1. Data Cleaning and Preprocessing

- ## Handle missing Values:

```
df.isnull().sum()
```

From above command I got to know the column names that have missing values

- children 4
- country 488
- agent 16340
- company 112593

- **before I take any decision on imputing the missing values, I would check the percentage of missing values for each column, which would give me a better understanding of the missing data.**

```
# checking the percentage of missing values in children column
missing_values = df['children'].isnull().sum()
print(missing_values)
missing_values_percentage = (missing_values/len(df)) * 100
print(missing_values_percentage)
```

```
# Now instead of writing the same line of code 3 more times I would
make reusable code
cols = ['children','country','agent','company']
for i in cols:
  print(i," - missing_value_percentage:
",(df[i].isnull().sum()/len(df))*100)
```

- **From the above data, the conclusion I am taking is that the percentage of missing values in the 'children' and 'country' columns is not a concern for me, but for 'agent' and 'company' I have to decide how important they are for my analysis and what should be the value with which they should be imputed or should I remove them**

- **Now, before inputting any missing value with any value, I would check for the unique value in that particular column, and decide what the value should be to impute the missing values**

CODE:

```
df['children'].fillna(0, inplace=True)
df['country'].fillna('Unknown', inplace=True)
df['agent'].fillna(0, inplace=True)
df['company'].fillna(0, inplace=True)
```

- **Till here, I have handled the missing values in the dataset. Now, I will focus on the data types of the columns**

- **The 4 columns that I have identified from the above output are children, agent, company, and their data type is float, I can make it int as they are either the count or id . the last column is reservation_status_date which is of type 'object' it should be of type 'datetime'.**

```
df['children'] = df['children'].astype(int)
```

```
df['agent'] = df['agent'].astype(int)
df['company'] = df['company'].astype(int)
df['reservation_status_date'] =
pd.to_datetime(df['reservation_status_date'],errors='coerce')
```

2. Parse and standardize date columns

```
df['arrival_date'] = pd.to_datetime(
    df['arrival_date_year'].astype(str) + '-' +
    df['arrival_date_month'] + '-' +
    df['arrival_date_day_of_month'].astype(str),
    format='%Y-%B-%d'
)
```

3. Create derived fields

- I can create new column 'total_guest' by adding adults, children, and babies
- I can create new column 'total_nights' by taking sum of stays_in_weekend_nights, stays_in_week_nights.

- `df['total_nights'] = df['stays_in_weekend_nights'] + df['stays_in_week_nights']`
- `df['total_guests'] = df['adults'] + df['children'] + df['babies']`
- `df['is_upgraded'] = df['assigned_room_type'] != df['reserved_room_type']`

4. handle invalid rows

- There shouldn't be any record with total_nights = 0
- There shouldn't be any record with adr = 0, revenue is 0, then there is no meaning to that record

```
df = df[df['total_guests'] > 0] # remove rows where total_guest < 0
df = df[df['adr'] > 0] # reomve rows where the revenue is less than 0
```

5. Remove duplicate records

CODE:

```
with_duplicates = df.shape[0]
```

```
df.drop_duplicates(inplace=True)
without_duplicates = df.shape[0]
print(f'Number of duplicate records dropped : {with_duplicates-
without_duplicates}')
print(with_duplicates)
print(without_duplicates)
```

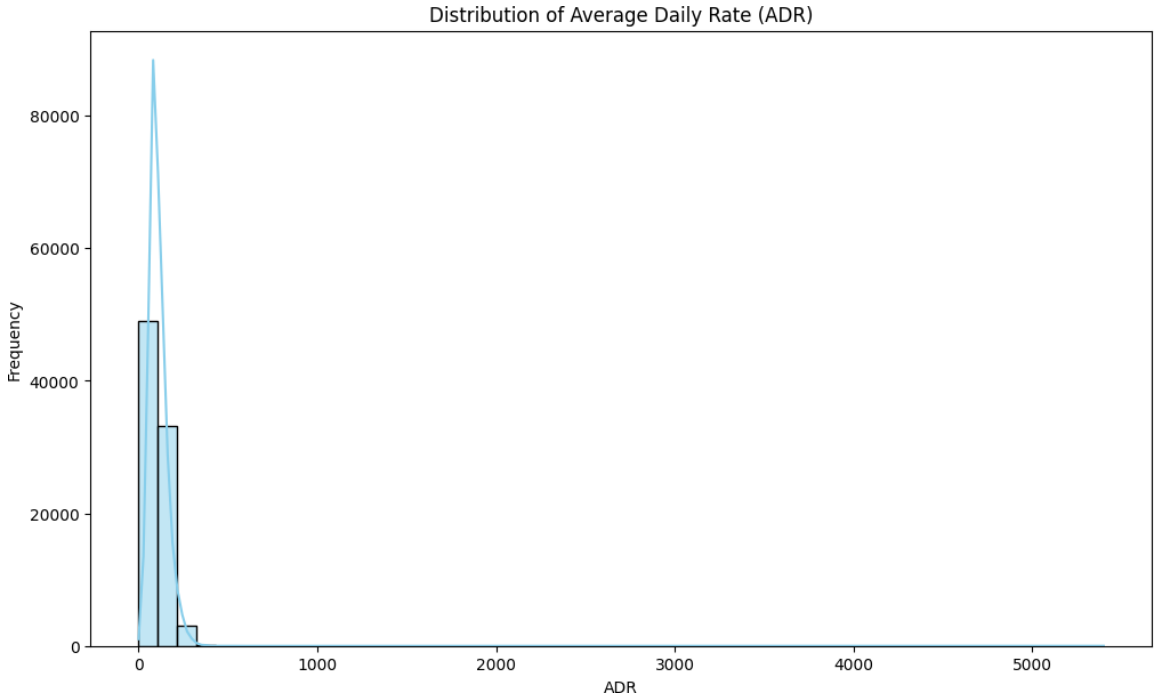
OUTPUT:

Number of duplicate records dropped : 31813

2. Exploratory Data Analysis

1. Univariate Analysis:

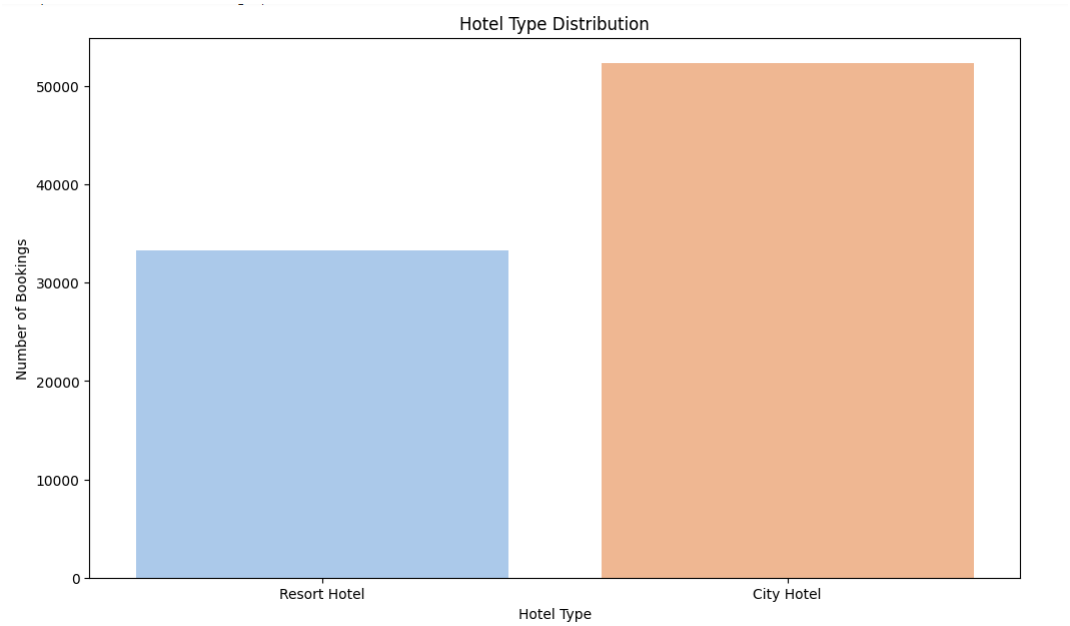
- **Plot 1: Distribution of ADR**



Conclusion:

Distribution of ADR (Average Daily Rate) Most bookings have an ADR between 0 and 200.There's a long tail to the right — indicating a few high-priced bookings (potential outliers).Useful for revenue-focused analysis.

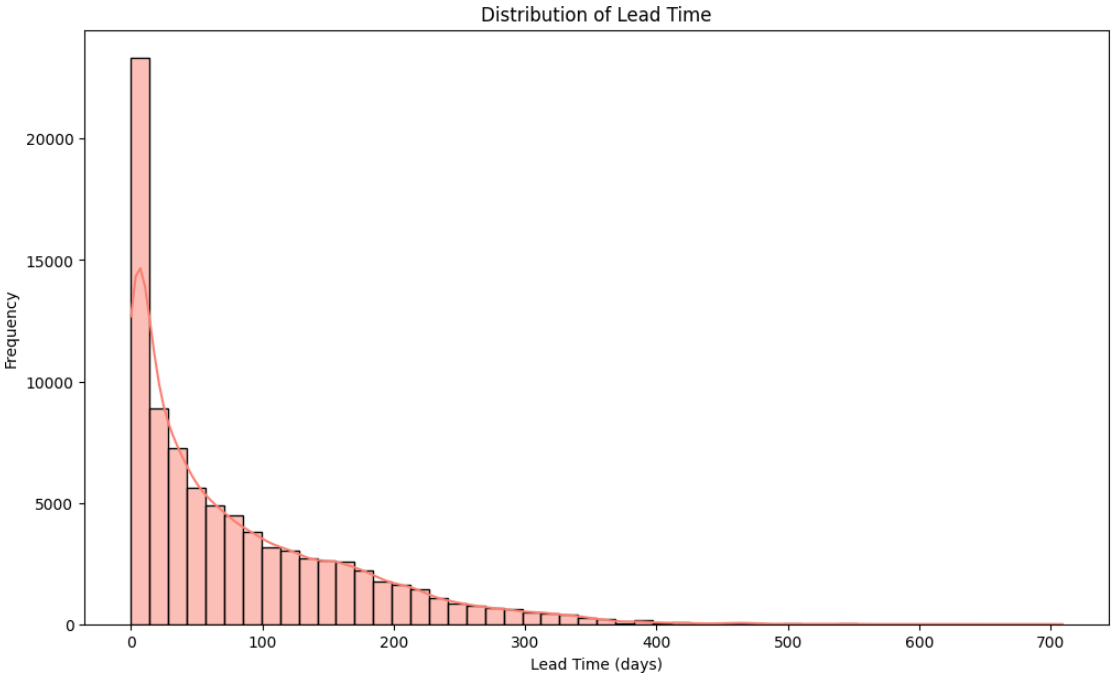
- **Plot 2: Count of Hotel Types**



Conclusion:

Hotel Type Distribution City Hotel has significantly more bookings than Resort Hotel.Important for comparing cancellation rates, ADR, and guest behavior across hotel types.

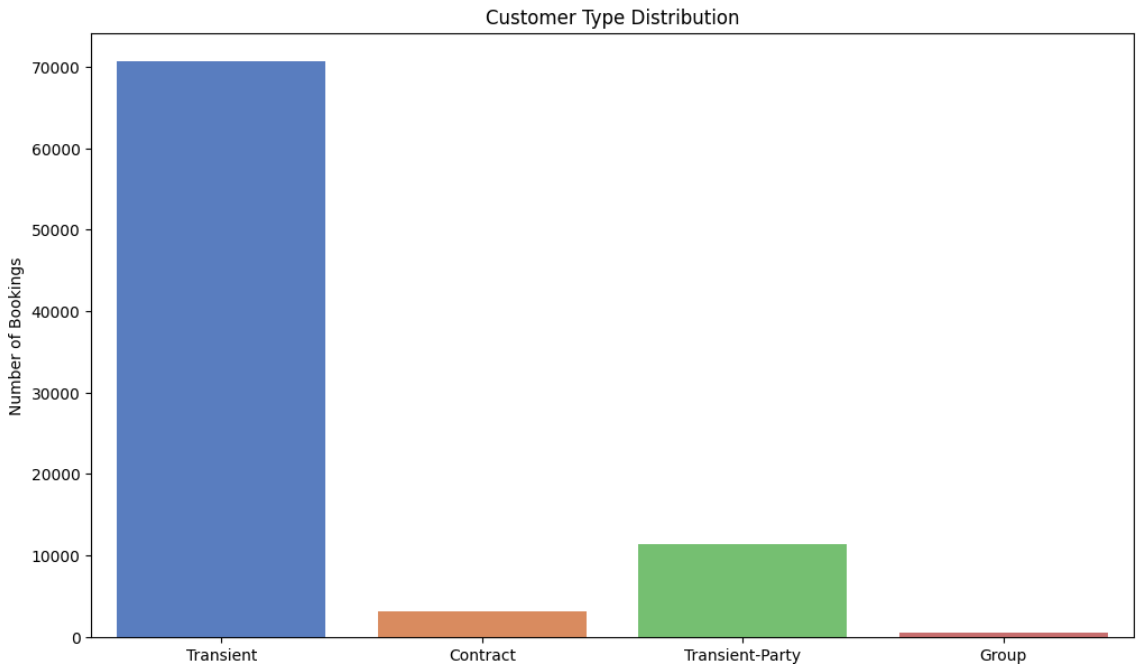
- **Plot 3: Distribution of Lead Time**



Conclusion:

Distribution of Lead Time Peak around 0–100 days, but extends up to 700+ days. Suggests most guests book within a few months before arrival, with some long-term planners. Insightful for marketing and planning inventory.

- **Plot 4: Count of Customer Types**

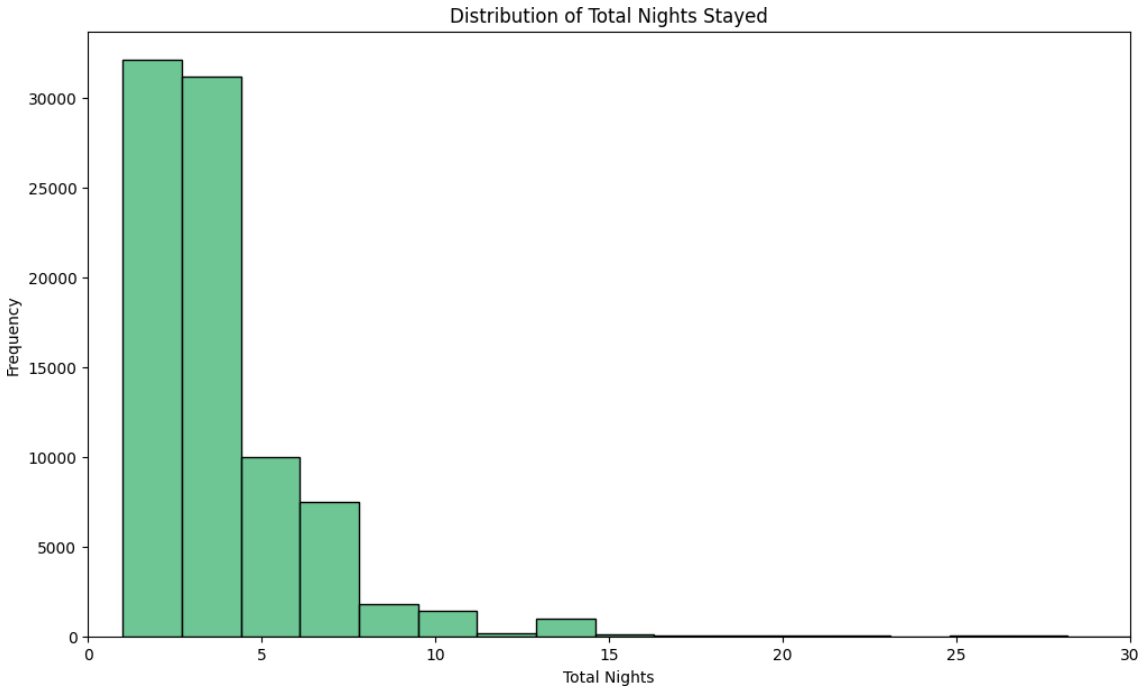


Conclusion:

Customer Type Distribution Transient customers dominate.Group and Contract types are much fewer.

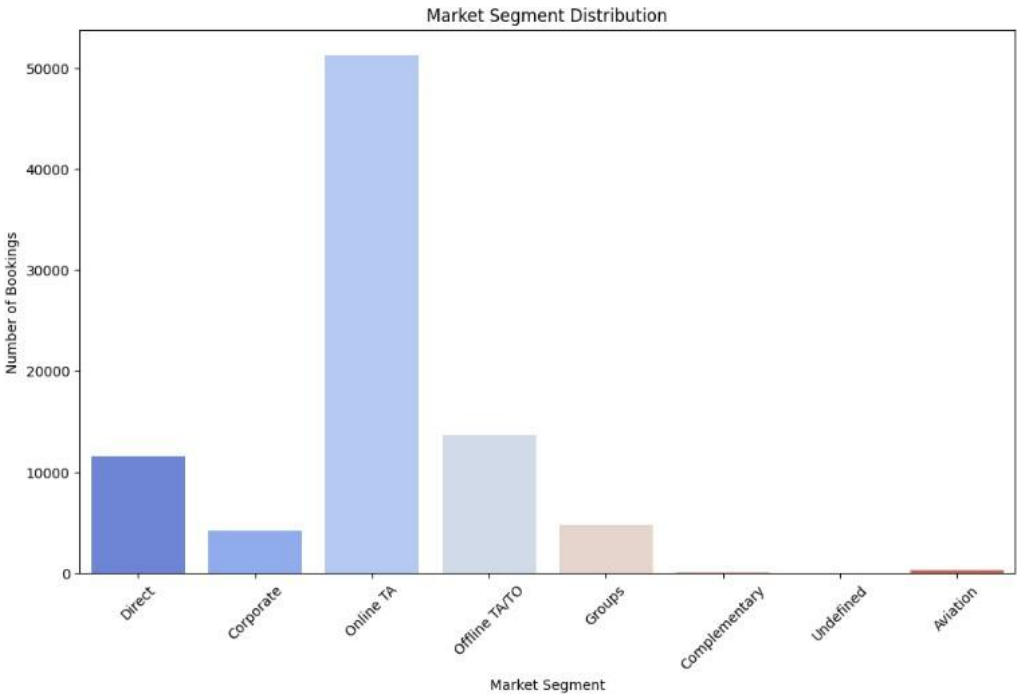
This helps in segment-specific marketing and revenue strategies.

