Hotel Booking Demand Analysis

Introduction

This report presents a short analysis of hotel booking data from Kaggle, which includes records from both City Hotel and Resort Hotel. The dataset is in CSV format with over 119,000 entries and 36 features related to booking dates, guest details, room preferences, cancellations and pricing.

Using Python libraries like Pandas**,** Matplotlib**,** Seaborne and Numpy for analysis.

Dataset Description

- **hotel**: Type of hotel – either "Resort Hotel" or "City Hotel".
- **is canceled**: Indicates whether the booking was canceled (1) or not (0).
- **lead_time**: Number of days between the booking date and the arrival date.
- **arrival_date_year**: Year of arrival.
- **arrival_date_month**: Month of arrival (e.g., January, February).
- **arrival_date_week_number**: Week number (1–52) of the year for the arrival date.
- **arrival_date_day_of_month**: Day of the month of arrival (1–31).
- **stays_in_weekend_nights**: Number of weekend nights (Saturday or Sunday) included in the stay.
- **stays_in_week_nights**: Number of weeknights (Monday to Friday) included in the stay.
- **adults**: Number of adults included in the booking.
- **children**: Number of children included in the booking.
- **babies**: Number of babies included in the booking.
- **meal**: Type of meal booked (e.g., BB = Bed & Breakfast).
- **country**: Country of origin of the guest.

market_segment: Market segment designation (e.g., Online TA, Groups).

distribution_channel: Channel through which the booking was made (e.g., Direct, Corporate).

is_repeated_guest: 1 if the guest is a returning customer; 0 otherwise.

previous_cancellations: Number of past bookings canceled by the customer.

previous_bookings_not_canceled: Number of past bookings not canceled by the customer.

reserved_room_type: Code of the room type initially reserved by the customer.

assigned_room_type: Code of the room type actually assigned at check-in.

booking_changes: Number of changes made to the booking after the initial reservation.

deposit_type: Type of deposit made – "No Deposit", "Refundable", or "Non Refund".

agent: ID of the travel agency that made the booking.

company: ID of the company that made the booking.

days_in_waiting_list: Number of days the booking stayed on the waiting list before confirmation.

customer_type: Type of customer (e.g., "Transient", "Contract").

adr: Average Daily Rate – revenue earned per occupied room per day.

required_car_parking_spaces: Number of parking spaces requested by the guest.

total_of_special_requests: Total number of special requests made by the guest (e.g., twin beds, high floor).

reservation_status: Final status of the reservation – "Check-Out", "Canceled", or "No-Show".

reservation_status_date: The date on which the reservation status was last updated.

name: Full name of the guest.

- **email**: Email address of the guest.

- **phone-number**: Contact number of the guest.

- **credit_card**: Encrypted or masked credit card number of the guest.

Data Cleaning and Sorting

In this analysis, missing values were found in the columns: children**,** country**,** agent**,** and company**.** To ensure data consistency, all rows containing null values were removed using df.dropna(). As a result, the dataset size was reduced drastically from 119,390 to just 217 rows, indicating a significant amount of incomplete data. The reservation_status_date column was converted to date time format for proper time-based analysis. Furthermore, dataset features were categorized into three groups: Numerical features (20 columns, e.g., lead_time, adr, total_of_special_requests), Categorical features (16 columns, e.g., hotel, meal, market_segment), and Date-related features (e.g., arrival_date_year, arrival_date_month, reservation_status_date).

Outlier Detection Using IQR

Before outlier removal, the dataset contained 192 rows and 35 columns**.** The Interquartile Range (IQR) method was applied to two key numerical features: adr (Average Daily Rate) and lead_time (number of days between booking and arrival). Outliers were identified as data points falling outside the range [Q1 - 1.5×IQR, Q3 + 1.5×IQR] and were removed to improve data quality and reduce skewness.

After applying the outlier removal function, the dataset was reduced to 160 rows**,** resulting in the removal of 32 rows**.**

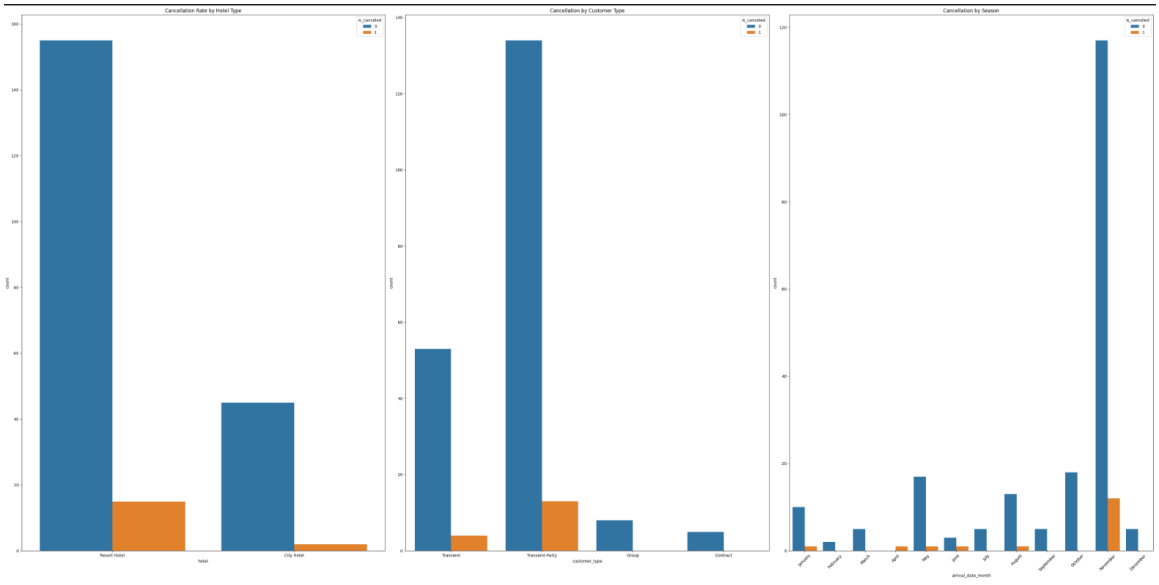
Interquartile Range (IQR) method following calculations were made:

Q1 = 25th percentile
Q3 = 75th percentile
IQR = Q3 - Q1
Lower Bound = Q1 - 1.5 * IQR
Upper Bound = Q3 + 1.5 * IQR

Data Visualizations

Booking Behavior Analysis Based on Count Plots

The following section presents a detailed analysis of hotel booking patterns across three dimensions—hotel type**,** customer type**,** and seasonality—based on count plot data. These visualizations help uncover patterns in booking volume, success rates, and cancellations, offering valuable insights for strategic planning and decision-making in the hospitality sector.



1**.** Hotel Type Performance

An analysis of booking success and cancellation across hotel types reveals that City Hotels outperform Resort Hotels in terms of reliability. City Hotels recorded a total of approximately 47 bookings, with only 3 cancellations, resulting in a cancellation rate of 6.4% and a success rate of 93.6%**.** In contrast, Resort Hotels accounted for approximately 170 bookings, of which 15 were cancelled, producing a cancellation rate of 8.8% and a success rate of 91.2%**.**

These findings suggest that City Hotels are slightly more reliable, potentially due to their appeal to business travelers or short-stay guests with less likelihood of changing plans. Resort Hotels, typically favored for vacation stays, may see higher cancellation rates due to fluctuating travel plans or seasonal factors. Nevertheless, both hotel types show

relatively low overall cancellation rates, highlighting consistent guest commitment across the sector.

2. Customer Type Reliability

Booking patterns also differ significantly across customer types. Group and Contract bookings emerge as the most reliable segments, with a 0% cancellation rate in both cases. Group bookings—typically made for organized tours or events—recorded 8 successful reservations, while Contract bookings, likely representing corporate clients, showed 5 successful bookings. These segments represent zero-risk categories and should be strategically prioritized.

On the other hand, Transient and Transient-Party customers, who represent individual and group leisure travelers respectively, show moderate cancellation rates. Transient customers booked 59 times, canceling 5 (8.5%), while Transient-Party customers made the most bookings (approximately 150), with 15 cancellations, yielding a cancellation rate of 10%**.** Though they form the largest customer segment by volume, their reliability is slightly lower, suggesting the need for engagement and retention strategies.

These patterns highlight a crucial insight: while leisure travelers drive volume**,** business and contract clients ensure booking stability, making them valuable for predictable revenue streams.

3. Seasonal Booking Trends

Booking activity displays a distinct seasonal pattern**,** with the month of November emerging as the dominant peak season**.** In November alone, the hotel registered approximately 130 bookings, 13 of which were cancelled, reflecting a 10% cancellation rate. October also showed heightened activity with 20 bookings and 2 cancellations. The months of August and September demonstrate a gradual increase leading up to the peak.

Conversely, February to April represent the lowest activity period, with minimal bookings (ranging from 1 to 2). March, June, July, and December also experienced very limited hotel activity**,** highlighting the off-season. January and May showed moderate levels of activity, with 12 and 17 bookings respectively.

