Technical Documentation

EDA Hotel Booking

Capstone Project – I

Submitted to AlmaBetter

Submitted By: Abhijeet Kumar

Email : abhikr7252@gmail.com/abhijeet.kr7252@gmail.com

GitHub:
https://github.com/abhikr11/EDA_Hotel_Booking_Analysis_Project

Contents

1. Abstract 2
2. Introduction 2
3. Problem Statement 2
4. What is EDA? ..... 3 - 4
5. Data Dictionary ..... 4 - 5
6. Data wrangling and feature engineering ..... 6 - 7
7. EDA and Data Visualization ..... 8 – 24
8. Solution to Business Objective 25
9. Conclusion ..... 25 - 26
10. Reference 26

Abstract

Hotel industry has become a competitive market in today's scenario. Every hotel manager need to know the demand and behaviour of bookings by customers. Hotel needs to prepare themselves for upcoming challenges and a way to generate more revenue. We have done Exploratory Data Analysis (EDA) on Hotel Booking dataset to find the general trends of booking, cancellation, other attributes of customers using Python.

The dataset was examined properly and cleaned for further analysis. Different categories were examined for any statistical pattern or behaviour which can be beneficial to the managers for decision making.

KEY WORDS: EDA, Hotel Booking Analysis, Data Analysis, Statistics, Python, NumPy, Pandas, Matplotlib, Seaborn

Introduction

Hotel Booking Analysis is a common data analysis task in the hospitality industry. The aim of this task is to extract meaningful insights from hotel booking data to make data-driven decisions. Exploratory Data Analysis (EDA) is a crucial step in this process as it helps in understanding the data, detecting patterns, and identifying relationships between variables. In this document, we will provide a technical overview of the EDA process for Hotel Booking Analysis.

Problem Statement

This data set contains booking information for a city hotel and a resort hotel, and 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. This hotel booking dataset can help to explore and analyse general trends in hotel bookings.

The Hotel Booking Analysis project aimed to conduct an exploratory data analysis(EDA) on hotel booking data to identify patterns and insights that could help hotel managers make informed decisions

What is Exploratory Data Analysis (EDA)?

Exploratory data analysis (EDA) is a method used to analyse and summarize a dataset in order to understand its characteristics and patterns. EDA can be used to clean and pre-process the data, as well as identify any outliers or anomalies that may be present. Some common techniques used in EDA include visualizing the data using graphs and plots, calculating summary statistics, and identifying correlations and relationships between variables.

The steps involved in EDA are as follows:

1. Data Collection

The first step in any data analysis project is to collect the data. In this case, we need to collect hotel booking data. The data obtained for two types of hotels City Hotels and Resort Hotels. Once the data is collected, it needs to be cleaned and pre-processed before we can start the EDA process.

2. Data Cleaning and Pre-processing

Data cleaning and pre-processing are important steps in the EDA process as they help in identifying and fixing errors, inconsistencies, and missing data. In this step, we need to remove duplicate records, handle missing values, and correct errors. We also need to standardize the data and convert it into a format that can be easily analyzed.

3. Exploratory Data Analysis

Once the data is cleaned and pre-processed, we can start the EDA process. The EDA process involves the following steps:

- **Univariate Analysis:** In this step, we analyse each variable in the dataset individually. We use descriptive statistics such as mean, median, mode, standard deviation, and range to summarize the data. We also use visualization techniques such as histograms, box plots, and density plots to understand the distribution of the data.

- **Bivariate Analysis:** In this step, we analyse the relationship between two variables. We use correlation analysis to measure the strength of the relationship between two variables. We also use scatter plots, line charts, and heatmaps to visualize the relationship between two variables.

- **Multivariate Analysis:** In this step, we analyse the relationship between multiple variables. We use techniques such as clustering, factor analysis, and principal component analysis to identify patterns and relationships between variables.

4. Data Visualization

Creating visual representations of the data to better understand and communicate the findings. This step can include creating charts to help convey the key insights of the analysis.

5. Conclusion

EDA is an important step in Hotel Booking Analysis as it helps in understanding the data and extracting meaningful insights. In this document, we provided a technical overview of the EDA process for Hotel Booking Analysis. The EDA process involves data collection, data cleaning and pre-processing, and exploratory data analysis. By following these steps, we can identify patterns, trends, and relationships in the data and make data-driven decisions.

Data Dictionary

Field	Description
Hotel	Type of hotel (City Hotel and Resort Hotel)
is_cancelled	(1) If the booking was cancelled or (0) for not cancelled
lead_time	time (in days) between the date of booking and the actual arrival
arrival_date_year	Year of arrival date
arrival_date_month	Month of arrival date
arrival_date_week_number	Week number of arrival date
arrival_date_day_of_month	Day of arrival date
stays_in_weekend_nights	Number of weekend nights(Saturday or Sunday) spent by guest in hotel
stays_in_week_nights	Number of week nights(Monday to Friday) spent by guest in hotel
adults	Number of adults
children	Number of children
babies	Number of babies

meal	Type of meal opted by guest
country	Country code of guest
market_segment	Which segment the customer belongs to
distribution_channel	from which channel customer accessed the stay - corporate booking/Direct/T.A.TO
is_repeated_guest	If the guest coming for first time or not (0 for first time and 1 for repeated)
previous_cancellations	Total number of previous cancelled bookings
previous_bookings_not_canceled	Total number of previous non-cancelled bookings
reserved_room_type	Type of room reserved by the customer
assigned_room_type	Type of room assigned to the customer
booking_changes	Total changes made to booking
deposit_type	Type of deposit(No deposit/ Refundable/ Not refundable)
agent	Booking ID of agent
company	Booking ID of company
days_in_waiting_list	Number of days in waiting list
customer_type	Type of customer(Transient, Group,Contract, Transient-Party)
adr	Average daily rate
required_car_parking_spaces	Total number of car parking required by customer
total_of_special_requests	Total number of additional special requirements
reservation_status	Status of reservation, if customer (Check-Out, Cancelled, No-Show)
reservation_status_date	Date of specific reservation status

Data Wrangling and Feature Engineering

1. Know the Data:

- First step is loading data in colab notebook.
- Check the schema of dataset using head() function. It will show the first 5 rows of data set.
- Check the total number of rows and columns using shape method. This dataset has total 119390 rows and 32 columns.
- Check info
- Check description of data using describe().
- Check for the columns
- Check for null values in dataset using isna().
- Check number of unique values for each columns

2. Remove Duplicate Rows:

- Check for duplicate rows using duplicated() method.
- Total 31994 duplicate rows were found.
- All duplicate rows were dropped using drop_duplicates() method.
- After dropping duplicate rows dataset had 87396 rows left.

3. Handling Null Values:

- There were total of 4 columns where null values were present.
- Company had 112593, agent had 16340, country had 488 and children had 4 null values.
- Company and agent contained ID's so it was filled with 0 using fillna().
- Country column was filled with "OTHER".
- Children column was filled with mean value.

4. Convert to Appropriate DataType:

- Check the data type of variables using dtypes function.
- Data type of children, agent and company was changed from float to int.
- Data type of reservation_status_date was changed from object to datetime.

5. Adding New Columns:

- New column was added for total guests, which is the summation of adults, children and babies.
- New column was added for total stay, which is summation of stays in weekend nights and stays in week nights.
- All rows which had no guest were dropped.

6. Drop Outlier:

- Box plot was plotted for ADR distribution for each month.
- Outlier was detected in February month.
- Outlier was dropped.

A copy of this dataset was made using copy() method.

A separate dataset was prepared for the bookings that were not cancelled and stored in variable not_canceled_df.

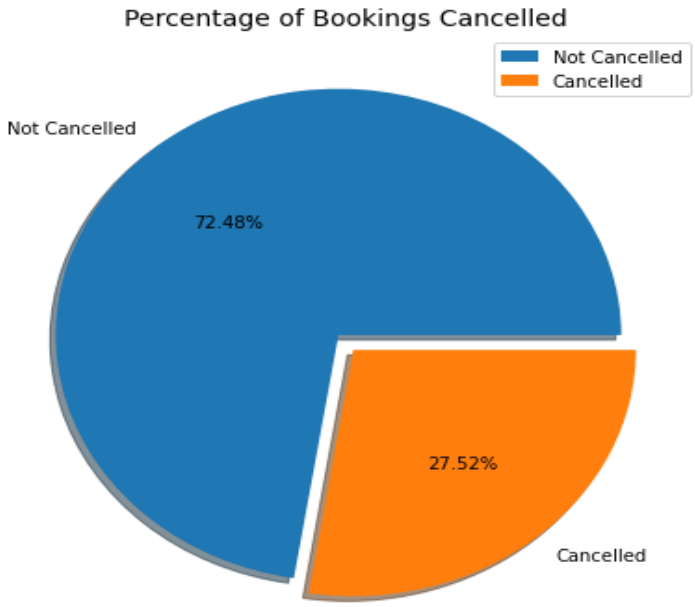
Our data is cleaned properly and all the manipulations were done, now it's time for final step i.e., exploratory data analysis, visualization and storytelling.

