Capstone Project-1

Exploratory Data Analysis (EDA) of Hotel Booking

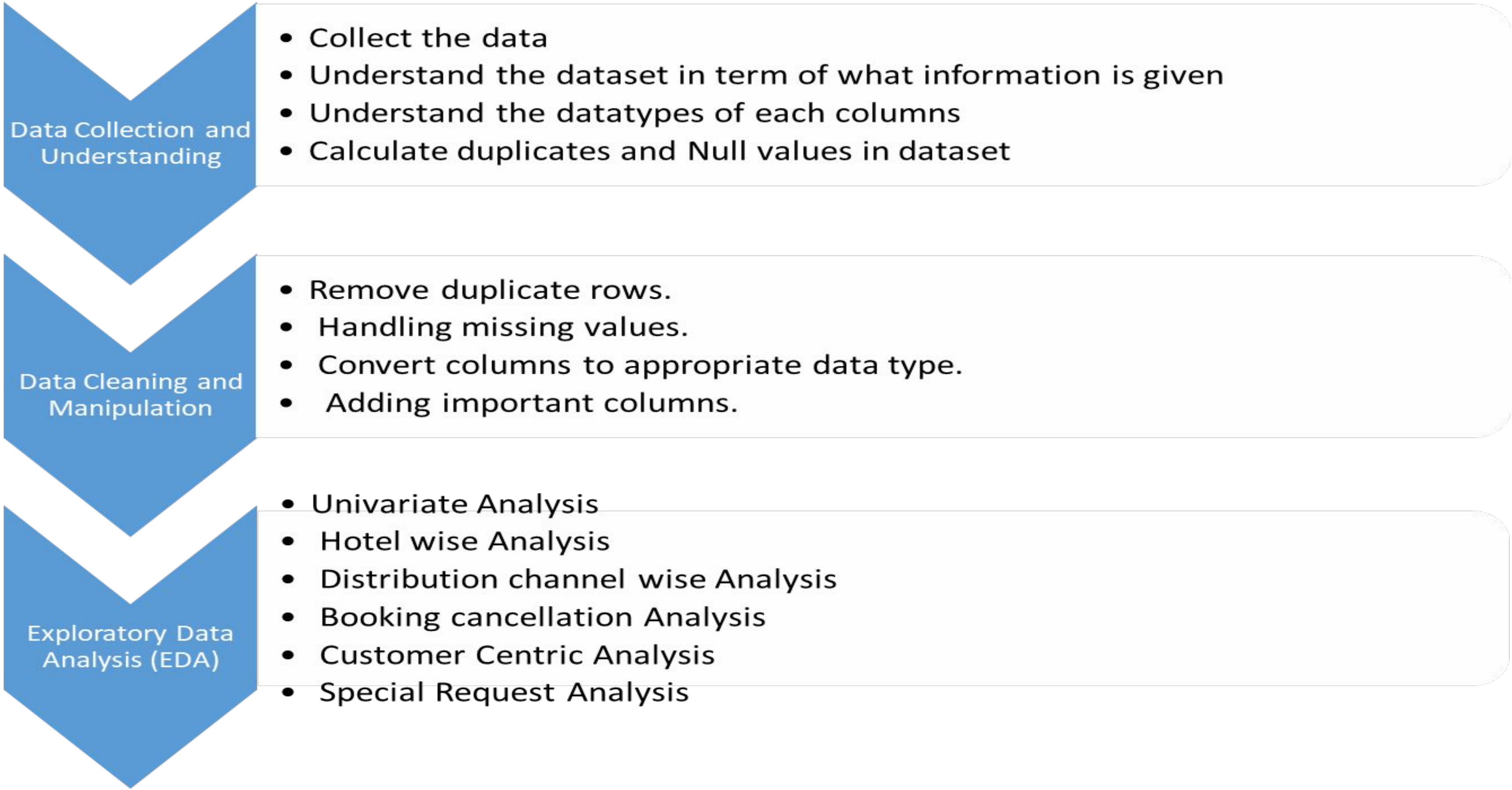
Team Members:

1. Aditya Kumar Saw
2. Parijat Krishna

List of content

1. ıntroduction
2. EDA process flow chart
3. Data Collection and Understanding:
4. Data Cleaning and Manipulation
5. Correlation Heat map
6. Exploratory Data Analysis (EDA)
7. Univariate Analysis
8. Conclusion

EDA process Flow Chart



Data Collection and Understanding

- Collect the data
- Understand the dataset in term of what information is given
- Understand the datatypes of each columns
- Calculate duplicates and Null values in dataset

Data Cleaning and Manipulation

- Remove duplicate rows.
- Handling missing values.
- Convert columns to appropriate data type.
- Adding important columns.

Exploratory Data Analysis (EDA)

- Univariate Analysis
- Hotel wise Analysis
- Distribution channel wise Analysis
- Booking cancellation Analysis
- Customer Centric Analysis
- Special Request Analysis

Data Collection and Understanding:

We have a Dataset of hotel booking analysis from years **2015 to 2017** and having **32 columns**. Our aim is to find the relevant insights from this dataset.

Data Description:

1. **hotel :** Hotel(Resort Hotel or City Hotel)
2. **is_canceled :** Value indicating if the booking was canceled (1) or not (0)
3. **lead_time :** Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
4. **arrival_date_year :** Year of arrival date
5. **arrival_date_month :** Month of arrival date
6. **arrival_date_week_number :** Week number of year for arrival date

7. arrival_date_day_of_month : Day of arrival date

8. stays_in_weekend_nights : Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

9. stays_in_week_nights : Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel

10. adults : Number of adults

11. children : Number of children

12. babies : Number of babies

13. meal : Type of meal booked. Categories are presented in standard hospitality meal packages:

14. country : Country of origin.`

15. market_segment : Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

16. distribution_channel : Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”

17. is_repeated_guest : Value indicating if the booking name was from a repeated guest (1) or not (0)

18. previous_cancellations : Number of previous bookings that were cancelled by the customer prior to the current booking

19. previous_bookings_not_canceled : Number of previous bookings not cancelled by the customer prior to the current booking

20. reserved_room_type : Code of room type reserved. Code is presented instead of designation for anonymity reasons.

21. assigned_room_type : Code for the type of room assigned to the booking.

22. booking_changes : Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation

23. deposit_type : Indication on if the customer made a deposit to guarantee the booking.

24. agent : ID of the travel agency that made the booking

25. company : ID of the company/entity that made the booking or responsible for paying the booking.

26. days_in_waiting_list : Number of days the booking was in the waiting list before it was confirmed to the customer

27. customer_type : Type of booking, assuming one of four categories

a) adr : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights

b) required_car_parking_spaces : Number of car parking spaces required by the customer

c) total_of_special_requests : Number of special requests made by the customer (e.g. twin bed or high floor)

d) reservation_status : Reservation last status, assuming one of three categories

Cancelled – booking was canceled by the customer

Check-Out – customer has checked in but already departed

No-Show – customer did not check-in and did inform the hotel of the reason why

1. reservation_status_date : Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel

Data Cleaning and Manipulation

Cleaning data is crucial step before EDA as it will remove the ambigous data that can affect the outcome of EDA.

While cleaning data we will perform following steps:

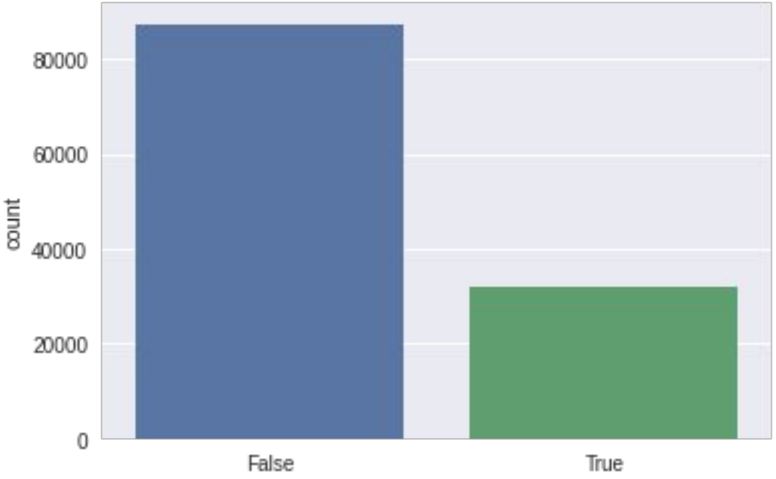
1.Removed duplicate rows.

2.Handling missing values.

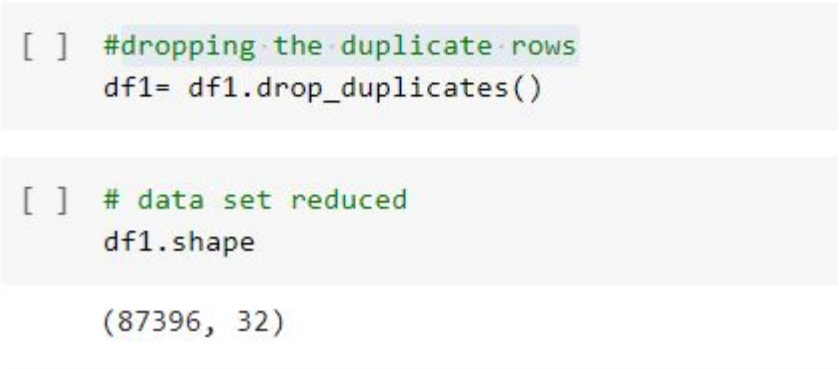
3.Convert columns to appropriate datatypes.

