Rakamin
Academy

Investigate Business Hotel using Data Visualization

Created by:
Nishrina Rawi
nishrinarawi@gmail.com
www.linkedin.com/in/nishrina-rawi

An enthusiastic Bachelor of Statistics graduate with a passion for pursuing a career as a Data Analyst. Possessing a strong foundation in statistical analysis, I am eager to contribute to data-driven decision-making processes. Experienced in conducting surveys and proficient in utilizing data analytics tools such as Ms. Excel, SQL, RStudio, and Python to extract meaningful insights from data.

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

Overview

It is very important for a company to always analyze its business performance. On this occasion, we will delve deeper into business in the hospitality sector. Our focus is to find out how our customers behave in making hotel reservations, and its relationship to the rate of cancellation of hotel reservations. We will present the results of the insights we find in the form of data visualization to make it easier to understand and more persuasive.

Data Preprocessing

To get quality insights, data preprocessing is very important because garbage in = garbage out. What is done at this stage is data exploration and data cleaning.

1. Data Exploration:

At this stage, exploration of the data is carried out with the aim of finding out the suitability of the data type and whether there are any suspicious values in the data.

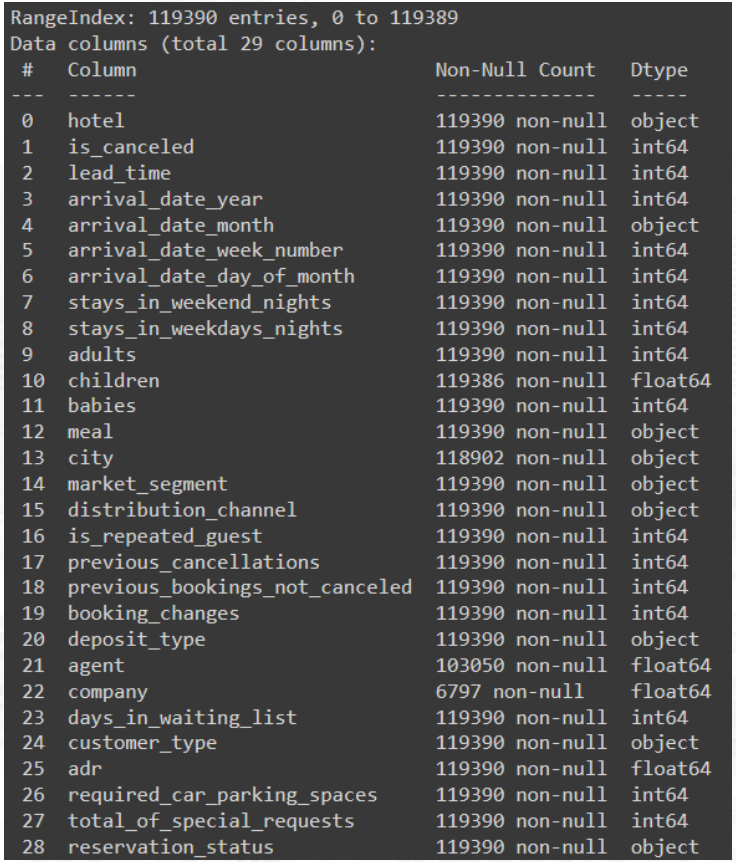
2. Data Cleaning:

After obtaining information through data exploration, a data cleaning process is carried out which includes handling missing values and handling inappropriate values.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

1. Data Exploration:



```
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 
```

- Data shape:

