

# Capstone Project

Hotel Booking data analysis and visualization

Pranay Umredkar  
Saitej Bhimnath  
Shweta Nayak  
Yuvaraj Mahajan

# Table of Contents:

- Objective
- Data Summary
- Data loading and exploration
- Data Wrangling
- Correlation heatmap
- Hotel wise analysis
- Hotel booking cancellation analysis
- Timewise analysis
- Distribution channel wise analysis
- Some other questions
- Challenges
- Conclusion

# Objective

We are going to analyse hotel bookings dataset for 3 years from 2015 - 2017.  
We will be discussing following steps in upcoming slides.

- Data loading and exploration.
- Data Wrangling
- Data analysis and visualization.
- Conclusion.

# Data Summary

The data table consists of 119,390 rows and 32 columns. Our analysis starts with defining each column and our understanding for each column mentioned below:

- **hotel** : Hotel type.
- **is\_canceled** : value indicates if the booking is canceled or not.
- **lead\_time** : How long in advance the booking was made.
- **arrival\_date\_year** : Customer arrival year.
- **arrival\_date\_month** : In which month of the year customer visited hotel.
- **arrival\_date\_week\_number** : In which week of the year customer arrived.
- **arrival\_date\_day\_of\_month** : Date of the month customer visited hotel.
- **stays\_in\_weekend\_nights** : Customer stayed or booked to stay in hotel during weekend nights.
- **stays\_in\_week\_nights** : Customer stayed in hotel during week nights.
- **adults** : Number of adults.
- **children** : number of children.
- **babies** : Number of babies.
- **meal** : Type of meal booked.:
- **country** : Country of origin of customer.
- **market\_segment** : where the bookings came from.
- **distribution\_channel** : Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” .

- **is\_repeated\_guest** : Value indicating if the booking name was from 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 that were cancelled by the customer prior to the current booking.
- **reserved\_room\_type** : Code of room type reserved. Code is presented instead of designation for anonymity reasons.
- **assigned\_room\_type** : Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due.
- **booking\_changes** : Number of changes/amendments made to the booking from the moment the booking was entered on the PMS.
- **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 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 as defined by dividing the sum of all lodging transactions by the total number of staying nights.
- **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, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out: customer check out from hotel, No show: Customer did not check-in hotel and informed hotel with reason.
- **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 cancelled or when did the customer checked out of the hotel.

We had added two columns for our own convenient analysis.

- **total\_stay** : Addition of stay\_in\_week\_nights + stay\_in\_weekend\_nights.
- **total\_people** : Addition of adults + children + babies.

# Data loading and Exploration

## Data Loading:

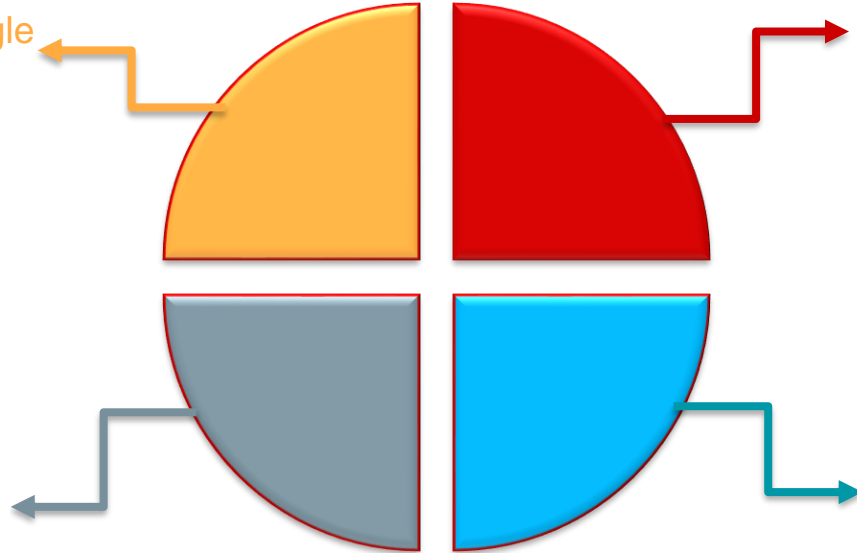
Loading data from google drive and reading in notebook

## Data Exploration:

Checking data in different columns

Numerical columns and categorical columns

Information and data types of columns



# Data Wrangling

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis.

