1. Introduction

This report presents a comprehensive analysis of hotel booking data using Python. The dataset contains records of hotel bookings for city and resort hotels, with features covering customer demographics, booking details, and financial data. The objective was to clean and preprocess the dataset, conduct statistical and visual analyses, and generate actionable insights regarding customer behavior and revenue metrics like ADR (Average Daily Rate).

---

2. Objectives

- Clean and prepare the dataset for analysis
- Handle missing values and identify outliers
- Explore booking patterns using univariate, bivariate, and multivariate analysis
- Analyze time-series trends
- Identify factors that influence ADR and cancellations
- Support business decision-making using correlation analysis and visual exploration

---

3. Project Summary

The dataset contains detailed information about hotel bookings. The data includes over 119,000 rows and 32 columns. The project was divided into several phases:

---

4. Data Cleaning and Preprocessing

Key steps:

- Imported essential libraries: pandas, numpy, seaborn, matplotlib, scipy, and statsmodels.
- Loaded the dataset: hotel_bookings.csv
- Inspected the dataset structure and identified missing values.
- Filled missing values:
    - Filled missing values in children, country, and agent columns using the mode.
    - Dropped the company column due to excessive missing values (>93%).

- Converted the arrival_date_month from month names to numerical format for time-series analysis.
- Created a new arrival_date column combining year, month, and day.
- Created new features such as total_stay = stays_in_week_nights + stays_in_weekend_nights.

Code Used:

```
df['children'].fillna(df['children'].mode()[0], inplace=True)
df['country'].fillna('Unknown', inplace=True)
df['agent'].fillna(df['agent'].mode()[0], inplace=True)
df.drop(columns='company', inplace=True)
df['total_stay'] = df['stays_in_week_nights'] + df['stays_in_weekend_nights']
```

---

5. Exploratory Data Analysis (EDA)

A. Univariate Analysis

- Histograms and count plots were used to examine the distribution of:
    - ADR
    - Lead time
    - Market segment
    - Customer type

B. Bivariate & Multivariate Analysis

- Boxplots and scatter plots revealed relationships between:
    - ADR and lead time
    - ADR and customer type
    - Booking changes and special requests

C. Time Series Analysis

- Booking trends were plotted over time using the arrival_date column.
- Monthly trends were observed to identify seasonality and peak periods.

Example code:

```
df['arrival_date'] = pd.to_datetime(df[['arrival_date_year', 'arrival_date_month', 'arrival_date_day_of_month']])
monthly_bookings = df.groupby(df['arrival_date'].dt.to_period('M')).size()
monthly_bookings.plot(kind='line')
```

---

6. Correlation Analysis

- Pearson correlation matrix was calculated.
- Heatmap was used to visualize correlations.
- Key findings:
    - adr has a moderate positive correlation with total_of_special_requests and lead_time.
    - Weak correlation between booking_changes and adr.

Example Code:

```
corr_matrix = df.corr(numeric_only=True)
sns.heatmap(corr_matrix[['adr']].sort_values(by='adr', ascending=False), annot=True)
```

---

7. Key Business Questions Answered (based on your analysis)

1.What influences ADR the most?

Ans: More guests (children/adults) → Higher ADR. Special requests moderately increase ADR. Longer stays slightly impact ADR. Parking and lead time have minimal effect. Booking changes have negligible impact.

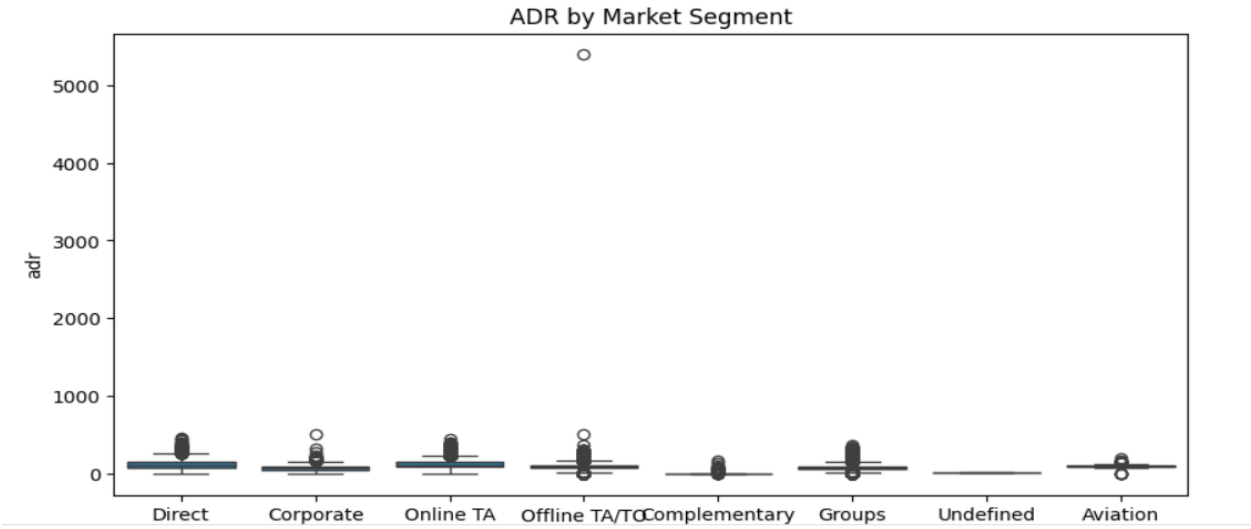
Code:

