April 2024

Forecasting hotel booking cancellations: analyzing customers' behaviors with data mining techniques

Universidad Politecnica De Madrid
Data Mining
Final project report

Matteo Del Prato
Federico Paschetta

Table of Contents

1. Introduction

Cancellation of bookings is a major problem in the hospitality industry. This report builds on the premise that an attempt to analyze hotel booking data has the objective of trying to predict cancellations as a strategy to help hotels better decide ways of stemming the losses.

At its core, the issue of hotel booking cancellations represents a significant economic burden for hotel owners and operators. Cancellations not only result in immediate revenue loss but also disrupt revenue forecasting, inventory management, and resource allocation. Moreover, the ripple effects extend beyond the financial realm, encompassing reputational damage and customer dissatisfaction, both of which can erode long-term competitiveness and brand loyalty.

According to industry reports, the average cancellation rate for hotel bookings typically ranges from 10% to 20%. This statistic can vary based on factors such as location, seasonality, and booking channel. A study conducted by STR revealed that hotel cancellations cost the global hospitality industry an estimated $10 billion annually. This figure encompasses direct revenue losses resulting from cancellations, as well as indirect costs associated with operational inefficiencies, such as overbooking and staffing adjustments. A study published in the International Journal of Hospitality Management found that hotels experience an average revenue loss of 5% to 15% due to cancellations, with variations based on factors such as booking lead time and room type.

In light of these sobering statistics, it becomes evident that addressing the issue of hotel booking cancellations requires a proactive and data-driven approach. By harnessing the power of predictive analytics and advanced data mining techniques, hotel owners can gain actionable insights into the underlying drivers of cancellations, enabling them to implement targeted strategies aimed at mitigating risks and optimizing revenue generation.

This will therefore yield patterns, future probabilities of cancellations and recommendation actions based on historical data of the hotel management. In particular, it will include several stages: descriptive, when it provides characteristics of the abandonment rate; diagnostic, due to the investigation of the factors that affect the abandonment rate; predictive, as the basis for trend predictions in abandonment behavior; and prescriptive.

2. Dataset description

For this project, the dataset "HOTEL_BOOKINGS.csv" was chosen as the primary and only source to work with. It contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

The data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The data was downloaded and cleaned by Thomas Mock and Antoine Bichat for #TidyTuesday during the week of February 11th, 2020.

The datset structured with 119,390 entries and 32 columns. Some of the key columns include details on the hotel type, cancellation status, booking dates, stay duration, customer type, and charges.

To summarize the cancellations specifically, I have calculate the total number of bookings that were canceled, which corresponds to entries where the value of the is_canceled column is 1. There are 44,224 cancellations recorded in the dataset.

Covering a timespan of bookings from July 2015 to August 2017, the dataset spans enough time to assess seasonal impacts, market evolution, and consumer trend shifts. Each entry encapsulates data points from the booking time up to the potential cancellation, offering a longitudinal perspective on the reservation lifecycle.
Among its myriad of fields, the dataset documents the lead time for bookings, length of stay, number of guests including adults, children, and babies, type of meal booked, and other key factors.

In addition to the raw booking data, the dataset incorporates derived variables such as the total number of stays on weekends and weekdays, enabling a nuanced analysis of stay patterns. Moreover, the inclusion of details regarding previous cancellations and bookings offers a retrospective lens to analyze customer reliability and booking patterns.

Below you can see the table consisting of the list of features, their description and type.

FEATURE	DESCRIPTION	TYPE
hotel	Hotel (H1 = Resort Hotel or H2 = City Hotel)	Categorical
is_canceled	Value indicating if the booking was canceled (1) or not (0)	Numerical
lead_time	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date	Numerical
arrival_date_year	Year of arrival date	Categorical
arrival_date_month	Month of arrival date	Categorical
arrival_date_week_number	Week number of year for arrival date	Categorical
arrival_date_day_of_month	Day of arrival date	Categorical
stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	Numerical
stays_in_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel	Numerical
adults	Number of adults	Numerical
children	Number of children	Numerical
babies	Number of babies	Numerical
meal	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal	Categorical
country	Country of origin. Categories are represented in the ISO 3155–3:2013 format	Categorical

FEATURE	DESCRIPTION	TYPE
market_segment	Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"	Categorical
distribution_channel	Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"	Categorical
is_repeated_guest	Value indicating if the booking name was from a repeated guest (1) or not (0)	Numerical
previous_cancellations	Number of previous bookings that were cancelled by the customer prior to the current booking	Numerical
previous_bookings_not_canceled	Number of previous bookings not cancelled by the customer prior to the current booking	Numerical
reserved_room_type	Code of room type reserved. Code is presented instead of designation for anonymity reasons.	Categorical
assigned_room_type	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due	Categorical
booking_changes	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS	Numerical
deposit_type	Indication on if the customer made a deposit to guarantee the booking.	Categorical
agent	ID of the travel agency that made the booking	Categorical
company	ID of the company/entity that made the booking or responsible for paying the booking.	Categorical

FEATURE	DESCRIPTION	TYPE
days_in_waiting_list	Number of days the booking was in the waiting list before it was confirmed to the customer	Numerical
customer_type	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking	Categorical
adr	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights	Numerical
required_car_parking_spaces	Number of car parking spaces required by the customer	Numerical
total_of_special_requests	Number of special requests made by the customer (e.g. twin bed or high floor)	Numerical
reservation_status	Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why calendar_today	Categorical
reservation_status_date	Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel	Categorical

Table 1: features, description and type

3. Business understanding

3.1 Business goals

The primary business goal of this project is to increase by 10 % overbookings rates in hotels by analysing booking cancellations.

Hotels can proactively implement management strategies to mitigate the adverse financial and logistical repercussions. These insights will enable hotels to optimize occupancy rates, refine overbooking practices, and enhance customer relationship management.

Understanding the propensities for cancellation is crucial for adjusting overbooking levels and defining cancellation policies that balance profitability with customer satisfaction. By leveraging historical booking data, this analysis seeks to reveal patterns and indicators that preemptively signal the likelihood of a booking being canceled.

By aligning inventory management with predictive cancellation insights, the hotel can better allocate rooms, adjust pricing dynamically, and manage customer expectations through personalized communication and offers. In essence, this analysis is not just about predicting cancellations but about enabling the hotel to engineer a robust booking system that maximizes occupancy and revenue while maintaining high standards of customer service.

3.2 Business questions

Regarding the analysis of booking cancellations, different questions were formulated to guide the analysis. Such questions fall under Descriptive, Diagnostic, Predictive and Prescriptive analyses categories.

3.2.1 Descriptive analysis

- What is the distribution of canceled versus non-cancelled bookings?
- What are the characteristics of bookings that are most frequently canceled (e.g., booking duration, time of year, customer type)?
- How do cancellation rates differ across different hotel types (resorts vs. city hotels)?
- What trends are observable in booking cancellations over the years or seasons?

3.2.2 Diagnostic analysis

- Is there a correlation between the lead time of a booking and its probability of being canceled?
- How do external factors such as holidays impact cancellation rates?
- How do different variables influence cancellation probabilities?
- Are there variables which are correlated?

3.2.3 Predictive analysis

- Which models can we build to best predict cancellation?
- How accurate are our predictive models?
- Which variables are the most explicative in our model?

3.2.4 Prescriptive analysis

- What strategies can hotels implement to reduce the number of cancellations based on the insights gained from predictive analysis?
- How can personalized marketing or tailored booking options impact cancellation rates?
- What operational changes could be made to minimize losses from cancellations?
- Can dynamic pricing or overbooking strategies be optimized based on predictive cancellation rates?

3.3 Data mining goal

The primary objective of the data mining effort in this project is to increase overbookings ratesin hotels by 10 %. The analysis aims to identify the key factors that significantly influence booking cancelations. To achieve this, a range of data mining techniques has been performed:

- Logistic regression
- Decision Tree
- Neural Network

The outcome of the report will help hotel managers in making data-driven decisions to avoid economic losses due to cancellations and increase overall profitability.

4. Data understanding: descriptive and diagnostic analysis

4.1 Project setup

The first steps consisted in uploading and analysing the .csv file. Firstly, I decided to drop "reservation_status" and "reservation_status_date" variables since they provide information about target variable, which we wouldn't have in a real use case scenario. I then checked if any null values were showing up. Below you can see the list of null values found, and the actions addressed:

FEATURE	NUMBER OF NULL VALUES	ACTION	REASON
children	4	dropped	the number of null values is very low compared to the total data (119,390)
country	488	dropped	the number of null values is very low compared to the total data (119,390)
agent	16.340	filled	features have value = 0, meaning costumer didn't book via a travel agent. I filled them using a new value ("no_agent")
company	112.593	filled	features have value = 0, meaning costumer didn't book via travel company. I filled them using a new value ("no_company")

Table 2: null features, number, action and reason

4.2 Outliers analysis

After pre-processing the data, I analysed the outliers in order to decide which actions to take with them. I considered all features, and took tailored made actions to aim for a better result at the end. To identify those outliers, I used the box plot so that I could see a visual summary of the distribution of the variable and easier detect those values that fall far outside the expected range.

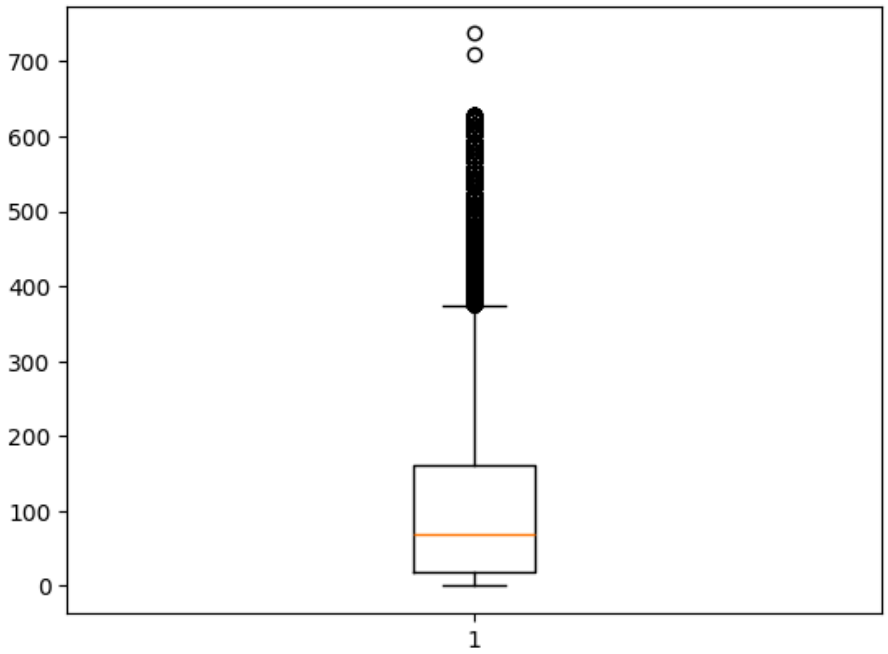

Figure 1: Box plot with outliers of the "lead_time" variable

For the “lead_time” variable, the rows with values 737 and 709 were dropped as I considered them too extreme for the analysis.

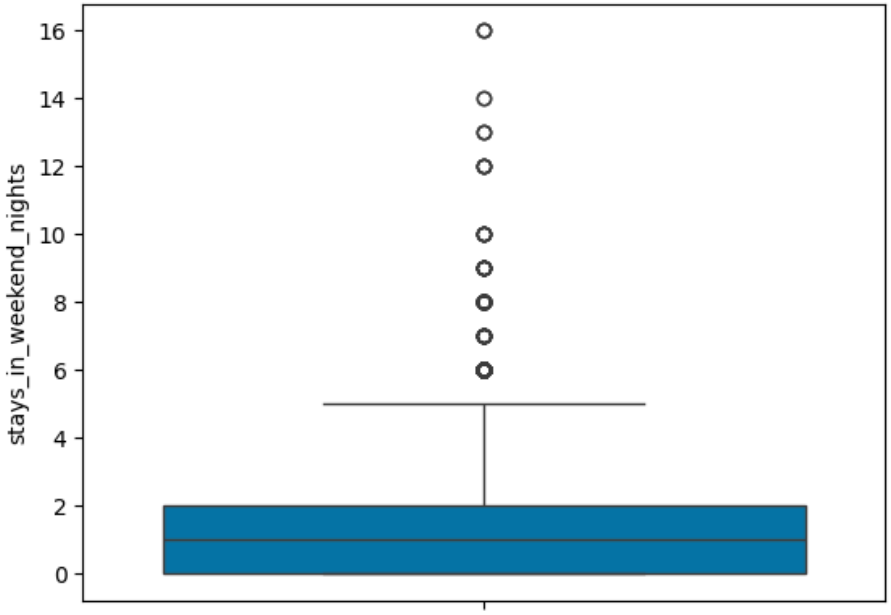

Figure 2: Box plot with outliers of the “stays_in_weekend_nights”variable

Regarding the “stays_in_weekend_nights” variable, I dropped a total amount of 13 rows, all containing values higher than 10. This was due to the fact that I only considered stays under one month and a half.

