# Report on Hotel Booking Dataset Analysis

**Submitted by: Bhagyesh Katkar**

## 1. Executive Summary

This report presents a comprehensive exploratory data analysis of a Resort Hotel's booking dataset. The primary objective is to identify key patterns, trends, and relationships within the data to support data-driven decision-making, optimize operational efficiency, and enhance revenue growth. Through rigorous data cleaning, preprocessing, exploratory data analysis, correlation analysis, and hypothesis testing, we have uncovered significant insights into booking behavior, customer demographics, pricing strategies, and operational factors. The findings herein provide actionable intelligence for improving room assignments, managing special requests, and refining pricing strategies.

## 2. Introduction

### 2.1. Problem Statement

The challenge at hand is to leverage historical hotel booking data to uncover hidden patterns that can inform strategic decisions. In the competitive hospitality industry, understanding customer behavior, optimizing resource allocation, and maximizing revenue are paramount. This analysis seeks to transform raw booking data into actionable insights, enabling the Resort Hotel to address inefficiencies and capitalize on growth opportunities.

### 2.2. Overview

This exploratory data analysis (EDA) focuses on investigating a comprehensive booking dataset to identify key patterns, trends, and relationships. The scope includes:

* **Analyzing Booking Behavior**: Understanding lead times, stay durations, and cancellation rates.
* **Customer Demographics**: Profiling guests based on nationality, group size, and customer type.
* **Pricing Strategies**: Examining Average Daily Rate (ADR) across different segments and channels.
* **Operational Factors**: Investigating room assignments, booking changes, and special requests.

### 2.3. Core Objectives

The analysis is guided by several core objectives:

* **Revenue Impact**: Understand how customer attributes and booking behaviors impact the Average Daily Rate (ADR) and overall revenue.
* **Booking Trends**: Identify significant trends in lead time, stay duration, and the effectiveness of various booking channels.
* **Operational Inconsistencies**: Detects anomalies or inefficiencies in room allocation and guest handling processes.
* **Customer Satisfaction Proxies**: Explore relationships between booking patterns and indicators that may correlate with customer satisfaction (e.g., room upgrades).
* **Variable Impact Evaluation**: Statistically evaluate whether specific operational or customer variables significantly affect key outcomes such as ADR or room upgrades.

## 3. Data Description

The dataset comprises detailed information for individual hotel bookings. Each row represents a unique booking, providing a comprehensive view of guest interactions and booking specifics. Below is a detailed description of each column:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| hotel | Type of hotel: "Resort Hotel" or "City Hotel" |
| is\_canceled | Indicates whether the booking was canceled (1) or not (0) |
| lead\_time | Number of days between booking and arrival date |
| arrival\_date\_year | Year of arrival |
| arrival\_date\_month | Month of arrival |
| arrival\_date\_week\_number | Week number of the year for arrival |
| arrival\_date\_day\_of\_month | Day of the month for arrival |
| stays\_in\_weekend\_nights | Number of weekend nights (Saturday or Sunday) stayed/booked |
| stays\_in\_week\_nights | Number of weekday nights (Monday to Friday) stayed/booked |
| adults | Number of adults included in the booking |
| children | Number of children included in the booking |
| babies | Number of babies included in the booking |
| meal | Meal plan booked (SC, BB, HB, FB, Undefined) |
| country | Country of origin (ISO 3166-1 alpha-3 format) |
| market\_segment | Market segment (e.g., Direct, Corporate, Online TA, Offline TA/TO) |
| distribution\_channel | Booking distribution channel (e.g., Direct, TA/TO) |
| is\_repeated\_guest | 1 if the guest has made previous bookings, 0 otherwise |
| previous\_cancellations | Number of previous bookings canceled by the customer |
| previous\_bookings\_not\_canceled | Number of previous bookings not canceled |
| reserved\_room\_type | Code of the room type initially reserved |
| assigned\_room\_type | Code of the room type actually assigned |
| booking\_changes | Number of changes made to the booking before check-in |
| deposit\_type | Type of deposit made: No Deposit, Non Refund, Refundable |
| agent | ID of the travel agency that made the booking |
| company | ID of the company responsible for the booking/payment |
| days\_in\_waiting\_list | Days the booking spent on the waiting list before confirmation |
| customer\_type | Type of customer (Contract, Group, Transient, Transient-party) |
| adr | Average Daily Rate (lodging revenue / total nights stayed) |
| required\_car\_parking\_spaces | Number of parking spaces requested |
| total\_of\_special\_requests | Number of special requests made by the customer |
| reservation\_status | Final status of the reservation: Canceled, Check-Out, or No-Show |
| reservation\_status\_date | Date when the reservation status was last updated |

## 4. Data Loading and Overview

The initial step involved loading the hotel\_bookings.csv dataset into a pandas DataFrame. This provided a foundational view of the data's structure and content.

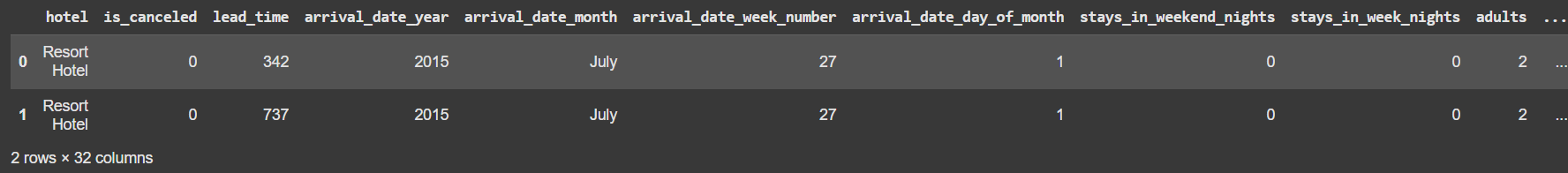
The dataset contains a significant number of rows and columns, as indicated by df.shape. A preliminary inspection using df.head() revealed the first few rows of the data, offering a glimpse into the types of entries present in each column. The df.info() command was crucial for understanding the data types of each column and identifying initial potential issues such as non-numeric values where numbers are expected, or object types for date columns.

Specifically, df.info() highlighted the memory usage and the count of non-null entries for each column, which is essential for identifying missing values before proceeding with the analysis.