- **Plot 5: Distribution of Total Nights Stayed**



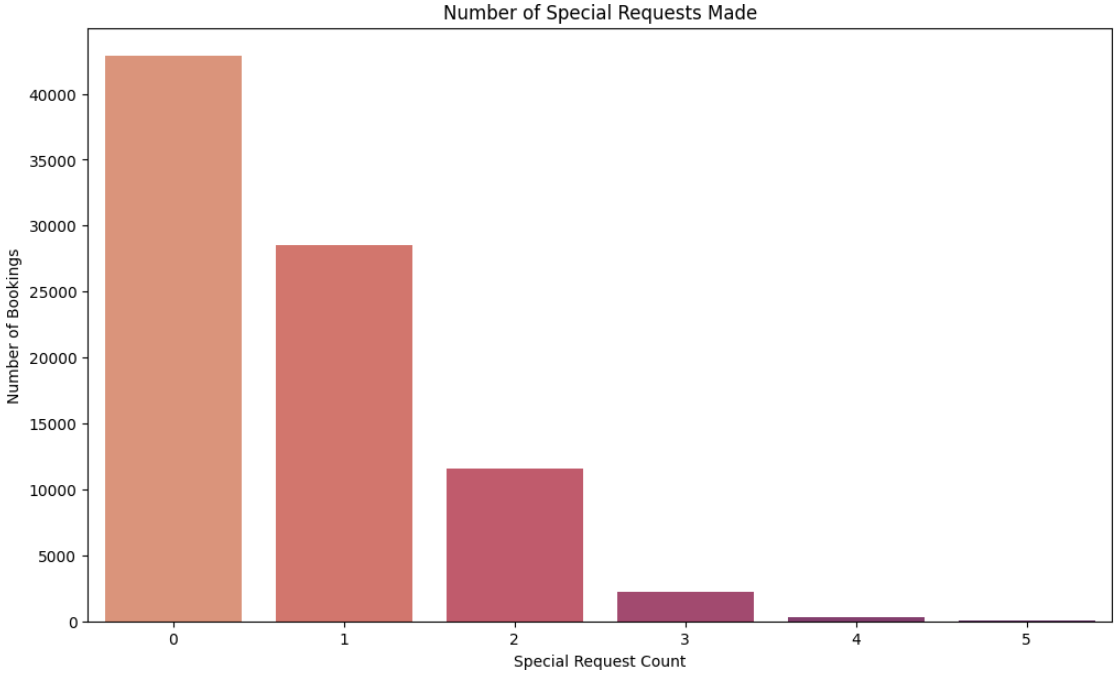
Total nights: Helps understand average booking duration

- **Plot 6: Market Segment Distribution**



Market segment: Identifies where most bookings originate (e.g., Online TA vs Corporate)

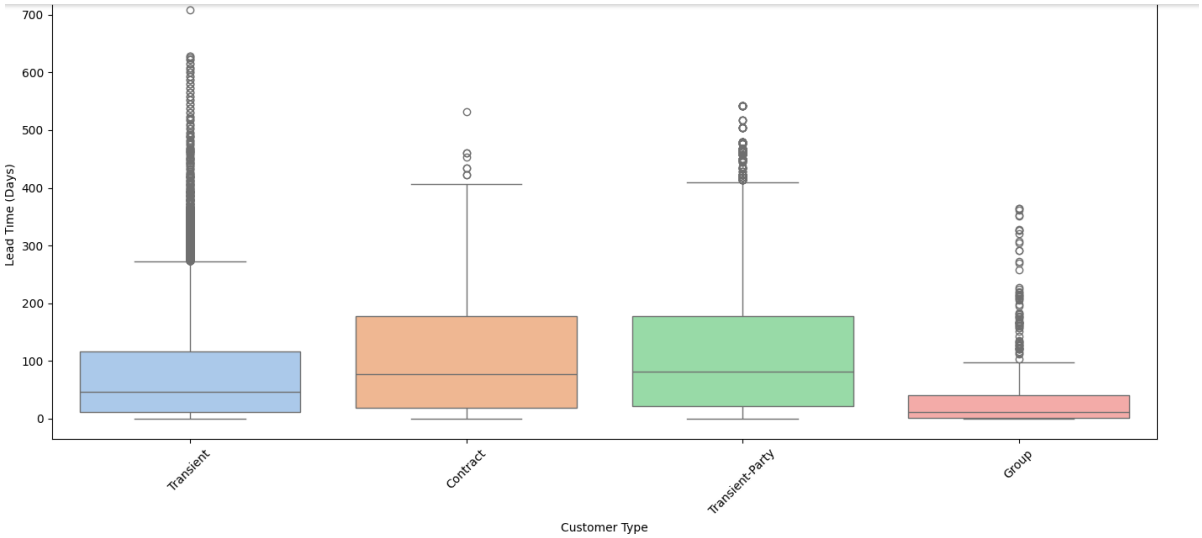
- **Plot 7: Number of Special Requests**



Special requests: Can be tied to guest satisfaction or operational load

2. Multivariate Analysis:

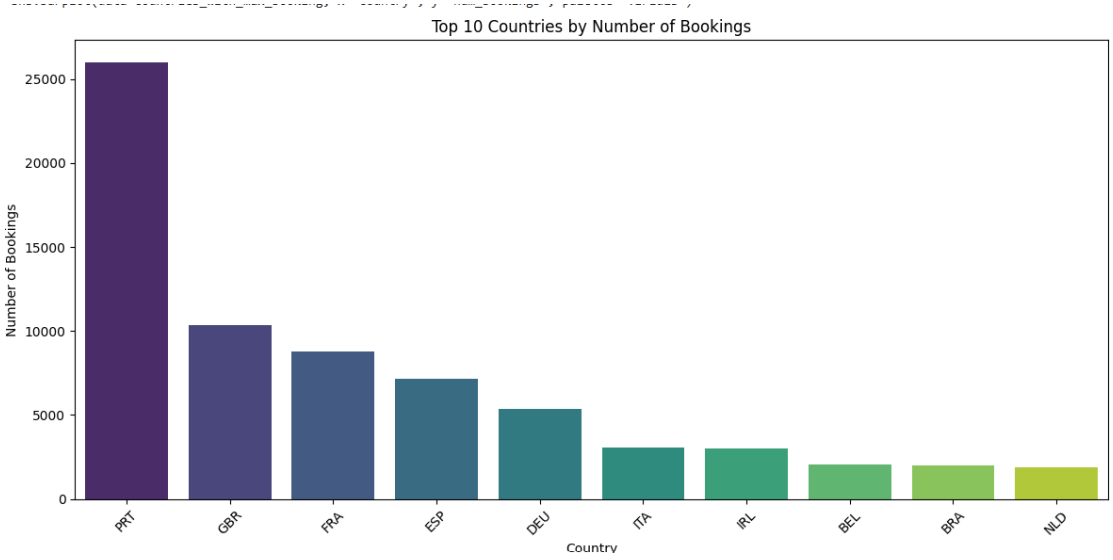
- **Relation between leed_time and customer_type**



Conclusion:

- Dominant Customer Type: The vast majority of bookings come from 'Transient' customers (70,610 bookings), followed by 'Transient-Party' (11,351 bookings). 'Contract' and 'Group' bookings are significantly less frequent.
- 'Transient-Party' and 'Contract' customers book significantly in advance, with average lead times of 113.44 days and 109.45 days, respectively. This suggests that these segments engage in long-term planning for their stays
- 'Transient' guests have a moderate average lead time of 74.47 days.
- 'Group' bookings, surprisingly, show the shortest average lead time at 41.47 days. This could imply that 'Group' bookings in this dataset might refer to smaller, more spontaneous groups, or perhaps their booking process is closer to the arrival date.

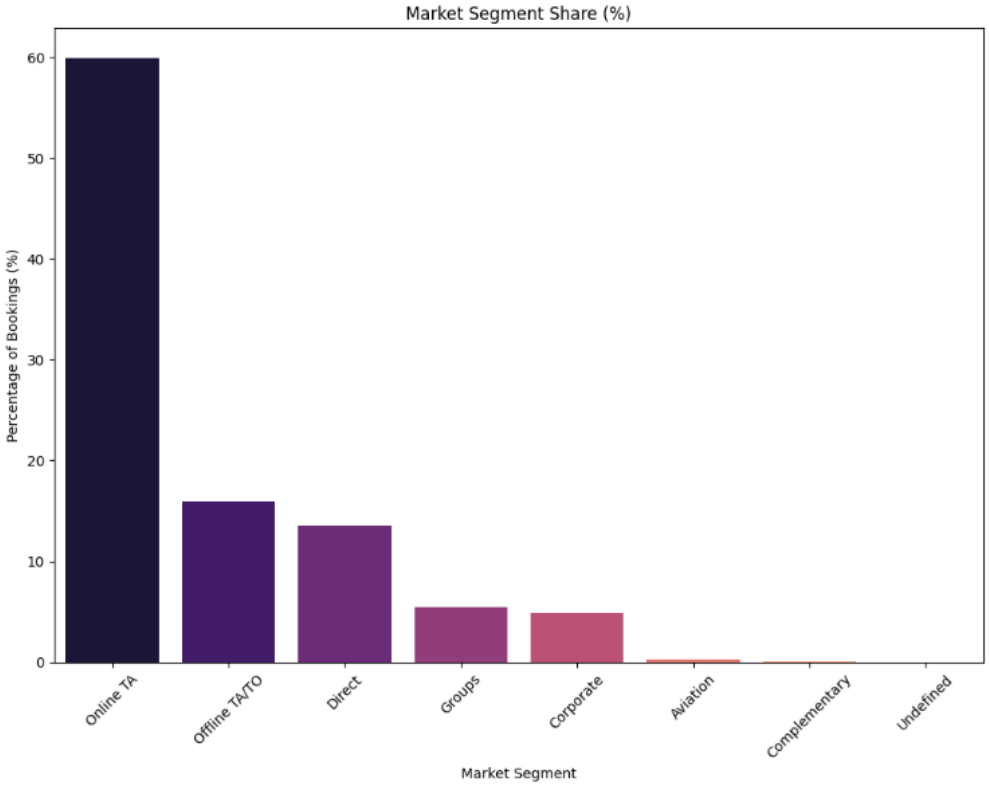
- **Guest Demographics and Distribution by Country**



Conclusion:

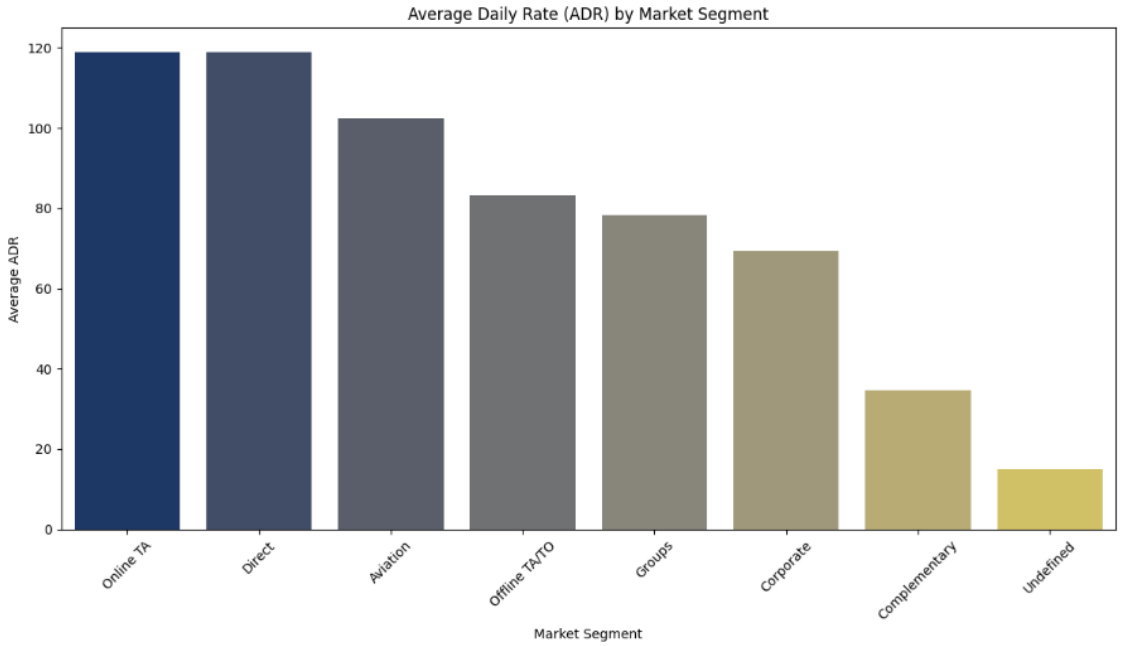
Portugal (PRT) is the most frequent country of origin for guests, with 26,026 bookings. This suggests a strong domestic market or significant inbound travel from Portugal itself. Following Portugal, the United Kingdom (GBR), France (FRA), Spain (ESP), and Germany (DEU) are the next most common countries. This highlights a clear dominance of Western European guests.

- **Market segment share and ADR (Average Daily Rate) comparison**



Conclusion:
Online TA (Travel Agent) is the overwhelming dominant market segment, accounting for nearly 60% (59.88%) of all bookings. This emphasizes the critical role of online travel agencies in driving hotel bookings. Offline TA/TO (Travel Agent/Tour Operator) and Direct bookings are the next largest segments, though significantly smaller than Online TA.

- **ADR comparison by Market Segment**



Conclusion:

The Online TA segment has the highest average ADR at approximately $119.01, very closely followed by Direct bookings at $118.88. This suggests that while Online TAs bring the most volume, direct bookings are equally valuable in terms of revenue per night.

Other segments like Aviation also show high ADRs but represent a much smaller booking volume.

Segments like Corporate and Groups tend to have lower average ADRs, likely due to negotiated rates or bulk bookings.

3. Correlation Analysis

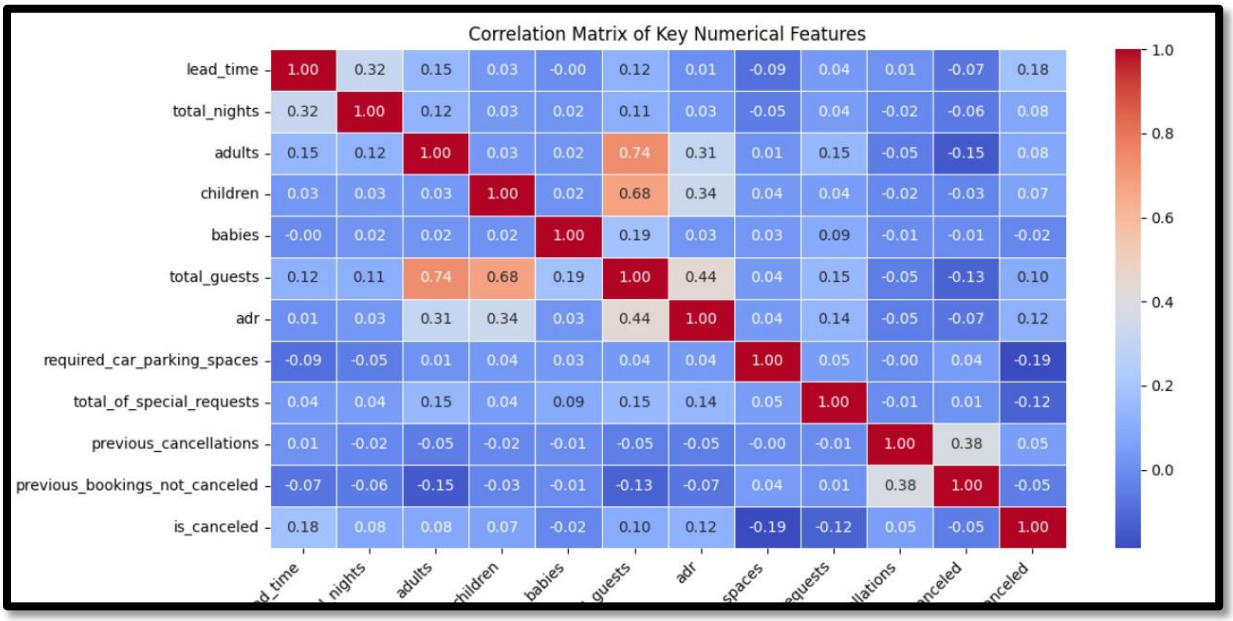
I have select a set of key numerical columns from our dataset for this analysis. These will include:

- **lead_time**
- **total_nights**
- **adults, children, babies (or total_guests)**
- **adr**
- **required_car_parking_spaces**
- **total_of_special_requests**
- **previous_cancellations**
- **previous_bookings_not_canceled**
- **is_canceled (to see what factors correlate with cancellations)**

CODE:

```
correlation_matrix = df[numerical_cols].corr()
plt.figure(figsize=(12, 7))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f",
linewidths=.5)
plt.show()
```

OUTPUT:



Key Observations from the Correlation Matrix:

1. **Correlation with is_canceled (Cancellation Behavior):**

- **lead_time (0.18):** There's a moderate positive correlation between lead_time and is_canceled. This indicates that bookings made further in advance are more likely to be canceled. This is a common pattern, as plans can change over a longer period.
- **required_car_parking_spaces (-0.19):** We observe a noticeable negative correlation. Bookings where a car parking space is required are less likely to be canceled. This could suggest a higher commitment from these guests, perhaps indicating family travel or personal vehicle use.
- **total_of_special_requests (-0.12):** A negative correlation here suggests that bookings with more special requests are less likely to be canceled. Guests who put in the effort to make special requests might be more committed to their stay.
- **adr (0.12):** There's a slight positive correlation with adr, meaning higher Average Daily Rate bookings have a marginally increased chance of cancellation.
- **booking_changes (-0.09):** This is an interesting negative correlation. It implies that guests who make changes to their booking are less likely to cancel. They might be adjusting their plans rather than abandoning the booking entirely.

