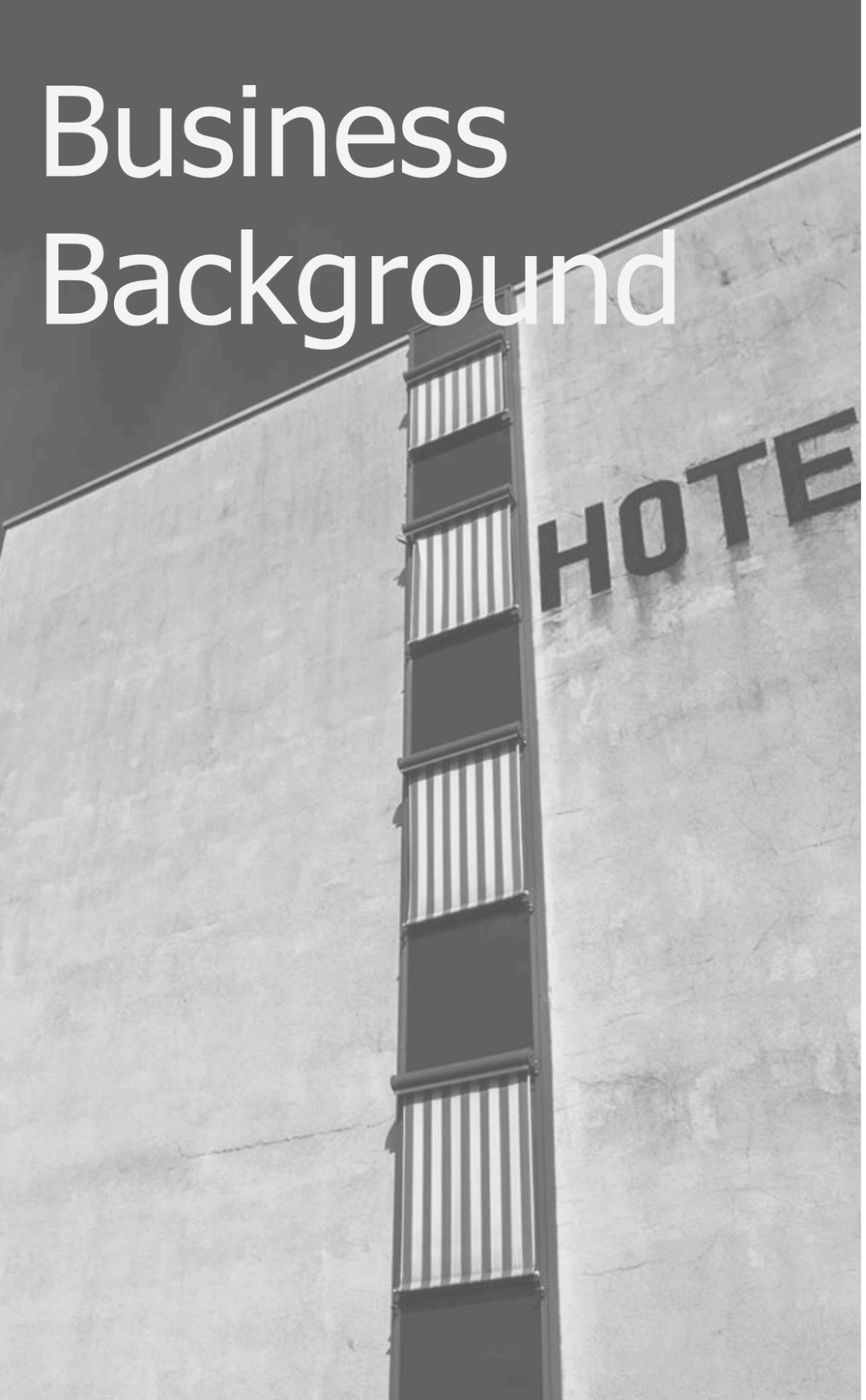
Hotel Booking Cancellation

Harshitha S

Outline

- Business Background
- Data Understanding & Data Preprocessing
- Exploratory Data Analysis
- Modeling Evaluation
- Business Recommendation

Business Background



Problem Statement

The hotel experienced a loss of revenue due to unexpected hotel cancellations made by customers. The company's goal is to handle this uncertainty by predicting which customers are likely to cancel their hotel bookings. The ultimate goal of doing this is to help the company identify customers with the highest probability of cancellation, so that the company can immediately anticipate this by providing proactive strategies such as providing incentives, special reminders and implementing other strategies to mitigate the occurrence of order cancellations, and ultimately increase company revenue

Objective

The main objective is to create a model that can accurately predict the possibility of a hotel booking being canceled or not. And with this prediction can provide valuable insights for companies that can help them allocate resources effectively and provide business recommendations to reduce the number of canceled bookings and increasing company revenue.

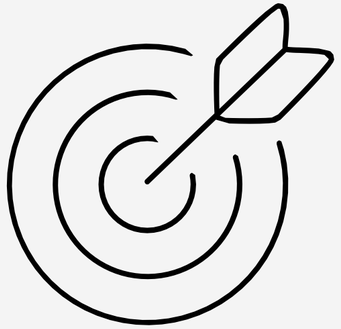
Data Understanding

83293

Rows

32

Columns



1

Target

Dataset contains 83293 rows and 33 columns

Data Understanding

Guest

- adults
- children
- babies
- country
- agent
- company
- is_repeated_guest
- previous_cancellations
- previous_bookings_not_canceled
- customer_type

Booking Transaction

- hotel
- stays_in_weekend_nights
- stay_in_week_nights
- meal
- reserved_room_type
- assigned_room_type
- booking_changes
- deposit_type
- adr
- reservation_status
- days_in_waiting_list
- bookingID

Target

- is_cancel

Times

- lead_time
- arrival_date_year
- arrival_date_month
- arrival_date_week_number
- arrival_date_day_of_month
- reservation_status_date

Request

- required_car_parking_spaces
- total_of_special_requests

Market

- market_segment
- distribution_channel

Data Cleaning

Missing Value

Company

78599 NaN Values

agent

11404 NaN Values

country

346 NaN Values

children

3 NaN Values

Duplicate

No Duplicate Data

Convert Type Data

reservation_status_date : DateTime
children : int

Convert Data

column meal:
undefined -> SC
column arrival_date_day_of_month:
29 -> 28

After Cleaning :

83293 Rows , 32 Columns

Data Cleaning

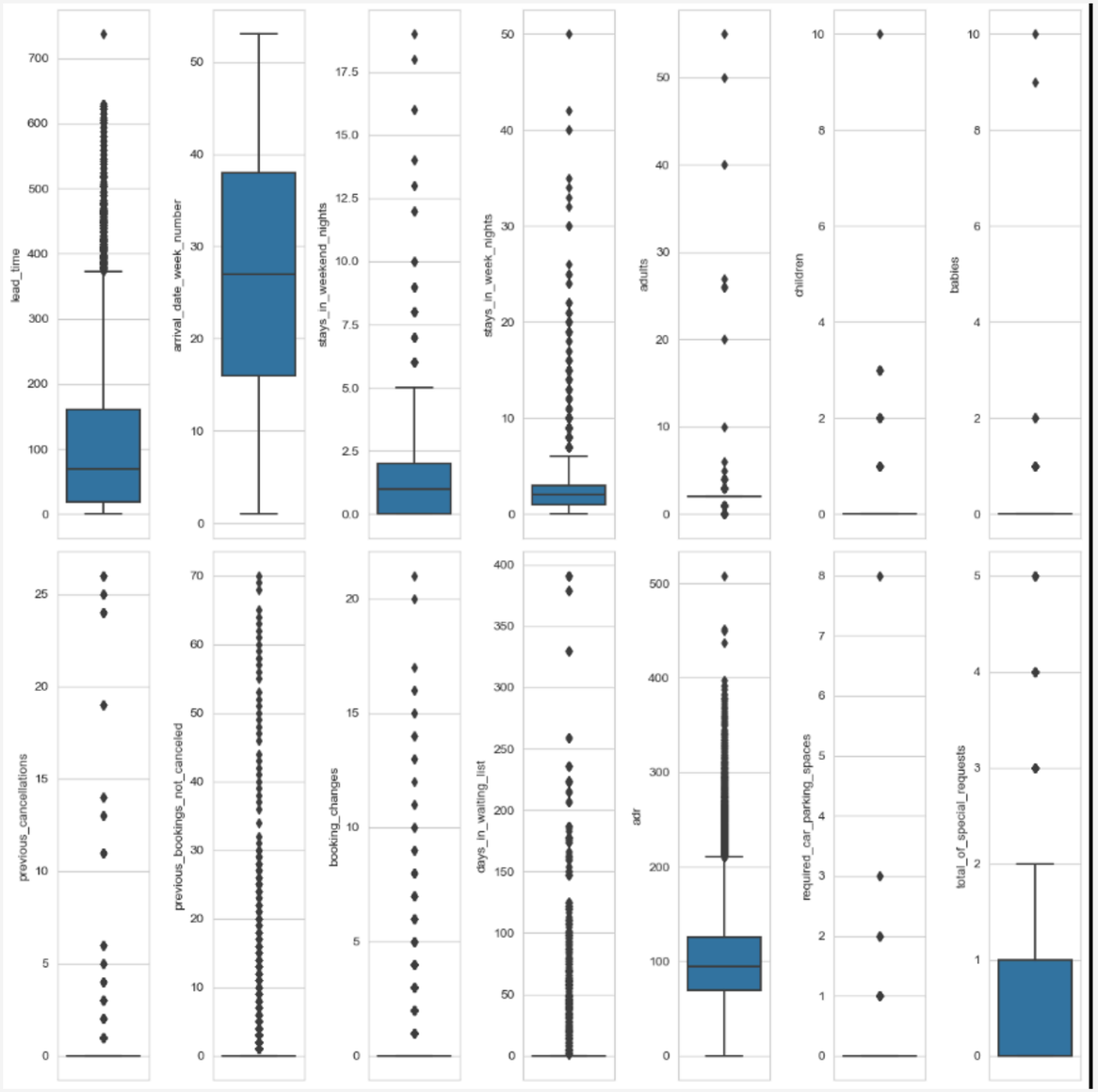
- Drop company Feature because it has NaN value and more than 90% overall Data
- agent feature is filled with 0 value because not all bookings go through agents and the percentage of missing values is only 13.69%
- country feature is filled with mode value because country feature is a categorical type and the percentage of missing values is only 0.42%
- children feature is filled with value 0 (mode) because the percentage of missing value is only 3

