Hotel Booking Case Study Report

Name- Sahil Karande

Id- 250240325059

1. Executive Summary

This case study conducts an Exploratory Data Analysis (EDA) on a hotel booking dataset to identify trends and patterns in guest behavior, pricing, and booking dynamics. The objective is to support improved revenue strategies and customer experience.

2. Objective

- Identify patterns in guest booking behaviour, such as lead time, booking changes, and special requests.
- Understand the factors influencing Average Daily Rate (ADR) and room upgrades.
- Analyse how customer types and distribution channels affect stay characteristics and pricing.
- Apply statistical tests and visual analytics to derive actionable insights for optimizing hotel revenue management and customer satisfaction.

3. Dataset Overview

- Data Source: `hotel_bookings.csv`
- Primary Fields Included:

```python
#Check the info of the dataset
#Here we get the info of all the columns
#Info like: Column name, Non_null Count, DataType of that column

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

4. Data Cleaning & Preprocessing

1. Dropped Column

- Dropped the company column due to **~94% null values**.

```
df.drop(['company'], axis=1, inplace=True, errors='ignore')
```

2. Handling Missing Values

- **agent**: Replaced nulls with 0 and converted to integer.

```
df['agent'] = df['agent'].fillna(0)
```

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

- **children**: Replaced nulls with 0.

```
df['children'] = df['children'].fillna(0)
```

- **country**: Replaced nulls with mode value 'PRT'.

```
df['country'] = df['country'].fillna(df['country'].mode()[0])
```

3. Data Type Conversion

- Converted selected columns to appropriate types (category, int):

```
dtype_map = {
    'hotel': 'category',
    'arrival_date_month': 'category',
    'meal': 'category',
    'country': 'category',
    'market_segment': 'category',
    'distribution_channel': 'category',
    'reserved_room_type': 'category',
    'assigned_room_type': 'category',
    'deposit_type': 'category',
    'customer_type': 'category',
    'reservation_status': 'category',
    'children': 'int16',
    'agent': 'int32',
```

```
}


for col, dtype in dtype_map.items():
    df[col] = df[col].astype(dtype)
```

4. Date Parsing and Formatting

- Converted reservation_status_date and created arrival_date from components:

```
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'].astype(str) + '-' +
    df['arrival_date_day_of_month'].astype(str),
    format='%Y-%B-%d',
    errors='coerce'
)


df['arrival_day_name'] = df['arrival_date'].dt.day_name()
df.drop(['arrival_date_year', 'arrival_date_month',
'arrival_date_day_of_month'], axis=1, inplace=True)
```

5. Feature Engineering (Derived Fields)

- Total stay, checkout date, and group size:

```
df['total_stay_nights'] = df['stays_in_weekend_nights'] +
df['stays_in_week_nights']

df['departure_date'] = df['arrival_date'] +
pd.to_timedelta(df['total_stay_nights'], unit='D')

df['total_members'] = df['adults'] + df['children'] + df['babies']
```

- Revenue generated from each booking:

```
df['revenue_generated'] = df['adr'] * df['total_stay_nights']
```

6. Removed Duplicates

- Removed duplicate rows (if any):

```
df = df.drop_duplicates(keep='first')
```

7. Outlier Detection (Visual Check)

- Boxplots for numerical columns to visually check for outliers:

```
numeric_cols = df.select_dtypes(include=['int64', 'int32', 'int16',
'float64']).columns


for col in numeric_cols:
    plt.figure(figsize=(6, 4))
    df.boxplot(column=col)
    plt.title(f'Boxplot of {col}')
```

