A logo for a university

Description automatically generated

**ASSIGNMENT COVER SHEET**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PROGRAMME | : | Master of Business Analytics | | |
| SUBJECT CODE AND TITLE | : | BAA5033 Python for Business Analytics | | |
| ASSIGNMENT TITLE | : | Understanding Hotel Booking Demand | | |
|  |  |  | | |
| LECTURER | : | Dr. Aaron Aw Teik Hong | ASSIGNMENT DUE DATE: | 18/10/2024 |

STUDENT’S DECLARATION

1. I hereby declare that this assignment is based on my own work except where acknowledgement of sources is made.
2. I also declare that this work has not been previously submitted or concurrently submitted for any other courses in Sunway University/College or other institutions.

[ Submit “Turn-it-in” report (please tick √): Yes \_\_\_\_\_ No \_\_√\_\_\_ ]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NO. | NAME | STUDENT ID NO. | SIGNATURE | DATE |
| 1. | Abdul Hakim Bin Kamalur Rahman | 24015257 | Abdul Hakim | 18/10/2024 |
| 2. | Annabel Ching Ke Xin | 24002685 | Annabel | 18/10/2024 |
| 3. | Chia Poh Yang | 23138555 | Chia | 18/10/2024 |
| 4. | Divani A/P Arumugam | 19058908 | Divani | 18/10/2024 |
| 5. | Harresh A/L Ragunathan | 19076090 | Harresh | 18/10/2024 |
| 6. |  |  |  |  |

E-mail Address / Addresses (according to the order of names above):

|  |  |
| --- | --- |
| 1. 24015257 @imail.sunway.edu.my | 4. 19058908@imail.sunway.edu.my |
| 2. 24002685@imail.sunway.edu.my | 5. 19076090 @imail.sunway.edu.my |
| 3. 23138555@imail.sunway.edu.my | 6. |

APPROVAL FOR LATE SUBMISSION OF ASSIGNMENT (If applicable)

IF extension is granted, what is the revised due date? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature of Lecturer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |
| --- |
| Marker’s Comments: |

Marks and / or Grade Awarded: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ADDENDUM**

**USE OF ARTIFICAL INTELLIGENCE (A.I.) DECLARATION**

Students are allowed to use AI to support completion of assessments. However, students are reminded to do so ethically and transparently. This is so that (a) submissions can be fairly and accurately marked; and (b) feedback can be provided on the content that reflects student ability, in order to help with future submissions. Students are also reminded that in accordance with the University’s Academic Malpractice Policy, Item 4.11.2, “*… the representation of work: written, visual, practical or otherwise, of any other person, including another student or* ***anonymous web-based material*** *[emphasis added], or any institution, as the candidate’s own*” is considered malpractice.

**Declaration**

[ ] I / We used the following A.I. tools to produce content in this submission:

|  |  |  |  |
| --- | --- | --- | --- |
| **Tool** | **Purpose** | **Prompts** | **Sections where AI output was used / Outcome(s) in the submission** |
| *ChatGPT* | *Generating coding reference* | *“How to manage with missing value in Python?”*  *“Suggest some visualization suitable for Exploratory Data Analysis”* | *The coding part in Python. All the answer provided by ChatGPT are for references.* |
| *Grammarly* | *Correcting grammar and spelling, improving sentence structure* | *N/A* | *Grammarly suggestions were used for all sections of the essay* |
|  |  |  |  |

*Note: Add additional rows if necessary.*

**OR**

[ ] I / We did not use any A.I. tools to produce any of the content in this submission.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NO. | NAME | STUDENT ID NO. | SIGNATURE | DATE |
| 1. |  |  |  |  |
| 2. |  |  |  |  |
| 3. |  |  |  |  |
| 4. |  |  |  |  |
| 5. |  |  |  |  |
| 6. |  |  |  |  |

E-mail Address / Addresses (according to the order of names above):

|  |  |
| --- | --- |
| 1. | 4. |
| 2. | 5. |
| 3. | 6. |

**Contents**

[**Part 1: Business Understanding 5**](#_Toc179766536)

[**Part 2: Exploratory Data Analysis 10**](#_Toc179766537)

[**a. Pre-processing 10**](#_Toc179766538)

[**b. Exploratory data analysis 14**](#_Toc179766539)

[**Part 3: Business Insights 17**](#_Toc179766540)

[**3.1 Insights on distributions, trends, and patterns in the data 17**](#_Toc179766541)

[**3.2 Business Insights and Analysis 21**](#_Toc179766542)

[**3.3 Menu application 39**](#_Toc179766543)

[**References 40**](#_Toc179766544)

**List of Figure**

[Figure 2.1: Missing Values 10](#_Toc179766425)

[Figure 2.2: Missing Values (Continued) 11](#_Toc179766426)

[Figure 2.3: Undefined Values 11](#_Toc179766427)

[Figure 2.4: Undefined Values (Continued) 11](#_Toc179766428)

[Figure 2.5: Null Values 11](#_Toc179766429)

[Figure 2.6: Before Cleaning Outliers 13](#_Toc179766430)

[Figure 2.7a: After Cleaning Outliers with 1.5 threshold 13](#_Toc179766431)

[Figure 2.7b: After Cleaning Outliers with 1.0 threshold 14](#_Toc179766432)

[Figure 2.8: Removing Outliers for Certain Columns 14](#_Toc179766433)

[Figure 2.9: Basic Statistics 15](#_Toc179766434)

[Figure 2.10: Correlation Matrix 16](#_Toc179766435)

[Figure 3.1: Categorical Variables 19](#_Toc179766436)

[Figure 3.2: Numerical Variables 20](#_Toc179766437)

[Figure 3.3: Total Number of Bookings Cancelled VS Not Cancelled for Both Hotel 21](#_Toc179766438)

[Figure 3.4: Number of Bookings Cancelled VS Not Cancelled by Resort Hotel 22](#_Toc179766439)

[Figure 3.5: Number of Bookings Cancelled VS Not Cancelled by City Hotel 23](#_Toc179766440)

[Figure 3.6: Global View of Booking Volumes 24](#_Toc179766441)

[Figure 3.7: Top 10 Countries with The Highest Bookings 24](#_Toc179766442)

[Figure 3.8: Top 10 Countries with The Highest Booking Volumes 25](#_Toc179766443)

[Figure 3.9:Average Daily Rate (ADR) and The Number of Bookings Over Time 26](#_Toc179766444)

[Figure 3.10: Monthly Average Daily Rate (ADR) Fluctuations for Both Hotels throughout The Year 28](#_Toc179766445)

[Figure 3.11: Top 10 Bookings made by Agents and Companies 29](#_Toc179766446)

[Figure 3.12: Total night VS The Number of Stays for Both Hotel 31](#_Toc179766447)

[Figure 3.13: The Number of Cancellation Rate based on The Guest Combination 32](#_Toc179766448)

[Figure 3.14: Cancellation Rate by Deposit Type 33](#_Toc179766449)

[Figure 3.15: Cancellation Rate by Repeated Guests 34](#_Toc179766450)

[Figure 3.16: Total Bookings by Reserved Room Type 35](#_Toc179766451)

[Figure 3.17: Actual Assigned Room Types 36](#_Toc179766452)

[Figure 3.18: Cancellation Rate for Top 5 Assigned Room Types 36](#_Toc179766453)

[Figure 3.19: Menu Application 39](#_Toc179766454)

**List of Table**

[Table 1.1: Variables Description 10](#_Toc179735833)

# Part 1: Business Understanding

**Dataset Name:** Hotel booking demand

**Description:**

This dataset provides detailed booking information for two types of hotels, a city hotel and a resort hotel. It comprises 119,391 observations, each representing a unique hotel booking between 1st of July 2015 and 31st of August 2017. The dataset includes variables related to the booking process, such as customer lead time, the number of nights stayed (both weekday and weekend), guest demographics (including adults, children, and babies), and type of meal booked. It also contains information on market segments, distribution channels, and whether the booking was made by a repeat guest.

In addition, the dataset tracks previous cancellations, previous non-canceled bookings, and various booking details such as the reserved and assigned room types, the number of booking changes, and the deposit type. Further variables include agent and company information, the days a booking is spent on the waiting list, average daily rate, and special requests.

The target variable is whether a booking was canceled, providing an opportunity to analyze the factors contributing to cancellations. This dataset is valuable for predicting booking cancellations, optimizing hotel booking systems, and improving customer retention.

**Source:**

The dataset was obtained from Kaggle which is a popular platform for solving data science challenges.

**URL:**

<https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand>

**Variables description:**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| **hotel** | Categorical | Hotel booked by each booking   * Resort hotel * City hotel |
| **is\_canceled** | Categorical | Value indicating if the booking was   * Canceled (1) * Not canceled (0) |
| **lead\_time** | Integer | Number of days that elapsed between the entering date of the booking into the property management system (PMS) and the arrival date |
| **arrival\_date\_year** | Integer | Year of arrival date |
| **arrival\_date\_month** | Categorical | Month of arrival date with 12 categories: “January” to “December” |
| **arrival\_date\_week\_number** | Integer | Week number of the arrival date |
| **arrival\_date\_day\_of\_month** | Integer | Day of the month of the arrival date |
| **stays\_in\_weekend\_nights** | Integer | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to staye at the hotel |
| **stays\_in\_week\_nights** | Integer | Number of weekday nights (Monday to Friday) the guest stayed or booked to staye at the hotel |
| **adults** | Integer | Number of adults |
| **children** | Integer | Number of children |
| **babies** | Integer | Number of babies |
| **meal** | Categorical | Type of meal booked. Categories are presented in standard hospitality meal packages:   * Undefined/SC - no meal package * BB – Bed & breakfast * HB – half board (breakfast and one other meal – usually dinner) * FB – full board (breakfast, lunch and dinner) |
| **country** | Categorical | Country of origin. Categories are represented in the ISO country codes 3166-3:2013 format |
| **market\_segment** | Categorical | Market segmentation designation.  In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators” |
| **distribution\_channel** | Categorical | Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” |
| **is\_repeated\_guest** | Categorical | Value indicating if the booking name was from   * Repeated guest (1) * Not a repeated guest (0) |
| **previous\_cancellations** | Integer | Number of previous bookings that were cancelled by the customer prior to the current booking.  In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and canceled. |
| **previous\_bookings\_not\_canceled** | Integer | Number of previous bookings not cancelled by the customer prior to the current booking.  In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and not canceled. |
| **reserved\_room\_type** | Categorical | Code of room type reserved. Code is  presented instead of designation for  anonymity reasons |
| **assigned\_room\_type** | Categorical | Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request.  Code is presented instead of designation for anonymity reasons. |
| **booking\_changes** | Integer | 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** | Categorical | 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   In case no payments were found the  value is “No Deposit”. If the payment was equal to or exceeded the total cost of stay, the value is set as “Non-Refund”. Otherwise, the value is set as “Refundable” |
| **agent** | Categorical | ID of the travel agency that made the booking |
| **company** | Categorical | ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons |
| **days\_in\_waiting\_list** | Integer | Number of days the booking was in the waiting list before it was confirmed to the customer |
| **customer\_type** | Categorical | 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** | Numerical | Average daily rate calculated by dividing the sum of all lodging transactions by the total number of staying nights |
| **required\_car\_parking\_spaces** | Integer | Number of car parking spaces required by the customer |
| **total\_of\_special\_requests** | Integer | Number of special requests made by the customer (e.g. twin bed or high floor) |
| **reservation\_status** | Categorical | 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 |
| **reservation\_status\_date** | Date | Date at which the last status was set.  This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel |

Table 1.1: Variables Description

# Part 2: Exploratory Data Analysis

## **a. Pre-processing**

In order to proceed with the visualization process, the dataset must first go through the pre-processing stage. The main objective of pre-processing is to make the dataset more accurate, which in turn will provide more accurate visualizations. This is done through eliminating unneeded columns, handling missing values, and handling outliers. These tasks were all performed using Python within Google Colab.

