Data Visualisation for Economics

Analysis taxes with OECD, IMF and EUROSTAT datasets

Lucas Simatovic Master 2 : Economie option DADEE Assigment 2

novembre 02, 2025

Abstract

Through this exercice, we are going to analyse the tax revenue accross OECD countries and also different dataset like the OECD obviously, IMF and EUROSTAT. We start by cleaning different datasets provide, then construct different types of graphs with OECD dataset. I put some annotations and I refine visualisations. Next, I'll try to replicate graphs from task 1 with IMF and EUROSTAT datas. And finally, I present a contradictory story from OECD datas.

Import of librairies

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(readxl)
library(scales)
library(patchwork)
library(ggrepel)
```

CLEANING OECD DATASET

```
oecd_raw <- read_excel("TAXES.xlsx", sheet = "Table", skip = 5)
names(oecd_raw)[1:3] <- c("Country", "Tax_category", "Tax_code")
oecd_raw <- oecd_raw %>% select(-17)
names(oecd_raw)[4:16] <- as.character(2010:2022)

oecd_long <- oecd_raw %>%
  pivot_longer(cols = `2010`:`2022`, names_to = "Year", values_to = "Value") %>%
  mutate(
    Tax_category = str_squish(str_remove_all(Tax_category, "·")),
    Year = as.numeric(Year),
    Value = as.numeric(Value)
  ) %>%
  filter(!is.na(Value), !is.na(Country))

oecd_long
```

```
## # A tibble: 4,446 x 5
##    Country   Tax_category      Tax_code  Year Value
##    <chr>     <chr>             <chr>    <dbl> <dbl>
##  1 Australia Total tax revenue TOTALTAX  2010  25.1
##  2 Australia Total tax revenue TOTALTAX  2011  25.8
##  3 Australia Total tax revenue TOTALTAX  2012  26.8
##  4 Australia Total tax revenue TOTALTAX  2013  27.0
##  5 Australia Total tax revenue TOTALTAX  2014  27.2
##  6 Australia Total tax revenue TOTALTAX  2015  27.8
##  7 Australia Total tax revenue TOTALTAX  2016  27.5
##  8 Australia Total tax revenue TOTALTAX  2017  28.5
##  9 Australia Total tax revenue TOTALTAX  2018  28.6
## 10 Australia Total tax revenue TOTALTAX  2019  27.7
## # i 4,436 more rows
```

CLEANING IMF DATASET

I could not find from IMF a better table, so as you can see Russian Federation appears but i clean it and only keep OECD countries.

```
imf_raw <- read_csv("imf1.csv")

imf_long <- imf_raw %>%
  select(COUNTRY, INDICATOR, `2010`:`2022`) %>%
  pivot_longer(cols = `2010`:`2022`, names_to = "Year", values_to = "Value") %>%
  mutate(
    Year = as.numeric(Year),
    Value = as.numeric(Value)
  ) %>%
  filter(!is.na(Value), !is.na(COUNTRY))

imf_long
```

```
## # A tibble: 1,677 x 4
##    COUNTRY            INDICATOR                                        Year  Value
##    <chr>              <chr>                                           <dbl>  <dbl>
##  1 Russian Federation Taxes on income, profits, and capital gains,~  2018  8070.
##  2 Russian Federation Taxes on income, profits, and capital gains,~  2019  8761.
##  3 Russian Federation Taxes on income, profits, and capital gains,~  2020  8517.
##  4 Russian Federation Taxes on international trade and transaction~  2018  3700.
##  5 Russian Federation Taxes on international trade and transaction~  2019  3049.
##  6 Russian Federation Taxes on international trade and transaction~  2020  1872.
##  7 Russian Federation Revenue, Transactions                          2018 39982.
##  8 Russian Federation Revenue, Transactions                          2019 43543.
##  9 Russian Federation Revenue, Transactions                          2020 44514.
## 10 Russian Federation Taxes on payroll and workforce, Transactions   2018     0
## # i 1,667 more rows
```

CLEANING GDP DATASET

```
gdp_raw <- read_csv("gdp_growth.csv",
                    col_types = cols(
                      `Reference area` = col_character(),
                      .default = col_double()
                    ))

gdp_raw <- gdp_raw %>%
  filter(!is.na(`Reference area`),
         `Reference area` != "© Terms & conditions")

names(gdp_raw)[1] <- "Country"
names(gdp_raw)[2] <- "2009"
names(gdp_raw)[3:15] <- as.character(2010:2022)

gdp_raw <- gdp_raw %>% select(-16)

gdp_long <- gdp_raw %>%
  pivot_longer(cols = `2009`:`2022`,
               names_to = "Year",
               values_to = "GDP_growth") %>%
  mutate(Year = as.numeric(Year)) %>%
  filter(Year >= 2010)

gdp_long
```

```
## # A tibble: 494 x 3
##    Country    Year GDP_growth
##    <chr>     <dbl>      <dbl>
##  1 Australia  2010       2.41
##  2 Australia  2011       3.92
##  3 Australia  2012       2.60
##  4 Australia  2013       2.60
##  5 Australia  2014       2.19
##  6 Australia  2015       2.76
##  7 Australia  2016       2.28
##  8 Australia  2017       2.88
##  9 Australia  2018       2.17
## 10 Australia  2019      -0.120
## # i 484 more rows
```

TASK 1

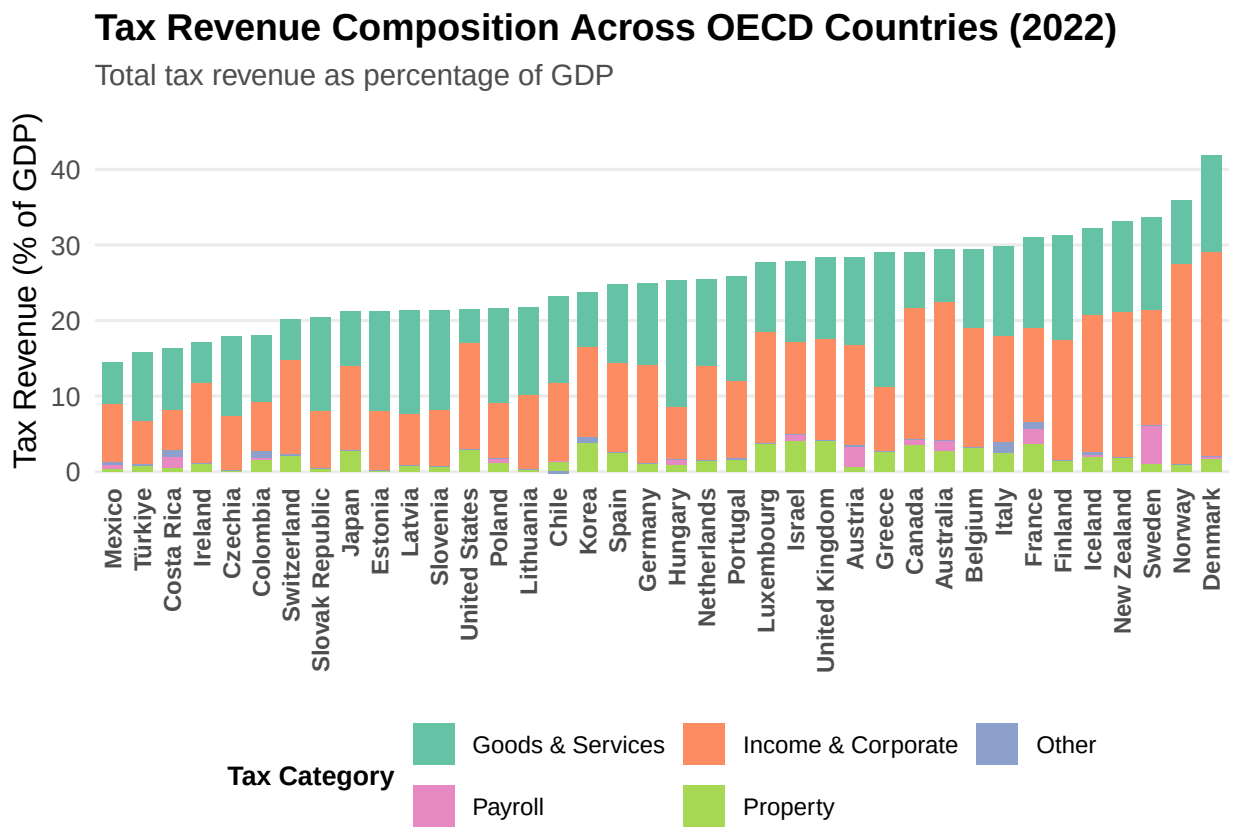
In this part we analyse the OECD datas which provide explanations about the compostion of taxes revenues, the trend of corporation taxes across years and the relationship between GDP growth and Tax accross countries.

TASK 1A

Firstly, we start with the compostion of the different taxes in OECD countries and a second graph shows the composition between Income and corporate taxes in the 38 OECD countries.

