Exploratory Data Analysis on Hotel Booking Dataset

DIPANJAN MAITY

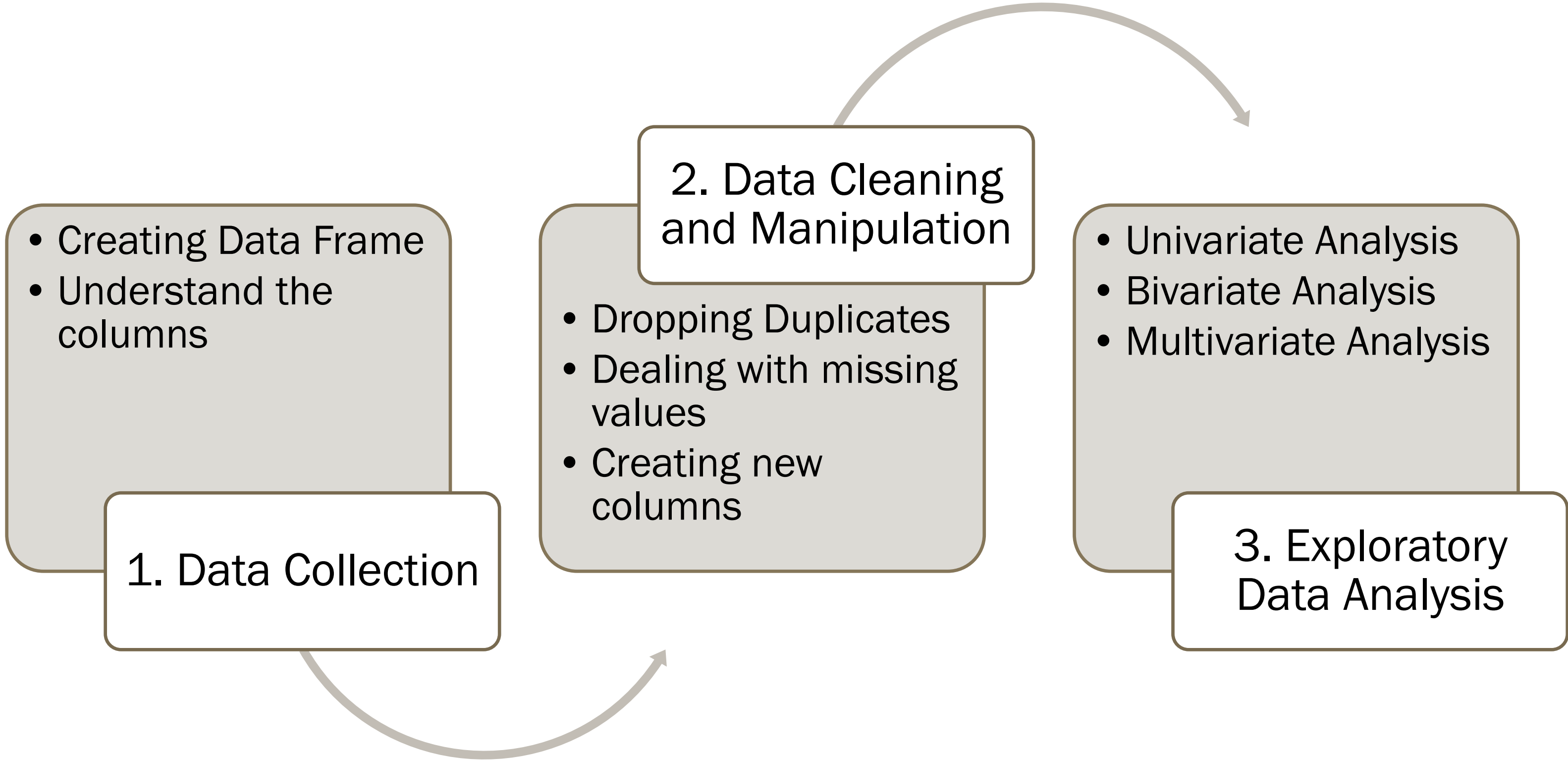
Problem Statement

❑ Bookings in the hotel sector are highly dependent on a variety of variables.

❑ Dataset from hotel bookings will be analyzed for this project. This dataset 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 and many things.

❑ The major goal of this project is to investigate and analyze data in order to identify significant elements that influence reservations and provide knowledge to hotel management so that it may implement various campaigns to increase sales and performance.

Process Chart



1. Data Collection

- Creating Data Frame
- Understand the columns

2. Data Cleaning and Manipulation

- Dropping Duplicates
- Dealing with missing values
- Creating new columns

3. Exploratory Data Analysis

- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

1. Data Collection and Examination

- ❑ I have imported the dataset into a Data frame using Pandas Library. This dataset had 119390 rows and 32 columns.

- ❑Now let's examine these 32 columns.
  - ❑ **hotel :** 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 into the PMS 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 week nights.
  - ❑ **adults :** Number of adults
  - ❑ **children :** Number of children

1. Data Collection and Examination

❑ **babies :** Number of babies

❑ **meal :** Type of meal booked.

❑**country :** Country of origin.

❑ **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 cancelled by customer prior to the current booking

❑**previous_bookings_not_canceled :** No of previous bookings not cancelled by 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 :** No of changes made to the booking from the booking was entered until check-in or cancellation.

❑ **deposit_type :** No Deposit, Non Refund , Refundable.

❑ **agent :** ID of the travel agency that made the booking

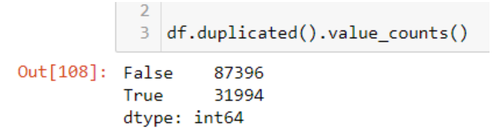
❑ **company :** ID of the company/entity that made the booking .

1. Data Collection and Examination

- ❑ **days_in_waiting_list :** Number of days the booking was in 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.

2. Data Cleaning and Manipulation

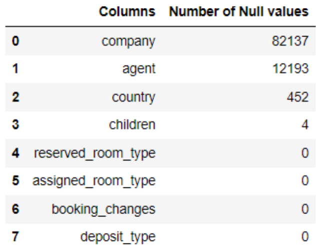
- ❑ The Dataset had 31994 duplicates value, So I dropped all the duplicate values from the dataset



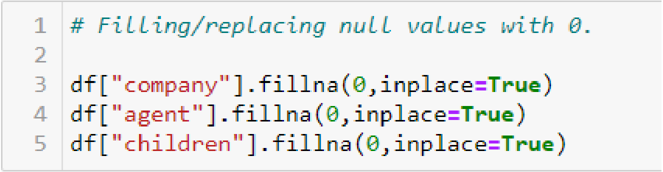
```

df.duplicated().value_counts()
```

```
Out[108]: False    87396
          True     31994
          dtype: int64
```

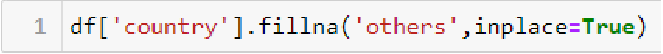
- ❑ The Dataset had 4 columns with missing values , I filled them with “0” or other appropriate values



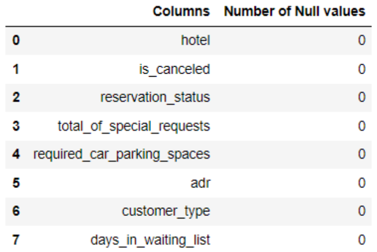
	Columns	Number of Null values
0	company	82137
1	agent	12193
2	country	452
3	children	4
4	reserved_room_type	0
5	assigned_room_type	0
6	booking_changes	0
7	deposit_type	0



```python
# Filling/replacing null values with 0.

df["company"].fillna(0,inplace=True)
df["agent"].fillna(0,inplace=True)
df["children"].fillna(0,inplace=True)
```

