Case Study 3: Visualization

AKSTA Statistical Computing

The .Rmd **and** *.html (or .pdf) should be uploaded in TUWEL by the deadline. Refrain from using explanatory comments in the R code chunks but write them as text instead. Points will be deducted if the submitted file is not in a decent form.*

DISCLAIMER: In case students did not contribute equally, include a disclaimer stating what each student's contribution was.

Data

Load the data set you exported in the final Task of Case Study 2. Eliminate all observations with missing values in the income status variable.

As a reminder, the data set includes world data from 2020, focusing on:

- **Education Expenditure (% of GDP)**
- **Youth Unemployment Rate (15-24 years)**
- **Net Migration Rate** (difference between the number of people entering and leaving a country per 1,000 persons)

for most world entities in 2020. The data was downloaded from https://www.cia.gov/the-world-factbook/about/archives/. Additional information on continent, subcontinent/region and income status was appended to the dataset in Case Study 2.

```
library(ggplot2)
```

```
## Warning: Paket 'ggplot2' wurde unter R Version 4.4.2 erstellt
```

```
library(dplyr)
```

```
## Warning: Paket 'dplyr' wurde unter R Version 4.4.2 erstellt
```

```
##
## Attache Paket: 'dplyr'
```

```
## Die folgenden Objekte sind maskiert von 'package:stats':
##
##     filter, lag
```

```
## Die folgenden Objekte sind maskiert von 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(forcats)
```

```
## Warning: Paket 'forcats' wurde unter R Version 4.4.2 erstellt
```

Tasks:

a. Education expenditure in different income levels

Using **ggplot2**, create a density plot of the education expenditure grouped by income status. The densities for the different groups are superimposed in the same plot rather than in different plots. Ensure that you order the levels of the income status such that in the plots the legend is ordered from High (H) to Low (L).

- The color of the density lines is black.
- The area under the density curve should be colored differently among the income status levels.
- For the colors, choose a transparency level of 0.5 for better visibility.
- Position the legend at the top center of the plot and give it no title (hint: use `element_blank()`).
- Rename the x axis as "Education expenditure (% of GDP)"

Comment briefly on the plot.

```
data_case_study2 <- read.csv("world_data_2020_tidy.csv", header = TRUE, sep = ";")
head(data_case_study2)
```

```
##        country.x iso_code.x continent.x     subcontinent.x net_migration_rate
## 1    Afghanistan        AFG        Asia      Southern Asia               -0,1
## 2        Albania        ALB      Europe    Southern Europe               -3,3
## 3        Algeria        DZA      Africa    Northern Africa               -0,9
## 4 American Samoa        ASM     Oceania          Polynesia              -26,1
## 5        Andorra        AND      Europe    Southern Europe                  0
## 6         Angola        AGO      Africa Sub-Saharan Africa               -0,2
##   youth_unempl_rate education_expenditure        income_group
## 1              17,6                   4,1          Low income
## 2              31,9                   3,6 Upper middle income
## 3              39,3                     . Upper middle income
## 4                 .                     .         High income
## 5                 .                   3,2         High income
## 6              39,4                   3,4 Lower middle income
```

```
nrow(data_case_study2)
```

```
## [1] 227
```

```
data_case_study2 <- data_case_study2[data_case_study2$income_group != ".", ]
nrow(data_case_study2)
```

```
## [1] 212
```

```
# transforming variables to numeric values
data_case_study2$education_expenditure <- as.numeric(gsub(",", ".", data_case_study2$education_expenditu
```

```
## Warning: NAs durch Umwandlung erzeugt
```

```
data_case_study2$net_migration_rate <- as.numeric(gsub(",", ".", data_case_study2$net_migration_rate))
data_case_study2$youth_unempl_rate <- as.numeric(gsub(",", ".", data_case_study2$youth_unempl_rate))
```

```
## Warning: NAs durch Umwandlung erzeugt
```

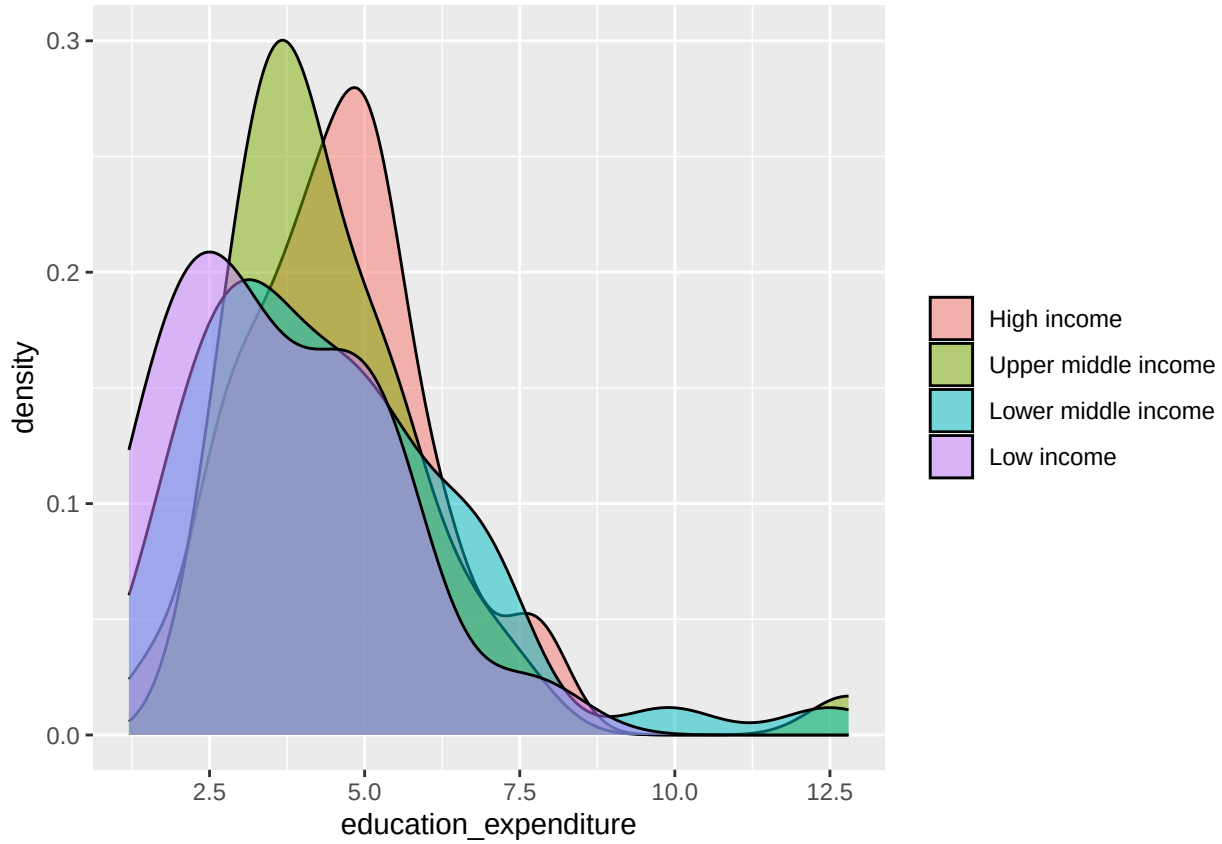
```
nrow(data_case_study2)
```

```
## [1] 212
```

```r
# ordering income groups like in task
data_case_study2$income_group <- factor(
  data_case_study2[["income_group"]],
  levels = c("High income", "Upper middle income", "Lower middle income", "Low income")
)

# density plot
ggplot(data_case_study2, aes(x = education_expenditure, fill = income_group)) +
  geom_density(color = "black", alpha = 0.5) +
  scale_fill_discrete(name = NULL)
```

```
## Warning: Removed 47 rows containing non-finite outside the scale range
## (`stat_density()`).
```




```r
  labs(x = "Education expenditure in % of GDP") +
  theme(
    legend.position = "top",
    legend.justification = "center",
    legend.title = element_blank()
  )
```

```
## NULL
```

Analyzing the plot, we can see that there are more countries in group “High income” and “Upper middle income”. Further, we can see that countries being in groups “lower middle income” and “upper middle income” spend the highest portion of their gdp for education. This makes sense, as education should be a basic need.

