Capstone Project 1

Hotel Booking Analysis

Priyanka K

Steps Involved

- Defining Problem statement
- Data Exploration
- Data Cleaning
- Exploratory Data Analysis
    - Univariate Analysis
    - Bivariate Analysis
    - Multivariate Analysis
- Visualizing the Data
- Driving Conclusions

Problem Statement

- A hotel data set is given to us, which contains booking information for a city hotel and a resort hotel, and includes information such as when are the bookings made, length of stay, the number of adults, children, and babies, and the number of available parking spaces, among other things.
- The hotel industry is highly volatile, with numerous factors influencing bookings. With the details provided in the data set, be expected to do an exploratory data analysis and find out the factors that govern the bookings, which hotel the customers prefer, and so on.

Data summary

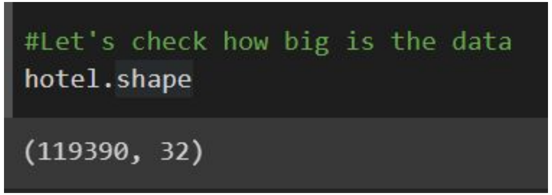
- **Hotel:** Whether the booking is for City hotel or Resort hotel
- **Is_canceled: Is** the booking canceled or not
- **Lead_time:** The number of days elapsed between the booking and the arrival date
- **Arrival_date_year:** Year of arrival date
- **Arrival_date_month:** Month of the arrival date
- **Arrival_date_week_number:** The week number for which the guest is going to visit.
- **Arrival_date_day_of_month:** Day of the arrival date
- **Stays_in_weekend_nights:** Number of weekend night stay
- **Stays_in_week_nights:** Number of weekday night stay
- **Adults:** Number of adults
- **Children:** Number of children
- **Babies:** Number of babies
- **Meal:** Type of meal preferred
- **Country**: Country code of the guest
- **Market_segment:** The market segment of the booking

- **Distribution_channel:** By which market segment customer access the stay
- **Is_repeated_guest:** Whether the guest stays for the first time or not
- **Previous_cancellations:** Are there any previous cancellations
- **Previous_bookings_not_canceled:** Count of the prior bookings canceled
- **Reserved_room_type:** Room type preferred by the guest
- **Assigned_room_type:** Assigned room for the guest
- **Booking_changes:** Count of changes made to the booking
- **Deposit_type:** Deposit type opted for the booking
- **Agent**: Agent data for the booking
- **Company:** Company to which the guest belongs
- **Days_in_waiting_list:** Number of days on the waiting list
- **Customer_type:** Customer type to which the booking belongs
- **ADR:** Revenue generated by the hotel through this booking
- **Required_car_parking_spaces:** Is car parking is required
- **Total_of_special_requests:** Number of special requests by the guest.
- **Reservation_status**: Reservation status of the booking
- **Reservation_status_date:** Date of the reservation status for the booking

Exploration of Data

Shape of the data:

The data set has 119390 observations and 32 columns

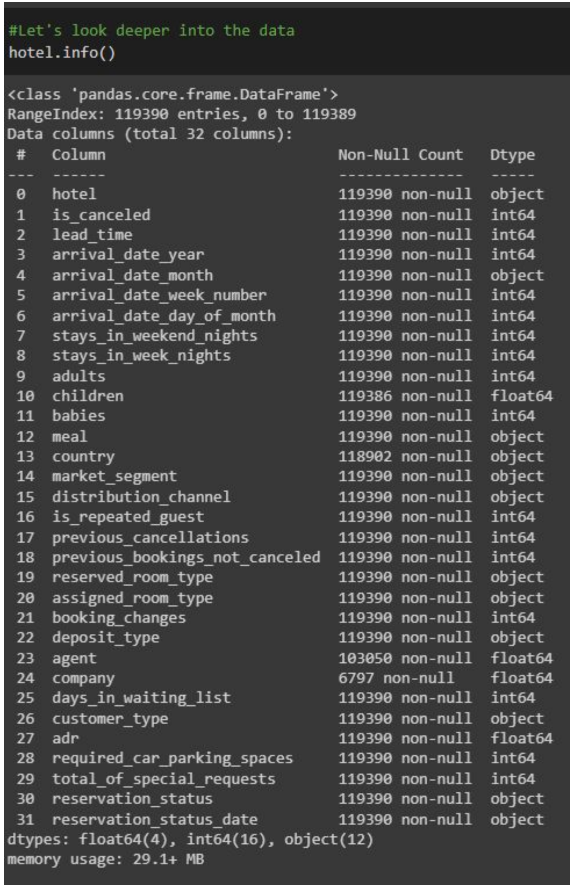


```
#Let's check how big is the data
hotel.shape

(119390, 32)
```

Information about columns:

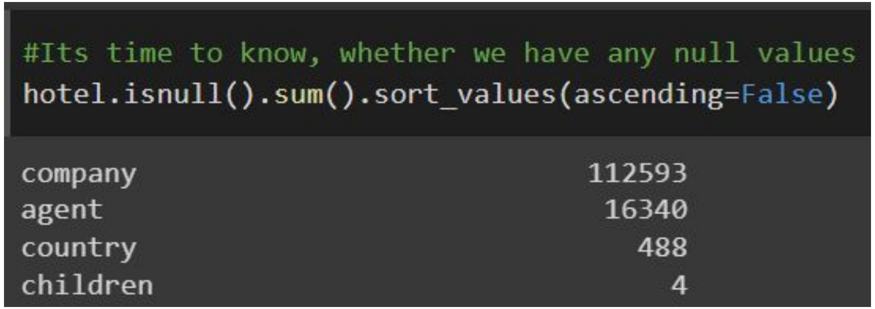
- There is a mix of datatypes in the data such as objects,int and float
- There are few columns which has null values



```
#Let's look deeper into the data
hotel.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 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
```

Dealing with Null values



```
#Its time to know, whether we have any null values
hotel.isnull().sum().sort_values(ascending=False)
```

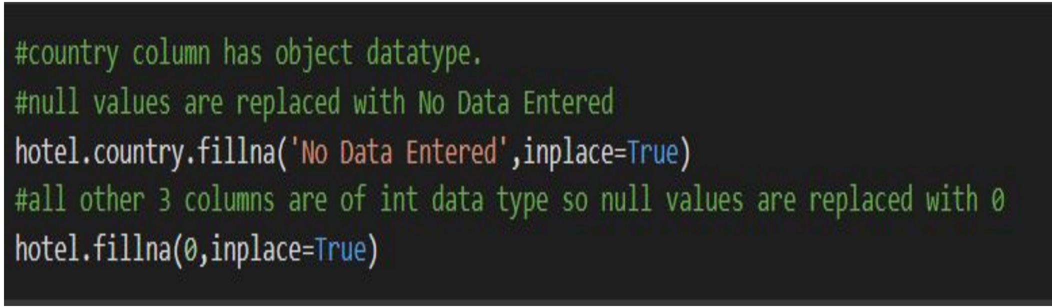
```
company           112593
agent              16340
country              488
children               4
```

Columns with null values:

There are 4 columns with null values.They are company,agent,country and children.

