GTC ML Project 1 – Data Cleaning & Preprocessing Challenge

Objective: Build a robust data preprocessing pipeline for a hotel booking cancellation prediction model.
Business Problem: The revenue team has identified that last-minute booking cancellations significantly impact profitability. Your task is not to build the final model, but to prepare the raw data for it. The quality of your data cleaning will directly determine the model's future success.

You are given a raw dataset (hotel_bookings.csv) direct from our Property Management System (PMS). Your goal is to transform it into a clean, machine-learning-ready dataset by completing the following phases:

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

- **Load the data** and **generate summary statistics** (.describe(), .info()).
- **Identify all missing values**. Create a visualization (e.g., a missingno matrix or a heatmap) to show the extent and pattern of missing data for each column.
- **Detect outliers** in key numerical columns (like adr and lead_time) using boxplots and the IQR method.
- **Document your findings** in your notebook. What are the main data quality issues?

Phase 2: Data Cleaning (The Core of the Project)

- **Handle Missing Values**: Develop and justify a strategy for each column.
  - For company and agent: Replace missing values with a label like "None" or 0.
  - For country: Impute with the mode (most frequent country) or a new "Unknown" category.
  - For children: A small number of missing values could be imputed with the median or mode.
- **Remove Duplicates**: Identify and drop any exact duplicate rows.
- **Handle Outliers**: Cap extreme values in columns like adr (e.g., any value above 1000 can be set to 1000) to prevent them from skewing future models. Justify your chosen method.
- **Fix Data Types**: Ensure date columns are correctly formatted.

Phase 3: Feature Engineering & Preprocessing

- **Create New Features**:
  - total_guests = adults + children + babies
  - total_nights = stays_in_weekend_nights + stays_in_week_nights
  - is_family = A binary flag (Yes/No) indicating if the booking includes children or babies.
- **Encode Categorical Variables**:
  - Use One-Hot Encoding for low-cardinality categories (e.g., meal, market_segment).
  - For high-cardinality features like country, use techniques like frequency encoding or grouping infrequent categories into an "Other" group.
- CRITICAL STEP: Remove Data Leakage: Immediately drop the columns reservation_status and reservation_status_date. These columns contain information that would not be available at the time of prediction and would make the model useless in a real-world scenario.
- Final Preparation: Split your cleaned dataset into training and testing sets (test_size=0.2, random_state=42).

The GTC Team