```
tax_categories_main <- c(
  "Taxes on income, profits and capital gains of individuals and corporations",
  "Taxes on payroll and workforce",
  "Taxes on property",
  "Taxes on goods and services",
  "Other taxes"
)

tax_labels <- c(
  "Taxes on income, profits and capital gains of individuals and corporations"
  = "Income & Corporate",
  "Taxes on payroll and workforce" = "Payroll",
  "Taxes on property" = "Property",
  "Taxes on goods and services" = "Goods & Services",
  "Other taxes" = "Other"
)

data_task1a <- oecd_long %>%
  filter(
    Year == 2022,
    Tax_category %in% tax_categories_main
  ) %>%
  mutate(Tax_category_short = recode(Tax_category, !!!tax_labels))

totals <- data_task1a %>%
  group_by(Country) %>%
  summarise(Total = sum(Value, na.rm = TRUE))

data_task1a <- data_task1a %>%
  left_join(totals, by = "Country")

plot_1a <- ggplot(data_task1a,
                  aes(x = reorder(Country, Total),
                      y = Value,
                      fill = Tax_category_short)) +
  geom_col(position = 'stack', width = 0.7) +
  labs(
    title = "Tax Revenue Composition Across OECD Countries (2022)",
    subtitle = "Total tax revenue as percentage of GDP",
    x = NULL,
    y = "Tax Revenue (% of GDP)",
    fill = "Tax Category",
  ) +
  scale_fill_brewer(palette = "Set2") +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14, hjust = 0),
    plot.subtitle = element_text(size = 11, color = "gray30", hjust = 0),
```

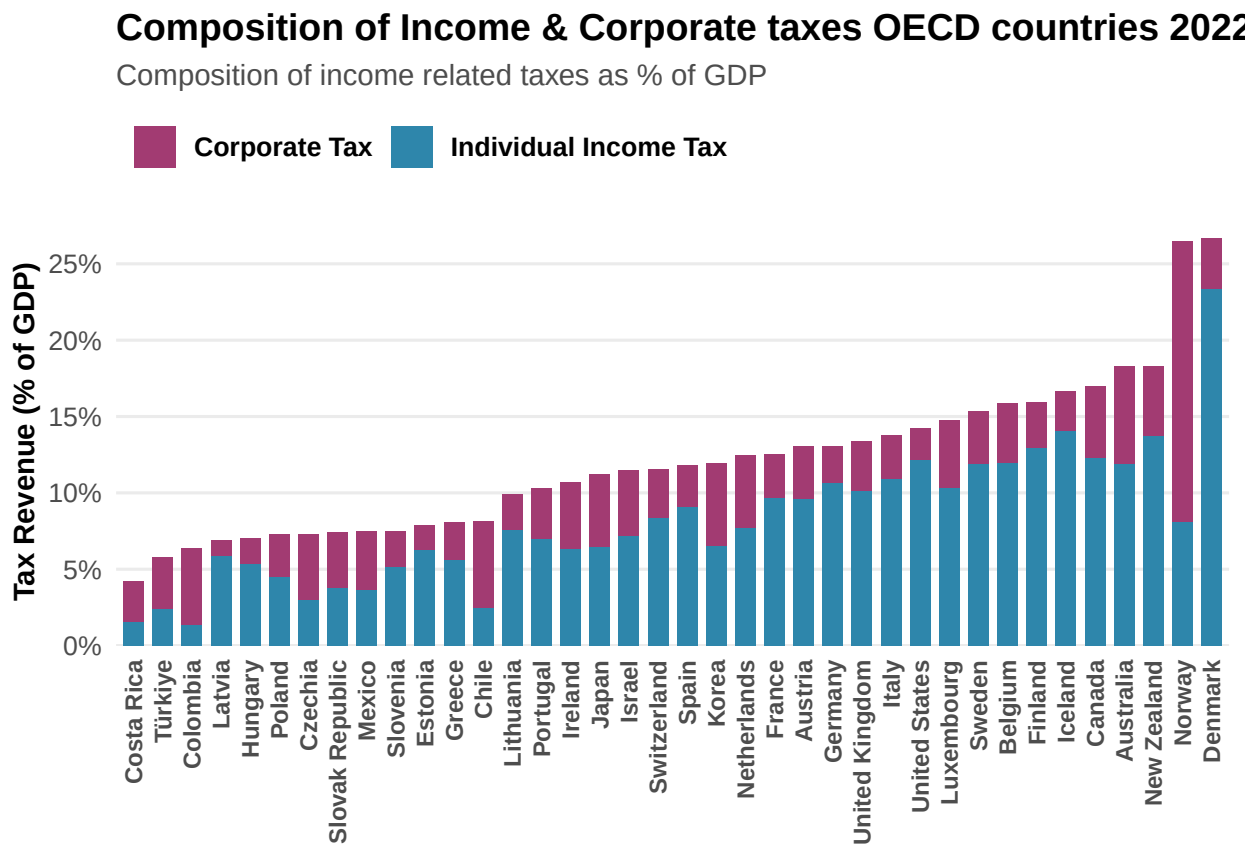
```
    axis.text.x = element_text(
      size = 9, face = "bold", angle = 90, vjust = 0.5, hjust = 1),
    axis.text.y = element_text(size = 10),
    legend.position = "bottom",
    legend.title = element_text(face = "bold", size = 10),
    legend.text = element_text(size = 9),
    panel.grid.major.x = element_blank(),
    panel.grid.minor = element_blank(),
    plot.caption = element_text(hjust = 0, size = 8, color = "gray50")
  ) +
  guides(fill = guide_legend(nrow = 2, byrow = TRUE))

print(plot_1a)
```




```
tax_categories_income_corp <- c(
  "Taxes on income, profits and capital gains of individuals",
  "Taxes on income, profits and capital gains of corporations"
)

tax_labels_income_corp <- c(
  "Taxes on income, profits and capital gains of individuals" =
    "Individual Income Tax",
  "Taxes on income, profits and capital gains of corporations" =
    "Corporate Tax"
)
```

```r
data_income_corp <- oecd_long %>%
  filter(
    Year == 2022,
    Tax_category %in% tax_categories_income_corp
  ) %>%
  mutate(Tax_category_short = recode(Tax_category, !!!tax_labels_income_corp))

totals_income_corp <- data_income_corp %>%
  group_by(Country) %>%
  summarise(Total_Income_Corp = sum(Value, na.rm = TRUE))

data_income_corp <- data_income_corp %>%
  left_join(totals_income_corp, by = "Country")

plot_income_corp_stacked <- ggplot(data_income_corp,
                                   aes(x = reorder(Country, Total_Income_Corp),
                                       y = Value,
                                       fill = Tax_category_short)) +
  geom_col(position = 'stack', width = 0.7) +
  labs(
    title = "Composition of Income & Corporate taxes OECD countries 2022",
    subtitle = "Composition of income related taxes as % of GDP",
    x = NULL,
    y = "Tax Revenue (% of GDP)",
    fill = NULL,
  ) +
  scale_fill_manual(
    values = c(
      "Individual Income Tax" = "#2E86AB",
      "Corporate Tax" = "#A23B72"
    )
  ) +
  scale_y_continuous(
    expand = expansion(mult = c(0, 0.05)),
    labels = label_number(suffix = "%"),
    breaks = seq(0, 30, 5)
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14, hjust = 0),
    plot.subtitle = element_text(size = 11, color = "gray30", hjust = 0),
    plot.caption = element_text(size = 9, color = "gray50", hjust = 0),
    axis.text.x = element_text(size = 9, face = "bold",
                               angle = 90, vjust = 0.5, hjust = 1),
    axis.text.y = element_text(size = 10),
    axis.title.y = element_text(face = "bold", size = 11),
    legend.position = "top",
    legend.title = element_blank(),
    legend.text = element_text(size = 10, face = "bold"),
    legend.justification = "left",
    panel.grid.major.x = element_blank(),
    panel.grid.minor = element_blank()
  )
```

