Python Project: Hotel Booking Cancellations Analysis

Vadim Lagresle, Arthur Rochas

Novembre 2024

1 Introduction

According to the dataset, 36.52% of bookings are cancelled, which may justify an analysis of these determinants to better anticipate and, for example, correct overbooking or improve hotel sales.

2 Descriptive statistics

2.1 Variables distribution

Checking the consistency of the distribution of variables, and the absence of anomalies, makes it possible to verify their quality, so as to avoid drawing the wrong conclusions from problems in the data.

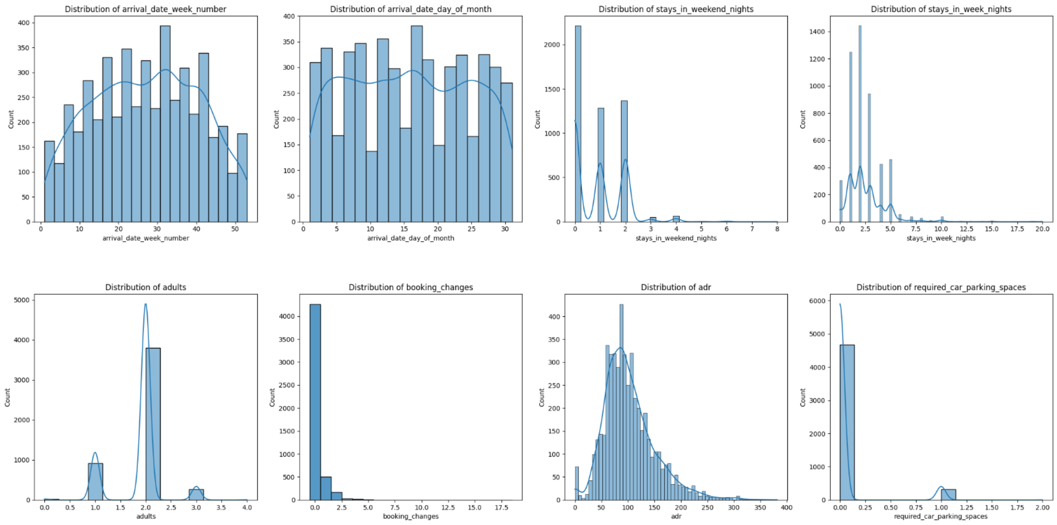


Figure 1

We can see that the numbers of arrival dates are consistent with the usual peak (summer) and off-peak (winter) seasons in the hotel industry. The number of weekend nights (around 0 and 2) is consistent with the number of weekday nights (around 0 and 5). The distribution of the number of adults is also consistent: mostly adult couples, sometimes a single adult and rarely 3 adults, very exceptionally only minors.

We can note that the distribution of the Average Daily Rate (ADR) can be approximated by a Log-normal distribution with positive skewness, which also seems to make sense: the hotel industry tries to match the quality offered (and therefore its prices) to the distribution of incomes.

