# Python for Data Science

**Khali El Mahrsi**

## Project 7: Hotel Booking Cancellation Analysis

Arthur Lemoine, Lucile Dubarry

21st November 2025

## I. Introduction

Customers who cancel their reservations are a major challenge in all service economies, particularly in the hotel sector. The following analysis aims to measure the extent of cancellations and identify patterns that explain this behavior from a dataset with detailed information on hotel bookings. The final goal would be to help anticipate cancellations by understanding the factors that drive them. Our key research question: *what customer characteristics can be used to predict whether a booking will be cancelled?*

The overall cancellation rate in the dataset is relatively high: on average, 37% of bookings are cancelled over the full year. As a first step, we examine the determinants of this global rate to identify potential trends, before developing a predictive model of client behavior.

Cancellations generate significant economic losses. Because the dataset does not include information on deposit amounts, we cannot estimate the revenue loss after accounting for deposit retention. However, while considering room price, the revenue loss due to cancellations amounts to $195,721 for the year. Cancellations correspond to 38% of the total revenue from bookings, a value very close to the overall cancellation rate.

## II. Bookings patterns

We first aim to understand the overall booking activity for this hotel brand and identify its main characteristics. This section provides a descriptive overview of the bookings made during the year.

**Overall booking statistics in the year**

- Total bookings: 4,918
- Total clients (adults): 9,221
- Total revenue: $512,235

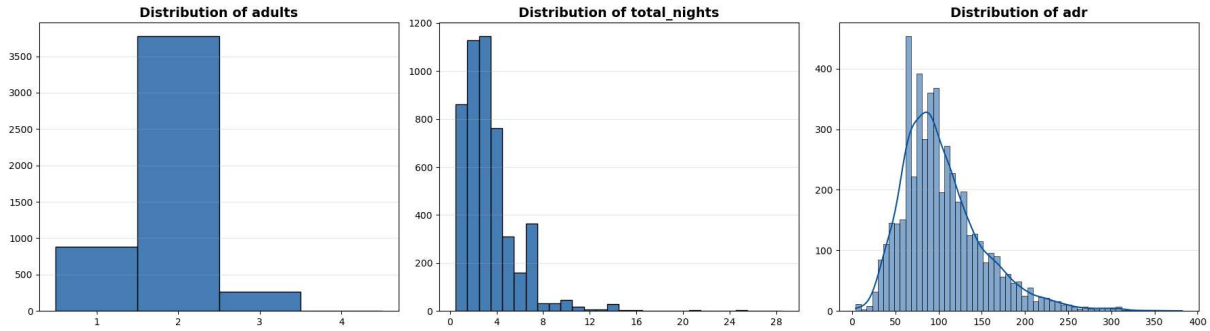
### a) Description of bookings

We describe the bookings using the most relevant variables, differentiating between categorical and numerical variables.

**Distribution of numerical variables**

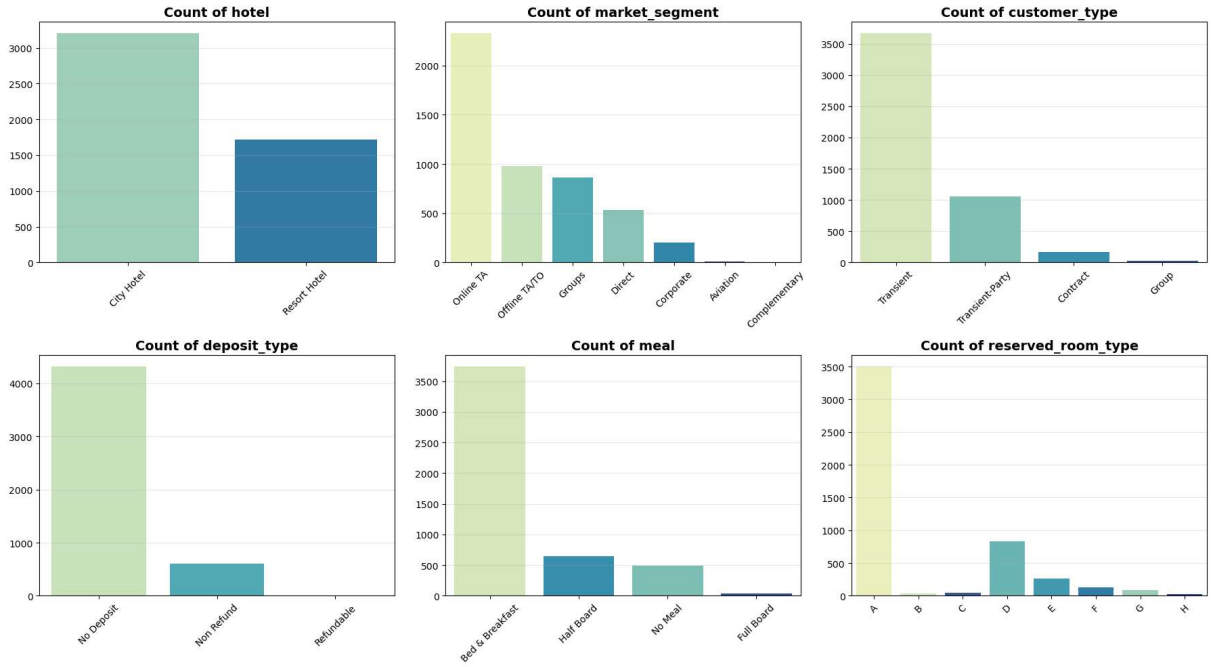
The most common booking profile is: 2 adults, staying 2 to 3 nights, with a room price of around $100 per night. Most clients do not submit special requests; neither do they

