Exploratory Data Analysis (EDA) of Hotel Bookings

This project analyzes hotel booking data using **Pandas, Matplotlib, and Seaborn** to uncover key insights. The goal is to understand **booking trends, cancellations, customer behavior, revenue patterns, and geographical distribution** of guests. By identifying peak seasons, cancellation reasons, and pricing impacts, this EDA helps hotels optimize their strategies for **better revenue management and customer satisfaction.**

Goals of the Project:

1. **Understand Booking Trends –** Analyze booking patterns across different months and hotel types.

2. **Identify Cancellation Factors –** Examine reasons and trends in reservation cancellations.

3. **Revenue Analysis –** Study the impact of cancellations and booking rates on revenue.

4. **Customer Segmentation –** Identify key customer groups based on demographics and booking behavior.

5. **Market Segment Insights –** Evaluate the distribution of customers across different market segments.

5. **Geographical Distribution –** Analyze the origin of guests and their booking preferences.

6. **Impact of Price on Bookings –** Investigate how the Average Daily Rate (ADR) affects reservation and cancellation rates.

7. **Seasonal Trends –** Detect peak and low booking periods for better planning and revenue management.

8. **Data Cleaning & Preprocessing –** Handle missing values, remove outliers, and structure the dataset for better analysis.

9. **Visual Insights –** Use Matplotlib and Seaborn to create visualizations that make data-driven insights easier to interpret.

Materials and Methods:

The data for this project comes from a hotel booking dataset, containing information about reservations, customer demographics, booking status, and hotel types. This dataset includes details such as check-in/check-out dates, market segments, cancellation status, lead time, and pricing (ADR). The analysis aims to understand booking patterns, customer preferences, seasonal demand, cancellation trends, and revenue insights to enhance decision-making in the hospitality industry.

General Part:

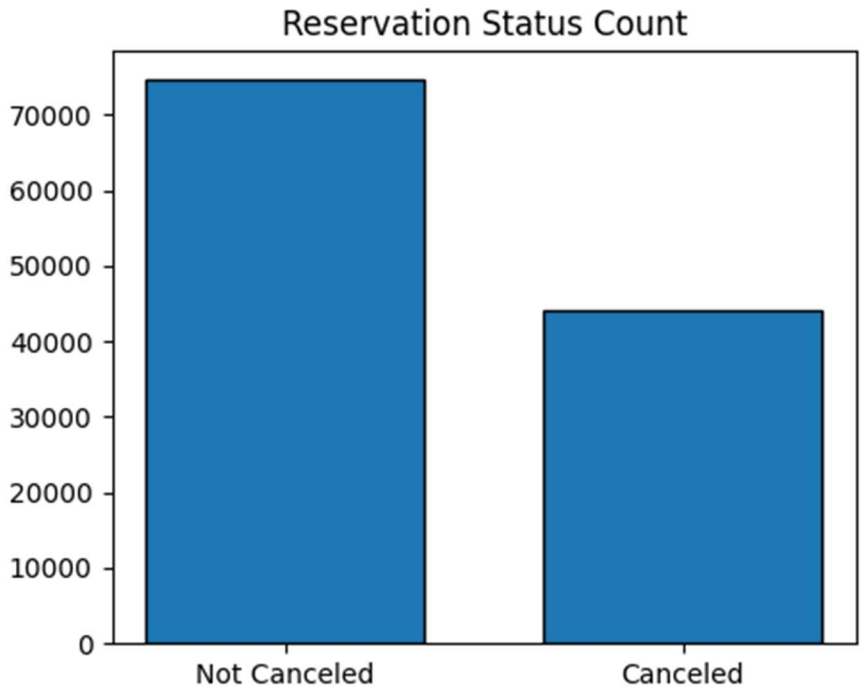
Libraries Import: Pandas, NumPy, Seaborn, Matplotlib
- **Dataset Exploration**: Initial dataset analysis, checking for missing values, duplicates, and generating summary statistics.
- **Feature Engineering**: Conversion of date columns, creation of new features like booking lead time and cancellation rates.
- **Visualization in Pandas**: Analysis of booking trends, cancellation patterns, customer segmentation, and revenue insights using various plots and charts.