4.Adding important columns

1- Checking for duplicated row



Dropping the duplicate rows

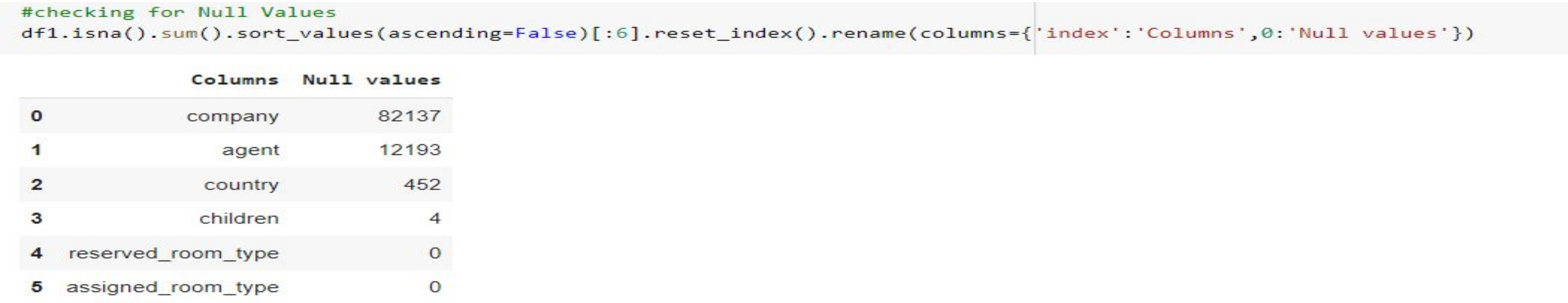


```
[ ]  #dropping the duplicate rows
     df1= df1.drop_duplicates()
```

```
[ ]  # data set reduced
     df1.shape
```

```
(87396, 32)
```

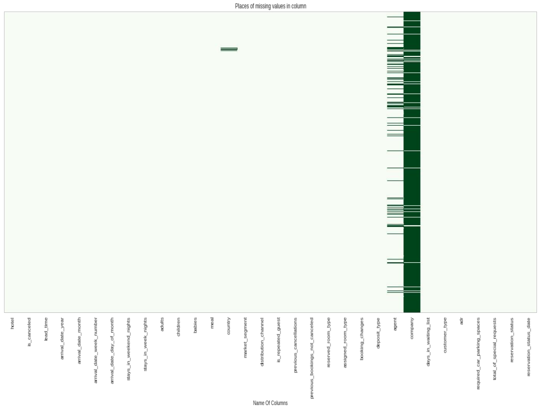
checking for Null Values



```
#checking for Null Values
df1.isna().sum().sort_values(ascending=False)[:6].reset_index().rename(columns={'index':'Columns',0:'Null values'})
```

	Columns	Null values
0	company	82137
1	agent	12193
2	country	452
3	children	4
4	reserved_room_type	0
5	assigned_room_type	0

Visualizing null values through heatmap



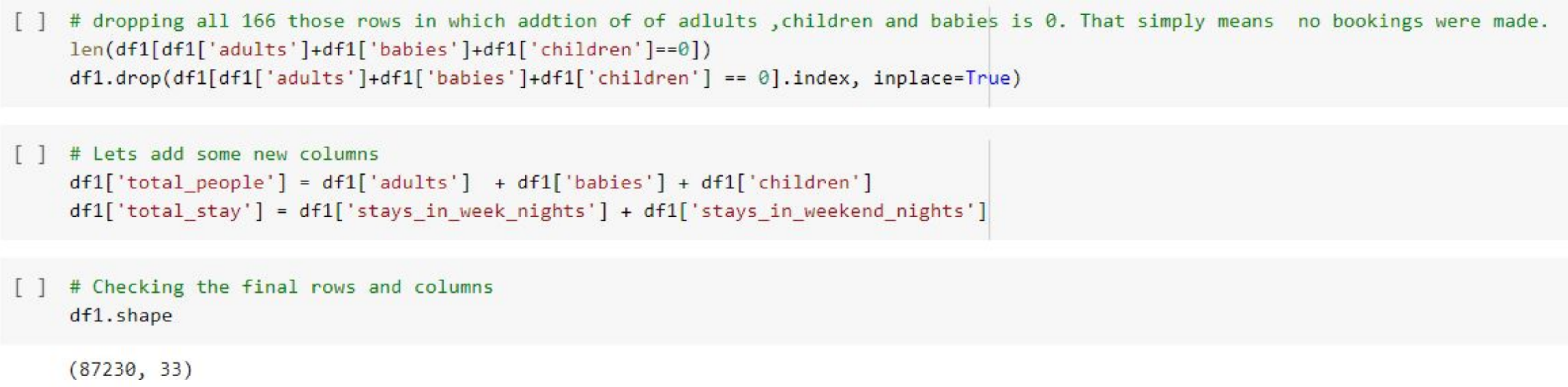
We have null values in Company, Agent, Children, Country:-

We Have Null values in columns- Company, agent, Country,children.

1.For company and agent I will fill the Missing values with 0.

2.For country I will fill Missing values with Object 'Others'. (assuming while collecting data country was not found so user selected the 'Others' option.)

3.As the counting of missing values in Children Column is only 4, so we can replace with 0 considering no childrens.



```
[ ]  # dropping all 166 those rows in which addtion of of adlults ,children and babies is 0. That simply means  no bookings were made.
     len(df1[df1['adults']+df1['babies']+df1['children']==0])
     df1.drop(df1[df1['adults']+df1['babies']+df1['children'] == 0].index, inplace=True)
```

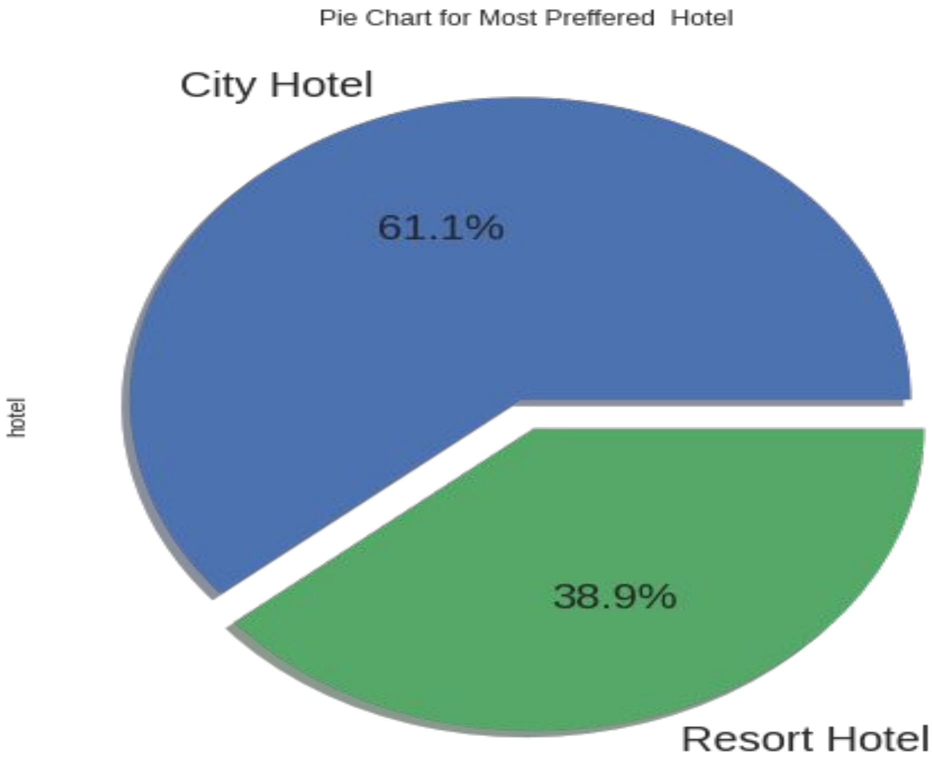
```
[ ]  # Lets add some new columns
     df1['total_people'] = df1['adults']  + df1['babies'] + df1['children']
     df1['total_stay'] = df1['stays_in_week_nights'] + df1['stays_in_weekend_nights']
```

```
[ ]  # Checking the final rows and columns
     df1.shape

     (87230, 33)
```

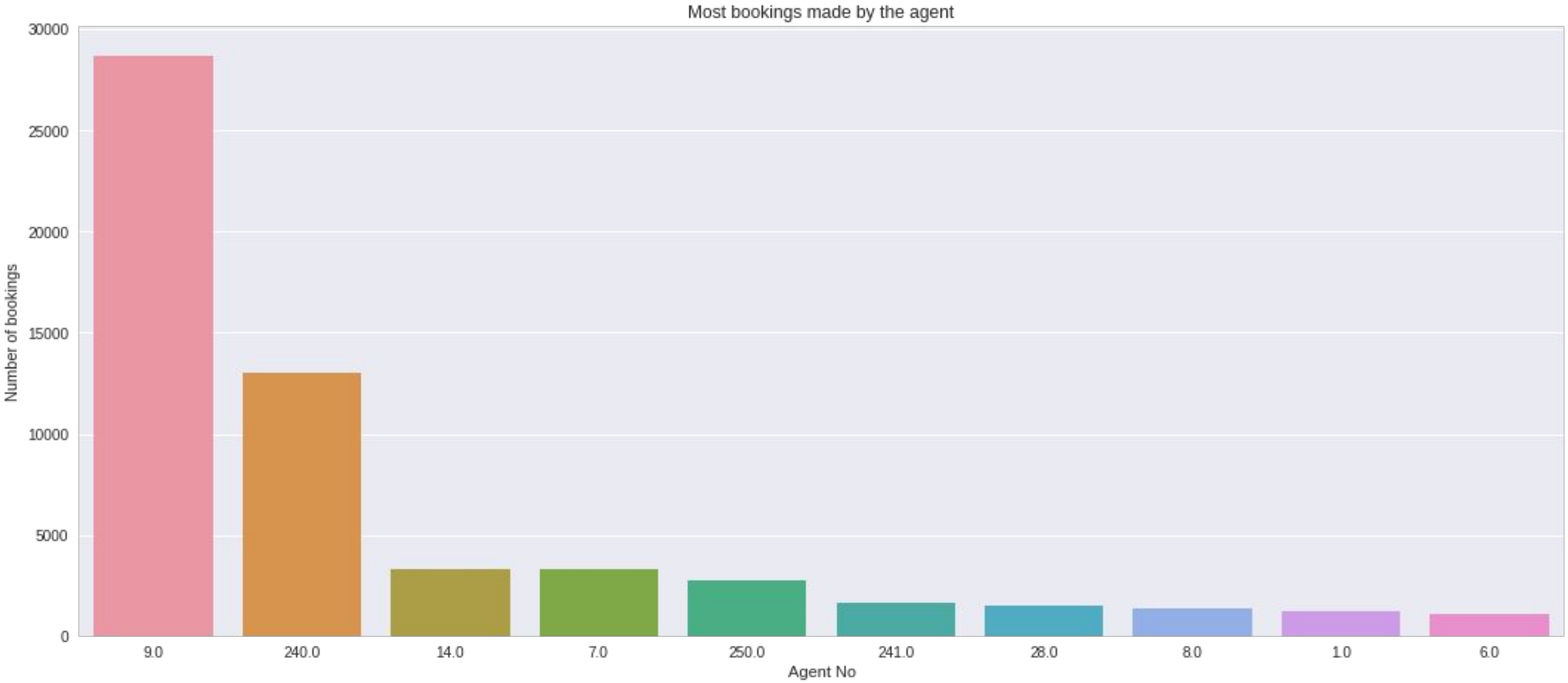
Exploratory Data Analysis (EDA)

