**Exploratory Data Analysis on Hotel Booking Analysis**

**Rushikesh shete**

**Data Science Trainee**

**Alma Better**

**Abstract:**

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal

length of stay 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?

Exploratory Data Analysis helps us in understanding the data and in exploring some

questions related to users on the data provided.

**Keywords:** Numpy, Pandas, EDA, Data Frames, Visualisations

**1. Problem Statement:**

This Dataset contains data that compares various booking information between two hotels, City

Hotel and Resort Hotel. We will be using the data to analyse the factors affecting the hotel

bookings. These factors can be used for reporting trends and predicting future bookings.

● hotel– There are two types of hotels one is City Hotel and another is Resort Hotel.

● Is\_canceled– Here 0 and 1 value indicates booking was cancelled(1) or not(0).

● lead\_time– Time-lapse between reservation and actual arrival date.

● 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(sat sun) guests are staying at

hotel.

● Stays\_in\_week\_nights– Number of week nights(mon-fri) guests are staying at the hotel.

● adults – Number of adults

● children– Number of children.

● babies – Number of babies.

● meal– Type of food provided at hotel and guests ordered.

● country– Country from where guests came

● Market\_segment– Which market

● segment is used for booking.

● Distribution\_channel – Booking Distribution Channel(TA/TO/Direct/Corporate etc.)

● Is\_repeated\_guests– Is customer made booking previously yes(1) or no(0)

● Previous\_cancellations– Number of previous cancellations.

● Previous\_bookins\_not\_canelled – Number of previous bookings which are not cancelled.

● reserved\_room\_type– Type of room customer reserved.

● assigned\_room\_type– Type of room assigned to customer.

**Questions we want to answer are and doing analysis:**

• Which Hotel has more bookings? (City or Resort)

• Percentage of bookings of cancellation

• Which month has more number of bookings

• Which year has more number of bookings?

• Which meal is preferred more by customers?

• From which country more guests came?

• How Long People Stay at hotels?

• Bookings are more on Weekdays or weekends?

• Do non repeated guests cancel more often than repeated ones ?

• Do customers who were on the waiting list for a long time have less cancellation compared to others ?

• Do a number of special requests have any correlation with having children/babies?

• Number of bookings per year for two hotels.

• Customer distribution based on type.

• Booking distribution based on assigned room type.

• Which Agent made the most bookings?

• Does a meal type have any correlation with Cancellation ?

• Which market segment has the most number of bookings?

• Which hotel has the most number of repeated guests/customers?

• Car Parking Space

**2. Introduction.**

Each observation represents hotel bookings. Bookings arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were cancelled. Since this is hotel real data,all data elements personal data or customer identification was deleted.

The dataset contains over 119,390 rows and 32 columns. Columns have the data type object,int64 and float64. It appears that a few columns contain some empty values since the Non-Null count for a few columns is lower than the total number of rows (119390)

**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**

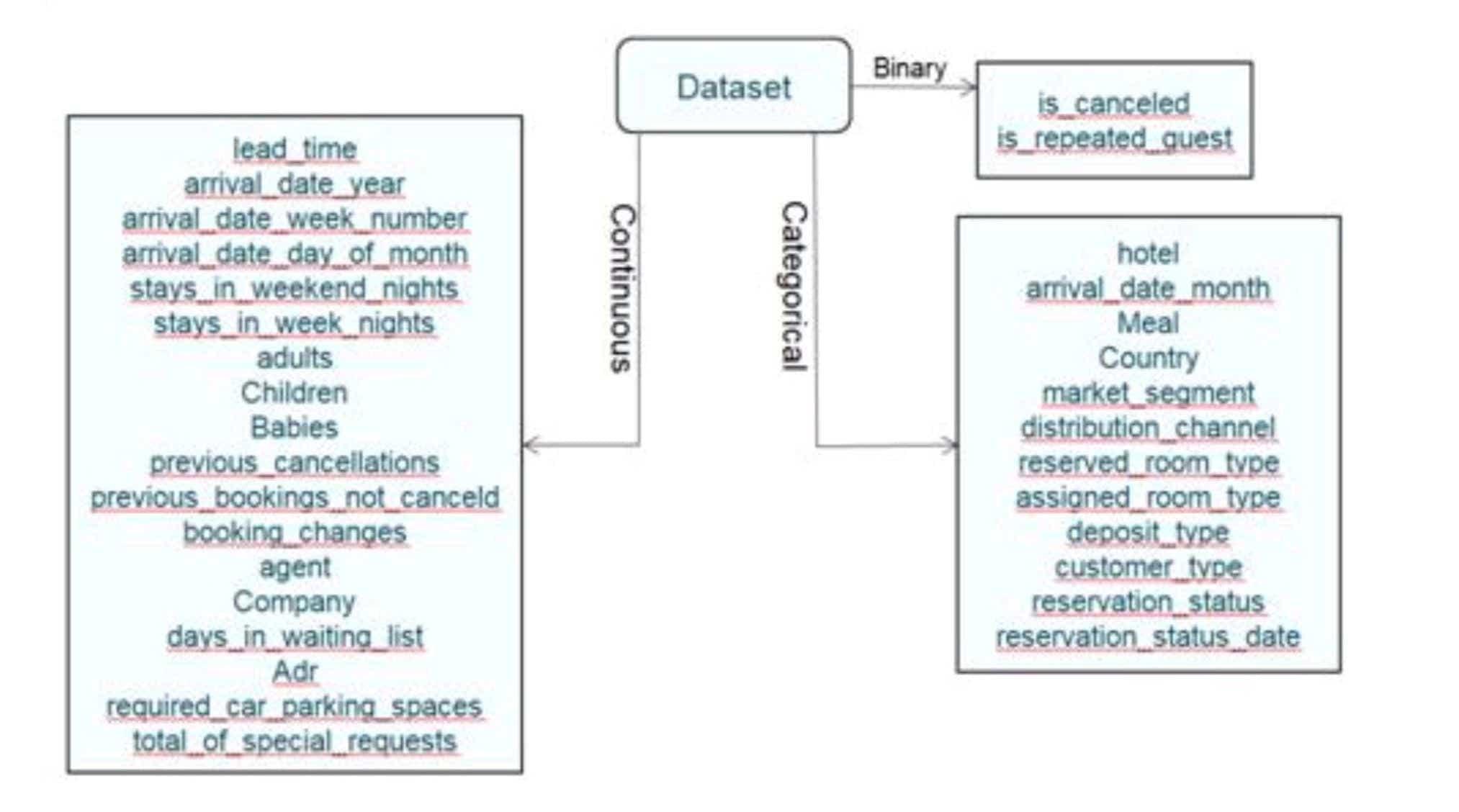
**30 reservation\_status 119390 non-null object**

**31 reservation\_status\_date 119390 non-null object**

We’ll need to deal with empty values. Filled all null values like company and agent with 0,

children with mean values of children and country with others as this is a string data type.

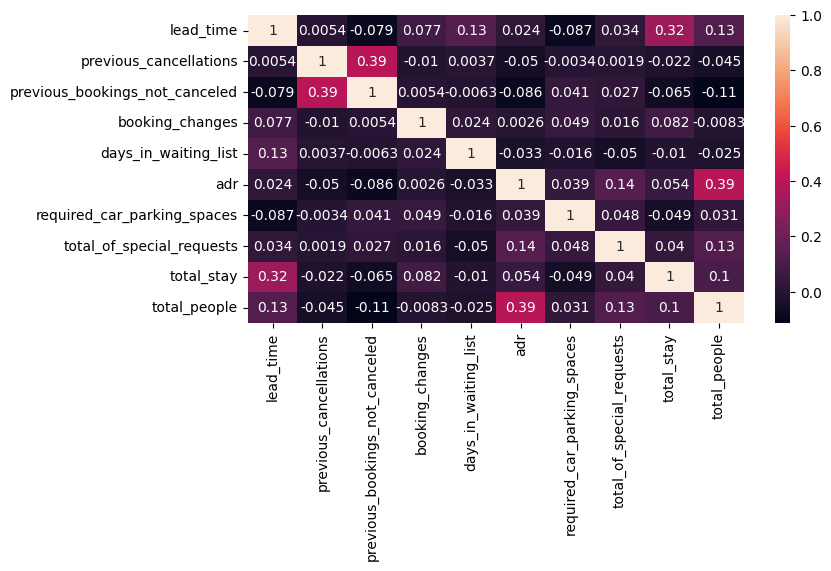
**Data Architecture-**



**3. Exploratory Data Analysis**

e will first find the relationship between the numerical data

Since, columns like 'is\_cancelled', 'arrival\_date\_year', 'arrival\_date\_week\_number', 'arrival\_date\_day\_of\_month', 'is\_repeated\_guest', 'company', 'agent' are categorical data having numerical type. So we don't need to check them for correlation.Also, we have added total\_stay and total\_people columns. So, we can remove adults, children, babies, stays\_in\_weekend\_nights, stays\_in\_week\_nights columns.

****

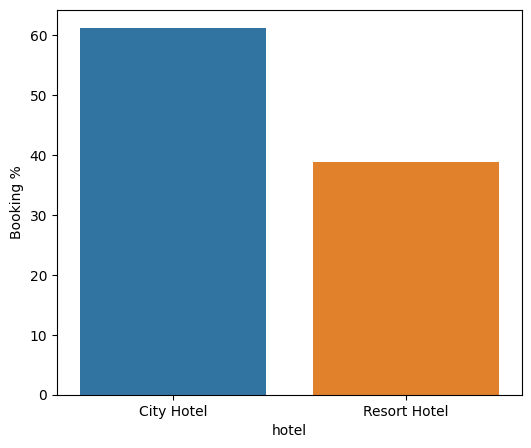
The total length of stay and the lead time have a slight correlation. This could imply that for longer hotel stays, people generally plan little ahead of time.

Ad revenue is slightly correlated with total people, which makes sense because more people means more revenue, and thus more ad revenue.

**3.1 Which Hotel has more bookings? (City or Resort)**

To compare this first we calculated the percentage of each hotel and then plotted it in a

graph with the help of seaborn library.

****

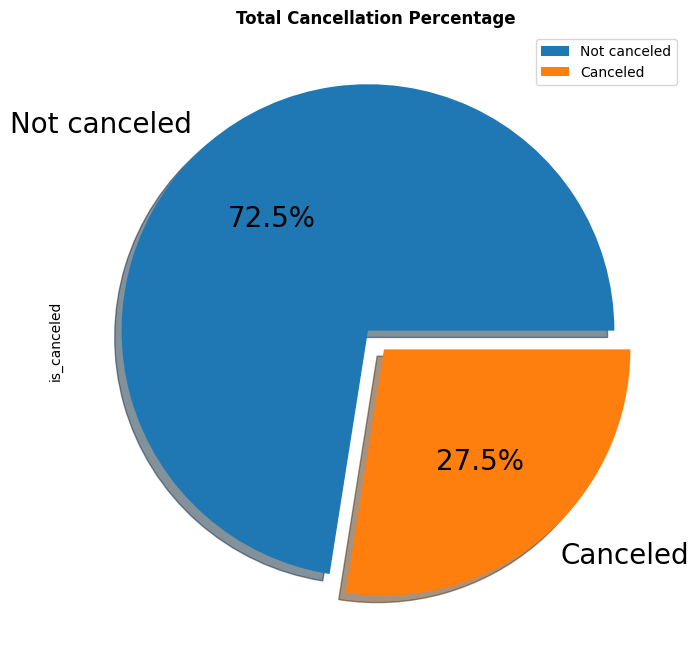
The figure 3.1a around 60% booking are for city hotels and 40% are for resort hotels. So

We can say city hotels have more bookings as compared to resort hotels.

**3.2 Percentage of bookings of cancellation**

Now we calculated the total percentage of booking cancelled with the help of the valuecount method and plotting pie charts using matplotlib library.

So here when we can see there is 27.5% of bookings were cancelled.0 indicates not cancelled bookings and 1 indicates cancelled bookings.

****

**3.3 Which month has more number of bookings?**

August and July have more bookings which is around 11257 and 10057 in both the hotels.

January has the least number of bookings which is 4693.

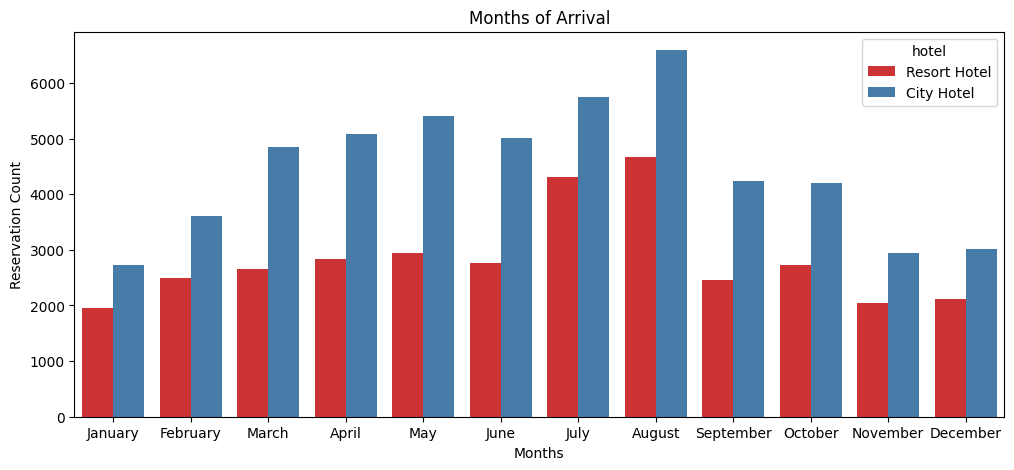
I plotted this in the graph below. Also we can have more bookings than resort hotels.

Figure 3.3a

**August 11257**

**July 10057**

**May 8355**

**April 7908**

**June 7765**

**March 7513**

**October 6934**

**September 6690**

**February 6098**

**December 5131**

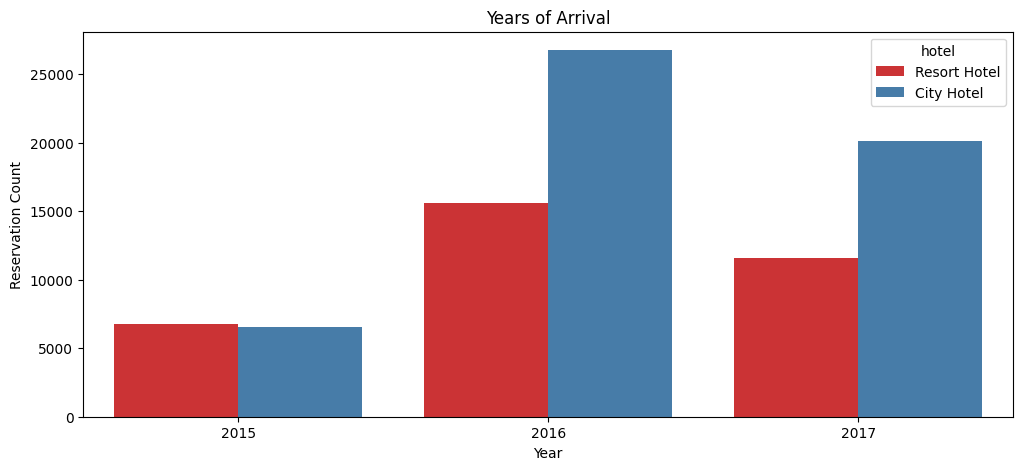
**November 4995**

**January 4693**

**3.4 Which year has more number of bookings?**

Here we did Analysis based on the year users belong to. There were 3 years of data

available in the given dataset from 2015-2017.

****

**Figure 3.4**

**2016 42391**

**2017 31692**

**2015 13313**

**Name: arrival\_date\_year,dtype: int64**

In the above plot we can see in the year 2016 there was more number of bookings as

compared to 2015 and 2017 in both the hotels.

**3.5 Which meal is preferred more by the customer?**

Here we are going to see which meal type is the favourite one of most of the customers.

We have plotted a pie chart and calculated the percentage of each meal type.

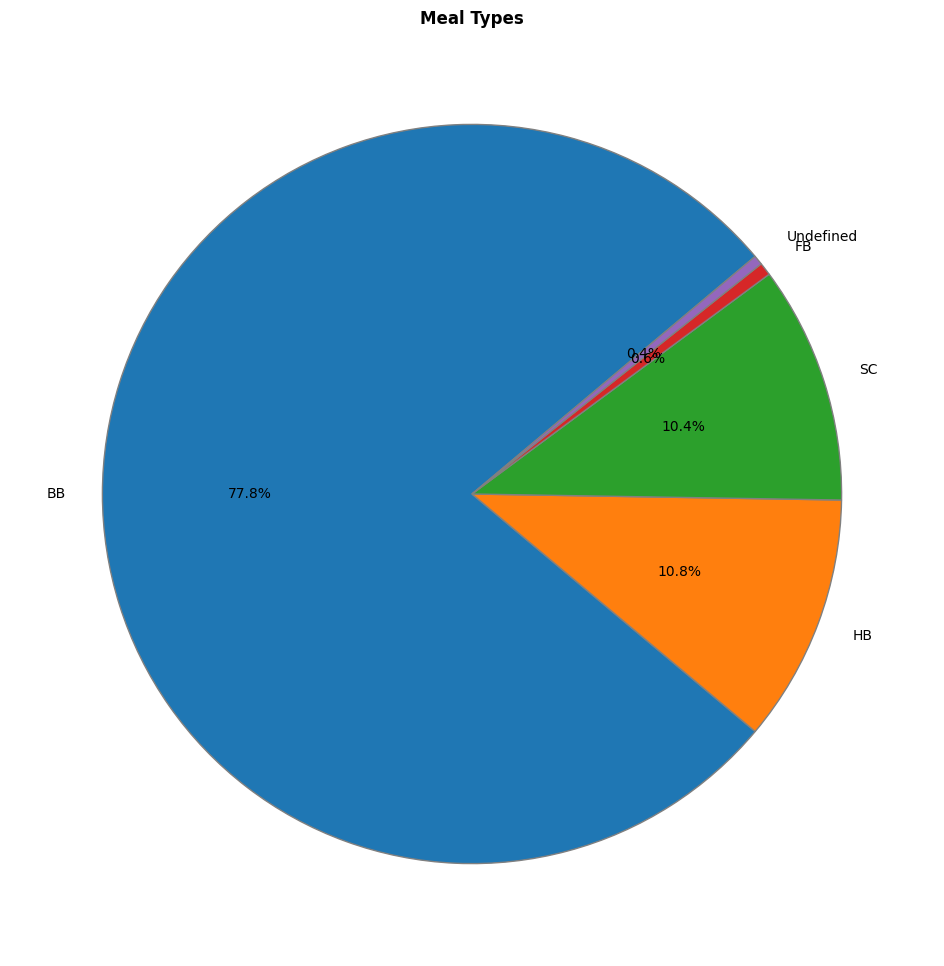
****

Figure 3.5a

It is clearly evident that users mostly prefer BB meals. FB is the least preferred meal and HB

and SC are equally preferred meals.

**3.6 From which country most guests are coming?**

Here we checked country guests that both the hotels in total have received. The data on

cancelled bookings will not be included here. We are going to do the same analysis on top

10 countries from where most guests are coming.

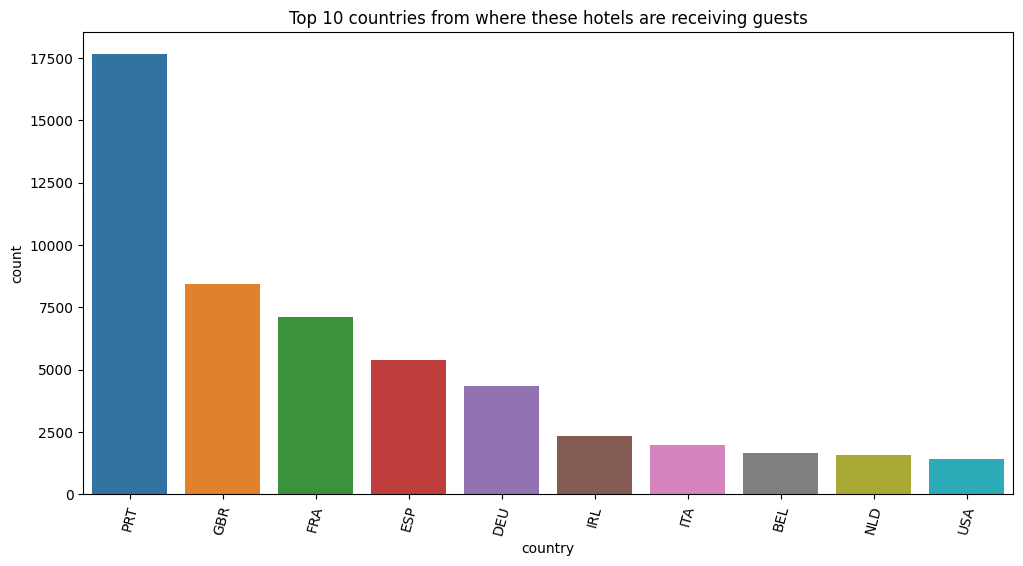


Figure 3.6a

Most of the customers are from European countries like PRT(Portugal), GBR(Great

Britain), FRA(France).

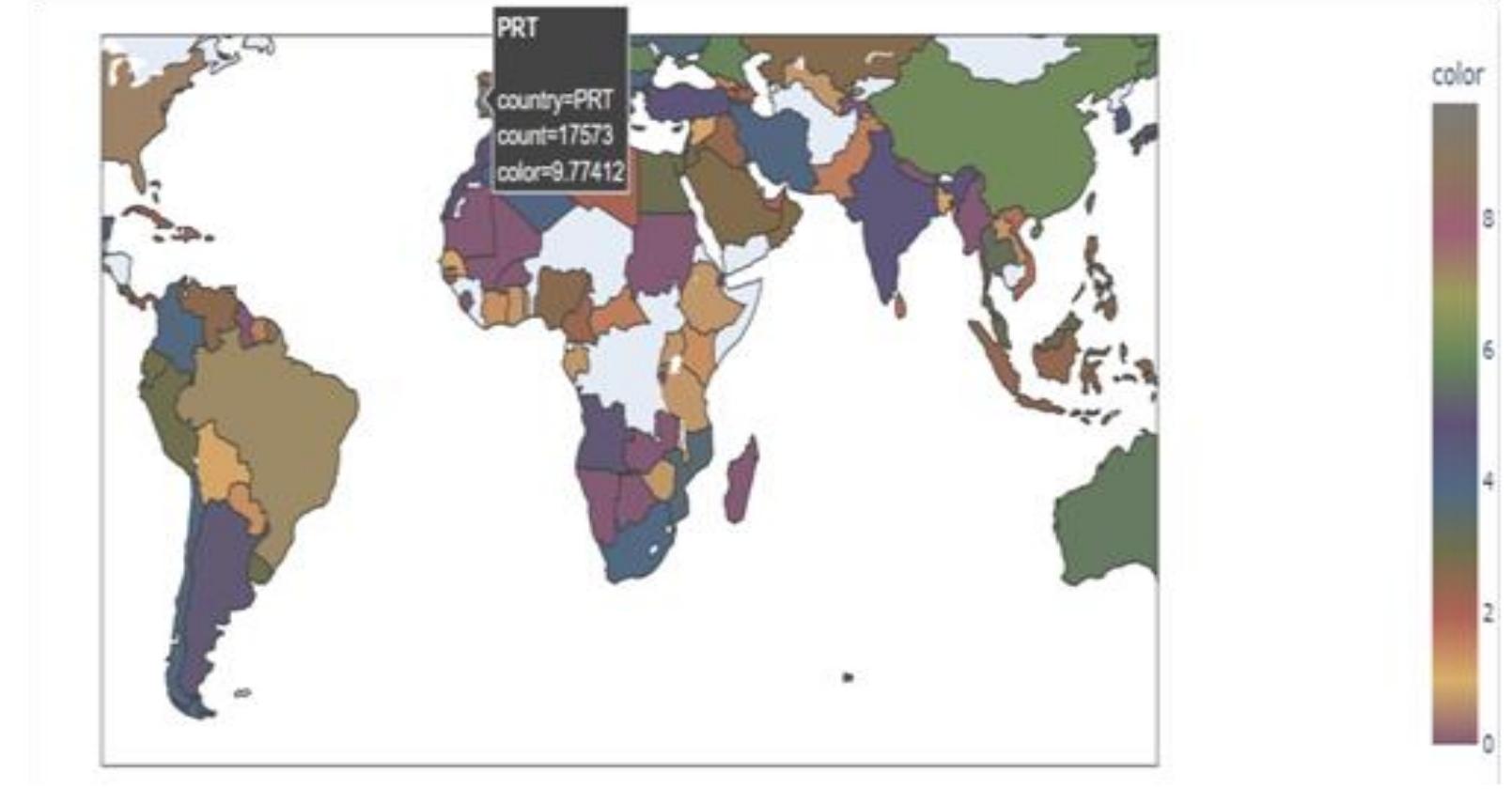


Figure 3.6b

Also plotted countries on the world map.

**3.7 How long do people stay at hotels?**

Now we want to perform an analysis to see how long people are staying at a hotel.

For this first we created a different table by combining stay in weekend nights and stay in week

days which are total nights spent by customers.

And then plot it on a bar plot.

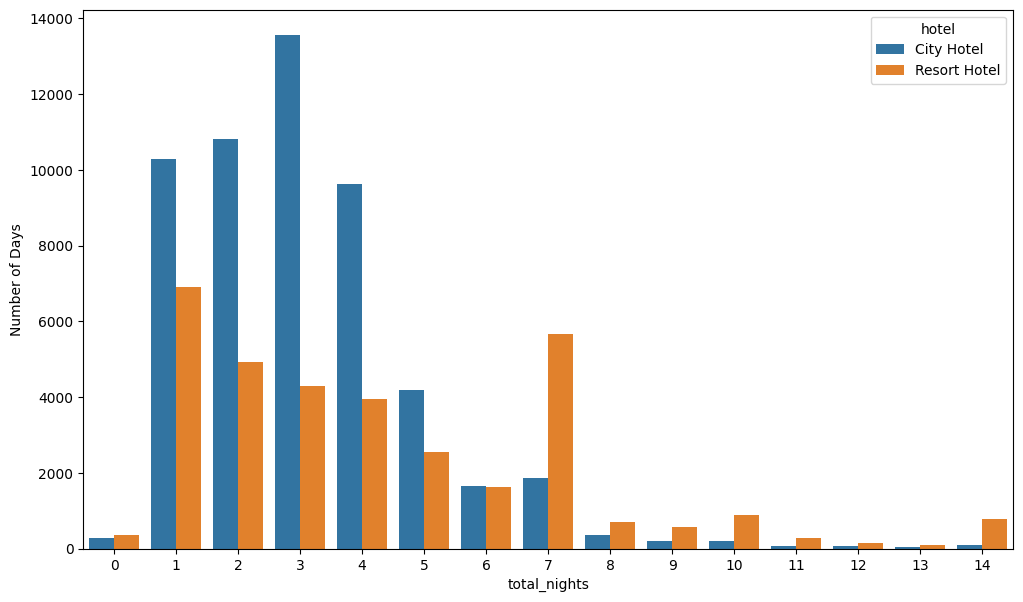
****

Figure 3.7a

Above figure indicates that most people stay at the hotel for less than 7 days.

**3.8 Bookings are more on weekdays or weekends?**

Now we calculated weekend and week days nights by sum function. And then plotted it in two different graphs weekdays and weekends with seaborn countplot.

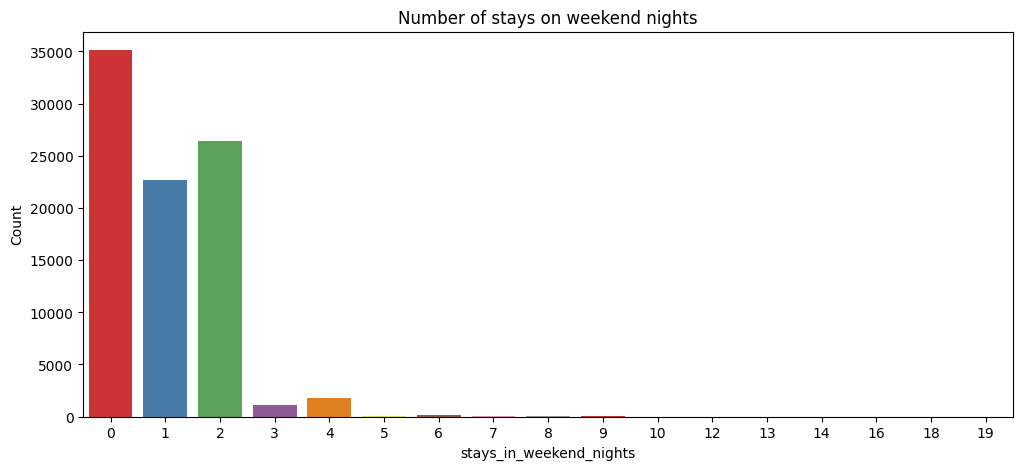
Below two figures indicate that there are more bookings at week nights as compared to weekend nights. Usually there are more outings on weekends but here it seems different.****

Figure 3.8a

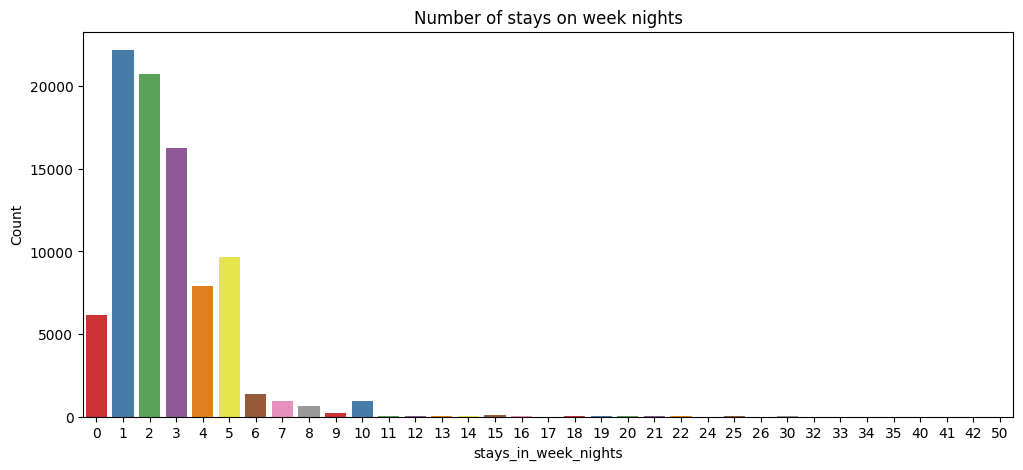
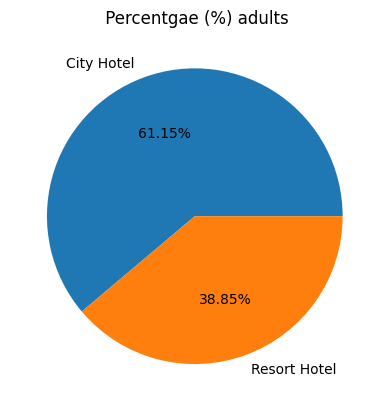
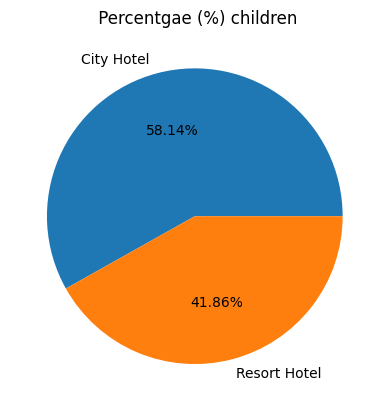
****