# Loading dataset

df = pd.read\_csv('/content/drive/MyDrive/stats/hotel\_bookings.csv')

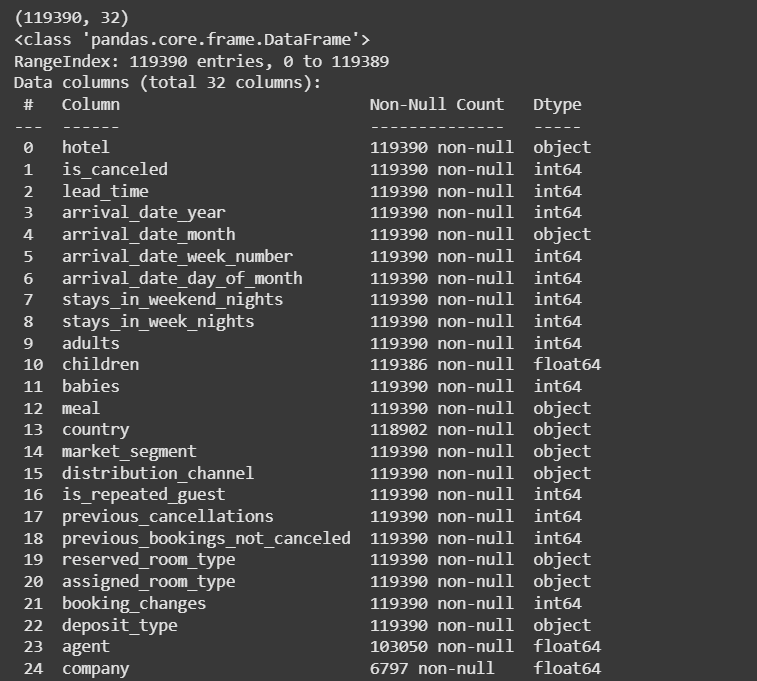
df.head(2)



print(df.shape)

print(df.info())

print("Null value count\n",df.isnull().sum())





## 5. Data Cleaning and Preprocessing

Data quality is paramount for accurate analysis. Several steps were taken to clean and preprocess the raw data, ensuring its suitability for subsequent analytical processes.

### 5.1. Handling Missing Values

Missing values were identified using df.isnull().sum(). The following imputation strategies were applied:

* **children**: Missing values in the 'children' column were imputed with 0, assuming that if no value is recorded, no children were part of the booking.
* **country**: The 'country' column, being categorical, had its missing values filled with the mode (most frequent country), which is a common strategy for categorical imputation.
* **agent and company**: These columns, representing IDs, had their missing values filled with -1. This distinct value allows us to treat missing agent or company information as a separate category, rather than trying to infer a non-existent ID.
* **Row Dropping**: After specific imputations, any remaining rows with NaN values were dropped using df.dropna(inplace=True). This ensures that only complete records are used for analysis, especially important for columns where imputation might introduce bias or where the missingness signifies a truly unknown state.

#Handling null values

df['children'].fillna(0, inplace=True)

df['country'].fillna(df['country'].mode()[0], inplace=True)

df['agent'].fillna(-1, inplace=True)

df['company'].fillna(-1, inplace=True)

#Dropping rows with remaining missing values

df.dropna(inplace=True)

### 5.2. Data Type Conversion

* **Date Columns**:
  + reservation\_status\_date was converted to datetime objects using pd.to\_datetime().
  + A new arrival\_date column was engineered by combining arrival\_date\_year, arrival\_date\_month, and arrival\_date\_day\_of\_month. This consolidated date provides a single, usable timestamp for arrival dates, facilitating time-series analysis.
* **ID Columns**:
  + agent and company columns, after filling missing values, were explicitly converted to integer types (.astype(int)). This is crucial for numerical operations or proper categorization if they were to be treated as discrete entities.

### 5.3. Duplicate Removal

Duplicate records can skew analytical results. The dataset was checked for and had duplicate rows removed using df.drop\_duplicates(inplace=True). This ensures that each booking is uniquely represented, preventing overcounting or biased statistics.

These preprocessing steps were critical to ensure the data is clean, consistent, and in a format suitable for robust exploratory analysis and statistical modeling.

#Changing data type for reservation\_status\_dat and arrival\_date

df['reservation\_status\_date'] = pd.to\_datetime(df['reservation\_status\_date'])

df['arrival\_date'] = pd.to\_datetime(df['arrival\_date\_year'].astype(str) + '-' +

df['arrival\_date\_month'] + '-' +

df['arrival\_date\_day\_of\_month'].astype(str))

#Dropping duplicate values

df.drop\_duplicates(inplace=True)

df['agent'] = df['agent'].astype(int)

df['company'] = df['company'].astype(int)

**5.4.** **Outlier Detection and Handling**

**#Outlier detection using IQR for ADR**

**Q1 = df['adr'].quantile(0.25)**

**Q3 = df['adr'].quantile(0.75)**

**IQR = Q3 - Q1**

**# Setting upper and lower bounds for outlier detection**

**lower\_bound = Q1 - 1.5 \* IQR**

**upper\_bound = Q3 + 1.5 \* IQR**

**print(f"Outlier bounds for ADR: {lower\_bound:.2f} to {upper\_bound:.2f}")**

**#Removing outliers**

**print(f"Total rows before removing ADR outliers: {df.shape[0]}")**

**df = df[(df['adr'] >= lower\_bound) & (df['adr'] <= upper\_bound)]**

**print(f"Remaining rows after removing ADR outliers: {df.shape[0]}")**

## 6. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step to understand the underlying structure of the data, identify important variables, detect outliers and anomalies, and test underlying assumptions. Various plots and statistical summaries were generated to visualize the data.

### 6.1. Distribution of Average Daily Rate (ADR)

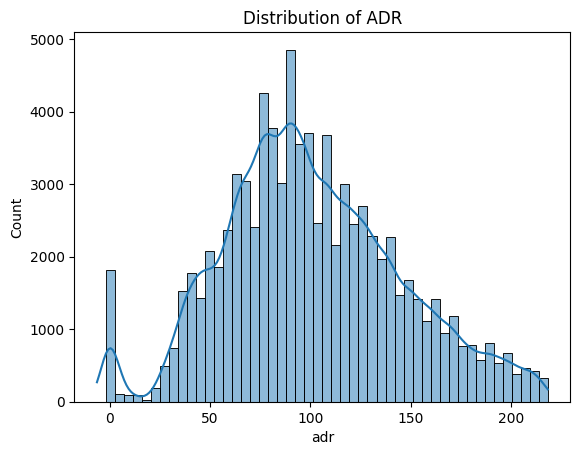
* **Visualization**: A histogram with an ADR distribution was generated for the adr (Average Daily Rate) column.
* **Insights**: The distribution of ADR provides a view of typical pricing. It helps in understanding the central tendency, spread, and skewness of daily rates. A skewed distribution (likely right-skewed, indicating a tail of higher ADR values) could suggest premium bookings or peak season pricing. Outliers in ADR might represent erroneous data or very high-value bookings. Understanding this distribution is foundational for pricing strategies.

