ML PROJECT

Hotel Bookings

Exploratory Data Analysis and Predictive Modeling of Hotel Booking Data

Understanding Hotel Booking Trends

Analyzing **historical booking data** reveals important trends that shape guest preferences and behaviors during peak seasons and for different types of accommodations.

Data Collection and Preparation

Proper data collection and **cleaning processes** are crucial for ensuring that the analysis accurately reflects the underlying trends in hotel bookings and guest demographics.

Key Variables

- hotel: The category of hotels, which are two resort hotel and city hotel.
- is_cancelled : The value of column show the cancellation type. If the booking was cancelled or not.
- lead_time : The time between reservation and actual arrival.
- stayed_in_weekend_nights: The number of weekend nights stay per reservation
- stayed_in_weekday_nights: The number of weekday nights stay per reservation.
- meal: Meal preferences per reservation.[BB, FB, HB, SC, Undefined]
  BB = Bed & Breakfast, HB = Half Board, FB=Full Board, SC = Self Catering

Key Variables

- Country: The origin country of guest.
- market_segment: This shows how reservation was made and what is the purpose of reservation. Eg, corporate means corporate trip, TA for travel agency.
- distribution_channel: The medium through booking was made. [Direct,Corporate,TA/TO,undefined,GDS.]
- Is_repeated_guest: Shows if the guest is who has arrived earlier or not.
- days_in_waiting_list: Number of days between actual booking and transact.
- customer_type: Type of customers(Transient, group, etc.)

Data Cleaning

Data cleaning is a crucial step before performing Exploratory Data Analysis (EDA), as it removes ambiguous and inconsistent data that could negatively affect the analysis results.

* Removing duplicate rows

* Converting columns to appropriate data types
  Changing datatype of column 'reservation_status_date' to date_time.

* Adding important or derived columns
  total people=adult+children+babies

Exploratory Data Analysis

It’s the process of looking at any raw data using charts, graphs, and summary statistics to understand what’s happening before applying any math or AI.

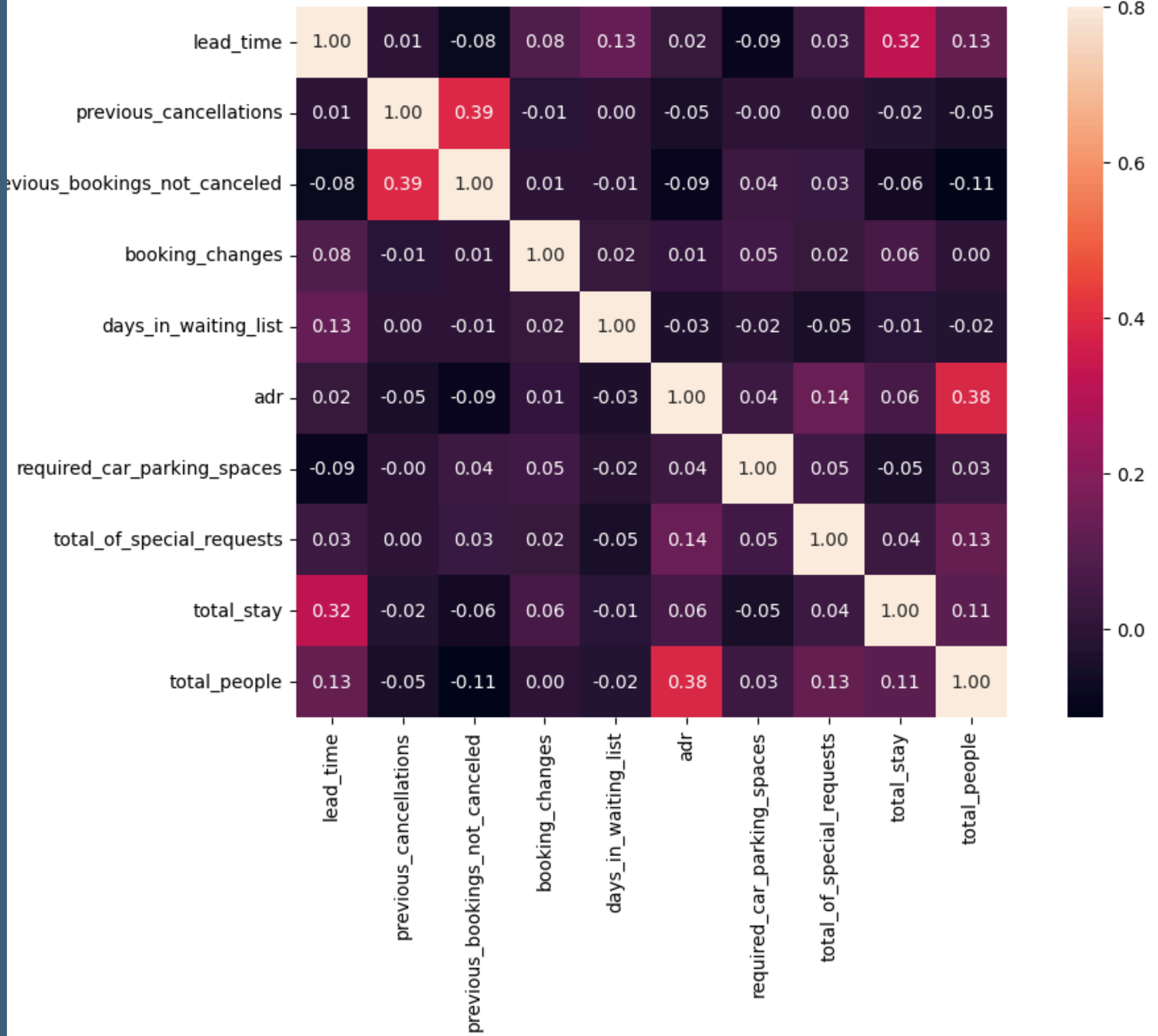
Heat Map

These are the findings:

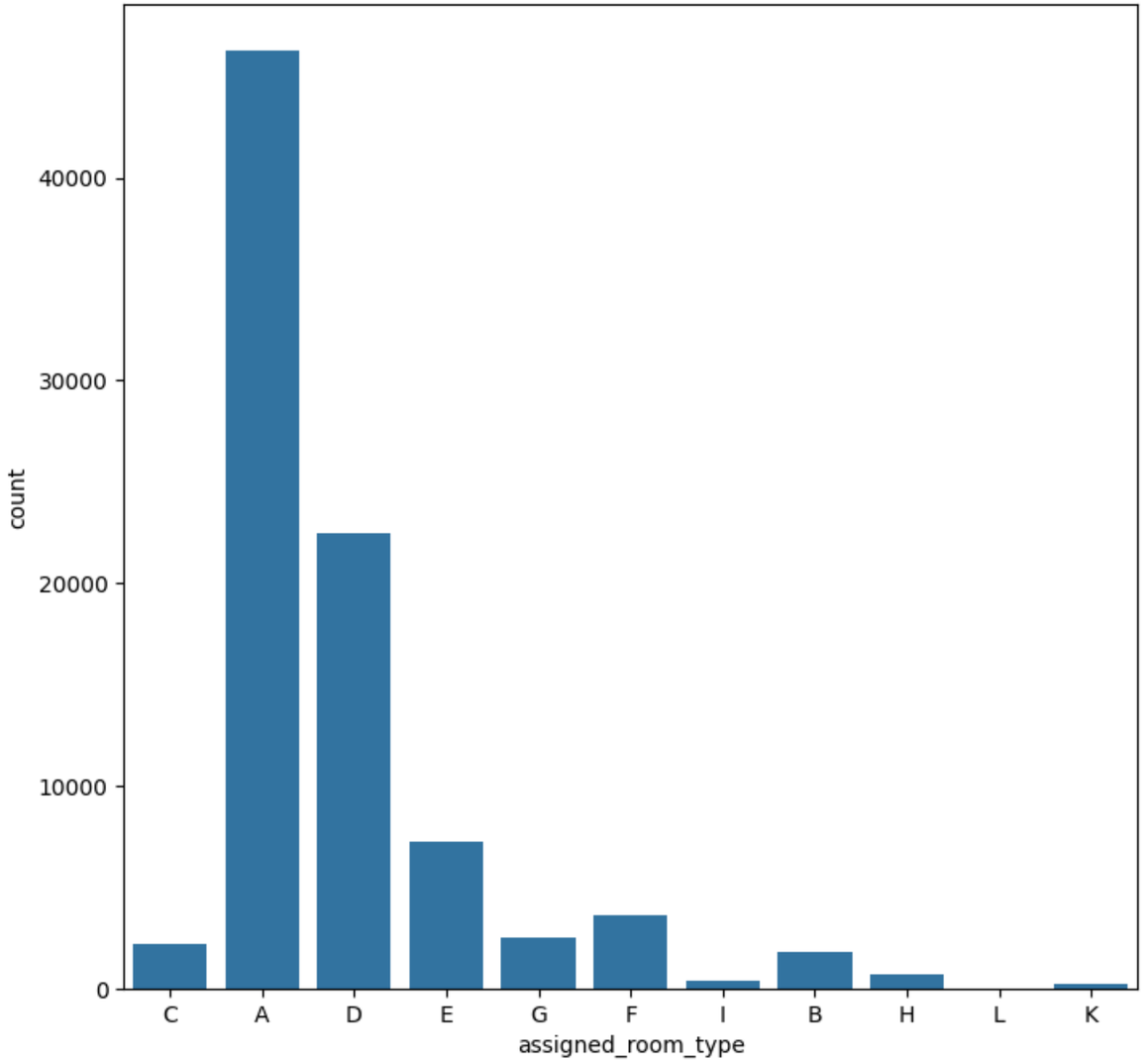
1. lead time is slightly correlated to total_stay that means for longer vacation people usually book ahead of time.

