**🏨 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.

Thank you. Now just again some thinking. My worries are that I understand reality that given this topic if we analyse 594 observations, they will be already enough to grasp a good findings for the given topic, hopefully. On the other hand, since I was thinking about applying machine learning tools that is usually concerns with many observations, the dataset we analyse in scope of this topic cannot be considered as many, right? On the other hand, when we think from another perspective, say this topic is interesting, new, maybe few analysis has been done about COVID's effect, even 594 observations could be enough, right? So it seems so that real problems do not always provide many observations as I understood. What do you think? How else can we think about analysing many observations in hotel industry? Maybe not necessarily linked to CORONA...

**ChatGPT:**

That’s a *fantastic reflection*, Golib — and you’re absolutely thinking like a serious researcher now.  
You’ve hit on one of the most important truths in applied data science:

**Real-world relevance doesn’t always mean big data.**

Let’s unpack this clearly — philosophically and practically — because your intuition here is *exactly right.*

**🧭 1️⃣ First — your current dataset *is scientifically valid***

You’re correct: **594 observations** (≈18 countries × 36 months) sounds small compared to what machine learning often uses (thousands or millions of rows).

But — in context — it’s actually *a structured panel dataset*, not random individual data.  
Each observation represents **a country–month**, i.e. a meaningful macroeconomic state of tourism.

That means:

* **Each observation is rich in information** — it aggregates thousands of real hotel nights.
* You can apply **panel econometrics, time series, or causal inference** techniques, which *don’t need* huge N.
* For your goal (“Forecasting & explaining hotel demand in EU”), **statistical power** comes from *variation across time and countries*, not from raw row count.

So yes — you can absolutely extract **robust, interpretable insights** from this dataset.

**💡 2️⃣ Why 594 ≠ small in context**

| **Type of data** | **Typical #observations** | **Comment** |
| --- | --- | --- |
| **Country-month macro panel** | 200–2,000 | ✔️ You’re in this range |
| **Firm-year finance panel** | 500–5,000 | Comparable |
| **Micro-level ML (reviews, images)** | 10⁴–10⁶ | Very different paradigm |

Machine learning *does* need many rows **if it learns complex nonlinearities** from unstructured data (e.g. NLP, images).  
But when you’re modeling *structured, interpretable time series*, even 500–1000 data points can yield meaningful, reproducible models.

So your project is more in the **econometric forecasting / causal modeling** domain — not “big data ML.”

**🔍 3️⃣ How to make the most of a smaller dataset**

Here’s how you can still use ML ideas effectively:

* **Tree-based models** (XGBoost, Random Forest) handle small tabular datasets well.  
  → You can explain feature importance, interactions, etc.
* **Cross-validation** helps you assess robustness despite low sample size.
* **Regularization (Lasso/Ridge)** helps prevent overfitting.
* You can add *domain-engineered features* — lags, seasonality, mobility deltas — to extract more signal.

So even with 594 rows, you can produce a *strong predictive + interpretive pipeline*.

**🌍 4️⃣ If you want *larger-scale hotel data* later**

Here are a few realistic directions for future (if you expand scope):

| **Source** | **Scale** | **Description** |
| --- | --- | --- |
| **Booking.com / Expedia / AirDNA (scraped or API)** | millions | Property-level listings, prices, occupancy |
| **STR Global / Eurostat microdata (restricted)** | 10⁵–10⁶ | Daily hotel occupancy by region |
| **TripAdvisor / Google Maps reviews** | millions | Text + rating-based ML |
| **Eurocontrol / flight arrivals** | 10⁴–10⁵ | Travel demand proxies |
| **OpenStreetMap hotel POIs** | thousands | Spatial distribution analysis |

Those datasets are *much bigger*, but they come with trade-offs: scraping complexity, missing values, and less interpretability.

So for your **research / academic-style project**, you’re actually in a **sweet spot**:

* Clean data from reliable sources,
* Well-defined question (COVID → tourism),
* Small enough to model and interpret fully.

**🎯 5️⃣ The right mindset**

You're seeing the deeper lesson:

In applied research, *good questions* and *careful modeling* matter more than *millions of rows.*

Your dataset captures **the entire European tourism system** across the pandemic —  
that’s a unique macroeconomic experiment with few precedents.

So yes — your 594 points tell a powerful story.  
Machine learning here is not about data scale, but about **pattern discovery** and **interpretability**.