Enhancing Hotel Revenue Management through Predictive Analytics

(Technical Report)

(Model masters)

Abstract

This report presents a data-driven analysis aimed at enhancing revenue management for Hotel A by investigating booking cancellations and no-shows. Utilizing a dataset of reservation details, customer demographics, and booking information, we employ exploratory data analysis (EDA) to identify key factors influencing reservation outcomes and quantify associated revenue loss. The analysis reveals relationships between booking characteristics, customer demographics, and reservation status, suggesting opportunities for targeted revenue management strategies. The report concludes with recommendations for Hotel A and suggestions for further investigation.

i. Table of content

ii. List of Figueres

1. Introduction

In the highly competitive hotel industry, effective revenue management is critical for maximizing profitability and ensuring optimal resource allocation. Hotels, a prominent chain with diverse property types including Airport Hotels, Resorts, and City Hotels, faces significant challenges due to unpredictable booking cancellations and no-shows. These events lead to substantial revenue loss from vacant rooms, inefficient resource utilization, and potential guest dissatisfaction.

1.1. Background of the Study

The hospitality sector operates in a dynamic environment where booking behaviors are influenced by numerous factors including hotel type, customer demographics, pricing strategies, and booking channels. Hotels seeks to leverage data-driven analytics to gain deeper insights into these relationships and develop predictive capabilities for reservation outcomes. Initial data exploration reveals important considerations, including instances where booking dates occur after expected check-in dates (as seen in rows 6, 18, 54, 56-58, 81-82, and 85), suggesting data quality issues requiring careful preprocessing. The dataset encompasses diverse customer segments across various demographic attributes such as ethnicity, income levels ranging from under $25,000 to over $100,000, and different educational backgrounds from mid-school to graduate level education.

Objectives

This study aims to develop an analytical solution that enhances Hotels's revenue management capabilities through the following objectives:

1. Conducting comprehensive exploratory data analysis to identify key factors influencing booking cancellations and no-shows, and quantifying associated revenue loss.
2. Developing a predictive model to accurately classify bookings as "Cancellation," "No-Show," or "Check-Out," comparing various machine learning models using the F1 score.
3. Creating a customer segmentation model to identify distinct guest profiles based on cancellation behavior.
4. Ensuring model explainability by providing clear insights into feature importance for actionable hotel management strategies.

1.2. Scope

This analysis utilizes the provided hotel booking dataset, which includes attributes related to reservation details, customer demographics, booking channels, and payment methods. The analytical techniques employed encompass statistical analysis, data visualization, machine learning classification, and clustering algorithms. The analysis specifically addresses unique cancellation patterns across Hotel's three property types while considering geographical and temporal dimensions.

1.3. Limitations

Several limitations impact this analysis. Data quality issues, including date inconsistencies and potential missing values, may affect findings. While the analysis identifies correlations between variables and booking outcomes, it may not establish definitive causation. Results may be specific to hotels and not directly generalizable to other hotel chains or regions. Additionally, external factors not captured in the dataset, such as economic conditions, seasonal variations, or unforeseen events, could influence booking cancellations but remain outside the scope of this analysis.

2. Methodology

The dataset utilized in this analysis comprises historical booking data from hotels, encompassing 27,500 reservation records with corresponding outcomes. This dataset was provided as part of the case study challenge to develop predictive solutions for the hotel's revenue management.

To select the best model for predicting booking outcomes (cancellation, no-show, or check-in), we experimented with several algorithms, including Logistic Regression, XGBoost, and SVM. Model performance was evaluated primarily using the F1-score, chosen to balance precision and recall given the potential for class imbalance. To ensure robust and generalizable results, we employed k-fold cross-validation.

For customer segmentation based on cancellation behavior, we used unsupervised learning methods such as K-Means and hierarchical clustering. Feature selection prioritized both behavioral (cancellation history) and demographic (age, income) attributes to create comprehensive customer profiles. The resulting clusters were then analyzed to identify distinguishing characteristics and derive actionable insights for operational improvements, such as targeted marketing and personalized service strategies.

The dataset contains comprehensive information about each booking, including reservation details, guest demographics, booking preferences, and final reservation status. Each record represents a unique booking identified by a "Reservation-id" and includes the following attributes:

2.1. Data Collection

2.1.1. Demographic Information

- Gender: Gender of the person making the reservation
- Age: Age of the reservation holder (ranging from 18 to 70 years)
- Ethnicity: Categorized as Latino, African American, Asian American, or Caucasian
- Educational Level: Highest education attained (Mid-School, High-School, College, or Grad)
- Income: Income bracket of the reservation holder (<25K, 25K-50K, 50K-100K, >100K)
- Country region: Geographic region of origin (North, South, East, West)

2.1.2. Reservation Details

- Hotel Type: Category of hotel property (City Hotels, Airport Hotels, Resorts)
- Expected checking: Anticipated date of arrival
- Expected checkout: Planned departure date
- Booking date: Date when the reservation was made
- Adults: Number of adults included in the reservation
- Children: Number of children included in the reservation
- Babies: Number of infants included in the reservation

2.1.3. Booking Preferences

- Meal Type: Selected meal plan (BB - Bed and Breakfast, HB - Half-board, FB - Full Board)

- Previous Cancellations: Indicator of whether the guest has previously canceled reservations
- Deposit type: Payment security method (No Deposit, Refundable, Non-Refundable)
- Booking channel: Reservation method (Direct, Online, Agent)
- Required Car Parking: Whether parking space was requested
- Use Promotion: Indicates if promotional offers were applied
- Discount Rate: Percentage discount applied, if any

2.1.4. Outcome and Pricing

- Reservation Status: Final status of the booking (1 - check-in, 2 - cancellation, 3 - no-show)
- Room Rate: Nightly rate charged for the room

2.2. Data Preparation

The initial phase involved preparing the raw dataset for analysis. This included loading the data into a Pandas Data Frame and conducting an initial inspection to identify potential data quality issues. The inspection revealed several areas requiring attention, including missing values primarily in the Babies column, the need for new feature creation, inconsistent date entries, negative values in the booking lead column, and duplicate entries in Reservation-id. To ensure data integrity and suitability for analysis, several corrective measures were applied.

2.2.1. Handling Missing Values

We addressed missing values, focusing primarily on the Babies column. Missing numerical values were imputed with column medians. Recognizing that a missing Babies entry likely signified the absence of infants, we replaced all missing values with 0.

2.2.2. Feature Engineering

Basic Features
To derive meaningful insights from the dataset, we created several new features:

- **Stay days:** Calculated as the difference in days between the Expected checkout and Expected checking dates. This represents the intended length of stay.
- **booking lead:** Represents the number of days between the Booking date and the Expected checking date, reflecting the booking lead time.
- **Total guests:** Calculated as the sum of Adults, Children, and Babies. This feature provides a total guest count for each reservation.

Aggregate Features

Previous cancellation frequency: Calculated to get previous cancellation history and used to measure guest behavior.

2.2.3. Data Transformation

- **Encoding Categorical Variables:** Since machine learning algorithms cannot directly process categorical data, we employed appropriate encoding techniques to convert these variables into numerical representations. Specifically, we used one-hot encoding for nominal categorical features (Ethnicity, Hotel Type, Country region) and label encoding for ordinal categorical features (Educational Level).

2.2.4. Handling Data Inconsistencies and Data Cleaning

We identified and addressed inconsistencies and cleaned the data within the dataset:

- **Negative booking lead Values:** Reservations exhibiting negative booking lead values, indicating illogical booking dates, were removed from the dataset to maintain data integrity.
- **Duplicate Reservation-id Entries:** We identified and removed duplicate entries based on the Reservation-id column to ensure each reservation is uniquely represented.

