5.761e-04
`Net migration`	6.901e-05
`Life expectancy at birth, total (years)`	3.422e+00
`School enrollment, secondary (% gross)`	3.777e-01
`Researchers in R&D (per million people)`	7.791e-03
`Secure Internet servers (per 1 million people)`	2.497e-04
`High-technology exports (% of manufactured exports)`	9.356e-01
`Population in the largest city (% of urban population)`	6.674e-01
`Population ages 0-14 (% of total population)`	3.245e+00
	t value
(Intercept)	1.999
`Urban population (% of total population)`	-0.234
`Surface area (sq. km)`	-0.215
`Population, total`	0.743
`GDP (current US$)`	-0.372
`CO2 emissions (metric tons per capita)`	1.250
`Tax revenue (% of GDP)`	-0.132
`Unemployment, total (% of total labor force) (modeled ILO estimate)`	-0.446
`Forest area (% of land area)`	0.430
`Gini index`	-0.211
`GDP per capita (current US$)`	-0.305
`Net migration`	-0.261
`Life expectancy at birth, total (years)`	-0.007
`School enrollment, secondary (% gross)`	1.252
`Researchers in R&D (per million people)`	0.649
`Secure Internet servers (per 1 million people)`	0.220
`High-technology exports (% of manufactured exports)`	0.171
`Population in the largest city (% of urban population)`	0.397
`Population ages 0-14 (% of total population)`	-1.122
	Pr(>|t|)
(Intercept)	0.0806
`Urban population (% of total population)`	0.8210

```
`Surface area (sq. km)`                                                    0.8350
`Population, total`                                                        0.4788
`GDP (current US$)`                                                        0.7192
`CO2 emissions (metric tons per capita)`                                   0.2465
`Tax revenue (% of GDP)`                                                   0.8983
`Unemployment, total (% of total labor force) (modeled ILO estimate)`      0.6677
`Forest area (% of land area)`                                             0.6786
`Gini index`                                                               0.8378
`GDP per capita (current US$)`                                             0.7679
`Net migration`                                                            0.8008
`Life expectancy at birth, total (years)`                                  0.9949
`School enrollment, secondary (% gross)`                                   0.2458
`Researchers in R&D (per million people)`                                  0.5343
`Secure Internet servers (per 1 million people)`                           0.8316
`High-technology exports (% of manufactured exports)`                      0.8685
`Population in the largest city (% of urban population)`                   0.7016
`Population ages 0-14 (% of total population)`                             0.2943

(Intercept)                                                                .
`Urban population (% of total population)`
`Surface area (sq. km)`
`Population, total`
`GDP (current US$)`
`CO2 emissions (metric tons per capita)`
`Tax revenue (% of GDP)`
`Unemployment, total (% of total labor force) (modeled ILO estimate)`
`Forest area (% of land area)`
`Gini index`
`GDP per capita (current US$)`
`Net migration`
`Life expectancy at birth, total (years)`
`School enrollment, secondary (% gross)`
`Researchers in R&D (per million people)`
`Secure Internet servers (per 1 million people)`
`High-technology exports (% of manufactured exports)`
`Population in the largest city (% of urban population)`
`Population ages 0-14 (% of total population)`
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.79 on 8 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.8434,    Adjusted R-squared:  0.4911
```

```
F-statistic: 2.394 on 18 and 8 DF,  p-value: 0.1046
```

Coefficients Table

- **Intercept (467.8)**: This suggests the expected Overall Science Score in 2018 when all predictors are zero. While close to statistical significance (p = 0.0806), it indicates a baseline level of science proficiency.
- **Variables**: None of the predictors are statistically significant at the conventional 0.05 level, as indicated by the lack of asterisks next to their p-values. This suggests no strong evidence from this model that these factors are linearly related to science scores in a statistically significant way within the dataset provided.

Model Fit

- **Adjusted R-squared (0.4911)**: Significantly lower than the Multiple R-squared (0.8434), this adjustment accounts for the number of predictors in the model, suggesting that some predictors may not contribute meaningfully to explaining the variance in science scores. Overall Science Scores is explained by the model (approximately 49.11%).
- **F-statistic (2.394) and p-value (0.1046)**: Tests the null hypothesis that all regression coefficients are zero. The p-value above the common threshold (0.05) suggests the model, as a whole, might not significantly predict science scores, indicating that, collectively, the predictors may not have a strong linear relationship with science scores.

Overall

- **Lack of Statistical Significance**: The absence of statistically significant predictors may suggest that either the variables chosen do not have a strong linear relationship with science scores or the model could benefit from refinement.

