Case Study 3: Visualization

AKSTA Statistical Computing

Carla Salazar Veneta Grigorova Juraj Simkovic

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.

```
## # A tibble: 6 x 11
##    ...1 country     ISO3  continent subcontinent              status expenditure
##   <dbl> <chr>       <chr> <chr>     <chr>                     <chr>        <dbl>
## 1     1 Afghanistan AFG   Asia      Southern Asia             L              4.1
## 2     2 Albania     ALB   Europe    Southern Europe           UM             3.6
## 3     6 Angola      AGO   Africa    Sub-Saharan Africa        LM             3.4
## 4     9 Argentina   ARG   Americas  Latin America and the Ca~ UM             5.5
## 5    10 Armenia     ARM   Asia      Western Asia              UM             2.7
## 6    12 Australia   AUS   Oceania   Australia and New Zealand H              5.1
## # i 4 more variables: youth_unempl_rate <dbl>, net_migr_rate <dbl>,
## #   low_yu <chr>, high_nmr <chr>
```

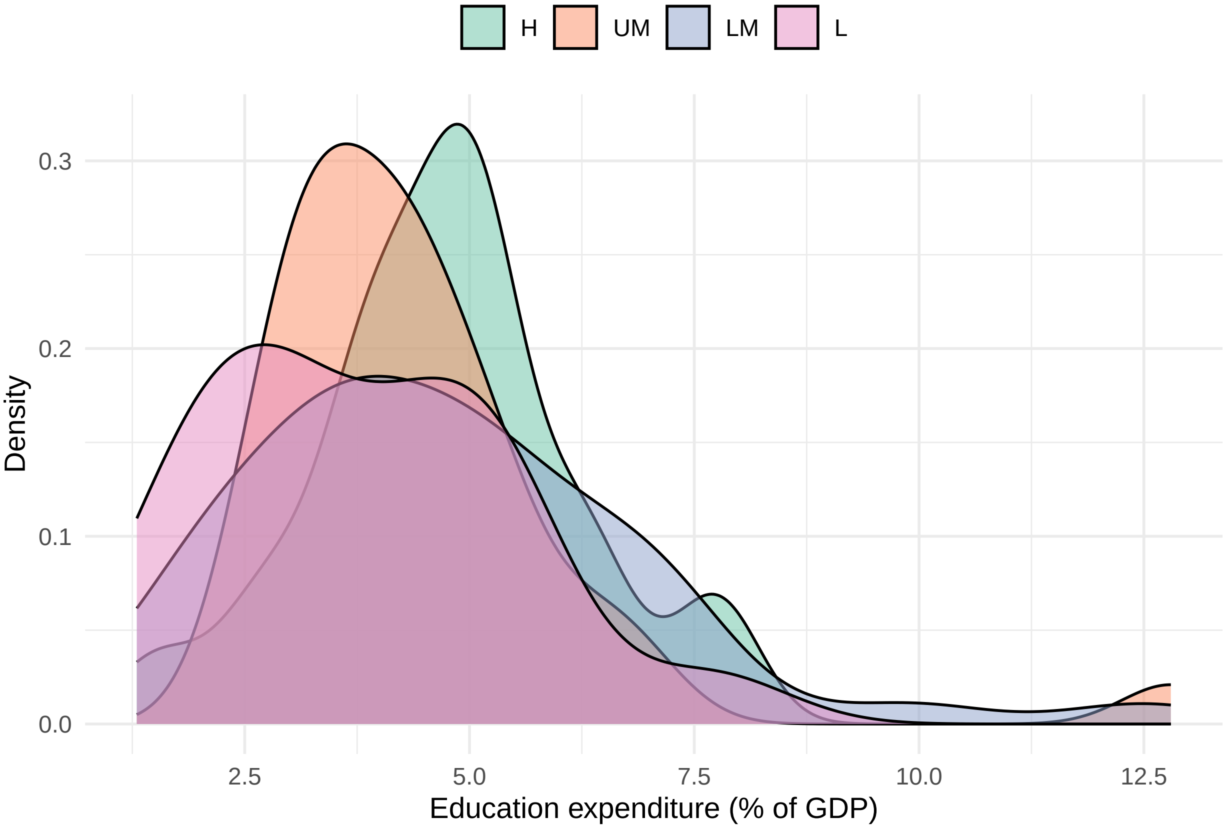
Tasks:

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.



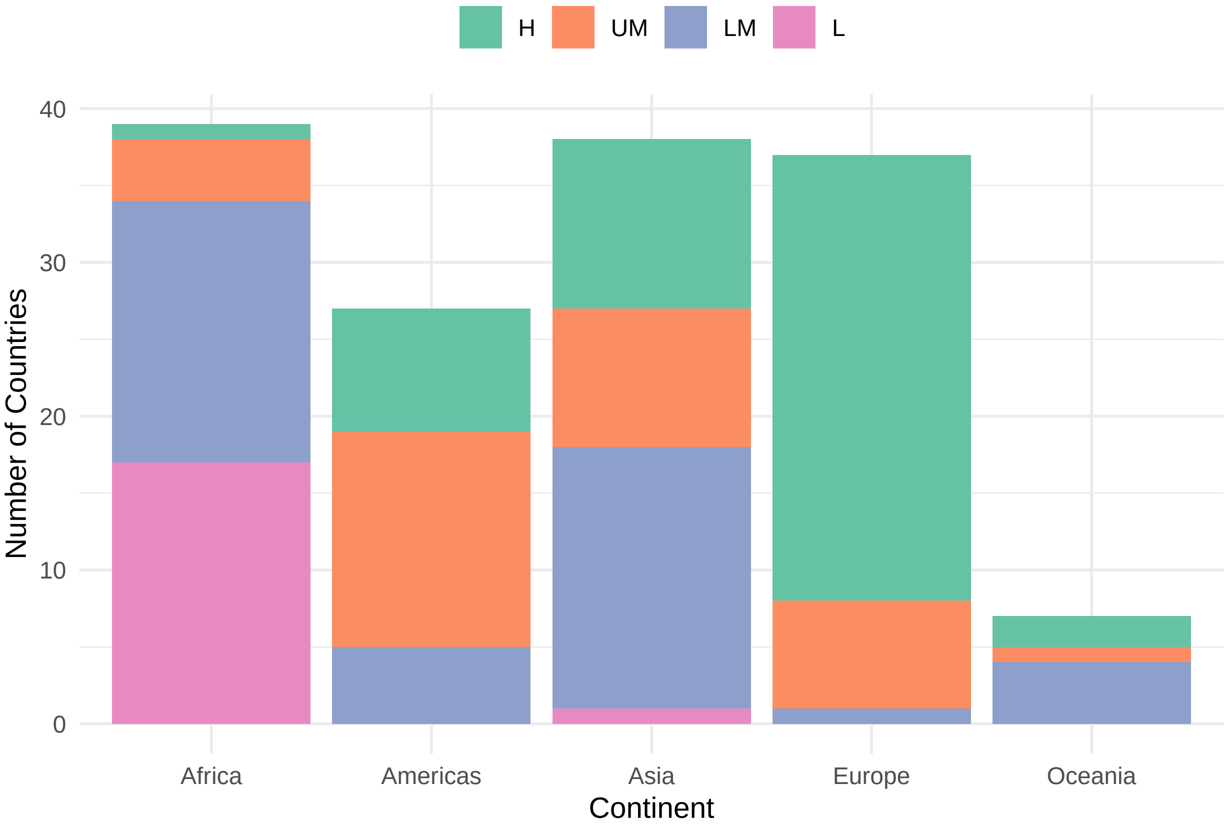
From the plot, we can see that higher-income countries (H and UM) usually spend a bit more on education compared to lower-income ones (LM and L). The curves for H and UM are narrower, meaning their values are more consistent. On the contrary, LM and L show more variation and tend to have lower values. It looks like richer countries invest more (as % of GDP) into education than poorer ones.