Cancellation rates during peak months remain consistently within the 8–10% range, indicating that despite increased volume, the likelihood of cancellations does not rise significantly. This provides a stable foundation for forecasting and capacity planning.

Strategic Business Implications

Revenue Management

The reliability of Group and Contract bookings (0% cancellations) presents an opportunity for targeted campaigns and long-term partnerships, especially in the off-season. November's peak requires focused capacity planning and pricing strategies, while City Hotels, being slightly more reliable, can be leveraged for consistent occupancy throughout the year.

Marketing Strategy

The Transient-Party segment, although high in volume, shows moderate cancellation behavior and should be the focus of customer engagement and loyalty programs. Seasonal patterns highlight the need for targeted promotions between February and April to improve off-season performance. The development of corporate contract relationships may enhance revenue stability during low-demand months.

Operational Planning

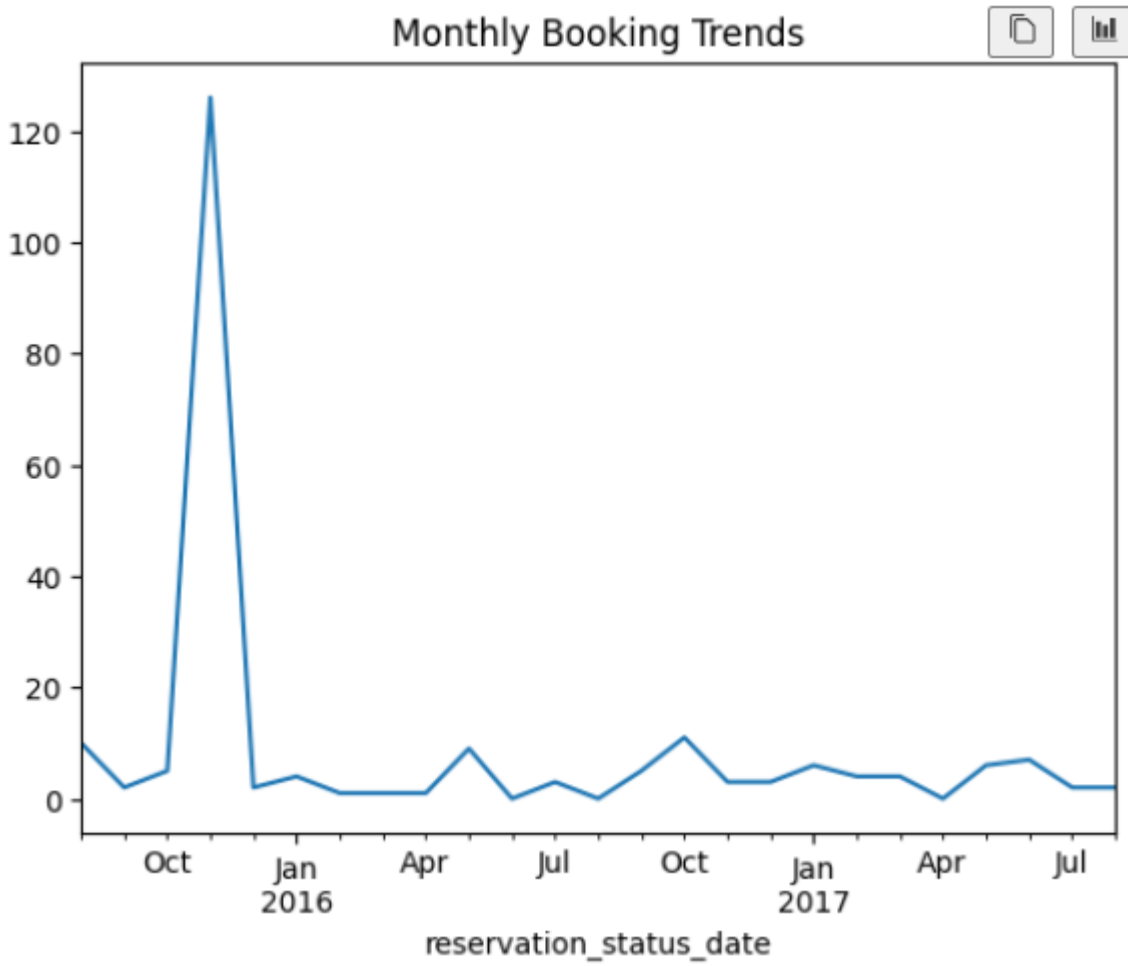
Insights from seasonal trends should guide staffing decisions, with peak deployments in November and scaled-down operations from February to April. Given the consistent 8–10% cancellation rate, an overbooking buffer can be safely incorporated into operational planning. Customer segmentation based on reliability should be used to inform booking policies and service customization.

Risk Assessment

Overall, the dataset shows low cancellation risk, generally within a manageable range of 6–10%. However, the heavy reliance on November bookings introduces seasonal concentration risk**,** underlining the importance of diversifying customer segments and spreading demand across the year.

This analysis clearly demonstrates that the hotel booking business is highly seasonal and that customer segments vary significantly in reliability**.**

Time series line plot



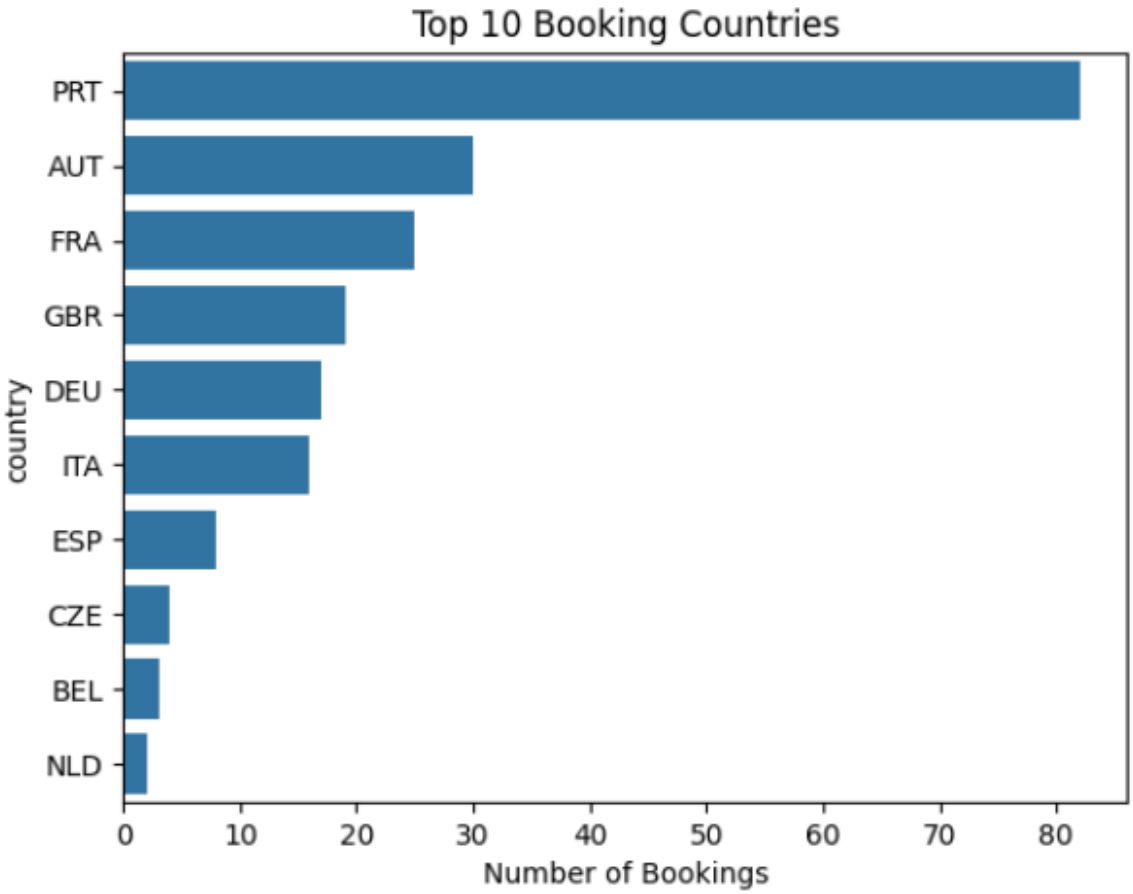
The monthly booking trend shows a highly seasonal pattern, with a sharp peak in November 2015 (~125 bookings), likely due to festivals or tourism campaigns. December 2015 follows with moderate bookings (~45), but a steep drop occurs in January 2016**,** falling to just 8–10 bookings—a decline of over 90%. From February to October 2016**,** bookings remain consistently low (2–5 per month), marking a prolonged off-season.

