Capstone Project

Hotel Booking EDA Analysis

Kaveri Shende

Arshad Aafaq D

Sakshi Chaturvedi

Vikas Kumar

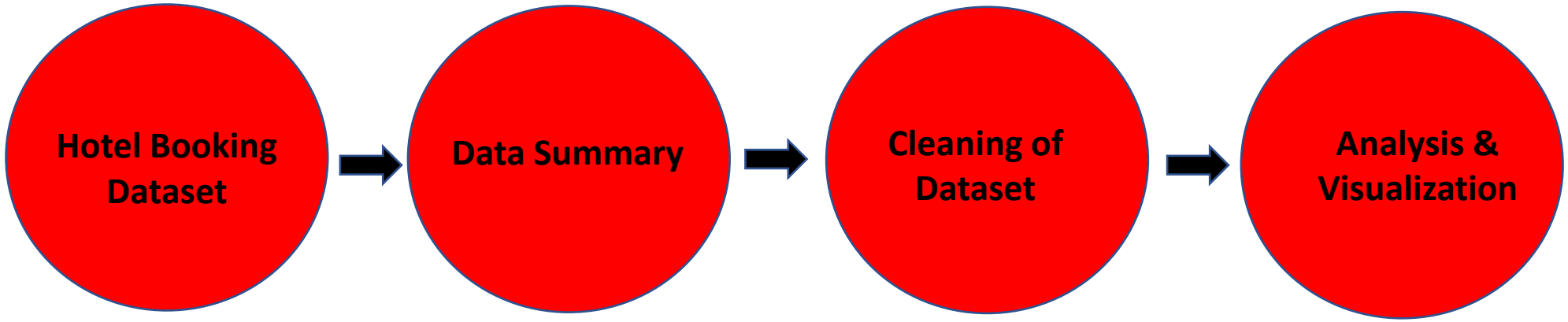
Yogesh Agre

Content:

- Problem Statement
- Why Hotel Booking Analysis and it’s data
- Data Summary
- Exploring our database
- Flow chart and EDA Process
- Cleaning of Data and Manupulation
- Data Manupulations
- Data Visualization
- Challenges faced during data exploration
- Conclusions of our analysis

Problem Statements:

- Our Agenda is to Analysis of various datasets that covariates governing the bookings of hotels.
- Hotel booking datasets has enormous potential to drive hotel booking business to success. Analysis of different insights can be drawn for Hotel owners to work on and capture the market growth.
- Our main objective is to perform EDA on the given dataset to discover key factors responsible for Hotel booking system and their success.
- We need to analyze the data and come up with meaningful insights that would actually help business to strategize their moves.



Hotel Booking Analysis and it’s data?

- The purpose of our project was to gather and analyze detailed information about hotels in order to provide insights and estimate the profit.
- The majority of Revenue Management research on demand forecasting and prediction issues is conducted in the tourism and travel-related industries.
- We have given two hotel data sets. i.e., the resort hotel is one of the hotels, and the city hotel is the other. There are 32 columns and 119390 rows.
- With out industry-specific data, it is impossible to completely understand the requirements and peculiarities of the remaining tourism and travel sectors, such as hospitality, cruising, theme parks, etc. To help overcome this restriction, two hotel datasets with demand data are given.
- Hotels will be able to identify the issue that is causing customers to cancel their bookings, as well as the reason for the cancellations, by utilizing the predictive
- It would be fantastic if the hotel management team could identify the root cause and develop a better strategy.
- The goal of our project was to collect and analyze detailed hotel information in order to provide insights and estimate profit.

Data Summary

- Analysis of the dataset over a span of Three years - 2015, 2016 and 2017

  - 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

Exploring our Database

In Hotel booking Dataset it contain basic information about two types of hotel i.e city hotel and resort hotels. We have given 32 rows and 119390 columns. The features of are:-

- **hotel** : Hotel type.
- **is_canceled** : value indicates if the booking is canceled or not.
- **lead_time** : How long in advance the booking was made.
- **arrival_date_year** : Customer arrival year.
- **arrival_date_month** : In which month of the year customer visited hotel.
- **arrival_date_week_number** : In which week of the year customer arrived.
- **arrival_date_day_of_month** : Date of the month customer visited hotel.
- **stays_in_weekend_nights** : Customer stayed or booked to stay in hotel during weekend nights.
- **stays_in_week_nights** : Customer stayed in hotel during week nights.
- **adults** : Number of adults
- **children** : number of children.

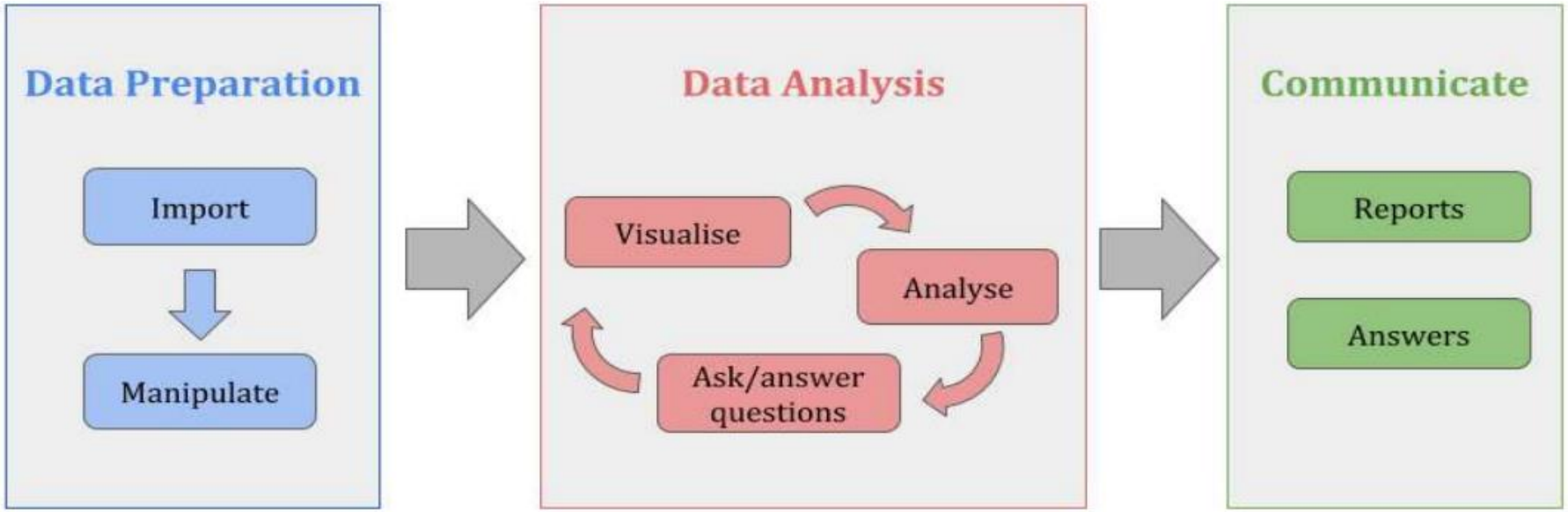
Exploring our Database Continued..

- **babies** : Number of babies.
- **meal** : Type of meal booked.:
- **country** : Country of orgin of cutomer.
- **market_segment** : where the bookings came from.
- **distribution_channel** : Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” .
- **is_repeated_guest** : Value indicating if the booking name was from a repeated guest (1) or not (0).
- **previous_cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking.
- **previous_bookings_not_canceled** : umber of previous bookings that were cancelled by the customer prior to the current booking.
- **reserved_room_type** : Code of room type reserved. Code is presented instead of designation for anonymity reasons.
- **assigned_room_type** : Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due.
- **booking_changes** : Number of changes/amendments made to the booking from the moment the booking was entered on the PMS.

Exploring our Database Continued..

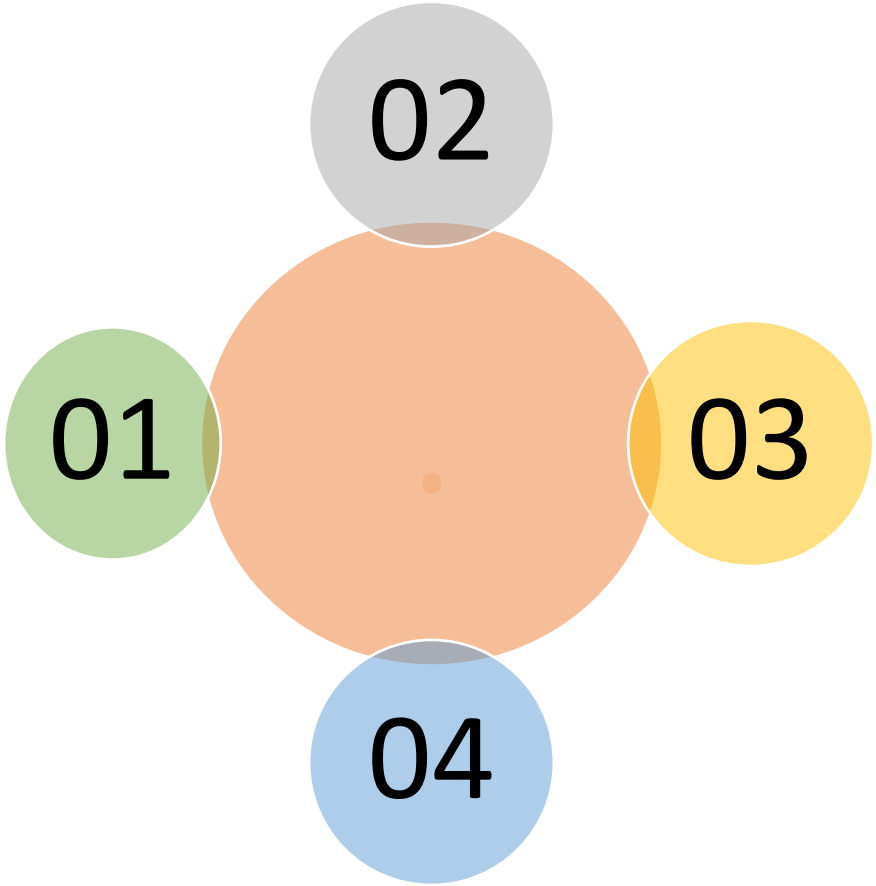
- **agent** : ID of the travel agency that made the booking.
- **company** : ID of the company/entity that made the booking or responsible for paying the booking.
- **days_in_waiting_list** : Number of days the booking was in the waiting list before it was confirmed to the customer.
- **customer_type** : Type of booking, assuming one of four categories.
- **adr** : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights.
- **required_car_parking_spaces** : Number of car parking spaces required by the customer.
- **total_of_special_requests** : Number of special requests made by the customer (e.g. twin bed or high floor).
- **reservation_status** : Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out: customer check out from hotel,No show: Customer did not check-in hotel and informed hotel with reason.
- **reservation_status_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 cancelled or when did the customer checked out of the