One can also see, that a few countries from groups “High income” and “Upper middle income” are spending-

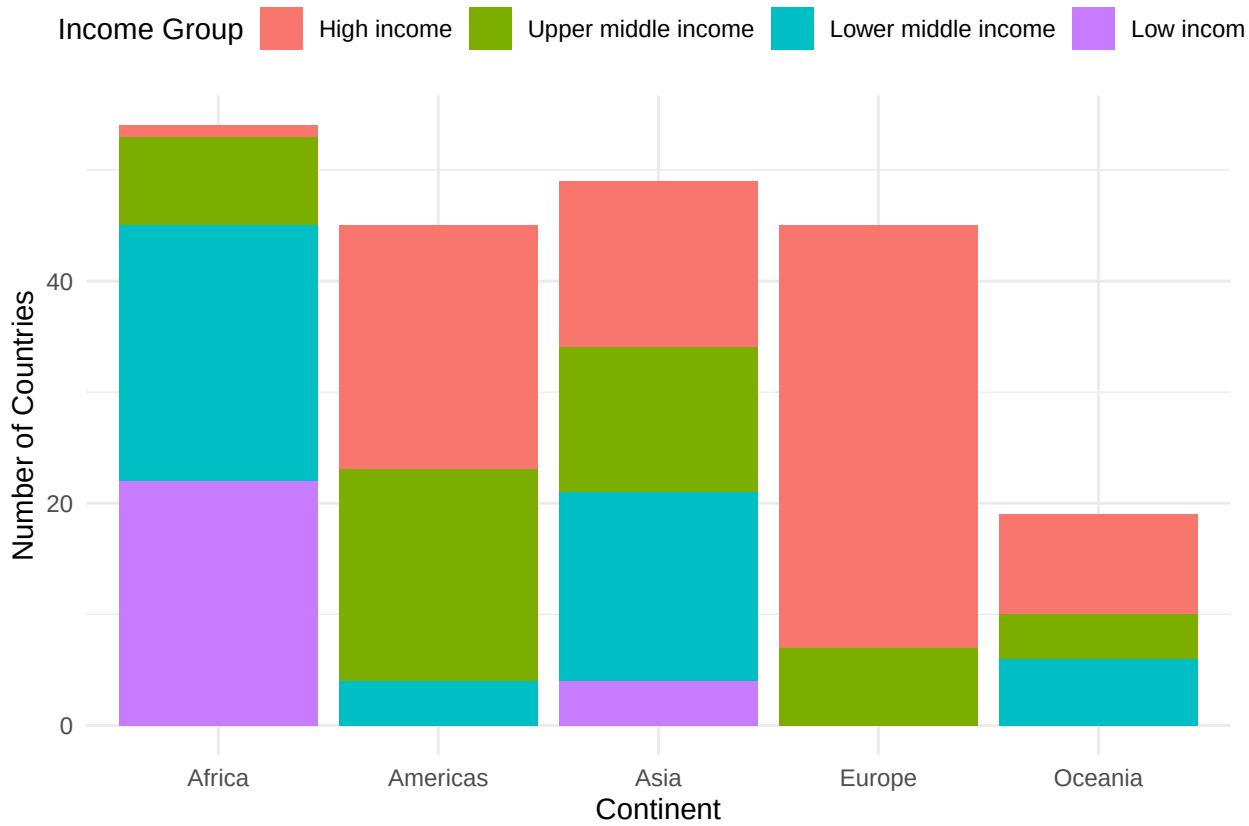
below 2,5% of their gdp on education, while many countries from groups “Lower middle income” and “Low income” spend below 2,5% of their gdp on education.

b. Income status in different continents

Investigate how the income status is distributed in the different continents.

- Using **ggplot2**, create a stacked barplot of absolute frequencies showing how the entities are split into continents and income status. Comment the plot.
- Create another stacked barplot of relative frequencies (height of the bars should be one). Comment the plot.
- Create a mosaic plot of continents and income status using base R functions.
- Briefly comment on the differences between the three plots generated to investigate the income distribution among the different continents.

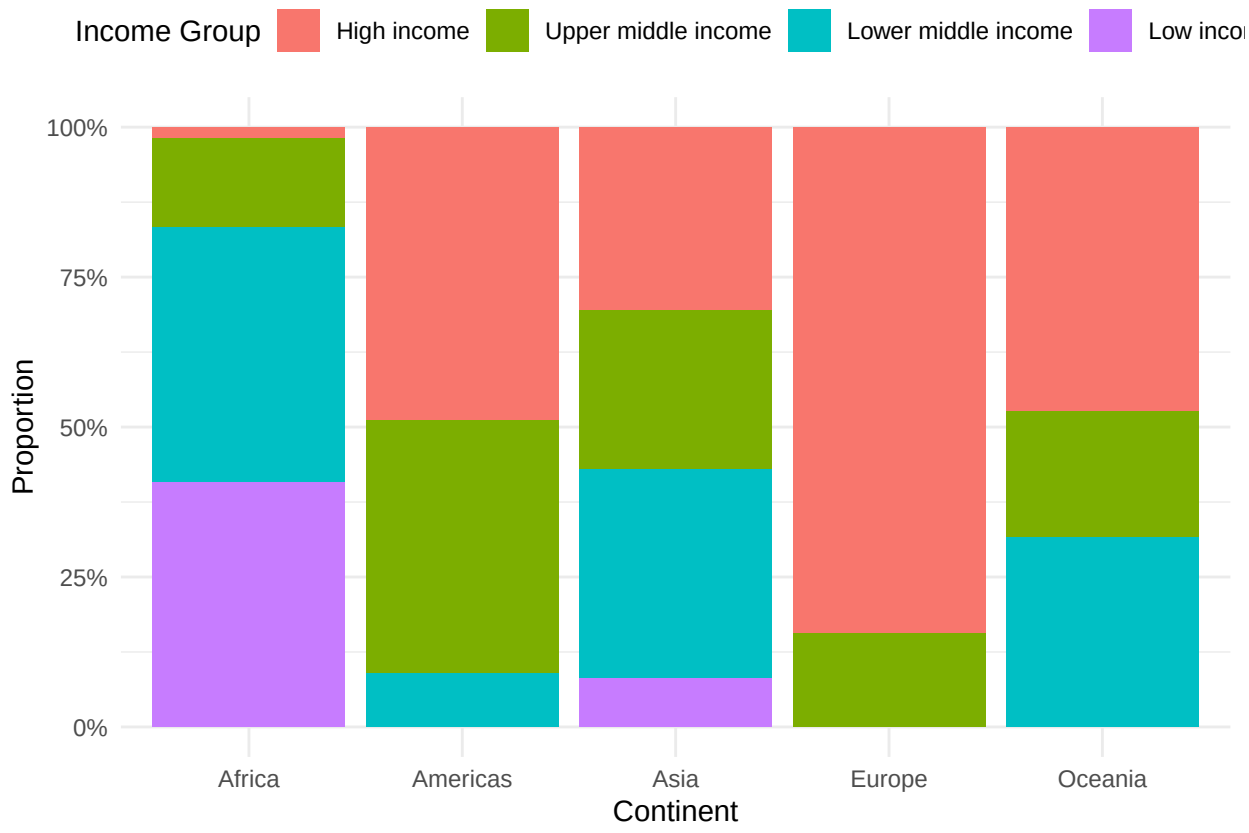
```
# stacked barplot with absolute frequencies with removed title of legend
ggplot(data_case_study2, aes(x = continent.x, fill = income_group)) +
  geom_bar(position = "stack") +
  labs(x = "Continent", y = "Number of Countries", fill = "Income Group") +
  theme_minimal() +
  theme(legend.position = "top")
```