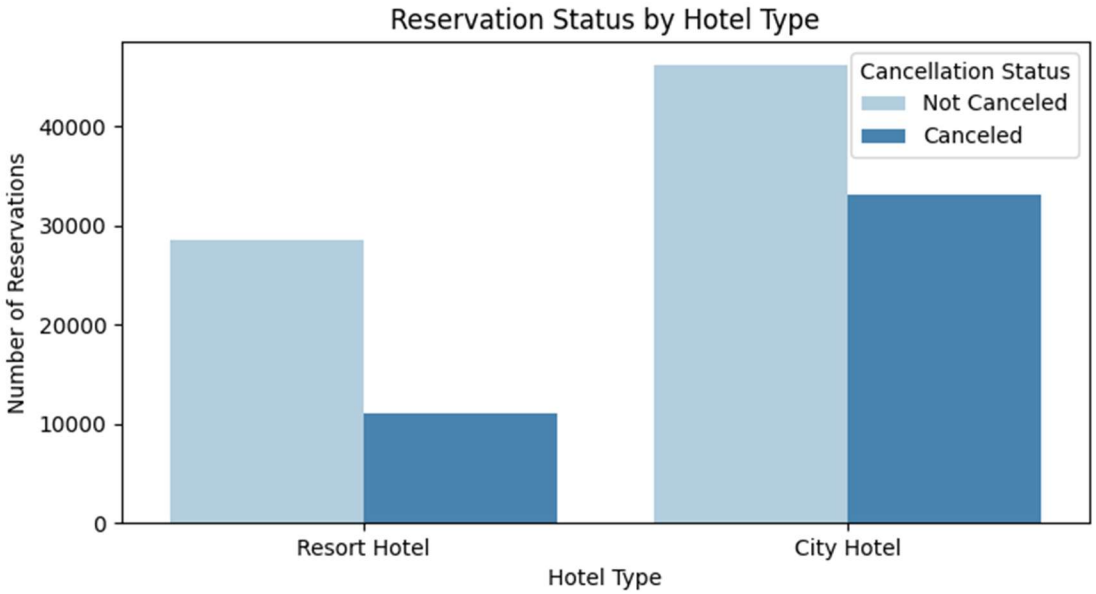
Project Outcome & Insights:

The project performs **Exploratory Data Analysis (EDA)** on the **hotel booking dataset** to gain meaningful insights into booking trends, customer behavior, and revenue generation. Below are the key outcomes:

1. Booking Trends & Hotel Performance

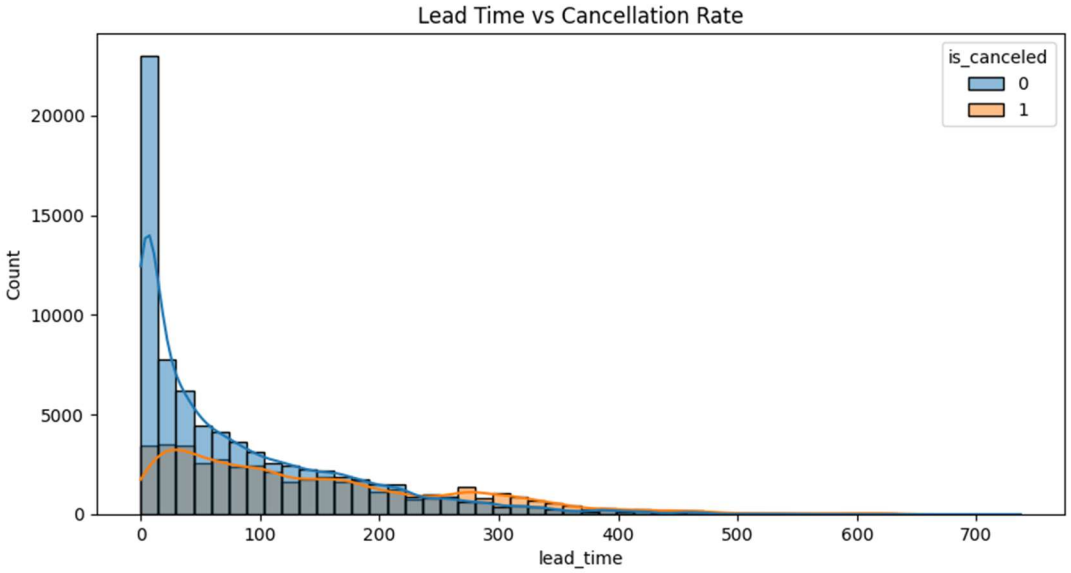
● **Hotel Type Analysis**: Identifies booking trends in **City Hotels vs. Resort Hotels** to determine occupancy patterns.
● **Seasonality & Peak Periods**: Analyzes booking trends over time, helping hotels optimize pricing and availability during peak seasons.
● **Market Segment Performance**: Evaluates which customer segments (e.g., corporate, leisure, online travel agents) drive the most bookings.