Flowchart and EDA Process



Cleaning of Data and Manupulation

Company and agent values replaced with zero



Dropped duplicate rows from the dataset

Converting datatype of colu mns 'children', 'company' and 'agent' from float to int

changing datatype of column 'reservation _status_date' to data_type.

Removing Duplicates

```
# Determining duplicate values in our Hotel Booking dataset
print(main_df[main_df.duplicated()].shape)
print(main_df.duplicated().sum())
```

```
(31994, 32)
31994
```

```
#Dropping the duplicate values from hotel booking dataset
df_hotel = df_hotel.drop_duplicates()
```

```
#Recheking the shape of our hotel booking dataset after dropping all the duplicates
df_hotel.shape
```

```
(87396, 32)
```

```
#Recheking our hotel booking dataset wheater they have any more duplicate values.
df_hotel.duplicated().sum()
```

```
0
```

Dealing with null values of Hotel Booking

There are 4 columns Children, Agent, Company, Country with Null values

```
[18] df_hotel.isna().sum().sort_values(ascending=False)
```

company	82137
agent	12193
country	452
children	4
reserved_room_type	0

```
#Company and agent values replaced with zero
df_hotel[['company','agent']] = df_hotel[['company','agent']].fillna(0)
```

```
] df_hotel['children'].fillna(df_hotel['children'].mean(), inplace = True)
```

```
] df_hotel['country'].fillna('others', inplace = True)
```

```
] null_detail(df_hotel)
```

Company and agent values is replaced with Zero, country is replaced with others

Converting Datatypes

```
[ ]  # Converting datatype of columns 'children', 'company' and 'agent' from float to int.
     df_hotel[['children', 'company', 'agent']] = df_hotel[['children', 'company', 'agent']].astype('int64')
```

```
# changing datatype of column 'reservation_status_date' to data_type.
df_hotel['reservation_status_date'] = pd.to_datetime(df_hotel['reservation_status_date'], format = '%Y-%m-%d')
```

Adding Required Columns

```
# Adding total staying days in hotels
df_hotel['total_stay'] = df_hotel['stays_in_weekend_nights']+df_hotel['stays_in_week_nights']

# Adding total people num as column, i.e. total people num = num of adults + children + babies
df_hotel['total_people'] = df_hotel['adults']+df_hotel['children']+df_hotel['babies']
```

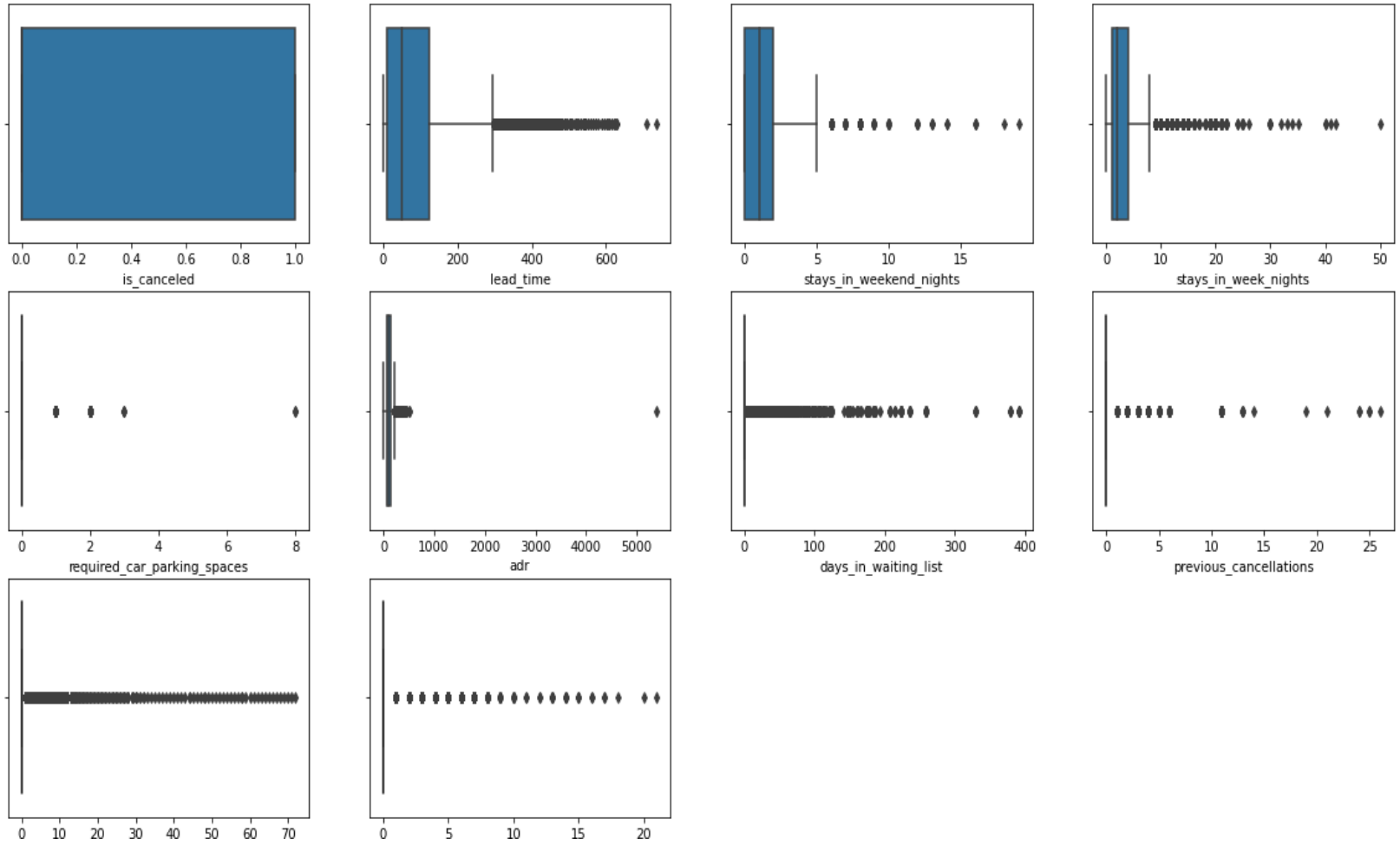
Data Visualization

```
correlation = df_hotel[['lead_time','previous_cancellations','previous_bookings_not_canceled','booking_changes',
                        'days_in_waiting_list','adr','required_car_parking_spaces','total_of_special_requests','total_stay','total_people']]
```

```
#correlation matrix
plt.figure(figsize=(12, 6))
sns.heatmap(correlation.corr(), annot=True)
```

- The total length of stay and the lead time have a slight correlation. This could imply that for longer hotel stays, people generally plan little ahead of time.
- Ad revenue is slightly correlated with total people, which makes sense because more people means more revenue, and thus more ad revenue.

Datavisualization



As we can see, this dataset contains a large number of outliers, implying that it is not very reliable.

Data Visualization

Lets take some insights which we have pulled from our analysis

1 what is the count of each type of Hotels ?

2 In which month maximum hotel were booked ?

3 What is the booking rate according to the population?

4 Which form of distribution do customers prefer most?

5 Which hotel will have long-term guests?

6 Which type of food is preferred by the guest?

7 Which hotel has a higher rate of returning customers?

8 which type of hotel is mostly preferred by adults , children or babies

9 Which hotel will have long-term guests?

10 Which hotel produces maximum revenue?

11 Which distribution route has given adr the most boost in terms of revenue?

12 Which room type has highest adr?

13 ADR across different market segment?

14 ADR across the different month.

15 Which month saw the most canceled reservations?

16 Which hotel has the highest cancellation rate, the city or the resort?

17 determining which countries have the most hotel cancellations in different type of hotels

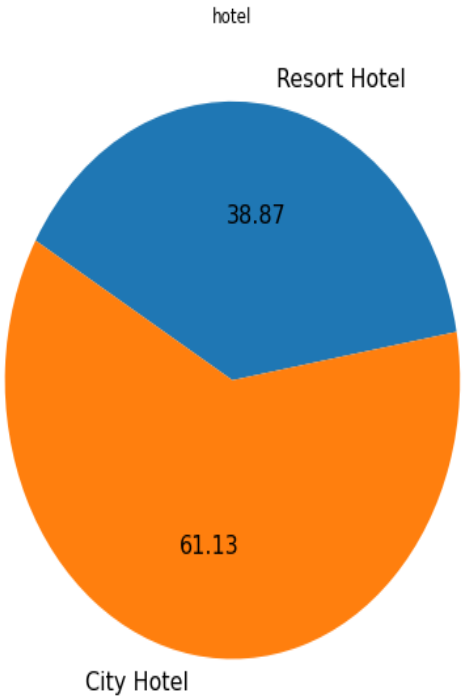
18 Does longer waiting period causes booking cancellation?