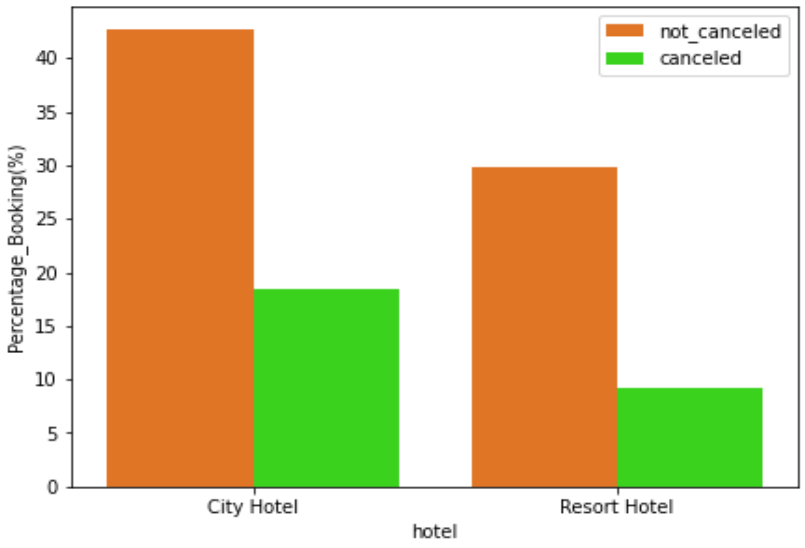
EDA and Data Visualization

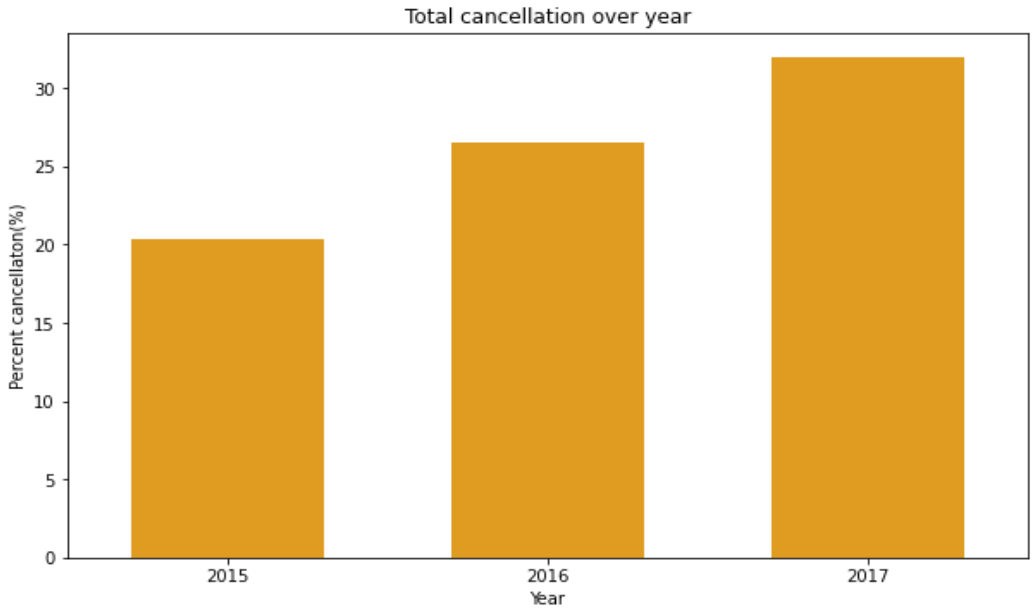
Exploratory Data Analysis (EDA) is an important step in the Data Analysis project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. We have used Python libraries like NumPy, Pandas, Matplotlib and Seaborn for data wrangling and visualization of various aspect of dataset through informative charts.

1. What is the percentage of bookings that got cancelled?

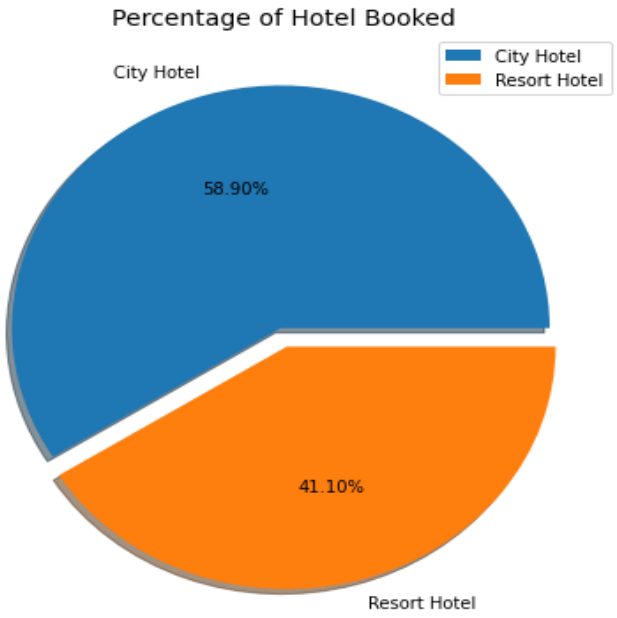






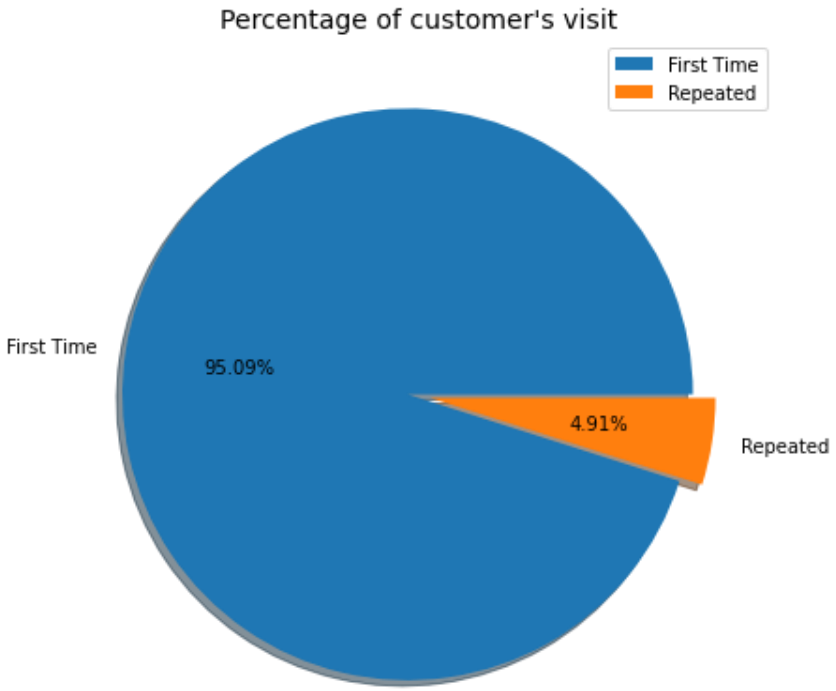
- 27.5% of the bookings were cancelled by the customers from which about 18% cancellation was from City Hotel and 9% from Resort Hotel.
- The highest booking was for City Hotels about 43% and about 30% for the Resort Hotels.
- Total cancellation rate increases over the year, it was lowest in 2015 and increases year after year.

2. Which hotel is most prefered by the guests?



- Most of the guests prefered City hotels over Resort hotels.

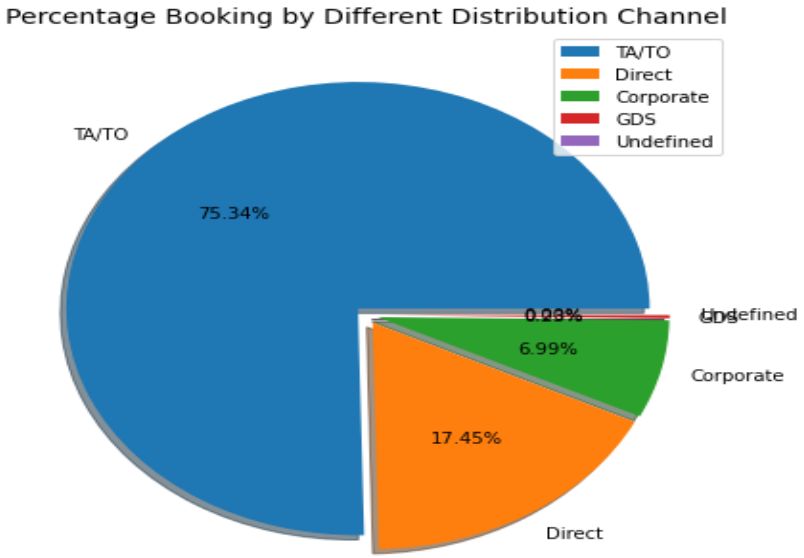
3. What is the percentage of repeated guest?