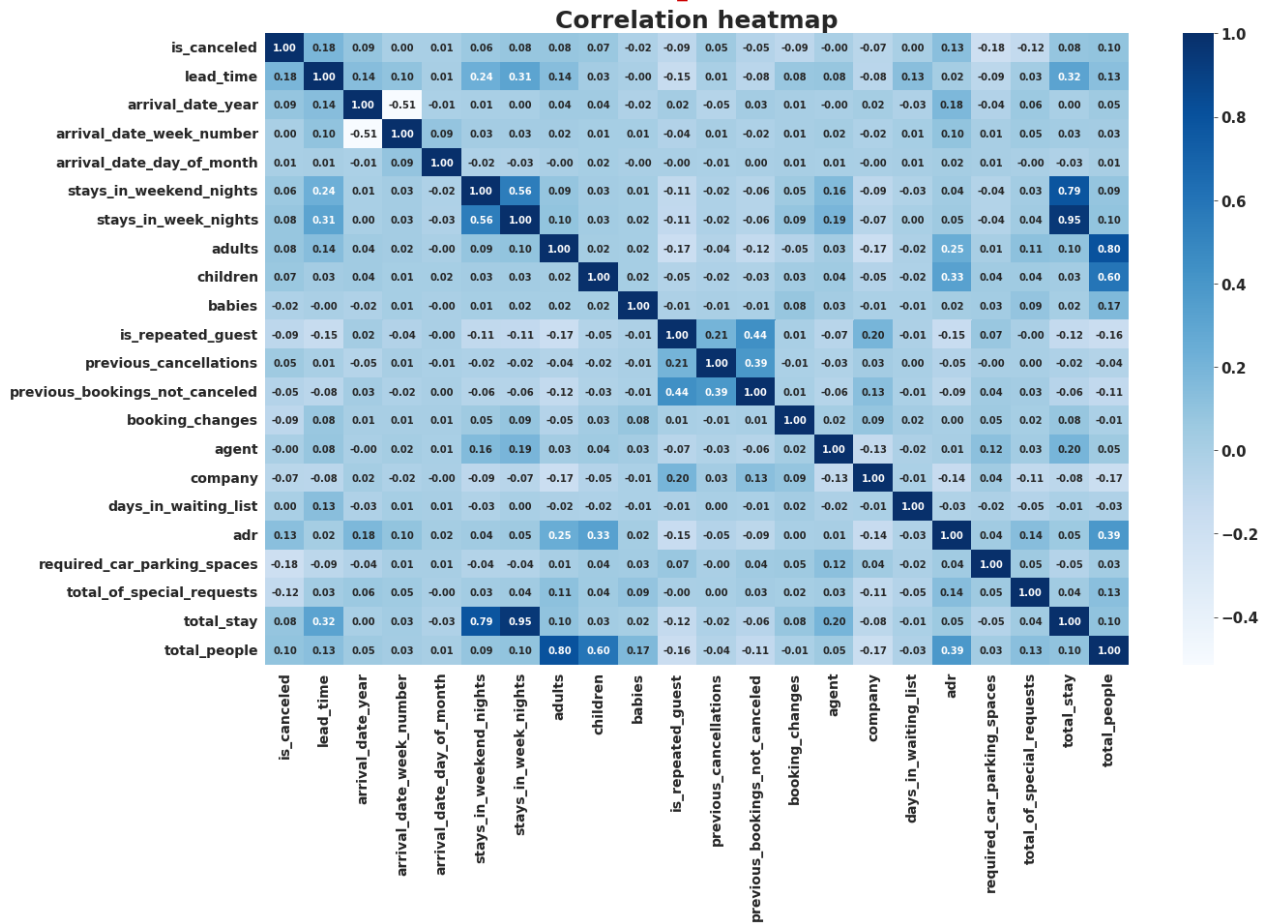
It includes following steps.

- Handling missing values.
- Removing duplicates data.
- Converting columns to proper dtype format.
- Adding or removing columns for analysis.



# Correlation heatmap

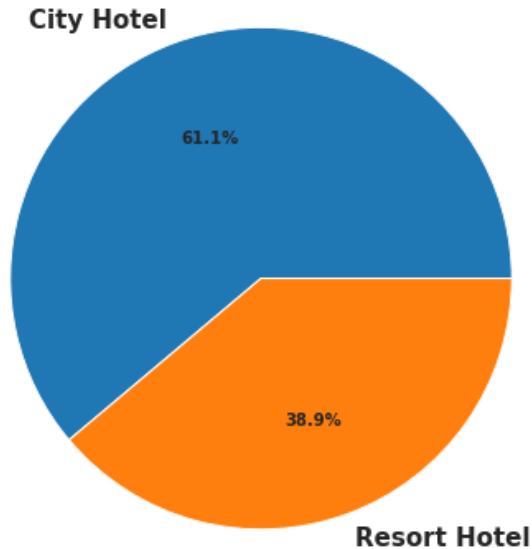
- Total\_stay and lead time have slight correlation it. This might means that customer plan reservation before their actual arrival.
- ADR (Average Daily Rate) is slightly correlated with total people, which makes sense as more number of people means more revenue.
- Previous booking not cancelled and repeated guest have high correlation as repeated guest are most likely to be who has not cancelled their previous bookings.
- We can also see that some columns have high correlation between them for example total people with adult and children it is because as those column are derived from them by addition.



# Hotel wise analysis

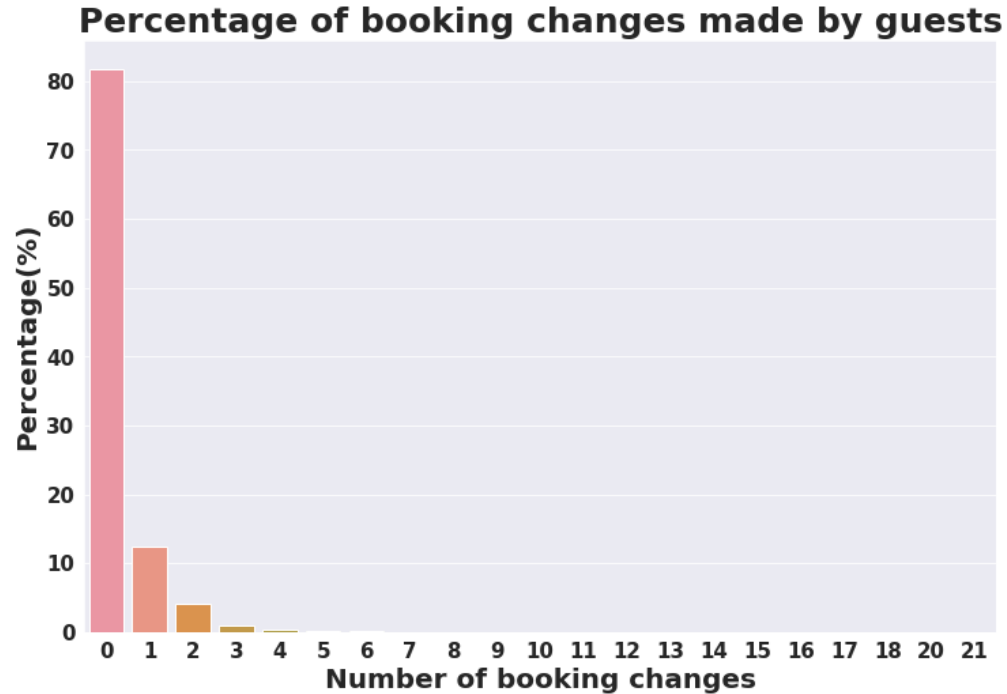
- Which type of hotels are most preferred by customer?