In subsequent years, the pattern flattens, with no major peaks and booking levels stabilizing around 5–8 per month. This suggests possible market saturation, competition, or internal issues. The business appears heavily dependent on November, with nearly half of total bookings coming from that month, posing serious revenue and operational risks**.**

To reduce dependency, the business should target off-season months, promote corporate and group bookings, and explore new marketing strategies. Long-term sustainability

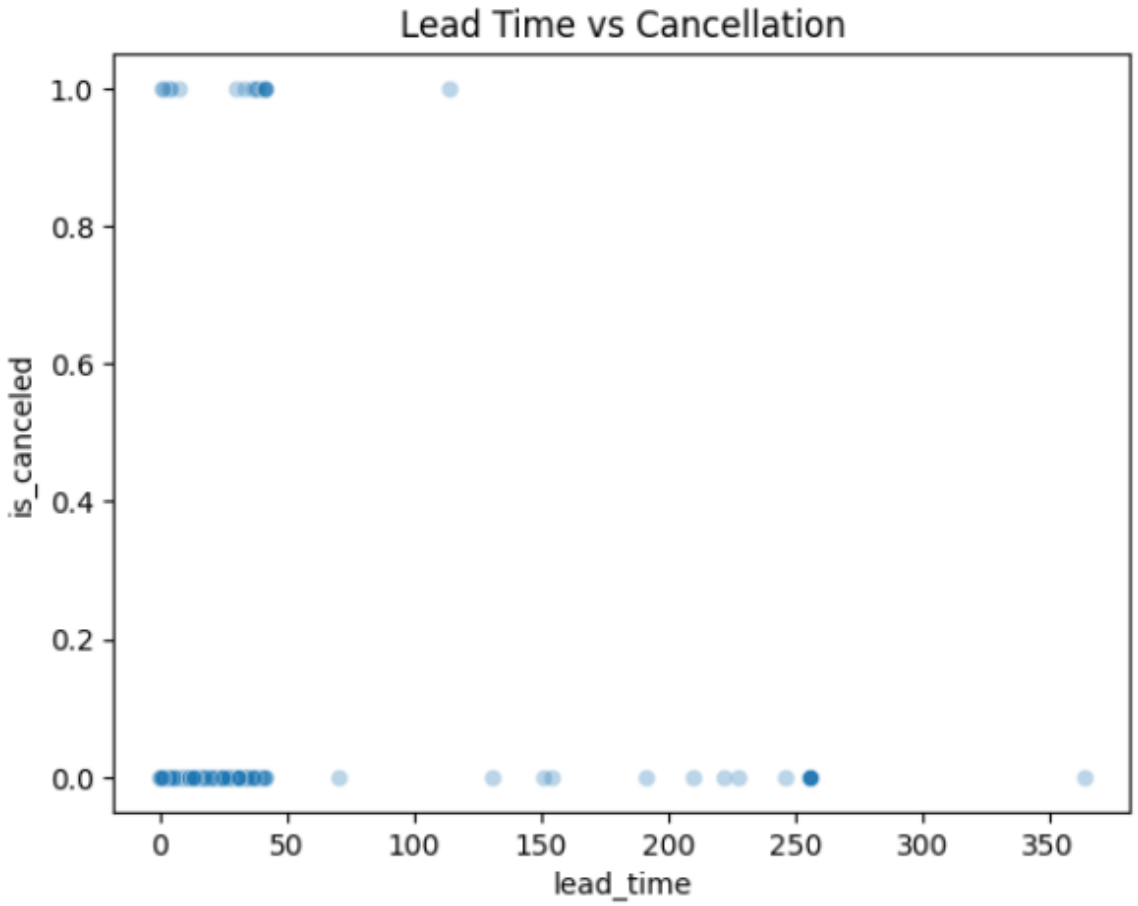
depends on diversification, seasonal planning, and financial flexibility to manage the uneven demand cycle.

Geographical analysis



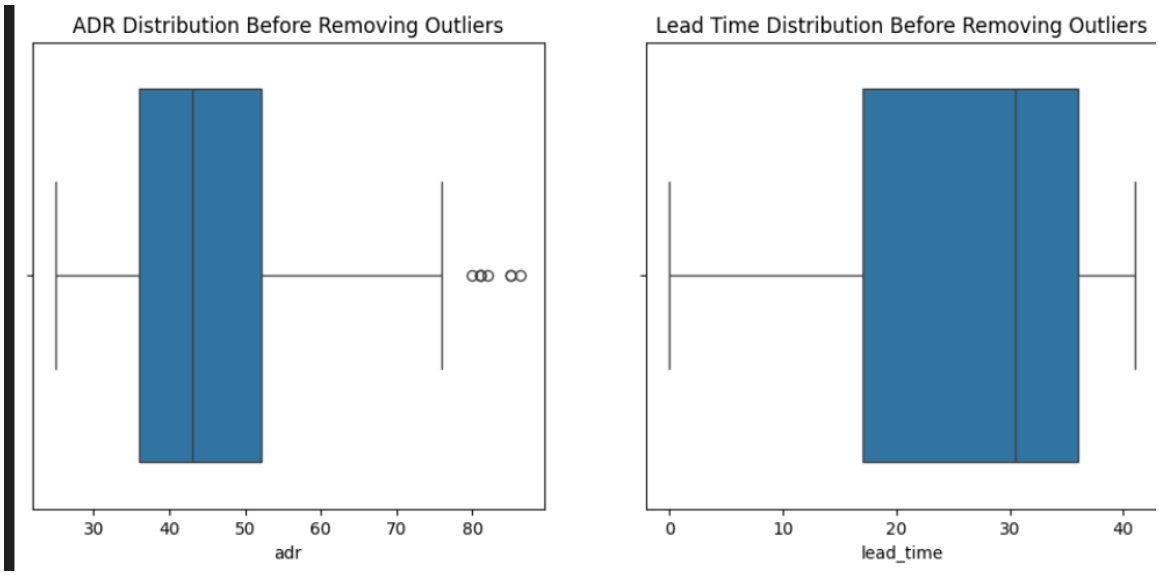
The majority of hotel bookings come from Portugal (PRT**)**, accounting for around 40% of total bookings, making it the dominant market. This is followed by Austria**,** France**,** Great Britain, and Germany**,** indicating a strong focus on Western and Central European customers. Countries like Italy, Spain, Czech Republic, Belgium, and Netherlands contribute smaller volumes. Importantly, all top booking sources are from Europe, with no representation from other continents, showing a lack of geographic diversity. This suggests an opportunity to expand into non-European markets to reduce regional dependency and increase global reach.

Scatter plot of Lead Time vs. Cancellation



The scatter plot shows that most bookings occur with a lead time of 0–50 days, which also includes a mix of both cancellations and successful stays, indicating higher uncertainty in short-term bookings. The 50–100 day window stands out as the most stable period, with decent booking activity and fewer cancellations, suggesting it as an ideal lead time range. Bookings beyond 150 days are rare and scattered, with unpredictable outcomes. Overall, this indicates that guests prefer short to mid-range advance planning, and optimizing pricing and policies around the 50–100 day window could help improve booking stability and reduce cancellation risk.