Data consist of 119390 rows and 29 columns

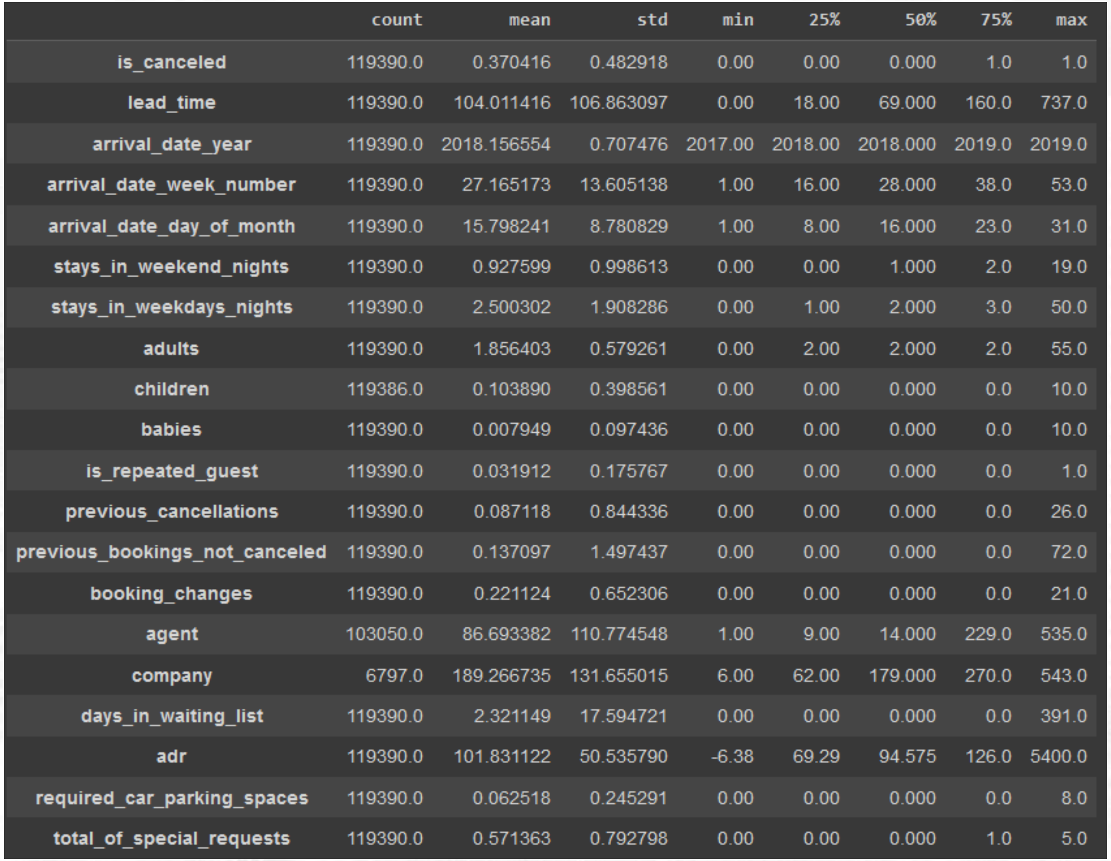
- Missing value

There are several columns that have missing values, including children, city, agent, and company

Untuk selengkapnya, dapat melihat jupyter notebook disini

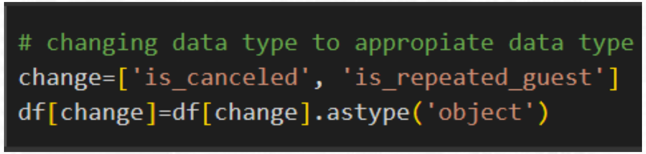
Data Preprocessing

1. Data Exploration:



	count	mean	std	min	25%	50%	75%	max
is_canceled	119390.0	0.370416	0.482918	0.00	0.00	0.000	1.0	1.0
lead_time	119390.0	104.011416	106.863097	0.00	18.00	69.000	160.0	737.0
arrival_date_year	119390.0	2018.156554	0.707476	2017.00	2018.00	2018.000	2019.0	2019.0
arrival_date_week_number	119390.0	27.165173	13.605138	1.00	16.00	28.000	38.0	53.0
arrival_date_day_of_month	119390.0	15.798241	8.780829	1.00	8.00	16.000	23.0	31.0
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.00	0.00	1.000	2.0	19.0
stays_in_weekdays_nights	119390.0	2.500302	1.908286	0.00	1.00	2.000	3.0	50.0
adults	119390.0	1.856403	0.579261	0.00	2.00	2.000	2.0	55.0
children	119386.0	0.103890	0.398561	0.00	0.00	0.000	0.0	10.0
babies	119390.0	0.007949	0.097436	0.00	0.00	0.000	0.0	10.0
is_repeated_guest	119390.0	0.031912	0.175767	0.00	0.00	0.000	0.0	1.0
previous_cancellations	119390.0	0.087118	0.844336	0.00	0.00	0.000	0.0	26.0
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.00	0.00	0.000	0.0	72.0
booking_changes	119390.0	0.221124	0.652306	0.00	0.00	0.000	0.0	21.0
agent	103050.0	86.693382	110.774548	1.00	9.00	14.000	229.0	535.0
company	6797.0	189.266735	131.655015	6.00	62.00	179.000	270.0	543.0
days_in_waiting_list	119390.0	2.321149	17.594721	0.00	0.00	0.000	0.0	391.0
adr	119390.0	101.831122	50.535790	-6.38	69.29	94.575	126.0	5400.0
required_car_parking_spaces	119390.0	0.062518	0.245291	0.00	0.00	0.000	0.0	8.0
total_of_special_requests	119390.0	0.571363	0.792798	0.00	0.00	0.000	1.0	5.0

`is_canceled` and `is_repeated_guest` columns have a minimum value of 0 and a maximum value of 1. This indicates that the two columns are not of integer type, but object.

