Universidad Externado

Bayesian Hotel Pricing Lab

Winter School Montevideo

Montevideo, Uruguay

CIENCIA DE DATOS

Objective

This lab aims to build **Bayesian models** to understand hotel price dynamics across Europe. Students will explore whether prices follow a global trend or are influenced by city-specific or country-specific effects. The main goal is to detect **global vs. local pricing patterns** and learn how **hierarchical models** can uncover hidden structures in real-world data.

Dataset

You will work with two datasets:

- `hotelseuropefeatures.csv`: general hotel info: `hotel_id`, `city`, `country`, `stars`, `rating`, `neighbourhood`, `accommodation_type`, etc.
- `hotelseuropeprice.csv` : pricing and availability: `price`, `offer`, `weekend`, `holiday`, `nnights`, `scarce_room`, etc.

These datasets must be merged using the `hotel_id` column.

Project Structure

1. Problem Definition

- Formulate a research question: *Do prices follow a global trend or do they differ by city/country?*

- Choose your grouping structure: **City** vs. **Country**
- Optional: restrict your dataset to one specific country if you prefer a more focused analysis.

2. Data Preparation

- Merge datasets
- Handle missing values
- Select and transform predictors (categorical/numerical)
- Choose target variable: `price`

3. Exploratory Analysis

Limit to a maximum of **three informative plots**, such as:

- Boxplots of `price` per city or country
- Correlations between numerical variables
- Scatterplots between predictors and target

4. Modeling Phase (Flexibility Encouraged)

You are **free to define the level of complexity** of your model based on your learning goals and data quality. Choose **at least two** of the following approaches:

a. Global Model

- Bayesian linear regression using all data
- No group structure
- Example:

$$price_i \sim \mathcal{N}(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots, \sigma)$$

b. Local Models by City or Country

- Separate models per city or country
- Analyze how posterior distributions vary across locations

c. Hierarchical Model

- Multilevel intercept or slope model by `city` or `country`
- Example:

$$\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha) price_i \sim \mathcal{N}(\alpha_{j[i]} + \beta x_i, \sigma)$$

5. Posterior Analysis & Visualization

- Plot posterior distributions (with `arviz`)
- Show 90% HDIs for parameters
- Boxplots of intercepts per group (city/country)
- Visual comparison of global vs. local parameters

6. Discussion

- What model structure explains price best?
- Are there major differences in price dynamics across cities?
- Does a global model overlook important group-specific effects?
- How sensitive are your results to prior assumptions?

Model Interpretation Guide

Model Type	**Description**	**Expected Insights**
Global Model	Fits a single linear regression across all observations without grouping.	Identifies general price drivers; may overlook local variations.
Local Models	Fits a separate model for each city or country independently.	Captures heterogeneity but lacks pooling; useful when city behavior is distinct.
Hierarchical Model	Models parameters (e.g., intercepts) as varying by city but sharing a common global prior.	Combines global and local structure; shrinks estimates in cities with few data points.

Table 1: Modeling options to explain hotel price dynamics.

How to Compare and Select Models

- **Posterior predictive checks:** Are predicted values consistent with observed data?
- **WAIC/LOO:** Use information criteria for model comparison.
- **Interpretability:** Can you explain results clearly?
- **Uncertainty:** Look at HDIs and trace plots for stable estimates.
- **Sample size per group:** Prefer hierarchical models when some groups are small.

Technical Requirements

- Python 3.10+
- Libraries: `pymc`, `arviz`, `pandas`, `matplotlib`, `seaborn`
- Use MCMC sampling via `pm.sample`
- Ensure good diagnostics: $\hat{R} < 1.01$, high ESS, trace plots

Deliverables

Each group must submit:

- A well-documented `.ipynb` notebook or `.py` script
- Clean visualizations
- Model interpretation
- Code reproducibility

Tip

You are **not required** to build the most complex model **build what you can justify** and what best answers your question. Quality of reasoning and clarity of interpretation matter more than complexity.

Presentation Guidelines

Each group will present their work in a **10-minute oral presentation**. The presentation should include:

Structure:

1. Problem motivation and research question
2. Dataset description and cleaning
3. Exploratory visualizations (max 3)
4. Modeling strategy and results
5. Interpretation of posteriors
6. Conclusions and reflections

Table 2: Variables in `hotel_features` dataset

Variable	Description
`hotel_id`	Unique identifier for each hotel (used for merging with price data).
`city`	City name as recorded in the listing. May differ from the verified location in `city_actual`.
`distance`	Distance from the hotel to the primary city center (`center1label`).
`stars`	Hotel classification in stars (typically 1–5).
`rating`	User rating score (0–10) from the booking platform.
`country`	Country where the hotel is located.
`city_actual`	Verified city location based on coordinates or official address.
`rating_reviewcount`	Number of reviews used to compute `rating`.
`center1label`	Name or label of the first city center point used for `distance`.
`center2label`	Name or label of the second city center point used for `distance_alter`.
`neighbourhood`	Neighborhood or district where the hotel is located.
`ratingta`	TripAdvisor rating score for the hotel.
`ratingta_count`	Number of TripAdvisor reviews used for `ratingta`.
`distance_alter`	Distance from the hotel to the alternative city center (`center2label`).
`accommodation_type`	Type of lodging (e.g., Hotel, Apartment, Hostel, B&B).

Table 3: Variables in `hotel_prices` dataset

Variable	Description
`hotel_id`	Unique identifier for the hotel (primary key to merge with `hotel_features`).
`price`	Price for the stay (in EUR) for the given date and number of nights.
`offer`	Binary indicator (0/1) for whether a special offer was listed for that hotel.
`offer_cat`	Categorical version of the offer type or discount level.
`year`	Year when the price was recorded.
`month`	Month when the price was recorded (1–12).
`weekend`	Binary indicator (0 = weekday, 1 = weekend) for the check-in date.
`holiday`	Binary indicator for whether the date falls on a public holiday in the country.
`nnights`	Number of nights for the booking.
`scarce_room`	Binary indicator (0/1) showing if the booking site flagged the room as “scarce” (e.g., “Only 2 rooms left!”).