Professional Document: Hotel Booking Analysis Project

Executive Summary

This document presents a comprehensive analysis of a hotel booking dataset, focusing on identifying key trends, customer behaviors, and factors influencing booking cancellations. The insights derived from this analysis aim to provide actionable recommendations for hotel management to optimize operations, enhance customer satisfaction, and ultimately improve revenue. The project involved extensive data preprocessing, exploratory data analysis (EDA), and feature engineering to uncover patterns related to hotel types, guest demographics, booking channels, and seasonal variations.

1. Introduction

The hospitality industry is highly dynamic, with various factors influencing booking patterns and cancellation rates. Understanding these dynamics is crucial for hotels to implement effective strategies for revenue management, operational efficiency, and customer retention. This project leverages a rich dataset of hotel booking information to delve into these aspects, providing a data-driven perspective on hotel performance.

1.1 Project Objectives

The primary objectives of this analysis were to:

- Gain a deep understanding of the hotel booking landscape.
- Identify significant factors contributing to booking cancellations.
- Analyze customer behavior patterns, including repeat guest rates and special requests.
- Determine optimal lengths of stay and their impact on average daily rates (ADR).
- Provide actionable recommendations to improve hotel performance and profitability.

1.2 Data Source and Scope

The dataset used for this analysis is sourced from the article "Hotel Booking Demand Datasets" by Nuno Antonio, Ana Almeida, and Luis Nunes (2019). It encompasses booking information for both a City Hotel and a Resort Hotel, covering a wide range of attributes such as arrival dates, lead times, guest demographics, meal plans, market segments, and reservation statuses. The analysis covers the period represented in the dataset, providing insights into historical booking trends.

2. Data Overview and Preprocessing

Effective data analysis begins with thorough data understanding and preprocessing. This section details the initial inspection of the dataset, the handling of missing values, and the creation of new features to enhance the analytical depth.

2.1 Dataset Description

The raw dataset comprised 119,390 entries and 32 distinct features. Each entry represents a unique hotel booking, with features providing granular details about the booking process and guest characteristics. Key features include:

- `hotel` : Type of hotel (City Hotel or Resort Hotel).
- `is_canceled` : Binary indicator for booking cancellation (1 = Yes, 0 = No).
- `lead_time` : Number of days between booking and arrival.
- `arrival_date_year` , `arrival_date_month` , `arrival_date_week_number` , `arrival_date_day_of_month` : Temporal booking details.
- `stays_in_weekend_nights` , `stays_in_week_nights` : Length of stay components.
- `adults` , `children` , `babies` : Guest demographics.
- `meal` : Type of meal package booked.
- `country` : Country of origin of the guest.
- `market_segment` , `distribution_channel` : Booking source details.
- `is_repeated_guest` : Indicator for repeat customers.
- `previous_cancellations` , `previous_bookings_not_canceled` : Historical booking behavior.
- `reserved_room_type` , `assigned_room_type` : Room type information.
- `booking_changes` : Number of modifications made to the booking.
- `deposit_type` : Type of deposit made.
- `agent` , `company` : IDs of travel agents or companies.
- `days_in_waiting_list` : Days on waiting list before confirmation.
- `customer_type` : Type of customer.
- `adr` : Average Daily Rate.
- `required_car_parking_spaces` , `total_of_special_requests` : Additional guest requirements.
- `reservation_status` , `reservation_status_date` : Final status of the booking.

2.2 Missing Value Treatment

Missing values are a common challenge in real-world datasets and can significantly impact the accuracy and reliability of analysis. A systematic approach was adopted to address these:

- **children** : This column had a negligible number of missing values (4 entries). Given the context, these were imputed with `0` , assuming that if the number of children was not specified, it implied no children were part of the booking.
- **country** : This feature, indicating the guest's country of origin, had 488 missing entries. To preserve the distribution and avoid bias, these missing values were filled using the mode (most frequent country) of the column. This approach is suitable for categorical data where a dominant category exists.
- **agent** **and** **company** : These columns, representing the IDs of travel agents and companies, had a substantial number of missing values (16,340 and 112,593 respectively). Filling these with `NaN` or dropping them could lead to loss of information or misinterpretation. Therefore, missing values in these columns were replaced with the string `'none'` . This treatment allows these entries to be considered as a distinct category, indicating bookings not associated with a specific agent or company, which can be a valuable insight in itself.

