AlmaBetter

CAPSTONE PROJECT- I

EDA on Hotel Booking Analysis

BY

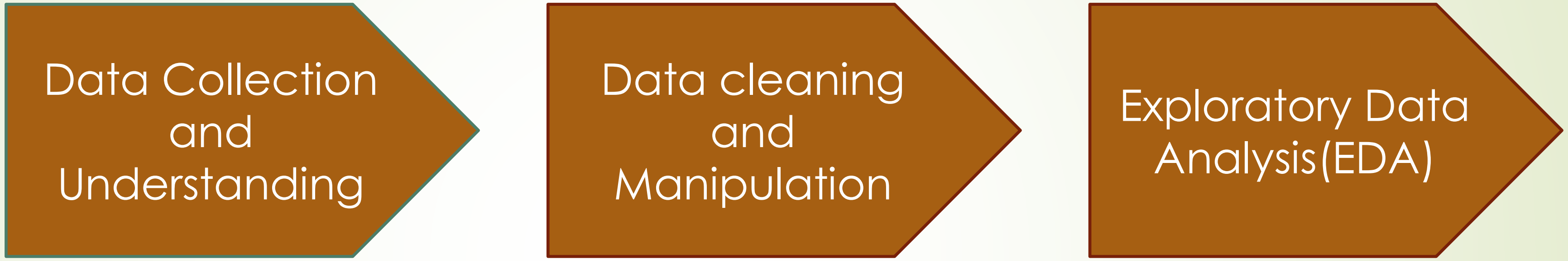
Monika Ransing

PROBLEM STATEMENT:

- In this project I am going to analyze Hotel Booking dataset. This dataset contains information of city hotel and resort hotel, and includes information of booking time, length of stay, number of adults, children and/or babies, also have information of available parking space, among other thing.
- The objective of this project is explore and analyze the data to discover important factors that govern the booking.

WORK FLOW:

- I divide my workflow into 3 steps.



Data Collection and Understanding

Data cleaning and Manipulation

Exploratory Data Analysis(EDA)

- **EDA is divided into 3 analysis:**

  1) Univariate Analysis: Here I am analysing only one variable.

  2) Bivariate Analysis: Here I am analysing two variables and their relationship.

  3) Multivariate Analysis: Here I am analysing more than two variables.

AlmaBetter

DATA COLLECTION AND UNDERSTANDING:

- Data collection and understanding are very important. So I have Hotel Booking data. This data contains **119390 rows** and **32 columns**. Let's understand the columns.
- **Dataset Description:**

○ **Hotel :** Type of hotel(City or Resort)

○ **is_canceled :** If the booking was canceled (1) or not (0)

○ **lead_time:** Number of days before the actual arrival of the guests

○ **arrival_date_year :** Year of arrival date

○ **arrival_date_month :** Month of arrival date

○ **arrival_date_week_number :** Week number of year for arrival date

○ **arrival_date_day_of_month :** Day of arrival date

○ **stays_in_weekend_nights :** Number of weekend nights (Saturday or Sunday) spent at the hotel by the guests.

○ **stays_in_week_nights :** Number of weeknights (Monday to Friday) spent at the hotel by the guests.

- **adults :** Number of adults among guests
- **children :** Number of children among guests
- **babies :** Number of babies among guests
- **meal :** Type of meal booked
- **country :** Country of guests
- **market_segment :** Designation of the market segment
- **distribution_channel :** Name of booking distribution channel
- **is_repeated_guest :** If the booking was from a repeated guest (1) or not (0)
- **previous_cancellations :** Number of previous bookings that were canceled by the customer prior to the current booking
- **previous_bookings_not_canceled :** Number of previous bookings not canceled by the customer prior to the current booking
- **reserved_room_type :** Code of room type reserved
- **assigned_room_type :** Code of room type assigned

- **booking_changes :** Number of changes/amendments made to the booking
- **deposit_type :** Type of the deposit made by the guest
- **agent :** ID of the travel agent who made the booking
- **company :** ID of the company that made the booking
- **days_in_waiting_list :** Number of days the booking was on the waiting list
- **customer_type :** Type of customer, 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
- **reservation_status :** Reservation status (Canceled, Check-Out or No-Show)
- **reservation_status_date :** Date at which the last reservation status was updated

