Capstone Project

Hotel Booking data analysis and Visualization

Kshitij Chaudhari

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” .

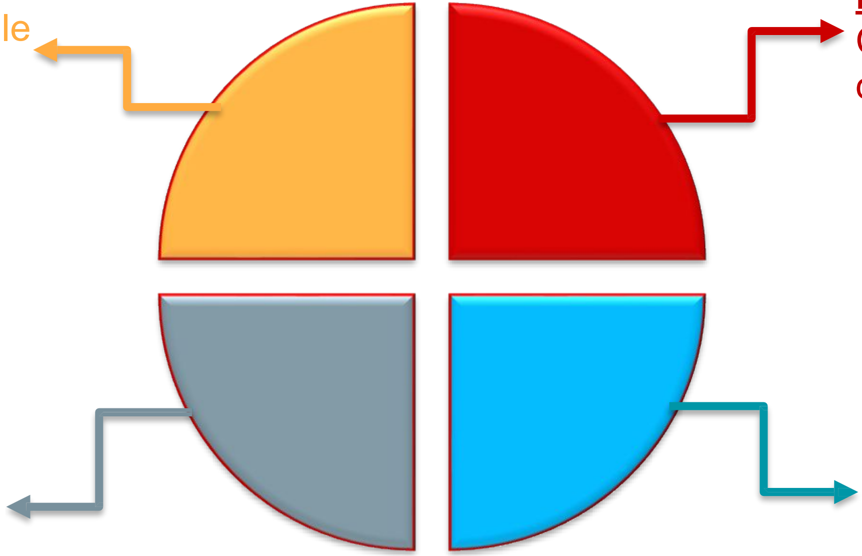
- **is_repeated_guest** : Value indicating if the booking name was from a repeated guest (1) or not (0).
- **previous_cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking.
- **previous_bookings_not_canceled** : Number of previous bookings that 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

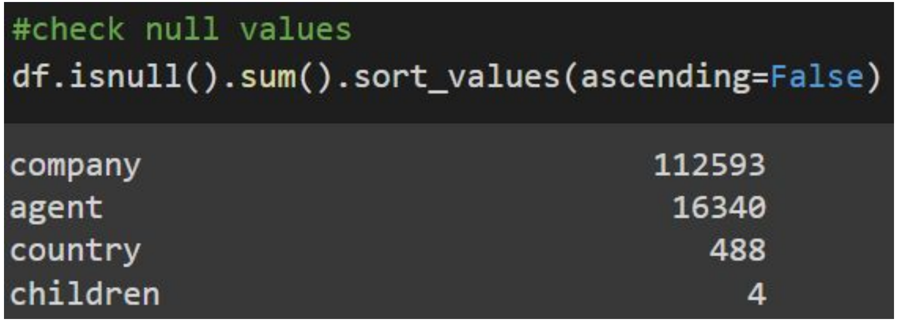
Information and data types of columns

Data Wrangling

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

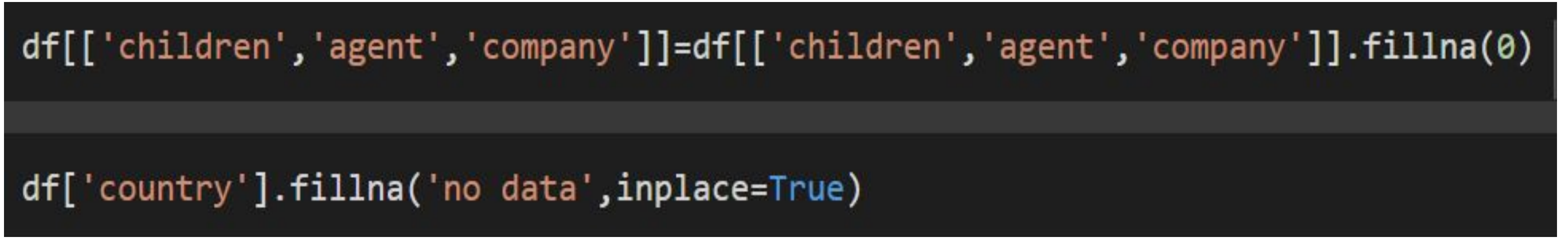
It includes following steps.

- Handling missing values.