2.3 Outlier Identification and Handling

Outliers, or extreme values, can skew statistical analyses and machine learning models. The analysis identified outliers in several numerical features, including `lead_time` , `stays_in_weekend_nights` , `stays_in_week_nights` , `adults` , `previous_cancellations` , `previous_bookings_not_canceled` , `booking_changes` , `days_in_waiting_list` , and `adr` . For instance, the `adults` column showed 26.06% outliers, and `booking_changes` had 18.12% outliers. While the specific method of outlier treatment (e.g., capping, transformation, or removal) is not explicitly detailed in the provided notebook, the recognition and quantification of these outliers are crucial steps in ensuring data quality and model robustness. The decision to handle or retain outliers depends on their nature and potential impact on the analysis objectives.

2.4 Feature Engineering

To derive more meaningful insights and create features that better represent the underlying patterns, several new features were engineered:

- **total_guests** : This feature was created by summing the `adults` , `children` , and `babies` columns. This provides a consolidated view of the total number of individuals in a booking, which can be a strong predictor for various outcomes, including room type requirements and resource allocation.

- **total_nights** : Calculated as the sum of stays_in_weekend_nights and stays_in_week_nights , this feature represents the total duration of a guest's stay. This is a critical metric for understanding booking duration preferences and their correlation with cancellation rates or ADR.
- **is_family** : A binary indicator (1 or 0) was created to identify bookings that include children or babies. This feature helps in segmenting customer types and understanding the specific needs and behaviors of family travelers, which can inform targeted marketing and service offerings.

These engineered features provide a more holistic view of the booking characteristics, enabling deeper analysis and more precise recommendations.

3. Key Insights and Visualizations

This section presents the core findings from the exploratory data analysis, supported by various visualizations that highlight significant trends and relationships within the dataset.

3.1 Hotel Type Performance

Analysis of booking volumes by hotel type revealed a consistent trend: the City Hotel consistently outperformed the Resort Hotel in terms of the number of bookings across all years covered in the dataset. This suggests a higher demand for urban accommodations, possibly due to business travel, city tourism, or better accessibility. Understanding this disparity is vital for resource allocation and marketing strategies for each hotel type.

3.2 Cancellation Behavior Analysis

Booking cancellations represent a significant challenge for hotels, leading to lost revenue and operational inefficiencies. The analysis revealed several critical insights into cancellation behavior:

- **Overall Cancellation Rate**: The dataset showed a substantial overall cancellation rate of approximately 37%. This high percentage underscores the importance of developing strategies to mitigate cancellations.
- **Hotel Type Impact**: City Hotels exhibited a notably higher cancellation rate (around 41.7%) compared to Resort Hotels (approximately 27.8%). This difference could be attributed to various factors, such as the nature of travel (business vs. leisure), booking flexibility, or target demographics for each hotel type.
- **Deposit Type Influence**: Bookings made with a 'Non Refund' deposit type showed an exceptionally high cancellation rate. This counter-intuitive finding suggests that guests who commit to non-refundable bookings might be more prone to last-minute changes or cancellations due to unforeseen circumstances, or perhaps they are less committed

to the booking in the first place, relying on the non-refundable nature to deter them from frivolous bookings. Further investigation into the reasons behind these cancellations is warranted.

- **Lead Time Correlation**: A strong positive correlation was observed between lead time (days between booking and arrival) and cancellation rates. Longer lead times were associated with higher probabilities of cancellation. This is logical, as guests booking far in advance have more time for plans to change or for alternative options to emerge.
- **Geographical Impact**: Guests from Portugal (PRT) demonstrated the highest cancellation rate. This geographical insight is crucial for targeted marketing and cancellation prevention strategies, potentially involving localized offers or communication.

3.3 Guest Behavior Patterns

Understanding guest behavior is fundamental to tailoring services and marketing efforts:

- **Repeat vs. New Guests**: The vast majority of bookings were made by new guests, with repeated guests constituting only a small percentage of the total. This highlights an opportunity for hotels to focus on loyalty programs and incentives to encourage repeat business.
- **Customer Type Distribution**: The 'Transient' customer type accounted for the largest share of bookings. This category typically refers to individual travelers without a specific contract or group affiliation, indicating a need for flexible booking options and efficient individual guest services.
- **Special Requests and Cancellations**: A compelling insight was the observation that bookings with special requests had a lower cancellation rate. This suggests that guests who take the time to make special requests are more committed to their stay. Furthermore, fulfilling these requests likely enhances guest satisfaction, further reducing the likelihood of cancellation. This finding emphasizes the importance of accommodating guest preferences.

