

ANALYSIS OF HOTEL BOOKING CANCELLATION

INTRODUCTION

In the hospitality industry, booking cancellations greatly impact demand-management decisions, hindering the production of accurate forecasts essential for effective revenue management. To address cancellations, hotels often enforce strict cancellation policies and adopt overbooking strategies, which can negatively affect revenue and social reputation, ultimately harming business performance. Overbooking can result in the hotel denying service provision, leading to a poor customer experience and damaging the hotel's reputation and immediate revenue. This can also result in future revenue loss as dissatisfied customers may not return. Conversely, rigid cancellation policies, particularly non-refundable ones, can reduce the number of bookings and revenue, as significant discounts may need to be applied to attract guests (Antonio et. al., 2017).

PROBLEM STATEMENT

The following project utilizes a hotel booking dataset containing various details surrounding two types of hotels namely City and Resort hotels.

In recent years, both the City Hotel and Resort Hotel have seen substantial increases in cancellation rates. This has led to challenges such as decreased revenue and underutilized rooms. Consequently, their primary goal is to lower cancellation rates to improve revenue generation. This report analyzes hotel booking cancellations and other factors that indirectly affect their business and annual revenue.

Assumptions

1. From 2015 to 2017, no significant events had a notable impact on the data considered.
2. The data is up-to-date and can be effectively utilized to assess potential hotel strategies.
3. There are no unexpected challenges with the hotel adopting any of the suggested approaches.
4. The recommended solutions are not already being implemented by the hotels.
5. Booking cancellations are the most significant factor influencing revenue generation.
6. Cancellations result in unoccupied rooms for the originally reserved period.
7. Clients generally cancel their hotel reservations within the same year they made them.

Research Questions

1. What factors influence hotel reservation cancellations?
2. How can we reduce hotel reservation cancellations?
3. How can hotels be assisted in making pricing and promotional decisions?

Hypothesis

1. Higher prices lead to more cancellations.
2. Longer waiting lists result in more frequent cancellations.
3. Most clients make their reservations through offline travel agents.

DATA ANALYSIS

The dataset, available on Kaggle, contains hotel demand information and originates from the article "*Hotel Booking Demand Datasets*" by Nuno Antonio, Ana Almeida, and Luis Nunes, published in the Journal "*Data in Brief - Volume 22*" in February 2019 (Elsevier Inc., 2018).

It includes records of hotel bookings made between July 1, 2015, and August 31, 2017, encompassing both arrivals and cancellations. The dataset combines data from two types of hotels: city and resort. Each dataset comprises 32 variables, with the resort hotel data containing 40,600 observations and the city hotel data containing 72,300 observations, resulting in a total of 119,300 observations. To protect privacy, all identifying information related to the hotels and customers has been removed.

The table below illustrates the information of all the thirty-two columns included in the dataset.

Index	Variable	Description
1	hotel	Type of hotel (Resort Hotel, City Hotel)
2	is_canceled	Reservation cancellation status (0 = not canceled, 1 = canceled)
3	lead_time	Number of days between booking and arrival
4	arrival_date_year	Year of arrival
5	arrival_date_month	Month of arrival
6	arrival_date_week_number	Week number of the year for arrival
7	arrival_date_day_of_month	Day of the month of arrival
8	stays_in_weekend_nights	Number of weekend nights (Saturday and Sunday) the guest stayed or booked
9	stays_in_week_nights	Number of week nights the guest stayed or booked
10	adults	Number of adults
11	children	Number of children
12	babies	Number of babies
13	meal	Type of meal booked (BB, FB, HB, SC, Undefined)
14	country	Country of origin of the guest
15	market_segment	Market segment designation
16	distribution_channel	Booking distribution channel
17	is_repeated_guest	If the guest is a repeat customer (0 = not repeated, 1 = repeated)
18	previous_cancellations	Number of previous bookings that were canceled by the customer
19	previous_bookings_not_canceled	Number of previous bookings that were not canceled by the customer
20	reserved_room_type	Type of reserved room
21	assigned_room_type	Type of assigned room
22	booking_changes	Number of changes made to the booking
23	deposit_type	Type of deposit made (No Deposit, Refundable, Non Refund)
24	agent	ID of the travel agent responsible for the booking
25	company	ID of the company responsible for the booking
26	days_in_waiting_list	Number of days the booking was in the waiting list
27	customer_type	Type of customer (Transient, Contract, Transient-Party, Group)
28	adr	Average Daily Rate
29	required_car_parking_spaces	Number of car parking spaces required
30	total_of_special_requests	Number of special requests made
31	reservation_status	Last reservation status (Check-Out, Canceled, No-Show)
32	reservation_status_date	Date of the last reservation status

