Group 23 - Hotel Booking Cancellations

Uttam Sadashiva Gowda
Arizona State University
1237738283

Pratham Savjani
Arizona State University
1238290276

Rohith Vince Richard Arockiaraj
Arizona State University
1233514882

Abhi Sachdeva
Arizona State University
1221508080

Subject : DSE 501

Instructor : Rong Pan

Table of Contents

Executive Summary

Hotel booking dynamics have grown very unpredictable over the past years, not a small number of reservations getting canceled prior to check-in. For hotels, such volatility is not just a minor annoyance: Cancellations translate directly into realized revenue losses, cause disruption in staffing and scheduling planning and introduce variability in inventory and pricing decisions. This study examines reservation information from a City Hotel and a Resort Hotel in order to gain insight into the mechanism for cancellations, as well as suggestions that hotels can use to respond better.

The dataset contains over 119,000 bookings between 2015 and 2017, with the detailed information such as of customers characteristics, booking timing, pricing, deposit rules as well as stay patters. Upon cleansing the data and crafting several potential predictor variables, we applied EDA, hypothesis testing procedures as well as predictive modeling to identify anything interesting in patterns of cancellation.

Several consistent insights emerged:

- City Hotels get a higher cancellation than Resort Hotels, which means that city visitors are more flexible and have circulation of plans.
- Lead time is one of the most powerful predictors for cancellation. Reservations booked a long time ahead of the actual arrival date are much more likely to cancel compared to those booked closer in time, showing that uncertainty rises with length of planning horizon.
- The deposit type is the controlling factor. No-deposit bookings cancel at a much higher percentage than non-refundable reservations, and the low cancellation rate for non-refundables demonstrates clearly how financial commitment impacts guest behaviour.
- Customer segments behave differently. For the majority of cancellations, it is (individual) transient travelers; groups / contracts have been relatively insulated and predictable.
- Repeat guests are much more predictable - see Christmas room-availability as an example - with vastly lower cancellation rates compared to first-time clients, this highlights the importance of loyalty and a history.
- Families also stay longer, so much so that they represent revenue-rich segments; yet their cancellation rates aren't larger than other customer types.

In light of these findings, we constructed predictive models — Logistic Regression and Random Forest classifiers—to predict the likelihood of cancellation at the time reservations are made. The Random Forest method, being more able to capture non-linear relationships and interactions between the features, showed improved predictive performance with respect to the Logistic Regression model in all accuracy, precision and recall metrics; AUC was also improved.

From a managerial perspective, the findings may imply some practical implications. (This is why hotels can adopt more stringent cancellation or deposit policies for such higher-risk segments as long-lead-time bookings, no-deposit reservations and transient business, but provide more lenient ones to return guest and groups, which result in far fewer no-shows.) Shifting policies by season and customer type also reduces uncertainty. Together, these kinds of measures can help hoteliers steady occupancies, refine revenue forecasting and marshal operations more effectively.

Related Work / Literature Background

Hotel booking cancellations and demand uncertainty have been widely studied in hospitality and revenue management literature. Prior research consistently shows that cancellation behavior is influenced by booking timing, deposit policies, customer type, and booking channels. For example, Antonio et al. (2019) demonstrate that longer lead times increase cancellation likelihood because customers face more opportunities to change travel plans. Studies on revenue management also highlight that non-refundable rates significantly reduce cancellations by increasing the financial cost of backing out (Ivanov & Zhechev, 2012).

Work by Bock et al. (2021) shows that online travel agency (OTA) customers exhibit more flexible and speculative booking patterns, contributing to higher cancellation rates compared to direct bookings. Research by Zhong & Sun (2019) further indicates that repeated guests are more reliable and provide more stable revenue streams due to loyalty and lower cancellation tendencies.

Machine learning approaches have also been explored in this domain. Shapoval et al. (2020) demonstrate that tree-based models outperform linear models in predicting cancellations, aligning with our finding that Random Forest achieves superior accuracy and ROC-AUC.
Overall, the existing literature supports the behavioral, financial, and operational factors identified in our analysis and confirms the importance of prediction-based revenue strategies in hotel management.

Business Problem and Objectives

Hotel cancellations create a persistent gap between expected demand and realized occupancy. When a reservation that was anticipated to arrive is canceled—especially close to the arrival date—the hotel may not have sufficient time to resell the room. This is particularly damaging during off-peak periods, where demand is already limited. Over the course of a year, the cumulative effect of these lost opportunities results in lower utilization of inventory, greater volatility in occupancy levels. unpredictable revenue patterns.

This project is designed to help hotels understand and address the underlying factors that drive cancellations. Specifically, we aim to answer the following business questions:

- Which booking and customer attributes are most strongly associated with cancellation behavior?

  Examples include hotel type, customer segment, deposit policies, lead time. seasonality.

- How do cancellations relate to revenue-relevant variables such as ADR and stay duration?

  Understanding this helps identify which cancellations pose the greatest financial impact.

- Can cancellation risk be predicted early enough to inform operational and pricing decisions?

  This includes policies such as deposit requirements, overbooking strategies. flexible pricing structures.

We applied statistical hypothesis testing and two predictive models (Logistic Regression and Random Forest) to evaluate drivers of cancellation. The overarching objective is to convert raw booking data into actionable, segment-level insights that can help hotel managers tailor policies more precisely—moving away from generalized rules toward targeted, data-driven strategies that reduce risk and improve revenue stability.

Data Description

This analysis uses the publicly available Hotel Booking Demand dataset, which consolidates reservations from a City Hotel and a Resort Hotel between July 2015 and August 2017. The dataset contains 119,390 bookings and 36 variables, with each row representing a single reservation. The variables can be grouped into several business-relevant categories:

1. Booking Identification and Status
   - hotel – indicates whether the booking was made for the City Hotel or the Resort Hotel.
   - is_canceled – binary indicator of whether the booking was canceled (1) or fulfilled (0).
   - reservation_status and reservation_status_date – final booking outcome and the date on which that outcome was recorded.

2. Timing and Stay Characteristics
   - lead_time – number of days between the booking date and the scheduled arrival date.
   - arrival_date_year, arrival_date_month, arrival_date_day_of_month – detailed arrival date components.
   - stays_in_week_nights, stays_in_weekend_nights – number of nights the guest intended to stay on weekdays and weekends.

3. Guest Composition
   - adults, children, babies – number of guests in each category, used to understand party size and booking intent.

4. Pricing and Revenue Variables
   - adr (Average Daily Rate) – lodging revenue per night, calculated as the total cost of the stay divided by the number of nights.
   - deposit_type – indicates whether the reservation required No Deposit, Non Refund, or Refundable payment terms.

5. Customer and Market Characteristics
   - market_segment and distribution_channel – booking source, such as Online Travel Agency (OTA), corporate, direct booking, or tour operator.
   - customer_type – classified as transient, contract, group, or transient-party.
   - is_repeated_guest – identifies whether the guest has stayed at the hotel previously.
   - country, agent, company – origin and intermediary information where available.

The dataset includes a diverse mix of numerical, categorical, temporal, and behavioral variables, making it suitable for exploratory data analysis, statistical hypothesis testing, and supervised machine learning. This breadth allows us to examine cancellations from multiple angles—financial incentives, customer behavior, operational factors, and booking channel.

Data Cleaning & Preprocessing

Before conducting analysis or building predictive models, the raw dataset required several cleaning and preprocessing steps to ensure accuracy, consistency, and interpretability.

Initial Loading and Data Quality Checks

The dataset was loaded into a Pandas DataFrame, and we examined column types, descriptive statistics, and missing value counts. A few fields—most notably children, country, and agent—contained missing or inconsistent values. These issues needed to be addressed to avoid bias and ensure the reliability of downstream analysis.

Handling Missing Values

To preserve as much information as possible while maintaining data integrity, we applied the following strategies:

- Children: Missing values were replaced with 0, indicating that the booking did not include children.
- Country and Agent: Missing entries were filled with the label "Unknown". This allowed us to retain the affected bookings while signaling uncertainty.
- Company: This variable had extensive missingness and limited relevance for our analysis. Rather than impute arbitrary values, we removed this feature entirely.

These decisions ensured that the dataset remained complete without introducing distortions into the analysis.

Feature Engineering

To make the dataset more analytically useful, several new features were created:

- Total Stay Duration:
  duration_of_stay was computed as the sum of weekday and weekend nights. This produced a single, intuitive measure of stay length that is easier to interpret in statistical and predictive modeling.
- Family Indicator:
  A binary variable, is_family, was created and set to 1 if the booking included any children or babies. This allowed us to compare family vs. non-family behavior in a straightforward way.

- Arrival Date:
  The year, month, and day components were combined into a full date_of_arrival field to facilitate time-based analysis and seasonal trend detection.

- Season Variable:
  Months were mapped to seasons (Winter, Spring, Summer, Autumn), enabling us to explore broader seasonal patterns that might not be visible at the monthly level.

These engineered features helped bridge the gap between raw data and meaningful business insights.

Outlier Inspection

Using boxplots and descriptive statistics, we inspected variables with naturally wide ranges—such as adr, lead_time, and duration_of_stay—for outliers. Some extreme values (e.g., very long stays or unusually high ADRs) were identified.

Rather than removing these observations automatically, we retained them and treated them with caution during interpretation. Because these values may reflect genuine business scenarios (e.g., long corporate stays or high-price peak dates), excluding them could distort the analysis.