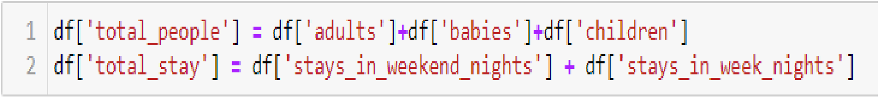


```python
df['country'].fillna('others',inplace=True)
```



	Columns	Number of Null values
0	hotel	0
1	is_canceled	0
2	reservation_status	0
3	total_of_special_requests	0
4	required_car_parking_spaces	0
5	adr	0
6	customer_type	0
7	days_in_waiting_list	0

- ❑ I have Created 2 new column
  - ❑ ‘Total_People’ : adding all the guest including Children, adults, babies.
  - ❑ ‘Total_stay’ : adding all the day including weekend nights and weekdays night



```python
df['total_people'] = df['adults']+df['babies']+df['children']
df['total_stay'] = df['stays_in_weekend_nights'] + df['stays_in_week_nights']
```

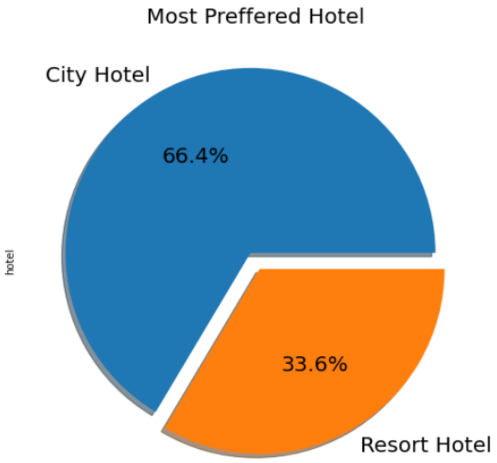
Exploratory Data Analysis (EDA)

- ❑ **Univariate analysis:** Univariate analysis is the simplest of the three analyses where the data I am analyzing depends on only one variable.

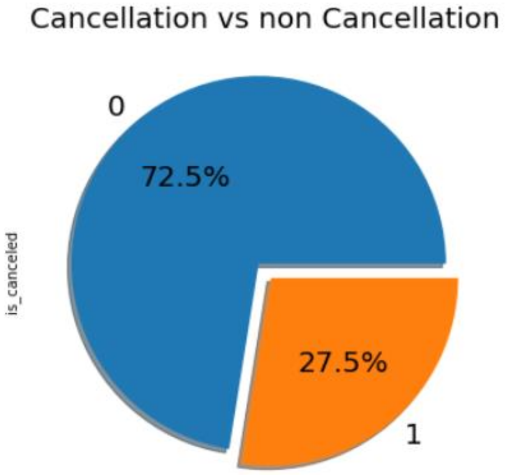
- ❑ **Bivariate analysis:** Bivariate analysis is where I am comparing two variables to study their relationships.

- ❑ **Multivariate analysis :** Multivariate analysis is similar to Bivariate analysis but I am comparing more than two variables.

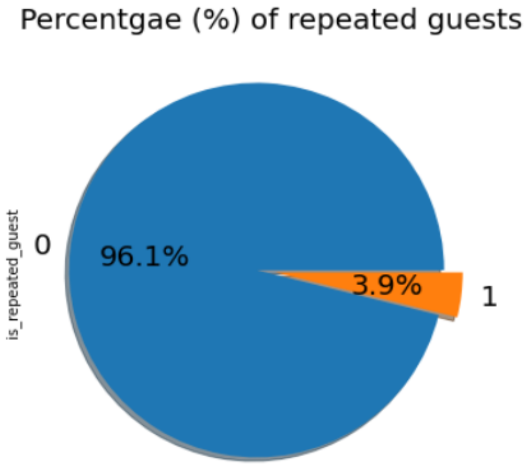
Univariate Analysis



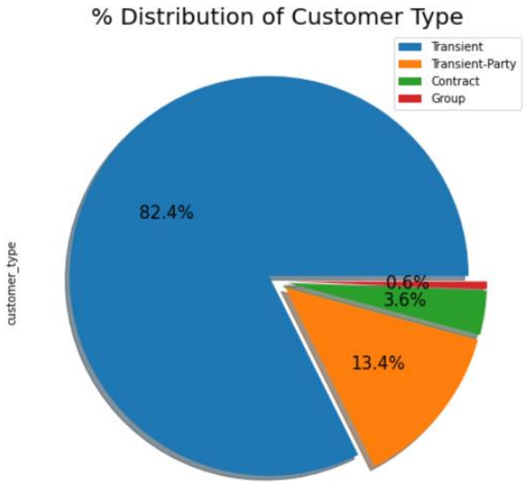
Most Preffered Hotel



Cancellation vs non Cancellation



Percentgae (%) of repeated guests



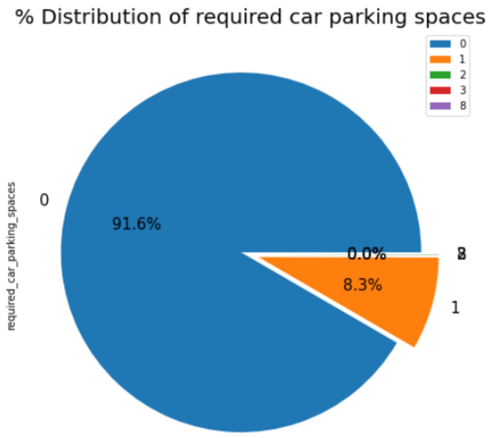
% Distribution of Customer Type

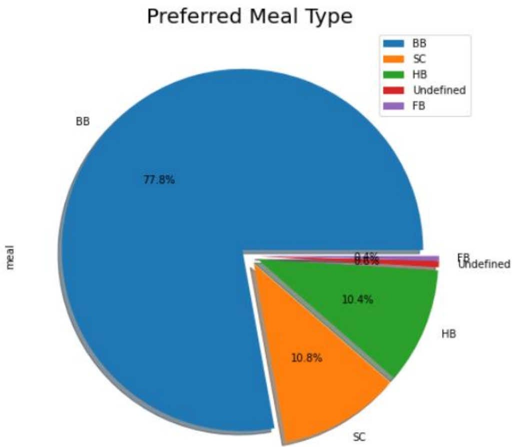
Findings:

1. Contract : when the booking has an allotment or other type of contract associated to it
2. Group : when the booking is associated to a group
3. Transient : when the booking is not part of a group or contract, and is not associated to other transient booking
4. Transient-party : when the booking is transient, but is associated to at least other transient booking

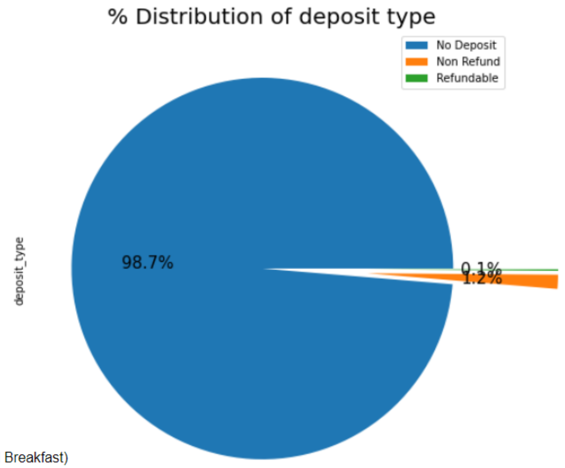
- ❑ We can clearly see that City Hotel is most preferred hotel by the visitors.
- ❑ 27.5 % of the bookings were cancelled out of all bookings.
- ❑ Only 3.90% of visitors are repeats, which is quite small.
- ❑ The percentage of transient customers is higher at 82.4%. A very small fraction of Bookings are connected to the Group.

