Cohort Azaadi - Capstone Project - 1

Hotel Booking Analysis

By:

Bhupathi Sampath

Samastha849@gmail.com

https://www.linkedin.com/in/bhupathi-sampath-a36574185

BhupathiSampath (Bhupathi Sampath) (github.com)

Problem Statement

- 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 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 analyse the data to discover important factors that govern the bookings.

About Dataset

- This dataset is representing the hotel bookings happened over the time.
- It contains 32 columns and 119390 records.

1. **Hotel** : Hotel(Resort Hotel or City Hotel)
2. **Is cancelled** : Value indicating if the booking was cancelled (1) or not (0)
3. **Lead time** : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
4. **Arrival date year** : Year of arrival date
5. **Arrival date month** : Month of arrival date
6. **Arrival date week number** : Week number of year for arrival date
7. **Arrival date day of month** : Day of arrival date
8. **Stays in weekend nights** : Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
9. **Stays in week-nights** : Number of week-nights (Monday to Friday) the guest stayed or booked to stay at the hotel
10. **Adults** : Number of adults
11. C**hildren** : Number of children
12. B**abies** : Number of babies
13. M**eal** : Type of meal booked. Categories are presented in standard hospitality meal packages
14. C**ountry** : Country of origin.`
15. M**arket segment** : Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
16. D**istribution channel** : Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"
17. I**s repeated guest** : Value indicating if the booking name was from a repeated guest (1) or not (0)
18. P**revious cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking
19. P**revious bookings not canceled** : Number of previous bookings not cancelled by the customer prior to the current booking.
20. R**eserved room type** : Code of room type reserved. Code is presented instead of designation for anonymity reasons.

Cont'd..

21. **Assigned room type** : Code for the type of room assigned to the booking.
22. **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
23. D**eposit type** : Indication on if the customer made a deposit to guarantee the booking.
24. A**gent** : ID of the travel agency that made the booking
25. C**ompany** : ID of the company/entity that made the booking or responsible for paying the booking.
26. D**ays in waiting list** : Number of days the booking was in the waiting list before it was confirmed to the customer
27. **Customer type** : Type of booking, assuming one of four categories
28. A**dr** : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
29. R**equired car parking spaces** : Number of car parking spaces required by the customer
30. T**otal of special requests** : Number of special requests made by the customer (e.g., twin bed or high floor)
31. R**eservation status** : Reservation last status, assuming one of three categories

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

32. R**eservation 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

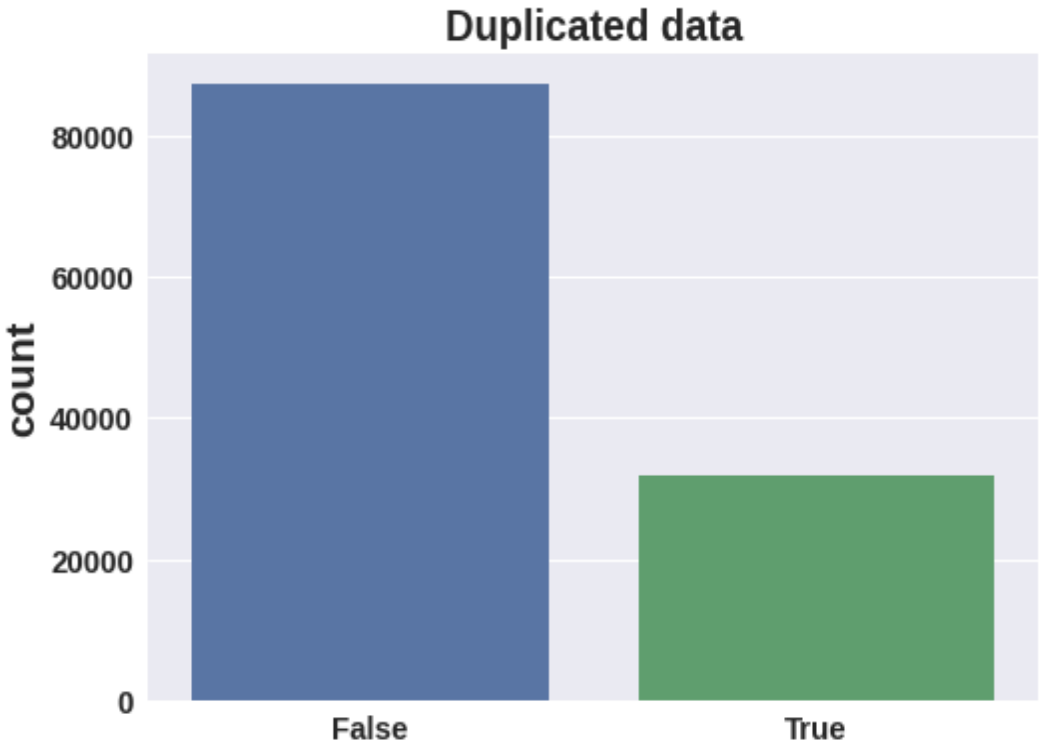
Data Cleaning

(I). <u>Duplicated Data:</u> We can see in the diagram in right side that dataset is containing 31994 duplicated data. So in order to get proper analysis of data I have deleted the duplicated data.

(II). <u>Null Values:</u>

1. *Company – 82137:* Fill null values with "**0**" by assuming company is others category.
2. *Agent – 12193:* Fill null values with "**0**" by assuming company is others category.
3. *Country – 452:* Fill null values with "**Others**" by assuming while collecting data user selected the option "Other".
4. *Children – 4:* Filled children column with "**0**" by assuming no children

(III). <u>Remove rows:</u> Found 166 records where children, adult and babies are having value 0. Which means there is no bookings made.
So I have decided to remove those 166 records to avoid inconsistency in data exploration.



