Final Project Submission Guidelines

Dataset

You must use the provided dataset that you can download here: https://drive.google.com/file/d/114pmqbn2NVeaQgAzXdd2KjQZEPm_u9K5/view?usp=sharing

Each row represents a **NUTS3 region**, with socio-economic, infrastructure, and environmental variables.

Important:

- The dataset may contain **missing values** that you need to handle
- Some variables may be **irrelevant, redundant, or highly correlated**

Data cleaning and feature selection are part of the project.

Do not change the dataset file name or file extension
Do not include the dataset in your submission (.zip folder)

Delivery

Each group must submit a zip folder with **three items**:

1. Final Report (PDF)

- **Maximum length: four pages** in double column ACM format (latex or word template here) .
  References and appendix do not count toward the four-page limit.
- Appendices are allowed but:
  - will **not be graded**
  - there is **no guarantee they will be read**

The report should clearly present:

- your **research question, hypothesis, related works in the existing literature**
- your **methodology**
- the **main results and their interpretation**

The objective is to explain **what drives the location of data centers**, following the type of analysis presented in class (data + model + interpretation).

2. Notebook

You must submit **one (and only one) Jupyter notebook**.

The notebook must:

- run from start to finish
- reproduce all results shown in the report
- include the full pipeline (data loading, cleaning, modeling, evaluation)

If the notebook is not reproducible, the project will be penalized.

3. README

Include a short README explaining:

- how to run the notebook
- required libraries
- any relevant setup step

Modeling Requirements

You can frame the problem in two ways:

- **Classification**: predict whether a region has at least one data center
- **Regression**: predict the number of data centers

In all cases, you must:

- Train **at least one interpretable model using `statsmodels`**
  (to identify the main drivers of data center location)
- Train **one additional model using `sklearn`**, focused on **maximizing predictive performance**
- Use **existing scientific literature as guidance** for selecting variables, forming hypotheses, and interpreting results

You are expected to evaluate and interpret your results.

Tip: **Start with simple models using a small number of features**, and iteratively explore different feature combinations to improve performance.

Evaluation

Projects will be evaluated based on:

- clarity of the report
- correctness of the analysis
- reproducibility
- quality of interpretation