Excellent — you’re thinking exactly like a data scientist preparing a research-grade econometric study. Let’s step back and set up a **clear, structured research plan** before coding further.

**🎯 1. Research goal**

**Main question:** How has the European hotel industry recovered after COVID-19, and have countries returned to their pre-pandemic equilibrium?

You already defined the **target variable**:

* nights\_spent = monthly hotel nights per country (proxy for tourism demand).

**🧩 2. Research sub-questions**

1. **Recovery path**
   * How fast did hotel activity rebound in each country?
   * Have levels returned to 2019 values?
2. **Determinants of recovery**
   * How do macroeconomic conditions (GDP, unemployment, turnover) relate to recovery?
   * What was the role of COVID incidence and policy restrictions?
3. **Structural change**
   * Did the relationship between hotel nights and macro drivers change post-COVID?
   * Are there persistent shifts (new equilibrium) or full mean reversion?

**🧮 3. Conceptual framework**

| **Category** | **Variables** | **Notes** |
| --- | --- | --- |
| **Tourism activity** | nights\_spent | Dependent variable |
| **Economic environment** | gdp, unemployment\_rate, turnover\_index | Demand-side conditions |
| **Health shocks** | covid\_cases | Pandemic intensity |
| **Temporal factors** | month, year, period | Seasonality & structural breaks |
| **Lags** | gdp\_lag1-3, covid\_cases\_lag1-3, … | Capture delayed impacts |
| **Potential additions** | HICP, exchange rate, air passengers, policy stringency | Optional enrichment |

**📊 4. Additional indicators (optional)**

You already have a **strong baseline**, but if you want to deepen interpretability:

| **Indicator** | **Why useful** | **Source** |
| --- | --- | --- |
| **HICP (inflation)** | Captures real income pressure on tourism demand; may explain slower recovery in high-inflation periods | Eurostat (monthly) |
| **Exchange rates (vs USD/EUR)** | Affects foreign tourism competitiveness | ECB / Eurostat |
| **Air passengers** | Leading indicator of tourism mobility (you already mentioned it in the pipeline) | Eurostat |
| **Policy stringency index (Oxford COVID)** | Explains differences in 2020–2021 restrictions | Our World in Data |

**Verdict:**  
👉 *HICP and air passengers* are the most meaningful additions. You could add them easily through your Eurostat pipeline.

**⏱️ 5. Lags — how to use them properly**

Lags help model the **temporal transmission** of shocks:

* **Economic lags (GDP, turnover):** hotels respond with 1–3 month delay to macro conditions.
* **COVID cases:** immediate and lagged negative impact (restrictions, fear).
* **Unemployment:** slower-moving, so longer lags (2–6 months) may capture it better.

✅ **Best practice:**

1. Test lags up to 3 months in EDA (correlations, partial autocorrelation).
2. In modeling:
   * For panel regression, include 1–2 significant lags only to avoid multicollinearity.
   * For ML models, you can include all and let regularization handle selection.

Example for panel regression:

hotel\_it = β0 + β1\*gdp\_it + β2\*gdp\_it-1 + β3\*covid\_it-1 + β4\*unemployment\_it + α\_i + δ\_t + ε\_it

where α\_i = country fixed effects, δ\_t = month or year dummies.

**⚙️ 6. Common modeling approaches**

| **Goal** | **Model** | **Description** |
| --- | --- | --- |
| Descriptive / econometric | **Fixed-effects panel regression (FE/RE)** | Quantify determinants & recovery drivers |
| Predictive baseline | **SARIMA / ARIMAX per country** | Captures local temporal structure |
| Multivariate forecasting | **Dynamic panel (Arellano-Bond), VAR** | Studies lagged interdependence |
| Machine learning | **XGBoost, Random Forest, LSTM, TFT** | Predictive accuracy across all countries |
| Structural recovery | **Difference-in-Differences / event study** | Pre- vs. post-COVID treatment effect |

For your stated goals (recovery analysis & country comparison), a **panel regression with period interactions** is the most interpretable and widely used in academic/Eurostat-style research.

**📅 7. Proposed research plan (workflow)**

**Phase 1 – EDA & Descriptive Analysis**

* Confirm coverage and seasonality.
* Compute *recovery ratios*:
* Visualize recovery trajectories by country and region group.
* Correlate recovery with macro variables.

**Phase 2 – Econometric Modeling**

* Estimate panel regression:
* Include period interactions to see if relationships changed post-COVID:
* df['covid\_effect'] = df['covid\_cases'] \* (df['period'] == 'COVID')
* Assess fixed vs random effects (Hausman test).

**Phase 3 – Forecasting / Scenario Analysis**

* Fit forecasting models (ARIMA or XGBoost) using lagged predictors.
* Project 2025–2026 values under different macro assumptions.

**✅ 8. Next actions**

1. Add optional **HICP and air passengers** if accessible.
2. Finalize EDA notebook — I’ll generate that template next.
3. Then build a **panel regression baseline** before moving to forecasting.

Would you like me to:

* (A) **Generate the EDA notebook now** (plots + preprocessing based on current dataset),  
  or
* (B) **Show how to extend your pipeline** to include *HICP and air passengers* first?