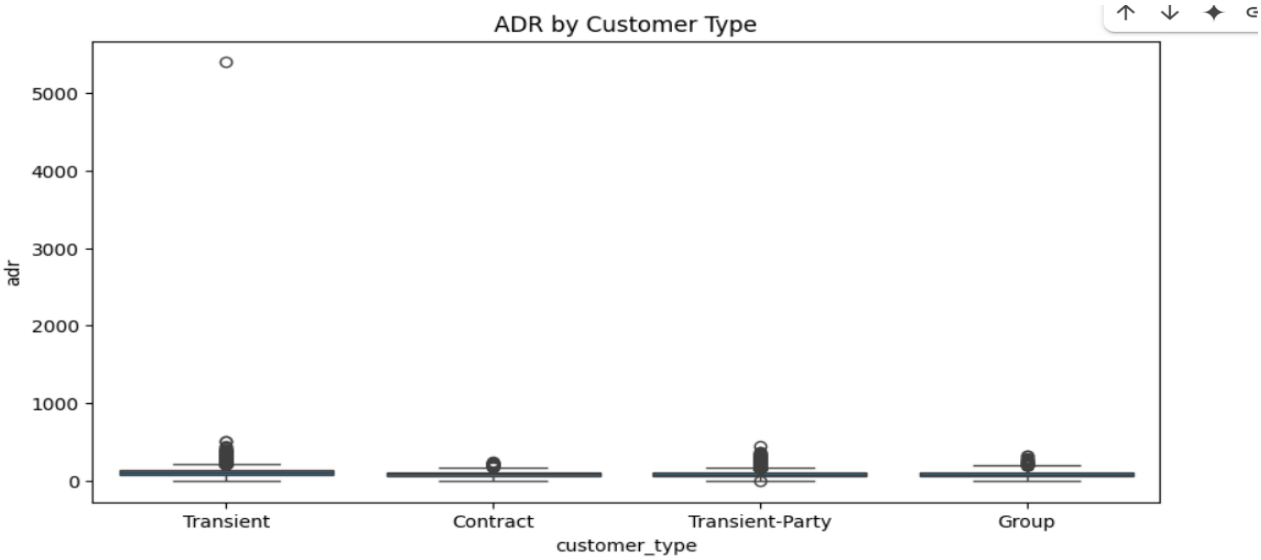
```
plt.figure(figsize=(10, 5))
sns.boxplot(x='market_segment', y='adr', data=df)
plt.title("ADR by Market Segment")
plt.show()

plt.figure(figsize=(10, 5))
sns.boxplot(x='customer_type', y='adr', data=df)
plt.title("ADR by Customer Type")
plt.show()
```

```
plt.figure(figsize=(10, 5))
sns.boxplot(x='hotel', y='adr', data=df)
plt.title("ADR by Hotel Type")
plt.show()

plt.figure(figsize=(10, 5))
sns.boxplot(x='reserved_room_type', y='adr', data=df)
plt.title("ADR by Reserved Room Type")
plt.xticks(rotation=45)
plt.show()
```





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

Ans: There is a very weak positive correlation between lead_time and booking_changes

Code:

np.corrcoef(df['lead_time'], df['booking_changes'])

Output:

array([[1.00000000e+00, 1.48830073e-04],

[1.48830073e-04, 1.00000000e+00]])