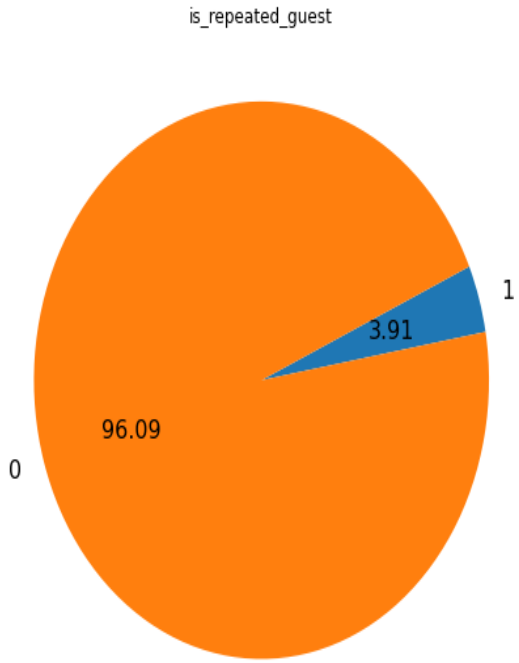
19 What is the percentage distribution of required_car_parking_spaces?

20 Which type of food is preffered by the guest?

Data Visualization

Which type of Hotel is mostly preferred by the customer ? No of repeated guests?

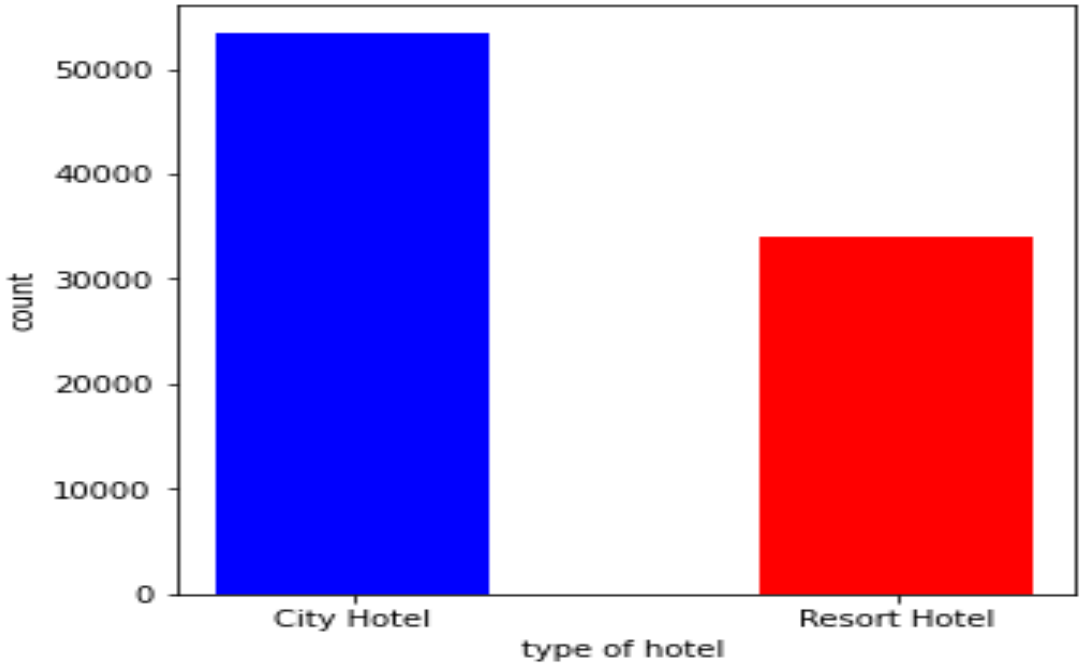




Resorts hotels were the preferred choice between city and resorts by the customer with 61.13 % booking. It could be attribute to good customer facility.
New Guests are more as 3.91% guests were repeated.

Data Visualization

What is the count of each type of Hotels ?

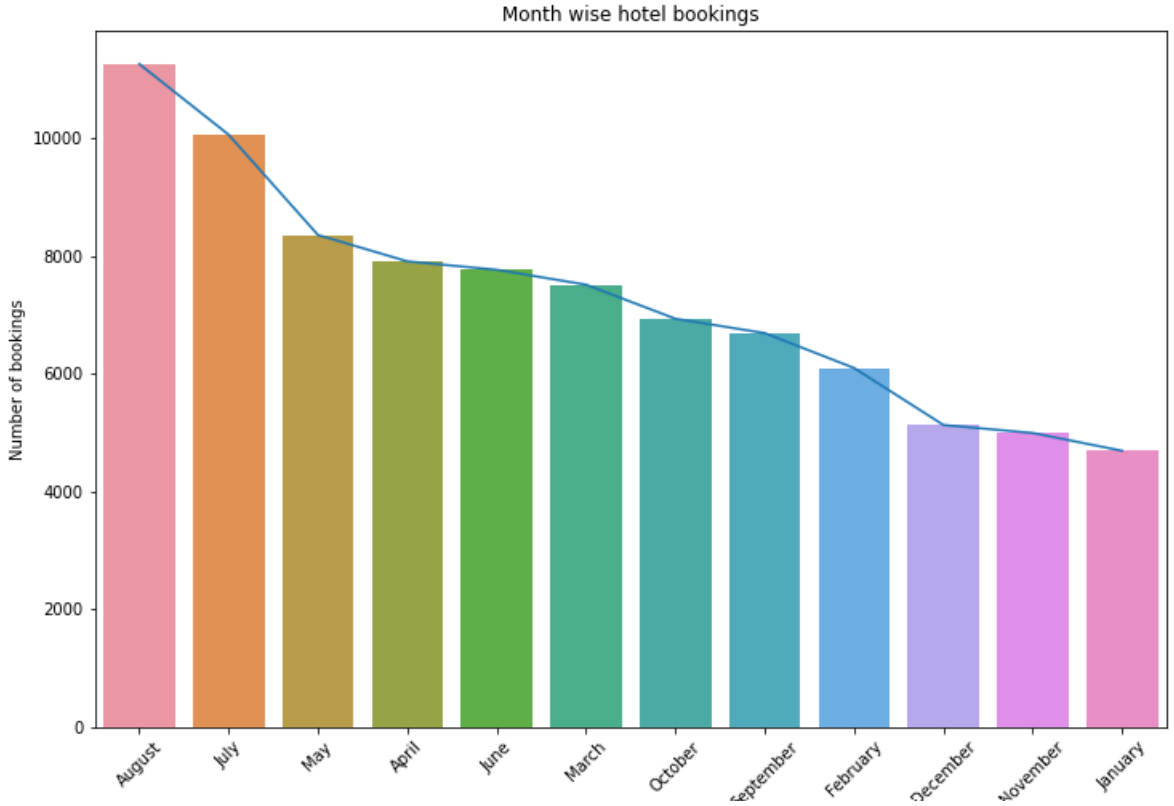


City Hotel	53427
Resort Hotel	33968

From the above, we learned that people are booking city hotels more than resort hotels. Now we will find out in which month the people book the hotel.