```
print(plot_income_corp_stacked)
```



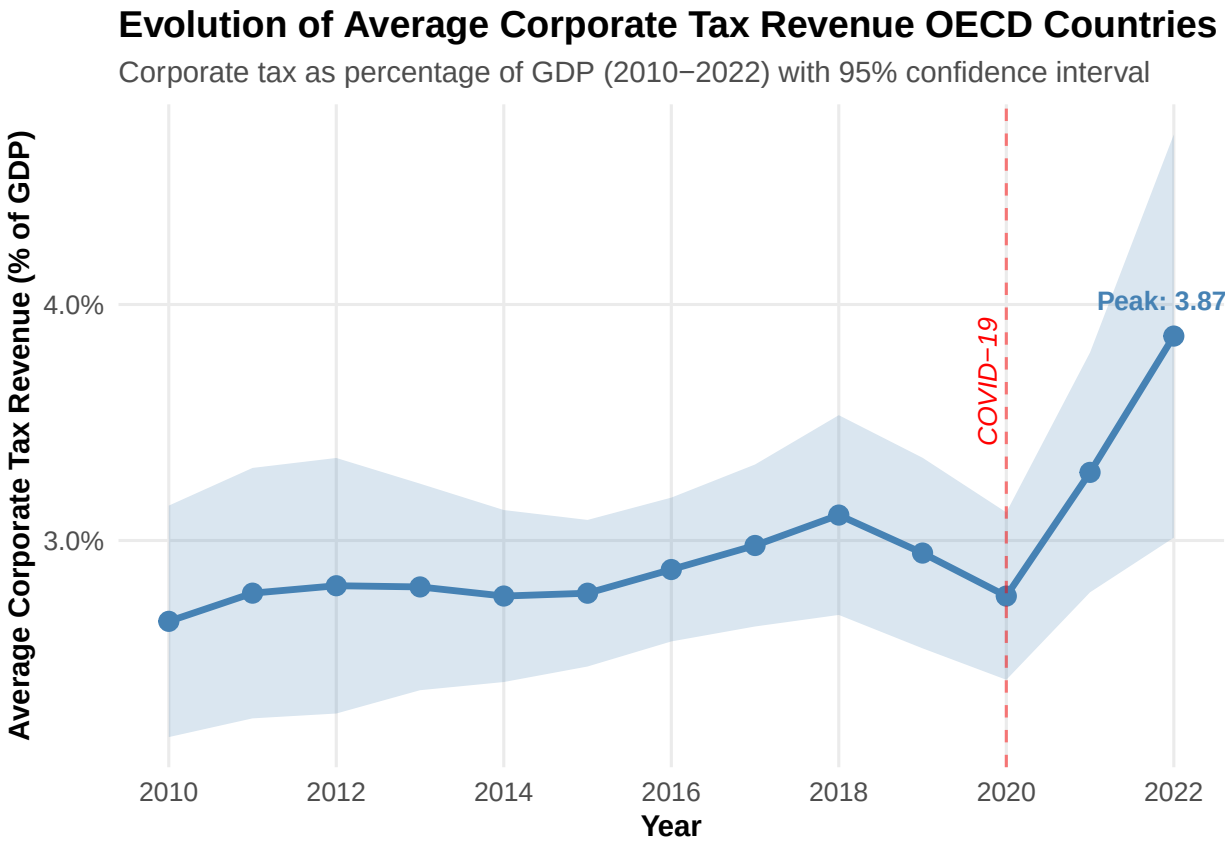
Why I choose this type of graph? Simply it effectively illustrates the compostion of taxes in countries, shows the fiscal structure acrross the different types of taxes and compare countries between us. Denmark and Norway tell us that they have a strong taxation on Income and Corporate that demonstarte they have a strong welfare-state financing model.

TASK 1B

We are going to represent the evolution of the average corporate tax revenue accross OECD countries and years.

```
data_task1b <- oecd_long %>%
  filter(
    Tax_category == "Taxes on income, profits and capital gains of corporations",
    Year >= 2010,
    Year <= 2022
  ) %>%
  group_by(Year) %>%
  summarise(
    Mean_Corp_Tax = mean(Value, na.rm = TRUE),
    SD = sd(Value, na.rm = TRUE),
```

```
    N = n()
  ) %>%
  mutate(
    SE = SD / sqrt(N),
    CI_lower = Mean_Corp_Tax - 1.96 * SE,
    CI_upper = Mean_Corp_Tax + 1.96 * SE
  )

peak_year <- data_task1b %>%
  filter(Mean_Corp_Tax == max(Mean_Corp_Tax)) %>%
  pull(Year)

peak_value <- data_task1b %>%
  filter(Year == peak_year) %>%
  pull(Mean_Corp_Tax)

plot_1b <- ggplot(data_task1b, aes(x = Year, y = Mean_Corp_Tax)) +
  geom_ribbon(aes(ymin = CI_lower, ymax = CI_upper),
              fill = "steelblue", alpha = 0.2) +
  geom_line(color = "steelblue", size = 1.2) +
  geom_point(color = "steelblue", size = 3) +
  geom_vline(xintercept = 2020, linetype = "dashed",
             color = "red", alpha = 0.5) +
  annotate("text", x = 2020, y = max(data_task1b$Mean_Corp_Tax) * 0.95,
           label = "COVID-19", angle = 90, vjust = -0.5,
           color = "red", size = 3.5, fontface = "italic") +
  annotate("text", x = peak_year, y = peak_value + 0.15,
           label = paste0("Peak: ", round(peak_value, 2), "%"),
           color = "steelblue", size = 3.5, fontface = "bold") +
  labs(
    title = "Evolution of Average Corporate Tax Revenue OECD Countries",
    subtitle = "Corporate tax as percentage of GDP (2010-2022) with 95% confidence interval",
    x = "Year",
    y = "Average Corporate Tax Revenue (% of GDP)",
  ) +
  scale_x_continuous(breaks = seq(2010, 2022, 2)) +
  scale_y_continuous(labels = label_number(suffix = "%", accuracy = 0.1)) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 11, color = "gray30"),
    axis.title = element_text(face = "bold", size = 11),
    panel.grid.minor = element_blank(),
    plot.caption = element_text(hjust = 0, size = 8, color = "gray50")
  )

print(plot_1b)
```



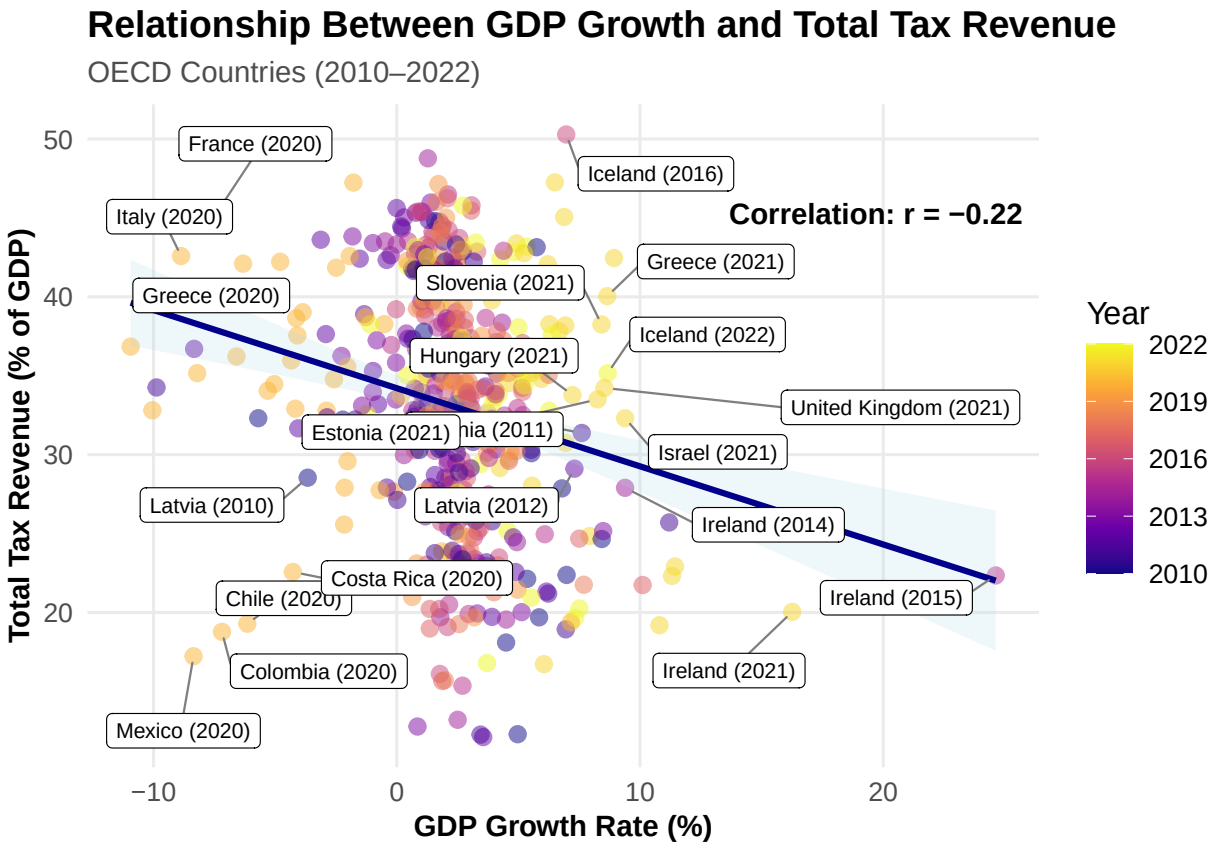
This graph shows us the evolution of corporate tax revenue. It illustrates the trend accross countries. It shows that, in 95% confidence interval (outliers are rejected), after COVID-19 there is a huge increase in corporation taxation in average from 2.8% in 2010 and in 2022 a peak at 3.87% in percentage of GDP, reflecting post-pandemic recovery policies and increased corporate profits.

TASK 1C

Finally, with OECD datas, we present a scatter plot which shows the relashionship between GDP growth and Tax revenue