1. Which type of hotels is mostly preferred by the guests?



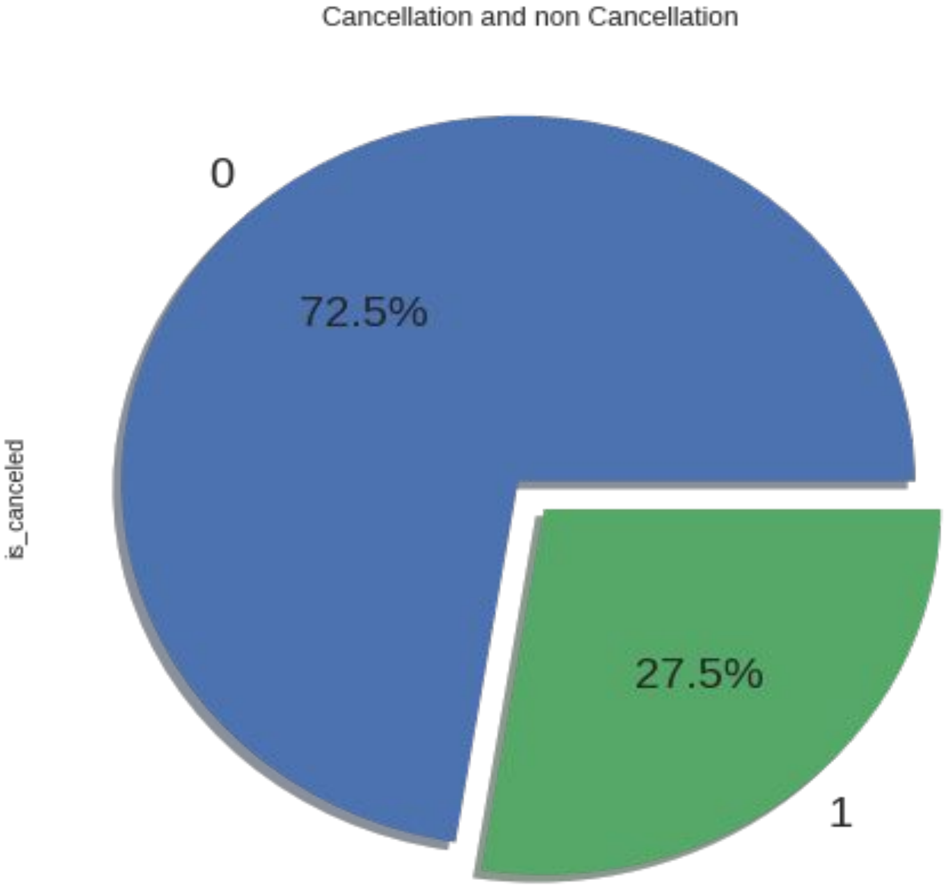
Observation 1: City Hotel is more preferred by guest as compared to Resort Hotel. Thus, City hotel has maximum bookings as compared to Resort Hotel.

2.Which agent made the most bookings?



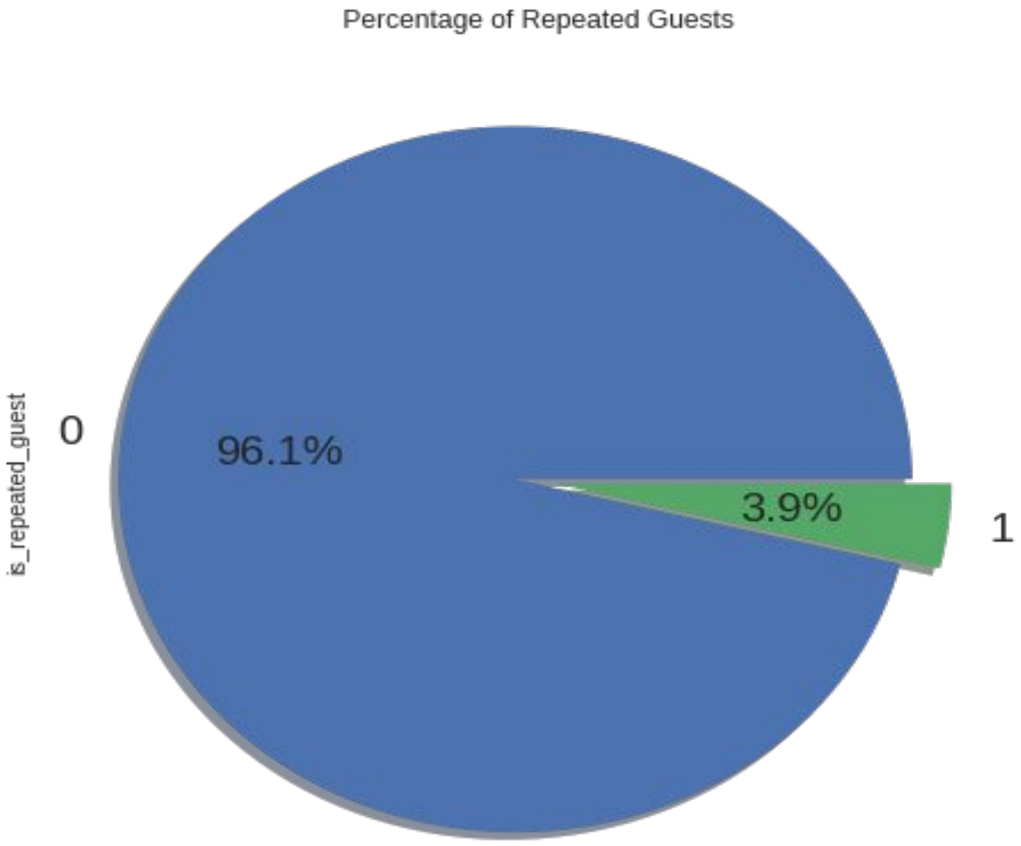
Observation 2: The Agent ID- 9 had done most of the bookings.

3.What is the percentage of cancellation of booked hotels?



Observation 3: 27.5 % of the bookings is cancelled.

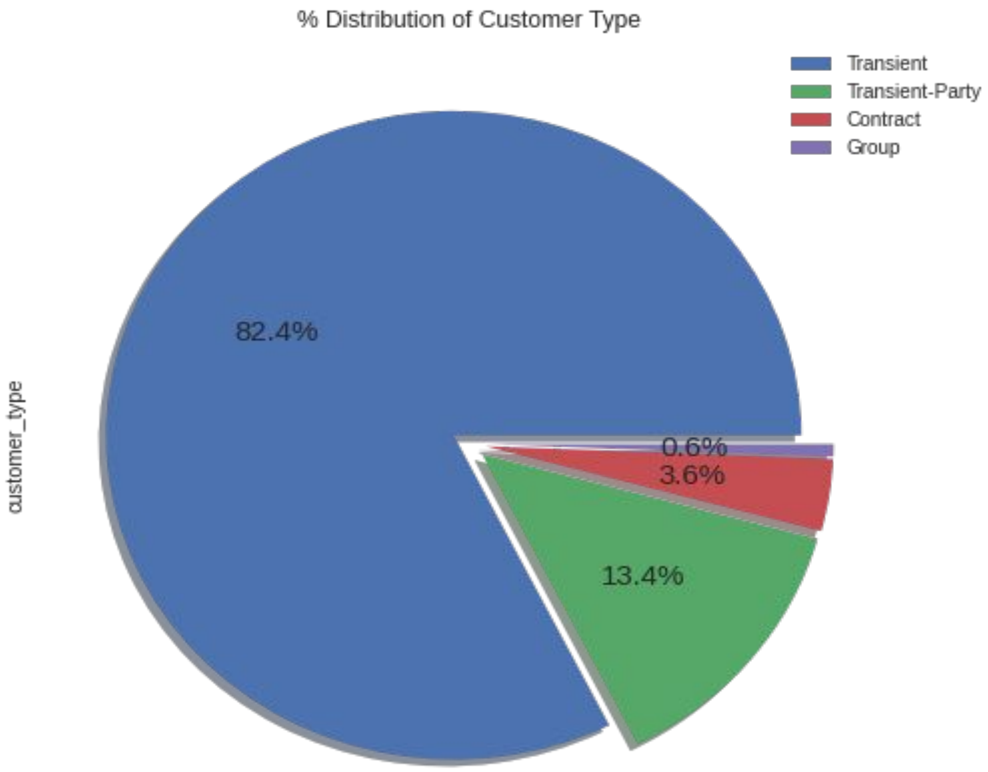
4. What is the percentage of repeated guests?



Observation 4: Repeated guest are less in numbers that is 3.9% only.

In order to retained the guests/customers, management should take feedbacks from guests and try to imporve the services.

5. What is the percentage distribution of *Customer Type*?



1.Contract: When the booking has an allotment or other type of contract associated to it

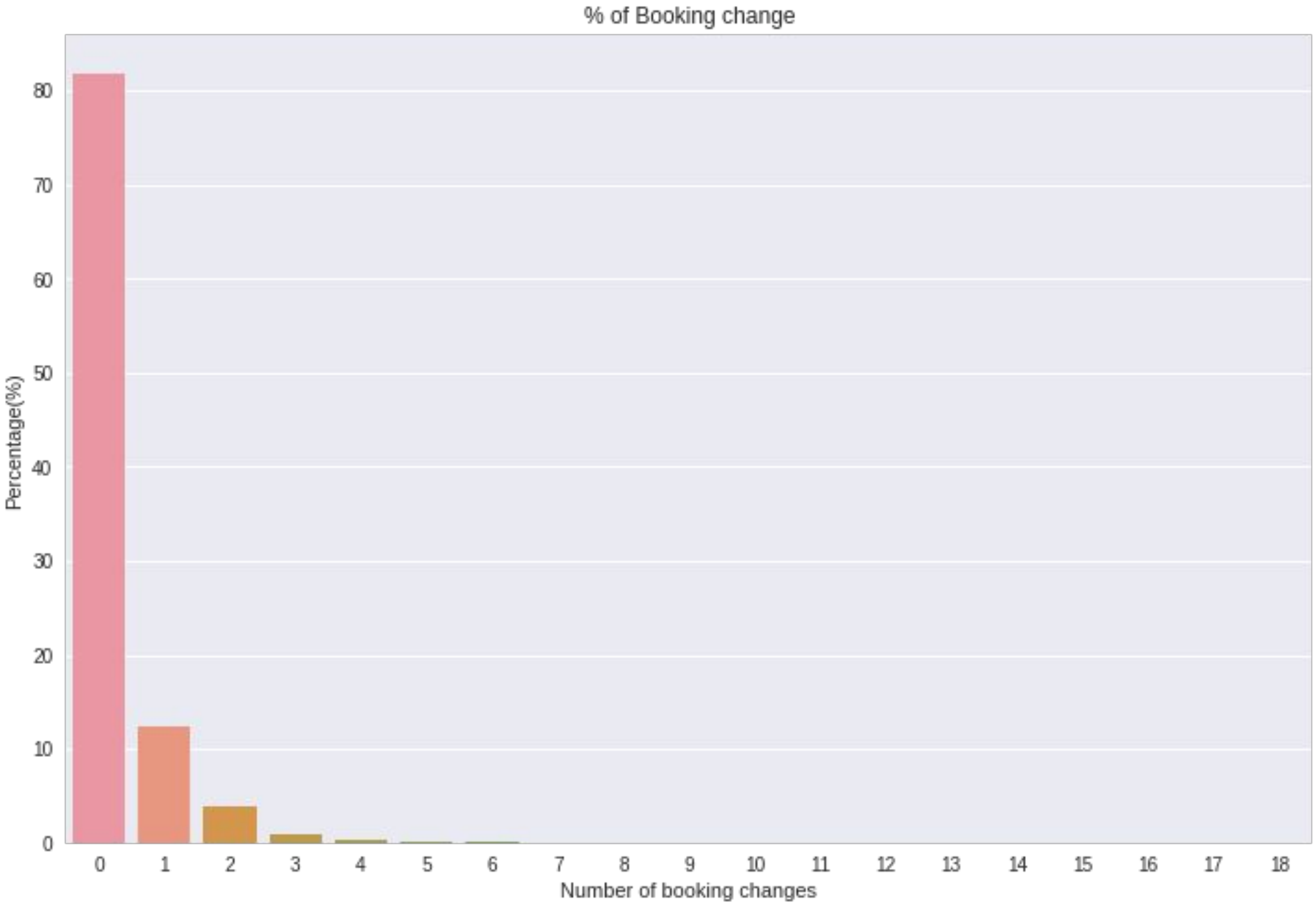
2.Group: When the booking is associated to a group

