Exploratory Data Analysis (EDA) Report on Hotel Bookings Dataset

Name : Omkar Shingade

Roll no : 250240325044

1. Introduction

This report presents an exploratory data analysis (EDA) of the **Hotel Bookings Dataset**, which contains information about hotel reservations, cancellations, customer demographics, and booking trends. The dataset includes **119,390 entries** with **32 features**, covering aspects such as lead time, stay duration, booking channels, and revenue metrics (ADR - Average Daily Rate).

Core Objectives

1. **Understand customer attributes and booking behaviors** impacting revenue.
2. **Identify trends** in lead time, stay duration, and booking channels.
3. **Detect inconsistencies or anomalies** in room allocation and guest handling.
4. **Explore relationships** between booking patterns and customer satisfaction indicators.
5. **Evaluate operational or customer variables** affecting outcomes like ADR or room upgrades.

2. Data Loading and Initial Exploration

Commands

- import numpy as np
- import pandas as pd
- import matplotlib.pyplot as plt
- import seaborn as sns
- df = pd.read_csv('hotel_bookings.csv')
- df.head()
- df.shape
- df.info()
- df.describe()

df.head() - Displays the first 5 rows to understand the dataset structure.

df.shape() - Shows dataset dimensions (rows, columns).

df.info() - Provides data types and missing values.

df.describe() - Gives statistical summaries (mean, min, max, etc.).

3. Data Cleaning and Preprocessing

- df = df.drop(columns='company') # 94% missing → irrelevant
- df['children'] = df['children'].fillna(0) # Replace with 0 (no children)
- df['country'] = df['country'].fillna(df['country'].mode()[0]) # Fill with most frequent country
- df['agent'] = df['agent'].fillna(0) # Replace missing agent IDs with 0

- company **column dropped** → Too many missing values (94.3%).
- children **filled with 0** → Assumes no children if data is missing.
- country **filled with mode** → Retains data distribution.
- agent **filled with 0** → Represents bookings without an agent.

Removing Duplicates

df.duplicated().sum() # Check duplicates (32,020 found)

df = df.drop_duplicates() # Remove duplicates

df.duplicated().sum() # Confirm removal (0 duplicates)

Duplicates can bias analysis → Removed **32,020 duplicate rows**.

Feature Engineering

```
# Convert date columns to datetime
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', errors='coerce')


# Drop redundant columns
df = df.drop(columns=['arrival_date_year', 'arrival_date_month', 'arrival_date_day_of_month', 'arrival_date_week_number'])


# Create new features
df['total_stays'] = df['stays_in_weekend_nights'] + df['stays_in_week_nights']
df['total_guests'] = (df['adults'] + df['children'].fillna(0) + df['babies']).astype(int)
df = df.drop(columns=['stays_in_weekend_nights', 'stays_in_week_nights', 'adults', 'children', 'babies'])
```

- arrival_date → Combines year, month, and day for easier analysis.
- total_stays → Sum of weekend and weekday nights.
- total_guests → Total guests per booking (adults + children + babies).

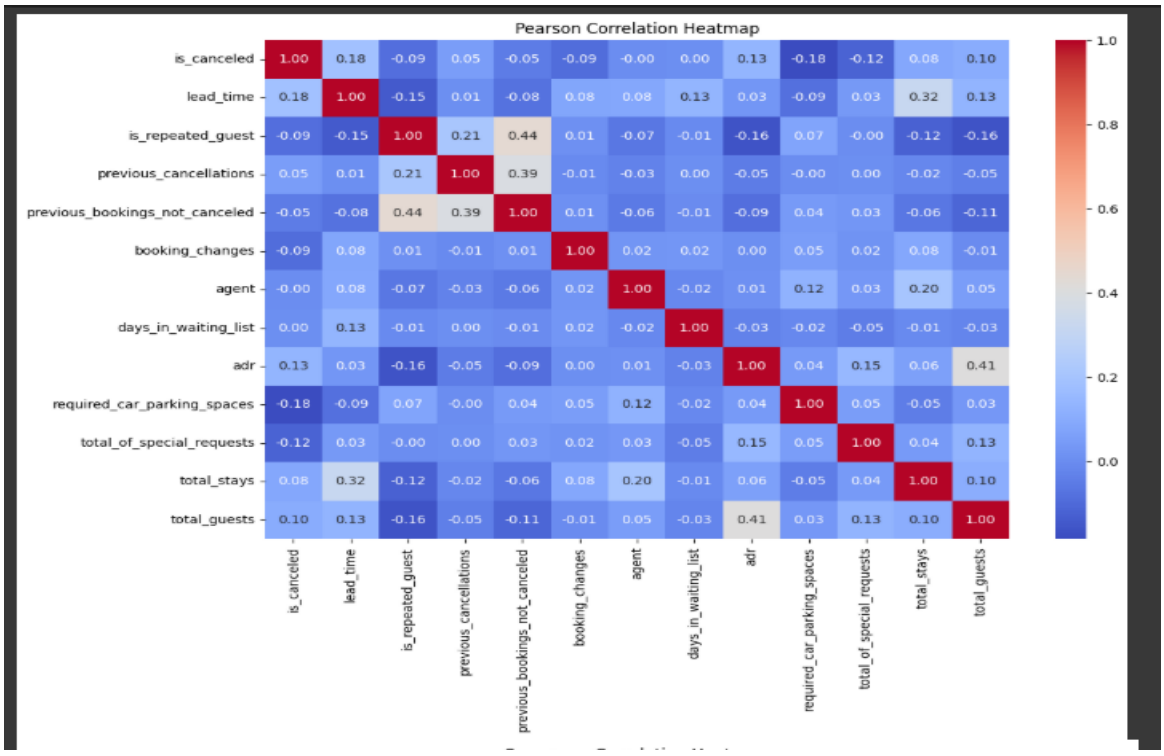
4. Exploratory Data Analysis (EDA)