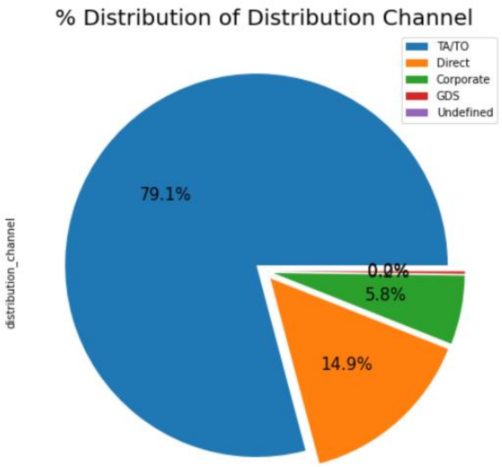
Univariate Analysis





- BB - (Bed and Breakfast)
- SC- (Self Catering)
- HB- (Half Board)
- FB- (Full Board)

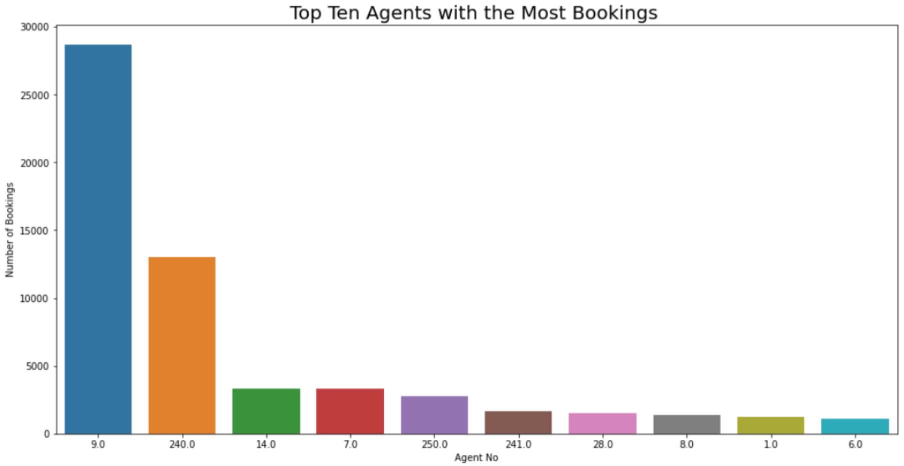


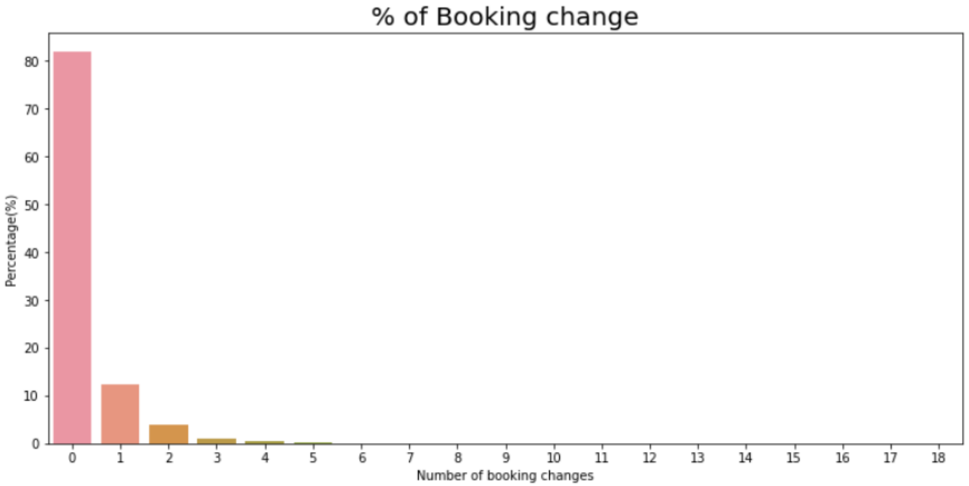


Findings :

- ❑ The parking space was not needed by 91.6% of the visitors. 8.3% of visitors only needed one parking space.
- ❑ Consequently, bed and breakfast(BB) is the most popular sort of meal among the visitors. Almost Equally desirable are HB- (Half Board) and SC- (Self Catering).
- ❑ The majority of visitors almost 98.7% prefer "No deposit" types of deposits which means the customer made no deposit to guarantee the booking.
- ❑ 79.1 % bookings were made through TA/TO (travel agents/Tour Operators).Second most channel is direct.

Univariate Analysis

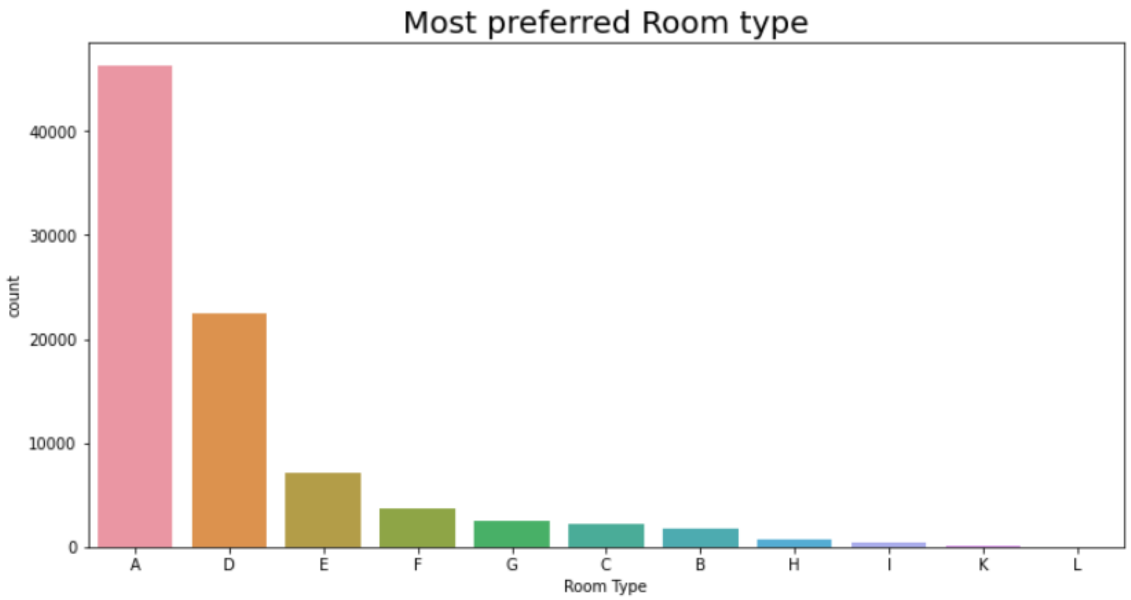




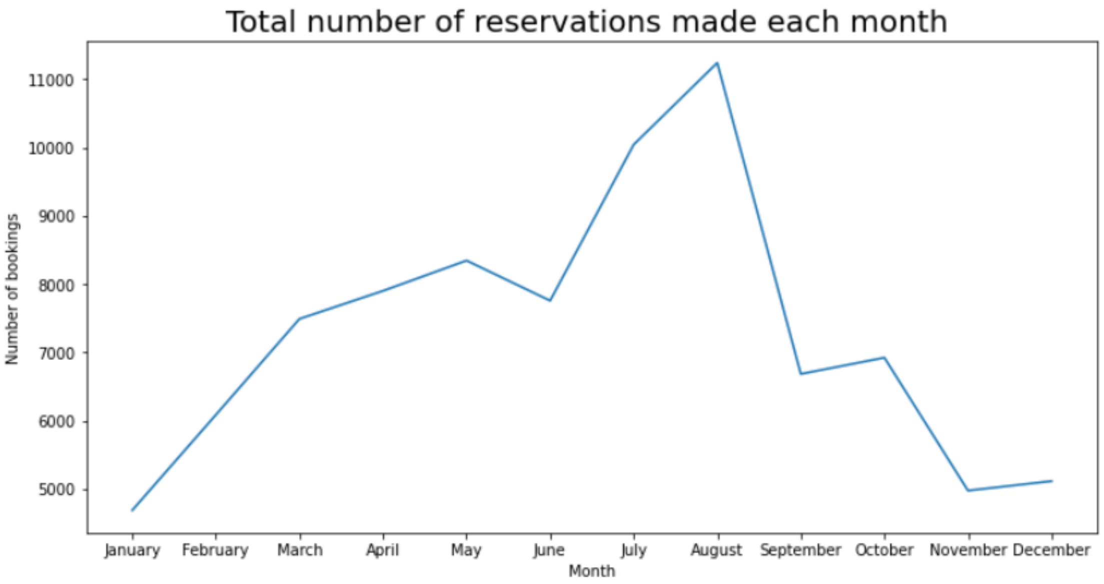
Findings :

