Capstone Project - 1

Hotel Booking Analysis

Team Members

Samarjeet singh chhabra
Priyadarshani Gaikwad
Devarshi Dwivedi
Jay Pardeshi
Mohd. Anas Ansari

Table Of Contents

Problem Statement

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests?

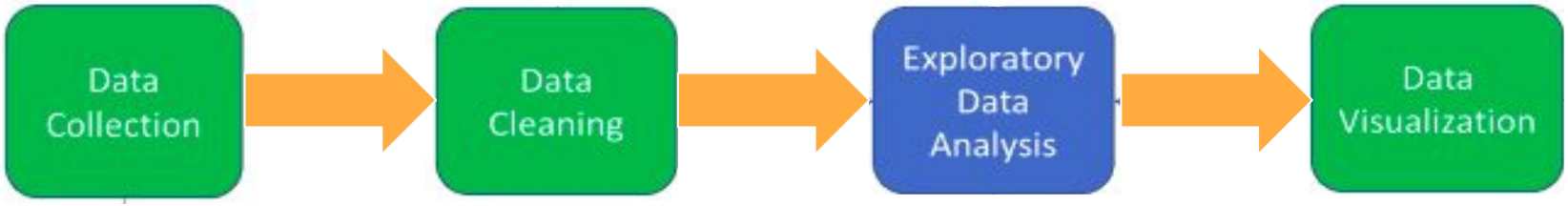
This hotel booking dataset can help you explore those questions!
This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

All personally identifying information has been removed from the data. Explore and analyze the data to discover important factors that govern the bookings.

Overview of the given data and problem

- We are provided with hotel bookings dataset of the following years – 2015 to 2017
- This dataset is unstructured, contains a lot of null values and needs cleansing.
- Other than that , there are going to be certain data columns that we won't be needing so filtering is required.
- After proper Filtering and cleansing, We are going to analyse this dataset and try to gain insight and analyse factors that govern these bookings.
- We will be using some libraries such as Numpy, Pandas and Matplotlib for different task such as managing arrays, working on dataframes and visualizing data..
- We will be using data visualization to depict everything graphically.

Steps Followed In Analysis



Data collection : We collected the hotel booking data on which EDA is to be done. We then understood the data, its columns/features and its content.

Data cleaning : We cleaned the data by dropping or replacing null values, deleting unwanted columns, checking data type and conversion to a data type of required column and we performed many other operations to get the required dataset

Steps Followed In Analysis

-Continued

EDA will be divided into following 3 analysis:

- **Univariate Analysis**: Univariate analysis is the simplest of the three analysis where the data you are analyzing is only having one variable.
- **Bivariate analysis**: In Bivariate analysis we will compare two variables to study their relationships.
- **Multivariate analysis**: Multivariate analysis is similar to Bivariate analysis here we will compare more than two variables.

Data visualization: We represented the insights through data visualization with the help of different types of graphs and charts.

Understanding The Dataset Provided

The data has 119390 rows and 32 columns or features. Now let's understand what these columns have.

All columns heading and data description:

- **hotel** : Hotel type.
- **is_canceled** : booking is canceled or not (0 & 1).
- **lead_time** : advance booking time
- **arrival_date_year** : guests arrival year.
- **arrival_date_month** : guests arrival month.
- **arrival_date_week_number** : guests arrival week.
- **arrival_date_day_of_month** : guests arrival day.
- **stays_in_weekend_nights** : weekend nights bookings
- **stays_in_week_nights** : weeknights bookings
- **adults** : Number of adults.
- **children** : number of children.
- **babies** : Number of babies.
- **meal** : Type of meals
- **country** : Country of origin

Understanding The Dataset Provided -Continued

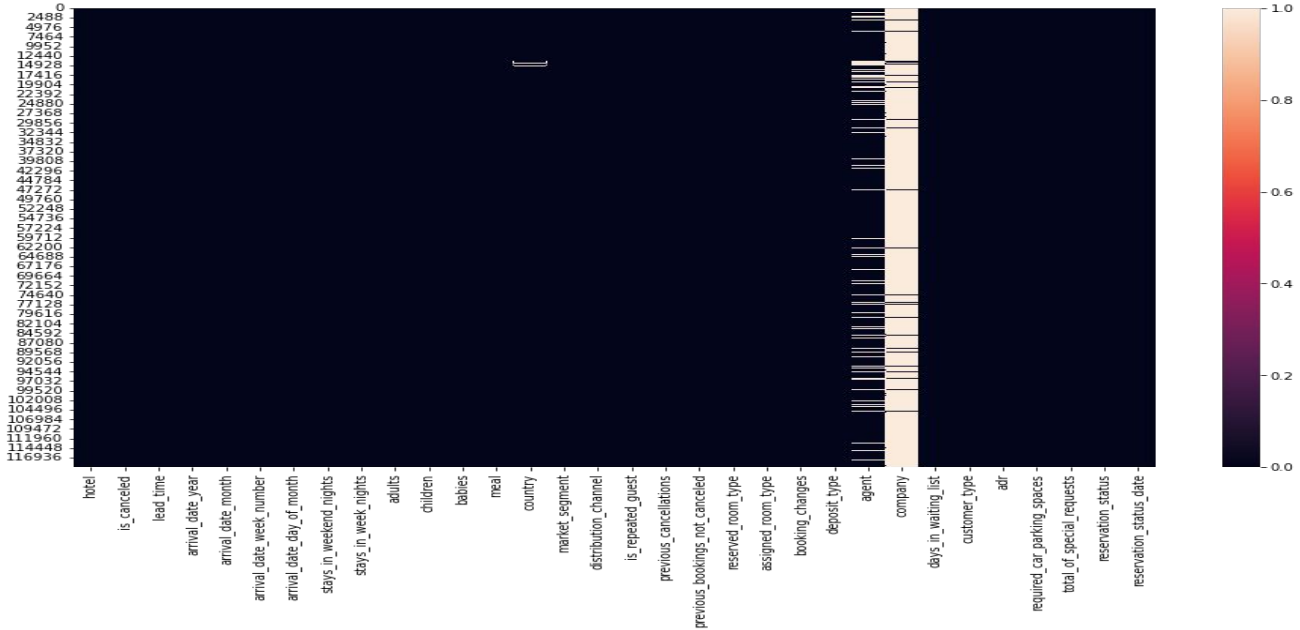
- **market_segment** : where the bookings came from.
- **distribution_channel** : Booking distribution channel.
- **is_repeated_guest** : repeated guest (1) yes or not (0).
- **previous_cancellations** : previous bookings that were cancelled
- **previous_bookings_not_canceled** : previous bookings that were not cancelled
- **reserved_room_type** : Code of room type reserved.
- **assigned_room_type** : Code for the type of room assigned to the booking
- **booking_changes** : Number of changes/amendments made to the booking
- **deposit_type** : Indication on if the customer made a deposit to guarantee the booking.
- **agent :** ID of the travel agency that made the booking.
- **company** : ID of the company/entity that made the booking
- **days_in_waiting_list** : Number of days the booking was in the waiting list
- **customer_type** : Type of booking, assuming one of four categories.
- **adr** : Average Daily Rate
- **required_car_parking_spaces** : Number of car parking spaces required to customer.
- **total_of_special_requests** : Number of special requests
- **reservation_status** : Reservation last status
- **reservation_status_date** : Date at which the last status was set.