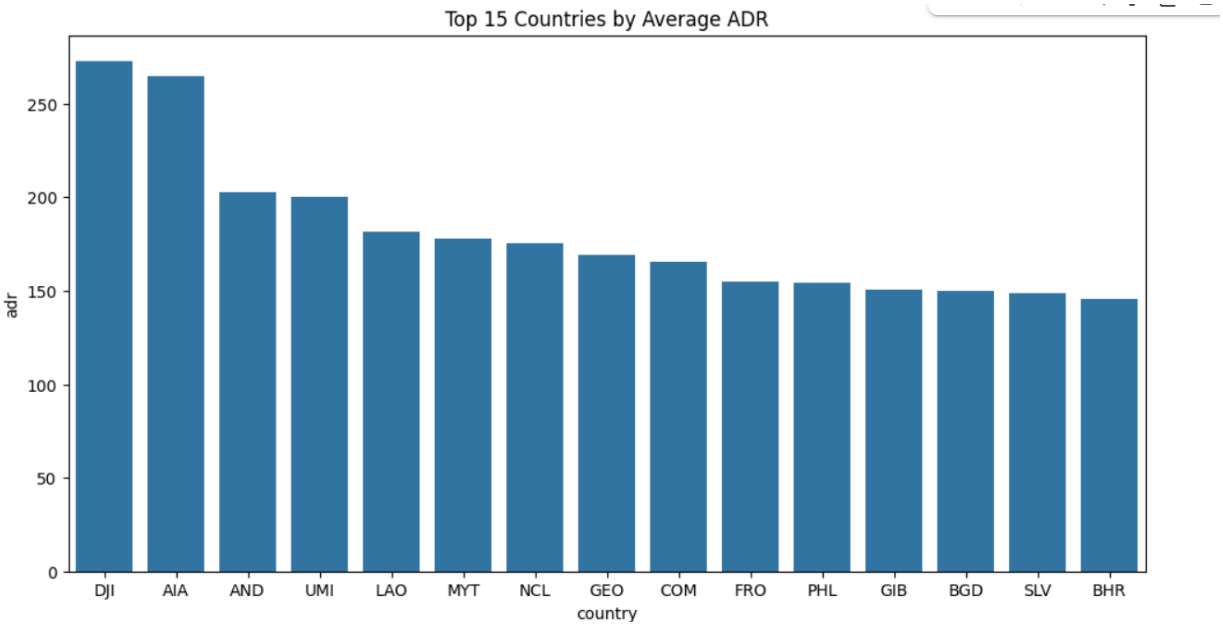
3.Are there pricing or booking differences across countries?

Ans:

Code:

```
country_stats = df.groupby('country')[['adr', 'lead_time']].mean()
top_countries = country_stats.sort_values(by='adr', ascending=False).head(15)
top_countries = top_countries.reset_index()
print(top_countries)

plt.figure(figsize=(12, 6))
sns.barplot(x='country', y='adr', data=top_countries)
plt.title("Top 15 Countries by Average ADR")
```

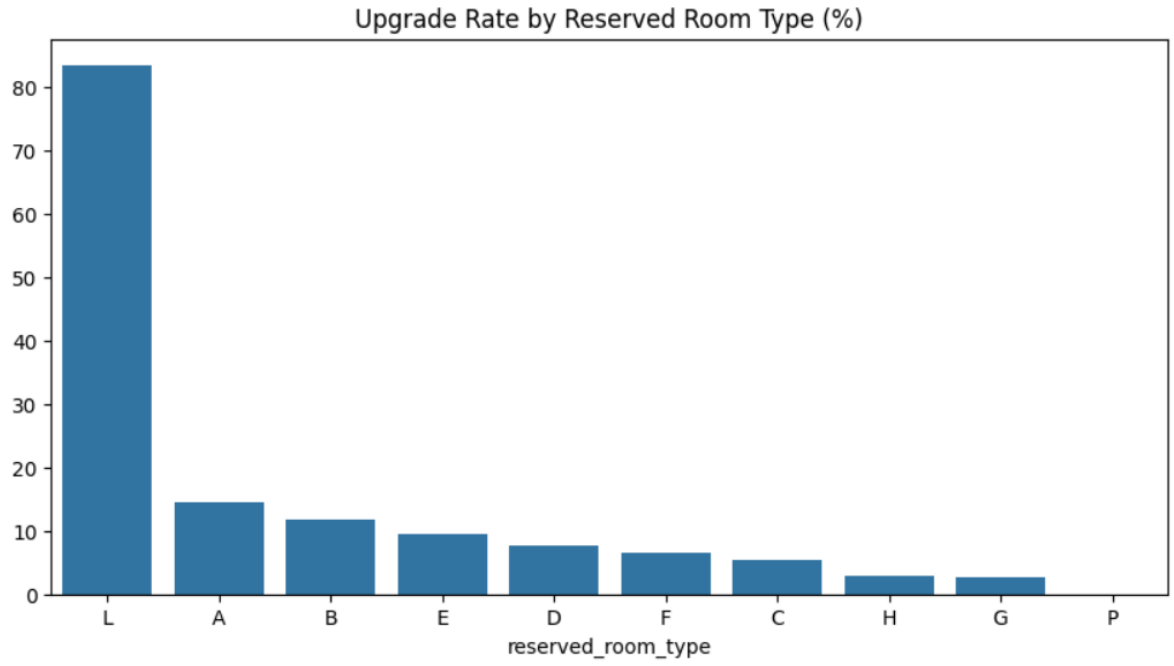


4. Is there a pattern in room upgrades or reassignment?

Ans:

Code:

```
df['is_upgraded'] = (df['reserved_room_type'] != df['assigned_room_type']).astype(int)
upgrade_by_room =
df.groupby('reserved_room_type')['is_upgraded'].mean().sort_values(ascending=False) * 100
plt.figure(figsize=(10, 5))
sns.barplot(x=upgrade_by_room.index, y=upgrade_by_room.values)
plt.title("Upgrade Rate by Reserved Room Type (%)")
plt.show()
```



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

Ans:

