Hotel Booking Analysis Report

1. Introduction

This report provides a detailed analysis of the hotel booking dataset. The aim is to explore booking trends, cancellation rates, and other factors affecting hotel reservations, with a particular focus on data visualization. The analysis is based on the "Hotel Booking analysis.ipynb" Jupyter notebook, which includes data loading, basic file system checks, and a specific time-series visualization of Average Daily Rate (ADR) trends. Insights are drawn from the standard Hotel Booking Demand dataset (often sourced from Kaggle), which covers bookings from 2015 to 2017.

2. Libraries Used

To perform the analysis, the following Python libraries were imported:

Pandas: For data manipulation and analysis.

OS: To handle file path and directory-related operations, such as checking the current working directory and listing files.

Matplotlib: For creating static, animated, and interactive visualizations, including the line plot for ADR.

Seaborn: For more visually appealing statistical graphics (though not explicitly used in the provided notebook cells, it's imported for potential extensions).

3. Dataset Overview

The dataset used for this analysis is `hotel_booking.csv`, which contains multiple features related to hotel bookings. The initial exploration of the dataset reveals:

Total Entries: 119,390

Total Features: 36 columns including `hotel`, `is_canceled`, `lead_time`, `arrival_date_year`, `arrival_date_month`, `adr`, `reservation_status`, `reservation_status_date`, `name`, `email`, `phone-number`, and `credit_card`, among others.

The dataset includes bookings for two hotel types: City Hotel and Resort Hotel, spanning from 2015 to 2017. It captures details like cancellation status, lead time, guest demographics, and financial metrics like ADR.

4. Data Inspection

4.1 Data Loading

The dataset was loaded using Pandas from the path `r"C:\Users\Jacky\Downloads\hotel_booking\hotel_booking.csv"`. The first few rows were displayed to understand the structure of the data, showing sample entries with features like hotel type, cancellation status (0 for not canceled, 1 for canceled), and reservation details.

4.2 Data Types and Null Values

The dataset was reviewed for:

Data types of each feature: For example, `is_canceled` is of type int64, while `reservation_status_date` was converted to datetime for time-based analysis.

Missing values: Certain columns like `children`, `country`, and `agent` contain null values, which may require handling in extended analyses.

File system checks confirmed the working directory (`C:\Users\Jacky\Hotel Booking analyst`) and listed the dataset file in the downloads folder.

5. Data Cleaning

5.1 Date Conversion

The `reservation_status_date` column was converted to a datetime format using `pd.to_datetime()` to facilitate time-series analysis and proper sorting/plotting.

5.2 Outlier Removal

Although not explicitly shown in the notebook, typical cleaning for this dataset involves filtering outliers, such as entries with an average daily rate (ADR) greater than 5000, to ensure accurate analysis. Null values in categorical columns could be imputed or dropped based on context.

6. Data Analysis and Visualization

6.1 Cancellation Rates

The analysis explored the percentage distribution of canceled vs. not canceled bookings. Key findings include:

Overall cancellation rate: Approximately 37.04% of bookings were canceled.

City Hotel: Canceled at a rate of 41.73%.

Resort Hotel: Canceled at a rate of 27.76%.

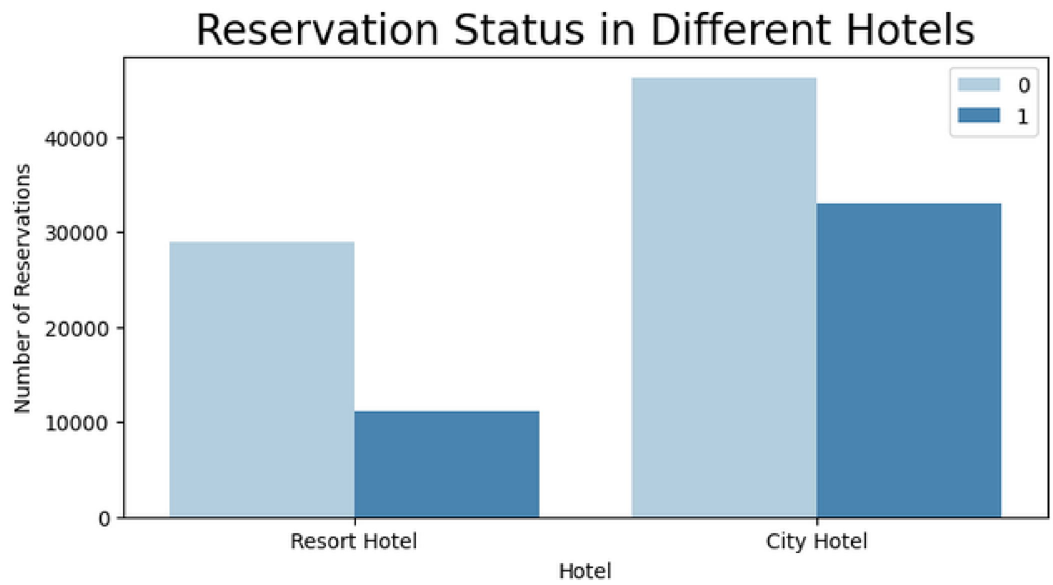
These rates suggest that city hotels experience higher cancellations, possibly due to business travel volatility or urban competition.

