PSTAT 195 Final Project

Data Analysis - Portugal Hotel Booking Demand and Revenue

Report by: Meiyu Li, Ziqing Luo

Director: Alexander Franks, Yekaterina Kharitonova

June 12th, 2020

1 Abstract

For hotel industry, revenue management is the key to the business's success. This analysis aims to provide a deeper insight into hotel's market performance and help improve revenue management strategies by exploring seasonality, booking cancellations rates, customer segmentation and average revenue prediction. The dataset used was obtained through a keyword search on Kaggle and contains booking information about a resort hotel and city hotel in Portugal. The methodology utilized includes data exploration composed by various data visualization and multi-linear prediction models. The results show that summer is the on-season for travel and the high booking cancellation rates have severe impacts on the hotel's revenue. Better overbooking policies and further research is still needed.

2 Introduction

Revenue management(RM) refers to a series of tactics and methods firms used to scientifically manage demand for their products and services.[3] The practice has grown from its origins in airlines to its status today as a mainstream business practice in a wide range of industry areas, including hospitality, energy, fashion retail, and manufacturing.[3] In tourism and travel related industries, revenue management strategies are of particular significance because they enable the hotels to anticipate demand and optimise pricing and occupancy in order to achieve the ultimate goal of maximizing profits. This work aims to help the hotel owners to understand the patterns and trends of hotel demand and contribute to the improvement of hotels' revenue management techniques. Specifically, it explores three main **questions**: when is the on-season and off-season of hotel visits, how does the cancellation rates vary and what affects them, as well as how we can predict the revenue based on the booking information.

The data set used consists of booking information of a resort hotel and a city hotel between July 1st, 2015 and August 31st, 2017, including bookings that effectively arrived and bookings that were cancelled. Some data variables include whether the reservation was cancelled, length of study, the citizenship of the customers, lead time, number of adults and children, among others. The data set is an ideal resource to solve our questions of interest because it not only allows us to analyze the major components influencing a hotel's revenue, such as price, demand, and cancellation rates, but also suitable for building prediction models.

3 Data

This data set, called ”Hotel Booking Demand”, is downloaded from Kaggle and it has a license. This data is originally from the article ”Hotel Booking Demand Datasets”, written by Nuno Antonio, Ana Almeida, and Luis Nunes, February 2019. [2]There are 119390 observations in total, which includes hotel type(a city hotel or a resort hotel in Portugal), when the booking was made, length of stay, the number of adults, children, and babies, and the number of available parking spaces, among other things. From our discussion, the listed variables may not be over-represented, but we think there are under-represented variables. For example, it would be better to include how many rooms are available in both hotels, since we are clear about the hotels' demand and it will be more comprehensible to know the hotels' supply. Besides these ethical considerations, we know that all personally identifying information has been removed from the data set posted on Kaggle. All in all, there is no big ethical issue in this data set.

We observe that ”company” column has nearly 94 missing values, which are the names of the hotels. it is hard for us to fill the ”company” column, since the information is missing too much. So, we decide to drop the company column. Also, the ”agent” column has 13.68% missing values. This percent is quite large, so we do not decide to drop this column since these missing values may have important information. In particular, there are 333 unique agent, which are too many to be predictable. The NA values can be the agents that are not listed in present 333 agents. Also, we have only 4 missing values in our ”children” column, thus we decide to fill NA as 0, which means these agents may not have children. Also, we create a new column called ”length of stays” which is the sum of days guests spend during weekends and the days guests spend during weekdays. We did not transform categorical variables in the data set, since the two-value categorical variables have already been transformed into binary variables.

4 Methods

First, we do EDA and plot the scree plot. When using PCA, we omit the categorical variables such as reservation status and deposit type, since the challenge with categorical variables is to find an appropriate way to represent distances between variable categories and individuals in the factorial space.[1] Then, we draw the scree plot. We see the first PC can explain nearly 80 percent of the variation. Actually in this scenario, the principal components are less interpretable and do not have real meaning since they are constructed as linear combination of the initial variables. Also, since we create a new column ”length of stays”, we use box-plot to examine the outliers for our further analysis of this new variable. We observe that most values are below 16, which means most guests choose to stay at both hotels for less than 16 days.

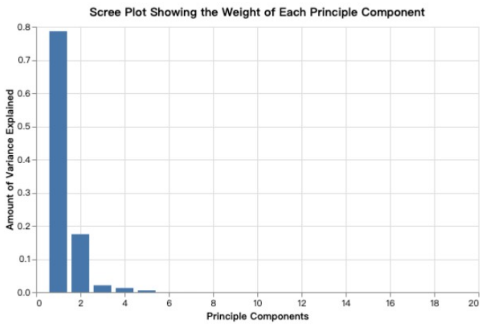

Figure 1: PCA

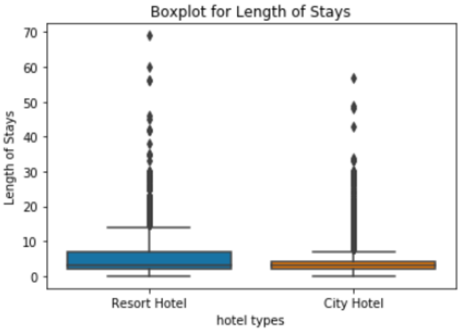

Figure 2: Boxplot for Length of Stays

We also want to explore some categorical variables. In figure 3, We analyze the reservation status for city and resort hotels. Most people are in check out status, this follows our intuition. Also, there are very few people in no-show status. Generally, the number of guests choosing city hotels is larger than those choosing resort hotel. In figure 4, shockingly, guests who choose non-refundable have nearly 99 percent chance to cancel their bookings. To explore this, we calculated the number of guests with non-refundable deposit type, and found only 12.22 percent of guests choose the non-refundable type. We may assume that the reason that the huge cancellation rate for non-refundable deposit type is because the numerator or the number of guests choosing non-refundable type is very small. In our first question, we applied line plots showing the prices and demand over time for both hotels, which allows us to clearly see the changes and patterns as well as comparing two hotels. For our second question, we employed a bar chart to display the monthly cancellation rates. Bar chart is suitable for this case as it not only tracks the changes of rate over months but also easily

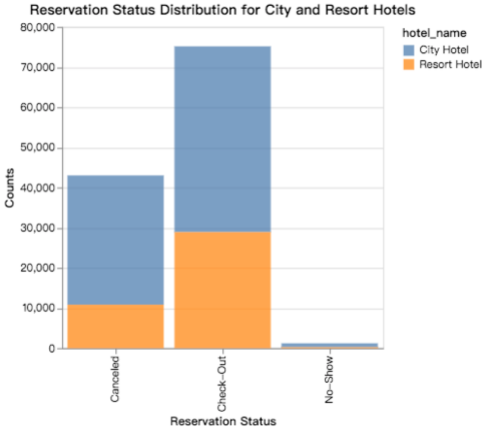

Figure 3: Reservation Status Distribution

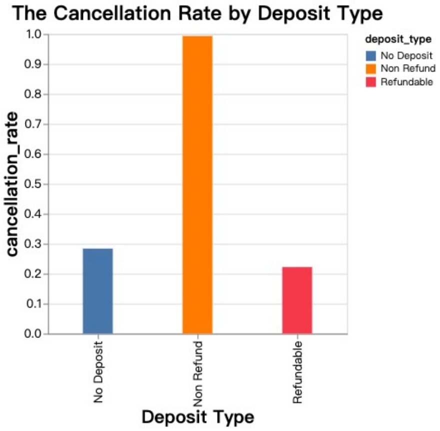

Figure 4: Deposit Type

compares between hotels. We also used transform regression to analyze the relationship between lead time and cancellation rates as it enables us to see how far the points are from the regression line. In our next question, we created a pie chart to explore the cancellations on different channels so that we could clearly observe how the cancellations are distributed by the size in each segment. Finally, we built a multi-linear regression line to predict the average revenue earned per room per day.

5 Results

5.1 Question of Interest 1: How Does Price and Demand Change over time?

Overall, the booking demand for city hotel is greater than the demand for resort hotel. Visitors tend to book more rooms of city hotel than of resort hotel in Portugal. Also, the highest demand for both hotels is at August, which indicates August is the most popular month for visitors. Besides August, April to June are also high-demand months for city hotel. December and January are the least popular months, we can see the demand for both two hotels during these two months are comparatively low. Generally, based on the demand curve for these two hotels, we can see that people would like to visit Portugal during summer and are not interested in visiting during winter. Interestingly, from August to September, the demand for both hotels decreases immediately. This may be because the summer vacation ends and Fall may not be a fun time in Portugal to attract visitors.

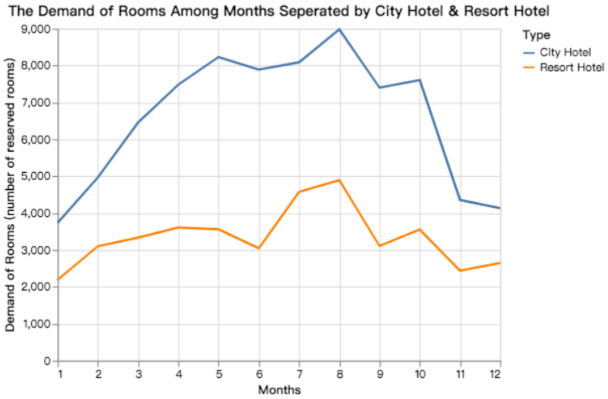

Figure 5: Demand By Month

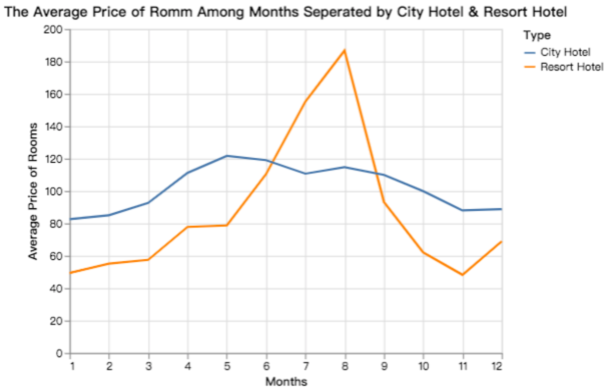

Figure 6: Price By Month

The price of city hotel changes smoothly over the months while the price of resort hotel changes very dramatically. Especially for the month August, the resort hotel price reaches peak 190e per night. Also this month is the highest demand month for both hotels. Interestingly, whatever the demand for city hotel changes over 12 months, the average price of city hotel does not exceed from 80e to 120e, which indicates that city hotel's price change may not depend too much on months. Instead, resort hotel located in some scenic spots of Portugal raises price very higher during popular months such as July and August.

5.2 Question of Interest 2: How much is the Cancellation Rate and How Can It Affect the Revenue?

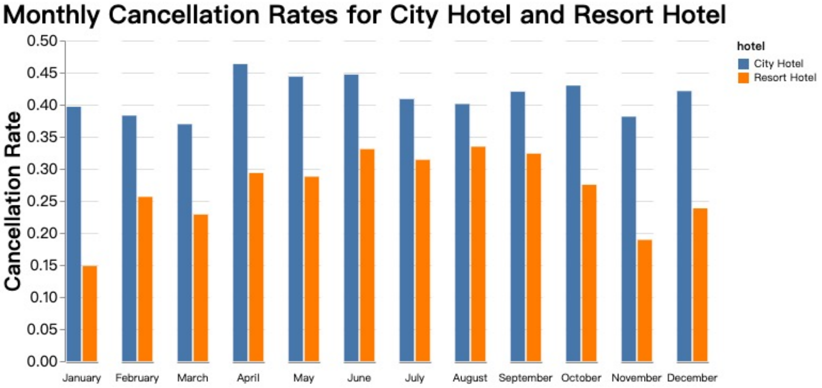

Figure 7: Cancellation Rate By Month

Cancellations can have a direct impact on the hotel's actual demand and thus its revenue. In Figure 3, there is noticeable difference between the monthly cancellation rates of two hotels: The cancellation rate of City Hotel is generally higher than that of Resort Hotel, fluctuating around 40% over the months. The cancellation rates of Resort Hotel vary a lot among months. The cancellation rates are higher than 30% during the summer but tend to be lower than 25% during the winter season.