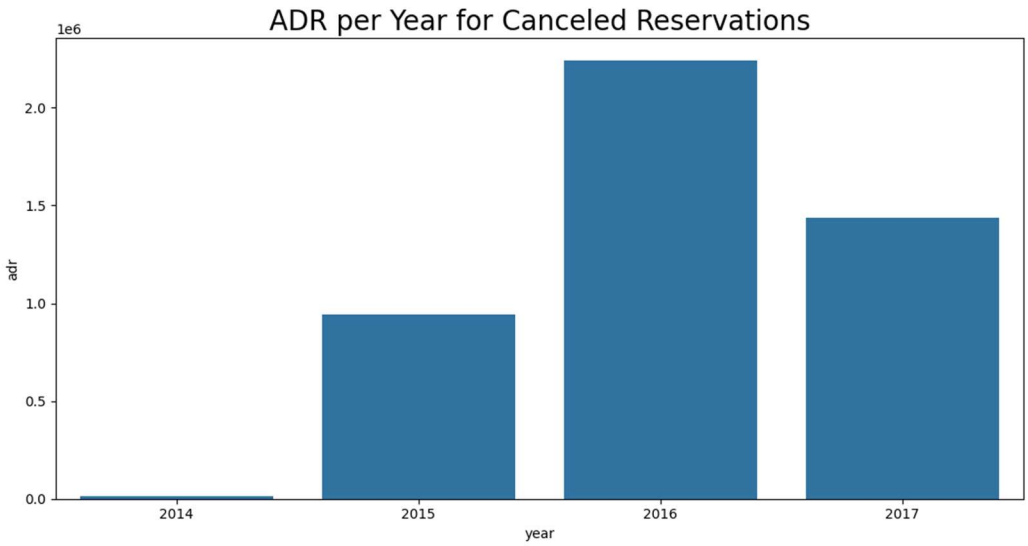


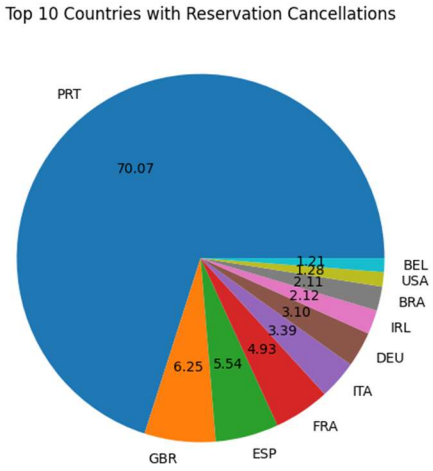


2. Customer Behavior Analysis

● **Booking Lead Time**: Examines how far in advance guests book their stays, aiding hotels in revenue management.
● **Cancellation Patterns**: Identifies factors contributing to booking cancellations, helping in policy adjustments to reduce cancellations.
● **Guest Preferences**: Analyzes preferences for meal plans, room types, and special requests, assisting in personalized service offerings.

