Arowwai
Industries

Hotel Booking Cancellation Prediction

By Raditya Erlang Arkananta

Table of Contents

About Me

Raditya Erlang Arkananta

Data Science Enthusiast

Experience

- PT Hekikai Indonesia — Sep 2024 - Apr 2025

QA/QC Japanese Translator Staff

- PT Indonesia Indikator — Jan 2024 - Sep 2024

HR Officer

Education

- Dibimbing — Nov 2024 - Jun 2025

Full Stack Data Science

- Technology Institute of Sepuluh Nopember — Aug 2019 - Sep 2023

Industrial Engineering

PREVIOUS PROJECTS

Pakistan E-Commerce Customer RFM Analysis

- Identified seasonal trends to optimize marketing during peak periods.
- Segmented customers to enable targeted marketing strategies.
- Found issues in payment methods causing errors and cancellations.
- Discovered a large customer segment with high potential for growth.

Customer Sentiment in Ticketing System

- Analyzed customer satisfaction using CSAT, CES, and NPS metrics.
- Extracted insights from customer comments to understand sentiment.
- Found areas for improvement in customer service.
- Identified low product referrals and the need to boost recommendations.

Indonesia Housing Prices Prediction Model

- Found location to be the strongest factor in predicting housing prices.
- Identified South Jakarta, North Jakarta, and Badung as top high-value areas.
- Analyzed that space efficiency boosts prices more than size alone.
- Recommended buyers, agents, and investors to prioritize location over features.

Main Project

Background

In order to maximize profit and improve operational efficiency, it is crucial for hotels to minimize last-minute cancellations that leave rooms unoccupied.
High cancellation rates can lead to revenue loss, poor inventory planning, and increased uncertainty in staff and resource allocation

Objectives

This project aims to analyze key features that influence customer cancellations, so hotel management can predict potential cancellations, refine booking policies, and make a Predictive Machine Learning Model in order to minimize profit losses due to Cancellations.

Data Understanding

About the Data

The dataset contains data on hotel booking records, including various details related to customer reservations, stay duration, demographics, booking behavior, and cancellation status.

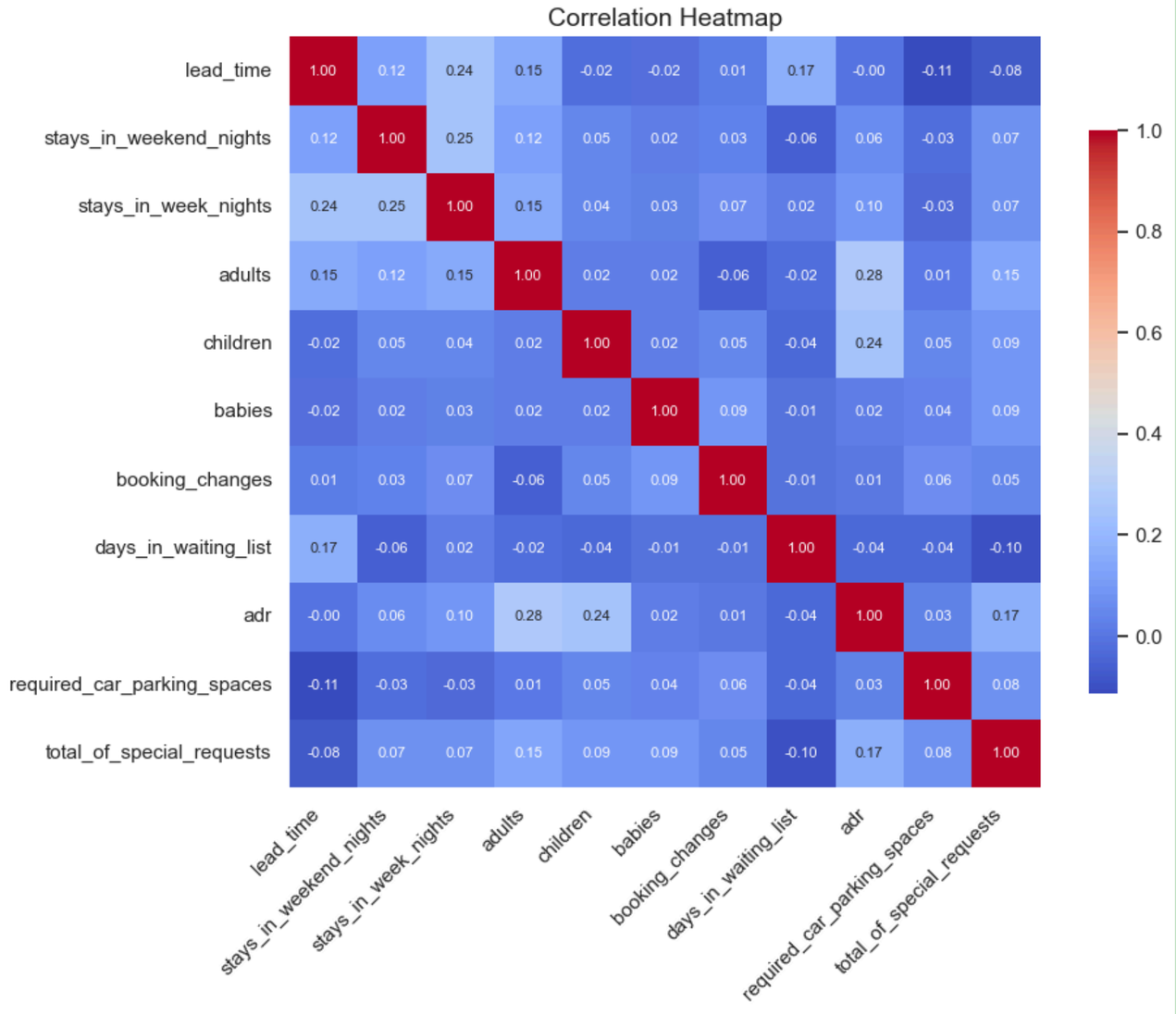
83293 Rows

33 Columns

Jul 2017 - Aug 2019

Multicollinearity Study

As shown in the heatmap, there is no indication of multicollinearity among the features, and the VIF (Variance Inflation Factor) scores are within acceptable limits;
Therefore, we will proceed with the EDA and modeling process.



Data Pre-Processing

1

Handling Duplicate

No Duplicate

2

Handling Missing Value

Imputed missing values: ‘children’ with median, ‘country’ with mode, and ‘agent(ID)’ with ‘0’.