Dealing with null values:

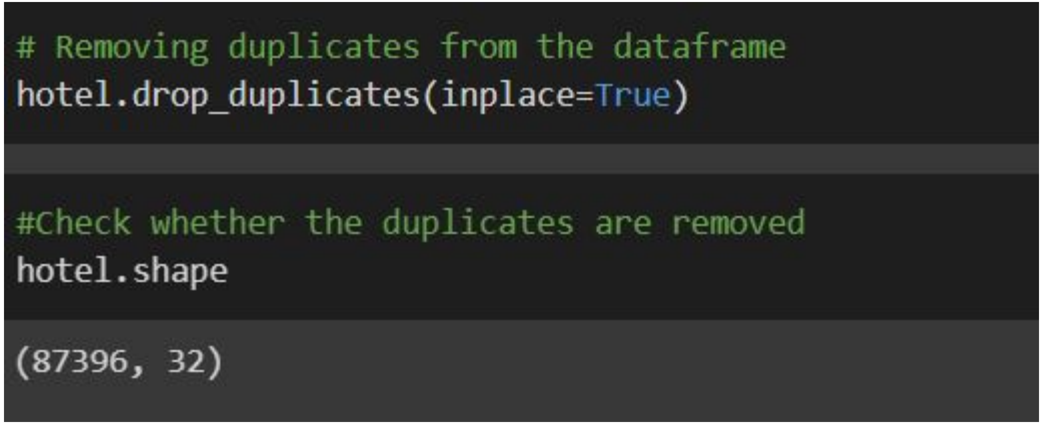
- All null values have been replaced as ‘No Data Entered’.
- Even children column with null values is replaced with text.Because, if we replace it with 0.It might affect the analysis later.



```
#country column has object datatype.
#null values are replaced with No Data Entered
hotel.country.fillna('No Data Entered',inplace=True)
#all other 3 columns are of int data type so null values are replaced with 0
hotel.fillna(0,inplace=True)
```

Dealing with Duplicates

- Duplicates from the dataframe are removed for more accurate analysis
- Shape of the dataframe is now changed after the removal of duplicates.



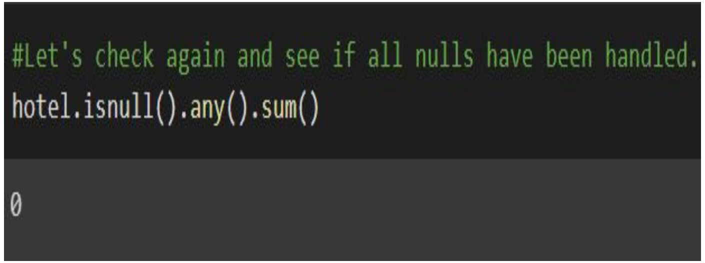
```
# Removing duplicates from the dataframe
hotel.drop_duplicates(inplace=True)
```

```
#Check whether the duplicates are removed
hotel.shape
```

```
(87396, 32)
```

Let’s check if all nulls are being replaced:

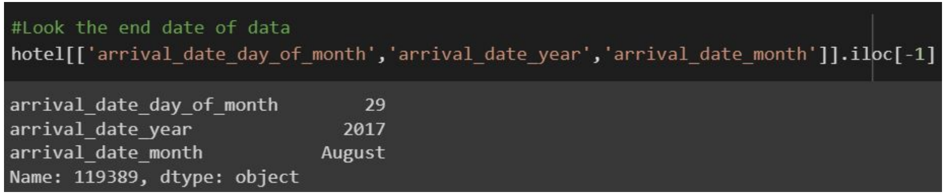
- Yes, all nulls have been replaced and it is checked with .isnull method



```
#Let's check again and see if all nulls have been handled.
hotel.isnull().any().sum()
```

```
0
```

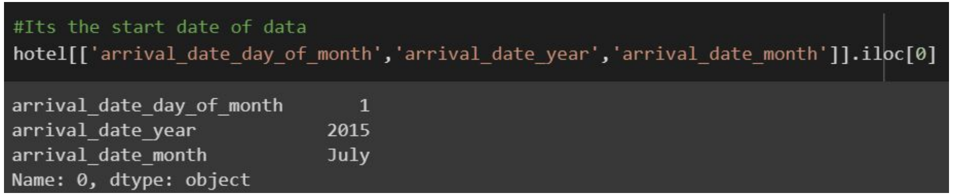
Start and end of the data:



```
#Look the end date of data
hotel[['arrival_date_day_of_month','arrival_date_year','arrival_date_month']].iloc[-1]
```

```
arrival_date_day_of_month        29
arrival_date_year              2017
arrival_date_month           August
Name: 119389, dtype: object
```

- The data starts from 1st of July 2015 and ends by 29th August 2017

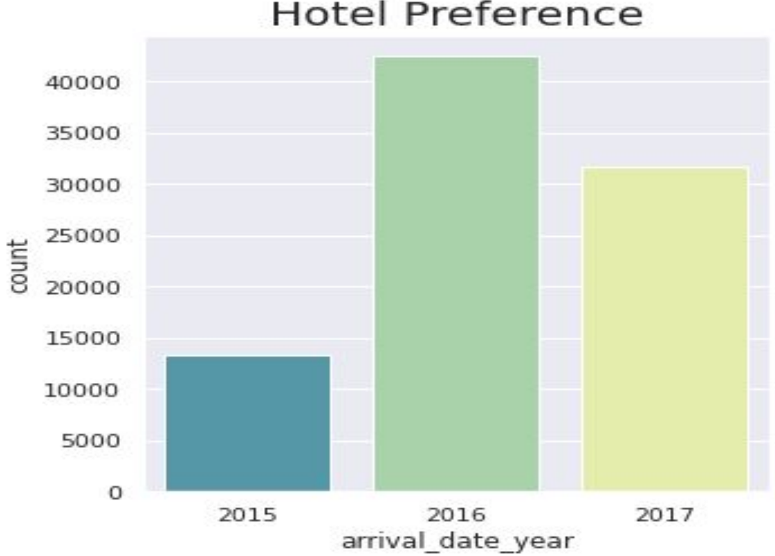


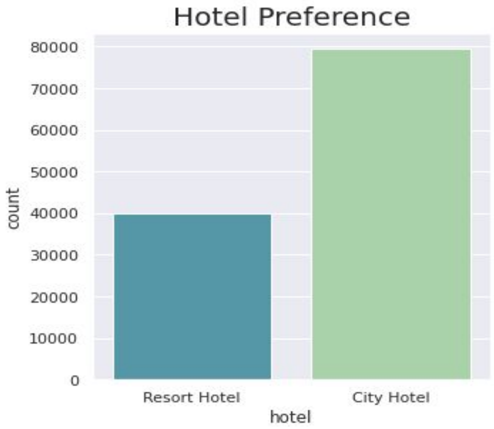
```
#Its the start date of data
hotel[['arrival_date_day_of_month','arrival_date_year','arrival_date_month']].iloc[0]
```

```
arrival_date_day_of_month       1
arrival_date_year            2015
arrival_date_month           July
Name: 0, dtype: object
```

Number of Bookings Over Years

- The data is spread over three years 2015,2016 and 2017.
- Total bookings in 2015 are 13313
- Total bookings in 2016 are 42391
- Total bookings in 2017 are 31692
- The bookings may be higher in 2016 because it has 12 months of data.

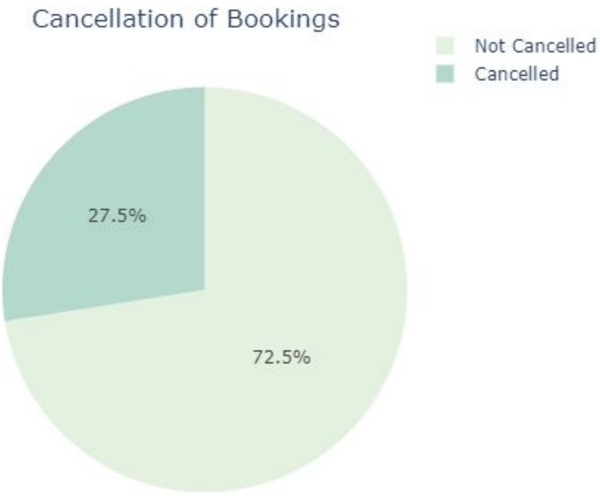




Hotel Preferences of customers

- More bookings are done for city hotels compared to resort hotels
- City hotel bookings count is 53428
- Resort hotel bookings count is 33968

