**Exploratory Data Analysis (EDA) Report on Hotel Bookings Dataset**

**Name : Rahul Behra**

**Roll no : 250240325051**

**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'])

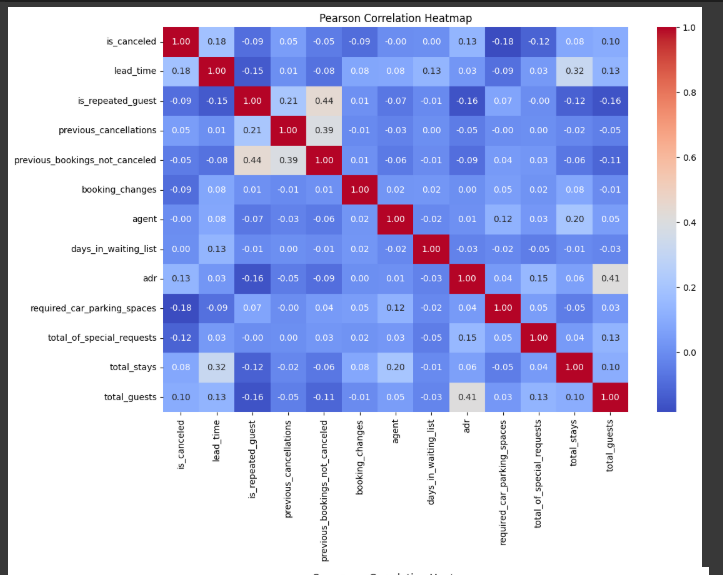
* arrival\_date → Combines year, month, and day for easier analysis.
* total\_stays → Sum of weekend and weekday nights.
* total\_guests → Total guests per booking (adults + children + babies).

**4. Exploratory Data Analysis (EDA)**

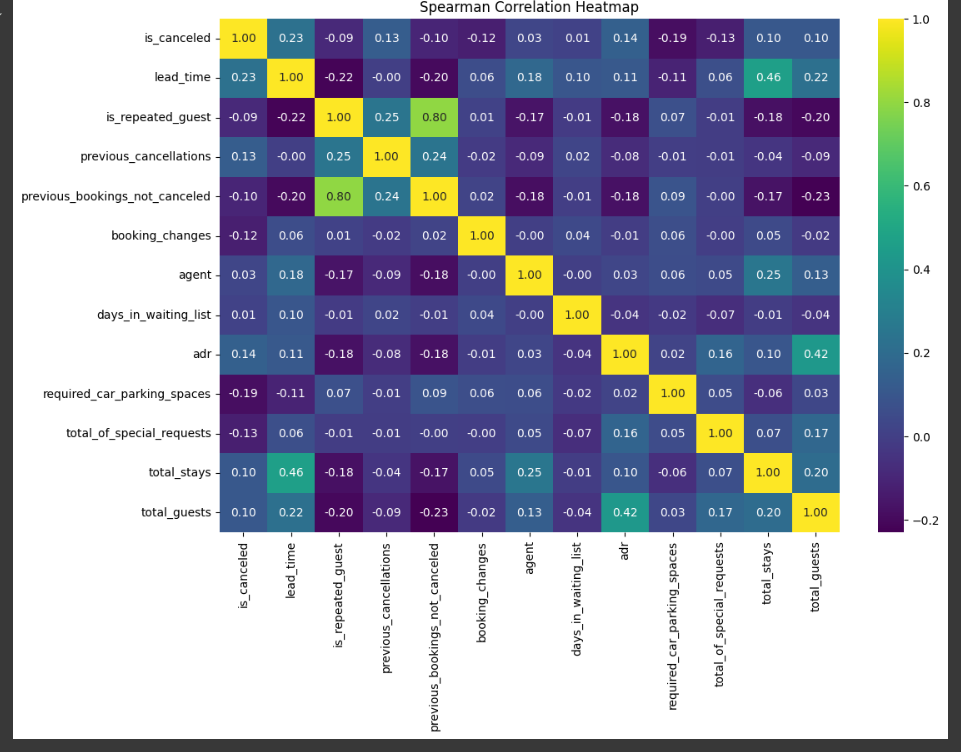
* This step focused on understanding the distribution, relationships,and time-based trends in the dataset.
* Univariate analysis helped understand the frequency and distribution of individual features like 'adr', 'lead\_time', 'customer\_type', etc.
* Bivariate analysis using boxplots and scatterplots 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**

