Report: Hotel Bookings Dataset

Name : Gautami Kadam

Roll no : 250240325026

# 1. Introduction

This report presents an in-depth analysis of the 'hotel\_bookings.csv' dataset. The dataset contains booking information for a city hotel and a resort hotel, including customer details, booking preferences, and financial data. The goal is to clean the data, handle missing values, engineer useful features, and prepare the dataset for further analysis or modeling.

**Step 1: Data Cleaning and Preprocessing**

* This initial step was dedicated to preparing the dataset for meaningful analysis.
* Handled missing values: filled numeric columns like 'children' with 0 and categorical ones like 'country' with the most common value.
* Converted string-based date columns into Python datetime format for accurate time-based computations.
* Removed outliers in 'adr' (Average Daily Rate) to avoid skewed interpretations in price-related analysis.
* Derived new features such as 'total\_nights' (total stay duration) and 'total\_guests' (total group size) to improve data granularity.
* Eliminated duplicate records to maintain data integrity.

**Step 2: 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.

**Step 3: Correlation Analysis**

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

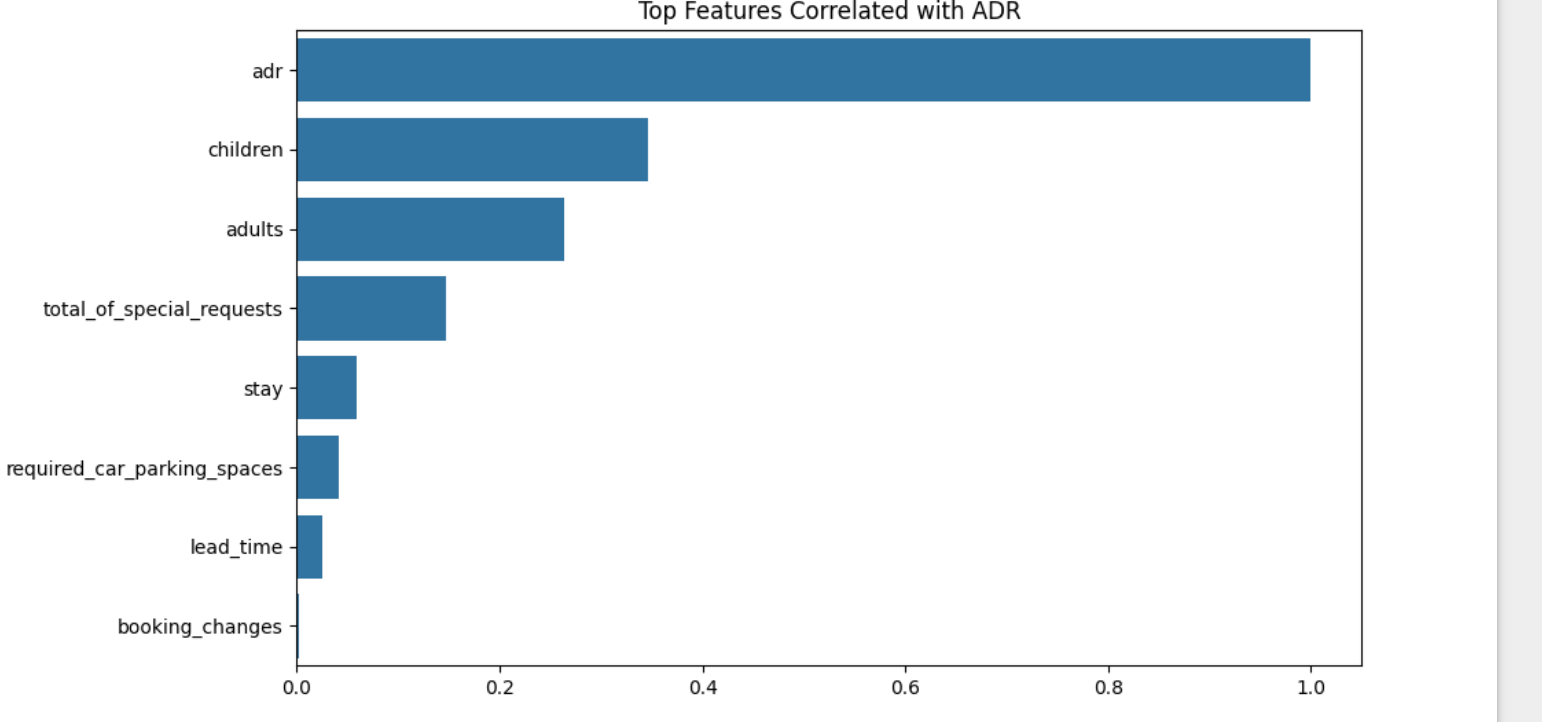
**Step 4: Hypothesis Testing**

* 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.
* 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.

