Investigate Business Hotel using Data Visualization

Created by:
Muhammad Aqajahs Al Qarani
Email : jalikarani@gmail.com
LinkedIn : linkedin.com/in/alqarani
Github : github.com/jalikarani

I have a **statistical background** that is relevant to my major in industrial science. Have an understanding in the fields of **data analysis, visualization dashboards, data science, and machine learning** with experience in the data field of approximately 3 years accompanied by a certificate of proficiency. Able to process data structurally, starting from **data collection, data preparation, data cleansing, pre-processing, modeling, prediction. Some commonly used tools include SQL, Python, PowerBI, Tableu, Looker Studio**

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

Business Understanding

Problem Statement:

A hospitality company needs to analyze customer behavior in hotel bookings and its relationship with the cancellation rate of hotel reservations. We will delve deeper to understand how certain factors influence customer decisions in choosing and canceling hotel reservations.

Goals

To improve operational efficiency and profitability of the hospitality company through a better understanding of customer behavior in hotel bookings and efforts to reduce the reservation cancellation rate.

Objectives

1. Collect and analyze customer data, including preferences, booking habits, cancellation history, and related information.
2. Present the findings through visualizations and storytelling.

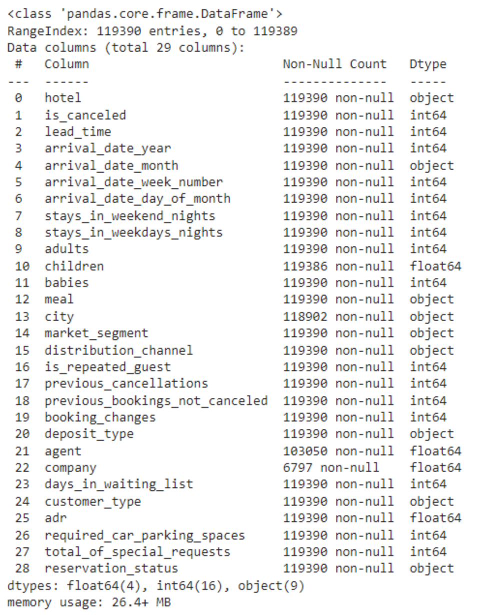
Data Preprocessing

Data Information:

- Based on the information provided, there are **29 features** with a total **of 119,390 rows**.
- There are missing values in **the 'children,' 'city,' 'agent,' and 'company'** features.
- The data types are as follows: **Float64 (4 features), int64 (16 features), object (9 features).**

Deskripsi Data :

Upon observation, it's evident that many features have outliers. This can be seen from the difference between the mean and median, as well as the significant gap between the maximum and Q3 (third quartile). Here are some features that show indications of outliers: **lead_time, stays_in_weekend_nights, stays_in_weekdays_nights, adults, children, previous_cancellations, previous_bookings_not_canceled, booking_changes, agent, days_in_waiting_list, adr.**



```
<class 'pandas.core.frame.DataFrame'>
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)
memory usage: 26.4+ MB
```

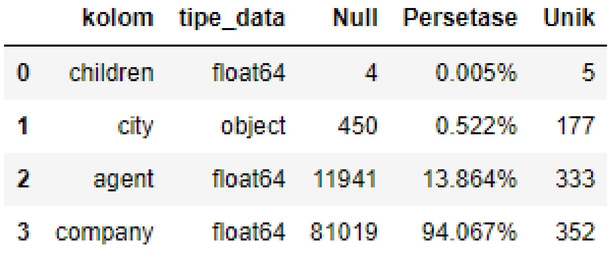
Data Preprocessing

Handling Duplicates:

There are 33,261 duplicate records, which need to be dropped to avoid any inaccuracies in the analysis.

