Final Project

STAE04/DABN19: Data Visualization

Anja Braeu

2021-10-28

1 Introduction

The Covid-19 pandemic forced tourism to a hold. Most bookings had to be cancelled or postponed indefinitely, leaving many hotels empty for a long time. As tourism slowly returns to normal, this project aims to visually analyze which characteristics may possibly lead to a cancelled booking when no global pandemic impedes travelling. This may help hotels to better forecast their capacities and adjust cancellation policies accordingly.

For this cause a dataset containing information on hotel bookings between July 2015 and August 2017 is used. The original data comes from the article `Hotel Booking Demand Datasets` by Nuno Antionio, Ana Almeida and Luis Nunes in 2019, but was tidied by Antoine Bichat and Thomas Mock in 2020 for #TidyTuesday. The tiedied version of the Hotel Booking Dataset can be found on Kaggle. It is a very extensive dataset, offering 119.390 observations and 32 variables of different datatypes. Table 1 gives an overview of the variables that are used in the following visual analysis.

Table 1: Overview of the used Variables, their Data Types and Description

Variable Name	Type	Description
is_canceled	integer	Dummy variable indicating whether booking was canceled (1) or not (0)
hotel	character	Type of Hotel (City or Resort)
lead_time	integer	Number of days between booking and arrival date
arrival_date_month	factor	Month of arrival
stays_in_weekend_nights	integer	Number of weekend nights booked (Saturday, Sunday)
stays_in_week_nights	integer	Number of week nights booked (Monday to Friday)
adults	integer	Number of Adults
children	integer	Number of Children
babies	integer	Number of Babies
country	character	Country of origin (ISO 3155–3:2013 format)
is_repeated_guest	integer	Dummy variable indicating whether booking is made from a repeated guest (1) or not (0)
adr	numeric	Average daily rate of the hotel

2 Data Analysis

In the following analysis, characteristics of bookings that are cancelled versus booking that are not cancelled are visually investigated.

2.1 Tidying and Wrangling the data

Since the dataset is optimized for visualisation and machine learning projects by the authors and further tidied for #TidyTuesday, it could potentially be used as it is for the exploratory analysis. However, some minor modifications are done. The dummy variables indicating if a booking was canceled and if the booking was made from a repeated guest were changed to a factor with expressive names for the levels. Further, the months of arrival were arranged in the right order. In the original dataset, the number of nights stayed at a hotel are split into weekend and week nights, which does not serve a purpose for this report. Thus, these variables are summarized into one, indicating the total number of nights stayed at a hotel. In order to then compute the total cost of the stay, the total number of nights stayed are multiplied with the average daily rate. Lastly, the variable stating the number of children is added to the `children` variable, as it would not yield any additional information for the cause of this analysis.

2.2 Repeated guest

First, the difference in cancellations between first-time and repeated guests is visualized. The mosaic plot in Figure 1 clearly shows that most guests are not repeated guests in a booked hotel. However, it is unsurprising that repeated guests seem less likely to cancel a booking than first-time guests. There are several possible reasons for this, for example that repeated guests already made a good experience with a certain hotel and make a more conscious decision for exactly this hotel than first-time guests that may even book several hotels for the same dates.

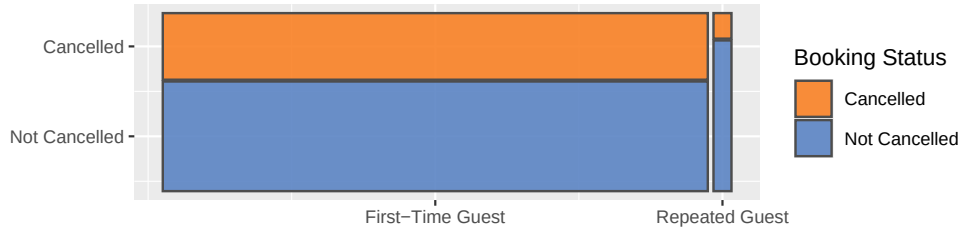

Figure 1: Booking Status for First-Time and Repeated Guests

2.3 Month of Arrival

Second, there may be differences in the number of hotel bookings and cancellations for each month, which the histogram in Figure 2 illustrates. The highest number of bookings is registered for the summer, whereas hotels didn't seem to be in high demand in the winter. This is probably due to the fact that most guests are from the European countries, where a majority of people take their holiday in the summer. Unfortunately, the dataset does not specify, in which country the booked hotels are, but this booking pattern leads to the assumption that the hotels may be situated in Europe. As is, the histogram further leads to the assumption that while the number of

