Investigate Hotel Business using Data Visualization

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

Rakamin
Academy

Created by:
Kenneth Wahyudi, S.Si.

kennethwahyudi48@gmail.com
https://www.linkedin.com/in/kenneth-wahyudi-80b886209/

I am an aspiring data scientist with interest in machine learning and paddling through the data lake. I have experiences as a research intern in BATAN, and several data science and machine learning projects under my belt. Specifically intermediate to advanced knowledge in Exploratory Data Analysis, Data Pre-processing, Supervised & Unsupervised Learning, as well as Data Visualization and Storytelling. And I am always looking forward to exploring new and uncharted territories that might allow myself to grow and improve.

Overview

Overview

It is of utmost importance for a company to consistently analyze its business performance. On this occasion, we will delve deeper into the realm of hospitality business. Our focal point is to understand the behavior of our customers when it comes to hotel reservations and its correlation with the cancellation rate. The insights we gather will be presented in the form of data visualization, making it more accessible and inherently persuasive.

All results are from outputs of codes written in Python and its packages, namely pandas, matplotlib, seaborn, etc. Jupyter Lab was the GUI used in this project.

Link to the dataset used in this project can be found in the following link :
https://drive.google.com/drive/folders/1CzpRqjjVzo_6E8e70gW_sym1gYQ3rzHt?usp=sharing

Data Preprocessing

Data Preprocessing

The dataset contains 29 columns and 119,390 rows. In this preprocessing, here are the outlines of what were done to the dataset :

- NULL values (or missing values) handling
- Invalid values handling
- Removal of unnecessary and redundant data

Now, let’s breakdown each of those outlined above even further.

The complete code and .ipynb file for the data preprocessing can be accessed in this link.

Data Preprocessing

- **NULL values (or missing values) handling**. After checking the data types and non-null value counts of each column, it is discovered that there are 4 columns that have missing values within them. They are the ‘children’, ‘city’, ‘agent’, and ‘company’ column. And after delving deeper by creating a data frame that houses column names, null value counts, and null values percentage to total dataset, it is discovered that the values of the ‘company’ column are around 94.3% missing. This means that the column is virtually useless. Therefore, for the other 3 columns, we only drop the null valued rows. But delete/drop the entire 'company' column.

Data Preprocessing

- **Invalid values handling**. We first displayed the value counts of object typed columns to see their values. After looking at the results, most of the columns have logical/sensible values. However, if we look closer to the 'meal' column, we can see that one of its values is 'Undefined'. This is practically impossible, since the other values have basically covered all the other options. The other values are 'Breakfast', 'Dinner', 'Full Board', and 'No Meal'. Now, common practice dictates that in a case like this, one should replace the values to match the mode value of the column. However, the mode value is breakfast. We can't just assume that all of those customers had the breakfast option. So logically speaking, the 'No Meal' value is more appropriate, since it is more logical to assume that they didn't have any meal than that they just have the breakfast option.

Data Preprocessing

- **Removal of unnecessary and redundant data**. We first display the mode value and its count from every column. From the results, we can see that some columns have very severe value/class imbalance. And the columns with mode value counts to total data percentage of over 95% are 'babies' at 99.26%, 'previous_bookings_not_canceled' at 99.04%, 'is_repeated_guest' at 98.6%, and 'days_in_waiting_list' at 96.56%. Out of all of the above columns, the 'babies' column is the most severe with 99.26% of its values are the majority value. And out of them all, the 'babies' column is the one that could be redundant, since there are also the 'adults' and 'children' column, and the mode value of the 'babies' column is 0. This means that an overwhelming most of the time (99.26% of the time to be exact), there are no babies as part of the guests in the hotel. Which would proof useless in the later part of the analysis, since virtually no information/insight could be gathered from such an imbalanced & redundant column. Therefore, it would be wise (and basically unharmful) to drop the 'babies' column from further analysis. And just use the 'adults' and 'children' columns for the amount of guests enlisted in a single booking.