The following section will skim through the four main stages of a data analytics project namely; Data Collection, Data Exploration, Data Pre-Processing, Data Analysis, Data Interpretation.

Data Collection

Import Necessary Libraries

Firstly, all the necessary libraries have been imported which will aid in data exploration, manipulation, and visualisation. These are commonly used in data analysis tasks and provide great functionality in handling and displaying data in Python. The chosen IDE is JupyterLab operated through anaconda.

```
# LIBRARIES

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

Importing the Dataset

The following code is used to import the dataset from a CSV file and store it into a data frame, named `df_hotel`, followed by displaying the first 10 rows of the dataset, in order to understand the structure and content of the dataset.

```
# Reading data in a dataframe
df_hotel = pd.read_csv('hotel_booking.csv')

# Print the first 10 rows of the dataset
df_hotel.head(10)
```

	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	...
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	..
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	..
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	..
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	..
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	..
5	Resort Hotel	0	14	2015	July	27	1	0	2	2	..
6	Resort Hotel	0	0	2015	July	27	1	0	2	2	..
7	Resort Hotel	0	9	2015	July	27	1	0	2	2	..
8	Resort Hotel	1	85	2015	July	27	1	0	3	2	..
9	Resort Hotel	1	75	2015	July	27	1	0	3	2	..

Data Exploration

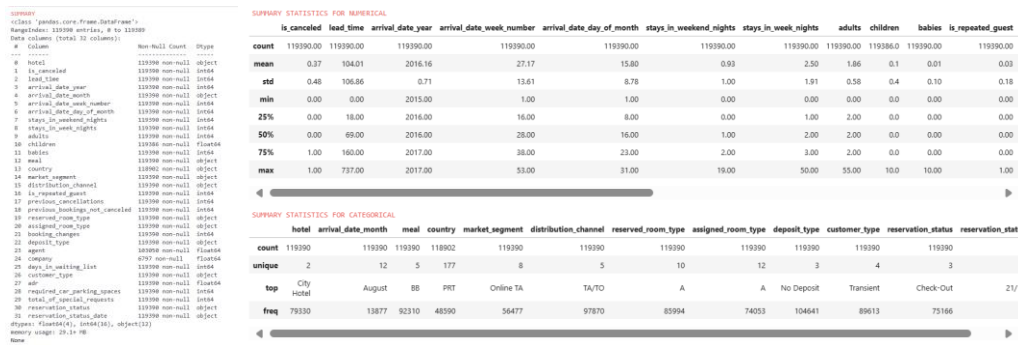
Exploratory Data Analysis involves analyzing and visualizing data to understand its key characteristics, uncover patterns, and identify relationships between variables. This method helps to explore datasets, discover patterns, find outliers, and identify variable relationships (EDA, 2024).

Descriptive and Summary Statistics

```
# Information on COLUMNS - dtype, non-null values and memory usage
print(Fore.RED + "SUMMARY" + Fore.BLACK)
display(df_hotel.info())
print()

# Display Summary Statistics for Continuous or Numeric Columns
print(Fore.RED + "SUMMARY STATISTICS FOR NUMERICAL")
display(df_hotel.describe().round(2))
print()

print(Fore.RED + "SUMMARY STATISTICS FOR CATEGORICAL")
display(df_hotel.describe(include='object'))
```



df_hotel.info()