3

Dropping Irrelevant & Missing Columns

Dropping 90% Missing Company Col

4

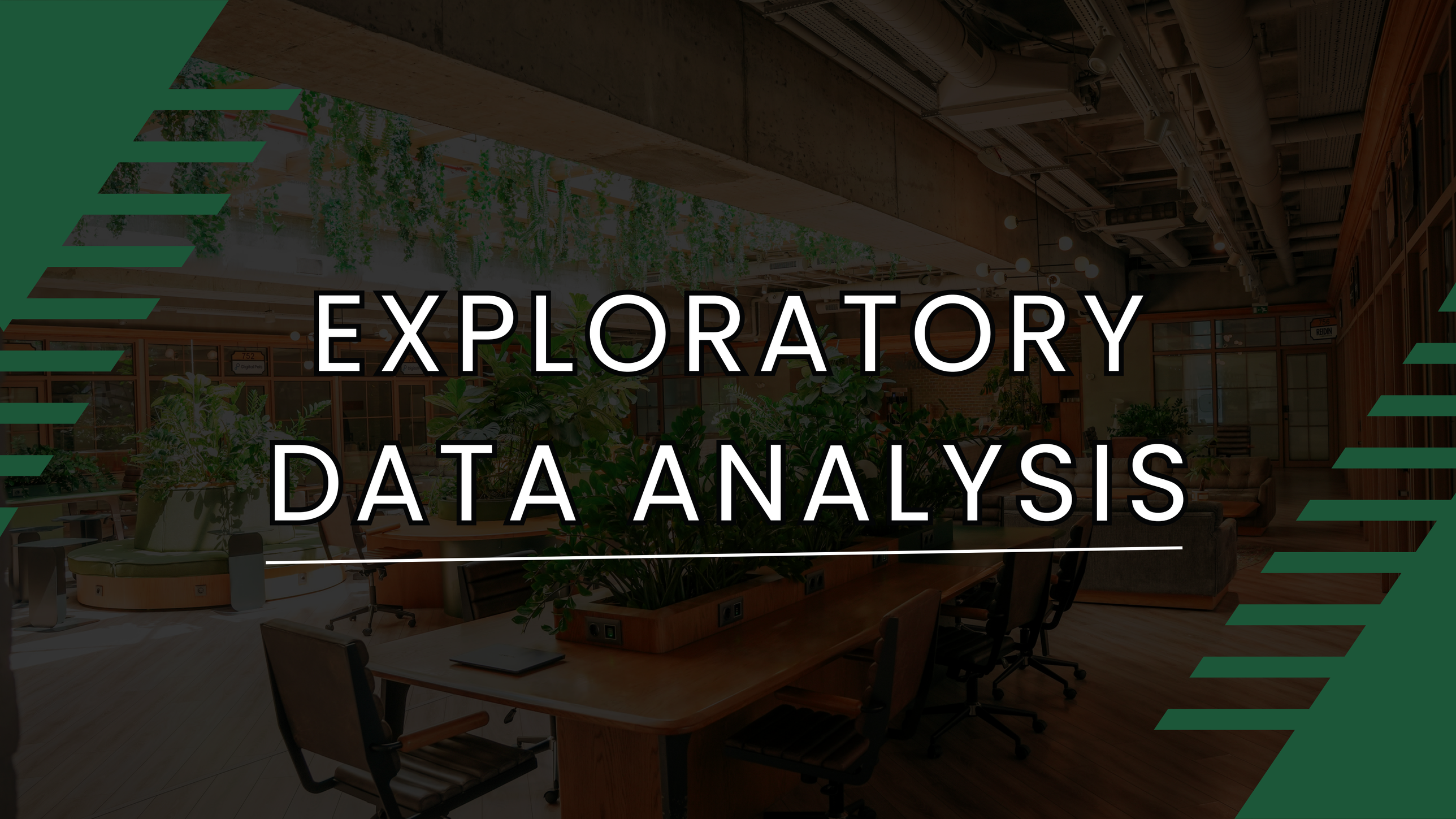
Outlier Handling

Capping was applied to limit extreme or unrealistic values in discrete count features, reducing outlier impact while preserving typical booking behavior.

