UNIVERSIDAD
POLITÉCNICA
DE MADRID

Hotel Booking Demand Visualisation

Data Visualisation Project Report

Authors:
Zhuo Cheng
Wenqi Jiang
Álvaro Ramírez Sixtos

25 January 2022

Part 1 Introduction

In this project we present an example of using various visualisation tools to show hotel booking demand data to generate knowledge. The dataset used in the project contains data on hotel reservations made during the period between July 1, 2015 and August 31, 2017. The dataset includes detailed information of the reservation, such as number of guests, type of rooms, number of nights, number of children accompanied and so on. From visualising and analysing the data, knowledge on how to improve hotel management and economic performance can be generated.

The structure of this report is described as follows: Part 2 is data description, where information regarding each variable can be found. Design abstraction is given in Part 3, where three different visualisations are explained following the abstraction levels.

The application is published at: http://alvaroame.shinyapps.io/BookingShinyApp.

Part 2 Data description

The dataset represents information regarding hotel reservation during the period between July 1, 2015 and August 31, 2017 for two types of hotel: resort-type hotel and urban hotel. Approximately 119,000 observations were included in the dataset. Each observation contains 32 attributes that include information such as when the reservation was made, length of stay, number of adults, children and/or babies, the number of available parking spaces, among other things. Both checked-in reservations (where guests actually arrived) and cancelled reservations were recorded.

The data is originally from the article Hotel Booking Demand Datasets, published by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The table below provides brief descriptions of each attribute and its type.

Variable	Type	Description
ADR	Quantitative	Average Daily Rate
Adults	Quantitative	Number of adults
Agent	Categorical	ID of the travel agency that made the booking
ArrivalDateDayOfMonth	Cyclic	Day of the month of the arrival date

ArrivalDateMonth	Cyclic	Month of arrival date with 12 values: “January” to “December”
ArrivalDateWeekNumber	Cyclic	Week number of the arrival date
ArrivalDateYear	Integer	Year of arrival date
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
Babies	Quantitative	Number of babies
BookingChanges	Quantitative	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
Children	Quantitative	Number of 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
Country	Categorical	Country of origin.
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

DaysInWaitingList	Quantitative	Number of days the booking was in the waiting list before it was confirmed to the customer
DepositType	Categorical	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: 1. No Deposit – no deposit was made; 2. Non Refund – a deposit was made in the value of the total stay cost; 3. Refundable – a deposit was made with a value under the total cost of stay.
DistributionChannel	Categorical	Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"
IsCanceled	Categorical	Value indicating if the booking was cancelled (1) or not (0)
IsRepeatedGuest	Categorical	Value indicating if the booking name was from a repeated guest (1) or not (0)
LeadTime	Quantitative	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
MarketSegment	Categorical	Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
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)
PreviousBookingsNotCanceled	Quantitative	Number of previous bookings not cancelled by the customer prior to the current booking

PreviousCancellations	Quantitative	Number of previous bookings that were cancelled by the customer prior to the current booking
RequiredCardParkingSpaces	Quantitative	Number of car parking spaces required by the customer
ReservationStatus	Categorical	Reservation last status, assuming one of three categories: Cancelled – booking was cancelled 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
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 cancelled or when did the customer checked-out of the hotel
ReservedRoomType	Categorical	Code of room type reserved. Code is presented instead of designation for anonymity reasons
StaysInWeekendNights	Quantitative	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
StaysInWeekNights	IntQuantitativeeger	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
TotalOfSpecialRequests	Quantitative	Number of special requests made by the customer (e.g. twin bed or high floor)
Hotel	Categorical	Hotel (H1 = Resort Hotel or H2 = City Hotel)

Part 3 Design abstraction

In this part, we follow the four abstraction levels for each visualisation idiom. We first characterise the specific problem in the domain, explain the variable used and define the task, encode the data and finally implement it. The objective is to use data visualisation to provide valuable information that can help improve hotel management and economic performance. Such information and knowledge can be of great interest to the regional manager of a major hotel chain that uses reservation information to make the best decisions regarding adjustments to its facilities, marketing strategies to attract more guests, and everything else needed to offer the best possible experience to its guests.

3.1 How is demand distributed throughout the year?

3.1.1 Problem characterization

Demand is clearly one of the key interests from the managerial point of view. Which months have the highest demand? Is there any cyclic pattern? Which season is relatively low? What strategies can be taken to meet different demands? Analysts want to find out answers to these questions because they are directly related to the economic performance of any hotel.