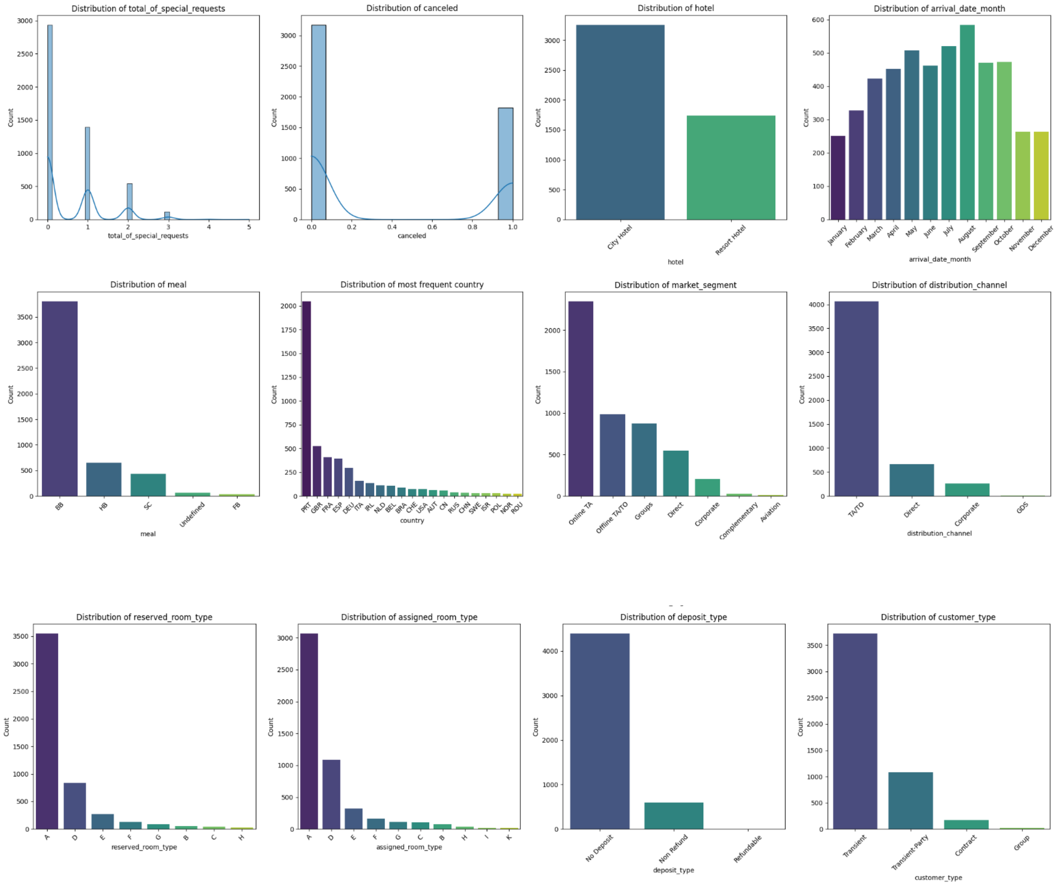


Figure 2

We can see that for many variables (country, market segment, hotel type, room type, meal, etc.), the distribution of numbers is fairly concentrated. The distribution of reserved room type is less concentrated than assigned room type, possibly due to overbooking in the more popular type A rooms.

2.2 Relationships between variables

The relationships between variables can be used to detect potential anomalies and confounding factors between the independent variables and the dependent variable.

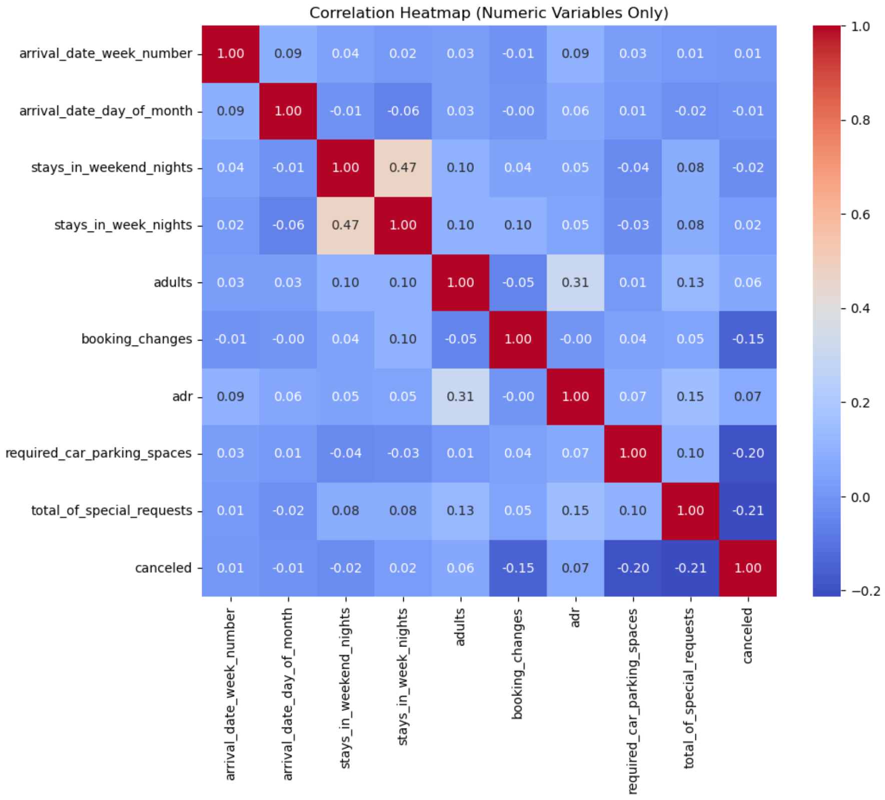


Figure 3

In our correlation matrix, the relationship between the number of adults and ADR is the most important non-trivial correlation between the numerical explanatory variables.