## Percentage of most preferred hotels



- Based on data, City Hotels are more preferred than Resort hotels

- What is the distribution of booking changes made by guests?

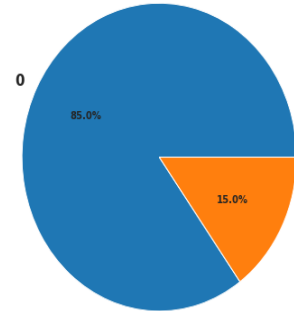
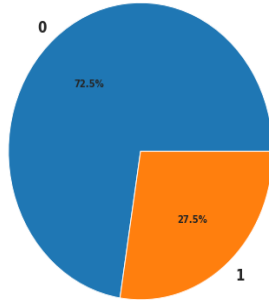


- More than 82% of bookings were not changed.
- Around 12% of guest changed 1 booking and less than 5% has made 2.

# Hotel Bookings Cancellation

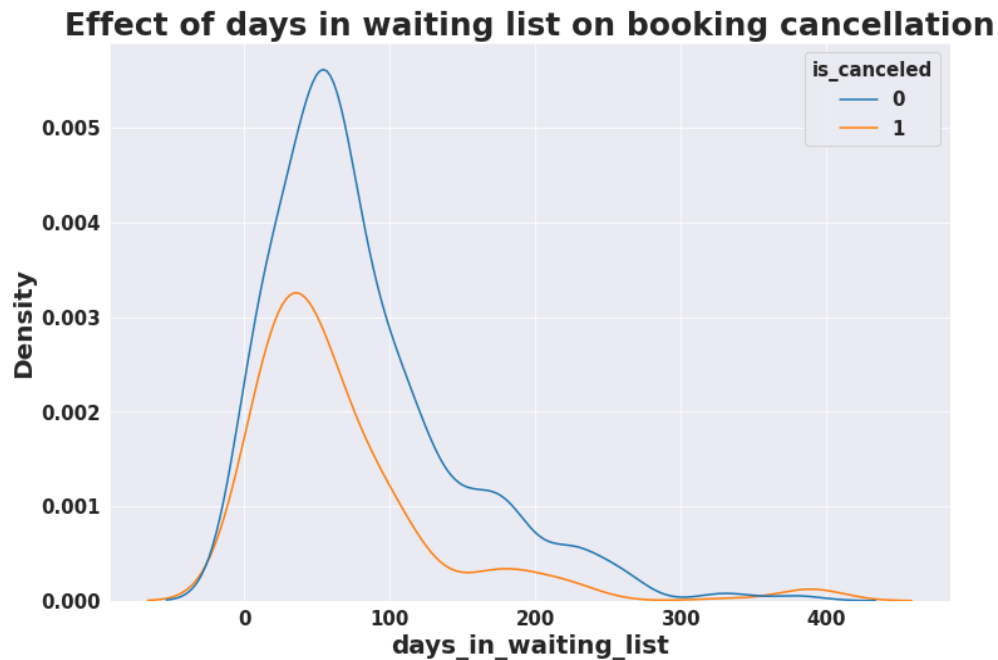
- What is the percentage of booking cancellation?
- Does not getting same room type as demanded causes cancellation?

Percentage of Hotel booking Cancellation vs Non cancellation    Percentage of bookings cancellation if same room as demanded not allotted



- 0 means not cancelled and 1 is cancelled.
- 27.5 % bookings were cancelled.
- Only 15 % of bookings were cancelled if not allotted same room type as demanded.

- Does longer waiting period causes booking cancellation?



- We can see that most bookings were cancelled which has less than 100 days in waiting list but also bookings were not cancelled for the same still density of cancelled is slightly higher for the same but there is no direct effect of longer waiting period with booking cancellation.

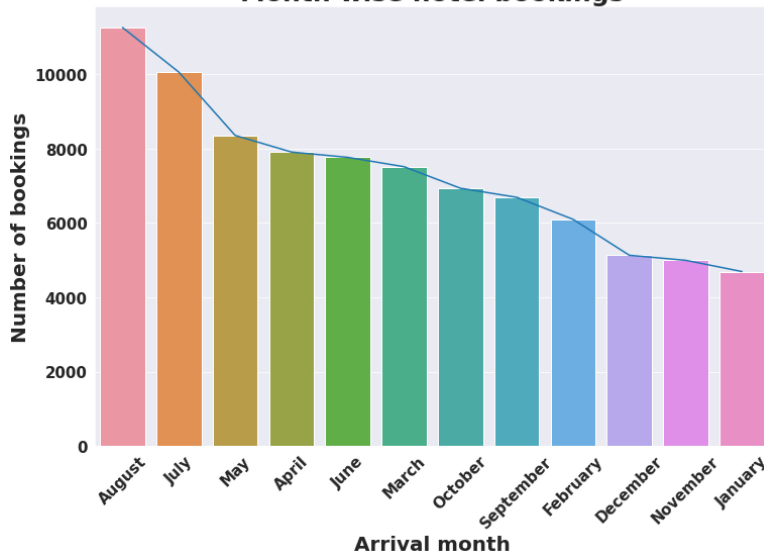
# Timewise analysis

- Which year has the highest bookings?
- Which month has highest number of bookings overall?