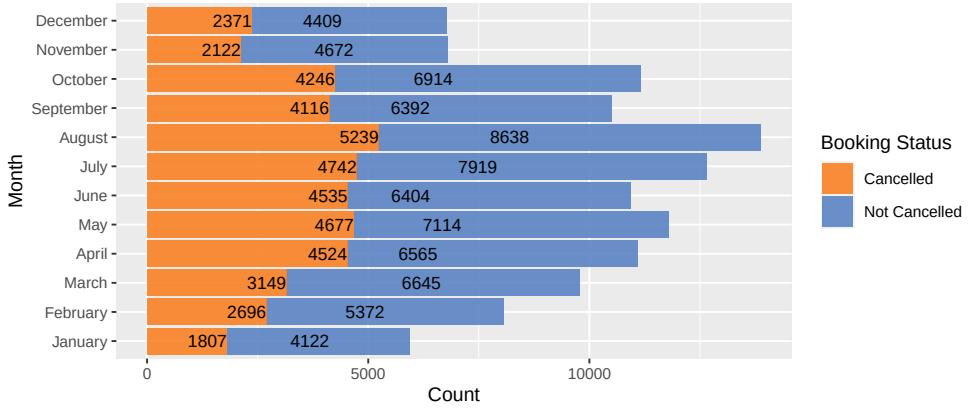

Figure 2: Number of Bookings and Cancellations by Month

bookings differs by month, the cancellation rate seems similar for each month. Thus, the cancellation rate is calculated by grouping the dataset by month and dividing the number of cancellations by the total number of bookings. Interestingly, June has the highest cancellation rate with 41.5% of total bookings being cancelled. It is followed by April, May and September. In contrast, the cancellation rates in the winter months are comparably low. The lowest cancellation rate is being recorded in January, with only 30.5% of total bookings cancelled. Overall, there is no clear pattern, but it seems that when the demand is not so high, bookings are less likely to be cancelled.

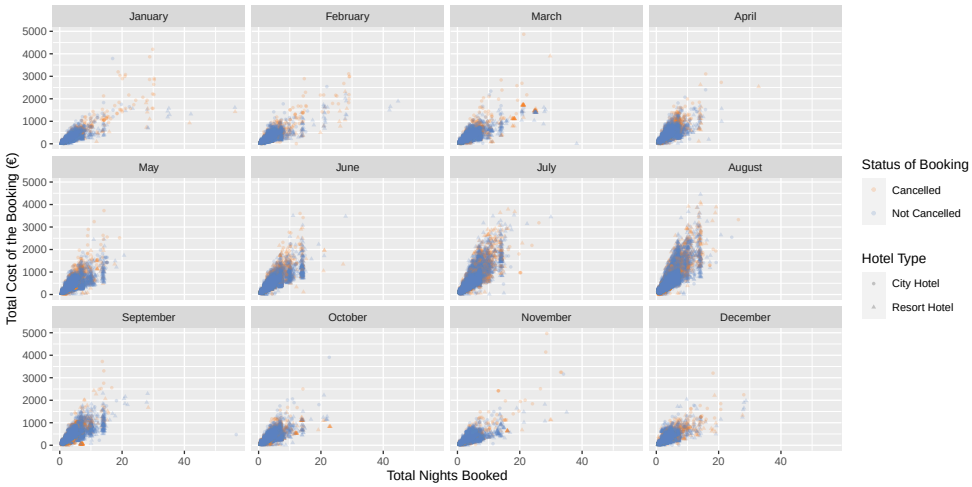

Figure 3: The Relationship between Cancellations, Total Nights Booked, Total Cost and Hotel Type by Month of Arrival

Figure 3 tries to detangle possible reasons for the comparably higher cancellations rates in high season. The number of nights is plotted against the total cost of the booking for each month. The colors indicate whether a booking was cancelled or not and the shapes indicate the type of hotel. It is evident that the total cost of the bookings increase towards August and decrease towards January. Further, Resort Hotels seem to be more in demand the summer month compared to the winter. Generally, guests seem

to book a longer and more costly stay in Resort compared to City Hotels. Generally, most guests seem to book stays between one and 20 nights, with the highest variety being in January and February. As there are many observations in this dataset, it is not too easy to see which bookings were cancelled or not.