- ❑ Top Agent is Agent ID no: 9 who has Most Number of Bookings which is 28721.
- ❑ Above 80% of the bookings were not changed by guests. Percentage of Single changes made was about 10%.

Univariate Analysis



Most preferred Room type

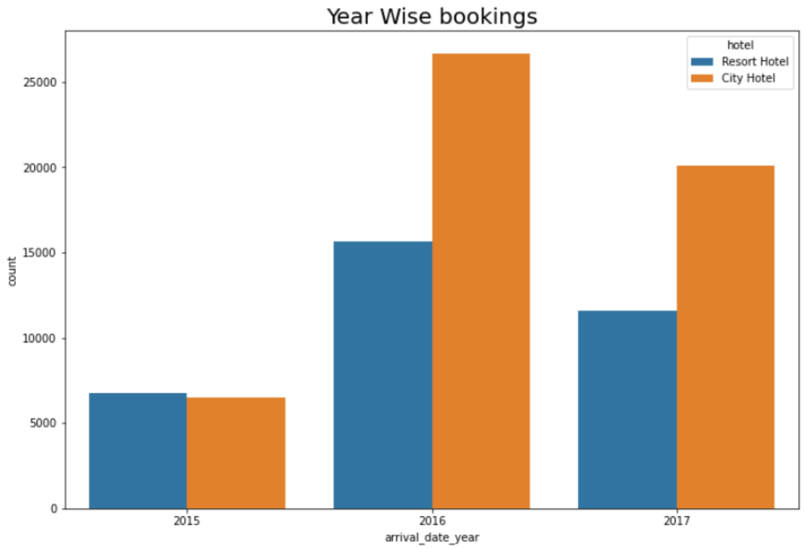


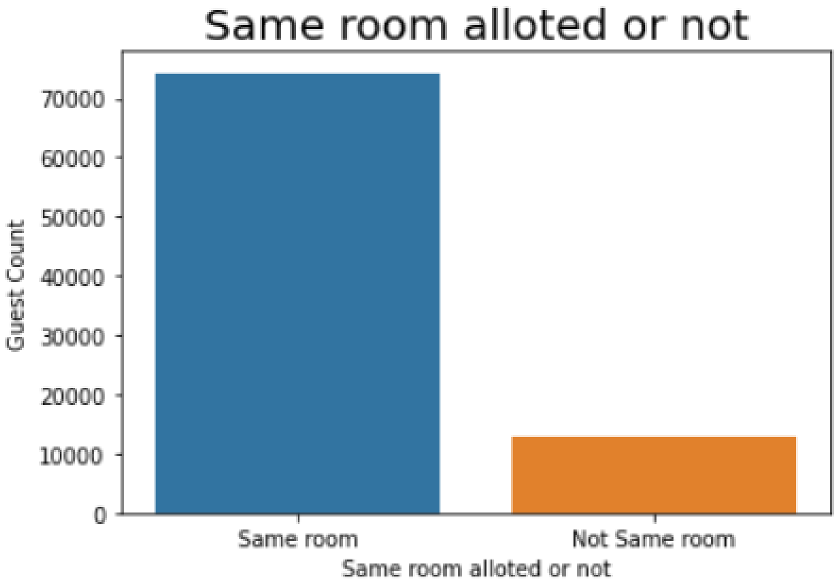
Total number of reservations made each month

Findings :

- ❑ The most preferred Room type is "A“ and second most preferred is ‘D’.
- ❑ The months with the most bookings were July and August. Bookings may have been made in anticipation of summer vacations.

Univariate Analysis





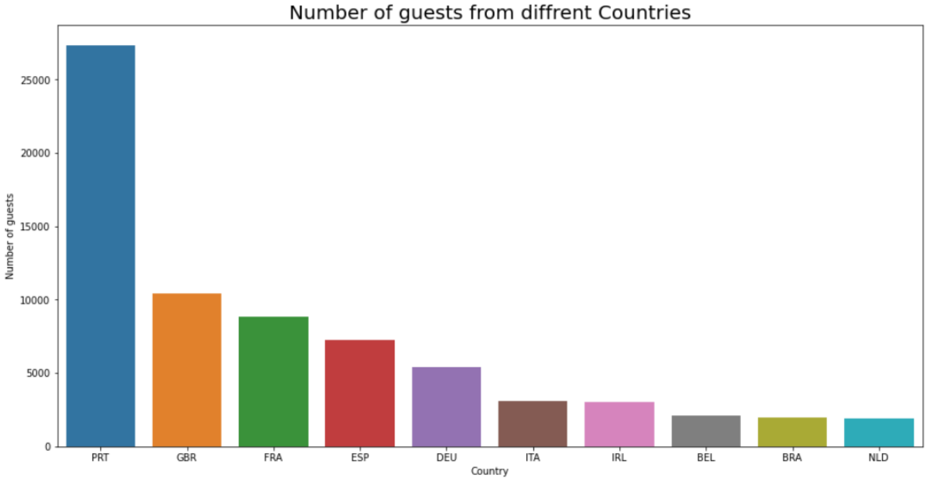
Findings :

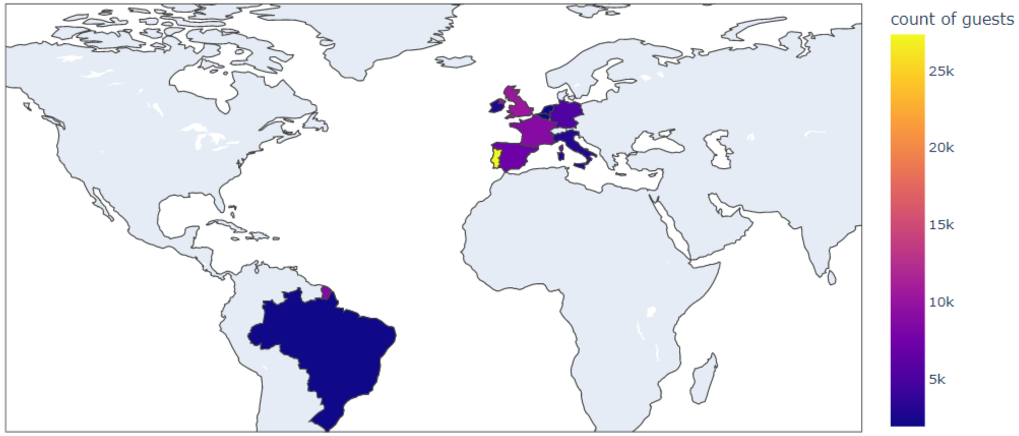
❑ City hotels had the most of the bookings.