**Step 5: Key Business Questions Answered**

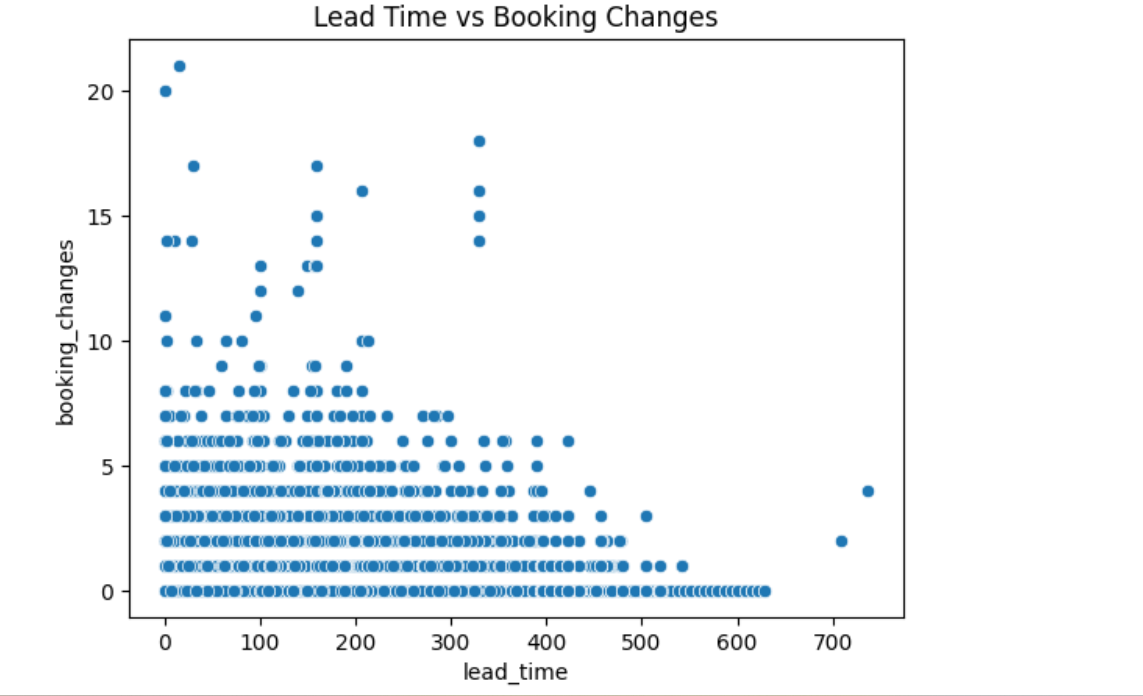
* **What influences ADR the most?**

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.



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

Ans: There is a very weak positive correlation between lead\_time and booking\_changes

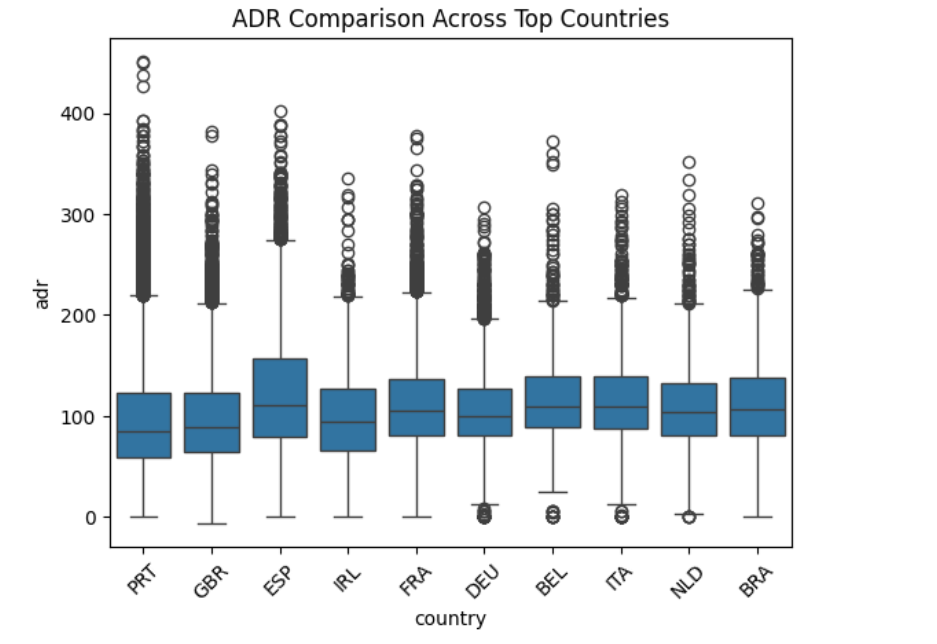


* **Are there pricing or booking differences across countries?**

Ans: Spain (ESP) has the highest median ADR, indicating higher spending.