```
plt.ylabel(col)
plt.grid(True)
plt.tight_layout()
plt.show()
```

5. Exploratory Data Analysis

- **Univariate Analysis**
  - Lead Time: Most bookings were made at the last minute (0–1 days), showing a right-skewed distribution.
  - ADR (Average Daily Rate): Majority of customers paid between ₹30–₹100; a few outliers paid much higher.
  - Hotel Type: City hotels had nearly double the bookings compared to resort hotels.
  - Country: Portugal (PRT) had the highest number of bookings.
  - Booking Status: Around 72.5% of bookings were not canceled, while 27.5% were canceled.
  - Customer Type: Most guests were Transient (solo travelers), followed by Transient-Party and Contract customers.
  - Arrival Day: Monday had the highest number of arrivals; weekend days (Friday, Saturday) were also popular.
- **Bivariate Analysis**
  - Cancellation vs Arrival Day: Cancellations were highest on Fridays and Saturdays.
  - Lead Time vs Cancellation: Bookings made far in advance were more likely to be canceled.

- Lead Time vs Booking Changes: Guests who booked earlier tended to make slightly more changes.
- Customer Type vs Lead Time: Contract and Transient-Party guests booked with the longest lead times.
- Cancellation Rate by Customer Type:
    - Highest: Transient customers (~30%)
    - Lowest: Group bookings (~10%)

- Multivariate Insights
    - Lead Time, Booking Changes, Special Requests:
        - Weak correlations with each other and with ADR.
        - More special requests are weakly associated with higher ADR and longer stays.
    - Room Reassignment:
        - ~15% of bookings had a room type change (upgrade or adjustment).
    - Stay Duration:
        - Average stay: 3.6 nights
        - Longest: Contract customers (~6 nights)
        - Shortest: Group customers (~3 nights)
    - Market Segment & Revenue:
        - Online TA and Direct bookings had the highest ADR.
        - Corporate and Complementary segments had the most booking consistency (least cancellations).

- Country-Based Analysis

- Portugal, UK, France led in booking volume.
