Exercise: Forest and GDP

Stephan.Huber@hs-fresenius.de

Data Science and Data Analytics (short mock exam)

Please answer all (!) questions in an R script. Normal text should be written as comments, using the '#' to comment out text. Make sure the script runs without errors before submitting it. Each task (starting with 1) is worth XXX points. You have a total of XXX minutes of editing time. Please do not forget to number your answers.

When you are done with your work, save the R script, export the script to pdf format and upload the pdf file.

Suppose you aim to empirically examine the impact of economic activity on the environment, i.e., forest area (% of land area). The data set that we use in the following is 'forest.Rdata'. I downloaded from the Worldbank. In 'data_forest.R' I document the data preperation. See:

`https://raw.githubusercontent.com/hubchev/courses/main/scr/data_forest.R`

Data description

gdp GDP (constant 2015 US$)

GDP at purchaser's prices is the sum of gross value added by all resident producers in the

economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2015 prices, expressed in U.S. dollars. Dollar figures for GDP are converted from domestic currencies using 2015 official exchange rates. For a few countries where the official exchange rate does not reflect the rate effectively applied to actual foreign exchange transactions, an alternative conversion factor is used. (NY.GDP.MKTP.KD)

gdp_growth GDP growth (annual %)

Annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2015 prices, expressed in U.S. dollars. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. (NY.GDP.MKTP.KD.ZG)

unemployment Unemployment, total (% of total labor force) (modeled ILO estimate)

Unemployment refers to the share of the labor force that is without work but available for and seeking employment. (SL.UEM.TOTL.ZS) See: https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS

income World Bank Country and Lending Groups

For the current 2022 fiscal year, low-income economies are defined as those with a GNI per capita, calculated using the World Bank Atlas method, of $1,045 or less in 2020; lower middle-income economies are those with a GNI per capita between $1,046 and $4,095; upper middle-income economies are those with a GNI per capita between $4,096 and $12,695; high-income economies are those with a GNI per capita of $12,696 or more.

forest Forest area (% of land area)

Forest area is land under natural or planted stands of trees of at least 5 meters in situ, whether productive or not, and excludes tree stands in agricultural production systems (for example, in fruit plantations and agroforestry systems) and trees in urban parks and gardens. (AG.LND.FRST.ZS)

pop Population, total - Spain

Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates. (SP.POP.TOTL)

unemployment_dif

Yearly change in unemployment: unemployment(t)-unemployment(t-1)

gdppc GDP per capita (gdppc = gdp/pop)

---

(1) Set your working directory.

(2) Clear your global environment.

(3) Install and load the following packages: 'tidyverse', 'sjPlot', and 'ggpubr'

(4) Download and load the data, respectively, with the following code:

```
load(url("https://github.com/hubchev/courses/raw/main/dta/forest.Rdata"))
```

If that is not working, you can also download the data from ILIAS, save it in your working directory and load it from there with:

```
load("forest.Rdata")
```

(5) Show for all numerical variables the summary statistics including the mean, median, minimum, and the maximum.

(6) Rename the variable 'country.x' to 'country' in the dataset 'df'.

(7) Create a variable that indicates the gdp in million US $ ('gdp' divided by 1,000,000). Name the variable 'gdp_mio'.

(8) Create a table showing the mean values of the variables 'gdp_mio', and 'forest' over time separately by region. Use the pipe operator. (Tip: See below for how your result should look like.)

```
## # A tibble: 7 x 3
##   region                     m_gdp_mio m_forest
##   <chr>                          <dbl>    <dbl>
## 1 East Asia & Pacific          623388.    48.2
## 2 Europe & Central Asia        382810.    29.4
## 3 Latin America & Caribbean    135439.    46.6
## 4 Middle East & North Africa   117698.     3.04
## 5 North America               9117487.    35.8
## 6 South Asia                   235150.    25.4
## 7 Sub-Saharan Africa            25952.    32.3
```

(9) Create a table showing the mean values of the variables 'gdppc', and 'forest' over time separately by region. Use the pipe operator. (Tip: See below for how your result should look like.)

```
## # A tibble: 7 x 3
##   region                     m_gdp_pc m_forest
##   <chr>                         <dbl>    <dbl>
## 1 East Asia & Pacific          12248.    48.2
## 2 Europe & Central Asia        21003.    29.4
## 3 Latin America & Caribbean     8951.    46.6
## 4 Middle East & North Africa   13937.     3.04
```

```
## 5 North America                46123.    35.8
## 6 South Asia                    2080.    25.4
## 7 Sub-Saharan Africa            1759.    32.3
```

(10) Investigate the relationship of economic activity measured by the GDP and the GDP per capita with a country's forest area. Therefore, graphically visualize the relationsship and consider things like correlation analysis and regression analysis.