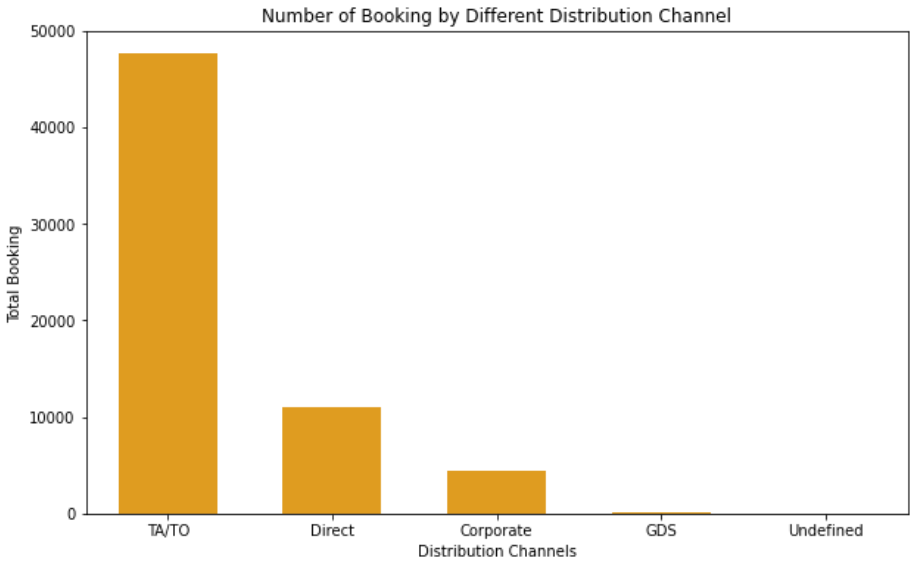


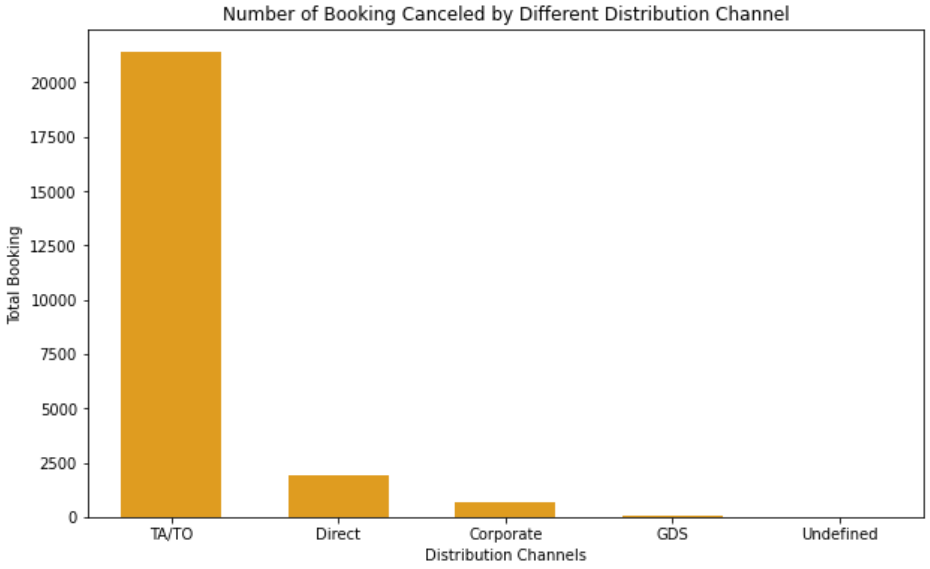
- Total of approx 5% customers are returning to same hotel after their first visit.

- The returning percentage of customer for Resort Hotels are higher than that of City Hotels. About 4% customers return to City Hotel and 6.3% customers return to Resort Hotel.

4. Which distribution channel was mostly used for booking? Which channel has most cancellation?



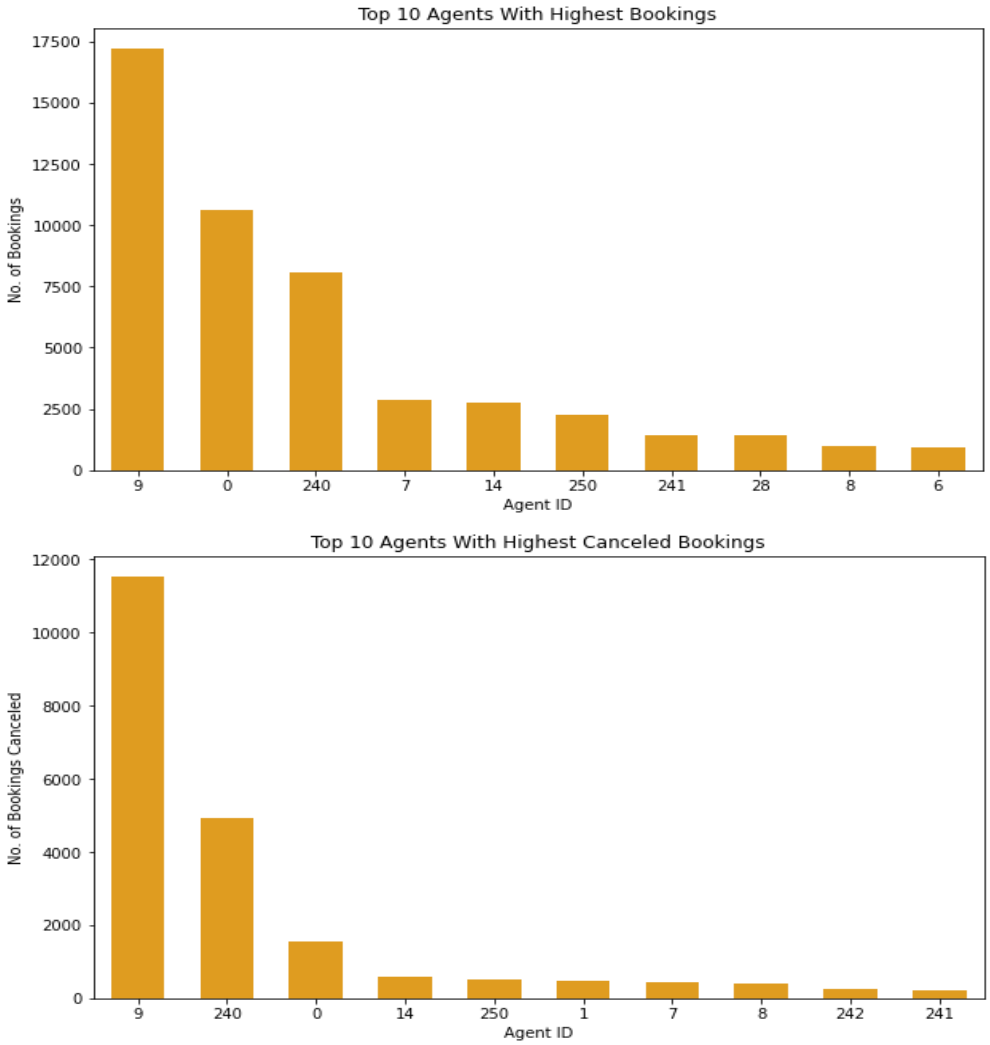




- Most of the bookings are done by TA/TO for about 75%. Also most of the cancellation are from TA/TO.

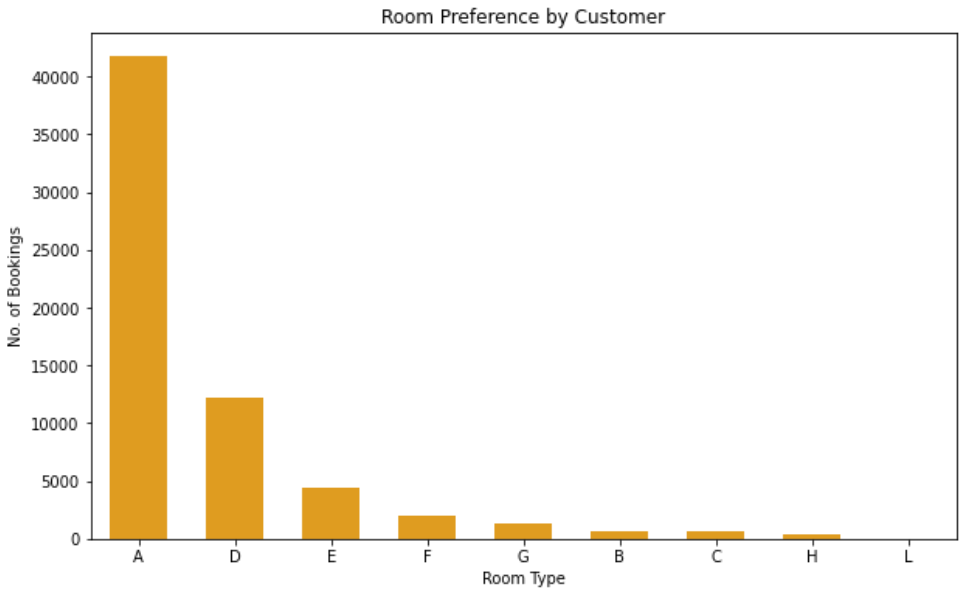
- The booking done directly by customer is 11031 which is 17.5% but the cancellation rate is lower (only 1923 bookings were cancelled) than that of TA/TO.