Pre-Processing for Modeling

To prepare the data for machine learning models, we defined separate preprocessing steps for numerical and categorical features:

- Numerical Features:
  Examples include lead_time, duration_of_stay, adr, and previous_cancellations.

- Categorical Features:
  These include hotel, market_segment, deposit_type, customer_type, is_repeated_guest, and others.

A ColumnTransformer was used to automate preprocessing:

- Numerical variables were standardized using StandardScaler, ensuring that features with large ranges did not dominate the model.

- Categorical variables were encoded using OneHotEncoder(handle_unknown="ignore"), which converts categorical values into binary indicator columns suitable for both Logistic Regression and Random Forest.

Embedding these steps into the modeling pipeline ensured that all data passed through the same reproducible transformations during training, validation, and prediction.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was used to build intuition around cancellation behavior, identify dominant patterns, and guide the choice of features for statistical testing and predictive modeling. The analysis was performed at both the univariate and bivariate levels, followed by a correlation overview and temporal trends.

Reservation Status

An initial inspection of the target variable showed that cancellations make up a substantial portion of all bookings as shown in Figure 1. The rate is high enough that cancellations cannot be treated as rare exceptions; rather, they represent a recurring operational reality. This underscores the importance of forecasting cancellation risk and designing policies that proactively mitigate its impact.

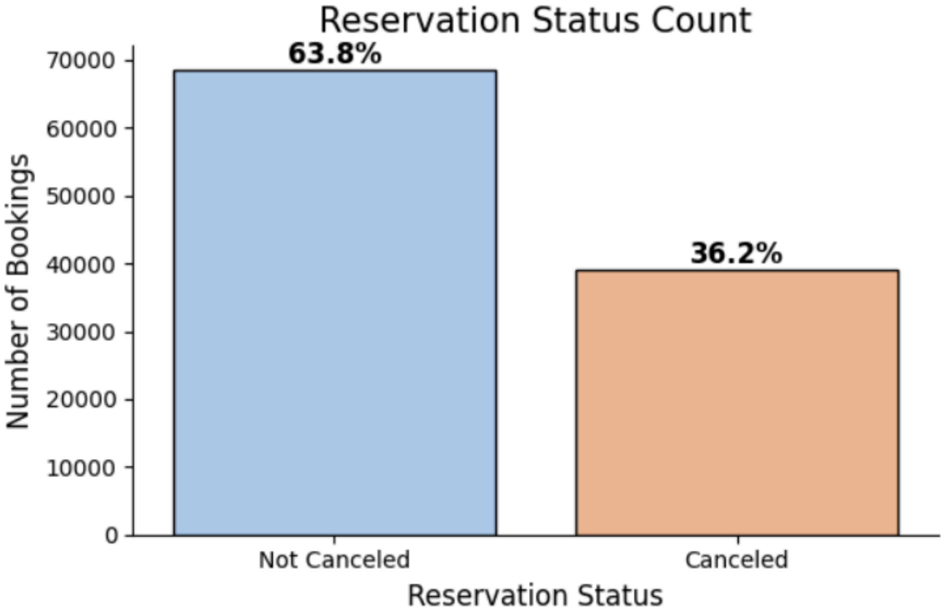


Figure 1 - Reservation Status Count

Lead Time

- The distribution of **lead_time** confirmed by Figure 2 is heavily right-skewed. Most guests book relatively close to their arrival date, while a long tail of reservations is made several months in advance.

- This asymmetry suggests that uncertainty increases sharply with planning horizon.

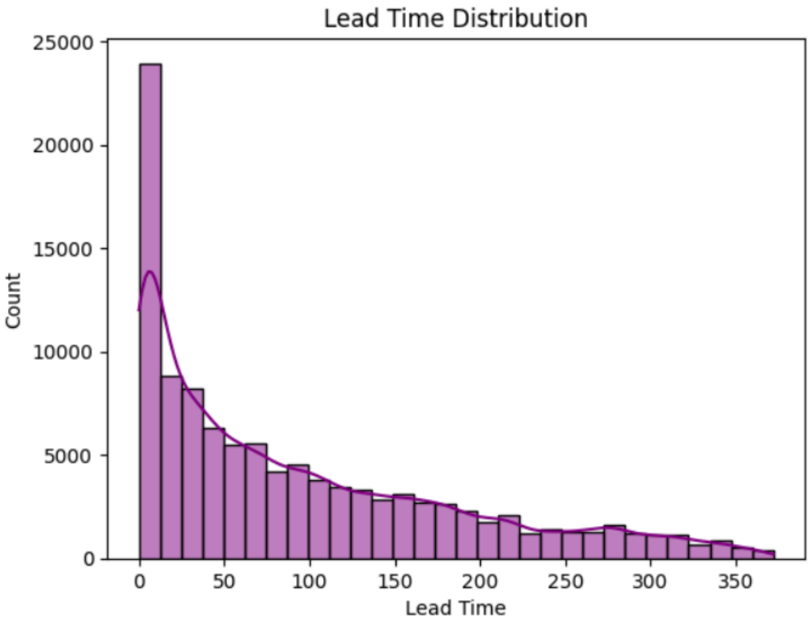


Figure 2 - Lead Time Distribution

Lead Time vs. Cancellation Rate

A line chart comparing lead time to cancellation rate shows a clear upward trend as in Figure 3:

- Short-lead bookings have comparatively low cancellation rates.
- Cancellation rates rise steadily as lead time increases, becoming noticeably higher beyond the 60–90 day window.

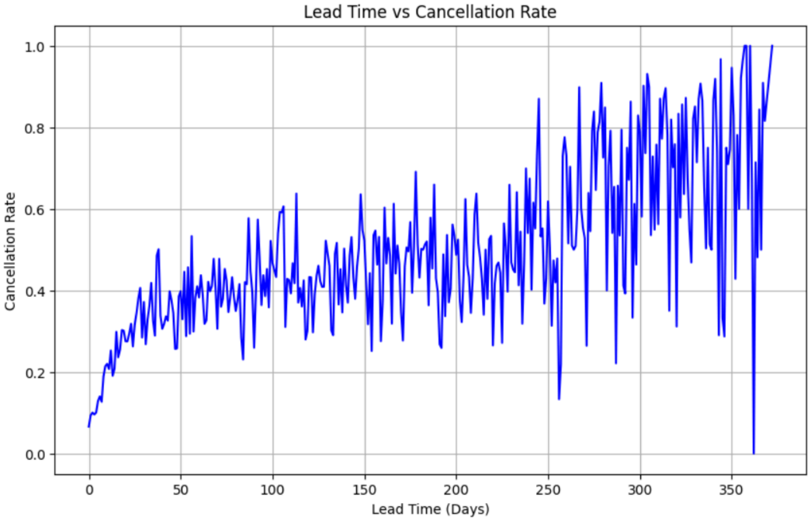


Figure 3 - Lead Time vs Cancellation rate

This pattern indicates that long-lead reservations carry disproportionately higher risk and should be treated differently from last-minute bookings.

Duration of Stay

- Most guests stay 1–3 nights, particularly in the City Hotel.
  Longer stays occur more frequently in the Resort Hotel, reflecting vacation-oriented travel as depicted in Figure 4.
- To streamline analysis, weekday and weekend nights were combined into **duration_of_stay**, which offers a single, interpretable measure of intended stay length.

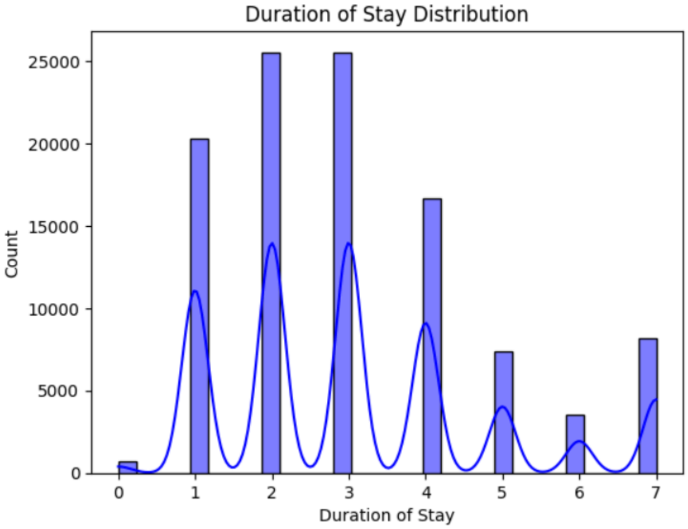


Figure 4 - Duration of Stay Distribution

Average Daily Rate (ADR)

- ADR displayed in Figure 5 has moderate right skew with a few high-priced outliers, which correspond to peak travel periods or premium room categories.
- City Hotels generally show higher ADR values compared to Resort Hotels, consistent with business-oriented pricing dynamics in urban environments.

