Debby Rofiko Malik

Data Analysis Portfolio

2025

Education

Tahun
2021 - 2025

Biomedical Eng. - Bachelor Degree
Universitas Indonesia

Work Experience

Tahun
2023 - 2024

3D Design Engineer
Covent Indonesia

Past Projects

- Hotel Booking Analysis
- Vendor Performance Analysis

Tools

Power BI

SQL

Hard Skills

Math and Statistics | Data Cleaning | Data Visualization | Machine Learning

Soft Skills

Critical Thinking | Problem Solving | Communication | Teamwork | Adaptability | Leadership | Time Management | Creativity

Certifications

- Data Science & Data Analysis (MySkill)
- Microsoft Excel, Word, and PowerPoint (MySkill)

Contact Me

Debby Rofiko Malik

debbymalik21@gmail.com

github.com/debbyrofikomalik

Hotel Booking Analysis

Python (Google Colab) | Data Cleaning | Exploratory Data Analysis | Data Visualization

</> Access Code Here

Problem

In recent years, City Hotel and Resort Hotel have **experienced high cancellation rates**. As a result, both hotels are facing several challenges, including **reduced revenue and suboptimal room utilization**. Therefore, lowering cancellation rates has become a primary objective for both hotels in order to **enhance efficiency and maximize revenue generation**.

About the Dataset

The dataset comprises 119.390 hotel bookings from two types of hotels: a City Hotel and a Resort Hotel. Each entry corresponds to a single booking made between July 1, 2015 and August 31, 2017, including reservations that were completed and those that were canceled. It is available on Kaggle.

Data Dictionary

Column	Explanation
hotel	Indicates whether the booking was for a City Hotel or a Resort Hotel
is_canceled	Shows whether the booking was canceled (1) or not (0)
lead_time	Number of days between booking date and arrival date
arrival_date_week_number, arrival_date_day_of_month, arrival_date_month	Week number, day date, and month number of arrival date

Column	Explanation
stays_in_weekend_nights, stays_in_week_nights	Total nights the customer booked for weekends (Saturday and Sunday) and weekdays (Monday to Friday)
adults, children, babies	Number of adults, children, babies booked for the stay
meal	BB – Bed & Breakfast, HB - Half Board, FB- Full Board, SC - Self Catering
market_segment	Booking market segment, where 'TA' represents Travel Agents and 'TO' represents Tour Operators
distribution_channel	Booking distribution channel, with 'TA' for Travel Agents and 'TO' for Tour Operators
is_repeated_guest	Binary value indicating if the booking name was from a repeated guest (1) or not (0)
previous_cancellations	Number of previous bookings that the customer canceled before the current booking
previous_bookings_not_canceled	Number of previous bookings made by the customer that were not canceled prior to the current booking

Column	Explanation
reserved_room_type	Code representing the type of room reserved. Codes are used instead of names for privacy reasons.
assigned_room_type	Code for the room type actually assigned. This may differ from the reserved room type due to operational reasons (e.g., overbooking) or customer requests.
booking_changes	Total number of amendments or modifications made to the booking.
deposit_type	Type of deposit associated with the booking:: 1. No Deposit – no deposit was paid, Non Refund – full-stay deposit paid, non-refundable., 3. Refundable – partial deposit paid, refundable.
agent	ID of the travel agency responsible for making the booking.
company	ID of the company responsible for the booking payment. IDs are anonymized.
days_in_waiting_list	Number of days the booking remained on the waiting list before confirmation.

Column	Explanation
customer_type	Type of customer for the booking: 1. Group – booking associated with a group, 2. Transient – individual booking not associated with a group or contract, 3. Transient-party – individual booking associated with at least one other transient booking.
adr	Average Daily Rate, calculated by dividing total lodging revenue by the total number of nights stayed.
required_car_parking_spaces	Number of car parking spaces requested by the customer.
total_of_special_requests	Number of special requests made by the customer (e.g., twin bed, high floor).
reservation_status	Final booking status: 1. Check-Out – customer checked in and has departed, 2. No-Show – customer did not check in and informed the hotel of the absence,
reservation_status_date	Date when the final booking status was recorded. This can be used alongside reservation_status to determine when a booking was canceled or when a customer checked out.