Box plot



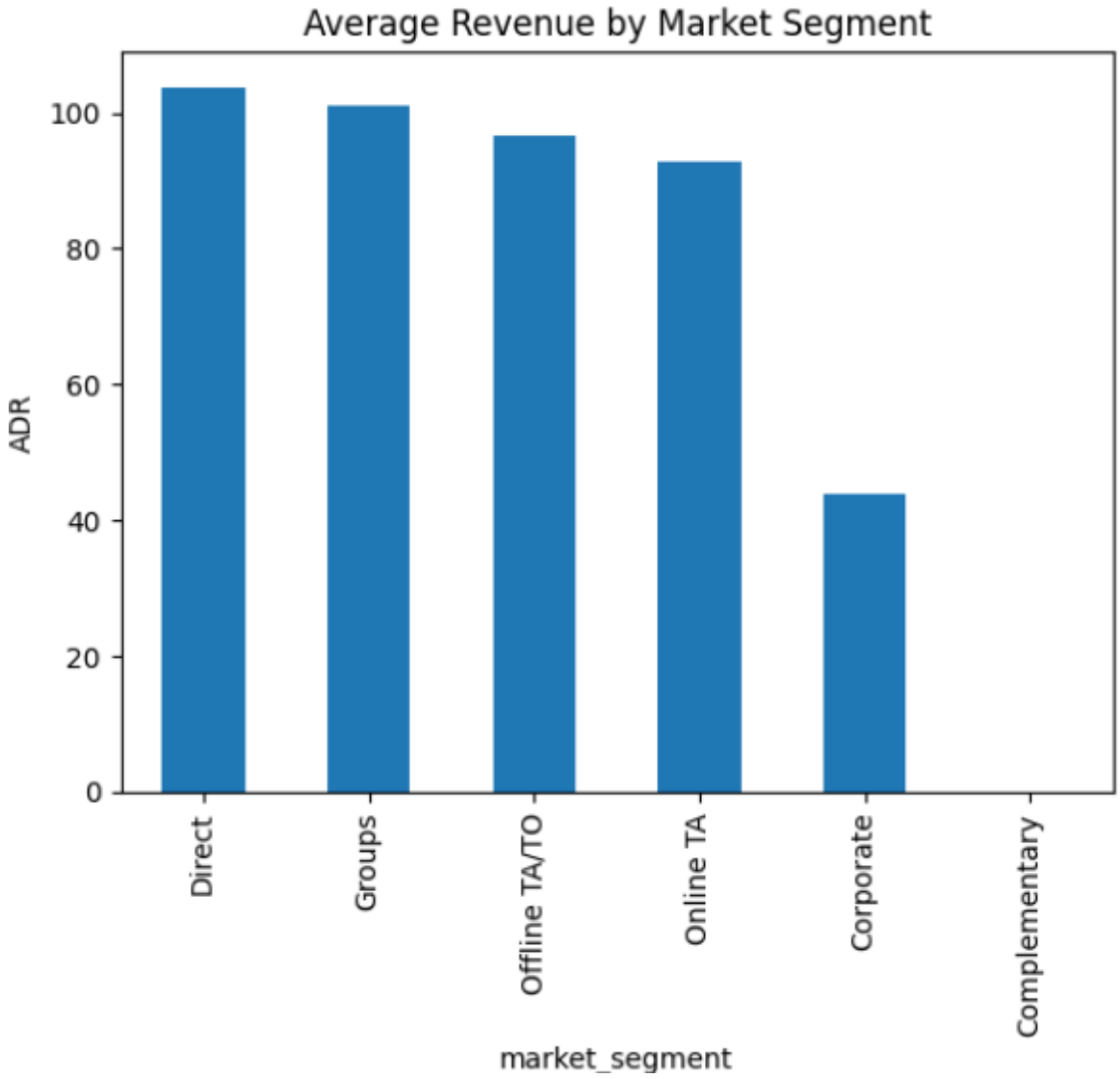
Before Removing Outliers

The ADR (Average Daily Rate) box plot shows that most bookings fall within the ₹45–₹60 range, with a median around ₹50–₹55. However, several bookings above ₹80 appear as outliers, indicating the presence of premium customers. Similarly, the lead time plot reveals a median of about 15–20 days, but with a few scattered values above 35 days, representing early planners. These outliers suggest a small segment of high-value or proactive customers, adding variability to overall booking behavior.

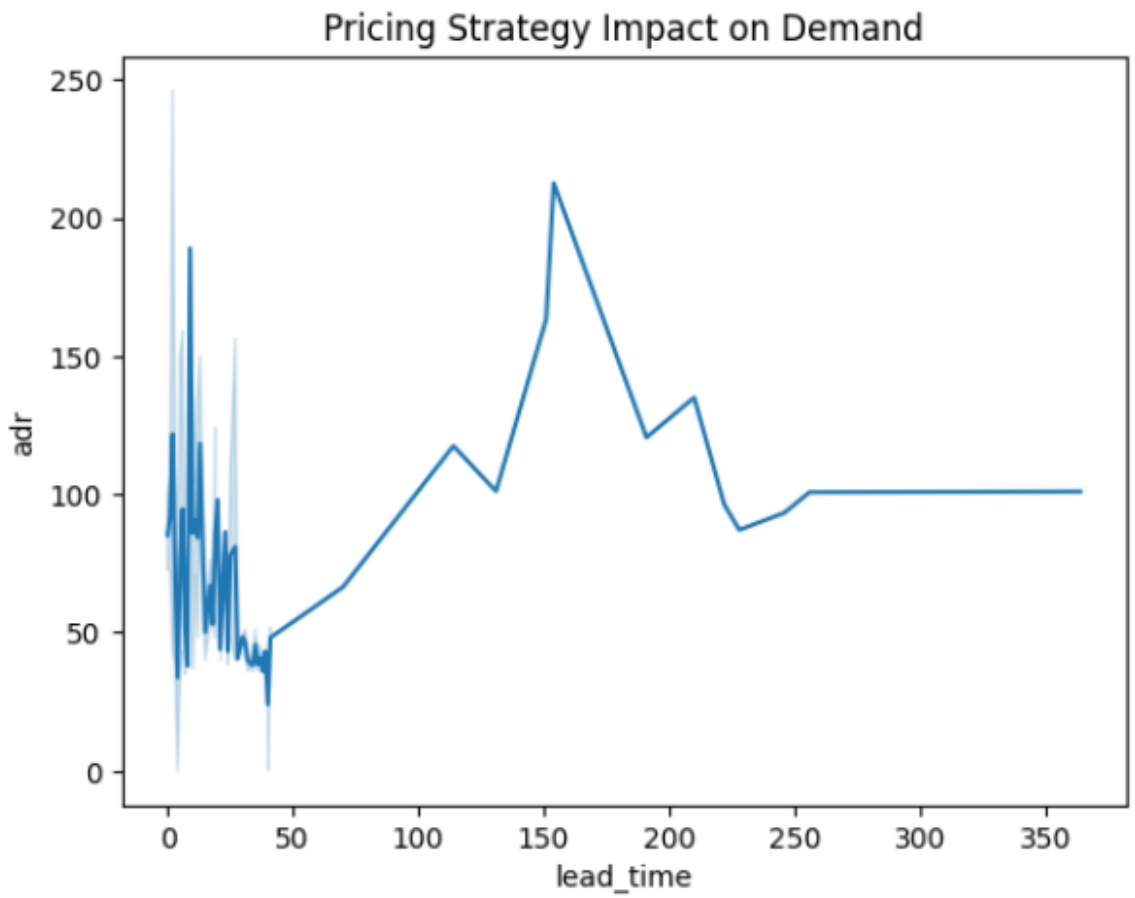
After Removing Outliers

Once outliers are removed, both ADR and Lead Time distributions appear more symmetrical and tightly focused. ADR is now clearly centered on ₹45–₹60, reflecting the typical pricing for most bookings. Lead time becomes more concentrated between 5 and 30 days, showing that the majority of guests book within a short to medium time frame. Removing outliers helps highlight the core business trends and simplifies operational planning and forecasting.

Revenue Analysis

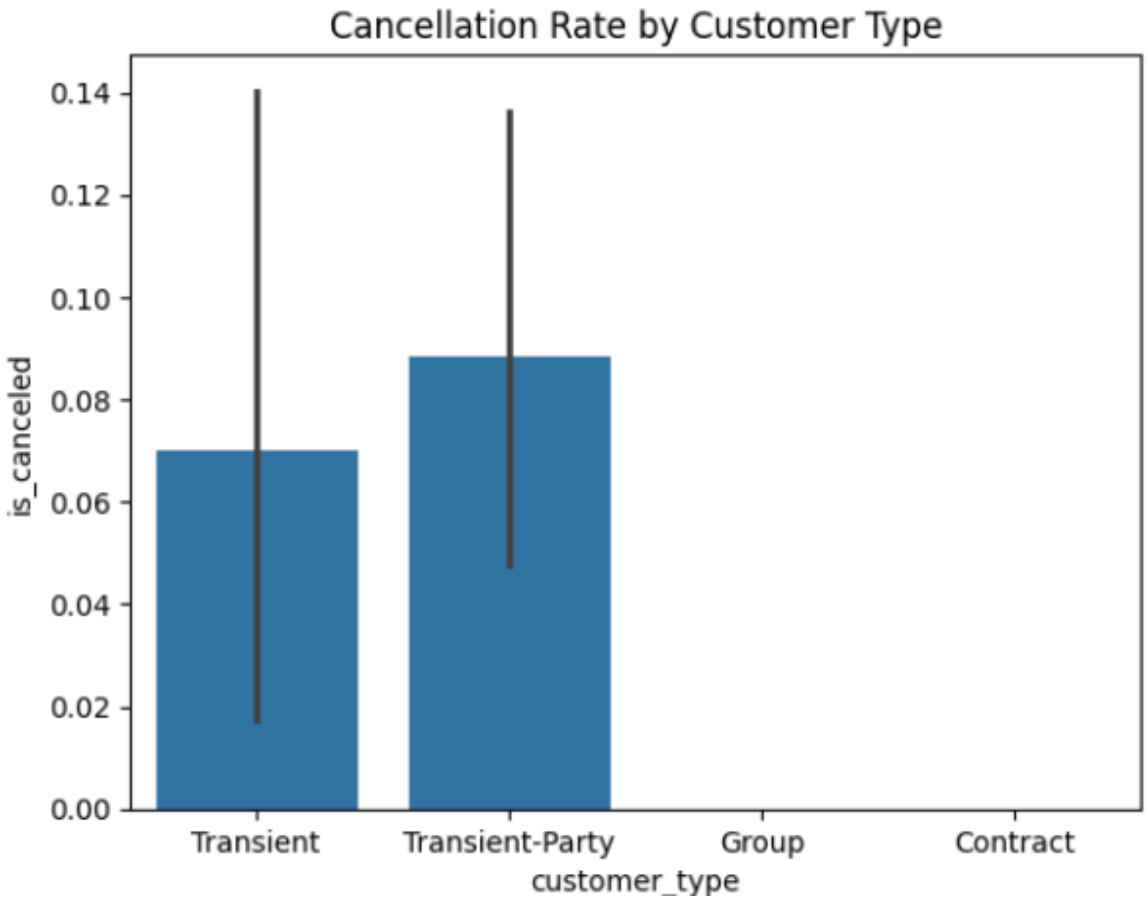


The bar chart showing Average Revenue by Market Segment reveals that *direct bookings* generate the highest revenue (ADR ~103–105) due to zero commission costs, making them the most profitable. Group bookings follow closely (ADR ~100–102), benefiting from bulk reservations. Mid-tier segments like *offline* and *online travel agents* bring moderate ADRs (~92–98), though online channels face higher commission cuts. Corporate bookings, with significantly lower ADR (~44–46), reflect discounted rates typical of B2B agreements, while complementary stays yield near-zero ADR, indicating promotional or free services that contribute little directly to revenue.