The following graphs illustrate the interactions between some variables:

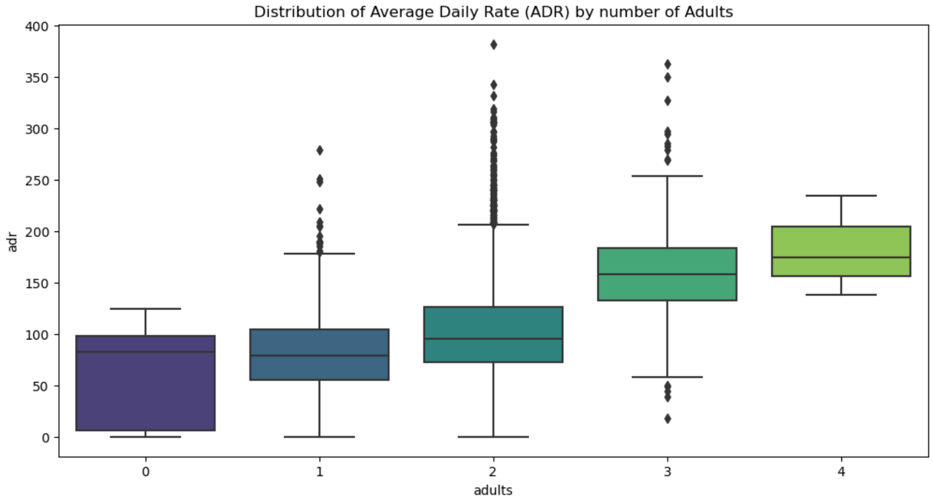
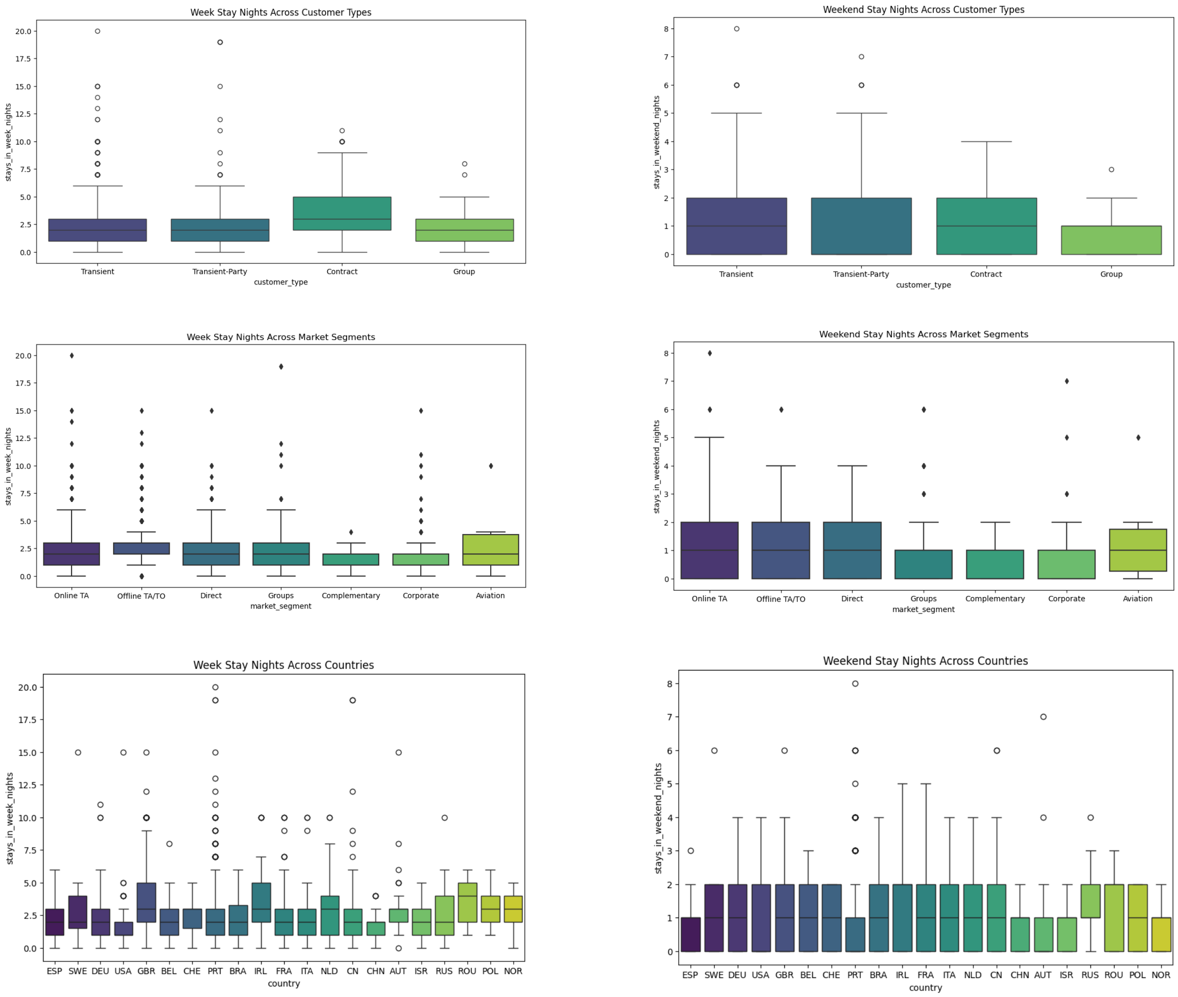
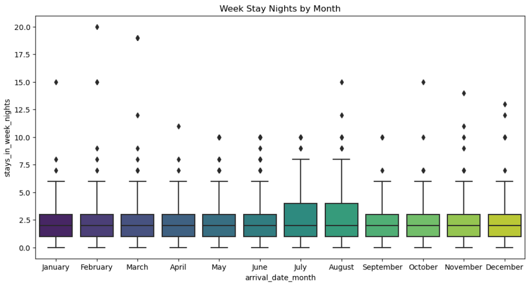
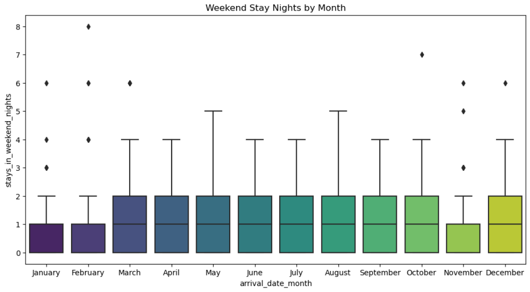
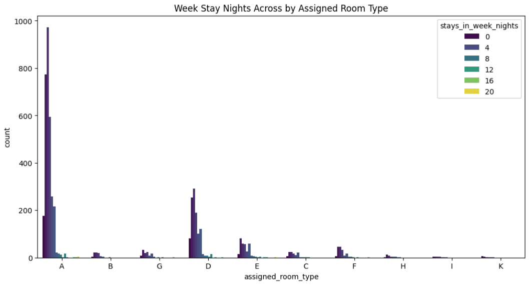


