Hotel Booking Data Analysis: Exploratory Data Analysis (EDA)

Submitted By:
Name: Shivam Verma **PRN:** 22070521132

Section: C **Semester:** 7

Batch: 2022-2026

Under the Guidance of: Dr. Piyush Chauhan **Course:** Machine Learning (CA1 Submission)

Institute: Symbiosis Institute of Technology (SIT), Nagpur

1. Introduction

The Technical Report and Project outline an end-to-end Exploratory Data Analysis (EDA) of hotel booking data. The data, collected from various sources within hotels, is structured to cover booking transactions, hotel details, room categories, and calendar information.

The primary goal of this analysis is to:

- Explore operational patterns within hotel bookings.
- Understand customer behaviour patterns.
- Derive insights into occupancy trends and room category preferences.
- Analyse booking behaviours across different timeframes.

The overarching objective is to provide data-driven observations that assist stakeholders in making informed decisions.

2. Dataset Description

This project utilizes a set of interconnected CSV files to conduct an Exploratory Data Analysis (EDA) on hotel booking data. These datasets provide a comprehensive view of booking transactions, hotel characteristics, room types, and temporal information.

-> **orignal_bookings_data.csv (df_bookings):** Contains detailed records of individual booking transactions, including booking IDs, dates (booking, check-in, check-out), number of guests, room categories, booking platforms, and generated/realized revenue.

-> **fact_aggregated_bookings.csv (df_agg_bookings):** Provides aggregated booking data, summarizing successful bookings and room capacities by property, check-in date, and room category.

-> **dim_date.csv (df_date):** A dimension table offering temporal attributes such as calendar date, month-year, week number, and day type (weekday/weekend).

-> **dim_hotels.csv (df_hotels):** A dimension table with meta-information about each hotel property, including property ID, name, category (e.g., Luxury, Business), and city.

-> **dim_rooms.csv (df_rooms):** A dimension table detailing room types with room IDs and their corresponding room classifications (e.g., Standard, Elite, Premium, Presidential).

Dataset Link: **Datasets Received By Company (View)**

GitHub Link: https://github.com/shivamverma9808/Machine-Learning-EDA.git

---

3. Importing Libraries

This section involves loading all essential Python libraries required for data analysis and visualization.

Libraries Imported:

- pandas (as pd) for data manipulation and analysis
- numpy (as np) for numerical operations
- matplotlib.pyplot (as plt) for plotting and visualization
- seaborn (as sns) for enhanced statistical data visualization
- plotly.express (as px) for interactive plotting

---

4. Data Exploration

Initial data exploration involved a brief review of each dataset's structure, including column names, data types, and sample records, to gain a clear overview before deeper analysis.

df_bookings (orignal_bookings_data.csv):

- **Head (Sample Records):** Displays the first few rows, showing columns like booking_id, property_id, booking_date, check_in_date, checkout_date, no_guests, room_category, booking_platform, ratings_given, booking_status, revenue_generated, and revenue_realized.
- **Shape:** The dataset contains 134,590 rows and 12 columns.
- **Unique Room Categories:** The unique room categories found are 'RT1', 'RT2', 'RT3', 'RT4'.
- **Unique Booking Platforms:** The unique booking platforms are 'direct online', 'others', 'logtrip', 'tripster', 'makeyourtrip', 'journey', 'direct offline'.
- **Descriptive Statistics (.describe()):**
    - no_guests shows a minimum value of -17.0, confirming negative entries.

- ratings_given has only 56,683 non-null entries out of 134,590, indicating a significant number of missing ratings.
- revenue_generated has a maximum of 2.856e+07, while revenue_realized maxes at 45220, highlighting the large outlier.

df_hotels (dim_hotels.csv):

- **Shape:** The dataset contains 25 rows and 4 columns.
- **Head (Sample Records):** Shows property_id, property_name, category (e.g., Luxury, Business), and city (e.g., Delhi, Mumbai).
- **City Distribution:** A bar chart of df_hotels.city.value_counts() was generated, showing the number of hotels in each city.

df_agg_bookings (fact_aggregated_bookings.csv):

- **Head (Sample Records):** Displays property_id, check_in_date, room_category, successful_bookings, and capacity.