Monthly Hotel Booking Analysis Based on Hotel Type

Monthly Hotel Booking Analysis Based on Hotel Type

In this analysis we’re going to analyze the monthly hotel bookings based on the hotel type. Here are the outlines of the process :

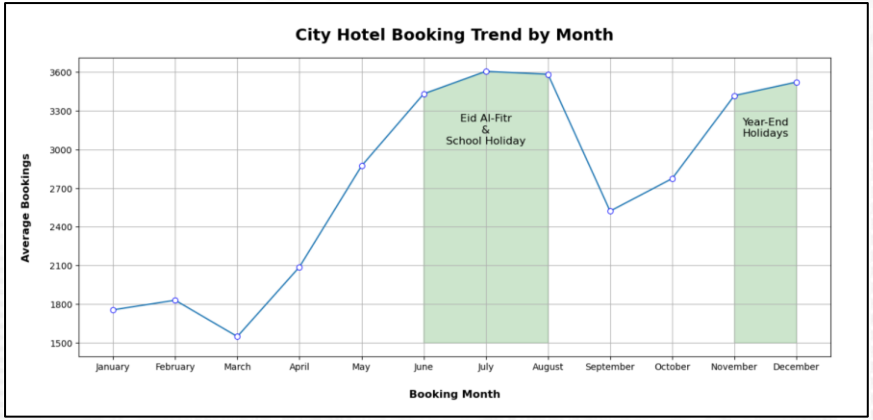
1. Creating an aggregate table that shows the comparison trend of amount of bookings between the city hotel and the resort hotel each month.

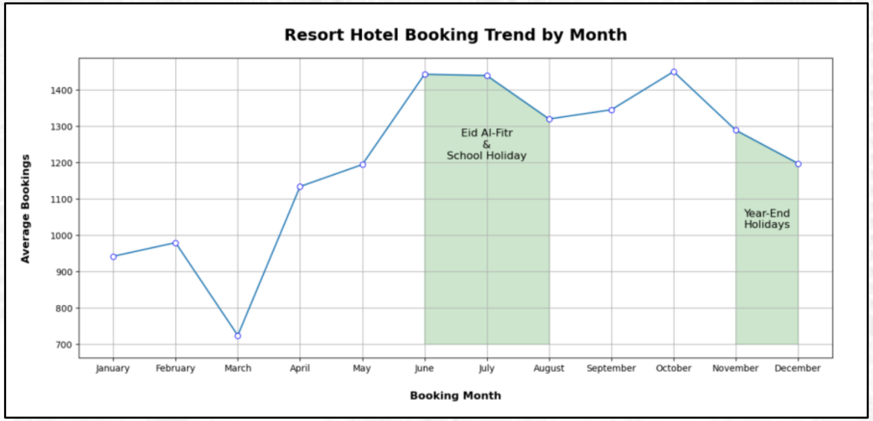
2. Normalizing the data, since out of all the 3 years (2017, 2018, and 2019), only September and October are present in all 3, whilst the other months are only present in 2 out of 3 of those years. Therefore, we’re going to only take the average of amounts of bookings of each month between the years it is present.

3. Sorting the month names so that they are in the correct order before finally plotting it into charts.

Now let’s look at the charts and analyze them.

Monthly Hotel Booking Analysis Based on Hotel Type

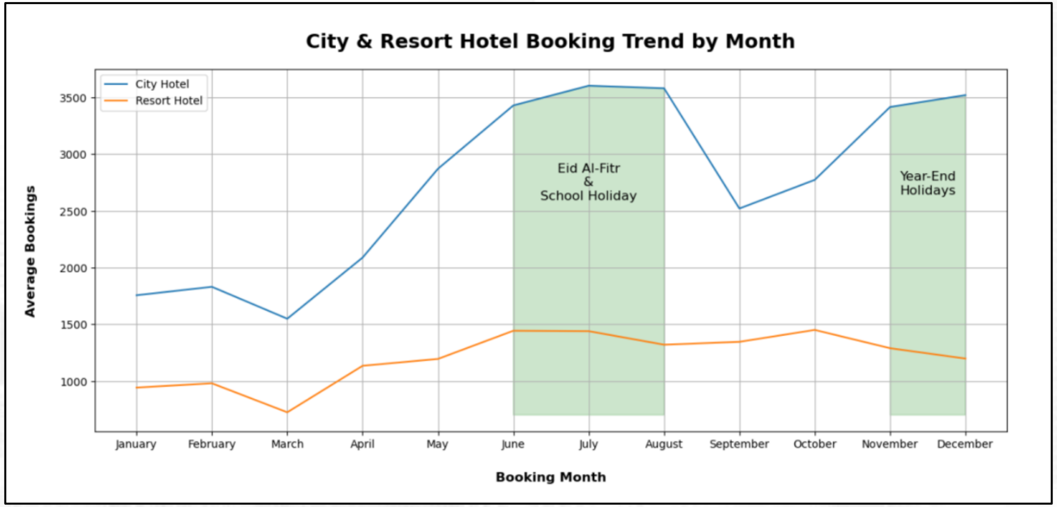




In the city hotel chart, the month with the least amount of bookings is march while the month with the most bookings is July, followed closely by August, December, June and November. In Indonesia, the most popular public holidays with lengthy paid leaves are the eid al-fitr (which falls around June in 2017, 2018, and 2019) and year-end holidays. Therefore it makes sense that these month intervals are the ones with the most bookings.