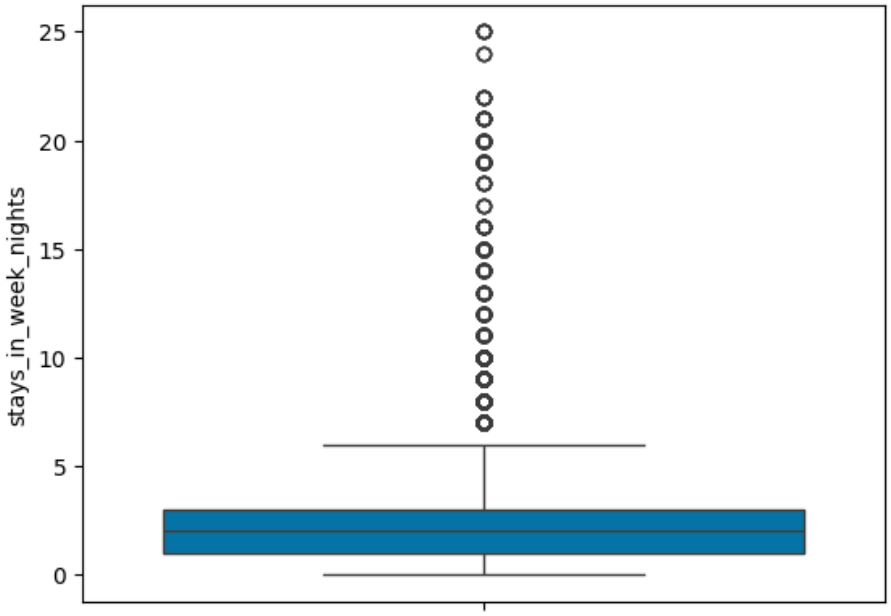

Figure 3: Box plot with outliers of the "stays_in_week_nights" variable

No actions were taken for the "stays_in_week_nights" variable since potential outliers for this feature were already dropped in the previous one.

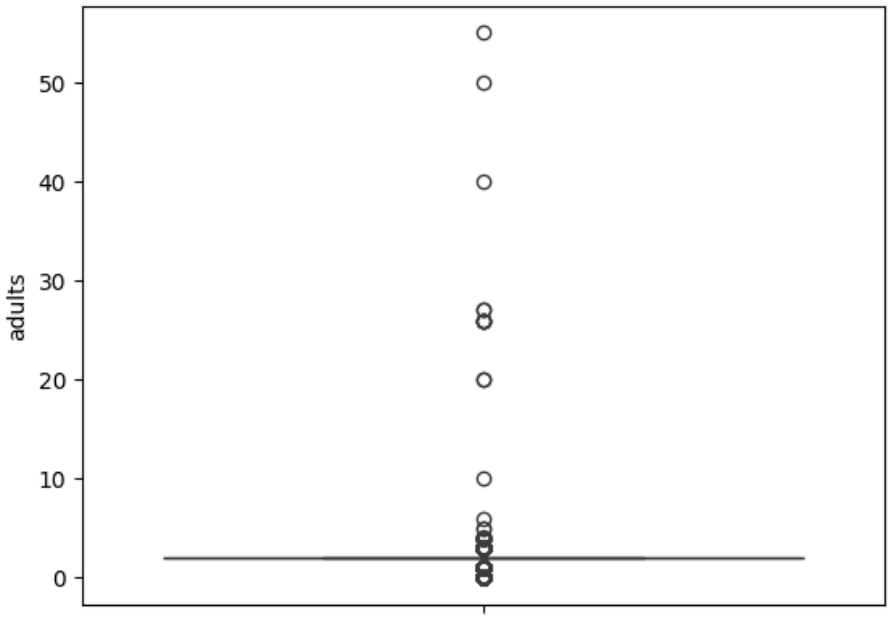

Figure 4: Box plot with outliers of the "adults" variable

Regarding the "adults" variable, a total amount of 13 outliers (rows with values >= 10) were dropped. Although more people cancelling equals a more money being lost, I decided to focus on the bigger cluster of bookings number - which included less than 10 adults - to get a more precise output.

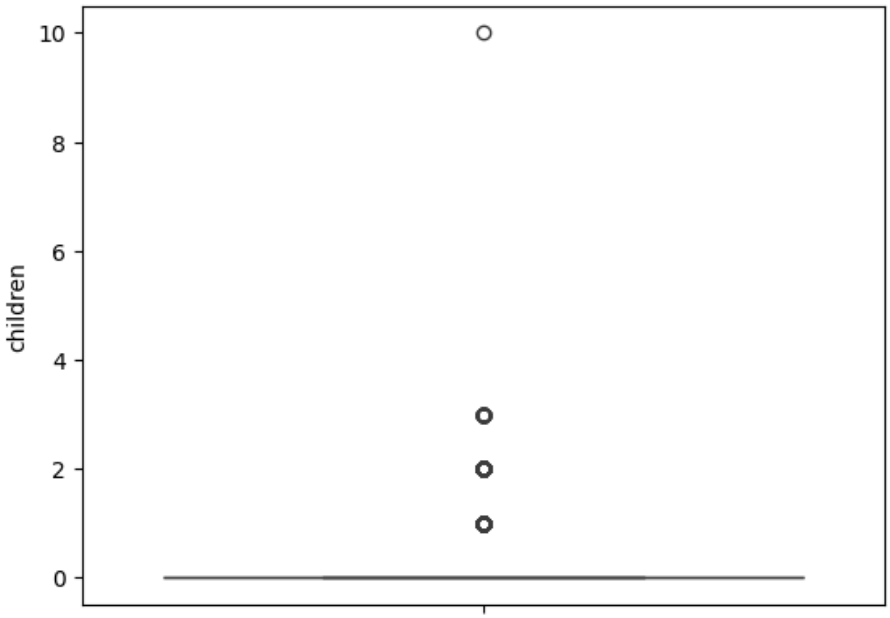

Figure 5: Box plot with outliers of the "children" variable

Considering the "children" variable, only one outlier (a single row with value = 10 children) was dropped, since it is too distant from the other values (including number of children from 0 to 3).

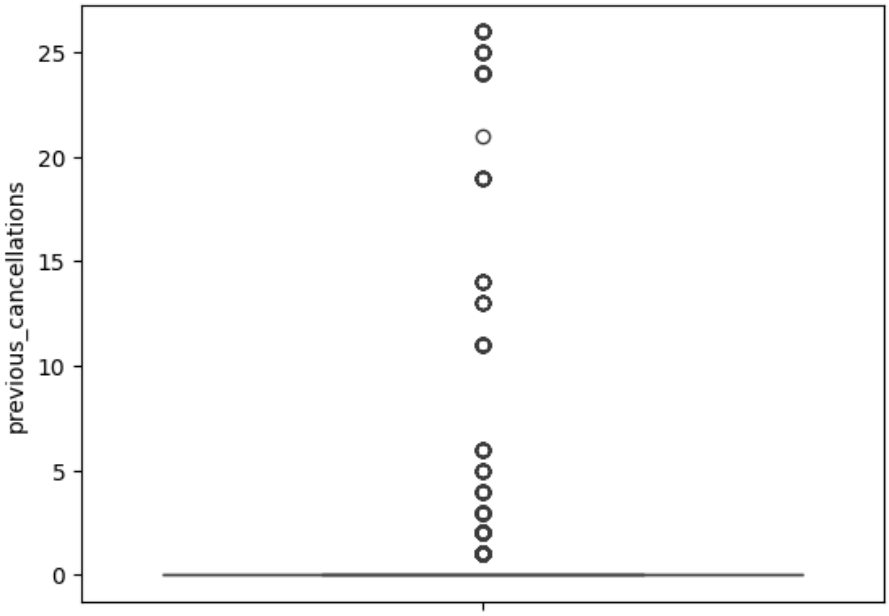

Figure 6: Box plot with outliers of the “previous_cancelations” variable

I decided not to drop any outlier from the “number_of_cancellations” feature in order to know how likely a consumer is to cancel again.

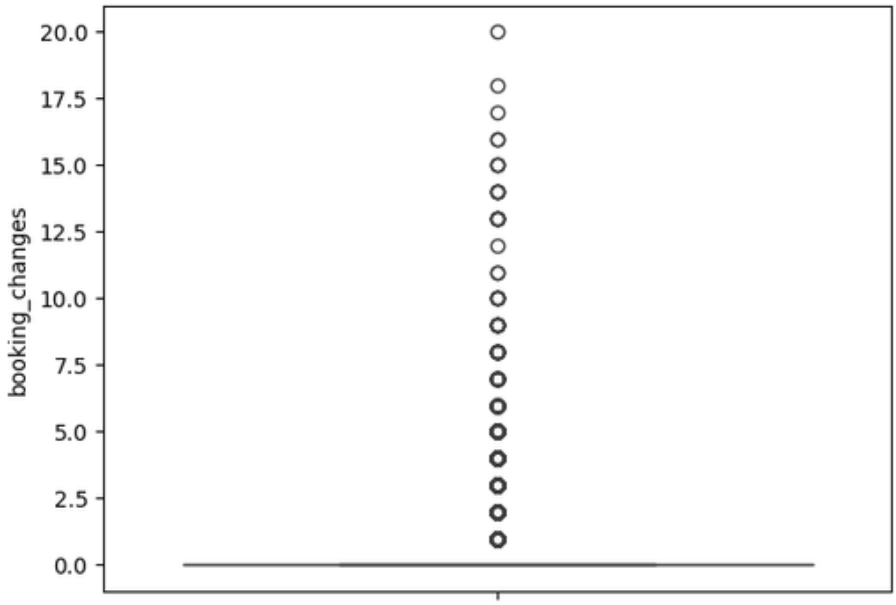

Figure 7: Box plot with outliers of the "booking_changes" variable

I dropped a total amount of 143 outliers (rows with value >5) in the "booking_changes" feature because the number of occurencies was so little it did not influence the overall result.

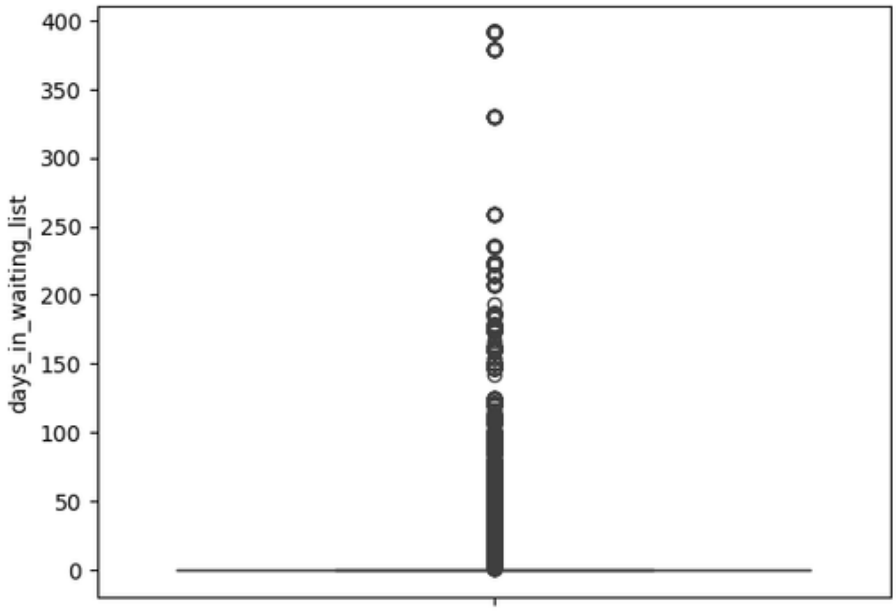

Figure 8: Box plot with outliers of the “days in waiting list” variable

Considering the “days in waiting list” feature, 85 outliers were dropped (rows with values >250) since they were considered as extreme outliers (a celar gap can be detected in the graph).

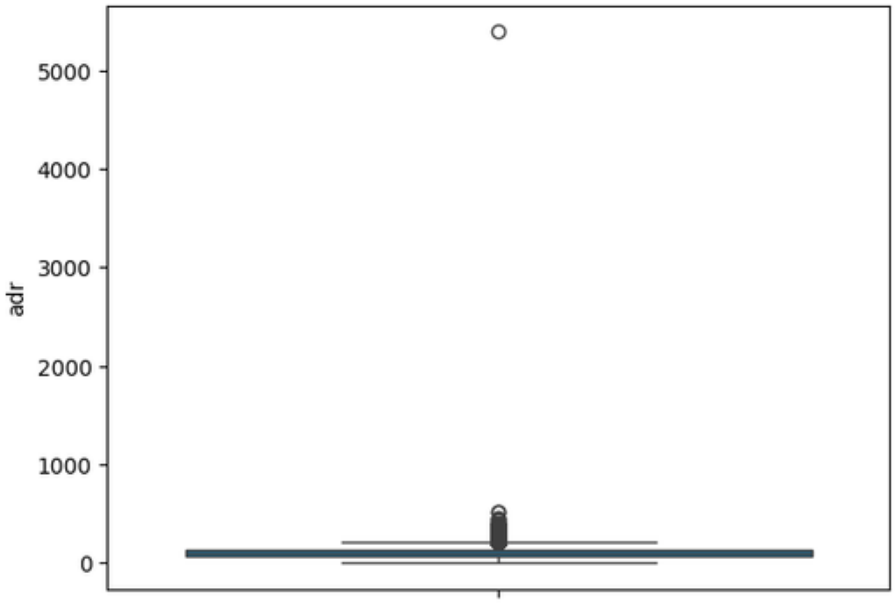

Figure 9: Box plot with outliers of the "adr" variable

Regarding the “adr” feature, I only dropped 1 outlier (row with value >1000) because it was too extreme considering the others values’ distribution.

