**Project Title:** Exploratory Data Analysis of Hotel Booking Data for Operational Insights and Revenue Optimization

**Submitted** By: Naresh Ramesh Khanderay

# **Introduction**

# This case study performs an Exploratory Data Analysis (EDA) of hotel booking data comprising records from a Resort Hotel and a City Hotel. The goal is to uncover operational patterns, guest behavior trends, and variables affecting revenue to enhance hotel management decisions.

# **Objectives**

* + Clean and preprocess the dataset for reliable analysis.
  + Explore trends in lead time, customer demographics, and booking behavior.
  + Investigate relationships between features and key metrics like ADR.
  + Conduct hypothesis tests to validate business assumptions.

**Project Summary**

The objective of this hotel booking analysis was to perform Exploratory Data Analysis (EDA) to uncover patterns, trends, and correlations within the booking data. The dataset contained over 119,390 records with detailed customer booking information from both City Hotels and Resort Hotels.

# **DATA CLEANING AND PREPROCESSING**

# Steps Taken:

# Import Libraries: Imported essential libraries such as pandas, numpy, matplotlib, and seaborn.

# Load Data: Loaded the dataset from hotel\_bookings.csv.

# Inspect Dataset: The dataset initially contained 119,390 rows and 32 columns.

# Handle Duplicates: Identified and removed 31,994 duplicate rows, resulting in a cleaned dataset with 87,396 records.

# Handle Missing Values:

# Columns children, country, and agent: Missing values were filled using the mode of each column.

# Column company: Dropped because it had more than 93% missing values.

# Data Types and Structure: Verified and corrected data types. Categorical and numerical variables were appropriately handled to prepare for analysis.

# Diagram Used:

# Bar Plot showing the number of missing values per column, helping identify which variables needed treatment.

# **Outlier Treatment:**

# Steps Taken for Outlier Treatment:

# Identified Outliers: Outliers were detected in the variables lead\_time and adr using:

# Boxplots (to visualize extreme values)

# Skewness distribution plots (to analyze the shape of data distribution)

# Applied IQR-Based Capping: Used the Interquartile Range (IQR) method to cap extreme values in lead\_time and adr, effectively reducing the influence of outliers and improving data quality for analysis.

# Diagram Used:

# Boxplots and Distribution Plots

# Before outlier removal: Highlighted the presence of extreme values.

# After outlier removal: Showed the effectiveness of IQR capping in normalizing the data.

# Steps Taken for Feature Engineering & Data Wrangling:

# Identified Variable Types: Categorized dataset columns into:

# Categorical Variables (e.g., hotel, customer\_type, reservation\_status)

# Discrete Numerical Variables (e.g., is\_canceled, agent, adr)

# Continuous Numerical Variables (e.g., lead\_time, adr)

# Analyzed Discrete Variables: Although they are numerical, discrete variables such as is\_canceled, agent, and adr were analyzed independently, considering their limited unique values.

# Handled Categorical Variables: Categorical features were treated appropriately for grouping, plotting, and summary statistics.

# Identified Continuous Variables: Variables like lead\_time and adr were identified for outlier treatment and statistical analysis.

# Diagram Used:

# Value Counts per Variable: Displayed the number of unique values in each column to support the classification into categorical, discrete, and continuous types.

**Exploratory Data Analysis (EDA)**

**Key Visualizations:**

* **Univariate Analysis:**  
  Plotted histograms for key features such as:

adr (Average Daily Rate)

lead\_time

customer\_type

market\_segment

and other categorical/numerical variables to observe individual distributions.

* **Bivariate Analysis:**

Boxplots to compare ADR across market segments.

Heatmap of correlation matrix to identify relationships among numerical variables and detect multicollinearity.

* **Time Series Analysis:**

Examined booking trends by month and year, revealing seasonality and demand patterns.

**Insights Gained:**

* Online Travel Agencies (OTA) dominate the booking volume across both hotel types.
* ADR (Average Daily Rate) shows noticeable variation across distribution channels.

Guests with higher lead times tend to make more booking changes, indicating longer decision windows.

**Correlation Analysis**

* adr (Average Daily Rate) shows moderate positive correlation with:

total\_of\_special\_requests

lead\_time

* Weak correlations observed between:

booking\_changes and adr

previous\_cancellations and is\_canceled

These results suggest that while some variables are mildly related to pricing or behavior, others may have minimal impact.

Diagram Used:

* Heatmap of the Correlation Matrix

Highlighted stronger and weaker relationships among numerical features.

Useful for identifying patterns and potential multicollinearity issues in the data.

**Hypothesis Testing**

Tests Performed:

1. **ADR: OTA vs Direct Bookings**

Null Hypothesis (H₀): There is no significant difference in ADR between Online Travel Agencies and Direct bookings.

Result: Rejected H₀ — A significant difference was found, indicating that booking channel influences room pricing.

1. **Room Upgrades vs Lead Time**

Null Hypothesis (H₀): Room upgrades are independent of the lead time.

Result: Failed to reject H₀ — No strong statistical evidence to suggest a relationship.

1. **Stay Duration vs Customer Type**

Null Hypothesis (H₀): There is no significant difference in stay duration across different customer types.