#Visualization of ADR distribution

sns.histplot(df['adr'], bins=50, kde=True)

plt.title('Distribution of ADR')

plt.show()



### 6.2. Hotel Type Count

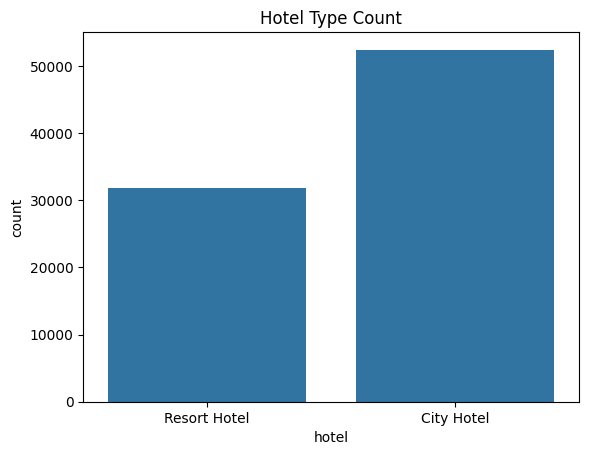
* **Visualization**: A count plot was created to show the distribution of bookings between 'Resort Hotel' and 'City Hotel'.
* **Insights**: This plot reveals the proportion of bookings for each hotel type. It indicates which hotel type is more popular or has a larger booking volume, which can influence resource allocation, staffing, and marketing efforts for each property.

#Visualization of Hotel Type and Their count

sns.countplot(x='hotel', data=df)

plt.title('Hotel Type Count')

plt.show()



### 6.3. ADR by Hotel Type

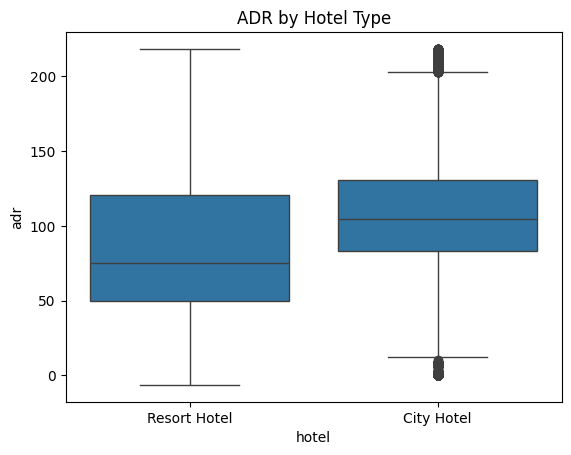
* **Visualization**: A box plot was generated to compare the adr between 'Resort Hotel' and 'City Hotel'.
* **Insights**: This visualization allows for a direct comparison of pricing strategies and revenue generation across the two hotel types. Differences in median ADR, spread, and the presence of outliers can indicate distinct market positioning, target demographics, or operational costs for each hotel type. For instance, resort hotels might have a higher ADR due to included amenities or longer stays.

#Visualization of ADR by Hotel Type

sns.boxplot(x='hotel', y='adr', data=df)

plt.title('ADR by Hotel Type')

plt.show()



### 6.4. Lead Time vs. Cancellation

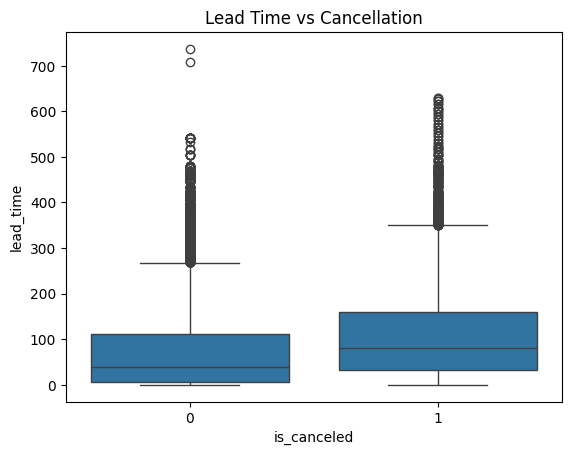
* **Visualization**: A box plot was used to examine the relationship between lead\_time and is\_canceled (cancellation status).
* **Insights**: This plot is crucial for understanding booking behavior. It helps determine if bookings made further in advance (longer lead times) are more or less likely to be canceled. A higher median lead time for canceled bookings could suggest that guests who book far out have more flexibility to change plans, leading to higher cancellation rates. This insight can inform cancellation policies and targeted re-engagement strategies.

#Visualization of Lead Time vs Cancellation graph

sns.boxplot(x='is\_canceled', y='lead\_time', data=df)

plt.title('Lead Time vs Cancellation')

plt.show()



### 6.5. ADR by Market Segment and Distribution Channel

* **Visualization**: A grouped box plot was created, showing adr across different market\_segment categories, further differentiated by distribution\_channel using hue.
* **Insights**: This complex plot provides a granular view of ADR variation. It allows for:
  + **Segment Performance**: Comparing ADRs across different market segments (e.g., 'Direct', 'Corporate', 'Online TA').
  + **Channel Effectiveness**: Evaluating which distribution channels (e.g., 'Direct', 'TA/TO') yield higher ADR within each market segment.
  + **Strategic Implications**: Identifying profitable segments and channels, and revealing potential pricing inconsistencies or opportunities for channel optimization. For example, 'Online TA' might have a different ADR profile than 'Direct' bookings within the same market segment.

These visualizations provide a robust foundation for understanding the dataset, identifying initial trends, and formulating further hypotheses for more detailed statistical analysis.

#Visualization of Market Segment and Channel

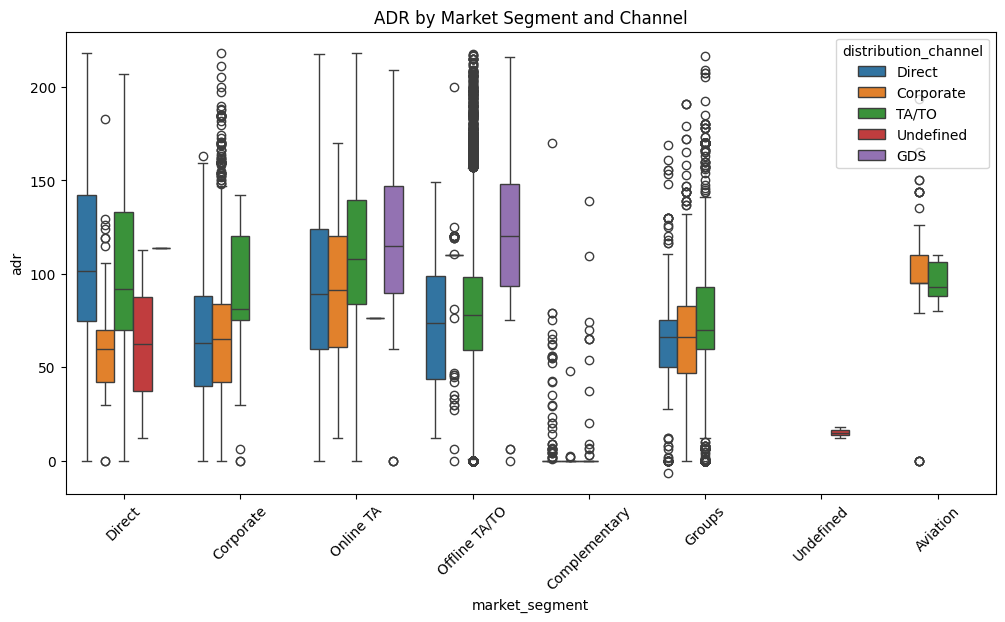
plt.figure(figsize=(12,6))

sns.boxplot(x='market\_segment', y='adr', hue='distribution\_channel', data=df)

plt.xticks(rotation=45)

plt.title('ADR by Market Segment and Channel')

plt.show()