b. Income status in different continents

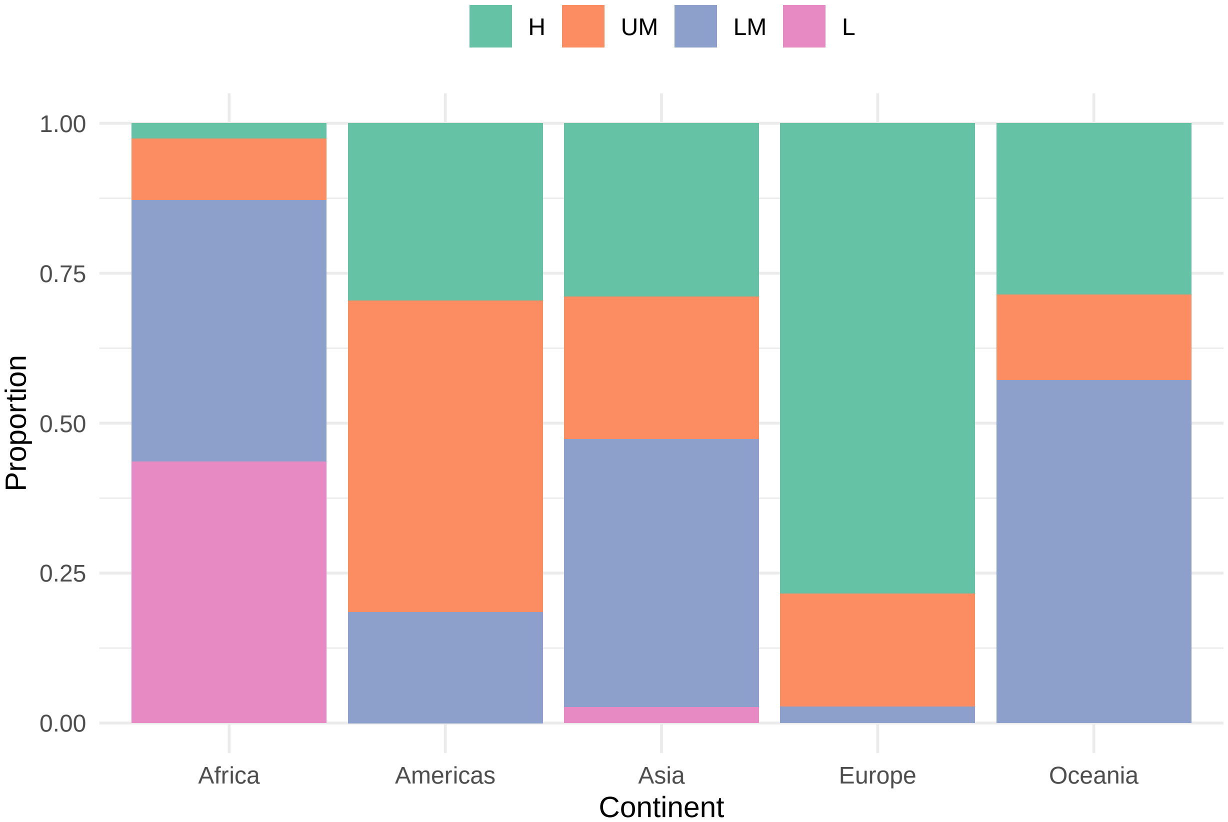
Investigate how the income status is distributed in the different continents.

- 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.

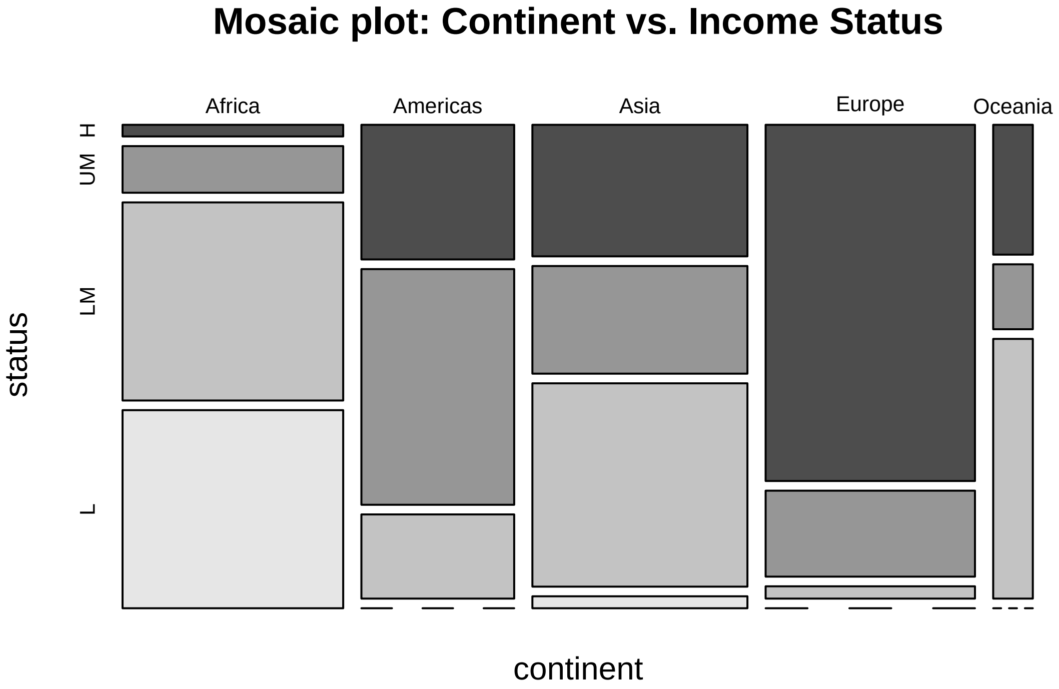
- Create a mosaic plot of continents and income status using base R functions.
- Briefly comment on the differences between the three plots generated to investigate the income distribution among the different continents.



This plot shows how many countries in each continent belong to each income group. Europe has the highest number of high-income (H) countries, while Africa has the most lower-middle (LM) and low-income (L) ones. Asia and the Americas have a more even mix. Oceania has just a few countries overall, mostly LM and UM.



