Capstone Project 1

EDA on Hotel Booking Analysis

By:
Prakhar Kumar

Problem statement:

- This data set 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, or babies, and the number of available parking spaces, among other things. All personally identifying information has from the data. We will perform exploratory data analysis the data to get insight from the it.

- The main object behind this project is to explore and analyze the data to discover important factors on which bookings in hotels depends on and give insights to management which help them to boost the business ,performance and earn high profit.

Steps to start :-

So we divide the steps in 3 parts:-

1. ***Collecting and understanding the data***
2. ***Data wrangling(cleaning & manipulation)***
3. ***Exploratory data analysis(EDA)***

EDA will be further divided into 3 parts:-

- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

Data Summary : After collecting data there are 119390 rows and 32 columns

- **hotel** :Two hotels are given: Resort Hotel City Hotel
- **is_canceled** : 1: Canceled 0: Not canceled
- **lead_time** : gap between booking and arrival
- **arrival_date_year** :arrival year
- **arrival_date_month** :arrival month
- **arrival_date_week_number** : arrival week
- **arrival_date_day_of_month** :arrival date
- **stays_in_weekend_nights** :count of nights the guests booked the hotel during Sat-Sun
- **stays_in_week_nights** :count of nights the guests booked the hotel during Mon-Fri
- **weekly_stays** :duration of stay including weekend nights and week nights stay
- **adults** :count of adults
- **children** :count of children
- **babies** :count of babies
- **meal** :meal type (no meal package; BB; HB; FB)

- **country** :country of guests
- **market_segment** :Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- **distribution_channel**:Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators
- **is_repeated_guest** :1: Yes 0: No
- **previous_cancellations**: count of previous bookings that were cancelled by the customer before final booking
- **previous_bookings_not_canceled**: count of no canceled bookings
- **reserved_room_type** : booked room category
- **assigned_room_type**: assigned room category
- **booking_changes**: count of changes made by the customer before final booking
- **deposit_type**: type of deposit made by the customer
- **agent** : travel agent id
- **company** : booking company id
- **days_in_waiting_list** :count of days the booking was in the waiting list before it was confirmed

- **customer_type** :Transient Contract Group Transient-party
- **adr**: average daily rate for the booking
- **price** :total price spent by a guest entity
- **required_car_parking_spaces** :count of car parking spaces allocated to the customer
- **total_of_special_requests** :count of special requests made by the customer
- **reservation_status** :status of reservation
- **reservation_status_date** :date corresponding to status of reservation
- **total_guests** : sum of adults,children,babies
- **arrival_date** : in date format(yyyy-mm-dd)

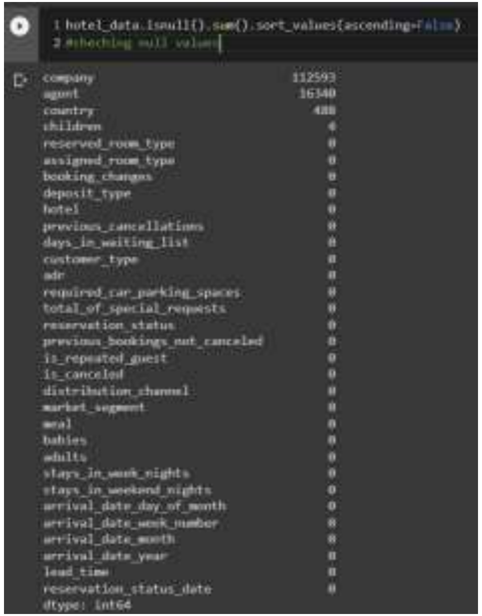
Data cleaning and Manipulation

First checking is there is there any null values or not

There are 4 columns that have null values i.e company,agent,country, children

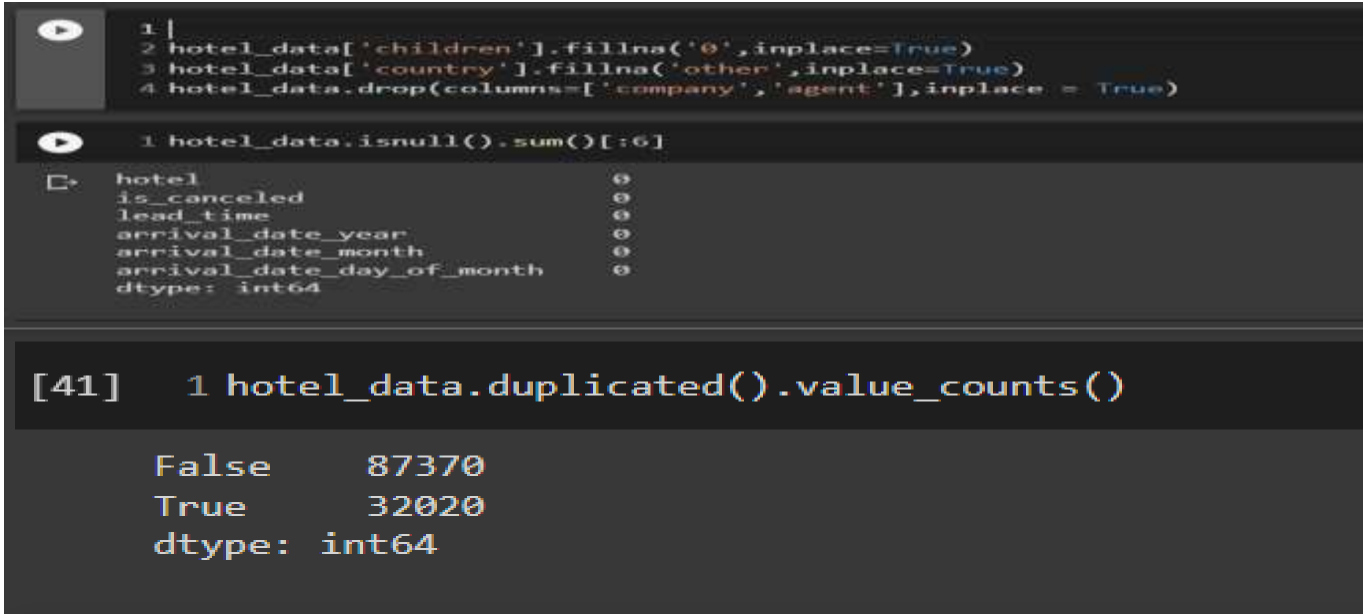
Dropping company and agent data as the they not necessary for the analysis

