# Capstone Project

## Hotel Booking Analysis

ANIL BHATT

# INDEX:

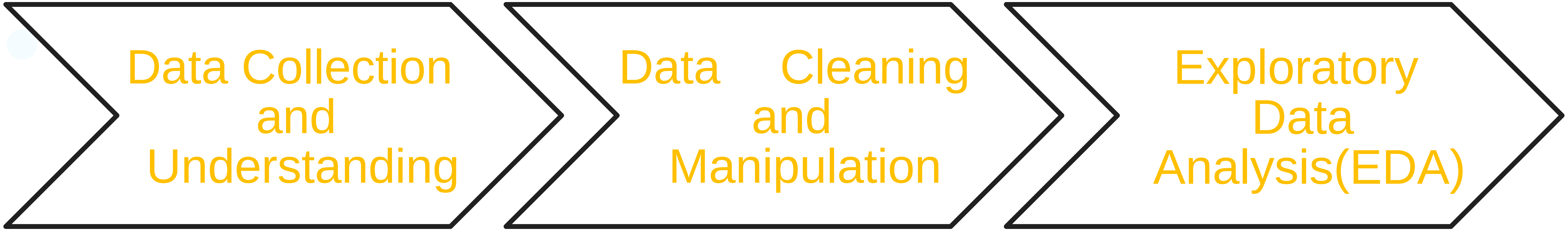
- Topic Discussion
- Work Flow
- Data Collection and Understanding
- Data Cleaning and Manipulation
- Removing outliers
- Correlation heatmap
- Exploratory Data Analysis
    - General Questions
    - Hotel Related Analysis
    - Distribution Wise Analysis
    - Time Related Analysis
- Conclusion

# TOPIC DISCUSSION:

- For this project we will be analyzing Hotel Booking data. 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.
- Hotel industry is a very volatile industry and the bookings depend on the above factors and many more.
- The main objective behind this project is to explore and analyze data to discover important factors that govern the bookings and give insights to hotel management, which can perform various campaigns to boost the business and performance.

# Work Flow :

- we will divide our work flow into following 3 steps.



Data Collection and Understanding

Data Cleaning and Manipulation

Exploratory Data Analysis(EDA)

EDA will be divided into the following 3 analyses. ( We have covered all 3 types in 22 Questions in collab)

1. **Univariate analysis:** Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable.
2. **Bivariate analysis:** Bivariate analysis is where you are comparing two variables to study their relationships.
3. **Multivariate analysis:** Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables.

# Data Collection and Understanding:

- After collecting data it’s very important to understand your data. So we had hotel Booking analysis data.  Which had 119390 rows and 32 columns. So let’s understand these 32 columns.

## Data Description:

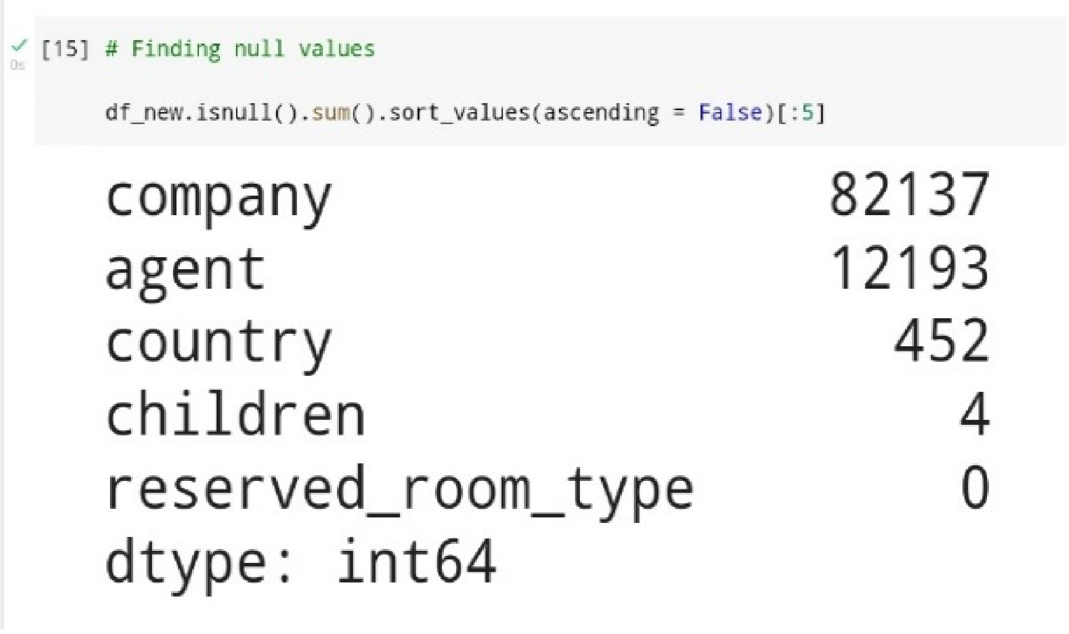
- **hotel** :Resort Hotel or City Hotel
- **is_canceled** : Value indicating if the booking was canceled (1) or not (0)
- **lead_time** : Number of days that elapsed between the entering date of the booking and the arrival date.
- **arrival_date_year** : Year of arrival date.
- **arrival_date_month** : 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.
- **stays_in_week_nights** : Number of weeknights.
- **adults** : Number of adults.
- **children** : Number of children.
- **babies** : Number of babies.
- **meal** : Type of meal booked.
- **country** : Country of origin.

# Data Collection and Understanding:

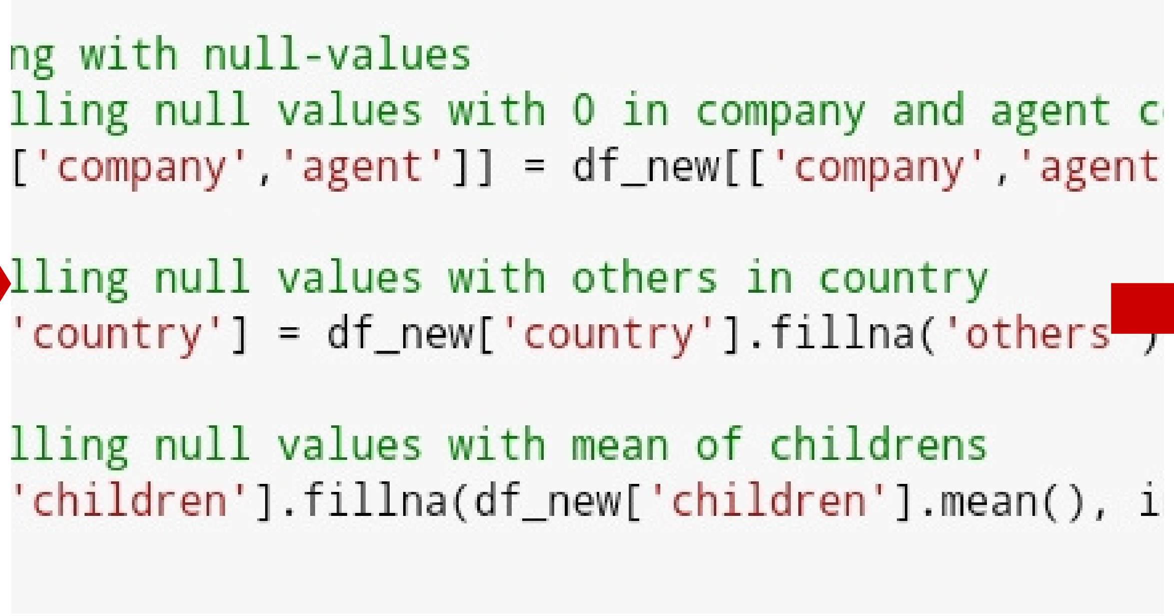
- **market_segment** : Market segment designation. (TA/TO)
- **distribution_channel** : Booking distribution channel.(T/A/TO)
- **is_repeated_guest** : is a repeated guest (1) or not (0)
- **previous_cancellations** : Number of previous bookings that were canceled by the customer prior to the current 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 for the type of room assigned to the booking.
- **booking_changes** : Number of changes made to the booking from the moment the booking was entered on the
- PMS until the moment of check-in or cancellation
- **deposit_type** : No Deposit, Non-Refund, Refundable.
- **agent** : ID of the travel agency that made the booking
- **company** : ID of the company/entity that made the booking.
- **days_in_waiting_list** : Number of days the booking was on the waiting list before it was confirmed to the customer
- **customer_type** : type of customer. Contract,Group,transient,Transient party.
- **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.

# Data Cleaning and Manipulation:

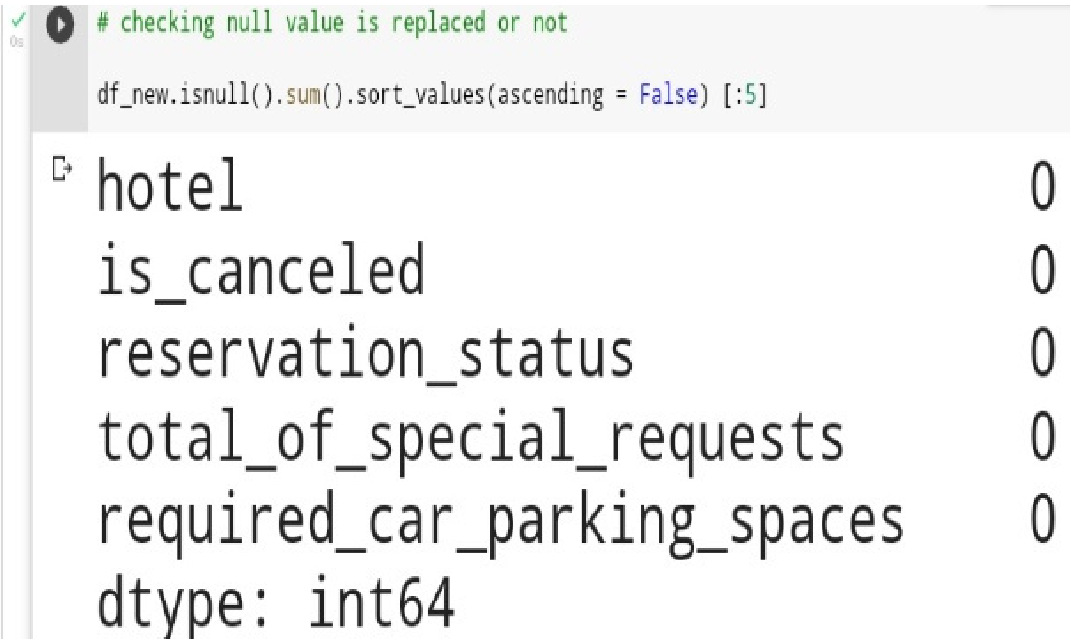
- There were 4 columns company, agent, country, and children with null values changing the value.



```
[15] # Finding null values

     df_new.isnull().sum().sort_values(ascending = False)[:5]

company              82137
agent                12193
country                452
children                 4
reserved_room_type       0
dtype: int64
```

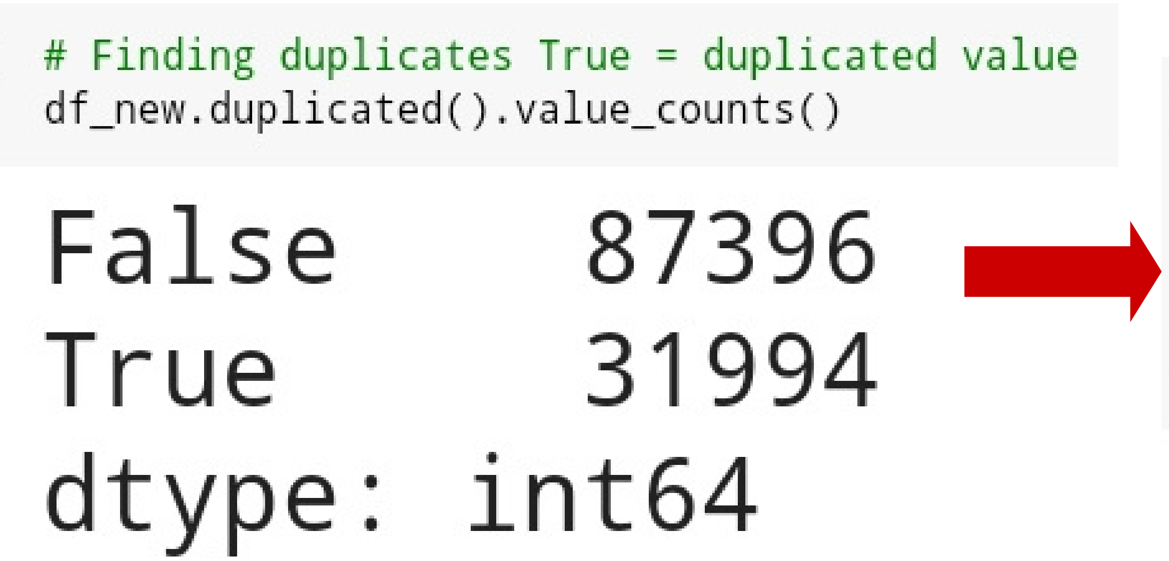


```
ng with null-values
lling null values with 0 in company and agent c
['company','agent']] = df_new[['company','agent

lling null values with others in country
'country'] = df_new['country'].fillna('others')

lling null values with mean of childrens
'children'].fillna(df_new['children'].mean(), i
```

