EDA Capstone Project-1

Hotel Booking Analysis

Individual Contributor

Anant Alok

Content

- Problem Statement
- Workflow
- Data Summary
- Data Cleaning
- Data Manipulation
- Exploratory Data Analysis
- Observation
- Conclusion

Problem Statement

- Travel has become an integral aspect of our lives today. In this interconnected world, we find ourselves on the move for various reasons, be it for professional commitments, exploration, quality time with family, and more.
- Have you ever pondered about the ideal timing for reserving a hotel room, or the most favorable duration of stay to secure the best daily rate? Perhaps you've also considered predicting whether a hotel is likely to receive an unusually high volume of special requests. This dataset on hotel bookings can provide insights into these intriguing inquiries!

Let’s Explore and analyze the data to discover important factors that govern the bookings.

Workflow

- Data Collection and understanding
- Data Cleaning
- Data Manipulation
- Exploratory Data Analysis
- Analyzing the result

Data Summary

Dataset Name : Hotel Bookings.csv

Shape : 119390

Columns : 32

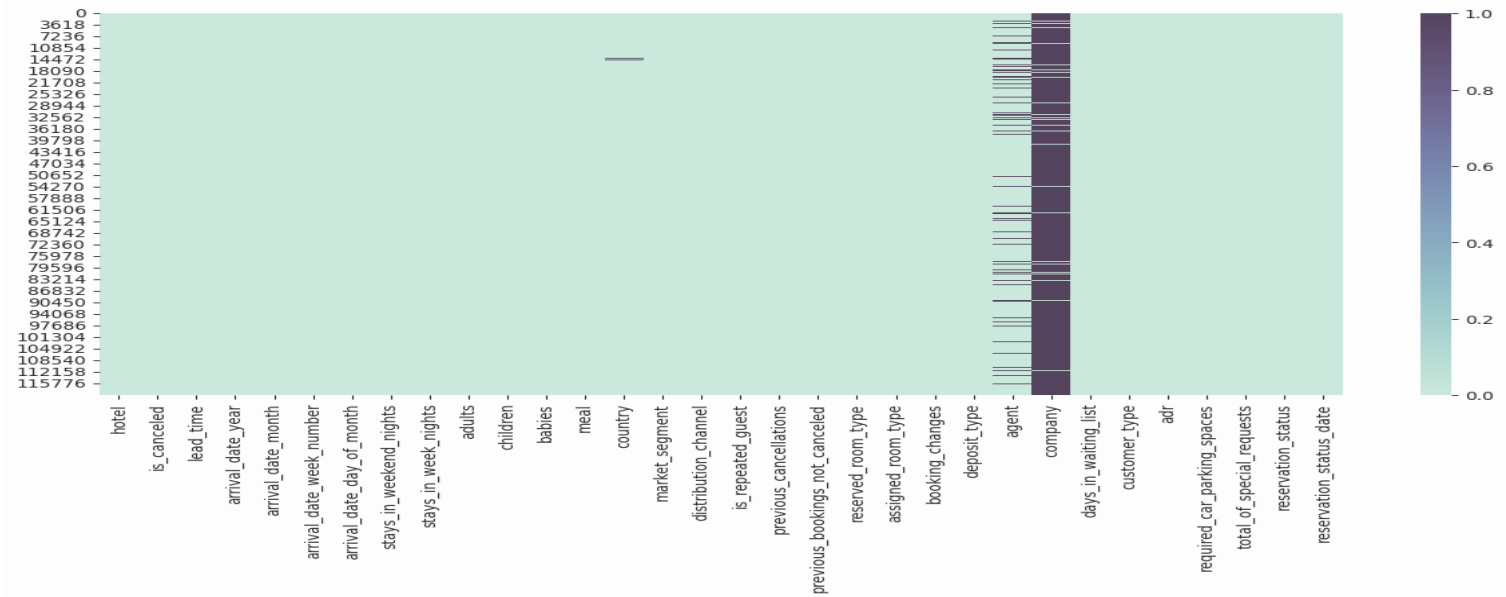
Features:

Hotel, Is_cancelled, Arrival_date_year, Arrival_date_month, Arrival_date_week_number, Day_of_the_month, Stays_in_weekend_nights, Stays_in_week_nights, Adults, Children, Babies, Meals, Country, Market_segment, Distribution Channel, Is_repeated_guest,Previous_cancellations, Previous_bookings_not_canceled, Reserved_room_type, Assigned_room_type, Booking_changes, Deposit_type, Agent, Lead_time, Days_in_waiting_list, Customer_type, Adr, Required_car_parking_spaces, Booking changes, Reservation_status, reservation_status_date

Data Cleaning

The dataset contains NaN values in few columns like:

- Company
- Agent
- Country
- Children



To handle the null values:

- I replace company and agent missing value with 0.
- Country value is replaced with 0 as well but then I casted it to string format.
- Children column had only 4 missing values, so replaced with rounded mean value.

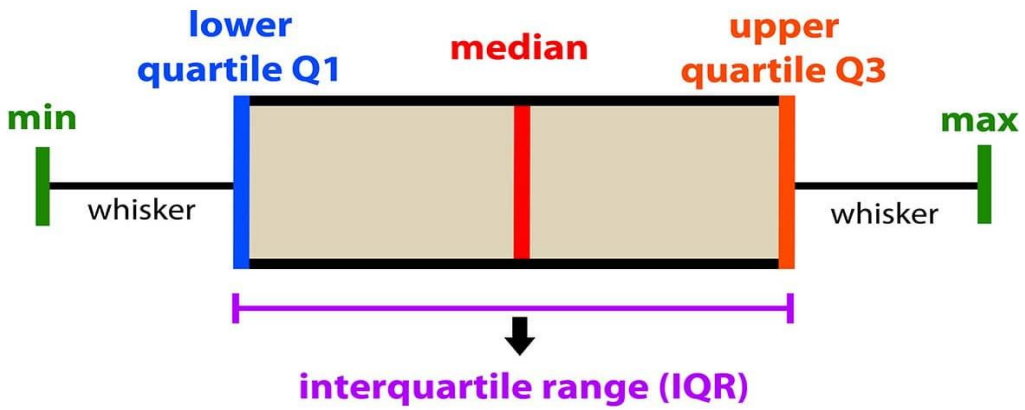
Data Cleaning

Outliers: An outlier is a data point in the dataset that is distant from all the observations.

To handle outliers I used IQR i.e. I considered all the datapoint, that lie below 1.5 times IQR from the lower quartile(Q1) and above 1.5 times IQR from the upper quartile(Q3) and replaced them with lower bound and upper bound.

Lower bound : (Q1 – 1.5* IQR)

Upper bound : (Q3 + 1.5* IQR)



Data Manipulation

Dropped rows where adults, babies and children sum up to zero because it doesn't make any sense to have them in any analysis.