```
Out[1]:
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column              Non-Null Count  Dtype  
---  -
 0   hotel               119390 non-null  object  
 1   is_cancelled        119390 non-null  bool     
 2   lead_time           119390 non-null  int64   
 3   is_reserved         119390 non-null  bool     
 4   arrival_date_year   119390 non-null  int64   
 5   arrival_date_month  119390 non-null  int64   
 6   arrival_date_week_number  119390 non-null  int64   
 7   arrival_date_day_of_month  119390 non-null  int64   
 8   stays_in_weekend_nights  119390 non-null  float64  
 9   stays_in_week_nights  119390 non-null  float64  
10   children            119390 non-null  float64  
11   babies             119390 non-null  float64  
12   is_repeated_guest   119390 non-null  bool     
13   meal               118802 non-null  object  
14   country            119390 non-null  object  
15   market_segment     119390 non-null  object  
16   distribution_channel  119390 non-null  object  
17   is_repeated_guest   119390 non-null  bool     
18   previous_bookings_not_cancelled  119390 non-null  bool     
19   reserved_room_type  119390 non-null  object  
20   assigned_room_type  119390 non-null  object  
21   booking_changes     119390 non-null  object  
22   deposit_type        119390 non-null  object  
23   agent              80809 non-null  object  
24   company            6797 non-null   float64  
25   days_in_waiting_list  119390 non-null  object  
26   customer_type       119390 non-null  object  
27   adp                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: 29.14 MB
```

SUMMARY STATISTICS FOR NUMERICAL

	is_cancelled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest
count	119390.00	119390.00	119390.00	119390.00	119390.00	119390.00	119390.00	119390.00	119390.00	119390.00	119390.00
mean	0.37	104.01	2016.16	27.17	15.80	0.93	2.50	1.86	0.1	0.01	0.03
std	0.48	106.86	0.71	13.61	8.78	1.00	1.91	0.58	0.4	0.10	0.18
min	0.00	0.00	2015.00	1.00	1.00	0.00	0.00	0.00	0.0	0.00	0.00
25%	0.00	18.00	2016.00	16.00	8.00	0.00	1.00	2.00	0.0	0.00	0.00
50%	0.00	69.00	2016.00	28.00	16.00	1.00	2.00	2.00	0.0	0.00	0.00
75%	1.00	160.00	2017.00	38.00	23.00	2.00	3.00	2.00	0.0	0.00	0.00
max	1.00	737.00	2017.00	53.00	31.00	19.00	50.00	55.00	10.0	10.00	1.00

SUMMARY STATISTICS FOR CATEGORICAL

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type	deposit_type	customer_type	reservation_status	reservation_status_date
count	119390	119390	119390	118902	119390	119390	119390	119390	119390	119390	119390	119390
unique	2	12	5	177	8	5	10	12	3	4	3	21
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	A	No Deposit	Transient	Check-Out	21/
freq	79330	13877	92310	48590	56477	97870	85994	74053	104641	89613	75166	

As it can be seen in the above figures, the code above provides a summary or information about the dataset. The code `df_hotel.info()` is used to print the information on the columns, data types, and memory usage. It can be noted that there are exactly 119,389 entries with 32 attributes as explained in the above sections. This is further followed by a statistical summary of numeric and categorical variables.

It can be noted that the dataset consists of 20 numerical variables and 12 categorical variables. The variable `reservation_status_date` is of object type when in fact should be of date type. Furthermore, variable *average daily rate (adr)* contains negative values and the maximum value shown is considerable high; implying the data contains certain outliers which need to be further explored.

Displaying Missing and Unique Values

```
# Display the Total Number of rows and columns
print(Fore.RED + "TOTAL NUMBER OF ROWS AND COLUMNS")
display(df_hotel.shape)
print()

# Display the number of unique values for each column
print(Fore.RED + "UNIQUE VALUES")
display(df_hotel.nunique())
print()

