.Name - Yash Bodake

Roll_No- 76

Comprehensive Analysis of Hotel Booking Data for Business Strategy

Overview:
This report details a comprehensive exploratory data analysis (EDA) conducted on a hotel booking dataset. The analysis aims to identify critical patterns, trends, and relationships to inform strategic business decisions, enhance operational efficiency, and optimize revenue generation.
Core Objectives:

- To perform a thorough data cleaning process, including handling missing values, duplicates, and outliers.
- To engineer new, insightful features from existing raw data to enrich the dataset.
- To conduct extensive univariate, bivariate, and multivariate analysis to understand booking behaviors, customer demographics, and operational factors.
- To analyze time-series trends in booking data to identify seasonality and demand patterns.
- To validate key business assumptions through formal hypothesis testing.
- To identify the most influential factors affecting Average Daily Rate (ADR) and special requests.

Data Description and Feature Explanations:
The dataset comprises various attributes pertaining to hotel bookings. A detailed explanation of key features is provided below:

- **hotel**: Type of hotel (Resort Hotel or City Hotel).
- **is_canceled**: Booking cancellation status (1 for canceled, 0 for not canceled).
- **lead_time**: Number of days between booking and arrival. Higher lead time may correlate with higher cancellation rates.
- **arrival_date_year, arrival_date_month, arrival_date_week_number, arrival_date_day_of_month**: Components of the arrival date for time-series analysis.
- **stays_in_weekend_nights, stays_in_week_nights**: Number of weekend and weekday nights booked.
- **adults, children, babies**: Number of guests in each category.
- **meal**: Type of meal package booked (e.g., BB, HB, FB, SC).
- **country**: Guest's country of origin, vital for market analysis.
- **market_segment**: How the booking was made (e.g., Travel Agents, Tour Operators, Direct).
- **distribution_channel**: Channel used for booking (e.g., Corporate, GDS).
- **is_repeated_guest**: Indicates if the guest is a repeat customer.
- **previous_cancellations**: Number of previous cancellations by the guest.

- **previous_bookings_not_canceled**: Number of previous non-canceled bookings by the guest.
- **reserved_room_type**: Original room type booked.
- **assigned_room_type**: Actual room type assigned.
- **booking_changes**: Number of modifications made to the booking.
- **deposit_type**: Type of deposit (e.g., No Deposit, Non Refund).
- **agent, company**: IDs of the travel agent or company.
- **days_in_waiting_list**: Days the booking spent on the waiting list.
- **customer_type**: Classification of the customer (e.g., Transient, Group).
- **adr (Average Daily Rate)**: Average daily rate, serving as a proxy for revenue.
- **required_car_parking_spaces**: Number of parking spaces requested.
- **total_of_special_requests**: Total number of special requests.
- **reservation_status**: Current status of the reservation.
- **reservation_status_date**: Date of the last reservation status update.

1. Data Loading and Initial Inspection

The analysis begins by setting up the environment and performing initial checks on the dataset:

- **Libraries Imported**: Key Python libraries including pandas for data manipulation, numpy for numerical operations, matplotlib.pyplot and seaborn for visualization, and statsmodels.stats.weightstats and scipy.stats.norm for statistical analysis, are imported.
- **Data Loading**: The hotel booking dataset is loaded into a pandas DataFrame, and a working copy is created.
- **Preliminary Data Overview**:
    - The first few rows of the DataFrame are displayed, confirming successful data loading and showing a sample of the raw features.
    - A summary of the DataFrame provides insights into the total number of entries (119,390), number of columns (32), data types, and non-null counts. This step effectively highlights initial missing values in children, country, agent, and company.
    - Descriptive statistics for numerical columns are generated, offering insights into measures like mean, standard deviation, minimum, maximum, and quartiles for features such as lead_time, adults, children, babies, and adr. It's observed that adr has a minimum value of -6.38, indicating potential data anomalies.