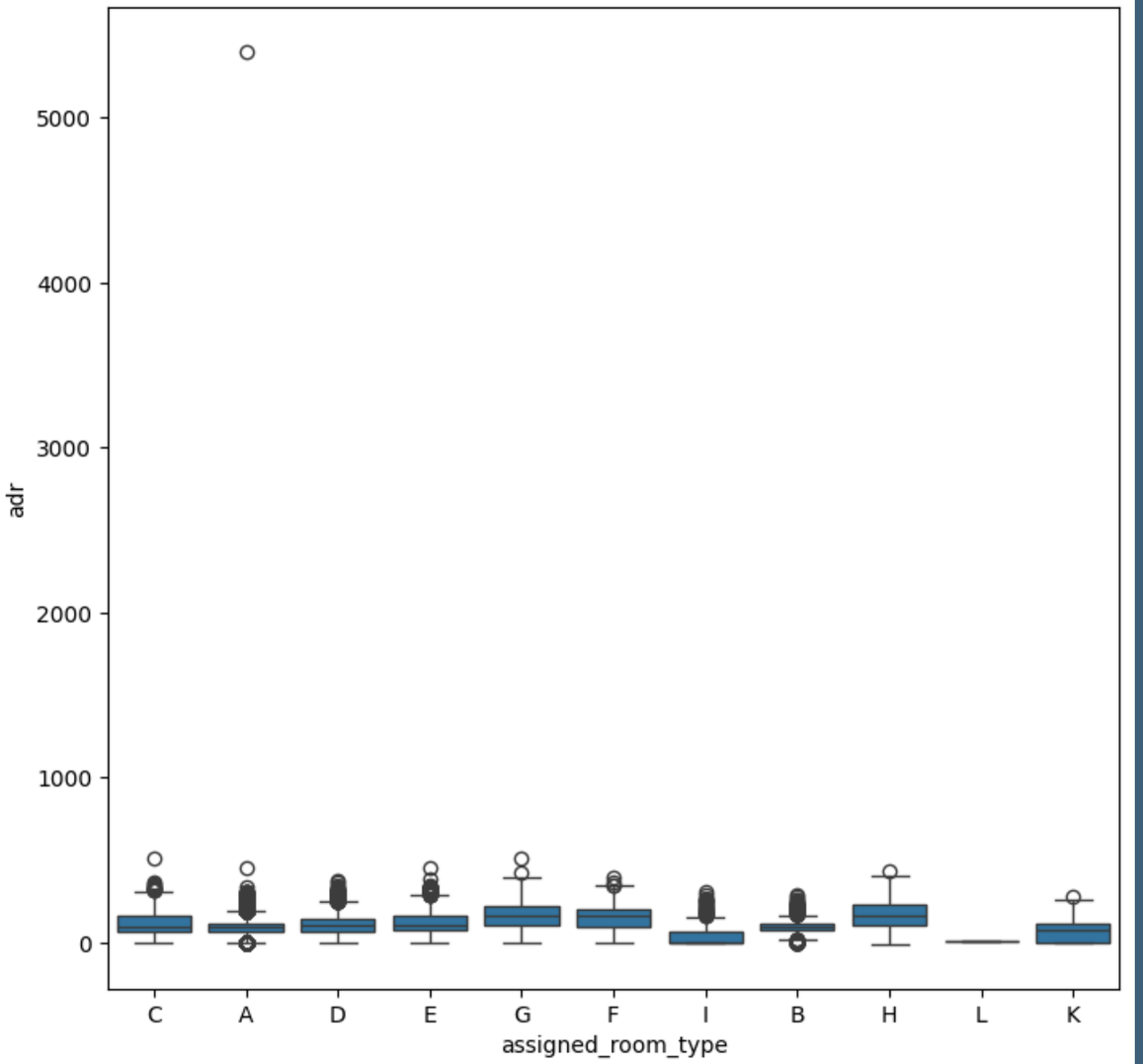
2. adr and total_people are also correlated as more number of people generates more revenue.

3. previous_cancellation has slightly correlation with previous_booking_not_canceled that means customer may have higher booking rate also have higher cancellation too.