Assumptions

No unusual occurrences between 2015 and 2017 have had a substantial impact on the data used

The information remains relevant and can be applied to efficiently analyze potential hotel strategies.

There are no unforeseen negative consequences for the hotels in implementing any of the recommended techniques.

The hotels are not currently using any of the suggested solutions.

The biggest factor affecting the effectiveness of earning income is bookig cancellations.

Cancellations result in vacant rooms for the booked length of time.

Clients make hotel reservations the same year they make cancellations.

Research Question & Hypothesis

What variables affect hotel reservation cancellations?

How can hotel reservation cancellations be reduced?

How can hotels be supported in making pricing and promotional decisions?

More cancellations occur when prices are higher.

When there is a longer waiting list, customers tend to cancel more frequently.

The majority of clients are coming from offline travel agents to make their reservations.

Exploratory Data Analysis & Data Cleaning

Q1: Are all columns necessary?

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 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_week_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  country                         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  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
 32  name                            119390 non-null  object
 33  email                           119390 non-null  object
 34  phone-number                    119390 non-null  object
 35  credit_card                     119390 non-null  object
dtypes: float64(4), int64(16), object(16)
memory usage: 32.8+ MB
```

Data Inspection

```
# Display the number of rows and columns in the dataset
df.shape
```

```
(119390, 36)
```

There are **119.390 rows** and **36 columns**

```
# Display the list of all column names in the dataset
df.columns
```

```
Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
       'arrival_date_month', 'arrival_date_week_number',
       'arrival_date_day_of_month', 'stays_in_weekend_nights',
       'stays_in_week_nights', 'adults', 'children', 'babies', 'meal',
       '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',
       'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date', 'name', 'email',
       'phone-number', 'credit_card'],
      dtype='object')
```

Identified Issues

The dataset **contains personally identifiable information** (name, email, phone number, credit_card).

Action Taken

Removed those columns to ensure privacy.

Result

```
df.shape
```

```
(119390, 32)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 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_week_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  country                         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  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

Identified Issues

- The **reservation_status_date** was initially stored in **incorrect data type** (object/text).
- This **prevents** performing **time-based analysis.**

Action Taken

Converted reservation_status_date to datetime format.

```
# Change 'reservation_status_date' to datetime format to enable time-based analysis
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'])
```

Result

```
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(4), int64(16), object(11)
memory usage: 29.1+ MB
```

Q2: Are data types correctly assigned for all features?

Data Pre-Processing

Q3: Are there any missing values or outliers?

```
Missing values per column:
hotel                                0
is_canceled                          0
lead_time                            0
arrival_date_year                    0
arrival_date_month                   0
arrival_date_week_number             0
arrival_date_day_of_month            0
stays_in_weekend_nights              0
stays_in_week_nights                 0
adults                               0
children                             4
babies                               0
meal                                 0
country                            488
market_segment                       0
distribution_channel                 0
is_repeated_guest                    0
previous_cancellations               0
previous_bookings_not_canceled       0
reserved_room_type                   0
assigned_room_type                   0
booking_changes                      0
deposit_type                         0
agent                            16340
company                         112593
days_in_waiting_list                 0
customer_type                        0
adr                                  0
required_car_parking_spaces          0
total_of_special_requests            0
reservation_status                   0
reservation_status_date              0
dtype: int64
```

Identified Issues

The children column has 4 missing entries, country has 488, agent has 16.340, and company has 112.593

Action Taken

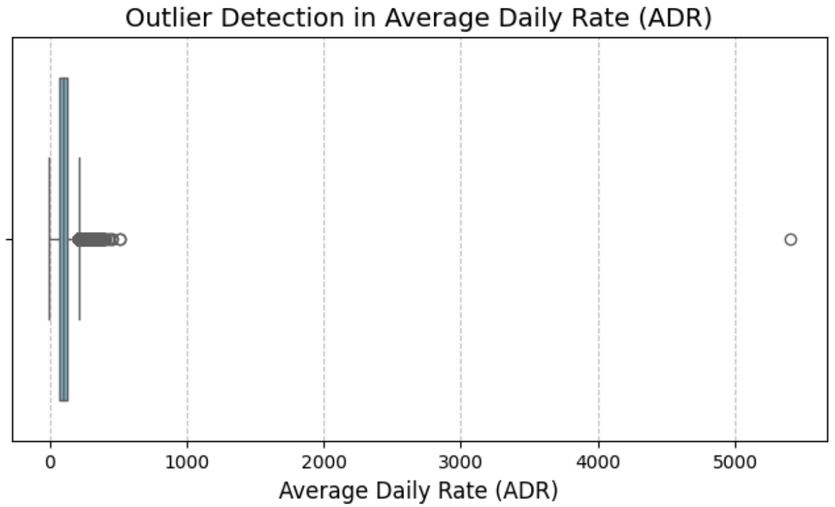
- The **missing values are completely at random** (MCAR), meaning the **missingness does not depend on other observed values** or variables.
- **Company column** contains **95% missing values making imputation impratical** and it is also considered non-critical to the analysis
- Other columns with missing values can be safely deleted since the dataset is large enough