Percentage of Bookings being cancelled

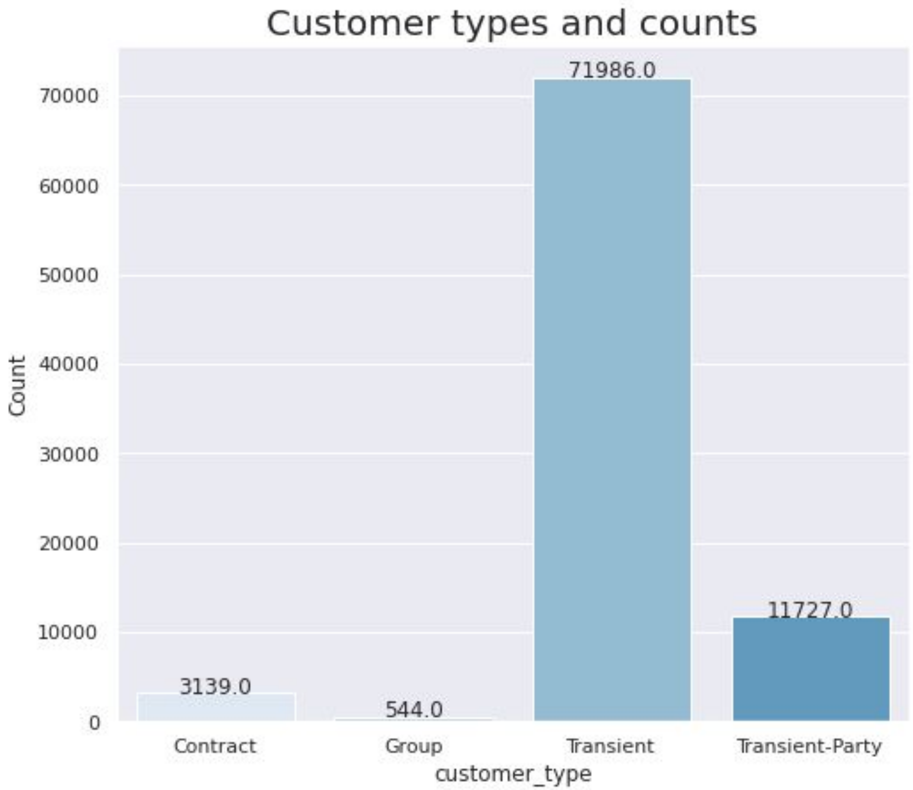


- Around 72% of the bookings are unchanged, and 27.5% of the total bookings are canceled.
- The number of bookings canceled is 24025
- The number of bookings not canceled is 63371

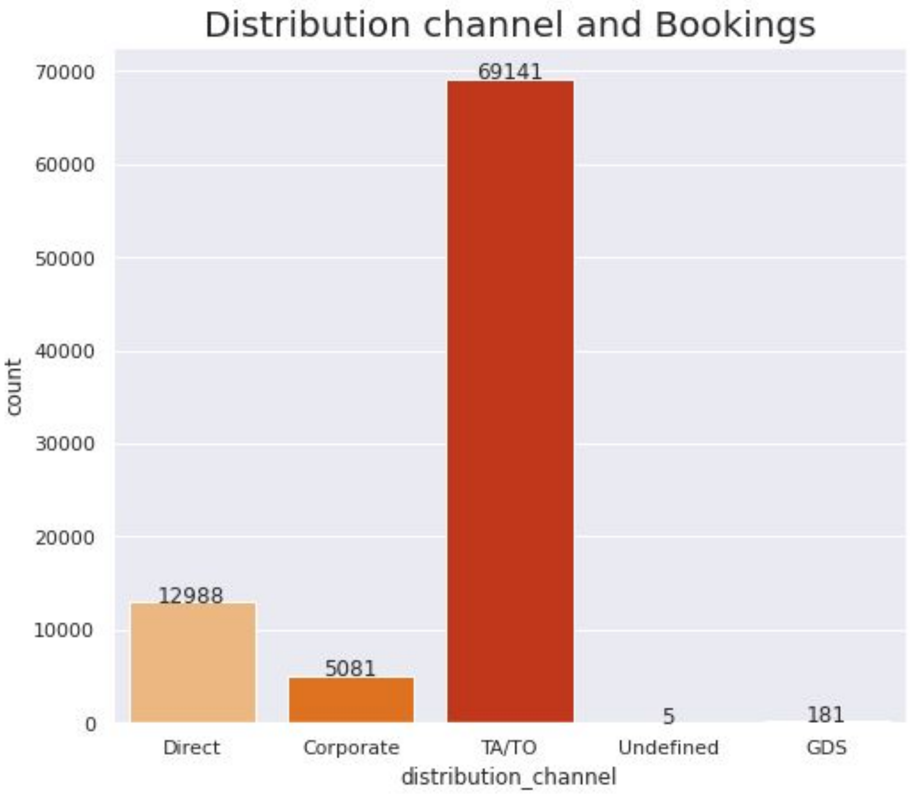
Different types of customers

There are four different customer types namely contract, group, Transient and Transient party. Most of the bookings are done. by transient type of customers.

- 71986 observations are of transient type.
- 11727 observations are from Transient party type customers.
- 3139 observations are from contract type of customers.
- 544 observations are from group type of customers.



Distribution Channel



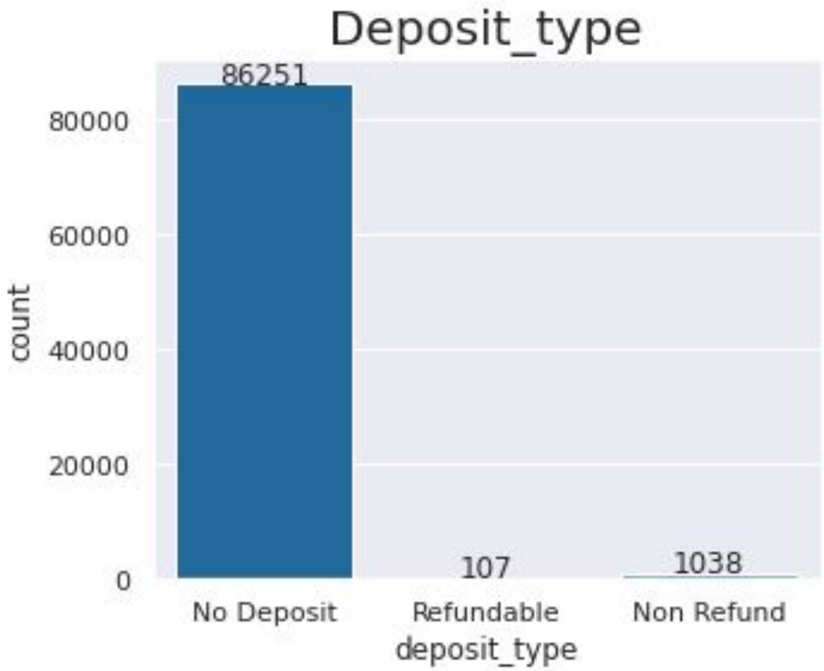
There are five different distribution channel in the data.They are direct,corporate, TA/TO, GDS and undefined.

- Direct - 12988 bookings
- Corporate - 5081 bookings
- TA/TO -69141 bookings
- GDS - 181 bookings
- undefined - 5 bookings

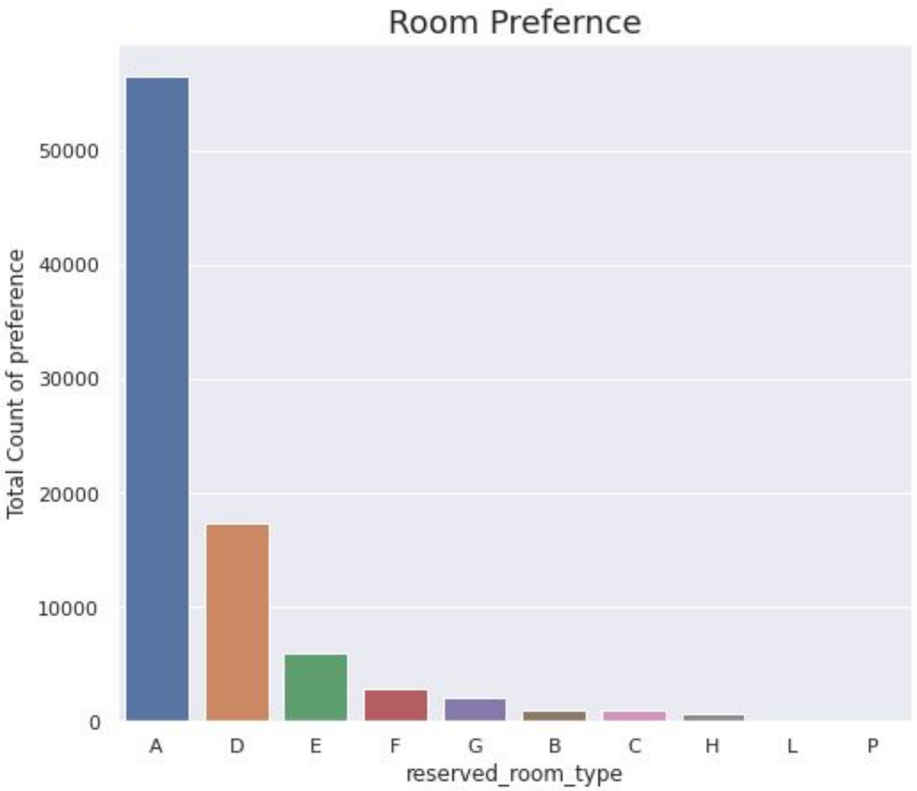
Different Deposit Types used for Bookings