Data Pre-Processing Result

83293 Rows
33 Columns

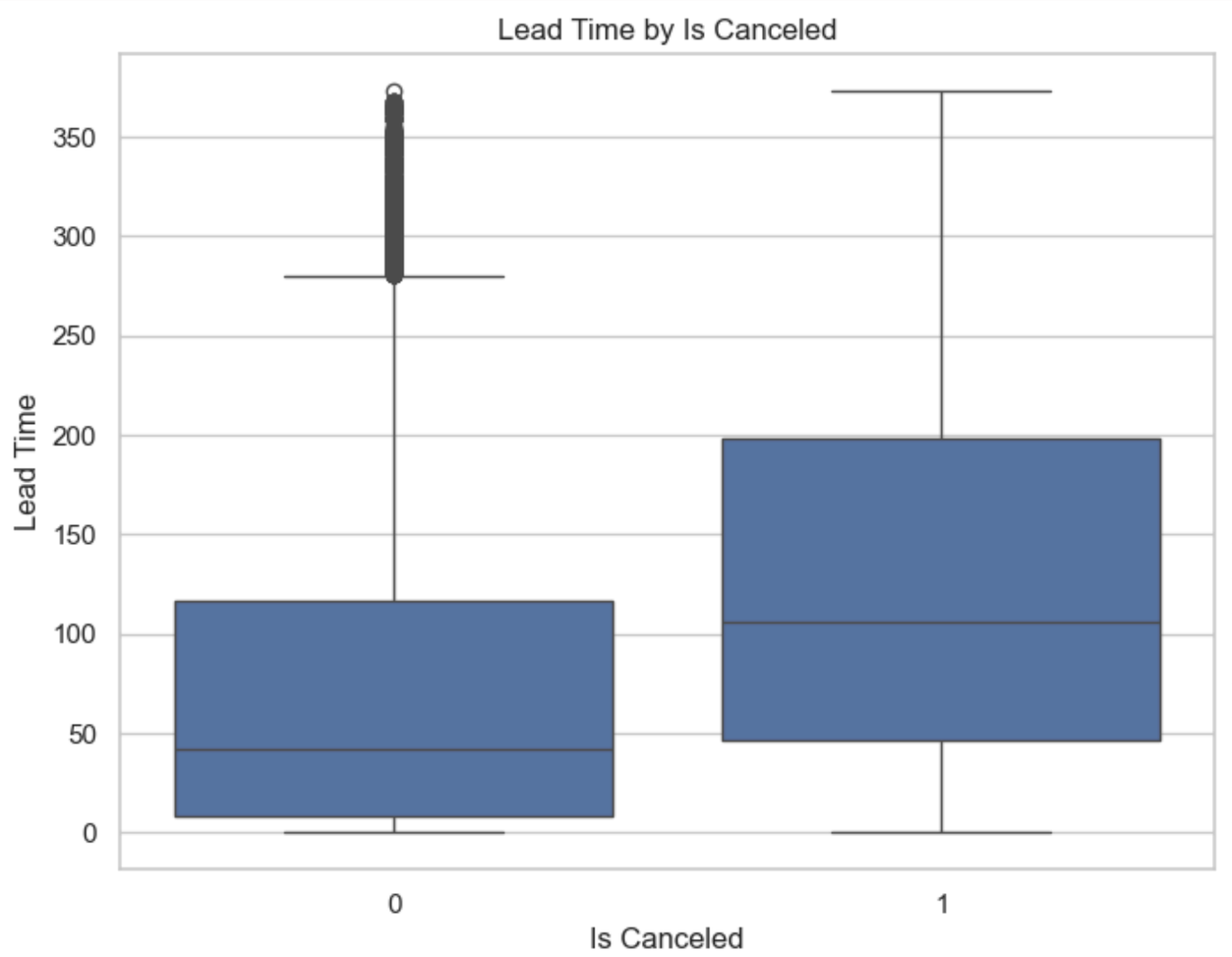
76289Rows
32 Columns



EXPLORATORY DATA ANALYSIS

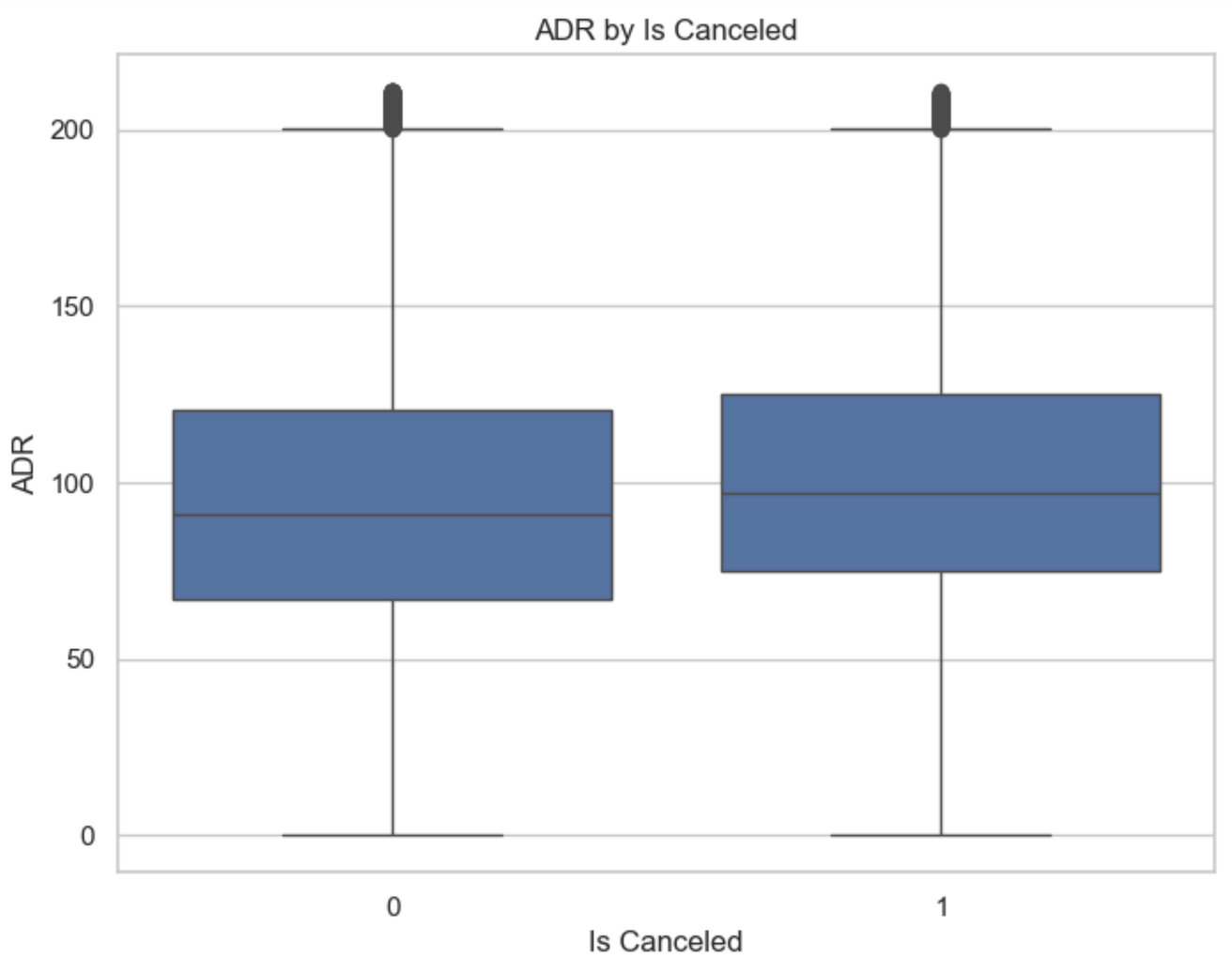
Cancellation by Lead Time

Guests with longer lead times are more likely to cancel their bookings. The extended time between reservation and arrival gives them more opportunity to change their plans, find better deals, or reconsider their stay.



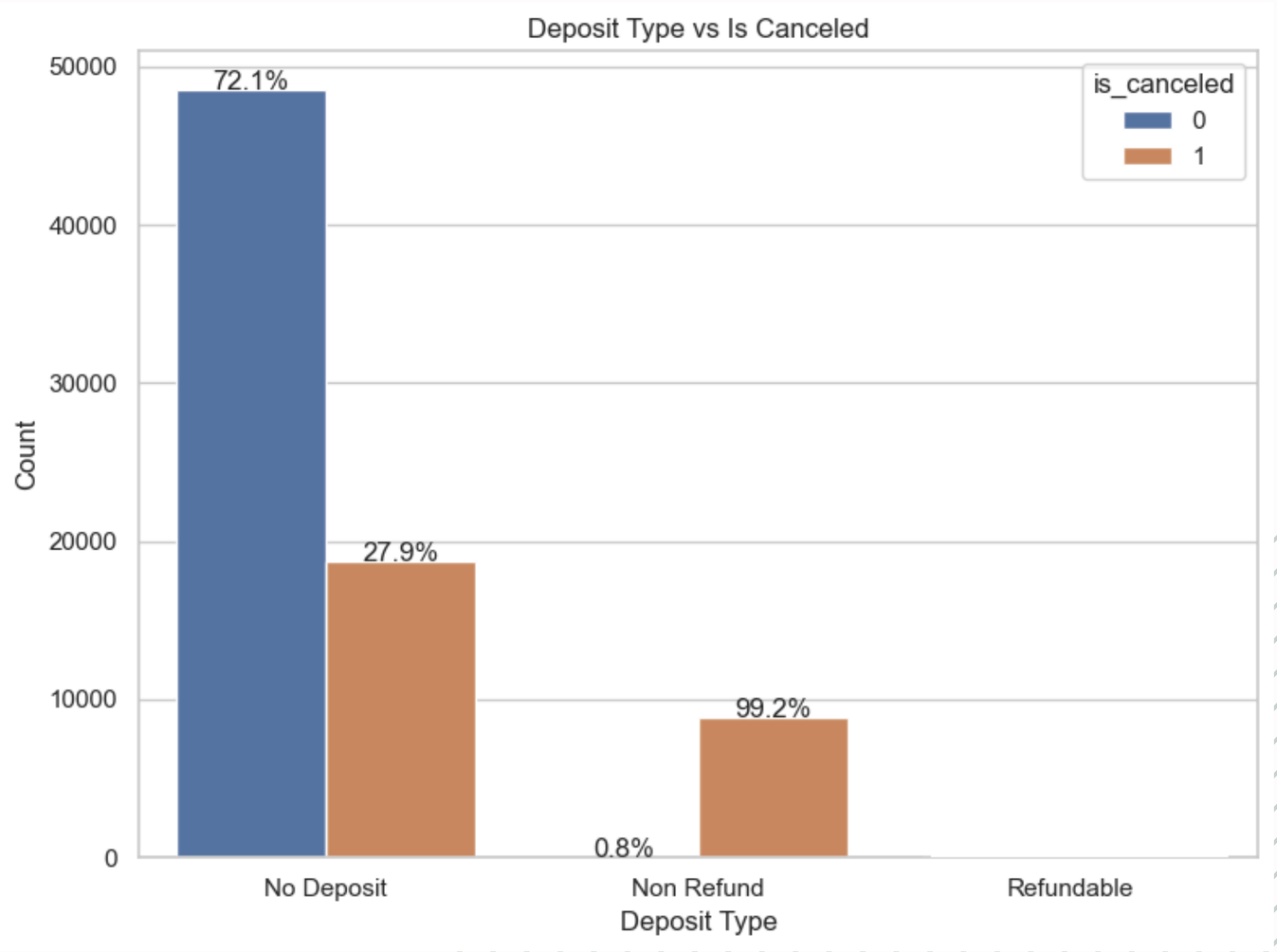
Cancellation by Daily Rate

While the average daily rate (ADR) for canceled bookings is slightly higher than for non-canceled ones, the difference is relatively small, and the overall distributions are quite similar. This suggests that ADR alone is not a strong predictor of cancellation behavior.