# Display the Missing Values
print(Fore.RED + "MISSING VALUES")
missing_data = pd.DataFrame({'Count': df_hotel.isnull().sum().sort_values(ascending=False), 'Percentage': ((df_hotel.isnull().sum() / len(df_hotel)) * 100).sort_values(ascending=False)})
display(missing_data)
print()
```

TOTAL NUMBER OF ROWS AND COLUMNS (113990, 32)		MISSING VALUES		Count Percentage	
UNIQUE VALUES					
hotel	2	company	112593	94.306893	
is_canceled	2	agent	10540	12.686238	
lead_time	479	country	458	0.407874	
arrival_date_year	3	children	4	0.003350	
arrival_date_month	12	reserved_room_type	0	0.000000	
arrival_date_week_number	53	assigned_room_type	0	0.000000	
arrival_date_day_of_month	31	booking_changes	0	0.000000	
stays_in_weekend_nights	17	deposit_type	0	0.000000	
stays_in_week_nights	35	hotel	0	0.000000	
adults	14	previous_cancellations	0	0.000000	
children	5	days_in_waiting_list	0	0.000000	
babies	5	customer_type	0	0.000000	
meal	5	adr	0	0.000000	
country	177	required_car_parking_spaces	0	0.000000	
market_segment	10	total_of_special_requests	0	0.000000	
distribution_channel	2	reservation_status	0	0.000000	
is_repeated_guest	15	previous_bookings_not_canceled	0	0.000000	
previous_cancellations	73	is_repeated_guest	0	0.000000	
previous_bookings_not_canceled	10	is_canceled	0	0.000000	
reserved_room_type	12	distribution_channel	0	0.000000	
assigned_room_type	18	reserved_room_type	0	0.000000	
booking_changes	21	market_segment	0	0.000000	
deposit_type	3	meal	0	0.000000	
agent	352	babies	0	0.000000	
company	128	adults	0	0.000000	
days_in_waiting_list	4	stays_in_week_nights	0	0.000000	
customer_type	8879	stays_in_weekend_nights	0	0.000000	
adr	5	arrival_date_day_of_month	0	0.000000	
required_car_parking_spaces	6	arrival_date_week_number	0	0.000000	
total_of_special_requests	5	arrival_date_year	0	0.000000	
reservation_status	3	lead_time	0	0.000000	
reservation_status_date	926	reservation_status_date	0	0.000000	
dtype: int64					

Data Pre-Processing

Convert Data Type

6

#	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	datetime64[ns]

As discussed above, *reservation_status_date* attribute should not be of object type. Hence, it was converted to date type, and the resulting dataset has been displayed to confirm the changes.

Handling Missing Values

```
# Remove columns 'Company' and 'Agent' as they contain high number of null values, which are difficult to deal with
df_hotel.drop(['company', 'agent'], axis = 1, inplace = True)