There are three different modes of deposit types in the data

- No Deposit - 86251
- Refundable Deposit - 107
- Non Refund - 1038



Room Preference of Guests



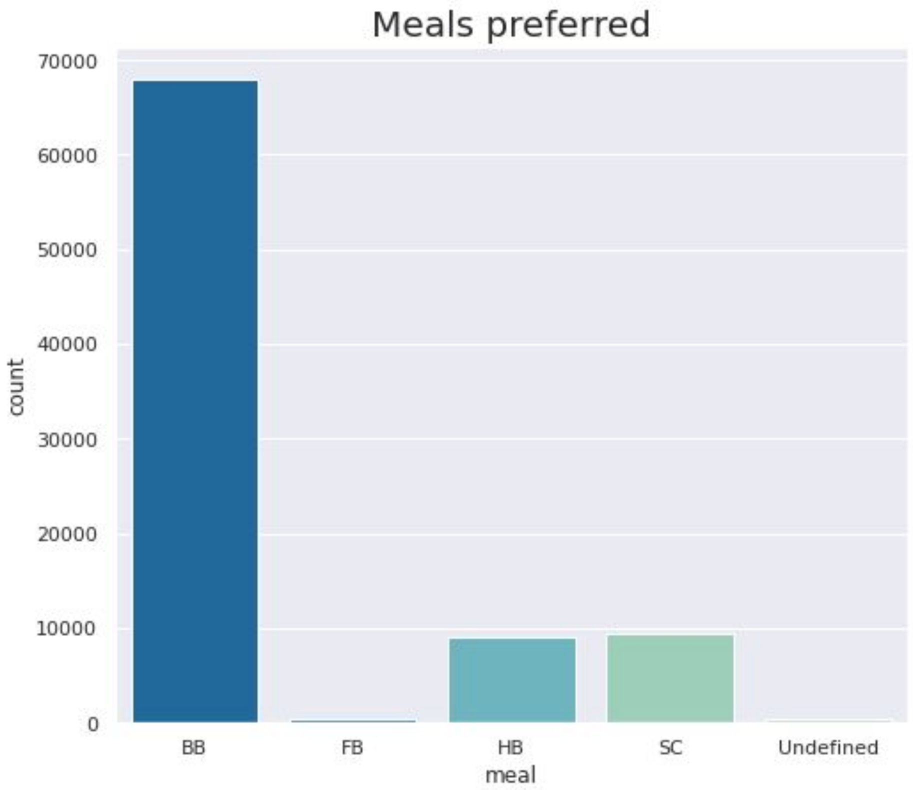
There are 9 different room options in the hotel. But the most preferred room is type A with more bookings.

- A - 56552
- B - 999
- C - 915
- D - 17398
- E - 6049
- F - 2823
- G - 2052
- H - 596
- L - 6
- P - 6

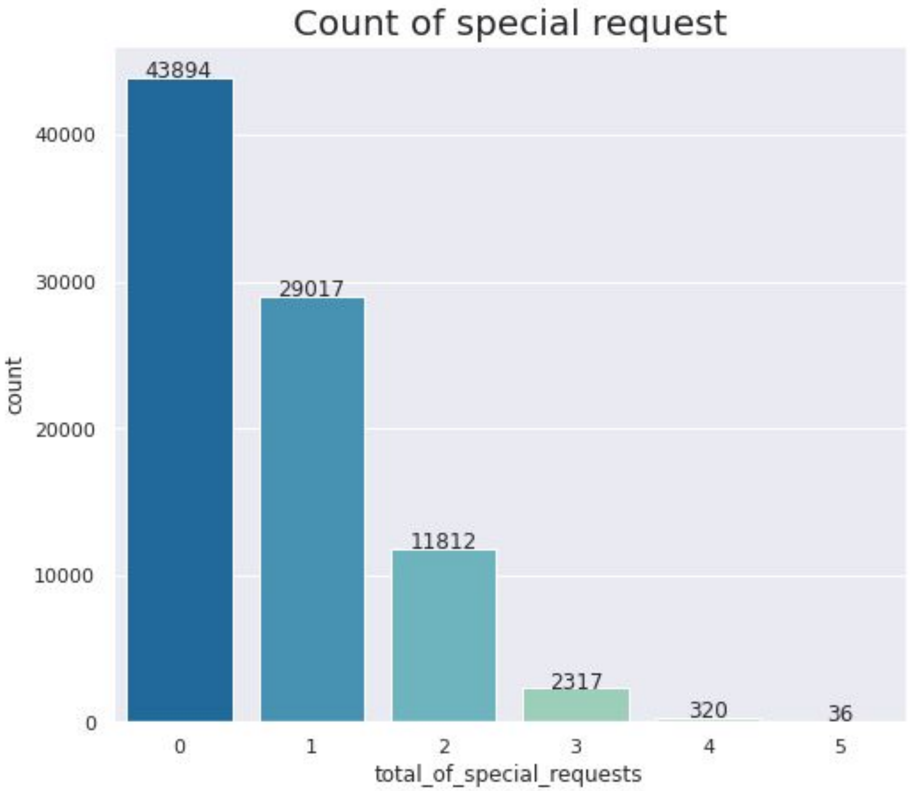
Meals preferred by the customers

There are different kinds of meals provided in the hotel, the preference counts are

- BB - 67978
- SC - 9481
- HB - 9285
- Undefined - 492
- FB - 360



How many clients have special requests



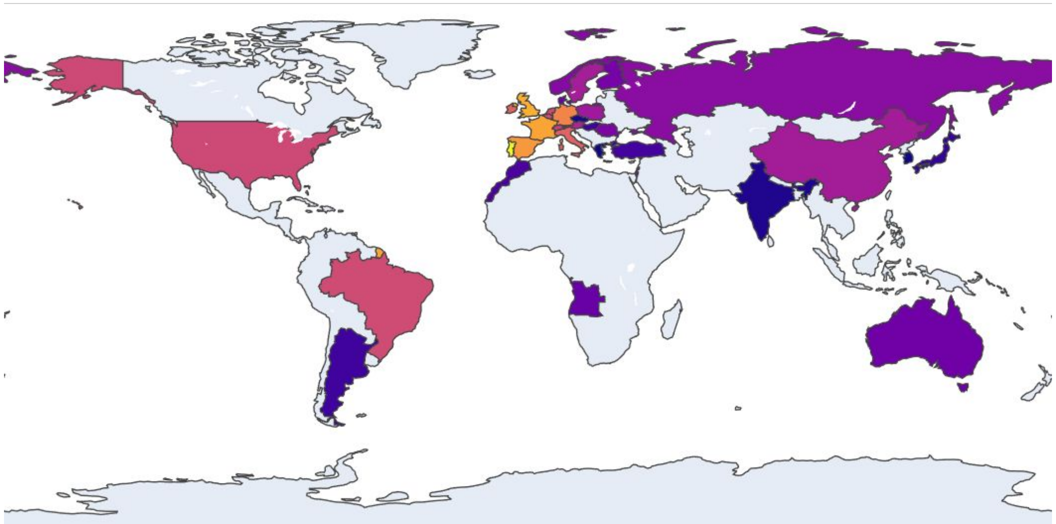
The maximum number of special request by the guest is 5. The count of special requests are