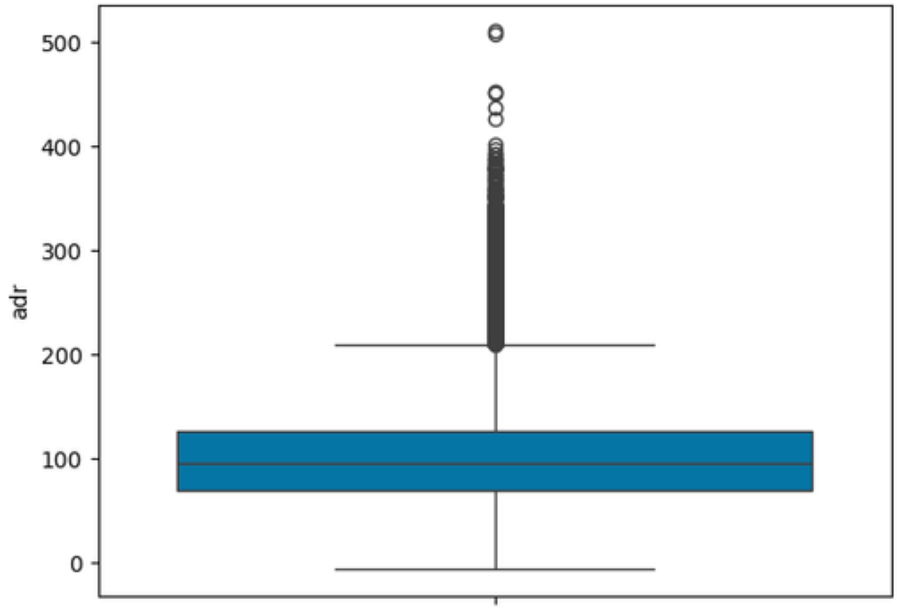

Figure 10: Resulting box plot of the “days in waiting list” variable after removing outlier

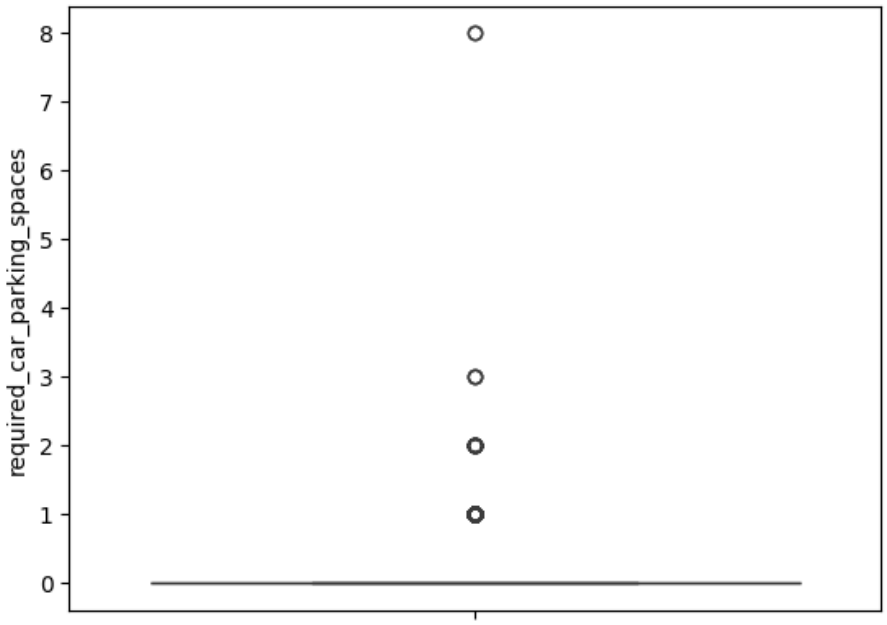

Figure 11: Box plot with outliers of the "required_car_parking_spaces" variable

Regarding the “required_car_parking_space” feature, I dropped a total amount of 2 outliers (rows with value >7) as they can be clearly seen as extrems just by looking at the graph.

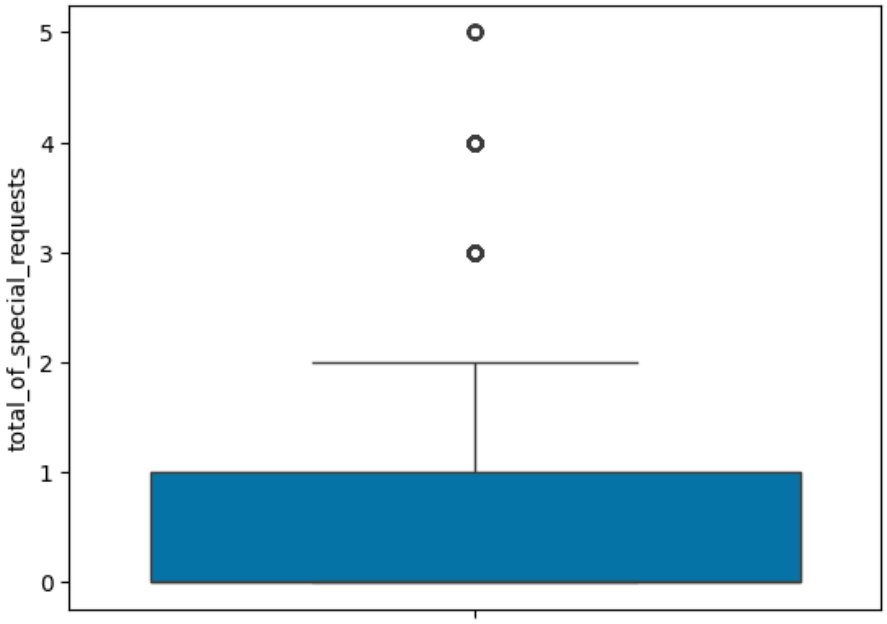

Figure 12: Box plot with outliers of the “total of special requests” variable

Regarding the “total of special requests” feature I dropped 375 outliers (rows with values >3) because they would affect the analysis too much resulting in less accuracy.

4.3 Answering descriptive analysis questions

After cleaning the datset from outliers, I considered only reservations, cancelations and not cancelled bookings. First of all, I inspected the distribution of canceled versus non-cancelled bookings.

4.3.1 Descriptive analysis / question 1

What is the distribution of canceled versus non-cancelled bookings?

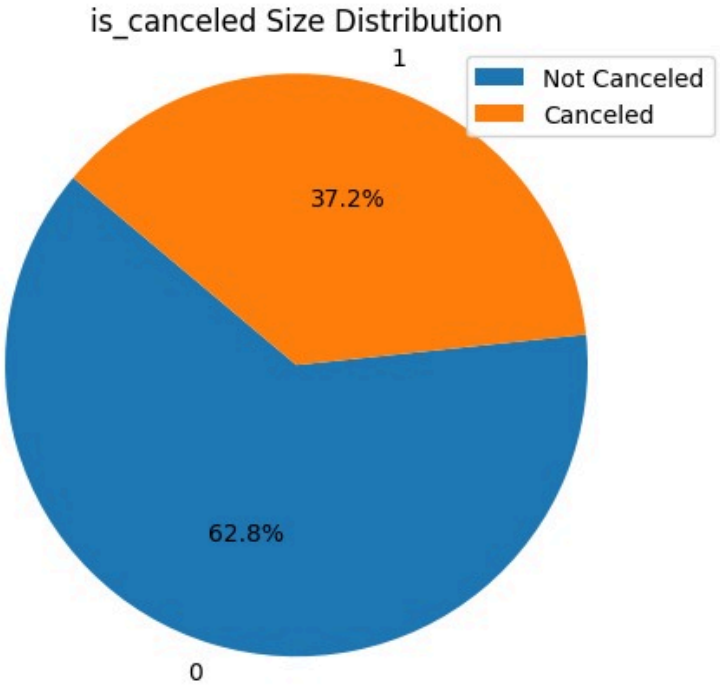

Figure 13: is_canceled size distribution (0=not canceled, 1=is canceled)

In terms of numbers, 37.12% of the data correspond canceled bookings, while 62.8% are those which did not get canceled. As we can see, the percentage of canceled bookings roughly accounts for 2/5 of total bookings.

Next on, I inspected which factors are most associated with cancelled bookings.

4.3.2 Descriptive analysis / question 2

What are the characteristics of bookings that are most frequently canceled (e.g., booking duration, time of year, customer type)?

To answer this question, I grouped both canceled and not_canceled bookings and analysed how these features would differ when related to other variables.

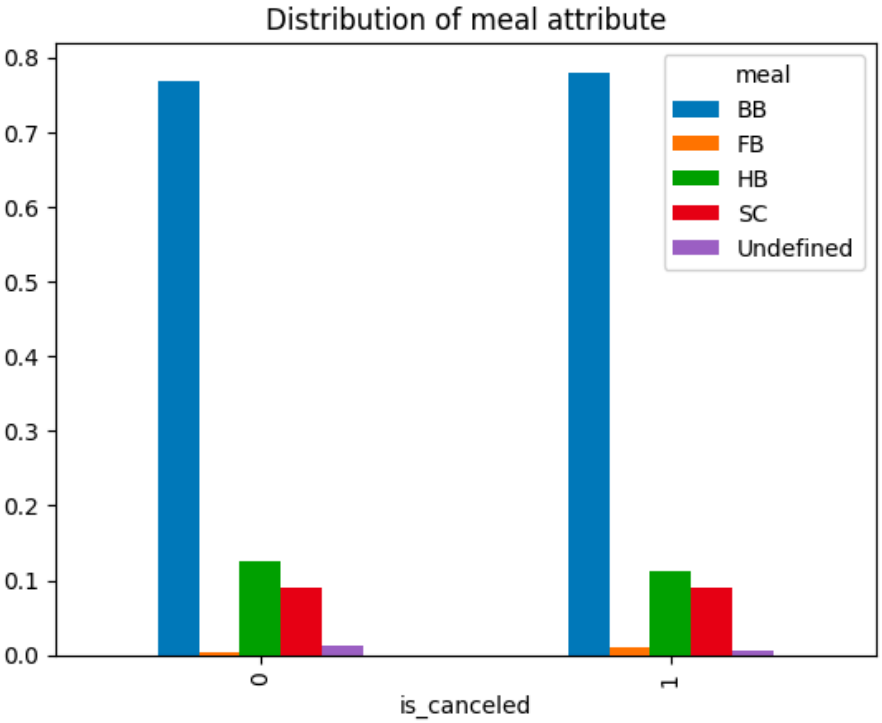

Figure 14: distribution of meal attribute (value in %)

More than 75% of BB meals are present in both canceled and not canceled booking, without any great change among the data. The same happens for HB meals (less than 1.5%) and SC ones (less than 1 %).

Only FB anf Undefined meals do change across cancelled and not cancelled bookings. FB meals are more associated with cancelled bookings, while the opposite happens for Undefined meals. Yet, discrepancies for these values are very low.

Bookings without meals included (SC - Self Catering) have a lower cancellation rate, which might suggest that guests who plan to dine outside the hotel are more certain about their travel plans or that these bookings are often made by guests with other commitments (business, family nearby) that make cancellations less likely

If we consider the country attribute, it is evident how the vast majority of bookings in the database belogns to portoguese people. While we can see how more than 28% of not cancelled bookings is related to portoguese costumers, they account for more than 60% of cancelled bookings. If we consider the other nationalities, we can easily understand how the trend shifts: the % of not cancelled bookings is always higher than the one addressing bookings which got cancelled.

Since guests from Portugal are more likely to cancel, a reason could be that bookings are most likely to be influenced by factors like distance, ease of travel or cultural tendencies.

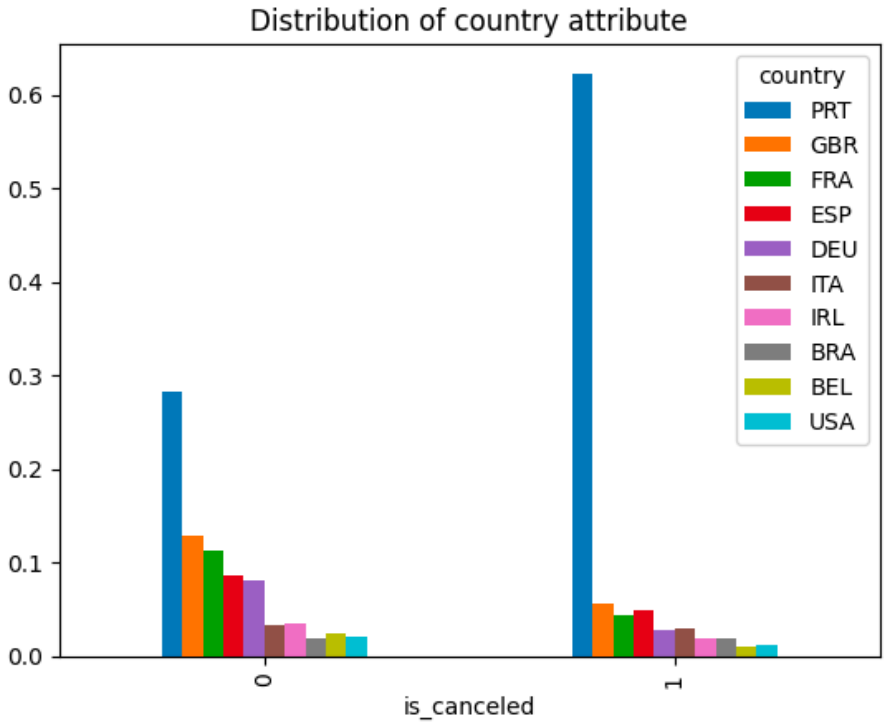

Figure 15: distribution of country attribute (value in %)

