

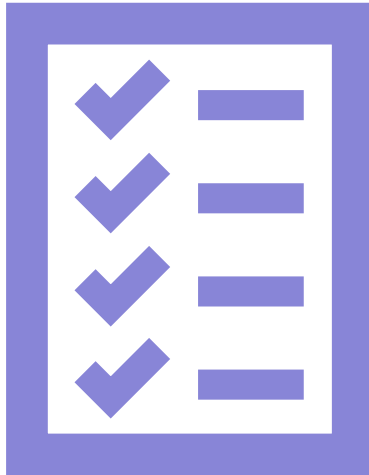
PREDICTING HOTEL BOOKING CANCELATION

By: Pushpinder Raghu

Email: Pushpinder.Raghu@gmail.com

GitHub ID: DevPushpinder





POINTS TO DISCUSS:

- Agenda
- Data summary
- Univariate analysis
- Hotel wise analysis
- Distribution Channel wise analysis
- Booking cancellation analysis
- Timewise analysis
- Correlation heatmap
- EDA Conclusion
- Model Building
- Model Pipeline
- Process to use Existing Model

AGENDA

To discuss the analysis of given hotel bookings data set from 2015-2017.

We'll be doing analysis of given data set in following ways :

- Univariate analysis
 - Hotel wise analysis
 - Distribution Channel wise analysis
 - Booking cancellation analysis
 - Timewise analysis
-

By doing this we'll try to find out key factors driving the hotel bookings trends.

DATA SUMMARY

The dataset consists of 119390 rows and 32 columns.

- **hotel:** categorical variable, information on category of hotel
- **is_cancelled:** target variable to be predicted.
- **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 (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
- **stays_in_week_nights:** Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel

DATA SUMMARY

(CONTD..)

- **adults:** Number of adults
- **children:** Number of children
- **babies:** Number of babies
- **meal:** Type of meal booked.
- **country:** Country of origin
- **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:** Value indicating if the booking name was from a repeated guest (1) or not (0)
- **previous_cancellations:** Number of previous bookings that were cancelled by the customer prior to the current booking.
- **previous_bookings_not_canceled:** Number of previous bookings not cancelled by the customer prior to the current booking.
- **reserved_room_type:** Code of room type reserved. Code is presented instead of designation for anonymity reasons.
- **assigned_room_type:** Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due

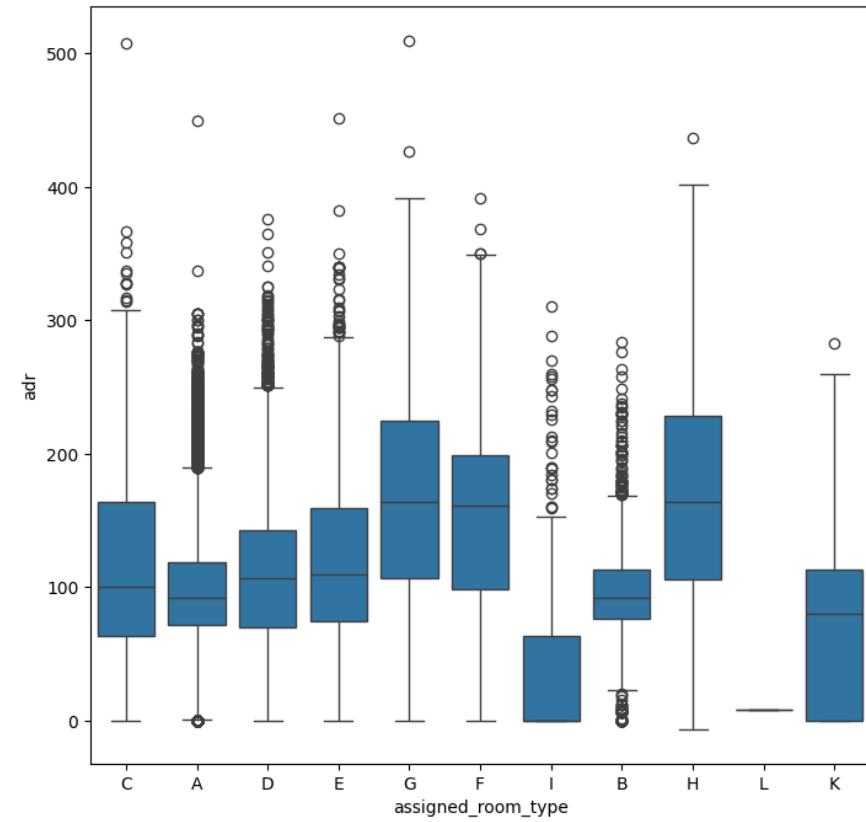
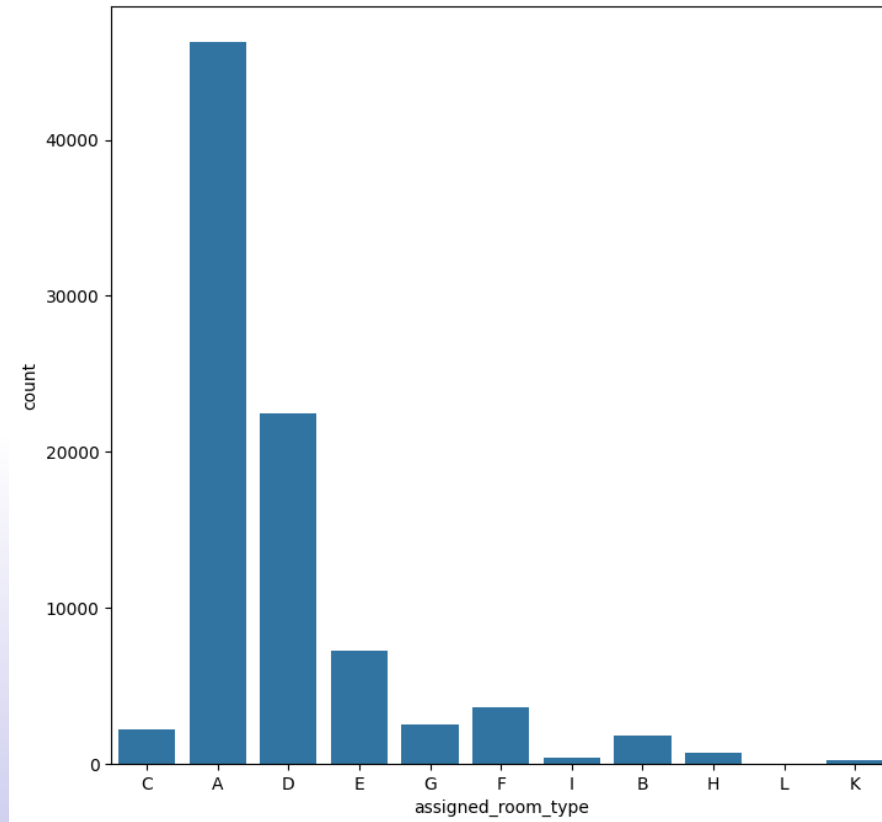
DATA SUMMARY

(CONTD..)

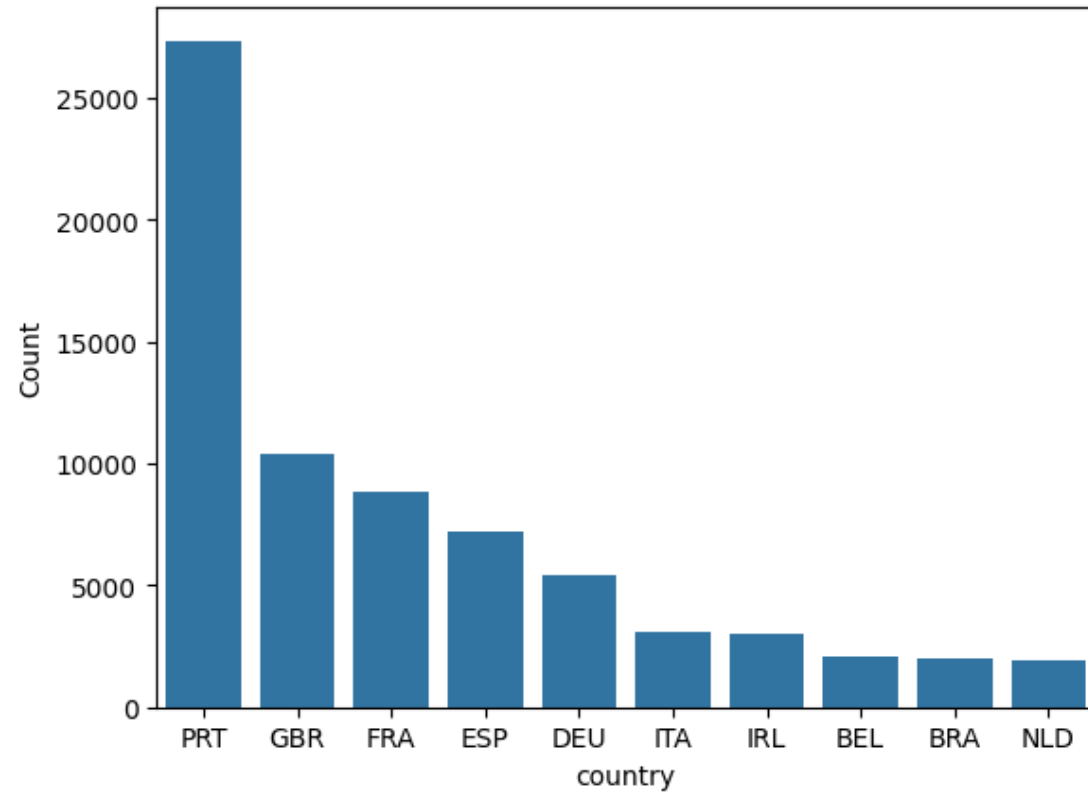
- **booking_changes:** Number of changes/amendments made to the booking from the moment the booking was entered on the PMS.
- **deposit_type:** Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No
- **agent:** ID of the travel agency that made the booking.
- **company:** ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for
- **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 booking, assuming one of four categories: Contract - when the booking has an allotment or other type of
- **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, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out
- **reservation_status_date:** Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to