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

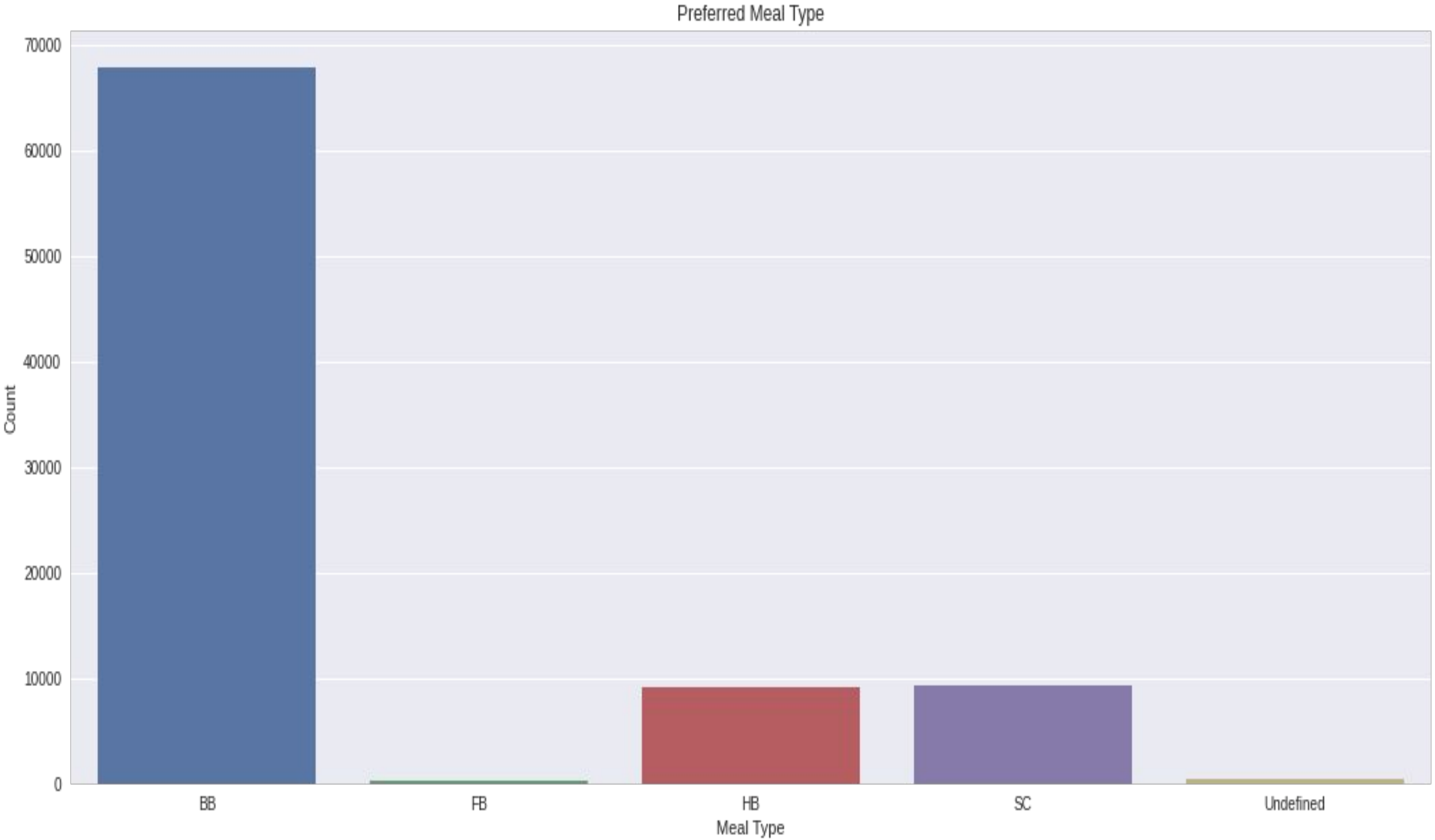
Observation: Transient customer type is most which is 82.4 % while percentage of bookings associated by other groups is vey low.

6. What is the most percentage of booking changes made by the customer?



Observation 6: Almost 82% of the bookings was not changed by guests while approx. 10% bookings was changed by guests.

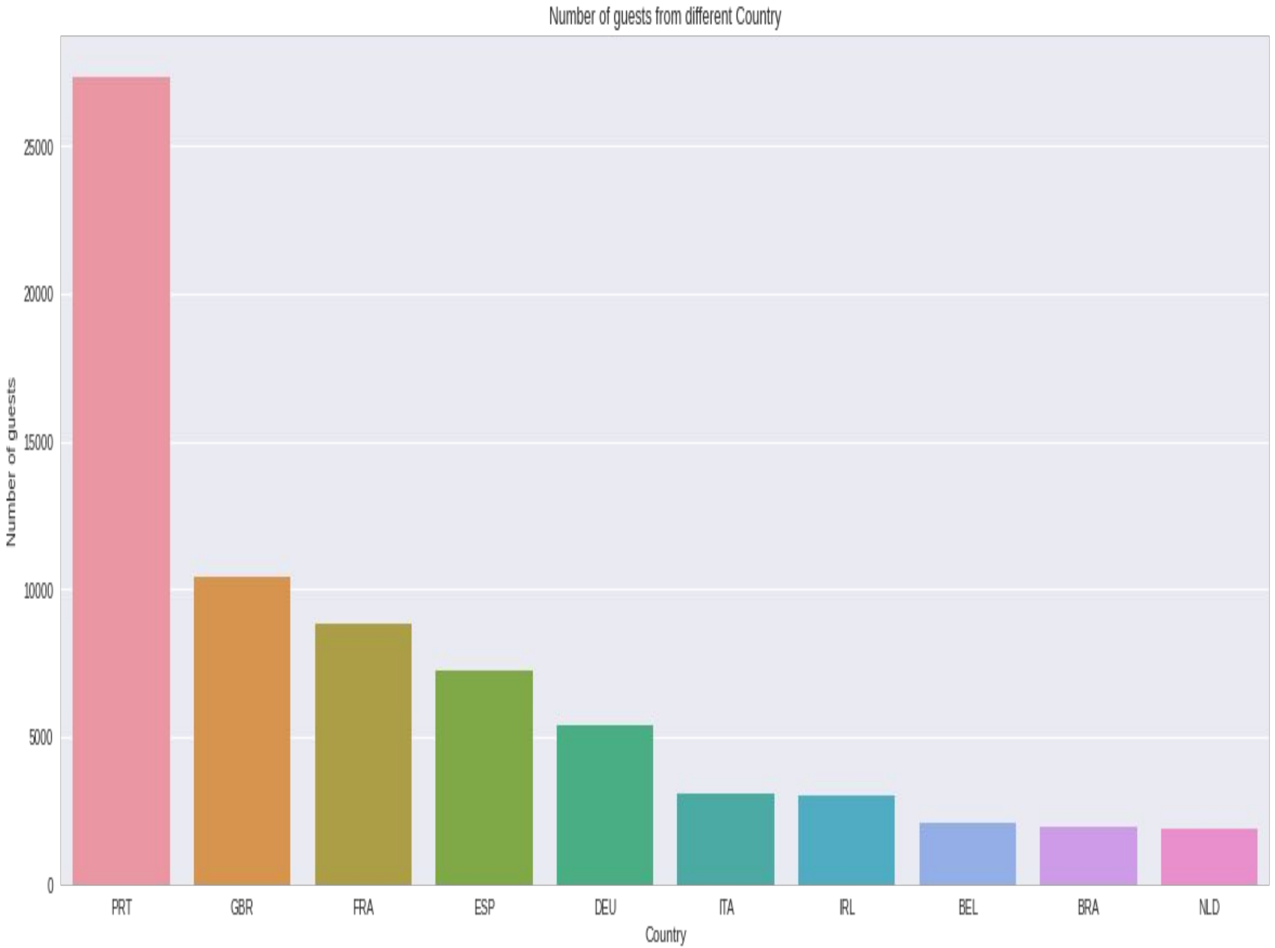
7. Most preferred type of food by guests?



Observation 7:

The most preferred food by the customers is BB (Bed and Breakfast). HB (Half Board) and SC (Self catering) are equally preferred.

8. From which country the most guests are coming?



Observation 8: Most of the guests are coming from Portugal (PRT) that is more than 25000 guests.

