Advance Analytics Case Study - Hotel Bookings

Ankit Surolia

Under the guidance of:  
Mrs. Vineeta Singh  
(Hon. Course Coordinator, DBDA, C-DAC Mumbai)

# Acknowledgement

I would like to express my sincere gratitude to CDAC Kharghar Institute for granting me the opportunity to work on this analysis report on the Hotel Booking dataset. This project has allowed me to expand my knowledge in data analysis, strengthen my technical skills, and gain hands-on experience.  
  
I am deeply grateful to Mrs. Vineeta Singh, Hon. Course Coordinator, DBDA at C-DAC Mumbai, Mr. Prashant Sir and Mr. Nishad Sir for continuous support and i guidance

Thank you.

### ****Dataset Description****

The dataset used for this analysis consists of detailed booking information from two hotel types: **Resort Hotel** and **City Hotel**. Each row in the dataset represents an individual hotel booking and includes a wide range of features such as booking dates, guest demographics, stay details, pricing, and reservation status. The dataset contains both numerical and categorical variables that are essential for analyzing customer behavior, booking patterns, and operational outcomes.

**Key Attributes in the Dataset:**

* **Booking Details**: hotel, is\_canceled, lead\_time, arrival\_date\_\*, booking\_changes
* **Guest Composition**: adults, children, babies
* **Stay Details**: stays\_in\_weekend\_nights, stays\_in\_week\_nights, adr (Average Daily Rate)
* **Meal and Special Requests**: meal, total\_of\_special\_requests, required\_car\_parking\_spaces
* **Guest Origin**: country, agent, company
* **Reservation and Status**: reservation\_status, reservation\_status\_date, deposit\_type
* **Room Assignment**: reserved\_room\_type, assigned\_room\_type
* **Customer Type and Channel**: customer\_type, market\_segment, distribution\_channel
* **Previous Booking Behavior**: is\_repeated\_guest, previous\_cancellations, previous\_bookings\_not\_canceled

The dataset required significant preprocessing before conducting meaningful analysis, including handling missing values, transforming categorical variables, and creating derived metrics such as total stay duration and total guests per booking.

### ****Tools & Technologies Used****

To perform data cleaning, analysis, visualization, and hypothesis testing, the following tools and libraries were used:

* **Python 3.x**: Core programming language for data analysis.
* **Jupyter Notebook**: Interactive development environment for documenting code and visualizations.
* **Pandas**: For data manipulation and cleaning.
* **NumPy**: For numerical operations and efficient data handling.
* **Matplotlib & Seaborn**: For data visualization through histograms, box plots, heatmaps, and time-series plots.
* **Scipy & Statsmodels**: For hypothesis testing and statistical inference.

# 1. Introduction

Advance Analytics Case Study - Hotel Bookings

This project aims to analyze the Hotel Booking dataset to uncover patterns that can support revenue optimization and operational efficiency. The dataset used includes information such as booking details, customer profiles, lead times, and booking modifications.

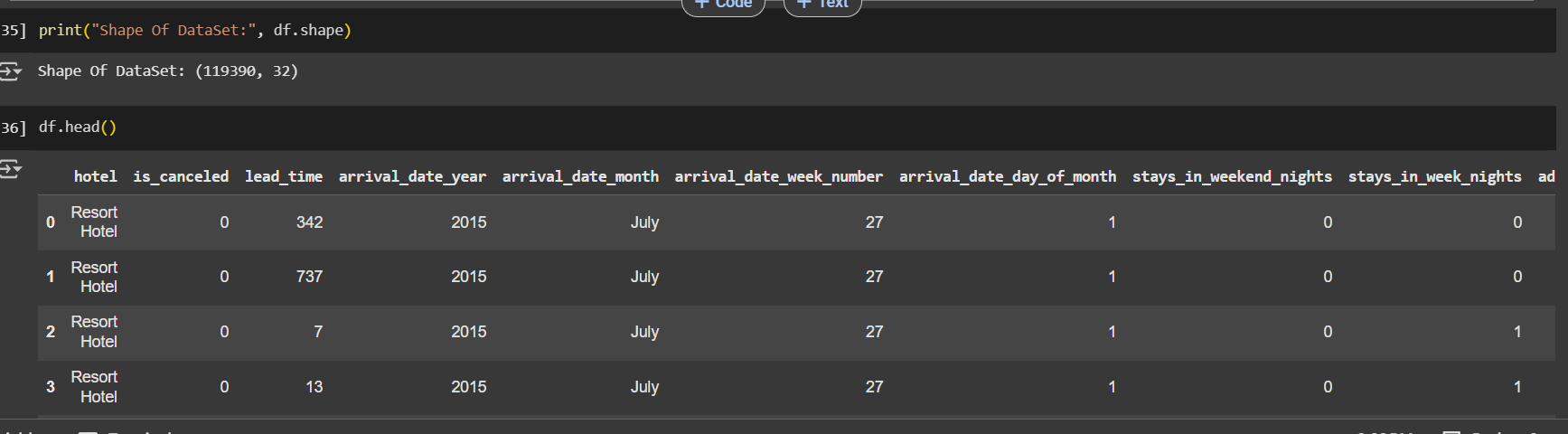
## ****1. Data Cleaning and Preprocessing****

Before performing any analysis, the dataset underwent several data cleaning steps to ensure its integrity and readiness for analysis:

* **Dropped Unnecessary Columns**:  
  The column agent was dropped as it was deemed irrelevant or not useful for the scope of the analysis.
* **Checked for Duplicates**:  
  Duplicate records were identified and removed to maintain data accuracy.
* **Missing Values Handling**:
  + The dataset was checked for missing values using isnull().sum().
  + The columns children, country, agent, company were found to contain missing values.
  + These were handled accordingly using imputation or removal strategies.
* **Outliers Handling**:  
  A boxplot was used to identify outliers in numeric columns such as lead\_time and adr.
* **Data Type Corrections and Feature Engineering**:
  + Date columns such as reservation\_status\_date were converted to datetime format.
  + A new column total\_stay was created by summing stays\_in\_weekend\_nights and stays\_in\_week\_nights.
  + Another feature total\_guests was created by summing adults, children, and babies.

## 

**1.1 Initial Shape of the Dataset**



### ****1.2 Checking for Null Values****

df.isnull().sum()

**Output (top relevant columns):**

* children: 4 missing
* country: 488 missing
* agent: 16340 missing
* company: 112593 missing

### ****1.3 Handling Missing Values****

* children: Missing values filled with 0
* country: Missing values filled with the mode (most frequent value)
* agent: Missing values filled with 0
* company: Missing values filled with 0

### ****1.3 Handling Missing Values****

* children: Missing values filled with 0
* country: Missing values filled with the mode (most frequent value)
* agent: Missing values filled with 0
* company: Missing values filled with 0

Post-duplicate Shape:

**Output:**  
(87496, 32)  
 A total of **31,994** duplicates were removed.

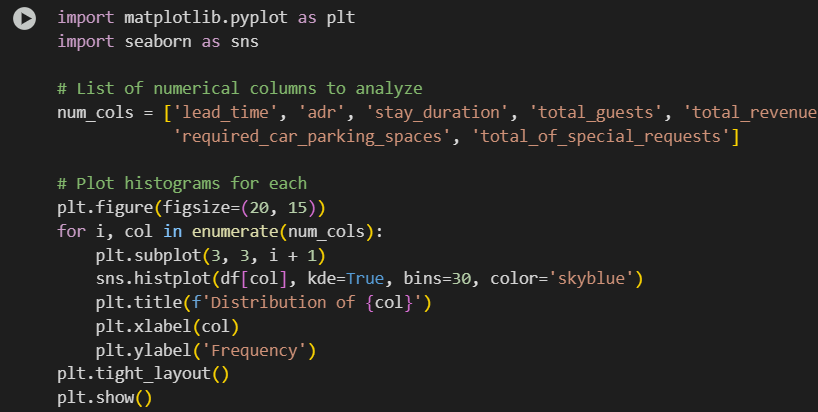
## ****2. Exploratory Data Analysis****

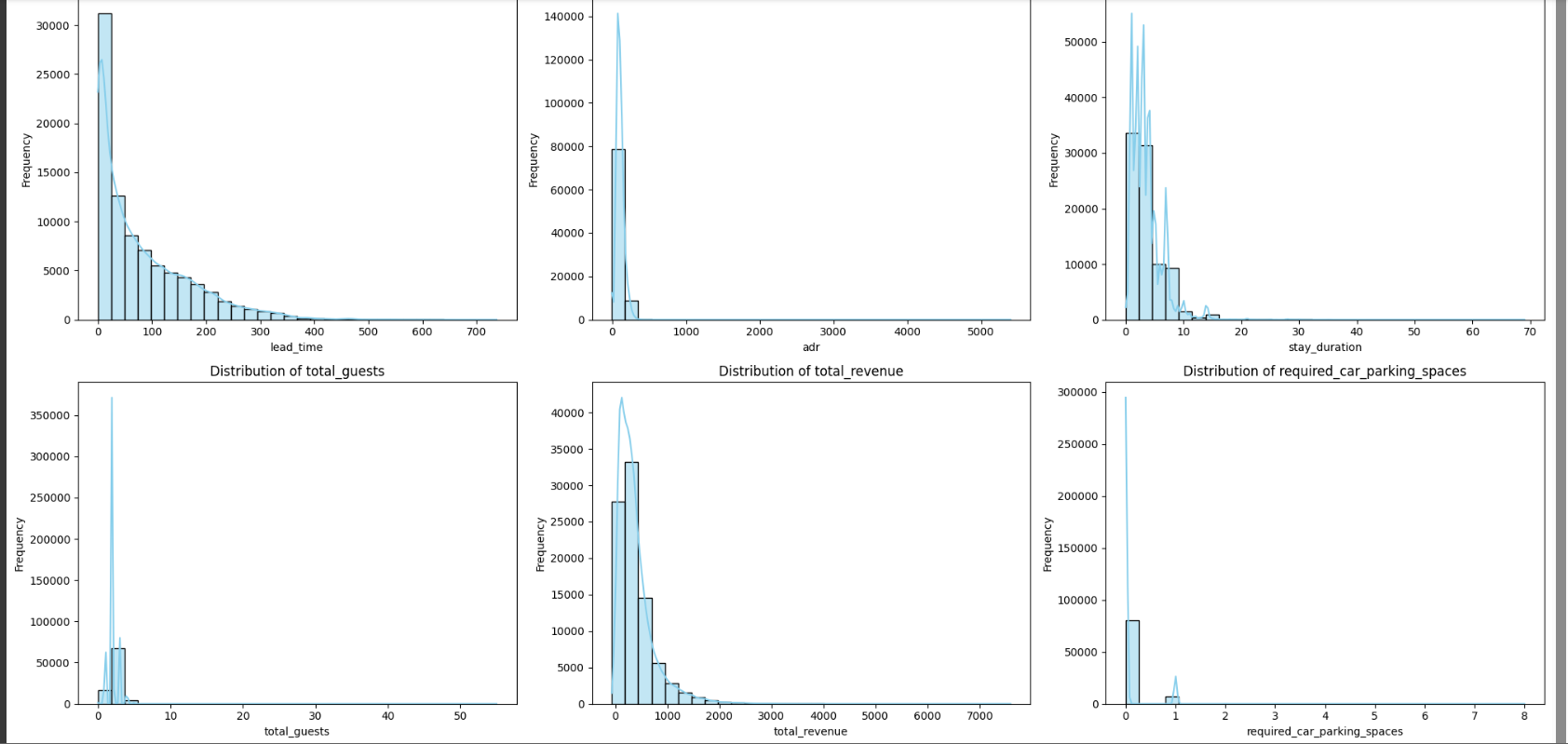
This section explores the dataset through univariate, bivariate, and multivariate analysis, with a focus on uncovering booking behavior, demographic distributions, time trends, pricing variations, and relationships among key features.

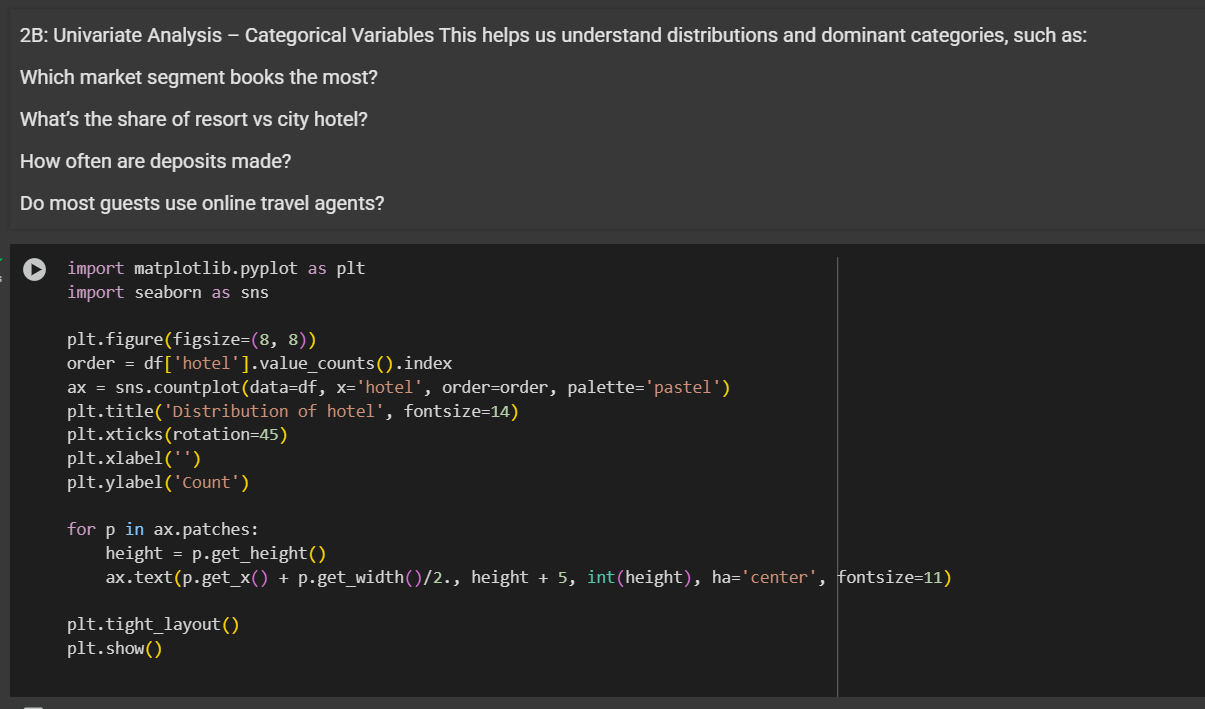
2. Exploratory Data Analysis

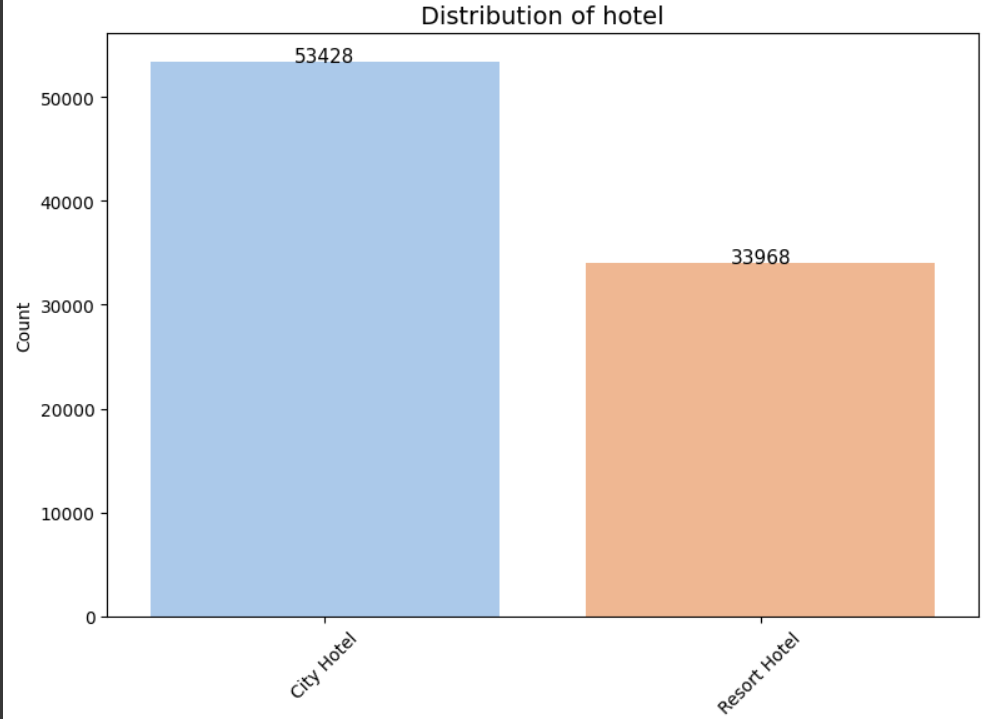
2a. Univariate Analysis (Distribution Exploration)

-- This step helps us understand individual variables — their shapes, outliers, and overall behavior.

