Investigate Business Hotel using Data Visualization

Created by:

Nur Imam Masri

Email : nurimammasri.01@gmail.com
LinkedIn : linkedin.com/in/nurimammasri
Github : github.com/nurimammasri
Portfolio : bit.ly/ImamProjectPortfolio

I am innovative and able to work together in a team orientation. Have an understanding of the fields of **Data Science, Data Analyst, Machine Learning, and Artificial Intelligence** with 5 years of learning experience and a specialist certificate (**TensorFlow Developer Certificate**). Often performs the data mining process by analyzing data, representing data, preprocessing data, making predictions. Excellent in understanding business operations and various data analytics tools (**Python, R, SQL, Pandas, Numpy, Matplotlib, Seaborn, TensorFlow, Sklearn, BigQuery, etc.**) and dashboarding using **Google Data Studio, Power BI, and Tableau.**

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Business Understanding

Rakamin Academy

📌 Problem Statement:

Business performance management is a metric for determining overall business progress toward goals. Business performance related to sales/marketing effectiveness. This performance is determined by the company's ability to set a strategy in order to be able to sell products/services that can meet customer expectations.

In the hospitality business, of course, it is closely related to the customer. The more customers who order, the more income the company. Therefore, analyzing the behavior of customers in booking hotels is very important. In addition to analyzing customer behavior in booking hotels, to measure the success of a hotel business, we can see the level of order cancellations. If many customers cancel their orders, this will adversely affect the hotel's business performance.

The purpose of this project is analyzing customer behaviors in hotel business. Provide insights related to hotel business performance. This insight can be obtained by exploring data, such as analyzing how customers behave in ordering hotel tickets or looking for factors that influence the cancellation of hotel ticket bookings. Then present the insights obtained using visualization and data storytelling.

📌 Goals

The company wants to improve business performance by analyzing how customers behave in ordering hotel tickets or looking for factors that influence the cancellation of hotel ticket bookings

📌 Objectives

Making reporting insights obtained to provide insight related to hotel business performance by using visualization and data storytelling.

✅ Data Description ✅

Rakamin Academy

📌 **Data provide by Rakamin - hotel_bookings data.csv** 📌

Hotel Bookings Dataset (link datasets)

Dataset Description:
This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has been removed from the data.

This dataset contains 119,390 samples. Contains 29 features :

- **hotel** - Hotel (H1 = Resort Hotel or H2 = City Hotel)
  - City Hotel is a type of hotel located in urban centers and is typically found in large cities.
  - Resort Hotel is an accommodation type usually built on the coastline or at the foot of mountain hills, offering scenic natural views.
- **is_canceled** - Value indicating if the booking was canceled (1) or not (0)
- **lead_time** - Number of days that elapsed between the entering date of the booking into the PMS (Property Management System) and the arrival date
- **arrival_date_year** - Year of arrival date
- **arrival_date_month** - Month of arrival date
- **arrival_date_week_number** - Week number of year for arrival date
- **arrival_date_day_of_month** - Day of arrival date
- **stays_in_weekend_nights** - Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

✅ Data Description ✅

Rakamin Academy

- **adults** - Number of adults
- **children** - Number of children
- **babies** - Number of babies
- **meal** - Type of meal booked. Categories are presented in standard hospitality meal packages:
  - Undefined/SC – no meal package
  - BB – Bed & Breakfast
  - HB – Half board (breakfast and one other meal – usually dinner)
  - FB – Full board (breakfast, lunch and dinner)
- **city** - City name
- **market_segment** - Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
- **distribution_channel** - Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"
- **is_repeated_guest** - Value indicating if the booking name was from a repeated guest (1) or not (0)
- **previous_cancellations** - Number of previous bookings that were cancelled by the customer prior to the current booking
- **previous_bookings_not_canceled** - Number of previous bookings not cancelled by the customer prior to the current booking
- **booking_changes** - 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
- **deposit_type** - Indication on if the customer made a deposit to guarantee the booking. This variable can assume three:
  - 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

