4/26/2020

**GROUP 8**

Amritha Rekha

Jaswanth Pedapalli

Ming Ming He

Sri Ram

Vijit Vengsarkar

**Data Visualization**

**Hotel Booking Demand**

1. **Executive Summary**

Our aim through this Visualization is to tell a story based on a hotel demand dataset by creating few visualizations. The dataset consists of data for two kinds of hotels (City/Resort). The story we are portraying here is from the hotel management perspective of which is the best time of the year to visit Portugal and what are the strategies that the hotel management should come up with in order to attract the customers.  
We were able to find some interesting relationships like the planned customers are less likely to cancel their reservations. If the hotel management is able to fulfill the customers special requests they are likely to stay at the hotel. Summer and late fall is the best time for hotel management as many of the tourists will be visiting the country due to festivals and best weather in Portugal.

1. **Dataset Link**

<https://www.kaggle.com/jessemostipak/hotel-booking-demand>

1. **Data Description**

The Dataset consists of two kinds of hotel demand data. One of the hotels (H1) is a resort hotel and the other is a city hotel (H2). Both datasets share the same structure, with 31 variables describing the 40,060 observations of H1 and 79,330 observations of H2.

Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled.

Since this is hotel real data, all data elements pertaining to the hotel or customer identification have been deleted.

Following is the table consisting of each variable and its description in the dataset:

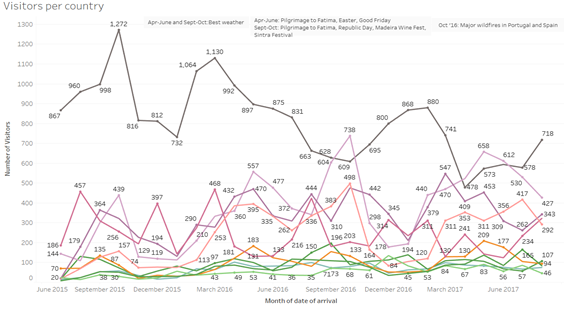
|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Description** |  |
| *ADR* | Numeric | Average Daily Rate |  |
| *Adults* | Integer | Number of adults |  |
| *Agent* | Character | ID of the travel agency that made the booking |  |
| *ArrivalDateDayOfMonth* | Integer | Day of the month of the arrival date |  |
| *ArrivalDateMonth* | Character | Month of arrival date |  |
| *ArrivalDateWeekNumber* | Integer | Week number of the arrival date |  |
| *ArrivalDateYear* | Integer | Year of arrival date |  |
| *AssignedRoomType* | Character | 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* | Integer | Number of babies |  |
| *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 |  |
| *Children* | Integer | Number of children |  |
| *Company* | Character | 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* | Character | Country of origin. Categories are represented in the ISO 3155–3:2013 format |  |
| *CustomerType* | Character | 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* | Integer | Number of days the booking was in the waiting list before it was confirmed to the customer |  |
| *DepositType* | Character | 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. |  |
| *DistributionChannel* | Character | Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” |  |
| *IsCanceled* | Character | Value indicating if the booking was canceled (1) or not (0) |  |
| *IsRepeatedGuest* | Character | Value indicating if the booking name was from a repeated guest (1) or not (0) |  |
| *LeadTime* | Integer | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date |  |
| *MarketSegment* | Character | Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators” |  |
| *Date of Arrival* | Character | Concatenation of *ArrivalDateMonth* and *ArrivalDateYear* |  |

1. **Data Preprocessing**

There was not much preprocessing required in the dataset as most of the data that was acquired had clean data. However as far as cleaning the data from the point of view of our project was concerned, we cleaned the data with respect to below mentioned points in Microsoft Excel:

1. Removed the variable *reserved\_room\_type*. This variable only had room type codes for anonymous reasons and would not have helped in any of the vsiualizations as no meaning can be derived just from room type codes
2. Removed the variable *assigned\_room\_type* for similar reason as *reserved\_room\_type*
3. Removed the variable company as there was no insight that could be derived using this variable. Also about 94% of this variable had null values.
4. Added a new variable *date of arrival* by concatenating variables *ArrivalDateMonth* and *ArrivalDateYear* to make it easy to visualize date of arrivals in Month-Year format
5. **Insights and Findings**

## **Visitors and country trend**

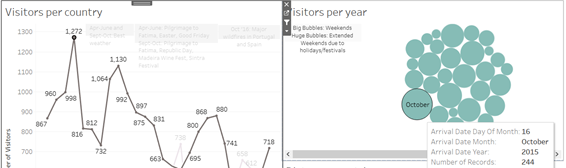


**Visitors per Country**

**Variables used:** Is\_cancelled -Describes whether a customer cancelled the booking or really visited the hotel, Country – Country of origin of the visitor, Date of Arrival – Visitor arrival date