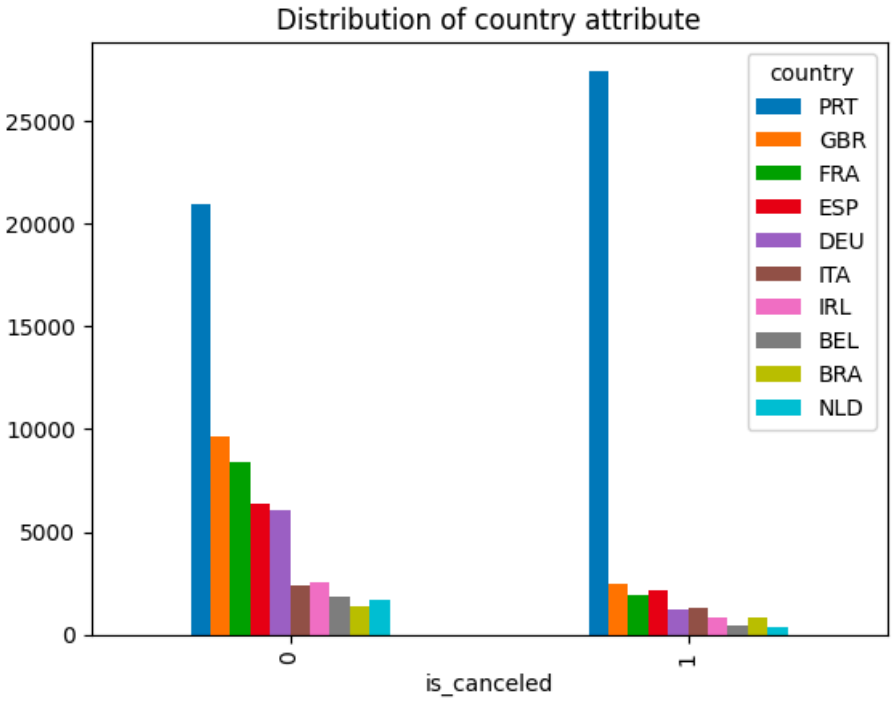

Figure 16: distribution of country attribute (total count)

When it comes to analysing the distribution of market segment attribute, groups booking are more likely to get canceled, accounting for almost 30 % of the total cancelations. The % value does not change when it comes to Online TA bookings,

while if we consider direct bookings it is clear how costumer who book autonomously are less likely to cancel (only 0.5%).

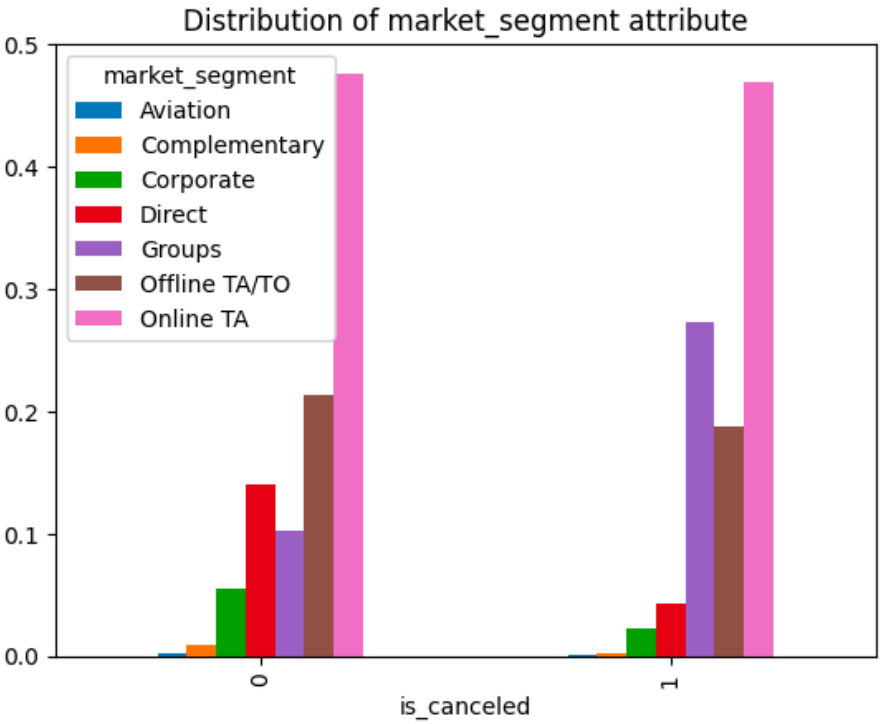

Figure 17: distribution of market segment attribute (value in %)

The online travel agent segment has a high cancellation rate, which might reflect the volatile nature of online bookings where comparisons and changes can be made quickly and impulsively.

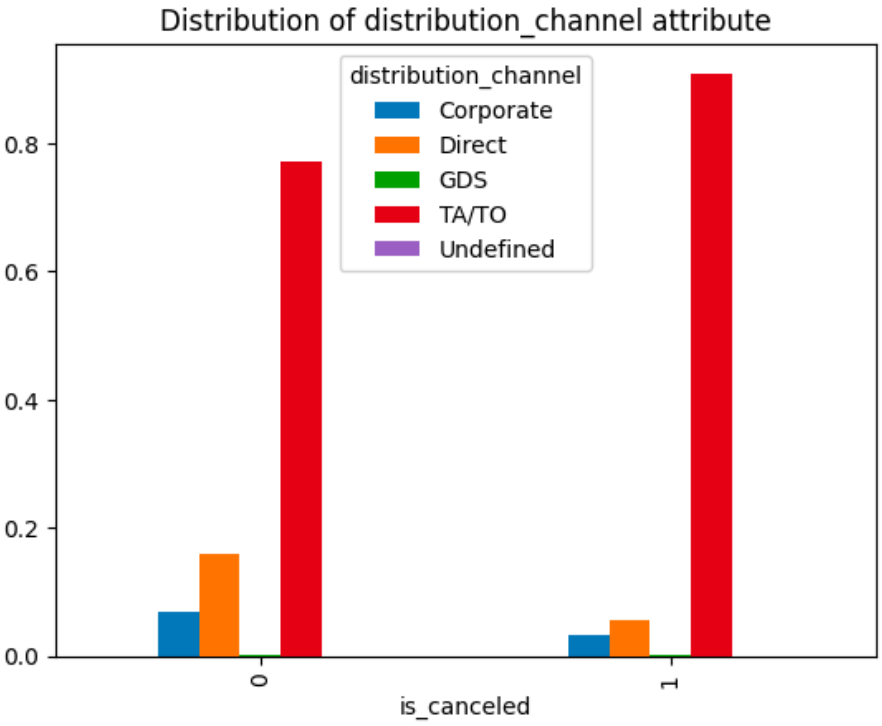

Figure 18: distribution of distribution channel attribute (value in %)

Regarding the distribution channel attribute, the graph shows no particular insight and the outcome depicted is as expected.

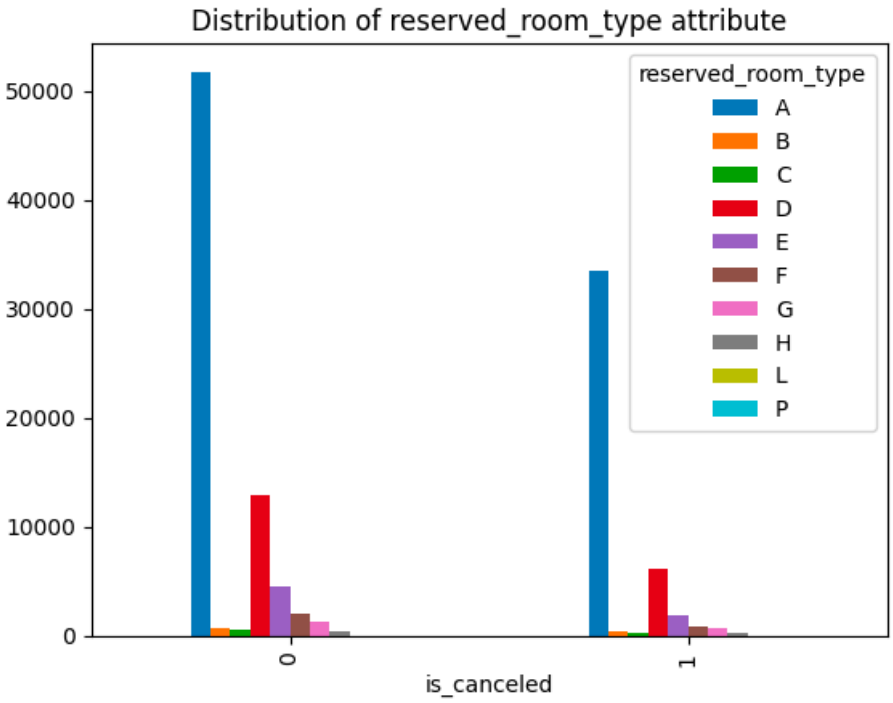

Figure 19: distribution of reserved room type attribute (total count)

The same can be said for the graph showing costumer type and assigned room type attributes.

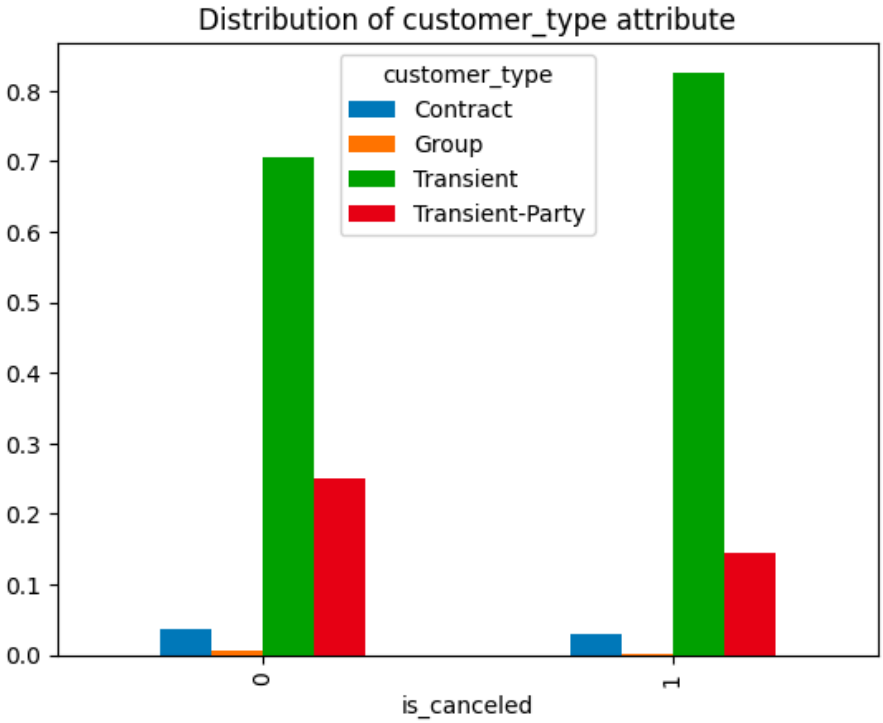

Figure 20: distribution of costumer type attribute (value in %)

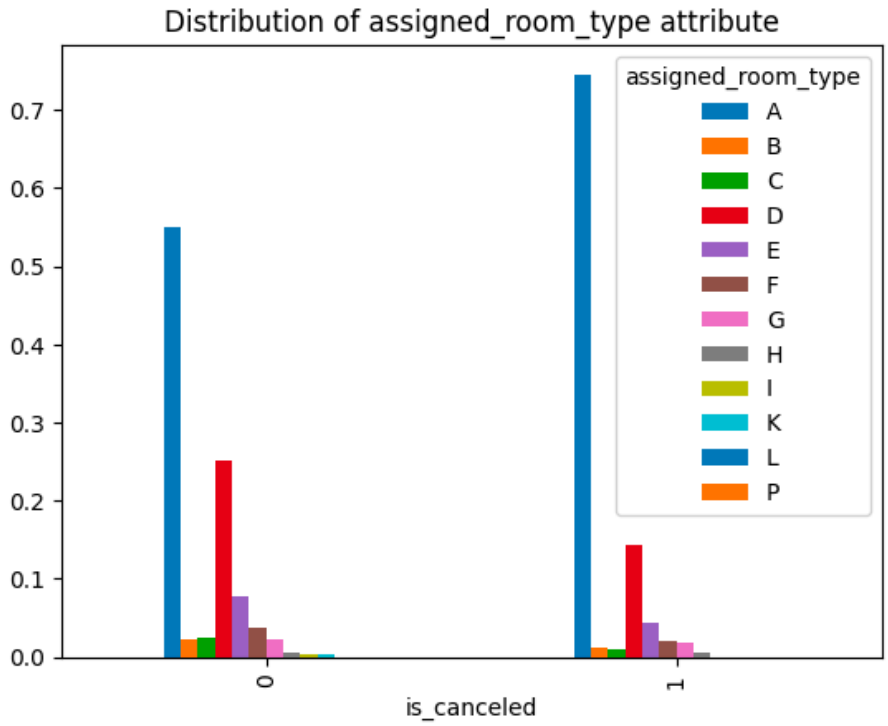

Figure 21: distribution of assigned room attribute (value in %)

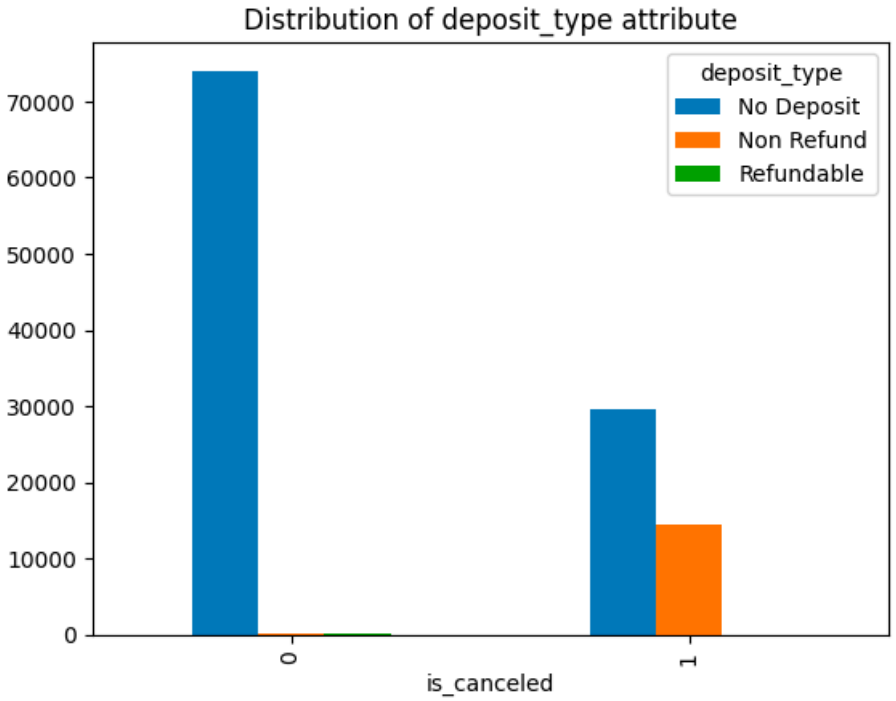

Figure 22: distribution of deposit type attribute (total count)

It is interesting how non refund bookings are only associated with canceled bookings. Probably not enough observation for these types of deposit attribute were taken, and data can be incomplete.

