Capstone Project

Booking.com – Hotel Booking Analysis

Project Type – EDA (Exploratory Data Analysis)

Contribution - Individual

BY - DEBARPAN CHOWDHURY

Table of Content

- Introduction of ‘Booking.com’.
- Data Description of Variables which are Mainly Used
- Summery
- Dataset Summery
- Libraries used
- Exploratory Data Analysis (EDA)
- Conclusion
- Challenges Faced

Introduction of ‘Booking.com’

- **Founded:** 1996, Amsterdam, Netherlands (Started as a small Dutch start-up)
- **Mission:** Make travel easier for everyone
- **Services:**
  - Connects travellers with:
    - Memorable experiences
    - Transportation options
    - Accommodation (hotels, homes, apartments, etc.)
  - Largest travel marketplace for properties of all sizes
  - Reaches global audience for property owners
- **Availability:** 43 languages
- **Listings:** Over 28 million total, including 6.6 million+ unique stays
- **Benefits:** Easy booking, 24/7 customer support

Data Description of Variables which are Mainly Used

- **hotel:** Name of hotel (City or Resort)
- **is_canceled:** Whether the booking is canceled or not (0 for no canceled and 1 for canceled)
- **lead_time:** time (in days) between booking transaction and actual arrival.
- **arrival_date_year:** Year of arrival
- **stays_in_weekend_nights:** No. of weekend nights spent in a hotel
- **stays_in_week_nights:** No. of weeknights spent in a hotel
- **country:** Country of origin of customers (as mentioned by them)
- **market_segment:** What segment via booking was made and for what purpose.
- **is_repeated_guest:** Whether the customer has made any booking before(0 for No and 1 for Yes)
- **deposit_type:** Type of deposit at the time of making a booking (No deposit/ Refundable/ No refund)
- **previous_cancellations:** No. of previous canceled bookings.
- **previous_bookings_not_canceled:** No. of previous non-canceled bookings.
- **agent:** Id of agent for booking
- **days_in_waiting_list:** No. of days on waiting list.
- **customer_type:** Type of customer(Transient, Group, etc.)
- **adr:** Average Daily Rate.
- **reservation_status:** Whether a customer has checked out or canceled,or not showed
- **reservation_status_date:** Date of making reservation status.

Summery

This analysis delved into a comprehensive dataset of **119,390 hotel bookings**, **aiming to uncover insights into customer behaviour and booking patterns.** The dataset **encompassed a wide range of variables, including booking details, room preferences, and cancellation information.**

Before diving into the analysis, the dataset underwent a **careful cleaning process**, **Duplicate records were eliminated**, **Corrected improper data type** format which was used in the data, and **missing values were addressed**. The 'country', 'children', and 'agent' variables were imputed with their respective modes, while the 'company' variable, due to its high percentage of missing values, was dropped.

The EDA phase involved a combination of statistical analysis and data visualization techniques:

Descriptive Statistics
Data Visualization

Dataset Summery

CATEGORICAL COLUMNS

- hotel
- arrival_date_month
- meal
- country
- market_segment
- distribution_channel
- reserved_room_type
- assigned_room_type
- deposit_type
- Customer_type
- reservation_status
- reservation_status_date

NUMERICAL 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
- agents
- days_in_waiting_list
- adr
- required_car_parking_spaces
- total_of_special_requests

Libraries Used

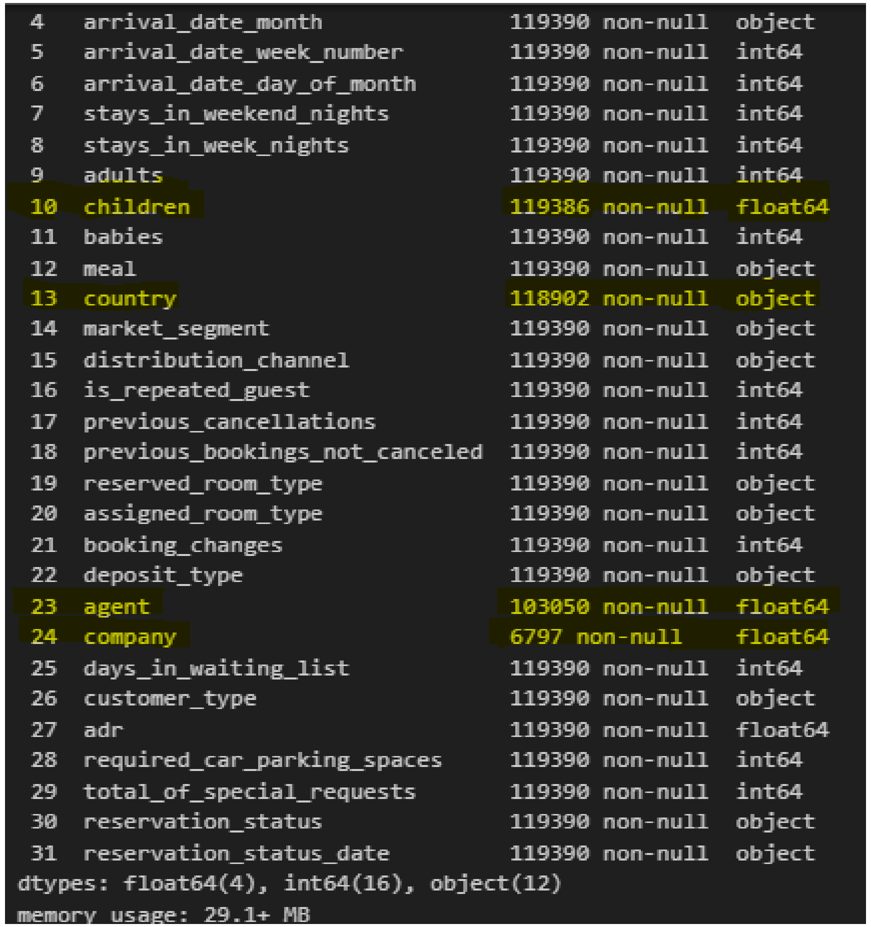
- **Pandas** - for data manipulation, aggregation
- **Matplotlib and Seaborn** - for visualisation and behaviour with respect to the target variable.
- **NumPy** - for computationally efficient operations

Exploratory Data Analysis (EDA)

1. Analyzing the Data
2. Ploting the DataFrame in the HeatMap to visualize total count of null values in each column and also if there any relationship between null values.
3. Replacing null values from Children, Country, & Agents and Dropped Country Column
4. Correcting data type
5. Removing Duplicate Rows
6. Data Visualization