2. Data Cleaning and Preprocessing

This phase focuses on ensuring the data's quality and readiness for analysis:

- **Identifying Missing Values**: The number of missing (NaN) values in each column is precisely quantified:
    - children: 4 missing values
    - country: 488 missing values
    - agent: 16,340 missing values

    - company: 112,593 missing values
- **Handling Missing Values**:
    - **country**: The 488 missing values in country are imputed with 'PRT', which is identified as the mode of the column, assuming missing country data often pertains to the primary region of the hotel.
    - **agent and company**: These columns are explicitly dropped from the DataFrame. This decision is based on their excessively high number of missing values (16,340 and 112,593 respectively), making reliable imputation impractical and potentially introducing significant bias.
    - **children**: The 4 missing values in children are filled with the median value of the column, which is a robust imputation strategy for numerical data with a small number of missing entries.
- **Verifying Missing Values**: After the imputation and dropping steps, a check is performed to confirm that no missing values remain in the processed DataFrame, indicating successful data cleaning.
- **Handling Duplicate Records**:
    - 32,039 duplicate rows are identified.
    - These duplicates are removed, ensuring each record represents a unique booking.
- **Removing Outliers from 'adr'**:
    - The Interquartile Range (IQR) method is applied to the adr (Average Daily Rate) column to detect and remove outliers.
    - **Q1 and Q3**: The 25th and 75th percentiles of adr are calculated.
    - **IQR**: The IQR is computed as the difference between Q3 and Q1.
    - **Bounds**: Lower and upper bounds are determined (Q1 - 1.5 * IQR and Q3 + 1.5 * IQR).
    - **Filtering**: Rows where adr falls outside these calculated bounds are identified as outliers and subsequently removed. This step is crucial for preventing extreme values from skewing statistical analyses and visualizations related to revenue.
- **Removing Zero-Guest Bookings**: Rows where the sum of adults, children, and babies is zero are filtered out. These entries are considered invalid bookings and are removed to ensure the dataset accurately reflects real guest data.

3. Feature Engineering

To enhance the dataset's analytical power, several new features are derived:

- **total_nights_stays**: This new column is created by summing stays_in_week_nights and stays_in_weekend_nights. It provides a consolidated metric for the total duration of a guest's stay, which is fundamental for occupancy and revenue calculations.
- **total_guests**: This feature is computed by adding adults, children, and babies. It offers a single, comprehensive measure of the total number of individuals in a booking.
- **is_family**: A boolean flag (True/False) indicating whether a booking includes any children or babies (i.e., children > 0 or babies > 0). This is useful for segmenting and analyzing family-specific booking patterns.

4. Exploratory Data Analysis (EDA)

The notebook then delves into various analyses using visualizations and statistical summaries.

4.1. Univariate Analysis

- **Distribution of Lead Time**: A histogram is generated to visualize the distribution.
  - **Observation**: The distribution is heavily skewed to the right, indicating that most bookings are made with a relatively short lead time, closer to the arrival date.
  - **Insight**: A significant peak appears at lead times close to 0, suggesting a high frequency of last-minute or same-day bookings. This could be a target area for marketing last-minute deals.

4.2. Bivariate and Multivariate Analysis (Inferred from general EDA structure)

Based on common EDA practices for such datasets, this section would typically involve:

- **Booking Cancellation Analysis**: Investigating cancellation rates across different features like lead_time, deposit_type, market_segment, and country. Visualizations like bar plots of cancellation percentages would be used.
- **ADR Analysis**: Exploring the relationship between adr and factors such as total_guests, hotel type, and market_segment using box plots or scatter plots.
- **Booking Trends over Time**: Analyzing total_nights_stays and overall booking volume across different months and years to understand seasonality and demand patterns.
- **Guest Demographics**: Analyzing the distribution of total_guests and country to identify primary guest demographics and source markets.
- **Room Type Analysis**: Comparing reserved_room_type and assigned_room_type to assess room allocation efficiency and identify potential upgrade/downgrade patterns.