---

5. Data Cleaning and Feature Engineering

This section addresses missing values, corrects data types, and handles inconsistencies in the datasets.

Cleaning df_bookings (Original Bookings Data):

1. **Handling Invalid Entries (no_guests):**
   - **Problem:** The no_guests column contained negative values (e.g., -3.0, -2.0, -10.0, -17.0).
   - **Action:** The .abs() function was used to convert all negative no_guests values to their positive equivalents.
2. **Handling Missing Values (no_guests):**
   - **Problem:** The no_guests column had 3 missing values, which is minimal.
   - **Action:** These missing values were filled using the median() of the no_guests column to maintain data consistency without significantly impacting analysis.
3. **Handling Missing Values (ratings_given):**
   - **Problem:** Out of 134,590 total values, 77,907 rows had null ratings.
   - **Action:** The ratings_given column was dropped entirely because a large number of missing values (over 50%) would make imputation unreliable, and the column was deemed not critical for the main analysis goals.
4. **Handling Outliers (revenue_generated):**
   - **Problem:** The revenue_generated column contained a significant outlier (e.g., 9,100,000 when revenue_realized was 9,100). This suggests data entry errors where values might have been incorrectly scaled.
   - **Action:** Instances where revenue_generated was greater than revenue_realized were corrected by setting revenue_generated equal to revenue_realized.
5. **Standardizing Date Formats (booking_date, check_in_date, checkout_date):**
   - **Problem:** Date columns (booking_date, check_in_date, checkout_date) had inconsistent formats (e.g., YYYY-MM-DD and D/M/YYYY mixed) and previous automatic inference (infer_datetime_format=True) resulted in many unparsed dates (NaT).
   - **Action:** A robust custom function parse_date_robust was implemented to attempt parsing dates using a predefined list of common formats (%d-%m-%Y, %d/%m/%Y, %Y-%m-%d, etc.). This

function was applied to all three date columns, ensuring all dates were successfully converted to datetime objects and then formatted uniformly as YYYY-MM-DD strings.

Cleaning df_agg_bookings (Fact Aggregated Bookings Data):

1. **Handling Missing Values (capacity):**
    - **Problem:** The capacity column had 2 missing values.
    - **Action:** These missing values were filled using the median() of the capacity column.
2. **Standardizing Date Format (check_in_date):**
    - **Action:** The check_in_date column was converted to datetime objects using its specific format (%d-%b-%y) and then formatted consistently as YYYY-MM-DD strings.

---

6. Data Transformation

This section involves creating derived columns to prepare the data for meaningful analysis.

Creating Occupancy Percentage (occ_pct) in df_agg_bookings:

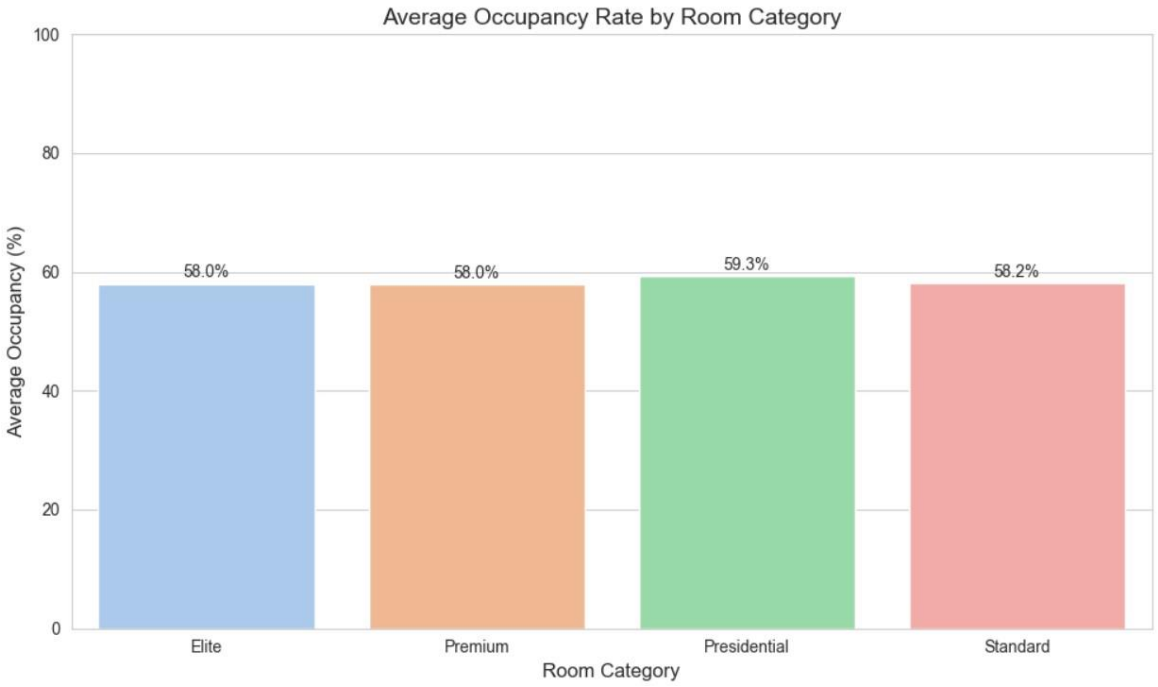
- A new column occ_pct was created by dividing successful_bookings by capacity.
- The raw occupancy values (e.g., 0.833333) were then converted to a percentage format (e.g., 83.33) by multiplying by 100 and rounding to two decimal places.

---

7. Meaningful Insights Generation

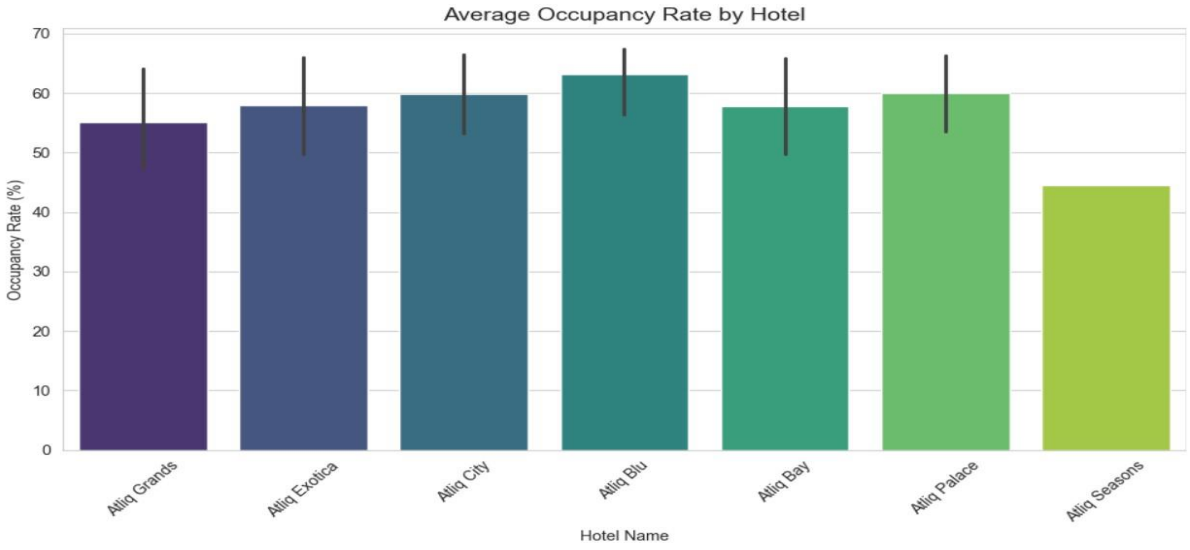
This section presents key business insights derived through various analyses and visualizations.

