



Exploratory Data Analysis on Hotel Booking Demand

February 21st, 2021

Hannah (Bohan) Chou

hannah.chou@student.ie.edu

<https://www.linkedin.com/in/chouhannah/>

Data Source & Background

The data used in this analysis is from [Kaggle](#), however the raw data comes from this science article [Hotel booking demand datasets](#). The dataset records 119,390 bookings made to 2 hotels in Portugal, one city hotel in Lisbon, and one resort hotel in Algarve, between 1st of July of 2015 and the 31st of August 2017.

The original data was pulled from the PMS (property management system). Excluding the hotel column, it has 31 attributes. These attributes can be segmented into 4 categories:

1. Time related columns: any column that is time-bound, e.g. *arrival_date_year*, *arrival_date_month*...
2. Guest information related columns: any column that could describe the person making the order, e.g. *country*, *adults*, *children*, *babies*, *is_repeated_guest*...
3. Order related columns: the columns that are not guest-specific, e.g. *market_segment*, *distribution_channel*, *reserved_room_type*...
4. Issue related columns: any column that is related to cancelation, e.g. *is_cancelled*, *previous_cancellation*...

Goal of the Analysis

The purpose of this analysis is to better understand the correlation between the attributes of the order and cancellation. For example, some of the questions that this analysis aims to answer are:

- Does the group size of the guests influence cancelation?
- Are there any specific nationalities of guests more likely to cancel?
- Does the demographic of the guests of these two hotels differ?
- Do the guests who have previously cancelled more likely to cancel again?

Further topic to explore

Cancellations are undesirable, but No-Shows (customers did not check-in and did inform the hotel of the reason why) are even more troublesome for hotels, because if with enough time in between arrival & cancellation, hotels are more likely to resell the rooms.

The focus of this analysis is on cancellation, however a deep dive on what correlates with No-Show could be beneficial & practical to hotel revenue management.

Approach

PySpark and its SQL function and SQL type packages are the tools that's being used for this analysis.

I. Data Cleaning & Basic Profiling

When the data was originally read into the Spark environment, it had no missing value. However, after investigation, in the column of *agent*, *company* and *children* there are a few null values which were encoded incorrectly for PySpark, so transformation was done on these columns. Several time related columns are also not encoded in the desired datetime format, they were also transformed.

Then the columns are grouped by their unique values to count their frequencies, so the composition of the data could be better understood. In the process, the distribution of their respective values shows, so it could be used to decide if this column has distinguishing power. Because if the majority of a column's values are heavily concentrated on value only, it does not give us additional information.

II. Create New Columns & Segments

Based on the understanding gained from the previous step, a list of new columns is created. For the binned columns, the thresholds are set according to their statistics & frequencies (and some common sense for explainability).

- **total_headcount:** adults + children + babies
- **guest_size:** bin the total_headcount into 4 categories: *solo*, *couple*, *small_group*, *big_group*
- **is_family:** whether there are children or babies in the order or not
- **is_agent:** if this booking comes via a travel agent
- **is_company:** whether this booking is paid by a company or not
- **adr_cat:** bin the adr column into 3 categories: *below_average*, *average*, *above_average*
- **lead_time_cat:** bin the lead_time into 4 categories: *same_day*, *short*, *medium*, *long*
- **total_stays:** stays_in_weekend_nights + stays_in_week_nights
- **stay_length_cat:** bin the total_stays into 5 categories: *one*, *two*, *three*, *medium*, *long*
- **booking_change_cat:** bin booking_changes into 4 categories: *none*, *one*, *some*, *many*

- **season_cat**: whether this order's arrival date is during high season, low season or other

The dataset is also broken down into two to have a more granular understanding of each hotel. For example, The *adr_cat* column needs to be hotel-specific, because each hotel has a different baseline for price.