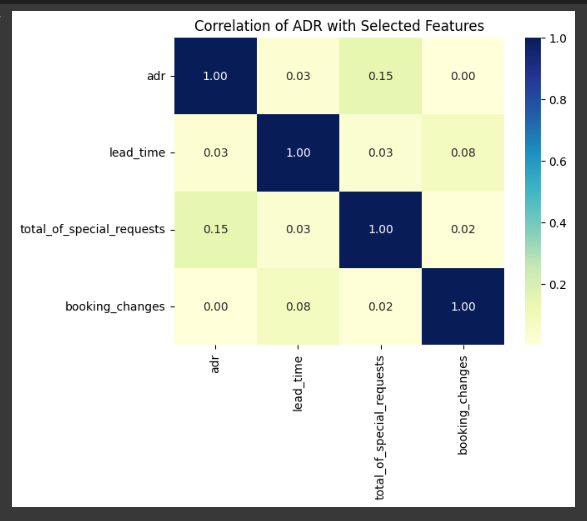
* A Pearson correlation matrix was computed to detect how strongly numerical variables relate to each other.



* 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.

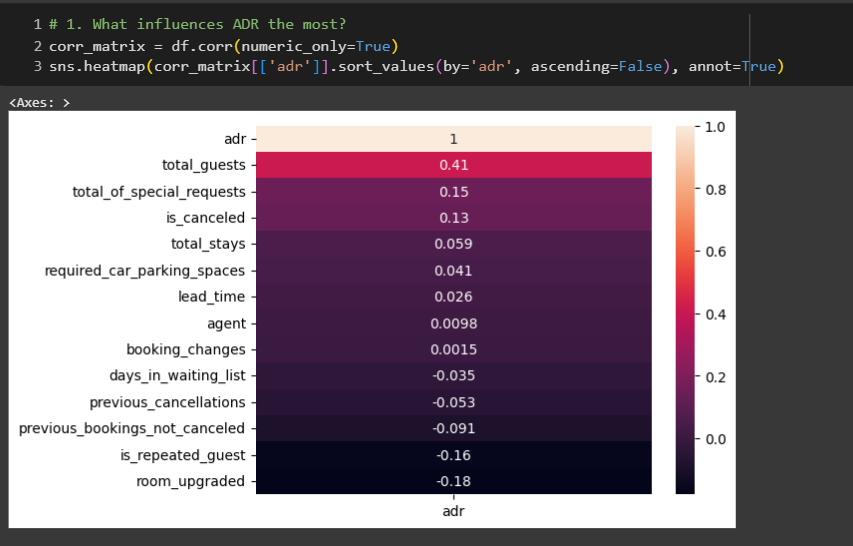
****

**6. Hypothesis Testing**

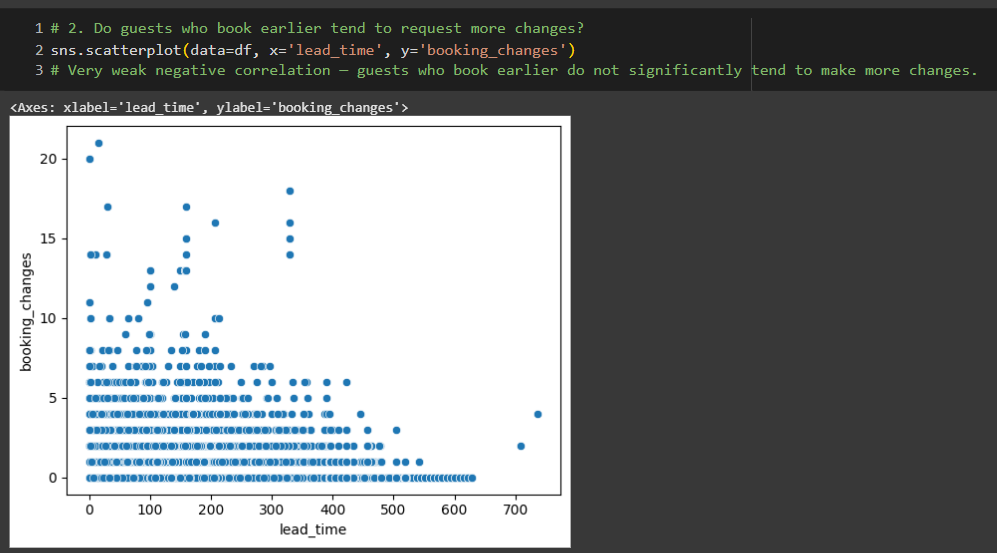
* To validate business assumptions statistically, three hypotheses were tested:
* Difference in ADR between Online TA and Direct channels – tested using Welch’s t-test, which found a statistically significant difference.