On the stacked barplot with absolute frequencies we can see how many countries exist in each income group on each continent.

```
# stacked barplot with relative frequencies with removed title of legend
ggplot(data_case_study2, aes(x = continent.x, fill = income_group)) +
```

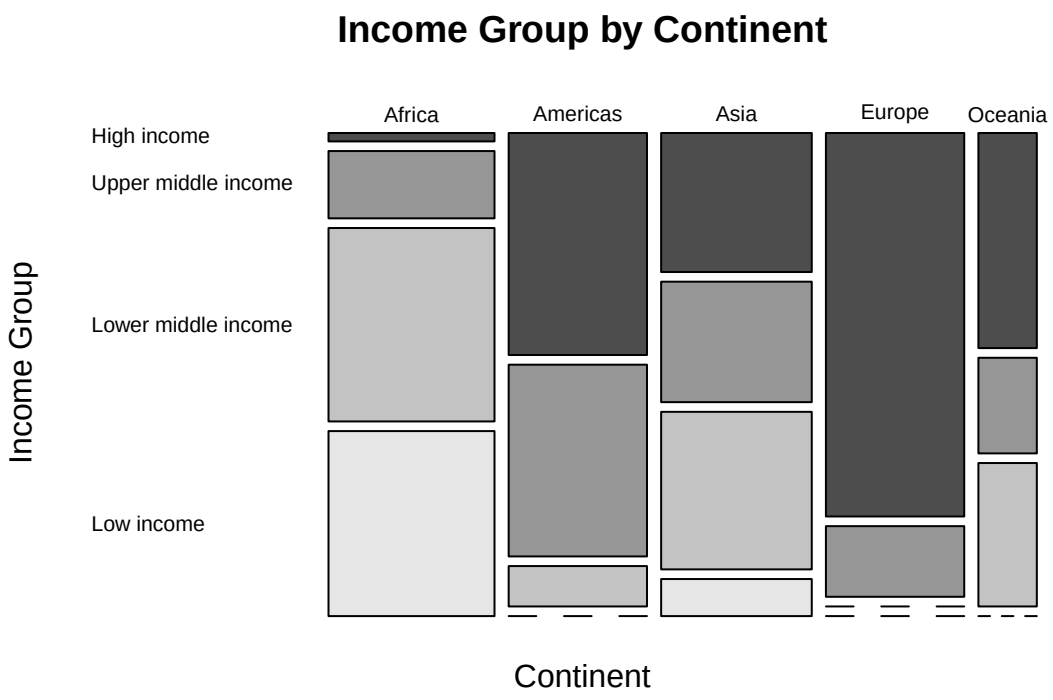
```
  geom_bar(position = "fill") +
  labs(x = "Continent", y = "Proportion", fill = "Income Group") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal() +
  theme(legend.position = "top")
```



On the stacked barplot with relative frequencies we can see what the proportion of countries in each income group on each continent is.

```
# contingency table for mosaic plot
tbl <- table(data_case_study2$continent.x, data_case_study2$income_group)

# mosaic plot
mosaicplot(tbl, main = "Income Group by Continent", xlab = "Continent", ylab = "Income Group", color = T
```



On the mosaic plot we can see how countries on each continent are split between all present income groups. The first plot shows the absolute amount of countries in each income group. THis plot is especially effective when a user needs to analyze a question with absolute numbers. The second plot plots the same, but in a relative manner. It can be used to compare the distribution of countries on each continent within each income group very fast and intuitive without computing number of countries in the brain. Each continent can be compared on its own with all other continents very effective.

The last mosaci plot shows also the relative distribution of countries of a continent within income groups. The main message and design is similar to the stacked barplot with relative frequencies. But there are still two differences that let us reject this type of plot in the following task. Firstly, the income groups are not defined as color with a legend anymore, but on the x-axis. The percentage of relative frequencies of a country is moved to the spectator to calculated the distribution of each continent in its brain. Secondly, the axis labeling of the y-axis has moved from bottom to the top, which is an unusual place.

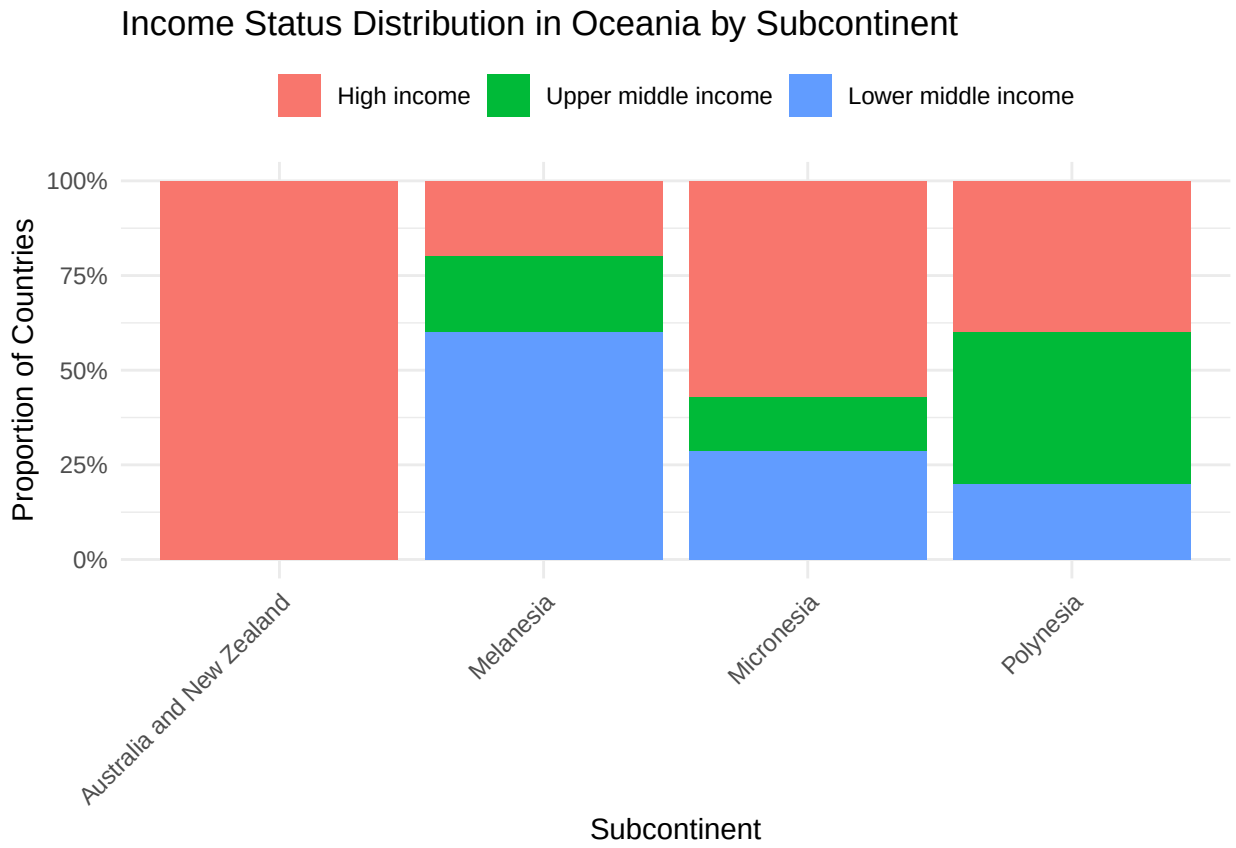
c. Income status in different subcontinents