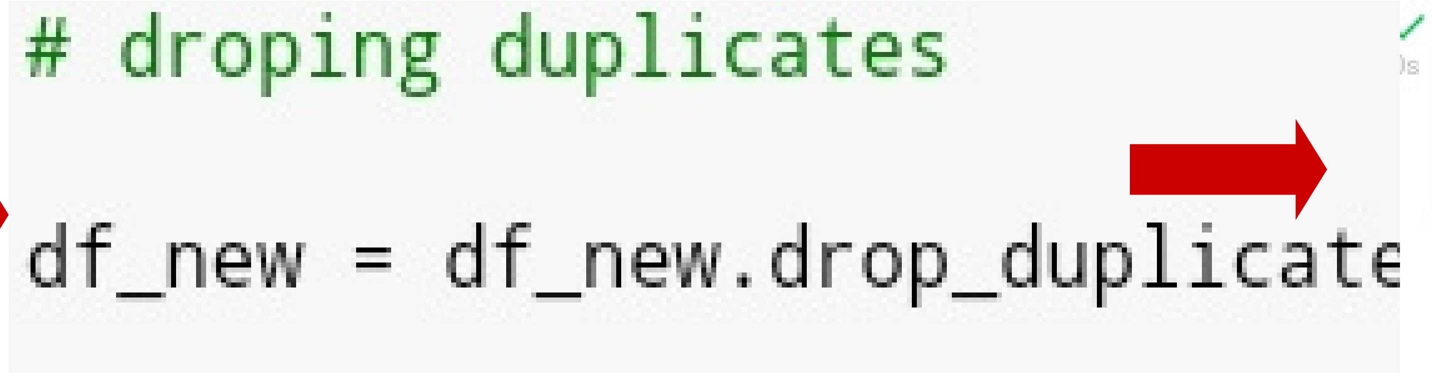


```
# checking null value is replaced or not

df_new.isnull().sum().sort_values(ascending = False) [:5]

hotel                          0
is_canceled                    0
reservation_status             0
total_of_special_requests      0
required_car_parking_spaces    0
dtype: int64
```

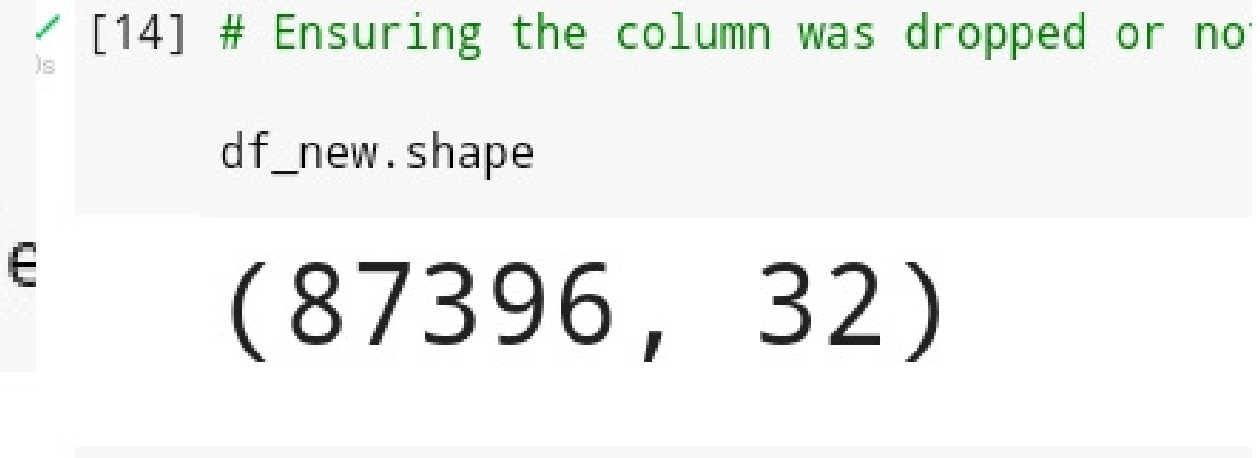
- Handling Duplicates: Data had 31994 duplicate values. So we dropped it from the data.



```
# Finding duplicates True = duplicated value
df_new.duplicated().value_counts()

False    87396
True     31994
dtype: int64
```



```
# droping duplicates

df_new = df_new.drop_duplicate
```

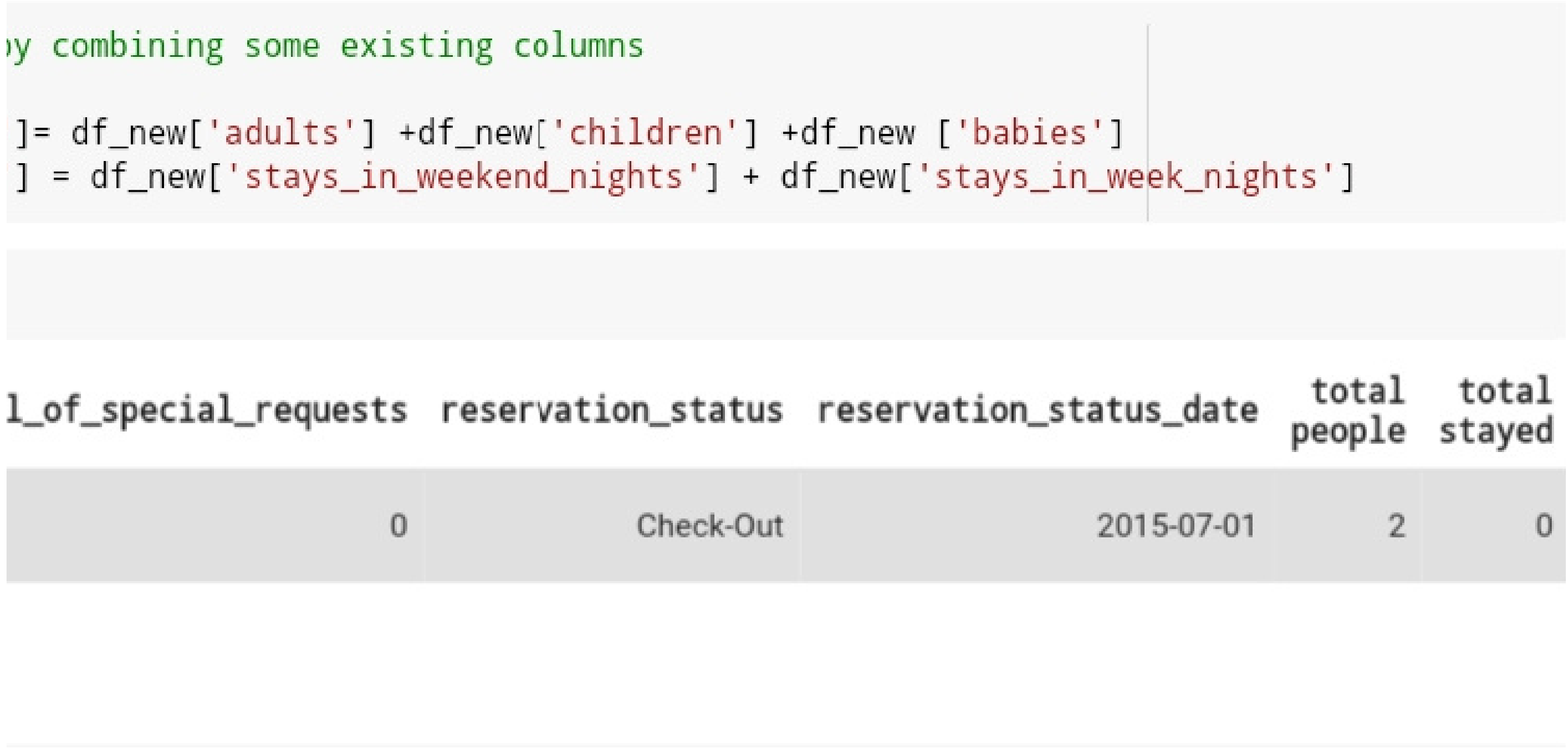


```
[14] # Ensuring the column was dropped or no

     df_new.shape

(87396, 32)
```

# Data Cleaning and Manipulation:

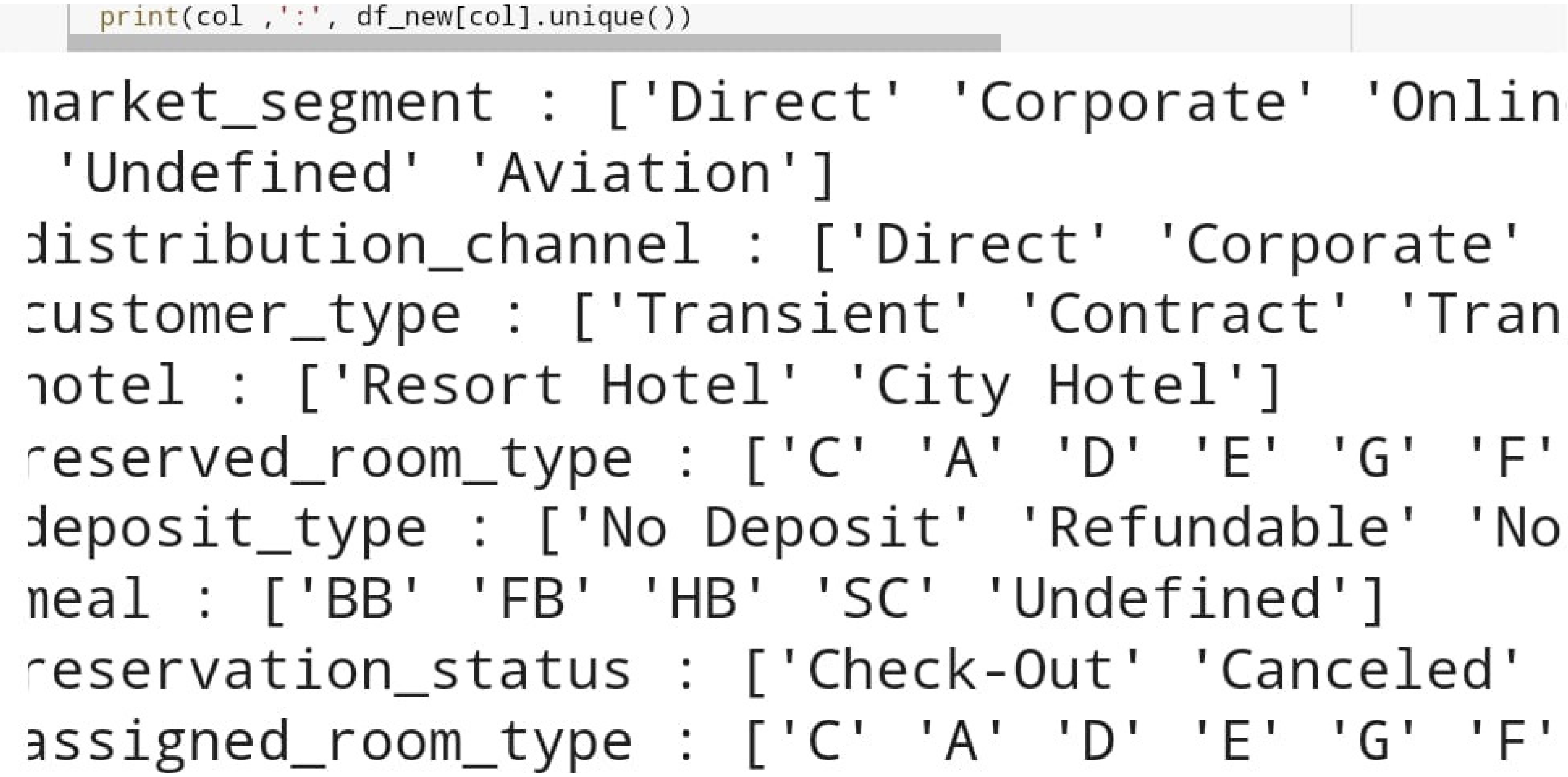
- We created 2 new columns
  1)‘Total People’ = from the Children, adults, and babies.
  2) ‘Total stayed’ = From weekend nights and weekdays night

- We saved columns containing object data type in a variable named categorical_cols  by subtracting numerical columns I.e. set(df.describe()) on that we used For loop to iterate these values  and we get unique values



```
y combining some existing columns

]= df_new['adults'] +df_new['children'] +df_new ['babies']
] = df_new['stays_in_weekend_nights'] + df_new['stays_in_week_nights']
```

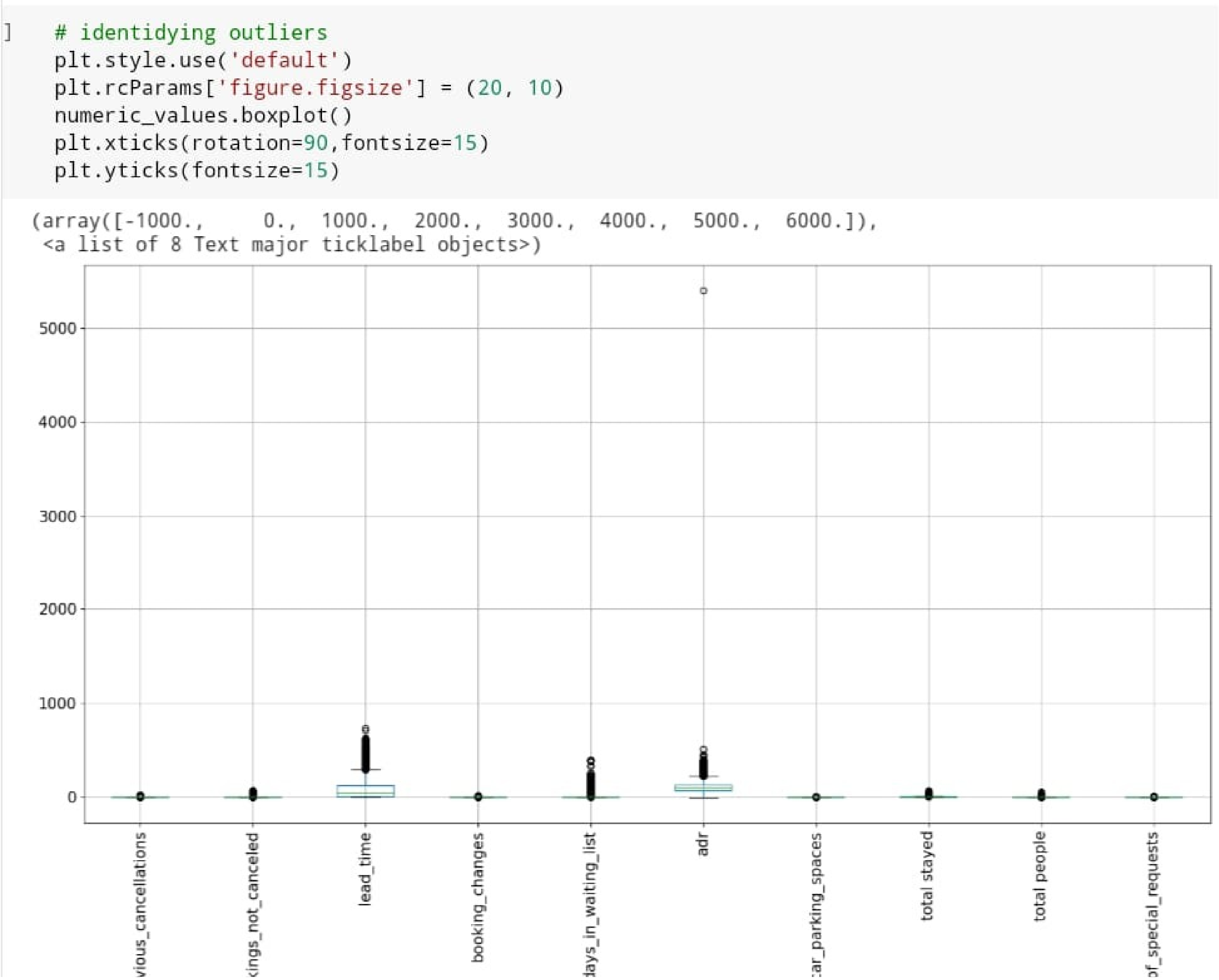
| l_of_special_requests | reservation_status | reservation_status_date | total people | total stayed |
|---|---|---|---|---|
| 0 | Check-Out | 2015-07-01 | 2 | 0 |



```
print(col ,':', df_new[col].unique())
```

```
narket_segment : ['Direct' 'Corporate' 'Onlin
 'Undefined' 'Aviation']
distribution_channel : ['Direct' 'Corporate'
customer_type : ['Transient' 'Contract' 'Tran
notel : ['Resort Hotel' 'City Hotel']
reserved_room_type : ['C' 'A' 'D' 'E' 'G' 'F'
deposit_type : ['No Deposit' 'Refundable' 'No
neal : ['BB' 'FB' 'HB' 'SC' 'Undefined']
reservation_status : ['Check-Out' 'Canceled'
assigned_room_type : ['C' 'A' 'D' 'E' 'G' 'F'
```

# Finding and removing outliers:

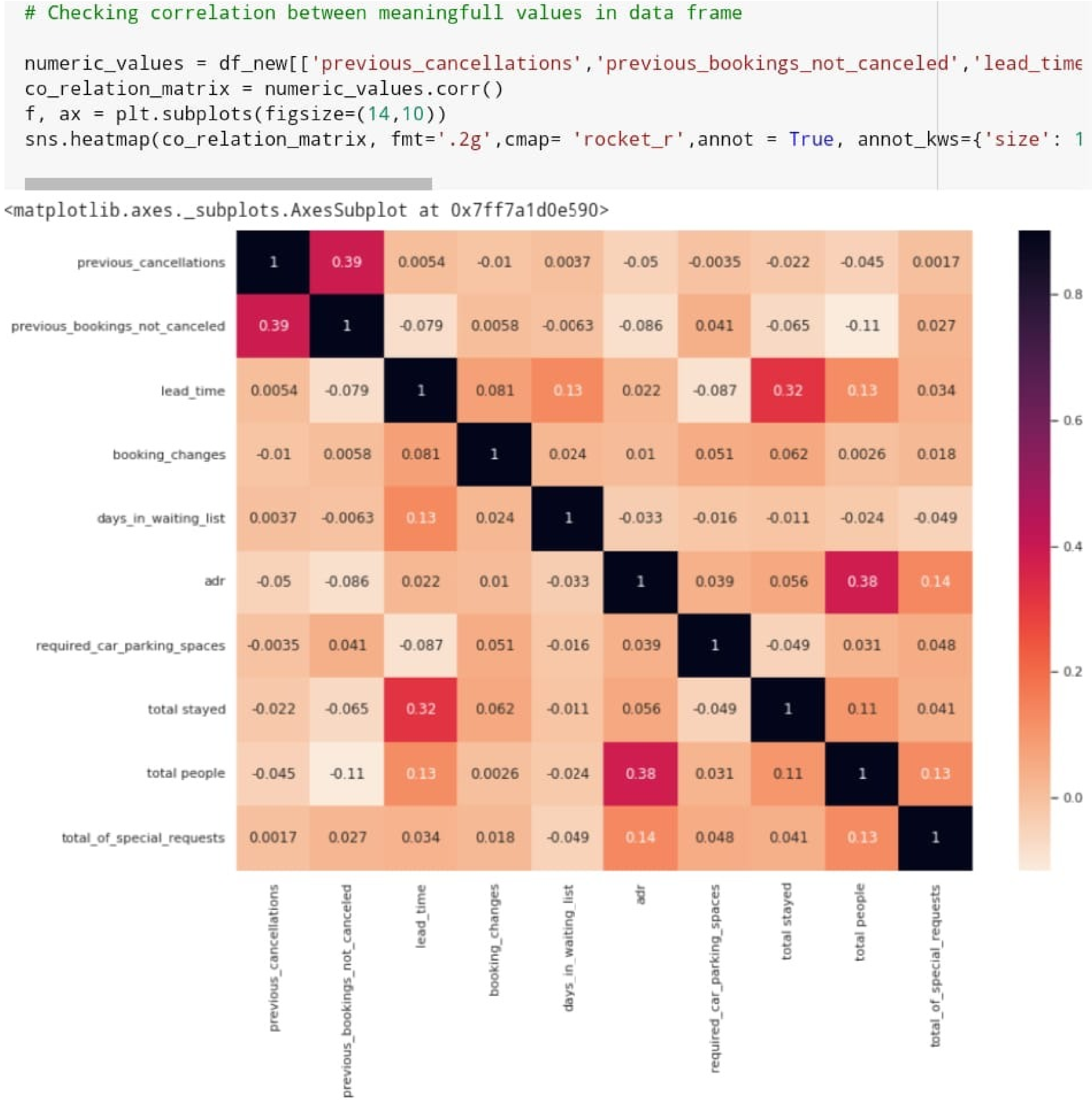


```
# identidying outliers
plt.style.use('default')
plt.rcParams['figure.figsize'] = (20, 10)
numeric_values.boxplot()
plt.xticks(rotation=90,fontsize=15)
plt.yticks(fontsize=15)
```

```
(array([-1000.,     0.,  1000.,  2000.,  3000.,  4000.,  5000.,  6000.]),
 <a list of 8 Text major ticklabel objects>)
```

We found the outlier by applying a box plot to it. Although it is a manual method, In this case, we found only one outlier in a dataset that's why we used the manual method after this we dropped the outlier values but if there were more outliers we can use IQR or Z score method it is more preferred methods and gives accurate outliers.

# Correlation and Heatmap:



```
# Checking correlation between meaningfull values in data frame

numeric_values = df_new[['previous_cancellations','previous_bookings_not_canceled','lead_time
co_relation_matrix = numeric_values.corr()
f, ax = plt.subplots(figsize=(14,10))
sns.heatmap(co_relation_matrix, fmt='.2g',cmap= 'rocket_r',annot = True, annot_kws={'size': 1
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff7a1d0e590>
```

| | previous_cancellations | previous_bookings_not_canceled | lead_time | booking_changes | days_in_waiting_list | adr | required_car_parking_spaces | total stayed | total people | total_of_special_requests |
|---|---|---|---|---|---|---|---|---|---|---|
| previous_cancellations | 1 | 0.39 | 0.0054 | -0.01 | 0.0037 | -0.05 | -0.0035 | -0.022 | -0.045 | 0.0017 |
| previous_bookings_not_canceled | 0.39 | 1 | -0.079 | 0.0058 | -0.0063 | -0.086 | 0.041 | -0.065 | -0.11 | 0.027 |
| lead_time | 0.0054 | -0.079 | 1 | 0.081 | 0.13 | 0.022 | -0.087 | 0.32 | 0.13 | 0.034 |
| booking_changes | -0.01 | 0.0058 | 0.081 | 1 | 0.024 | 0.01 | 0.051 | 0.062 | 0.0026 | 0.018 |
| days_in_waiting_list | 0.0037 | -0.0063 | 0.13 | 0.024 | 1 | -0.033 | -0.016 | -0.011 | -0.024 | -0.049 |
| adr | -0.05 | -0.086 | 0.022 | 0.01 | -0.033 | 1 | 0.039 | 0.056 | 0.38 | 0.14 |
| required_car_parking_spaces | -0.0035 | 0.041 | -0.087 | 0.051 | -0.016 | 0.039 | 1 | -0.049 | 0.031 | 0.048 |
| total stayed | -0.022 | -0.065 | 0.32 | 0.062 | -0.011 | 0.056 | -0.049 | 1 | 0.11 | 0.041 |
| total people | -0.045 | -0.11 | 0.13 | 0.0026 | -0.024 | 0.38 | 0.031 | 0.11 | 1 | 0.13 |
| total_of_special_requests | 0.0017 | 0.027 | 0.034 | 0.018 | -0.049 | 0.14 | 0.048 | 0.041 | 0.13 | 1 |

- For longer duration of stays people generally plan little before the actual arrival.
- Around 13% of total people have made a special request in their booking.
- Barely any transaction has occurred which led to the cancellation afterward.