# For the columns 'Country' and 'Children', null value rows will just be eliminated
df_hotel.dropna(inplace = True)

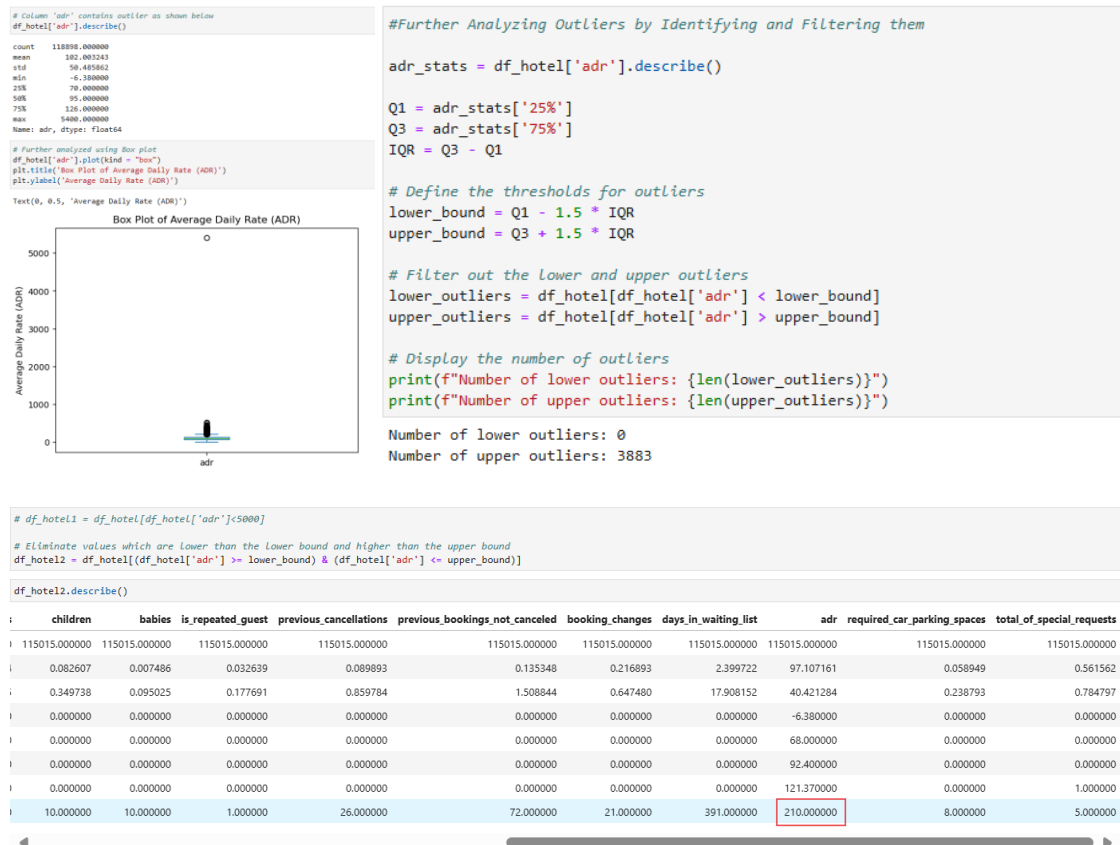
# Confirm changes
df_hotel.isnull().sum()
```