```
data_task1c <- oecd_long %>%
  filter(Tax_category == "Total tax revenue") %>%
  left_join(gdp_long, by = c("Country", "Year")) %>%
  filter(!is.na(GDP_growth), !is.na(Value))

correlation <- cor(data_task1c$GDP_growth, data_task1c$Value, use = "complete.obs")

outliers <- data_task1c %>%
  filter(GDP_growth < quantile(GDP_growth, 0.05) |
           GDP_growth > quantile(GDP_growth, 0.95) |
           Value < quantile(Value, 0.05) |
           Value > quantile(Value, 0.95))
```

```r
plot_1c <- ggplot(data_task1c, aes(x = GDP_growth, y = Value)) +
  geom_point(aes(color = Year), alpha = 0.5, size = 2.5) +
  geom_smooth(method = "lm", se = TRUE, color = "darkblue",
              fill = "lightblue", alpha = 0.2) +
  ggrepel::geom_label_repel(
    data = outliers,
    aes(label = paste0(Country, " (", Year, ")")),
    size = 2.9,
    box.padding = 0.5,
    point.padding = 0.3,
    max.overlaps = 12,
    segment.color = "gray50",
    segment.size = 0.4,
    min.segment.length = 0,
    color = "black",
    fill = "white",
    label.size = 0.2
  ) +
  annotate("text",
         x = max(data_task1c$GDP_growth) * 0.8,
         y = max(data_task1c$Value) * 0.9,
         label = paste0("Correlation: r = ", round(correlation, 2)),
         color = "black",
         size = 4,
         fontface = "bold") +
  labs(
    title = "Relationship Between GDP Growth and Total Tax Revenue",
    subtitle = "OECD Countries (2010-2022)",
    x = "GDP Growth Rate (%)",
    y = "Total Tax Revenue (% of GDP)",
    color = "Year",
    ) +
  scale_color_viridis_c(option = "plasma") +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 11, color = "gray30"),
    axis.title = element_text(face = "bold", size = 11),
    legend.position = "right",
    panel.grid.minor = element_blank(),
    plot.caption = element_text(hjust = 0, size = 8, color = "gray50")
  )

print(plot_1c)
```



This scatter plot is mandatory to explain the relationship between GDP and taxes. We can deduce from this graph that there is a negative correlation (-0.22). I try to put some outliers names like Ireland in 2015 (which have a low total tax revenue and high GDP growth compared to other countries) and countries who present more difficulties during COVID-19 like Italy. High concentration with a postive but low GDP growth (0-5%) and a total tax revenue (25-40%) in percentage of GDP.

Overall, these three visualizations provide complementary insights into OECD fiscal structures. While tax composition highlights structural diversity, the time series of corporate taxes shows recent upward trends, and the GDP–tax scatter plot reveals the weak inverse relationship between economic growth and tax pressure.

TASK 2

For each visualisation, I've directly implemented vertical x axis for each, use "plasma" palette type in the relashionship to showing the scale (lower (2010) = blue to higher (2022) = yellow) and in terms of annotations with the correlation result. The principal event is the COVID 19 who's the most important shock that can be noted. And so the graph B shows us the increase of corporate tax revenues for example in this case also can be applicable for an other tax type and show an increase too.

TASK 3

Here, we start analysing the difference between IMF and OECD datas and then EUROSTAT and OECD datas in a same that in the task 1.

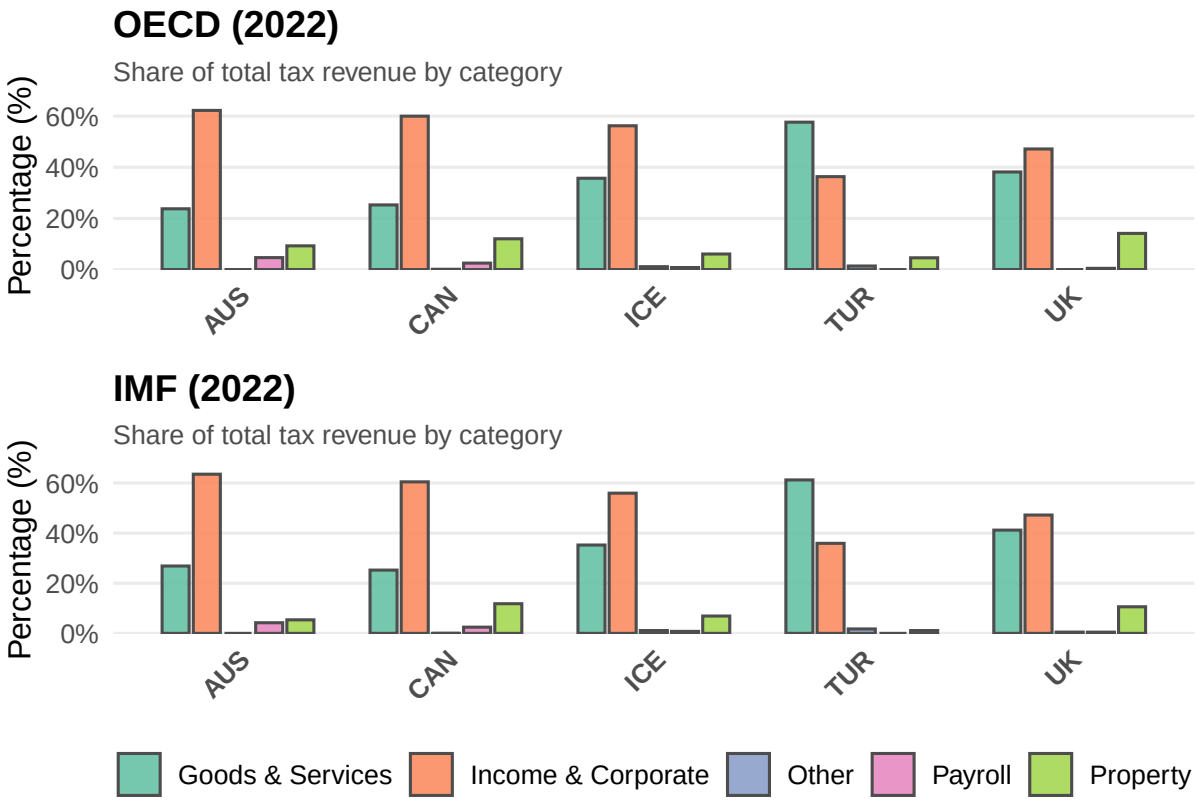
TASK 3A

We begin to compare if the final graphs differ because of the datas. I select 5 countries: Australia, Canada, Icelande, Turkeye and United Kingdom.