- Guests from UK, Ireland, Germany booked furthest in advance (100+ days).
- Spain, Portugal had more last-minute bookings.
- ADR and lead time varied significantly across countries.

- **Notable Relationships**
  - More special requests → Higher ADR and longer stays.
  - Guests with more booking changes often booked earlier.
  - Guests from certain countries (ESP, PRT) booked shorter stays.
  - Bookings with more lead time had a slightly higher cancellation probability

6. Correlation Analysis

- **Pearson Correlation**: Measures **linear** relationships between numeric variables.
- **Spearman Correlation**: Measures **monotonic** relationships (increasing or decreasing trends), even if not linear.
- **Heatmaps**: Visualized correlations to identify patterns and possible multicollinearity.

Key Findings

1. ADR (Average Daily Rate) Correlations

- **ADR vs Total Special Requests**:
  - Weak **positive** correlation. More requests → slightly higher ADR.
- **ADR vs Lead Time**:

    - Very weak positive correlation. Longer planning time might slightly increase ADR.

- **ADR vs Booking Changes**:
    - Negligible correlation. Booking changes have little to no effect on price.

2. Lead Time Correlations

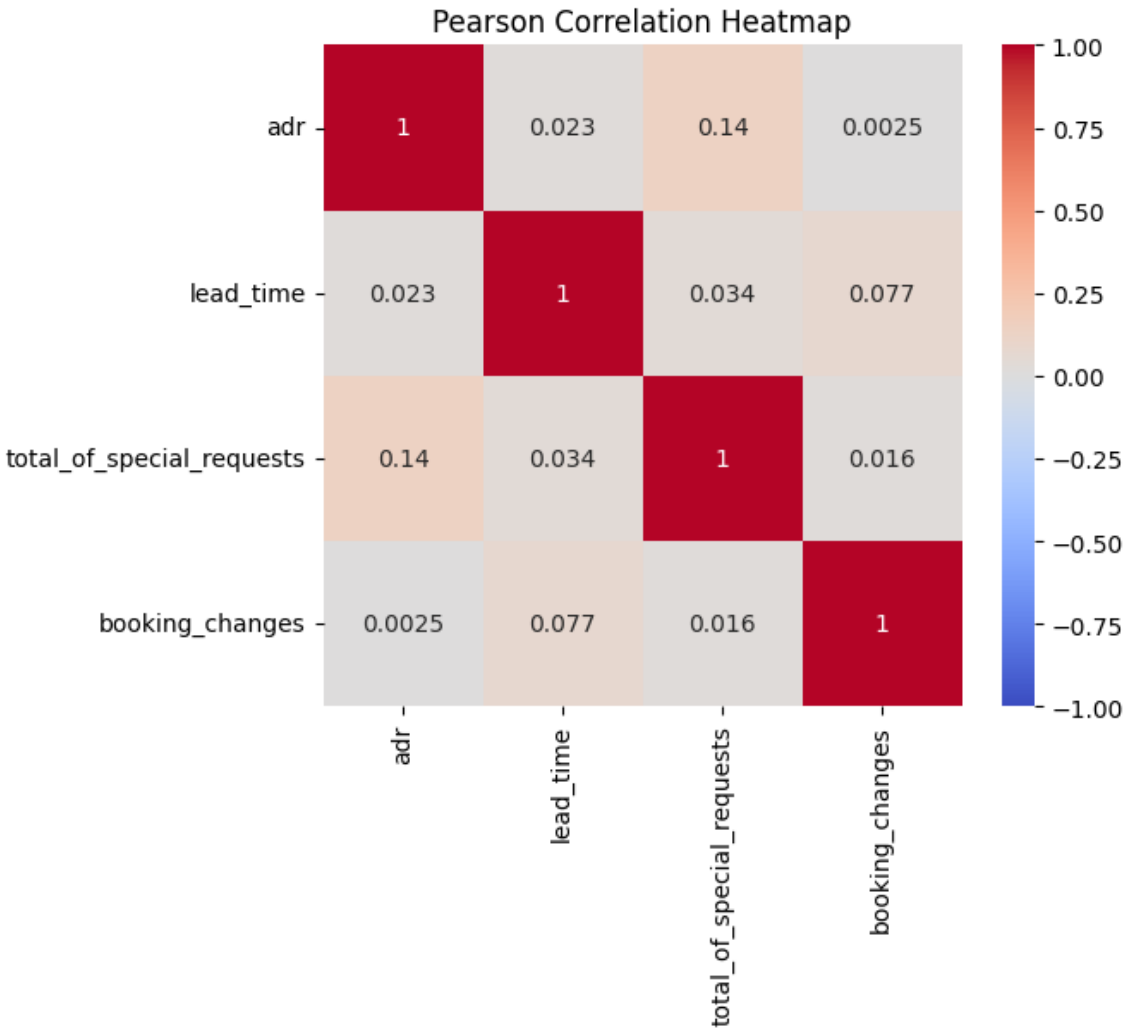
- **Lead Time vs Booking Changes**:
    - Very weak positive correlation. Guests who book earlier tend to make slightly more changes.
- **Lead Time vs Special Requests**:
    - Very minor positive relationship. Longer lead times slightly associated with more requests.

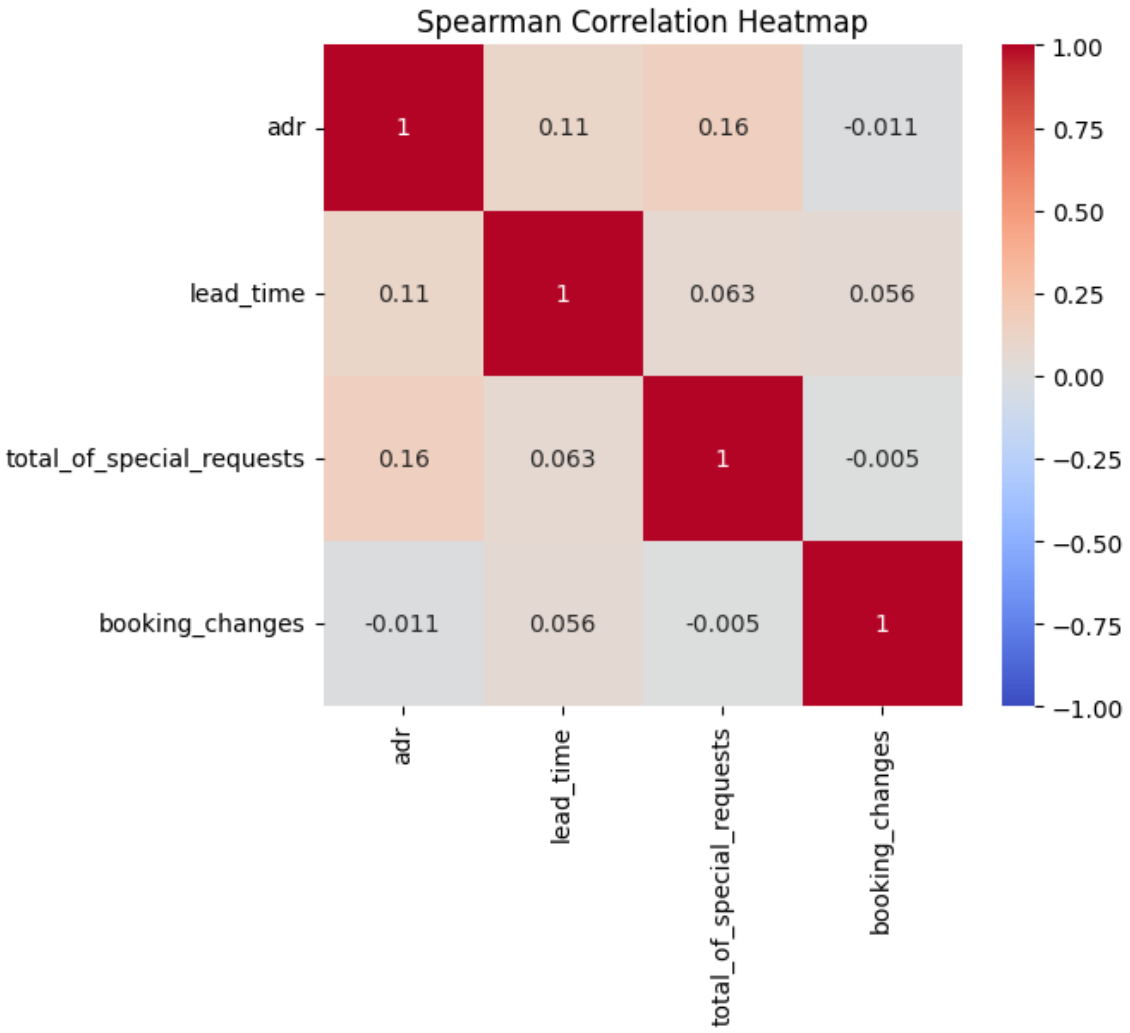
3. Special Requests vs Booking Changes

- Slight positive correlation. More requests → slightly more booking changes.

Interpretation

- Most correlations are **very weak**, indicating that:
    - There is **no strong linear or monotonic** relationship between these variables.
    - Variables like lead time, booking changes, and special requests do **not strongly influence ADR** on their own.


Pearson Correlation Heatmap
adr
lead_time
total_of_special_requests
booking_changes
1
0.023
0.14
0.0025
0.023
1
0.034
0.077
0.14
0.034
1
0.016
0.0025
0.077