Result

```
hotel                            0
is_canceled                      0
lead_time                        0
arrival_date_year                0
arrival_date_month               0
arrival_date_week_number         0
arrival_date_day_of_month        0
stays_in_weekend_nights          0
stays_in_week_nights             0
adults                           0
children                         0
babies                           0
meal                             0
country                          0
market_segment                   0
distribution_channel             0
is_repeated_guest                0
previous_cancellations           0
previous_bookings_not_canceled   0
reserved_room_type               0
assigned_room_type               0
booking_changes                  0
deposit_type                     0
days_in_waiting_list             0
customer_type                    0
adr                              0
required_car_parking_spaces      0
total_of_special_requests        0
reservation_status               0
reservation_status_date          0
dtype: int64
```

Q3: Are there any missing values or outliers?

```
# Outlier detection for 'adr' (Average Daily Rate)
plt.figure(figsize=(8, 4))
sns.boxplot(x=df['adr'], color=sns.color_palette("Blues", 1)[0])
plt.title('Outlier Detection in Average Daily Rate (ADR)', fontsize=14)
plt.xlabel('Average Daily Rate (ADR)', fontsize=12)
plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.show()
```



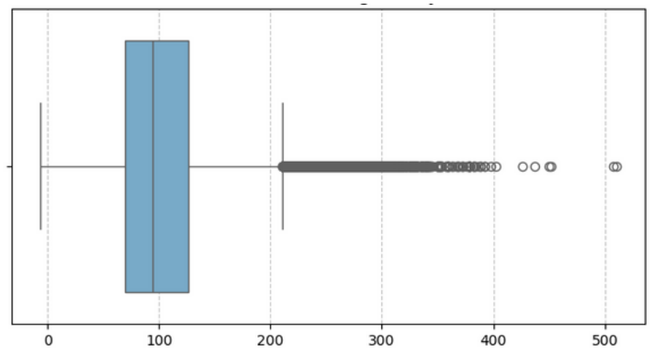
Identified Issues

From the visualization above, it can be seen that **outliers are present** at very high ADR values, far above the majority of the data distribution (**> 5000**)

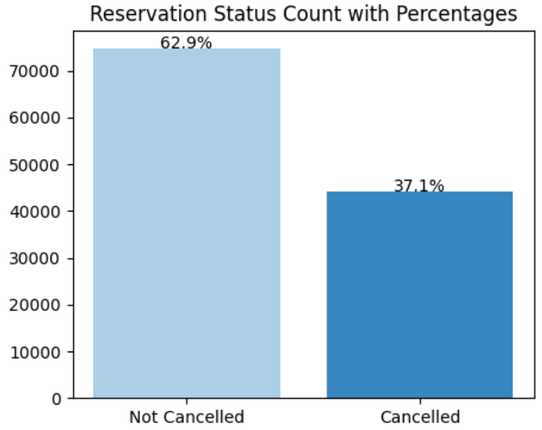
Action Taken

All rows where ADR ≥ 5000 were removed. This decision ensures that the analysis and models are more representative of typical customer behavior while reducing the impact of atypical extreme values.