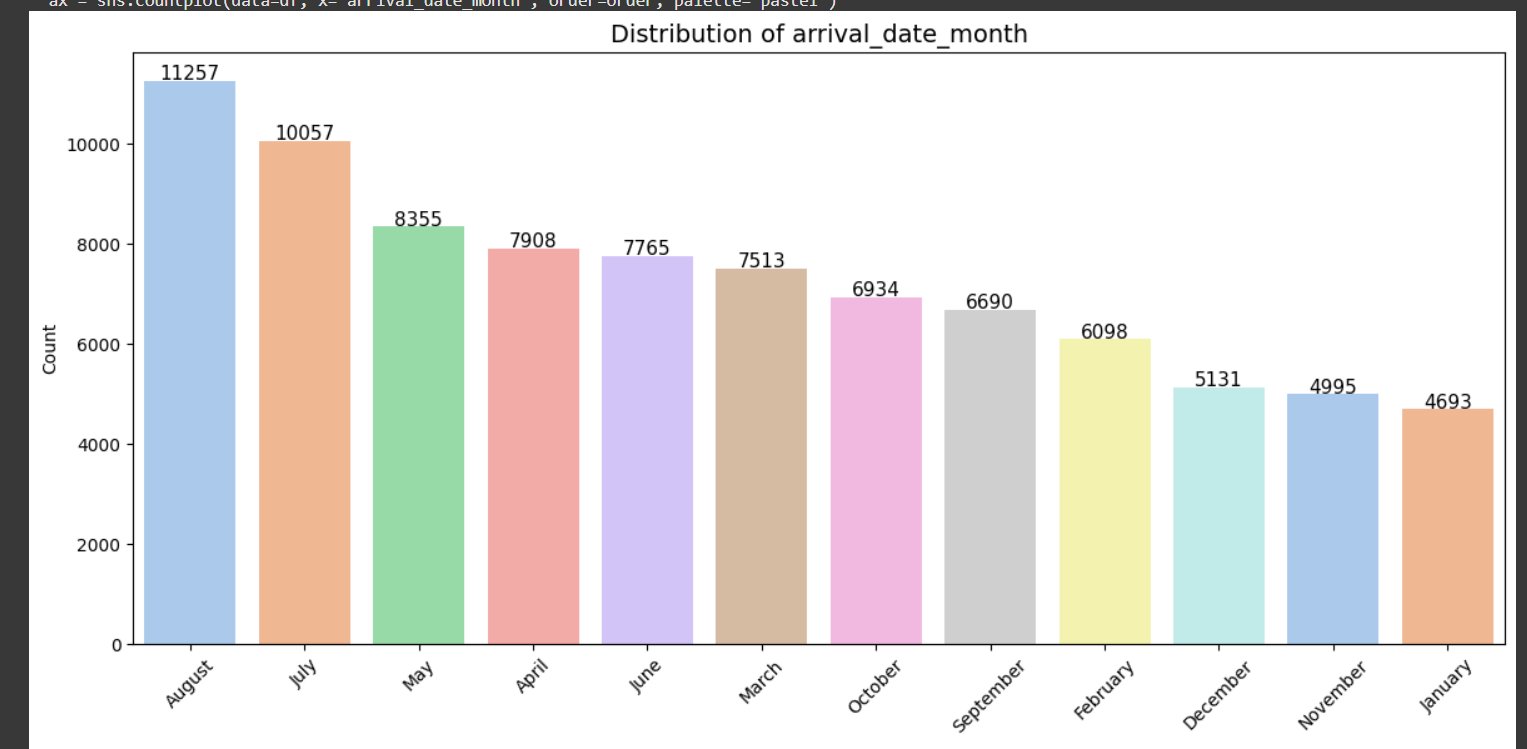


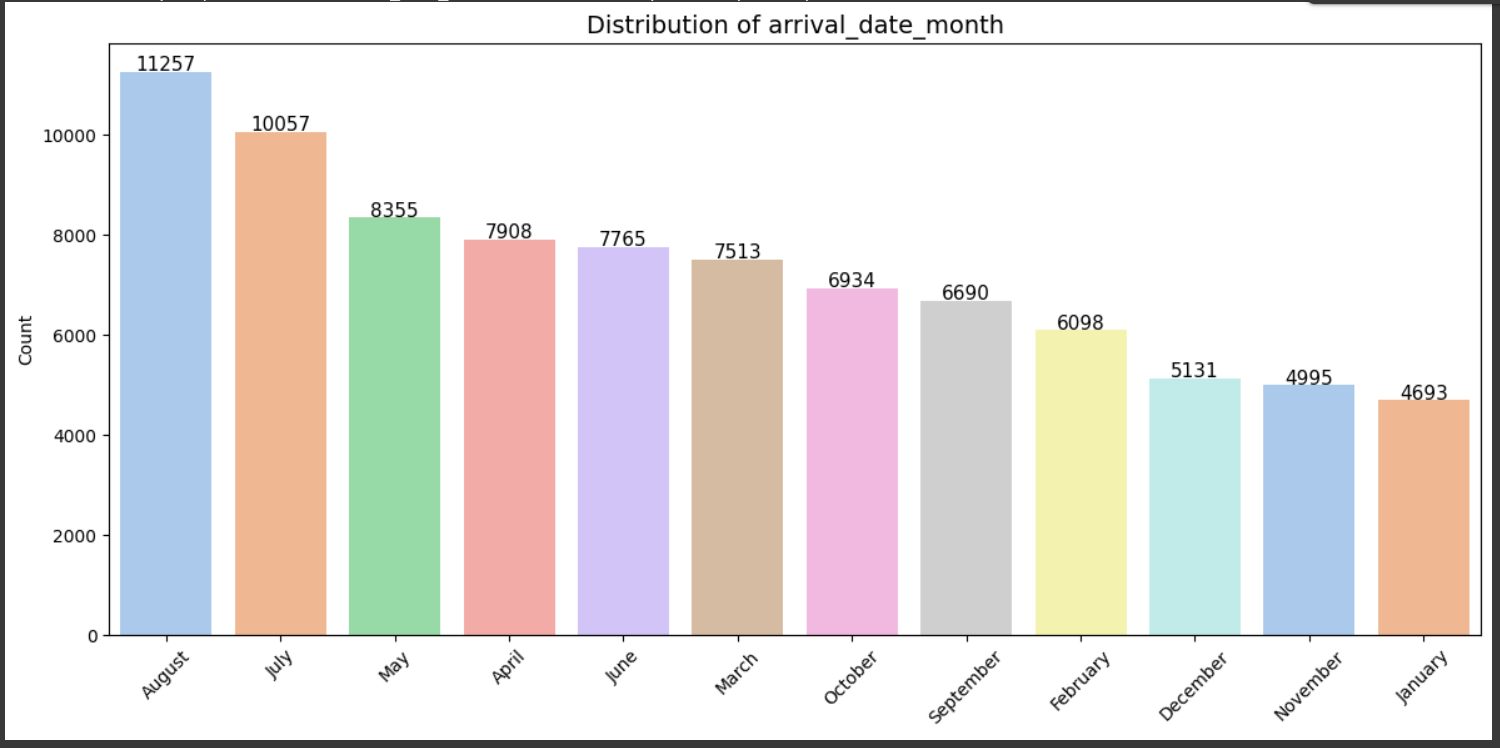


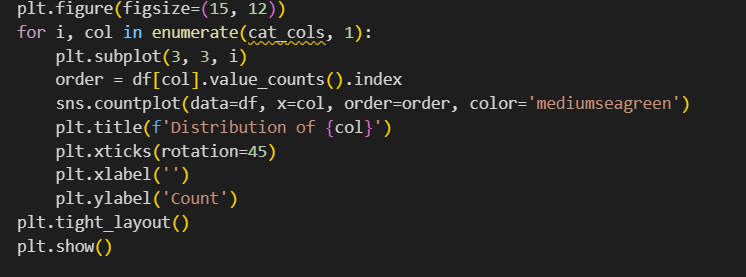


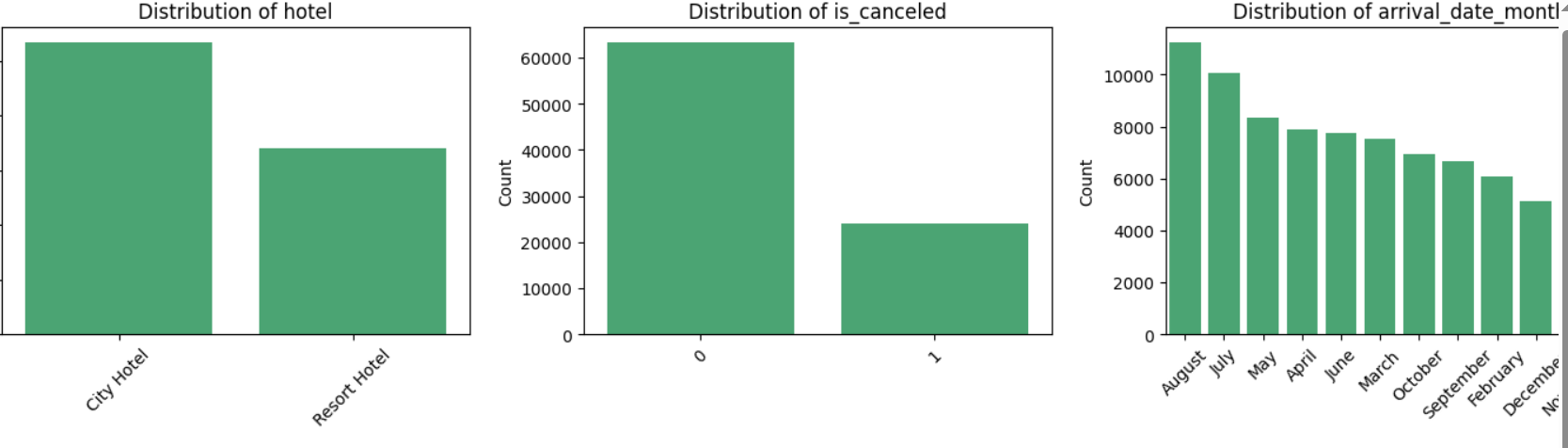
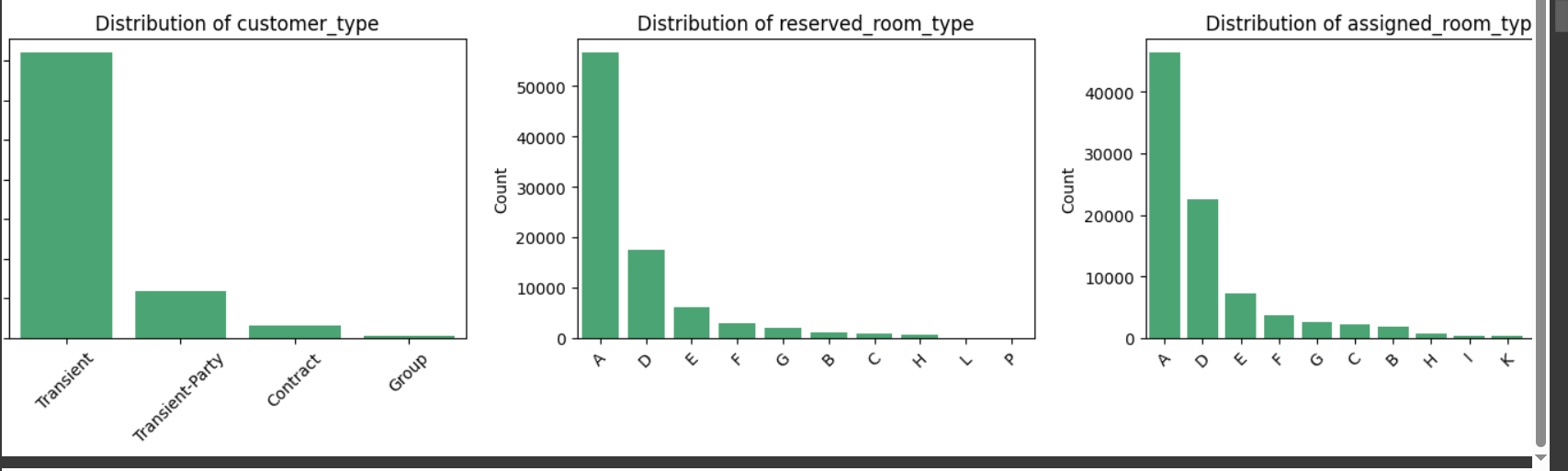
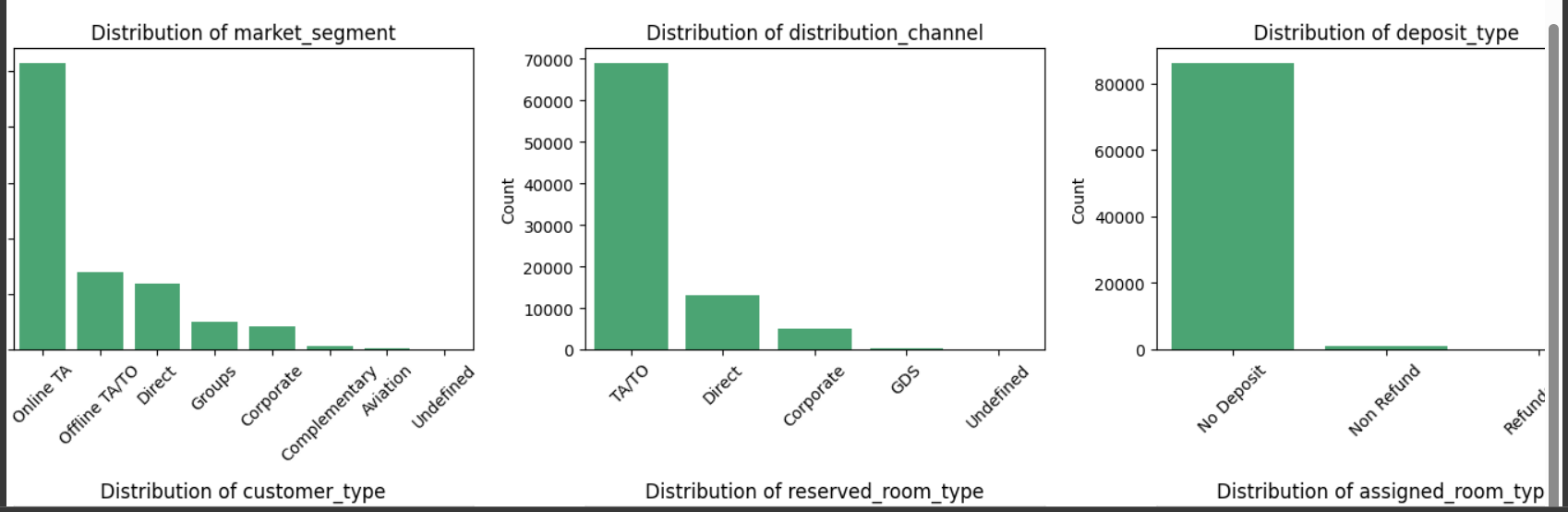