Created two separate dataframe of resort and city hotels:

- resort = df[(df["hotel"] == "Resort Hotel") & (df["is_canceled"] == 0)]
- city = df[(df["hotel"] == "City Hotel") & (df["is_canceled"] == 0)]

Now let's look at some visualizations

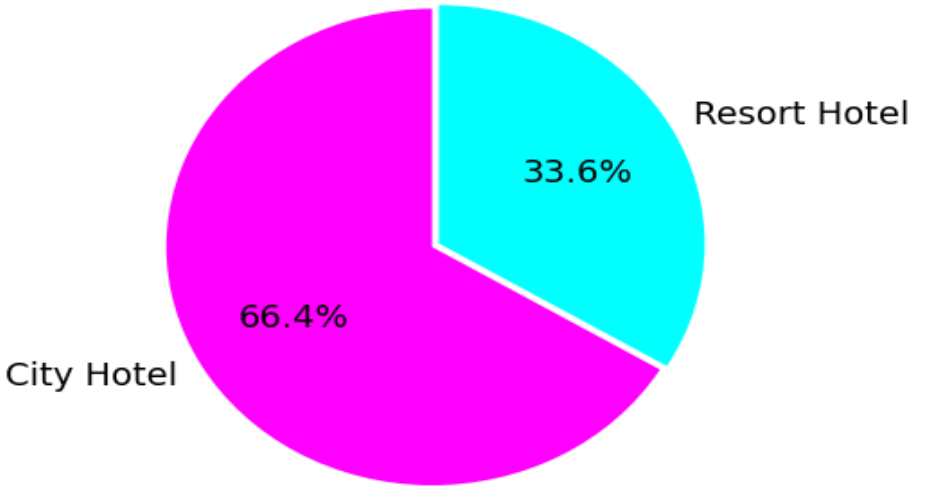
Exploratory Data Analysis

This dataset comprises details regarding reservations made at both a city hotel and a resort hotel. It encompasses information such as the booking date, duration of the stay, the count of adults, children, and infants, and the availability of parking spaces, among other data points.

I have conducted analyses for both city hotels and resort hotels, both individually and collectively.

Nearly two-thirds of people show a preference for

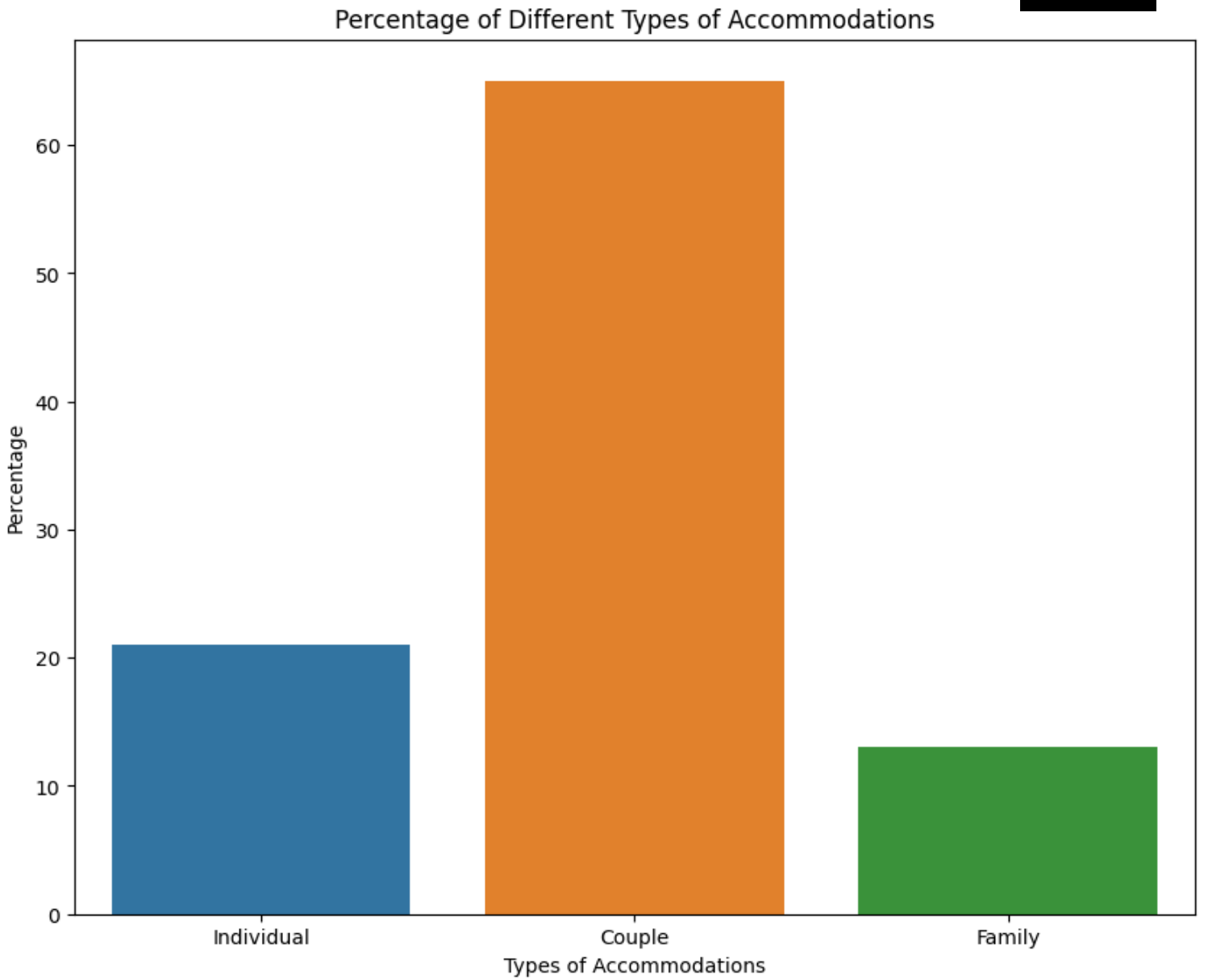
City hotels over resort hotels.



