NOVA IMS INFORMATION MANAGEMENT SCHOOL

SEPTEMBER 2025

"Where Should I Live?"

Leveraging Data to Optimize Quality of Life Choices

Group Project
Programming for Data Science 2025/26

I. *"Where Should I Live?"*

In the wake of increasing mobility across Europe, fueled by remote work, international education, and shifting social and economic priorities, there has been growing demand for tools that help citizens make better-informed decisions about where to live. Recognizing this, the European Commission's Directorate-General for Employment, Social Affairs and Inclusion (DG EMPL) has launched a new initiative focused on improving access to comparative information about living conditions, opportunities, and quality of life across member states.

To support this initiative, a data science task force was assembled, comprised of researchers and developers across Europe. You were selected as part of this effort to design a tool that translates complex country-level data into clear, actionable insights for everyday people: students seeking affordable education, professionals looking for better work-life balance, retirees interested in safety and healthcare, and more.

Your project, "Where Should I Live", aims to simplify this decision-making process by analyzing key characteristics of European countries, such as cost of living, safety, and employment. Through exploratory analysis, visualizations, and interpretable insights, your goal is to build a framework that empowers users to evaluate which countries align best with their personal values and life goals.

II. PROJECT OBJECTIVES

Working in data science goes beyond technical expertise; it demands strong critical thinking and problem-solving skills.

The goals for this project are three-fold:

Data Wrangling and Analysis: In this phase, raw data is transformed into clean, structured datasets. You'll apply fundamental techniques to explore the data and extract relevant insights about European countries.

Working with Advanced Topics: Here, you go beyond the basics by integrating advanced concepts such as web scraping, geographical data, or other relevant data types. This phase reflects the real-world challenges of gathering and handling diverse data.

Data Science in Action: Finally, you'll synthesize your work into a coherent analysis. Using techniques of your choice, you'll explore patterns, visualize relationships, and generate actionable frameworks and conclusions. This phase is **open-ended**, allowing for creativity, critical thinking, and deeper exploration.

Data Wrangling and Analysis

Your first task is to **import, comprehend, and preprocess** the dataset. This will then enable you to identify key aspects of the cities under study and create compelling visualizations. Moreover, you should be able to answer the following questions:

1. How did you handle missing values and duplicate records in the dataset? Justify your approach.

2. a) Which *country* appears most frequently in the dataset? How many *cities* are associated with it?
b) How many *cities* are present in total? How many are associated with *Greece*?
c) Which is the *least spoken language* in the dataset? Which are the *top 3* most spoken languages?

3. a) Entries uploaded before *April 2023* need to be updated. Which cities would require an update?
b) How many *days* ago was the last update? On what *day*, *month*, and *year* did it occur?

4. a) How are the *Unemployment Rate* and *GDP per Capita* distributed and related? What does this relationship suggest?
Provide a **visual representation**.

b) Which are the *top 5 cities* with the largest difference between the *Average Monthly Salary* and *Average Cost of Living*?
What about the *top 5 countries* with the smallest *average* difference?
Show these results with **meaningful visualizations**.

c) Which is the *best city* for someone seeking:
- an *average monthly salary* above €1600,
- a *cost of living* below €900, and
- a country suitable for starting a family (with a relatively *larger youth population*)?

5. What are **three additional insights** you find meaningful when comparing the given cities?

Advanced Topic - Building an Interactive Map

The goal of this section is to create an **interactive map of Europe** where users can explore cities and view relevant information such as country, population, average salary, and cost of living. This task is divided into two main parts:

1. Web Scraping

- Starting from the Wikipedia Main Page (*https://en.wikipedia.org/wiki/Main_Page*), extract the geographical coordinates of each city.
- The coordinates must be obtained only through web scraping of this source
- Ensure that each city is matched with its corresponding dataset entry accurately.

2. Interactive Map

- Using the scraped coordinates, build an interactive map of Europe that highlights all cities.
- Each city marker should display the following when hovered/clicked:
  - Country
  - Population
  - Average monthly salary
  - Average cost of living
- You may use any tools covered in class (e.g., plotly, geopandas, etc).
- A baseline example is shown below. You are free to personalize the design (colors, popups, etc.), as long as the required information is included.