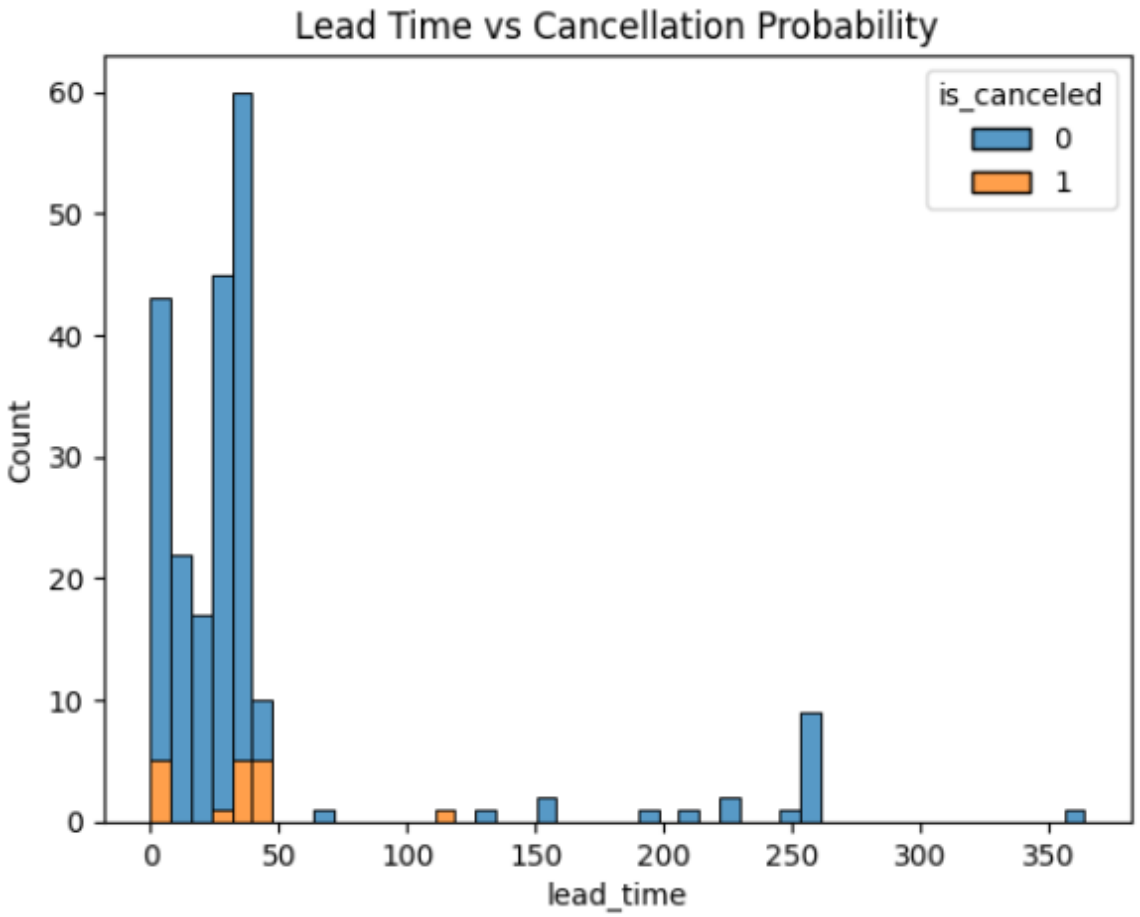


The line plot of Lead Time vs. ADR illustrates a clear dynamic pricing trend. In the 0–50 days range, ADR fluctuates widely (30–190+), showing effective last-minute premium pricing during high demand. Between 50–150 days, ADR increases steadily (~40–120), reflecting a well-executed early-bird strategy. For bookings made 150–250+ days in advance, ADR reaches premium levels (~180–220), likely for events or exclusive periods, though volume is low. Beyond 250 days, ADR stabilizes (~100), indicating standard pricing. This pattern highlights the success of lead time-based pricing, where both urgency and advance planning are leveraged for revenue optimization.

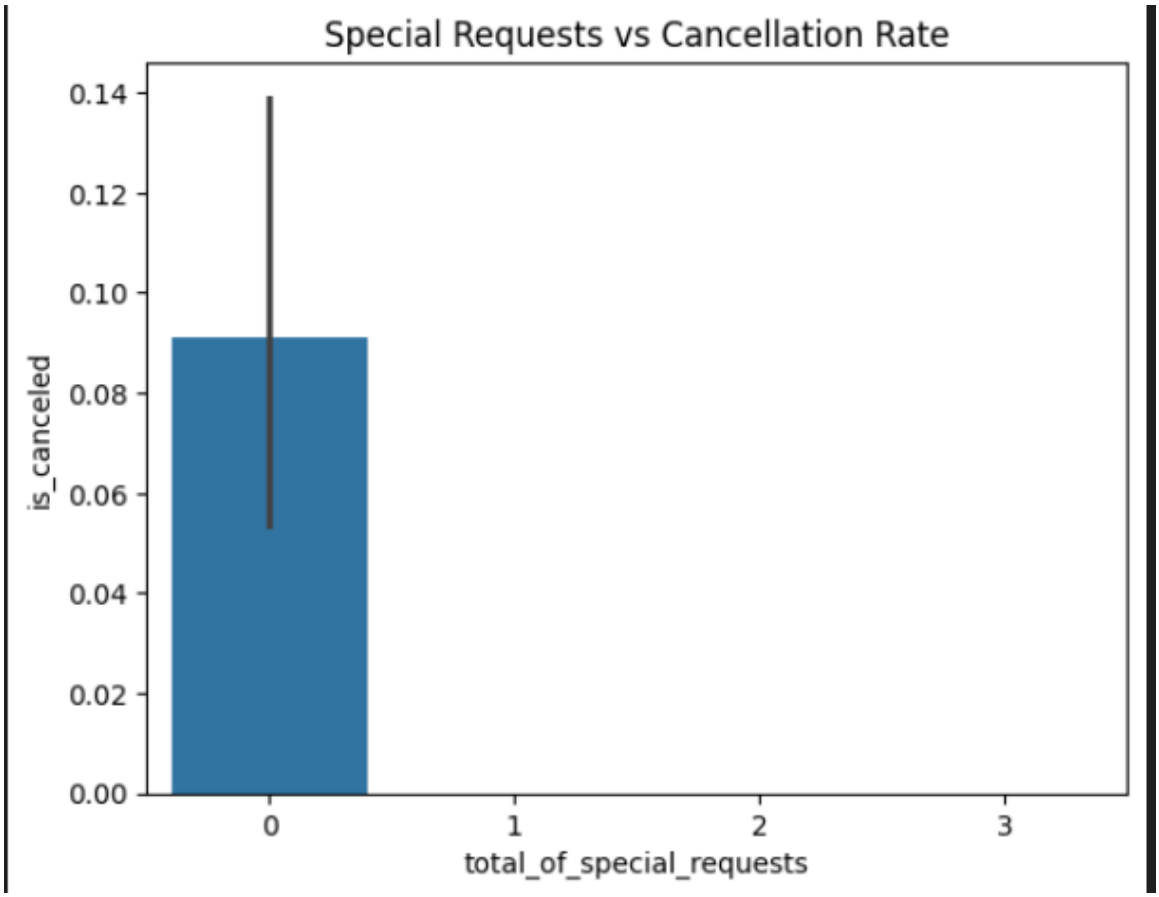
Customer Behavior Analysis



The cancellation rate by customer type shows that *Transient-Party* bookings carry the highest risk (~9%), followed by *Transient* individual travelers (~7%), both of which represent leisure segments. In contrast, *Group* and *Contract* customers have a 0% cancellation rate, indicating exceptional reliability—likely due to business or institutional planning. This suggests that targeting and retaining group or contract customers can reduce operational uncertainty.



