Prof. Dr. Thomas Schultz
Jonathan Lennartz (jlen@uni-bonn.de)
Johannes Grün (gruen@cs.uni-bonn.de)

University of Bonn
B-IT/Computer Science II
April 4, 2023

Summer term 2023

Visual Data Analysis

Assignment Sheet 1

Solution has to be submitted via eCampus by April 10, 2023, 10:00 a.m.

If you have questions concerning the exercises, please use the forum on eCampus.

- Please work on this exercise in **small groups** of 3 students. Submit each solution only once, but clearly indicate who contributed to it by forming a team in eCampus. Remember that all team members have to be able to explain all answers.
- Please submit your answers in PDF format, and your scripts as *.py/*.ipynb files. If you are using Jupyter notebook, please also export your scripts and results as PDF.

Help with Installing the Required Software
If you encounter problems with your installation, or are looking for a group, do not hesitate to use the eCampus forum. There will *not* be a regular exercise class on April 5, 2023. The first exercise class will be Wednesday next week.

Exercise 1 (Install and try Python, *10 Bonus Points*)

The programming tasks in this lecture will be in Python. Doing them requires some open source software, which is available for Linux, Windows, and Macs.

- If you are new to Python, one option to install the required software is via the Python 3.10 version of the Anaconda distribution, which you can download for free from `https://www.anaconda.com/distribution/`.
- If you already have a recent version of Python, you should be able to continue using your preferred environment, which should include a Python 3 interpreter, and recent versions of packages such as NumPy, SciPy, matplotlib, pandas, and seaborn.
- Please submit all coding exercises so that we can easily run your code. E.g. use relative paths to the data (which should be contained in your submission) you are using instead of full paths. Further, use comments to help us follow your implementation.

Jupyter notebook is quite useful for running quick experiments with Python. For writing scripts, we recommend using an integrated development environment that supports Python. One possible choice is Spyder. If you decide to install Anaconda, all this software will be included by default.

Seaborn provides a simple way to plot data in Python. Start reading the seaborn tutorial at `https://seaborn.pydata.org/tutorial.html`, and playing around with this software. Use it to explore the file `solar_panels.csv`, which is available from eCampus and contains power generation data of solar panels, along with environmental data.[1]

[1] Source: `https://www.kaggle.com/datasets/vipulgote4/solar-power-generation`

Use visualizations to answer the following questions. You can obtain up to 10 bonus points by submitting your answers along with the visualizations and code:

a) To gain initial insights, plot a histogram for each column of the data.

b) Use a visualization to show how the amount of generated power changes over the year. At which day is the maximum power generated?

c) To understand more about the interaction between weather and generated power, cluster the data into two groups based on the average sky cover per day for a suitable threshold. Is there a connection between sky cover and power generation? Does this connection change over the year?
Hint: Answering this second question requires aggregating the data before visualization. A powerful Python package for these operations is pandas. Its documentation can be found at https://pandas.pydata.org/docs/user_guide/index.html

Hint: In case you require a more detailed introduction to Python itself, please refer to the Python 3 tutorial at https://docs.python.org/3.10/tutorial/.

Good Luck!