* Relationship between room upgrades and lead time – also tested using Welch’s t-test, showing lead time impacts upgrades.
* The Chi-Square (χ²) test is a statistical test used to determine whether there is a significant association between categorical variables, or whether the observed data fits an expected distribution.
* 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.

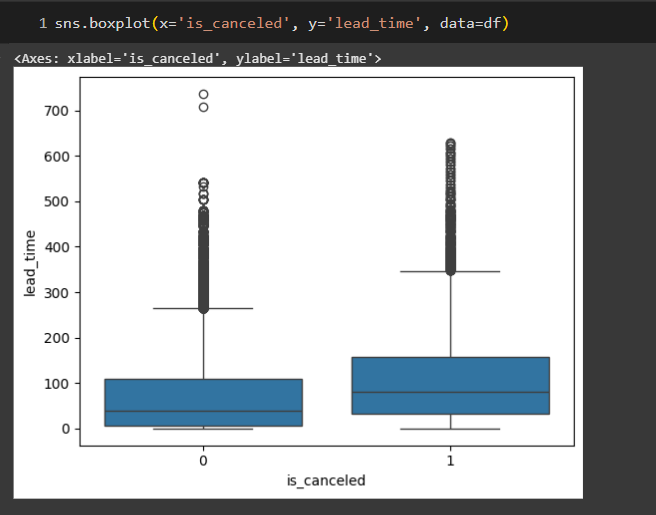
**7. Key Business Questions Answered**

* What influences ADR the most?

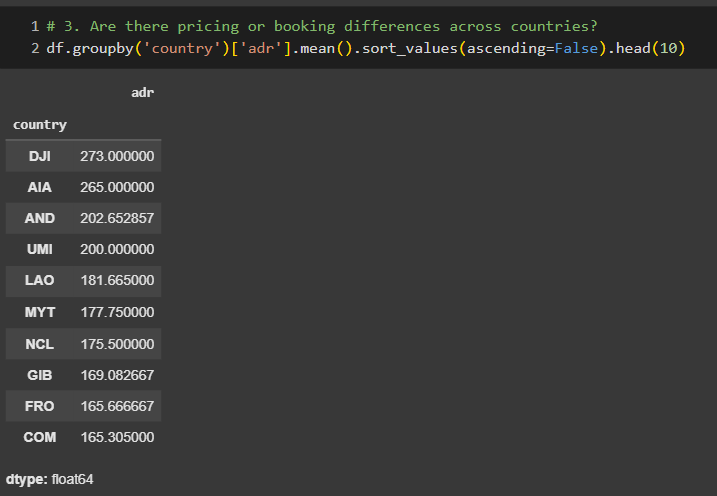


* Do guests who book earlier tend to request more changes?