Code:

```
df['room_matched'] = (df['reserved_room_type'] == df['assigned_room_type']).astype(int)
match_counts = df['room_matched'].value_counts(normalize=True).rename({1: 'Matched', 0:
'Mismatched'}) * 100
print("Room Match Status (%):\n", match_counts)
```

Output:

```
Room Match Status (%):
 room_matched
```

Matched 87.505654
Mismatched 12.494346
Name: proportion, dtype: float64

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

```
group_size_counts = df['total_guests'].value_counts().sort_index()
top_nationalities = df['country'].value_counts().head(10)
top_nationalities, group_size_counts.head(10)
```

Output:
(country
PRT 48590
GBR 12129
FRA 10415
ESP 8568
DEU 7287
ITA 3766
IRL 3375
BEL 2342
BRA 2224
NLD 2104
Name: count, dtype: int64,
total_guests
0.0 180
1.0 22581
2.0 82051
3.0 10495
4.0 3929
5.0 137
6.0 1
10.0 2
12.0 2
20.0 2
Name: count, dtype: int64)

7 Do guest types influence booking behavior?

Ans:

```
df['is_canceled'] = df['is_canceled'].astype(int)
customer_stats = df.groupby('customer_type')[['adr', 'lead_time', 'is_canceled']].mean()
customer_counts = df['customer_type'].value_counts()
customer_stats['booking_count'] = customer_counts
customer_stats = customer_stats.sort_values(by='adr', ascending=False).reset_index()
customer_stats
```

Output:

	customer_type	adr	lead_time	is_canceled	booking_count
0	Transient	107.013621	93.295515	0.407463	89613
1	Contract	87.549637	142.969823	0.309617	4076
2	Transient-Party	86.084253	137.037056	0.254299	25124
3	Group	83.488579	55.057192	0.102253	577

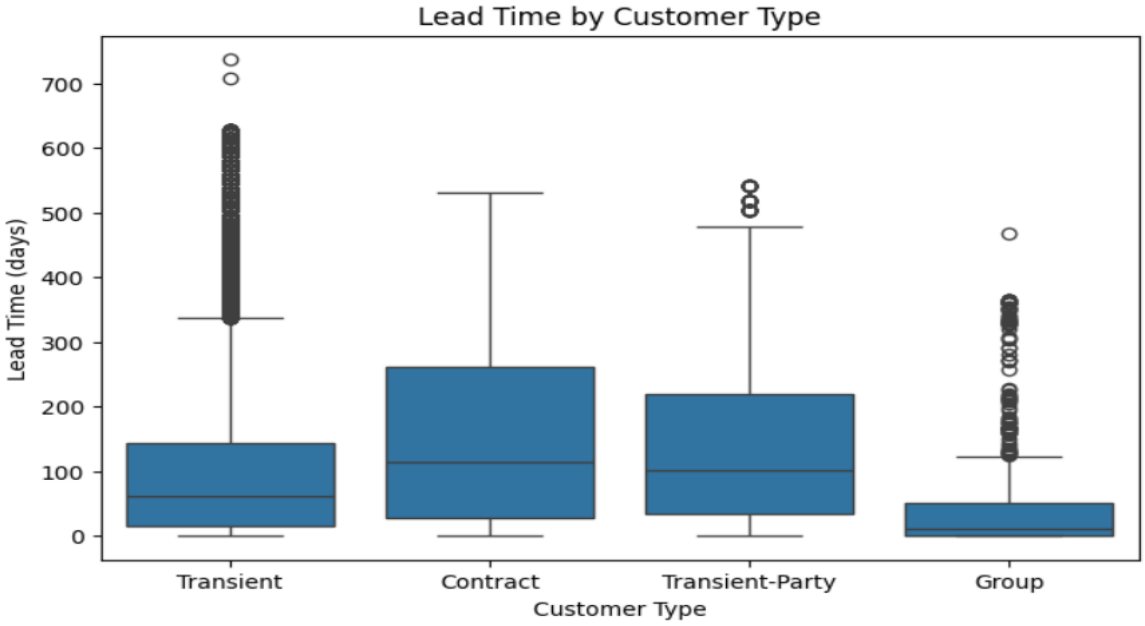
8. How does lead time vary by customer type and country?

