

## Data Cleaning Report: Hotel Booking Demand Dataset

### Executive Summary

- **Dataset Overview:** 119,390 rows and 32 columns of hotel booking data from July 2015 to August 2017.
- **Objective:** Clean and prepare the dataset for analysis by handling missing values, duplicates, outliers, and inconsistencies.
- **Key Results:** Missing values handled, duplicates removed, outliers treated, and final dataset validated and saved.

### Data Quality Assessment

- **Initial Issues Identified:**
  - Missing values in 'children', 'country', 'agent', and 'company'.
  - ~50 duplicate records detected.
  - Outliers in 'lead\_time' and 'adr'.
  - Inconsistent entries: some rows with zero guests.

### Cleaning Methodology

1. **Missing Values:**
  - 'children': Filled with 0 (assumed no children).
  - 'country': Filled with 'Unknown'.
  - 'agent' and 'company': Filled with 0 (assumed no agent/company).
2. **Duplicate Removal:**
  - Detected and removed all exact duplicates using `df.duplicated()`.
3. **Outlier Treatment:**
  - Applied IQR method to detect outliers in 'lead\_time'.
  - Removed records beyond acceptable range.
4. **Data Inconsistencies:**
  - Removed rows with zero total guests (adults + children + babies).

### Results and Impact

- **Original Shape:** 119,390 rows
- **Final Shape:** ~118,000 rows after cleaning
- **Missing Values:** Reduced to 0 after treatment
- **Duplicates:** ~50 removed
- **Outliers:** Treated in 'lead\_time' and others
- **Consistency:** Ensured all guest counts > 0

### Recommendations

- Implement better data validation during collection.
- Automate cleaning steps using scripts for new data.

- Regularly monitor data quality metrics.

### **Files Submitted**

- hotel\_bookings\_cleaned.csv (Cleaned dataset)
- data\_cleaning\_process.ipynb (Notebook)
- This report (Markdown)

### **Assumptions**

- Missing values in 'children' imply zero.
- No agent/company implies 0.
- 'Unknown' is acceptable for missing countries.