In the resort hotel chart, the month with the least amount of bookings is march while the month with the most bookings is June, followed closely by July and October. The popularity of the interval between June and August makes sense, however there is a slight anomaly that October has more bookings than November, and November has more bookings than December. This is the Resort Hotel after all, so it could be that because of how high the prices are in the year-end holidays, people tend to choose city hotels in that interval.

Monthly Hotel Booking Analysis Based on Hotel Type



And above are the trends of the city hotel and resort hotel side by side. It is visible that they follow similar trends until their trends diverge in the year-end interval, as previously mentioned.

The complete code and .ipynb file for the plot making and analysis can be accessed in this link.

Rakamin Academy

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates

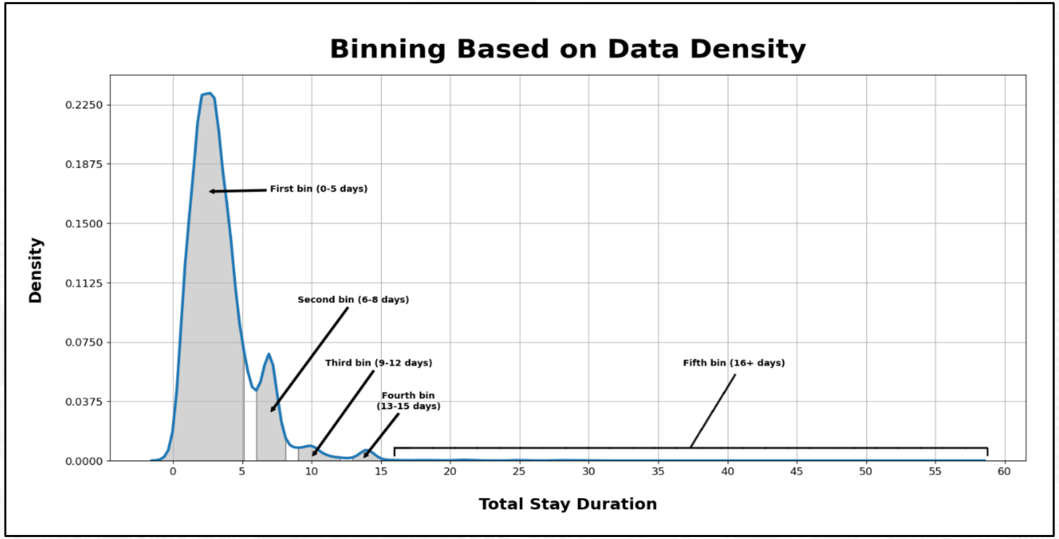
In this analysis, we’re going to analyze the impact analysis of stay duration on hotel bookings cancellation rates. Here are the outlines of the process involved in the analysis :

1. Making new columns that sums the total duration of stay, and binning/grouping the stay durations to get a more significant impact difference.
2. Grouping the stay duration bin values by cancellation and hotel type.
3. Making the plots of each stay duration’s cancellation ratio by hotel type.

Let us now proceed to the charts and analysis.