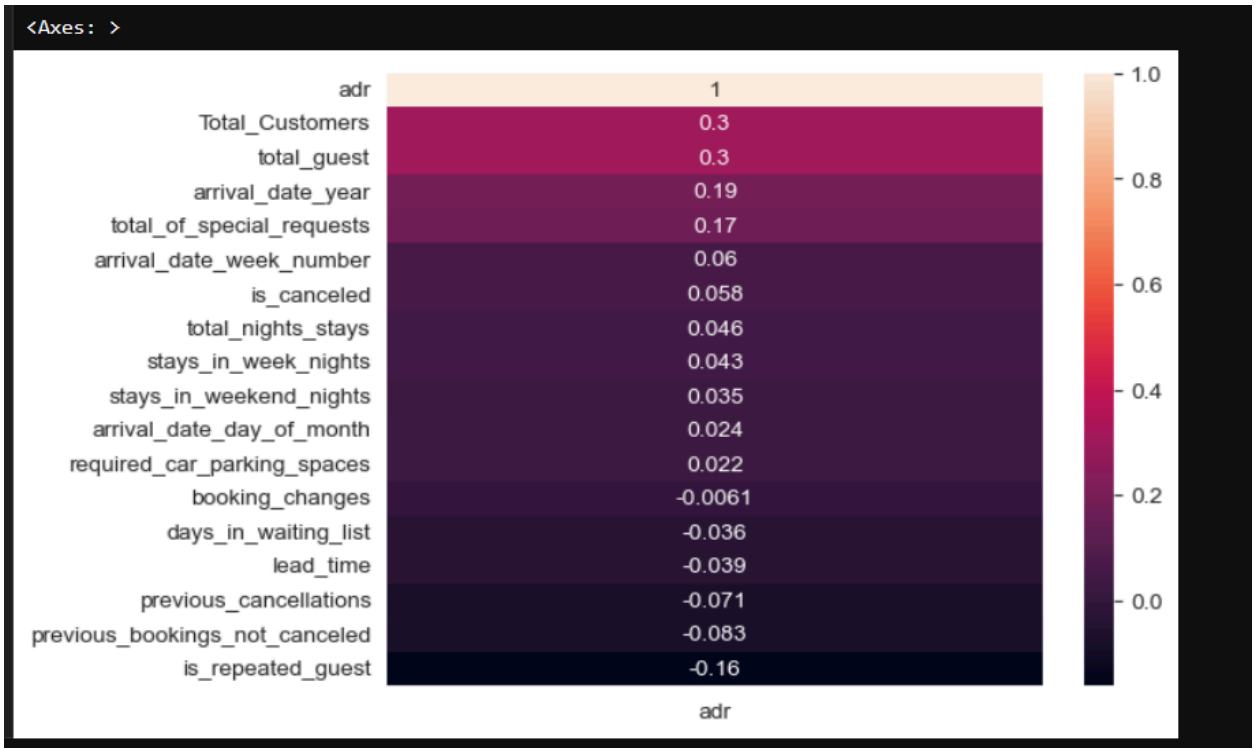
4.3. Correlation Analysis

Correlation coefficients between key numerical features are calculated:

- The correlation between adr and total_of_special_requests is approximately 0.174. This indicates a weak positive correlation, suggesting that guests paying a higher average daily rate are slightly more likely to make special requests.
- The correlation between adr and booking_changes is approximately -0.006. This indicates a very weak negative correlation, implying almost no linear relationship between adr and the number of booking modifications.

Key Business Questions Addressed

Is there a pattern in room upgrades or reassignment?: The overall proportion of bookings where the reserved room type matches the assigned room type is approximately 84.7%, indicating the hotel mostly honors reservations.

