Capstone Project - I

EDA on Hotel Booking Data

Team members

Shadab Shaikh

Isha Jadhav

Process Overview:

I. Problem Definition

II. Initializing Data:

1. Importing Libraries and Loading Dataset
2. Viewing and Understanding Dataset

III. Data Analysis:

1. Data Cleaning
2. Data Visualization

IV. Conclusion

I. Problem Definition

Just like every other business, hotel businesses also heavily benefit from the customer behavior data. The hotels can understand different factors to optimize their booking numbers and increase their business numbers. These factors can decide what their selling point can be to attract more number of guests or how to cater a variety of different customers or what improvements their facilities need to make in order to increase customer attraction.

Our job is to go through one such dataset and analyze its different aspects and come up with the accurate conclusion of the analysis.

Let’s understand our dataset…

Basic Dataset Information

- **Dataset Name:** Hotel Booking Dataset
- **Time Period of available data:** July 2015 to August 2017 (Arrival) and
  August 2014 to September 2017 (Reservation)
- **Size and Shape:** The shape of dataset is (119390 x 32)
  i.e, 119390 rows and 32 columns

Let’s understand our dataset...

Feature Information

- **hotel:** Type of Hotel (City Hotel / Resort Hotel)
- **is_canceled:** Whether the booking is canceled(1) or not canceled(0)
- **lead_time:** Time difference between reservation and arrival of guests
- **arrival date_year:** Year of arrival
- **arrival date_month:** Month of arrival
- **arrival date_week_number:** Week number of arrival
- **arrival date_day_of_the_month:** Day of month of arrival
- **stays_in_weekend_nights:** Number of stays in weekend nights
- **stays_in_week_nights:** Number of stays in week nights
- **adults:** Number of adult guests

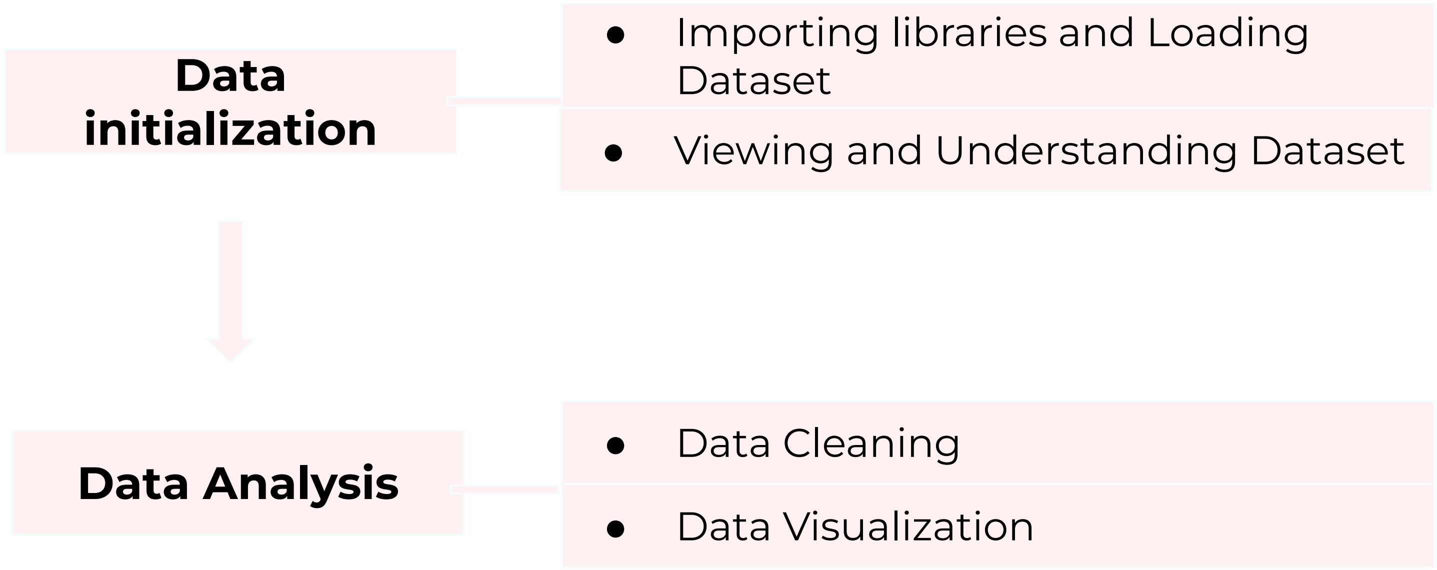
Let’s understand our dataset...

- **children:** Number of children
- **babies:** Number of babies
- **meal:** Type of meal booked
- **country:** Country of origin of guests
- **market_segment:** Purpose of and way of booking
- **distribution_channel:** Mode of reservation
- **is_repeated_guest:** Whether the guest is repeated(1) or not(0)
- **previous_cancellation:** Whether the guest had cancelled previously(1) or not(0)
- **previous_bookings_not_canceled:** Whether the previous booking cancelled
- **reserved_room_type:** Type of room reserved by guests
- **assigned_room_type:** Type of room assigned to the guests
- **booking_changes:** Number of changes made to the booking

Let’s understand our dataset...

- **deposite_type:** Type of deposit made (refundable/non-refundable/no deposit)
- **agent:** ID of agent that booked the hotel
- **company:** ID of company from which the booking was made
- **days_in_waiting_list:** Number of days in waiting list
- **customer_type:** Type of customers
- **adr:** Average Daily Rate (Average revenue made by the hotel per room per day)
- **required_car_parking_spaces:** Number of car parking spaces required
- **total_of_special_requests:** Number of special requests made by guest
- **reservation_status:** Status of reservation (Canceled/Check-Out/No-Show)
- **reservation_status_date:** Date of latest reservation status

