Final project part I - Generating a believable dataset

Brunno Vanelli

The final project of the course will be divided into two parts for three reasons: to practice the concepts learned so far, to evaluate the progress of the students, and to divide the work during the semester.

The result of the first part will be a dataset that will be used in the second part of the final project, and will also be graded separately.

The goal of this project is that you are able to work as a team on solving a real-world problem. Often during development, tasks can come with quite vague requirements and descriptions, even though the final result might be tangible and easy to understand. Your goal as a developer is to come up with a **methodology** to solve the problem, **create a plan**, then execute the plan, **reporting on the progress and the decisions made**.

Requirements

- **Form groups of two students** — to work independently, you should agree with your partner how each task will be divided, how you are going to collaborate with each other.
- Your deliverable will be **a dataset** and a **report**. The report will be one Jupyter notebook that will contain the dataset generation, as well as the written explanation of the decisions made on how to generate the dataset. The dataset will be a CSV file (or any other preferable format for data storage). Important is: when running your script, a dataset should be generated by the end of the file, and exported.
- Deliver both the dataset and the report as a zip file on Ilias until the submission deadline.

The problem to solve

Pretend you work for the IT Department of Albstadt-Sigmaringen University. You have now been assigned a big project for generating a sustainability report.

You have been asked by your team lead to evaluate specifically the Scope 2 emissions exclusively for the purchase of electricity. Here, you are interested in the electricity consumption by the entire university, including all electricity used by powering devices, heating water, lighting the campus, etc.

There is just one problem: **the data currently does not exist at all**!

You have discussed this within the department and agreed on the following initial plan:

- The data doesn't exist at all, so you will have to start with a simulation of what this data should look like.
- You want to get instantaneous hourly consumption, measured in kWh.
- You are interested in the projections for an entire year, so the data will be generated for the year 2024: from the 1st of January to the 31st of December.
- You are also interested in a more detailed view of the electricity consumption, for a basic reason: during the day, the electricity consumption is very different from the night. Even during the day, there are peak hours of consumption. So you have decided within the team to have data **sampled every hour**. This means that, for the year of 2024, every single hour of every single day should have one individual measurement.

Important to ask yourself:

- You are not reinventing the wheel, there are people that have already studied how consumption changes during the day — can you use their estimations?
- You might not have an idea what the total estimates are for the entire university, but a comparable university might have already published similar data — Can you use their data and scale it to your needs?

Example

I have a computer with a 65 W power supply, and I use it for work for 8 hours a day (assuming from 9am to 5pm).

I now want to generate a dataset that will display the consumption for a single day. Here is what it would look like:

```
import numpy as np
import pandas as pd
from datetime import datetime, timedelta

# There are 24 hours in a day
hours = np.arange(
    datetime(2025, 1, 1),
    datetime(2025, 1, 2),
    timedelta(hours=1)
).astype(datetime)
# Dataframe with default values
df = pd.DataFrame(index=hours, columns=['power'])
df["power"] = 0.0
# The power supply has 65 W of power: in 1h, it consumes 65 Wh or 0.065 kWh
df.loc[(df.index.hour >= 9) & (df.index.hour < 17), "power"] = 0.065
# Display results
print(df)
# Export to a CSV file
df.to_csv("consumption_data.csv")
```

Here, I made many assumptions: that the power supply is always on, that the power draw is constant, that I do not use the computer outside the hours of work, etc.

Grading

You will be graded on:

- The explanations of the decisions made on how to generate the dataset.
- If you have done multiple iterations, the thought process behind each decision: why was each decision made?
- Does your final dataset *look* correct — are there any gaps, is the data consistent?
- Is your notebook well-formatted and easy to read? Is it well-organized? Do the explanations make sense with the code?

You **will not** be graded on how accurate the dataset is: here, the goal is to build a proof of concept, not to generate a perfect dataset. In fact, your final result will differ quite a lot from reality. The goal here is to understand the process of generating a believable, parameterized dataset. If your assumptions are correct, you could always tweak the algorithm with more information to correct the accuracy later on.