```
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
```

As mentioned previously, the columns *Company* and *Agent* have a very high number of null values which can be difficult to deal with. As these columns are not necessarily required for the purpose of this project, they will be dropped from the dataset.

The remaining two columns, namely *Country* and *Children* have a lower number of null values, which will not have a big impact on the analysis. Hence, the rows containing null values will just be dropped from the dataset.

Outlier Detection

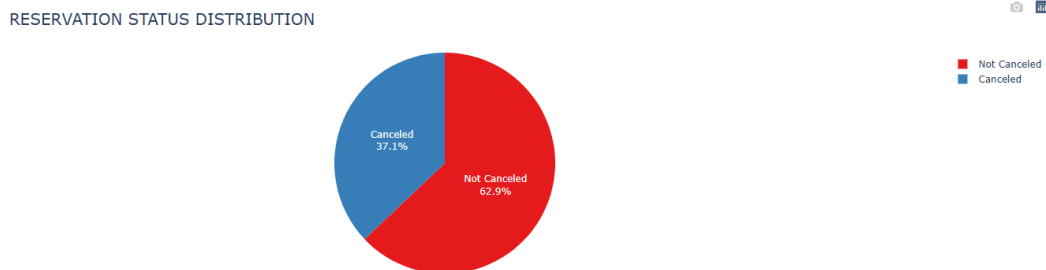


The column *adr* contains outlier values, indicated by a skewed box plot distribution. To address this, the Interquartile Range (IQR) between the first quartile (Q1) and third quartile (Q3), was calculated to identify the bounds. The results indicated 3,883 upper outliers and no lower outliers. Values outside these bounds were eliminated, thus ensuring a dataset free of outliers, ready for analysis.

Data Analysis and Visualization

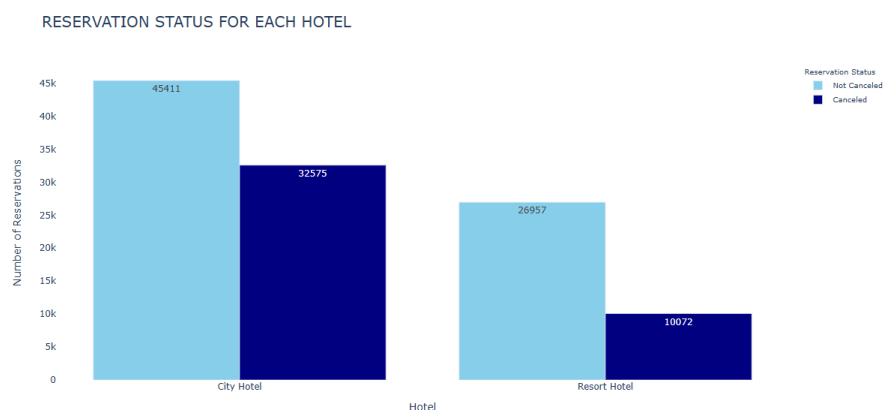
This section includes various visualizations designed to further analyze the reasons for hotel cancellations and to validate the hypotheses and assumptions.

Count of Reservation Status



The bar graph illustrates the percentage of reservations that were canceled versus those that were not. It's evident that a substantial number of reservations remain active. However, 37% of clients canceled their reservations, significantly impacting the hotels' revenue.

Count of Reservation Status for Each Hotel



