Capstone Project-1

Hotel Booking Analysis

by - Swapnil Aher

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

EDA will be divided into following 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.

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-

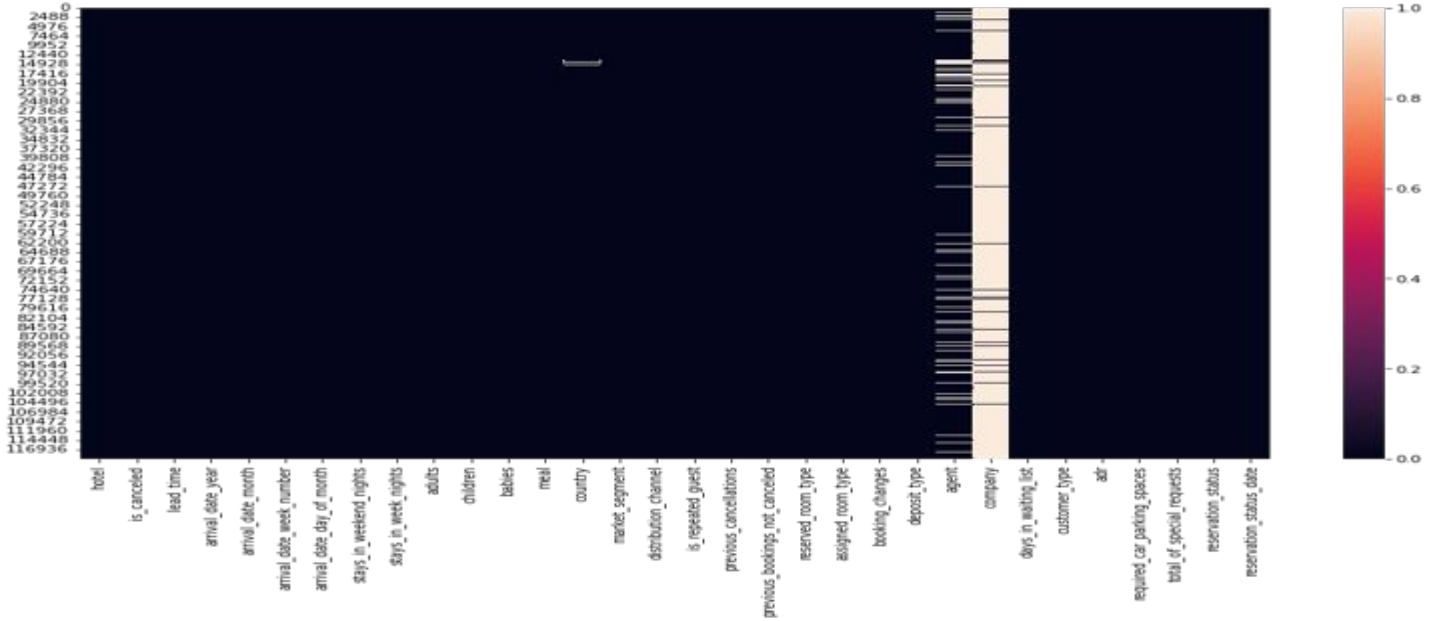
- **market_segment** : where the bookings came from.
- **distribution_channel** : Booking distribution channel.
- **is_repeated_guest** : repeated guest (1) yes or not (0).
- **previous_cancellations** : previous bookings that were cancelled
- **previous_bookings_not_canceled** : previous bookings that 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.**

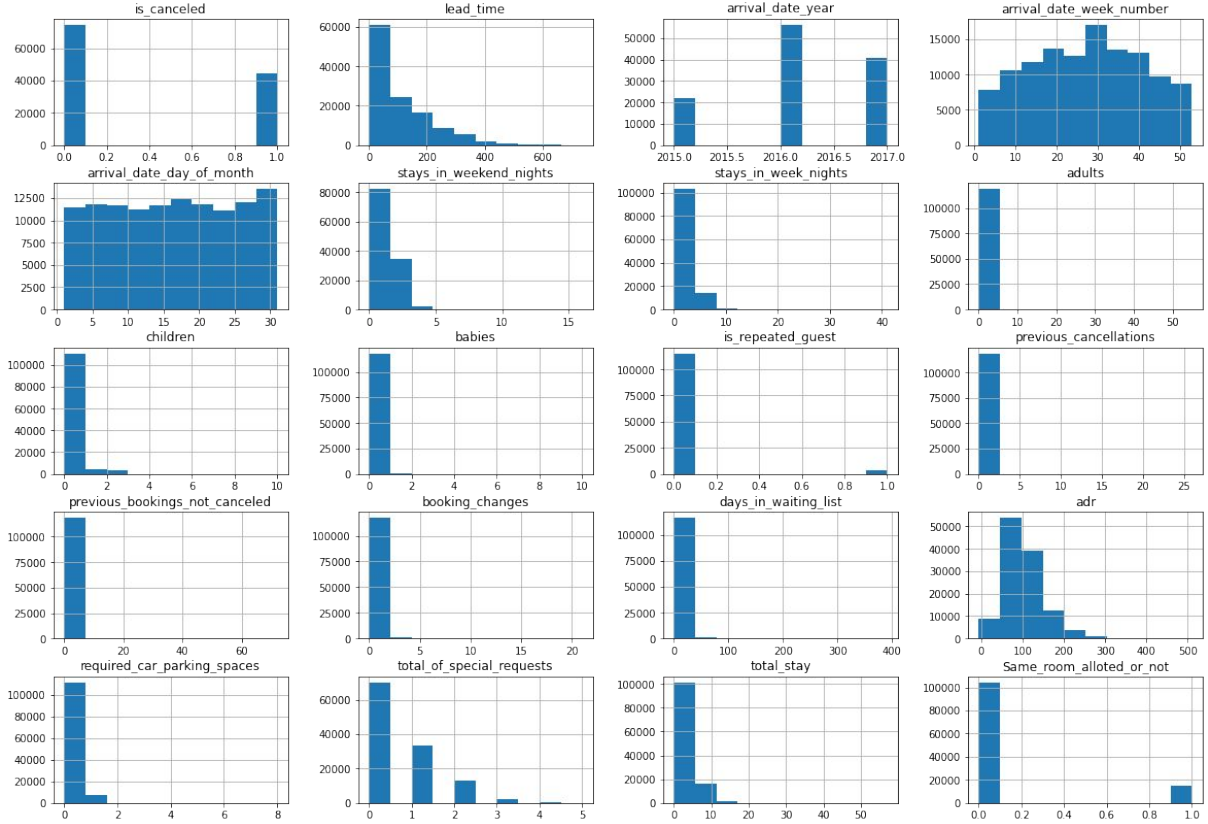
Data Cleaning



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

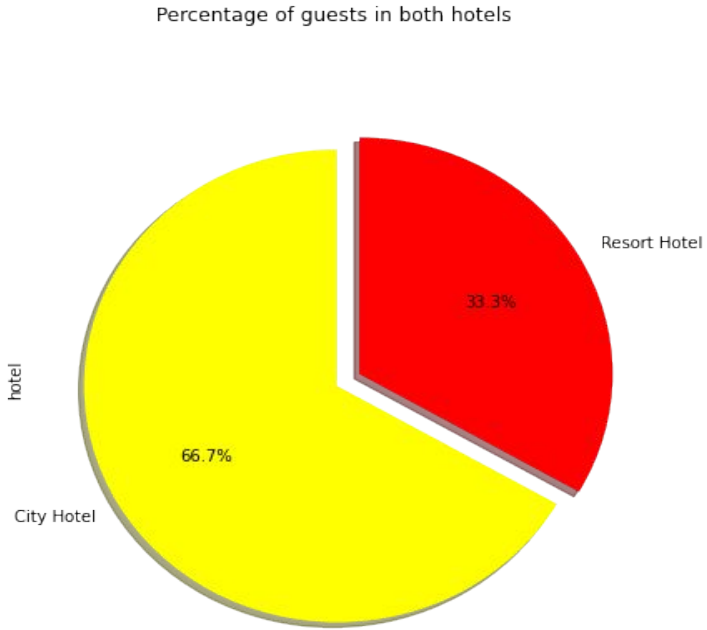


Brief of various column trends:

- 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

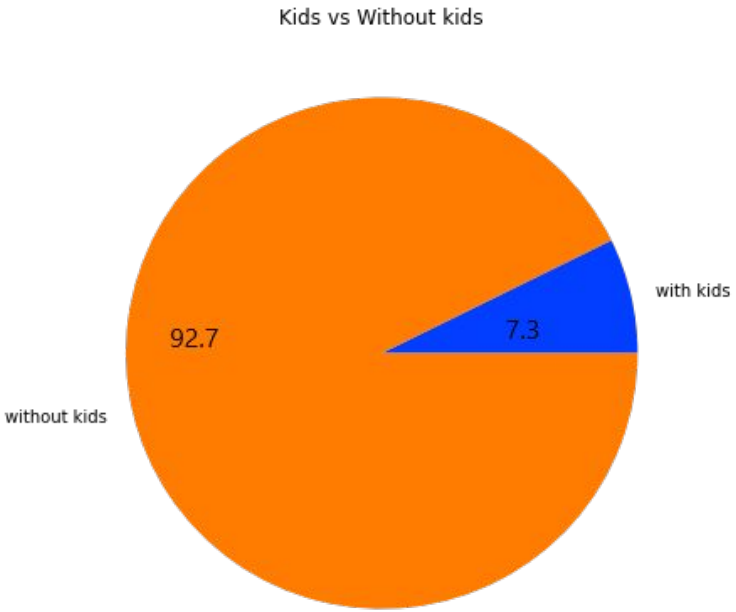
1. Most Preferred Hotel By The Guests



Ratio of Booking

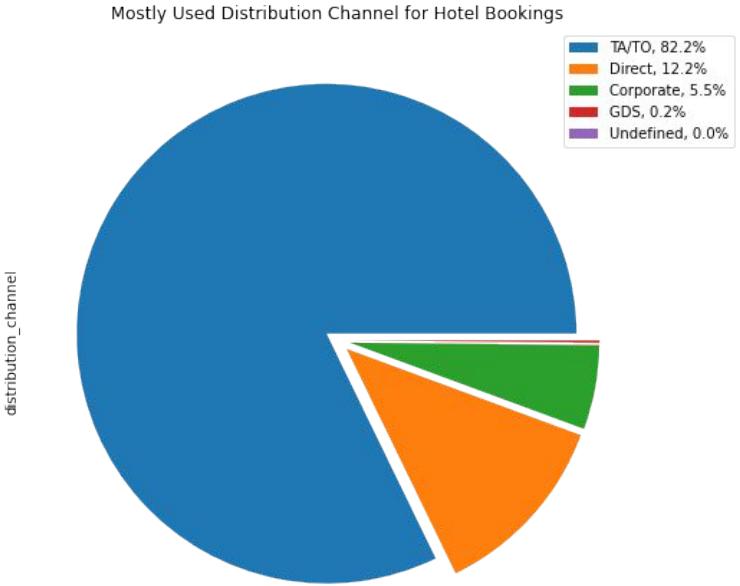
- City Hotel is most preferred hotel by the guests having **66.7%** weightage.
- Resort Hotel is less preferred having **33.3%** weightage.

2. Adults Travelling with Kids Or without Kids



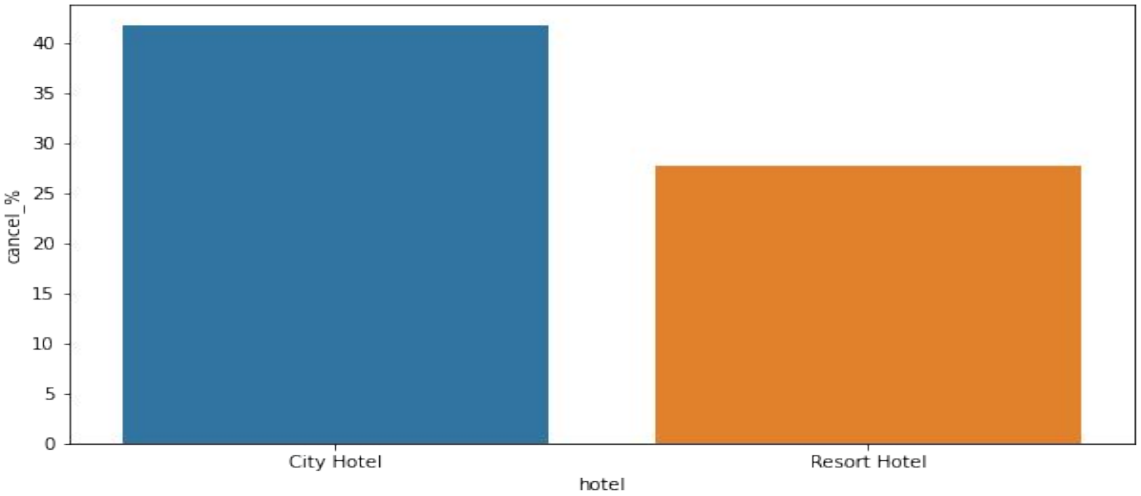
- Maximum Adults are traveling without any kids which are almost **92.7%.**
- Only **7.3%** Adults are traveling with kids.

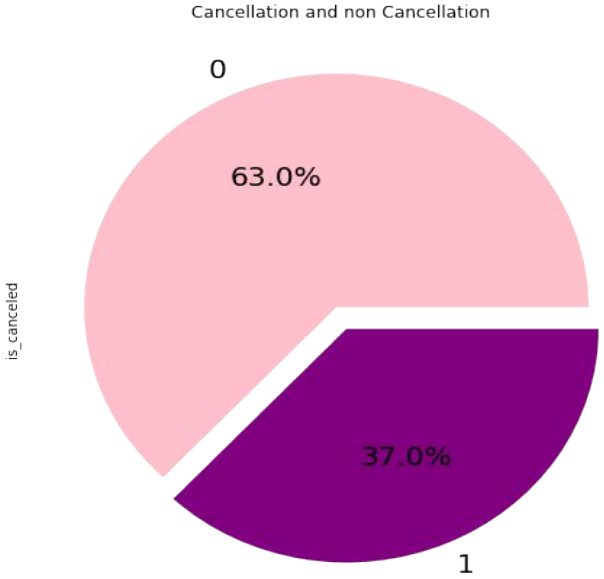
3. Most Preferred Distribution Channel For Hotel Booking



- **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.

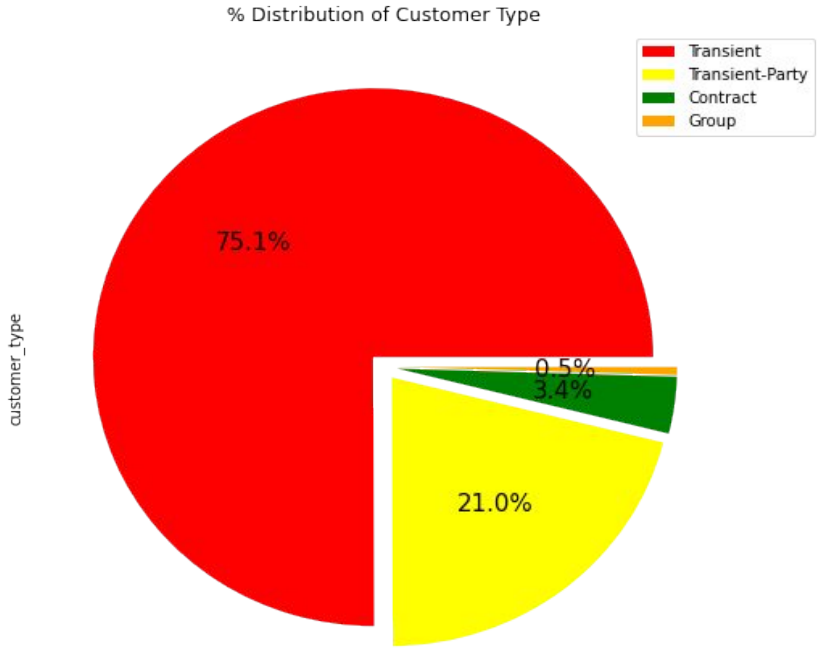
4. Hotel Booking Cancellation rate





- City Hotel Booking cancellation rate is **41.73%.**
- Resort Hotel Booking cancellation rate is **27.76%.**
- From above observation it is clear that City Hotel has higher cancellation rate.
- 0=cancelled, 1= not cancelled
- Overall 37% bookings were cancelled

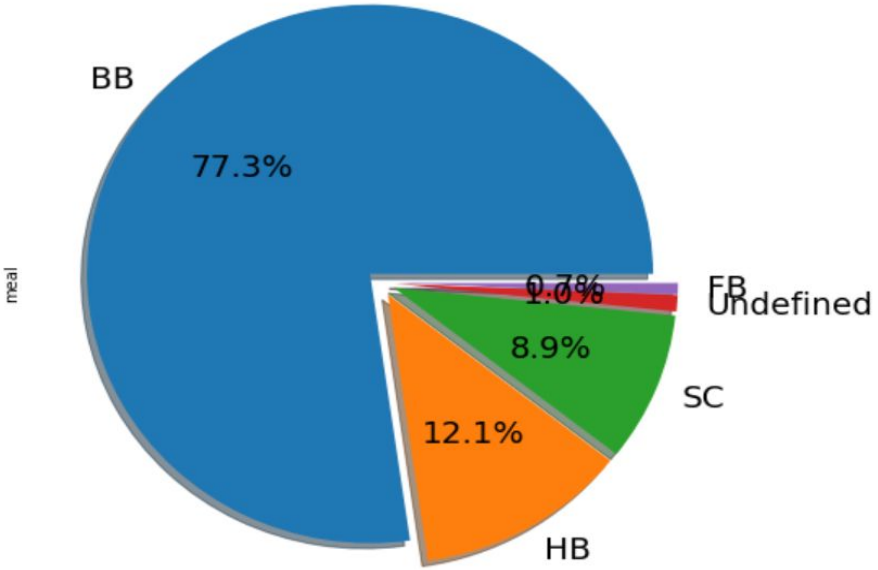
5. Distribution of Customer Type

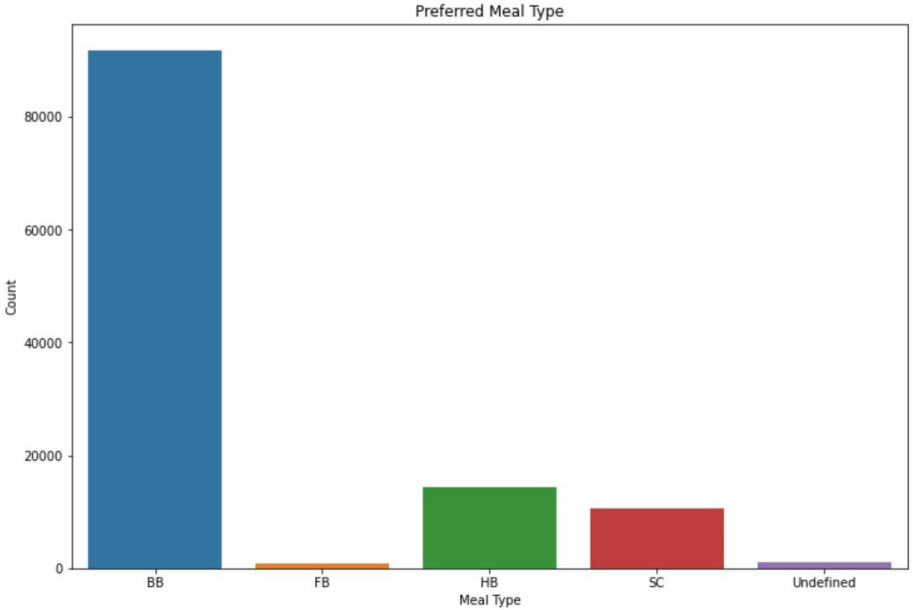


- From Above Graph it is clear that Transient customer type is more which is **75% .**
- Percentage of Booking associated by the group is very low.

- **Contract:**
  When the booking has an allotment or other type of contract associated to it.
- **Group:**
  When the booking is associated to a group.
- **Transient:**
  When the booking is not part of a group or contract , and is not associated to other transient booking.
- **Transient-Party:**
  When the booking is transient, but is associated to at least other transient booking.

6. Meal Preference



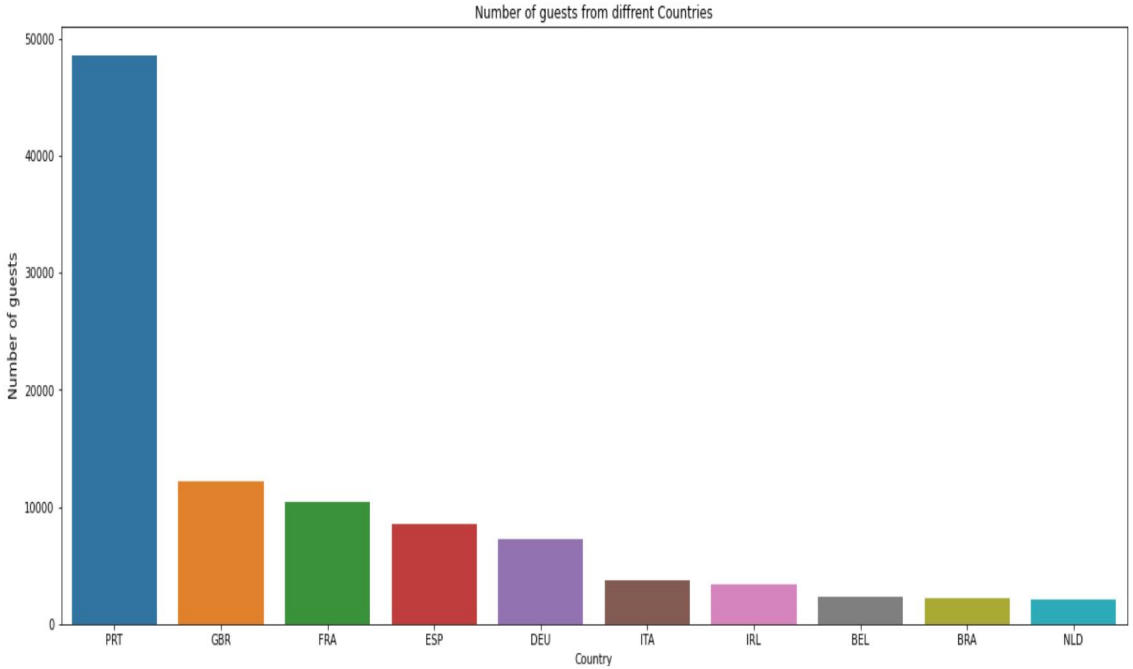


Types of meal in hotels:

- BB - (Bed and Breakfast), HB- (Half Board), FB- (Full Board), SC- (Self Catering)

- Most preferred meal type by the guests is BB(Bed and Breakfast), HB- (Half Board) and SC- (Self Catering) are equally preferred.

7. Top 10 Countries

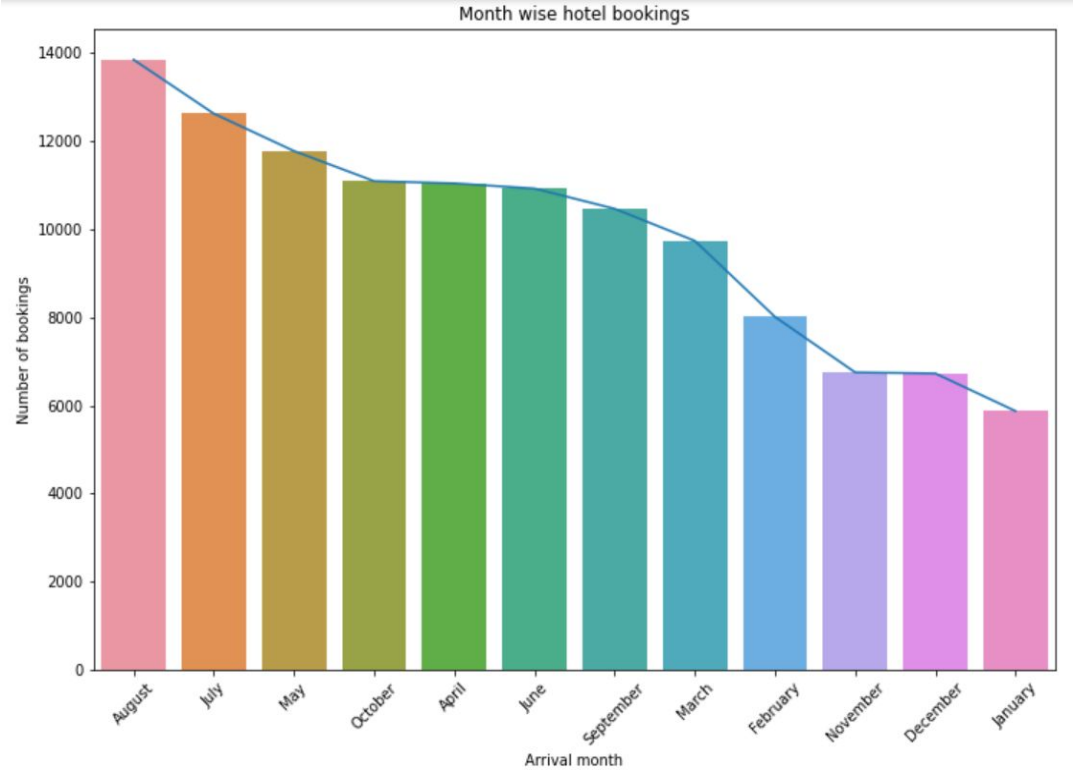


- PRT- Portugal
- GBR- United Kingdom
- FRA- France
- ESP- Spain
- DEU- Germany
- ITA- Italy
- IRL- Ireland
- BEL- Belgium
- BRA- Brazil
- NLD- Netherlands

- Most of the guests are coming from Portugal i.e. **4800** guests.