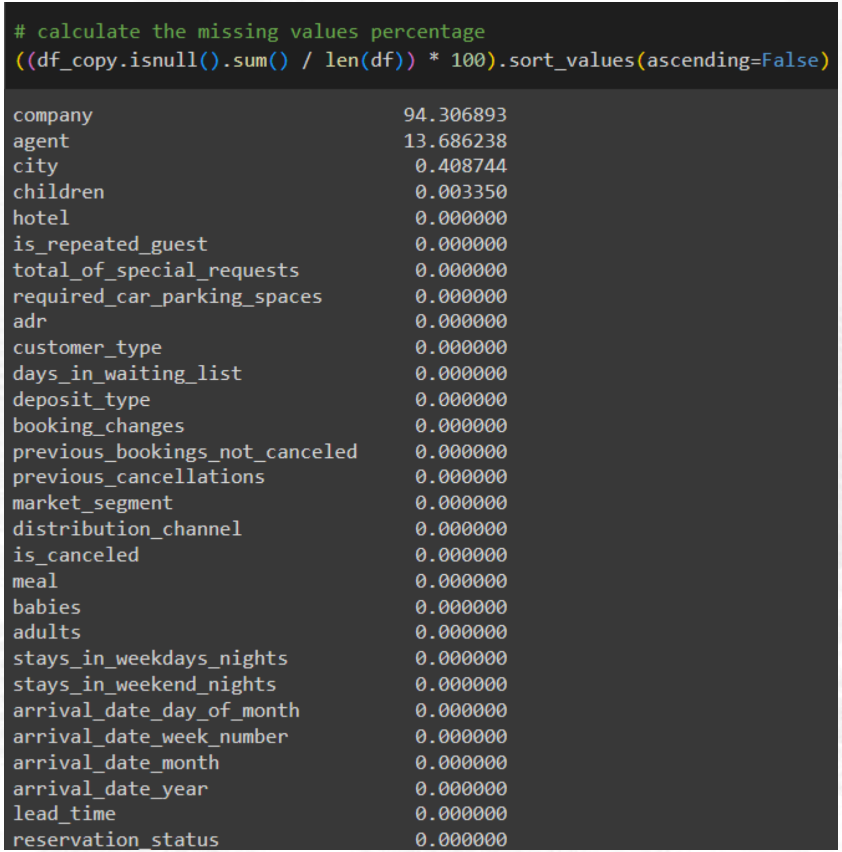


```
# changing data type to appropiate data type
change=['is_canceled', 'is_repeated_guest']
df[change]=df[change].astype('object')
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

1. Data Exploration:



```
# calculate the missing values percentage
((df_copy.isnull().sum() / len(df)) * 100).sort_values(ascending=False)
```

```
company                           94.306893
agent                             13.686238
city                               0.408744
children                           0.003350
hotel                              0.000000
is_repeated_guest                  0.000000
total_of_special_requests          0.000000
required_car_parking_spaces        0.000000
adr                                0.000000
customer_type                      0.000000
days_in_waiting_list               0.000000
deposit_type                       0.000000
booking_changes                    0.000000
previous_bookings_not_canceled     0.000000
previous_cancellations             0.000000
market_segment                     0.000000
distribution_channel               0.000000
is_canceled                        0.000000
meal                               0.000000
babies                             0.000000
adults                             0.000000
stays_in_weekdays_nights           0.000000
stays_in_weekend_nights            0.000000
arrival_date_day_of_month          0.000000
arrival_date_week_number           0.000000
arrival_date_month                 0.000000
arrival_date_year                  0.000000
lead_time                          0.000000
reservation_status                 0.000000
```

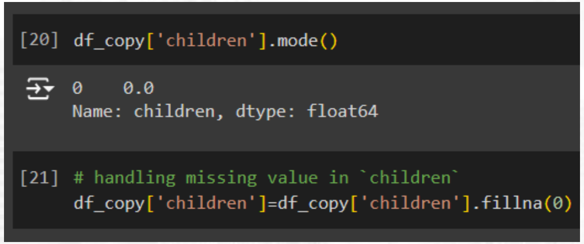
- In the `children` column there are 4 (0.00335%) data that have a null value
- In the `city` column there are 488 (0.408744%) data that have a null value
- In the `agent` column there are 16340 (13.69%) data that have a null value
- In the `company` column there are very large missing values, namely 94.30% of the data with NaN values.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

2. Data Cleaning:

a. Missing values handling



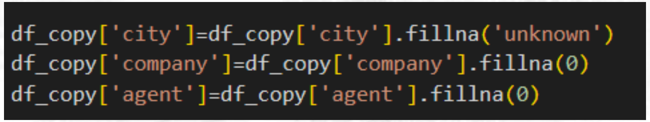
```
[20] df_copy['children'].mode()

     0    0.0
     Name: children, dtype: float64

[21] # handling missing value in `children`
     df_copy['children']=df_copy['children'].fillna(0)
```

Fill missing value in children with mode value

To resolve missing values in the children column, fill it in with the mode value in that column. Most customers do not bring children, therefore the missing value is filled with the value **0**.