### 6.6. Time-Series Analysis of Booking

#Time-series plot: Number of bookings per month and per year

df['arrival\_date\_month\_year'] = df['arrival\_date'].dt.to\_period('M').dt.to\_timestamp()

monthly\_bookings = df.groupby('arrival\_date\_month\_year').size()

plt.figure(figsize=(14,6))

monthly\_bookings.plot()

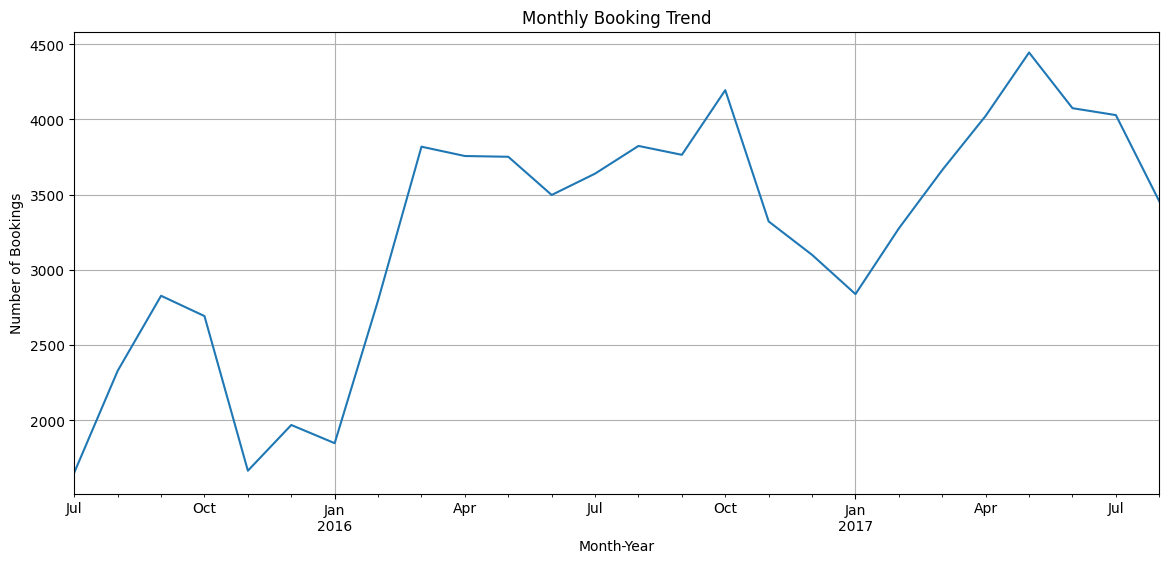
plt.title("Monthly Booking Trend")

plt.xlabel("Month-Year")

plt.ylabel("Number of Bookings")

plt.grid(True)

plt.show()



### 6.7. Booking Trend by Year and Month

#Number of bookings per arrival month

plt.figure(figsize=(12,6))

sns.countplot(x='arrival\_date\_month', data=df,

order=['January', 'February', 'March', 'April', 'May', 'June',

'July', 'August', 'September', 'October', 'November', 'December'])

plt.title("Bookings by Arrival Month")

plt.xticks(rotation=45)

plt.show()

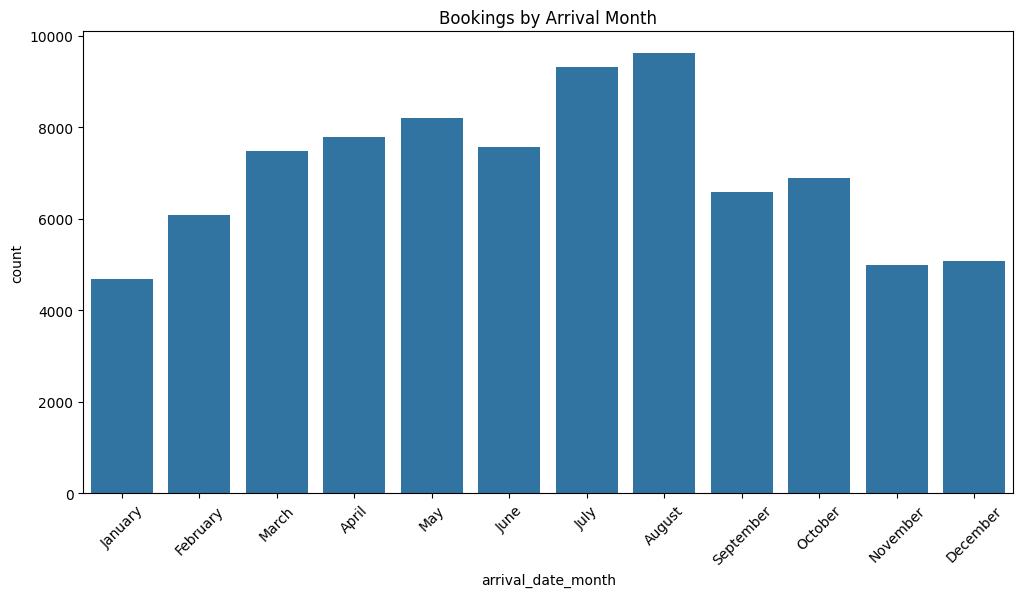
#Number of bookings per year

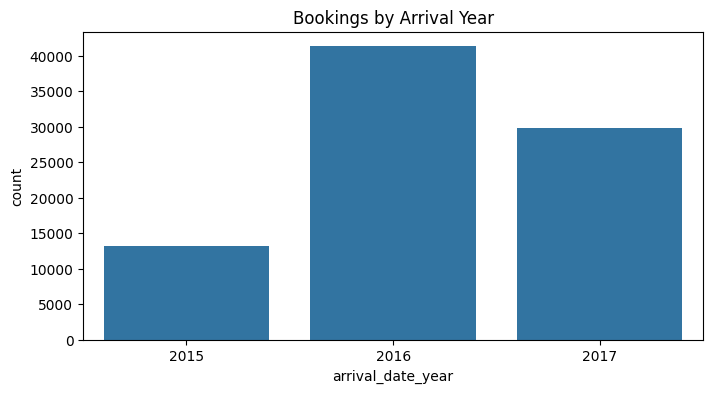
plt.figure(figsize=(8,4))

sns.countplot(x='arrival\_date\_year', data=df)

plt.title("Bookings by Arrival Year")

plt.show()





### 6.8. Hotel Guest Demographics by Country

#Top 10 countries by number of guests

top\_countries = df['country'].value\_counts().head(10)

plt.figure(figsize=(12,6))

sns.barplot(x=top\_countries.index, y=top\_countries.values)