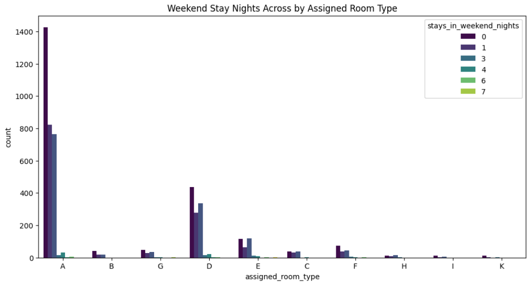
Figure 4: Numerical variables correlations

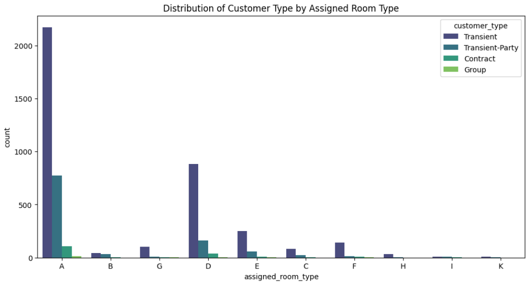


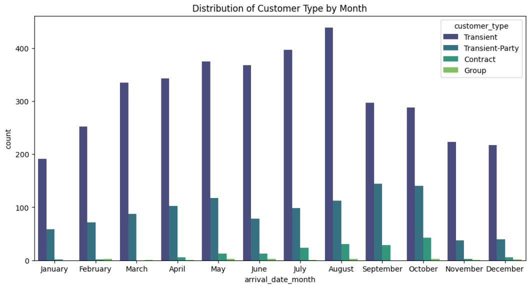












We also test relationships between potential confounding factors (Figures 5 to 10).

