Capstone Project

Hotel Booking data analysis and Visualization

By Suvendu Nayak
Cohort geneva

Table of Contents:

- Objective
- Data Summary
- Data loading and exploration
- Data Wrangling
- Correlation heatmap
- Hotel wise analysis
- Timewise analysis
- Miscellaneous questions
- Hotel booking cancellation analysis
- Challenges
- Conclusion

Objective

I am going to analyse hotel bookings dataset for 3 years from 2015 - 2017.
I shall 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. Each column is defined 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” .

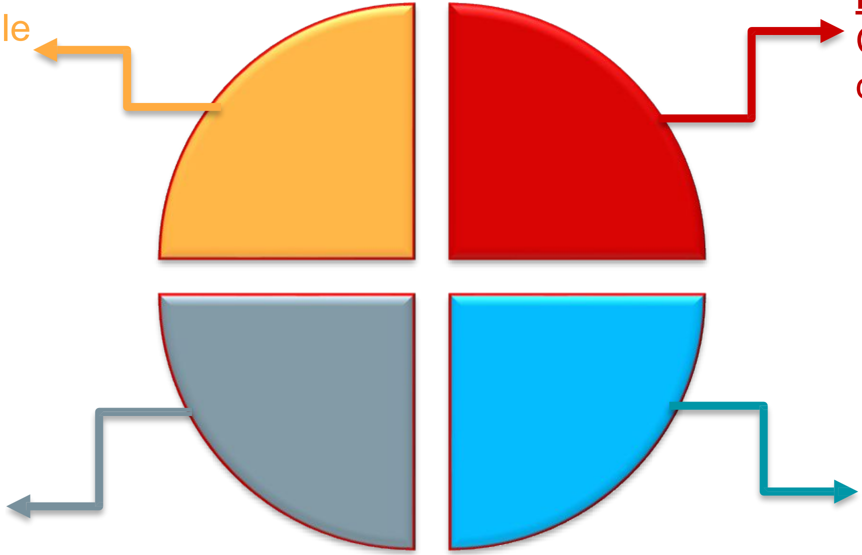
- **is_repeated_guest** : Value indicating if the booking name was from a repeated guest (1) or not (0).
- **previous_cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking.
- **previous_bookings_not_canceled** : Number of previous bookings that were cancelled by the customer prior to the current booking.
- **reserved_room_type** : Code of room type reserved. Code is presented instead of designation .
- **assigned_room_type** : Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type.
- **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

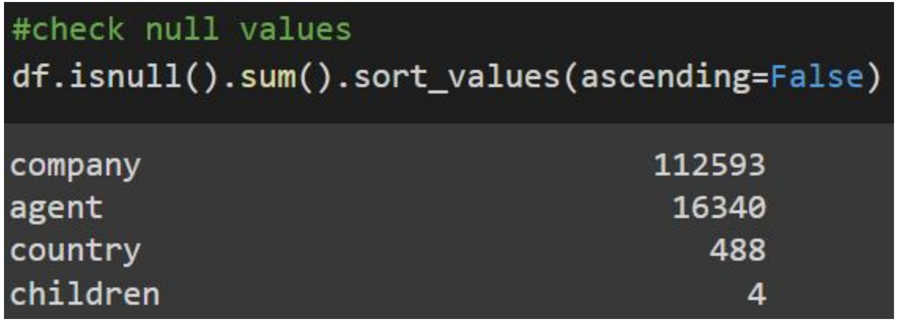
Information and data types of columns

Data Wrangling

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

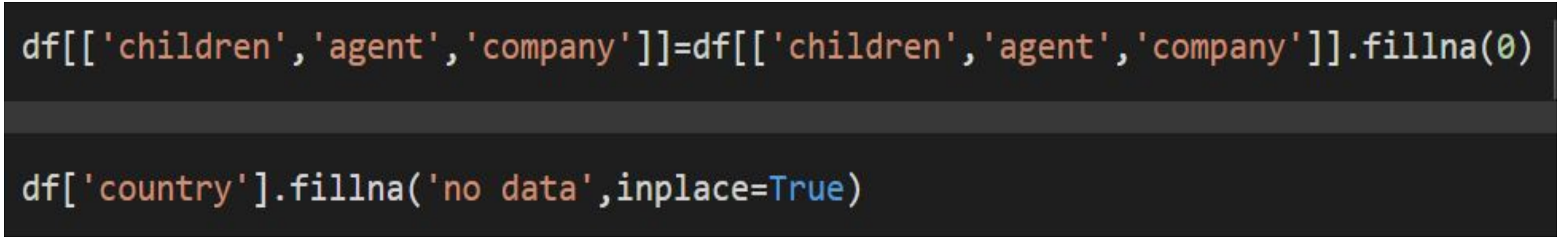
It includes following steps.