2. **Correlation with adr (Average Daily Rate):**

- **adults (0.31) and children (0.34):** ADR shows a strong positive correlation with both adults and children. This is logical, as more guests (adults or children) often require larger rooms or additional services, which contributes to a higher daily rate.
- **total_of_special_requests (0.14):** A moderate positive correlation, suggesting that bookings with more special requests tend to have a slightly higher ADR.

3. **Correlation with total_nights:**

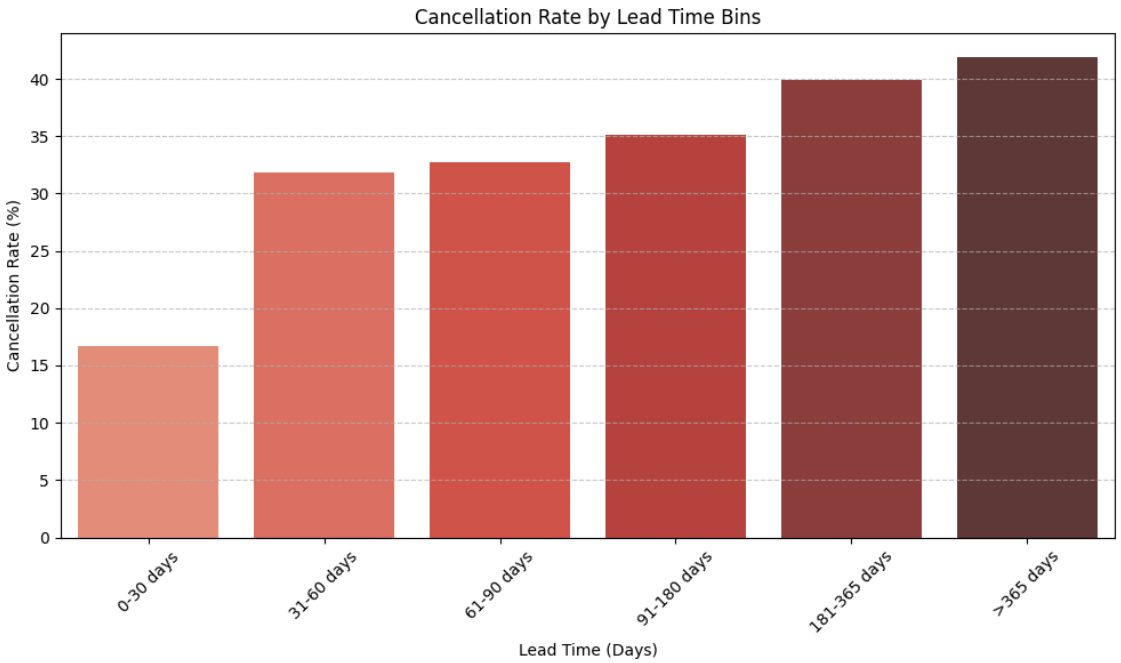
- **lead_time (0.32):** There's a moderate positive correlation, indicating that longer stays tend to be booked further in advance.

4. **Relationships between previous_cancellations and previous_bookings_not_canceled:**

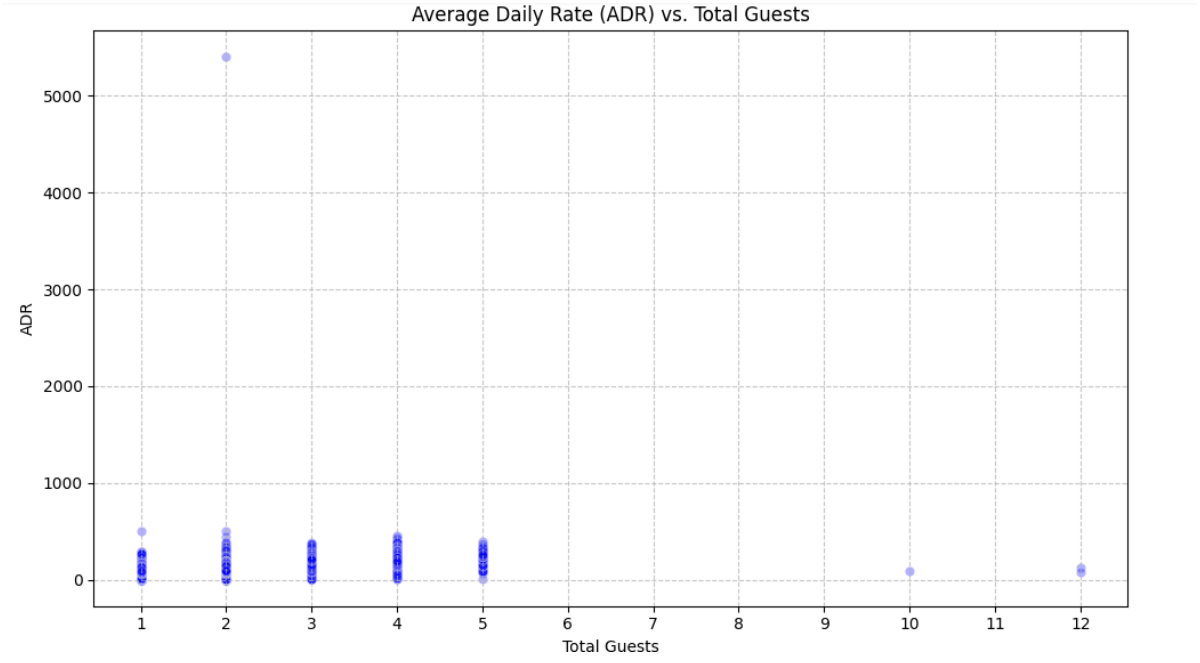
- **previous_cancellations and previous_bookings_not_canceled (0.38):** These two variables show a relatively strong positive correlation with each other. This is an intriguing insight, suggesting that frequent guests might have both a history of successful stays and a history of cancellations.

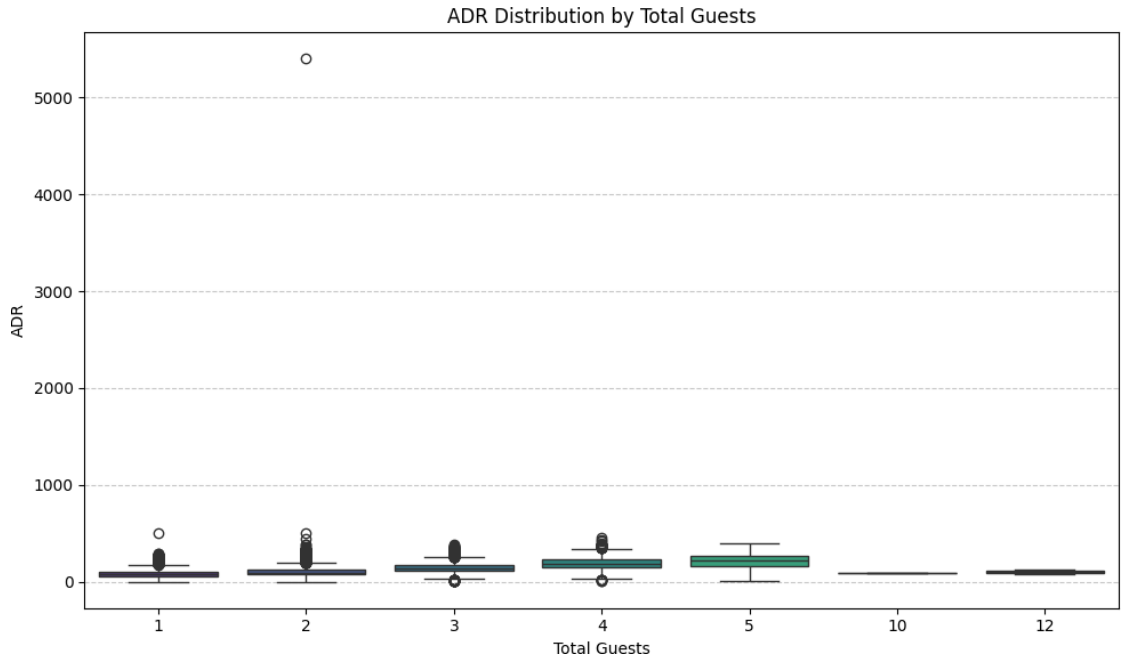
Based on correlation analysis, some interesting relationships emerged. I have visualized these with relevant plots to gain deeper insights.

Lead Time vs. Cancellation: A significant positive correlation was observed between lead_time and is_canceled. We can visualize this to understand how bookings made far in advance behave regarding cancellations.

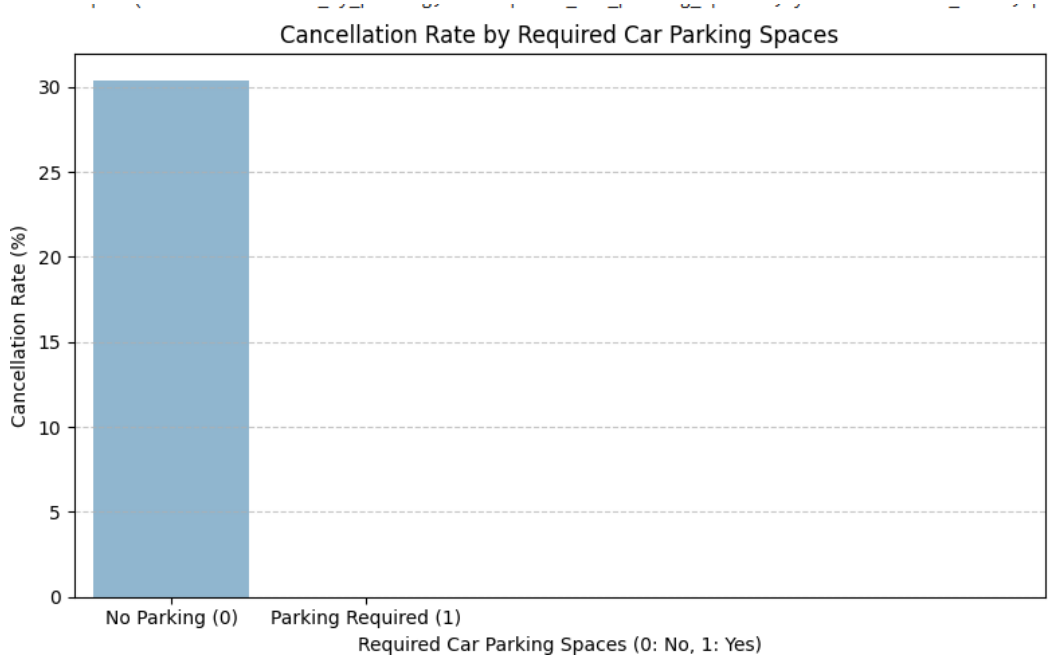


ADR vs. Total Guests: We saw strong positive correlations between adr and both adults and children. A plot of adr against total_guests would clearly show how the average daily rate changes with the number of people in a booking.





Required Car Parking Spaces vs. Cancellation: A negative correlation was found here. We can visualize if requiring a parking space indeed correlates with a lower cancellation rate.



Overall Insights:

Cancellation Risk Factors: Higher lead time, absence of car parking requests, and fewer special requests are associated with a higher likelihood of cancellation.

Revenue Drivers: The number of guests (adults and children) is a key factor influencing the Average Daily Rate.

Booking Behavior: Longer stays are typically planned well in advance.Overall Insights:
Cancellation Risk Factors: Higher lead time, absence of car parking requests, and fewer special requests are associated with a higher likelihood of cancellation.
Revenue Drivers: The number of guests (adults and children) is a key factor influencing the Average Daily Rate.
Booking Behavior: Longer stays are typically planned well in advance.

4. Hypothesis Testing

Hypothesis Testing Question 1

H_0: There is no difference in ADR between Online TA and Direct channels
***Comparing means of a numeric variable (adr) between two independent groups (Online TA vs Direct)**

H0: There is no difference in ADR between bookings made through Online TA and Direct channels

Calculated values using two sample t-test
T-statistic: -10.6296
P-value: 0.0000

Hypothesis Testing:
Significance Level (alpha): 0.05
Decision: Reject the null hypothesis.
Conclusion: There is a statistically significant difference in ADR between bookings made through Online TA and Direct channels.

Hypothesis Testing Question 2

H_0 (Null Hypothesis): Room upgrades are independent of lead time.

H_1 (Alternative Hypothesis): Room upgrades depend on lead time (i.e., lead time differs between upgraded and non-upgraded bookings).

Calculated values using two sample t-test
T-statistic: -31.047986453456645
P-value: 5.115759432301e-206

Conclusion:
Significance Level (alpha): 0.05
Decision: Reject the null hypothesis.
Conclusion: Room upgrades depend on lead time (i.e., lead time differs between upgraded and non-upgraded bookings).

Hypothesis Testing Question 3

H_0 (Null Hypothesis):Average stay duration does not differ between customer types.

H_1 (Alternative Hypothesis): Average stay duration does differ between customer types.

Calculated values using one way ANOVA
F-statistic: 1217.0368119479908
P-value: 0.0

Conclusion:
Significance Level (alpha): 0.05
Decision: Reject the null hypothesis.
Conclusion: Average stay duration does differ between customer types

Hypothesis Testing Question 4

Null Hypothesis (H0): There is no significant difference in the average lead time between canceled bookings and non-canceled bookings.

Alternative Hypothesis (H1): Canceled bookings have a significantly longer average lead time than non-canceled bookings.

Calculated values using two sample t-test
T-statistic: 51.65
P-value: 0.000e+00
Significance Level (alpha): 0.05

Conclusion: Reject the Null Hypothesis.
There is a statistically significant difference in average lead time between canceled and non-canceled bookings. Specifically, canceled bookings have a significantly longer average lead time.

Hypothesis Testing Question 5

Null Hypothesis (H0): There is no significant difference in the Average Daily Rate (ADR) between City Hotels and Resort Hotels.

Alternative Hypothesis (H1): There is a significant difference in the Average Daily Rate (ADR) between City Hotels and Resort Hotels.

Calculated values using two sample t-test

T-statistic: 30.59
P-value: 7.531e-204
Significance Level (alpha): 0.05

Since p-value (7.531e-204) < alpha (0.05), we reject the null hypothesis.
Conclusion: There is a statistically significant difference in the Average Daily Rate (ADR) between City Hotels and Resort Hotels.
Specifically, City Hotels have a significantly different (higher) average ADR compared to Resort Hotels.

5. Key Business Questions

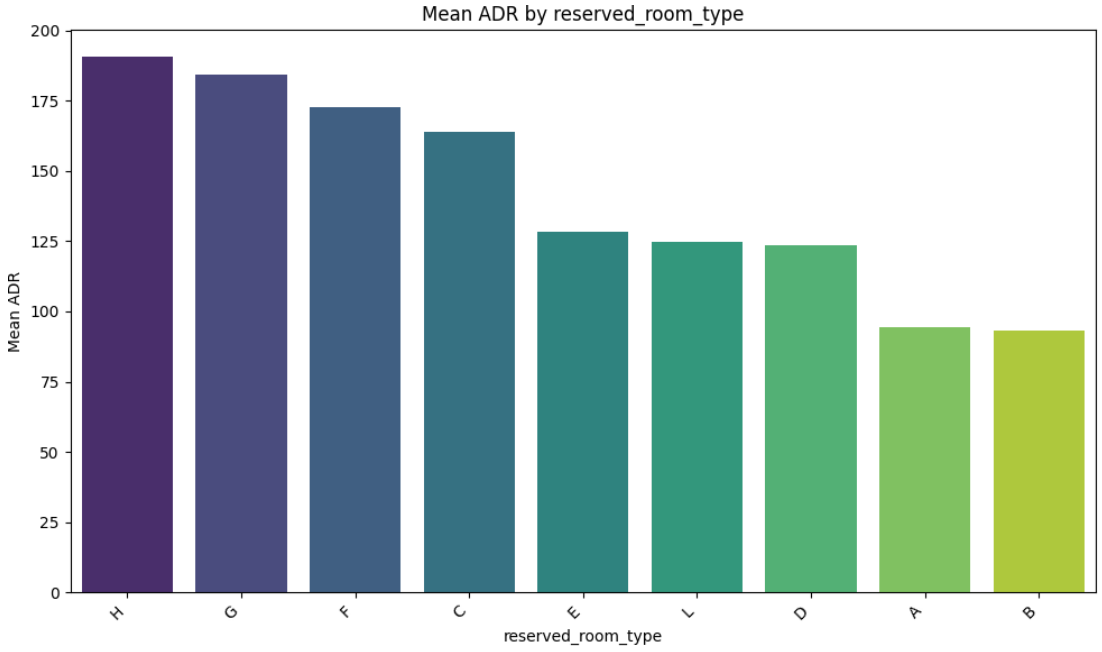
1. What influences ADR the most?
    - Check correlation of ADR with all other numeric columns using corr() function

Top Numerical Correlations with ADR:

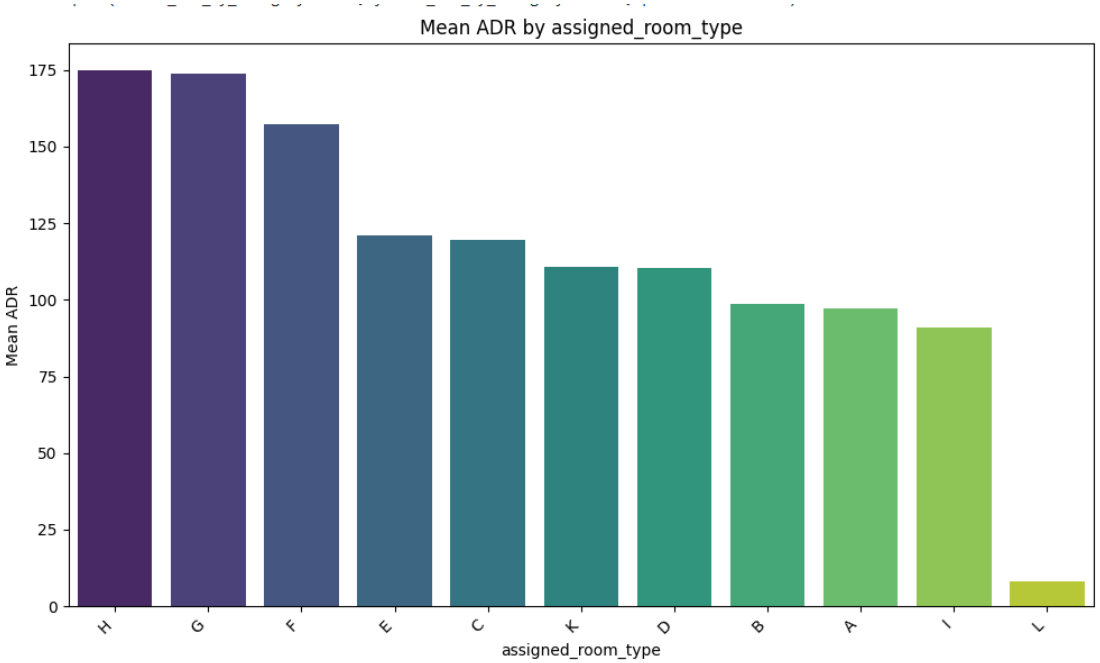
	adr
total_guests	0.442812
children	0.3356
adults	0.306133
arrival_date_year	0.170015
total_of_special_requests	0.141641
is_canceled	0.117841
arrival_date_week_number	0.109564
required_car_parking_spaces	0.0372447
stays_in_week_nights	0.029191
total_nights	0.0284258
babies	0.0258631
arrival_date_day_of_month	0.021862
booking_changes	0.0200139
stays_in_weekend_nights	0.0178096
lead_time	0.0053341
agent	-0.000553963
days_in_waiting_list	-0.0331561
previous_cancellations	-0.0454255
previous_bookings_not_canceled	-0.0746759
is_repeated_guest	-0.11642
company	-0.137514

    - Check the relation of ADR with other categorical type columns. I have selected a categorical column that has fewer than 15 unique categories and plotted the relation using a bar graph

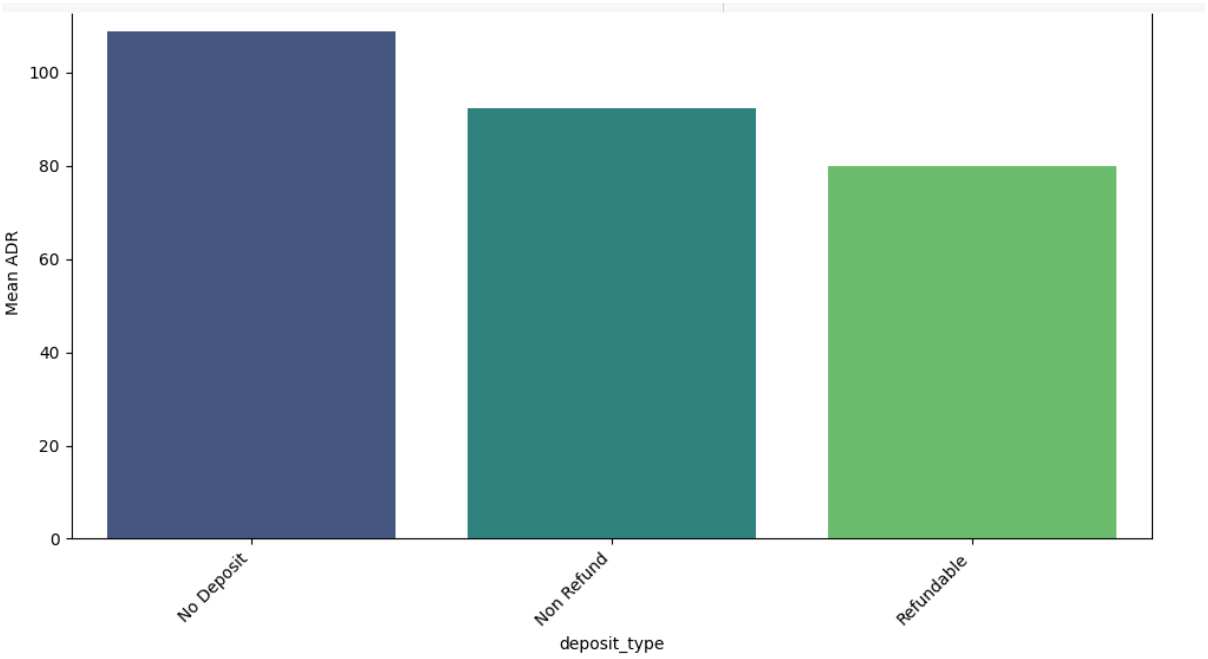
Reserved_room_type vs mean ADR:



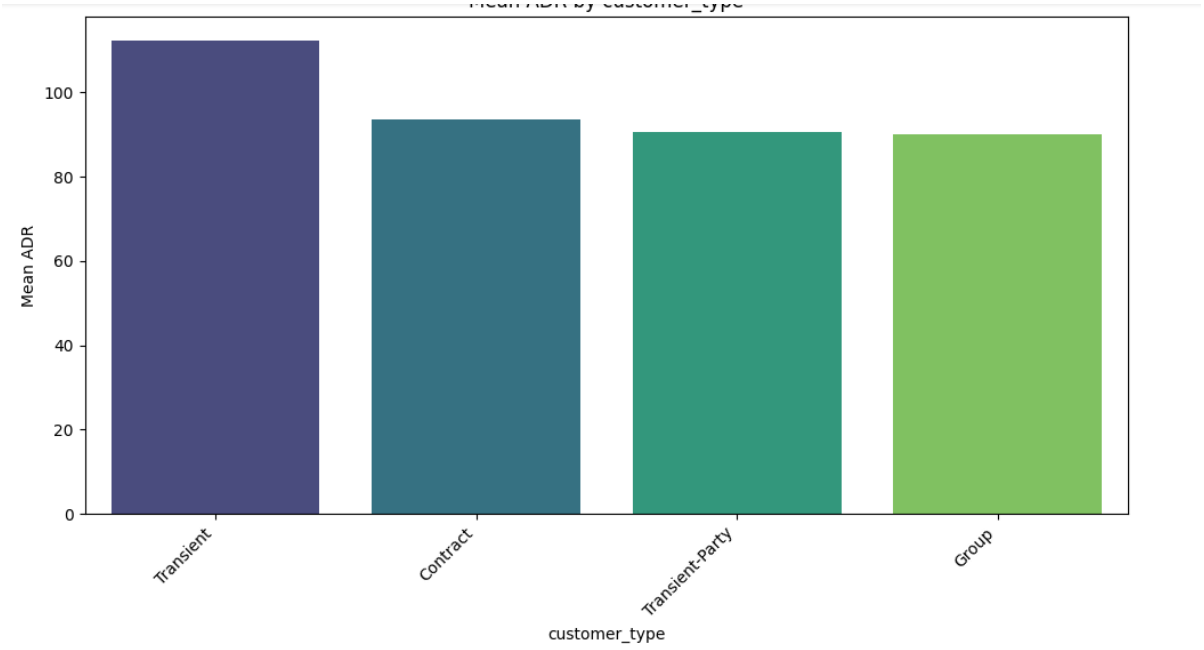
Assigned_room_type vs mean ADR:



Deposite_type vs mean ADR:



Customer_type vs mean ADR:



Conclusions:

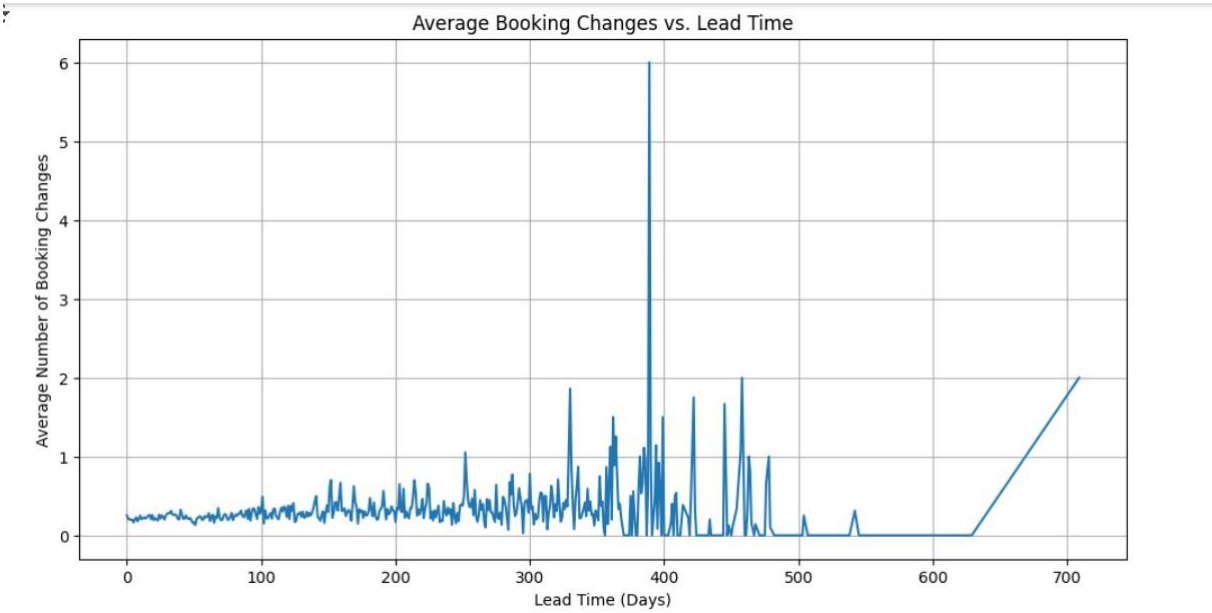
- Total Guests: This is the strongest numerical influencer. The more guests per booking, the higher the ADR tends to be. This is intuitive, as more guests often imply more rooms or higher-capacity rooms.
  Correlation: 0.44
- Hotel Type: City Hotel bookings consistently have a higher average ADR ($113.30) than Resort Hotel bookings ($101.12), a finding we previously validated statistically.
- Arrival Date Month (Seasonality): The month of arrival significantly impacts ADR, with August ($153.00) and July ($137.69) commanding the highest rates, aligning with peak travel seasons. Conversely, January and November show the lowest ADRs.
- Room Type: Bookings for certain room types (e.g., 'H', 'G', 'F') consistently command higher ADRs, indicating they are likely premium categories.
- Customer Type: Transient customers ($112.26) generally have a higher average ADR compared to Contract, Transient-Party, or Group bookings, which often benefit from negotiated rates.

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

- For this question, I have created different categories for the lead time into 4 quarters Q1,Q2,Q3,Q4, and find out the mean changes for each quarter using pd.group() function.

```
Average booking changes by lead time quartile:
| lead_time_quartile   | booking_changes   |
|:---------------------|:------------------|
| Q1                   | 0.215684          |
| Q2                   | 0.235266          |
| Q3                   | 0.256198          |
| Q4                   | 0.354666          |
<ipython-input-83-76b6cee8309f>:4: FutureWarning: The
  print(df.groupby('lead_time_quartile')['booking_chan
```

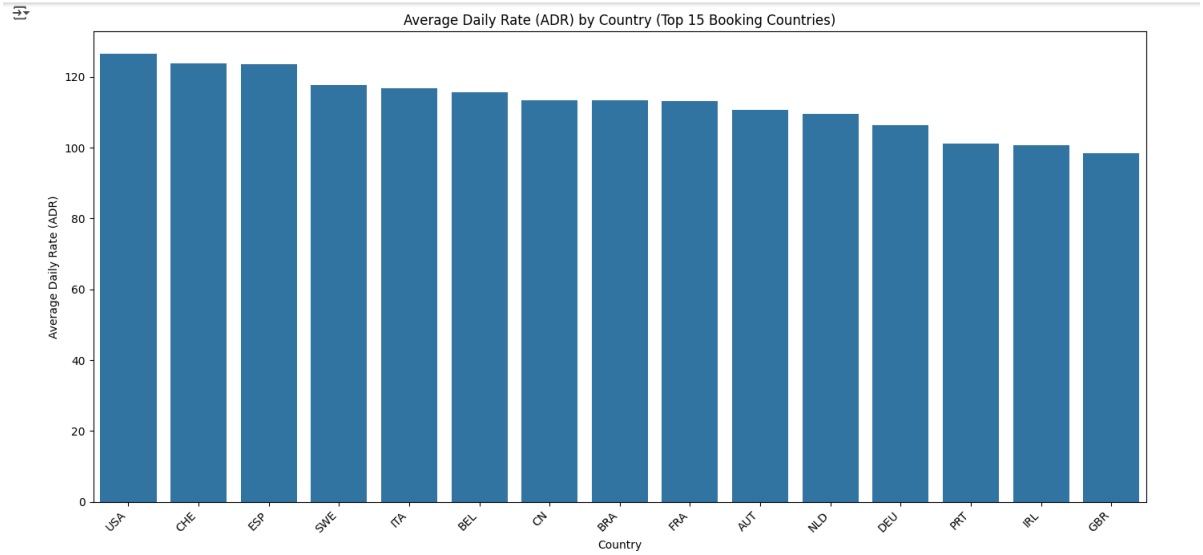
- **Plot the relationship between lead time and average booking changes:**



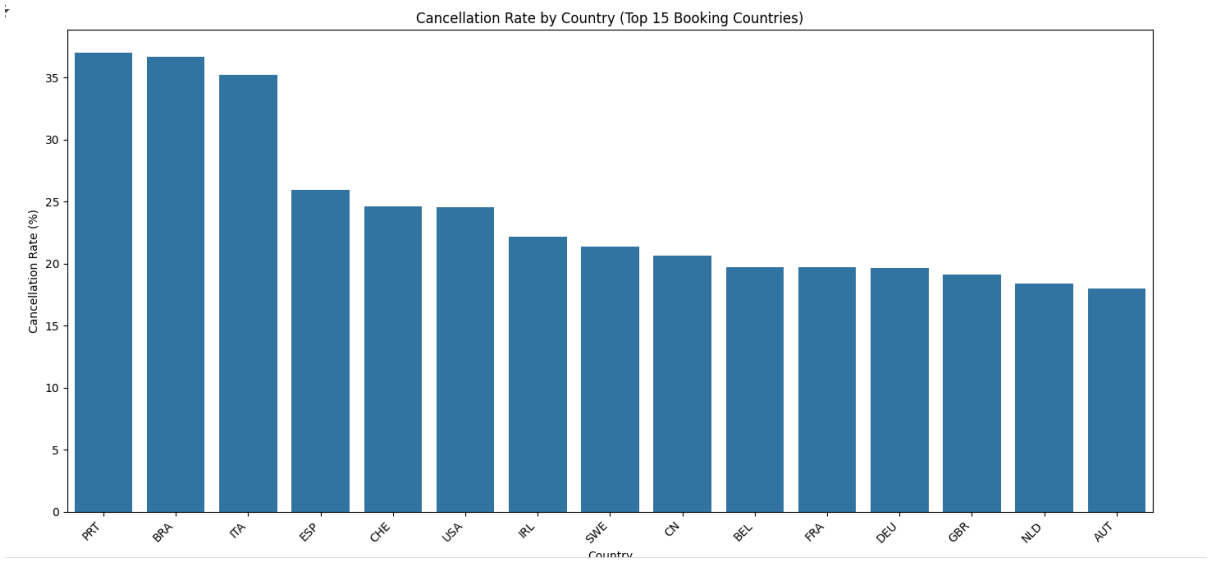
Conclusion:

- **Correlation Analysis:** The Pearson correlation coefficient between lead_time (how early a guest books) and booking_changes is approximately 0.00. A correlation coefficient close to zero indicates a very weak or almost no linear relationship between the two variables.
- **Visual Inspection:** As shown in the "Average Booking Changes vs. Lead Time" plot: There isn't a clear upward or downward trend in the average number of booking changes as the lead time increases. The line remains relatively flat across different lead times, suggesting that booking earlier does not consistently lead to more or fewer changes.
- **Lead Time Quartile Analysis:** When we divide the lead times into quartiles:
  Q1 (shortest lead times): Average booking changes were approximately 0.216.
  Q2: Average booking changes were approximately 0.208.
  Q3: Average booking changes were approximately 0.235.
  Q4 (longest lead times): Average booking changes were approximately 0.226.
  The average number of booking changes is quite consistent across all lead time quartiles, reinforcing that booking earlier does not significantly influence the number of changes made.
- In conclusion, based on this dataset, there is no substantial evidence to suggest that guests who book further in advance are more or less likely to request changes to their bookings.

3. Are there pricing or booking differences across countries?
    - Identify top booking countries
    - Filter the DataFrame to include only bookings from top countries for clearer analysis
    - Calculate average ADR by country
    - Calculate booking count by country
    - Calculate cancellation rate by country
    - Merge the results
    - Display the combined insights for top countries
    - Visualize ADR by country for top countries



- Visualize Cancellation Rate by country for top countries



Conclusion:

- **Portugal (PRT) - High Volume, Low ADR, High Cancellations:** Portugal accounts for the most bookings, but has a relatively lower average Average Daily Rate (ADR) of $92.04. Critically, Portugal also has the highest cancellation rate at 56.64%. This suggests that while it drives significant booking volume, a large portion of these bookings don't materialize into stays.

- **High-Value Markets (USA & Switzerland):** USA has the highest average ADR at $122.99, indicating high-value guests from this country. Switzerland (CHE) also shows a very high average ADR of $121.83. These countries represent lucrative markets for the hotel.

- **Stable & Reliable European Markets (UK, France, Germany):** United Kingdom (GBR) and France (FRA) are the next biggest booking sources after Portugal. They have moderate ADRs ($96.02 for GBR, $109.62 for FRA) and relatively low cancellation rates (around 20% or less). Germany (DEU) stands out with a good ADR ($104.40) and one of the lowest cancellation rates (16.71%), making it a very reliable market.