Convert Type & Data

- Convert reservation_status_date feature to DateTime type because it was previously object type.
- Convert children feature to Integer because rest feature such as adults and babies is integer
- Convert Data on meal feature because based on dataset information Undefined = SC means no meal.
- Convert Data on date_arrival_day_month because after checking the date there is an odd date on 29 February 2018 and the number is 65 rows, it cannot be leap year so the date is changed to 28 February 28 2018.

Exploratory Data Analysis

Univariate Analysis (Boxplot Outlier)



Outlier Handling

It can be seen that almost all numerical data has outliers, but here we will remove unreasonable outliers such as the adult, children, babies columns and then check whether the columns when summed are 0 because it is impossible for customers who book a hotel to have no one. Also check the total number of stays whether there is anomalous data or not.

Drop Data

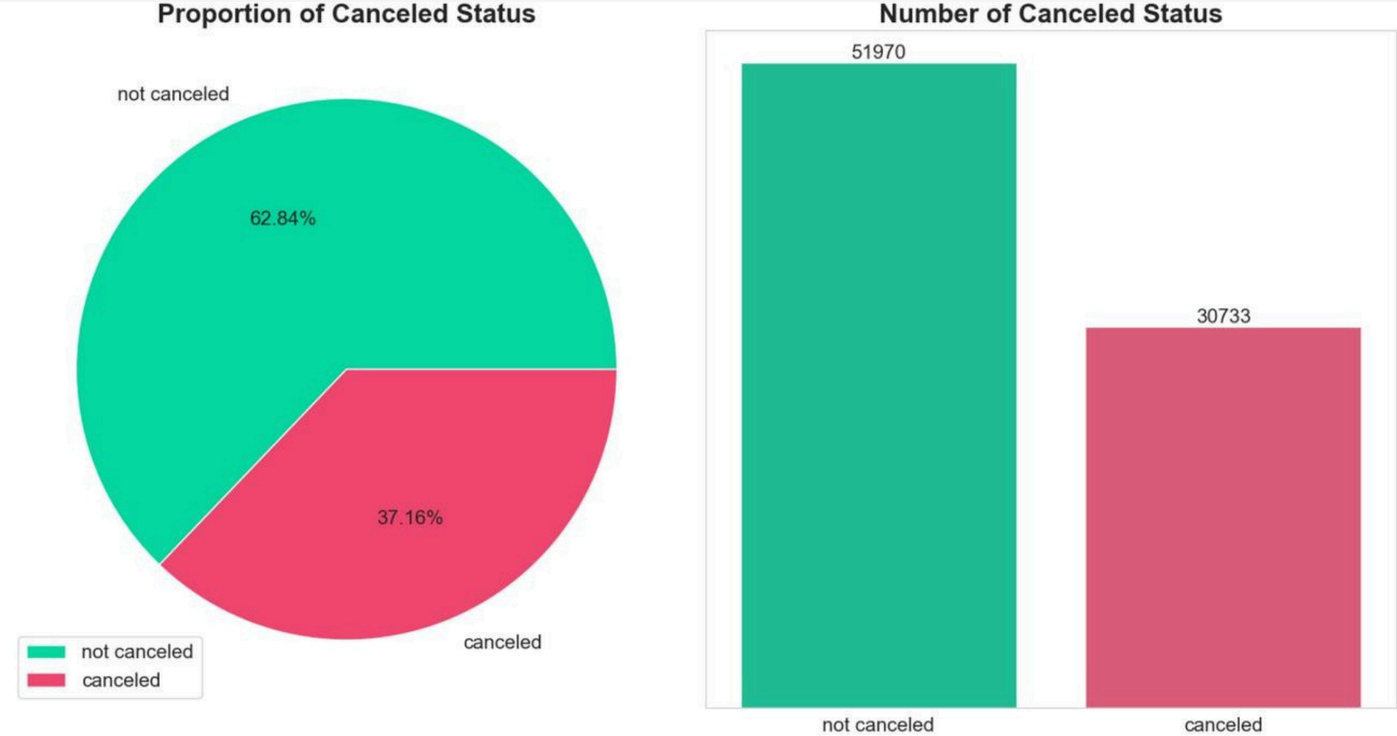
- **Drop 127 anomalous hotel guest data** (adults + children + babies) whose total number of people is 0
- **Drop 3 anomalous babies and children data**
- **Drop 13 anomalous adults data** where for 1 hotel room is usually filled with a maximum of only 4 adults while there are extreme values above 4
- **Drop 445 anomalous stays_in_weekend_nights & stays_in_week_nights data that total 0. It could bepossibility of data errors or forgotten input from the hotel.**
  Drop 2 required_car_parking_spaces anomaly data because the number of guests is only 2 but requires 8 car parking spaces.