Handling the duplicated data where True means duplicate items and dropping them



```
hotel_data.isnull().sum().sort_values(ascending=False)
#checking null values
```

```
company                           112593
agent                              16340
country                              488
children                               4
reserved_room_type                     0
assigned_room_type                     0
booking_changes                        0
deposit_type                           0
hotel                                  0
previous_cancellations                 0
days_in_waiting_list                   0
customer_type                          0
adr                                    0
required_car_parking_spaces            0
total_of_special_requests              0
reservation_status                     0
previous_bookings_not_canceled         0
is_repeated_guest                      0
is_canceled                            0
distribution_channel                   0
market_segment                         0
meal                                   0
babies                                 0
adults                                 0
stays_in_week_nights                   0
stays_in_weekend_nights                0
arrival_date_day_of_month              0
arrival_date_week_number               0
arrival_date_month                     0
arrival_date_year                      0
lead_time                              0
reservation_status_date                0
dtype: int64
```



```

hotel_data['children'].fillna('0',inplace=True)
hotel_data['country'].fillna('other',inplace=True)
hotel_data.drop(columns=['company','agent'],inplace = True)
```

```
hotel_data.isnull().sum()[:6]
```

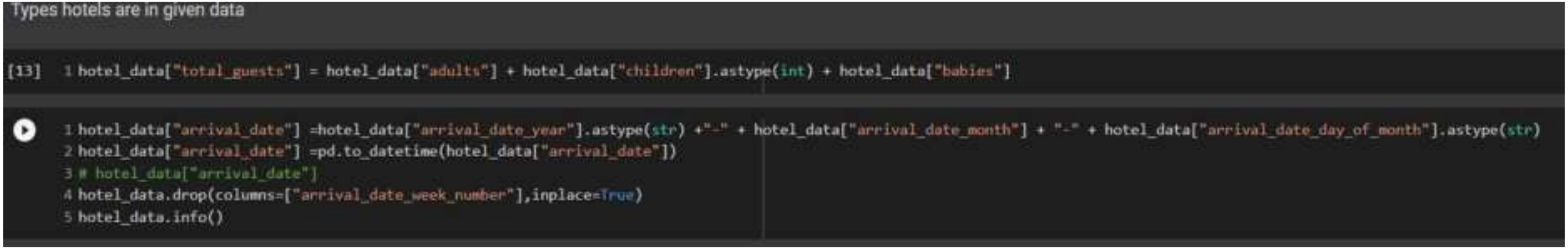
```
hotel                        0
is_canceled                  0
lead_time                    0
arrival_date_year            0
arrival_date_month           0
arrival_date_day_of_month    0
dtype: int64
```

```
[41]  hotel_data.duplicated().value_counts()
```

```
False    87370
True     32020
dtype: int64
```

Creating three new columns :

1) Total_guests: contains sum of adults,children and babies
2) Arrival_date: contains arrival date (yyyy-mm-dd)
3) Weekly_stays: contains sum of stays_in_week_nights and stys_in_weekend_nights

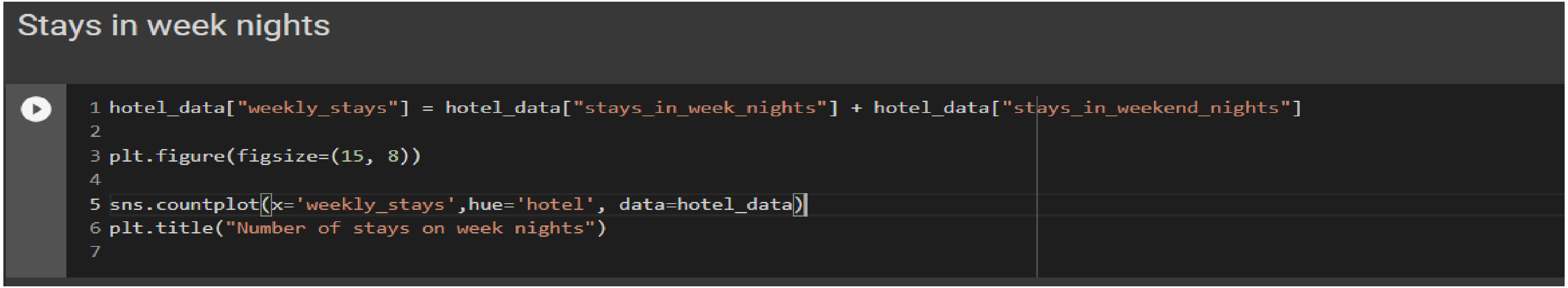


Types hotels are in given data

[13]

```
hotel_data["total_guests"] = hotel_data["adults"] + hotel_data["children"].astype(int) + hotel_data["babies"]
```

```
hotel_data["arrival_date"] =hotel_data["arrival_date_year"].astype(str) +"-" + hotel_data["arrival_date_month"] + "-" + hotel_data["arrival_date_day_of_month"].astype(str)
hotel_data["arrival_date"] =pd.to_datetime(hotel_data["arrival_date"])
# hotel_data["arrival_date"]
hotel_data.drop(columns=["arrival_date_week_number"],inplace=True)
hotel_data.info()
```

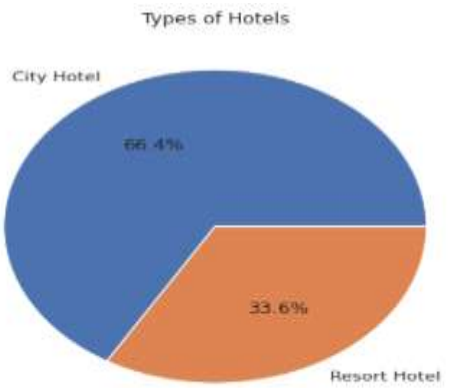


Stays in week nights

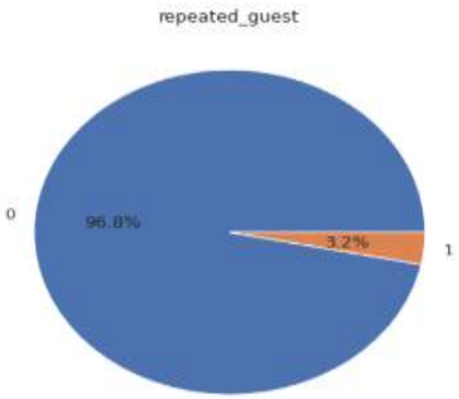
```
hotel_data["weekly_stays"] = hotel_data["stays_in_week_nights"] + hotel_data["stays_in_weekend_nights"]

plt.figure(figsize=(15, 8))

sns.countplot(x='weekly_stays',hue='hotel', data=hotel_data)
plt.title("Number of stays on week nights")