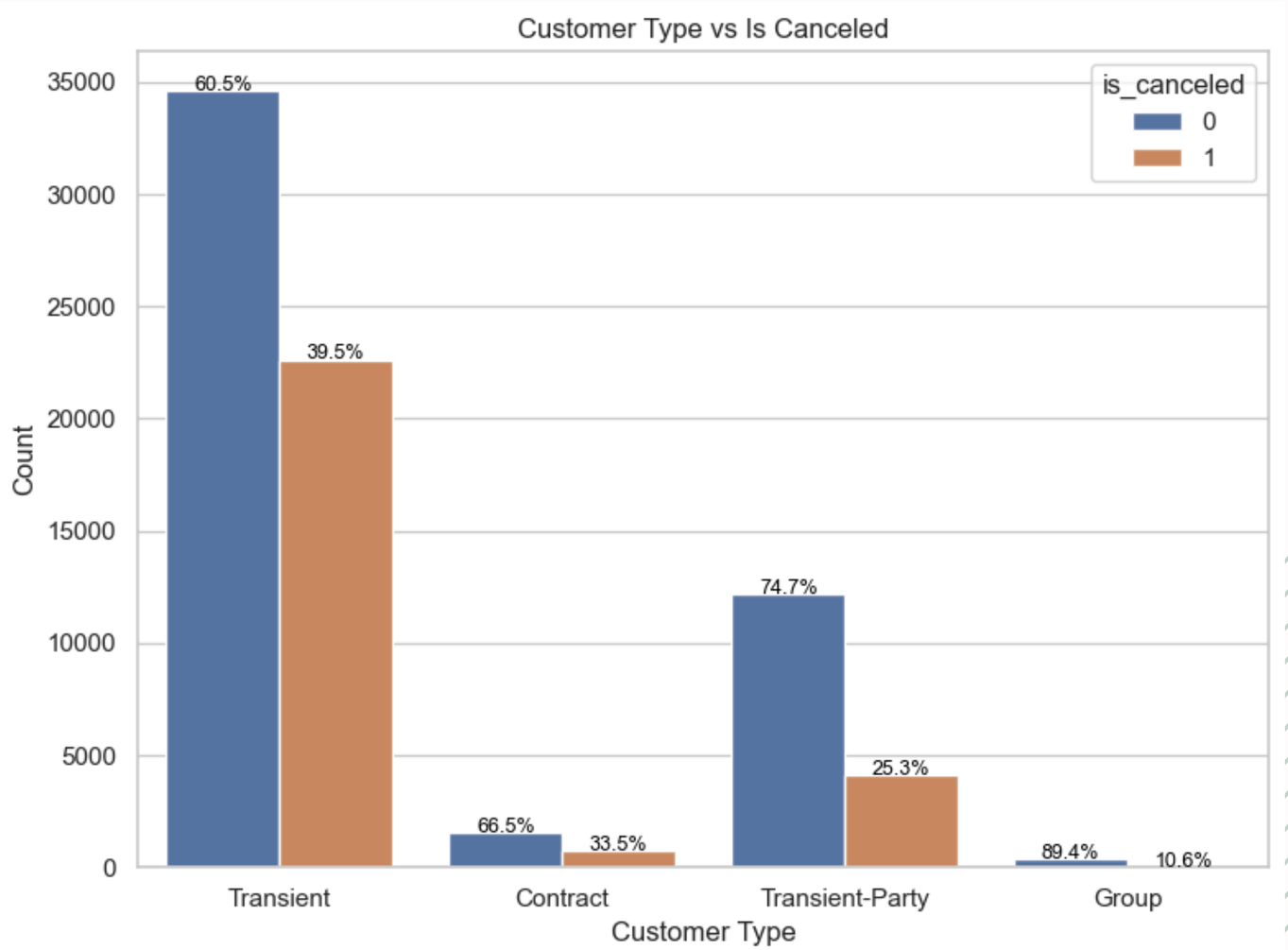
Cancellation by Deposit

While "No Deposit" bookings show the highest cancellations due to zero financial risk, a surprisingly high number of cancellations also occur for "Non Refund" bookings, possibly due to unavoidable circumstances or guest misjudgment.



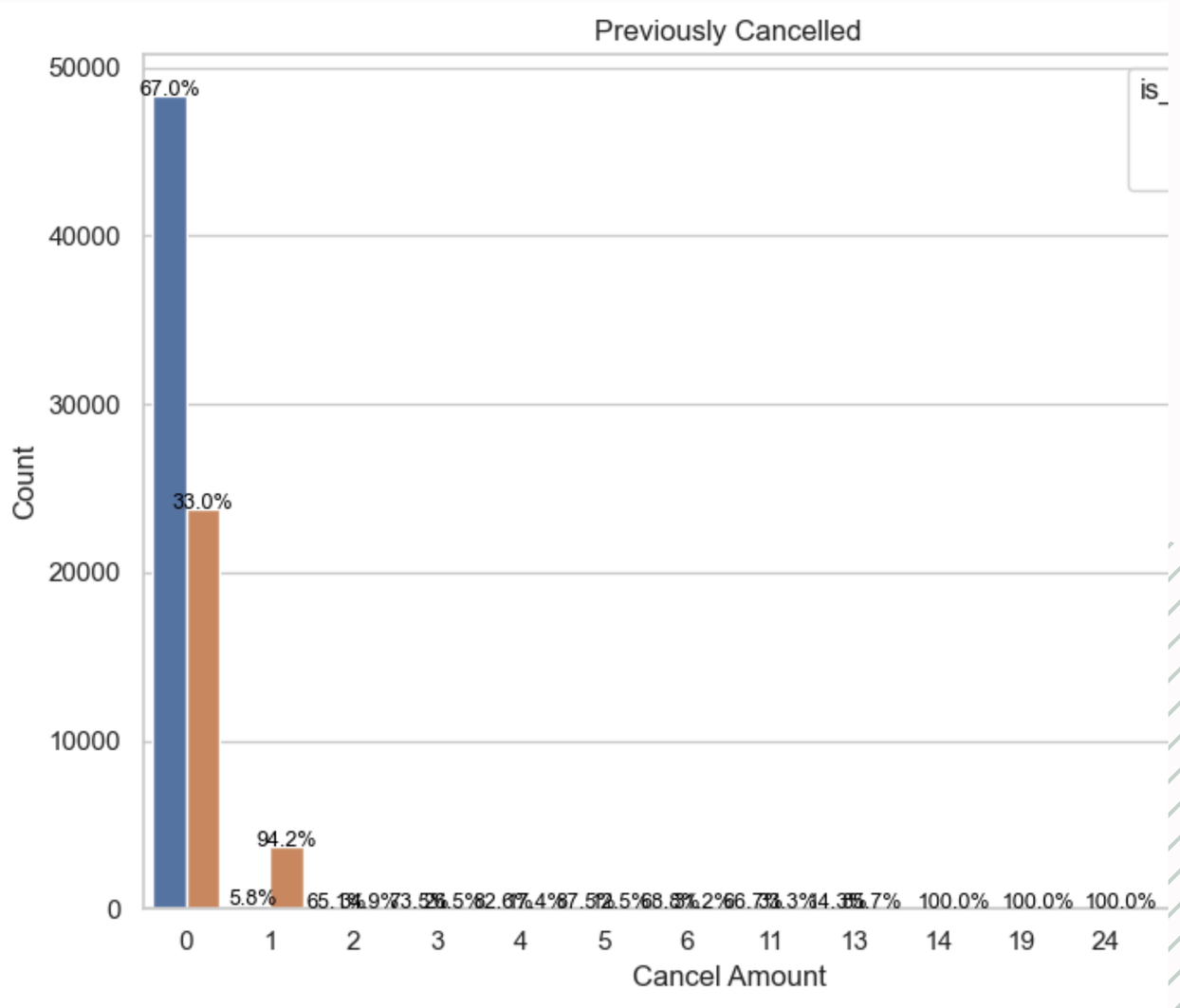
Cancellation by Customer Type

- Transient customers make up the majority of bookings but also have the highest cancellation rate, indicating a need for targeted retention strategies.
- Contract customers have the lowest cancellation rate, suggesting they are the most reliable customer segment.
- Transient-Party customers show a moderate volume with a significant number of cancellations, possibly due to coordination difficulties in group travel.
- Group bookings are rare and show minimal cancellations, though the low volume may limit impact.