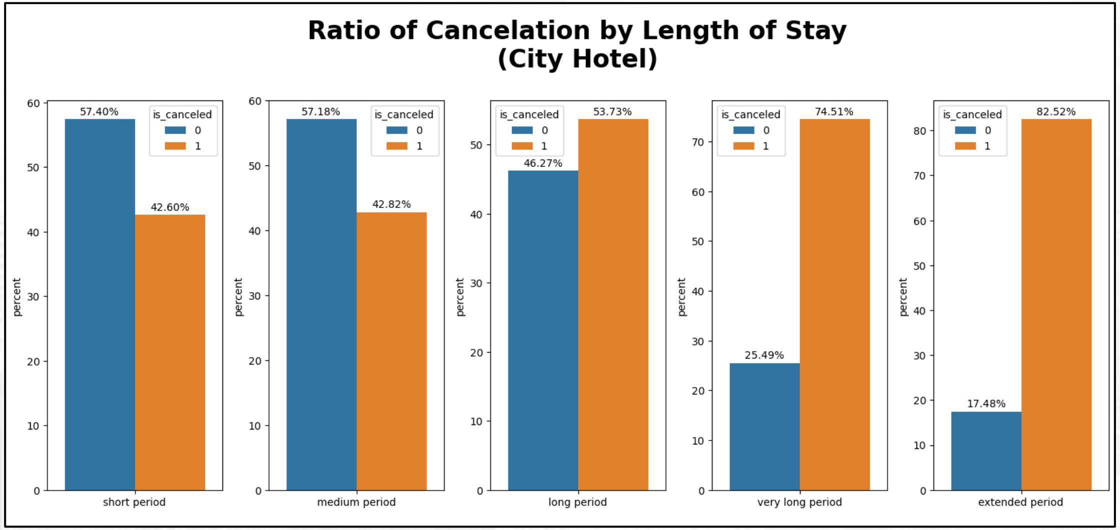
The complete code and .ipynb file for the plot making and analysis can be accessed in this link.

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates



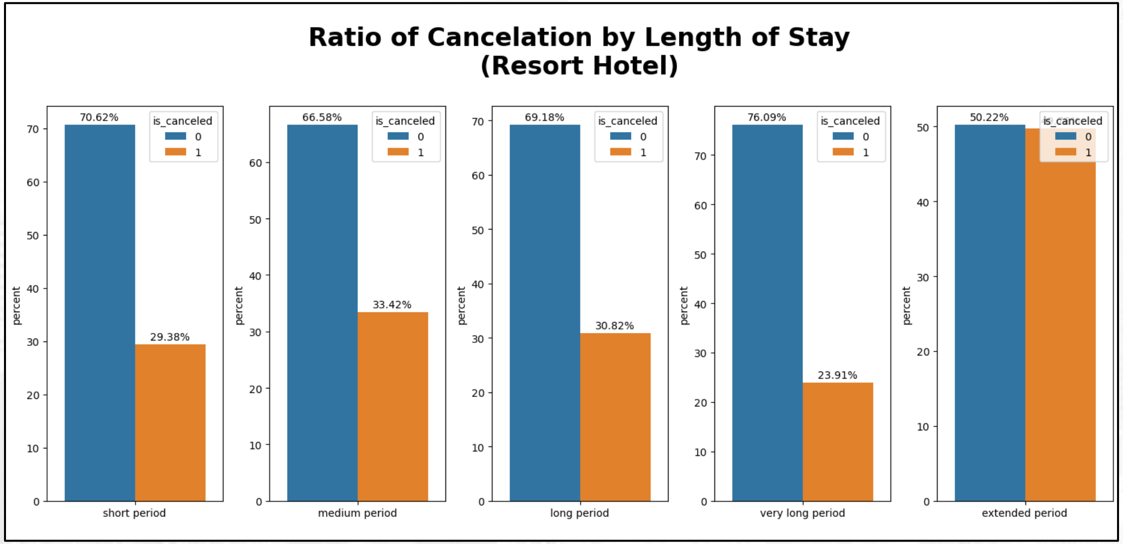
In order to get a more significant result, we're grouping or binning the total stay duration according to its peak densities. And in the chart above, we have binned the total stay duration by its peak densities, and getting 5 bins in total. The first bin is for 0-5 days of stay, the second bin is for 6-8 days of stay, the third bin is for 9-12 days of stay, the fourth bin is for 13-15 days of stay, and finally the fifth bin is for 16+ days of stay.

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates



In the city hotel chart above, we can see that stay duration does have an effect towards cancellation rate. With short and medium periods have around 43% cancelation rates. But as the periods get longer, the cancelation rates sky rocketed. From 54% in the long period to 75% in the very long period, before finally reaching 83% in the extended period. Further analysis is required as to why these are the cases.

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates



In the resort hotel chart above, we can see that the stay duration also has an effect towards cancellation rates. But contrary to the city hotel, the resort hotel's stay durations do not have meaningful impact other than the extended period stay durations. Where in the other shorter periods, the cancelation rates are around 24-34%, but in the extended period the cancelation rate sky rocketed to 50%. Again, further analysis is required as to why these are the cases.

