

## Visualization

Prof. Bernhard Schmitzer, Uni Göttingen, summer term 2023

### Problem sheet 3

- *Submission by 2023-06-20 18:00 via StudIP as a **single PDF**. Please combine all results into one PDF. If you work in another format (markdown, jupyter notebooks), convert them to PDF before submission.*
- *Use Python 3 for the programming tasks as shown in the lecture. If you cannot install Python on your system, the GWDG jupyter server at <https://jupyter-cloud.gwdg.de/> might help. Your submission should contain the final images as well as the code that was used to generate them.*
- *Work in groups of two to three. Clearly indicate names and matrikelnr of all group members at the beginning of the submission.*

#### Exercise 3.1: smoking and life expectancy.

1. The file `smokers.npz` contains an array `data` of dimensions  $N_{\text{pers}} \times 3$  of type `int` which contains information about  $N_{\text{pers}} = 20,000$  persons from  $N_{\text{countries}} = 20$  countries. Each row represents one person. The first column encodes the country that they live in, by an integer from 0 to 19. The second column encodes whether that person was a regular smoker (`=1`) or not (`=0`). The third column gives the age in full years that this person reached at the time of their death. Import this array into python.
2. Plot histograms over ages for the total population, for smokers and non-smokers (with absolute counts in each bin). In addition, plot the normalized histograms (where entries in all bins sum to one), which represent an approximate probability density function.
3. For each country, determine the average life expectancy of people and the fraction of smokers. Visualize this information.
4. For each country, determine the life expectancy of smokers and non-smokers. Visualize this information.
5. Generate a 2d histogram of people over their country and their age, for smokers and non-smokers. Find a way to visualize this in a single plot as a multi-color image.  
*Hints:* Think about a good way of normalizing the color channels. Think about a reasonable ordering of the countries.
6. For smokers and non-smokers, the relation ‘country that a person lived in’ to ‘age of that person’ is a stochastic functional relation. Visualize this relation for both groups (smokers, non-smokers) to obtain a single chart which conveys the information how long smokers and non-smokers tend to live in various countries and how large the variation of ages is.

#### Exercise 3.2: salary trends.

1. The file `salaries.npz` contains an array `salaries` of dimensions  $N_{\text{pers}} \times N_{\text{years}}$  which contains the yearly salaries in Euros of  $N_{\text{pers}} = 200$  persons over  $N_{\text{years}} = 20$  years, from 2001 to 2020. In addition, it contains an array `inflation_factors` of dimensions  $N_{\text{years}} - 1$  with the inflation rate for Euros *in percent* for the years 2001 to 2019 (careful: it is formatted as array of shape  $1 \times (N_{\text{years}} - 1)$ ). Import both arrays into python.
2. Compute the effective deflated value of one Euro in each year from 2001 to 2020 in terms of 2001 Euros.
3. You are a consultant for the governing party in the fictional country. From the data create a chart that demonstrates that the average salary has increased substantially over the last 20 years.
4. Now you work for the opposition party. Create a chart that shows that while salaries have increased substantially for high-income groups, for a large fraction of employees the effective salaries have stagnated. Explain what data you show in your chart (1-2 sentences).