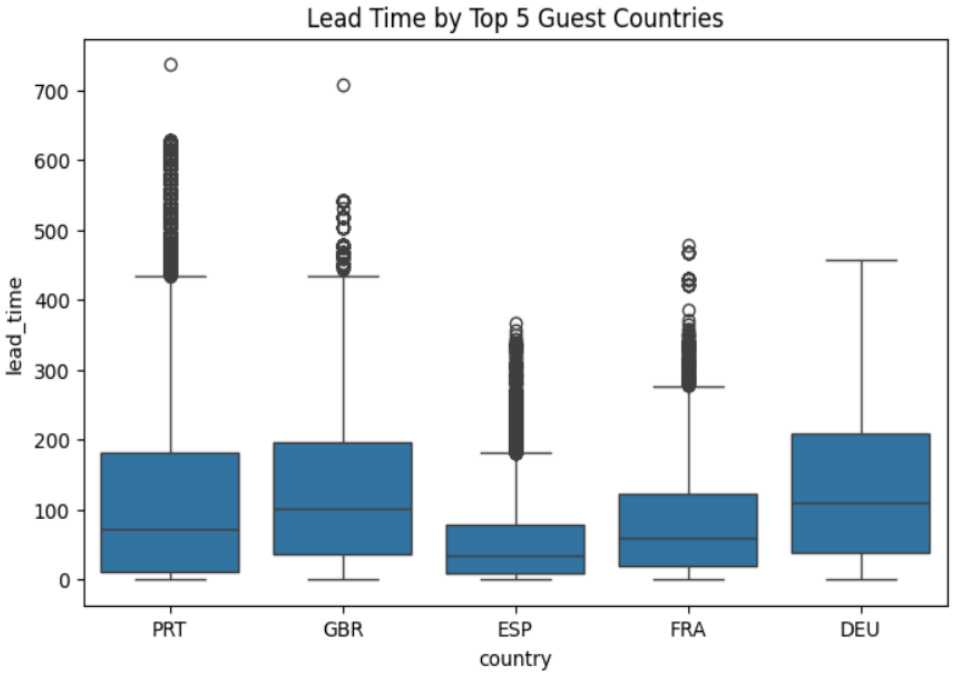
Ans:

```
plt.figure(figsize=(8, 5))
sns.boxplot(x='customer_type', y='lead_time', data=df)
plt.title("Lead Time by Customer Type")
plt.ylabel("Lead Time (days)")
plt.xlabel("Customer Type")
```

```
top_countries = df['country'].value_counts().head(5).index
df_top_countries = df[df['country'].isin(top_countries)]
plt.figure(figsize=(8, 5))
sns.boxplot(x='country', y='lead_time', data=df_top_countries)
plt.title("Lead Time by Top 5 Guest Countries")
```

Output:


Lead Time by Customer Type
Lead Time (days)
0
100
200
300
400
500
600
700
Transient
Contract
Transient-Party
Group
Customer Type


Lead Time by Top 5 Guest Countries
lead_time
0
100
200
300
400
500
600
700
PRT
GBR
ESP
FRA
DEU
country

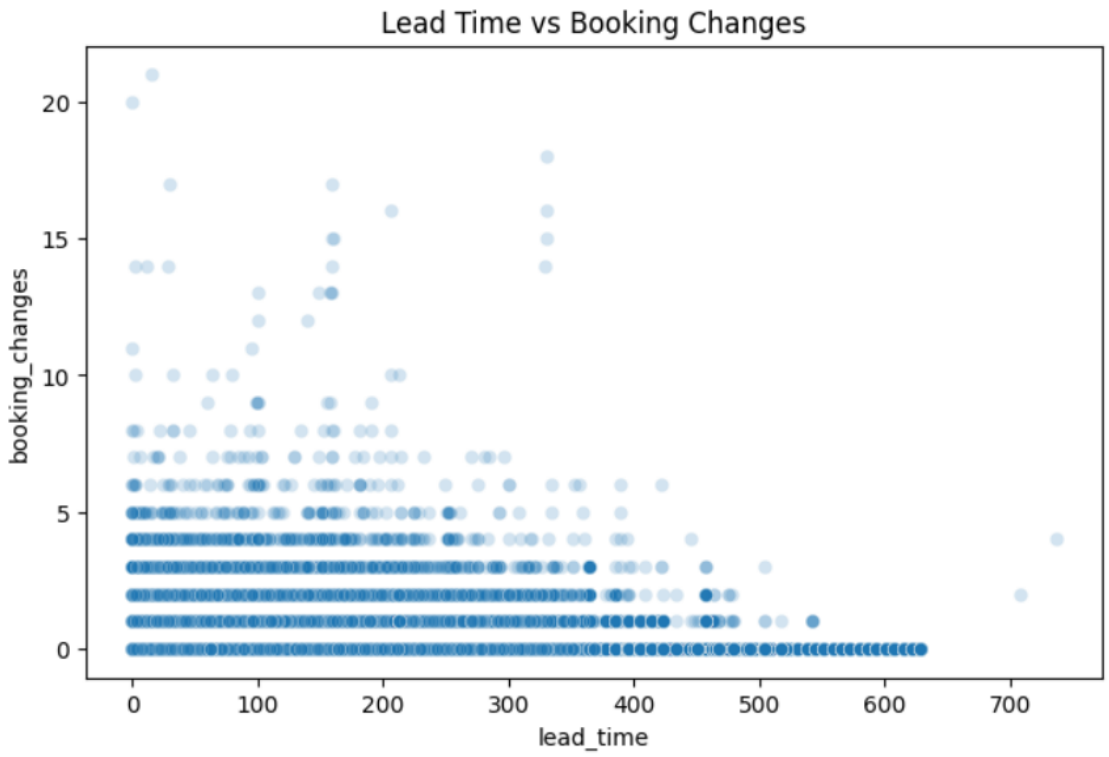
9. Are longer lead times associated with fewer changes or cancellations?