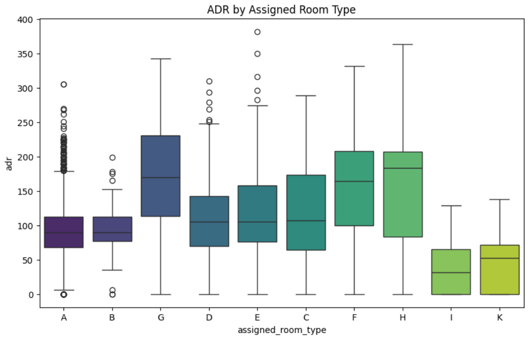


Figure 5: ADR by room

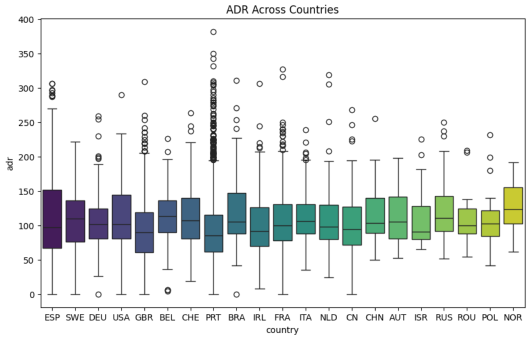


Figure 6: ADR by country

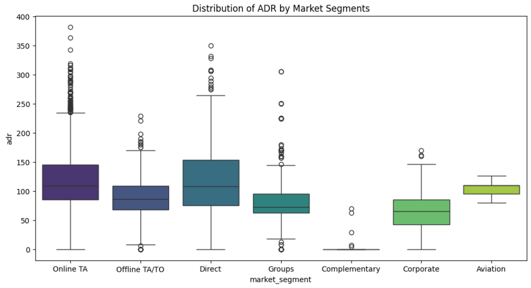


Figure 7: ADR by market segments

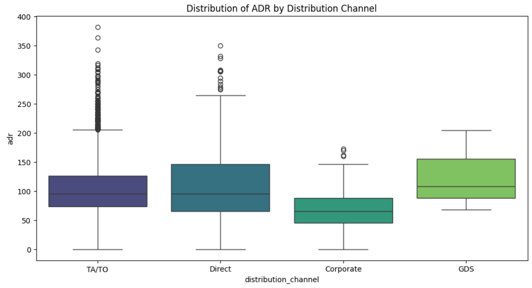


Figure 8: ADR by distribution channel

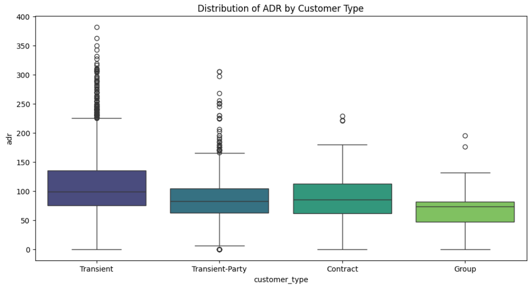


Figure 9: ADR by customer type

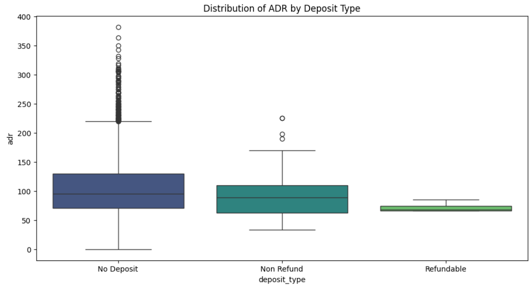
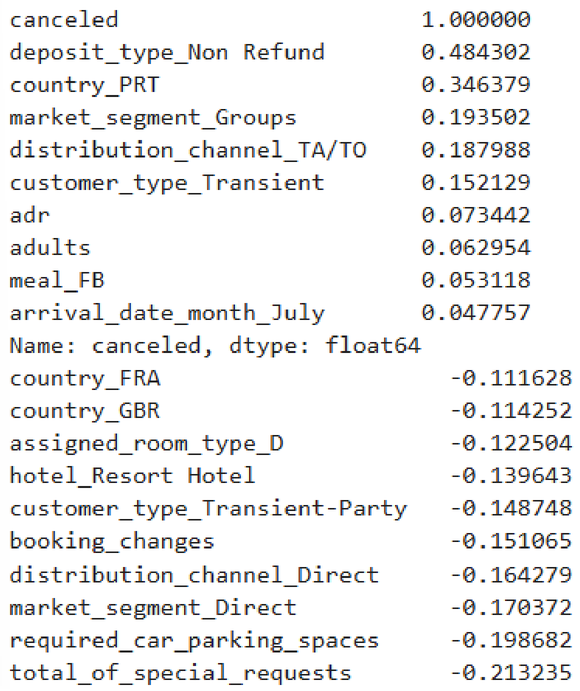