```
country_mapping <- c(
  "AUS" = "Australia",
  "CAN" = "Canada",
  "ICE" = "Iceland",
  "TUR" = "Türkiye",
  "UK"  = "United Kingdom"
)

country_mapping_imf <- c(
  "AUS" = "Australia",
  "CAN" = "Canada",
  "ICE" = "Iceland",
  "TUR" = "Türkiye, Republic of",
  "UK"  = "United Kingdom"
)

imf_categories <- c(
  "Taxes on income, profits, and capital gains, Transactions" =
    "Income & Corporate",
  "Taxes on payroll and workforce, Transactions" = "Payroll",
  "Taxes on property, Transactions" = "Property",
  "Taxes on goods and services, Transactions" = "Goods & Services",
  "Other taxes, Transactions" = "Other"
)

oecd_5pays <- oecd_long %>%
  filter(
    Country %in% country_mapping,
    Tax_category %in% tax_categories_main,
    Year == 2022
  ) %>%
  mutate(
    Country = names(country_mapping)[match(Country, country_mapping)],
    Tax_category_short = recode(Tax_category, !!!tax_labels),
    Source = "OECD"
  ) %>%
  group_by(Country) %>%
  mutate(Percentage = Value / sum(Value) * 100) %>%
  ungroup()

imf_5pays <- imf_long %>%
  filter(
```

```
    COUNTRY %in% country_mapping_imf,
    INDICATOR %in% names(imf_categories),
    Year == 2022
  ) %>%
  mutate(
    Country = names(country_mapping_imf)[match(COUNTRY, country_mapping_imf)],
    Tax_category_short = recode(INDICATOR, !!!imf_categories),
    Source = "IMF"
  ) %>%
  group_by(Country) %>%
  mutate(Percentage = Value / sum(Value) * 100) %>%
  ungroup()

country_order <- c("AUS", "CAN", "ICE", "TUR", "UK")

for(df in list(oecd_5pays, imf_5pays)){
  df$Country <- factor(df$Country, levels = country_order)
  df$Tax_category_short <- factor(
    df$Tax_category_short,
    levels = c("Income & Corporate", "Payroll",
               "Property", "Goods & Services", "Other"))
}

plot_oecd <- ggplot(oecd_5pays,
                    aes(x = Country, y = Percentage,
                        fill = Tax_category_short)) +
  geom_col(position = position_dodge(width = 0.75), width = 0.65,
           color = "gray30", alpha = 0.9) +
  labs(
    title = "OECD (2022)",
    subtitle = "Share of total tax revenue by category",
    x = NULL, y = "Percentage (%)", fill = "Tax Category"
  ) +
  scale_fill_brewer(palette = "Set2") +
  scale_y_continuous(labels = label_number(suffix = "%", accuracy = 1),
                     expand = expansion(mult = c(0, 0.05))) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14, hjust = 0),
    plot.subtitle = element_text(size = 10, color = "gray30", hjust = 0),
    axis.text.x = element_text(angle = 45, hjust = 1,
                               face = "bold", size = 10),
    legend.position = "bottom",
    legend.title = element_blank(),
    panel.grid.major.x = element_blank(),
    panel.grid.minor = element_blank()
  )

plot_imf <- ggplot(imf_5pays,
                   aes(x = Country, y = Percentage,
                       fill = Tax_category_short)) +
  geom_col(position = position_dodge(width = 0.75), width = 0.65,
           color = "gray30", alpha = 0.9) +
```

```
  labs(
    title = "IMF (2022)",
    subtitle = "Share of total tax revenue by category",
    x = NULL, y = "Percentage (%)", fill = "Tax Category"
  ) +
  scale_fill_brewer(palette = "Set2") +
  scale_y_continuous(labels = label_number(suffix = "%", accuracy = 1),
                     expand = expansion(mult = c(0, 0.05))) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14, hjust = 0),
    plot.subtitle = element_text(size = 10, color = "gray30", hjust = 0),
    axis.text.x = element_text(angle = 45, hjust = 1,
                               face = "bold", size = 10),
    legend.position = "bottom",
    legend.title = element_blank(),
    panel.grid.major.x = element_blank(),
    panel.grid.minor = element_blank()
  )

combined_plot <- plot_oecd / plot_imf +
  plot_layout(guides = "collect") & theme(legend.position = "bottom")
print(combined_plot)
```



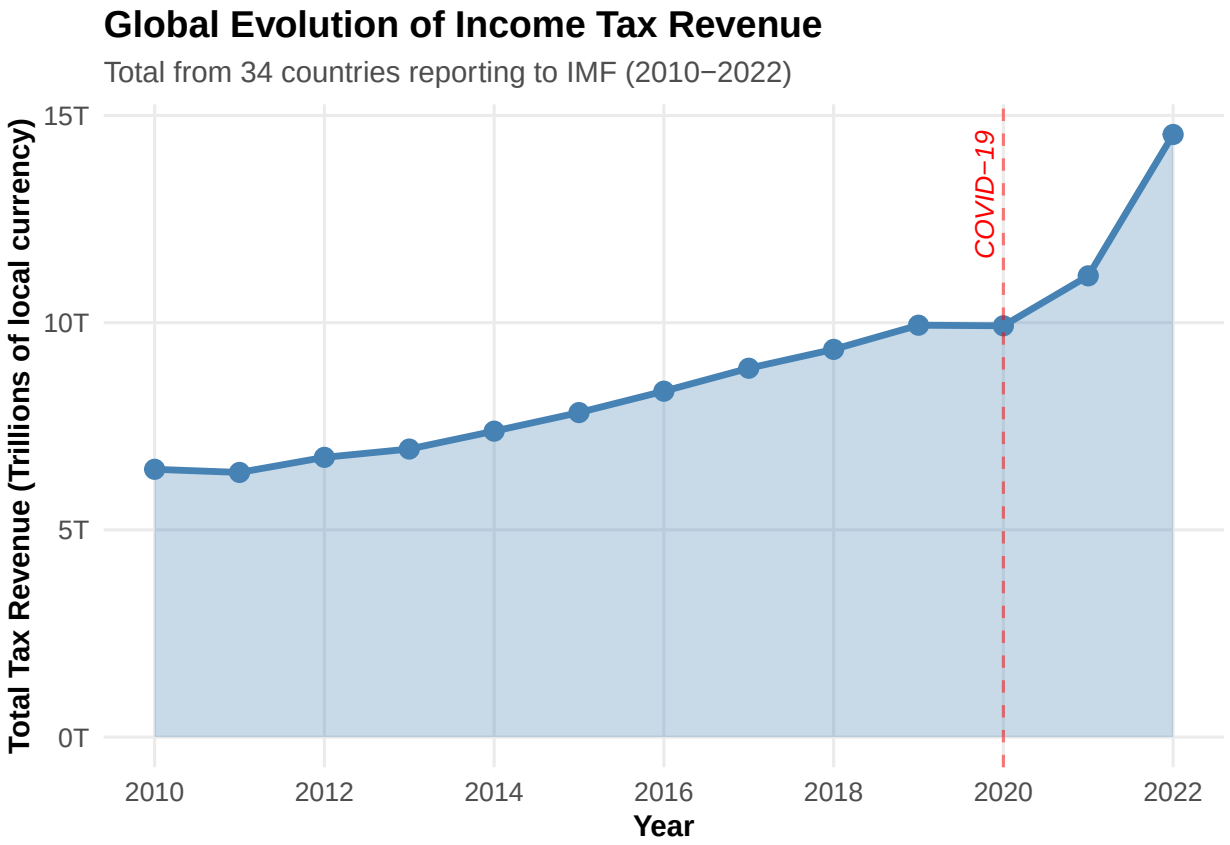
For this first comparison, i create two bar chart which provide informations about the fiscal

structure but they are not stack. And then I placed the two graphs: one above the other for a visual comparison. We deduce they are almost the sames but not. So yes they differ a bit. Data are not the same between IMF and OECD.

TASK 3B

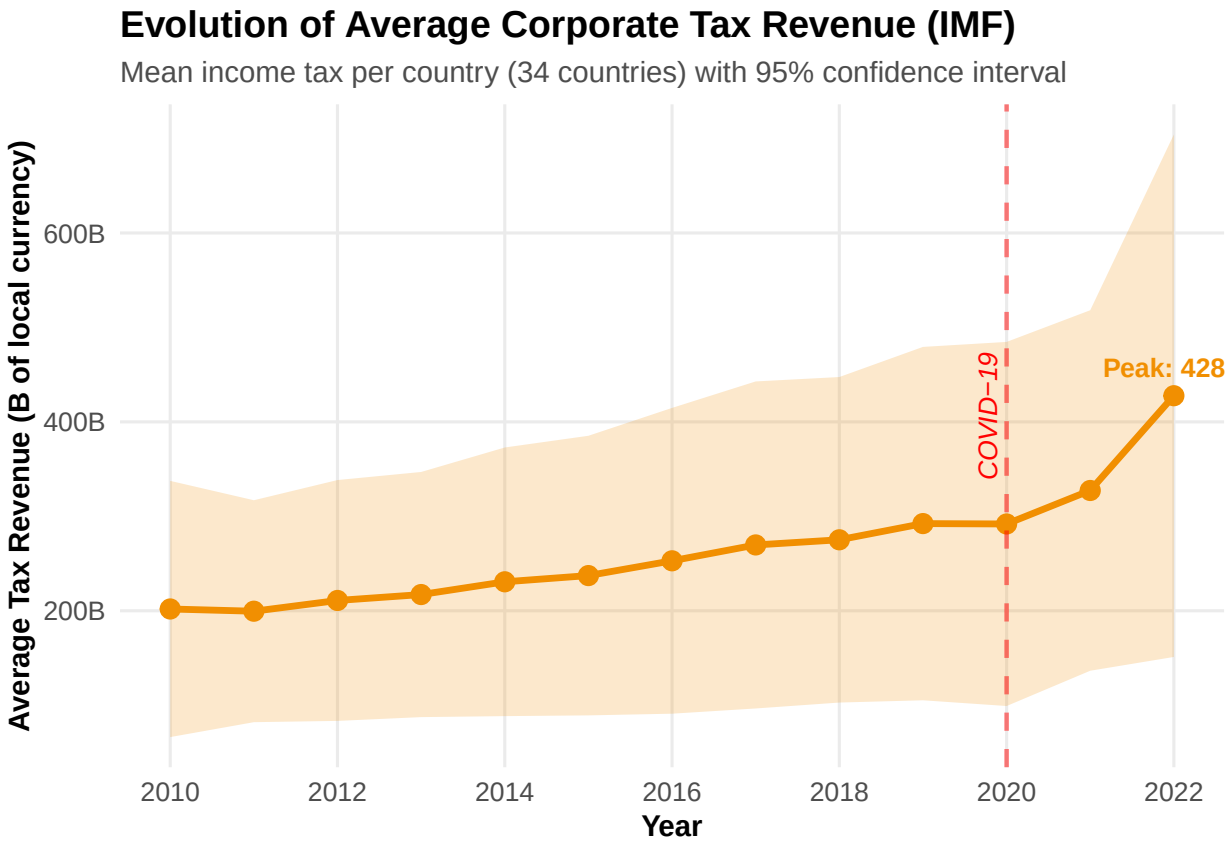
Now, we display the same type of graph like in task 1B in application with the IMF data.