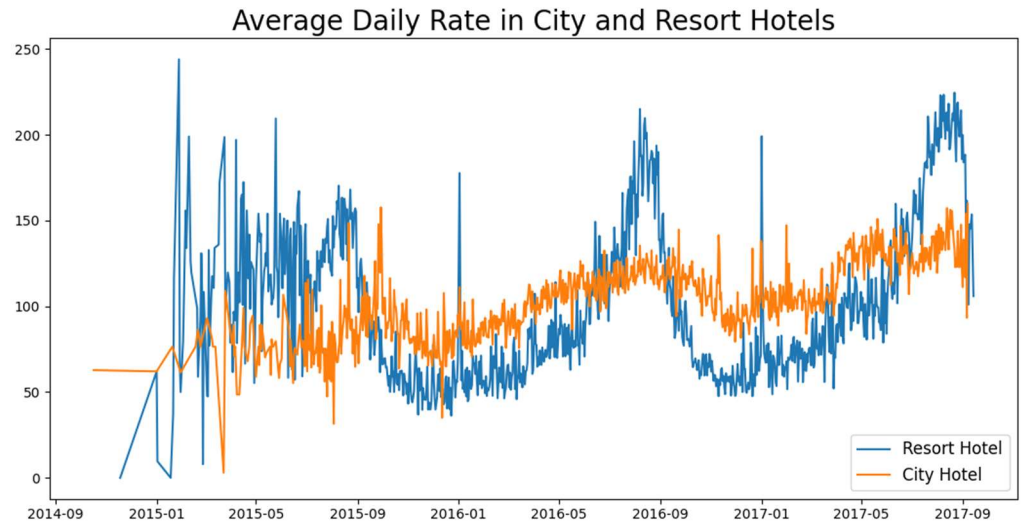


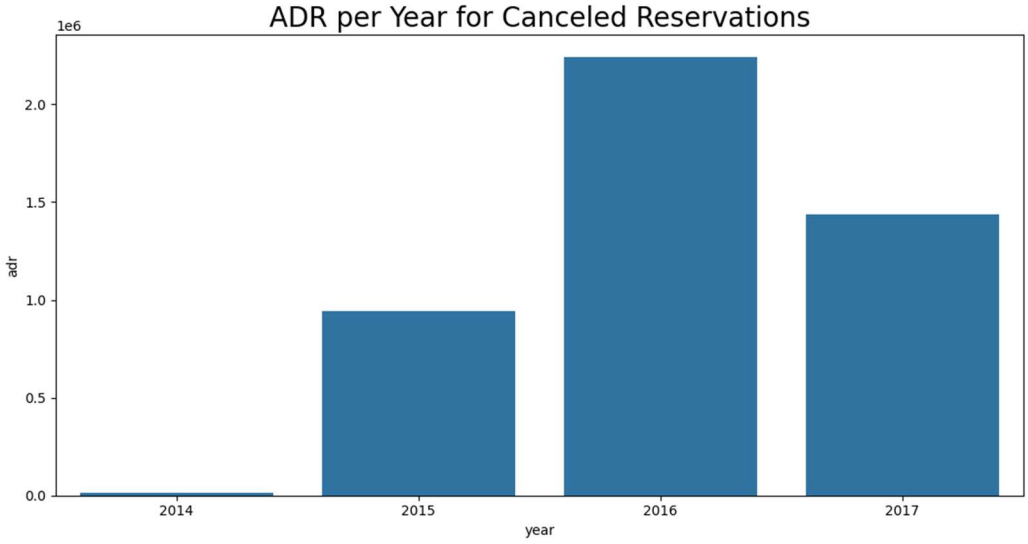


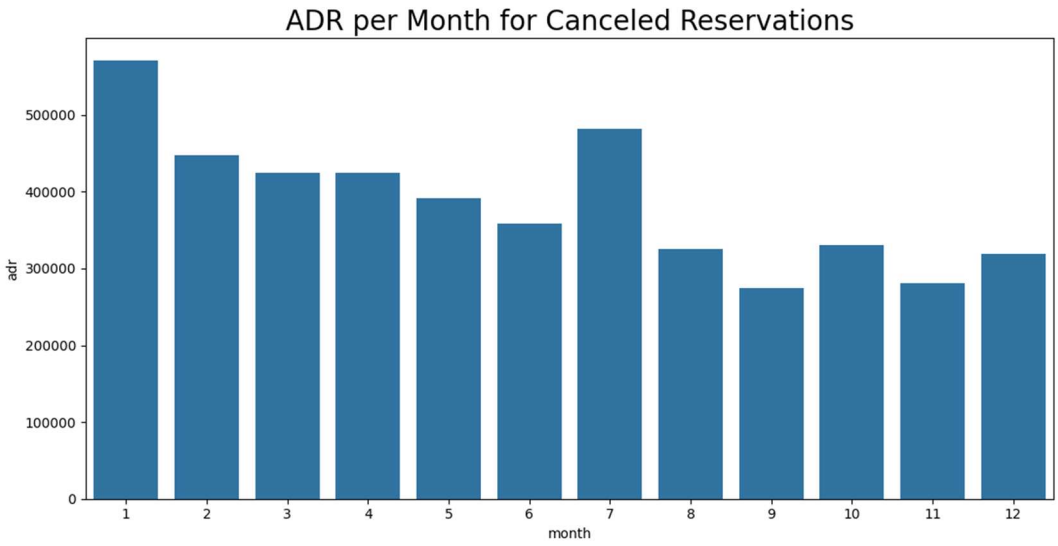


3. Revenue & Profitability Insights

● **Average Daily Rate (ADR) Analysis**: Tracks the ADR for different hotel types and booking sources to maximize revenue.
● **Revenue Contribution by Customer Type**: Identifies which guest categories (business, leisure, groups) contribute the most revenue.
● **Year-over-Year Revenue Growth**: Tracks yearly revenue trends to support business growth and financial planning.