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

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

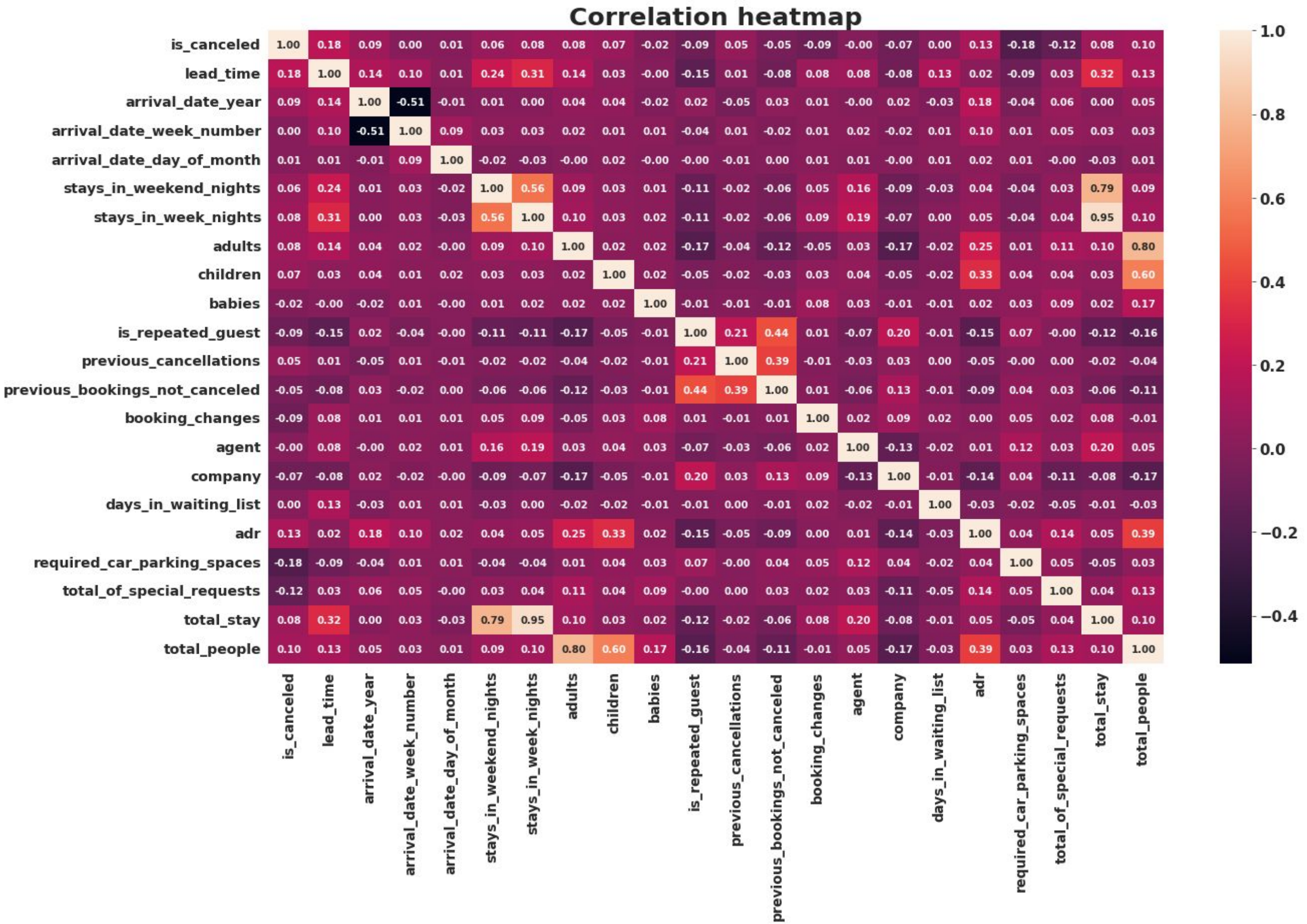


```
df[['children','agent','company']]=df[['children','agent','company']].fillna(0)
```

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

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.

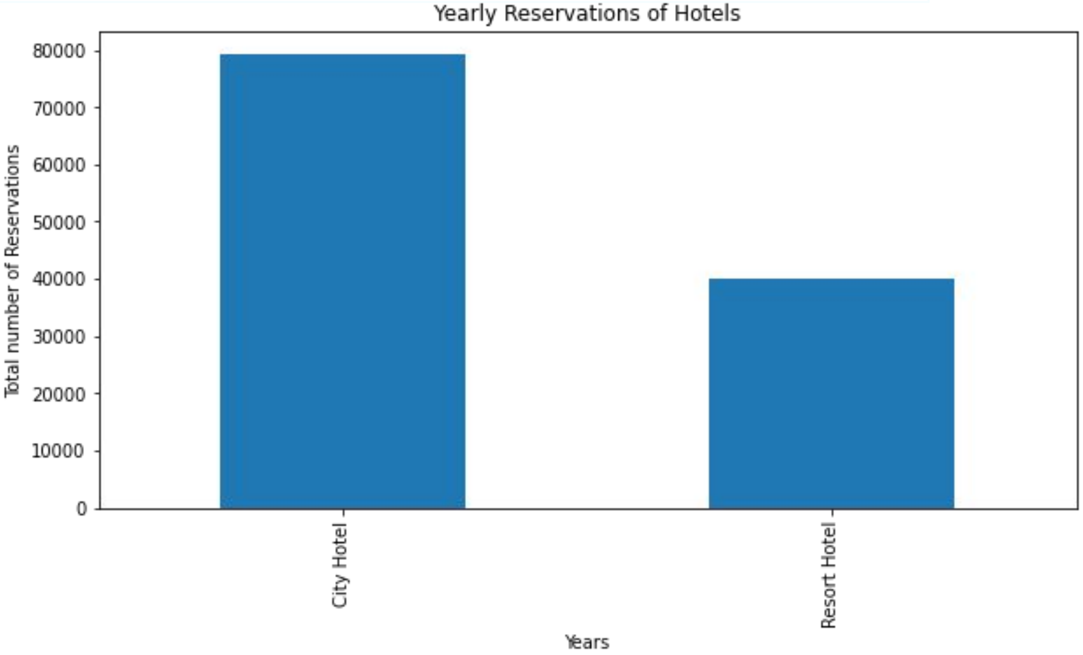


Correlation heatmap

	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

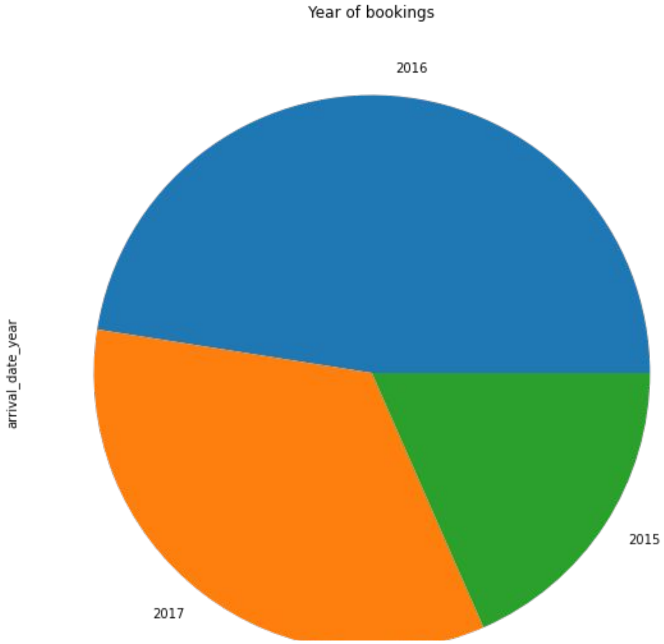
Hotel wise analysis

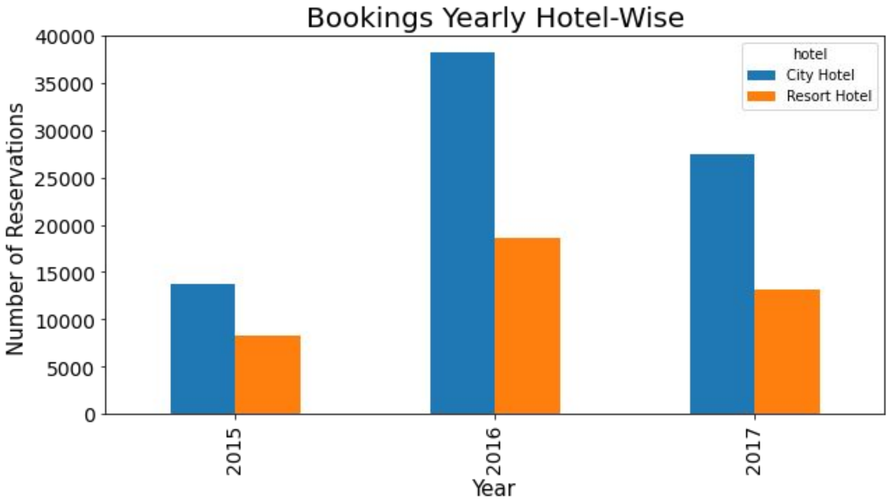
- **Which hotel has most number of bookings?**



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

- Which Year had the Maximum number of bookings throughout & on Hotel Basis?