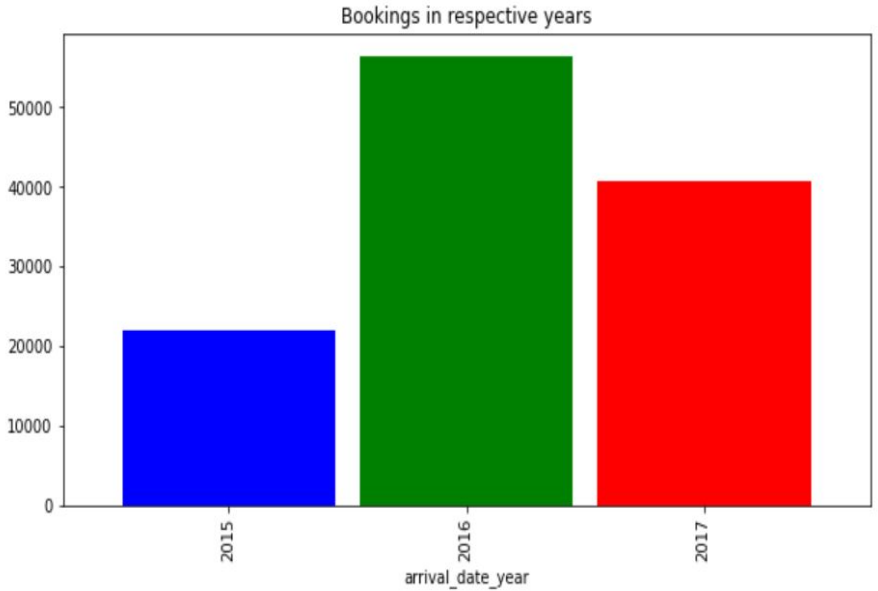
Bivariate Analysis

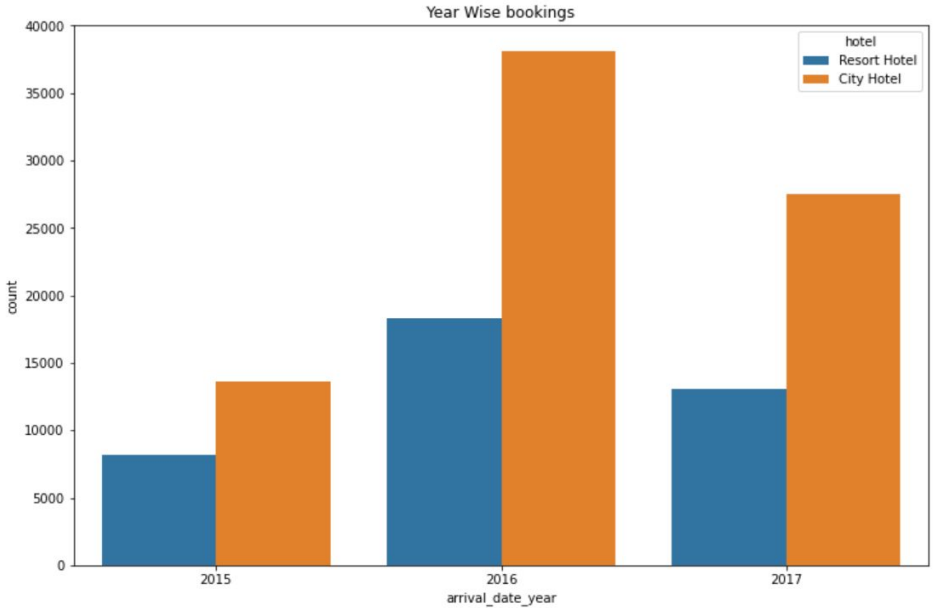
1. Most Bookings in Month



- July and August months had the highest number of Bookings.
- Summer vacation can be the reason for the bookings.

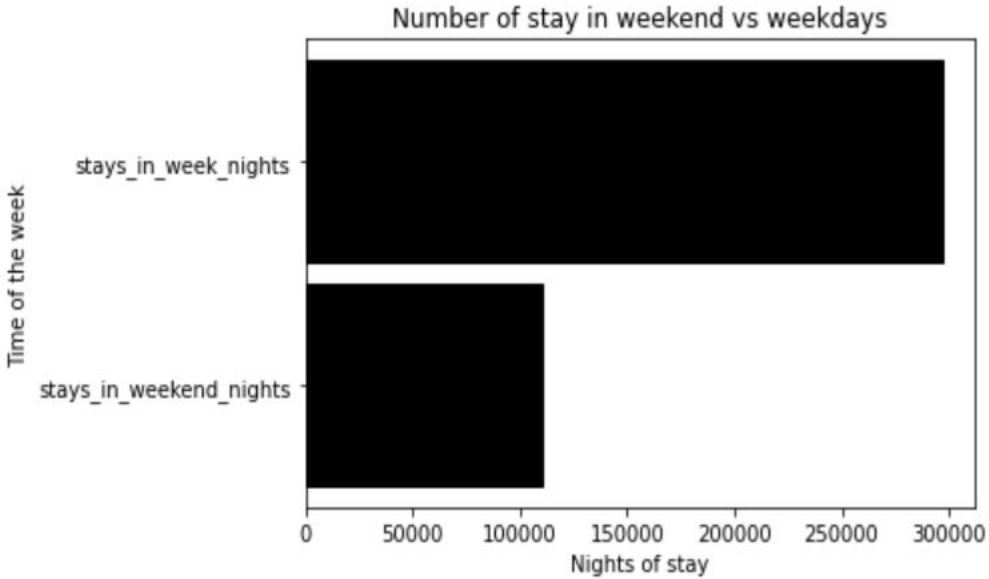
2. Highest Bookings in Hotel/Year





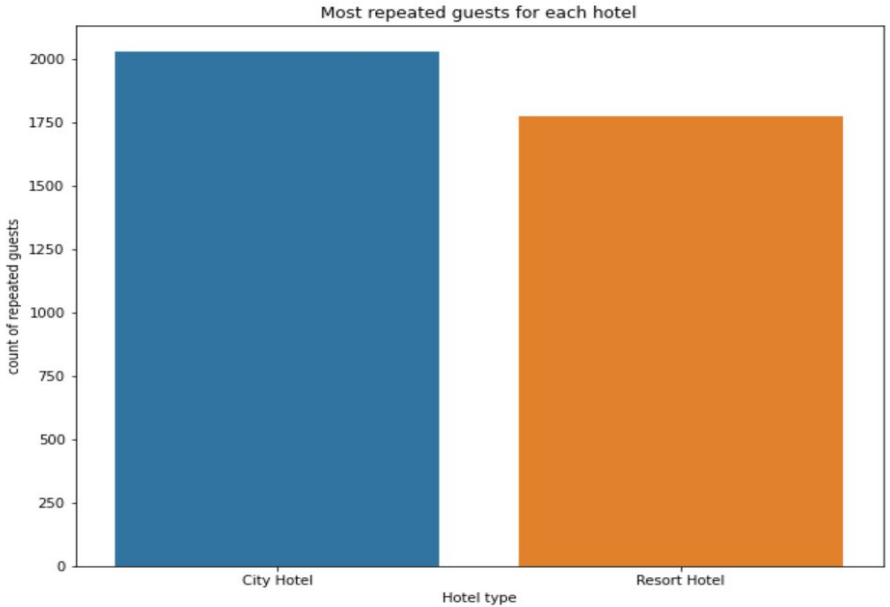
- 2016 had the highest bookings.
- 2015 had the lowest bookings.
- Overall City hotels had the most of the bookings.

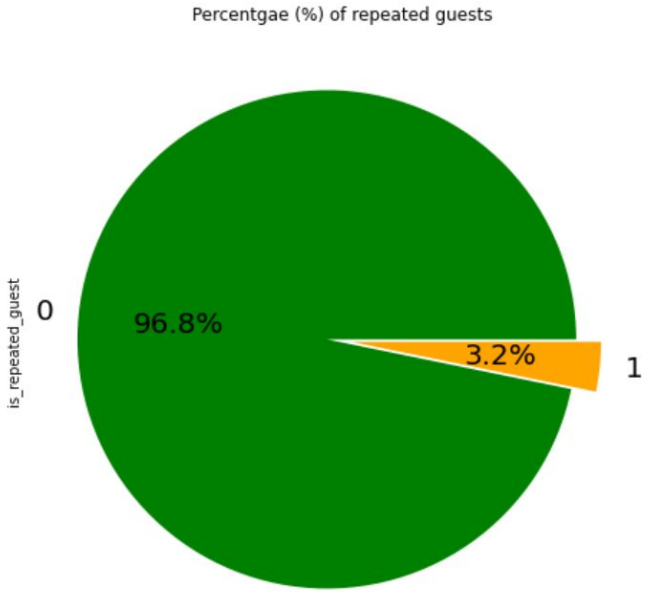
3. Highest Stays in Week or Weekend Nights



- **297499 stays days** were booked on weekdays and only 110444 stays days were booked on weekends.
- Guests Stays more in week nights than weekend nights

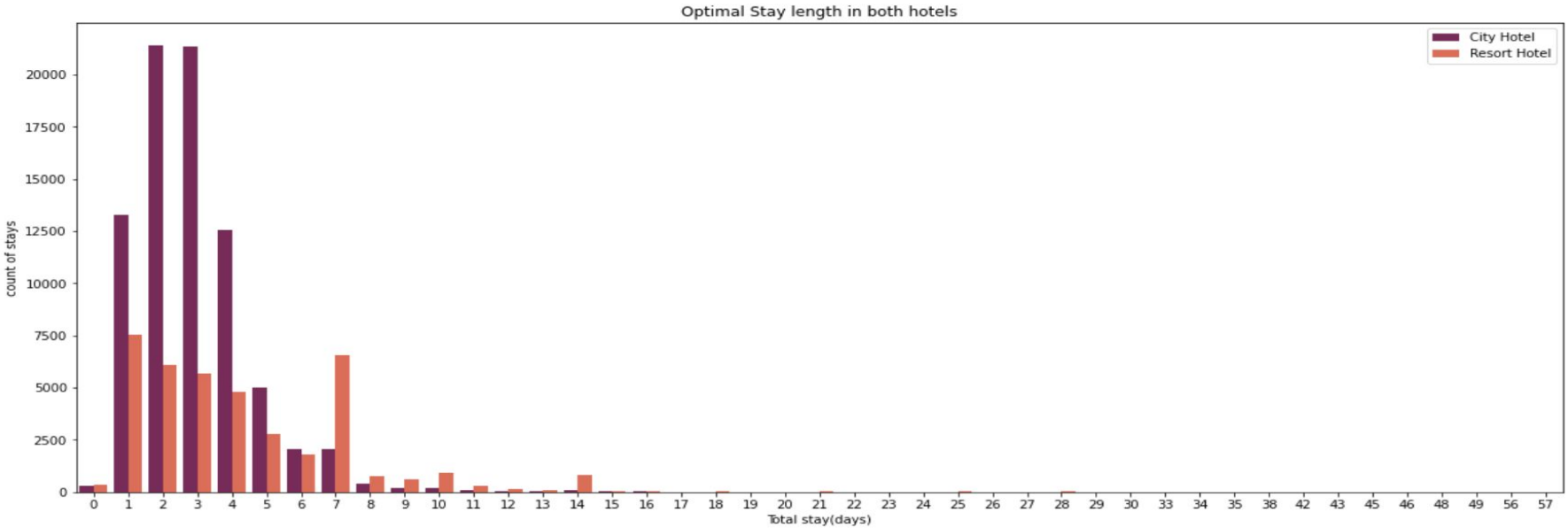
4. Hotel with Repeated Guest





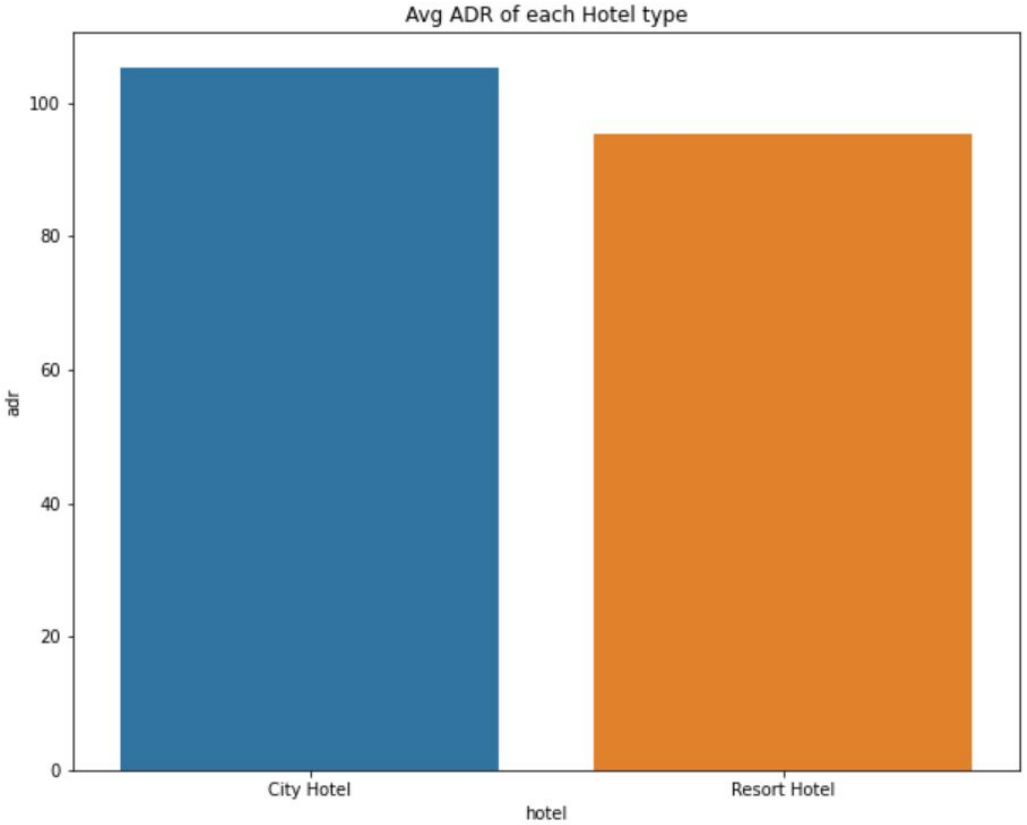
- Resort Hotel has slightly more repeated guests than the City Hotels.
- It is almost similar for both hotels.
- Overall repeated guests are very few which only 3.2 %.
- In order to retained the guests, management should take feedbacks from guests and try to improve the services

5. Optimal Stay Length in Hotels



- Optimal stay length in both hotels 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.
- On an average customer preferred to stay 1 to 4 days.

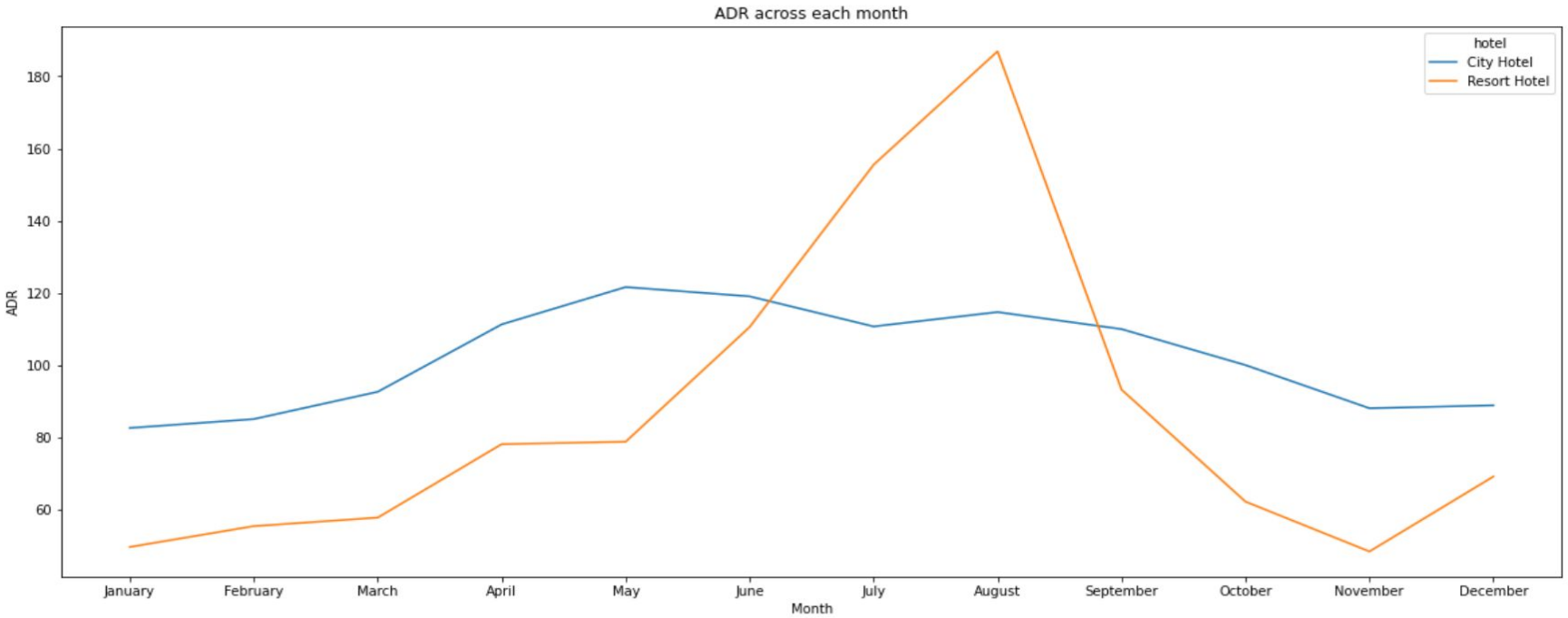
6. Hotel having highest ADR



- City hotel having highest ADR that means city hotels are generating more revenues than the resort hotels.
- More the ADR more is the revenue.