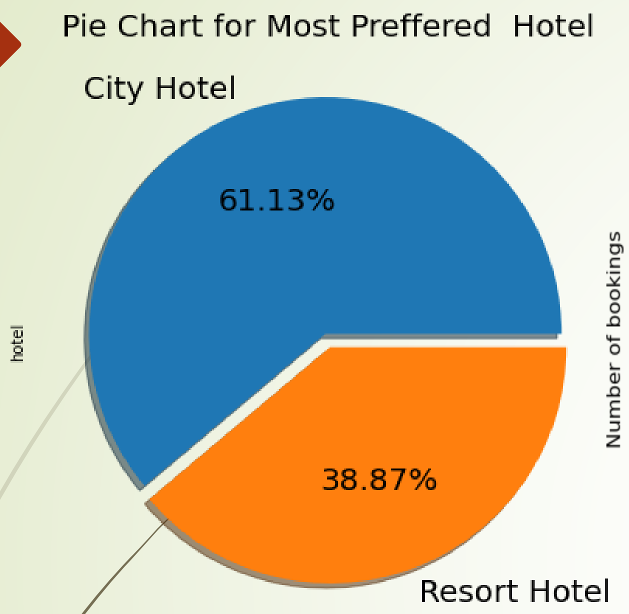
AlmaBetter

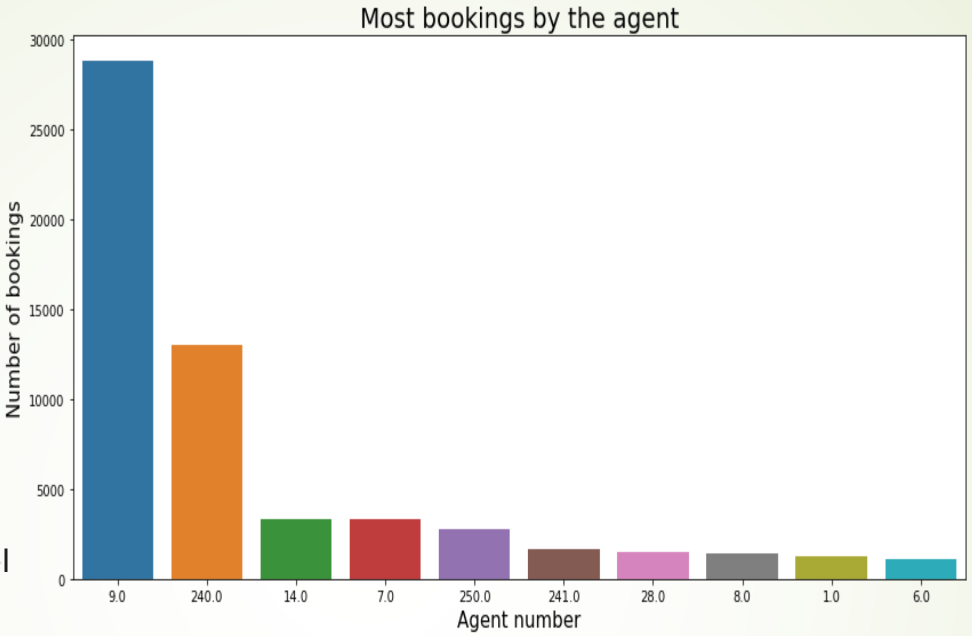
DATA CLEANING AND MANIPULATION:

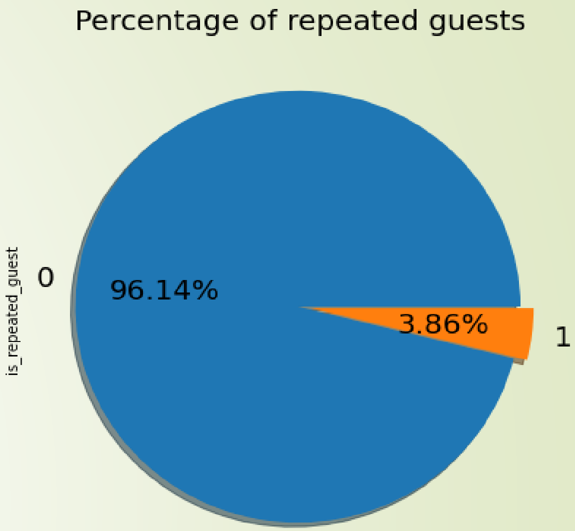
- Given data has **31994** duplicate values. So the duplicate values are dropped by using **drop_duplicates()**.
- Data has 4 columns that have missing values and those columns are **company(82137)**, **agent(12193)**, **country(452)** and **children(4)**. So the missing values of columns company, agent and children are replaced by **0** and the missing values of column country are replaced by others by using **.fillna()**.
- There are some rows in data that have the total number of adults, children or babies equal to zero this means there is no booking made. So I remove such rows.
- In the dataset I add 2 important columns and those columns are **"Total stay"** and **"Total People"**. For **total stay** column I add **'stays_in_weekend_night'** and **'stays_in_week_night'** columns and for **total people** column I add **'adults', 'children', 'babies'** columns.