**Handle Missing values**

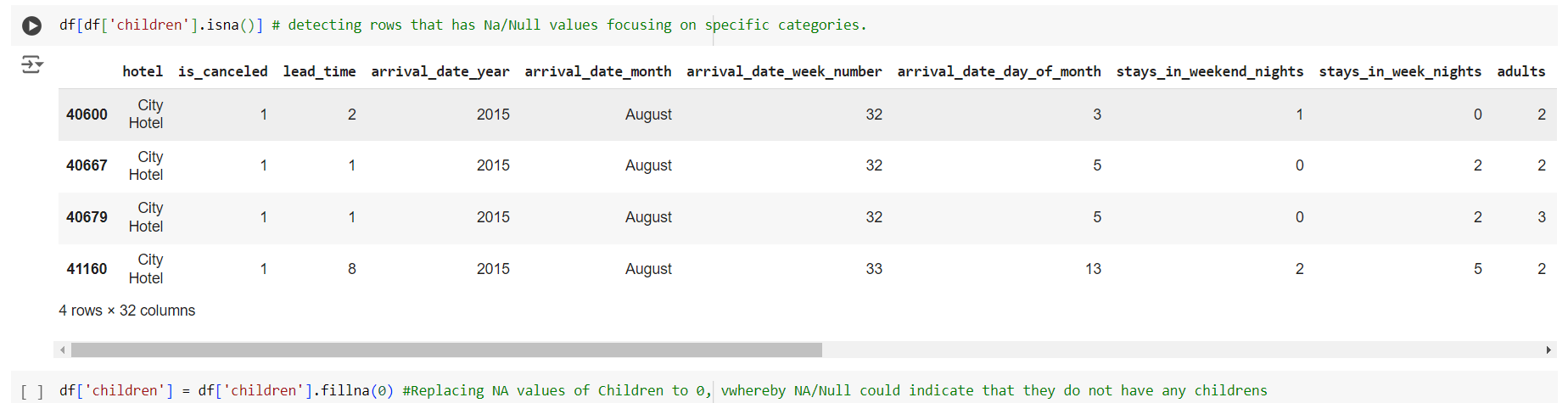


Figure 2.1: Missing Values

It may be observed that the “children” column contains both “NA” and “0” values. To address this, we replace all “NA” values in the “children” column with “0”, as “NA” suggests that the booking did not involve any children.

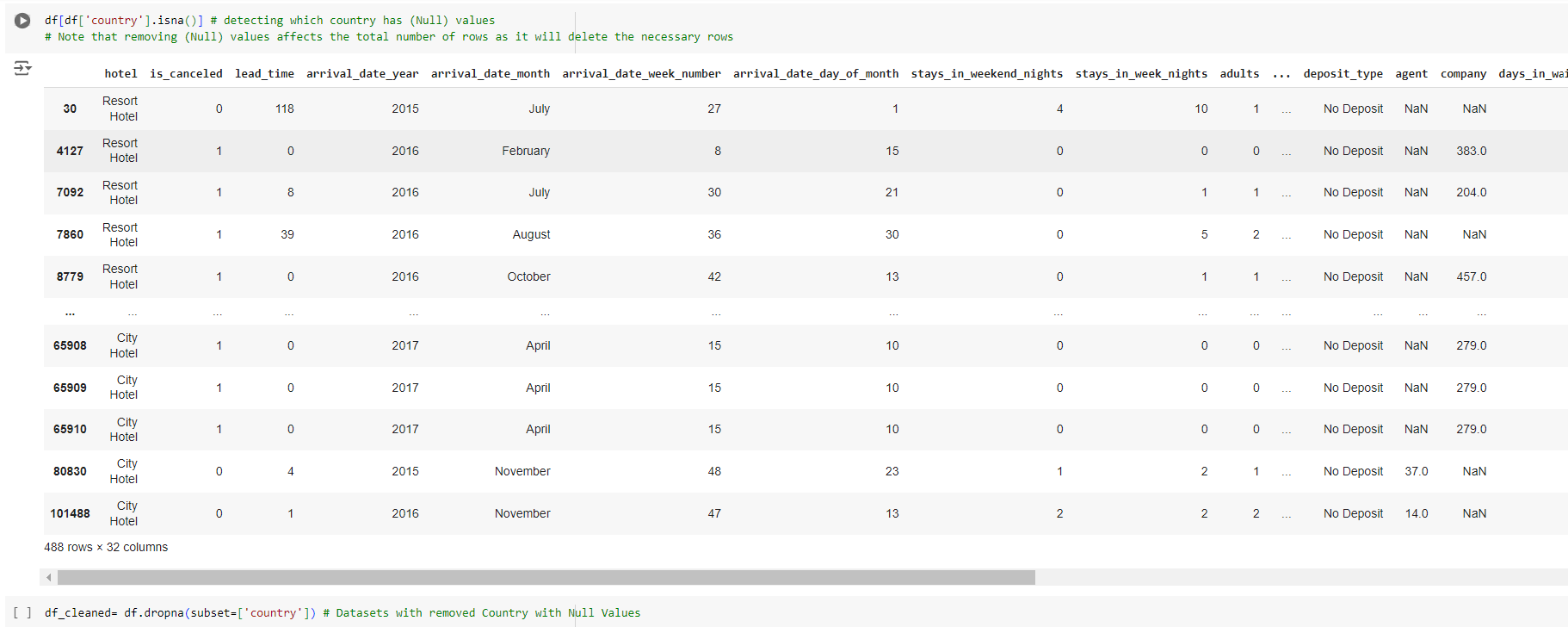


Figure 2.2: Missing Values (Continued)

The rows that contain null values in the “country” column were removed.



Figure 2.3: Undefined Values

It may be observed that there are “undefined” values in the distribution channel and market segment column. These values are dropped as it will not provide any insights.



Figure 2.4: Undefined Values (Continued)

According to the data description, “undefined” and “SC” both share the same meaning that there is no meal ordered for that particular booking. Therefore, the values “undefined” are replaced with “SC”.



Figure 2.5: Null Values

According to Antonio, De Almeida, & Nunes (2019), the data was extracted directly from the hotels’ Property Management System (PMS) database. This system ensures that there is no missing data in the database tables. As a result, the values “NULL” should not be regarded as missing data but rather as “not applicable”. For example, if the “Agent” value is marked as “NULL”, it signifies that the booking was made directly by the customer rather than through a travel agent. Therefore, we will replace all “NULL” values with “Self-Bookers” for easier identification.

**Outlier**

**Identify outlier**

After addressing missing values, the outliers are detected using the Interquartile Range (IQR) method, a technique for identifying outliers in continuous data (Vinutha, Poornima, & Sagar, 2018). The IQR represents the range between the 25th and 75th percentiles (Q1 and Q3, respectively), capturing the middle 50% of the data (Barbato, Barini, Genta, & Levi, 2011). With the IQR, Q1, and Q3 values, the lower and upper fences can be calculated. These fences are essential for identifying outliers, as any value that falls below the lower fence or above the upper fence is considered an outlier. In this case, we apply a –1/+1 threshold using the following formulas:

Lower Fence = Q1 - 1IQR

Upper Fence = Q3 + 1IQR

Initially, a –1.5/+1.5 threshold was used to handle the outliers. However, it was observed that not all the outliers were eliminated, which lead to the decision of applying a –1/+1 threshold. Before handling the outliers, the dataset had 119,390 rows and 32 columns. After processing, 99,765 rows and 32 columns remain.

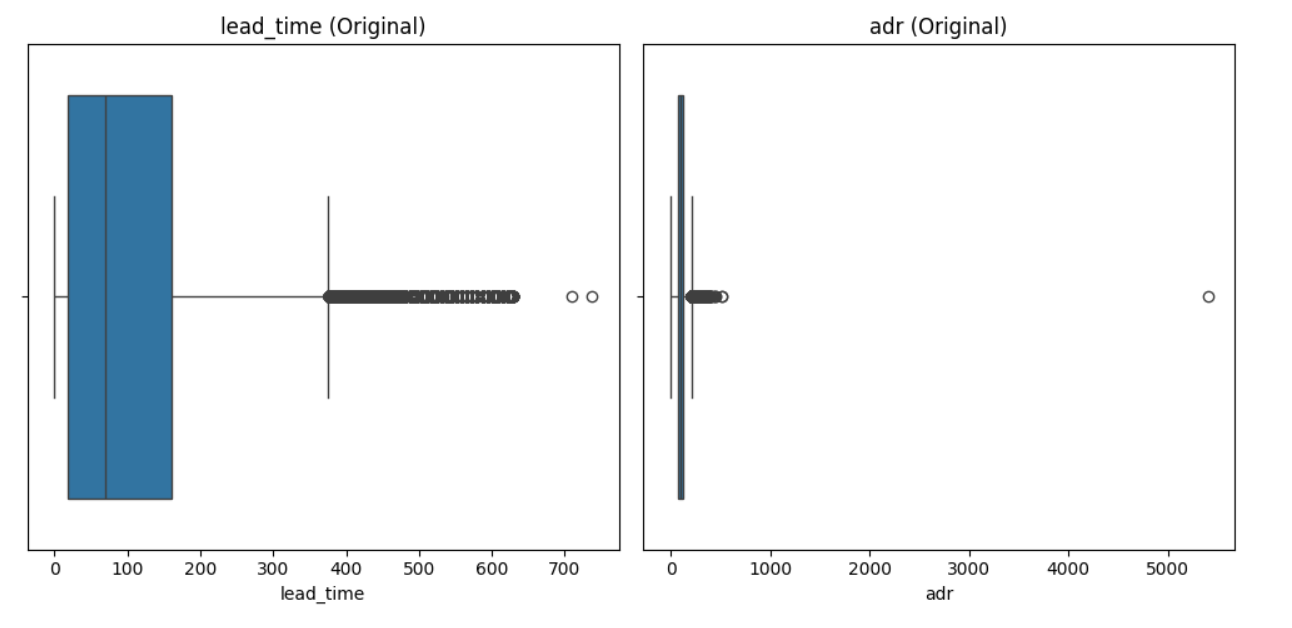


Figure 2.6: Before Cleaning Outliers

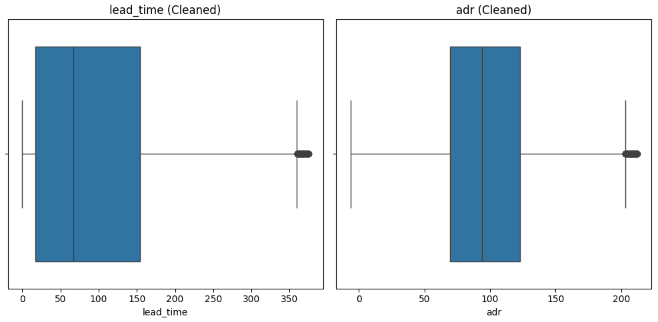


Figure 2.7a: After Cleaning Outliers with 1.5 threshold

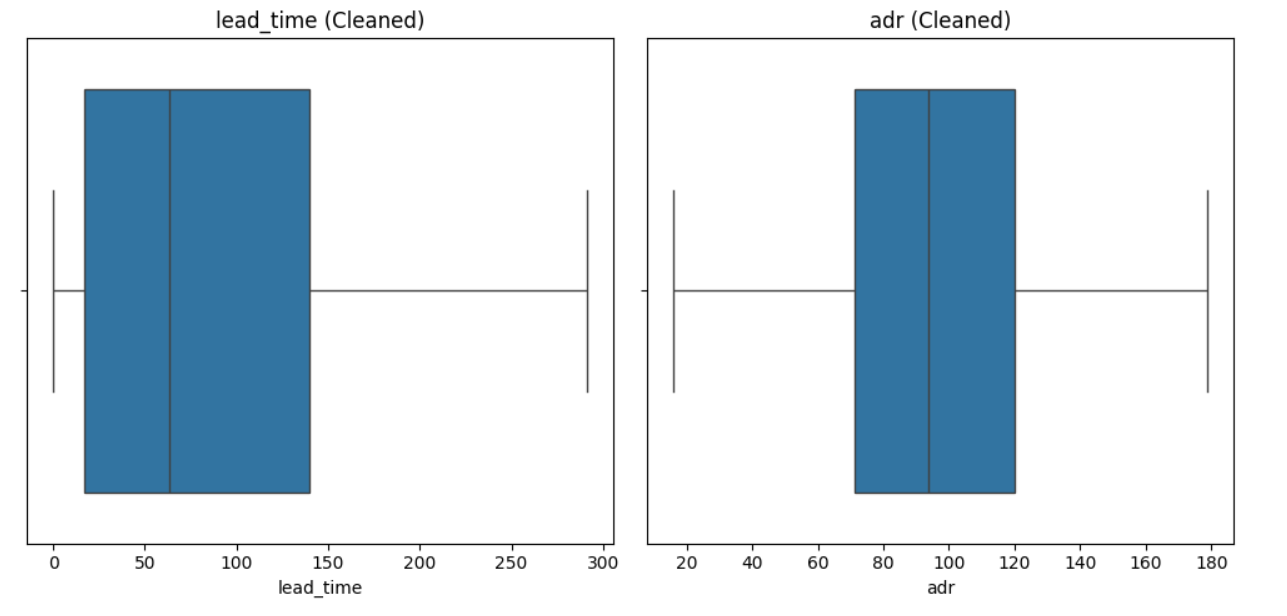


Figure 2.7b: After Cleaning Outliers with 1.0 threshold



Figure 2.8: Removing Outliers for Certain Columns

The outliers in the “children” and “babies” columns were also identified. As a result, a decision was made to remove these outliers based on the following conditions, where if the number of children exceed 2 or the number of babies exceed 3.

