🏨 Hotel Booking Cancellations - Data Cleaning & Preprocessing

📌 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.

🔎 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` and `agent` → replaced with `0` or `"None"`.

  - `country` → filled with mode `"PRT"` or encoded using frequency encoding (rare countries grouped as `"Other"`).

  - `children` → filled using median.

- **Duplicates**

  - Removed exact duplicate rows.

- **Outliers**
  - Capped `adr` at 1000 (`df['adr'] = df['adr'].clip(upper=1000)`).

- **Fix Data Types**
  - Converted arrival date columns into a single column `arrival_date` in datetime format.
  - Converted `children` and `babies` columns into integers (`Int64`) with null support.

⚠ Data Quality Issues

1. **Missing Values**
   - `country` had missing values → imputed with "Unknown" or frequency encoding.
   - `children` column had NaNs → filled with median.

2. **Duplicates**
   - Found duplicate rows → dropped.

3. **Outliers**
   - Extreme values in `adr` (above 5000) capped at 1000.
   - Checked `lead_time` outliers with boxplot & IQR.

4. **Data Types**
   - Created `arrival_date` column with proper datetime type.
   - Children/babies columns fixed as `Int64`.

5. **Potential Data Leakage**
   - Dropped `reservation_status` and `reservation_status_date` (contain info only available after booking).

6. **Inconsistent Categories**

   - Rare values in categorical columns merged into `"Other"`.

🔧 Phase 3: Feature Engineering & Preprocessing

- **New Features**

  - `total_guests = adults + children + babies`

  - `total_nights = stays_in_weekend_nights + stays_in_week_nights`

  - `is_family = 1 if children > 0 or babies > 0 else 0`

- **Encoding Categorical Variables**

  - One-Hot Encoding for low-cardinality columns (`meal`, `market_segment`).

  - Frequency Encoding for high-cardinality (`country`).

- **Removed Data Leakage**

  - Dropped `reservation_status` and `reservation_status_date`.

- **Final Preparation**

  - Train-test split: `train_test_split(df, test_size=0.2, random_state=42)`

🎯 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.

🗂️ Files in Repository

- `hotel_bookings.csv` → raw dataset.
- `notebook.ipynb` → data cleaning, preprocessing, EDA.
- `README.md` → project documentation.