UNIVARIATE ANALYSIS

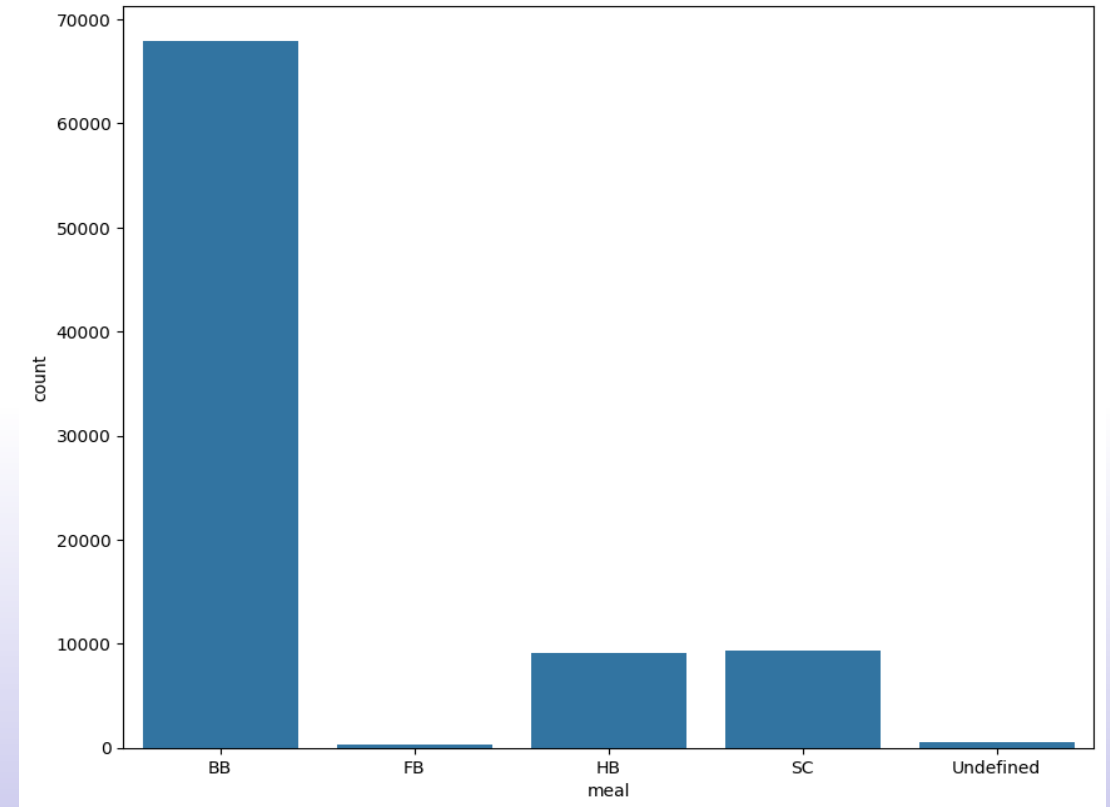
Most demanded room type is A, but better adr rooms are of type H, G and C also. Hotels should increase the no. of room types A and H to maximise revenue.



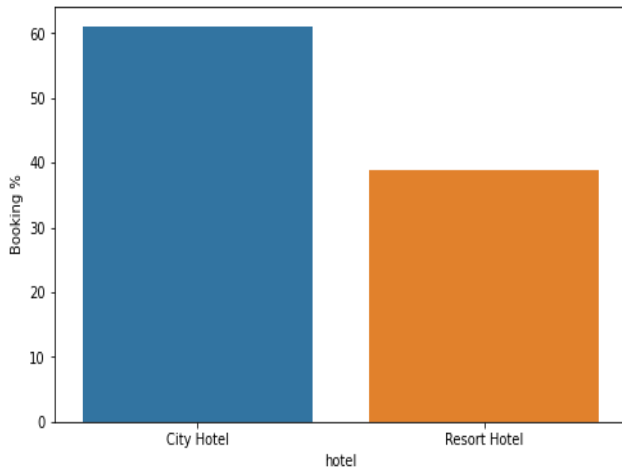
Most guest are from Portugal and other European countries.



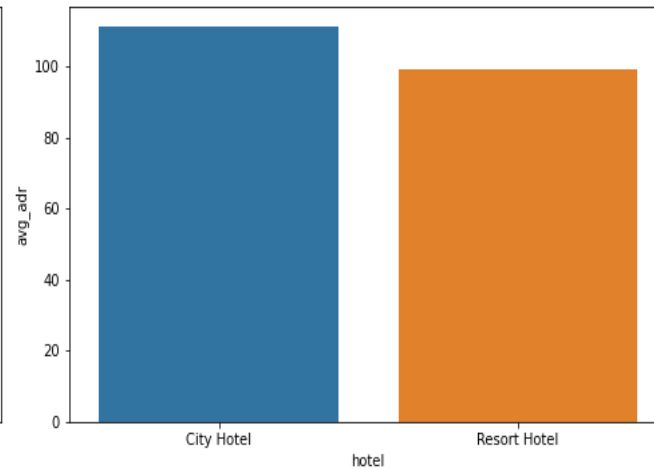
Most preferred meal type is BB (Bed and breakfast).



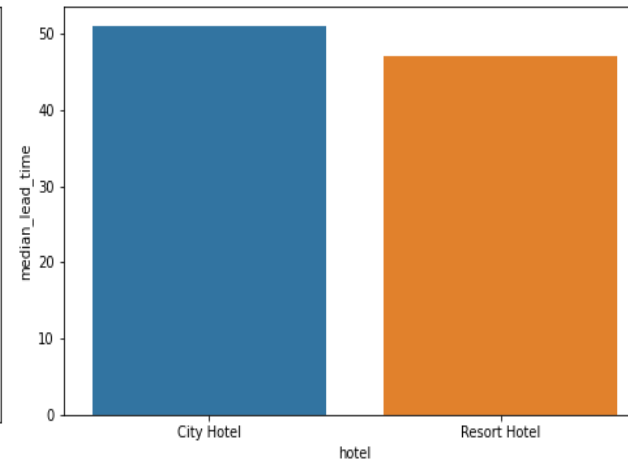
HOTEL WISE ANALYSIS



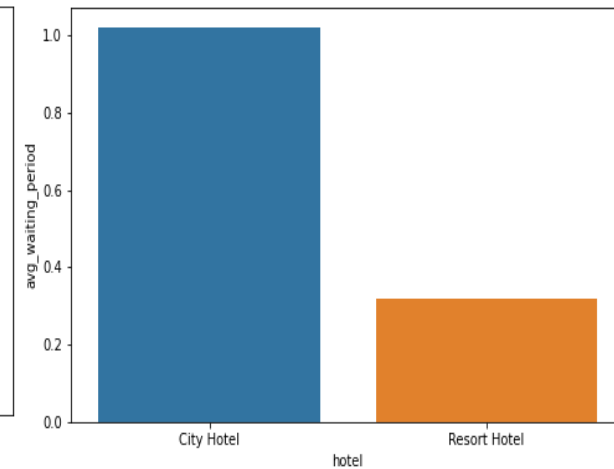
Around 60% bookings are for City hotel and 40% bookings are for Resort hotel.



