Capstone Project

EDA on Hotel Booking Analysis

By Arjun Domle

Data Science Trainee
AlmaBetter

Points to Discuss

- Data Pipeline
- Problem Statement
- Technologies and Libraries used
- Description of individual variables
- About Dataset
- Data Wrangling
- EDA and Chart Visualization
- Conclusion

Data Pipeline

Data Reading

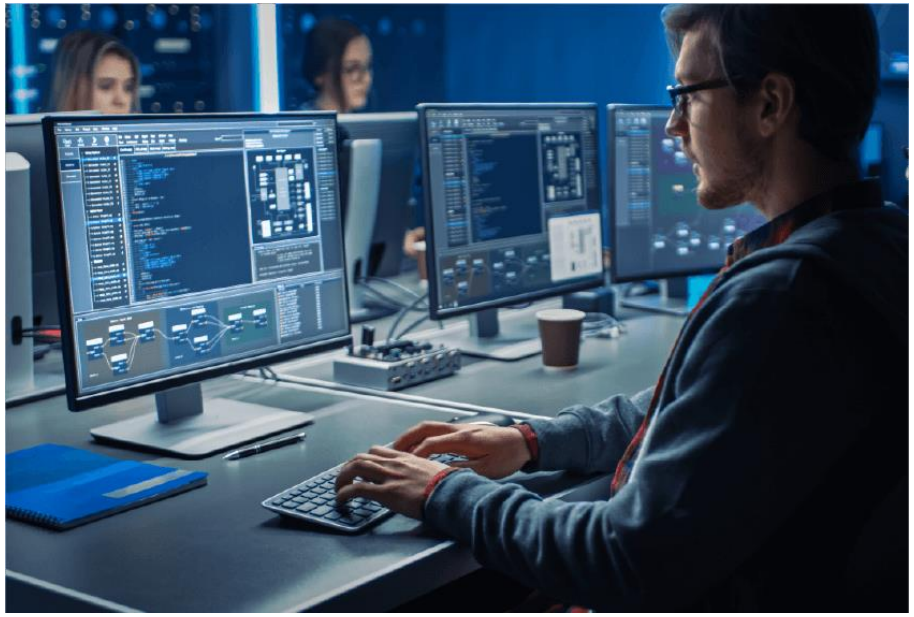
Data Preparing

Data Cleaning

Exploring Data Analysis

Data Visualization

Conclusion



Data Pipeline :

A data pipeline is a set of tools and processes used to automate the movement and transformation of data between a source system and a target repository.

Data Reading :

read and write tabular data using pandas functions.

Data preparation :

Data, when initially obtained, must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format (known as structured data) for further analysis, often through the use of spreadsheet or statistical software.

Cleaning the Data :

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

Exploratory data analysis :

Once the datasets are cleaned, they can then be analyzed. Analysts may apply a variety of techniques, referred to as exploratory data analysis, to begin understanding the messages contained within the obtained data.

Data visualization :

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Problem Statement

- Have you ever been curious about when is the best time to reserve a hotel room or the ideal duration of stay to get the best daily rate?
- Or, have you ever wanted to predict if a hotel is likely to receive a higher number of special requests?
- This hotel booking dataset can assist you in answering these questions.
- The dataset includes booking details for both a city hotel and a resort hotel, such as booking date, stay duration, number of adults, children, babies and parking spaces available. All personal identification information has been removed from the data.
- By analyzing and exploring the data, you can uncover important factors that influence bookings.

Technologies and Libraries Used

- ✓**Python : the programming language**
- ✓**NumPy : for computational Operations**
- ✓**Pandas : for Data manipulation and aggregation**
- ✓**Matplotlib : for plotting charts and visualizations**
- ✓**Seaborn : for high API visualizations**
- ✓**Google Collab : for writing and performing Coding and EDA**

❑**Reference :**

- https://www.w3schools.com/
- https://stackoverflow.com/
- https://www.almabetter.com/courses/full-stack-data-science

Description of individual Variable

1. hotel : Name of the hotel (Resort Hotel or City Hotel)

2. is_canceled : If the booking was canceled (1) or not (0)

3. lead_time: Number of days before the actual arrival of the guests

4. arrival_date_year : Year of arrival date

5. arrival_date_month : Month of month arrival date

6. arrival_date_week_number : Week number of year for arrival date

7. arrival_date_day_of_month : Day of arrival date

8. stays_in_weekend_nights : Number of weekend nights (Saturday or Sunday) spent at t he hotel by the guests.

9. stays_in_week_nights : Number of weeknights (Monday to Friday) spent at the hotel by the guests.
10. adults : Number of adults among guests
11. children : Number of children among guest
12. babies : Number of babies among guest
13. meal : Type of meal booked
14. country : Country of guests
15. market_segment : Designation of market segment
16. distribution_channel : Name of booking distribution channel
17. is_repeated_guest : If the booking was from a repeated guest (1) or not (0)
18. previous_cancellations : Number of previous bookings that were cancelled by the customer prior to the current booking

19. previous_bookings_not_canceled : Number of previous bookings not cancelled by the customer prior to the current booking
20. reserved_room_type : Code of room type reserved
21. assigned_room_type : Code of room type assigned
22. booking_changes : Number of changes/amendments made to the booking
23. deposit_type : Type of the deposit made by the guest
24. agent : ID of travel agent who made the booking
25. company : ID of the company that made the booking
26. days_in_waiting_list : Number of days the booking was in the waiting list

27. customer_type : Type of customer, assuming one of four categories

28. adr : Average Daily Rate, as defined by dividing the sum of all lodging transactions by the total number of staying nights

29. required_car_parking_spaces : Number of car parking spaces required by the cust omer

30. total_of_special_requests : Number of special requests made by the customer

31. reservation_status : Reservation status (Canceled, Check-Out or No-Show)

32. reservation_status_date : Date at which the last reservation status was updated

Answer Here

About Dataset

- ✓This data set contains a single file which compares various booking information between two hotels: a city hotel and a resort hotel.
- ✓Includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.
- ✓ The dataset contains a total of 119390 rows and 32 columns.
- ✓Dataset Contains duplicated items i.e. 31944 which is removed later.
- ✓In this dataset we find data types of every columns i.e. (Int, float ,string) and observe that some columns data types is not accurate and remove later.
- ✓We find unique value of every columns it means what actual values in every columns.

```python
# Import Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
from datetime import datetime
import seaborn as sns
import ast
```

- **Importing Libraries**

- **Loading Dataset**

Dataset Loading

```
[2] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```python
# Load Dataset
database = "/content/drive/MyDrive/Almabetter1/Module 1/Week 4/Capstone Project/Hotel Bookings.csv"
hotel_booking_df =pd.read_csv(database)
```

```
# Dataset First Look
hotel_booking_df
```

	hotel	is_canceled	lead_time	arrival_date_year	a
0	Resort Hotel	0	342	2015	
1	Resort Hotel	0	737	2015	
2	Resort Hotel	0	7	2015	
3	Resort Hotel	0	13	2015	
4	Resort Hotel	0	14	2015	
...	...	...	...	...	
119385	City Hotel	0	23	2017	
119386	City Hotel	0	102	2017	
119387	City Hotel	0	34	2017	
119388	City Hotel	0	109	2017	
119389	City Hotel	0	205	2017	

119390 rows × 32 columns

Dataset First View

Rows and Columns

```
[5] # Dataset Rows & Columns count
    print(f'Total rows are : {hotel_booking_df.shape[0]}')

    print(f'Total columns are : {hotel_booking_df.shape[1]}')
```

```
Total rows are : 119390
Total columns are : 32
```

• Dataset information

```
[6]  # Dataset Info
     hotel_booking_df.info()
```

```
<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
```

• Dataset Columns

```
# Dataset Columns
df_column = hotel_booking_df.columns
df_column
```

```
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'],
      dtype='object')
```

```
# Dataset Describe
hotel_booking_df.describe()
```

	is_canceled	lead_time	arrival_date_year	arri
count	87396.000000	87396.000000	87396.000000	
mean	0.274898	79.891368	2016.210296	
std	0.446466	86.052325	0.686102	
min	0.000000	0.000000	2015.000000	
25%	0.000000	11.000000	2016.000000	
50%	0.000000	49.000000	2016.000000	
75%	1.000000	125.000000	2017.000000	
max	1.000000	737.000000	2017.000000	