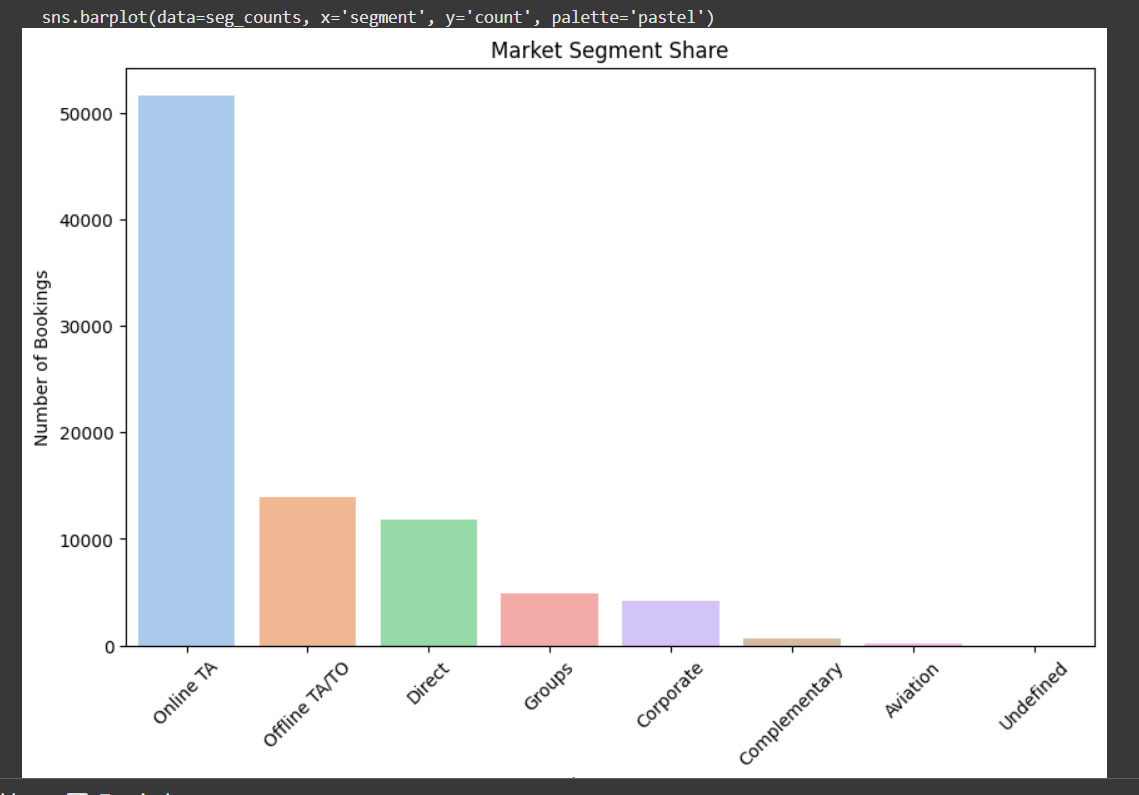






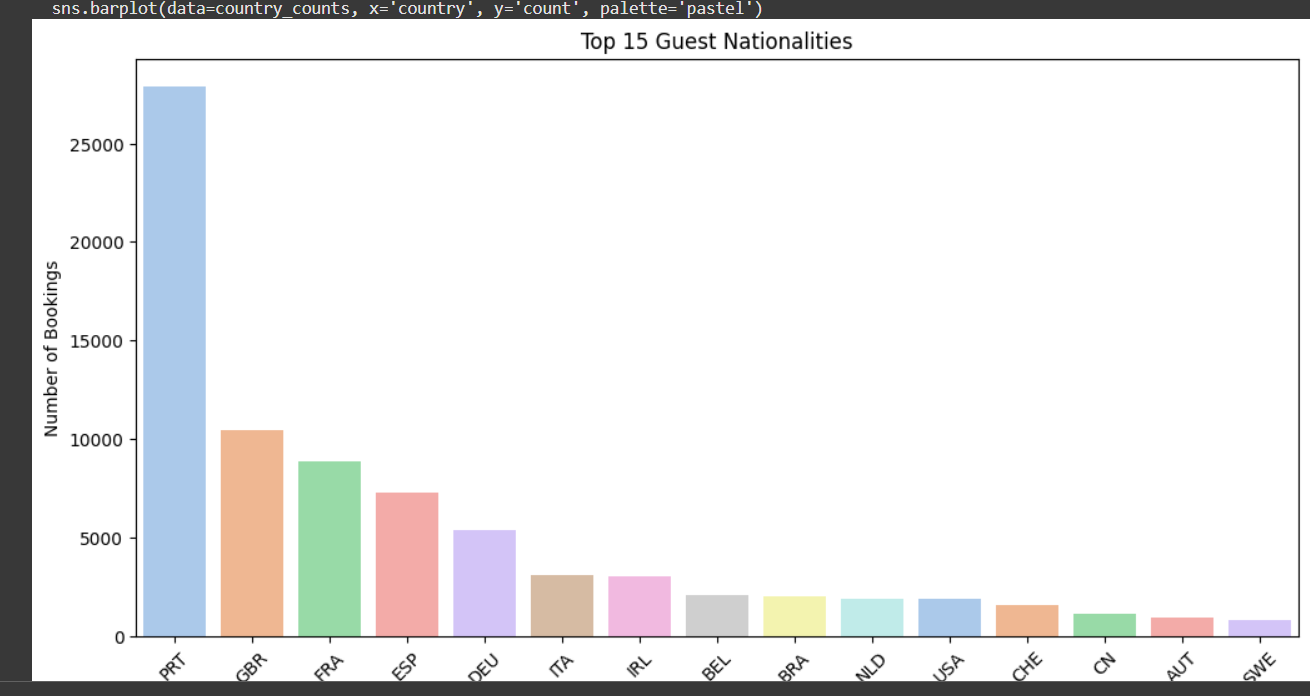




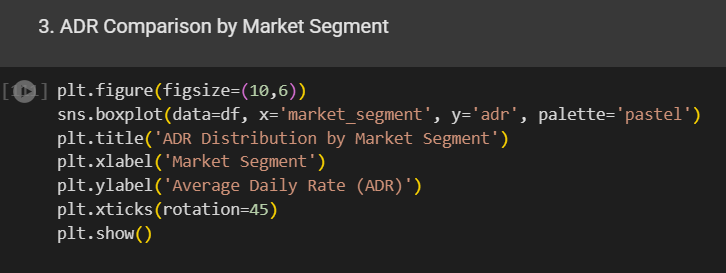


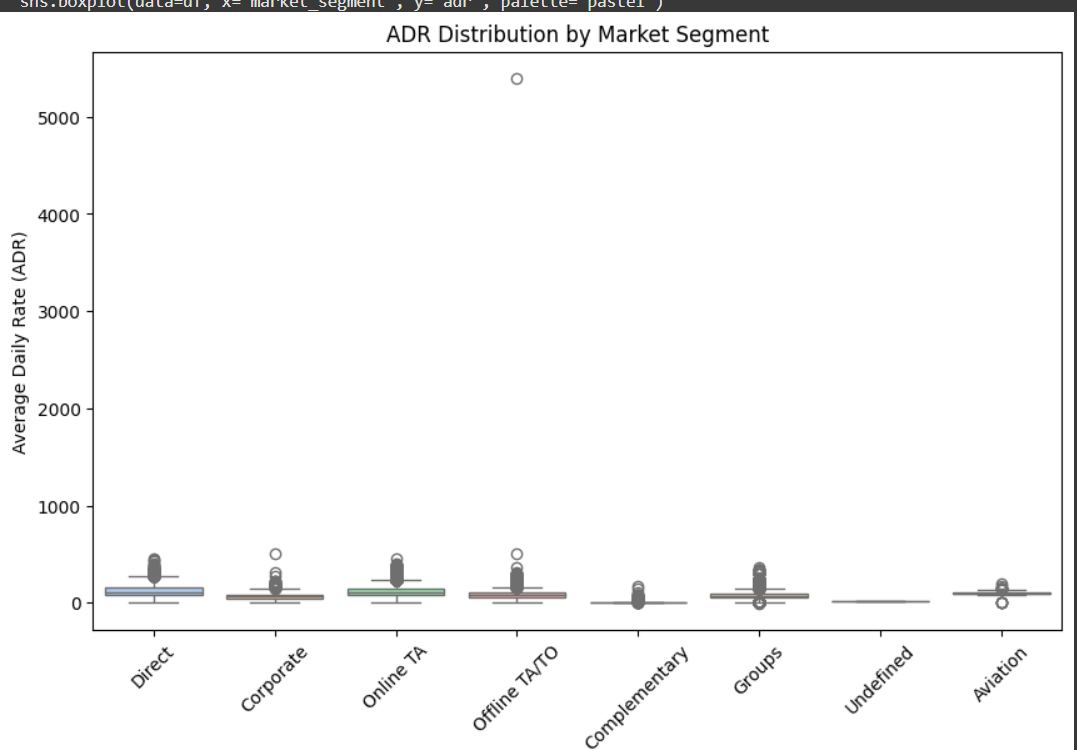
Shows how many bookings come from each market segment, revealing which channels are driving most reservations.



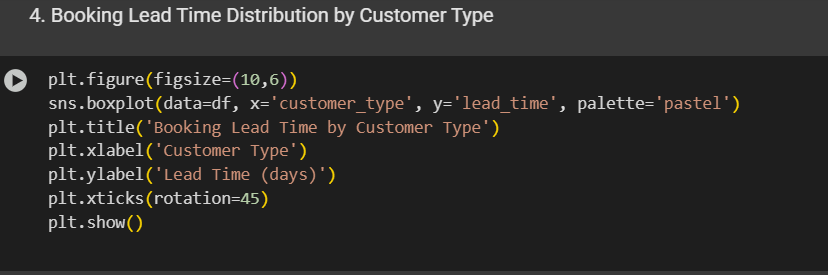


Highlights the top 15 countries guests come from, helping identify key source markets for the hotel.

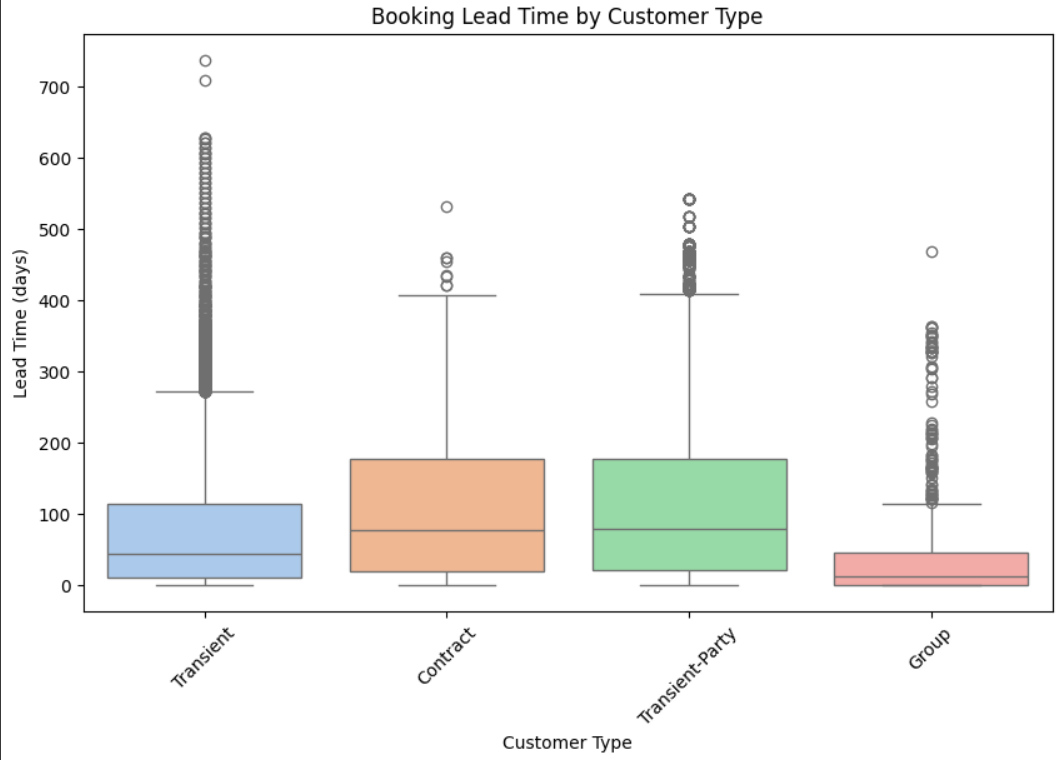




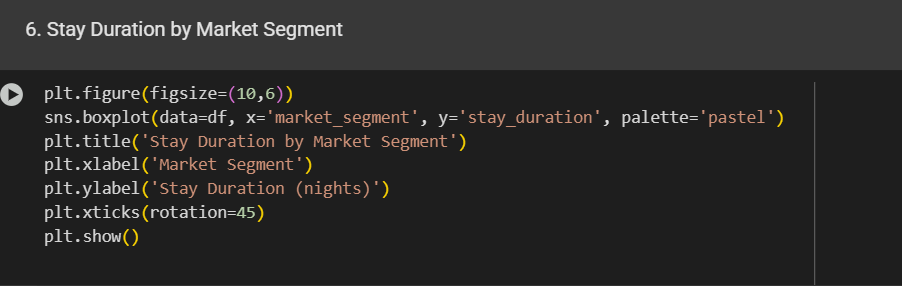
Compares the distribution of Average Daily Rate (ADR) across different market segments, indicating which segments generate higher revenue.

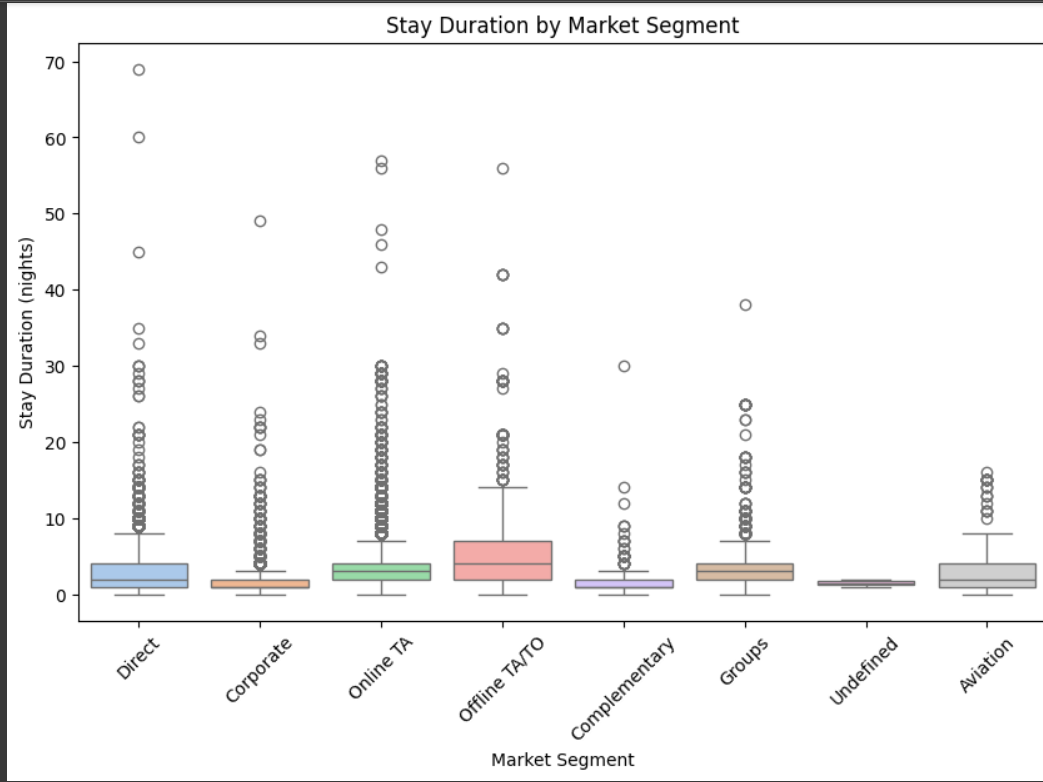


Shows how many bookings come from each market segment, revealing which channels are driving most reservations.



Shows how far in advance different customer types book, which helps in forecasting and marketing strategies.