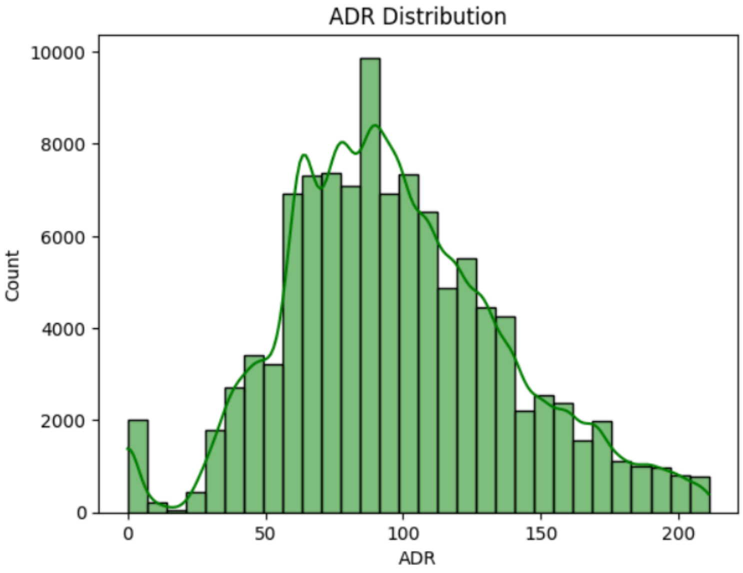


Figure 5 - ADR Distribution

This aligns with typical urban pricing behavior where business travel demands can elevate nightly rates.

Cancellations by Segment and Hotel Characteristics

Hotel Type

- City Hotels exhibit considerably higher cancellation rates than Resort Hotels as shown in Figure 6. This suggests that business or short-notice urban travel is more prone to change, reinforcing hotel type as a relevant predictor.

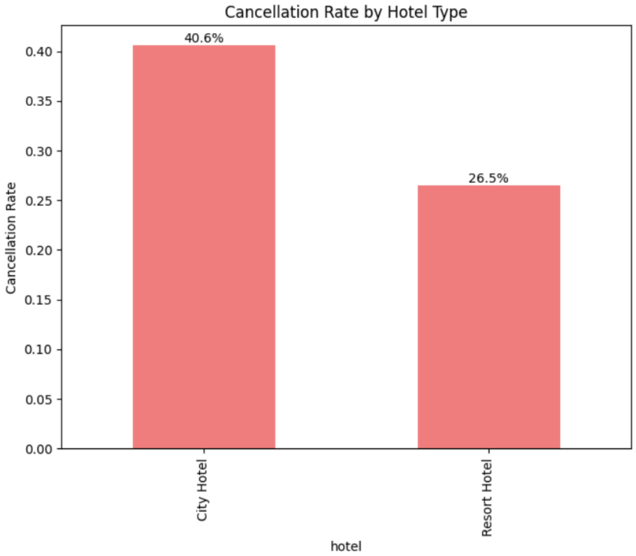


Figure 6 - Cancellation Rate by Hotel Type

Market Segment and Distribution Channel

A comparison of cancellation rates across market_segment is depicted in Figure 7. It shows pronounced differences:

- Online Travel Agencies (OTAs) contribute both a large volume of bookings and disproportionately high cancellation rates.
- Direct and corporate channels show markedly lower cancellation rates.

This mirrors well-known industry behavior in which OTA customers often make speculative reservations.

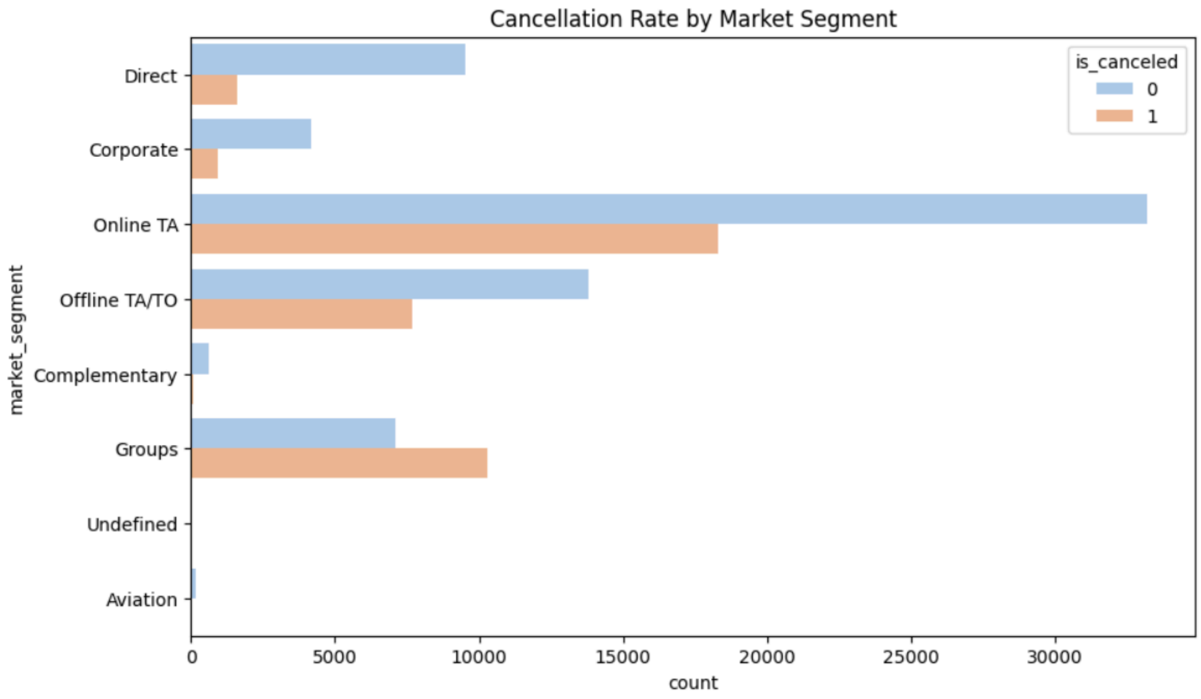


Figure 7 - Cancellation Rate by Market Segment

Customer Type

Cancellation behavior varies sharply by customer_type as shown in Figure 8:

- Transient guests account for most cancellations.
- Group and Contract customers show significantly lower volatility.
- Transient-party bookings fall between the two extremes.

These differences underscore the importance of customer segmentation in policy design.

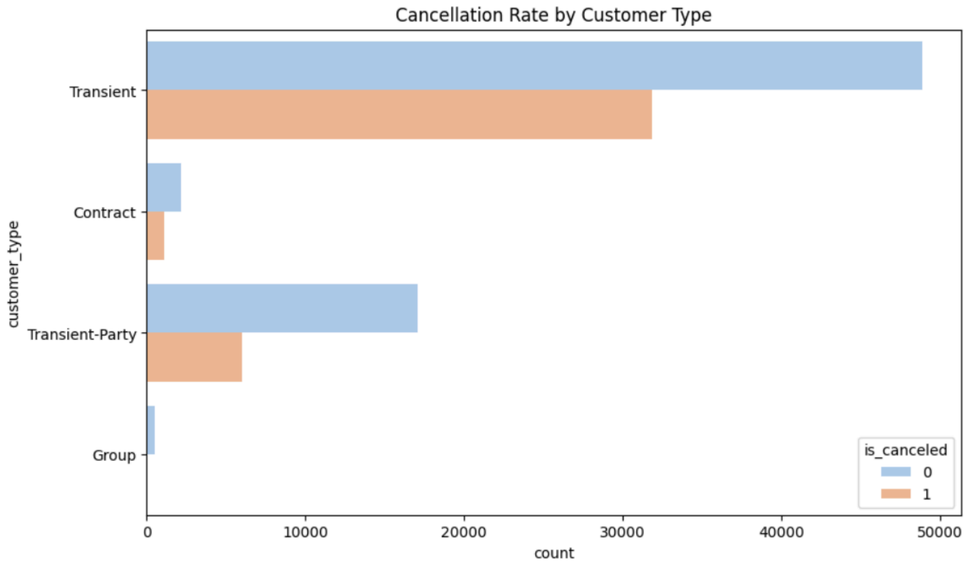


Figure 8 - Cancellation Rate by Customer Type

Deposit Type

Deposit rules remain one of the strongest behavioral drivers:

- No-deposit bookings cancel at very high rates.
- Non-refundable bookings rarely cancel.
- Refundable bookings exhibit moderate cancellation levels.

These distinctions highlight deposit type as a key lever for influencing guest commitment.

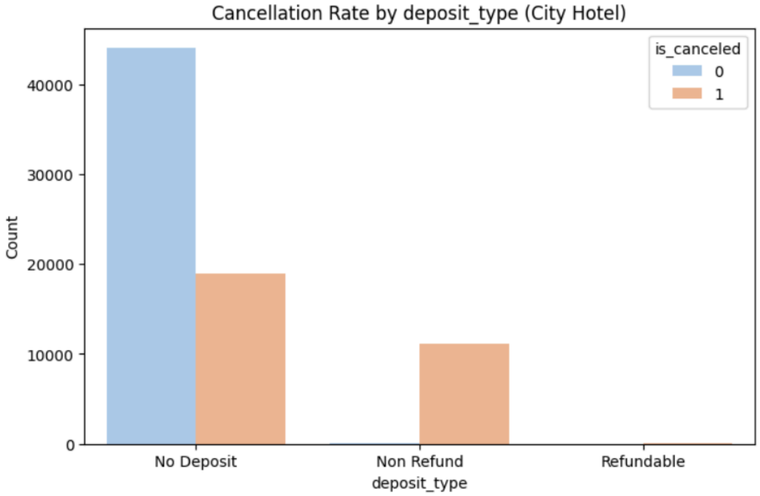


Figure 9 - Cancellation Rate by Deposit Type

Top Countries by Cancellations

A country-level breakdown is shown in Figure 10 and it conveys that a few countries account for a large proportion of cancellations, while many others contribute marginally.
This can support targeted marketing or communication for high-impact source markets.