- This step focused on understanding the distribution, relationships, and time-based trends in the dataset.

- Univariate analysis helped understand the frequency and distribution of individual features like 'adr', 'lead_time', 'customer_type', etc.
- Bivariate analysis using boxplots and scatterplots revealed how factors like hotel type, market segment, and customer type impact ADR.
- Multivariate analysis combined multiple features to uncover more complex patterns (e.g., ADR by hotel and customer type together).
- Time series plots highlighted booking trends across months and seasons.

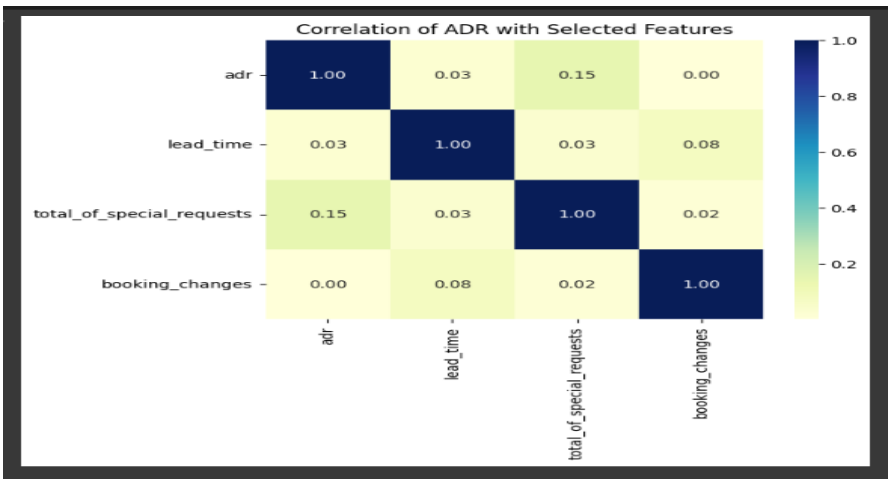
5. Correlation Analysis

- A Pearson correlation matrix was computed to detect how strongly numerical variables relate to each other.



- A full or enhanced correlation matrix that includes strong, clear visualizations, color-coded relationships, and insights across many features — like a “super” version of a regular correlation matrix.

- 'adr' showed positive correlation with variables like 'lead_time', 'total_guests', and 'special_requests'.
- The correlation heatmap visually showed these relationships, helping to identify multicollinearity and influential variables.



6. Hypothesis Testing

- To validate business assumptions statistically, three hypotheses were tested:
- Difference in ADR between Online TA and Direct channels – tested using Welch's t-test, which found a statistically significant difference.
- Relationship between room upgrades and lead time – also tested using Welch's t-test, showing lead time impacts upgrades.
- Variation in stay duration across customer types – tested using one-way ANOVA, confirming that different customer types stay for different lengths.
- Result: All three null hypotheses were rejected ($p < 0.05$), proving significant associations in the data.

7. Key Business Questions Answered

- What influences ADR the most?

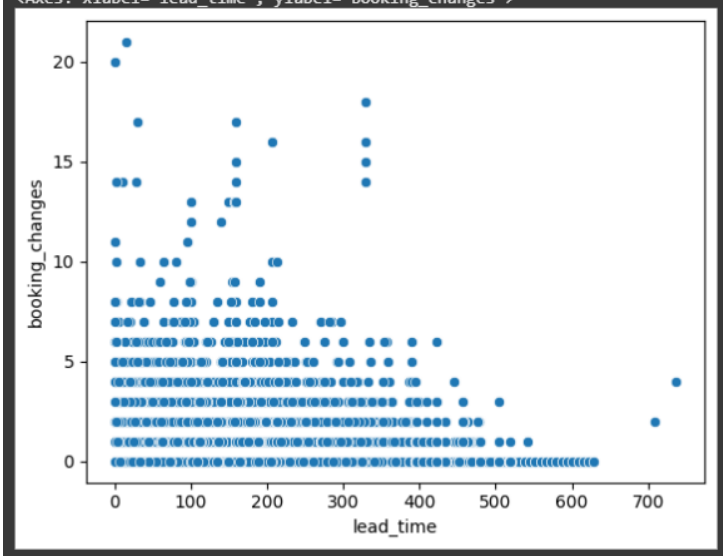
```
1 # 1. What influences ADR the most?
2 corr_matrix = df.corr(numeric_only=True)
3 sns.heatmap(corr_matrix[['adr']].sort_values(by='adr', ascending=False), annot=True)
```

```
<Axes: >
```

- Do guests who book earlier tend to request more changes?