Total Drop Data : 590 (0.71%)

After Outlier Handling

82703 Rows

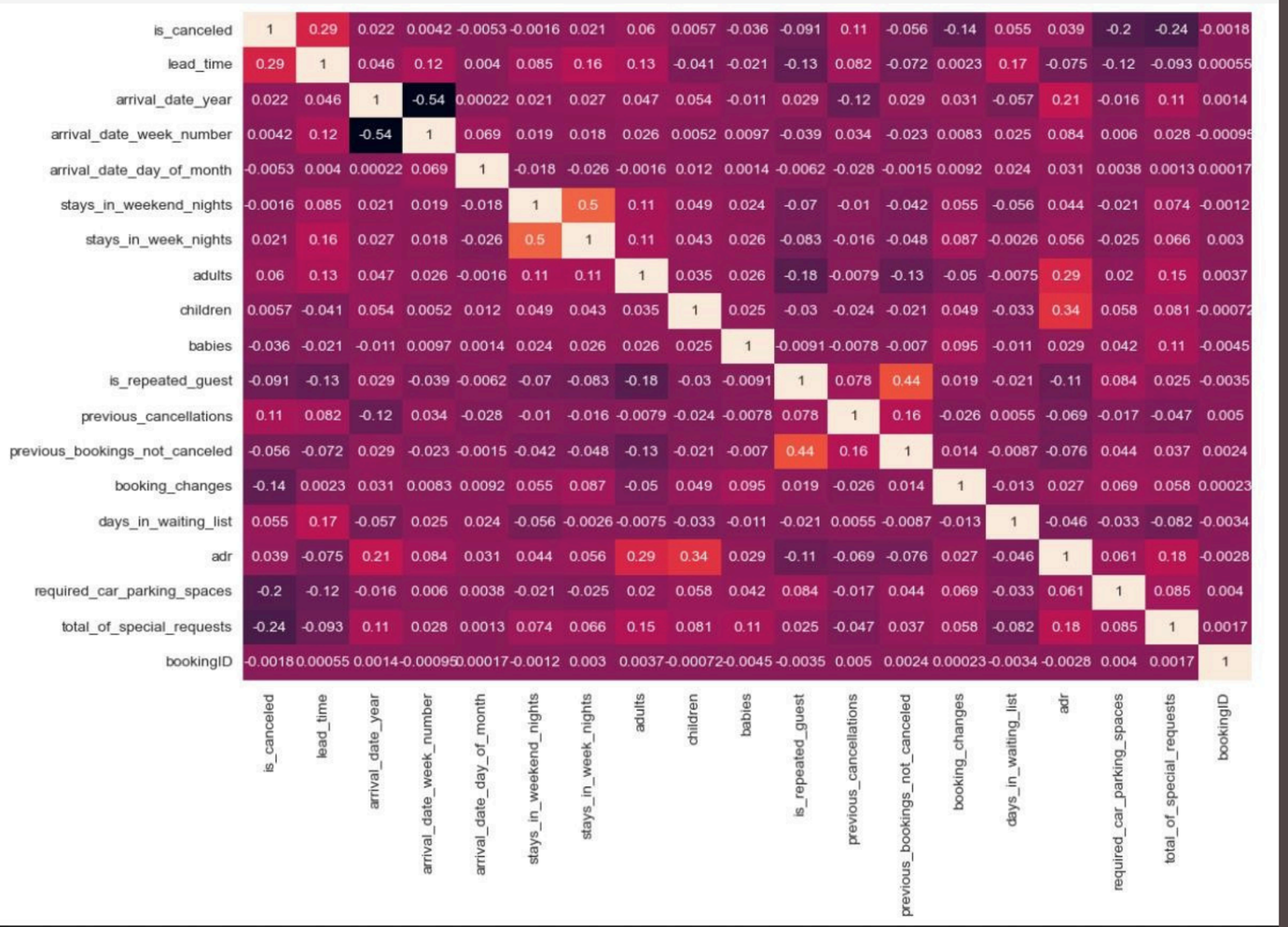
Univariate Analysis (Proportion Target)



How is the proportion between bookings that cancel and not cancel ?