Data Visualization

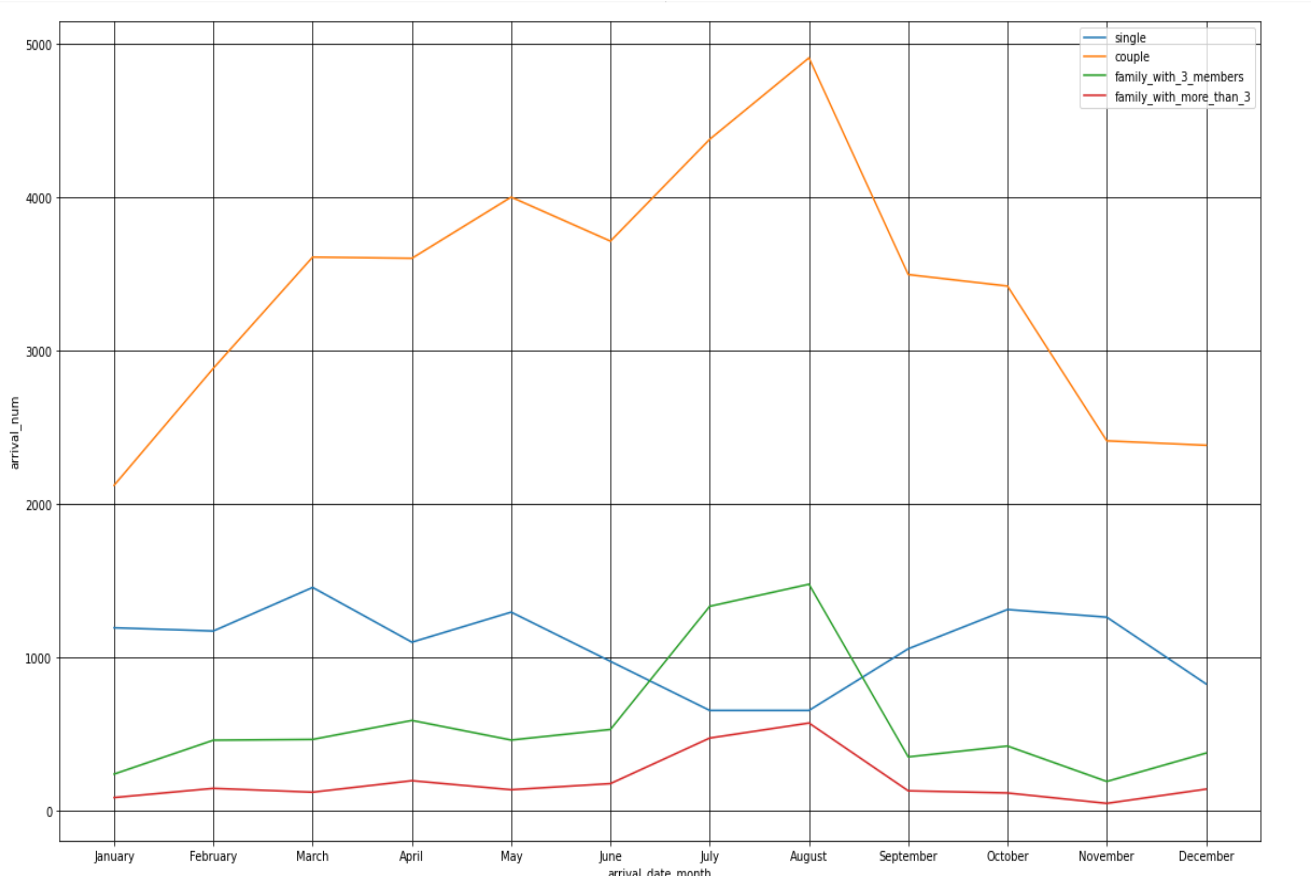
Which of the two hotels is preferred by customers, and in which month most hotels were booked?



As a result of this , we learned that city hotels were in high demand, with bookings peaking in August.

Data Visualization

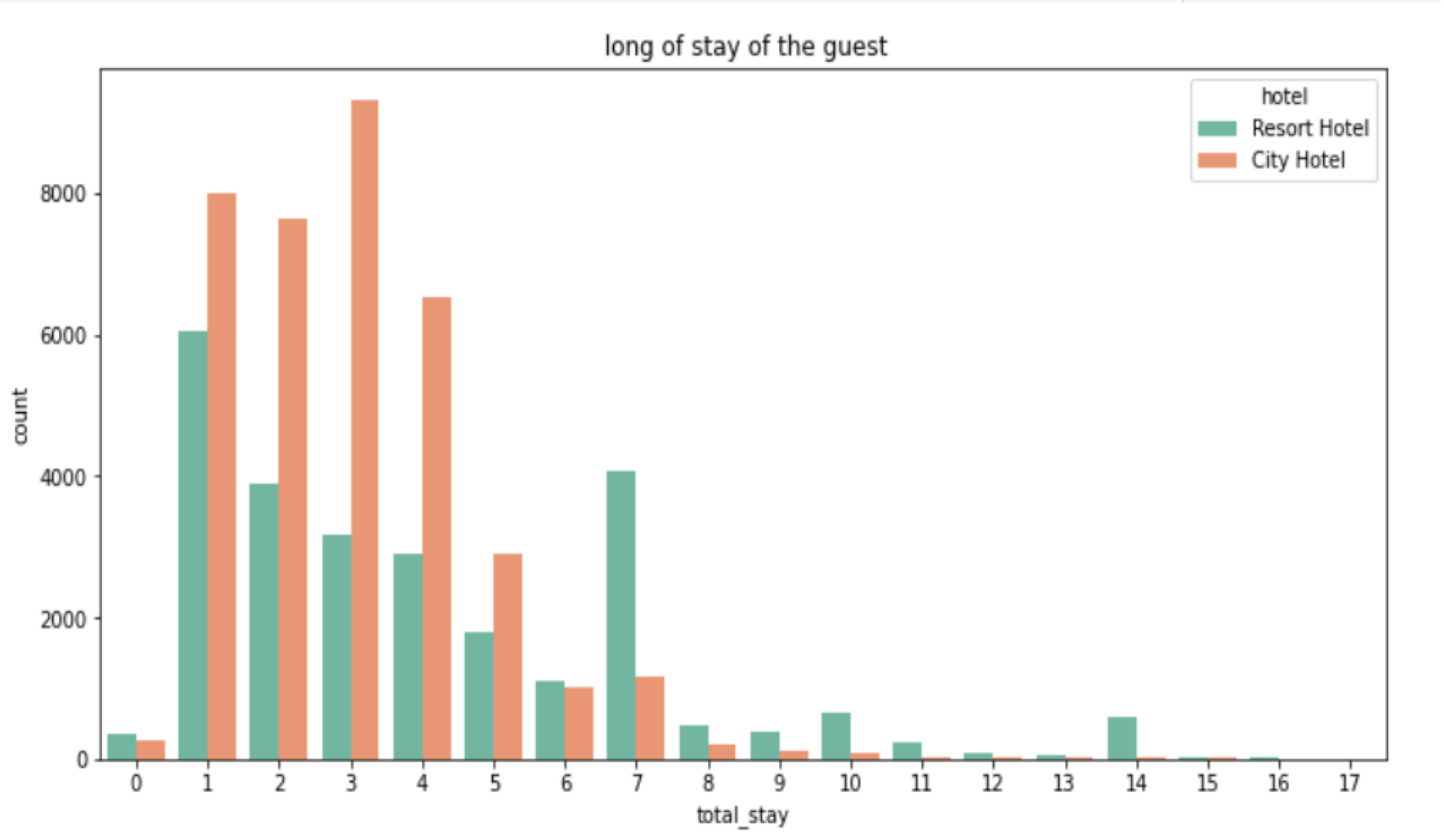
What is the booking rate according to the population?



- It appears that couples made the majority of reservations.
- The month of August saw the greatest number of reservations.
- Bookings for hotels increased in the months of June, July, and August.
- Bookings for families with three or more members are the least expensive.

Data Visualization

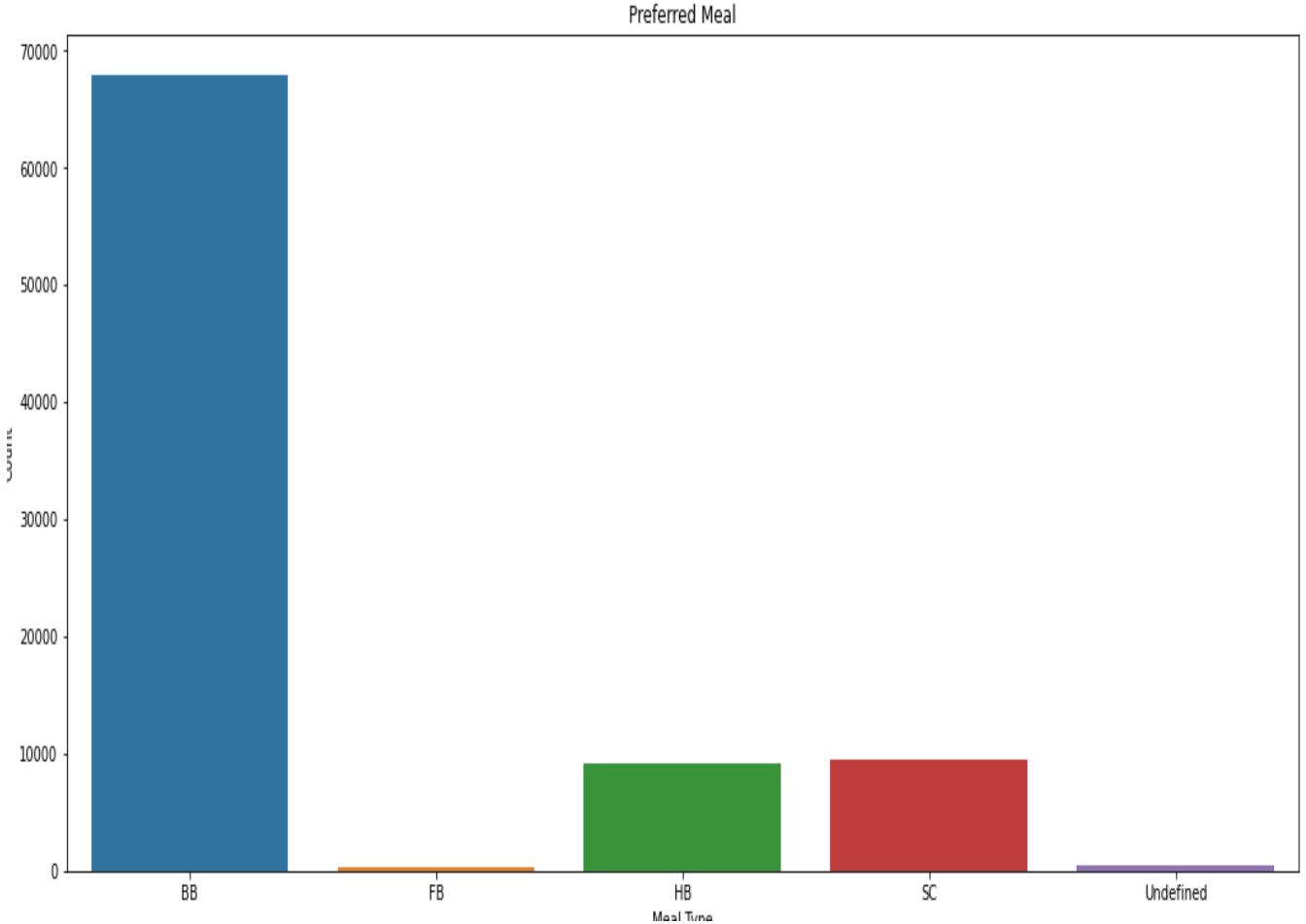
Which hotel will have long-term guests?



