Capstone Project

Hotel Booking Analysis(EDA)

By Kishore kumar

Introduction

A few decades ago, traveling was not a part of everyday life. But today travel is an enormous budding industry of 8.8 trillion economy. This directly affects the hotel industry which is highly competitive.

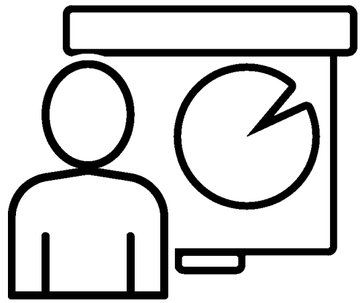
We are here with a compact data to study about the hotel industry, mainly the booking. We are focussing on two types of hotels in this study. This data set contains different hotel types, countries located, guest, stays. Also the study have some factors that affect booking like wait time, lead time , months, average daily rate etc.

By the end we will conclude the study with following insights:-

- Best time for booking
- Optimal duration of stay
- Distribution Segment and market segments to be focused to increase revenue
- Factors leading to cancellation which affects the revenue.
- Factors like meals, special requests etc. which might affect in the increase of ADR and revenue.

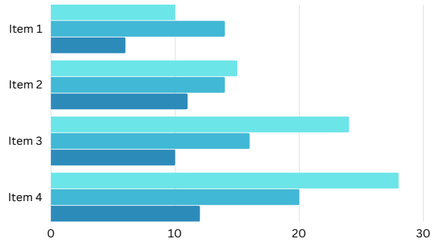
Steps involved in analysis

Data Collection & Inspection

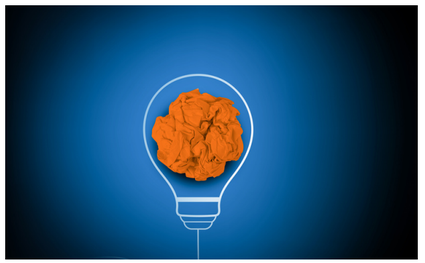


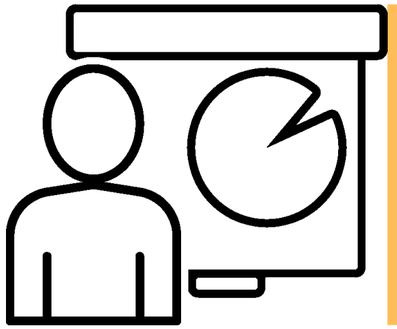
Data Cleaning & Manipulation

Exploratory Data Analysis



Conclusion





DATA COLLECTION
& INFORMATION

Features information:

This data has 119390 rows and 32 columns. Here are the columns:

- **hotel** : Hotel(Resort Hotel or City Hotel)
- **is_canceled** : Value indicating if the booking was canceled (1) or not (0)
- **lead_time** : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
- **arrival_date_year** : Year of arrival date
- **arrival_date_month** : Month of arrival date
- **arrival_date_week_number** : Week number of year for arrival date
- **arrival_date_day_of_month** : Day of arrival date
- **stays_in_weekend_nights** : Number of weekend nights (Saturday or Sunday) the guest stayed or booked hotel
- **stays_in_week_nights** : Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- **adults** : Number of adults
- **children** : Number of children
- **babies** : Number of babies
- **meal** : Type of meal booked. Categories are presented in standard hospitality meal packages:
- **country** : Country of origin.

Features information(contd):

- **market_segment** : Market segment designation.(TA/TO)
- **distribution_channel** : Booking distribution channel.(TA/TO)
- **is_repeated_guest** : 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** : Number of previous bookings not cancelled by the customer prior to the current booking
- **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 from the moment the booking was entered on the PMS until the moment of check-in or cancellation
- **deposit_type** :No Deposit,Non Refund,Refundable
- **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
- **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

Data Briefing

The dataset is quite large and upon using the “shape” method I am able to find the initial number of rows and columns present in the dataset.

Now, regarding the extraction of the data information from the dataset, I used “info()” method which gave me the information about data types(dtypes), count of non-null values & memory usage by the dataset.

- There were 4 columns with float64 dtypes, 12 columns with object dtypes & 16 columns with int64 dtypes.

I used “column” method to get the list of all the column label names and used “isnull()” method to get the count of null values if any in the dataset.

- 4 null values found in ‘children’ column as total 0.003% missing values.
- 488 null values found in ‘country’ column as total 0.409% missing values.
- 16340 null values found in ‘agent’ column as total 13.686% missing values.
- 112593 null values found in ‘company’ column as total 94.307% missing values.

Data summary:

- is_canceled
- lead_time
- arrival_date_year
- arrival_date_week_number
- arrival_date_day_of_month
- stays_in_weekend_nights
- stays_in_week_nights adults
- children
- babies
- is_repeated_guest
- previous_cancellations
- previous_bookings_not_canceled
- booking_changes
- agent
- company
- days_in_waiting_list
- adr required_car_parking_spaces total_of_special_requests

Numeric Variable

Data set

Categorical Variable

- hotel
- arrival_date_month
- meal
- country
- market_segment
- distribution_channel
- reserved_room_type
- assigned_room_type
- deposit_type
- customer_type
- reservation_status
- reservation_status_date

Data Cleaning & Manipulation

Data Wrangling

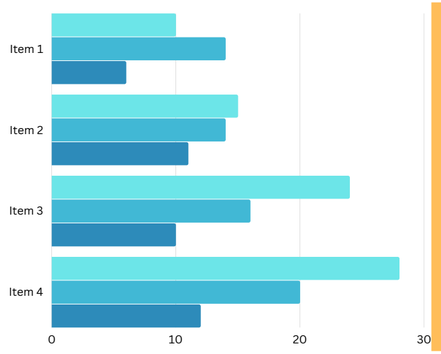
Data Cleaning: There were 4 columns company, agent, country and children with missing values.

- Dropping Columns with having Maximum Null values– agent and company
- Columns with nominal null values have been manipulated by filling them with
  - ➔ Numerical column:- Mean (Number of children)
  - ➔ String column:- Mode (Country)

Handling Duplicates: Data had 32020 duplicates values, so we dropped it from data.

Data Manipulation:

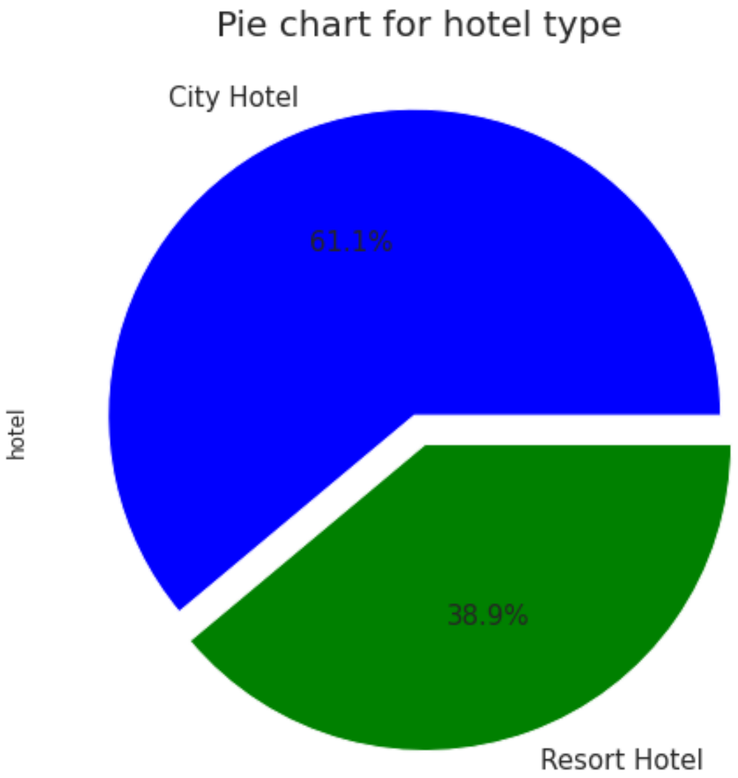
- Combining columns for an effective study
  - ➔ Total_members = adults + children + babies
  - ➔ total_night_stays = stays_in_weekend_nights+ stays_in_week_nights



EXPLORATARY DATA ANALYSIS(EDA)

Univariate Analysis:

1. Which type of hotel is mostly prefered by the guests?



Observation:

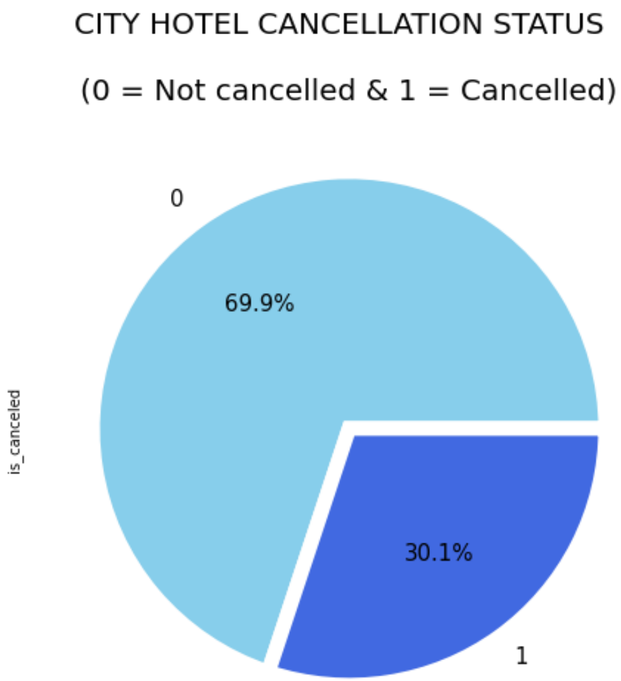
City Hotel is most preffered by guests. Thus city hotels has maximum bookings.

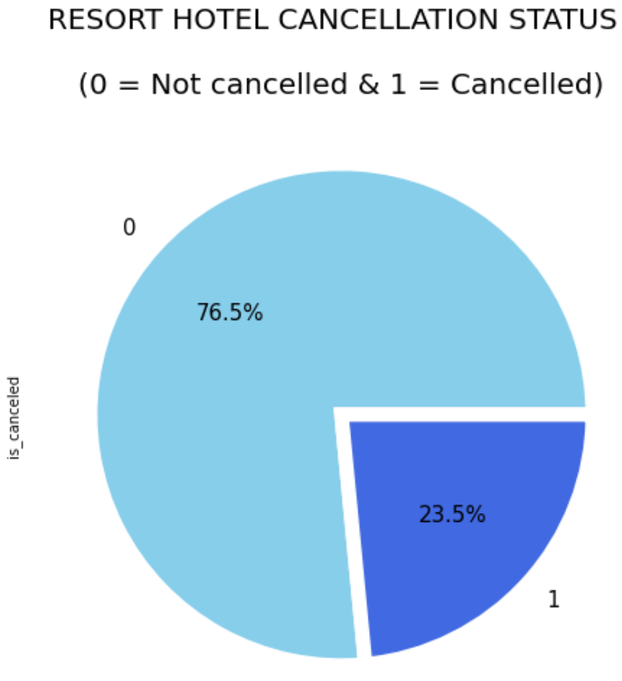
Total both Type of hotel:

- City Hotel 61.07518 %
- Resort Hotel 38.92482 %

Univariate Analysis:

(2) What is the pecentage of cancellation in City Hotel and Resort hotel?





Observation:

Maximum cancellation of city hotel as below:

cancellation_city_hotel:

- Not Cancelled(0) - 69.89%
- Cancelled(1) - 30.11 %

Observation:

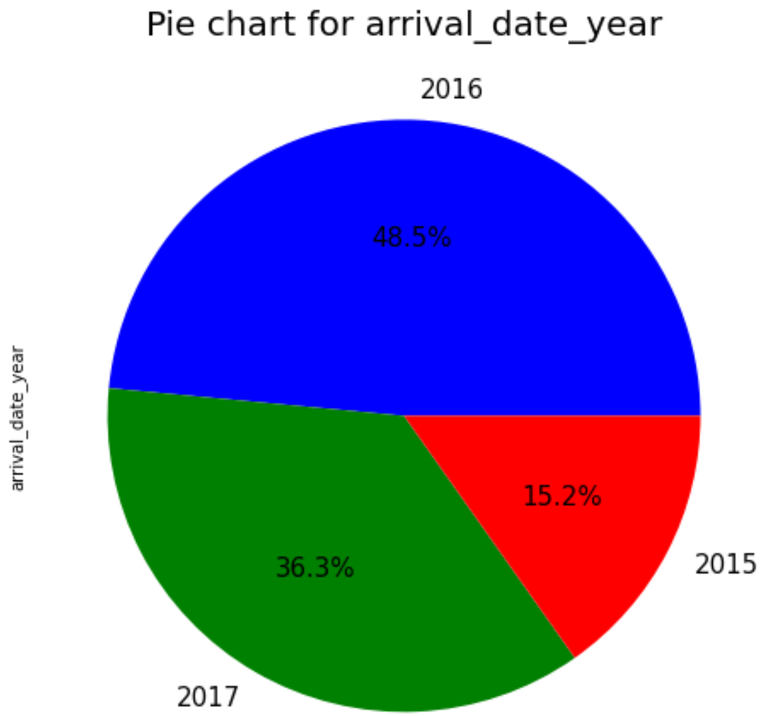
Maximum cancellation of resort hotel as below:

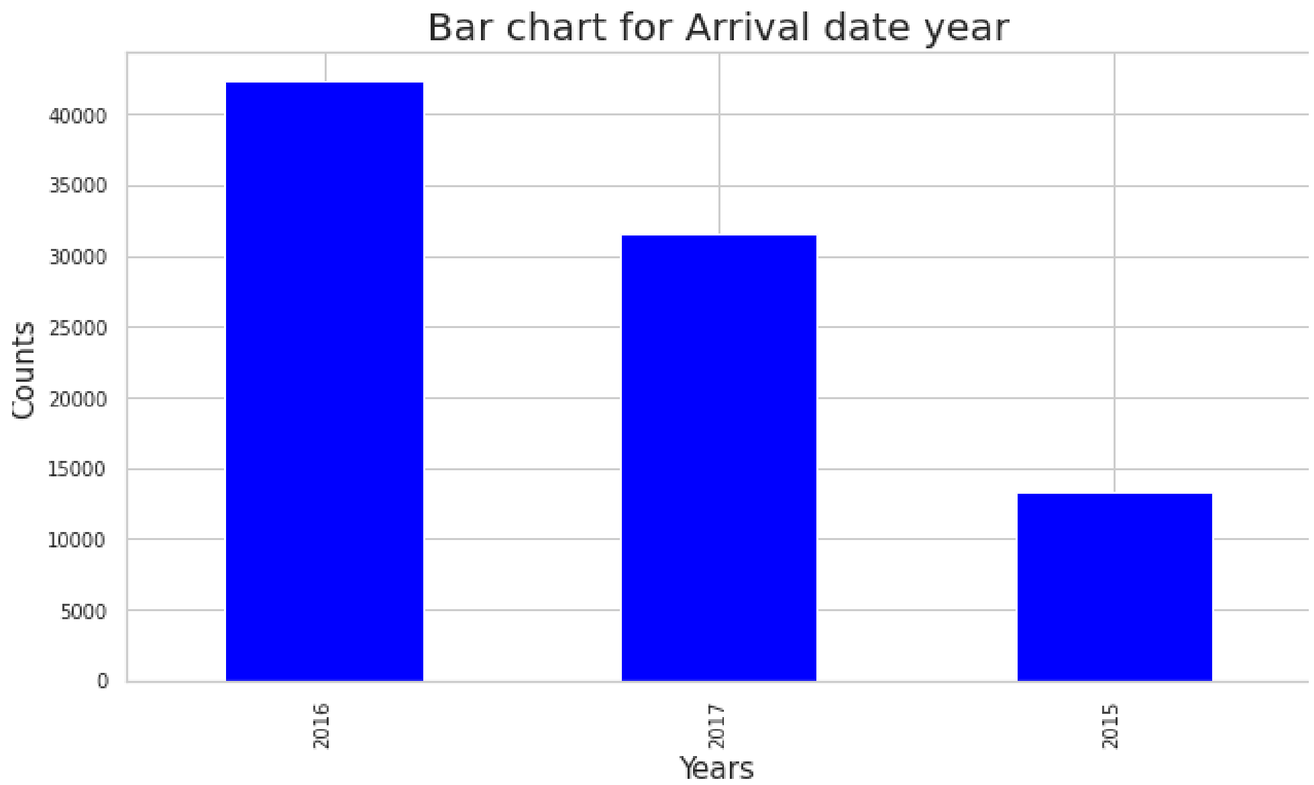
cancellation_Resort_hotel:

- Not Cancelled(0) - 76.50%
- Cancelled(1) - 23.50%

Univariate Analysis:

(3) Which year had the highest bookings of hotels?



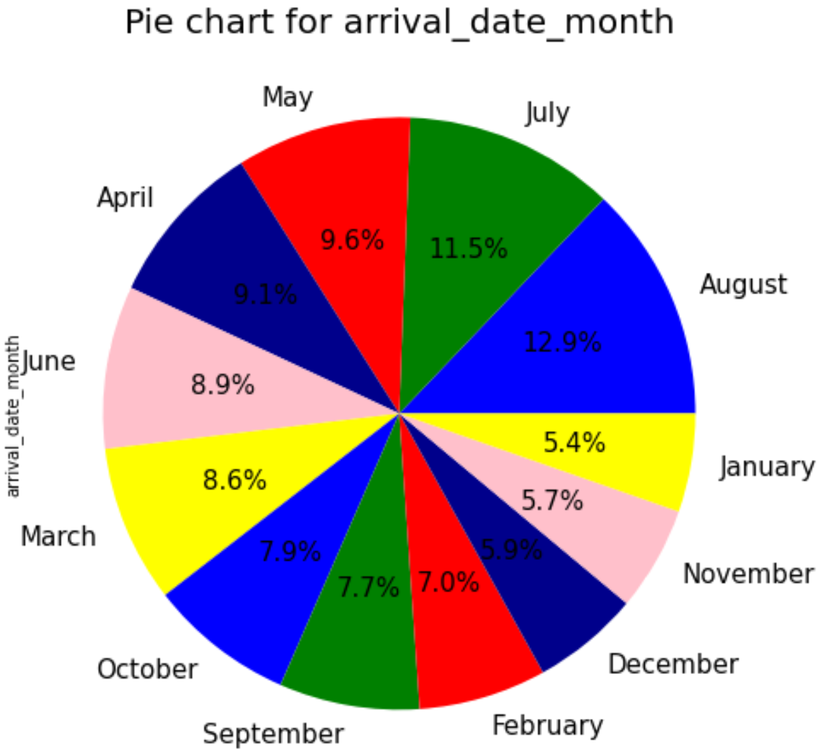


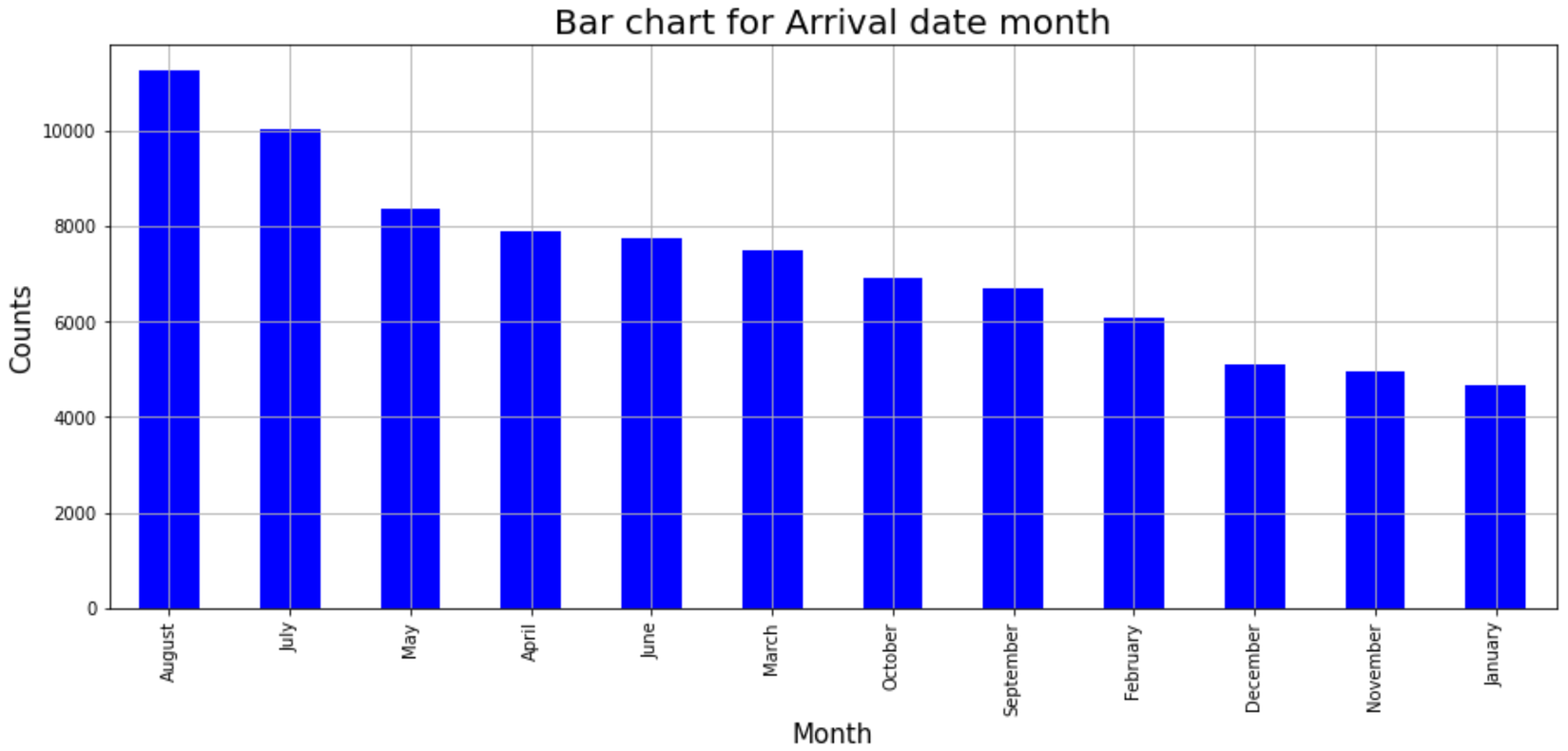
Observation:

- 2016 had the 48.6% highest bookings which having more than 40000 number of guests.
- 2015 had 15.2% less bookings which having 13000 number of guests.
- overall City hotels had the most of the bookings

Univariate Analysis:

(4) In which month most of the bookings happened?





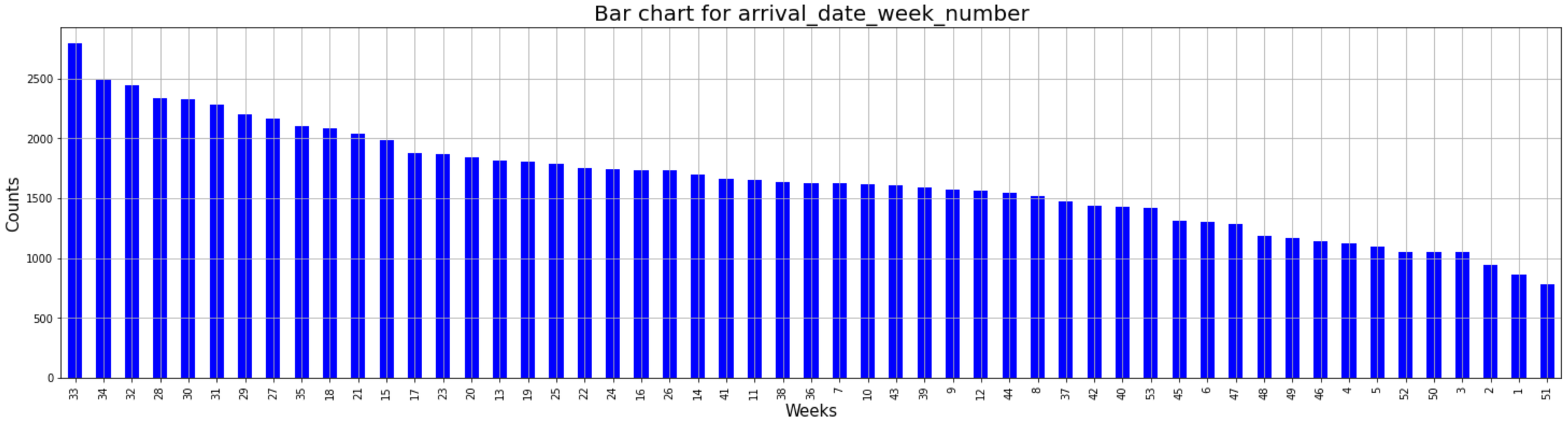
Observation:

Mostly July & August month have more booking the hotels as compared to January month by guest as below:

- July - 11.50% - 10000
- August - 12.9% - 10000+
- January - 5.4% - 4800

Univariate Analysis:

(5) In which week most of the bookings happened?



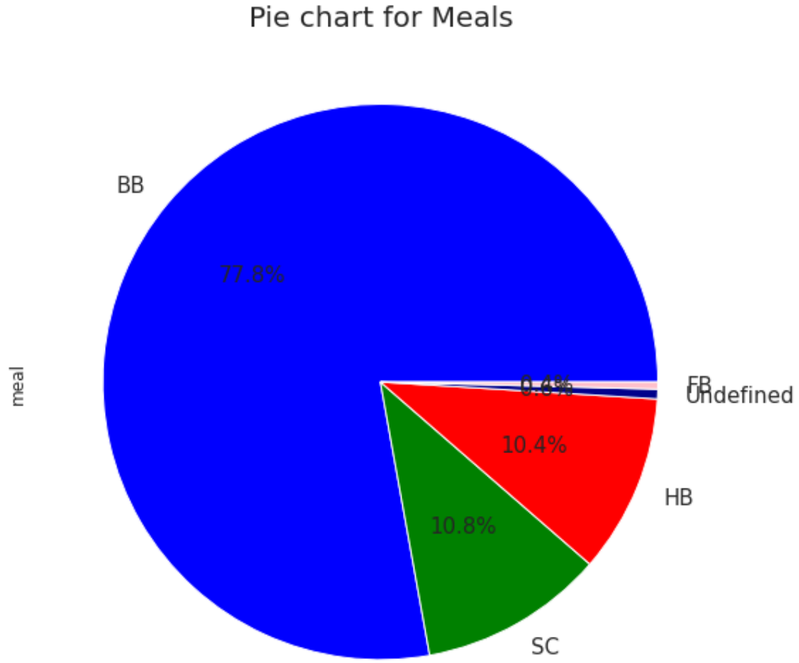
Observation:

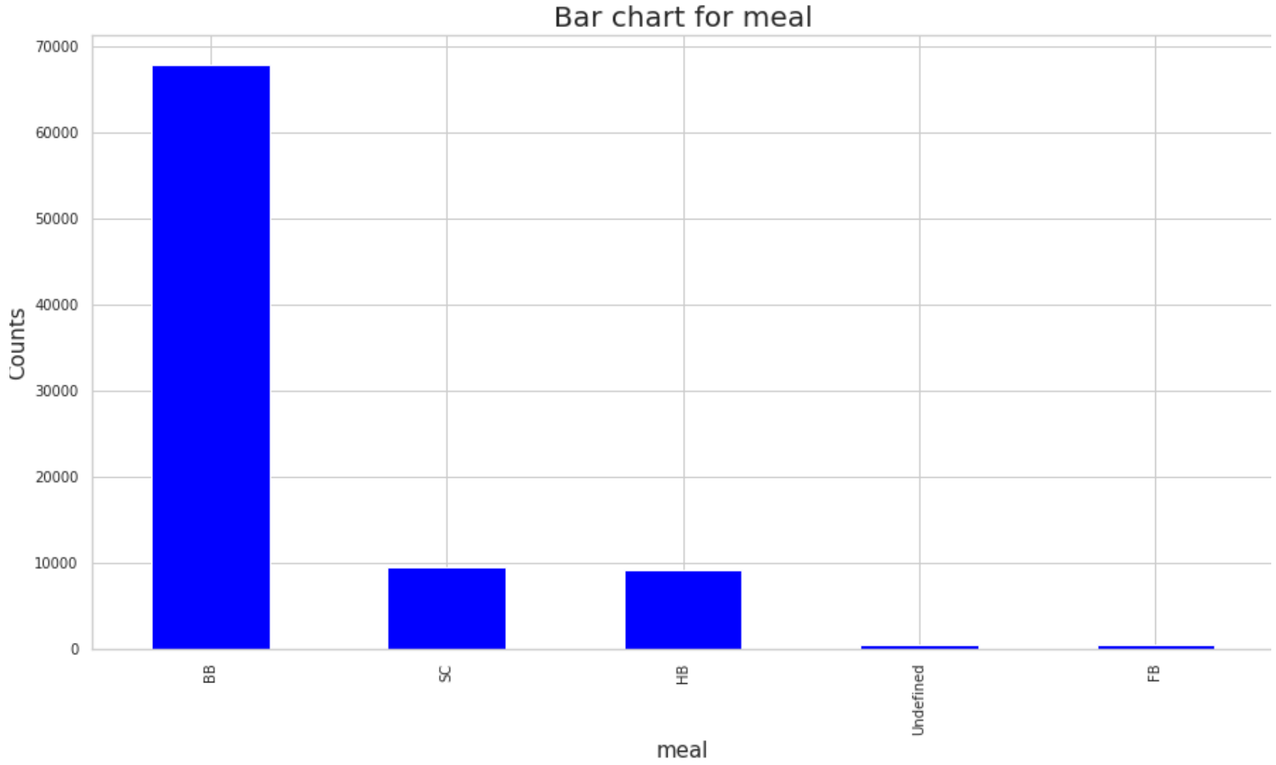
Mostly 33th week having most of the booking and 51th week having least booing see below:

- 33 weeks-3.2% heights booking.
- 51 week-0.9% leasts bookings.

Univariate Analysis:

(6) Which type of food is mostly preferred by the guests?



