Capstone Project

‘EDA on Hotel booking data’

By-Deepak Kumar Dubey(cohort Austin)

CONTENTS

- Introduction
- Data summary
- Results of EDA analysis
- Conclusion

INTRODUCTION

We have a hotel booking data for a duration of around 3 years at our hands which we need to analyse.Before beginning our analysis , let's have a look at some important Key performance indicators on which we will base our analysis on.

Important KPI's in hotel business-

1)Average daily rate

2)occupancy rate

3)Total bookings

4)Average length of stay

5)Meals ordered per booking

Data summary-

Column information-We have total 32 columns out of which only 3 contains null values .Let's have a look at some important of them one by one.

1)Hotel -Tells us about the type of hotel.We have two types of hotel in our data.

- City hotel
- Resort hotel

2)is_canceled- Weather booking was canceled or not

- 0 not canceled
- 1 canceled

3)lead_time -

- No of hours that elapsed between entering date of booking into property management system and arrival date

4)arrival date year

- 2015
- 2016
- 2017

5)arrival_date_month- january to december

6)arrival_date_week_number

7)arrival_date_day_of_month

8)stays_in_weekend_nights-

- No of weekend nights (Sat/Sun) the guest stayed or booked to stay at the hotel

9)stays_in_week_nights

- No of week nights (Mon - Fri) the guest stayed or booked to stay at the hotel

10)Adults

11)children

12)Babies

13)meal

- Type of meal booked. Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)

14)country

15)market_segment-(a group of people who share one or more common characteristics, lumped together for marketing purposes)

- TA: Travel agents
- TO: Tour operators

16)distribution_channel-(A distribution channel is a chain of businesses or intermediaries through which a good or service passes until it reaches the final buyer or the end consumer)

- TA/TO
- Corporate
- GDS
- Direct

17)is_repeated_guest (value indicating if the booking name was from repeated guest)

- 1: yes
- 0:no

18)previous_cancellations-

- Number of previous bookings that were cancelled by the customer prior to the current booking

19)previous_bookings_not_canceled-

- Number of previous bookings not cancelled by the customer prior to the current booking

20)deposit_type-

- Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.

21)agent-

- ID of the travel agency that made the booking

22)customer_type-

- 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

23)adr (average daily rate)-

- Average daily rate=(sum of all lodging transaction/total number of staying night)

24)required_parking_space

25)reservation_status-

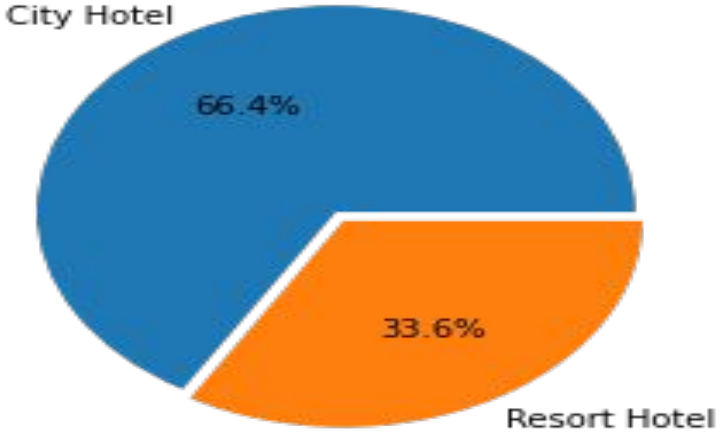
- Canceled – booking was canceled by the customer;
- Check-Out – customer has checked in but already departed;
- No-Show – customer did not check-in and did inform the hotel of the reason why

26)reservation_status_date-

- Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel

Which hotel???

Let’s begin our analysis with ,between two types of hotel-City hotel and resort hotel which was booked more in these 3 years through a pie chart.



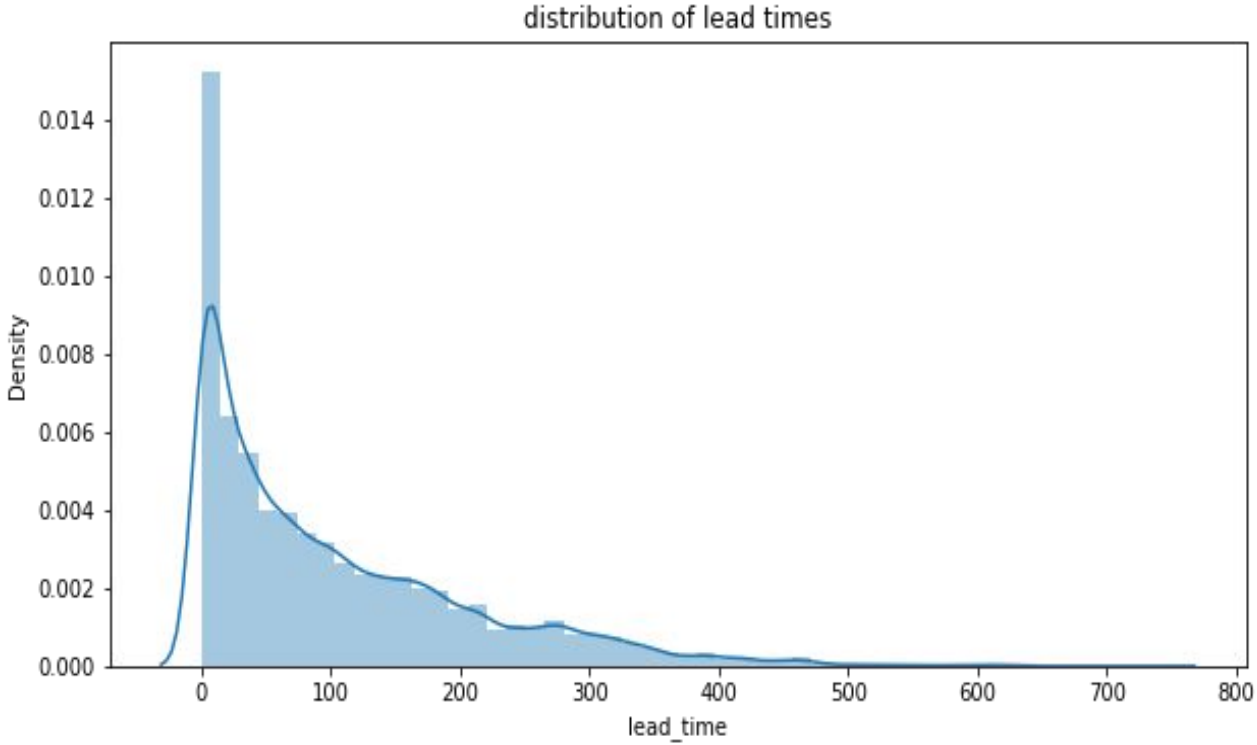
Conclusion-we can clearly see that city hotels were booked more number of times as compared to resort hotel.

Should we pre-book the Hotel??

Lead time variation and its distribution plot-

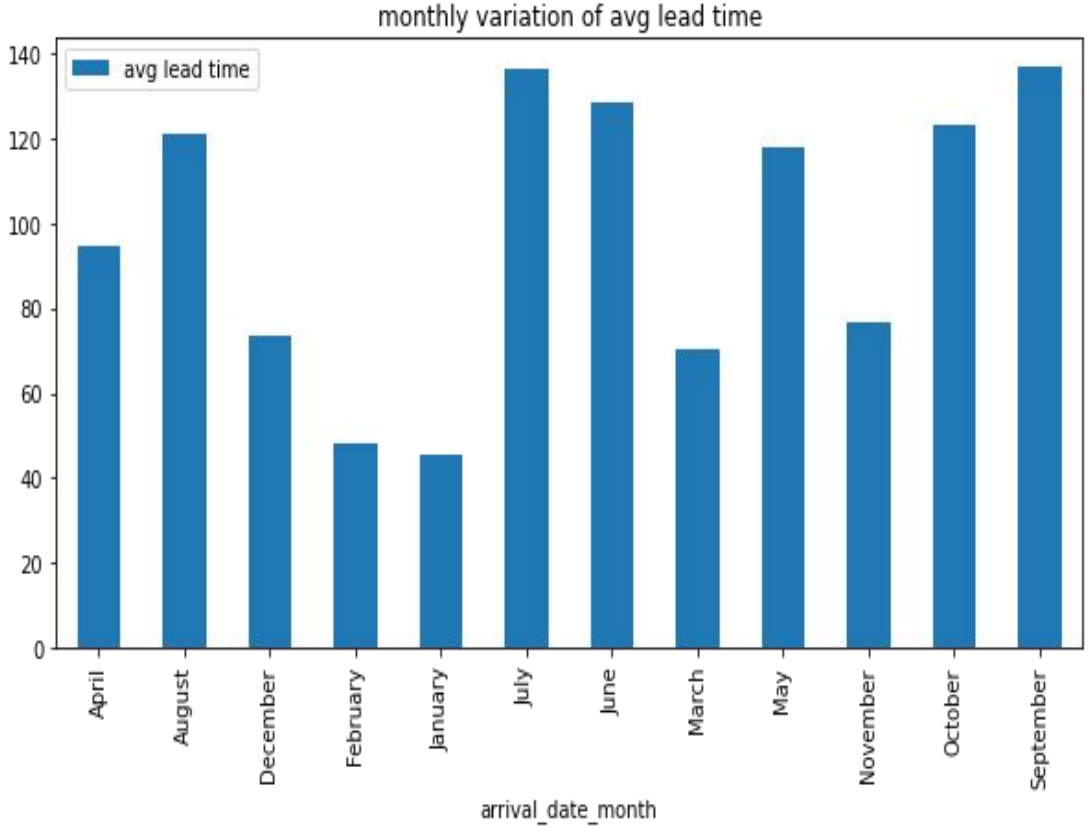
Let’s have a look at the distribution plot for avg monthly lead time

We can see from the distribution plot of lead times that most of the customers have there lead times between 0 and 100 but one customers have lead time as high as 700 hours and many have greater than 100.



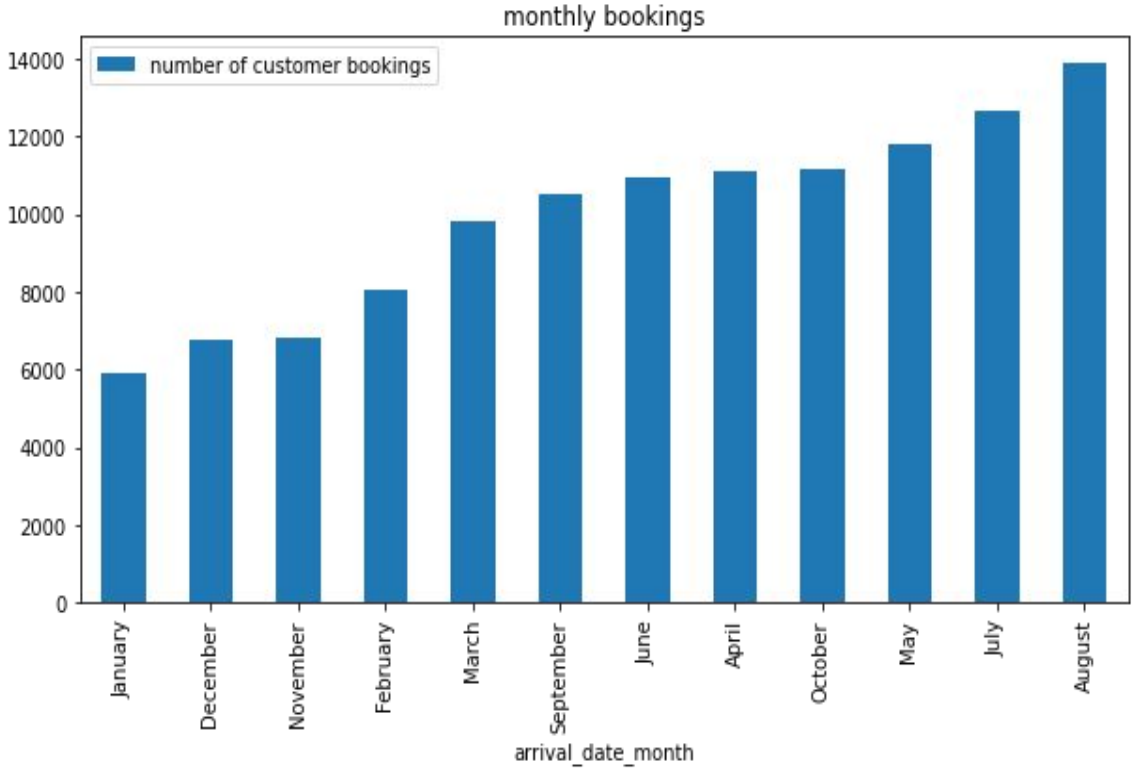
Which month has highest average lead time??

