Investigate Business Hotel using Data Visualization

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

Created by:
Muhammad Cikal Merdeka
Email : mcikalmerdeka@gmail.com
LinkedIn : linkedin.com/in/mcikalmerdeka
Github : github.com/mcikalmerdeka

Dedicated entry-level data scientist with analytical and experimental background of Physics. My graduation 2023, a pivotal year marked by significant advancements in artificial intelligence with the introduction of GPT-4 and other generative AI models, has fueled my curiosity and excitement to delve into the field of data. I have comprehensive grasp of data science methodology from business understanding to modelling process with proficiency in **Python, SQL, Tableau, Power BI, Looker Studio and other tools** related to data analytics workflow from several coursework and bootcamps.

Project Overview

The hospitality industry encompasses businesses that provide accommodation for guests, including room reservations, meals, and other facilities. It can range from small to large enterprises, from one-star to five-star hotels. Success in this industry depends on factors such as location, service quality, and competitive pricing. The demand for comfortable and affordable accommodation makes the hotel business a viable option for entrepreneurs. It's important to understand the target market and the health of the hotel business to devise appropriate strategies and provide services and facilities tailored to their needs.

This project will analyze the hotel business by processing historical data to obtain information about hotel bookings, length of stay, and the time interval between booking and cancellation. In this project we will be using **Python** for main programming language with the help of several libraries such as **numpy** and **pandas** for data preprocessing while for data visualization will be done with the help of **matplotlib** and **seaborn**.

Data Description

Rakamin Academy

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.

Feature Information :

- hotel : Hotel (Resort Hotel or City Hotel)
- 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 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

For more details, you can view the Jupyter notebook here.

Data Description

- Stays_in_weekend_nights : Number of weeked nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- Stays_in_weekdays_nights : Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- Adults : number of adults
- Children : number of children
- Babies : number of babies
- Meal : type of meal booked
- City : city of origin
- Market_segment : Market segment designation. Context : the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- Distribution_channel : ooking distribution channel.
- Is_prepeated_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

For more details, you can view the Jupyter notebook here.

Data Description

- Previous_bookings_not_cancelled : 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 categories:

– 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

- Agent : ID of the travel agency that made the booking
- Company : ID of the company that made the booking
- Days_in_waiting_list : Number of days the booking was in the waiting list before it was confirmed to the customer

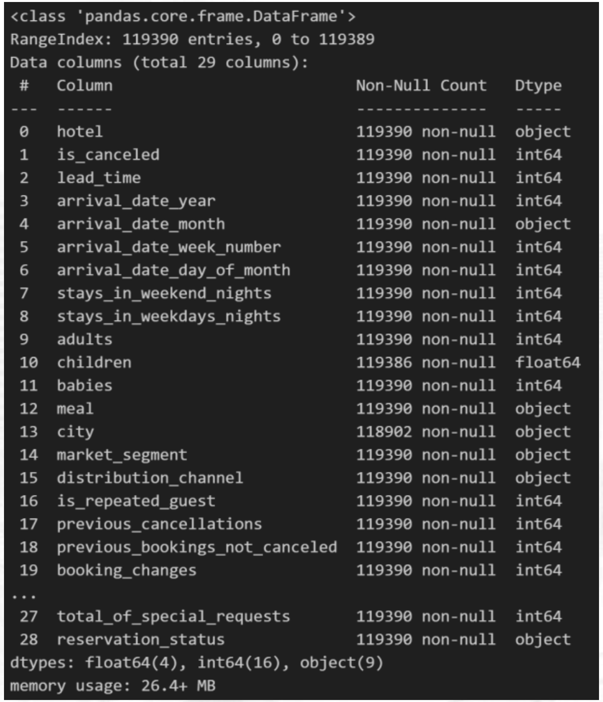
For more details, you can view the Jupyter notebook here.

Data Description

- 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_space : Number of car parking spaces required by the customer
- Total_of_special_request : Number of special request made by the customer
- 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

For more details, you can view the Jupyter notebook here.

Data Cleansing and Preprocessing



```
<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  
...
 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
```

Dataset Information :

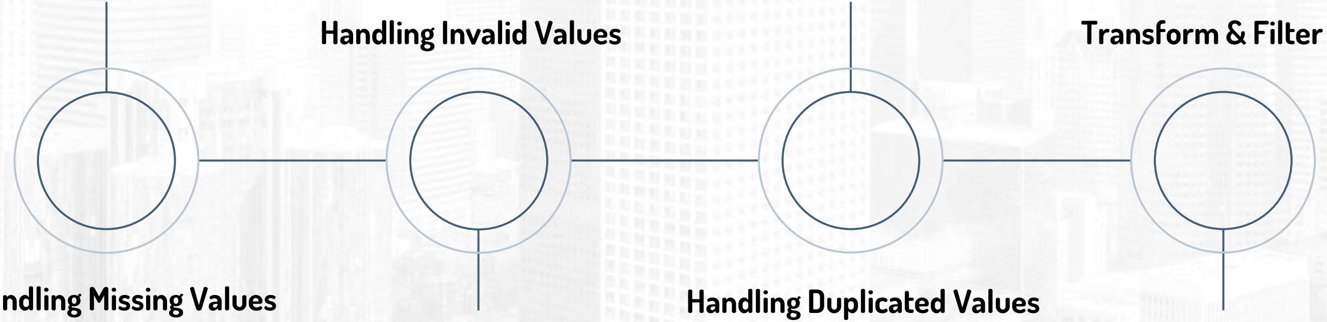
The dataset comprises 29 columns and 119,390 rows of data, featuring three types of data: int64, object, and float64. Notably, several columns contain missing values :

- Company : 94% null values, affecting 112,593 rows.
- agent : 13% null values, impacting 16,340 rows.
- City : 0.4% null values, affecting 488 rows.
- Children : contains 0.003% null values, impacting 4 rows.

For more details, you can view the Jupyter notebook here.

Data Cleansing and Preprocessing

Rakamin Academy



Handling Missing Values

- Imputation with 0 for company, agent, and children
- Imputation with mode for city

Handling Invalid Values

- Transform for ‘undefined’ values in meal, market_segment, and distribution_channel
- meal → No Meal, market_segment → Online TA, distribution_channel → TA/TO

Handling Duplicated Values

- Duplicate rows will not be dropped due to not having the unique identifier that we can rely on like customer ID or order ID

Transform & Filter Data

- Filtering by total guest (number of guest <=0), stay_duration (night duration < 0), and ADR < 0 (drop due to calculation error)

For more details, you can view the Jupyter notebook here.