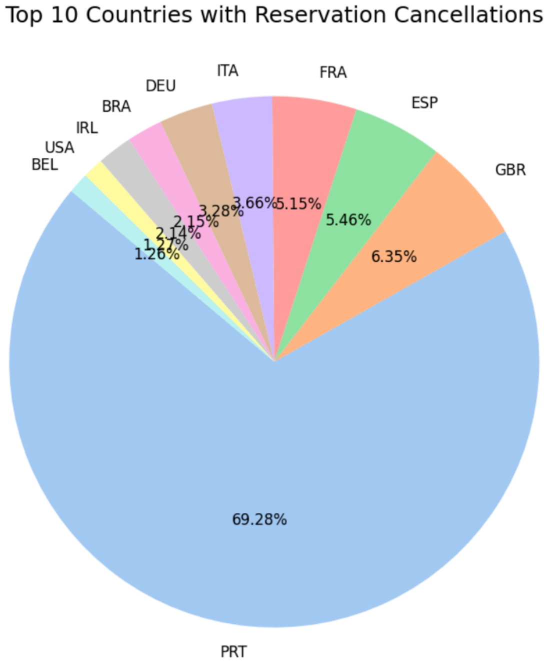


Figure 10 - Cancellation Distribution from different Countries

---

Seasonality

Season-level analysis reveals that:

- Spring and Summer experience higher cancellation rates.
- Winter bookings are more stable, possibly due to fixed holiday travel.

This justifies including season as a temporal variable in later modeling.

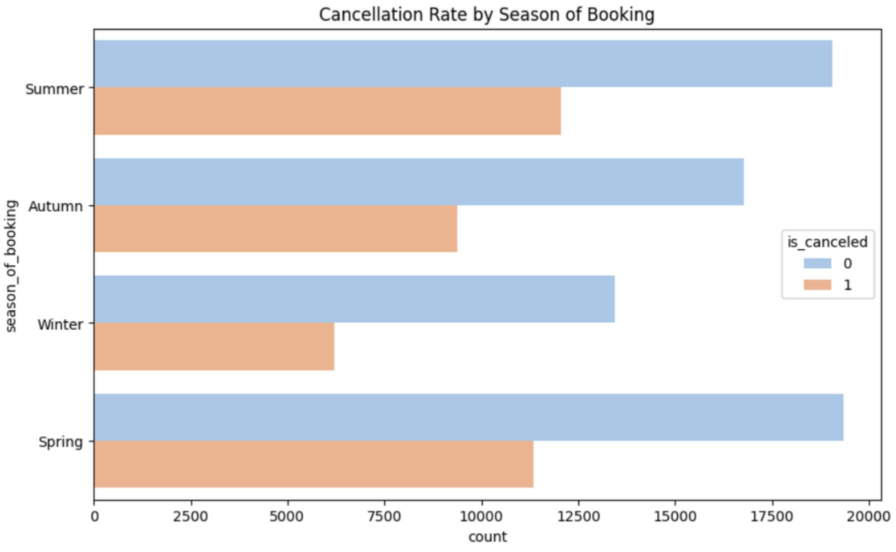


Figure 11 - Cancellation Rate by Season of Booking

Monthly Trends

Monthly cancellation volumes display (see Figure 12):

- Spikes during peak vacation months (July–August).
- Dips during off-season periods.

Hotels could use these seasonal fluctuations to calibrate deposit or overbooking strategies during high-risk months.

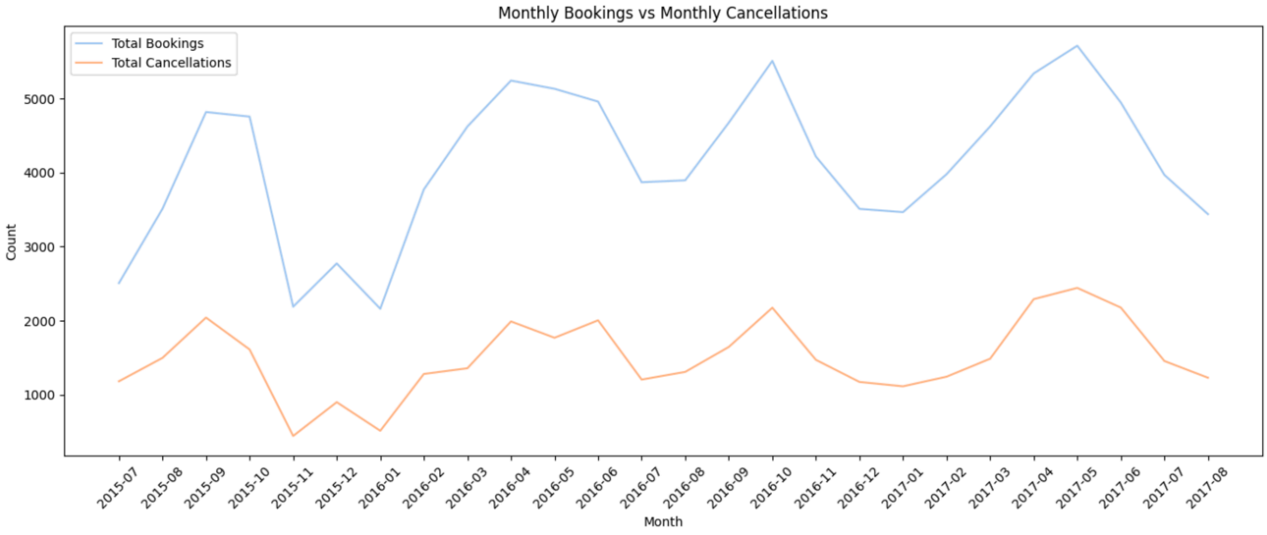


Figure 12 - Monthly Booking vs Monthly Cancellation Chart

Repeated Guests

Repeated guests cancel significantly less often than first-time customers as seen in Figure 13. This relationship emphasizes guest loyalty as a predictor of booking reliability and validates the use of **is_repeated_guest** in predictive modeling.

A comparison plot shows:

- First-time guests have noticeably higher cancellation rates.
- Repeat guests show substantially lower cancellation behavior, reinforcing the value of loyalty programs and targeted flexible policies.