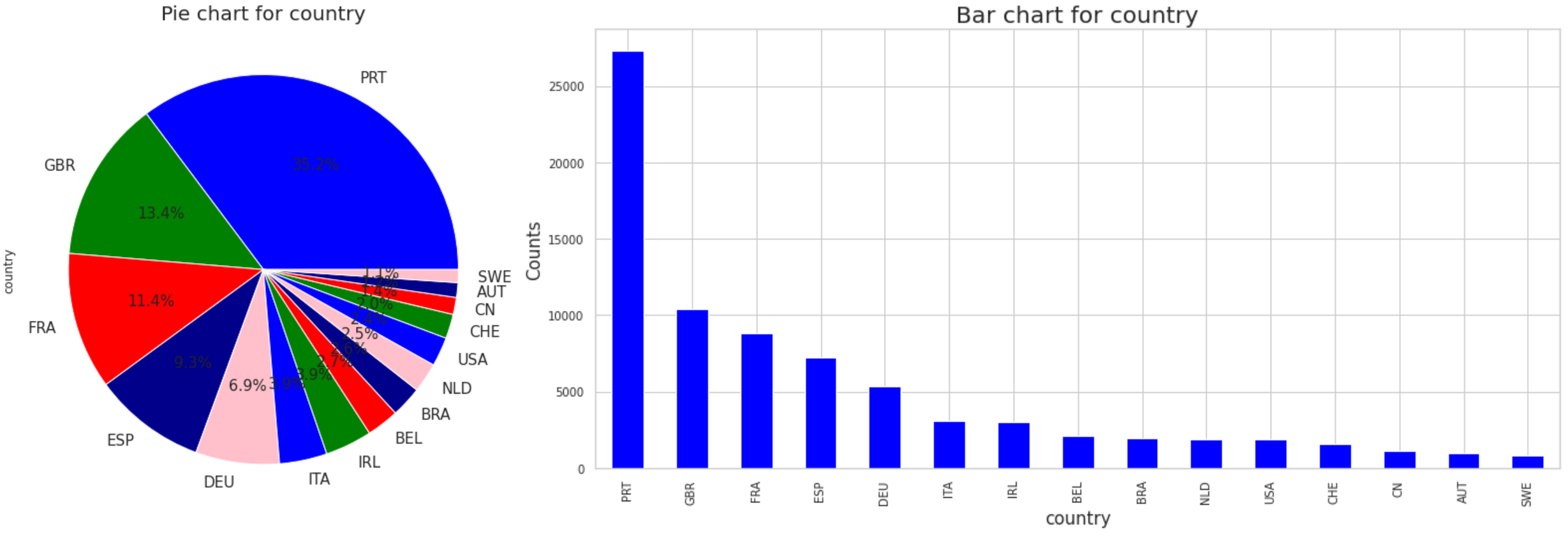


Observation:

- So the mostly preferred by the food type by the guests is BB(Bed and Breakfast)-77.8%
- HB- (Half Board) and SC- (Self Catering) are equally preferred.
- Types of meals available in hotels & mostly meals preferred by guests in percentages :
  ➔BB - (Bed and Breakfast) - 77.84%
  ➔HB- (Half Board) - 10.41%
  ➔ FB- (Full Board) - 0.41%
  ➔ SC- (Self Catering) - 10.41

Univariate Analysis:

(7) From which country the most guests are coming?

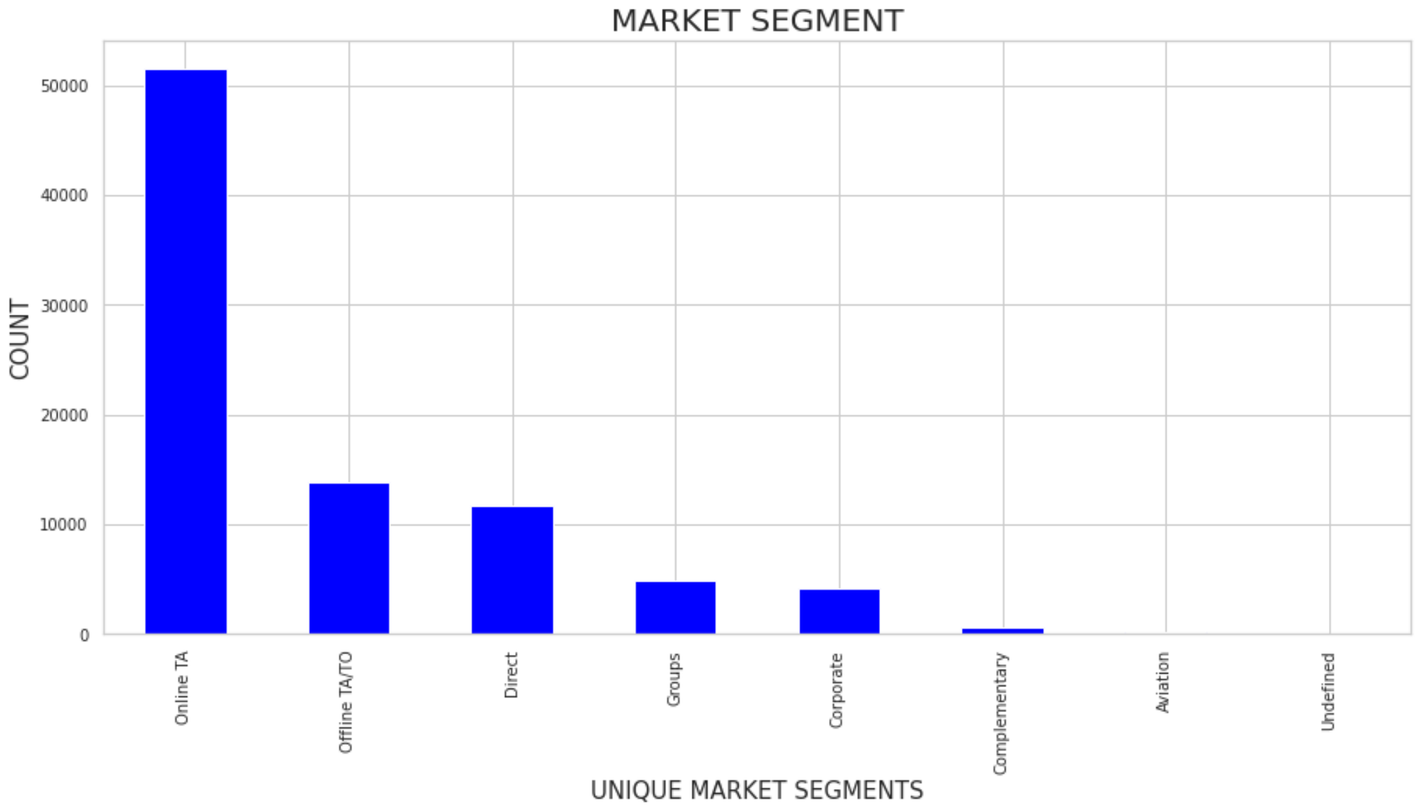


Observation:

- Most of the guests are coming from Portugal(PRT) i.e. 35.20% guests are from portugal.
- Least guests are coning from Sweden(SWE) i.e. 1.1% only.

Univariate Analysis:

(8) Which type of market segments are used by guest?

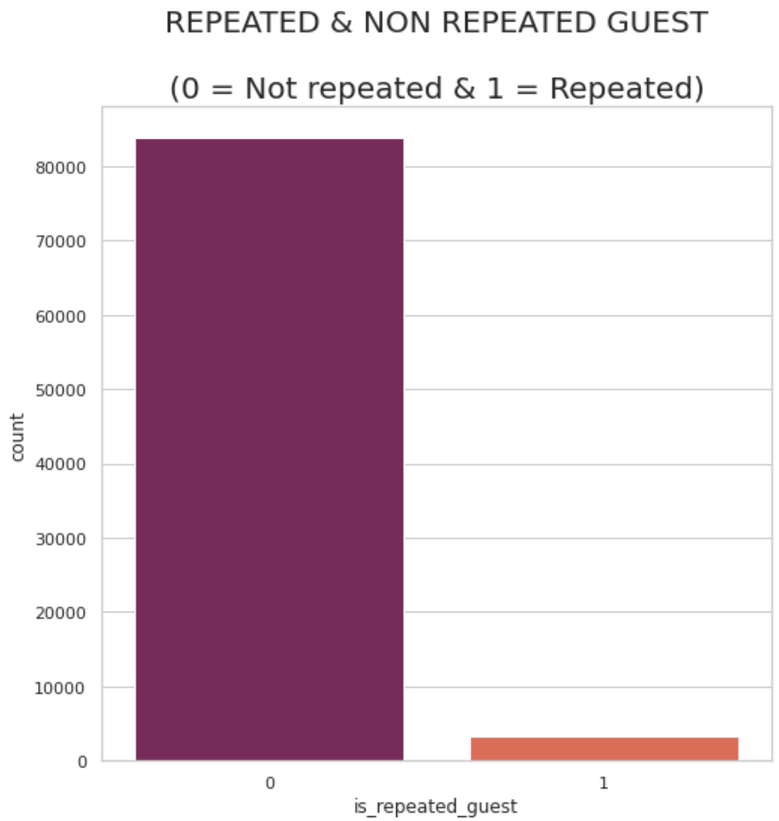


Observation:

- As per above Mostly Online TA(Travel Agents) market segmentation available-51544 Nos.

Univariate Analysis:

(9) What is the Percentage of repeated guests?

