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.

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
- youth unemployment rate
- net migration rate (difference between the number of persons entering and leaving a country during the year per 1,000 persons – based on midyear population)

for most world entities. The data was downloaded from https://www.cia.gov/the-world-factbook/about/archives/.

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

Tasks:

a.

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
- `rawdata_347.txt` which contains (estimated) net migration rate per country.

b.

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

c.

You will acquire more country level information such as the classification of the country based on income. Such an information can be found at https://datahelpdesk.worldbank.org/knowledgebase/articles/906519. From there extract the classification for 2020 into low/lower-middle/upper-middle/high income countries.

d.

Merge this information to the data set in b.

1. What are the common variables? Can you merge using them? Why or why not?
2. A reliable merging for countries are ISO codes as they are standardized across data sources. Download the mapping of ISO codes to countries from https://www.cia.gov/the-world-factbook/references/country-data-codes/ and load it into R.
3. Merge the data sets using the ISO codes.

e.

Introduce into the data set information on continent for each country and subcontinent (region). You should find a way to gather this data. You can find an appropriate online resource, download the data and merge the information with the existing data set. Name the merged data set `df_vars`.

f.

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.

g.

Make a frequency table for the status variable in the merged data set. Briefly comment on the results.

h.

What is the distribution of income status in the different continents? Compute the absolute frequencies as well as the relative frequency of status within each continent. Briefly comment on the results.

i.

From h. identify the countries which are the only ones in their respective group. Explain in few words the output.

j.

For each continent count the number of sub-regions in the data set. How granular are the subcontinents that you employ in the analysis?

k.

Look at the frequency distribution of income status in the subregions of Nort- and South-Americas. Comment on the results.

l.

Dig deeper into the low-middle income countries of the Americas. Which ones are they? Are they primarily small island states in the Caribbean? Comment.

m.

Create a table of average values for median age, 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.

n.

Look also at the standard deviation instead of the mean in m. Do you gain additional insights? Briefly comment the results.

o.

Repeat the analysis in m. for each income status and continent combination. Discuss the results.

p.

Identify countries which are doing **well** in terms of both youth unemployment and net migration rate (in the top 25% of their respective continent in terms of net migration rate and in the bottom 25% of their respective continent in terms of youth unemployment).

r.

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 .html (or .pdf).