Data Pipeline



Data initialization

- Importing libraries and Loading Dataset
- Viewing and Understanding Dataset

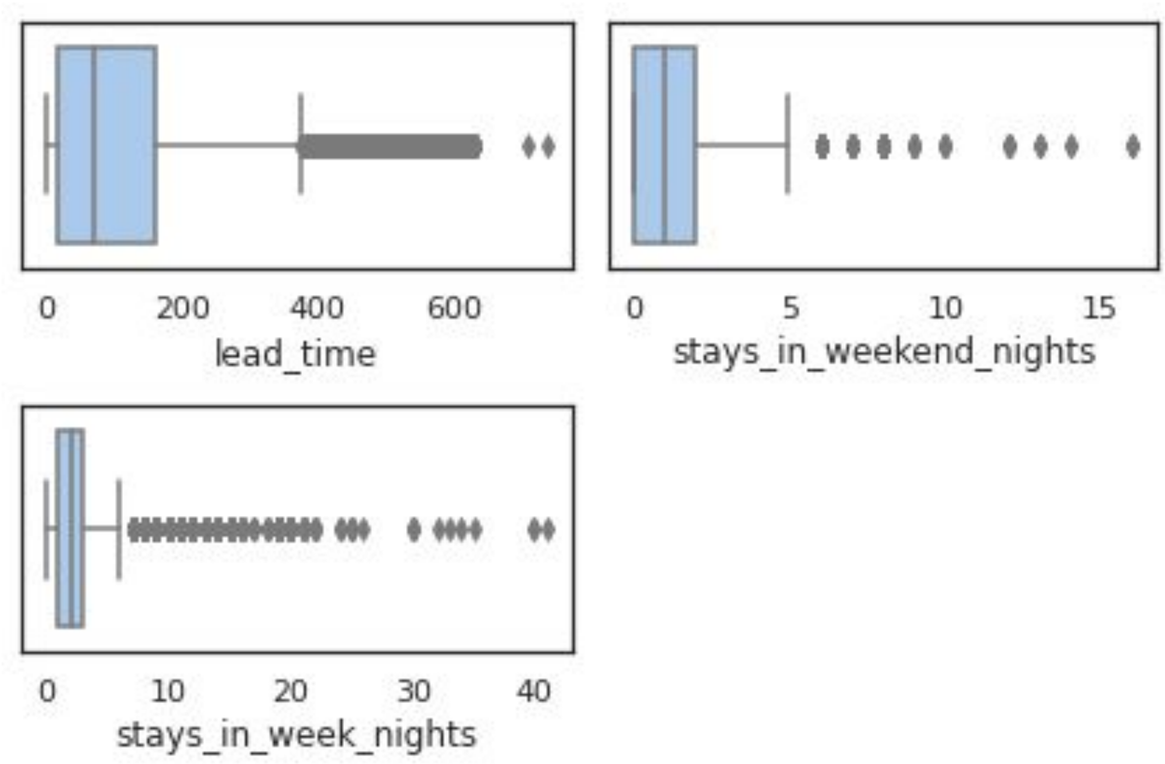
Data Analysis

- Data Cleaning
- Data Visualization

Data Pipeline

- **Data Initialization:** In this initial stage, we go through the dataset thoroughly and try to get maximum insights about the data.
  - **Importing Libraries and Loading Dataset:** We use Python commands to import required libraries and the dataset is loaded in iPython notebook in CSV format.
  - **Viewing and understanding the Dataset:** Here we get to know the size and shape of data, different aggregate values for numeric data, and also check for null values, etc.
- **Data Analysis:** In this stage, we deal with major tasks required for analysis of the dataset.
  - **Data Cleaning:** We use different parameters to make data clean and workable by removing null values, replacing insignificant data type with relevant ones and treating outliers.
  - **Data Analysis:** We use different plots to visualize the data in different columns and their relation with each other.

Outlier detection



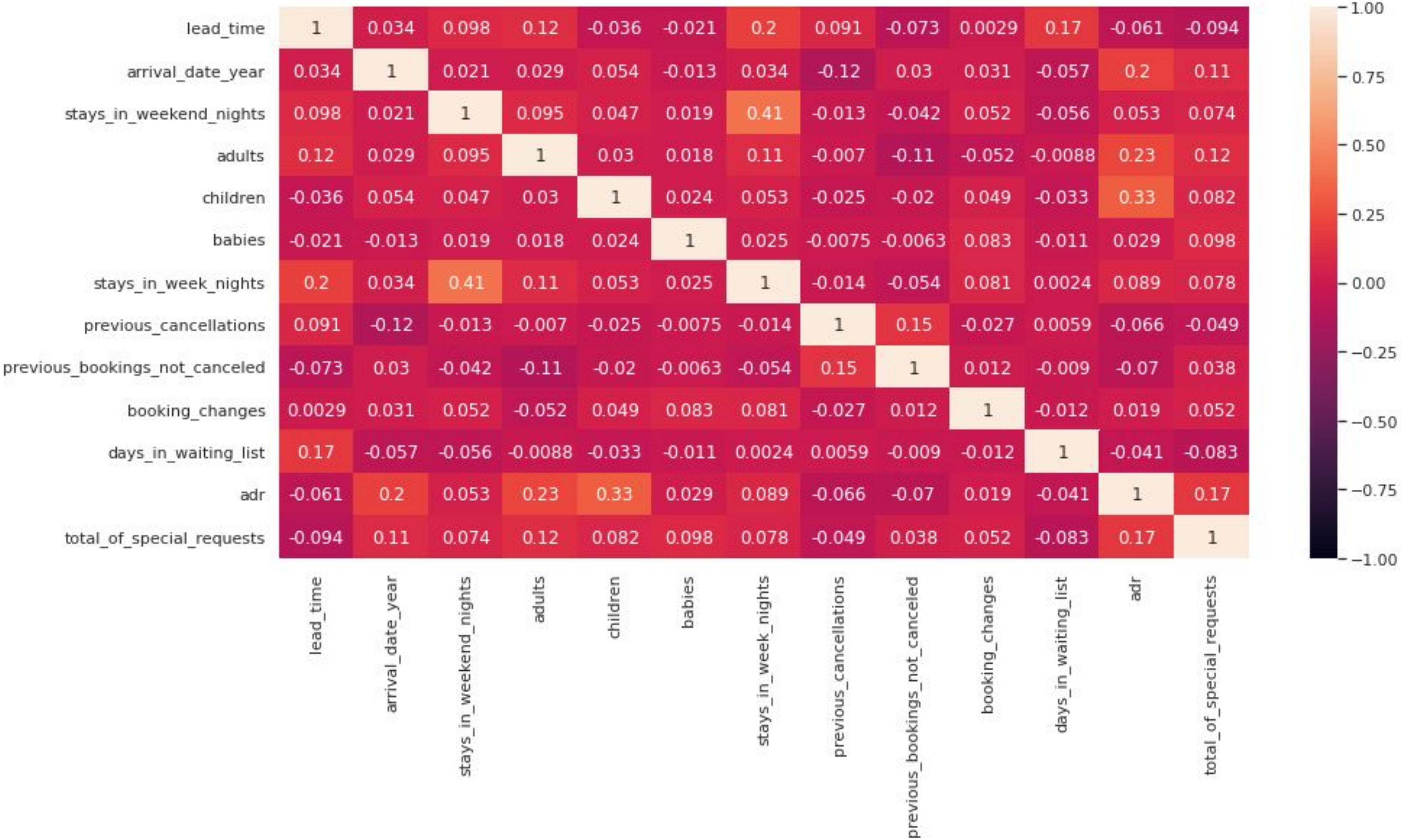
This is a part of data cleaning and processing stage.