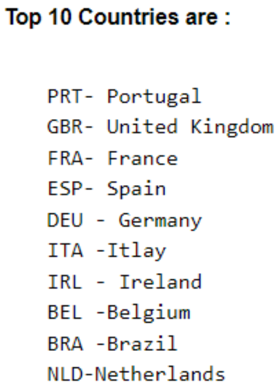
❑ Year 2016 had the highest bookings and 2015 had the lowest bookings.

❑ The majority of the time, guests receive the exact accommodation they have reserved.

Univariate Analysis







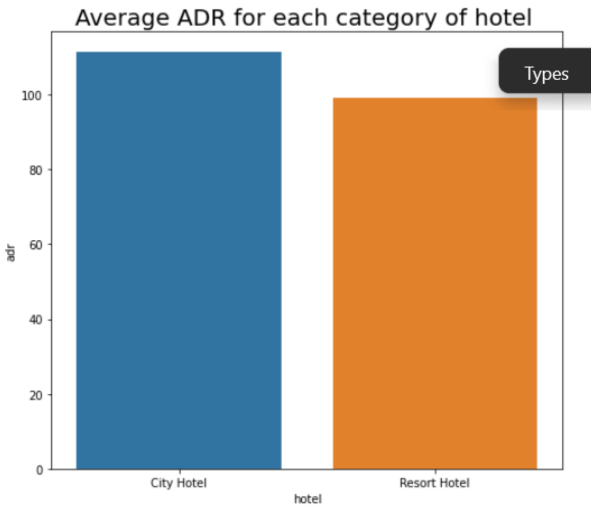
Top 10 Countries are :

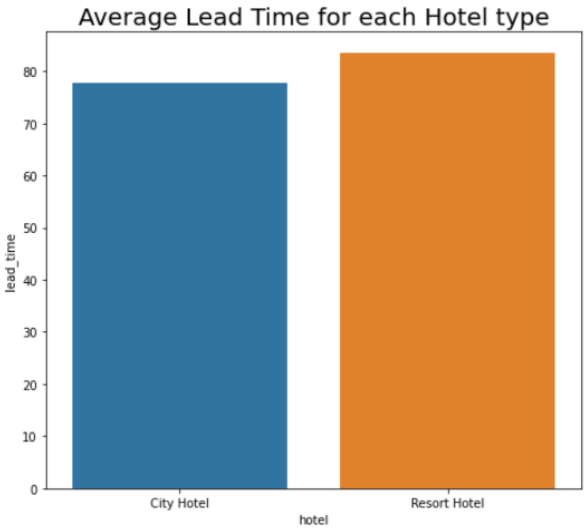
```
PRT- Portugal
GBR- United Kingdom
FRA- France
ESP- Spain
DEU - Germany
ITA -Itlay
IRL - Ireland
BEL -Belgium
BRA -Brazil
NLD-Netherlands
```

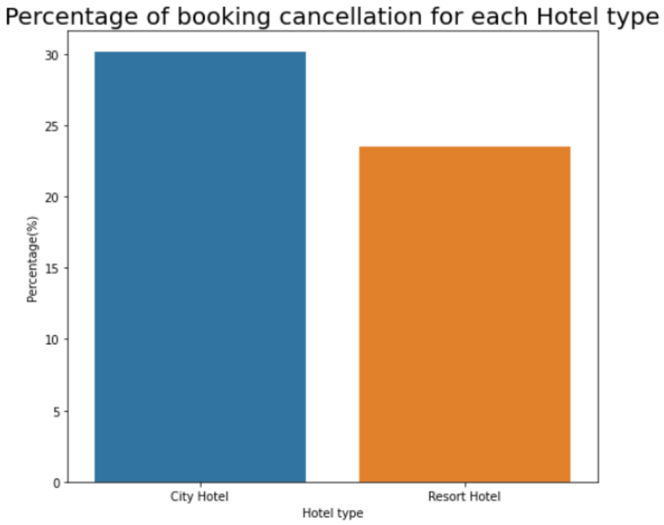
Findings :

- ❑ Most of the visitors are coming from Portugal.
- ❑ Following Portugal, GBR (Great Britain), France, and Spain are the nations from which the majority of the visitors originated.

Bivariate and Multivariate Analysis







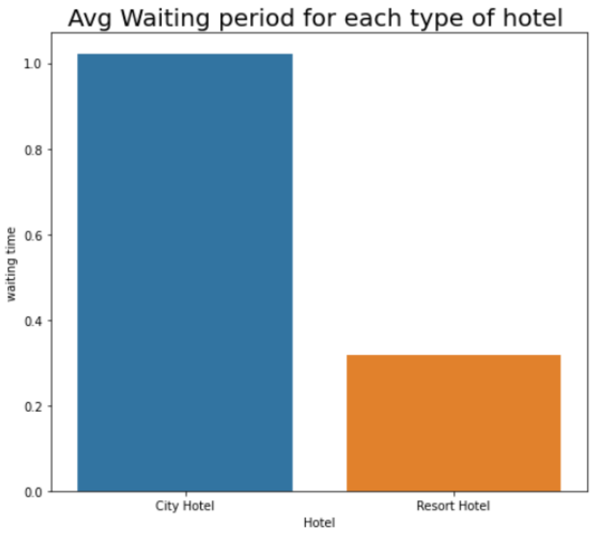
Findings :

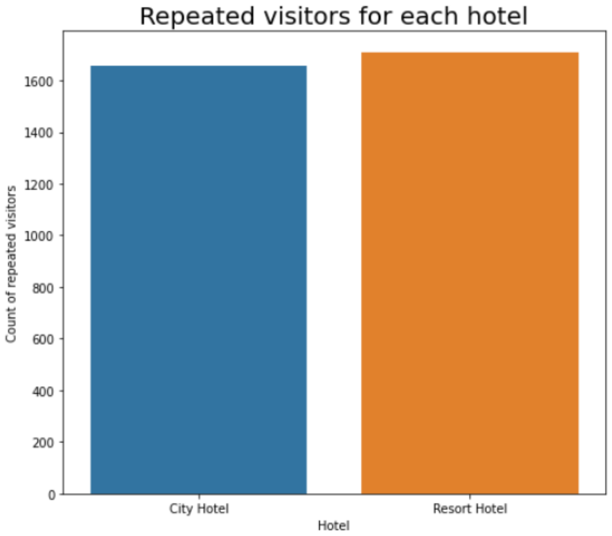
❑ The highest ADR is at the City Hotel. This indicates that city hotels make more money than resort hotels.

❑Hotel resorts have a slightly longer average lead time. Customers must make very early travel plans as a result.

❑ City hotel experiences the highest rate of cancellations.

Bivariate and Multivariate Analysis

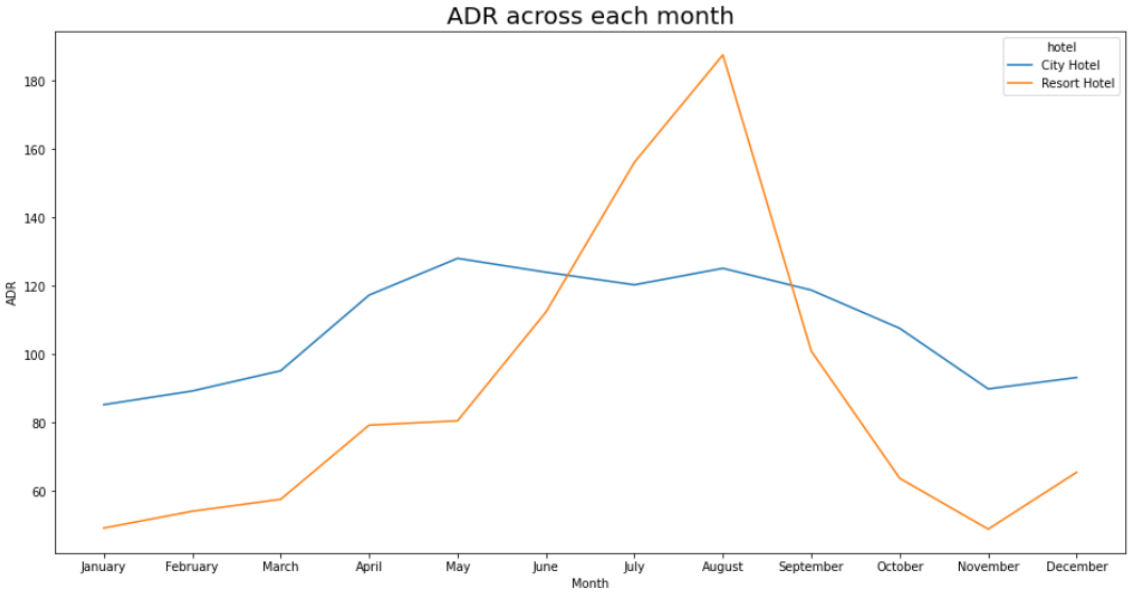

Avg Waiting period for each type of hotel


Repeated visitors for each hotel

Findings :