5. Which agent has highest booking and highest cancellation?



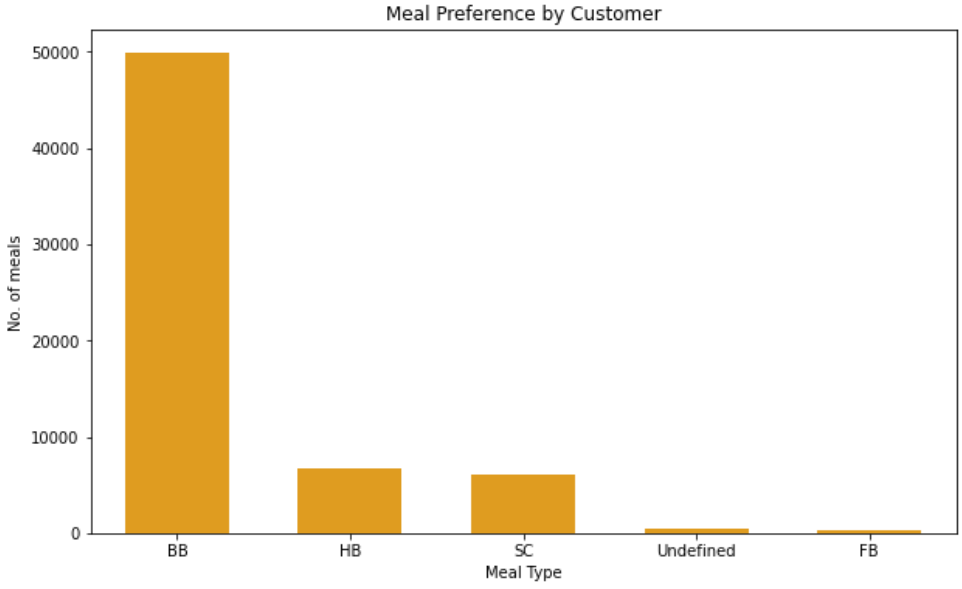
- Agent 9 has most number of bookings but also has highest number of cancellation.
- Second most booking is from anonymous (Agent ID = 0), as we have filled 0 in place of null values.
- Agent has 8084 not cancelled bookings but 4944 cancelled bookings.
- All other agents have less number of bookings.

6. What is the most preferred room type by the customers?



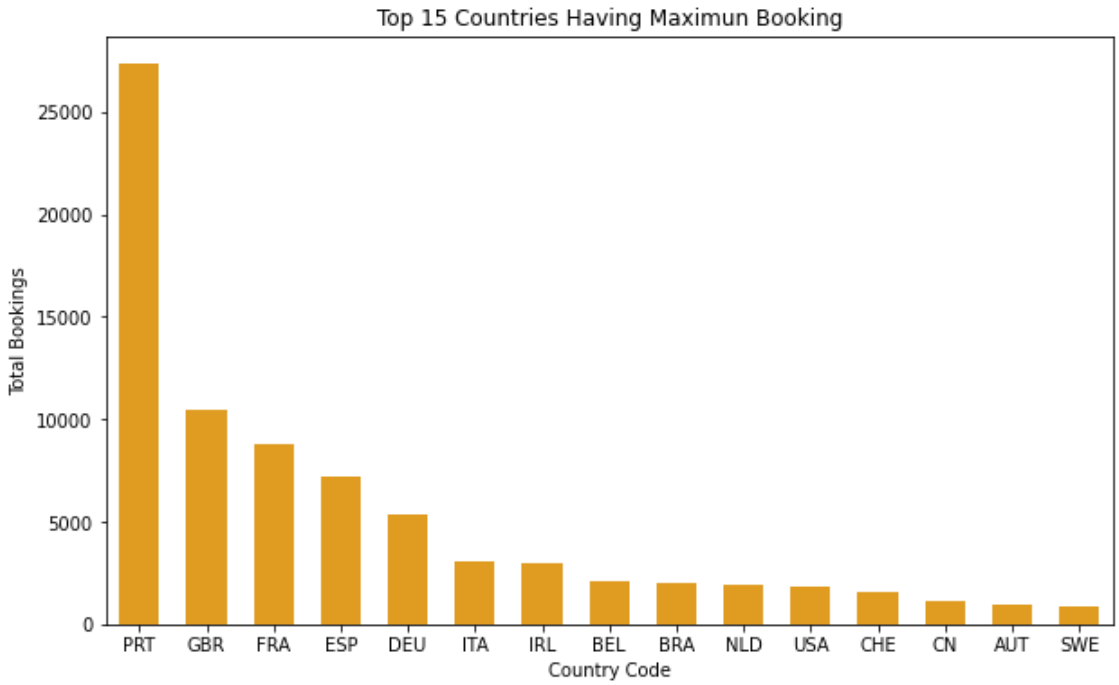
- The most preferred room type is A having more than 41000 bookings, second preferred room type is D having 12000 bookings, than E and F.
- Least preferred room type is L.

7. What is the most preferred meal type?



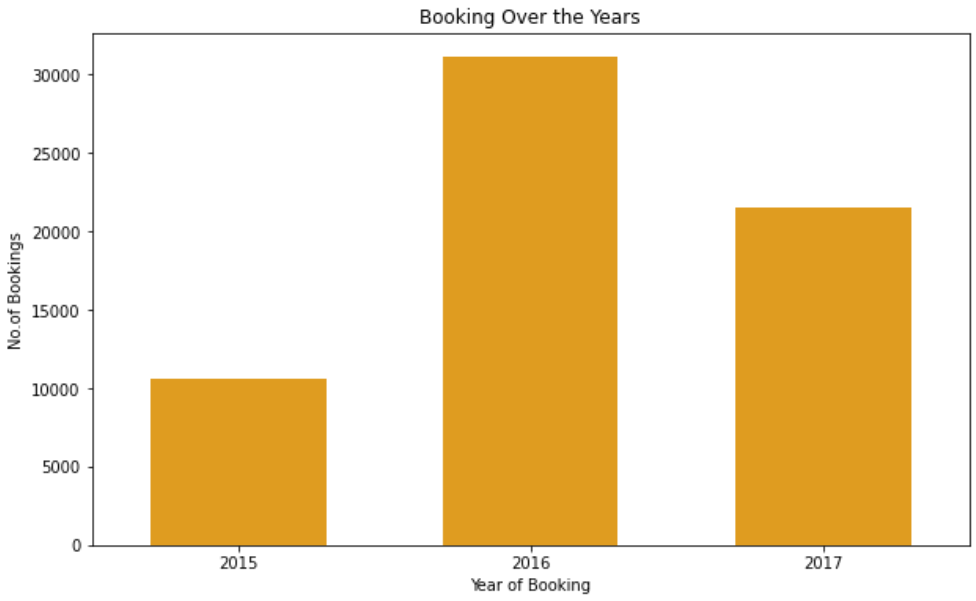
- BB meal type is most preferred by customers and FB meal type is least preferred.

8. Which country do most customers belong to?



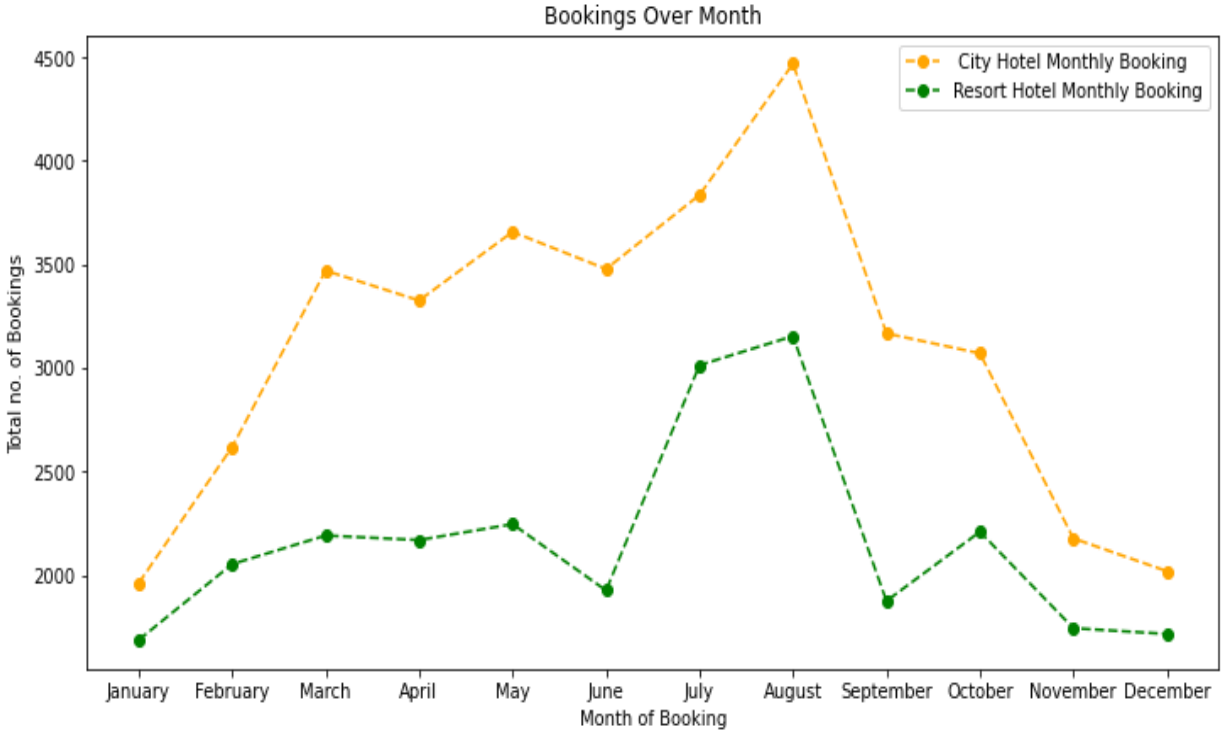
- Portugal has highest number of bookings followed by United Kingdom, France, Spain and Germany.

