**EDA – Individual Dataset**

Amma A. Boa-Amponsem

College of Professional Studies, Northeastern University

ALY 6040: Data Mining Applications

Professor Hema Seshadri

April 19, 2024

**EDA – Individual Dataset**

**Introduction**

The hotel business is a dynamic and highly competitive sector which thrives on optimizing operations, understanding customer behavior and demand, and to maximize profit. In this report, we will explore and obtain insights in booking patterns and trends, generating revenue, and customer demographics using a hotel booking demand dataset, containing valuable details including hotel types, guest profiles, lead times, bookings, and cancellations.

In this analysis, the main objective is identifying key attributes or factors that impact hotel booking cancellations and to recommend strategies to optimize booking management and enhance customer satisfaction. Specifically, our objective was divided into three (3) key research questions:

1. What is the overall cancellation rate of bookings varying between resort hotels and city hotels?
2. Are there any seasonal patterns in the monthly rates of hotel occupancy?
3. Is there a relationship between the distribution of lead times and booking cancellations?

The concept is to understand the determining factors that could contribute to cancellations in hotel bookings which is vital for hotel managers to make strategic decisions in minimizing cancellations while optimizing profit and customer satisfaction.

By conducting a thorough exploratory data analysis on the dataset, we aim to explore the relationships between the dependent and predictor variables illustrated by visualizations that could give actionable insights for further analysis and recommendations.

**Analysis**

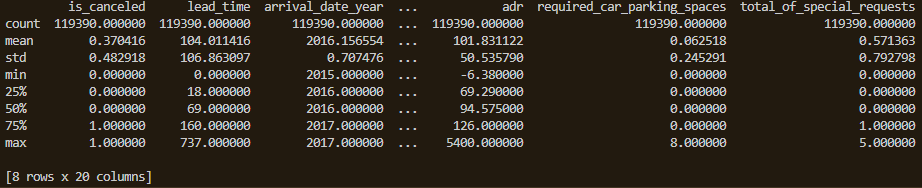
**Descriptive Statistics**

There are 119,390 entries (rows) and 32 variables that includes categorical, numerical, and binary values within the dataset, recording booking dates from 2015 to 2017. To explore the hotel booking demand dataset, we implemented the python application to create codes and run our outcomes. We loaded the data into a pandas Data Frame, named it as ‘hotel\_bookings\_df’, and explored the structure of the data by displaying the first few rows to obtain an initial understanding of its content. We proceeded to perform descriptive statistics to understand the distributions and central tendencies of the numerical variables.

A screen shot of a computer code

Description automatically generated

Samples from the results show that the ‘lead\_time’ ranges from 0 to 737 days, with a mean lead time of approximately 104 days, whereas the average daily rate ‘adr’ ranged from -6.38 to 5400 which seems abnormally high, with a mean ADR estimate of 101.83. Other variables such as the ‘required\_car\_parking\_spaces’ had a maximum value of 8 which could potentially be an outlier or a valid entry for large groups, and the ‘total\_of\_special\_requests’ variable had a maximum value of 5, which could be a valid entry for multiple special requests by customers or potential outliers.



We further performed a data cleaning process using the ‘numpy’ library by identifying and handling missing values in the dataset, replacing them with their appropriate measures, that is, modes for categorical columns and means for numerical columns. There were no duplications found in the provided code snippet.

A screen shot of a computer code

Description automatically generated

**Exploratory Data Analysis (EDA)**

After cleaning the dataset, we continued to the main part of our analysis which was perform an exploratory data analysis. We explored the three (3) key objectives in this report to uncover trends or patterns and relationships within the dataset. We will breakdown their analyses showing our findings and illustrations using visualization techniques. These visualization tools utilized for each research question are:

1. What is the overall cancellation rate of bookings varying between resort hotels and city hotels?

* Bar graph was utilized by applying ‘matplotlib’ and ‘seaborn’ libraries to compare the overall cancellation rate between both hotels.

1. Are there any seasonal patterns in the monthly rates of hotel occupancy?

* Line graph was employed by using ‘matplotlib’ and ‘seaborn’ libraries to illustrate the number of bookings or rates of occupancy against different months of the specified years.

1. Is there a relationship between the distribution of lead times and booking cancellations?

* Box plot was used to visualize the comparison of lead time distributions between canceled and non-canceled bookings. We employed the ‘matplotlib’ and ‘seaborn’ libraries to generate our codes.

**Cancellation Rate by Hotel Type**

For our first objective, we analyzed the ‘is\_canceled’ and ‘hotel’ variables to compute the overall cancellation rate for all bookings and compared it between resort hotels and city hotels, providing insights into booking patterns and possible differences in cancellation performance between both hotel types. Resort hotels are usually located in leisure or scenic destinations including tourist attractions sites, mountains, or beachfronts, often found in remote and secluded areas away from urban places. On the other hand, city hotels are situated in urban areas or cities and are conveniently found near business districts, residential places, transportation hubs and tourist attractions.

A computer code with colorful text

Description automatically generated

**Observations/Interpretations**

Based on our findings in *Figure 1*, we observed that the overall cancellation rates for city hotels are lower in comparison with bookings that were not canceled. Similarly, the number of bookings canceled are drastically lower than those that were not. Although, varying the cancellation rates between both hotel types, we can see the number of bookings canceled at city hotels were substantially higher than that of resort hotels.

A graph of a number of bars

Description automatically generated

*Figure 1: Bar graph of Cancellation Rate by Hotel Type*

This observation indicates that a greater overall demand for accommodation at city hotels than resort hotels, potentially leading to a greater propensity for cancellations. The increased number of cancellations at city hotels could possibly be attributed to causes such as last-minute change in plans, bigger guest pools leading to likely higher cancellations, or dynamic nature of urban travels.