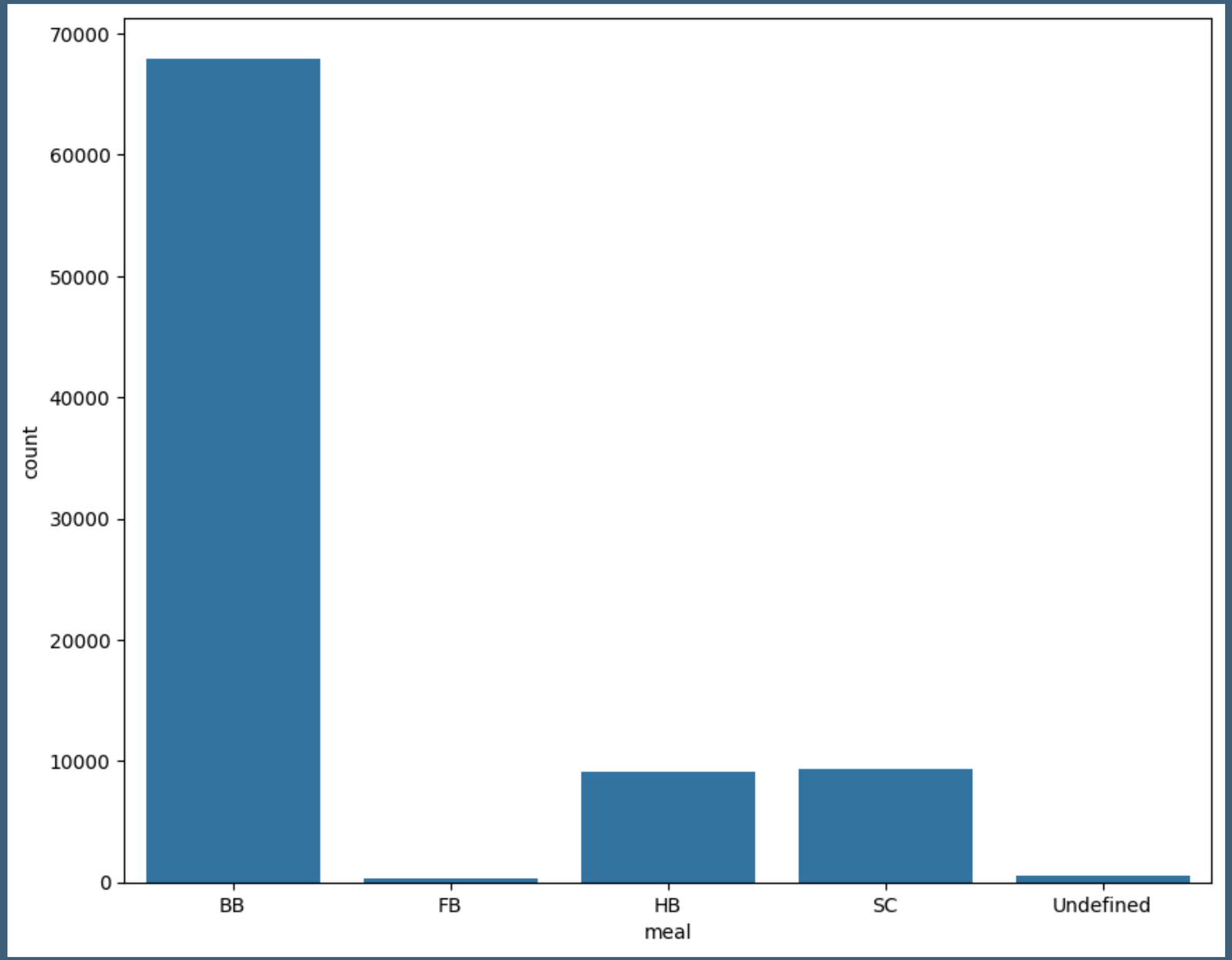
Count plot and box plot





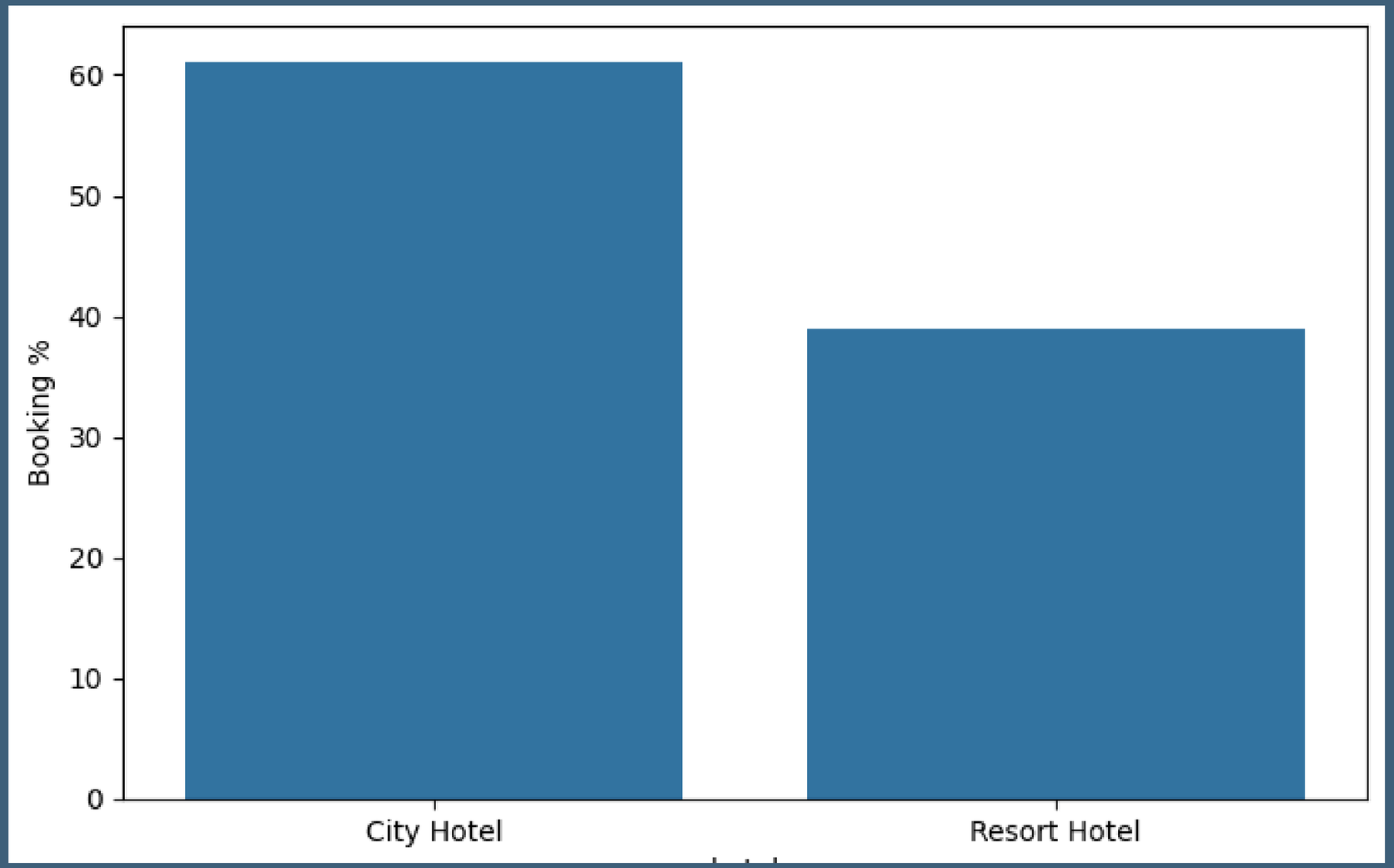
Univariate Analysis

Most preferred meal type is
BB(Bead and Breakfast)