```
imf_corporate <- imf_long %>%
  filter(
    INDICATOR == "Taxes on income, profits, and capital gains, Transactions",
    Year >= 2010,
    Year <= 2022
  )

imf_world_avg <- imf_corporate %>%
  group_by(Year) %>%
  summarise(
    Countries = n(),
    Total_Tax = sum(Value, na.rm = TRUE)
  )

plot_3b <- ggplot(imf_world_avg, aes(x = Year, y = Total_Tax / 1e6)) +
  geom_area(fill = "steelblue", alpha = 0.3) +
  geom_line(color = "steelblue", size = 1.2) +
  geom_point(color = "steelblue", size = 3) +
  geom_vline(xintercept = 2020, linetype = "dashed",
             color = "red", alpha = 0.5) +
  annotate("text", x = 2020, y = max(imf_world_avg$Total_Tax / 1e6) * 0.9,
           label = "COVID-19", angle = 90, vjust = -0.5,
           color = "red", size = 3.5, fontface = "italic") +
  labs(
    title = "Global Evolution of Income Tax Revenue",
    subtitle = paste0("Total from ", max(imf_world_avg$Countries),
                      " countries reporting to IMF (2010-2022)"),
    x = "Year",
    y = "Total Tax Revenue (Trillions of local currency)",
  ) +
  scale_x_continuous(breaks = seq(2010, 2022, 2)) +
  scale_y_continuous(labels = label_number(scale = 1, suffix = "T")) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 11, color = "gray30"),
    axis.title = element_text(face = "bold", size = 11),
    panel.grid.minor = element_blank(),
    plot.caption = element_text(hjust = 0, size = 8, color = "gray50", lineheight = 1.2)
  )

print(plot_3b)
```




```
imf_corporate_stats <- imf_long %>%
  filter(
    INDICATOR == "Taxes on income, profits, and capital gains, Transactions",
    Year >= 2010,
    Year <= 2022
  ) %>%
  group_by(Year) %>%
  summarise(
    Countries = n(),
    Mean_Tax = mean(Value, na.rm = TRUE),
    SD = sd(Value, na.rm = TRUE),
    Total_Tax = sum(Value, na.rm = TRUE),
    N = n()
  ) %>%
  mutate(
    SE = SD / sqrt(N),
    CI_lower = Mean_Tax - 1.96 * SE,
    CI_upper = Mean_Tax + 1.96 * SE,
    Mean_Tax_Billions = Mean_Tax / 1e3,
    CI_lower_Billions = CI_lower / 1e3,
    CI_upper_Billions = CI_upper / 1e3
  )

peak_year_imf <- imf_corporate_stats %>%
  filter(Mean_Tax_Billions == max(Mean_Tax_Billions, na.rm = TRUE)) %>%
  pull(Year)
```

```r
peak_value_imf <- imf_corporate_stats %>%
  filter(Year == peak_year_imf) %>%
  pull(Mean_Tax_Billions)

plot_imf_corporate_ci <- ggplot(imf_corporate_stats,
                                aes(x = Year, y = Mean_Tax_Billions)) +
  geom_ribbon(aes(ymin = CI_lower_Billions, ymax = CI_upper_Billions),
              fill = "#F18F01", alpha = 0.2) +
  geom_line(color = "#F18F01", linewidth = 1.2) +
  geom_point(color = "#F18F01", size = 3) +
  geom_vline(xintercept = 2020, linetype = "dashed",
             color = "red", alpha = 0.5, linewidth = 0.8) +
  annotate("text", x = 2020, y = max(imf_corporate_stats$Mean_Tax_Billions) * 0.95,
           label = "COVID-19", angle = 90, vjust = -0.5,
           color = "red", size = 3.5, fontface = "italic") +

  annotate("text", x = peak_year_imf, y = peak_value_imf + 30,
           label = paste0("Peak: ", round(peak_value_imf, 0), "B"),
           color = "#F18F01", size = 3.5, fontface = "bold") +
  labs(
    title = "Evolution of Average Corporate Tax Revenue (IMF)",
    subtitle = paste0("Mean income tax per country (",
                      max(imf_corporate_stats$Countries),
                      " countries) with 95% confidence interval"),
    x = "Year",
    y = "Average Tax Revenue (B of local currency)",
  ) +
  scale_x_continuous(breaks = seq(2010, 2022, 2)) +
  scale_y_continuous(labels = label_number(suffix = "B", scale = 1)) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 11, color = "gray30"),
    axis.title = element_text(face = "bold", size = 11),
    panel.grid.minor = element_blank(),
    plot.caption = element_text(hjust = 0, size = 8, color = "gray50")
  )