Most visitors of resort hotel stayed for one day, however most city hotel guests spent anywhere between one and seven days.

Data Visualization

Which type of food is preferred by the guest?



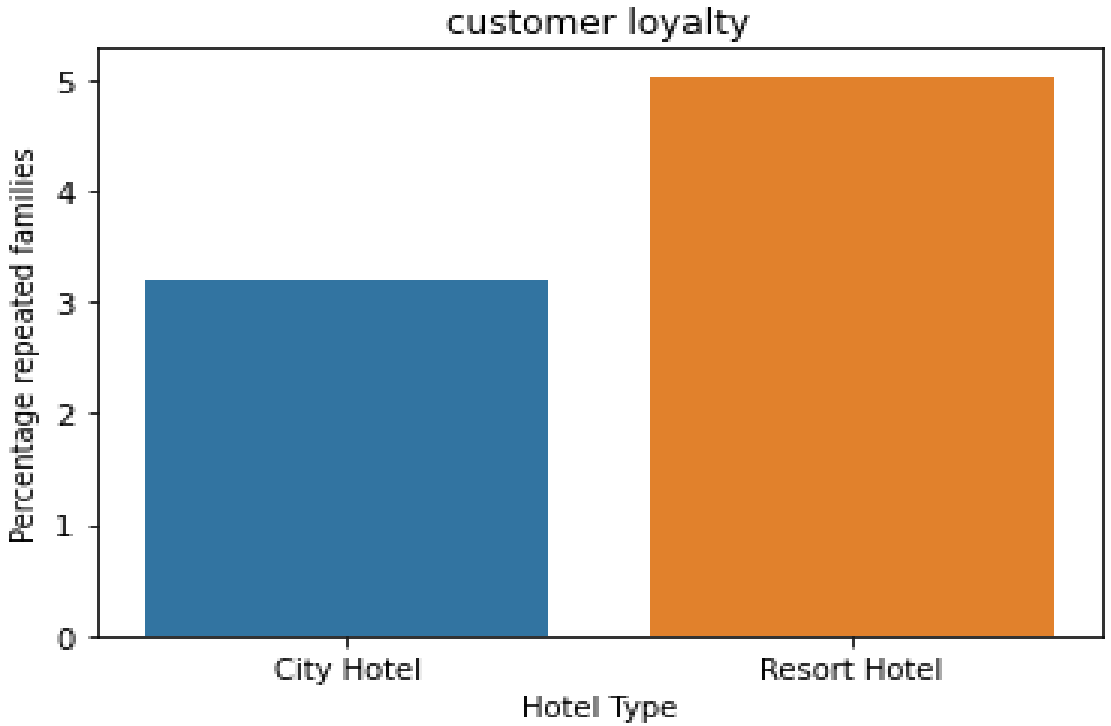
Meal types in hotels:
BB - (Bed and Breakfast)
HB- (Half Board)
SC- (Supplemental Committee)
FB- (Full Board) (Self Catering)

As a result, the most popular meal type among guests is BB (Bed and Breakfast), followed by HB (Half Board) and SC (Self Catering).

Data Visualization

Which hotel has a higher rate of returning customers?

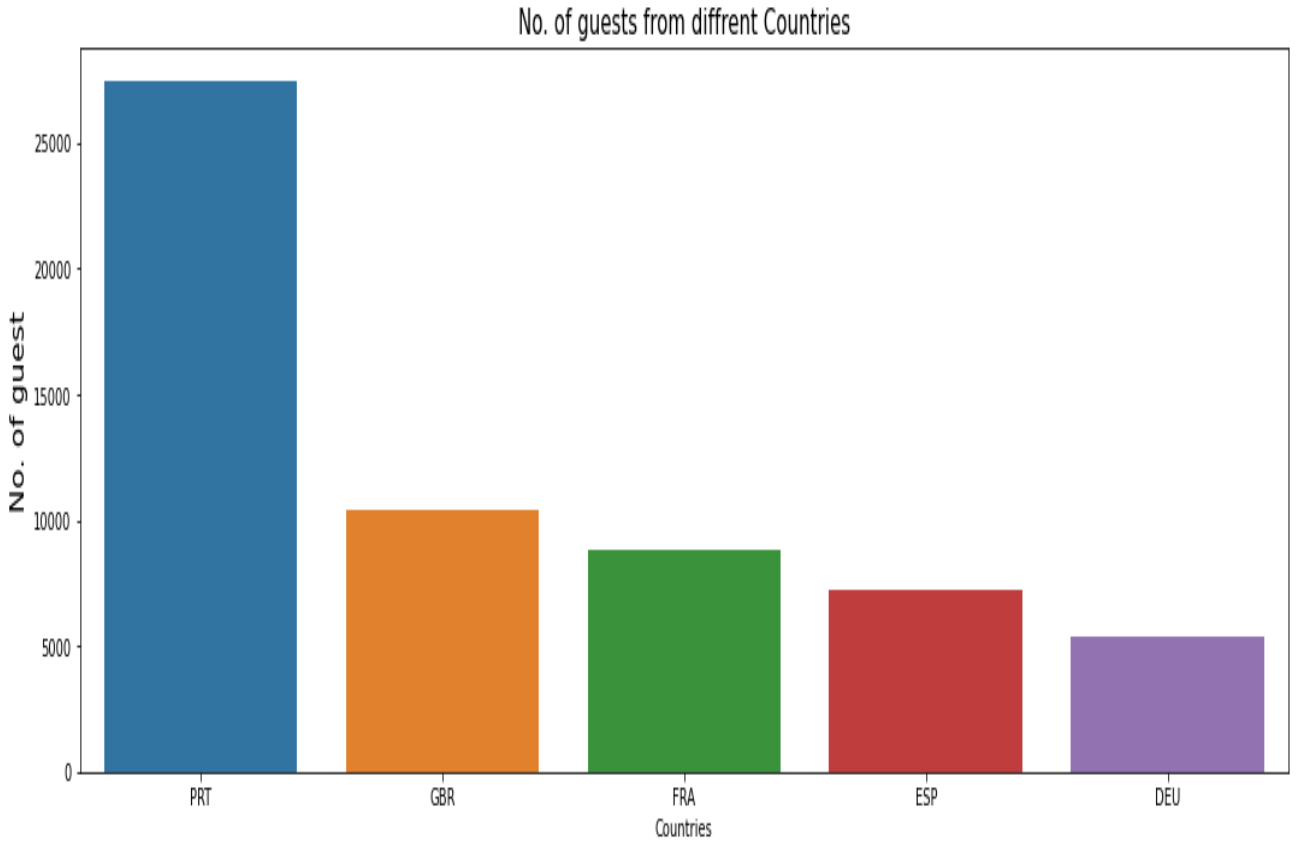
	Hotel Type	Total Families	Repeated Families	Percentage repeated families
0	City Hotel	53427	1708	3.20
1	Resort Hotel	33968	1707	5.03



From the above graph it is clear that highest rate of returning customers are from the resort hotel.

Data Visualization

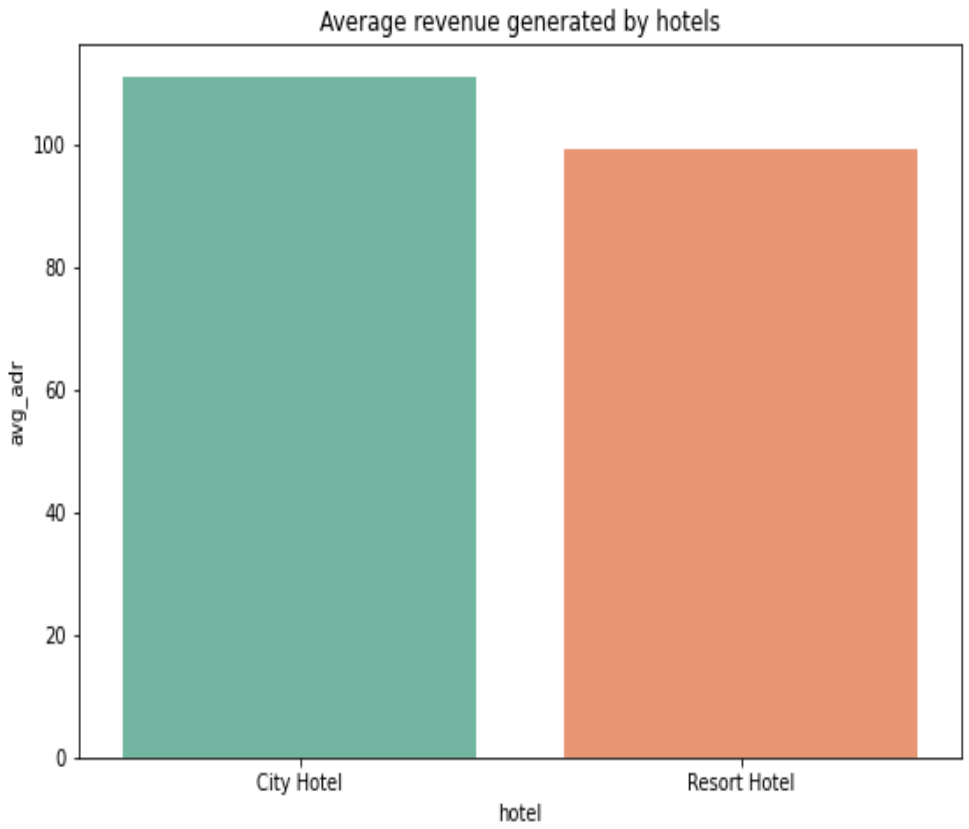
The maximum number of guests are from which country?