However, there are three conclusions to be drawn from Figure 3: First, bookings for City Hotels seem to be cancelled more easily than for Resort Hotels. This is confirmed by calculating the cancellation rate for each Hotel Type. Overall, 41.7% of bookings for City Hotels are cancelled, whereas only 27.8% of bookings for Resort Hotels are cancelled. Second, costly stays seem to be cancelled more often than cheaper stays, which is confirmed when splitting the variable `Total cost` into categories (not shown). This does not necessarily apply to July and August. Last, the findings from Figure 2 are emphasized and partly explained by the differences in prices and demand.

2.4 Time between Booking and Arrival Date

Third, the time between booking and arrival date may influence whether a guest cancels a booking or not. The boxplots for cancellations by hotel type based on lead time show a similar picture for both resort and city hotels. Comparing the medians of cancelled and not cancelled bookings, the chances for cancellation are higher if the time elapsed between booking and arrival is fairly long. This may be connected to cancellation policies and that guests have more time to change their plans or find better deals. Interestingly, the medians of Resort Hotels lie before the medians of City Hotels, meaning that guests generally seem to book Resort Hotels more spontaneously than City Hotels.

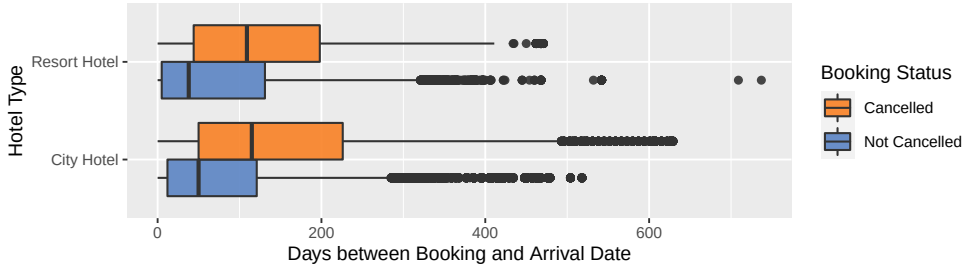

Figure 4: Cancellations by Hotel Type based on Time between Booking and Arrival Date

2.5 Country of Origin

Fourth, there may be differences in cancellation rates based on a guest's country of origin, which the worldmap in Figure 5 illustrates. To ensure representability, no rate was calculated for countries with less than 10 observations. Based on this map, guests from European countries except for Portugal, as well as from the United States and Australia have fairly low cancellation rates between 16% (Germans) and 30% (Norwegians). Guests from Portugal cancel 57% of all their bookings. Since there are not many observations from guests with other nationalities, it is refrained from generalizing the cancellation behavior for guests from other countries. Nevertheless, some tendencies are evident.