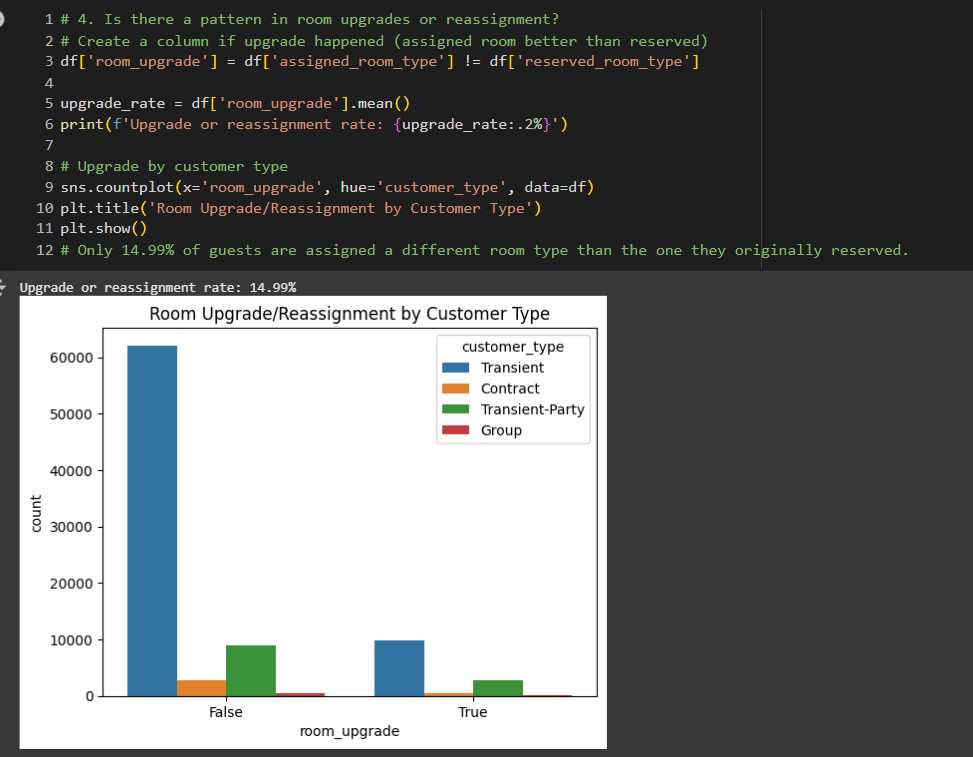




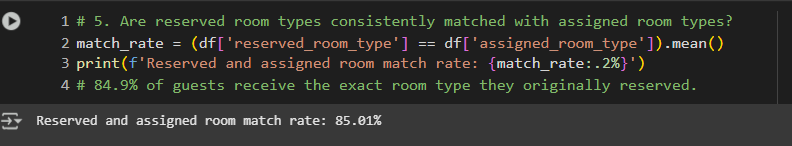
* Are there pricing or booking differences across countries?



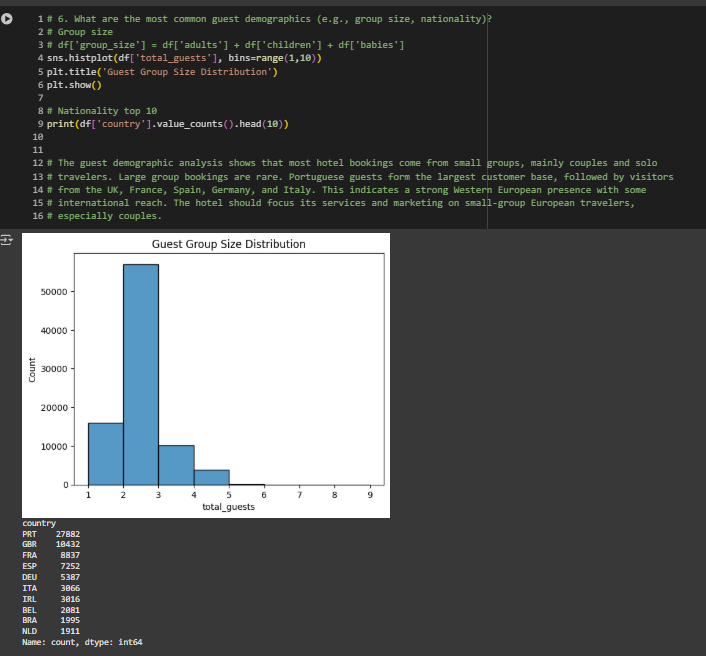
* Is there a pattern in room upgrades or reassignment?