2.2.5. Data Splitting

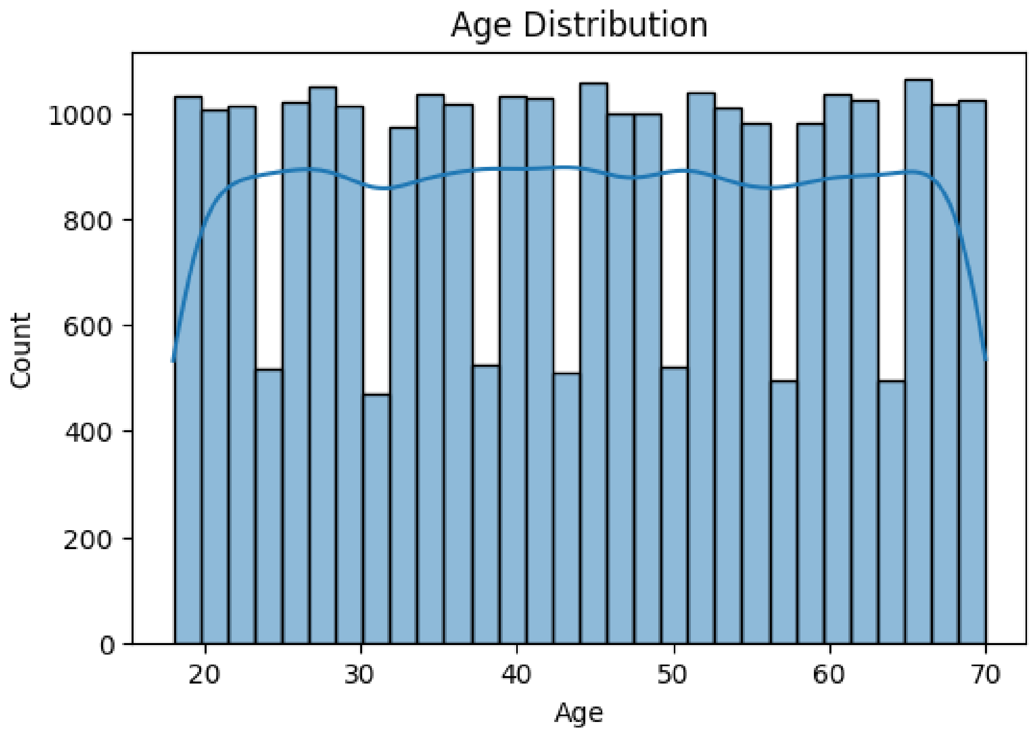
To ensure unbiased model evaluation, we split the dataset into three subsets:

- **Training Set:** Used to train the machine learning models.
- **Validation Set:** Used to tune model hyperparameters and assess performance during training.
- **Testing Set:** Used to evaluate the final performance of the trained model on unseen data.

By completing these preparation steps, including the careful handling of the Babies column and the removal of problematic data points, the dataset was transformed into a clean, consistent, and enriched format, ready for subsequent exploratory analysis and predictive modeling. This careful preparation ensures the reliability and accuracy of our results, ultimately contributing to actionable insights for hotels.

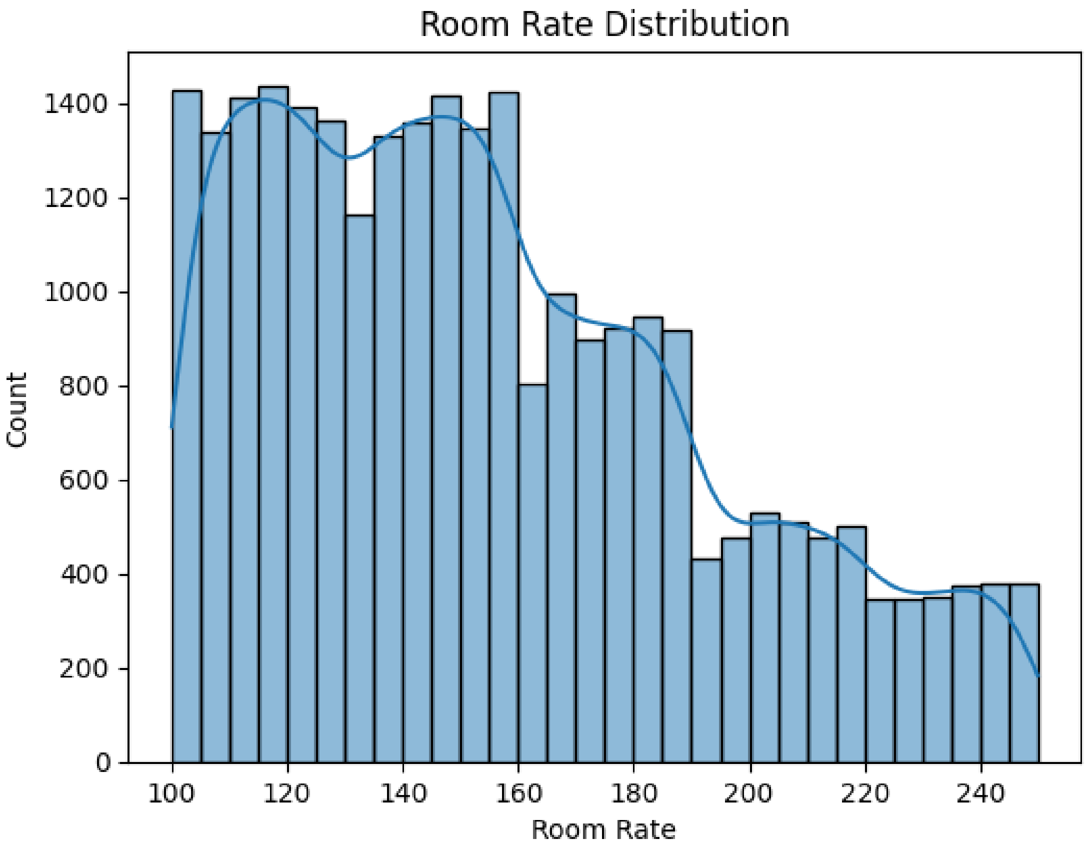
3. Descriptive analysis

3.1 Age distribution



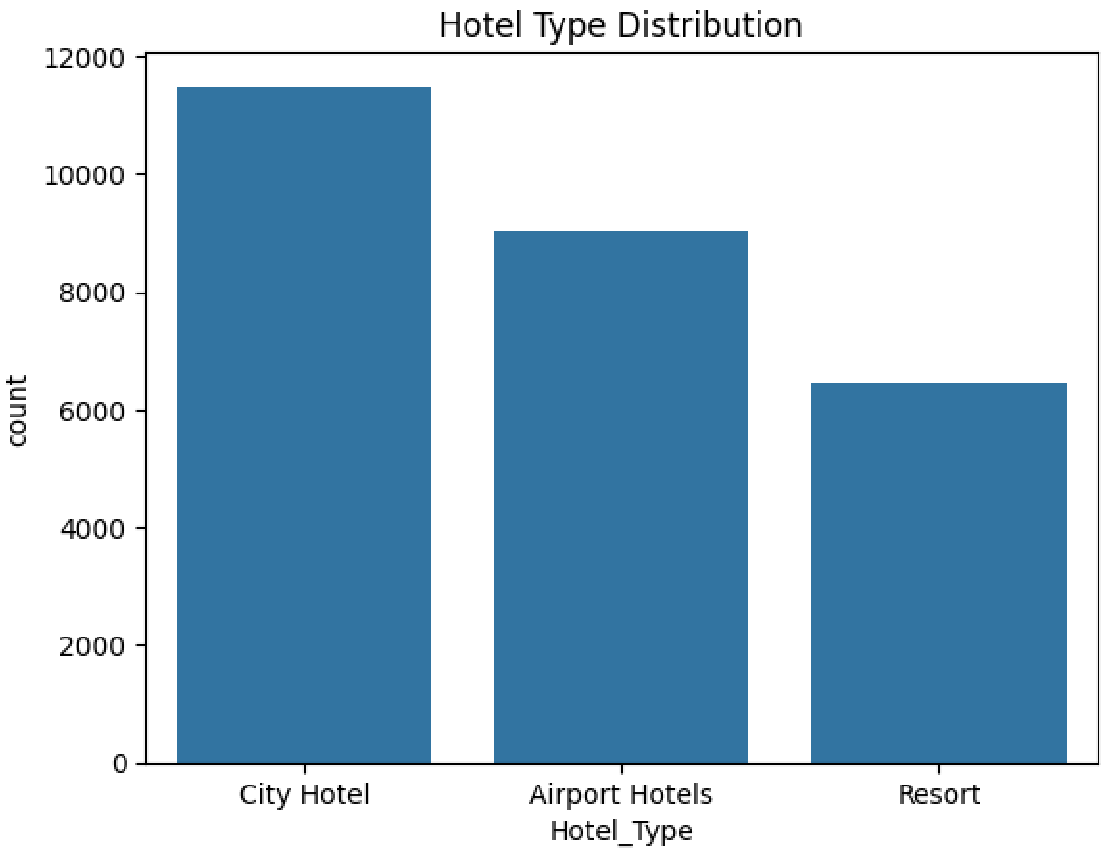
The age distribution of hotel guests is relatively uniform across the range of 18 to 70, indicating a diverse customer base with no single dominant age demographic. This suggests that while age alone may not strongly predict cancellations, it remains a valuable factor when combined with other variables like income and hotel type to inform marketing and service strategies.