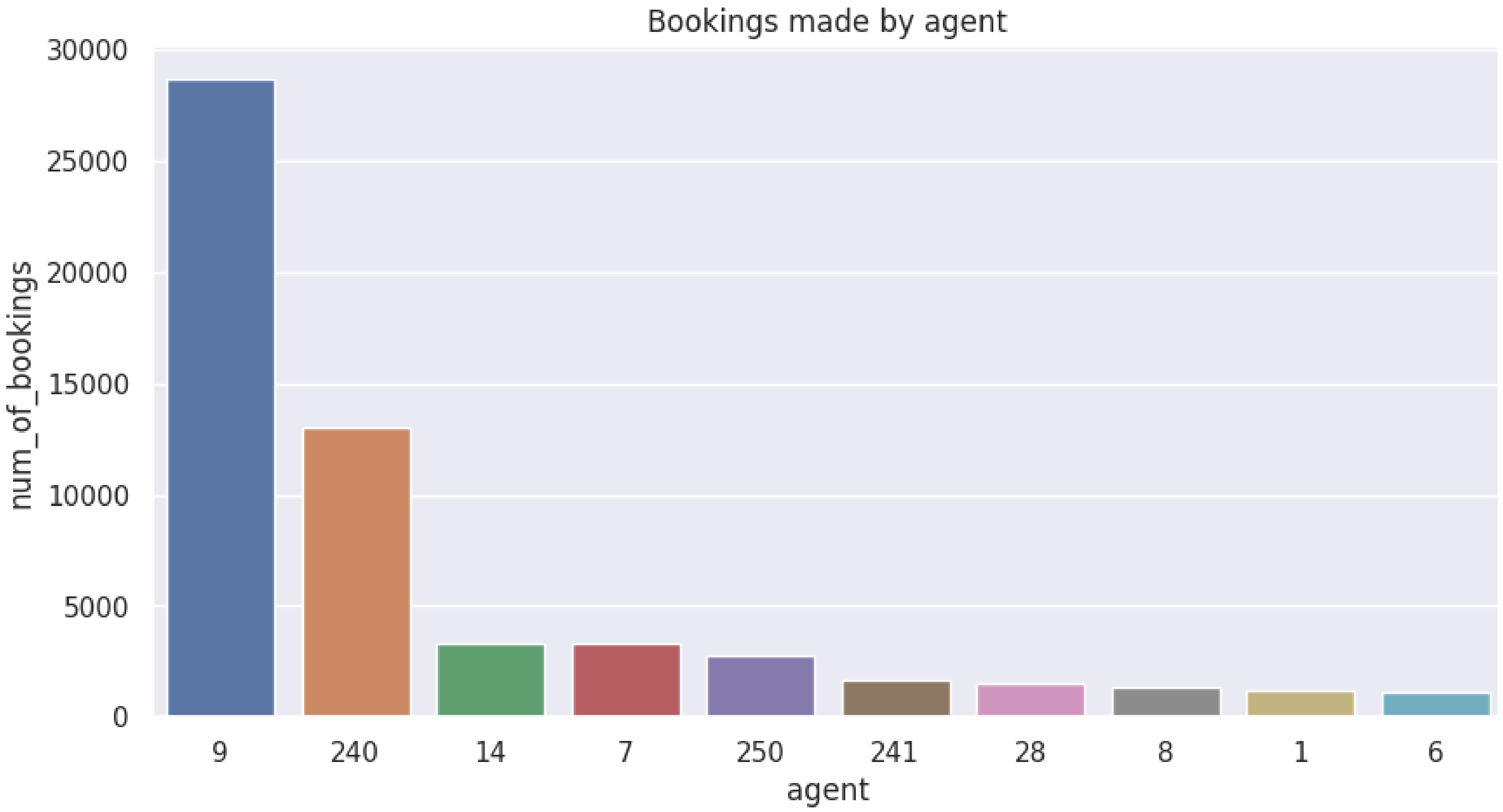
EDA

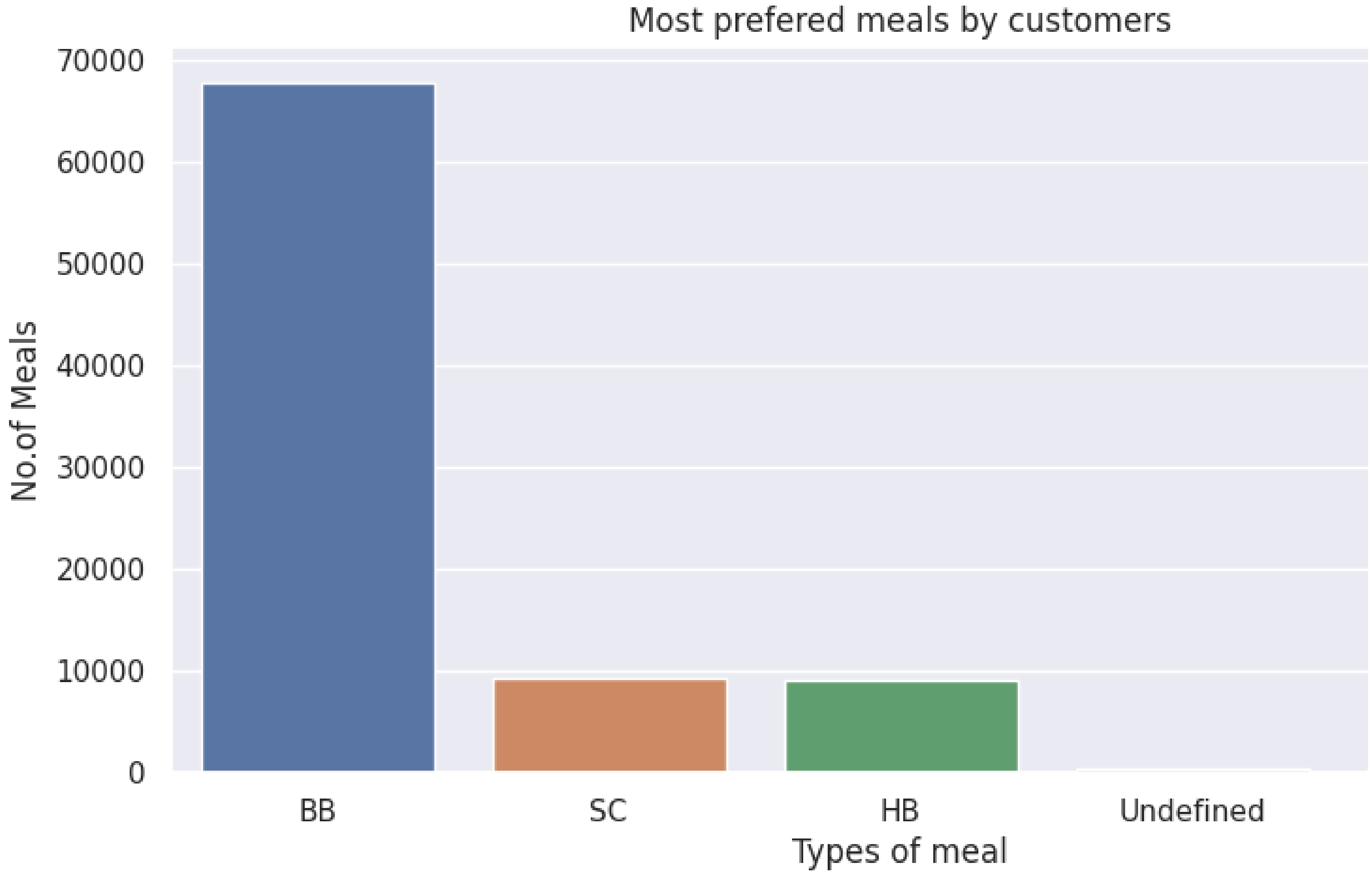
# Data Analysis: General Questions

- Which Agent makes most no. of bookings?
- Most preferred room type by the customers?
- From which country, do guests visit the most?
- Most preferred meal by customers?
- Is there any preference for the ‘Reserved room type’?

**Observation :**

- Agent no.9 makes most of the bookings.





Types of meals:

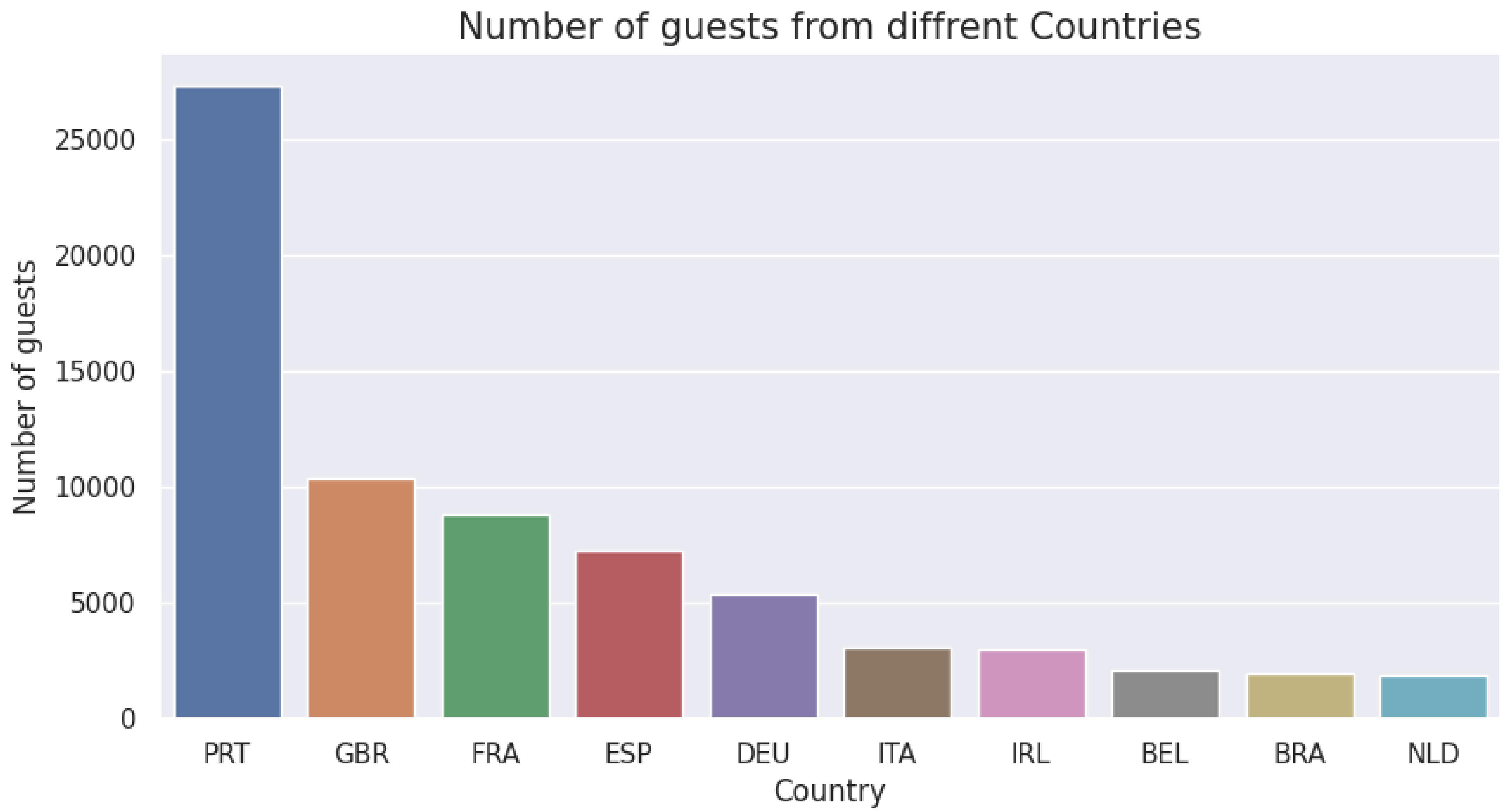
1. BB - (Bed and Breakfast)
2. HB - (Half Bread)
3. FB - (Full Board)
4. SC - (Self Catering)

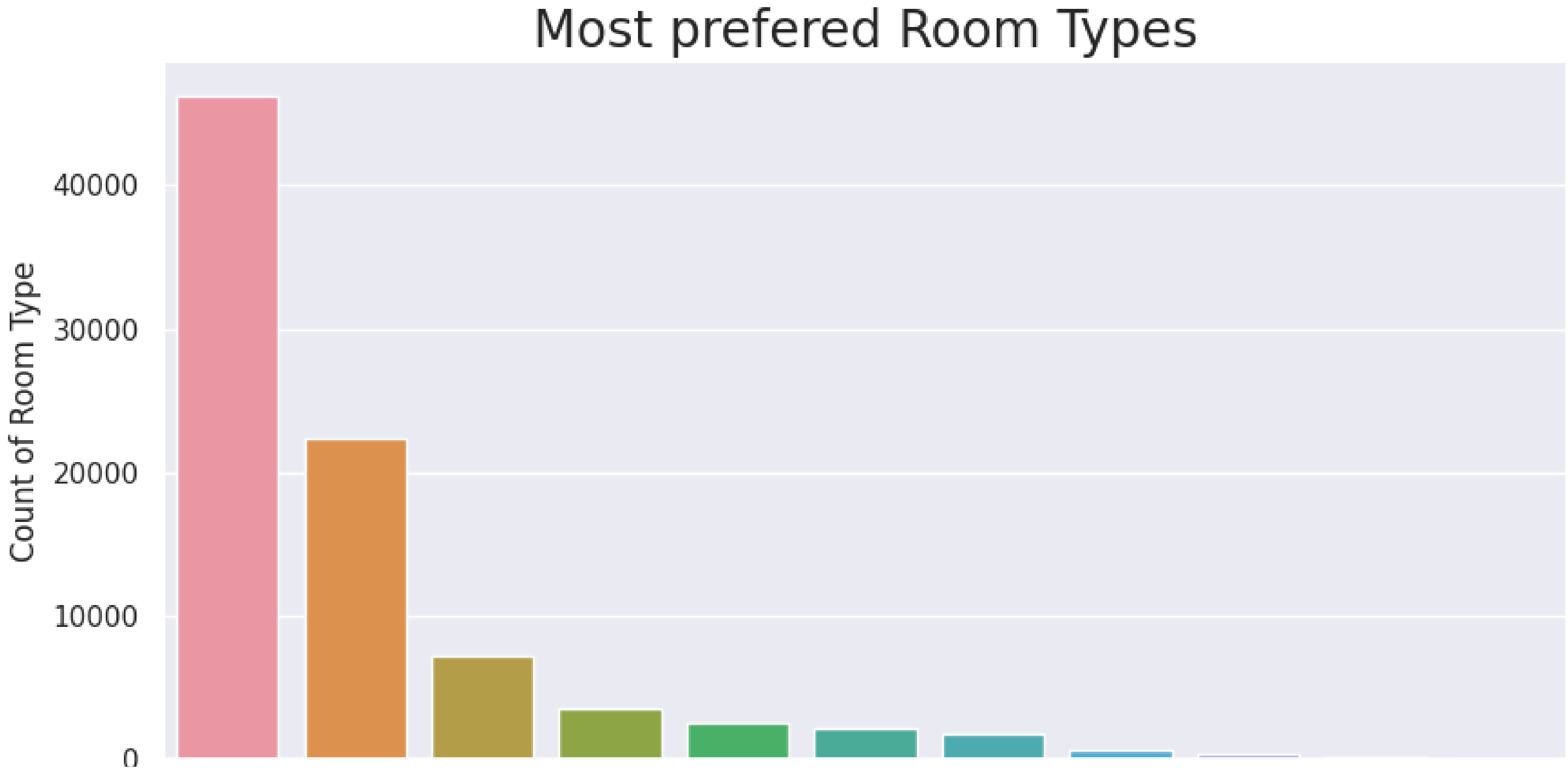
Observation :

The above chart shows most of the customers prefer BB-type Meals.