Data Overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
```

We took a overview of the data together by using many methods such as **.head() , .tail() , .describe(), shape , .info()** and **etc.**

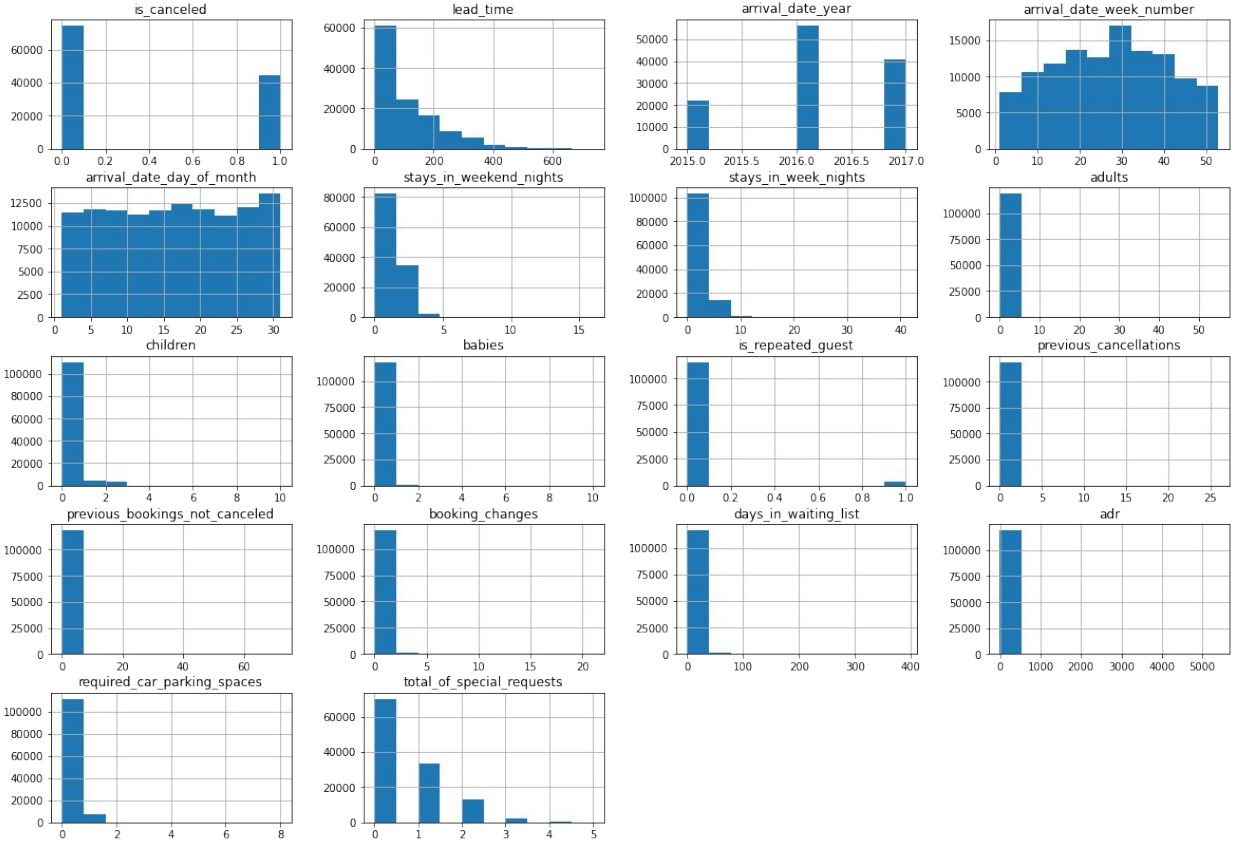
Data Cleaning



Found null values in following columns and took actions accordingly

1. **Children** - replaced all missing 4 values with **0(int64)**.
2. **Country** - replaced all the missing values with “not mentioned”.
3. **Company** - Deleted the column as it was not useful.
4. **Agent**- Deleted the column as it was not useful.

EDA On Dataset



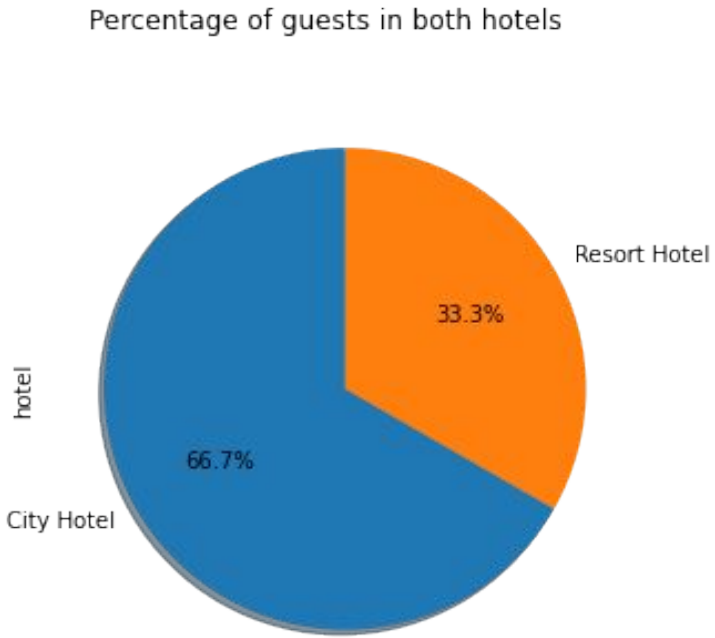
Brief of various column trends

Before we start getting insights from data here are histograms to have a brief picture of various column trends and data.

All columns with data type int64 are represented in histograms

EDA- Univariate Analysis

1. Booking ratio of both hotels

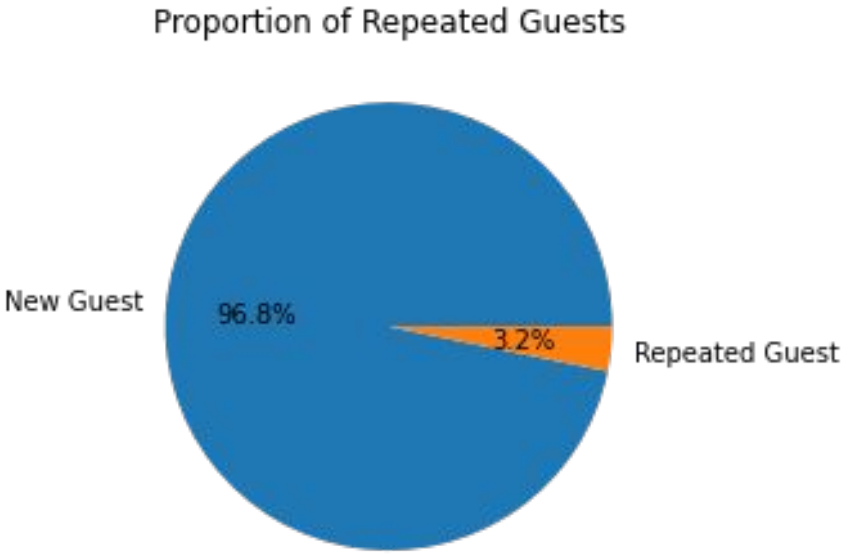


Ratio of Booking:

City Hotel have **66.7%** weightage

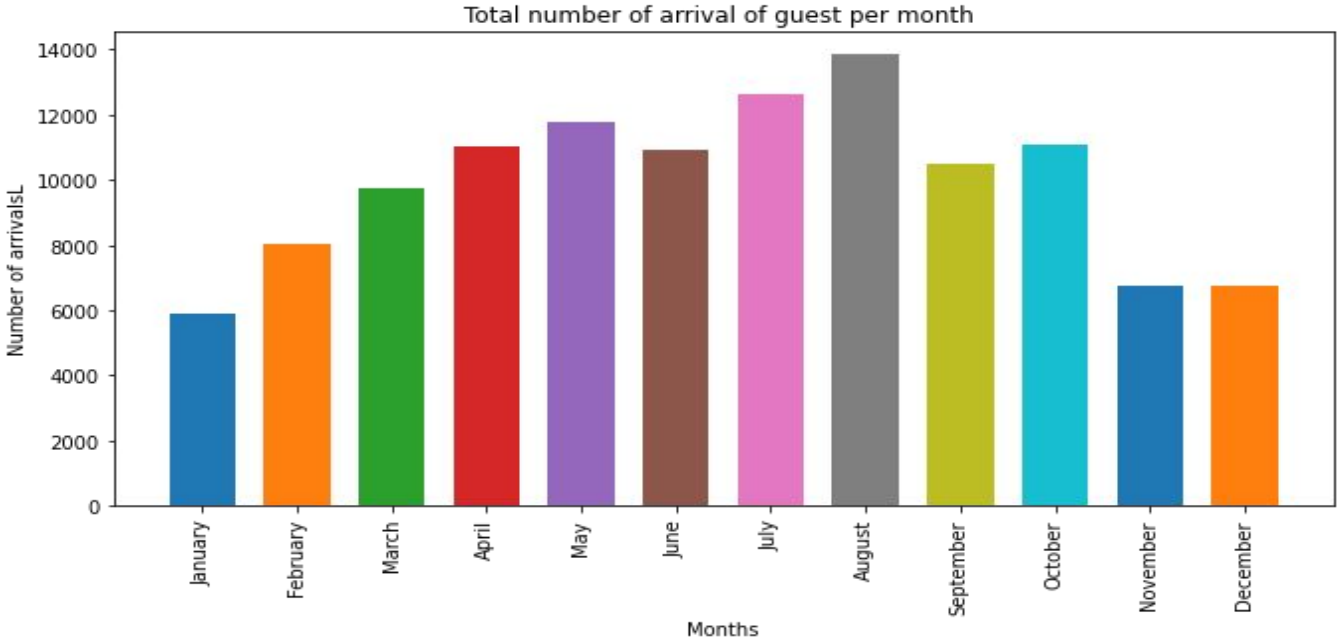
Resort Hotel have **33.3%** weightage

2. Repeated guests ratio.



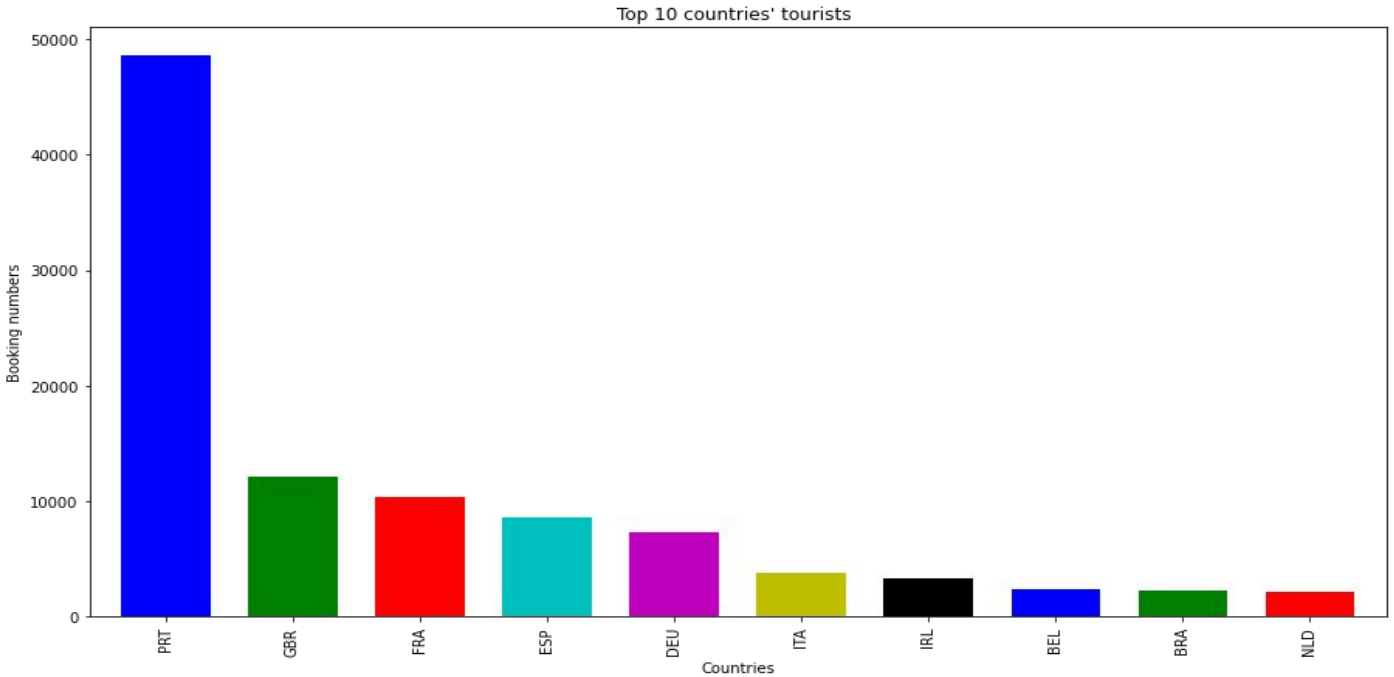