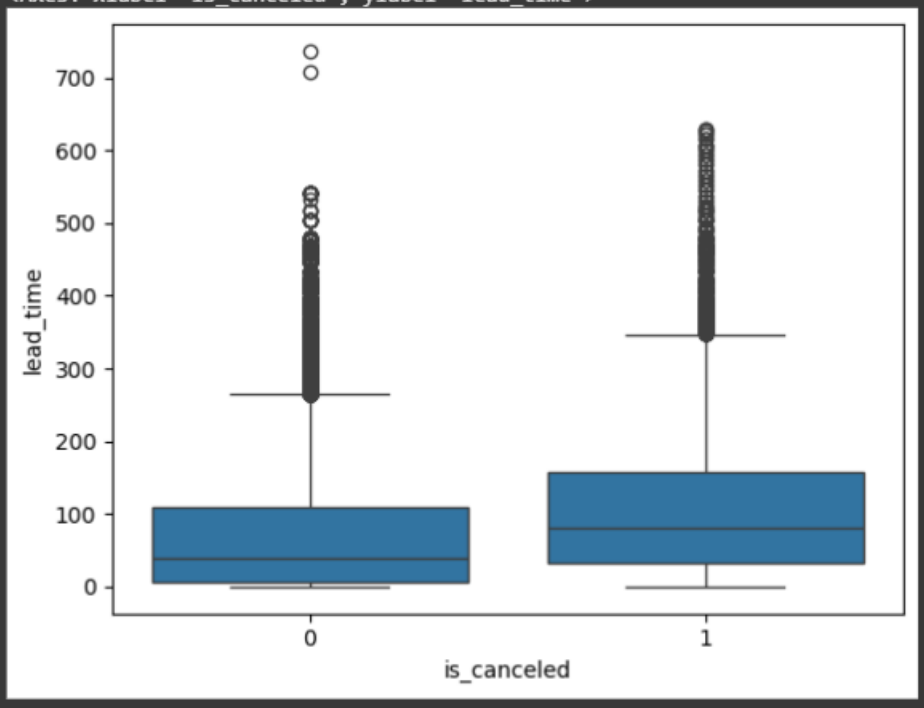
```
1 # 2. Do guests who book earlier tend to request more changes?
2 sns.scatterplot(data=df, x='lead_time', y='booking_changes')
3 # Very weak negative correlation – guests who book earlier do not significantly tend to make more changes.
```

```
<Axes: xlabel='lead_time', ylabel='booking_changes'>
```



```
1 sns.boxplot(x='is_canceled', y='lead_time', data=df)
```

```
<Axes: xlabel='is_canceled', ylabel='lead_time'>
```



- Are there pricing or booking differences across countries?

```
1 # 3. Are there pricing or booking differences across countries?
2 df.groupby('country')['adr'].mean().sort_values(ascending=False).head(10)
```

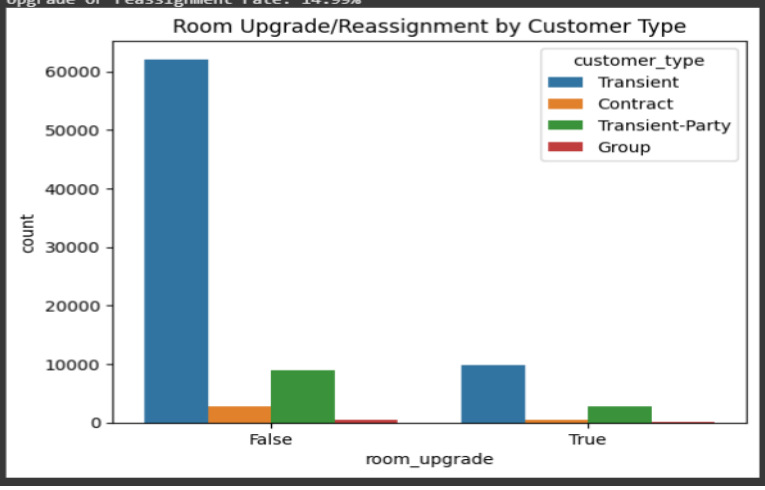
country	adr
DJI	273.000000
AIA	265.000000
AND	202.652857
UMI	200.000000
LAO	181.665000
MYT	177.750000
NCL	175.500000
GIB	169.082667
FRO	165.666667
COM	165.305000