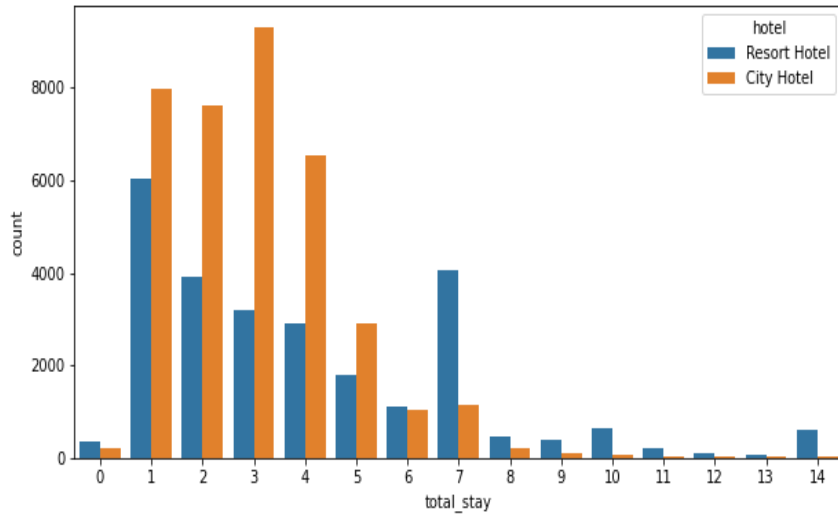
Avg adr of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue.



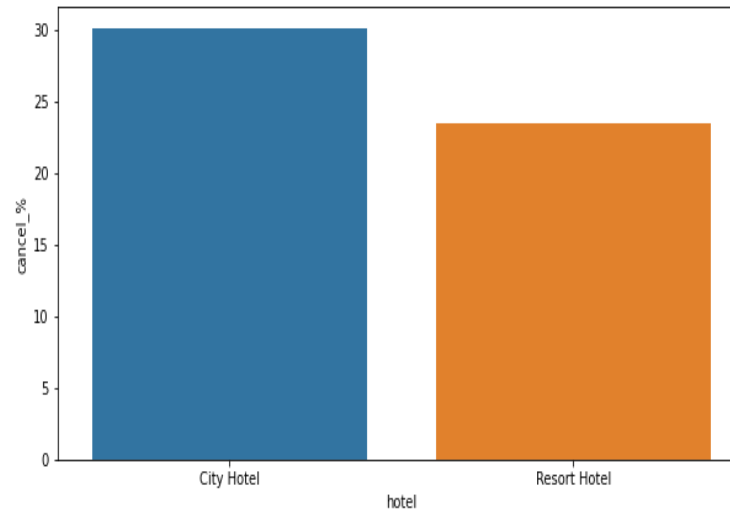
City hotel has slightly higher median lead time. Also median lead time is significantly higher in each case, this means customers generally plan their hotel visits way to early.



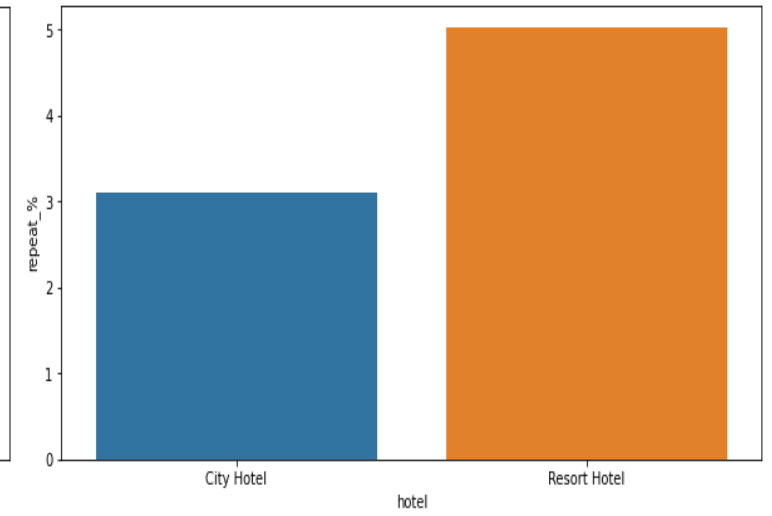
City hotel has significantly longer waiting time, hence City Hotel is much busier than Resort Hotel.



Most common stay length is less than 4 days and generally people prefer City hotel for short stay, but for long stays, Resort Hotel is preferred.

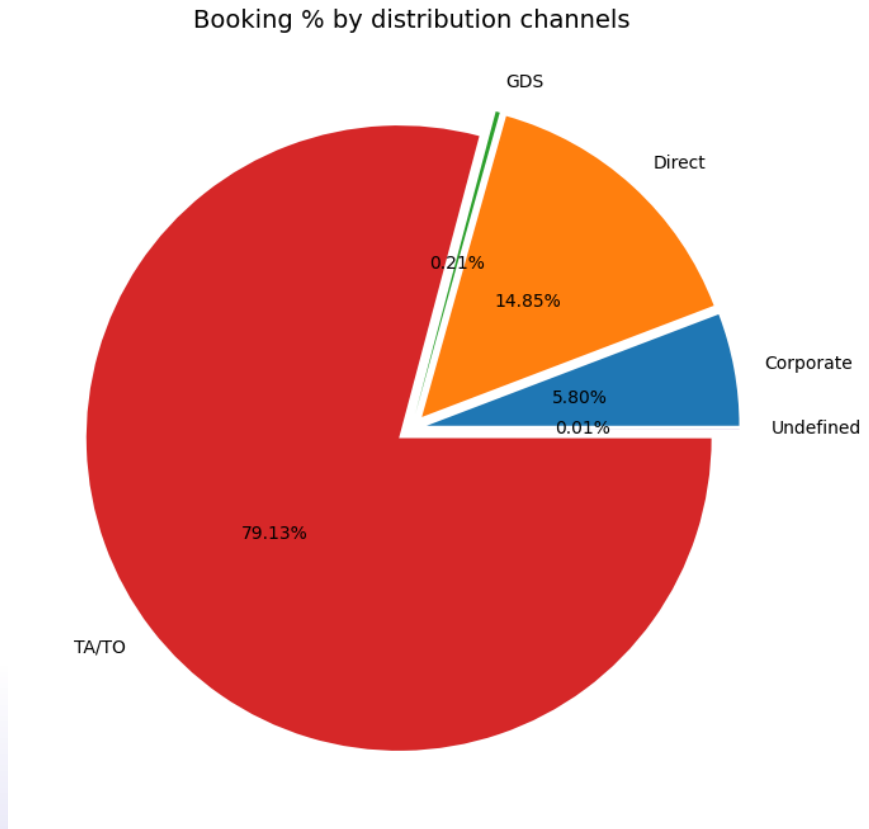


Almost 30 % of City Hotel bookings got canceled.



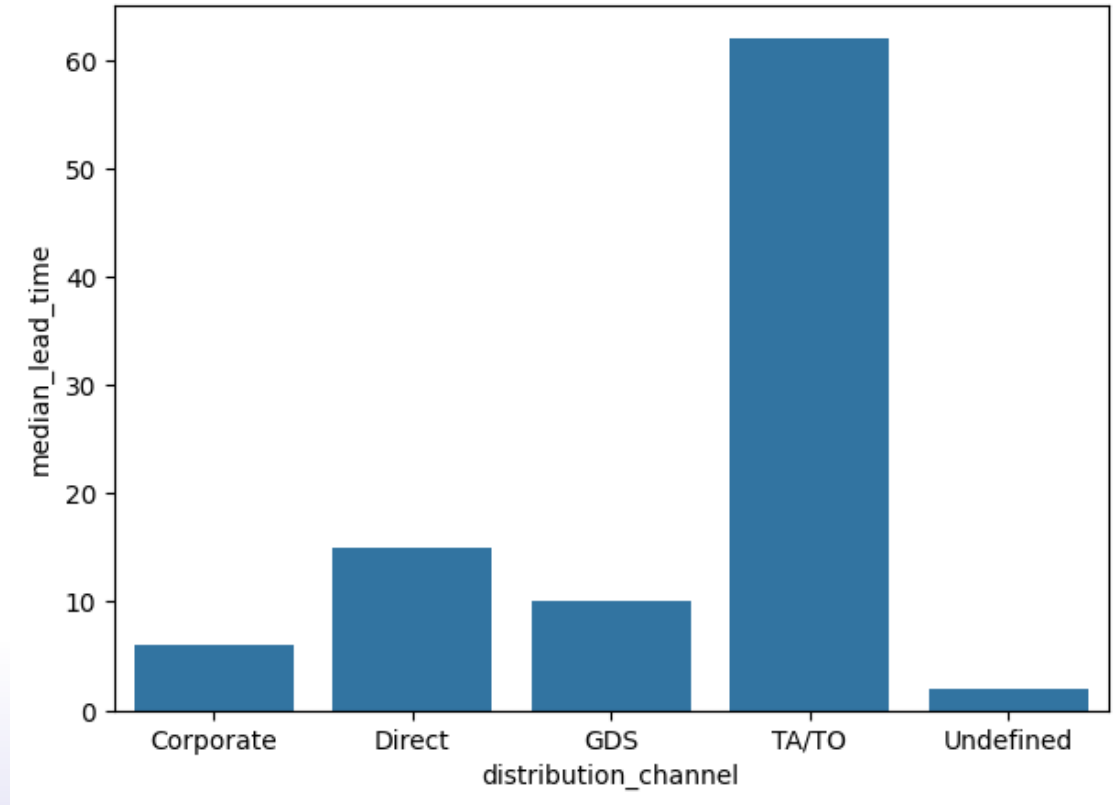
Both hotels have very small percentage that customer will repeat, but Resort hotel has slightly higher repeat % than City Hotel.