1. **What is an average occupancy rate in each of the room categories?**



- **Analysis:** The average occupancy percentage (occ_pct) was calculated for each room_category after mapping them to more meaningful names (Standard, Elite, Premium, Presidential).
- **Key Insight:** This chart visually represents how popular each room category is in terms of average occupancy, highlighting which categories are most frequently booked.

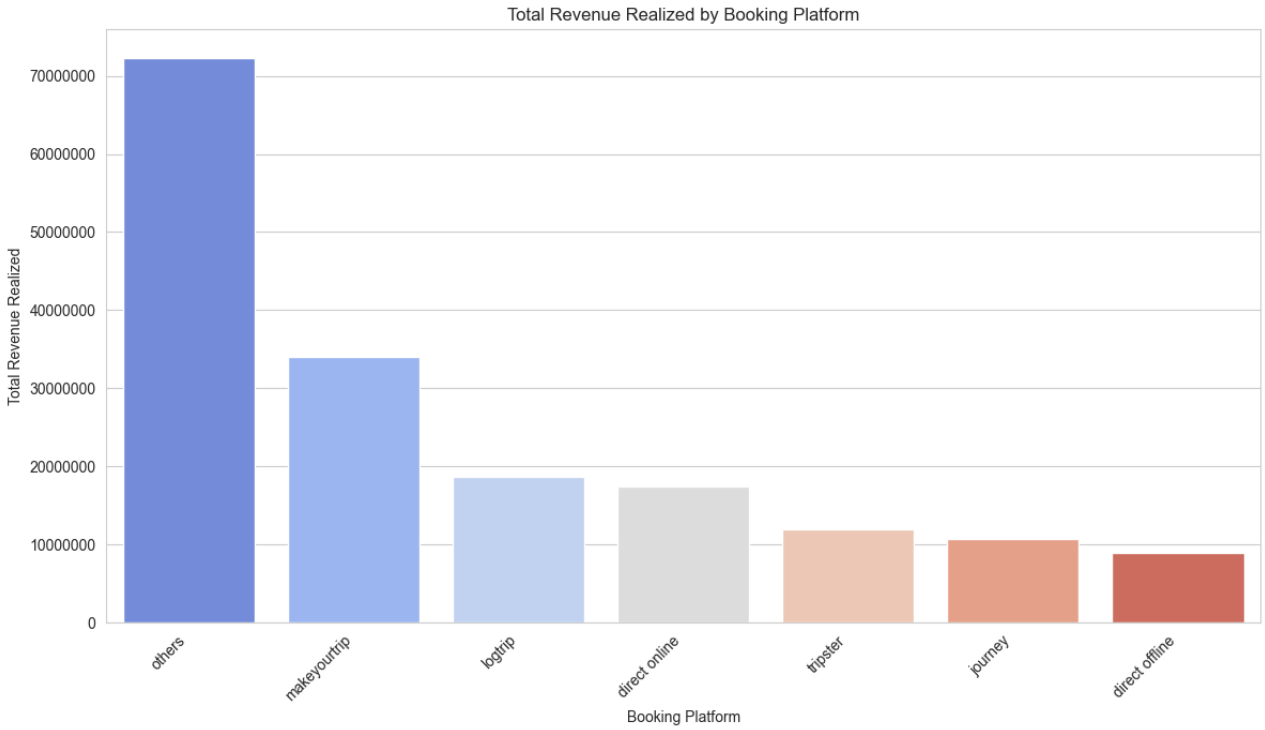
2. **Which hotel has the highest occupancy rate?**



- **Analysis:** The average occupancy rate was calculated for each property_id and then merged with df_hotels to include property_name.