Number of bookings in different year

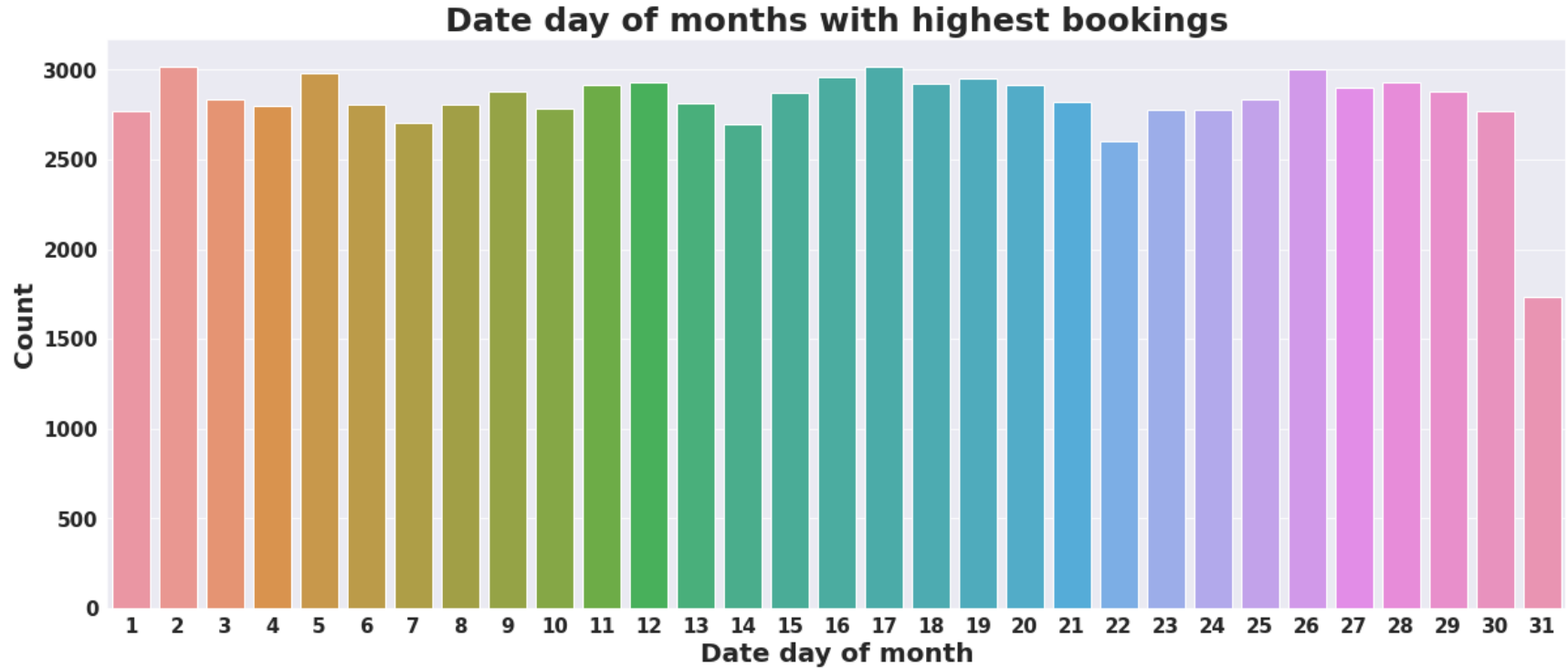


Month wise hotel bookings



- year 2016 has the highest number of bookings (42391) followed by 2017 (31692) and 2015 (13313). It means 2016 and 2017 was a very good year for hotel industry and In 2016 and 2017 more booking was in city hotels and in 2015 it is slightly lower than resort hotel.
- It seems that august has the highest number of bookings followed by July and may and least bookings are in November and January.

- which date day of month has the highest number of booking overall?

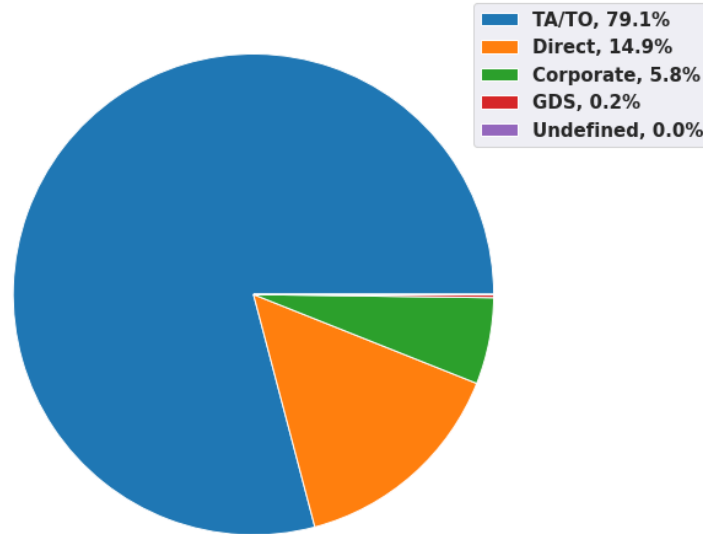


- We can see that the bookings are equally distributed amongst date day of month and only less numbers of bookings are on date 31.

# Distribution channel wise analysis

- Which distribution channel is mostly used for hotel bookings?