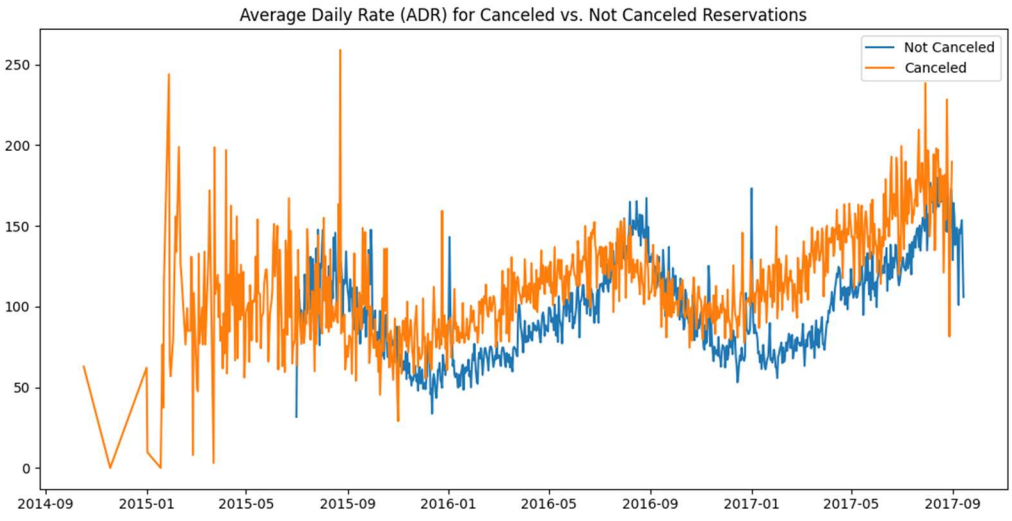

Average Daily Rate in City and Resort Hotels
250
200
150
100
50
0
2014-09
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
Resort Hotel
City Hotel

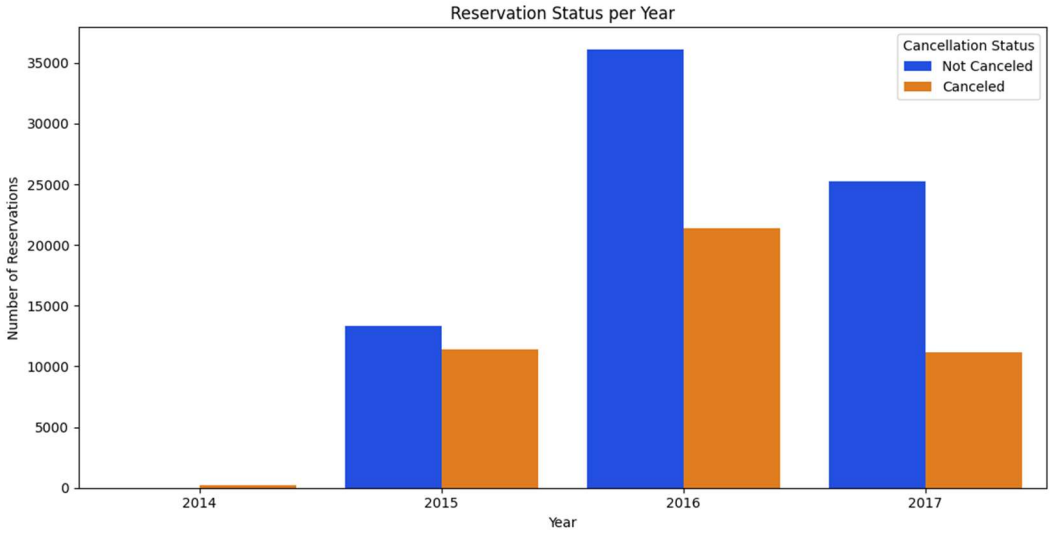

ADR per Year for Canceled Reservations
1e6
2.0
1.5
1.0
0.5
0.0
adr
2014
2015
2016
2017
year