print(plot_imf_corporate_ci)
```



So, I create previouly in task 1B a graph and now, i create two graphs that represent the sum on average of tax corporations by country accross years. It is representative and allow to compare the direction of the graphs. We see that the COVID-19 accelerate once again the tax revenue of corporations with a peak in 2022 at 428B. Yes it is mentionned that I have 34 countries and OCDE is composed of 38, but I don't find a good table from IMF who regroup these 38 countries. And then the results are not perfectly the same but follow the same direction.

TASK 3C

CLEANING EUROSTAT DATASET

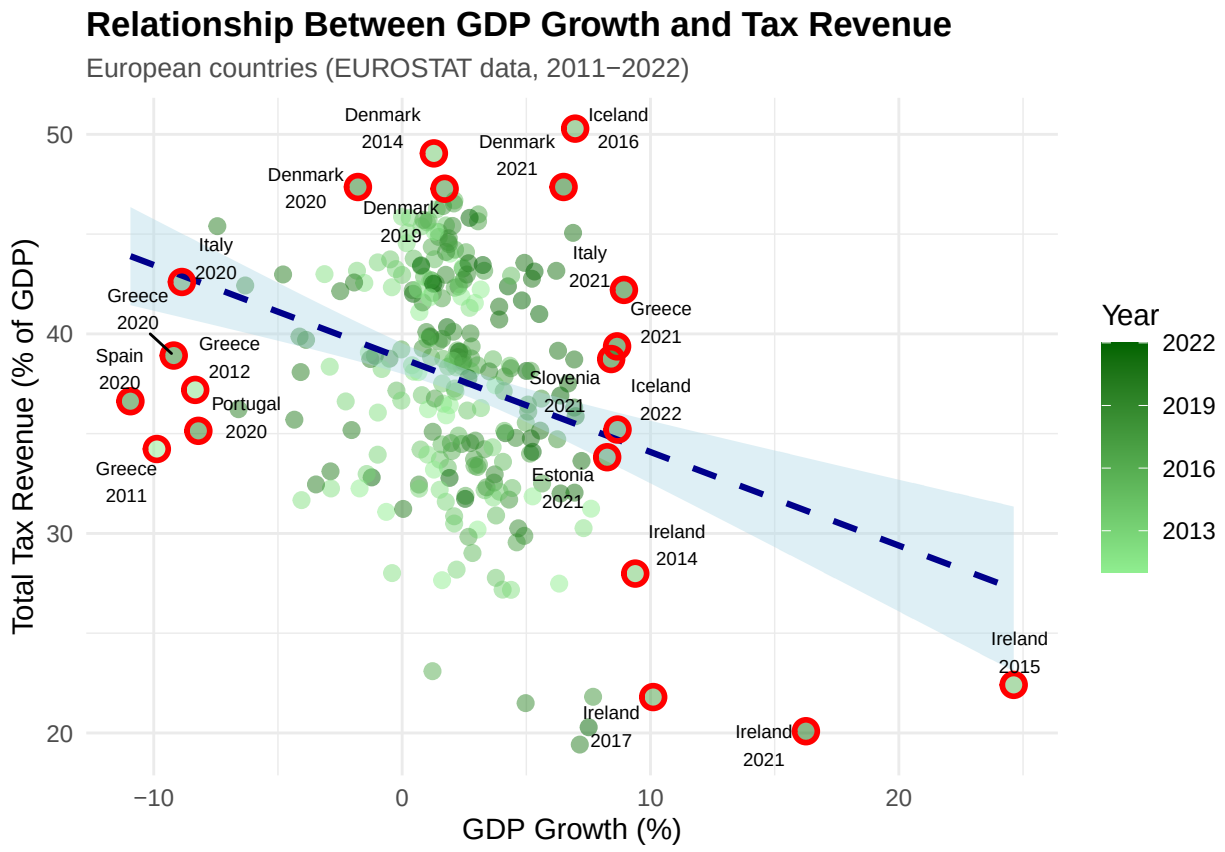
```
eurostat_raw <- read_excel("tax-main-aggregates.xlsx",
                           sheet = "Table 1",
                           skip = 2)

names(eurostat_raw)[1] <- "Country"
eurostat_raw <- eurostat_raw %>%
  filter(!is.na(Country),
         !str_detect(Country, "Source:|Data extracted|^NA$|EU-27|EA-20"))

eurostat_raw <- eurostat_raw %>%
  select(1:14)
```

```r
names(eurostat_raw)[2:14] <- as.character(2011:2023)

eurostat_long <- eurostat_raw %>%
  pivot_longer(cols = `2011`:`2023`,
               names_to = "Year",
               values_to = "Total_tax") %>%
  mutate(
    Year = as.numeric(Year),
    Total_tax = as.numeric(Total_tax)
  ) %>%
  filter(!is.na(Total_tax))

eurostat_long <- eurostat_long %>%
  mutate(Country = case_when(
    Country == "Czechia" ~ "Czech Republic",
    Country == "Türkiye" ~ "Turkey",
    TRUE ~ Country
  ))

eurostat_long
```

```
## # A tibble: 377 x 3
##    Country  Year Total_tax
##    <chr>   <dbl>     <dbl>
##  1 Belgium  2011      44.0
##  2 Belgium  2012      45.1
##  3 Belgium  2013      45.8
##  4 Belgium  2014      45.4
##  5 Belgium  2015      45.1
##  6 Belgium  2016      44.4
##  7 Belgium  2017      44.9
##  8 Belgium  2018      44.8
##  9 Belgium  2019      43.3
## 10 Belgium  2020      43.0
## # i 367 more rows
```

```r
eurostat_gdp <- eurostat_long %>%
  left_join(gdp_long, by = c("Country", "Year"))# Données européennes uniquement
europe_data <- eurostat_gdp %>%
  filter(!is.na(GDP_growth), !is.na(Total_tax)) %>%
  filter(Year >= 2011, Year <= 2022)

outliers_eu <- europe_data %>%
  filter(abs(GDP_growth) > 8 | Total_tax > 47)

p3c <- ggplot(europe_data,
              aes(x = GDP_growth, y = Total_tax)) +
  geom_point(aes(color = Year), alpha = 0.5, size = 2.5) +
  geom_smooth(method = "lm", se = TRUE, color = "darkblue",
              fill = "lightblue", linetype = "dashed", size = 1.1) +
  geom_point(data = outliers_eu, color = "red", size = 3, shape = 1, stroke = 1.5) +
  geom_text_repel(data = outliers_eu,
```

```
                  aes(label = paste(Country, Year, sep = "\n")),
                  size = 2.5, max.overlaps = 20) +
  labs(
    title = "Relationship Between GDP Growth and Tax Revenue",
    subtitle = "European countries (EUROSTAT data, 2011-2022)",
    x = "GDP Growth (%)",
    y = "Total Tax Revenue (% of GDP)",
    color = "Year",
  ) +
  scale_color_gradient(low = "#90EE90", high = "#006400") +
  theme_minimal(base_size = 11) +
  theme(
    plot.title = element_text(face = "bold", size = 13),
    plot.subtitle = element_text(size = 10, color = "gray30"),
    plot.caption = element_text(size = 8, color = "gray50"),
    legend.position = "right"
  )

