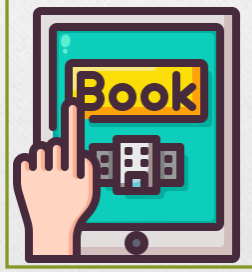
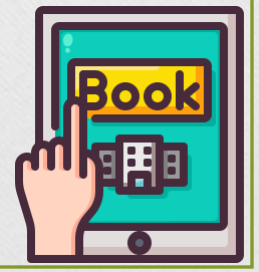
# Capstone Project

## Hotel Booking Analysis - EDA

**Insights & Suggestions for**

**Optimal Booking Management**




**By – Javed Ali**

**Email: jvdali.e@gmail.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 in order to get the best daily rate? What if you wanted to predict whether or not 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 analyze the data to discover important factors that govern the bookings.

# Business Objective

Analyse the data on bookings of City Hotel and Resort Hotel to gain insights on the different factors that affect the booking. This is undertaken as an individual project.

# Contents

**Step 1 : Understanding Problem Statement (Business Problem)**

**Step 2 : Data Acquisition**

**Step 3 : Data Processing/Formatting (includes Data Cleaning & Modeling and Feature Engineering)**

**Step 4 : Exploratory Data Analysis (EDA)**


**EDA is divided into 3 analysis:**

- **a) Univariate Analysis: Analysing only one variable.**
- **b) Bivariate Analysis: Analysing two variables and their relationship.**
- **c) Multivariate Analysis: Analysing more than two variables.**

# Hotel Booking Dataset Overview

- **The dataset contains information about bookings in two types of hotels. One of the hotels is a Resort Hotel and the other is a City Hotel.**
- **This data contains 1,19,390 rows and 32 columns. Out of which 40,060 observations are for Resort Hotel and 79,330 observations are for City Hotel.**
- **Each observation/record represents a hotel booking.**
- **Bookings recorded from July 1, 2015, to August 31, 2017, include successful and cancelled bookings.**
- **All data elements related to hotel and customer identification have been removed to ensure data privacy.**

**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

## Dataset Description:

- **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

## Dataset Description:

- **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

# Data cleaning and manipulation:

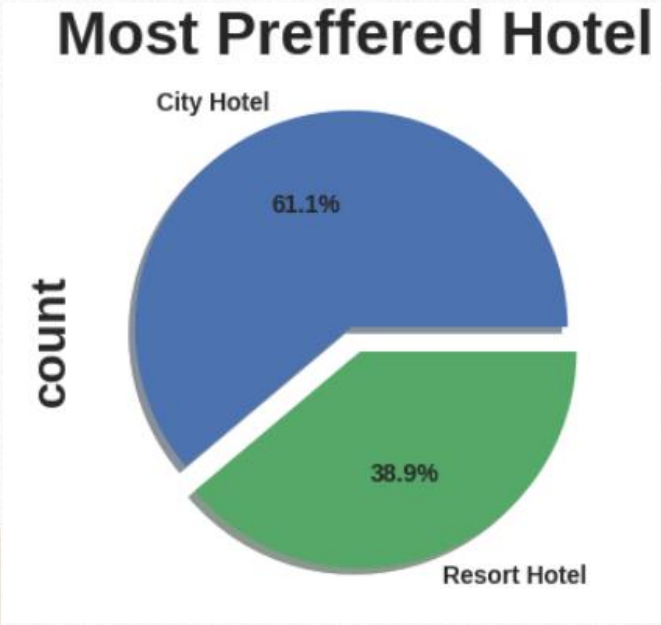
1. Given data has 31,994 duplicate values. So, the duplicate values are dropped by using drop_duplicates()