- Handling missing values.



```
#check null values
df.isnull().sum().sort_values(ascending=False)
```

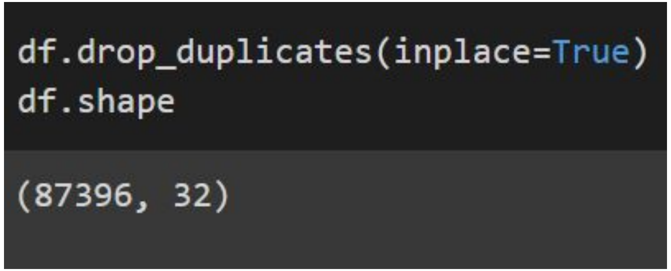
```
company          112593
agent             16340
country             488
children              4
```



```
df[['children','agent','company']]=df[['children','agent','company']].fillna(0)
```

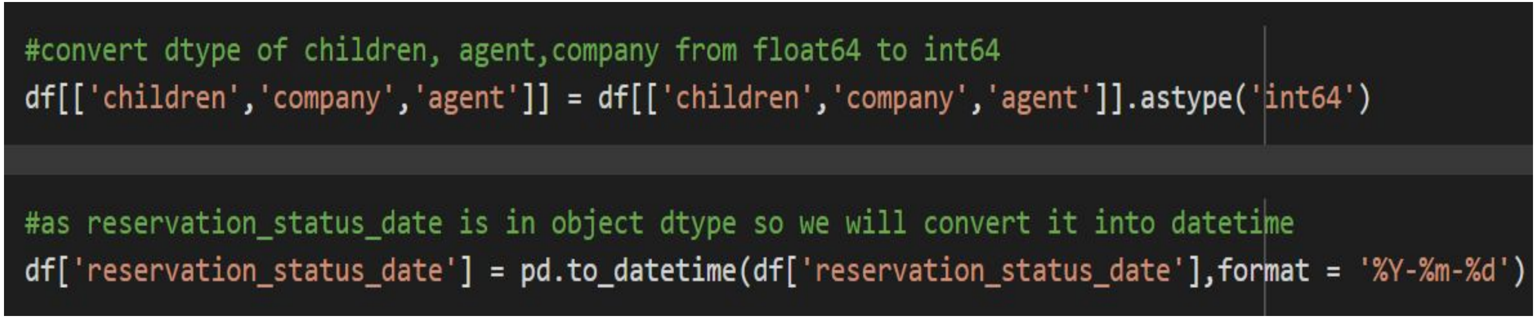
```
df['country'].fillna('no data',inplace=True)
```

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



```
df.drop_duplicates(inplace=True)
df.shape
```

```
(87396, 32)
```

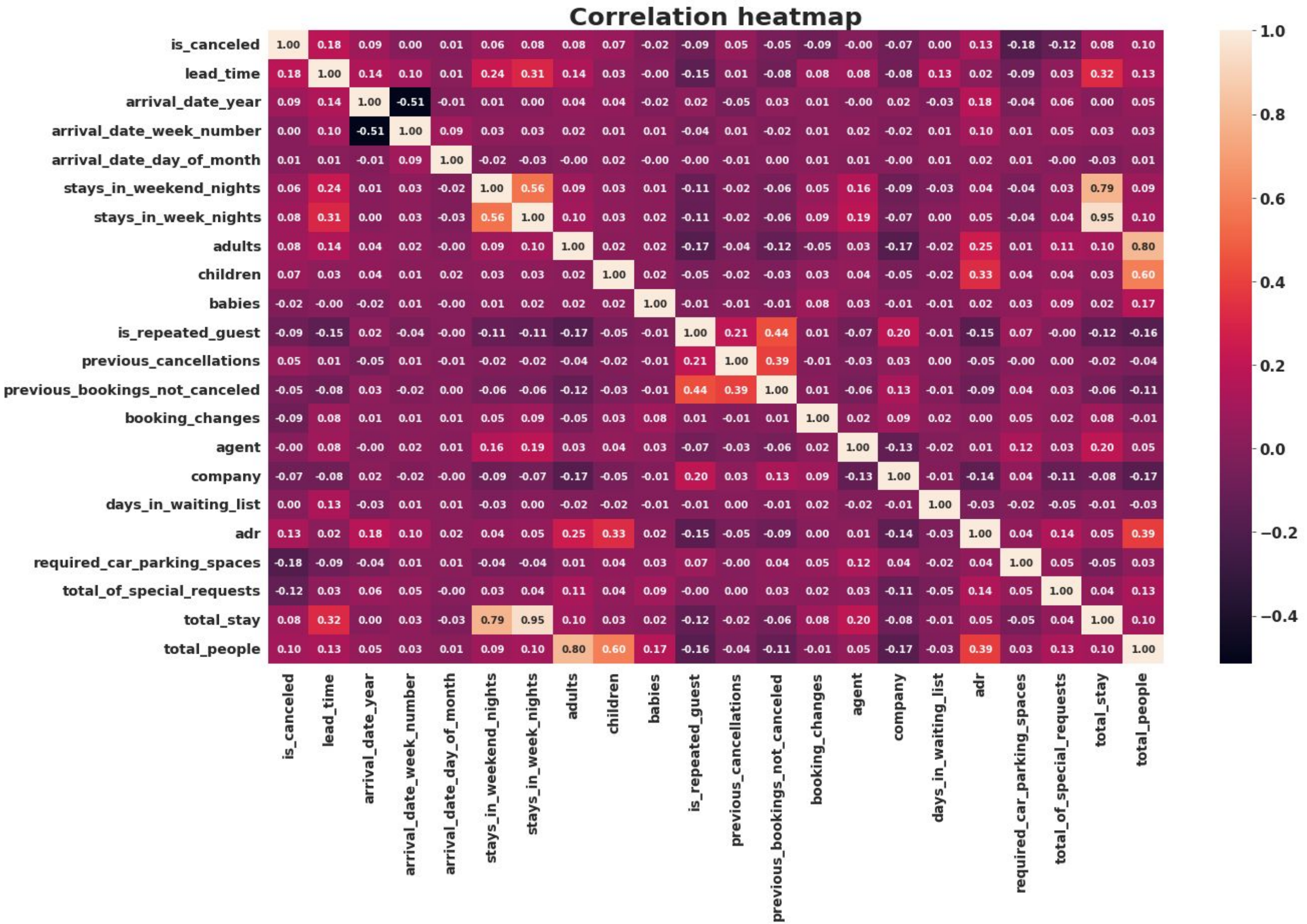


```
#convert dtype of children, agent,company from float64 to int64
df[['children','company','agent']] = df[['children','company','agent']].astype('int64')
```

```
#as reservation_status_date is in object dtype so we will convert it into datetime
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'],format = '%Y-%m-%d')
```

Correlation heatmap

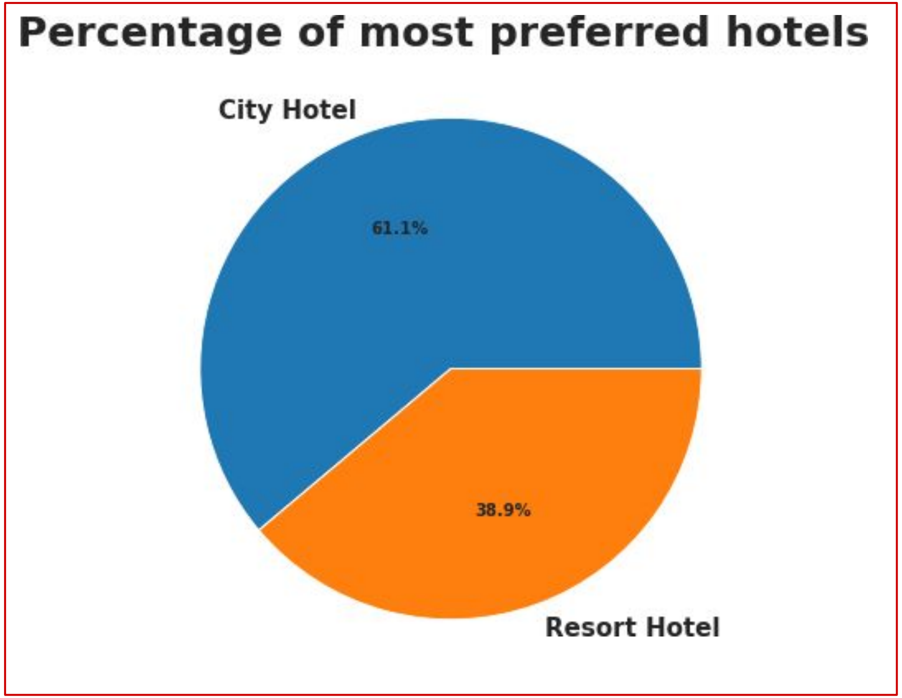
- Total_stay and lead time have slight correlation it. This might means that customer plan reservation before their actual arrival.
- Adr(Average Daily Rate) is slightly correlated with total people, which makes sense as more number of people means more revenue.
- Previous booking not cancelled and repeated 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.



	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	total_stay	total_people
is_canceled	1.00	0.18	0.09	0.00	0.01	0.06	0.08	0.08	0.07	-0.02	-0.09	0.05	-0.05	-0.09	-0.00	-0.07	0.00	0.13	-0.18	-0.12	0.08	0.10
lead_time	0.18	1.00	0.14	0.10	0.01	0.24	0.31	0.14	0.03	-0.00	-0.15	0.01	-0.08	0.08	0.08	-0.08	0.13	0.02	-0.09	0.03	0.32	0.13
arrival_date_year	0.09	0.14	1.00	-0.51	-0.01	0.01	0.00	0.04	0.04	-0.02	0.02	-0.05	0.03	0.01	-0.00	0.02	-0.03	0.18	-0.04	0.06	0.00	0.05
arrival_date_week_number	0.00	0.10	-0.51	1.00	0.09	0.03	0.03	0.02	0.01	0.01	-0.04	0.01	-0.02	0.01	0.02	-0.02	0.01	0.10	0.01	0.05	0.03	0.03
arrival_date_day_of_month	0.01	0.01	-0.01	0.09	1.00	-0.02	-0.03	-0.00	0.02	-0.00	-0.00	-0.01	0.00	0.01	0.01	-0.00	0.01	0.02	0.01	-0.00	-0.03	0.01
stays_in_weekend_nights	0.06	0.24	0.01	0.03	-0.02	1.00	0.56	0.09	0.03	0.01	-0.11	-0.02	-0.06	0.05	0.16	-0.09	-0.03	0.04	-0.04	0.03	0.79	0.09
stays_in_week_nights	0.08	0.31	0.00	0.03	-0.03	0.56	1.00	0.10	0.03	0.02	-0.11	-0.02	-0.06	0.09	0.19	-0.07	0.00	0.05	-0.04	0.04	0.95	0.10
adults	0.08	0.14	0.04	0.02	-0.00	0.09	0.10	1.00	0.02	0.02	-0.17	-0.04	-0.12	-0.05	0.03	-0.17	-0.02	0.25	0.01	0.11	0.10	0.80
children	0.07	0.03	0.04	0.01	0.02	0.03	0.03	0.02	1.00	0.02	-0.05	-0.02	-0.03	0.03	0.04	-0.05	-0.02	0.33	0.04	0.04	0.03	0.60
babies	-0.02	-0.00	-0.02	0.01	-0.00	0.01	0.02	0.02	0.02	1.00	-0.01	-0.01	-0.01	0.08	0.03	-0.01	-0.01	0.02	0.03	0.09	0.02	0.17
is_repeated_guest	-0.09	-0.15	0.02	-0.04	-0.00	-0.11	-0.11	-0.17	-0.05	-0.01	1.00	0.21	0.44	0.01	-0.07	0.20	-0.01	-0.15	0.07	-0.00	-0.12	-0.16
previous_cancellations	0.05	0.01	-0.05	0.01	-0.01	-0.02	-0.02	-0.04	-0.02	-0.01	0.21	1.00	0.39	-0.01	-0.03	0.03	0.00	-0.05	-0.00	0.00	-0.02	-0.04
previous_bookings_not_canceled	-0.05	-0.08	0.03	-0.02	0.00	-0.06	-0.06	-0.12	-0.03	-0.01	0.44	0.39	1.00	0.01	-0.06	0.13	-0.01	-0.09	0.04	0.03	-0.06	-0.11
booking_changes	-0.09	0.08	0.01	0.01	0.01	0.05	0.09	-0.05	0.03	0.08	0.01	-0.01	0.01	1.00	0.02	0.09	0.02	0.00	0.05	0.02	0.08	-0.01
agent	-0.00	0.08	-0.00	0.02	0.01	0.16	0.19	0.03	0.04	0.03	-0.07	-0.03	-0.06	0.02	1.00	-0.13	-0.02	0.01	0.12	0.03	0.20	0.05
company	-0.07	-0.08	0.02	-0.02	-0.00	-0.09	-0.07	-0.17	-0.05	-0.01	0.20	0.03	0.13	0.09	-0.13	1.00	-0.01	-0.14	0.04	-0.11	-0.08	-0.17
days_in_waiting_list	0.00	0.13	-0.03	0.01	0.01	-0.03	0.00	-0.02	-0.02	-0.01	-0.01	0.00	-0.01	0.02	-0.02	-0.01	1.00	-0.03	-0.02	-0.05	-0.01	-0.03
adr	0.13	0.02	0.18	0.10	0.02	0.04	0.05	0.25	0.33	0.02	-0.15	-0.05	-0.09	0.00	0.01	-0.14	-0.03	1.00	0.04	0.14	0.05	0.39
required_car_parking_spaces	-0.18	-0.09	-0.04	0.01	0.01	-0.04	-0.04	0.01	0.04	0.03	0.07	-0.00	0.04	0.05	0.12	0.04	-0.02	0.04	1.00	0.05	-0.05	0.03
total_of_special_requests	-0.12	0.03	0.06	0.05	-0.00	0.03	0.04	0.11	0.04	0.09	-0.00	0.00	0.03	0.02	0.03	-0.11	-0.05	0.14	0.05	1.00	0.04	0.13
total_stay	0.08	0.32	0.00	0.03	-0.03	0.79	0.95	0.10	0.03	0.02	-0.12	-0.02	-0.06	0.08	0.20	-0.08	-0.01	0.05	-0.05	0.04	1.00	0.10
total_people	0.10	0.13	0.05	0.03	0.01	0.09	0.10	0.80	0.60	0.17	-0.16	-0.04	-0.11	-0.01	0.05	-0.17	-0.03	0.39	0.03	0.13	0.10	1.00

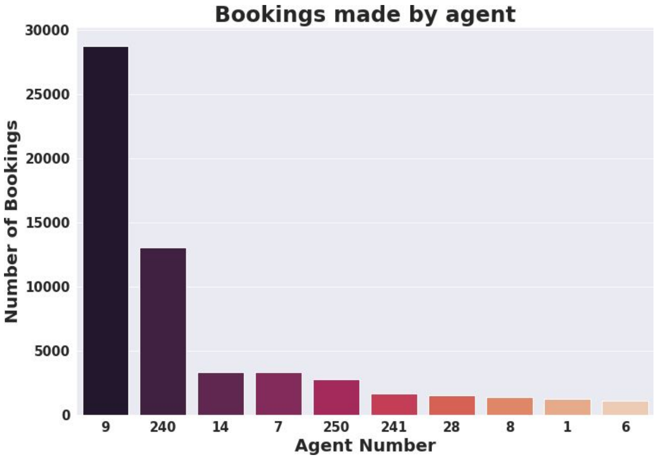
Hotel wise analysis

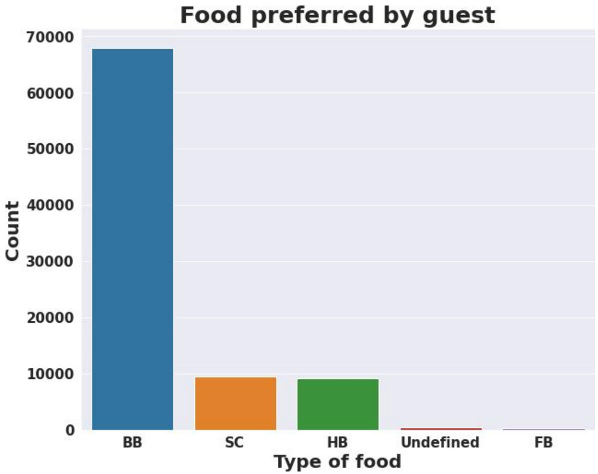
- **Which type of hotels are most preferred by customer?**



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

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

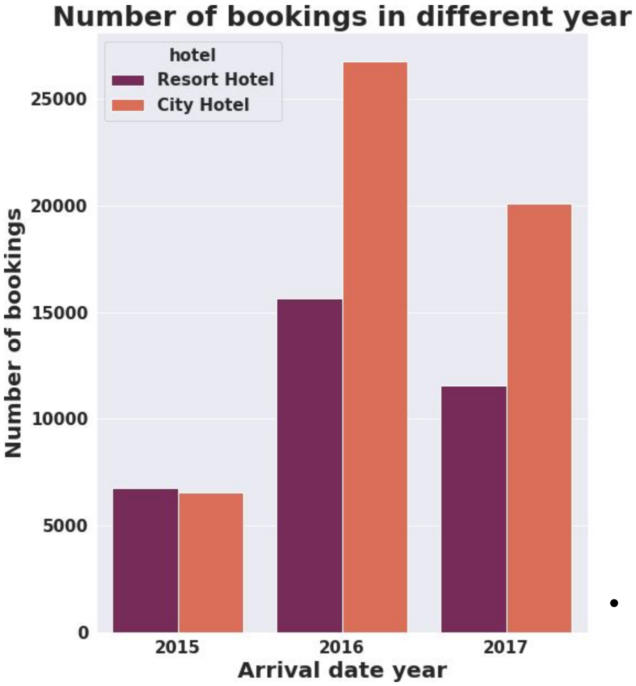


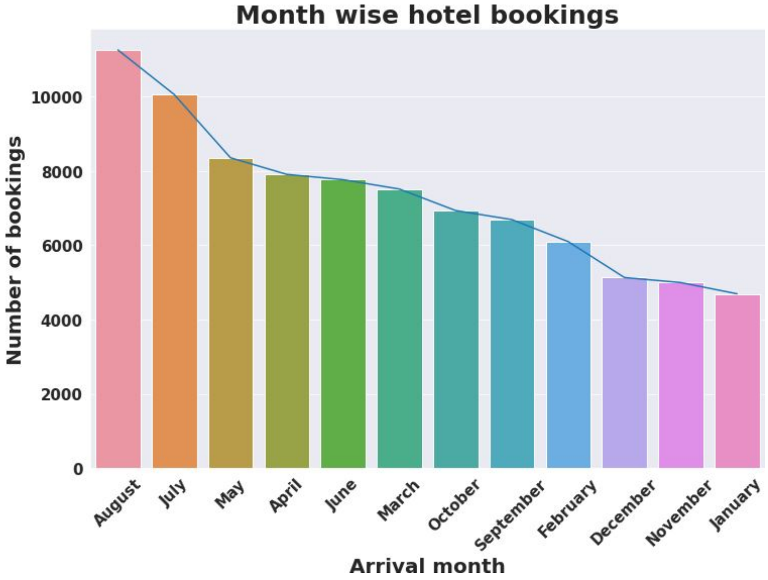


- Id Number 9 agent has made most numbers of bookings 28759.
- Id Number 240 has most number of bookings after ID 9 followed by 14 and 7.
- Most preferred meal type by the guest is BB (Bed and Breakfast).
- SC and HB are equally preferred, undefined and Fb are very less.
- BB - (Bed and Breakfast)
-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

Timewise analysis

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

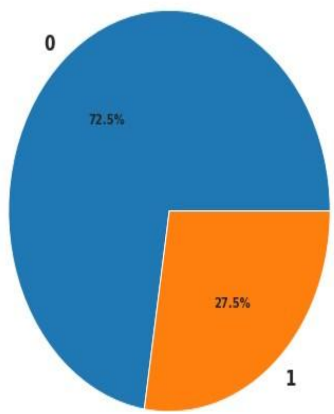




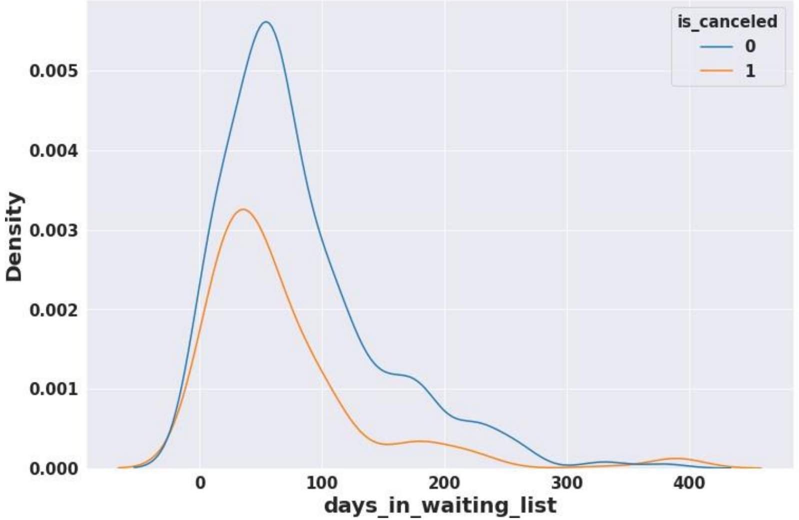
- year 2016 has the highest number of bookings (42391) followed by 2017 (31692) and 2015 (13313). It means 2016 and 2017 was a very good year for hotel industry and In 2016 and 2017 more booking was in city hotels and in 2015 it is slightly lower than resort hotel.
- It seems that august has the highest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

Hotel Bookings Cancellation

- **What is the percentage of booking cancellation?**
- **Does longer waiting period causes booking cancellation?**
- **Percentage of Hotel booking Cancellation vs Non concellation**

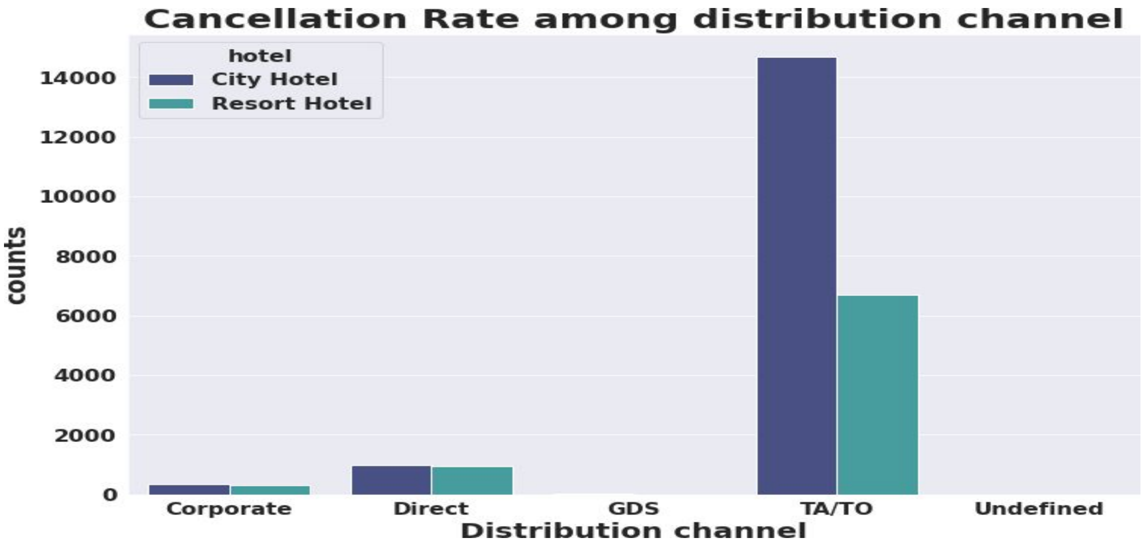


Effect of days in waiting list on booking cancellation



- 0 means not cancelled and 1 is cancelled.
- 27.5 % bookings were cancelled.
- 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.

- ## Which distribution channel has the highest cancellation rate?



- It seem that TA/TO has the highest cancellation rate among distribution channel and more cancellation are for city hotels.

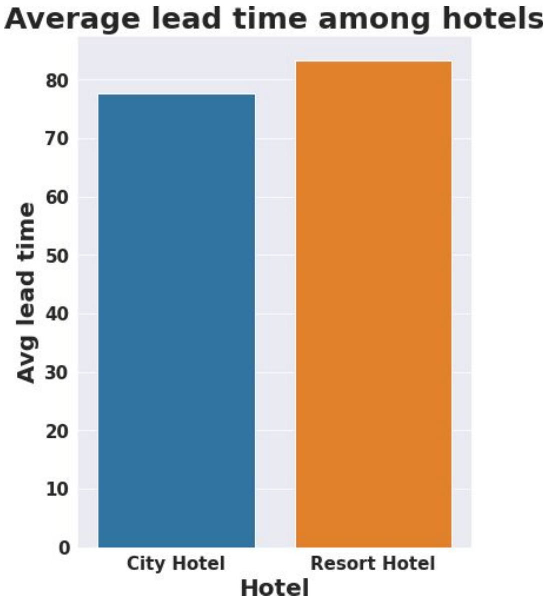
•**Corporate-** These are corporate hotel booking companies which makes bookings possible.