require parking spaces, nor do they modify their booking. This highlights the low relevance of those variables in assessing booking behavior.




The distribution of the number of adults has a clear mode of 2, while the distribution of both the number of nights per stay and the average daily rate (adr) seems log-normal.

## Distribution of categorical variables




Bookings made at city hotels represent roughly two thirds of all bookings; the remaining third goes to resorts. In terms of market segment, half of the customers are categorized as online TA, 20% as offline TA/TO, 18% are groups, 11% direct bookings and the remaining going to corporate, aviation and complementary segments.

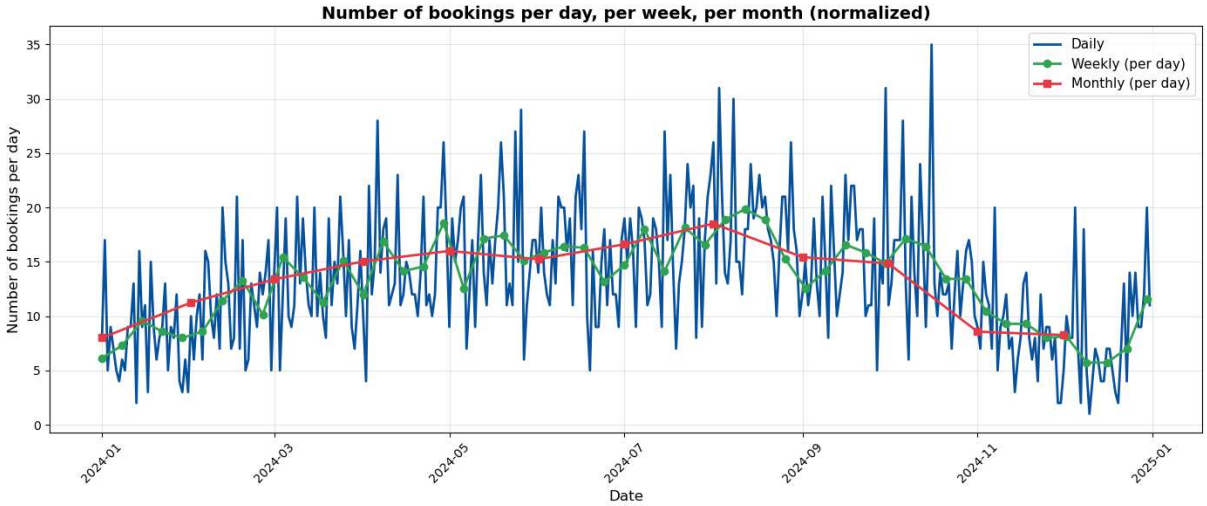
Most bookings do not require any deposit, but when they do, they are never refundable.

A vast majority of bookings include breakfast but no other meal. Similarly, more than 70% of bookings are made to a room of category A.

b) Seasonality in bookings

To better understand how demand fluctuates throughout the year, we analyze booking patterns from a seasonal perspective, as we expect it may be linked to cancellation rates.

**For all hotels**




We plot the number of bookings per day at different time scales: daily, weekly (normalized by the number of days per week), monthly (normalized by the number of days per month). Overall, bookings per month follow a relatively stable seasonal pattern, with slightly higher demand in spring and early summer compared with winter. However, within-month and within-week variations remain high, indicating a large degree of short-term fluctuation.