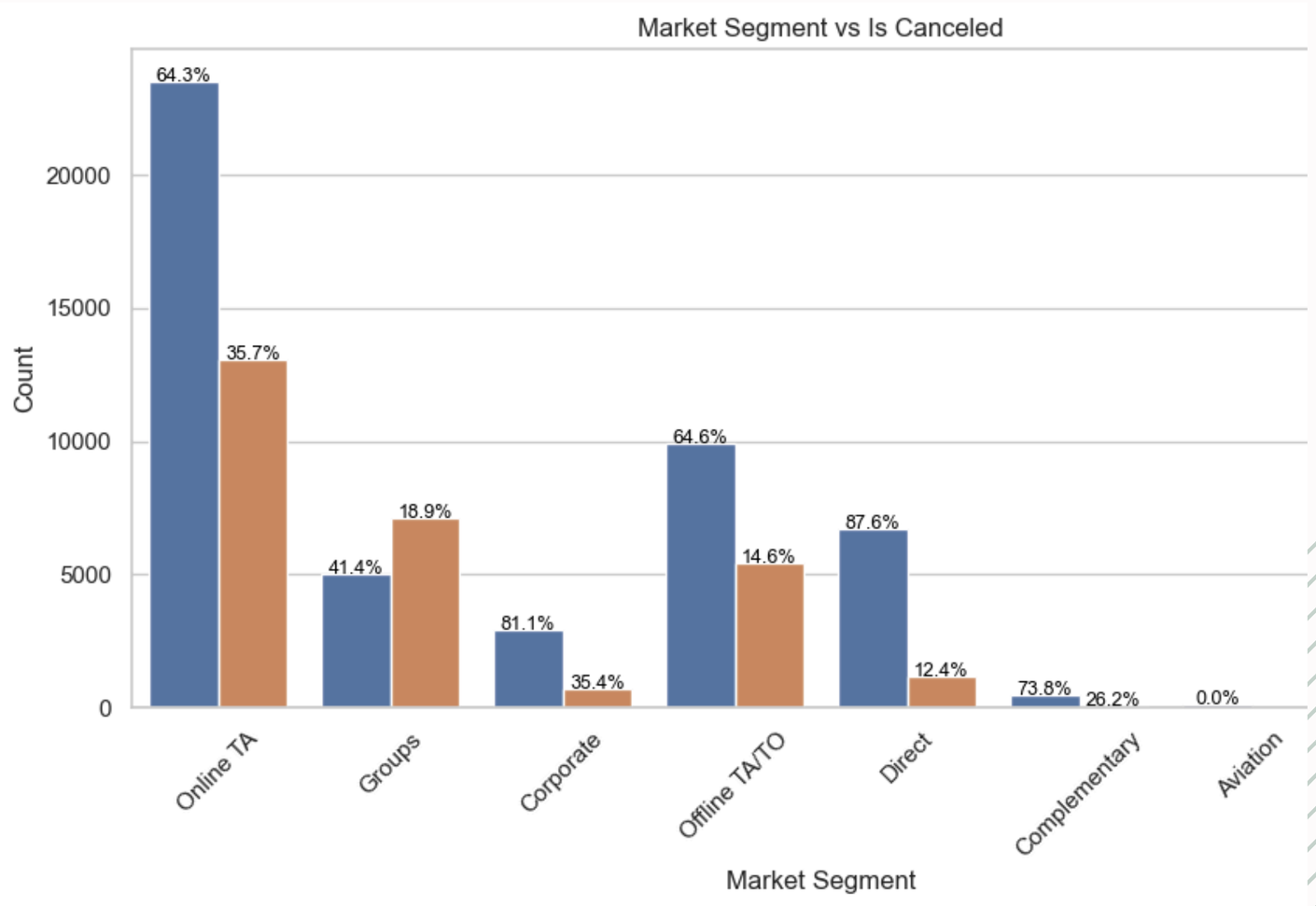
Cancels by Booking Changes

Bookings with no changes are the most common, but still show a high number of cancellations.
As booking changes increase, the likelihood of cancellation also rises—especially for bookings with two or more changes. This suggests that frequent modifications may indicate uncertainty and a higher risk of cancellation.



Cancellation per Segment

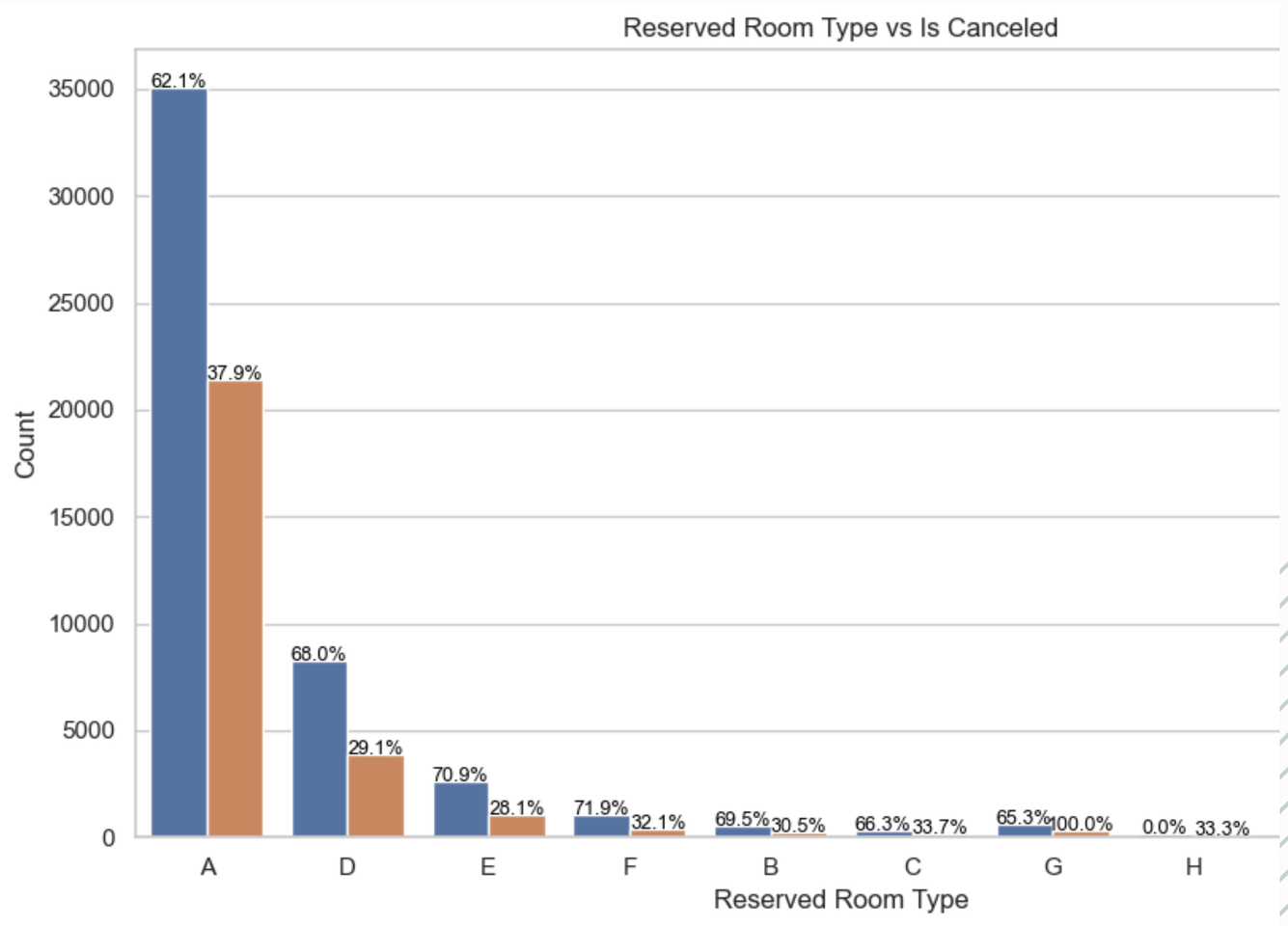
- Online TA (Travel Agencies) accounts for the highest volume of bookings and also shows a high cancellation rate, indicating a volatile customer base.
- The Groups segment has a higher cancellation count than completions, making it a high-risk segment.
- Corporate and Direct bookings have low cancellation rates, suggesting they are more stable and reliable.
- Offline TA/TO (Travel Agents/Tour Operators) bookings are moderately high with a noticeable, but more balanced, cancellation pattern.