Observation : More than 25000 people, or the majority of the attendees, are from Portugal.
Abbreviations for nations:
PRT- Portugal GBR- United Kingdom FRA- France ESP- Spain DEU - Germany

Data Visualization

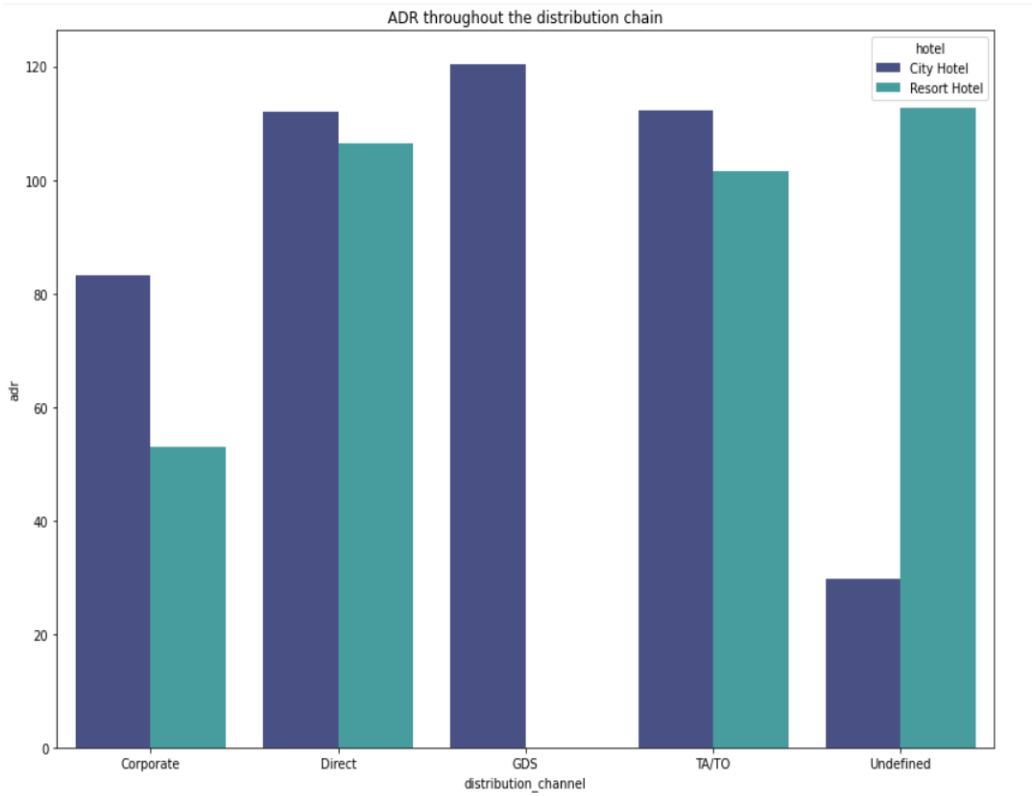
Which hotel produces maximum revenue?



According to the above graph, the average revenue of city hotels is higher than that of resort hotels Observation

Data Visualization

Which distribution route has given ADR the most boost in terms of revenue?



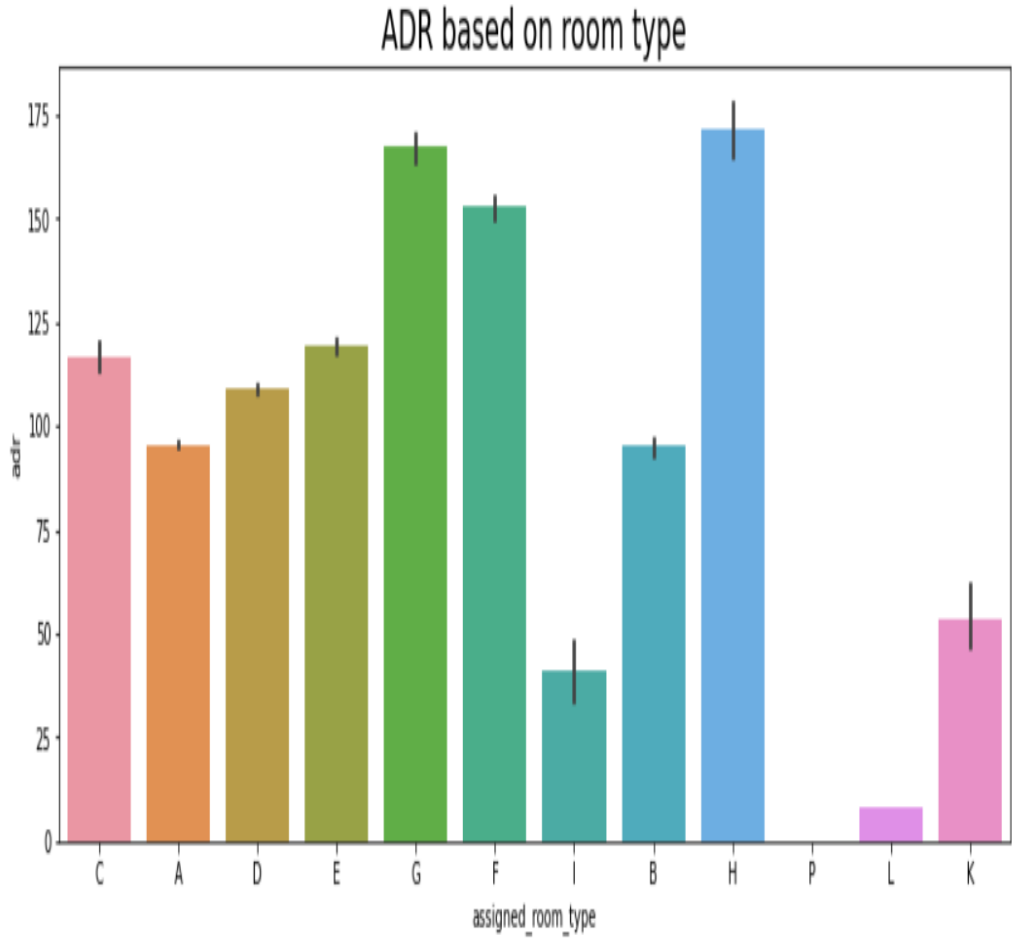
Corporate - These are companies that help businesses make hotel reservations.
GDS-GDS - serves as a global link between travel agents and suppliers, including hotels and other lodging establishments. It enables automated transactions and provides real-time product, price, and availability data to travel agencies and internet booking engines.
Direct - refers to making reservations with the specific hotels directly.
TA/TO - Bookings are made through travel agents or travel operators.
Undefined – Reservations are not defined. Maybe customers made their reservations when they arrived.

Inference

•In both types of hotels, "Direct" and "TA/TO" have contributed to adr about equally.
•GDS made a significant contribution to adr of the "City Hotel" type.
•GDS contributing little to hotel reservations at resorts.