3.2. Room rate distribution



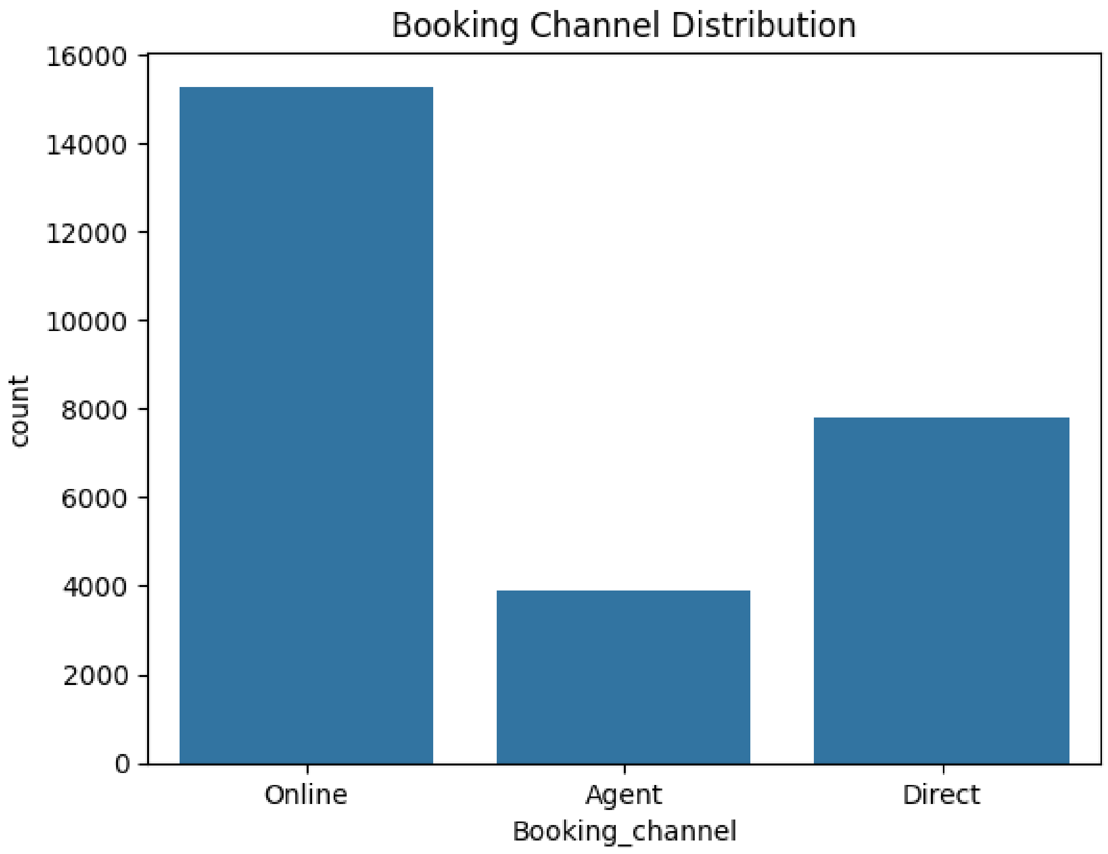
The room rate distribution shows a right-skewed, multimodal pattern, indicating a tiered pricing strategy with distinct room categories, which can inform revenue management and pricing optimization strategies.

3.3. Hotel type distribution



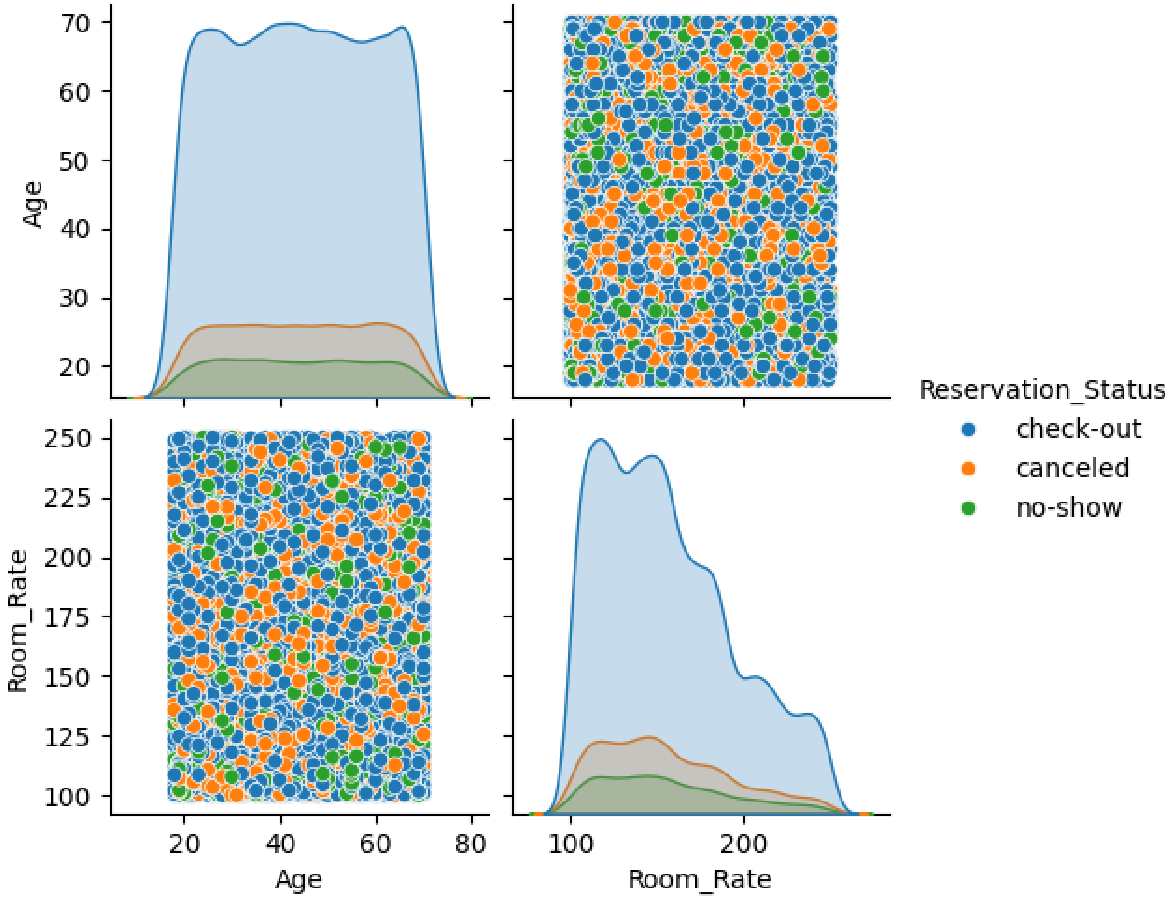
The Hotel Type Distribution bar chart reveals that City Hotels have the highest number of bookings, followed by Airport Hotels, and then Resorts, providing a clear picture of the relative proportions of each hotel type within the dataset. This distribution indicates that City Hotels account for the largest share of bookings, while Resorts represent the smallest proportion.

3.4 Booking channel distribution



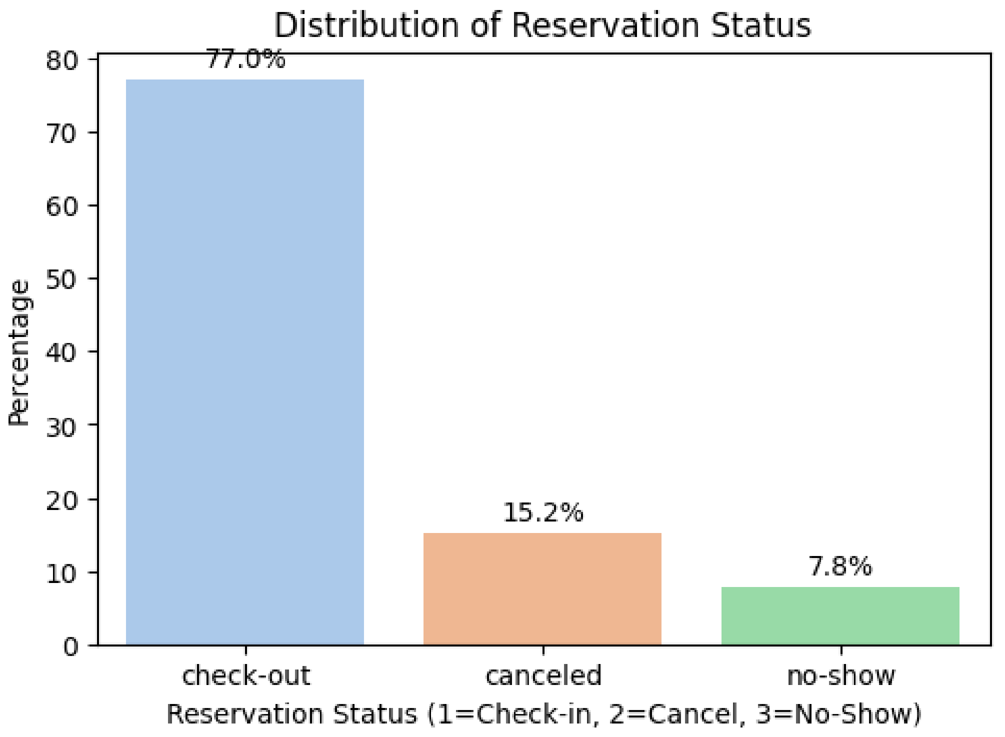
The Booking Channel Distribution bar chart shows that the "Online" channel has the highest number of bookings, followed by "Direct" and then "Agent" bookings, providing a clear comparison of booking counts across these channels. This distribution indicates that Online bookings dominate the dataset, with Direct bookings being twice as frequent as Agent bookings.