Result



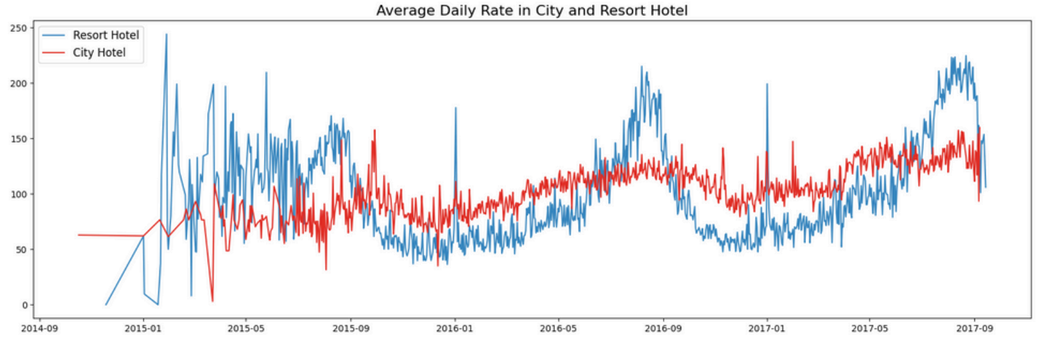
Data Visualization & Analysis Insight



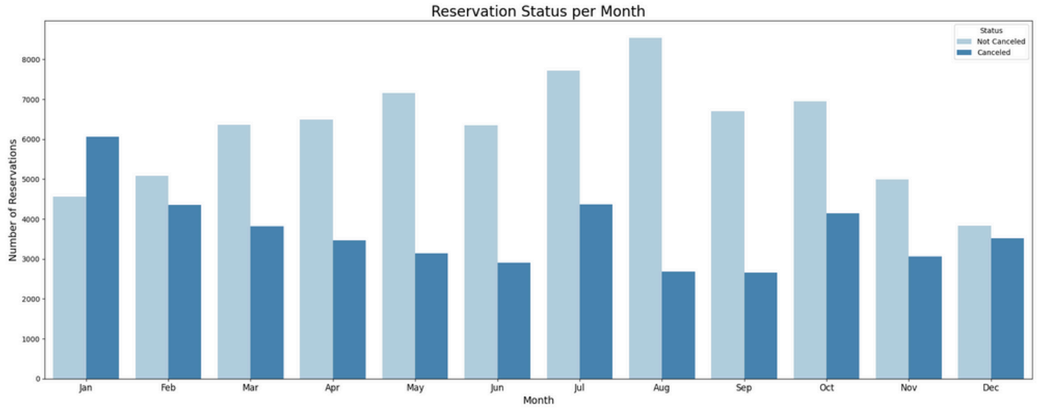
- The bar chart illustrates the distribution of reservation outcomes, distinguishing between cancellations and confirmed bookings.
- Notably, **37% of clients canceled their reservations**, representing a substantial proportion that can materially affect hotel revenue performance.

Despite this, t**he majority of reservations remain intact, highlighting both the opportunity for sustained earnings and the financial risk** posed by cancellation behavior.

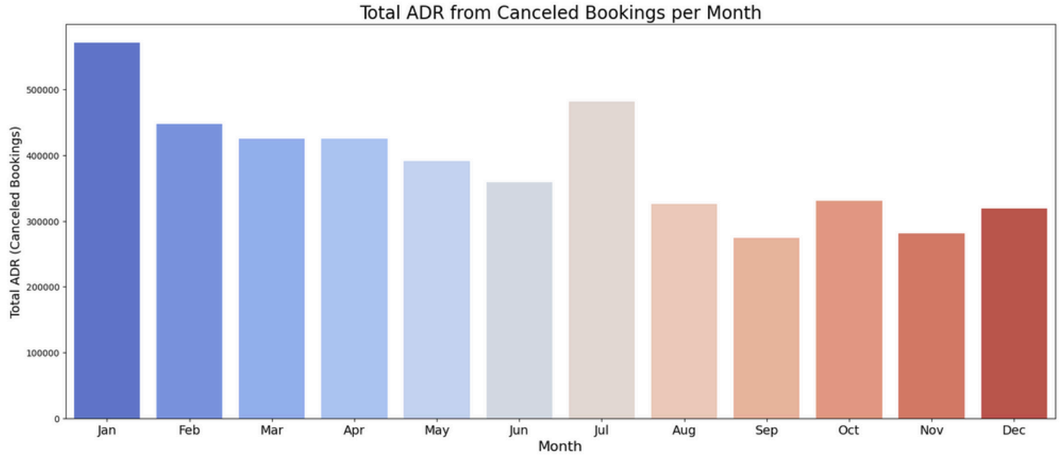
In comparison to resort hotels, **city hotels have more bookings**. It's possible that **resort hotels are more expensive** than those in cities.