## **b. Exploratory data analysis**

**Basic statistics (mean, median, standard deviation)**

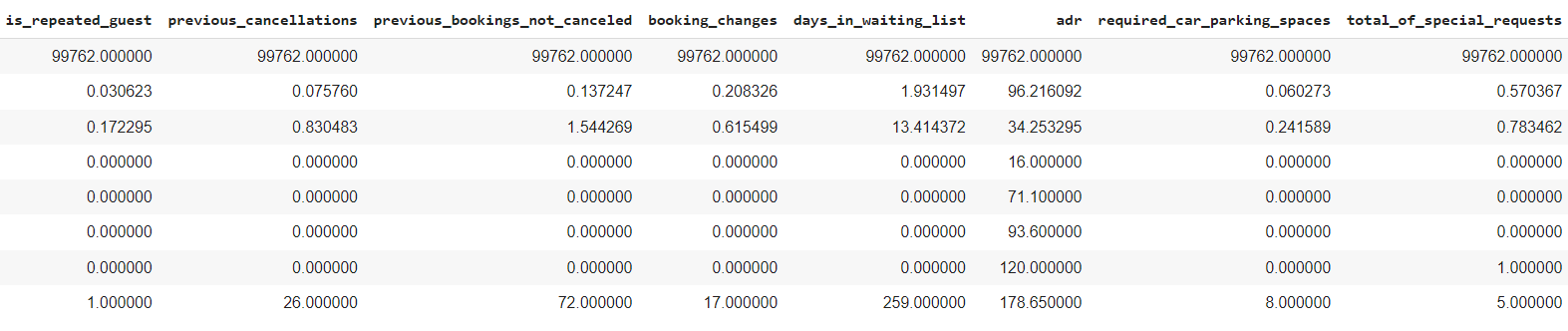
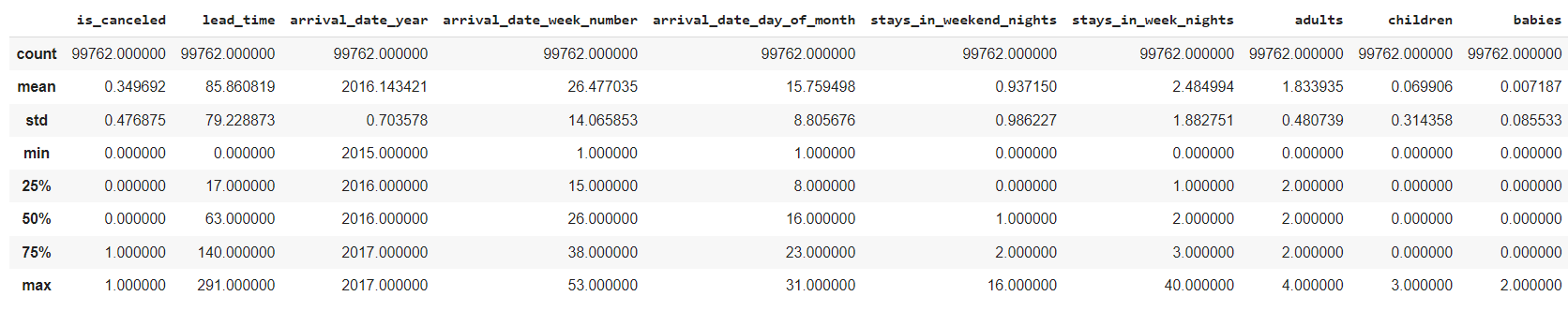


Figure 2.9: Basic Statistics

The dataset contains 99,762 records and provides statistical summaries for various variables. For example, the average lead time for bookings is around 85.86 days, with a median of 63 days, and a standard deviation of 79.23, indicating some variability in the booking patterns. The mean number of adults per booking is approximately 1.83, with a median of 2, and a standard deviation of 0.48, suggesting most bookings include two adults. For children and babies, the averages are low at 0.07 and 0.007 respectively, with medians of zero, indicating that the majority of bookings are made without children or babies.

Outliers are visible in some variables, such as “previous\_cancellations” which has a maximum value of 26 compared to a median of 0, and “days\_in\_waiting\_list” which ranges up to 259, although the median is 0, showing most bookings do not involve a waiting period. Additionally, the “adr” (average daily rate) has a mean of 96.22 and a standard deviation of 34.25, with the highest recorded value at 178.65. These statistics offer insights into booking behaviors and patterns within the dataset.

**Correlation matrix**

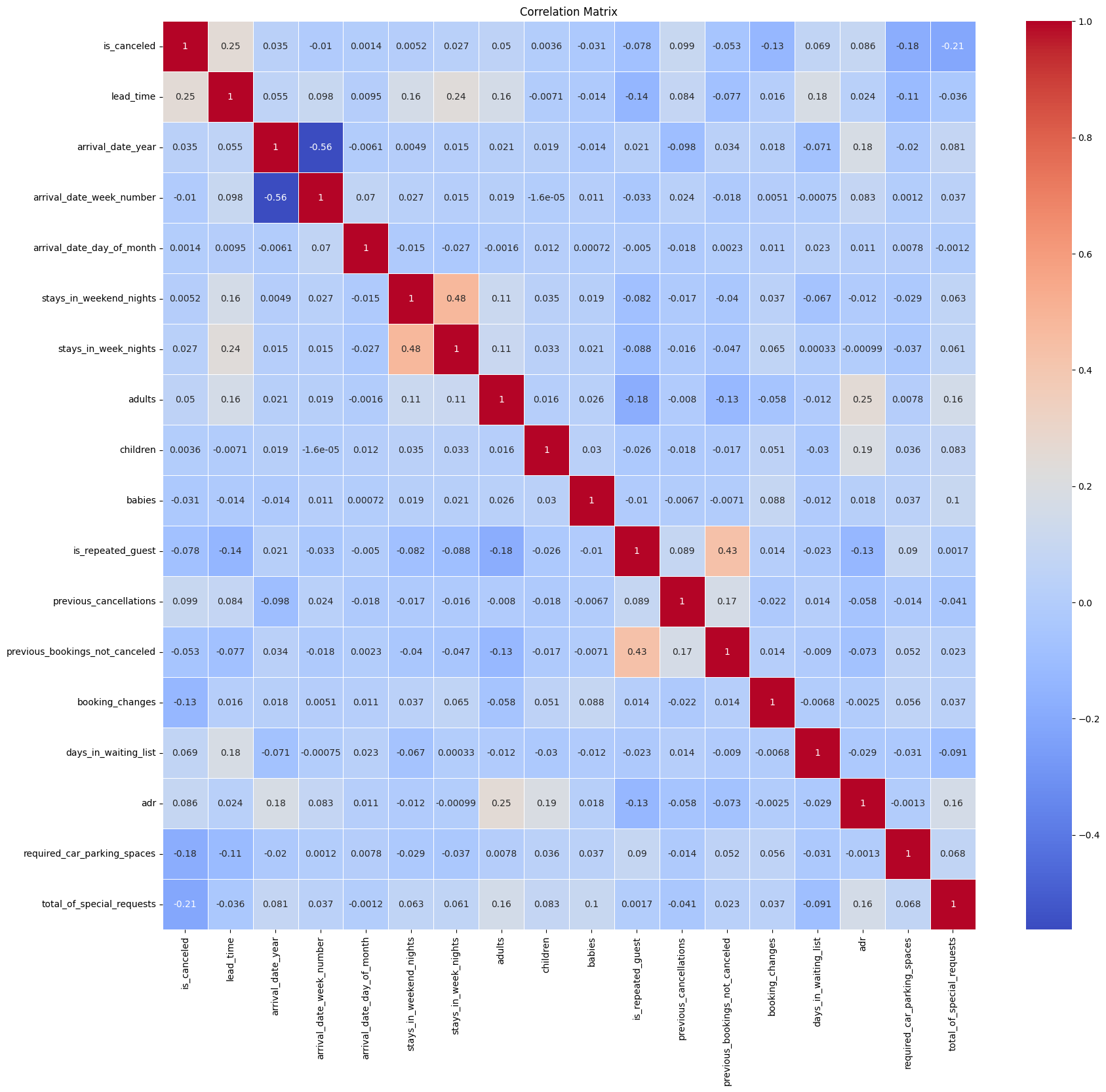


Figure 2.10: Correlation Matrix

The correlation matrix provides insights into the relationships between various variables in the dataset. Notably, the “is\_canceled” variable shows a moderate positive correlation with “lead\_time” (0.25) and “previous\_cancellations” (0.10), indicating that longer lead times and prior cancellations increase the likelihood of a booking being canceled. It also shows a moderate negative correlation with “is\_repeated\_guest” (-0.08), suggesting that repeat guests are less likely to cancel. Another significant relationship is between “stays\_in\_week\_nights” and “stays\_in\_weekend\_nights” (0.48), implying that bookings with more weeknight stays tend to have more weekend stays as well. “Previous\_bookings\_not\_canceled” has a moderate positive correlation with “is\_repeated\_guest” (0.43), suggesting repeat guests often have a history of not canceling their bookings. Overall, most correlations are relatively low, indicating that many variables do not have strong linear relationships with one another.