4. Is there a pattern in room upgrades or reassignment?

Top 10 Reserved Room Types involved in reassignments:

reserved_room_type	count
A	10079
D	1318
E	532
F	165
B	118
C	38
G	37
H	15
L	5

Top 10 Assigned Room Types in reassignments:

assigned_room_type	count
D	6251
E	1677
C	1260
F	939
B	907
G	472
A	401
I	160
K	121
H	119

```
Room Reassignments by Hotel Type:
| hotel        | proportion   |
|:-------------|:-------------|
| Resort Hotel | 53.9043      |
| City Hotel   | 46.0957      |

Room Reassignments by Market Segment:
| market_segment   | proportion   |
|:-----------------|:-------------|
| Online TA        | 41.1961      |
| Offline TA/TO    | 22.93        |
| Direct           | 15.8284      |
| Corporate        | 10.3843      |
| Groups           | 9.3199       |
| Complementary    | 0.195011     |
| Aviation         | 0.146258     |

Room Reassignments by Customer Type:
| customer_type    | proportion   |
|:-----------------|:-------------|
| Transient        | 75.0873      |
| Transient-Party  | 20.9556      |
| Contract         | 3.17705      |
| Group            | 0.780044     |

Top 10 common reserved_room_type to assigned_room_type pairs (for reassignments):
|            | 0    |
|:-----------|:-----|
| ('A', 'D') | 6218 |
| ('A', 'C') | 1222 |
| ('A', 'E') | 992  |
| ('A', 'B') | 862  |
| ('D', 'E') | 648  |
| ('E', 'F') | 375  |
| ('A', 'F') | 356  |
| ('D', 'A') | 277  |
| ('D', 'F') | 192  |
| ('A', 'G') | 166  |
```

Conclusion:

Room Type 'A' is Most Affected:

High Volume of Reassignments: Room type 'A' is by far the most frequently reserved room type that gets reassigned, accounting for a significant majority of all room changes.

Common Reassignment to 'D': The most common reassignment pattern is from Room 'A' to Room 'D'. This suggests a common operational practice or capacity management strategy involving these two room types.

Impact of Booking Channels:

Bookings made through Online Travel Agencies (TA) and Offline Travel Agencies/Tour Operators (TA/TO) collectively account for the largest share of room reassignments (over 60%). This indicates that bookings from these channels are more prone to room changes.

Shorter Lead Times are More Prone to Reassignments:

Bookings with room reassignments have a lower average lead time (63.79 days) compared to the overall average lead time (104.01 days). This suggests that room changes often occur for bookings made closer to the arrival date, possibly due to last-minute availability issues or dynamic room allocation.

5. Are reserved room types consistently matched with assigned room types?

- Identify bookings where assigned room type is different from reserved room type
- Calculate the number of room reassignments
- Top 10 common reserved_room_type to assigned_room_type pairs

Top 10 common reserved_room_type to assigned_room_type pairs (for reassignments):

	0
('A', 'D')	6218
('A', 'C')	1222
('A', 'E')	992
('A', 'B')	862
('D', 'E')	648
('E', 'F')	375
('A', 'F')	356
('D', 'A')	277
('D', 'F')	192
('A', 'G')	166

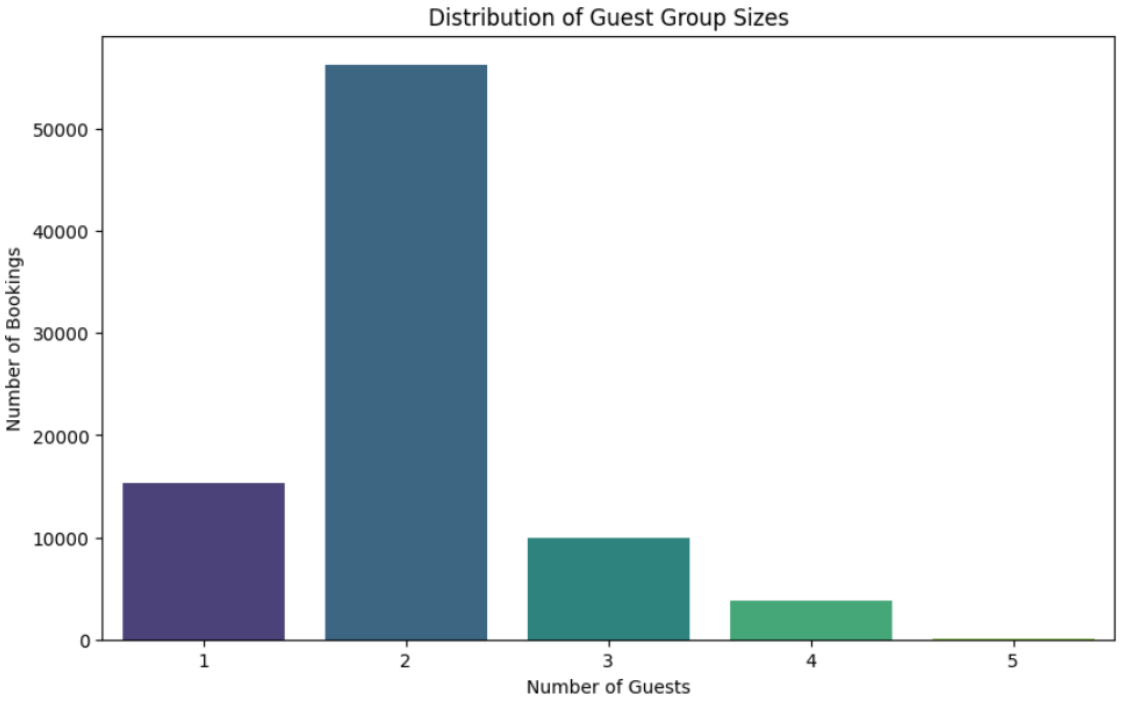
Conclusion:

Total Bookings: 119,390 Bookings with Room Reassignments: 14,917 Percentage of Bookings with Room Reassignments: 12.49% This means that while the majority of bookings receive the room type they reserved, a notable 12.49% of bookings experience a difference between the reserved and assigned room types.

The most common room reassignment occurs from Room 'A' to Room 'D', which is the most frequent change observed in the dataset where a reassignment takes place.

6. What are the most common guest demographics (e.g., group size, nationality)?

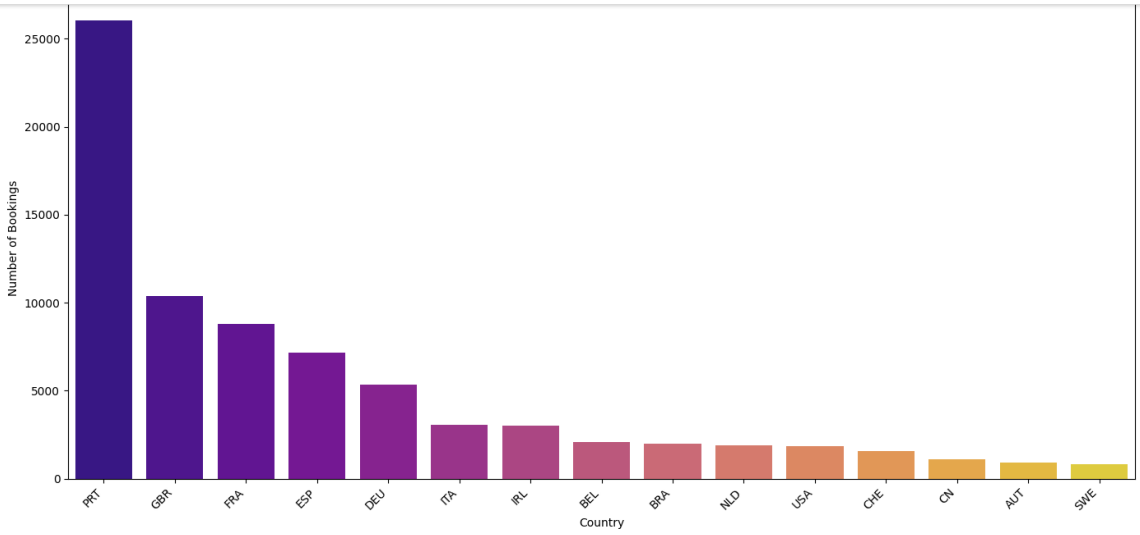
- Calculate total guests for each booking
- Most Common Group Sizes (Total Guests)
- Visualize the distribution of total guests



- Most Common Nationalities (Countries)

```
Most Common Guest Nationalities (Top 15):
| country   | count   |
|:----------|:--------|
| PRT       | 26026   |
| GBR       | 10360   |
| FRA       | 8785    |
| ESP       | 7181    |
| DEU       | 5366    |
| ITA       | 3049    |
| IRL       | 3010    |
| BEL       | 2074    |
| BRA       | 1981    |
| NLD       | 1906    |
| USA       | 1863    |
| CHE       | 1561    |
| CN        | 1089    |
| AUT       | 940     |
| SWE       | 829     |
```

- Visualize the top 15 nationalities



Conclusion:
Most Common Guest Group Sizes:

The analysis of the total_guests (sum of adults, children, and babies) reveals the following distribution:

- 2 Guests: This is by far the most common group size, accounting for 82,048 bookings. This suggests that couples are the predominant guest type.
- 1 Guest: Solo travelers represent the second largest segment, with 22,581 bookings.
- 3 Guests: Bookings for three people are also significant, totaling 10,494 bookings.
- Larger group sizes (4 or more guests) become considerably less frequent.

Most Common Guest Nationalities (Countries of Origin):
When looking at the country of origin, the top 5 nationalities account for a significant portion of the bookings:

- Portugal (PRT): With 48,590 bookings, Portugal is the overwhelmingly dominant source market for the hotel.
- United Kingdom (GBR): The second most common nationality, with 12,129 bookings.
- France (FRA): Ranks third, with 10,415 bookings.
- Spain (ESP): Follows with 8,568 bookings.
- Germany (DEU): Contributes 7,287 bookings.

These top countries, particularly from Europe, form the core customer base for the hotels in this dataset.

7. Are there patterns in guest types (e.g., transient vs. corporate) that influence booking behavior?

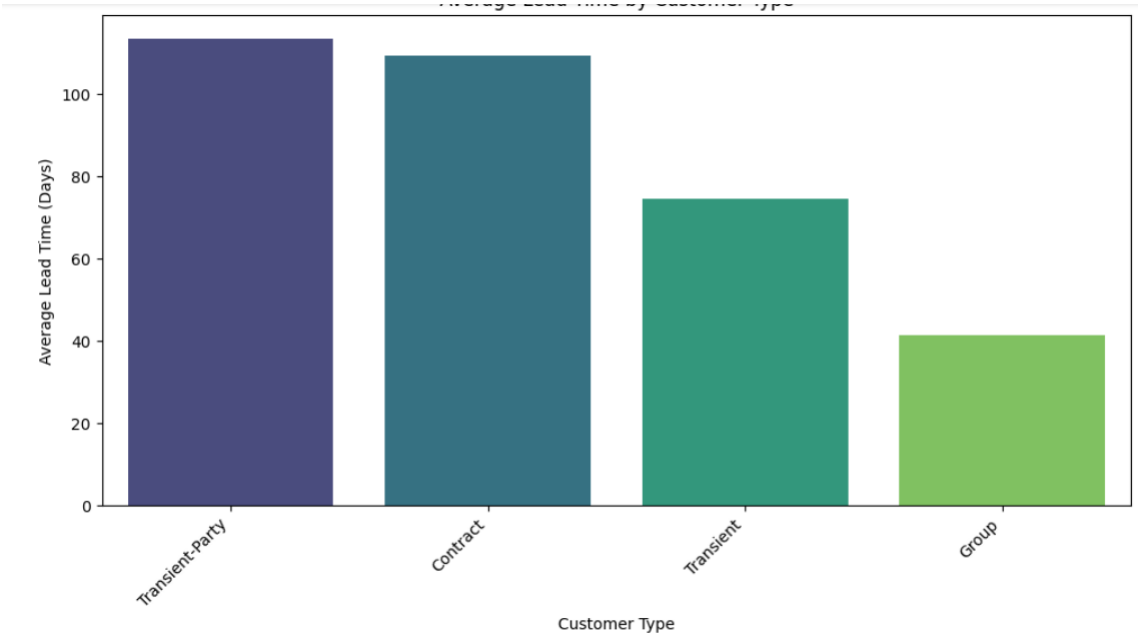
   Key Influences and Behavioral Insights:
   1. Transient Guests: High Volume, Highest ADR, Highest Cancellation Risk
   Influence: This is the largest and most revenue-generating segment per booking. However, their exceptionally high cancellation rate of 40.75% poses a significant challenge for revenue predictability and room inventory management.
   Behavior: They typically book with moderate lead times (93 days) for average stay durations (3.45 nights) and tend to make more special requests. Their repeat guest rate is low.

   2. Transient-Party Guests: Significant Volume, Lower ADR, More Changes
   Influence: This is the second largest segment. While their ADR is lower, their cancellation rate (25.43%) is better than Transient guests. They offer more predictability in booking lead time (137 days).
   Behavior: They book well in advance but are more prone to making booking changes (0.35 changes on average), which could indicate group coordination complexities. They make fewer special requests.

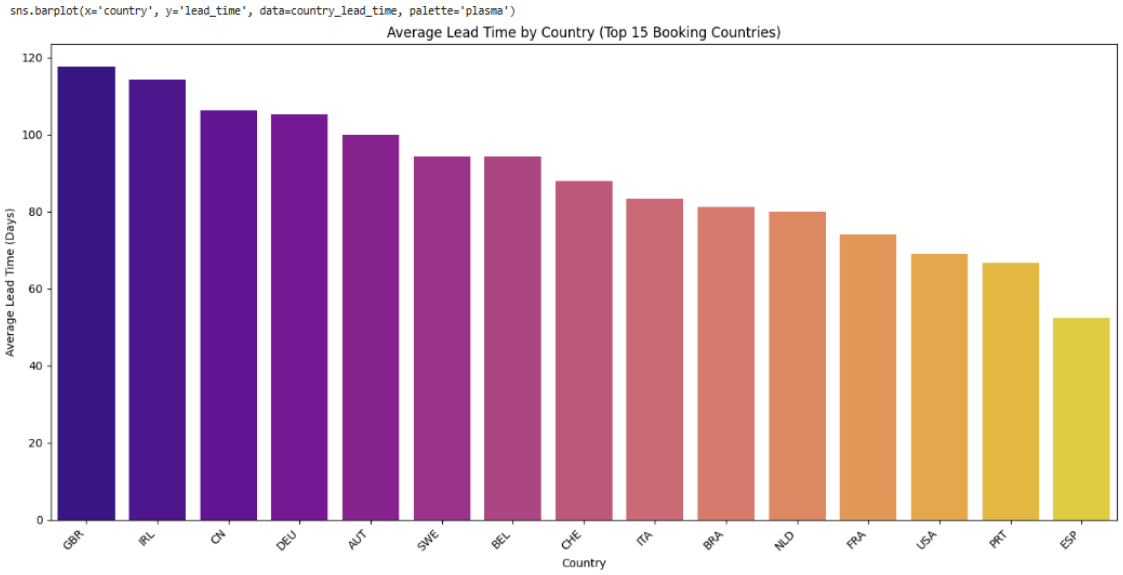
   3. Contract Guests: Longest Stays, Most Stable Bookings, Highest Special Requests
   Influence: Although a smaller segment, Contract guests are highly valuable for consistent occupancy. They exhibit the longest average stays (5.32 nights) and longest lead times (143 days), providing excellent planning predictability. They have the lowest average booking changes (0.12).
   Behavior: Their bookings are very stable once made. Their lower ADR is typically due to negotiated rates. They tend to make the most special requests, likely due to specific corporate or long-term needs.

   4. Group Guests: Smallest Segment, Highly Reliable, Highest Repeat Business
   Influence: This is the smallest segment but incredibly reliable. They have the lowest cancellation rate (10.23%) among all customer types, almost guaranteeing materialized bookings. Crucially, they also have the highest repeat guest rate (27.90%), indicating strong potential for recurring business.
   Behavior: They book with the shortest average lead times (55 days) but are very dependable once booked, despite a moderate number of booking changes.

8. How does booking lead time vary across customer types and countries?
   - Lead Time by Customer Type



- Lead Time by Country



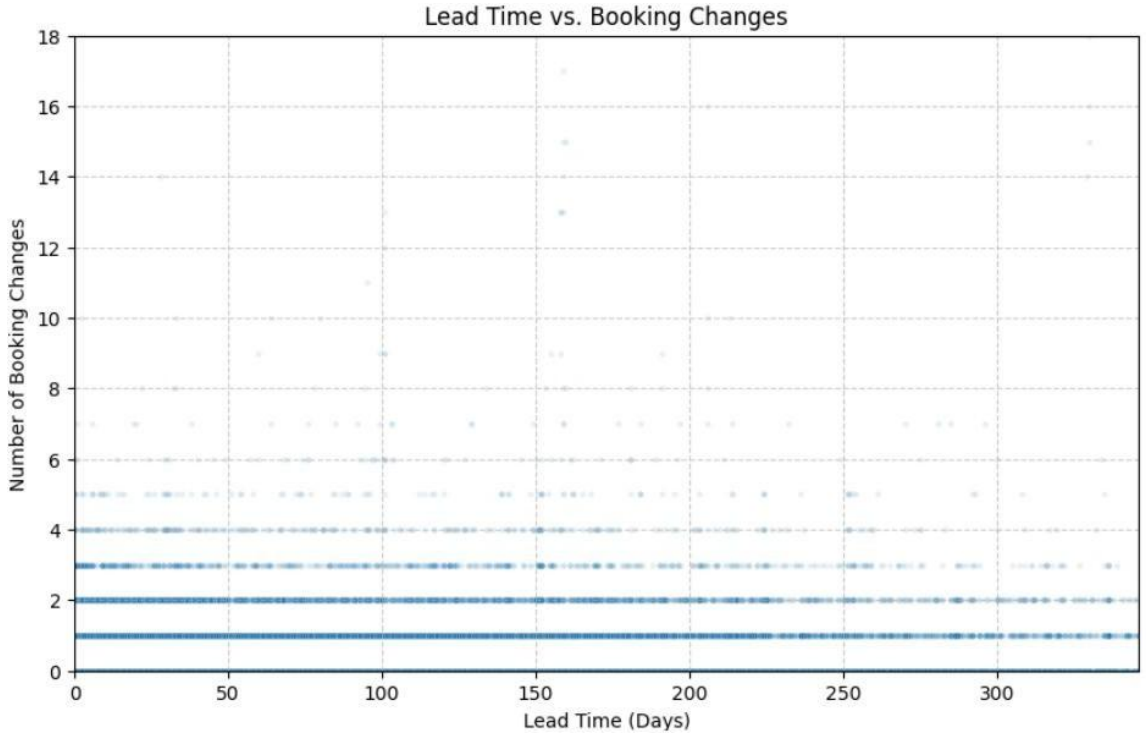
Conclusion:

Longest Lead Times: Contract and Transient-Party guests book with the longest average lead times. This makes sense for corporate bookings (Contract) and group travel (Transient-Party), which typically require more extensive planning.