(IV). After the data cleaning, the final dataset shape is 32 columns with 87230 records are there

Categorical columns & their unique values

1. ***Columns:*** ['customer_type', 'meal', 'hotel', 'distribution_channel', 'reserved_room_type', 'market_segment', 'deposit_type', 'assigned_room_type', 'reservation_status']
2. **Values:**

- *Customer type:* ['Transient', 'Contract', 'Transient-Party', 'Group'] Unique
- *Meal:* ['BB', 'FB', 'HB', 'SC', 'Undefined']
- *hotel:* ['Resort Hotel', 'City Hotel']
- *distribution channel:* ['Direct', 'Corporate', 'TA/TO', 'Undefined', 'GDS']
- *reserved room type:* ['C', 'A', 'D', 'E', 'G', 'F', 'H', 'L', 'B']
- *market segment:* ['Direct', 'Corporate', 'Online TA', 'Offline TA/TO', 'Complementary', 'Groups', 'Undefined', 'Aviation']
- *deposit type:* ['No Deposit', 'Refundable', 'Non-Refund']
- *assigned room type:* ['C', 'A', 'D', 'E', 'G', 'F', 'I', 'B', 'H', 'L', 'K']
- *reservation status:* ['Check-Out', 'Canceled', 'No-Show']

Extra added columns

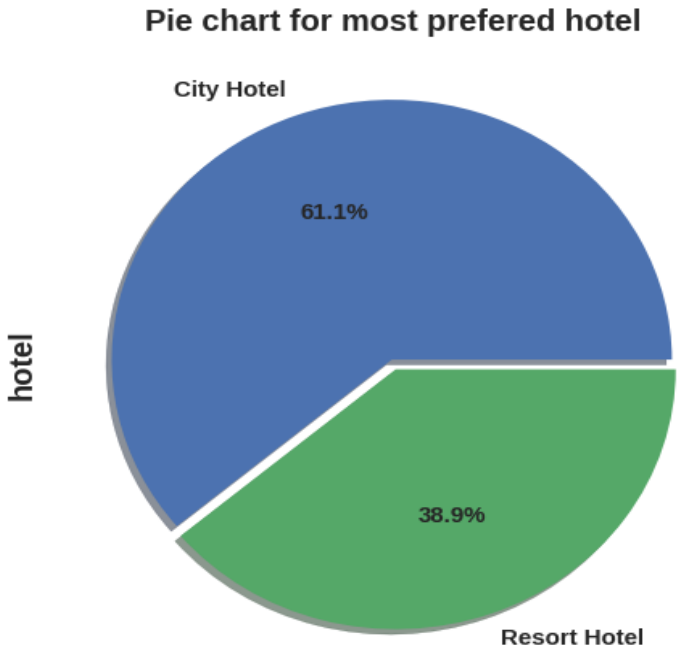
1. ***Total Members:*** Total number of guests in every booking. By adding (Adults, Children and Babies)
2. ***Total Stays:*** Total number of days guests stayed in hotel. By adding (Stays in week-nights and Stays in weekend-nights)

Questions

1. Which hotel type is most preferred by costumers?
2. Which type of customers has more bookings?
3. Which type of rooms are most preferred rooms?
4. What is the cancellation of bookings with respect to Distribution Channel?
5. What are the reasons for cancellation of bookings?
6. Which agent made more bookings?
7. What is the percentage of repeated guests?
8. What is the percentage distribution of required car parking spaces?
9. What is the percentage of booking changes made by the customer?
10. Which distribution channel is made more bookings?
11. Which distribution channel is used for early bookings?
12. Which distribution channel making good revenue generating with respect to hotel?
13. What is the adr across the market segments?
14. What is the number of bookings by month wise?
15. What is the adr across the months?

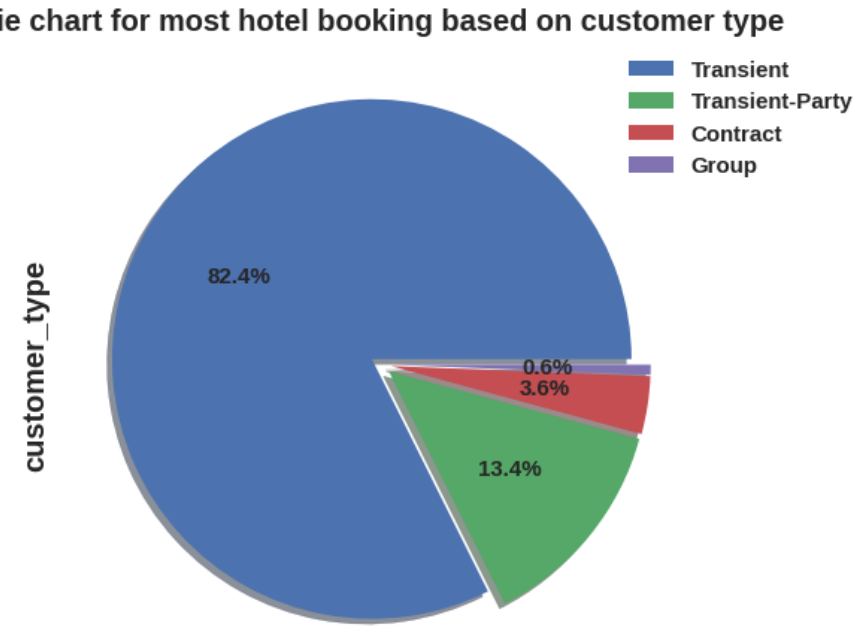
Exploratory Analysis

Most preferred hotel



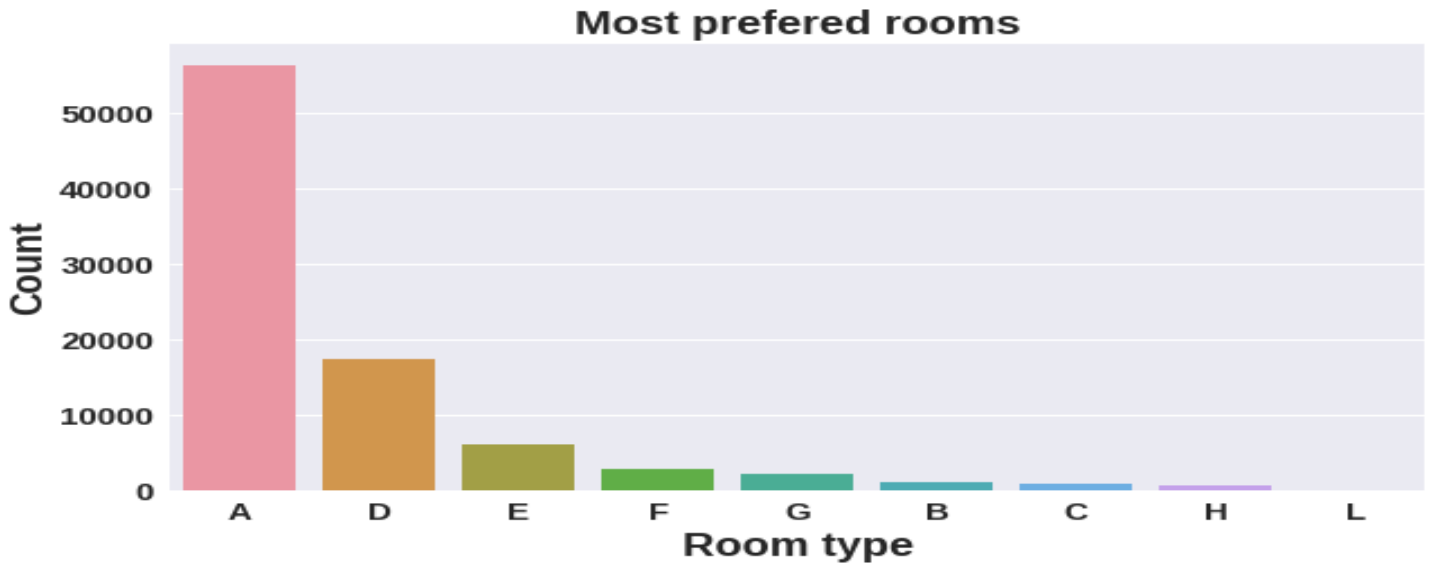
- City hotel is the most preferred hotel where Resort hotel is less preferred with 61.1%.
- Hence, we can say city hotel is the busiest hotel.