EDA

City hotel gets around 60% and Resort hotel gets around 40% of booking

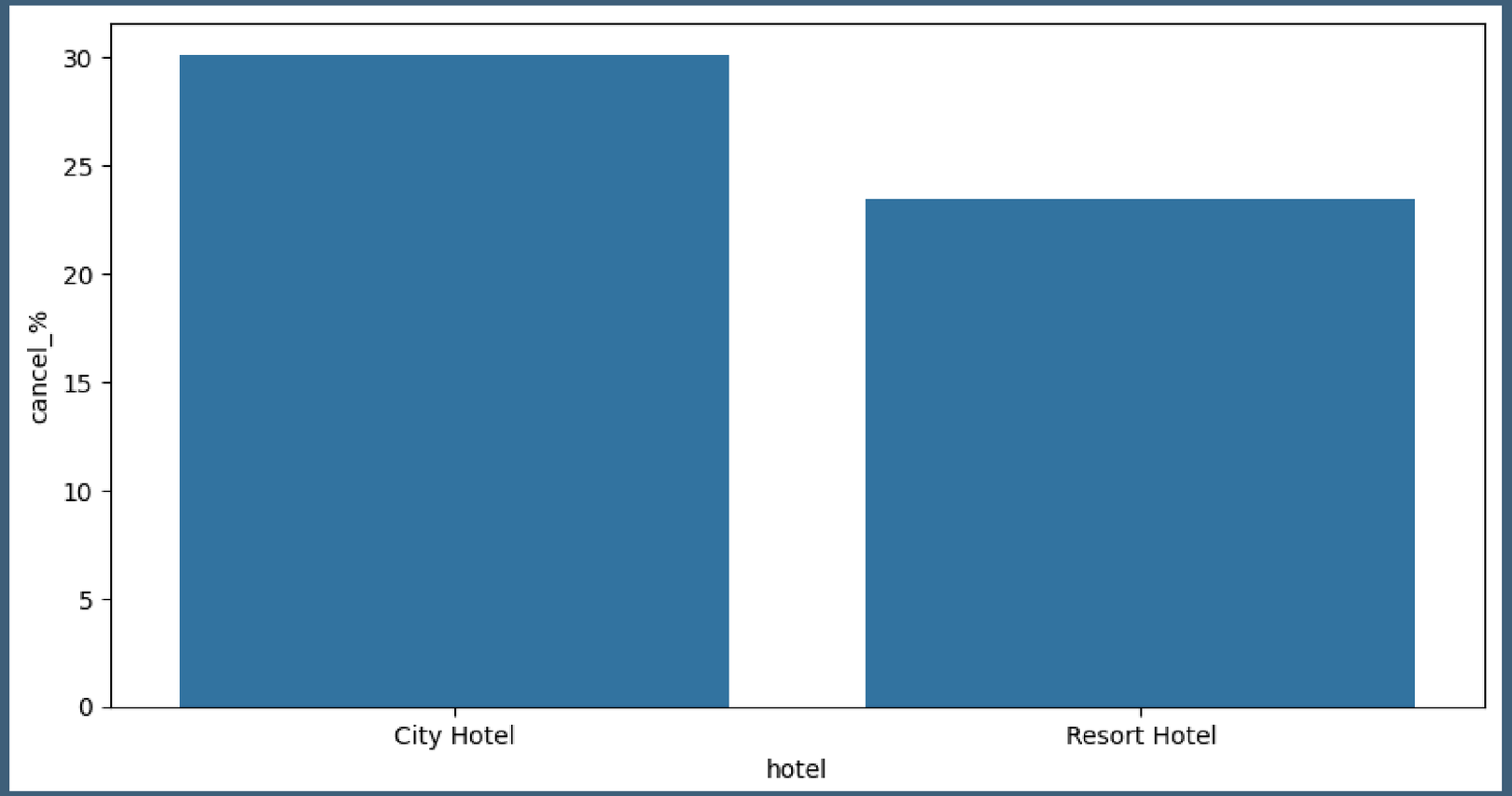


EDA

This shows City hotel is making more revenue compared to Resort hotel

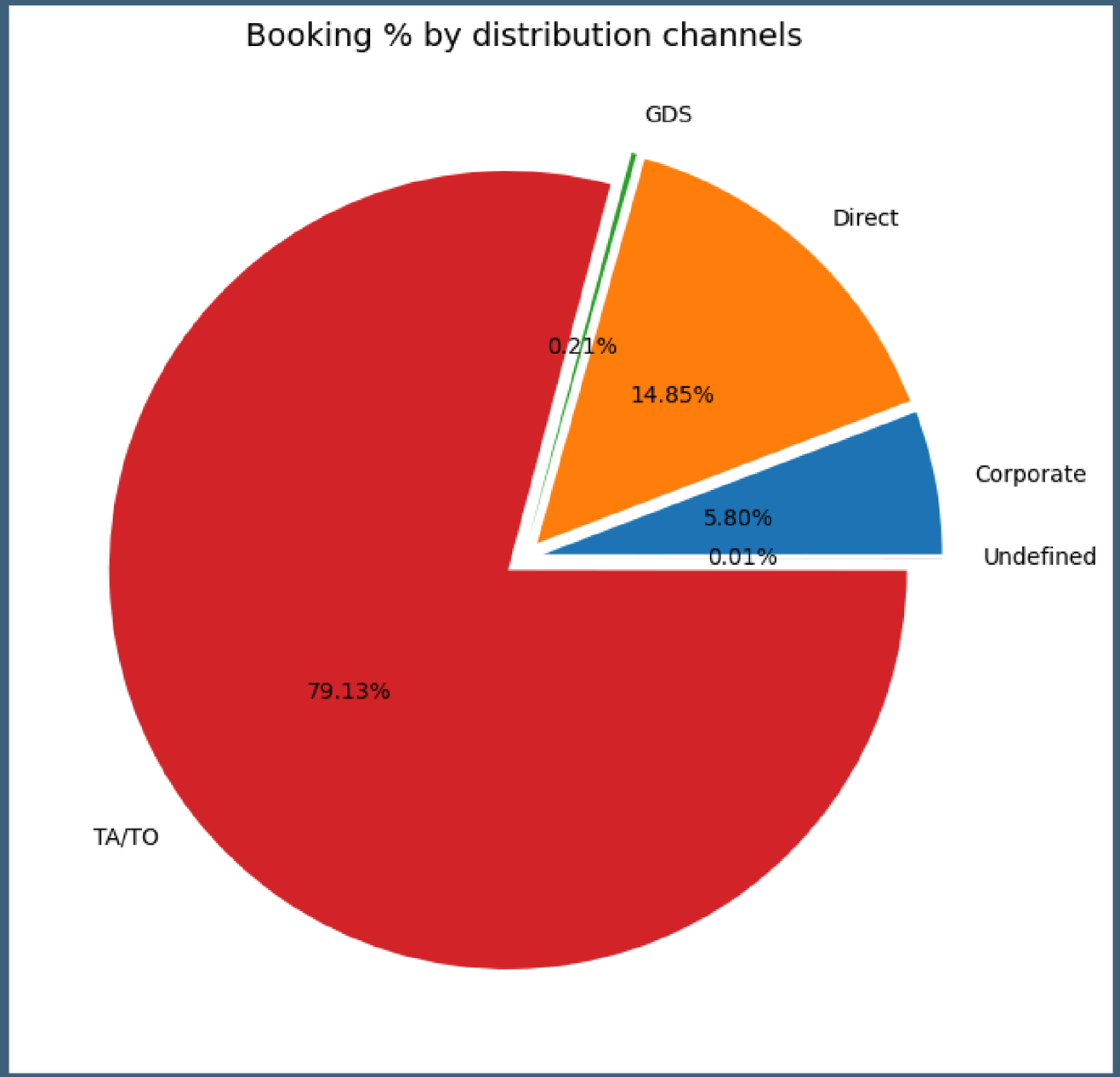
EDA

Almost 30 % of City Hotel bookings got canceled.