Handling Missing Value :

- children: 0.005% (4) of the total data
- city: 0.522% (450) missing from the total data
- agent: 13.864% (11,941) missing from the total data
- company: 94.067% (81,019) missing from the total data



	kolom	tipe_data	Null	Persetase	Unik
0	children	float64	4	0.005%	5
1	city	object	450	0.522%	177
2	agent	float64	11941	13.864%	333
3	company	float64	81019	94.067%	352

Cara Handling :

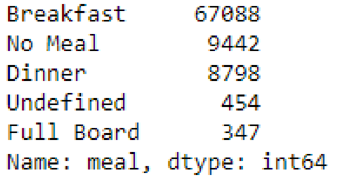
1. For the children feature, it is filled with 0, as it is possible that the customer does not have children.
2. For the agent feature, it is filled with 0, as there is no agent involvement.
3. For the company feature, it is filled with 0, as no company is defined.
4. For the city feature, it will be filled with ‘Unknown’, as the city is not defined.

***Additionally, we will change the data types of numerical features 'children,' 'agent,' and 'company' from float to integer.**

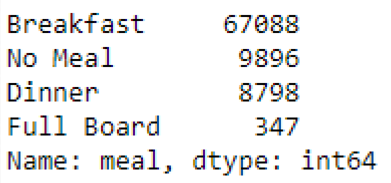
Data Preprocessing

Handling Invalid Value :

If we observe, there is an inconsistent category labeled **'Undefined'** in the 'meal' feature. This category is assumed to indicate that customers did not order anything, so it will be categorized as **'No Meal'**.



```
Breakfast     67088
No Meal        9442
Dinner         8798
Undefined       454
Full Board      347
Name: meal, dtype: int64
```



```
Breakfast     67088
No Meal        9896
Dinner         8798
Full Board      347
Name: meal, dtype: int64
```

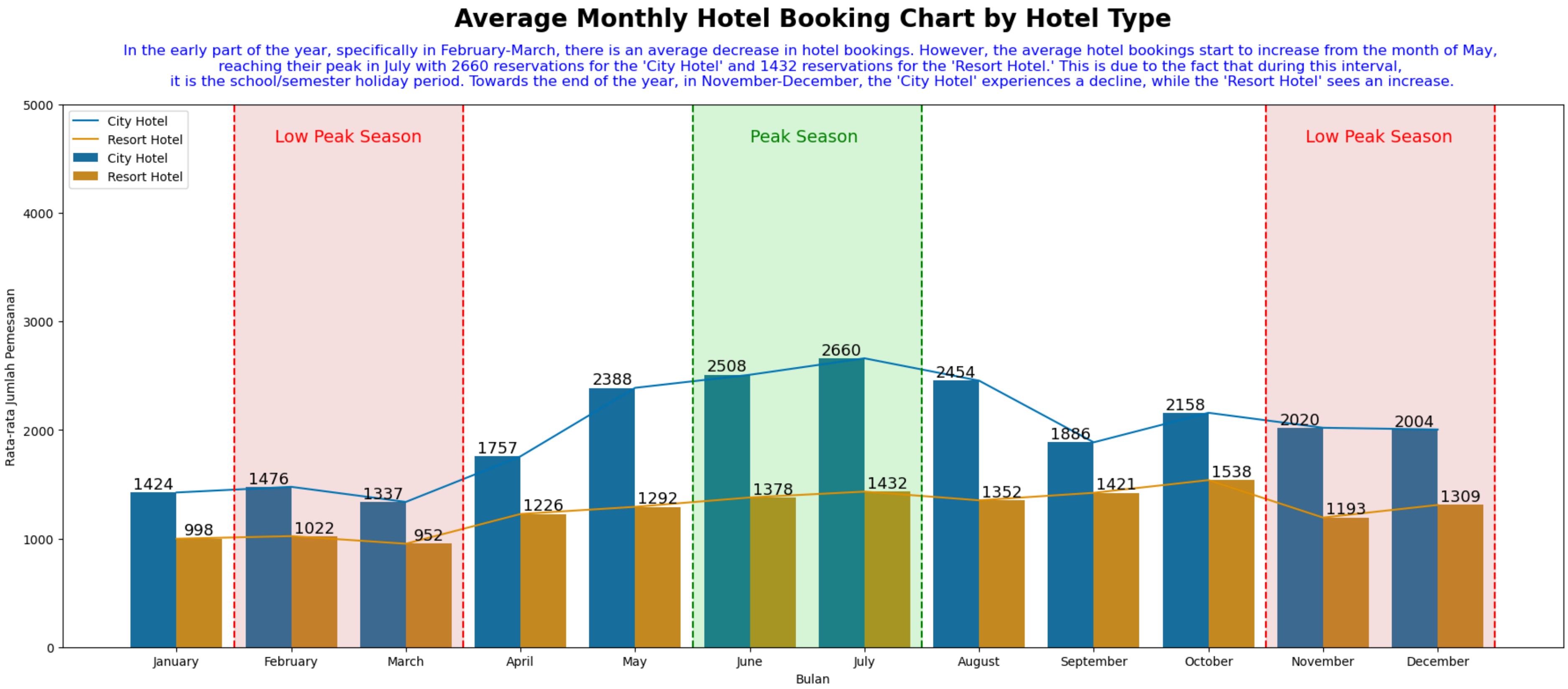
Featues Engineering :