# Part 3: Business Insights

## **3.1 Insights on distributions, trends, and patterns in the data**

**3.1.1 Categorical variables**

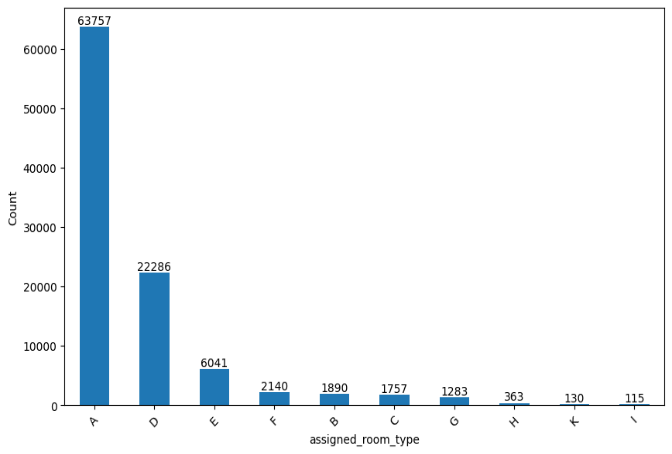
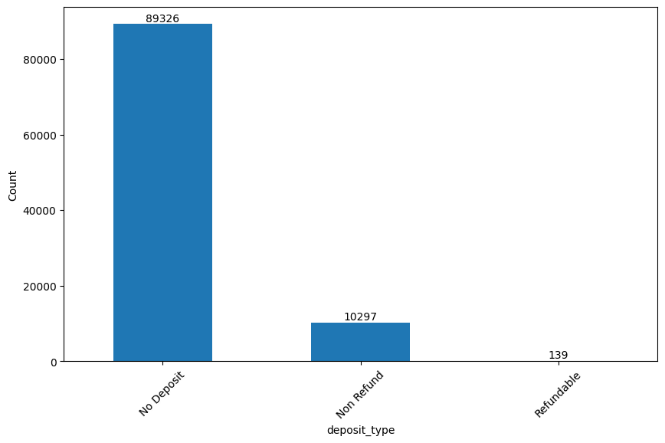
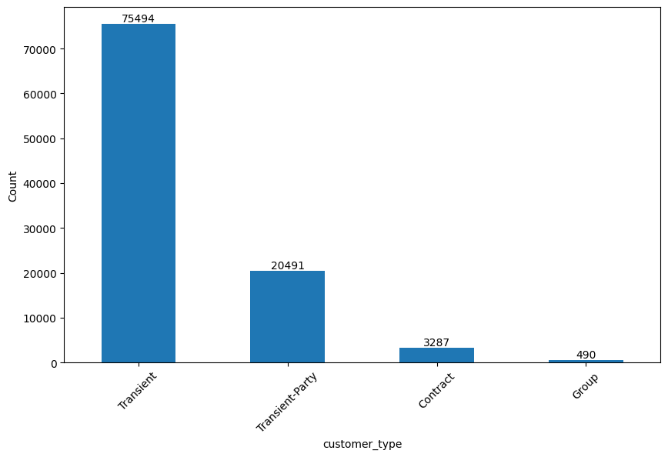
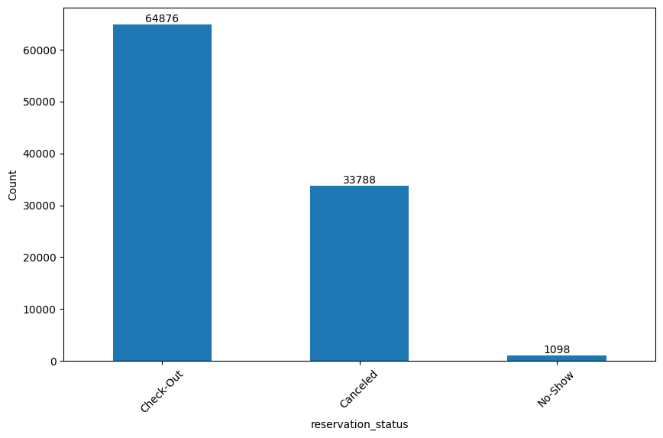
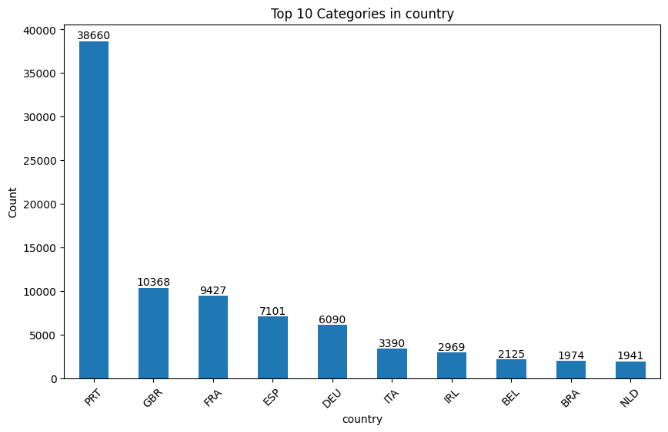
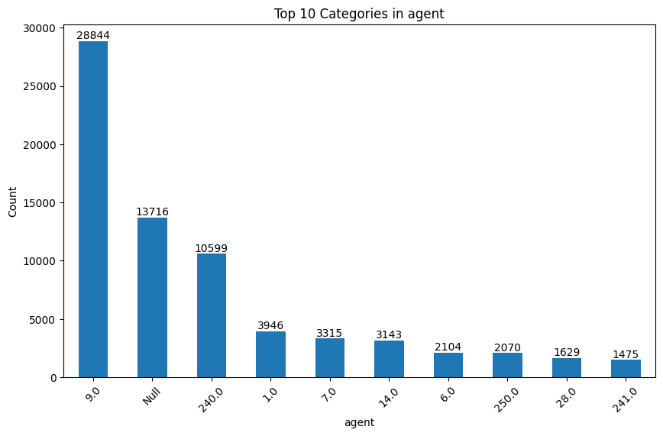
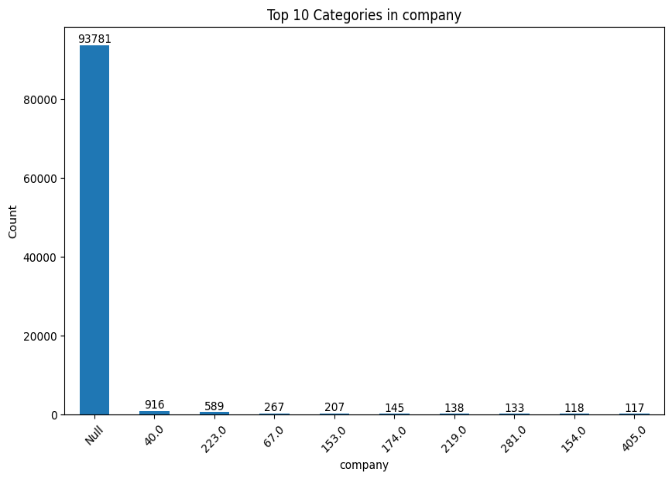
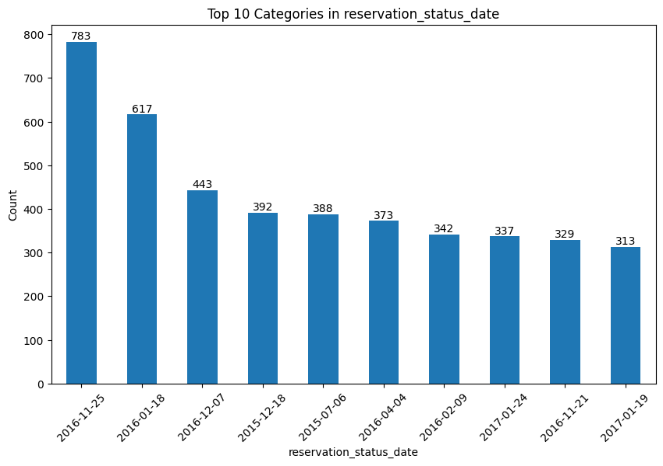
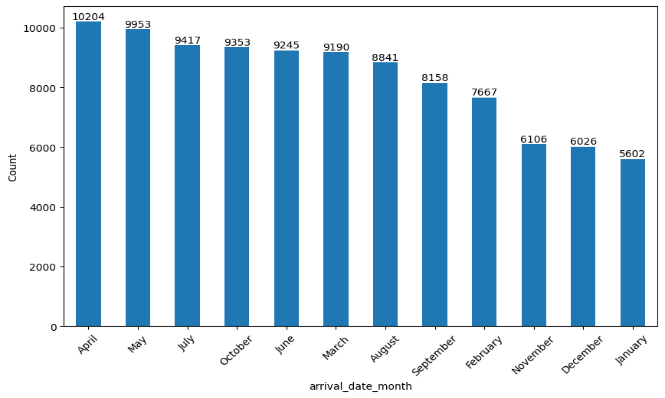
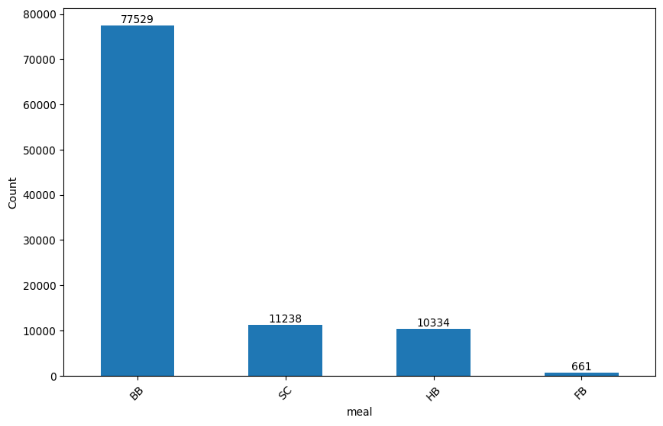
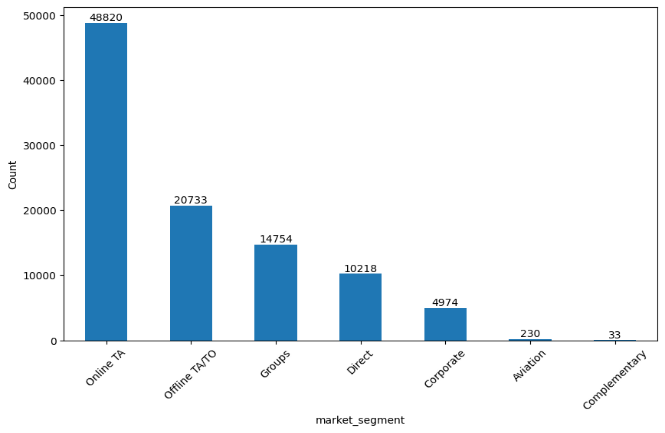
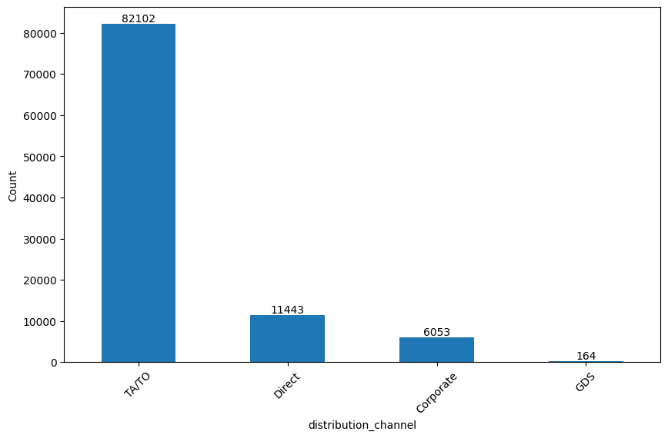
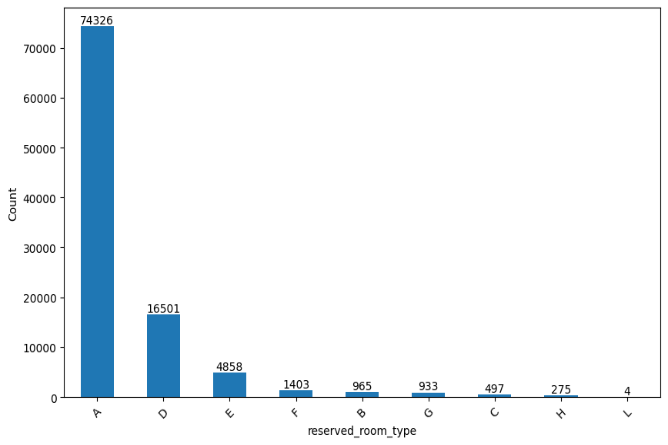
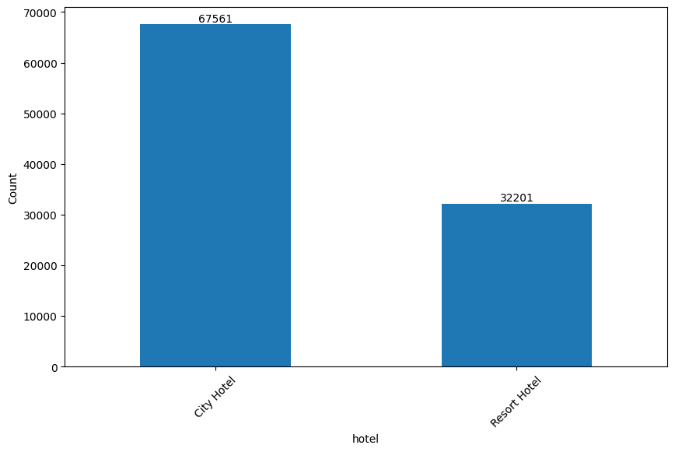


Figure 3.1: Categorical Variables

The count plot above highlights the distribution of reservations across different hotel types, with City Hotels having a higher count compared to Resort Hotels. The reservation status data shows that most reservations were checked out, followed by cancellations and no-shows. The company and agent categories reveal that a significant number of reservations were made without specifying a company or agent, indicating a large volume of direct bookings. The country data indicates that most reservations were made by guests from Portugal (PRT), followed by the United Kingdom (GBR) and France (FRA). The customer type data highlights that transient customers were the majority, while the deposit type data shows that most bookings did not require a deposit. Additionally, the distribution channel data reveals that most bookings were made through travel agents and tour operators (TA/TO), and the market segment data shows a high number of online travel agent (OTA) bookings. The meal data indicates that bed and breakfast (BB) was the most common meal plan. Lastly, the arrival date data shows a relatively even distribution of arrivals throughout the year, with a slight peak in the summer months.