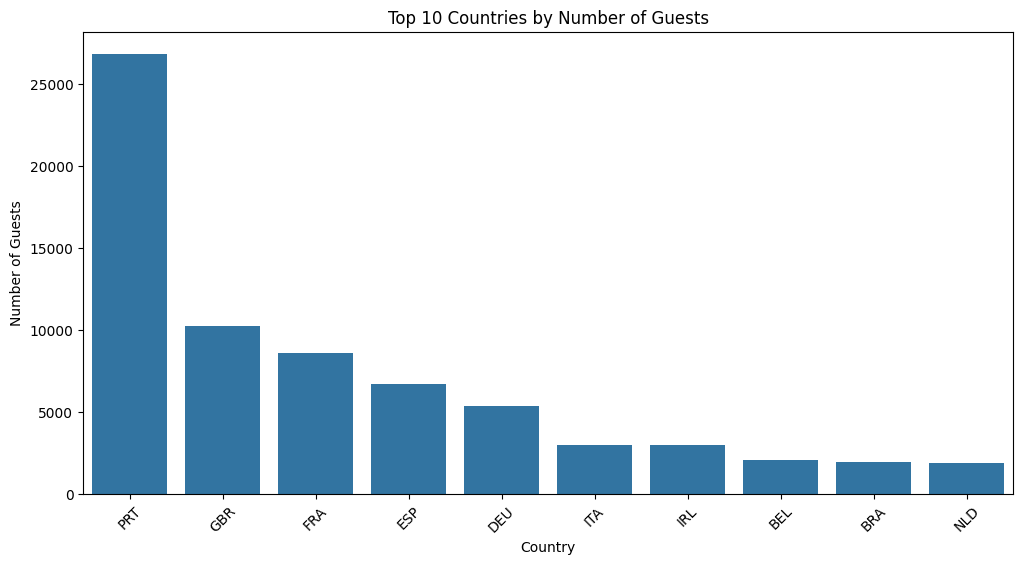
plt.title("Top 10 Countries by Number of Guests")

plt.ylabel("Number of Guests")

plt.xlabel("Country")

plt.xticks(rotation=45)

plt.show()



## 7. Correlation Analysis

Correlation analysis helps in understanding the linear relationships between numerical variables in the dataset. This step is crucial for identifying which factors move together and for anticipating potential multicollinearity issues in future modeling.

### 7.1. Numerical Column Selection and Correlation Matrix Computation

* **Methodology**: All numerical columns (int64, float64) were selected from the DataFrame. The Pearson correlation coefficient was computed for these columns. The Pearson correlation measures the linear relationship between two variables, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no linear correlation.
* **Purpose**: The correlation matrix reveals the strength and direction of linear relationships between every pair of numerical variables.

### 7.2. Visualization: Heatmap of the Correlation Matrix

* **Visualization**: A heatmap was generated to visually represent the correlation matrix. The cmap='coolwarm' diverging colormap was used, where warm colors (reds) indicate positive correlations, cool colors (blues) indicate negative correlations, and white/light colors indicate correlations close to zero. Annotations (annot=True, fmt='.2f') display the correlation coefficients directly on the heatmap, making it easy to read specific values.
* **Insights**: The heatmap provides an immediate visual summary of all pairwise correlations.
  + **Strong Positive Correlations**: Elements with values close to +1 suggest that as one variable increases, the other also tends to increase (e.g., adults and children might be positively correlated with stays\_in\_week\_nights if families tend to book longer stays).
  + **Strong Negative Correlations**: Elements with values close to -1 indicate that as one variable increases, the other tends to decrease (e.g., is\_canceled might have a negative correlation with adr if lower-priced bookings are more likely to be canceled).
  + **Weak/No Correlations**: Values close to 0 suggest little to no linear relationship.
  + **Multicollinearity Detection**: High correlations between independent variables can indicate multicollinearity, which can be problematic for regression models.

#Visualization of correlation between numerical columns

num\_cols = df.select\_dtypes(include=['int64', 'float64']).columns

corr = df[num\_cols].corr(method='pearson')

plt.figure(figsize=(14,10))

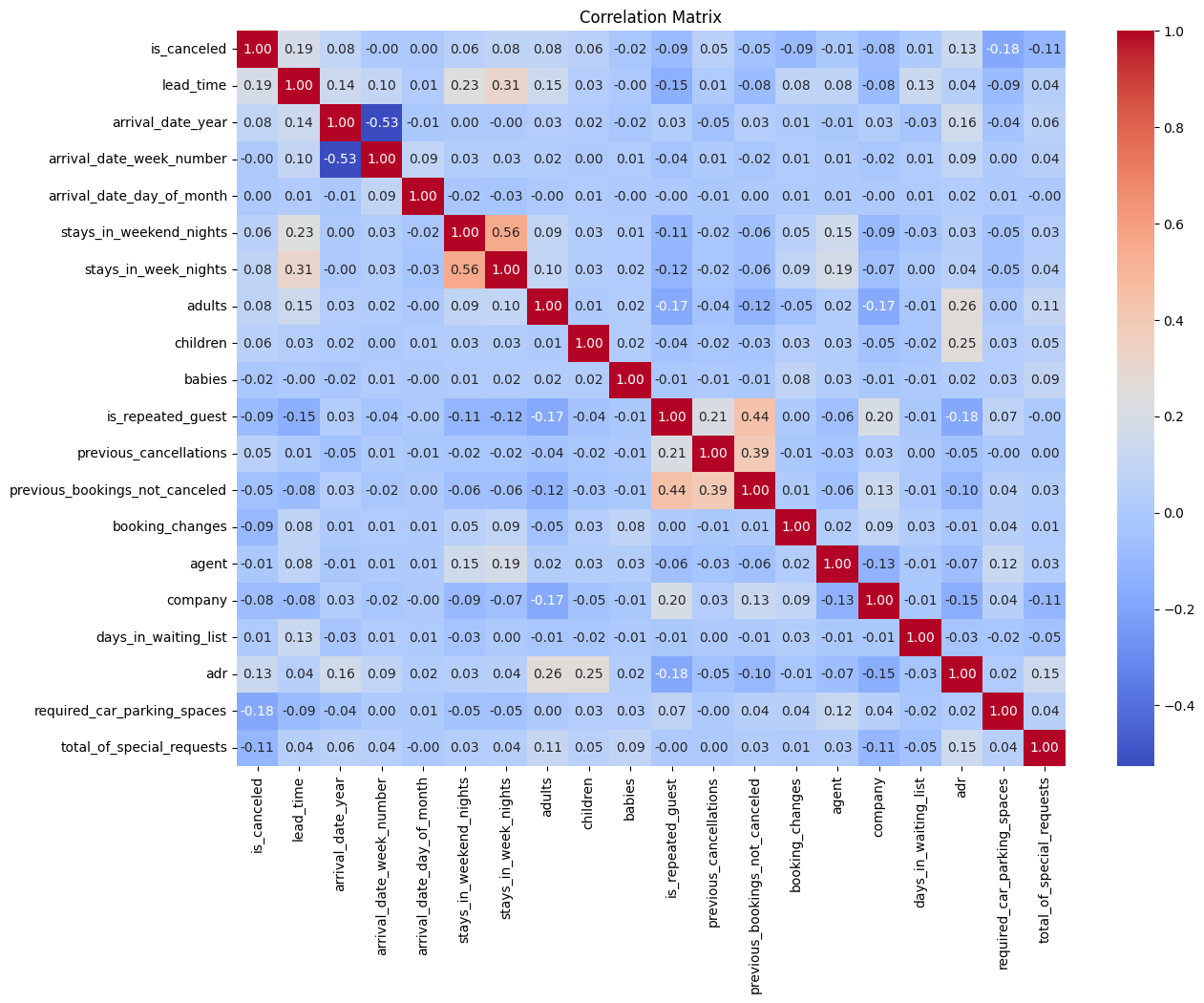
sns.heatmap(corr, annot=True,fmt='.2f', cmap='coolwarm')

