PGAD: Programación para el Análisis de Datos
Final Project

October 08, 2025

Title: *Urban Data Analytics: Exploring the Relationship Between Mobility, Environment, and Quality of Life*

Instructions:

- The project can be developed in pairs of students. It is your responsibility to find a co-worker.
- The project has three parts: Written report: 45%, Computer Work: 35%, and Final Presentation: 20%.
- The maximum project delivery date is **December 05, 23:59 Hrs**, in Moodle.
- December 05 (10am): Final Course Exam (Practice and Theory) (To be confirmed).
- December 12 (10am): Final Presentation of the project during the class session.

Outcomes:

By the end of the project, students should demonstrate the ability to:

- Integrate and analyze urban or environmental data using Python, R, and SQL.
- Apply sound preprocessing and analysis techniques to extract insights.
- Communicate analytical results clearly and effectively using visual and written formats.

Objectives:

- Formulate a clear and relevant analytical question related to urban life and sustainability.
- Design and populate a relational database using `SQL`.
- Apply data cleaning and transformation procedures to prepare the dataset for analysis.
- Perform exploratory and statistical analyses to discover patterns and relationships.
- Develop a final report integrating code, visualizations, and interpretation of findings.

Description:

This project motivates to explore how patterns of mobility, environmental conditions, and urban characteristics relate to the quality of life in one or more cities.

The project integrates the use of `Python`, `R`, and `SQL` to manage the complete data analysis pipeline: from data collection and preprocessing to exploratory analysis, statistical interpretation, and visualization.

The project emphasizes the following skills:

- Integration of heterogeneous data sources.
- Application of preprocessing techniques (cleaning, imputation, normalization, encoding).
- Use of `SQL` for data management and querying.
- Exploratory Data Analysis (EDA) using `Python` or `R`.
- Communication of results through reproducible reports and visualizations.

Possible Data Sources:

- Bogotá Open Data Portal — mobility, environment, public transportation, traffic accidents.
- World Air Quality Index — air pollution and air quality indicators.
- OpenWeatherMap API — meteorological and temperature data.
- World Bank Open Data — urban, environmental, and sustainability indicators.
- Kaggle Datasets — city and environment-related datasets.

Project Questions:

- Is there a relationship between air pollution levels and the use of public transportation?
- What urban factors best predict citizen well-being or satisfaction?
- How do weather conditions correlate with mobility patterns or accident frequency?
- Which city areas show higher environmental or transport inequality?

Stages and Deliverables:

No.	Stage	Description	Deliverable
0	Data Selection	Choose a theme and dataset aligned with project objectives.	One-page proposal.
1	Integration and SQL	Design and populate a SQL database. Use Python or R to query data.	SQL script + connection code.
2	Data Cleaning	Apply preprocessing (missing values, outliers, normalization).	Python notebook.
3	Exploratory Analysis	Perform descriptive statistics and visual exploration.	R or Python notebook.
4	Inferential or Predictive Analysis	Apply regression, correlation, or basic ML models.	Technical report.
5	Communication	Build a dashboard or reproducible report.	Final presentation.

Evaluation Criteria:

Criterion	Weight
Relevance and clarity of research question	15%
Correct use of Python, R, and SQL	25%
Quality of preprocessing and data cleaning	20%
Depth of exploratory and visual analysis	20%
Reproducibility and clarity of final report/presentation	20%

Final Remarks

This project represents the culmination of the course *Programming for Data Analysis*, integrating theoretical knowledge and practical skills in data management, programming, and analytical reasoning.

Each project should reflect a comprehensive understanding of the data analysis pipeline, from raw data to meaningful interpretation, demonstrating the student's ability to think critically, program efficiently, and reason analytically in the context of data science.

Report Template Requirement

The written report must follow the structure and style described in the document *“Project Writeup – Introduction to Machine Learning, MIT”*, provided as a reference template for this course.

Each one may choose to write the report either in **English** or **Spanish**, but the organization and content must adhere to the sections and guidelines outlined in the template.