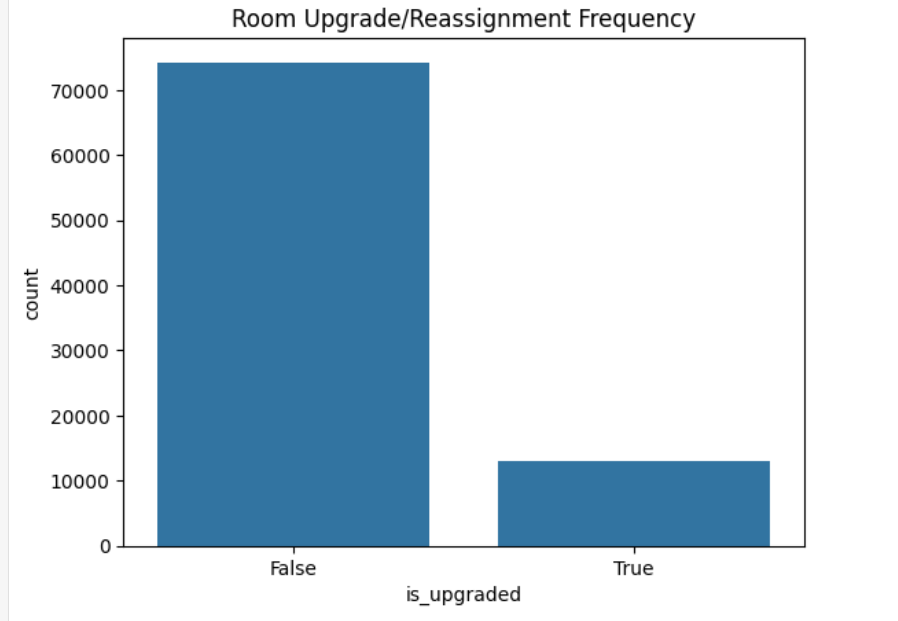
All countries show significant ADR outliers, suggesting opportunities for high-end bookings.

Portugal (PRT), UK (GBR), and France (FRA) show stable mid-range ADRs, ideal for consistent revenue targeting.



* **Is there a pattern in room upgrades or reassignment?**

Ans: 15% of the bookings resulted in a room upgrade or reassignment.



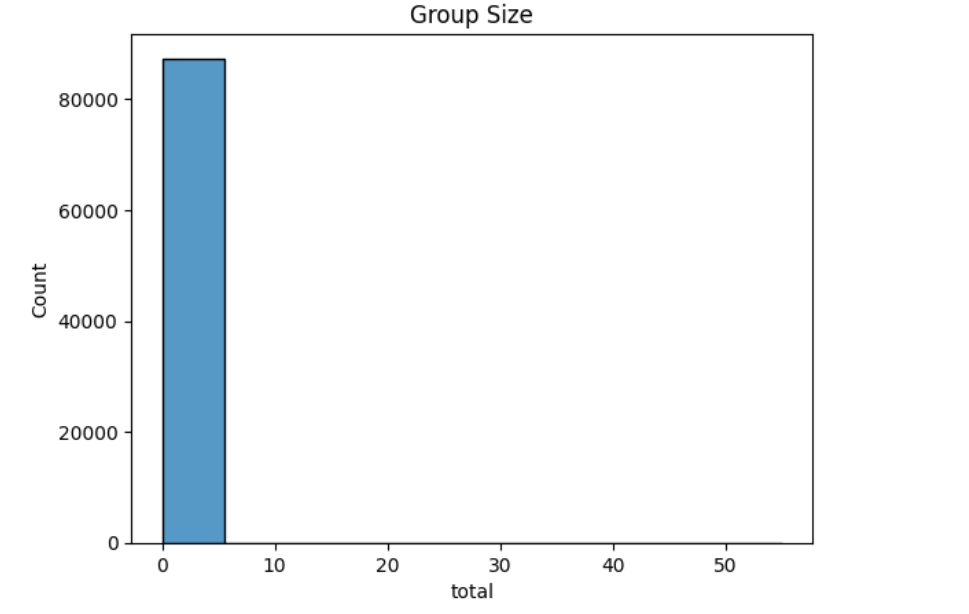
* **Are reserved room types consistently matched with assigned room types?**

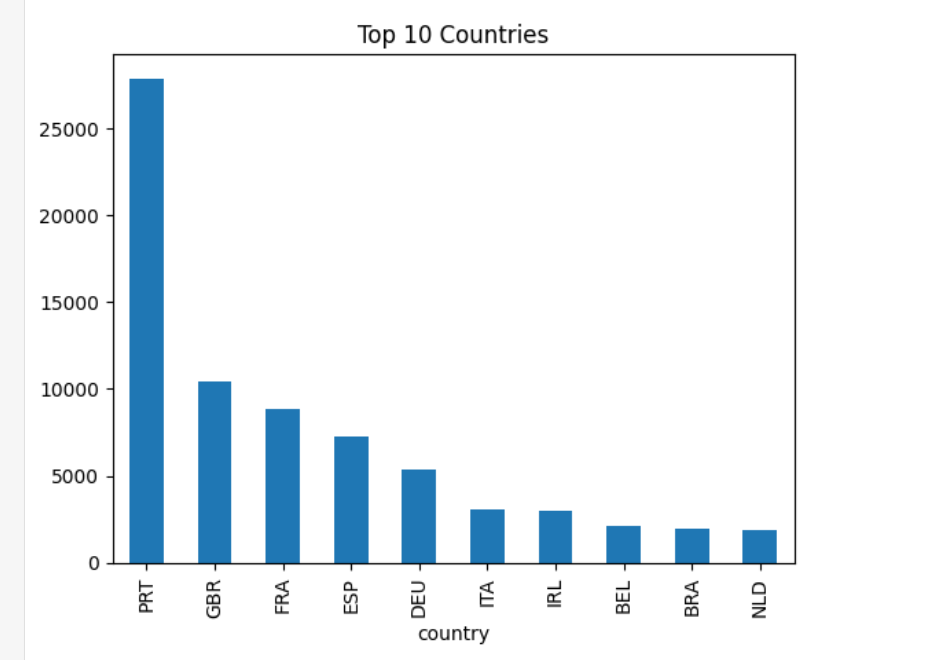
Ans: 85% guest were assigned the rooms that they booked without any upgrade.