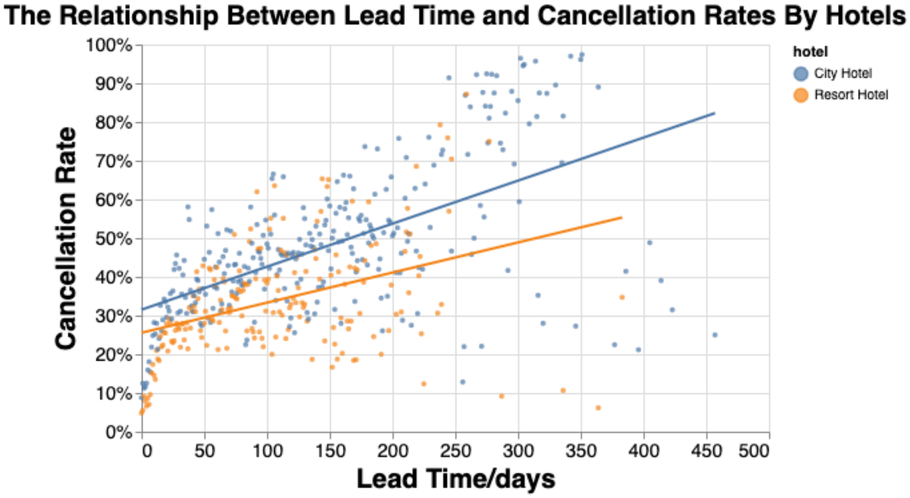

Figure 8: Cancellation Rate VS Lead Time

To have a deeper insight into what might affect the cancellation rates, we assumed that there might be an association between the lead time and the guests' cancellations. As we can see from Figure 8, there is a relationship between the lead time and the cancellation rate. The cancellation rate tends to increase when the lead time is longer for both hotels, which makes sense because unexpected things that may prevent customers from traveling are more likely to occur when the guests make the booking much in advance. Interestingly, City Hotel has a higher cancellation rate than Resort Hotel when the lead time is the same, implying that City Hotel's customers are more likely to cancel their booking reservations. However, we also notice that there are many outliers which are far from the linear regression lines, especially when the lead time is larger than 360. This is because few people make reservations almost 1 year before the travel so lead time greater than 360 has relatively smaller frequencies. As we know, outliers can be misleading because they tend to overestimate or underestimate the true cancellation rates and this is why we want to exclude them here.

5.3 Question of Interest 3: What Are the Patterns of Consumer Behaviors Look Like?

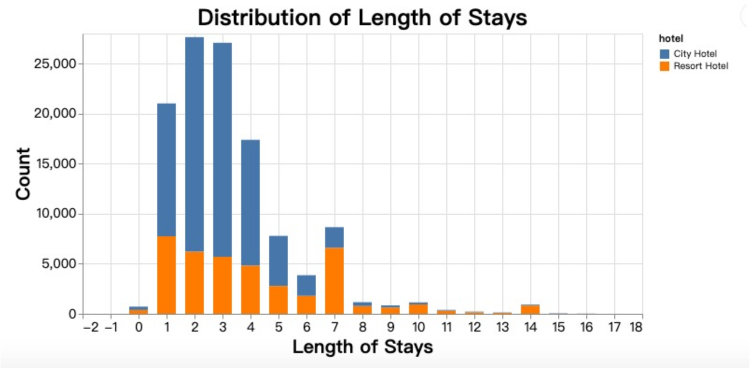

Figure 9: Distribution of Length of Stays for Two Hotels

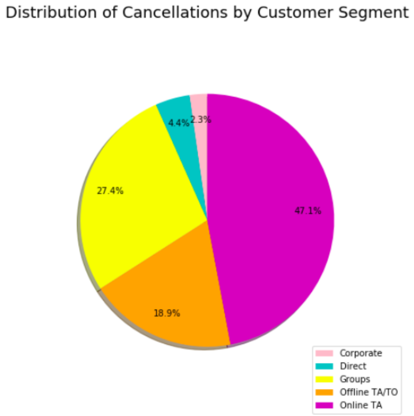

Figure 10: Cancellation for Market Segments

For this question, first we would like to know how long customers usually stay in both hotels. In Figure 9, we exclude the length of stay which are greater than 16, since the counts for length above 16 are relatively small and thus can be considered as outliers. (This is also demonstrated in the Exploratory Data Analysis). We are more interested in the length of stay that most people would like to spend. From Figure 9, people tend to stay longer in Resort Hotel. While the reservations in City Hotel for fewer than 5 days are significantly greater than those in Resort Hotel, the situation reverses for reservations longer than 1 week. More people choose to stay in Resort Hotel for longer vacation, especially for 7 days. This is because the resort hotel provides more privacy, entertainment and experiences, attracting customers to stay longer. For City Hotel, most people choose to stay for 2 to 3 days, and this may be due to the fact that people traveling for business purposes are inclined to choose City Hotel for lodging and their stay usually lasts for a short time.

For the next step, we explore the variability of cancellations among different customer segments. From Figure 10, the majority of the booking cancellations come from three market segments: Online TA (Travel Agents), Groups and Offline TA/TO. In particular, 47.1 percent of the cancellations come from Online TA segment, making up almost half of the total cancellations. The rates for Groups and Offline TA/TO sectors are 27.4 percent and 18.9 percent respectively. The high cancellations rates in segment Online TA is because online agents, such as hotel booking apps, tend to encourage guests to reserve a room even when there are still lots of uncertainties for their travel plans, with the message that they can take advantage of cancellation policies later. However, this cancellation rates can severely affect the hotel's demand forecasting as they distort the demand by inflating them to an artificial level. Consequently, this unrealistic demand might reduce the actual hotel fillings as well as the total revenue.