AlmaBetter

EDA: UNIVARIATE ANALYSIS



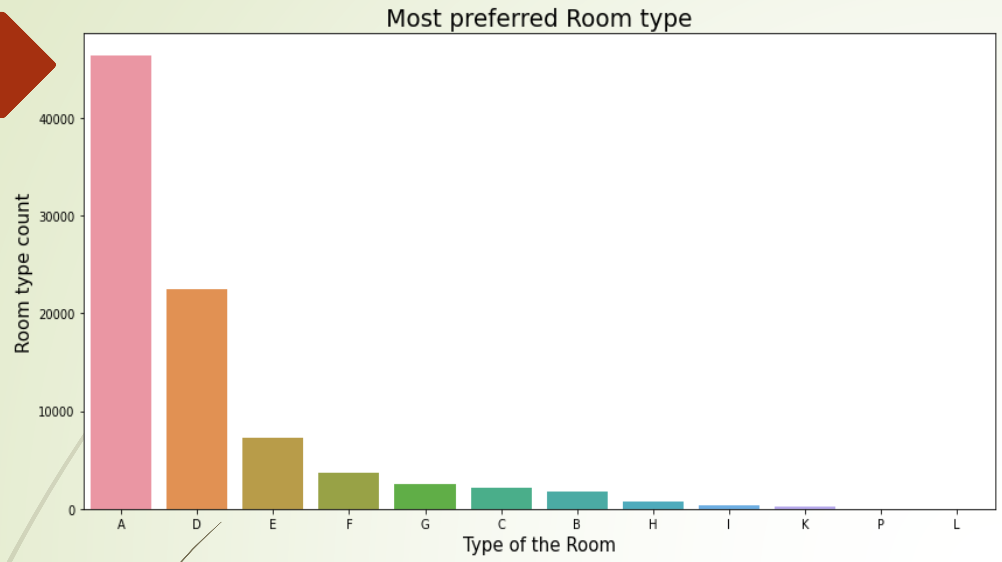


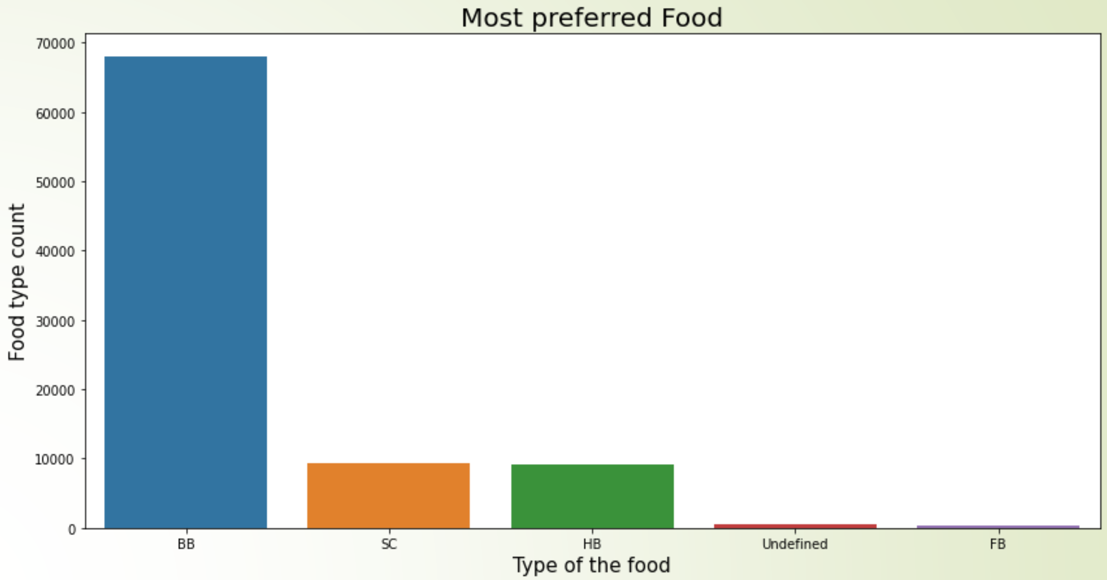


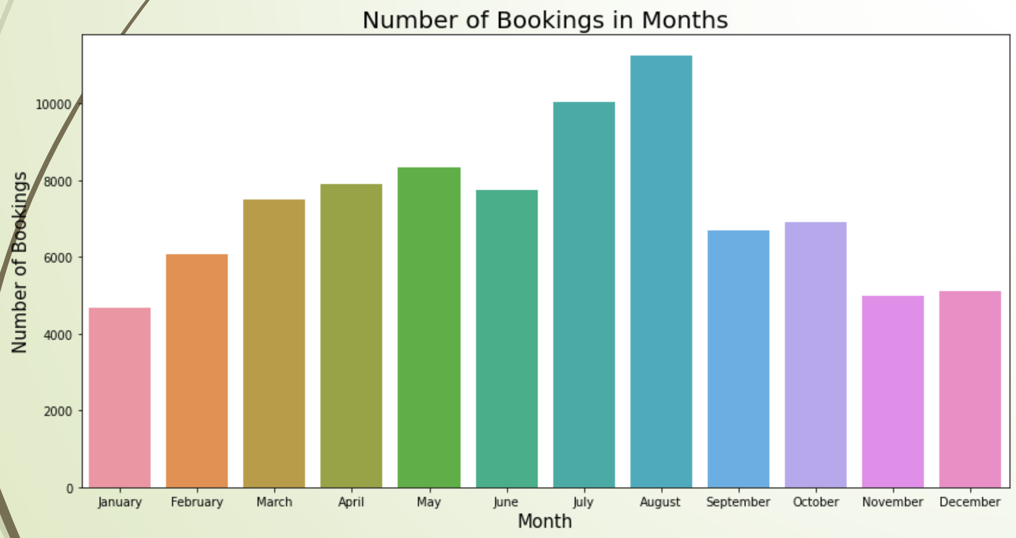
Conclusion:

1) **City hotel** is the most preferred hotel and the percentage is **61.13%** means city hotel is the busier hotel type.
2) **Agent no. 9** made the most bookings and the count is **28759** means agent no 9 is the most preferred agent for booking.
3) Percentage of repeated guests is very less which is **3.86%** means repeated guests do not prefer the same hotel for their stay.