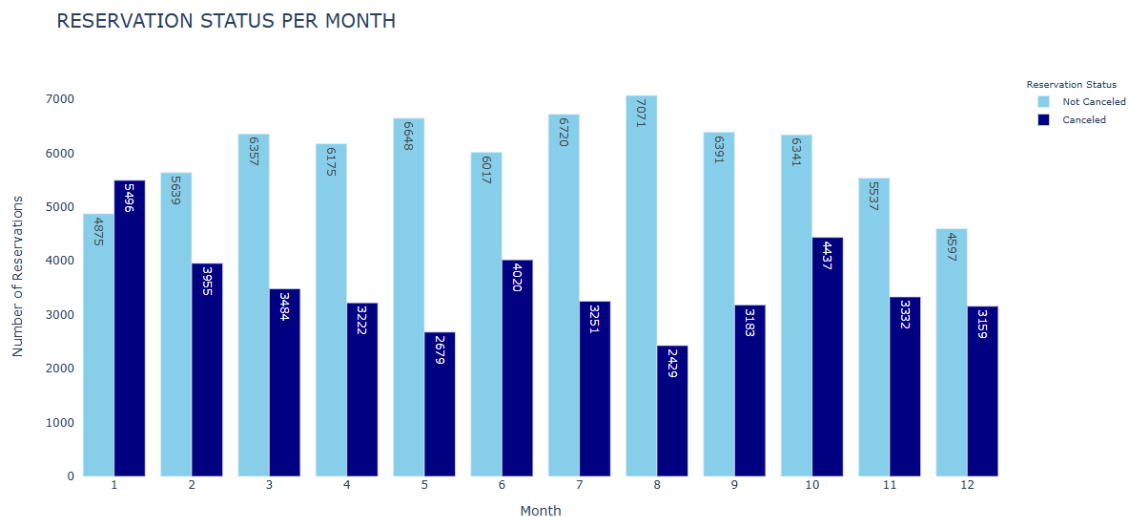
City hotels have more bookings compared to resort hotels, possibly because resort hotels are more expensive. However, cancellations in City hotels (Cancelled - 42%, Not Cancelled – 58%) are higher than those of Resort Hotels (Cancelled – 28%, Not Cancelled – 72%).

Calculate the Average Daily Rate for Each Hotel



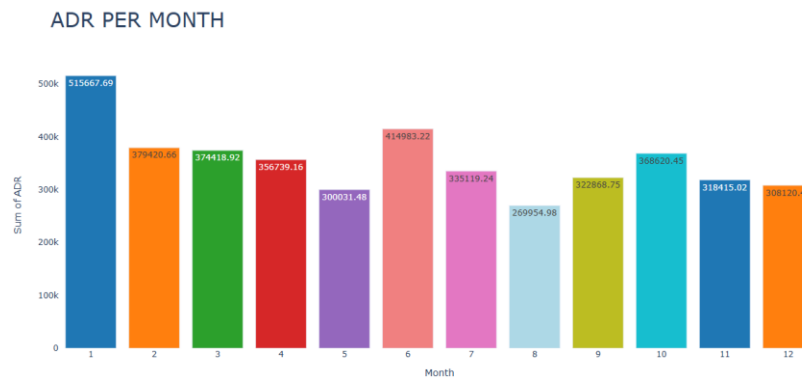
The line graph indicates that on some days, the average daily rate for city hotels is lower than that of resort hotels, and on other days, it is even lower. It is clear that weekends and holidays are likely to cause an increase in resort hotel rates.

Analysis of Reservation Status on a Monthly Basis



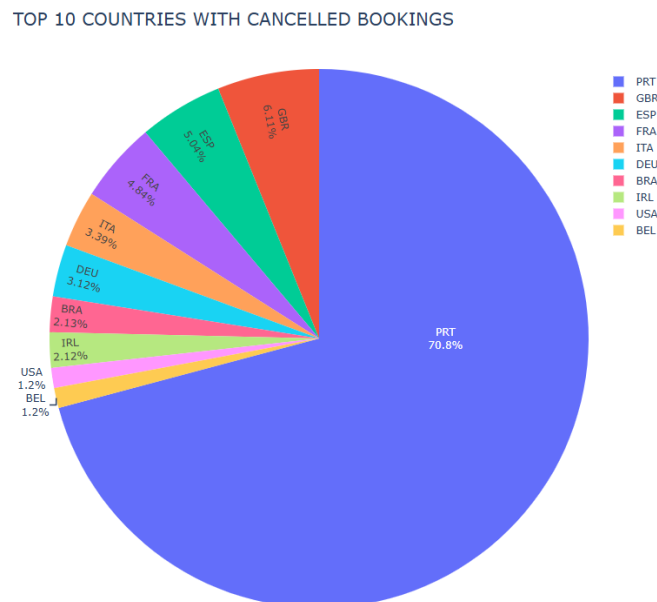
A grouped bar graph was created to analyze the months with the highest and lowest reservation levels based on reservation status. It shows that both confirmed and canceled reservations peak in August, while January has the most canceled reservations.

Analysis of Average Daily Rate of Cancelled Bookings on a Monthly Basis



As shown, ADR in August was the lowest, whilst ADR in January was the highest. The above bar graph indicates that cancellations are most frequent when prices are highest and least frequent when prices are lowest. Thus, the cost of accommodation is the primary factor driving cancellations.

Top 10 Countries with Cancelled Bookings



The top country is Portugal with the highest number of cancellations, amounting to 70%. This is followed by the UK (GBR), Spain (ESP), and France (FRA).

Analysis of Market Segment