0.016
1
adr
lead_time
total_of_special_requests
booking_changes
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



7. Hypothesis Testing

Performed several statistical tests to validate business assumptions using t-tests, chi-square, and ANOVA.

1. ADR by Distribution Channel

Hypothesis:

- H_0: There is no difference in ADR between bookings made via Online TA and Direct channels.
- Test Used: Independent two-sample t-test
- Result:

    - p-value = 0.3303 (greater than 0.05)
    - Failed to reject H_0
- Conclusion: No significant difference in average ADR between Online TA and Direct bookings.

2. Room Upgrade vs Lead Time

Hypothesis:

- H_0: Room upgrades are independent of lead time.
- Test Used: Chi-Square Test of Independence
- Result:
    - p-value < 0.0001 (less than 0.05)
    - Rejected H_0
- Conclusion: There is a significant association between room upgrades and how early the booking was made.

3. Stay Duration by Customer Type

Hypothesis:

- H_0: Average stay duration is the same for all customer types.
- Test Used: One-Way ANOVA
- Result:
    - p-value < 0.0001 (less than 0.05)
    - Rejected H_0
- Conclusion: Stay duration significantly differs across customer types (e.g., Contract vs. Transient).

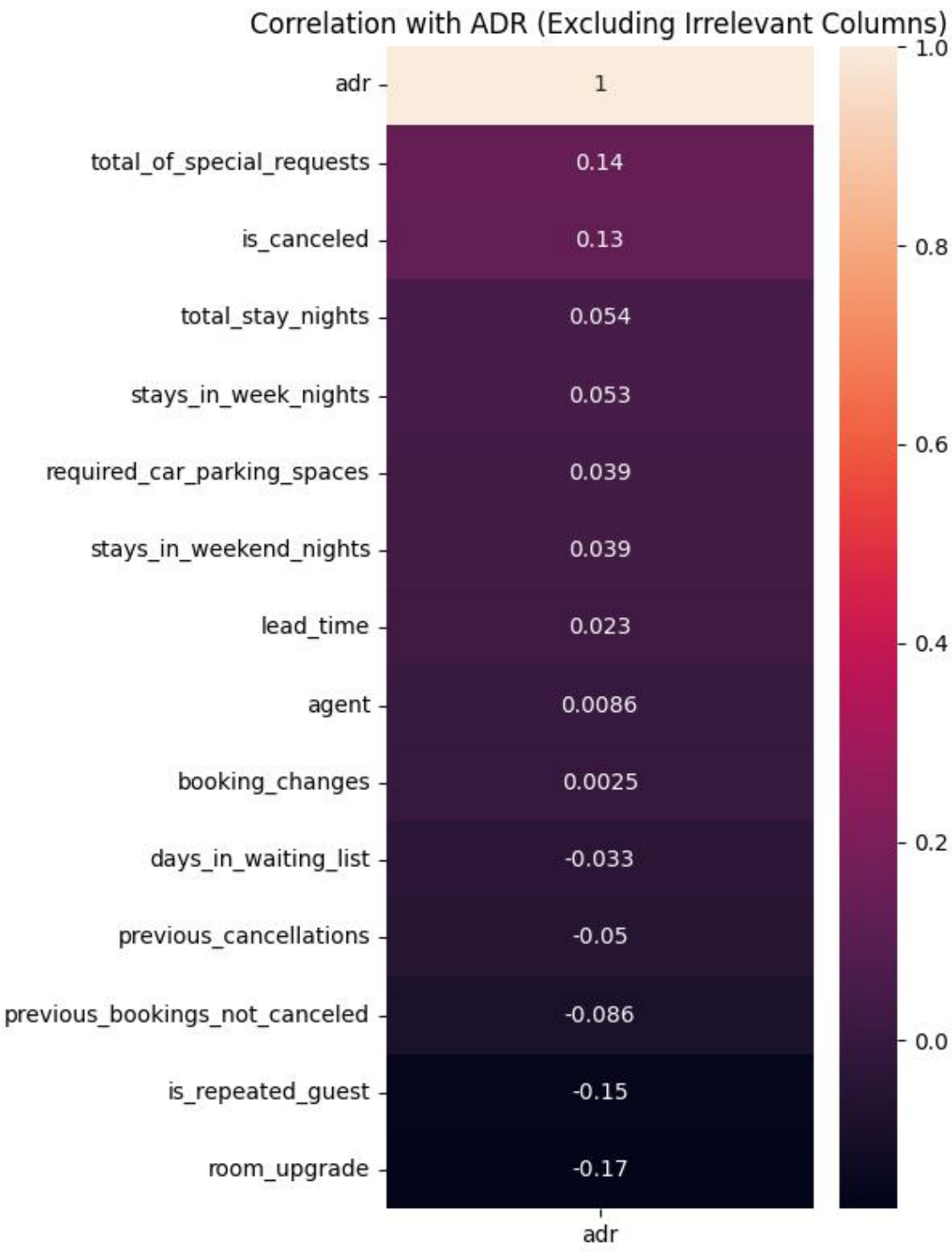
Overall Insights

- Lead time influences room upgrades.
- Customer type affects stay duration.

- Distribution channel does not significantly affect pricing (ADR)

8. Key Business Findings

 # 1. What influences ADR the most?

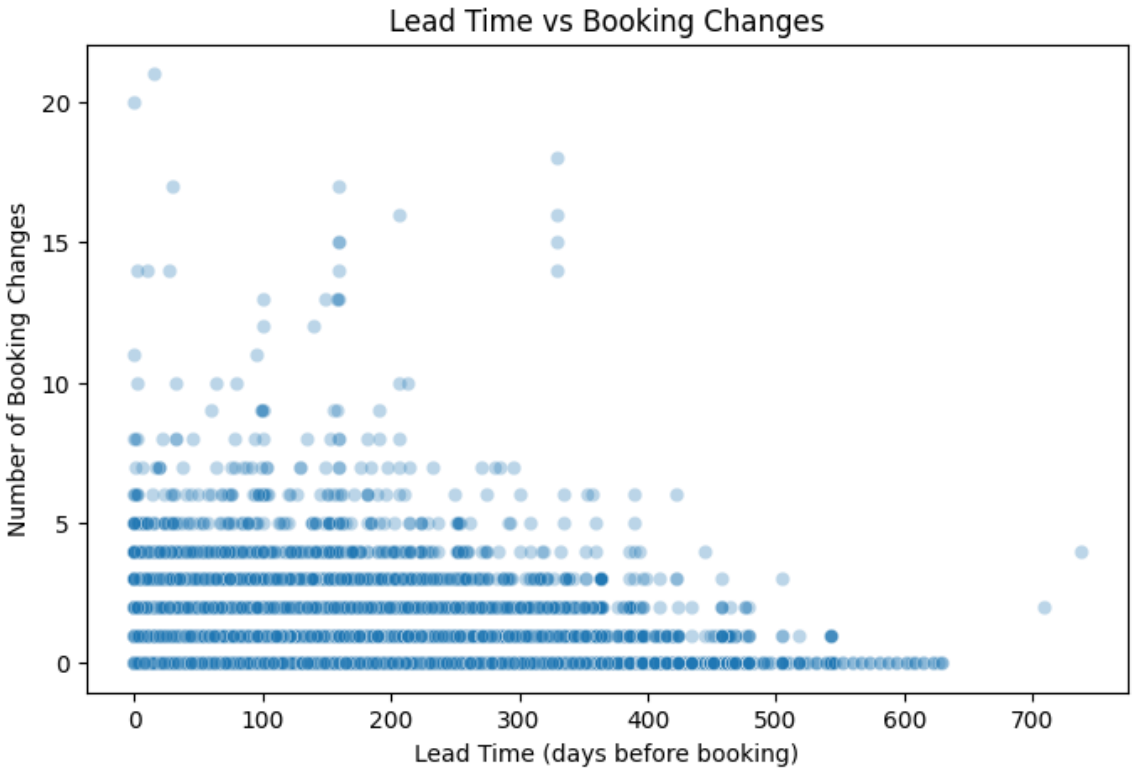


After excluding irrelevant columns, the correlation analysis shows that total special requests have the strongest (but still weak) positive influence on ADR. Lead time has a slight positive effect, while booking changes have negligible impact on pricing.

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

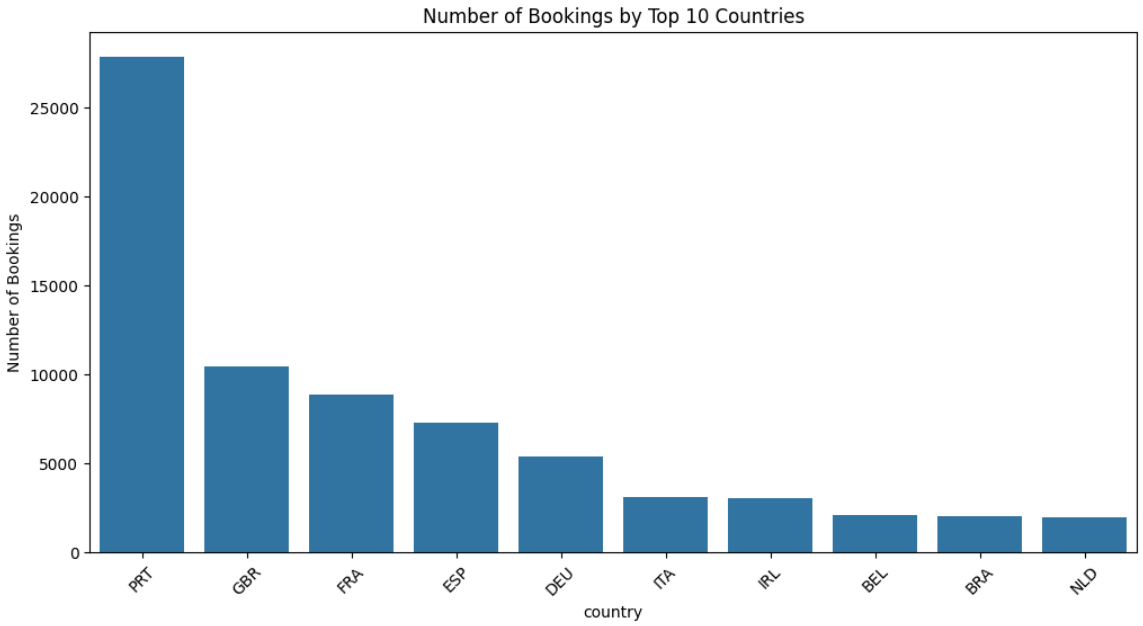
* Guests who book earlier tend to request a bit more changes.

* However, the effect size is very small (correlation ~0.08), so lead time is not a strong predictor of booking changes by itself.