DISTRIBUTION CHANNEL WISE ANALYSIS

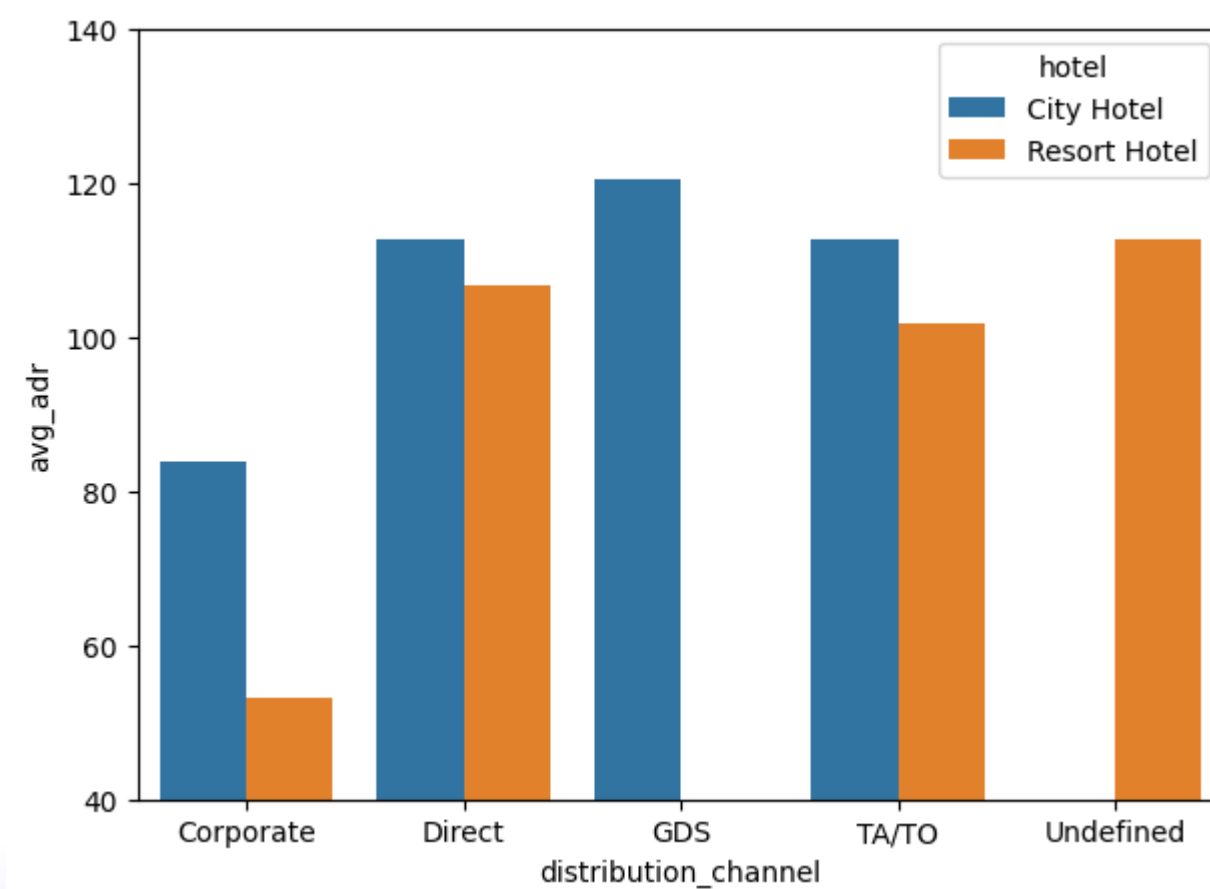


Here we can see that the most of guest are making reservation through TA/TO channels which is travel agency and tour operator.

Second most used channel is direct.



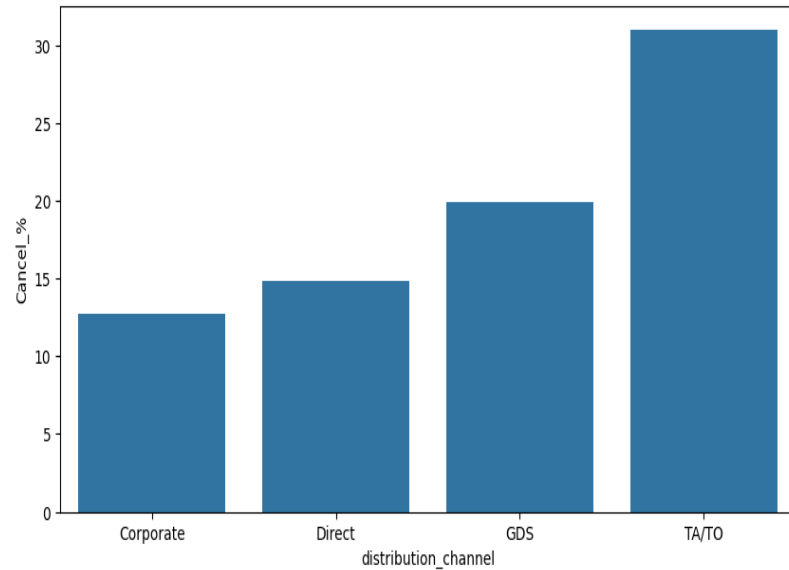
Channel which is mostly used for early booking of hotels is also TA/TO.



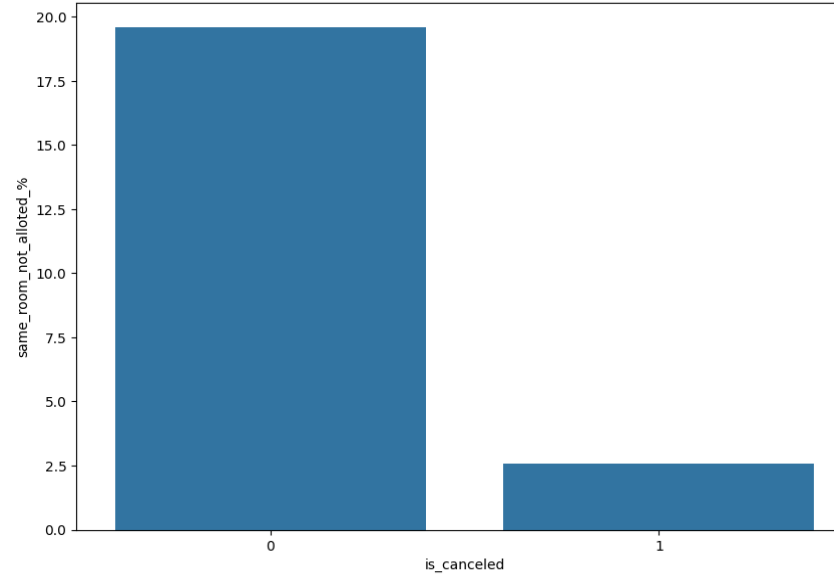
GDS channel brings higher revenue generating deals for City hotel, in contrast to that most bookings come via TA/TO and Direct. City Hotel can work to increase outreach on GDS channels to get more higher revenue generating deals.

Resort hotel has more revenue generating deals by direct and TA/TO channel. Resort Hotel need to increase outreach on GDS channel to increase revenue.

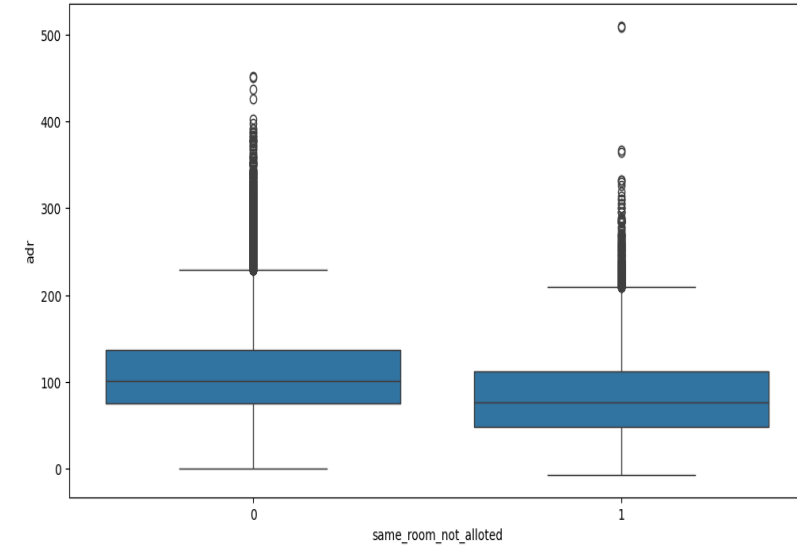
BOOKING CANCELLATION ANALYSIS



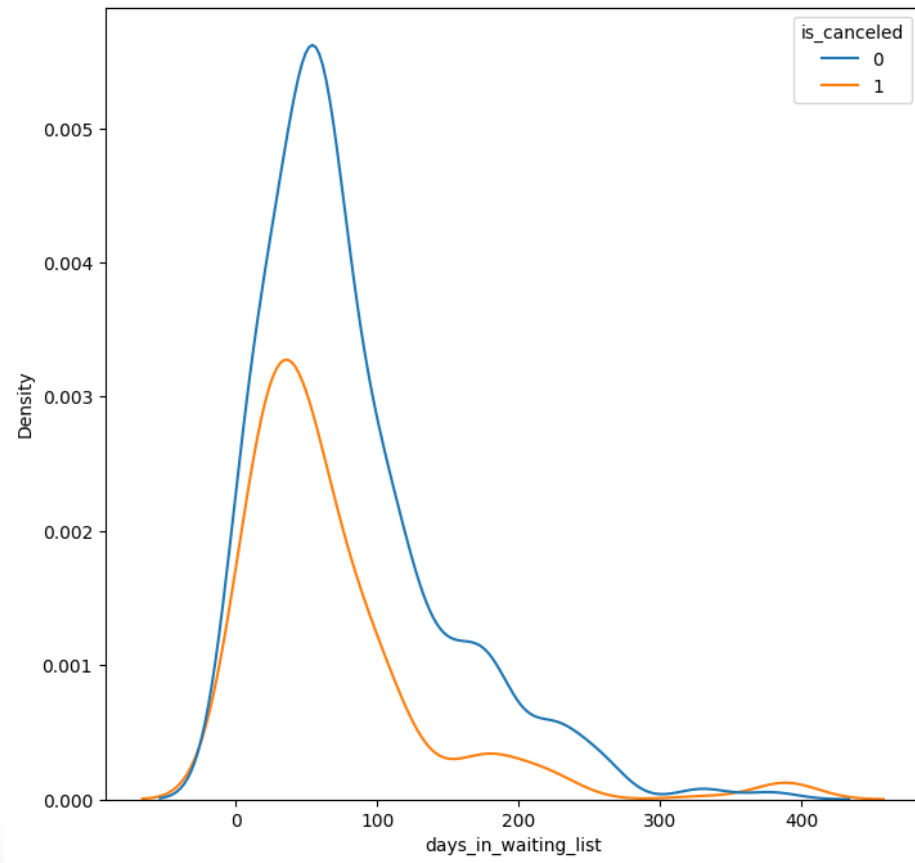
TA/TO has highest booking cancellation %. Therefore, a booking via TA/TO is 30% likely to get cancelled.



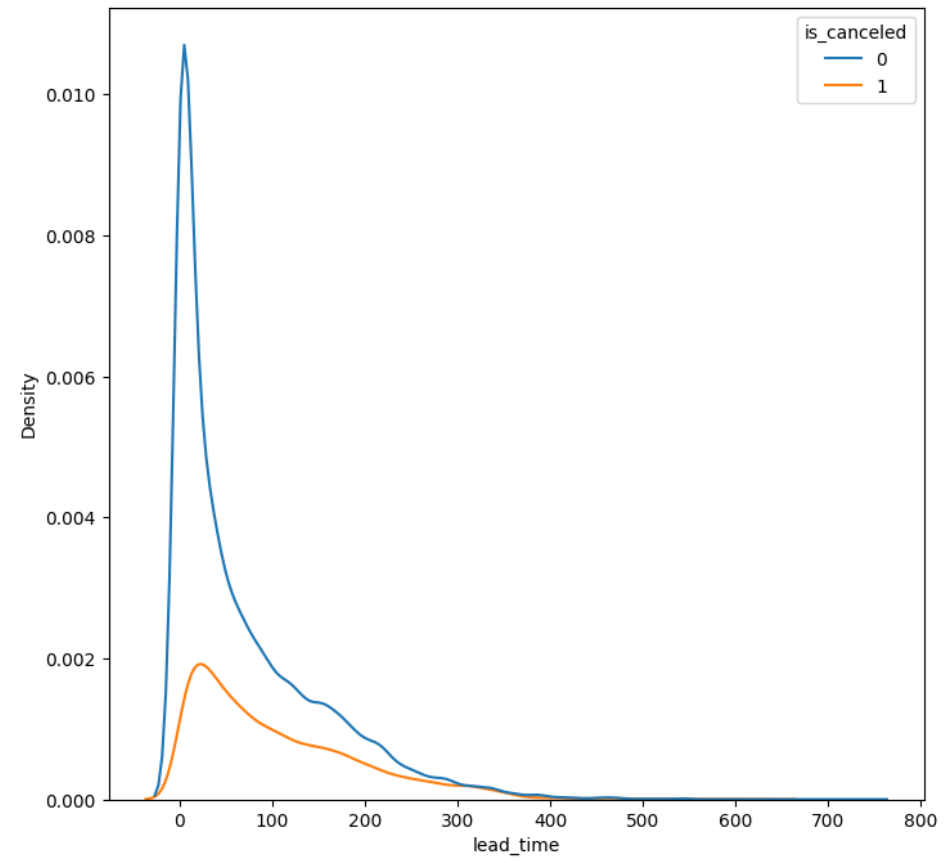
Not getting same room as demanded is not the case of cancellation of rooms. A significant percentage of bookings are not cancelled even after getting different room as demanded.



customers who didn't get same room have paid a little lower adr, except for few exceptions.

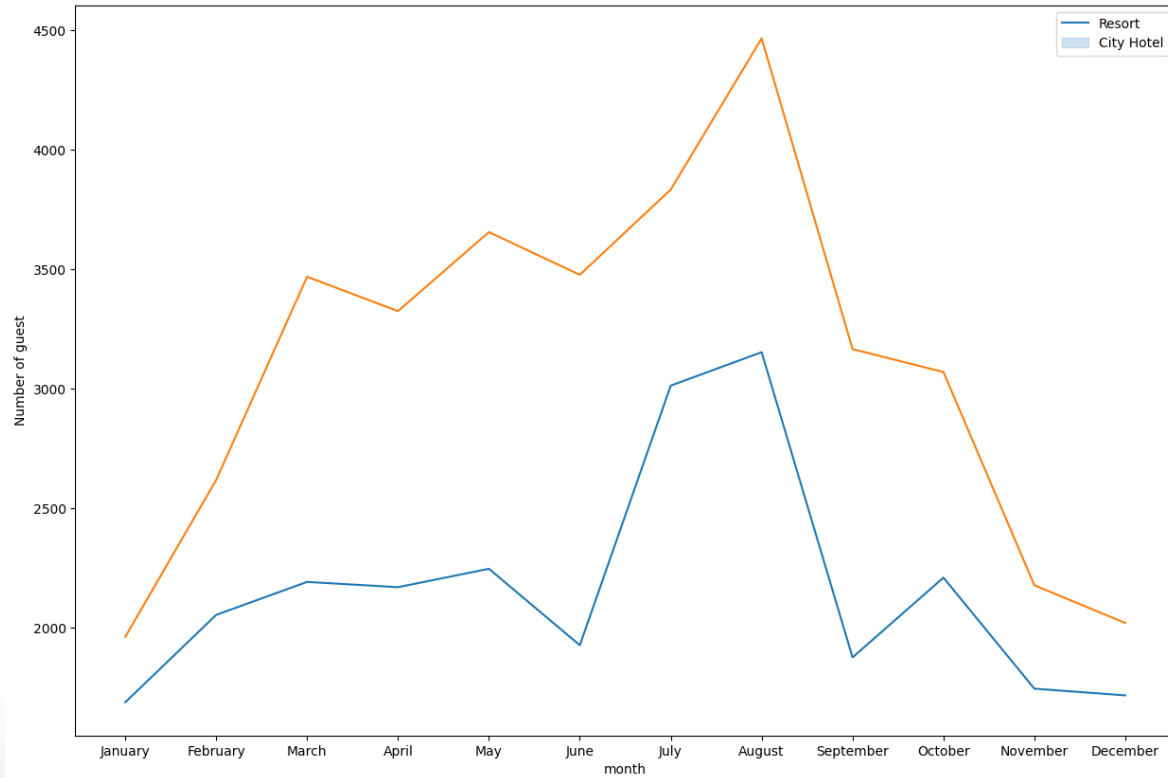


Most of the bookings that are cancelled have waiting period of less 150 days but also most of bookings that are not cancelled also have waiting period of less than 150 days. Hence this shows that waiting period has no effect on cancellation of bookings.