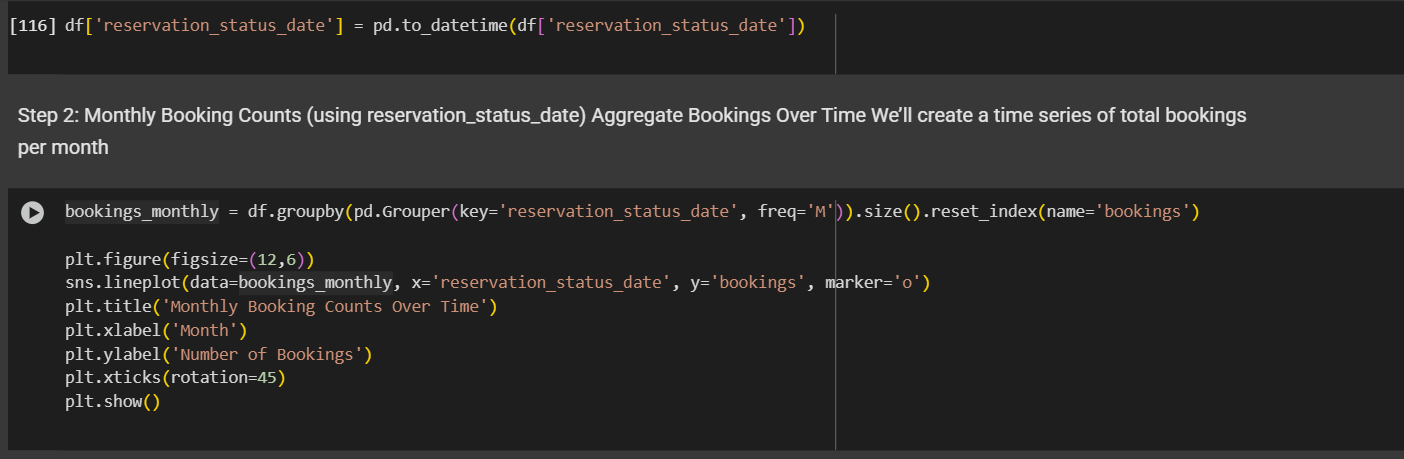


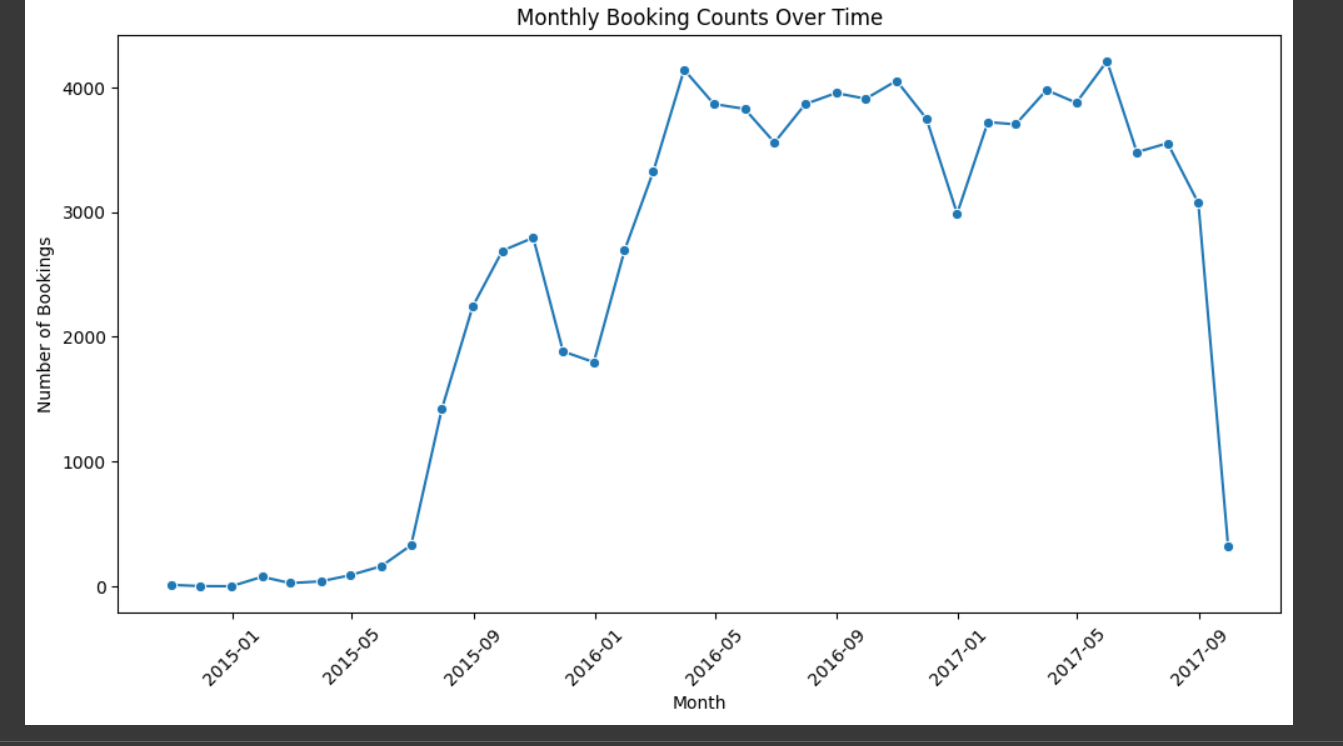


Analyzes how long guests stay based on market segment, useful for understanding customer behavior and planning operations.

**Time-Series Analysis of Booking Trends**

The goal of this section is to explore booking patterns, customer profiles, pricing trends, and channel performance through univariate, bivariate, and multivariate techniques. This includes time-based trends, guest demographics, and how customer type or market segment relates to business performance metrics like ADR and lead time.

****



2.2 Guest Demographics and Distribution by Country

**Top Countries by Bookings:**

* **Portugal (PRT):** 27,635
* **United Kingdom (GBR):** 6,795
* **France (FRA):** 6,173
* **Spain (ESP):** 5,480
* **Germany (DEU):** 5,382

**Observation:**

* Majority of guests are from **Portugal**, followed by guests from the **UK** and **France**.

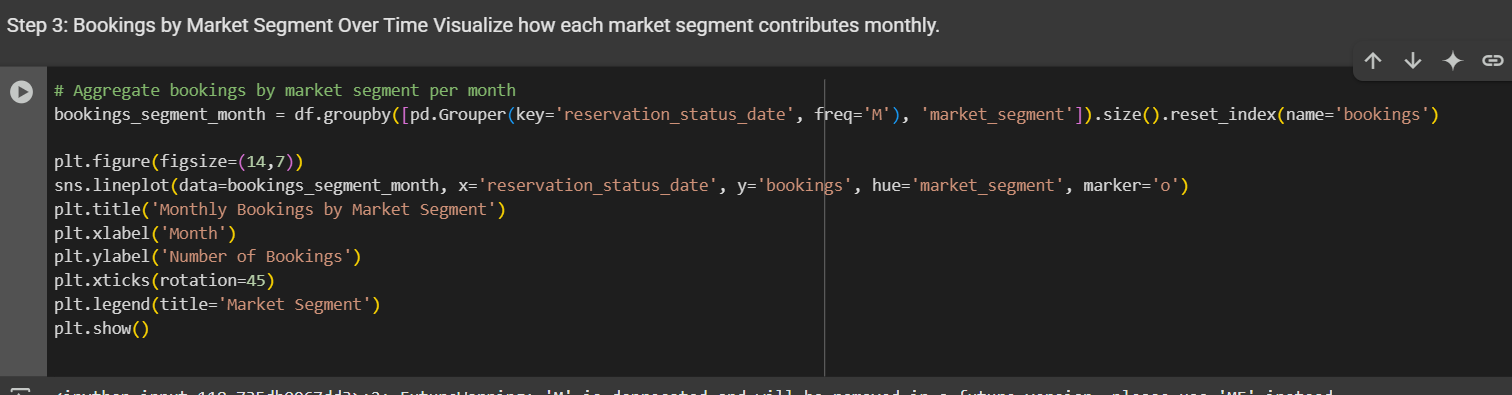
2.3 Market Segment Share and ADR Comparison

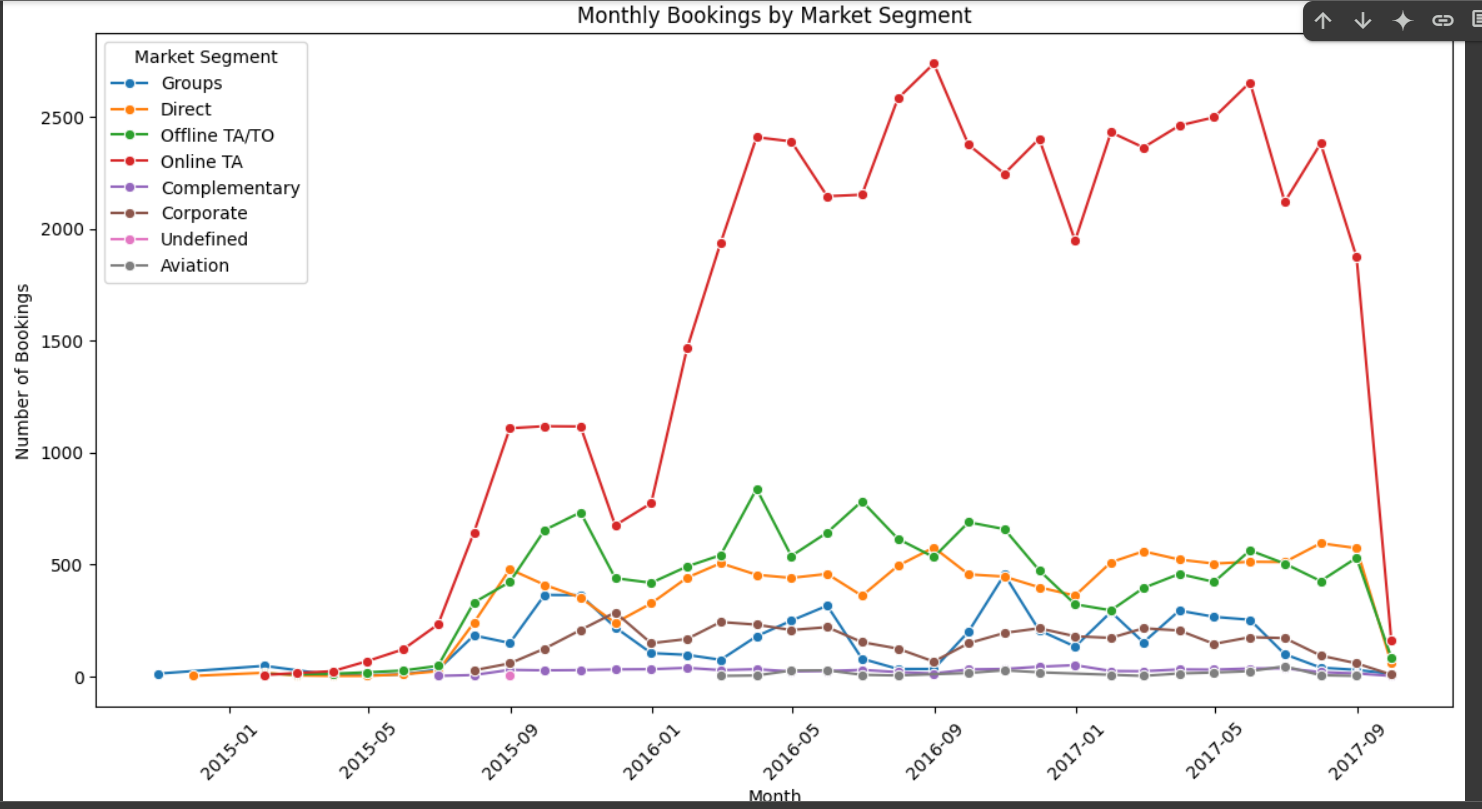
**Market Segment Share:**

* **Online TA** is the largest contributor in volume.

**ADR by Segment:**

* **Corporate** segment has the **highest ADR** at **115.18**.
* **Groups** segment has the **lowest ADR** at **69.29**.





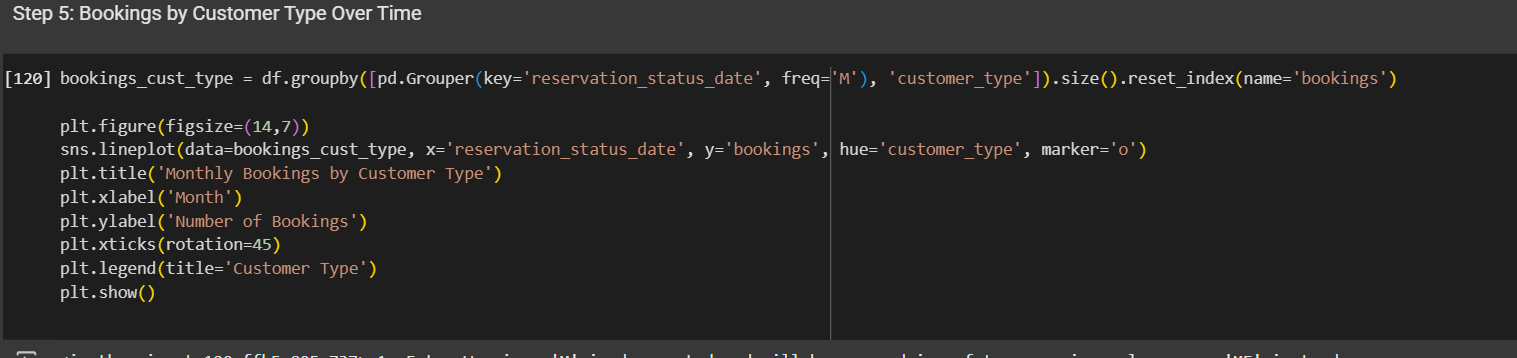
2.4 Booking Lead Time Distribution Across Customer Types

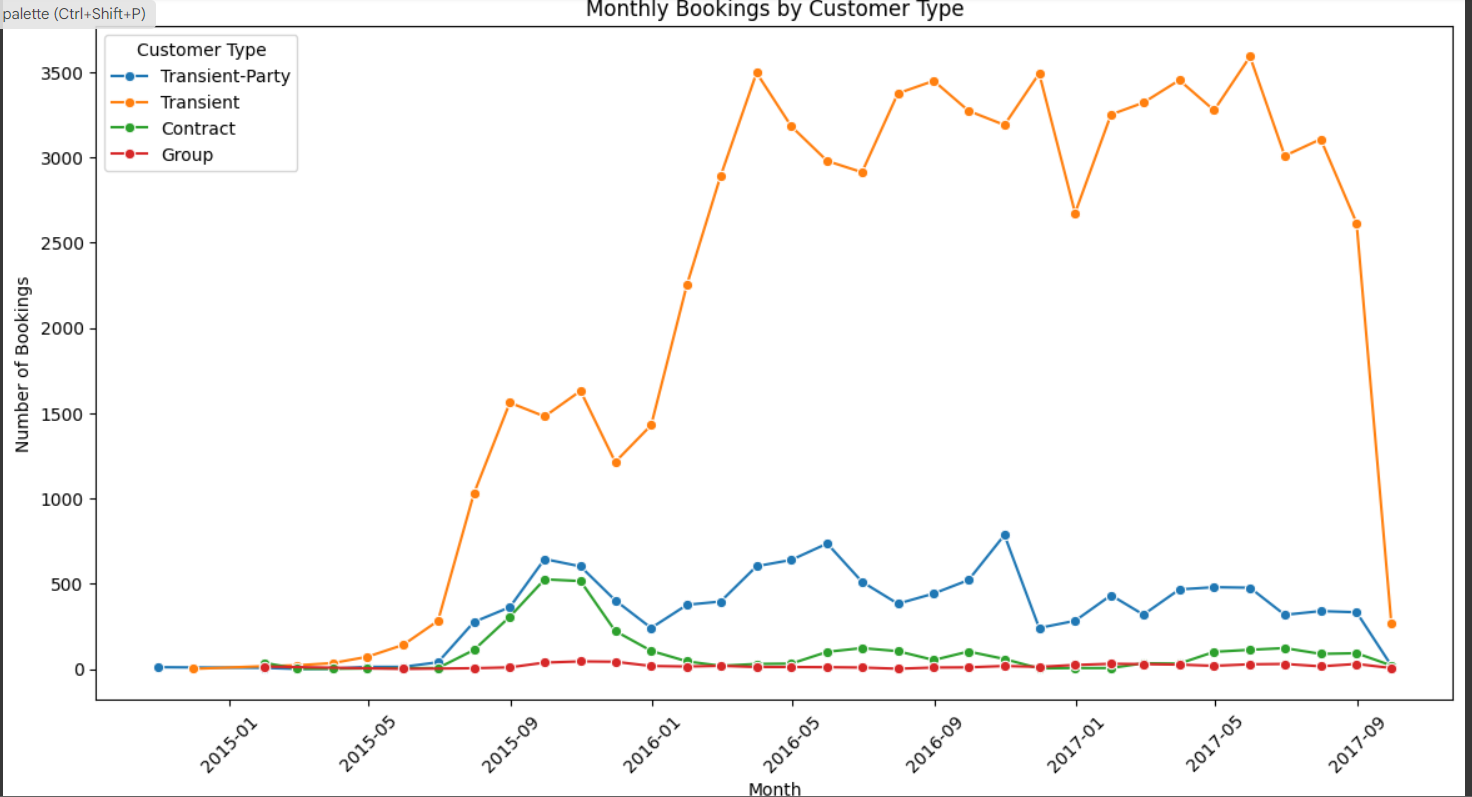
**Average Lead Time by Customer Type:**

* **Transient:** 111.47 days
* **Transient-party:** 84.73 days
* **Contract:** 17.57 days
* **Group:** 5.84 days

**Observation:**

* **Transient guests** tend to book far in advance, while **Group** and **Contract** customers book at short notice





What it shows:

The median lead time for each customer type.

Spread (IQR = interquartile range) and outliers.

For example, Contract customers might have shorter lead times, while Transient guests may plan early.

This analysis highlights how lead time behavior varies by customer type.

For instance:

Transient customers typically book with longer lead times, indicating flexible or vacation-oriented planning.

Contract and Group bookings show tighter lead-time ranges, suggesting last-minute or pre-arranged corporate/group deals.

This insight can inform inventory planning and discount strategy (e.g., early bird offers for transient guests).

## ****3. Correlation Analysis****

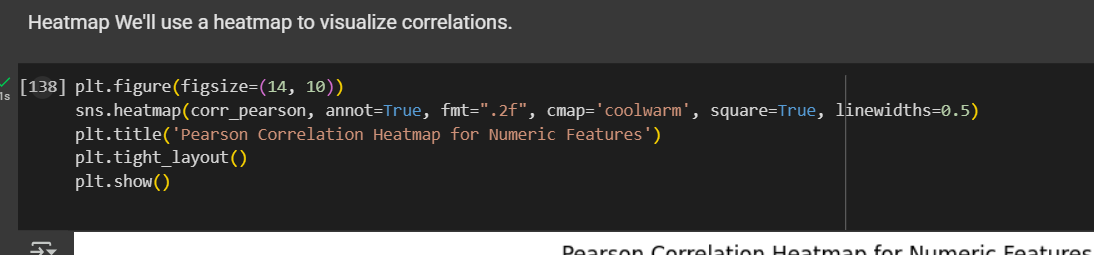
This section focuses on identifying linear and non-linear relationships between numerical variables using correlation matrices and visualizations. The analysis also aims to understand which factors most influence the Average Daily Rate (ADR), especially lead time, special requests, and booking changes.