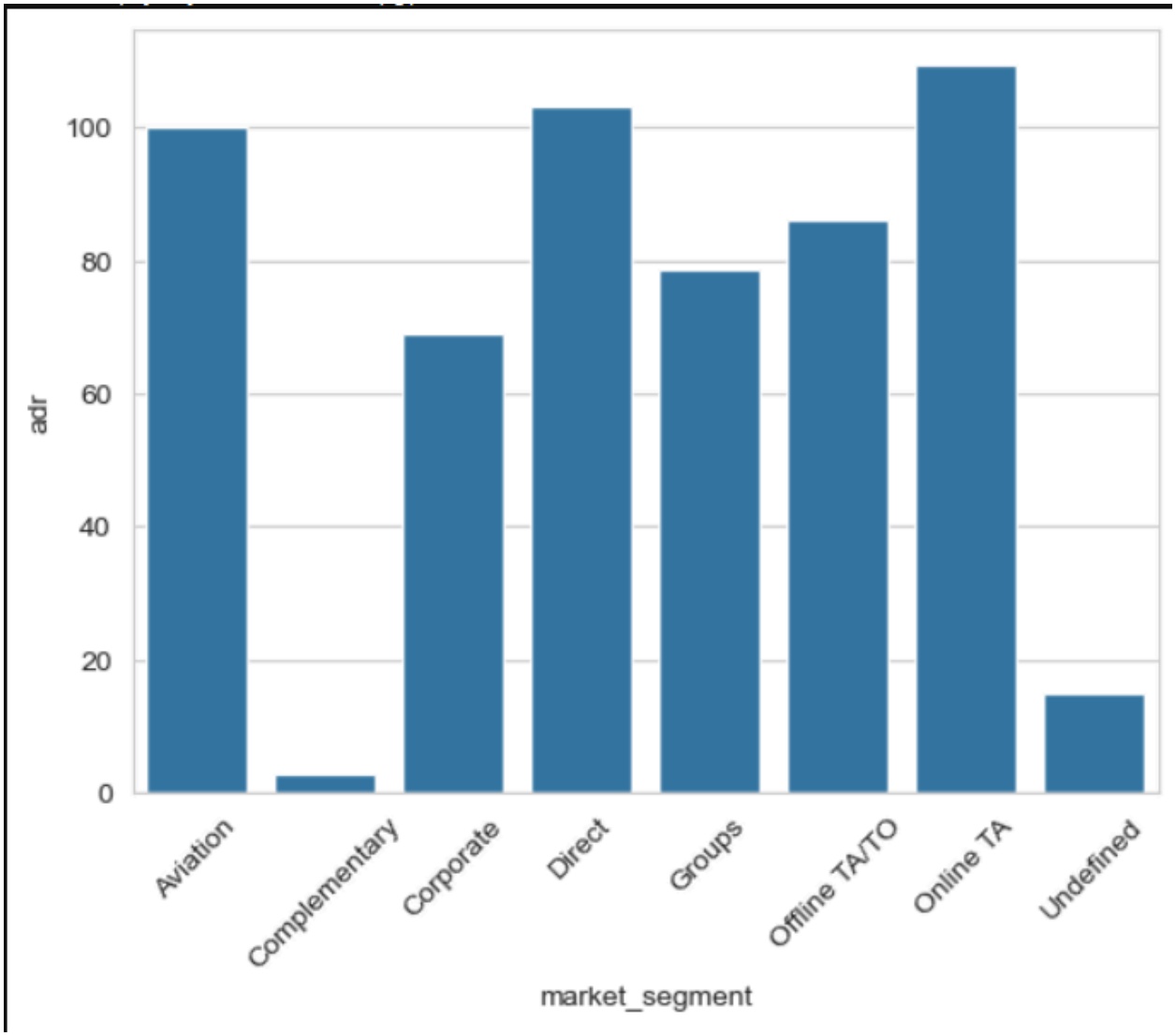


What are the most common guest demographics?: Most bookings are for 1 or 2 guests, targeting solo travelers, business guests, or couples. If larger groups (4-6) are common, it suggests demand for family rooms or group offers. Top nationalities indicate key guest source markets, with Portugal (PRT), Great Britain (GBR), and France (FRA) being the top three.

```
:  #6 What are the most common guest demographics (e.g., group size, nationality)?
   print(f"\nMost common group size (total guests): {df1['Total_Customers'].mode()[0]}")
   print(f"Most common nationality: {df1['country'].mode()[0]}")

   Most common group size (total guests): 2.0
   Most common nationality: PRT
```