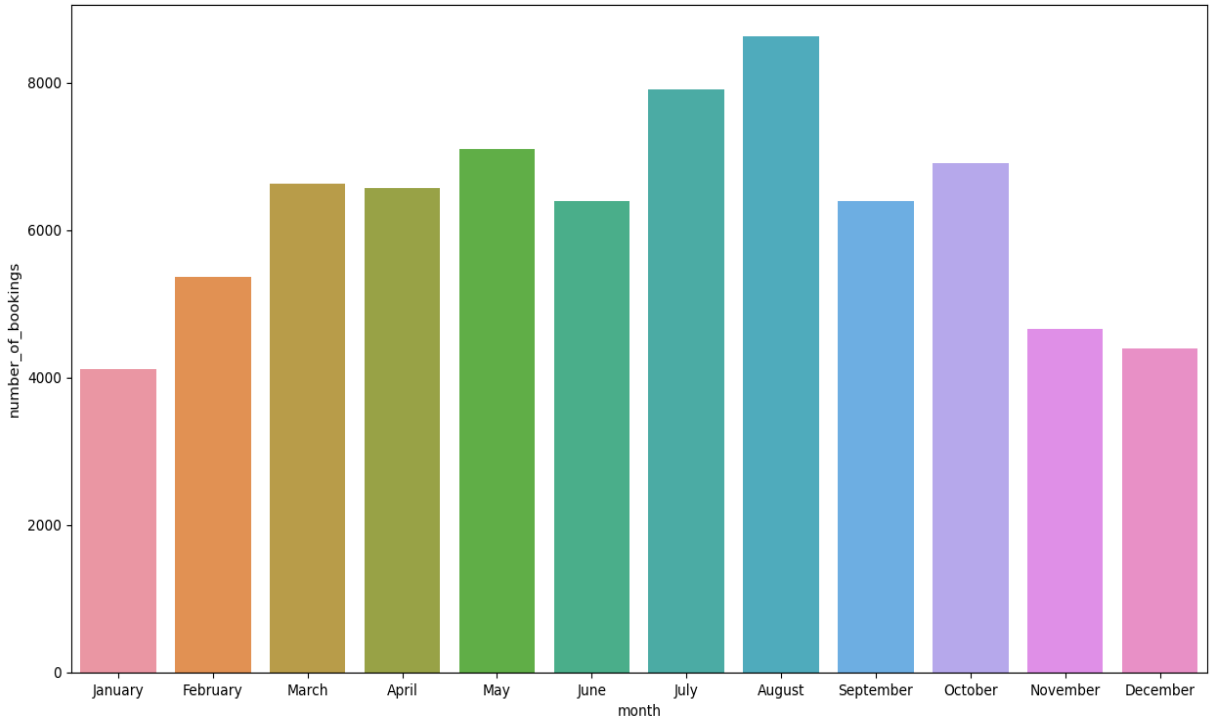
Exploratory Data Analysis

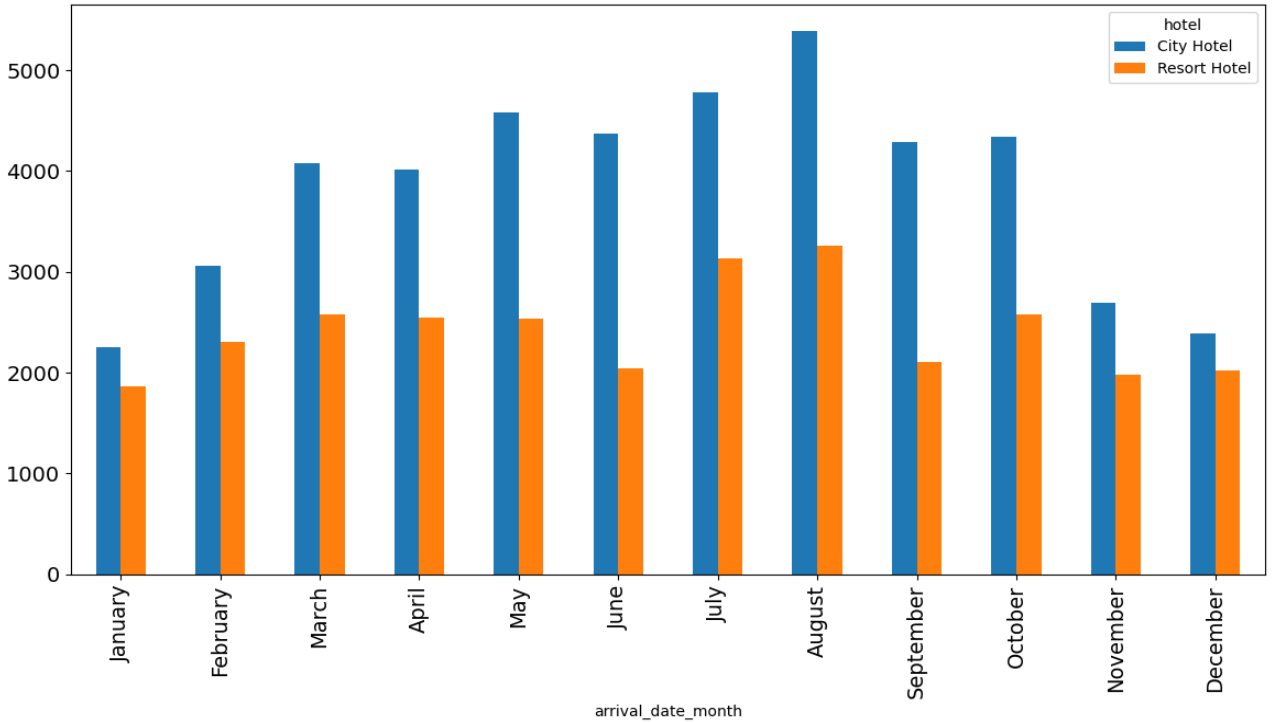
Types of Accommodations

- Couples account for over 60% of hotel bookings, making them the largest contributing group in terms of percentage.
- Families make the least number of bookings, just slightly below 15%.
- A little over 20% of bookings are made by individuals.