  ADR- Average Daily Rate

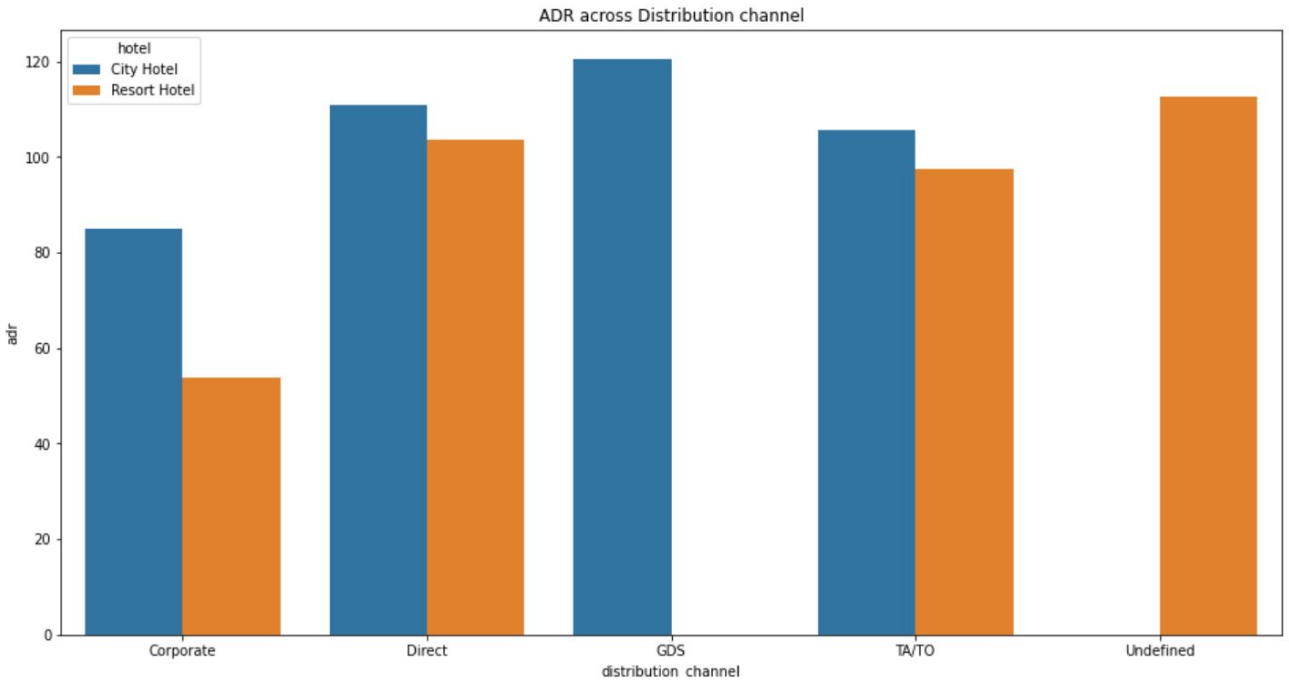
7. Hotel generating more Revenue



- For Resort hotel is ADR is high in the months June, July, August as compared to City Hotels. May be Customers/People wants to spend their Summer vacation in Resorts Hotels.

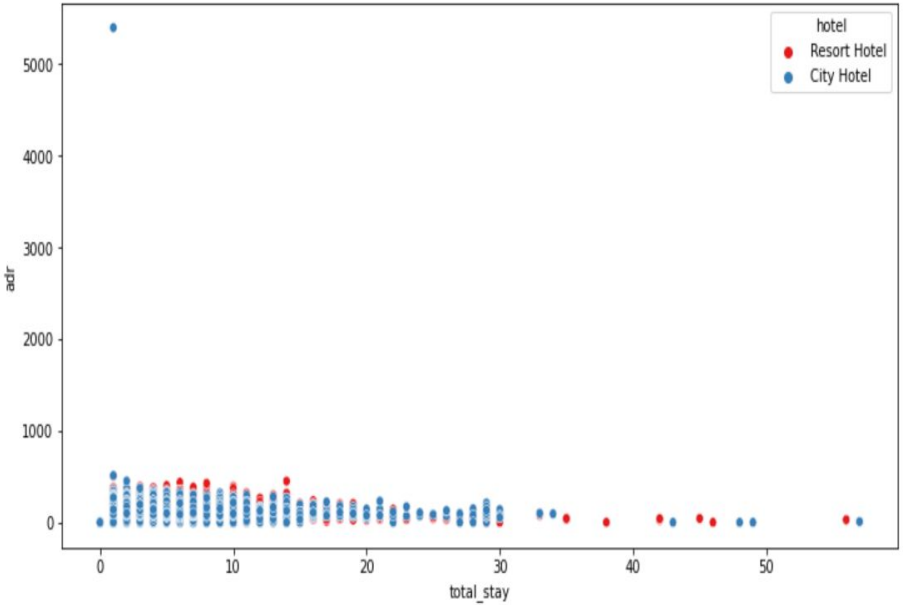
- The best time for guests to visit Resort or City hotels is January, February, March, April, October, November and December as the average daily rate in this month is very low.

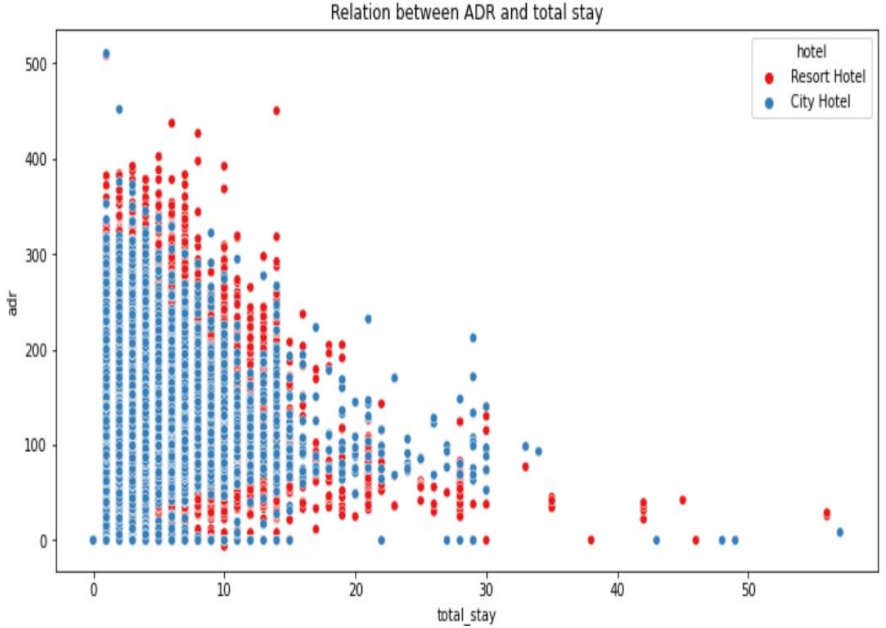
8. Distribution channel contributed in Income



- 'Direct' and 'TA/TO' has almost equally contributed in ADR in both type of hotels i.e. 'City Hotel' and 'Resort Hotel'.
- GDS has highly contributed in ADR in 'City Hotel' type.
- GDS needs to increase Resort Hotel bookings.