Describing Dataset

```
# Check Unique Values for each variable.
print(hotel_booking_df.apply(lambda col: col.unique())) # We have describes unique
```

```
hotel                                [Resort Hotel, City Hotel]
is_canceled                                             [0, 1]
lead_time                [342, 737, 7, 13, 14, 0, 9, 85, 75, 23, 35, 68...
arrival_date_year                                [2015, 2016, 2017]
arrival_date_month       [July, August, September, October, November, D...
arrival_date_week_number [27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 3...
arrival_date_day_of_month[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...
stays_in_weekend_nights  [0, 1, 2, 4, 3, 6, 13, 8, 5, 7, 12, 9, 16, 18,...
stays_in_week_nights     [0, 1, 2, 3, 4, 5, 10, 11, 8, 6, 7, 15, 9, 12,...
adults                   [2, 1, 3, 4, 40, 26, 50, 27, 55, 0, 20, 6, 5, 10]
children                            [0.0, 1.0, 2.0, 10.0, 3.0, nan]
babies                                          [0, 1, 2, 10, 9]
meal                                [BB, FB, HB, SC, Undefined]
country                  [PRT, GBR, USA, ESP, IRL, FRA, nan, ROU, NOR, ...
market_segment           [Direct, Corporate, Online TA, Offline TA/TO, ...
distribution_channel            [Direct, Corporate, TA/TO, Undefined, GDS]
```

Unique values

```
# Visualizing the missing values
miss_values =hotel_booking_df.isnull().sum().sort_values(ascending=False)
miss_values # We have check the count of null value in individual columns
```

```
company                         82137
agent                           12193
country                           452
children                            4
reserved_room_type                  0
assigned_room_type                  0
booking_changes                     0
deposit_type                        0
hotel                               0
previous_cancellations              0
days_in_waiting_list                0
customer_type                       0
```

Missing Values/null values

Data Wrangling

❑Data Manipulation

----Addition of columns----

We have seen that there are few columns required in Data to analysis purpose which can be evaluated from the given columns.

a) **Total Guests:** This columns will help us to evaluate the volumes of total guest and revenue as well. We get this value by adding total no. of Adults, Children & babies

b) **Revenue:** We find revenue by multiplying adr & total guest. This column will use to analyze the profit and growth of each hotel.

----Delete of columns----

a)**company:** As we have seen that this columns has almost Null data. so we have delete this column as this will not make any impact in the analysis.

----Replace of Values in columns----

a)**is_canceled, is_not_canceled & is_repeated_guest:** We have seen, that these columns contains only 0,1 as values which represent the status of booing cancellation. We replace these values (0,1) from 'Canceled' & 'Not canceled. In the same way for column 'is_repeated_guest', we replace 0,1 from 'Repeated' & 'Not repeated'. Now this values will help to make better understanding while visualization.

----Changes in data type of values in columns----

a)**Agent & Children:** We checked that these columns contains float values, which is not making any sense in data as this values represent the count of guest & ID of agent. So we have changed the data type of these columns from 'float' to 'Integer'.

----Removed is_null values & duplicate entries----

a)Before visualize any data from the data set we have to do data wrangling. For that, we have check the null value in all the columns. After checking, when we are getting a column which has more number of null values, dropped that column by using the 'drop' method. In this way, we have dropped the 'company' column. When we find minimal number of null values, filling these null values with necessary values as per requirement by using .fillna().

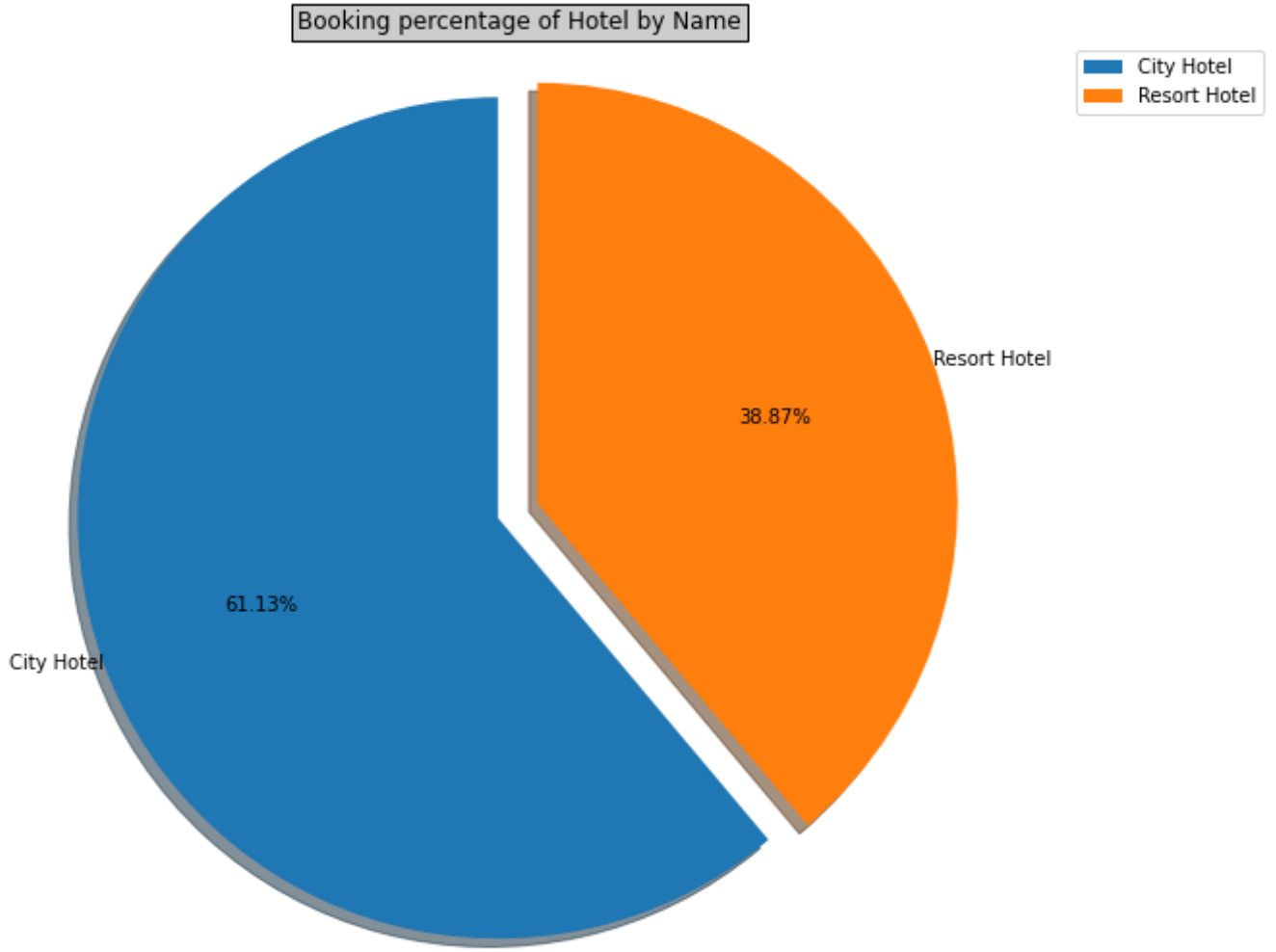
b) In the same, we have checked if there is any duplicity in data & we found that there are few rows which have duplicate data. So we have removed those row from data set by using .drop_duplicates() method.

In this way, we have removed unnecessary data & make our data clean and ready to analyze.