1. EDA –Analyzing Data

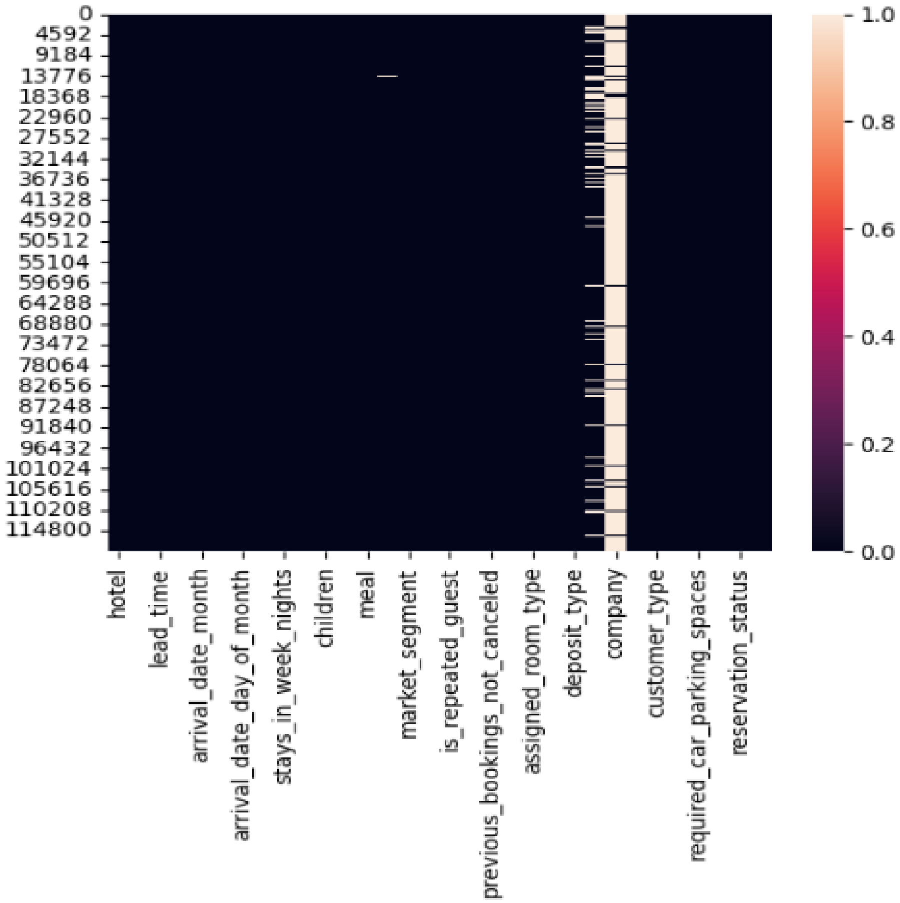
- In the DataFrame there are 119390 rows and 32 columns.
- And after that use info() function to see the datatype and the non-null count of every columns
- After calling the info() function we find that among 32 columns, 4 contain null value which are children, country, agent and company
- Then we count total number null value present in each column which contain null values using is_null() and finded
  - children column have 4 rows with null value
  - country column have 488 rows with null value
  - agent column have 16340 rows with null value
  - Company column have 112593 rows with null values



```
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

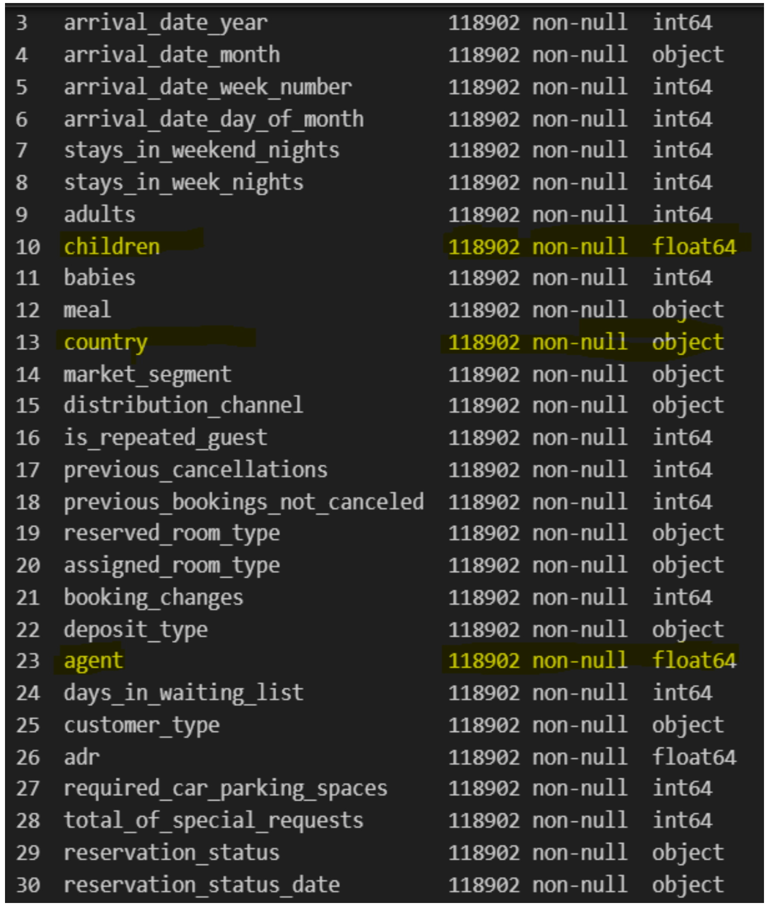
2. EDA – Plotted DataFrame.isnull() in Heatmap

- I have ploted DataFrame.isnull() in the heatmap so that I can examine weather there are any relation between the columns null values.
- So, after visualizing I don’t able to find any correlation between null value of different columns.



3. EDA – Replacing Null Values and Droping Column

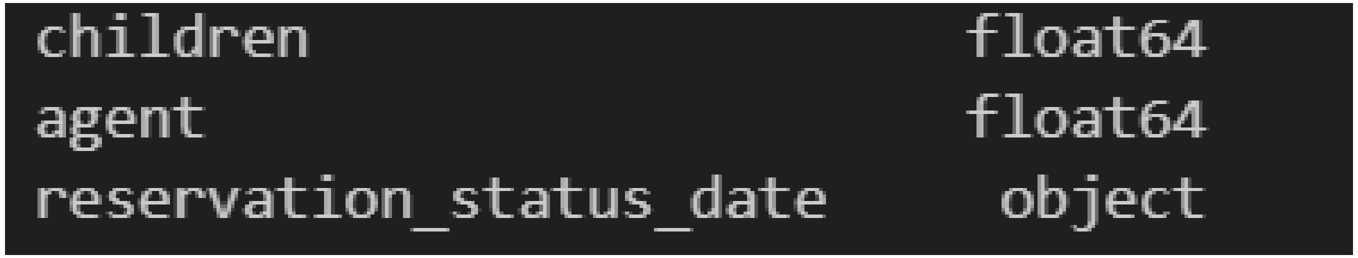
- For the children column which contain 4 rows with null values, I have replaced them with 0.0. Because when I am calculating mean of existing rows which have some value present in this column the result is 0.0.
- For the country column which contain 488 rows with null values, I have dropped the null values rows with dropna(). Because it is a categorical column. In categorical column we have only two option, 1) Either replace the null values with the existing value having maximum value_counts() or 2) If the number of rows are small as compare to the to total number of rows in the column/dataset, so, preferable I to drop the rows.
- For the agent column which contain 16340 rows with null values, I have replaced them with 9.0. Because the agent who have maximum number of booking is 9.0 considering all the booking consisting null values in agent column is done by the agent 9.0.
- For the company column which contain 112593 rows with null values, I have dropped then company column. Because more then 95% rows in this column is filled with null values.
- After all this modifications of cleaning data I have a DataFrame containing 118902 rows and 31 columns



```
3   arrival_date_year               118902 non-null  int64
4   arrival_date_month              118902 non-null  object
5   arrival_date_week_number        118902 non-null  int64
6   arrival_date_day_of_month       118902 non-null  int64
7   stays_in_weekend_nights         118902 non-null  int64
8   stays_in_week_nights            118902 non-null  int64
9   adults                          118902 non-null  int64
10  children                        118902 non-null  float64
11  babies                          118902 non-null  int64
12  meal                            118902 non-null  object
13  country                         118902 non-null  object
14  market_segment                  118902 non-null  object
15  distribution_channel            118902 non-null  object
16  is_repeated_guest               118902 non-null  int64
17  previous_cancellations          118902 non-null  int64
18  previous_bookings_not_canceled  118902 non-null  int64
19  reserved_room_type              118902 non-null  object
20  assigned_room_type              118902 non-null  object
21  booking_changes                 118902 non-null  int64
22  deposit_type                    118902 non-null  object
23  agent                           118902 non-null  float64
24  days_in_waiting_list            118902 non-null  int64
25  customer_type                   118902 non-null  object
26  adr                             118902 non-null  float64
27  required_car_parking_spaces     118902 non-null  int64
28  total_of_special_requests       118902 non-null  int64
29  reservation_status              118902 non-null  object
30  reservation_status_date         118902 non-null  object
```

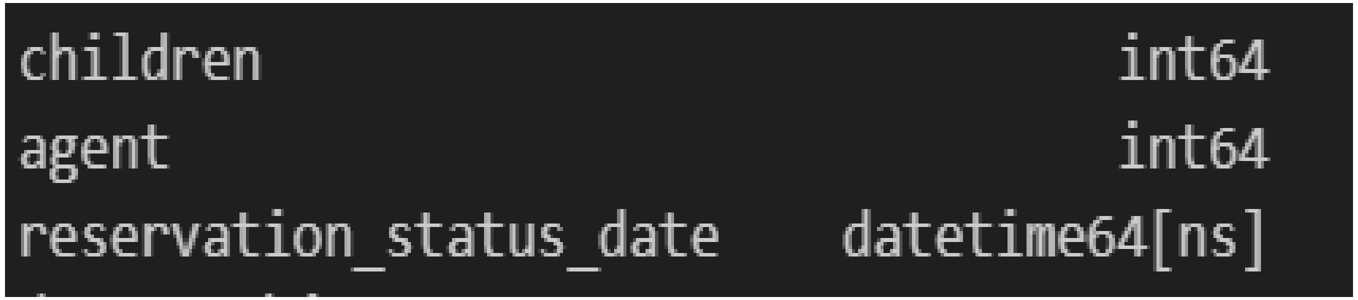
4. EDA - Correcting Data Type

- Some of the columns in the DataFrame like children, agent, reservation_status_date having incorrect Data Type