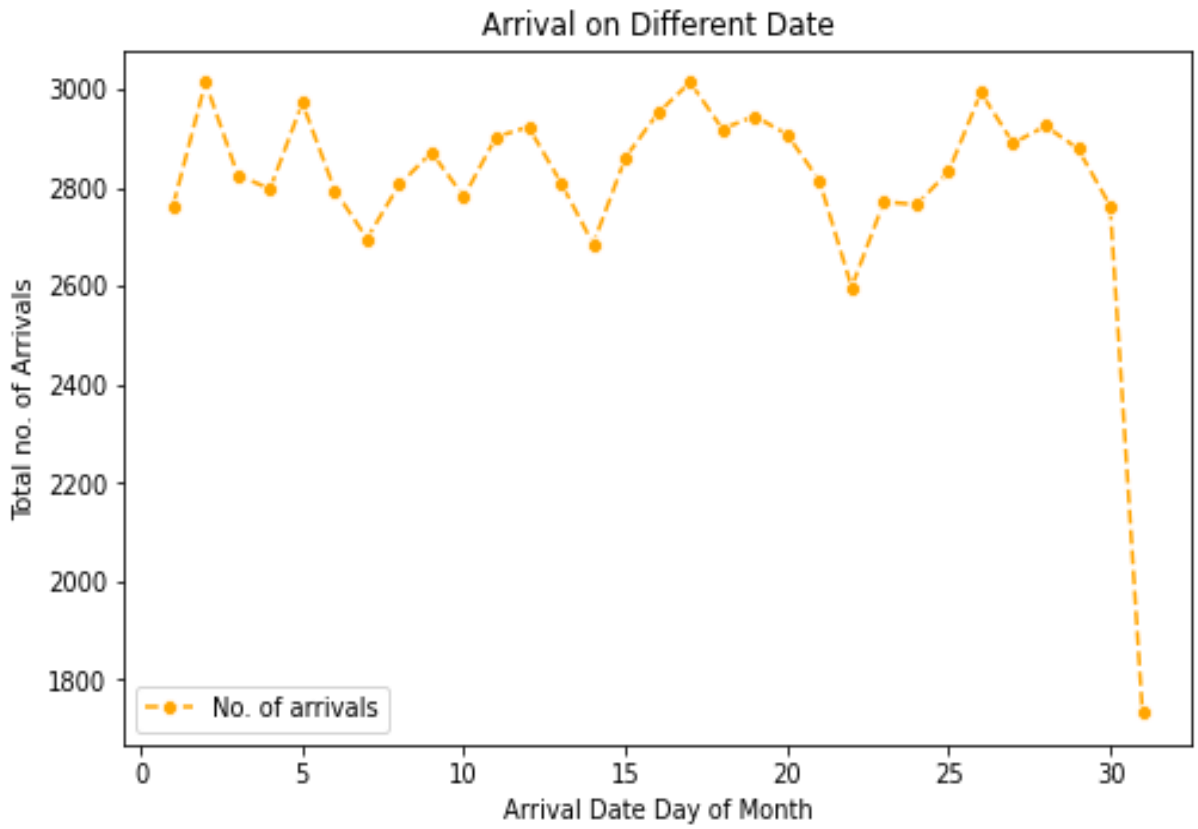
9. Which year had the highest confirmed bookings?



- 2016 had the highest no. of bookings and 2015 had lowest bookings.
- No. of bookings increased in 2016, then it decreased next year in 2017.

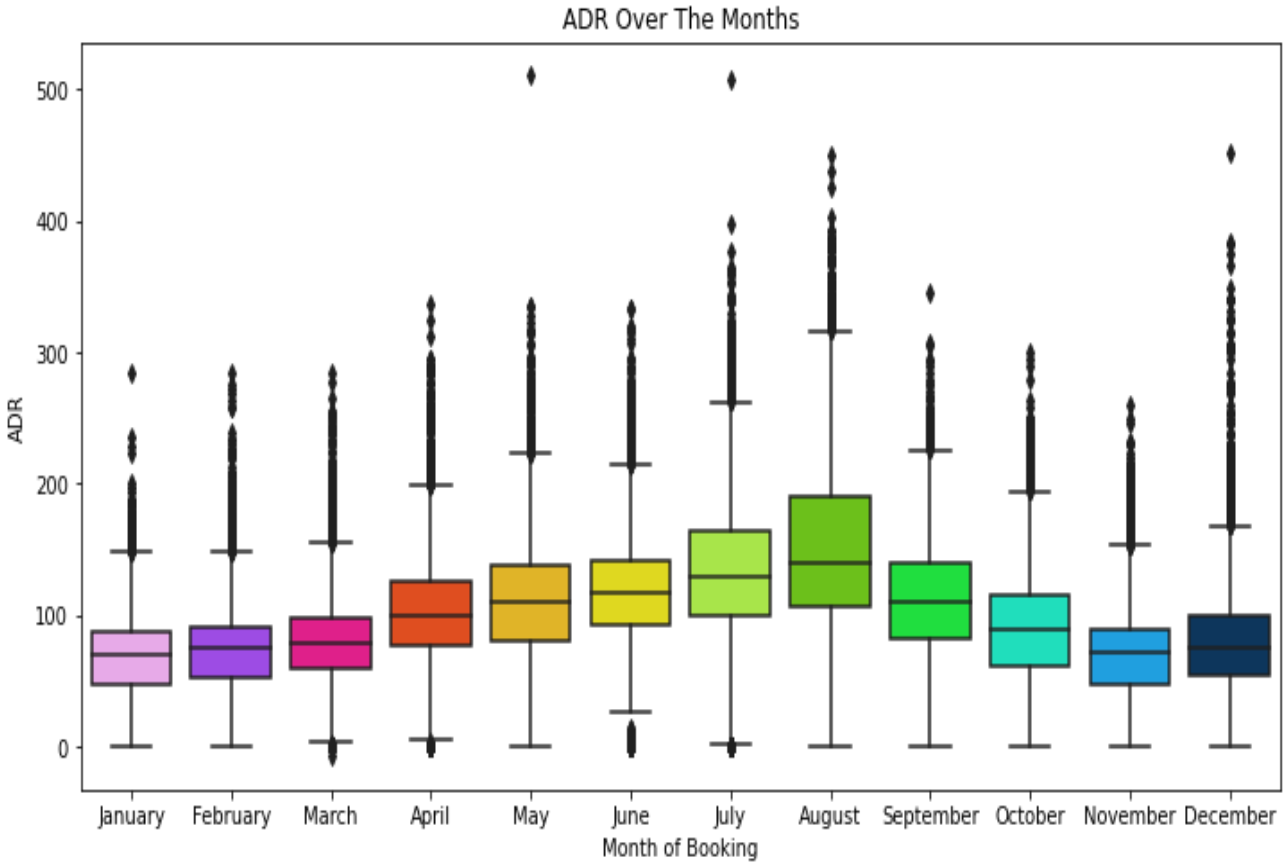
10. Which month likely to have more rush?





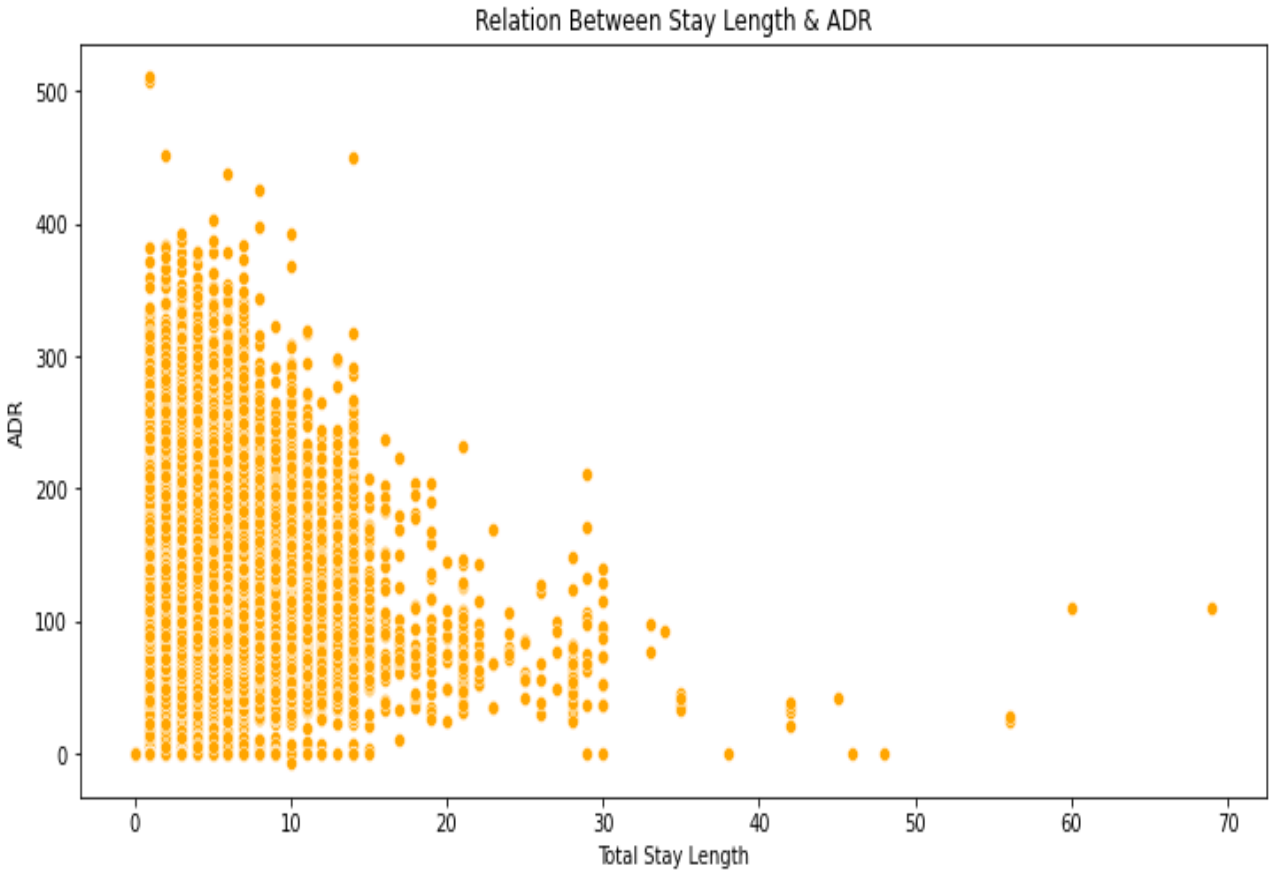
- The no. of bookings at the starting and end of the year is quite low but it has highest in the middle of the year.
- July and August are the busiest month for both the hotels.
- From date vs total arrival graph, we can see there is jump in total arrival at regular interval of 5-6 days. So, we can conclude that these jumps can over weekends. There will be more rush on weekends.