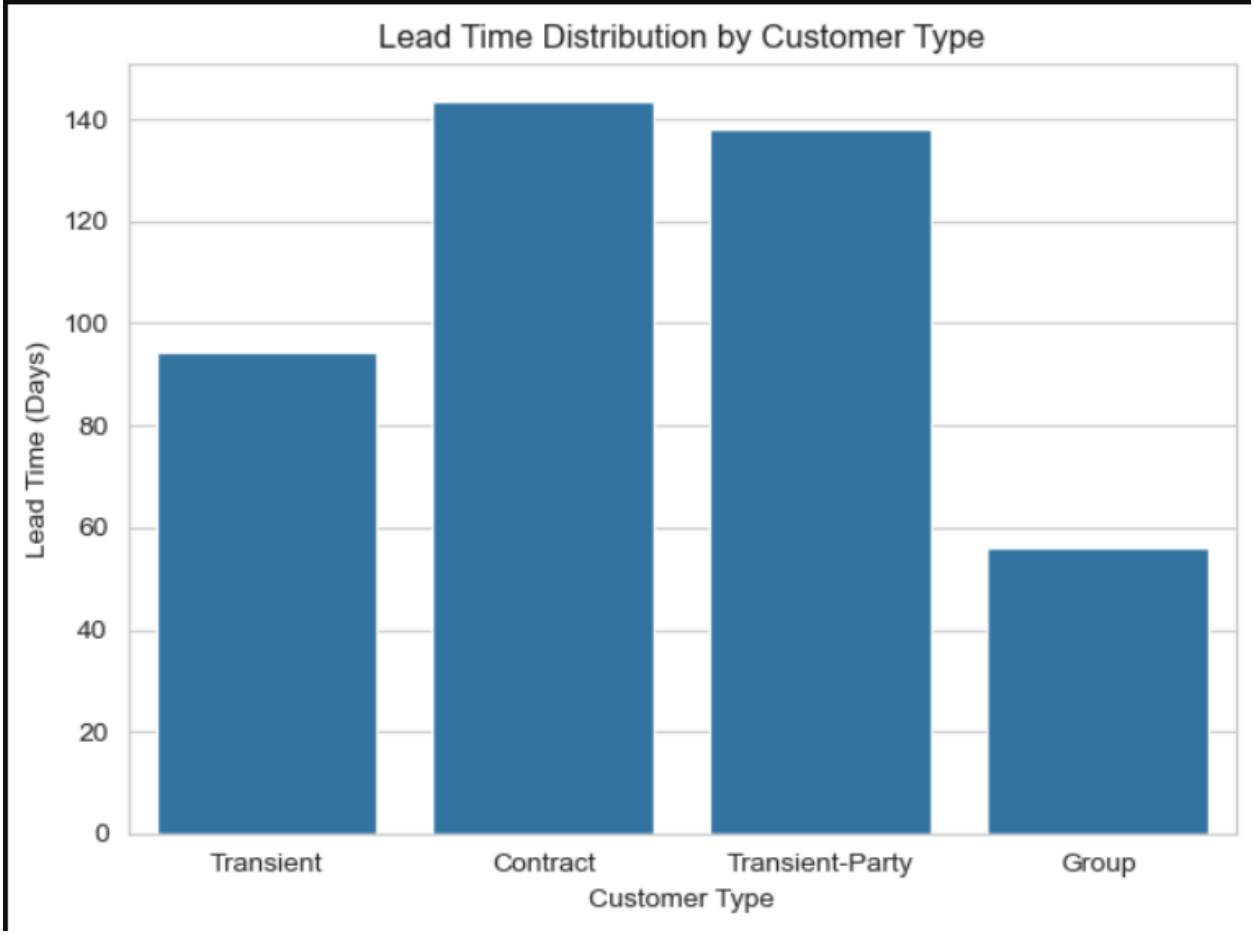
Do certain market segments or distribution channels show higher booking consistency or revenue?: The unstacked table of average ADR by market segment and distribution channel shows variations in revenue across different combinations, indicating some combinations yield higher average ADR. 'Transient' customers have the highest average ADR.



Do bookings with more lead time or from specific countries yield higher ADR?: The scatter plot of lead-time vs. ADR shows a weak positive trend, suggesting slightly higher ADR for bookings with more lead time. Bookings from certain countries have significantly higher average ADRs.

```
#8 How does booking lead time vary across customer types and countries?

sns.barplot(x='customer_type', y='lead_time', data=df1,errorbar= None)
plt.title('Lead Time Distribution by Customer Type')
plt.xlabel('Customer Type')
plt.ylabel('Lead Time (Days)')
plt.tight_layout()
plt.show()
```



Are guests with higher ADR more likely to request special services or make booking modifications?: Scatter plots of ADR vs. `total_of_special_requests` and ADR vs. `booking_changes` suggest a weak positive relationship, indicating that guests with higher ADR tend to request slightly more special services and make a few more booking changes.

5. Conclusion and Recommendations (Inferred and Consolidated)

The thorough cleaning, preprocessing, and exploratory analysis of the hotel booking dataset provide valuable insights for strategic decision-making:

- **Enhanced Data Quality**: By meticulously handling missing values, removing duplicates,

and addressing outliers in adr, the dataset is now robust and reliable for further modeling or business intelligence.

- **Understanding Lead Time Dynamics**: The dominance of short lead times highlights an opportunity for targeted marketing campaigns for last-minute bookings. Conversely, understanding longer lead times for certain segments could inform early-bird offers.
- **Impact of ADR on Special Requests**: The weak positive correlation between adr and total_of_special_requests suggests that while higher-paying guests might slightly more often request special services, this relationship is not strong enough to be a primary driver for offering premium services solely based on adr.
- **Booking Modifications and ADR**: The negligible correlation between adr and booking_changes indicates that the pricing strategy does not significantly influence a guest's propensity to modify their booking.
- **Stay Duration Analysis**: The total_nights_stays feature provides a clearer picture of stay durations, which is crucial for optimizing room allocation, staff scheduling, and identifying trends for short-term vs. long-term guests.

This report summarizes the key data processing and initial analytical steps, offering foundational insights into the hotel booking patterns. These insights can be further expanded upon for more advanced predictive modeling or detailed business strategy formulation.