EDA and Chart Visualization

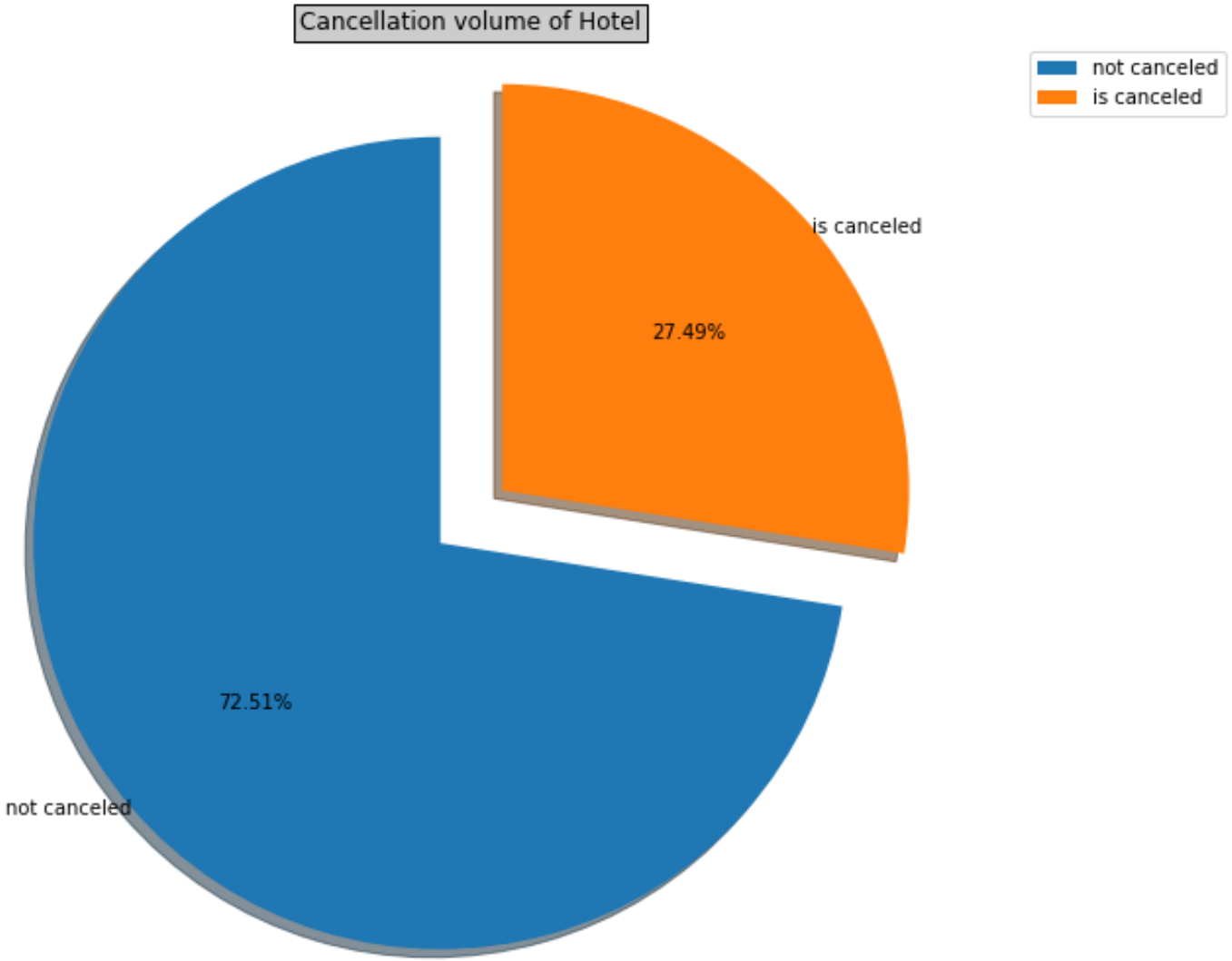
❑ hotel with the most booking percentage.

- ✓ As we see in chart, we can conclude that City Hotel (61.13%) is dominating over Resort Hotel (38.87%)



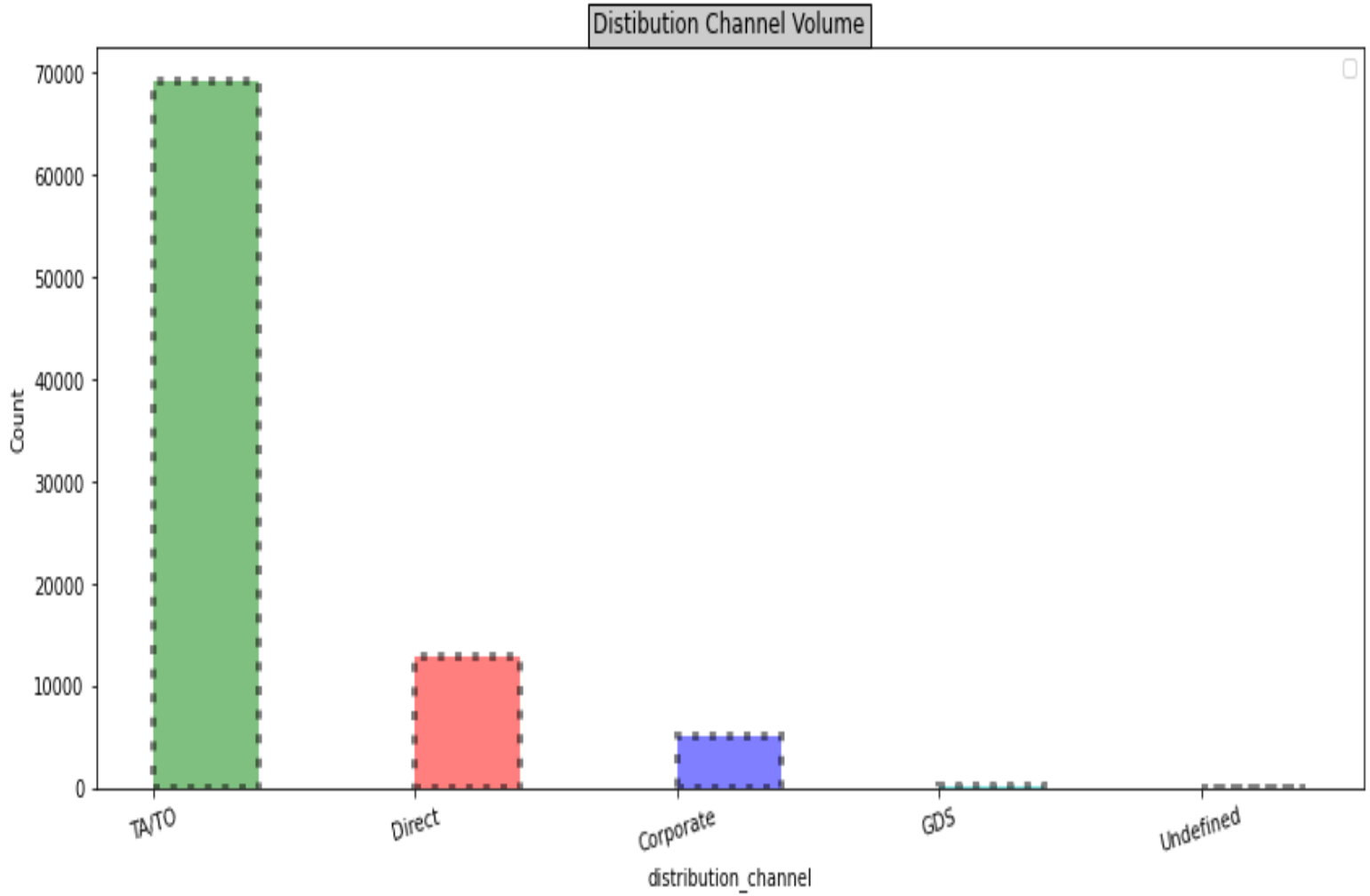
❑ the cancellation rate of the hotels booking.

- ✓ **Here, we found that overall more than 25% of booking got cancelled. Upto(27.49%).**