Here, we will check the number of customers based on their orders, so customer categories need to be summed first. Afterward, we will calculate the total orders based on whether they are on weekends or weekdays. This will result in creating new features called "**total_guests**" and "**stay_duration.**"

***Next, we will perform filtering to display only the data that has a customer count based on orders (removing rows with a count of 0).**

Code selengkapnya bisa di cek [disini]

Monthly Hotel Booking Analysis Based on Hotel Type



Monthly Hotel Booking Analysis Based on Hotel Type

Interpretation:

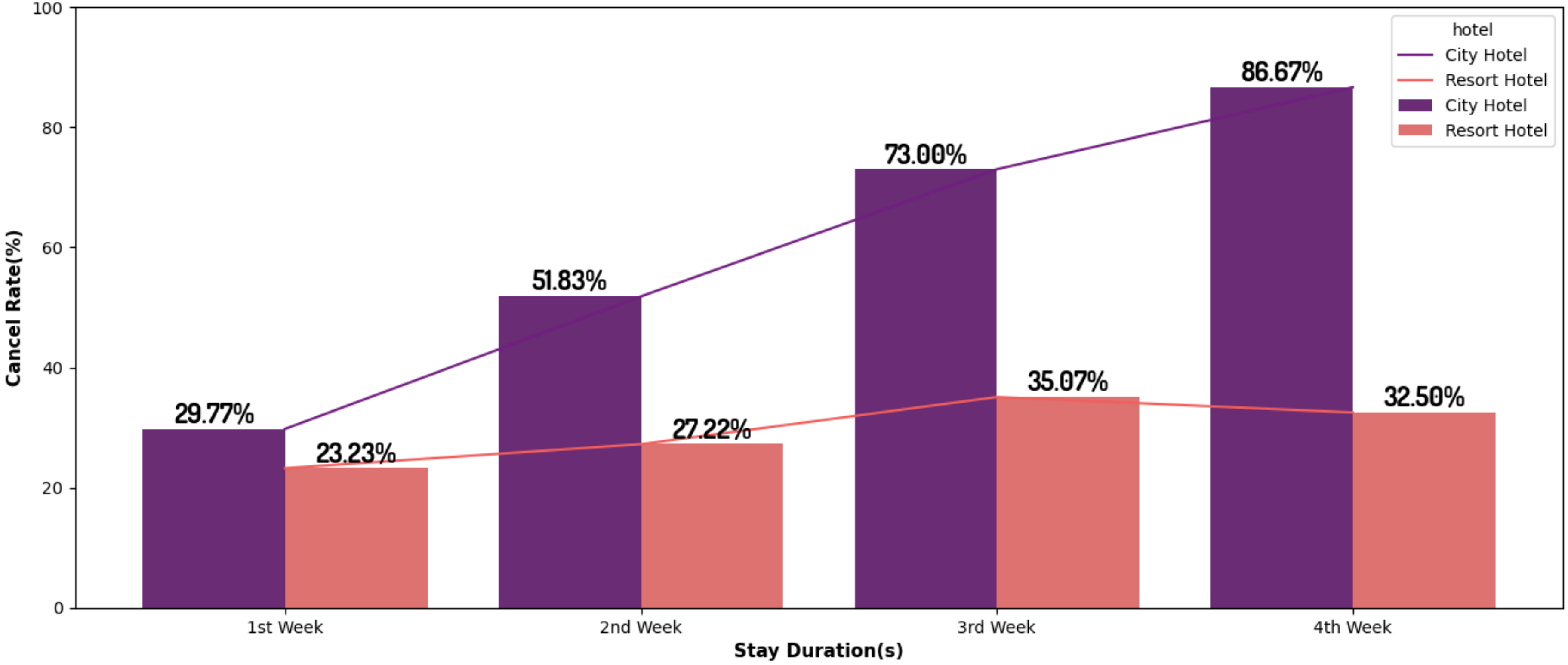
- At the beginning of the year, it can be observed that the average hotel bookings experienced a decline, specifically **from February to March**. This trend is consistent for both types of hotels. This decrease may be attributed to the start of academic activities in schools/universities and offices, leading to fewer hotel bookings.
- The peak increase in average bookings occurs in **July**, with a continuous rise starting from **May to July**. During this period, it is the mid-year holiday or mid-semester break, prompting many people to take vacations, resulting in increased hotel bookings.
- Towards the end of the year, the trends for city and resort hotels diverge. **City hotels experience a slight decrease in average bookings**, although not significantly compared to the previous month. **Meanwhile, Resort hotels see a slight increase**, although it is not highly significant either.