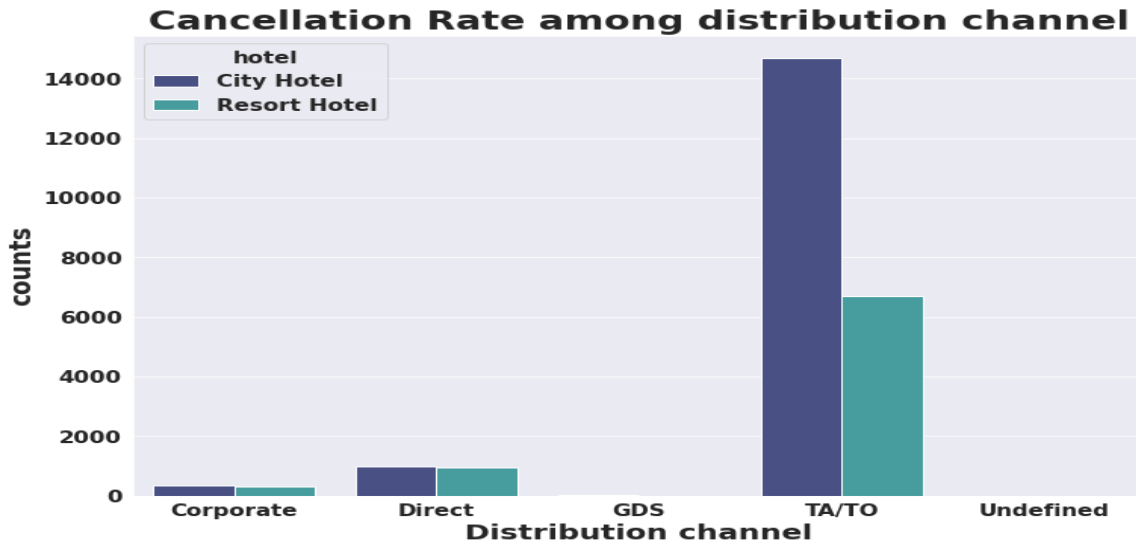
Distribution channel which is mostly used for hotel bookings



- TA/TO is mostly used for booking hotels (79.1%) followed by Direct Channel (14.9%).
- TA/TO: Travel Agent / Tour Operator.



- Which distribution channel has the highest cancellation rate?

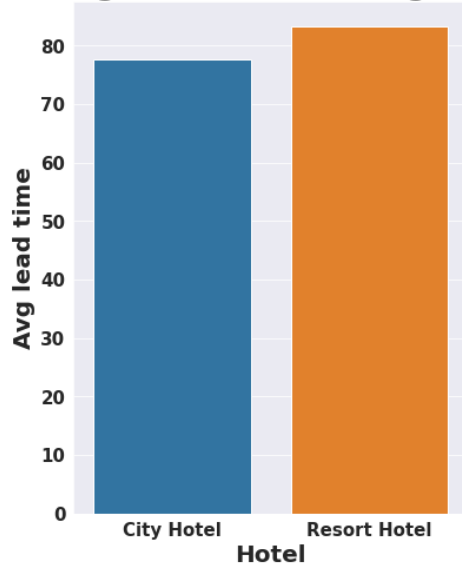


- It seems that TA/TO has the highest cancellation rate among distribution channels, and more cancellations are for city hotels.

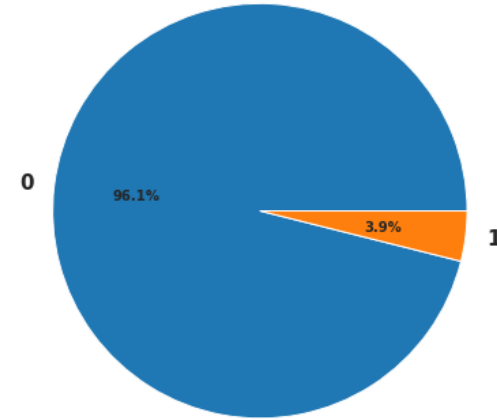
- Corporate**- These are corporate hotel booking companies which make bookings possible.
- GDS**- GDS is a worldwide conduit between travel bookers and suppliers, such as hotels and other accommodation providers. It communicates live product, price and availability data to travel agents and online booking 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.

- What is the percentage of repeated guest?
- Which hotel has the more lead time?

Average lead time among hotels

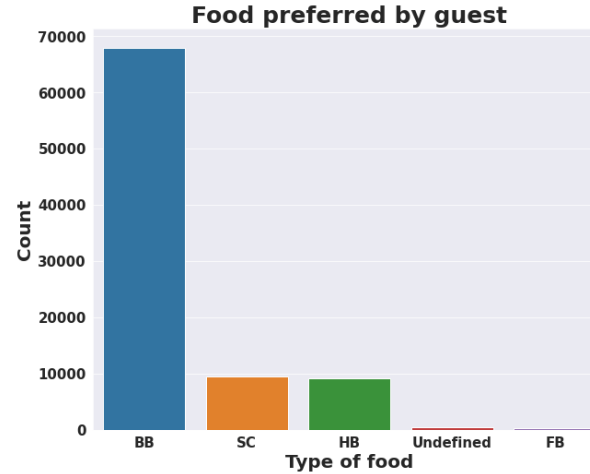
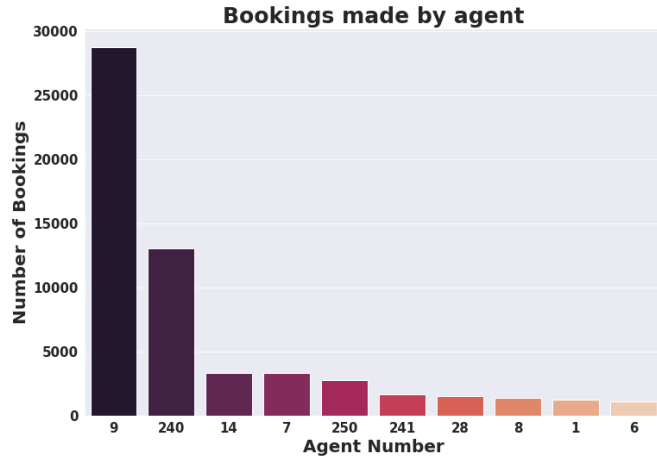


Percentage of repeated guest



- It seems that only 3.9% guests are repeated.
- Resort hotels has slightly high average lead time that means customer plans trip early than city hotels.