Data Visualization

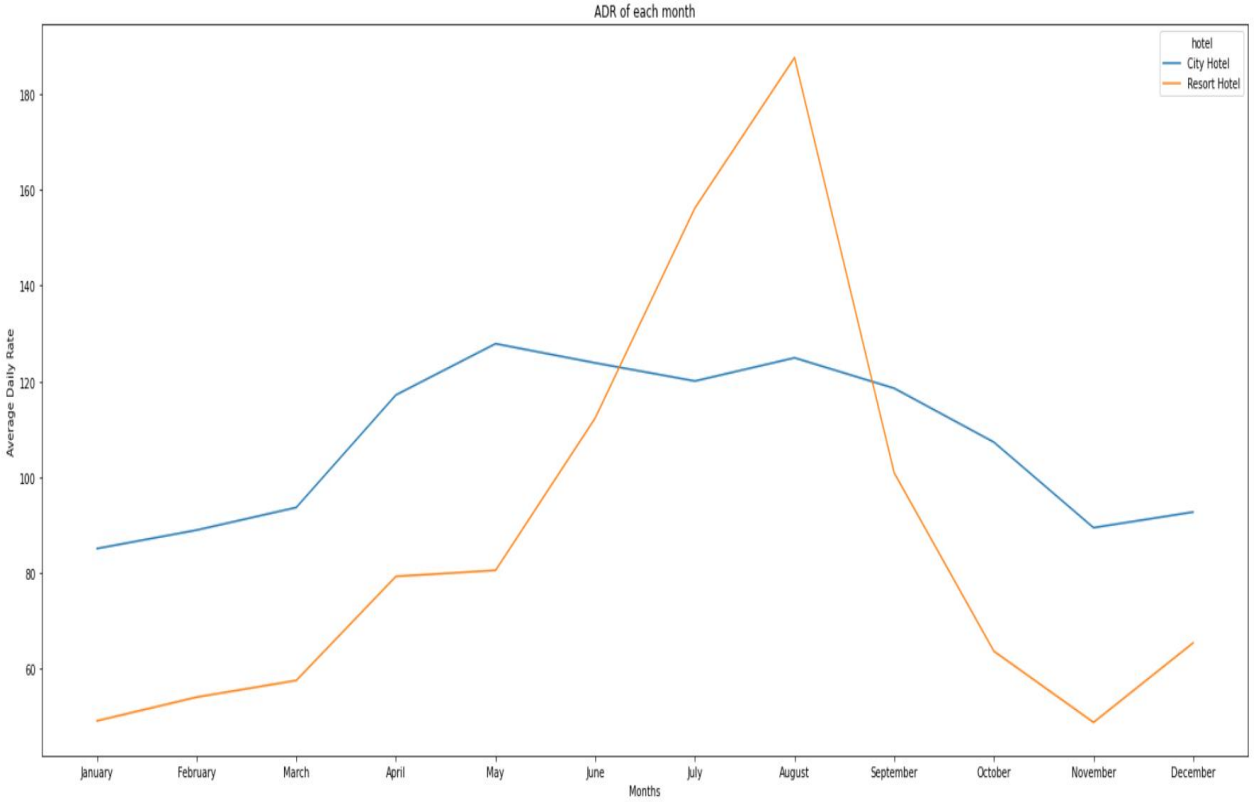
Which room type has the highest average daily rate?



H type has the highest Average daily rate followed by G type and F type

Data Visualization

In which month do the hotels have the highest ADR?

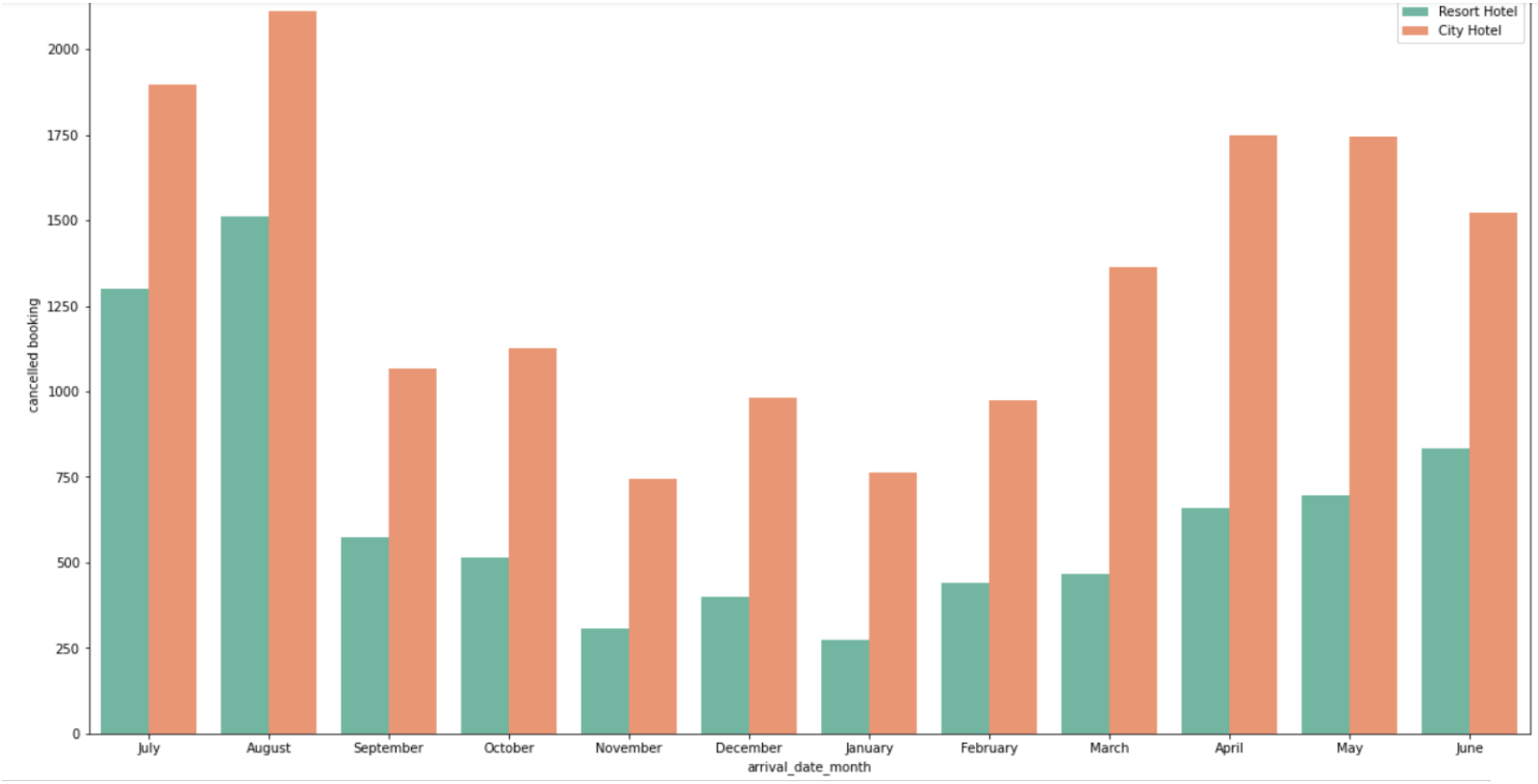


In comparison to City Hotels, the ADR for Resort Hotel is higher in the months of June, July, and August. Perhaps clients/people wish to vacation in resort hotels this summer.

January, February, March, April, October, November, and December are the ideal months for visitors to resort or city hotels because of the low average daily rate throughout these months.

Data Visualization

Which month saw the most canceled reservations?



For hotels in cities, the majority of cancellations occurred in the month of October, but for hotels in resort areas, the majority occurred in the month of August. Additionally, similar cancellations of reservations for both hotel types occurred in the month of August. City hotels had greater cancellations of reservations overall.

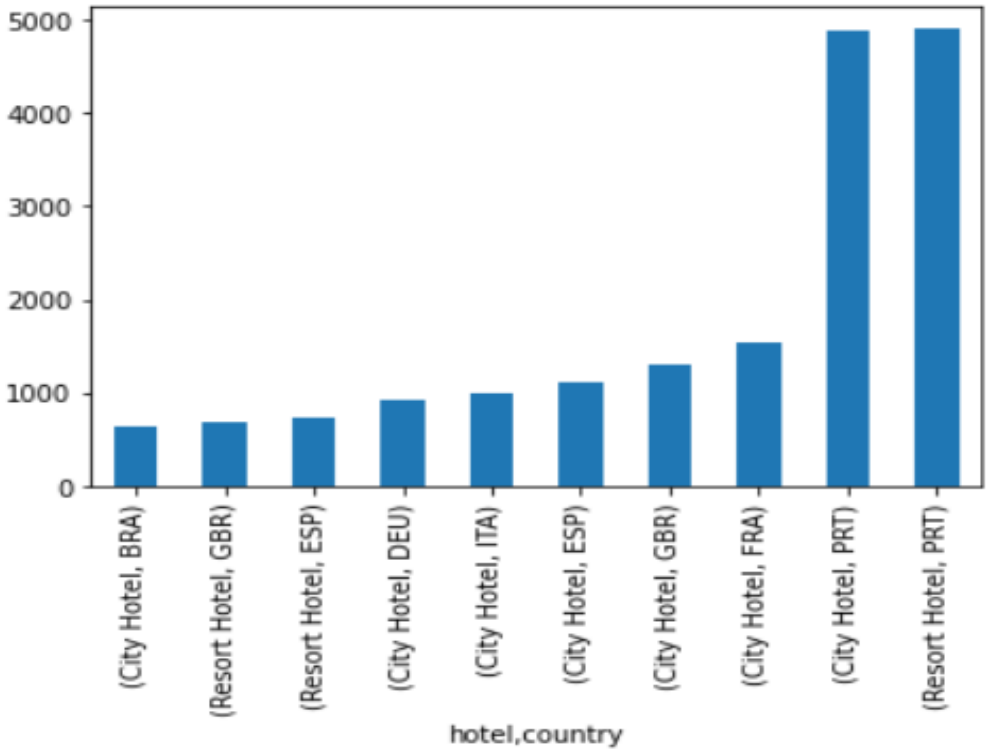
Data Visualization

Which hotel has the highest cancellation rate, the city or the resort?

About 30% of hotel reservations for city hotels and 24% for resort hotels are cancelled.