Ans:

```
plt.figure(figsize=(8, 5))
sns.scatterplot(x='lead_time', y='booking_changes', data=df, alpha=0.2)
plt.title("Lead Time vs Booking Changes")
```

Output:

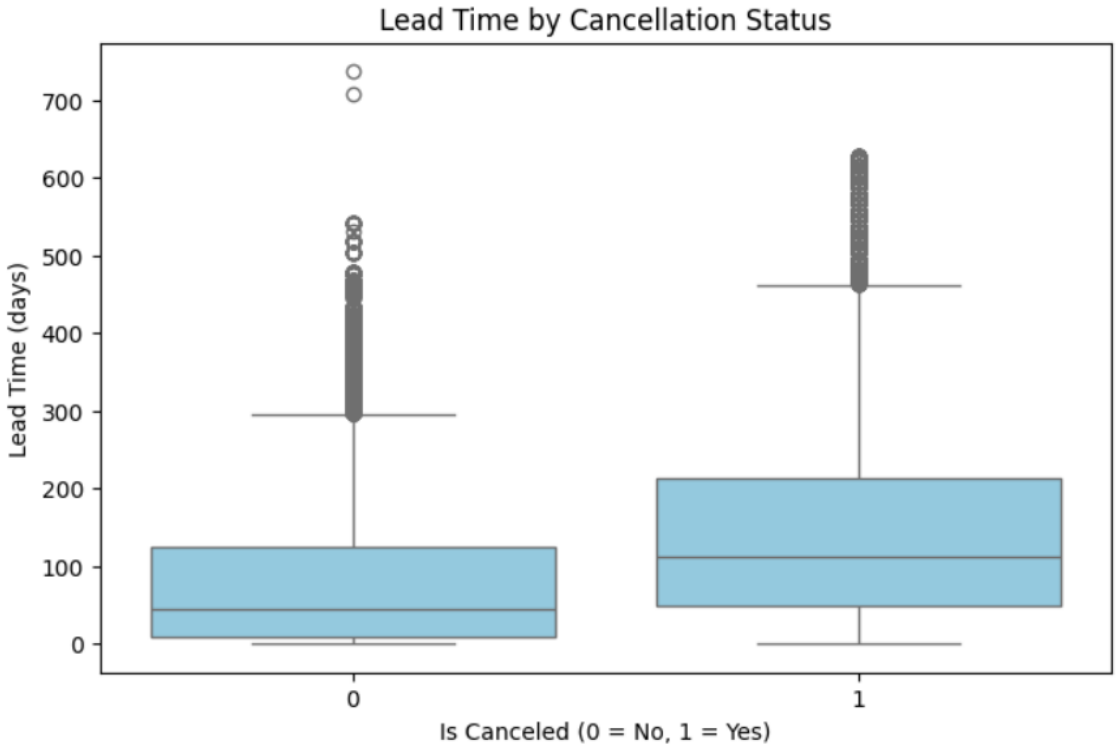
```
Text(0.5, 1.0, 'Lead Time vs Booking Changes')
```




```
plt.figure(figsize=(8, 5))
sns.boxplot(x='is_canceled', y='lead_time', data=df, color='skyblue')  # or remove `palette`
plt.title("Lead Time by Cancellation Status")
plt.xlabel("Is Canceled (0 = No, 1 = Yes)")
plt.ylabel("Lead Time (days)")
```

Output:

```
Text(0, 0.5, 'Lead Time (days)')
```



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

Ans:

stay_by_customer_type = df.groupby('customer_type')['total_stay'].mean().sort_values(ascending=False)

stay_by_market_segment = df.groupby('market_segment')['total_stay'].mean().sort_values(ascending=False)

stay_by_customer_type, stay_by_market_segment

Output:

```
(customer_type
 Contract            5.320658
 Transient           3.447145
 Transient-Party     3.064719
 Group               2.882149
 Name: total_stay, dtype: float64,
 market_segment
 Offline TA/TO     3.903877
 Aviation          3.607595
 Online TA         3.573986
 Direct            3.205775
 Groups            2.992529
 Corporate         2.092918
 Complementary     1.647376
 Undefined         1.500000
 Name: total_stay, dtype: float64)
```

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

Ans:

```
upgrade_distribution = df['is_upgraded'].value_counts(normalize=True).rename({0: 'No Upgrade', 1: 'Upgraded'}) * 100

upgrade_distribution
```