- Retention rate of hotel is very low.
- Only **3.2%** Guests are coming again and **96.8%** doesn't.

3. Total number of arrival of guests per month



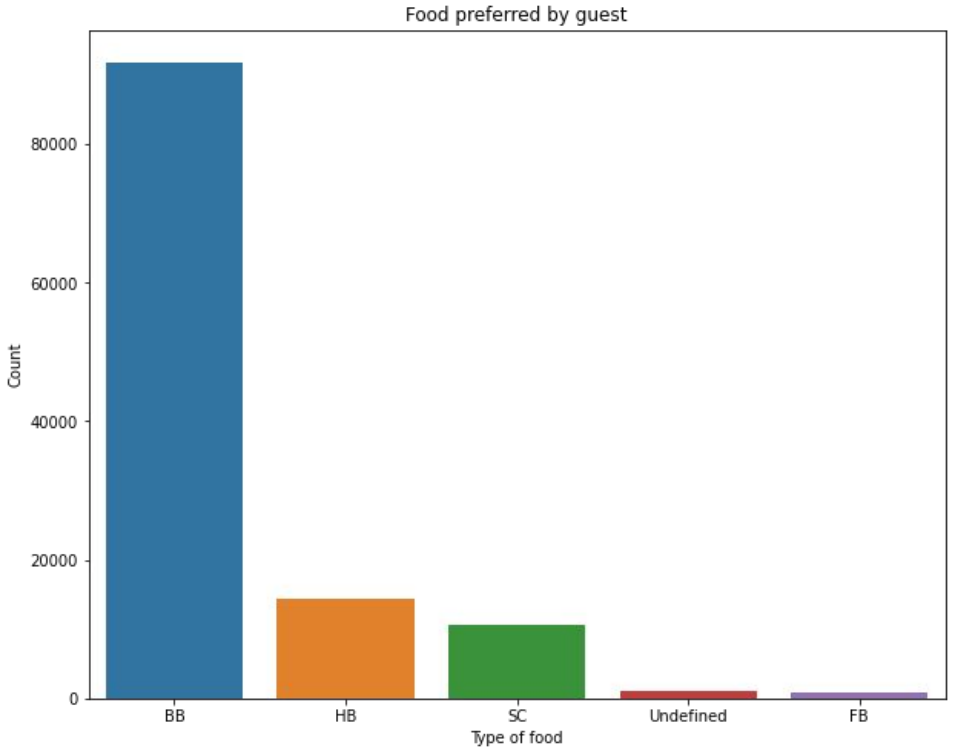
- Most number of guest comes in month of **August**.
- **August** is the busiest month for hotels.
- **January** is the least busiest month for hotels.

4. Top 10 Countries Guests.



- Most guest are from **Portugal** and other **European countries**
- We can conclude that most tourist who have booked rooms were from **portugal** and after that we have **GBR**(United kingdom) tourist.

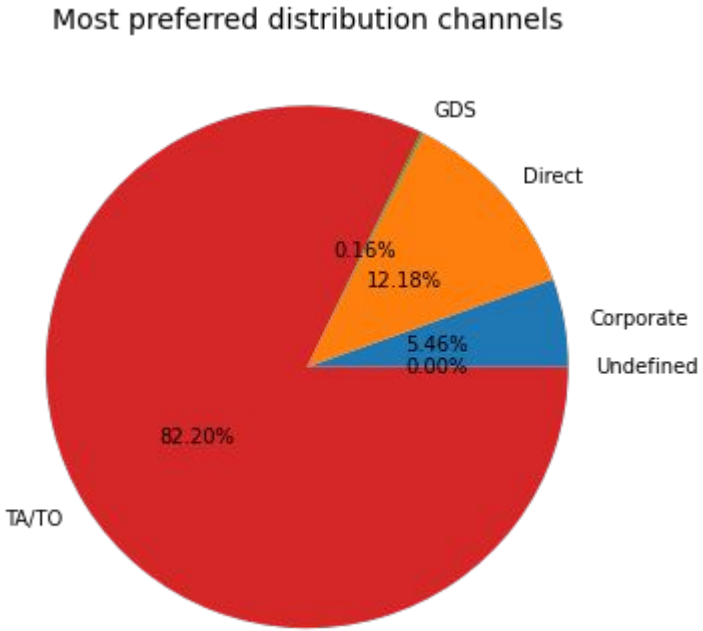
5. Meal Preference



- Most preferred meal by guests is **BB**(Bed and Breakfast) followed by **HB(**Half Board).
- Least preferred is **FB**(Full Board).

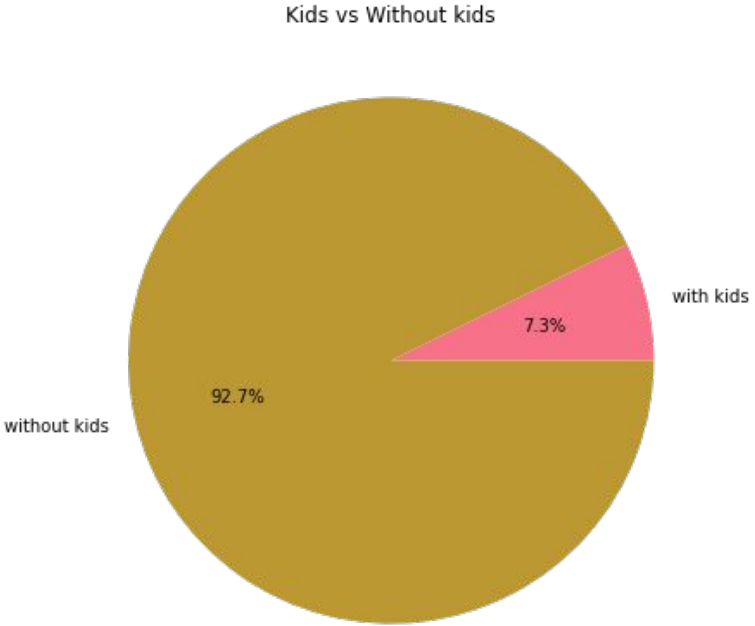
6. Most preferred channel for Hotel booking

- **82.2%** booking were done from channel **TA/TO**(“TA” means “Travel Agents” and “TO” means “Tour Operators”).
- 2nd most preferred channel for booking is **Direct** booking.



EDA- Bivariate Analysis

1. Number of adults traveling with kids/babies

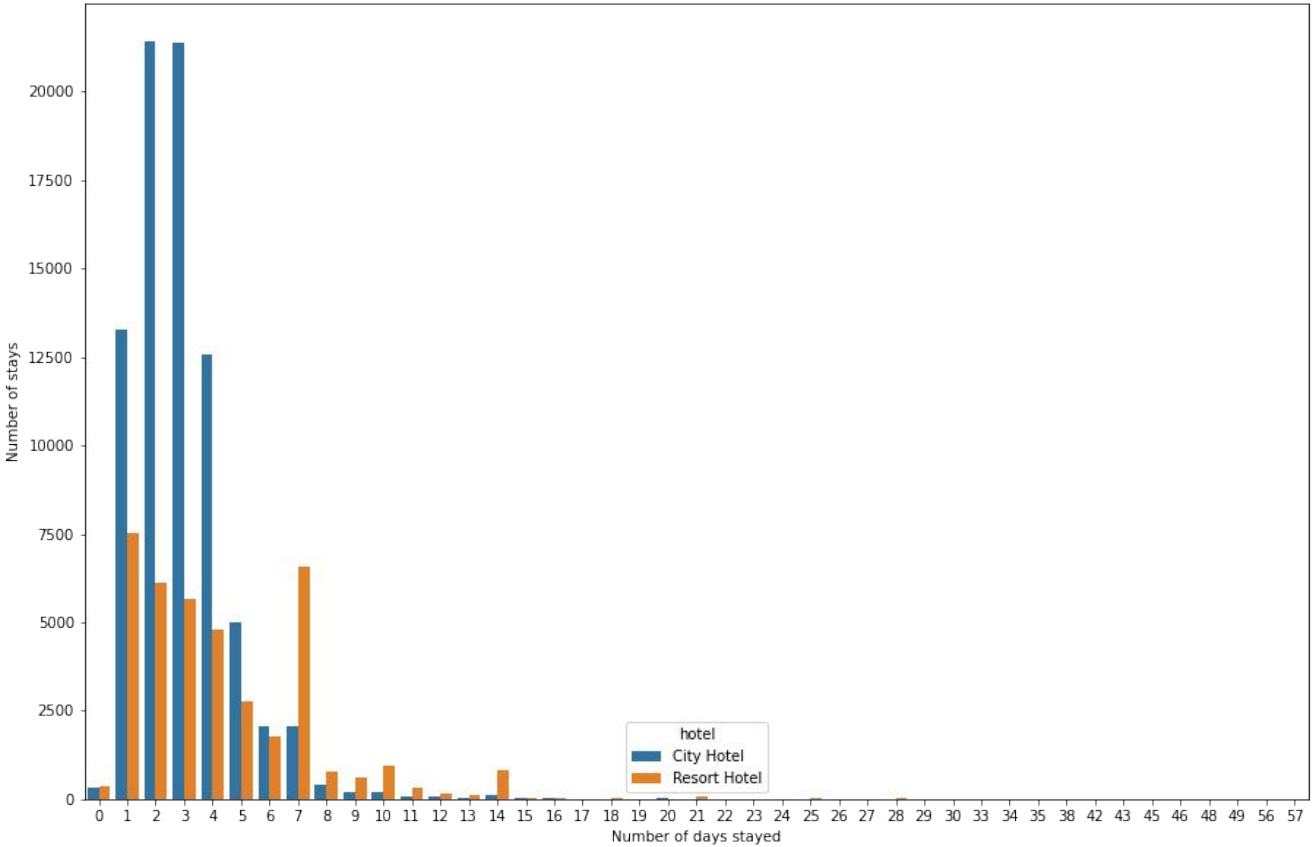