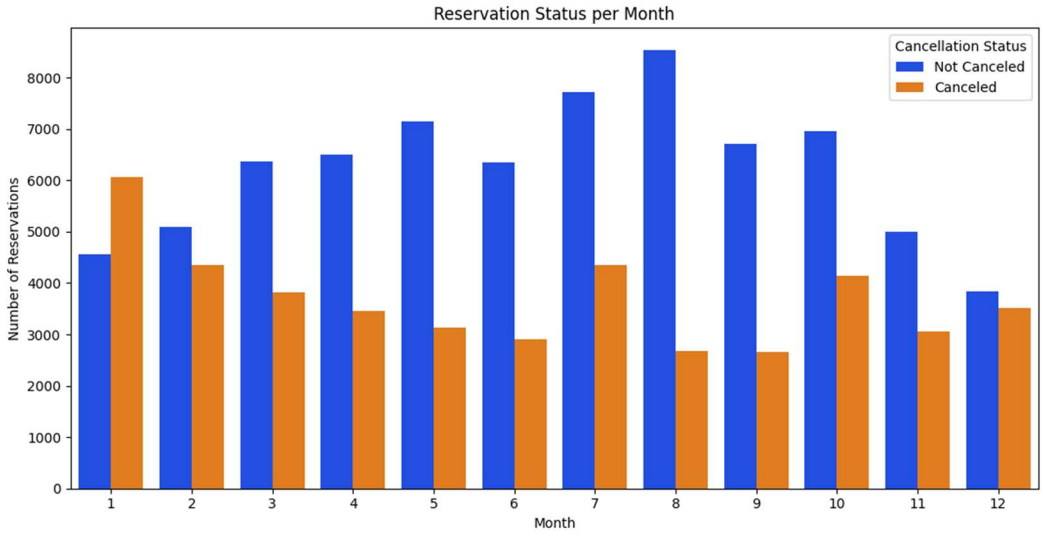


4. Operational Efficiency & Logistics

● **Length of Stay Impact**: Evaluates how stay duration affects occupancy and revenue, helping optimize pricing strategies.
● **Effect of Special Requests on Operations**: Analyzes the impact of customer special requests on operational efficiency.
● **Reservation Status Trends**: Examines confirmed vs. canceled vs. no-show reservations to optimize hotel management strategies.







Key Questions and Insights to be Addressed:

1.What is the cancellation rate?

cancellation_rate = df['is_canceled'].mean() * 100

print(f"Overall Cancellation Rate: {cancellation_rate:.2f}%")

Answer:

Overall Cancellation Rate: 37.13%

2.Which hotel type has more cancellations?

```
cancellation_by_hotel = df.groupby('hotel')['is_canceled'].mean() * 100
print(cancellation_by_hotel)
```

Answer:

hotel

City Hotel 41.708175

Resort Hotel 27.975048

Name: is_canceled, dtype: float64

3.Which month has the highest number of cancellations?

```
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'])
df['month'] = df['reservation_status_date'].dt.month
cancellations_by_month = df[df['is_canceled'] ==
1]['month'].value_counts().sort_index()
print(cancellations_by_month)
```

Answer:

month

1 6060

2 4351

3 3818

4 3464

5 3138

6 2901

7 4360

8 2684

9 2658

10 4141

11 3058

12 3519

Name: count, dtype: int64

4.Which country has the most cancellations?

```
cancelled_countries = df[df['is_canceled'] ==
1]['country'].value_counts().head(10)
print(cancelled_countries)
```

Answer:

country

PRT 27514

GBR 2453

ESP 2177

FRA 1934

ITA 1333

DEU 1218

IRL 832

BRA 830

USA 501

BEL 474

Name: count, dtype: int64

5.How does ADR (Average Daily Rate) vary for canceled vs. non-canceled bookings?

```
adr_comparison = df.groupby('is_canceled')['adr'].mean()
print(adr_comparison)
```

Answer:

is_canceled

0 100.210618

1 104.917985

Name: adr, dtype: float64

6.What is the most common market segment for bookings?

```
market_segment_counts = df['market_segment'].value_counts()
print(market_segment_counts)
```

Answer:

market_segment

Online TA 56402

Offline TA/TO 24159

Groups 19806

Direct 12448

Corporate 5111

Complementary 734

Aviation 237

Name: count, dtype: int64

7.Does lead time impact cancellation rate?

```
plt.figure(figsize=(10,5))
sns.histplot(data=df, x='lead_time', hue='is_canceled', bins=50, kde=True)
plt.title("Lead Time vs Cancellation Rate")
plt.show()
```

Answer: