CASE STUDY: ENHANCING HOTEL REVENUE MANAGEMENT THROUGH PREDICTIVE ANALYTICS

DXP13

Contents

Abstract

The hotel industry faces significant challenges due to booking cancellations and no-shows, which contribute to revenue loss, inefficient resource utilization, and reduced guest satisfaction. These issues are particularly pronounced in large hotel chains, where the unpredictability of bookings can disrupt operations and strain resources. This case study focuses on utilizing predictive analytics to understand and mitigate the impact of cancellations and no-shows for Hotel A, a prominent hotel chain operating three distinct property types: Airport Hotels, Resorts, and City Hotels. With over 25,000 bookings annually across its diverse portfolio, Hotel A has experienced consistent disruption to its operations due to these factors. The primary objective of this report is to explore the underlying causes of cancellations and no-shows by analyzing historical booking data and identifying key factors such as booking behaviors, seasonal trends, and pricing strategies.

The study also aims to quantify the financial implications of cancellations and no-shows by calculating the associated revenue loss, which in 2024 was estimated to be over $2 million. By developing a predictive model to classify bookings as cancellations, no-shows, or check-ins, the report seeks to provide Hotel A with a robust forecasting tool that can help anticipate disruptions and make proactive adjustments to pricing, promotions, and reservation policies. The predictive model will be evaluated based on its accuracy and the ability to explain the factors contributing to each prediction, ultimately empowering hotel management to make informed, data-driven decisions.

Additionally, customer segmentation techniques will be employed to analyze guest behaviors related to cancellations and no-shows, enabling Hotel A to tailor its strategies to different guest profiles. This will help improve customer retention and enhance operational efficiency by aligning marketing efforts and policies with specific guest preferences and cancellation patterns. By integrating both predictive analytics and customer segmentation, this report will offer actionable insights to optimize resource allocation, reduce cancellations and no-shows, and enhance overall guest satisfaction. The insights gained from the final model will ultimately support the hotel chain's efforts to improve its operational efficiency, reduce revenue loss, and foster long-term customer loyalty.

Introduction

The hotel industry faces significant challenges in managing booking cancellations and no-shows, which lead to revenue loss, inefficient resource utilization, and lower guest satisfaction. Hotel A, a chain operating three types of properties—airport hotels, resorts, and city hotels—seeks a data-driven approach to predict and mitigate these issues.

Hotel A, a prominent chain with diverse offerings including airport hotels, resorts, and city hotels, has identified this challenge as a key area for improvement. With properties spanning different locations and catering to various types of travelers, the complexity of managing cancellations and no-shows is further heightened. To address these issues, Hotel A has turned to a data-driven approach, leveraging advanced analytics and machine learning models to gain deeper insights into the patterns and behaviors behind these booking disruptions.

This report aims to:

1. Investigate the key factors contributing to booking cancellations and no-shows.
2. Develop predictive models for classifying reservations as cancellations, no-shows, or check-ins.
3. Implement customer segmentation based on cancellation behaviors.
4. Provide actionable insights to enhance operational efficiency and customer retention.

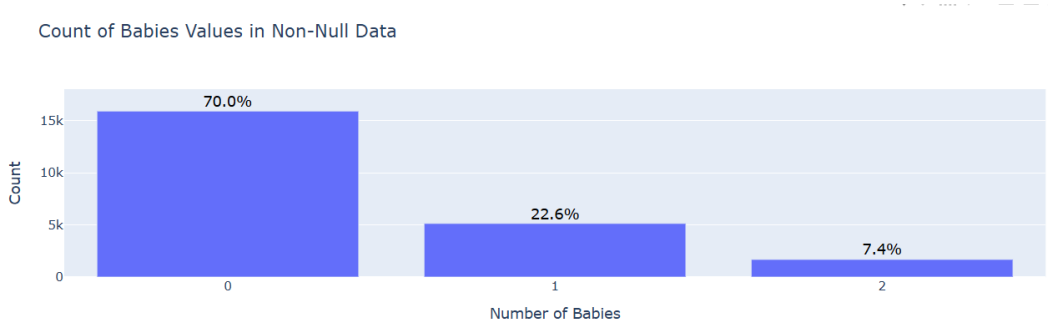
The dataset used in this study is vast and multifaceted, encompassing a wide range of features that provide valuable insights into the booking and cancellation behaviors of guests. It includes various types of information, capturing both transactional and demographic data, as well as operational metrics from different hotel properties. The richness of the dataset allows for a comprehensive analysis, with variables that reflect multiple aspects of the booking process, guest preferences, pricing strategies, and seasonal trends. This diversity in data provides a solid foundation for building robust predictive models and conducting detailed segmentation, enabling a deeper understanding of the factors influencing cancellations and no-shows across Hotel A's different property types.

Methods

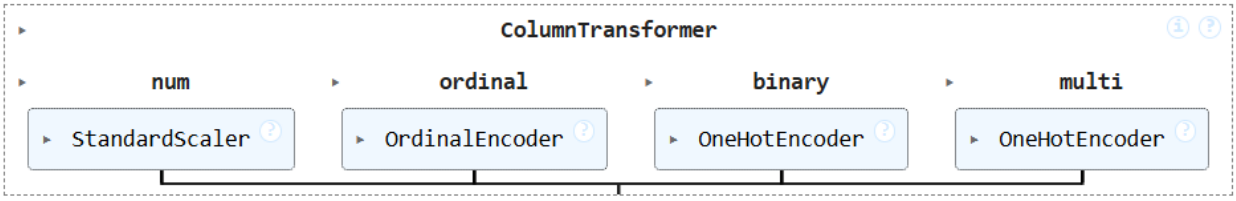
Data Preprocessing

The dataset, provided by Hotel A, contains 27,500 booking records with over 20 attributes. We conducted several data preprocessing steps, including:

- **Handling Duplicates**: Duplicate Reservation_ID entries were kept as distinct records because each reservation contained unique attributes (e.g., gender, age, and meal type). Aggregation was not possible due to the variation in these attributes.
- **Missing Values**: The 'Babies' column had 17.29% missing values, which were filled with zeros based on the most frequent value in different groupings of other features, such as Age, Hotel_Type and Income. This decision was justified by the fact that for every combination of these groupings, the mode of the 'Babies' column was consistently 0, indicating that, in general, bookings with babies were less common across all segments. As a result, filling the missing values with zeros aligned with the overall distribution of the data and avoided introducing any bias. All other columns were well-balanced in terms of missing values, ensuring that the data integrity was maintained.



Scaling & Encoding



Several encoding and scaling techniques were applied to prepare the data for modeling. The StandardScaler was used to scale continuous variables, ensuring they have a mean of 0 and a standard deviation of 1. OrdinalEncoder was applied to ordinal features such as education and income, where the education categories were ordered as 'Mid-School' < 'High-School' < 'College' < 'Grad', and income categories were ordered as '<25K' < '25K--50K' < '50K--100K' < '>100K'. OneHotEncoder was used for binary categorical variables, and also for multi-class categorical features, dropping the first category to avoid multicollinearity. These preprocessing steps were performed before model training to ensure the data is properly formatted and scaled, allowing the model to learn efficiently and accurately.

Feature Engineering

1. Date-Time Features

The dataset contains several date columns, including Expected_checkin, Expected_checkout, and Booking_date. These columns were converted into datetime objects for easier manipulation and analysis. Several new features were derived from these columns:

- Expected Check-in Year, Month, Day, Day of the Week: Extracted from the Expected_checkin date to capture temporal patterns related to the booking date.
- Expected Checkout Year, Month, Day, Day of the Week: Extracted from the Expected_checkout date to understand the checkout behavior.
- Booking Date Year, Month, Day, Day of the Week: Extracted from the Booking_date to gain insights into booking patterns.

In addition, duration-based features were calculated to better understand booking behavior:

- Duration Check-in to Checkout: This feature represents the length of stay by calculating the number of days between Expected_checkin and Expected_checkout. This feature is useful in understanding how long guests are staying at the hotel.
- Booking to Check-in Duration: This feature represents the number of days between the Booking_date and Expected_checkin. This metric can be useful to identify last-minute bookings or those made well in advance.

2. Number of People and Rooms

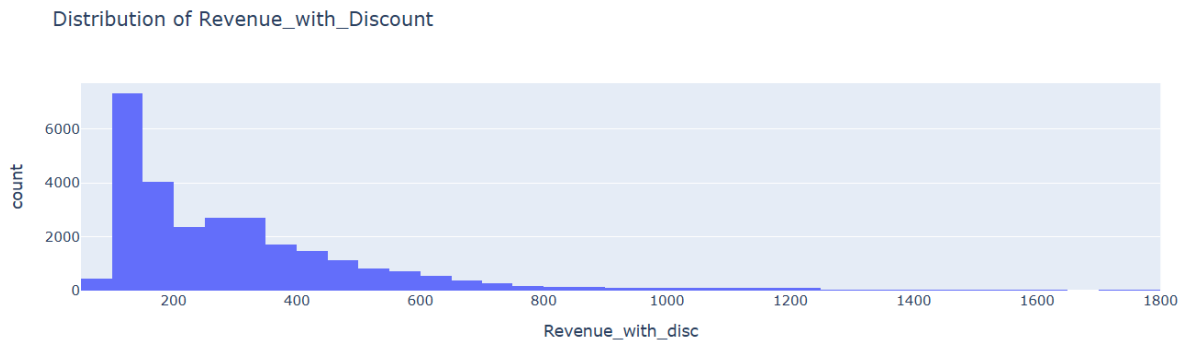
To quantify reservation size, the **No_of_ppl_without_babies** feature sums the number of adults and children, excluding babies, to provide an accurate count of people. The **No_of_Rooms** feature calculates the required number of rooms based on this count, assuming each room accommodates up to 5 people; if the count exceeds 5, additional rooms are assigned accordingly.

3. Revenue Calculation

Several revenue-related features were derived to estimate the total revenue from the booking. The **Revenue** feature is calculated as a product of Room_Rate, Duration_checkin_checkout, and No_of_Rooms, representing the total revenue before any discounts.

- **Revenue_with_Discount = Revenue × (1 - Discount Rate/100)**

The above metric provides a more accurate estimate after discounts.

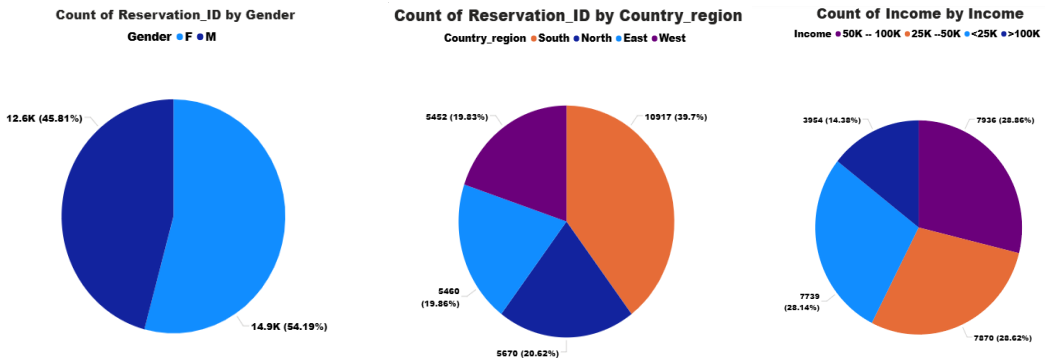


To handle outliers and skewed distributions, a Log-Transformed Revenue with Discount feature was created by applying a logarithmic transformation to Revenue_with_Discount, reducing skewness and stabilizing variance, which helps normalize the data for better performance in machine learning models (skewness after transformation = 0.49).

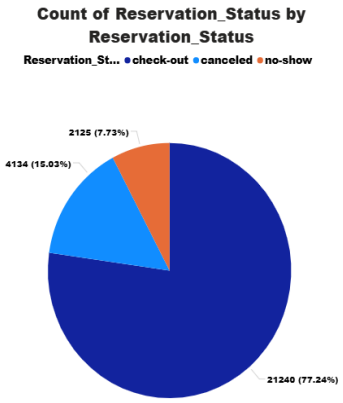
EDA

Univariate Analysis

The dataset is well-balanced across categorical variables. Females slightly outnumber males, and ethnic diversity is notable, with Latino, Caucasian, and African American groups leading. Higher education levels are predominant, and income is mainly in the middle range. Regionally, the South has the highest concentration, with other areas more evenly distributed. City Hotels are the most preferred, followed by Airport Hotels and Resorts. Most guests opt for Bed and Breakfast meals, are first-time visitors, and prefer No Deposit and online bookings. Car Parking and promotions are commonly requested. Overall, the dataset reflects diverse guest profiles, making it suitable for analyzing booking patterns.



The Reservation_Status variable shows the distribution of booking outcomes. The majority of reservations resulted in a check-out (21,240), indicating successful stays. Cancellations follow, with 4,134 bookings canceled, and the smallest group is no-shows (2,125), where guests did not show up. This distribution highlights that most reservations are fulfilled, with a smaller portion either canceled or resulting in no-shows. The pie chart visually illustrates this distribution, emphasizing the proportion of each outcome and helping to understand the extent of cancellations and no-shows in relation to completed stays.



Numerical Features

	Age	Adults	Children	Babies	Discount_Rate	Room_Rate
mean	43.98	2.23	1.74	0.31	8.29	155.65
min	18.00	1.00	1.00	0.00	0.00	100.00
25%	31.00	1.00	1.00	0.00	0.00	124.00
50%	44.00	2.00	2.00	0.00	5.00	149.45
75%	57.00	3.00	2.00	0.00	10.00	181.00
max	70.00	5.00	3.00	2.00	30.00	250.00

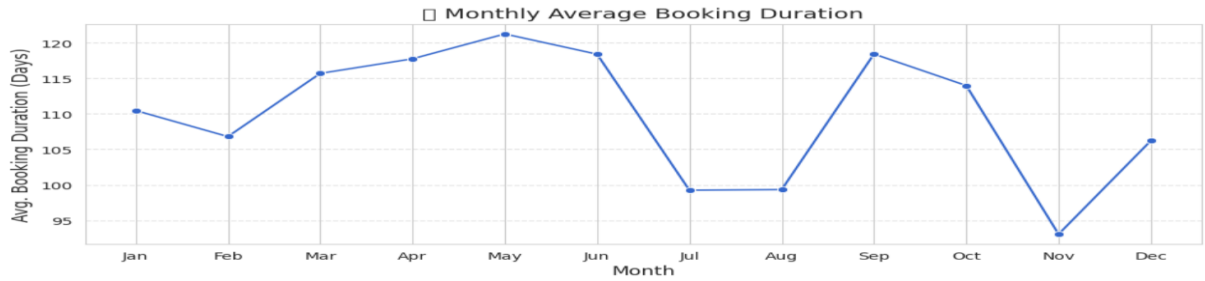
The numerical features in the dataset reveal key characteristics of bookings. The average guest age is 44 years, with most bookings involving 2 adults and 1 to 2 children. Babies are less common, with an average of 0.31 babies per booking. Discount rates typically range from 0% to 30%, with a mean of 8.29%. The room rate averages 155.65, with most bookings falling between 124 and 181. These statistics provide insight into the typical composition of guests and booking details, including room pricing and discount usage.

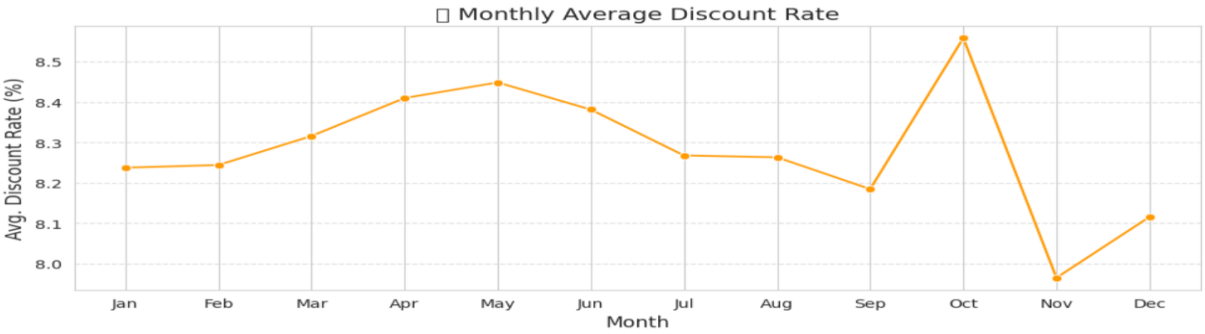
Several ANOVA, Chi-square tests, and Spearman correlation analyses were conducted to uncover key insights in the dataset. The ANOVA tests were used to examine the impact of categorical variables on numerical outcomes, while the Chi-square tests assessed the relationships between categorical variables. Spearman correlation was applied to identify any monotonic relationships between continuous features. These statistical analyses helped to identify significant patterns and relationships in the data, providing valuable insights for understanding booking behaviors and predicting cancellations.

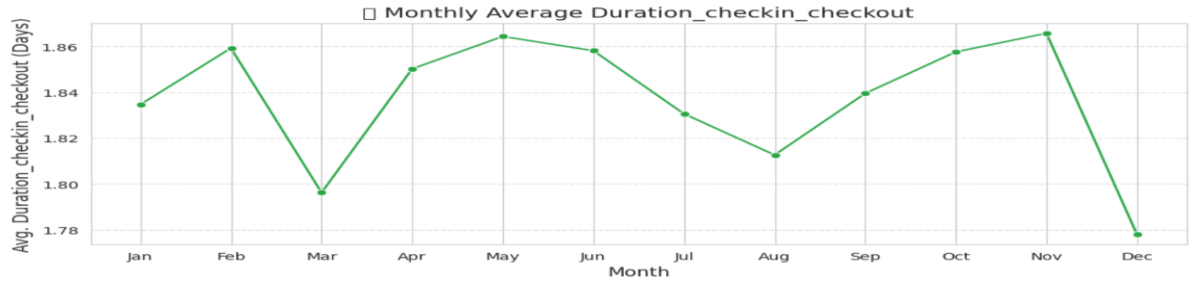
Results

Business Questions

Observations from the Monthly Trends

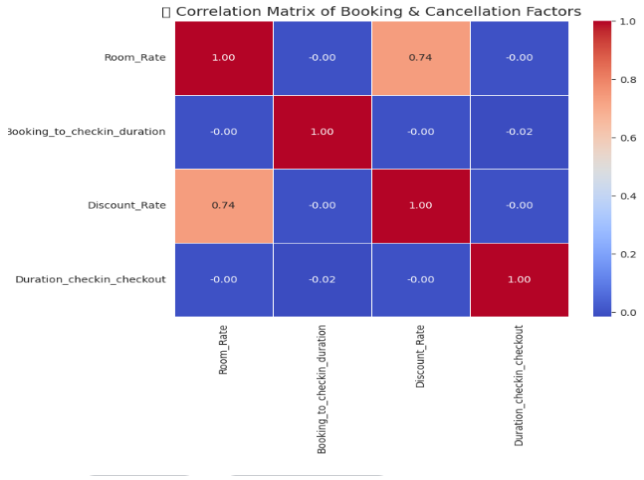




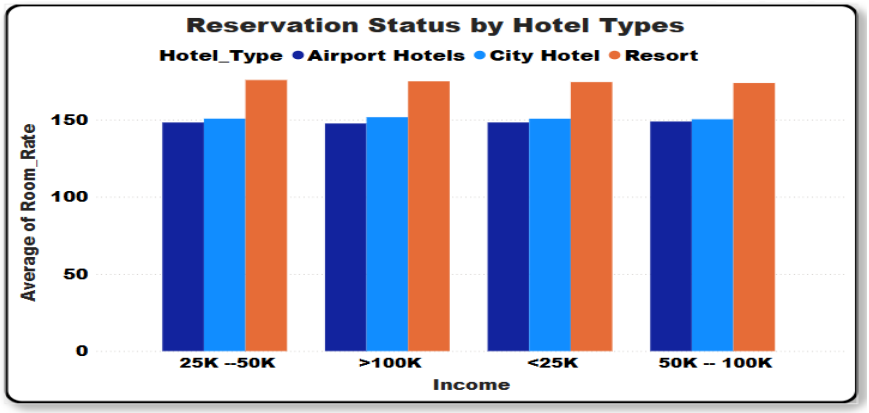


Analyzing key trends reveals the following insights. The **cancellation rate** peaks around August at 20%, while **booking duration** decreases, indicating that shorter stays are more prone to cancellations. However, as the cancellation rate drops to 12%, booking duration increases, suggesting that longer stays are less likely to be canceled. In **October**, there is a significant spike in the **discount rate** (around 8.5%), and the cancellation rate remains low (~12%), suggesting that discounts may encourage guests to keep their reservations. However, in **November**, the cancellation rate peaks again, and the discount rate is lower (~8%), which could imply that higher

prices or other factors contribute to cancellations. Additionally, both **booking duration** and **check-in duration** peak in **May-June** and **October**, indicating that people tend to stay longer during these months. In contrast, **December** sees a sharp drop in booking duration, suggesting that people book shorter stays during the holiday period. A correlation analysis revealed that there is a strong positive relationship between **room rate** and **discount rate**, with a correlation value of 0.74, suggesting that higher room rates are associated with higher discounts. However, there is no correlation between **Booking to Check-in Duration**, indicating that the price of the room does not influence how early a guest books. Similarly, discounts do not seem to impact how early guests book, and the correlation between **Booking to Check-in Duration** and **Duration from Check-in to Checkout** is very low (-0.02), suggesting that booking in advance does not affect the length of stay.



Do High-Income Guests Book More Expensive Rooms?



We analyzed the historical booking data to explore the relationship between income levels and the average room rates booked. The data revealed that the average room rate remained fairly consistent across all income groups. Whether a guest earned less than $25K or more than $100K, the room

prices booked appeared to be similar. It was also observed that resorts had slightly higher average room rates compared to other hotel types. However, within each hotel type, income did not seem to significantly influence the price paid for rooms. To further investigate this relationship, a Spearman Correlation Test was conducted, which resulted in a high p-value (>0.05), indicating that there was no statistically significant correlation between income and room rate.

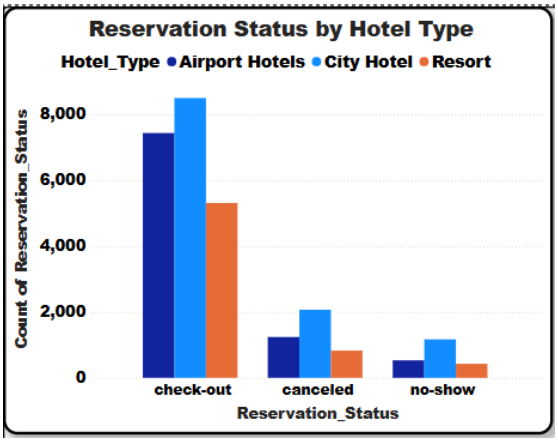
Do some hotel types have higher cancellation/no-show rates?

The analysis of cancellation and no-show rates across hotel types reveals that **City Hotels** have the highest total bookings (11,731) and the most cancellations (2,067). However, when considering the cancellation rates, **Resorts** exhibit a rate of around 12.6% (827 cancellations out of 6,563 bookings), and **Airport Hotels** have a higher cancellation rate of 13.5% (1,240 cancellations out of 9,205 bookings). These findings suggest that the cancellation and no-show rates vary across hotel types.

H_0: There is no relationship between hotel type and reservation status (cancellations happen randomly across all hotel types

H_1: There is a relationship between hotel type and reservation status (some hotel types are more prone to cancellations/no-shows)

To statistically validate whether these differences are significant, a **Chi-Square Test** was conducted to determine the relationship between **Hotel_Type** and **Reservation_Status** (whether the reservation was canceled, a no-show, or a check-in). The null hypothesis (H_0) states that there is no relationship between hotel type and reservation status, meaning cancellations occur randomly across all hotel types.

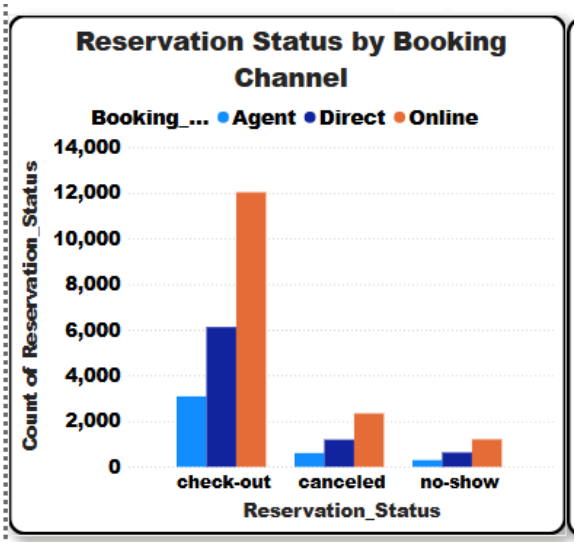


Contingency Table:

Reservation_Status	canceled	check-out	no-show	Total
Hotel_Type				
Airport Hotels	1240	7434	531	9205
City Hotel	2067	8496	1168	11731
Resort	827	5310	426	6563
Total	4134	21240	2125	27499

Chi-Square Test Results:
Chi-Square Statistic: 288.55
Degrees of Freedom: 4
P-value: 0.00000

Do online bookings have higher cancellation rates than direct bookings?



Contingency Table:

Reservation_Status	canceled	check-out	no-show	Total
Booking_channel				
Agent	603	3087	293	3983
Direct	1191	6122	629	7942
Online	2340	12031	1203	15574
Total	4134	21240	2125	27499

Based on the above results we can observe that cancelation rates are nearly identical across all booking channels

- Agent – 603/3983 = 15.14%
- Direct – 1191/7942 = 15.0%
- Online – 2340/15574 = 15.02%

While No shows rates are also very similar across the Booking Channels

- Agent – 293/3983 = 7.3%
- Direct -629/7942 = 7.92%
- Online – 2340/15574 – 7.72%

These findings suggest that cancellation and no-show rates do not vary significantly across the different booking channels.

```
Chi-Square Test Results:
Chi-Square Statistic: 1.19
Degrees of Freedom: 4
P-value: 0.87972
```

To confirm this observation, a **Chi-Square Test** was conducted to determine the relationship between **Booking_Channel** and **Reservation_Status** (whether the reservation was canceled, a no-show, or a check-in). The null hypothesis (H_0) states that there is no relationship between booking channel and reservation status, meaning cancellations and no-shows occur randomly across all booking channels. The alternative hypothesis (H_1) posits that there is a relationship, indicating some channels may have higher cancellation or no-show rates.

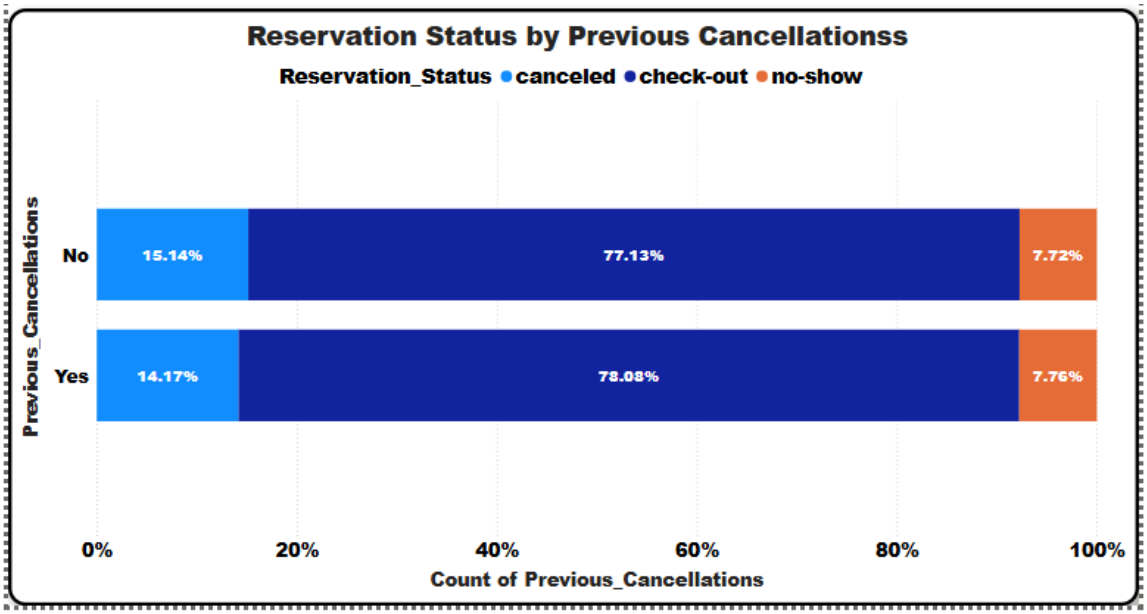
Do customers with previous cancellations tend to cancel again?

The analysis of cancellations based on previous cancellation history shows that customers with no prior cancellations had a cancellation rate of 15.14%, while customers with previous cancellations

had a slightly lower cancellation rate of 14.17%. At first glance, this suggests that there is no substantial difference in the cancellation behavior between the two groups. The contingency table supports this observation, implying that past cancellations might not significantly influence future cancellations.

```
Chi-Square Test Results:
Chi-Square Statistic: 2.08
Degrees of Freedom: 2
P-value: 0.35316
```

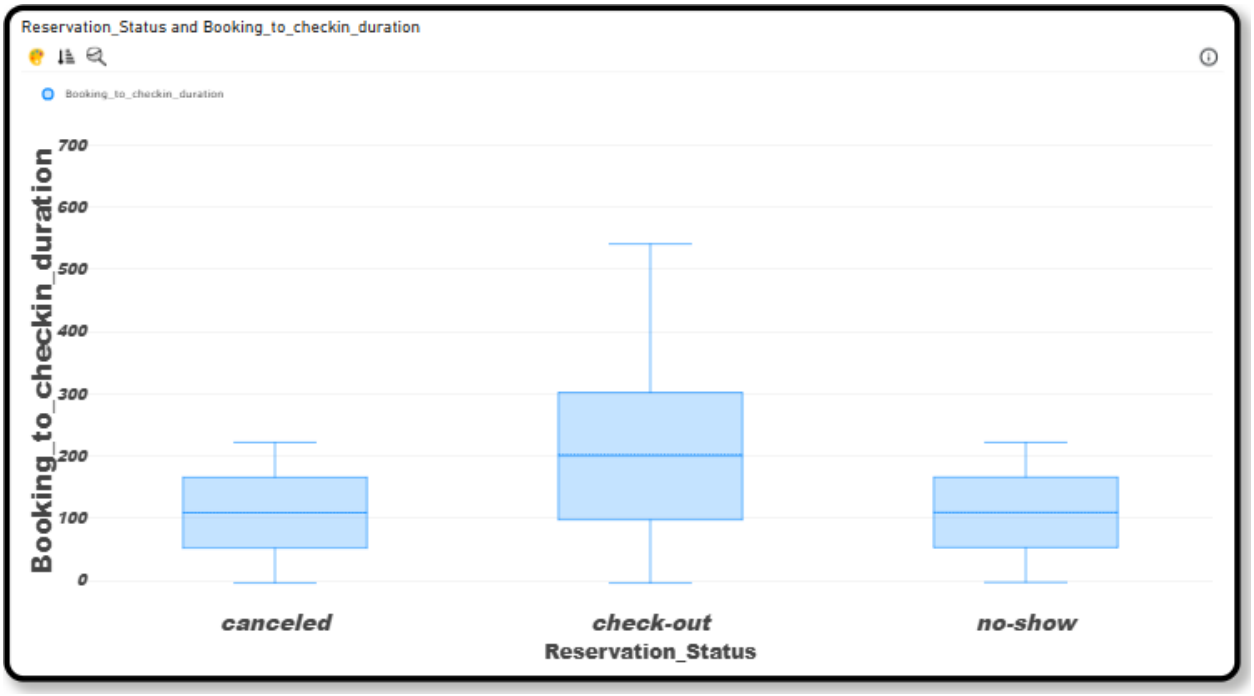
To confirm whether this small difference is just random variation or a real pattern, a **Chi-Square Test** was conducted to assess the relationship between **Past_Cancellations** and **Future_Cancellations**. The null hypothesis (H_0) posits that there is no relationship between past cancellations and future cancellations, while the alternative hypothesis (H_1) suggests that there is a relationship, and customers who previously canceled are more likely to cancel again.



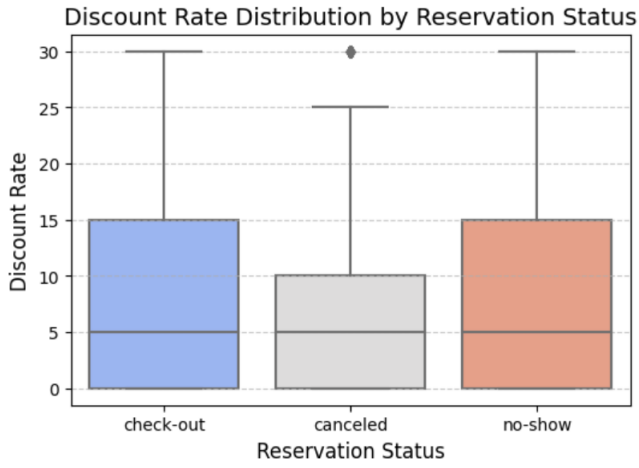
Does lead time (booking to check-in duration) impact cancellations?

The box plots show that cancellations and no-shows are associated with shorter lead times (i.e., bookings made closer to check-in). The boxplots for cancellations and no-shows are clustered near zero, indicating that these reservations tend to have shorter booking-to-check-in durations. This suggests that last-minute bookings are more likely to result in cancellations or no-shows. The data implies that shorter lead times may increase the risk of cancellations.

To further test this observation, an **ANOVA** test was conducted to assess whether the **Booking_to_Checkin_Duration** varies significantly between cancellations, check-outs, and no-shows. The null hypothesis (H_0) states that the lead time does not differ significantly across the three reservation statuses, while the alternative hypothesis (H_1) suggests that lead times vary depending on reservation status.



Do promotions and discounts increase cancellations?



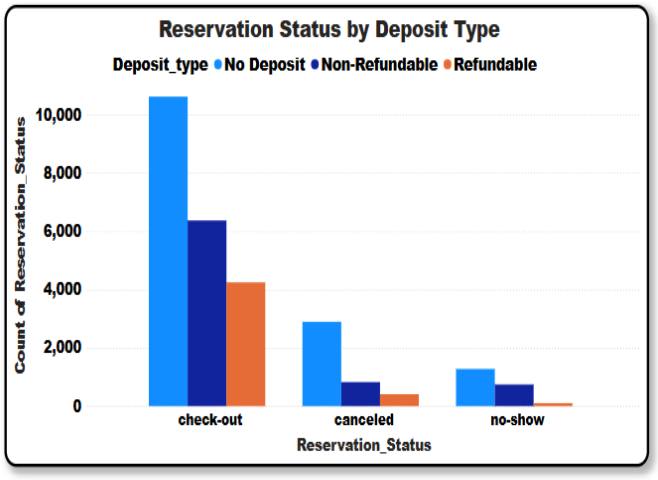
Based on the results we can see Cancellations tend to have lower discount rates on average, with a smaller interquartile range (IQR), No-show reservations also have a wide range, similar to check-outs. This suggest that the guest who received little or no discount may be likely to cancel, and discounts alone may not prevent no-shows, to check if discount rate significantly differs across reservation statuses we conduct ANOVA test,

```
ANOVA Test Results:
F-Statistic: 1.66
P-value: 0.18997
```

Based on the ANOVA test results, since the p value is greater than 0.05, we fail to reject the null hypothesis which means, There is no statistically significant difference in the average discount rate across canceled, check-out, and no-show reservations, Discounts alone are not a strong predictor of cancellations or no-shows

Does the deposit type affect cancellations?

The analysis of deposit types and their relationship to cancellation and no-show rates reveals key patterns.



Contingency Table:

Reservation_Status	canceled	check-out	no-show	Total
Deposit_type				
No Deposit	2895	10620	1276	14791
Non-Refundable	826	6372	743	7941
Refundable	413	4248	106	4767
Total	4134	21240	2125	27499

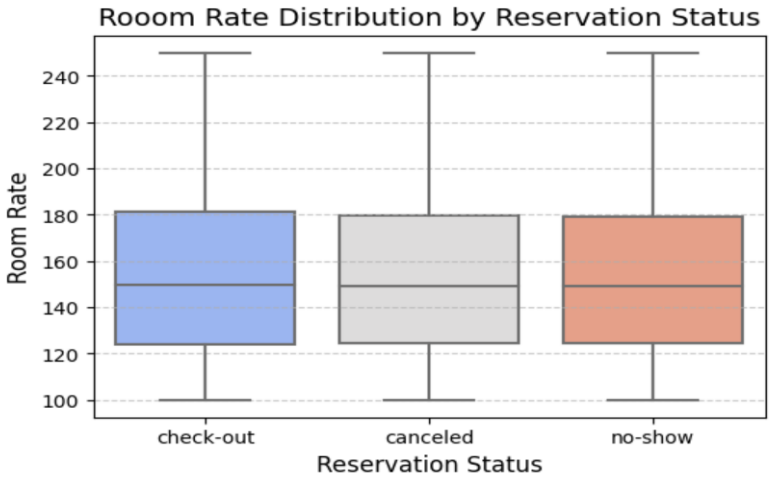
No Deposit bookings have the highest cancellation rate (19.6%) compared to other deposit types and Non-Refundable deposits have the highest No-Show rate (9.4%), likely due to customers being unable to cancel without losing their money.

```
Chi-Square Test Results:
Chi-Square Statistic: 827.25
Degrees of Freedom: 4
P-value: 0.00000
```

The above results are from chi- squared test, since the p value is less than 0.005, we can reject the null hypothesis, which means there is a significant relationship between Deposit Type and Reservation Status.

- Non-Refundable Deposits have fewer cancellations
- Refundable Deposits & No Deposit have higher cancellations
- No-Shows are more common for No Deposit types

Do guests with higher room rates tend to cancel or no-show more often?

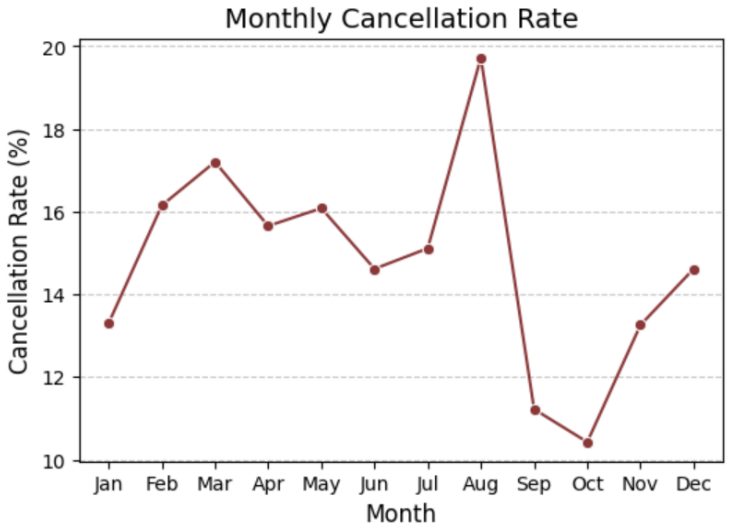


The median room rate is similar across all three group, The range of room rates (min to max) is nearly identical, meaning cancellations occur across all price levels. No significant outliers or skewed distributions in any category

To confirm whether room rate differences are statistically significant, we performed an ANOVA test: Since the p-value is greater than 0.05, we fail to reject the null hypothesis. Room rates do not significantly impact whether a guest cancels, checks out, or no-shows, rice is NOT a strong predictor of cancellations or no-shows

Seasonality Impact on Cancellations

The line chart represents the monthly cancellation rate (%) throughout the year, the cancellation rate steadily increases, peaking at around 17% in March, A sharp spike to nearly 20%, the highest cancellation rate of the year. This suggests that many people book summer vacations but cancel last minute and A dramatic drop to below 12%, possibly due to off-season stability A gradual rise, likely due to holiday season plans changing. let's link the chart with the Chi-Square analysis.



```
Chi-Square Test Results:
Chi-Square Statistic: 190.85
Degrees of Freedom: 22
P-value: 0.00000
```

Based on the chi square results the p values are less than 0.005 we can reject the null hypothesis, and we can conclude that Reservation status (whether a booking is canceled, checked out, or no-show) is significantly influenced by the month.

Understanding Revenue Loss Due to Cancellations and No-Shows

By calculating Revenue and Revenue with Discounts, we could estimate the financial impact of cancellations and no-shows based on the lost revenue.

Every hotel booking holds revenue potential. However, cancellations and no-shows significantly impact the bottom line. Our objective was to quantify this impact and understand the trends behind revenue loss. To determine the financial impact, we leveraged the dataset containing historical reservation data. Specifically, we focused on the Reservation Status column to filter out reservations that were either "Canceled" or "No-Show", We then summed the Revenue and Revenue_with_disc values associated with these bookings to estimate.

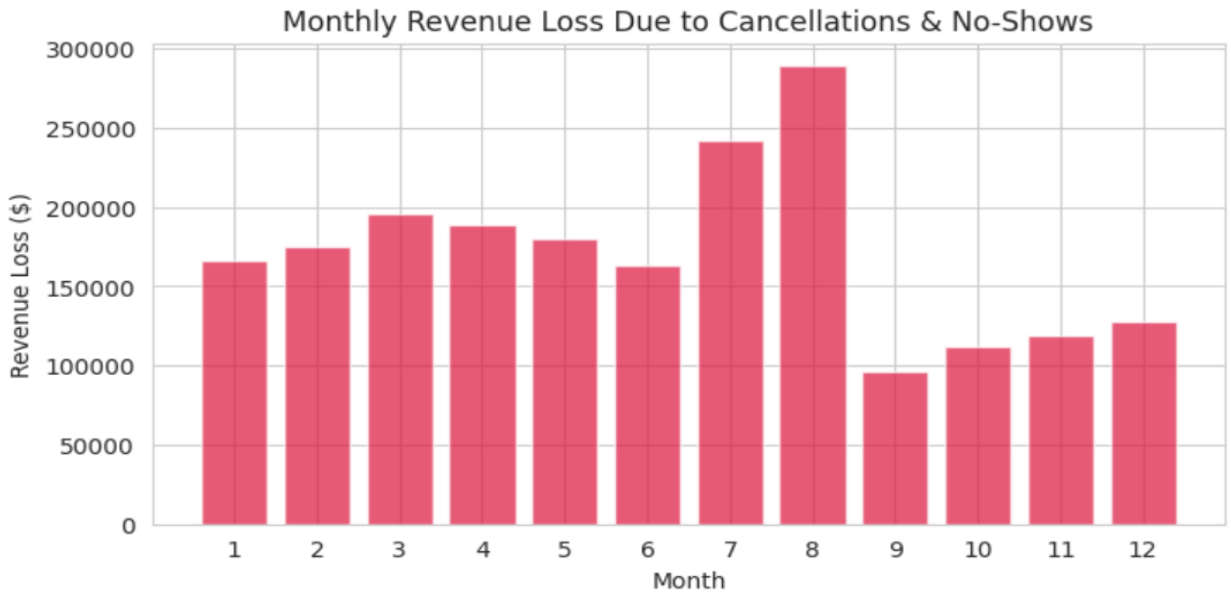
- Total revenue loss without discounts
- Total revenue loss after applying discounts

The findings are

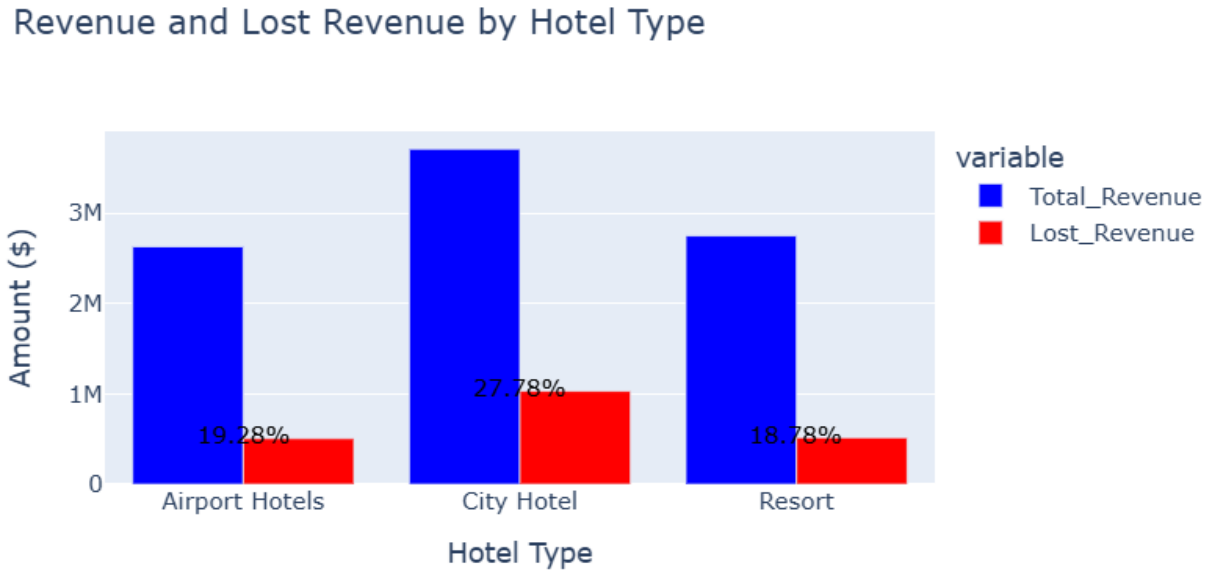
• Total revenue loss without discounts	$2,054,757.77
• Total revenue loss after applying discounts	$1,857,965.52

These figures indicate that cancellations and no-shows resulted in a direct revenue loss exceeding $2 million, with the application of discounts reducing the loss to approximately $1.85 million. This loss highlights the importance of strategies to mitigate cancellations, as the financial impact is significant, even with the application of discounts.

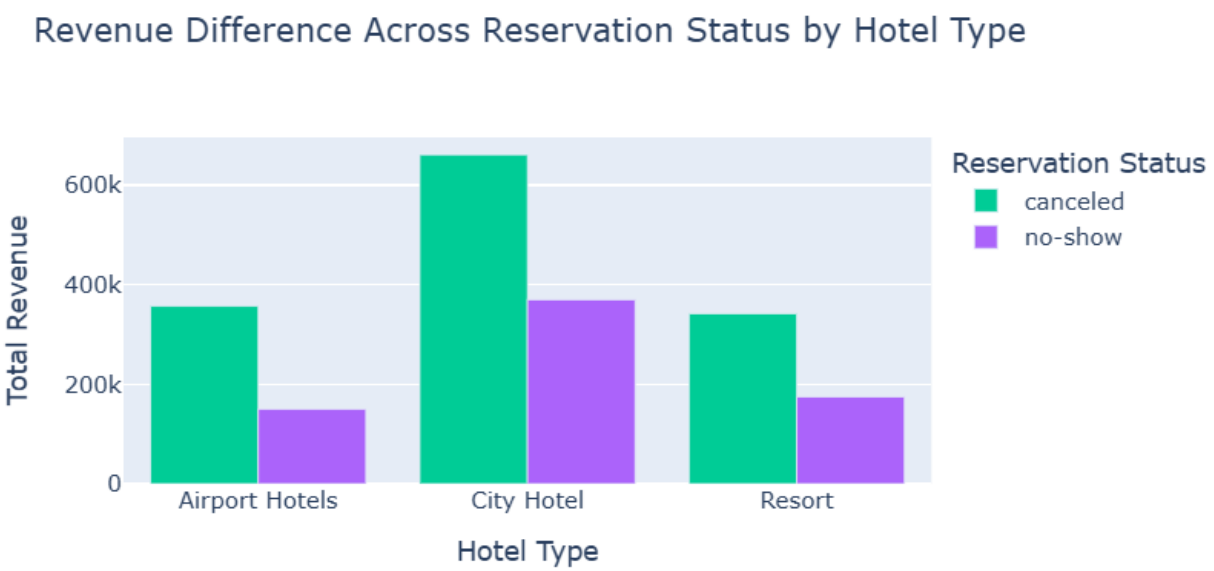
Next, we analyzed revenue loss across different months to see when cancellations and no-shows were at their highest, based on the below graph we can observe July and August saw the highest revenue loss, possibly due to seasonal travel fluctuations and high booking demand leading to more last-minute cancellations, September recorded the least revenue loss, potentially due to off-peak travel periods.



We further analyzed how different hotel types were impacted by revenue loss



City Hotel	$1,031,132
Airport Hotels	$507,100
Resort Hotels	$516,525



City hotels experienced the highest revenue loss, likely due to frequent business travelers who may cancel last-minute due to schedule changes and Airport hotels also had significant losses, suggesting a link between flight cancellations and hotel cancellations. Resorts had slightly lower losses, possibly because vacationers are more committed to their trips compared to business travelers

Predictive Model

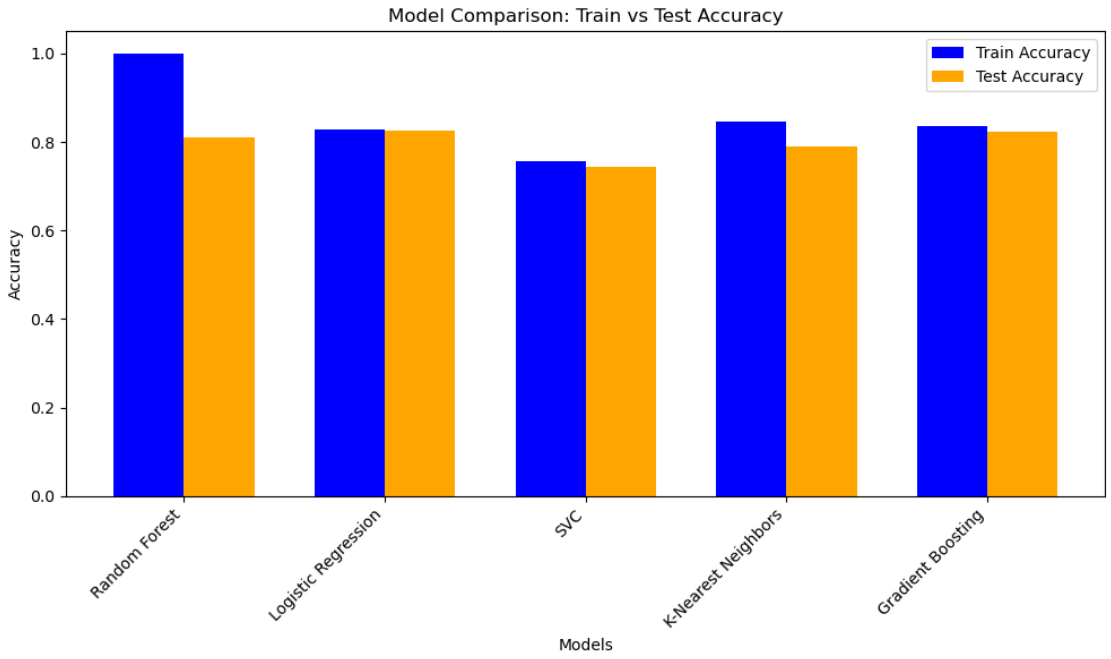
The target variable is Reservation_Status, which takes on three values:

1 – Confirmed reservation,

2 – Canceled reservation,

3 – No-show.

The initial evaluation of the models without SMOTE showed varied performance. The Random Forest model achieved an accuracy of 82%, with a precision of 0.67 and a recall of 0.56. The confusion matrix revealed a tendency to misclassify some of the minority classes, especially the no-show category. Logistic Regression had a slightly higher accuracy of 84%, with a precision of 0.72, but similarly struggled with the minority class. The SVC model achieved an accuracy of 76%, and KNN performed with an accuracy of 81%. Gradient Boosting, another powerful model, achieved an accuracy of 84%, showing competitive performance compared to Logistic Regression.



Without oversampling, the class imbalance in the dataset had a significant impact on the model's performance, particularly in terms of the F1-score. The dataset was heavily skewed towards the majority class, leading to a bias in predictions. This imbalance caused the models to perform poorly in predicting the minority classes, especially the no-show category. As a result, the F1-scores for the minority classes were substantially lower, reflecting the models' inability to correctly classify these instances. The overall performance, measured by accuracy, was decent, but the imbalance heavily skewed the precision and recall for the underrepresented classes, resulting in a suboptimal F1-score. This issue was particularly noticeable in models such as Logistic Regression and SVC, where the lack of class balance led to poor recall for the minority classes, thereby diminishing the effectiveness of the models in practical, real-world applications.

When SMOTE was applied to the dataset to balance the class distribution, the models showed improved performance. The Random Forest model, after hyperparameter tuning, emerged as the best performer, with an accuracy of 92.92% on the resampled data. The tuned model demonstrated a balanced precision, recall, and F1-score for all three classes, outperforming other models in terms of overall accuracy and class balance. Logistic Regression, despite benefiting from SMOTE, still performed lower than Random Forest, with an accuracy of 76%, while KNN, SVC, and Gradient Boosting also showed improvements but did not surpass Random Forest.

Hyperparameter Tuning

To optimize the performance of the Random Forest model, hyperparameter tuning was applied using GridSearchCV with a 5-fold cross-validation approach. A total of 216 candidates were tested, resulting in 1080 individual fits. The best hyperparameters identified through this process were: bootstrap: False, max_depth: None, min_samples_leaf: 1, min_samples_split: 2, and n_estimators: 200. These parameters significantly improved the model's performance, yielding a test accuracy of 92.92% on the resampled data. The classification report for the optimized model demonstrated strong results across all classes, with precision and recall consistently high for each class. Specifically, Class 1 achieved a precision of 0.94 and recall of 0.90, Class 2 achieved a

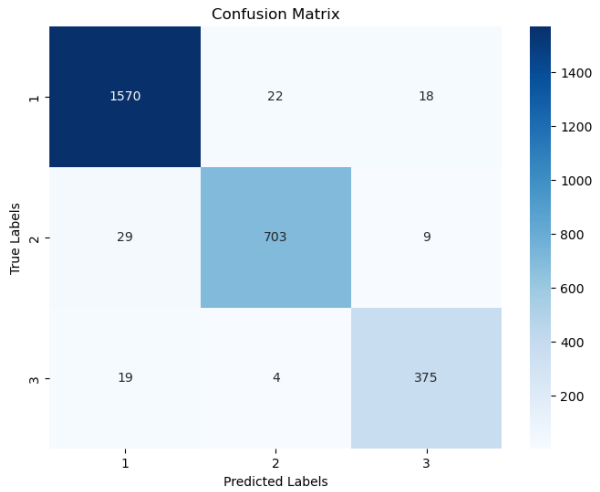
precision of 0.91 and recall of 0.93, and Class 3 achieved a precision of 0.93 and recall of 0.96. Overall, the model performed well with balanced results across the classes, as reflected by the macro and weighted average scores. This hyperparameter tuning process successfully enhanced the model's ability to classify all categories accurately, addressing earlier challenges caused by class imbalance.

Model

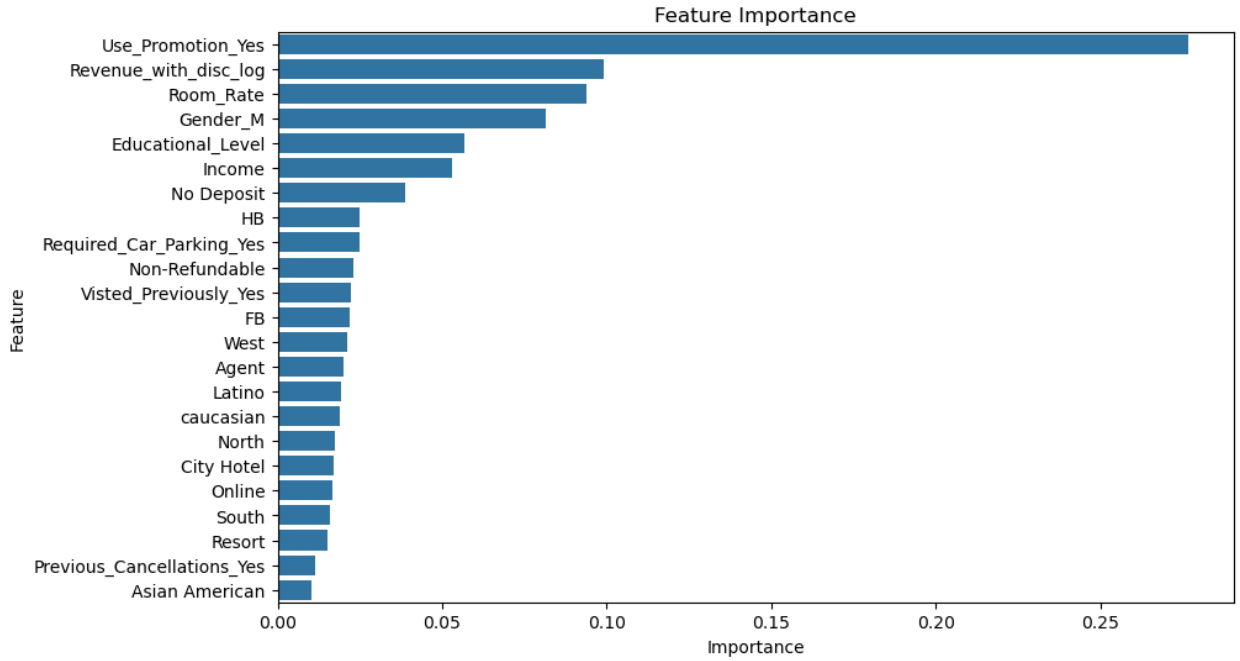
Classification Report:

	precision	recall	f1-score	support
1	0.97	0.98	0.97	1610
2	0.96	0.95	0.96	741
3	0.93	0.94	0.94	398
accuracy			0.96	2749
macro avg	0.96	0.96	0.96	2749
weighted avg	0.96	0.96	0.96	2749

The model performed exceptionally well on the new dataset, achieving an accuracy of 96.33%. The classification report reflects strong performance across all three classes. Specifically, for Class 1, the model achieved a precision of 0.97, recall of 0.98, and an F1-score of 0.97, demonstrating its ability to accurately predict the majority class. Class 2 had a precision of 0.96, recall of 0.95, and an F1-score of 0.96, indicating balanced performance in identifying this class as well. Class 3 saw slightly lower but still impressive values, with precision at 0.93, recall at 0.94, and F1-score at 0.94. The model's macro and weighted averages were both 0.96, confirming its overall robustness and balanced performance across all classes.

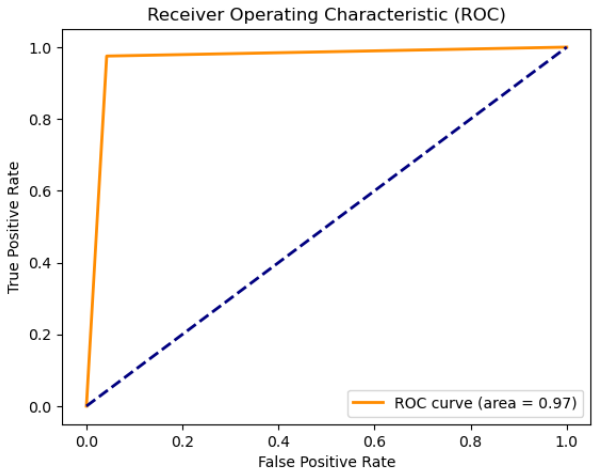


The confusion matrix provides further insight into the model's predictions. It shows that Class 1 had only 22 false positives and 18 false negatives out of 1610 total predictions, Class 2 had 29 false positives and 9 false negatives out of 741, and Class 3 had 19 false positives and 4 false negatives out of 398. These low values for false positives and false negatives further highlight the model's strong classification capabilities.



Additionally, feature importance analysis reveals that certain features significantly influenced the model's predictions. The most influential feature was "Use_Promotion_Yes," with an importance score of 0.2766, followed by "Revenue_with_disc_log" (0.0993) and "Room_Rate" (0.0939). Other important features included "Gender_M" (0.0814), "Educational_Level" (0.0569), and "Income" (0.0530). These findings suggest that promotional offers, pricing, and demographic information are critical factors in predicting reservation status.

Feature	Importance
Use_Promotion_Yes	0.276563
Revenue_with_disc_log	0.099261
Room_Rate	0.093922
Gender_M	0.081396
Educational_Level	0.056888
Income	0.052965
No Deposit	0.038831



An ROC (Receiver Operating Characteristic) value of **0.97** indicates excellent performance by the model. The ROC curve assesses the model's ability to distinguish between the positive and negative classes, with the **Area Under the Curve (AUC)** quantifying this ability.

Customer segmentation

The goal is to divide our customers into clusters and analyze those clusters in order to gain a better understanding of the hotel guests holistically and also based on cancellation behviour.

Data Preprocessing and Feature Engineering

To enhance our analysis and modeling, we performed multiple preprocessing and feature engineering steps to extract meaningful insights from the dataset.

1. **Time-Based Feature Extraction:**
   We extracted year, month, day, weekday, and week-of-year from key date columns (Expected_checkin, Expected_checkout, and Booking_date). These features help capture seasonality patterns, booking behaviors, and potential influences on cancellation rates.

2. **Stay Duration Analysis:**
   We calculated the number of weekdays and weekends within each booking's stay duration. Since hotel demand can vary significantly between weekdays and weekends (e.g., business travelers vs. vacationers), this distinction allows us to analyze customer preferences and trends in cancellation behavior.

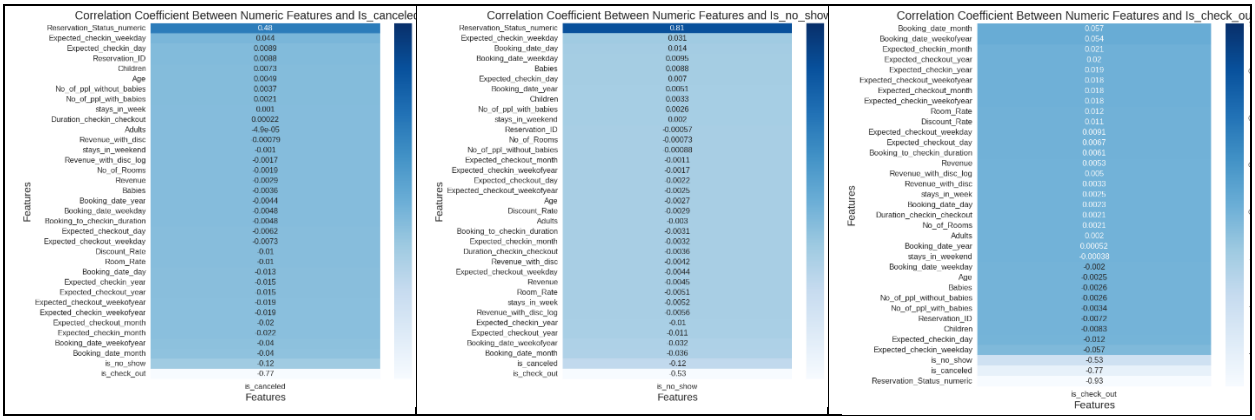
3. **Handling Negative Booking Durations:**
   Some records showed negative values for Booking_to_checkin_duration, indicating potential data entry errors or inconsistencies. To correct this, we replaced such erroneous Expected_checkin values with Booking_date and recalculated the duration. This ensures that all booking records maintain logical consistency.

4. **Revenue Calculation:**
   We computed total revenue per booking based on room rate, stay duration, and the number of rooms reserved. Additionally, we incorporated discounts to derive Revenue_with_disc, reflecting actual revenue post-discount. To normalize revenue

distribution and mitigate skewness, we applied a log transformation (Revenue_with_disc_log). This transformation enhances model performance by reducing extreme variations.

Through these steps, we refined the dataset for better predictive performance and analytical insights, ensuring accurate and reliable customer segmentation based on cancellation behavior.

Feature importance on cancellation behavior

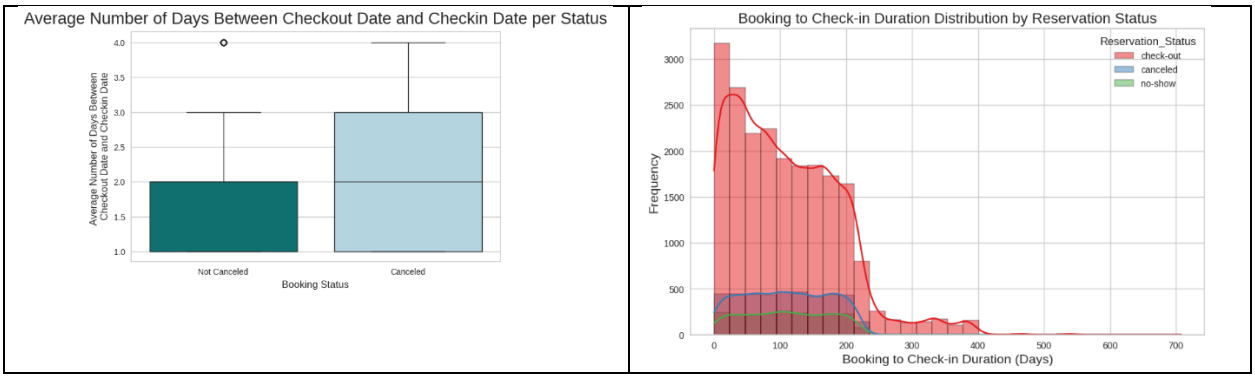


The target variables were encoded such as is_canceled, is_checkout and is_noshow for individual analysis. For is_canceled, the most influential feature is Expected_checkin_weekday (0.04), suggesting that the day of the week impacts cancellation likelihood. However, other booking-related features like Booking_to_checkin_duration (-0.0048) and Discount_Rate (-0.01) have weak correlations, indicating minimal influence.

For is_no_show, Expected_checkin_weekday (0.03) has a slightly positive correlation, implying that check-in timing might play a role in guest behavior. Other features like Booking_date_day (0.01) and Babies (0.008) also show weak positive correlations.

For is_check_out, booking-related factors like Booking_date_month (0.057) and Booking_date_weekofyear (0.054) have the highest positive correlations, suggesting a seasonal trend in confirmed stays. Expected_checkin_weekday (-0.056) negatively correlates, indicating that weekday check-ins might reduce the likelihood of successful check-outs.

Overall, booking patterns, seasonal trends, and minor demographic factors contribute to reservation outcomes, but their influence is relatively weak.



The lead time for canceled bookings follows a normal distribution but has minimal influence on cancellations. Interestingly, longer expected stay durations increase cancellation probability. A Box plot reveals that canceled bookings have a wider spread, with a maximum expected stay of 4 days, compared to 3 days for non-canceled bookings. This suggests extended stays may heighten uncertainty or risks, leading to cancellations.

Clustering can be approached by segmenting bookings based on influential features like expected check-in weekday, booking lead time, and expected stay duration. Since seasonal trends and minor demographics impact outcomes, clustering can group guests by booking behavior, cancellation risk, and check-out likelihood, enabling tailored strategies to optimize retention and reduce no-shows.

Clustering

The approach used for preparing the dataset for clustering is robust, combining several preprocessing and dimensionality reduction techniques to make the data suitable for clustering algorithms. The first step involves handling categorical variables: ordinal encoding is applied to columns such as "Educational_Level" and "Income," while one-hot encoding is used for nominal variables like "Country_region" and "Hotel_Type." This ensures that all features are represented numerically, which is crucial for machine learning models.

Next, the data is standardized using StandardScaler, which normalizes the feature range and centers the data, ensuring that all features have equal weight in clustering algorithms. This step is particularly important when using algorithms like KMeans, which rely on distance metrics.

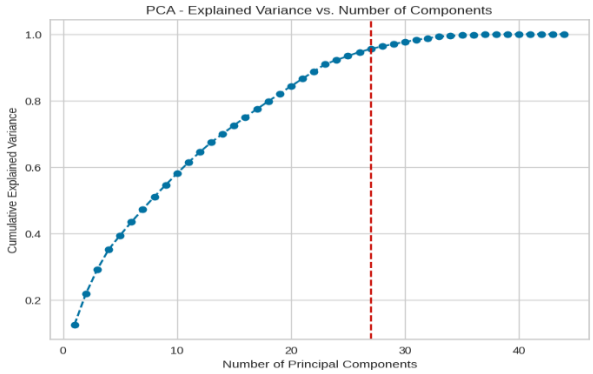
Dimensionality reduction through PCA is applied to capture the most important features while reducing the complexity of the dataset. By retaining 95% of the variance, PCA allows the algorithm to focus on the most informative components, thus improving computational efficiency and potentially clustering performance. And finally UMAP and t-SNE dimensionality reduction are used for visualizations purposes only.

Models used

KMeans partitions data into k clusters by minimizing within-cluster variance, ideal for well-separated, spherical clusters. DBSCAN groups points based on density, marking noise as outliers, but requires careful parameter tuning. Agglomerative Clustering builds hierarchical clusters, suitable for small to medium datasets, but struggles with high-dimensional data. Gaussian Mixture Models (GMM) assumes clusters follow Gaussian distributions, working well for such datasets, but may underperform with non-Gaussian data. Spectral Clustering uses graph Laplacian to transform data for non-linear separations and is effective with high-dimensional data after PCA. KMeans and Spectral Clustering are best for this dataset, while DBSCAN and Agglomerative Clustering may struggle.

Models Evaluation

Initially the models are evaluate for all three reservations status. Since we are analyzing cancellation behavior, the cancelled reservation status are filters for segmentation.



Out of 47 variable, PCA resulted in 27 component retaining 95% of variance. The pca_data is used for clustering and the results were

Clustering Algorithm	**Silhouette Score**	**Calinski-Harabasz Index**	**Davies-Bouldin Index**	**Notes**
KMeans	0.0880	287.69	2.64	Moderate silhouette score indicating average cluster cohesion. The Calinski-Harabasz index suggests decent compactness, while the Davies-Bouldin index indicates moderate within-cluster variance.
Agglomerative Clustering	0.0662	239.02	2.96	Low silhouette score indicating poor separation of clusters. The Calinski-Harabasz index is lower, suggesting less distinct clusters. The Davies-Bouldin index is higher, indicating more variance within clusters.
Gaussian Mixture Models (GMM)	0.0786	270.58	3.23	The GMM has a lower silhouette score, showing poor separation. The Calinski-Harabasz index is decent but not optimal, and the Davies-Bouldin index is the highest among all algorithms, indicating poor clustering structure.
Spectral Clustering	0.1026	321.16	2.46	Spectral clustering provides the highest silhouette score and Calinski-Harabasz index, indicating better separation and compactness compared to the other algorithms. The Davies-Bouldin index is

Clustering Algorithm	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index	Notes
				still high, but it is the best among the algorithms.

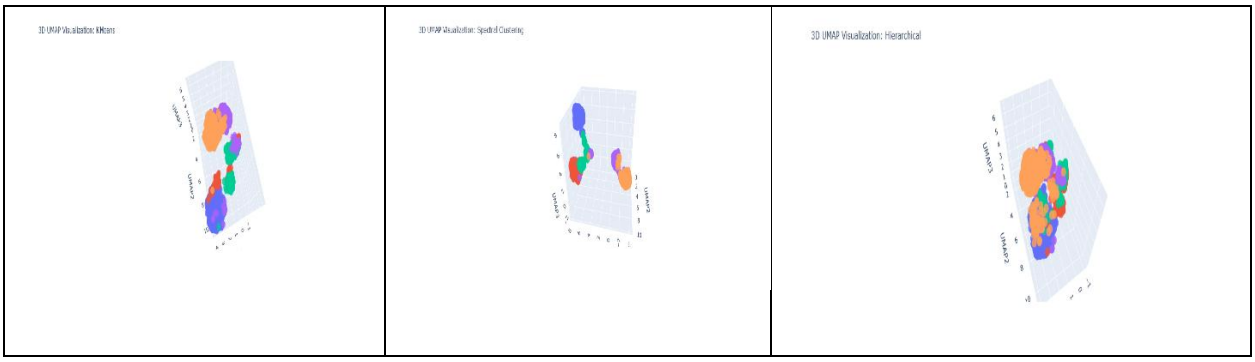
Recommendation:

1. **Silhouette Score**: **Spectral Clustering** performs the best with the highest silhouette score (0.1026), indicating the best cluster separation.
2. **Calinski-Harabasz Index**: **Spectral Clustering** also has the highest value (321.16), showing its ability to form compact and well-separated clusters.
3. **Davies-Bouldin Index**: While still high for all models, **Spectral Clustering** performs better than others with the lowest Davies-Bouldin index (2.46).

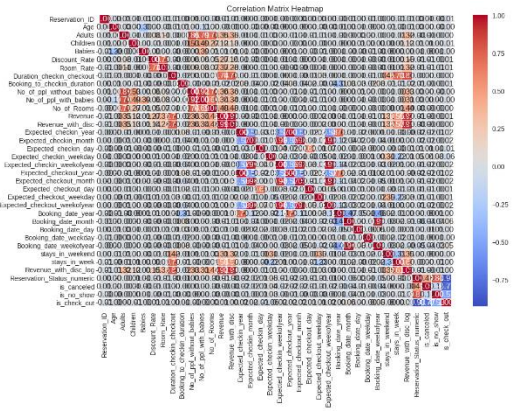
Given these metrics, **Spectral Clustering** emerges as the most reliable algorithm for this clustering task, with decent cluster separation and compactness. **KMeans** may still be considered but would likely need further tuning to improve its performance. **Agglomerative Clustering** and **GMM** show less promising results based on the metrics.

How Spectral clustering perfoms?

Spectral Clustering performs well on dataset because of its ability to handle **complex structures** and **non-linear separations** in the data. After PCA, the data is reduced to a lower-dimensional space, but the relationships between points may still be complex. Spectral Clustering can capture these relationships effectively. Unlike KMeans (which assumes spherical clusters) or GMM (which assumes Gaussian distributions), Spectral Clustering does not make strong assumptions about the shape or size of clusters.



The 3d plot of Kmeas, spectral and Agglomorative shows how the labeling has been done where in spectral , the labels are much distinct compared to other two with 5 clusters. This could be due to spectral exploiting non linear relationship and the variables also not much correlated depicted by the corr plot.



ANOVA was performed on a **booking cancellation dataset**, the insights can be interpreted based on how features relate to the likelihood of cancellations and cluster grouping. Below are the insights based on the results:

Feature Category	**Examples**	**Insights**
Highly Significant	Deposit_type, Room_Rate, Booking_to_checkin_duration, Hotel_Type	These features strongly influence clustering. High Room_Rate and Deposit_type indicate high-spending customers. Long Booking_to_checkin_duration suggests early planners.
Moderately Significant	Duration_checkin_checkout, Booking_channel_Direct	These features contribute moderately. Duration_checkin_checkout reflects stay length, and Booking_channel_Direct indicates booking preferences.
Non-Significant	Educational_Level, Income, Babies, stays_in_weekend	These features do not differentiate clusters. They can be removed to simplify the clustering process.
Constant/Undefined	cluster_dbscan, cluster_spectral	These features are constant or have no variability, making them irrelevant for clustering.

Based on these significant variables, cluster characteristics are identified explained in discussion.

Discussion

Business Implications

The analysis of cancellation rates and booking durations provides valuable insights for the hotel business, particularly regarding the impact of short-term stays and peak seasons. Shorter stays, especially during peak months like August, are more likely to be canceled, suggesting that guests with uncertain or last-minute plans are prone to cancellations. This insight could inform business strategies, emphasizing the need to create flexible booking policies for shorter stays or consider implementing incentives to encourage guests to honor these bookings.

Additionally, the significant spike in discounts during October, accompanied by a low cancellation rate, highlights the effectiveness of promotions in keeping reservations intact. This suggests that the hotel should explore more targeted discount strategies, especially during peak seasons, to minimize cancellations. However, the increase in cancellations in November despite lower discounts indicates that factors other than price—such as guest satisfaction or external influences like events or weather—may also play a significant role. Therefore, the business should also consider non-price-based strategies, such as personalized customer engagement or loyalty incentives, to reduce cancellations.

The correlation analysis supports these patterns, showing that discounts are strongly correlated with room rates but do not significantly impact when or how long guests book their stays. This insight suggests that while discounts may help retain guests, other factors such as guest preferences or hotel-specific policies could be more influential in determining booking behaviors. The hotel should focus on refining its offerings to better align with guest preferences and improve the overall guest experience.

The analysis also reveals that income level does not play a significant role in determining the room rates booked by guests. Despite the assumption that higher-income guests would opt for more expensive rooms, the data suggests otherwise. This finding implies that price segmentation based on income might not be as effective as anticipated. Instead, focusing on other guest characteristics, such as preferred room types or hotel amenities, could yield better results in tailoring offerings to different market segments.

Regarding the impact of hotel type on cancellations and no-shows, the Chi-Square test confirms a significant relationship between hotel type and reservation status. This finding suggests that certain hotel types, such as Airport Hotels, are more prone to cancellations than others like Resorts. The business can leverage this insight by adjusting strategies to reduce cancellations at high-risk hotels, such as offering non-refundable deposits or implementing stricter cancellation policies. Tailoring cancellation management strategies to the unique characteristics of each hotel type will likely enhance the overall efficiency of hotel operations.

Although booking lead time (from booking to check-in) appears to be linked to cancellations and no-shows at first glance, the ANOVA test results show no statistically significant relationship. This suggests that rather than focusing on enforcing strict policies for last-minute bookings, the hotel could benefit from offering **flexible rebooking options** or encouraging guests to modify their stays rather than canceling. This approach could help mitigate potential revenue loss from last-minute cancellations and improve customer satisfaction.