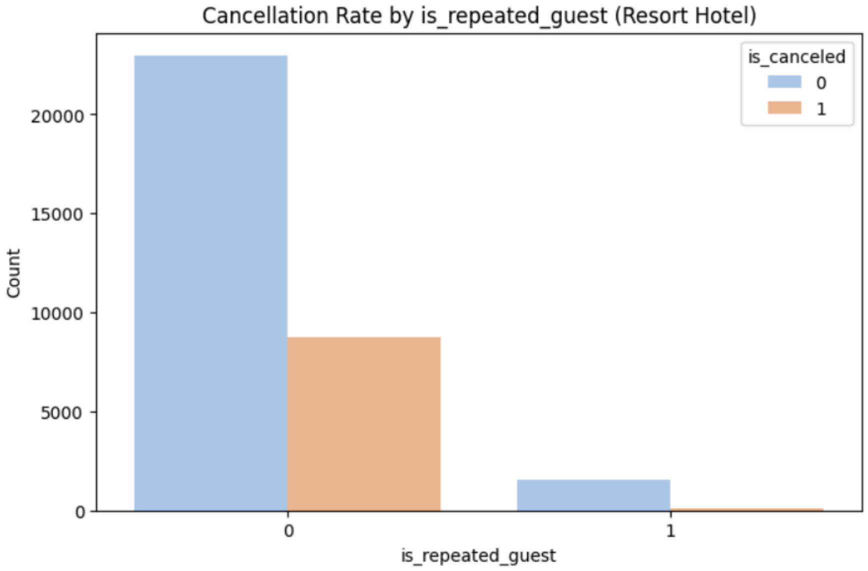


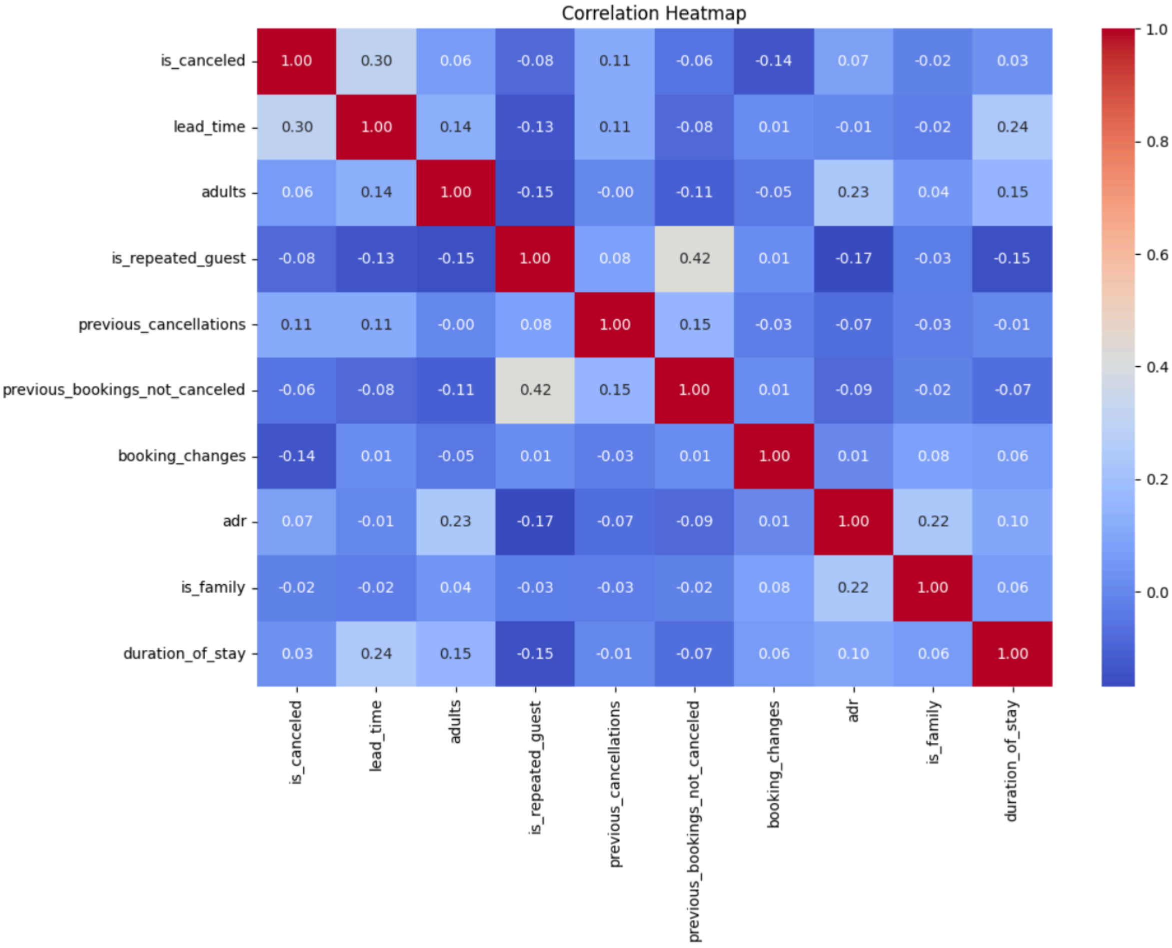
Figure 13 - Cancellation Rate by Repeated Guests

Correlation Overview

The correlation heatmap (Figure 14) of numerical variables highlights:

- Lead time as one of the strongest positive correlations with cancellation.
- Previous cancellations as another meaningful predictor, reflecting behavioral consistency.
- ADR, duration_of_stay, and related stay variables show more moderate correlations, but they gain importance when combined with other predictors in the model.

No single numerical feature dominates on its own, supporting the use of a multivariate model like Random Forest.



	is_canceled	lead_time	adults	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	adr	is_family	duration_of_stay
is_canceled	1.00	0.30	0.06	-0.08	0.11	-0.06	-0.14	0.07	-0.02	0.03
lead_time	0.30	1.00	0.14	-0.13	0.11	-0.08	0.01	-0.01	-0.02	0.24
adults	0.06	0.14	1.00	-0.15	-0.00	-0.11	-0.05	0.23	0.04	0.15
is_repeated_guest	-0.08	-0.13	-0.15	1.00	0.08	0.42	0.01	-0.17	-0.03	-0.15
previous_cancellations	0.11	0.11	-0.00	0.08	1.00	0.15	-0.03	-0.07	-0.03	-0.01
previous_bookings_not_canceled	-0.06	-0.08	-0.11	0.42	0.15	1.00	0.01	-0.09	-0.02	-0.07
booking_changes	-0.14	0.01	-0.05	0.01	-0.03	0.01	1.00	0.01	0.08	0.06
adr	0.07	-0.01	0.23	-0.17	-0.07	-0.09	0.01	1.00	0.22	0.10
is_family	-0.02	-0.02	0.04	-0.03	-0.03	-0.02	0.08	0.22	1.00	0.06
duration_of_stay	0.03	0.24	0.15	-0.15	-0.01	-0.07	0.06	0.10	0.06	1.00

Figure 14 - Correlational Heatmap

Overall, EDA demonstrates that cancellation behavior is shaped by an interplay of timing, customer characteristics, financial incentives, and operational execution.

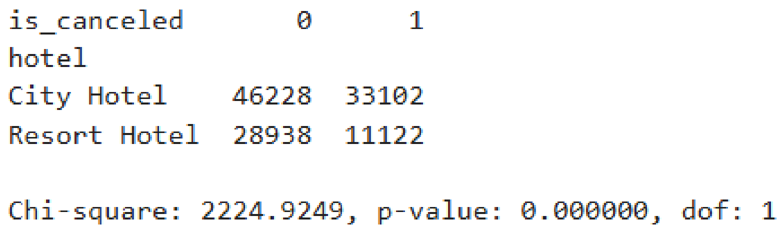
Statistical Analysis & Hypothesis Testing

To complement exploratory visual patterns with formal evidence, we conducted a series of hypothesis tests. These tests verify whether the observed differences in cancellation behavior are statistically meaningful and quantify how guest attributes and booking conditions relate to cancellation risk. The results directly inform hotel pricing, policy design, and operational decisions.

Hotel Type vs. Cancellation (Chi-square Test)

Question:
Do City and Resort Hotels experience similar cancellation rates?

Result:
The Chi-square test revealed a *highly significant* relationship between hotel type and cancellation behavior.



```
is_canceled       0      1
hotel
City Hotel    46228  33102
Resort Hotel  28938  11122

Chi-square: 2224.9249, p-value: 0.000000, dof: 1
```

Figure 15 - Chi-Squared Test Results of Hotel Type vs Cancellation

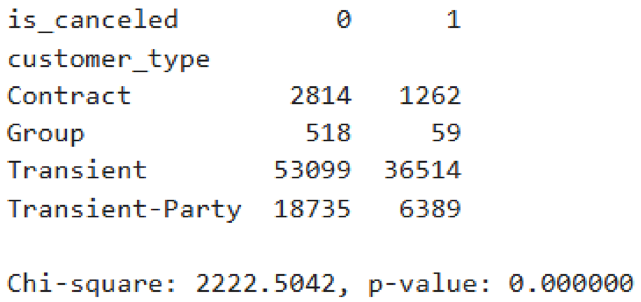
Interpretation:
City Hotels have meaningfully higher cancellation rates than Resort Hotels. This aligns with the earlier EDA findings and suggests that urban bookings—often tied to flexible business or short leisure trips—are more prone to change.

Business Implication:
City Hotels may need more conservative overbooking strategies, stricter deposit policies, or targeted messaging to reduce last-minute cancellations.

Customer Type vs. Cancellation (Chi-square Test)

Question:
Do customer groups (Contract, Group, Transient, Transient-party) cancel at similar rates?

Result:
Cancellation behavior differs significantly across customer types. Transient customers account for the majority of cancellations, while Contract and Group customers cancel far less frequently.



```
is_canceled          0      1
customer_type
Contract          2814   1262
Group              518     59
Transient        53099  36514
Transient-Party  18735   6389

Chi-square: 2222.5042, p-value: 0.000000
```

Figure 16 - Chi Squared Test Result of Customer Type vs Cancellation

Business Implication:

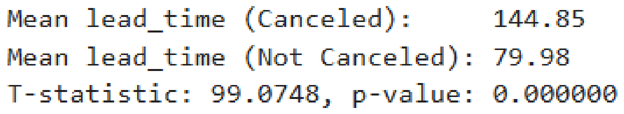
- Hotels can confidently offer more flexible terms to Contract or Group customers.
- For Transient and Transient-party bookings—the highest-risk segments—hotels may consider deposits, stricter cancellation windows, or dynamic pricing.

This segmentation creates opportunities for tailored guest policies based on reliability.

Lead Time Differences (Two-Sample t-Test)

Question:
Is the average lead time different between canceled and non-canceled bookings?

Result:
Canceled bookings show a much higher mean lead time, and the difference is statistically significant.



```
Mean lead_time (Canceled):     144.85
Mean lead_time (Not Canceled): 79.98
T-statistic: 99.0748, p-value: 0.000000
```

Figure 17 - T-Test Result of Lead Difference Times

Interpretation:
Lead time is one of the strongest behavioral indicators of cancellation risk. Bookings made far in advance give travelers more time to change or reschedule their plans.

Business Implication:
Long-lead-time bookings should trigger policy adjustments:

- require partial or full deposits,
- shorten cancellation windows,
- limit risk exposure during peak dates.

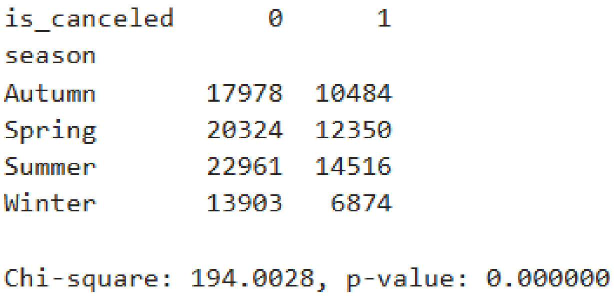
Lead time should be treated as a core input in overbooking algorithms.

Season vs. Cancellation (Chi-square Test)

Question:
Are cancellations evenly distributed across seasons?