Acronym used for Countries-

PRT- Portugal

GBR- United Kingdom

FRA- France

ESP- Spain

DEU - Germany

ITA -Itlay

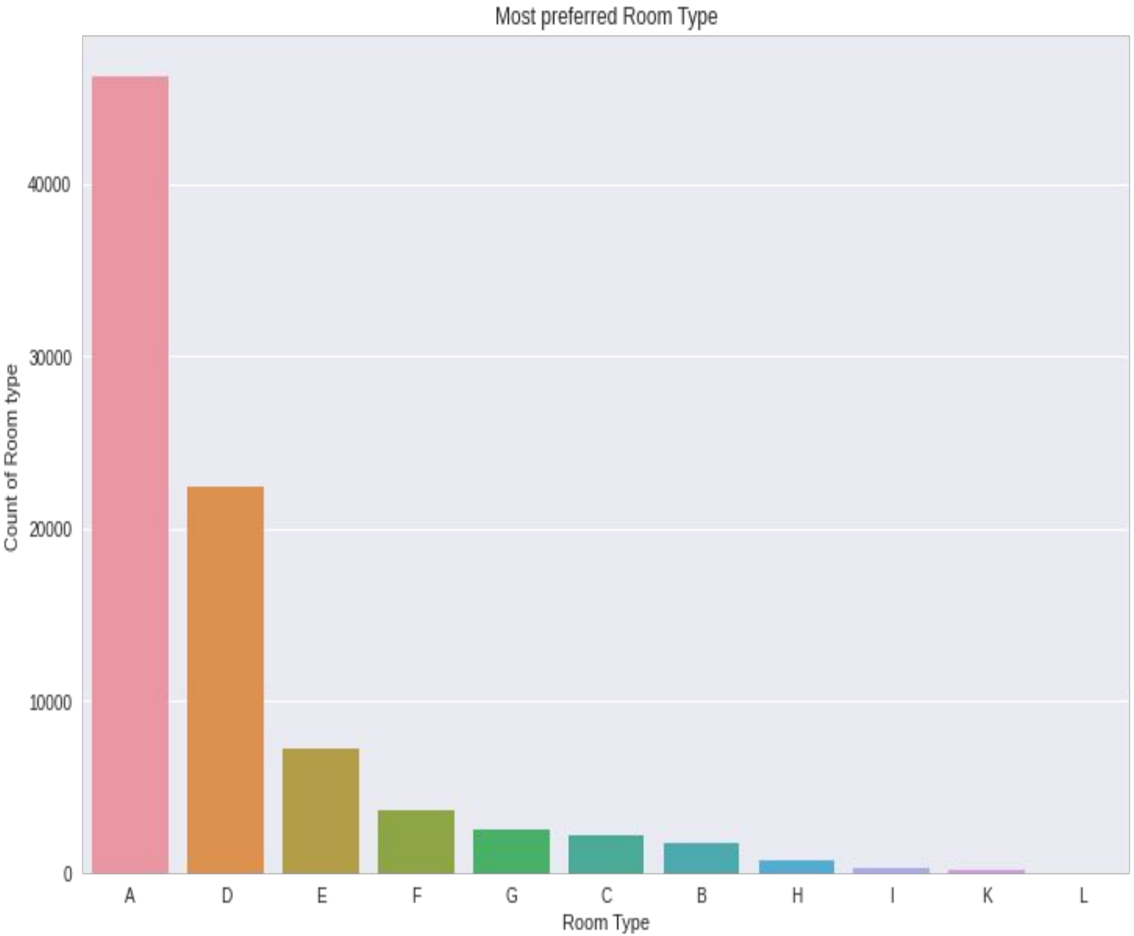
IRL - Ireland

BEL -Belgium

BRA -Brazil

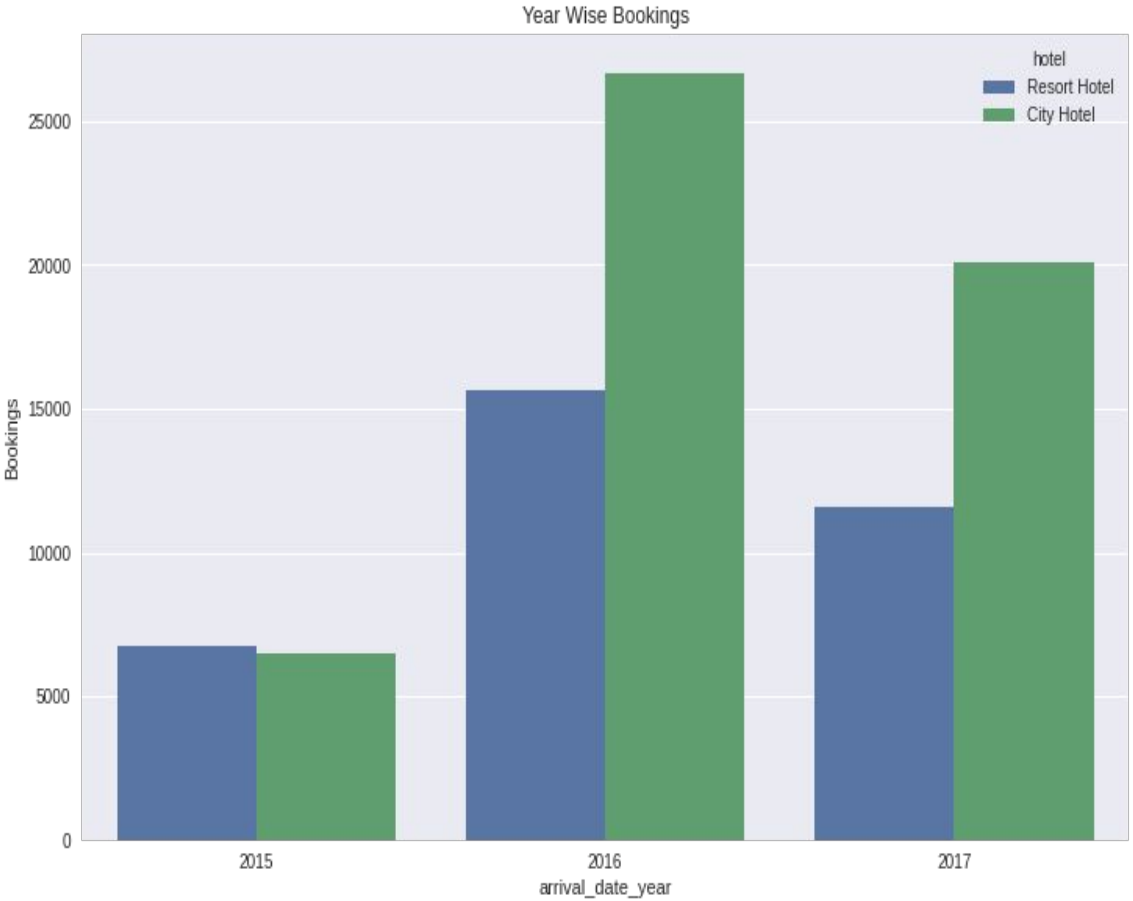
NLD-Netherlands

9. The most preferred room type by the customers?



Observation 9: The most preferred ***Room Type is A***

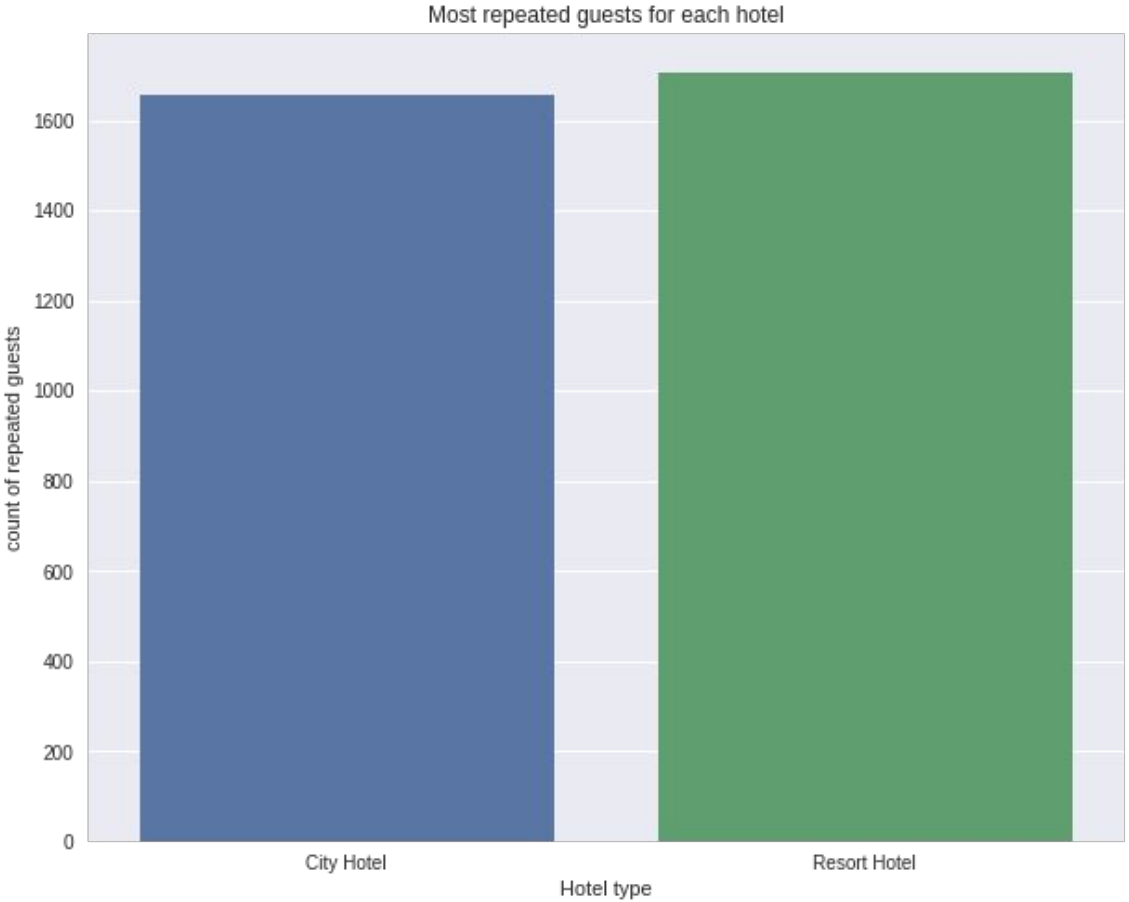
10. Which year has the most hotel bookings?



Observation 10:

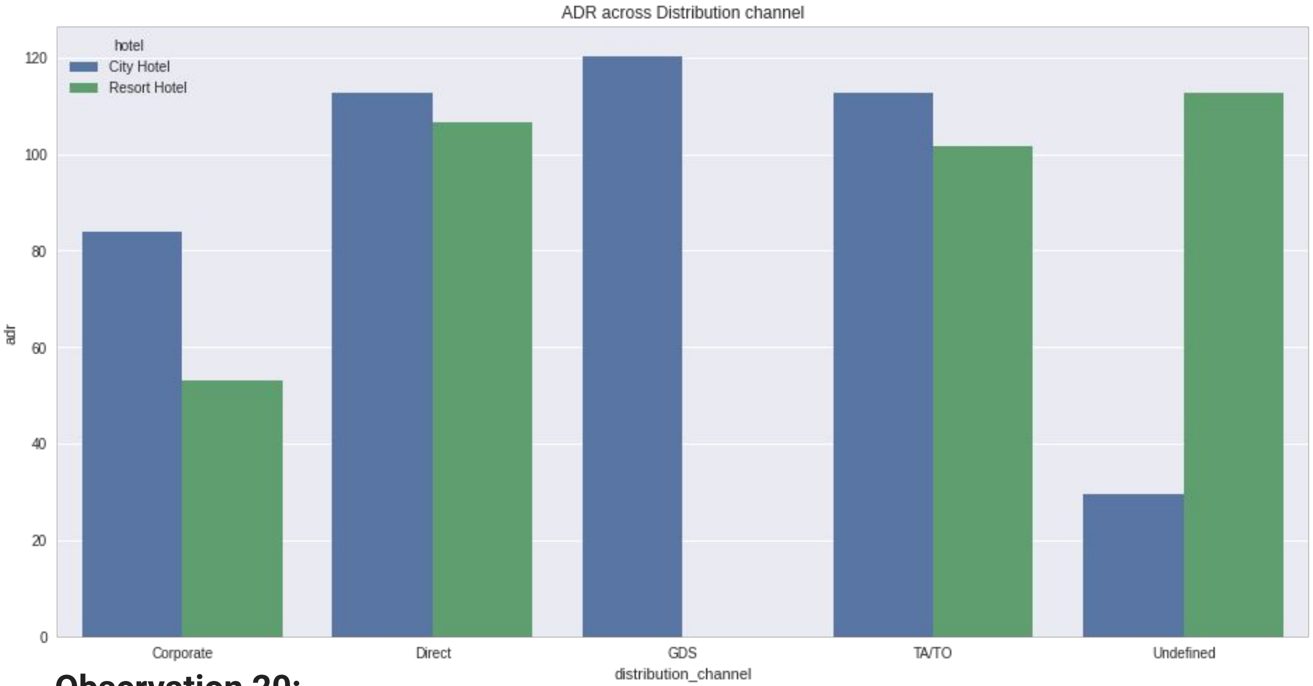
1. 2016 Year has the most bookings.
2. 2015 has less than 10,000 bookings.
3. While overall city hotels had the most of the bookings.

11. Which hotels have the most repeated guests?



Observation 11: Resort Hotel has slighlty more repeated guests as compared to City Hotel.