✅ Data Description ✅

Rakamin Academy

- **agent** - ID of the travel agency that made the booking
- **company** - ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
- **days_in_waiting_list** - Number of days the booking was in the waiting list before it was confirmed to the customer
    - customer_type - 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
- **adr**- Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
- **required_car_parking_spaces** - Number of car parking spaces required by the customer
- **total_of_special_requests** - Number of special requests made by the customer (e.g. twin bed or high floor)
- **reservation_status** - Reservation last status, assuming one of three categories:
    - Canceled – booking was canceled by the customer;
    - Check-Out – customer has checked in but already departed;
    - No-Show – customer did not check-in and did inform the hotel of the reason why

Tools

Rakamin Academy

- Open-source IDE : Jupyter Notebook

- Programming Languages : Python

- Data Preprocessing & Cleaning Library : Pandas & Numpy

- Data Visualization Library : Matplotlib & Seaborn

📌 Data Cleansing/Preprocessing 📌

Rakamin Academy

```
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 29 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_weekdays_nights        119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  city                            118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  booking_changes                 119390 non-null  int64
 20  deposit_type                    119390 non-null  object
 21  agent                           103050 non-null  float64
 22  company                         6797 non-null    float64
 23  days_in_waiting_list            119390 non-null  int64
 24  customer_type                   119390 non-null  object
 25  adr                             119390 non-null  float64
 26  required_car_parking_spaces     119390 non-null  int64
 27  total_of_special_requests       119390 non-null  int64
 28  reservation_status              119390 non-null  object
dtypes: float64(4), int64(16), object(9)
```

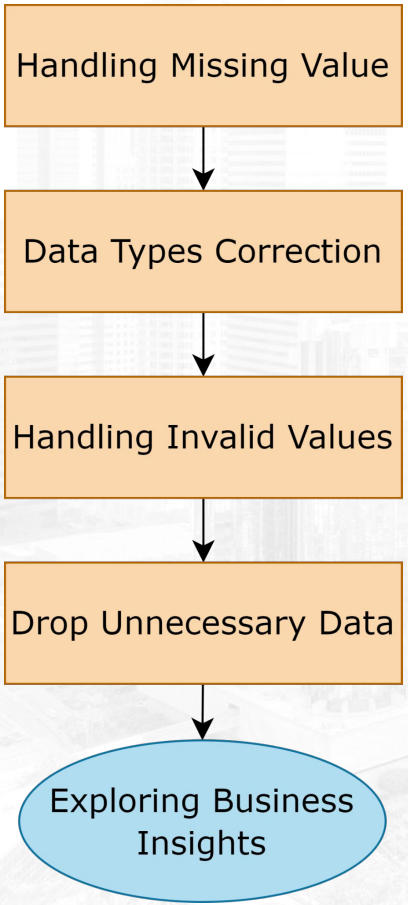
☕ Basic Datasets Information ☕ :

- The dataset consists of **29 columns** and **119,390 rows** of data.
- There are 3 types of data: **int64, object, float64.**
- There are some **missing values in the following columns**:
  - **company** with a total of 94% null values, amounting to 112,593 rows.
  - **agent** with a total of 13% null values, amounting to 16,340 rows.
  - **city** with a total of 0.4% null values, amounting to 488 rows.
  - **children** with a total of 0.003% null values, amounting to 4 rows.

For the detail of my codes, you can see here

Data Cleansing/Preprocessing

Rakamin Academy



Handling Duplicate Rows

- **This dataset has many duplicates, totaling 33,261 rows.** However, in this process, we will not drop the duplicate rows, assuming that it is due to the absence of customer IDs in the data

Handling Missing Value

- **company** with a total of **94% null values**, amounting to 112,593 rows.
  Filling zero values in the company column since no company is involved.
- **agent** with a total of **13% null values**, amounting to 16,340 rows.
  Filling zero values in the agent column since no agent is involved.
- **city** with a total of **0.4% null values**, amounting to 488 rows.
  Filling 'unknown' for unavailable city entries.
- **children** with a total of **0.003% null values**, amounting to 4 rows.
  Filling zero values for children, as it is likely that the customers have no children.