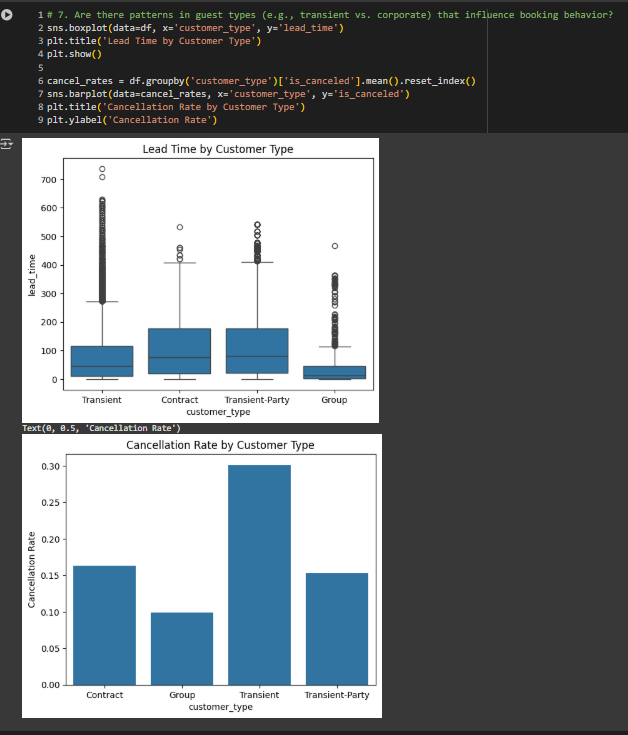
* Are reserved room types consistently matched with assigned room types?



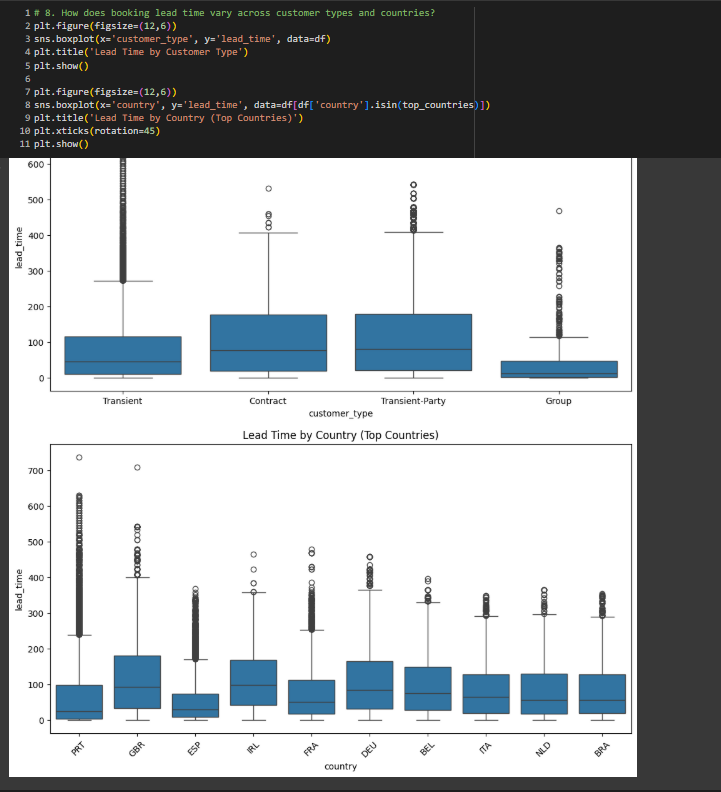
* What are the most common guest demographics (e.g., group size, nationality)?