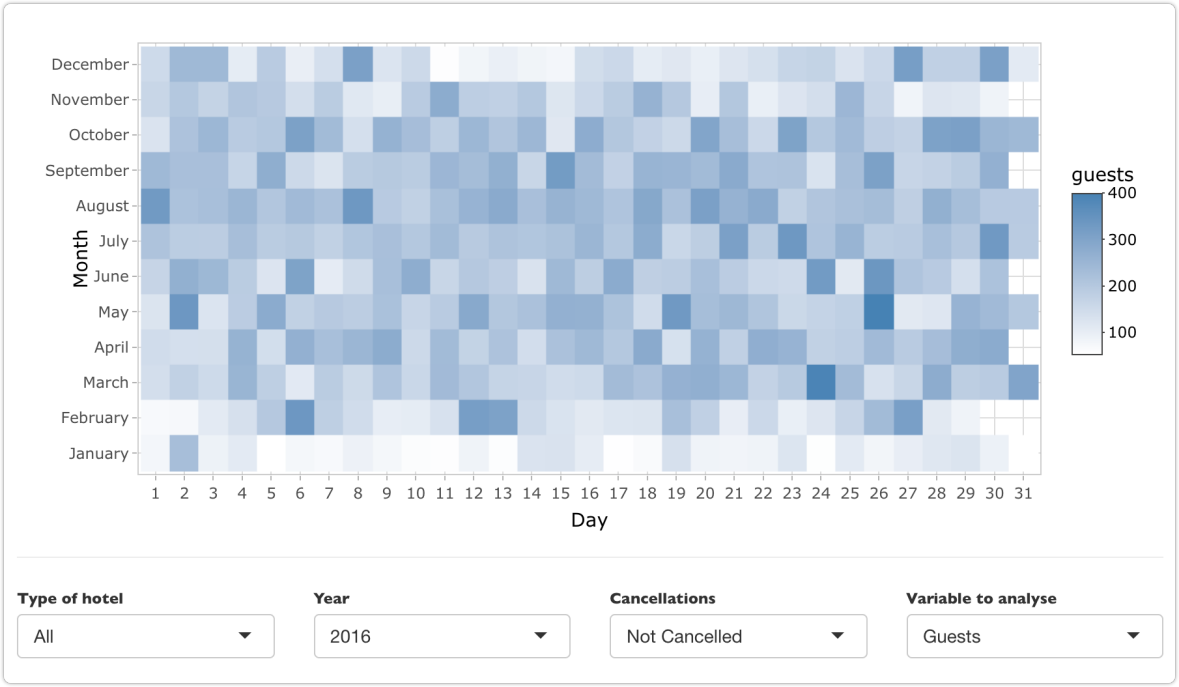
3.1.2 Data and task abstraction

In order to see how demand is distributed, the following variables are selected in the presentation: *ArrivalDateDayOfMonth*, *ArrivalDateMonth*, *ArrivalDateYear*, *Hotel* and *IsCanceled*. New variables *guests* and *nights* are created by summing *Adults*, *Children*, and *Babies*, and by summing *StaysInWeekendNights* and *StaysInWeekendNights*. The arrival date variables are used as the key attribute and they correspond to a specific date of a year. All the variables except Hotel and IsCanceled are numeric.

The task here is to present the demand information to the management in order to discover new knowledge. Multiple attributes are used because the idea is to find the temporal distribution of demand. Filters and facets are used to make the visualisation informative.

3.1.3 Interaction and visual encoding

The idiom used for this task is heatmap. Its two axes are month and day of the month. Other variables are given in the drag-down menu, where users can choose the specific combination of the variables. Heatmap is a good tool to use in this case because it gives a very straightforward overview of demand during a year directly by the darkness of the color.



The design needs to be validated by checking for instance if the data is complete, is there any outlier and so on. Once it is validated, analysts can proceed to the encoding and implementation stage. One drawback of heatmaps is that it doesn't provide aggregated information directly. For example, it is difficult to tell which month has the largest demand. It is also not possible to capture cyclic patterns from heatmaps.

The following table summarises the design abstraction process:

Idiom	Heatmap
What	Table: date used as key attribute, one quantitative variable used in each view, multiple views (number of guests/reservations/nights)
How	Color and grid
Why	To show peaks and valleys

3.2 Is there any correlation between lead time and the number of total nights the guest stayed in the hotel?

3.2.1 Problem characterization

The business problem here is to find the best time to launch marketing campaigns. From the previous task, it is possible to visualise the peak of demand. Based on this information, analysts

may want to find the relation between time of making reservations and the number of nights, if there is any. Such information can be used to guide the timing of marketing.