```

Exploratory Data Analysis(EDA)

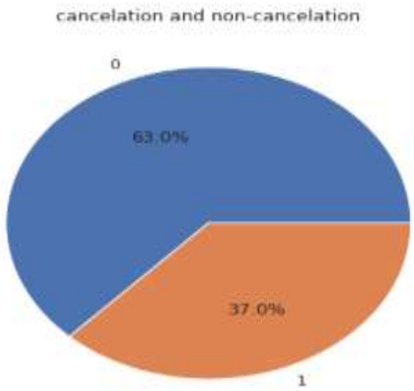
TYPES OF HOTELS



REPEATED GUESTS



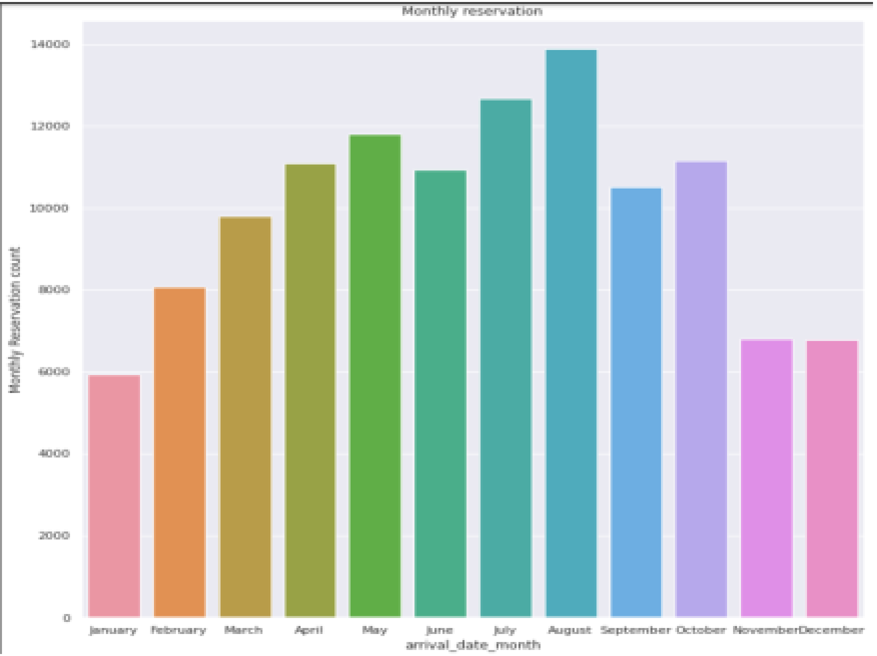
CANCELATION AND NON-CANCELATION

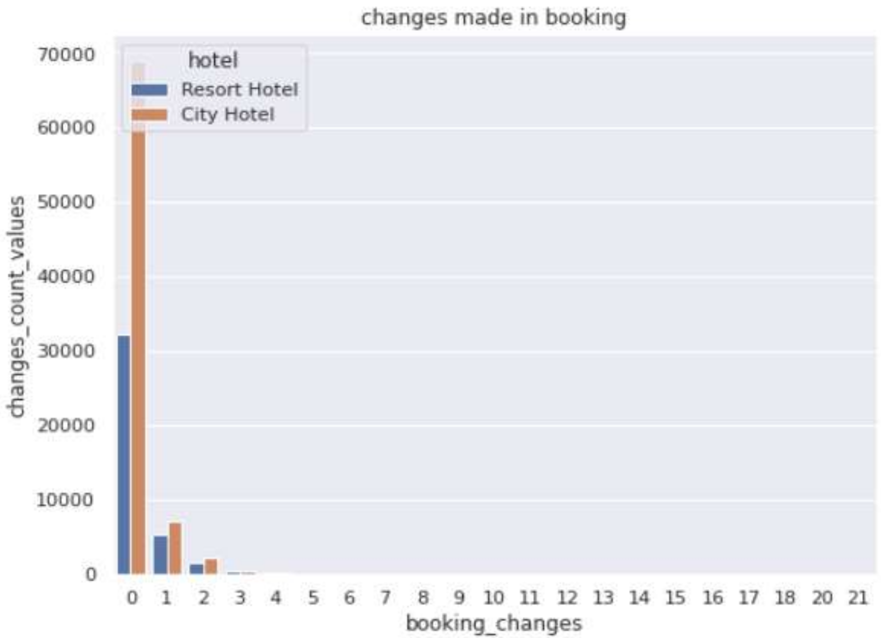


CONCLUSION :

- City hotels are most preferred hotel type by the guest by 66.4%
- Only 3.2% person were revisited the hotels,rest 96% were new guest
- 37% booking were got cancelled out of all the bookings

EDA(continued):