**By hotel type**

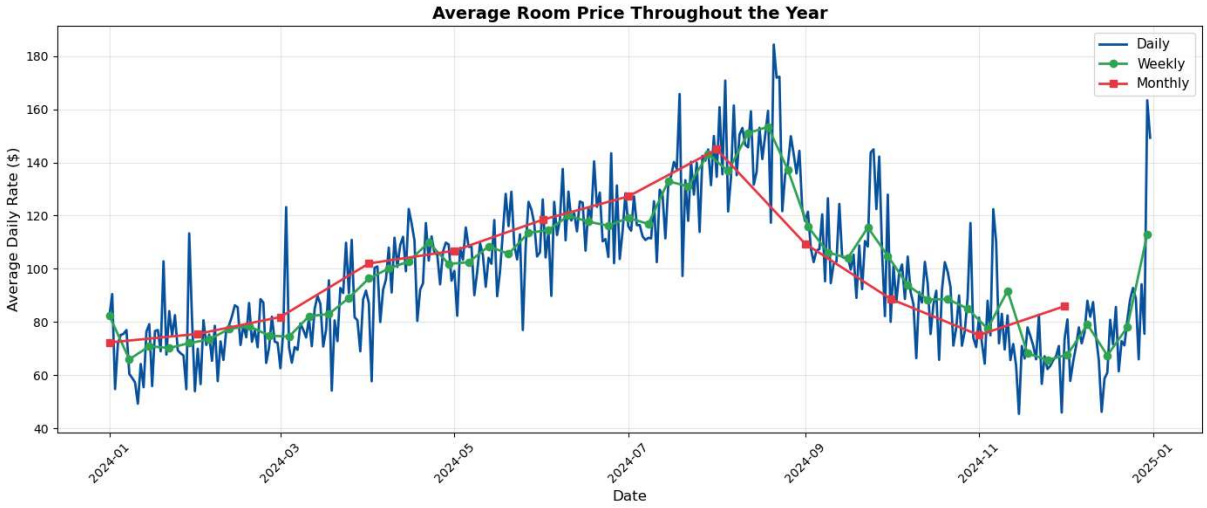
Because seasonal demand may differ across hotel categories, with resorts being more popular in summer, we repeat the analysis separately for city hotels and resort hotels.

The two hotel types do not share the same seasonal pattern. The overall increase in bookings during spring and early summer is entirely driven by city hotels. On the contrary, resorts show a relatively constant demand throughout the year. This first result seems rather counter-intuitive as we would except people to favor resorts as a holiday destination.

c) Price analysis

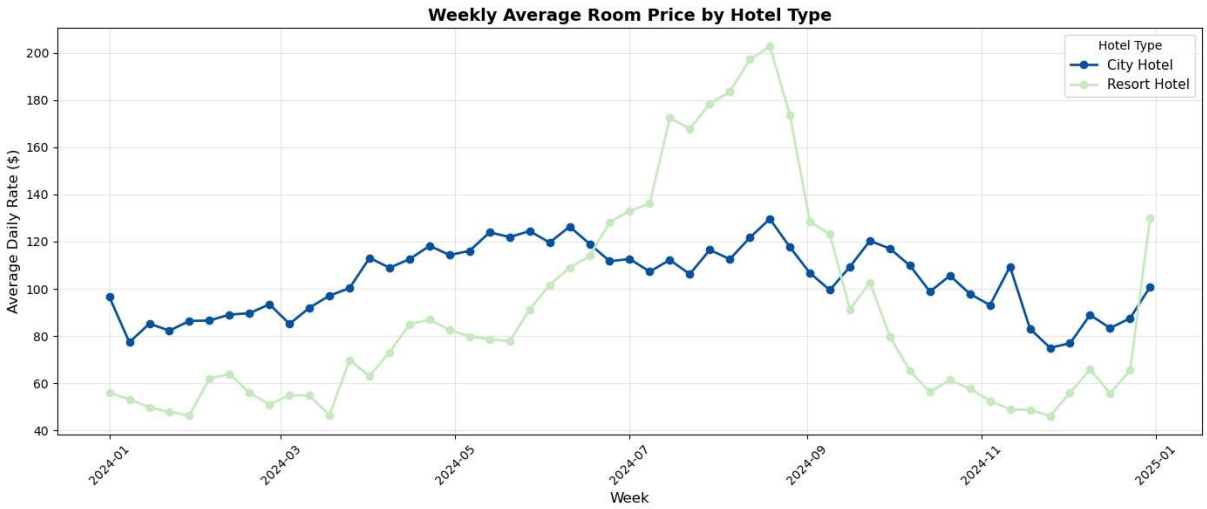
**For all hotels**

As a 101 class in economics could predict, the average room price follows the overall demand for hotel rooms. In other words, the higher the demand, the higher the price. The increase in price around New Year's Eve is strikingly high and suggests that factors outside demand drive the price.