We can see using these box plots, the outliers present in the following columns.
-lead_time
-stays_in_weekend_
nights
-stays_in_week_nights

Correlating Different Parameters

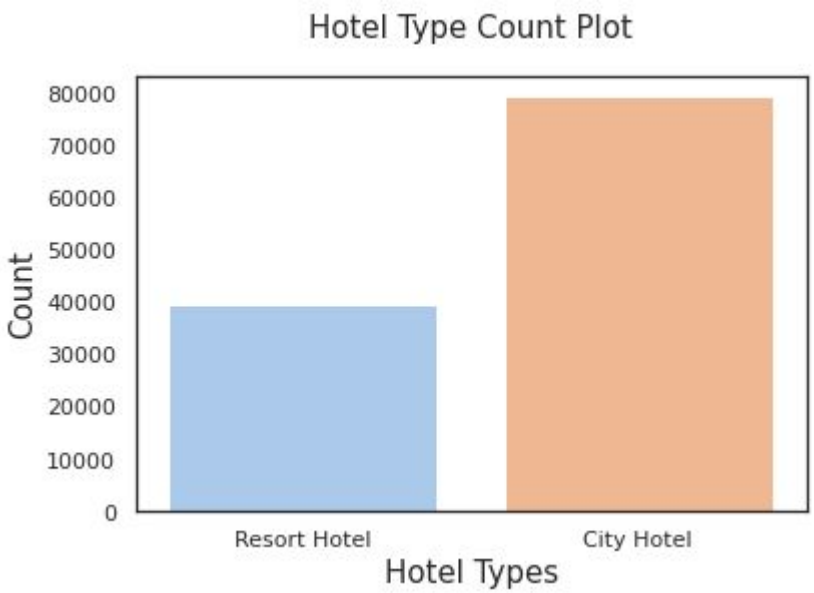
Correlation is helpful to understand the relationship between different parameters in the dataset, how they affect each other and the overall business.
Understanding correlation is essential to get a brief overview of how the different parameters simultaneously vary with each other. This gives us a brief idea of the overall dataset and how the values of each table vary throughout.

Using the heatmap, we can easily go through this correlation...



	lead_time	arrival_date_year	stays_in_weekend_nights	adults	children	babies	stays_in_week_nights	previous_cancellations	previous_bookings_not_canceled	booking_changes	days_in_waiting_list	adr	total_of_special_requests
lead_time	1	0.034	0.098	0.12	-0.036	-0.021	0.2	0.091	-0.073	0.0029	0.17	-0.061	-0.094
arrival_date_year	0.034	1	0.021	0.029	0.054	-0.013	0.034	-0.12	0.03	0.031	-0.057	0.2	0.11
stays_in_weekend_nights	0.098	0.021	1	0.095	0.047	0.019	0.41	-0.013	-0.042	0.052	-0.056	0.053	0.074
adults	0.12	0.029	0.095	1	0.03	0.018	0.11	-0.007	-0.11	-0.052	-0.0088	0.23	0.12
children	-0.036	0.054	0.047	0.03	1	0.024	0.053	-0.025	-0.02	0.049	-0.033	0.33	0.082
babies	-0.021	-0.013	0.019	0.018	0.024	1	0.025	-0.0075	-0.0063	0.083	-0.011	0.029	0.098
stays_in_week_nights	0.2	0.034	0.41	0.11	0.053	0.025	1	-0.014	-0.054	0.081	0.0024	0.089	0.078
previous_cancellations	0.091	-0.12	-0.013	-0.007	-0.025	-0.0075	-0.014	1	0.15	-0.027	0.0059	-0.066	-0.049
previous_bookings_not_canceled	-0.073	0.03	-0.042	-0.11	-0.02	-0.0063	-0.054	0.15	1	0.012	-0.009	-0.07	0.038
booking_changes	0.0029	0.031	0.052	-0.052	0.049	0.083	0.081	-0.027	0.012	1	-0.012	0.019	0.052
days_in_waiting_list	0.17	-0.057	-0.056	-0.0088	-0.033	-0.011	0.0024	0.0059	-0.009	-0.012	1	-0.041	-0.083
adr	-0.061	0.2	0.053	0.23	0.33	0.029	0.089	-0.066	-0.07	0.019	-0.041	1	0.17
total_of_special_requests	-0.094	0.11	0.074	0.12	0.082	0.098	0.078	-0.049	0.038	0.052	-0.083	0.17	1

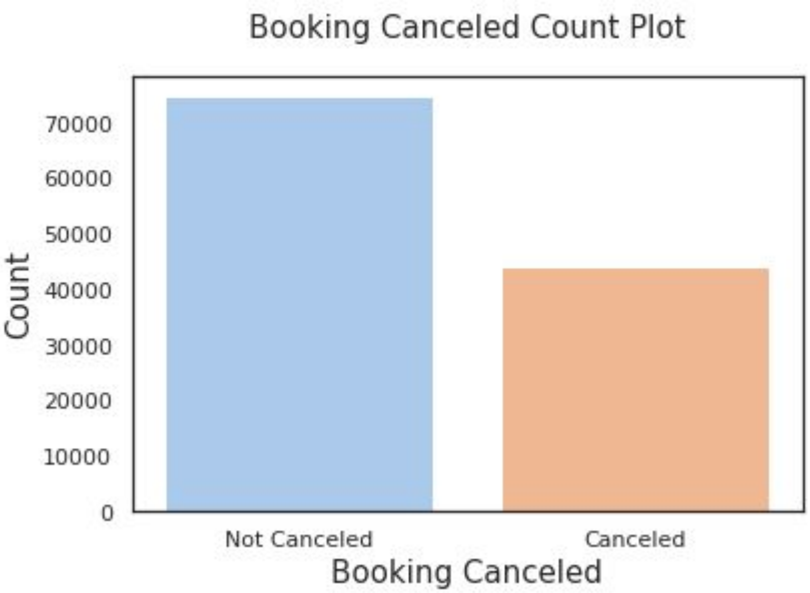
Types of Hotels



We have two types of Hotels in the dataset we are working with.
-Resort Hotel
-City Hotel

We can see that the City Hotels are preferred by the most number of guests.