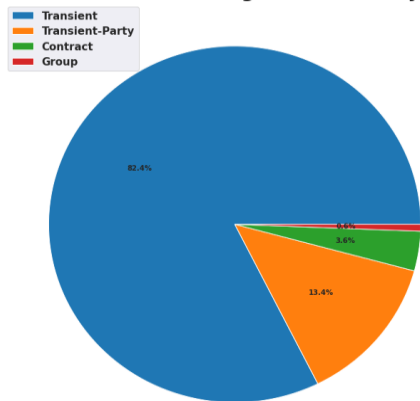
- Which agent has made most bookings?
- Which type of food is mostly preferred by the guests?



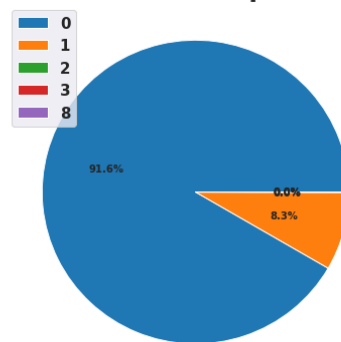
- Id Number 9 agent has made most numbers of bookings 28759.
- Id Number 240 has most number of bookings after ID 9 followed by 14 and 7.
- Most preferred meal type by the guest is BB (Bed and Breakfast).
- SC and HB are equally preferred, undefined and Fb are very less.

- What is the percentage distribution of customer type?
- What is the percentage distribution of car parking space?

Distribution Percentage of customer type

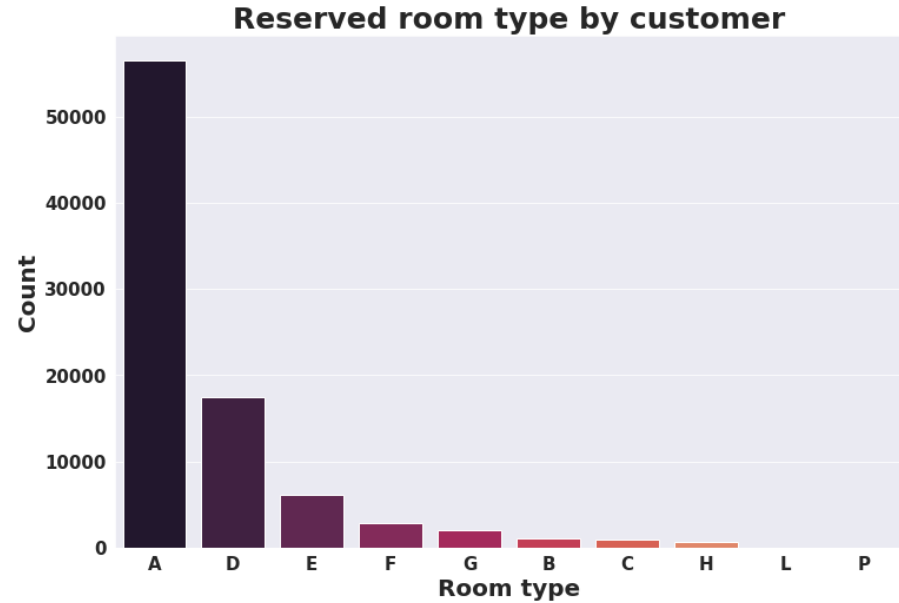
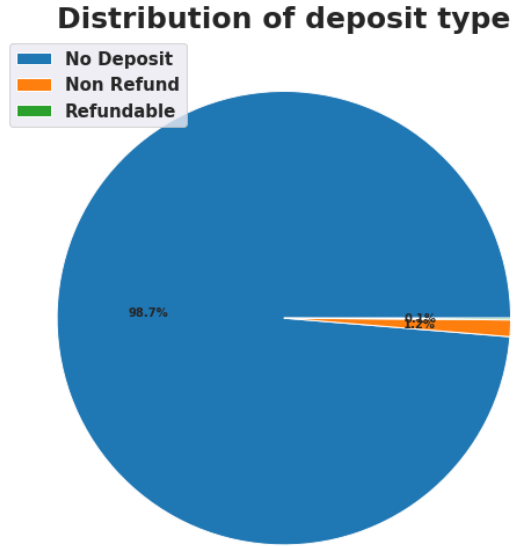


Percentage distribution of car parking space required



- 91.6 % guests did not required car parking space. Only 8.3 % guests required 1 car parking space and after that it is very negligible.
- Transient customer has the highest 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, amongst non-group category are transient.
- **Transient-party**: when the booking is transient, but it is associated to at least other transient booking.