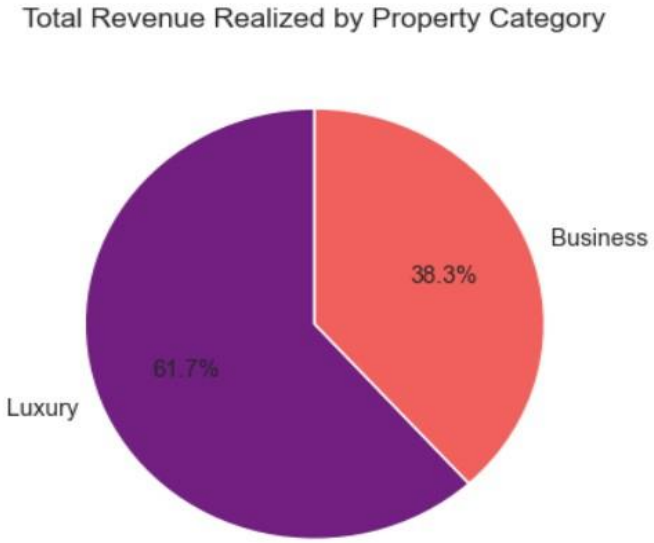
- **Key Insight:** This chart allows for direct comparison of occupancy performance across different hotels, identifying top-performing properties.

3. **What's the revenue realized per booking platform?**



- **Analysis:** The total revenue_realized was summed for each booking_platform.
- **Key Insight:** This visualization helps identify which booking platforms are most effective in generating revenue, which can inform marketing and partnership strategies.