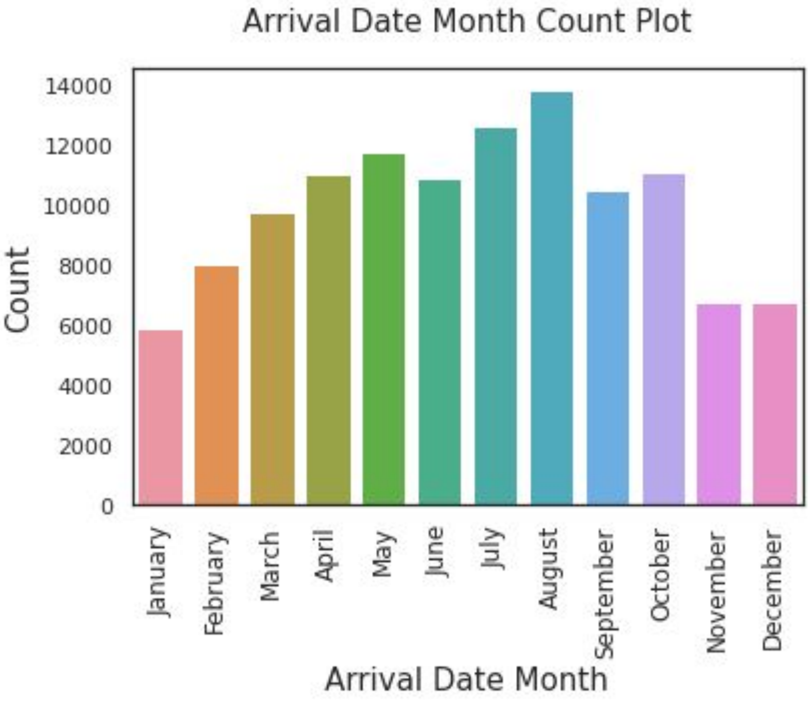
Cancelled Bookings



We can see the number of cancelled bookings with the help of this visualization.

Majority of guests do not cancel, once the bookings are finalised

Arrival Date Month

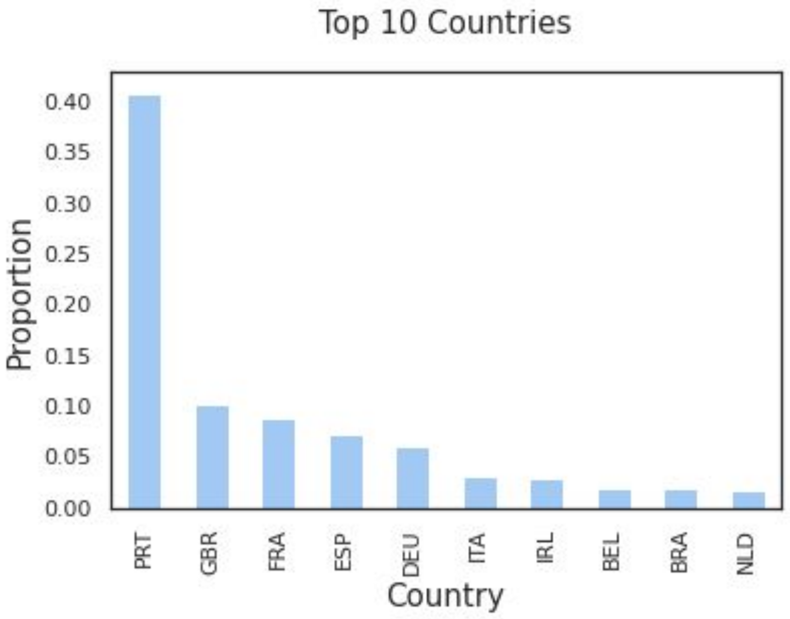


We can see the trend of arrival of guests in different months of the year.

Most number of guests arrive in the month of August followed by July and May.

Leasts guest arrival is at the time of year end and the initial months of the year

Guests from Countries

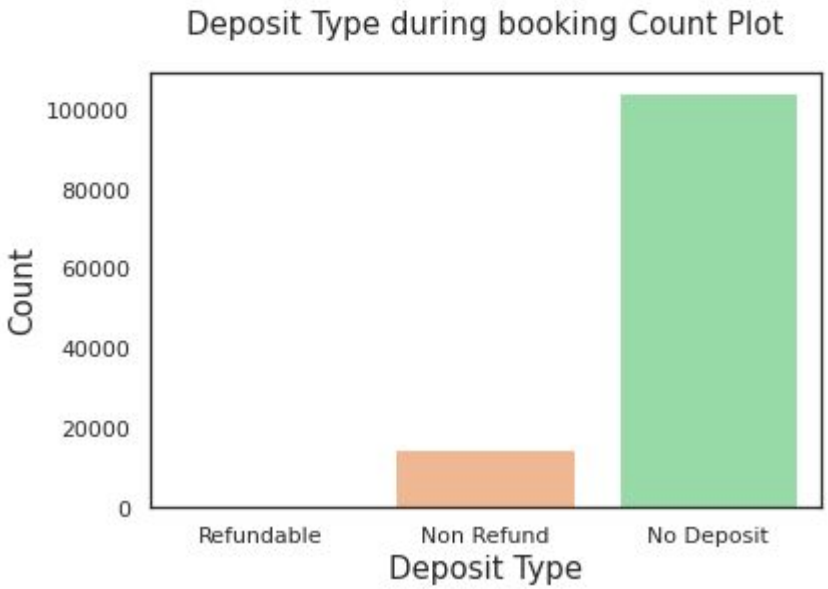


Our dataset implies that the guests arrive from all over the world.

With the help of this visualization we can see the top 10 countries from which the hotels receive most number of guests.

We can see Portugal tops the charts with being the most number of guests from there.