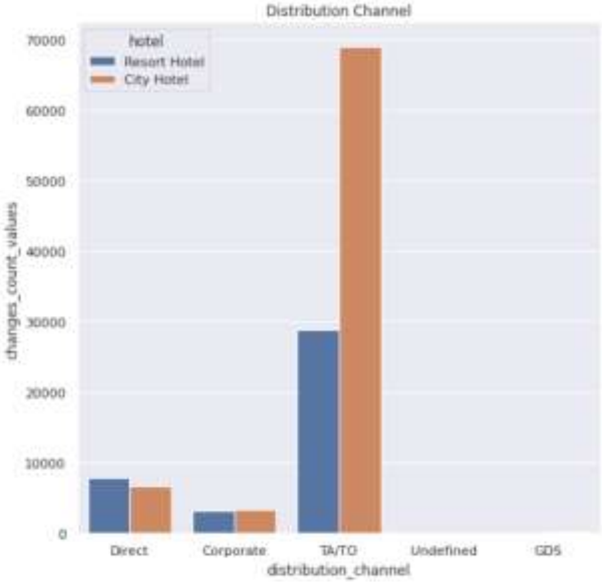


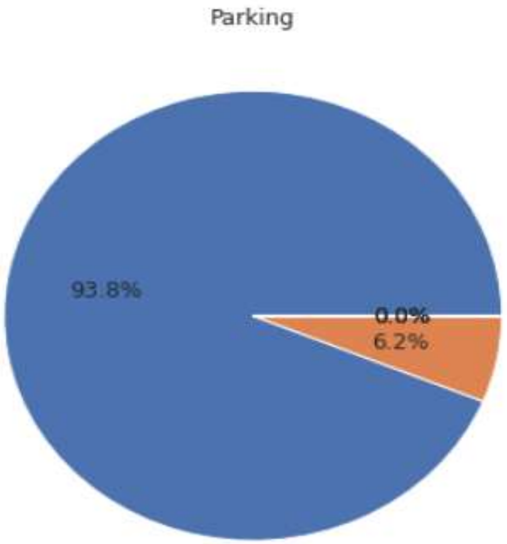


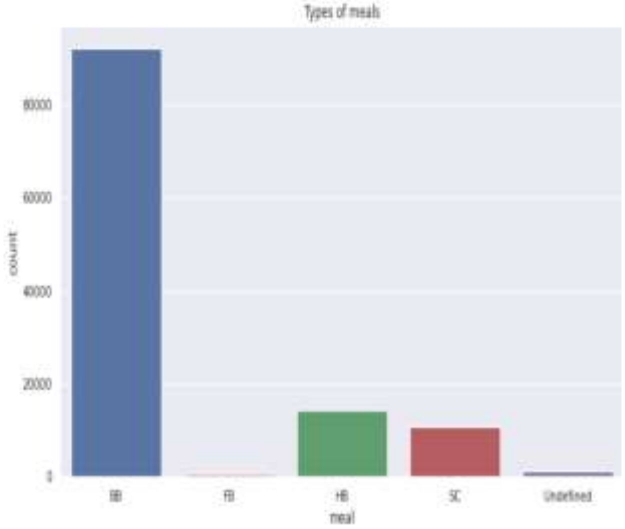
CONCLUSION :

- As we can see in graph,from march to september there is increase in booking of hotels and decreases after september.
- There are more number of guests who makes no changes,as we see in graph there are more changes in city hotel than resort as there are high amount of bookings.

EDA(continued):



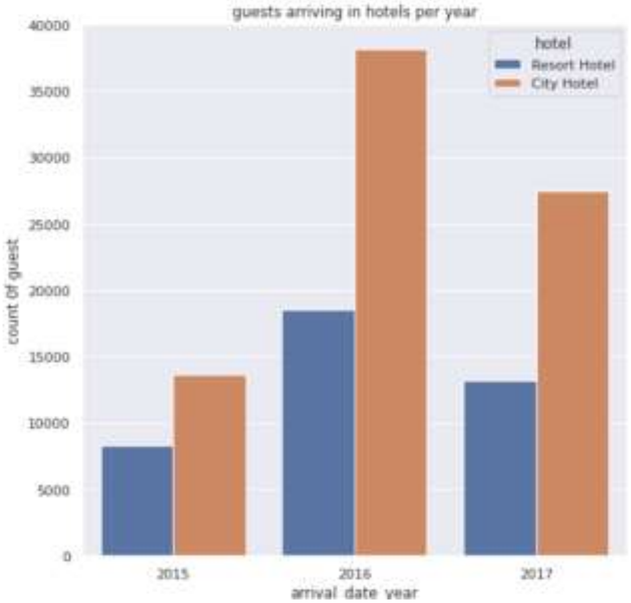


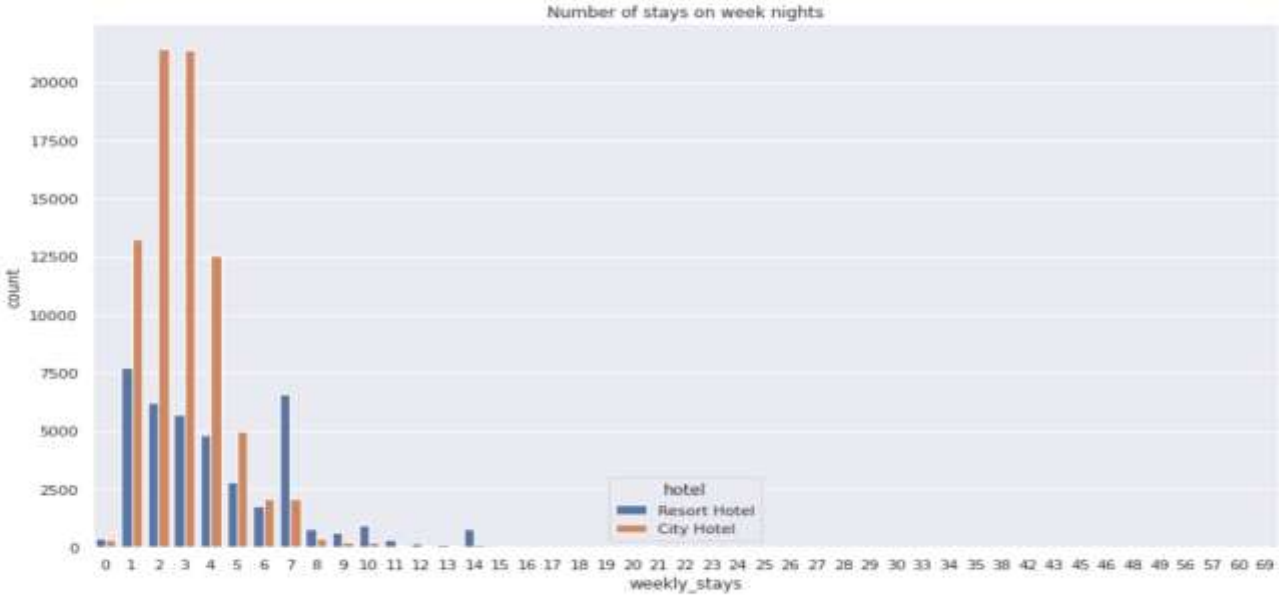


CONCLUSION :

- Majority of dIstribution channel were Travel agencies/Travel operators(TA/TO)..
- Most of guests about 93.8% does not required parking space in hotel and about 6.2% guests required parking space in hotel.
- Out of the meals, BB (Bed & Breakfast) is the most ordered meal, followed by HB(Half Board), SC(no meal package), Undefined and FB (Full Board).

