REPORT ON-Analyzing Hotel Booking Demand: A Data-Driven Approach to Understanding Customer Behavior and Cancellation Patterns

NAME: RAJANNA ADHIKARY  
Date: 11 March 2025

TABLE OF CONTENTS

* - Brief description of the data set and its attributes.
* - Initial exploration plan.

* - Data cleaning and feature engineering actions.
* - Key findings and insights from the Exploratory Data Analysis.
* - Three hypotheses about the data.
* - Significance test for one hypothesis with results discussion.
* - Recommendations for next analysis steps.

- Summary of data quality and request for additional data

1. BRIEF DESCRIPTION OF THE DATASET AND ATTRIBUTE SUMMARY

> The dataset consists of 119,390 records of hotel bookings from two types of hotels: City Hotel and Resort Hotel. It includes 32 features, categorized as follows:

Lead Time: The number of days between the booking date and the arrival date. Longer lead times can indicate higher cancellation risks.

Customer Type: Indicates whether the booking is transient (one-time), part of a group, contract-based, or associated with a repeated guest.

Market Segment: The distribution channel used for booking (Direct, Online Travel Agent, etc.).

Cancellation Status: A binary column indicating if a booking was canceled (1) or not (0).

This dataset provides valuable insights into customer behavior, booking trends, and cancellation patterns.

Summary of Attributes

* Hotel Type: City Hotel or Resort Hotel
* Booking Details: Lead time, arrival date, and stay duration
* Guest Information: Number of adults, children, babies
* Reservation Status: Canceled or not, previous cancellations
* Financials: Average Daily Rate (ADR), deposit type
* Special Requests: Parking, meal preferences, booking changes

This dataset provides insights into customer preferences and booking trends, helping hotels optimize operations and improve customer experience.

# 2. INITIAL PLAN FOR DATA EXPLORATION

1. Load & Inspect Data – Check structure, data types, and missing values.

2. Clean Data – Handle missing values, remove duplicates, and format dates.

3. Explore Trends – Analyze booking patterns, cancellations, and seasonality.

4. Hypothesis Testing – Test assumptions about cancellations and booking behavior.

5. Visualize Insights – Use charts to highlight key findings.

3. ACTIONS TAKEN FOR DATA CLEANING AND FEATURE ENGINEERING

Data Cleaning Steps

Handled Missing Values:

Children had missing values. → Replaced with most frequent value.

country had missing values → categorized as "Unknown" to prevent data loss.

Removed Duplicates: Checked for duplicate bookings and removed them if found.

Feature Engineering

Created total\_nights by summing stays\_in\_week\_nights and stays\_in\_weekend\_nights.

Converted categorical variables like hotel and customer\_type into numerical format for statistical analysis.

4. KEY FINDINGS AND INSIGHTS FOR--Exploratory Data Analysis (EDA)

1. Peak Seasonality & Booking Trends

Observation: The highest number of bookings occur in July and August, while January has the lowest bookings.

Business Impact: Demand-based pricing strategies can help maximize revenue during peak months and offer discounts in off-season to attract more guests.

2️Higher Cancellation Rates for Resort Hotels

Observation: Resort hotels have a higher cancellation rate than city hotels.

Possible Reasons: Vacation plans are more flexible, leading to more cancellations.

Recommendation: Implement stricter cancellation policies or non-refundable discounts to reduce losses.

3️ impact of Lead Time on Cancellations

Observation: Longer lead times (bookings made well in advance) show higher cancellation rates.

Why?: Guests may book multiple hotels and cancel later.

Actionable Step: Introduce deposit-based reservations for bookings made months in advance.

4️.Online Travel Agents (OTA) Drive More Cancellations

Observation: Bookings made through OTAs (like Expedia, Booking.com) have higher cancellation rates than direct bookings.

Why?: OTAs often offer free cancellations, leading to speculative bookings.

Business Strategy: Encourage direct bookings by offering loyalty discounts and exclusive deals on the hotel’s official website.

# 5. HYPOTHESIS FORMULATION

1. Customers with longer lead times are more likely to cancel bookings.

2. City hotels receive more bookings than resort hotels throughout the year.

3. Guests booking through Online Travel Agents (OTA) have a higher cancellation rate than those booking directly.

# 6. FORMAL SIGNIFICANT TEST

Hypothesis: Longer lead times lead to a higher probability of cancellation.

Test Performed: Chi-Square Test

Null Hypothesis (H₀): There is no relationship between lead time and cancellation.

Alternative Hypothesis (H₁): Longer lead times lead to a higher probability of cancellation.

Chi-Square Test Result (Python Code):

from scipy.stats import chi2\_contingency

import pandas as pd

# Create a contingency table

contingency\_table = pd.crosstab(df['is\_canceled'], df['lead\_time'] > 100)

# Perform the Chi-Square test

chi2, p, dof, expected = chi2\_contingency(contingency\_table)

print(f"Chi-Square p-value: {p}")

If p < 0.05, we reject H₀, meaning lead time significantly impacts cancellation.

If p > 0.05, we fail to reject H₀, meaning lead time does not influence cancellations.

# 7. RECOMMENDATIONS FOR NEXT ANALYSIS STEP

1. Explore Other Factors:

Analyze additional variables like booking channels, customer demographics, and duration of stay to further understand cancellation patterns.

2. Predictive Modeling:

Develop a predictive model (e.g., logistic regression or decision trees) to predict cancellations based on lead time, hotel type, and other features.

3. Seasonal Trends:

Investigate how seasonal factors or local events affect cancellation rates and booking behavior.

4. Customer Segmentation:

Segment the data by customer types (e.g., families, business travelers) to identify specific patterns in cancellations and preferences.

5. Outlier Detection:

Identify and analyze outliers in the data, especially in lead times, to ensure robustness in modeling and interpretation.

# 8. SUMMARY OF DATA QUALITY AND REQUEST FOR ADDITIONAL DATA

Data Quality Summary

The dataset is generally well-structured, with key attributes such as hotel type, lead time, cancellation status, and family presence clearly defined. However, it is missing some important aspects, including booking channels, customer demographics, and seasonal trends, which could enhance the depth of the analysis. While the data is sufficient for basic exploratory analysis, greater granularity could lead to more accurate modeling and insights.

Request for Additional Data

To improve the analysis and enhance predictive capabilities, it would be beneficial to obtain additional data on booking channels, customer demographics (such as age and nationality), seasonal trends, and external factors (like local events). This information would enable a more comprehensive understanding of booking behavior and cancellation patterns.