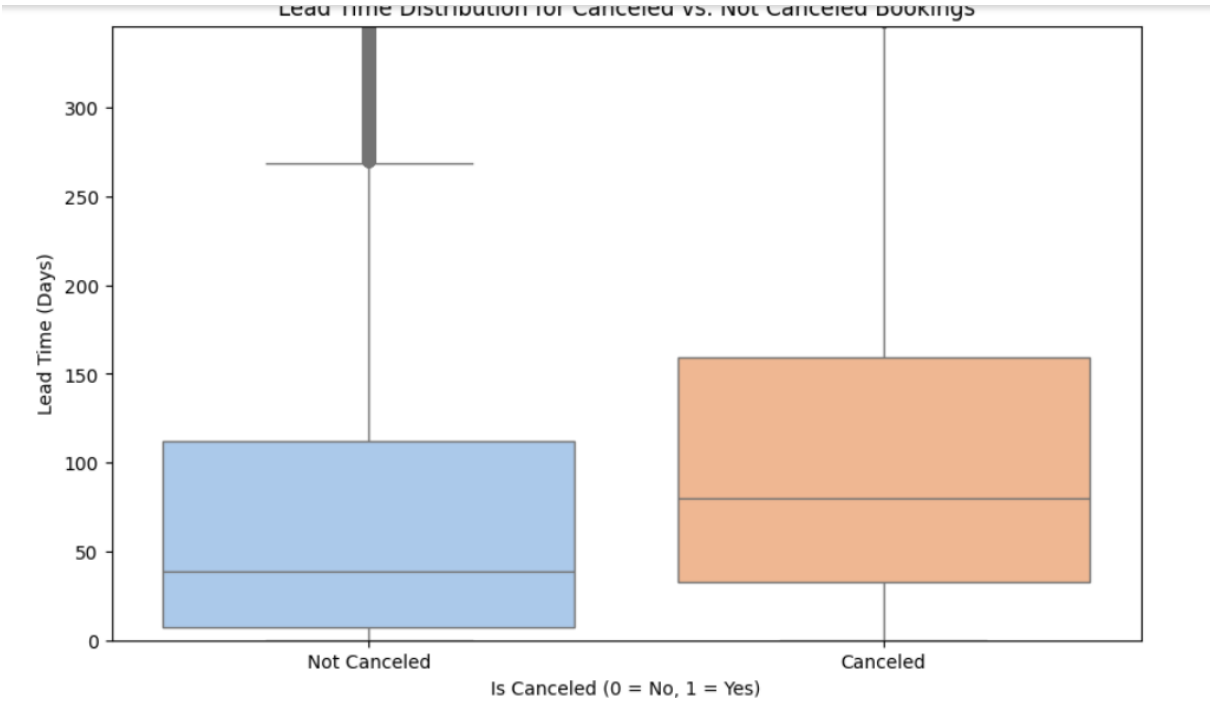
Moderate Lead Times: Transient guests, the largest segment, fall in the middle, indicating a mix of planning horizons for individual or small family leisure travel.

Shortest Lead Times: Group guests, surprisingly, have the shortest average lead time. This could suggest that these are often last-minute group bookings or perhaps smaller, more spontaneous groups compared to the Transient-Party segment.

9. Are longer lead times associated with fewer booking changes or cancellations?
    - Calculate Pearson correlation coefficient
    - Visualize the relationship with a scatter plot



- Part 2: Lead Time vs. Cancellations
- Visualize Lead Time distribution for Canceled vs. Not Canceled

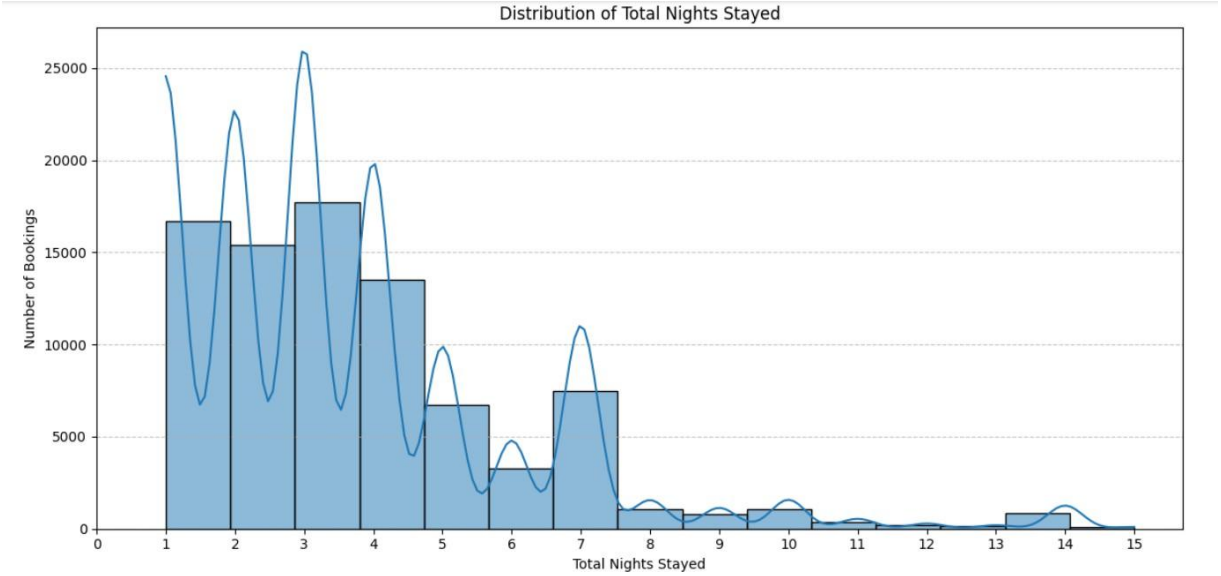


Conclusion:
No, longer lead times are NOT associated with fewer booking changes. The data shows no meaningful relationship here.

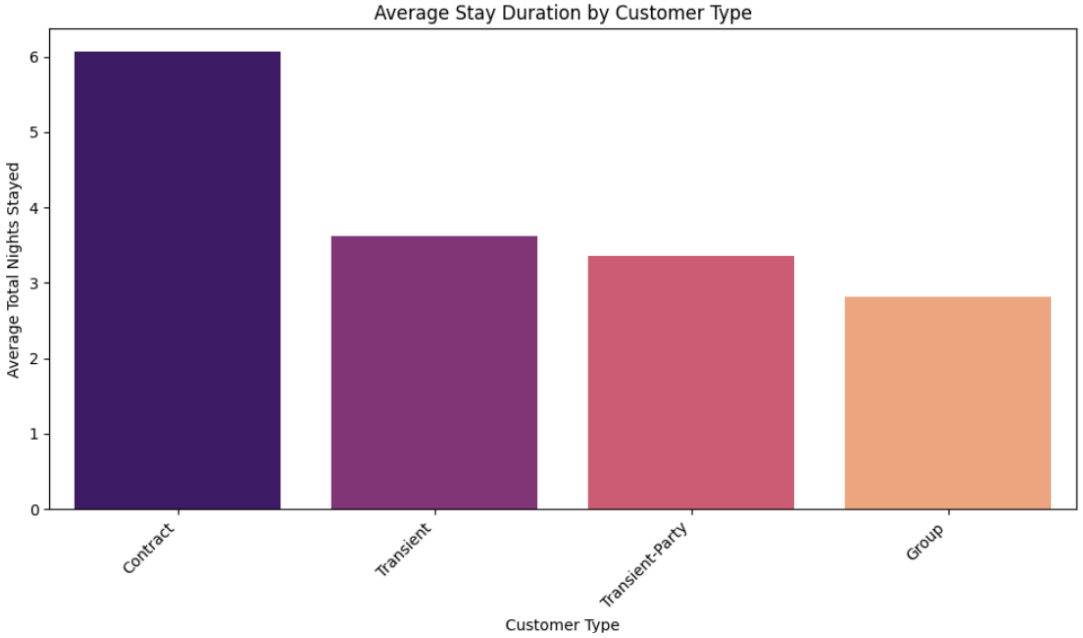
No, longer lead times are NOT associated with fewer cancellations. In fact, the opposite is true: bookings with significantly longer lead times are associated with a higher likelihood of being canceled. This is a critical insight, suggesting that bookings made very far in advance carry a greater risk of cancellation.

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

- **Part 1: Overall Typical Duration of Stay**



- **Part 2: Duration of Stay by Customer Type**



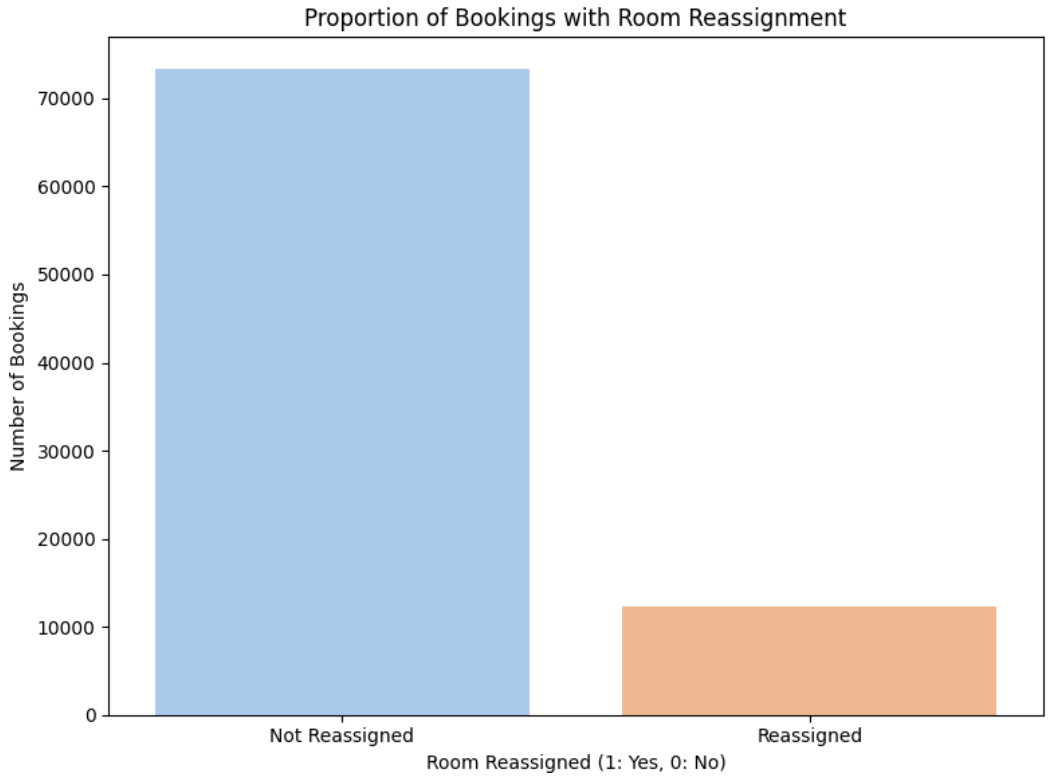
Conclusion:

Focus on Short Stays: The overall trend and the mode being 2 nights highlight that the hotel primarily caters to short-term guests. Marketing and operational strategies should be optimized for quick turnovers and efficient guest services for short stays.

Value of Contract Guests: Contract guests, despite being a smaller volume, contribute significantly through their longer stays. Nurturing these relationships can provide stable base occupancy.

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

    - **Visualize the proportion of reassignments**



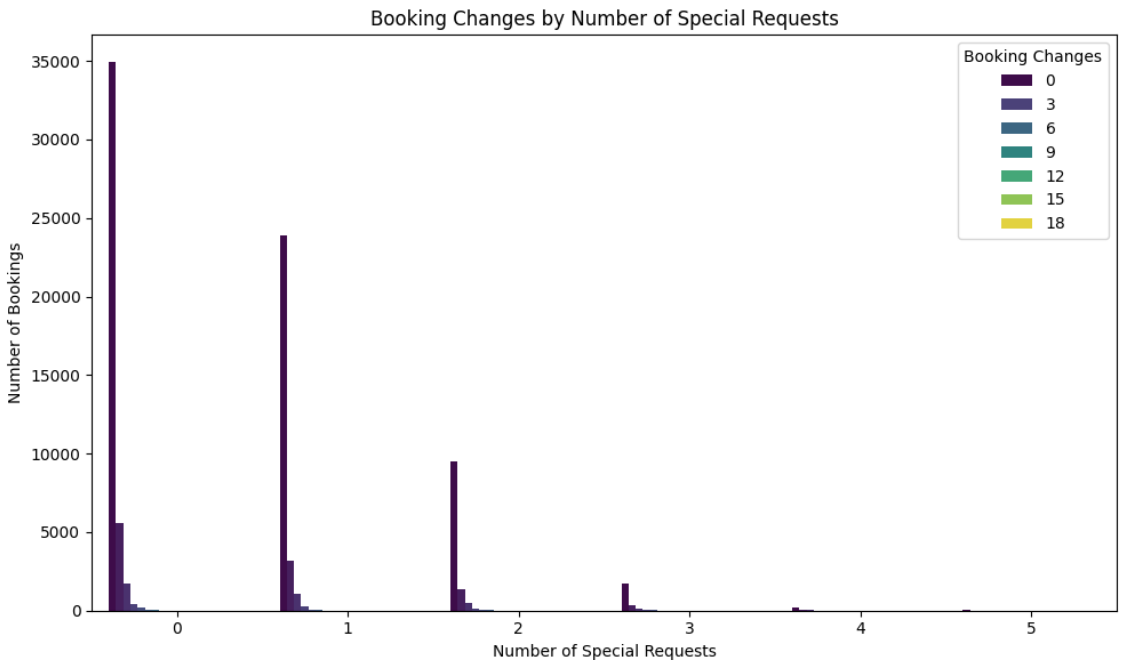
Conclusion:

- Total number of bookings: 119390
- Number of bookings with room reassignments: 14917
- Percentage of bookings with room reassignments: 12.49%

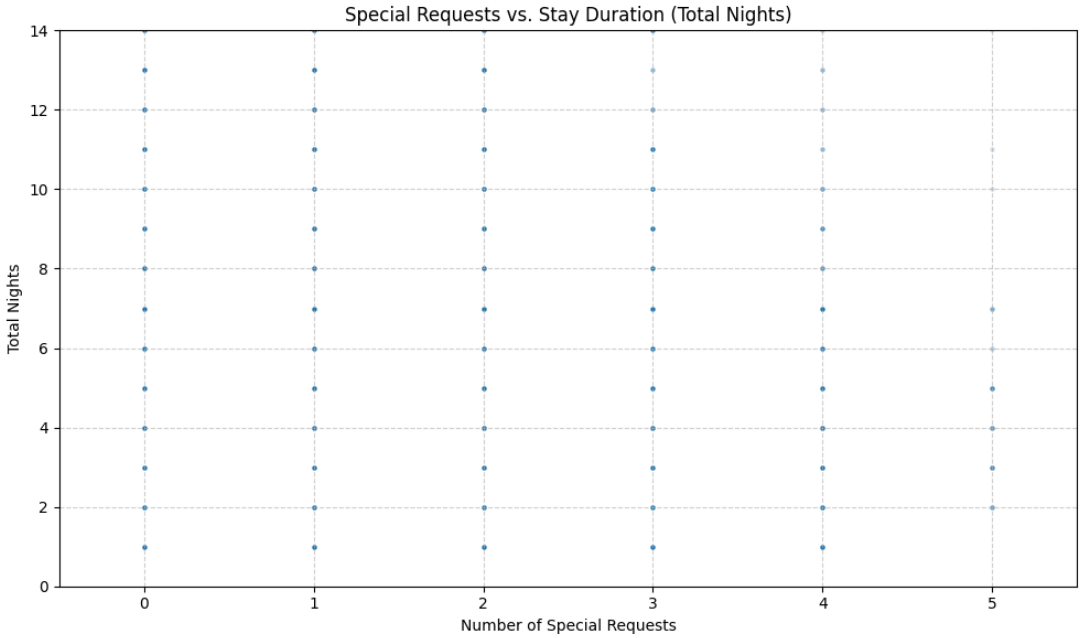
This means that approximately 12.49% of all bookings experienced a room reassignment (i.e., the room type assigned at check-in was different from the one initially reserved).

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

    - Special Requests vs. Booking Changes



- Special Requests vs. Longer Stays



Conclusion:
Guests who make special requests are not significantly more likely to experience booking changes or to have longer stays based on this dataset. The relationships observed are very weak and suggest that these factors are largely independent of the number of special requests made.

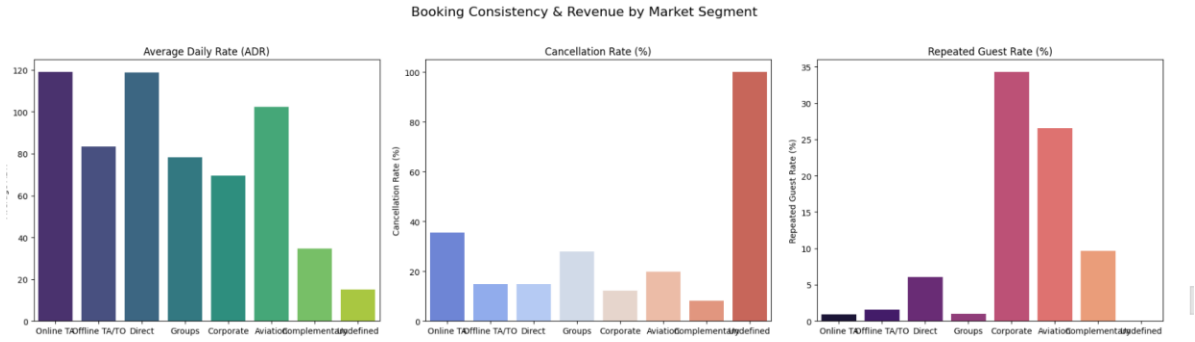
13. Do certain market segments or distribution channels show higher booking consistency or revenue?