**Method:** We have used the date of arrival to plot the number of visitors visiting each month of the year from 2015 July to 2017 July. The different color trend lines are the visitors from different countries with the number of visitors plotted on Y axis and Date of arrival on X Axis

**Insights and Findings:** The Visitors per country worksheet emphasizes on the visitors visiting from different countries of origin to Portugal at what time of the year using a trend line graph.  
We can see that from 2015 to 2017 there is an obvious higher number of visitors from Portugal as these can be considered as local tourists.  
The general trend observed with respect to all the countries was the rise in the graph observed specially in the months from April to June and September to October. This is because this is the best time to visit Portugal given the weather conditions are the best during these months.   
Also all the major festivals in Portugal such as Pilgrimage to Fatima, Easter, Madeira Wine Festival are celebrated during these months.  
The only surprising trend was observed in the fall of Portugal tourists in the month of October 2016 as compared to October 2015. This can be credited to outbreak of major wildfires in Portugal as also in Spain in October 2016. However this did not stop the tourists from countries such as UK, Germany, Ireland and Italy visiting Portugal. These are the major insights that can be inferred from this visualization.

****

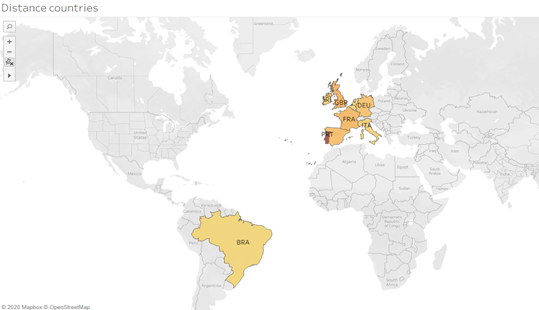
**Visitors per Year**

**Variables used:** Arrival Date Month – Month of Arrival of Visitor, Arrival Date Year – Year of Arrival of visitor, Country, Arrival Date Day of Month – Individual Visitory arrival day

**Method:** We have made use of Bubble chart here to show the number of visitors visiting on each day of the year from July 2015 to July 2017. It becomes easy to visualize the larger group with larger bubbles

**Insights and Findings:** This worksheet shows the number of visitors that have visited Portugal on each day every year from 2015 July to 2017 July. This worksheet is an extension to Visitors per country in a sense that a general trend can be observed based on when the visitors ideally visit Portugal.

For example for any high spike in visitors a general trend is observed that the maximum number of visitors have either visited Portugal on a weekend or during an extended weekend. By extended weekend it means that following a weekend there is either a public holiday or some festival. For example in the figure above we can see 16th October 2015 had the highest number of visitors in Portugal for the spike that we see for October 2015 to the left.  
This is because October 16th 2015 was a Friday and the following Monday was the beginning of the festival Pilgrimage to Fatima



**Distance Countries**

**Variables used:** Country

**Method:** We have simply plotted the country codes in the dataset based on their geographical locations on the map

**Insights and Findings:** This visualization shows that maximum visitors that come to Portugal belong to EU and UK. All these countries are closest in distance to Portugal as compared to all the other countries. The only surprising inclusion is that of Brazil. Brazil has similar social and cultural ties as Portugal and hence the general high number of visitors from Brazil to Portugal can be observed.

## **Reservation Status Statistics based on Arrival Date Trend**

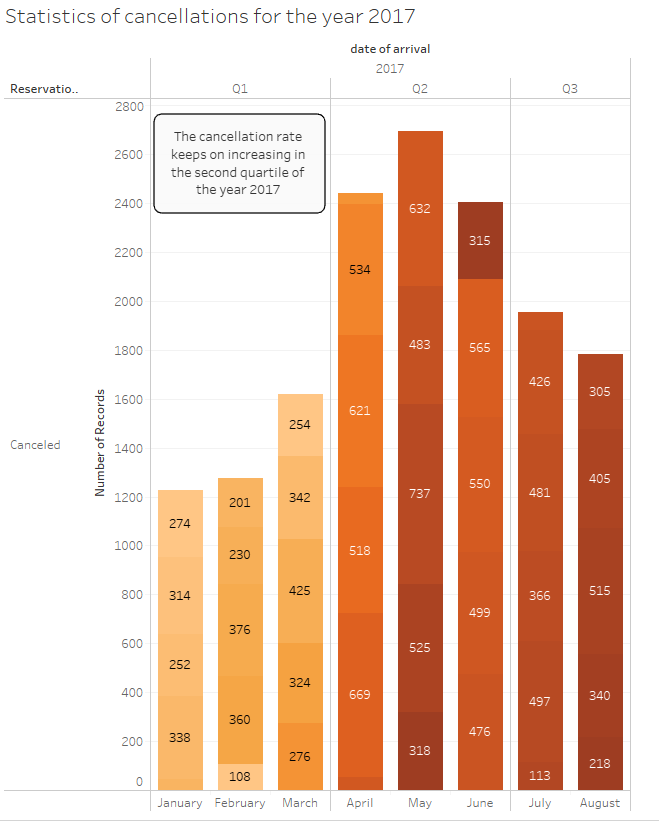
**Reservation Status per month**

**Variables used:** Reservation\_status – Last reservation status. It can be assumed to be in one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed and No-Show – customer did not check-in, Date of Arrival