PISA Science 2022

```
summary(full_science22)
```

```
Call:
lm(formula = Overall_Science_Score_2022 ~ ., data = science22)

Residuals:
     Min       1Q   Median       3Q      Max
-21.6243  -7.9124   0.8814   7.2924  23.7126
```

```
Coefficients:
                                                                                Estimate
(Intercept)                                                                    5.621e+02
`Urban population (% of total population)`                                     5.876e-01
`Surface area (sq. km)`                                                       -3.520e-05
`Population, total`                                                            1.716e-06
`GDP (current US$)`                                                           -3.198e-11
`Tax revenue (% of GDP)`                                                       5.353e-01
`Unemployment, total (% of total labor force) (modeled ILO estimate)`         -5.167e+00
`Forest area (% of land area)`                                                 6.812e-01
`Gini index`                                                                  -2.257e+00
`GDP per capita (current US$)`                                                 5.452e-04
`Net migration`                                                                9.463e-05
`Life expectancy at birth, total (years)`                                     -1.038e+00
`School enrollment, secondary (% gross)`                                       1.122e-01
`Researchers in R&D (per million people)`                                     -1.697e-03
`High-technology exports (% of manufactured exports)`                          4.356e-01
`Population in the largest city (% of urban population)`                       9.257e-01
`Population ages 0-14 (% of total population)`                                -2.492e+00
                                                                           Std. Error
(Intercept)                                                                    2.122e+02
`Urban population (% of total population)`                                     7.521e-01
`Surface area (sq. km)`                                                        4.877e-05
`Population, total`                                                            8.126e-07
`GDP (current US$)`                                                            1.841e-11
`Tax revenue (% of GDP)`                                                       1.391e+00
`Unemployment, total (% of total labor force) (modeled ILO estimate)`          2.592e+00
`Forest area (% of land area)`                                                 4.583e-01
`Gini index`                                                                   1.522e+00
`GDP per capita (current US$)`                                                 5.568e-04
`Net migration`                                                                9.578e-05
`Life expectancy at birth, total (years)`                                      3.111e+00
`School enrollment, secondary (% gross)`                                       4.922e-01
`Researchers in R&D (per million people)`                                      5.513e-03
`High-technology exports (% of manufactured exports)`                          1.146e+00
`Population in the largest city (% of urban population)`                       5.676e-01
`Population ages 0-14 (% of total population)`                                 3.353e+00
                                                                           t value
(Intercept)                                                                      2.648
`Urban population (% of total population)`                                       0.781
`Surface area (sq. km)`                                                         -0.722
`Population, total`                                                              2.112
`GDP (current US$)`                                                             -1.737
```

```
`Tax revenue (% of GDP)`                                                0.385
`Unemployment, total (% of total labor force) (modeled ILO estimate)`  -1.993
`Forest area (% of land area)`                                          1.487
`Gini index`                                                           -1.483
`GDP per capita (current US$)`                                          0.979
`Net migration`                                                         0.988
`Life expectancy at birth, total (years)`                              -0.334
`School enrollment, secondary (% gross)`                                0.228
`Researchers in R&D (per million people)`                              -0.308
`High-technology exports (% of manufactured exports)`                   0.380
`Population in the largest city (% of urban population)`                1.631
`Population ages 0-14 (% of total population)`                         -0.743
                                                                      Pr(>|t|)
(Intercept)                                                             0.0293
`Urban population (% of total population)`                              0.4571
`Surface area (sq. km)`                                                 0.4910
`Population, total`                                                     0.0677
`GDP (current US$)`                                                     0.1206
`Tax revenue (% of GDP)`                                                0.7103
`Unemployment, total (% of total labor force) (modeled ILO estimate)`   0.0813
`Forest area (% of land area)`                                          0.1754
`Gini index`                                                            0.1763
`GDP per capita (current US$)`                                          0.3562
`Net migration`                                                         0.3521
`Life expectancy at birth, total (years)`                               0.7472
`School enrollment, secondary (% gross)`                                0.8254
`Researchers in R&D (per million people)`                               0.7661
`High-technology exports (% of manufactured exports)`                   0.7139
`Population in the largest city (% of urban population)`                0.1416
`Population ages 0-14 (% of total population)`                          0.4786

(Intercept)                                                            *
`Urban population (% of total population)`
`Surface area (sq. km)`
`Population, total`                                                    .
`GDP (current US$)`
`Tax revenue (% of GDP)`
`Unemployment, total (% of total labor force) (modeled ILO estimate)` .
`Forest area (% of land area)`
`Gini index`
`GDP per capita (current US$)`
`Net migration`
`Life expectancy at birth, total (years)`
```

```
`School enrollment, secondary (% gross)`
`Researchers in R&D (per million people)`
`High-technology exports (% of manufactured exports)`
`Population in the largest city (% of urban population)`
`Population ages 0-14 (% of total population)`
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.62 on 8 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.7537,    Adjusted R-squared:  0.2611
F-statistic:  1.53 on 16 and 8 DF,  p-value: 0.2765
```

Coefficients

- **Intercept (562.1)**: Predicted Overall Science Score in 2022 when all predictors are zero. The intercept is significant (p = 0.0293), suggesting a baseline level of science proficiency.
- **Variables**: **`Population, total`** has a positive effect (1.716e-06), with its p-value (0.0677) suggesting a marginal significance level. This implies that countries with larger populations tend to have higher science scores, albeit the evidence is not strong enough at the conventional 0.05 significance level. This is a pattern that is similar to the 2022 Math result.

Model Fit

- **Adjusted R-squared (0.2611)**: Significantly lower than the Multiple R-squared (0.7537), this metric adjusts for the number of predictors, indicating that many predictors do not significantly contribute to the model. Approximately 26.11% of the variability in Science Scores is explained by the model's predictors.
- **F-statistic (1.53) and p-value (0.2765)**: The F-statistic and associated p-value test the null hypothesis that none of the predictors are significantly related to the response variable. A p-value greater than 0.05 suggests that, collectively, the predictors may not significantly explain the variability in science scores.

Overall

- **Predictor Significance**: Most predictors are not statistically significant, suggesting that they may not have a strong linear relationship with science scores in 2022 or that the model could be missing key explanatory variables or interactions.

Conclusion

From the analysis above, we found that most predictors are not statistically significant, suggesting that they may not have a strong linear relationship with PISA science and math score in 2018 and 2022. It potentially means that out model could be missing key explanatory variables or interactions.

The only variable we found to have some systematic effect is **`Population, total`** in year 2022. The relationship showed that larger countries is more possible to have higher PISA science/math score. It could be due to direct effects (e.g., larger talent pools, more investment in education) or indirect effects (e.g., larger countries may have more varied educational policies, more urbanization which could correlate with better educational facilities).

However, this is important to note the relationship is weak, and our overall model performance is not ideal. A better model structure as well as better variable selection are more important.

Q2: Changes from 2018 to 2022

In this question, we would like to know whether the PISA math score of OECD countries changed between 2018 and 2022. Further, we aimed to understand what country characteristic affect the PISA math score change for year 2018 and 2022 of the OECD countries.

Data preperation

```
# For t-test

# Fill missing value with 0
Math = data.frame(M2009$Country, data_2018_clean$Overall_Math_Score_2018, data_2022_clean$Ov
colnames(Math) <- c('Country','Math2018','Math2022')

# Wide to long
Math_long <- gather(Math, Year, Score, 'Math2018','Math2022')
```

```
# For Regression

# Merge the two datasets on the country identifier
math_scores_combined <- merge(data_2018_clean[, c("Country", "Overall_Math_Score_2018")], da

# Calculate the change in scores
math18$Score_Change = math_scores_combined$Overall_Math_Score_2022 - math_scores_combined$Ov
math22$Score_Change = math_scores_combined$Overall_Math_Score_2022 - math_scores_combined$Ov
```

```r
# Remove unnecessary variable
om18 = c("Overall_Math_Score_2018")
math18 = math18[, !(names(math18) %in% om18)]
om22 = c("Overall_Math_Score_2022")
math22 = math22[, !(names(math22) %in% om22)]
```

t-test

```r
library(rstatix)
```

```
Warning: package 'rstatix' was built under R version 4.3.3
```

```
Attaching package: 'rstatix'
```

```
The following object is masked from 'package:stats':

    filter
```

```r
get_summary_stats(group_by(Math_long, Year), Score, type = "mean_sd")
```

```
# A tibble: 2 x 5
  Year     variable     n  mean    sd
  <chr>    <fct>    <dbl> <dbl> <dbl>
1 Math2018 Score       38  487.  33.5
2 Math2022 Score       37  472.  33.9
```

```r
# Conduct a paired t-test
t_test_results <- t.test(math_scores_combined$Overall_Math_Score_2018, math_scores_combined$

t_test_results
```

```
    Paired t-test

data:  math_scores_combined$Overall_Math_Score_2018 and math_scores_combined$Overall_Math_Sc
t = 9.4893, df = 36, p-value = 2.473e-11
```

```
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 11.59947 17.90537
sample estimates:
mean difference
       14.75242
```

t-test Result

A paired-samples t-test was conducted to compare PISA math score in 2018 and 2022. There was a significant difference in the PISA math score for 2018 (M = 486.982, $SD = 33.53$) and 2022 (M = 427.327, $SD = 33.922$) ; $t(36) = 9.4893$, $p < 0.001$. The PISA math score of OECD countries are significantly decreased from 2018 to 2022. It implied that some major global event (e.g., COVID) may lead to this result.