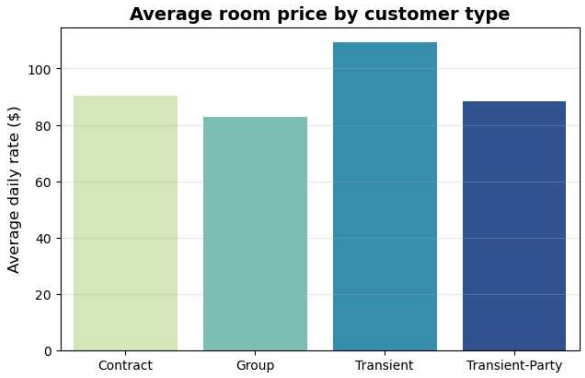

**By hotel type**

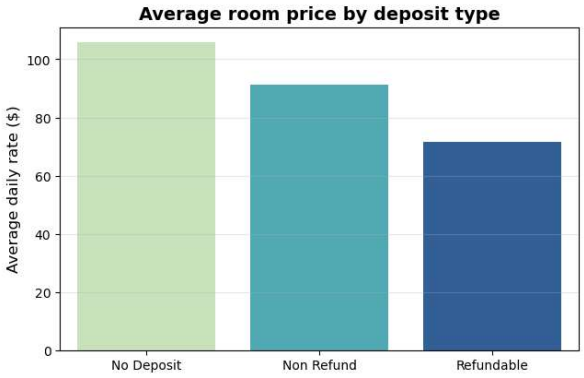
The impact of factors other than demand on prices is confirmed when differentiating between city hotels and resorts. In fact, the overall increase in price during summer is driven by an increase in the daily rate for rooms in resort hotels, while the number of bookings for this type of hotel is relatively flat throughout the year. In contrast, the higher demand for city hotels between June and August does not result in an increase in price.

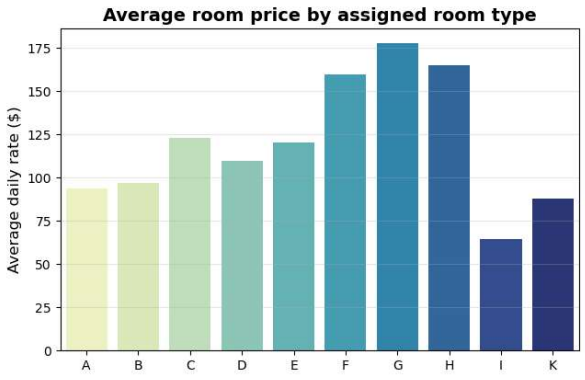



A possible explanation for this strange pattern could lie in the fact that the increase in price is so high that people choose to spend their holiday time in cities rather than resorts, resulting in a suboptimal equilibrium.


Finally, we checked if there was a noticeable difference in terms of room prices between weekdays and weekends. We find only small differences, well in the range of a 95% confidence interval. Thus, there does not seem to be a large difference in prices between weekdays and weekends, conditional on the hotel type.







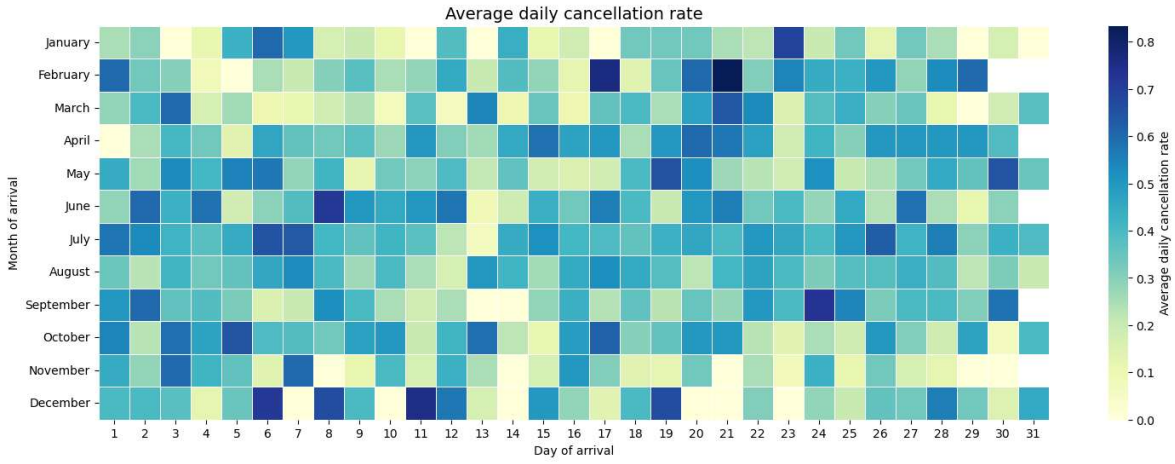
In fact, while there is not much variance in average daily rate among customer types, there are large differences in prices across deposit types and room types. Room quality, as proxied by room price, seems to increase with the alphabet, but only until H. Without further information on the room types, it is difficult to analyze this relation any further. While we could easily explain that rooms with no deposit have a higher price on average, to compensate for the risk of profit loss, it is harder to explain the fact that refundable rooms are much cheaper than non-refundable ones.

