CAPSTONE PROJECT-01
EXPLORATORY DATA ANALYSIS

HOTEL BOOKING ANALYSIS

By: Mohammed Ajmal Sabir M A
(Cohort Canberra)

Agenda:

- Data Analysis Steps
- Data Summary
- Sorting Variable Types
- Data Visualization:
  - Hotel type
  - Type of hotel that most guests preferred
  - Average price per room per day (monthly wise)
  - Highest number of guests in year wise
  - Nationality of guests
  - Distribution by market segment
  - Type of customer made more type of bookings
  - Agent with most number of bookings
  - Deposit type preferred by customer
  - Correlation graph
- Conclusion

Data Analysis Steps:

Import Libraries:

In this part, we had imported the required libraries to perform Exploratory Data Analysis for the hotel booking dataset.

Descriptive Statistics:

In this part, we start by looking at descriptive statistic parameters for the dataset. We will use describe() for this

Missing Value Imputation:

We will now check for missing values in our dataset. In case there are any missing entries, we will impute them with appropriate values

Graphical Representation:

We will do Univariate and Multivariate Analysis .So we use Matplotlib and Seaborn libraries of Python for various graphical representations.

Data Summary:

The summary of findings and understandings about various columns after the data exploration of dataset is as follows:

This data in the dataset is of hotel bookings of the year 2015, 2016 and 2017. it contains 32 columns, the description of the data present in the columns are as follows:

- **hotel** : Type of hotel (City or Resort).
- **is_canceled** : Whether the booking is canceled or not (0 for not canceled and 1 for canceled).
- **lead_time** : Time in days between booking transaction and actual arrival.
- **arrival_date_year** : The year of arrival (2015, 2016, 2017).
- **arrival_date_month** : The month of arrival.
- **arrival_date_week_number** : The week number of the year of arrival date.
- **arrival_date_day_of_month** : The day of month of arrival date.
- **stays_in_weekend_nights** : Number of nights spent in a hotel on weekends.
- **stays_in_week_nights** : Number of nights spent in a hotel on Weekdays.

- **adults** : Number of adults in single booking record.
- **children** : Number of children in single booking record.
- **babies** : Number of babies in single booking record.
- **meal** : Type of meal chosen [BB-Bed and Breakfast, FB- Full Board , HB- Half Board SC-Self Catering]
- **country** : Country of origin of customers.
- **market_segment** : By which segment was booking made and for what purpose [Direct, Corporate, Online TA, Offline TA/TO, Complementary, Groups, Undefined, Aviation].\
- **distribution_channel** : Via which medium booking was made(Direct', Corporate, TA/TO, Undefined, GDS).
- **is_repeated_guest** : Whether the customer has made any booking before(0 for No and 1 for Yes).
- **previous_cancellations** : Number of previous canceled bookings.
- **previous_bookings_not_canceled** : Number of previous non-canceled bookings.
- **reserved_room_type** : Room type reserved by a customer(Room Type=[C, A, D, E, G, F, H, L, P, B]).
- **assigned_room_type** : Room type assigned to the customer (Room Type=[C, A, D, E, G, F, H, L, P, B]).

- **booking_changes** : Number of booking changes done by customers.
- **deposit_type** : Type of deposit at the time of making a booking (No deposit, Refundable, No refund).
- **agent** : Id of agent for booking.
- **company**: Id of the company making a booking.
- **days_in_waiting_list** : Number of days on waiting list.
- **customer_type** : Type of customer(Transient, Contract, Transient-Party, Group).
- **adr** : Average Daily rate=(Rooms revenue/Total rooms sold).
- **required_car_parking_spaces** : Car parking slots (0, 1, 2, 8, 3)
- **total_of_special_requests** : total number of special request.
- **reservation_status** : Whether a customer has checked out or canceled, or not showed.
- **reservation_status_date** : Date of making reservation status.

Sorting Variable Types:

Categorical:

- hotel, arrival_date_month, meal, country, market_segment, distribution_channel, reserved_room_type, assigned_room_type, deposit_type,
- customer_type, reservation_status.