dtype: float64

- Is there a pattern in room upgrades or reassignment?

```
# 4. Is there a pattern in room upgrades or reassignment?
# Create a column if upgrade happened (assigned room better than reserved)
df['room_upgrade'] = df['assigned_room_type'] != df['reserved_room_type']

upgrade_rate = df['room_upgrade'].mean()
print(f'Upgrade or reassignment rate: {upgrade_rate:.2%}')

# Upgrade by customer type
sns.countplot(x='room_upgrade', hue='customer_type', data=df)
plt.title('Room Upgrade/Reassignment by Customer Type')
plt.show()
# Only 14.99% of guests are assigned a different room type than the one they originally reserved.
```

```
Upgrade or reassignment rate: 14.99%
```



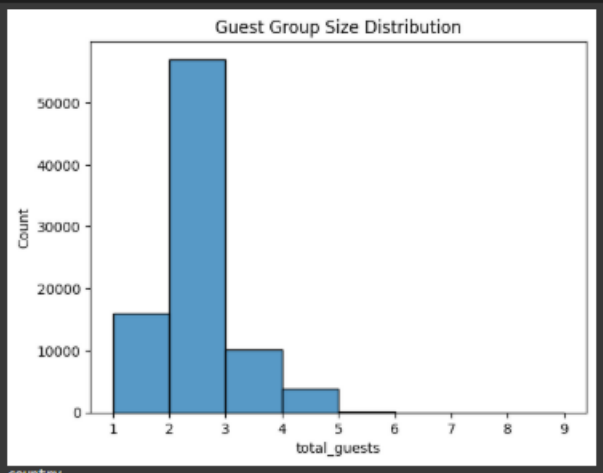
- Are reserved room types consistently matched with assigned room types?

```
# 5. Are reserved room types consistently matched with assigned room types?
match_rate = (df['reserved_room_type'] == df['assigned_room_type']).mean()
print(f'Reserved and assigned room match rate: {match_rate:.2%}')
# 84.9% of guests receive the exact room type they originally reserved.
```

```
Reserved and assigned room match rate: 85.01%
```

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

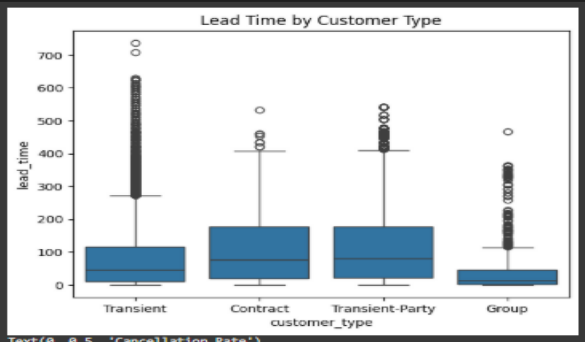
```python
# 6. What are the most common guest demographics (e.g., group size, nationality)?
# Group size
# df['group_size'] = df['adults'] + df['children'] + df['babies']
sns.histplot(df['total_guests'], bins=range(1,10))
plt.title('Guest Group Size Distribution')
plt.show()

# Nationality top 10
print(df['country'].value_counts().head(10))


# The guest demographic analysis shows that most hotel bookings come from small groups, mainly couples and solo
# travelers. Large group bookings are rare. Portuguese guests form the largest customer base, followed by visitors
# from the UK, France, Spain, Germany, and Italy. This indicates a strong Western European presence with some
# international reach. The hotel should focus its services and marketing on small-group European travelers,
# especially couples.
```



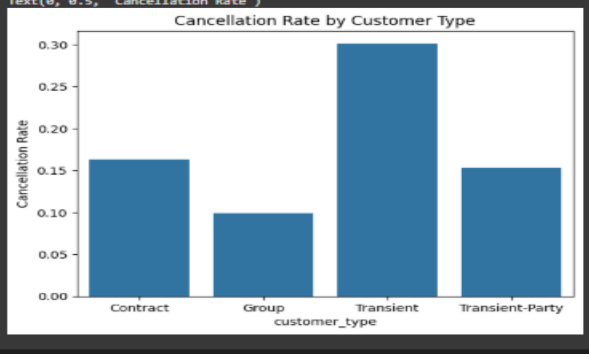

```
country
PRT    27882
GBR    10432
FRA     8837
ESP     7252
DEU     5387
ITA     3066
IRL     3016
BEL     2081
BRA     1995
NLD     1911
Name: count, dtype: int64
```

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

```
# 7. Are there patterns in guest types (e.g., transient vs. corporate) that influence booking behavior?
sns.boxplot(data=df, x='customer_type', y='lead_time')
plt.title('Lead Time by Customer Type')
plt.show()

cancel_rates = df.groupby('customer_type')['is_canceled'].mean().reset_index()
sns.barplot(data=cancel_rates, x='customer_type', y='is_canceled')
plt.title('Cancellation Rate by Customer Type')
plt.ylabel('Cancellation Rate')
```