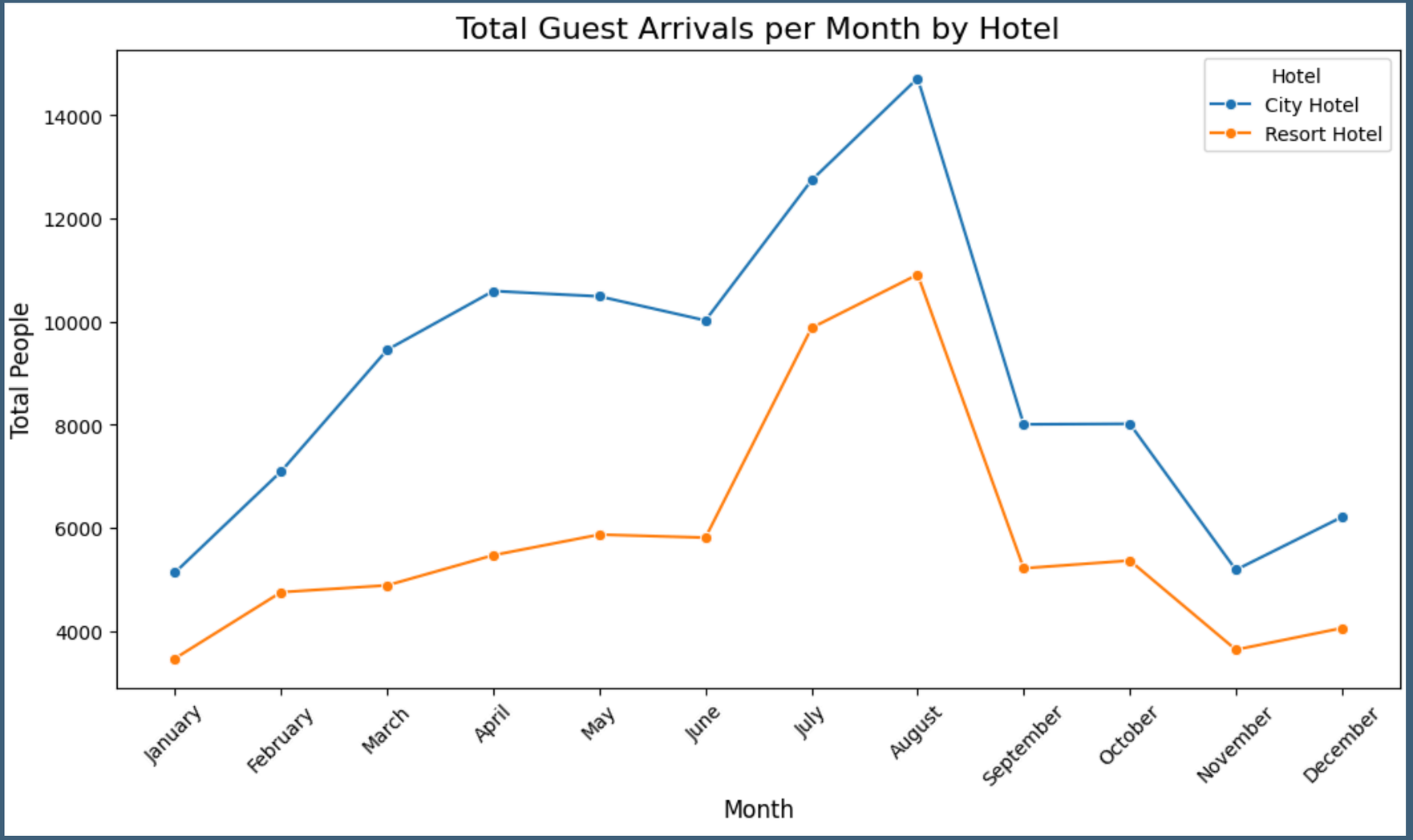
EDA

Which is the most common channel for booking hotels?



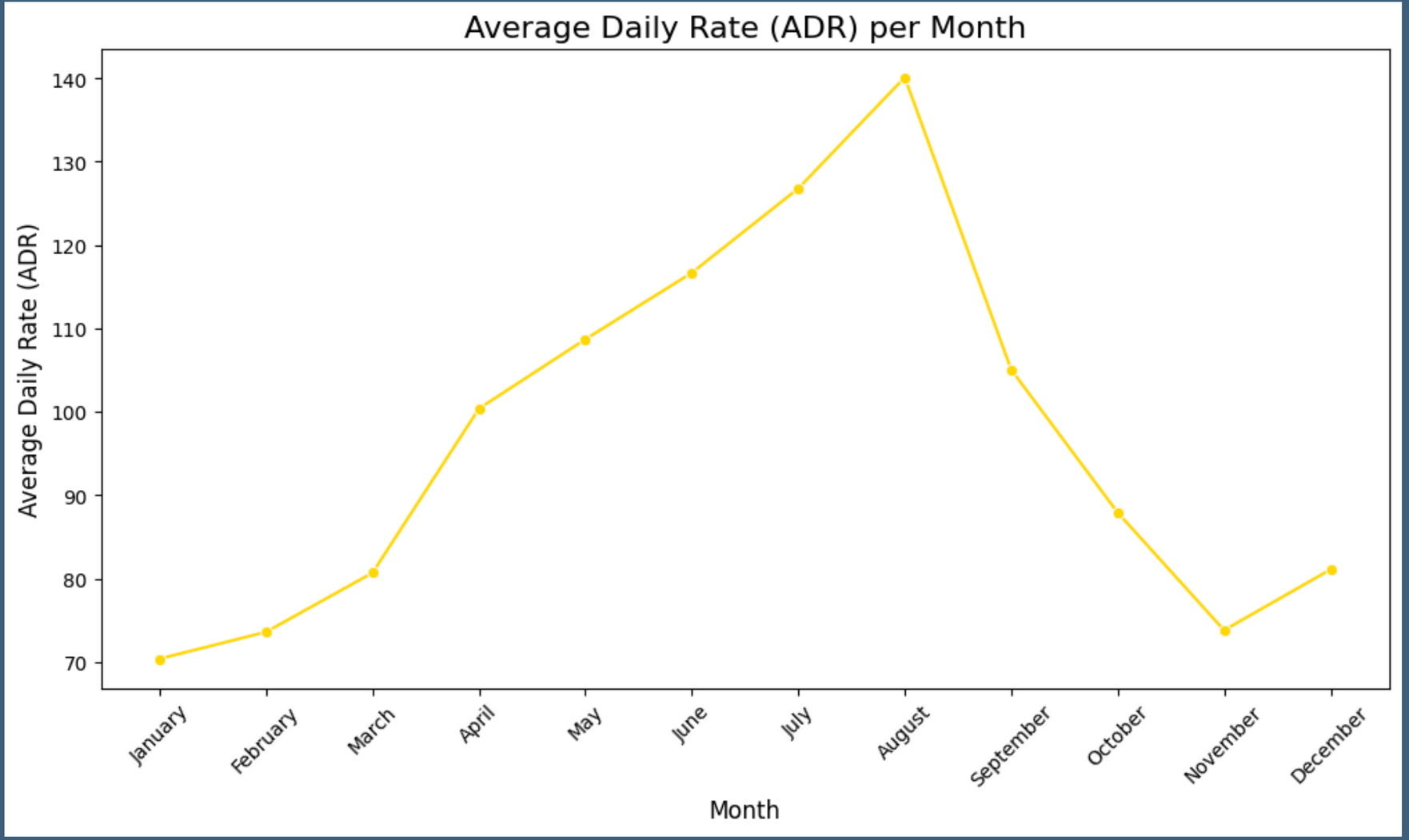
EDA

July and August are the peak months.