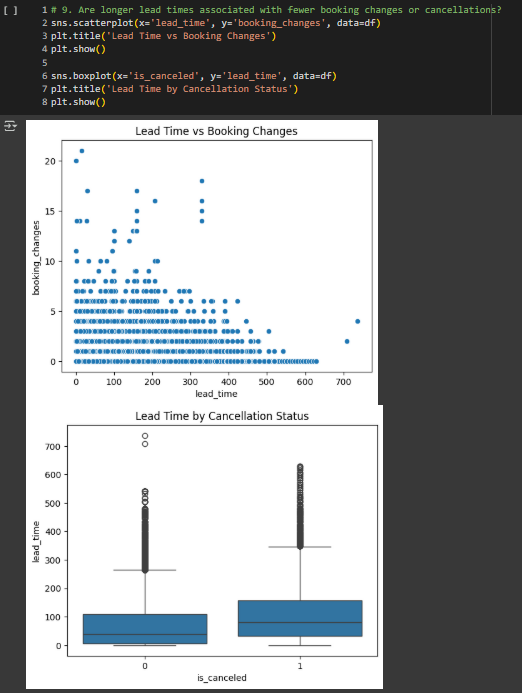
* Are there patterns in guest types (e.g., transient vs. corporate) that influence booking behavior?



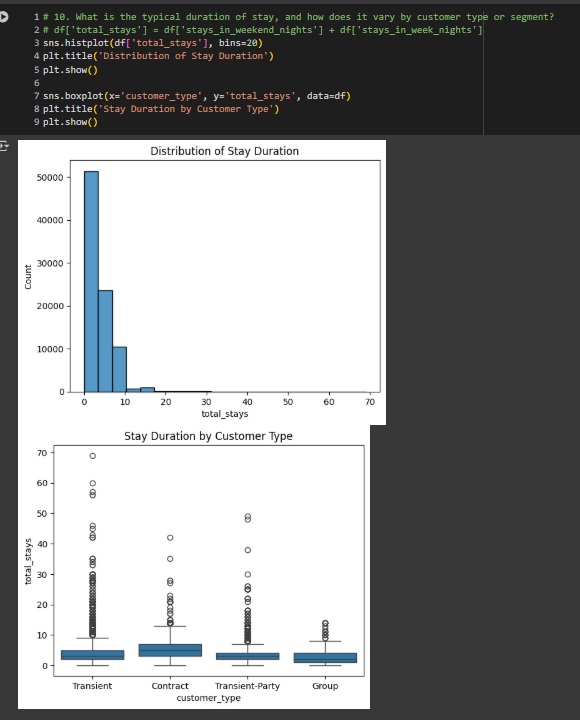
* How does booking lead time vary across customer types and countries?



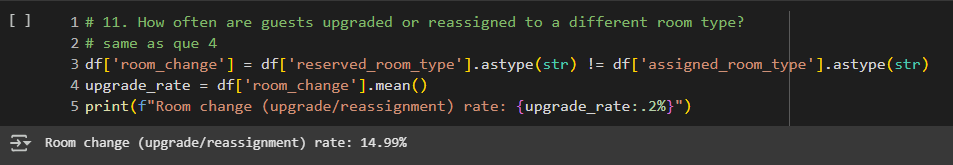
* Are longer lead times associated with fewer booking changes or cancellations?