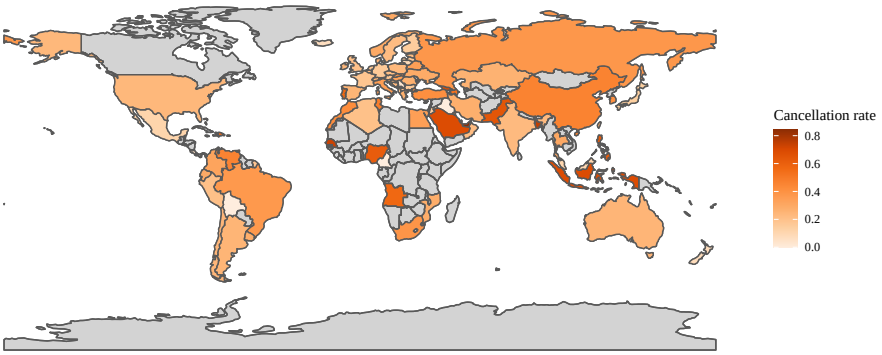

Figure 5: Cancellation Rate by Country of Origin

2.6 Category of Traveller(s)

Last, the category of traveller(s) may be a characteristic that influences the cancellation behaviour. Figure 6 plots the cancellations based on category of traveller, hotel type, lenght of stay and the total cost of the booking. Overall, previous findings are reconfirmed, for example that bookings for City Hotels are more likely to be cancelled than for Resort Hotels among all traveller categories. Additionally, most bookings are made for two adults. These tend to prefer Resort over City Hotels, where they tend to book more costly stays. A group of adults seems to have the lowest cancellation rates for Resort Hotels among all traveller categories, but they tend to not stay longer than one or two weeks. Naturally, the total cost of the booking seems higher for groups and families than for guests travelling alone, whereas two adults can spend a considerable amount on a hotel. It can be assumed that most of them are couples, who might enjoy a relaxing holiday with some luxury.

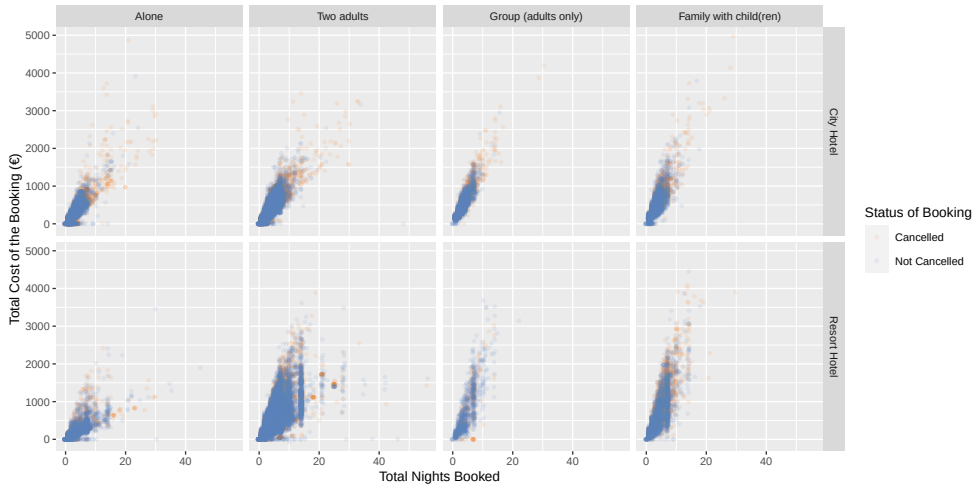

Figure 6: Cancellations based on Category of Traveller(s), Hotel Type, Total Nights Booked and Total Cost

3 Conclusion

In conclusion, the visual analysis of the `Hotel bookings` dataset reveals some insights on characteristics that influence whether a booking os canceled or not. A guest that books a specific hotel for the first time is more likely to cancel than a repeated guest. Further, the month of arrival seems to make a difference, such that bookings for the summer have higher cancellation rates than for the winter. In addition, the cancellation rates for Resort Hotels are considerably lower compared to City Hotels. Also, the total cost of a booking influences the cancellation behavior such that costly bookings are more likely to be cancelled than comparably cheaper stays. The length of a stay on the other hand does not seem to majorly determine cancellations. Looking at the time between booking and arrival, stays that are shortly after the booking date are less likely to get canceled than if there is a long time in between booking and arrival. One should be careful when making assumptions of cancellation rates by country of origin based on the used dataset. Among guests from European countries, Germans have the lowest cancellation rate, whereas Portuguese guests are more likely to cancel their booking than not. Lastly, the category of traveller(s) does not reveal major differences in the cancellation behavior of different types of travellers.

In sum, a group of repeated guests that book a relatively cheap stay in the off-season for a Resort Hotel close to the arrival date seem to be least likely to cancel their booking. Possibly, they are from a European country, such as Germany. However, as the used dataset contains a large number of observations, some plots are hard to interpret regarding the differences between cancelled and not cancelled bookings. Also, one should be careful when generalizing these findings, as there may be a relationship between the investigated variables and cancellations, but these could also be due to some variables that are excluded from the dataset (omitted variable bias). Nevertheless, this analysis revealed some interesting insights and may be useful when trying to explain why a booking was cancelled.