Visualization
Prof. Bernhard Schmitzer, Uni Göttingen, summer term 2024

Exam 2024-07-15

General remarks.

- Work with Python and jupyter notebooks, e.g. via the GWDG jupyter server at `https://jupyter-cloud.gwdg.de/`.
- Your submission should consist of (multiple) notebook files (optionally in text format, e.g. via jupytext), combined into a single zip/tar archive.
- Do not include the data files in your submission.
- Submit via StudIP, similar to the problem sheets. Submission deadline: 2024-07-26 23:59
- Indicate your enrolment numbers clearly at the beginning of each notebook file.
- Groups of two students do not have to complete problems 1.4 and 2.3.
- To judge the expected amount of work in the assignments, an approximate number of suggested charts is given in parentheses. There is no single right solutions to the assignments. Sometimes more or even less charts may be appropriate, depending on your approach.
- If you do not understand a problem description, do not hesitate to contact `schmitzer@cs.uni-goettingen.de`.

Problem 1: Census data, Germany 2022.

In this problem we work with census data obtained in Germany in 2022. The original data is available at `https://www.zensus2022.de/EN/Census_results/_inhalt.html`. But the data itself is only documented in German. Consequently, for the exam we provide a translated and slightly preprocessed version of the data as `census.csv` with some documentation and meta data in `census_description.pdf`.

1. **Import and pre-processing.** Import `census.csv` into Python as a pandas dataframe, process missing values and particular formatting issues and make sure that each column in the end has the appropriate data type. For this problem we will only work with data on the lowest regional level (municipality, German: 'Gemeinde'). Filter the data accordingly.
2. **Analysis with respect to municipality size.** First, we analyze the variation of municipalities with respect to their size (=number of inhabitants).
   (a) Show the distribution of municipalities with respect to size. (1 chart)
   (b) Analyze and visualize the relation between the fraction of non-German inhabitants and municipality size. There is a noticeable trend, and your charts should represent it clearly. (2 charts)
   (c) Similarly, analyze and visualize the relation between the fraction of women (or men) and municipality size. (2 charts)

3. **Relation between old age, fraction of women, and fraction of widowers.** Analyze and visualize the relation between the fraction of people aged 75 or older, the fraction of women, and the fraction of widowers, over the set of municipalities. What is the relation between the three quantities? (3-4 charts)

4. **Dimensionality reduction with respect to age distribution.**

   (a) Add new columns to the original dataset that contain the cumulative age distribution, i.e. add columns `age_cumul_xx` for `xx` $\in \{\texttt{01}, \ldots, \texttt{11}\}$ that contain the relative cumulative number of people in age brackets `01` to `xx` in a given municipality (in particular `age_cumul_11` should always have value 1).

   (b) Perform principal component analysis on the variables

   $$(\texttt{age\_cumul\_01}, \ldots, \texttt{age\_cumul\_11}).$$

   (c) Show the explained variance of each eigenvector and the cumulative variance of the first k eigenvectors with respect to $k \in \{1, 2, \ldots\}$. Visualize the first two eigenvectors and give a brief interpretation. (one sentence per eigenvector, 2-3 charts)

   (d) Show a scatter plot of the 2d PCA embeding of the municipalities, i.e. their projection to the two leading eigenvectors. Which PCA coordinate is correlated strongly with the fraction of people aged 75 or older? Confirm this visually. (2 charts)

Problem 2: Drought Monitor Germany. In this problem we work with data on drought in the topsoil (uppermost 25cm) in Germany between 1951 until 2022, as provided by the the Helmholtz Centre for Environmental Research. The original data is available at https://www.ufz.de/index.php?en=37937. For the exam we provide the data in convenient numpy format in the file `drought.npz` and a description in `drought_description.txt`. Some details on the definition of the soil moisture index (SMI) can be found (in German and English) in the original source description.

1. **Basic heatmap of SMI.**

   (a) Visualize the SMI across Germany as a heatmap / image for a given year and month. Make the plot interactive, such that year and month can be chosen dynamically by the user. Build a similar plot where for a given year the heatmap of the average SMI over this year is shown. (2 charts)

   (b) For better orientation, mark the following cities on the map: Berlin, Cologne, Frankfurt, Göttingen, Hamburg, Leipzig, Munich, Stuttgart. To determine the location of a city on the heatmap, use the `latitude` and `longitude` arrays of the dataset, and simply place each city on the grid cell with the most similar latitude and longitude coordinates. Make sure that the names of the cities can be inferred form the plot. (1 chart, or add to previous charts)

2. **Country-wide averages.**

   (a) For each observed month, compute the average SMI over the whole country (in the following referred to as monthly averages). In addition, for each year, compute the average SMI over the months and the whole country (yearly averages).

(b) Show the yearly averages between 1951 and 2022, visually emphasize the trend in the evolution, and highlight several years of exceptional drought. (1 chart)

(c) For all years between 1951 and 2022, show the monthly averages from January to December, with a focus on the exceptionally dry years identified in the previous chart. (1 chart)

3. **Drought thresholds.** Researchers have suggested various thresholds of the SMI for the definition of droughts. We adopt the following conventions:

SMI range	convention
$(0.20, 0.30]$	abnormally dry
$(0.10, 0.20]$	moderate drought
$(0.05, 0.10]$	severe drought
$(0.02, 0.05]$	extreme drought
$[0.00, 0.02]$	exceptional drought

For each month determine the fraction of area of Germany that falls within each of these categories (this fraction can be computed via the relative number of pixels of the masked image area). Show these fractions over time. To make the plot more legible, apply some additional aggregation, either by averaging over years, or (more sophisticated) by computing a rolling mean over 12 months. (1-2 charts)