3.5 distributions of Age and Room Rate



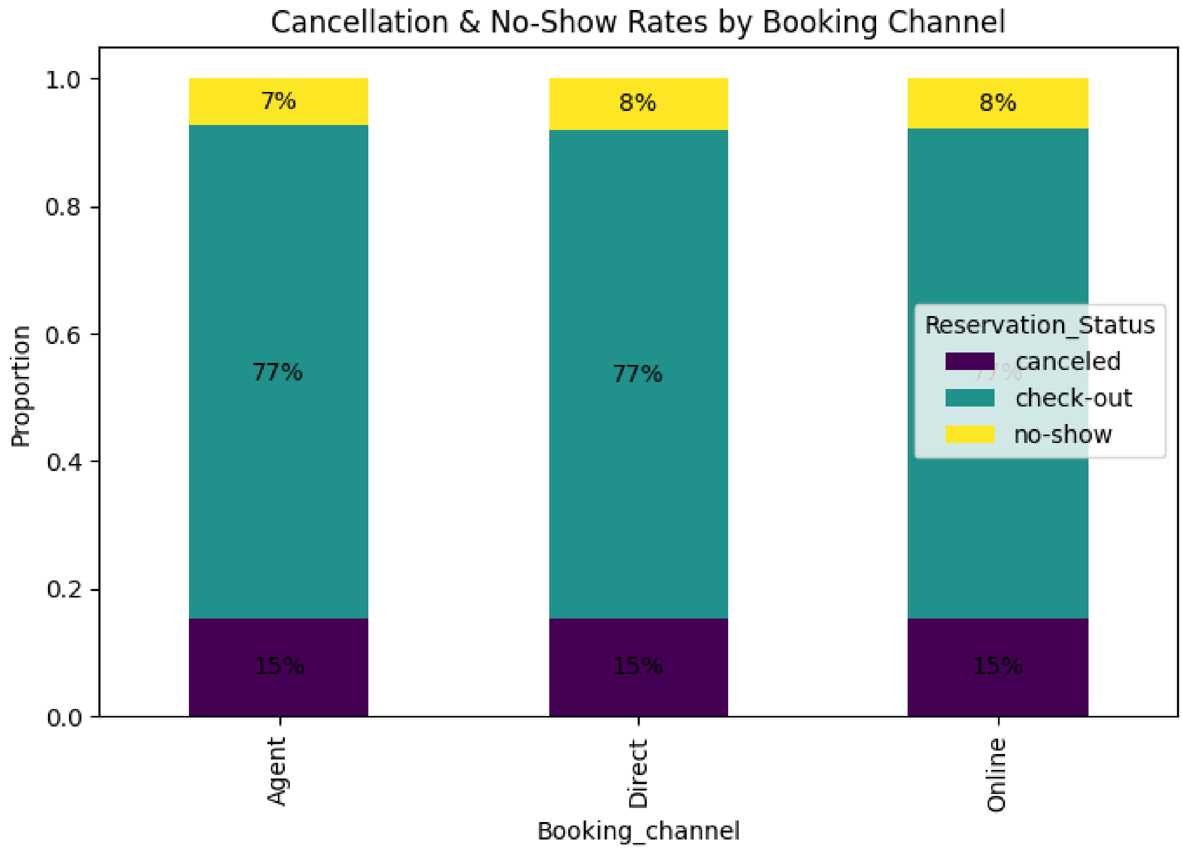
The pair plot visually presents the distributions of Age and Room Rate, categorized by Reservation Status, the scatterplots between Age and Room Rate reveal no clear linear relationship, with substantial overlap across different reservation statuses.

3.6 Bar chart of reservation status



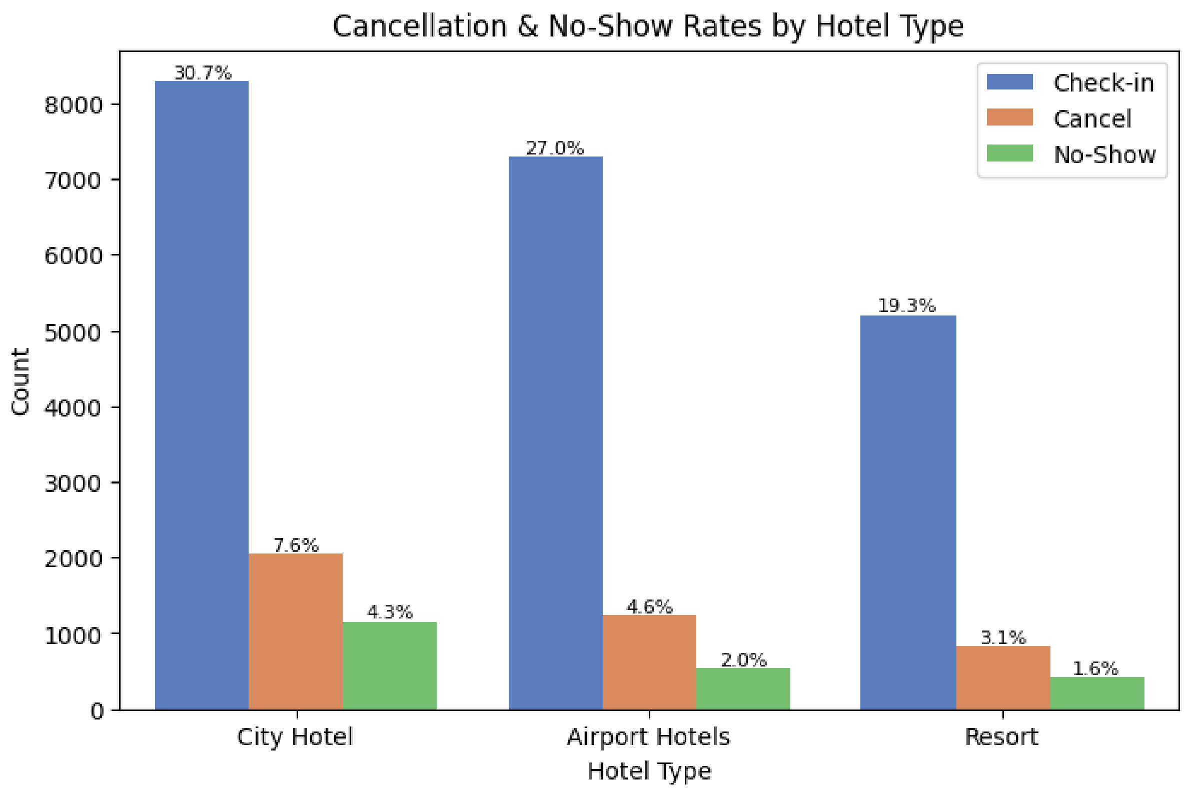
The bar chart displays the distribution of reservation statuses, with check-outs accounting for 77.0% of bookings, followed by cancellations at 15.2%, and no-shows at 7.8%. This distribution indicates that check-outs are the dominant outcome, while cancellations and no-shows represent smaller proportions of the total bookings.