However, when we look at the impact of price on cancellation status, there does not seem to be any link between the two variables, hinting towards the fact that cancellations are driven by other factors.

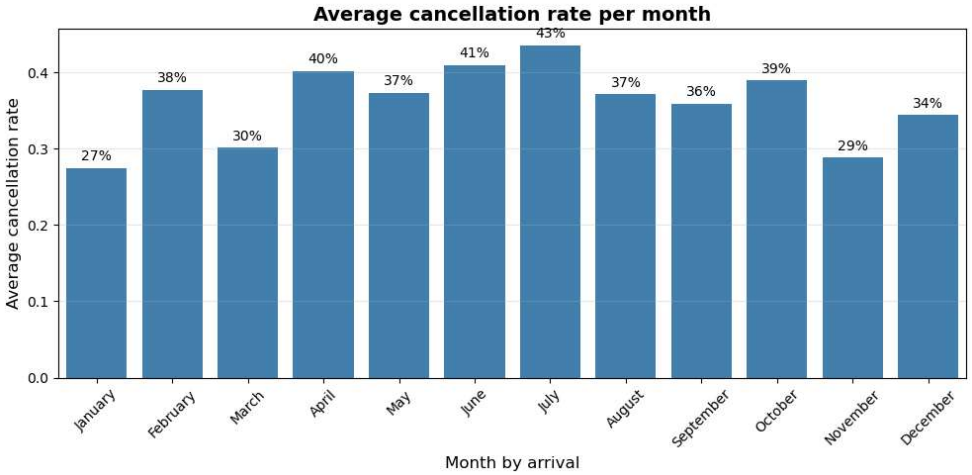
## III. Cancellation patterns

a) Seasonality in cancellations

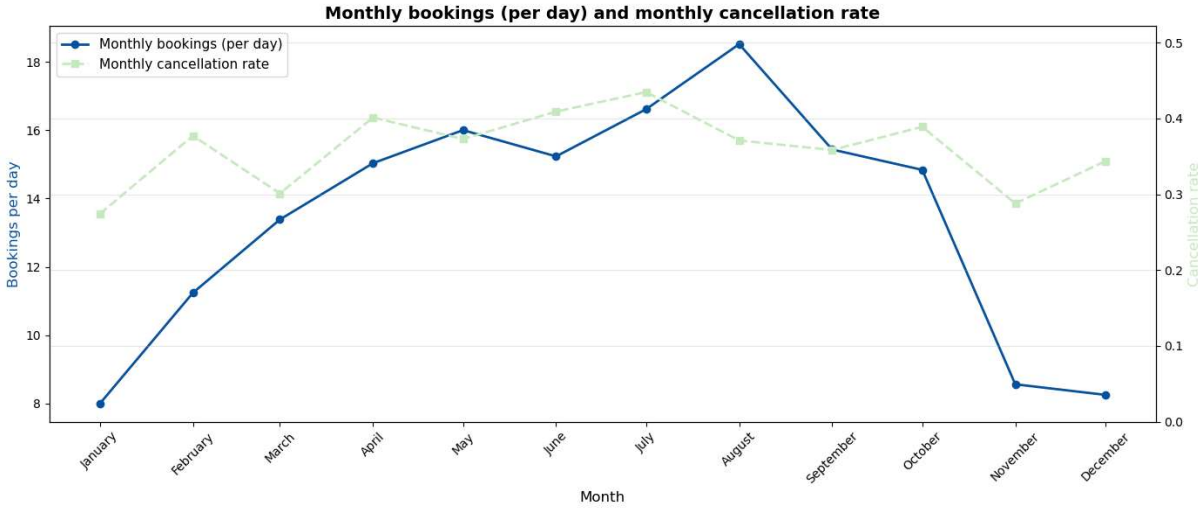
Having examined how bookings evolve across seasons, we now analyze if the cancellation rate follows similar patterns.




When looking at daily cancellation rates, we cannot identify any patterns. Days with high cancellation rates appear to be random, with no visible association with specific dates such as holidays (New Year's Eve, Christmas).




At the monthly level, we observe a slight increase in cancellation rates during spring and early summer. Winter months show the lowest cancellation rates. This may reflect the fact that months with higher booking levels also tend to have higher cancellation rates.