3.2.2 Data and task abstraction

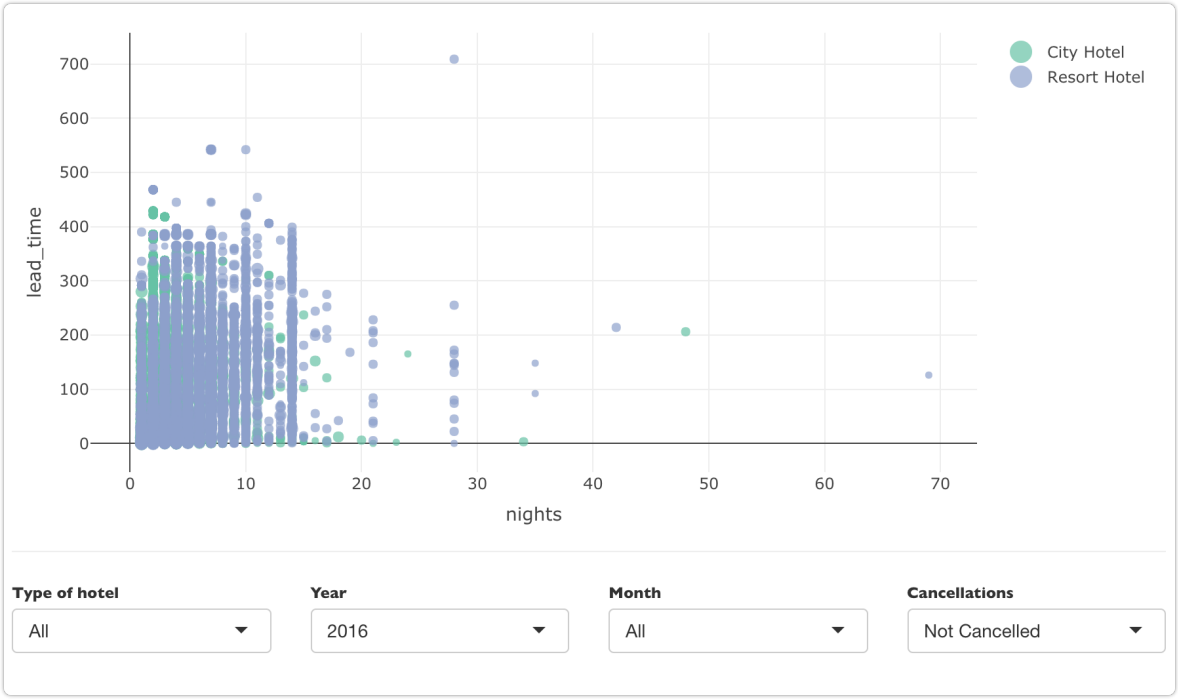
Two variables are used: lead time and nights. Lead time is defined as the number of days that elapsed between the entering date of the booking into the PMS and the arrival date. Both of the two variables are numeric.

The task is to discover the correlation between these two variables, if there is any. Multiple attributes are used because correlation measures the relationship between two variables. In terms of design choices, data is encoded by dots. Colors and sizes are incorporated to add more visualisation options.

3.2.3 Interaction and visual encoding

The idiom used is scatterplot. Nights and lead time are plotted in the two axes. Variables type of hotel, year, month and cancellation are given in the drag-down menu, where users can choose the specific combination of the variables. Type of hotel is distinguished by color, and the size of the scatter indicates the number of guests.

The design needs to be validated by communicating with the users, because from the result it doesn't seem to have any correlation. Therefore, there might be model misspecification risk. Due to the large number of observations and their distribution, scatterplot does not provide a user-friendly view. These could be improved by respecify the model and choosing a better idiom.



The following table summarises the design abstraction process:

Idiom	Scatterplot
What	Table: two quantitative variables used in each view, multiple views (type of hotel, year, month and cancellation)
How	dots, color, size
Why	Scatterplot can show correlation in a straightforward way

3.3 How has the market and the distribution channel evolved?

3.3.1 Problem characterization

From a marketing standpoint, the selection of distribution channels is obviously a major consideration. Which channels produce the best outcomes and which produce the worst? Is there a consistent trend between direct reservations from consumers and marketing platforms? What strategies may be used to determine the differences between marketing channels and which quarters or months require additional marketing? The analysts wanted to be able to evaluate the data to discover answers to these issues, which are critical to the hotel's operating costs.