3.7 Bar chart of cancellation & no-show rates by booking channel



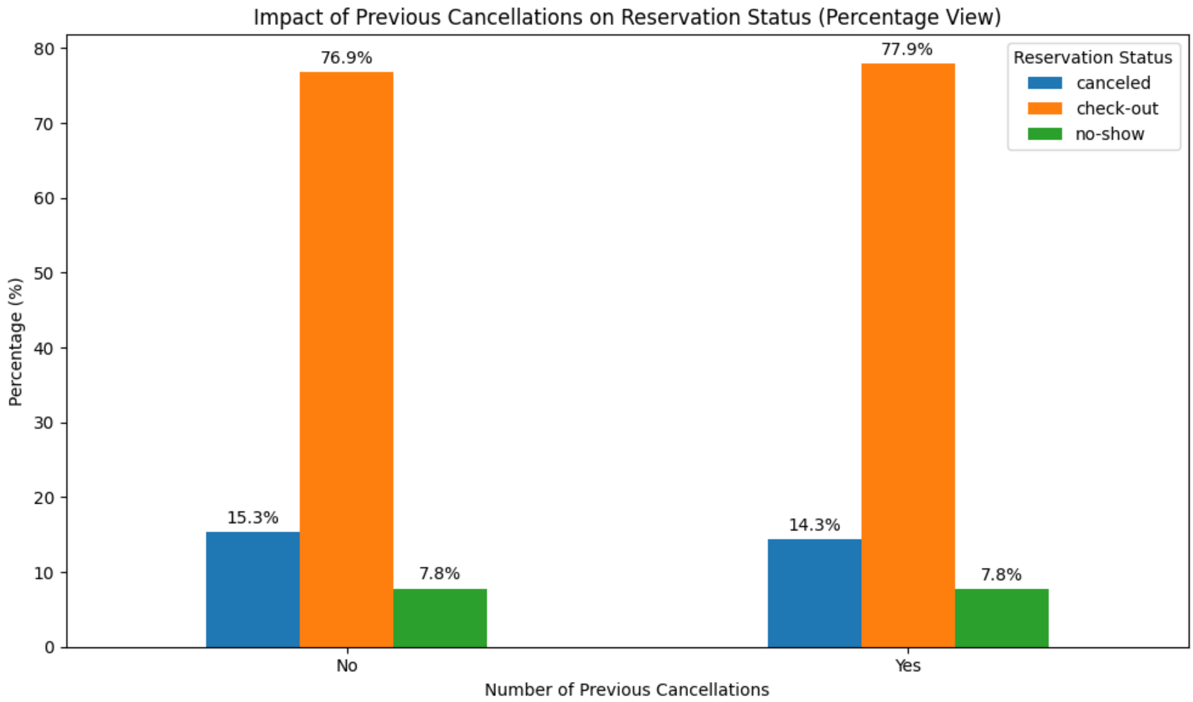
The stacked bar chart visually presents the distribution of reservation statuses (check-out, canceled, no-show) across three booking channels (Agent, Direct, Online), showing consistent proportions for check-outs and cancellations across all channels. The proportions of no-shows are also very similar, with a slight variation in the Agent channel.

3.8 Cancellation & no-show rates by hotel type



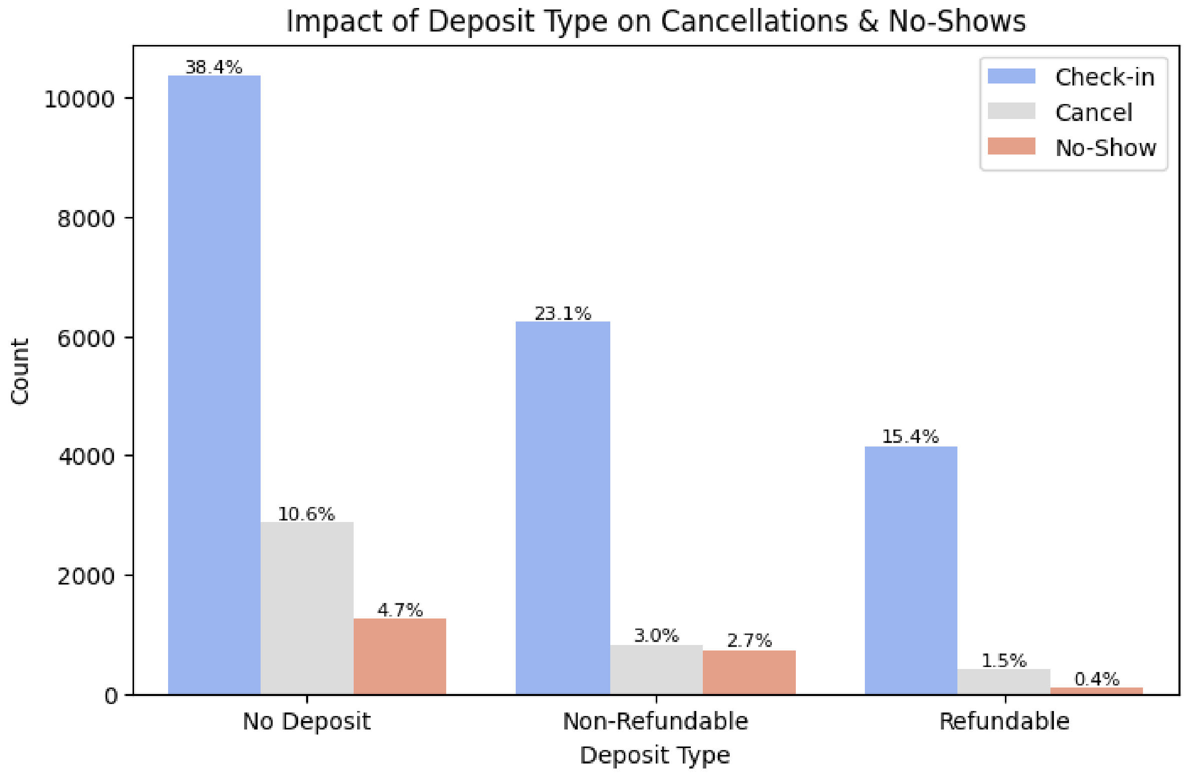
The grouped bar chart displays the counts of Check-ins, Cancellations, and No-Shows for each hotel type, with City Hotels having the highest counts across all statuses, followed by Airport Hotels, and then Resorts. The chart also provides percentage labels for each status within each hotel type, allowing for a comparison of proportions specific to each type.

3.9 Impact of previous cancellation on reservation status



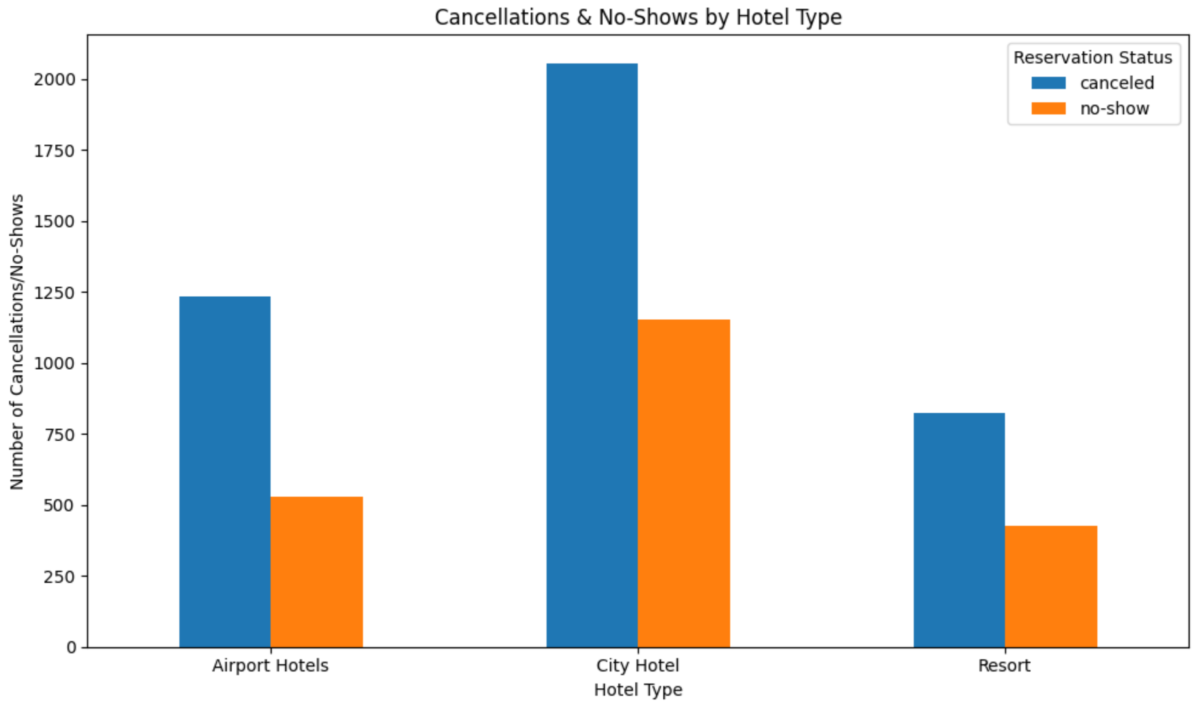
The grouped bar chart displays the percentage distribution of reservation statuses (canceled, check-out, no-show) for bookings with and without previous cancellations, showing that the majority of bookings in both groups resulted in check-outs. The percentages of cancellations, check-outs, and no-shows are similar between the two groups, with slight variations in cancellation rates.

