Visualization
Prof. Bernhard Schmitzer, Uni Göttingen, summer term 2025

Problem sheet 5

- *Submission by **2025-06-13** 18:00 via StudIP **as a single PDF/ZIP**. Please combine all results into one PDF or archive. If you work in another format (markdown, jupyter notebooks), add a PDF converted version to your 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 up to three. Clearly indicate names and enrollment numbers of all group members at the beginning of the submission.*

Exercise 5.1: temperature data.
We use data from the Deutscher Wetterdienst (DWD). The original data is available at https://www.dwd.de/DE/leistungen/cdc/cdc_ueberblick-klimadaten.html, but all data required for the exercise is once more provided in the zip file of the problem sheet.

1. The file `temperature_data_processed.csv` contains condensed information on temperature measurements at 80 selected measurement stations in Germany from 1781 to 2024. The column `stationid` indicates the id of the corresponding station, `date` contains the date of the measurement in the format YYYYMMDD, `time` contains the time in the format HH, and `temp` contains the temperature measured at the given station at the given date and time in degree Celsius according to some standardized protocol (that has however slightly changed over the years). Import this into Python as a dataframe. Parse the `date` column to add explicit `year`, `month`, and `day` columns. Temperature values of -999 indicate missing data. Remove these rows from the dataframe.
2. Examine and visualize how many stations contributed measurements in each year covered in the dataset.
3. Identify the stations that were active both in 1960 and 2020. We will refer to these as *reference stations*.
4. Filter the data for the intersection (logical and) of the following conditions: The station must be a reference station, the year must be in the interval $[1960, 2020]$, and the time must be either 12 or 14. In the following work with the filtered data.
5. Show for each year, how many measurements are available at time 12 and at 14.
6. Show monthly temperature trends (i.e. warm in summer, cold in winter) by visualizing (a summary of) the distribution of all temperature measurements in a given month. Try different methods and include at least two different charts.
7. Show the long term temperature trend between years 1960 and 2020. Do this for the combined data of all months of each year, and only for the data for a selected month

(e.g. 6). Given the large number of years, boxplot or violin plots will not be appropriate in this setting. Find another good representation.

Hint: By using the split-apply-combine paradigm in pandas, all processing and filtering required for this problem can be done quite easily.