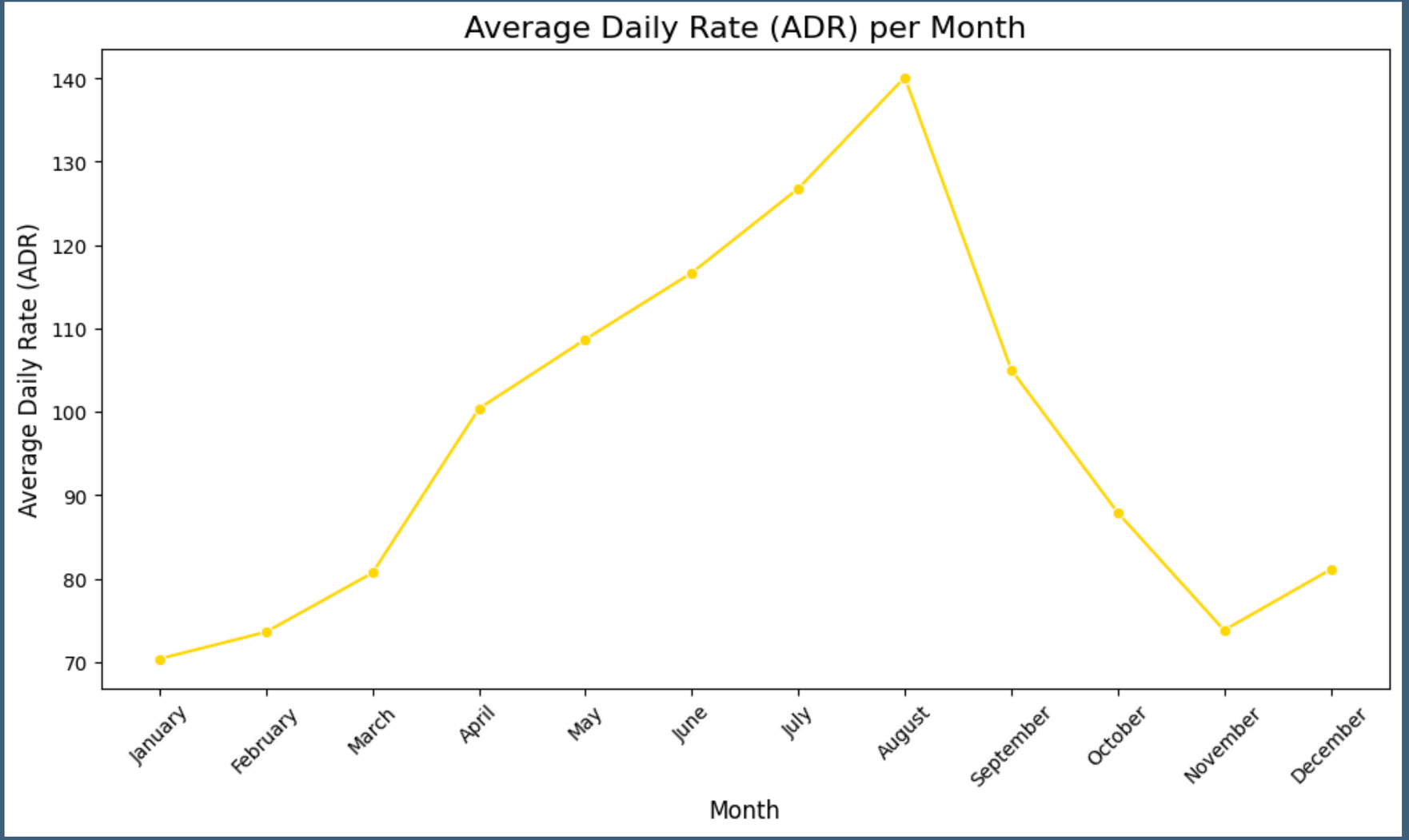
EDA

It also show during july and August adr is high as visitors are also high



EDA

It also show during july and August adr is high as visitors are also high



IMPUTATION

This dataset had several “holes“ where information was missing, specifically in columns like children, agent, and company.

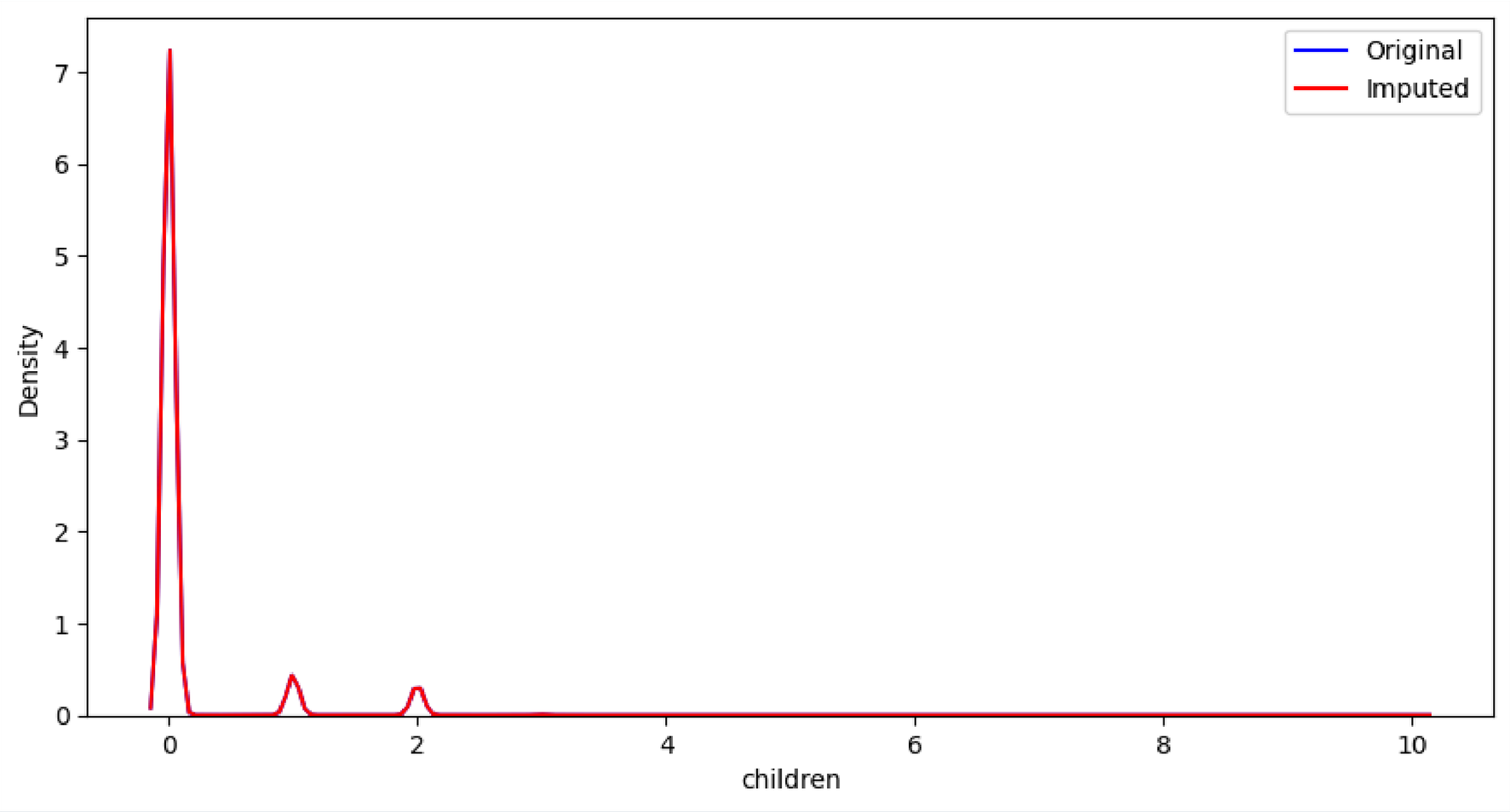
Divided data into 80% Training and 20% Testing sets using train test split method.

Strategies:

- Median: Filled numerical gaps (Children, People) to minimize outlier impact.
- Mode: Used the “Most Frequent“ value for categorical gaps (Agent, Country).
- Constant: Filled ’Company’ gaps with 0 (representing “No Company“).
- ColumnTransformer: Applied specific rules to specific columns simultaneously for cleaner code.

Density vs children

The overlapping curves confirm that our imputation preserved the data's original distribution without introducing bias or distorting the results.



Encoding

- One-Hot Encoding: Turned unordered text (Hotel, Meal, Market) into binary columns (0 and 1).
- Ordinal Encoding: Ranked “Deposit Type“ to show the level of financial commitment.
- Leakage Prevention: Removed “Reservation Status“ to prevent the model from “cheating.“
- Data Consolidation: Merged numerical and encoded features into a single high-performance dataset.

Scaling

- Standardization: Scales values to a mean of 0 and variance of 1.
- Balance: Ensures large values (Price) don’t dominate small values (Children).
- Efficiency: Accelerates model training and improves accuracy.
- Consistency: Applies identical scaling logic to both Train and Test sets.

Model Training & Evaluation

Utilized Random Forest, an ensemble of 100 decision trees for robust classification.