3.10 Impact of deposit type on cancellations & no-show



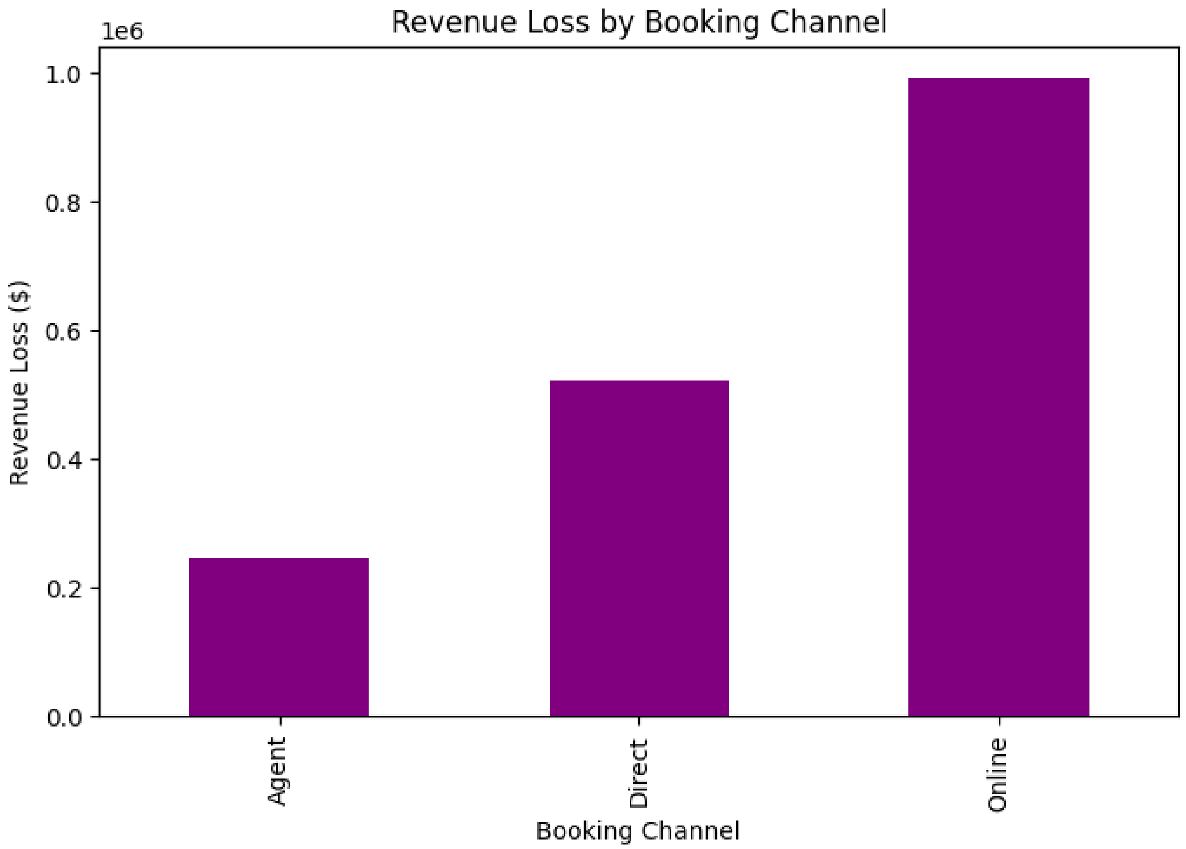
The grouped bar chart displays the counts of Check-ins, Cancellations, and No-Shows for each deposit type, with "No Deposit" bookings having the highest counts across all statuses, followed by "Non-Refundable," and then "Refundable." The chart also provides percentage labels for each status within each deposit type, allowing for a comparison of proportions specific to each type.

3.11 Cancellations & no-shows by hotel type



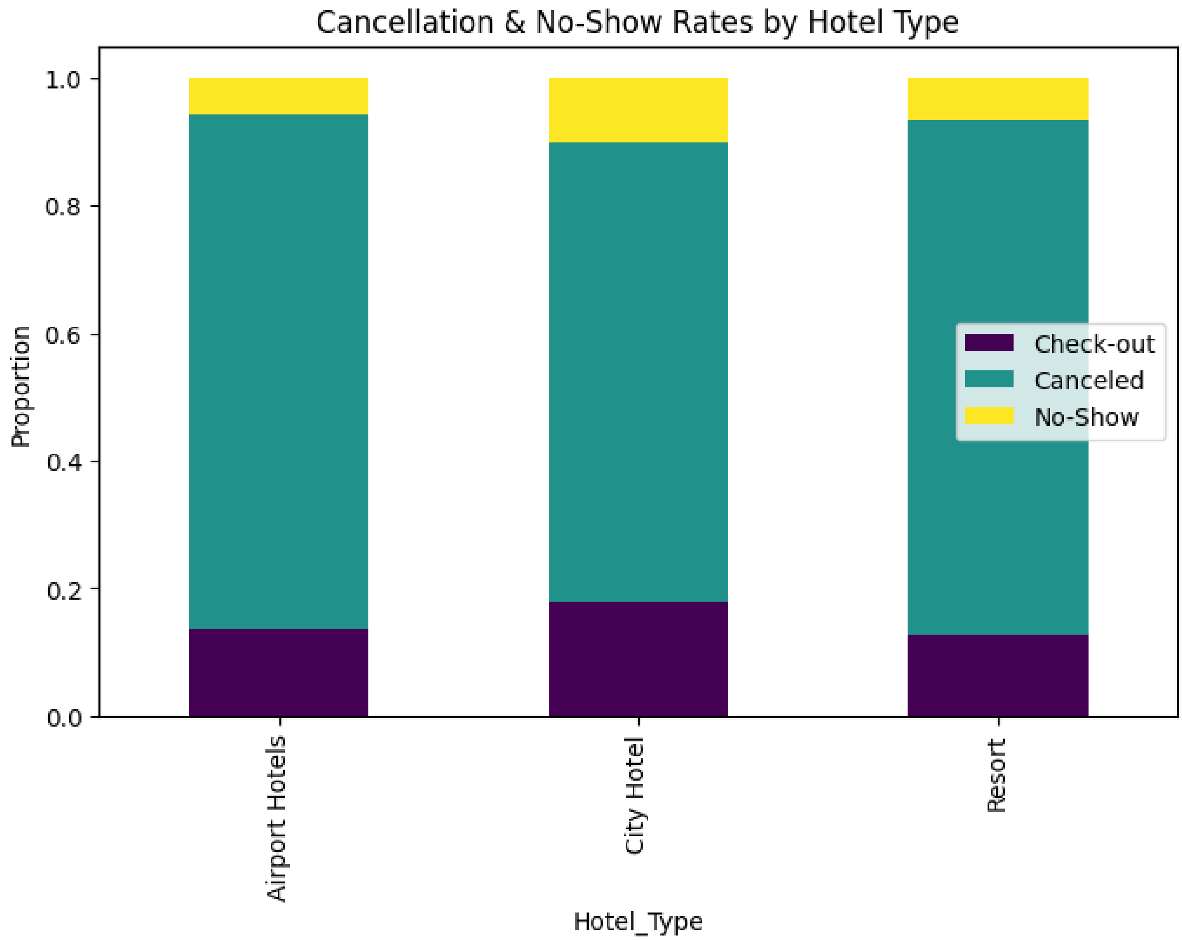
The grouped bar chart displays the counts of cancellations and no-shows for each hotel type, with City Hotels having the highest counts for both, followed by Airport Hotels, and then Resorts. For all hotel types, the number of cancellations exceeds the number of no-shows.