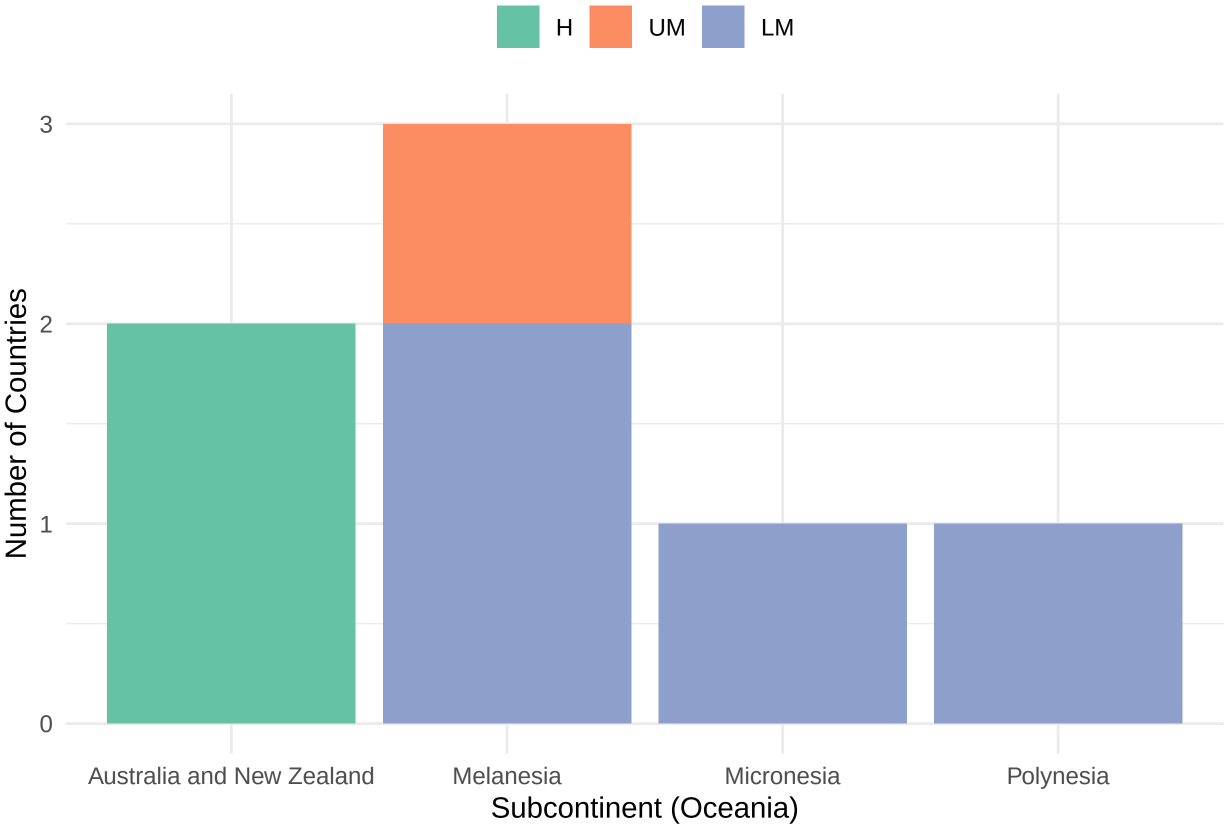
This plot makes it easier to compare proportions across continents. In Africa, the majority of countries are LM or L. Europe is mostly H. Asia has a mix, but LM is most common. The Americas lean more toward UM, while Oceania is mostly LM. This plot highlights how uneven the global income status distribution is.


Mosaic plot: Continent vs. Income Status

The mosaic plot shows the same info but in a different format. The widths of the boxes show the number of countries per continent, and the height shows how those countries are split by income. It confirms what we saw before: Africa has mostly LM/L, Europe has mostly H, and Asia is very mixed. It's a bit harder to read than the barplots but still useful.

c. Income status in different subcontinents

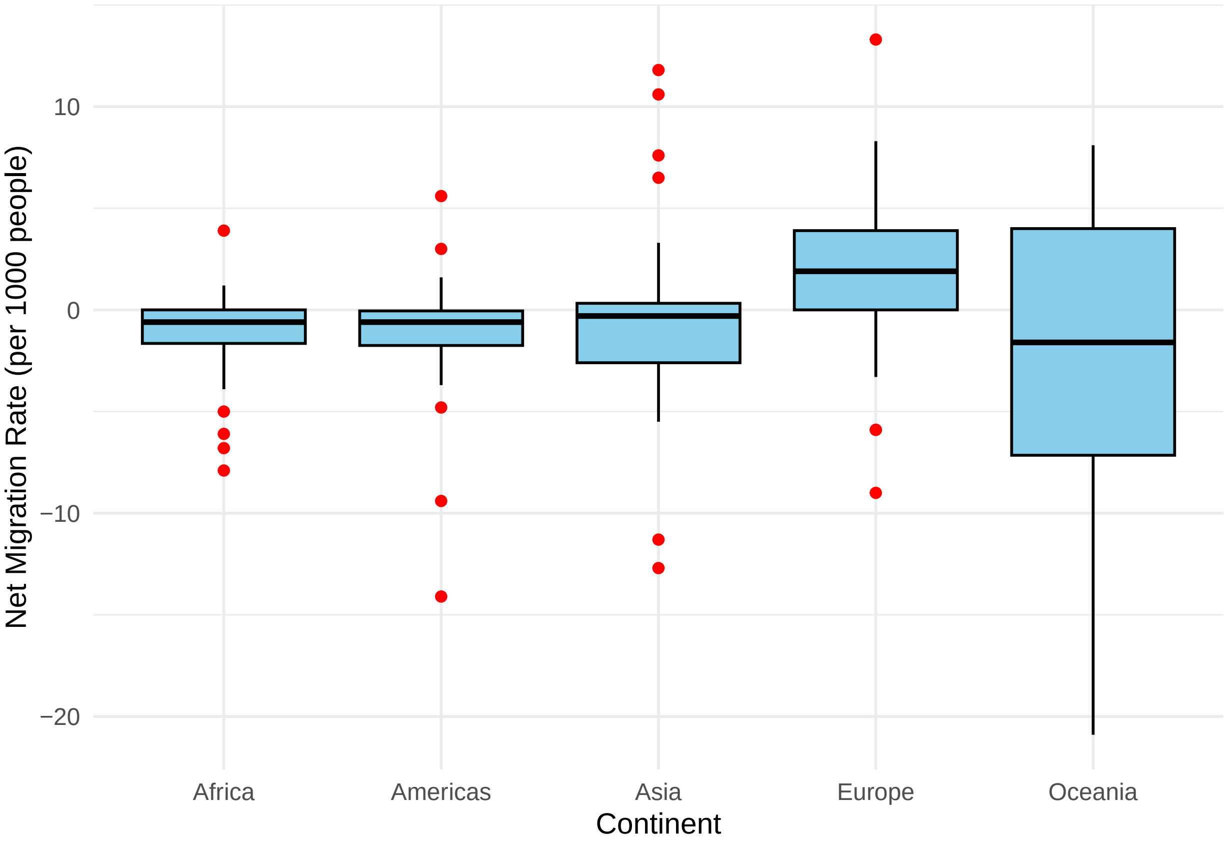
For Oceania, investigate further how the income status distribution is in the different subcontinents. Use one of the plots in b. for this purpose. Comment on the results.



This barplot shows how income status is distributed across Oceania's subregions. Australia and New Zealand are both high-income (H), which makes sense. Melanesia has a mix of LM and UM, while Micronesia and Polynesia are both entirely LM. This shows a clear economic divide between the more developed and less developed parts of Oceania.