EDA(continued):



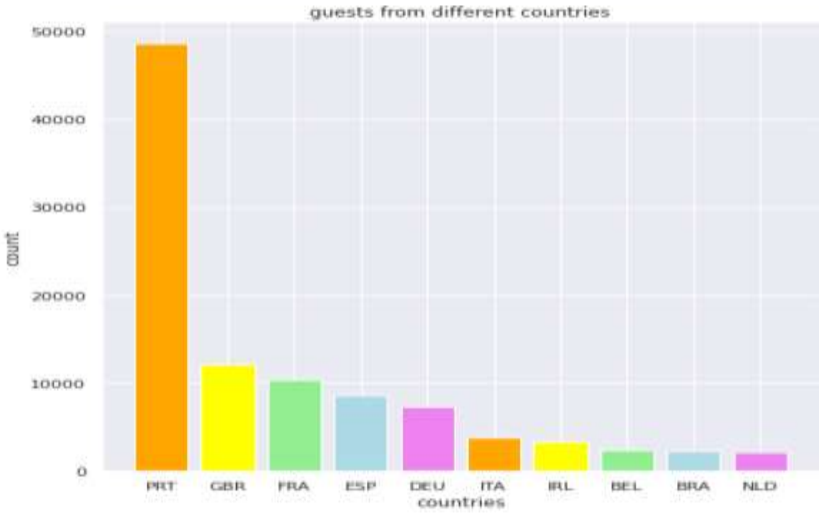


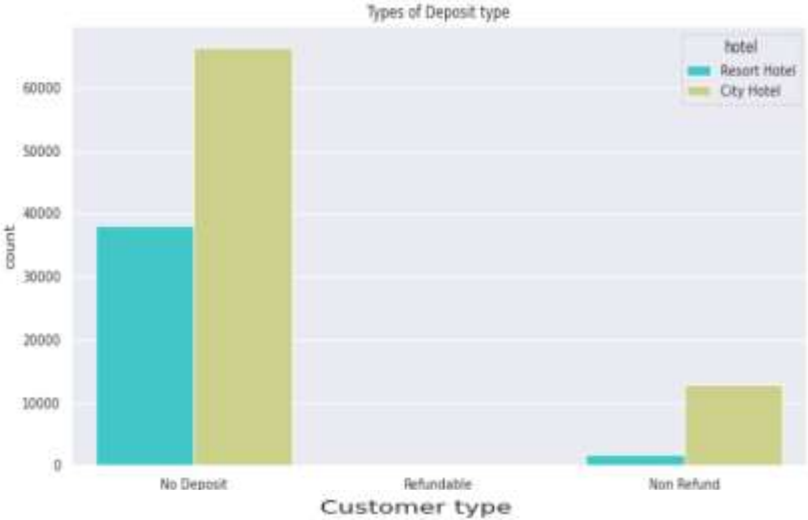
CONCLUSION :

- As per data in 2016 max. number of guests(56707) followed by years 2017(40687) and 2015(21996) have booked there hotels in which they preferred city hotels.
- Guests stay at least 2 to 3 nights on an avg in city hotels.

As the days increases guest choose resorts over the city hotels.

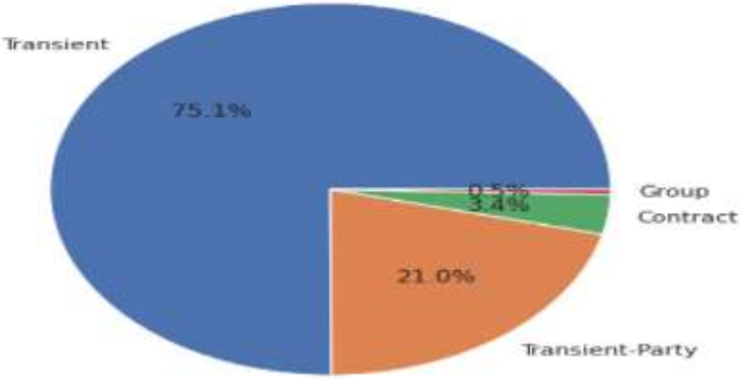
EDA(continued):



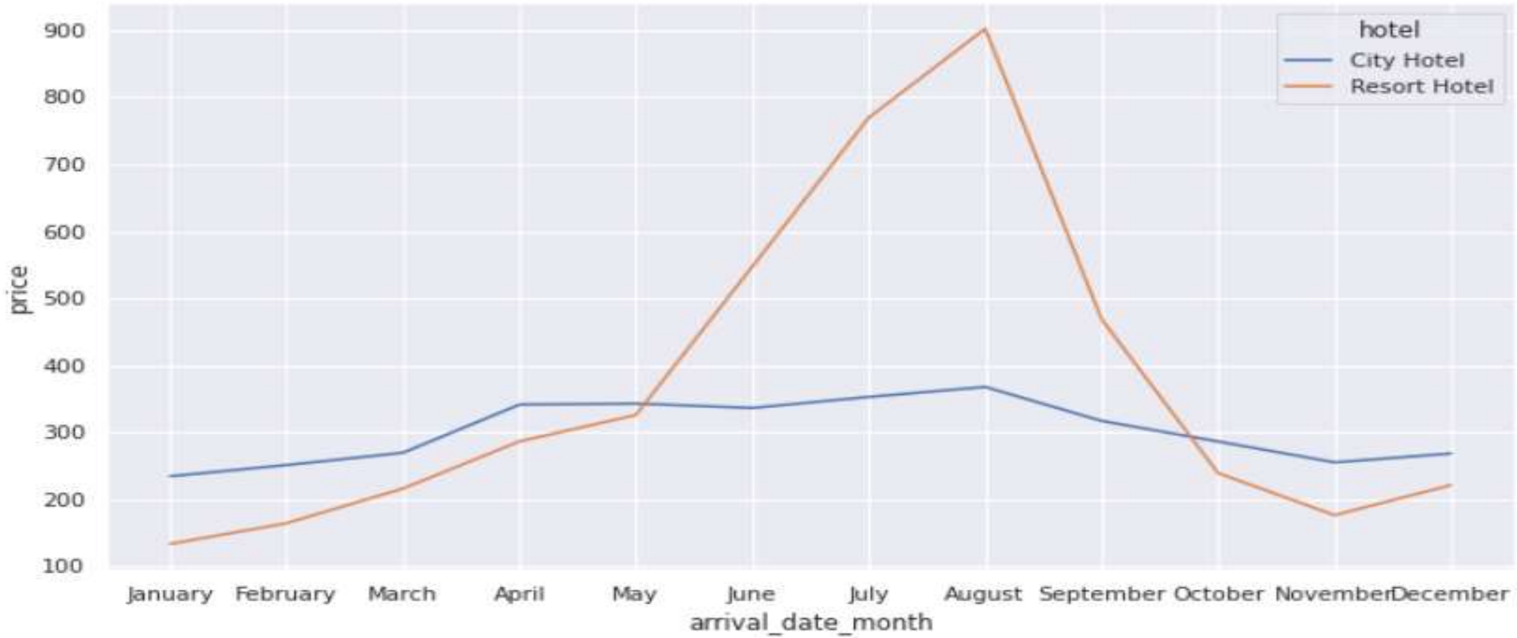


CONCLUSION :