Accuracy: 0.8523443769345409

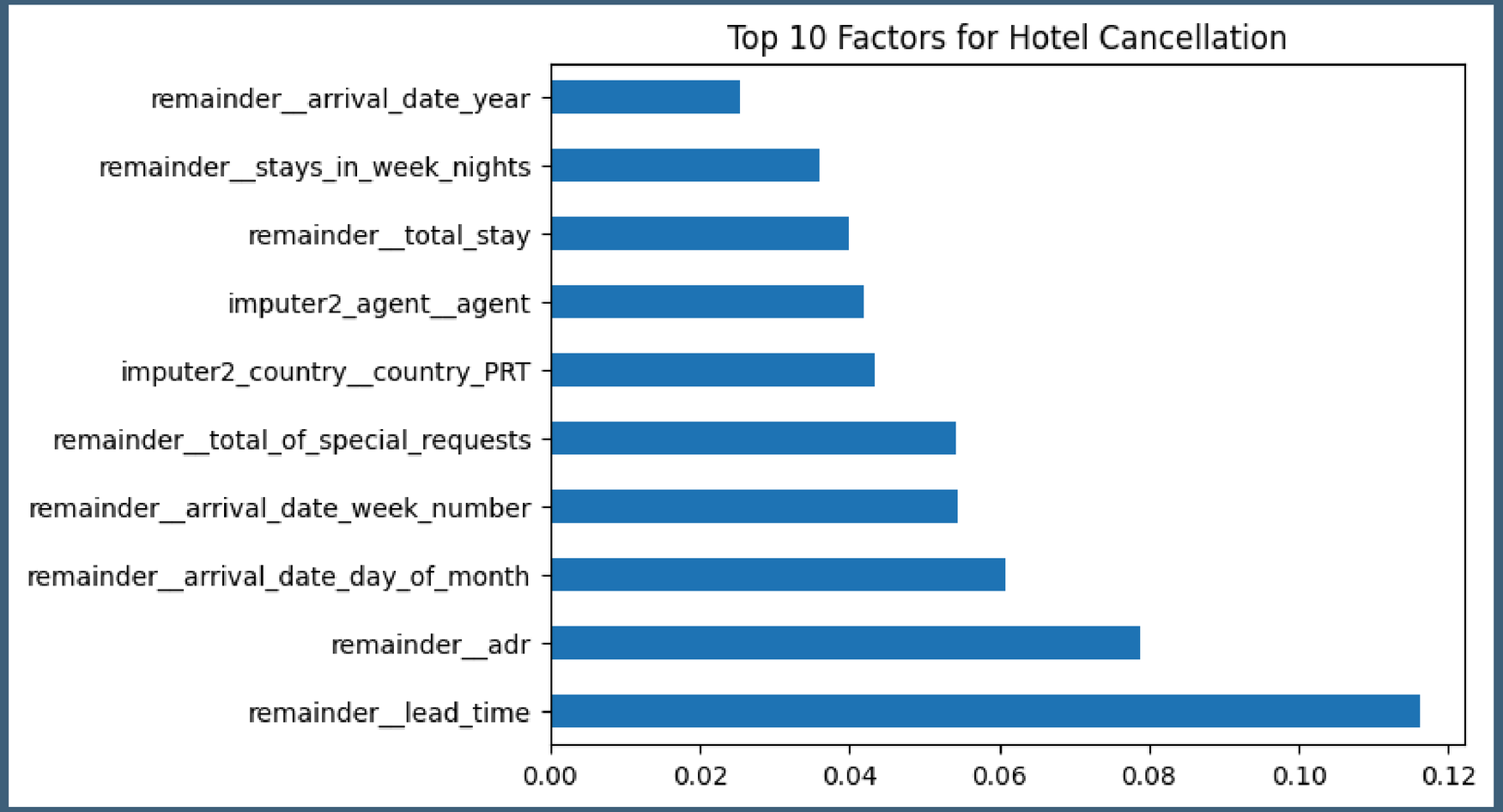
	precision	recall	f1-score	support
0	0.87	0.93	0.90	12728
1	0.78	0.63	0.70	4718
accuracy			0.85	17446
macro avg	0.83	0.78	0.80	17446
weighted avg	0.85	0.85	0.85	17446

- Accuracy (85%): Correct overall prediction rate for all bookings.
- Precision (78%): Reliability of “Cancellation“ alerts; few false alarms.
- Recall (63%): Ability to capture guests who actually cancel.
- Class 0 Performance: 93% success in identifying guests who will stay.

FINAL EVALUATION

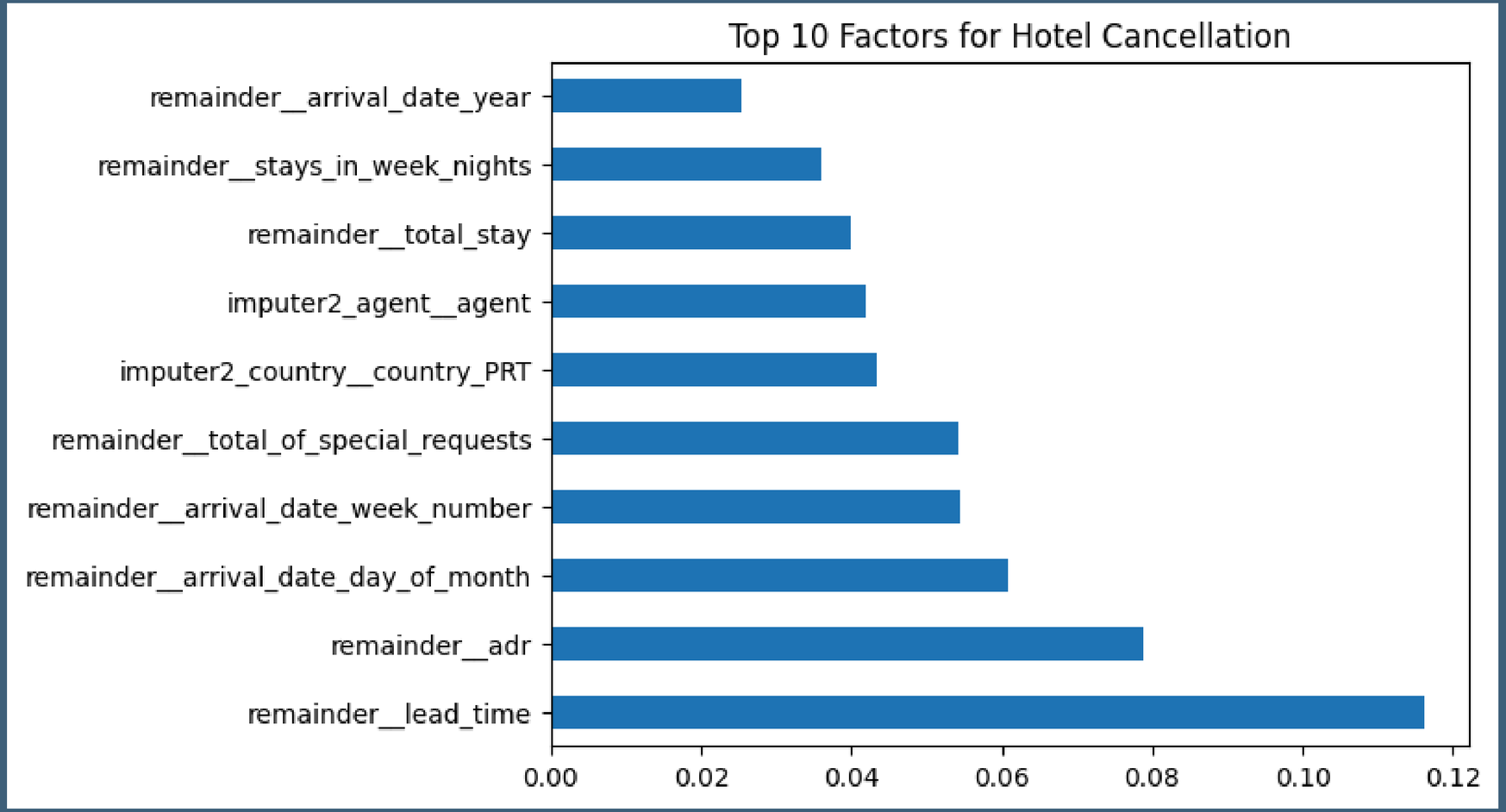
Factors affecting Cancellation of Hotel Booking:

1. Lead Time: Guests who book a long time in advance are much more likely to cancel or have their plans change compared to last-minute bookers..