3.4 Average Daily Rate (ADR) Dynamics

Average Daily Rate (ADR) is a key performance indicator for hotels. The analysis revealed that ADR fluctuates significantly based on the month of arrival and the type of hotel. Generally, ADR tends to be higher during peak seasons, such as the summer months, reflecting increased demand. This seasonality in pricing is a critical factor for revenue management, allowing hotels to optimize their pricing strategies to maximize income during high-demand periods and attract guests during off-peak times.

3.5 Optimal Length of Stay

The analysis of stay durations indicated that the most common length of stay for guests was between 1 and 3 nights. This short-stay preference is typical for business travelers or short leisure trips. Conversely, longer stays, particularly those extending to 7 nights or more, were more prevalent in Resort Hotels. This aligns with the nature of resort vacations, which often involve extended leisure periods. These insights can help hotels tailor packages and promotions based on typical stay durations for different hotel types and customer segments.

4. Methodologies Applied

The project employed a standard data science methodology, encompassing data loading, preprocessing, exploratory data analysis, and feature engineering. The tools and techniques utilized ensured a robust and insightful analysis.

4.1 Data Loading and Initial Inspection

The dataset was loaded using the pandas library, a fundamental tool for data manipulation in Python. Initial inspection involved:

- `df.info()` : To understand data types, non-null counts, and memory usage, providing a quick overview of data completeness.
- `df.describe()` : To generate descriptive statistics for numerical columns, including count, mean, standard deviation, min, max, and quartiles, which helped in identifying potential outliers and data distributions.
- `df.isna().sum()` : To quantify missing values for each column, guiding the missing value imputation strategy.
- `df.duplicated().sum()` : To identify and count duplicate rows, ensuring that each booking record was unique and preventing skewed analysis results.

4.2 Data Cleaning

Data cleaning was a crucial step to ensure the quality and reliability of the dataset. This involved:

- **Handling Missing Values**: As detailed in Section 2.2, specific strategies were applied to `children` , `country` , `agent` , and `company` columns to address missing data effectively.
- **Removing Duplicates**: Duplicate rows were identified and removed using `df.drop_duplicates(inplace=True)` . This step is essential to prevent overrepresentation of certain booking patterns and ensure that each observation contributes uniquely to the analysis.

4.3 Feature Engineering

Feature engineering played a vital role in transforming raw data into more informative attributes. The creation of `total_guests`, `total_nights`, and `is_family` (as discussed in Section 2.4) allowed for a more nuanced understanding of booking characteristics and customer segments. These features are particularly valuable for building predictive models or for more granular segmentation of guests.

4.4 Exploratory Data Analysis (EDA) and Visualization

EDA was performed using a combination of statistical methods and data visualization techniques. The primary libraries used were `matplotlib.pyplot`, `seaborn`, and `plotly.express`:

- **`sns.countplot`**: Used to visualize the distribution of categorical variables, such as `is_repeated_guest` and `hotel` types, providing quick insights into the frequency of different categories.
- **`px.bar`**: Utilized for creating interactive bar charts, particularly effective for comparing metrics across different categories or time periods, such as booking volumes by month or cancellation rates by deposit type.
- **Line Plots**: Employed to illustrate trends over time, such as the number of bookings per month or the variation in ADR across different periods. These plots helped in identifying seasonality and long-term patterns.

These visualizations were instrumental in uncovering the key insights presented in Section 3, making complex data patterns easily interpretable.

5. Recommendations for Improving Hotel Performance

Based on the comprehensive analysis of the hotel booking dataset, the following strategic recommendations are proposed to enhance hotel performance, optimize revenue, and improve guest satisfaction.

5.1 Cancellation Management Strategies

The high overall cancellation rate, particularly for City Hotels and bookings with long lead times, necessitates proactive management:

- **Tiered Cancellation Policies**: Implement dynamic cancellation policies based on lead time. For bookings made far in advance, consider offering a tiered non-refundable option with varying discounts. For example, a slightly lower non-refundable rate for very long lead times, increasing closer to the arrival date. This can help secure revenue while still offering some flexibility.