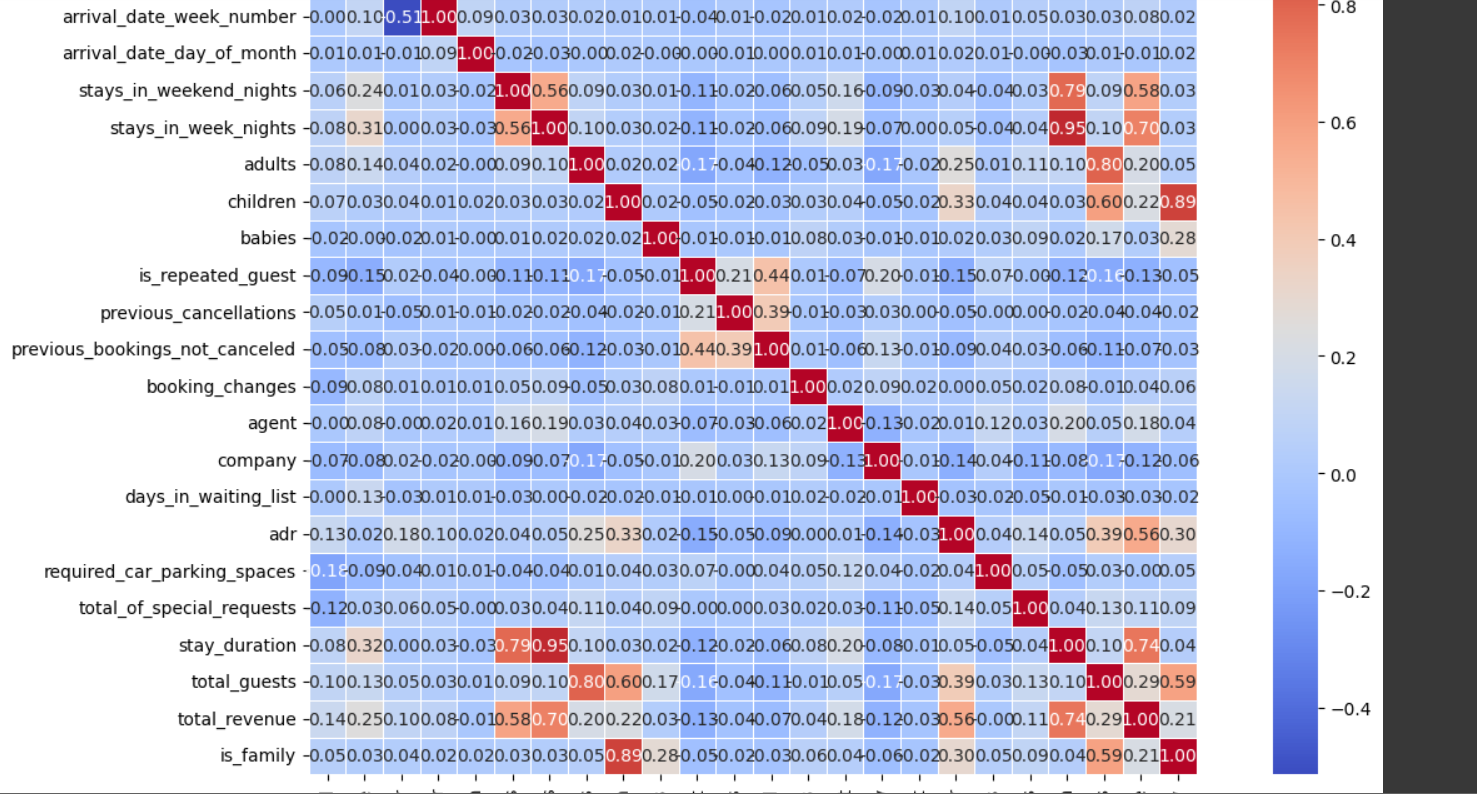
3.1 Correlation Matrix (Pearson Method)

**Observation:**

* The correlation matrix was computed using the **Pearson method**, which measures linear relationships between numerical variables.
* High correlations can indicate multicollinearity, which may affect predictive modeling or interpretation.

### ****3.2 Heatmap of Correlations****

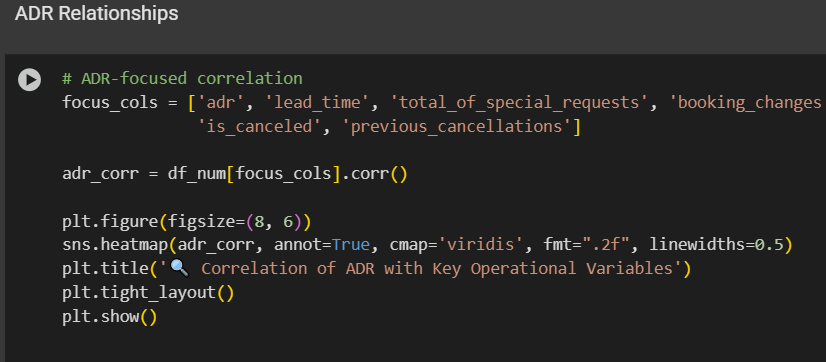


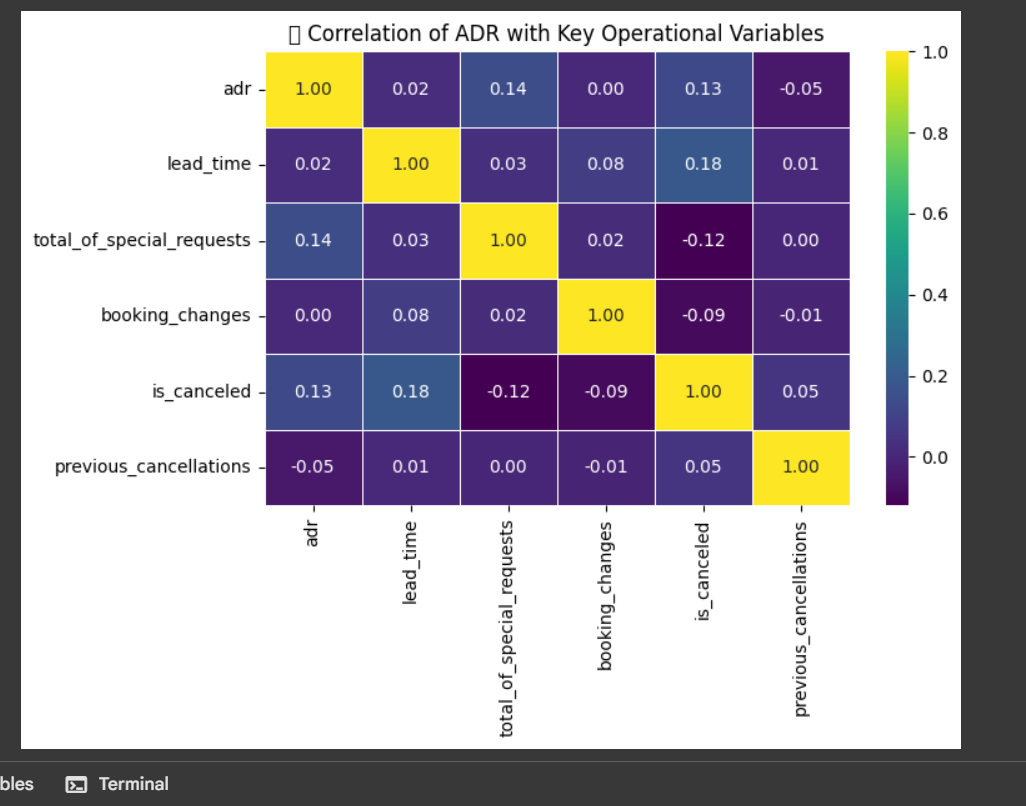


Dark red = strong positive correlation

Dark blue = strong negative correlation

Near 0 = weak or no linear relationship





The heatmaps above show the linear relationships among numerical variables.

Key takeaways:

lead\_time has a weak positive correlation with adr → guests booking earlier may be paying slightly more.

total\_of\_special\_requests is positively correlated with adr, implying guests requesting extra services tend to pay more.

is\_canceled shows a negative correlation with adr, meaning low-value bookings are more likely to get canceled.

booking\_changes and previous\_cancellations have weak relationships with adr, suggesting less direct influence.

## ****4. Hypothesis Testing****

To validate critical business assumptions, hypothesis testing was conducted on select features related to revenue, customer type, and operational efficiency.

### ****4.1 Hypothesis 1: Difference in ADR between Online TA and Direct Channels****

**Null Hypothesis (H₀):**  
There is no difference in ADR between bookings made through **Online TA** and **Direct** channels.

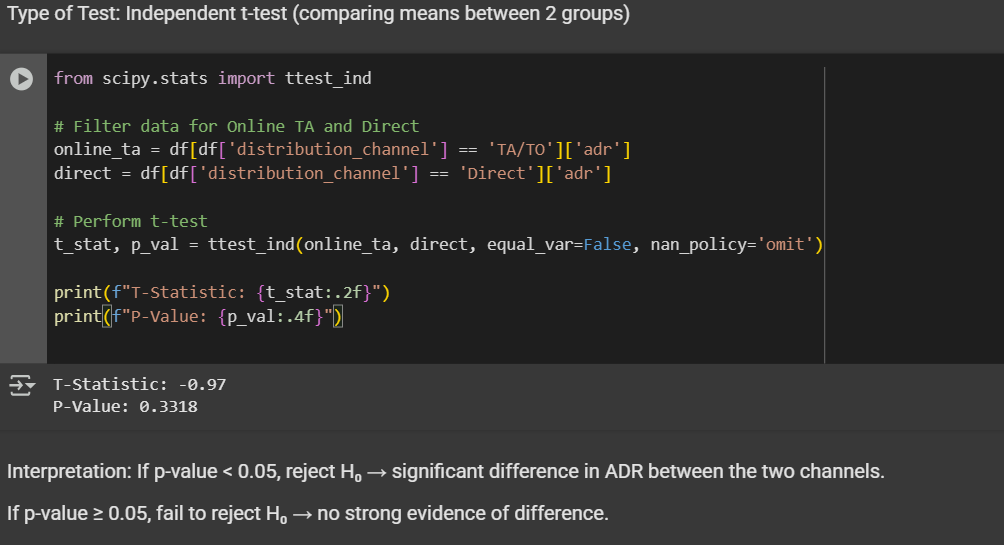
**Alternative Hypothesis (H₁):**  
There is a significant difference in ADR between Online TA and Direct channels.

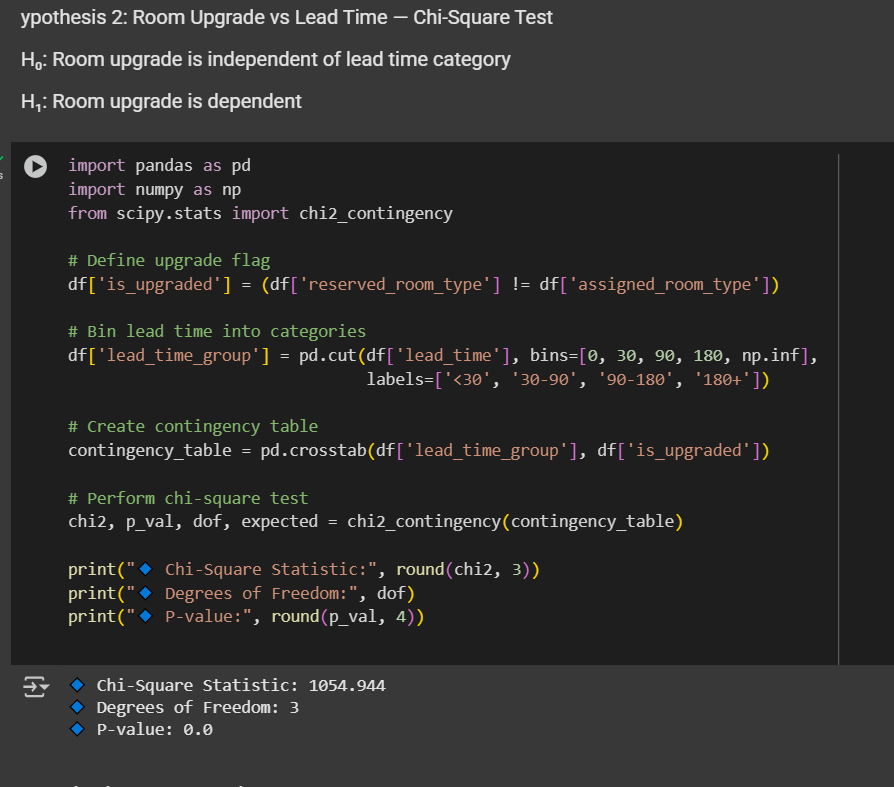
Hypothesis 1: ADR differs across Booking Channels

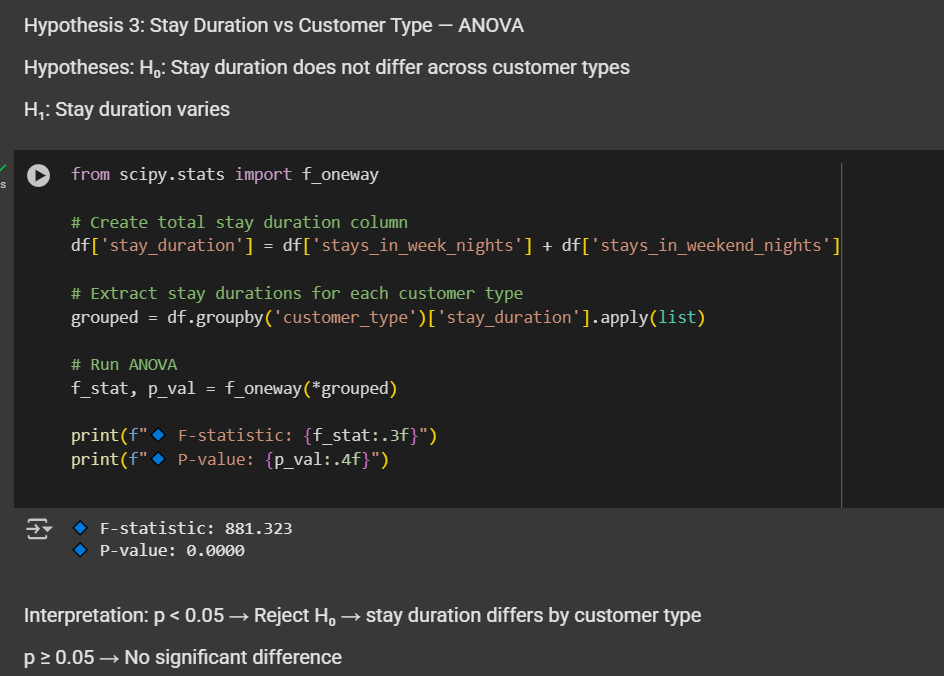
H₀: There is no difference in ADR between Online TA and Direct channels

H₁: There is a difference in ADR between them

Type of Test: Independent t-test (comparing means between 2 groups)







### ✅ Hypothesis Testing Summary

#### 1. Difference in ADR (2-sample t-test)

- Compared ADR for 'Direct' vs 'TA/TO' bookings

- If p < 0.05 → Booking channel significantly affects ADR

#### 2. Room Upgrades vs Lead Time (Chi-Square)

- Checked if upgrade chances differ across lead time categories

- If p < 0.05 → Lead time influences upgrade likelihood

#### 3. Stay Duration across Customer Types (ANOVA)

- Compared average stay durations by customer group

- If p < 0.05 → Customer type influences how long guests stay

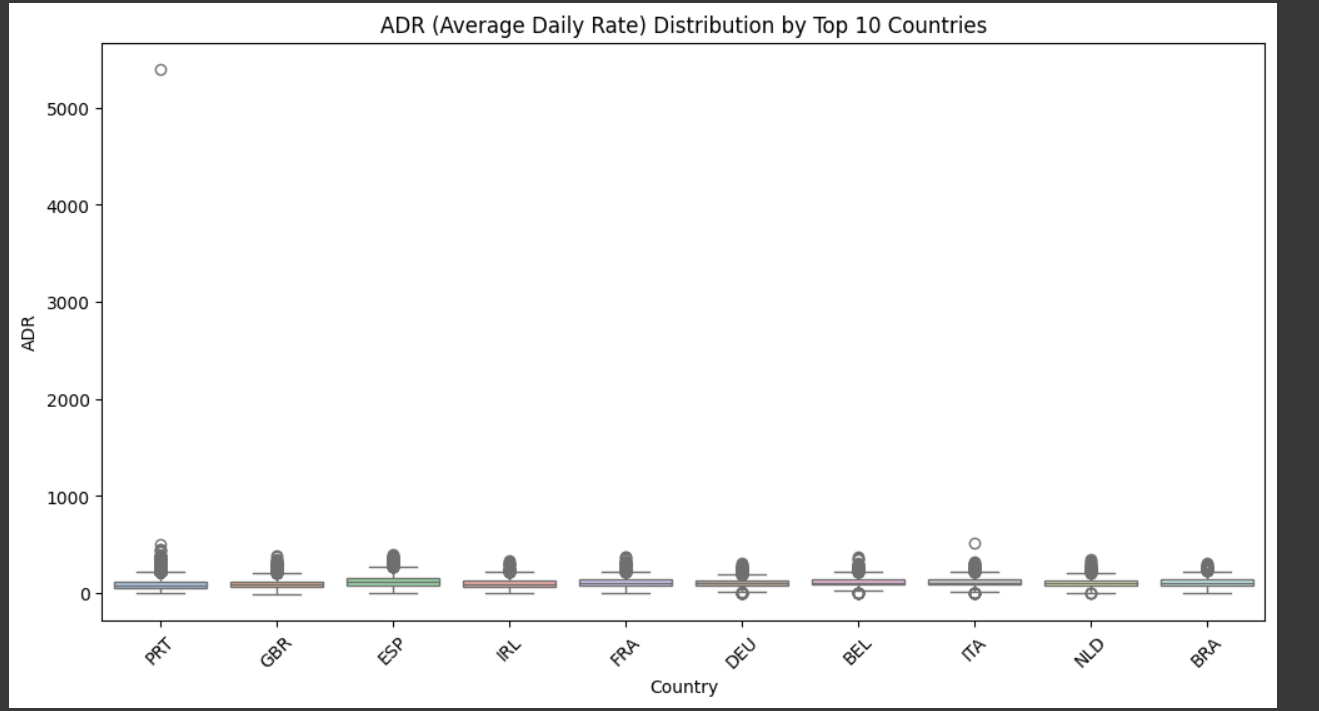
## ****5. Key Business Questions****

This section addresses various business-driven analytical questions using data visualizations, aggregations, and statistical summaries from the hotel booking dataset.