EDA: UNIVARIATE ANALYSIS



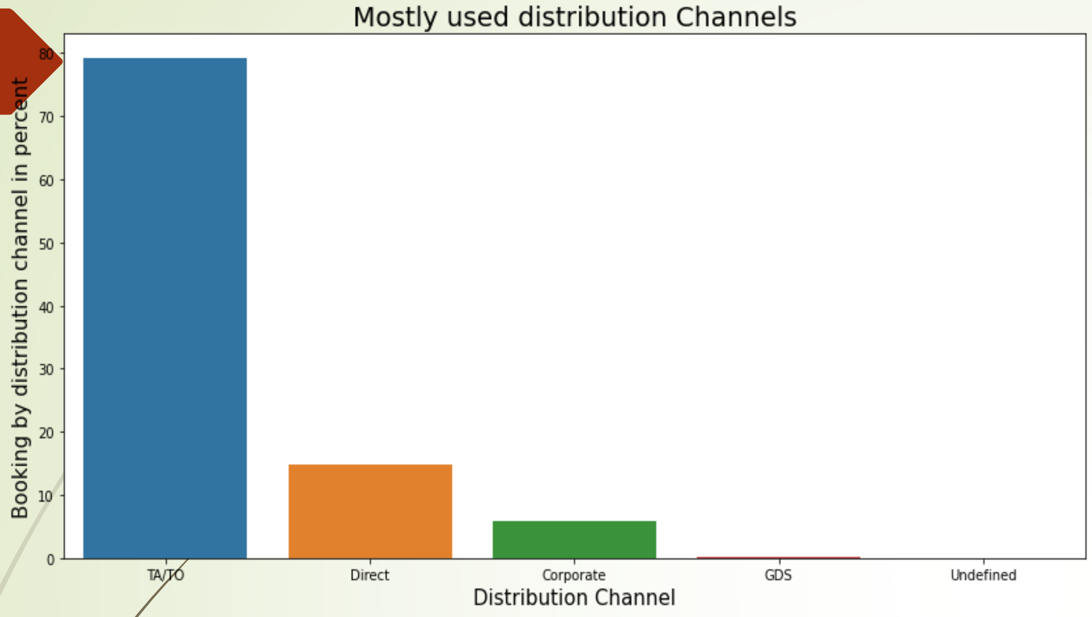


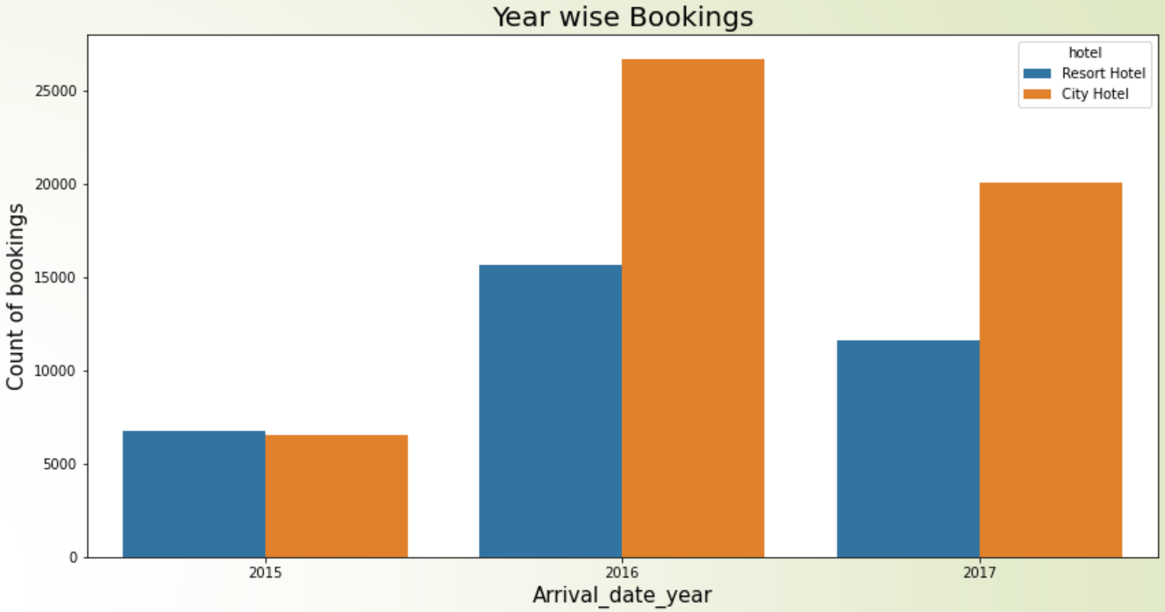


Conclusion:

1) **Room type A** is the most preferred room type and bookings are **46283**.
2) Most preferred food type is **BB type** and **67907** guests preferred BB type food.
3) **August** month has a maximum number of bookings and the count is **11242** means the august month is the busiest month in the year.

EDA: UNIVARIATE ANALYSIS

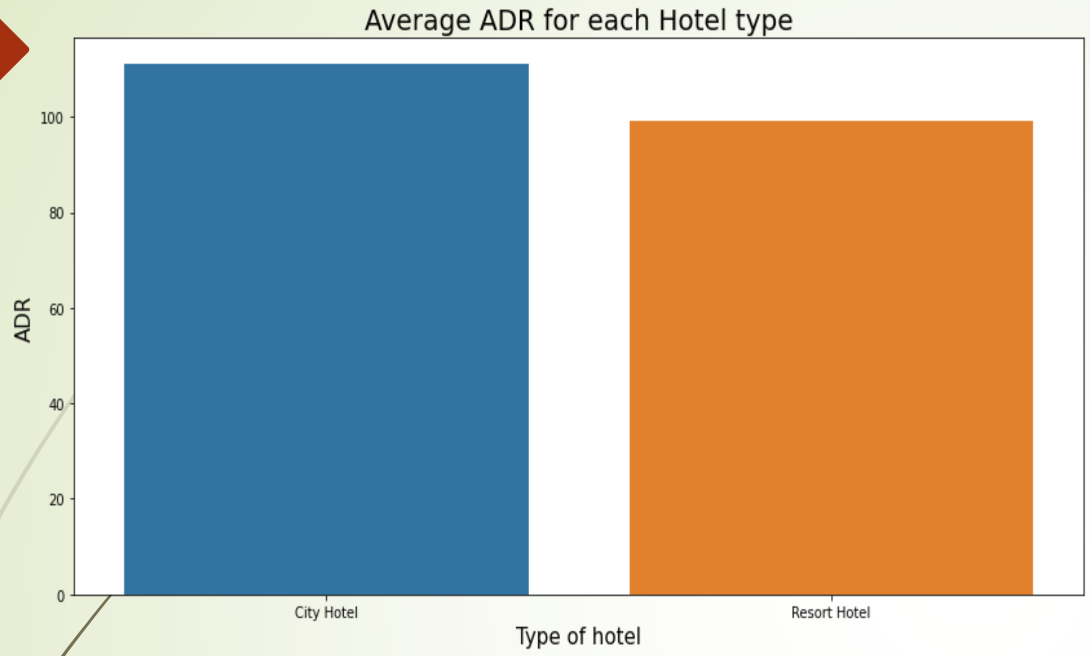


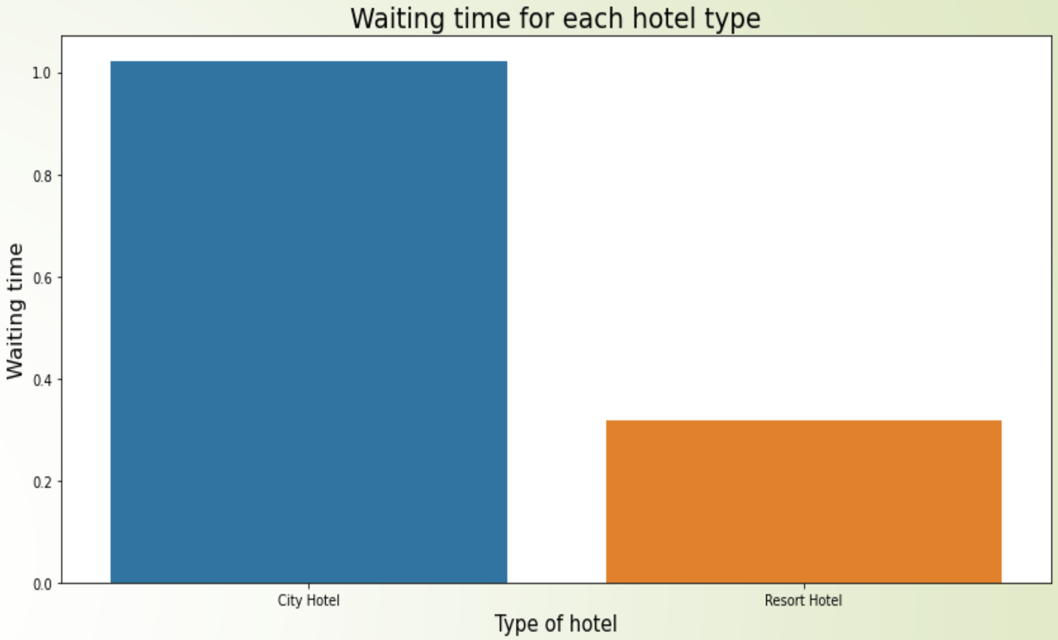


Conclusion:

1) Mostly preferred booking channel is **TA/TO** which has a percentage of booking is **79.13%** and the second preferred channel is the **Direct** channel which made **14.85%** of bookings
2) The **Year 2016** had the highest booking in the city as well as in resort hotel type and the bookings are **42313** and the year **2015** had **fewer bookings**.

EDA: BIVARIATE AND MULTIVARIATE ANALYSIS

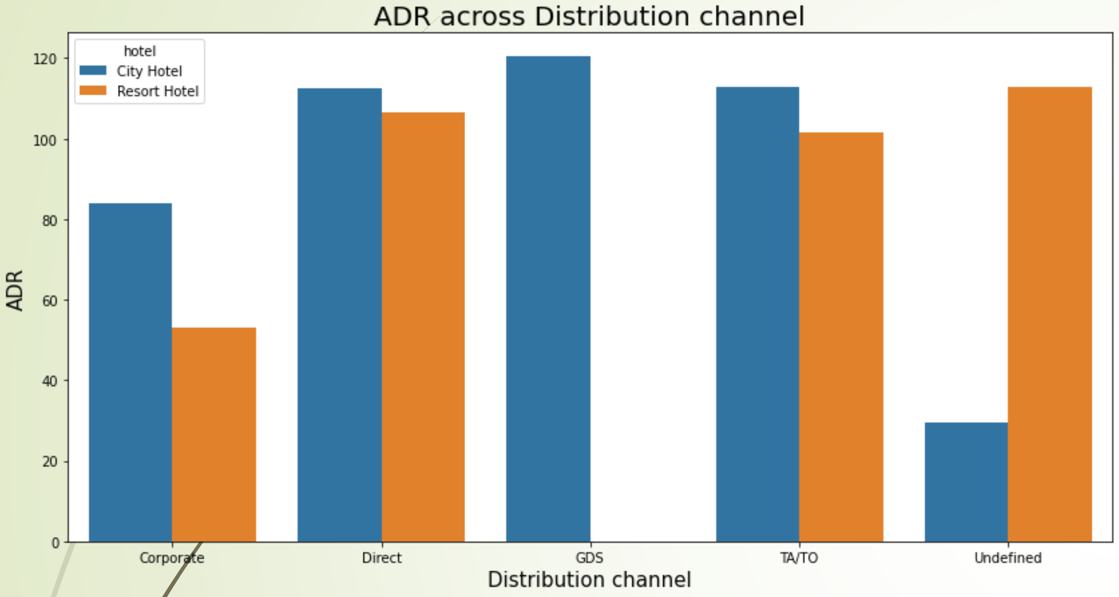


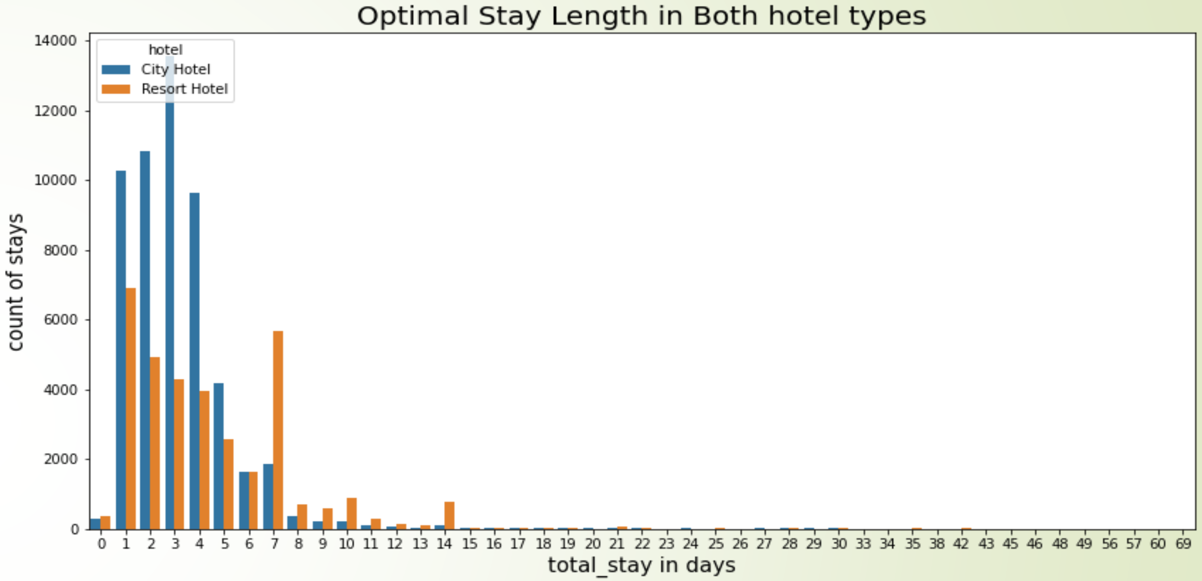


Conclusion:

1) **City hotel** has the highest ADR and the average ADR for city hotels is **110.98**. High ADR means high revenue so the city hotel has high revenue.
2) **City hotel** has a long waiting time so the city hotel is the busier hotel.