Result: Rejected H₀ — Significant variation in stay durations among customer categories (e.g., transient vs. contract).

**Diagram Used:**

* Boxplots – To compare distributions of ADR and stay durations across groups.
* Grouped Bar Charts – To visualize categorical comparisons and summarize group-based differences effectively.

**Key Findings**

Hotel and Booking Trends

Here are the key trends observed in hotel operations and bookings:

* Hotel Type Preference: City hotels are significantly more popular, receiving 61.07% of all bookings compared to resort hotels.
* Average Daily Rate (ADR) Drivers: The ADR is highest for guests who:

Make special requests.

Book with longer lead times.

* Cancellation Rate: Approximately one in every four reservations (25%) results in a cancellation.
* Most Favored Meal Plan: BB (Bed & Breakfast) is the preferred meal plan among guests.
* Primary Booking Platform: Bookings are predominantly made through the Online Travel platform.
* Leading Distribution Channel: TA/TO (Travel Agents/Tour Operators) are the leading distribution channel for bookings.
* Busiest Months: September and October are the busiest months for bookings.
* Minimal Wait Times: July experiences the shortest wait times for reservations.
* Room Assignment Mismatches: While room assignment mismatches are common, they do not significantly impact the Average Daily Rate (ADR) or lead to cancellations.

**Challenges Faced**

* + High dimensional dataset with categorical and numerical mix.
  + Null values and duplicate records.
  + Outlier skew affecting distribution shapes.
  + Complex feature relationships.
  + The data contained a large number of duplicates.
  + The improper data type format was used for the data.
  + It was challenging to select the best visualization techniques.
  + The dataset contained a large number of null values.

**Solution to Business Objective**

**Here's a well-formatted breakdown of the proposed strategies for hotel optimization:**

Strategic Recommendations for Hotel Optimization

**1. Optimize Resort Hotel Bookings**

Since city hotels receive more bookings, we should introduce targeted packages and seasonal promotions specifically designed to attract more customers to our resort locations.

**2. Diversify Meal Plan Preferences**

While BB (Bed & Breakfast) is the most requested meal type, we must maintain its high quality. To diversify preferences, ease kitchen load, and encourage variety, we should offer discounts on other meal plans.

**3. Streamline Online Booking Channels**

Given that most bookings originate from online platforms, we can improve efficiency by reducing or eliminating underperforming segments like complementary and aviation, which contribute minimally to overall bookings.

**4. Enhance Distribution Channel Strategy**

* Invest more in our top-performing channels: TA/TO (Travel Agents/Tour Operators) and Corporate channels, as they generate the highest number of bookings.
* Consider phasing out the GDS channel due to its very low activity.

**5. Boost Repeat Bookings**

With only 3.86% of guests being repeat customers, we need to implement strategies to improve retention:

* Offer loyalty programs or repeat booking incentives.
* Personalize marketing efforts based on guest history.
* Identify and target priority customers for specific retention initiatives.

**6. Rationalize Room Type Management**

* Maintain and enhance Room Types A and D, as they are the most preferred by guests.
* Promote Room Types E, F, and G through discounts to balance demand across our inventory.
* To reduce operational costs, eliminate Room Types B, C, H, and L, which are less frequently booked.

**7. Encourage Advance Deposits**

Guests currently avoid pre-deposit bookings. We need to educate guests on the benefits of advance deposits and actively promote them to:

* Accelerate revenue recognition.
* Reduce cancellations and no-shows.

**8. Optimize Parking Space Allocation**

Since spaces for 3 and 8 cars are rarely booked, we should limit our offerings to 1 or 2 spaces to optimize resources and maximize space efficiency.

**9. Improve Room Assignment Accuracy**

With 15% of guests not receiving their reserved room, it's crucial to ensure better room allocation accuracy to significantly improve customer satisfaction.

**10. Mitigate High Cancellation Rate**

Given the 25% cancellation rate, we should implement several measures:

* Establish a flexible yet firm cancellation policy.
* Offer discounts for non-refundable bookings.
* Send reminder notifications to guests to help minimize cancellations.

**11. Promote Group Bookings**

As most bookings are currently for two people, we can increase revenue by actively promoting family and group stays. This can be achieved by offering:

* Bundle deals.
* Group discounts.
* Event-specific promotions.

# **Conclusion**

This EDA provided insights into guest booking behavior and hotel operations. By cleaning data, visualizing key variables, and testing hypotheses, the study highlighted actionable areas for business improvements: - Optimize pricing by segment and channel. - Improve room allocation to reduce mismatches. - Target guests with higher ADR profiles.

**Tools & Libraries Used**

**Programming Language:**

Python

**Python Libraries:**

Pandas – for data manipulation and analysis

NumPy – for numerical operations

Matplotlib & Seaborn – for data visualization

Scipy – for statistical testing and hypothesis analysis

**Development Environment:**

**Jupyter Notebook** – for interactive coding, analysis, and visual documentation

# **References**

UCI Hotel Booking Demand Dataset

Assignment brief from CDAC Mumbai (AAS Case Study Feb 2025)

**END OF REPORT**