**3.1.2 Numerical variables**

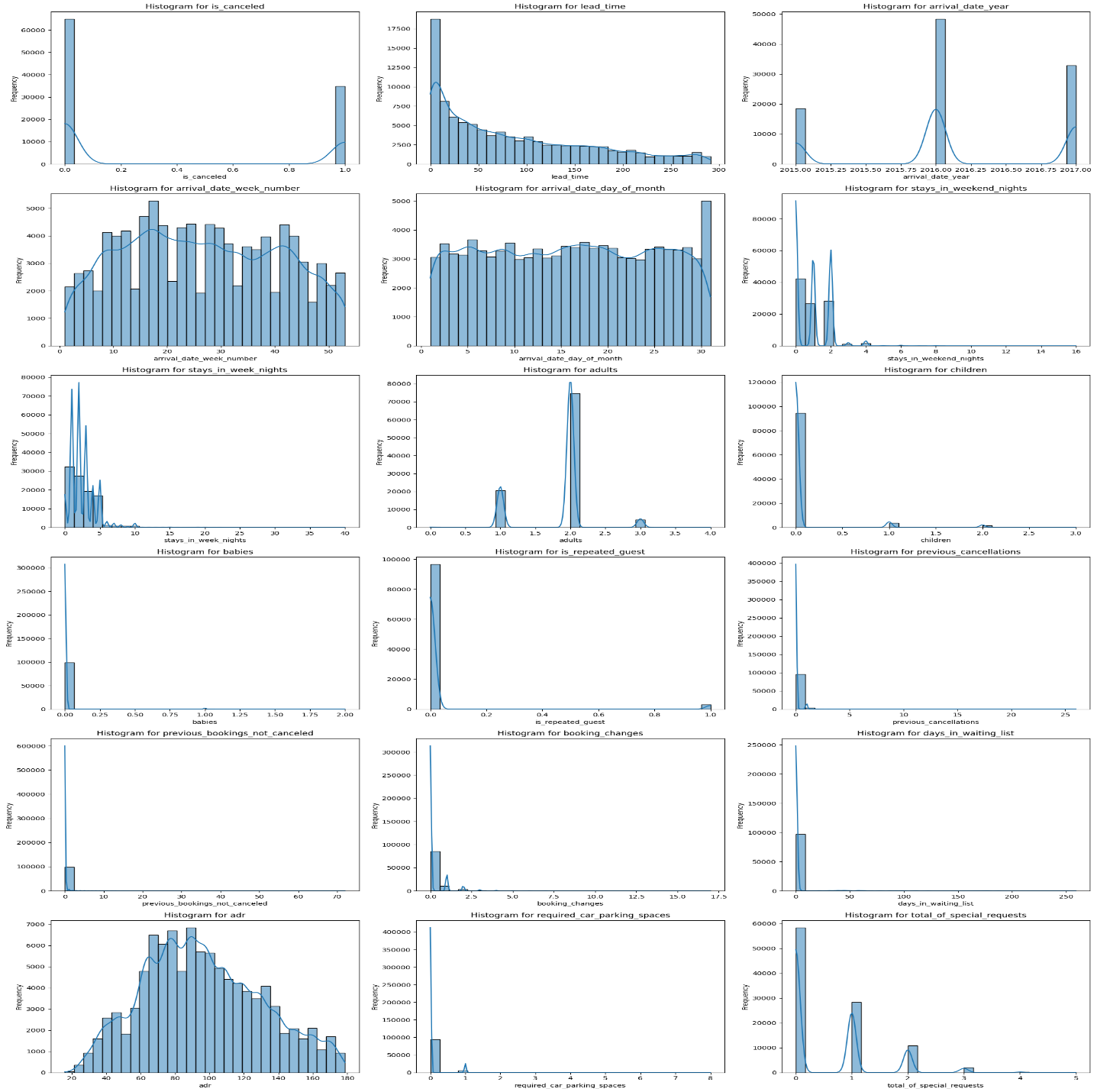


Figure 3.2: Numerical Variables

The histograms provide a detailed view of the distribution of numerical variables in the dataset. The “is\_canceled” variable reveals a higher frequency of non-canceled bookings which encoded as 0 compared to canceled bookings which encoded as 1. The “lead\_time” variable shows a right-skewed distribution, with most bookings made within shorter lead times, though some have significantly longer lead times. The “Arrival\_date\_year” reflects more bookings in 2016 since the data collection spanned from 1st of July 2015 and 31st of August 2017, meaning only the bookings from 2016 were fully captured. The distributions for “stays\_in\_week\_nights” and “stays\_in\_weekend\_nights” show that most stays are short, typically ranging from 1 to 3 nights. The "adults" and "children" variables indicate that most bookings include 2 adults and no children. Variables such as “previous\_bookings\_not\_canceled”, “is\_repeated\_guest”, and “previous\_cancellations” are heavily skewed toward 0, suggesting that most guests are not repeat customers and have few cancellations. The “ADR” (Average Daily Rate) distribution is approximately normal, with a slight right skew. Additionally, most bookings do not request extra parking spaces or special services. These distributions shed light on key patterns in hotel bookings, including the prevalence of short stays, few special requests, and low cancellation rates among repeat guests.

## **3.2 Business Insights and Analysis**

**3.2.1 What is the distribution of cancellations and non-cancellations across different months of the year?**

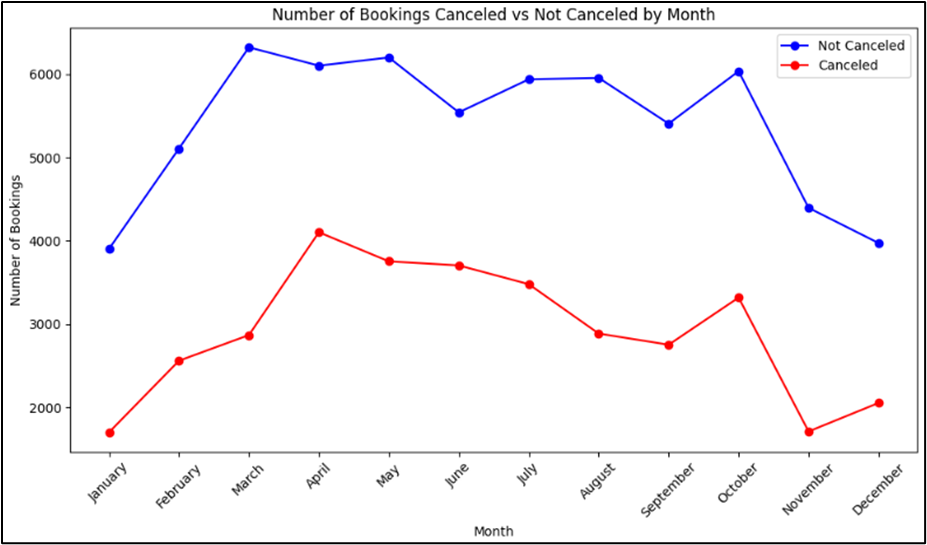


Figure 3.3: Total Number of Bookings Cancelled VS Not Cancelled for Both Hotel

The blue line, representing non-cancelled bookings, shows the trend in successful bookings. Based on the observation, it is noticeable that non-cancelled bookings generally peaked between March and October, with slight variations. For instance, the highest number of successful bookings can be seen consistently in March and May. However, towards the latter part of the line, a noticeable dip can be seen starting October, with a significant decrease in bookings by November and December.

The red line shows the bookings which are cancelled and from the graph, it follows an interesting trend. In the month of April and October, the number of cancellations is relatively high, after which there is a decline. In other words, the graph shows that more bookings are cancelled in the early and toward end months of the year, with the peak being in April and decreasing towards the end of the year.

This graph highlights how seasonal trends affect the rate of both successful and cancelled bookings. This information can be useful for the hotel managements to adjust their strategies accordingly, either by offering incentives to curb the cancellation issues during peak season or organizing marketing campaigns during off-peak times to attract more bookings.

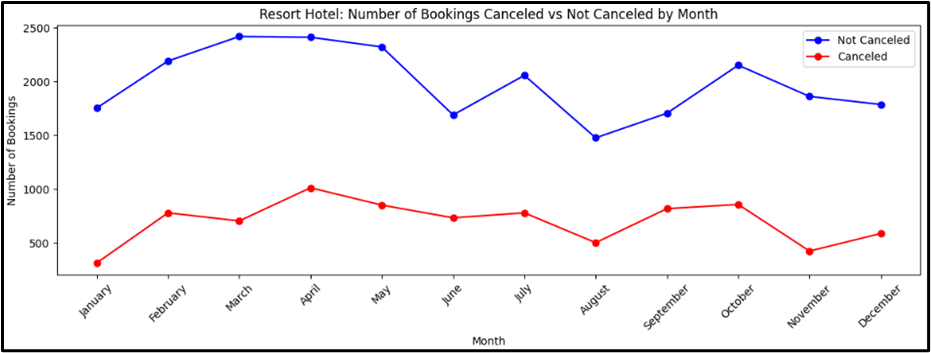


Figure 3.4: Number of Bookings Cancelled VS Not Cancelled by Resort Hotel

There are some key insights that can be explored from this graph. In the month of April and July, during the warmer months, resort hotels tend to have a higher volume of non-cancelled bookings. However, there is a slight decrease after these months, but the numbers remain steady throughout the year. For cancellations, the graph shows a consistent pattern, with higher cancel rate during the early and middle part of the year which are in April and July. The cancellation rate then slightly decreases towards the end of the year.

By looking at the graph, a higher number of bookings during peak vacation months can be seen in the resorts hotel as people prefer leisure travel. The cancellation trend signals that customers might cancel their booking because of unforeseen circumstances such as sickness or change of plans (Antonio et al., 2017). This information can help the resort management to be more flexible with booking policies and have marketing strategies that provide promotions for last minute bookings.

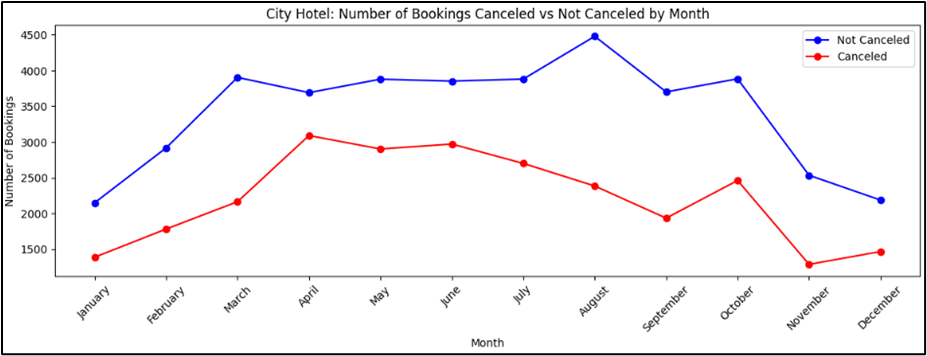


Figure 3.5: Number of Bookings Cancelled VS Not Cancelled by City Hotel

There is a similar pattern for non-cancellation bookings for the resort hotel, with the highest successful bookings rate occurring between March and August. However, towards the end of the year (November and December), there is a sharper decline.

Like resort hotels, the number of cancellations bookings peak in the month of April. However, city hotels have a constant level of cancellations throughout the year, although the number begins to drop towards the end of the year.

City hotels are experiencing significant bookings during the spring and summer months likely due to business travel and tourism (Bausch et al., 2021). Uncertainty in business travel or last-minute changes may contribute to the higher rate of cancellation.

These charts provide a better understanding into how the customer behaviour differs throughout the year and how these patterns fluctuate between resort and city hotels. Having insights on the rate of cancellation is crucial for city hotels as they can use this data to optimize their operations.

**3.2.2 What is the geographical distribution of hotel guests and which countries do the majority of guests come from?**



Figure 3.6: Global View of Booking Volumes

The above map provides an overview of global bookings with various shades of colours indicating the number of bookings. The higher booking volumes are represented by darker shades whereas the lower booking volumes or no bookings are represented by lighter shades. From this map, the regions with the most significant hotel booking activities are in Europe, South America, and parts of Asia.



Figure 3.7: Top 10 Countries with The Highest Bookings

The top ten countries with the highest booking volumes which are Portugal, Great Britain, France, Spain, Germany, Italy, Ireland, Belgium, Brazil, and the Netherlands. These countries are highlighted to represent their booking intensity. Countries in Europe, such as Portugal and Spain, along with countries South America, such as Brazil, have a darker shade signifying higher bookings. This map helps the hotel management to identify the countries which dominate hotel bookings, thus can assist them in marketing and promotional efforts towards these regions.