Exploratory Data Analysis

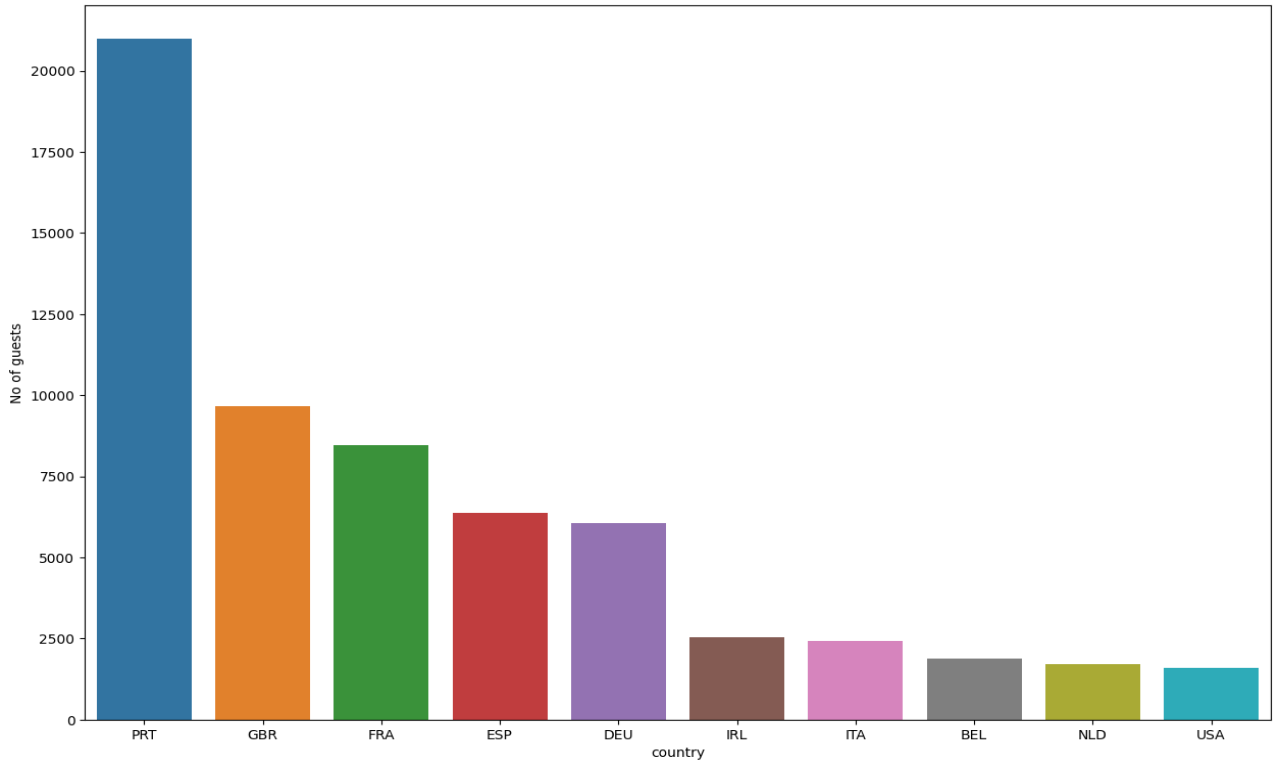


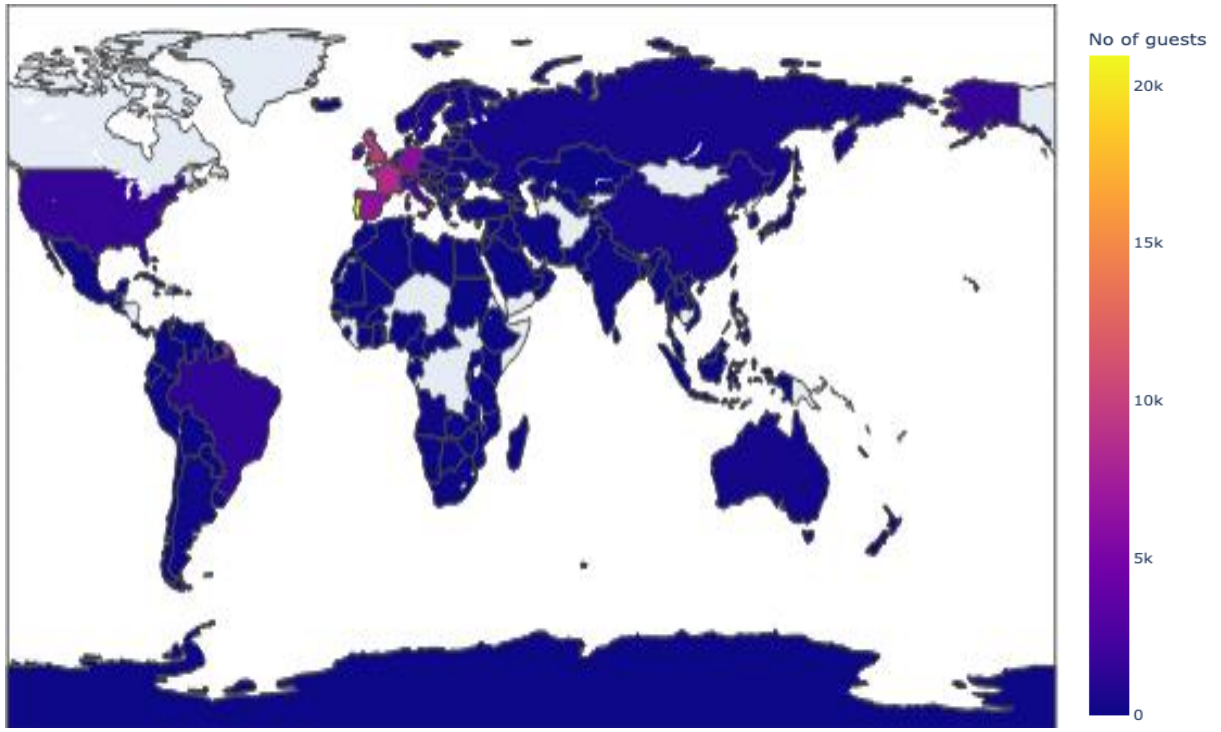


- July and August witness the highest number of bookings, while January and December have the lowest booking rates.

- City hotel bookings exceed those of resort hotels in every month.