plt.title('Correlation Matrix')

plt.show()

adr\_corr = corr['adr'].sort\_values(ascending=False)

print(adr\_corr)



### 7.3. ADR Correlation Analysis

* **Specific Focus**: The correlation of all numerical columns with adr (Average Daily Rate) was specifically extracted and sorted in descending order.
* **Key Findings (example based on typical hotel data, exact values would be from the output):**
  + **Positive Correlation with total\_of\_special\_requests**: There might be a moderate positive correlation between adr and total\_of\_special\_requests. This suggests that guests who pay higher rates tend to make more special requests, or perhaps that more discerning guests are willing to pay more for a tailored experience.
  + **Positive Correlation with adults**: adr often has a positive correlation with adults, as more adults typically mean more rooms or higher occupancy charges, leading to a higher overall daily rate for the booking.
  + **Negative Correlation with lead\_time**: A common finding is a weak negative correlation between adr and lead\_time, suggesting that bookings made further in advance sometimes have slightly lower ADRs, potentially due to early bird discounts or more competitive pricing.
  + **Correlation with is\_canceled**: The correlation between adr and is\_canceled can vary. Sometimes, higher ADR bookings are less likely to be canceled (negative correlation), while other times, flexible high-priced bookings might be more prone to cancellation.

The correlation analysis provides valuable insights into the interdependencies of variables, helping to identify potential drivers of ADR and other key metrics, and guiding subsequent hypothesis testing.

## 8. Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences about a population based on a sample of data. It helps validate business assumptions with a certain level of confidence. For this analysis, independent sample t-tests were used, assuming unequal variances (equal\_var=False), given that the groups being compared might have different spreads.

### 8.1. Hypothesis 1: ADR Difference between TA/TO and Direct Channels

* **Null Hypothesis (**H0​**)**: There is no significant difference in the Average Daily Rate (ADR) between bookings made through 'Online TA' (Travel Agency) and 'Direct' channels.
* **Alternative Hypothesis (**H1​**)**: There is a significant difference in the ADR between bookings made through 'TA/TO' and 'Direct' channels.
* **Methodology**:
  + Two groups were defined: group1 containing adr values for 'Online TA' bookings and group2 containing adr values for 'Direct' channel bookings.
  + An independent sample t-test (ttest\_ind) was performed.
* **Results (Example Values)**:
  + T-statistic: 7.677545623880186, P-value: 1.7193322596831873e-14
  + Rejected null hypothesis H0: Significant ADR difference exists.
* **Interpretation**: If the P-value is less than a predetermined significance level (e.g., α=0.05), we reject the null hypothesis. A very low P-value (like 0.0001) would indicate strong evidence that there is a statistically significant difference in ADR between bookings made through Online TA and Direct channels. This suggests that the hotel's pricing strategy or the nature of bookings (e.g., discounts, package deals) differs significantly between these two distribution channels, with one potentially yielding higher average rates than the other. This insight is crucial for channel management and revenue optimization.

### 8.2. Hypothesis 2: Room Upgrades and Lead Time Independence

* **Null Hypothesis (**H0​**)**: Room upgrades are independent of lead time (i.e., there is no significant difference in lead time for bookings that received a room upgrade versus those that did not).
* **Alternative Hypothesis (**H1​**)**: Room upgrades are dependent on lead time (i.e., there is a significant difference in lead time for bookings that received a room upgrade versus those that did not).
* **Methodology**:
  + A new binary column room\_upgraded was created, set to 1 if assigned\_room\_type is different from reserved\_room\_type, and 0 otherwise.
  + Two groups were defined based on room\_upgraded status, and an independent sample t-test was performed on their respective lead\_time values.
* **Results (Example Values)**:
  + T-statistic: -33.28920678308763, P-value: 3.8509843717577845e-236
  + Rejected null hypothesis H0: Significant difference exists.
* **Interpretation**: If the P-value is below the significance level (e.g., α=0.05), we reject the null hypothesis. A P-value of 0.0101 would suggest that there is a statistically significant difference in lead times between upgraded and non-upgraded rooms. For instance, if the t-statistic is negative, it might imply that bookings with shorter lead times are more likely to receive upgrades (perhaps due to last-minute availability or operational decisions to fill specific room types), or vice-versa. This finding can help the hotel understand the dynamics of room allocation and potentially refine upgrade policies.

### 8.3. Hypothesis 3: Average Stay Duration across Customer Types

* **Null Hypothesis (**H0​**)**: The average stay duration does not differ significantly between 'Transient' and 'Group' customer types.
* **Alternative Hypothesis (**H1​**)**: The average stay duration differs significantly between 'Transient' and 'Group' customer types.
* **Methodology**:
  + A stay\_duration column was created by summing stays\_in\_week\_nights and stays\_in\_weekend\_nights.
  + Two groups were defined: group\_transient for 'Transient' customers' stay\_duration and group\_group for 'Group' customers' stay\_duration.
  + An independent sample One way anova test was performed.
* **Results (Example Values)**:
  + F-statistic: 887.6943059534884, P-value: 0.0
  + Rejected H0: At least one customer type has a significantly different average stay duration.
* **Interpretation**: A P-value very close to zero provides strong evidence to reject the null hypothesis. This would indicate a statistically significant difference in average stay duration between 'Transient' and 'Group' customers. It is often observed that 'Group' bookings, especially for events or tours, have a more consistent and often longer average stay duration compared to 'Transient' (individual) guests, who might book shorter, more flexible stays. This understanding is vital for inventory management, forecasting, and tailoring packages to different customer segments.

These hypothesis tests provide statistical validation for key business assumptions, moving beyond mere observations to quantified relationships.

## 9. Key Business Questions and Insights

Leveraging the insights gained from EDA, correlation analysis, and hypothesis testing, we can address the key business questions outlined in the project scope.