- 0 request- 43894
- 1 request- 29017
- 2 requests-11812
- 3 requests-2317
- 4 requests- 320
- 5 requests -36

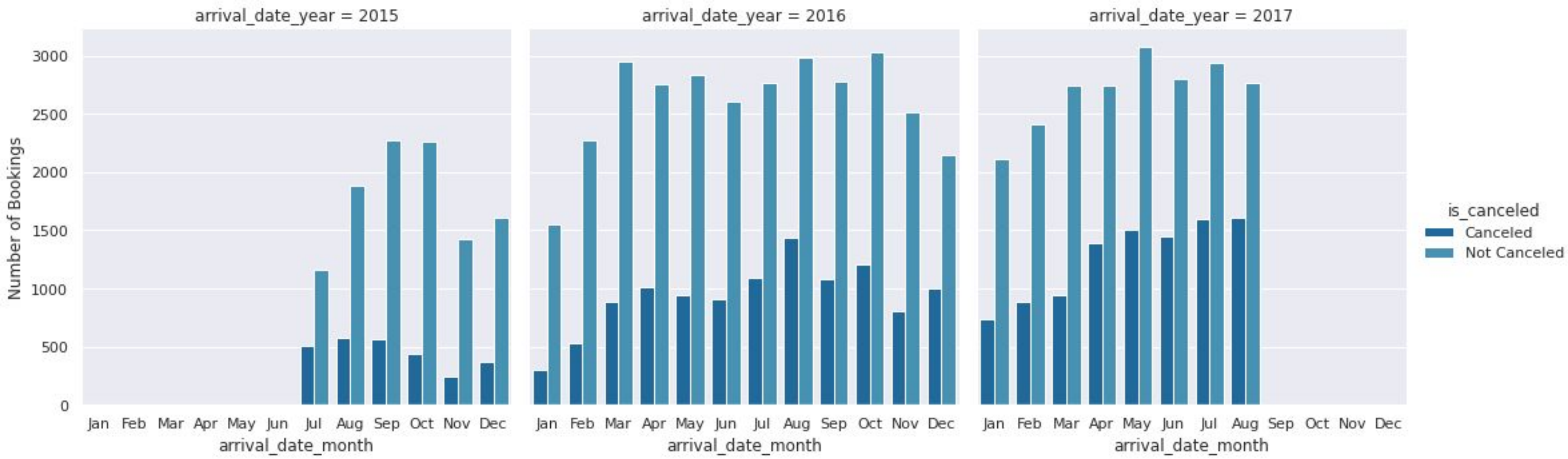
Country from which most clients are from

There are a lot of countries that contribute to bookings. Let's take a look at top few:

- Portugal - 27453
- United Kingdom - 10433
- France - 8837
- Spain - 7252
- Germany - 5387
- Italy - 3066
- Ireland - 3016
- Belgium - 2081
- Brazil - 1995
- Netherlands - 1911
- United States - 1875
- Switzerland - 1570



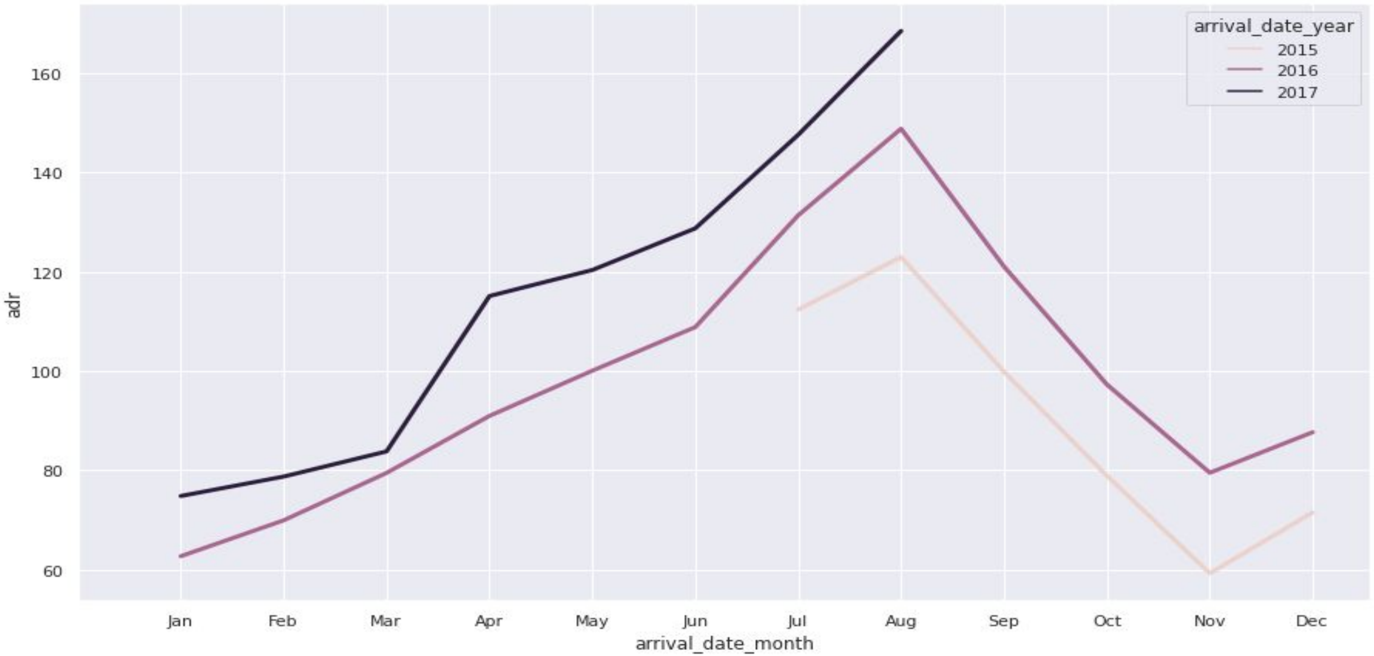
Number of cancellations over the years



- There is a good rise in the number of bookings not canceled when we see the data for all three years. There is a higher number of bookings from May to November. The bookings at year start and end are comparatively low compared to other months. We should put in some work to increase the bookings for the month with fewer bookings.

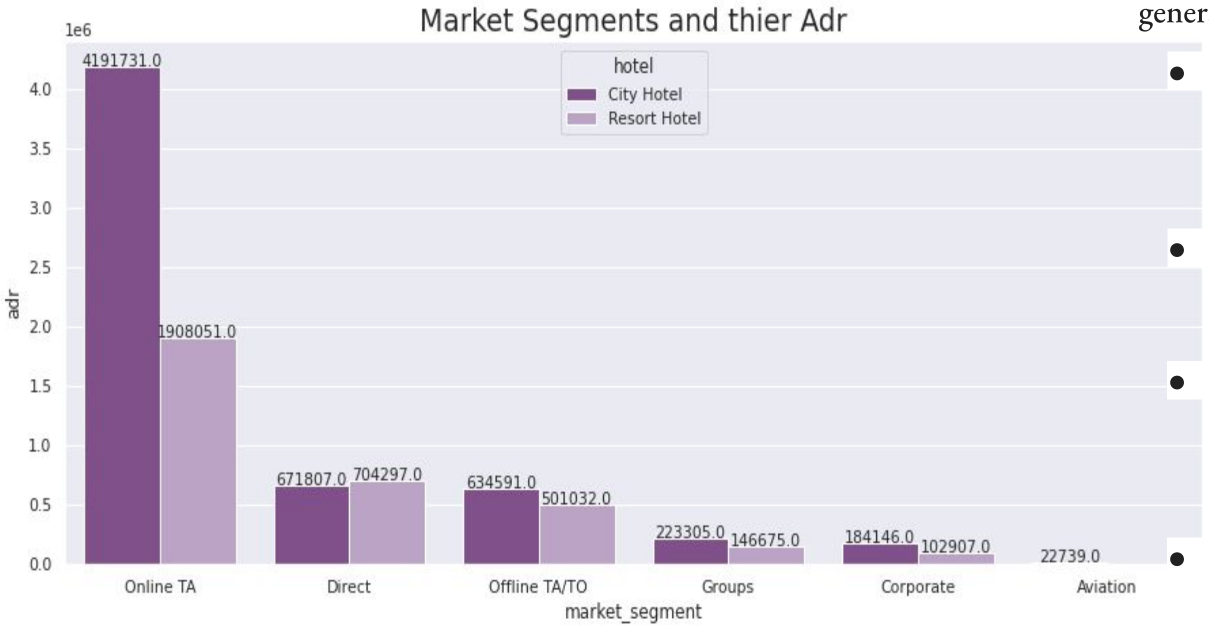
- A consistent rise in booking cancellations is from the start to the end of the data. And few months have about 40% of bookings canceled compared to total bookings, which is not great for the business. We should learn the ropes about the factors for the increase in booking cancellations and take needy actions.