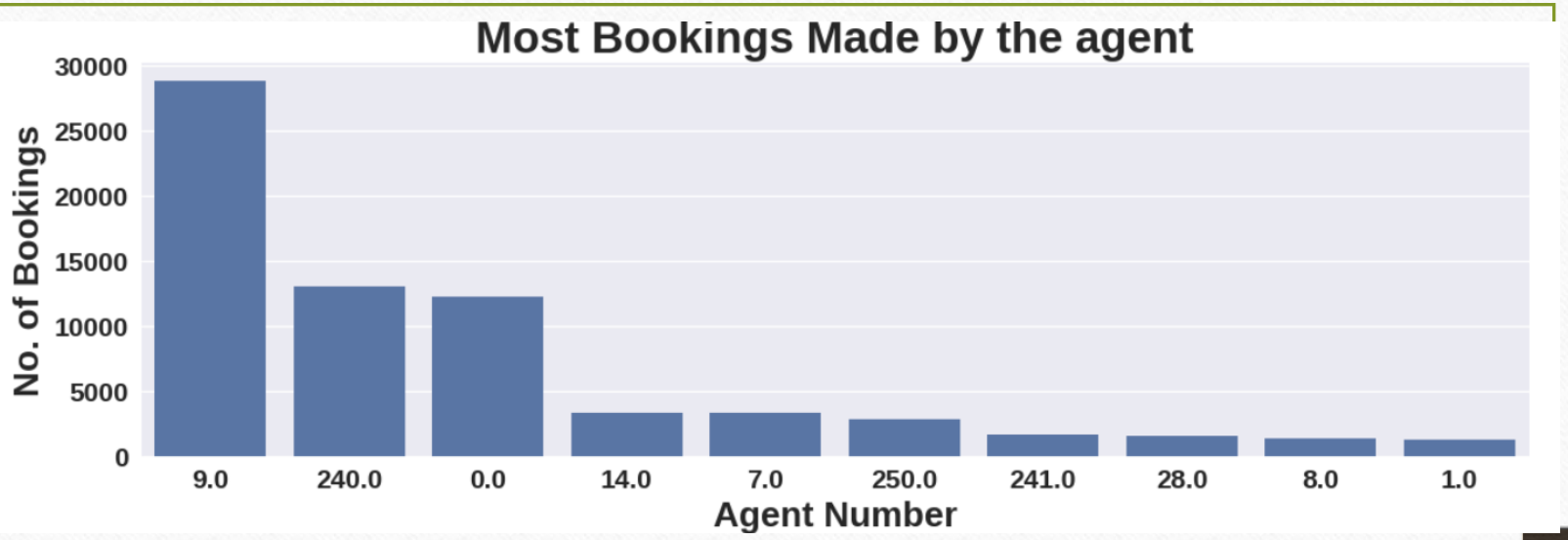
2. Data has 4 columns that have missing values, and those columns are company (82,137), agent(12,193) , country(452) and children(4). So, the missing values of columns company, agent and children are replaced by 0 and the missing values of column country are replaced by using .fillna()

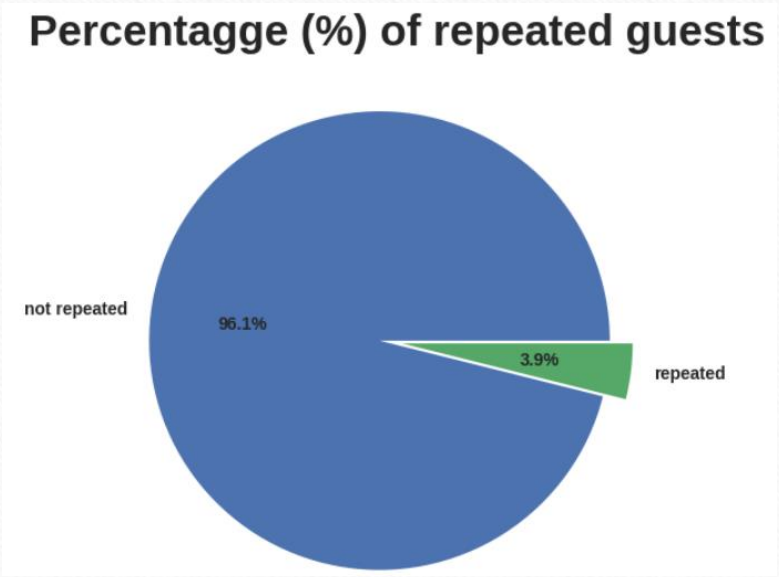
3. There are some rows in data that have the total number of adults, children or babies equal to zero this means there is no booking made. So, I removed such rows.