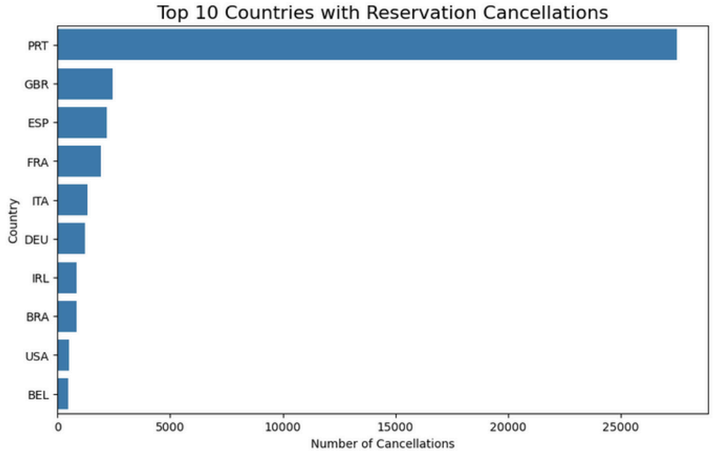
The line graph illustrates that on **some days the average daily rate of a city hotel falls below that of a resort hotel**, **while on other days the difference is even greater**. Naturally, resort hotel rates tend to **increase during weekends and holidays**.



A grouped bar graph was created to examine the months with the highest and lowest reservation levels based on reservation status. **It is evident that both confirmed and canceled reservations peak in August**, **while January records the highest number of canceled reservations**.



This bar graph indicates that **cancellations occur most** frequently **when prices are highest** and least frequently when prices are lowest. Hence, **the accommodation cost appears to be the main factor** influencing cancellations.

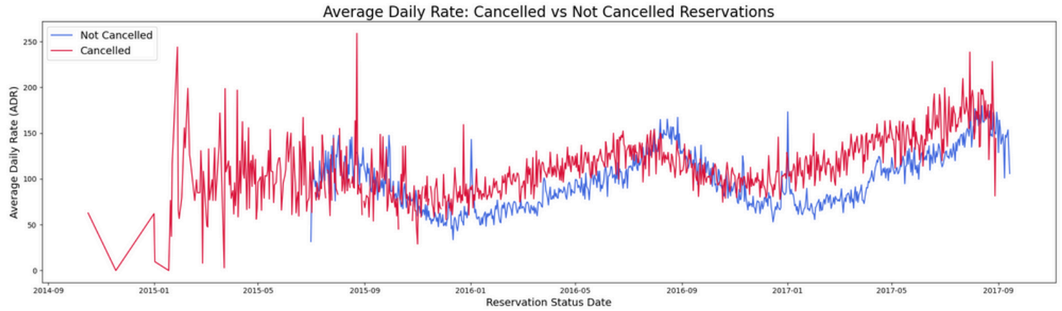


The next step is to identify which country has the highest number of reservation cancellations. The results show that **Portugal ranks first with the greatest number of cancellations**.

market_segment	proportion
Online TA	0.474377
Offline TA/TO	0.203193
Groups	0.166581
Direct	0.104696
Corporate	0.042987
Complementary	0.006173
Aviation	0.001993

It is also important to know: from **which sources do guests make their reservations** — Direct, Groups, or Online/Offline Travel Agents?

It can be seen that **around 46%** of guests book through **online travel** agencies, **27% through groups**, and only **4% make direct bookings**.



The next analysis explores the **relationship between price and cancellation trends.** As seen in the graph, **reservations are more likely to be canceled when the average daily rate is higher**, indicating that higher prices contribute to increased cancellations.

Insights and Recommendations

Cancellation Rates Rise as the Price Does

To reduce reservation cancellations, hotels can adjust their pricing strategies by offering location-based rate reductions and providing discounts.

High Cancellations in January

In January, when cancellations peak, hotels can launch marketing campaigns or special promotions to increase revenue.

Higher Cancellations in Resort Hotels

Offering reasonable discounts on room prices during weekends or holidays could help reduce cancellations.

Improving Quality to Reduce Cancellations

Enhancing hotel quality and service standards, especially in Portugal, can help reduce the cancellation rate.

THANK YOU