11. What is the distribution of ADR over months?



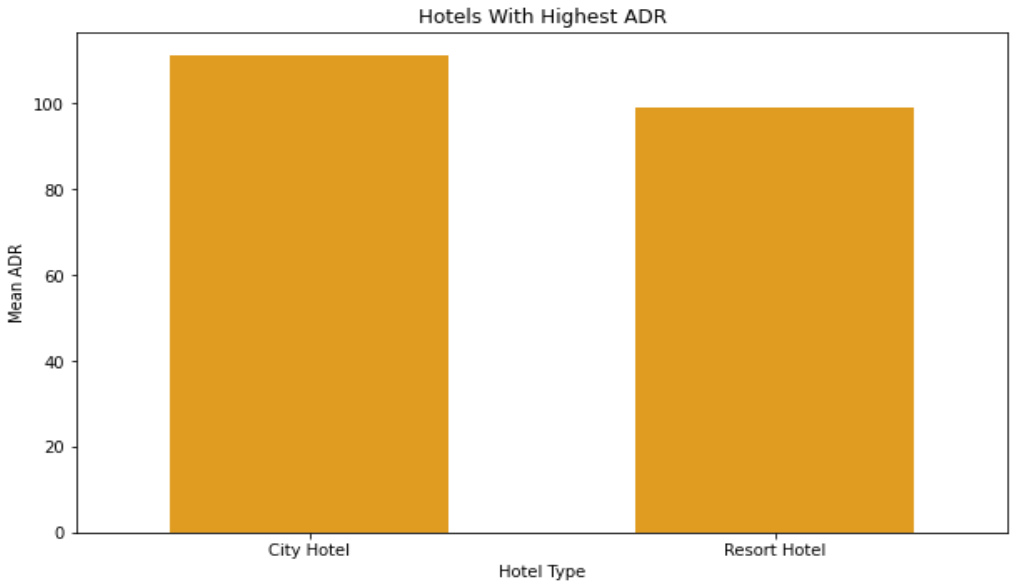
- Average ADR increases till August and then lowers to the end of year. Hotels made good profit deals with high adr at end of year in December.
- Average ADR is lowest at the starting and end of month, it gradually increased till middle of the year but then decreased.

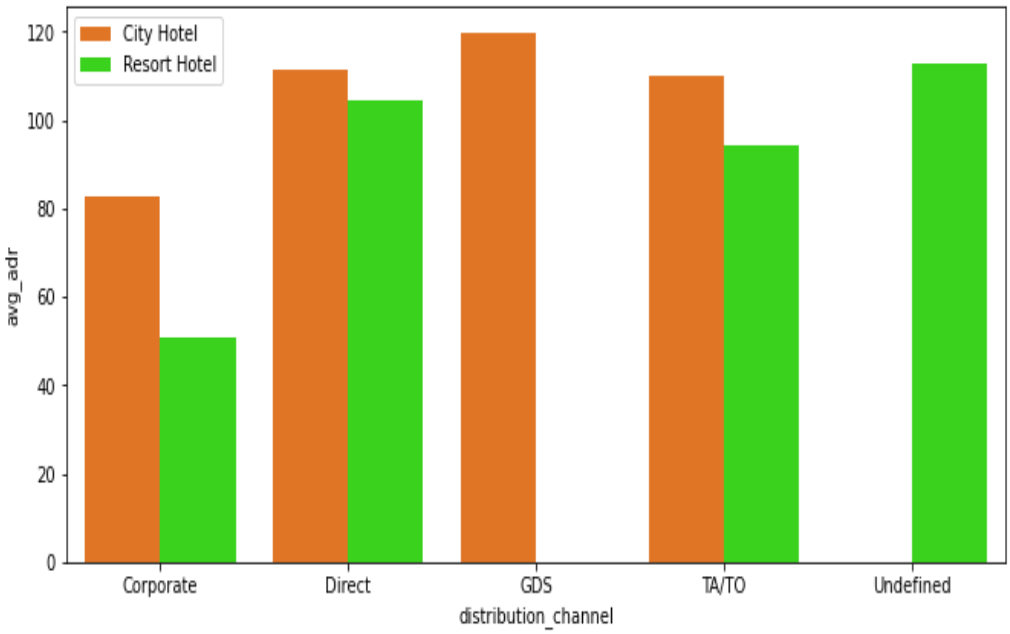
12. Does the length of stay affect ADR?



- For the longer duration of stay, the ADR decreases.
- For smaller duration of stay ADR can be as high as 500.
- Customers can get better deals on longer stay which will increase the revenue.

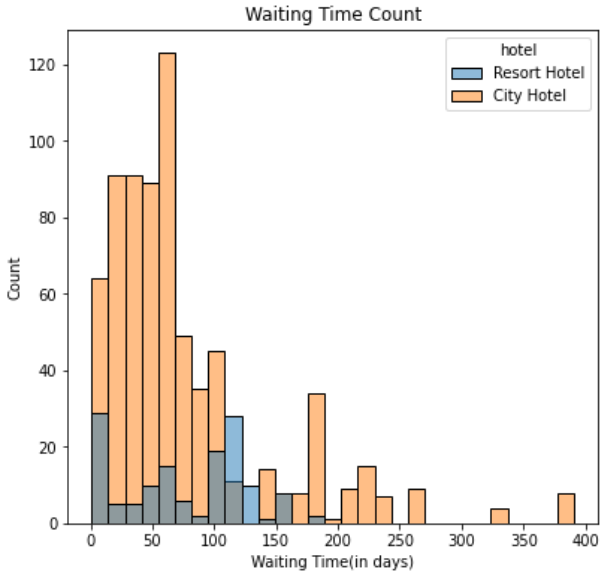
13. Which hotel and distribution channel has highest mean ADR?

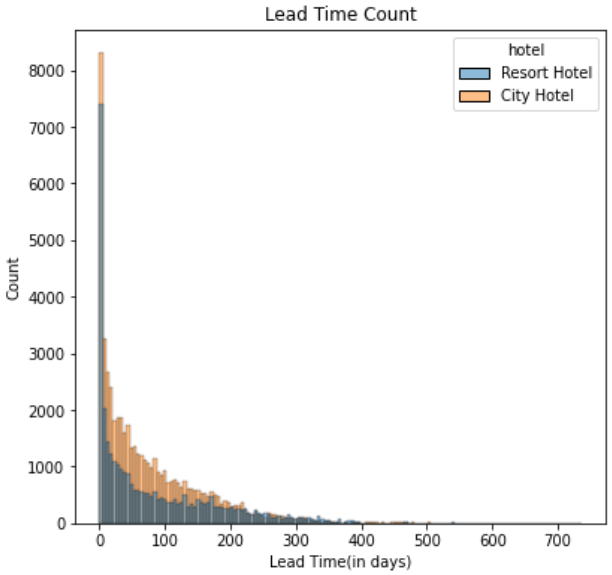




- City Hotel has more average ADR than Resort Hotels.
- For City Hotels direct, GDS and TA/TO have higher ADR.
- For Resort Hotels direct, TA/TO, undefine channels have higher ADR.