The proportion of not canceled and canceled imbalance, it can be seen that the majority is more not canceled but the number of canceled of 37.16% can be said to be quite a lot, that if there are 100 people who book a hotel the possibility of canceling an order is 37 people. With a significant number of canceled, this can affect hotel revenue and profit and make the hotel suffer losses.

Multivariate Analysis (Feature Correlation)



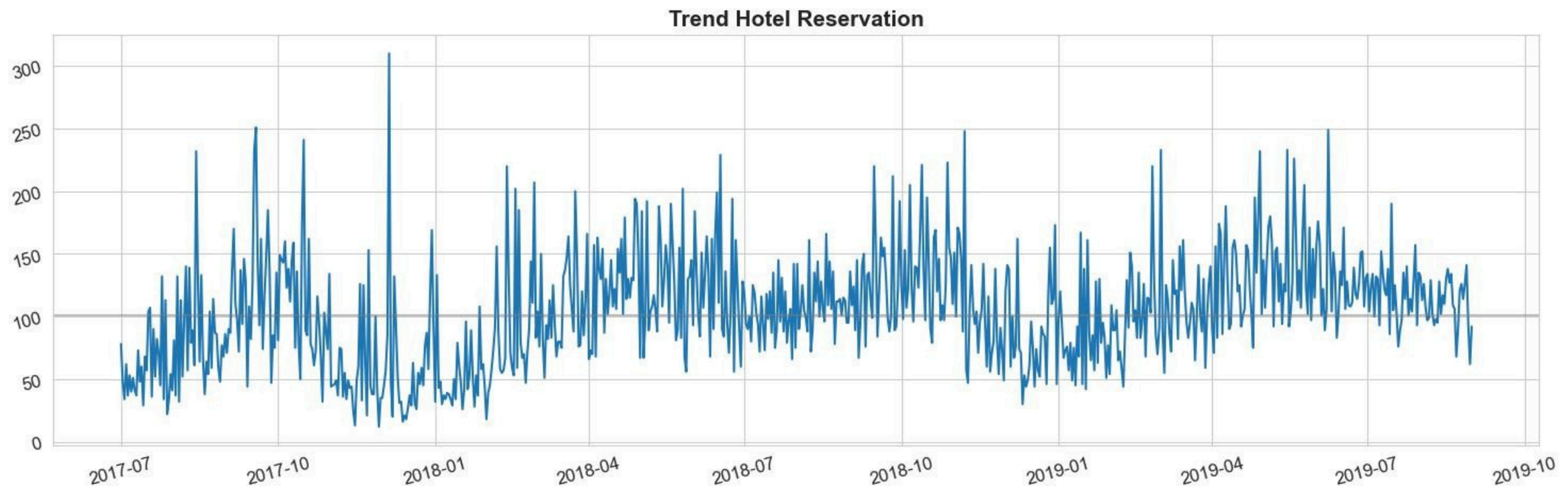
How is the correlation between features and the target?

- Based on the heatmap of each feature, there is no very strong correlation, a fairly strong positive correlation is found in previous_bookings_not_canceled with is_repeated_guest, meaning that with a booking that is not canceled, it is likely that the customer will book again and vice versa.
- Then there is also a fairly strong correlation between children, adults and adr, meaning that with an increase in the number of bookings staying or bringing children, adr will also increase,
- while for is_canceled and lead_time the correlation is 0.29, meaning that the possibility of canceling occurs if the waiting time is higher and vice versa.