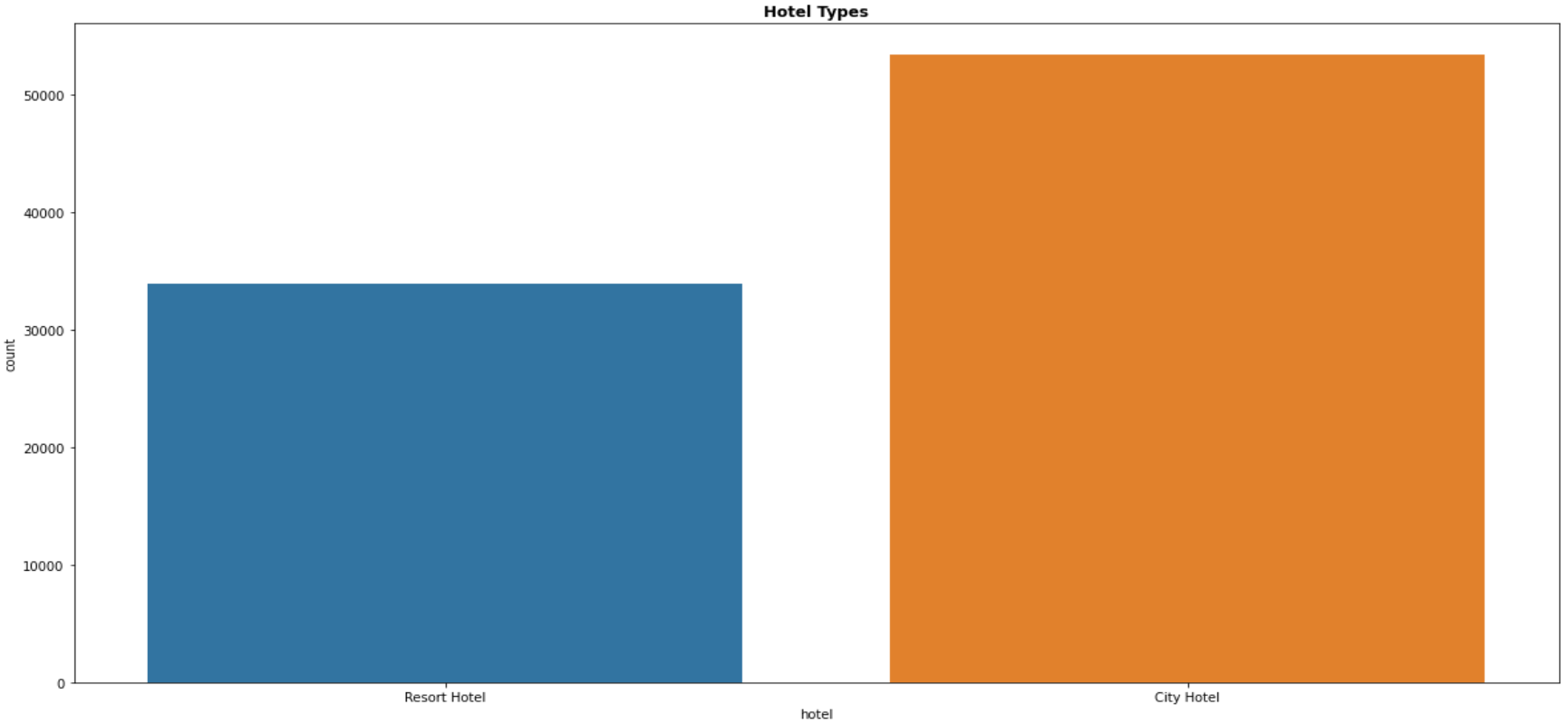
Numeric:

- arrival_date_year, arrival_date_week_number, arrival_date_day_of_month, stays_in_weekend_nights, stays_in_week_nights, booking_changes, adults, children, babies, agent, adr, company, days_in_waiting_list, required_car_parking_spaces, total_of_special_requests.

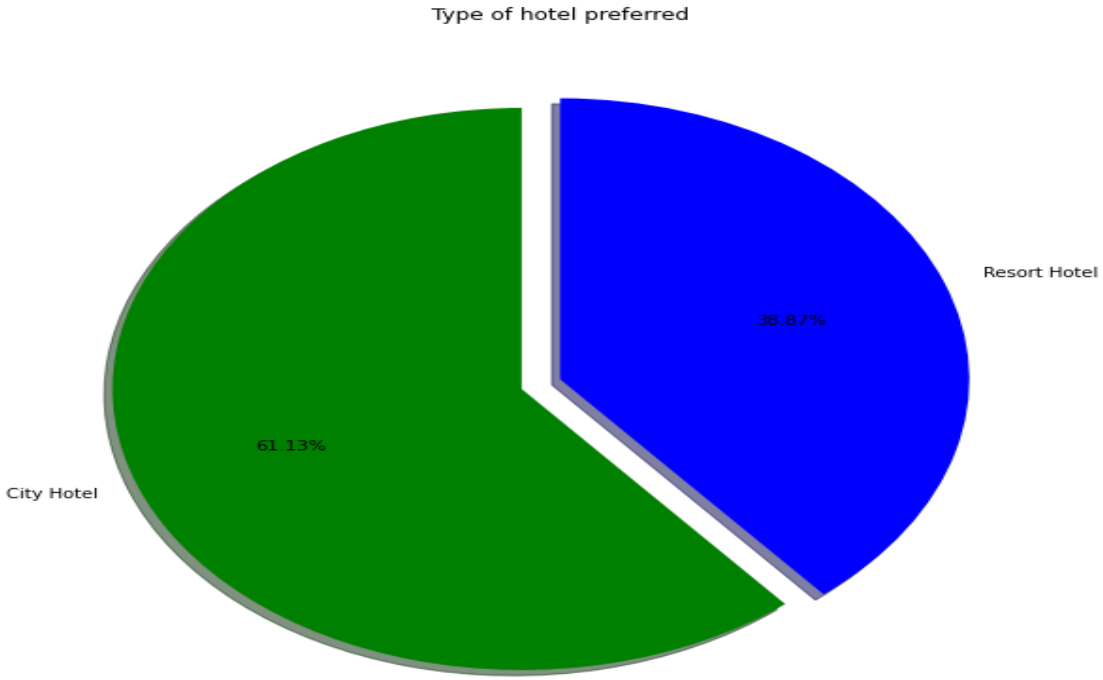
Binary:

- is_canceled, is_repeated_guest.