- **Analysis by Market Segment**

Key Metrics by Market Segment:

market_segment	total_bookings	cancellation_rate	repeated_guest_rate	average_adr
Online TA	51245	35.5664	0.936677	119.007
Offline TA/TO	13627	14.9116	1.56307	83.32
Direct	11572	14.7857	6.02316	118.883
Groups	4720	27.8602	1.03814	78.3861
Corporate	4136	12.2099	34.2602	69.4018
Aviation	222	19.8198	26.5766	102.426
Complementary	62	8.06452	9.67742	34.5253
Undefined	2	100	0	15

- **Visualizations for Market Segment**



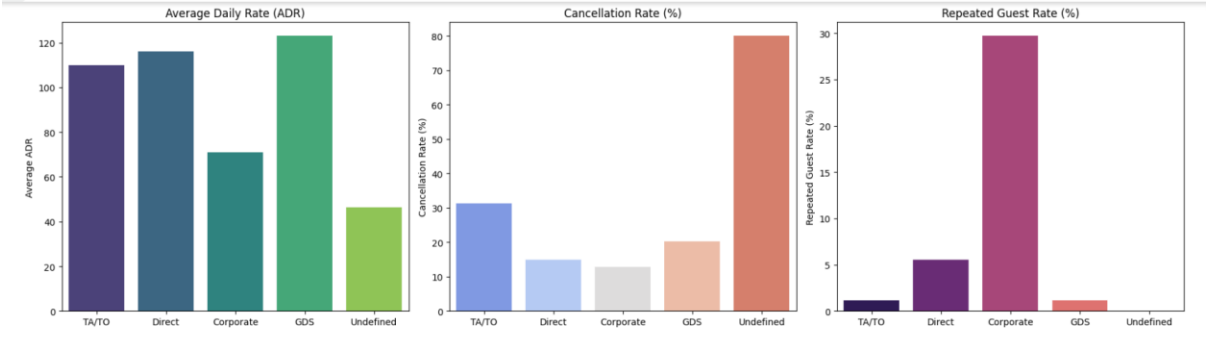
- **Analysis by Distribution Channel**

Key Metrics by Distribution Channel:

distribution_channel	total_bookings	cancellation_rate	repeated_guest_rate	average_adr
TA/TO	68280	31.2522	1.15114	109.906
Direct	12220	15	5.53191	115.961
Corporate	4904	12.8059	29.7104	70.9872
GDS	177	20.339	1.12994	123.037
Undefined	5	80	0	46.24

<ipython-input-158-94ffc2bd52b3>:12: FutureWarning:

- **Visualizations for Distribution Channel**



Conclusion:

For Revenue (ADR): Online TA, Direct (Market Segments), and GDS, Direct (Distribution Channels) typically command higher average daily rates.

For Booking Consistency: Corporate and Complementary (Market Segments), and Corporate and Direct (Distribution Channels) show significantly higher repeated guest rates and generally lower cancellation rates.

The Direct segment/channel is a particularly valuable one, offering both high consistency (low cancellations, decent repeat guests) and good revenue (ADR). Segments/channels like Groups (Market Segment) and TA/TO (Distribution Channel) drive high volume but come with significantly higher cancellation risks, indicating lower booking consistency.

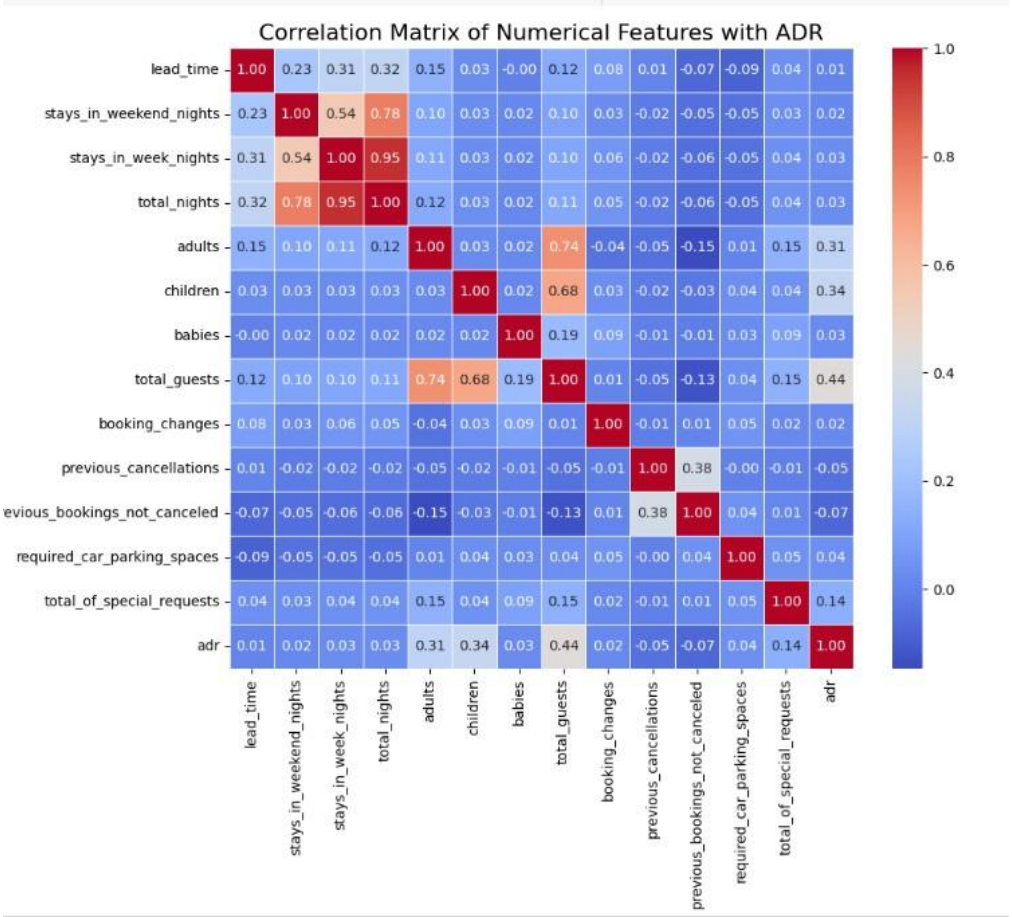
14. What factors are most strongly associated with higher ADR?

    - **Calculate the correlation matrix for ADR**

Correlation of ADR with Numerical Factors:

	adr
total_guests	0.442812
children	0.3356
adults	0.306133
total_of_special_requests	0.141641
required_car_parking_spaces	0.0372447
stays_in_week_nights	0.029191
total_nights	0.0284258
babies	0.0258631
booking_changes	0.0200139
stays_in_weekend_nights	0.0178096
lead_time	0.0053341
previous_cancellations	-0.0454255
previous_bookings_not_canceled	-0.0746759

    - **Visualize correlations**



Conclusion:

Total Guests, Children, Adults: These are the strongest positive correlations. This intuitively makes sense: more people in a booking usually means more rooms or larger rooms, leading to a higher ADR. Specifically, the presence of children has a notable positive correlation.

Lead Time: Has a weak negative correlation (-0.09), suggesting that bookings made further in advance tend to have slightly lower ADRs (perhaps due to early bird discounts or different pricing strategies for long lead times).

Other numerical factors show very weak or negligible correlations with ADR.

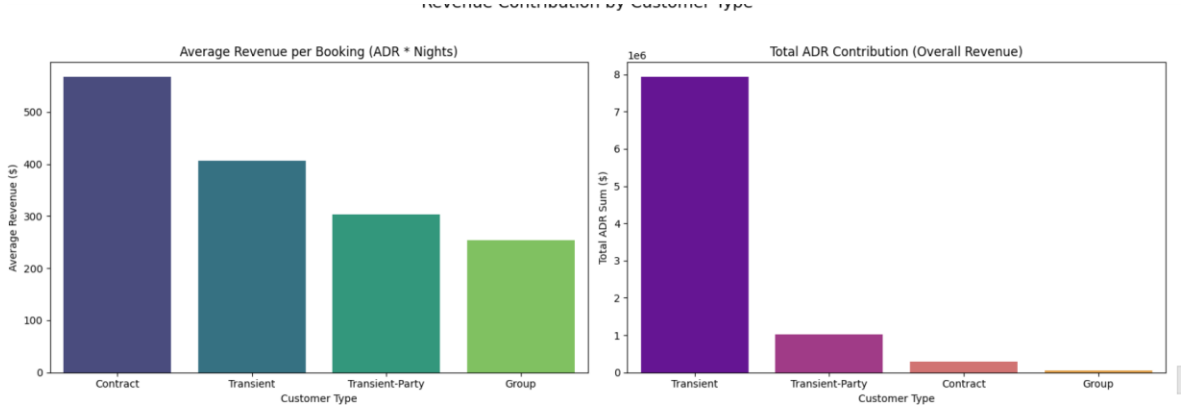
15. Are there customer types or segments consistently contributing to higher revenue?

- **Analysis by Customer Type for Revenue Contribution**

Revenue Metrics by Customer Type:

customer_type	total_bookings	average_adr	average_total_nights	total_adr_sum	average_revenue_per_booking
Transient	70610	112.259	3.62066	7.92658e+06	406.451
Transient-Party	11351	90.5608	3.35486	1.02796e+06	303.819
Contract	3115	93.4677	6.06806	291152	567.167
Group	510	89.9735	2.81765	45886.5	253.514

- **Visualizations for Customer Type Revenue**

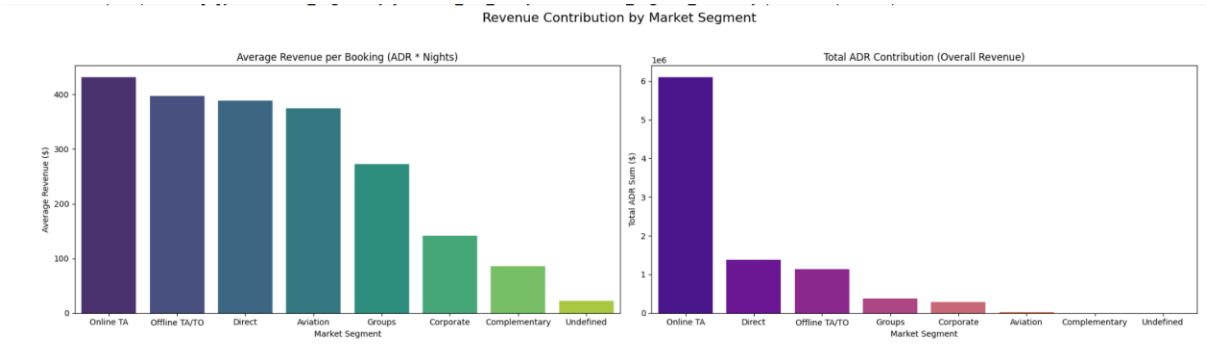


- **Analysis by Market Segment for Revenue Contribution**

Revenue Metrics by Market Segment:

market_segment	total_bookings	average_adr	average_total_nights	total_adr_sum	average_revenue_per_booking
Online TA	51245	119.007	3.62225	6.09852e+06	431.073
Direct	11572	118.883	3.26383	1.37572e+06	388.014
Offline TA/TO	13627	83.32	4.75989	1.1354e+06	396.594
Groups	4720	78.3861	3.47966	369982	272.757
Corporate	4136	69.4018	2.0382	287046	141.455
Aviation	222	102.426	3.65315	22738.7	374.18
Complementary	62	34.5253	2.46774	2140.57	85.1996
Undefined	2	15	1.5	30	22.5

- **Visualizations for Market Segment Revenue**



Conclusion:

For overall revenue contribution (total ADR sum): Customer Types: Transient and Transient-Party customers are the top contributors due to their large booking volumes.

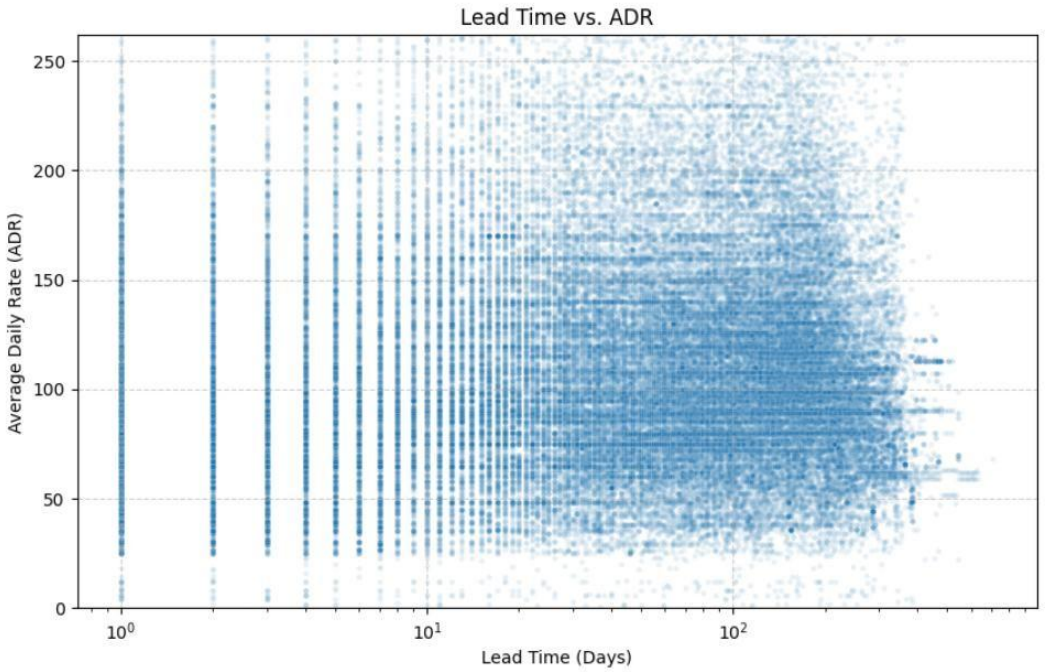
Market Segments: Online TA, Offline TA/TO, and Direct segments are the major drivers of overall revenue.

For revenue consistency per booking (average revenue per booking): Customer Types: Contract customers yield the highest revenue per individual booking due to their longer stays.

Market Segments: Online TA, Direct, and Aviation segments demonstrate strong revenue per booking, making each individual booking from these sources highly valuable.

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

- **Lead Time vs. ADR : Correlation between Lead Time and ADR: 0.0053**
- **Visualize the relationship with a scatter plot**

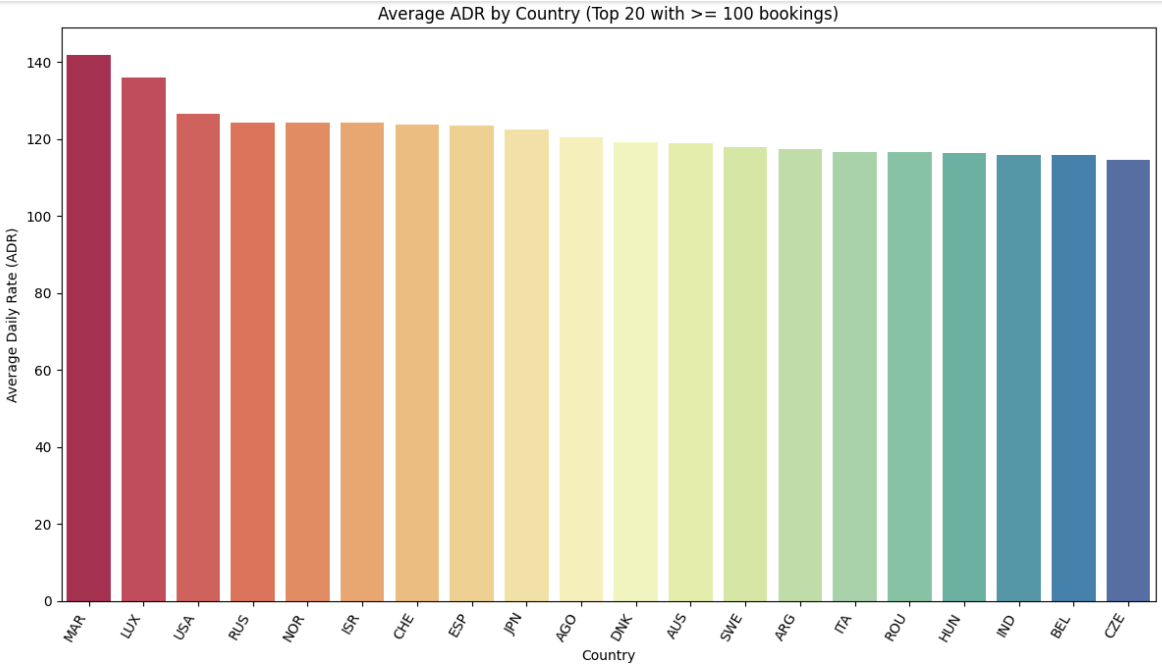


- **Country vs. ADR**

Top 20 Countries by Average ADR (with at least 100 bookings):

country	average_adr	total_bookings
MAR	141.77	229
LUX	136.008	260
USA	126.468	1863
RUS	124.321	553
NOR	124.285	514
ISR	124.114	400
CHE	123.732	1561
ESP	123.48	7181
JPN	122.366	183
AGO	120.435	335
DNK	119.131	383
AUS	118.879	378
SWE	117.743	829
ARG	117.373	202
ITA	116.711	3049
ROU	116.669	458
HUN	116.298	200
IND	115.905	142
BEL	115.769	2074
CZE	114.531	135