Deep Dive Analysis

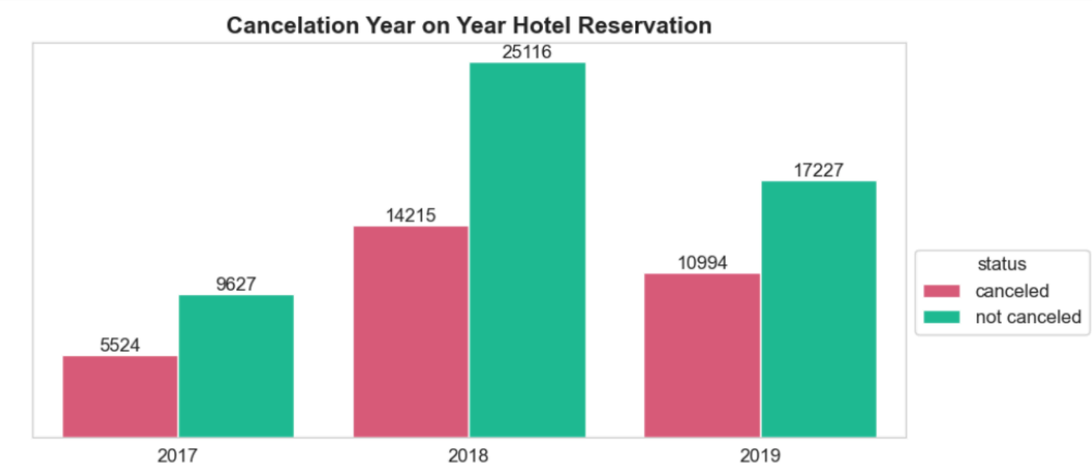
Question 1 :
How is the trend of hotel bookings?
When do cancellations happen?

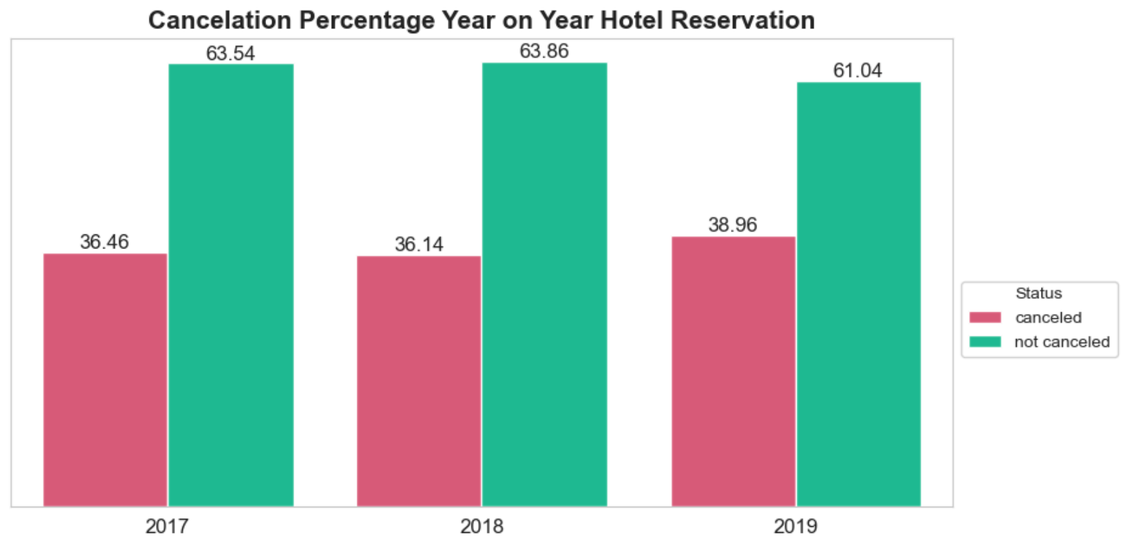
Trend Daily Hotel Reservation



There are 101 hotel orders a day (median)

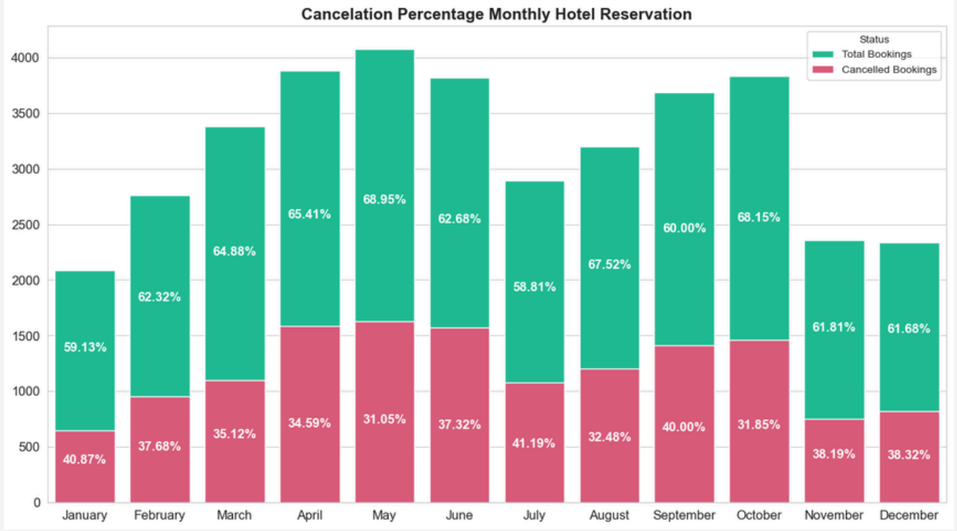
Year on Year Hotel Reservation





It can be seen that hotel orders were highest in 2018 and the number of cancellations was high in that year. But the number of cancelation percentages actually increases every year. Because in 2019 the percentage of cancelations is higher than in previous years