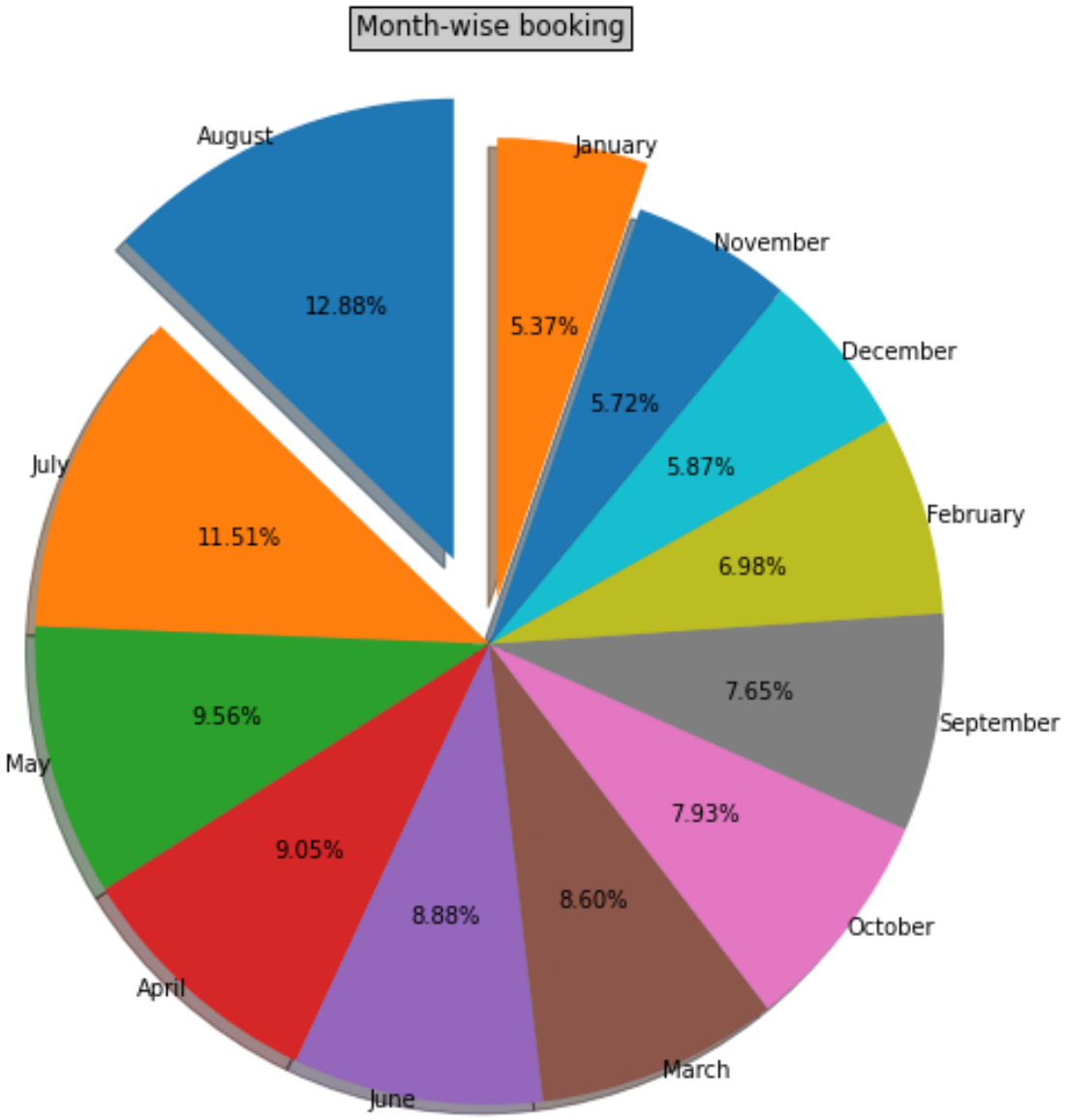
❑ distribution_channel With maximum volume of booking.

- ✓ As clearly seen TA/TO(Tour of Agent & Tour of operator) is highest, recommending to continue booking through TA/TO.



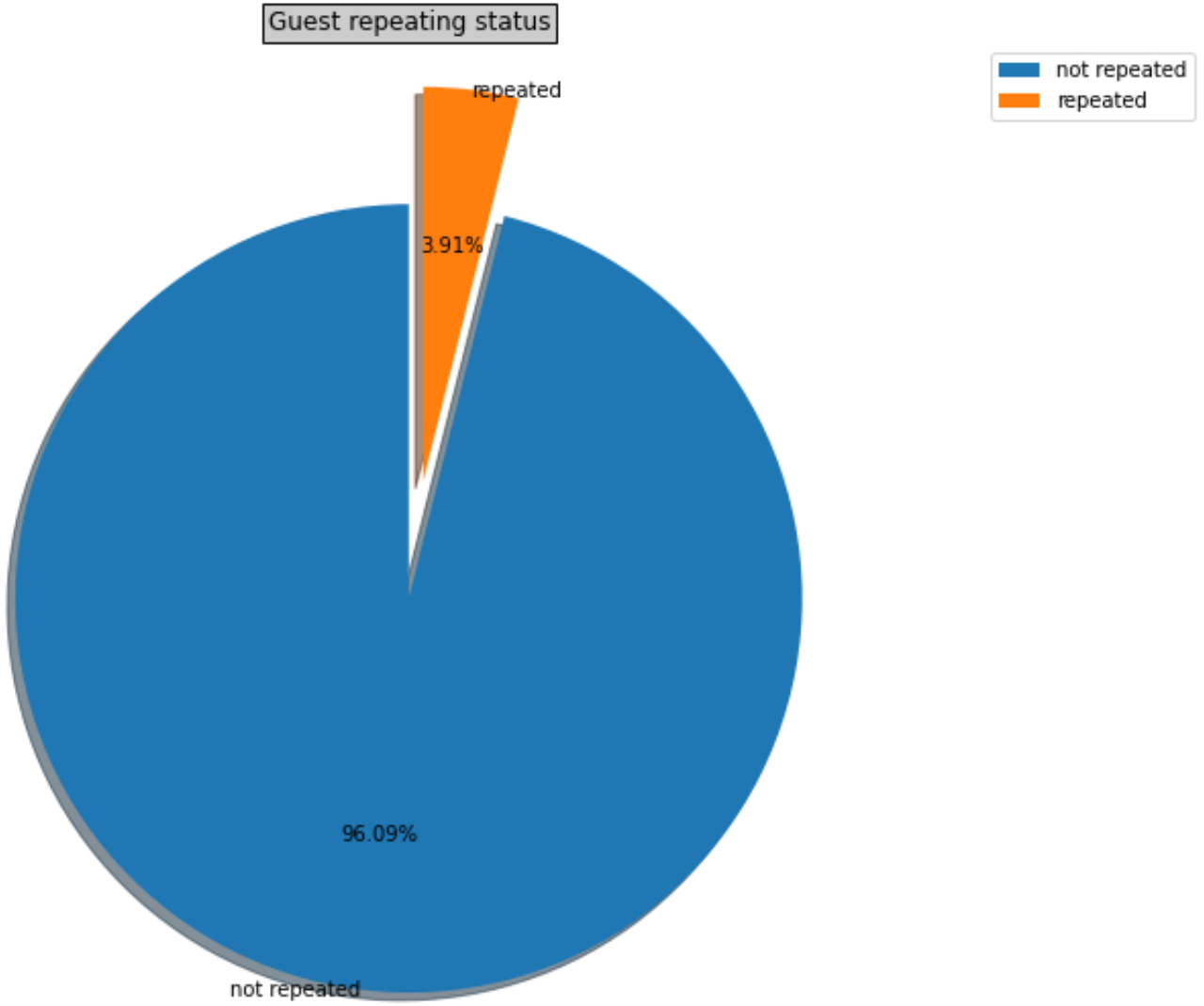
❑ percentage share of booking in each month, on overall level

- ✓ The above percentage shows month May, July and Aug are the highest booking months.

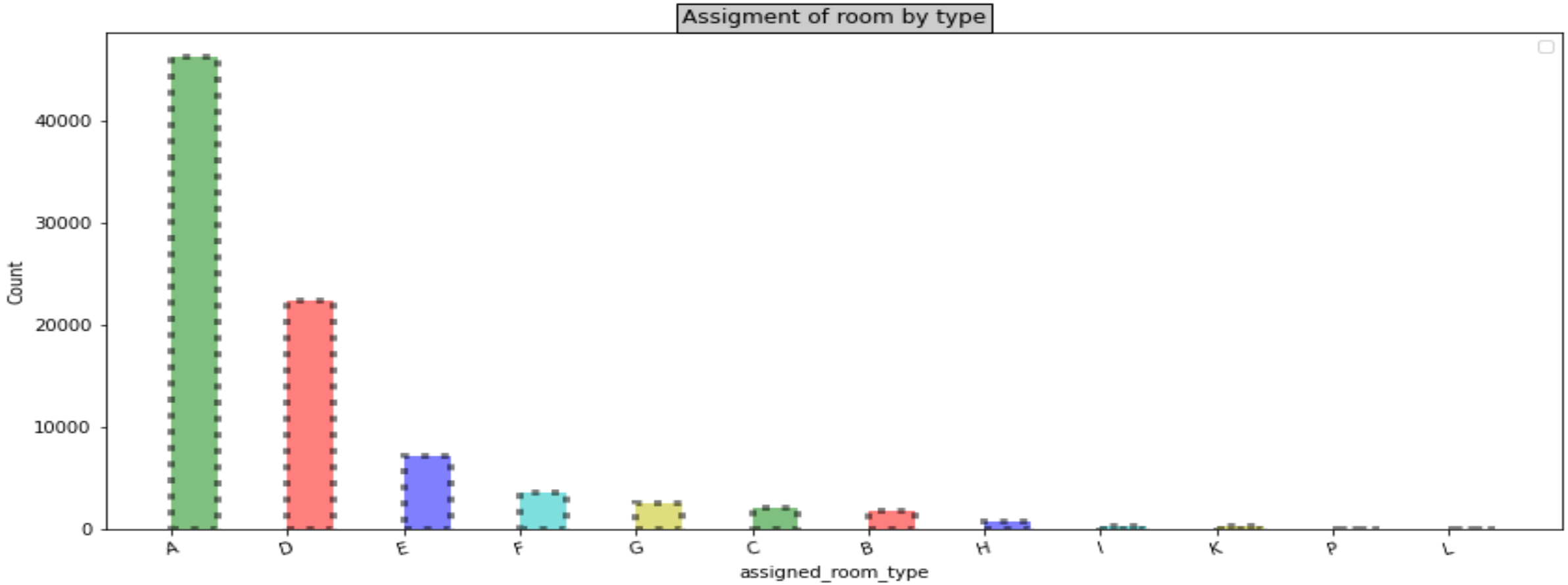


❑ percentage share of repeated & non-repeated guests.