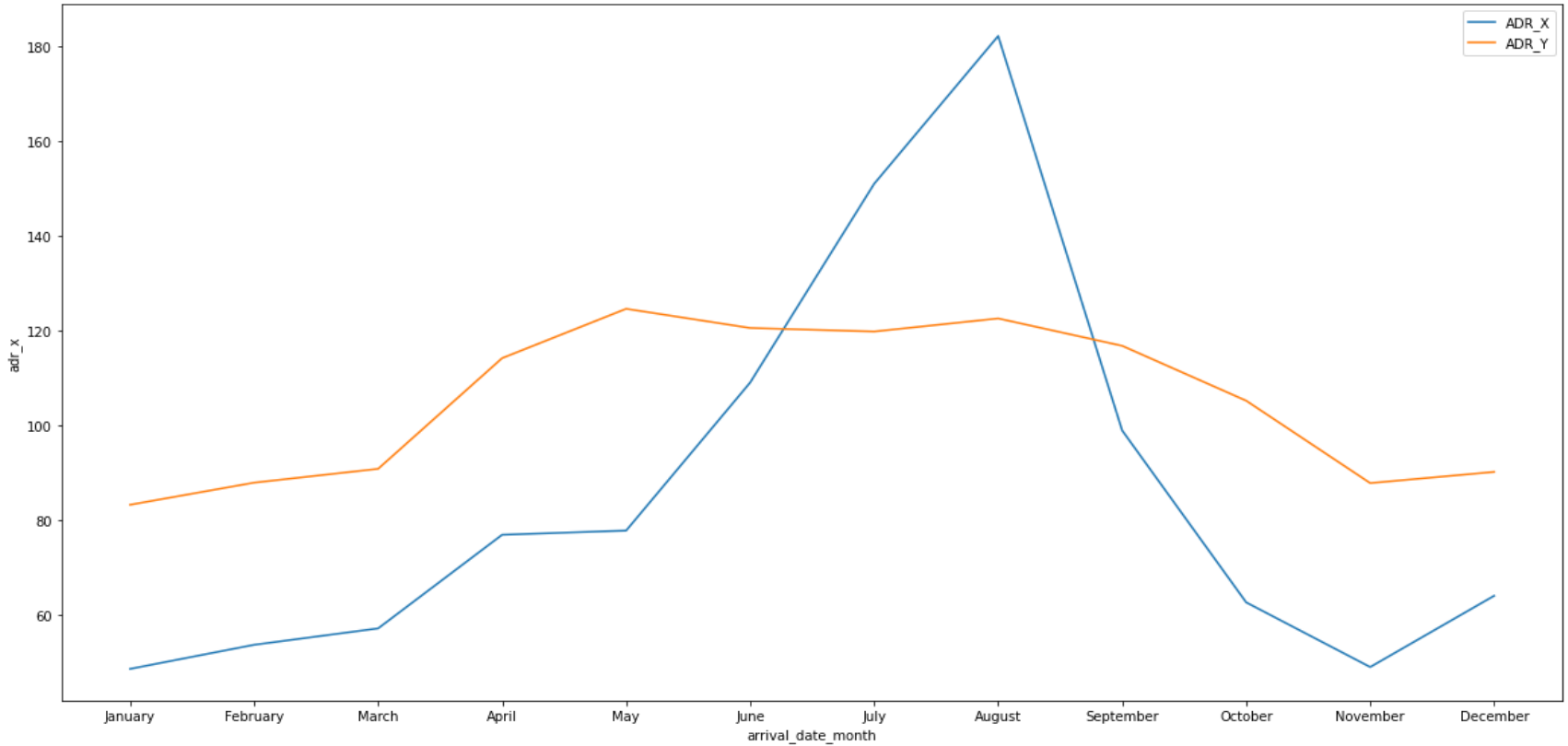
Hotel Type Visualization:



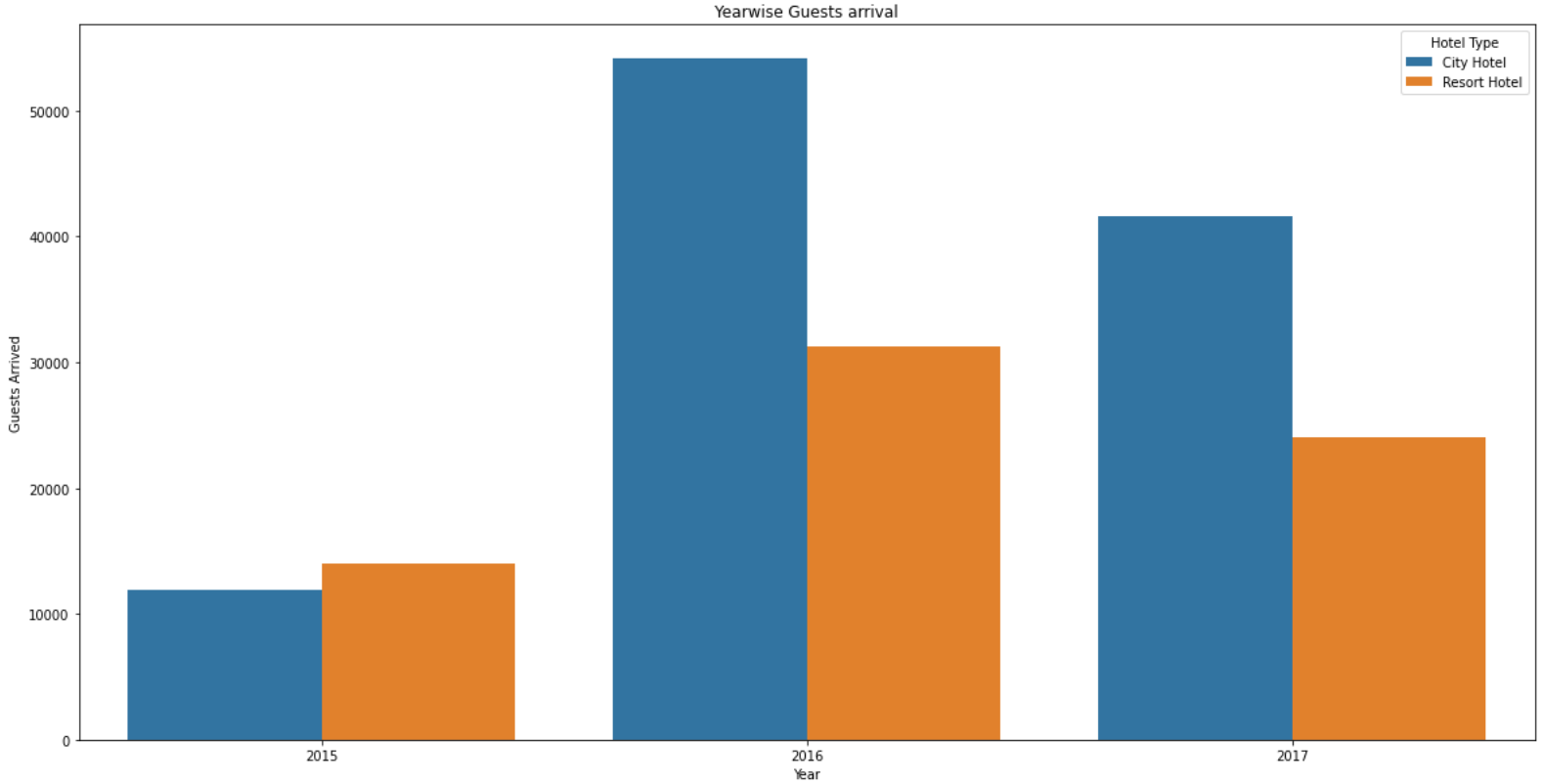
Type of hotel that most guests preferred:



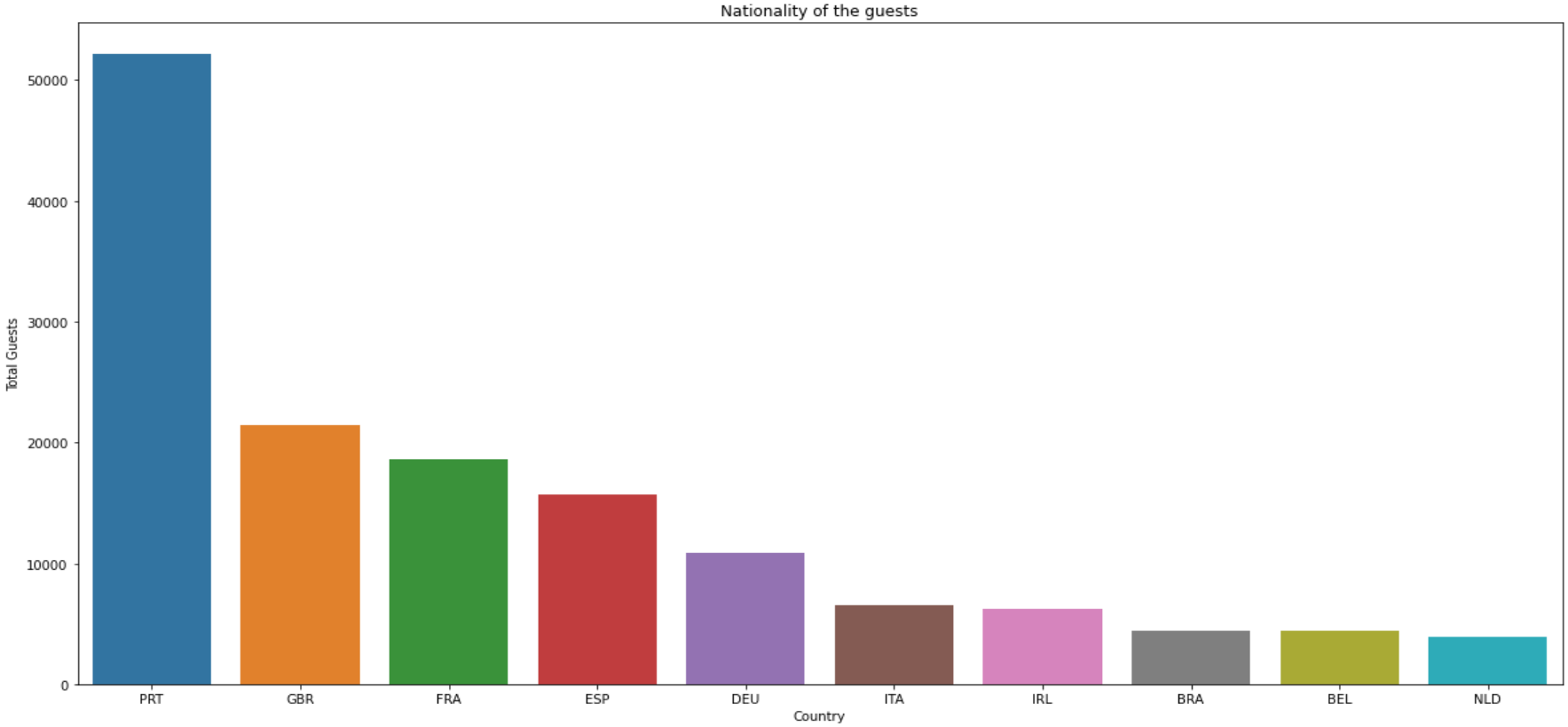
Average price for room per day in month wise:



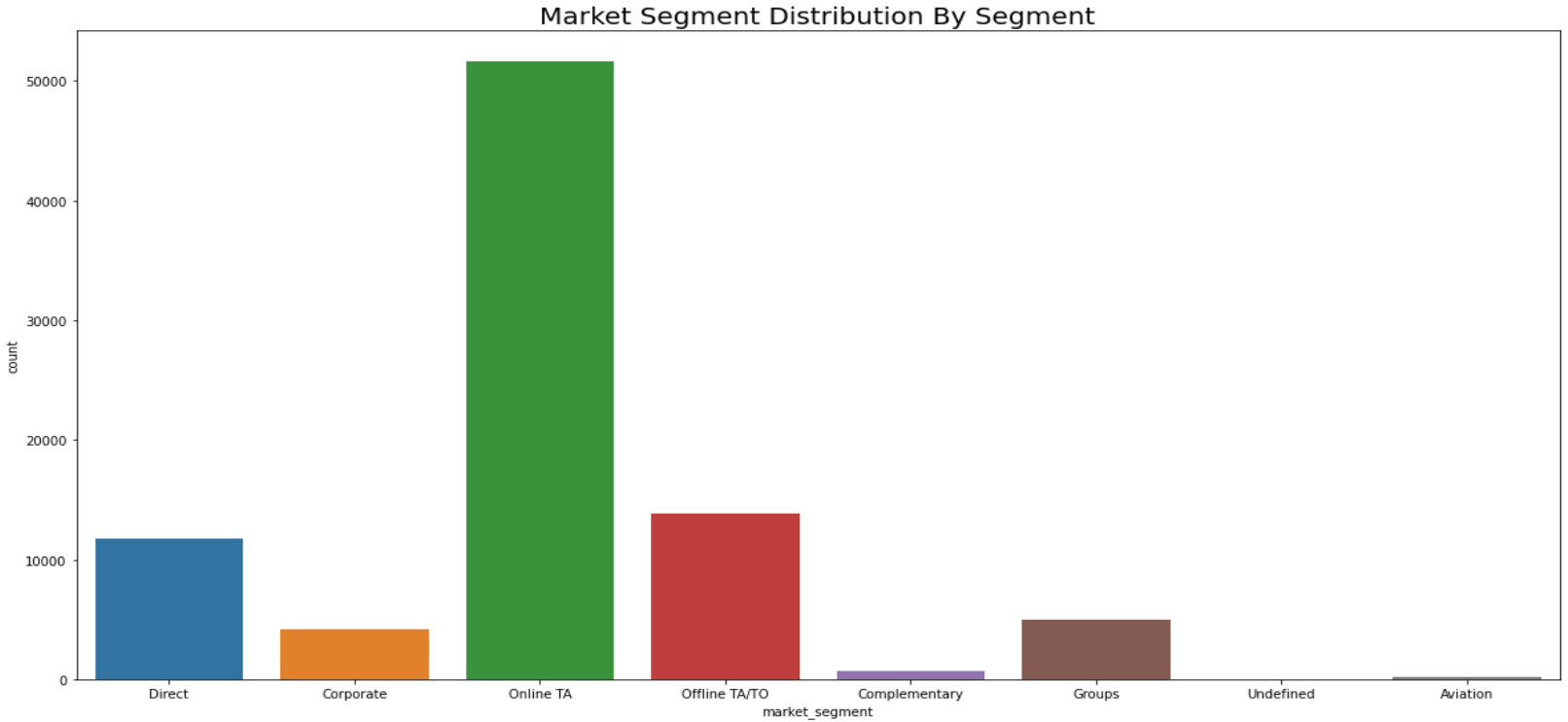
Highest number of guests in year wise:



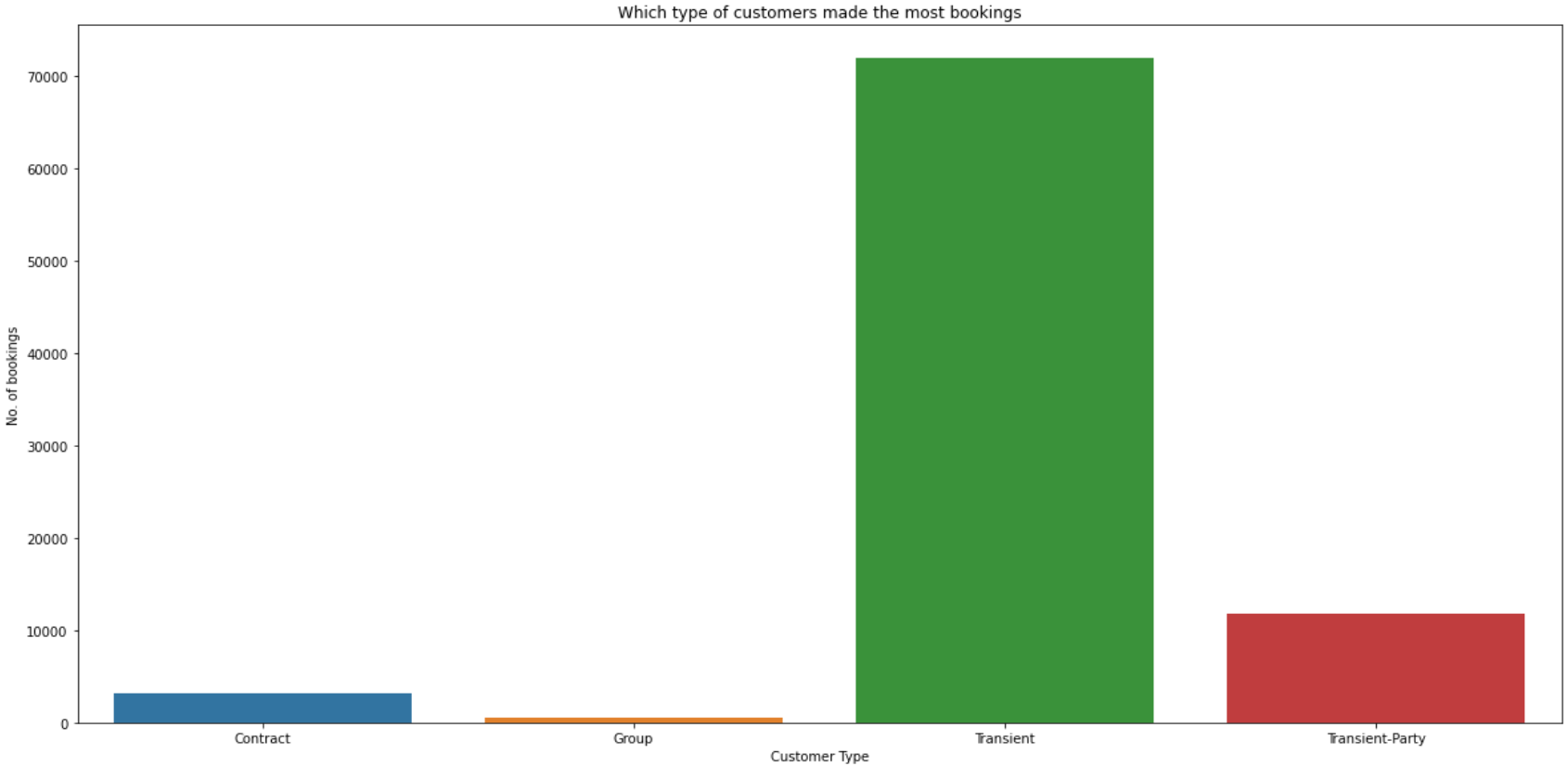
Nationality of guests:



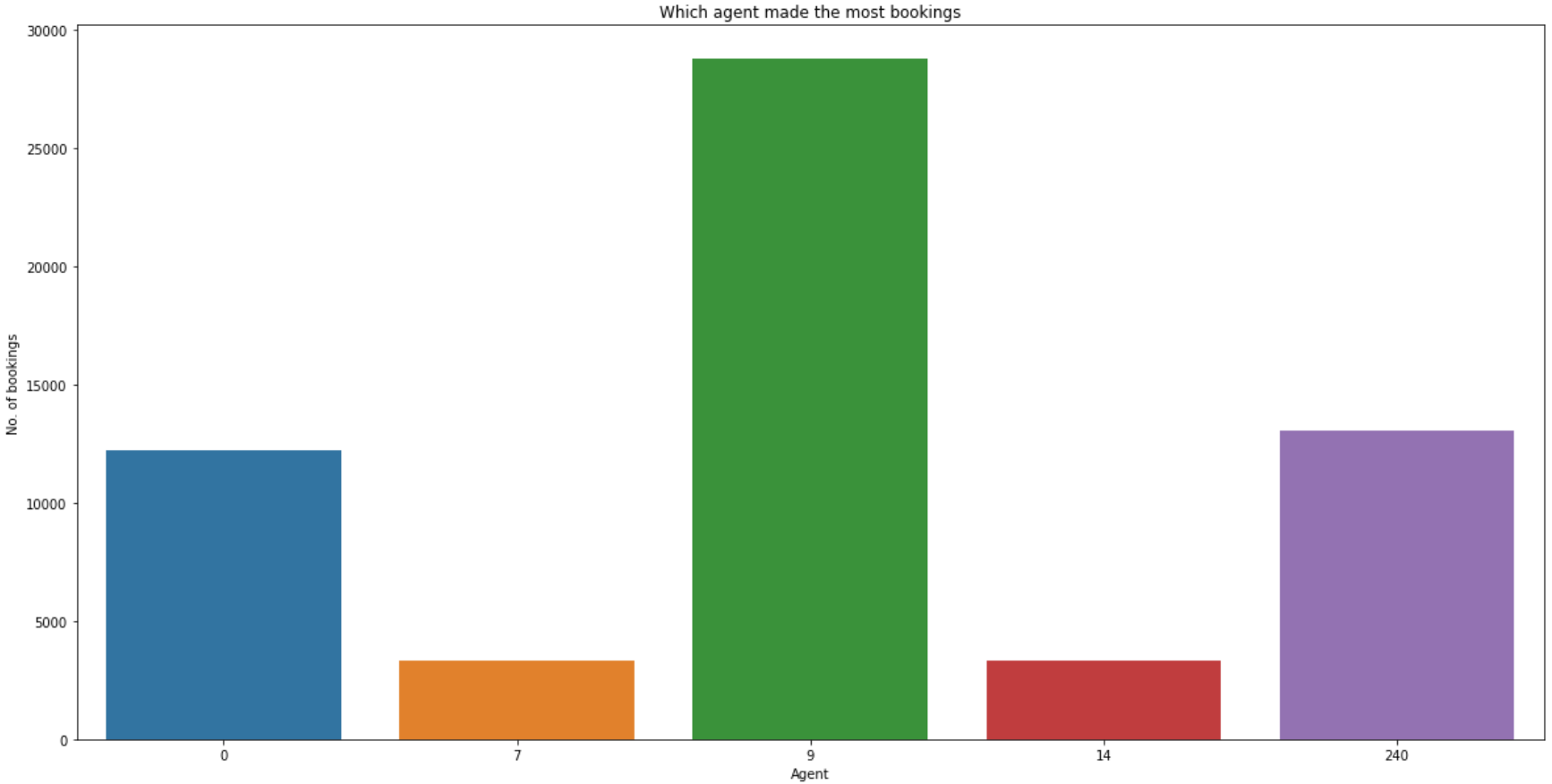
Distribution by market segment:



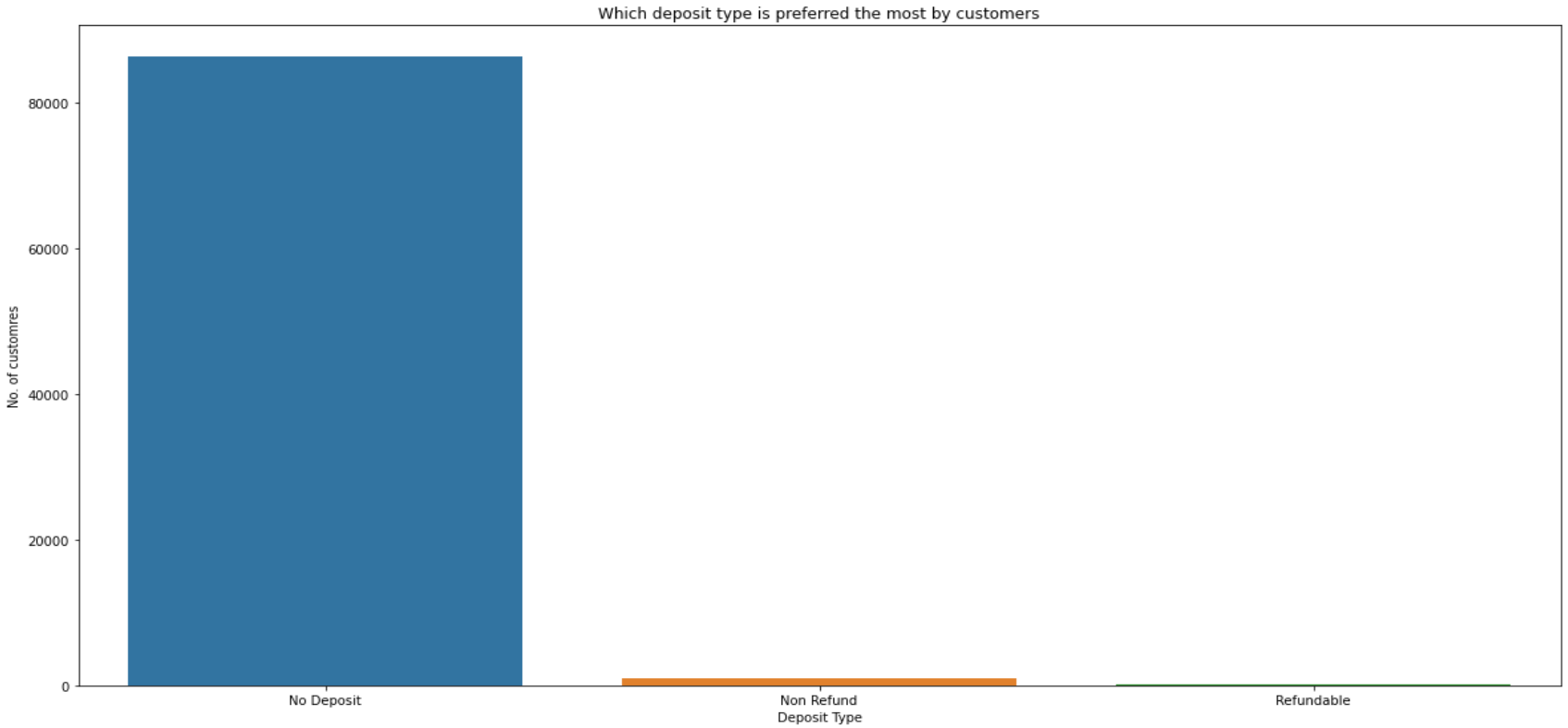
Type of customers made more bookings:



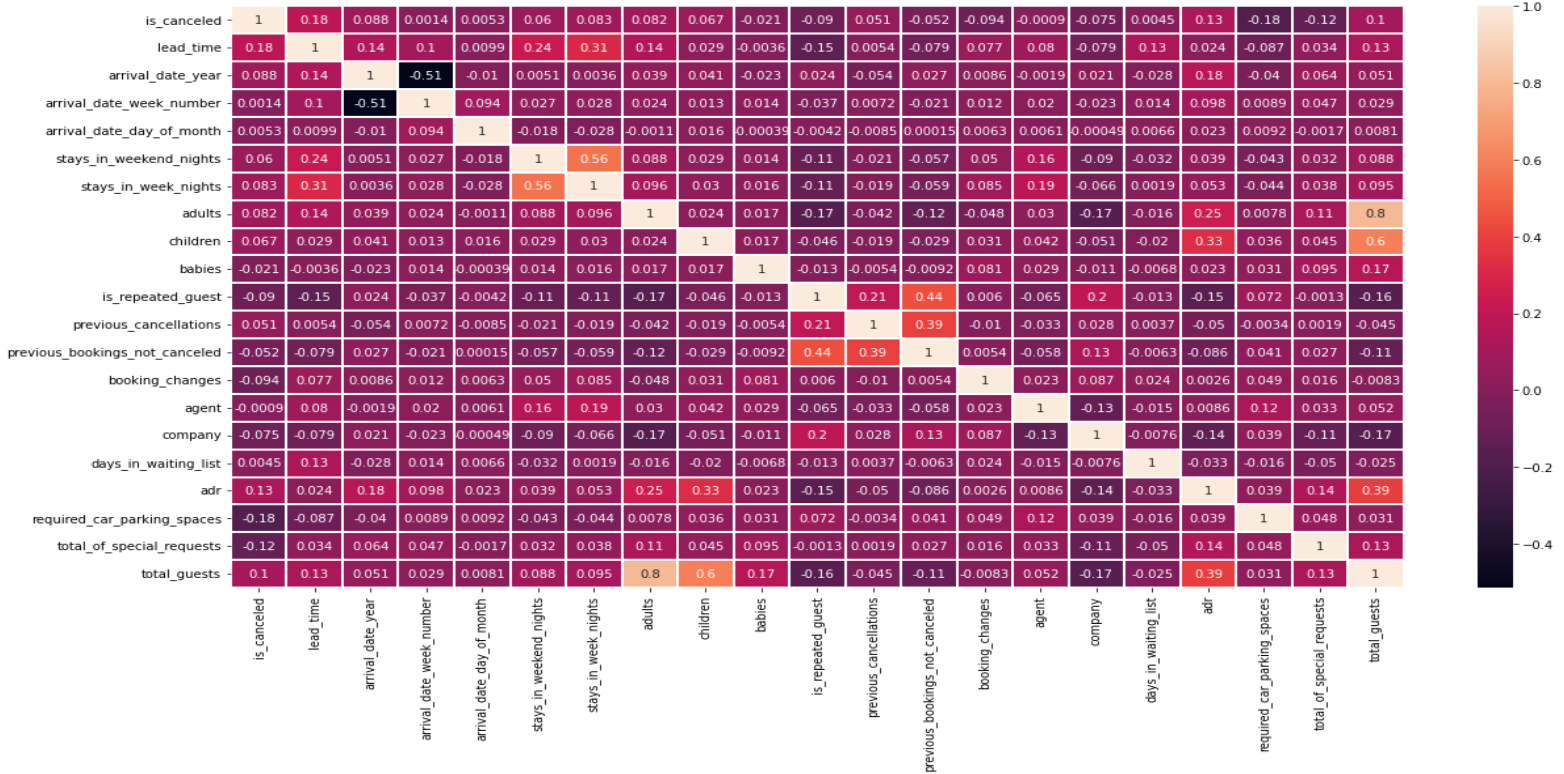
Agent with most No. of bookings:



Deposit type preferred by customers:



Correlation heat map:



Conclusion:

From the analysis of hotel booking we can conclude:

- The Data consists of information on bookings of two hotel type which is City Hotel and Resort Hotel.
- The dataset has the data of hotel bookings for the year 2015, 2016 and 2017.
- City Hotels are the most preferred than Resort hotels.
- Room rent is high in month of August, Hence the revenue of hotels are high.
- In the year 2016 had highest number of guests in hotels.
- The most numbers of guests are from European countries such as Portugal followed by Great Britain and France.
- Online travel agency had more distribution by market Segment.
- Transient type customer made most number of bookings.
- Agent ID 9 bought most number of bookings.
- Most of the customers preferred 'No Deposit' type bookings, so they're flexible and risk free in cancellation.
- Heat map provides correlation between all the variables of our dataset.

Challenges:

- Data set are wide and complex. It is hard to extract data from the respective column.
- There were NaN values, null values and duplicate rows which need to be handled.
- Designing multiple visualizations to summarize the information in the dataset and successfully communicate the results and tends to the reader.

Thank You