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

Ans : Almost all the groups were of the size of 0 to 5 people.

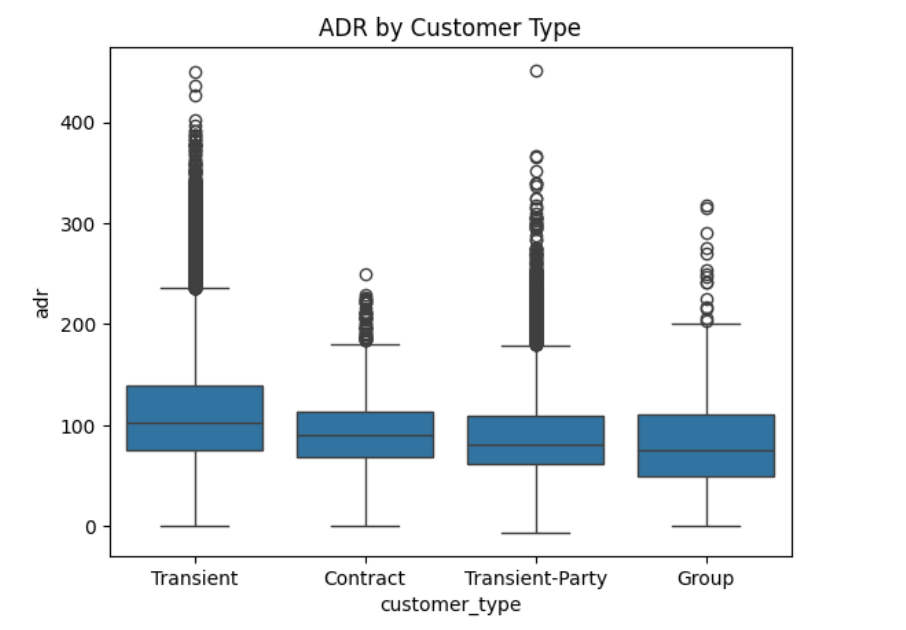
The highest number of guests were from PRT followed by GBR and France and so on.





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

Ans: All customer types show outliers, but Transient guests have the widest ADR spread, suggesting variable pricing or booking flexibility.



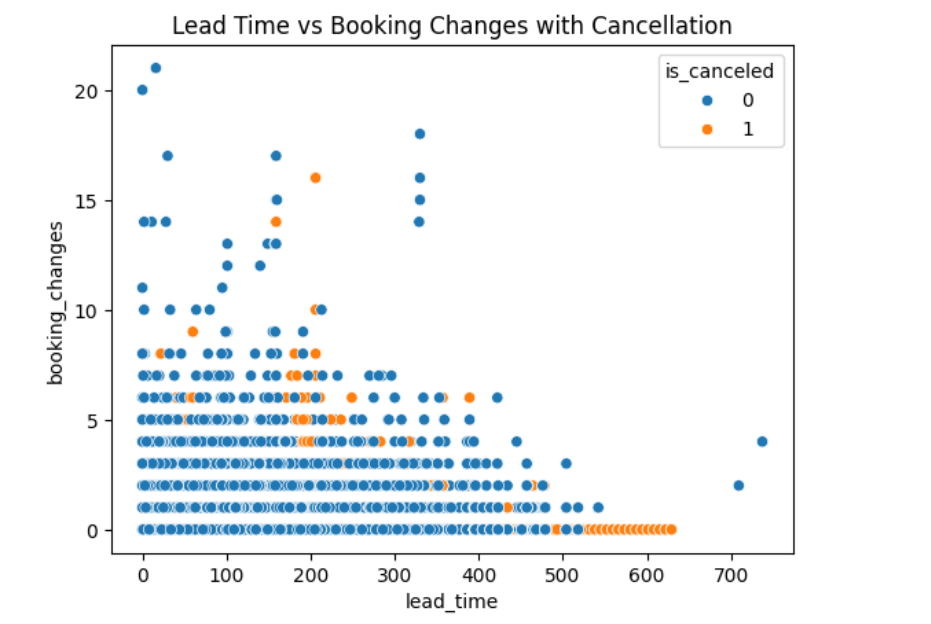
* **How does booking lead time vary across customer types and countries?**

Ans: Longer booking lead times slightly lead to higher average daily rates.