4. In the dataset, I added 2 important columns, and those columns are "Total stay" and "Total People". For "Total stay" column I added 'stays_in_weekend_night' and 'stays_in_week_night' columns and for "Total People" column I added 'adults', 'children', 'babies' columns.

# EDA: UNIVARIATE ANALYSIS



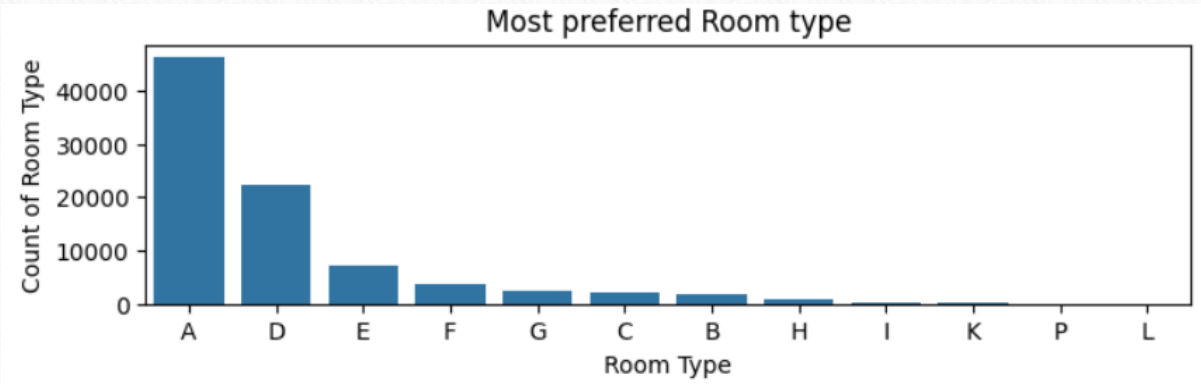

**Conclusion:**

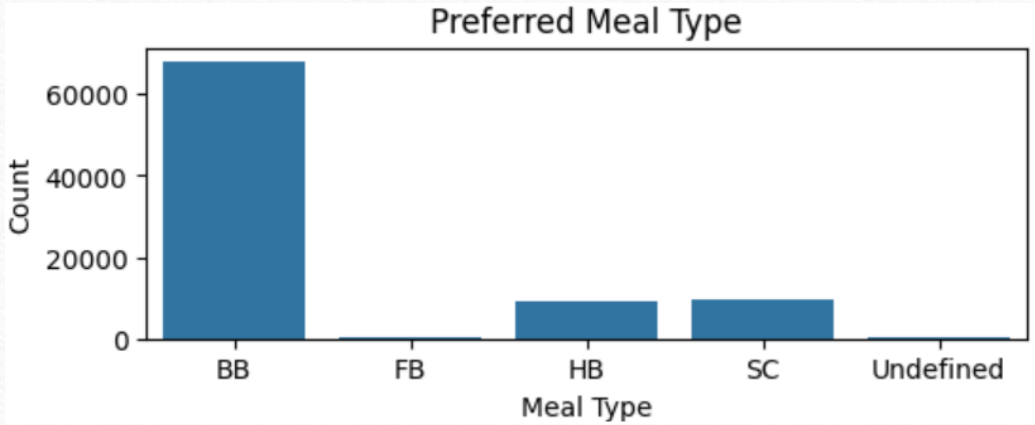
**1) City hotel is the most preferred hotel and the percentage is 61.13% means city hotel is the busier hotel type.**

