Case Study 2

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.*

The CIA World Factbook provides intelligence on various aspects of 266 world entities, including history, people, government, economy, energy, geography, environment, communications, transportation, military, terrorism, and transnational issues. This case study involves analyzing 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)

The data was sourced from the CIA World Factbook Archives. You are required to use `dplyr` for data manipulation, while any package can be used for importing data.

Tasks:

a. Data Import and Cleaning

Load the following datasets from TUWEL and ensure that missing values are handled correctly and column names are clear. Each dataset should ultimately contain only two columns: **country** and the respective variable. Note that some data sets also contain information on the year when the value was last updated.

- `rawdata_369.txt` which contains the (estimated) public expenditure on education as a percent of GDP. *Pay attention! The delimiter is 2 or more white spaces (one space would not work as it would separate country names which contain a space); you have to skip the first two lines.*
- `rawdata_373.csv` which contains the (estimated) youth unemployment rate (15-24) per country
- `rawdata_347.txt` which contains (estimated) net migration rate per country.

b. Merging Raw Data

Merge the datasets using `dplyr` on a unique key and retain the union of all observations.

- What key are you using for merging?
- Return the dimensions of the merged dataset.

c. Enriching Data with Income Classification

Obtain country income classification (low, lower-middle, upper-middle, high) from the World Bank and merge it with the dataset.

- Identify common variables between datasets. Can they be used for merging? Why or why not?
- Since ISO codes are standardized, download and use the CIA country data codes for merging. Make sure you are not losing any of the countries in your original data set when merging.

d. Adding Geographical Information

Introduce continent and subcontinent (or region) data for each country.

- Find and download an appropriate online resource.
- Merge this information into the dataset, naming the final dataset `df_vars`. Make sure you are not losing any of the countries in your original data set when merging.

e. Data Tidiness and Summary Statistics

- Evaluate the tidiness of `df_vars` (observational units, variables, fixed vs. measured variables). Make adjustments to tidy the data, if necessary.
- Create a frequency table for the income status variable and briefly interpret the results.
- Analyze the distribution of income status across continents by computing absolute and relative frequencies. Comment on the findings.
- Using the distribution of income status across continents, identify which countries are the only ones in their income group across the continent. Discuss briefly.

f. Further Summary Statistics and Insights

- Create a table of average (mean and median) values for expenditure, youth unemployment rate and net migration rate separated into income status. Make sure that in the output, the ordering of the income classes is proper (i.e., L, LM, UM, H or the other way around). Briefly comment the results and any differences between the mean and median.
- Look at the standard deviation and the interquartile range of the variables per income status instead of the location statistics above. Do you gain additional insights? Briefly comment the results.
- Extend the analysis of the statistics median and IQR to **each income status and continent combination**. Play around with displaying the resulting table. Use `pivot_longer()` and/or `pivot_wider()` to generate different outputs. Discuss the results as well as the readability of the different tables.
- Identify countries performing well in terms of both **youth unemployment** and **net migration rate** (top 25% in net migration and bottom 25% in youth unemployment within their continent).

g. Conditional probabilities

Estimate the following based on the observed frequencies in the data:

- What is the (posterior or conditional) probability that a European country belongs to the high income group? What is the prior probability that a country belongs to the high income group?
- Given a country has high youth unemployment (above %25), what is the probability that it also has negative net migration?

h. Simpson's Paradox Analysis

Investigate whether an overall trend in youth unemployment rate in the high and low income groups reverses when analyzed at the continent level. E.g., does the youth unemployment rate appear lower in low-income countries overall, but higher when controlling for continent? Explain the results and possible reasons behind this paradox.

i. Data Export

Export the final tidy dataset from e. as a **CSV** with:`;` as a separator; `.` representing missing values; no row names included. Upload the `.csv` to TUWEL, together with the submission.