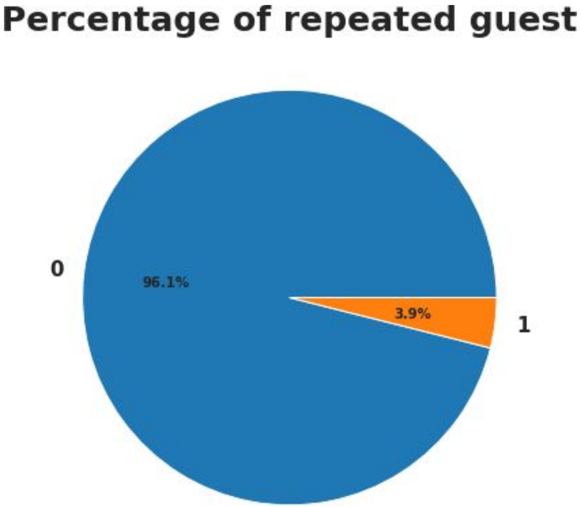
•**GDS-** GDS is a worldwide conduit between travel bookers and suppliers, such as hotels and other accommodation providers. It communicates live product, price and availability data to travel agents and online booking engines, and allows for automated transactions.

•**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.

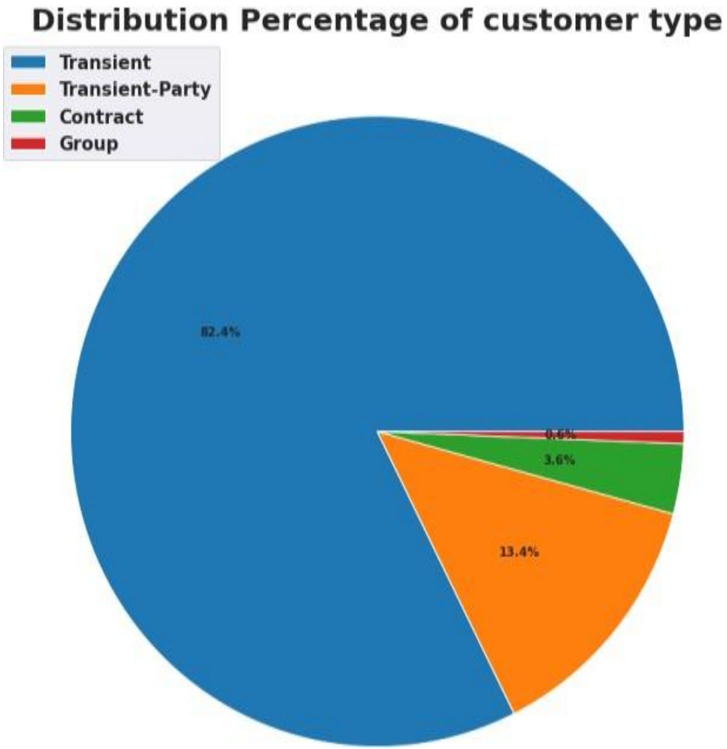
- **Which hotel has the more lead time?**
- **What is the percentage of repeated guest?**
-





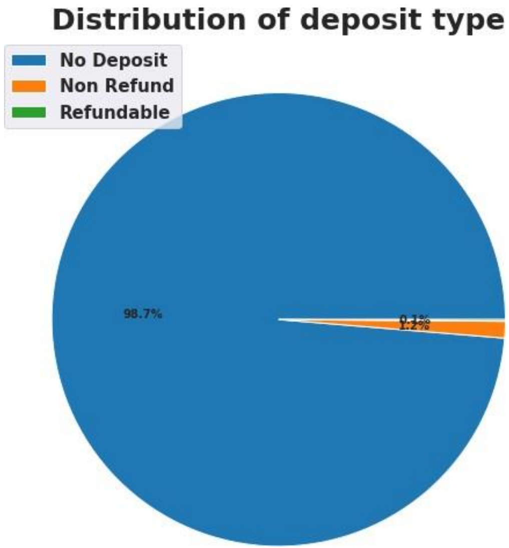
- It seems that only 3.9% guests are repeated.
- Resort hotels has slightly high average lead time that means customer plans trip early than city hotels.