3. Are there pricing or booking differences across countries?

Top 10 countries by bookings show varying average daily rates (ADR). Portugal (PRT) leads in bookings with 27,879, having an average ADR of about 95. Guests from Spain (ESP) and Italy (ITA) tend to pay higher rates, with average ADRs above 115. Overall, while booking volumes differ, pricing varies significantly across countries.

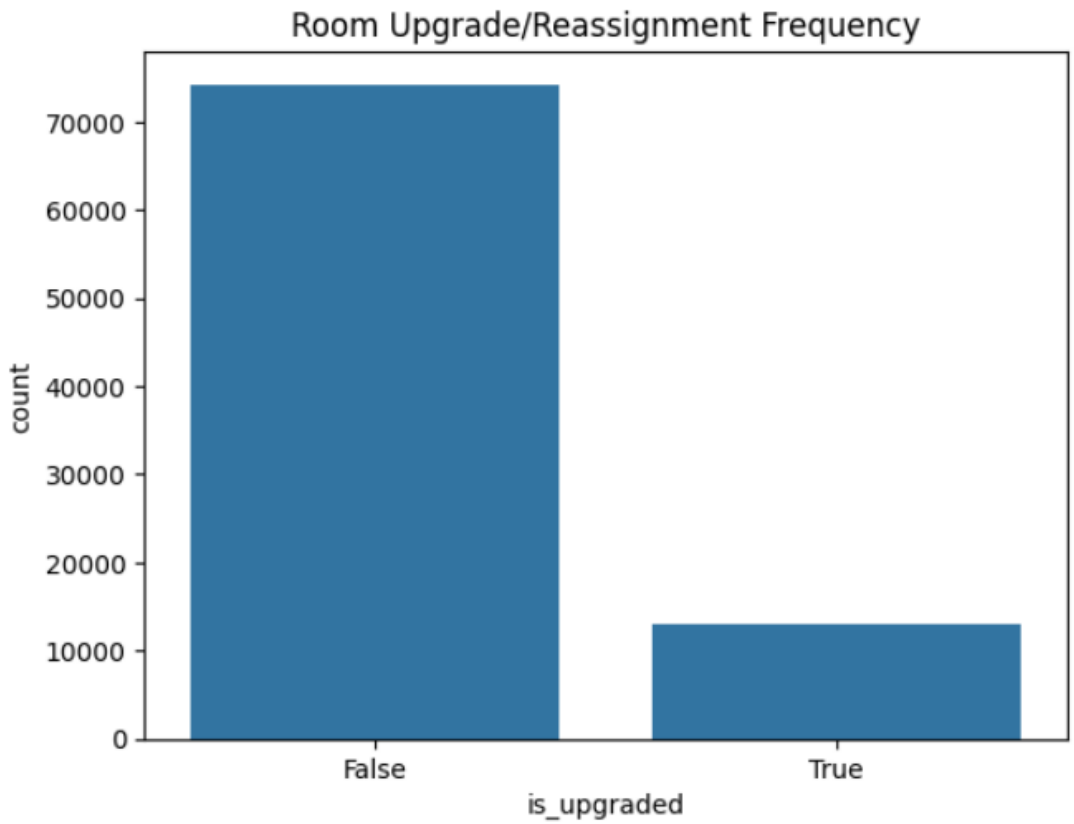


* Average daily rates (ADR) differ significantly by country, showing guests from different countries pay different prices.

* The top booking countries are Portugal, Great Britain, and France.

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

14.99

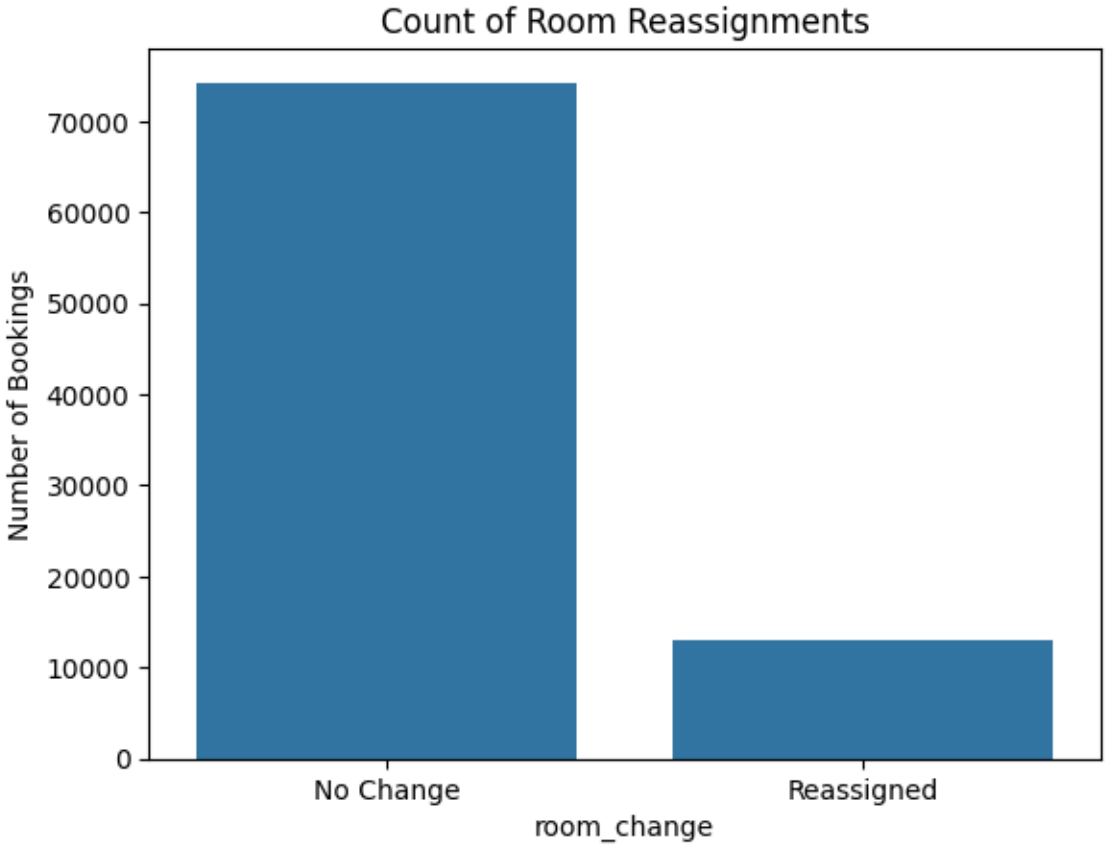


To analyze room upgrades or reassignment trends, a new column is introduced to flag discrepancies between the Reserved room type and the Assigned room type. A count plot is then used to illustrate the frequency of these mismatches. The key insight here is to determine how often guests are assigned a room that differs from what they originally booked, which could highlight patterns related to upgrades, overbooking, or operational changes.

About 15% of bookings involved a room reassignment, where the assigned room type differed from the originally reserved room type.

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

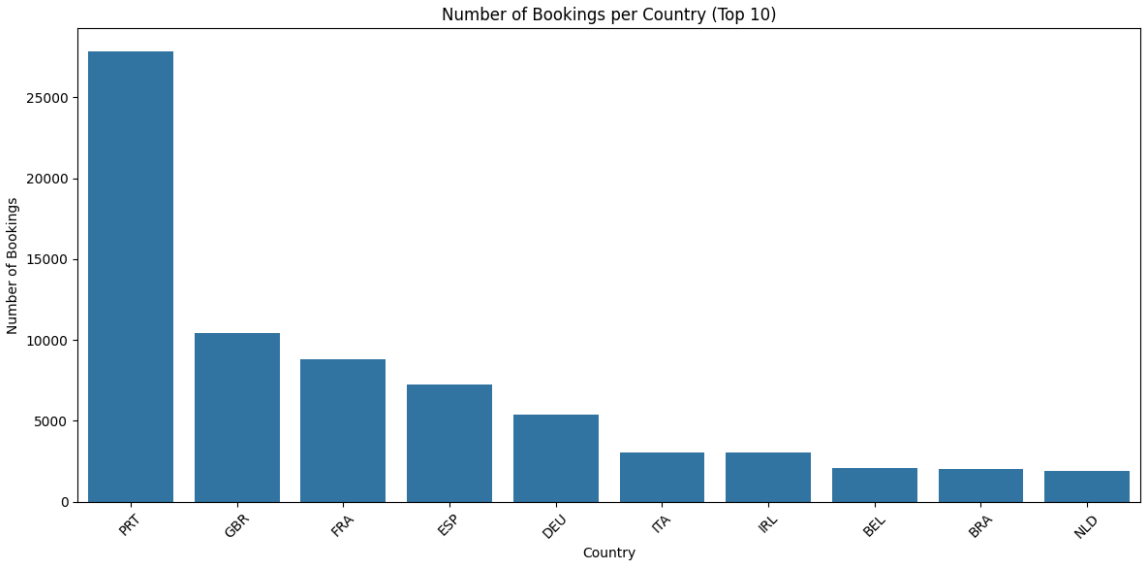
Most bookings (about 74,266) had no room change, while around 13,097 bookings experienced room reassignment. Room changes occurred in roughly 15% of the cases.



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

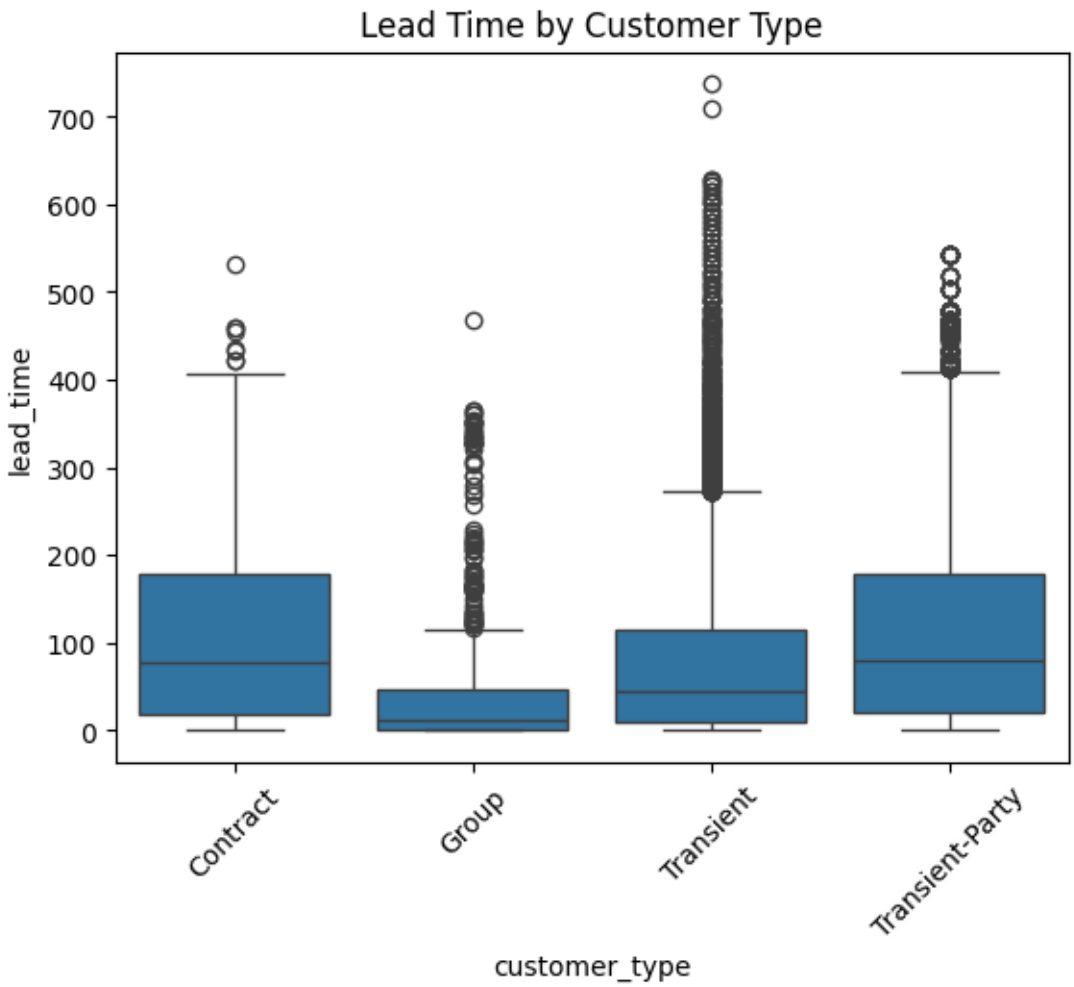
The top 10 countries with the highest number of bookings are led by Portugal (PRT) with 27,879 bookings, followed by Great Britain (GBR) with 10,432 bookings, and France (FRA) with 8,837 bookings. Other

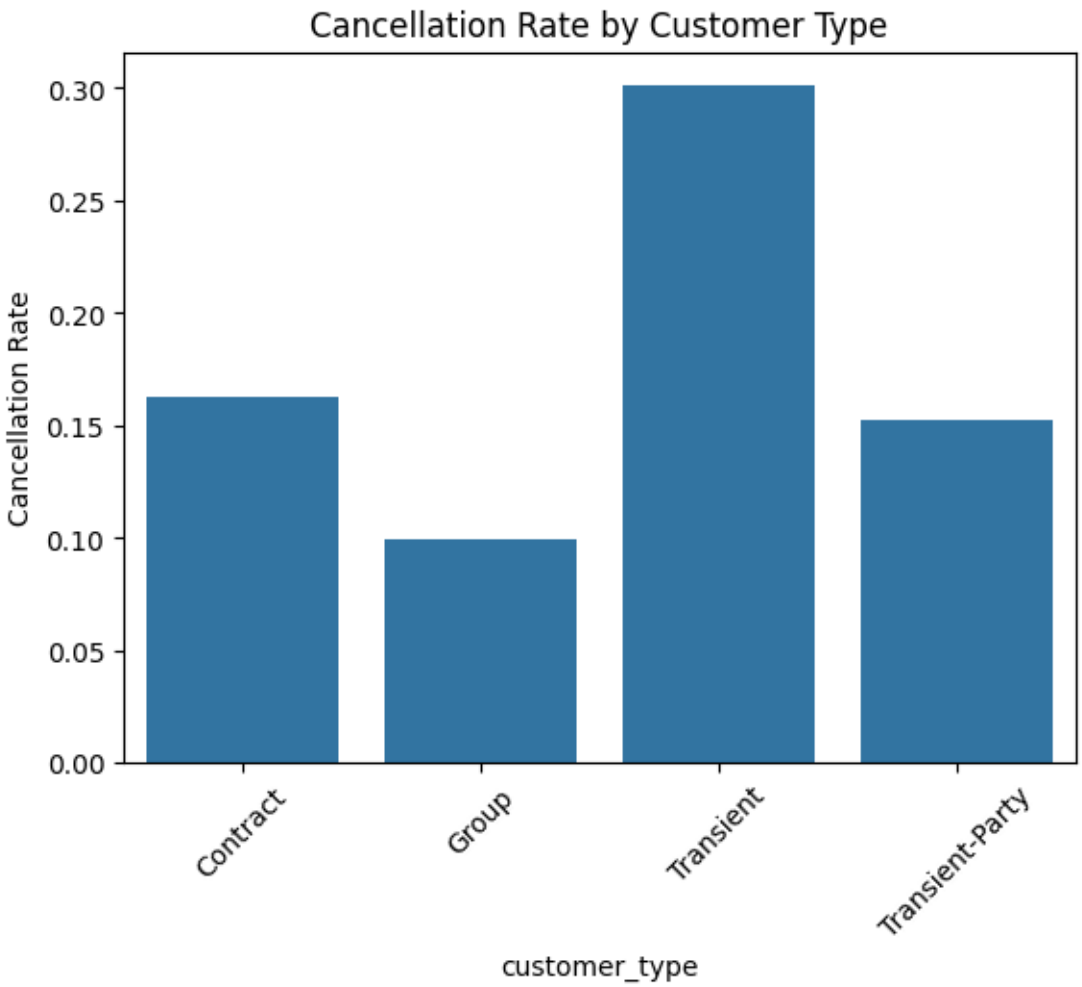
countries like Spain, Germany, Italy, Ireland, Belgium, Brazil, and the Netherlands also contribute significantly, each with bookings ranging from around 1,900 to 7,200.