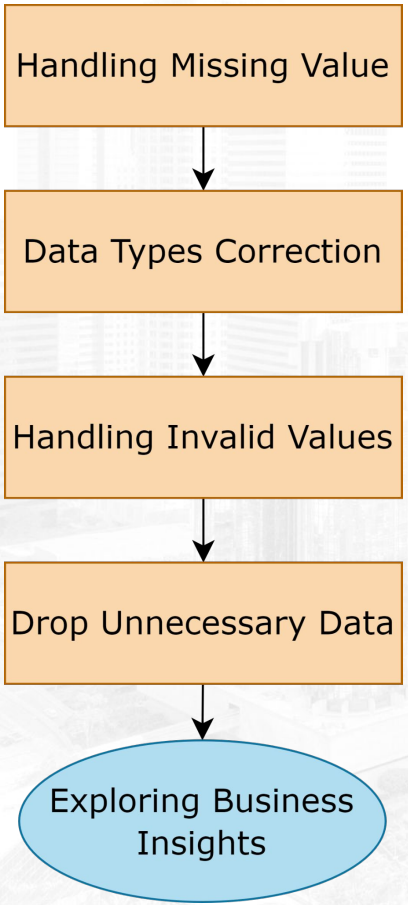
Data Types Correction

- Change the data type of **float64** which had null before, **children, agent, and company** to **int64**

For the detail of my codes, you can see here

📌 Data Cleansing/Preprocessing 📌

Rakamin Academy



📌 Handling Invalid Values

- Replacing incorrect values in meal column. These values will be replaced with **'No Meal' since it is assumed that 'Undefined'** indicates customers who did not order any meals

📌 Drop Unnecessary Data

- So, we will filter some things, namely:
  - **Total Guest (Number of Customers / Guests) <= 0**, or there are no guests at all.
  - **Total duration of the night <= 0**, or not available in the data for the duration of the stay.
  - **If there is a single data entry for adr (Average Daily Rate),** it might be due to a data calculation error. Since there is only one row, it will be dropped to avoid errors in the analysis.

For the detail of my codes, you can see here

📌 Monthly Hotel Booking Analysis Based on Hotel Type

In the hotel industry, customer behavior in hotel bookings plays a crucial role as it directly impacts the company's revenue. Analyzing customer behavior when booking hotels is essential. For instance, we can identify which types of hotels are most popular among customers and correlate this with the seasonal conditions when the hotels are booked.

Therefore, the goal of this task is to compare the number of hotel bookings each month based on the hotel types and analyze when there are increases or decreases in hotel bookings during different months or seasons.