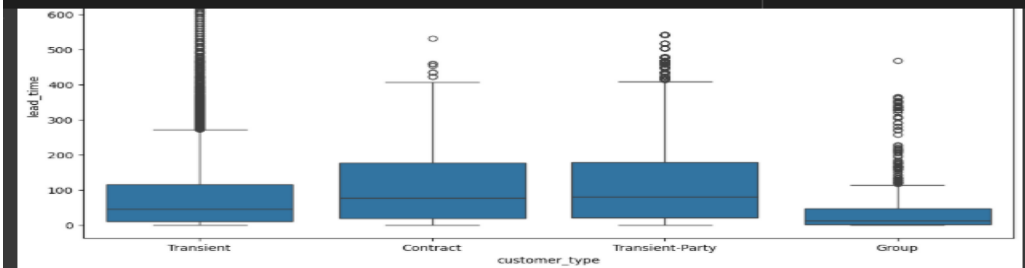


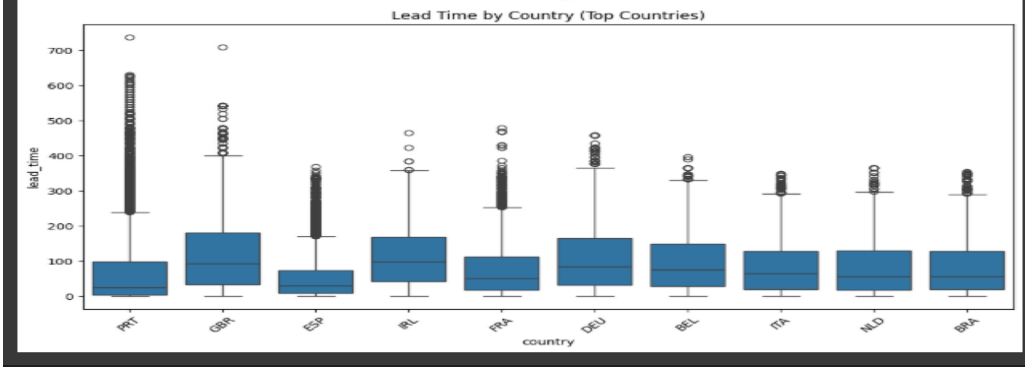

```
Text(0, 0.5, 'Cancellation Rate')
```



- How does booking lead time vary across customer types and countries?

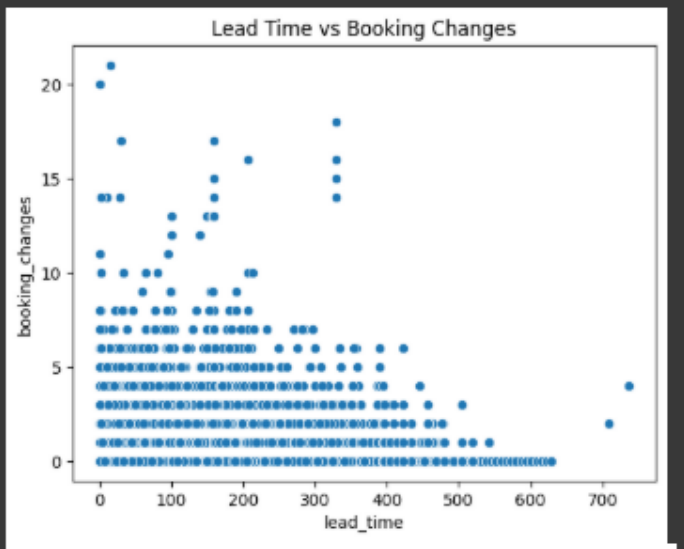
```
# 8. How does booking lead time vary across customer types and countries?
plt.figure(figsize=(12,6))
sns.boxplot(x='customer_type', y='lead_time', data=df)
plt.title('Lead Time by Customer Type')
plt.show()

plt.figure(figsize=(12,6))
sns.boxplot(x='country', y='lead_time', data=df[df['country'].isin(top_countries)])
plt.title('Lead Time by Country (Top Countries)')
plt.xticks(rotation=45)
plt.show()
```

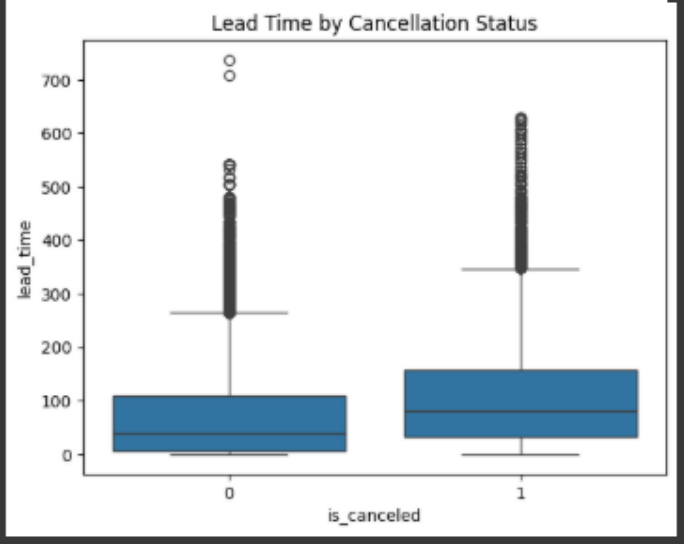




- Are longer lead times associated with fewer booking changes or cancellations?