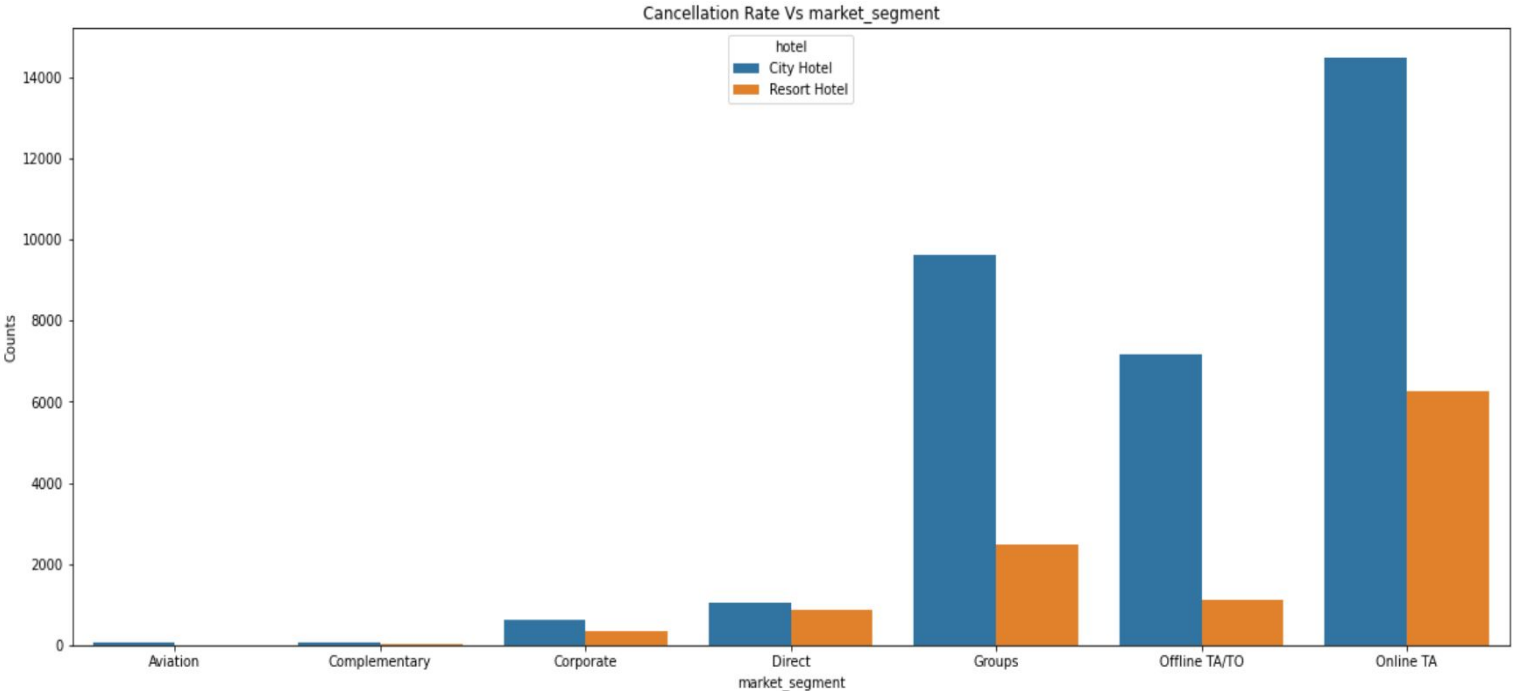
9. ADR affected by length of Stay-





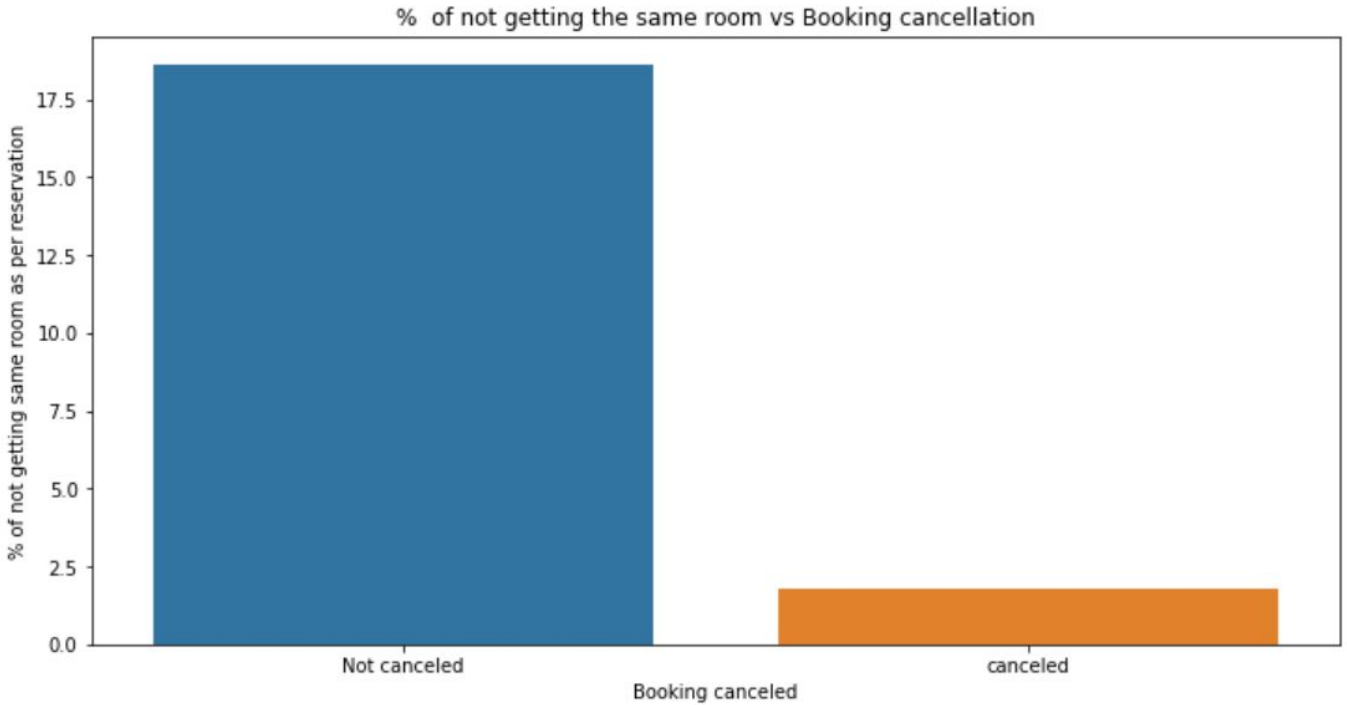
As the total stay increases the ADR(Average Daily Rate) Decreases.

10. Market Segment with Highest Cancellation Rate



- 'Online T/A' has the highest cancellation in both type of cities.
- In order to reduce the booking cancellations hotels need to set the refundable/ no refundable and deposit policies.

11. Allotment of Room type as Reserved



There's not much effect on cancellation of the bookings even if the guests are not assigned with rooms which they reserved during booking.

