Case Study 2

AKSTA Statistical Computing

The .Rmd and .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 .PDF is not in a decent form.

Data

The CIA World Factbook provides basic intelligence on the history, people, government, economy, energy, geography, environment, communications, transportation, military, terrorism, and transnational issues for 266 world entities.

In this case study you will work with world data from 2020 which contains information on

- median age
- population growth rate

for most world entities.

For data manipulation, **dplyr** functions should be used. For importing data, any package can be used.

Tasks:

1. From https://datahub.io/core/country-codes obtain the csv containing information related to the countries and load it in R. Keep only the columns containing official country name in English, ISO 3166 country codes with 2 and 3 alpha-numeric characters, development status (developing vs developed) and region and sub-region.
2. Load in R the following data sets which you can find in TUWEL. For each data set, ensure that missing values are read in properly, that column names are unambiguous. Each data set should contain at the end only two columns: country and the variable.

- `rawdata_343.txt` which contains the (estimated) median age per country. *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. Hint*: you can look into function `read.fwf` or the **readr** corresponding function. It might also be useful to use **tidyr** functions to unite some columns back or separate them.
- `rawdata_373.csv` which contains the (estimated) youth unemployment rate (15-24) per country

3. Merge the two data sets containing raw data using **dplyr** function on the unique keys. Keep the union of all observations in the two tables. What key are you using for merging?

4. Merge the resulting data set above with the data set containing country information (from point 1) using **dplyr** functions on the unique keys. Name this new object `df_vars`.

- Inspect the country names and check if it would be a reliable variable for matching. Why or why not?
- In TUWEL you have the file `CIA_factbook_matching_table_iso.xlsx` which provides reliable matching information based in ISO codes. Use this for the final merging of the data sets.

5. Discuss on the tidyness of the data set `df_vars`. What are the observational units, what are the variables? What can be considered fixed vs measured variables? Tidy the data if needed.
6. Count the number of developing vs. developed countries in the merged data set.
7. Count how many countries per region does the merged data set contain.
8. Count the number of developing vs. developed countries for each region.
9. Create a table of average values and the standard deviation for both median age and youth unemployment rate separated into developing and developed countries (hint: eliminate observations with missing development status beforehand). Comment briefly on the results.
10. Repeat the analysis in the previous task for each development status and region combination.
11. In `df_vars` create two additional indicator variable `above_average_median_age` which contains a `yes` is the country's median age lies above the **region average** and `no` otherwise. Create another `above_average_yu` which contains the same information but for the youth unemployment variable.
12. Export the final data set to a csv with ";" separator and "." as a symbol for missing values; no rownames should be included in the csv. Upload the .csv to TUWEL together with your .Rmd and PDF.