Table 1
Variables description.

Variable	Type	Description	Source/Engineering
ADR	Numeric	Average Daily Rate as defined by [5]	BO, BL and TR / Calculated by dividing the sum of all lodging transactions by the total number of staying nights
Adults	Integer	Number of adults	BO and BL
Agent	Categorical	ID of the travel agency that made the booking[a]	BO and BL
ArrivalDateDayOfMonth	Integer	Day of the month of the arrival date	BO and BL
ArrivalDateMonth	Categorical	Month of arrival date with 12 categories: "January" to "December"	BO and BL
ArrivalDateWeekNumber	Integer	Week number of the arrival date	BO and BL
ArrivalDateYear	Integer	Year of arrival date	BO and BL
AssignedRoomType	Categorical	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons	BO and BL
Babies	Integer	Number of babies	BO and BL
BookingChanges	Integer	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation	BO and BL/Calculated by adding the number of unique iterations that change some of the booking attributes, namely: persons, arrival date, nights, reserved room type or meal
Children	Integer	Number of children	BO and BL/Sum of both payable and non-payable children
Company	Categorical	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons	BO and BL.
Country	Categorical	Country of origin. Categories are represented in the ISO 3155–3:2013 format [6]	BO, BL and NT
CustomerType	Categorical	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	BO and BL
DaysInWaitingList	Integer	Number of days the booking was in the waiting list before it was confirmed to the customer	BO/Calculated by subtracting the date the booking was confirmed to the customer from the date the booking entered on the PMS
DepositType	Categorical	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made;	BO and TR/Value calculated based on the payments identified for the booking in the transaction (TR) table before the booking's arrival or cancellation date. In case no payments were found the value is "No Deposit".
		Non Refund – a deposit was made in the value of the total stay cost;	If the payment was equal or exceeded the total cost of stay, the value is set as "Non Refund".
		Refundable – a deposit was made with a value under the total cost of stay.	Otherwise the value is set as "Refundable"

Table 1 (*continued*)

Variable	Type	Description	Source/Engineering
DistributionChannel	Categorical	Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"	BO, BL and DC
IsCanceled	Categorical	Value indicating if the booking was canceled (1) or not (0)	BO
IsRepeatedGuest	Categorical	Value indicating if the booking name was from a repeated guest (1) or not (0)	BO, BL and C/ Variable created by verifying if a profile was associated with the booking customer. If so, and if the customer profile creation date was prior to the creation date for the booking on the PMS database it was assumed the booking was from a repeated guest
LeadTime	Integer	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date	BO and BL/ Subtraction of the entering date from the arrival date
MarketSegment	Categorical	Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"	BO, BL and MS
Meal	Categorical	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)	BO, BL and ML
PreviousBookingsNotCanceled	Integer	Number of previous bookings not cancelled by the customer prior to the current booking	BO and BL / In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and not canceled.
PreviousCancellations	Integer	Number of previous bookings that were cancelled by the customer prior to the current booking	BO and BL/ In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and canceled.
RequiredCardParkingSpaces	Integer	Number of car parking spaces required by the customer	BO and BL
ReservationStatus	Categorical	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	BO

Table 1 (*continued*)

Variable	Type	Description	Source/Engineering
ReservationStatusDate	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	BO
ReservedRoomType	Categorical	Code of room type reserved. Code is presented instead of designation for anonymity reasons	BO and BL
StaysInWeekendNights	Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	BO and BL/ Calculated by counting the number of weekend nights from the total number of nights
StaysInWeekNights	Integer	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel	BO and BL/Calculated by counting the number of week nights from the total number of nights
TotalOfSpecialRequests	Integer	Number of special requests made by the customer (e.g. twin bed or high floor)	BO and BL/Sum of all special requests

[a] ID is presented instead of designation for anonymity reasons.

Not all variables in these datasets come from the bookings or change log database tables. Some come from other tables, and some are engineered from different variables from different tables. A diagram presenting the PMS database tables from where variables were extracted is presented in Fig. 1. A detailed description of each variable is offered in the following section.

2. Experimental design, materials and methods

Data was obtained directly from the hotels' PMS databases' servers by executing a TSQL query on SQL Server Studio Manager, the integrated environment tool for managing Microsoft SQL databases [4]. This query first collected the value or ID (in the case of foreign keys) of each variable in the BO table. The BL table was then checked for any alteration with respect to the day prior to the arrival. If an alteration was found, the value used was the one present in the BL table. For all the variables holding values in related tables (like meals, distribution channels, nationalities or market segments), their related values were retrieved. A detailed description of the extracted variables, their origin, and the engineering procedures employed in its creation is shown in Table 1.

The PMS assured no missing data exists in its database tables. However, in some categorical variables like Agent or Company, "NULL" is presented as one of the categories. This should not be considered a missing value, but rather as "not applicable". For example, if a booking "Agent" is defined as "NULL" it means that the booking did not came from a travel agent.

Summary statistics for both hotels datasets are presented in Tables 2–7. These statistics were obtained using the 'skimr' R package [7].

A word of caution is due for those not so familiar with hotel operations. In hotel industry it is quite common for customers to change their booking's attributes, like the number of persons, staying duration, or room type preferences, either at the time of their check-in or during their stay. It is also common for hotels not to know the correct nationality of the customer until the moment of check-in. Therefore, even though the capture of data took considered a timespan prior to arrival date, it is understandable that the distribution of some variables differ between non canceled and canceled bookings. Consequently, the use of these datasets may require this difference in distribution to be taken into account. This difference can be seen in the table plots of Fig. 2 and Fig. 3. Table plots are a powerful visualization method and were produced with the tabplot R package [8] that allow for the exploration and analysis of large multivariate datasets. In table plots each column represents a variable and each row a bin with a pre-defined number of observations. In these two figures, each bin