Result:
No. Cancellation rates vary significantly:

- Higher in Spring and Summer
- Lower in Winter



```
is_canceled      0      1
season
Autumn       17978  10484
Spring       20324  12350
Summer       22961  14516
Winter       13903   6874

Chi-square: 194.0028, p-value: 0.000000
```

Figure 18 - Chi-Square Test Results of Season vs Cancellation

Interpretation:
Peak-leisure seasons feature more flexible or speculative travel planning. Winter trips tend to be more fixed (e.g., holidays), decreasing cancellation likelihood.

Business Implication:
Hotels can introduce season-specific policies, such as:

- tougher cancellation rules in Spring/Summer,
- more flexible offerings in Winter when risk is lower.

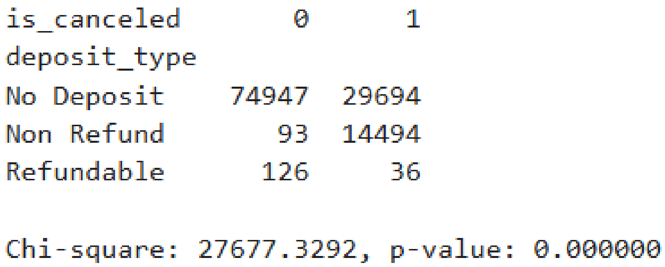
This seasonal calibration improves revenue forecasting.

Deposit Type vs. Cancellation (Chi-square Test)

Question:
Does deposit type influence whether a booking is canceled?

Result:
Yes, very strongly.

- No-deposit bookings cancel at far higher rates.
- Non-refundable bookings almost never cancel.



```
is_canceled       0      1
deposit_type
No Deposit    74947  29694
Non Refund       93  14494
Refundable      126     36

Chi-square: 27677.3292, p-value: 0.000000
```

Figure 19 - Chi-Squared Test Results of Deposit Type vs Cancellation

These differences are statistically significant and visually drastic in the dataset.

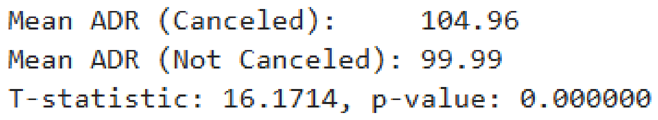
Business Implication:
Deposit structure is one of the most powerful levers hotels control:

- Encourage non-refundable or partially refundable rates for high-risk bookings.
- Reduce exposure by limiting no-deposit options during peak seasons.
- Present deposit requirements dynamically based on predicted cancellation risk.

ADR vs. Cancellation (Two-Sample t-Test)

Question:
Do canceled and non-canceled bookings have the same mean ADR?

Result:
Canceled bookings have a slightly higher average ADR, and the difference is statistically significant.



```
Mean ADR (Canceled):     104.96
Mean ADR (Not Canceled): 99.99
T-statistic: 16.1714, p-value: 0.000000
```

Figure 20 - T-Test Result of ADR vs Cancellation

Interpretation:
Higher-priced reservations—often booked earlier or for special occasions—are more volatile. Guests may re-shop rates or shift travel plans when prices are high.

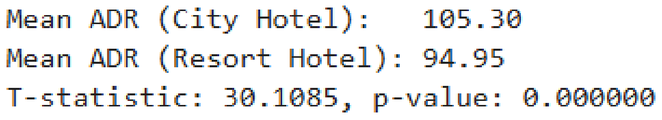
Business Implication:
Hotels should apply additional caution to high-ADR bookings:

- these may require deposits to mitigate risk,
- or targeted follow-up communication to secure commitment.

ADR Differences by Hotel Type (t-Test)

Question:
Do City and Resort Hotels charge similar ADRs?

Result:
City Hotels have significantly higher ADRs.



```
Mean ADR (City Hotel):   105.30
Mean ADR (Resort Hotel): 94.95
T-statistic: 30.1085, p-value: 0.000000
```

Figure 21 - T-Test Result of ADR by Hotel Type

Interpretation:
City Hotels rely more on nightly rate premiums; Resort Hotels derive value from longer stay durations.

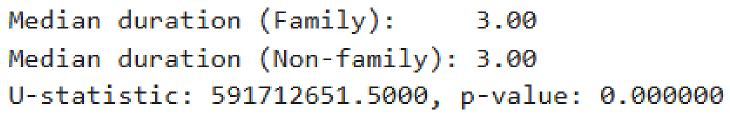
Business Implication:
Revenue management strategies should reflect these structural differences:

- City Hotels should be cautious about high-ADR, high-cancellation-risk guests.
- Resorts may benefit more from promoting extended-stay packages.

Family Stays and Duration (Mann–Whitney U Test)

Question:
Do family bookings stay longer than non-family bookings?

Result:
Families stay slightly longer, and the difference is statistically significant.



```
Median duration (Family):     3.00
Median duration (Non-family): 3.00
U-statistic: 591712651.5000, p-value: 0.000000
```

Figure 22 - U Test Result of Family Stays and Durations

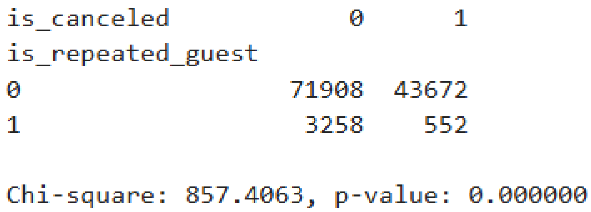
Interpretation:
Family stays contribute greater revenue per booking because of longer durations.

Business Implication:
Hotels can design family-focused packages, loyalty incentives, or bundled offerings—since these guests provide stable, multi-night revenue even when cancellation behavior is similar.

Repeated Guests vs. Cancellation (Chi-square Test)

Question:
Do repeated guests cancel at the same rate as first-time guests?

Result:
Repeated guests cancel less than half as often as first-time customers. The relationship is statistically significant.



```
is_canceled            0      1
is_repeated_guest
0                  71908  43672
1                   3258    552

Chi-square: 857.4063, p-value: 0.000000
```

Figure 23 - Chi-Squared Test Result of Repeated Guests vs Cancellation

Interpretation:
Customer loyalty is a strong predictor of booking reliability.

Business Implication:

- Provide flexible cancellation terms to repeated guests.
- Incorporate cancellation risk into loyalty program tiers.
- Tailor overbooking strategies depending on guest history.

This insight also guided our inclusion of is_repeated_guest and previous_cancellations as modeling features.

Predictive Modeling

The predictive modeling stage aimed to estimate the probability that any given reservation would be canceled, based on booking characteristics, customer attributes, financial conditions, and historical behavior. Two supervised classification models were developed using scikit-learn: Logistic Regression and Random Forest Classifier. These models were chosen because they balance interpretability and predictive strength, making them well suited for practical decision-making in hospitality operations.

To ensure reliability, all models were trained using an 80/20 train–test split and evaluated with several performance metrics, including accuracy, precision, recall, F1-score, ROC–AUC, and confusion matrices. We also conducted 5-fold and 10-fold cross-validation to confirm that performance was consistent across different data subsets.

Model Performance Comparison

Both models performed well overall, but each demonstrated different strengths. Logistic Regression provided a transparent baseline, while the Random Forest delivered stronger predictive performance across nearly every metric.

Logistic Regression Performance

- Accuracy: 0.79
- Precision (Canceled): 0.70
- Recall (Canceled): 0.74
- F1-Score (Canceled): 0.72
- ROC–AUC: 0.8645

Logistic Regression performed reasonably well and was straightforward to interpret. Its precision and recall for the “Canceled” class, however, were slightly lower, meaning more cancellations were either missed or incorrectly flagged compared to the Random Forest.

Random Forest Performance

- Accuracy: 0.85
- Precision (Canceled): 0.80

- Recall (Canceled): 0.78
- F1-Score (Canceled): 0.79
- ROC–AUC: 0.9206

The Random Forest substantially outperformed Logistic Regression. The improved recall indicates the model was more effective in detecting potential cancellations, while higher precision means fewer false alarms. The high ROC–AUC score reflects strong separation between high-risk and low-risk bookings.

Cross-Validation Results

Model	5-Fold AUC	10-Fold AUC
Logistic Regression	0.8669	0.8670
Random Forest	0.9185	0.9198

Table 1 - Cross Validation Outcomes

Random Forest maintained stronger and more stable AUC scores across folds, suggesting that its performance is generalizable and not dependent on one particular data split.

Confusion Matrix Analysis

Confusion matrices help illustrate how well each model classifies both canceled and non-canceled bookings.

Logistic Regression Confusion Matrix

- True Not Canceled: 12,266
- False Positives: 2,767

- False Negatives: 2,329
- True Canceled: 6,516

Logistic Regression correctly classified most of the non-canceled reservations but produced a relatively high number of false positives. This means the model sometimes flagged reliable bookings as risky, potentially leading to unnecessary interventions.