We can see from the chart may ,june,july, August have high average lead as compared to other months but why are we seeing such trend.Is it somehow related to demand??

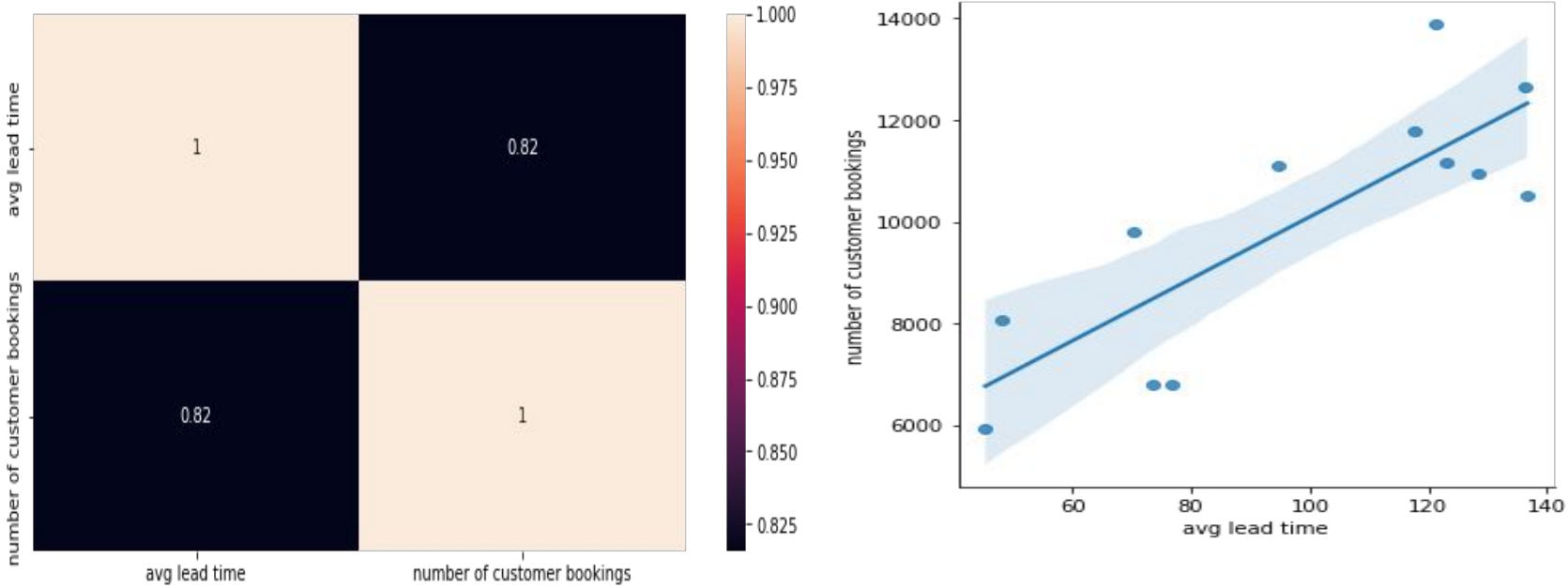


Which months of the year Hotel are relatively crowded??

This chart shows the monthly bookings .We can see that april,may ,june july receive highest amount of bookings.Aren't these the same months which had high avg lead times.



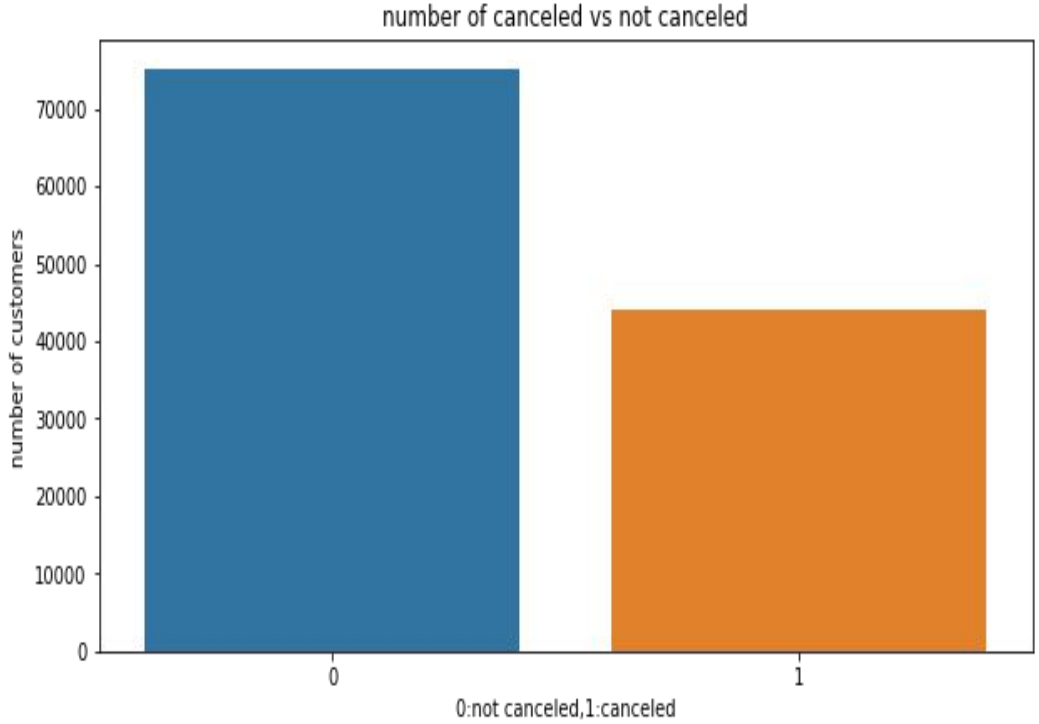
So is there any correlation between monthly bookings and lead time??



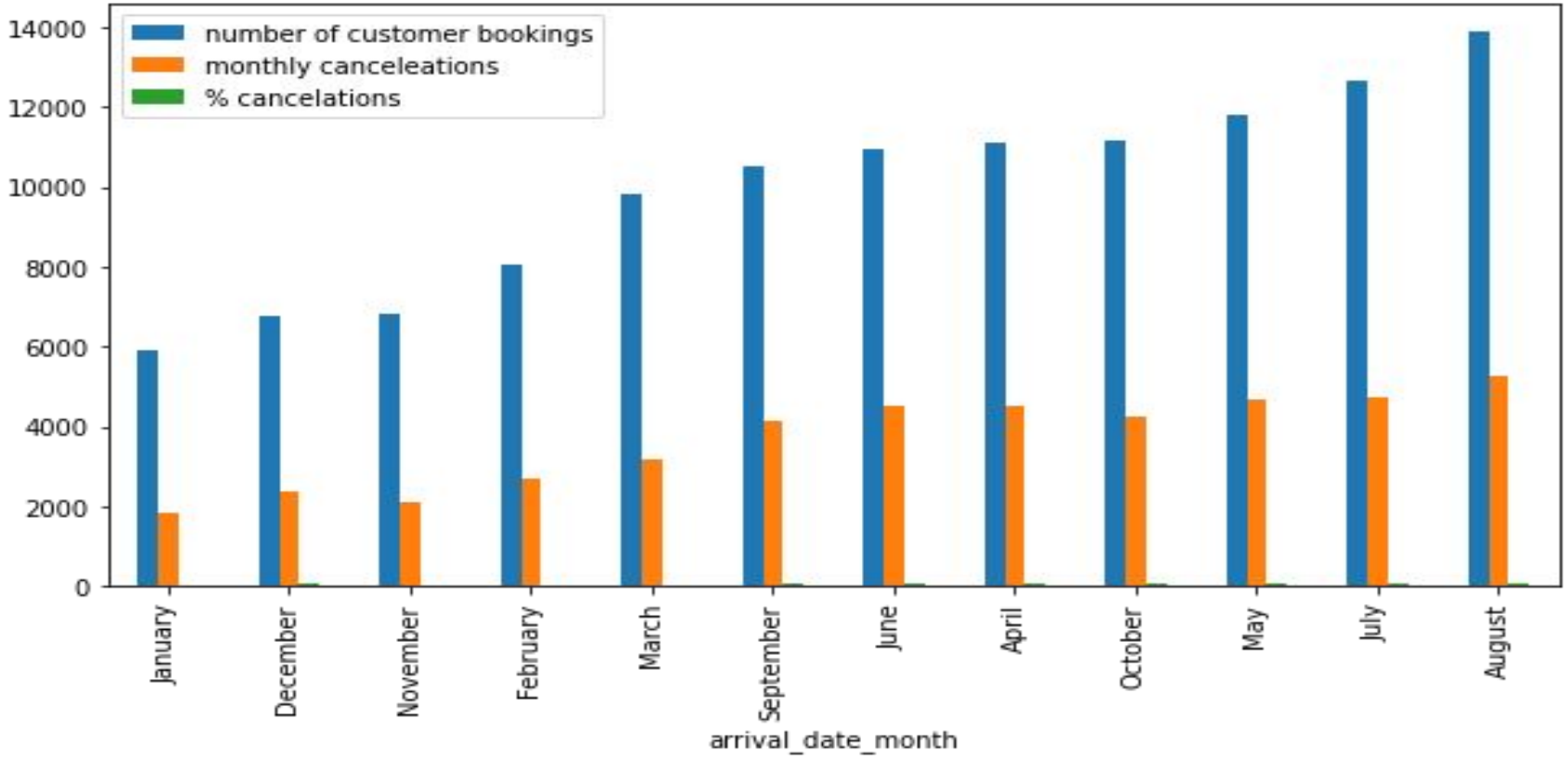
So we can see that correlation coefficient between these two parameters is 0.82.This implies that a fairly strong positive relationship.

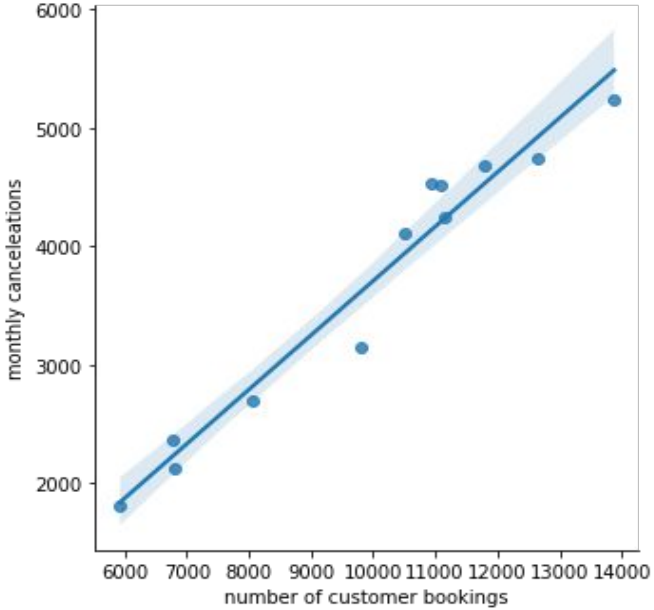
How many bookings get canceled ??

We can see from this chart that almost ⅓ rd of bookings get canceled.