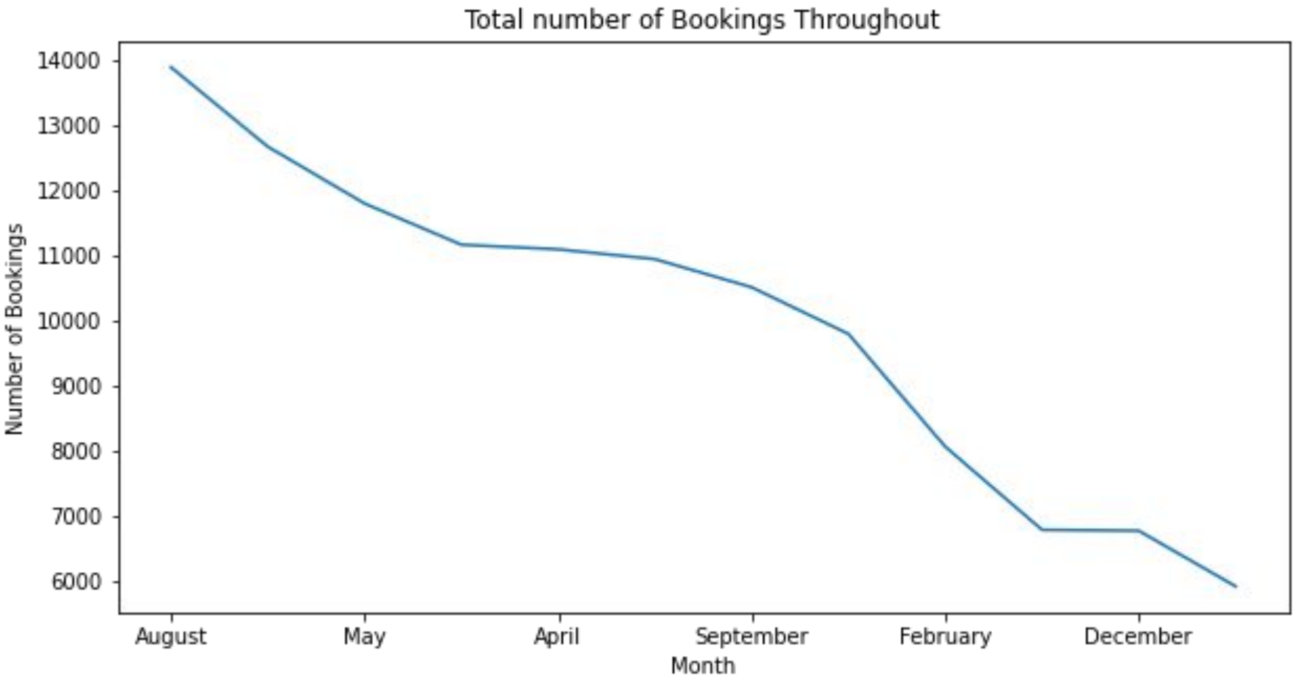




The Maximum number of bookings Are in the Year 2016 for “Both Hotels”