Exploratory Data Analysis

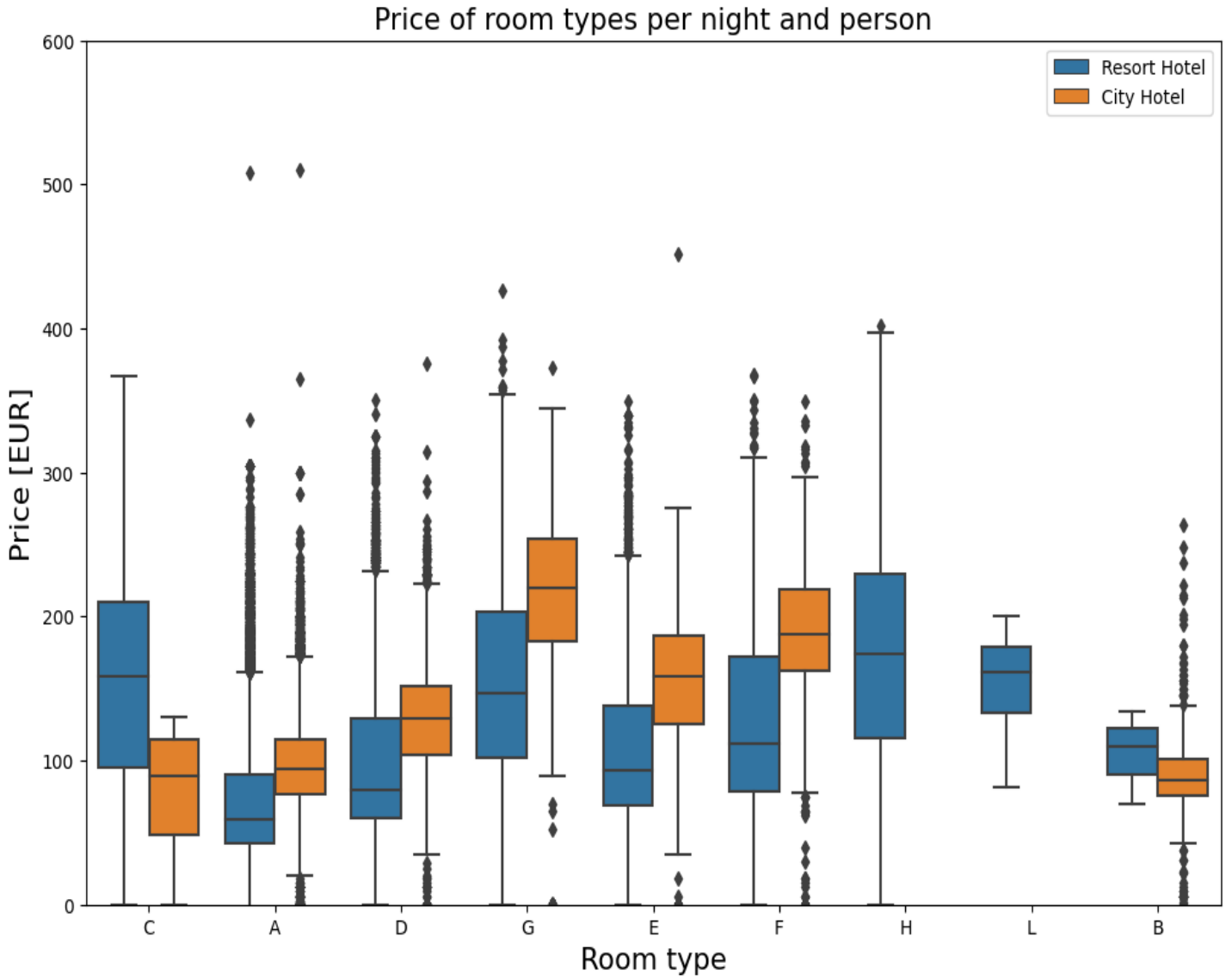




- The majority of guests originate from Portugal, with Great Britain and France following closely behind in terms of guest numbers.
- Demographically, it's evident that a significant proportion of the guests hail from European countries.

Exploratory Data Analysis

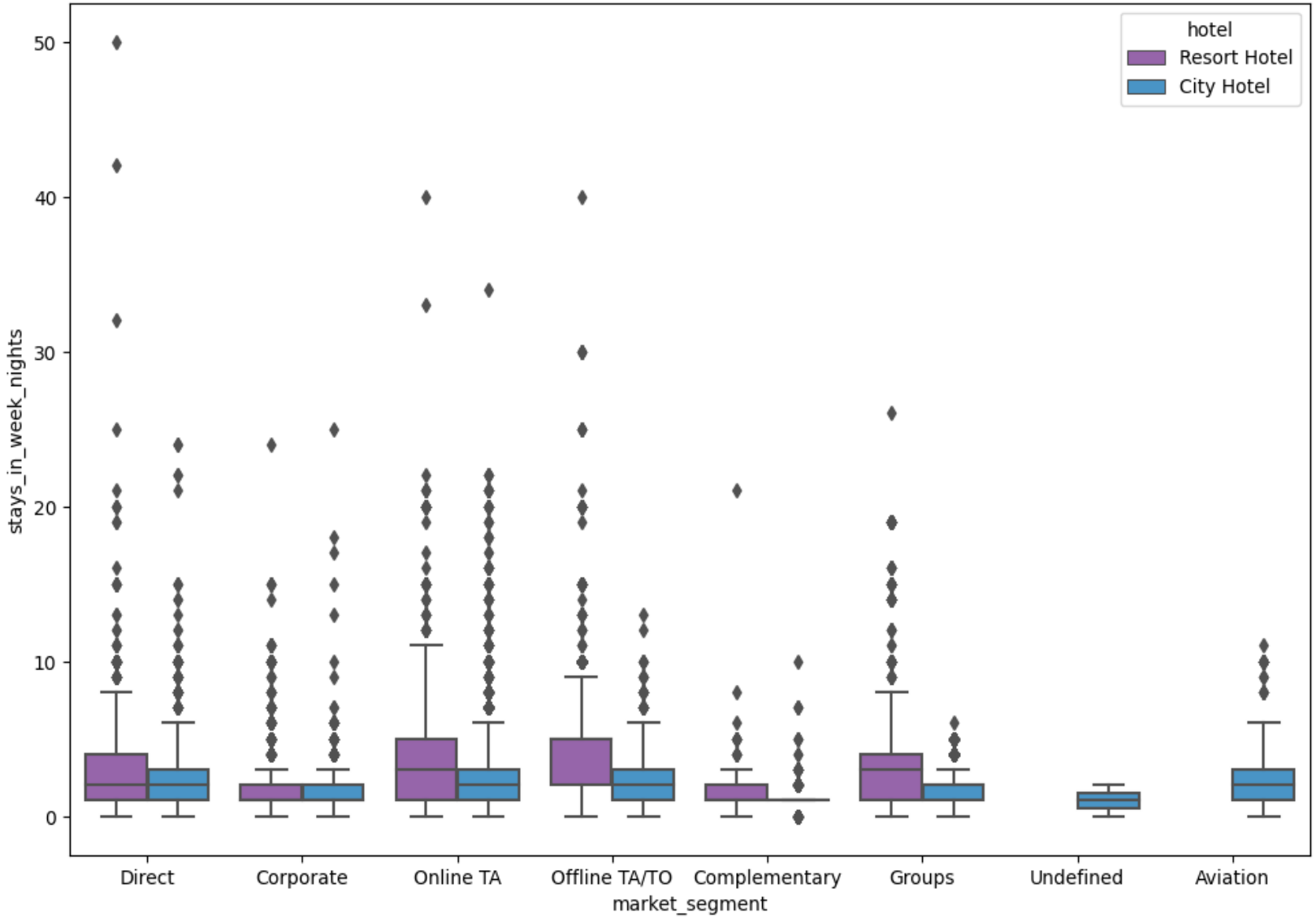
- The boxplot indicates substantial price distribution variations across various room types and hotels.
- In both hotel categories, the P type stands out as the most expensive room option, closely followed by the L type. Conversely, the G type is consistently the most affordable room choice in both hotel types.



Exploratory Data Analysis

- The chart provides insights into the length of stay in nights across different market segments and hotel types.