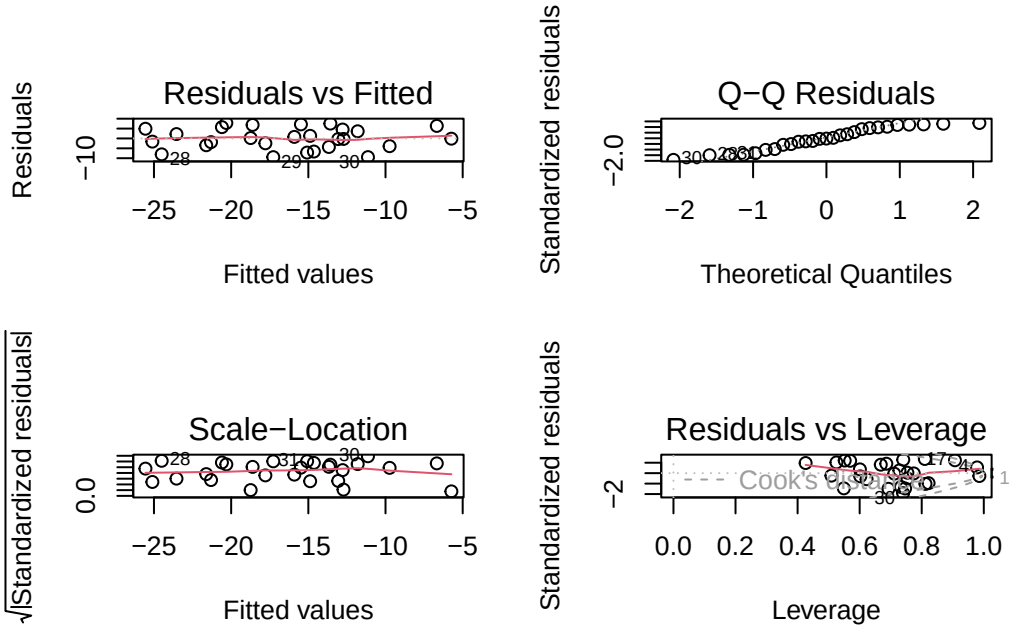
We further utilize regression analysis to explore possible influencing factors.

Assumption check

Math Score Difference with 2018 Features

```
mathdiff18 <- lm(Score_Change ~ ., data = math18)
par(mfrow = c(2,2))
plot(mathdiff18)
```

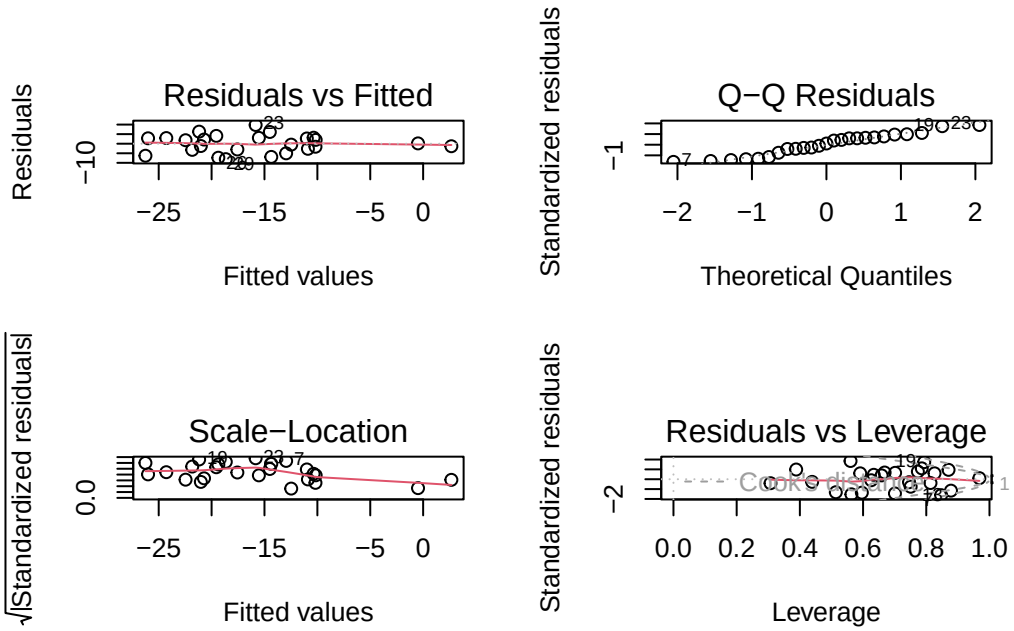
```
Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Math Score Difference with 2022 Features

```
mathdiff22 <- lm(Score_Change ~ ., data = math22)
par(mfrow = c(2,2))
plot(mathdiff22)
```

```
Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Assumption check: Summary

- Both of the models have the similar **assumption results**:
  - **Residuals vs Fitted**. Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, which is good.
  - **Normal Q-Q**. Used to examine whether the residuals are roughly normally distributed. The residuals points roughly follow the straight dashed line, which showed our data follows a normal distribution.
  - **Scale-Location.** Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity. This is not the case in our example, where we have a heteroscedasticity problem. But this is much better than what we saw in Q1 data.
  - **Residuals vs Leverage**. Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. We can see there is no extreme cases in our data. Which is also an improvement in comparison to Q1 data.

Base on the assumption, the data we have in Q2 is much proper to do a linear regression analysis. Which make it easy to interpret and compare the results, given we have so many variables in the dataset.