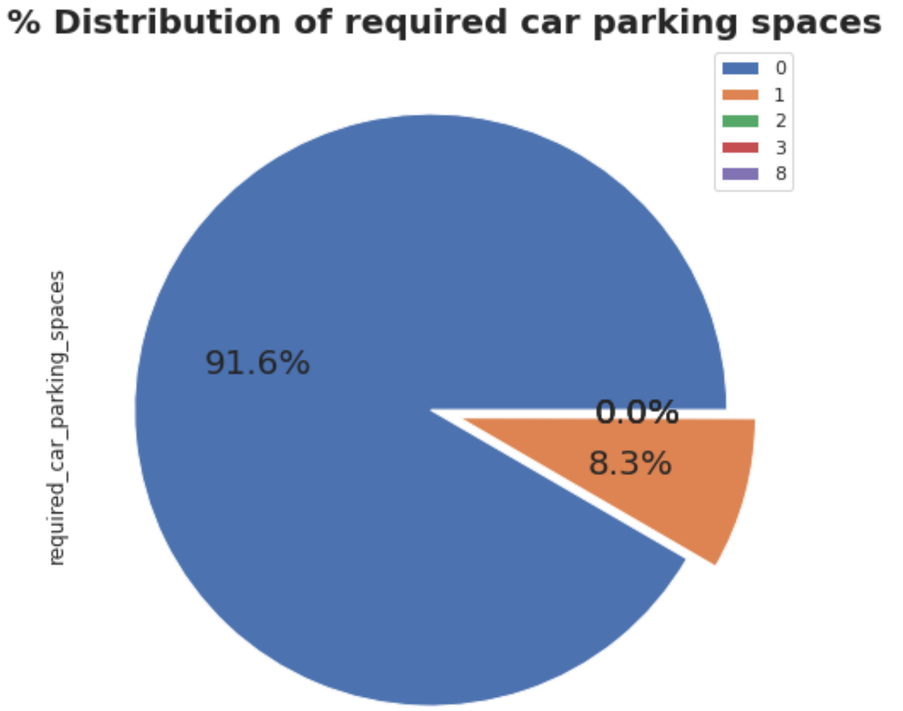


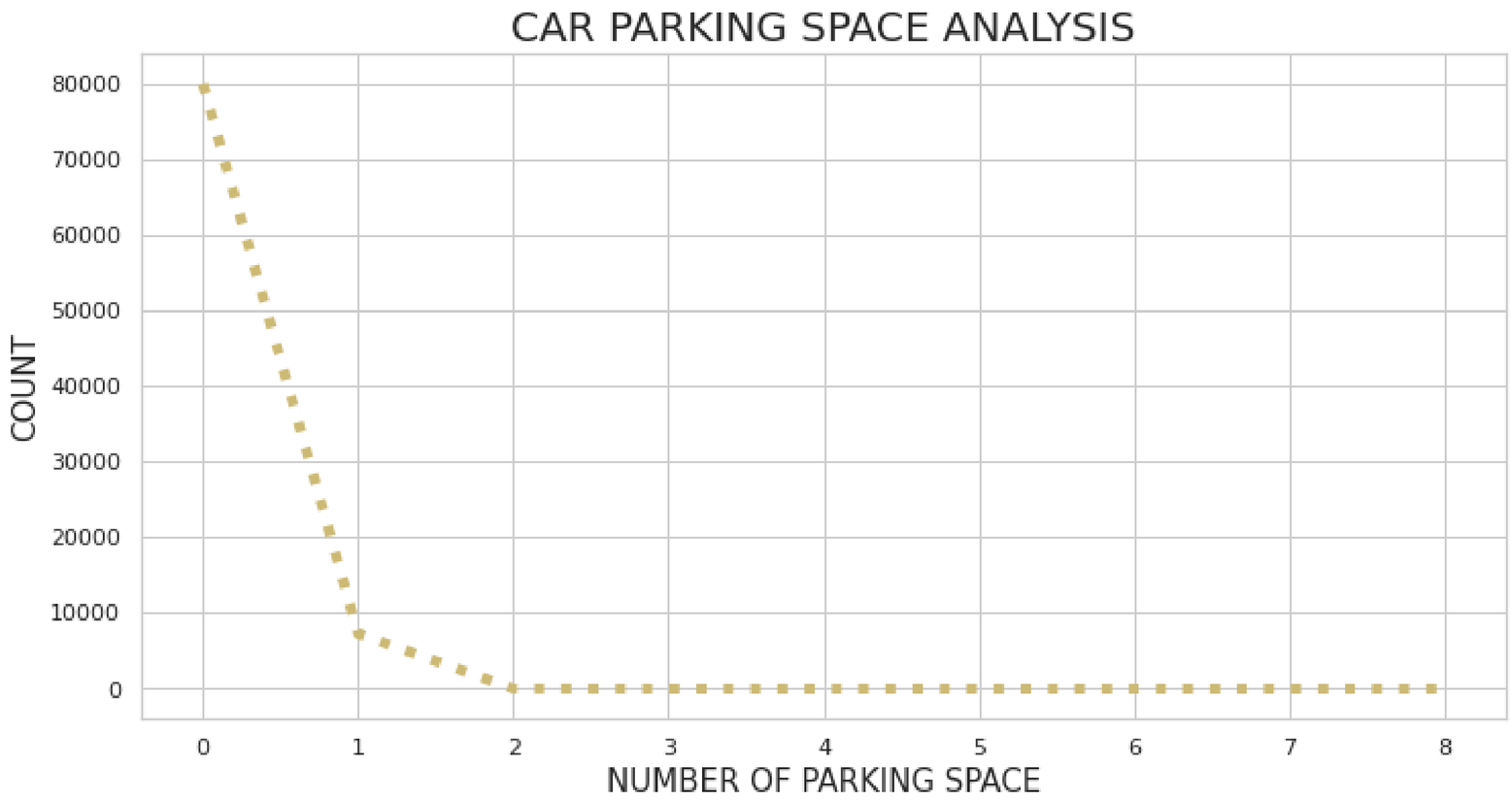
Observation:

- Maximum repeated guest are coming for hotel as below:
- is_repeated_guest:
  - ➔ Not repeated(0) - 96.14%
  - ➔ Repeated(1) - 03.85 %
- 3364 people are repeaated guests. The retention guest rate is very low.

Univariate Analysis:

(10) What is the percentage distribution of required_car_parking_spaces?





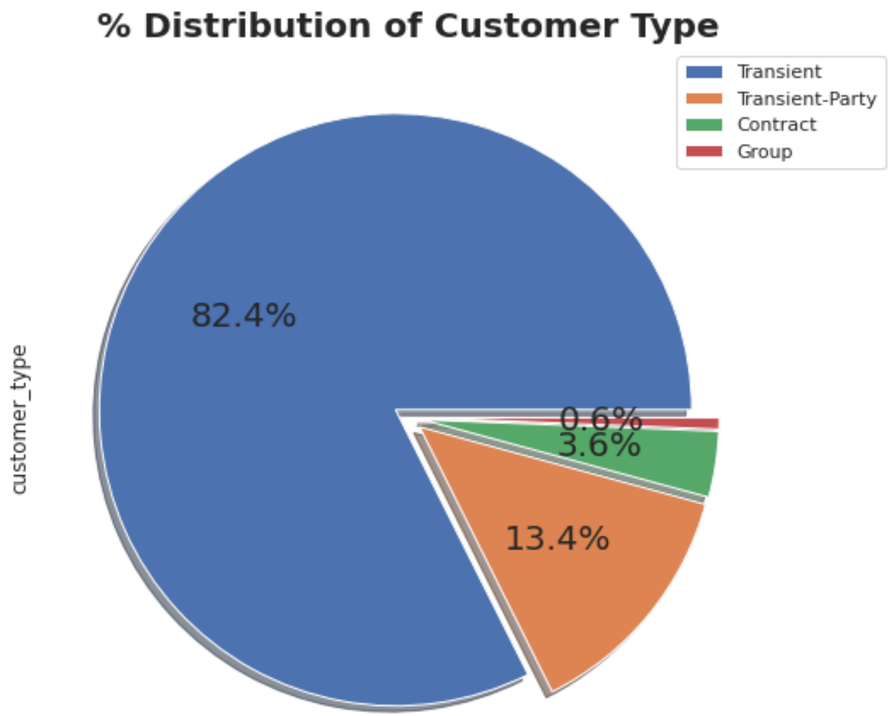
Observation:

Mostly of guests having not required car parking spaces.

- Parking space not required - 91.60%.
- Parking space required - 08.30%.

Univariate Analysis:

(11) What is the percentage distribution of "Customer Type"?

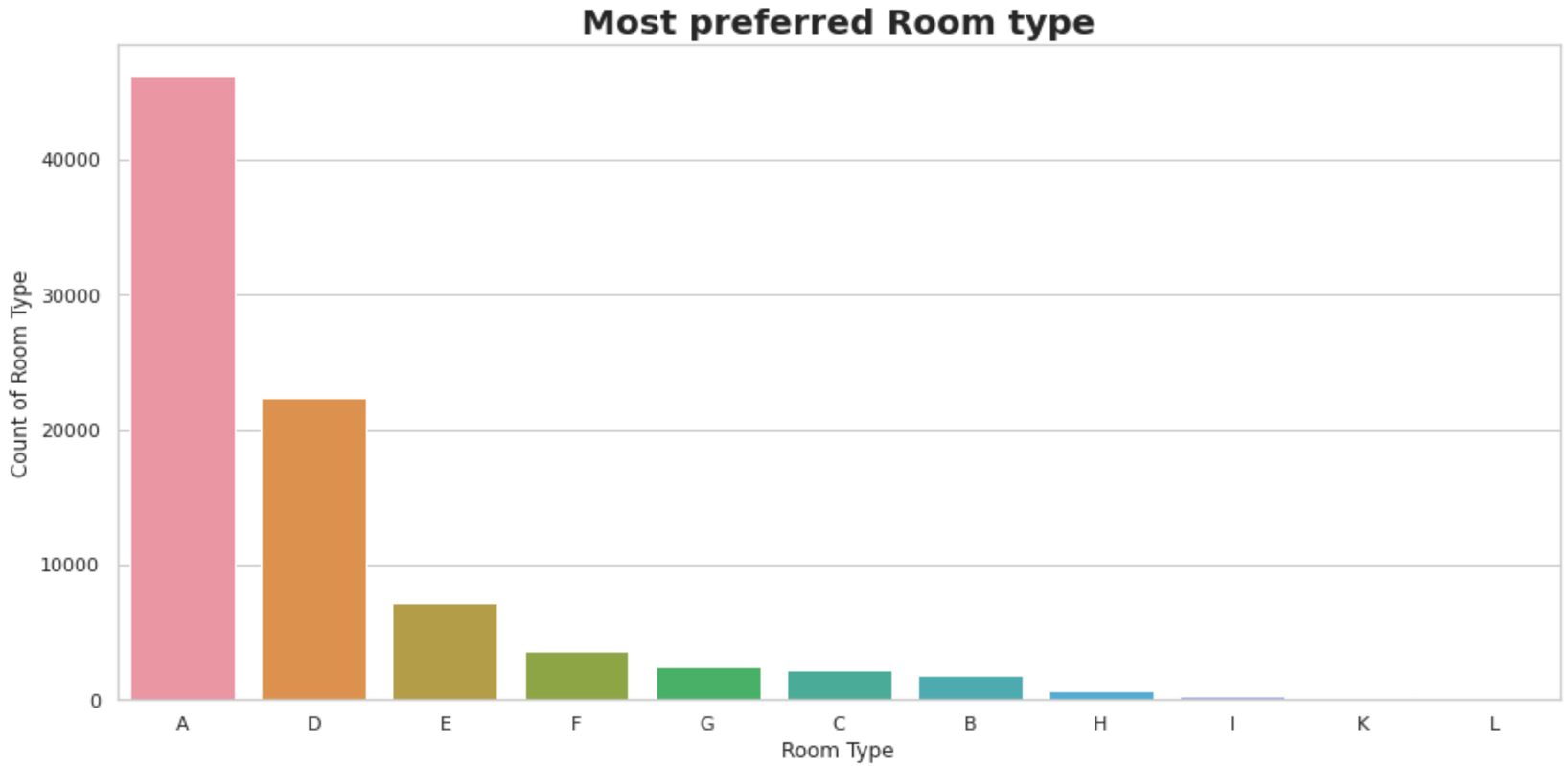


Observation:

- Transient customer type is more which is 82.4 %.
- Group customer type is vey low- 0.6%.