- Maximum Adults are traveling without any kids which are almost **92.7%.**
- Only **7.3%** Adults are traveling with kids.

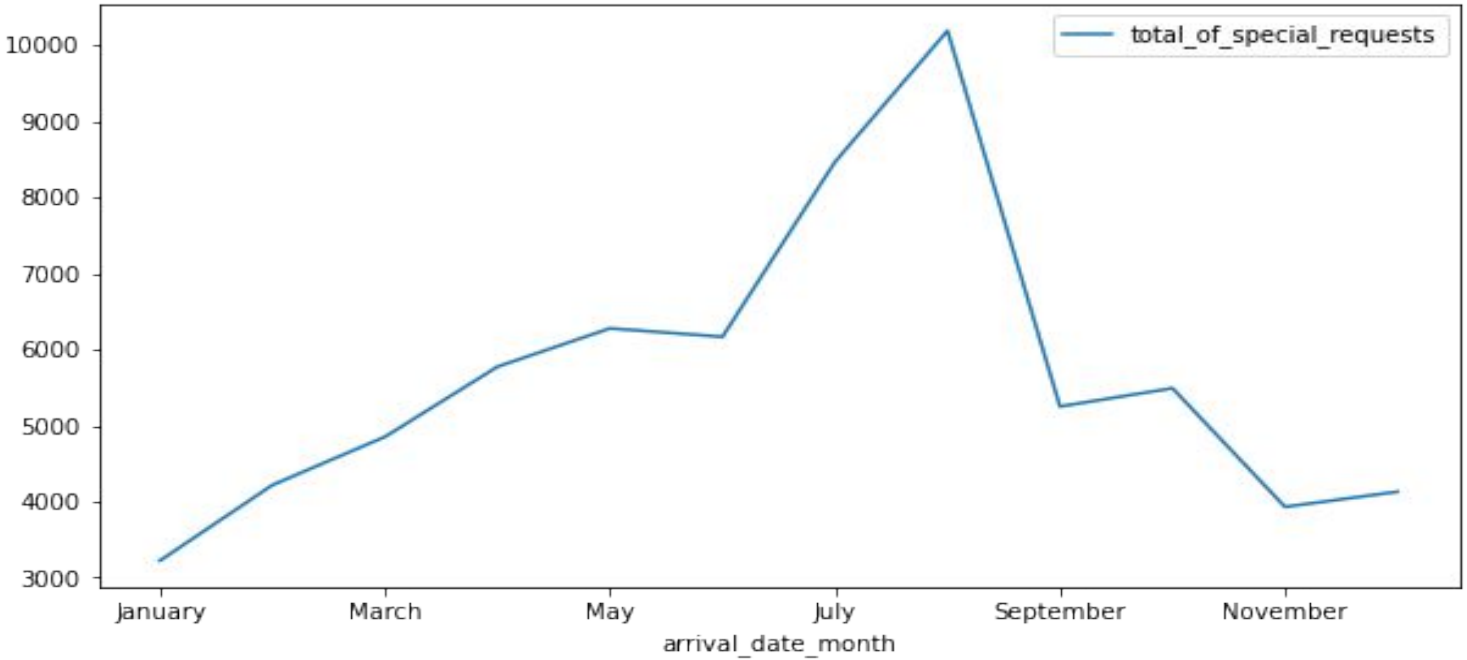
2. Optimal stay length in both hotel type

- Optimal stays in both the hotel is less than 7 **days** usually people stays for a week.

- For stay more than 7 days people like to stay in Resort hotel as we can see after 7 days City hotel booking are very less as compared to Resort hotel

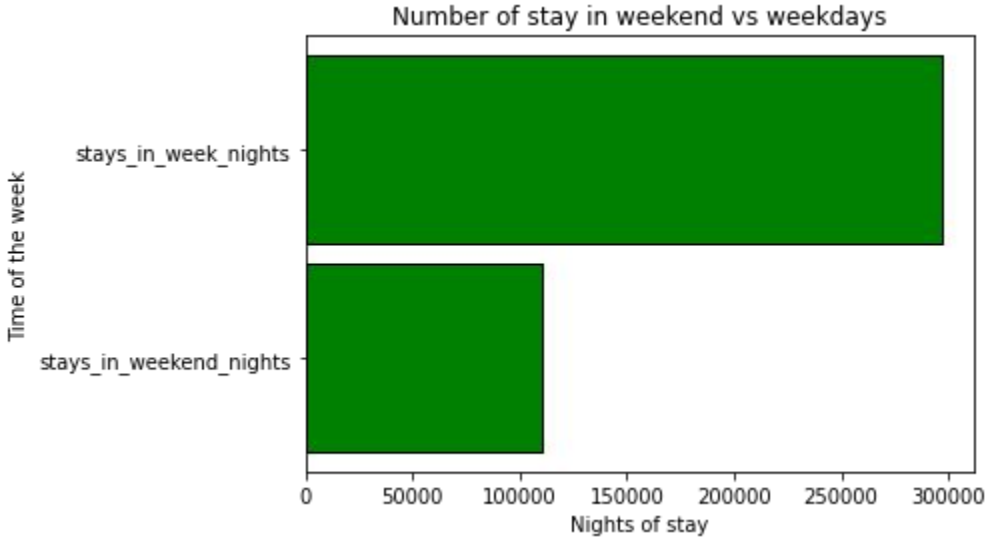


3. Number of special requests Month wise.



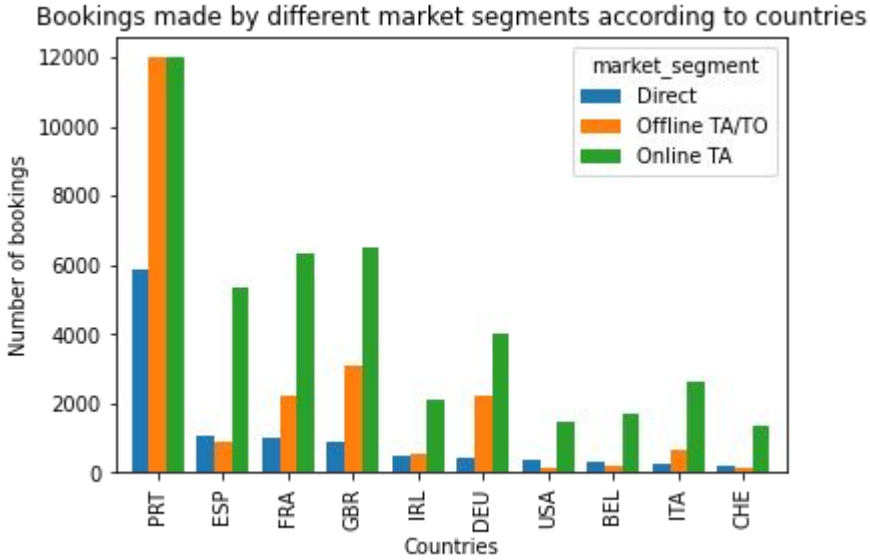
- Maximum number of special requests are made between **july** and **september**.
- Least requests are in **january**.
- we can state that it depends on number of guests arrival.

4. Plotting number of stays in weeknights vs weekend night.



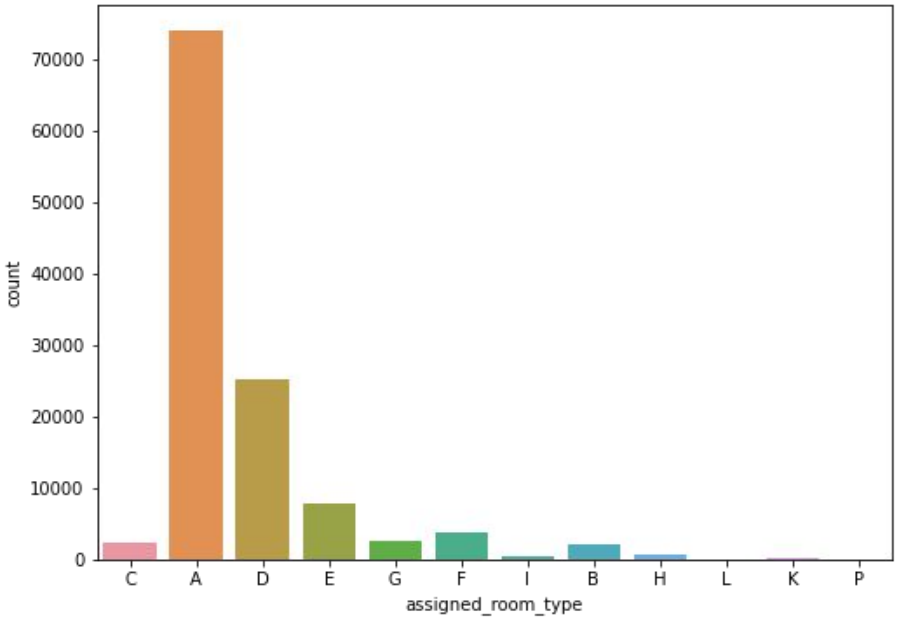
- **297499 stay days** were booked on weekdays and only **110444 stay days** were bookings were of weekends.

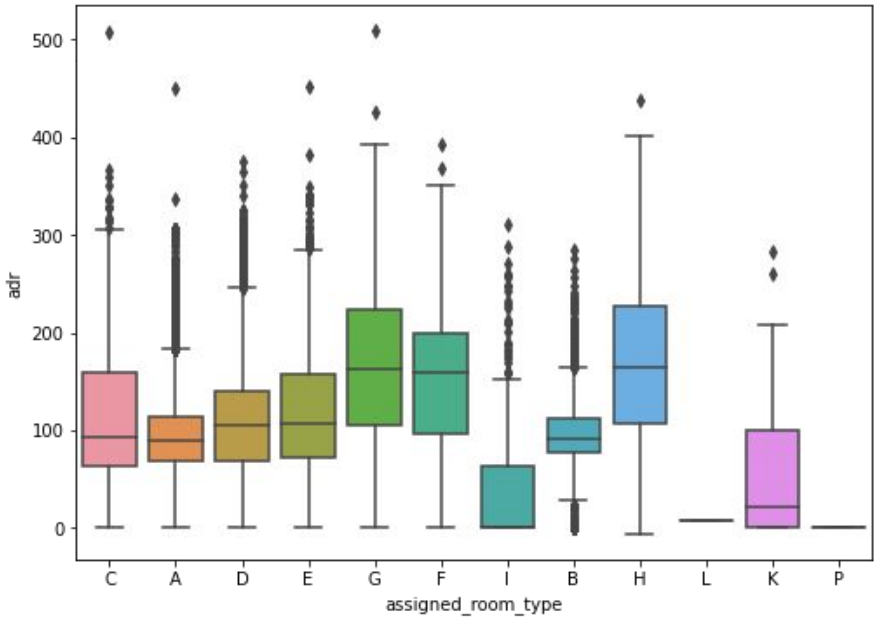
5. Bookings made through three main market segments from different countries.