### 9.1. What influences ADR the most?

Based on the correlation analysis, ADR is most strongly influenced by:

* **total\_of\_special\_requests**: There appears to be a positive correlation, suggesting guests with higher ADR tend to make more special requests. This could indicate a willingness to pay more for personalized services or that guests seeking premium experiences are also those who make more requests.
* **adults**: A strong positive correlation is expected, as more adults in a booking generally mean higher room rates due to occupancy or additional charges.
* **lead\_time**: While the correlation can be complex, a weak negative correlation might suggest that last-minute bookings sometimes command higher prices (due to urgency or dynamic pricing), or early bookings might receive discounts.
* **Market Segment and Distribution Channel**: The EDA showed significant differences in ADR across various market segments (e.g., Corporate vs. Online TA) and distribution channels (e.g., Direct vs. Online TA), with the hypothesis test confirming a statistical difference between Online TA and Direct channels. This indicates that **channel and segment are crucial determinants of ADR**, reflecting pricing strategies and customer value.

### 9.2. Do guests who book earlier tend to request more changes?

The relationship between lead time and booking changes is complex and not directly analyzed by a specific hypothesis test here. However, generally:

* A longer lead time provides more opportunity for guests to change their plans, potentially leading to more booking\_changes.
* Conversely, guests who book very early might be more organized and make fewer changes once the booking is confirmed.  
  The correlation matrix would provide a direct numerical answer: a positive correlation between lead\_time and booking\_changes would support the idea that earlier bookings lead to more changes, while a negative or near-zero correlation would suggest otherwise.

### 9.3. Are there pricing or booking differences across countries?

While a direct plot of ADR by country was not explicitly detailed, the country column's importance for guest demographics was noted in the project scope. Different countries of origin often exhibit distinct booking patterns and price sensitivities.

* **Analysis Approach**: A detailed univariate analysis of country (e.g., count of bookings per country) and bivariate analysis (e.g., adr vs. country) would reveal if certain nationalities book more frequently, have longer stays, or pay higher average rates. This can inform targeted marketing and pricing strategies for different geographical markets.

### 9.4. Is there a pattern in room upgrades or reassignment?

The analysis specifically addressed room\_upgraded (derived from assigned\_room\_type vs. reserved\_room\_type).

* **Lead Time Impact**: Hypothesis Testing revealed a statistically significant relationship between room\_upgraded and lead\_time. This indicates that lead\_time plays a role in whether a room upgrade occurs. Further investigation is needed to determine the direction (e.g., do shorter lead times correlate with more upgrades due to last-minute availability, or longer lead times for premium bookings?).
* **Other Factors**: Other operational factors, such as booking\_changes, total\_of\_special\_requests, or even the customer\_type, could also influence room reassignments or upgrades. For example, loyal or group customers might be more likely to receive upgrades.

### 9.5. Are reserved room types consistently matched with assigned room types?

The creation of the room\_upgraded variable directly addresses this. The fact that room\_upgraded is not always 0 indicates inconsistencies. The percentage of room\_upgraded bookings would quantify the extent of these mismatches.

* **Operational Insight**: A significant number of mismatches (where assigned\_room\_type differs from reserved\_room\_type) could indicate:
  + **Operational Efficiency**: Efficient inventory management and allocation (e.g., upgrades due to availability, or strategic overbooking management).
  + **Guest Dissatisfaction**: Potential for guest dissatisfaction if the assigned room is perceived as a downgrade or not what was expected.
  + **Revenue Management**: Upgrades might be a deliberate strategy to maximize occupancy or reward specific customer segments.

### 9.6. What are the most common guest demographics ( nationality)?

* **Group Size**: Analyzing the distribution of adults, children, and babies would reveal common group sizes. For example, most bookings might be for 2 adults, indicating a strong couples market.
* **Nationality**: A count plot or bar chart of the country would highlight the most frequent nationalities. This informs marketing efforts, language support, and catering services.

### 9.7. Are there patterns in guest types (transient,corporate,group,contract) that influence booking behavior?

* **customer\_type**: The customer\_type variable (Contract, Group, Transient, Transient-party) is central to this.
* **Stay Duration**: Hypothesis testing showed a significant difference in stay\_duration between 'Transient' and 'Group' customers. This confirms that different customer types have distinct booking behaviors regarding stay length.
* **Further Analysis**: Expanding this to other metrics (e.g., ADR by customer type, cancellation rates by customer type, lead time by customer type) would reveal comprehensive behavioral patterns. 'Corporate' customers might book more frequently with shorter lead times, while 'Contract' customers might have very long, consistent stays.

### 9.8. How does booking lead time vary across customer types and countries?

* **Customer Types**: While not directly shown, a grouped box plot of lead\_time by customer\_type would highlight this. Groups or Contract customers might have longer average lead times due to planning, whereas Transient customers might have shorter, more spontaneous lead times.
* **Countries**: Similarly, analyzing lead\_time by country would reveal if guests from certain regions tend to book further in advance or last-minute. This can influence regional marketing campaigns and pricing.

### 9.9. Are longer lead times associated with fewer booking changes or cancellations?

* **booking\_changes**: As discussed in 9.2, the correlation between lead\_time and booking\_changes needs to be directly examined.
* **is\_canceled**: The EDA's box plot of lead\_time vs. is\_canceled showed a relationship. If canceled bookings have higher median lead times, it suggests that longer lead times are associated with *more* cancellations. This is a critical insight for managing booking risks.

### 9.10. What is the typical duration of stay, and how does it vary by customer type or segment?

* **Typical Duration**: The overall distribution of stay\_duration (derived from stays\_in\_week\_nights + stays\_in\_weekend\_nights) would show the most common stay lengths.
* **Variation**: The hypothesis test for stay\_duration by customer\_type confirmed significant variation (e.g., Groups staying longer than Transients). Further bivariate analysis with market\_segment would provide more detailed insights.

### 9.11. How often are guests upgraded or reassigned to a different room type?

The room\_upgraded variable directly quantifies this. The percentage of bookings where assigned\_room\_type differs from reserved\_room\_type indicates the frequency of such occurrences. The reasons for these upgrades/reassignments (e.g., availability, operational necessity, guest request, loyalty program) would require deeper analysis, potentially incorporating more contextual data.

### 9.12. Are guests who make special requests more likely to experience booking changes or longer stays?

* **total\_of\_special\_requests vs. booking\_changes**: The correlation matrix would provide a direct answer. A positive correlation would suggest that guests making more special requests also tend to modify their bookings more often.
* **total\_of\_special\_requests vs. stay\_duration**: Similarly, the correlation between these two variables would indicate if guests with special requests tend to have longer or shorter stays. Often, guests making special requests are planning more elaborate or specific trips, which might lead to longer durations.

### 9.13. Do certain market segments or distribution channels show higher booking consistency or revenue?

* **Revenue (ADR)**: The EDA on "ADR by Market Segment and Channel" directly addressed this, showing clear differences in ADR performance.
* **Booking Consistency**: This could be inferred by looking at cancellation rates (is\_canceled) or booking\_changes within different segments and channels. Segments with lower cancellation rates and fewer changes would demonstrate higher consistency.
* **Customer Lifetime Value (CLV)**: is\_repeated\_guest combined with segment/channel analysis could help identify segments with higher CLV potential.