print(p3c)
```



We notice that the graph looks like the previous graph in task 1C. Outliers like Ireland in 2015 are also here. A negative correlation is here. But we focus on a certain period ; so the correlation differ if we take an other. The same concentration of point in terms of Total tax revenue and GDP are present. We don't have perfectly the same graph beacause of data. To comment the results, we saw that Ireland with a low taxation rate have a huge GDP growth on this period.

Overall, this comparative analysis demonstrates that, although the OECD, IMF, and EUROSTAT datasets differ slightly in definitions and coverage, they reveal consistent fiscal patterns. Cross-source validation strengthens the credibility of the findings and confirms that key economic shocks — particularly COVID-19 — have left a visible imprint on global and regional tax structures.

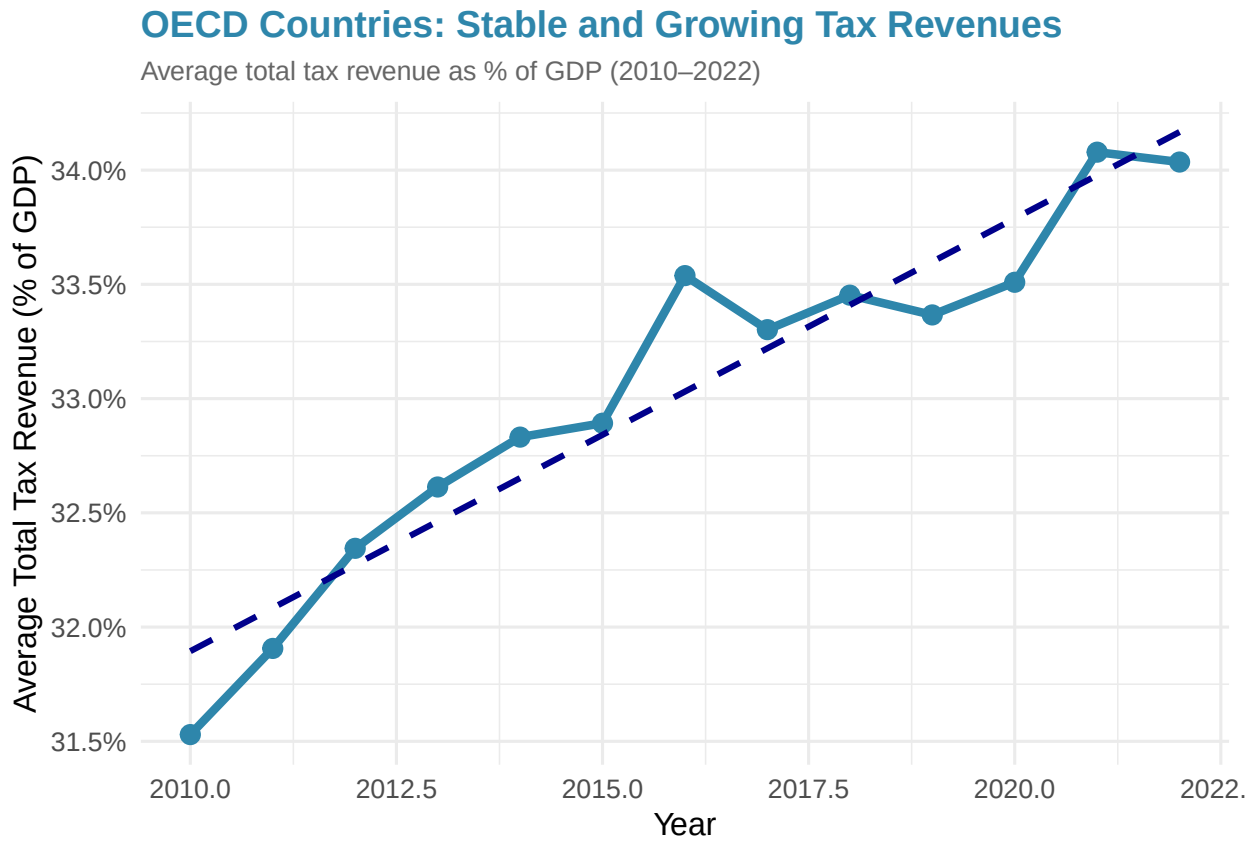
TASK 4

After these tasks with the OECD dataset, we imagine two visualisations wichh ones tell us a contradictory story.

TASK 4A

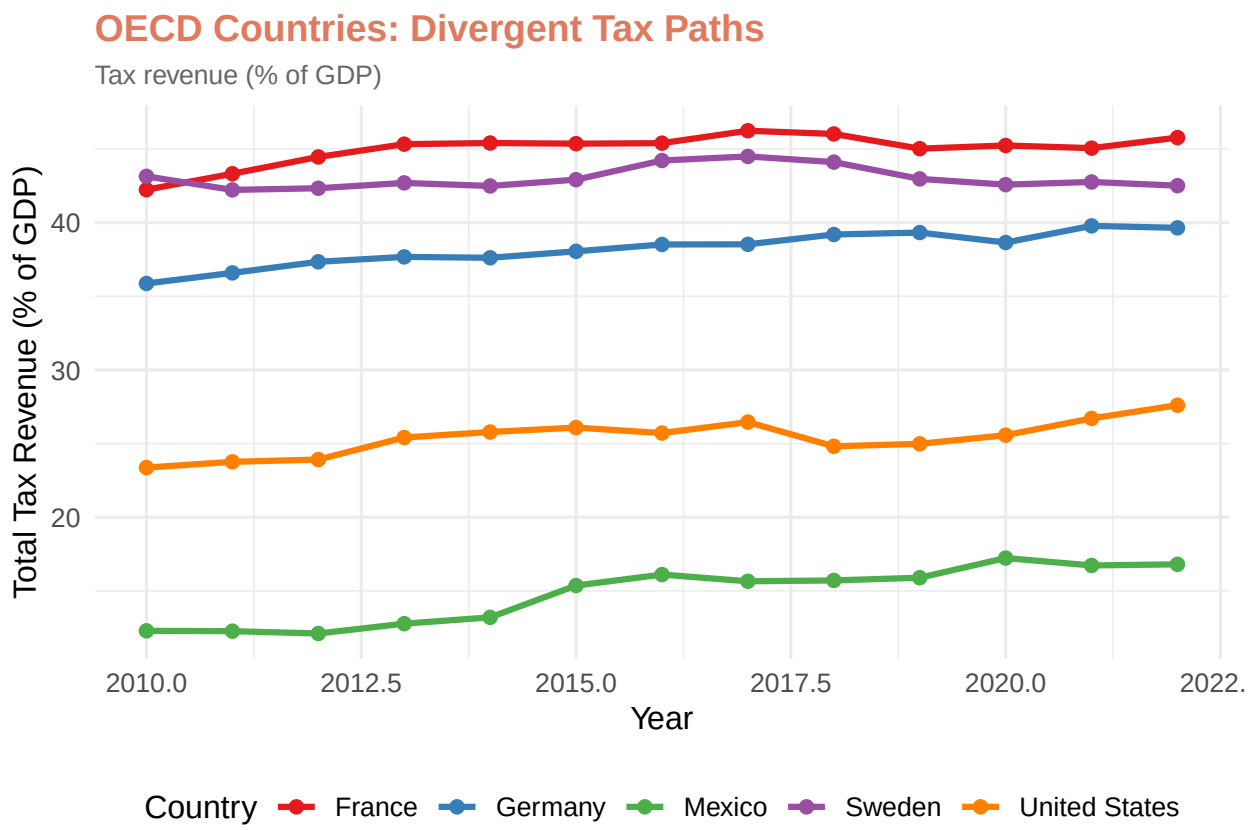
We start by construct two visualisations in optimistic and a pessimistic way. Lets me show you.

```
data_task4_opt <- oecd_long %>%
  filter(Tax_category == "Total tax revenue",
         Year >= 2010, Year <= 2022) %>%
  group_by(Year) %>%
  summarise(Average_Tax = mean(Value, na.rm = TRUE))

plot_optimistic <- ggplot(data_task4_opt, aes(x = Year, y = Average_Tax)) +
  geom_line(color = "#2E86AB", size = 1.5) +
  geom_point(color = "#2E86AB", size = 3) +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed",
              color = "darkblue") +
  labs(
    title = "OECD Countries: Stable and Growing Tax Revenues",
    subtitle = "Average total tax revenue as % of GDP (2010-2022)",
    x = "Year",
    y = "Average Total Tax Revenue (% of GDP)",
  ) +
  scale_y_continuous(labels = label_number(suffix = "%")) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14, color = "#2E86AB"),
    plot.subtitle = element_text(size = 10, color = "gray40"),
    plot.caption = element_text(size = 8, color = "gray50", hjust = 0)
  )

print(plot_optimistic)
```




```r
top_countries <- c("France", "Germany", "United States", "Mexico", "Sweden")

data_task4_pess <- oecd_long %>%
  filter(Tax_category == "Total tax revenue",
         Country %in% top_countries,
         Year >= 2010, Year <= 2022)

plot_pessimistic <- ggplot(data_task4_pess, aes(x = Year, y = Value, color = Country)) +
  geom_line(size = 1.1) +
  geom_point(size = 2) +
  labs(
    title = "OECD Countries: Divergent Tax Paths",
    subtitle = "Tax revenue (% of GDP) ",
    x = "Year",
    y = "Total Tax Revenue (% of GDP)",
    color = "Country",
  ) +
  scale_color_brewer(palette = "Set1") +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14, color = "#E07A5F"),
    plot.subtitle = element_text(size = 10, color = "gray40"),
    legend.position = "bottom",
    plot.caption = element_text(size = 8, color = "gray50", hjust = 0)
  )
```

```
print(plot_pessimistic)
```



TASK 4B

These two visualisations suggest that:

- The first visualization tells a story of resilience and predictable growth. By aggregating all OECD countries into a single average trend line, individual volatility disappears, revealing a steady upward trajectory from of GDP. The blue color palette evokes trust and stability, while the linear regression emphasizes positive momentum. This design choice focus on the mean rather than individual countries suggests that OECD tax systems are fundamentally sound and capable of weathering economic shocks like COVID-19. The smooth line creates a narrative of institutional strength and fiscal convergence across member states.
- The second visualization reveals stark fiscal divergence among OECD members. By disaggregating into five distinct country, it exposes persistent inequalities: France and Sweden maintain 40-45% GDP taxation while Mexico remains below 20%. This multi-line format highlights volatility—particularly Greece's dramatic post-crisis swings challenging the notion of a unified model between OECD members. The design deliberately fragments the data to tell a story of structural inequalities and question whether meaningful fiscal convergence exists when member countries follow fundamentally different paths over the same period.