4.3.3 Descriptive analysis / question 3

How do cancellation rates differ across different hotel types (resorts vs. city hotels)?

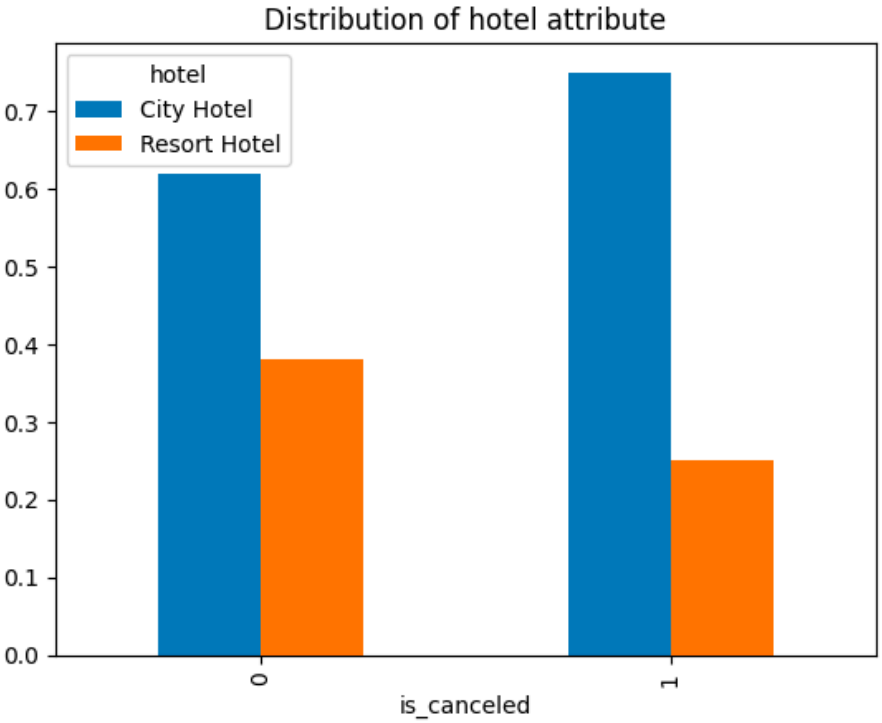

Figure 23: distribution of hotel attribute (total count)

City Hotels - although accounting for most of the bookings overall - account for more than 70% cancelled bookings and 60 % of not cancelled bookings. This probably indicates that City Hotels experience a higher traffic of bookings overall, and might be due to a variety of factors such as the nature of travel (business vs. leisure), booking policies, location dynamics, or clientele differences.

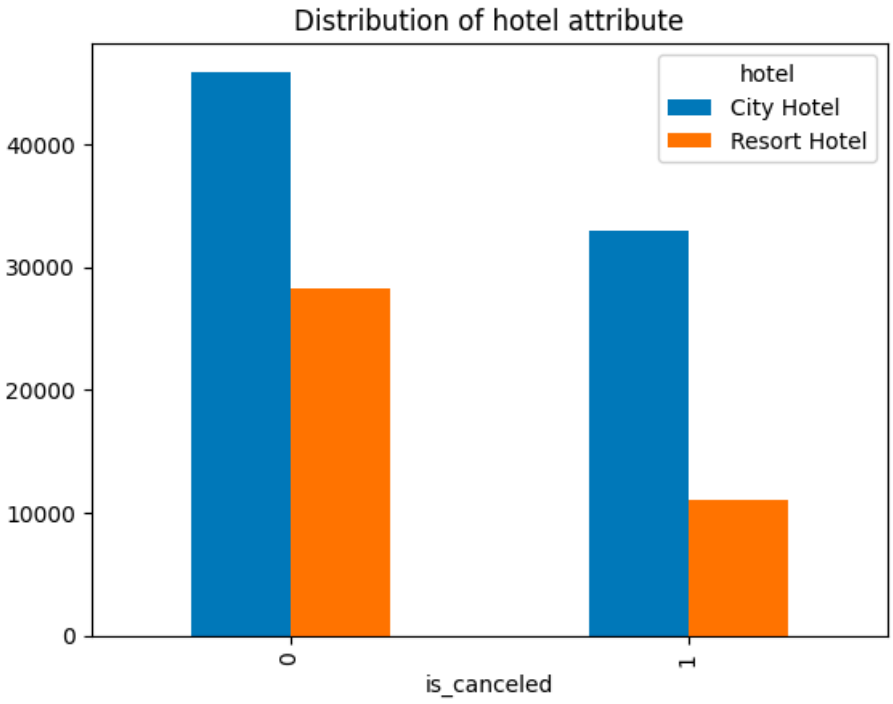

Figure 24: distribution of hotel attribute (value in %)

4.3.4 Descriptive analysis / question 4

What trends are observable in booking cancellations over the years or seasons?

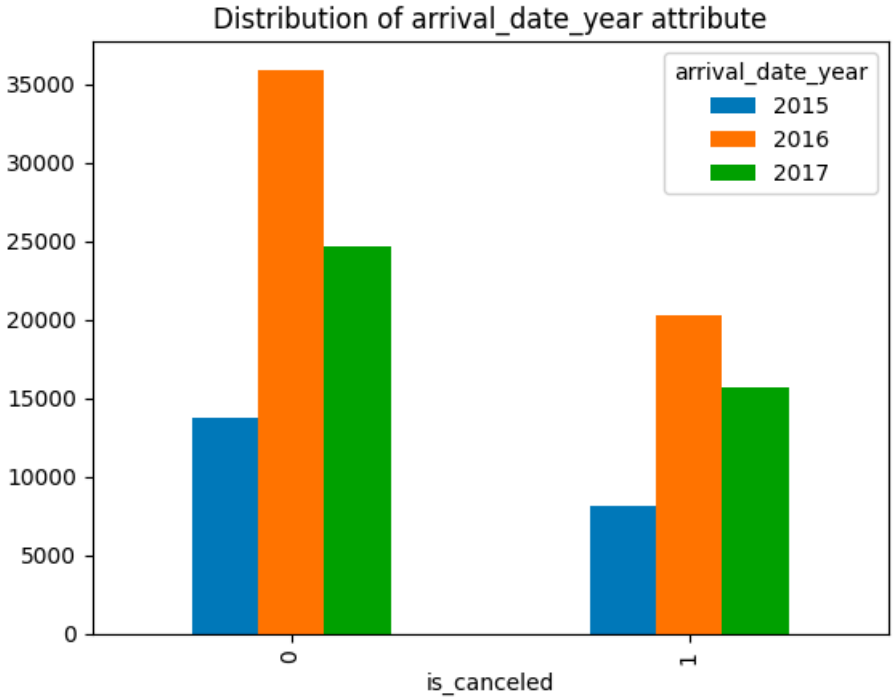

Figure 25: distribution of arrival_date_year attribute (total count)

We can clearly see a trend that reflects a rise and subsequent fall in both non-cancellations and cancellations over the years. In 2016, there's a notable peak in the overall number of bookings made, with the highest volume of both successful stays and cancellations. This suggests that while the hotel industry experienced a boost in business, it also faced a proportionate increase in the number of guests who opted not to fulfill their bookings. The decline in both non-cancellations and cancellations in 2017 could be attributed to a variety of factors, including potential market saturation, enhanced booking policies, or even broader economic trends impacting travel and hospitality. Interestingly, despite the decrease in raw numbers from 2016 to 2017, the proportion of cancellations remains higher in 2017 compared to 2015, indicating that the issue of booking cancellations is becoming more pronounced relative to the total bookings. This trend could signal a need for hotels to examine the factors contributing to the elevated cancellation rates. It could be postulated that the ease of online booking and the flexibility of cancellation policies might contribute to this trend, encouraging customers to book without certainty in their plans. Seasonal factors may also play a role, with certain times of the year possibly being more prone to cancellations due to the variability in weather conditions, holiday periods, or special events which can disrupt travel plans.

4.4 Answering diagnostic analysis questions

After performing a descriptive analysis, I went on to answer the questions from the diagnostic analysis.

4.4.1 Diagnostic analysis / question 1

Is there a correlation between the lead time of a booking and its probability of being canceled?

I wanted to see which is the lead time variable distribution and how data changes across cancelled and not cancelled bookings.

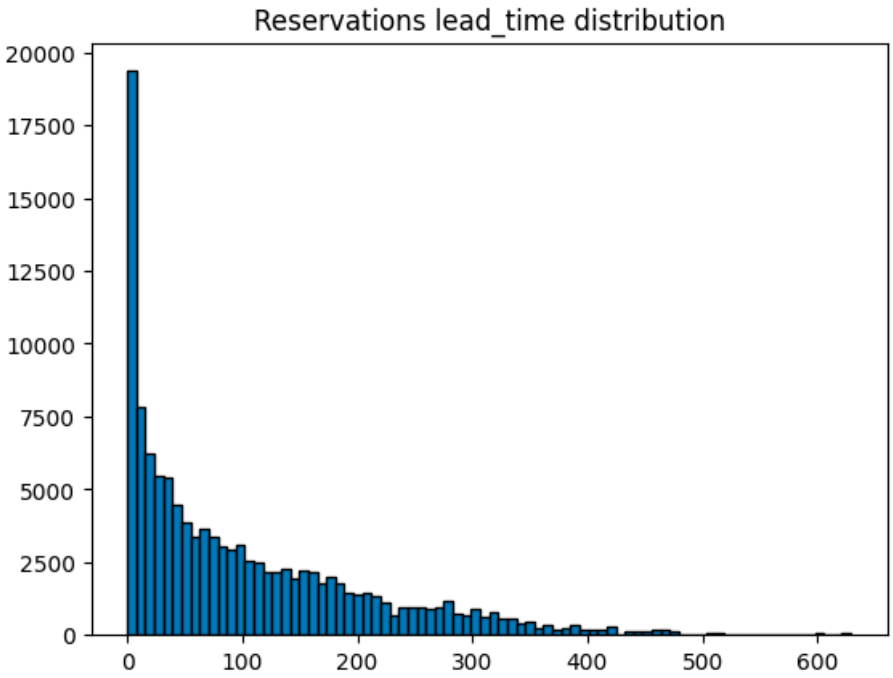

Figure 26: reservations lead_time distribution

The graph shows the distirbution across all reservations made.

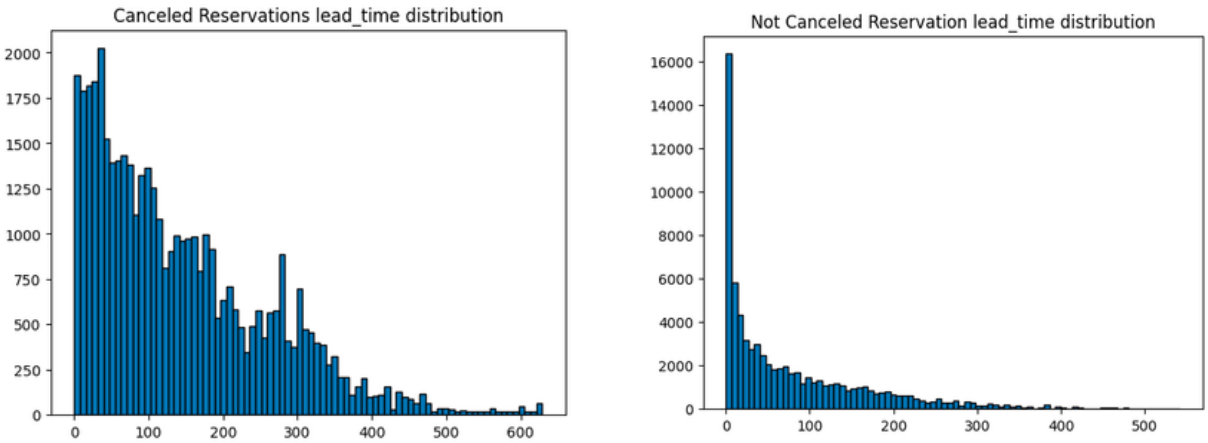

Figure 27: canceled reservations lead_time distribution (left) vs not canceled reservations lead_time_distribution (right)

If we analyse the graphs, we can clearly notice how the vast majority of bookings with a low lead time value are not cancelled. The third graph displays 16.000 bookings (not canceled) with lead_time = 0 while the second one shows how, for the same lead_time value, canceled reservations are less than 2.000. This means aproximately 1 booking out of 9 (with a low lead_time value) is likely to be cancelled.