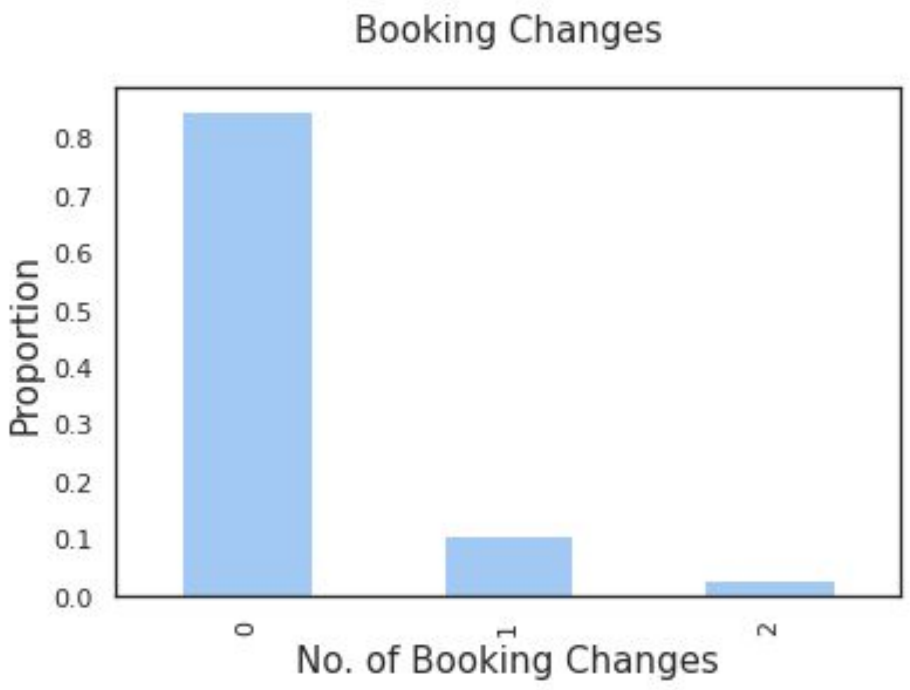
Deposit Type



There are three types of deposits made at the time of booking.
-Refundable
-Non-Refund
-No Deposit

We can see that most people prefer not to make any deposits at the time of booking.

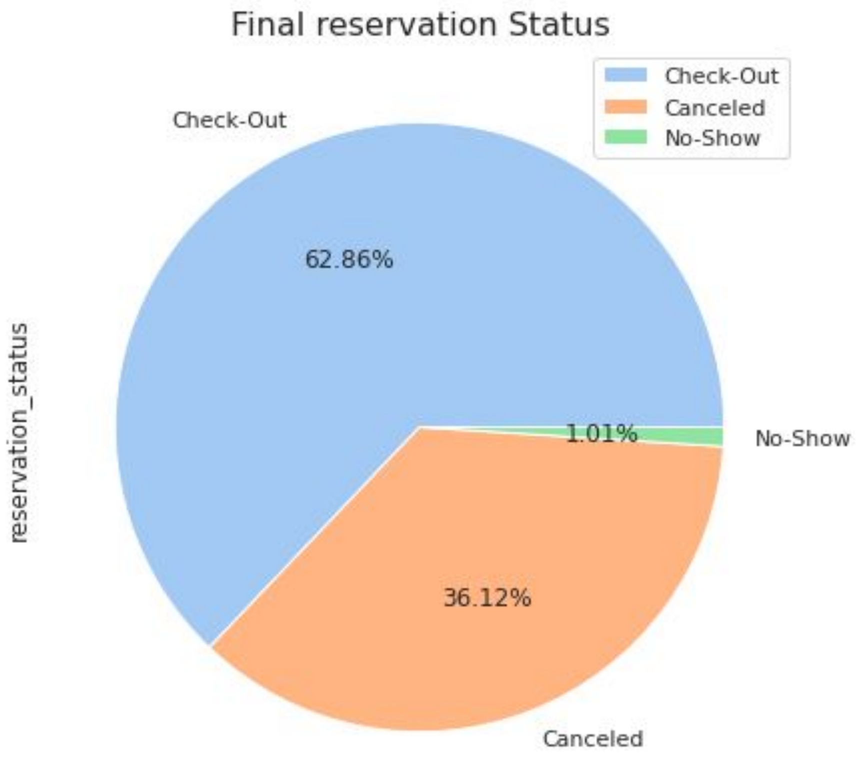
Booking Changes



We can see the trend of number booking changes at the time of booking.

While majority of customers do not make any booking changes, a significant number of customers make at least one booking changes.

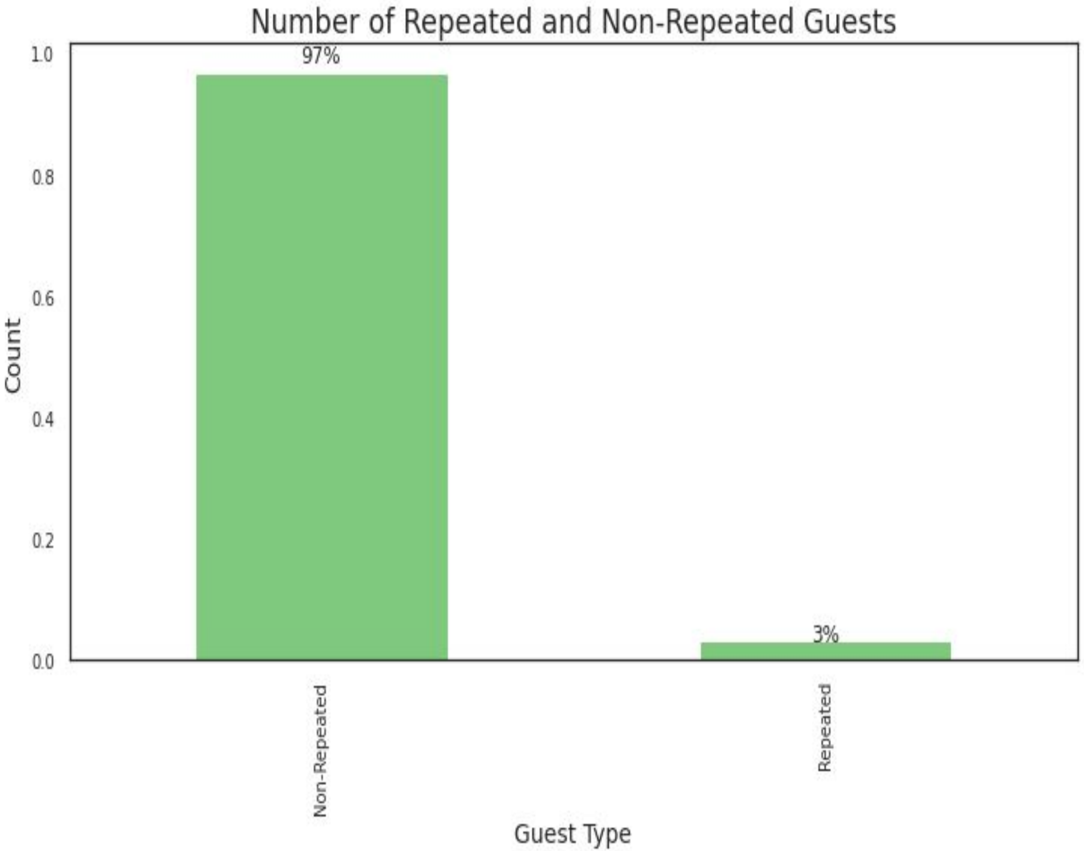
Reservation Status



Here we can see the bifurcation of the reservation status of the guests in the hotels.