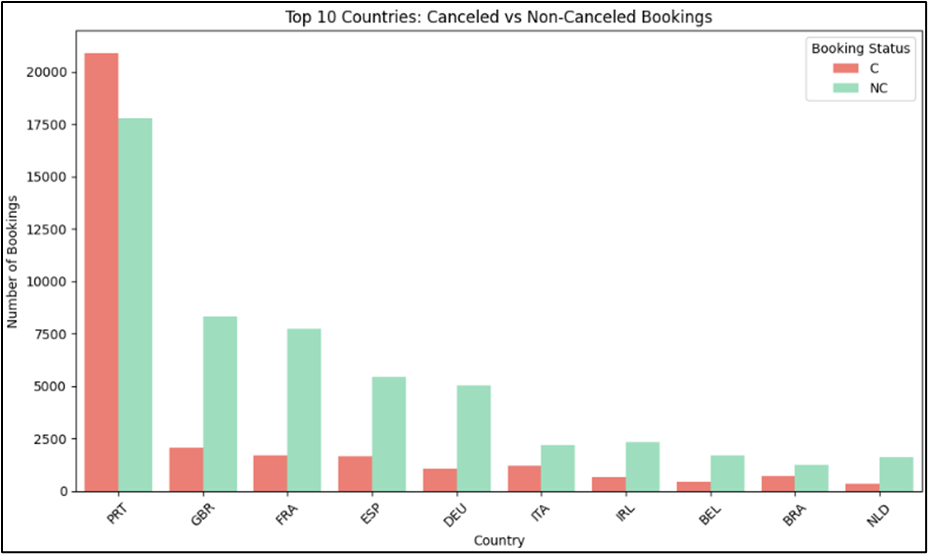


Figure 3.8: Top 10 Countries with The Highest Booking Volumes

The above bar chart shows the number of cancelled and non-cancelled bookings in the top 10 countries. By looking at the chart, Portugal (PRT) has the highest overall booking volume and the rate of cancelled is more than non-cancelled. The high number of cancellations suggest that the customers may make last minutes changes of their plans, or the hotel has flexible booking policies.

Great Britain (GBR) and France (FRA) have notable rates of cancellation, but the gap between cancelled and non-cancelled bookings is lower than Portugal’s. In contrast, countries like Italy (ITA) and Ireland (IRL) have a reverse pattern where the number of non-cancelled is higher than cancelled bookings, which indicate that these countries have a reliable reservation. Brazil (BRA) and Netherlands (NED) have relatively lower total bookings.

By looking at the booking patterns by country, Portugal, Great Britain, and France dominates in terms of total bookings. However, these countries also experience high cancellation rates. This can be explained by the flexible cancellation policies or other factors that drive last-minute booking changes by guests from these countries. Countries like Ireland and Italy show a more stable booking patterns with fewer cancellation. By understanding these booking cancellation patterns, hotel managements can have specific marketing strategies for specific regions. For instance, countries like Portugal may benefit from a campaign that incentivizes customers to book their stay early or impose penalties for last-minute cancellations. Countries like Italy and Ireland could maintain their booking policies as these countries have more reliable customers.

**3.2.3 How do hotel rates fluctuate across the year, and during which periods can guests secure the most affordable bookings?**

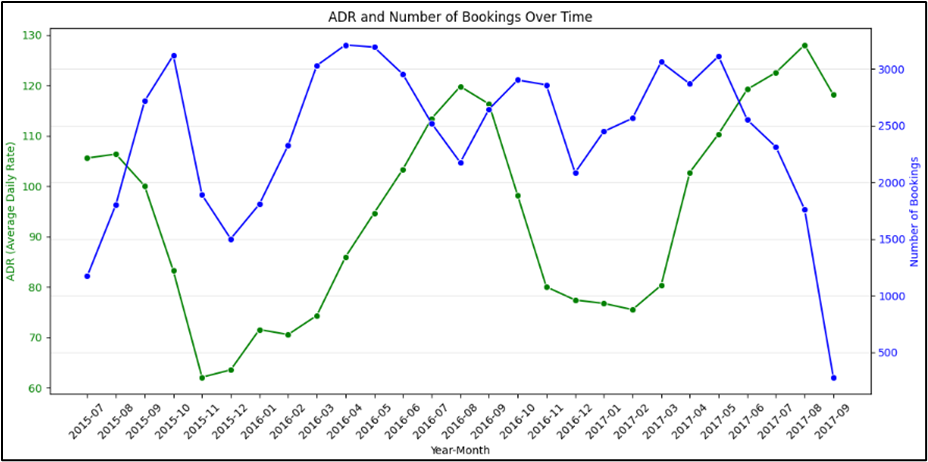


Figure 3.9:Average Daily Rate (ADR) and The Number of Bookings Over Time

The chart above highlights the key insight for hotel management in identifying the seasonal trends and demand patterns. According to Boffa & Succurro, (2012), factors such as holidays, special events, and seasonal trends may fluctuate the occupancy in the hotels. This type of data will provide useful information for the hotel management to make strategic decision-making, especially in staffing, pricing and marketing efforts.

From the chart above, there are two significant trends, which are the number of booking and the Average Daily Rate (ADR). The relationship between these two variables provides useful information in understanding how demand influences pricing. The green line represents the average daily rate over time. This trend gives a better picture in observing the price fluctuations across the months and years. The blue line represents the number of monthly bookings over time, excluding cancellations.

Upon analysing the charts, there are several observations and key trends that can help hotel management to optimize hotel operations. Firstly, one significant observation in the chart is the inverse relationship between the number of bookings and the ADR. For instance, during the high booking periods, the ADR tends to increase but eventually plateaus or even drops as the number of bookings start to decline. By looking at this trend, it can be said that when the demand is high, hotels capitalize on the increase of occupancy by increasing the rate. However, when they push the prices too high, it may discourage further bookings, thus leading to a drop in the number of bookings. The understanding in optimizing the rate and the demand is critical for revenue management.

Secondly, the chart demonstrates repeating seasonal trends which suggest specific months or seasons have higher amounts of occupancy. For example, the graph shows a sharp increase in bookings followed by a steep declined. This can be explained by looking at specific months or seasons when demand surges, possibly due to holidays, tourist seasons, or local events. The hotels respond by increasing their ADR during the peak season to maximize revenue. However, the ADR tends to decrease during the low-demand periods. A strategy that may be used by hotels is to offer lower rates to attract more bookings.

The next observations that can be seen in the chart are the demand fluctuation and pricing strategy. The ADR line fluctuations align with booking trends but there are some instances where ADR remains relatively stable, even when the number of bookings drop significantly. This trend can be explained by understanding the pricing strategy made by the hotels. The hotels are maintaining the price despite lower demand, potentially due to the operational cost or contractual obligations with corporate clients (Boffa & Succurro, 2012). By having these limitations, the hotels must adjust the price in a way that will both not hinder them and attract more guests.

Towards the latter part of the line graph, there is a sharp decline in both ADR and the number of bookings. This explains the notable drop in demand which may be attributed to post-peak season or economic factors that affect consumer behaviour. This sudden drop in both trends may prompt hotel management to reassess their marketing campaigns, special promotions, and pricing strategy to help increase the number of bookings during the off-peak periods.

This chart also helps the hotel managers to look for valuable insights regarding forecasting future demands. By analysing the past trends in bookings and ADR, they can predict future peaks and low seasons, which can help them to make decisions regarding staff, inventory management, and marketing efforts. On top of that, the graph provides insights into how flexible pricing strategies can be used to maintain occupancy rates during off-peak periods while maximizing revenue during peak times. By having a good analyzation on these metrics, hotels can fine-tune their operations and improve overall profitability.

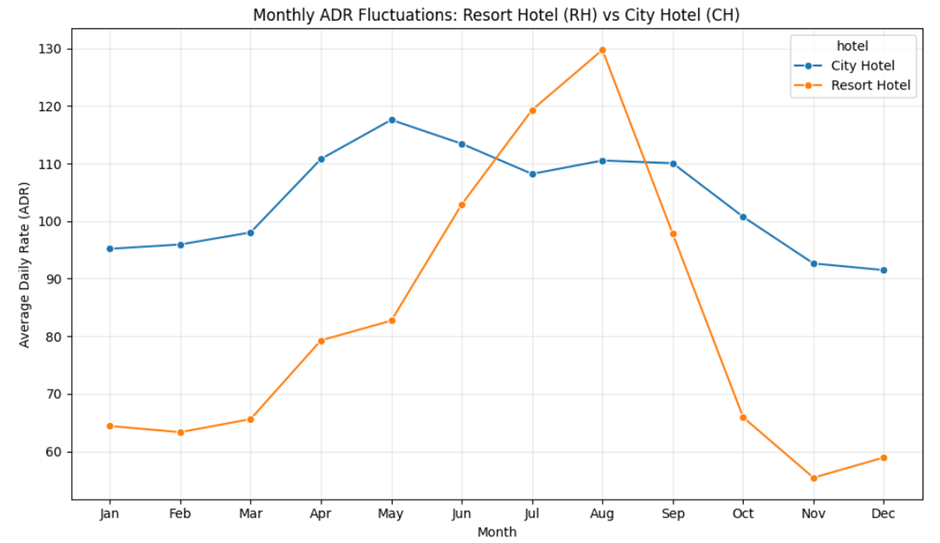


Figure 3.10: Monthly Average Daily Rate (ADR) Fluctuations for Both Hotels throughout The Year

The line chart above shows the Monthly Average Daily Rate (ADR) fluctuations for Resort and City Hotels throughout the year.

**Analysis for City Hotel ADR Fluctuations:**

The ADR for City Hotels is relatively stable compared to Resort Hotels, as the price fluctuates within a small range, which is between 90 and 120. This indicates that there is a steady demand for City Hotels, most likely because of business travellers and activities that happen throughout the year, which are less affected by seasonal changes.

**Analysis for Resort Hotel ADR Fluctuations:**

The ADR fluctuates significantly for Resort Hotels, with a sharp increase starting from May and peaking in August at an ADR of 130. Then, it declines rapidly after September, and it reached the lowest ADR of approximately 30 in November. This pattern portrays that Resort Hotels are much more popular and in higher demand in the summer than at other times of the year.

**Comparing Winter and Summer Trends for Resort Hotel:**

During the winter months, ranging from January to March and October to December, the ADR for Resort Hotels is much lower, indicating reduced demand. On the other hand, during the summer months, from June to September, the ADR for Resort Hotels has increased notably and reached its peak in August, suggesting that the price increases when demand is high. This trend aligns with the findings by Lai et al. (2022), which suggests that summer is the most popular season for travel in most European countries due to two key factors. Firstly, many people take vacations during the summer holidays as schools and businesses are often closed or operating at a reduced capacity. Secondly, the warm temperatures serve as a major attraction. The study by Lopes et al. (2021) further supports that warm temperatures significantly contribute to attracting tourists in Portugal, as they generally felt comfortable exploring and visiting the country, which in turn, facilitated greater mobility.

**3.2.4 How does booking through agents and companies influence a hotel’s decision to tailor its marketing strategies?**

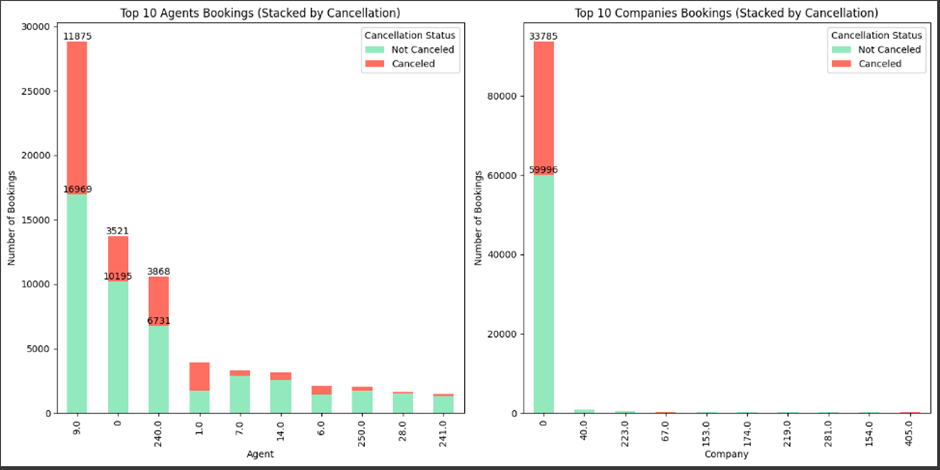


Figure 3.11: Top 10 Bookings made by Agents and Companies

The visualization above consists of two bar charts representing the top 10 bookings made by agents and companies, respectively. It is crucial to study the number of bookings made by agents and companies to tailor marketing strategies.

**Analysis from Agent Bookings Bar Chart:**

Agent 9 has the highest total number of bookings with 28,844 bookings and is significantly higher than any other agents. From the 28,844 bookings, 11,875 bookings were cancelled and the remaining 16,969 bookings were cancelled. The rank is then followed by Self Bookers and Agent 240, with 13,716 and 10,599 bookings respectively. Self Bookers have 3,521 and 10,195 cancelled and non-cancelled bookings respectively, whereas Agent 240 has 3,868 and 6,731 non-cancelled bookings respectively. Despite being ranked the second, it should be noted that Agent 9’s bookings are double of Self Bookers bookings. Furthermore, there is a steep decline after the top three agents, starting with Agent 1. The remaining agents (Agents 7, 14, 6, 250, 28, and 241) have fewer than 5,000 bookings each.