Hotel bookings by customer type



Customer type Transient has more bookings with 82.4%.	Transient-party guests has 13.4%	Contract type guests has 3.6%.	Group type has less with 0.6%.
1. Transient: When the booking is not a part of group or contract, and it is not associated with other transient booking	**2.** Transient-party: When booking is transient, and associated with at least one other transient booking	**3.** Contract: When the booking has allotment or other type of contract associated to it.	4. Group: When the booking is associated to group

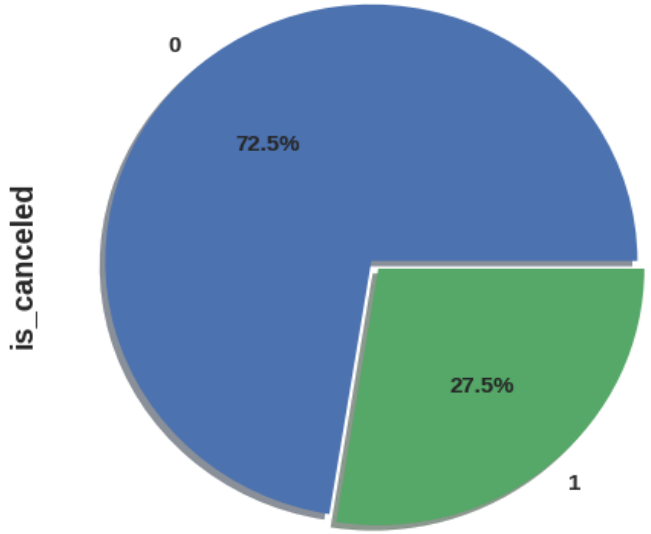
Exploratory Analysis



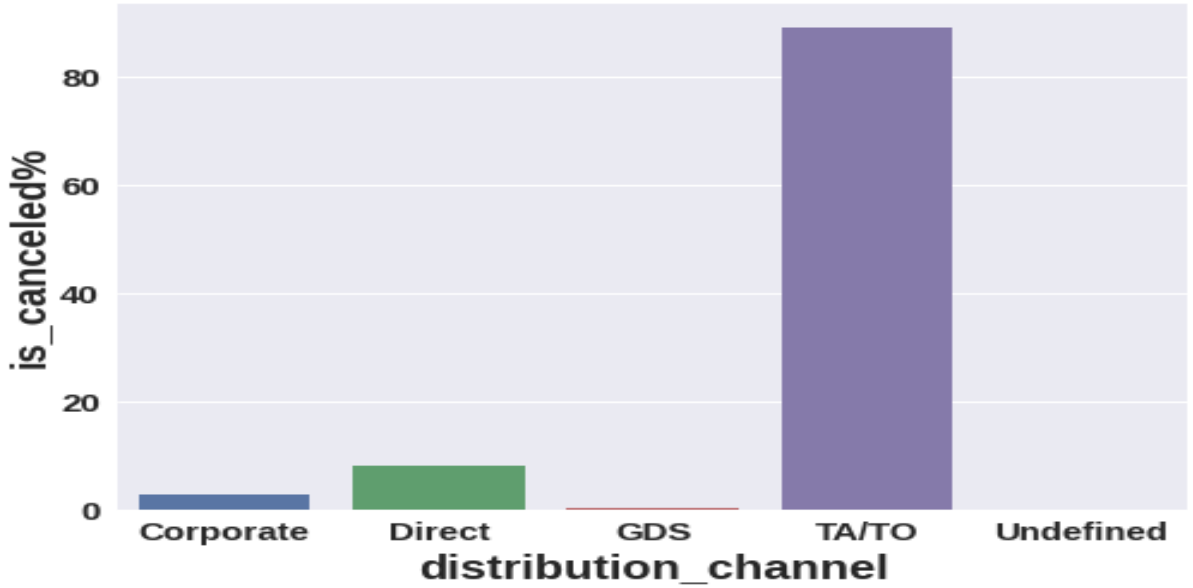
- Most of the guests are preferring the rooms "A"(Code of room type). So, Code "A" type rooms can be increased to increase the bookings

- By looking at above result we can see agent with ID Number: 9 has done more bookings 28721.