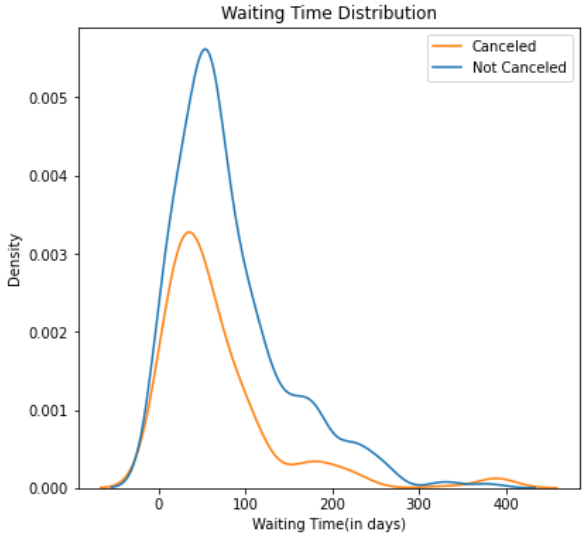
14. Which hotel has highest waiting and leading time?

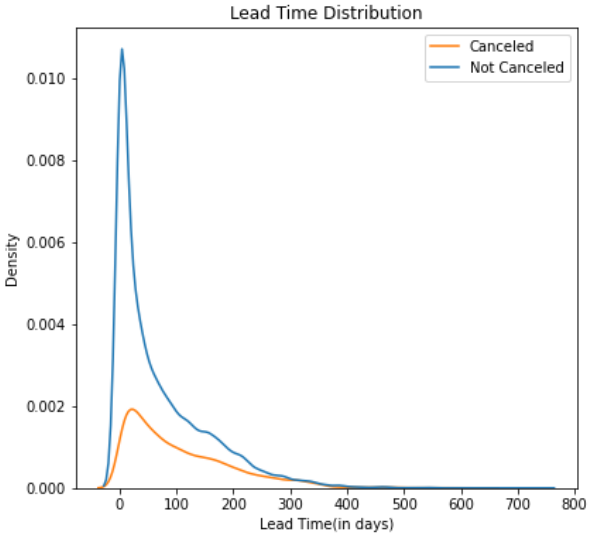




- For most of the City Hotels, the waiting time is usually less than 100 days but for Resort Hotels waiting time can be more, even more than 100 days.
- The lead time for city hotel is high which mean City hotels are booked in advance.

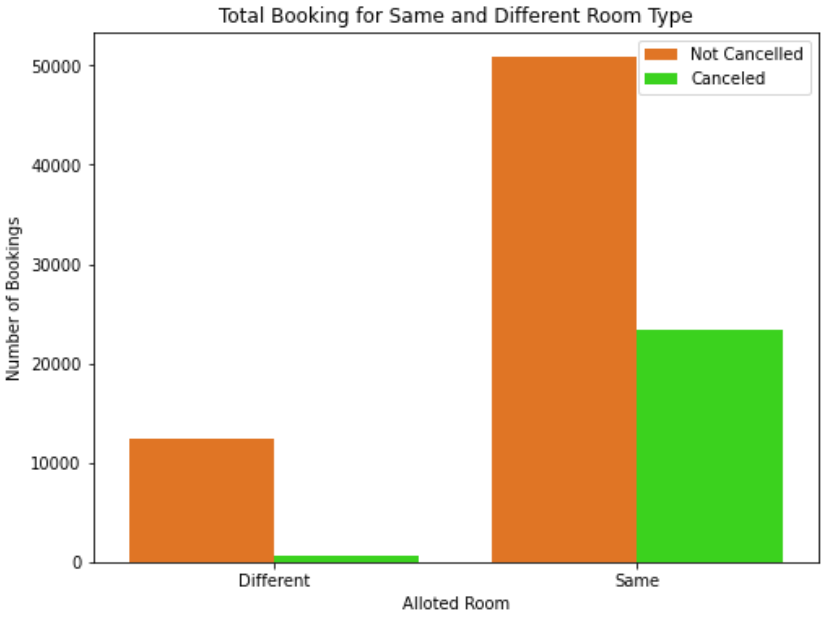
15. Does the longer waiting time or longer lead time is the reason for cancellation of bookings?

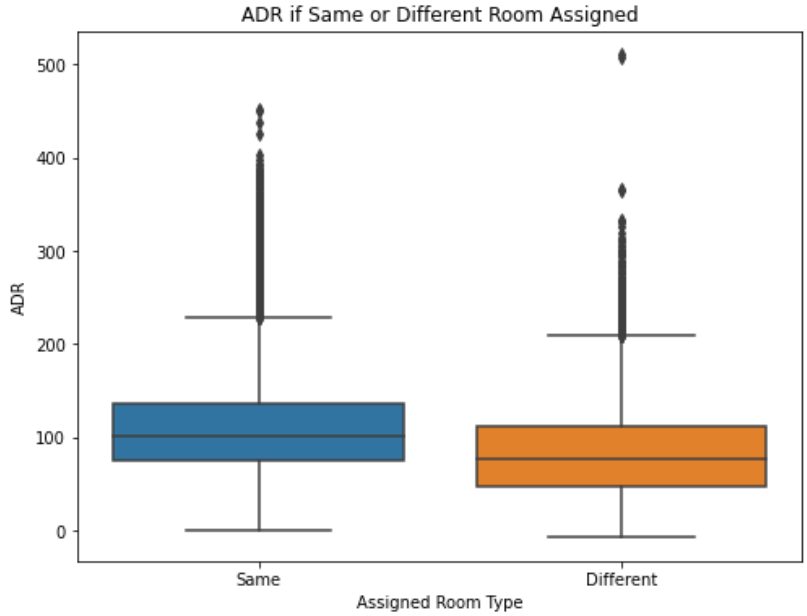




- Most of the bookings were cancelled which had waiting time of less than 150 days but for the bookings which were not cancelled also had waiting time of less than 150 days. So, waiting time doesn't have significant role for cancellation of bookings.
- Same is the case for lead time for both maxima cancelled and maximum not cancelled is for lead time less than 100 days.

16. If not allotting the same room type as reserved, results in cancellation or affects ADR?



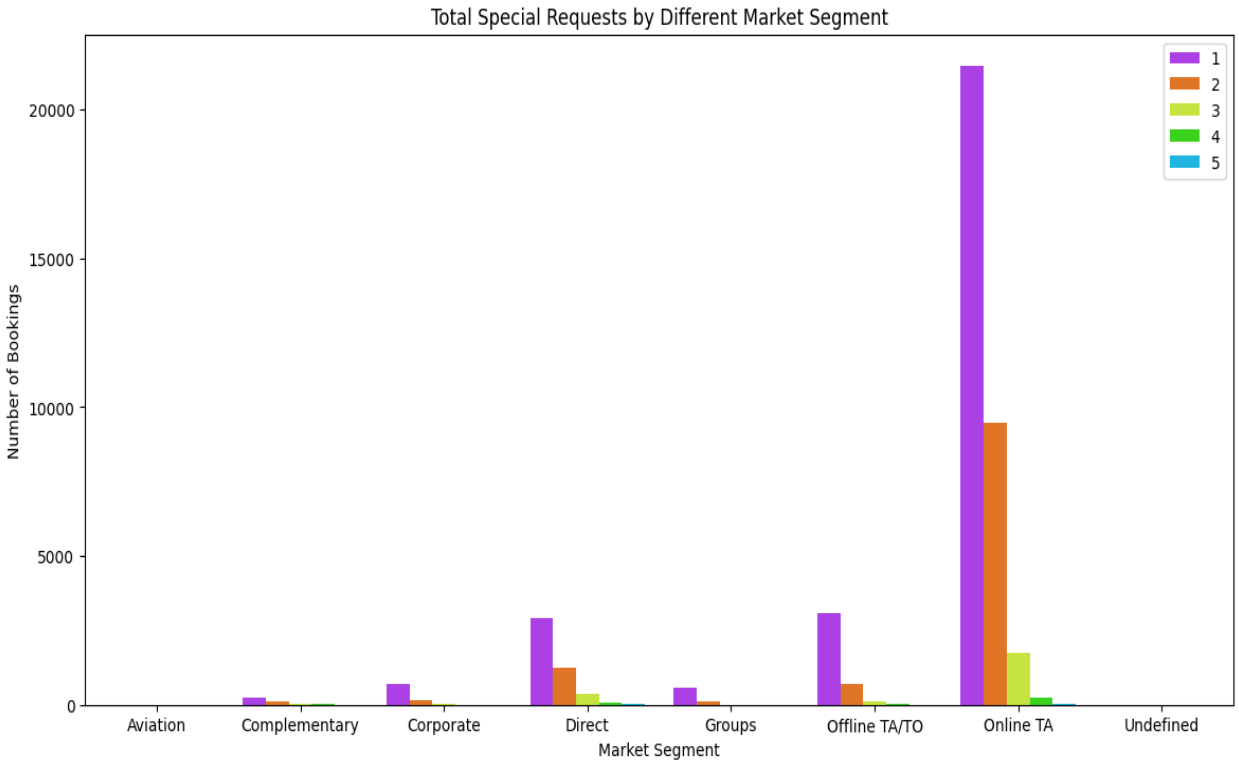


- Allocating different room is not the cause for cancellation, as very few customers (616) had cancelled the booking even after getting different room.

- A significant number (23392) of customers had cancelled the booking even after getting same room as they wanted.