- ✓ we can see that the number of repeated guests is very less as compared to overall guests.
- ✓ This can be very worst news in-respective of profit for an Organization, as the number has surpassed 95%.
- ✓ Team should really focus on strategy to make customer visit again.



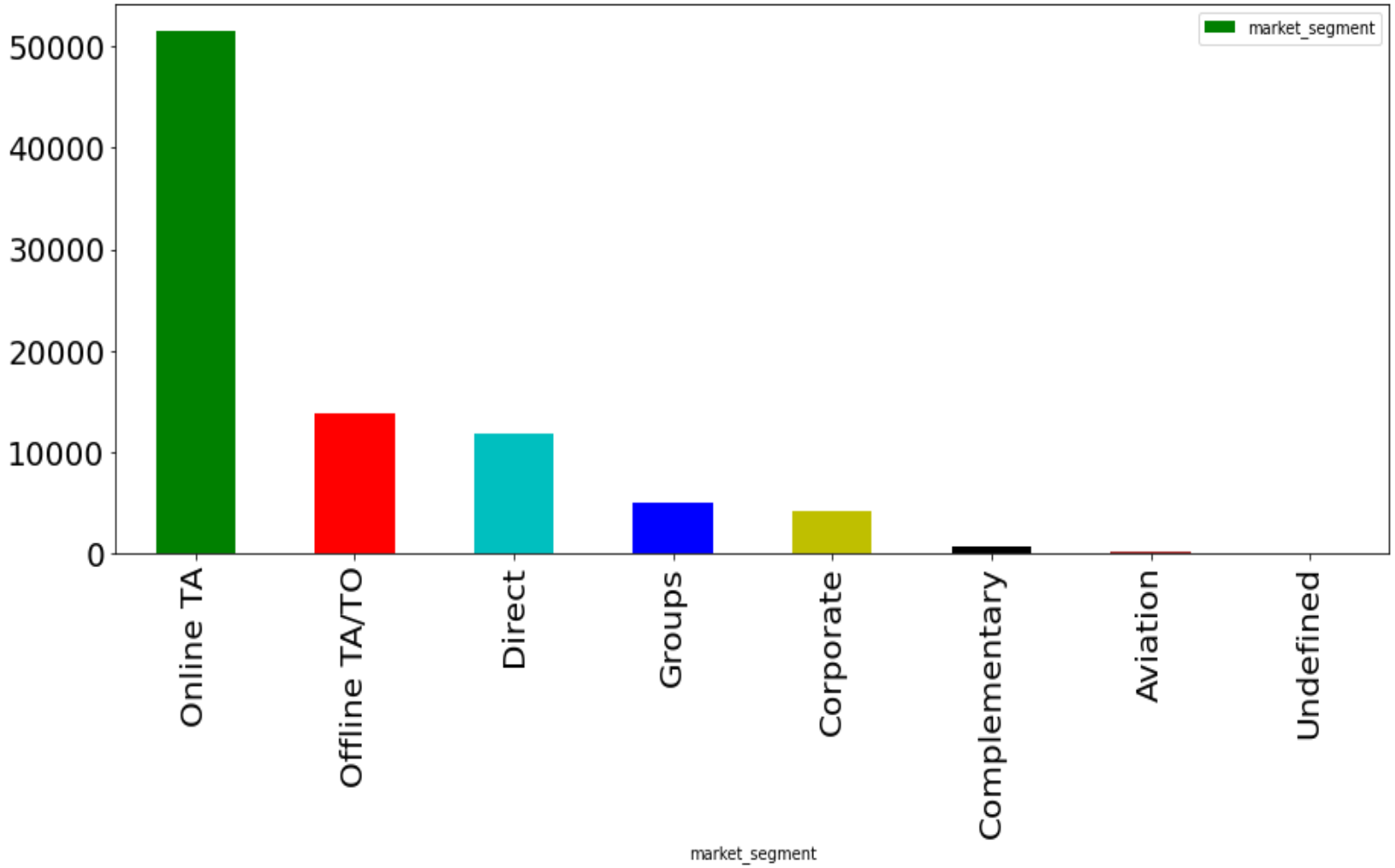
❑ The room type most preferred by guest.



- ✓ room type 'A' is most preferred by guest.
- ✓ Whereas room type 'D' is below the half of room type 'A.'

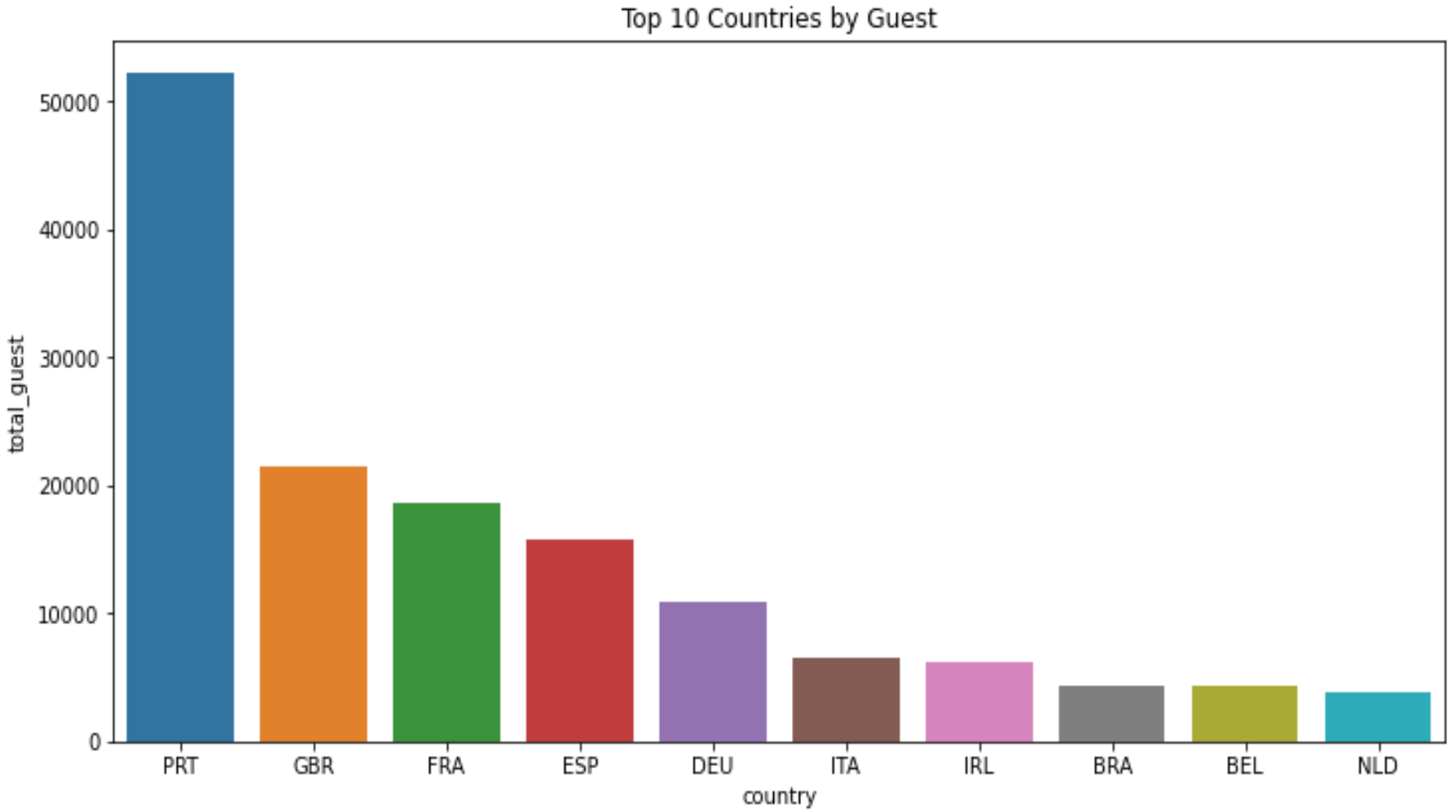
❑ market segment used most frequently to book hotel by the guest.