Linear regression

Math Score Difference with 2018 Features

```
summary(mathdiff18)
```

```
Call:
lm(formula = Score_Change ~ ., data = math18)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4551 -3.6734 -0.1138  4.7144  7.8296

Coefficients:
                                                                        Estimate
(Intercept)                                                            2.356e+01
`Urban population (% of total population)`                            -8.077e-02
`Surface area (sq. km)`                                                3.175e-07
`Population, total`                                                    2.645e-08
`GDP (current US$)`                                                   -5.694e-13
`CO2 emissions (metric tons per capita)`                              -2.750e-01
`Tax revenue (% of GDP)`                                              -1.742e-01
`Unemployment, total (% of total labor force) (modeled ILO estimate)`  1.882e-01
`Forest area (% of land area)`                                        -1.142e-01
`Gini index`                                                          -2.052e-01
`GDP per capita (current US$)`                                        -3.816e-05
`Net migration`                                                        1.080e-05
`Life expectancy at birth, total (years)`                             -4.119e-01
`School enrollment, secondary (% gross)`                               3.399e-02
`Researchers in R&D (per million people)`                             -5.467e-04
`Secure Internet servers (per 1 million people)`                       1.792e-05
`High-technology exports (% of manufactured exports)`                 -3.484e-01
`Population in the largest city (% of urban population)`               2.557e-01
`Population ages 0-14 (% of total population)`                         6.213e-01
                                                                      Std. Error
(Intercept)                                                            1.145e+02
`Urban population (% of total population)`                             3.247e-01
`Surface area (sq. km)`                                                2.206e-06
`Population, total`                                                    1.703e-07
`GDP (current US$)`                                                    3.087e-12
`CO2 emissions (metric tons per capita)`                               1.829e+00
```

```
`Tax revenue (% of GDP)`                                                 8.234e-01
`Unemployment, total (% of total labor force) (modeled ILO estimate)`  8.756e-01
`Forest area (% of land area)`                                           1.784e-01
`Gini index`                                                             1.222e+00
`GDP per capita (current US$)`                                           2.819e-04
`Net migration`                                                          3.377e-05
`Life expectancy at birth, total (years)`                                1.675e+00
`School enrollment, secondary (% gross)`                                 1.848e-01
`Researchers in R&D (per million people)`                                3.812e-03
`Secure Internet servers (per 1 million people)`                         1.222e-04
`High-technology exports (% of manufactured exports)`                    4.578e-01
`Population in the largest city (% of urban population)`                 3.266e-01
`Population ages 0-14 (% of total population)`                           1.588e+00
                                                                        t value
(Intercept)                                                               0.206
`Urban population (% of total population)`                               -0.249
`Surface area (sq. km)`                                                   0.144
`Population, total`                                                       0.155
`GDP (current US$)`                                                      -0.184
`CO2 emissions (metric tons per capita)`                                 -0.150
`Tax revenue (% of GDP)`                                                 -0.212
`Unemployment, total (% of total labor force) (modeled ILO estimate)`     0.215
`Forest area (% of land area)`                                           -0.640
`Gini index`                                                             -0.168
`GDP per capita (current US$)`                                           -0.135
`Net migration`                                                           0.320
`Life expectancy at birth, total (years)`                                -0.246
`School enrollment, secondary (% gross)`                                  0.184
`Researchers in R&D (per million people)`                                -0.143
`Secure Internet servers (per 1 million people)`                          0.147
`High-technology exports (% of manufactured exports)`                    -0.761
`Population in the largest city (% of urban population)`                  0.783
`Population ages 0-14 (% of total population)`                            0.391
                                                                       Pr(>|t|)
(Intercept)                                                               0.842
`Urban population (% of total population)`                                0.810
`Surface area (sq. km)`                                                   0.889
`Population, total`                                                       0.880
`GDP (current US$)`                                                       0.858
`CO2 emissions (metric tons per capita)`                                  0.884
`Tax revenue (% of GDP)`                                                  0.838
`Unemployment, total (% of total labor force) (modeled ILO estimate)`     0.835
`Forest area (% of land area)`                                            0.540
```

```
`Gini index`                                                    0.871
`GDP per capita (current US$)`                                  0.896
`Net migration`                                                 0.757
`Life expectancy at birth, total (years)`                       0.812
`School enrollment, secondary (% gross)`                        0.859
`Researchers in R&D (per million people)`                       0.890
`Secure Internet servers (per 1 million people)`                0.887
`High-technology exports (% of manufactured exports)`           0.469
`Population in the largest city (% of urban population)`        0.456
`Population ages 0-14 (% of total population)`                  0.706

Residual standard error: 9.682 on 8 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.4954,    Adjusted R-squared:  -0.6399
F-statistic: 0.4364 on 18 and 8 DF,  p-value: 0.9313
```

Coefficients

- **Intercept (23.56)**: The expected change in math scores when all predictor variables are held at zero. Its high p-value (0.842) indicates that the intercept is not statistically significant, suggesting that when all other factors are zero, the change in math scores is not significantly different from zero.
- **Variables**: No variables are marked with significance codes, implying none are statistically significant at conventional levels (e.g., 0.05, 0.01).

Model Fit

- **Adjusted R-squared (-0.6399)**: A negative adjusted R-squared indicates that the model fits worse than a horizontal line; in other words, the predictors do not explain the variation in score changes effectively, and the model likely includes too many predictors relative to the number of observations.
- **F-statistic (0.4364) and p-value (0.9313)**: The F-statistic tests the null hypothesis that all regression coefficients are zero. The very high p-value suggests that, collectively, the predictors do not significantly explain the variability in the change in math scores, reinforcing the conclusion drawn from the adjusted R-squared.

Overall

- **Lack of Statistical Significance**: The absence of statistically significant predictors in this model suggests that these factors may not have a direct linear relationship with the change in math scores from 2018, or that the model is over-specified given the data available.

- **Model Complexity**: The negative adjusted R-squared and the overall model's lack of significance indicate an overly complex model for the sample size. This complexity could be contributing to the model's inability to effectively predict changes in math scores.

Math Score Difference with 2022 Features

```
summary(mathdiff22)
```

```
Call:
lm(formula = Score_Change ~ ., data = math22)

Residuals:
   Min     1Q Median     3Q    Max
-8.070 -3.020  0.159  2.952  9.788

Coefficients:
                                                                  Estimate
(Intercept)                                                     -3.003e+01
`Urban population (% of total population)`                      -4.467e-02
`Surface area (sq. km)`                                         -2.421e-05
`Population, total`                                              6.931e-07
`GDP (current US$)`                                             -1.752e-11
`Tax revenue (% of GDP)`                                         1.242e-01
`Unemployment, total (% of total labor force) (modeled ILO estimate)`  3.066e-01
`Forest area (% of land area)`                                   1.117e-01
`Gini index`                                                     5.182e-01
`GDP per capita (current US$)`                                   8.346e-05
`Net migration`                                                  5.018e-05
`Life expectancy at birth, total (years)`                       -7.588e-02
`School enrollment, secondary (% gross)`                        -2.250e-01
`Researchers in R&D (per million people)`                        1.610e-03
`High-technology exports (% of manufactured exports)`            4.921e-01
`Population in the largest city (% of urban population)`         9.310e-02
`Population ages 0-14 (% of total population)`                   2.855e-01
                                                                Std. Error
(Intercept)                                                      8.212e+01
`Urban population (% of total population)`                       2.910e-01
`Surface area (sq. km)`                                          1.887e-05
`Population, total`                                              3.144e-07
`GDP (current US$)`                                              7.125e-12
```