We can see that 62.86% of guests go forward with their booking, about 36.12% of guests cancel their reservation. And about 1.01% of guests never show up for their visit.

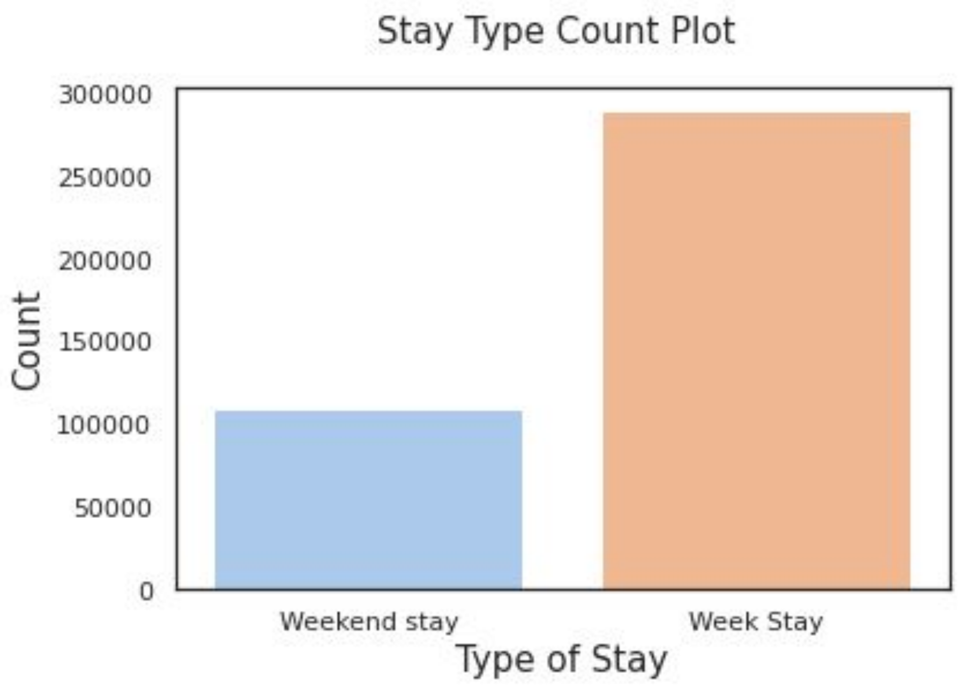
Repeated Guests



This plot shows us the trend of repeated and non repeated guests.

We can see that about 3% of people repeat their visit to the same hotel.

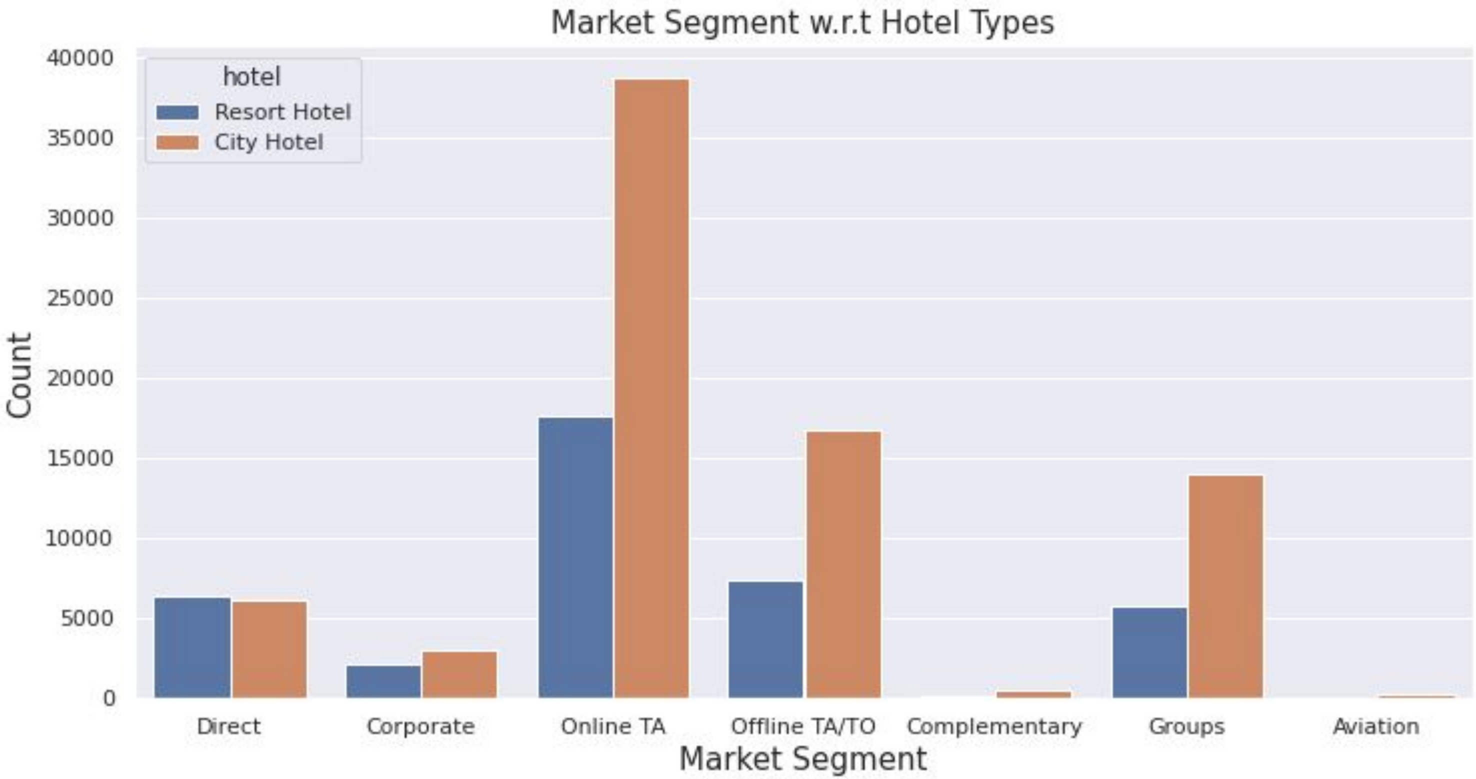
Stays in Weekend Nights Vs Week Nights



This chart helps us visualize the variation of stays in weekend nights and stays in week nights.