**Observation :**

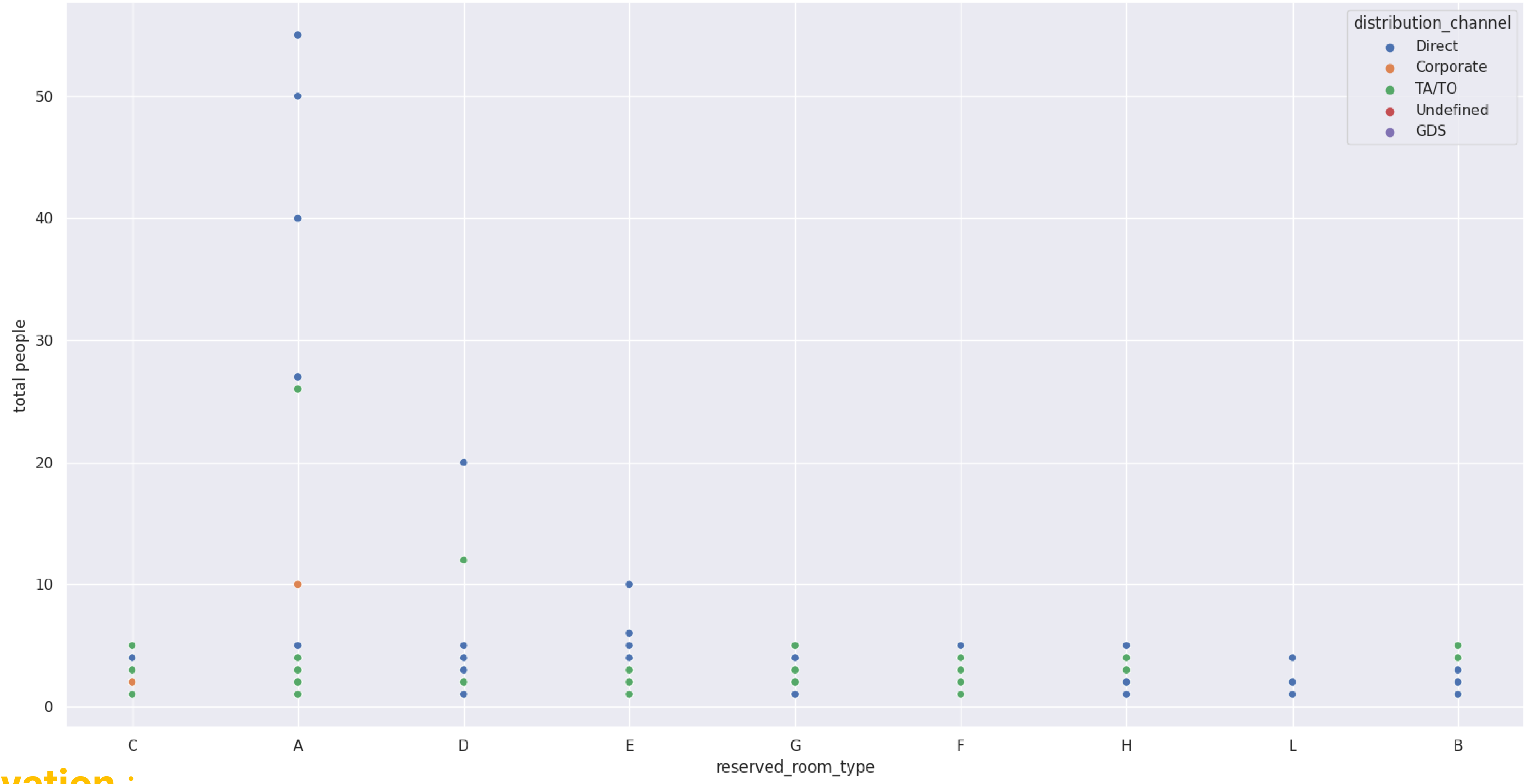
- Most of the guests are coming from Portugal, i.e., more than 25,000 guests are from Portugal.


Number of guests from diffrent Countries


Most prefered Room Types

**Observation :**

- Most preferred room type is A by customers.



## Observation :

- There is a strong propensity for the A, D, and E reserved_room_type.
- Also, for prescriptive analysis, the P category can be evaluated when the hotel gets more years of data.
- Reserved_room_type A and B is the preferred choice for "Corporate".

EDA

# Hotel related analysis:

While doing a hotel-related analysis of a given hotel booking dataset, we answered the following questions:

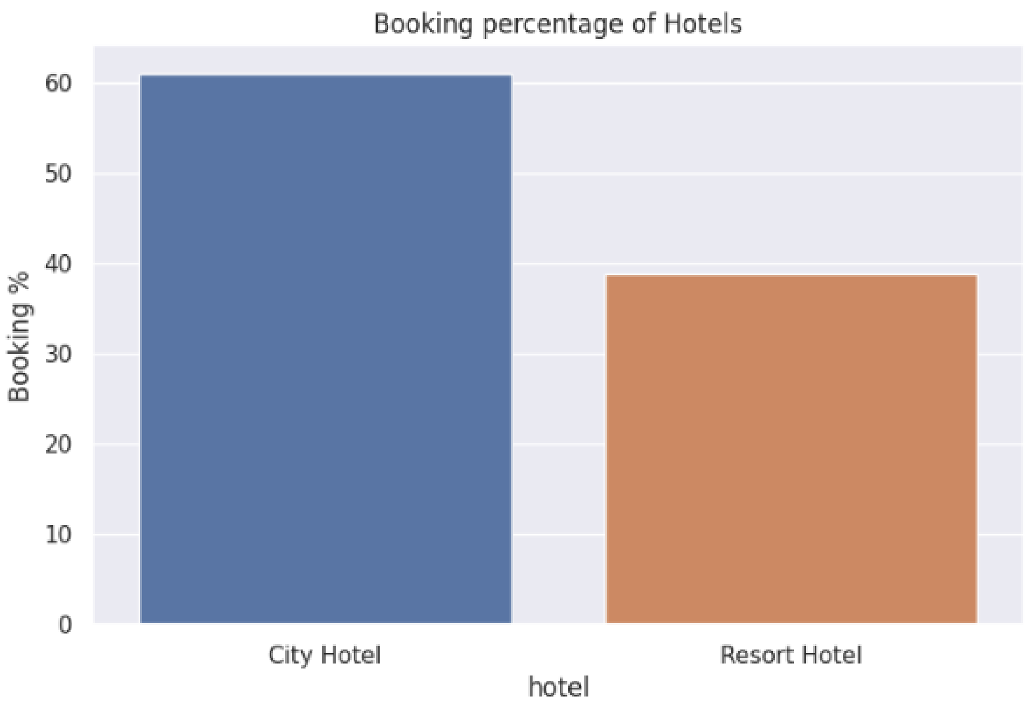
1. Percentage of bookings in each hotel?

2. Avg hotel ADR?

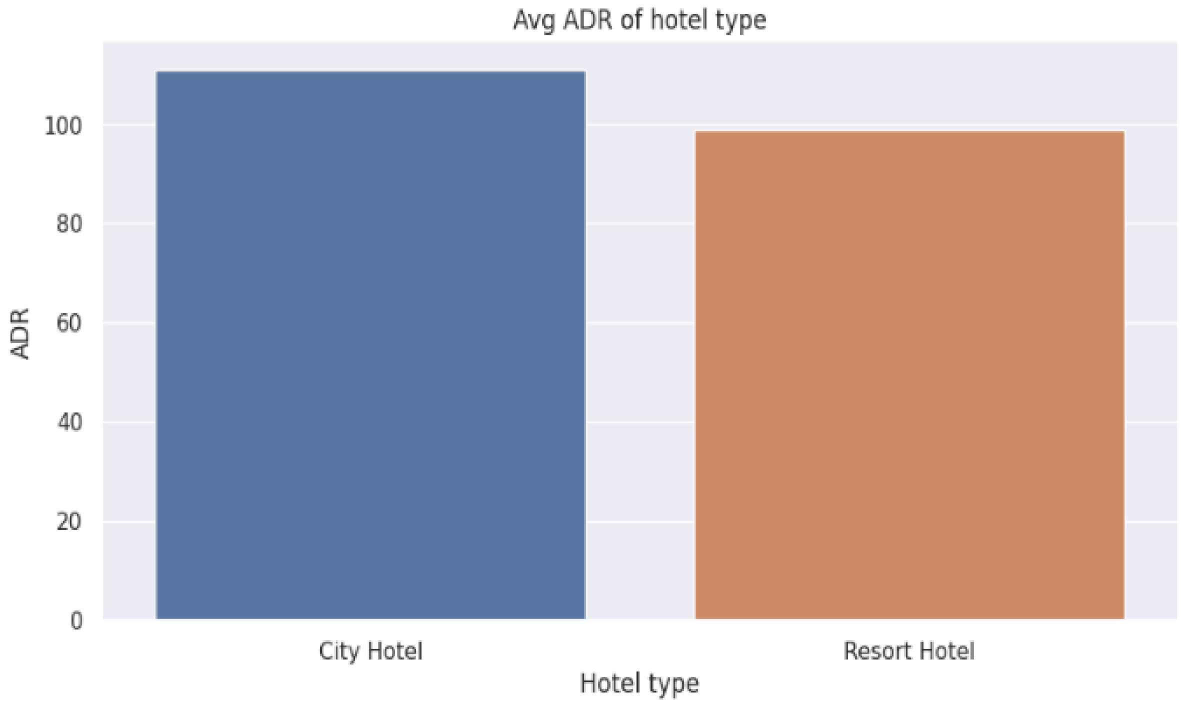
3. Avg lead time by each hotel type?

4. Which Hotel has a high chance that its customer will return for another stay?

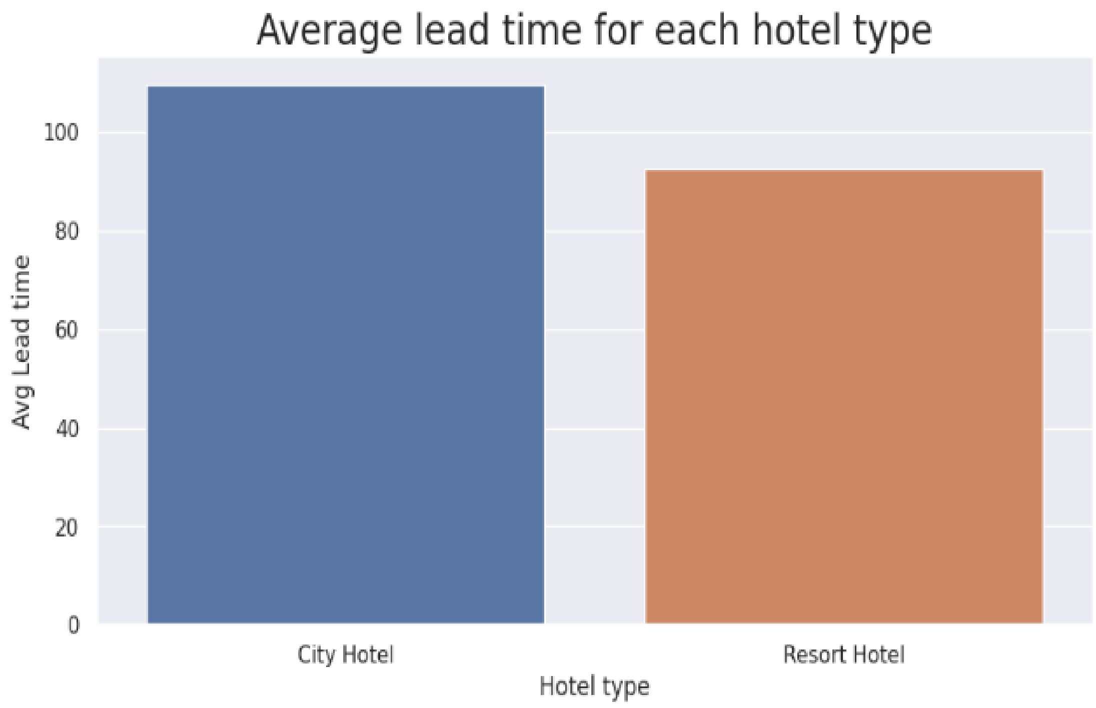
1.



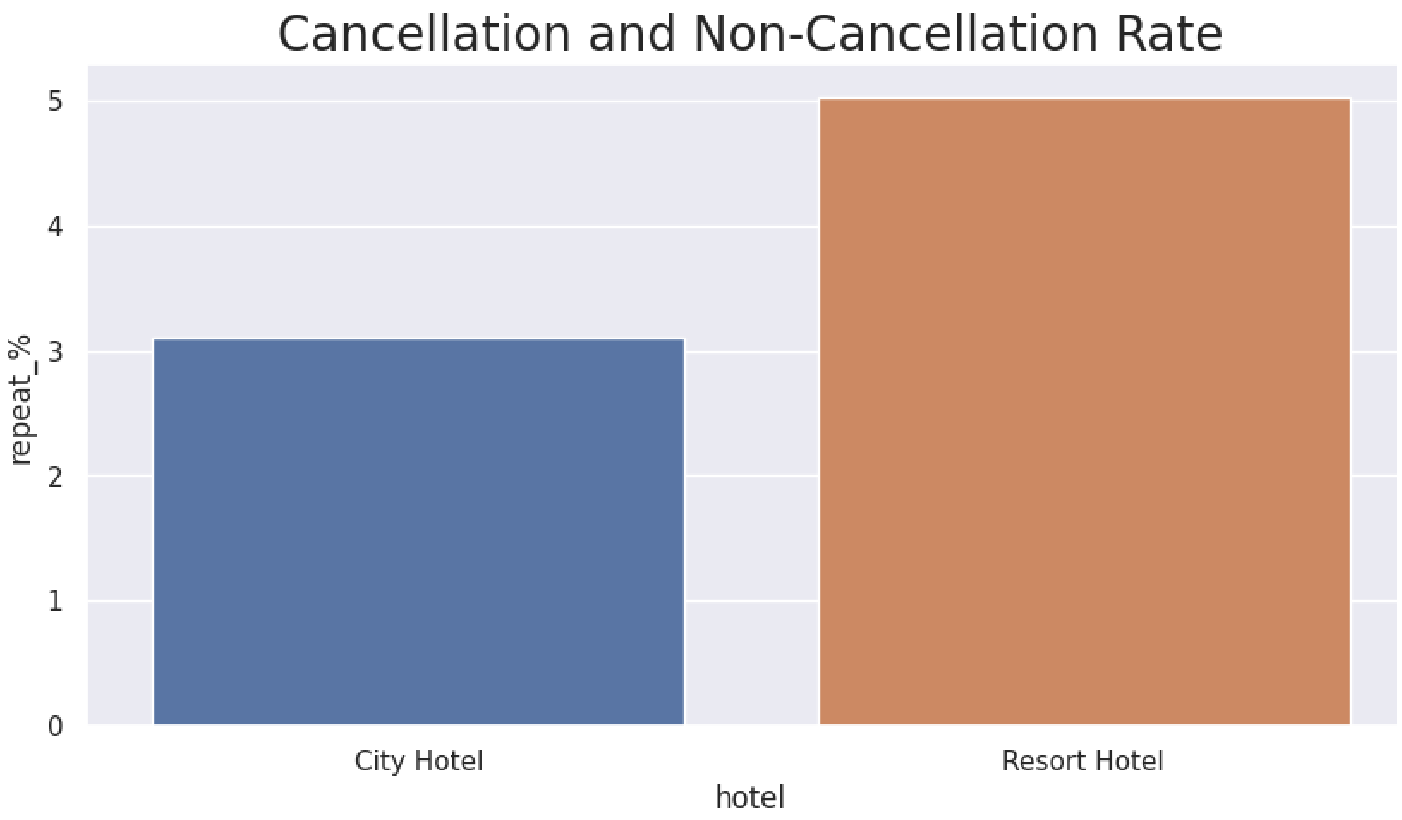
2.