- On average '**Online TA**' is the most preferred channel and on average least preferred is '**Direct**' channel.
- Maximum bookings are from **Portugal** country, followed by country **GRB**(United kingdom)

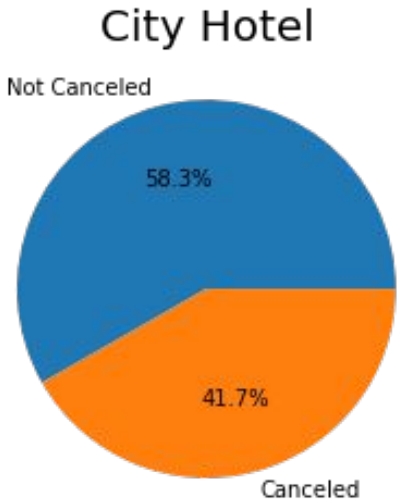
6. Plotting in demand room and which room generates more ADR



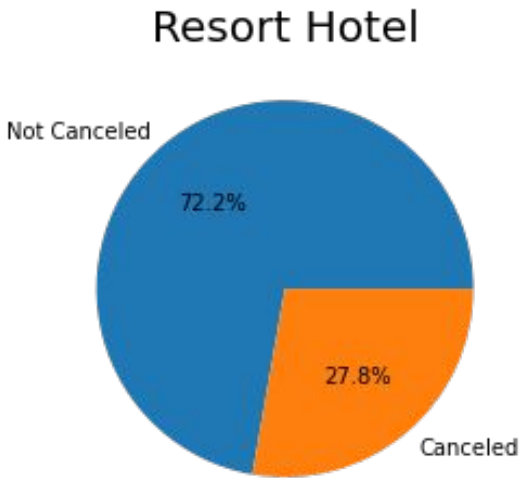


We can see that '**A**' type room is most in demand but on contrary room type '**H**', '**G**' and '**F**' are most ADR generating rooms respectively

7. Cancellation ratio of both hotels.



City hotels get **41.7%** cancellation and **58.3%** confirm booking out of all bookings.

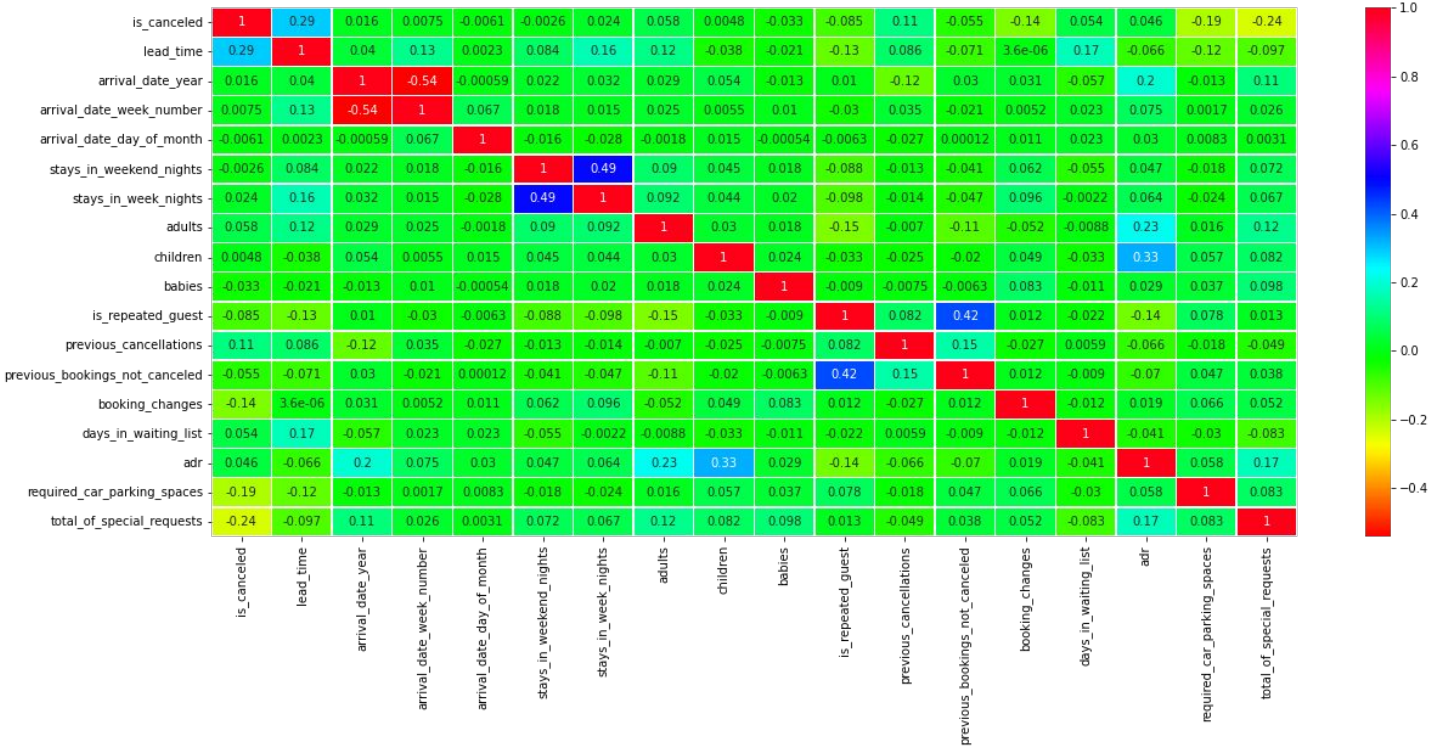


Resort hotels get **28.8%** cancellation and **72.2%** confirm booking out of all bookings.

- City hotels are canceled more as compared to resort hotels.

EDA- Multivariate Analysis

1. Correlation heatmap of data.

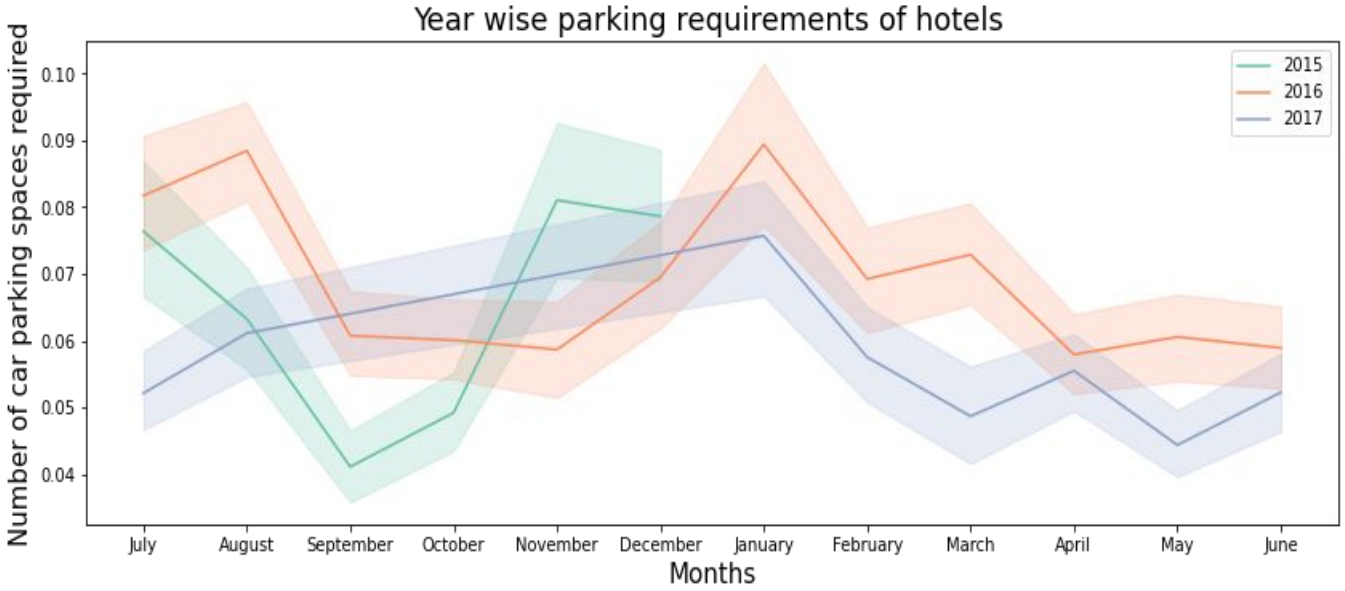


1. Correlation heatmap of data.

-Continued

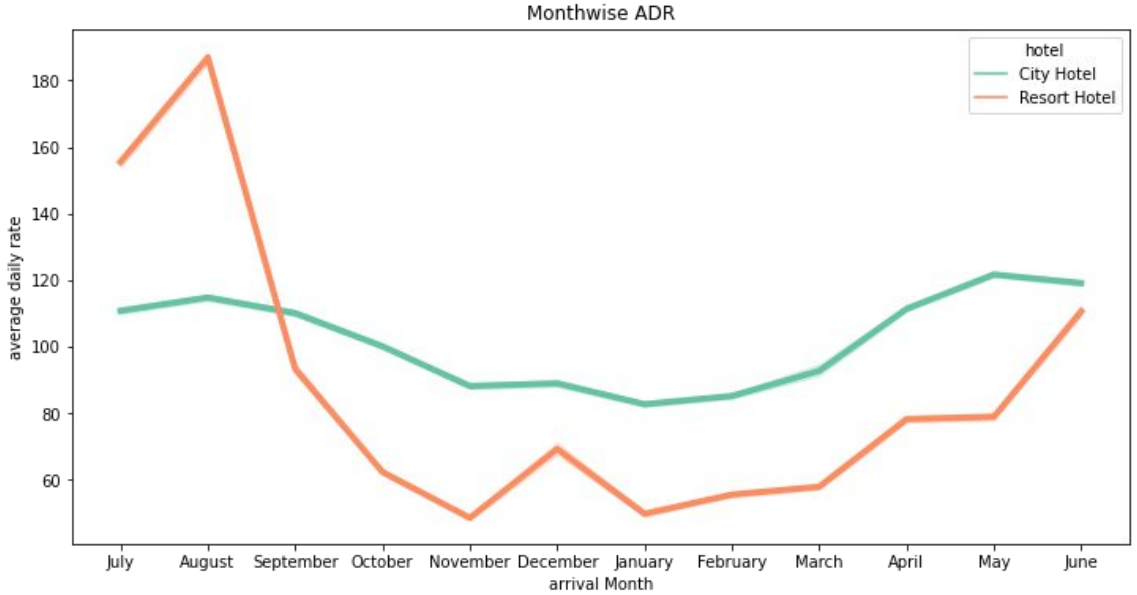
1. **ADR**(Average Daily Rate) and **guests with children** have slight positive correlation. which means more the kids, more is the ADR.