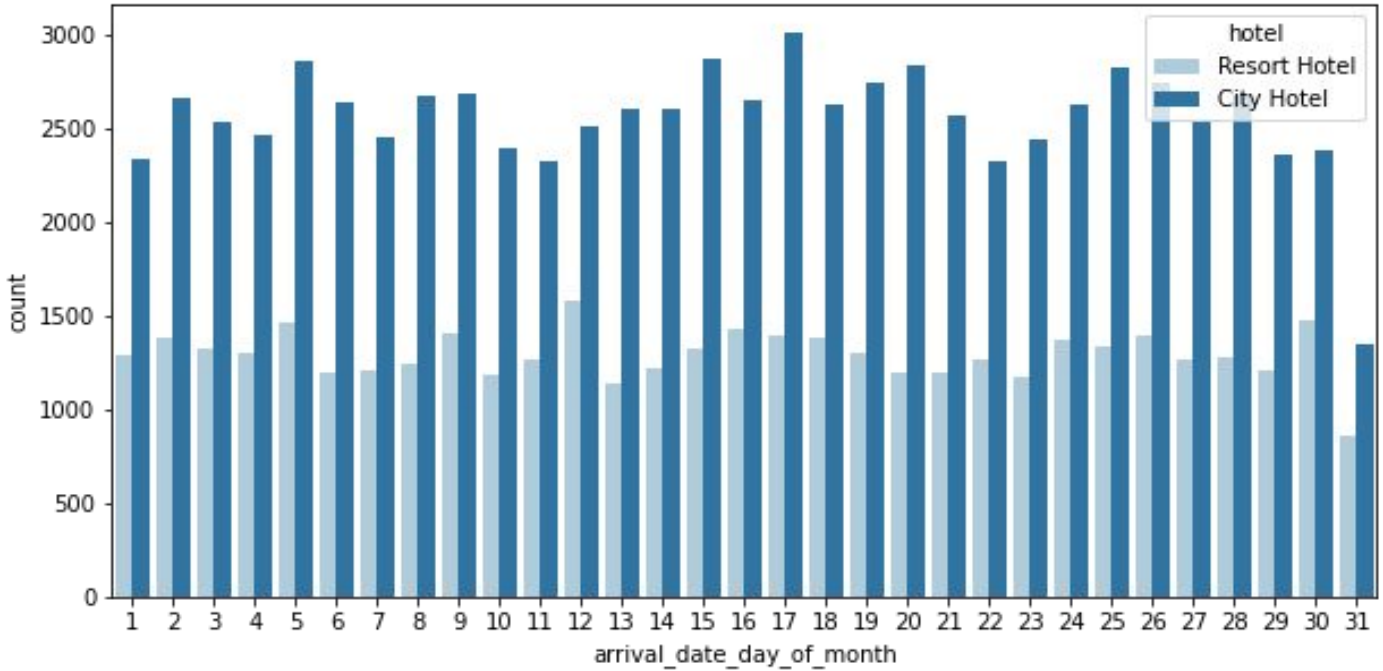
What's the correlation between cancellations and bookings??





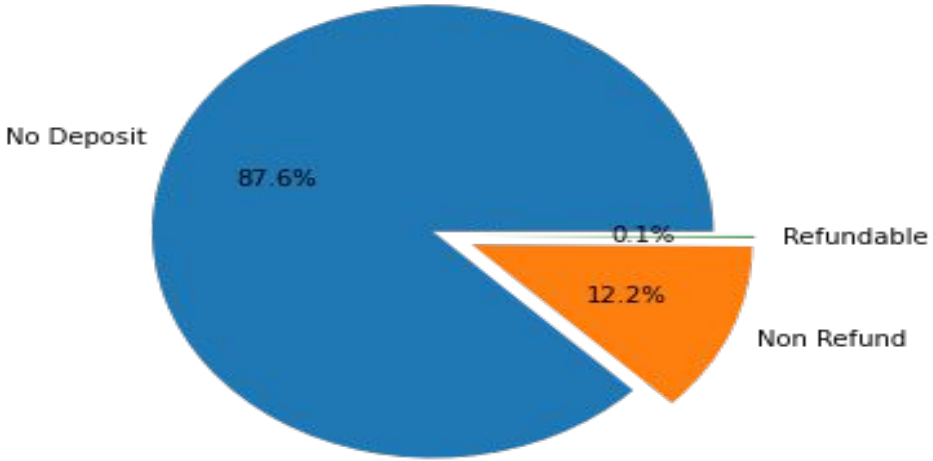
We can see there is a strong correlation coefficient of 0.98 which is like almost perfect positive correlation.

Which days of month are busiest??



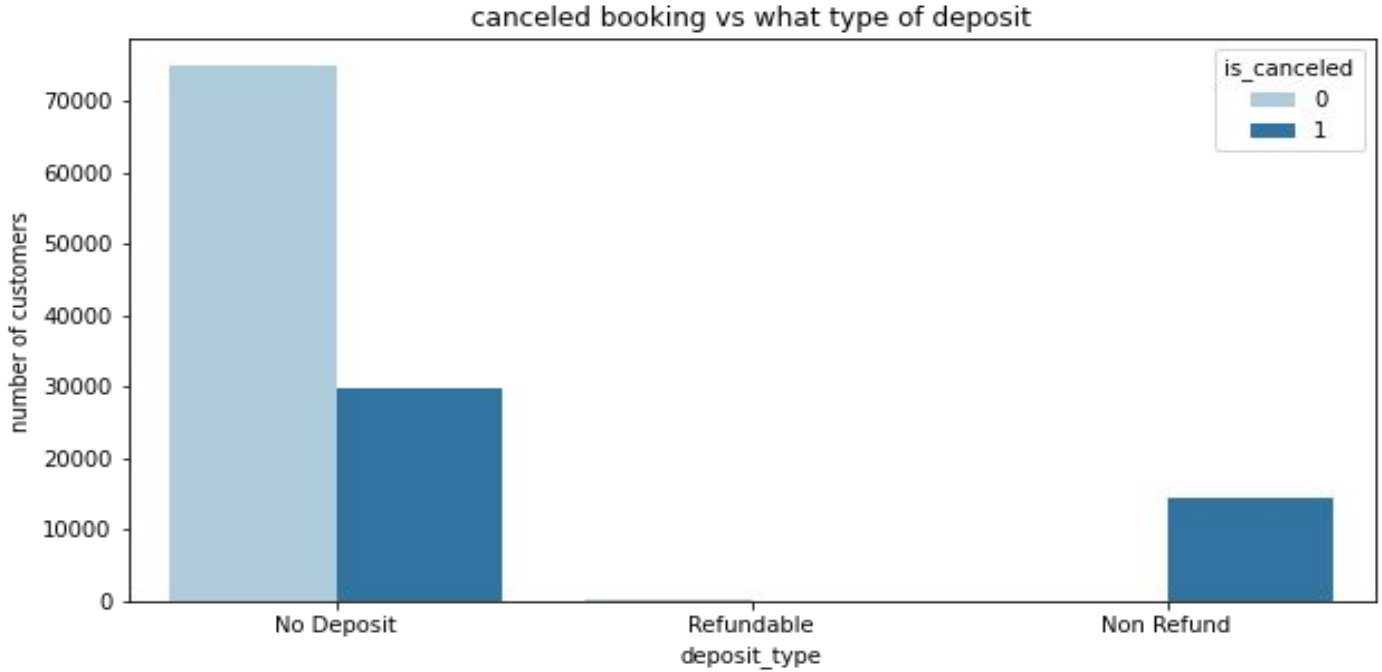
All days of month are busy except last 2-3 days and 1st day.

What kind of deposit do customers prefer while booking hotels???



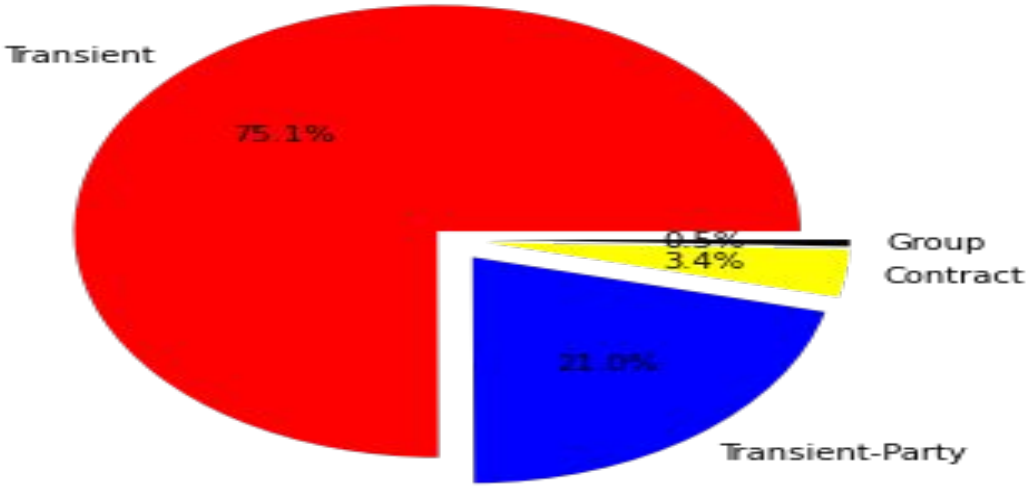
So 87.6% people prefer no deposit,12.2% prefer Non-refund type deposit,meagre 0.1% prefer Refundable type of deposit.

Canceled booking??Refundable deposit or Non refundable



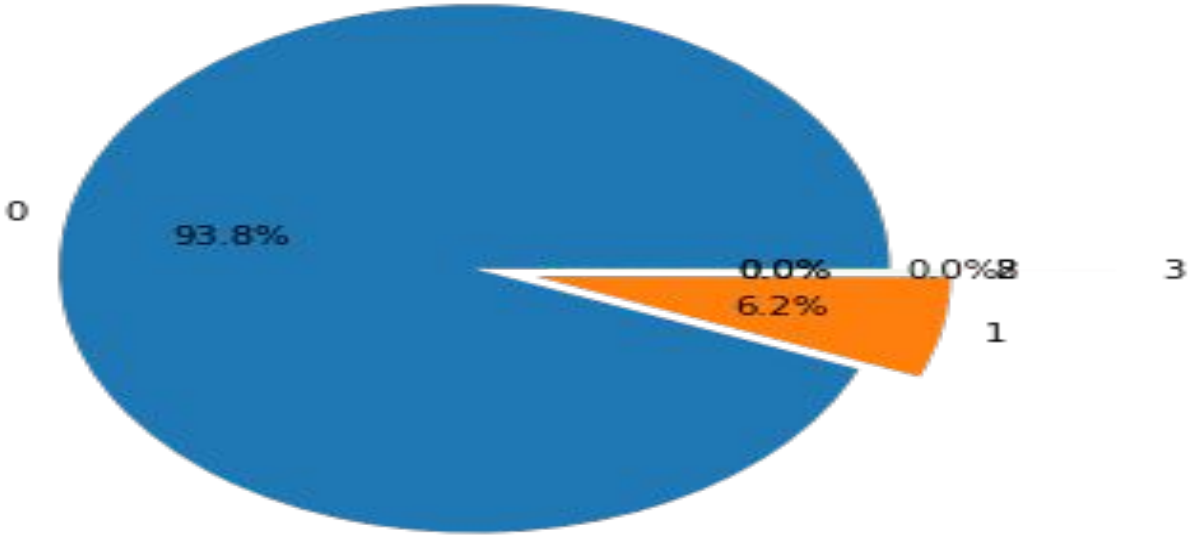
Those who have deposits and cancelled are of non-refund type as we can see from the rightmost bar.98 people have Refundable type out of which only 5 cancel.

What type of customer visits most??