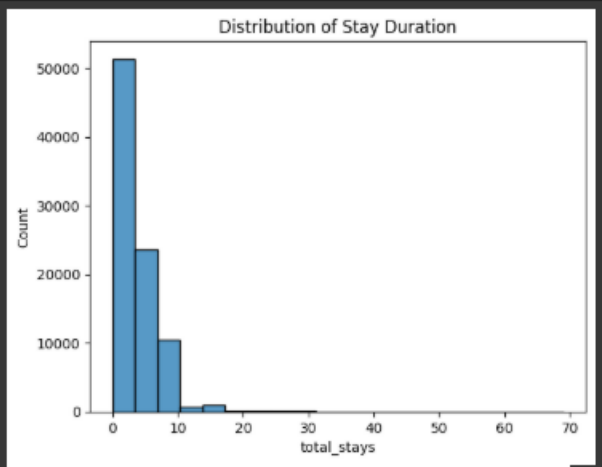
```
[ ]  1 # 9. Are longer lead times associated with fewer booking changes or cancellations?
     2 sns.scatterplot(x='lead_time', y='booking_changes', data=df)
     3 plt.title('Lead Time vs Booking Changes')
     4 plt.show()
     5
     6 sns.boxplot(x='is_canceled', y='lead_time', data=df)
     7 plt.title('Lead Time by Cancellation Status')
     8 plt.show()
```

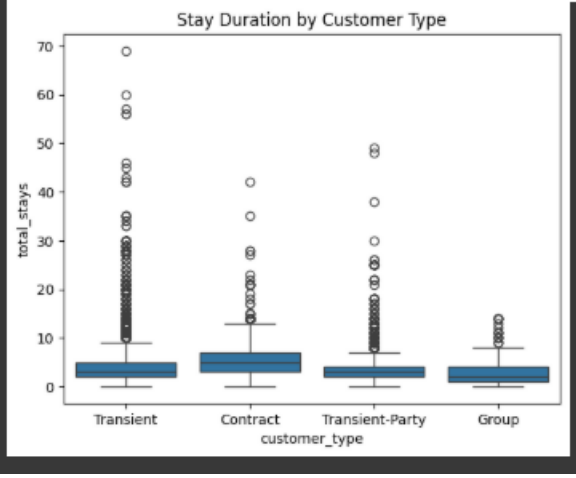




- What is the typical duration of stay, and how does it vary by customer type or segment?

```python
# 10. What is the typical duration of stay, and how does it vary by customer type or segment?
# df['total_stays'] = df['stays_in_weekend_nights'] + df['stays_in_week_nights']
sns.histplot(df['total_stays'], bins=20)
plt.title('Distribution of Stay Duration')
plt.show()

sns.boxplot(x='customer_type', y='total_stays', data=df)
plt.title('Stay Duration by Customer Type')
plt.show()
```





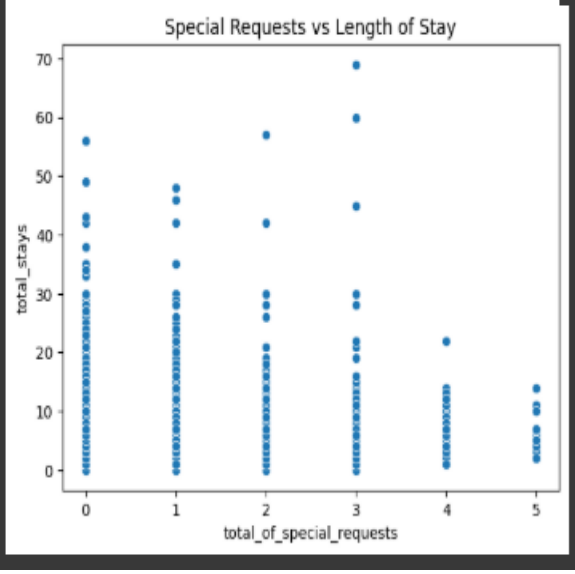
- How often are guests upgraded or reassigned to a different room type?

```python
# 11. How often are guests upgraded or reassigned to a different room type?
# same as que 4
df['room_change'] = df['reserved_room_type'].astype(str) != df['assigned_room_type'].astype(str)
upgrade_rate = df['room_change'].mean()
print(f"Room change (upgrade/reassignment) rate: {upgrade_rate:.2%}")
```

```
Room change (upgrade/reassignment) rate: 14.99%
```