4. **What is the total revenue contribution of each property category?**

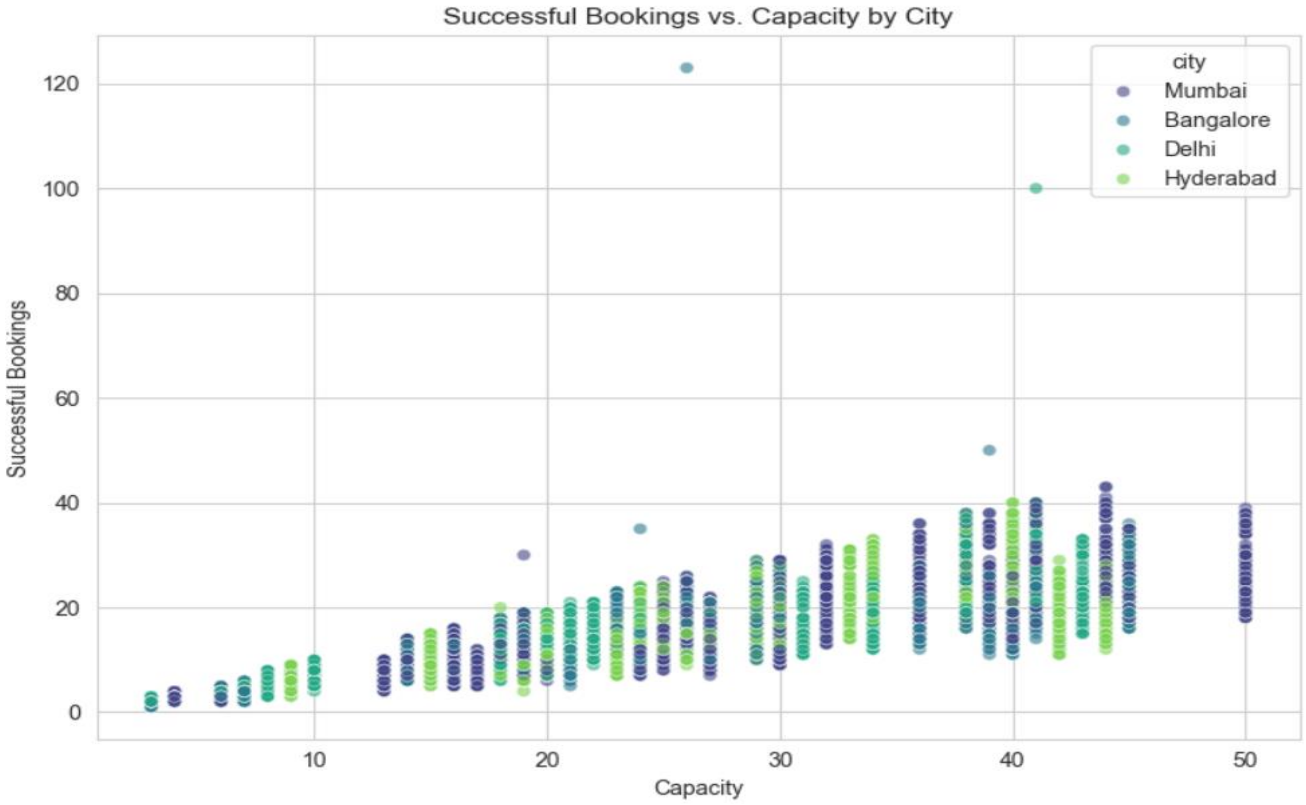


- **Analysis:** The total revenue_realized was summed for each category (Luxury, Business).
- **Key Insight:** This pie chart provides a clear overview of the revenue distribution across property categories, indicating which category is the primary revenue driver.

5. **What is the monthly trend in revenue realized over time?**

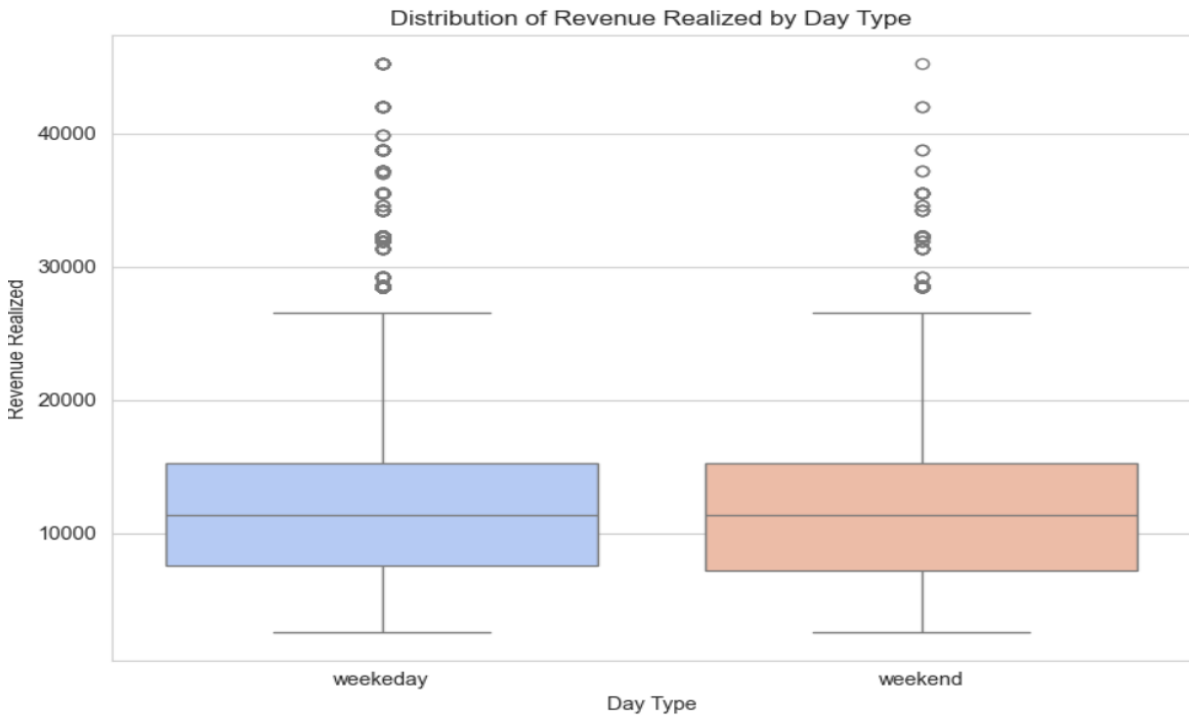
- **Analysis:** The check_in_date was used to extract the month and year (booking_month_year), and then the total revenue_realized was summed for each month.
- **Key Insight:** This trend line helps in understanding seasonality in revenue, identifying peak seasons, and observing any growth or decline patterns over the analyzed period.

