Hotel Booking Demand

Introduction

Hotels are one of the important hospitality industries, which has increased its demand in the past years. The demand for hotels impacts its decisions on pricing, reliability, and availability. However, it is easy to book a reliable hotel if we understand the process of demand trends, which helps to be economical and secure. Even for the employees, accurate forecasting of booking demand helps to improve their service to the customers and to increase their revenue. Hotel booking demand is based on various factors like seasonal trends, number of days staying, weekends, and weekdays. Many hotels face the issue of understanding and forecasting the demand for bookings. This project helps to examine the key components and their effects on hotel booking demand, even though it's a challenging task. To address this challenge, this research aims to investigate the factors that drive hotel booking demand and develop a predictive model to forecast future demand. With this project, we can get to know the patterns and trends which provide better insights for customers and hotel managers.

Related work

Many works have been done on this problem and gave different solutions for this. But the very solution that stood out from others is that, given the intensifying competition among hotels, enhancing hotel management and putting information construction into practice is definitely a smart move. Mr. Wei Wei and Zhengwei Lou have done extensive research about this and published a paper called 'Design and Implementation of Hotel Room Demand Management System' (2019). The main inspiration for them to work on this project is the unclear demands of hotel room bookings in different seasons and how there are no standard operatives in the industry or trend to predict it. They have used administrator attributes, room information, booking information, and statement information as key feature tables and implemented a relationship between them. Overall, they created the design and implemented the hotel room management system by analyzing market demands.

Method

For this project, I have used a dataset from Kaggle (Hotel booking demand) by Jesse Mostipak. However, the data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. This dataset contains various records in a city, namely hotels, which explains the type of hotel, like resort or city hotel, lead time (the number of days between booking and arrival date), ADR (Average Daily Rate), and market segment (the segment from which the booking was made, such as Online Travel Agents (OTA) or Direct).

Prior to analysis, data cleaning and preprocessing—which involved several procedures—were essential. First, missing values were resolved; for example, the mode was assigned to categorical variables like "country," and the median values were imputed to numerical attributes like "children." Unstandardized column entries and mixed date formats were among the data format anomalies in the dataset that were fixed to guarantee consistency.

In order to assure meaningful research, it was imperative that the dataset be segmented by pertinent criteria, such as market sectors, hotel types, and client types. This entailed removing records that contained inaccurate or missing information, such as entries with undefined client groups or negative lead times. Data integrity was ensured by reducing the dataset from its initial size of 120,000 records to 119,390 rows following cleaning.

The cleaning and transformation processes were performed using Python libraries such as Pandas and NumPy. For instance, outlier detection and removal were automated using z-score thresholds, ensuring statistical anomalies did not skew results. Aggregations such as mean, median, and mode were used to summarize customer demographics and booking trends.

The dataset was ready for an exploratory analysis that sought to identify patterns and practical insights in hotel booking behaviors after these pretreatment and analysis processes were finished. The foundation for a thorough comprehension of the dataset and its possible uses was established by these discoveries and visualizations.

Exploratory Data Analysis (EDA)

Using the data source option, the hotel booking demand .csv file has been uploaded. All the data in the file has been loaded and be sectorized by different columns. From the overview we can observe the total records present in the dataset is huge with 119,390 records which makes difficult to analyze as the huge data size tends to more outliers. In this data we can observe that missing data is present in some columns like children, agent, company. We can overcome these errors by using data cleaning i.e. keeping '0' in the null value to remove any future complications. Some outliers can be observed in 'adr' column with lowest being '-6.38$' nd highest being '54000$' which are not possible in real life.

We can do univariate analysis on some of attributes in the dataset, for example using Total children we can say that customers with 0 children are highly probable to book the room compared to others.

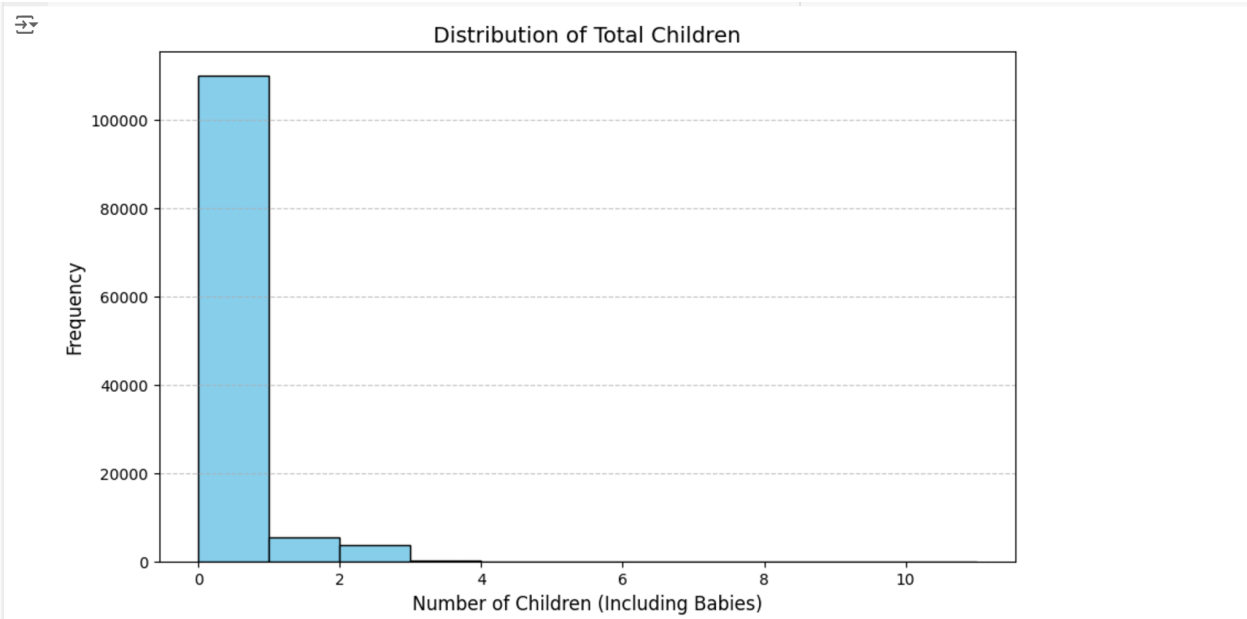


Fig 1: explains the univariate analysis example of total children and their booking pattern

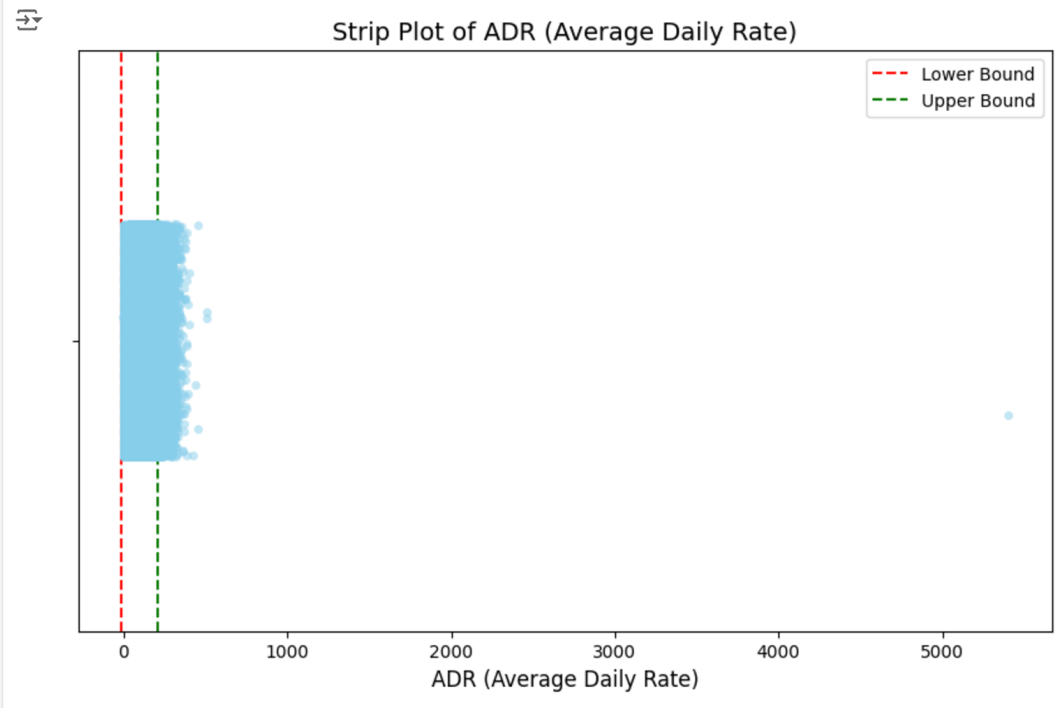


Fig 2: Strip plot to show the outliers of the ADR (average daily rate)

Using pandas' data cleaning has been done which removed the outliers in 'adr' column. The records with anomality have been reduced from 119390 to 115597 records, thus removing 3793 records.

Creation of two new columns has been done for hypothesis purpose namely total_children, total_nights. Total_children was created using 'children' and 'babies' Columns and total_nights was created using stays_in_week_nights and stays_in-weekend_nights columns. These are further explained in the hypothesis which helps to answer it sufficiently.

Hypotheses

1.Families traveling with children are more likely to extend their hotel stays, suggesting that the presence of children transforms a routine trip into a longer, leisure-focused experience