Code selengkapnya bisa dilihat disini

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates

Cancellation Percentage per Length of Stay by Hotel Type (%)

In general, the longer the duration of stay, the higher the likelihood of customers canceling their reservations.
Based on the City Hotel type, the highest cancellation rate occurs for stays in the fourth week (86.67%).
Meanwhile, for the Resort Hotel type, the peak is in the third week (35.07%).



Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates

Interpretation:

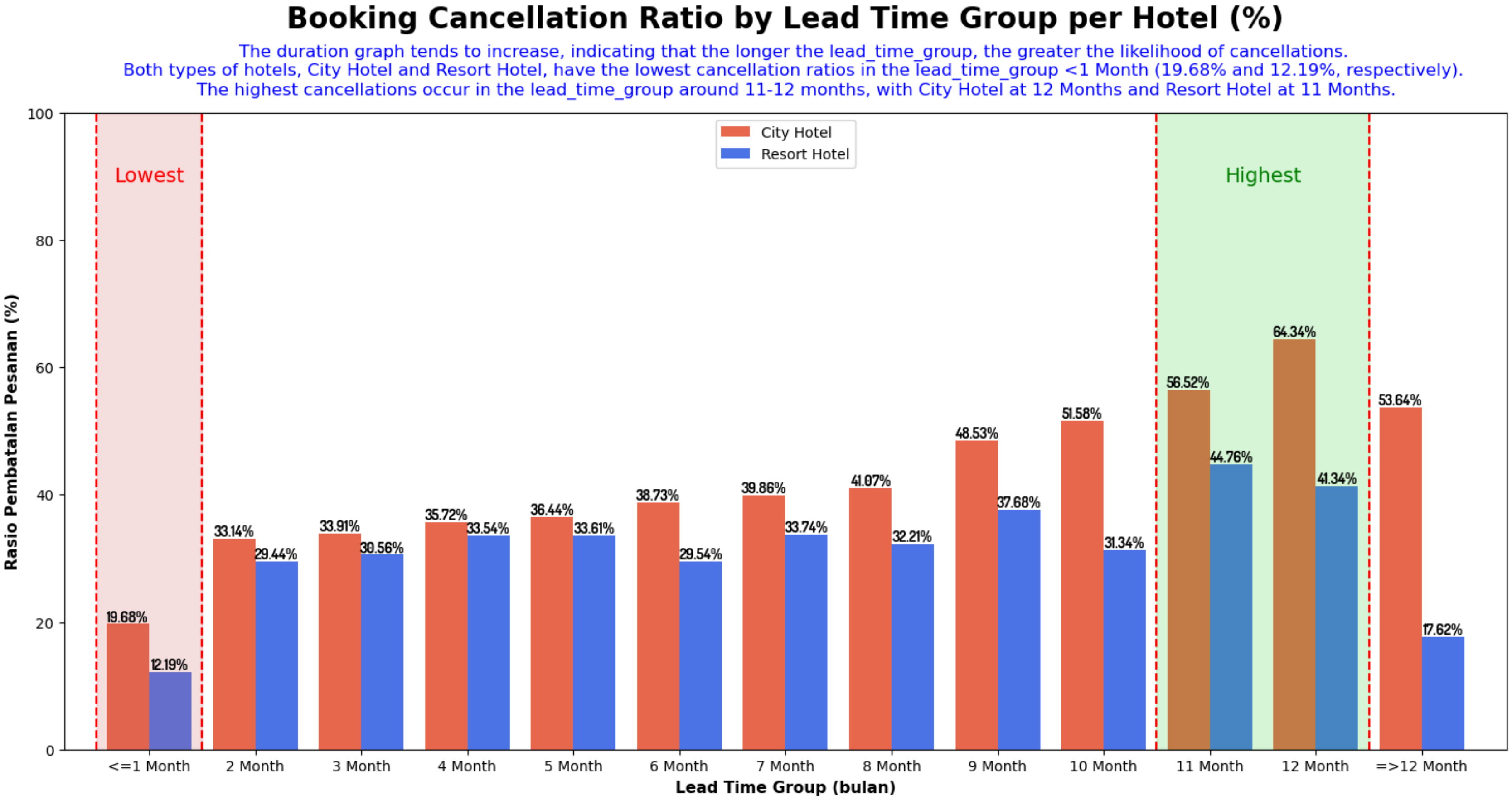
- **Cancellation rates for City Hotels show a significant** variation when analyzed based on the length of stay, with a similar **but not highly significant trend** observed for Resort Hotels.

- The highest cancellation rate for City Hotels occurs for **stays longer than 4 weeks**, whereas for Resort Hotels, the highest cancellation rate is seen for **stays of 3 weeks**.

- Both types of hotels, City and Resort, exhibit a **positive correlation between the length of stay and the cancellation percentage**. This means that as customers stay for a longer duration, the cancellation rate tends to increase.

Code selengkapnya bisa dilihat disini

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate



Booking Cancellation Ratio by Lead Time Group per Hotel (%)

The duration graph tends to increase, indicating that the longer the lead_time_group, the greater the likelihood of cancellations.
Both types of hotels, City Hotel and Resort Hotel, have the lowest cancellation ratios in the lead_time_group <1 Month (19.68% and 12.19%, respectively).
The highest cancellations occur in the lead_time_group around 11-12 months, with City Hotel at 12 Months and Resort Hotel at 11 Months.

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate

Interpretation:

- The graph tends to show an **upward trend in the data**, with a temporary decline for bookings made more than 12 months in advance.

- **As the lead time for booking increases, the likelihood of cancellation also increases**.

- Both types of hotels, City and Resort, have the lowest cancellation rates in the **'lead_time_group <1 Month' category**, with percentages **of 19.68% and 12.19%,** respectively.

- The highest cancellations occur in the 'lead_time_group around **11-12 months**,' with City Hotels at 12 months having a cancellation rate of **64.34%,** and Resort Hotels at **11 months** having a rate of **44.76%.**

Code selengkapnya bisa dilihat disini