If we proceed to consider increased lead_time values, the data does not show particular insights but is dispalyed accordingly to what was expected: the likelihood of cancellations increases (for lead_time tending to 0, the probability of cancellation tends to 0).

4.4.2 Diagnostic analysis / question 2

How do external factors such as holidays impact cancellation rates?

I started the analysis by considered the following holidays:

- 1 January
- 6 January
- 27 March
- 5 April
- 16 April
- 1 May
- 15 August
- 31 October
- 1 November
- 24 December
- 25 December
- 31 December

Then, I computed the number of cancelled reservations divided by the total number of reservations for each day.

DAY / MONTH	CANCELLATIONS (%)
1 January	0.35
1 May	0.41
1 November	0.52

DAY / MONTH	CANCELLATIONS (%)
15 August	0.41
16 April	0.38
24 December	0.27
25 December	0.32
27 March	0.33
31 December	0.33
31 October	0.37
5 April	0.33
6 January	0.27

Table 3: % of booking cancellations related to holidays

The table shows how, throughout the years considered in the database, the holiday occurences which have the highest oercentage of cancelled bookings are the 1st of November, (0,52 %) followed by the 1st of May (0,41%) and the 15th of August (0,41%).

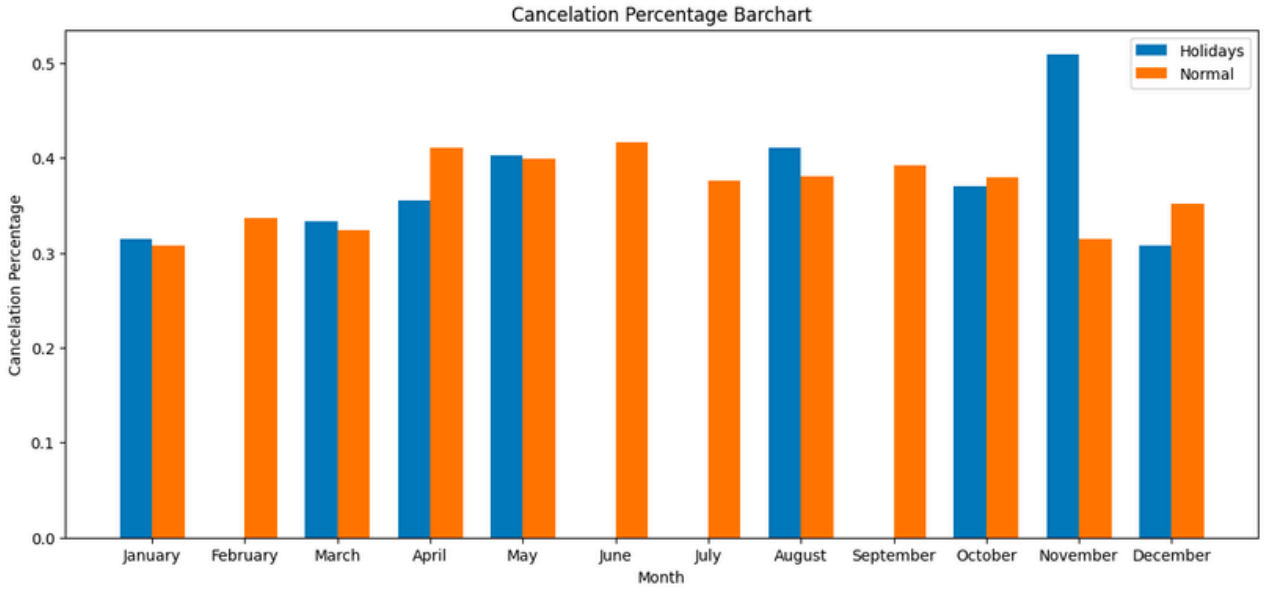

Figure 28: cancelation percentage barchart

The increase in cancellation rates around traditional holiday periods, particularly at the end of the year in December, might be influenced by the uncertainty associated with weather conditions or the increased likelihood of guests making last-minute changes to their travel plans to be with family.

The graph also depicts how the percentage of cancelled bookings during holidays is higher than the percentage of cancelled bookings during working days in January and March. This trend may suggest that guests are more likely to change their plans around holidays, potentially due to the flexibility of holiday schedules or the competing demands of family obligations and other holiday events.

4.4.3 Diagnostic analysis / question 3

How do different variables influence cancellation probabilities?

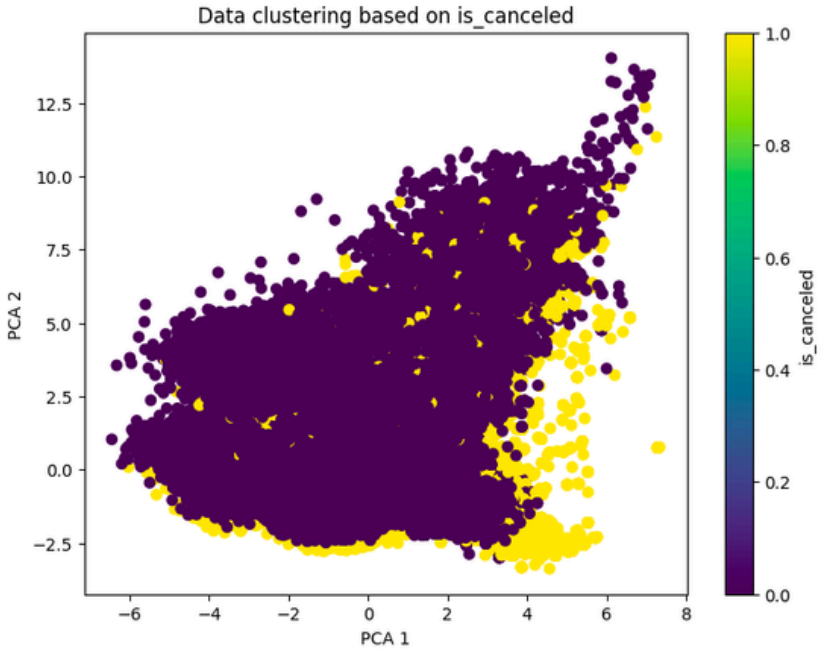

Figure 29: Data clustering based on is_canceled attribute

I did a Principal Component Analysis (PCA) on my dataframe in order to make a more understandable visualization. I reduced components to two, so that I can visualize the data easily with only two axis and then I applied a KMeans clustering to look for clusters in my data.

As visible in the figure, the 'is_canceled' values do not create distinct clusters but only a huge cloud in the centre. Anyway appears clear that leftier the points are, the more probable it does not belong to a 1 class in 'is_canceled' attribute and vice versa.

The majority of purple points, compared to the yellow ones are, of course, given by the prevalence of not canceled rows in the data over the canceled ones.

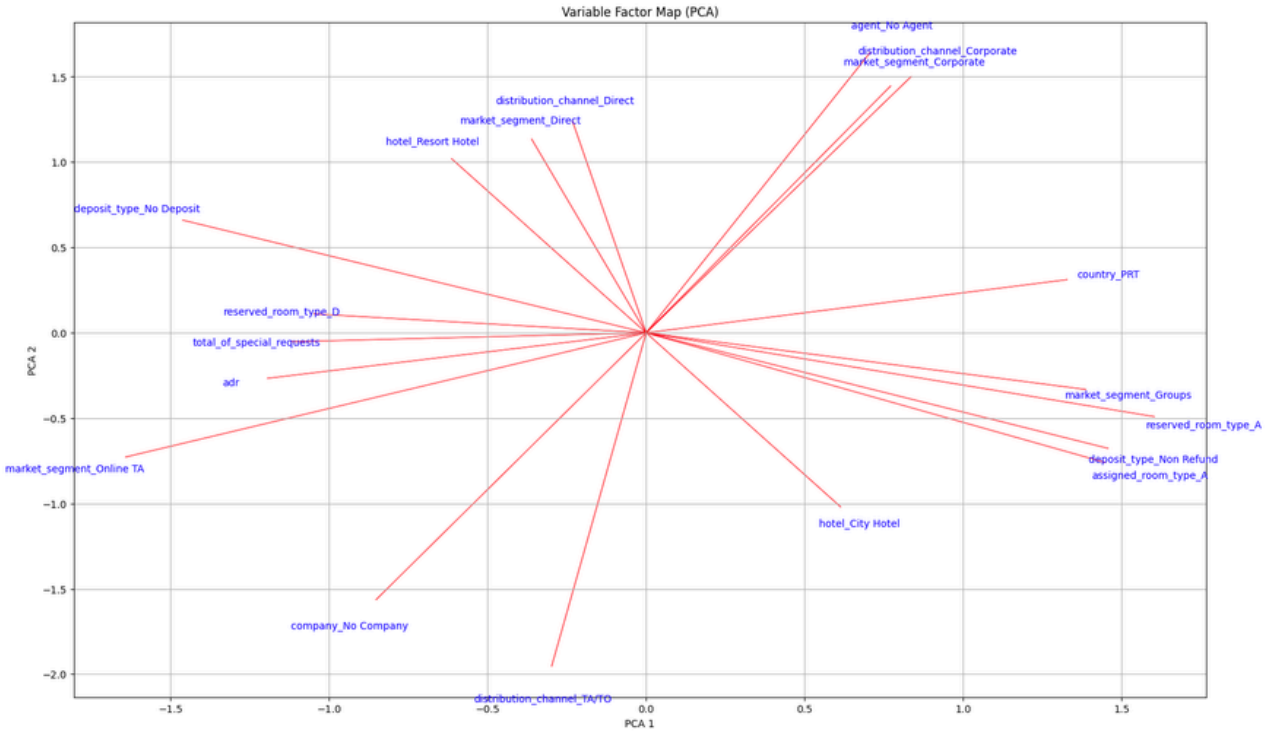

Figure 30: Variable factor map (PCA)

The following variable factor map shows the most important factors (as we selected them using a threshold, to consent a proper visualization) to divide the two classes and we can see, as example, that reservations coming from Portugal are shifted more to the right and, as we can see, are more likely to belong to cancelled ones, perfectly in line with what previous analyses showed.

4.4.4 Diagnostic analysis / question 4

Are there variables which are correlated?

Using a correlation matrix, I looked for correlations between couples of numerical variables, in order to see if variance (and so predictive power) explained by variables is quite the same. Inspecting the correlation matrix, I noticed that, except the obvious 1s in the main diagonal, there isn't any extreme value (0.48 being the highest and -0.17 as the lowest). This result shows us that variables aren't strongly correlated one with each other.