Univariate Analysis:

(12) Which is the most preferred room type by the customers?

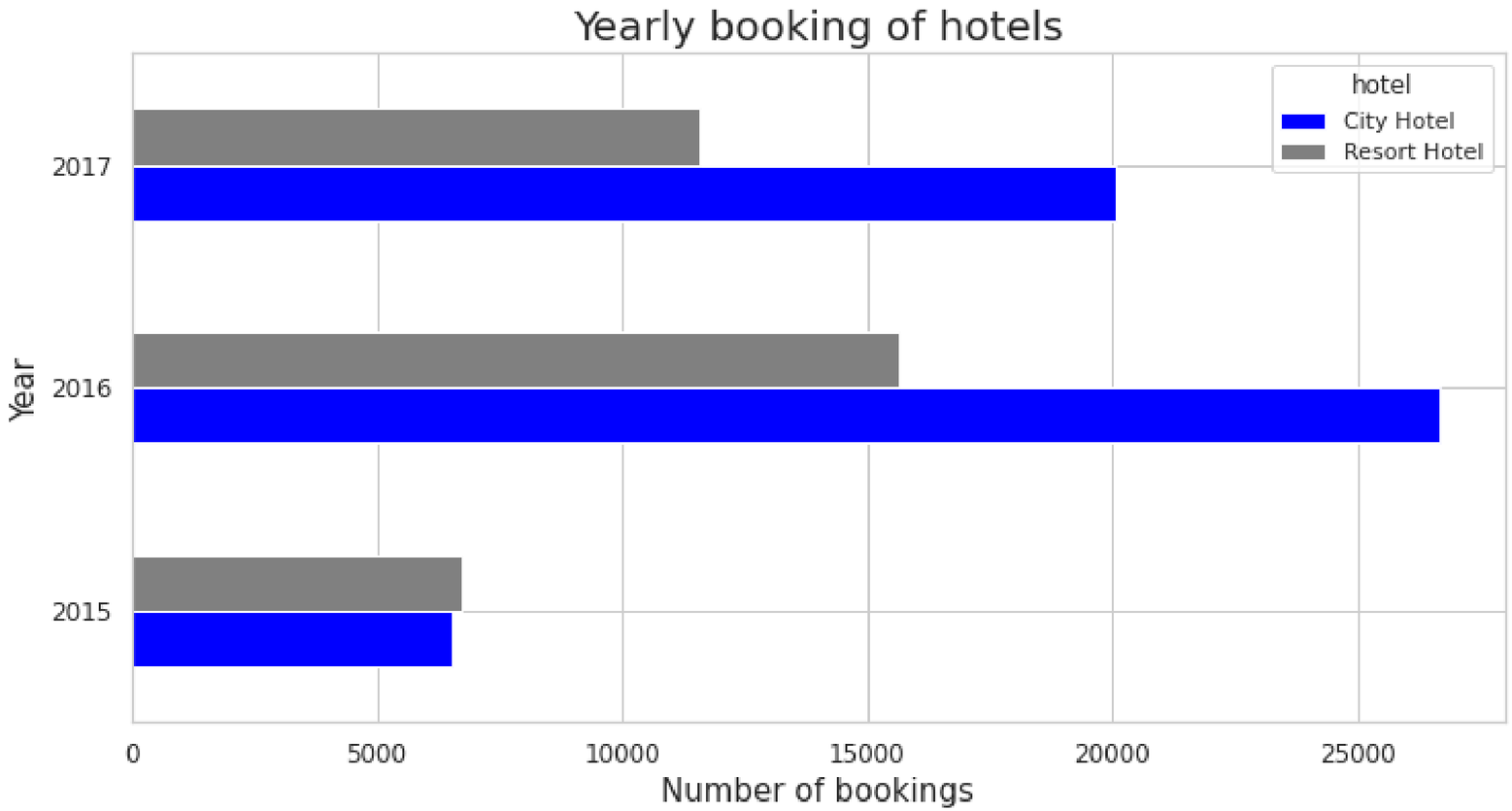


Observation:

- So the most preferred Room type is "A".

Bivariate Analysis:

(1) Which type of hotel is mostly guests comes in a year or booking in a year?



Observation:

- In 2016 year mostly guests are booking by:
  - ➔ City hotel- 26677
  - ➔ Resort hotel- 15626

Bivariate Analysis:

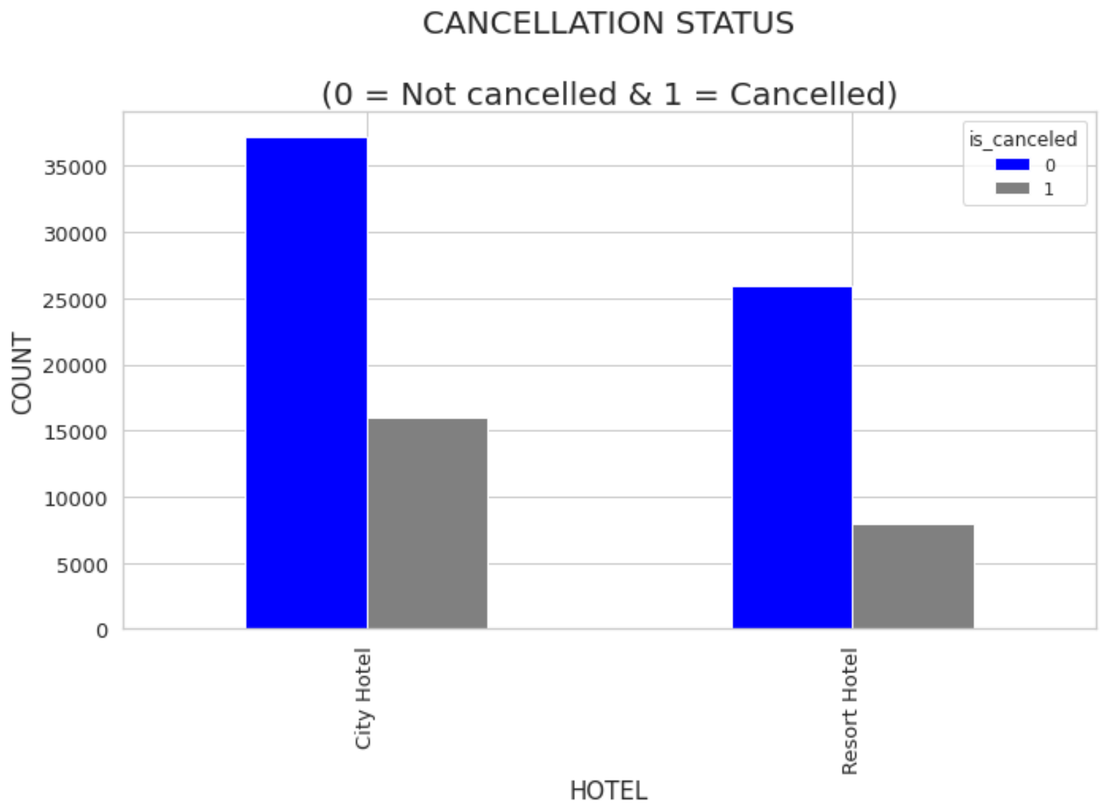
(2) Which type of hotel is mostly guests comes in a month or books in a month?

Observation:

- In July & August month maximum guests are booking by:
- July Month:
  - ➔ City hotel- 5727
  - ➔ Resort hotel- 4312
- August Month:
  - ➔ City hotel- 6575
  - ➔ Resort hotel- 4664

Bivariate Analysis:

(3) Which type of hotel has highest booking cancellation?

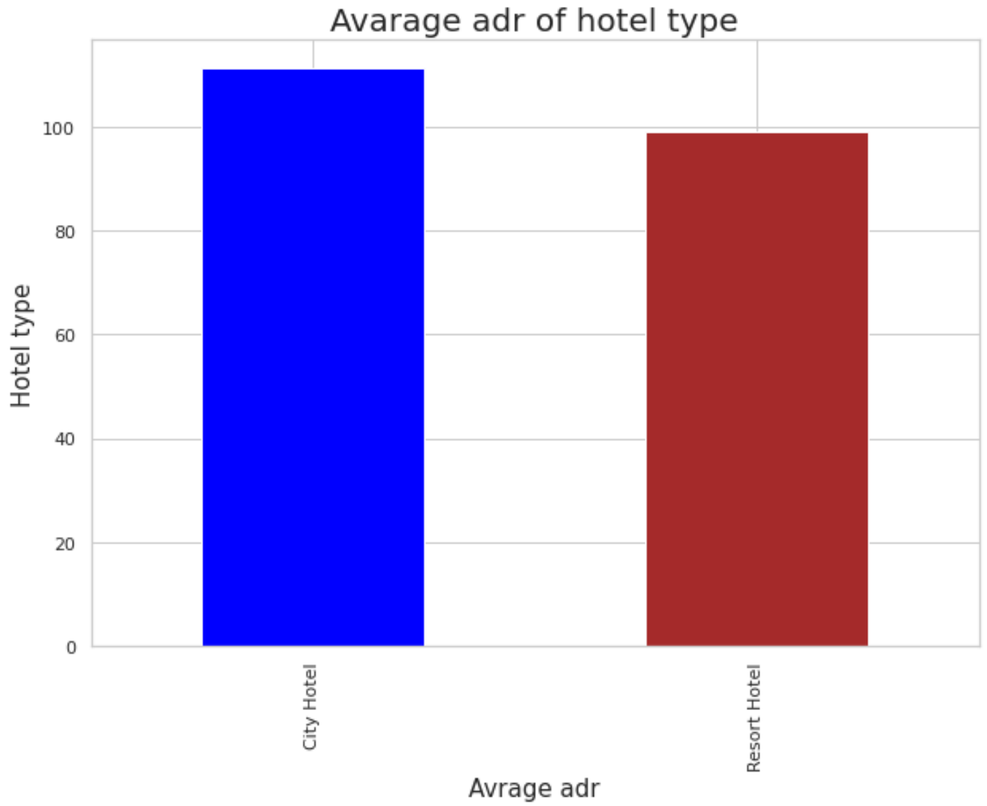


Obsevation:

- Maximum city hotel has 37227 of booking cancellation done.
- Not cancelled:
  - ➔ City Hotel - 37227
  - ➔ Resort Hotel - 25972
- Cancelled:
  - ➔ City Hotel - 16033
  - ➔ Resort Hotel - 7932

Bivariate Analysis:

(4) Which type of hotel has the highest ADR?