**Analysis from Companies Bookings Bar Chart:**

Self Bookers dominate the company booking numbers, with 93,781 bookings, where they clearly have the most bookings than all other companies combined. From the 93,781 bookings, 33,785 bookings were cancelled and the remaining 59,996 were not cancelled. Company 40 ranked the second and there is a significant gap between Self Bookers and company-mediated bookings. The remaining companies’ (Company 40, 223, 67, 153, 174, 219, 281, 154 and 405) booking numbers declined rapidly with very few bookings.

**Tailored Marketing Strategies:**

Based on the analysis of agent and company bookings, the hotel should strengthen their relationships with top agents and companies like Agent 9 and Agent 240 to maintain high booking levels. At the same time, hotels should also enhance direct booking experience for Self Bookers by improving their hotel’s official website, offering special discounts, and loyalty programmes. Research by Augustine and Adnan (2020) has shown that website trust affects consumers’ booking intentions positively, emphasizing the significance of developing a trustworthy website to attract more direct bookings. In short, by maintaining good connections with agents and companies as well as enhancing the direct booking experience, the hotel can increase bookings and improve customer retention.

**3.2.5 Other Discussion**

**Total Nights**

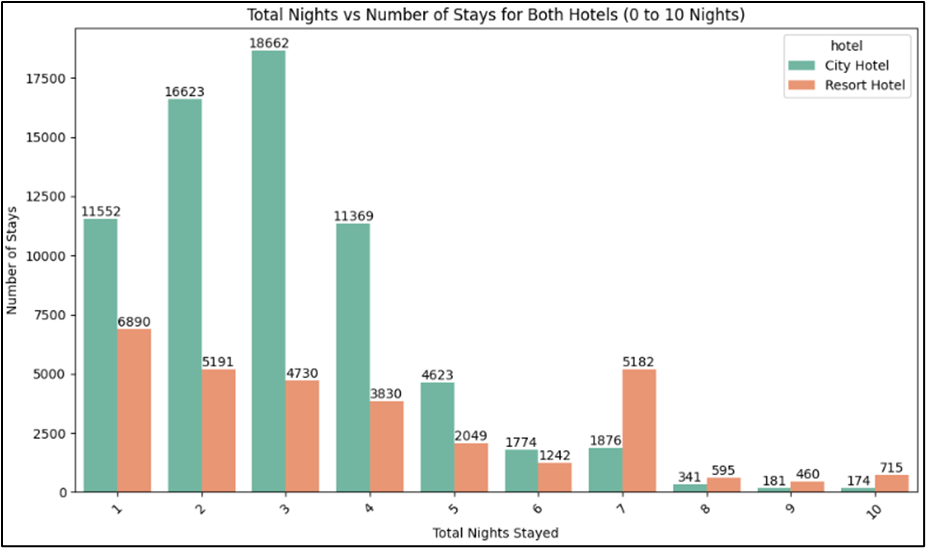


Figure 3.12: Total night VS The Number of Stays for Both Hotel

The bar chart above displays the total nights spent at resort and city hotel, over a period ranging from 1 to 10 nights. The number of nights stayed is represented by the x-axis and the number of stays corresponding to the duration is represented by the y-axis.

By analysing the bar chart, there are several notable patterns that have emerged. Firstly, the most common length of stay for both city and resort hotels are 3 nights and 1 night respectively. This suggests that city hotels cater a higher proportion of people who needed more time to stay, likely business travellers or tourists who plan for a longer visit. This can be attributed to the urban locations of city hotels, where businesspeople can easily meet each other and for tourists to have better access to transportation systems for travelling.

As the number of nights increase, the stays at both hotels start to decline. For city hotel, the trend starts to decrease significantly after 4 nights. This downward trend suggests that on average, guests who spend their night(s) at city hotels are unlikely to extend their stay after the fourth night, possibly because majority of the guests are businesspeople, where they will only stay for a few nights for business purposes.

In contrast, resort hotels show a different pattern for longer stays. From 7 to 10 nights, resort hotels start to show a higher proportion of stays compared to city hotels. For instance, on the seventh night, resort hotels have 5,182 stays while city hotels have 1,876 stays. This trend shows that resort hotels are more appealing for extended stays, possibly due to its leisure-oriented services and tranquil environments that they offered, resulting in guests to opt for a longer stay.

In short, while city hotels attracted shorter stays, resorts hotels are more popular for extended visits. This is due to the different guest behaviours that likely reflect the nature of the locations and facilities associated with each hotel type.

**Guest combinations**

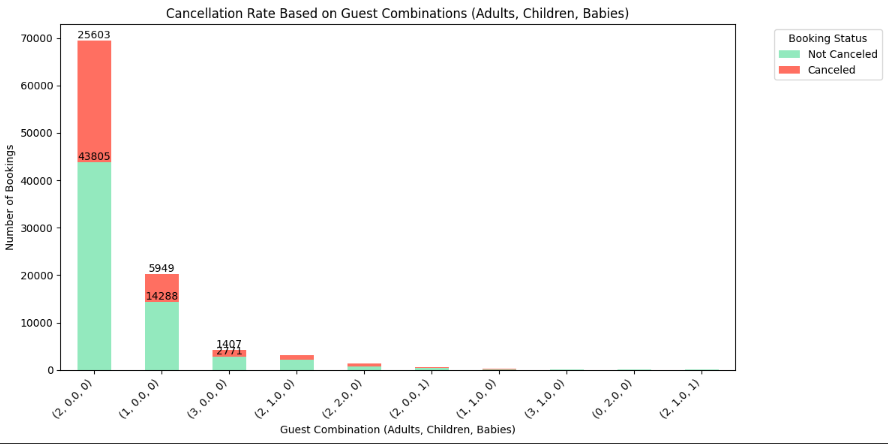


Figure 3.13: The Number of Cancellation Rate based on The Guest Combination

The above stacked chart shows the cancellation rate based on the different guest combinations which comprise of adults, children, and babies. The type for guest combinations is shown on the x-axis while the number of bookings is shown on the y-axis. The red colour bars represent the bookings that are cancelled, whereas the green colour bars represent the bookings that are successful.

There are some important observations that may be observed. Firstly, the largest group consists of 2 adults, 0 children, and 0 babies (2,0,0) with a total of 113,213 bookings. Among this group combination, 43,805 are not cancelled while 69,408 are cancelled. By looking at these trends, it shows that this type of group has a significant cancellation rate. This guest combination usually represents typical adult travellers such as couple or business professionals, who seems to have a higher uncertainty, which results in a high likelihood of cancelling their bookings.

For bookings which have children and babies, it has different trends compared to the guest combinations with adults only. As portrayed in the graph, the group with children and babies have a relatively small number of cancellations compared to the groups that only have adults. Although the number of guest combinations with children and babies are small, they have the lowest number of cancellations compared to others. This shows that booking with children might face a slightly lower number of cancellations as family trips are usually well planned (Chen et al., 2023).

**Deposit type**

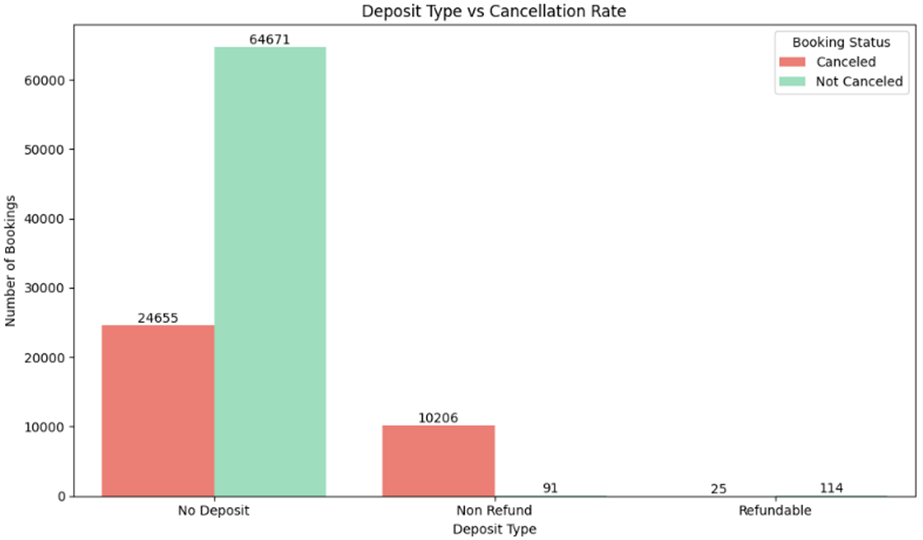


Figure 3.14: Cancellation Rate by Deposit Type

The clustered bar chart above shows the types of deposits (No Deposit, Non Refund, and Refundable) made by customers and their corresponding booking status, whether cancelled or not cancelled. The total number of bookings with ‘No Deposit’ type is 89,326, of which 24,655 were cancelled and 64,671 were not cancelled. The high cancellation rate for bookings with no deposit indicates that customers are willing to cancel when no financial commitment is involved. For bookings, with ‘Non Refund’ deposit type, there were 10,206 cancellations and 91 customers who did not cancel. Even though the bookings were non-refundable, many customers still cancelled. This portrays that having a non-refundable deposit reduces the probability of cancellation, but it cannot be eliminated completely. This might be due to reasons such as changes in travel plan or any emergency events prior to the travel that leads to cancellation. Furthermore, there were 139 bookings in this category with 25 customers cancelling and 114 not cancelling. Although the total bookings are small, the cancellation rate is lower. This suggests that customers feel comfortable cancelling because they are confident that there is a guarantee of getting their money back.

**Repeated guest**

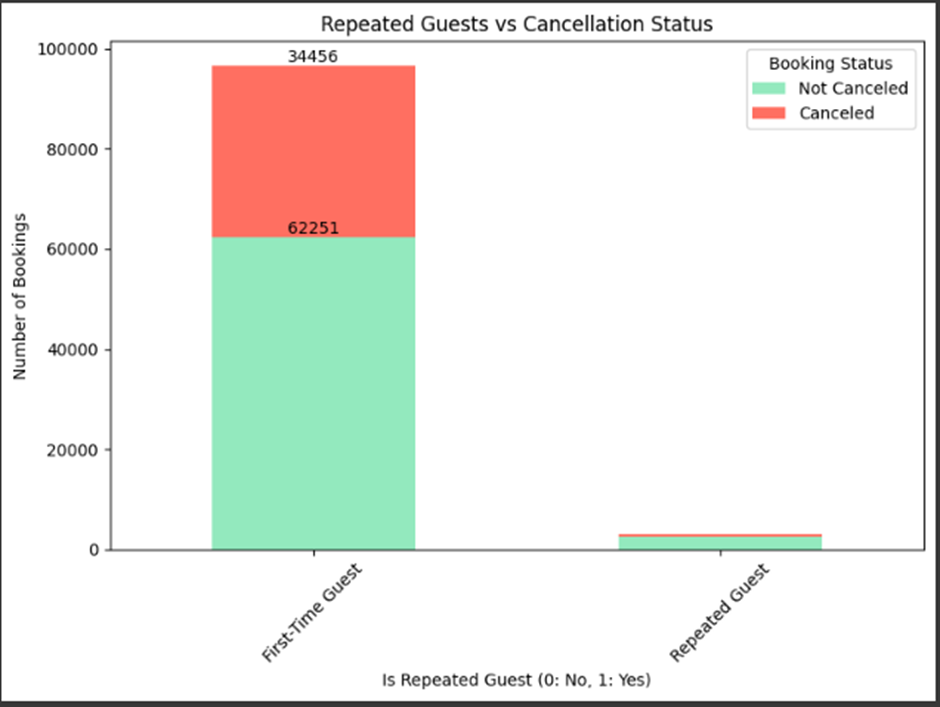


Figure 3.15: Cancellation Rate by Repeated Guests

The stacked bar chart above shows a comparison of the number of bookings between first-time guests and repeated guests. For first-time guests, out of the 96707 bookings made, 62,251 were not cancelled, while the remaining 34,456 bookings were cancelled. This gives a cancellation rate of approximately 35.6% which is relatively high. For repeated guests, the portion of cancelled bookings is visibly much smaller, suggesting a significantly lower cancellation rate. The lower cancellation rate among repeated guests indicates higher customer satisfaction with their previous experiences. This is consistent with the findings of Rather et al. (2021), who highlighted that repeat customers often have higher level of emotional engagement and value cocreation which increases the likelihood of customers revisiting. On the other hand, the higher cancellation rate among first-time guests shows that they might be less confident to maintain their booking.