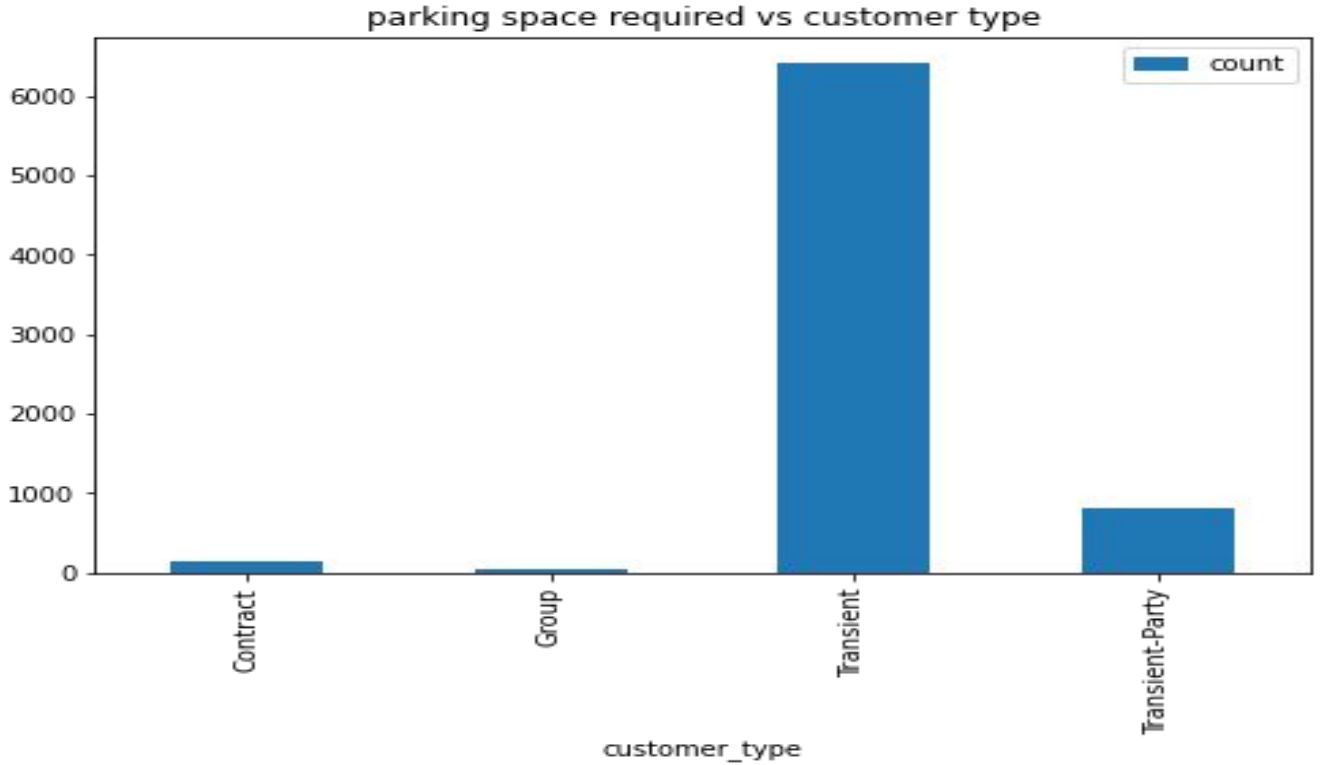
We can conclude from the pie chart that most customers visiting are of Transient type followed by Transient party then contract type and at last minimum number of customers visit in group.

How many customers visiting required parking space??



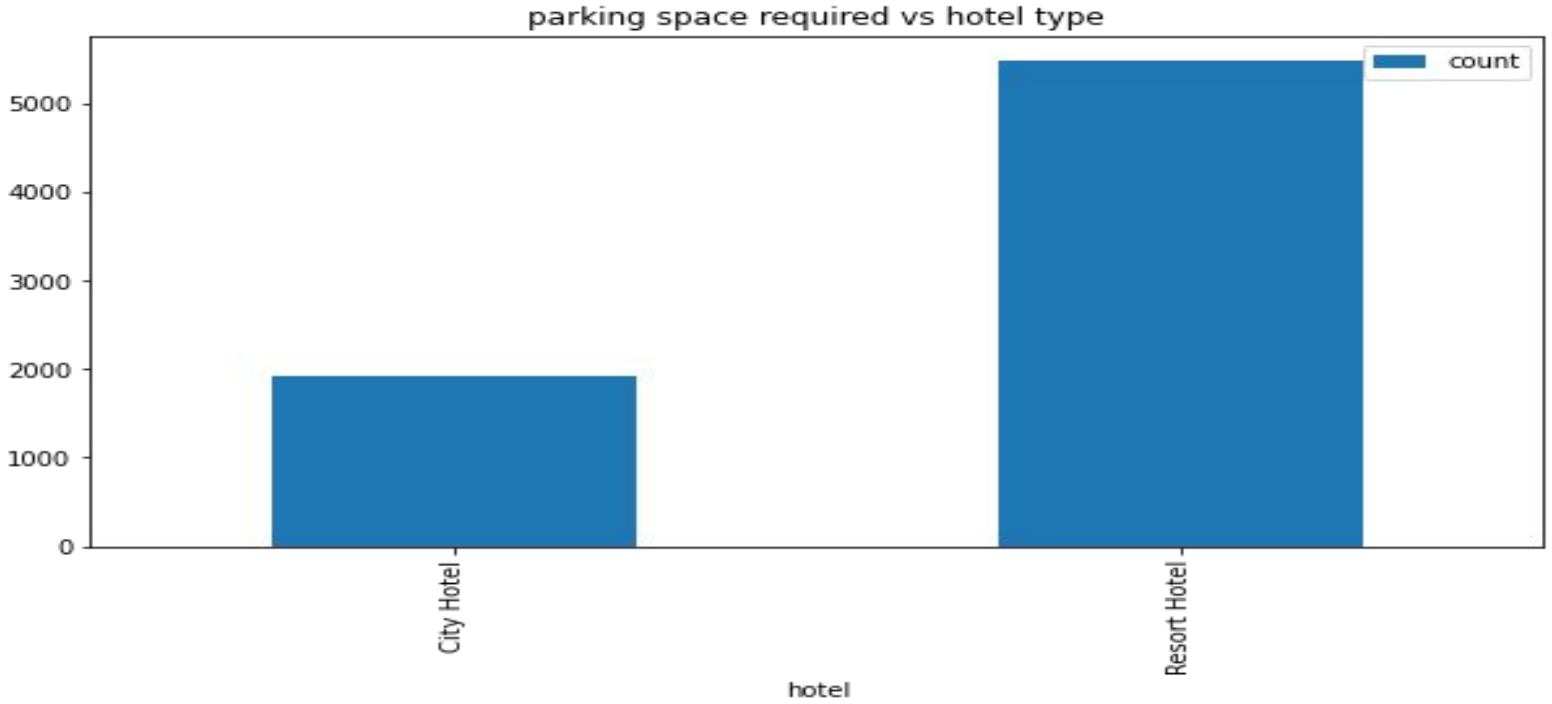
Most of customers visiting don’t require parking spaces.Only 6.2% people require one parking space.

What type of customers require parking space??



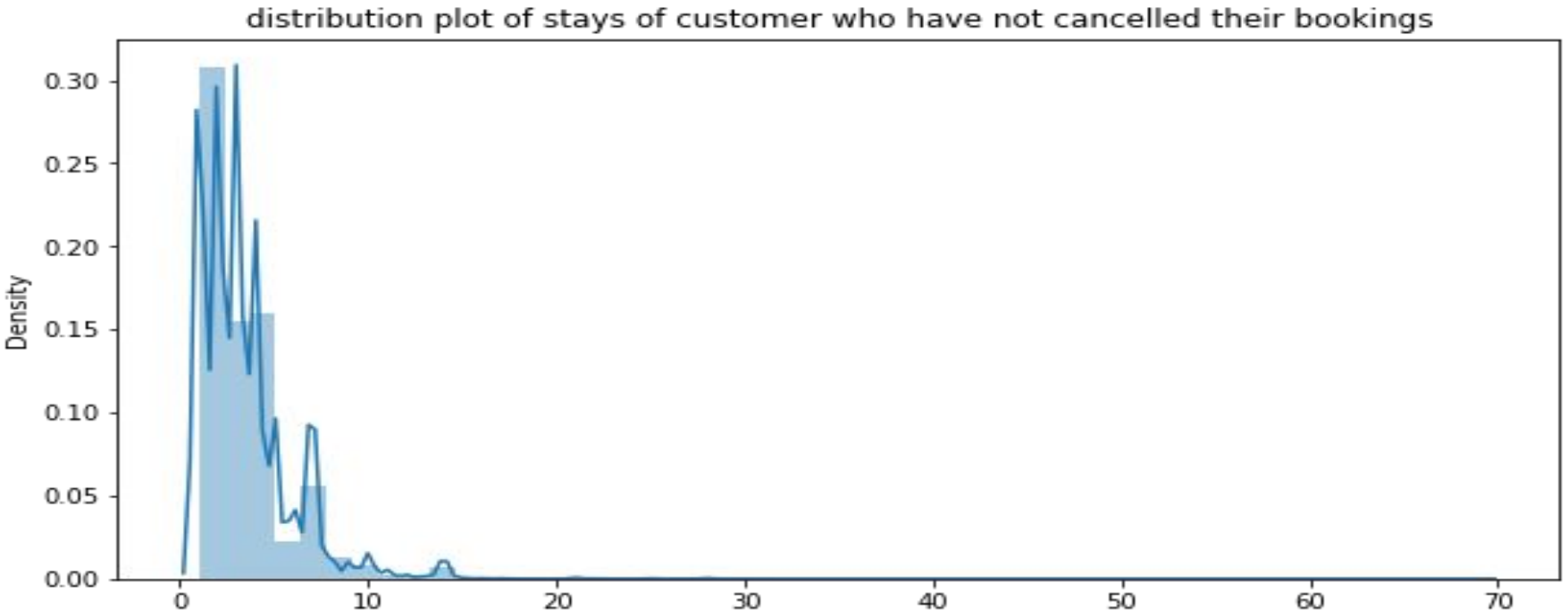
Transient type customers visit most and they require parking space most followed by transient-party customers.

customer visiting which hotel requires parking-City or Resort??

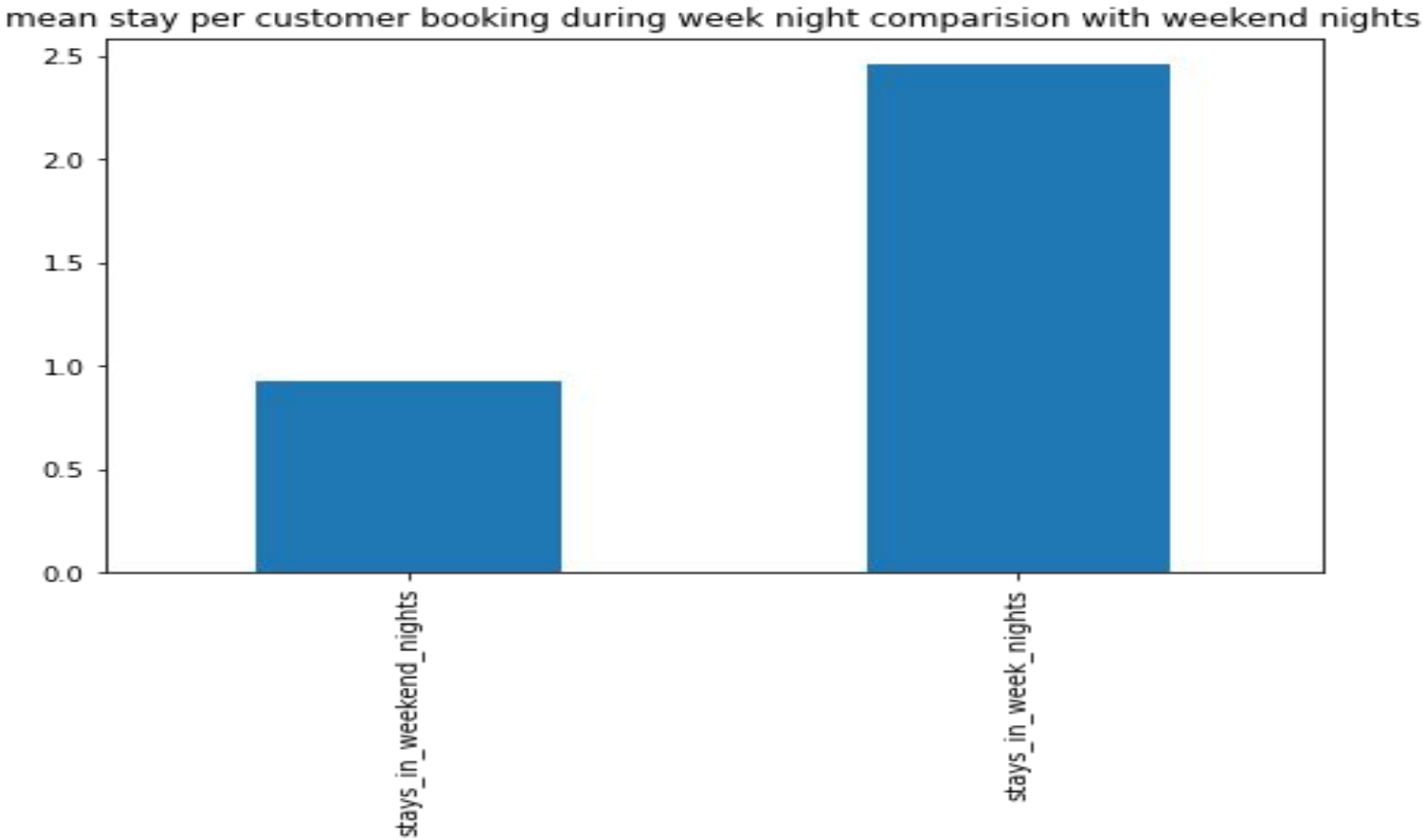


Among those who require parking space most are visiting Resort hotel.