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

Guests in the Transient-Party and Contract customer types tend to book the earliest, with average lead times of about 113 and 109 days respectively. Transient guests book with a moderate lead time of around 73 days, while Group bookings have the shortest lead time, averaging about 52 days. The majority of bookings come from Transient guests.


Lead Time by Customer Type
lead_time
0
100
200
300
400
500
600
700
Contract
Group
Transient
Transient-Party
customer_type



Transient customers have the highest cancellation rate (~30%).

Contract and Transient-Party customers show moderate cancellation rates (~16% and ~15%).

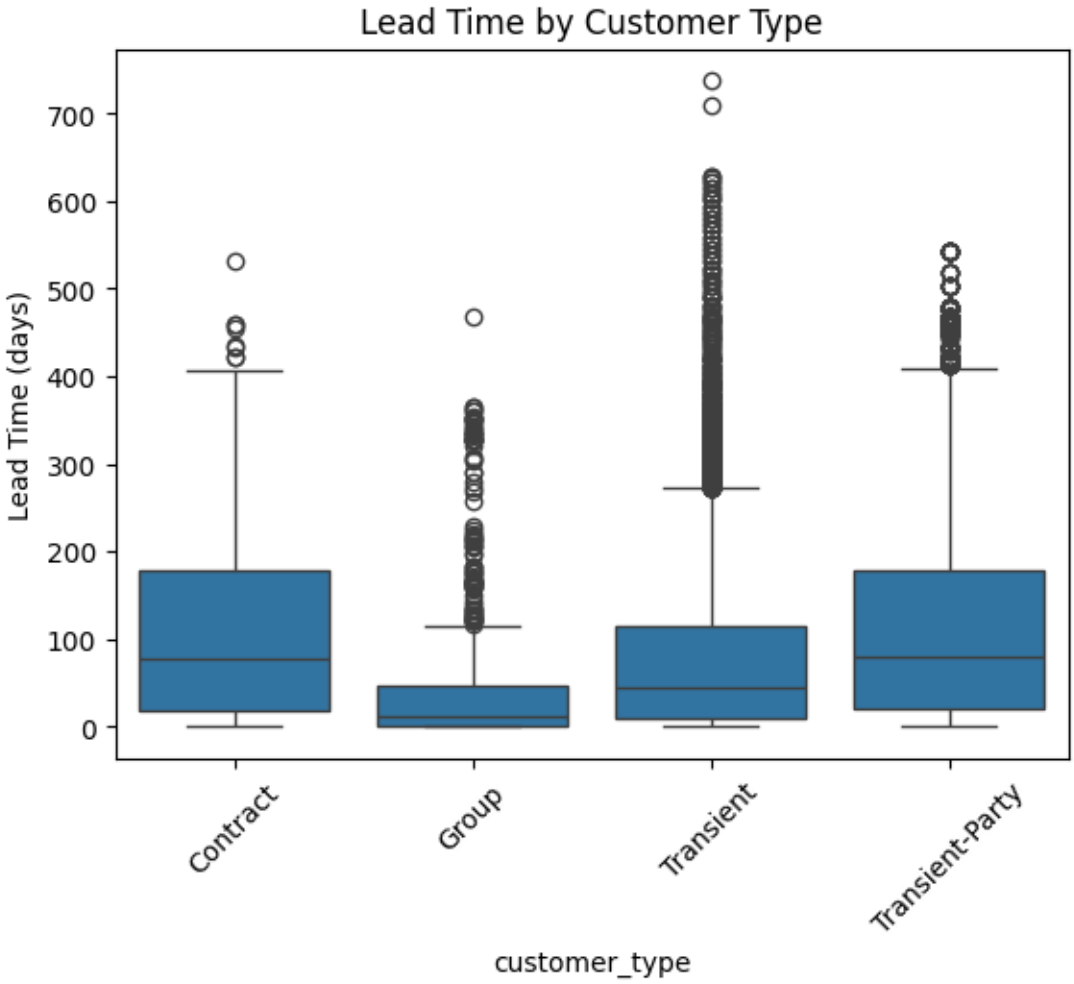
Group bookings have the lowest cancellation rate (~10%).

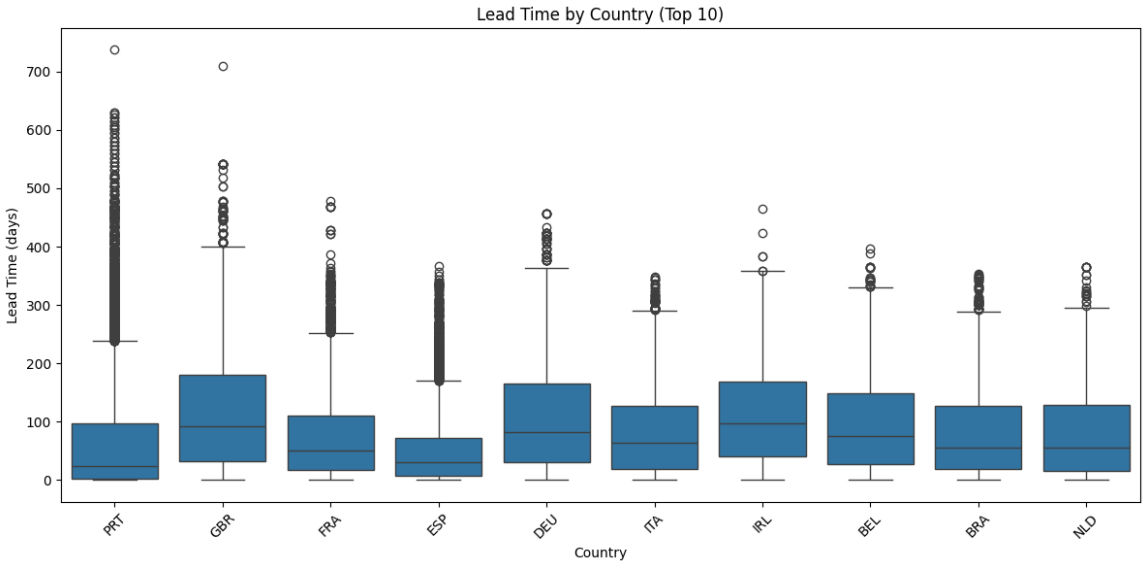
8.How does booking lead time vary across customer types and countries?

Transient-Party and Contract customers have the longest lead times (113 & 109 days).

Transient bookings are most common but have a shorter lead time (~73 days).

Group bookings have the shortest lead time (~52 days) and are least frequent.

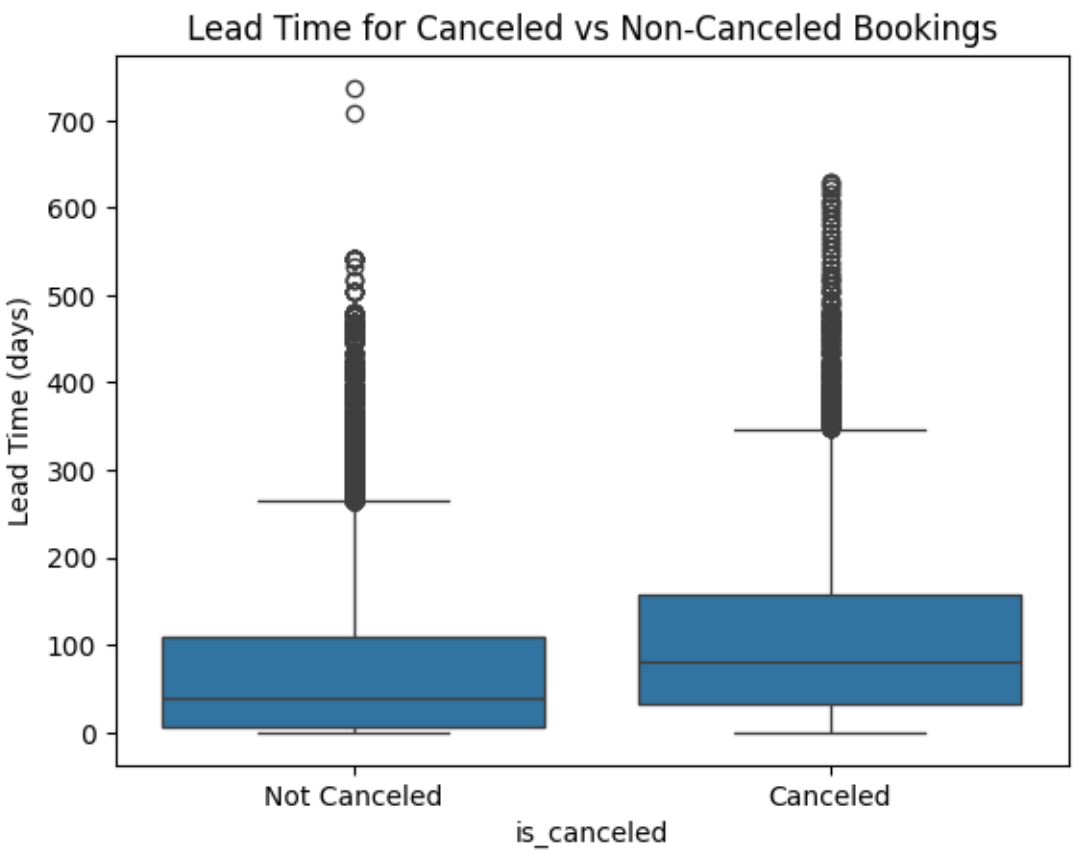


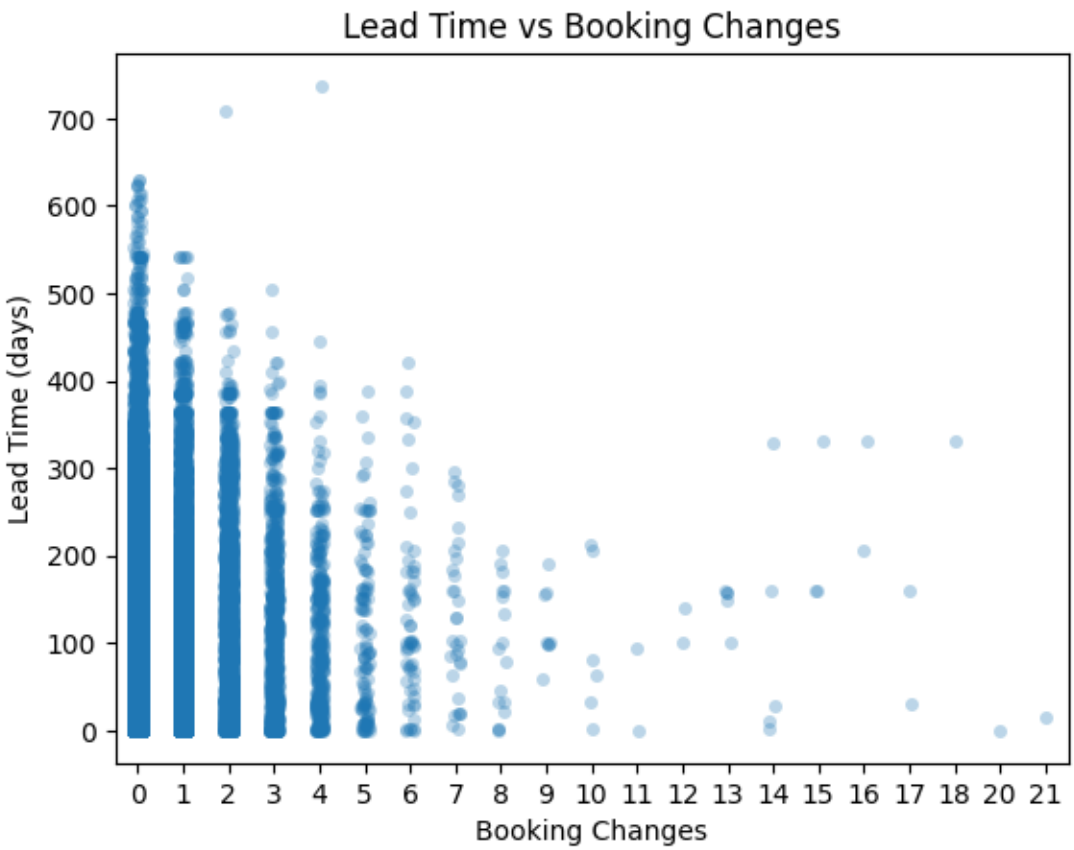


- Longest lead times: GBR (117 days), IRL (114), DEU (105) – early planners.
- Shortest: ESP (52), PRT (64) – more last-minute bookings.
- Moderate: FRA, ITA, BRA, NLD – around 74–83 days.
- PT Portugal leads in booking count but books moderately early.
- Conclusion: Lead time varies widely by country; UK & Ireland book earliest, Spain latest.

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

Guests who canceled booked much earlier (about 106 days) than those who didn't (about 70 days). Most bookings did not get canceled.


Lead Time for Canceled vs Non-Canceled Bookings
Lead Time (days)
0
100
200
300
400
500
600
700
Not Canceled
Canceled
is_canceled



People who change their bookings more tend to book earlier. Most don't change, but those with multiple changes have longer lead times.

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