For Oceania, investigate further how the income status distribution is in the different subcontinents. Use one of the plots in b. for this purpose. Comment on the results.

```
# creating oceania  dataset
oceania_data <- data_case_study2 %>%
  filter(continent.x == "Oceania", !is.na(subcontinent.x), !is.na(income_group))

# ordering incpome group data
oceania_data$income_group <- factor(oceania_data$income_group,
                                    levels = c("High income", "Upper middle income",
                                               "Lower middle income", "Low income"))
```

```r
# plotting relative frequencies of income group distribution
ggplot(oceania_data, aes(x = subcontinent.x, fill = income_group)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(x = "Subcontinent",
       y = "Proportion of Countries",
       fill = "Income Group",
       title = "Income Status Distribution in Oceania by Subcontinent") +
  theme_minimal() +
  theme(legend.position = "top",
        legend.title = element_blank(),
        axis.text.x = element_text(angle = 45, hjust = 1))
```



We have chosen the stacked barplot with relative frequencies of countries, because it is a good way to compare the distribution income groups of each subcontinent. Analyzing Oceania a main continent, we can see that its Auuuuuuuuuuuuuuuustralia and New Zealand subcontinent part has only countries that belong into the group "High income".

The distribution of countries in income groups of the other three subcontinents can easily be compared by each other using that type of plot. From that, we can see fast that Micronesia has the highest proportion of high income countries and Melanesia has the lowest proportion of high income countries of those three chosen subcontinents except of "Australia and New Zealand". ## d. Net migration in different continents

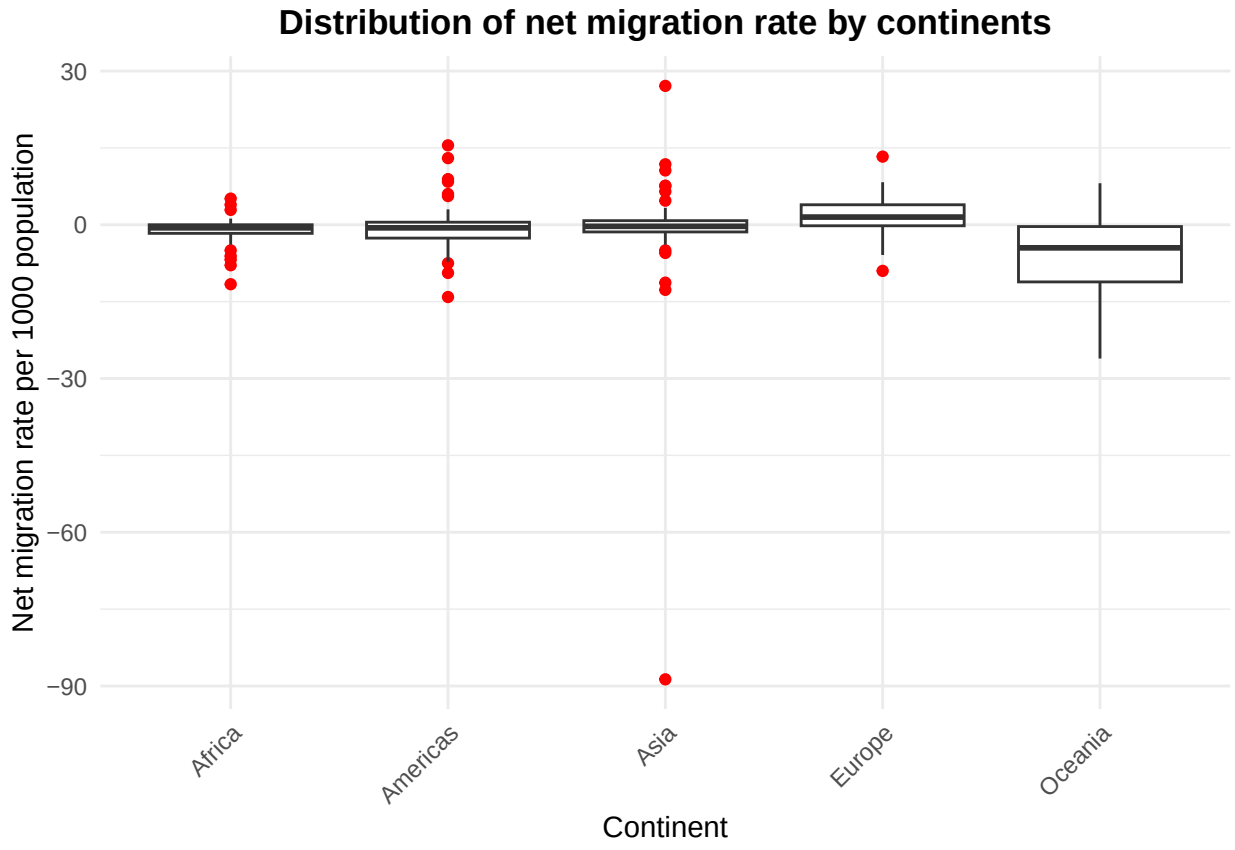
- Using **ggplot2**, create parallel boxplots showing the distribution of the net migration rate in the different continents.
- Prettify the plot (change y-, x-axis labels, etc).
- Identify which country in Asia constitutes the largest negative outlier and which country in Asia

constitutes the largest positive outlier.

- Comment on the plot.

```
data_box <- data_case_study2 %>%
  filter(!is.na(continent.x), !is.na(net_migration_rate))

# boxplott by continent
ggplot(data_box, aes(x = continent.x, y = net_migration_rate)) +
  geom_boxplot(outlier.color = "red") +
  labs(
    title = "Distribution of net migration rate by continents",
    x = "Continent",
    y = "Net migration rate per 1000 population"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(face = "bold", hjust = 0.5)
  )
```




```
asia_data <- data_box %>% filter(continent.x == "Asia")

# calc iqr ranges
q1 <- quantile(asia_data$net_migration_rate, 0.25, na.rm = TRUE)
q3 <- quantile(asia_data$net_migration_rate, 0.75, na.rm = TRUE)
iqr <- q3 - q1
```

```r
lower_bound <- q1 - 1.5 * iqr
upper_bound <- q3 + 1.5 * iqr

# investigate outliers
asia_outliers <- asia_data %>%
  filter(net_migration_rate < lower_bound | net_migration_rate > upper_bound)

# biggest positive and negative outlier
asia_outliers %>%
  arrange(net_migration_rate) %>%
  select(country.x, net_migration_rate) %>%
  slice(c(1, n()))
```

```
##   country.x net_migration_rate
## 1   Lebanon              -88.7
## 2     Syria               27.1
```

Comparing the boxplots with each other, we can see that oceania has the biggest proportion of people migrating from its continent. Asia has most widely distributed outliers to both positive and negative borders of net migration rate.

We calculated both outliers, the largest positive one and the largest negative one. The negative outlier was Lebanon, which says many people could be migrating from Lebanon. The positive outlier is Syria, which means that many people could be migrating to Syria.