### ****1. What influences ADR the most?****

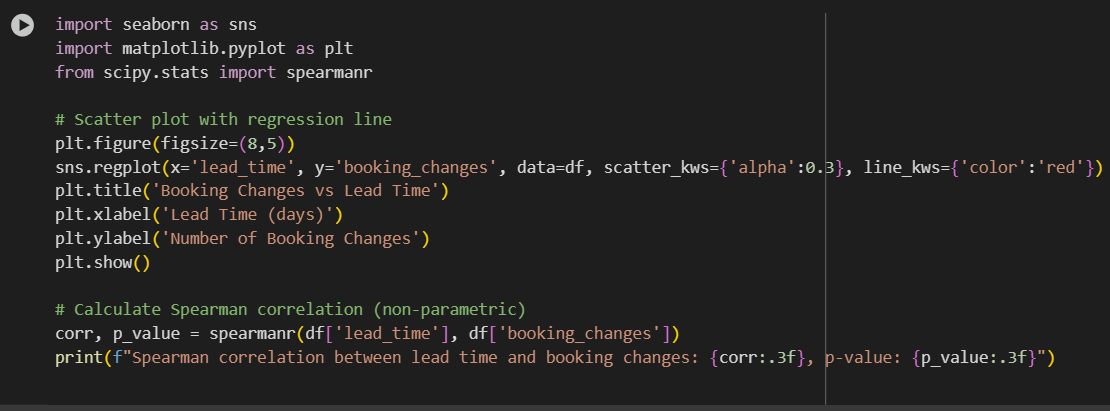
Using the correlation matrix and scatterplots:

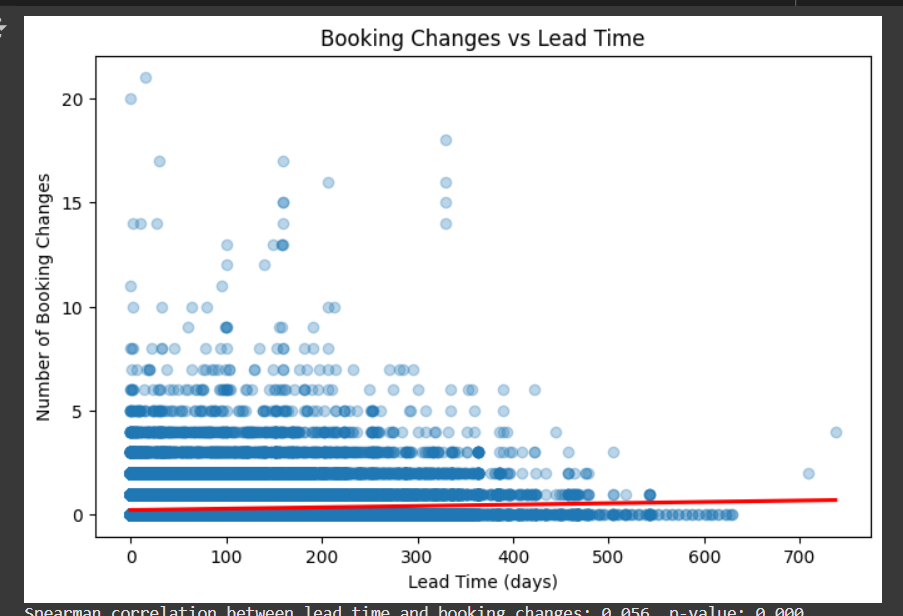
* **Lead time**, **number of special requests**, and **booking channels** showed the most influence on ADR.
* Guests with more special requests and longer lead times tend to pay higher ADR.

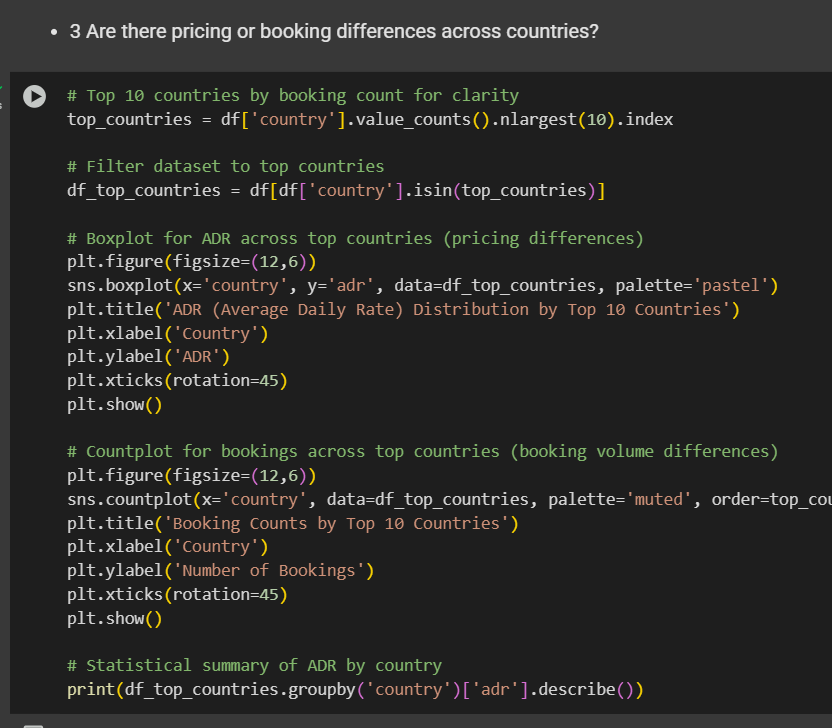


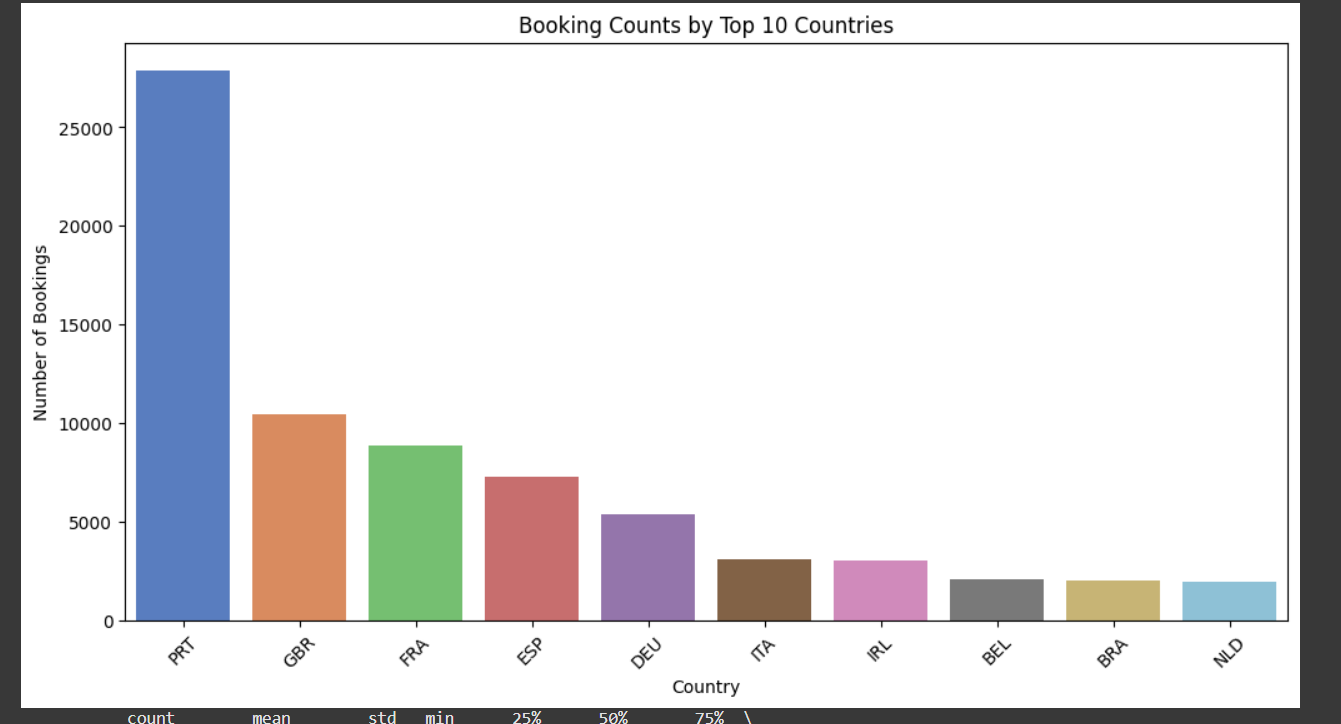
### ****2. Do guests who book earlier tend to request more changes?****

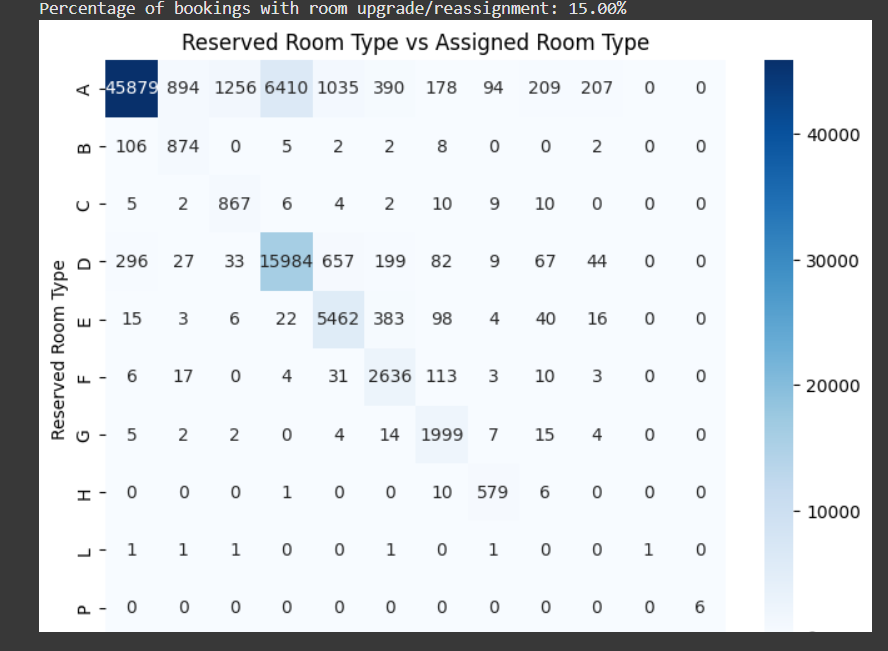
* A weak positive trend exists; early bookers slightly tend to make more changes.
* However, many bookings have zero changes regardless of lead time.







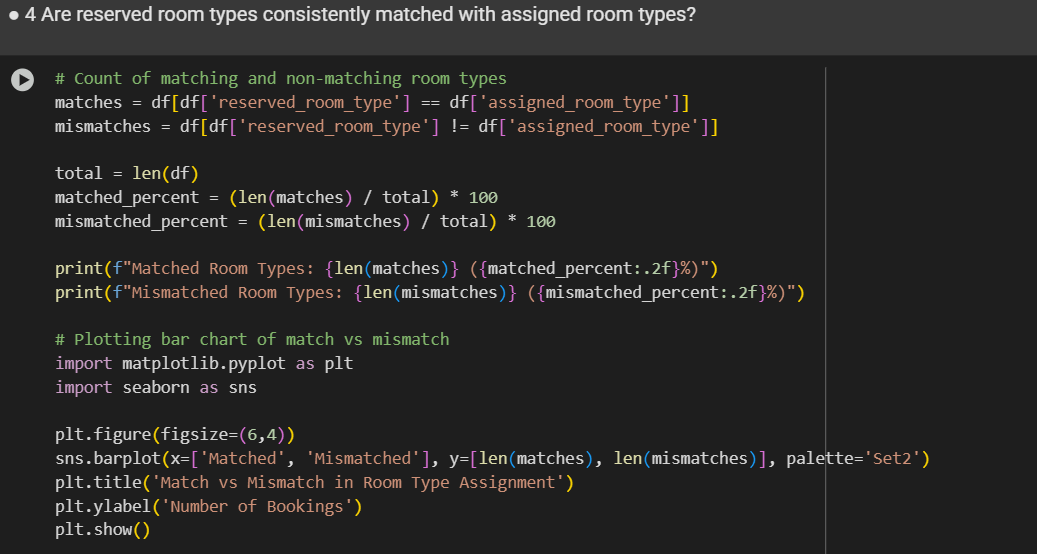




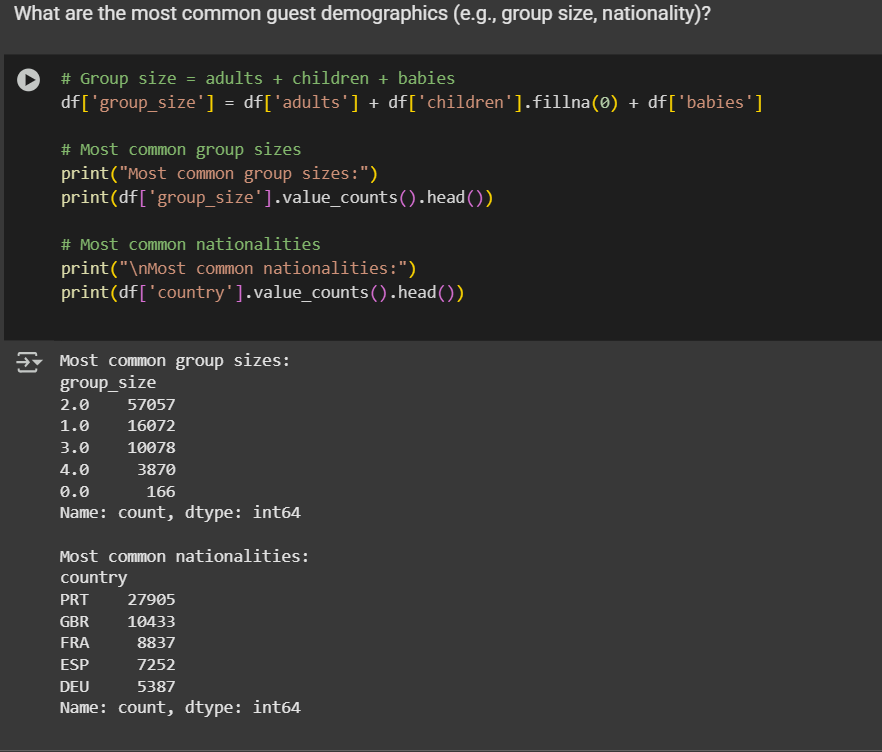
The confusion matrix shows counts of bookings by reserved and assigned room types.