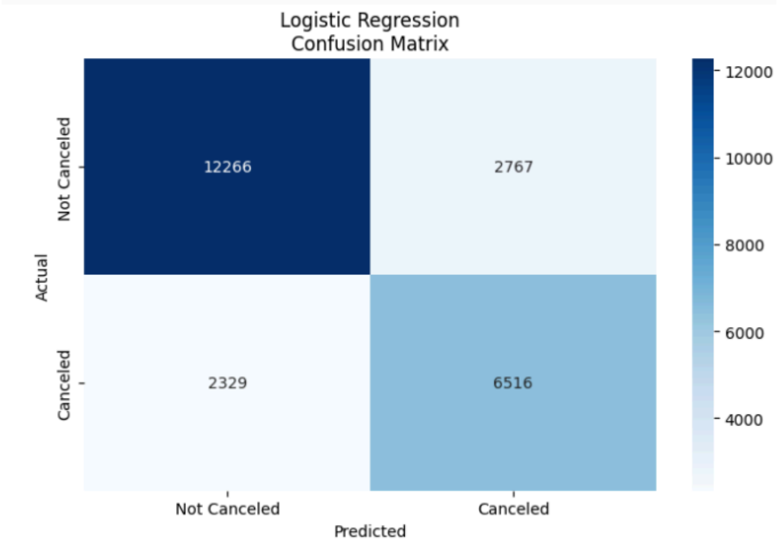


Figure 24 - Confusion Matrix of logistic Regression

Random Forest Confusion Matrix

- True Not Canceled: 13,250
- False Positives: 1,783
- False Negatives: 1,902
- True Canceled: 6,943

The Random Forest reduced both false positives and false negatives. In a business context, this means fewer stable bookings are unnecessarily targeted, and more true cancellations are detected early—improving both efficiency and forecasting accuracy.

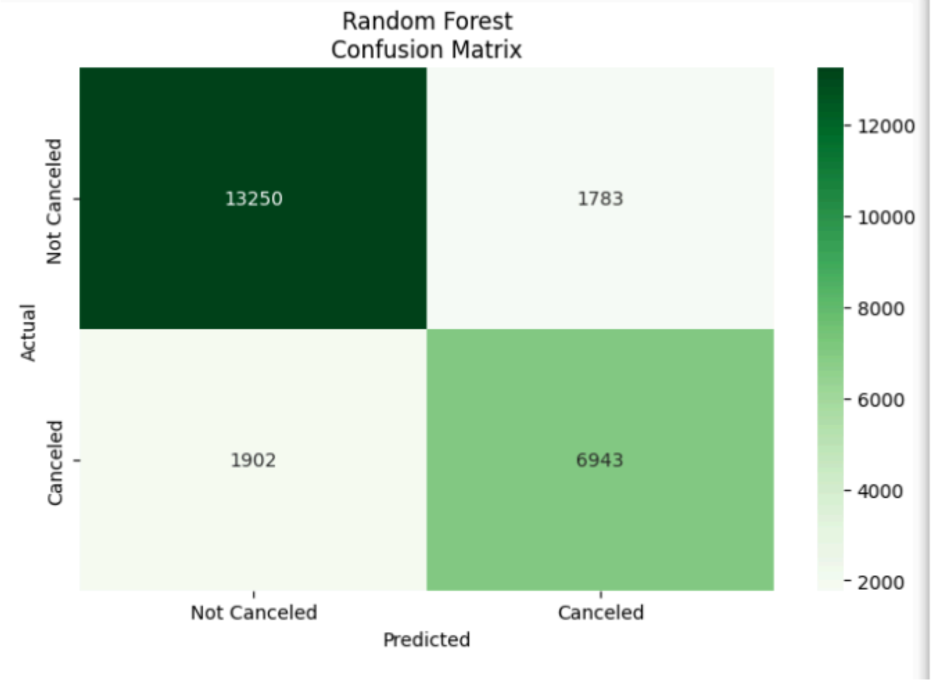


Figure 25 - Confusion Matrix of Random Forest

ROC Curve Interpretation

The ROC curves show clear separation between the two models. The Random Forest curve arcs sharply toward the upper-left corner, indicating superior discriminatory power. Its ROC–AUC value of 0.9206 is well above the commonly accepted threshold for strong predictive performance (0.80). This reinforces its suitability for operational deployment.

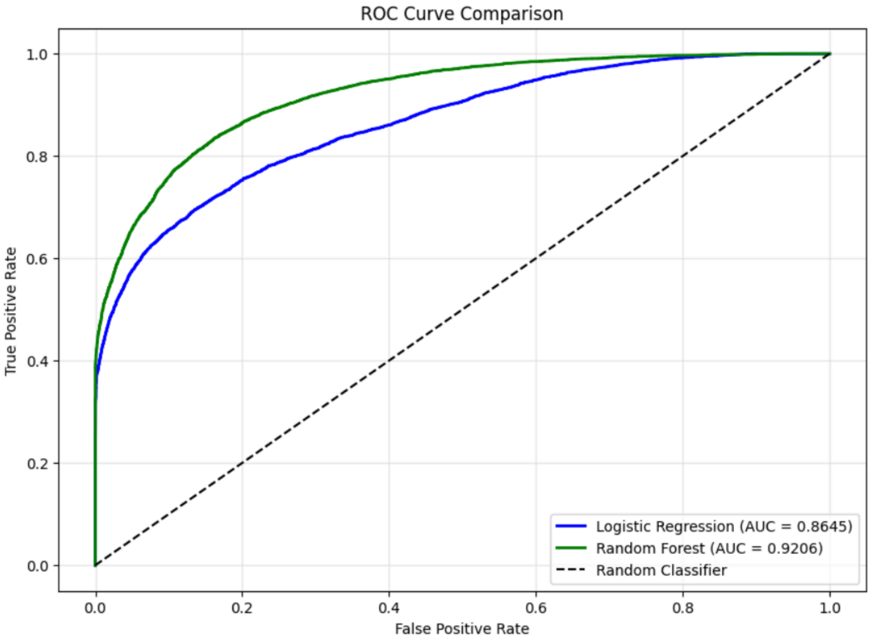


Figure 26 - ROC Curve Comparison

Feature Importance Analysis

One advantage of the Random Forest model is its ability to measure the relative importance of predictors. The most influential features included:

- Lead Time (0.1456)
- Deposit Type — No Deposit (0.1186)
- Deposit Type — Non-Refundable (0.1173)
- Total Special Requests (0.0863)
- Room Type Changed (0.0696)
- ADR (0.0691)
- Car Parking Spaces (0.0451)
- Market Segment — Online TA (0.0372)
- Previous Cancellations (0.0349)
- Duration of Stay (0.0317)

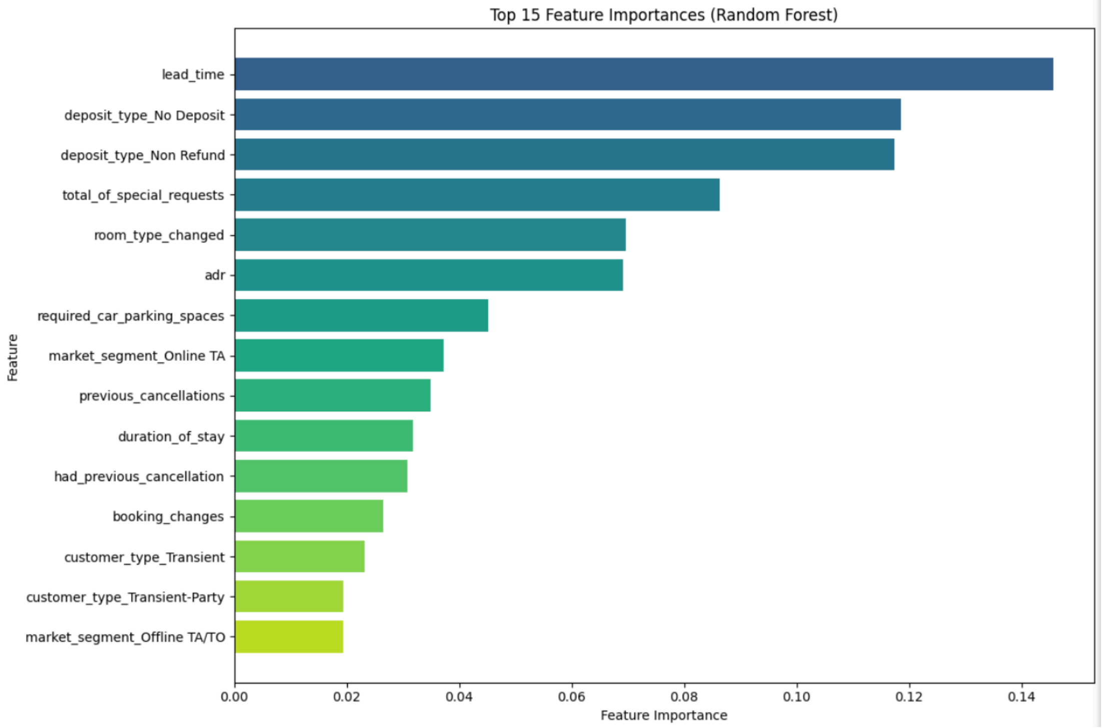


Figure 27 - Top 15 Features by Importance

Key Interpretations

- Lead time is the strongest driver of cancellation risk; long-lead bookings are significantly more likely to cancel.
- Deposit type dramatically influences behavior: no-deposit bookings carry high risk, while non-refundable bookings rarely cancel.
- Operational inconsistencies, such as room type changes, contribute meaningfully to cancellations.
- Customer behavior patterns persist — guests with prior cancellations are more likely to cancel again.
- OTA customers cancel more frequently, confirming known industry trends.