III. Comparison Across Dimensions

With the metrics created above, cancellation rate can be compared across many different dimensions:

- Guests profile regarding family & group size
- Booking through Company & customer type
- Distribution channel, market segment & agents
- Nationality
- Meals
- Lead time & number of changes made the booking
- Lead time & Price
- Lead time & deposit
- Previous cancellation
- Seasonality

Conclusion

I. Insights

- Group & Transient group orders have much lower cancellation rate across both hotels, while contract orders, though small, have a high cancellation rate especially for leisure travellers (not paid by companies) in the city hotel.
- Non-family couple travellers (cannot assume relationship, just mean it's a party size of 2) are the most likely to cancel for the city hotel, maybe due to the fact that they are the biggest composition in terms of demographic for this hotel. For the resort hotel, the most likely to cancel are the non-family big groups, though they are little in number of orders. The more significant demographic which is prone to cancel is small families with children.
- Orders that come from travel agents are more likely to cancel for both hotels. The lists of the agents who have contributed to high cancellation rates

& sizable orders are provided in the notebook, and the lists for each hotel differs.

- Offline travel agents have lower cancellation rates for the resort hotel, while not so for the city hotel. The segmentation of Groups is bringing sizable orders for both hotels but also has a higher than average cancellation rate.
- Lists of the nationality of the guest with higher cancellation rates & sizable orders have been listed in the notebook, while again the lists are quite different for each hotel, while both have domestic guests as most likely to cancel, due to domestic travellers being the majority for both hotels.
- Guests who have booked full-board meal plans are highly prone to cancellation, though not big in size for both hotels.
- The longer the lead time is, the more likely the order is to be canceled, which is very natural. However, if the cancellation gap is large enough, there are high chances that the hotel could resell the rooms.
- If the guests made just 1 or a few changes to the booking, it has a much lower cancellation rate. Whereas if the guests made either no or many change(s), they are more likely to cancel.
- For the resort hotel, the higher the average daily rate is, the more likely the order will be cancelled. However, for the city hotel in the couple segment, it is not the case. The super long lead time with a below average adr combination has a cancellation for 84%, which is much higher than the baseline (41%).
- For the non-refundable orders, curiously for both hotels, have a very high cancellation rate (99% for city hotels and 95% for resort hotels). Further attributes were considered but cannot seem to find a confounding factor.
- For both hotels, orders arriving in the low season (month of February, November, December and January) have lower cancellation rates, especially more apparent for the resort hotel.
- Those who have canceled previously one time are very likely to cancel again for both hotels. However for the city hotel, those who have canceled 2 or 3 times are actually less likely to cancel compared to a new guest.

II. Recommendation

Based on the insights gained the analysis, two strategies are recommended:

I. Check up with the guest 2 weeks prior to the arrival date

As seen in the booking changes breakdown, guests who have made 1 or few changes are the least likely to cancel. Therefore the hotels could contact the guests beforehand (e.g. 2 weeks before) if there's any changes that shall be made, maybe those who want to cancel would cancel then, and hence cause less no-show and give the hotel more time to resell the room.

II. Overbook when orders have high-cancellation traits


The nationality, size of the group, meal plans and agents are all indicators of how likely these orders will be cancelled. When the demand is high and the rooms are fully booked, the hotel revenue management team could look into the composition of the orders and decide if they should increase supply.

Annex

I. Column Legend

Column Name	Description
hotel	Hotel (H1 = Resort Hotel or H2 = City Hotel)
is_canceled	
lead_time	
arrival_date_year	
arrival_date_month	
arrival_date_week_number	
arrival_date_day_of_month	
stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
stays_in_week_nights	
adults	
children	
babies	
meal	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)
country	Country of origin. Categories are represented in the ISO 3155-3:2013 format
market_segment	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
distribution_channel	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”

is_repeated_guest	
previous_cancellations	
previous_bookings_not_cancelled	
reserved_room_type	Code of room type reserved. Code is presented instead of designation for anonymity reasons.
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 to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
booking_changes	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
deposit_type	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: - No Deposit – no deposit was made; - Non Refund – a deposit was made in the value of the total stay cost; - Refundable – a deposit was made with a value under the total cost of stay.
agent	ID of the travel agency that made the booking
company	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
days_in_waiting_list	
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
adr	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
required_car_parking_spaces	
total_of_special_requests	
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
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