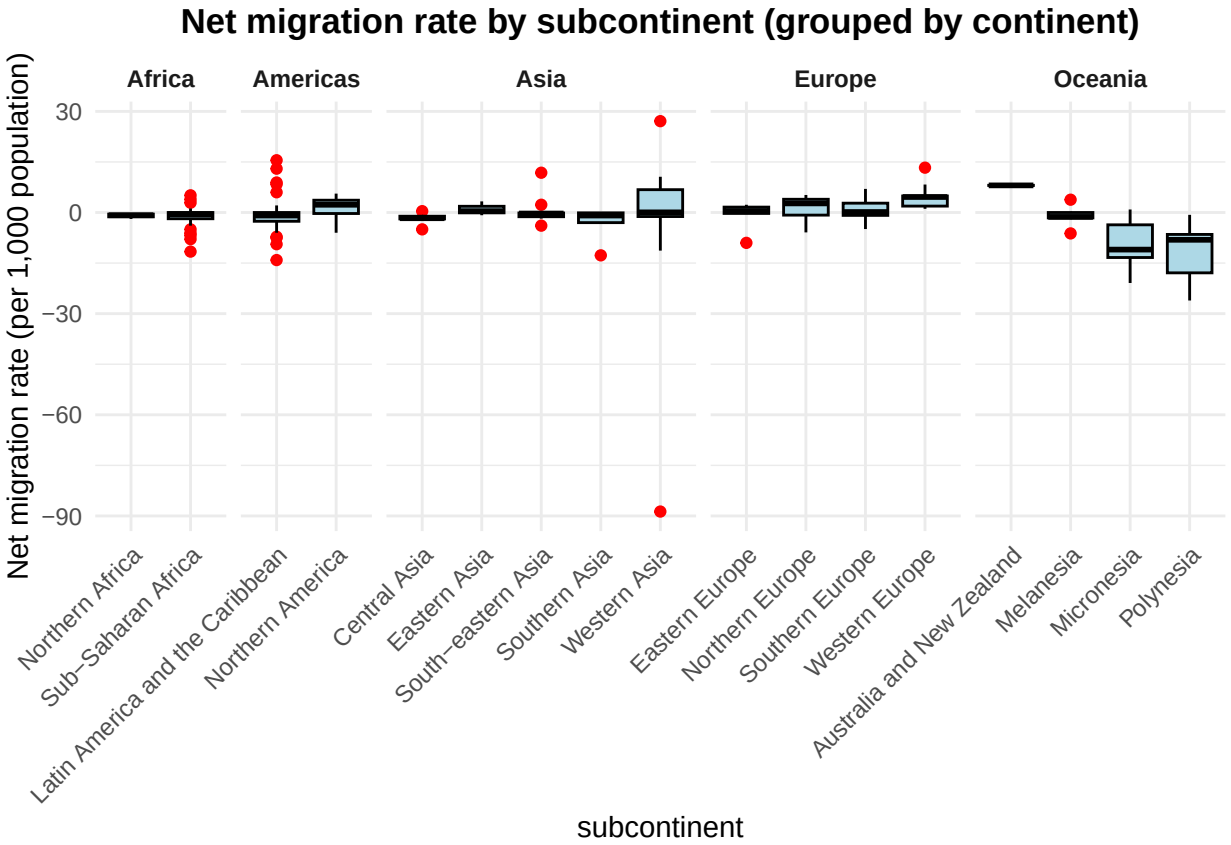
e. Net migration in different subcontinents

The graph in d. clearly does not convey the whole picture. It would be interesting also to look at the subcontinents, as it is likely that a lot of migration flows happen within the continent.

- Investigate the net migration in different subcontinents using again parallel boxplots. Group the boxplots by continent (hint: use `facet_grid` with `scales = "free_x"`).
- Remember to prettify the plot (rotate axis labels if needed).
- Describe what you see.

```r
migration_data <- data_case_study2 %>%
  filter(!is.na(continent.x), !is.na(subcontinent.x), !is.na(net_migration_rate))

# boxplott by subcontinent
ggplot(migration_data, aes(x = subcontinent.x, y = net_migration_rate)) +
  geom_boxplot(fill = "lightblue", color = "black", outlier.color = "red") +
  facet_grid(. ~ continent.x, scales = "free_x", space = "free_x") +
  labs(
    title = "Net migration rate by subcontinent (grouped by continent)",
    x = "subcontinent",
    y = "Net migration rate (per 1,000 population)"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    strip.text.x = element_text(face = "bold"),
    plot.title = element_text(face = "bold", hjust = 0.5)
  )
```

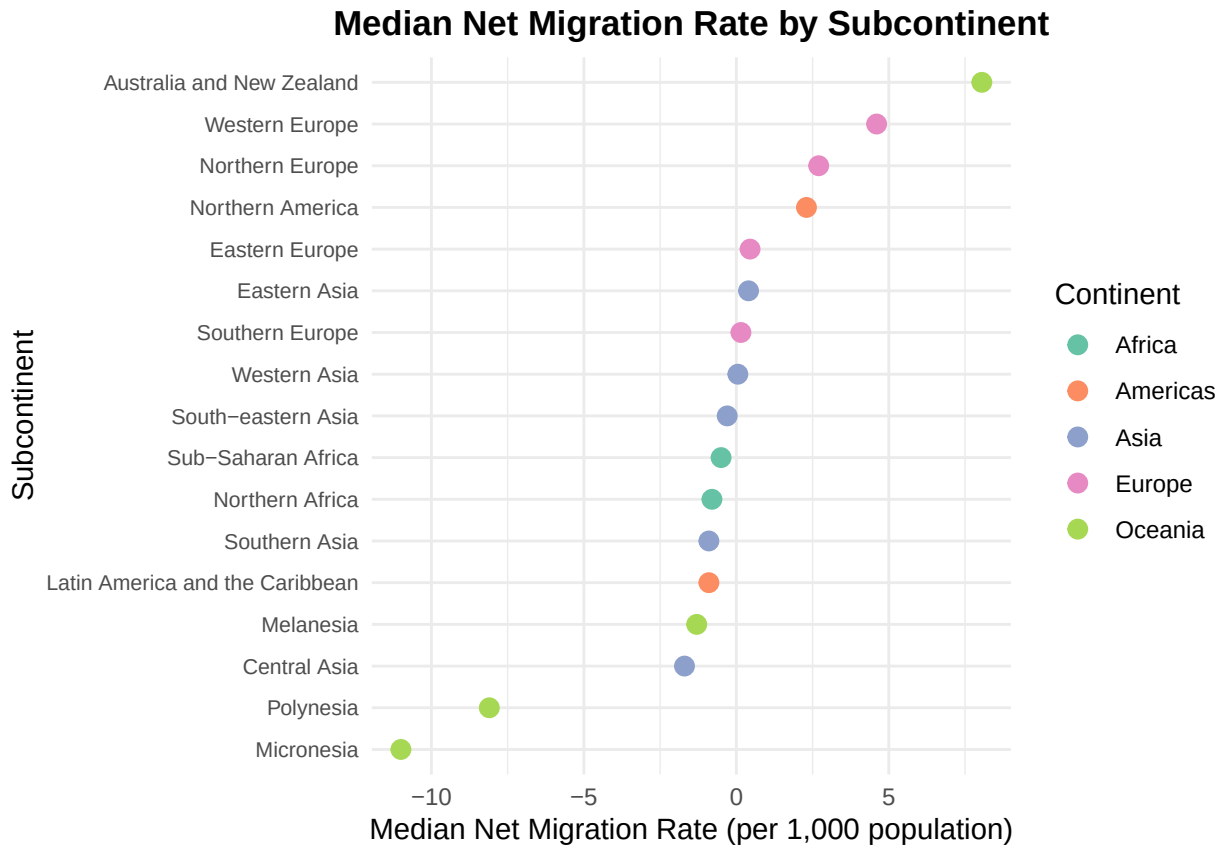


On the boxplots showing the net migration rate for each subcontinent, we can see that there is a flow between oceanian subcontinents that indicates a regional movement especially from Micronesia and Polynesia (lowest net migration rate) to Australia and New Zealand (largest net migration rate in considering continent Oceania and overall!). The people from *Micronesia* and *Polynesia* could also move to other continents, while poeple from other continents could move to Australia and New Zealand. So, to justify that hypothesis, we would need to know the connections of those movements.