- ✓ As in chart, Online TA has been used most frequently to book hotel by the guest.



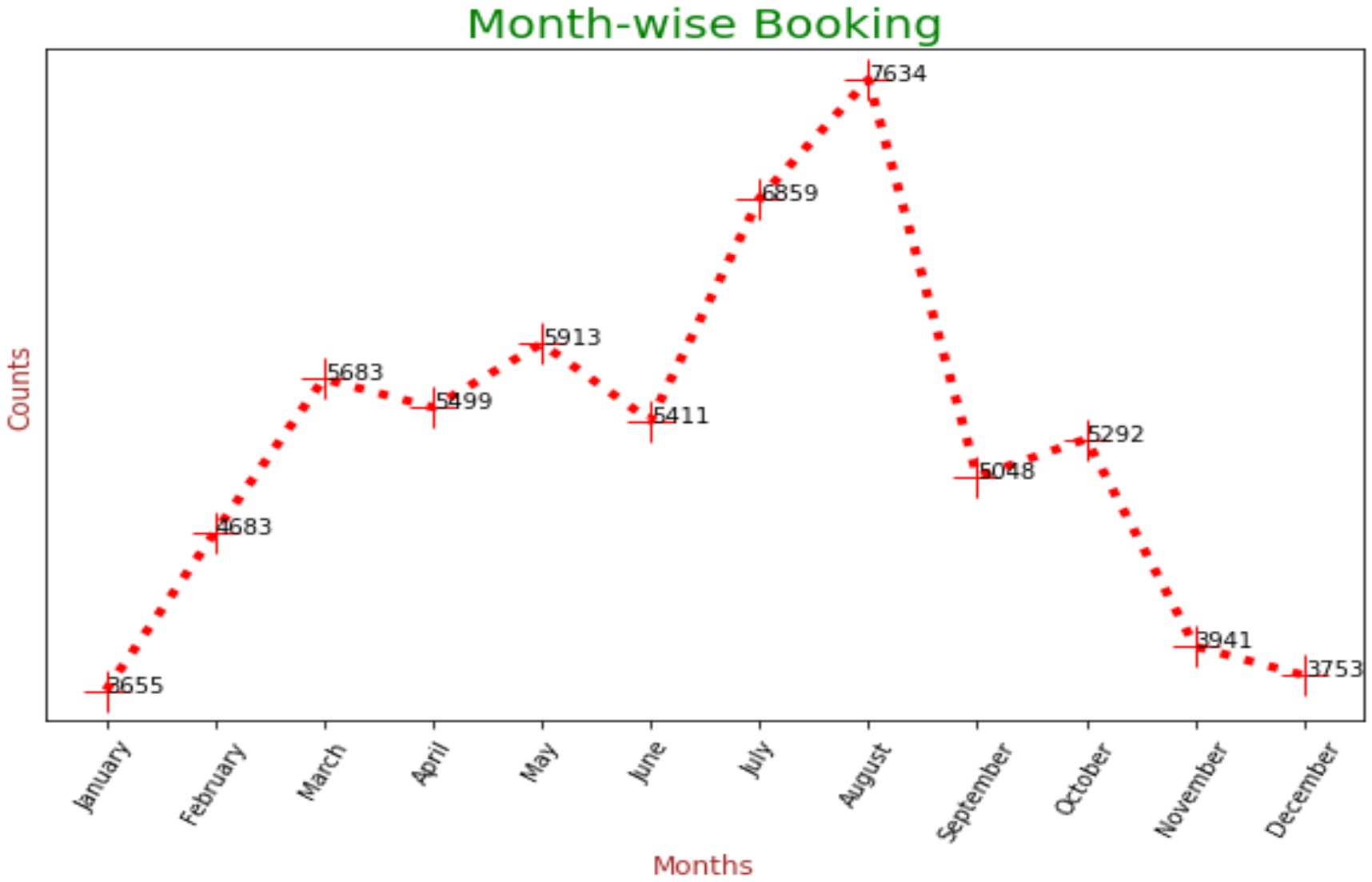
❑ maximum guest is coming from which country.

- ✓ **As we can see, that maximum guest is coming from Portugal.**



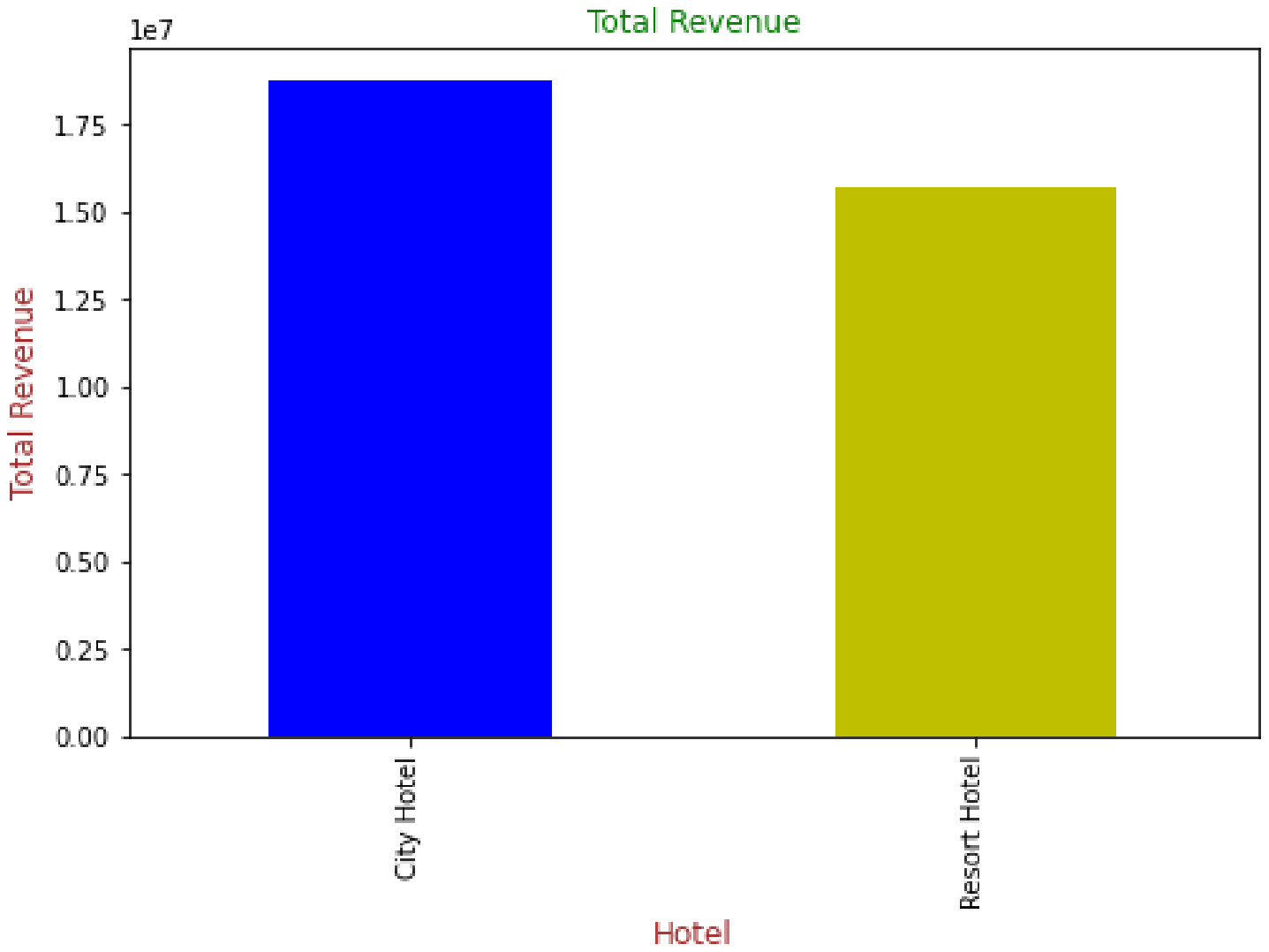
❑ months the maximum number of bookings happens.

- ✓ As we see, Month of July and August has the maximum no. of bookings.