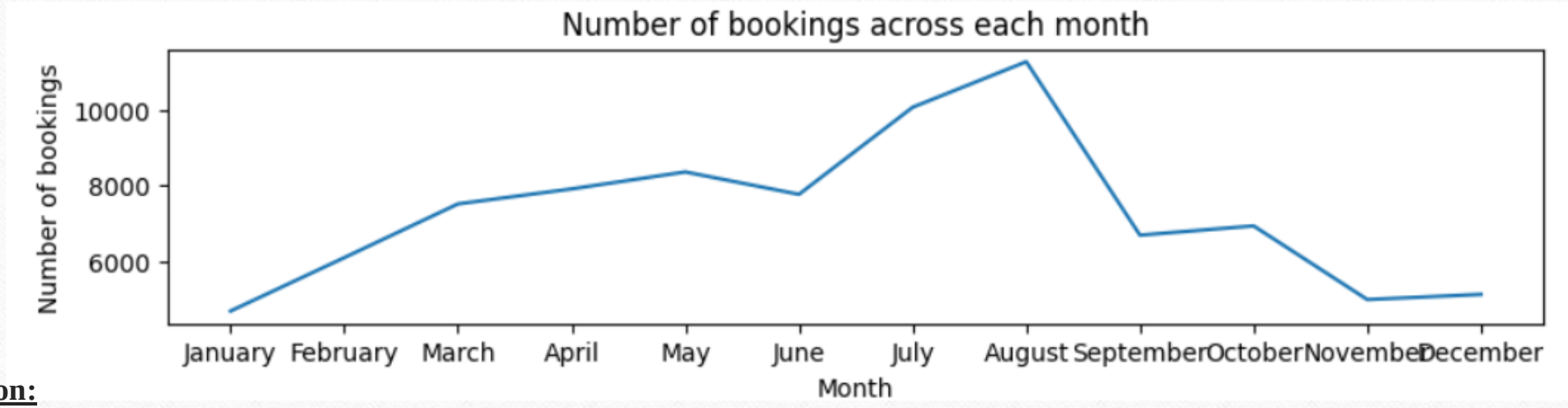
**2) Agent no. 9 made the most bookings and the count is 28,759 means agent no#9 is the most preferred agent for booking.**

**3) Percentage of repeated guests is very less which is 3.86%, it means guests do not prefer the same hotel for their stay.**