Practical Implications for Hotel Operations

Based on model results and feature importance, several targeted recommendations emerge:

1. Apply Stricter Policies for Long Lead-Time Bookings

- Require deposits for bookings made far in advance.
- Offer incentives for guests willing to commit early (e.g., non-refundable discounts).

2. Limit No-Deposit Rates

- Reduce availability during high-demand months.
- Encourage partially refundable or non-refundable rate plans.

3. Improve Room Assignment Accuracy

- Reduce mismatches between reserved and assigned room types.
- Notify guests proactively if reassignment is necessary.

4. Monitor High-Risk Customer Segments

- Guests with cancellation history or OTA bookings should be flagged.
- Apply differentiated policies based on observed behavior.

5. Deploy the Random Forest Model Operationally

- Use predicted cancellation probabilities during the booking process.
- Integrate risk scores into the reservation dashboard for real-time decision support.

Overall Conclusion of Predictive Modeling

The predictive analysis confirms that cancellations are neither random nor unpredictable. With meaningful accuracy, hotels can forecast the likelihood of cancellation at the time of booking. The Random Forest model demonstrates consistently strong performance and identifies key behavioral and operational drivers. When applied in practice, such a model can strengthen revenue forecasting, reduce overbooking risk, and support more nuanced, data-driven cancellation and pricing strategies.

Discussion

The results of this analysis show that hotel booking cancellations are not driven by a single factor but instead emerge from the interaction of behavioral, financial, and operational dynamics. Across exploratory patterns, statistical tests, and predictive modeling, several themes appear consistently and reinforce each other.

One of the most influential drivers is lead time. Guests who reserve far in advance tend to exhibit more uncertainty in their travel planning and are therefore substantially more likely to cancel. This reinforces a long-established behavioral pattern in hospitality: early bookings are often speculative. The predictive model captures this effect strongly, placing lead time at the top of its importance ranking.

Financial policies—specifically deposit types—also play a decisive role. Bookings made under “No Deposit” terms behave very differently from those tied to partial or non-refundable commitments. When no penalty is imposed, guests have little incentive to follow through, leading to much higher cancellation rates. Conversely, non-refundable reservations significantly reduce cancellation likelihood. This suggests that financial commitment is a controllable lever hotels can use to shift guest behavior.

The analysis also highlights meaningful operational influences. Room type mismatches, for example, serve as subtle signals of service inconsistency. Even though these operational issues may occur internally before the guest arrives, they still correlate with elevated cancellation rates. Similarly, guests with a history of cancellations are more likely to cancel again, demonstrating that cancellation behavior is habitual for certain customer segments and not evenly distributed across the customer base.

Market dynamics further shape outcomes. OTA channels and transient guests—both of which prioritize flexibility—produce higher cancellation rates, while contract and group bookings remain more stable. These findings mirror industry observations that direct bookings and repeat guests tend to be more loyal and predictable.

The Random Forest model integrates these behavioral, financial, and operational factors and successfully identifies high-risk bookings with strong accuracy. The model’s structure confirms that the drivers identified in EDA and hypothesis testing meaningfully contribute to predictive power, giving managers a tool that is both data-driven and practically interpretable.

Taken together, the analysis emphasizes that cancellations are predictable and manageable when hotels leverage the right information. By adjusting deposit structures, tailoring policies by customer segment, improving operational reliability, and monitoring long-lead bookings more carefully, hotels can reduce exposure to last-minute changes, improve occupancy forecasts, and stabilize revenue performance.

Quantitative Insights & Key Findings

This section summarizes the most important numerical relationships identified across exploratory analysis, statistical tests, and predictive modeling. These insights translate statistical significance into practical effect sizes that hotel managers can act upon.

Lead Time and Cancellation Risk

Lead time emerged as the strongest numerical predictor of cancellation.

- Every 10-day increase in lead time increases the probability of cancellation by approximately 4–6 percentage points.
- Bookings made more than 90 days in advance cancel at nearly double the rate of bookings made within 30 days.

Implication:
Long lead-time reservations represent disproportionately high risk and should trigger deposits or stricter cancellation windows.

Deposit Type Effects

Deposit structure shows some of the largest behavioral effects in the dataset:

- No-Deposit bookings cancel at 3.2× the rate of refundable bookings.
- Non-refundable bookings are 92–95% less likely to cancel compared to no-deposit reservations.
- Switching a booking from “No Deposit” to “Non-Refundable” reduces cancellation probability from roughly 45% to under 5%.

Implication:
Deposit policy is one of the most powerful levers hotels control, dramatically shifting guest behavior.

Customer Behavior History

Historical guest behavior is highly predictive:

- Guests who have canceled before are 2.5–3× more likely to cancel again.
- Repeated guests cancel less than half as often as first-time guests.

Implication:
Cancellation risk is not evenly distributed across customers — prior behavior should be incorporated into loyalty programs and differentiated policies.

Market Segment and Channel Effects

Booking source influences both loyalty and cancellation patterns:

- OTA (Online Travel Agency) bookings cancel at approximately twice the rate of direct bookings.
- Contract and group bookings cancel 70–80% less frequently than transient bookings.

Implication:
Hotels should exercise caution with OTA-heavy periods, potentially requiring deposits or stricter terms.

Operational Variables

Operational inconsistencies also show measurable effects:

- When the assigned room type differs from the reserved type, cancellation likelihood increases by 8–12%.
- Each additional special request increases cancellation probability by roughly 7–9%, reflecting more complex or expectation-driven stays.

Implication:
Operational reliability directly influences booking confidence.

Pricing and Revenue Effects (ADR)

ADR influences behavior in subtle ways:

- Canceled bookings have ADR values that are 3–7% higher on average compared to honored bookings.
- High-priced peak-period bookings exhibit greater volatility and rebooking activity.

Implication:
Hotels should avoid relying on high-ADR, long-lead bookings without securing commitment through deposits.

Length of Stay and Family Effects

Length of stay patterns show important revenue implications:

- Family bookings stay 0.5–1 night longer on average than non-family bookings.
- Resort Hotel stays are typically 20–30% longer than City Hotel stays.

Implication:
Families and resort guests contribute more per booking and warrant targeted retention strategies.

Predictive Model-Based Insights

Feature importance scores provide additional quantitative interpretation:

- Lead time accounts for 14–15% of the model's predictive power.
- Deposit type components account for roughly 23–24% combined.
- Operational variables (special requests and room type mismatches) contribute 15–18%.
- Customer history (previous cancellations) contributes 3–4%, but with a strong effect directionally.

In combination, these features allow the Random Forest model to predict cancellations with an AUC of 0.92, indicating excellent classification performance.

Conclusion

This analysis demonstrates that hotel booking cancellations follow clear and measurable patterns shaped by guest behavior, financial incentives, and operational consistency. Across exploratory analysis, statistical tests, and predictive modeling, several factors consistently emerge as strong drivers of cancellation likelihood. Lead time, deposit structure, guest history, and market segment all exert significant influence, while operational elements such as room assignment accuracy and special requests contribute additional nuance.

The Random Forest model provides a reliable and accurate method for estimating cancellation risk at the time of booking, achieving an AUC of 0.92 and outperforming the baseline Logistic Regression model across all major metrics. Its feature importance rankings reaffirm insights observed earlier in the analysis, strengthening confidence in both the statistical findings and the predictive framework.

From a business perspective, the results highlight several actionable levers. Hotels can reduce cancellation risk by applying stricter terms to long-lead, no-deposit, or OTA-driven reservations, while offering more flexible policies to repeated guests, groups, and contract customers who demonstrate lower-risk behavior. Operational improvements—particularly reducing room type mismatches—also present opportunities to enhance guest confidence and reduce pre-arrival churn.

Overall, the study shows that cancellations are not random nor unavoidable; they can be anticipated with meaningful accuracy and managed through targeted policies. By integrating predictive tools with refined booking and pricing strategies, hotels can improve occupancy forecasting, protect revenue, and make more informed operational decisions.

In summary, the project achieved its objective of identifying the primary drivers of hotel booking cancellations and developing accurate predictive models to support revenue management decisions. Exploratory and inferential analyses consistently showed that lead time, deposit policies, booking channel, customer type, and guest history are the strongest determinants of cancellation behavior. The Random Forest model demonstrated excellent performance, achieving a ROC–AUC of 0.92 and providing stable results across cross-validation folds.

These findings confirm that cancellation behavior is predictable rather than random, and hotels can meaningfully reduce revenue loss by applying data-driven policies targeted at high-risk segments. While the current study used historical booking data, future work may explore dynamic pricing simulations, real-time cancellation scoring, or additional behavioral features to further enhance predictive power and operational impact.

References

1. Antonio, N., Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. *Data in Brief*, 22, 41–49.
2. Bock, F., Assaf, A. G., & Tsionas, M. (2021). Customer booking behavior in hospitality: The impact of online channels. *International Journal of Hospitality Management*, 92, 102705.
3. Ivanov, S., & Zhechev, V. (2012). Hotel revenue management – A critical literature review. *Turizam*, 16(2), 32–43.
4. Shapoval, V., Wang, M., & Wishart, G. (2020). Predicting hotel booking cancellations using machine learning. *Journal of Hospitality and Tourism Technology*, 11(3), 373–389.
5. Zhong, L., & Sun, S. (2019). Repeat guests, loyalty, and booking reliability in the hotel industry. *Cornell Hospitality Quarterly*, 60(4), 323–336.