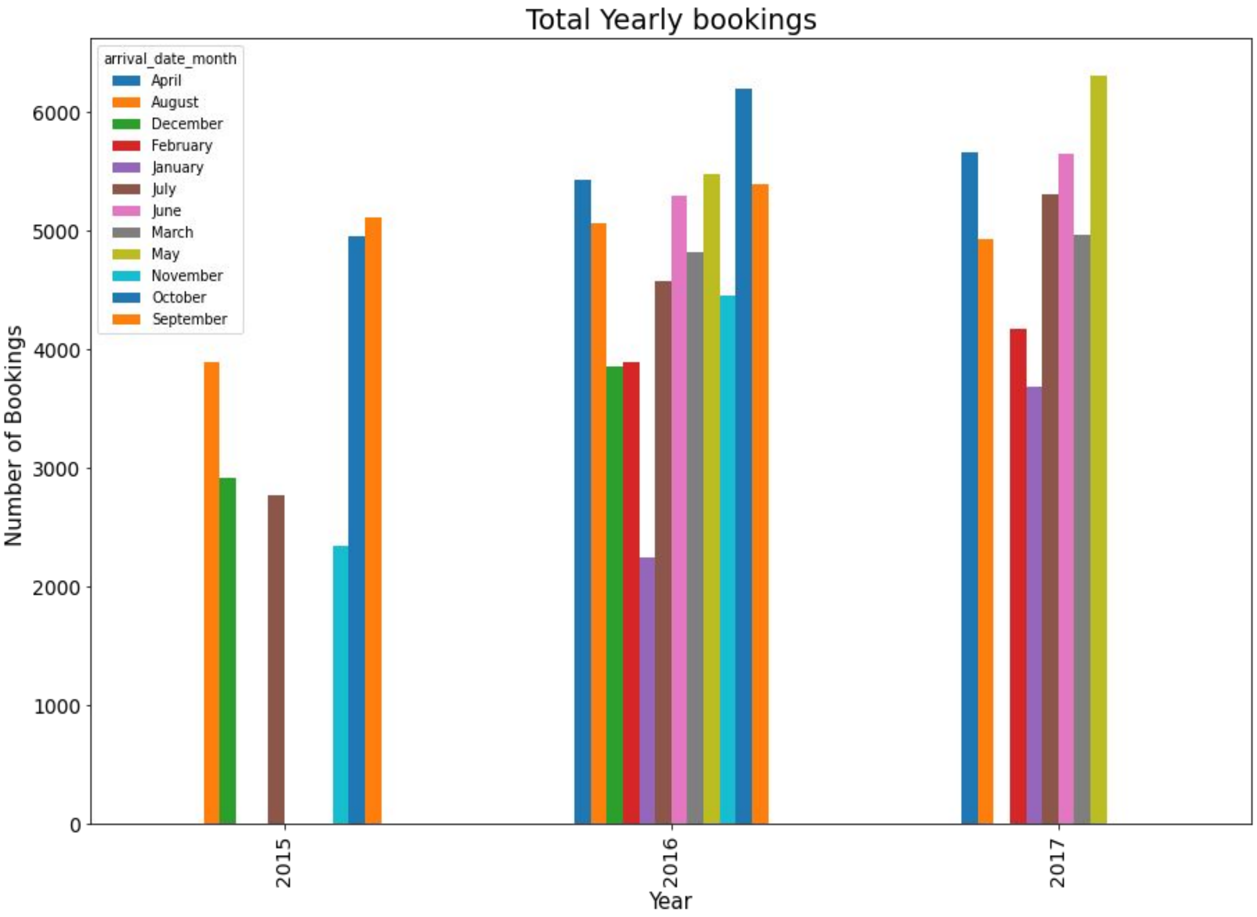
Bookings in the year 2016 has been increased tremendously as compared to the year 2015, but can also see that in the year 2017 we see a fall in the bookings compared to 2016 but Greater than in the year 2015.

- **Which Month had the Maximum number of bookings throughout?**



You can Clearly observe that throughout The period i.e from 2015 to 2017 we have the Maximum numer of bookings in the month of August and the Least Bookings are being totalled in the Month of December.

- ## Which Month had the maximum number of bookings in each year?