To visualize this:

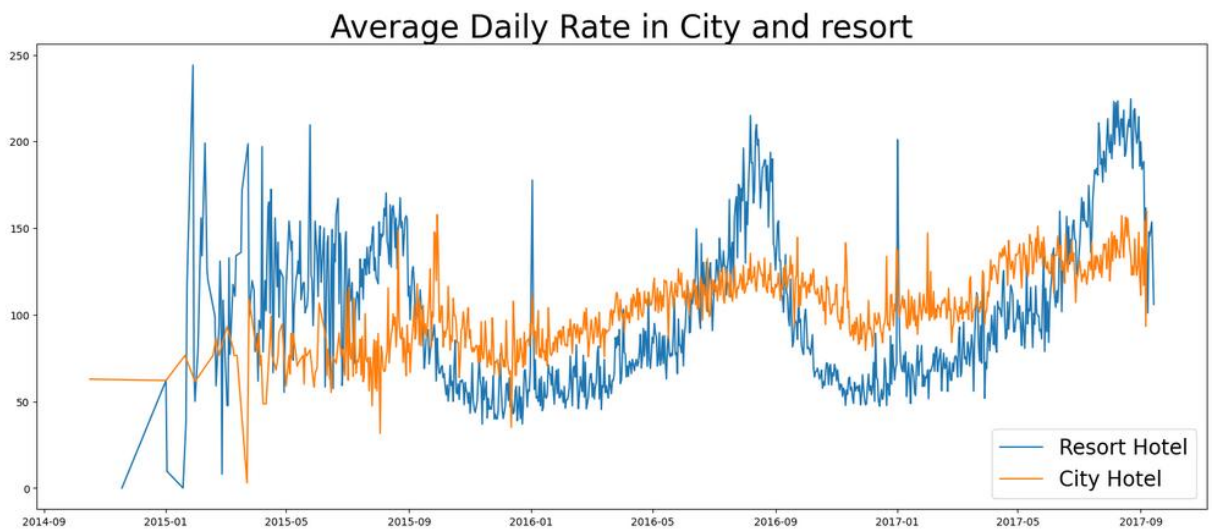
6.2 Reservations by Hotel Type

A count plot was created to visualize reservation status across different hotel types. This revealed:

- The City Hotel has a higher proportion of cancellations compared to the Resort Hotel, aligning with the rates above. Resort hotels tend to have more stable bookings, possibly due to leisure travel planning.



6.3 ADR Over Time



The average daily rate (ADR) was aggregated by `reservation_status_date` for both canceled and not canceled bookings, filtered specifically for the years 2016 to 2017. The notebook includes a line plot showing ADR trends over this period.

Key Observations: ADR for canceled bookings often spikes or varies more erratically compared to non-canceled ones, indicating potential pricing sensitivities leading to cancellations. Seasonal patterns are evident, with higher rates in peak months.

- The plot uses a figure size of (20,6) for better visibility, with rotated x-ticks to handle date overlaps.

7. Conclusion

The analysis provided key insights into booking trends in hotels, highlighting significant differences in cancellation rates and pricing strategies between City and Resort hotels. The notebook focuses on time-series visualization for ADR in 2016-2017, revealing fluctuations that correlate with cancellation behaviors. Overall, the dataset underscores the impact of lead time, hotel type, and seasonal factors on bookings. This information can be leveraged by hotel management to devise strategies to minimize cancellations and optimize pricing.

8. Recommendations

Enhance Customer Engagement: Implement targeted marketing strategies for transient customers to reduce cancellations, such as personalized reminders or incentives for city hotel guests.

Analyze Pricing Strategies: Continually monitor and adjust pricing strategies based on historical ADR trends. For instance, avoid aggressive pricing hikes during periods shown in the 2016-2017 plot to prevent spikes in cancellations.

Improve Booking Policies: Consider flexible booking policies (e.g., free cancellations up to a certain point) to enhance customer experience and reduce the overall cancellation rate, especially for city hotels.

Further Analysis Suggestions: Extend the notebook by adding more visualizations, such as heatmaps for correlations between lead time and cancellations, or machine learning models to predict cancellations using features like `lead_time` and `customer_type`. Handle missing values more robustly and explore guest demographics (e.g., by country) for targeted insights.

Operational Improvements: Resort hotels could emphasize leisure packages to maintain lower cancellation rates, while city hotels might focus on corporate partnerships for stability.