Figure 10: ADR by deposit type

3 Main variables associated with cancellations

The variables most correlated (positively or negatively) with cancellations are described here:



```
canceled                        1.000000
deposit_type_Non Refund         0.484302
country_PRT                     0.346379
market_segment_Groups           0.193502
distribution_channel_TA/TO      0.187988
customer_type_Transient         0.152129
adr                             0.073442
adults                          0.062954
meal_FB                         0.053118
arrival_date_month_July         0.047757
Name: canceled, dtype: float64
country_FRA                       -0.111628
country_GBR                       -0.114252
assigned_room_type_D              -0.122504
hotel_Resort Hotel                -0.139643
customer_type_Transient-Party     -0.148748
booking_changes                   -0.151065
distribution_channel_Direct       -0.164279
market_segment_Direct             -0.170372
required_car_parking_spaces       -0.198682
total_of_special_requests         -0.213235
```

Customer commitments to their reservation (number of special requests, car parking spaces requests, booking changes) are strongly negatively correlated with future cancellations.

ADR is an indicator positively correlated with cancellations. This is intuitive, as it is linked to customer cost, and customers should logically seek to change hotels if they find a cheaper option, but seems a minor cue compared to other variables.

However, making simple correlations confronts us with a potential problem of confounding variables. For example, number of adults booking is also positively correlated with cancellations, but this is probably due to a confounding factor. Intuitively, this may be due to the fact that the number of adults and ADR are positively correlated (see Figure 4). In other words, from a causal point of view, some variables would be strongly correlated with cancellations because they are correlated with variables that have a true causal impact on the probability of cancellation. Conversely, perhaps certain market segments or customers are more demanding, which both decreases ADR and increases cancellation probabilities, potentially creating a bias. Figure 7 illustrates a potential negative bias between ADR and cancellation, using the “direct” and “groups” segments as examples. The same applies to figure 8 for the “TA/TO” and “Direct” distribution channels, and figure 9 for “transient” customers.

A first classification model with random forest and data set seems to support, although a random forest is not in itself a method of causal inference, the idea that it is indeed ADR that mainly has a direct effect on cancellation and not other variables, and serves as a “mediator”.

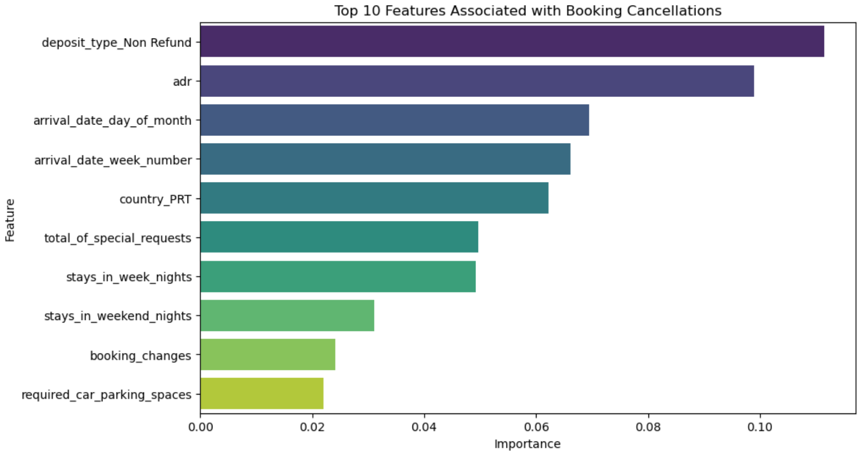


Figure 11

3.1 Data error handling

The previous ML model shows that the “Non Refund” guarantee deposit is the main predictor of cancellation, but the effect seems too large not to be due to an error in the data: 99

We therefore generated a second model, without the data with a “Non Refund” guarantee deposit, as this data is potentially erroneous.