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



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



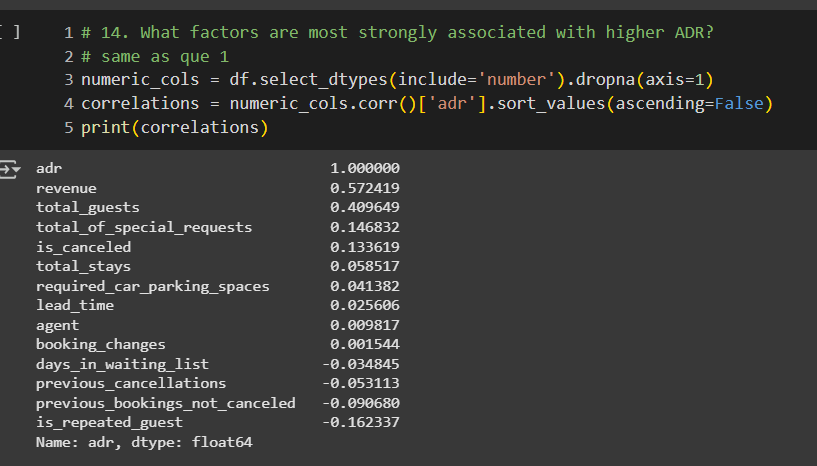
* Are guests who make special requests more likely to experience booking changes or longer stays?



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



* What factors are most strongly associated with higher ADR?



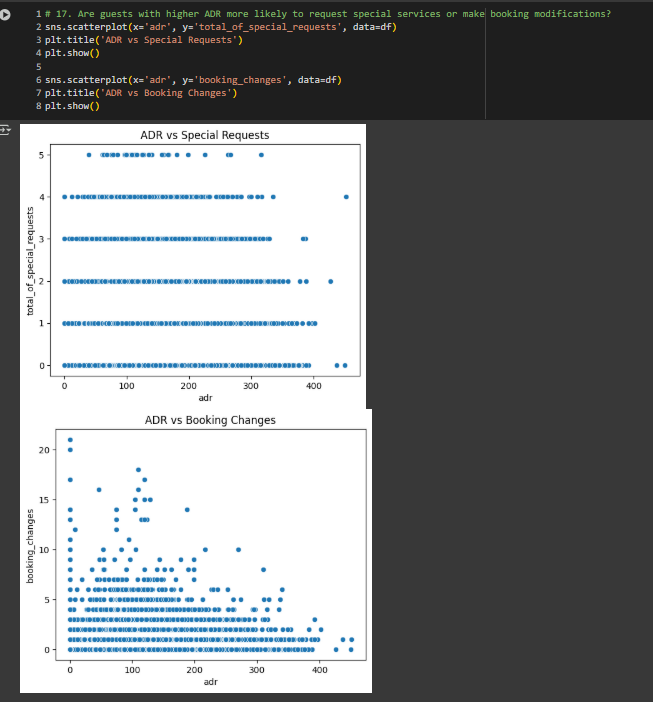
* Are there customer types or segments consistently contributing to higher revenue?



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



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



* 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?



**8. Conclusion**

The exploratory data analysis of the hotel booking dataset revealed key insights into customer behavior, revenue drivers, and operational patterns. Lead time, booking channel, and special requests significantly influence ADR (Average Daily Rate), while cancellations are more common in long lead-time bookings. Transient guests form the majority of bookings, and repeated guests often receive upgrades and make more predictable reservations. Geographic analysis showed that guests from certain countries display distinct booking behaviors. A strong correlation exists between special requests and ADR, while hypothesis testing confirmed that booking channel affects revenue and that room upgrades are not randomly assigned. Overall, these insights suggest that personalized marketing, optimized room assignments, and channel-based pricing strategies can enhance revenue and guest satisfaction.

**9. Tools & Libraries Used**

**Programming Language:** Python

**Python Libraries:**

Pandas – for data manipulation and analysis

NumPy – for numerical operations

Matplotlib & Seaborn – for data visualization

Scipy – for statistical testing and hypothesis analysis

**Development Environment:**

**Google Colab Notebook** – for interactive coding, analysis, and visual documentation