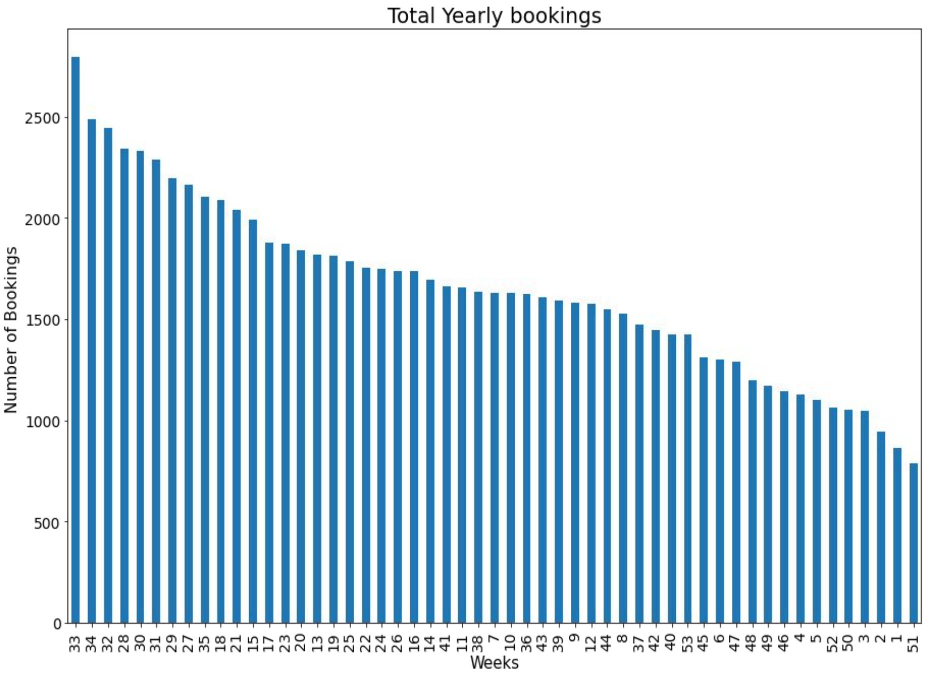


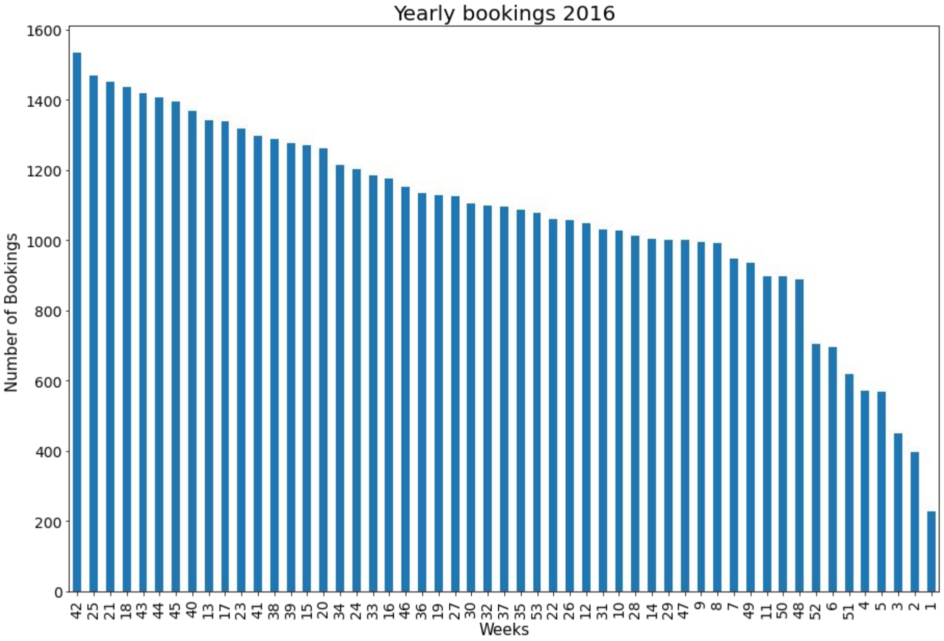
1.) September month with 5,114 bookings had the Maximum Number of bookings in the year 2015.

2.) October month with 6,203 bookings had the Maximum Number of bookings in the year 2016.

3.) May month with 6,313 bookings had the Maximum Number of bookings in the year 2017.

- **Which Week had the maximum number of bookings Throughout as well as in the year with maximum bookings i.e (YEAR-2016)?**