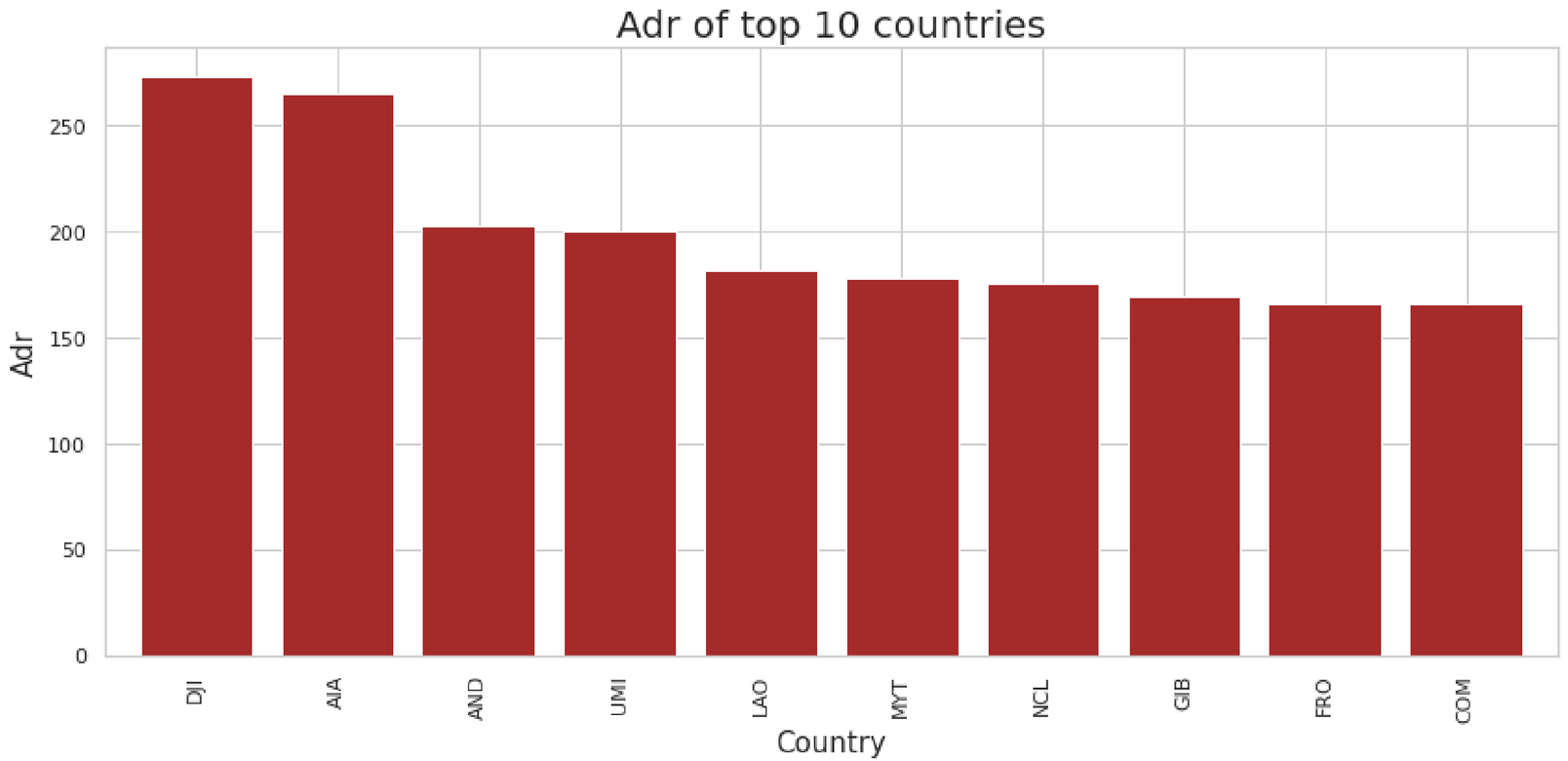


Observation:

- City hotel has the highest ADR. That means city hotels are generating more revenues than the resort hotels. More the ADR more is the revenue.
- ADR:
  - ➔ City hotel- 111.28
  - ➔ Resort hotel- 99.08

Bivariate Analysis:

(5) Which country has the highest ADR?

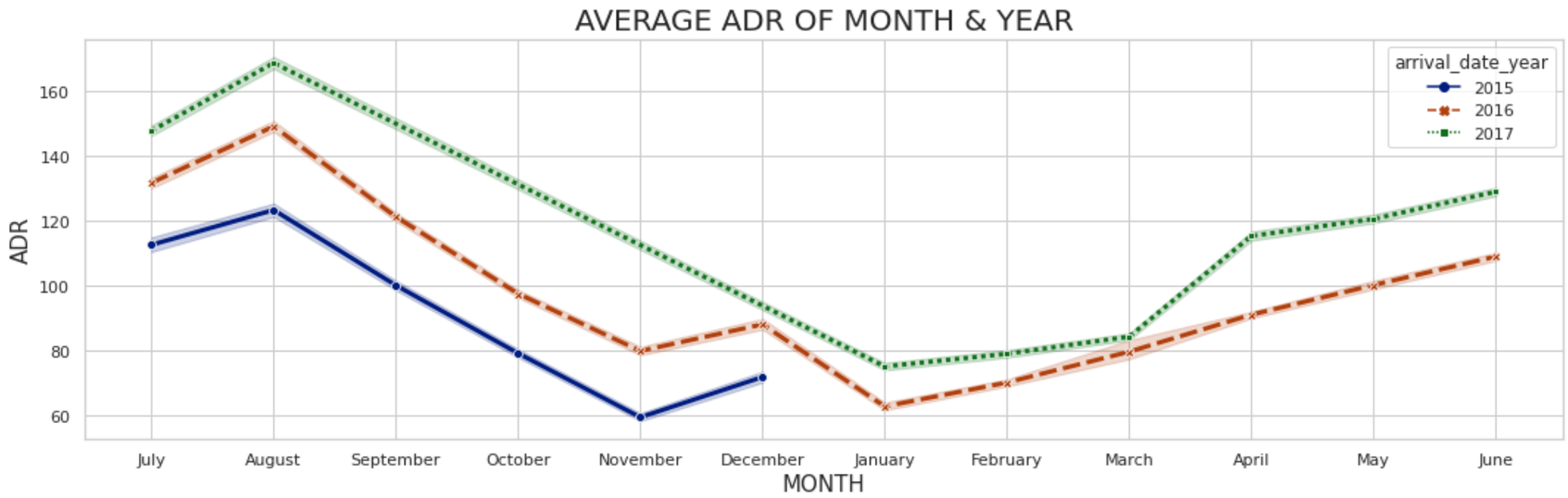


Observation:

- DJI has the highest ADR. More the ADR more is the revenue.

Bivariate Analysis:

(6) Which year & month has the highest ADR?



Observation:

- The highest ADR as mention below, More the ADR more is the revenue.
- In August month:
  - ➔ 2015 Year -123.29
  - ➔ 2016 Year -149.04
  - ➔ 2017 Year -168.60

Bivariate Analysis:

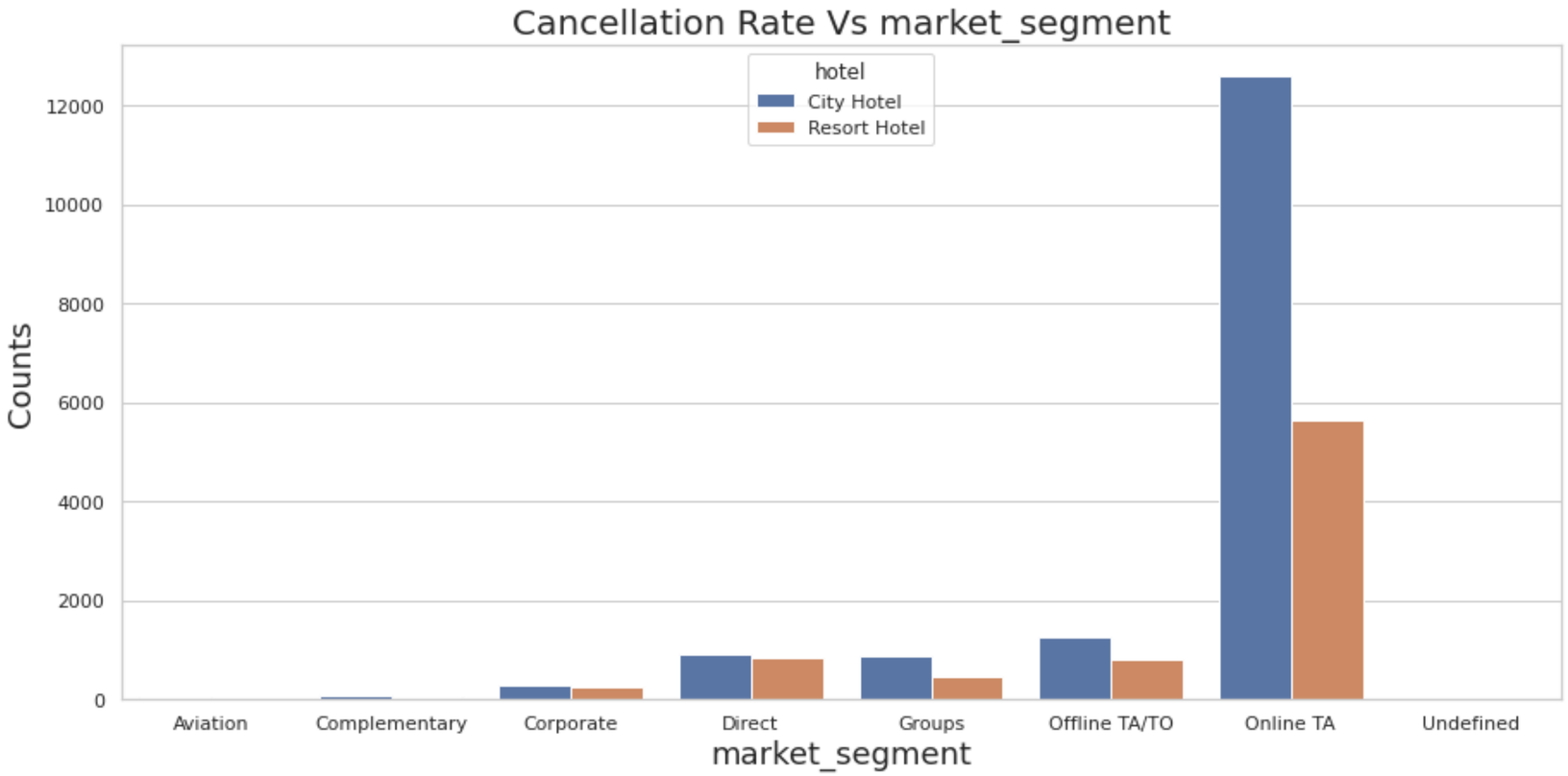
(7) Which Hotels has the most repeated guests?