```
df_copy['city']=df_copy['city'].fillna('unknown')
df_copy['company']=df_copy['company'].fillna(0)
df_copy['agent']=df_copy['agent'].fillna(0)
```

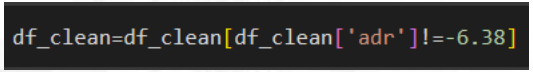
•For orders where the city value is unknown, fill in the value ***unknown***.
•The ID of the company responsible for placing orders with a null value is filled with the value **0**
•The ID of the agent responsible for placing orders with a null value is filled with the value **0**

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

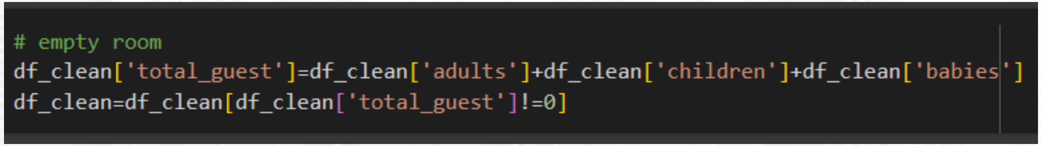
2. Data Cleaning:

b. Unsuitable value handling



```
df_clean=df_clean[df_clean['adr']!=-6.38]
```

The ‘adr’ or average daily rate value cannot be negative, therefore data with negative adr value is deleted.



```
# empty room
df_clean['total_guest']=df_clean['adults']+df_clean['children']+df_clean['babies']
df_clean=df_clean[df_clean['total_guest']!=0]
```

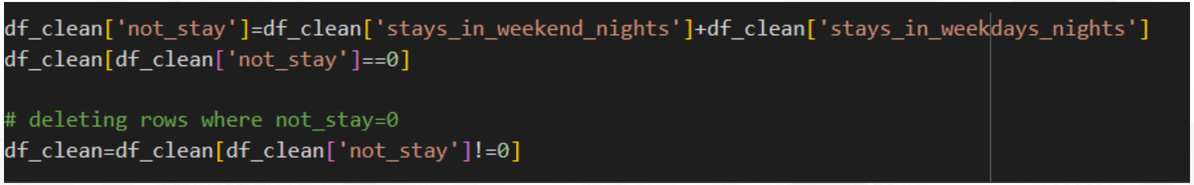
By adding up the adult, children, and babies columns, we get the total guests. It is impossible for the total number of guests from one hotel booking to be 0, which means there are no guests. Therefore data with total guests=0 is deleted.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

2. Data Cleaning:

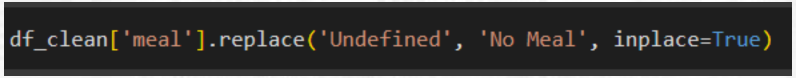
b. Unsuitable value handling



```
df_clean['not_stay']=df_clean['stays_in_weekend_nights']+df_clean['stays_in_weekdays_nights']
df_clean[df_clean['not_stay']==0]