How long do customers stay for??

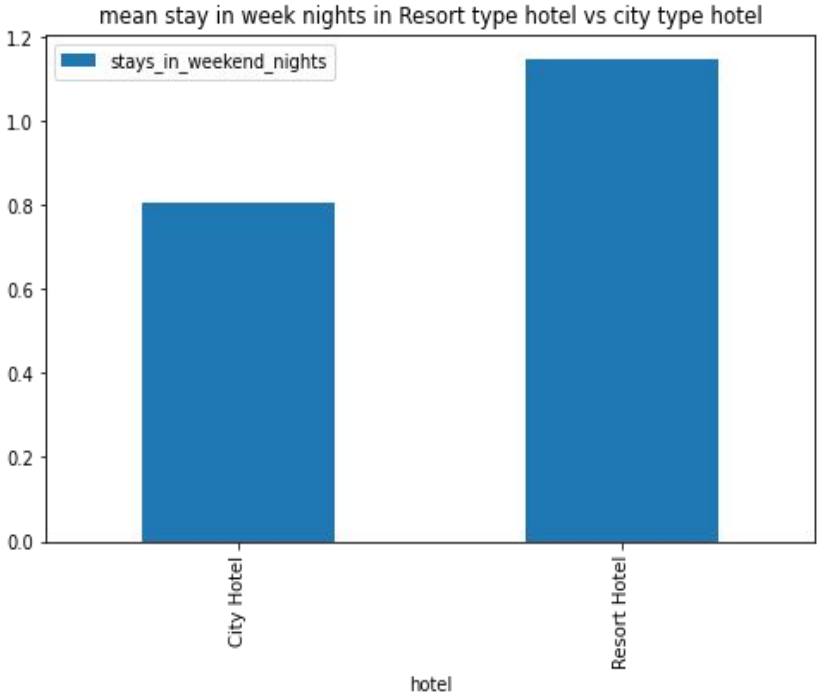


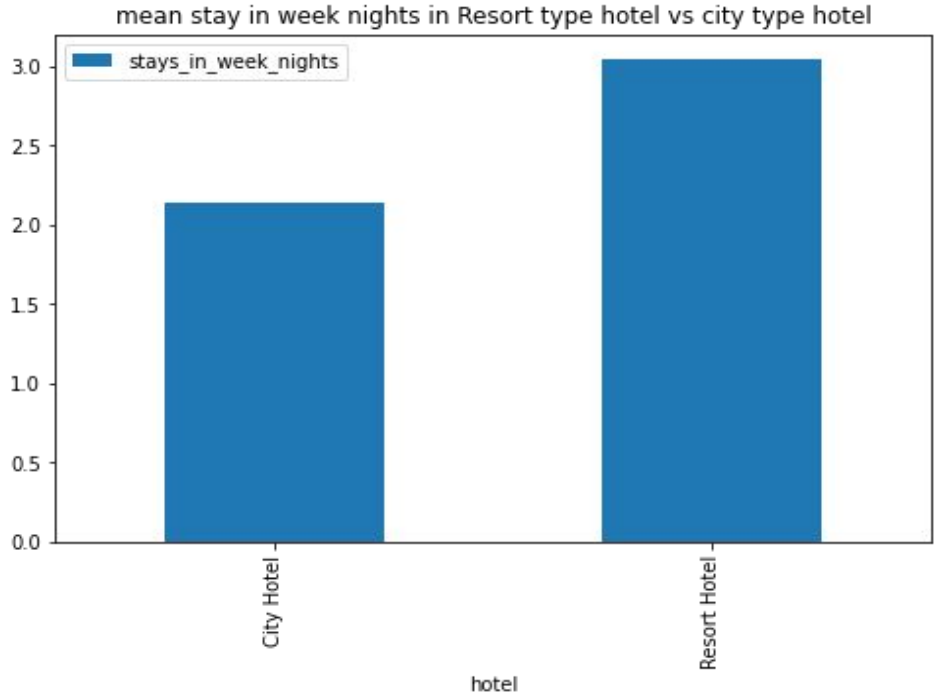
Mean stay of customers is 3.42 days maximum being 69 days and mode being 1 days.



Out of 119390 bookings,stay during weekend nights was around 1 night while stay during week nights wa around 2.5 nights.

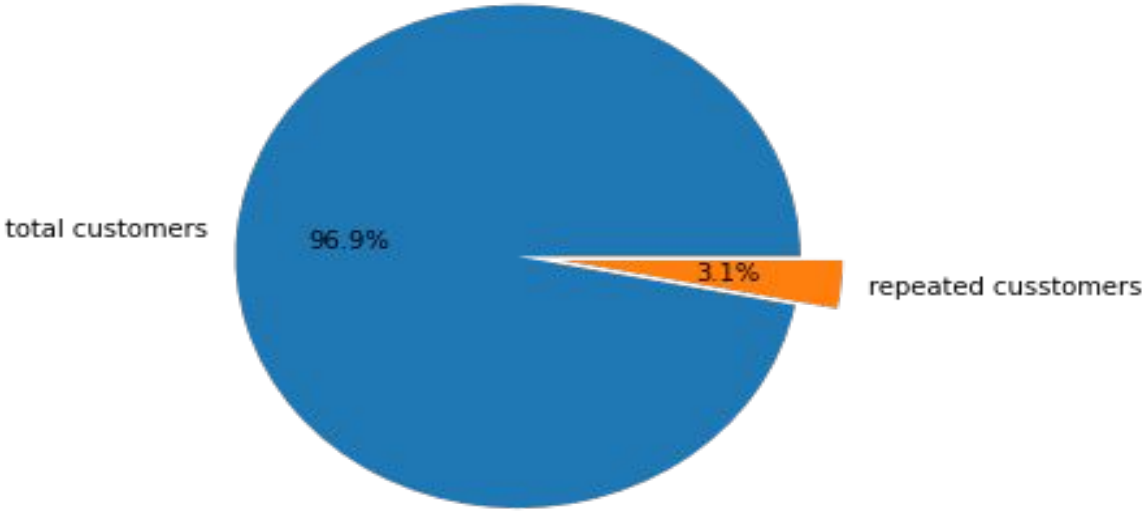
Which hotel do you prefer for longer stay??





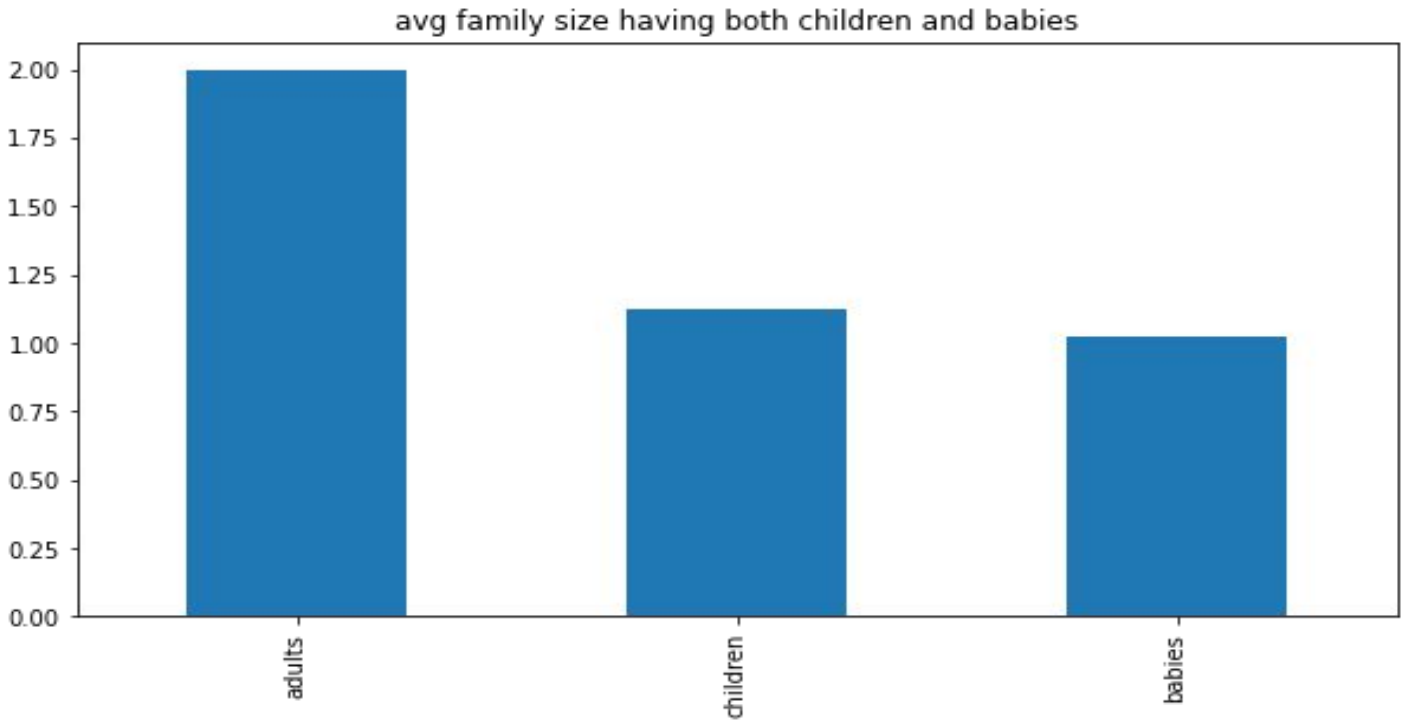
Clearly we can conclude that mean stay of customers who opted for Resort hotel stay for longer time. Even the person who stayed the longest i.e 69 days also stayed in resort type hotel.

How many customers Revisit??

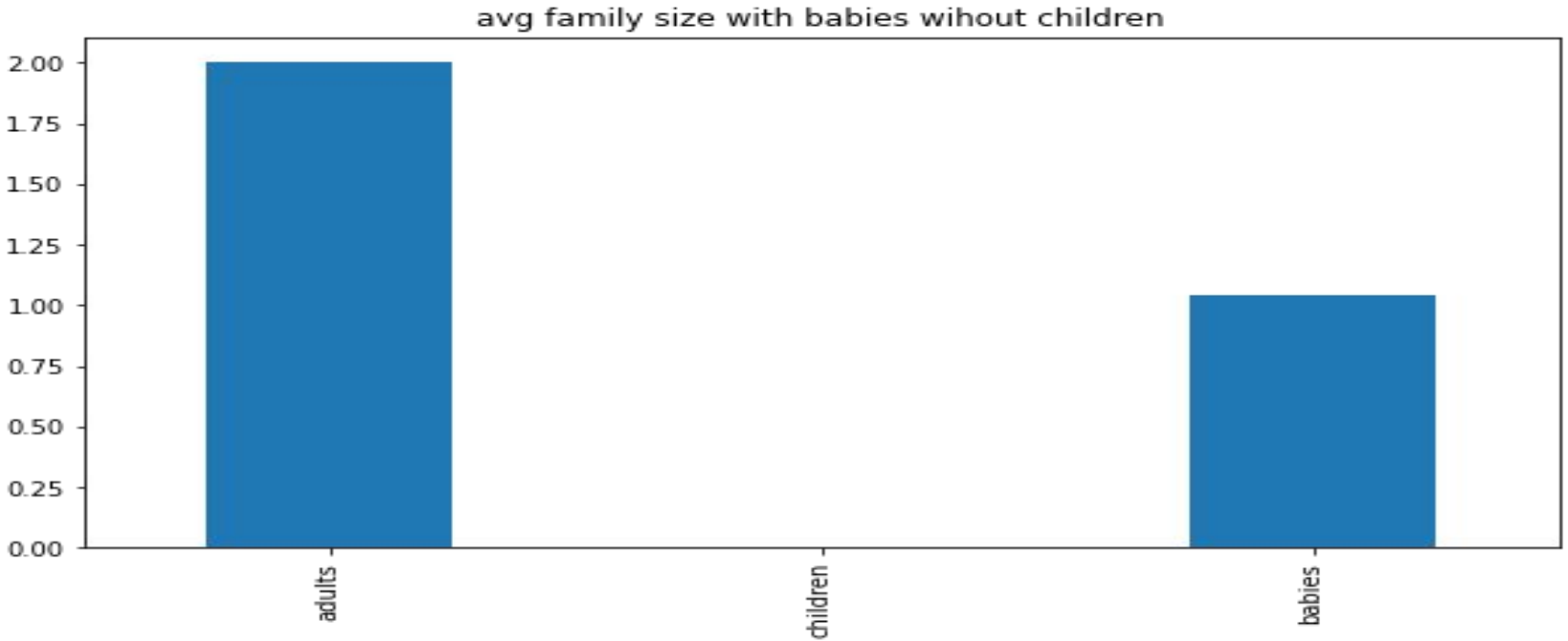


Just 3.1 % are repeated while 96.9% are new customers.Repeated customers stay for mean 1.85 days max being 45 days.