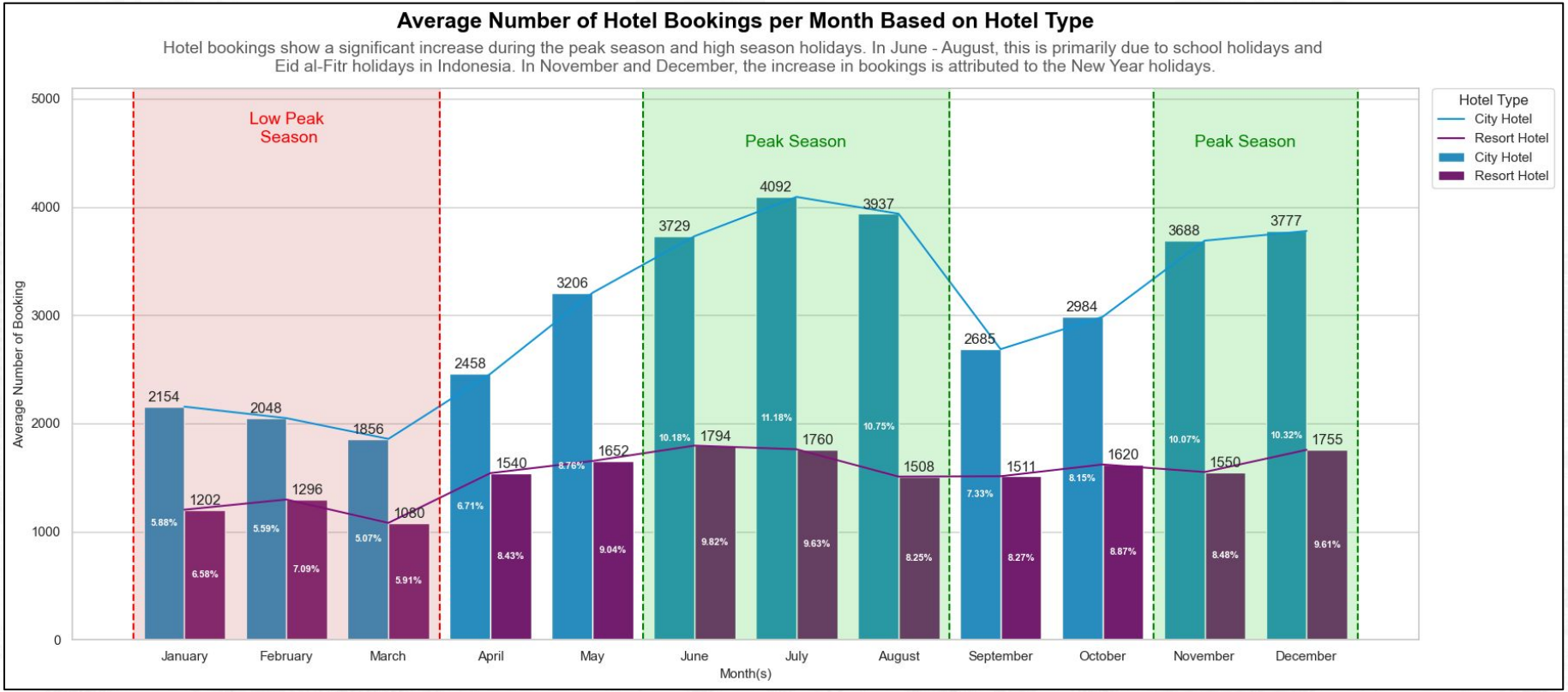
Steps taken

1. Create an aggregate table showing the comparison of the number of hotel bookings each month based on the hotel types. We will pay attention to the year of arrival data.
2. Normalize the data. We will pay attention to the data for September and October months.
3. Sort the data based on the months, paying attention to the correct spelling of the month names for easier visualization.
4. Create a line plot to show the changes in the increase or decrease in the number of hotel bookings each month based on the hotel types.
5. Interpretation

For the detail of my codes, you can see here

Monthly Hotel Booking Analysis Based on Hotel Type

Rakamin Academy



For the detail of my codes, you can see here

Monthly Hotel Booking Analysis Based on Hotel Type

Interpretation:

- Both hotels experience a similar increasing trend in the average number of hotel bookings. However, the highest peak is observed in the City Hotel.
- Hotel bookings show a significant increase during the peak season and high season holidays. The high peak season occurs in June - August for both City and Resort Hotels, mainly due to school holidays and the extended Eid al-Fitr holidays in Indonesia.
- Additionally, there is another high peak season in November and December, but it has a shorter duration compared to the peak season in June and July, likely due to the New Year holidays and the end of annual leave allotments. To optimize hotel room bookings during the low peak season, the hotel can implement New Year's promotions to attract more visitors.
- The Low Peak Season occurs throughout January - March and also in August - September, particularly in the City Hotel, where a significant decrease in bookings is observed. This decrease can be attributed to the start of new school and office seasons, as students and workers are focused on their academic and professional activities. During the low season, the hotel can offer discounts or vouchers to entice customers to continue visiting the establishment.

For the detail of my codes, you can see here

📌 Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates

In addition to analyzing customer behavior in booking hotels, the success of a hotel business can be measured by the booking cancellation rate. If many customers cancel their reservations, it can negatively impact the hotel's performance. Therefore, it is essential to explore the factors influencing booking cancellations. In this phase, we will investigate how the duration of stay can affect the hotel booking cancellation rate.