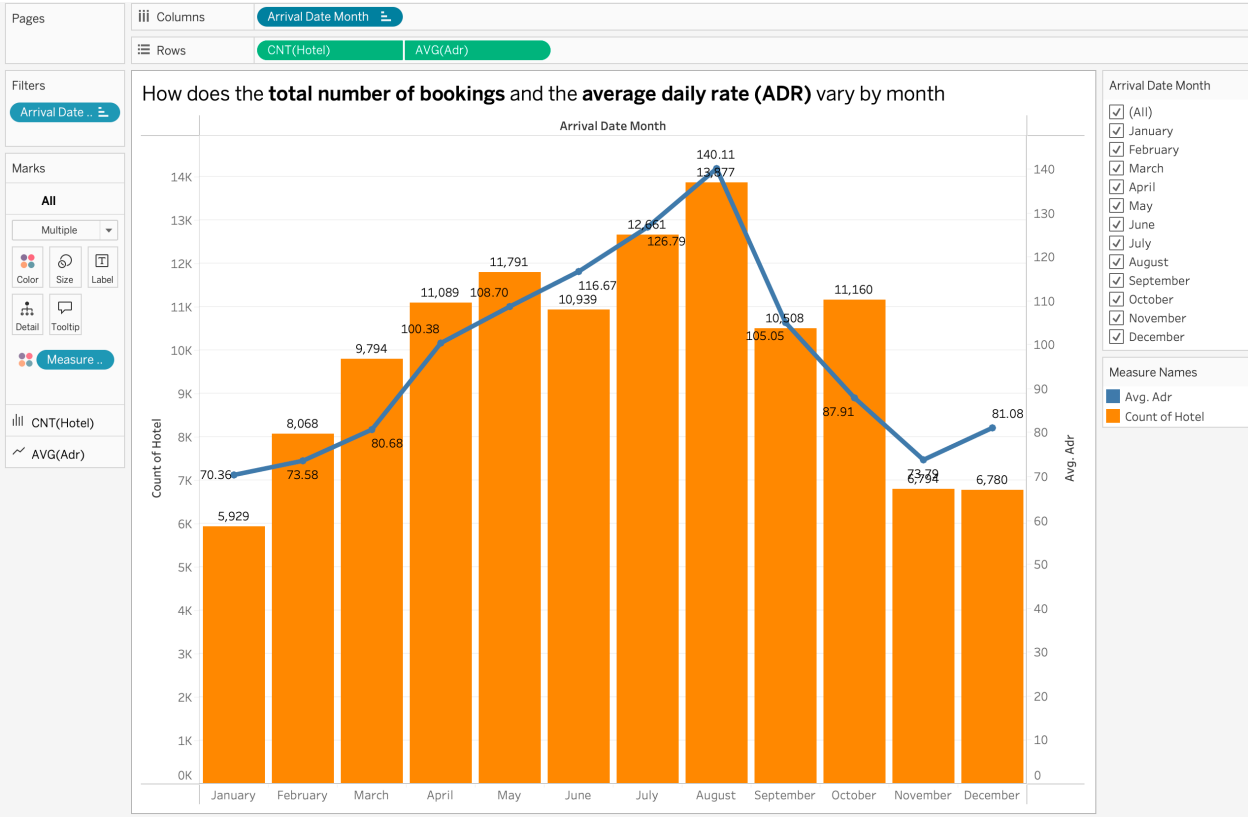
Result

For verifying my hypothesis, I have used tableau to present the visualization and python to generate it in Colab interface. I have used scatter plot for this hypothesis as it gives a good overlook for two numerical variables while also giving other dimensions information. The size of circles in the graph describes average daily rent and each color represent the type of costumer like blue being the contract type, orange for group customers. On the X- axis the total children are taken, and total nights stayed are on Y-axis.

Discussion

We can observe that Majority of stays are shorter from 1-10 nights. And most families with children tend to stay more days. Adr decreases as the number of children increases, this reflects the pricing strategy delivered to families. Transient customers (red circles) dominate this category, reflecting individuals or couples with brief stays. Customers with 3 kids or more are tend to stay long compared to others. We can see that the adr price decreases as the party size increases which suggest that there may be a discounting strategy. Our hypothesis is correct from the above visualization.

2. The total number of bookings peaks during summer months (June-August), while the Average Daily Rate (ADR) is higher during holiday seasons and winter months (December-February) due to increased demand for premium services.