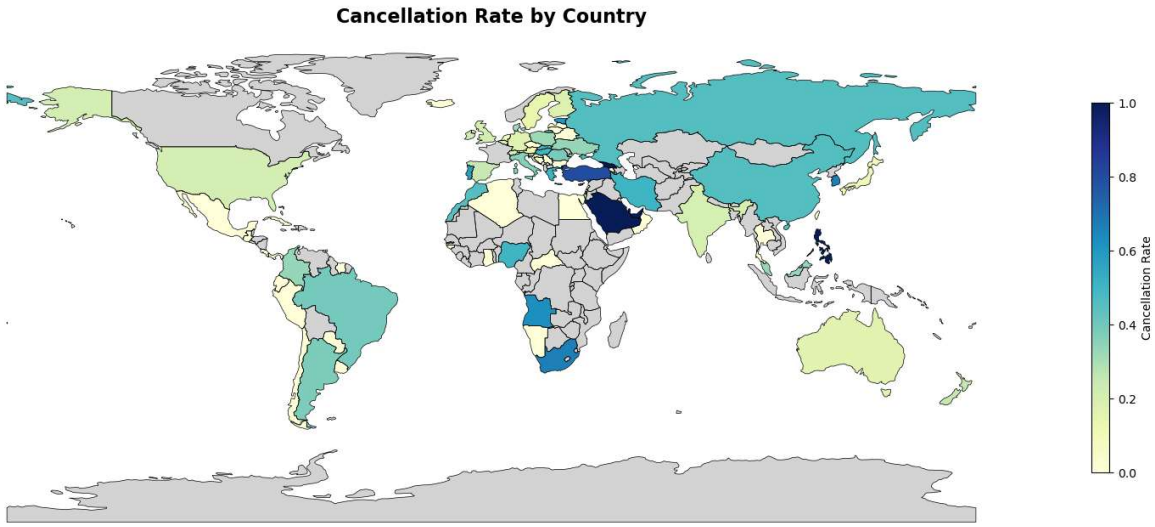



However, when we plot monthly bookings per day and monthly cancellation rates together, we see that the seasonal variation in cancellation rates is very small, relative to the seasonal pattern in bookings. While the average number of bookings per day more than doubles between January (8) and August (18), the cancellation rate does not show a comparable increase.

b) Interaction of cancellations with other variables

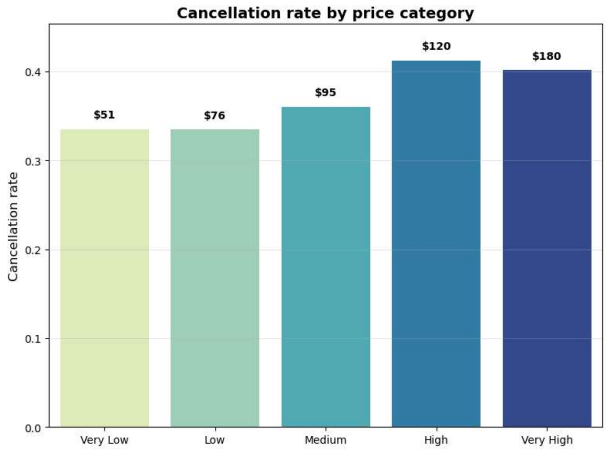
From the previous section we can conclude that there does not seem to be much seasonality in the cancellation rate. Let's now focus on other potential drivers.

## By country of residence



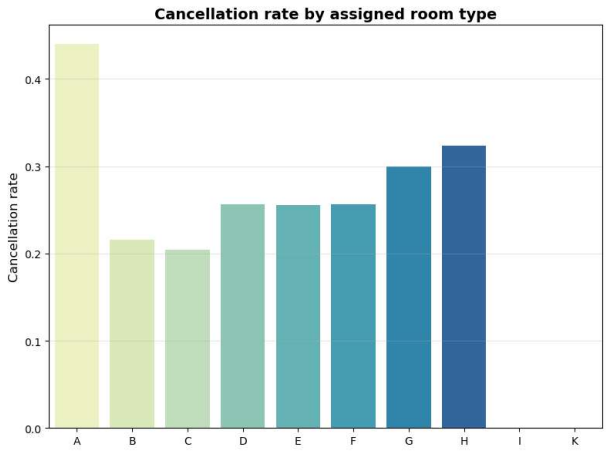

While there are large differences in cancellation rates across countries of residence, we do not observe any striking patterns arising from larger regions or income levels.

## By room price



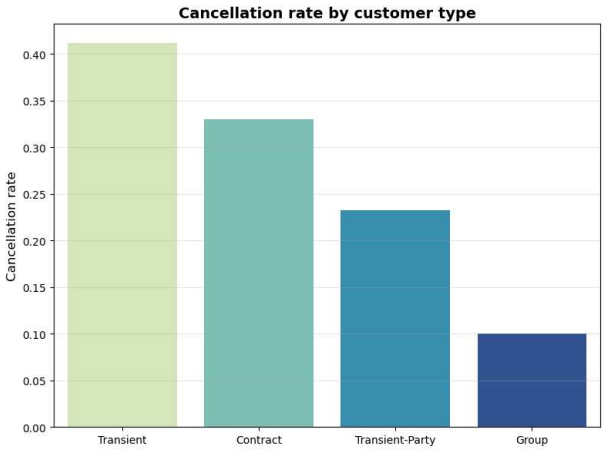

When clustering rooms in 5 categories from very cheap, with an average price of $51, to very expensive (average price of $180), the cancellation rate increases. However, this increase is not monotonous – there is a small decrease in the highest category and rather small in magnitude.