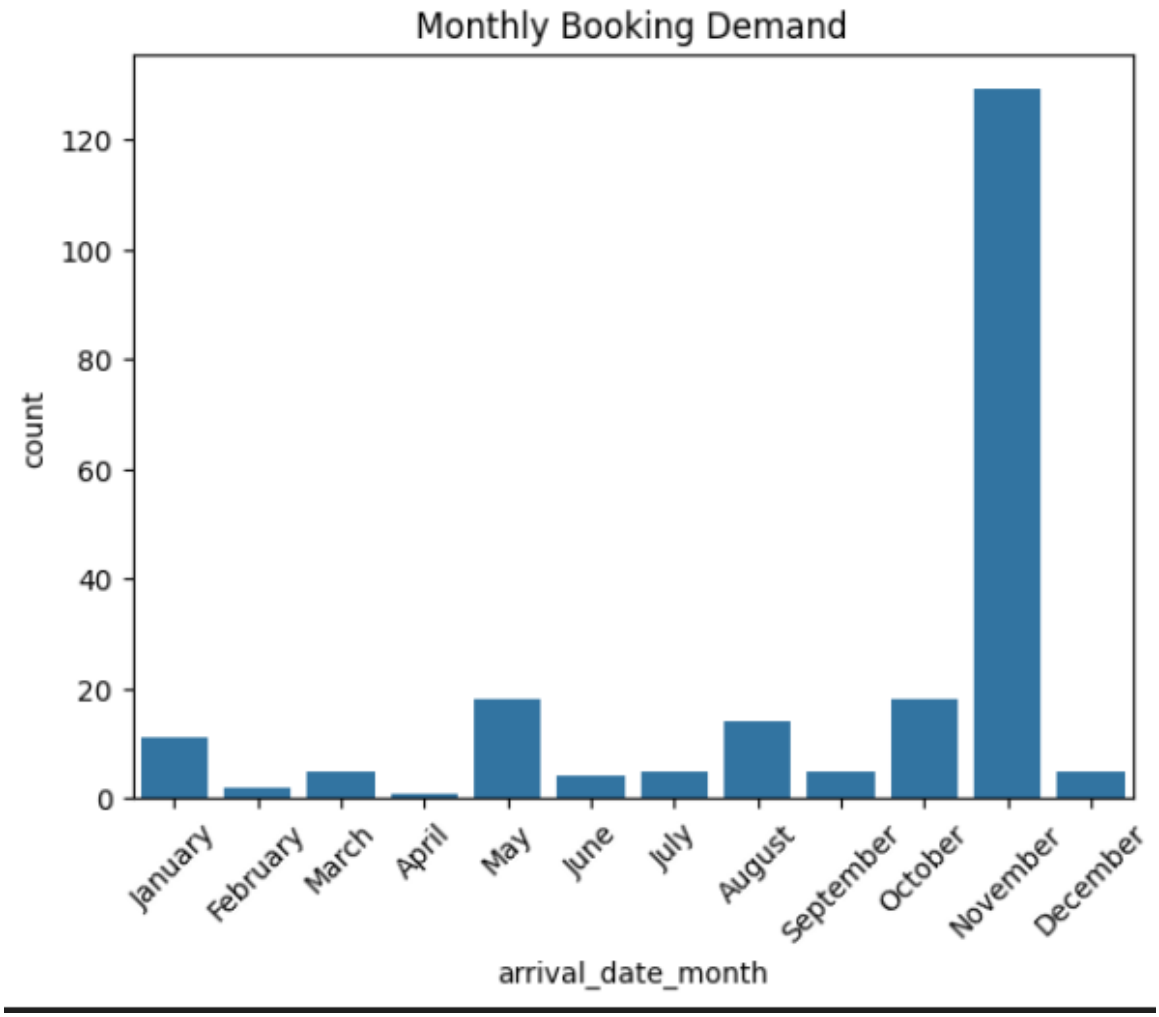
The lead time vs. cancellation histogram reveals that the majority of bookings occur within 0–50 days, but this window also carries a mixed cancellation pattern. The 50–150 day range offers a sweet spot with moderate booking volume and lower cancellation probability, making it an ideal target for promotional offers. Beyond 150 days, booking volume is very low and cancellations are unpredictable, requiring cautious handling and possibly tailored packages.



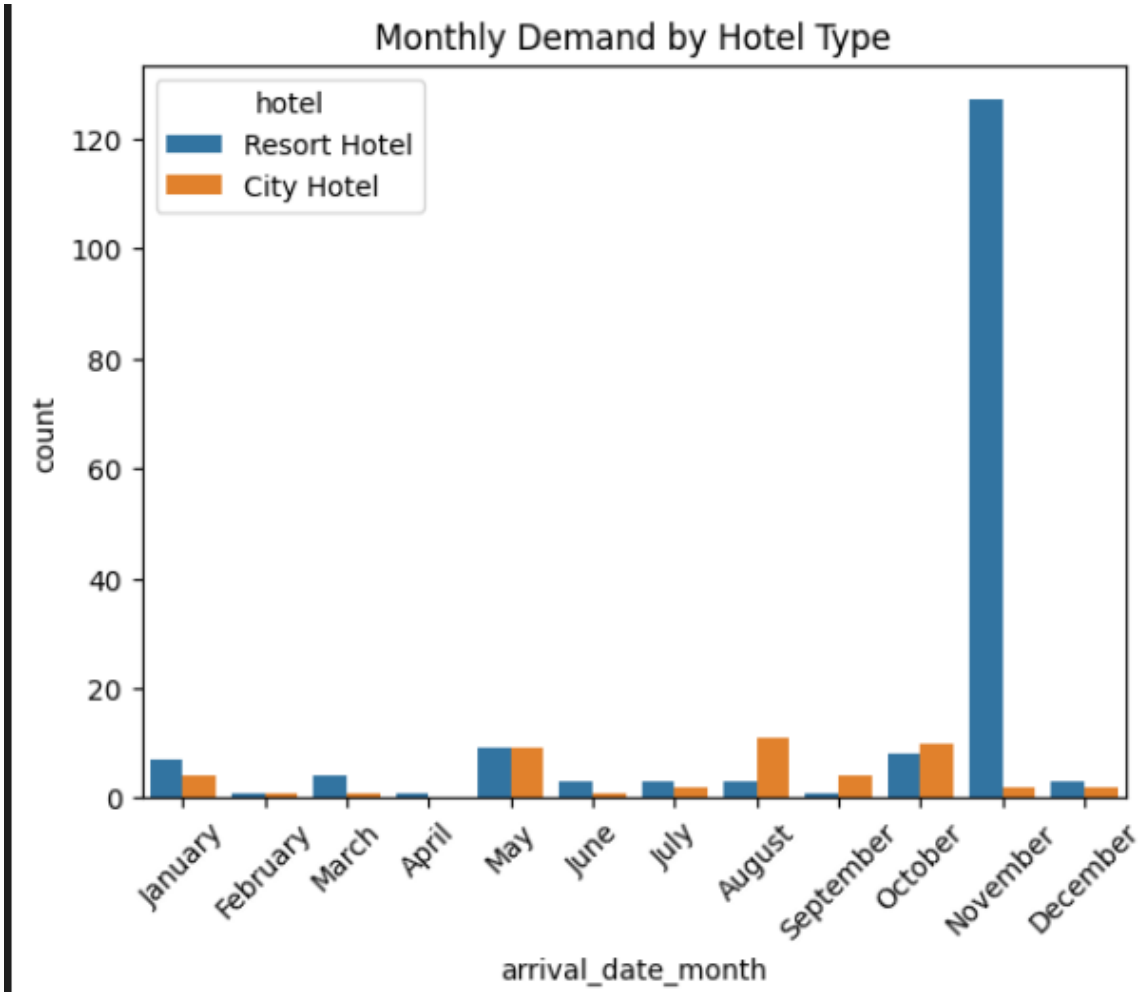
The special requests vs. cancellation chart shows a strong link between customer engagement and reliability. Guests with no special requests have a cancellation rate of ~9%, while those with one or more requests show virtually 0% cancellation. This highlights the importance of encouraging customization during the booking process, as it increases commitment and reduces cancellations—making it a valuable tool for customer relationship management and risk reduction.