**Method:** The above line chart shows the trend line of the cancellation for the entire duration specific to months. We used reservation status and the date of arrival to do segmentation on the data. The X-axis denotes the date of arrival and Y-axis denotes the number of cancellations

**Insights and Findings: A**mong the three consecutive years, the year 2017 shows the highest peak in the cancellations that is 2695 for May 2017. For the year 2015, we could see the cancellation reached its highest point for ***September 2015*** which is due to the West Nile virus outbreak. Health departments had cautioned people visiting endemic areas are at high risk. Older people and the people with weakened immune systems were expected to be at increased risk of getting ill. Peak transmission occurred during the summer months. For the year 2016 we could see the cancellation reached its highest peak for ***October 2016*** which is due to the RHDV2 - rabbit hemorrhagic disease virus 2.   
For the year 2017, we can see that the cancellation reached its highest peak for ***May 2017*** due to the Measles outbreak. The Lisbon area had been alerted due to this one of the most highly contagious infectious human diseases and can cause serious illness, lifelong complications, and death. We can say that the cancellations that occurred during these months of the year were due to disease outbreak that in Portugal.

**Status of cancellations for the year 2017**

****

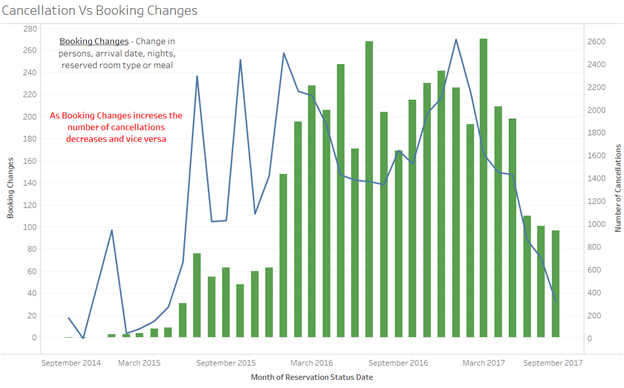
**Cancellations 2017**

**Variables used:** Reservation\_status, Arrival Date, Arrival Date Week Number – Week number of visitor arrival in ap articular year

**Method:** The stacked bar chart shows the cancellations for the year 2017 based on the three quartiles. The Q2 is having the maximum number of cancellations for the year 2017. It is clear from the graph that the no of cancellation is in an increasing trend from the starting of the Q2 (April) and keeps on increasing till May and then affects a slow drop.

**Insights and Findings:** As we could see from the previous chart that the year 2017 had the highest cancellation peak. We moved on to investigate more further in the year 2017 by dividing them into quartile segments.  
It is clear from the above stacked bar chart that the second quartile of the year 2017 has the highest cancellations for the year. Hence we could confirm that the cancellations that happened in May are the highest for Quartile 2.

**Cancellations Vs Request Trend**

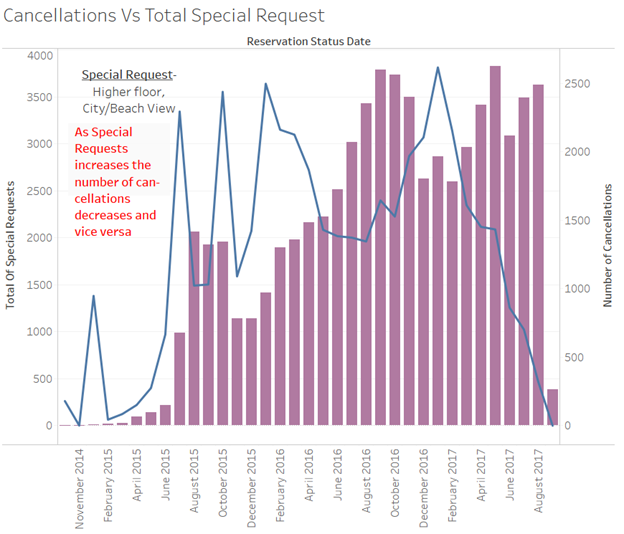
****

**Cancellation Vs Booking Changes**

**Variables Used:** Reservation Status Date, Booking Changes - Change in persons or arrival date or nights or reserved room type or meal, Is Cancelled

**Method:** The above chart is a dual Combination chart, it’s a combination of bar and line chart. The bars indicate the number of booking changes made by the customers. The line chart indicates the cancellation made by customers. The x- axis denotes the months and year of registration date.