Further, we can see that *Polynesia*, *Micronesia* and *Western Asia* have a huge variation due to wide interquartile ranges.

We also can see that Latin America and the Carribean and Sub-Sahara Africa have the most outliers which are not extreme, but clustered in similarly closely in both the positive and negative direction from the median.

f. Median net migration rate per subcontinent.



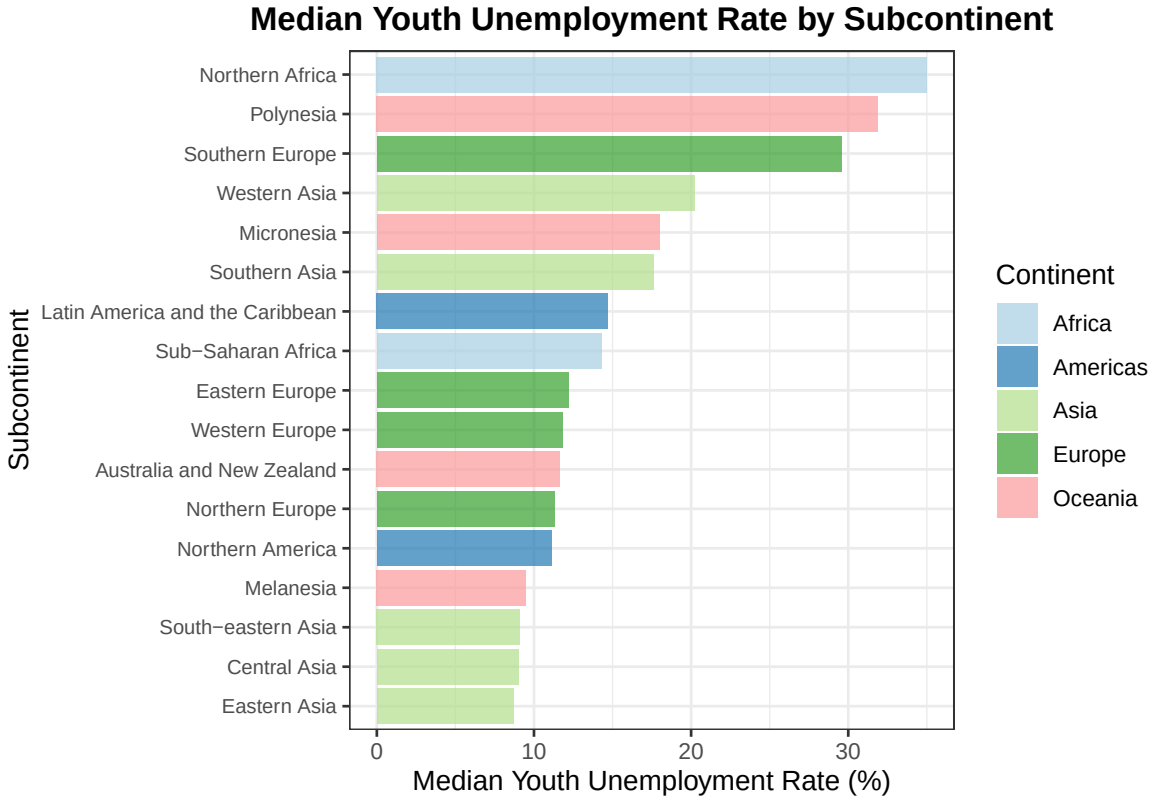
Regions with the most influx: Australia and New Zealand (Oceania) exhibits the highest median net migration rate, indicating it is the subcontinent with the most significant influx of people among those plotted. Western Europe, Northern Europe (both Europe), and Northern America (Americas) also show notable positive median net migration rates, signifying substantial population inflows. Eastern Europe (Europe) also has a positive median net migration rate.

Regions with the most outflux: Micronesia and Polynesia (both Oceania) show the lowest (most negative) median net migration rates, indicating these regions experience the most significant outflux of people. Central Asia (Asia) and Melanesia (Oceania) also have negative median net migration rates, suggesting a net outflow of population. Latin America and the Caribbean (Americas) also shows a negative median net migration rate. General Observations:

European subcontinents (Western, Northern, Eastern, Southern Europe) generally show positive net migration, with Southern Europe being close to zero but still positive. Oceanian subcontinents show a strong divergence: Australia and New Zealand has the highest influx, while Micronesia, Polynesia, and Melanesia all experience outflux. Asian subcontinents (Eastern Asia, Western Asia, South-eastern Asia, Southern Asia, Central Asia) are mostly clustered around a zero net migration rate or show a slight outflux (Central Asia, Southern Asia) or slight influx (Eastern Asia, Western Asia, South-eastern Asia). African subcontinents (Sub-Saharan Africa, Northern Africa) are also positioned near a zero net migration rate. The Americas show a contrast, with Northern America experiencing influx and Latin America and the Caribbean experiencing outflux.

g. Median youth unemployment rate per subcontinent

- Comment on the plot. E.g., what are the regions with the highest vs lowest youth unemployment rate?

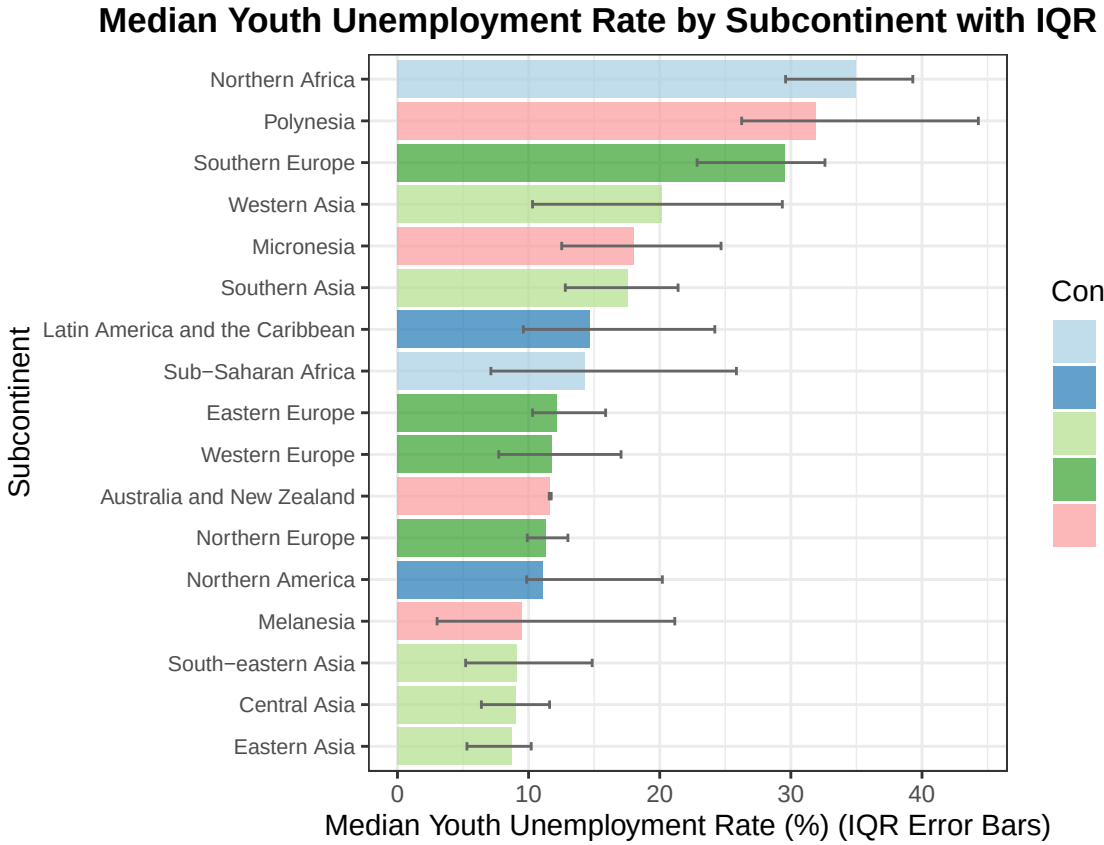


The bar chart displays the median youth unemployment rate for various subcontinents. The bars are colored by continent and are ordered along the y-axis from the subcontinent with the lowest median youth unemployment rate at the bottom to the one with the highest rate at the top.