## By room type



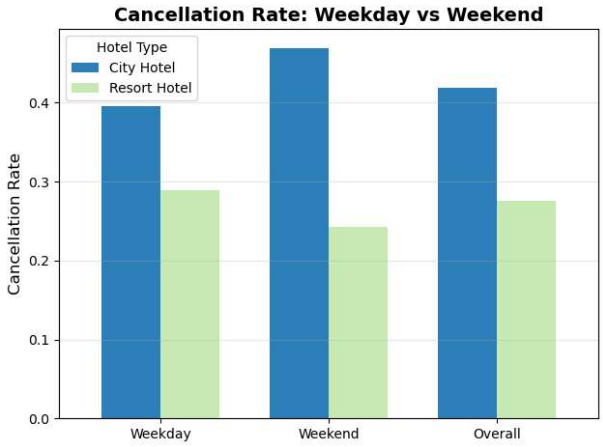

Similarly to the increase in room price with the alphabet, we observe an increase in cancellation rate, except for room type A that seems to be an outlier with a well above average cancellation rate of 43%.

### By customer type



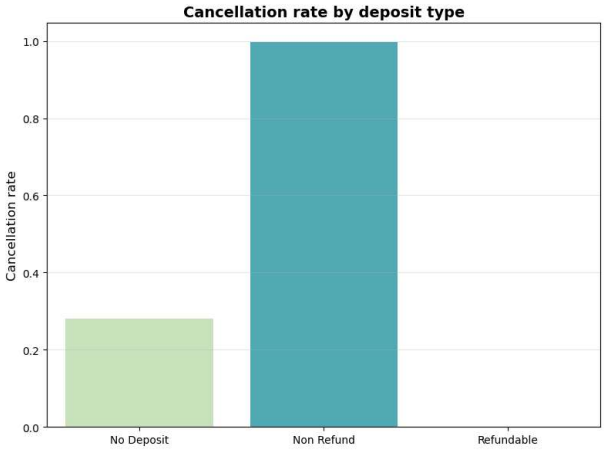

On the one hand, transient customers are guests who are predominantly on-the-move and seek short and often urgent hotel stays (source: Xotels) which can explain that they have above average cancellation rate. On the other hand, group reservations tend to be less canceled, probably because they are planned.

### By hotel type




We observe a much higher cancellation rate in city hotels than resorts. Moreover, the difference is exacerbated when we differentiate between weekdays and weekends.

### By deposit type

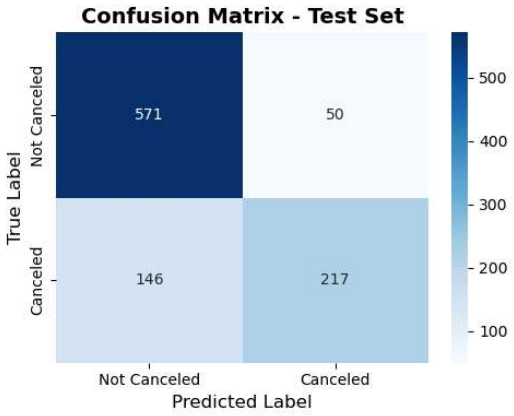



Finally, looking at the cancellation rate by deposit type seems to be the key. While non-refundable rooms have a cancellation rate of 99.7%. The cancellation rate for rooms with no deposit is 28%, which is also well below average. This seems odd as it would mean that the threat of non-refundable deposits plays against hotels. There are no non-refundable bookings in our data, explaining the absolute 0 for this category.

## IV. Machine Learning: predicting cancellations with XGBoost

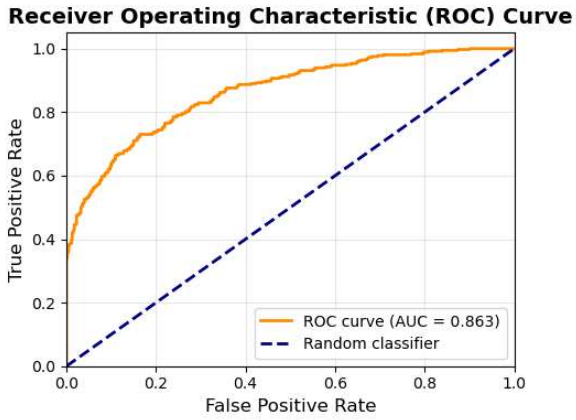
We develop a predictive model of client behavior using a gradient boosting framework (XGBoost algorithm). The objective is to anticipate cancellations by identifying the factors that most strongly drive them. Our key question: *what customer characteristics can be used to predict whether a booking will be cancelled?*