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.



```
## # A tibble: 2 x 2
##   country net_migr_rate
##   <chr>           <dbl>
## 1 Lebanon         -88.7
## 2 Syria            27.1
```

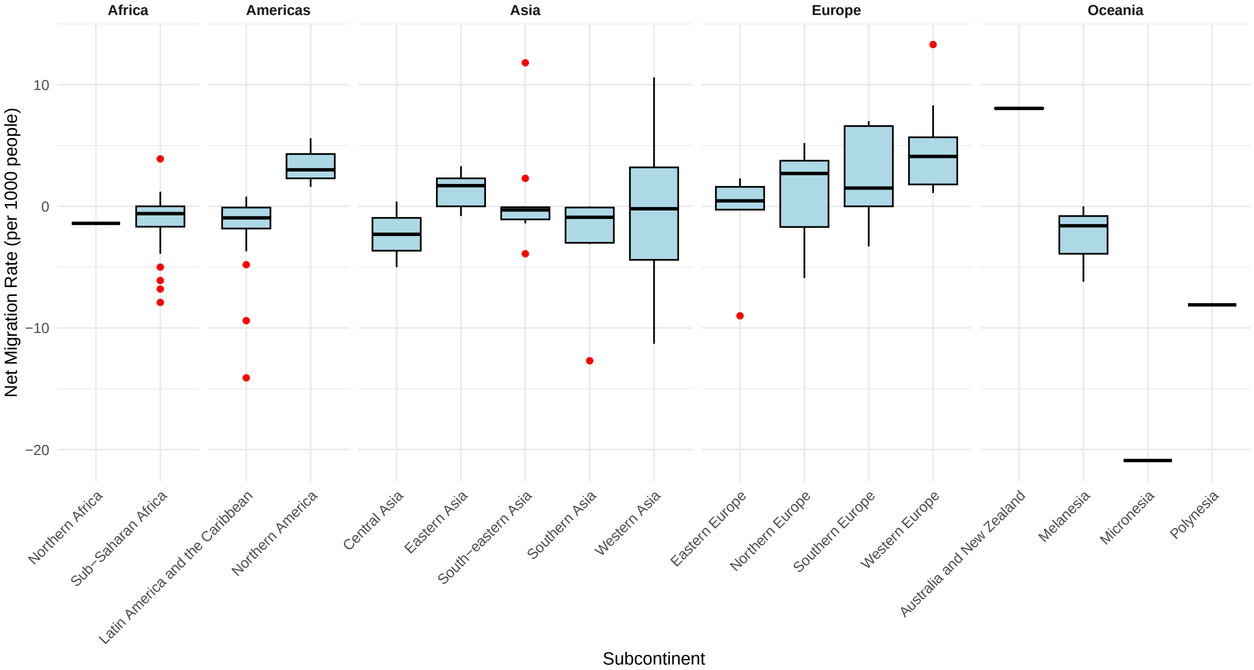
The boxplot shows how net migration rate differs across continents. Europe and Oceania have the highest median migration rates, meaning more countries in those regions have net immigration. Africa and the Americas mostly have lower or negative values. Asia is interesting because it has both large positive and large negative outliers, meaning there's a big variation between countries.

In Asia, Lebanon has the largest negative net migration rate (−88.7), indicating massive outmigration, likely due to political or economic instability. Syria has the largest positive rate (27.1), possibly due to post-war return migration or data anomalies. These outliers show the high variability in migration patterns across Asia.

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.



This faceted boxplot shows net migration rates split by subcontinent and grouped by continent. The variation within continents is really clear here.

In Africa, Northern and Sub-Saharan Africa have similar medians, but Sub-Saharan shows more negative outliers. In the Americas, the Caribbean and Latin America have low/negative medians, while Northern America has higher positive values.

Asia shows huge variation. Western Asia has both very negative and very positive cases, while Central and Eastern Asia are more centered around zero.

In Europe, all subregions have positive medians, with Western Europe having the highest.

In Oceania, Australia and New Zealand are net immigration hubs, while Micronesia has the most negative median (major outmigration), and Polynesia is also clearly negative.

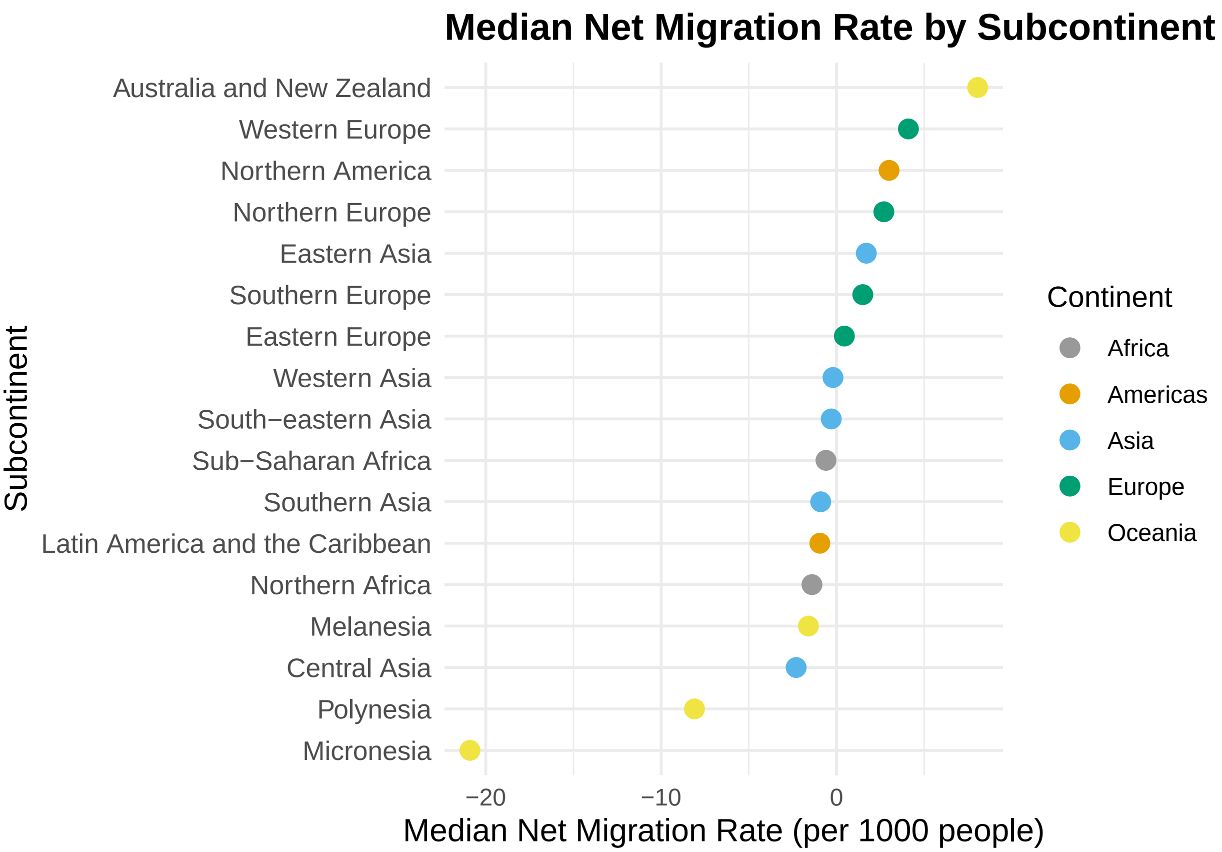
f. Median net migration rate per subcontinent.

The plot in task e. shows the distribution of the net migration rate for each subcontinent. Here you will work on visualizing only one summary statistic, namely the median.

For each subcontinent, calculate the median net migration rate. Then create a plot which contains the sub-regions on the y-axis and the median net migration rate on the x-axis.

- As geoms use points.
- Color the points by continent – use a colorblind friendly palette (see e.g., here).
- Rename the axes.
- Using `fct_reorder` from the **forcats** package, arrange the levels of subcontinent such that in the plot the lowest (bottom) subcontinent contains the lowest median net migration rate and the upper most region contains the highest median net migration rate.
- Comment on the plot. E.g., what are the regions with the most influx? What are the regions with the most outflux?

```
## `summarise()` has grouped output by 'subcontinent'. You can override using the
## `.groups` argument.
```



Australia and New Zealand, Western Europe, Northern America, and Northern Europe have the highest positive median values, which means they are regions with the most influx of people.

On the other hand, Micronesia, Polynesia, Central Asia, Melanesia, and Northern Africa have the most outflux, with negative medians—indicating more people are leaving than arriving.