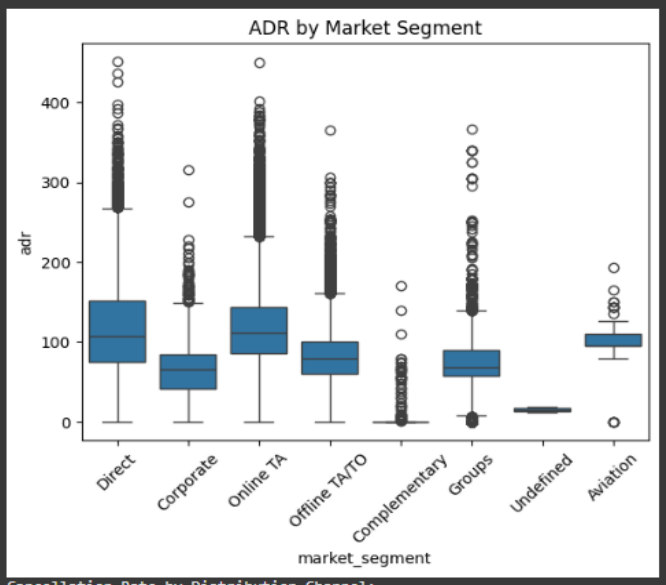
- Are guests who make special requests more likely to experience booking changes or longer stays?

```python
# 12. Are guests who make special requests more likely to experience booking changes or longer stays?
sns.boxplot(x='booking_changes', y='total_of_special_requests', data=df)
plt.title('Special Requests vs Booking Changes')
plt.show()

sns.scatterplot(x='total_of_special_requests', y='total_stays', data=df)
plt.title('Special Requests vs Length of Stay')
plt.show()
```



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

```python
# 13. Do certain market segments or distribution channels show higher booking consistency or revenue?
sns.boxplot(x='market_segment', y='adr', data=df)
plt.title('ADR by Market Segment')
plt.xticks(rotation=45)
plt.show()

booking_consistency = df.groupby('distribution_channel')['is_canceled'].mean()
print("Cancellation Rate by Distribution Channel:\n", booking_consistency)
```




```
Cancellation Rate by Distribution Channel:
 distribution_channel
Corporate     0.127785
Direct        0.148305
GDS           0.198895
TA/TO         0.309707
Undefined     0.800000
Name: is_canceled, dtype: float64
```

- What factors are most strongly associated with higher ADR?

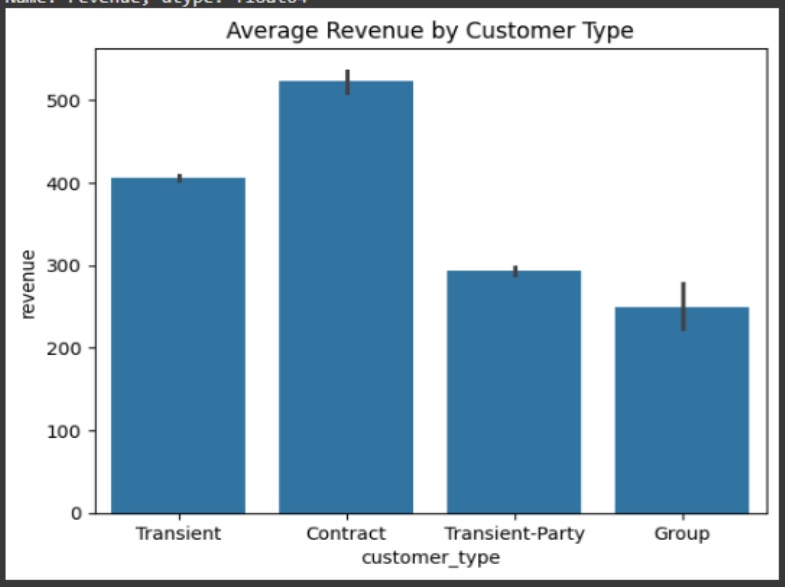
```python
# 14. What factors are most strongly associated with higher ADR?
# same as que 1
numeric_cols = df.select_dtypes(include='number').dropna(axis=1)
correlations = numeric_cols.corr()['adr'].sort_values(ascending=False)
print(correlations)
```

```
adr                             1.000000
revenue                         0.572419
total_guests                    0.409649
total_of_special_requests       0.146832
is_canceled                     0.133619
total_stays                     0.058517
required_car_parking_spaces     0.041382
lead_time                       0.025606
agent                           0.009817
booking_changes                 0.001544
days_in_waiting_list           -0.034845
previous_cancellations         -0.053113
previous_bookings_not_canceled -0.090680
is_repeated_guest              -0.162337
Name: adr, dtype: float64
```