To analyze the correlation between the duration of stay and the hotel booking cancellation rate.

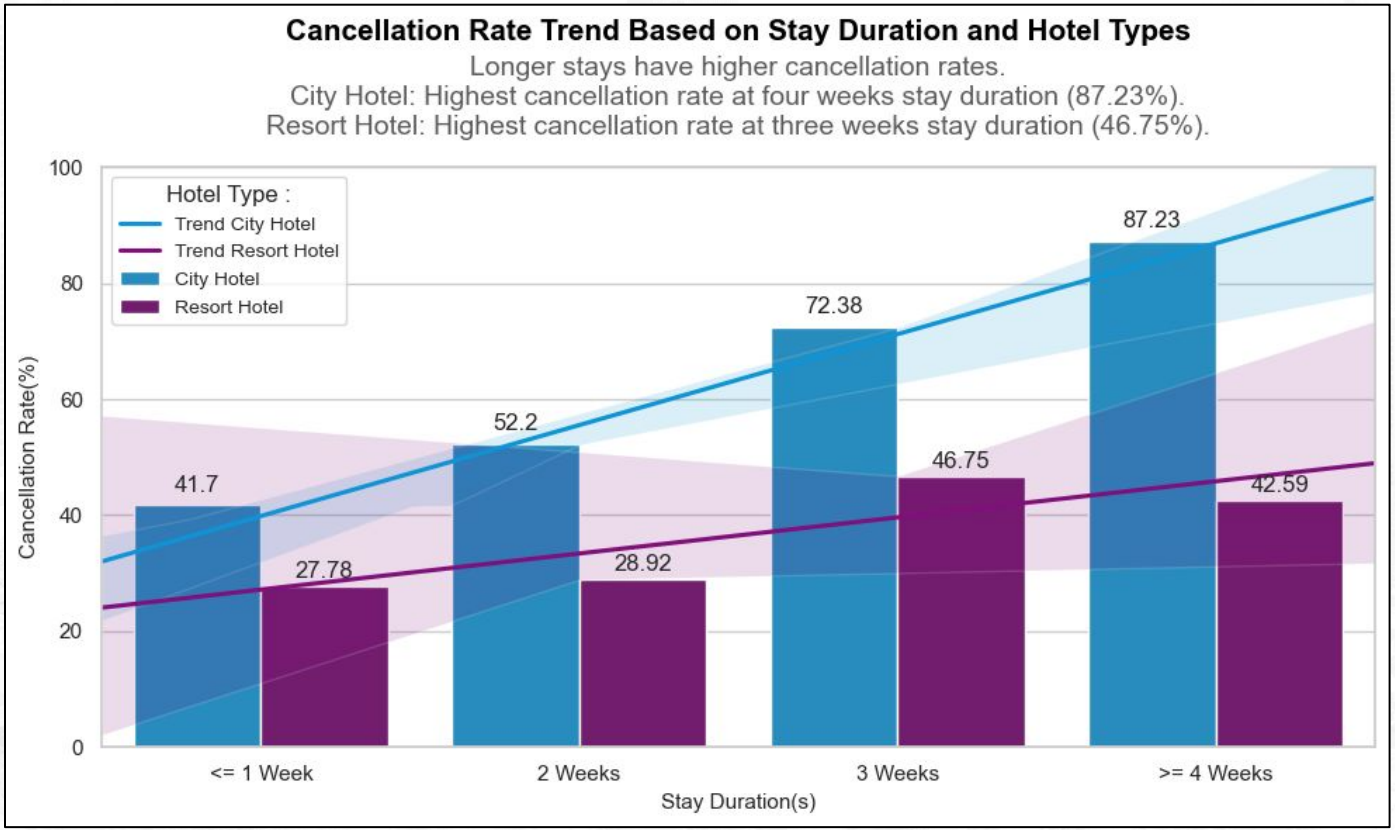
Steps taken

1. Pay attention to the "stay duration" column, which is obtained by summing the weekdays and weekend nights. Next, we will examine the data distribution to simplify the grouping process accordingly.
2. Group the values of the new column obtained in the previous step to make it more significant, taking into account the data distribution for meaningful categorization.
3. Create an aggregate table that compares the number of canceled hotel bookings based on the duration of stay for each hotel type, focusing on the proportion of canceled bookings.
4. Create a bar plot to display the cancellation ratio of bookings based on the duration of stay for each hotel type, emphasizing the proportion of canceled bookings.
5. Interpretation

For the detail of my codes, you can see here

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates

Rakamin Academy



For the detail of my codes, you can see here

📌 Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates

Rakamin Academy

Interpretation :

- The City Hotel experiences the highest number of cancellations with a significant increasing trend, while the Resort Hotel also shows an increasing trend, but not as steady.
