**🏨 Forecasting and Explaining Hotel Demand Recovery in Europe (2015–2025)**

**📖 Project Overview**

This project analyzes and forecasts hotel demand across European regions before, during, and after the COVID-19 pandemic (2015–2025).  
Using open data from **Eurostat**, combined with **policy**, **mobility**, and **air-traffic** indicators, we apply **interpretable machine learning models** (Ridge regression, XGBoost) to:

* Forecast monthly hotel nights at regional level
* Quantify the impact of COVID-19 and subsequent recovery trends
* Identify key drivers of post-pandemic tourism rebound

This project builds on prior work in *Forecasting Migration Flows with Machine Learning*, extending the same analytical approach to the **hospitality and tourism sector**, where professional experience provides domain insight.

**🧭 Research Questions**

1. How accurately can we forecast hotel demand using pre-pandemic patterns?
2. Which factors best explain differences in recovery intensity across countries and regions?
3. How heterogeneous was the rebound across Europe (2020–2025)?

**🗂️ Repository Structure**

hotel-demand-recovery/

├─ src/

│ ├─ hotel/

│ │ ├─ eurostat\_download.py # download Eurostat hotel data

│ │ ├─ oxcgrt\_download.py # download COVID policy stringency data

│ │ ├─ mobility\_download.py # download Google mobility data

│ │ └─ merge\_datasets.py # merge raw datasets into a clean panel

│ └─ utils/

│ ├─ io\_utils.py # load/save CSVs and pickles

│ ├─ feature\_utils.py # lag and growth-rate helpers

│ ├─ metrics\_utils.py # evaluation metrics (MAE, sMAPE, RMSE)

│ └─ plot\_utils.py # reusable plotting functions

│

├─ data/

│ ├─ raw/ # unprocessed data downloaded from sources

│ ├─ processed/ # merged & cleaned datasets ready for analysis

│ └─ external/ # optional reference data (NUTS boundaries, etc.)

│

├─ notebooks/

│ ├─ 01\_data\_exploration.ipynb

│ ├─ 02\_feature\_engineering.ipynb

│ ├─ 03\_baseline\_and\_xgboost.ipynb

│ └─ 04\_evaluation\_and\_visualization.ipynb

│

├─ outputs/

│ ├─ models/ # trained .pkl models

│ ├─ figures/ # generated plots & maps

│ └─ reports/ # summary tables or short analyses

│

├─ requirements.txt

└─ README.md

**🚀 Quickstart**

**1️⃣ Installation**

git clone https://github.com/<your-username>/hotel-demand-recovery.git

cd hotel-demand-recovery

pip install -r requirements.txt

**2️⃣ Data Download and Preparation**

Run the data pipeline scripts:

python src/hotel/eurostat\_download.py

python src/hotel/oxcgrt\_download.py

python src/hotel/mobility\_download.py

python src/hotel/merge\_datasets.py

The cleaned and merged dataset will be stored in  
data/processed/hotels\_merged\_panel.csv (or .parquet).

**Sources:**

* Eurostat “Tourist accommodation – nights spent at hotels and similar establishments”
* [Oxford COVID-19 Government Response Tracker (OxCGRT)](https://github.com/OxCGRT/covid-policy-tracker?utm_source=chatgpt.com)
* Google Community Mobility Reports
* EUROCONTROL Aviation Statistics

**🧱 Analysis Workflow**

After the data are ready, analysis continues in notebooks:

| **Notebook** | **Purpose** |
| --- | --- |
| **01\_data\_exploration.ipynb** | Inspect and visualize trends, missing values, and variable distributions. |
| **02\_feature\_engineering.ipynb** | Create lag features, growth rates, and seasonal dummies; export processed feature dataset. |
| **03\_baseline\_and\_xgboost.ipynb** | Train Ridge regression and XGBoost models, compare forecasting accuracy, and save trained models. |
| **04\_evaluation\_and\_visualization.ipynb** | Load saved models, compute metrics, generate SHAP analyses, and produce recovery maps and figures. |

All intermediate datasets and models are stored in data/processed/ and outputs/models/.

**📊 Outputs**

* Forecast performance metrics (MAE, RMSE, sMAPE)
* SHAP feature importance explaining recovery drivers
* Choropleth maps of regional recovery (2020–2025)
* Interactive or static plots comparing actual vs. predicted demand

**🧠 Skills Demonstrated**

* End-to-end data pipeline design (raw → processed → model)
* Time-series and panel-data forecasting with conventional ML tools
* Feature engineering and lag creation for temporal models
* Model interpretability (SHAP, feature importance)
* Reproducible, well-structured codebase organization

**🗓️ Suggested One-Week Timeline**

| **Day** | **Focus** | **Outcome** |
| --- | --- | --- |
| 1 | Data download and cleaning | Raw → processed dataset |
| 2 | Exploratory analysis | Trends, missing values |
| 3 | Feature engineering | Final feature table |
| 4 | Baseline models | Linear / Ridge results |
| 5 | XGBoost model | Tuned model + metrics |
| 6 | SHAP + visualization | Feature importance, maps |
| 7 | Wrap-up | README, plots, GitHub push |

**🧩 Next Steps**

* Add country-level deep dives (Spain, France, Germany).
* Integrate Airbnb data for hotel-vs-STR substitution analysis.
* Apply causal ML to quantify the impact of COVID-policy stringency on demand recovery.

**📬 Citation**

If you use or adapt this work, please cite:

Author: [Your Name]  
Project: *Forecasting and Explaining Hotel Demand Recovery in Europe (2015–2025)*  
Year: 2025, GitHub Repository.