```
df_hotel2['market_segment'].value_counts()

Online TA      53638
Offline TA/TO  24052
Groups         19709
Direct         11541
Corporate       5104
Complementary   734
Aviation        237
Name: market_segment, dtype: int64

df_hotel2['market_segment'].value_counts(normalize = True)

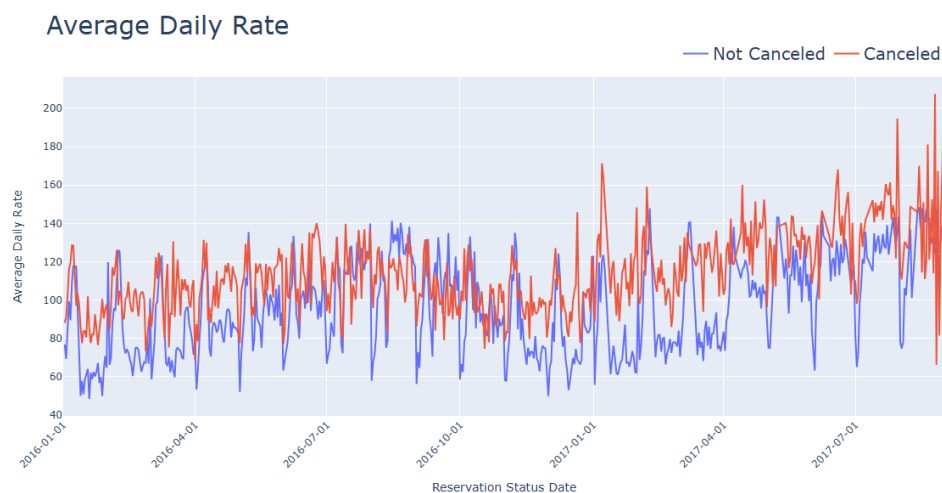
Online TA      0.466357
Offline TA/TO  0.209121
Groups         0.171360
Direct         0.100343
Corporate       0.044377
Complementary   0.006382
Aviation        0.002061
Name: market_segment, dtype: float64

cancelled_data['market_segment'].value_counts(normalize = True)

Online TA      0.456679
Groups         0.282763
Offline TA/TO  0.193636
Direct         0.040706
Corporate       0.022886
Complementary   0.002110
Aviation        0.001219
Name: market_segment, dtype: float64
```

There are four ways for hotel guests to book their reservations: Direct, Groups, Online, or Offline Travel Agents. Approximately 47% of clients book through online travel agencies, while 27% come from groups. Only 4% of clients make direct bookings by visiting the hotels. As shown above, most of the cancellations are from those booked via online agencies, amounting to 46%.

Analysis of ADR for Cancelled and Non-Cancelled Bookings



The graph also shows that reservations are more likely to be canceled when the average daily rate is higher, reinforcing the analysis that higher prices lead to higher cancellation rates.

Data Interpretation

1. Adjust Pricing Strategies:

- As prices increase, so do cancellation rates. To reduce cancellations, hotels should implement dynamic pricing strategies, potentially lowering rates for specific locations and times. Offering targeted discounts and special promotions to consumers during off-peak periods can also help retain bookings.

2. Flexible Booking Options:

- Since the cancellation rate is higher for resort hotels compared to city hotels, resorts should consider offering discounts on room prices during weekends and holidays. Additionally, providing flexible cancellation policies and allowing guests to change booking dates without penalties can reduce cancellations.

3. Targeted Marketing Campaigns:

- Given that cancellations peak in January, hotels could launch seasonal marketing campaigns to boost revenue during this month. Focused promotions on special packages, events, or discounted rates can attract more guests and decrease cancellations.
- Launch geographical campaigns aimed at countries with high cancellation rates, such as Portugal, the UK, Spain, and France. Tailor the marketing messages to address the specific needs and preferences of guests from these regions.

4. Enhance Direct Booking Incentives:

- Encourage direct bookings by offering exclusive discounts, loyalty points, or additional perks for guests who book through the hotel's website or in person. Enhancing the online booking experience can also make direct bookings more appealing and reduce cancellations.

5. Improve Quality of Services and Facilities:

- Improving the quality of hotels and services, especially in Portugal, can help reduce the cancellation rate. Investing in upgrading hotel facilities and

enhancing service quality can increase guest satisfaction and loyalty, making them less likely to cancel.

- Actively seek guest feedback and address any issues promptly to maintain a positive reputation and attract more bookings.

6. Engage with Customers:

- Engage with guests through personalized communication, offering tailored suggestions and keeping them well-informed about their stay. Building stronger relationships with guests can reduce the likelihood of cancellations.
- Utilize guest feedback and reviews to continuously improve services and address potential concerns that may lead to cancellations.

REFERENCES

<https://www.sciencedirect.com/science/article/pii/S2352340918315191#section-cited-by>

<https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data>

<https://www.datacamp.com/blog/what-is-data-analysis-expert-guide>

<https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/>

Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model -

<https://ieeexplore.ieee.org/document/8260781>