EDA: BIVARIATE AND MULTIVARIATE ANALYSIS



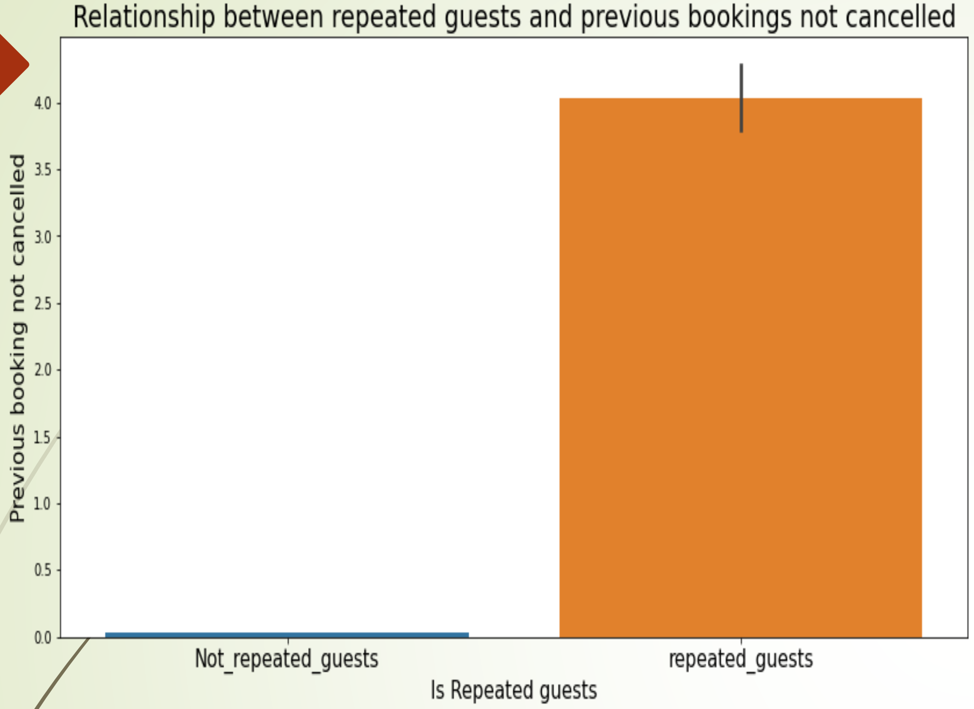


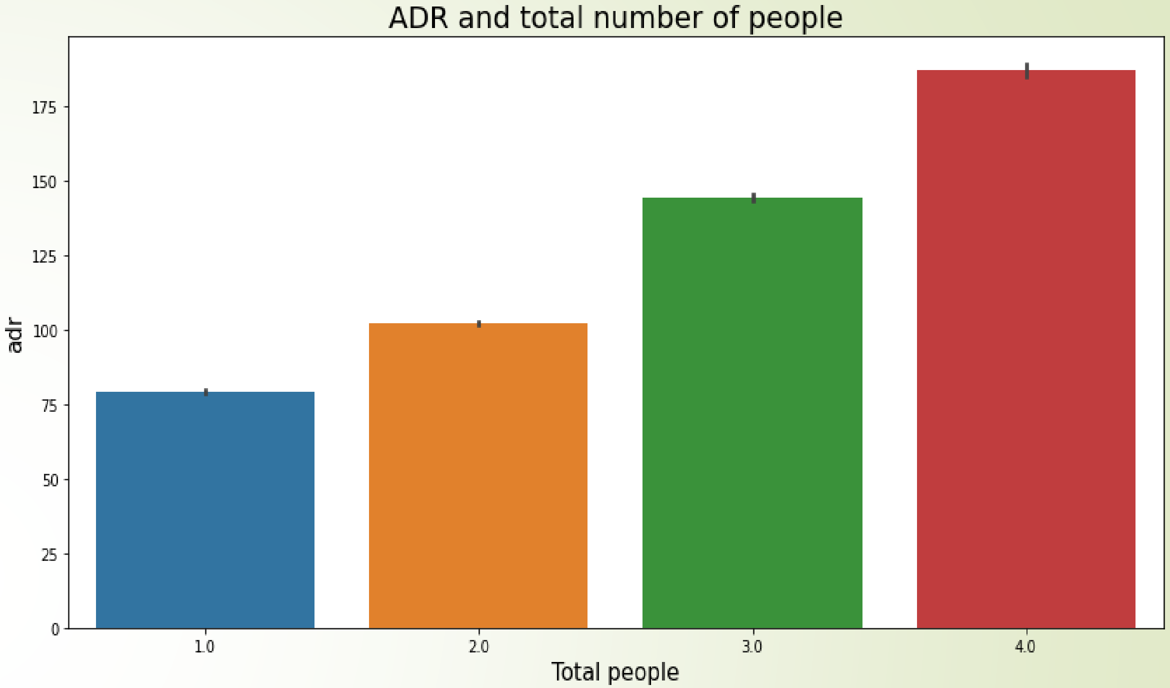
Conclusion:

1) **GDS distribution channel** contributed more to ADR in a city hotel and the **Direct & TA/TO** distribution channel has nearly equal contribution to ADR in both hotel types.

2) Optimal stay length in both hotel types is less than **7 days**.