ADR over different months of the years



- The graph shows a continuous growth of ADR from the start of the year till August when the revenue generation is at its peak. Then there is a drop in revenue till November and shows growth in December. The revenue generation is showing consistency year over year for every month.

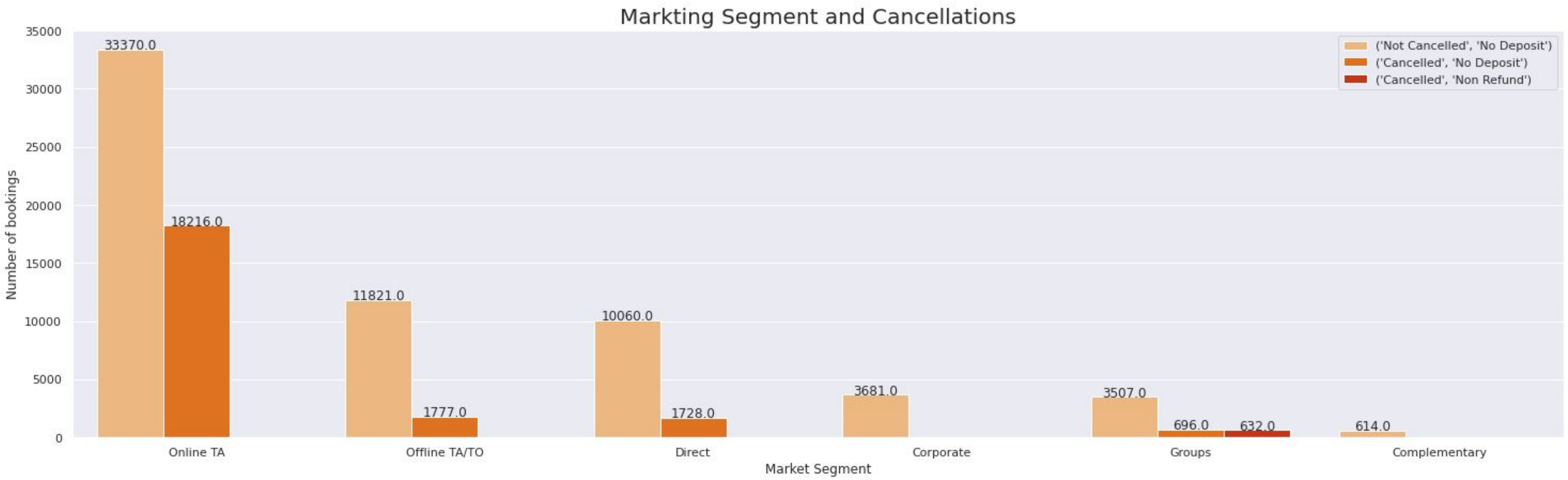
Market Segments and Adr



The different marketing segments and their revenue generation are:

- Online TA is the market segment that generates higher adr for both the hotels.
  - City hotel - 4191731
  - Resort hotel - 190805
- Direct generates the second highest adr
  - City hotel - 671807
  - Resort hotel - 704297
- Offline market segment is the third segment in generating adr
  - City hotel - 634591
  - Resort hotel - 501032
- All other segments such as, Groups, Corporate and Aviation contribute a little towards the adr generation.

Market Segments and Cancellations



The majority of the bookings prefer no deposit from different market types. The factor behind the booking cancellations may be the deposit type. As many bookings are from no deposit type, guests tend to cancel more often. For instance: More than 35% of the bookings from online market segment is cancelled because of no deposit type