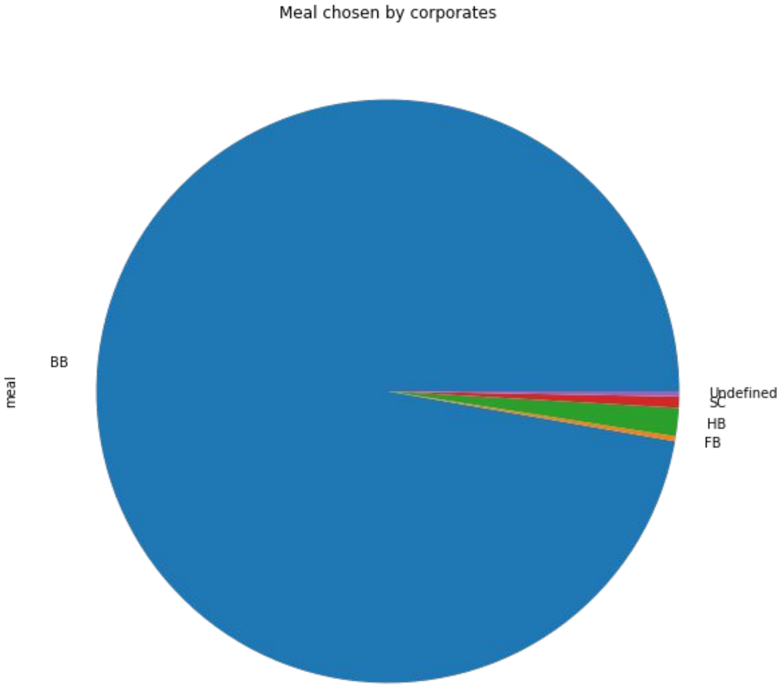

Total Yearly bookings

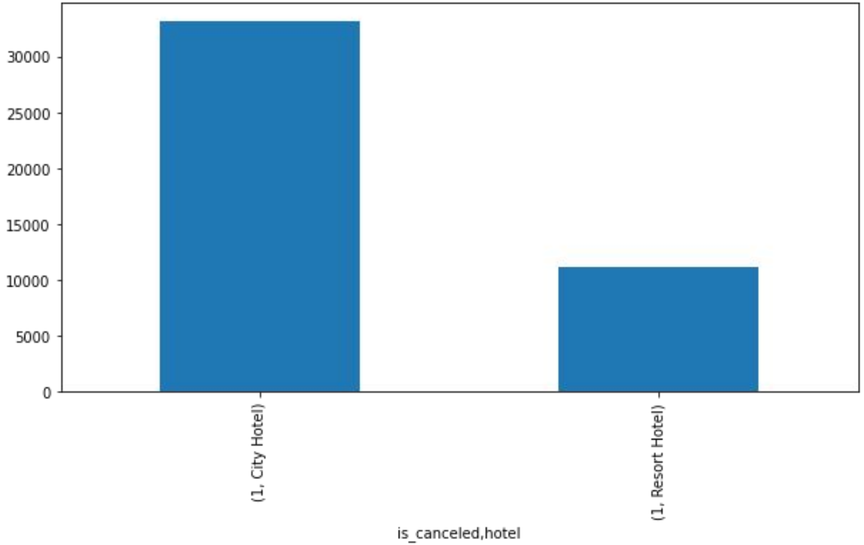

Yearly bookings 2016

For the Year 2016 you can see That in the 42nd week we had the Maximum Number of Bookings

Whereas for The whole period we see That In the 33rd week we had the Most Bookings.

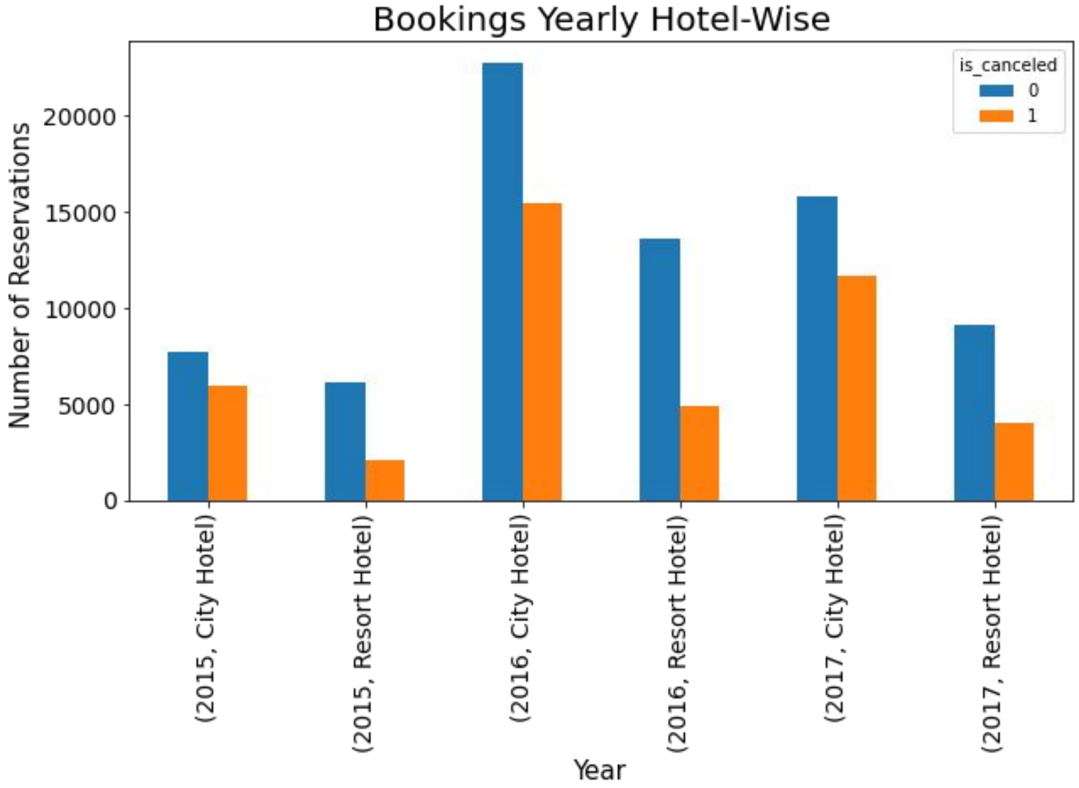
- What are the types of meals that the corporates prefer in their visit?
- Which hotel had Most Number of cancellations throughout?