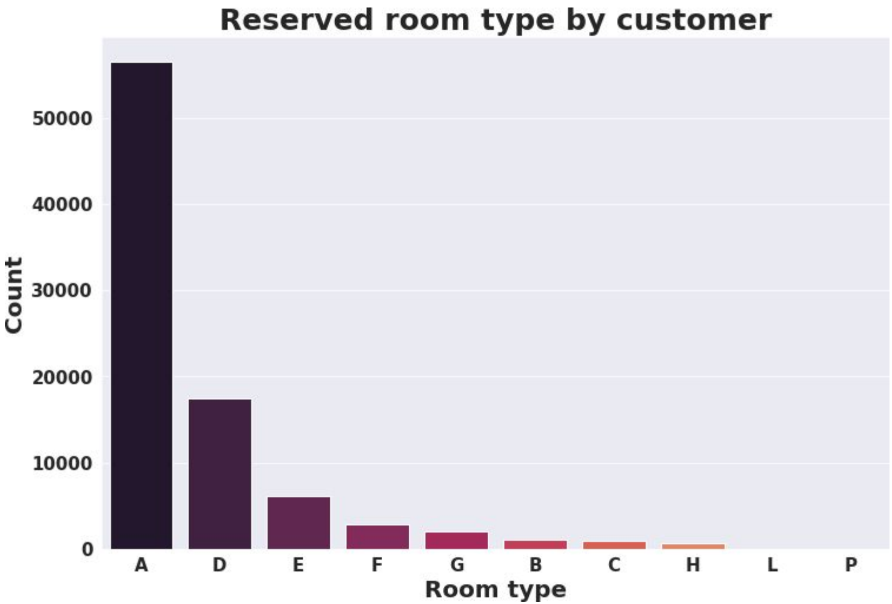
- **What is the percentage distribution of customer type?**



- 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.

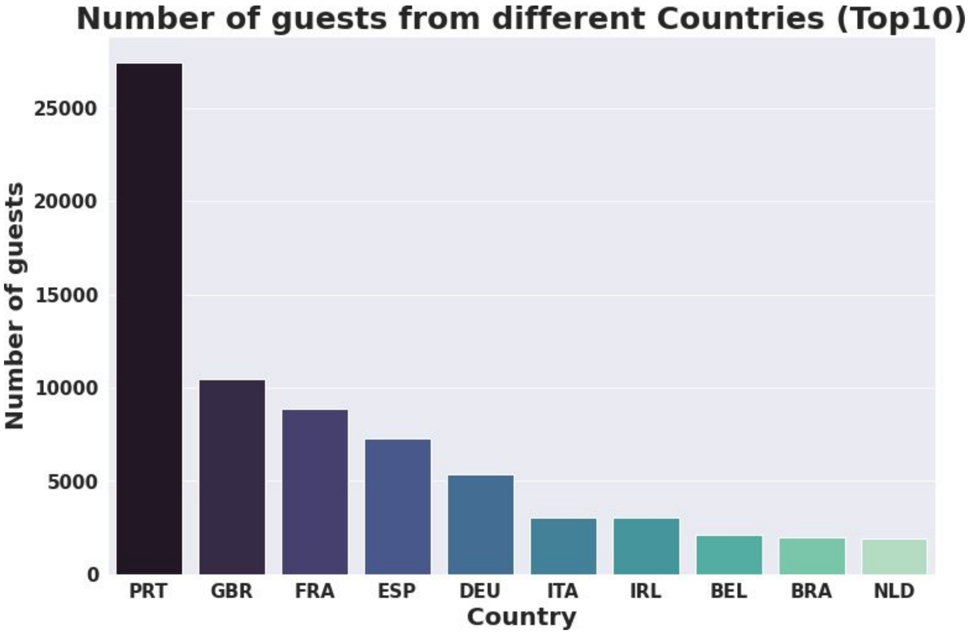
- **What is the Percentage distribution of deposit type?**
- **Which is the most reserved room type by customer?**

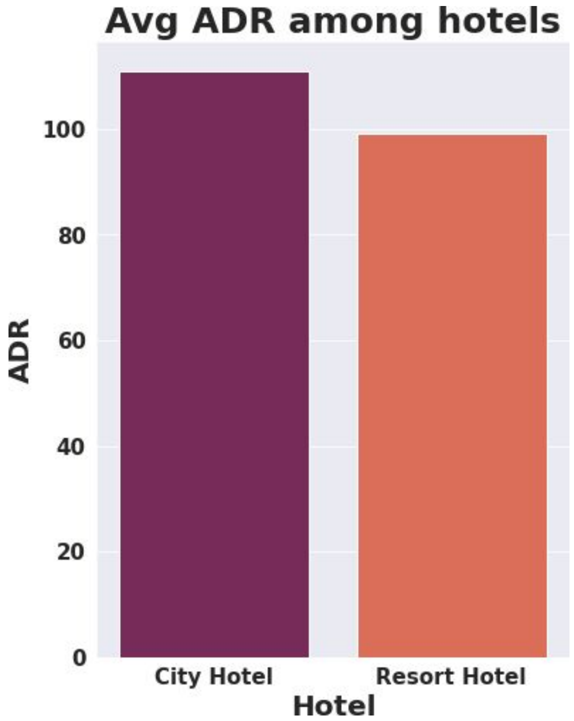

Distribution of deposit type


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?**





- 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.

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.
- 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.

Thank you