Output:

is_upgraded	proportion
No Upgrade	87.505654
Upgraded	12.494346

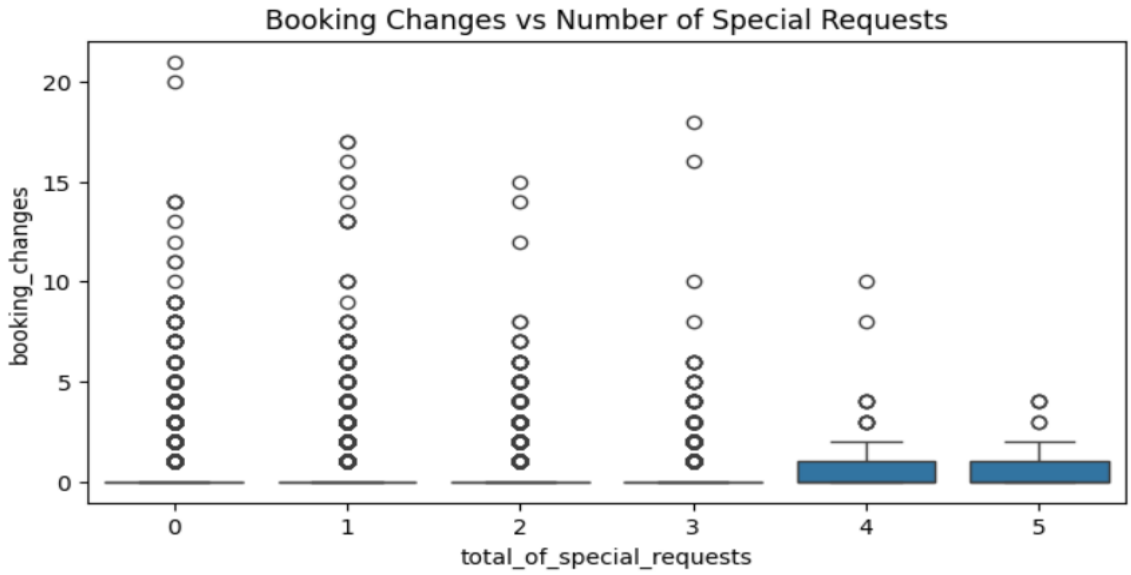
dtype: float64

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

Ans:

```
plt.figure(figsize=(8, 4))
sns.boxplot(x='total_of_special_requests', y='booking_changes', data=df)
plt.title("Booking Changes vs Number of Special Requests")
plt.show()
```

Output:



Code:
np.corrcoef(df.total_of_special_requests, df.booking_changes),
np.corrcoef(df.total_of_special_requests, df.total_stay)

Output:

```
(array([[1.        , 0.05283344],
        [0.05283344, 1.        ]]),
 array([[1.        , 0.07925878],
        [0.07925878, 1.        ]]))
```

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

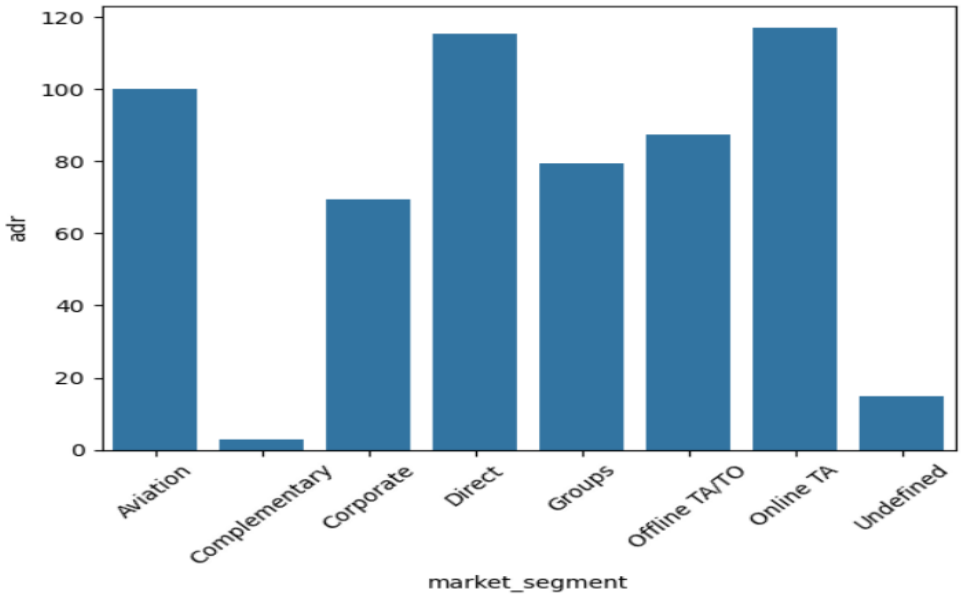
Ans:

Code

```
segment_stats = df.groupby('market_segment')[['adr', 'is_canceled']].mean()
channel_stats = df.groupby('distribution_channel')[['adr', 'is_canceled']].mean()
sns.barplot(x=segment_stats.index, y=segment_stats['adr']);
plt.xticks(rotation=45)
```

Output:

([0, 1, 2, 3, 4, 5, 6, 7], [Text(0, 0, 'Aviation'), Text(1, 0, 'Complementary'), Text(2, 0, 'Corporate'), Text(3, 0, 'Direct'), Text(4, 0, 'Groups'), Text(5, 0, 'Offline TA/TO'), Text(6, 0, 'Online TA'), Text(7, 0, 'Undefined')])



Online TA gives highest ADR but higher risk of cancellations.
Direct bookings offer high ADR and low cancellation — most reliable.
Group bookings cancel often — need caution when forecasting.
GDS & Direct are high-value channels with acceptable cancellation rates.
TA/TO has high revenue but high volatility.