### Model performance



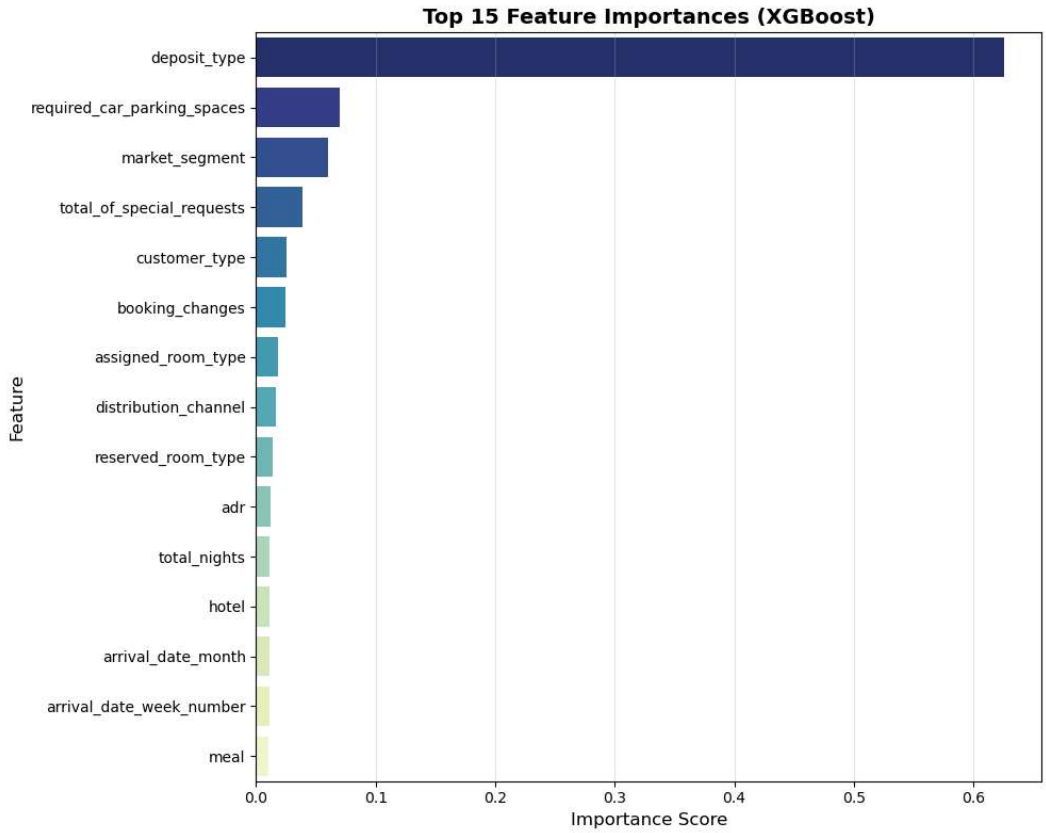

Using the confusion matrix, we compute indicators on how well the model performs:

- Accuracy (share of correct predictions) = 0.8
- Precision (share of correct predictions among predictions cancellations) = 0.8
- Recall (share of cancellations correctly identified) = 0.6




The ROC curve is used to evaluate the performance of a binary classification model. It shows the share of correctly identified cancellations relative to false positive ones. The diagonal line shows the performance of a random classifier. Our model's ROC curve is significantly above this diagonal. The area under the curve (AUC = 0.86) indicates good predictive performance.

**Model findings** through the feature importance analysis




The model reveals that the most important driver of cancellation behavior –and by very far– is deposit type. This is coherent with our descriptive analysis done above:

- Bookings with non-refundable deposits are predicted to be cancelled at a much higher rate (consistent with our finding that non-refundable rooms have an average cancellation rate of 99.7%).
- On the other hand, bookings without a deposit are predicted to be cancelled less frequently (average cancellation rate of 28%).

Beyond deposit type, all other factors have much lower importance on client behavior. The second most influential factor –still 10 times less important than deposit type– is the number of required car parking spaces. However, most clients do not request parking, so this factor has a limited relevance in practice. Other variables –number of nights, customer type, room price– have very low importance in predicting cancellation behavior.

This result appears counterintuitive. We would expect customers to cancel more when they face no financial cost (no deposit or a refundable one). Because this result contradicts with standard economic intuition, we cannot provide clear guidance to the hotel chain on whether they should avoid non-refundable deposits. One hypothesis –but unlikely– is that customers who choose refundable or no-deposit option might be more risk-averse, and therefore less inclined to cancel their booking.