2. **Total stay** and **lead time** have positive correlation.
3. **Adr(**Average Daily Rate) is positively correlated with **total guests**. Which states more the guest will generate more ADR
4. **Repeated guests** and **previous bookings not canceled** has strong positive correlation. Repeated guests are more likely to not cancel their bookings.
5. **Company** and **agents** are slightly more correlated
6. **Stays in week night** and **total stay** are positively correlated, even more than **weekend nights** which says, longer stays are in week time only.
7. **Lead time** and **total stay** are positively correlated. That means more is the stay of customer more will be the lead time.
8. **Adults, children, Babies, total stay** and **ADR** has positive correlation which means more the people, longer the stay which will hike **ADR.**

2. Requirement of car parking space -Year and Month wise



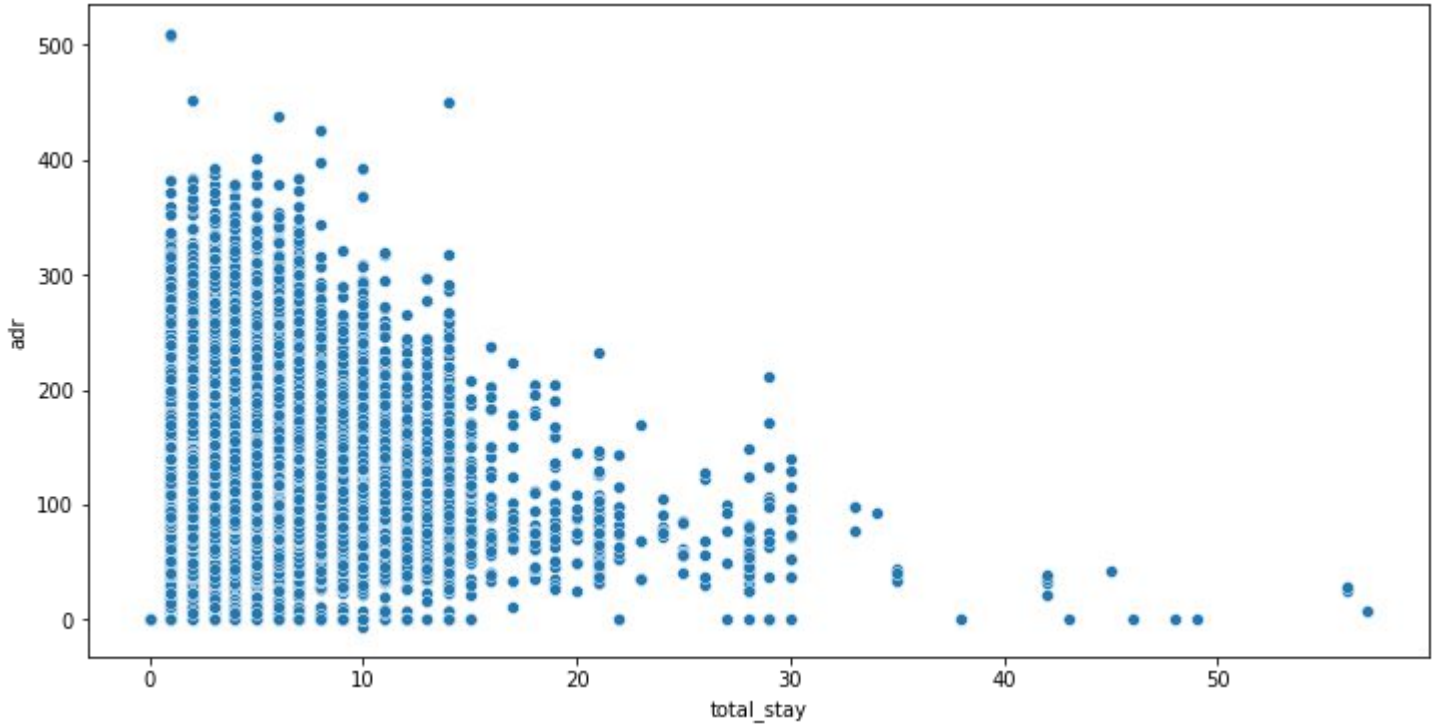
- We can see in **2015**-most parking spaces were needed in november and least in september.
- In **2016**- most parking spaces were needed in january and least in november.
- In **2017** most parking spaces were needed in january and least in may.