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

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

df['revenue'] = df['adr'] * df['total_stays']
revenue_by_customer_type = df.groupby('customer_type')['revenue'].mean()
print(revenue_by_customer_type)

sns.barplot(x='customer_type', y='revenue', data=df)
plt.title('Average Revenue by Customer Type')
plt.show()
```

```
customer_type
Contract          523.047729
Group             249.829890
Transient         406.283735
Transient-Party   293.004813
Name: revenue, dtype: float64
```



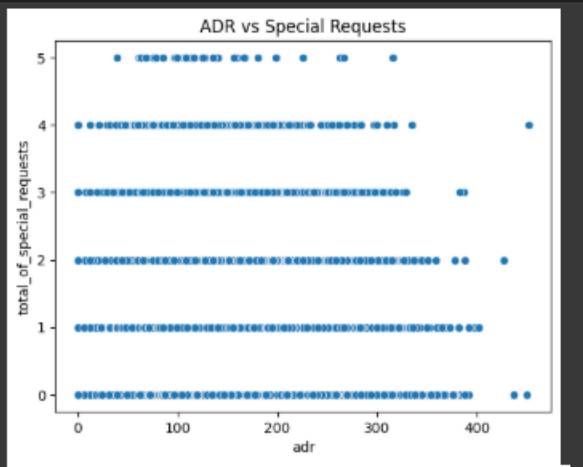
- Do bookings with more lead time or from specific countries yield higher ADR?

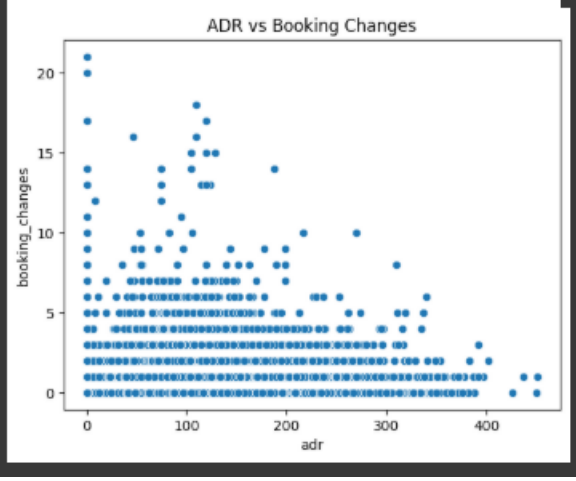
```python
# 16. Do bookings with more lead time or from specific countries yield higher ADR?
# same as que 1 and 3
top_countries = df['country'].value_counts().head(10).index
adr_by_country = df[df['country'].isin(top_countries)].groupby('country', observed=True)['adr'].mean().sort_values(ascending=False)
print("ADR by Top Countries:\n", adr_by_country)
```

```
ADR by Top Countries:
 country
ESP    122.283610
ITA    115.897678
BEL    115.379188
BRA    112.726937
FRA    112.531958
NLD    109.204359
DEU    105.939399
IRL    100.523140
GBR     97.660918
PRT     95.094359
Name: adr, dtype: float64
```

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

```python
# 17. Are guests with higher ADR more likely to request special services or make booking modifications?
sns.scatterplot(x='adr', y='total_of_special_requests', data=df)
plt.title('ADR vs Special Requests')
plt.show()

sns.scatterplot(x='adr', y='booking_changes', data=df)
plt.title('ADR vs Booking Changes')
plt.show()
```

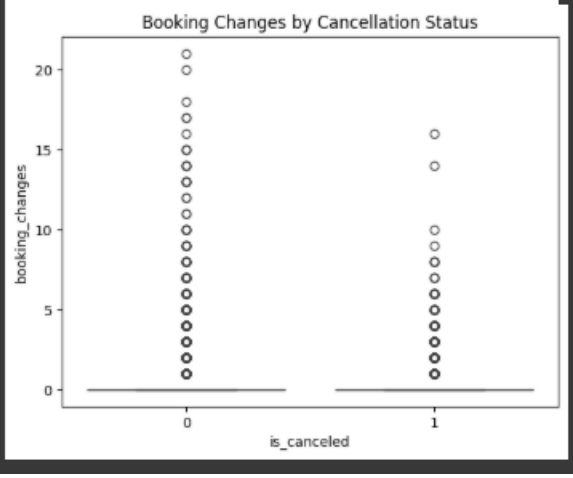




- Do guests from different countries behave differently in terms of booking timing or stay length?

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

```python
# 19. Are guests who make booking changes more likely to request additional services or cancel?
sns.scatterplot(x='booking_changes', y='total_of_special_requests', data=df)
plt.title('Booking Changes vs Special Requests')
plt.show()

sns.boxplot(x='is_canceled', y='booking_changes', data=df)
plt.title('Booking Changes by Cancellation Status')
plt.show()
```



8. Conclusion

This EDA revealed critical trends in hotel bookings, including cancellation drivers, revenue patterns, and booking channel preferences. The insights can guide pricing strategies, marketing efforts, and operational improvements to maximize revenue and customer satisfaction.