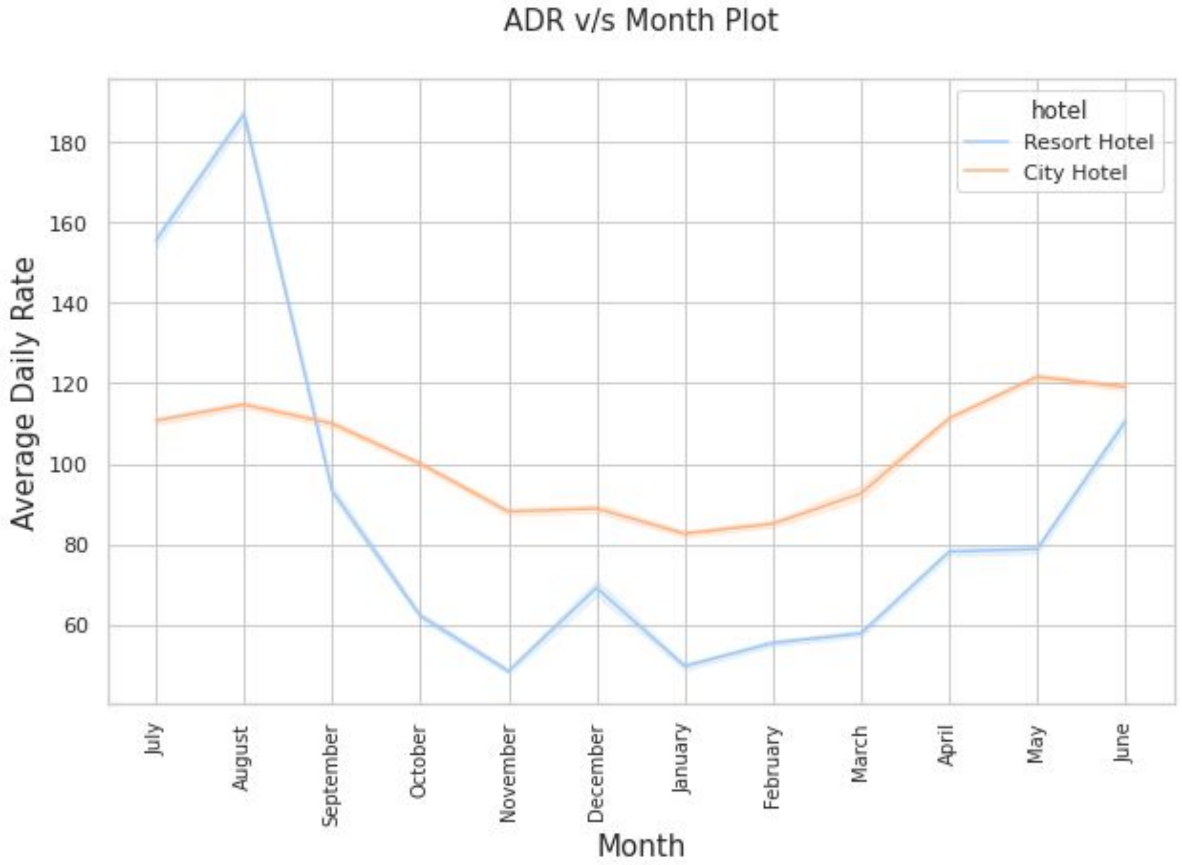
We can see that the majority stays by the guests is during week nights.

Market Segment Vs Hotel Types



This shows that “Online TA” is prevalent in “City Hotel” Type.