3.



4.



**Observation :**

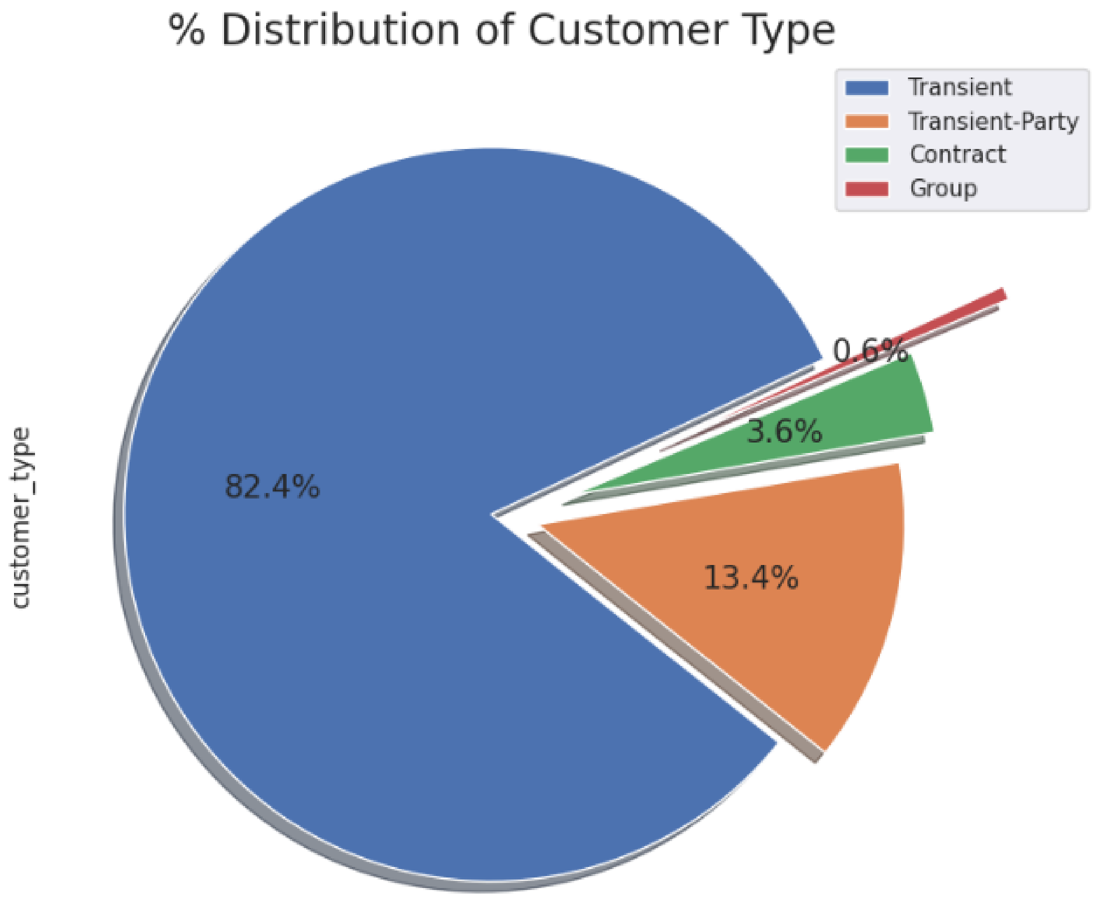
1. Around 60% of people prefer City hotels and 40% of people prefer Resort hotels
2. City Hotel has the highest ADR. This means City Hotels are generating more revenue than the resort hotels. More the ADR the more is the revenue
3. City hotel has more lead time than a resort hotel.
4. Resort Hotel has slightly more chance of its customer revisiting as compared to City Hotel.

# EDA

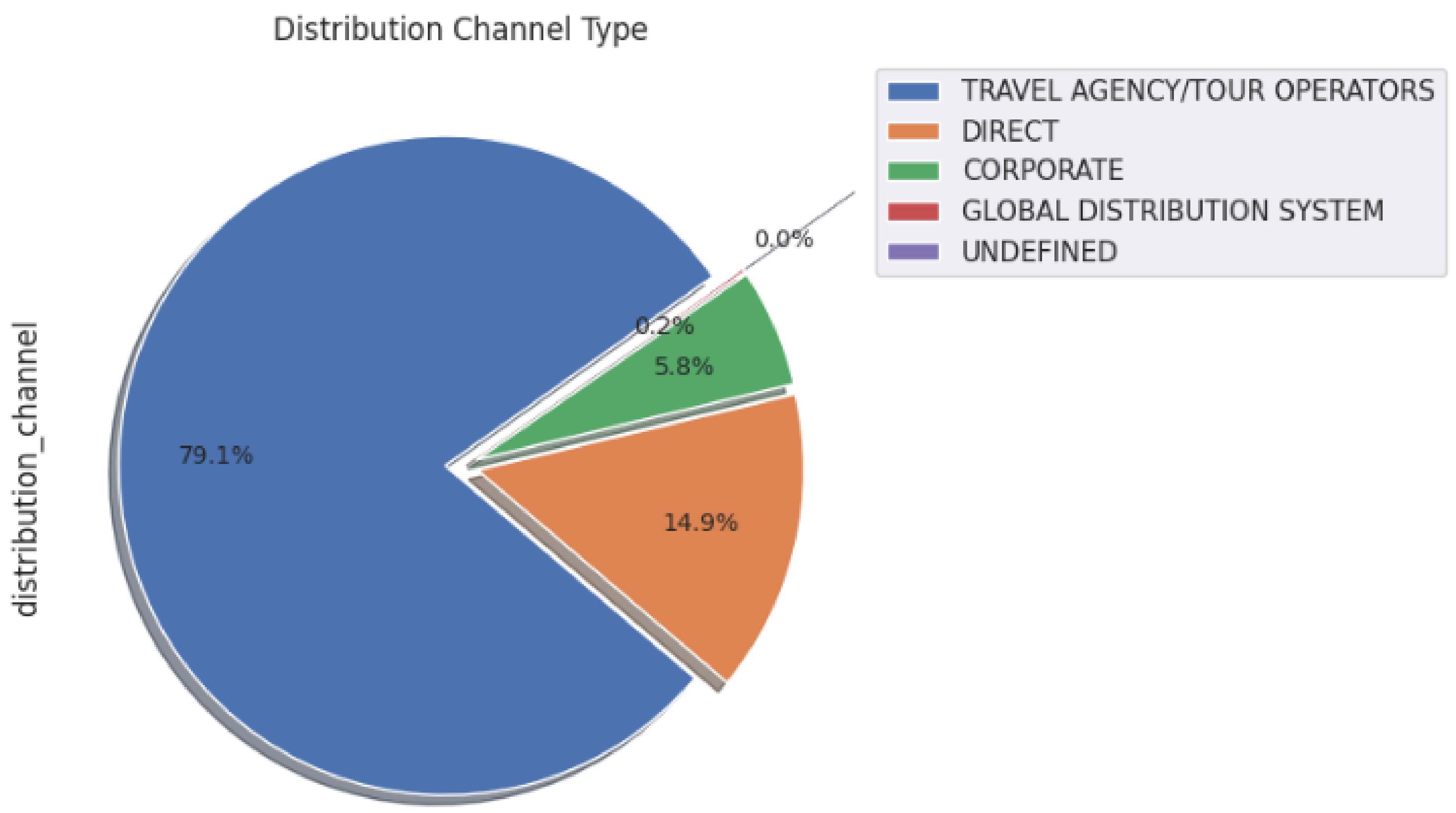
## Distribution-related analysis:

1. What is the percentage distribution of "Customer Type?
2. What is the Percentage distribution of the Deposit type?
3. Which distribution channel is used by the customer for hotel booking?
4. counts of repeated guests?
5. Hotel cancelation rate?
6. Percentage of Booking changes made by the customer?

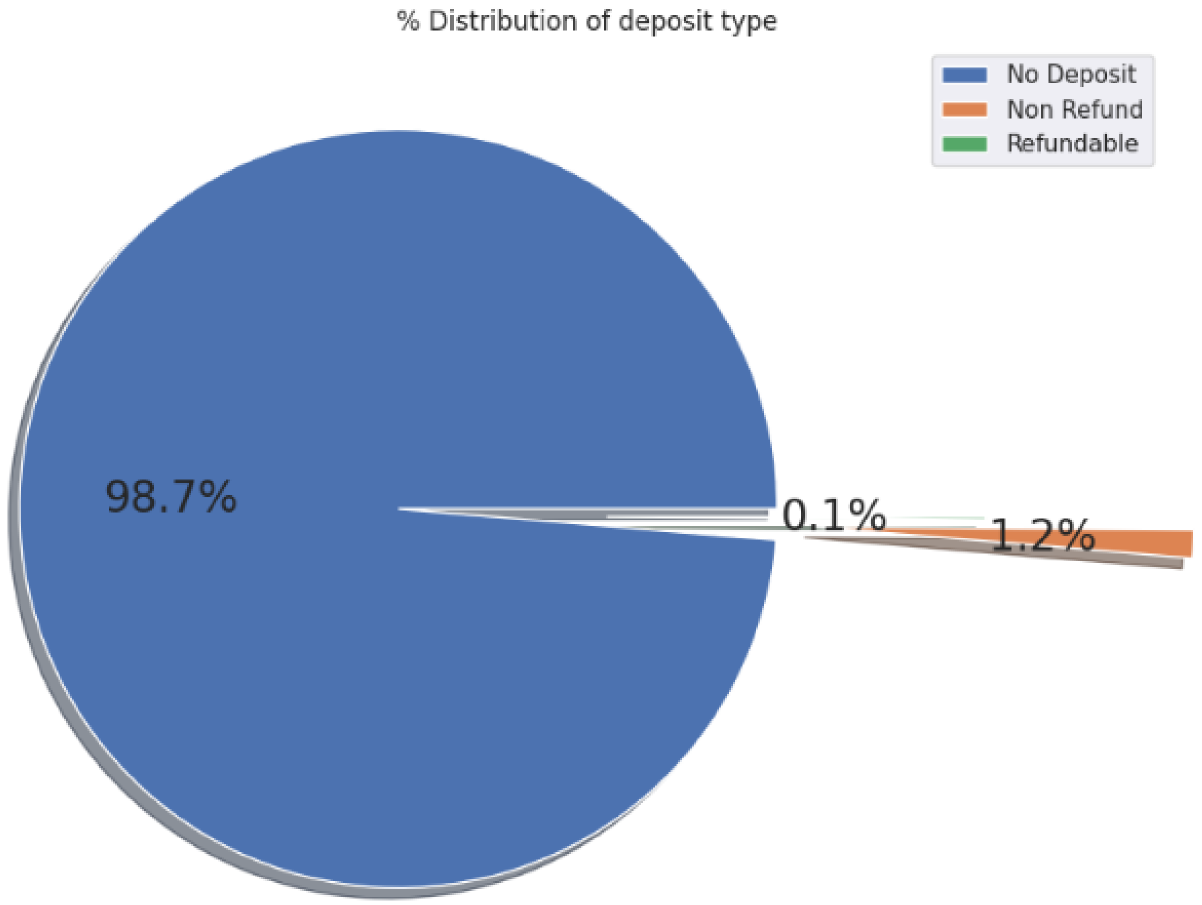
1.



2.