EDA-Multivariate Analysis

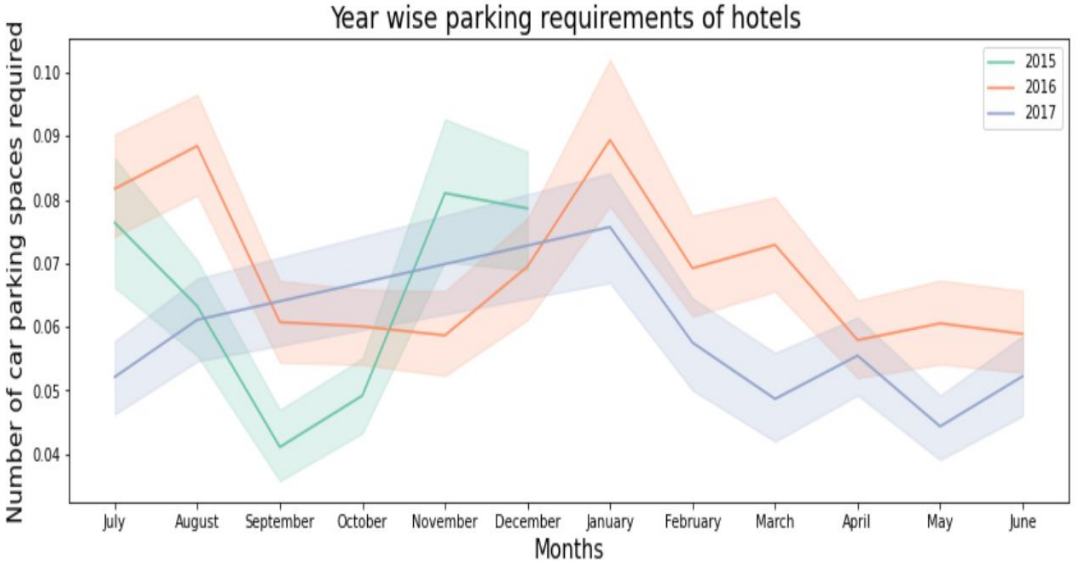
1. Correlation heatmap of data

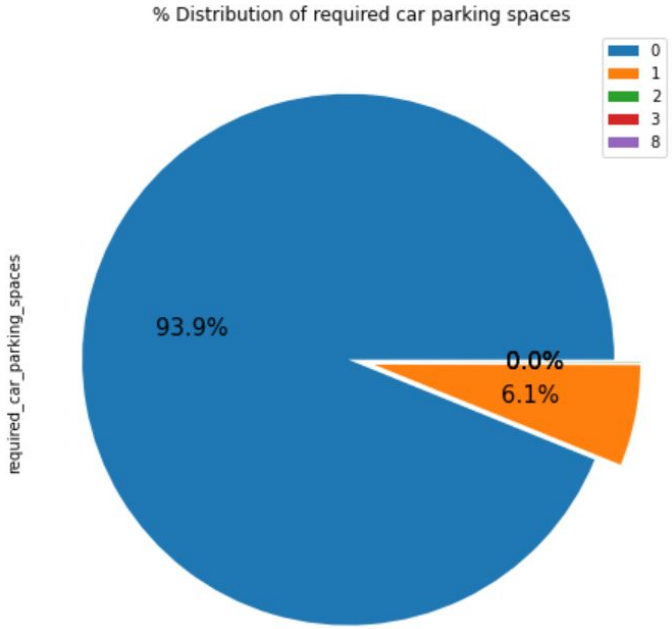
1. Correlation heatmap of data

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.**

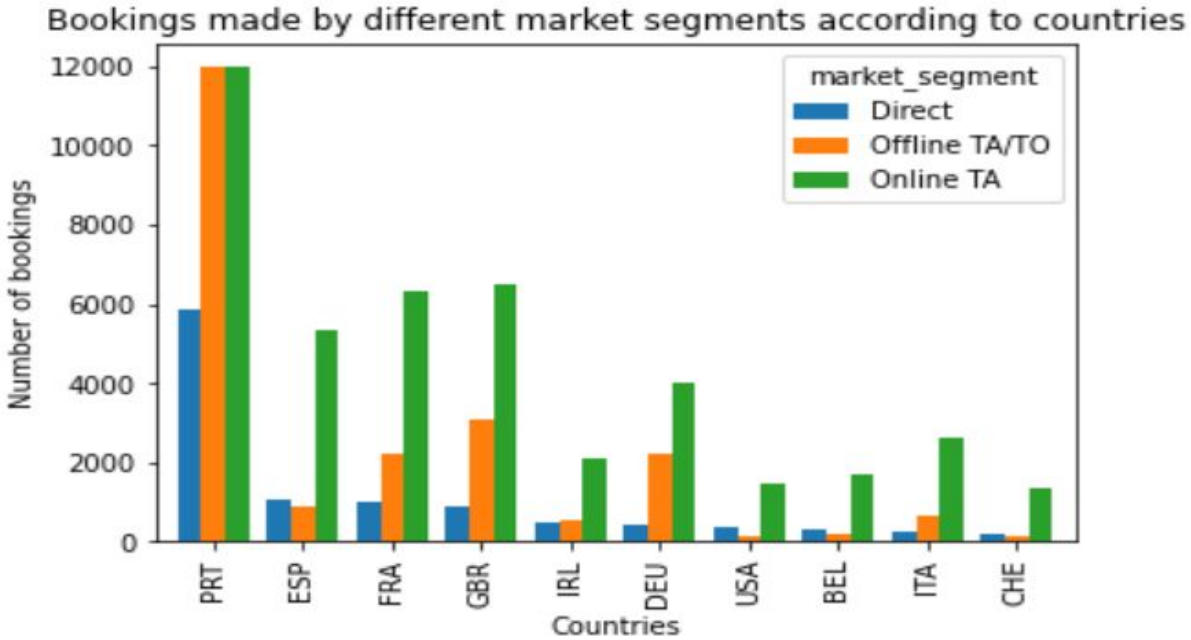
2. Car Parking Space





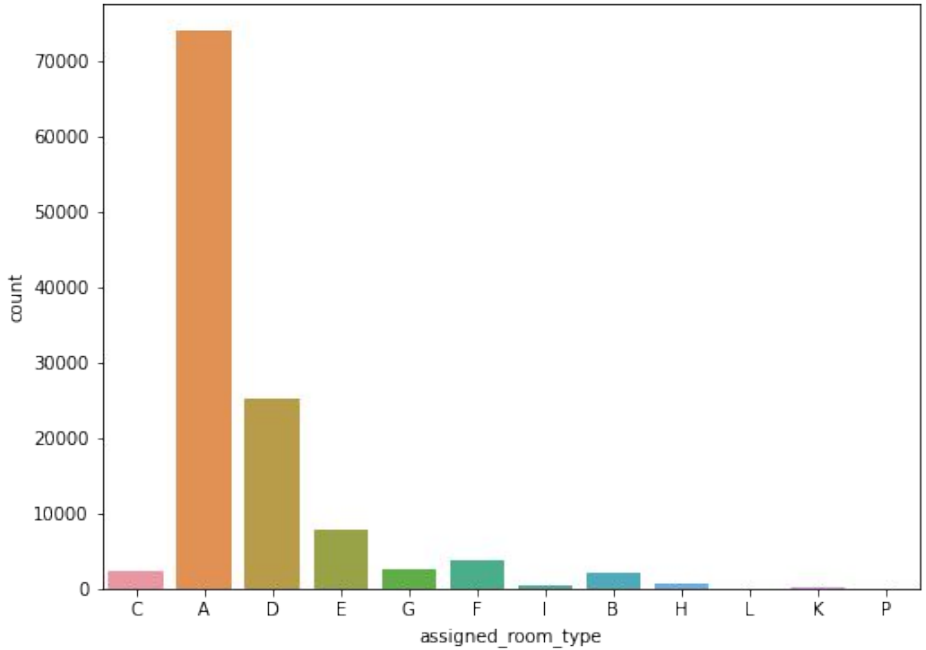
- In **year 2016 January** months required highest number of car parking space.
- In year 2015 September month required lowest number of parking space.
- Overall 93.9 % guests did not required the parking space.
- 6.1 % guests required only 1 parking space.

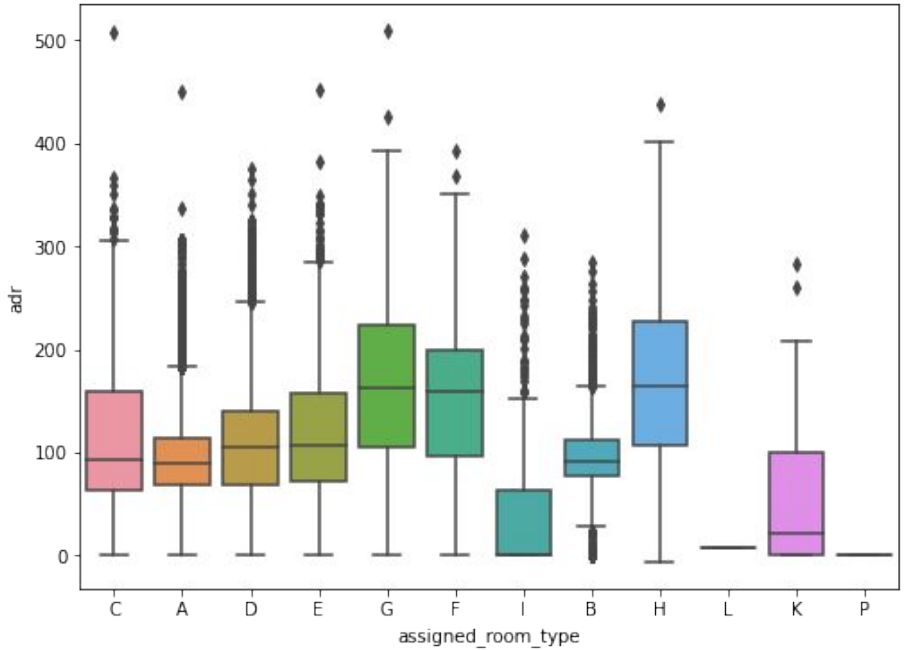
3. Number of Bookings from Different Countries



- 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)

4. Plotting in demand room and which room generate more ADR





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

Conclusion

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).

1. Most customers booked rooms online so they can be targeted with proper seasonal discounts and vacay-ads.

1. Since ADR was least during Nov and Jan, winter discounts(assumption) or off season discounts might help.

1. For retention, they should introduce Portuguese meals(sea foods and meat) and Eastern European meals as guests are more from there.

1. They should encourage direct bookings by offering some special discounts as online bookings cancellation is high.

1. 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.

1. Since resort hotels are less preferred, they should look into the factors- might be High cost or guests requirements.

Thank You