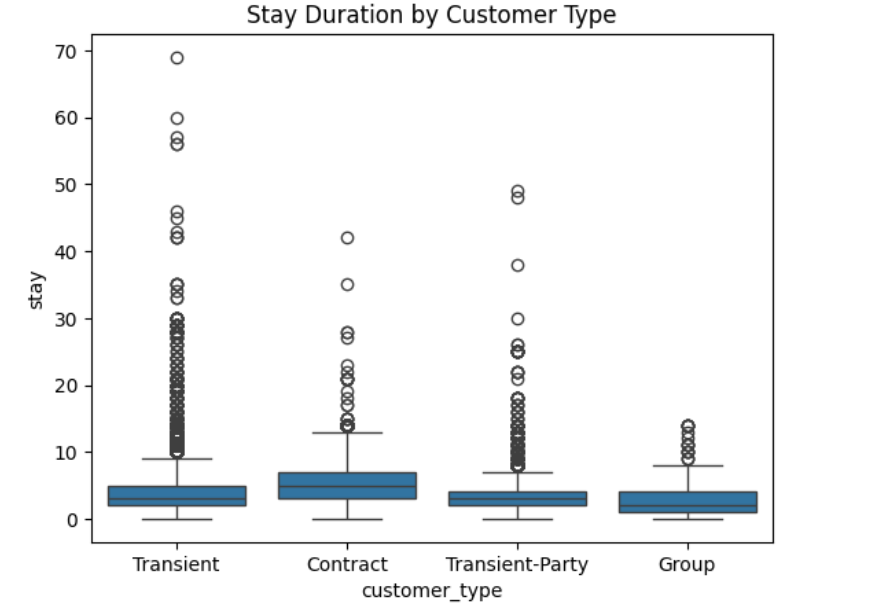
* **Are longer lead times associated with fewer booking changes or cancellations?**

Ans: longer lead times results in fewer booking changes or cancellations



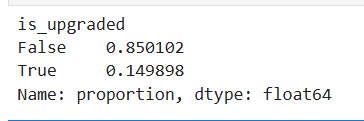
* **What is the typical duration of stay, and how does it vary by customer type or segment?**

Ans : transient type of customer tend to stay at a longer duration as comapre to group while the patter with contract and transient-party customers have a consitent stay duration.



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

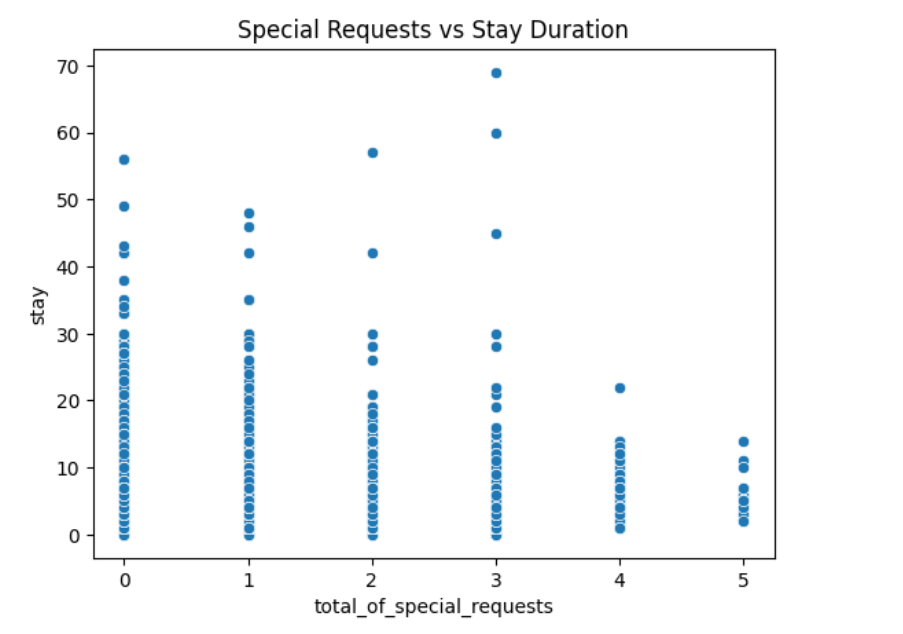
Ans: Its very rare that customers are upgraded



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

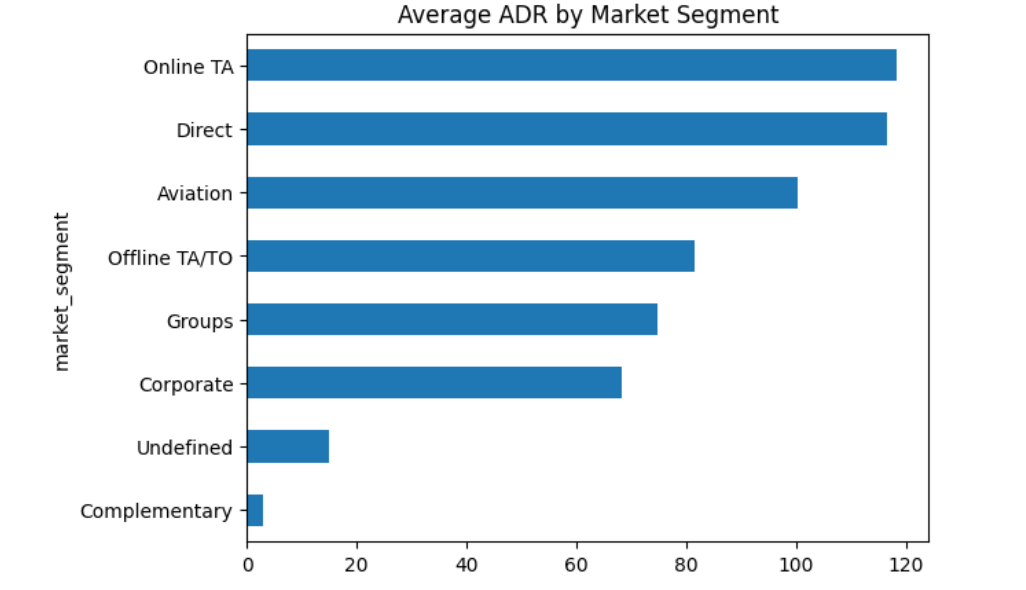
Ans: guests who make special requests more likely to experience booking changes. guests who make special requests more likely to stay for avg no of days