`Tax revenue (% of GDP)`	5.381e-01
`Unemployment, total (% of total labor force) (modeled ILO estimate)`	1.003e+00
`Forest area (% of land area)`	1.773e-01
`Gini index`	5.890e-01
`GDP per capita (current US$)`	2.155e-04
`Net migration`	3.706e-05
`Life expectancy at birth, total (years)`	1.204e+00
`School enrollment, secondary (% gross)`	1.905e-01
`Researchers in R&D (per million people)`	2.133e-03
`High-technology exports (% of manufactured exports)`	4.436e-01
`Population in the largest city (% of urban population)`	2.197e-01
`Population ages 0-14 (% of total population)`	1.298e+00
	t value
(Intercept)	-0.366
`Urban population (% of total population)`	-0.154
`Surface area (sq. km)`	-1.283
`Population, total`	2.204
`GDP (current US$)`	-2.459
`Tax revenue (% of GDP)`	0.231
`Unemployment, total (% of total labor force) (modeled ILO estimate)`	0.306
`Forest area (% of land area)`	0.630
`Gini index`	0.880
`GDP per capita (current US$)`	0.387
`Net migration`	1.354
`Life expectancy at birth, total (years)`	-0.063
`School enrollment, secondary (% gross)`	-1.181
`Researchers in R&D (per million people)`	0.755
`High-technology exports (% of manufactured exports)`	1.109
`Population in the largest city (% of urban population)`	0.424
`Population ages 0-14 (% of total population)`	0.220
	Pr(>|t|)
(Intercept)	0.7241
`Urban population (% of total population)`	0.8818
`Surface area (sq. km)`	0.2355
`Population, total`	0.0586
`GDP (current US$)`	0.0394
`Tax revenue (% of GDP)`	0.8233
`Unemployment, total (% of total labor force) (modeled ILO estimate)`	0.7676
`Forest area (% of land area)`	0.5463
`Gini index`	0.4046
`GDP per capita (current US$)`	0.7086
`Net migration`	0.2128
`Life expectancy at birth, total (years)`	0.9513

```
`School enrollment, secondary (% gross)`                                    0.2713
`Researchers in R&D (per million people)`                                   0.4721
`High-technology exports (% of manufactured exports)`                       0.2996
`Population in the largest city (% of urban population)`                    0.6828
`Population ages 0-14 (% of total population)`                              0.8313

(Intercept)
`Urban population (% of total population)`
`Surface area (sq. km)`
`Population, total`                                                         .
`GDP (current US$)`                                                         *
`Tax revenue (% of GDP)`
`Unemployment, total (% of total labor force) (modeled ILO estimate)`
`Forest area (% of land area)`
`Gini index`
`GDP per capita (current US$)`
`Net migration`
`Life expectancy at birth, total (years)`
`School enrollment, secondary (% gross)`
`Researchers in R&D (per million people)`
`High-technology exports (% of manufactured exports)`
`Population in the largest city (% of urban population)`
`Population ages 0-14 (% of total population)`
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.98 on 8 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.706, Adjusted R-squared:  0.1179
F-statistic: 1.201 on 16 and 8 DF,  p-value: 0.4129
```

Coefficients

- **Intercept**: The intercept (-30.03) represents the expected change in math scores when all predictors are held at zero. The intercept is not statistically significant in this context ($p = 0.7241$), which is typical for models where the intercept doesn't have a meaningful interpretation.
- **Variables**:
  - **`Population, total`** has a positive coefficient (6.931e-07), with a p-value close to significance (0.0586), suggesting a trend where countries with larger populations tend to have a slight increase in math scores, although this effect is marginal.

- `GDP (current US$)` shows a negative coefficient (-1.752e-11), significant at the 0.05 level (p = 0.0394), indicating that an increase in GDP is associated with a decrease in math scores over the period. This finding is counterintuitive and suggests a complex relationship between economic size and educational outcomes that may warrant further investigation.

Model Fit

- **Adjusted R-squared (0.1179)**: Much lower than the Multiple R-squared (0.706), reflecting the penalty for including a large number of predictors relative to the sample size. This suggests that many predictors may not contribute meaningfully to the model. Indicates that approximately 11.79% of the variability in the change in math scores is explained by the included predictors.
- **F-statistic (1.201) and p-value (0.4129)**: The overall model's F-test is not significant, indicating that collectively, the predictors may not significantly explain the variation in score changes across countries.

Overall

- **Complex Relationships**: The significant negative relationship between GDP and score changes suggests that economic factors and educational outcomes may interact in complex ways, potentially mediated by how resources are allocated to education.
- **Marginal Effects**: The marginal significance of `Population, total` highlights the need for careful consideration in interpreting predictors close to significance thresholds, particularly in the context of policy or educational interventions.

Conclusion

From the t-test result, we see PISA math score of OECD countries are significantly decreased from 2018 to 2022. Though we try to use regression method to capture the influencing factors, but we failed to establish a good model to predict the result from selected country characteristics in 2018 and 2022.

We did found some interesting variable in 2022 country characteristics to relate to the change in score. Where we found the significant negative relationship between GDP and score changes. This result changes in math scores challenges simplistic narratives about economic development and educational improvement, indicated that we need to have a further investigation on the impacting factors on GDP and how this can interact with math education performance.

Similar to Q1, we also found marginal effect on total population. This could similarly due to direct effects (e.g., larger talent pools, more investment in education) or indirect effects (e.g., larger countries may have more varied educational policies, more urbanization which could correlate with better educational facilities).

Note: I know non-significant means no relationship, but I really need something to write for the project.

Q3: Characteristics Affecting Scores Above 510 (2009, 2018)

In this question, we aimed to understand what country characteristic affect whether the PISA score is above 510 (math and science) for that designated year (2009 and 2018) for the OECD countries.