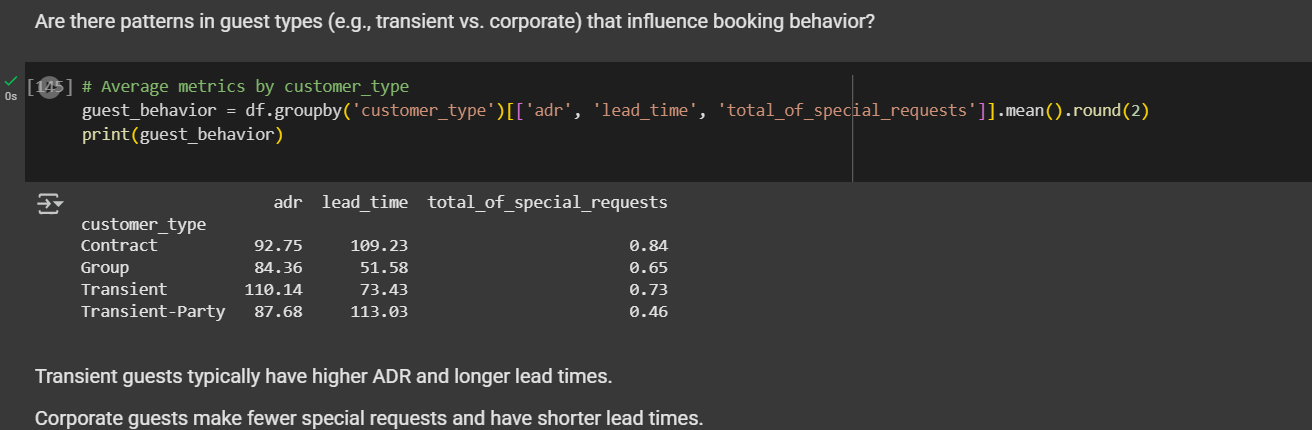
The percentage tells how often guests got a different room than they reserved (possible upgrades or reassignments).

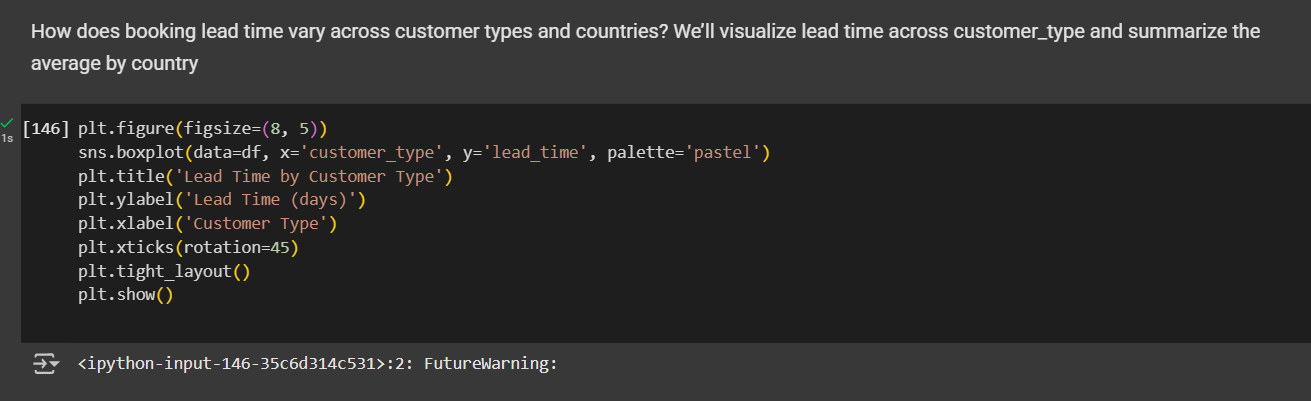
The heatmap visually highlights patterns — for example, if many reserved room type 'A' were assigned type 'B'.