- There is a lengthier wait time at city hotels than at resort hotels. As a result, we can conclude that city hotels are significantly busier than resort hotels.
- Compared to City Hotels, Resort Hotel has a little bit more repeat visitors.

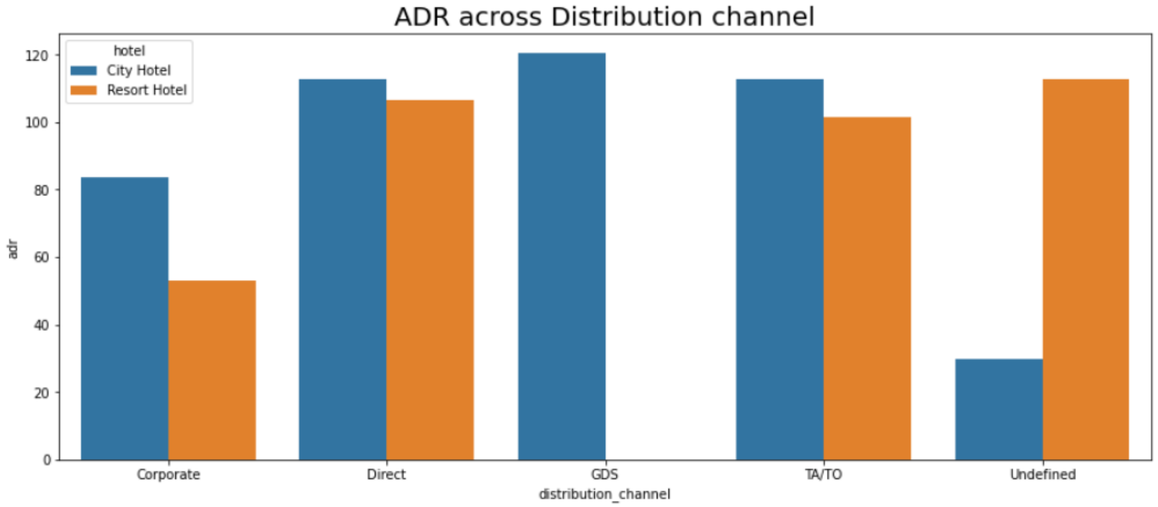
ADR (Average Daily Rate) Analysis

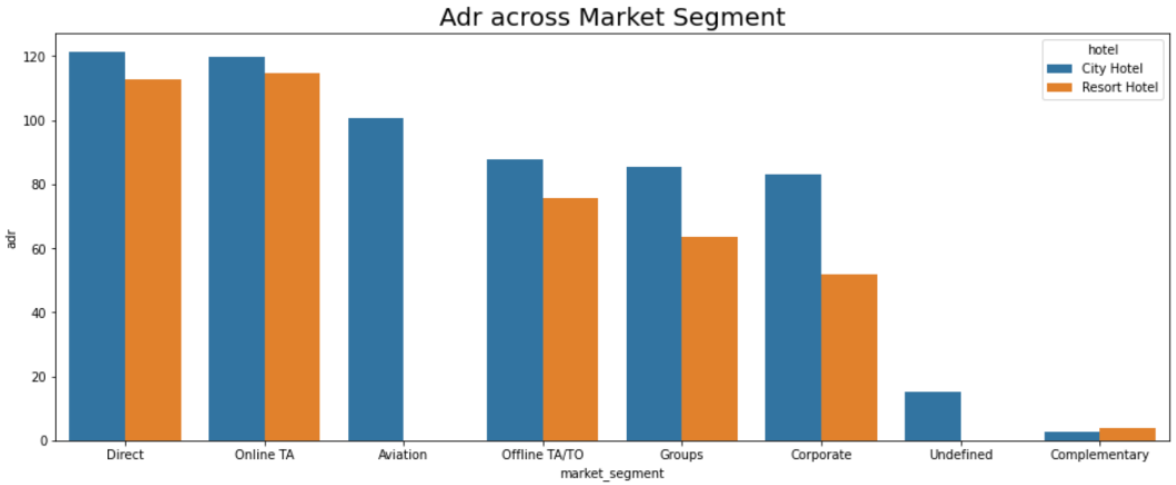


Findings :

- ❑ In comparison to City Hotels, the ADR for Resort Hotel is higher in the months of July and August. Perhaps people wish to vacation in resort hotels this summer.
- ❑ January, February, March, April, October, November, and December are the ideal months for visitors to resort or city hotels because of the low average daily rate throughout these months.