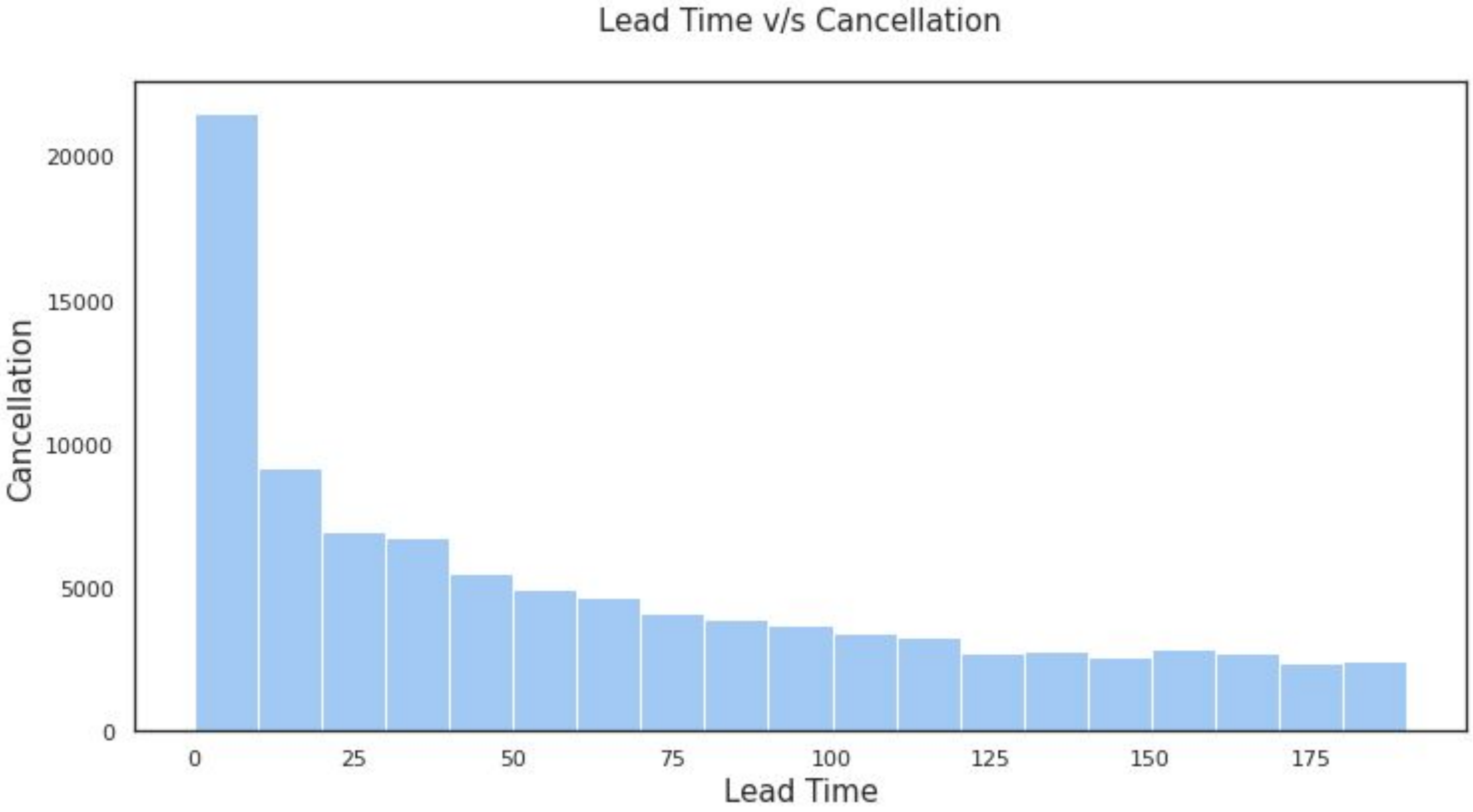
Average Daily Rate Vs Month



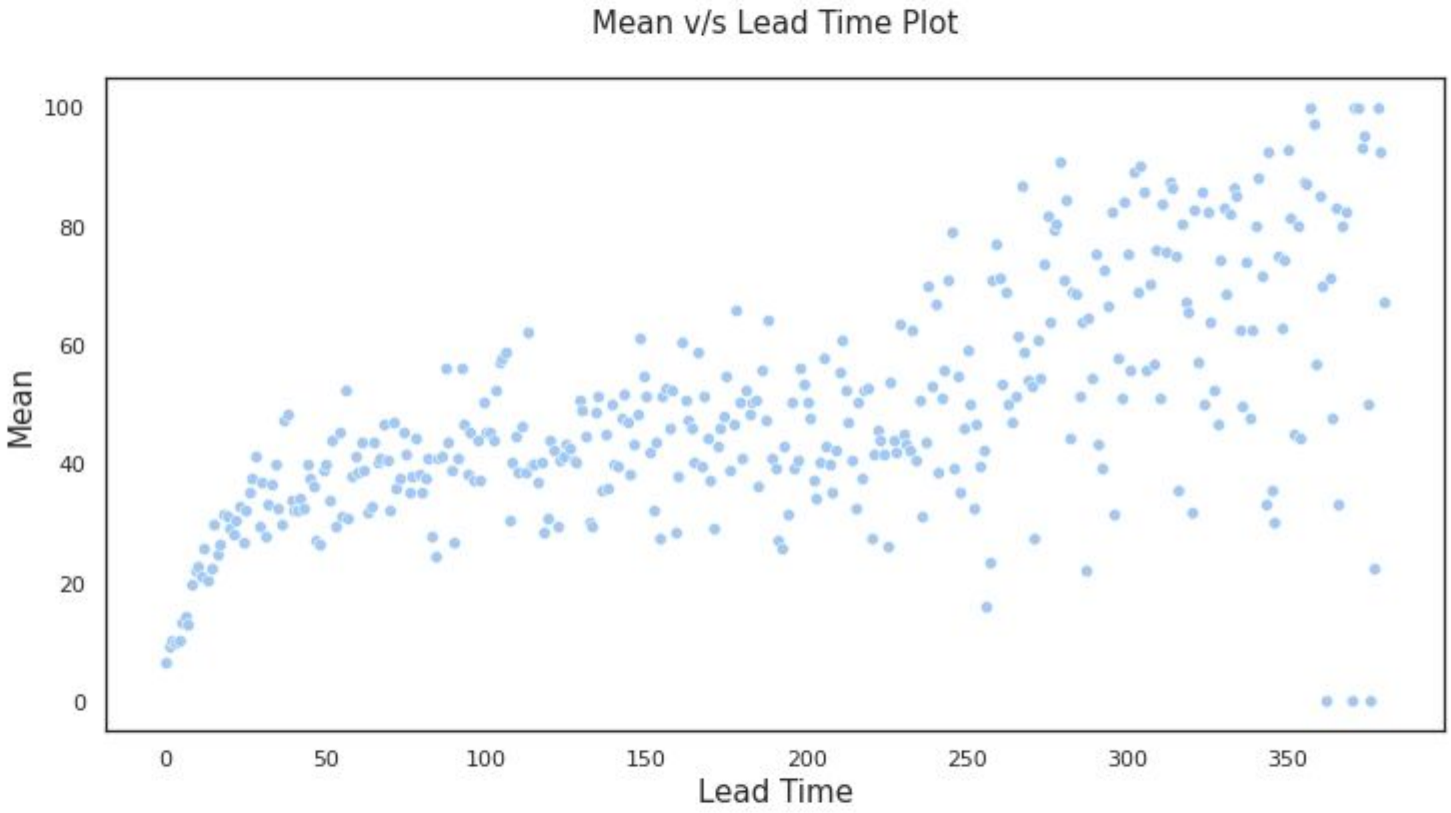
This plot shows us the variation of ADR w.r.t month for different hotel types.

We can see that ADR is highest for the month of August in Resort Hotel type and for the month of May in City Hotel type.

Lead Time v/s Cancellation (Histogram)



Mean v/s Lead Time (Scatter Plot)



Conclusion

- Room type 'A' is the most reserved room with 71.99% and Room type 'B' is the least reserved room type.
- 'August' is the month of Highest and 'January' is the month of Lowest Customer Arrival count with 11.6% and 4.9% respectively.
- 'Transient' customer is the most common and 'Group' is the least common customer type with 75% and 0.47% respectively.
- Around 84% customers have not made any changes and around 16% customers made booking changes.

- **More than 68% guests have only 2 members, around 18% are traveling alone. Some are travelling in groups more than 30.**
- **Around 37% customers cancelled their bookings.**
- **'Resort' average daily price is very fluctuating compared to 'Hotel' which is stable.**

THANK YOU