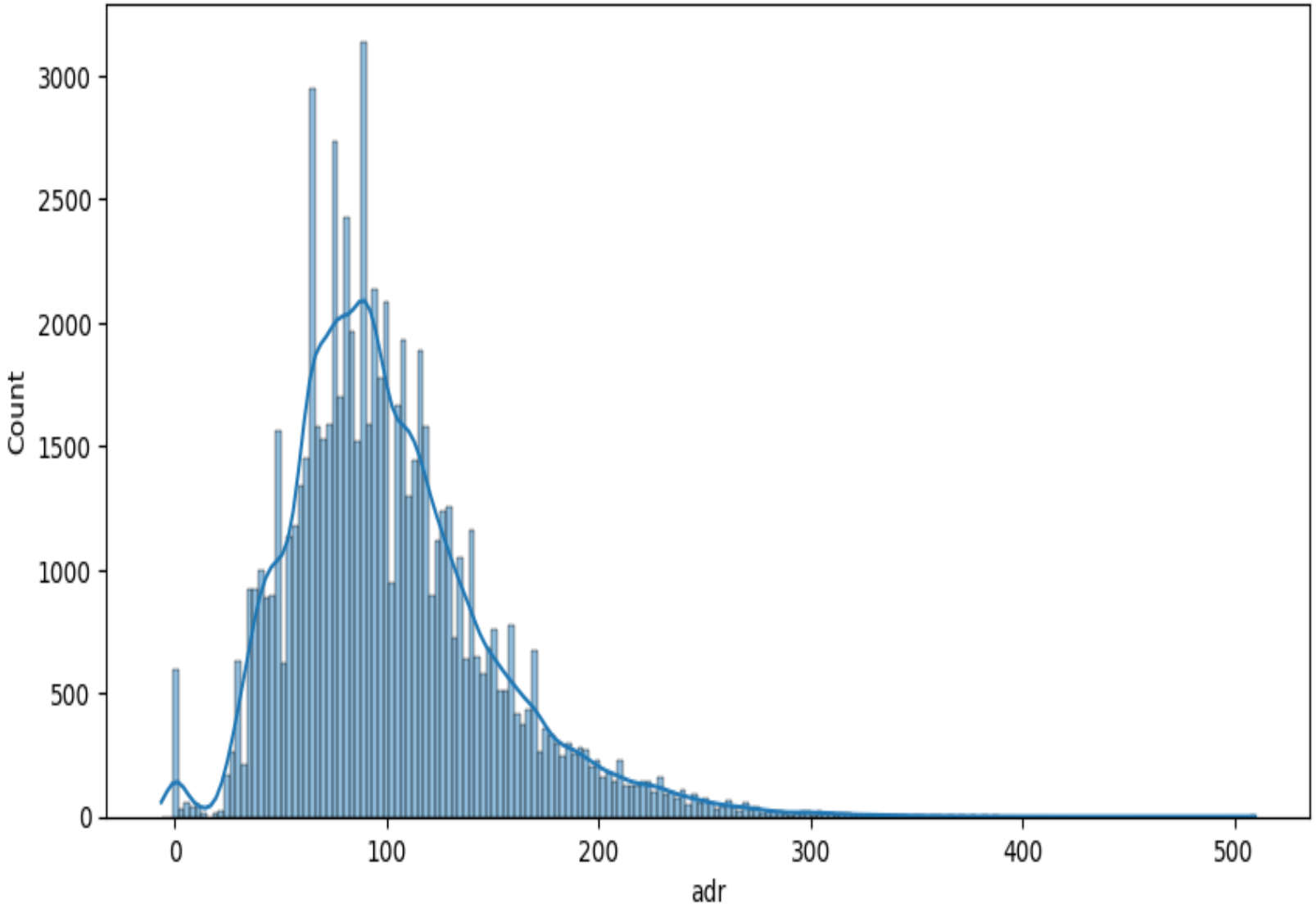
- These findings can be leveraged to enhance the business strategy in several ways, such as customizing promotions for market segments with extended stays or introducing special packages aimed at encouraging longer stays to boost revenue.