On average, people stay about 3.6 nights, and usually, the typical stay is around 3 nights.

People on contract tend to stay the longest (about 6 nights), while groups stay the shortest (around 3 nights). Transient and transient-party customers fall in between with stays of about 3.3 to 3.6 nights.

Guests booking through Offline travel agents stay the longest (about 4.7 nights), while those from Undefined or Complementary segments stay

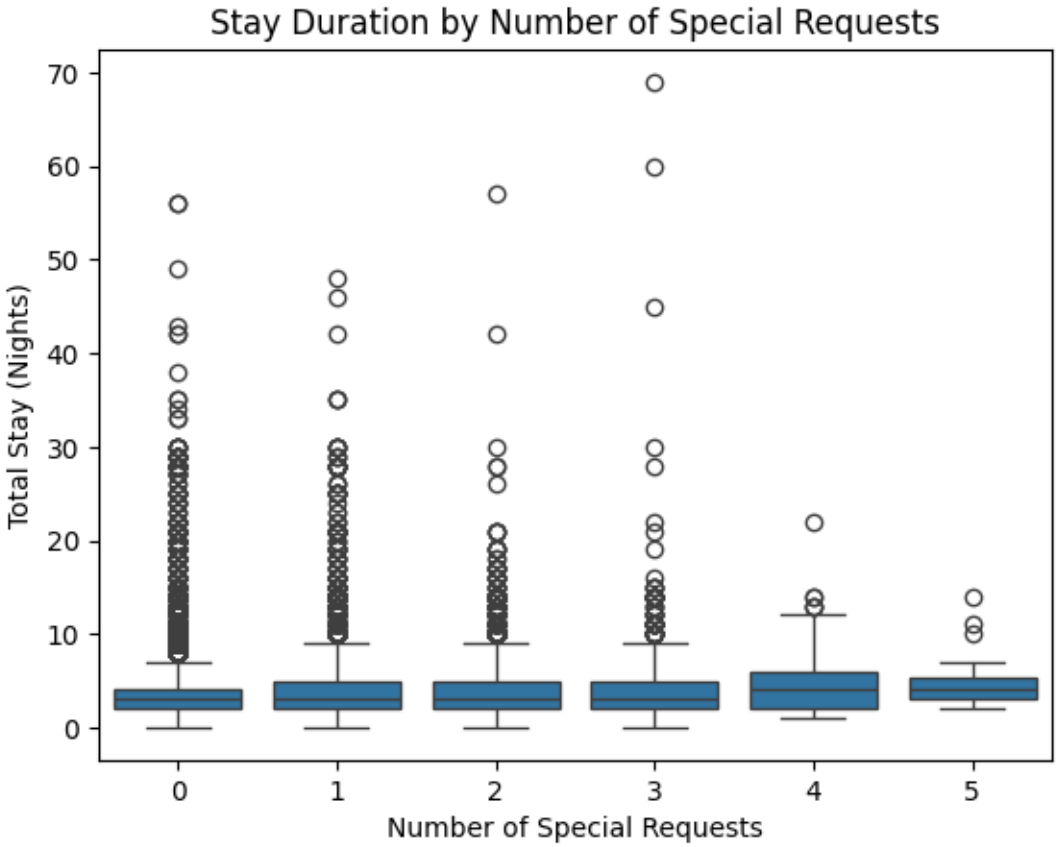
the shortest (around 1.5 to 1.7 nights). Others like Online travel agents, Aviation, and Groups have stays between 3 to 3.6 nights. Corporate guests stay for about 2 nights on average.

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

About 15% of guests get their rooms changed from what they originally booked.

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

People who make more special requests tend to change their bookings more often. For example, those with 5 special requests change bookings the most, while those with none or just one request change less.



Guests who make more special requests tend to stay longer. For example, those with no special requests stay about 3.5 nights on average, while guests with 5 special requests stay almost 5 nights. So, more requests usually mean a longer stay.

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

Most bookings are consistent (meaning guests stick to their original plans) in the Corporate and Complementary segments, with about 87% consistency. Direct and Offline TA/TO are close behind at around 85%. Groups and Online TA have lower consistency, and the Undefined segment has no consistent bookings.

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

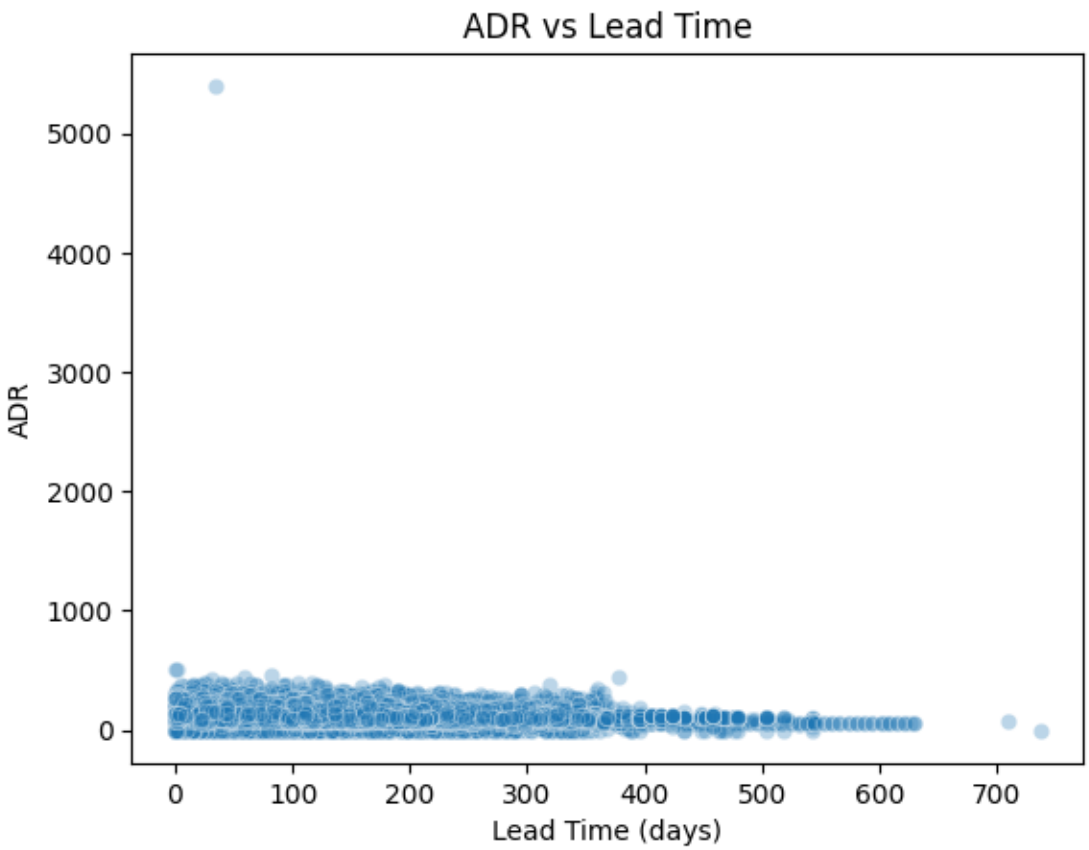
ADR is mostly influenced by revenue, number of guests, and special requests. Other factors like cancellations and stay length have minor effects.

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

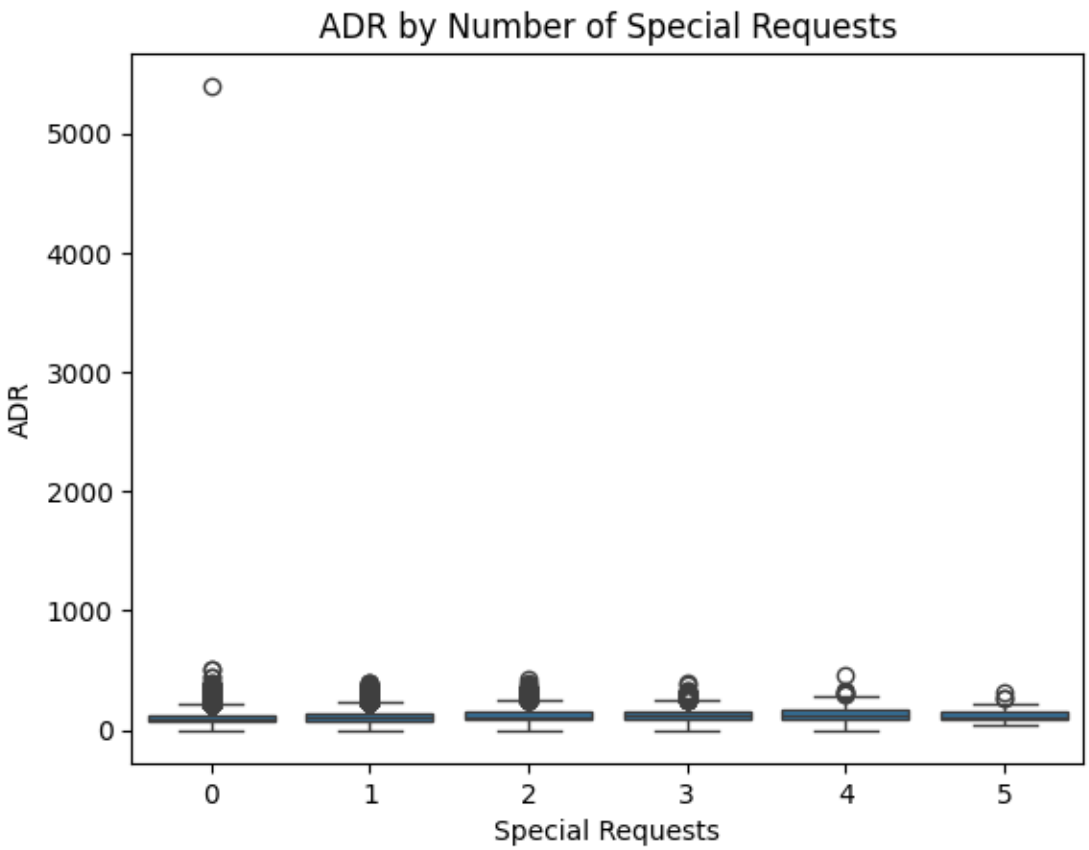
Transient customers contribute to higher revenue

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