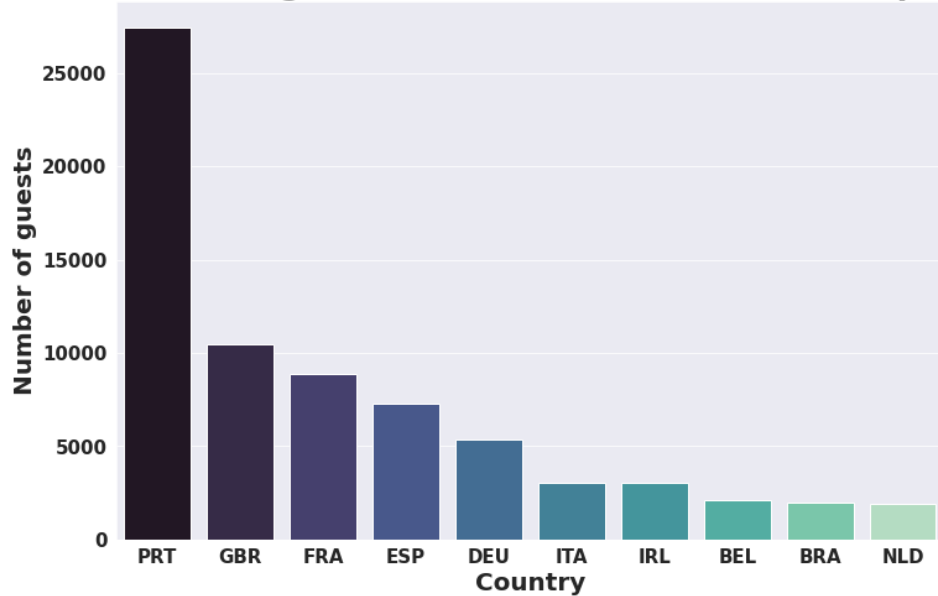
- What is the Percentage distribution of deposit type?
- Which is the most reserved room type by customer?



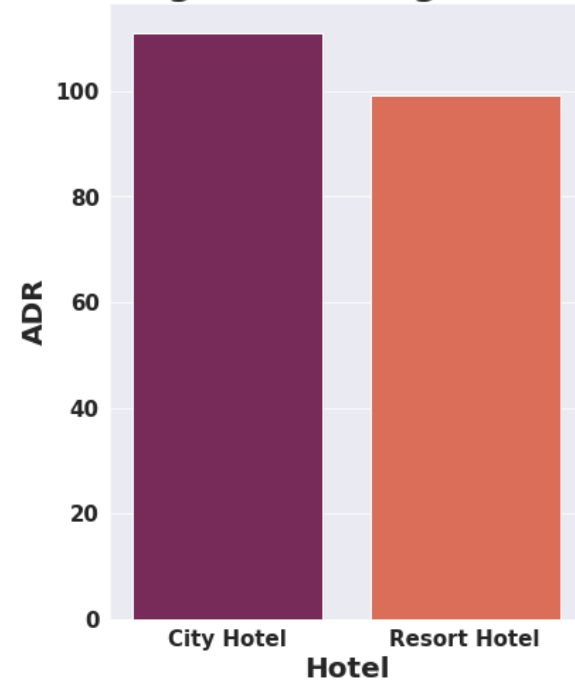
- Around 98.7% booking required no deposit and 1.2 % has no refund and only 0.1% has refundable.
- Most reserved room type is A and least preferred are L and P.

- From which country the most guests are coming?
- Which hotel type has the highest ADR?

Number of guests from different Countries (Top10)