Exploratory Analysis



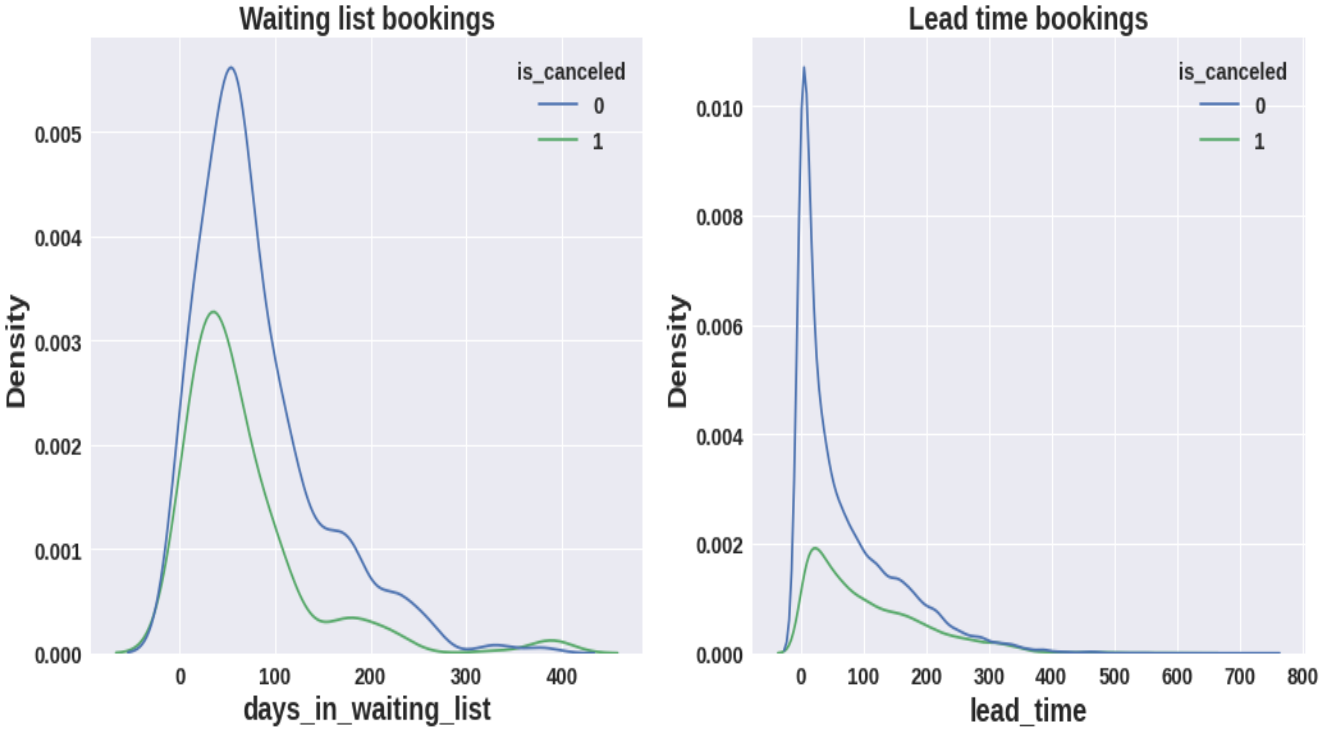
- 0 = Not cancelled
- 1 = Cancelled
- And we can see in above result that 27.5% bookings has been cancelled



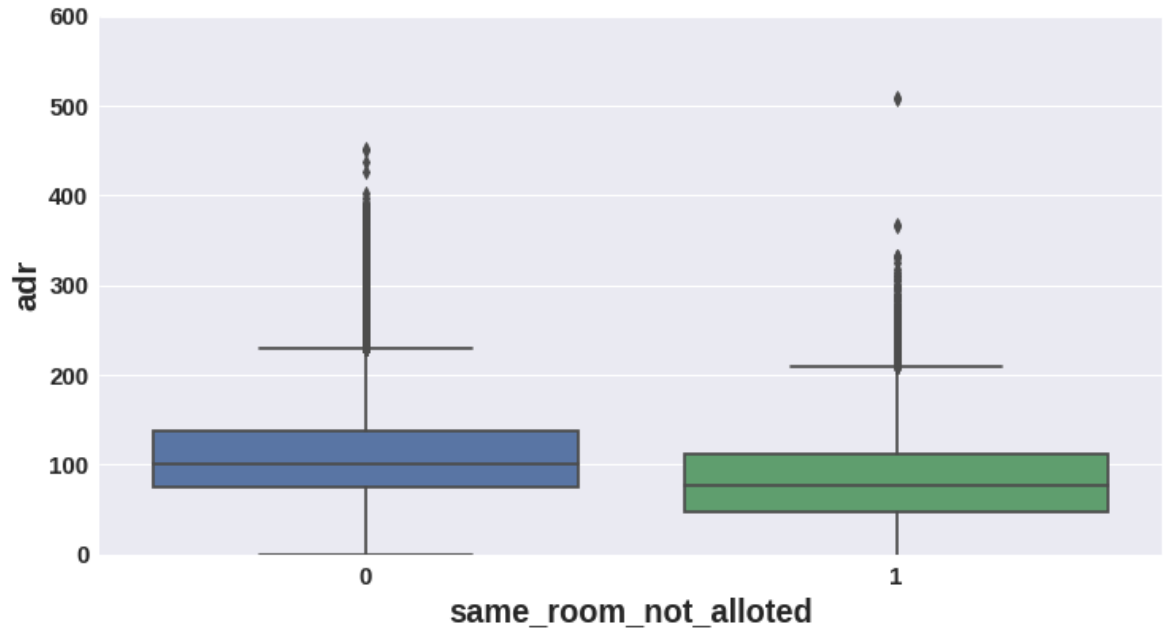
- By looking at above graph we can see distribution channel "TA/TO" has more cancellations with 83.9%.
- Direct has 8.3%
- Corporate has 2.69%
- GDS channel has 0.15% and Others has 0.02%

•Corporate- These are corporate hotel booing companies which makes bookings possible.
•GDS-A GDS is a worldwide conduit between travel bookers and suppliers, such as hotels and other accommodation providers. It communicates live product, price and availability data to travel agents and online booking engines and allows for automated transactions.
•Direct- means that bookings are directly made with the respective hotels
•TA/TO- means that bookings are made through travel agents or travel operators.
•Undefined- Bookings are undefined. may be customers made their bookings on arrival.

Exploratory Analysis



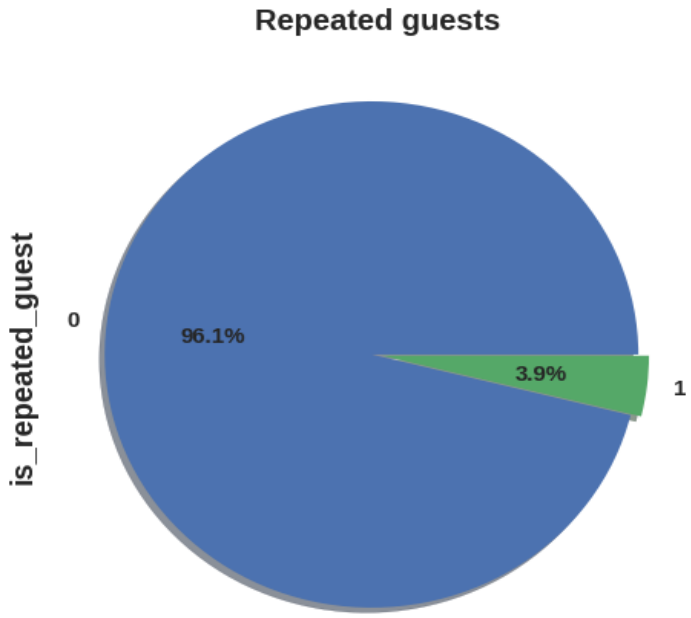
1. By looking at waiting list plot we can see less than 150 days waiting list has canceled and also there is not canceled bookings also more in less than 150 days waiting list. Hence, we can say days in waiting list is not much affecting the cancelation.
2. By looking at least time plot we can observe both the lines are similar with respect to the days in waiting list. Hence lead time also not much affecting the cancelation of bookings