ADR (Average Daily Rate) Analysis

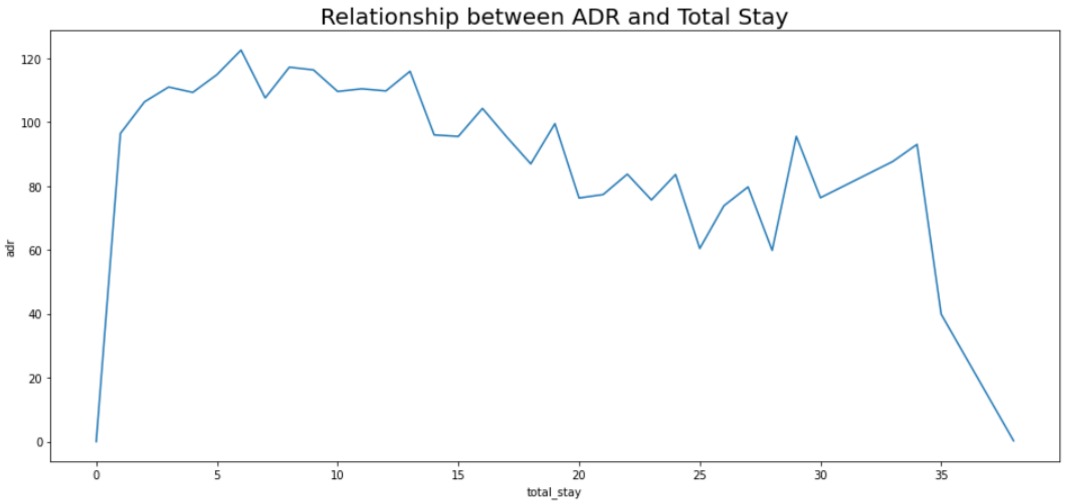


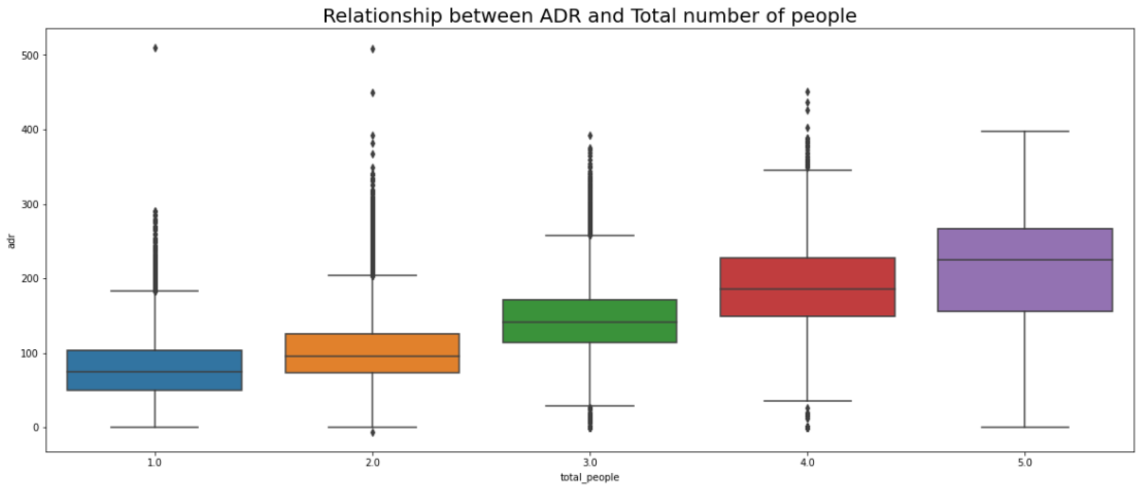


Findings :

- ❑ Distribution Channel:
  - ❑ "Direct" and "TA/TO" has almost equal ADR in both type of Hotel which is high among other channels
  - ❑ GDS scores highly in the "City Hotel" category. Bookings for Resort Hotels must rise at GDS. This indicates that "Direct" and "TA/TO" are outperforming the other channels in terms of revenue generation.
- ❑ Market Segment
  - ❑ In both types of hotels, "Direct" and "Online TA" are making the most contributions.

ADR (Average Daily Rate) Analysis


Relationship between ADR and Total Stay

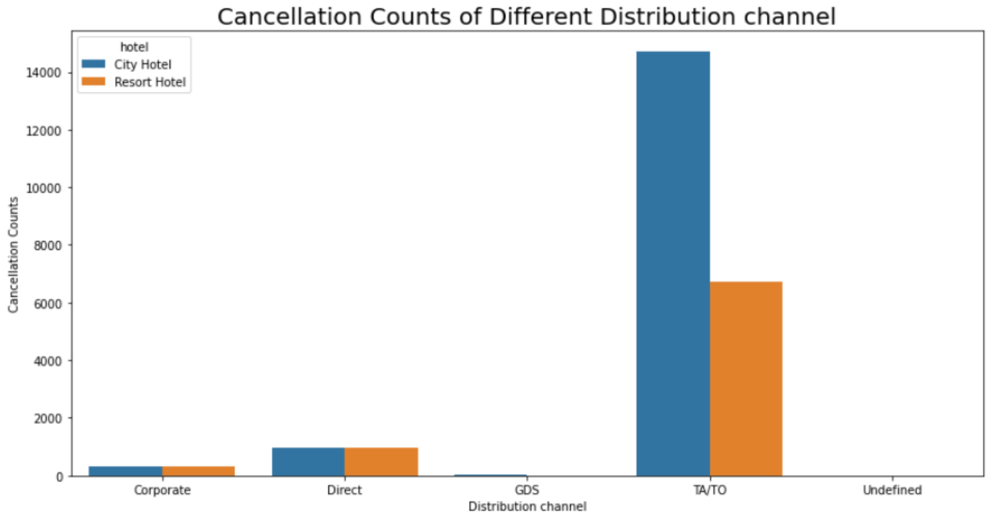

Relationship between ADR and Total number of people

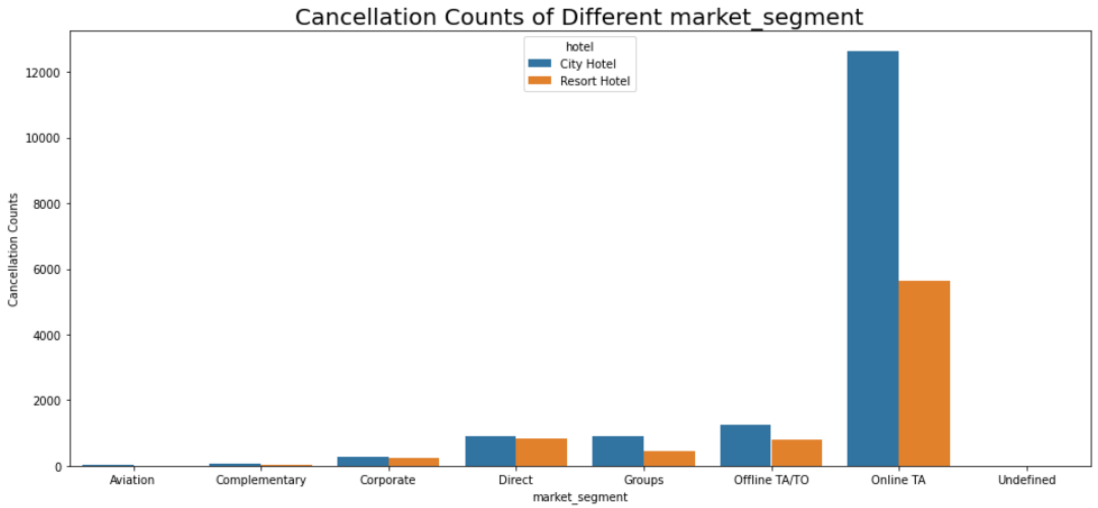
Findings :

❑ The ADR (average daily rate) rises along with the length of stay. However, the ADR decreases if a guest stays for a longer time.

❑ ADR also rises in proportion to the Total number of people. Thus more the people more will revenue of the hotels.

Cancellation Count Analysis


Cancellation Counts of Different Distribution channel


Cancellation Counts of Different market_segment

Findings :

❑ Distribution Channel:

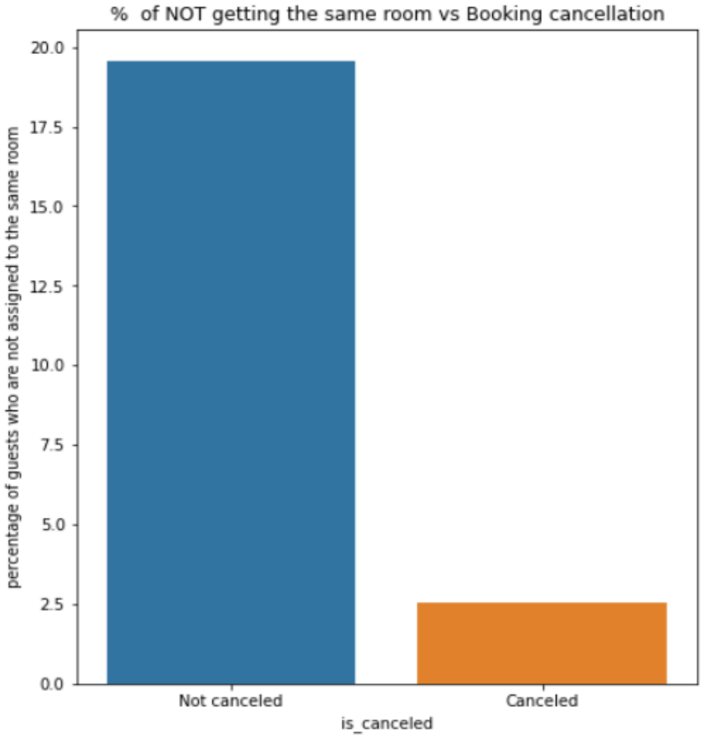
  ❑ Online TA/TO distribution channel has the highest cancellation in both type of Hotels. They should strengthen their deposit and cancellation rules to lower the number of cancellations.