Figure 3.8b

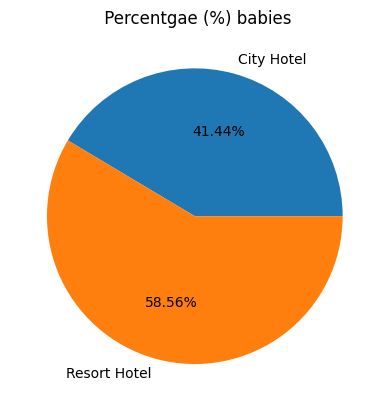
**3.9) Which type of hotel is mostly preferred by adults , children or babies?**

****

From above pie chart it is clearly mentioned that 61.15 % of Adults preferring city hotel



From above pie chart it is clearly mentioned that 58.14 % of children preferring city hotel



From above pie chart it is clearly mentioned that 41.44% of babies preferring city hotel which is less than resort hotel

From above it is clear that city hotels are mostly preferred by adults and children

**3.10 Number of bookings per year for two hotels**.

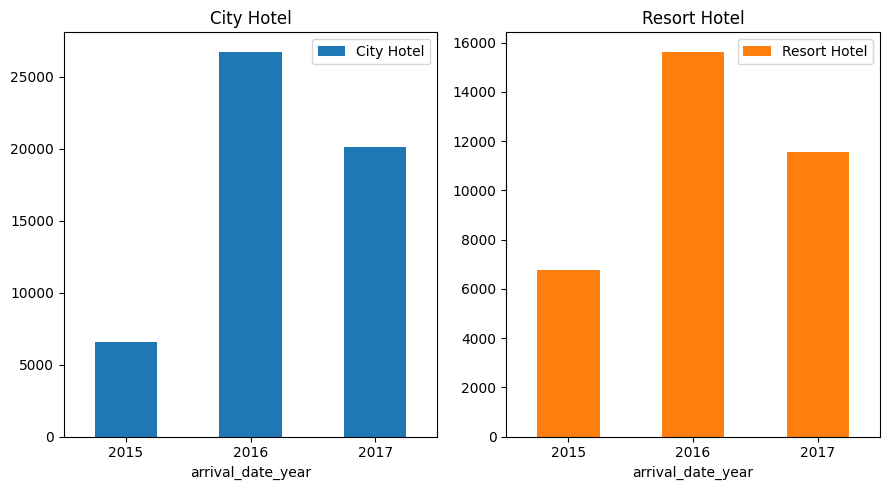


Figure 3.10a

Overall bookings of City Hotel is higher when compared to Resort Hotel we can observe

the percent of growth in number of booking from 2015 to 2017 is much higher in City Hotel

**3.11 Customers distribution based on type**

Customer Type distribution has the same trend in both types of Hotels, no

significant difference. The Transient type of hotel has more bookings.

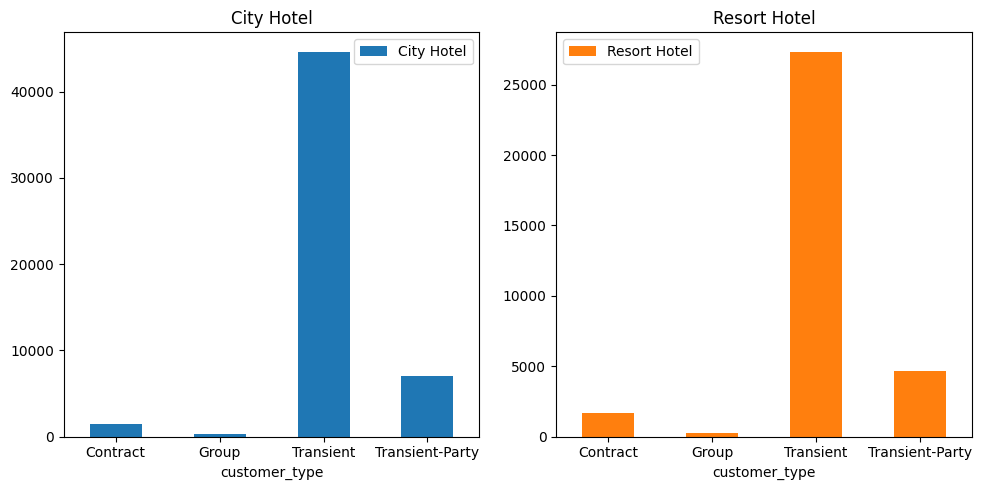


Figure 3.11a

**3.12 Booking distribution based on assigned room type.**

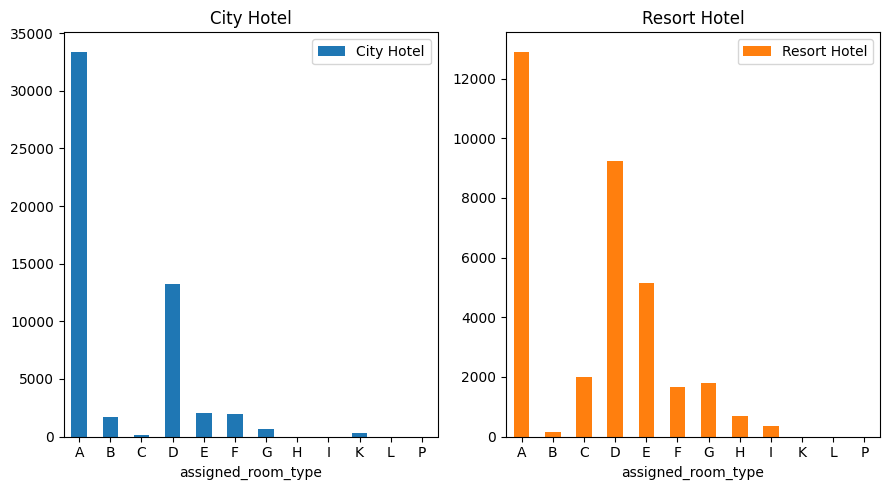
****

Figure 3.12a

Type A rooms are more preferred room type by customers.One clear difference in assigned

room type between City Hotel and Resort hotel is that B type rooms are assigned more than C

type rooms in City Hotel but in Resort Hotel it is the reverse.

**3.13 Which agent made the most bookings?**

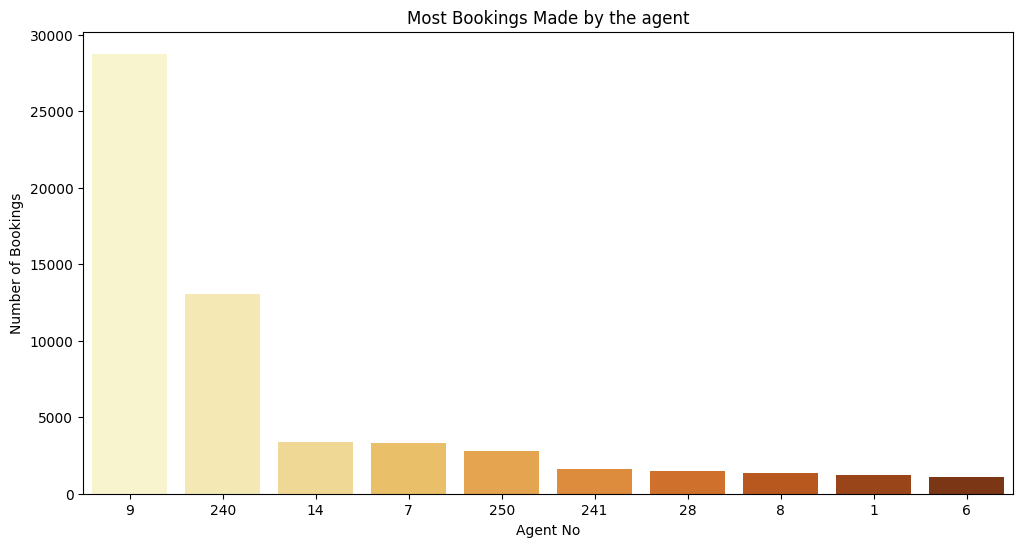


Figure 3.13a

Agent ID no: 9 made most of the bookings

Agent ID no: 6 made the least number of bookings

Agent ID no: 14 and 7 have almost same number of bookings

Agent ID no: 28, 8 and 1 have almost same number of bookings

**3.14 Does a meal type have any correlation with cancellation?**

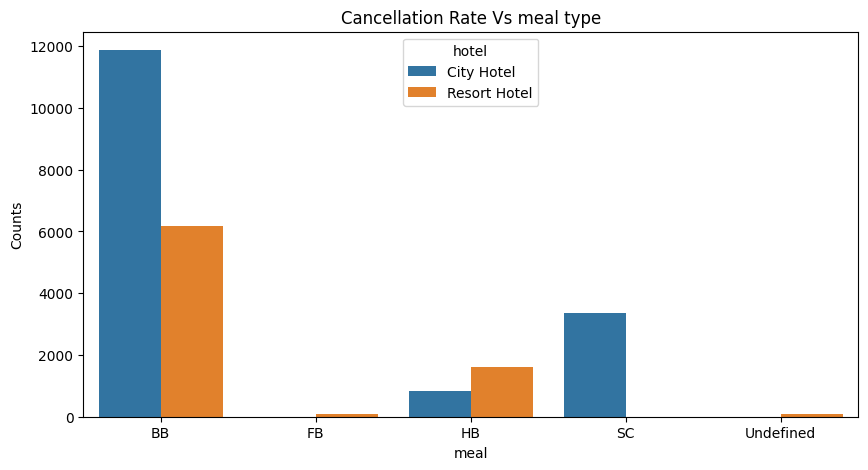


Figure 3.14a

Above figure indicates Most cancelled bookings preferred BB meal.And more cancelled

bookings are from resort hotels only for HB meals there is slight high cancellation from resort hotels.

**3.15 Which market segment has the most number of bookings?**

Most bookings are through Online TA. But it is really important to see how many people

actually show up after booking through Online TA. So we next compare Cancellation rates among different market segments