3.12 Revenue Loss by booking channel



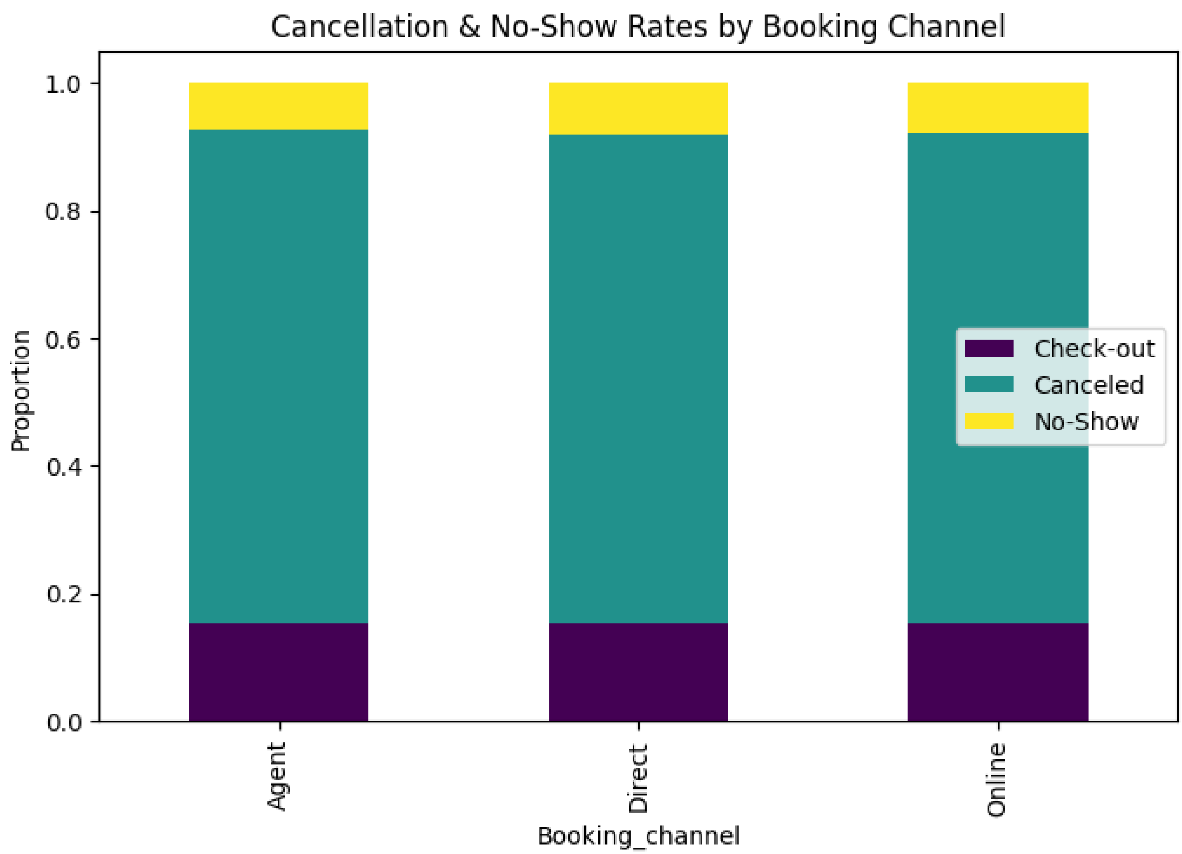
The bar chart displays the revenue loss associated with each booking channel, with the Online channel having the highest loss, followed by Direct, and then Agent. The visual comparison shows that the Online channel's revenue loss is significantly higher than the others, with the Agent channel experiencing the lowest loss.

3.13 Cancellation & no-show rates by hotel type



The stacked bar chart displays the proportional distribution of reservation statuses (check-out, canceled, no-show) for each hotel type, showing similar check-out proportions across all types. The chart visually suggests variations in cancellation and no-show proportions, with Resorts appearing to have the smallest proportions for both, although precise comparisons are challenging without specific labels.

3.14 Cancellation & no-show rates by booking channel



The stacked bar chart displays the proportional distribution of reservation statuses (check-out, canceled, no-show) for each booking channel, showing that the proportions of these statuses appear to be virtually identical across Agent, Direct, and Online channels. This visual consistency indicates that the relative proportions of check-outs, cancellations, and no-shows are similar regardless of the booking channel used.

4. Inferential analysis

4.1. Cancellation rates across hotel types

Chi-Square Statistic: 285.14
P-Value: 0.0000

The hypothesis test, supported by a Chi-Square statistic of 285.14 and a P-Value of 0.0000, rejects the null hypothesis, indicating a significant difference in cancellation rates across hotel types. This statistical analysis confirms that at least one hotel type has a cancellation rate that is significantly different from the others.

4.2. Cancellation rates between booking channels

Chi-Square Statistic: 1.15
P-Value: 0.8869

The hypothesis test, with a Chi-Square statistic of 1.15 and a P-Value of 0.8869, fails to reject the null hypothesis, indicating no significant difference in cancellation rates between booking channels. This statistical analysis suggests that the cancellation rates for Agent, Direct, and Online bookings are not significantly different from each other.

4.3. Cancellation rates based on deposit type

Chi-Square Statistic: 819.32
P-Value: 0.0000

The hypothesis test, with a Chi-Square statistic of 819.32 and a P-Value of 0.0000, rejects the null hypothesis, indicating a significant difference in cancellation rates based on deposit type. This statistical analysis confirms that deposit type has a significant impact on cancellation rates.

4.4. cancellation on previous cancellations

T-Statistic: -1.51
P-Value: 0.1323

The hypothesis test, with a T-Statistic of -1.51 and a P-Value of 0.1323, fails to reject the null hypothesis, indicating no significant difference in previous cancellations between the groups. This statistical analysis suggests that the number of previous cancellations does not significantly vary between the compared groups.

4.5. Difference in stay duration among reservation types

F-Statistic: 0.15
P-Value: 0.8601

The hypothesis test, with an F-Statistic of 0.15 and a P-Value of 0.8601, fails to reject the null hypothesis, indicating no significant difference in stay duration among reservation types. This statistical analysis suggests that stay durations are not significantly different across the various reservation types.

4.6. Effect of income level on cancellations

Chi-Square Statistic: 3.67
P-Value: 0.7208

The hypothesis test, with a Chi-Square statistic of 3.67 and a P-Value of 0.7208, fails to reject the null hypothesis, indicating no significant effect of income level on cancellations. This statistical analysis suggests that income level does not have a significant impact on cancellation rates.

5. Predictive model

5.1. Xgboost

- **Hyperparameter Tuning:** To optimize model performance, we performed a randomized search with cross-validation to find the best combination of hyperparameters. 5-fold cross-validation was used.
- **Best Hyperparameters:** The best hyperparameters identified through the tuning process were:

    - subsample: 1.0
    - Reg lambda: 1
    - Reg alpha: 2
    - N estimators: 100
    - Max depth: 4
    - Learning rate: 0.1
    - gamma: 0.1
    - Col sample by tree: 1.0

- **Performance Metrics:** The trained XGBoost model achieved the following performance:
    - **Test Accuracy:** 0.839
    - **Training Accuracy:** 0.955

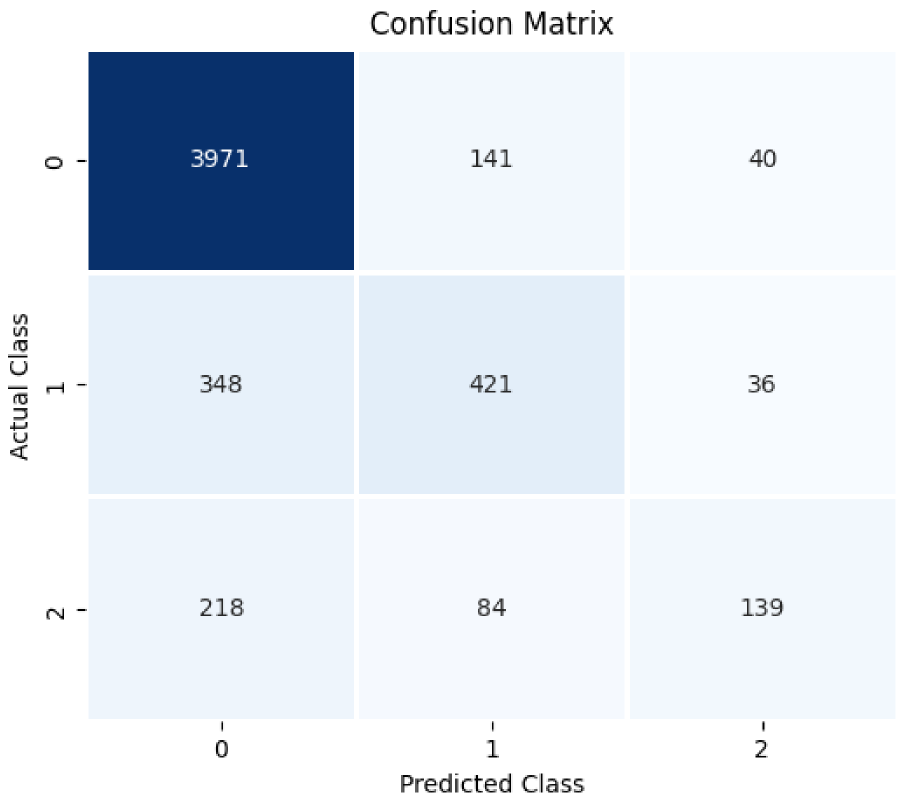
The XGBoost model achieved a test accuracy of 82.6% in predicting reservation status (check-out, cancellation, or no-show). While the model demonstrated strong performance in predicting check-out reservations (Class 1), it exhibited lower precision and recall for cancellations (Class 2) and no-shows (Class 3).