Data preparation

```
# For Math scores in 2009
data_2009_clean$High_Math_Score = ifelse(data_2009_clean$Overall_Math_Score_2009 > 510, 1, 0)
S09 = c("Overall_Math_Score_2009", "Overall_Science_Score_2009", "Country", "High_Science_Sc
math09 = data_2009_clean[, !(names(data_2009_clean) %in% S09)]

# For Science scores in 2009
data_2009_clean$High_Science_Score = ifelse(data_2009_clean$Overall_Science_Score_2009 > 510
M09 = c("Overall_Math_Score_2009", "Overall_Science_Score_2009", "Country", "High_Math_Score
science09 = data_2009_clean[, !(names(data_2009_clean) %in% M09)]

# For Math scores in 2018
data_2018_clean$High_Math_Score = ifelse(data_2018_clean$Overall_Math_Score_2018 > 510, 1, 0)
S18 = c("Overall_Math_Score_2018", "Overall_Science_Score_2018", "Country", "High_Science_Sc
math18 = data_2018_clean[, !(names(data_2018_clean) %in% S18)]

# For Science scores in 2018
data_2018_clean$High_Science_Score = ifelse(data_2018_clean$Overall_Science_Score_2018 > 510
M18 = c("Overall_Math_Score_2018", "Overall_Science_Score_2018", "Country", "High_Math_Score
science18 = data_2018_clean[, !(names(data_2018_clean) %in% M18)]
```

Logistic Regression Analysis

PISA Math 2009

```
# Logistic regression for high math scores in 2009
model_math_2009 <- glm(High_Math_Score ~ ., data = math09, family = binomial)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
# To view the summary of each model
summary(model_math_2009)
```

```
Call:
glm(formula = High_Math_Score ~ ., family = binomial, data = math09)

Coefficients:
                                                                                Estimate
(Intercept)                                                                    2.035e+01
`Urban population (% of total population)`                                     1.464e+00
`Surface area (sq. km)`                                                       -6.514e-06
`Population, total`                                                           -1.135e-06
`GDP (current US$)`                                                            4.682e-11
`CO2 emissions (metric tons per capita)`                                       1.566e+01
`Tax revenue (% of GDP)`                                                       1.563e+00
`Unemployment, total (% of total labor force) (modeled ILO estimate)`          4.482e+00
`Forest area (% of land area)`                                                 1.524e+00
`Gini index`                                                                   5.825e+00
`GDP per capita (current US$)`                                                 1.376e-03
`Net migration`                                                               -1.694e-04
`Life expectancy at birth, total (years)`                                     -7.611e+00
`School enrollment, secondary (% gross)`                                       9.086e-02
`Researchers in R&D (per million people)`                                     -1.087e-02
`High-technology exports (% of manufactured exports)`                         -3.023e-01
`Population in the largest city (% of urban population)`                      -2.151e+00
`Population ages 0-14 (% of total population)`                                 3.511e+00
                                                                            Std. Error
(Intercept)                                                                    9.279e+06
`Urban population (% of total population)`                                     2.315e+04
`Surface area (sq. km)`                                                        5.830e-02
`Population, total`                                                            4.858e-02
`GDP (current US$)`                                                            1.309e-06
`CO2 emissions (metric tons per capita)`                                       9.970e+04
`Tax revenue (% of GDP)`                                                       3.706e+04
`Unemployment, total (% of total labor force) (modeled ILO estimate)`          1.096e+05
`Forest area (% of land area)`                                                 1.660e+04
`Gini index`                                                                   1.067e+05
`GDP per capita (current US$)`                                                 1.914e+01
`Net migration`                                                                3.662e+00
`Life expectancy at birth, total (years)`                                      1.438e+05
`School enrollment, secondary (% gross)`                                       2.306e+04
```

```
`Researchers in R&D (per million people)`                                1.461e+02
`High-technology exports (% of manufactured exports)`                    5.969e+04
`Population in the largest city (% of urban population)`                 3.954e+04
`Population ages 0-14 (% of total population)`                           3.688e+05
                                                                        z value
(Intercept)                                                                   0
`Urban population (% of total population)`                                    0
`Surface area (sq. km)`                                                       0
`Population, total`                                                           0
`GDP (current US$)`                                                           0
`CO2 emissions (metric tons per capita)`                                      0
`Tax revenue (% of GDP)`                                                      0
`Unemployment, total (% of total labor force) (modeled ILO estimate)`         0
`Forest area (% of land area)`                                                0
`Gini index`                                                                  0
`GDP per capita (current US$)`                                                0
`Net migration`                                                               0
`Life expectancy at birth, total (years)`                                     0
`School enrollment, secondary (% gross)`                                      0
`Researchers in R&D (per million people)`                                     0
`High-technology exports (% of manufactured exports)`                         0
`Population in the largest city (% of urban population)`                      0
`Population ages 0-14 (% of total population)`                                0
                                                                       Pr(>|z|)
(Intercept)                                                                   1
`Urban population (% of total population)`                                    1
`Surface area (sq. km)`                                                       1
`Population, total`                                                           1
`GDP (current US$)`                                                           1
`CO2 emissions (metric tons per capita)`                                      1
`Tax revenue (% of GDP)`                                                      1
`Unemployment, total (% of total labor force) (modeled ILO estimate)`         1
`Forest area (% of land area)`                                                1
`Gini index`                                                                  1
`GDP per capita (current US$)`                                                1
`Net migration`                                                               1
`Life expectancy at birth, total (years)`                                     1
`School enrollment, secondary (% gross)`                                      1
`Researchers in R&D (per million people)`                                     1
`High-technology exports (% of manufactured exports)`                         1
`Population in the largest city (% of urban population)`                      1
`Population ages 0-14 (% of total population)`                                1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.6402e+01  on 22  degrees of freedom
Residual deviance: 3.0256e-10  on  5  degrees of freedom
  (15 observations deleted due to missingness)
AIC: 36

Number of Fisher Scoring iterations: 25
```