12. Which distribution channel contributed more to ADR in order to increase the the income?



Observation 20:

1. 'Direct' and 'TA/TO' has almost equally contributed in ADR in both types of hotels.
2. GDS has highly contributed in ADR in 'City Hotel' type.
3. GDS need to increase Resort Hotel Bookings.

Corporate- These are corporate hotel booing companies which makes bookings possible.

GDS-A GDS is a worldwide conduit between travel bookers and suppliers, such as hotels and other accommodation providers. It communicates live product, price and availability data to travel agents and online booking engines, and allows for automated transactions.

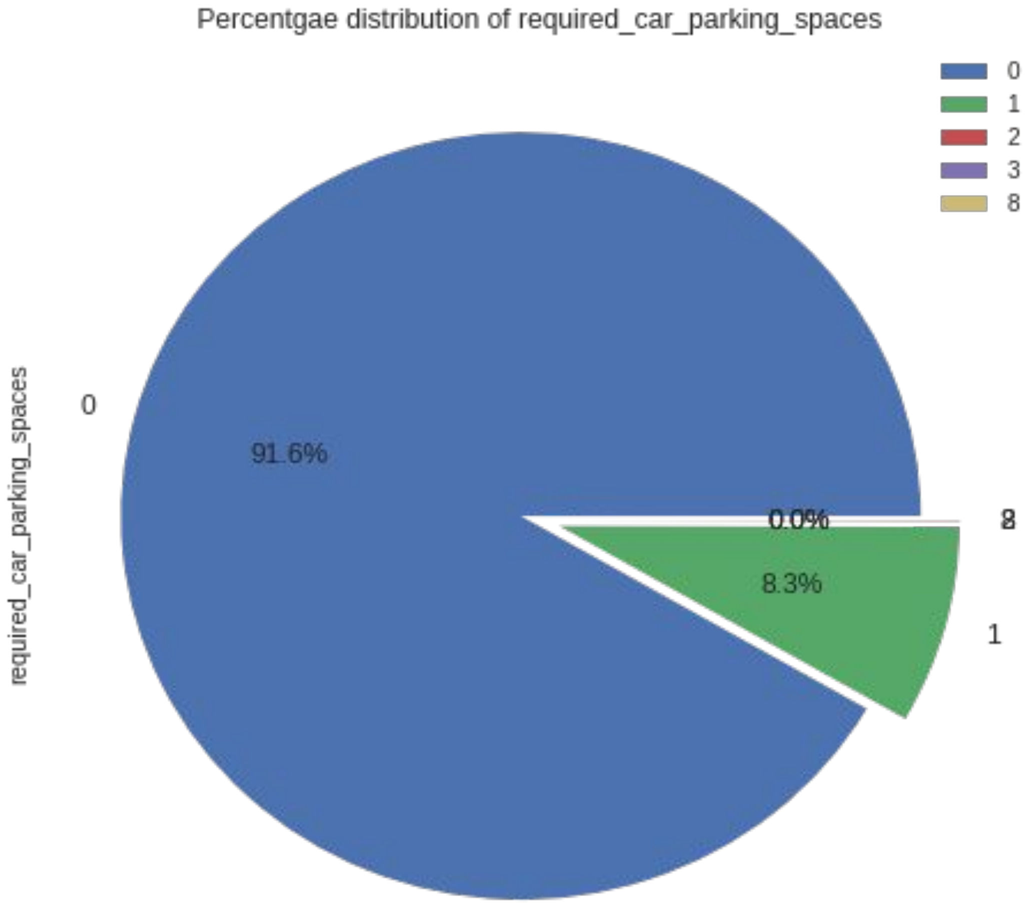
Direct- means that bookings are directly made with the respective hotels.

TA/TO- means that bookings are made through travel agents or travel operators.

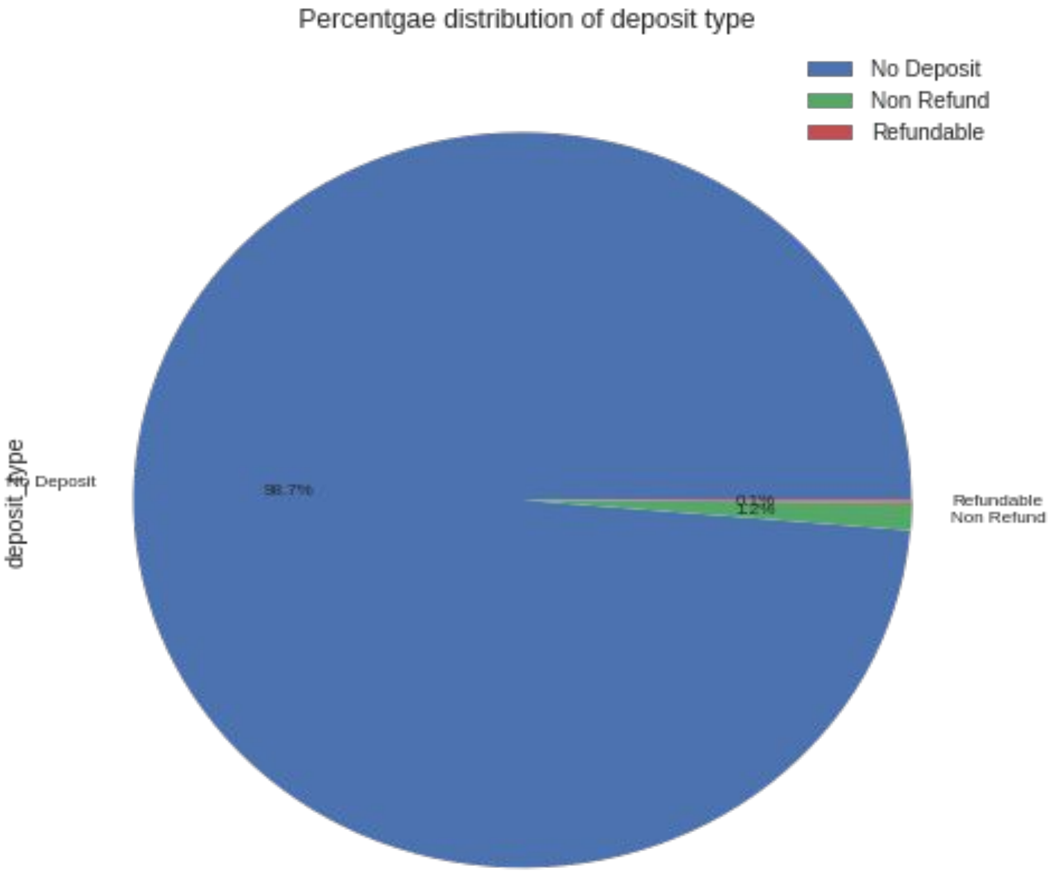
Undefined- Bookings are undefined. may be customers made their bookings on arrival.

13. What is the percentage distribution of required car parking spaces?

Observation 13: 93.8 % guests did not required the parking space. Only 6.2 % guests required only 1 parking space