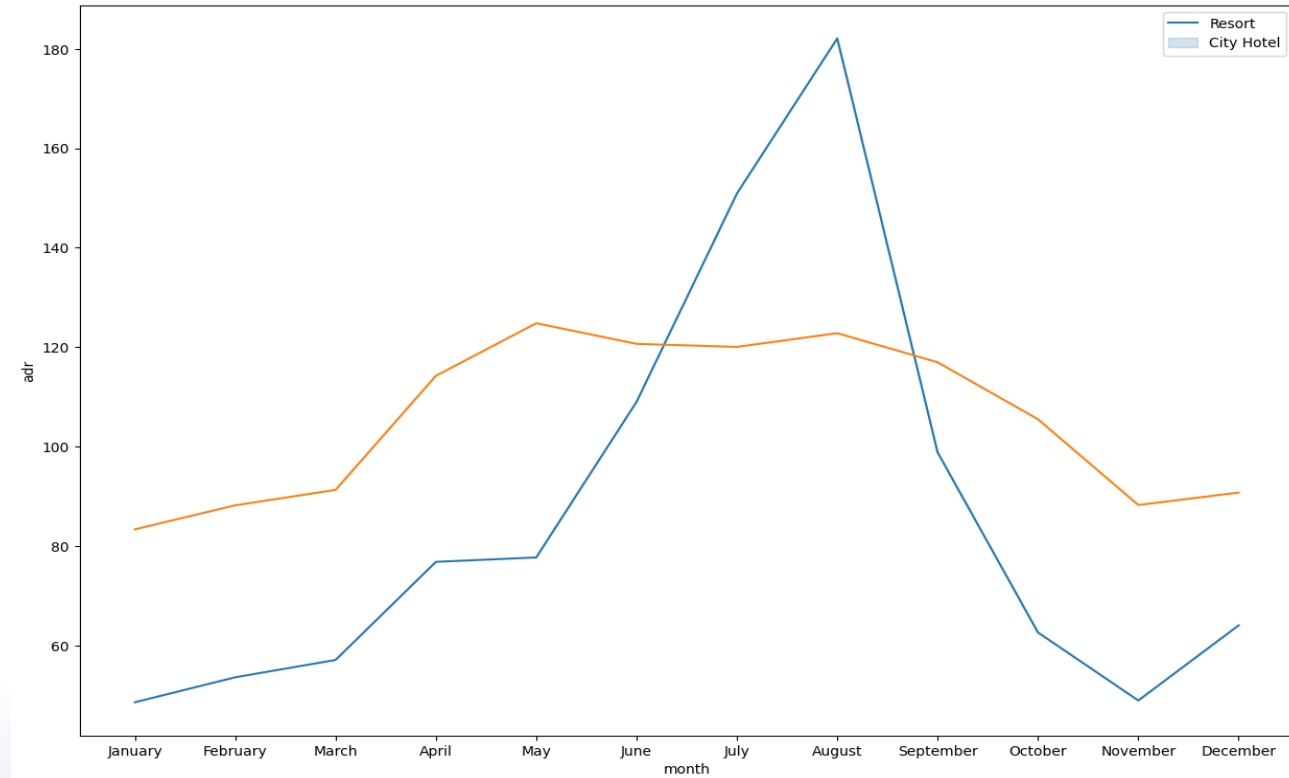


Lead time has no effect on cancellation of bookings, as both curves of cancellation and not cancellation are similar for lead time too.

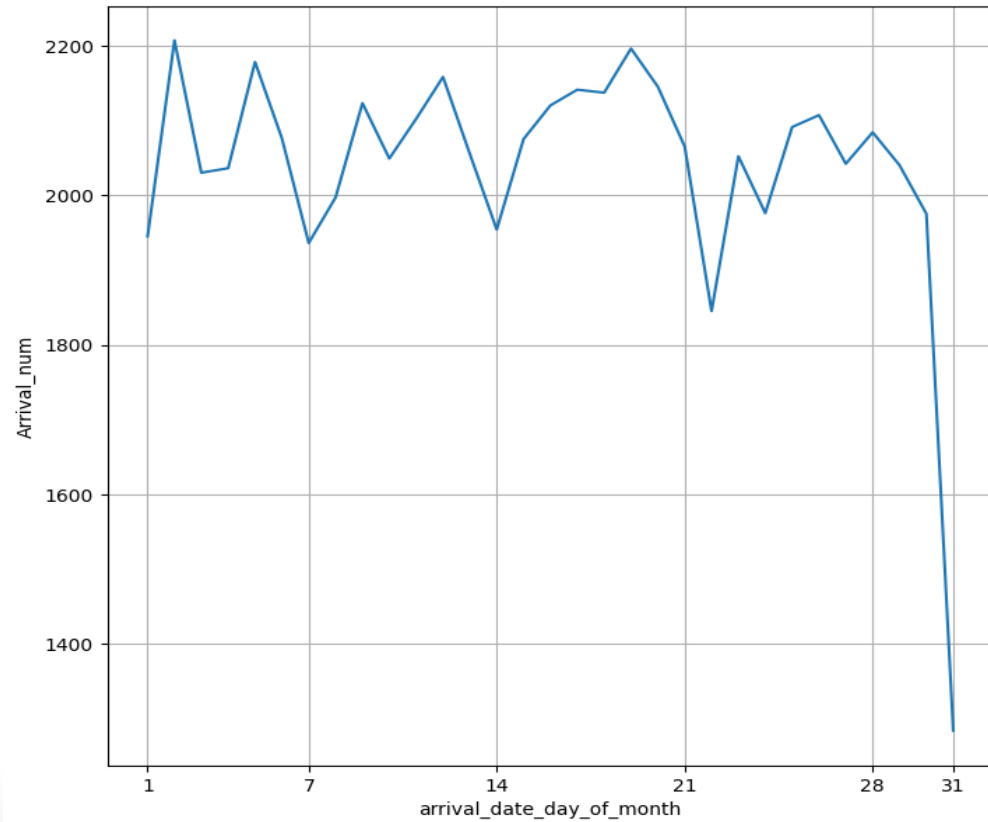
TIME-WISE ANALYSIS



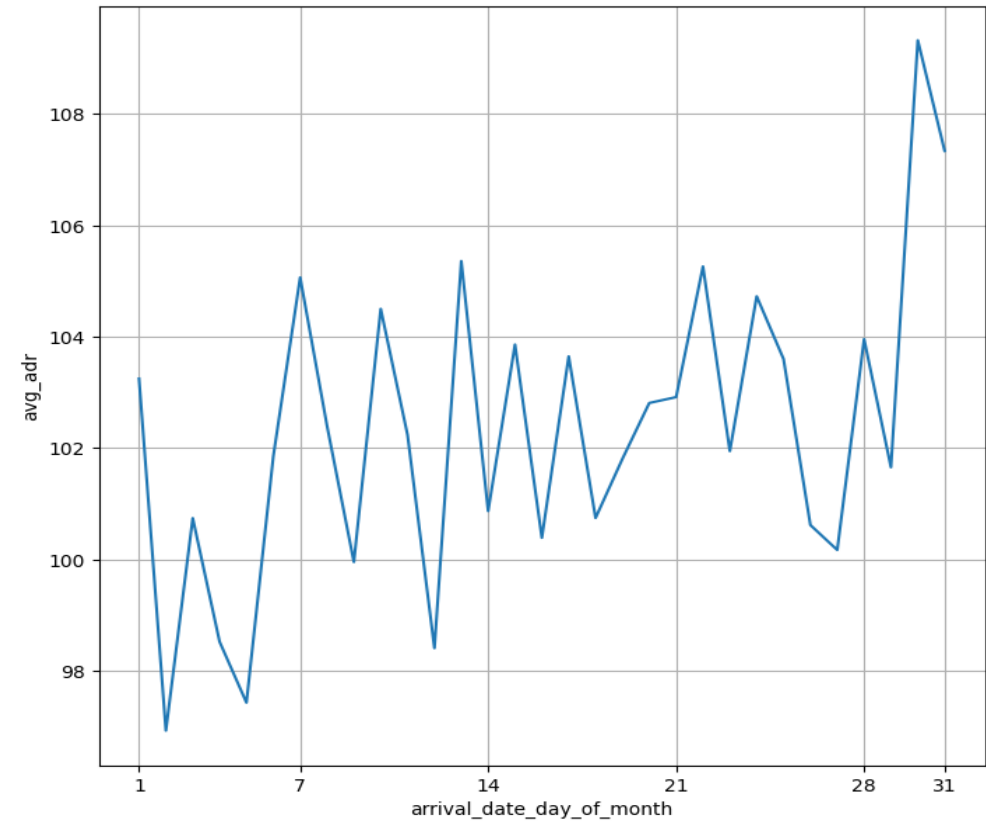
From the month of July to August the number of bookings increased and in August, City Hotel got most number of guests.