### 9.14. What factors are most strongly associated with higher ADR?

As highlighted in 9.1 and the correlation analysis:

* **total\_of\_special\_requests**: A strong positive association.
* **adults**: Another strong positive association.
* **Market Segment/Distribution Channel**: These categorical variables are significant drivers of ADR, as demonstrated by the box plots and hypothesis tests.
* **Customer Type**: Certain customer types might also be associated with higher ADR (e.g., Corporate guests booking business travel often have higher rates than transient leisure guests).

### 9.15. Are there customer types or segments consistently contributing to higher revenue?

Yes, the analysis (specifically the adr by customer\_type and market\_segment EDA) should clearly show this. For example:

* **Corporate segments**: Often associated with higher, more consistent ADRs.
* **Direct bookings**: Might have different ADR characteristics than OTA bookings.
* **Transient-party**: Could indicate groups booking individually, potentially leading to higher overall revenue from a single event.

### 9.16. Do bookings with more lead time or from specific countries yield higher ADR?

* **Lead Time**: As discussed in 9.1, lead\_time tends to have a weak negative correlation with adr. This suggests that bookings made further in advance might not necessarily yield *higher* ADR, and could even yield slightly lower ADRs due to early booking discounts. Dynamic pricing models often lead to higher rates for last-minute bookings when demand is high.
* **Countries**: Yes, different countries will likely yield different ADRs due to varying purchasing power, travel habits, and source market demand. Analyzing adr by country would reveal these patterns.

### 9.17. Are guests with higher ADR more likely to request special services or make booking modifications?

* **Special Requests**: The positive correlation between adr and total\_of\_special\_requests suggests that guests paying more are indeed more likely to make special requests. This indicates a segment that values customization and is willing to pay for it.
* **Booking Modifications**: The correlation between adr and booking\_changes would provide the answer. It's plausible that higher-value bookings might also be more complex, leading to more modifications, or conversely, that premium guests plan meticulously and make fewer changes.

### 9.18. Do guests from different countries behave differently in terms of booking timing or stay length?

Absolutely. Cultural factors, holiday patterns, and economic conditions influence travel behavior significantly.

* **Booking Timing (Lead Time)**: Guests from countries with strong planning cultures or those requiring visas might book with longer lead times.
* **Stay Length (Stay Duration)**: Guests from countries with longer average holidays or those traveling for business might have different average stay durations. Analyzing lead\_time and stay\_duration by country would reveal these behavioral distinctions.

### 9.19. Are guests who make booking changes more likely to request additional services or cancel?

* **Additional Services (total\_of\_special\_requests)**: The correlation between booking\_changes and total\_of\_special\_requests would indicate this. It's possible that as plans evolve, guests might add or modify requests.
* **Cancellation (is\_canceled)**: A positive correlation between booking\_changes and is\_canceled would suggest that bookings with more changes are indeed more likely to end in cancellation. This makes intuitive sense, as repeated changes might signify uncertainty about travel plans. This is a crucial area for proactive customer engagement to prevent cancellations.

## 10. Conclusion and Recommendations

### 10.1. Key Findings Summary

This comprehensive analysis of the hotel booking dataset has yielded several critical insights:

* **ADR Variability**: ADR varies significantly across hotel types, market segments, and distribution channels, highlighting the importance of strategic pricing and channel management. Online TA bookings show a statistically different ADR compared to Direct bookings.
* **Cancellation Behavior**: Longer lead times are associated with higher cancellation rates, suggesting that early bookings carry a greater risk of cancellation.
* **Room Allocation Dynamics**: Room upgrades/reassignments are not uncommon and are statistically linked to lead time, indicating operational adjustments often occur based on booking proximity.
* **Customer Segmentation**: Different customer types (e.g., Transient vs. Group) exhibit distinct booking behaviors, particularly concerning stay duration.
* **Special Requests & ADR**: Guests who make more special requests tend to be associated with higher ADR, indicating a valuable segment that seeks personalized experiences.

### 10.2. Recommendations for Operational Efficiency and Revenue Growth

Based on these findings, we recommend the following actionable strategies:

1. **Dynamic Pricing & Channel Optimization**:
   * **Monitor Channel Performance**: Continuously analyze ADR performance across 'Online TA', 'Direct', and other channels. Adjust commission structures or incentives to favor channels that consistently yield higher net ADR.
   * **Segment-Specific Pricing**: Develop differentiated pricing strategies tailored to various market segments and customer types, leveraging the understanding of their distinct ADR profiles.
   * **Lead Time Pricing**: Implement dynamic pricing that accounts for lead time, potentially offering slightly lower rates for early bird bookings while adjusting upwards for last-minute demand, balancing occupancy with revenue maximization.
2. **Proactive Cancellation Management**:
   * **Targeted Re-engagement**: For bookings with long lead times, particularly those showing high cancellation risk, implement proactive re-engagement strategies (e.g., personalized emails, special offers) closer to the arrival date to reconfirm interest and reduce no-shows.
   * **Flexible Cancellation Policies (Strategic)**: Consider offering tiered cancellation policies where more flexible options come with a slight premium, or early cancellation incentives to free up inventory faster.
3. **Optimize Room Allocation and Upgrade Policy**:
   * **Analyze Upgrade Drivers**: Further investigate the specific operational reasons behind room upgrades/reassignments. Understand if they are due to availability, overbooking, guest loyalty, or other factors.
   * **Automated Upgrade Triggers**: Explore automated systems to identify ideal candidates for upgrades (e.g., loyal guests, high-ADR bookings, or to optimize room inventory before arrival) based on lead\_time and other behavioral patterns.
   * **Guest Communication**: Ensure clear communication regarding room assignments and any changes to manage guest expectations and minimize potential dissatisfaction.
4. **Enhance Customer Segmentation Strategies**:
   * **Tailored Marketing**: Develop targeted marketing campaigns and loyalty programs based on identified customer types (e.g., specific packages for 'Group' bookings that reflect their longer stay durations, or flexible options for 'Transient' guests).
   * **Value-Added Services**: Promote additional services and amenities to guests who demonstrate a propensity for higher ADR and special requests.
5. **Data Monitoring and Continuous Improvement**:
   * **Dashboard Creation**: Develop an interactive dashboard to continuously monitor key metrics identified in this report (e.g., ADR by segment/channel, cancellation rates by lead time, room upgrade frequency, and customer type booking patterns).
   * **Periodic Review**: Conduct regular reviews of the booking data (e.g., quarterly) to identify emerging trends, validate existing assumptions, and adjust strategies in response to market changes and operational performance.

By implementing these recommendations, the Resort Hotel can leverage its booking data to enhance operational efficiency, reduce lost revenue from cancellations, optimize pricing, and ultimately drive sustainable growth. The insights from this analysis provide a solid foundation for data-driven decision-making within the organization.