EDA: BIVARIATE AND MULTIVARIATE ANALYSIS

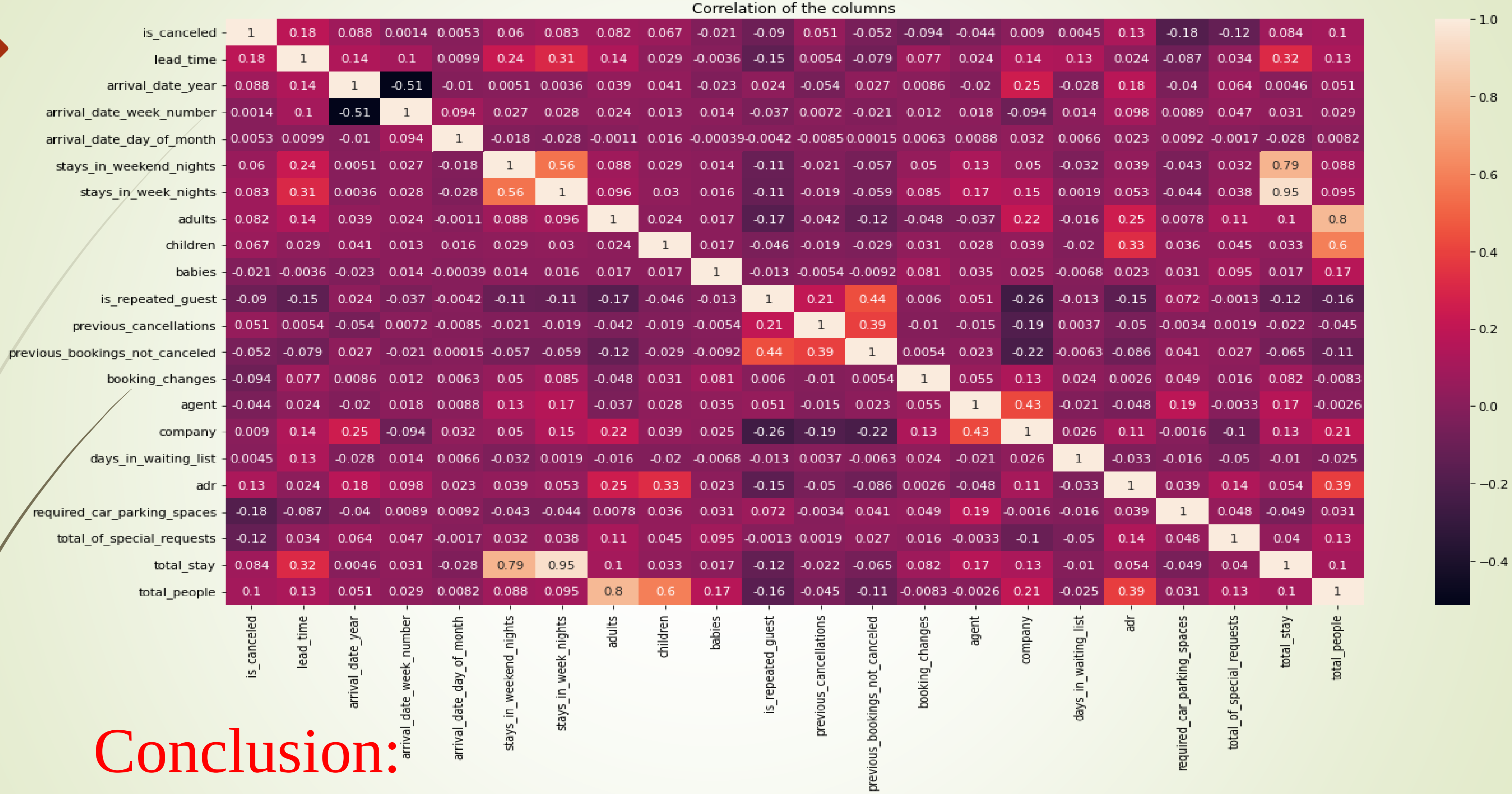




Conclusion:

1) Repeated guests do not cancel their previous bookings but non-repeated guests cancel their bookings.
2) If the number of people increases ADR is increasing and due to this revenue also increases.