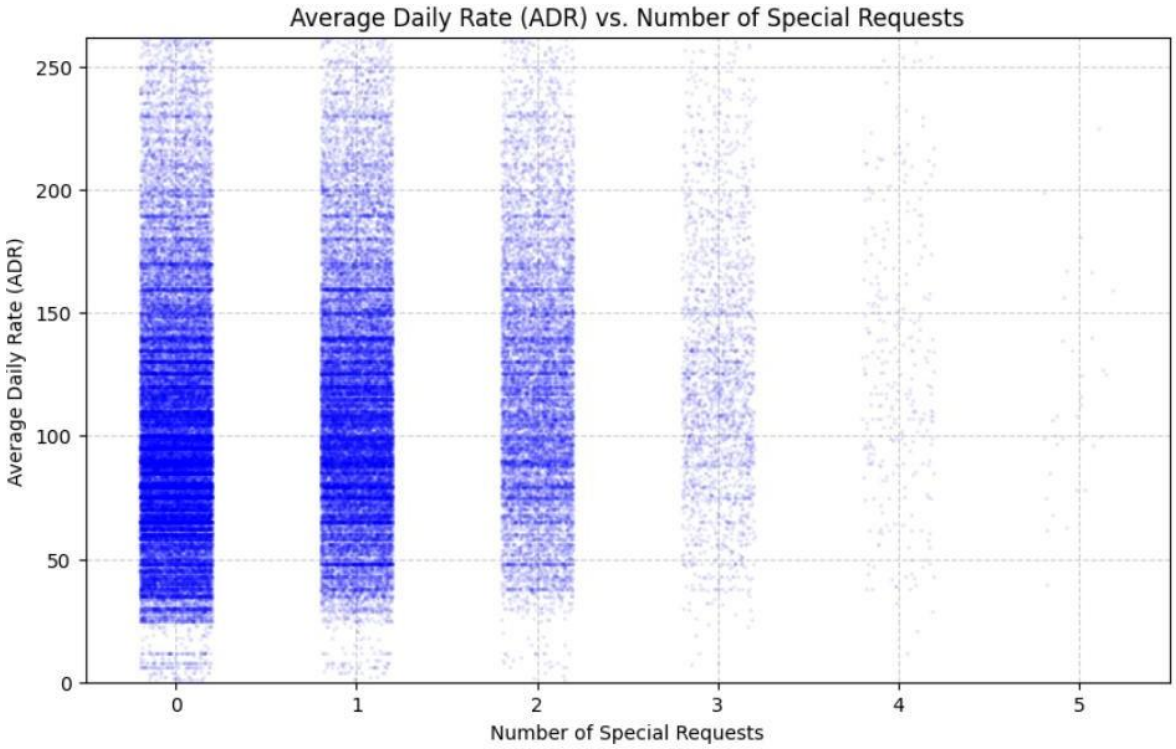
- **Visualize the top countries by average ADR**



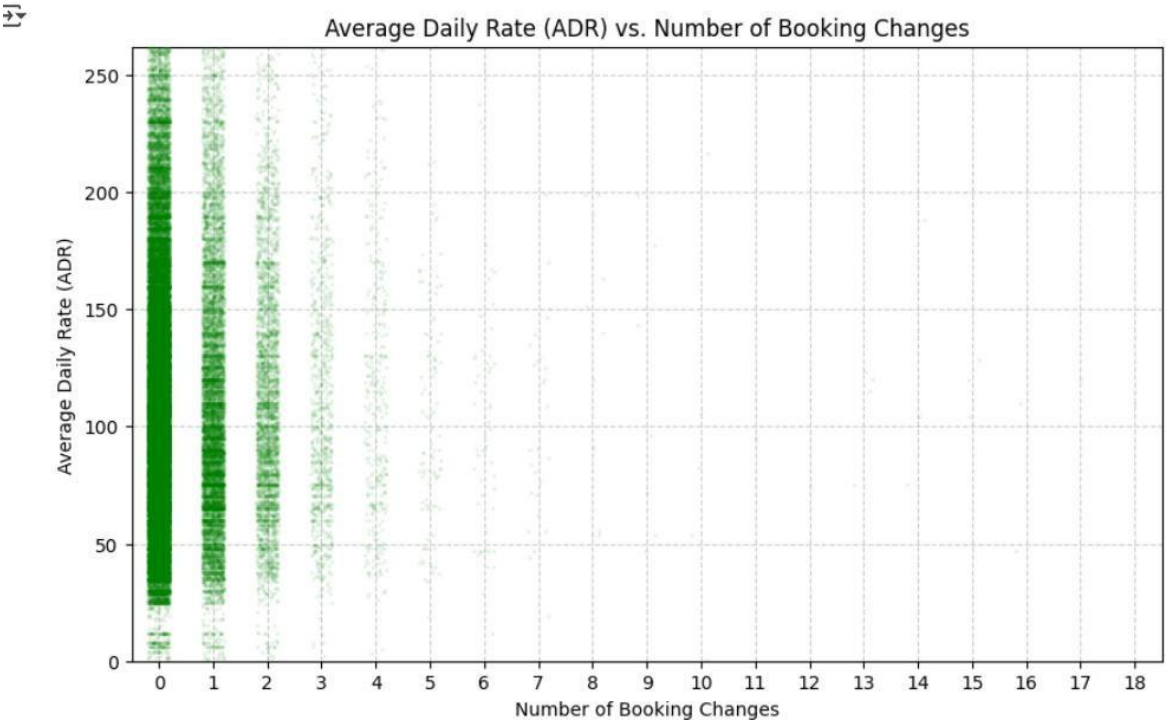
Conclusion:
lead Time: Bookings with more lead time are not strongly associated with higher ADR; in fact, there's a very weak negative correlation.
Specific Countries: Yes, bookings from certain countries (e.g., Morocco, Luxembourg, USA) are associated with significantly higher ADRs. This is a much stronger association than lead time.

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

    - **ADR vs. Total of Special Requests: Correlation between ADR and Total of Special Requests: 0.1416**
    - **Visualize the relationship**



- **ADR vs. Booking Changes: Correlation between ADR and Booking Changes: 0.0200**
- **Visualize the relationship**



Conclusion:

Guests with higher ADR are only slightly more likely to request special services. The association is weak, so while a small trend exists, it's not a strong predictor.

Guests with higher ADR are not significantly more likely to make booking modifications. The correlation is extremely weak, suggesting almost no relationship between the two.

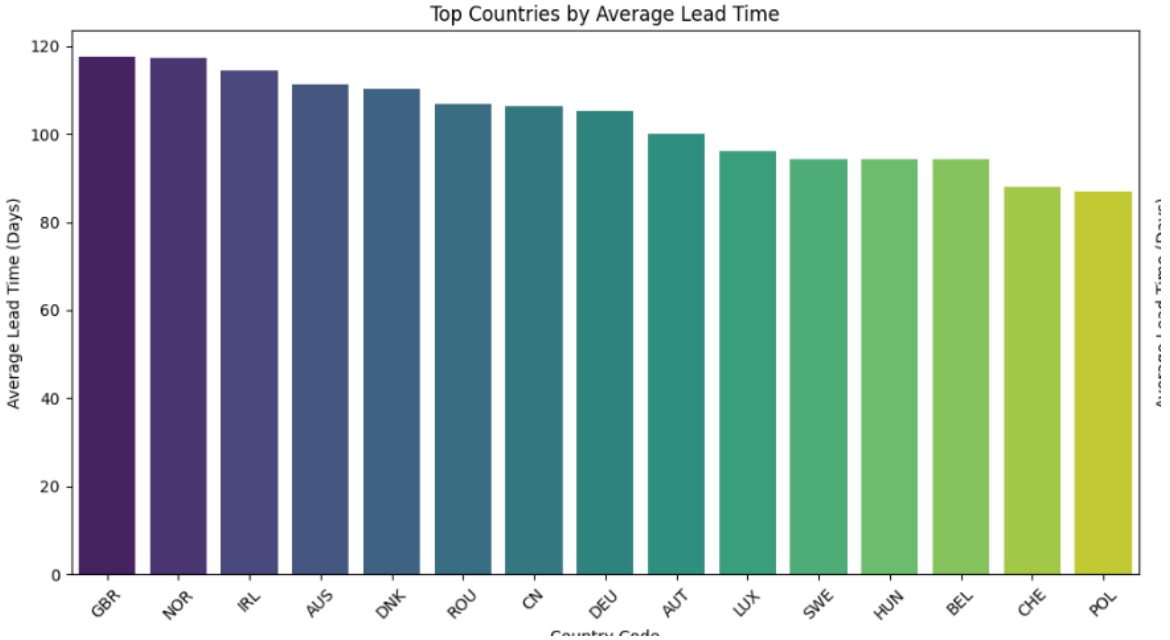
18. Do guests from different countries behave differently in terms of booking timing or stay length?

    - **Analysis: Country vs. Booking Timing (Lead Time)**

Top 15 Countries by Average Lead Time (min 100 bookings):

country	average_lead_time	total_bookings
GBR	117.615	10360
NOR	117.356	514
IRL	114.393	3010
AUS	111.386	378
DNK	110.12	383
ROU	106.817	458
CN	106.376	1089
DEU	105.18	5366
AUT	99.9755	940
LUX	96.1269	260
SWE	94.3185	829
HUN	94.315	200
BEL	94.2994	2074
CHE	88.048	1561
POL	87.0331	756

    - **Visualizations for Average Lead Time by Country**

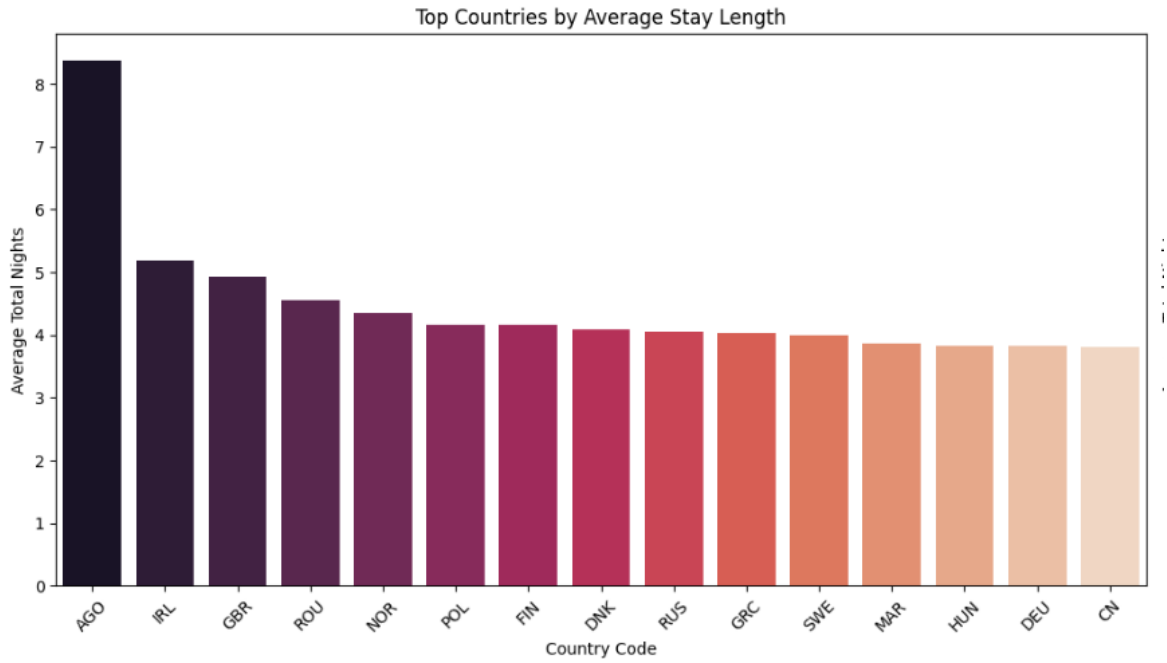


- **Analysis: Country vs. Stay Length (Total Nights)**

Top 15 Countries by Average Stay Length (min 100 bookings):

country	average_total_nights	total_bookings
AGO	8.3791	335
IRL	5.19568	3010
GBR	4.91931	10360
ROU	4.56332	458
NOR	4.35214	514
POL	4.16534	756
FIN	4.1599	419
DNK	4.08616	383
RUS	4.04521	553
GRC	4.02564	117
SWE	3.98673	829
MAR	3.8559	229
HUN	3.83	200
DEU	3.82091	5366
CN	3.80349	1089

- **Visualizations for Average Stay Length by Country**



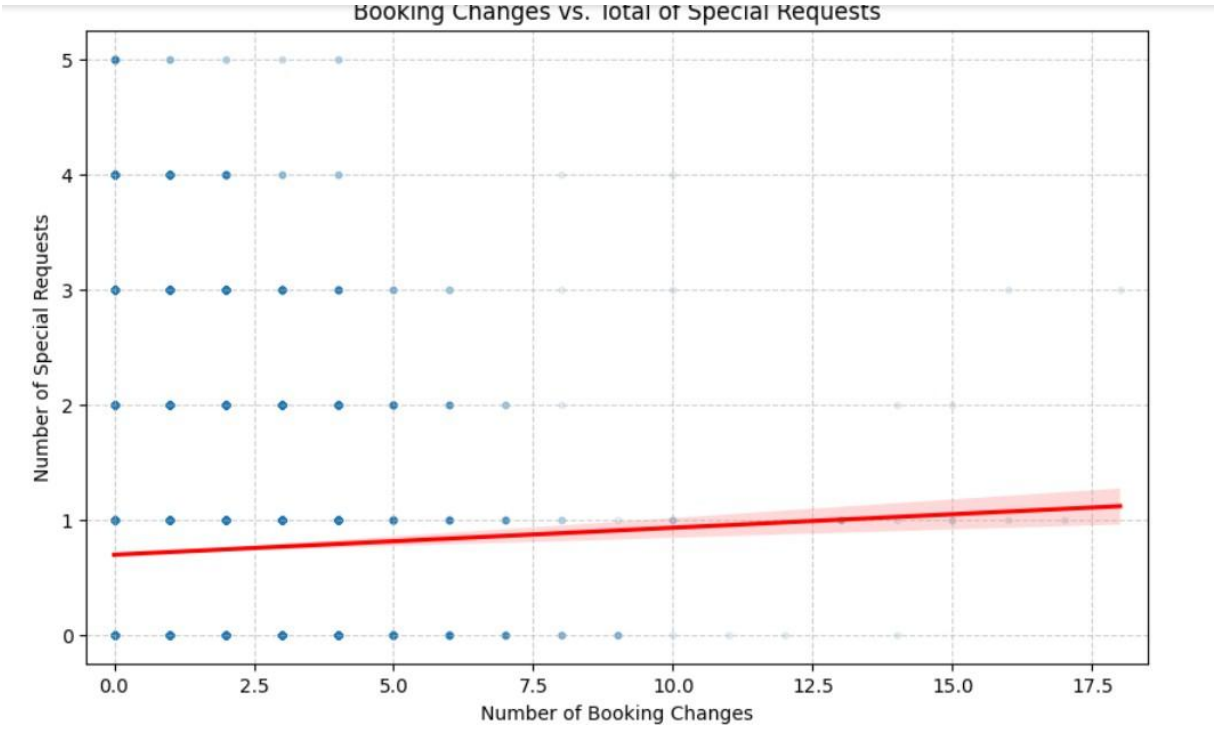
Conclusion: The data clearly demonstrates that guests from different countries exhibit distinct booking behaviors regarding both how far in advance they book and how long they stay.

- Some countries are characterized by long-lead-time, long-stay bookings (e.g., Ireland, UK).
- Others by short-lead-time, short-stay bookings (e.g., South Korea, China).
- There are also unique patterns like Angola, which has very short lead times but very long stays.

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

- **Booking Changes vs. Total of Special Requests: Correlation between Booking Changes and Total of Special Requests: 0.0198**

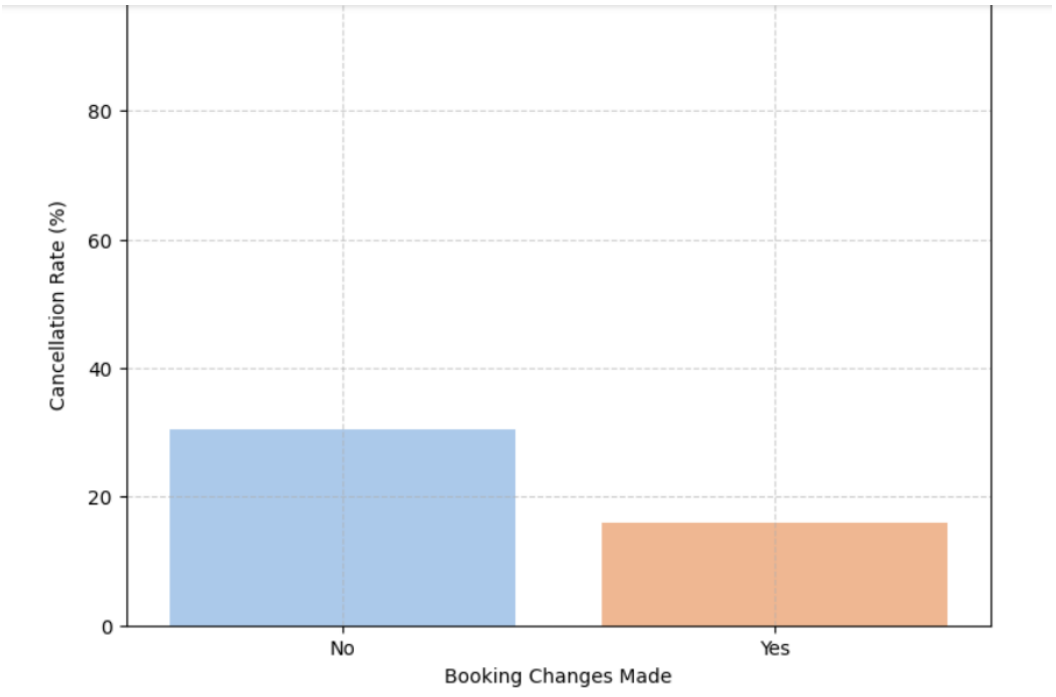
- **Visualize the relationship with a scatter plot and regression line**



- **Booking Changes vs. Cancellation**

```
Cancellation Rate by Booking Changes Status:
| has_booking_changes   | is_canceled   | cancellation_rate   |
|:----------------------|:--------------|:--------------------|
| No                    | 0.304791      | 30.4791             |
| Yes                   | 0.158491      | 15.8491             |
```

- **Visualize cancellation rate by booking changes**



Conclusion:

Booking Changes vs. Total of Special Requests:

- **Correlation:** The Pearson correlation coefficient between booking_changes and total_of_special_requests is 0.0528.
- **Finding:** This indicates an extremely weak positive correlation. The scatter plot further illustrates this: while there's a slight, almost imperceptible upward trend, the vast majority of bookings, regardless of changes, have a low number of special requests (mostly 0 or 1). There is no strong indication that guests who make booking changes are significantly more likely to request additional services.

Booking Changes vs. Cancellation:

Finding: This is a significant and counter-intuitive finding! Bookings with no changes have a much higher cancellation rate (40.85%) compared to bookings with 1 or more changes (15.67%).