```
children                    float64
agent                       float64
reservation_status_date      object
```

- After corrected the data Type:

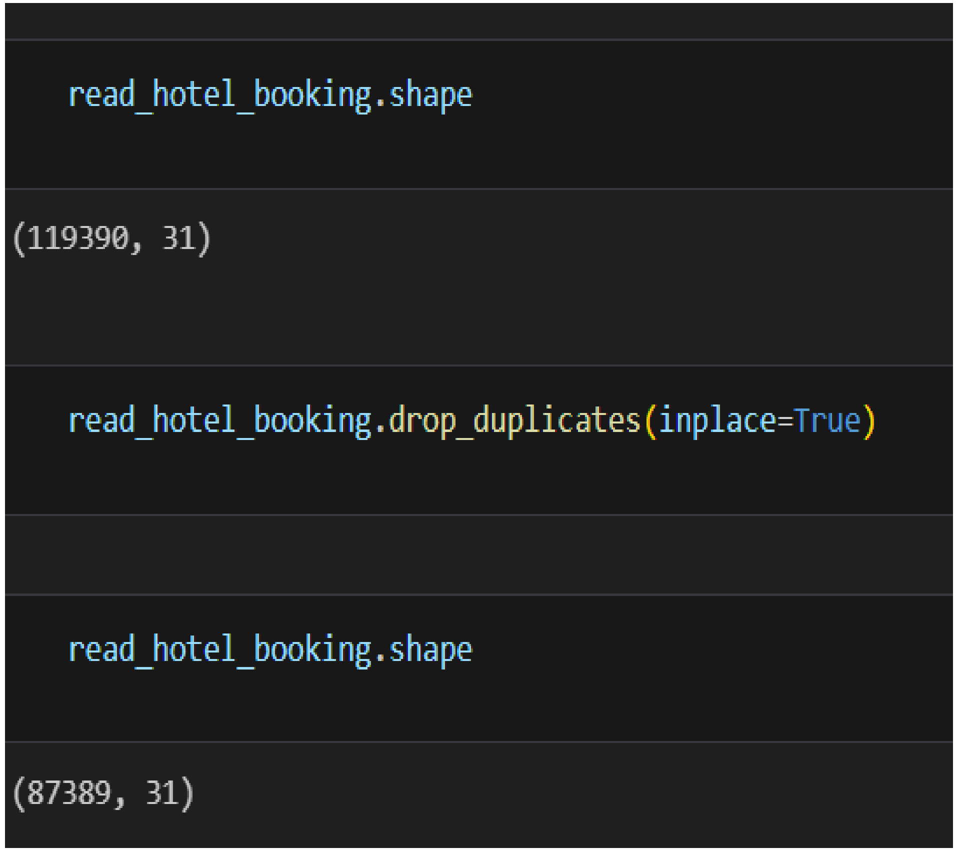


```
children                              int64
agent                                 int64
reservation_status_date     datetime64[ns]
```

5. EDA - Removing Duplicate Rows

- Before Removing duplicate rows total number of rows and columns are (118902, 31)
- After Removing duplicate rows total number of rows and columns are (86937, 31) using drop_duplicate() function.

- So, the total number of rows and column after Analyzing and Cleaning Data and correctin data type are (87389, 31). This data have no null value, no duplicate data, no incorrect data type.



```
read_hotel_booking.shape
```

```
(119390, 31)
```

```
read_hotel_booking.drop_duplicates(inplace=True)
```

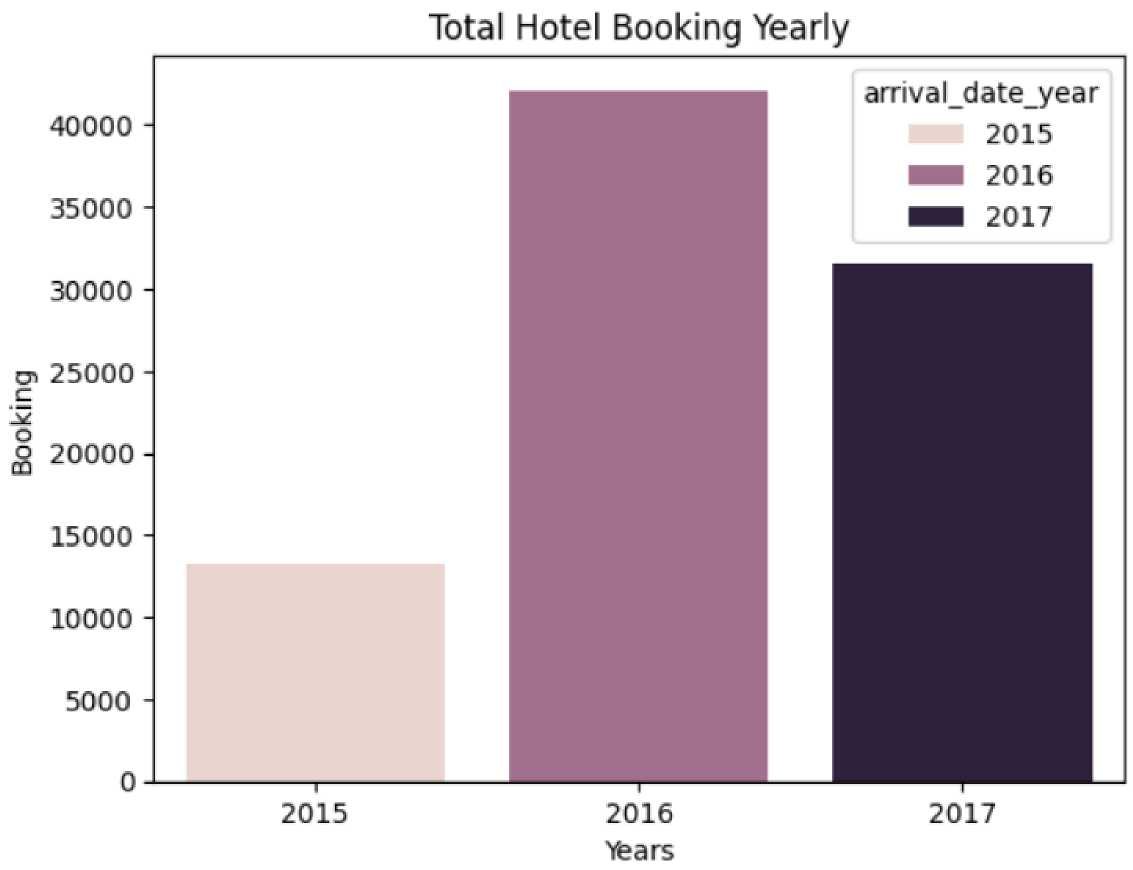
```
read_hotel_booking.shape
```