# EDA: UNIVARIATE ANALYSIS



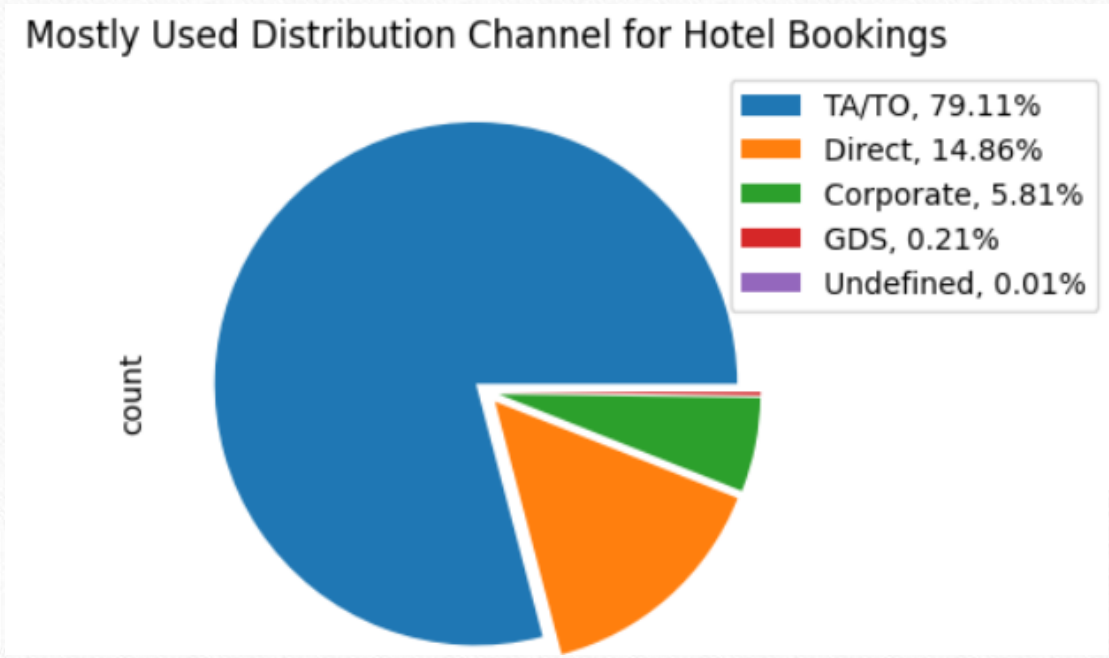

**Conclusion:**

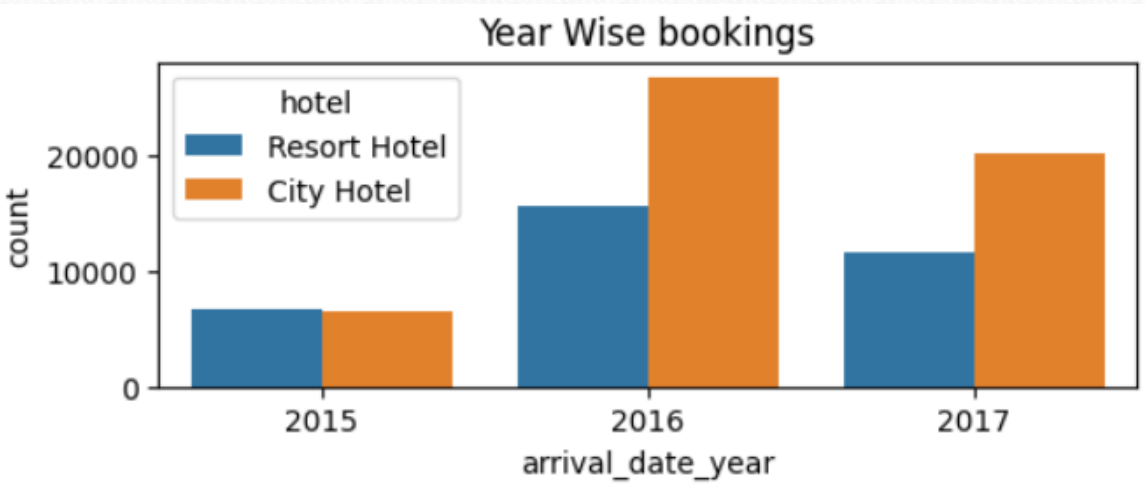
1) **Room type A** is the most preferred room type and bookings are **46,283**.

2) Most preferred food type is **BB type(Bed & Breakfast)** and **67,907** guests preferred BB type food.

3) **August month** has a maximum number of bookings, and the count is 11,242 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.**

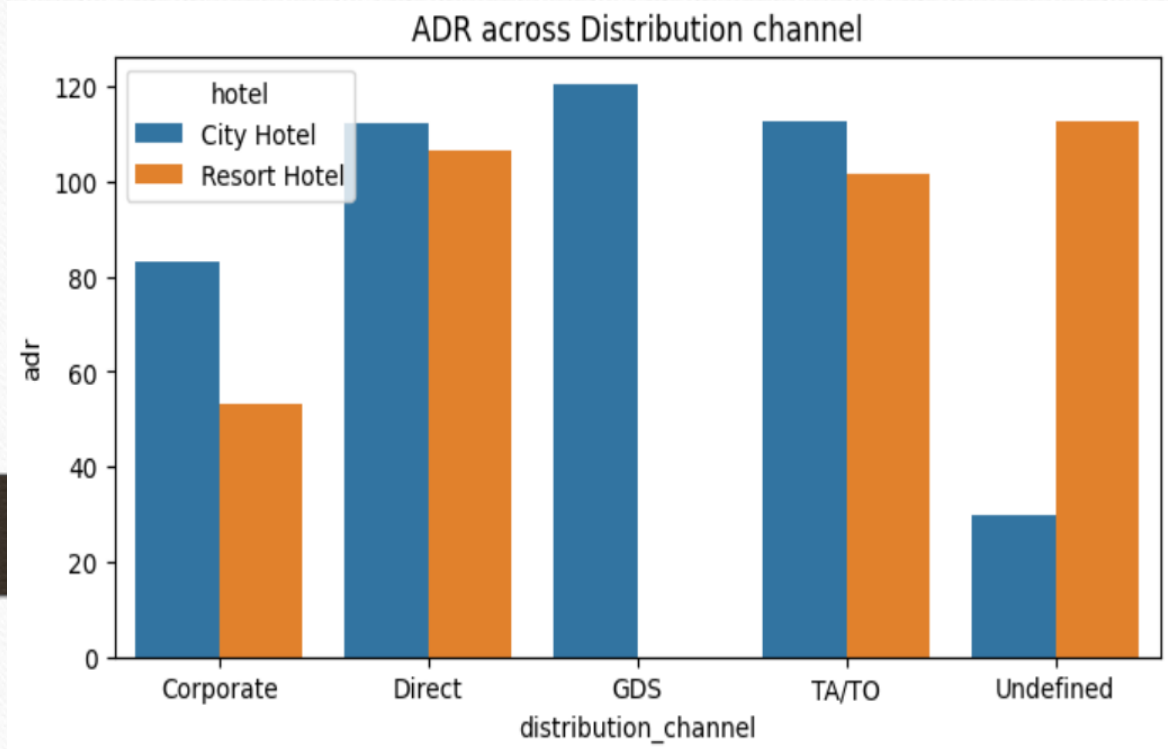
# EDA: BIVARIATE AND MULTIVARIATE ANALYSIS