Observation:

- Resort Hotel has slightly more repeated guests than the City Hotels. It is almost similar for both hotels.

Multivariate Analysis:

(1) Which Market Segment has the highest cancellation rate?

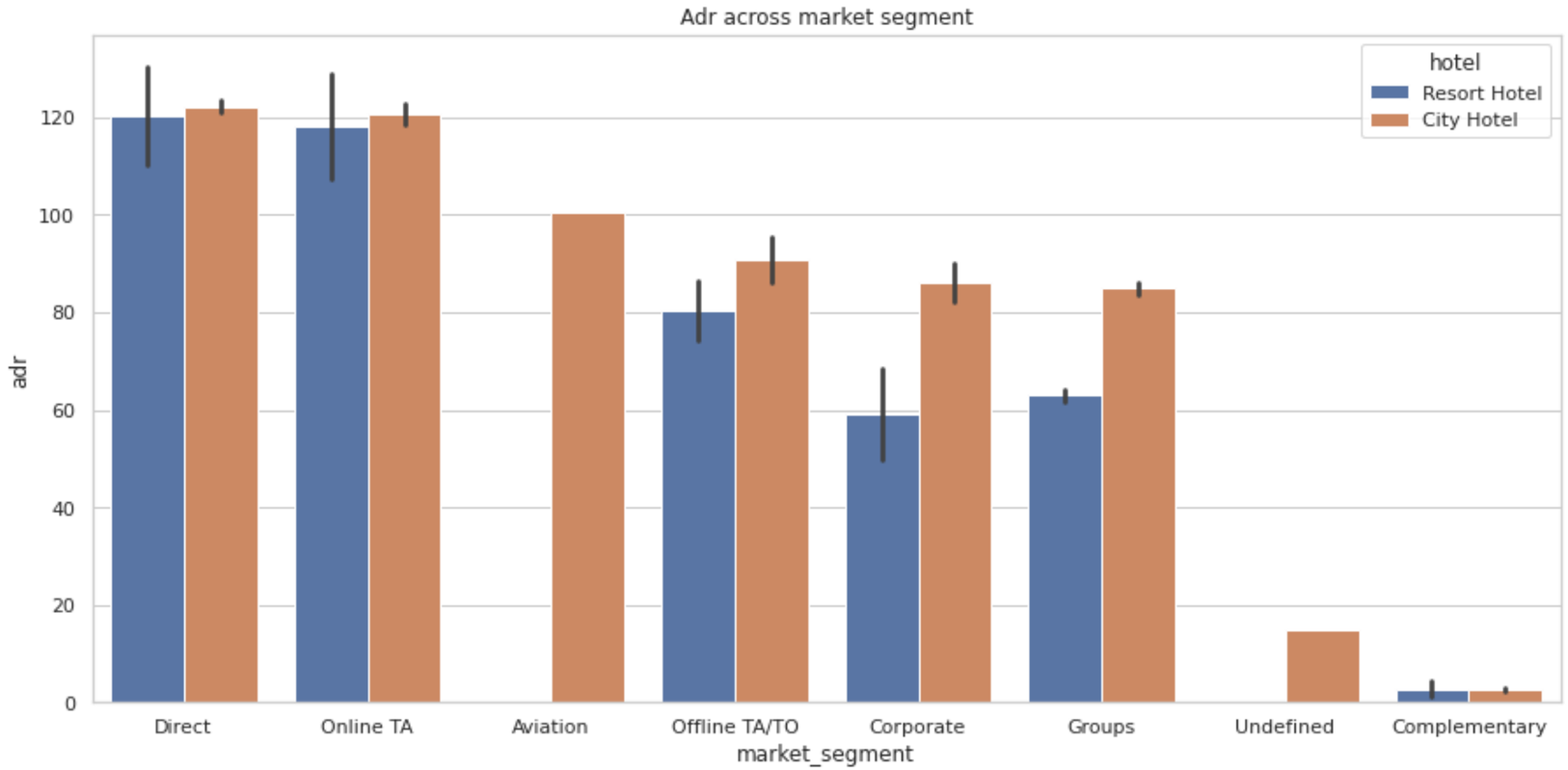


Observation:

- '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.

Multivariate Analysis:

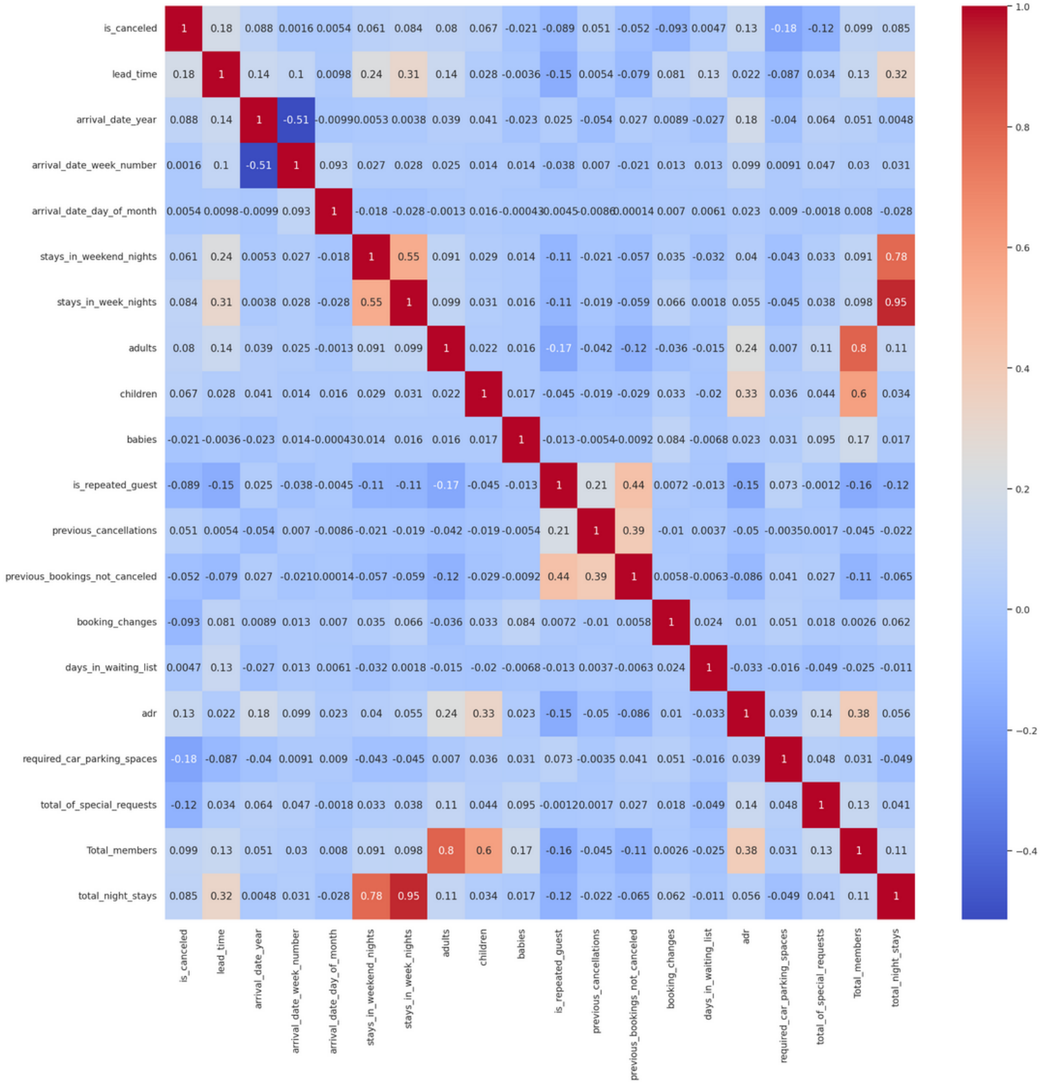
(2) ADR across different market segment?



Obseravtion:

- 'Direct' and 'Online TA' are contributing the most in both types of hotels.
- Aviation segment should focus on increasing the bookings of 'City Hotel'.

Correlation Analysis:



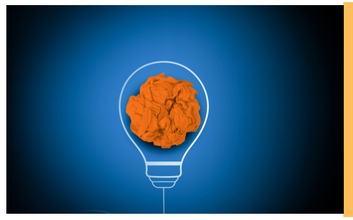
Observation:

- is_canceled and same_room_alloted_or_not are negatively corelated. That means customer is unlikely to cancel his bookings if he don't get the same room as per reserved room. We have visualized it above.

- lead_time and total_stay is positively corelated.That means more is the stay of cutsomer more will be the lead time.

- adults,childrens and babies are corelated to each other. That means more the people more will be adr.

- is_repeated guest and previous bookings not canceled has strong corelation. may be repeated guests are not more likely to cancel their bookings.



Conclusions

Conclusions:

- 'City hotels' and 'Resort hotels' are two types of hotels present in the dataset, out of which, 'City hotels'are more preffered by the customers than the latter. (61.1% customer prefers 'City hotels' whereas 38.9% customer prefers 'Resort hotels').

- Dataset contains booking data of 3 different years(2017,2016,2015), out of which, maximum hotel bookings took place in 2016 & 2015 witnessed the least number of hotel bookings.

- Out of all months, 'August' witnessed highest number of hotel bookings whereas 'January' witnessed the least.

- Among all the countries in the dataset, PRT(Portugal) has got the maximum number of hotel bookings.

- Its observed that 'City hotels' was more cancelled as compared to 'Resort hotels'.

- Coming to the analysis of market segment, 'Online TA' brings maximum bookings.

- The value of ADR of hotels is approximately 111.10.

- Average ADR of 'City hotels' is more than that of 'Resort hotels'. 'DJI' has the highest average ADR among all countries.

Conclusions(contd):

- Considering all the three years, 'August' has got the highest average ADR in each year.
- Its observed that average ADR is incrementing every year from 2015 to 2017, which clearly states that hotel business is scaling up every year.
- After analysing the meal data, its found that 77.84% of customers prefers BB(Bread & Breakfast).
- Most of the hotels have 0 to 1 car parking space.
- etc..

THANK YOU!