```
(87389, 31)
```

6. EDA – Data Visualization

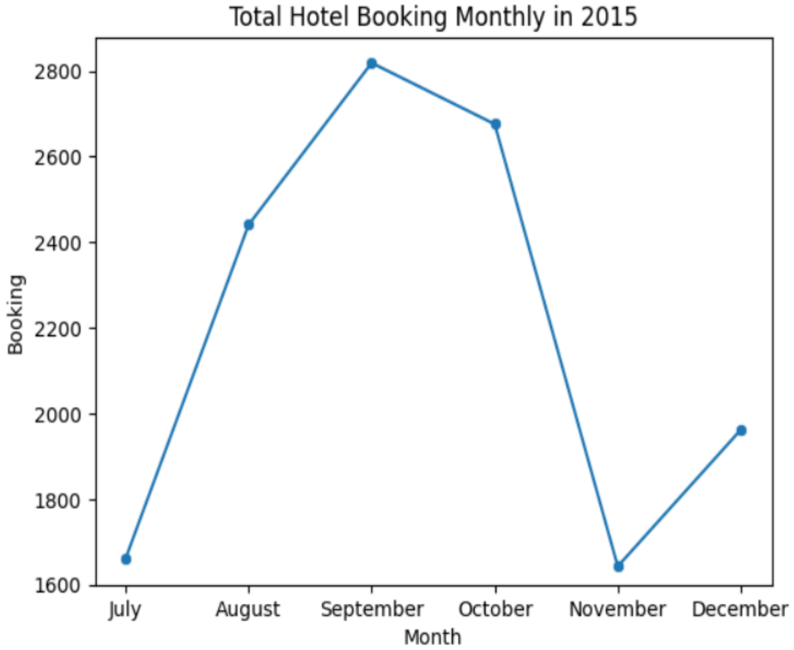
1. What are the total number of hotel booked Yearly?
2. Total Monthly Booking in 2015, 2016 & 2017?
3. Which Type of Customer Booked Maximum Hotels?
4. Which Top Country Makes the most reservation?
5. Which Hotel type is the Busiest?
6. Total number of reservation cancelled in each type of Hotel?
7. Which Market Segment has maximum and Minimum number of booking?
8. Top 5 agent makes the most number of booking.
9. How many customer don't wish to make a booking with a pre-deposit.
10. Total Number of Repeated Guest both hotels combined.
11. Correlation Heatmap of Numerical columns
12. Pair plot showing correlation between Guests and Booking Length.

Data Visualization 1
What are the total number of hotel booked Yearly?



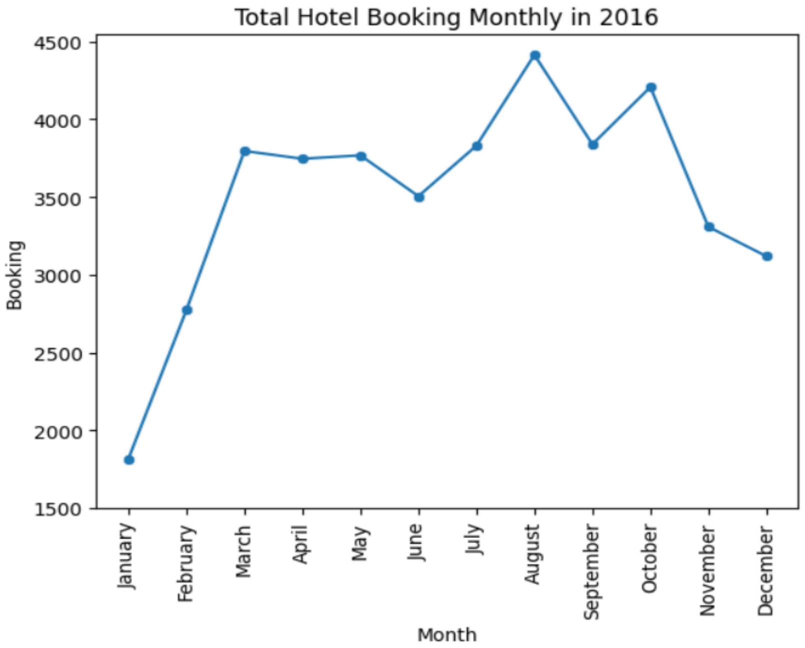
The maximum booking Yearly is in 2016 with approx. 40000 bookings as per the Dataset

Data Visualization 2
Total Monthly Booking in 2015, 2016 & 2017?



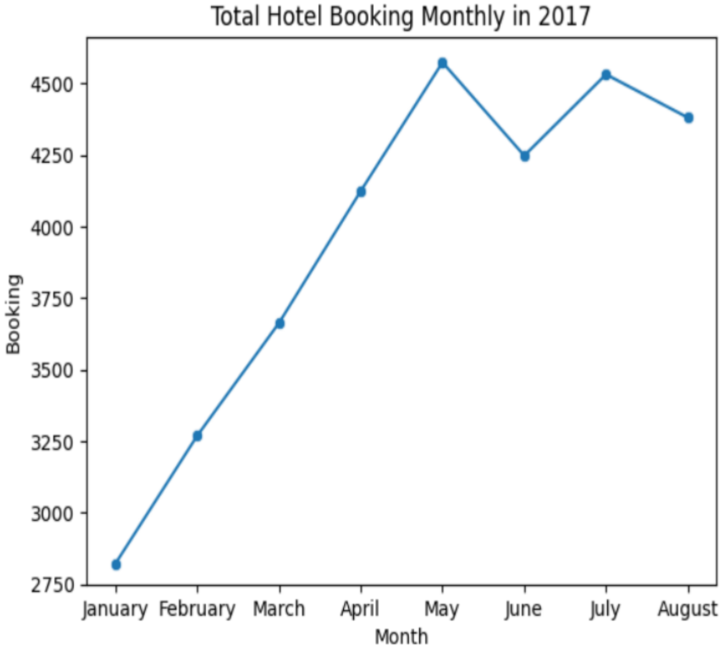