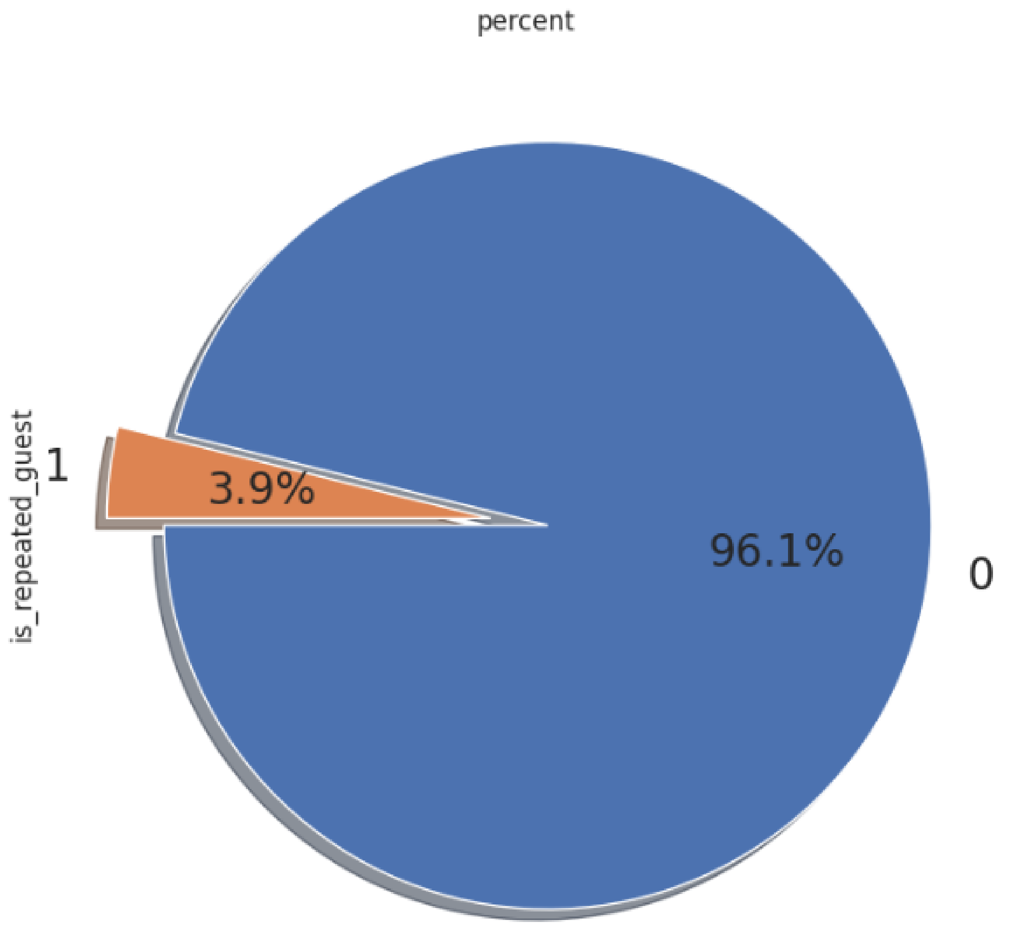
3.



## Observation :

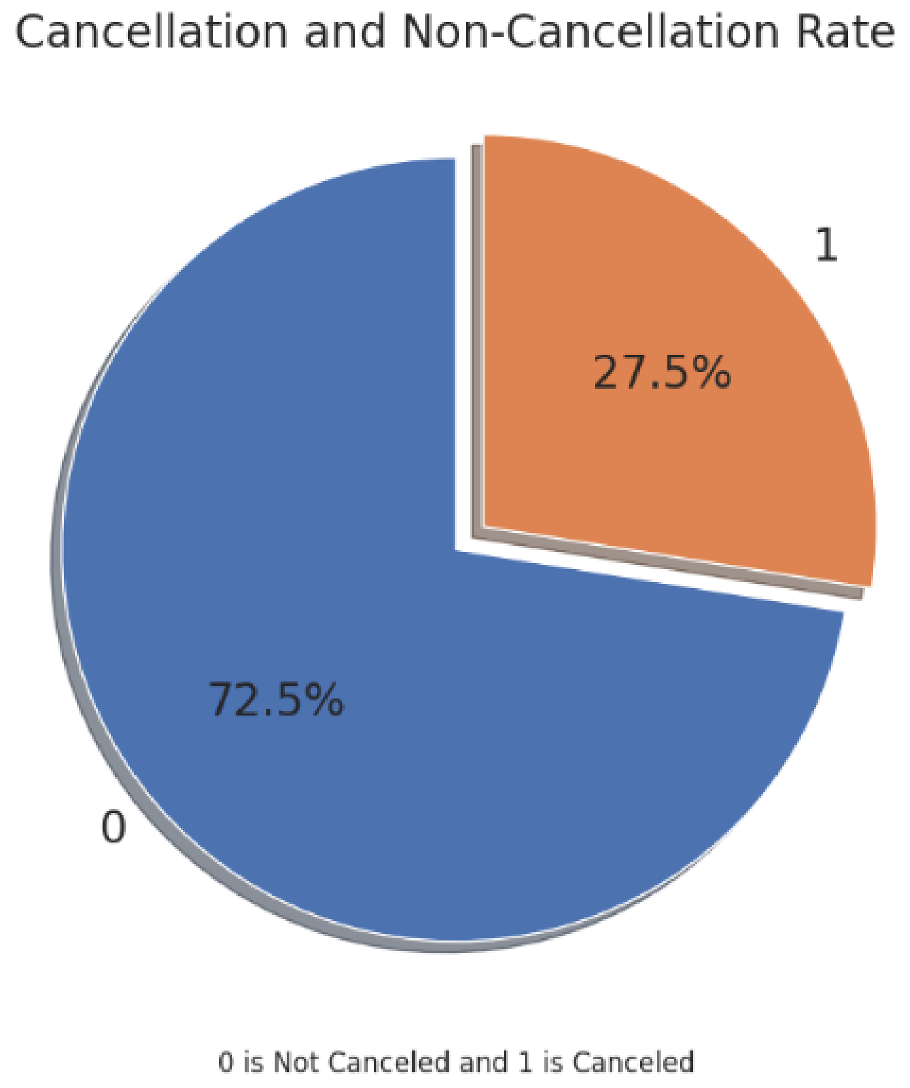
1. Transient customer type is more which is 82.4 %. The percentage of Bookings associated with the Group is very low
2. Channel type distribution
   - 79.1% of customer uses Travel Agency (TA)/Tour Operators(TO).
   - 14.9% of customers direct hotel.
   - 5.8% of customers booking through Corporate.
   - 0.2of % booking comes through the Global Distribution System.
3. 98.7 % of the guests prefer the “No deposit” type of deposit.

4.



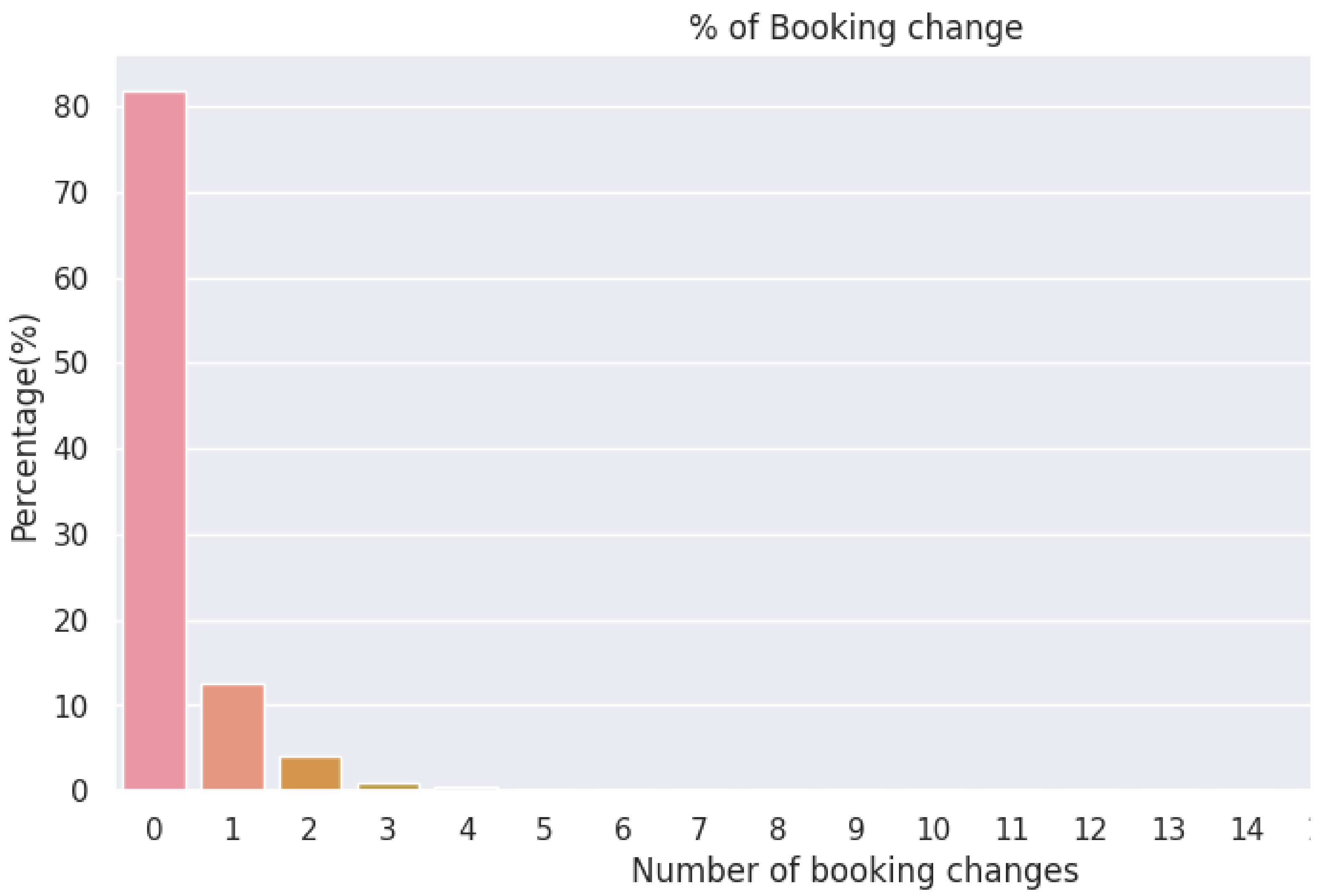
percent

5.



Cancellation and Non-Cancellation Rate

6.



% of Booking change

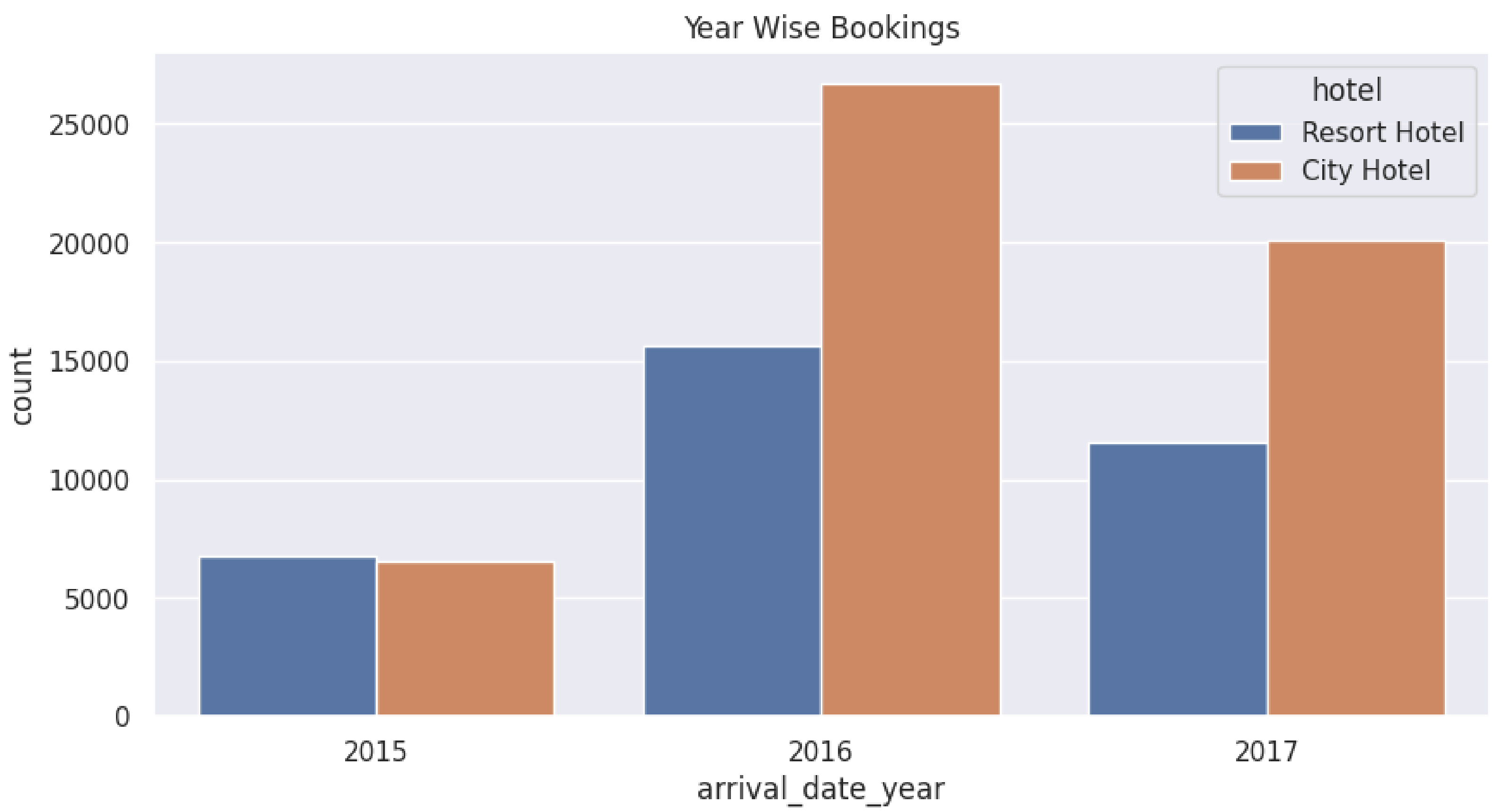
**Observation :**

1. Repeated guests are very few 3.9%.
2. 27.5% of people canceled their booking. We have to find out the reasons why these people canceled their bookings. Also, feedback can be taken from guests who cancel their bookings.
3. Almost 82% of the bookings were not changed by guests