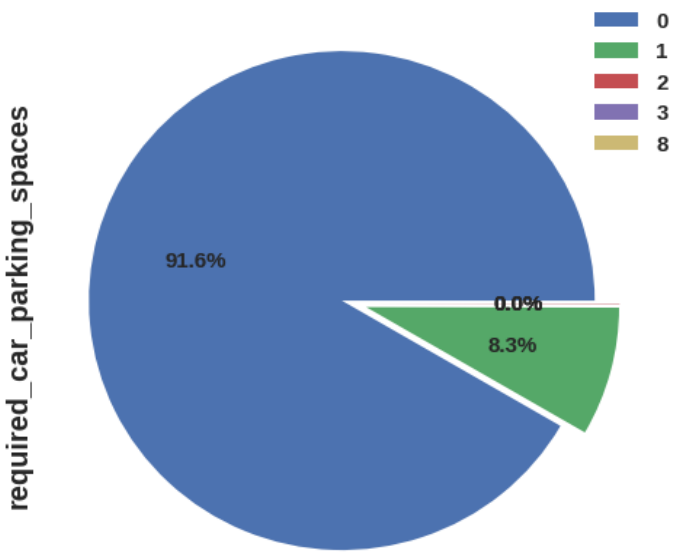
0: is same room allotted
1: is Same room is not allotted

- We can see in the plot same room allotted and not allotted are almost similar. Hence, Not getting same room is affecting the daily "adr". Guests who are not getting the same room are paying the less adr compared to same room allotted.

Exploratory Analysis

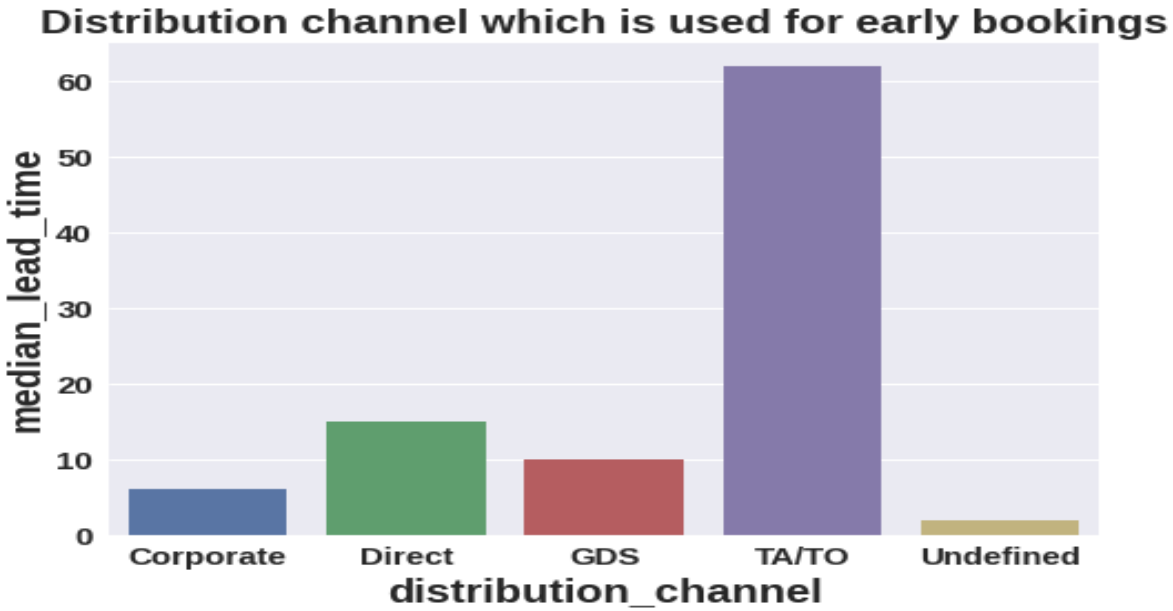


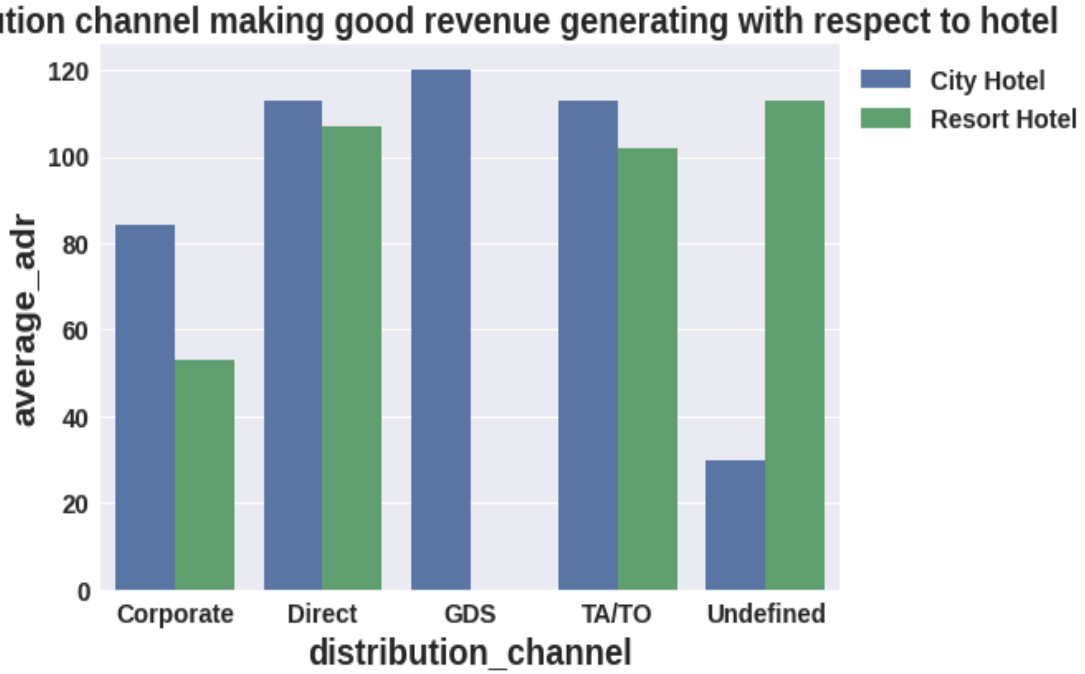
- 0 = Not repeated
- 1 = Repeated
- There is only few guest are repeated which is 3.9%.
- And 96.1% are not repeated