# deleting rows where not_stay=0
df_clean=df_clean[df_clean['not_stay']!=0]
```

By adding up the stays_in_weekend_nights and stays_in_weekdays_nights columns, we get the number of weeks the customer stayed at the hotel. Customers with a value of 0 in both columns means they have never stayed at a hotel. Therefore, this data was deleted.

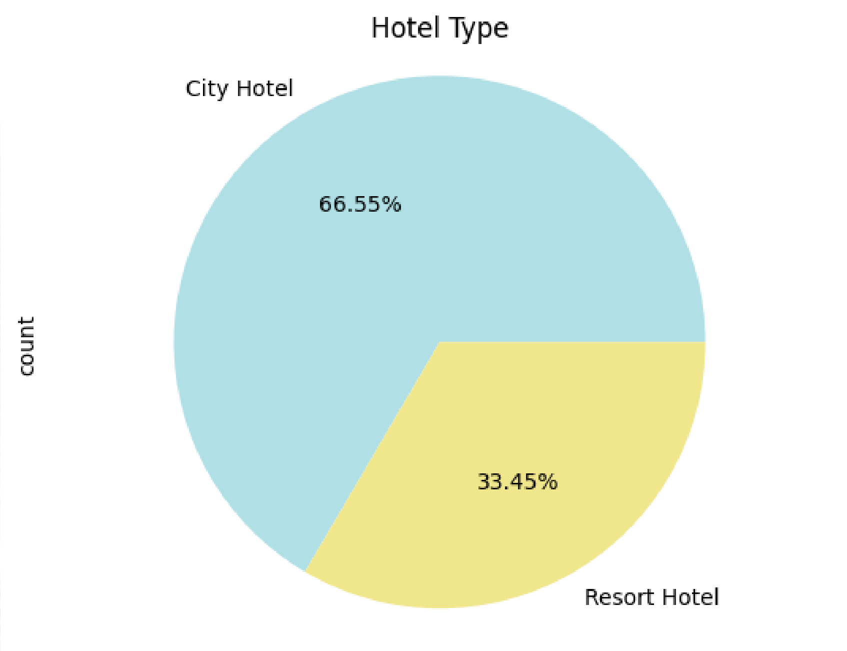


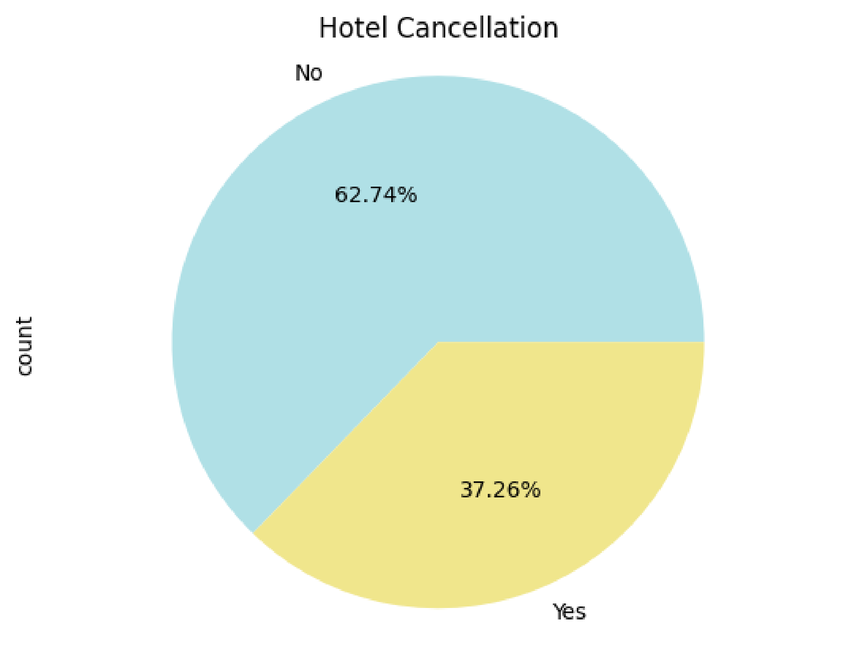
```
df_clean['meal'].replace('Undefined', 'No Meal', inplace=True)
```

For the meal column with a value of Undefined, replace it with No Meal

Untuk selengkapnya, dapat melihat jupyter notebook disini

Exploratory Data Analysis

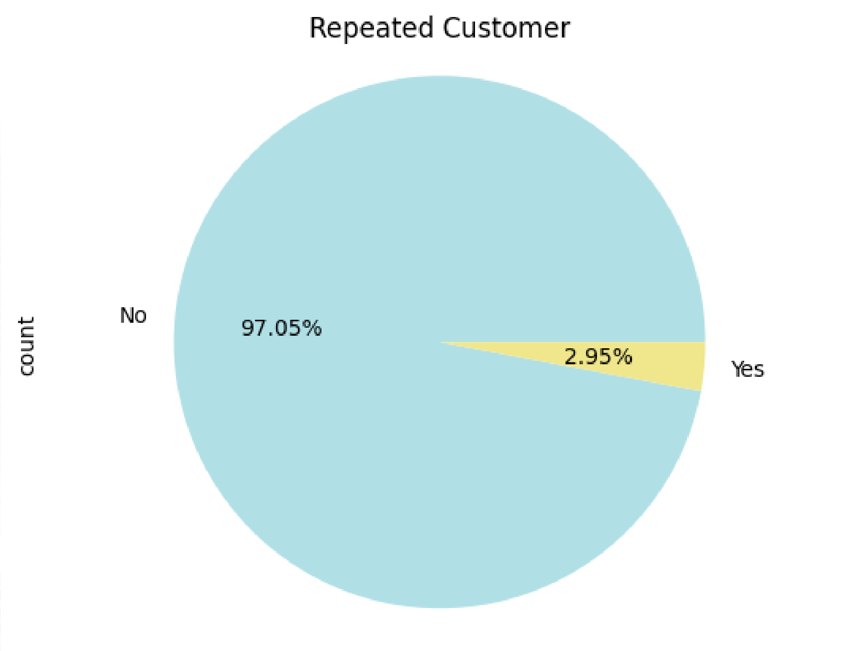


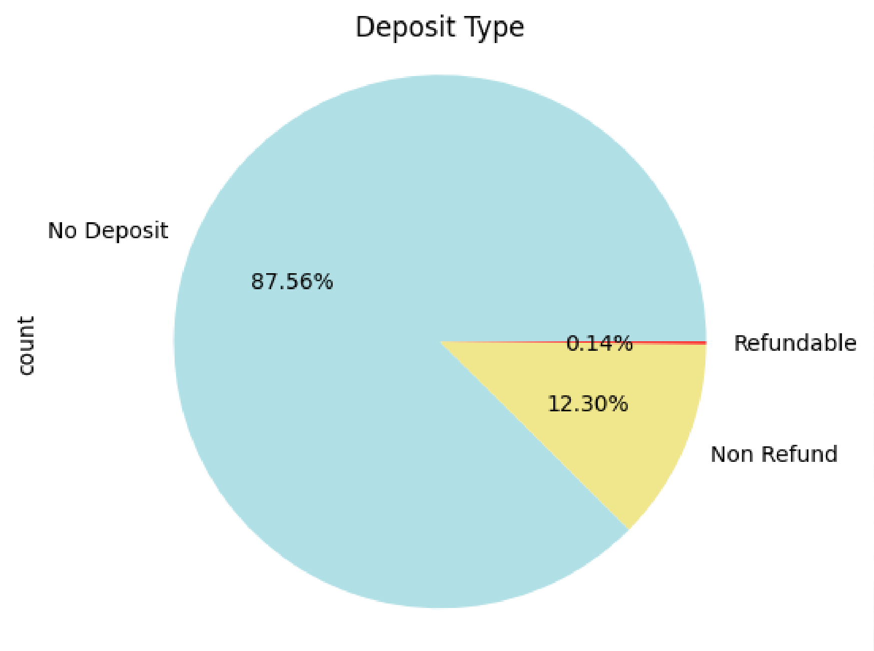


City hotels are booked more often than resort hotels. 66.55% of the booking history consists of city hotel bookings, while the rest are resort hotel bookings.