5.4 Question of Interest 4: What is the Final Model to Predict the Average Revenue Per Room Per Day?

First, we calculate the correlation values corresponding to different variables. The figure below shows the comparative correlations between revenue and different variables. The correlation between revenue and the number of children guests is 0.32 and the correlation between revenue and the arrival year is 0.20. Besides these, the number of adults, the number of special requests, and the categorical variable "are they repeated guests" also have strong correlation with the hotel revenue. Thus, we use these five variables to predict our revenue. R:Hotel Revenue, Y:Arrival Year, A :the Number Of Adults, C:the Number Of Children, RG:whether the guest is a repeated Guest(0 for no, 1 for yes), Req:the

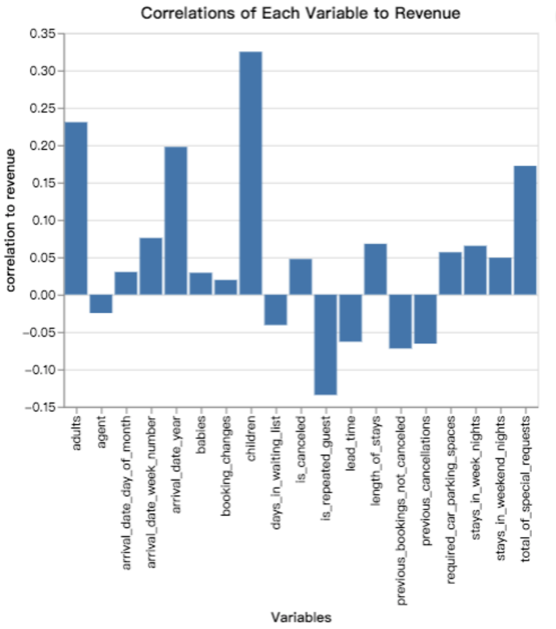

Figure 11: Correlations

Number Of Guest's Requests. After constructing
the model, we get our Root Mean Squared Error: $RMSE = 45.06$. Since we explore the cancellation(CL), lead time(LT), length of stays(LS) above, we decide to add these variables into our multi-linear model to see whether they diminish our root mean squared error. Our final model became:

$$R = 11.95Y + 17.01A + 37.06C - 30.46RG + 7.42Req - 0.05LT + 0.46LS + 8.59CL$$

We calculate our Root Mean Squared Error again, and found $RMSE = 44.67$, the loss has been diminished by 0.85%.

6 Conclusions and Future Work

Overall, this project makes some significant contributions to achieve the original goal of analyzing the hotels' market performance and helping develop better revenue management strategies. Firstly, it discovers the similarities and difference between two hotels in terms of price trend, seasonality and cancellation rates. The price for City Hotel is generally higher than that for Resort Hotel during the two years, except for the summer when the price for Resort Hotel suddenly boomed. Overall, the change in demand shares the similar pattern with summer being the peak travel season. In addition, it is also found that both hotels' demand have been inflated by the high cancellation rates. Hence, in order to improve revenue, more strategic cancellation policies on channels with large amount of cancellations should be implemented. Last but not least, this work constructs a revenue regression model for hotel managers to refer to and have a better understanding of the association of total revenue among other variables.

However, as with all the analyses, this work presents some limitations. The first stems from the depth of the exploration. Although the analysis shows the difference in market performance between two different types of hotels, it didn't provide explanations about the reasons behind this variance, which can be critical for hotel managers to determine the revenue management tactics. Secondly, this work utilizes multi-linear regression model to make predictions so only numerical variables are included into the training. However, categorical values, such as room types and distribution channels, can also affect the revenue earned for each transaction. Therefore, future research should also address the problem of including categorical values into the predicting process as well as using additional supporting datasets to study the reasons that can explain the difference between two hotels in demand and cancellation aspects in order to achieve the ultimate goal of helping develop more desirable revenue management strategies.

References

[1] Zakaria Jaadi. A Step by Step Explanation of Principal Component Analysis. *Builtin*, 4th September, 2019.

[2] Ana Almeida Nuno Antonio and Luis Nunes. Hotel Booking Demand Datasets. *Data in Brief*, 22:41–49, Febuary, 2000.

[3] K. T. Talluri and G. J. van Ryzin. *The Theory and Practice of Revenue Management.* Springer, Boston, MA, 2004.