Hotel Booking Dataset – Exploratory Data Analysis (EDA) Report

Submitted By – Shubham Patil

1. Executive Summary

This report presents an end-to-end Exploratory Data Analysis (EDA) of the Hotel Booking dataset. The objective is to uncover patterns in customer booking behavior, pricing (ADR), cancellations, room upgrades, and other operational factors. Data preprocessing, statistical tests, and visual analysis have been performed to answer key business questions. The insights can help improve revenue, operations, and guest satisfaction.

**2. Problem Statement**

**Uncovering Patterns in Hotel Booking Data for Operational Efficiency and Revenue Growth**

The hospitality industry generates a vast amount of booking and customer data. However, this data is often underutilized when it comes to deriving actionable insights. The goal of this analysis is to explore and understand historical hotel booking data in order to uncover key patterns that impact pricing (ADR), customer behavior, room allocation, cancellations, and operational decisions.

By performing exploratory data analysis (EDA), we aim to identify:

* Factors that influence revenue through pricing trends
* Guest demographics that contribute to high-value bookings
* Market segments that are more consistent and profitable
* Operational inefficiencies such as frequent room upgrades or reassignments
* Booking patterns that lead to cancellations or special service needs

This analysis is crucial for helping hotels make data-driven decisions to improve forecasting, enhance customer satisfaction, reduce revenue leakages, and optimize resource planning. The insights gained can directly support both short-term operational efficiency and long-term revenue growth strategies.

3. Introduction

This project is part of a CDAC DBDA module assignment. It focuses on analyzing historical hotel booking data. The goal is to understand what factors influence booking prices, customer preferences, and hotel operations. We use Python, Pandas, Seaborn, and statistical methods to extract insights.

4. Dataset Overview

* Source: hotel\_bookings.csv
* Rows: 119,390
* Columns: 32
* Key Columns:
  + adr (Average Daily Rate)
  + lead\_time
  + booking\_changes
  + customer\_type
  + market\_segment
  + reserved\_room\_type
  + assigned\_room\_type
  + is\_canceled
  + total\_of\_special\_requests
  + country

5. Data Cleaning & Preprocessing

* Dropped column company due to 94% missing values.
* Filled agent with 0 and converted to integer.
* Filled country with mode.
* Filled children with 0 and converted to integer.
* Converted object columns (like hotel, meal) to category to save memory.
* Created arrival\_date column using year, month, and day.
* Derived columns created:
  + total\_stay = stays\_in\_week\_nights + stays\_in\_weekend\_nights
  + total\_guests = adults + children + babies
  + is\_upgraded = 1 if reserved ≠ assigned room type
* Removed duplicate rows.
* Outliers clipped at 99th percentile in lead\_time and adr.

6. Exploratory Data Analysis (EDA)

EDA was structured into three parts:

6.1 Univariate Analysis

* Histograms for numeric columns like adr and lead\_time.
* Countplots for categorical features like market\_segment and customer\_type.

6.2 Comparative Analysis

* Boxplots:
  + adr by market\_segment and customer\_type
  + lead\_time by customer\_type
  + adr vs is\_upgraded
* Observed upgrade patterns and price differences.

6.3 Time-Series & Demographics

* Monthly booking trends using arrival\_date\_month.
* Guest distribution by country.
* Top countries: Portugal, UK, France.

7. Correlation Analysis

* Pearson correlation matrix computed.
* Visualized with a heatmap.
* adr was positively correlated with:
  + total\_of\_special\_requests
  + booking\_changes
  + is\_upgraded
* Weak or no correlation with is\_canceled.

7. Hypothesis Testing

Three key hypotheses were tested:

H₀ 1:

ADR is the same between Online TA and Direct channels

* Test: 2-Sample T-Test
* p-value = 0.611
* Result: Fail to reject H₀ → No significant difference

H₀ 2:

Room upgrades are independent of lead time

* Test: Chi-Square Test of Association
* p-value ≈ 0
* Result: Reject H₀ → Lead time influences upgrades

H₀ 3:

Stay duration is the same across customer types

* Test: One-Way ANOVA
* p-value = 0.0
* Result: Reject H₀ → Stay duration varies by customer type

8. Key Business Questions & Insights

* What influences ADR the most?  
  → lead\_time, special\_requests, market\_segment, customer\_type
* Do guests who book early make more changes?  
  → Weak correlation; no strong link
* Do countries differ in ADR and lead\_time?  
  → Yes; Portugal, UK, and France show variation
* Is there a pattern in room upgrades?  
  → Yes; some customer types are upgraded more often
* Common guest profiles?  
  → Mostly 2 guests from Portugal, UK, and France
* Does guest type affect behavior?  
  → Transient guests book earlier and make more requests
* Does lead\_time vary by customer type or country?  
  → Yes, clear variation observed
* Are longer bookings more stable?  
  → Not always; weak relation with cancellations or changes
* Which segments are more consistent?  
  → Corporate and transient are stable with high ADR
* Do high-ADR guests request more?  
  → Yes, positively correlated

9. Conclusion

This EDA revealed several key insights into hotel booking behavior. ADR is strongly influenced by guest type, market segment, and lead time. Room upgrades are common, especially for certain customer types. Booking patterns vary by country and guest demographics. These insights help in customer targeting, demand forecasting, and revenue management.