Data Visualization

Determining which countries have the most hotel cancellations in different type of hotels

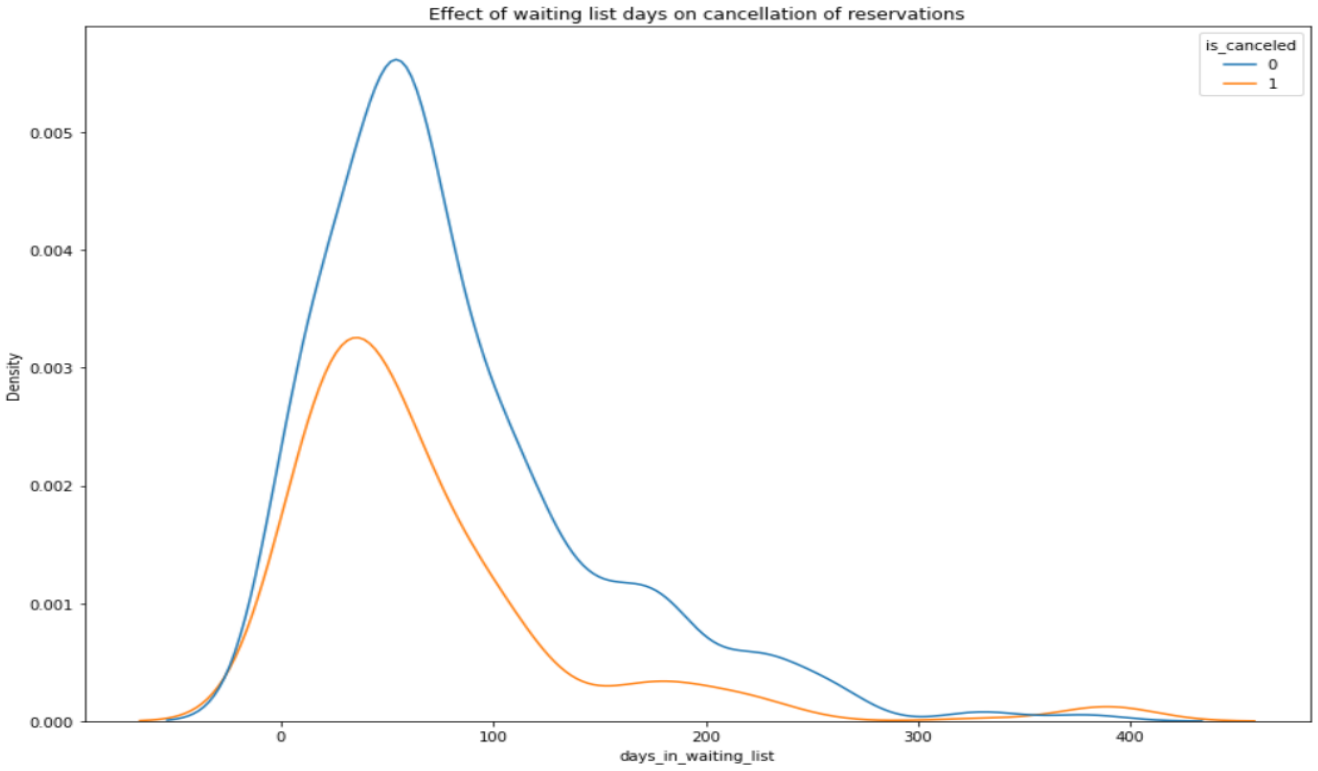


Here we can say that PRT(Portugal) country made the highest number of cancelation under the city hotel

Here we understand that the PRT country has made the large number of cancelation in resort type hotel.
so after the compare with both bar chart then we get that both hotel got most cancelation by PRT country.

Data Visualization

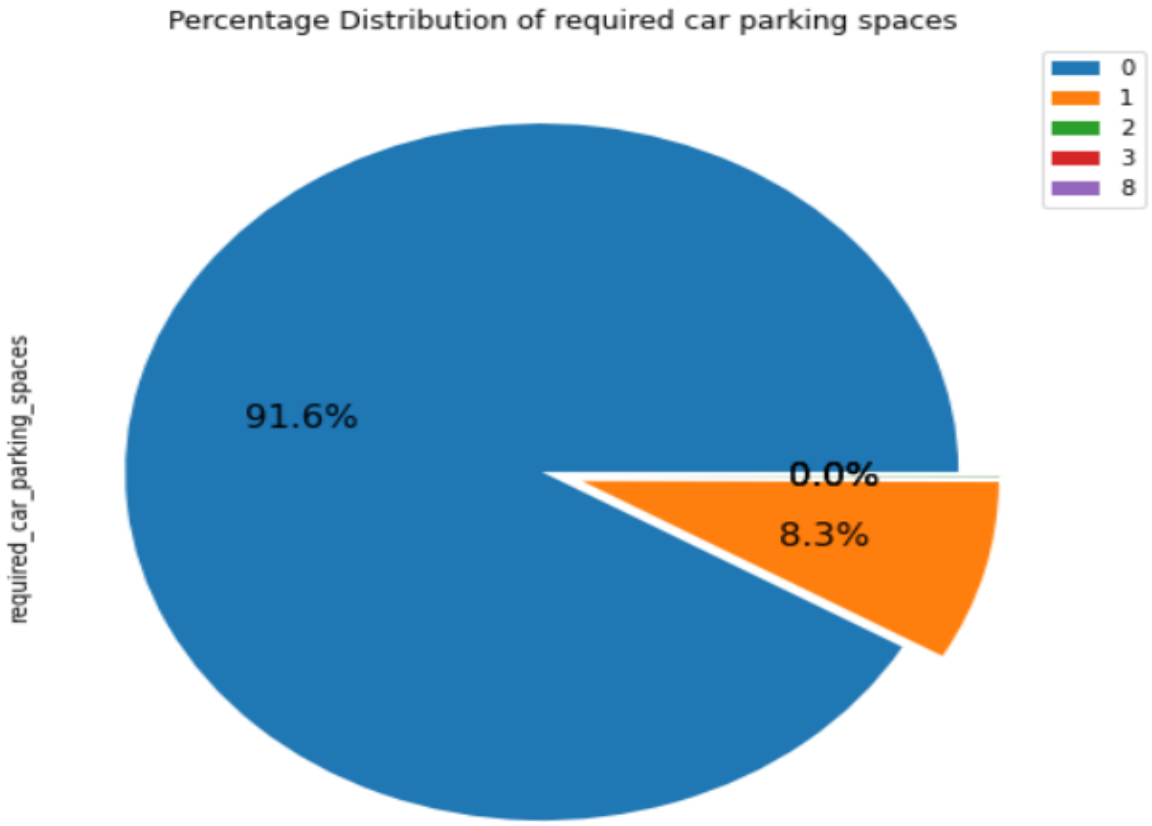
Does a longer waiting period result in cancelled bookings?



There is no direct correlation between a longer waiting period and booking cancellation, as can be seen from the fact that the majority of reservations that had less than 100 days on the waiting list were cancelled. However, reservations that had more than 100 days on the waiting list were also cancelled at a slightly higher rate.

Data Visualization

What is the percentage distribution of required_car_parking_spaces?



Observation:
91.6 % guests does not required the parking space. only 8.3 % guests required parking space.

So lets talk about challenges that we faced

- One of the major challenge was to clean the datasets as a lot of scattered information was present especially in the 1st data set.
- Handling the error, duplicate and NaN values in the dataset.
- To draw meaningful insights, we had to design multiple visualizations like scatterplot, jointplot etc. without compromising the results and trends.

To overcome above challenges, we followed

- Alma Better class material
- Pandas and NumPy
- GeeksforGeeks
- Analytics Vidhya
- Stack Overflow

Conclusion

- City Hotels are most preferred by the customer.
- Maximum number of bookings were done in month of august so we can put more offers during this month.
- As compared to couples ,single and family with more than 3 members is less expensive.
- TA/TO distribution channel is more preferred by the customer as compared to others, In order to grow their business, hotels might partner with these agents and operators or promote using them as a medium.
- Most visitors of resort hotel stayed for one day, however most city hotel guests spent anywhere between one and seven days.
- the most popular meal type among guests is BB (Bed and Breakfast), followed by HB (Half Board) and SC (Self Catering).From the above graph it is clear that highest rate of returning customers are from the resort hotel
- From above it is clear that city hotels are mostly preferred by babies, adults and children
- More than 25000 people, or the majority of the attendees, are from Portugal. he average revenue of city hotels is higher than that of resort hotels.

- H type has the highest Average daily rate followed by G type
- January, February, March, April, October, November, and December are the ideal months for visitors to resort or city hotels because of the low average daily rate throughout these months.
- For hotels in cities, the majority of cancellations occurred in the month of October, but for hotels in resort areas, the majority occurred in the month of August.
- Additionally, similar cancellations of reservations for both hotel types occurred in the month of August. City hotels had greater cancellations of reservations overall.
- For hotels in cities, the majority of cancellations occurred in the month of October, but for hotels in resort areas, the majority occurred in the month of August.
- Additionally, similar cancellations of reservations for both hotel types occurred in the month of August. City hotels had greater cancellations of reservations overall.
- About 30% of hotel reservations for city hotels and 24% for resort hotels are cancelled.
- here we can say that PRT country made the highest number of cancelation under the city hotel
- here we understand that the PRT country has made the large number of cancelation in resort type hotel.
- There is no direct correlation between a longer waiting period and booking cancellation, as can be seen from the fact that the majority of reservations that had less
- than 100 days on the waiting list were cancelled. However, reservations that had more than 100 days on the waiting list were also cancelled at a slightly higher rate.
- 91.6 % guests does not required the parking space. only 8.3 % guests required parking space.
- As a result, the most popular meal type among guests is BB (Bed and Breakfast), followed by HB (Half Board) and SC (Self Catering).


ThankYou