Regions with the highest youth unemployment rate: Northern Africa (Africa) shows the highest median youth unemployment rate, exceeding 30%. Polynesia (Oceania) follows with a rate around 30%, and Southern Europe (Europe) also has a notably high rate, above 25%. Regions with the lowest youth unemployment rate: Eastern Asia and Central Asia (both Asia) exhibit the lowest median youth unemployment rates, appearing to be below 10%. South-eastern Asia (Asia) and Melanesia (Oceania) also show relatively low rates. General Observations: Asian subcontinents (Eastern Asia, Central Asia, South-eastern Asia) predominantly feature at the lower end of the youth unemployment scale. African subcontinents, particularly Northern Africa, along with parts of Oceania (Polynesia, Micronesia) and Southern Europe, are positioned towards the higher end. The Americas (Latin America and the Caribbean, Northern America) and other European subcontinents (Eastern Europe, Western Europe, Northern Europe) fall into the middle to lower-middle range of median youth unemployment rates. Australia and New Zealand (Oceania) also falls into this middle range.

h. Median youth unemployment rate per subcontinent – with error bars

Repeat the plot from Task g. but include also error bars which reflect the 25% and 75% quantiles. You can use



`geom_errorbar` in **ggplot2**.

This bar chart enhances the previous plot from task g by incorporating error bars that represent the interquartile range (IQR) – the range between the 25th and 75th percentiles – for youth unemployment within each subcontinent. These error bars provide insight into the variability and precision of the median youth unemployment rates.

Precision of Medians and Variability: Subcontinents such as Eastern Asia, Central Asia, and South-eastern Asia (all Asia) display relatively short error bars. This suggests that the youth unemployment rates among countries within these subcontinents are fairly consistent, making the median a precise representation of the typical rate. Australia and New Zealand (Oceania) and Northern Europe (Europe) also show comparably short error bars. Conversely, Polynesia (Oceania) exhibits a very long error bar, indicating a wide dispersion in youth unemployment rates among its constituent countries; the median, while high, does not capture the full spectrum of experiences, with some countries likely having much lower or even higher rates. Northern Africa (Africa), which has the highest median, also shows a substantial error bar, implying considerable variation within the subcontinent. Similarly, Latin America and the Caribbean (Americas), Sub-Saharan Africa (Africa), Micronesia (Oceania), and Southern Europe (Europe) have noticeable error bars, reflecting a broader range of youth unemployment rates. Impact on Interpretation: The error bars confirm that while some subcontinents have high median youth unemployment, the actual rates can vary significantly from country to country within those subcontinents (e.g., Polynesia, Northern Africa). For regions with shorter error bars (e.g., Eastern Asia), the median value is a more robust indicator of the general situation across the subcontinent. This added information on variability is crucial for a more nuanced understanding beyond just the central tendency.

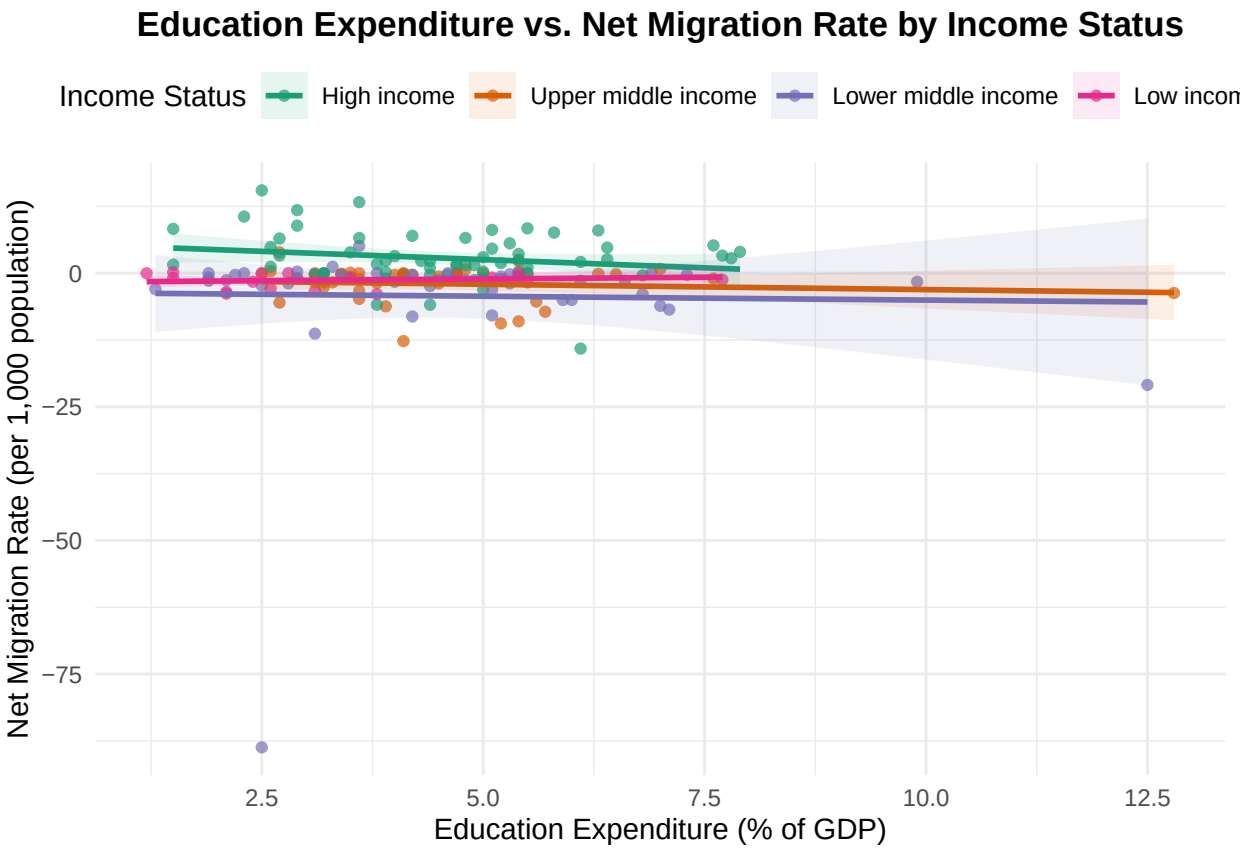
i. Relationship between education expenditure and net migration rate

Using **ggplot2**, create a plot showing the relationship between education expenditure and net migration rate.

- Color the geoms based on the income status.
- Add a regression line for each development status (using `geom_smooth()`).

Comment on the plot. Do you see any relationship between the two variables? Do you see any difference among the income levels?

```
## `geom_smooth()` using formula = 'y ~ x'
```



This scatter plot visualizes the relationship between education expenditure (as a percentage of GDP) and the net migration rate (per 1,000 population). Individual countries are represented as points, colored according to their income status, and a linear regression line with a confidence interval is shown for each income group.