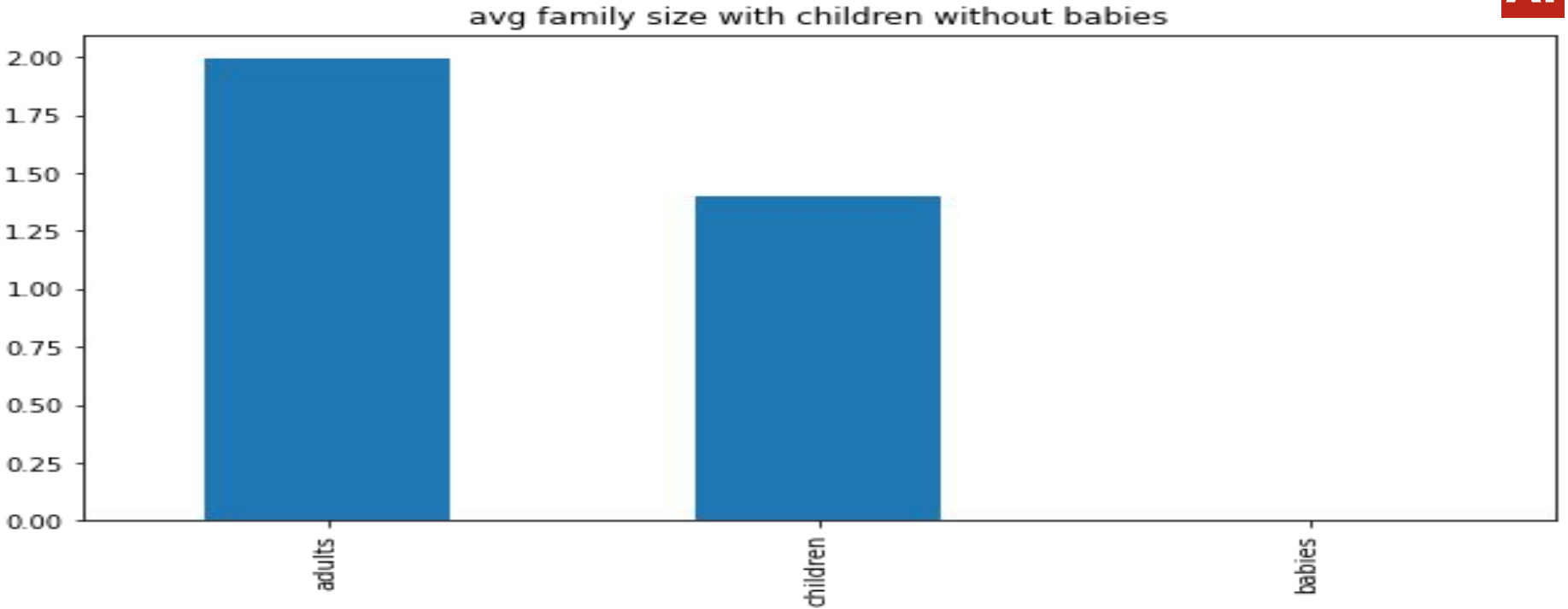
What is the size of the family of customers who visit hotel??



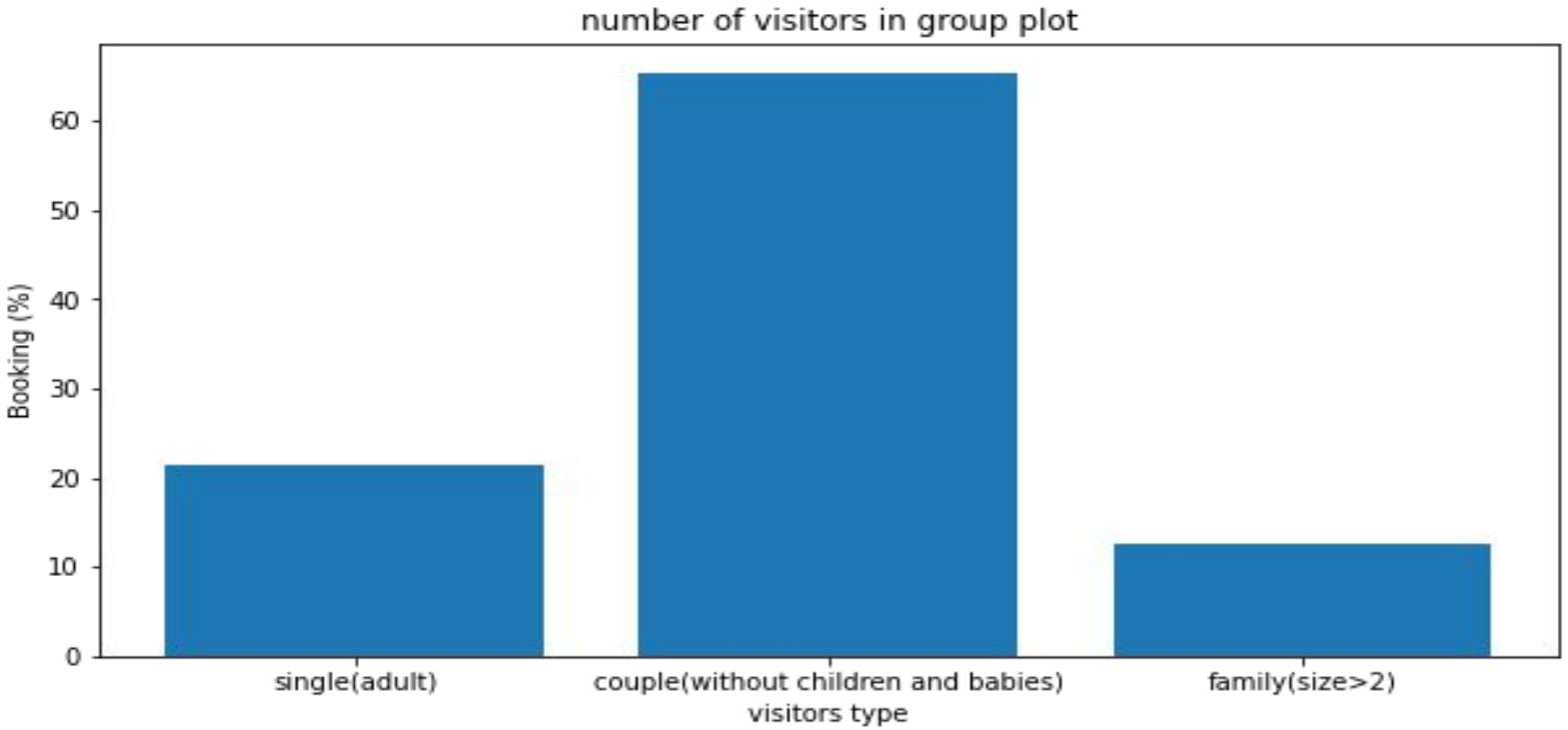
Average size of family who have both children and babies is around 4 members and consists of 2 adults ,1 children and 1 baby.



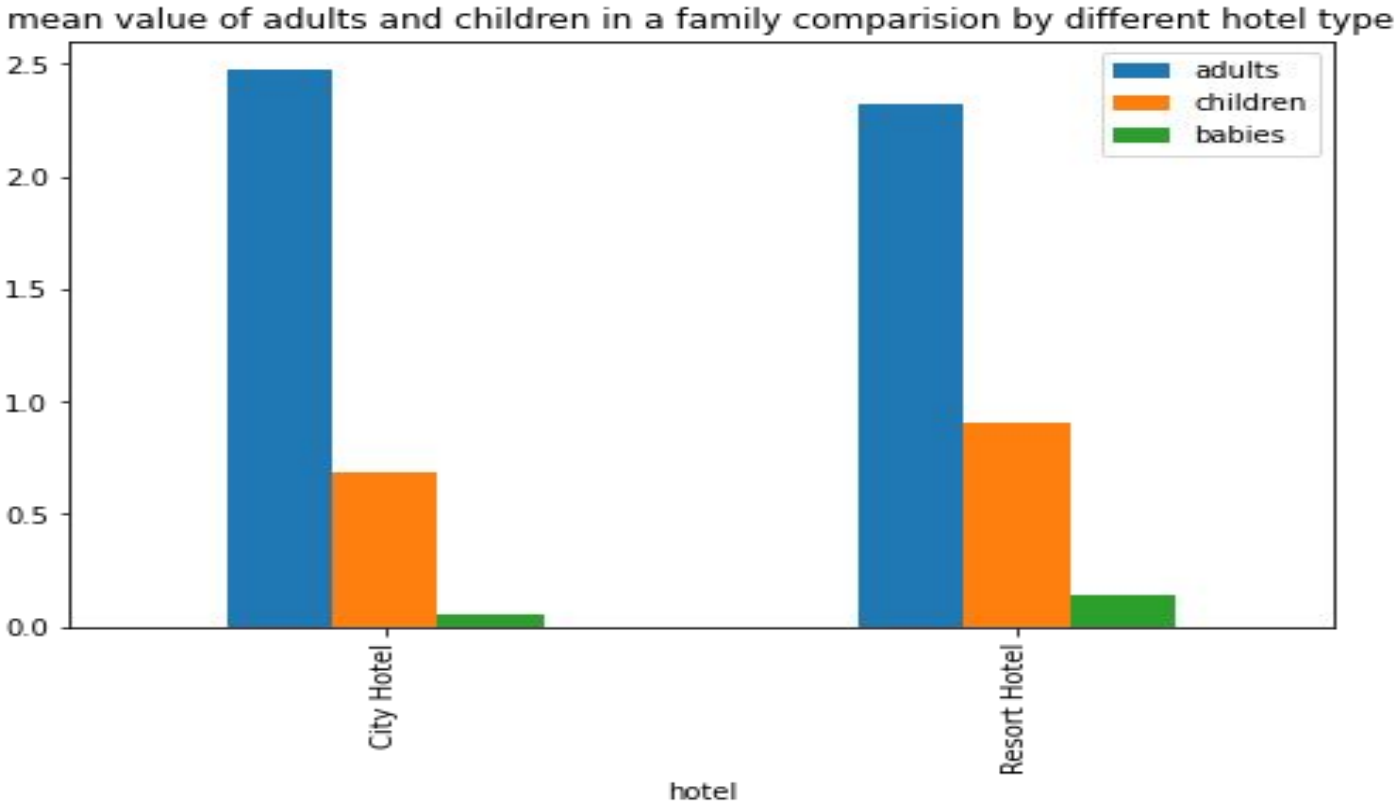
Average size of family having no children but baby is 3 members consists of two adults and one baby.



Average size of family having no babies but children is around 3.5 members having 2 adults and 1.4 children.