Booking pattern

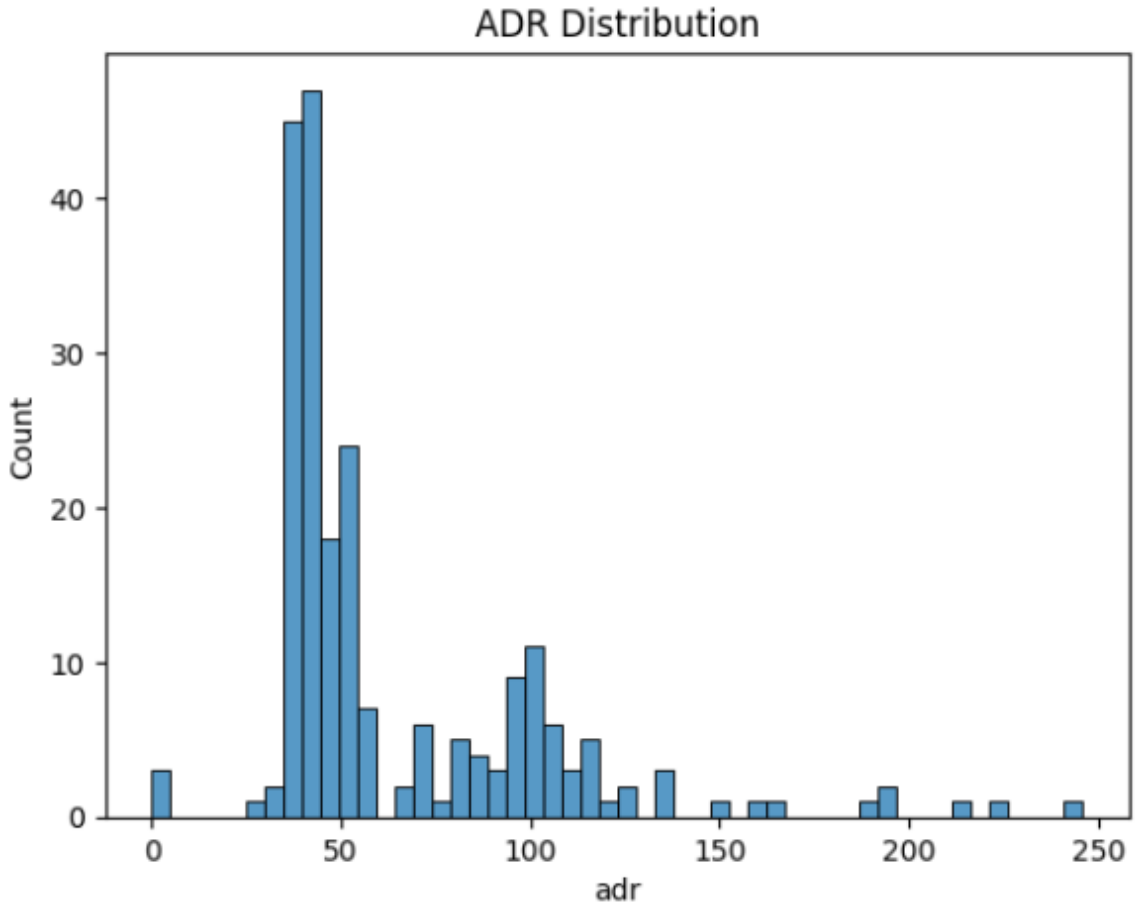
Seasonal Trends and Booking Peaks:
The data reveals a pronounced seasonal pattern, with November as the peak month registering over 130 bookings—far exceeding other months. This indicates a highly seasonal demand, likely driven by tourism, holidays, or events specific to that time. Aside from a minor spike in May (~18 bookings), the rest of the year sees consistently low booking volumes, suggesting a business model heavily reliant on a single high-demand season.



Hotel Preference and Market Focus:
Resort Hotels dominate the booking landscape, especially during peak months, while City Hotels receive minimal bookings year-round. This trend indicates that the business primarily serves leisure travelers seeking resort experiences, rather than business travelers who typically opt for city accommodations. The data emphasizes the importance of tailoring services, marketing, and operations around the resort segment.

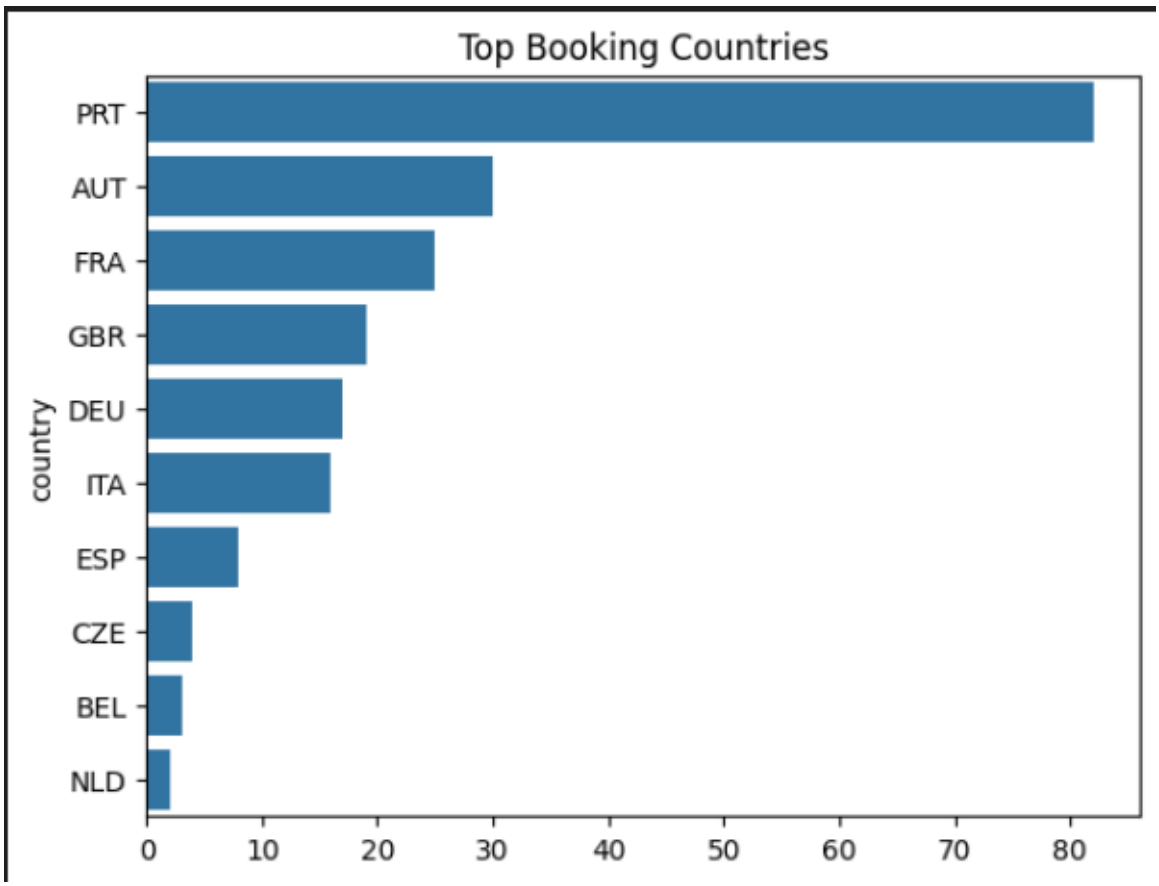


Pricing Patterns and Revenue Strategy:
The ADR (Average Daily Rate) distribution shows that most bookings fall in the 40–60 range, peaking around 45–50, which reflects a mid-range pricing strategy. While a small number of high-rate bookings exist (up to 250), they are infrequent, pointing to limited success in attracting premium customers. This suggests an opportunity to explore premium pricing or upscale packages during peak demand periods like November while maintaining value-focused strategies during low seasons.