Lastly, the analysis of deposit types shows a significant relationship between deposit type and cancellation rates. **Non-refundable deposits** are linked to fewer cancellations, while both **refundable deposits** and **no-deposit bookings** have higher cancellation rates. The higher no-show rate associated with no-deposit bookings indicates that customers are more likely to skip their reservations when no financial commitment is required. To reduce cancellations and no-shows, the business should consider revising its deposit policies, offering both refundable and non-refundable options to cater to different guest preferences while ensuring a commitment to the booking.

Finally, the seasonal variations in cancellation rates underscore the importance of understanding **seasonality**. With a significant relationship between reservation status and the month of the year, the business can use this insight to better anticipate periods of high or low cancellations and optimize booking strategies accordingly. By aligning marketing campaigns and operational plans with seasonal trends, the hotel can maximize revenue and minimize cancellations during peak and off-peak seasons.

Predictive Model

The model achieved strong results across all performance metrics, demonstrating its high ability to classify instances correctly. The **accuracy of 96.33%** reflects the model's general robustness, while the precision, recall, and F1-score across the three classes further reinforce its effectiveness in both identifying the correct class and minimizing false positives and false negatives. Class 1 had the highest performance, with precision and recall both near 1.0, which indicates that it was very successful at predicting this class. Class 3 had slightly lower scores, which is typical in imbalanced classification tasks, but the model still maintained respectable performance with an F1-score of 0.94.

The **ROC AUC score of 0.97** confirms the model's strong ability to separate between classes, with an excellent tradeoff between sensitivity (recall) and specificity (1 - false positive rate). The higher the AUC, the better the model is at distinguishing between the classes, and a score of 0.97 places the model in the "excellent" category for classification tasks.

Overall, the predictive model is highly effective for the task at hand, achieving both high accuracy and strong class-wise performance, making it suitable for deployment in real-world scenarios. The feature importance analysis indicated that attributes like "Use_Promotion_Yes", "Revenue_with_disc_log", and "Room_Rate" were crucial to the model's decision-making process, suggesting that these features play a significant role in predicting the target variable. Further tuning and testing on other datasets could improve performance even more, especially for challenging classes.

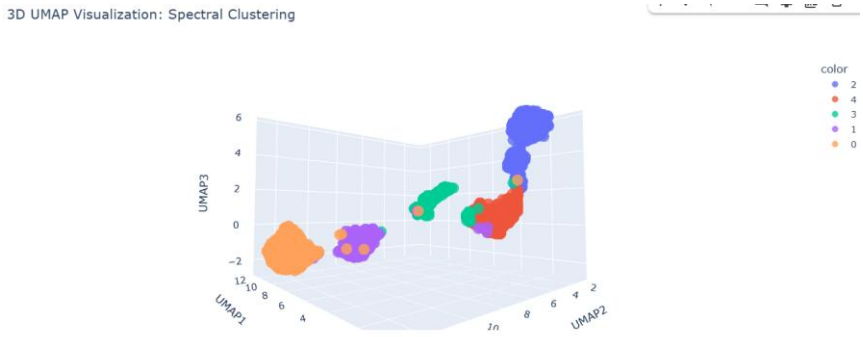
Random Forest is an ideal algorithm for interpreting hotel booking data, such as cancellations, no-shows, and check-ins, because it is a tree-based model that provides transparency into how decisions are made. Each decision tree within the Random Forest splits the data based on different features, making it easy to trace how specific factors contribute to the final prediction. This tree structure inherently supports interpretability, as the decision-making process can be visualized at each node, where feature values are used to guide decisions.

The feature importance analysis further enhances this interpretability by showing the relative impact of each feature on the model's predictions. In this case, the most important feature is **Use_Promotion_Yes**, with an importance score of 0.2808, indicating that promotional offers

significantly affect booking behavior, including cancellations and no-shows. Following closely are **Gender_M** (0.0936) and **Revenue_with_disc_log** (0.0932), which suggest that demographic factors and revenue adjustments (like discounts) also play crucial roles in influencing customer decisions. Other features such as **Room_Rate** (0.0882) and **Educational_Level** (0.0567) further indicate that pricing and customer demographics are also significant predictors.

The strength of Random Forest lies in its ability to aggregate multiple decision trees to produce a robust and accurate prediction while maintaining interpretability. Each tree provides a clear breakdown of how different features influence the outcome, and by examining the feature importance, hotel management can gain actionable insights into which factors are driving cancellations or no-shows. This transparency helps inform better operational strategies, such as refining promotional offers or adjusting pricing, ultimately leading to more data-driven decision-making.

Customer Segmentation Based on Cancellation Behavior



The customer base has been segmented into five distinct clusters based on cancellation behavior, each exhibiting unique characteristics. The following provides an overview of each cluster, their key features, cancellation patterns, and recommended strategies to reduce cancellations and enhance customer retention.

Cluster 0: Budget-Conscious Families

Cluster 0 consists of 1,130 bookings and represents budget-conscious families primarily booking City Hotels. This segment is characterized by a diverse range of Latino families, with 429 out of 1,130 bookings attributed to this ethnic group. Most individuals in this cluster hold a college-level education (475 out of 1,130), and the majority have an income ranging from $25K to $50K (330 out of 1,130). The average booking includes 1.8 adults and 1.67 children, and the average revenue per booking is $279.98. Furthermore, these families tend to book well in advance, with an average booking-to-check-in duration of 78.88 days.

Cancellation behavior within this cluster is likely driven by financial constraints or the allure of better deals elsewhere, given the relatively low-income levels and moderate revenue per booking. To address these challenges, offering flexible payment plans or early-payment discounts could reduce financial strain and improve customer loyalty. Additionally, providing loyalty programs could encourage repeat bookings from this price-sensitive segment.

Cluster 1: Middle-Income Planners

Cluster 1 includes 901 bookings and represents middle-income planners who primarily book City Hotels. Similar to Cluster 0, the majority of customers in this segment are Latino (356 out of 901) and have a college-level education (354 out of 901). This group typically has an income ranging from $50K to $100K (266 out of 901), and bookings are made with an average of 1.96 adults and 1.75 children per booking. The average revenue per booking is slightly lower than that of Cluster 0, at $278.82, and the average booking-to-check-in duration is 106.24 days.

The cancellations observed in this cluster are likely influenced by changing plans or unforeseen circumstances, as indicated by the relatively higher income levels. Offering free cancellation policies or options to reschedule bookings could cater to this group's tendency to change plans. Sending timely reminders closer to the check-in date may also help mitigate last-minute cancellations.

Cluster 2: Low-Income Long-Term Planners

Cluster 2 represents 984 bookings from low-income families who exhibit long-term planning behaviors, often booking City Hotels far in advance, with an average booking-to-check-in duration of 146.51 days. A significant portion of this segment is Latino (380 out of 984), and most individuals are college-educated (390 out of 984). With incomes primarily below $25K (287 out of 984), this cluster faces financial constraints, and the average revenue per booking is $276.72. Bookings in this cluster tend to include 1.77 adults and 1.71 children on average.

Cancellations within this group may stem from financial instability or the inability to commit to long-term plans. Offering flexible payment plans or discounts for early payments could help alleviate financial stress, while loyalty programs may encourage customers to return despite their financial challenges.

Cluster 3: Large Group Coordinators

Cluster 3 consists of 332 bookings and is characterized by larger groups, with an average of 4.19 adults and 2.29 children per booking. These bookings, which primarily involve City Hotels, tend to generate higher revenue, with an average of $672.16 per booking. The customers in this cluster are predominantly Latino (158 out of 332) and have a college-level education (132 out of 332). The income level of this group is primarily in the range of $25K to $50K (104 out of 332), and bookings are typically made with an average duration of 110.53 days prior to check-in.

Cancellation behavior in this segment is likely influenced by coordination issues within large groups or changes in group dynamics. To improve the customer experience and reduce cancellations, offering group discounts, customized packages, or dedicated support for large bookings may help ensure smoother coordination and greater commitment from the group.

Cluster 4: Resort Seekers

Cluster 4 includes 787 bookings from middle-income families, predominantly booking resorts. Customers in this cluster are primarily of Asian American descent (551 out of 787) and have a college-level education (315 out of 787). Their income typically falls within the $25K to $50K range (237 out of 787). The average revenue per booking is $377.44, and customers generally book resorts well in advance, with an average booking-to-check-in duration of 108.12 days. These bookings also tend to involve 2.93 adults and 1.73 children on average.

Cancellations in this group are likely driven by changing preferences or the availability of better alternatives. To mitigate cancellations, it is recommended to highlight unique features of the resorts to reduce the likelihood of customers switching to alternatives. Offering exclusive perks, such as free upgrades or additional amenities, may also help retain these customers and reduce cancellations.

Conclusion

In conclusion, the hotel industry, particularly Hotel A with its diverse property types, faces a complex set of challenges when managing booking cancellations and no-shows, which have a significant impact on revenue, resource utilization, and guest satisfaction. Through data analysis and predictive modeling, several key insights were uncovered. Firstly, seasonal trends, pricing strategies, and customer deposit preferences all contribute to fluctuations in cancellation and no-show rates. By recognizing these patterns, the hotel can implement more targeted strategies, such as flexible payment options, early-booking incentives, and customized marketing campaigns, to mitigate the financial impact of cancellations.

The predictive models developed proved highly effective, achieving impressive accuracy and performance metrics. The model's ability to classify reservations into cancellations, no-shows, and check-ins with high precision (96.33% accuracy) suggests it can be a powerful tool in predicting and mitigating cancellations in the future. The classification model's strong performance across precision, recall, and F1-scores for all three classes reinforces its robustness in identifying cancellation risks and minimizing both false positives and false negatives. This model can be deployed to proactively identify at-risk bookings and allow for preemptive measures to be taken, such as targeted offers or reminders to guests.

Customer segmentation also played a critical role in understanding booking behaviors, particularly around cancellations. Clustering analysis, particularly using Spectral Clustering, successfully identified distinct customer profiles, with each cluster exhibiting unique cancellation patterns. By segmenting guests based on factors like booking lead time, demographic characteristics, and hotel type, the hotel can tailor its approach to each group. For example, offering flexible cancellation policies or loyalty programs for high-risk segments, or providing exclusive perks to encourage retention, could effectively reduce cancellations and no-shows.

Additionally, statistical tests confirmed the significant influence of factors like deposit types and hotel type on cancellations and no-shows, providing valuable insights for refining booking policies. The relationship between non-refundable deposits and fewer cancellations, for instance, suggests that the hotel could adjust its deposit structures to strike a balance between guest convenience and revenue assurance.

In light of these findings, the hotel should consider refining its policies around deposit types, booking flexibility, and seasonal trends to optimize both revenue and customer satisfaction. Implementing these data-driven recommendations, supported by predictive classification models and clustering insights, would not only minimize cancellations and no-shows but also strengthen the overall customer experience, driving long-term loyalty and enhancing operational efficiency.

Acknowledgement

We would like to express my sincere gratitude to the Society of Statistics UOSJ, for organizing this competition, providing an opportunity to apply data-driven approaches to real-world challenges. We would also like to acknowledge the hard work and dedication of all the participants, whose contributions and collaborative spirit made this experience both enriching and inspiring. Special thanks to my mentors, peers, and the data science community for their invaluable insights and guidance throughout this project. Lastly, I would like to thank the anonymous data sources used in this competition, which allowed for a deep and meaningful analysis.