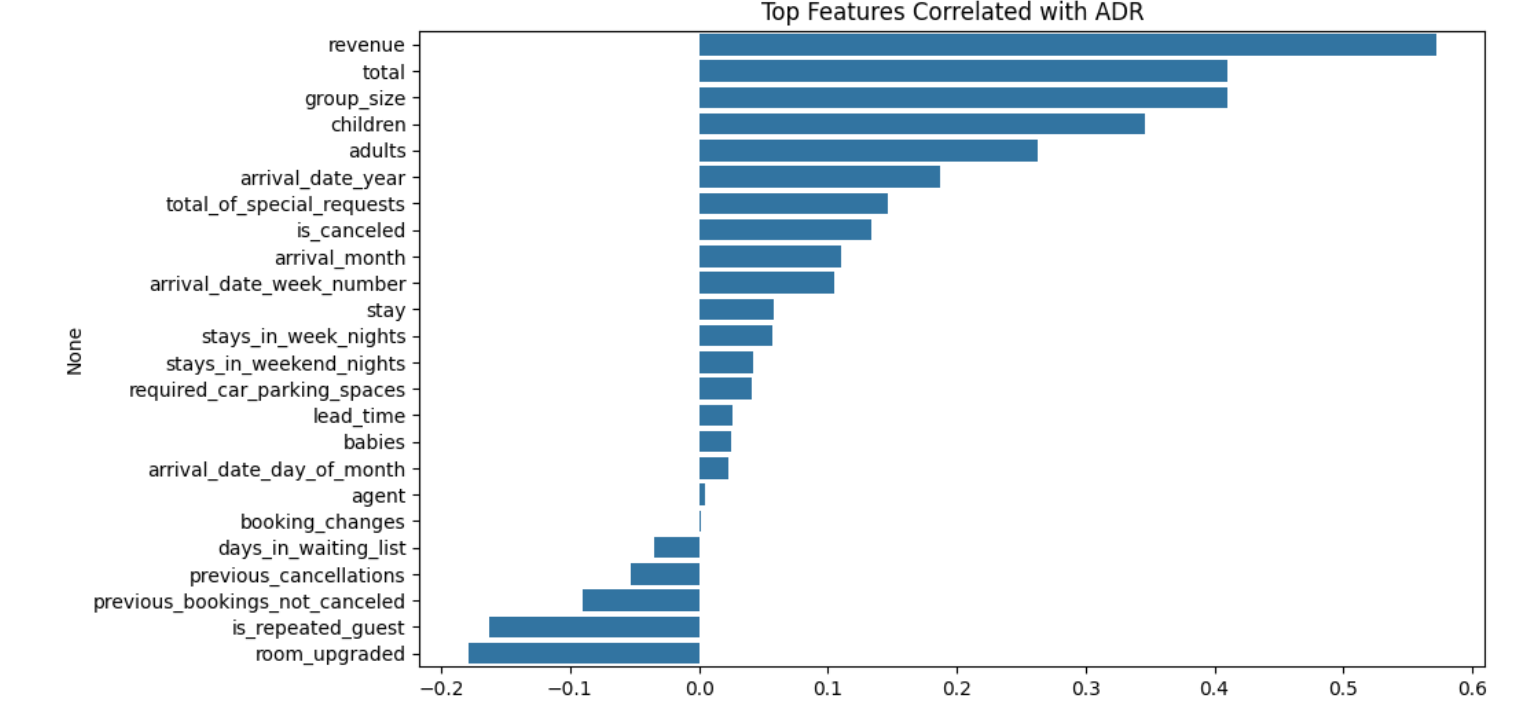
* **Do certain market segments or distribution channels show higher booking consistency or revenue?**