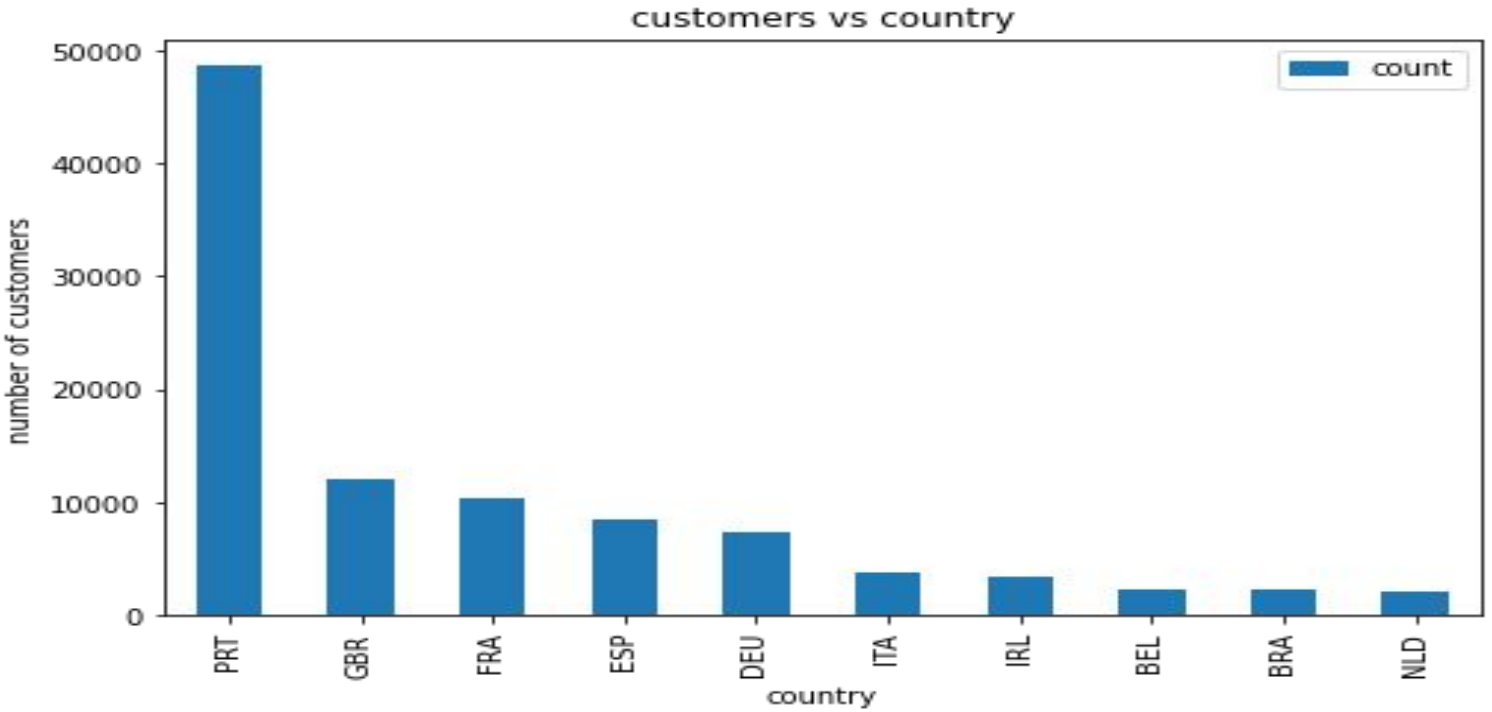


Around 60 % of people who visit hotels are couple having no children or babies followed by single adult followed by families having size 3 or above.



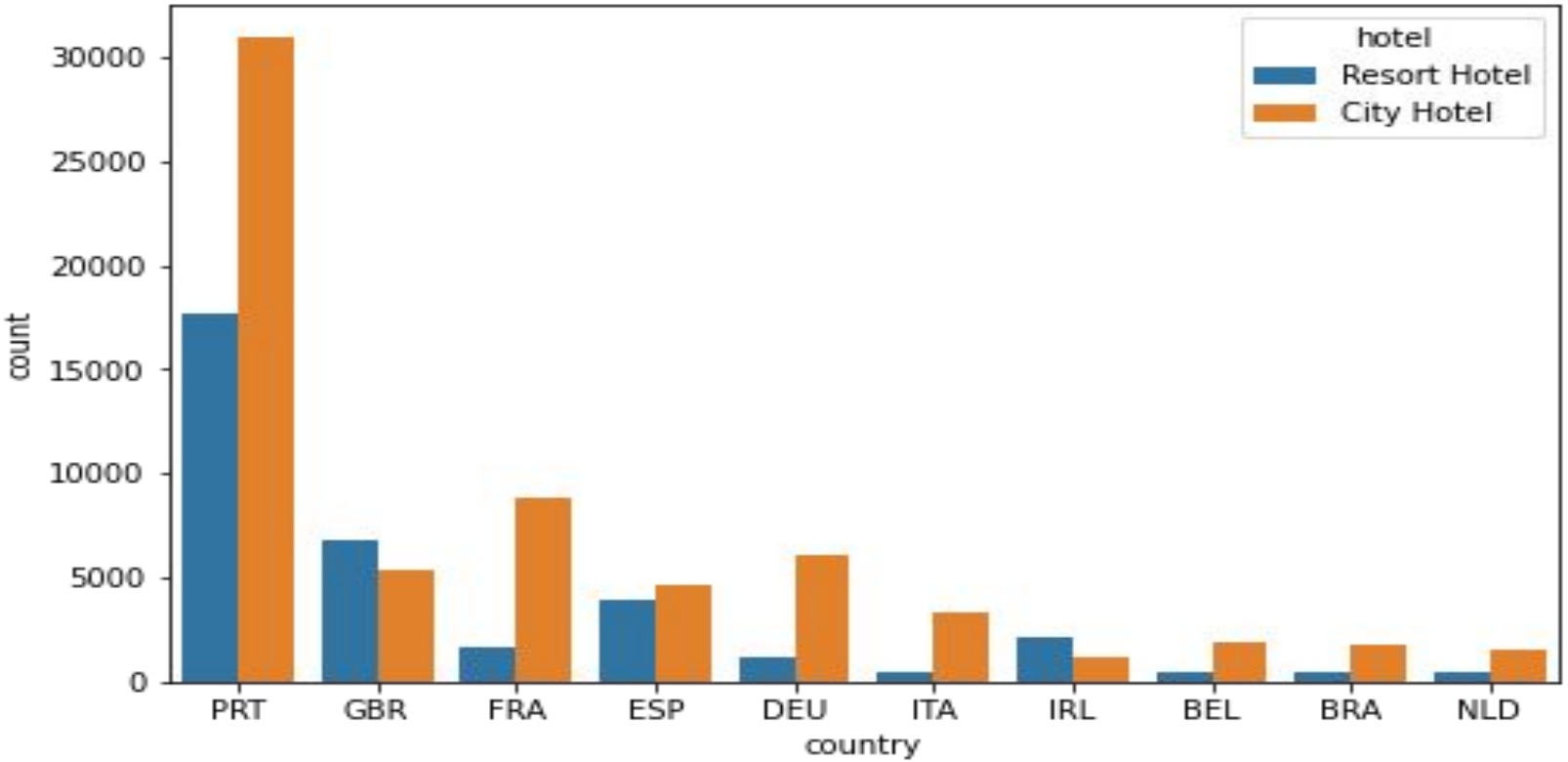
Families having more number of childrens and babies prefer resort type hotel.

What countries is this data from??

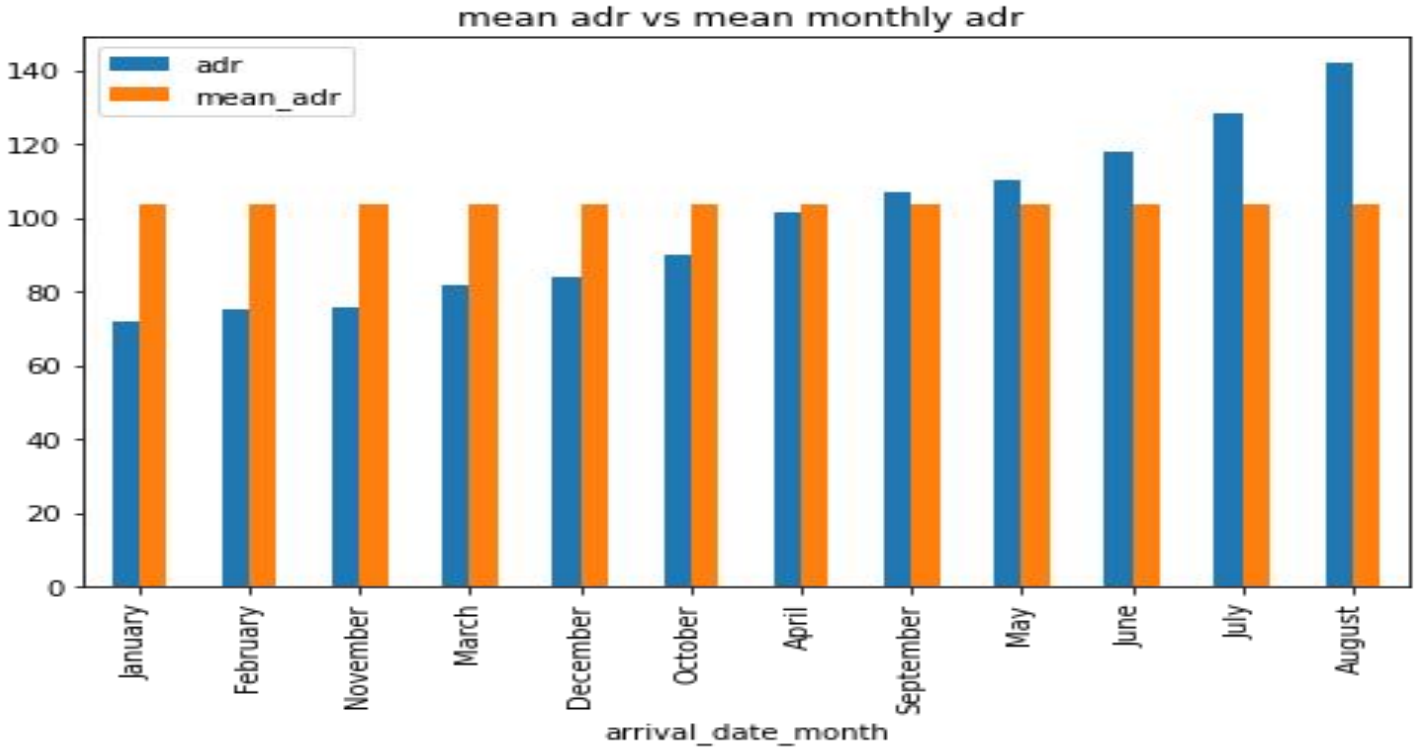


This data is from around 177 countries but almost 80% data is from european countries.