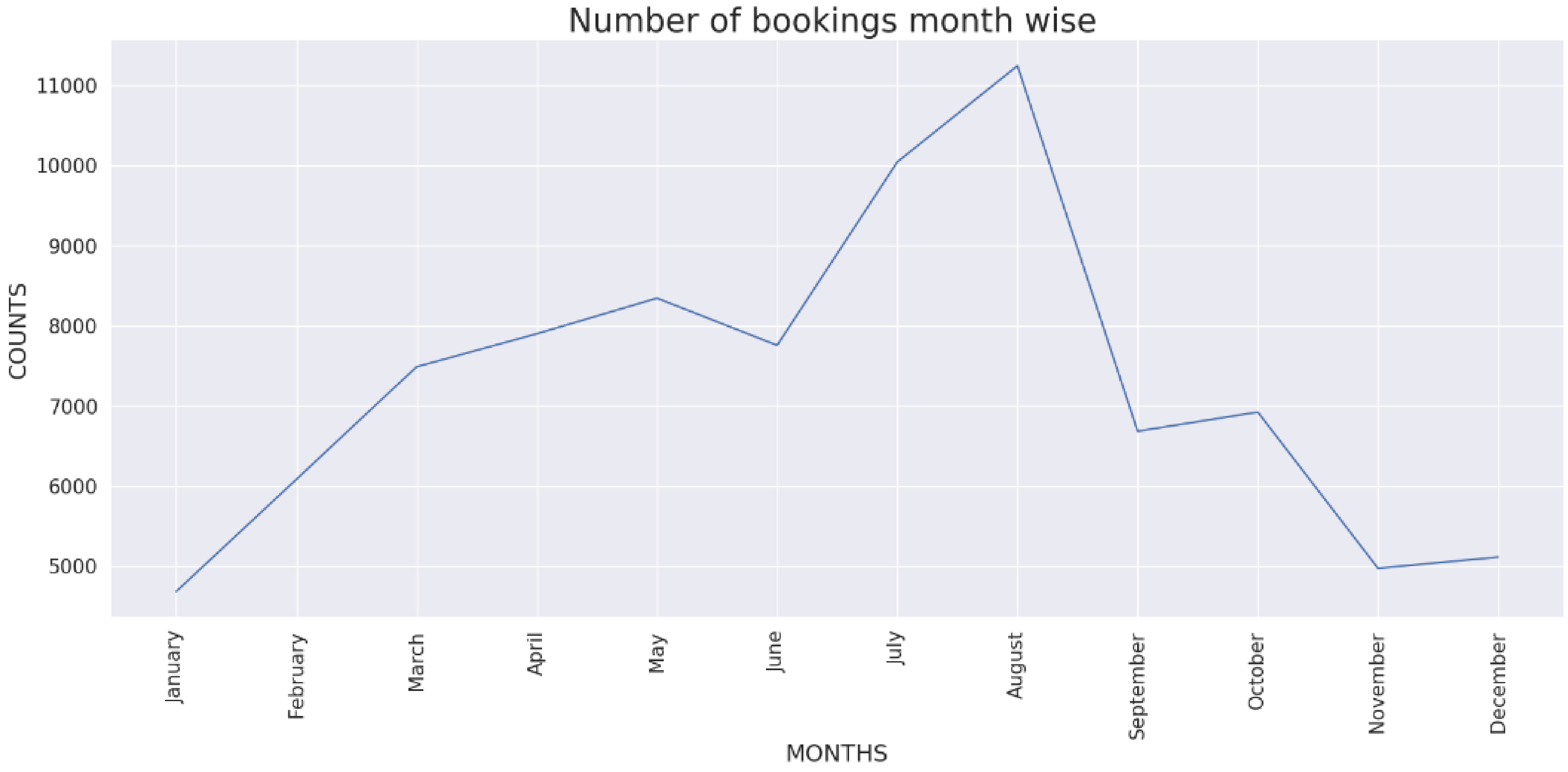
# Time related analysis:

1. Which year had the highest bookings?
2. In which month do most of the bookings happen?
3. ADR across the different months?
4. What is the monthly occupancy of the hotel for the three years of data that we have with us, i.e., for 2015, 2016, and 2017?

1.



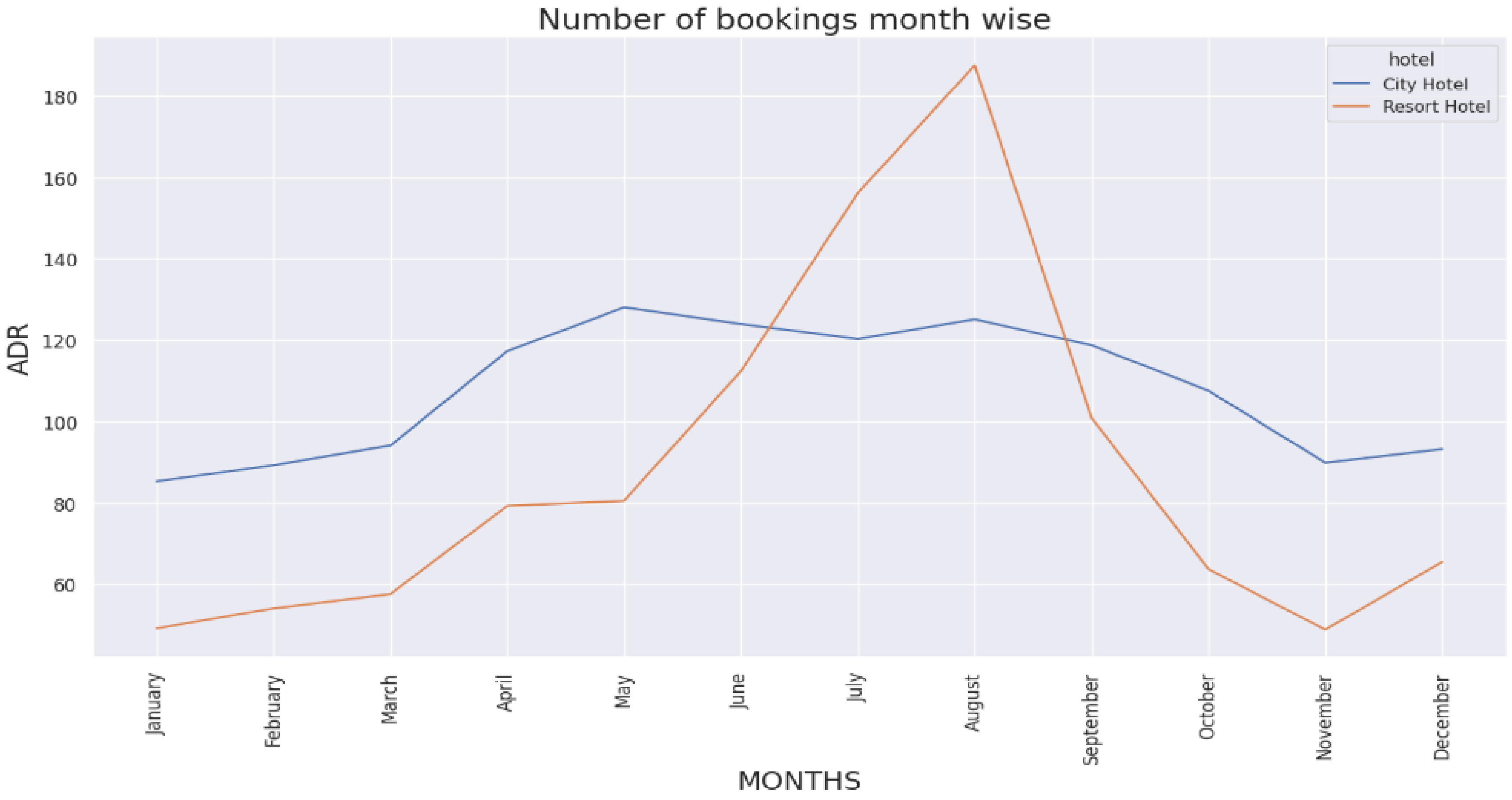
2.



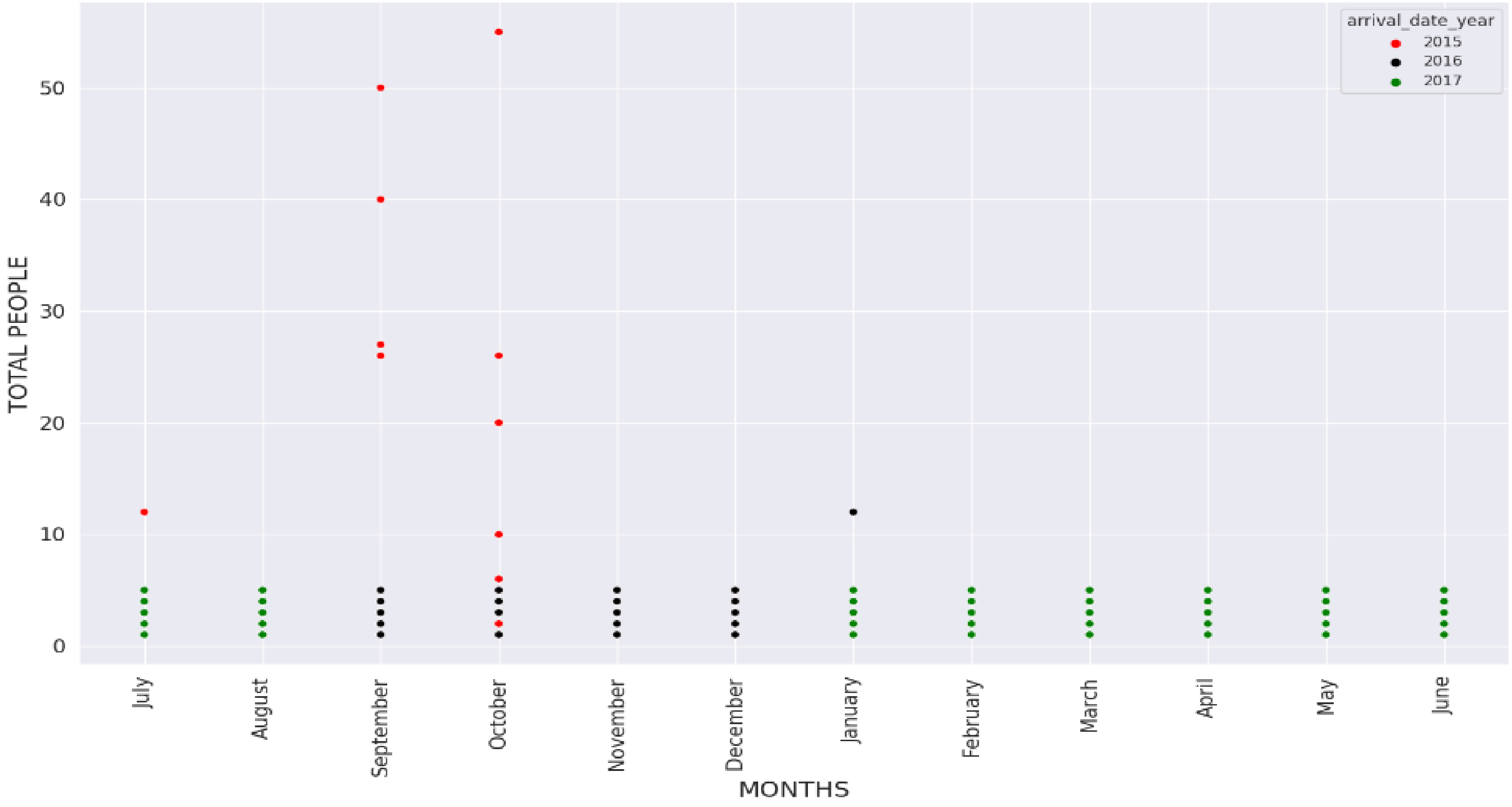
## Observation :

1. Month wise booking count :
   - 2015 had less than 7000 bookings.
   - 2016 had the highest bookings.
   - Overall, City Hotels had the most of the bookings.
2. July and August months had the most Bookings. Summer vacation can be the reason for the bookings.

1.



2.



## Observation :

1. ADR month-wise:
   - For Resort Hotel, ADR is high in June, July, and August as compared to City Hotels. Maybe Customers/People want to spend their Summer vacation in Resorts Hotels.
   - The best time for guests to visit Resort or City hotels is January, February, March, April, October, November, and December as the average daily rate in this month is very low.
2. Monthly occupancy of the hotel for the three years
   - For 2015, a maximum number of occupancy was there in September and October (2 months).
   - Further for 2016, a maximum number of occupancy was there in September, October, November, and December (4 months), which was better than the previous year, 2016.
   - Then for the year 2017, a good number of occupancy was there in the hotel from January to August. But for the months from Sept to Dec, there was a decline in occupancy of the Hotel.

# Conclusion

Based on the above observation, we concluded that this capstone project on “Hotel Booking Analysis”

- It helps a hotel in upscaling their business by improving their techniques, technologies, and many small requirements which guests required during their stay in this hotel.
- It also helps people to understand how things work in real life and the need & importance of data science in the real world.
- It helps increase in efficiency and effectiveness of various methods or processes involved in the hotel booking system.
- It increases the competition and productivity among various hotels to provide a better facility and comfort that the customer needed.

# Thank You