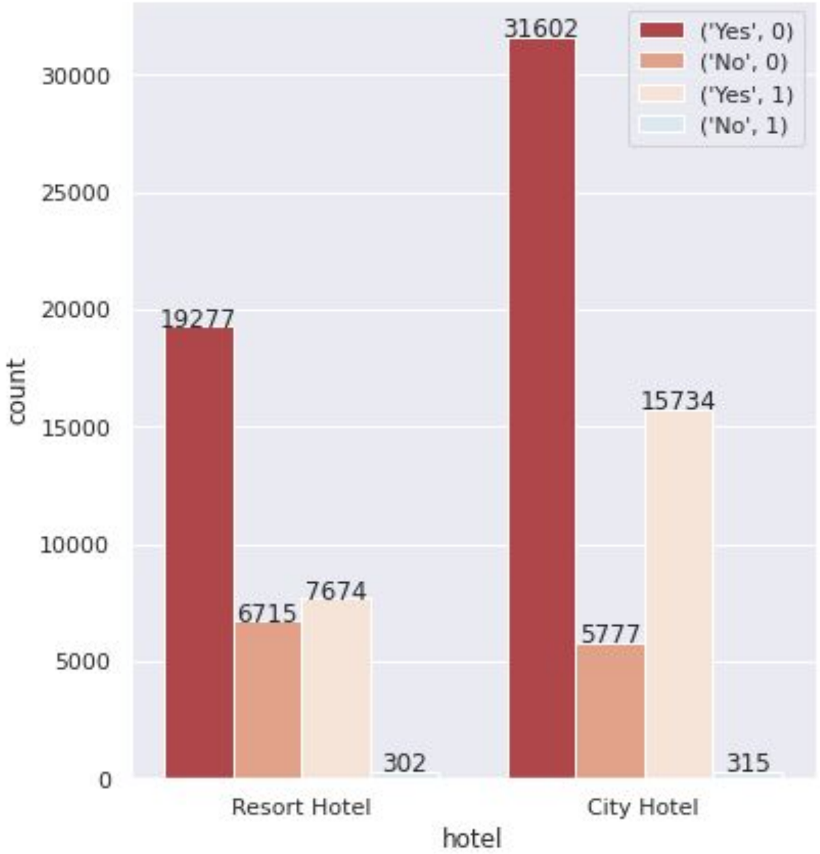
Rooms allotment

Resort Hotel

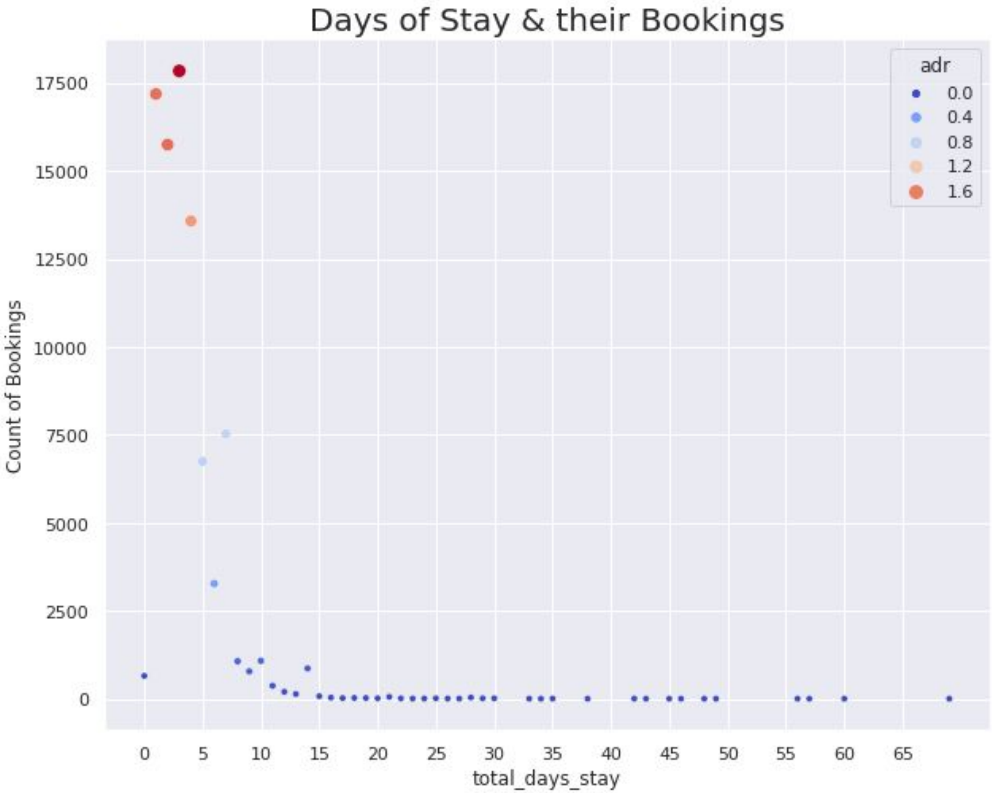
- Same room allotted and not cancelled - 19277
- Same room not alloted and not cancelled - 6715
- Same room allotted and cancelled - 7674
- Same room not alloted and cancelled - 302

City Hotel

- Same room allotted and not cancelled - 31602
- Same room not alloted and not cancelled - 5777
- Same room allotted and cancelled - 15734
- Same room not alloted and cancelled - 315



Adr and total stay



- Most of the bookings are for stay less than 5 days have high adr and number of bookings
- Number of bookings ranges from 12000 to 18000 bookings
- The next highest revenue generating stay is between 5 to 10 days.
- Longer stay has very few number of bookings and generates lesser revenue.

Agent, Booking Count and ADR

Agent 9

- Market Segment - Online TA
- Number of Bookings - 28751
- ADR - 3552171

Agent 240

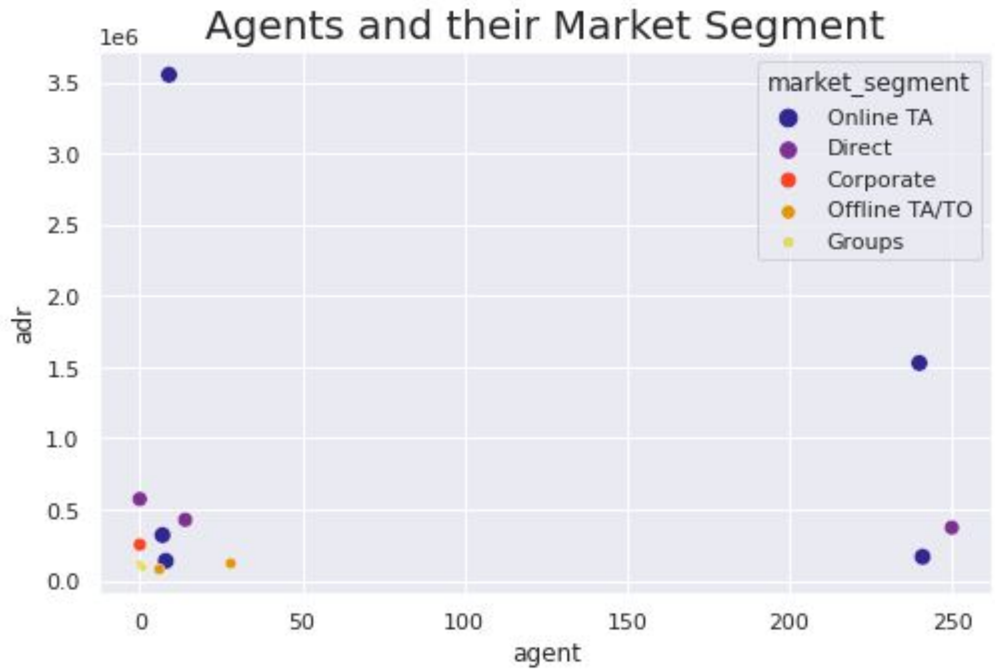
- Market Segment - Online TA
- Number of Bookings - 12997
- ADR - 1527619