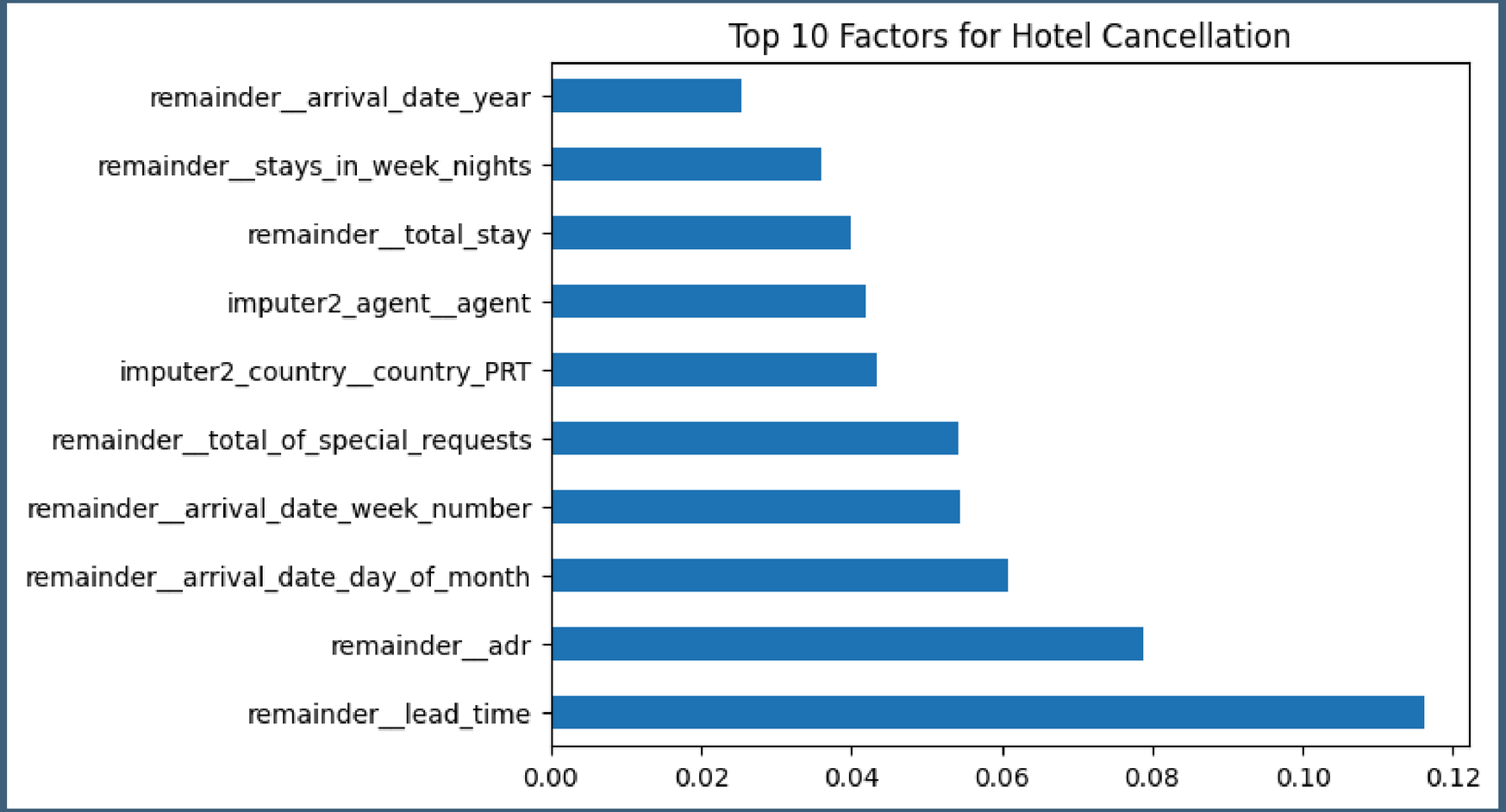
FINAL EVALUATION

2.ADR (Average Daily Rate):Higher-priced rooms often see different cancellation patterns.



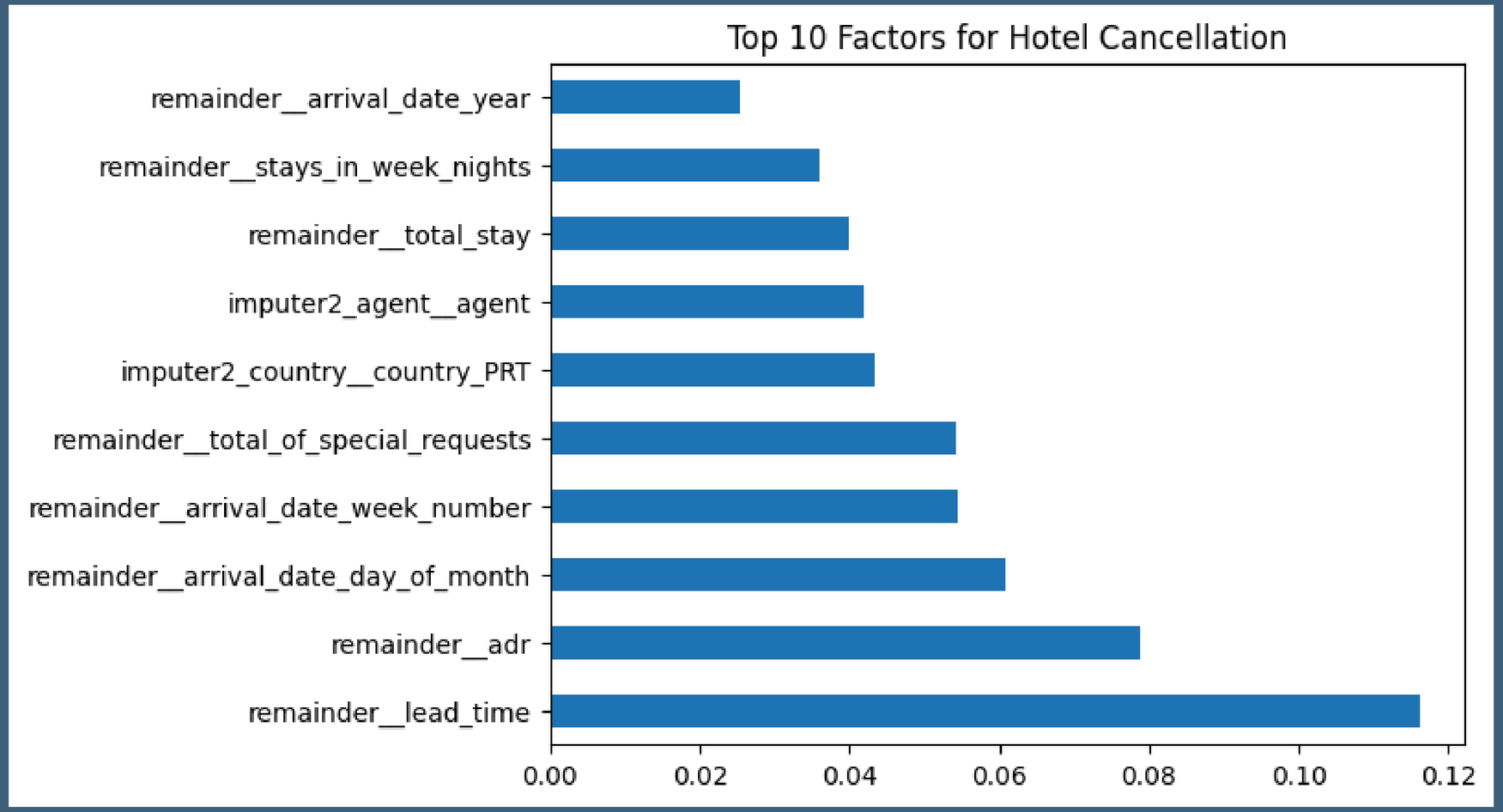
FINAL EVALUATION

3.Total Special Requests: Interestingly, guests who ask for specific things (high floor, extra pillows) are usually more “invested“ in their stay and less likely to cancel.



FINAL EVALUATION

4.Country_PRT: Local vs. International travel patterns are showing up as a significant indicator as this hotel lies on portugal.



ANY QUERIES?

PRESENTED BY:

ASBIN BHATTARAI