Fig. 1. Example of the interactive map (zoomed view of Portugal with city markers).

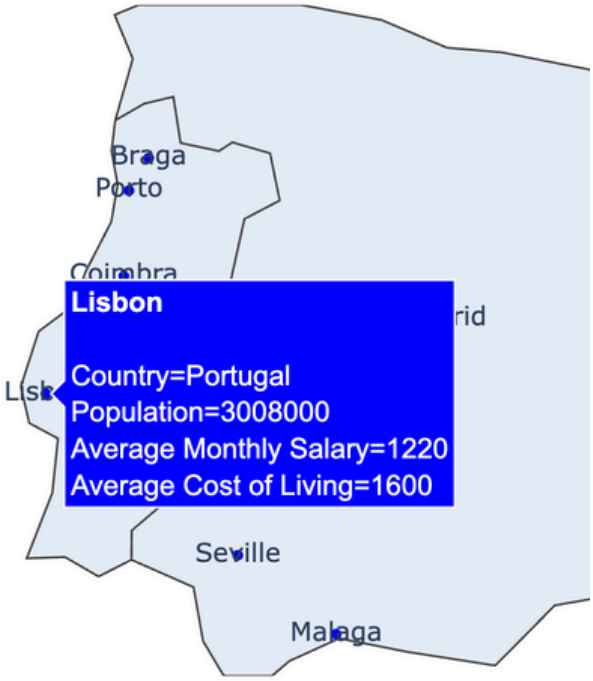

Fig. 2. Example of city hover text (Lisbon).

Data Science In Action

In this final phase, you will bring everything together by integrating your pre-processed dataset and transforming it into **meaningful insights or tools**. This is your opportunity to be creative: choose techniques you find most appropriate and develop something that would **genuinely help people compare European cities and decide where to live.**

Your work will be evaluated not only on technical execution but also on the **relevance, originality, and clarity** of your analysis or tool. Higher levels of complexity, critical thinking, and real-world applicability will be rewarded.

You are **free to choose your own direction**, but here are some possible starting points:

- Recommendation System: Allow users to specify the characteristics they care about (e.g., cost of living, salary, safety, education opportunities) and suggest the cities that best match their profile.
- Interactive Dashboard: Combine your dataset with additional external sources (e.g., education, healthcare, transportation) and design a dashboard that allows users to visually explore and compare cities.
- Comparative Analysis: Perform a deep dive into one or more dimensions (such as cost of living vs. salary, unemployment vs. GDP, population demographics, etc.) and present clear visualizations and interpretations that highlight meaningful trade-offs.

III. PROJECT DATA

The file ***city_data.csv*** contains information about European cities. Each row corresponds to a city, and each column provides a specific characteristic.

Variable	Description
City	City name and respective Country
Population Density	Population density (inhabitants per km2)
Population	Total Population of city municipality.
Working Age Population	Working age population group (15-64 years)

Youth Dependency Ratio	Youth dependency ratio (-15 over population 15-64)
Unemployment Rate	Unemployment rate (unemployed population 15-64 over labour force 15-64)
GDP per Capita	GDP per Capita in USD ($)
Days of very strong heat stress	Days of very strong heat stress (UTCI > 38°C)
Main Spoken Languages	Main spoken languages in the city
Average Monthly Salary	Estimated average monthly net salary in euros (€)
Average Rent Price	Estimated average rent price for a 1 bedroom apartment at the city center, in euros (€)
Average Cost of Living	Estimated average cost of living for 1 person living in the city (including rent).
Last Data Update	Date the data for the city was last updated.

IV. DELIVERABLES

- A **.ipynb notebook** (or zip of multiple notebooks) featuring all the code you used throughout the project to:
  a. Decide on your final solutions for the problem at hand.
  b. Obtain your final results (code that helped you make decisions, but does not directly contribute to reaching the final solution, should be included but commented).
  c. Any additional datasets required to run the notebook should be included in a zip file. (File naming format: **GroupXX_PDS_2526.ipynb**).