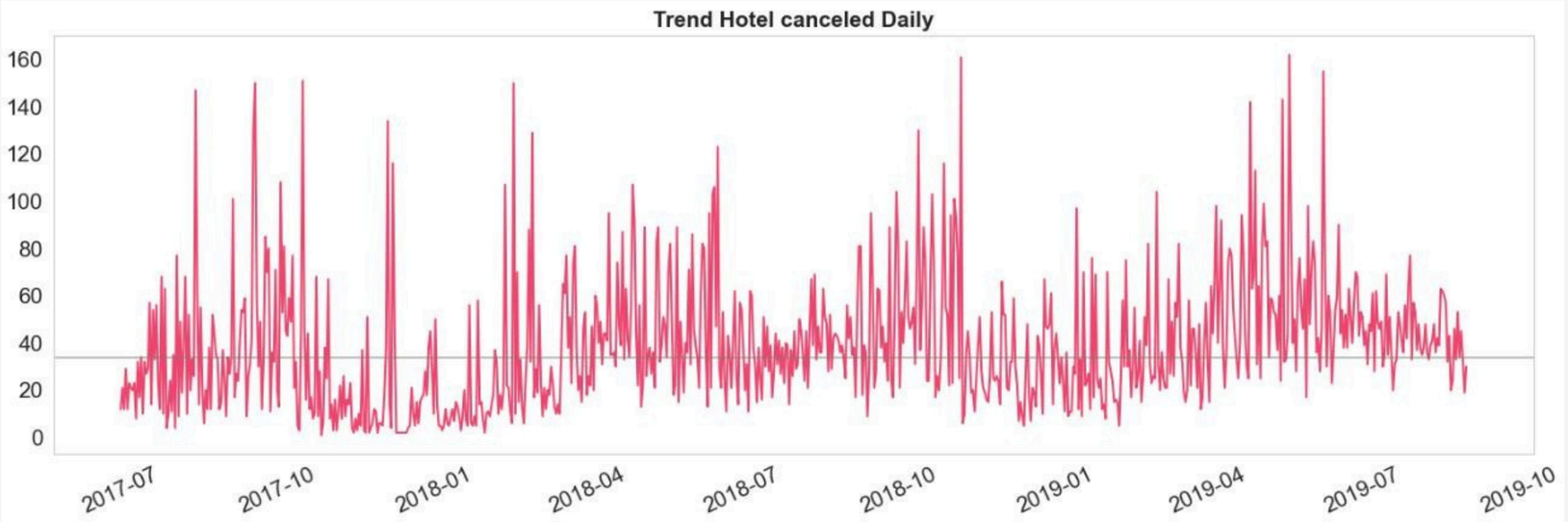
Monthly Hotel Reservation



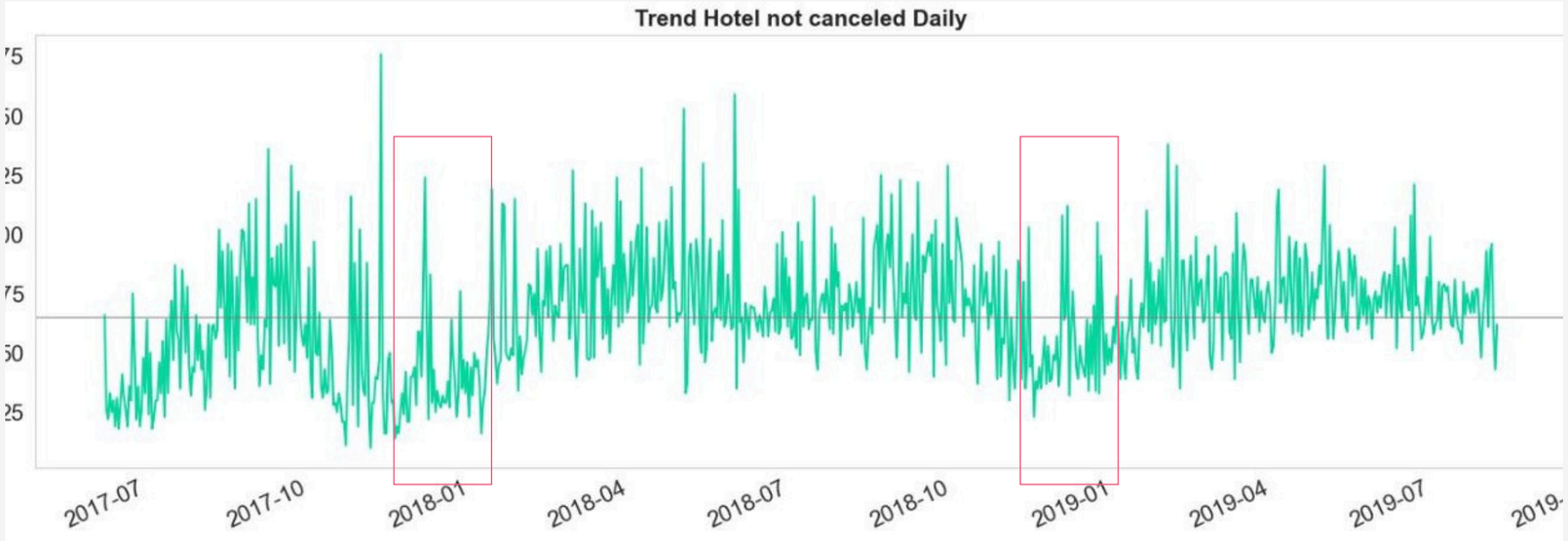
- The highest hotel reservations each year are in September 2017, October 2018 and May 2019 and along with high reservations in those months, high cancellation also happened. In addition, there is also a pattern that the decline in hotel reservations happens at the beginning of the year (January).