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

```
correlation = df.corr(numeric_only=True)['adr'].sort_values(ascending=False)
print("Top correlations with ADR:\n", correlation.head(10))
```

Output:

```
Top correlations with ADR:
 adr                          1.000000
total_revenue                 0.565766
total_guests                  0.368105
children                      0.324853
adults                        0.230641
arrival_date_year             0.197580
total_of_special_requests     0.172185
room_matched                  0.138133
arrival_date_week_number      0.075791
total_stay                    0.067945
Name: adr, dtype: float64
```

As the magnitude of co-relation coefficient of total_revenue vs adr is close to 1, so we can consclude that total_revenue influences adr the most

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

```
revenue_by_customer =
df.groupby('customer_type')['total_revenue'].mean().sort_values(ascending=False)
revenue_by_customer
```

Output:

customer_type	total_revenue
Contract	451.196700
Transient	381.631057
Transient-Party	260.493376
Group	243.813328

dtype: float64

Contract and Transient customers contribute the highest per-booking revenue. Transient bookings offer both volume and value.
Online TA and Direct channels bring in the highest revenue. Complimentary bookings contribute nearly zero.

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

```
df['lead_time_bucket'] = pd.cut(df['lead_time'], bins=[0, 30, 90, 180, 365, df['lead_time'].max()],
                                labels=['0–30d', '31–90d', '91–180d', '181–365d', '365+d'])
adr_by_lead_time = df.groupby('lead_time_bucket', observed=True)['adr'].mean()
adr_by_lead_time
```

Output:

lead_time_bucket	
0–30d	101.922165
31–90d	106.572426
91–180d	109.125540
181–365d	95.073570
365+d	78.811569

dtype: float64

Bookings made 1–6 months in advance have the highest ADR. Very early bookings (1+ year) tend to be cheaper — likely discounted group/contract rates.

International guests (ESP, FRA) tend to pay more than local (PRT). Portugal may get discounts due to domestic proximity.

17. Are guests with higher ADR more likely to request special services or make booking modifications?
Ans:
np.corrcoef(df.adr, df.total_of_special_requests), np.corrcoef(df.adr, df.booking_changes)

Output:
```
(array([[1.        , 0.17218526],
        [0.17218526, 1.        ]]),
 array([[1.        , 0.01961767],
        [0.01961767, 1.        ]]))
```

Guests who pay higher tend to make more special requests, but not more booking changes.

18. Do guests from different countries behave differently in terms of booking timing or stay length?
Ans:
```
country_behavior = df.groupby('country')[['lead_time',
'total_stay']].mean().sort_values(by='lead_time', ascending=False)
country_behavior.head(10)
```

Output:

country	lead_time	total_stay
FJI	322.000000	3.000000
FRO	286.400000	12.000000
BEN	274.000000	2.333333
LCA	268.000000	5.000000
KNA	251.500000	2.000000
MYT	208.000000	3.500000
MKD	198.200000	3.700000
BRB	192.000000	2.000000
DOM	185.285714	4.214286
GUY	180.000000	3.000000

```
country_behavior = df.groupby('country')[['lead_time',
'total_stay']].mean().sort_values(by='total_stay', ascending=False)
country_behavior.head(10)
```

country		
FRO	286.400000	12.000000
SEN	55.727273	8.818182
AGO	23.096685	8.116022
TGO	62.000000	8.000000
GNB	40.888889	7.111111
PLW	169.000000	7.000000
BHS	159.000000	7.000000
SLE	84.000000	7.000000
RWA	102.000000	6.500000
CPV	25.000000	6.083333

Countries like PRT (Portugal) might show shorter lead times, suggesting local or last-minute bookings.
Guests from countries farther away (like USA, BRA, GBR) may have longer lead times due to international travel planning.

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

```
contingency = pd.crosstab(df['booking_changes'] > 0, df['is_canceled'])
chi2, p, dof, expected = stats.chi2_contingency(contingency)
chi2, p
```

Output:
(4169.746849050339, 0.0)

Ho: Guests who make booking changes are not more likely to cancel
As p 0 < 0.05, we reject.
H0, Guests who make booking changes are more likely to request services or cancel.

Code:

```
contingency = pd.crosstab(df['booking_changes'] > 0, df['total_of_special_requests'])
chi2, p, dof, expected = stats.chi2_contingency(contingency)
chi2, p
```

Output:
(311.0949136567454, 4.120380110344912e-65)

Ho: Guests who make booking changes are likely to request additional services
As p 0 < 0.05, we reject H0 ⇒ Guests who make booking changes are likely to request additional services.