- About 41% of total guests are from Portugal followed by Great Britain(10%) and France(9%).
- Maximum guests preferred no deposit for City hotel where as Resorts hotels had some non-refundable deposits.
- There are 75% of customer are of transient type followed by transient-party(21%),contract(3%)



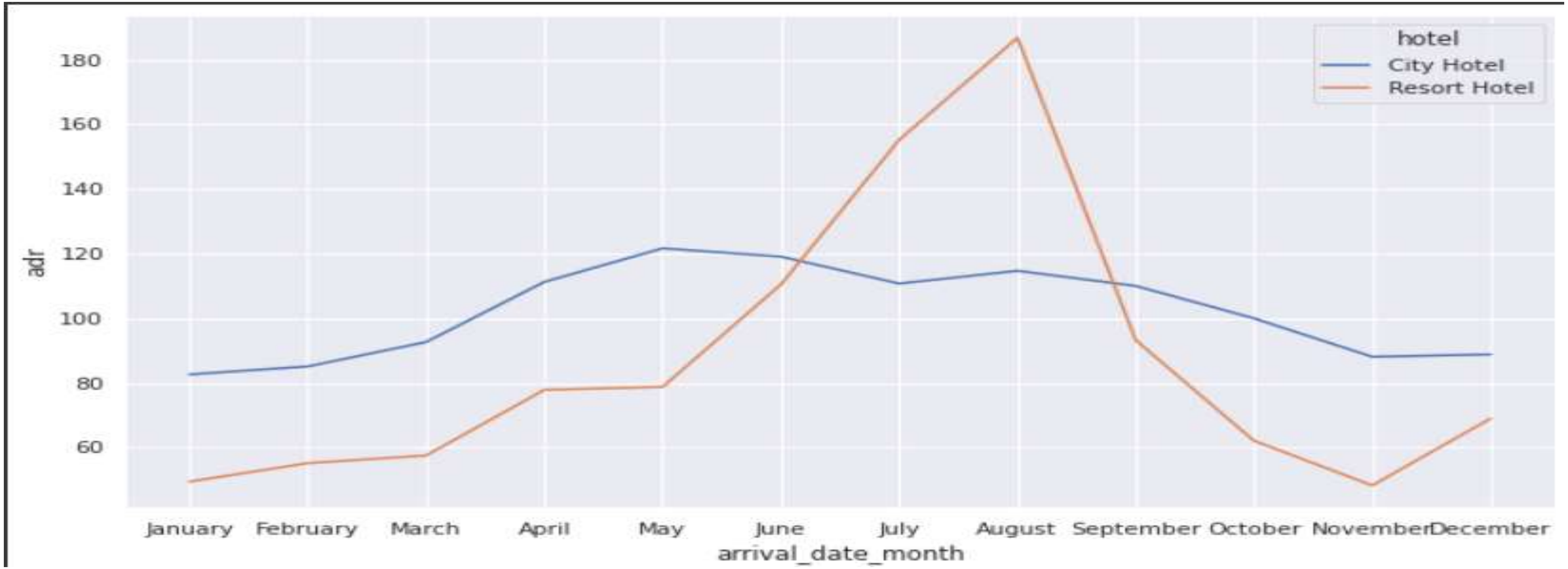
EDA(continued):



CONCLUSION :

- For Resort Hotel, ADR is more expensive during July, August & September and for City Hotel, ADR is slightly more during March, April & May.
- Thus we can say that , in January,February,March,April,November and December are good months to get cheaper hotels.

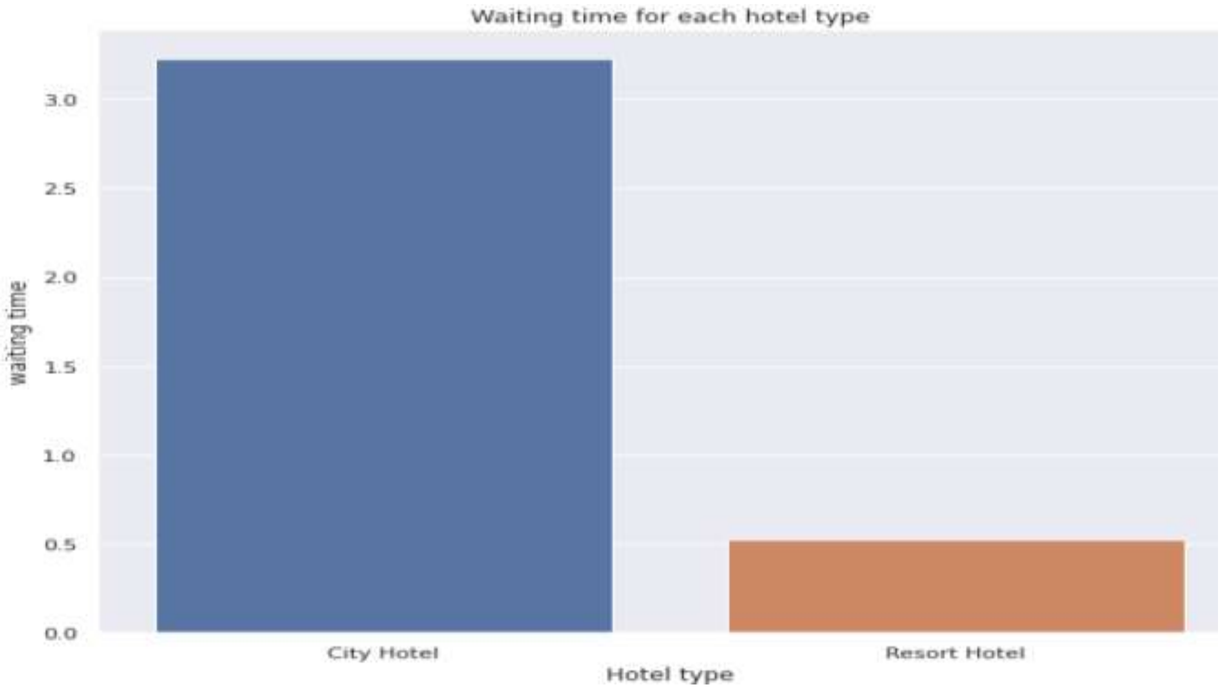
EDA(continued):

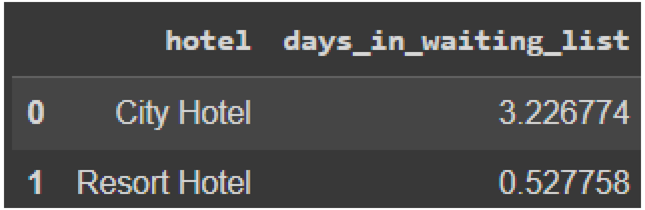


CONCLUSION :

- The avg arrival of guest in city hotels are increasing from the month of May to August .
- The avg arrival of guest in resort hotels is slightly increases in the month of April to June .
- Thus we can say that most of guests preference is city hotels and on an avg there are almost same amount of guest arrived in resort hotels in different months in an year.