In 2015 from July till December -

- Maximum number of booking is in September with approx. 2800 bookings
- Minimum number of booking in July & November with Approx. 1600 bookings.



In 2016 -

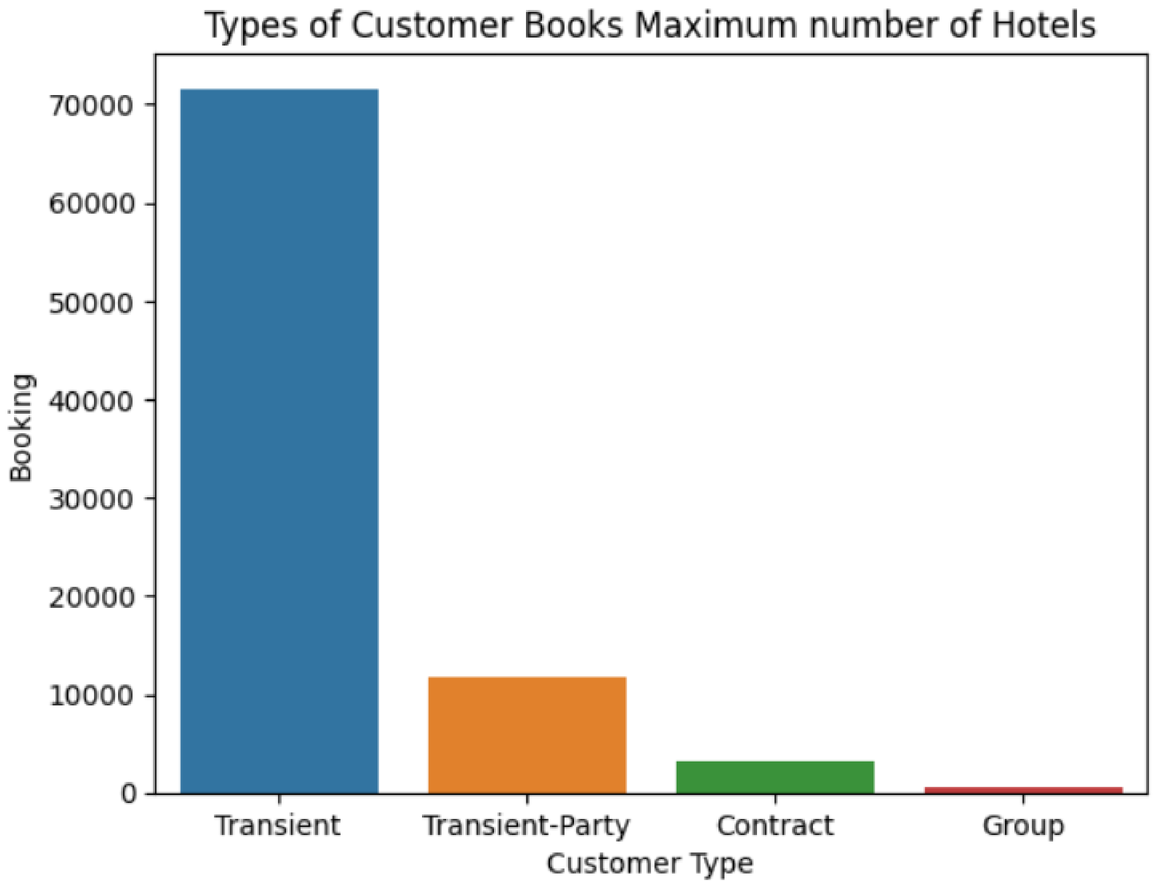
- Maximum number of booking is in August with approx. 4400 bookings
- Minimum number of booking in January with Approx. 1800 bookings.



In 2017 from January till August -
* Maximum number of booking is in May with approx. 4500 bookings
* Minimum number of booking in July & November with Approx. 2800 bookings.

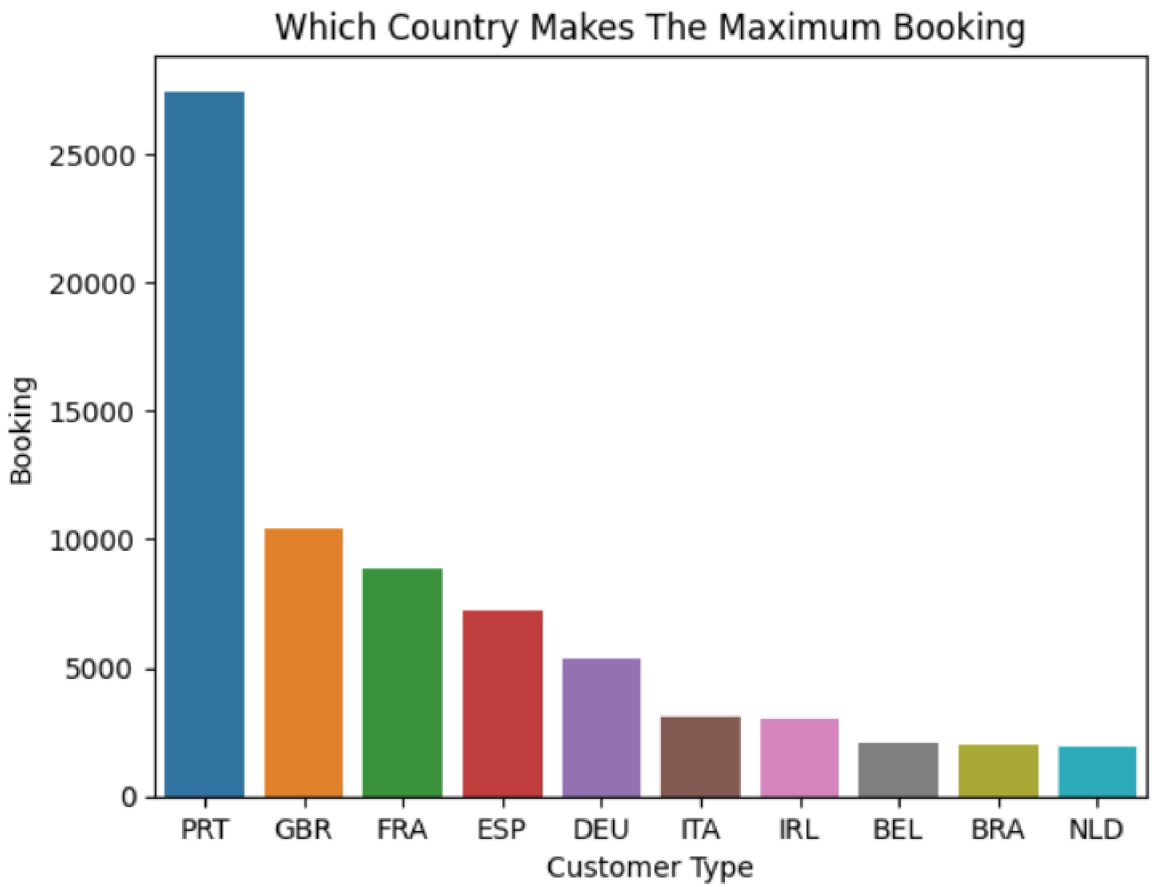
The busiest months for hotels are August, September and October.

Data Visualization 3
Which Type of Customer Booked Maximum Hotels?



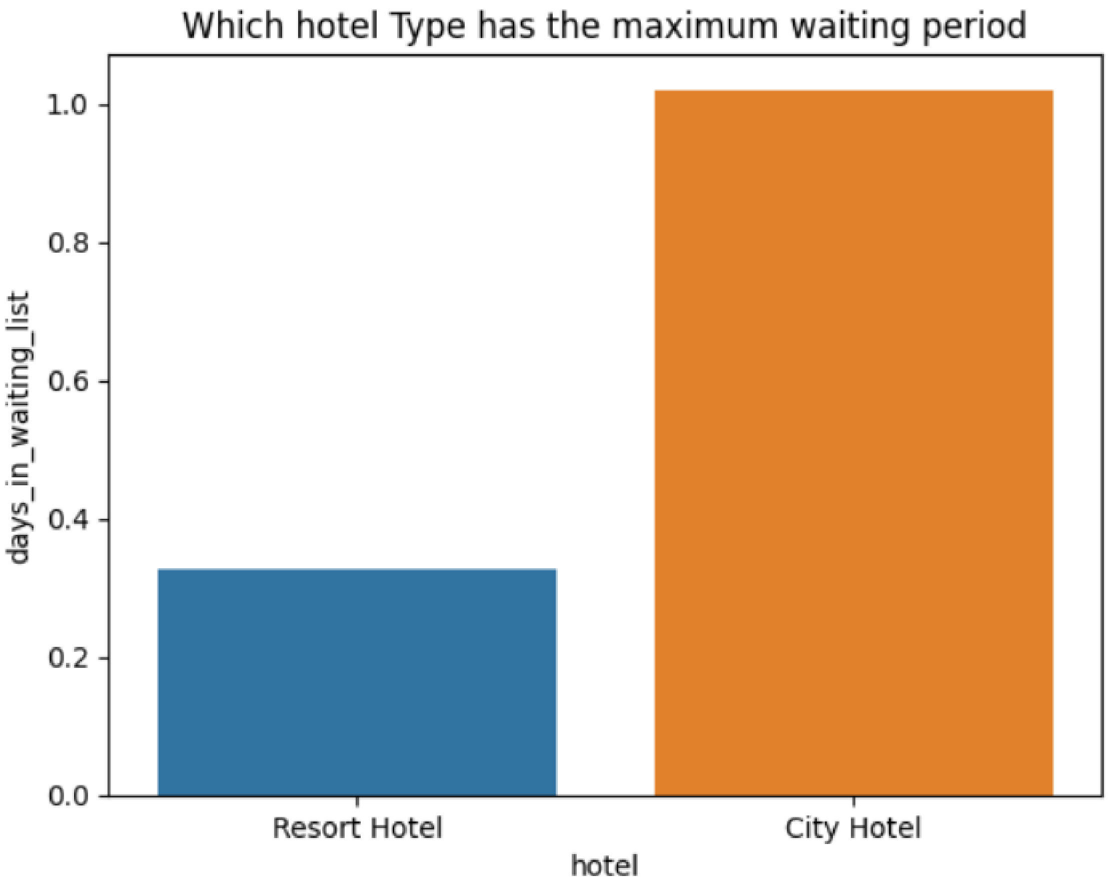
The type of customer who booked maximum number of booking – Transient with Approx. 70000

Data Visualization 4
Which Top Country Makes the most reservation?



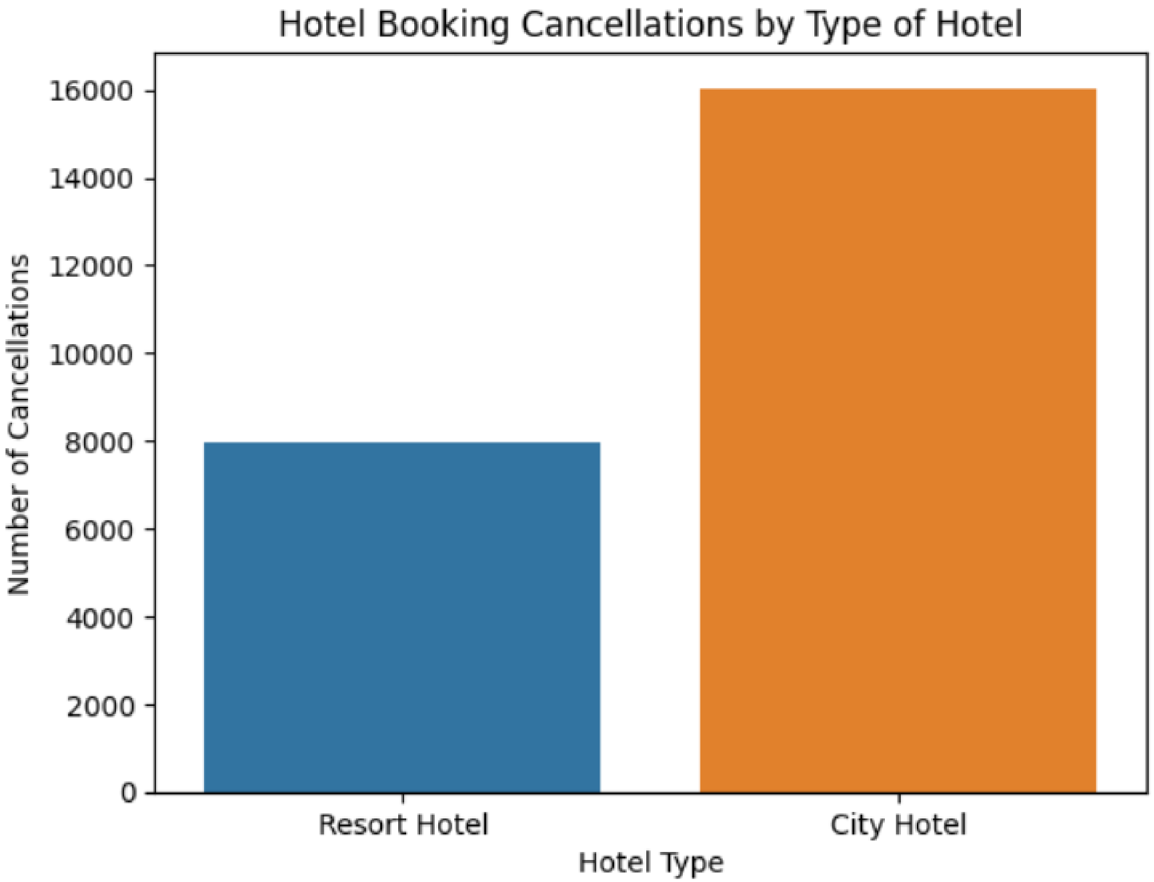