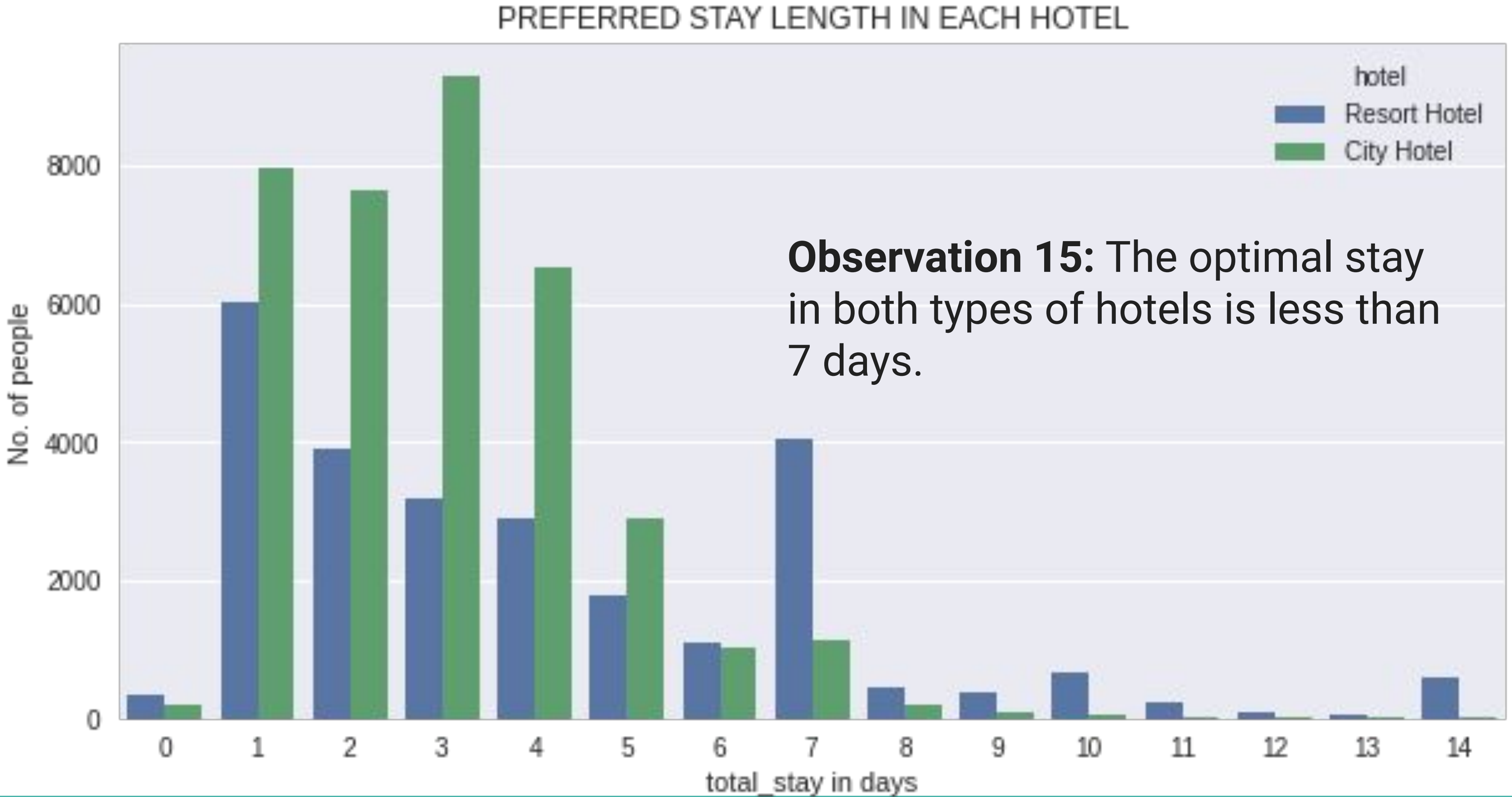


14. What is Percentage distribution of Deposit type?



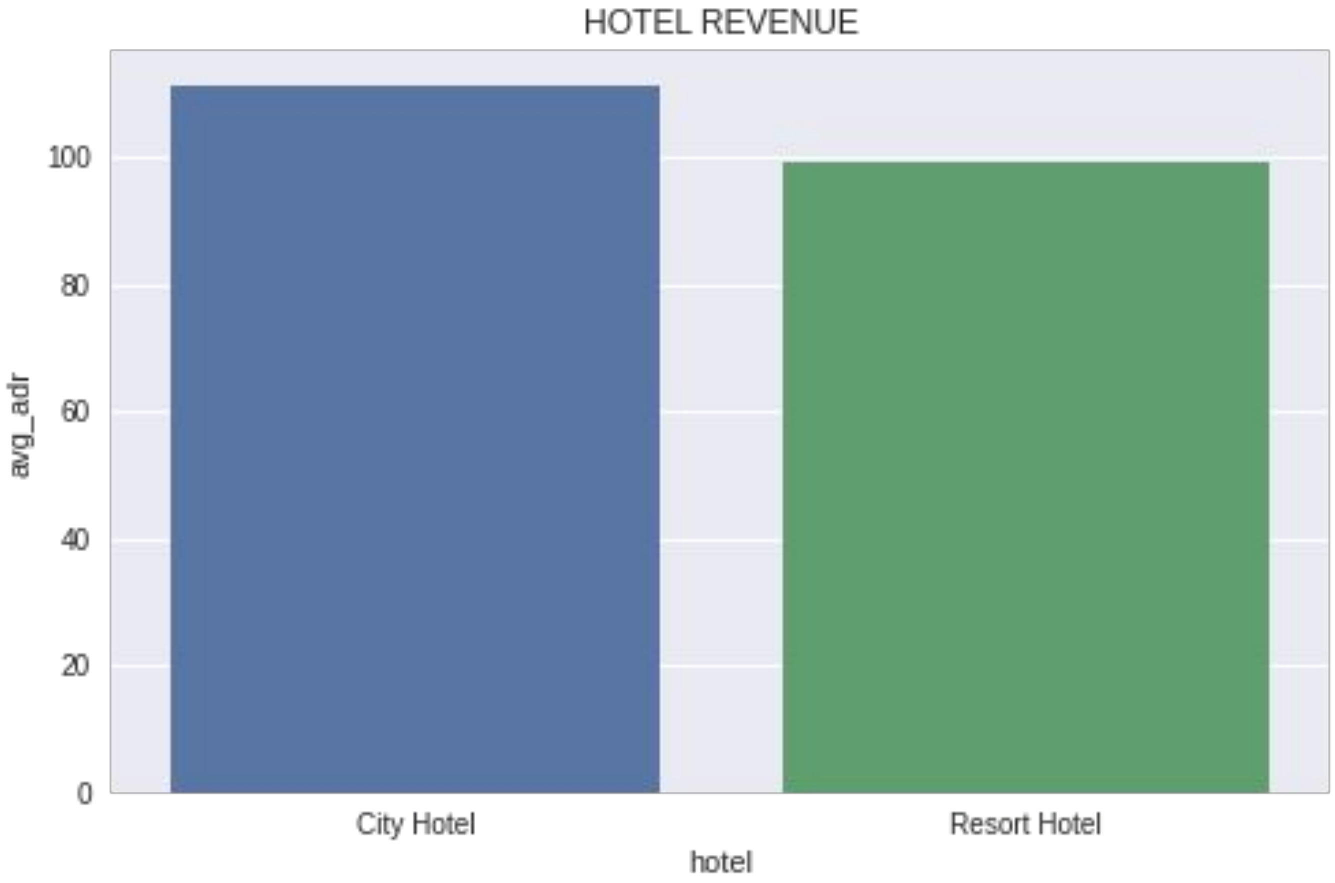
Obdervation 14: 98.7 % of the guests prefer **No deposit type** of stay.

15. What is most preferred stay length in each hotel?



Observation 15: The optimal stay in both types of hotels is less than 7 days.

16. Which hotel makes more revenue?

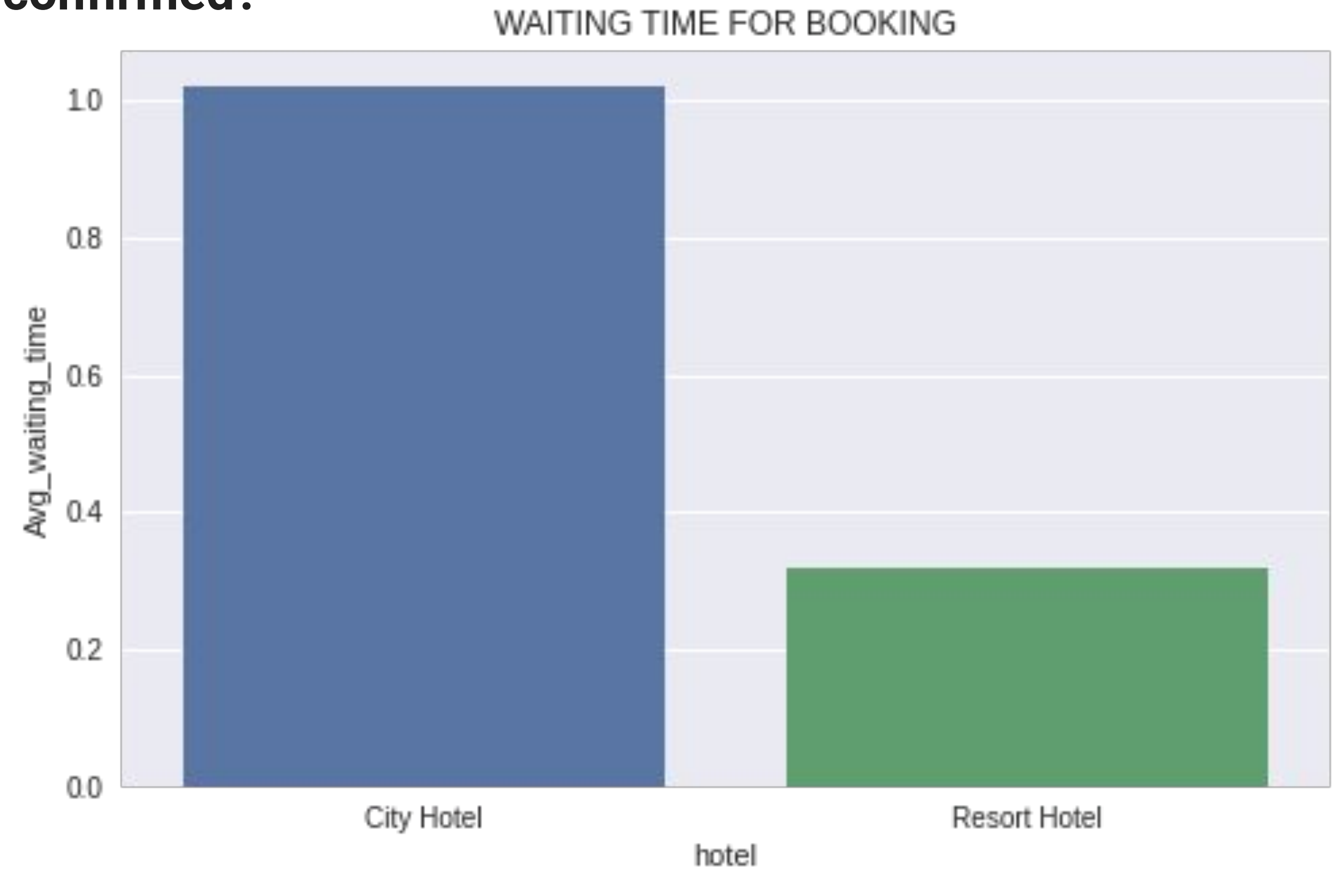


Observartion 16:

1.City hotels has slightly high average lead time than resort hotel.

2.Hence, city hotel makes slightly more revenue then resort hotel.

17. For which hotel, does people have to wait longer to get a booking confirmed?

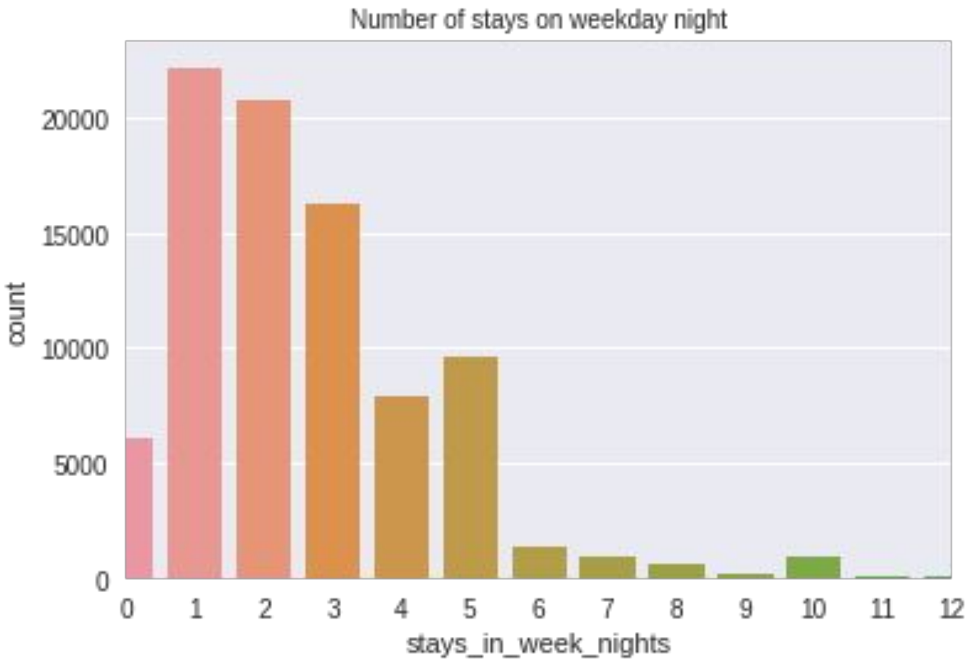


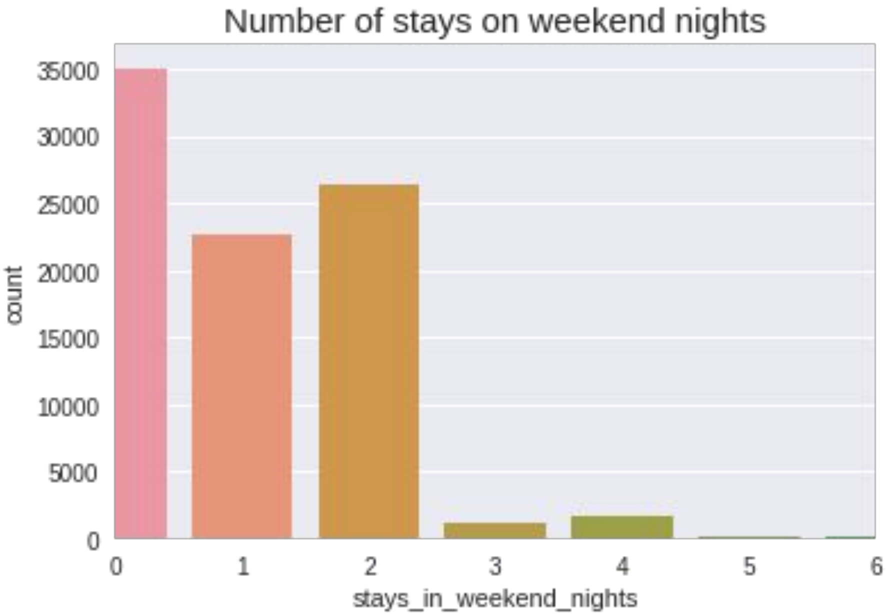
Observation 17:

1.City hotel has significantly longer waiting time then resort hotel.