This highlights clear differences in migration trends between regions: wealthier and more stable areas tend to attract more people, while others are losing population through emigration.

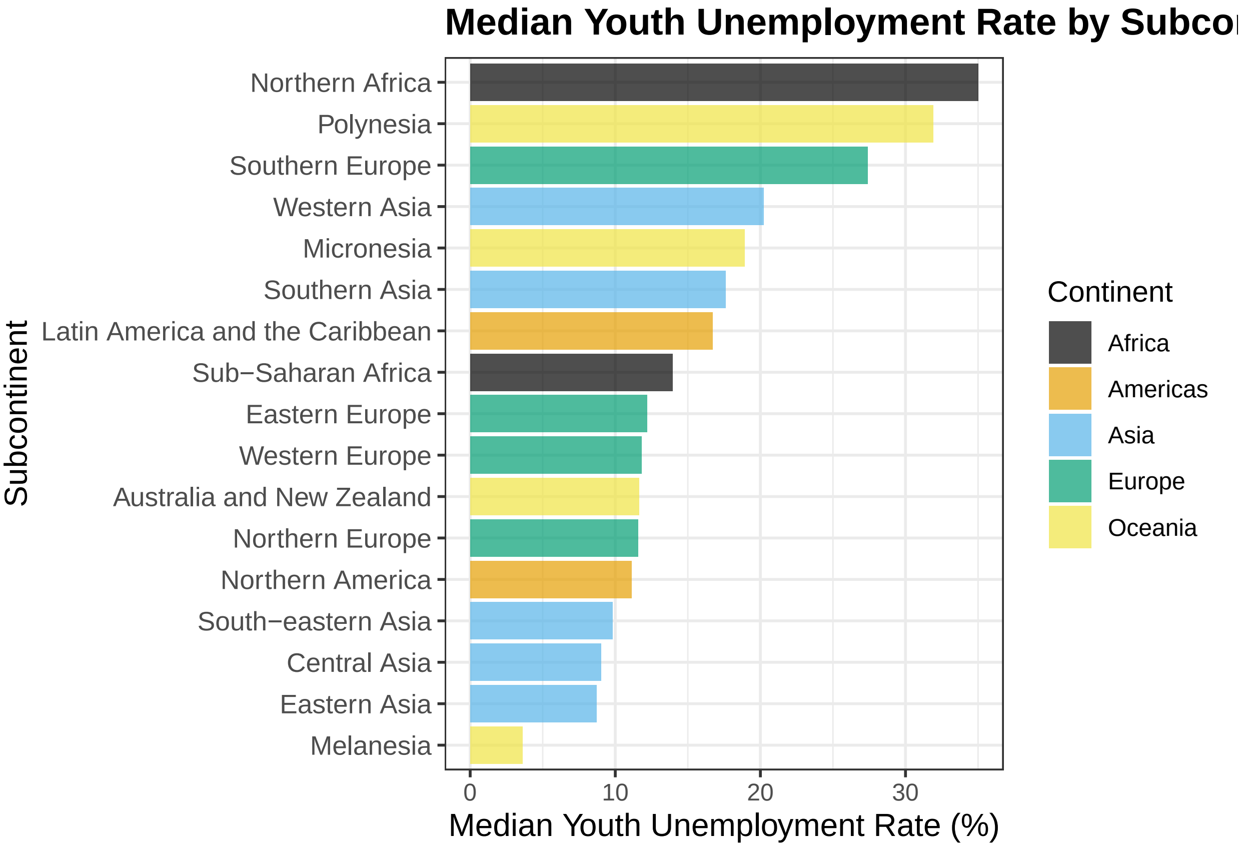
g. Median youth unemployment rate per subcontinent

For each subcontinent, calculate the median youth unemployment rate. Then create a plot which contains the sub-regions on the y-axis and the median unemployment rate on the x-axis.

- Use a black and white theme (`?theme_bw()`)
- As geoms use bars. (hint: pay attention to the statistical transformation taking place in `geom_bar()` – look into argument `stat="identity"`)
- Color the bars by continent – use a colorblind friendly palette.
- Make the bars transparent (use `alpha = 0.7`).
- Rename the axes.

- Using `fct_reorder` from the **forcats** package, arrange the levels of subcontinent such that in the plot the lowest (bottom) subcontinent contains the lowest median youth unemployment rate and the upper most region contains the highest median youth unemployment rate.
- Comment on the plot. E.g., what are the regions with the highest vs lowest youth unemployment rate?

```
## `summarise()` has grouped output by 'subcontinent'. You can override using the
## `.groups` argument.
```



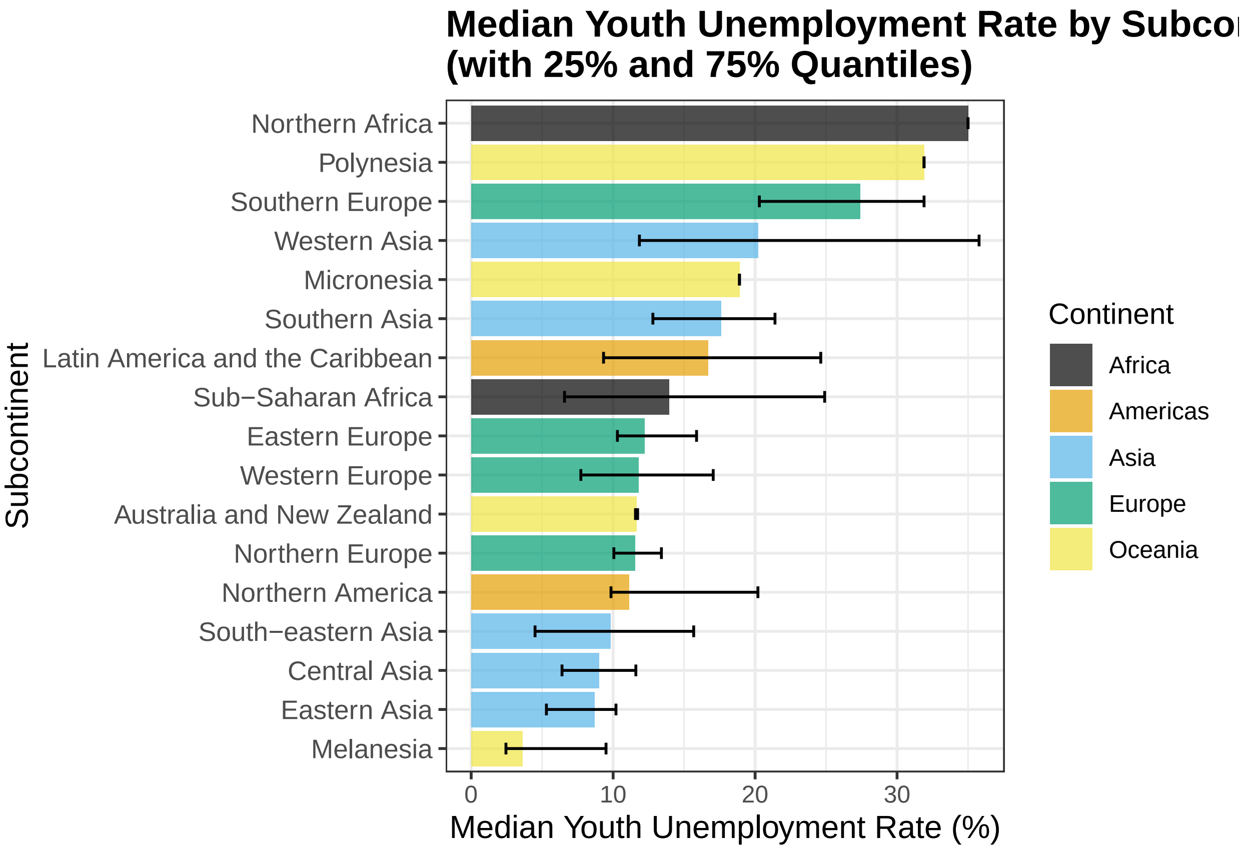
Regions like northern africa and polynesia show the highest median youth unemployment rates (above 30%). On the other end, melanesia, eastern asia, and central asia have the lowest median youth unemployment rates (all under 10%).