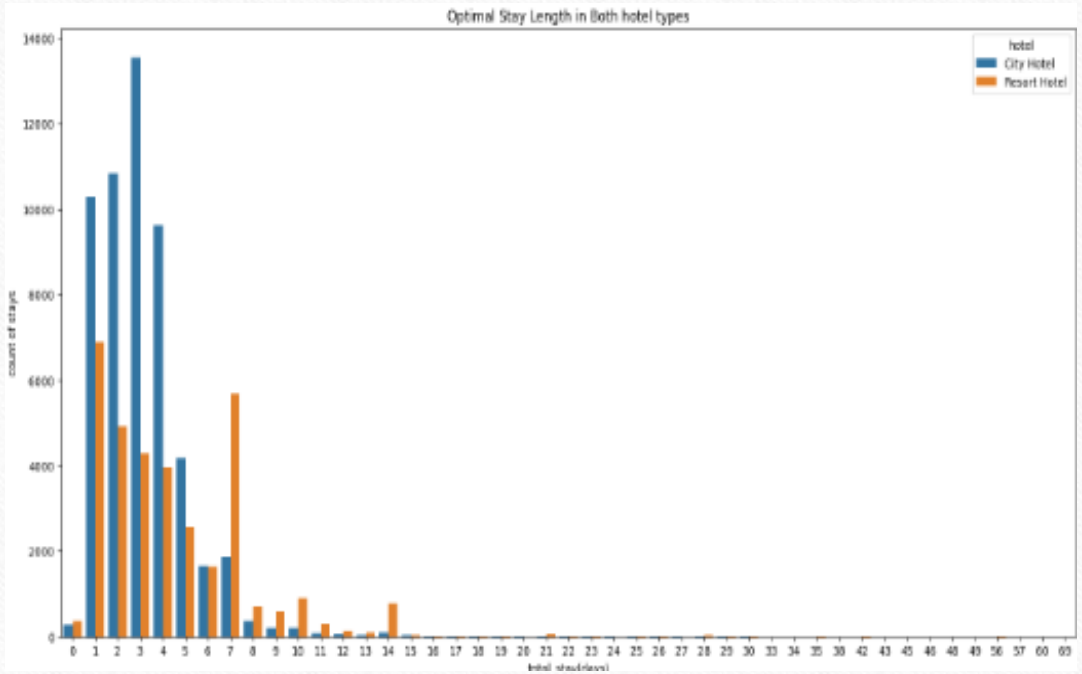


**Conclusion:**

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

# EDA: BIVARIATE AND MULTIVARIATE ANALYSIS



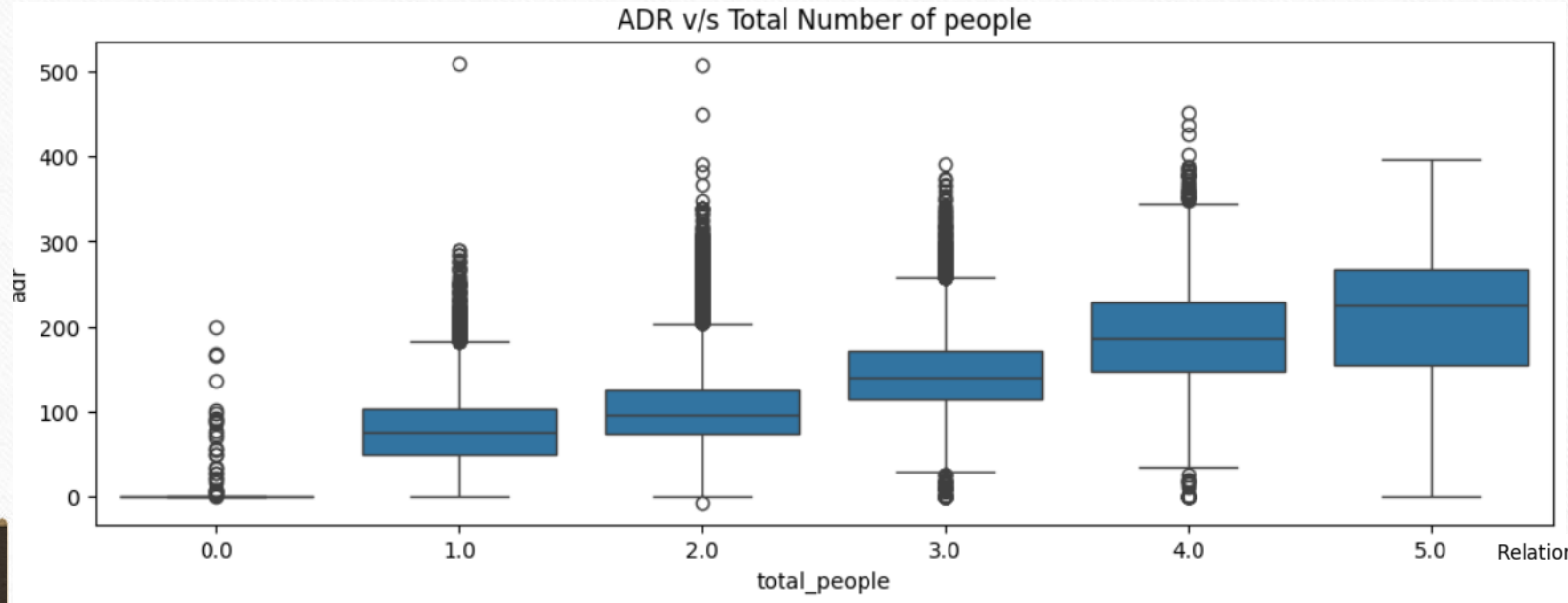




**Conclusion:**

**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



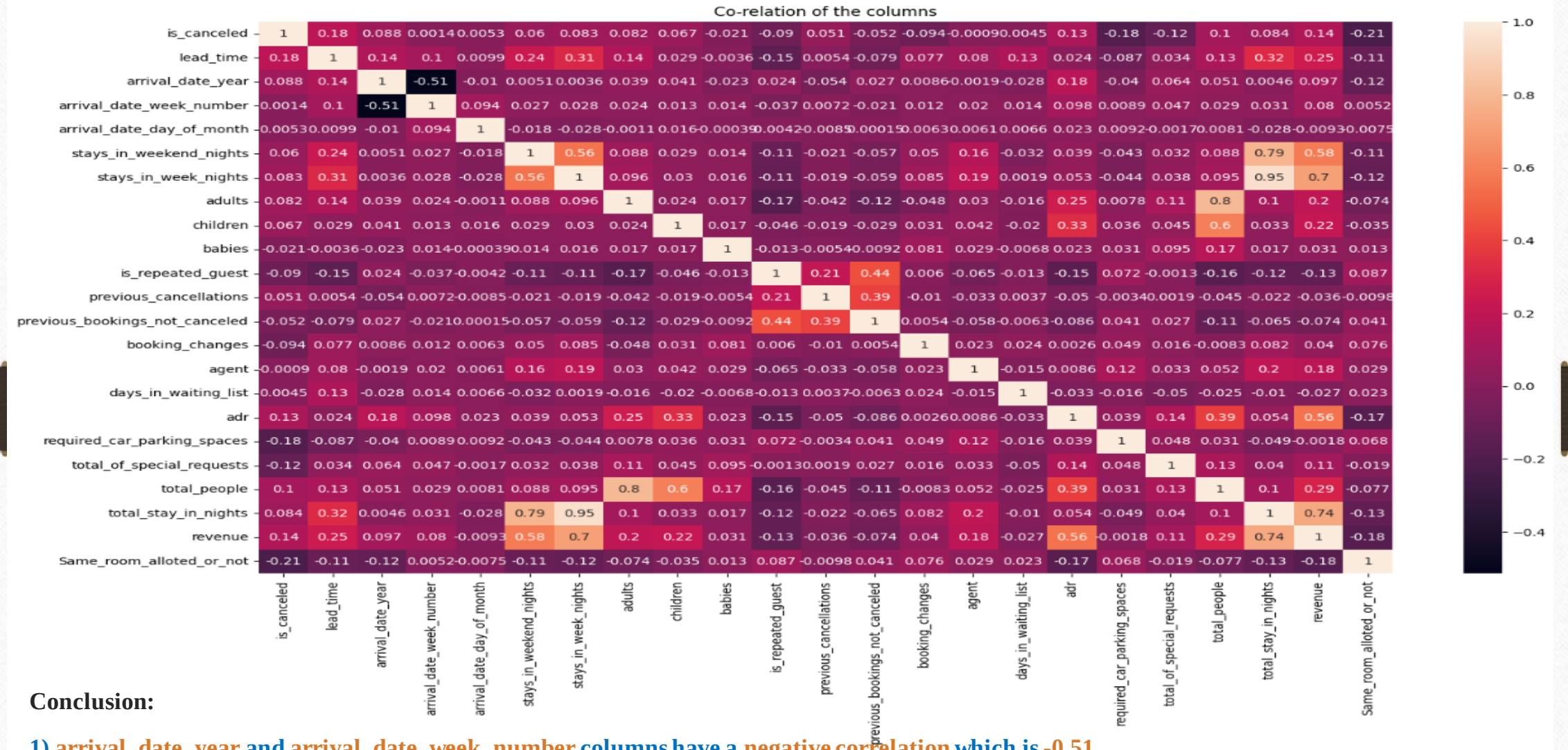


**Conclusion:**

**1) Repeated guests do not cancel their previous bookings but non-repeated/unique guests cancel their bookings.**

**2) If the number of people increases ADR is increasing ,due to this revenue also increases.**

# EDA: BIVARIATE AND MULTIVARIATE ANALYSIS




**Conclusion:**

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

2) **stays_in_week_nights and total_stay_in_nights columns have 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 hotels are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
2. 27.5 % bookings were got cancelled out of all the bookings.
3. Only 3.9 % people were revisited the hotels. Rest 96.1 % were new guests. Thus, retention rate is very low.
4. The percentage of 0 changes made in the booking was more than 82 %. Percentage of single changes made was about 10%.
5. Most of the customers (91.6%) do not require car parking space.
6. TA/TO (travel agents/Tour operators) distribution channel is mostly used, and the percentage age is 79.13%.
7. BB( Bed & Breakfast) is the most preferred type of meal by the guests.
8. Maximum number of guests were from Portugal, i.e. more than 25,000 guests.
9. Most of the bookings for City hotels and Resort hotel were happened in 2016.
10. Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
11. Booking cancellation rate is high for City hotels which almost 30 %.
12. Average lead time for resort hotel is high.
13. Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.
14. Optimal stay in both the type hotel is less than 7 days. Usually, people stay for a week.
15. Almost 19 % people did not cancel their bookings even after not getting the same room which they reserved while booking hotel. Only 2.5 % people cancelled the booking.
16. 2016 year had the highest number of bookings and bookings were 42,313.
17. arrival_date_year and arrival_date_week_number columns have a negative correlation which is -0.51.
18. stays_in_week_nights and total_stay_in_nights columns have a positive correlation which is 0.95.