Ans : higher booking consistency or revenue are showed by Online TA and Direct



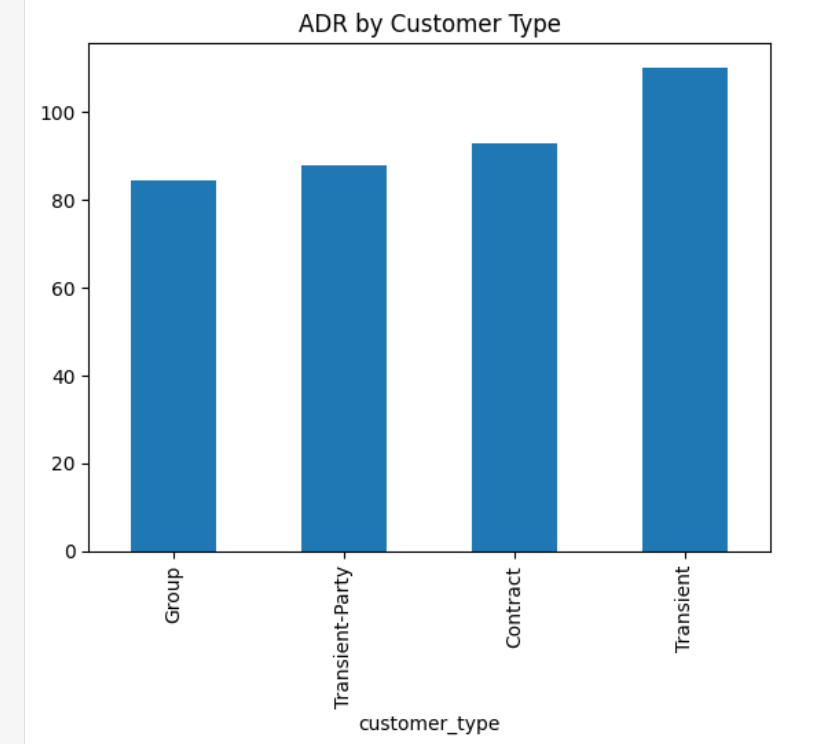
* **What factors are most strongly associated with higher ADR?**

Ans : factors like revenue, total no of guest,group\_size are the reasons for stronh ADR.



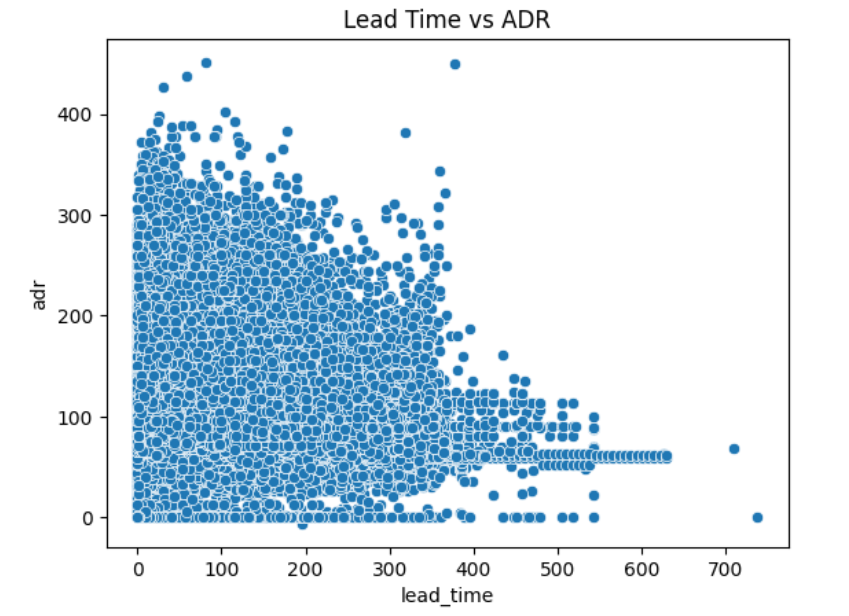
* **Are there customer types or segments consistently contributing to higher revenue?**

Ans : Transient customers contribute to higher revenue



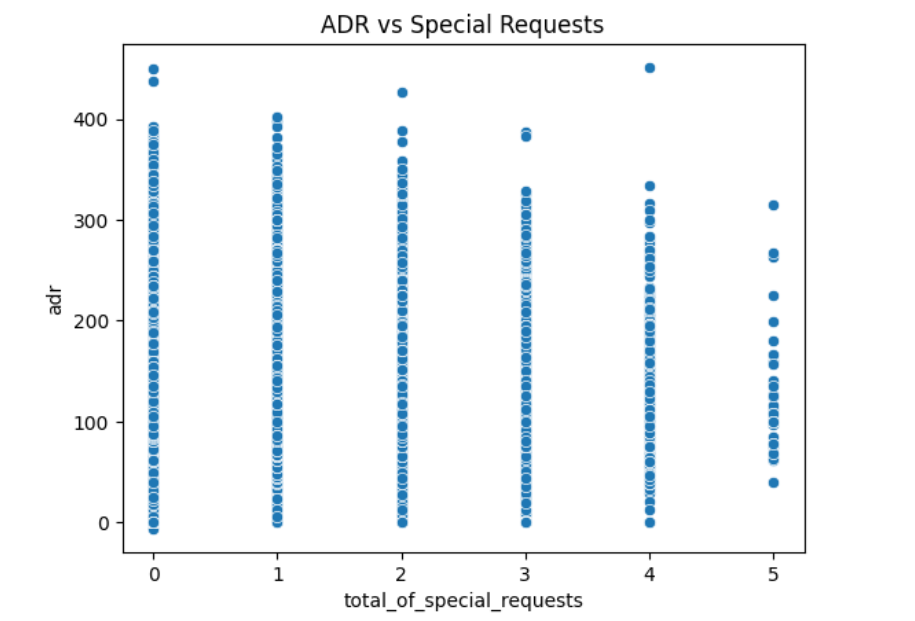
* **Do bookings with more lead time or from specific countries yield higher ADR?**