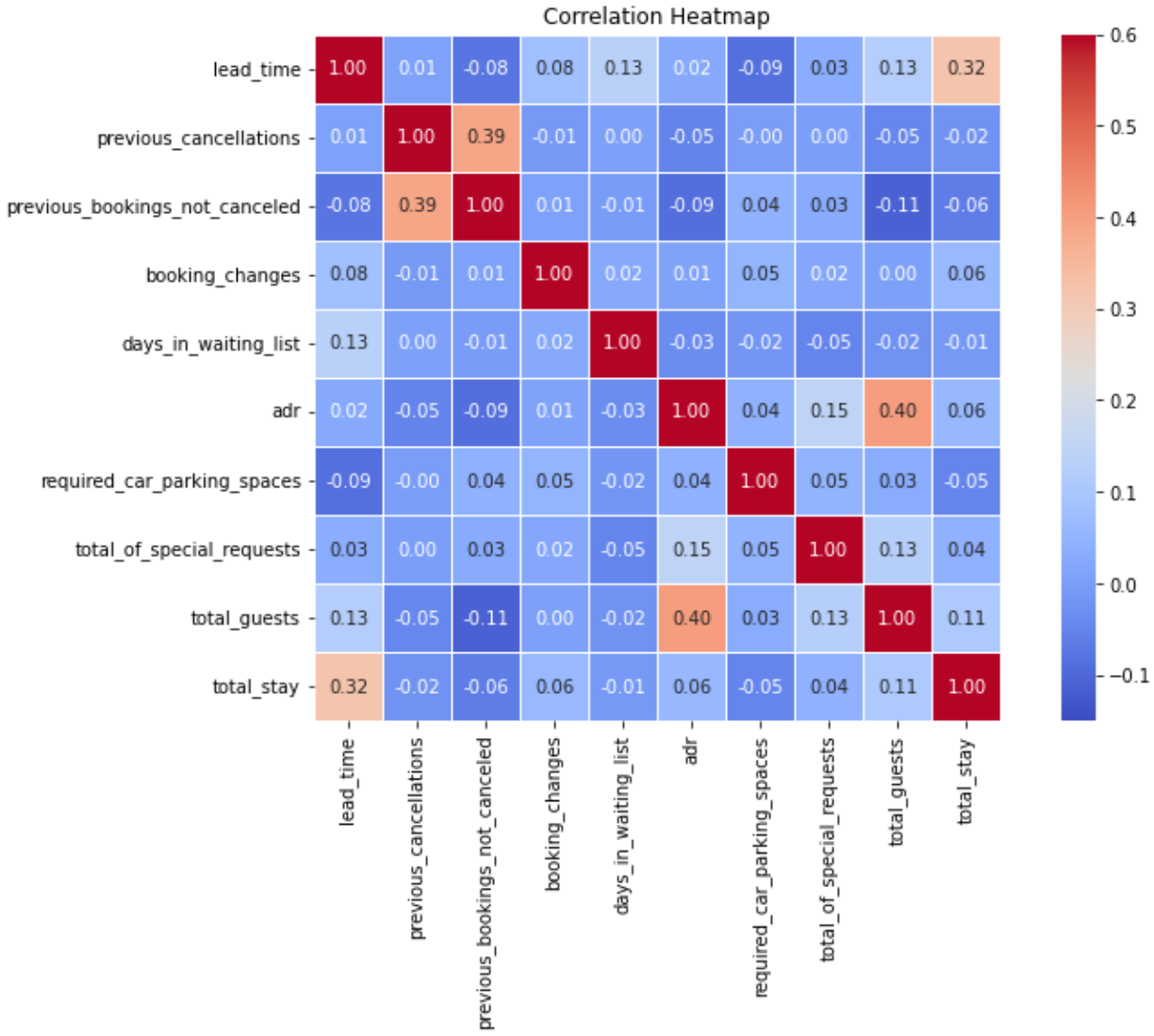
- Getting different room does affect the ADR, as customer tends to pay less if they get different room type.

17. Which market segment has more special requests?



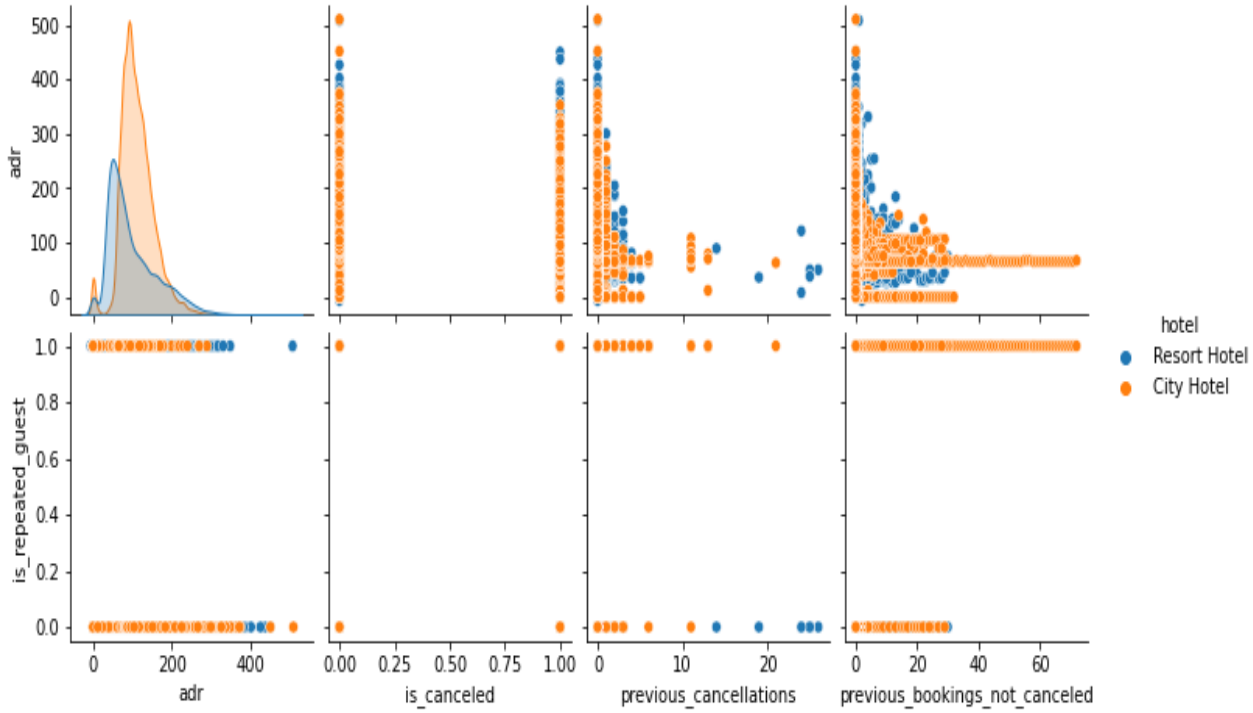
- Online TA made more number of special requests.
- Most of the Online TA customers asked for at least one special service.
- Online TA has huge number of special requests which can have negative business impact.

18. Correlation Heatmap



- Total Guests is positively correlated to ADR; this means as number of customer increases ADR increases.
- Total Stay is positively correlated to Lead Time, this means for longer stay customers booked the hotel in advance.
- Total Guests and Previous Bookings not cancelled is negatively correlated, this means if there are more number of guests then the chance of cancellation is really low.

19. Pair Plot



- Guests who have more number of previous cancellations tends to pay low ADR, it means people usually cancels if they get better deals with low ADR.
- For the guests who don't often cancels previous bookings pays low ADR.
- First time visitor has more number of previous cancellations and pays more (high ADR).
- Repeated guests have more number of previous booking not cancelled and pays low (low ADR).

Solution to Business Objective

- Ask guests for feedback, it will show reason for increasing cancellation rate over years.
- Increase service quality for more repeated guests.
- Give more exposure to TA/To and GDS channel, they have most bookings and ADR.
- Increase A type room and give good quality Bed Breakfast.
- Increase staff for June, July and August month as these months have high rush.
- Try to allot same room type as reserved by customer.
- Give early bird offers for advance booking for longer stay.
- New customers prone to more cancellations so define roadmap for new customers.
- Analyse cancellation when happens.

Conclusion

- 60% guests preferred City Hotels but Resort Hotels have more percentage of repeated guests.
- Percentage of cancellation increased from 20% in 2015, 26% in 2016 and 32% in 2017.
- Longer waiting and lead time is not the reason for cancellation.
- Getting different room type then allotted doesn't lead to cancellation but guests paid less if they were allotted different room.
- If there is more number of total guests, then chance of cancellation is low.
- Most of the guests cancelled booking if they get better deal elsewhere with low ADR.
- Only 5% guests re-booked Hotels.
- TA/TO channel and agent 9 has most bookings.

- Preferred room types are A, D, E and least preferred room types are B, C, H, L.
- Most preferred meal type is BB.
- Larger number of customers are from Portugal, UK, France, Spain and Germany.
- July and August tends to have high traffic and hotels can usually expect rush on weekends than weekdays.
- ADR increased till middle of the year (till August) then decreased. For short stay ADR is usually high but lower for longer stay.
- Online TA segments have high special requests.
- Guests booked earlier in advance for longer stay.
- Old guests paid less ADR.

Reference

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

Hands-On Exploratory Data Analysis with Python, by Suresh Kumar Mukhiya, Usman Ahmed.

Hotel Booking data was collected from AlmaBetter.