2.Hence, City Hotel is much busier than Resort Hotel.

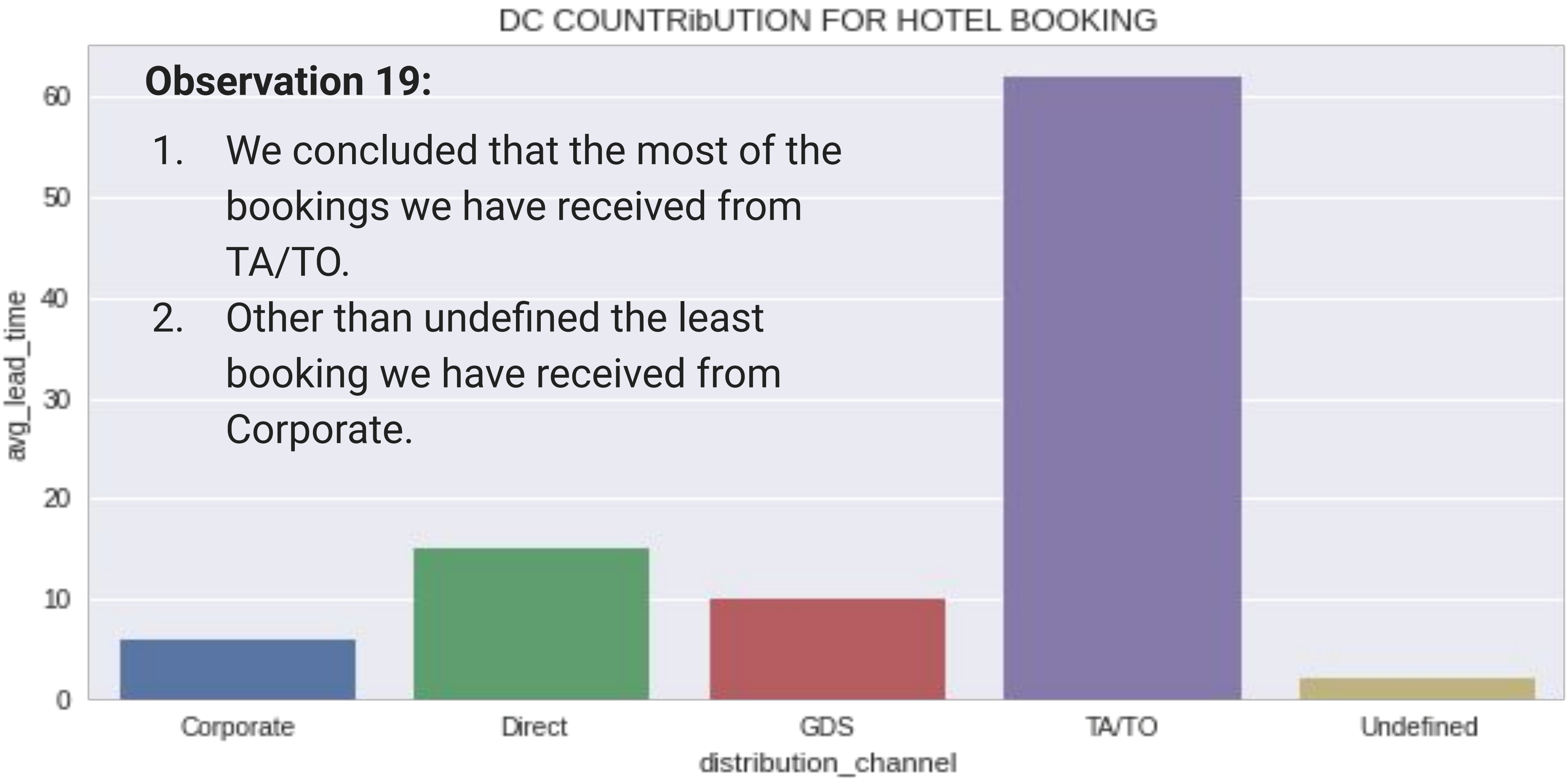
18. Whether Stay is over a weekend or weekday?





Observation 18: Majority of the stays are over the weekday's night. Whatever we saw for the chart on day of the month was random.

19. Which channel is contributing most for early booking of the hotel?

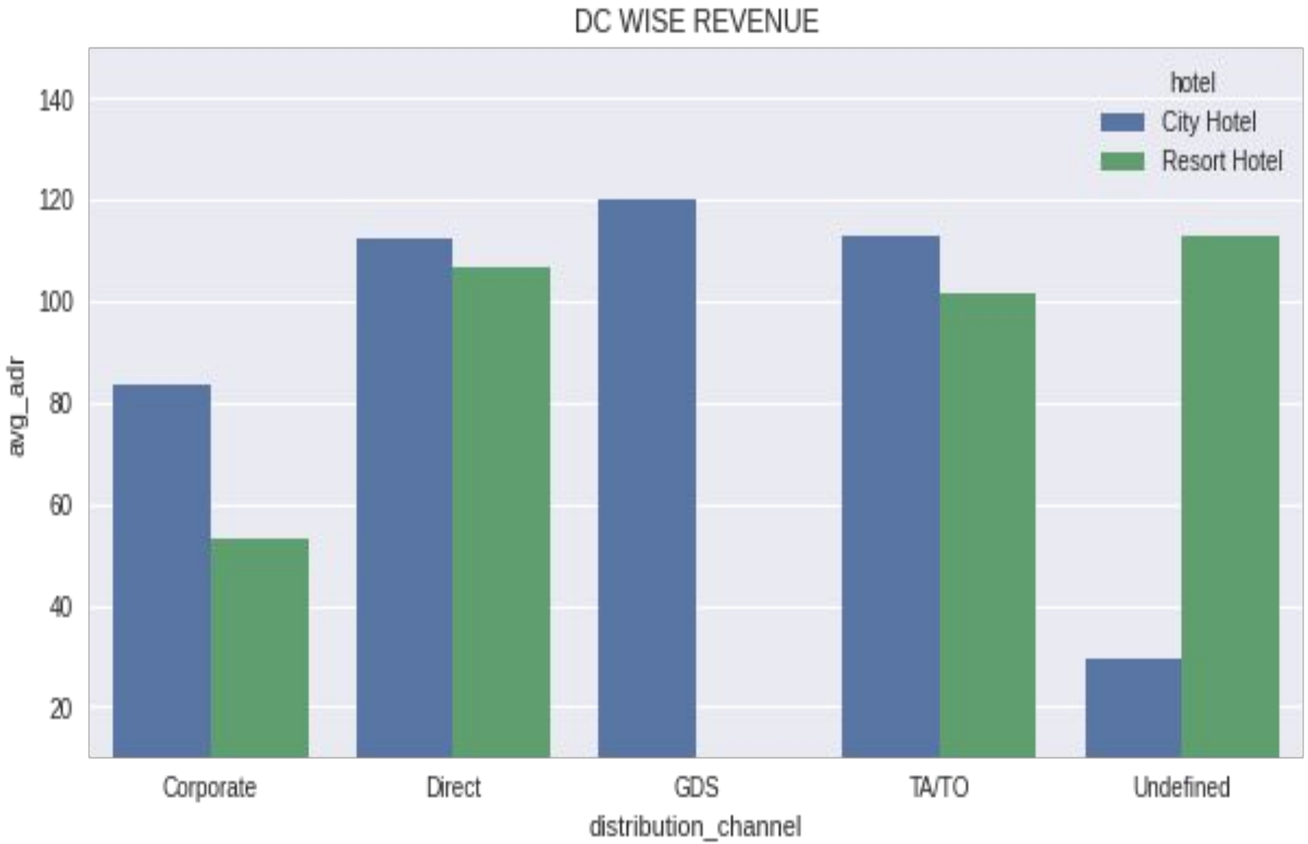


DC COUNTRibUTION FOR HOTEL BOOKING

Observation 19:

1. We concluded that the most of the bookings we have received from TA/TO.
2. Other than undefined the least booking we have received from Corporate.

20.Which distribution channel brings better revenue generating deals for hotels?



Observation 20:

1.In terms of revenue GDS is the most revenue generating Channel but its only for City hotel. For Resort Hotel its contribution is negligible as compared to other channels distribution.

2.Undefined can be associated to multiple channel distribution channels whose data is not provided so after undefined bookings from TA/TO are generating most revenue for the Resort Hotel.

3.Apart from other ditribution channel Direct bookings are also playing the crucial role in terms of revenue generation but we need to focus more on other less revenue generating mediums in order to increase the overall revenue.

Conclusion:

- City Hotel is more preferred by guest as compared to Resort Hotel. Thus, City hotel has maximum bookings as compared to Resort Hotel.
- The Agent ID- 9 had done most of the bookings.
- 27.5 % of the bookings is cancelled.
- Repeated guest are less in numbers that is 3.9% only. In order to retained the guests/customers, management should take feedbacks from guests and try to imporve the services.
- Transient customer type is most which is 82.4 % while percentage of bookings associated by other groups is vey low.
- Almost 82% of the bookings was not changed by guests while approx. 10% bookings was changed by guests.
- a) The most preferred food by the customers is BB (Bed and Breakfast).
  b)HB (Half Board) and SC (Self catering) are equally preferred.
- Most of the guests are coming from Portugal (PRT) that is more than 25000 guests.
- The most preferred room type is A.
- July and August months had the most Bookings. Summer vaccation can be the reason for bookings.
- a) 2016 Year has the most bookings.
  b) 2015 has less than 10,000 bookings
  c) While overall city hotels had the most of the bookings.
- City Hotels have the highest percentage of booking cancellation as compared to Resort Hotels.
- 93.8 % guests did not required the parking space. Only 6.2 % guests required only 1 parking space.

- 98.7 % of the guests prefer No deposit type of stay.
- The optimal stay in both types of hotels is less than 7 days.
- a) City hotels has slightly high average lead time than resort hotel.
  b) Hence, city hotel makes slightly more revenue then resort hotel.
- a) City hotel has significantly longer waiting time then resort hotel.
  b) Hence, City Hotel is much busier than Resort Hotel.
- Majority of the stays are over the weekday's night. Whatever we saw for the chart on day of the month was random.
- a) We concluded that the most of the bookings we have received from TA/TO.
  b) Other than undefined the least booking we have received from Corporate.
- a) In terms of revenue GDS is the most revenue generating Channel but its only for City hotel. For Resort Hotel its contribution is negligible as compared to other channels distribution.
  b) Undefined can be associated to multiple channel distribution channels whose data is not provided so after undefined bookings from TA/TO are generating most revenue for the Resort Hotel.
  c) Apart from other ditribution channel Direct bookings are also playing the crucial role in terms of revenue generation but we need to focus more on other less revenue generating mediums in order to increase the overall revenue.

THANK YOU