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

The value displayed in the barplot in g. is the result of an aggregation, so it might be useful to also plot error bars, to have a general idea on how precise the median unemployment is. This can be achieved by plotting the error bars which reflect the standard deviation or the interquartile range of the variable in each of the subcontinents.

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**.

```
## `summarise()` has grouped output by 'subcontinent'. You can override using the
## `.groups` argument.
```



Some subcontinents like northern africa, western asia, and sub-saharan africa have wide error bars, meaning there's a lot of variation between countries. others like northern europe, eastern europe, and melanesia are more consistent.

Even though northern africa and polynesia have the highest medians, the error bars show that some countries in those regions might be doing better or worse than the average. overall, this plot helps to see where youth unemployment is both high and uneven.

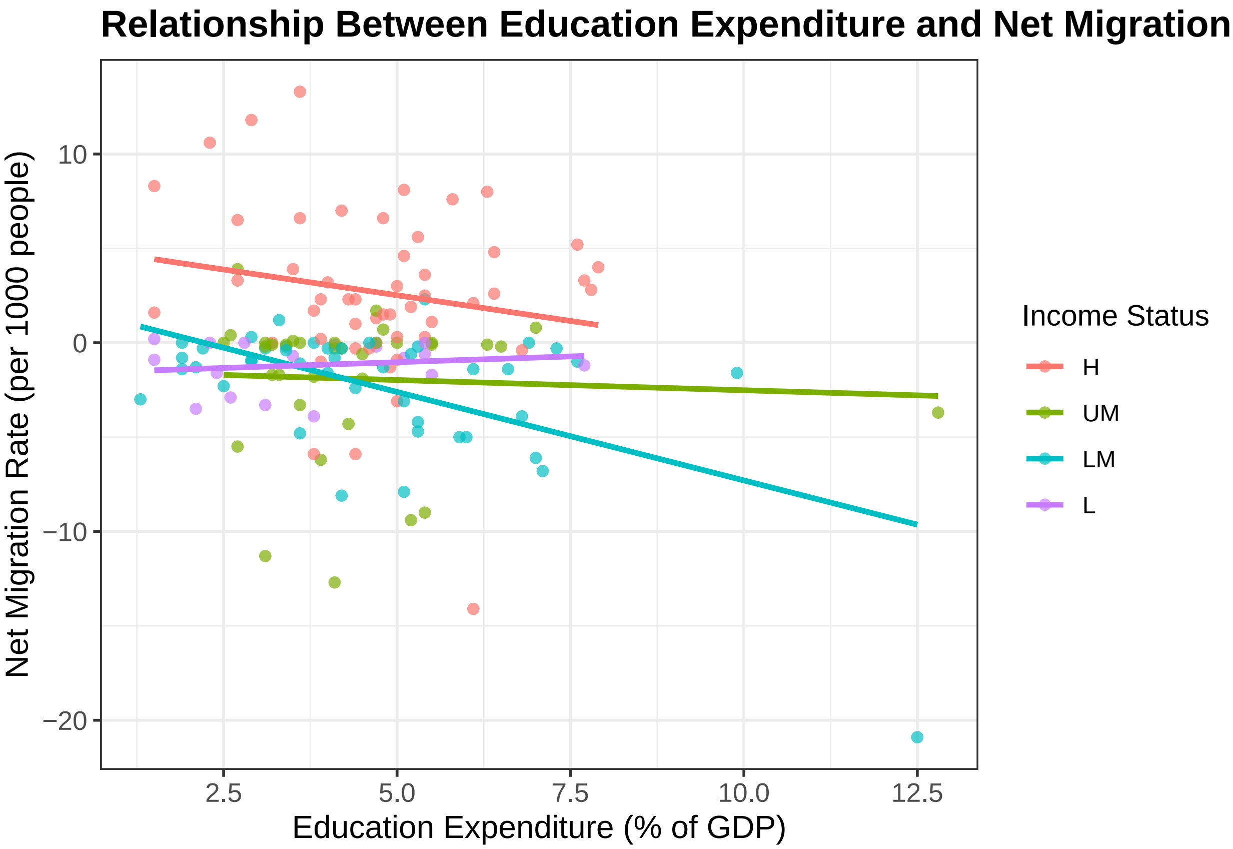
i. Relationship between education expenditure and net migration rate

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'
```



The plot shows the relationship between education expenditure and net migration rate, separated by income status. Overall, the trends are pretty weak.

For high-income countries (red), there's a slight negative trend. More spending is linked with a small drop in net migration. low-income countries (purple) show almost no trend at all.

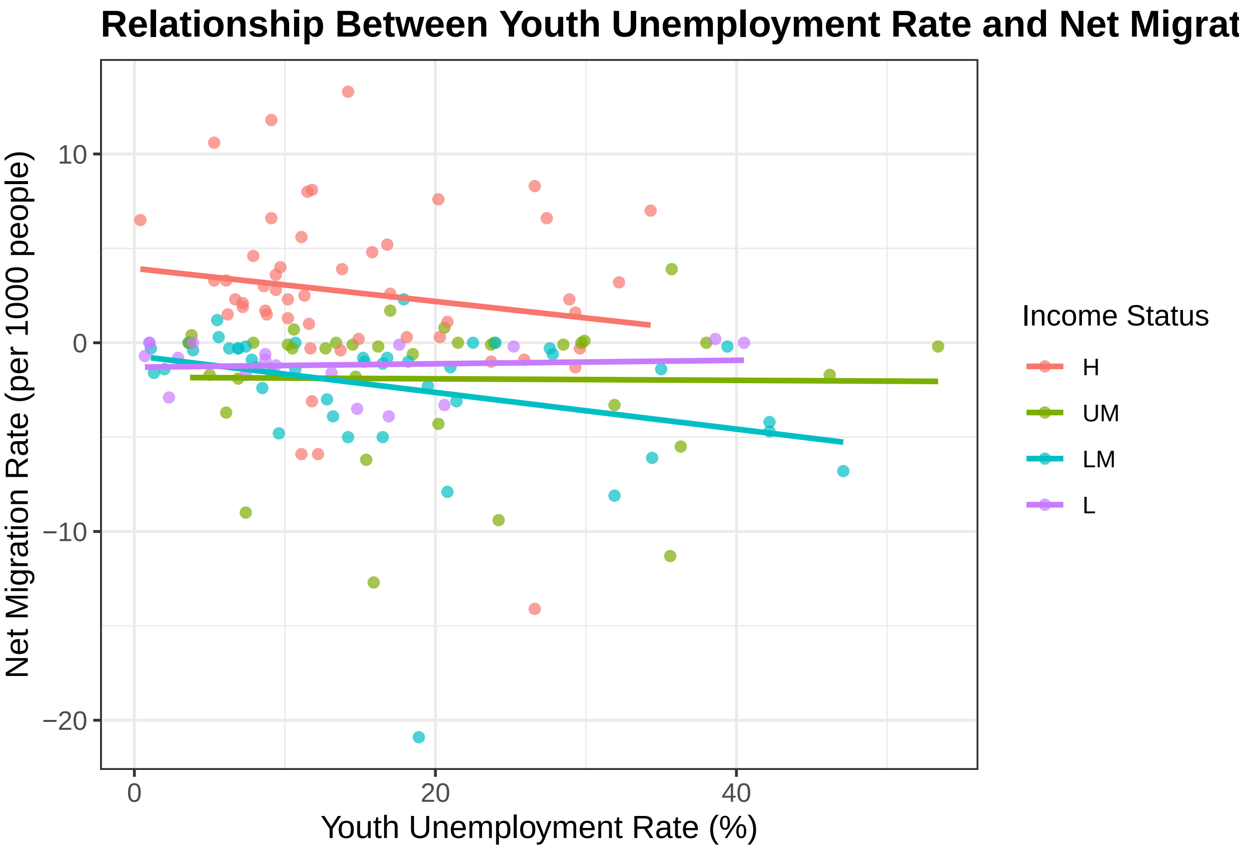
The most visible slope is for lower-middle-income countries (blue), where higher education spending is linked with a strong decrease in net migration.

This suggests that in some income groups, education spending might influence migration patterns, but it's not a consistent effect across all groups.

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'
```