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rates

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate

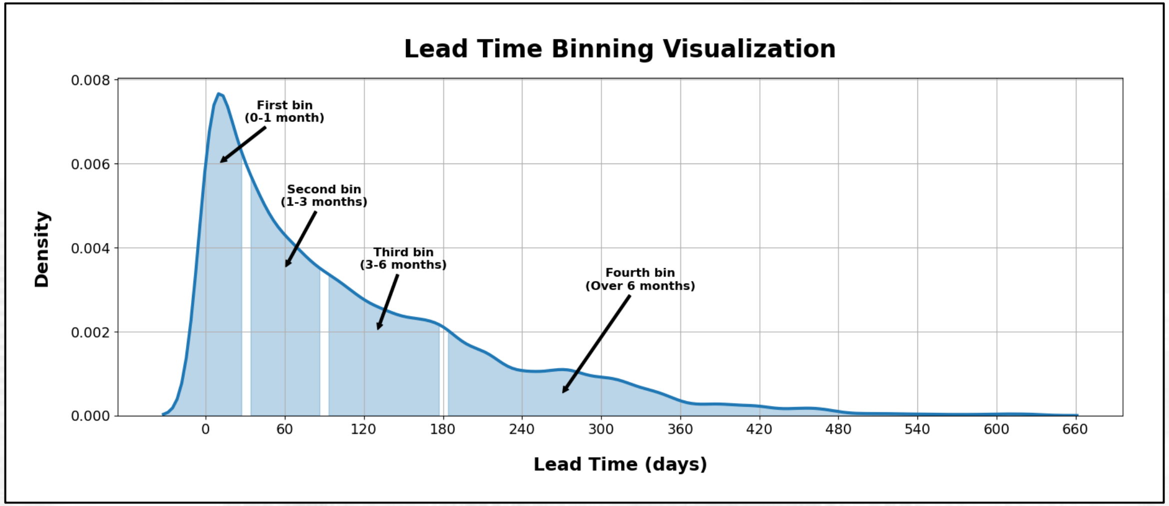
In this analysis, we're going to analyze the impact analysis of lead time on hotel bookings cancellation rates. Lead time is the time interval between bookings and the actual stay. Here are the outlines of the process involved in the analysis :

1. Segmenting the lead time values using their quartile values as guidelines. Then creating a new column in the dataset that houses the lead time segments.
2. Grouping the lead time bins/segments by the cancellation and hotel type.
3. Creating the plots of each lead time's cancellation ratio by hotel type.

Let us now proceed to the charts and analysis.

The complete code and .ipynb file for the plot making and analysis can be accessed in this link.

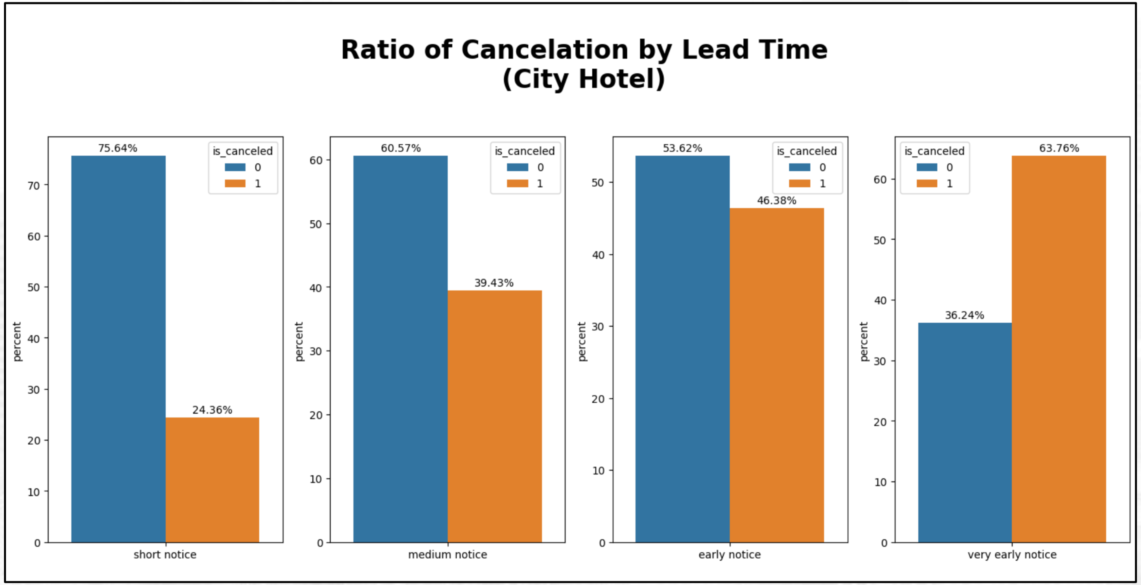
Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate



In order to get a more significant result, we need to bin/group the values of the lead_time column into segments. The distribution of the lead_time column values are not multi modal like the total_stays column. That means that it only has one peak and a smooth distribution. Therefore, we cannot just pick the bins by looking at the kde plot.