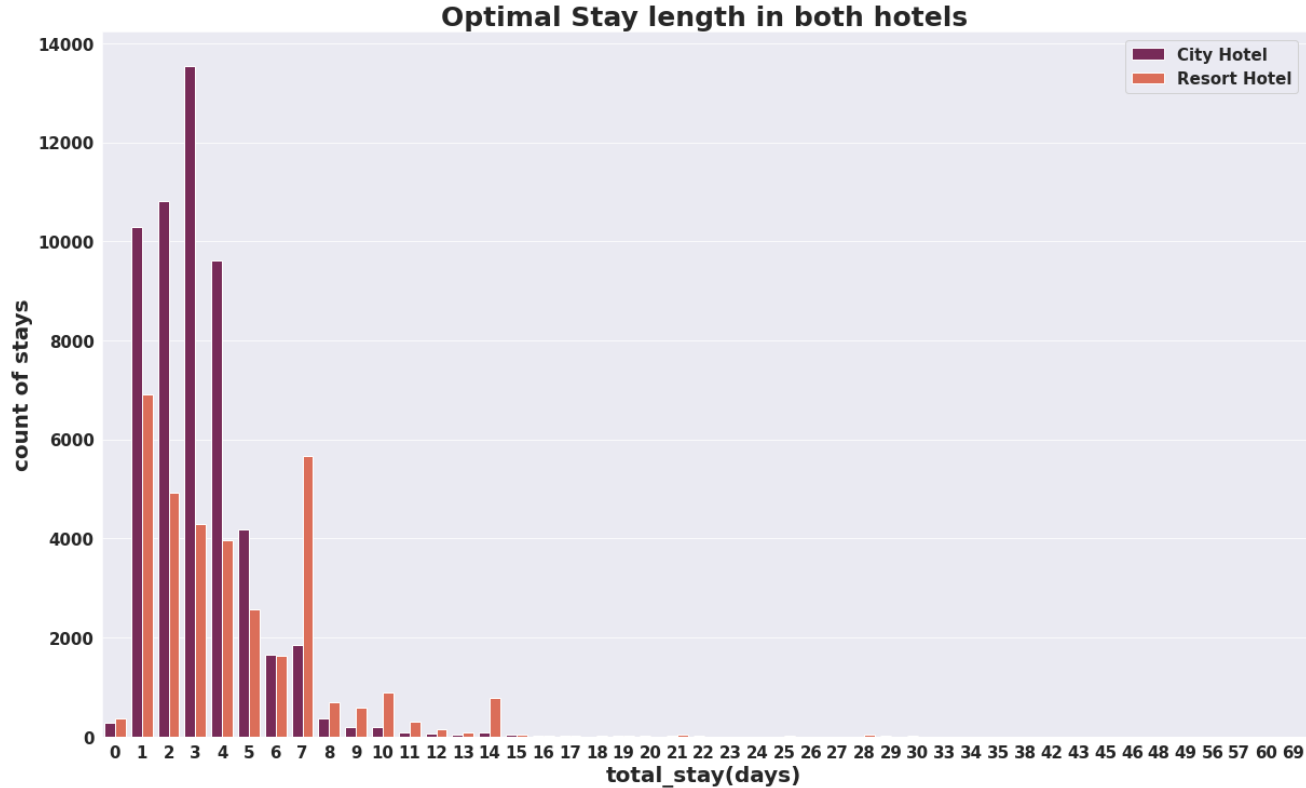


Avg ADR among hotels



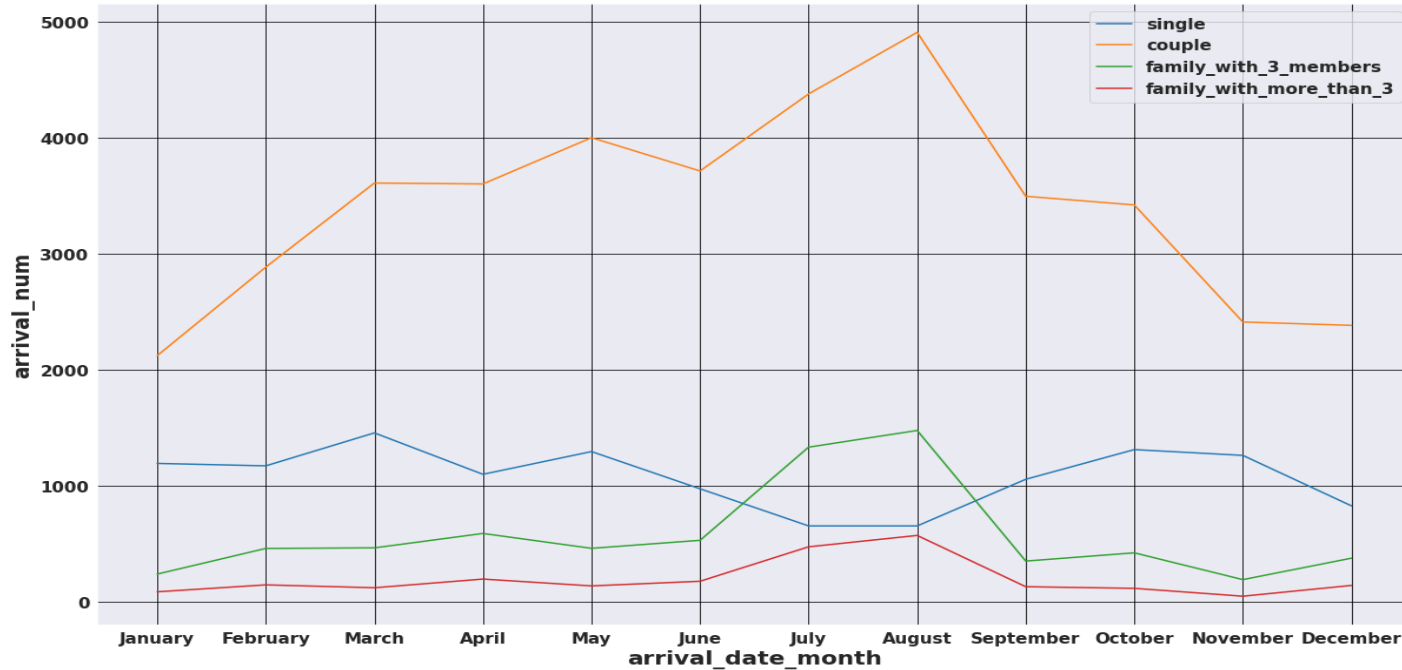
- Most number of guests are coming from European countries like Portugal, great Britain , France and Spain.
- City hotels has the slightly highest ADR that means city hotels are generating more revenue as compared to resort hotels.

- What is the Optimal stay length in both types of hotels ?



Optimal stay length in both hotels are less than 7 days, on an average customer preferred to stay 1 to 4 days.

- What are the booking percentage according to number of peoples?



- It seems that most number of bookings were made by couples but we are not sure that they are couple as data doesn't have any specific about couple or family.
- We can also see that most number of bookings were in august month.
- Bookings with family having 4 member or above is least.



# Challenges

- Handling null values and finding a way to replace them with something meaningful so that it doesn't affect analysis.
- Choosing visualization for different analysis.

# Conclusion

- City hotels are mostly preferred by guests they occupy 61.1 % of market share.
- Peak Months for hotel bookings were June, July and August. People preferred to spend more time in hotel during summer vacation and most numbers of bookings were in 2016.
- More than 82 % customer made no booking changes and 91.6 guests required no car parking space.
- Cancellation rate are high when bookings done through online TA/TO compared to direct bookings.
- The number of repeated guests is too low (3.9%), thus retention rate is low.
- BB (Bed and Breakfast) is the most preferred meal type by the guest (more than 70%) .
- City hotels has highest ADR (Average Daily Rate), thus city hotels are generating more revenue and resort hotels have slightly higher lead time.
- The majority of guests come from western Europe countries.
- Optimal stay length for both the hotels are for a week or less than week.

**Thank you**