3. Average daily rate(ADR) Month wise.



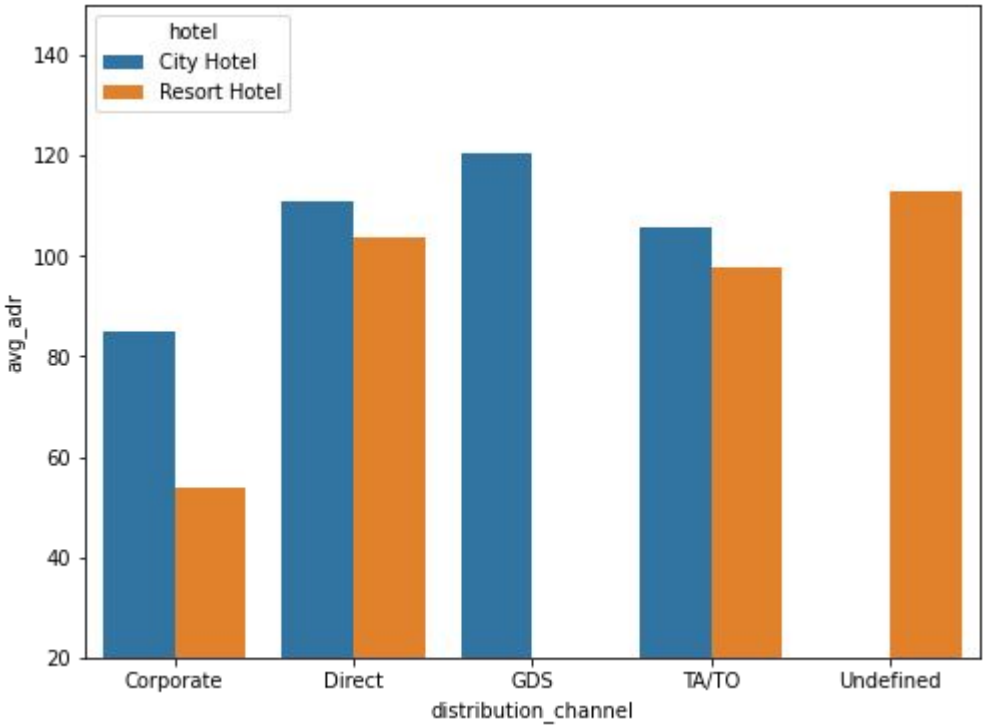
- By comparing the "Total number of arrival of guest per month" chart with "Month wise ADR" chart we can see that as the no. of guest increases **ADR** increases respectively.
- ADR is maximum in August for **Resort hotels** and in May for **City hotels**.

4. How length of stay affects the ADR.



- From this we can see that as the **length of stay** increases **ADR** decreases

5. Distribution channel with better revenue generating deals for hotels.



By comparing "Most preferred distribution channels" with This chart we can say that :

- GDS channel brings higher revenue generating deals for City hotel, in contrast to that most bookings come via **TA/TO**. City Hotel can work to increase outreach on **GDS** channels to get more higher revenue generating deals.
- Resort hotel has more revenue generating deals by direct and **TA/TO** channel. After direct channel.
- Resort Hotel need to increase outreach on **GDS** channel to increase revenue

Conclusions

After careful analysis , we can conclude that the hotel industry can benefit a lot by studying the type of customers, their booking mode, the booking month and the seasons. The hotel industry market, their ADR and bookings are based on the type of customers,the month, types of meal, hotel type ,their country of origin, Room types, booking medium and many others.

Suggestions

1. The hotel industry can take the advantage of seasons and months as ADR was highest in august (rainy season).
2. Most customers booked rooms online so they can be targeted with proper seasonal discounts and vacay-ads.
3. Since ADR was least during Nov and Jan, winter discounts(assumption) or off season discounts might help.
4. For retention, they should introduce portuguese meals(sea foods and meat) and eastern european meals as guests are more from there.
5. They should encourage direct bookings by offering some special discounts as online bookings cancellation is high.
6. Since room A is booked more, they should take into account the factors how it is different from other rooms and implement the same in other rooms as well.
7. Since resort hotels are less preferred, they should look into the factors- might be High cost or guests requirements.

Thank You