**Difference between reserved room type and assigned room type**

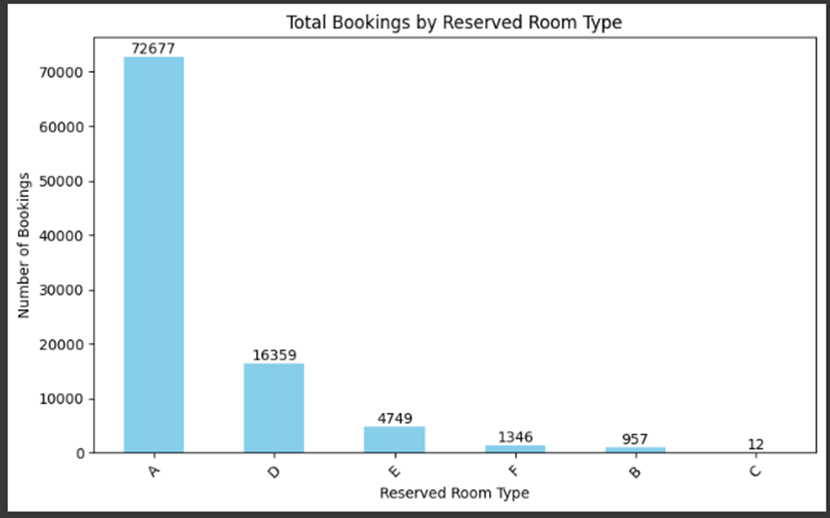


Figure 3.16: Total Bookings by Reserved Room Type

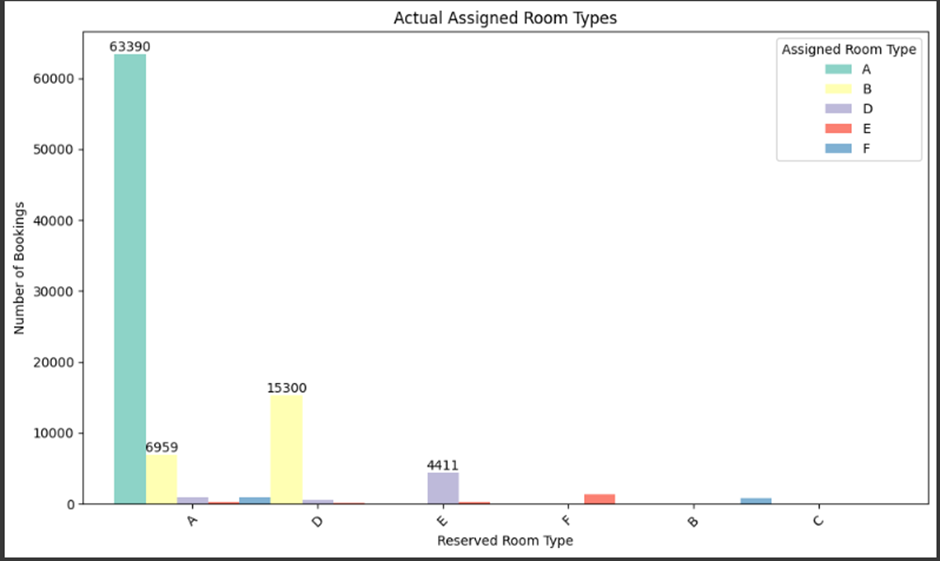


Figure 3.17: Actual Assigned Room Types

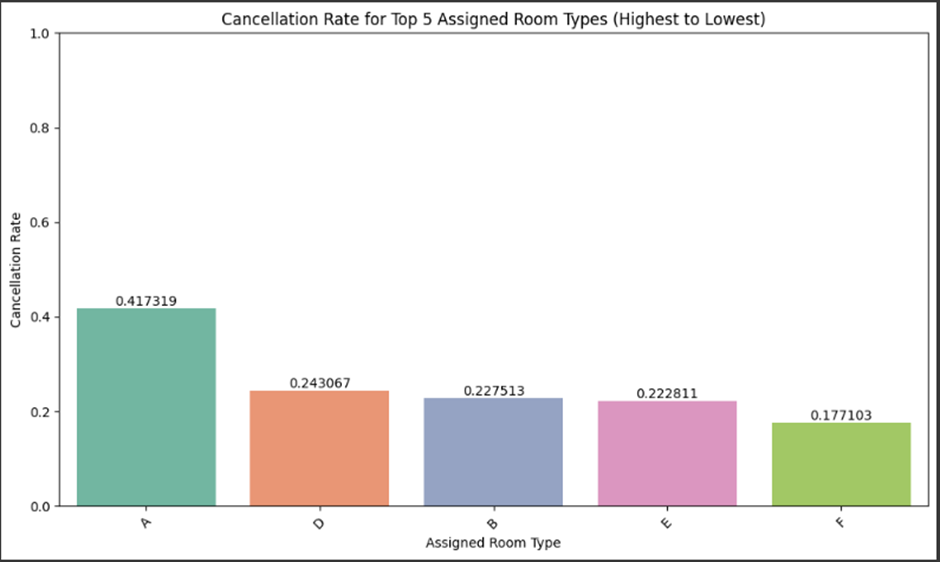


Figure 3.18: Cancellation Rate for Top 5 Assigned Room Types

**Analysing total bookings by reserved room type:**

Figure 3.16 shows the number of bookings made for different room types, which are A, B, C, D, E, and F. Room A has the highest number of bookings with 72677 bookings. This shows that Room A is the most popular and highly preferred room type. Room D comes in second with 16359 bookings. The gap between Room A and Room D is significant, suggesting that Room A is customers’ favourite which might be due to affordability, availability, and comfort. Next is Room E which has 4749 bookings, and this room type has a much smaller number of bookings compared to the first two room types. Room F, B, and C have fewer bookings with 1346, 957 and, 12 bookings respectively. These three room-types are not preferred possibly due to high pricing or discomfort; therefore, the hotel should consider improving these aspects to attract more customers.

**Analysing actual assigned room types:**

Figure 3.17 displays the actual room types that were assigned to customers. Out of 72677 customers who reserved Room A, 63390 customers received Room A, giving a percentage of 87.2% of customers who were given the room they wished for. The remaining 9287 rooms were assigned to different rooms than originally booked. For Room D, out of 16359 bookings, 15300 were assigned to Room D, which gives a percentage of 93.5%. This is followed by Room E, where the number of bookings that were actually assigned to the same room was 4411 out of 4749 total bookings giving a percentage of 92.9%. The high assignment rates for Rooms A, D and E suggest effective management of availability and booking systems for these room types.

**Analysing cancellation rate for the top five assigned room types:**

The chart shows the cancellation rates for the top five assigned room types. Room A has the highest cancellation rate, at around 0.42. This suggests that a significant number of customers who booked this room type cancelled their reservations. Room D, B and E have almost the similar cancellation rate between 0.22 and 0.24, indicating a relatively low cancellation rate. Room F has the lowest cancellation rate which is about 0.18.

**Implications for business decision making**

Based on the analysis through visualizations, the business-related problems are able to be answered. First of all, the distributions of cancellations and non-cancellations were plotted out through various line charts, enabling the ability to compare the number of cancellations and cancellations throughout the months of the year. Exploring further, more specific line charts were also created to compare the number of cancellations and non-cancellations based on the hotel, one chart for Resort Hotel and one chart for City Hotel. By analyzing these line charts, comparisons are able to be made along with identifying trends. Based on the findings, the majority of cancellations were made in April. The management may identify this trend and make decisions based on this trend, such as adjusting their marketing strategies or providing incentives during the spring season.

Looking into the geographical distribution of hotel guests, maps have been plotted, illustrating where the majority of guests are from. Based on the visualizations, it was found that the majority of bookings were made by guests from Europe, South America, and Asia. Additionally, a bar chart has been plotted to compare the top 10 countries with the most bookings. By observing the bar chart, the management can identify which country has the most bookings made from, in this case, Portugal. Based on this information, the management may want to invest more in target marketing and promoting towards these regions.

The hotel rates have been explored throughout the year in order to identify the most affordable booking periods for guests. This was achieved by plotting line charts that compare the ADR and number of bookings of both hotels over time. By comparing the ADR line with the bookings line, observations regarding how demand affects price may be made. Based on observations, the relationship between ADR and the number of bookings is inverse, meaning that if one were to increase, the other would decrease. This information can be useful for the management as they may want to optimize their prices based on the number of bookings in order to maximize revenue. This information may also be useful for guests as they will be able to identify when the best time is to make bookings.

Finally, in terms of analyzing the influence of how booking through agents and companies affect a hotel’s marketing strategies, bar charts were plotted, comparing the top 10 bookings made by agents and companies. By comparing these bar charts, Agent 9 has the highest number of bookings, and in terms of booking companies, the majority of guests are self-bookers, meaning they do not book through a company. This information is useful for the hotel management as they may adjust their marketing strategies by strengthening their relationship with the highest booking agents and enhancing their direct booking experience for self-bookers.

## **3.3 Menu application**

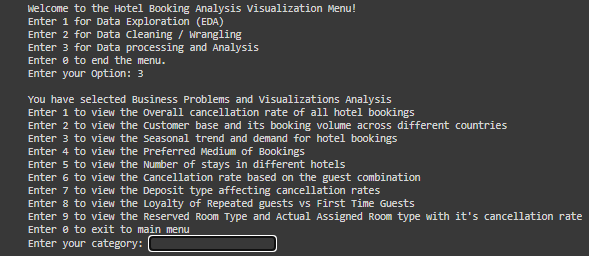


Figure 3.19: Menu Application

The figure above is a screenshot showcasing the menu application that was developed using Python. Upon executing the selected code, a main menu interface, referred to as the "menu gate," will be displayed. This interface allows users to interact with the application by selecting a specific option from the menu, each assigned a unique number. Users can navigate through various sections of the application, which include preprocessing stage, Exploratory Data Analysis (EDA) as well as business problems and visualization for insights. Additionally, there are options to visualize key findings, trends, and metrics through interactive charts and graphs. The sub-menu here will guide users to different types of visualizations related to business problem-solving. Users can select the exit option to terminate the application when they are done exploring the sections or wish to close the program. Each main session within the menu application is designed with a corresponding sub-menu, providing a clear and structured way for users to navigate and perform specific tasks.

# References

Antonio, N., Almeida, A. de, & Nunes, L. (2017). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. Tourism & Management Studies, 13(2), 25–39. <https://doi.org/10.18089/tms.2017.13203>

Antonio, N., De Almeida, A., & Nunes, L. (2019). Big data in hotel revenue management: Exploring cancellation drivers to gain insights into booking cancellation behavior. *Cornell Hospitality Quarterly*, *60*(4), 298-319.

Augustine, A. A., & Adnan, W. H. (2020). The effects of perceived price, website trust and online reviews on online hotel booking intention in Kuala Lumpur. *GSJ*, *8*(6).

Bausch, T., Gartner, W. C., & Humpe, A. (2021). How weather conditions affect guest arrivals and duration of stay: An alpine destination case. International Journal of Tourism Research, 23(6), 1006–1026. <https://doi.org/10.1002/jtr.2459>

Boffa, F., & Succurro, M. (2012). The impact of search cost reduction on seasonality. Annals of Tourism Research, 39(2), 1176–1198. <https://doi.org/10.1016/j.annals.2012.01.006>

Chen, S., Ngai, E. W. T., Ku, Y., Xu, Z., Gou, X., & Zhang, C. (2023). Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction. Decision Support Systems, 170. <https://doi.org/10.1016/j.dss.2023.113959>

Lai, S., Sorichetta, A., Steele, J., Ruktanonchai, C. W., Cunningham, A. D., Rogers, G., ... & Tatem, A. J. (2022). Global holiday datasets for understanding seasonal human mobility and population dynamics. *Scientific Data*, *9*(1), 17.

Lopes, H. S., Remoaldo, P. C., Ribeiro, V., & Martin-Vide, J. (2021). Perceptions of human thermal comfort in an urban tourism destination–A case study of Porto (Portugal). *Building and Environment*, *205*, 108246.

Rather, R. A., Hollebeek, L. D., & Rasoolimanesh, S. M. (2022). First-time versus repeat tourism customer engagement, experience, and value cocreation: An empirical investigation. *Journal of Travel Research*, *61*(3), 549-564.