**Insights and Findings:** As we can see from the above chart that as the number of booking changes increases, the number of cancellations made by the customer’s decreases, and vice versa.   
If the hotel agrees to the customer proposed changes then the number of cancellations will come down thereby maximizing the occupancy.

****

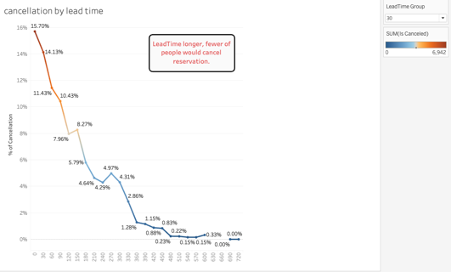
**Cancellation Vs Total Special Request:**

**Variables Used:** Special Request- Higher floor, City/Beach View, Reservation Status

**Method:** The above chart is a dual Combination chart, it’s a combination of bar and line chart. The bars indicate the number of special requests made by the customers. The line chart indicates the cancellation made by customers. The x- axis denotes the months and year of registration date.

**Insights and Findings:** As we can see from the above chart that accepting special requests from customers decreases the cancellations. If the hotel agrees to the customer requested changes then the number of cancellations will come down there by maximizing the occupancy.  
As there is an increase in the number of requests made by customers accepted by the hotel, the number of cancellations come down, so the hotels should try to accommodate the changes requested by the customers to increase the hotel occupancy.

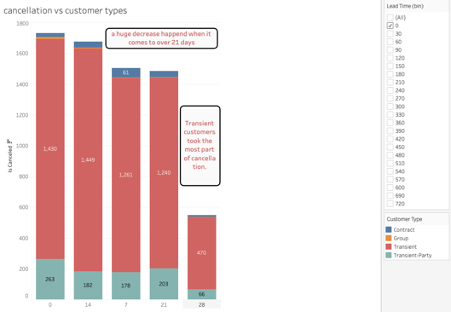
## **Cancellation vs Lead Time Trend**

**Cancellation by Lead Time**

**Variables Used:** Lead time - Days between the entering date of the booking into the PMS and the arrival date, Is Cancelled

**Method:** We used the scatter line to show the trend of cancellations. Colors were also used to show the sum of cancellations. Orange represented the high quantities, while blue represented the low number of cancellations. We used these 2 methods because we would like to see if there any trends of cancellations when the lead time changes. The scatter line obviously showed the negative linear relationships in the graph.

**Findings and Insights:** Based on this graph, we found if the lead time goes longer, the number of cancellations became fewer. Most of the cancellations happened from 0 to 30 group. It means customers who made a hotel reservation 30 or fewer days ahead had took a large part of cancelling reservations.



**Cancellation vs Customer Type**

**Variables Used:** Lead Time, Customer Type - Transient customers are the travelers who are walk-in guests, last-minute booking guests or Simply individual guests requiring a short stay at the hotel, Is Cancelled

**Method:** We used the bar chart to visualize the cancellation numbers within 30 days of lead time. We sorted the cancellation data by descending to rank the groups of lead time. We also used 4 different colors to represent the different customer types. The combined bar chart made it clear to find out which kind of customers made the most cancellations.

**Findings and Insights:** Based on this graph, we found a sharp decrease of cancellations happened when the lead time was over 21 days. It means that people who made reservations 21 or less days ahead of the arrival date preferred to canceling their reservations. We also found that transient customers contributed most of the cancellations in each lead time group. 21 days group had almost 800 transient customers more than the guests in 28 days group. It can be one of the reasons to explain the sharp decrease in the bar chart.

From these two charts, we can say that planned people are less likely to cancel their reservations. Most of the transient customers are walk-in customers or guests who only need a short stay in the hotel. They made decisions in a shorter time compared to planned people. So, the hotels should set planned people as their target customers.

1. **Conclusion**

Through the visualization, we have drawn insights to increase the occupancy of the hotels located in Portugal. The insights which are discussed are based on the bookings made by customer, country and season. So, hotels should focus more on the peak travel months that are April - June and September-October for maximum occupancy demand and the check-out peak is re-occurring every year for the month of May / October due to the visit to Pilgrimage to Fatima.   
Also, If the hotel is trying to accommodate the requests made by the customers then the cancellation made by customer’s are decreasing. As well as, the Hotel could reduce the openings for the transient guests. In this way, hotels could decrease the risk of high cancellations of booking.

The Hotels should focus on these insights to increase the maximum occupancy demand.

1. **References**
2. https://www.kaggle.com/jessemostipak/hotel-booking-demand/data#
3. https://www.sciencedirect.com/science/article/pii/S2352340918315191