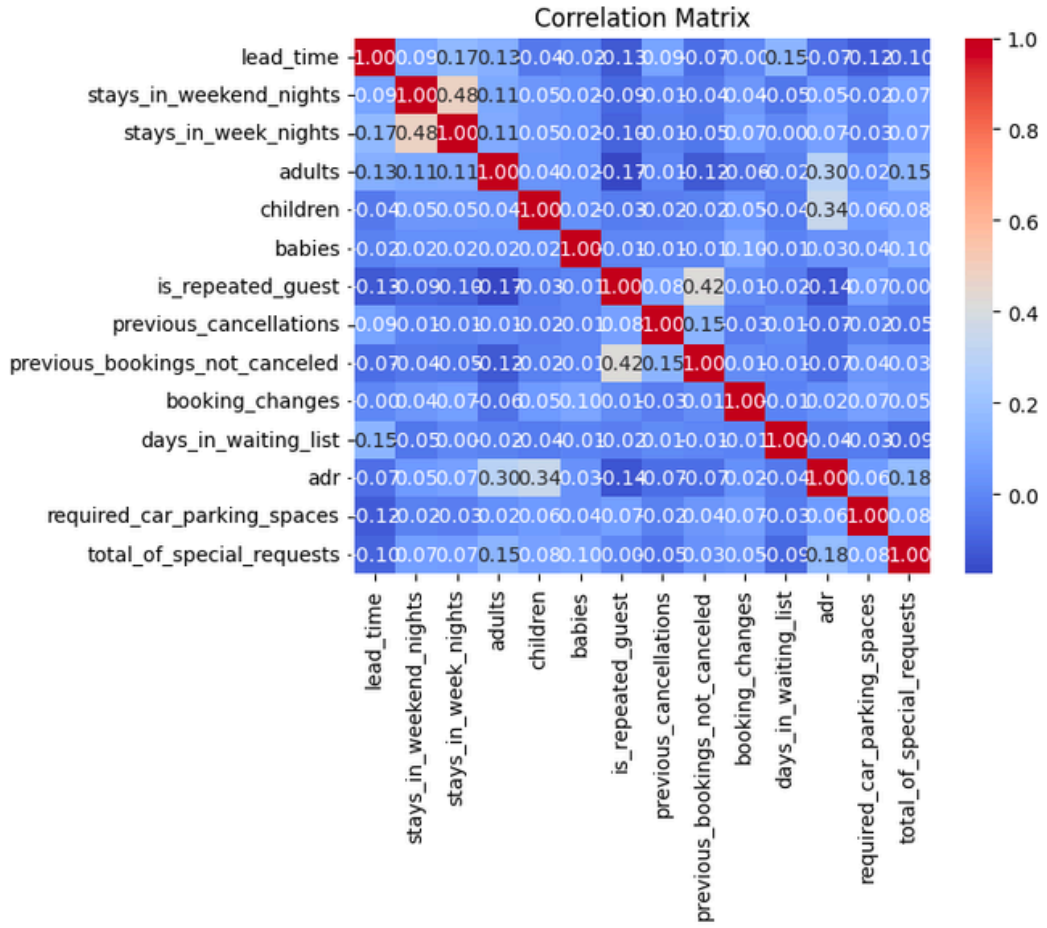

Figure 31: Correlation matrix

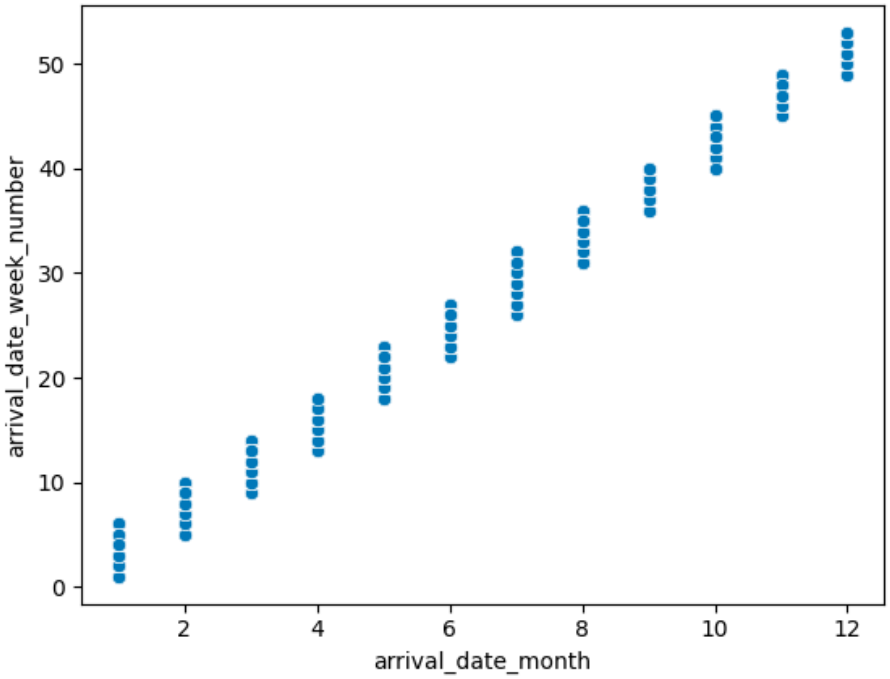

Figure 32: Scatterplot graph

Among categorical variables, there are ordinal variables like arrival_date_month or arrival_date_week_number. As imaginable, those two variables represent quite the same information, so, after converting month categories to their number, I inspected the correlation between them with a scatter plot. As visible in the graph, the points leads to an almost perfect identity function, meaning that they provide almost the same predictive power and so I can proceed with the deletion of 'arrival_date_week_number'

5. Modelling: predictive analysis

For the predictive analysis, I tried to find the best model available among the ones I could use. As I need to predict the probability of belonging to one category or another one (which are 1 and 0, meaning canceled and not canceled), I could have used Logistic Regression, Decision Tree or Neural Network.

I then decided to investigate all of them and take the best isnights from each model.

5.1 OneHot Encoding

As whichever model can deal only with boolean and numeric values, I need to encode categorical values in a way that the machine can understand. In order to do that I use OneHotEncoding, so, for every categorical variable, every category becomes a column in the dataframe, having True (or 1) only in the rows belonging to that category and False in each other.

5.1.1 Predictive analysis / question 1

Which models can we build to best predict cancellation?

5.1.1.1 Logistic regression

At first I tried with the Logistic Regression model, as it is the easiest one among the three mentioned above.

I started by doing the train-test split with a 80/20 split ratio as it looks reasonable for a dataset like ours. Then I created a pipeline with, at first, a MinMaxScaling for the numerical variables, in order to compress the values to make them usable by the model. Then I create the Logistic Regression model and fit it to the dataset. The results are promising as we got an overall 82.96% of accuracy. Recall, precision and, as a consequence, F1-Score are a bit lower, due to the slightly unbalanced dataset.

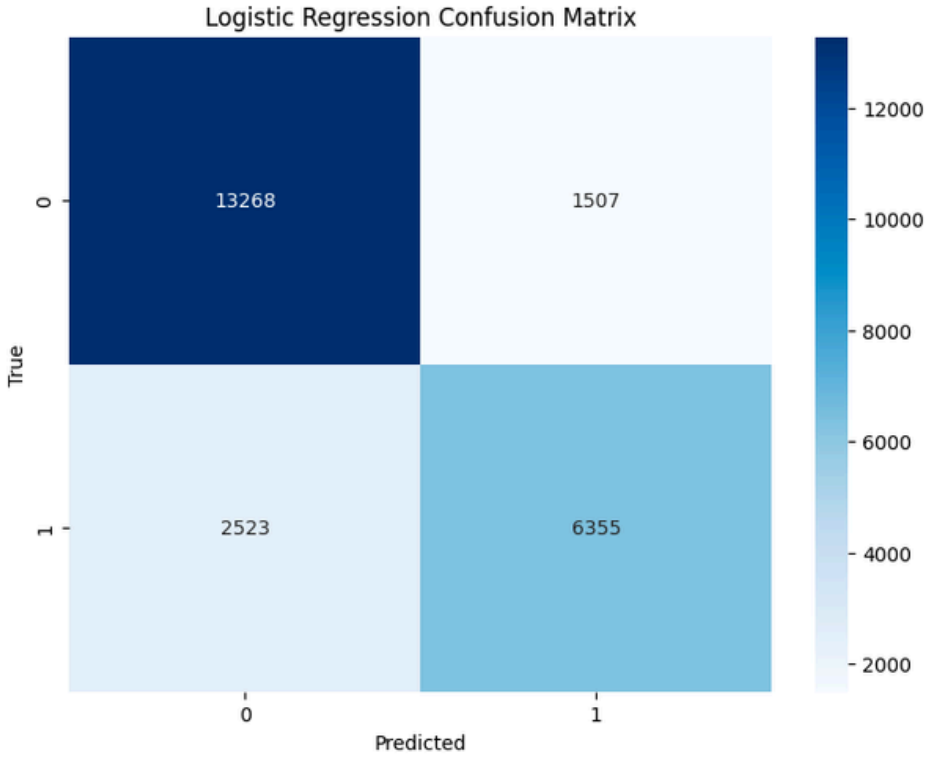

Figure 33: Logistic regression confusion matrix diagram

True Negatives (TN): 13.268

The model correctly predicted 'not canceled' (0) bookings 13,268 times. This is the largest number in the matrix, which indicates strong performance for this category.

False Positives (FP): 1.507

These are the cases where the model incorrectly predicted cancellations (predicted 'canceled' when the booking was actually not canceled). The impact here is that the hotel might overestimate the number of cancellations and could potentially overbook leading to an unexpected surplus of guests.

False Negatives (FN): 2.523

This number represents the bookings that were actually canceled (1), but the model predicted them as not canceled (0). This is critical because it means the hotel would expect these guests, potentially turning away other business only to have these guests cancel.

True Positives (TP): 6.355

The model successfully identified 6,355 cancellations. This insight is vital as it suggests that the hotel can rely on the model to identify a substantial number of potential cancellations, allowing them to take preemptive actions to mitigate the impact on occupancy and revenue

5.1.1.2 Decision Tree

Secondly, I used the Decision Tree model, as alternative to the previous one. I did the same process of train-test split, with the same ratio, MinMaxScaling of variables and then I fit the model.

The results are way better than the previous model, with a 86.54% of accuracy. I select this model also to visualize and understand the tree structure of decisioning the model is making but, due to the OneHotEncoding, the attributes of the dataset explodes and any kind of visualization is impossible, so we have to stick with Logistic Regression significativity variable explanation.

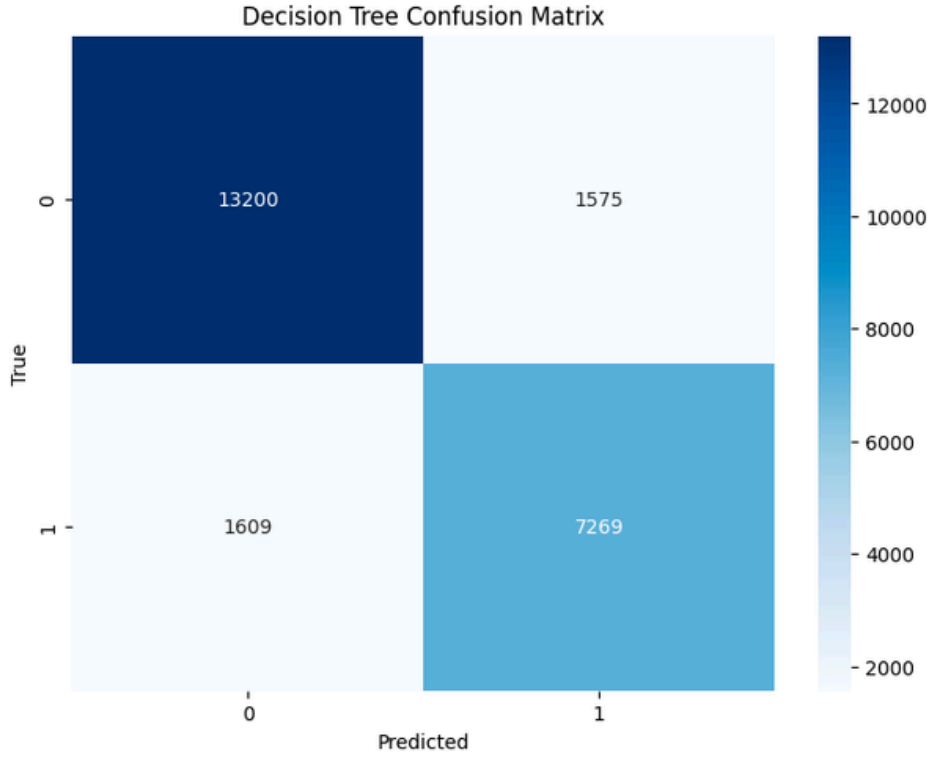

Figure 34: Decision Tree confusion matrix diagram

True Negatives (TN): 13.200

The model correctly predicted the 'not canceled' bookings 13,200 times, indicating its strength in identifying bookings that will likely proceed without a cancellation.

False Positives (FP): 1.575

Here, the model predicted that bookings would be canceled when they actually weren't. While lower than the True Negatives, this number still indicates potential room for improvement, as these mispredictions could lead to unnecessary adjustments in hotel resource allocation.

False Negatives (FN): 1.609

These are bookings that were actually canceled but were predicted as not canceled by the model. This figure is crucial for hotels as it could represent lost revenue or missed opportunities to resell the booking space.

True Positives (TP): 7.269

The model successfully identified 7,269 cancellations. This is a key performance indicator, as it shows the model's utility in predicting which bookings may not result in actual stays.

5.1.1.3 Neural Network

Last, in order to get better performances, I moved to the neural network model, aware that whichever result I will get, it wil be quite a black box model, which is exactly what I got.

After some tries and some different architecture models, I manage to choose the best model, being a dense neural network with 4 layers of 64-32-16-1 neurons, having L2 reguralizers in each layer, BatchNormalization and LeakyReLU as activation function. Then, in order to minimize overfitting, I add one Dropout layer at each hidden layer. I optimize all with a Nadam optimizer, early stopping and model checkpoint to reduce learning rate dynamically. With this architecture I manage to get 87.15% of accuracy with an epoch value of 80, but stopping already at epoch 69 due to early stopping.

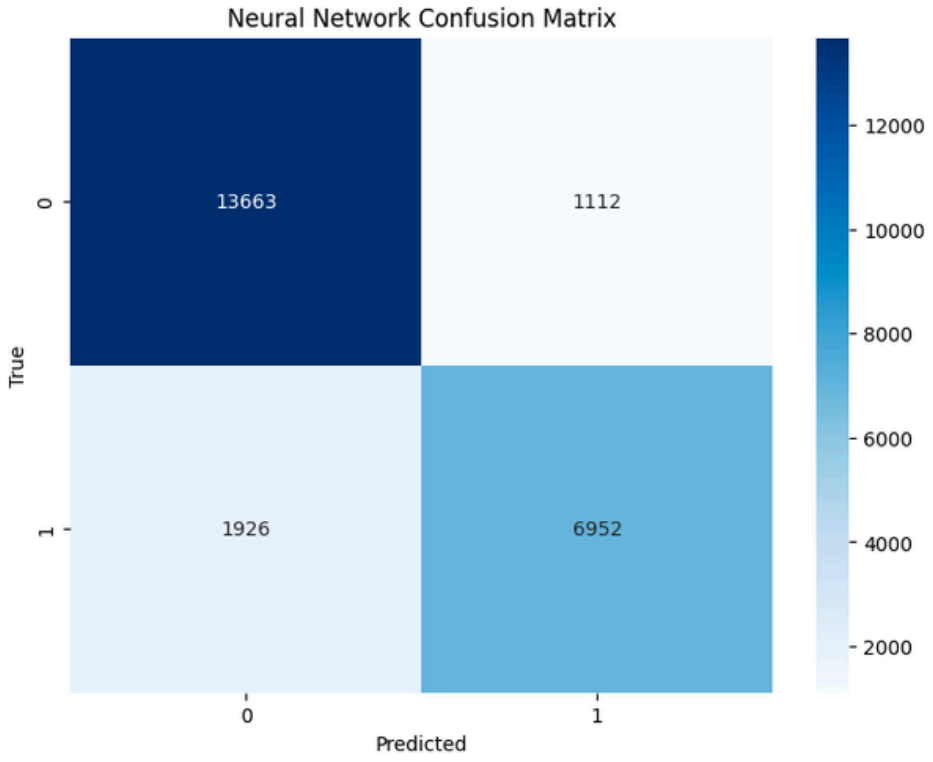

Figure 35: Neural Network confusion matrix diagram

True Negatives (TN): 13.663

The model has correctly predicted a significant number of non-cancellations, suggesting a strong ability to identify true negative cases.