- There is a positive correlation between Stay Duration and Cancellation Ratio, indicating that the longer customers stay, both in City and Resort Hotels, the higher the cancellation rate.
- The City Hotel's highest cancellation rate is observed for stays above 4 weeks, whereas for the Resort Hotel, the highest cancellation rate occurs for stays of 3 weeks.
- To address this issue, the hotel company can pay closer attention to the causes of booking cancellations and implement stricter cancellation policies. Additionally, offering special promotions may be effective in mitigating cancellations.

For the detail of my codes, you can see here

📌 Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate

The objective of this business insight exploration is to analyze the correlation between the lead time for hotel bookings and the rate of hotel booking cancellations. In the hotel industry, customers are usually allowed to book hotels before their arrival dates, and the lead time can vary from a few days to several months.

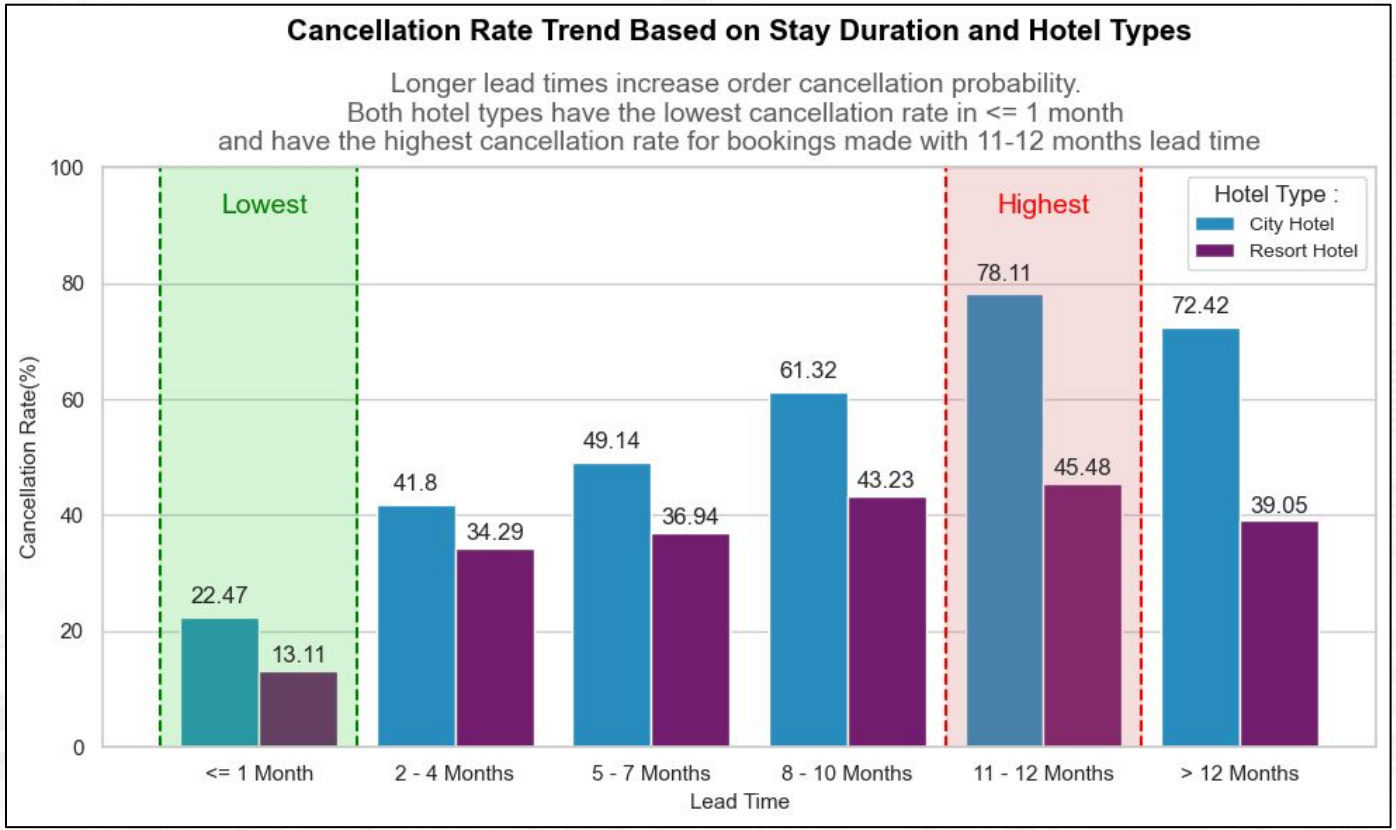
The task is to examine whether the lead time between the hotel booking and the arrival date influences the rate of hotel booking cancellations.