The majority of reservations are made through country PRT. Customers make the most bookings in the following top 5 countries: PRT, GBR, FRA, ESP, and DEU.

Data Visualization 5
Which Hotel type is the Busiest?



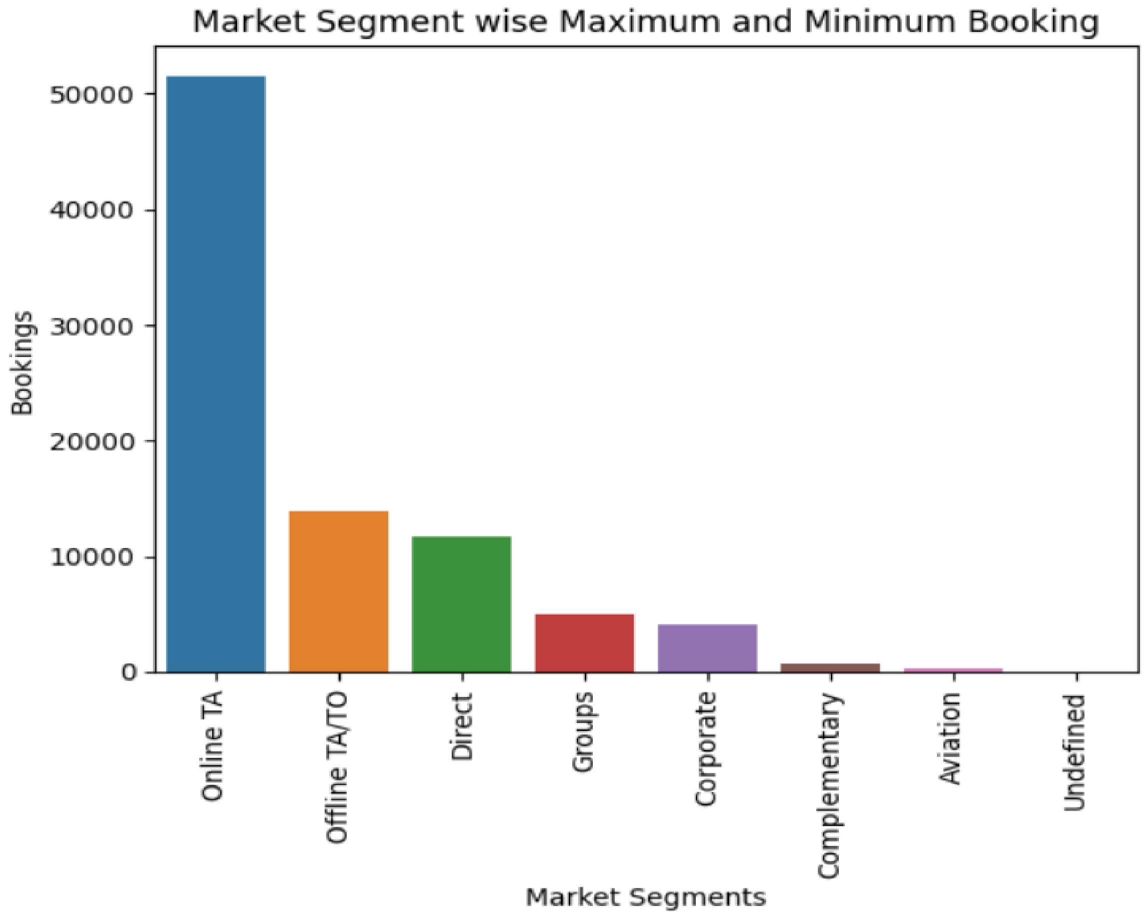
A City Hotel is more busier as compare to Resort Hotel.

Data Visualization 6
Total number of reservation cancelled in each type of Hotel?



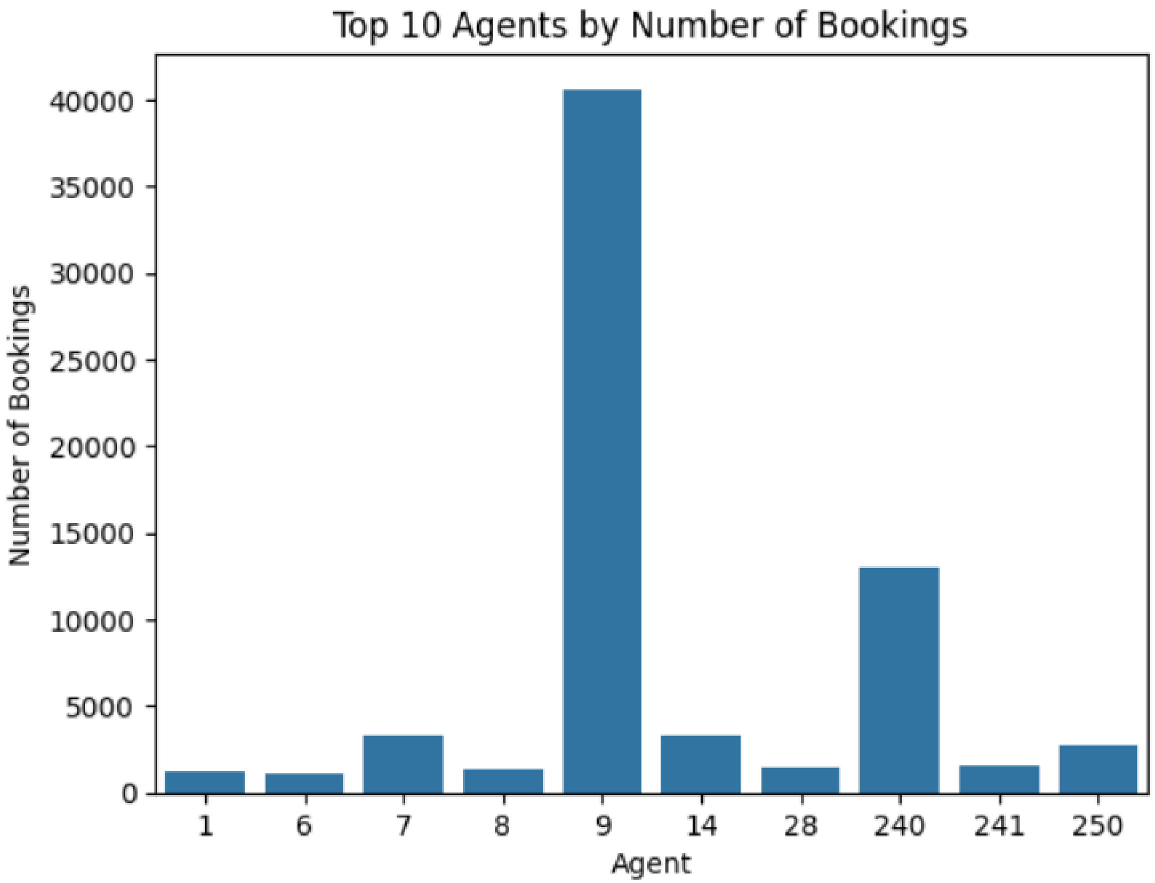
City Hotel have maximum number of booking cancelled as compared to Resort Hotel

Data Visualization 7
Which Market Segment has maximum and Minimum number of booking?



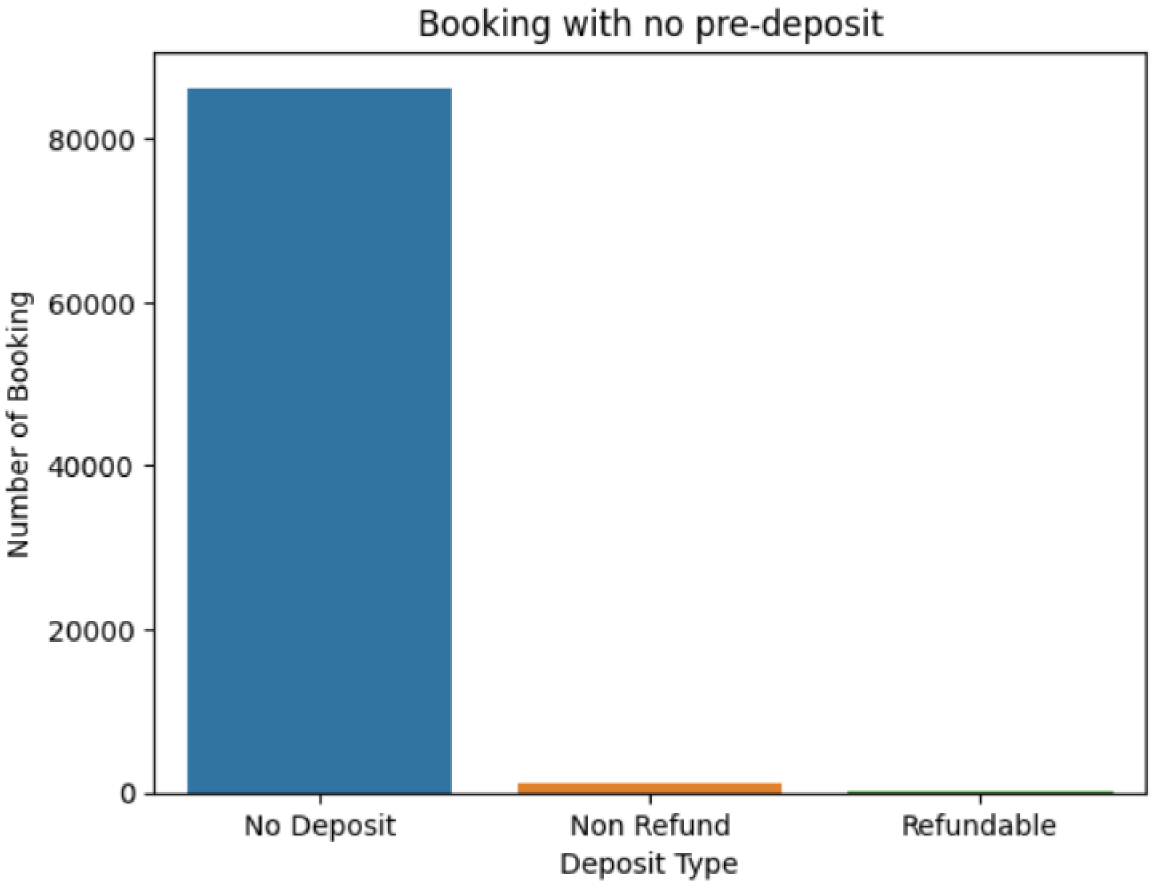
Market Segment wise Online TA have maximum number of Booking and Aviation have approximately Minimum number of booking

Data Visualization 8
Top 5 agent makes the most number of booking.



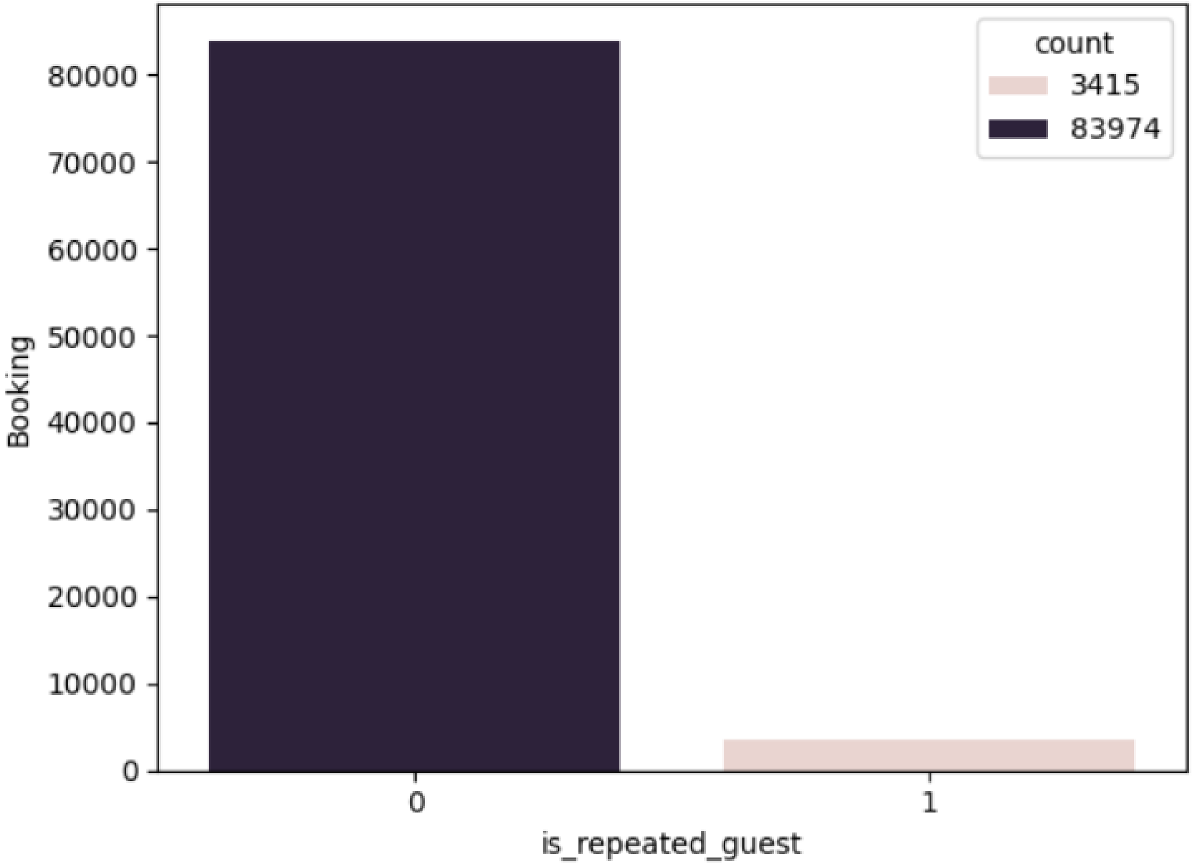
Agent number 9 made most number of bookings. 9, 240, 7, 14 and 250 are the top 5 agents by number of bookings made.

Data Visualization 9
How many customer don't wish to make a booking with a pre-deposit.