PISA Science 2009

```
# Logistic regression for high science scores in 2009
model_science_2009 <- glm(High_Science_Score ~ ., data = science09, family = binomial)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model_science_2009)
```

```
Call:
glm(formula = High_Science_Score ~ ., family = binomial, data = science09)

Coefficients:
                                                                    Estimate
(Intercept)                                                        3.546e+02
`Urban population (% of total population)`                         1.513e+00
`Surface area (sq. km)`                                           -9.800e-06
`Population, total`                                               -3.337e-09
`GDP (current US$)`                                                1.102e-11
`CO2 emissions (metric tons per capita)`                           1.626e+01
`Tax revenue (% of GDP)`                                           2.560e+00
`Unemployment, total (% of total labor force) (modeled ILO estimate)` 1.623e+00
`Forest area (% of land area)`                                     1.841e+00
`Gini index`                                                       9.414e+00
`GDP per capita (current US$)`                                     1.618e-03
`Net migration`                                                    3.215e-05
`Life expectancy at birth, total (years)`                         -1.359e+01
`School enrollment, secondary (% gross)`                          -2.795e-01
`Researchers in R&D (per million people)`                         -9.428e-03
`High-technology exports (% of manufactured exports)`              3.321e+00
`Population in the largest city (% of urban population)`          -1.747e+00
```

```
`Population ages 0-14 (% of total population)`                          1.042e+00
                                                                       Std. Error
(Intercept)                                                             4.806e+06
`Urban population (% of total population)`                              2.636e+04
`Surface area (sq. km)`                                                 5.738e-02
`Population, total`                                                     4.416e-02
`GDP (current US$)`                                                     1.162e-06
`CO2 emissions (metric tons per capita)`                                7.186e+04
`Tax revenue (% of GDP)`                                                2.580e+04
`Unemployment, total (% of total labor force) (modeled ILO estimate)`   9.546e+04
`Forest area (% of land area)`                                          1.168e+04
`Gini index`                                                            5.275e+04
`GDP per capita (current US$)`                                          1.992e+01
`Net migration`                                                         2.832e+00
`Life expectancy at birth, total (years)`                               7.631e+04
`School enrollment, secondary (% gross)`                                1.481e+04
`Researchers in R&D (per million people)`                               1.063e+02
`High-technology exports (% of manufactured exports)`                   4.713e+04
`Population in the largest city (% of urban population)`                2.523e+04
`Population ages 0-14 (% of total population)`                          1.955e+05
                                                                        z value
(Intercept)                                                                   0
`Urban population (% of total population)`                                    0
`Surface area (sq. km)`                                                       0
`Population, total`                                                           0
`GDP (current US$)`                                                           0
`CO2 emissions (metric tons per capita)`                                      0
`Tax revenue (% of GDP)`                                                      0
`Unemployment, total (% of total labor force) (modeled ILO estimate)`         0
`Forest area (% of land area)`                                                0
`Gini index`                                                                  0
`GDP per capita (current US$)`                                                0
`Net migration`                                                               0
`Life expectancy at birth, total (years)`                                     0
`School enrollment, secondary (% gross)`                                      0
`Researchers in R&D (per million people)`                                     0
`High-technology exports (% of manufactured exports)`                         0
`Population in the largest city (% of urban population)`                      0
`Population ages 0-14 (% of total population)`                                0
                                                                       Pr(>|z|)
(Intercept)                                                                   1
`Urban population (% of total population)`                                    1
`Surface area (sq. km)`                                                       1
```

```
`Population, total`                                                    1
`GDP (current US$)`                                                    1
`CO2 emissions (metric tons per capita)`                               1
`Tax revenue (% of GDP)`                                               1
`Unemployment, total (% of total labor force) (modeled ILO estimate)`  1
`Forest area (% of land area)`                                         1
`Gini index`                                                           1
`GDP per capita (current US$)`                                         1
`Net migration`                                                        1
`Life expectancy at birth, total (years)`                              1
`School enrollment, secondary (% gross)`                               1
`Researchers in R&D (per million people)`                              1
`High-technology exports (% of manufactured exports)`                  1
`Population in the largest city (% of urban population)`               1
`Population ages 0-14 (% of total population)`                         1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.6402e+01  on 22  degrees of freedom
Residual deviance: 3.2989e-10  on  5  degrees of freedom
  (15 observations deleted due to missingness)
AIC: 36

Number of Fisher Scoring iterations: 25
```

PISA Math 2018

```
model_math_2018 <- glm(High_Math_Score ~ ., data = math18, family = binomial)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model_math_2018)
```

```
Call:
glm(formula = High_Math_Score ~ ., family = binomial, data = math18)

Coefficients:
                                                        Estimate
(Intercept)                                            2.816e+02
`Urban population (% of total population)`            -1.387e+00
```

`Surface area (sq. km)`	-8.299e-06
`Population, total`	-5.908e-07
`GDP (current US$)`	-3.184e-12
`CO2 emissions (metric tons per capita)`	1.770e+01
`Tax revenue (% of GDP)`	5.511e+00
`Unemployment, total (% of total labor force) (modeled ILO estimate)`	3.137e+00
`Forest area (% of land area)`	2.435e-01
`Gini index`	8.167e+00
`GDP per capita (current US$)`	1.280e-04
`Net migration`	7.493e-05
`Life expectancy at birth, total (years)`	-9.431e+00
`School enrollment, secondary (% gross)`	2.540e-01
`Researchers in R&D (per million people)`	5.598e-03
`Secure Internet servers (per 1 million people)`	-1.150e-04
`High-technology exports (% of manufactured exports)`	2.552e+00
`Population in the largest city (% of urban population)`	-3.208e+00
`Population ages 0-14 (% of total population)`	2.125e+00
	Std. Error
(Intercept)	7.675e+06
`Urban population (% of total population)`	1.139e+04
`Surface area (sq. km)`	1.069e-01
`Population, total`	7.478e-03
`GDP (current US$)`	1.489e-07
`CO2 emissions (metric tons per capita)`	7.415e+04
`Tax revenue (% of GDP)`	3.991e+04
`Unemployment, total (% of total labor force) (modeled ILO estimate)`	3.901e+04
`Forest area (% of land area)`	1.252e+04
`Gini index`	3.929e+04
`GDP per capita (current US$)`	1.733e+01
`Net migration`	2.701e+00
`Life expectancy at birth, total (years)`	9.818e+04
`School enrollment, secondary (% gross)`	1.102e+04
`Researchers in R&D (per million people)`	1.697e+02
`Secure Internet servers (per 1 million people)`	3.936e+00
`High-technology exports (% of manufactured exports)`	3.602e+04
`Population in the largest city (% of urban population)`	1.321e+04
`Population ages 0-14 (% of total population)`	1.209e+05
	z value
(Intercept)	0
`Urban population (% of total population)`	0
`Surface area (sq. km)`	0
`Population, total`	0
`GDP (current US$)`	0

```
`CO2 emissions (metric tons per capita)`                                    0
`Tax revenue (% of GDP)`                                                    0
`Unemployment, total (% of total labor force) (modeled ILO estimate)`       0
`Forest area (% of land area)`                                              0
`Gini index`                                                                0
`GDP per capita (current US$)`                                              0
`Net migration`                                                             0
`Life expectancy at birth, total (years)`                                   0
`School enrollment, secondary (% gross)`                                    0
`Researchers in R&D (per million people)`                                   0
`Secure Internet servers (per 1 million people)`                            0
`High-technology exports (% of manufactured exports)`                       0
`Population in the largest city (% of urban population)`                    0
`Population ages 0-14 (% of total population)`                              0
                                                                    Pr(>|z|)
(Intercept)                                                                 1
`Urban population (% of total population)`                                  1
`Surface area (sq. km)`                                                     1
`Population, total`                                                         1
`GDP (current US$)`                                                         1
`CO2 emissions (metric tons per capita)`                                    1
`Tax revenue (% of GDP)`                                                    1
`Unemployment, total (% of total labor force) (modeled ILO estimate)`       1
`Forest area (% of land area)`                                              1
`Gini index`                                                                1
`GDP per capita (current US$)`                                              1
`Net migration`                                                             1
`Life expectancy at birth, total (years)`                                   1
`School enrollment, secondary (% gross)`                                    1
`Researchers in R&D (per million people)`                                   1
`Secure Internet servers (per 1 million people)`                            1
`High-technology exports (% of manufactured exports)`                       1
`Population in the largest city (% of urban population)`                    1
`Population ages 0-14 (% of total population)`                              1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.2652e+01  on 26  degrees of freedom
Residual deviance: 2.9529e-10  on  8  degrees of freedom
  (11 observations deleted due to missingness)
AIC: 38

Number of Fisher Scoring iterations: 25
```