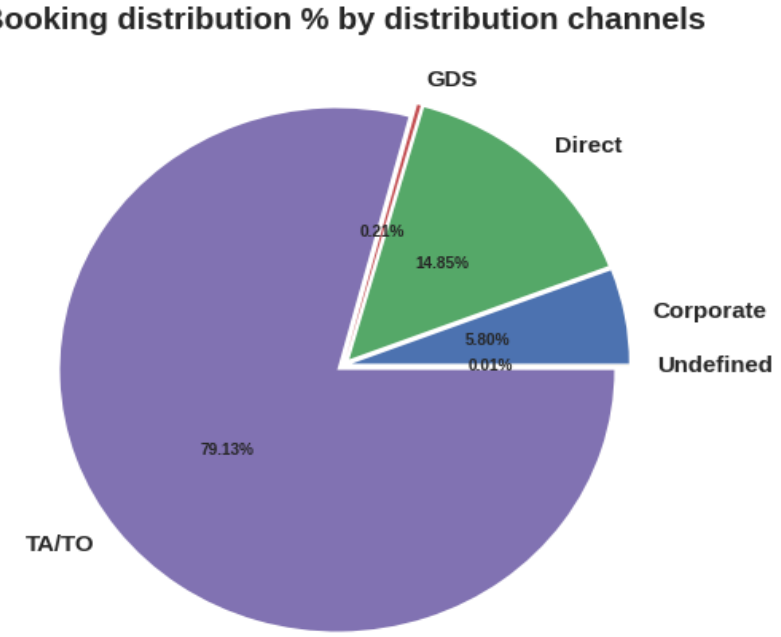


- 91.6% guests did not require the parking space where only 8.3% of guests required inly 1 parking space