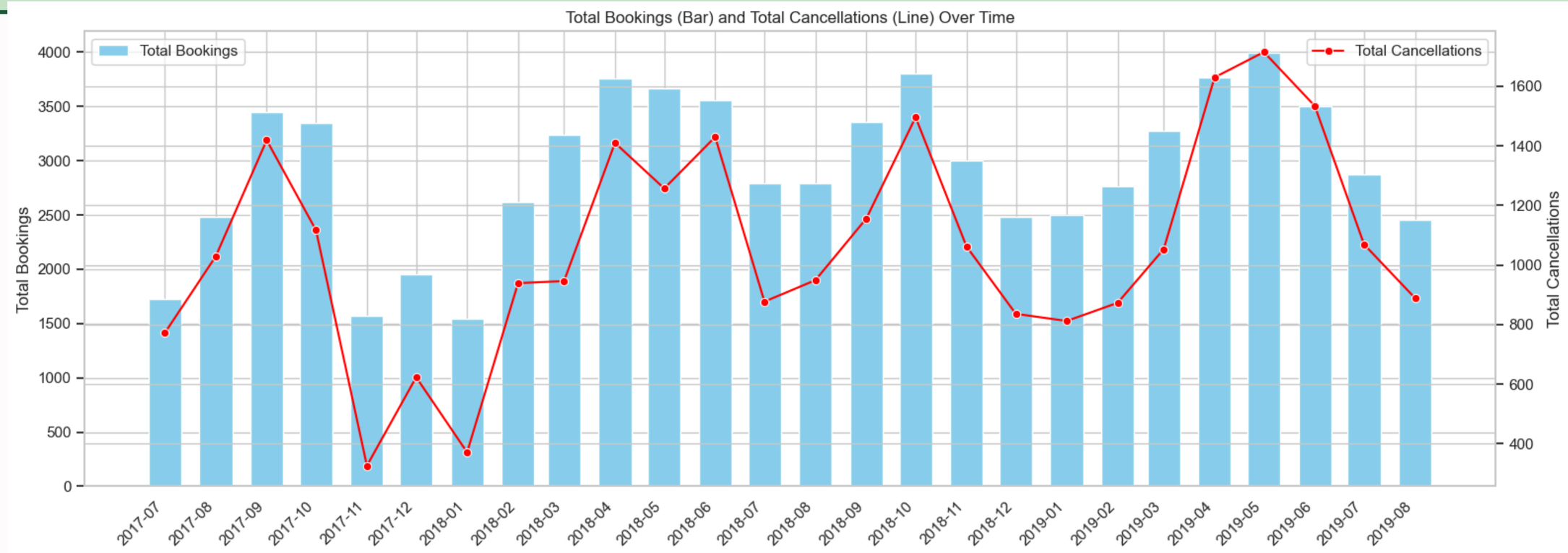
Cancellation by Reserved Room

Room type A is the most frequently reserved and also has the highest number of cancellations, indicating it may be the default or most affordable option.

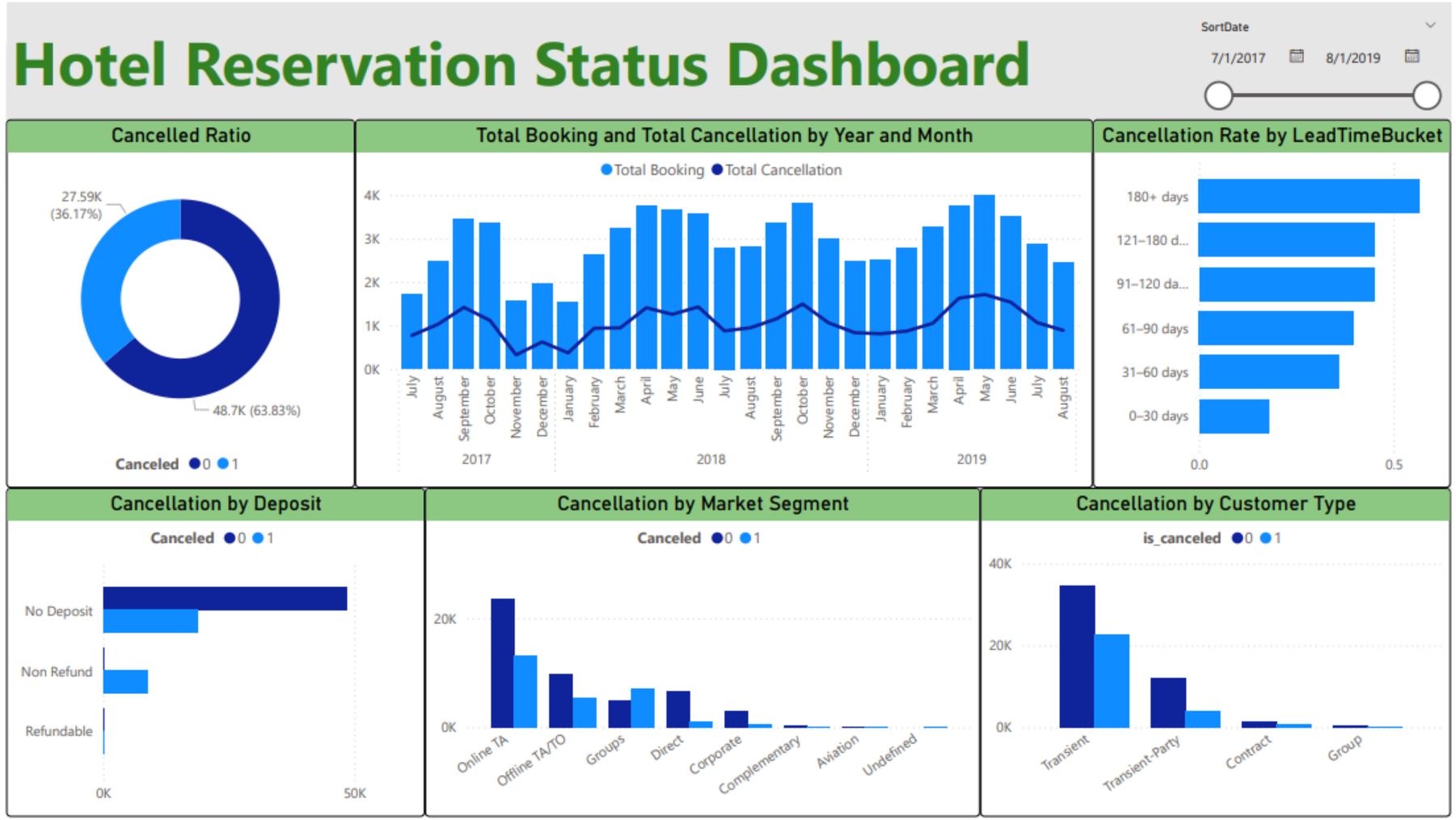
The pattern suggests that popular or standard room types experience higher cancellation rates, possibly due to overbooking, reassignment issues, or less commitment from cost-sensitive guests.