PISA Science 2018

```
model_science_2018 <- glm(High_Science_Score ~ ., data = science18, family = binomial)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model_science_2018)
```

```
Call:
glm(formula = High_Science_Score ~ ., family = binomial, data = science18)

Coefficients:
                                                                  Estimate
(Intercept)                                                     -4.144e+01
`Urban population (% of total population)`                      -1.525e+00
`Surface area (sq. km)`                                         -1.863e-06
`Population, total`                                              2.312e-07
`GDP (current US$)`                                             -7.871e-12
`CO2 emissions (metric tons per capita)`                         1.090e+01
`Tax revenue (% of GDP)`                                         1.871e+00
`Unemployment, total (% of total labor force) (modeled ILO estimate)`  1.852e+00
`Forest area (% of land area)`                                   6.740e-01
`Gini index`                                                     4.133e+00
`GDP per capita (current US$)`                                  -8.420e-04
`Net migration`                                                  1.726e-05
`Life expectancy at birth, total (years)`                       -2.793e+00
`School enrollment, secondary (% gross)`                         1.261e+00
`Researchers in R&D (per million people)`                        5.404e-03
`Secure Internet servers (per 1 million people)`                 2.950e-04
`High-technology exports (% of manufactured exports)`            6.247e-01
`Population in the largest city (% of urban population)`        -1.084e+00
`Population ages 0-14 (% of total population)`                  -3.242e+00
                                                                Std. Error
(Intercept)                                                      6.647e+06
`Urban population (% of total population)`                       8.988e+03
`Surface area (sq. km)`                                          1.327e-01
`Population, total`                                              1.316e-02
`GDP (current US$)`                                              1.776e-07
`CO2 emissions (metric tons per capita)`                         1.119e+05
`Tax revenue (% of GDP)`                                         3.557e+04
```

```
`Unemployment, total (% of total labor force) (modeled ILO estimate)`  4.284e+04
`Forest area (% of land area)`                                         6.380e+03
`Gini index`                                                           7.089e+04
`GDP per capita (current US$)`                                         1.522e+01
`Net migration`                                                        1.943e+00
`Life expectancy at birth, total (years)`                              6.198e+04
`School enrollment, secondary (% gross)`                               8.803e+03
`Researchers in R&D (per million people)`                              1.626e+02
`Secure Internet servers (per 1 million people)`                       6.625e+00
`High-technology exports (% of manufactured exports)`                  4.399e+04
`Population in the largest city (% of urban population)`               1.121e+04
`Population ages 0-14 (% of total population)`                         9.498e+04
                                                                       z value
(Intercept)                                                                  0
`Urban population (% of total population)`                                   0
`Surface area (sq. km)`                                                      0
`Population, total`                                                          0
`GDP (current US$)`                                                          0
`CO2 emissions (metric tons per capita)`                                     0
`Tax revenue (% of GDP)`                                                     0
`Unemployment, total (% of total labor force) (modeled ILO estimate)`        0
`Forest area (% of land area)`                                               0
`Gini index`                                                                 0
`GDP per capita (current US$)`                                               0
`Net migration`                                                              0
`Life expectancy at birth, total (years)`                                    0
`School enrollment, secondary (% gross)`                                     0
`Researchers in R&D (per million people)`                                    0
`Secure Internet servers (per 1 million people)`                             0
`High-technology exports (% of manufactured exports)`                        0
`Population in the largest city (% of urban population)`                     0
`Population ages 0-14 (% of total population)`                               0
                                                                       Pr(>|z|)
(Intercept)                                                                  1
`Urban population (% of total population)`                                   1
`Surface area (sq. km)`                                                      1
`Population, total`                                                          1
`GDP (current US$)`                                                          1
`CO2 emissions (metric tons per capita)`                                     1
`Tax revenue (% of GDP)`                                                     1
`Unemployment, total (% of total labor force) (modeled ILO estimate)`        1
`Forest area (% of land area)`                                               1
`Gini index`                                                                 1
```

```
`GDP per capita (current US$)`                                          1
`Net migration`                                                         1
`Life expectancy at birth, total (years)`                               1
`School enrollment, secondary (% gross)`                                1
`Researchers in R&D (per million people)`                               1
`Secure Internet servers (per 1 million people)`                        1
`High-technology exports (% of manufactured exports)`                   1
`Population in the largest city (% of urban population)`                1
`Population ages 0-14 (% of total population)`                          1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.2652e+01  on 26  degrees of freedom
Residual deviance: 2.9265e-10  on  8  degrees of freedom
  (11 observations deleted due to missingness)
AIC: 38

Number of Fisher Scoring iterations: 25
```

Conclusion

We found the same result for ALL models. All model summaries have the same waring "glm.fit: fitted probabilities numerically 0 or 1 occurred". The warning means that when R is computing probabilities internally, as part of the fitting process, they sometimes "underflow/overflow" - that is, they're so close to 0 or 1 that they can't be distinguished from them when using R's standard 64-bit floating-point precision (e.g. values less than about 1e-308 or greater than about 1-1e-16).

Coefficients

- **Z-value and P-value**: The z-values are near zero, and the p-values are 1 for all predictors, indicating no statistical significance. This unusual result, with p-values exactly equal to 1, suggests potential issues with the data or model specification.

Overall

1. **Large Standard Errors and P-values**: The large standard errors and corresponding p-values of 1 across all predictors might indicate issues such as perfect separation in logistic regression, where one or more predictors perfectly predict the outcome, leading to infinite estimates. This situation often arises in datasets with small sample sizes or when the outcome variable has limited variability.
2. **Model Complexity**: Given the small number of observations and large number of predictor, the model may be too complex, leading to overfitting.

In conclusion, we failed to identify the influencing factor that related to high math/science performance for OECD countries in 2009 and 2018 from the given variables. This could due to the fact that some defining variables are missing.