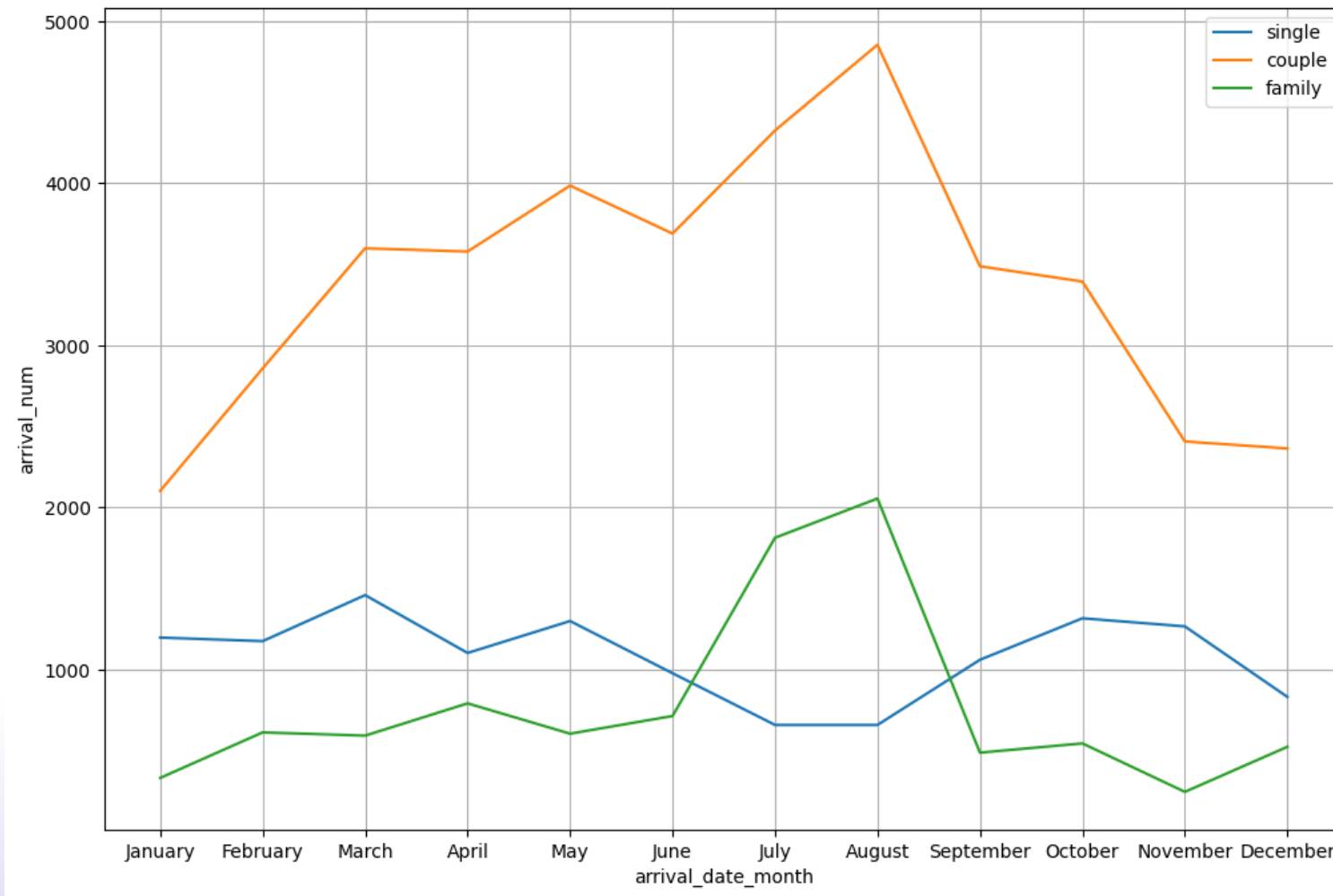
The revenue aspect looks different, the Resort Hotels receives more revenue with respect to City Hotel. From May to August there was rapid increase in adr. August recorded the highest.



We can see that graph Arrival_num has small peaks at regular interval of days. This can be due to increase in arrival weekend.



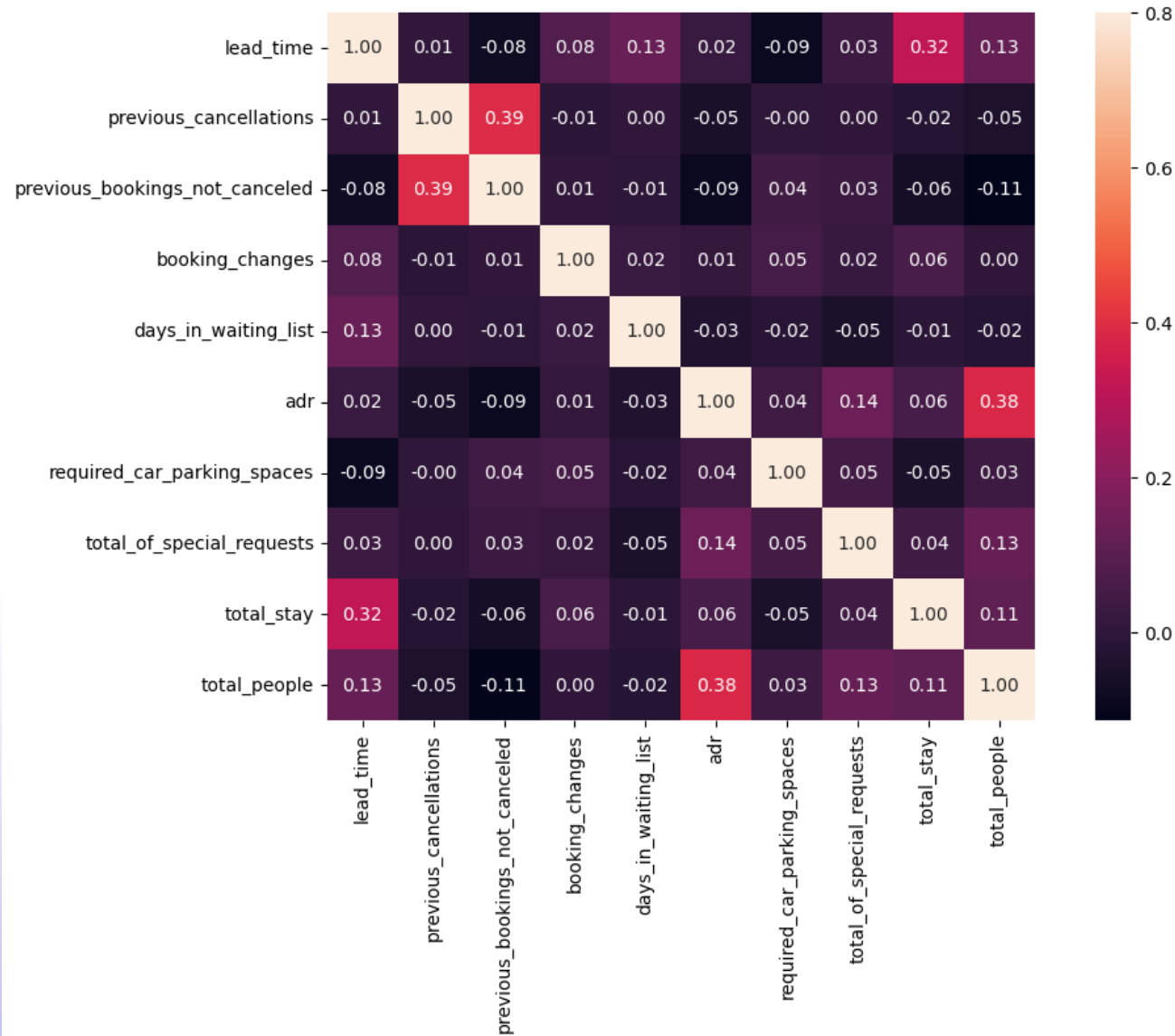
The avg adr tends to go up as month ends. Therefore charges are more at the end of month.



Mostly bookings are done by couples.

It is clear from graph that there is a sudden surge in arrival num of couples and family in months of July and August. So better plans can be planned accordingly at that time for these type of customers.

CORRELATION HEATMAP



- Total stay length and lead time are slightly correlated. This may infer that for longer hotel stays, people generally plan little before the actual arrival.
- adr is slightly correlated with total_people, which makes sense as more no. of people means more service to deliver, therefore more adr.

EDA CONCLUSION

- Around 60% bookings are for City hotel and 40% bookings are for Resort hotel, therefore City Hotel is busier than Resort hotel. Also, the overall adr of City hotel is slightly higher than Resort hotel.
- Mostly guests stay for less than 5 days in hotel and for longer stays Resort hotel is preferred.
- Both hotels have significantly higher booking cancellation rates and very few guests less than 3 % return for another booking in City hotel. 5% guests return for stay in Resort hotel.
- Most of the guests came from European countries, with most no. of guest coming from Portugal.
- Guests use different channels for making bookings out of which most preferred way is TA/TO.
- For hotels higher adr deals come via GDS channel, so hotels should increase their popularity on this channel.
- Almost 30% of bookings via TA/TO are cancelled.
- Not getting same room as reserved, longer lead time and waiting time do not affect cancellation of bookings. Although different room allotment do lowers the adr.
- July-August are the most busier and profitable months for both of hotels.
- Within a month, adr gradually increases as month ends, with small sudden rise on weekends.
- Couples are the most common guests for hotels, hence hotels can plan services according to couples needs to increase revenue.