6. **How does hotel capacity relate to the number of successful bookings across different cities?**



- **Analysis:** A scatter plot was used to visualize the relationship between capacity and successful_bookings, with data points colored by city.
- **Key Insight:** This plot helps identify if properties are utilizing their full capacity effectively and if there are differences in booking-to-capacity ratios across different cities. It can also highlight properties that are over or underperforming relative to their capacity.

7. **What is the variation in revenue realized between weekdays and weekends?**



- **Analysis:** The distribution of revenue_realized was examined across day_type (weekday/weekend).
- **Key Insight:** This box plot provides insights into the central tendency, spread, and outliers of revenue generation on weekdays versus weekends, indicating if one period significantly outperforms the other.

7. Future Scope

The Exploratory Data Analysis has established a robust understanding of the hotel booking data, laying a crucial foundation for more advanced analytical endeavours. The insights gained from data cleaning, transformation, and initial metric analysis directly pave the way for leveraging Machine Learning to extract predictive and prescriptive intelligence.

For the future scope of this project and to further enhance decision-making within the hospitality domain, the following Machine Learning algorithms can be implemented:

- **Booking Cancellation Predictions**
  Algorithm: Logistic Regression, Random Forest.
  Use booking features (e.g., lead time, room type, previous cancellations) to predict the likelihood of a booking being canceled. This helps in proactive overbooking strategies.

- **Customer Segmentation**
  Algorithm: K-Means Clustering.
  Segment guests based on behaviour (e.g., length of stay, special requests, room preferences) to enable targeted marketing and personalized services.

- **Recommendation System**
  Algorithm: Collaborative Filtering, Content-Based Filtering.
  Recommend suitable room types or packages based on customer history and preferences to improve customer satisfaction and retention.

- **Sentiment Analysis**
  Algorithm: NLP techniques like Naive Bayes, SVM, or BERT.
  Analyze customer feedback to understand service quality and areas of improvement.

8. Conclusion

The comprehensive Exploratory Data Analysis performed on the hotel booking dataset has yielded crucial insights into operational efficiency, customer behaviour, and revenue dynamics within the hospitality sector. The rigorous processes of data cleaning and transformation ensured the reliability and quality of the data for robust analysis.

This EDA confirms such main conclusions:

- **Varied Occupancy Across Room Categories and Hotels:** Analysis revealed distinct average occupancy rates for different room categories (e.g., Standard, Elite, Premium, Presidential) and significant variations in occupancy performance among individual hotel properties. These highlights demand patterns for specific room types and identifies top-performing hotels.
- **Revenue Driven by Specific Platforms and Categories:** The total revenue realized shows a clear dependence on booking platforms, indicating which channels are most profitable. Furthermore, property categories (Luxury vs. Business) contribute differently to overall revenue, informing strategic focus.
- **Pronounced Monthly Revenue Seasonality:** The analysis of monthly revenue trends indicates strong seasonal patterns in bookings and revenue generation. Understanding these fluctuations is crucial for forecasting, resource allocation, and targeted marketing campaigns.
- **Utilization Patterns of Hotel Capacity:** A scatter plot of successful bookings versus capacity, segmented by city, revealed insights into how effectively hotels are utilizing their available rooms. This helps identify properties that might be under-utilized or consistently operating at high capacity.
- **Revenue Discrepancies Between Weekdays and Weekends:** A distinct distribution of realized revenue was observed when comparing weekdays and weekends, suggesting varying booking values or volumes during these periods.
- **Importance of Data Quality in Hospitality Analytics:** The EDA process necessitated extensive cleaning, including handling negative guest counts, imputing missing capacity values, managing outlier revenue entries, and standardizing diverse date formats. This underscores the critical role of data preparation in deriving accurate and actionable business insights.

This EDA serves as a foundational analytical asset, providing actionable insights into hotel performance, customer preferences, and market dynamics, which are essential for strategic decision-making and operational optimization in the hotel industry.