Steps taken

1. Create a new column that categorizes the booking lead time by creating intervals for the categorization.
2. Create an aggregated table comparing canceled hotel bookings based on the booking lead time for each hotel type, focusing on the proportion of canceled bookings.
3. Create a plot illustrating the cancellation ratio of bookings based on the booking lead time for each hotel type, using an appropriate plot type.
4. Interpretation

For the detail of my codes, you can see here

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate



For the detail of my codes, you can see here

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate

Rakamin Academy

Interpretation :

- In general, the longer the lead time, the higher the probability of order cancellation. Lead time refers to the number of days between the booking entry into the Property Management System (PMS) and the arrival date, where longer lead times are associated with higher cancellation rates.
- Both hotel types have the lowest cancellation rate for lead times of <= 1 month, with City Hotel at 22.47% and Resort Hotel at 13.11%.
- Both hotel types have the highest cancellation rate for bookings made with 11-12 months lead time, with City Hotel at 77.41% and Resort Hotel at 43.5%.
- Both Resort and City Hotels experience the highest cancellation rate around a 1-year lead time. This could be due to customers' vacation plans getting canceled or forgetting about their hotel reservation if the lead time is too long. Hotels can provide reminders to customers to reduce cancellations and implement a strict cancellation policy for every reservation to prevent such occurrences.

For the detail of my codes, you can see here

Summary

- Both hotels show a similar increasing trend in bookings, with the City Hotel experiencing the highest peak. Hotel bookings significantly increase during peak and high season holidays, particularly in June - August and November - December. The City Hotel also observes a significant decline in bookings during January - March and August - September. To optimize bookings during low peak seasons, implementing promotions and discounts can be effective strategies for both hotels.
- City Hotel has the highest cancellations with a significant increasing trend. There is a positive correlation between Stay Duration and Cancellation Ratio for both hotels. The City Hotel's highest cancellation rate is for stays above 4 weeks, while the Resort Hotel's highest rate is for stays of 3 weeks. Implementing stricter cancellation policies and offering special promotions can help mitigate cancellations.
- Longer lead times are associated with higher cancellation rates for both City and Resort Hotels. The lowest cancellation rates are observed for lead times of <= 1 month, while the highest rates occur for lead times of 11-12 months. Both Resort and City Hotels experience the highest cancellation rate around a 1-year lead time. Implementing reminders and strict cancellation policies can help reduce cancellations and improve overall booking efficiency.