As lead time increases (how early someone books), the average price per night (ADR) generally goes up — peaking around 120 to 180 days in advance. People who book last-minute (0–30 days) or very early (after 210 days) tend to pay a bit less on average. Most bookings happen within the first 30 days.



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



As the number of special requests increases, the average price paid per night (ADR) also goes up. Guests with no requests paid around ₹99, while those with 4–5 special requests paid over ₹130 on average. This suggests guests asking for more tend to book higher-priced rooms.

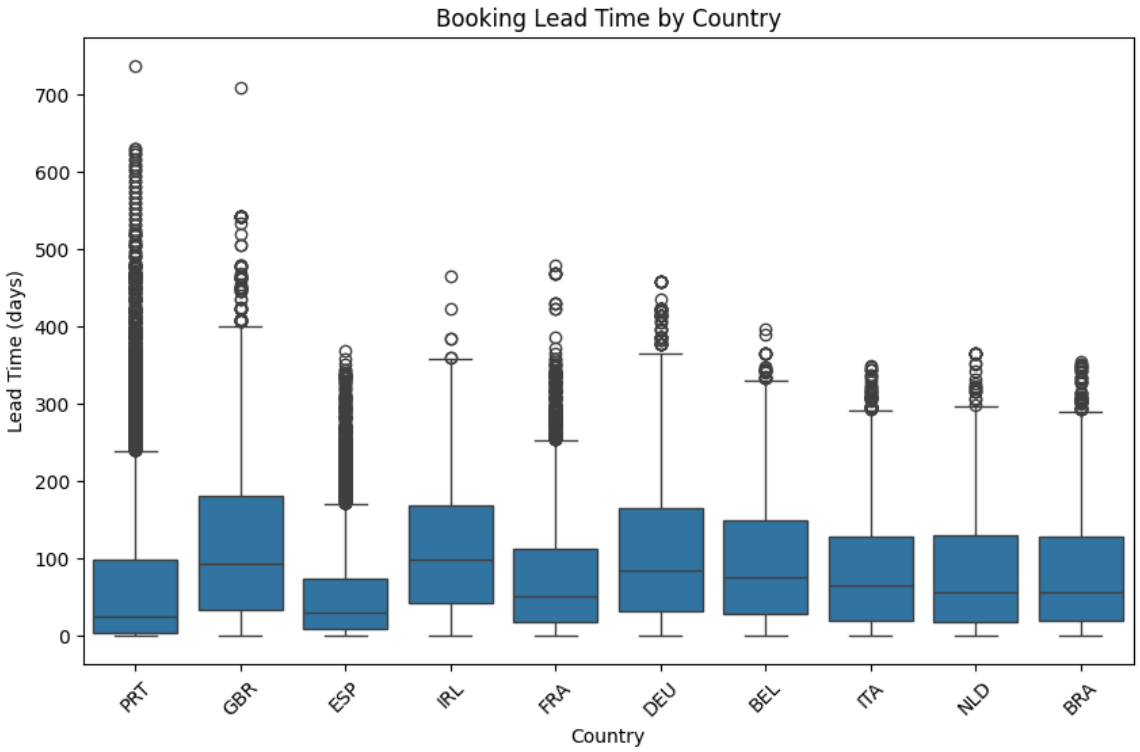
18. Do guests from different countries behave differently in terms of booking timing or stay length?

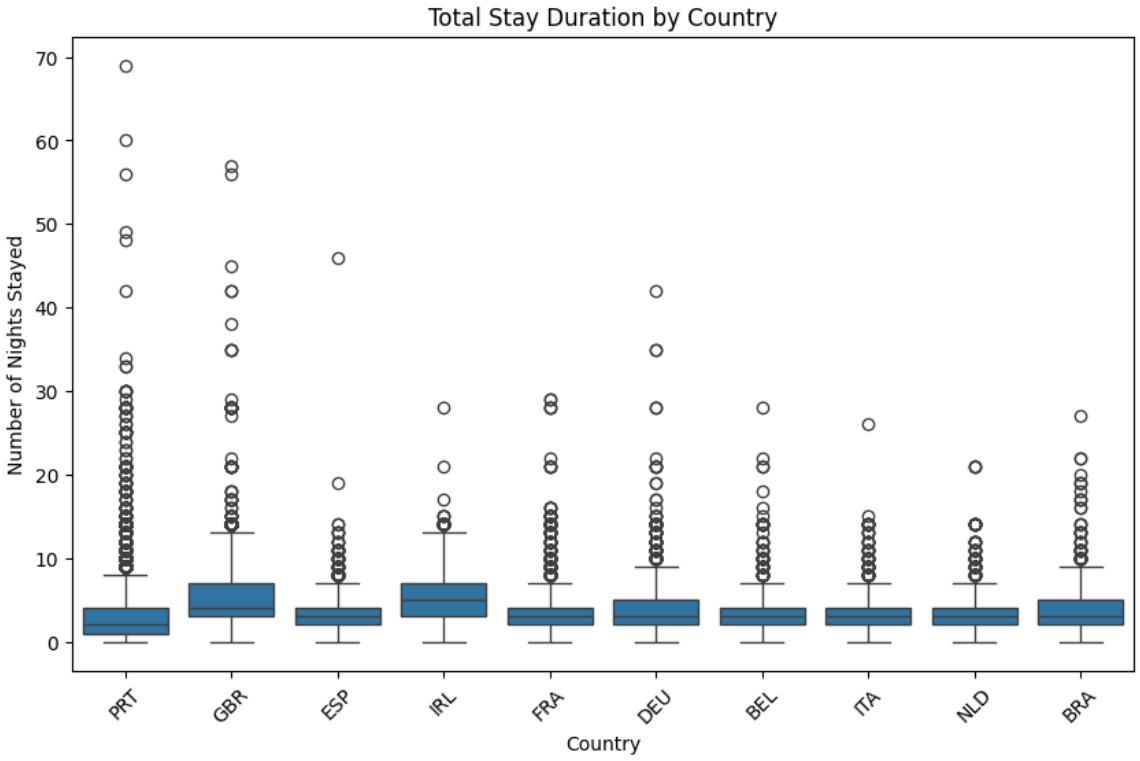
Almost all the countries have max lead time that is greater than 200

All the countries guest stay time is in between 10 to 20 days on average

So as to conculde I can say countries do not behave differently in terms of booking timing or stay length

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





Here's a short, layman-style summary:

PT **Portugal** had the most bookings, but people booked with shorter notice (avg \~64 days).

GB **UK**, IE **Ireland**, and DE **Germany** had much longer planning times—over 100 days on average.

ES **Spain** and BR **Brazil** had shorter lead times, showing more last-minute bookings.

Overall, lead time varies a lot by country—some plan way ahead, others book closer to their stay.

9. Conclusion

The analysis shows that booking behavior varies by country, customer type, and market segment. Most guests book their stays 1–3 months in advance, with longer lead times often linked to cancellations. Transient guests and online platforms dominate bookings but also show higher cancellation rates. More special requests usually mean longer stays and higher prices. Countries like the UK and Germany plan way ahead, while others book closer to their stay. Overall, planning patterns, guest types, and booking sources greatly influence hotel revenue and operations.