False Positives (FP): 1.112

These instances were predicted as cancellations by the model but did not actually result in cancellation. While lower than the number of True Negatives, it's essential to minimize this to avoid overestimating cancellations.

False Negatives (FN): 1.926

The bookings that were canceled but predicted by the model as non-cancellations. Reducing this number is critical as these cases could lead to unexpected losses in revenue.

True Positives (TP): 6.952

The model has a strong performance in correctly identifying cancellations, which is crucial for effective resource management and planning

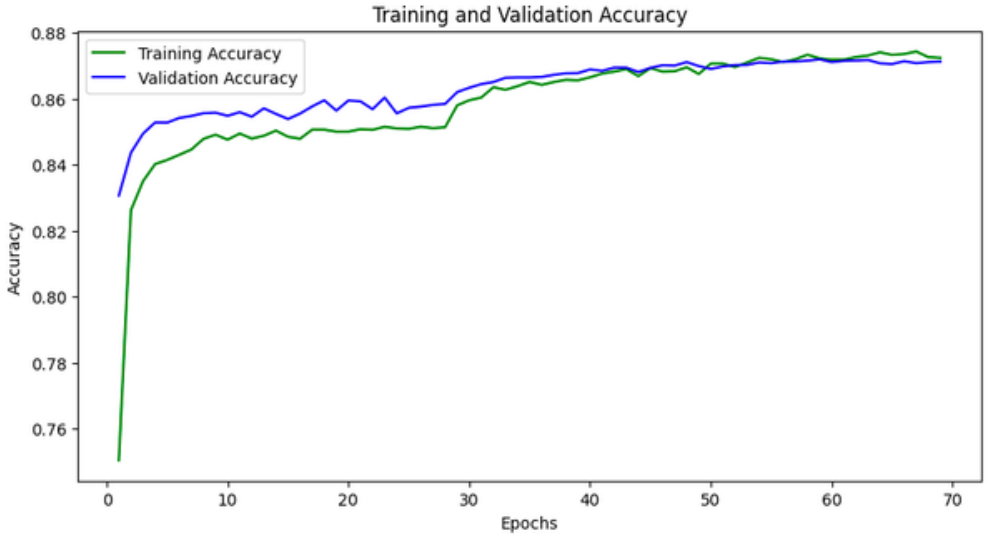

Figure 36: Neural Network training and validation accuracy diagram

The graph illustrates the training and validation accuracy over the epochs for a neural network model.

We can understand that the training accuracy starts high and continues to improve slightly over epochs, indicating the model is learning and able to fit the training data well. The validation accuracy tracks closely with the training accuracy but is consistently slightly lower: this is expected as models tend to perform a bit better on the data they have seen (training data) compared to new data (validation data).

Both accuracies plateau towards the end, suggesting that additional training epochs beyond this point may not result in significant performance gains, which aligns with the early stopping at epoch 69. The close convergence of the two lines indicates good generalization, meaning the model is not overfitting the training data significantly.

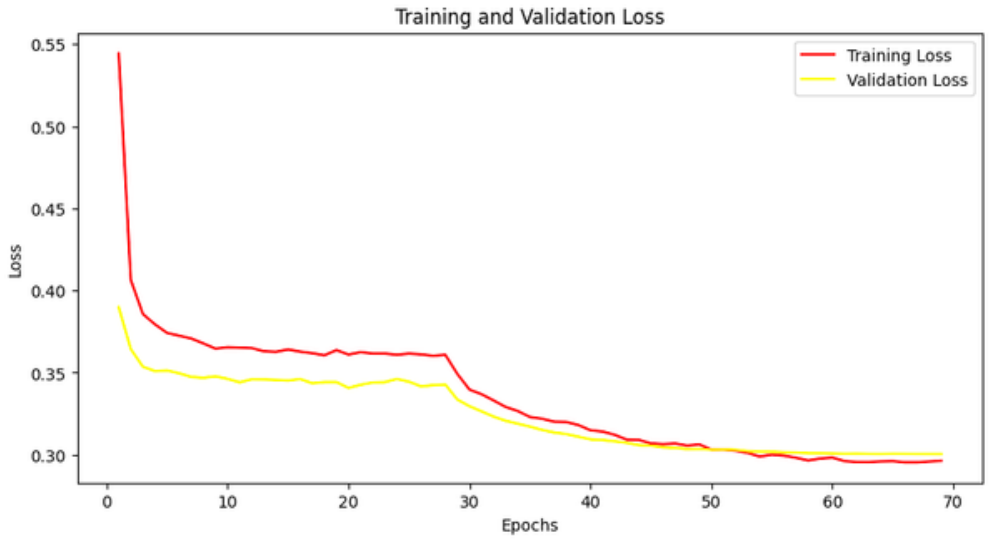

Figure 36: Neural Network training and validation loss diagram

Both training and validation loss decrease sharply at the beginning, which is typical as the model begins to learn from the data. After the initial drop, both losses continue to decrease but at a slower rate, leveling off as the epochs increase. The validation loss slightly increases towards the end, suggesting the model might be beginning to overfit; however, this increase is not significant, indicating the regularizers and dropout layers might be effectively mitigating overfitting. The consistent decline in training loss shows the model is learning effectively, and the leveling off of validation loss suggests it has reached a point of optimization given the current architecture and data.

5.1.2 Predictive analysis / question 2

How accurate are our predictive models?

After running the models and getting the results back, I compared more in depth perfromance parameters.

Below you can see a table listing all values for every model.

Logistic regression

PARAMETER	VALUE (%)
Accuracy	83.06
Precision	80.83
Recall	71.58
F1 - score	75.92

Decision Tree

PARAMETER	VALUE (%)
Accuracy	86.53
Precision	82.19
Recall	81.87
F1 - score	82.03

Neural Network

PARAMETER	VALUE (%)
Accuracy	87.15
Precision	86.21
Recall	78.30
F1 - score	82.06

Table 4,5,6: Predicitve models' parameters and values (%)

If we consider just the Logistic Regression model performance, we can evaluate that:

PARAMETER	VALUE (%)
Accuracy	83.06
Precision	80.83
Recall	71.58
F1 - score	75.92

Table 7: Logistic Regression model parameters and values (%)

Accuracy

Overall, the model's accuracy is approximately 83%, considering all predictions made, both correct and incorrect.

Precision

Precision for the 'canceled' prediction is about 80.8% (TP / (TP + FP)), indicating the proportion of correct 'canceled' predictions out of all 'canceled' predictions made.

Recall

The recall for the 'canceled' bookings is about 71.6% (TP / (TP + FN)). This is an essential metric for a hotel because it measures the model's ability to capture all potential cancellations.

F1-Score

The F1-Score would be calculated from the precision and recall, providing a single measure of the model's predictive power for cancellations, especially when the data is slightly unbalanced as mentioned

PARAMETER	VALUE (%)
Accuracy	86.53
Precision	82.19
Recall	81.87
F1 - score	82.03

Table 8: Decision Tree model parameters and values (%)

Accuracy

This high accuracy indicates that the Decision Tree model is correctly identifying a significant majority of both cancellations and non-cancellations.

Precision

A value of over 82% is good, meaning that when the Decision Tree model predicts a booking will be canceled, it is correct about 82% of the time.

Recall

A recall rate of nearly 82% indicates the Decision Tree is quite effective at catching cancellations. For the hotel industry, a high recall can be more important than precision, as failing to identify a cancellation could be costlier than wrongly identifying one.

F1-Score

The F1-score is the harmonic mean of precision and recall, and at over 82%, it indicates a strong balance between the two. It suggests that the Decision Tree does not overly favor one class over the other and is consistent in its predictions.

Regarding the Neural Network model performance, we can say that:

Test Accuracy

The Neural Network model shows a slightly higher accuracy than the Decision Tree. This means the model has a better overall rate of correct predictions. Given the complexity of Neural Networks, this higher accuracy could come from the model's

PARAMETER	VALUE (%)
Accuracy	87.15
Precision	86.21
Recall	78.30
F1 - score	82.06

Table 9: Neural Network model parameters and values (%)

ability to capture complex patterns in the data.

Precision

This model has a higher precision than the Decision Tree. It indicates a strong relevance of the predictions — a smaller proportion of the bookings predicted to be canceled were actually not canceled. This can be particularly useful when precise targeting of promotions or policies aimed at reducing cancellations is required.

Recall

The recall is slightly lower for the Neural Network model, meaning it misses more actual cancellations than the Decision Tree. This could be an area of concern because it means the model fails to flag all potential cancellations, leading to potential lost revenue.

F1-Score

Despite a lower recall, the F1-score of the Neural Network is very similar to that of the Decision Tree, indicating that the increase in precision compensates for the drop in recall. The F1-score being high means the model maintains a balanced performance between precision and recall.

Overall, we can say that the Logistic Regression model shows the lowest accuracy among the three models at approximately 83%.

The Decision Tree model shows a marked improvement over Logistic Regression in all metrics, with an accuracy of 86.54%, precision of 82.19%, and recall nearly matching precision. This suggests a good balance between identifying true cancellations and not falsely labeling non-cancellations as cancellations. Its F1-score is also solid, indicating a strong balance between precision and recall.

The Neural Network, though, is the best perfromer in terms of accuracy (87.16%) and

precision (86.21%). Its recall, while lower than the Decision Tree, is still higher than Logistic Regression. The higher precision suggests that when it predicts cancellations, it is correct more often than the other models. The F1-score is very close to that of the Decision Tree, which shows that despite a lower recall, the Neural Network maintains a good balance between recall and precision.

The Neural Network, with the highest accuracy and precision, would be the model of choice if the goal is to minimize false positive (meaning the hotel wants to avoid preparing for guests who end up canceling).

The Decision Tree, with the highest recall, is preferable if it's most important to catch as many cancellations as possible, even at the risk of some false positives.

5.1.3 Predictive analysis / question 3

Which variables are the most explicative in our model?

I decided to inspect more the the Linear Regression model, in order to look for the contribution each variable gave to it.

I look for the F-Statistic of the different variables and I manage to get the following results, showing only the 20 most significant variables in the model. As the F-Statistic confirms, lead_time is one of the most significant variables, with a high F-statistic value. So do deposit_type and country but, as I said previously, this can happen also because of the contemporary presence of these two categories, as Portuguese groups tend to make reservations, without refund, but to cancel then.

FEATURE	F - STATISTIC
lead_time	8935.109487
deposit_type_Non Refund	28656.575573
deposit_type_No Deposit	28094.189768
country_PRT	12023.316388

FEATURE	F - STATISTIC
assigned_room_type_D	1530.890236
assigned_room_type_A	3776.324812
distribution_channel_TA/TO	2961.347771
distribution_channel_Direct	2151.956034
market_segment_Groups	4827.071465
booking_changes	1963.821820
market_segment_Direct	2244.406246
customer_type_Transient	1703.541970
hotel_Resort Hotel	1758.282386
required_car_parking_spaces	3777.073557
total_of_special_requests	5530.010865
hotel_City Hotel	1758.282386
customer_type_Transient-Party	1474.591874
country_GBR	1330.458087
country_FRA	1307.964583
previous_cancellations	1178.641789

Table 10: List of the most 20 significant features along with F-Statistic values

6. Evaluation: prescriptive analysis

Here, the predictive insights are transformed into concrete, actionable strategies that hotels can implement to reduce the incidence of cancellations.

The insights gathered reveal the potential for targeted interventions, such as leveraging lead-time data to identify bookings with a high risk of cancellation and offering flexible booking conditions and rates to accommodate changes in plans.

Deposit policies can be restructured, promoting non-refundable deposits with competitive rates to minimize casual bookings that lead to cancellations. Distinct strategies need to be crafted for different customer segments, providing group-specific packages that encourage completion of bookings and personalized offers to transient guests, enhancing their commitment to their reservations.

Personalized marketing, harnessing the wealth of customer data, will enable the creation of compelling, tailored offers that speak directly to potential guests' preferences, reducing the likelihood of cancellations. Proactive and dynamic communication strategies should be implemented to maintain engagement with guests identified by the model as having a higher cancellation probability, offering them timely reminders, reassurances, or additional incentives to maintain their booking.

Operationally, predictive cancellation insights should inform staffing and inventory management, optimizing resource allocation, and controlling costs. Moreover, waitlists for periods identified as having high cancellation probabilities can quickly fill rooms that become available last minute, ensuring optimal occupancy rates.

The pricing strategy must be dynamic, incorporating predicted cancellation trends and adjusting rates accordingly during high-risk periods to offset potential revenue losses, or offering discounts to attract more assured bookings. Overbooking strategies should be calibrated by integrating predictive insights, balancing the risk of guest cancellations against the risk of service denial, ensuring that the number of overbookings is just right to maximize revenue without impairing guest experience.

It's crucial to emphasize that these prescriptive measures are not static; they require continuous monitoring and refinement in response to their real-world performance, shifts in market trends, and changes in consumer behavior. The effective implementation of a prescriptive analytics approach is a strategic investment in the hotel's resilience, equipping it to not only withstand the unpredictability of booking cancellations but also to navigate towards a horizon of enhanced profitability and heightened customer satisfaction.