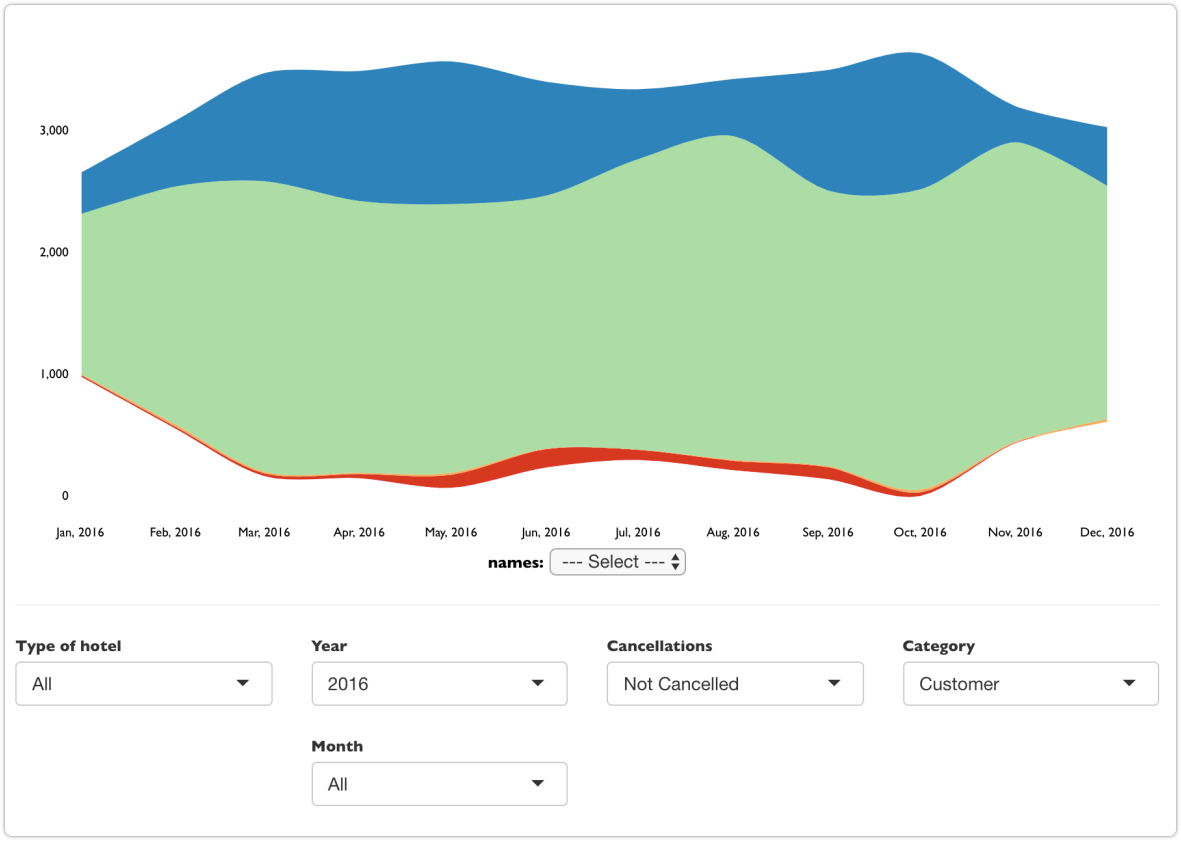
3.3.2 Data and task abstraction

The relevant variables were chosen in the presentation to analyse the trends of different booking distribution channels and client types: distribution channel, customer type, market segment. We may create the Stream Graph by collecting orders from several types of channels.

The aim here is to demonstrate to management the distribution channels and differences in the temporal distribution of orders from various types of users in revealing new opportunities. Furthermore, more specific information can be given by hotel type, date, and booking cancellation.

3.3.3 Interaction and visual encoding

The stream graphs are a more complex generalisation of stacked bar charts, where the main axis (horizontal or vertical) represents an ordered attribute (date) and the complementary spatial dimension represents a categorical key attribute together with a quantitative attribute. It may see orders from different channels for a certain year or month by utilising the drop-down choices below the graph. Managers can easily determine which channel is bringing in the most customers based on the area of the graph.



Before implementation, the data must be pre-processed, null value or outliers must be filtered out in advance. If the data has been validated, the next step is to develop and deploy it. Although Stream Graphs are easy to read and intuitive, they do have some downsides. As with any stacked area chart, they can easily get cluttered when there are too many categories at once and Stream Graphs is that the lack of a y axis makes it impossible to read the scale at a glance.

The following table summarises the design abstraction process:

Idiom	Stream Graph
What	Multidimensional table: date used as key attribute one quantitative value – attribute order number one categorical attribute – distribution_channel/customer_type
How	Colour and derived geometry
Why	Show trend and differences in channels

LINK TO THE PROJECT: http://alvaroame.shinyapps.io/BookingShinyApp