Based on hotel type

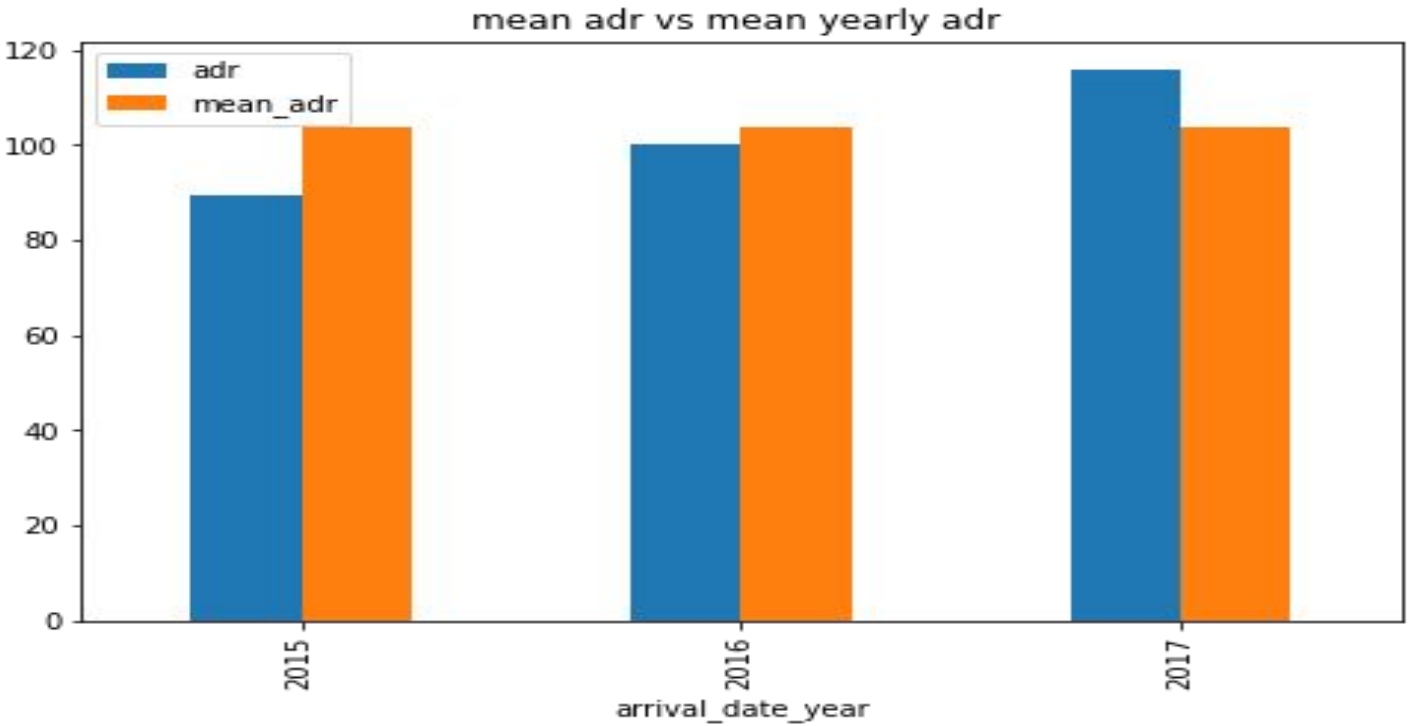


What about average daily rate??



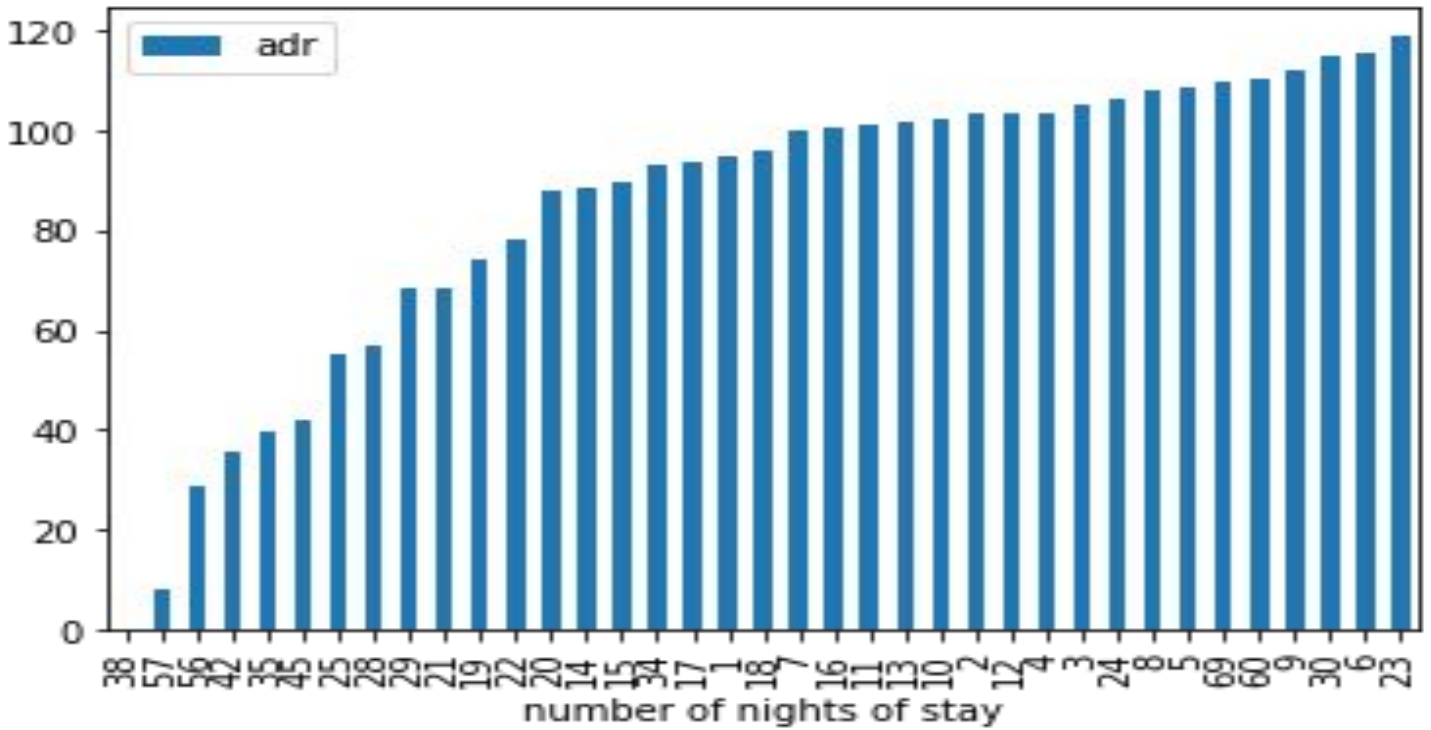
We can see from the chart that months may,june,july,august,september which see higher demand also have higher average daily rate.

Yearly variation of average daily rate??



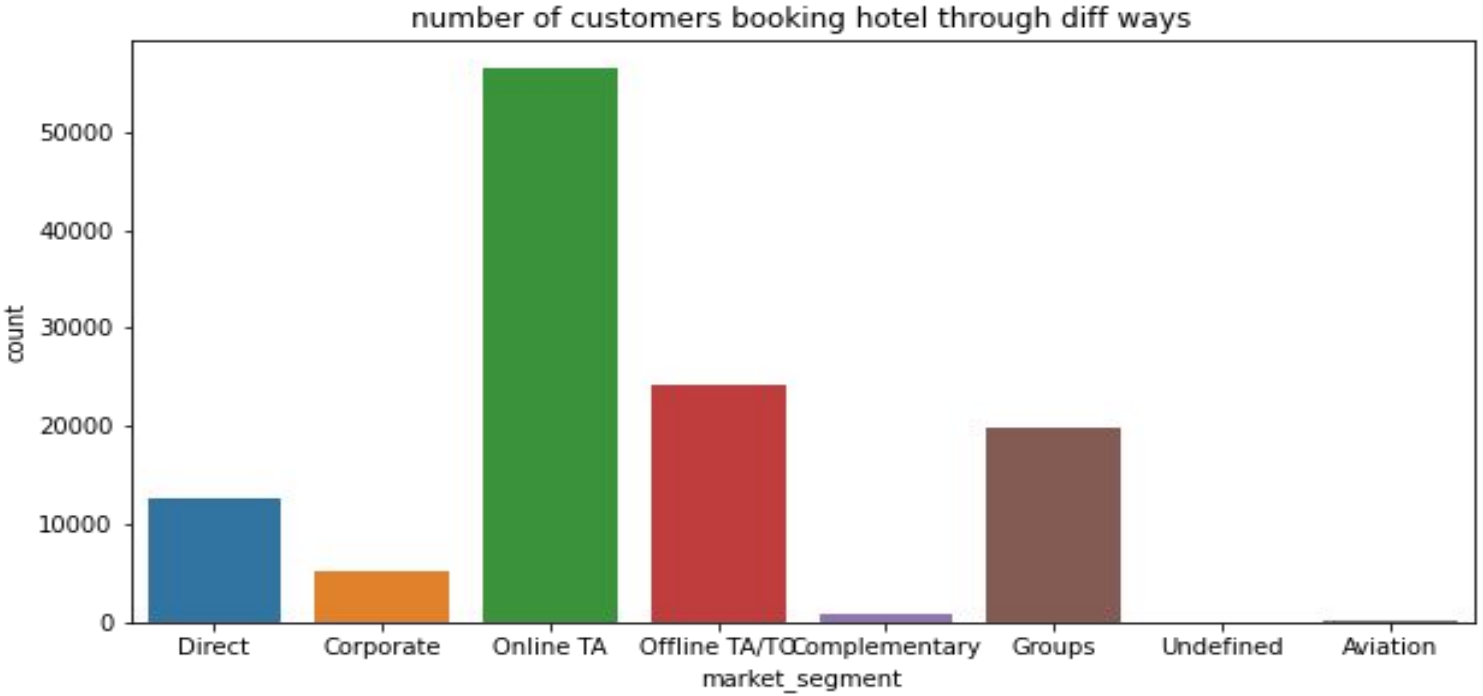
We can see that the most recent year has higher adr as compared to previous ones.This might be because of inflation.

Variation of adr with number of nights of stay



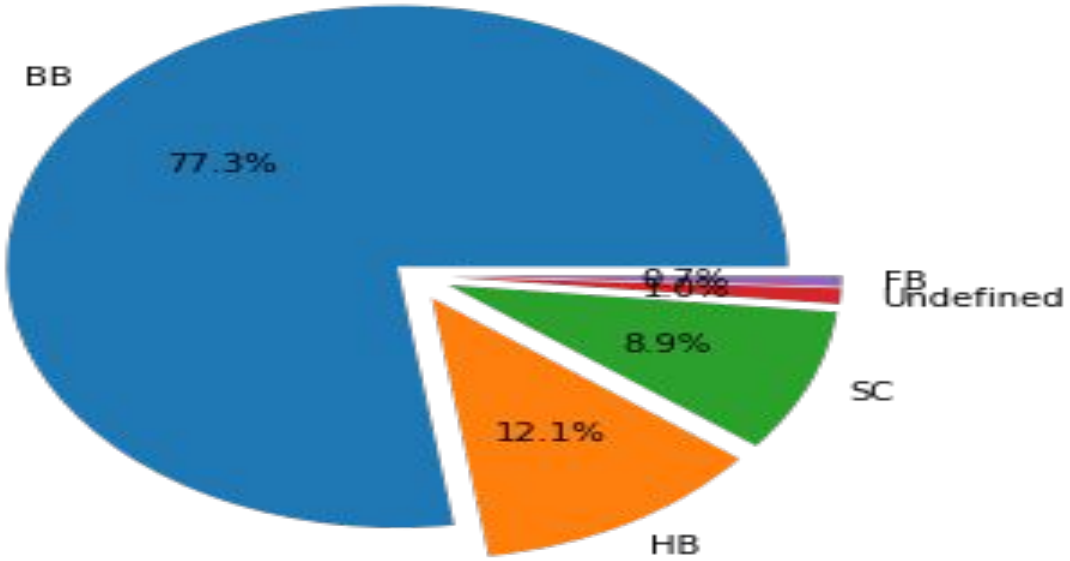
There isn’t any trend followed in this variation.It’s tough to find any conclusion from this data.

Market segment vs customer bookings



Online travelling agent is most preferred type followed by offline TA .

How many bookings opted for a meal??



Around 90% of the bookings opted for meal and breakfast and bed (BB) was the most preferred type of meal.

Conclusion

- May to September months were getting the most amount of bookings and average adr for these months was also high as compared to other months.
- Correlation coefficient came out around 0.82 which represents strong positive correlation.
- Around 30% of bookings were getting canceled.
- Correlation coefficient of 0.98 was found out between monthly cancellations and bookings which represents almost perfect correlation.
- City type hotels were getting twice the number of bookings than Resort type.
- People who ought to stay longer preferred Resort type hotels.
- Found out families having children and babies preferred Resort type hotels over City types a little more.
- Just around 6-7% of people required parking spaces and most of them opted for Resort type hotels.
- Couples(no babies and children) booked around 60% of the bookings followed by single adults.
- Most of the bookings are done through online traveling agents.
- Around 3 % people had to wait .Since most people opt for city type hotels ,city type has more avg waiting time.
- Almost 98 % of customers opted for meals and most opted for bed and breakfast time i.e BB type.
- 3.42 nights .Distribution plot was drawn for clear visualization.

THANK YOU