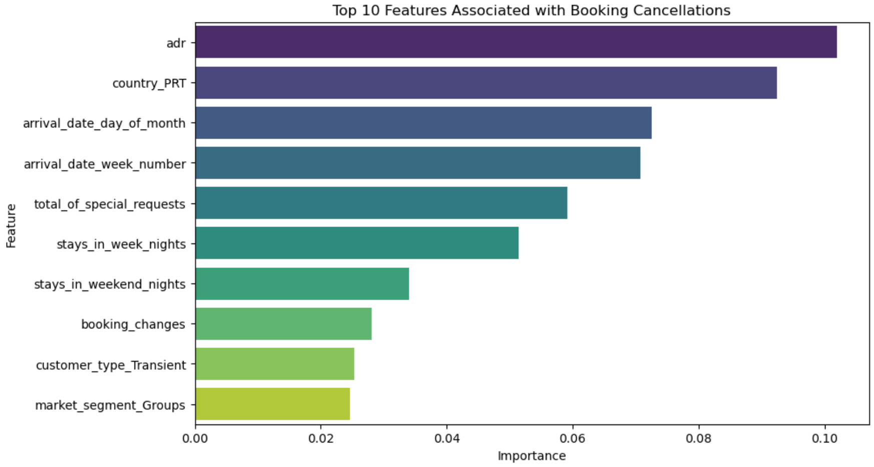
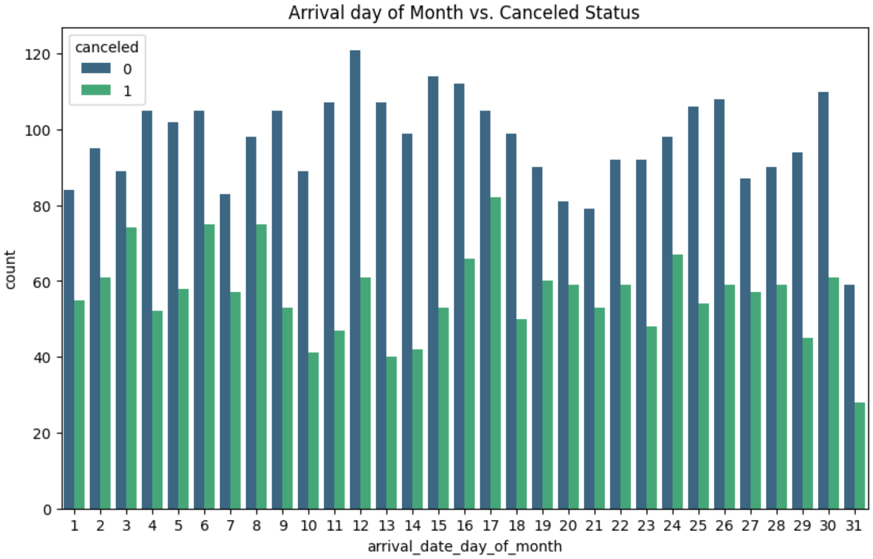


Figure 12

We can see that ADR is potentially the main mediator of cancellations. We then see that, according to this model, a Portuguese customer is an important predictor of cancellations. This could be explained either by cultural or geographical reasons (given that Portuguese customers are the main ones), or by measurement error or a confounding factor. Other explanatory factors, such as special requests or booking changes, appear to be a measure of the degree of customer involvement, and reduce the risk of cancellation for these customers.



The relationship between day of arrival and cancellation is not statistically significant according to a chi-square test (p-value greater than 0.05).

4 Conclusions

ADR seems to be the main mediator in booking cancellations. The hotel must therefore weight its price to obtain the best trade-off between ADR and number of cancellations. The Portuguese nationality of the customer seems to be a risk factor for cancellations, if we consider that the correlation found is not due to a confounding factor or an error in the data.

The hotel can also use the predictions of our models. In the test sample, representing around a third of the initial sample, our two models identified 67

The analysis could be extended with an index of market concentration in the geographical area where the hotel is located to describe its interaction with ADR, since according to the Hotteling model, if a hotel is isolated from its competitors, it could afford to implement a high ADR while limiting the number of cancellations.