Most income groups show a slight negative trend. As youth unemployment increases, net migration tends to drop.

This trend is most noticeable for lower-middle-income (LM) countries (blue), where higher youth unemployment is linked to more people leaving the country.

For high-income (H) and upper-middle-income (UM) countries, the relationship is weaker but still slightly negative. low-income (L) countries (purple) show almost no change.

In general, higher youth unemployment is associated with more outmigration, especially in LM countries.

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.

```
## # A tibble: 6 x 11
##    ...1 country     ISO3  continent subcontinent              status expenditure
##   <dbl> <chr>       <chr> <chr>     <chr>                     <fct>        <dbl>
## 1     1 Afghanistan AFG   Asia      Southern Asia             L              4.1
## 2     2 Albania     ALB   Europe    Southern Europe           UM             3.6
## 3     6 Angola      AGO   Africa    Sub-Saharan Africa        LM             3.4
## 4     9 Argentina   ARG   Americas  Latin America and the Ca~ UM             5.5
## 5    10 Armenia     ARM   Asia      Western Asia              UM              2.7
## 6    12 Australia   AUS   Oceania   Australia and New Zealand H               5.1
## # i 4 more variables: youth_unempl_rate <dbl>, net_migr_rate <dbl>,
## #   low_yu <chr>, high_nmr <chr>
```

```
## # A tibble: 6 x 4
##   ISO3   Rank Country       Population_2022_thousands
##   <chr> <int> <chr>                             <dbl>
## 1 IND       1 India                           1417173
## 2 CHN       2 China                           1412175
## 3 USA       3 United States                    333288
## 4 IDN       4 Indonesia                        275501
## 5 PAK       5 Pakistan                         235825
## 6 NGA       6 Nigeria                          218541
```

```
## # A tibble: 6 x 11
##    ...1 country     ISO3  continent subcontinent              status expenditure
##   <dbl> <chr>       <chr> <chr>     <chr>                     <fct>        <dbl>
## 1     1 Afghanistan AFG   Asia      Southern Asia             L              4.1
## 2     2 Albania     ALB   Europe    Southern Europe           UM             3.6
## 3     6 Angola      AGO   Africa    Sub-Saharan Africa        LM             3.4
## 4     9 Argentina   ARG   Americas  Latin America and the Ca~ UM             5.5
## 5    10 Armenia     ARM   Asia      Western Asia              UM             2.7
## 6    12 Australia   AUS   Oceania   Australia and New Zealand H              5.1
## # i 4 more variables: youth_unempl_rate <dbl>, net_migr_rate <dbl>,
## #   low_yu <chr>, high_nmr <chr>
```

```
## Joining with `by = join_by(ISO3)`
```

```
## # A tibble: 6 x 14
##    ...1 country     ISO3  continent subcontinent              status expenditure
##   <dbl> <chr>       <chr> <chr>     <chr>                     <fct>        <dbl>
## 1     1 Afghanistan AFG   Asia      Southern Asia             L              4.1
## 2     2 Albania     ALB   Europe    Southern Europe           UM             3.6
## 3     6 Angola      AGO   Africa    Sub-Saharan Africa        LM             3.4
## 4     9 Argentina   ARG   Americas  Latin America and the Ca~ UM             5.5
## 5    10 Armenia     ARM   Asia      Western Asia              UM             2.7
## 6    12 Australia   AUS   Oceania   Australia and New Zealand H              5.1
## # i 7 more variables: youth_unempl_rate <dbl>, net_migr_rate <dbl>,
## #   low_yu <chr>, high_nmr <chr>, Rank <int>, Country <chr>,
## #   Population_2022_thousands <dbl>
```

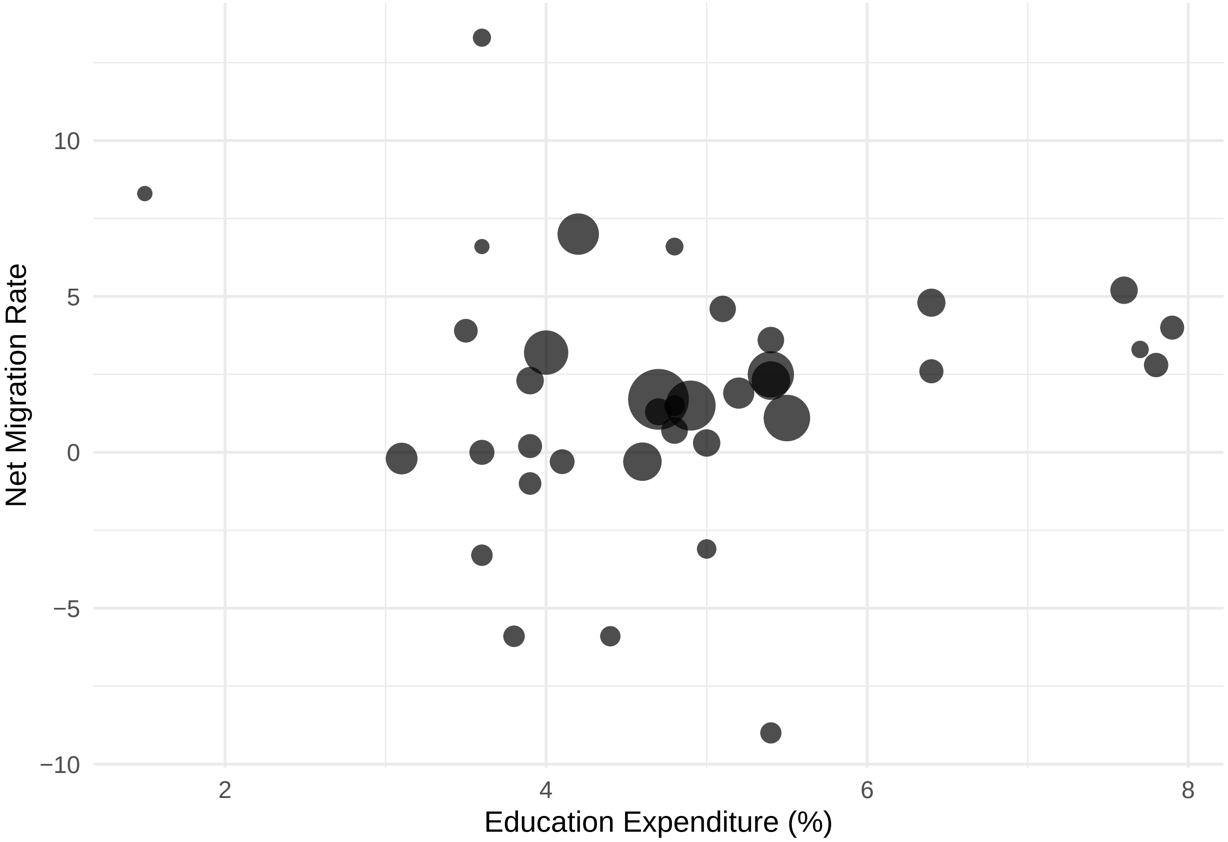
This join makes sense, because the larger dataframe df_clean_2020 contains more countries than the original clean_df dataframe. All of the rows in the clean_df have a match in the right dataframe, we therefore see no NAs in the new dataframe.