A detailed breakdown of performance by class is as follows:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Check-Out (1)	0.88	0.94	0.91	4152
Cancellation (2)	0.61	0.50	0.55	805
No-Show (3)	0.51	0.34	0.40	441

The macro-average F1-score was 0.62, and the weighted-average F1-score was 0.81. These results indicate that, while the model performs well overall in the dominant class (check-outs), it struggles to accurately identify cancellations and no-shows, likely due to class imbalance and the inherent difficulty in predicting these less frequent events. Furthermore, a difference of about 12% between the test accuracy and the training accuracy (95.5%) suggests possible overfitting.

5.1.1. Confusion matrix



The model performs exceptionally well in predicting check-outs, correctly classifying 3,971 instances with a high recall of 94%. However, it faces challenges in distinguishing cancellations and no-shows, with lower recall rates of 50% and 34%, respectively.

5.2. Logistic Regression

The Logistic Regression model achieved an overall accuracy of 83.4% in predicting reservation status (check-out, cancellation, or no-show). While this overall accuracy provides a general indication of performance, a more detailed examination of class-specific results is necessary to understand the model's strengths and weaknesses.

A breakdown of the model's performance by class is as follows:

Class	Precision	Recall	F1-Score	Support
Check-Out (1)	0.87	0.96	0.91	4152
Cancellation (2)	0.64	0.52	0.58	805
No-Show (3)	0.70	0.23	0.35	441

- **Check-Out Prediction (Class 1):** The model demonstrates strong performance in predicting check-out reservations, achieving high precision (0.87) and recall (0.96). This suggests that the model is effective at identifying the features most strongly associated with successful check-outs.
- **Cancellation Prediction (Class 2):** The model achieves moderate performance in predicting cancellations, with a precision of 0.64 and a recall of 0.52. This indicates that while the model correctly identifies some cancellations.
- **No-Show Prediction (Class 3):** The model exhibits the weakest performance in predicting no-shows, with a relatively high precision of 0.70 but a very low recall of only 0.23. This indicates that while the model is fairly accurate when it *does* predict a no-show.

5.3. The Support Vector Machine (SVM)

The Support Vector Machine (SVM) model achieved an overall accuracy of 83% in predicting reservation status (check-out, cancellation, or no-show). While this provides a general performance overview, a more detailed examination of class-specific results is essential for understanding the model's strengths and weaknesses.

The SVM model's performance by class is summarized below:

Class	Precision	Recall	F1-Score	Support
Check-Out (1)	0.85	0.97	0.90	4187
Cancellation (2)	0.67	0.44	0.53	829
No-Show (3)	0.84	0.15	0.26	382

Key Observations:

- **Check-Out Prediction (Class 1):** The SVM model demonstrates excellent performance in predicting check-out reservations. It achieves a high precision of 0.85 and a very high recall of 0.97, indicating that it effectively identifies nearly all check-out bookings while also minimizing false positive errors.
- **Cancellation Prediction (Class 2):** The model shows moderate performance in predicting cancellations, with a precision of 0.67 and a recall of 0.44. This suggests that when the model predicts a cancellation, it is correct about 67% of the time. However, it only captures 44% of all actual cancellations, indicating a significant number of false negatives (cancellations that the model fails to predict).
- **No-Show Prediction (Class 3):** The SVM model exhibits the weakest performance in predicting no-show reservations. While it achieves a high precision of 0.84, indicating that when it predicts a no-show, it is usually correct, the recall is extremely low at only 0.15.

6. Discussion

This report presented a data-driven analysis to enhance revenue management for hotels by investigating booking cancellations and no-shows. By combining exploratory data analysis (EDA), predictive modeling, and customer segmentation, we aimed to uncover key factors influencing reservation outcomes and provide actionable recommendations for hotel management.

6.1. Implications for hotels

These findings have several important implications for hotels:

1. **Targeted Interventions:** The hotel can leverage these insights to implement targeted interventions based on booking characteristics, customer demographics, and booking channel. For example, guests booking through the Online channel with "No Deposit" could be offered incentives to confirm their reservation or switch to a "Non-Refundable" deposit.
2. **Revenue Optimization:** By analyzing the room rate distribution and booking channel performance, the hotel can optimize its pricing strategies and marketing efforts to maximize revenue.
3. **Resource Allocation:** The insights into hotel type distribution can inform resource allocation decisions, such as staffing levels and inventory management. City Hotels, which account for the largest share of bookings, may require more resources than Airport Hotels or Resorts.
4. **Improved Deposit Policies:** The hotel can refine its deposit policies to reduce cancellations and no-shows, while also ensuring that the policies are fair and attractive to guests.

7. Conclusion

This report provides valuable insights into the key factors impacting reservation outcomes at hotels, highlighting the importance of data-driven decision-making for effective revenue management. The findings emphasize the need for a comprehensive approach that addresses data quality, optimizes booking strategies, and leverages advanced analytical techniques. By implementing the recommendations outlined in this report, Hotels can reduce revenue loss, improve resource allocation, and enhance overall operational efficiency. Further investigation should be undertaken to more effectively and accurately measure and capture factors contributing to the cancellation history. Moreover, additional analysis would be of value if more detailed external information could be acquired. In all, the key insights and recommendations will serve as a springboard toward more comprehensive revenue management and operational strategies in the future.

7.1. Key Findings

1. **Importance of Exploratory Data Analysis (EDA):** The EDA phase revealed several important insights into the characteristics of hotel's bookings. The age distribution of guests was relatively uniform, suggesting that age alone may not be a strong predictor of cancellations but is still valuable for marketing and service strategies. The room rate distribution showed a tiered pricing strategy, which can inform revenue management and pricing optimization strategies.
2. **Dominance of City Hotels and Online Bookings:** The analysis indicated that City Hotels accounted for the largest share of bookings, and the "Online" channel was the most popular booking method. This highlights the importance of focusing revenue management efforts on these segments.
3. **Impact of Deposit Type on Cancellations and No-Shows:** The results showed that "No Deposit" bookings had the highest counts of cancellations and no-shows, while "Non-Refundable" deposits were associated with lower cancellation rates. This suggests that deposit policies play a significant role in influencing booking behavior.
4. **Revenue Loss by Booking Channel:** The analysis revealed that the Online channel had the highest revenue loss associated with cancellations and no-shows, despite being the most popular booking method. This finding suggests a need for targeted interventions to reduce cancellations and no-shows within this channel.
5. **No strong correlations**
6. **No clear relationship with reservation status**

8. Acknowledgements

The authors gratefully acknowledge the support and guidance provided by the Statistics Society of the University of Sri Jayewardenepura, the organizing committee for this case study challenge. Their efforts in providing the dataset and facilitating the competition were instrumental in making this project possible.

9. Reference

Python Software Stack:

- Python (version 3.9+) - https://www.python.org
- Pandas (version 1.5+) - https://pandas.pydata.org/docs/
- NumPy (version 1.23+) - https://numpy.org/doc/
- Matplotlib (version 3.7+) - https://matplotlib.org/stable/contents.html
- Seaborn (version 0.12+) - https://seaborn.pydata.org/

- Scikit-learn (version 1.2+) - https://scikit-learn.org/stable/
- XGBoost (version 1.7+) - https://xgboost.readthedocs.io/en/stable/
- imbalanced-learn (version 0.11+) - https://imbalanced-learn.org/stable/
- jupyter Notebook - https://jupyter.org/