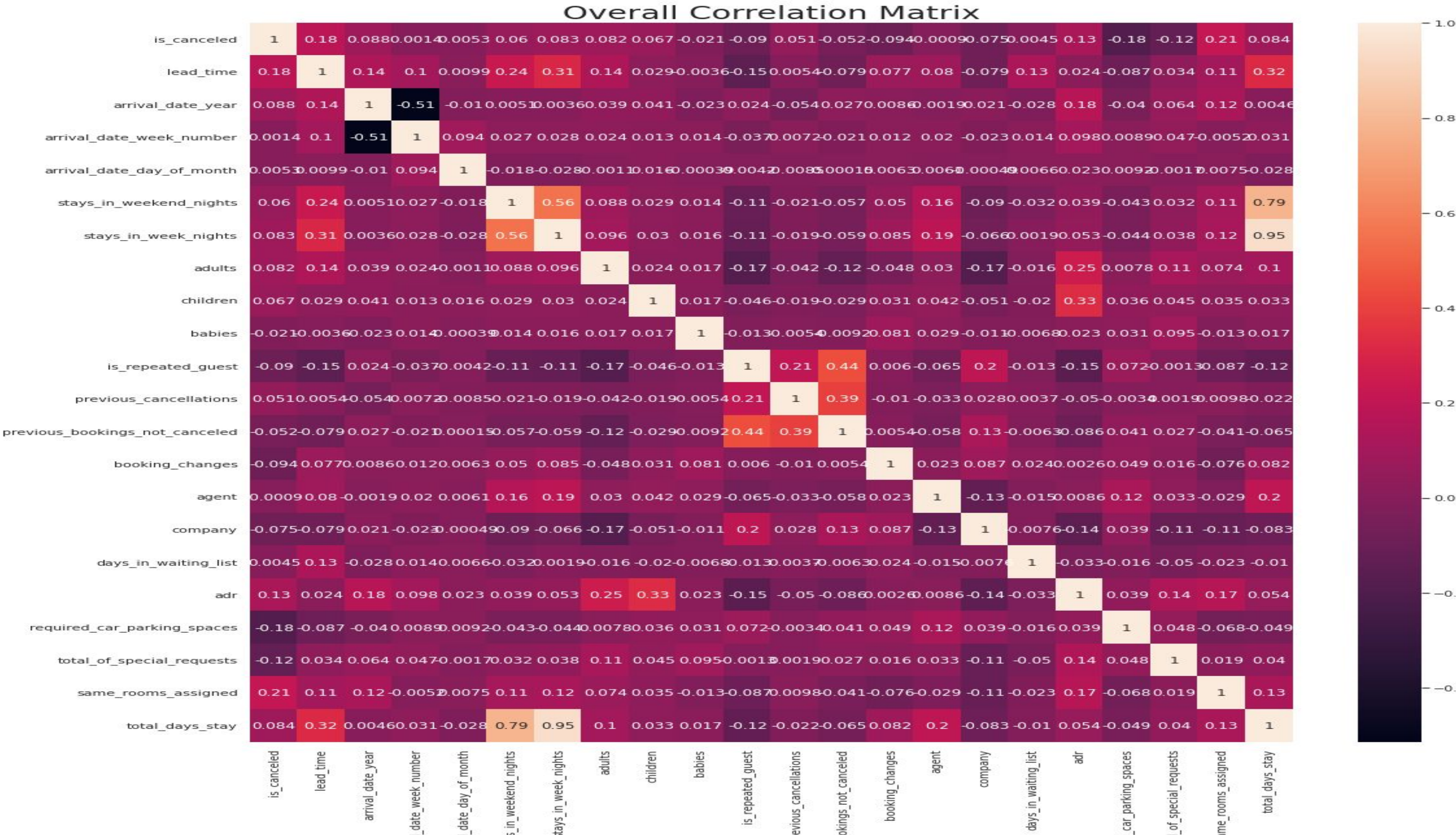
Agent 14

- Market Segment - Direct
- Number of Bookings - 3349
- ADR - 426354

Agent 250

- Market Segment - Direct
- Number of Bookings - 2776
- ADR - 371413





Overall Correlation Matrix

	is_canceled	lead_time	arrival_date_year	_date_week_number	_date_day_of_month	s_in_weekend_nights	tays_in_week_nights	adults	children	babies	is_repeated_guest	evious_cancellations	okings_not_canceled	booking_changes	agent	company	days_in_waiting_list	adr	car_parking_spaces	l_of_special_requests	me_rooms_assigned	total_days_stay
is_canceled	1	0.18	0.088	0.0014	0.0053	0.06	0.083	0.082	0.067	-0.021	-0.09	0.051	-0.052	-0.094	0.0009	-0.075	0.0045	0.13	-0.18	-0.12	0.21	0.084
lead_time	0.18	1	0.14	0.1	0.0099	0.24	0.31	0.14	0.029	-0.0036	-0.15	0.0054	-0.079	0.077	0.08	-0.079	0.13	0.024	-0.087	0.034	0.11	0.32
arrival_date_year	0.088	0.14	1	-0.51	-0.01	0.0051	0.0036	0.039	0.041	-0.023	0.024	-0.054	0.027	0.0086	-0.0019	0.021	-0.028	0.18	-0.04	0.064	0.12	0.0046
arrival_date_week_number	0.0014	0.1	-0.51	1	0.094	0.027	0.028	0.024	0.013	0.014	-0.037	0.0072	-0.021	0.012	0.02	-0.023	0.014	0.098	0.0089	0.047	-0.0052	0.031
arrival_date_day_of_month	0.0053	0.0099	-0.01	0.094	1	-0.018	-0.028	0.0011	0.016	0.00039	0.0042	0.0085	0.00015	0.0063	0.0061	0.00049	0.0066	0.023	0.0092	0.0017	0.0075	-0.028
stays_in_weekend_nights	0.06	0.24	0.0051	0.027	-0.018	1	0.56	0.088	0.029	0.014	-0.11	-0.021	-0.057	0.05	0.16	-0.09	-0.032	0.039	-0.043	0.032	0.11	0.79
stays_in_week_nights	0.083	0.31	0.0036	0.028	-0.028	0.56	1	0.096	0.03	0.016	-0.11	-0.019	-0.059	0.085	0.19	-0.066	0.0019	0.053	-0.044	0.038	0.12	0.95
adults	0.082	0.14	0.039	0.024	-0.0011	0.088	0.096	1	0.024	0.017	-0.17	-0.042	-0.12	-0.048	0.03	-0.17	-0.016	0.25	0.0078	0.11	0.074	0.1
children	0.067	0.029	0.041	0.013	0.016	0.029	0.03	0.024	1	0.017	-0.046	-0.019	-0.029	0.031	0.042	-0.051	-0.02	0.33	0.036	0.045	0.035	0.033
babies	-0.021	-0.0036	-0.023	0.014	0.00039	0.014	0.016	0.017	0.017	1	-0.013	-0.0054	-0.0092	0.081	0.029	-0.011	-0.0068	0.023	0.031	0.095	-0.013	0.017
is_repeated_guest	-0.09	-0.15	0.024	-0.037	0.0042	-0.11	-0.11	-0.17	-0.046	-0.013	1	0.21	0.44	0.006	-0.065	0.2	-0.013	-0.15	0.072	-0.0013	-0.087	-0.12
previous_cancellations	0.051	0.0054	-0.054	0.0072	0.0085	-0.021	-0.019	-0.042	-0.019	-0.0054	0.21	1	0.39	-0.01	-0.033	0.028	0.0037	-0.05	-0.0034	0.0019	0.0098	-0.022
previous_bookings_not_canceled	-0.052	-0.079	0.027	-0.021	0.00015	-0.057	-0.059	-0.12	-0.029	-0.0092	0.44	0.39	1	0.0054	-0.058	0.13	-0.0063	-0.086	0.041	0.027	-0.041	-0.065
booking_changes	-0.094	0.077	0.0086	0.012	0.0063	0.05	0.085	-0.048	0.031	0.081	0.006	-0.01	0.0054	1	0.023	0.087	0.024	0.0026	0.049	0.016	-0.076	0.082
agent	0.0009	0.08	-0.0019	0.02	0.0061	0.16	0.19	0.03	0.042	0.029	-0.065	-0.033	-0.058	0.023	1	-0.13	-0.015	0.0086	0.12	0.033	-0.029	0.2
company	-0.075	-0.079	0.021	-0.023	0.00049	-0.09	-0.066	-0.17	-0.051	-0.011	0.2	0.028	0.13	0.087	-0.13	1	-0.0076	-0.14	0.039	-0.11	-0.11	-0.083
days_in_waiting_list	0.0045	0.13	-0.028	0.014	0.0066	-0.032	0.0019	-0.016	-0.02	-0.0068	-0.013	0.0037	-0.0063	0.024	-0.015	-0.0076	1	-0.033	-0.016	-0.05	-0.023	-0.01
adr	0.13	0.024	0.18	0.098	0.023	0.039	0.053	0.25	0.33	0.023	-0.15	-0.05	-0.086	0.0026	0.0086	-0.14	-0.033	1	0.039	0.14	0.17	0.054
required_car_parking_spaces	-0.18	-0.087	-0.04	0.0089	0.0092	-0.043	-0.044	0.0078	0.036	0.031	0.072	-0.0034	0.041	0.049	0.12	0.039	-0.016	0.039	1	0.048	-0.068	-0.049
total_of_special_requests	-0.12	0.034	0.064	0.047	-0.0017	0.032	0.038	0.11	0.045	0.095	-0.0013	0.0019	0.027	0.016	0.033	-0.11	-0.05	0.14	0.048	1	0.019	0.04
same_rooms_assigned	0.21	0.11	0.12	-0.0052	0.0075	0.11	0.12	0.074	0.035	-0.013	-0.087	0.0098	-0.041	-0.076	-0.029	-0.11	-0.023	0.17	-0.068	0.019	1	0.13
total_days_stay	0.084	0.32	0.0046	0.031	-0.028	0.79	0.95	0.1	0.033	0.017	-0.12	-0.022	-0.065	0.082	0.2	-0.083	-0.01	0.054	-0.049	0.04	0.13	1

Solution to Business Objective

- The number of bookings and adr is high in August. We can promote bookings in other months with exciting offers while the bookings are low.
- The number of booking cancellations is due to the no-deposit option in the deposit type. Implement deposit rate for bookings will reduce cancellations.
- The contribution of marketing channels is low other than the online channel. Develop a marketing strategy to promote bookings.
- Most of the bookings are through online channels, so effective advertisement and online presence drive more bookings.

Conclusion

- The number of bookings is high for city hotels compared to Resort hotels.
- The percentage of bookings canceled is 27.5%, and 72.5% remains unchanged.
- Transient type of customers makes up the majority of the bookings.
- TA/TO, Direct, and Corporate is the distribution channel for the hotel.
- Maximum bookings are with no deposit type, which is one of the factors for cancellation.
- The most preferred room in both the hotel is a type A room.
- The most preferred meal is BB meal.
- The maximum number of special requests is 5. The majority of the bookings don't have any special requests.

Conclusion

- European countries like Portugal, the United Kingdom, France, and Spain tend to make more bookings.
- The cancellations show consistent growth year on year alongside the increase in bookings for the hotel.
- The highest revenue generated is in August compared to other months.
- The market segment with a higher booking is Online TA, but also with a higher number of cancellations compared to other market segments.
- Staying less than ten days has more bookings and generates higher revenue.
- Days on the waiting list is not having a relationship with cancellation.

Q&A…