EDA(continued):



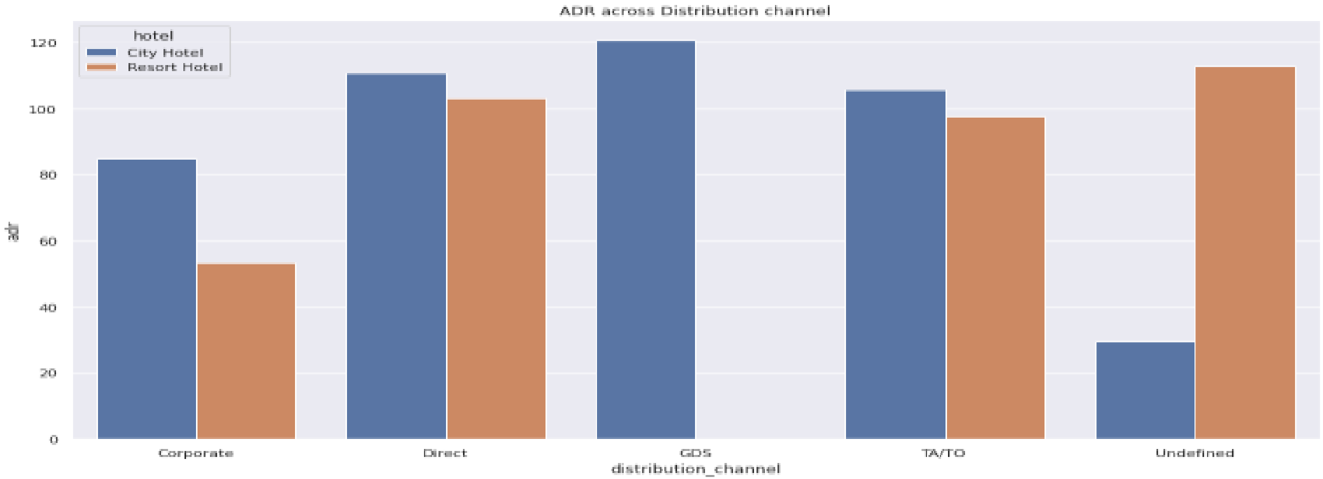


	hotel	days_in_waiting_list
0	City Hotel	3.226774
1	Resort Hotel	0.527758

CONCLUSION :

- Waiting time period for City hotels is high as compared to Resort hotels with the avg of 3.22 days in City hotels and approx ½ day of waiting in Resort hotels.
- That means City hotels are much busier than Resort hotels.

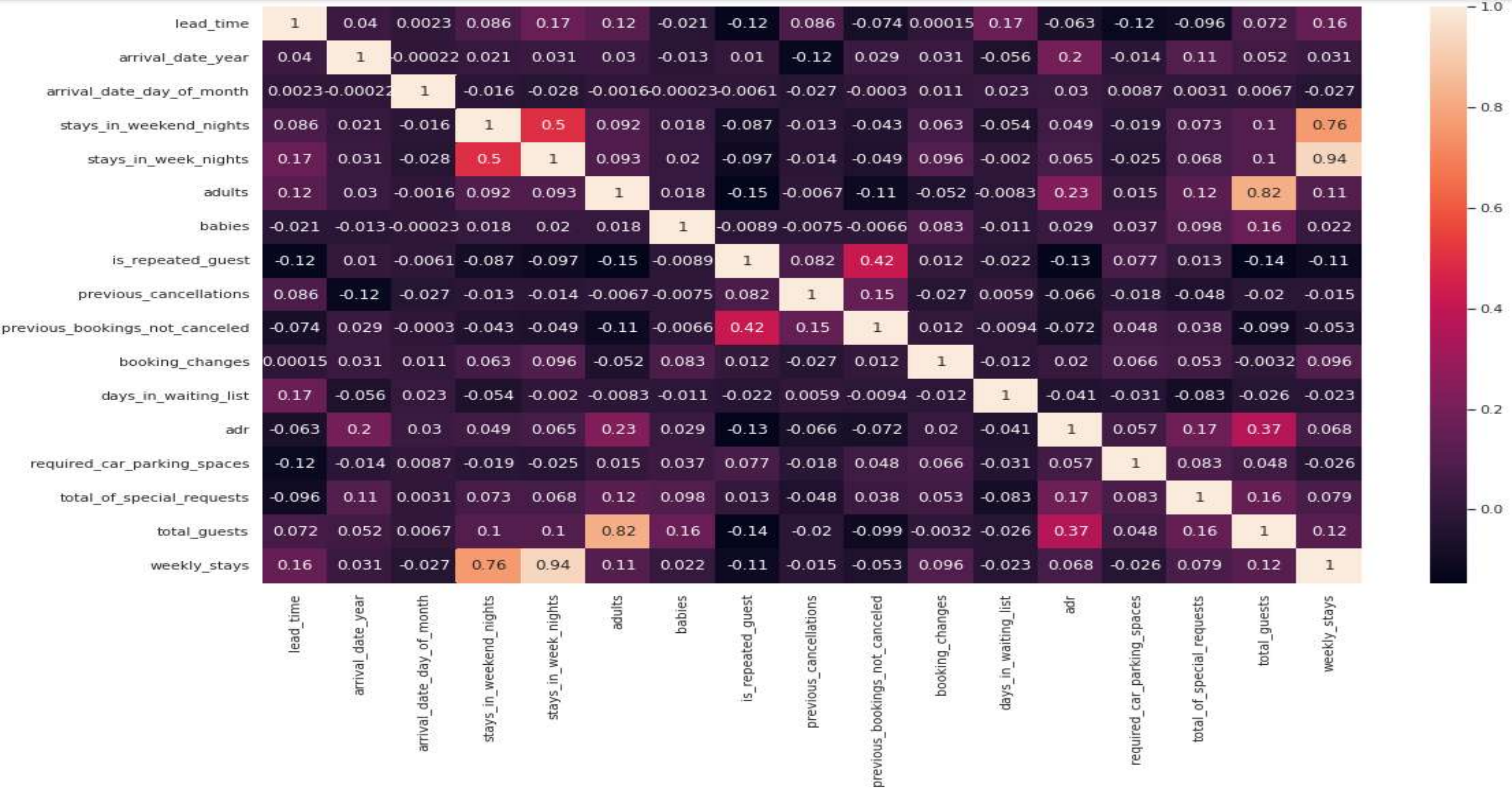
EDA(continued):