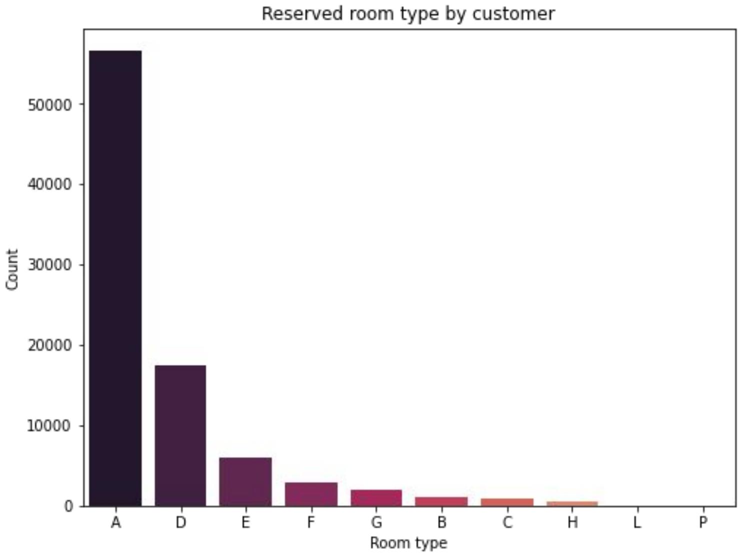




- Seems that corporates mostly prefer Bed and Breakfast package at the hotel, i.e. they check-in at night and check-out by the next morning.
- City Hotels had Maximum cancellations over the Years

- Which year has the highest number of cancellations Hotel-Wise?
- Which is the Most Reserved Room Type by customer ?

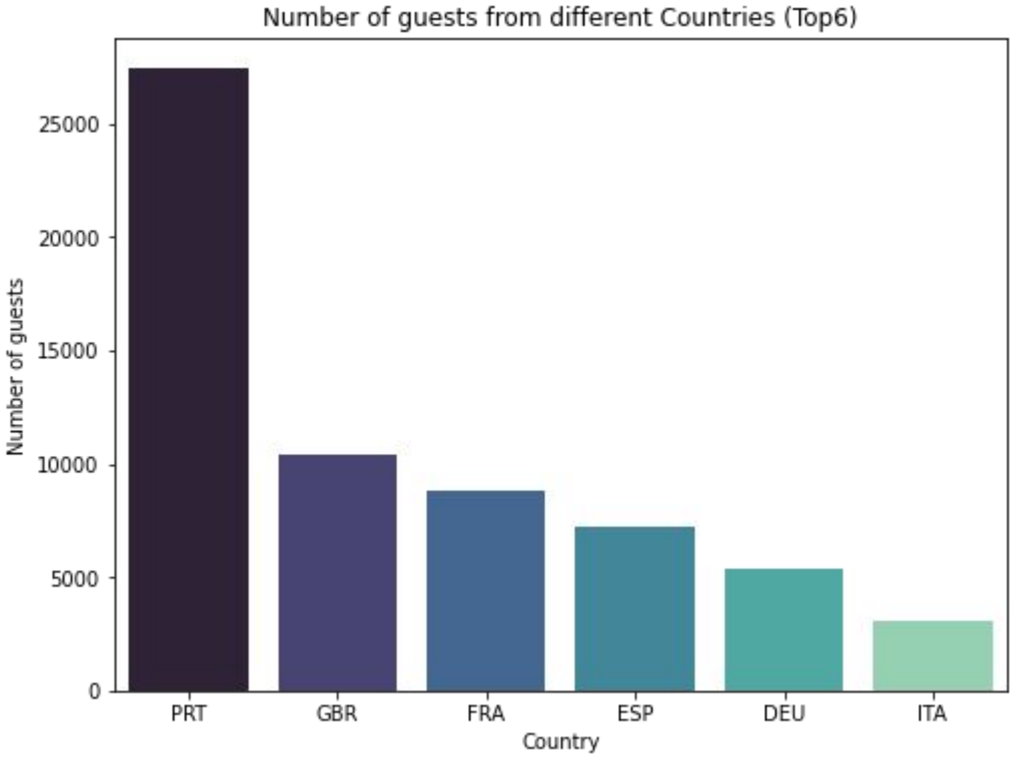




- Most reserved room type is A.
- Least preferred are L and P.

Year 2016 had most number of cancellations for City Hotel as well as Resort Hotel

- ## Which countries had most number of customers ?



PORTUGAL(PRT) had the Most Number of Bookings with 48,590 Bookings, Out of which 30,960 are City hotel bookings and 17,630 are Resort hotel Bookings

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