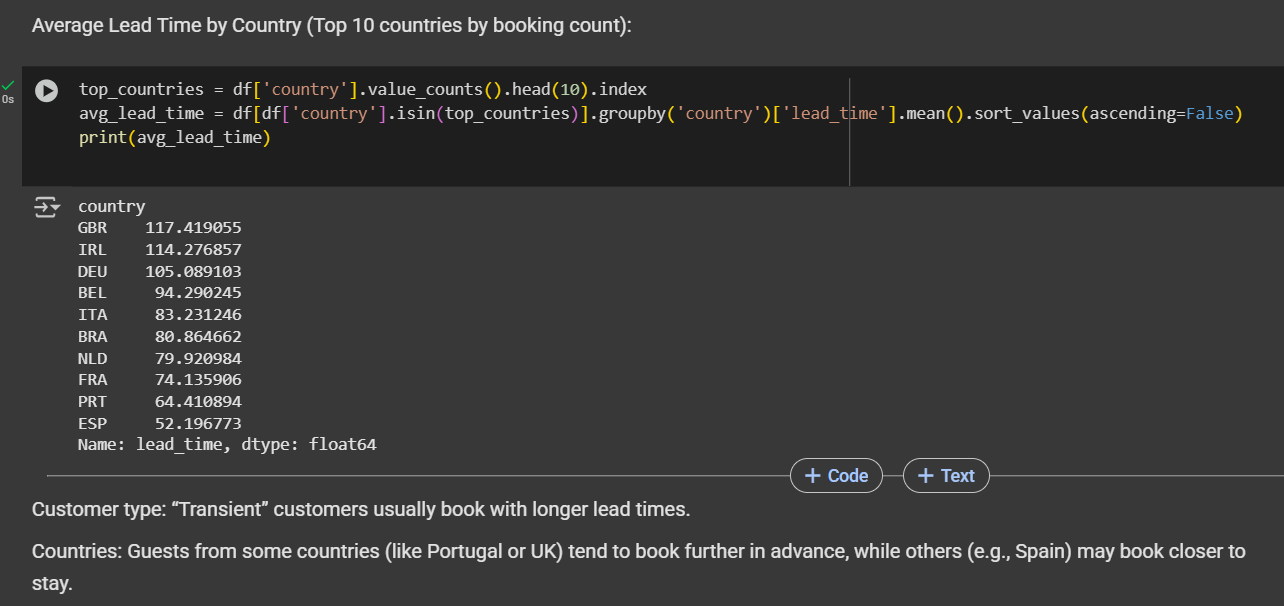


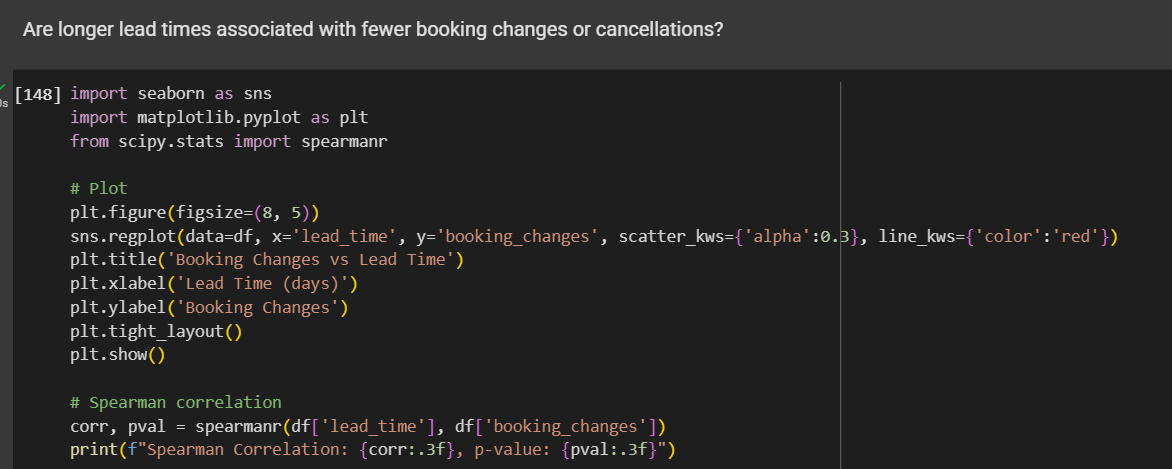


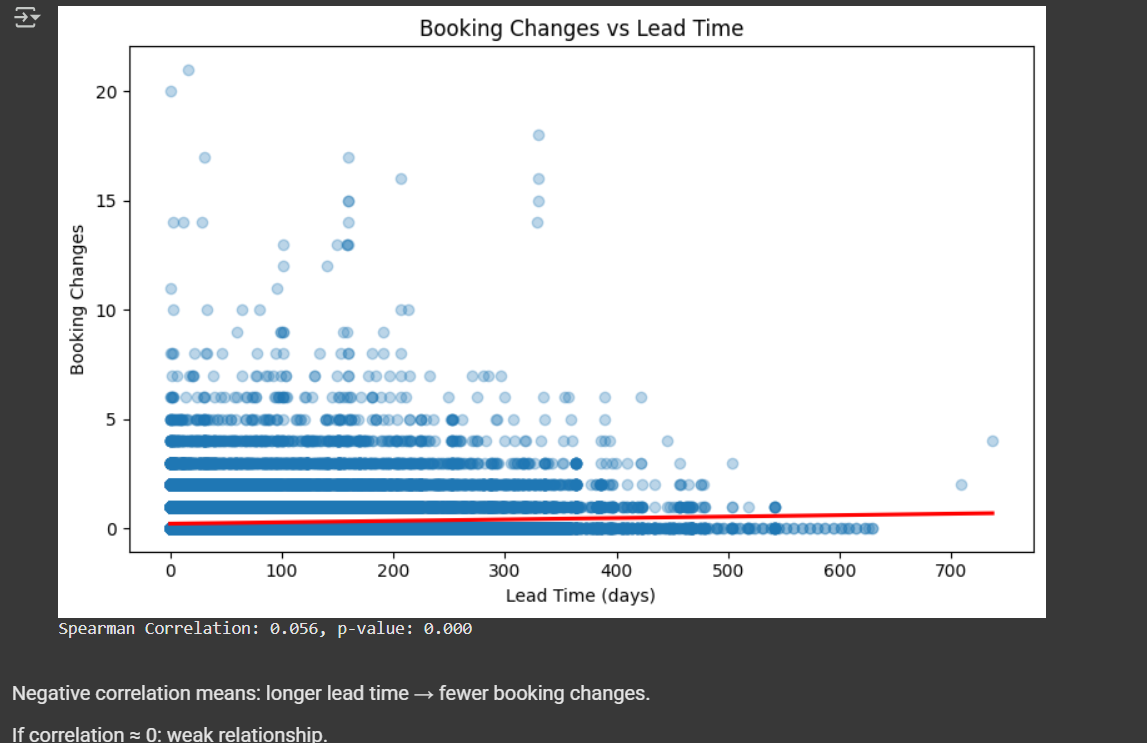


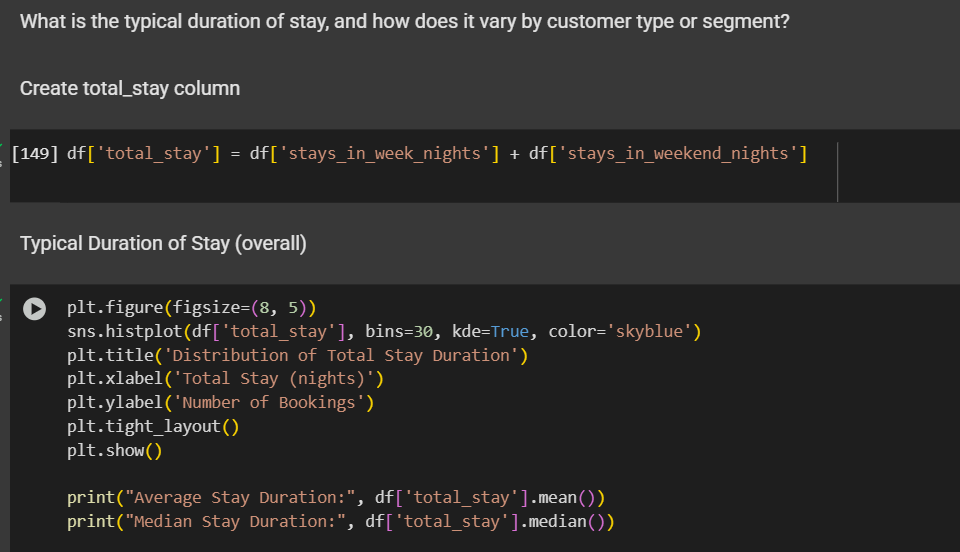


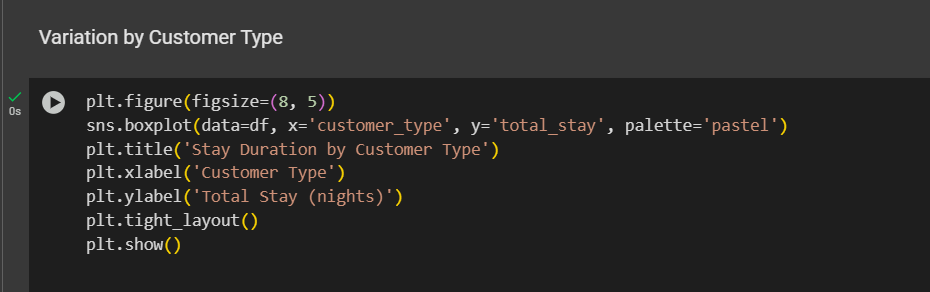




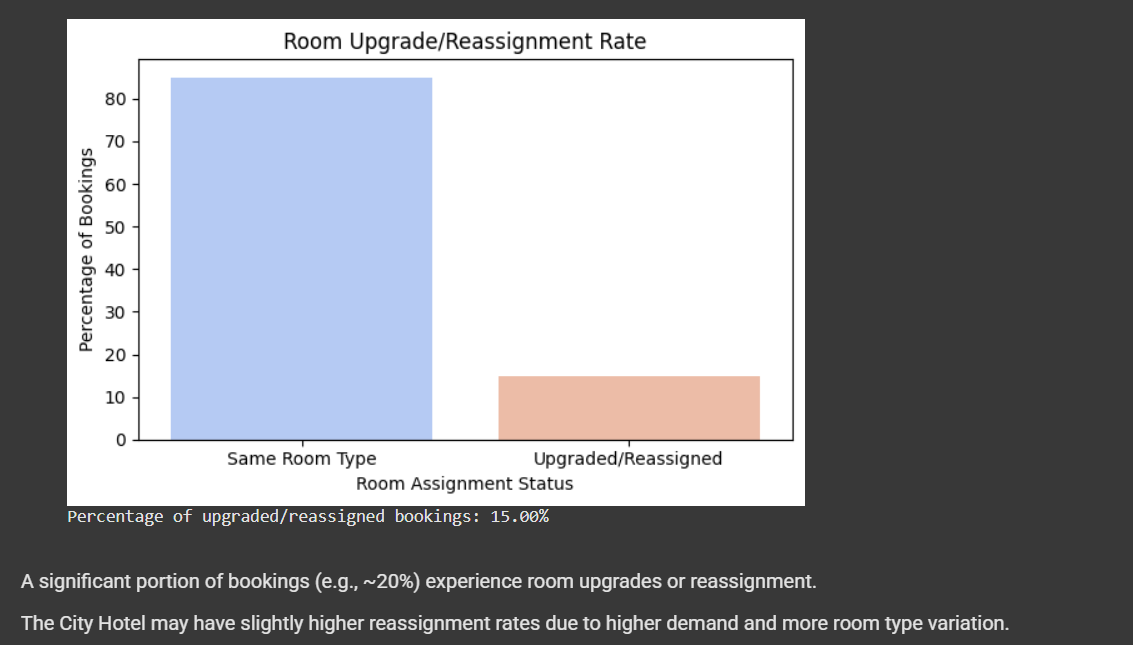


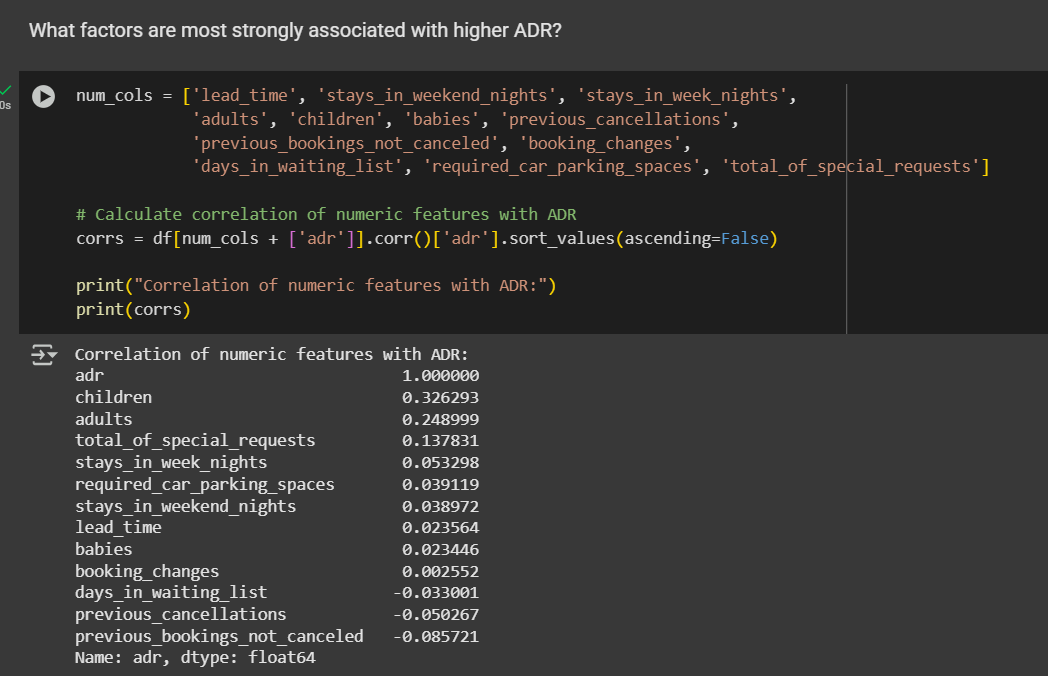


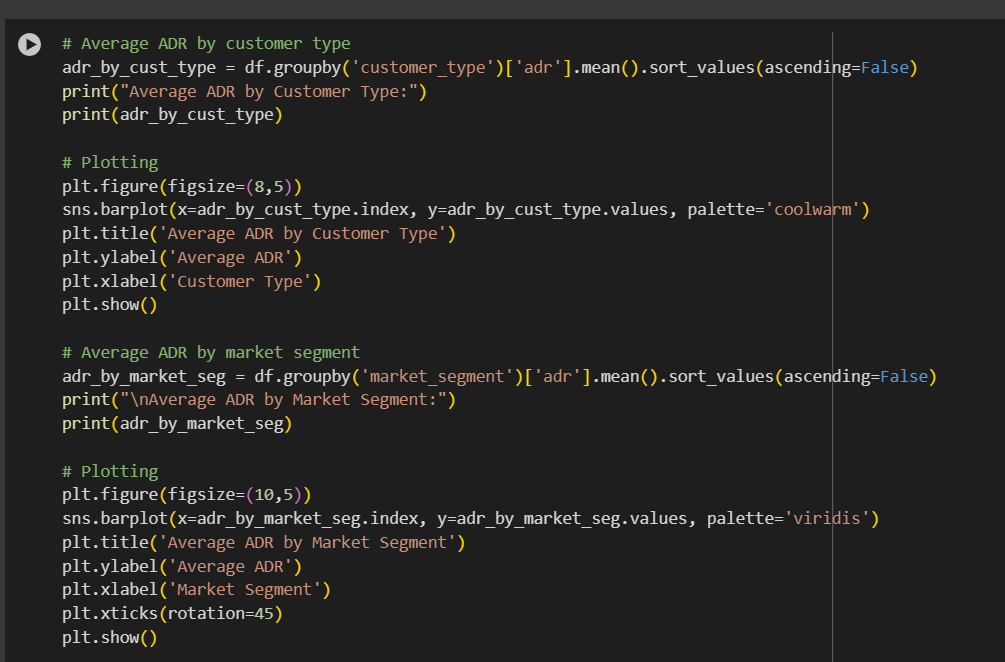


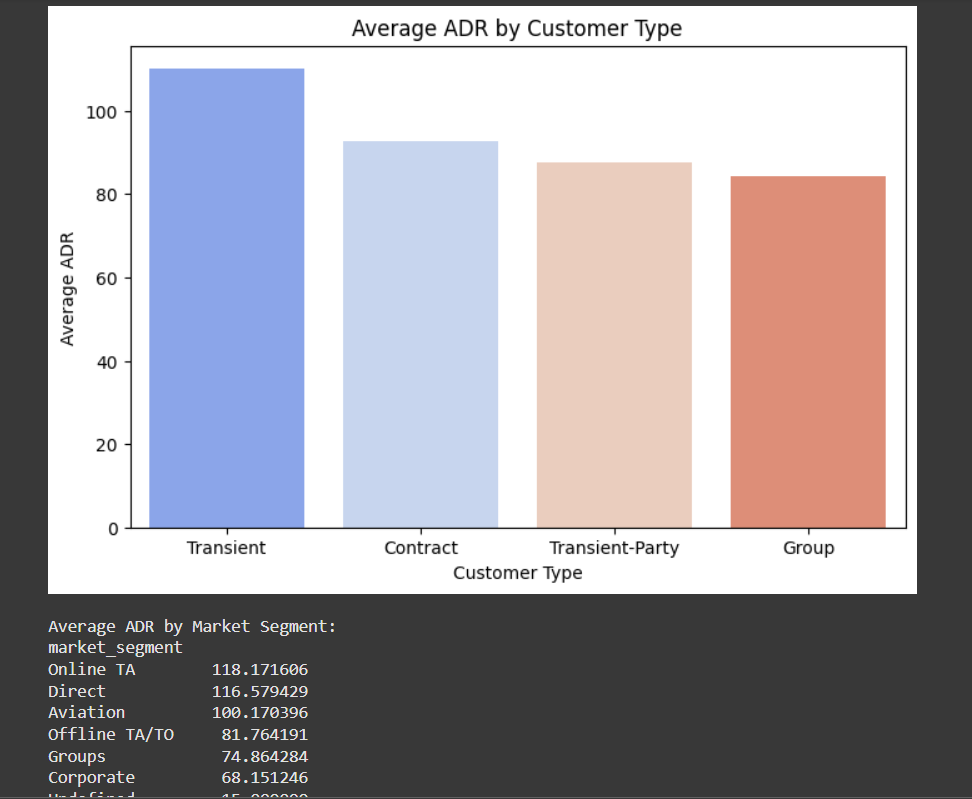


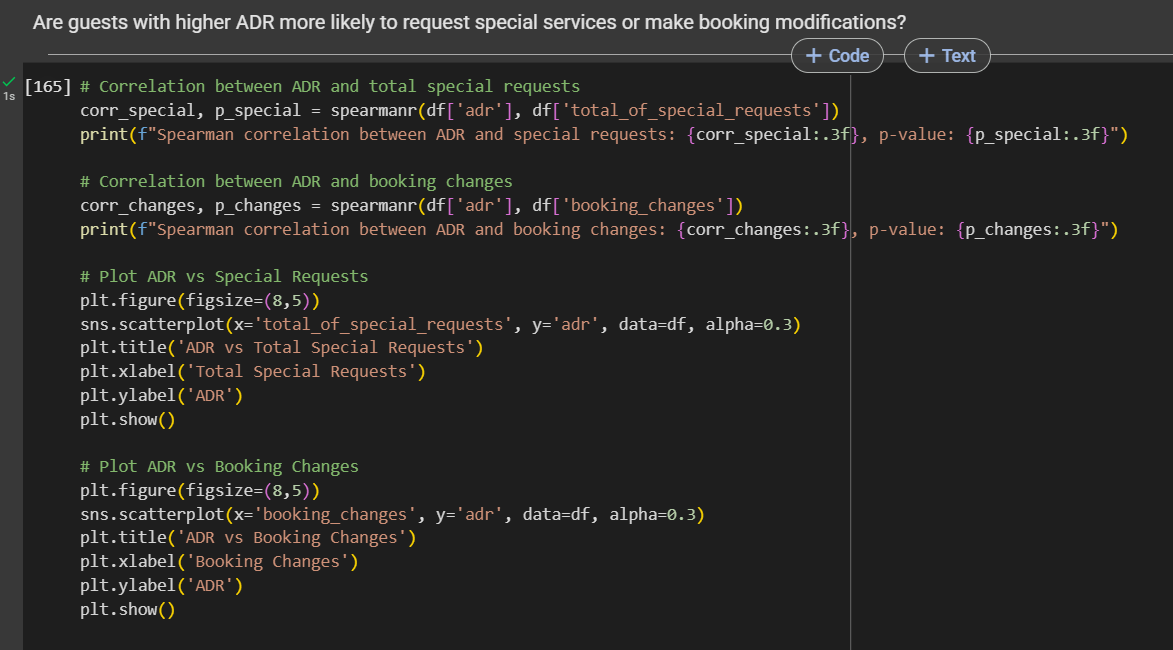


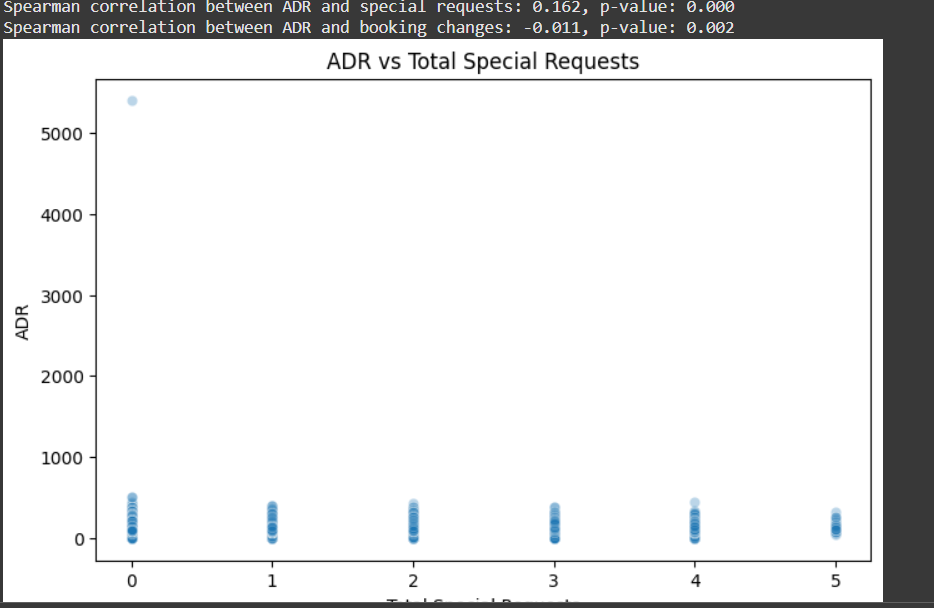


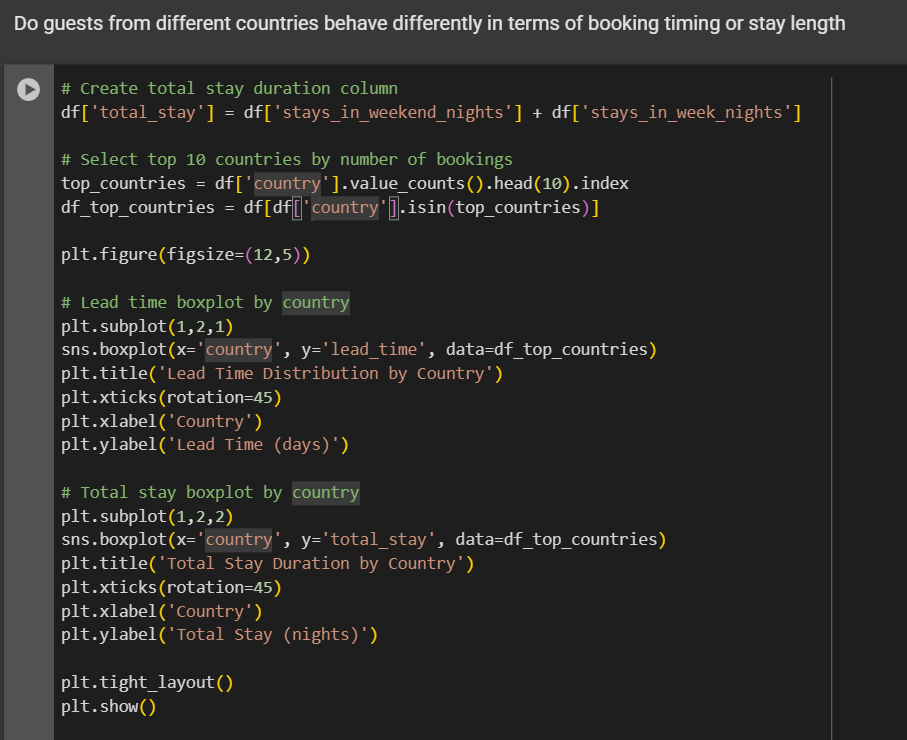


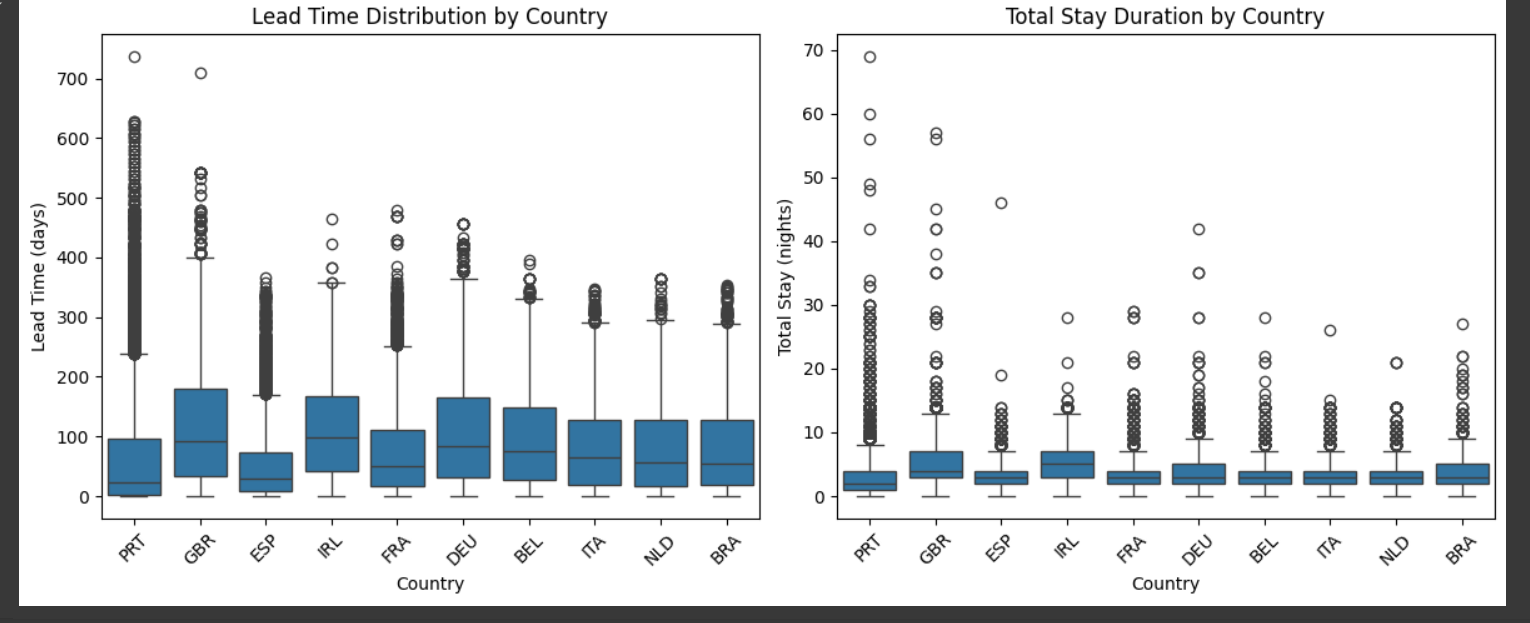


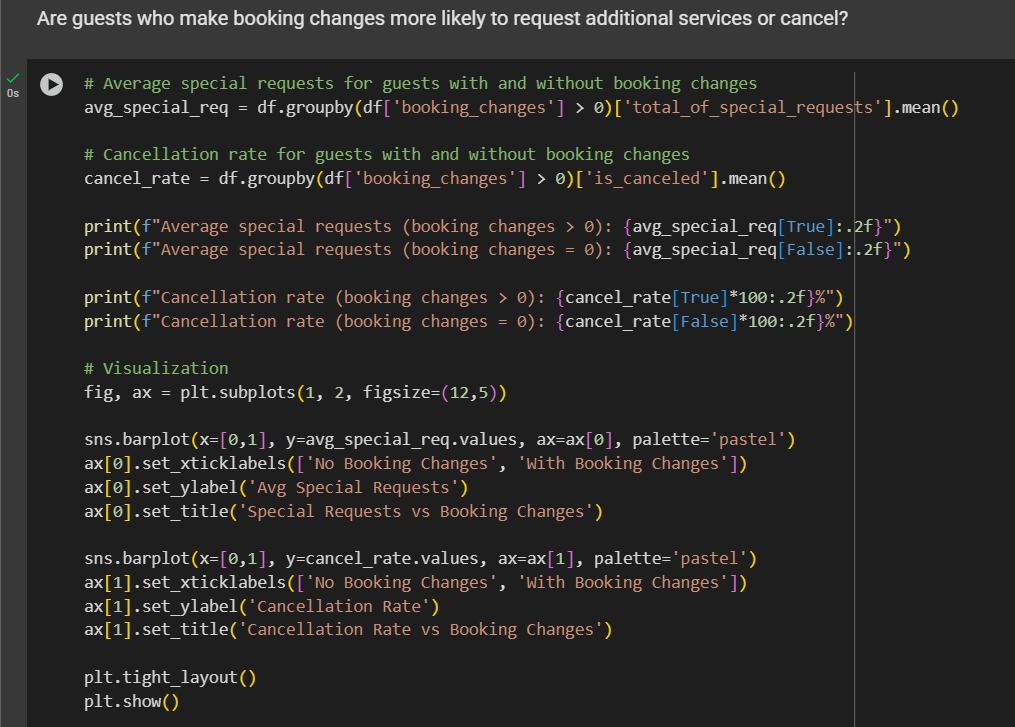














Explanation:

Guests who change bookings tend to request more special services and have a higher chance of canceling.

This insight can help hotel operations manage such guests proactively.