CONCLUSION :

- Direct and TA/TO has almost equal in adr and high from other distribution channels in both hotels.
- GDS has high adr in City hotels type. GDS needs to increase in resort type.
- From this we can say that Direct and TA/TO are generating more revenue than other distribution channels.

EDA(continued):



	lead_time	arrival_date_year	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	babies	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests	total_guests	weekly_stays
lead_time	1	0.04	0.0023	0.086	0.17	0.12	-0.021	-0.12	0.086	-0.074	0.00015	0.17	-0.063	-0.12	-0.096	0.072	0.16
arrival_date_year	0.04	1	-0.00022	0.021	0.031	0.03	-0.013	0.01	-0.12	0.029	0.031	-0.056	0.2	-0.014	0.11	0.052	0.031
arrival_date_day_of_month	0.0023	-0.00022	1	-0.016	-0.028	-0.0016	-0.00023	-0.0061	-0.027	-0.0003	0.011	0.023	0.03	0.0087	0.0031	0.0067	-0.027
stays_in_weekend_nights	0.086	0.021	-0.016	1	0.5	0.092	0.018	-0.087	-0.013	-0.043	0.063	-0.054	0.049	-0.019	0.073	0.1	0.76
stays_in_week_nights	0.17	0.031	-0.028	0.5	1	0.093	0.02	-0.097	-0.014	-0.049	0.096	-0.002	0.065	-0.025	0.068	0.1	0.94
adults	0.12	0.03	-0.0016	0.092	0.093	1	0.018	-0.15	-0.0067	-0.11	-0.052	-0.0083	0.23	0.015	0.12	0.82	0.11
babies	-0.021	-0.013	-0.00023	0.018	0.02	0.018	1	-0.0089	-0.0075	-0.0066	0.083	-0.011	0.029	0.037	0.098	0.16	0.022
is_repeated_guest	-0.12	0.01	-0.0061	-0.087	-0.097	-0.15	-0.0089	1	0.082	0.42	0.012	-0.022	-0.13	0.077	0.013	-0.14	-0.11
previous_cancellations	0.086	-0.12	-0.027	-0.013	-0.014	-0.0067	-0.0075	0.082	1	0.15	-0.027	0.0059	-0.066	-0.018	-0.048	-0.02	-0.015
previous_bookings_not_canceled	-0.074	0.029	-0.0003	-0.043	-0.049	-0.11	-0.0066	0.42	0.15	1	0.012	-0.0094	-0.072	0.048	0.038	-0.099	-0.053
booking_changes	0.00015	0.031	0.011	0.063	0.096	-0.052	0.083	0.012	-0.027	0.012	1	-0.012	0.02	0.066	0.053	-0.0032	0.096
days_in_waiting_list	0.17	-0.056	0.023	-0.054	-0.002	-0.0083	-0.011	-0.022	0.0059	-0.0094	-0.012	1	-0.041	-0.031	-0.083	-0.026	-0.023
adr	-0.063	0.2	0.03	0.049	0.065	0.23	0.029	-0.13	-0.066	-0.072	0.02	-0.041	1	0.057	0.17	0.37	0.068
required_car_parking_spaces	-0.12	-0.014	0.0087	-0.019	-0.025	0.015	0.037	0.077	-0.018	0.048	0.066	-0.031	0.057	1	0.083	0.048	-0.026
total_of_special_requests	-0.096	0.11	0.0031	0.073	0.068	0.12	0.098	0.013	-0.048	0.038	0.053	-0.083	0.17	0.083	1	0.16	0.079
total_guests	0.072	0.052	0.0067	0.1	0.1	0.82	0.16	-0.14	-0.02	-0.099	-0.0032	-0.026	0.37	0.048	0.16	1	0.12
weekly_stays	0.16	0.031	-0.027	0.76	0.94	0.11	0.022	-0.11	-0.015	-0.053	0.096	-0.023	0.068	-0.026	0.079	0.12	1

CONCLUSION :

- is_canceled and same_room_alloted_or_not are negatively correlated. That means customer is unlikely to cancel his bookings if he don't get the same room as per reserved room. We have visualized it above.
- lead_time and total_stay is positively correlated.That means more is the stay of coutsmer more will be the lead time.
- adults,childrens and babies are correlated to each other. That means more the people more will be adr.r.

SUMMARY:

- Most bookings of City and Resort hotels are made in year 2016.
- City hotels are most preferred by guests as it was booked by 66.4% of guests.Thus we can say that city hotels are the busiest hotel.
- About 37% of guest cancelled there reservation hotels in which City hotels have high cancelation rate than Resorts hotels.
- Majority of guest are from western Europe i.e Portugal,France,Great Britain
- Most guests are not preferred to made there changes in current booking.
- Only 3.2% people were revisited the hotels,rest 96.8% were new guests,thus retention rate is low.
- About 93.8% of guests does not required parking.
- BB(Bed and Breakfast) are the most preferred type of meal choose by guests.
- Booking made by TA/TO higher than the other distribution channels.
- On an avg guests stay for 2 to 3 nights in city hotels but as night of stays increases guests prefer resort hotels rather than city hotels.
- About 55.6% guests of city hotels and 32% guests of resort hotels choose no deposit as deposit type.
- Most of customers are Transient or Transient-party type of guests
- Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
- Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.
- Direct and TA/TO are generating more revenue than other distribution channels.

- Set Non-refundable Rates, Collect deposits, and implement more rigid cancellation policies.
- Encourage Direct bookings by offering special discounts
- Monitor where the cancellations are coming from such as Market Segment, distribution channels
- We should also target months between May to Aug. Those are peak months due to the summer period.
- Given that we do not have repeated guests, we should target our advertisement on guests to increase returning guests.
- Take feedbacks from guests and try to improve according to it .

Thank You