Relationship between variables and differences among income levels:

High income (green): For high-income countries, there appears to be a weak negative relationship. The regression line shows a slight downward slope, suggesting that as education expenditure increases, the net migration rate tends to decrease very modestly. These countries generally exhibit higher education expenditure (mostly between 2.5% and 7.5% of GDP) and have a wide range of net migration rates, many of which are positive. Upper middle income (orange): The regression line for upper-middle-income countries is nearly flat and close to a zero net migration rate. This indicates very little to no linear relationship between education expenditure and net migration for this group. The points are somewhat clustered, showing generally lower education expenditure than high-income countries. Lower middle income (grey/purple): Similar to upper-middle-income countries, the regression line is relatively flat, hovering slightly below zero for net migration. This suggests no strong linear association between the two variables for this income category. Low income (pink): The regression line for low-income countries is also quite flat and very close to a zero

net migration rate, indicating no clear linear relationship. These countries tend to have lower education expenditure. One outlier with very high education expenditure (around 12.5% of GDP) and a negative net migration rate is visible in this group. Overall Observations:

There isn't a strong, uniform relationship between education expenditure and net migration rate that applies across all income levels. High-income countries show the widest spread in both variables and a slight negative trend. The other three income groups (Upper middle, Lower middle, and Low income) generally have lower and less variable education expenditure, and their regression lines are flatter and closer to a zero net migration rate, indicating a weaker or non-existent linear relationship within these groups. The confidence intervals for the regression lines, particularly for the non-high-income groups, are relatively wide, suggesting considerable uncertainty about the precise slope of the relationship.

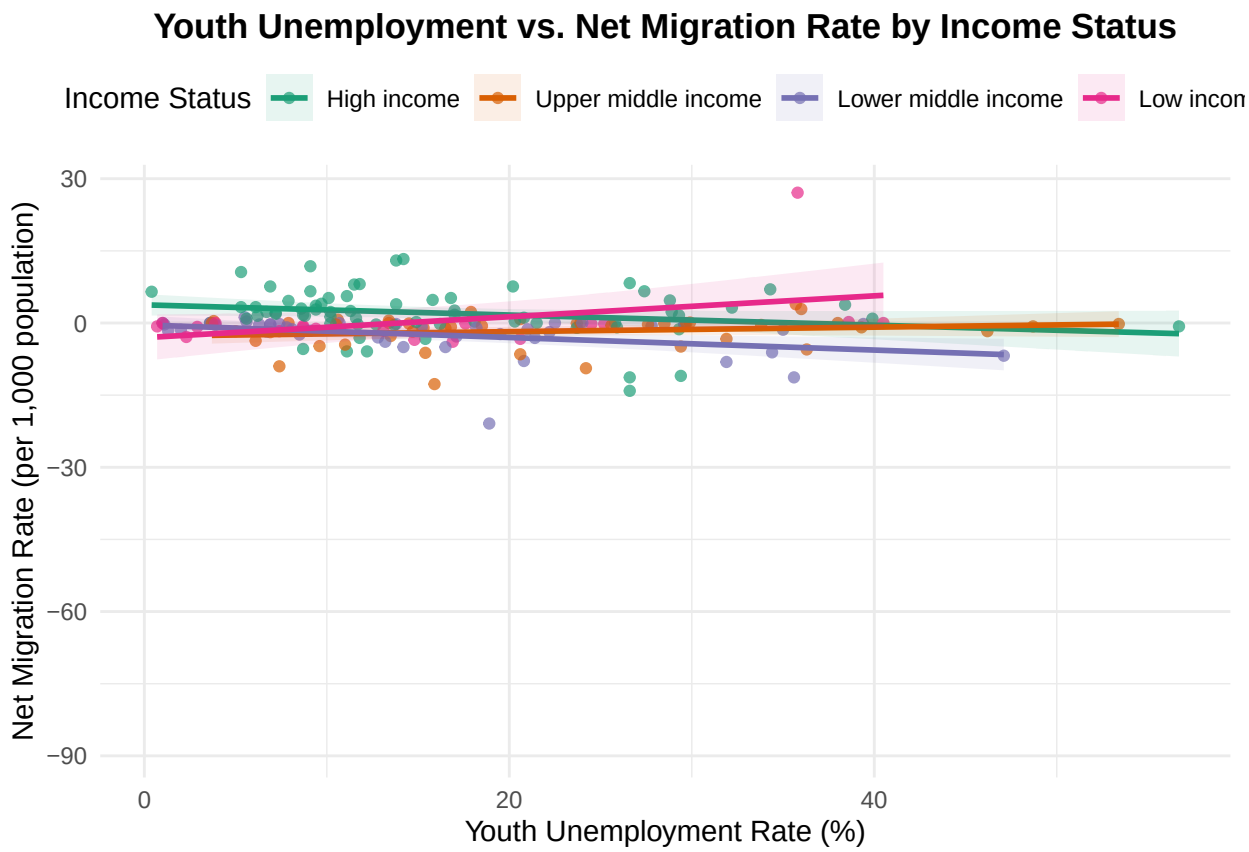
j. Relationship between youth unemployment and net migration rate

Create a plot as in Task i. but for youth unemployment and net migration rate. Comment briefly.

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 37 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 37 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



This scatter plot displays the relationship between youth unemployment rate and net migration rate, with countries color-coded by income status and individual regression lines fitted for each group.

For High income (green), Upper middle income (orange), and Lower middle income (grey/purple) countries, there is a general tendency for the regression lines to show a slight negative slope. This suggests that higher youth unemployment rates may be associated with lower net migration rates (or greater outflux) in these

groups, though the relationship appears relatively weak for high and upper-middle-income countries, and slightly more pronounced for lower-middle-income countries. For Low income countries (pink), the regression line is nearly flat or shows a very slight positive slope, indicating little to no linear relationship, or perhaps a marginal tendency for higher youth unemployment to be associated with slightly higher net migration, which contrasts with the other income groups. Overall, the points are quite dispersed, particularly for the low-income group, and the confidence intervals around the regression lines are wide, suggesting that youth unemployment is likely one of many factors influencing net migration and its impact varies across income levels.

k. Merging population data

Go online and find a data set which contains the 2020 population for the countries of the world together with ISO codes.

- Download this data and merge it to the dataset you are working on in this case study using a left join. (A possible source: World Bank))
- Inspect the data and check whether the join worked well.

l. Scatterplot of education expenditure and net migration rate in Europe

Make a scatterplot of education expenditure and net migration rate for the countries of Europe.

- Scale the size of the points according to each country's population.
- For better visibility, use a transparency of `alpha=0.7`.
- Remove the legend.
- Comment on the plot.

m. Interactive plot

On the merged data set from Task k., using function `ggplotly` from package **plotly**
re-create the scatterplot in Task l., but this time for all countries. Color the points according to their continent.

When hovering over the points the name of the country, the values for education expenditure, net migration rate, and population should be shown. (Hint: use the aesthetic `text = Country`. In `ggplotly` use the argument `tooltip = c("text", "x", "y", "size")`).

n. Parallel coordinate plot

In **parallel coordinate plots** each observation or data point is depicted as a line traversing a series of parallel axes, corresponding to a specific variable or dimension. It is often used for identifying clusters in the data.

One can create such a plot using the **GGally** R package. You should create such a plot where you look at the three main variables in the data set: education expenditure, youth unemployment rate and net migration rate. Color the lines based on the income status. Briefly comment.

o. World map visualisation

Create a world map of the education expenditure per country. You can use the vignette https://cran.r-project.org/web/packages/rworldmap/vignettes/rworldmap.pdf to find how to do this in R. Alternatively, you can use other packages (such as **ggplot2**, **sf** and **rnaturalearthdata**) to create a map.