V. EVALUATION

Your final project will be graded out of **20v**, according to the following criteria:

Criteria	Description	Max Grade (out of 20)
Data Wrangling and Analytics	Quality of data preprocessing and exploration. Correct answers to the assigned questions, as well as the ability to extract additional meaningful insights.	7v
Advanced topic	Mastery of the advanced concepts introduced in the course. Successful completion of the given tasks. Originality and creativity of the final solution will also be considered.	6v
Data Science in Action	Pertinence and effectiveness of your chosen approach. Originality, depth, and complexity of your analysis or tool will be recognized in the grade.	5v
Notebook Quality	Code organization, clarity, commenting, efficiency, and overall readability of your notebook.	2v

- The notebook must begin with a Markdown cell listing the names and student numbers of all group members.
- Groups that include in the notebook a **working, public GitHub repository link** used during project development will receive +**0.5v**.
- Students can submit all deliverables with a **maximum delay of 3 days**, incurring a penalty of **1v per day**. Beyond these three days, submissions will not be accepted.

Project Deadline: December 21st, 18h00

Your grade will reflect our assessment of the quality, correctness, clarity, and efficiency of your work. Below you can find a **general description** of what is expected in each evaluation component:

1. Notebook Quality (2 points) A strong notebook should be clear, well-structured, and easy to follow, even for someone reading it for the first time. We expect:

- Organization: logical sequencing of steps, clean formatting, and consistent structure.
- Clarity: appropriate use of comments and Markdown explanations to show that you understand what is being done at each stage.
- Readability: code that is concise, efficient, and avoids unnecessary repetition.

Your notebook should make it clear what you did at each step, why you chose that approach, and what you learned from it.

2. Data Wrangling & Analysis (7 points) This component evaluates your ability to clean, prepare, and understand the dataset. We expect you to:

- Preprocess data correctly: handle missing values, duplicates, inconsistencies, and justify your approach.
- Answer the assigned questions fully: provide clear reasoning behind each answer, not just the result.
- Extract meaningful insights: go beyond surface-level observations and identify patterns that are genuinely useful for understanding European cities.

3. Advanced Topics (6 points): This section assesses your mastery of the advanced concepts introduced in the course and your ability to apply them to real-world data. Specifically, we will evaluate:

- Correctness: whether you completed the assigned advanced tasks (e.g., web scraping, building an interactive map).
- Understanding: whether your approach shows that you grasp the underlying concepts rather than simply applying code mechanically.
- Originality: whether you added creative elements, personalization, or improvements beyond the baseline requirements.

4. Data Science in Action (5 points): The final component evaluates how you synthesize your work into a meaningful outcome. We will consider:

- Pertinence: does your tool, analysis, or insight genuinely help someone understand European cities or make better living decisions?
- Complexity and ambition: how challenging or non-trivial was your approach? Did you aim for an advanced, well-thought-out solution or just stick to the basics?
- Strategy and execution: did you use an appropriate methodology and structure your analysis logically?
- Initiative: did you bring in additional data, tools, or techniques beyond what was provided in class, or did you rely solely on the starter dataset and instructions?

VI. FINAL NOTES

1. Make sure to have a notebook that is understandable to someone reading it for the first time. A good structure, with appropriate comments showing that you understand what is being done at every block of code, goes a long way into that.
2. The trustworthiness of the information you provide is key. If you look for information outside the materials we provided, you should cite the source of the materials appropriately.
3. Before submitting, run your notebook from the start one last time, please comment unnecessarily lengthy cells that take too much time to run.
4. All the code you used that is unneeded to highlight the points you want to convey should be part of your submitted notebook(s), but it should be commented.
5. We will run your Jupyter Notebooks. So, please make sure we can run the notebook from start to finish in one go. Notebooks that do not fulfil this condition will be penalized.
6. **The notebook code will pass through a process of plagiarism and AI generation checking.**
7. **To avoid situations where we have conflicting versions, please make sure that you show, in the notebook, the version of the package you are using for each package you use.**
8. When determining the grade for your work, there will be a comparative component between your work and the works presented by your peers.

Attendance at the defense is mandatory for approval of the project. The discussion has a group component and an individual component. Considering this, a student's final grade can change during the defense depending on their performance, **without any limitations.**

The date for the defense will be selected during the semester.