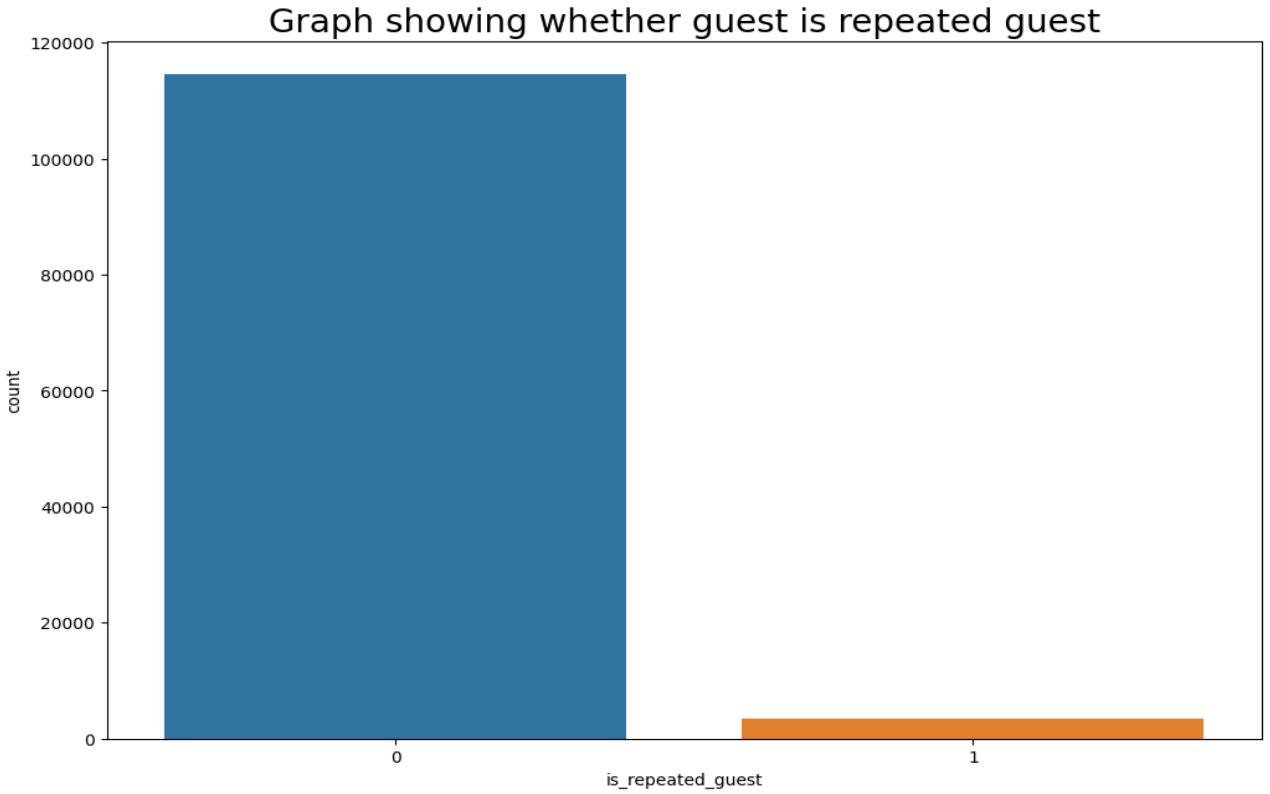
Exploratory Data Analysis

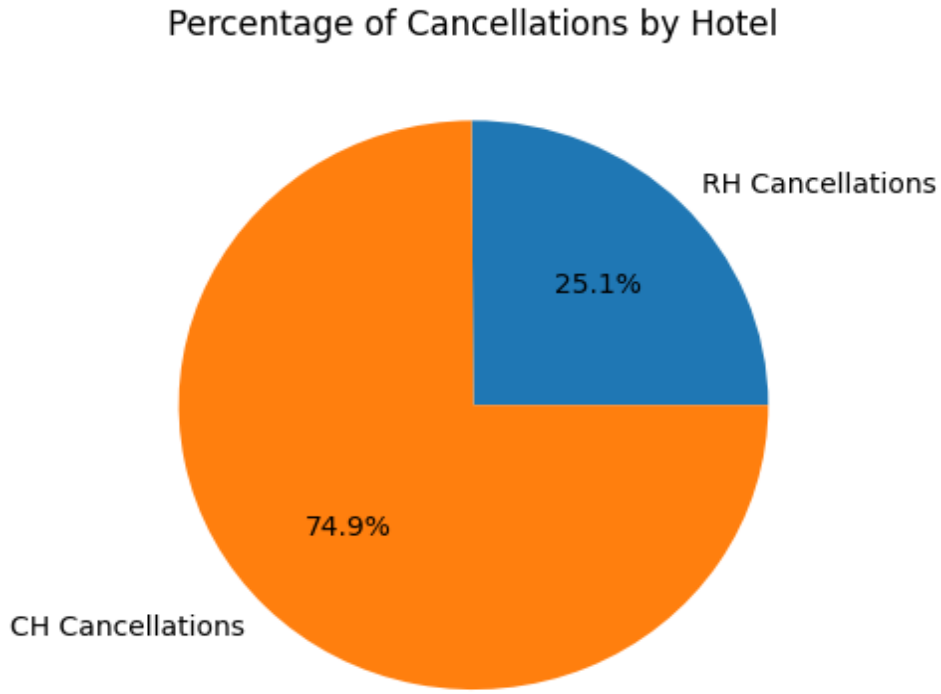
- The histogram clearly illustrates that the majority of ADR (Average Daily Rate) values are clustered in the range of 0 to 200 Euros, with a notable peak around 100 Euros.

- This indicates that most hotel room rates are quite budget-friendly, although there are a few exceptions with higher rates.