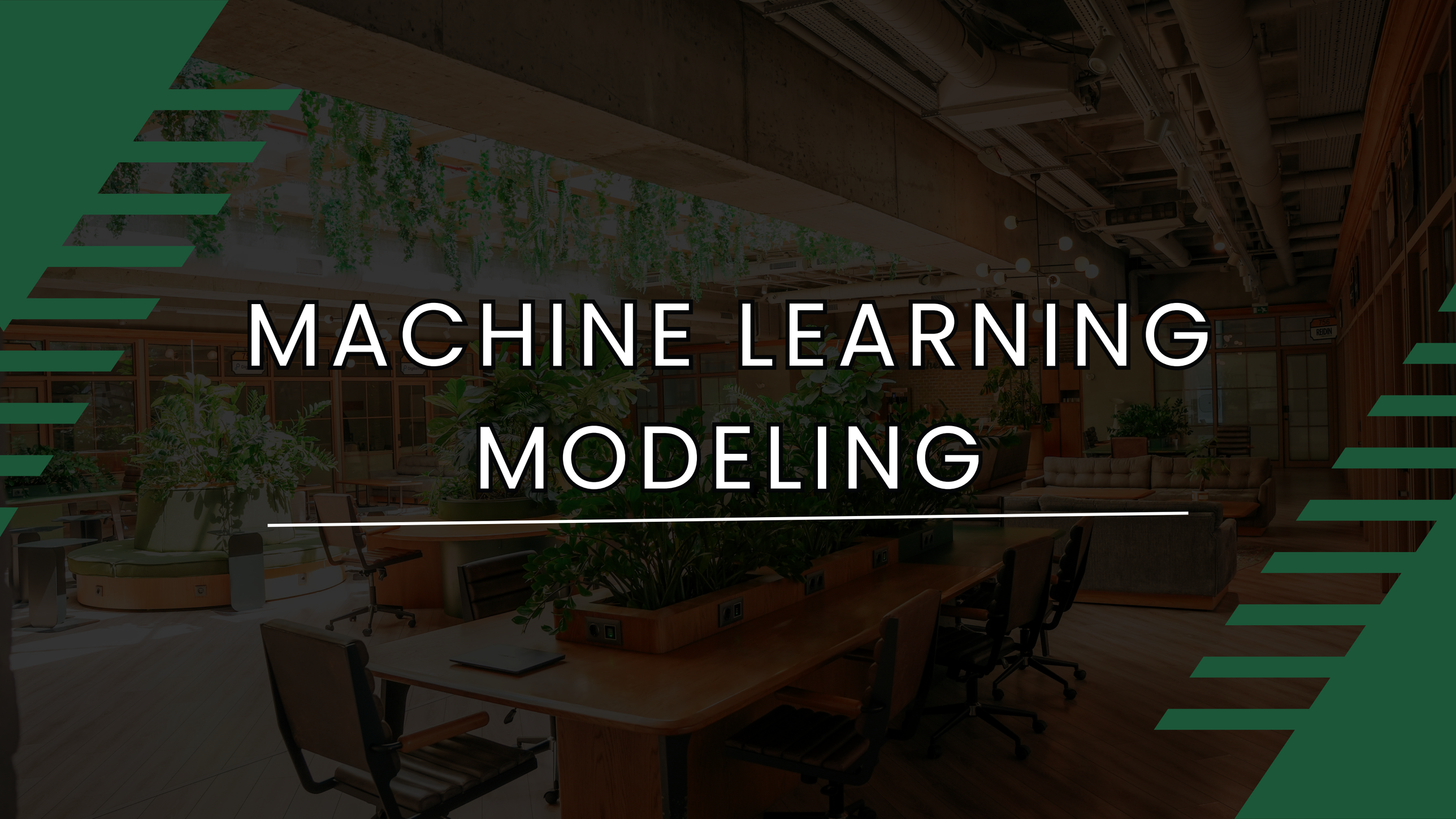
Bookings & Cancellation Over Time



Total bookings peak between August and October, with consistently high activity from April to June. However, in April 2019, highest cancellations also happened during highest total bookings.

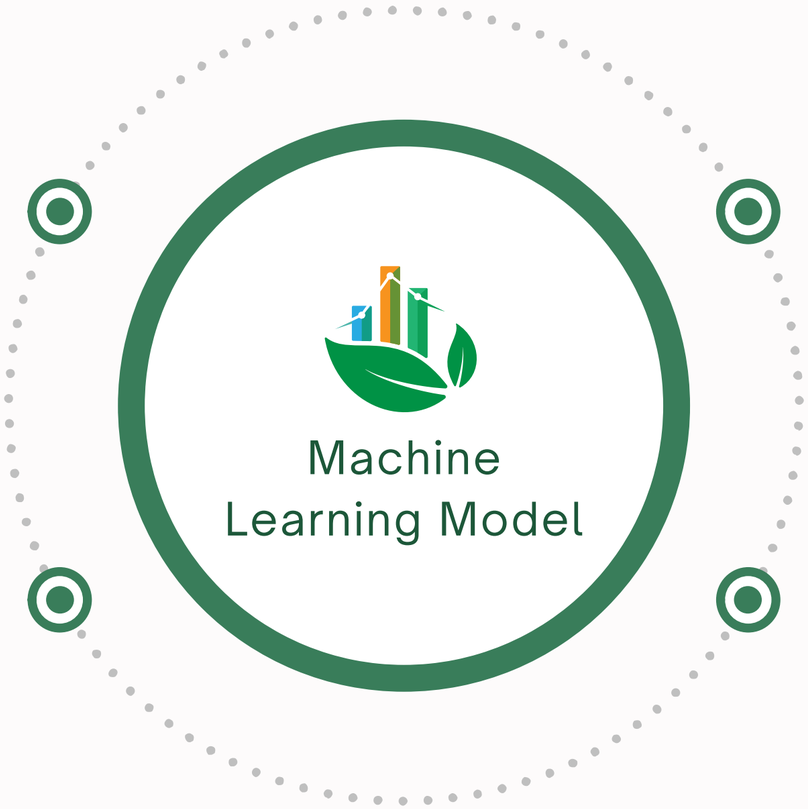


This dashboard presents key metrics such as cancellation rate, lead time impact, deposit type, market segment, and customer type to help stakeholders easily identify booking trends and cancellation patterns. A date slicer is included to filter data over time for better analysis