From the entire hotel booking history, it is known that 37.26% of the bookings were canceled by customers.

Exploratory Data Analysis

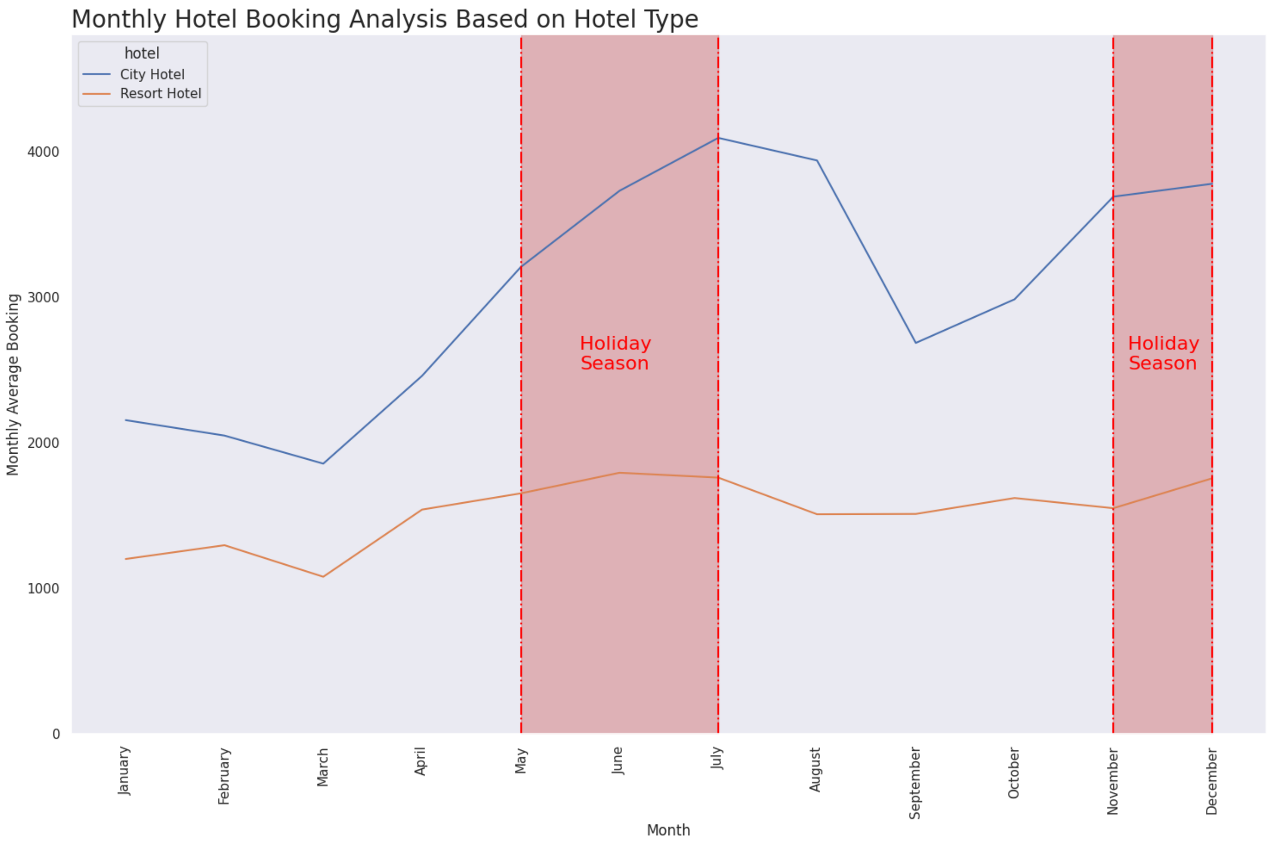




Only 2.95% of bookings are made by repeat customers.

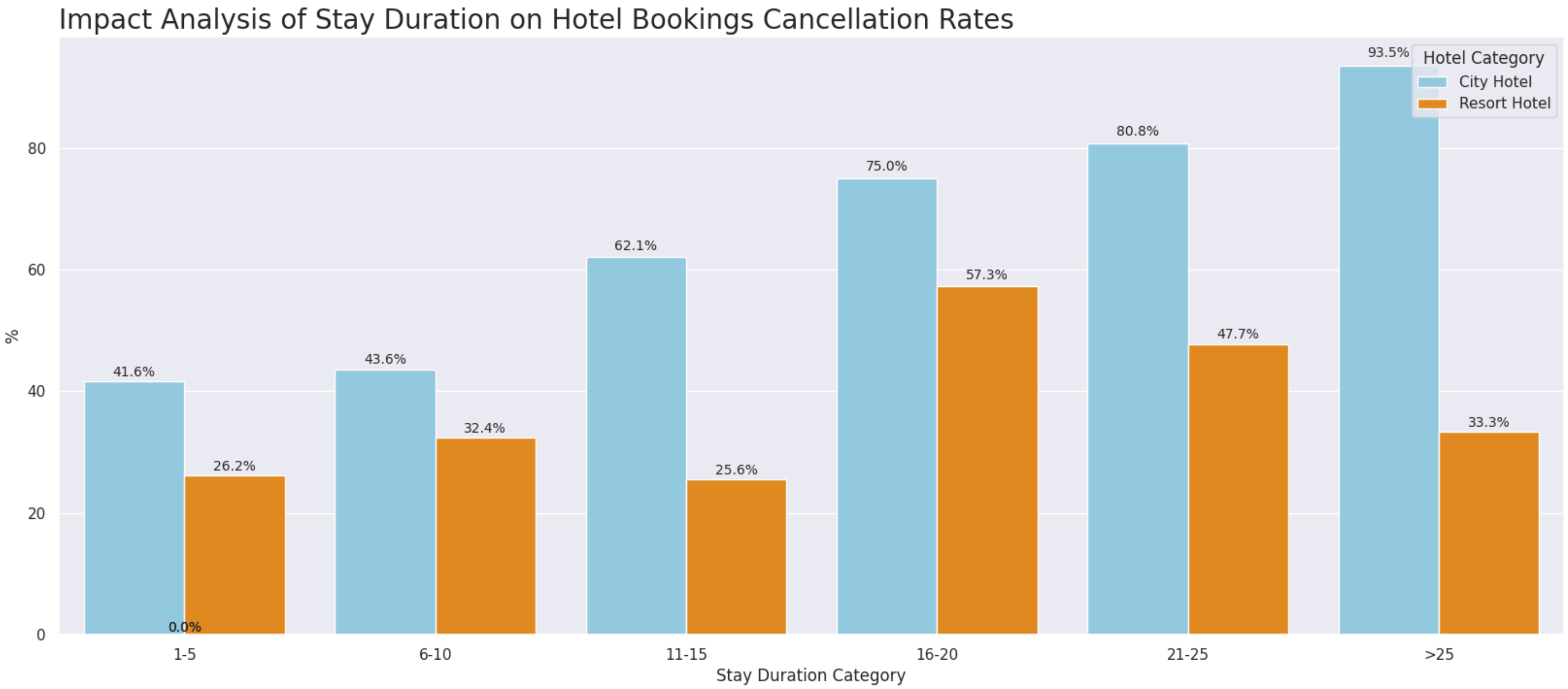
87.56% of bookings are made without a deposit. To lower the cancellation rate, management could implement a deposit policy.

Monthly Hotel Booking Analysis Based on Hotel Type



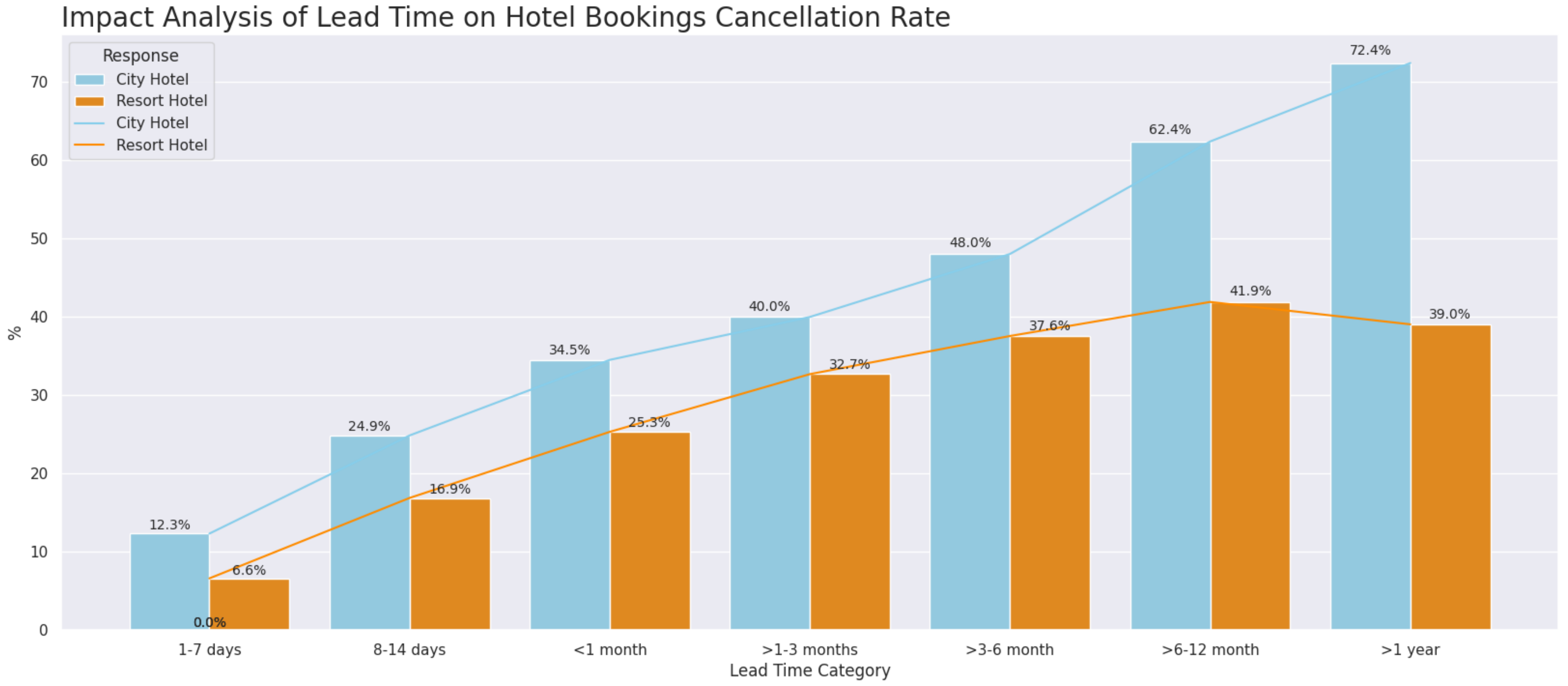
The patterns in average monthly bookings for City Hotels and Resort Hotels are similar, both experiencing fluctuations from month to month. However, **the average monthly bookings for City Hotels are consistently higher than those for Resort Hotels**, due to the greater volume of bookings at City Hotels. Notably, there **are increases during the holiday seasons in Indonesia, particularly from May to July and at the end of the year**. The peak in average hotel bookings occurs in July.

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates



The cancellation rates of resort hotels fluctuate based on stay duration categories, whereas for city hotels, the cancellation rate increases linearly with stay duration. The longer the stay duration at booking, the higher the cancellation rate of those reservations. For stays longer than 25 days, the cancellation rate reaches 93.5%.

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate



Both city hotel and resort hotel cancellation rates are more likely to increase as the time between booking and the hotel check-in date lengthens (lead time).

Business Recommendation

Recommendation

- Offering discount vouchers to customers who have booked and stayed at the hotel multiple times is a strategy to encourage repeat bookings.
- Implementing a deposit policy can help reduce the cancellation rate. When customers are required to pay a deposit, they are more likely to commit to their bookings, as they have a financial stake in the reservation. This reduces the likelihood of cancellations and helps ensure more stable occupancy rates.
- Implement dynamic pricing and flexible cancellation policies to manage fluctuations in bookings, especially during peak holiday seasons. For City Hotels, consider offering discounts or incentives for longer stays to mitigate the higher cancellation rates associated with extended bookings.
- Customers with a good booking history may be less likely to cancel their reservations compared to first-time or infrequent guests. By prioritizing these customers, hotels can potentially reduce the risk of cancellations during high-demand periods, ensuring more stable occupancy rates.