EDA: BIVARIATE AND MULTIVARIATE ANALYSIS



Correlation of the columns

	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	0.18	0.088	0.0014	0.0053	0.06	0.083	0.082	0.067	-0.021	-0.09	0.051	-0.052	-0.094	-0.044	0.009	0.0045	0.13	-0.18	-0.12	0.084	0.1
lead_time	0.18	1	0.14	0.1	0.0099	0.24	0.31	0.14	0.029	-0.0036	-0.15	0.0054	-0.079	0.077	0.024	0.14	0.13	0.024	-0.087	0.034	0.32	0.13
arrival_date_year	0.088	0.14	1	-0.51	-0.01	0.0051	0.0036	0.039	0.041	-0.023	0.024	-0.054	0.027	0.0086	-0.02	0.25	-0.028	0.18	-0.04	0.064	0.0046	0.051
arrival_date_week_number	0.0014	0.1	-0.51	1	0.094	0.027	0.028	0.024	0.013	0.014	-0.037	0.0072	-0.021	0.012	0.018	-0.094	0.014	0.098	0.0089	0.047	0.031	0.029
arrival_date_day_of_month	0.0053	0.0099	-0.01	0.094	1	-0.018	-0.028	-0.0011	0.016	-0.00039	-0.0042	-0.0085	0.00015	0.0063	0.0088	0.032	0.0066	0.023	0.0092	-0.0017	-0.028	0.0082
stays_in_weekend_nights	0.06	0.24	0.0051	0.027	-0.018	1	0.56	0.088	0.029	0.014	-0.11	-0.021	-0.057	0.05	0.13	0.05	-0.032	0.039	-0.043	0.032	0.79	0.088
stays_in_week_nights	0.083	0.31	0.0036	0.028	-0.028	0.56	1	0.096	0.03	0.016	-0.11	-0.019	-0.059	0.085	0.17	0.15	0.0019	0.053	-0.044	0.038	0.95	0.095
adults	0.082	0.14	0.039	0.024	-0.0011	0.088	0.096	1	0.024	0.017	-0.17	-0.042	-0.12	-0.048	-0.037	0.22	-0.016	0.25	0.0078	0.11	0.1	0.8
children	0.067	0.029	0.041	0.013	0.016	0.029	0.03	0.024	1	0.017	-0.046	-0.019	-0.029	0.031	0.028	0.039	-0.02	0.33	0.036	0.045	0.033	0.6
babies	-0.021	-0.0036	-0.023	0.014	-0.00039	0.014	0.016	0.017	0.017	1	-0.013	-0.0054	-0.0092	0.081	0.035	0.025	-0.0068	0.023	0.031	0.095	0.017	0.17
is_repeated_guest	-0.09	-0.15	0.024	-0.037	-0.0042	-0.11	-0.11	-0.17	-0.046	-0.013	1	0.21	0.44	0.006	0.051	-0.26	-0.013	-0.15	0.072	-0.0013	-0.12	-0.16
previous_cancellations	0.051	0.0054	-0.054	0.0072	-0.0085	-0.021	-0.019	-0.042	-0.019	-0.0054	0.21	1	0.39	-0.01	-0.015	-0.19	0.0037	-0.05	-0.0034	0.0019	-0.022	-0.045
previous_bookings_not_canceled	-0.052	-0.079	0.027	-0.021	0.00015	-0.057	-0.059	-0.12	-0.029	-0.0092	0.44	0.39	1	0.0054	0.023	-0.22	-0.0063	-0.086	0.041	0.027	-0.065	-0.11
booking_changes	-0.094	0.077	0.0086	0.012	0.0063	0.05	0.085	-0.048	0.031	0.081	0.006	-0.01	0.0054	1	0.055	0.13	0.024	0.0026	0.049	0.016	0.082	-0.0083
agent	-0.044	0.024	-0.02	0.018	0.0088	0.13	0.17	-0.037	0.028	0.035	0.051	-0.015	0.023	0.055	1	0.43	-0.021	-0.048	0.19	-0.0033	0.17	-0.0026
company	0.009	0.14	0.25	-0.094	0.032	0.05	0.15	0.22	0.039	0.025	-0.26	-0.19	-0.22	0.13	0.43	1	0.026	0.11	-0.0016	-0.1	0.13	0.21
days_in_waiting_list	0.0045	0.13	-0.028	0.014	0.0066	-0.032	0.0019	-0.016	-0.02	-0.0068	-0.013	0.0037	-0.0063	0.024	-0.021	0.026	1	-0.033	-0.016	-0.05	-0.01	-0.025
adr	0.13	0.024	0.18	0.098	0.023	0.039	0.053	0.25	0.33	0.023	-0.15	-0.05	-0.086	0.0026	-0.048	0.11	-0.033	1	0.039	0.14	0.054	0.39
required_car_parking_spaces	-0.18	-0.087	-0.04	0.0089	0.0092	-0.043	-0.044	0.0078	0.036	0.031	0.072	-0.0034	0.041	0.049	0.19	-0.0016	-0.016	0.039	1	0.048	-0.049	0.031
total_of_special_requests	-0.12	0.034	0.064	0.047	-0.0017	0.032	0.038	0.11	0.045	0.095	-0.0013	0.0019	0.027	0.016	-0.0033	-0.1	-0.05	0.14	0.048	1	0.04	0.13
total_stay	0.084	0.32	0.0046	0.031	-0.028	0.79	0.95	0.1	0.033	0.017	-0.12	-0.022	-0.065	0.082	0.17	0.13	-0.01	0.054	-0.049	0.04	1	0.1
total_people	0.1	0.13	0.051	0.029	0.0082	0.088	0.095	0.8	0.6	0.17	-0.16	-0.045	-0.11	-0.0083	-0.0026	0.21	-0.025	0.39	0.031	0.13	0.1	1

Conclusion:

1) arrival_date_year and arrival_date week_number columns has na egative correlation which is -0.51.

2) Stays_in_week_nights and total_stay has a positive correlation which is 0.95.

BUSINESS OBJECTIVE:

1) To increase hotel business some factors are important like high revenue, generation, customer satisfaction, facilities provided by the hotel, etc.

2) I am able to achieve the same things by showing the client which hotel is most preferred, the percentage of repeated guests, mostly preferred food by guests, then which hotel has the highest ADR, etc.

3) Most preferred room type is achieved by counterplot so the client can be well prepared in advance and this insight helps the client for further enhancement of their hospitality.

4) I am able to show which food type is mostly preferred so the client can offer the most preferred food to the guests.

5) Most preferred months are shown by barplot so the client can be well prepared in advance so that minimum grievances would be faced by the client.

6) Using barplot I am able to show which hotel type has high adr so the client can analyze which hotel has a high income.

7) I am able to show which hotel is the busiest hotel so the client can do relatable changes in facilities in less busy hotel types.

8) I am able to show the relationship between repeated guests and previous bookings not canceled so the client can prefer repeated guests.

9) Using barplot relationship between ADR and the total number of people is shown so the client can prefer the maximum number of people.

CONCLUSION:

1) City hotel is mostly preferred hotel by guests.

2) Agent no. 9 made the most bookings.

3) Percentage of repeated guests is less which is 3.86%.

4) Room type A is mostly the preferred room type.

5) Mostly preferred food type is BB type food.

6) August month has the most bookings and after august july has the most bookings.

7) TA/TO distribution channel is mostly used and the percentage age is 79.13%.

8) City hotel has the highest ADR. The highest ADR means more revenue.

9) 2016 year had the highest number of bookings and bookings were 42313.

10) City hotel has higher waiting time means city hotel is the busier hotel.

11) GDS distribution channel contributed most in ADR in city hotels but no contribution in the resort hotel.

12) Optimal stay length in both hotel types is less than 7 days.

13) Repeated guests do not cancel their bookings but not repeated guests cancel.

14) If the number of people is more then ADR also increases means revenue increases.

15) arrival_date_year and arrival_date_week_number columns have a negative correlation which is -0.51.

16) stays_in_weel_nights and total_stay columns have a positive correlation which is 0.95.