MACHINE LEARNING MODELING

Tested Models



- Logistic Regression
- Decision Tree
- Random Forest
- LinearSVM
- K-Nearest Neighbor

In the context of predicting hotel booking cancellations, the primary goal is to correctly identify as many guests who are likely to cancel as possible, so that the hotel can take proactive measures to minimize lost revenue and optimize room occupancy.
So our primary metric to see is “Recall"

Model Evaluation

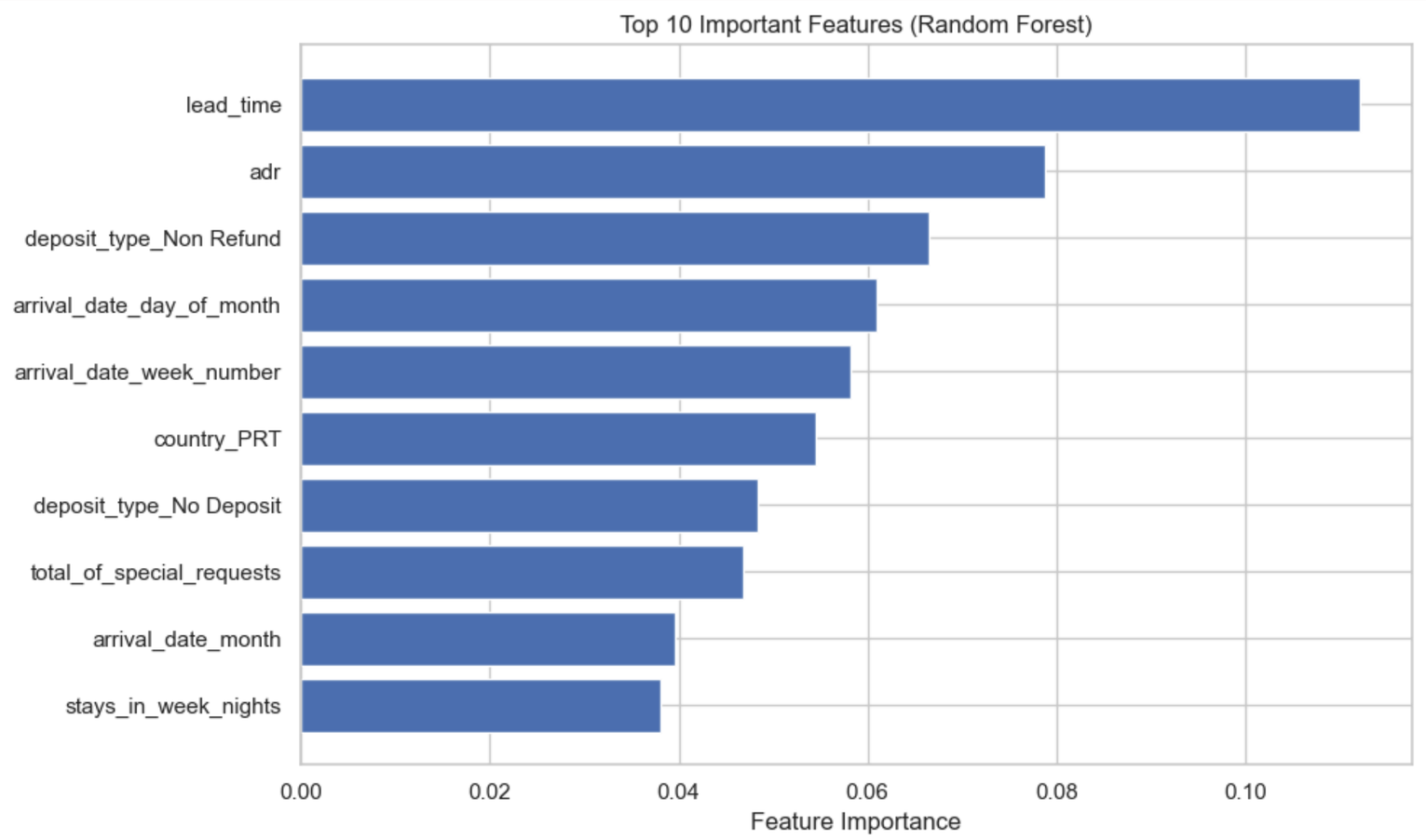
Model	Dataset	Accuracy	Precision (0)	Recall (0)	F1-Score (0)	Precision (1)	Recall (1)	F1-Score (1)
Logistic Regression	Train	0.82	0.82	0.91	0.86	0.81	0.65	0.72
Logistic Regression	Test	0.81	0.82	0.91	0.86	0.80	0.64	0.71
Decision Tree	Train	1.00	1.00	1.00	1.00	1.00	0.99	0.99
Decision Tree	Test	0.84	0.87	0.87	0.87	0.77	0.78	0.78
Random Forest	Train	1.00	1.00	1.00	1.00	1.00	0.99	0.99
Random Forest	Test	0.88	0.88	0.94	0.91	0.88	0.78	0.83
Linear SVM	Train	0.81	0.82	0.91	0.86	0.81	0.64	0.71
Linear SVM	Test	0.81	0.81	0.91	0.86	0.80	0.62	0.70
KNN	Train	0.99	1.00	1.00	1.00	0.99	0.99	0.99
KNN	Test	0.82	0.87	0.85	0.86	0.74	0.77	0.76

Model Evaluation

- Recall for cancellation (class 1) is the most important metric, as it shows how well the model identifies actual cancellations.
- High recall reduces missed cancellations (false negatives), which is more critical than mistakenly predicting a cancellation (false positive).
- Random Forest achieved the highest recall (0.78) on the test data, matching Decision Tree and slightly outperforming K-Nearest Neighbors.
- Although all three models showed some overfitting, Random Forest generalized better and remains the most reliable model for predicting cancellations.

Most Important Feature

- Lead time is the most important factor—longer wait times make cancellations more likely.
- ADR (price) also matters, suggesting that higher prices may lead to more cancellations.
- Deposit types like Non Refund and No Deposit are also key, showing that financial commitment affects cancellation behavior.
- Arrival timing (day, week, month) plays a role too, hinting at seasonal patterns in cancellations.



Insight & Recommendation

Long Lead Time Strongly Predicts Cancellations

Guests with longer lead times are significantly more likely to cancel.

"No Deposit" and "Non Refund" Cancellations is HigherRisk

No Deposit" bookings unsurprisingly show high cancellations due to zero financial commitment, "Non Refund" also experience considerable cancellations

Booking Behavior and Channel Significantly Affect Cancellation Risk

Booking channel and behavior strongly influence cancellations—Online TA and frequent booking changes signal higher risk, while Corporate and Direct bookings are more stable.

Seasonality and External Events Can Cause Booking Anomalies

There’s a clear seasonal pattern with a peak from August to October, during Highest Seasonality Period, there are also an influx of Cancellations.

Insight & Recommendation

Implement incentives for early bookers to commit (e.g., small discounts for non-refundable options) and consider stricter cancellation policies for bookings made far in advance.

Promote partially refundable or prepaid booking options to reduce cancellations from "No Deposit" guests, while offering optional insurance for "Non Refund" bookings to ease customer concerns.

Focus marketing and loyalty efforts on Corporate and Direct booking channels. For Online TA and bookings with multiple changes, implement early follow-up or retention tactics.

Adjust staffing and inventory strategies during peak months (Aug–Oct). And prepare incentives to reduce the amount of cancellations.

Arowwai
Industries

THANK YOU

Raditya Erlang Arkananta

wa.me/6281218900315

LinkedIn

Github Repository

erlang_work@yahoo.com