# **Hotel Booking Cancellations**

## **Project Overview**

This project focuses on \*\*data cleaning and preprocessing\*\* for a hotel booking dataset.

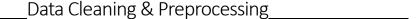
The ultimate goal is to prepare the dataset for building a \*\*cancellation prediction model\*\*.

Since last-minute cancellations significantly impact hotel profitability, a clean and reliable dataset is critical.

#### **Business Problem**

The hotel revenue team identified that \*\*last-minute cancellations\*\* severely reduce profits.

This project builds the \*\*data foundation\*\* to allow predictive models to forecast cancellations and help hotels improve planning. ---



## Phase 1: Exploratory Data Analysis (EDA) & Data Quality Report

- Loaded data and generated summary statistics using `.describe()` and `.info()`.
- Identified missing values and visualized them using \*\*missingno\*\* and \*\*heatmaps\*\*.
- Detected outliers in numerical columns ('adr', 'lead\_time') using boxplots and IQR method.

\_\_\_

#### **Phase 2: Data Cleaning**

- \*\*Missing Values\*\*
- `company`  $\rightarrow$  drop these column because it have many none value and will not be usfule in the prediction

and 'agent' → replaced with '0'

- `country` → filled with unknown as it have few non values
- `children` → filled using mode .
- \*\*Duplicates\*\*
- -There was 31994 rows
- Removed exact duplicate rows.

- \*\*Outliers\*\*
- Capped `adr` at 1000 (`df['adr'] = df['adr'].clip(upper=1000)`).
- \*\*Fix Data Types\*\*
- Converted (arrival\_date\_year, arrival\_date\_month, arrival\_date\_of\_month)columns into a single column `arrival\_date` in datetime format.
- Converted `children` and `babies` columns into integers (`Int64`) with null support.
- . \*\*Potential Data Leakage\*\*
- Dropped `reservation\_status` and `reservation\_status\_date` (contain info only available after booking).

## **Phase 3: Feature Engineering & Preprocessing**

- \*\*New Features\*\*
- -Make new colomn 'total\_guests' by summing the number of the children, adult, babies columns
- Make new colomn 'total-nights by summing the number of the stay in week night + stay in weekend night
- Make new colomn 'is family'  $\rightarrow$  A binary flag (Yes/No) indicating if the booking includes children or babies.
- \*\*Encoding Categorical Variables\*\*
- One-Hot Encoding for low-cardinality columns ('meal', 'market', 'segment').
- Frequency Encoding for high-cardinality ('country').
- \*\*Final Preparation\*\*
- Train-test split: `train\_test\_split (df, test\_size=0.2, random\_state=42)`

---