After looking at the descriptive statistics of the lead_time columns, it seems like it makes sense to use the quartiles as bins for the lead_time column. Since the quartiles lie near a reasonable value. Like 25% at 26 days (near a month), 50% at 79 days (near 3 months), and 75% at 169 days (near 6 months). Therefore, we're going to use the quartiles as a guideline in making the bins, and the bins will be 0-1 month, 1-3 months, 3-6 months, and over 6 months. And the following is the visualization of the binning process.

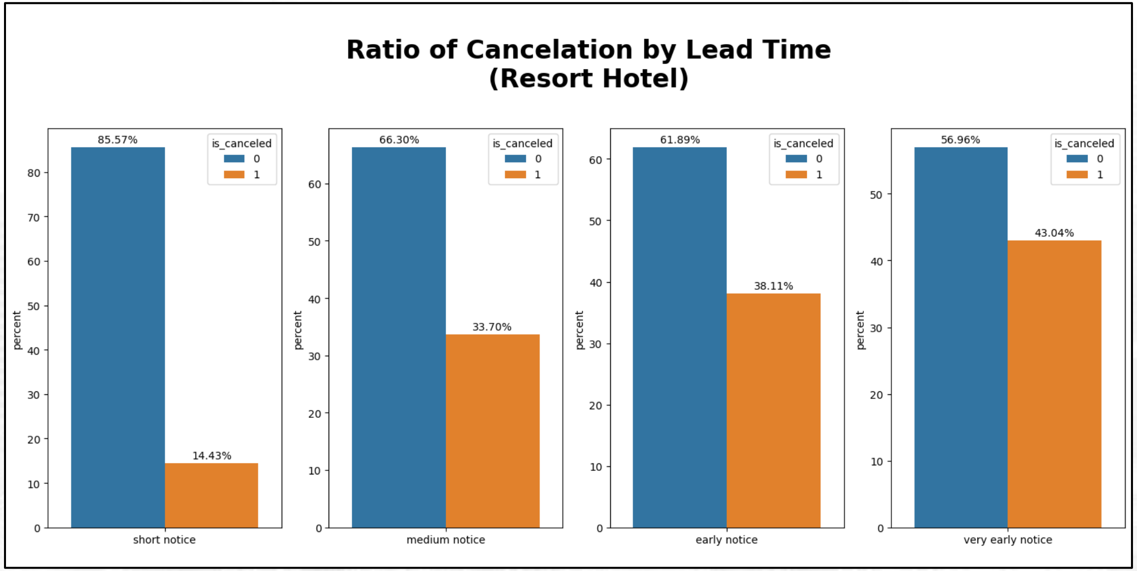
Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate



In the city hotel chart on the left, we can see that lead time does have a linear affect towards cancellation rates. The higher the lead time, the higher the chances of booking cancellation. With the short notice having a 24.36% cancellation rate, the medium notice having a 39.43% cancellation rate, the early notice having a 46.38% cancellation rate, and finally the very early notice having a 63.76% cancellation rate.

This is actually quite logical, as we can assume that when people plan a long way ahead of the due date, they tend to cancel it simply because it's too far into the future that other things have enough time to get in their way of actually being there to stay at the hotel.

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate



In the resort hotel chart on the left, we can also see that the lead time does have a linear affect towards cancellation rates. The higher the lead time, the higher the chances of booking cancellation. With short notice having a 14.43% cancellation rate, medium notice having a 33.7% cancellation rate, early notice having a 38.11% cancellation rate, and finally very early notice having a 43.04% cancellation rate.

Again, this is actually quite logical, as we can assume that when people plan a long way ahead of the due date, they tend to cancel it simply because it's too far into the future that other things have enough time to get in their way of actually being there to stay at the hotel.

Remarks

Remarks

The cancellation rates (both by lead time and stay duration) are lower in the resort hotel compared to the city hotel in every segment of the lead time & stay duration. It could be that resort hotels are generally more expensive, prestigious, and have better access to the holiday sights, therefore most resort hotel bookers might actually be holiday travelers, and they tend to commit and persist with their holiday plans. Compared to business travelers that have very undetermined travel itineraries. Although further analysis is still required to determine the other causes.