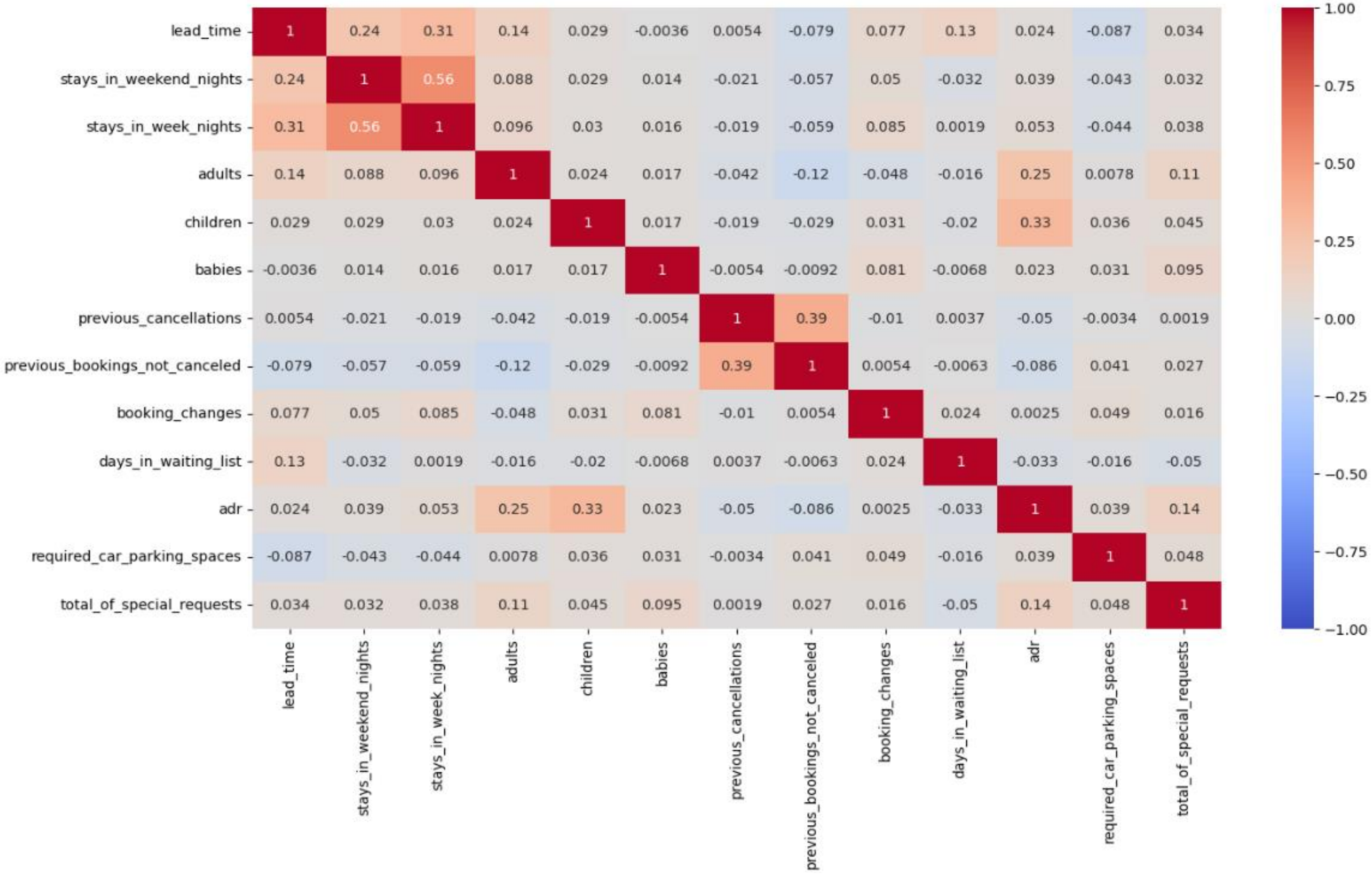
Customers do not wish to make a bookings with a pre-deposit.

Data Visualization 10
Total Number of Repeated Guest both hotels combined.



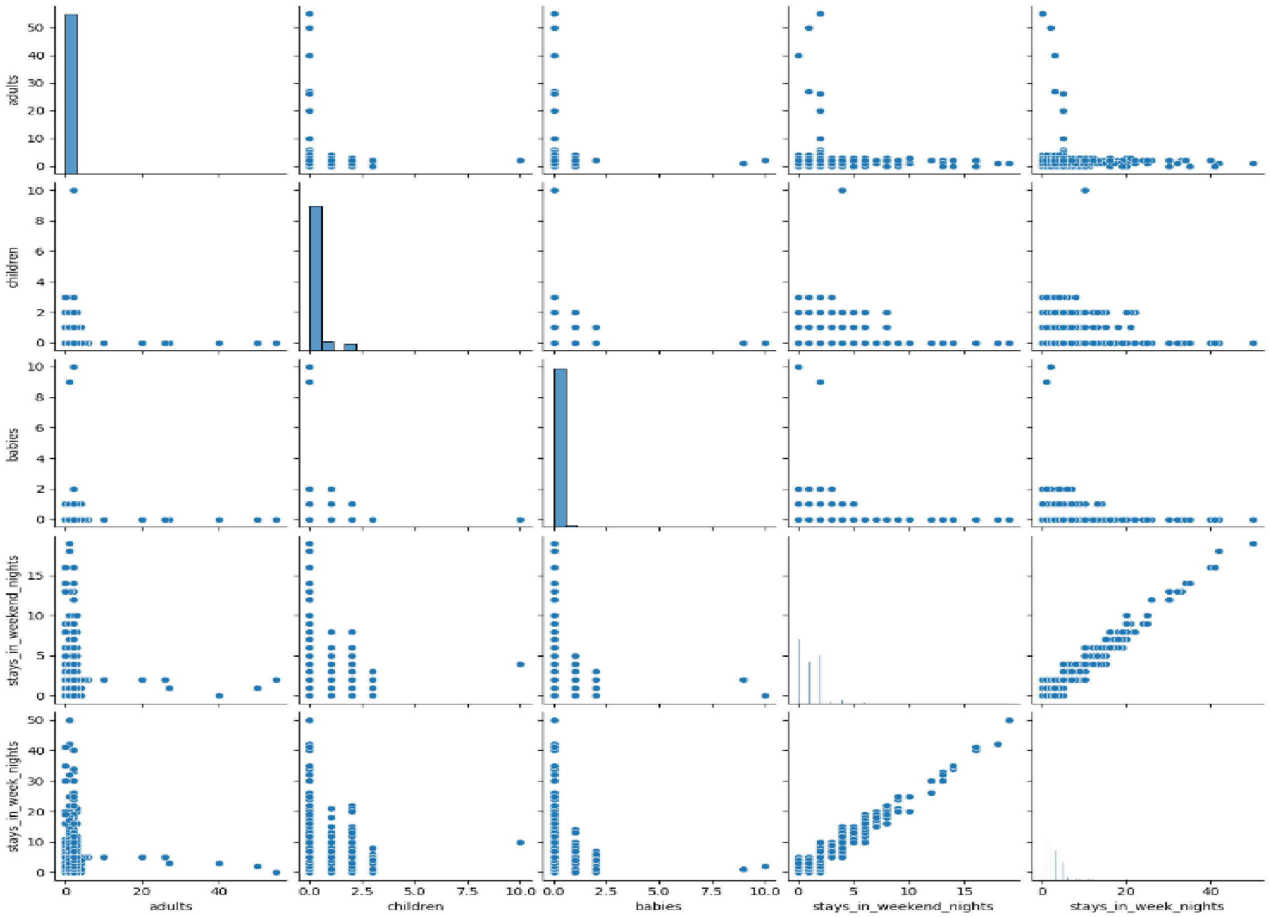
- The Maximum hotel bookings are made by new guests. Only less then 5% guests returned around 3400 approx.

Data Visualization 11
Correlation Heatmap of Numerical columns



- Guests often book multiple nights, especially for weekends.
- Longer lead times can result in longer wait times.
- Guests with more special requests are less likely to need parking.
- Bookings with more adults tend to have higher rates.
- Guests who make more changes might have shorter wait times.

Data Visualization 12
Pair plot showing correlation between Guests and Booking Length.



- Guests often book multiple nights, both on weekends and weekdays.
- that bookings with more adults often include children as well.

Conclusion

- The top country with the most number of bookings is PRT.
- The number one agent with the most number of bookings is 9.
- The Maximum hotel bookings are made by new guests. Only less then 5% guests returned.
- The Online (internet) platform is used to make the majority of bookings.
- A city hotel is busier than a resort.
- The busiest months for hotels are August, September and October.
- Customers do not wish to make a bookings with a pre-deposit.
- City Hotel have maximum number of booking cancelled as compared to Resort Hotel
- Guests often book multiple nights, especially for weekends.
- Longer lead times can result in longer wait times.

Challenges Faced

- The data contained a large number of duplicates.
- It was challenging to select the best visualization techniques.
- The dataset contained a large number of null values.
- The improper data type format was used for the data.

Thank You…

Name: Debarpan Chowdhury
Email: debarpancmm725@gmail.com
Linkedin: www.linkedin.com/in/debarpan-chowdhury-65093820b
GitHub: https://github.com/Debarpan200

GitHub Project File Link:
https://github.com/Debarpan200/EDA-Hotel-Booking-Analysis