Exploratory Analysis

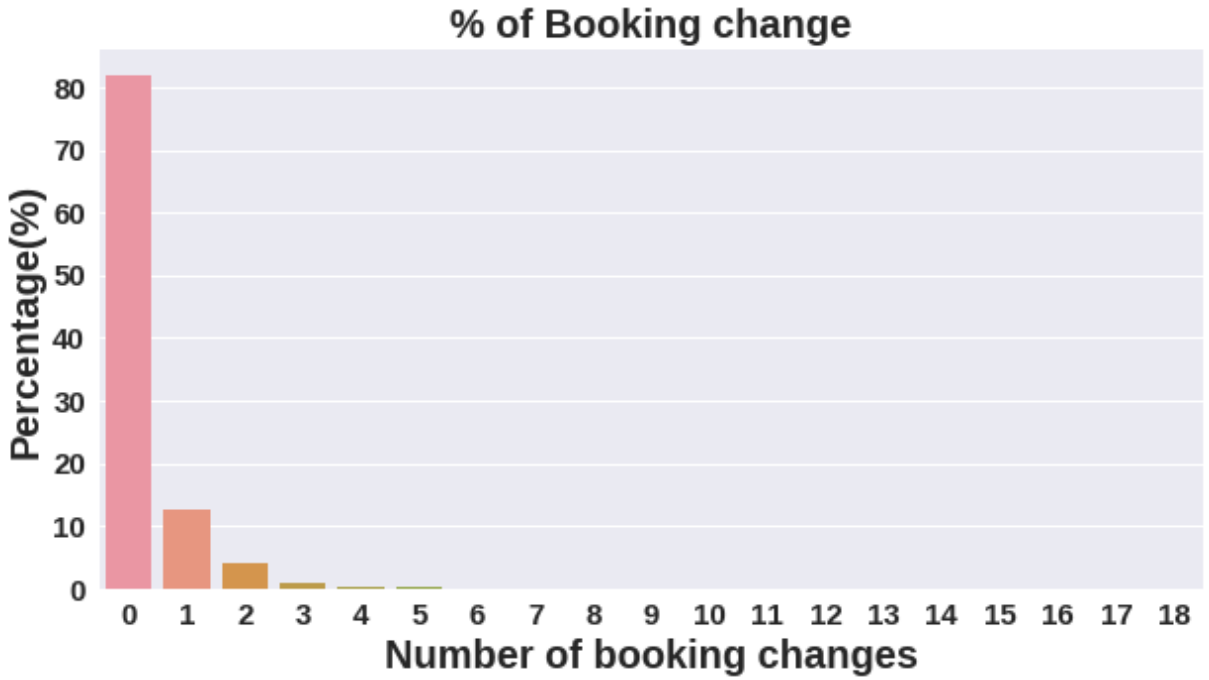


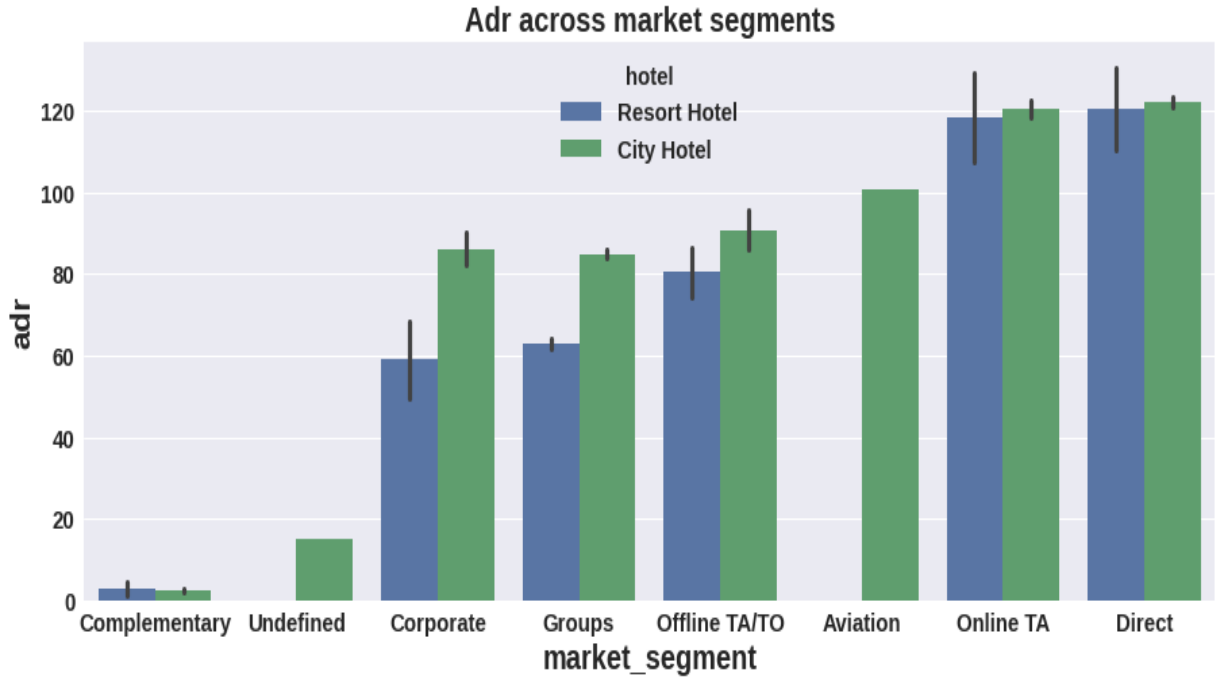




1. Distribution channel TA/TO has highest lead time. Hence, we can say TA/TO has the highest earlier bookings where Corporate has less lead time.
2. Distribution channels 'Direct' and 'TA/TO' are contributing the most in both types of hotels. GDS distribution channel should focus on increasing the bookings of 'City Hotel'.
3. Most of the bookings are made by the distribution channel TA/TO with 79.13% where Direct-14.85%, Corporate-5.80, GDS-0.21% and Undefined-0.0.1%

Exploratory Analysis

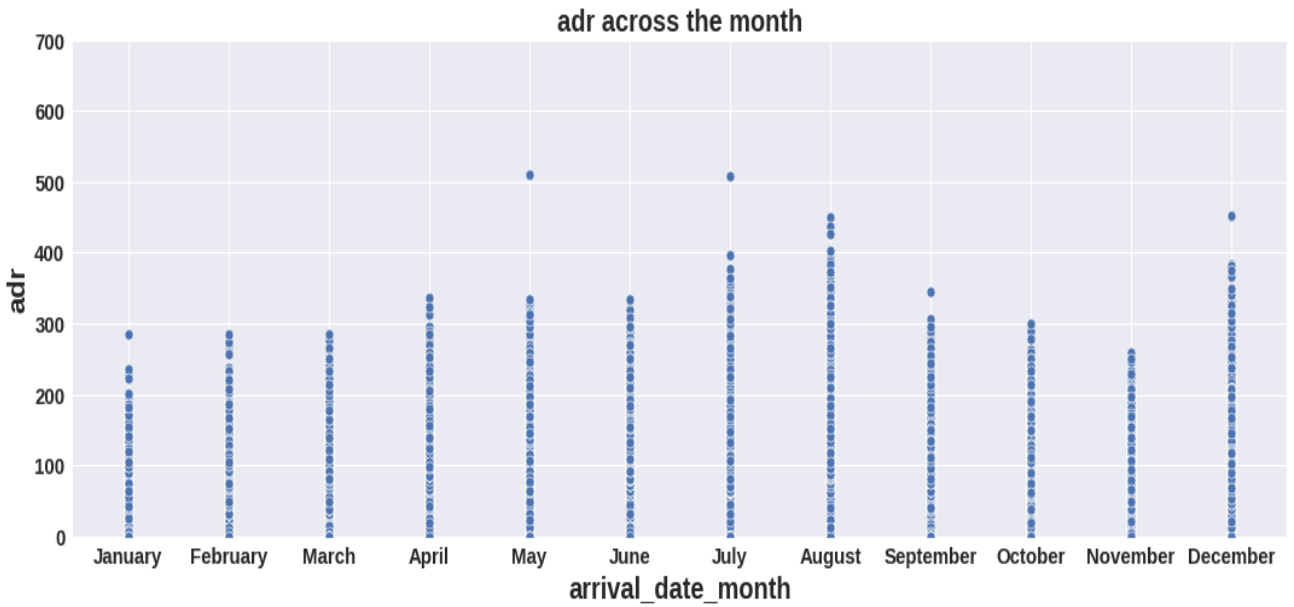


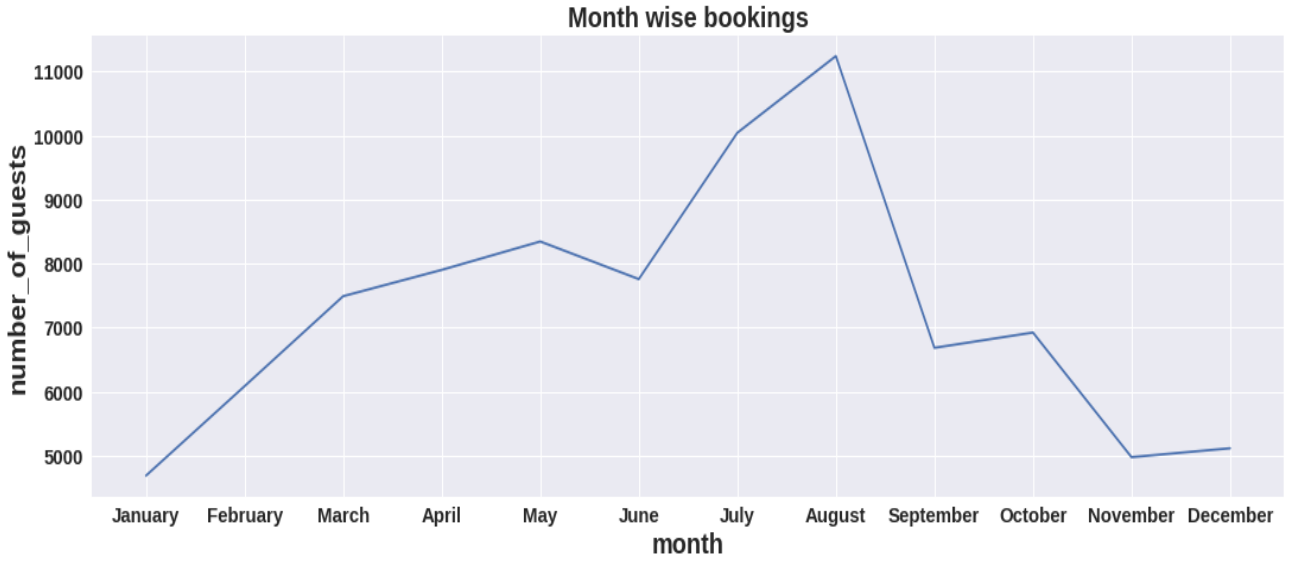


1. The percentage of "0" changes made in the bookings was more than 82%. And percentage of singe bookings was made 10% with second highest.
2. 'Direct' and 'Online TA' are contributing the most in both types of hotels. Aviation segment should focus on increasing the bookings of 'City Hotel'