Exploratory Data Analysis

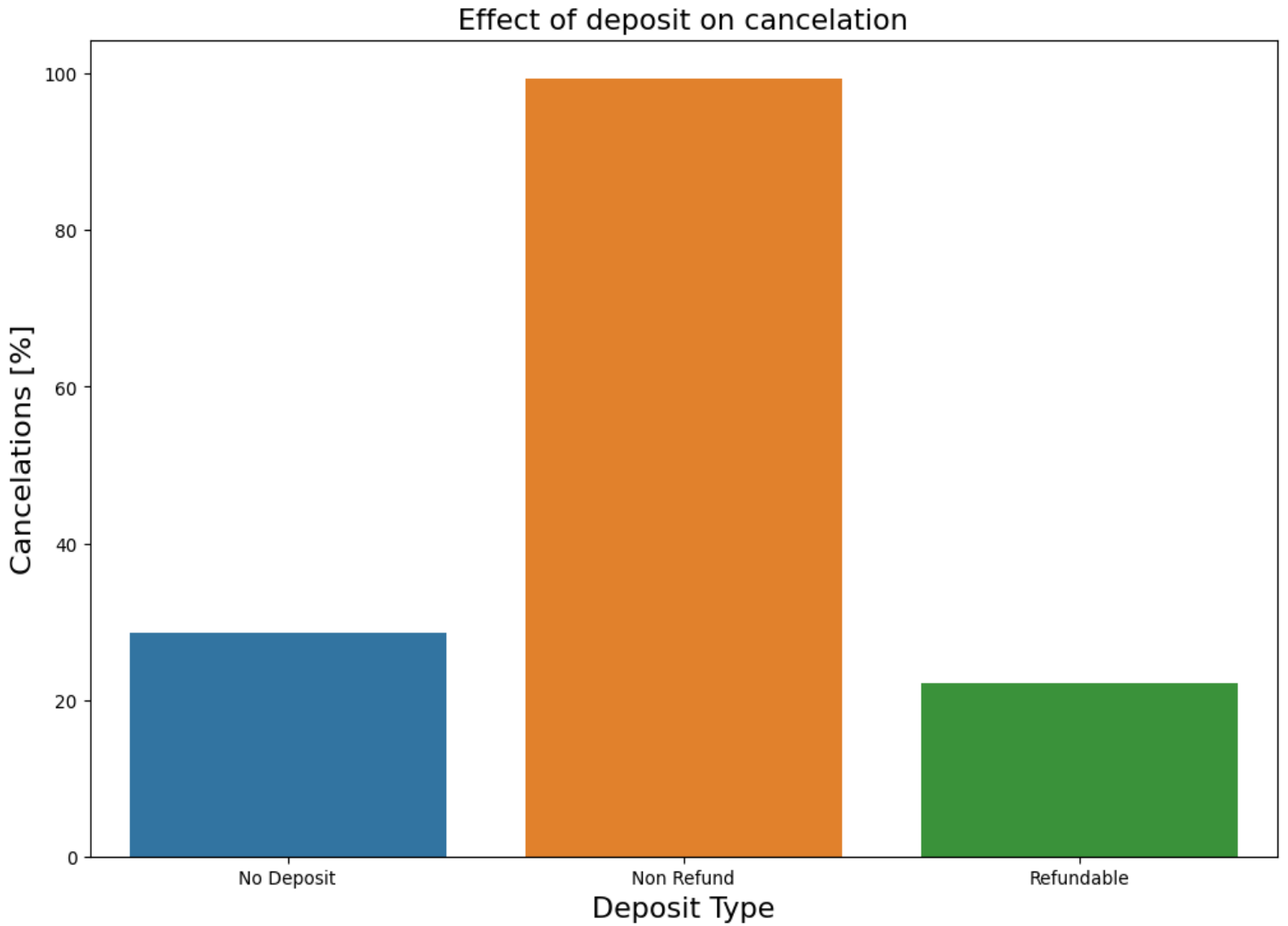




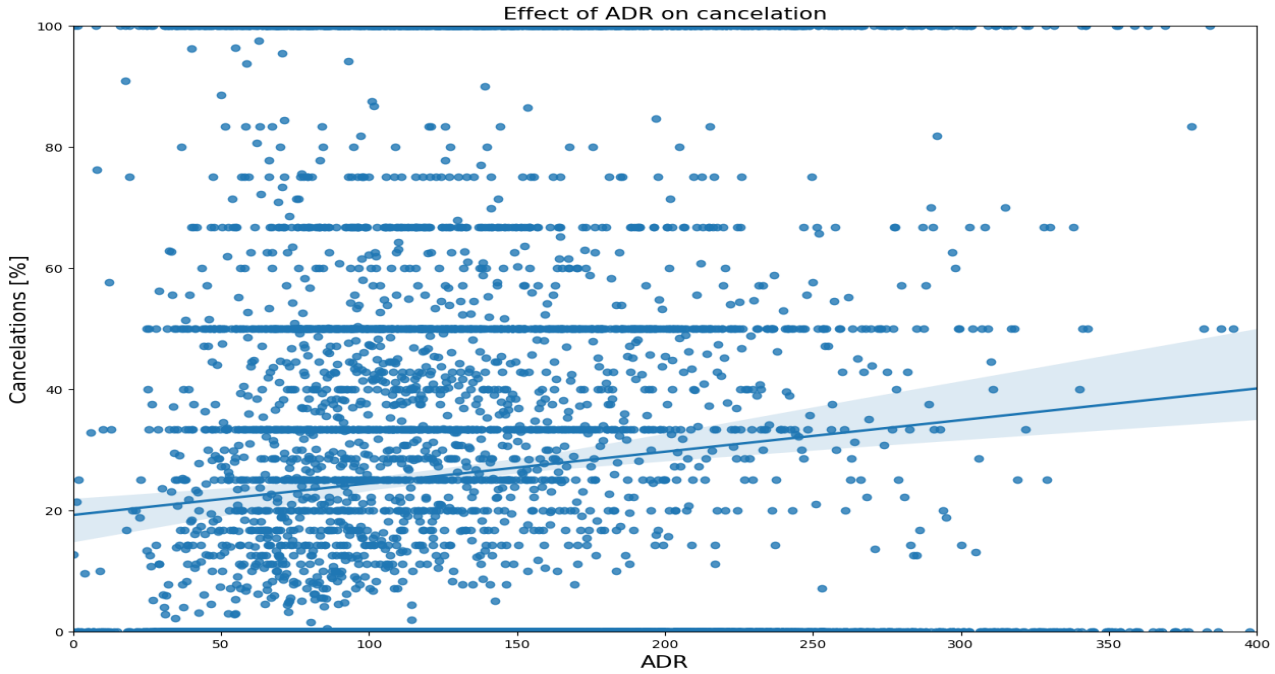
- The count of repeat customers is notably low.
- The percentage of cancellations in city hotels is significantly higher compared to resort hotels.

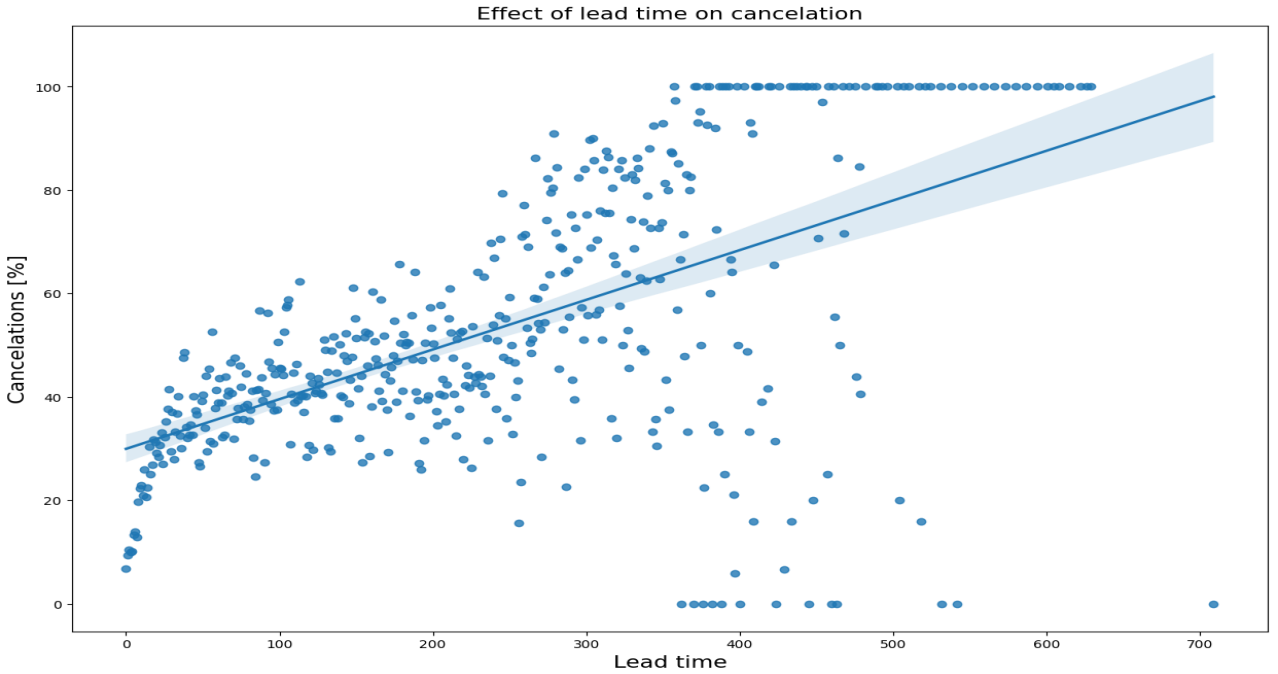
Exploratory Data Analysis

- The high occurrence of non-refundable cancellations can deter guests from booking rooms and potentially lead to customer dissatisfaction.
- Therefore, it is crucial for the hotel to meticulously assess its deposit policies and strike a balance between minimizing cancellations and not discouraging prospective customers.



Exploratory Data Analysis





- The graph indicates that as lead time increases, there is a corresponding increase in the percentage of cancellations. This suggests that customers are more likely to cancel reservations when they have a longer lead time.
- Additionally, there is a positive correlation between ADR (Average Daily Rate) and cancellations, implying that as ADR rises, the number of booking cancellations also tends to increase.

Observation

- City hotels are the top choice among guests, making them the most frequented type of accommodation. Consequently, city hotels can be regarded as the most bustling in terms of occupancy.
- July and August stand out as the peak months in terms of guest activity.
- The majority of guests originate from Europe.
- City hotels experience the highest cancellation rates, with a significant portion of these cancellations being non-refundable.
- Lead time and ADR emerge as the most influential factors contributing to cancellations.

Conclusion

- Higher cancellation likelihood is associated with hotels not taking deposits, thus it's advisable for hotels to implement minimal deposit requirements to reduce cancellation rates.
- The peak period for bookings falls between May and August; therefore, hotels should consider offering enticing promotions to attract more bookings during the off-season.
- With a decline in the number of returning guests, it is recommended that hotel management actively solicit and act upon customer feedback to enhance their facilities and increase customer loyalty.

Thank You