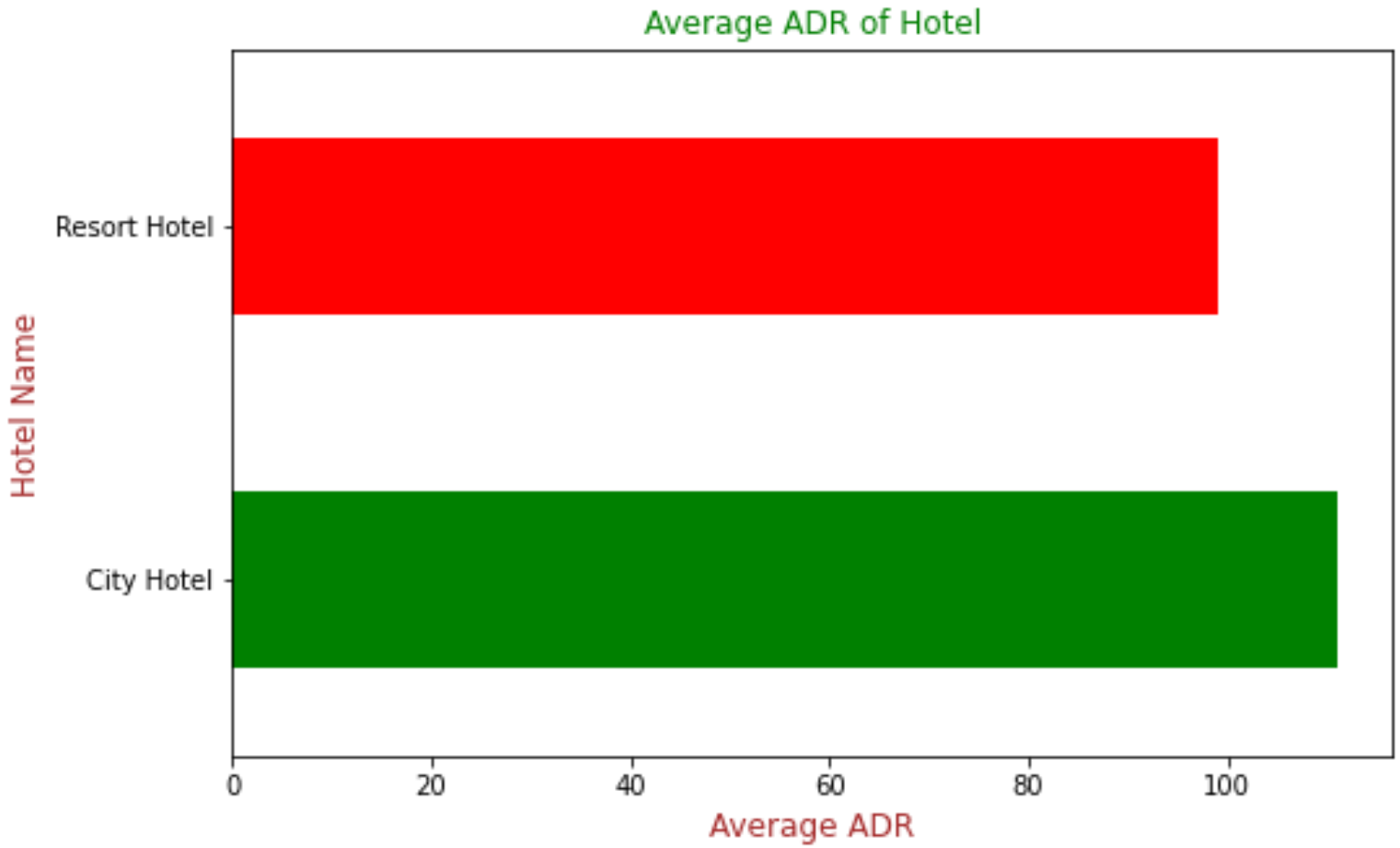
❑ hotel with high Revenue.

- ✓ City hotel has the high revenue than Resort hotel.
- ✓ However there's just slight difference between them.



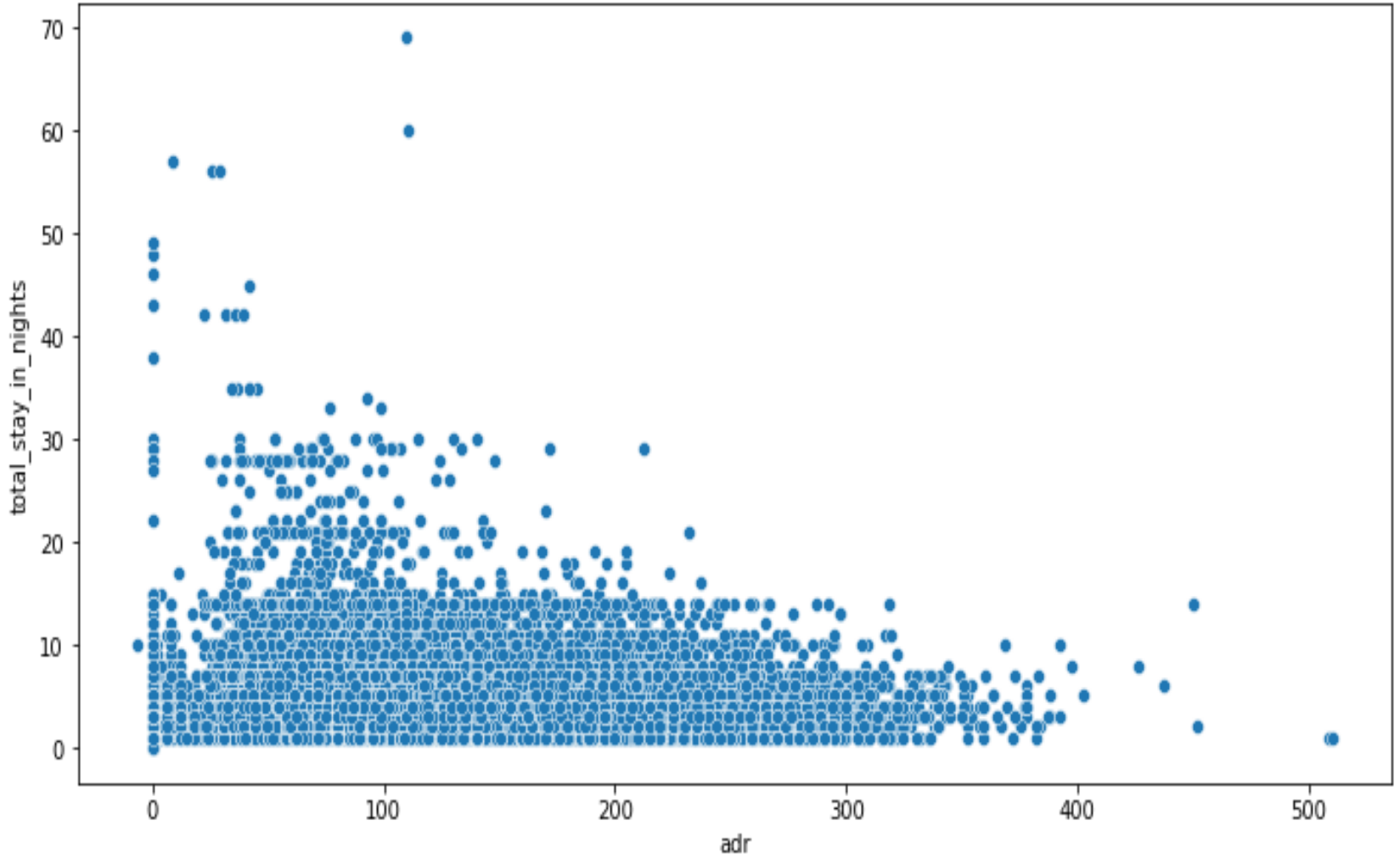
❑ Specifying the average ADR for both hotels.