MODEL BUILDING

DATA CLEANING

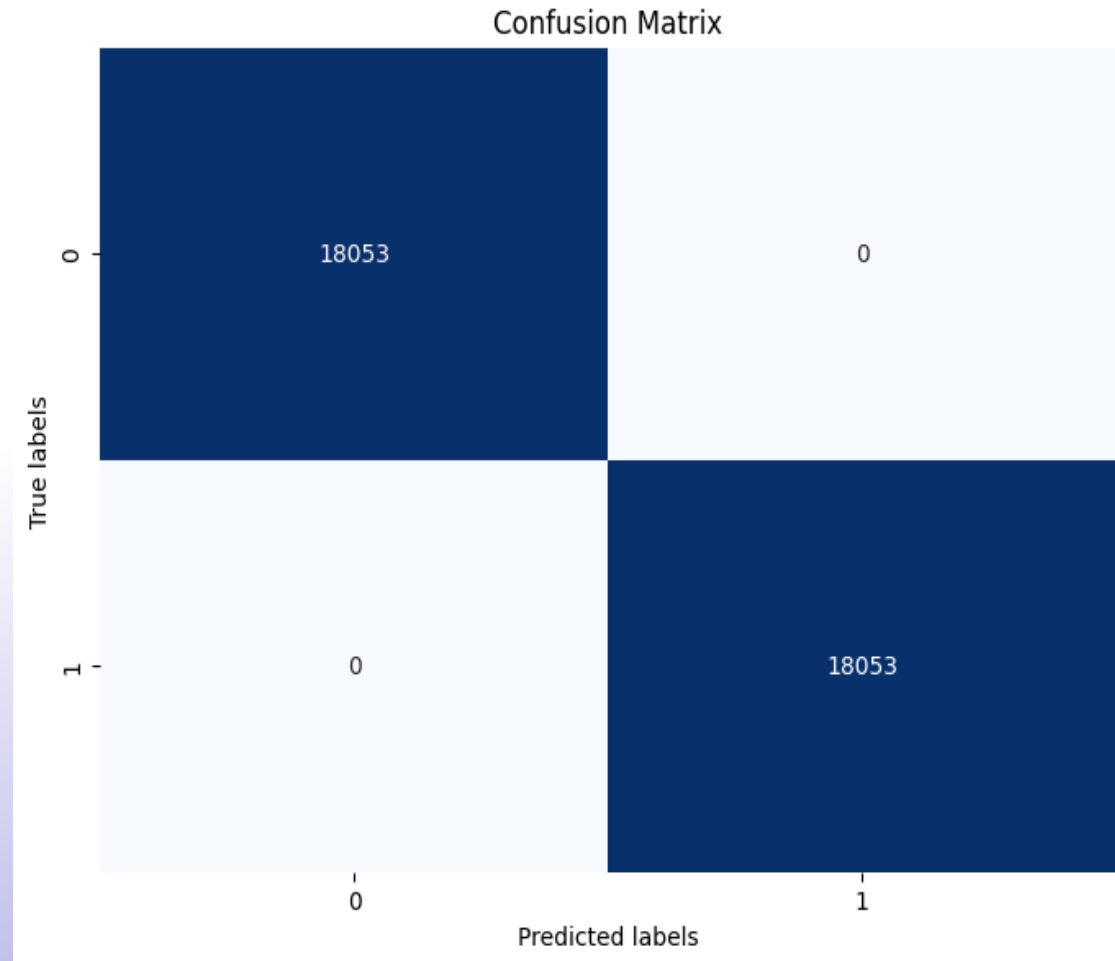
- 31994 duplicates samples in the database are removed, which is 26.8%
- There are 651(0.74%) records where stays_in_weekend_nights and stays_in_week_nights marked as zero which are removed,
- There are 166(0.18%) records where adults, children and babies marked as zero, which are removed.
- Dataset has imbalanced class of cancellation with 24025(24.49%) records in which
 - Cancelations in resort hotel= 23.48 %
 - Cancelations in city hotel= 30.04 %
 - SMOTE is implemented as part of treatment
- Features dropped are arrival_date_year, arrival_date_week_number, reservation_status_date Agent, company.
- Missing values in children, country feature are imputed using isolation forest algorithm.
- Hyperparameters used for SVC model after fine tuning are:
 - ❑ 'C': 0.1,
 - ❑ 'gamma': 1,
 - ❑ 'kernel': 'linear'
- Model performance
 - ❑ Accuracy: 1.0
 - ❑ Precision: 1.0
 - ❑ Recall: 1.0

MODEL BUILDING..CONTD

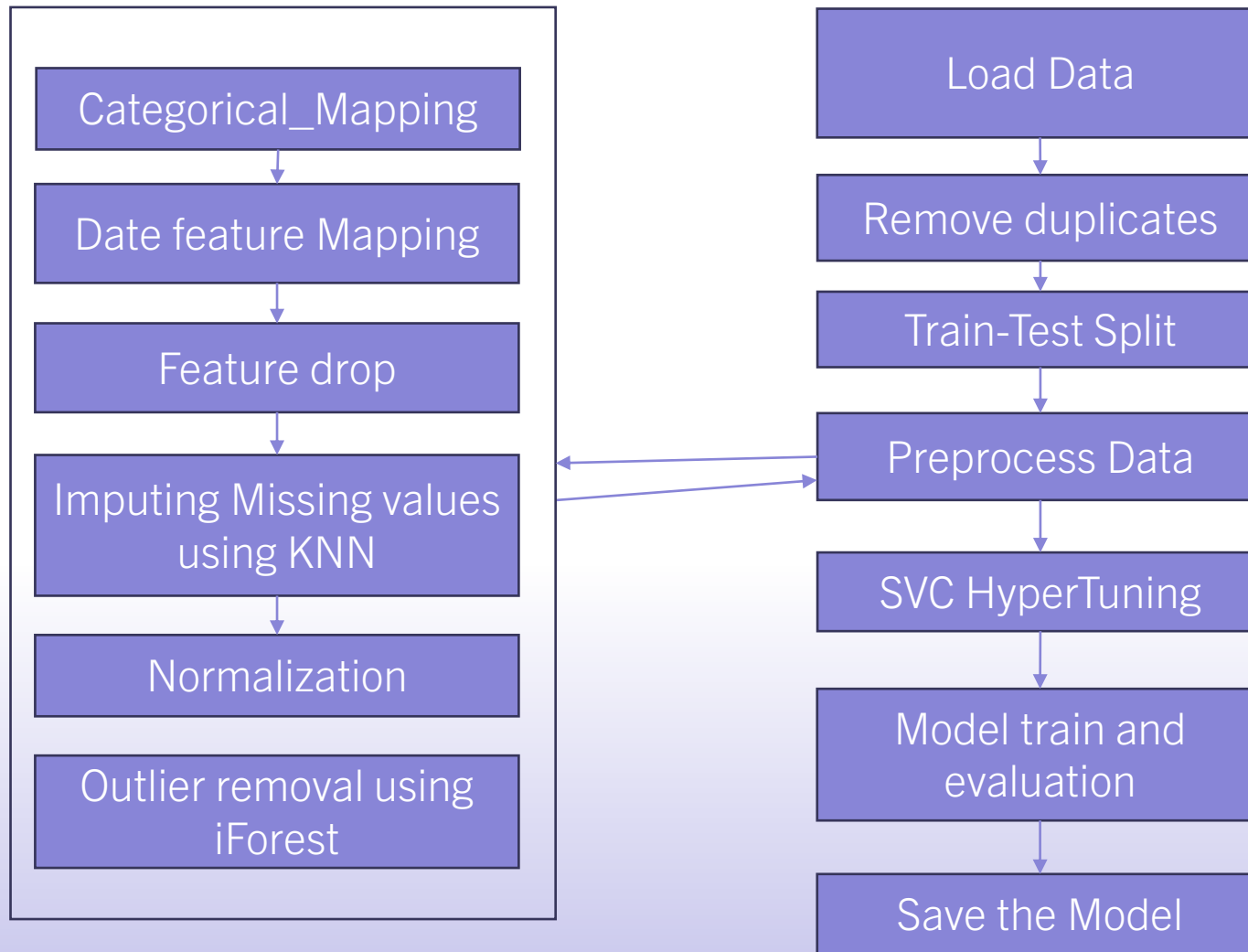
❑ F1 Score: 1.0

❑ Training time: 0.6826984882354736

❑ Inference time: 0.14650464057922363








MODEL PIPELINE



PROCESS TO USE EXISTING MODEL

List of files included in package:

Name	Date modified	Type	Size
 requirements.txt	3/1/2024 4:45 PM	Text Document	1 KB
 finalized_model.sav	3/10/2024 5:20 PM	SAV File	58 KB
 hotel_bookings_models_inputs.csv	3/1/2024 4:43 PM	Microsoft Excel Co...	5 KB
 Load_model.ipynb	3/1/2024 4:46 PM	IPYNB File	4 KB
 Pre_processing.py	3/1/2024 3:45 PM	Python File	3 KB

Process to use the model:

Step1: Create a new environment and Install the required dependencies using requirements.txt.

Step2: Open Load_model.ipynb jupyter notebook

Step3: Inputs can be provided either using csv file or using dictionary, if using csv file set 'csv_selected' variable value to True.

Step4: Run the Jupiter notebook to predict.