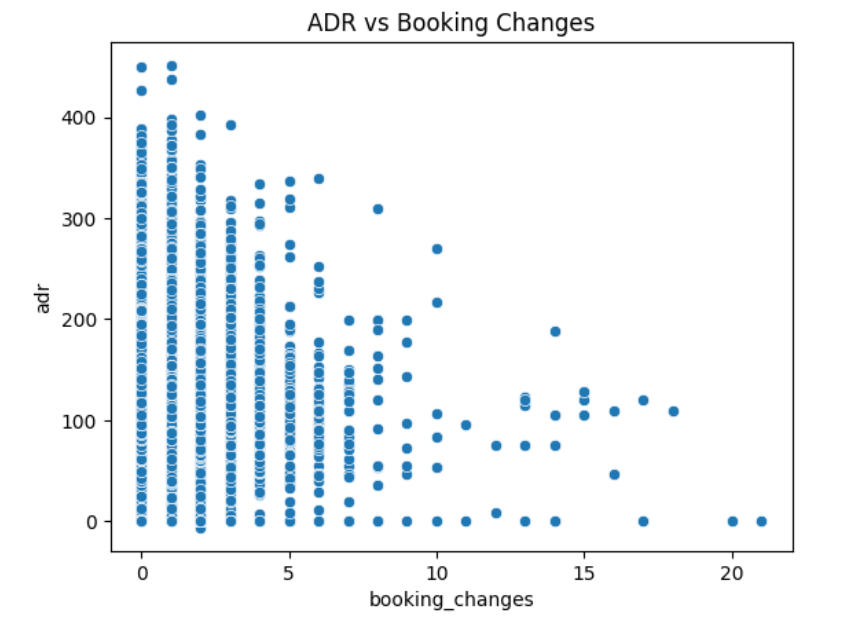
Ans: booking with lower lead time generate adr with higher ADR as comapred to more lead time



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

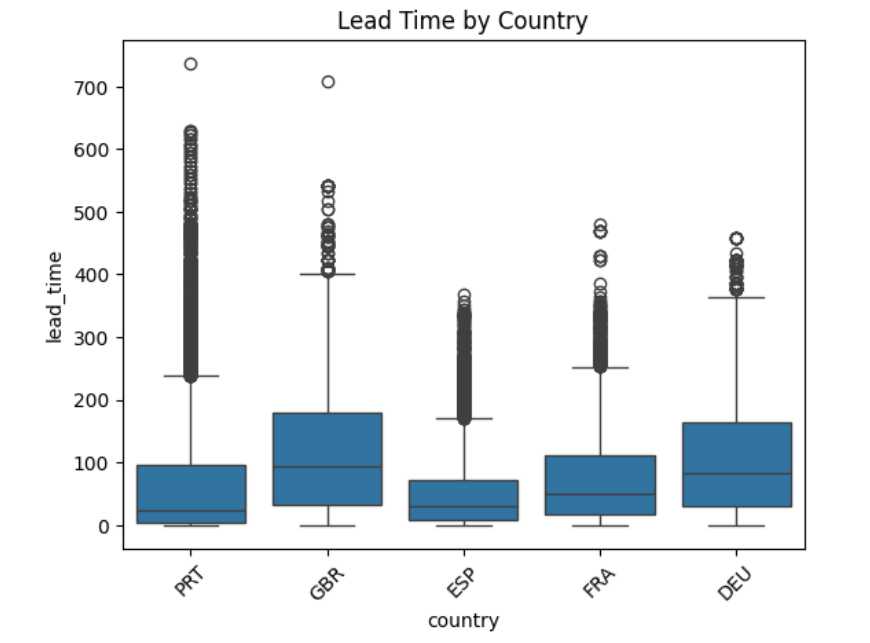
Ans : In general less number of special request result in higher ADR. IIn general less number of booking\_changes result in higher ADR





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

Ans : 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



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

Ans : More the changes less is the chance of it getting cancelled



**Conclusion:**

The analysis of the hotel bookings dataset yields several important insights that can help guide strategic business decisions:

**Data Quality and Integrity**: After handling missing values and cleaning the dataset, we ensured the data is reliable for analysis. Columns like agent, children, and country were properly imputed, while irrelevant ones like company were removed to enhance clarity.

**Customer Demographics**: The total number of guests per booking (total) helps in understanding customer types—solo travelers, families, or groups. This can guide personalized marketing and service offerings.

**Booking and Stay Patterns**: The engineered stay column reveals the length of stays, crucial for identifying trends in short- vs long-term visits. This supports optimized pricing strategies and room availability planning.

**Revenue Estimation**: By calculating the revenue column, we get a proxy for the financial value of each booking. This is essential for forecasting income, identifying high-value customers, and optimizing advertising spend.

**Seasonality**: With the creation of the arrival\_month and unified date columns, we can analyze bookings across seasons and identify peak periods. This enables targeted promotions and staff allocation.