**Seasonal Booking Patterns**

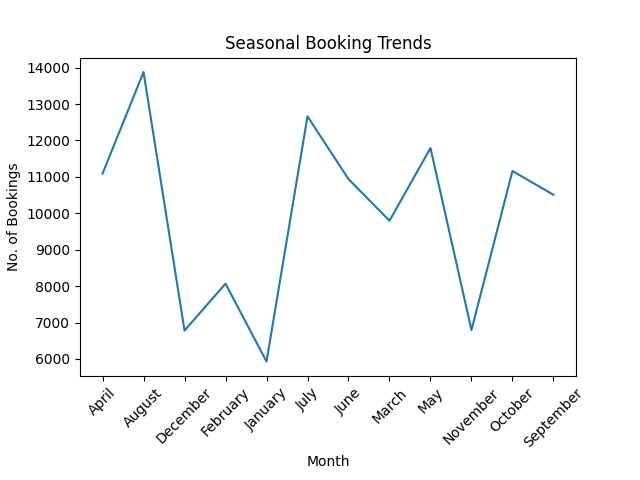
By analyzing the ‘arrival\_date\_month’ variable to identify booking trends and rates of occupancy over several months of the specified years. Seasonal patterns can be observed and analyzed in hotel demand and assist hotel managers in optimizing pricing and resource allocation for peak seasons.

A screen shot of a computer code

Description automatically generated

**Observations/Interpretations**

Our findings in *Figure 2* show distinct seasonal trends in booking activity where peak booking seasons were in August, followed closely by July and May. Conversely, the lowest booking season was in January, with a noticeable decline in the specified dates between 2015 and 2017. The peak seasons in May, July and August may possibly be popular travel seasons for summer vacations or holidays, while the lowest season in January could have potentially been due to inclement weather conditions in certain regions or countries.



*Figure 2: Line plot of Seasonal Booking Trends*

**Lead Times and Booking Cancellations**

We examined the distribution of ‘lead\_time’ variable, that is, the number of days between booking and arrival date, and compared it to the ‘is\_canceled’ variable to explore whether bookings that have longer lead times are more liable to be canceled than the shorter lead times. This will help us to understand the influence of lead time on customer behavior and booking stability.

A computer code on a black background

Description automatically generated

**Observations/Interpretations**

The analysis in *Figure 3* showcases a relationship between lead time distribution and cancellation rates. Results show that as lead times rise, the possibility of booking cancellations also rises, leading to higher rates in cancellation. On the other hand, bookings that were not canceled are monitored to be more dominant with shorter lead times, showing two possible outlier points at the top left-side of the boxplot. These findings could suggest that longer lead times may allow guests with enough flexibility to change their plans, leading to more likely cancellations. Conversely, bookings that are made closer to the arrival dates have shorter windows to change plans and are more likely to proceed as planned, resulting in fewer cancellation rates.

A diagram of a lead time distribution

Description automatically generated

*Figure 3: Boxplot of Lead Time Distribution by Cancellation Status*

**Conclusion**

In conclusion, our analyses on the hotel booking demand dataset intriguingly uncovers different crucial insights that can inform hotel managers to make strategic decisions in hotel management:

1. City hotels have higher overall rates of cancellation compared to resort hotels, suggesting a possible increase in propensity for cancellations in cities or urban areas. Although both hotel types face cancellations, the greater number of booking cancellations at city hotels indicate a need to strategize to minimize cancellations while optimizing revenue.
2. Patterns or trends in booking differ throughout the periodic dates, showing peak seasons in August, July, and May with the lowest booking season in January. Understanding these seasonal patterns is vital for effectual pricing strategies, resource allocation, and marketing efforts specifically during inclement weather conditions to maximize occupancy rates and revenue.
3. There is a noticeable correlation between lead time and booking cancellations, as longer lead time distributions correlate with higher cancellation rates. Understanding this is essential for implementing flexible policies relating to booking cancellation.

**Proposed Next Steps and Recommendations**

Some further proposed steps and recommendations were made based on these findings for further investigation on the dataset and making informed decisions in hotel management:

1. To conduct a more comprehensive review of the categorical variables to spot and handle some inconsistencies or invalid entries, investigate outliers and extreme values.
2. Create additional attributes such as overall length of stay, weather conditions, and nearness to facilities to advance data analysis and modeling.
3. Perform further in-depth analysis of other key variables or relationships within the dataset. Leverage the cleaned dataset to build predictive models for forecast demand, improving customer satisfaction, and optimizing pricing.

A few recommendations can be:

1. To offer guests the choice to adjust or cancel their bookings with lenient penalties, specifically for bookings that were made well in advance.
2. Modify pricing strategies to align with seasonal booking trends and lead time to maximize profits during peak periods and reduce losses due to increased cancellations in the lowest booking seasons.
3. Improve marketing strategies by creating target promotional campaigns and special incentives during off-peak seasons for more customer engagement to boost occupancy rates.
4. Enhance communication by keeping guests informed and updated regularly about policies on bookings, cancellations, and any other changes in availability to instil trust and reliability.

**References**

Antonio, N., Almeida, A., & Nunes, L. (2019). Hotel Booking Demand Datasets. Data in Brief, Volume 22. <https://www.sciencedirect.com/science/article/pii/S2352340918315191>

Mostipak, J. (2020). Hotel booking demand. Kaggle. <https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data>

Evans, J. R. (2012). Statistics, Data Analysis, and Decision Modeling 5th Edition. Pearson Publishing, ISBN: 9780132744287.

Bluman, A. (2017). Elementary Statistics 10th Edition. McGraw Hill ISBN 978-130-7494-327