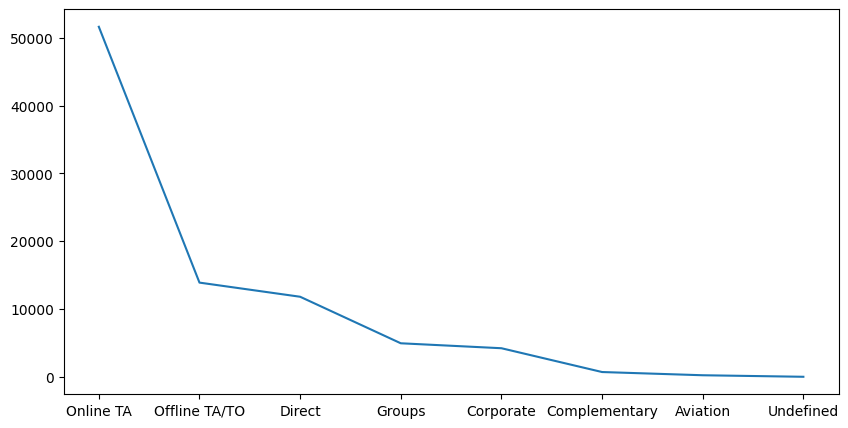


Figure 3.15a

**3.16 Which market segment has the most cancellation?**

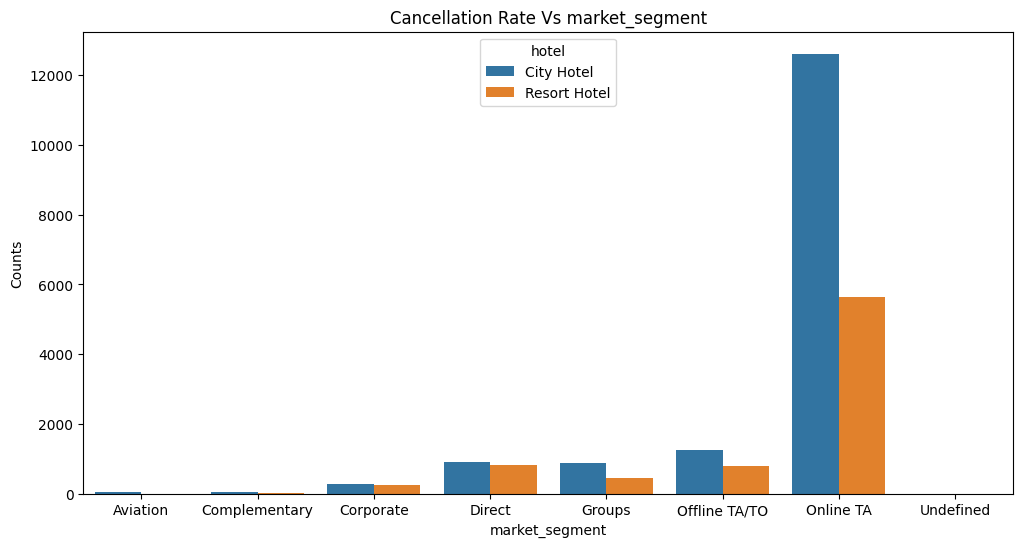
****

Figure 3.16a

Online TA' has the highest cancellation rate in both types of hotels. To avoid this the

hotels can give additional discounts or offers to people who book through Online TA's.

**3.17 Which hotel has the most number of repeated guests/customers?**

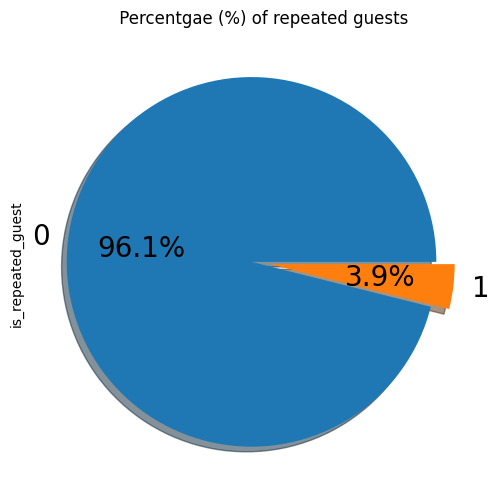
****

Figure 3.17a

Repeated guests are very few, only 3.9 %. In order to retain the customers,the

management should take feedback and try improving their services.They should also keep a

track of reviews from guests and try to improve the services.

**3.18 Correlation between car parking space and cancellation.**

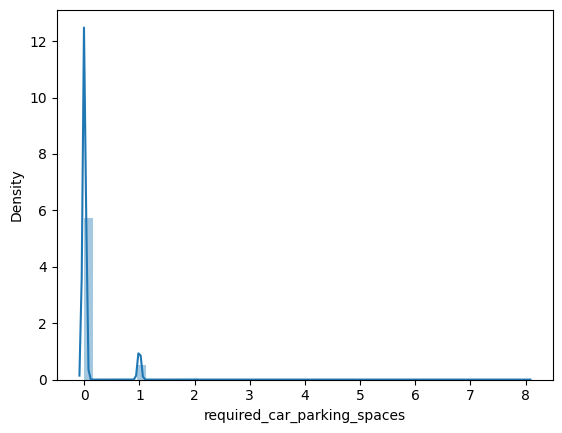
****

Figure 3.18a

We notice that most of the records require no or at most 1 car parking spaces, though

There are records that show more parking spaces. This could be families travelling together and larger business trips.

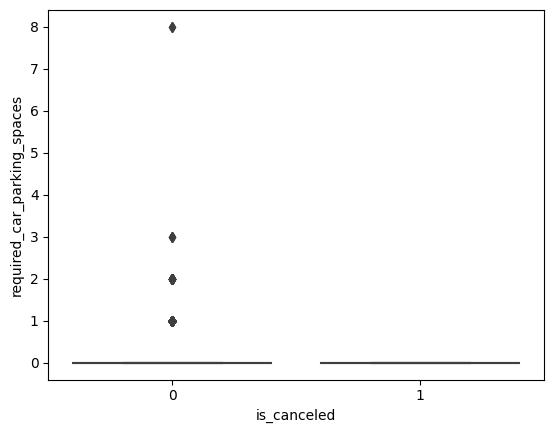


Figure 3.18b4

We see that non-cancelled bookings required more number of car parking spaces compared to

cancelled bookings

**4. Conclusion:**

That's it! We reached the end of our analysis. The following are our observations:

● Majority of hotel bookings are from city hotels.

● The cancellation rate for hotels is 27.5%

● We should target July and August as most of the bookings. Those are peak months due

to the summer period.

● The number of bookings seems to be high in 2016 while the bookings seem to be less in

2015 and 2017. This is majorly due to more data points being available in 2016.

● BB meal is the most preferred meal by customers.

● It appears that a disproportionately high number of bookings are from Portugal, probably

because the hotel is located in Portugal itself.

● Most people do not seem to prefer to stay at the hotel for more than 1 week.

● We observe that Weekday bookings are higher than the Weekend numbers. That is an

interesting finding.

● When we find a correlation between assigning different rooms for having children/babies

there is a low probability of getting a room in a resort hotel as compared to a city hotel.

● The percentage of cancellation for Non Repeated User is higher when compared to

repeated users for both City and Resort Hotels.

● There's some positive correlation between lead\_time and cancellation status i.e., the

higher the lead time the more chances of cancellation and avg lead time of cancelled

booking is 40 days more when compared to not cancelled bookings.

● Having children/babies have less special requests in Resort Hotel when compared to

City hotel.

● Even though overall bookings of City Hotel are higher when compared to Resort Hotel

the percent of growth in number of booking from 2015 to 2016 is much higher in City

Hotel.

● Customer Type distribution has the same trend in both types of Hotels, as transient type

of customers are higher in both hotels.

● Type A room is assigned to most of the customers in both the hotels.

● Agent 9 is made most bookings in both the hotels.So we can focus more on agent 9 for

more profitable business.

● Most bookings are through Online TA. Also 'Online T A' has the highest cancellation rate

of both types of hotels. To avoid this the hotels can give additional discounts or offers to

people who book through Online TA's.

● Repeated guests are very few which only 3.9 %.we should target our advertisement on

guests to increase returning guests. Also should take feedback and reviews and should

work on that.

● We notice that most of the records require no or at most 1 car parking spaces, though

There are records that show more parking spaces. This could be families

travelling together and larger business trips.

**References-**

1. Numpy, Pandas, Matplotlib & seaborn documentation.

2. Alma Better recorded classes

3. Articles on Towards Data Science.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* **Thank you \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***