❑ Market Segment :

  ❑ 'Online TA/TO' market segment has highest cancellations for city hotels.

Cancellation Count Analysis

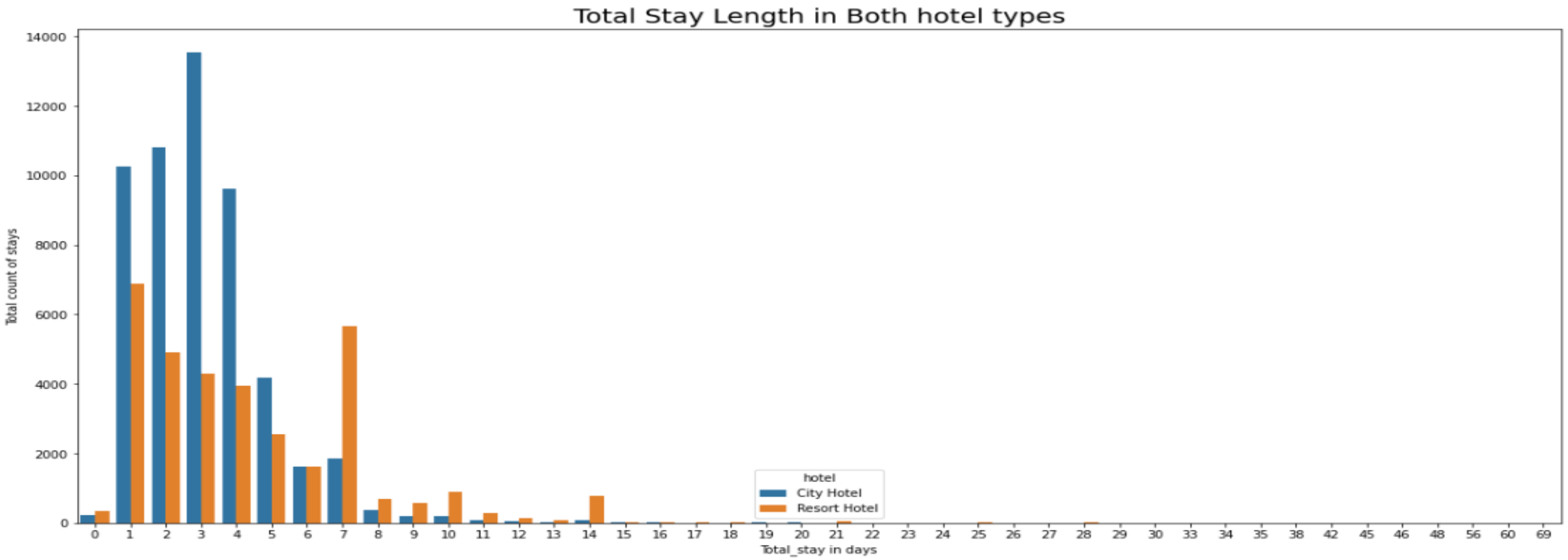
The Relationship between booking cancellation and the percentage of guests who are not assigned to the same room that they reserved :



Findings:

- ❑ Almost 19 % people did not canceled their bookings even after not getting the same room which they reserved while booking hotel. Only 2.5 % people cancelled the booking.
- ❑ It is evident that even if customers are not given the rooms they requested throughout the booking process, there is no appreciable impact on cancellations of reservations.

Ideal stay length Analysis

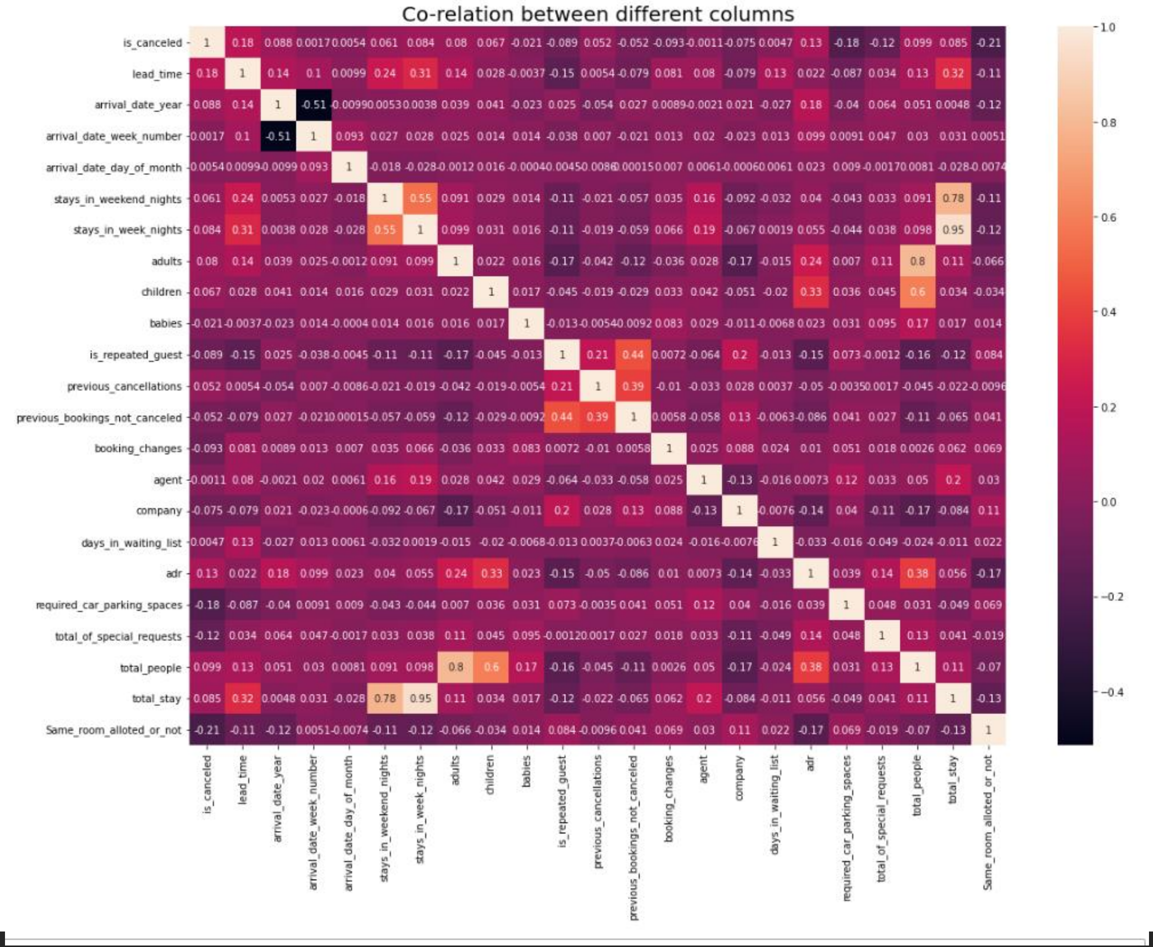


Findings :

❑ The ideal stay in either sort of hotel is under seven days.

❑ People prefer to stay in resort hotels for stays longer than seven days. As we can see, compared to resort hotels, reservations for city hotels have drastically decreased after 7 days.

Correlation Heatmap



Findings:

- ❑ 'is_canceled' and 'same_room_alloted_or_not' are negatively corelated. So, even if a guest doesn't get the exact accommodation he reserved, the customer is unlikely to change his reservations. It is already proven, to the last analysis.
- ❑Total people and adr are positively Correlated that means more number of people increases the adr.
- ❑"is_repeted guest" and "previous_bookings_not_canceled" have a strong positive correlation.
- ❑Lead time and total stay have a positive relationship. This implies that the lead time will increase as the customer stays longer.

Thank You