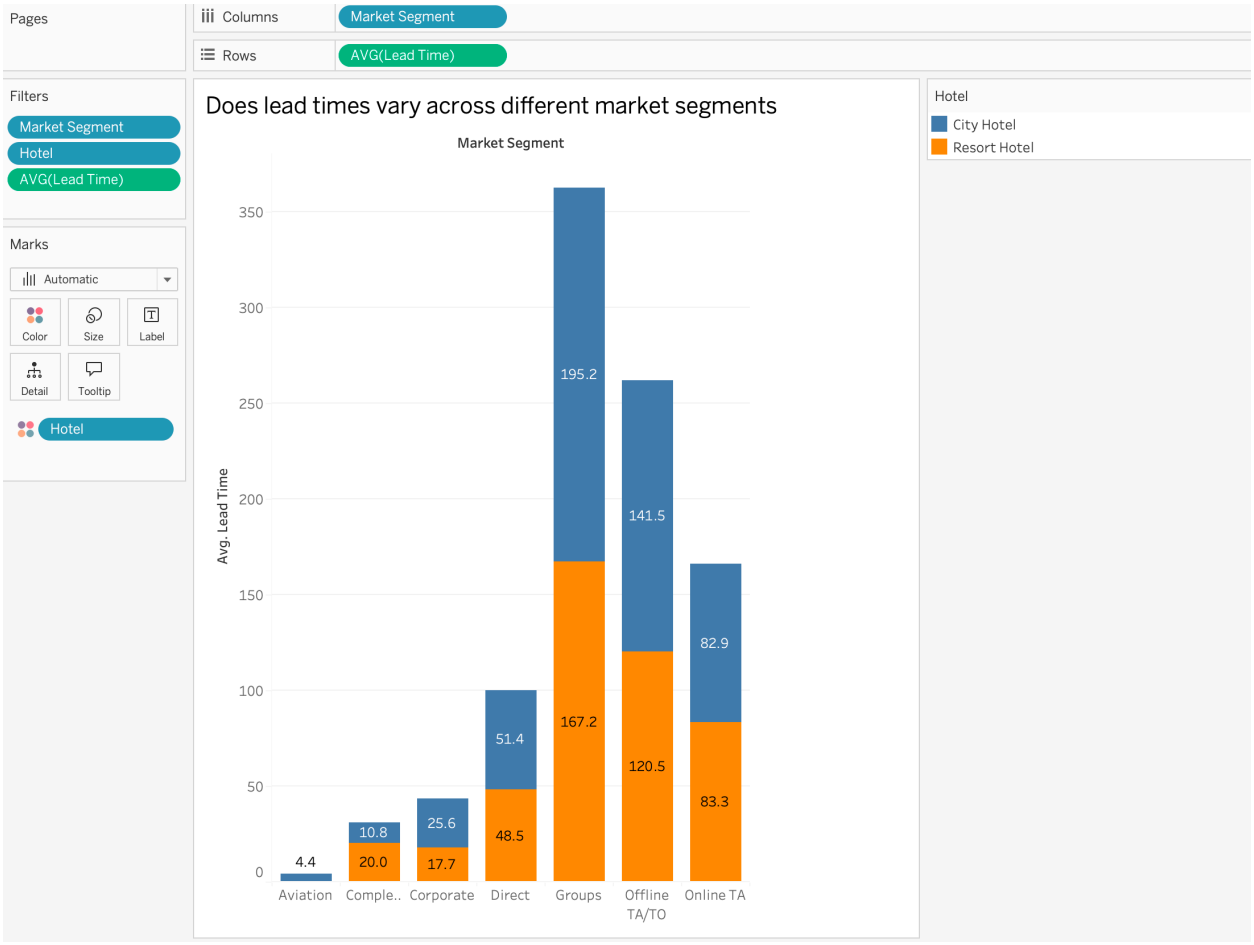
Result

It is a dual-axis combination graph, which helps us to compare two metric scales on the same timeline. I took the primary Y-axis (orange bars) for the total count of hotels and the secondary Y-Axis (blue line) to represent adr. On the X-axis, the month's arrival date can be seen. Based on the graph, we can say that the total bookings were lowest in January and increased gradually till August (5,929 to 13,877). Summer can be considered as the busiest month for hotel management throughout the year. Winter is considered as the lowest time happening for hotels. The average daily rate is highest in August (140.11) and lowest in January(70.36). Adr has been increased from January to August. Based on seasons, the summer brings higher bookings and higher adr.

Discussion

Our hypothesis that bookings will peak in summer has been proved and the adr is higher during this season is also confirmed. We can say that high adr in August reflects hotels leveraging the summer holiday season to maximize revenue through demand-driven pricing. Early summer bookings are suggested by the consistent increase in bookings from March to May and the rising Adr. Some useful suggestions might be offering early booking discounts in March and might help the customers during their summer stay. Having a seasonal package in winter will help the hotels to improve their bookings in the offseason.

3. Lead times vary significantly across different market segments, with group bookings and offline travel agencies showing longer lead times compared to direct and online travel agent bookings.



Result

Based on the above visualization, we can observe that the market segment with groups has the longest lead time of 195.2 days, showing that advance planning for group reservations is important. And we can see that offline travel agents or travel operators also have high lead times, with 141.5 days for city hotels and 120.5 days for resort hotels. In general, direct bookings have less lead time of 51.4 days for cities and 48.5 days for resorts, which shows people are more spontaneous in booking in this category. However, aviation shows the lowest lead time with 4.4 days among all other market segments.

Discussion

The hypothesis is correct, as we can observe that lead time varies across different market segments. The long lead time indicates that the customers are properly planned and way ahead of arrival day. Group and offline TA's have a long lead time compared to others, maybe because of package deals, especially for city hotels. Meanwhile, online TA has a short lead time, which suggests that those customers who are much more flexible tend to do booking at the last minute. When we see it from the point of view of the type of hotel, city hotels have higher lead times. From the above visualization, hotels should focus on early marketing and promotions to secure these bookings well in advance.

Future work

The insights from the project help to understand the relationship between customer behavior and hotel room demand management. We can extend the project in various advanced ways and applications. Machine learning helps to do the predictive analysis that helps to predict key factors like adr and how a new relationship can be formed by discount based on their past data. This helps to regulate the prices and staff burden in seasonal time. The data set that has been used only contains 32 columns/attributes, which sometimes restricts the big view. Adding extra attributes like customer reviews and local holidays will help to improve the efficiency. Dynamic pricing plays a key role in hotel booking demand. Advanced pricing regulation will help to maximize the revenue for hotels.

References

The data is originally from the article **Hotel Booking Demand Datasets**, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.

The data was downloaded and cleaned by Thomas Mock and Antoine Bichat for #TidyTuesday during the week of February 11th, 2020.

Latest Data set was derived from https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data