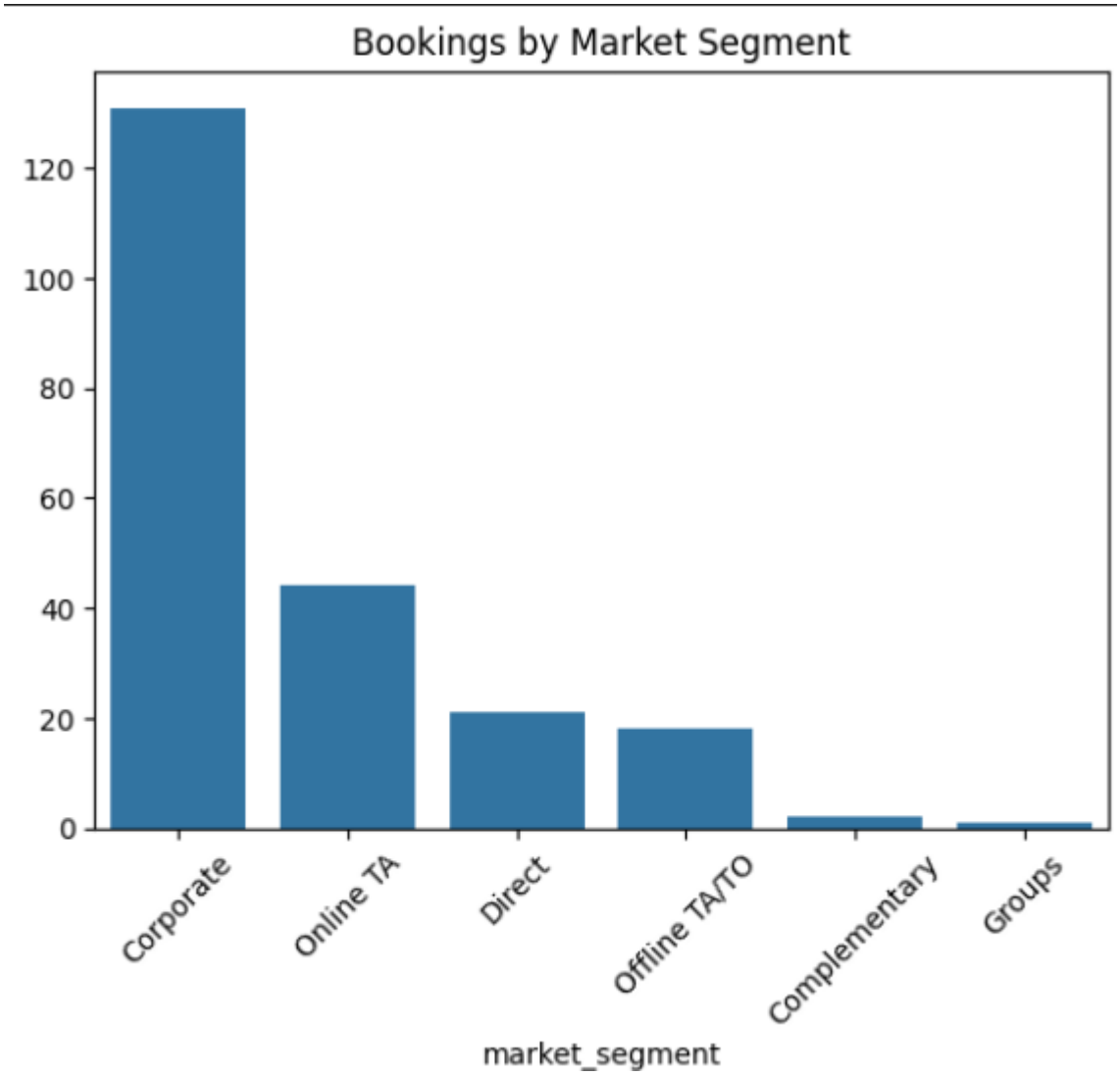


Market Analysis

Portugal is the dominant source market, contributing nearly 40% of total bookings with over 75 reservations, highlighting a strong domestic or local traveler base. Austria and France follow with about 25 bookings each, while Great Britain, Germany, and Italy contribute moderately. The remaining European countries have minimal impact, indicating a highly concentrated European customer base focused mainly on Western and Central Europe.

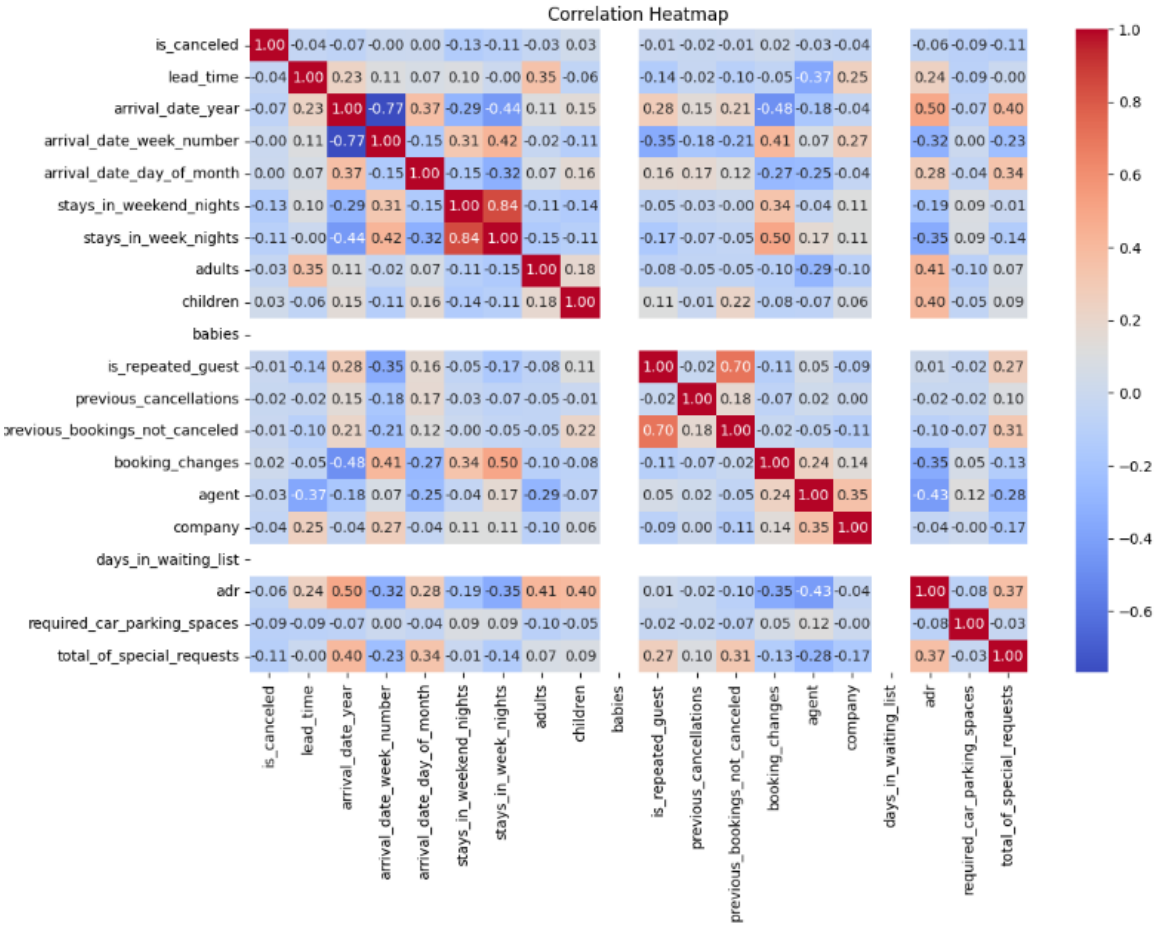


Corporate bookings lead the market segment, making up around 65-70% of total reservations, which suggests a strong business and group event presence even at resort properties. Online Travel Agencies represent the next largest segment, while Direct and Offline Travel Agents have smaller shares. This concentration on corporate and OTA channels reveals opportunities to increase direct bookings for better profitability and suggests geographic diversification—especially beyond Portugal—could reduce risk and improve revenue stability.



Correlation Analysis

The correlation heat map illustrates the linear relationships between multiple variables related to hotel booking patterns using Pearson correlation coefficients. Each coefficient ranges from -1 to +1, where values closer to +1 indicate a strong positive correlation, values closer to -1 indicate a strong negative correlation, and values around 0 suggest little to no linear relationship. The insights derived from this correlation analysis are categorized into distinct thematic areas for better interpretation and strategic application.



1. Stay Duration Relationships

A particularly strong positive correlation (0.84) is observed between weekend and weekday nights, indicating that properties attracting weekend guests also tend to attract weekday guests. This suggests the presence of long-stay guests or consistent guest preference across the week. Furthermore, adult guests are positively correlated with both weekend (0.31) and weekday stays (0.42), while the number of children shows a negative correlation with both weekend (-0.15) and weekday nights (-0.32), suggesting that families tend to book shorter stays.

2. Guest Composition Patterns

Guest composition variables show varying degrees of association. The number of adults is positively correlated with babies (0.18) and special requests (0.07), reflecting more requests from larger or more complex groups. The number of children has a perfect correlation (1.00) with babies, which is expected as families typically travel with both.

Cancellations are less likely in bookings involving children (-0.14), suggesting higher commitment among family-oriented bookings.

3. Cancellation Behavior

Lead time and cancellation have a weak negative correlation (-0.04), indicating minimal direct influence. However, longer lead times correlate moderately with fewer cancellations (-0.44), suggesting that early planners are less likely to cancel. Additionally, weekend stays show a weak negative correlation with cancellations (-0.32), implying that weekend bookings are more stable. Repeat guests also tend to cancel less frequently (-0.14), adding further insight into reliable customer segments.

4. Seasonal and Arrival Patterns

A strong negative correlation is observed between arrival year and arrival week number (-0.77), signifying consistent seasonal booking behaviors across years. This seasonal pattern is further supported by a similar correlation between lead time and week number (-0.77), suggesting that advance bookings are often associated with specific times of the year. The correlation between arrival date year and lead time (0.23) also reinforces the presence of seasonal planning among guests.

5. Lead Time and Booking Behavior

Lead time exhibits several meaningful correlations. It negatively correlates with booking changes (-0.48), indicating that early bookings are more stable. Additionally, lead time and special requests show almost no correlation (-0.00), suggesting that advance planning is not necessarily linked to personalized needs. Interestingly, repeat guests show shorter lead times (-0.14), likely due to familiarity with the property.

6. Booking Channel and Guest Type Insights

Distinct patterns emerge among booking channels. Agent bookings show a negative correlation with company bookings (-0.37), confirming that these are generally exclusive channels. Agent bookings also exhibit shorter lead times (-0.37) and higher average daily rates (ADR) (0.24), likely due to commission structures. In contrast, company bookings display fewer special requests (-0.10) and have a weak positive correlation with weekend stays (0.11), indicating limited customization but potential leisure overlaps.

7. Financial Metrics and Special Requests

ADR (Average Daily Rate) is moderately correlated with lead time (0.24), suggesting that earlier bookings may fetch higher prices. Additionally, ADR correlates positively with special requests (0.37), reflecting a pricing premium for personalized services. Interestingly, weekend stays negatively correlate with ADR (-0.19), which may indicate discounts or lower pricing to drive weekend occupancy. Agent bookings, which yield higher ADRs (0.24), are associated with fewer special requests (-0.28), supporting the theory of standardized package deals.

8. Operational Metrics

Car parking availability shows a slight negative correlation with ADR (-0.08) and special requests (-0.03), suggesting limited financial or behavioral impact. However, special requests increase slightly with group size, as indicated by the adult guest correlation (0.07). While seemingly minor, these insights help inform service planning and resource allocation.