- ✓ **As we can see the average ADR of City hotel is higher than Resort hotel, so the profit and revenue will be higher for city hotel.**



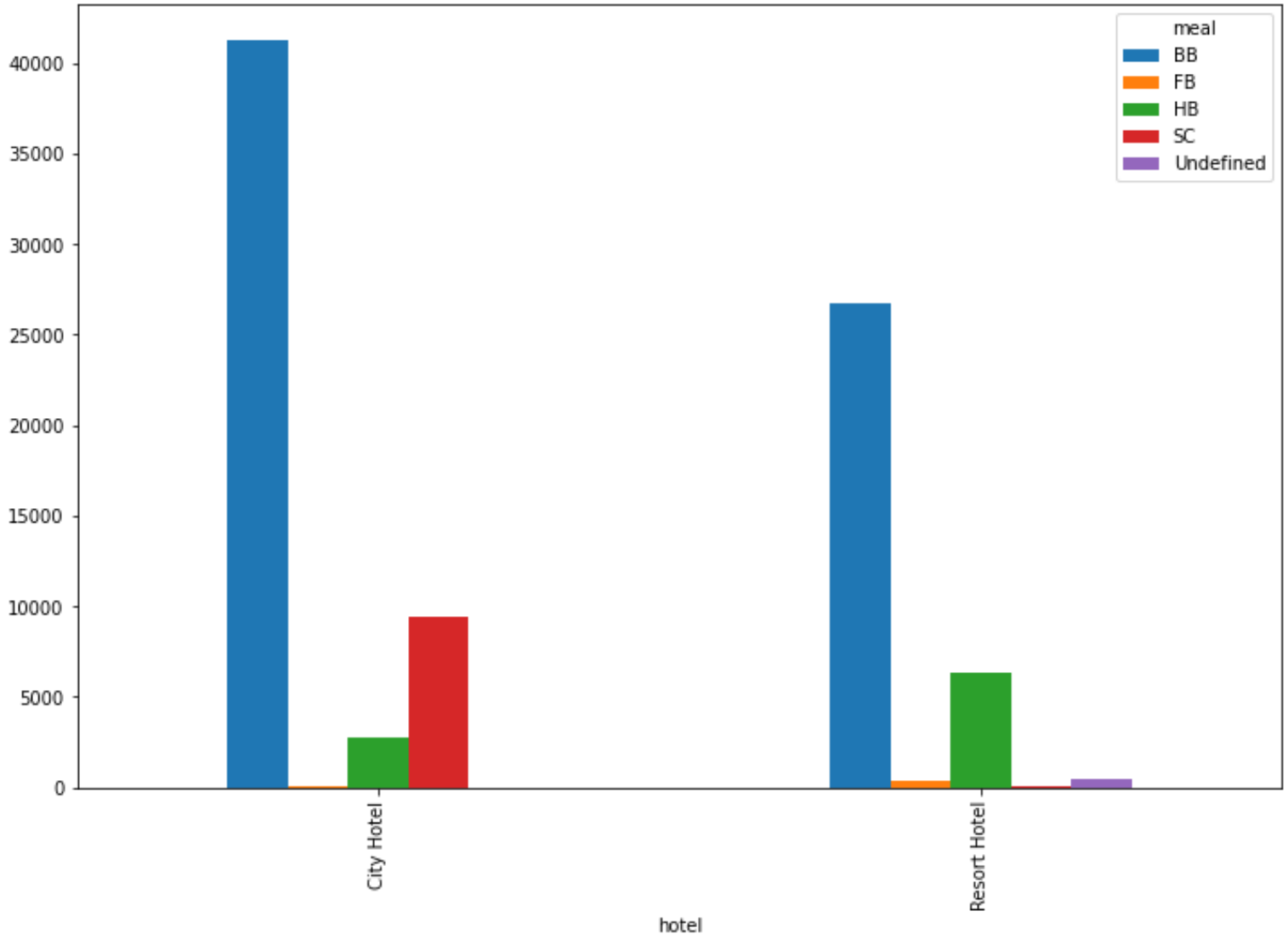
❑ comparison & effect of total stay days vs ADR.

- ✓ **Here, we found that if guest's stay days is getting decreased, ADR is getting high.**



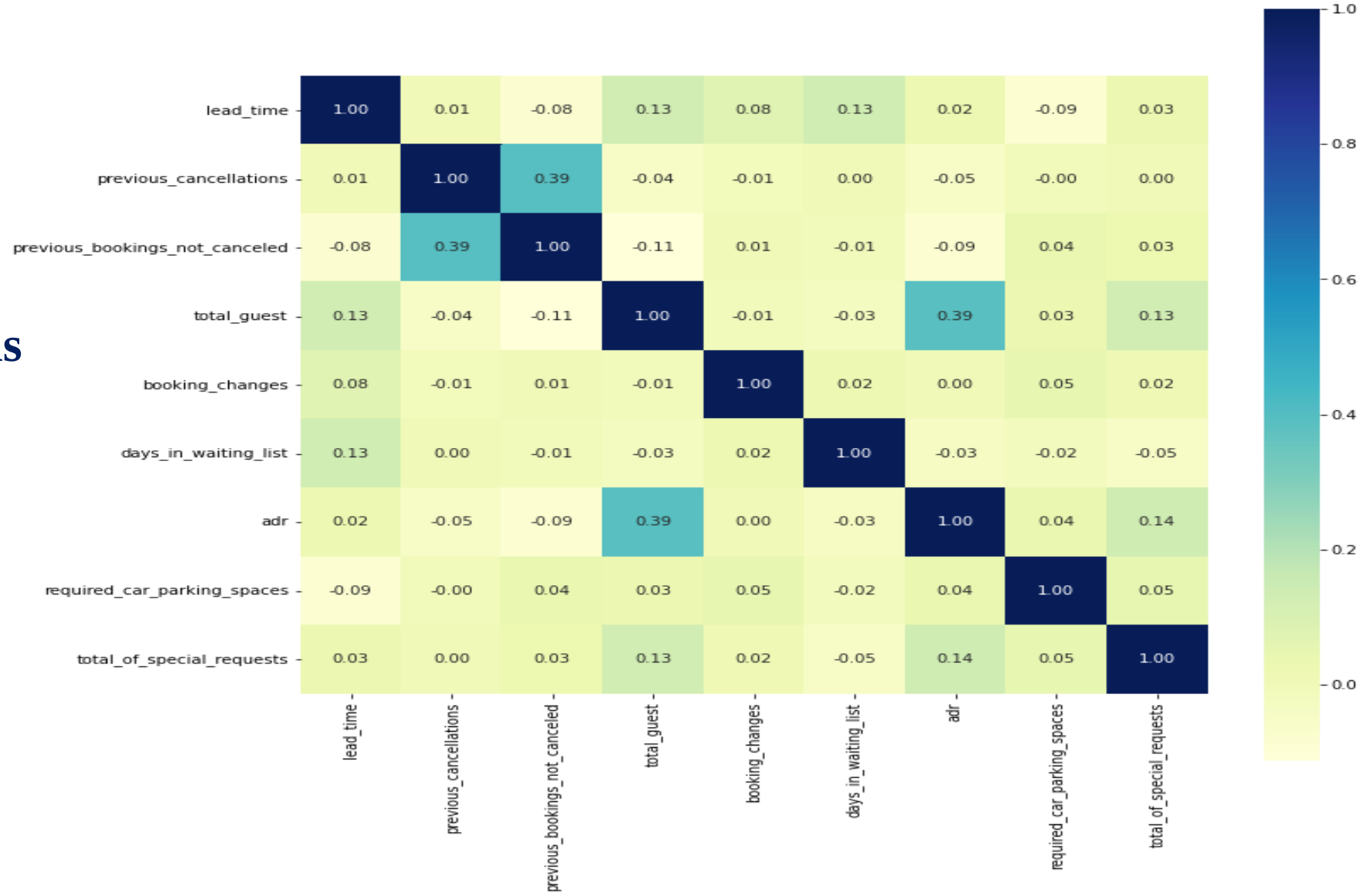
❑ meal most preferred by guests. Hotel-wise!

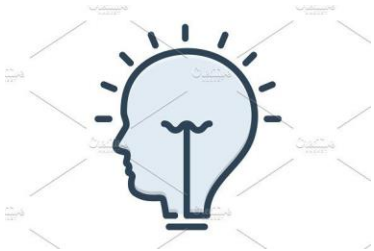
- ✓ As we can see, BB (Bed & breakfast) meal is most preferred by guests in both the hotels.
- ✓ So Hotel can give more delicious dishes in this meal to get customer repeat & attracts new customer.



❑ understand the relationship between different numerical values.

- ✓ **Highest correlation value between axis is 0.39% positive & lowest correlation value between the axis is -0.9% negative.**





Conclusion

1. City Hotel seems to be more preferred among travelers and it also generates more revenue & profit.
2. Most number of bookings are made in July and August as compared rest of the months.
3. Room Type A is the most preferred room type among travelers.
4. Most number of bookings are made from Portugal & Great Britain.
5. Most of the guest stays for 1-4 days in the hotels.
6. City Hotel retains more number of guests.

Conclusion cont...

7. Around one-fourth of the total bookings gets cancelled. More cancellations are from City Hotel.

8. New guest tends to cancel bookings more than repeated customers.

9. Lead time, number of days in waiting list or assignation of reserved room to customer does not affect cancellation of bookings.

10. Corporate has the most percentage of repeated guests while TA/TO has the least whereas in the case of cancelled bookings TA/TO has the most percentage while Corporate has the least.

11. The length of the stay decreases as ADR increases probably to reduce the cost.

Thank You