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.



The scatterplot shows that most European countries spend between 3% and 6% of GDP on education. Net migration rates vary, with most values between −5 and +10, but there's no strong visible trend. Countries with higher education spending don't clearly have higher or lower net migration. Population size (point size) also doesn't reveal a clear pattern. Larger countries are spread across the plot. Overall, the relationship between education spending and net migration in Europe appears weak or non-existent.

m. Interactive plot

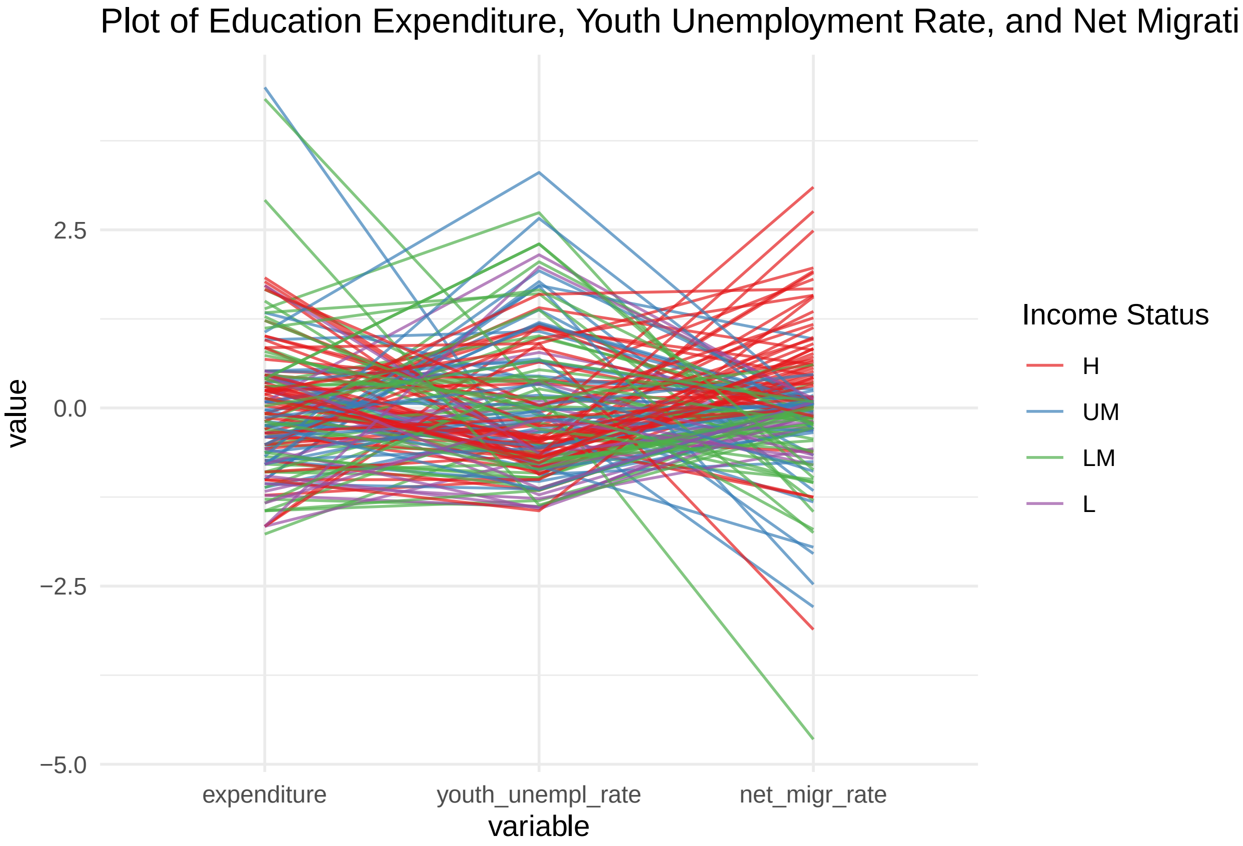
On the merged data set from Task k., using function `ggplotly` from package **plotly**
re-create the scatterplot in Task l., but this time for all countries. Color the points according to their continent.

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.



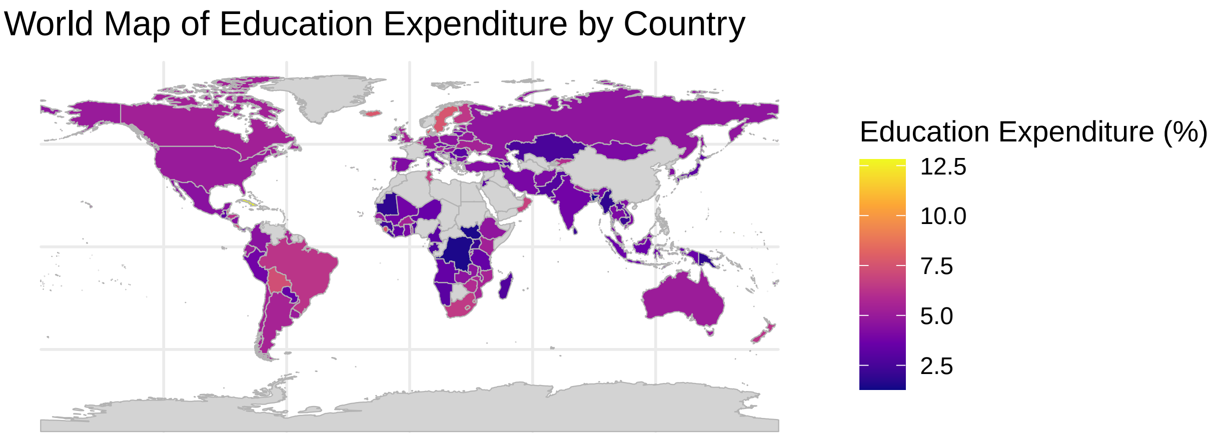
This plot shows how countries differ in education expenditure, youth unemployment, and net migration, depending on their income status. the lines are colored by income group, and we can see that high-income (red) countries generally have higher education spending, lower youth unemployment, and mostly positive net migration.

Lower-income countries (green and purple) tend to spend less on education and have more variation in youth unemployment and migration. their lines are more spread out, which shows that there's a lot of difference between countries in these groups.

Overall, the plot helps show that higher-income countries have more stable indicators, while lower-income ones vary more.

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.


World Map of Education Expenditure by Country
Education Expenditure (%)
12.5
10.0
7.5
5.0
2.5

The map shows that education expenditure varies a lot between countries. Some African countries spend a low percentage of their gdp on education, while many European, North American, and Asian countries spend around 4–6%. The grey areas mean there’s no data available.