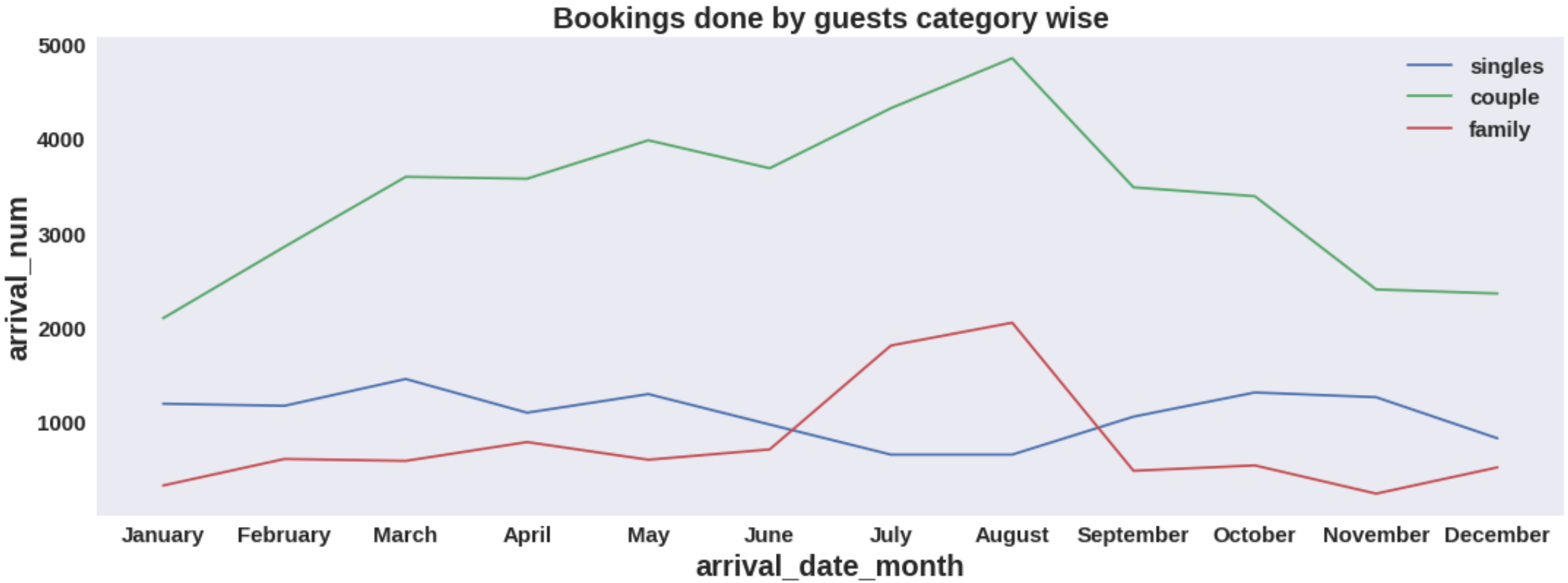
Exploratory Analysis

1. Bookings has been increasing ti ll the mid of year and we can se e in August bookings went to hi ghest. Hense, Most of the peopl e are planning trips in august m onth.
2. Avg adr rises from beginning of year up to middle of year and reaches peak in August and then lowers to the end of year. But hotels do make some good deals with high adr at end of year also

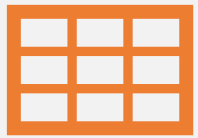




Exploratory Analysis



1. Mostly bookings are done by couples(although we are not sure that they are couple as data doesn't talk about that)
2. It is clear from graph that there is a sudden surge in arrival num of couples and family in months of July and August. So better plans can be planned accordingly at that time for these type of customers.



CREATING MORE COLUMNS BASED ON EXISTED DATA.

DIFFICULTY IN SELECTION OF COLUMNS TO GET INSIGHTS FROM DATA.

SELECTING THE VISUALIZATION.

Challenges

Conclusion

- City hotels are most preferred hotels. Thus, City hotels are busiest than Resorts.
- Transient customer type is more percentage of booking which is **82.4%** and Group type is low which is **0.6%.**
- Most of the guests are preferring the rooms "A"(Code of room type). So, Code "A" type rooms can be increased to increase the bookings.
- 27.5% bookings has been cancelled.
- Distribution channel "TA/TO" has more cancellations with 89.13%.
- Less than 150 days waiting list has cancelled and there is not cancelled bookings also more in less than 150 days waiting list. Hence, we can say days in waiting list is not much affecting the cancelation.
- Lead time also not much affecting the cancelation of bookings.
- same room allotted and not allotted are almost similar. Hence, Not getting same room is affecting the daily "adr". Guests who are not getting the same room are paying the less adr compared to same room allotted.
- Agent with ID Number: 9 has done more bookings.
- There is only few guest are repeated which is 3.9%.
- 91.6% guests did not require the parking space where only 8.3% of guests required inly 1 parking space.

Cont'd..

- Most of the guests are preferring the rooms "A"(Code of room type). So, Code "A" type rooms can be increased to increase the bookings
- distribution channel "TA/TO" has done more bookings and used for early bookings.
- Distribution channels 'Direct' and 'TA/TO' are contributing the most in both types of hotels. GDS distribution channel should focus on increasing the bookings of 'City Hotel'.
- Market segment "Direct" has the high adr in both Resort and City hotels where "Complementary" has less. Aviation segment can focus on City hotel.
- Bookings has been increasing till the mid of year and we can see in August bookings went to highest. Hense, Most of the people are planning trips in august month.
- Avg adr rises from beginning of year up to middle of year and reaches peak in August and then lowers to the end of year. But hotels do make some good deals with high adr at end of year also.
- Mostly bookings are done by couples(although we are not sure that they are couple as data doesn't talk about that).
- It is clear from graph that there is a sudden surge in arrival num of couples and family in months of July and August. So better plans can be planned accordingly at that time for these type of customers.

Thank you