- **Incentivize Non-Cancellation**: Introduce small incentives for guests who maintain their bookings, especially those with long lead times. This could include a complimentary welcome drink, a minor room upgrade, or a discount on in-hotel services. Such gestures can significantly improve guest commitment and reduce last-minute cancellations.
- **Targeted Communication**: For bookings with high cancellation risk (e.g., long lead times, specific market segments), initiate proactive communication. This could involve sending personalized reminders, offering flexible rebooking options, or confirming special requests to reinforce the value of the booking.
- **Review Non-Refundable Deposit Policy**: The observation of high cancellation rates for 'Non Refund' deposit types warrants a re-evaluation of this policy. It might be counterproductive if guests are booking these rates without full commitment, leading to administrative overhead and lost opportunities. Consider adjusting the terms or offering more attractive refundable alternatives.

5.2 Revenue Optimization through Dynamic Pricing

Leveraging the insights on ADR variations and seasonality is crucial for maximizing revenue:

- **Dynamic Pricing Models**: Implement sophisticated dynamic pricing models that adjust room rates based on real-time demand, seasonality, competitor pricing, and historical booking patterns. This ensures that hotels can capture maximum revenue during peak demand periods and remain competitive during off-peak times.
- **Seasonal Promotions**: Develop targeted promotional campaigns for off-peak seasons. This could include special packages, discounts for extended stays, or partnerships with local attractions to stimulate demand when occupancy is typically lower. For Resort Hotels, promoting longer stays during off-peak times could be particularly effective.
- **Event-Based Pricing**: Monitor local events, conferences, and holidays that might influence demand. Adjust pricing proactively for these periods to capitalize on increased booking interest.

5.3 Enhancing Guest Loyalty and Retention

Given the low percentage of repeated guests, fostering loyalty is a significant opportunity:

- **Robust Loyalty Programs**: Develop and actively promote a comprehensive loyalty program that rewards repeat guests with exclusive benefits, discounts, and personalized experiences. Tiers based on stay frequency or spending can encourage guests to climb the loyalty ladder.
- **Personalized Guest Experiences**: Utilize data on guest preferences (e.g., special requests, room types) to offer personalized experiences. Acknowledging and fulfilling

special requests not only reduces cancellations but also significantly enhances guest satisfaction, leading to positive reviews and repeat bookings.

- **Post-Stay Engagement**: Implement a post-stay engagement strategy, including personalized thank-you notes, feedback surveys, and exclusive offers for future stays. This maintains a connection with guests and encourages them to choose the hotel again.

5.4 Strategic Marketing and Channel Management

Optimizing marketing efforts and distribution channels can drive more profitable bookings:

- **Geographic Targeting**: Focus marketing campaigns on countries with historically lower cancellation rates and higher booking volumes. For regions with high cancellation rates (e.g., Portugal), consider tailored campaigns that address specific concerns or offer more flexible booking terms.
- **Channel Performance Analysis**: Continuously analyze the performance of different market segments and distribution channels. Allocate marketing budgets and resources to the most profitable channels, while identifying areas for improvement or divestment in underperforming ones.
- **Online Travel Agent (OTA) Optimization**: While OTAs are important for reach, analyze their profitability. Negotiate favorable terms and consider strategies to drive direct bookings, which typically have lower acquisition costs.

5.5 Operational Efficiency and Guest Satisfaction

Streamlining operations can directly impact guest satisfaction and reduce negative experiences:

- **Waiting List Management**: Implement an efficient system for managing waiting lists, ensuring timely communication with guests and prompt confirmation once rooms become available. Minimizing waiting times can prevent guests from seeking alternative accommodations.
- **Room Assignment Accuracy**: Regularly review discrepancies between `reserved_room_type` and `assigned_room_type`. Strive to assign the reserved room type whenever possible, as deviations can lead to guest dissatisfaction. If changes are necessary, communicate proactively and offer suitable alternatives or compensation.
- **Staff Training**: Ensure staff are well-trained in handling special requests, managing cancellations, and providing exceptional customer service. Empower front-line staff to resolve issues quickly and efficiently.

6. Conclusion

This hotel booking analysis project has provided valuable insights into various aspects of hotel operations, from booking trends and cancellation behaviors to guest preferences and revenue dynamics. By understanding the factors that influence booking decisions and cancellations, hotels can implement data-driven strategies to improve their performance. The recommendations outlined in this document, focusing on cancellation management, revenue optimization, guest loyalty, strategic marketing, and operational efficiency, offer a roadmap for hotels to enhance profitability and deliver superior guest experiences. Continuous monitoring of these metrics and adaptation of strategies based on evolving market conditions will be key to sustained success in the competitive hospitality industry.

Author: Manus AI

Date: September 05, 2025