- The highest percentage of hotel reservations is in May and the cancellation rate is low, the high reservations happen because March - May is the right time for vacation, while the highest cancellation rate is in July and January.

Daily Hotel Reservation



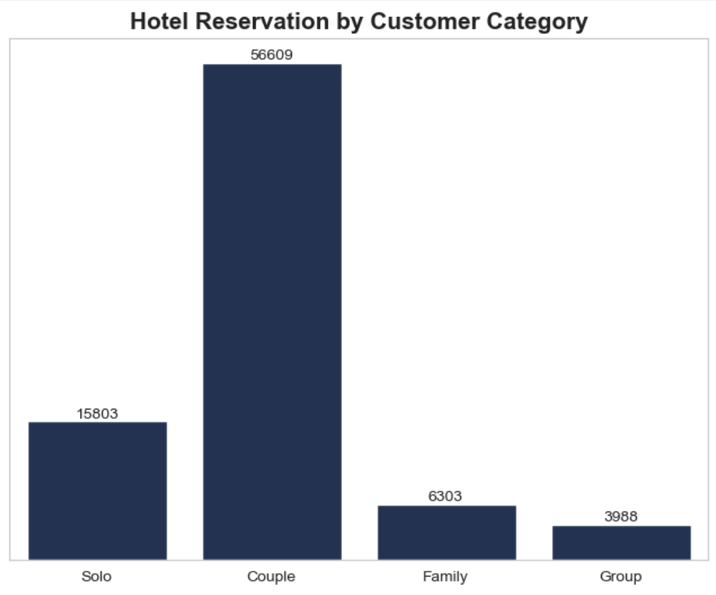
There are 34 cancellations per day (median)

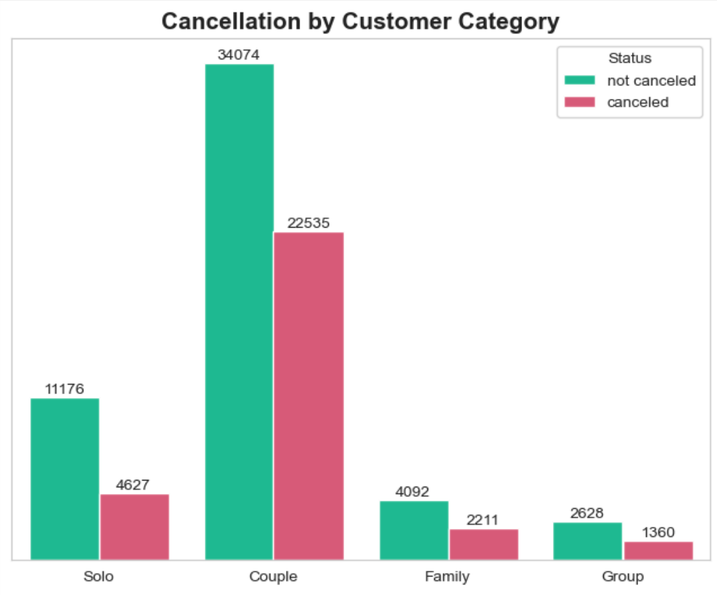


There are 65 non-cancellations every day but there is a pattern that almost non-cancellations decrease in the month at the beginning of each year (January).

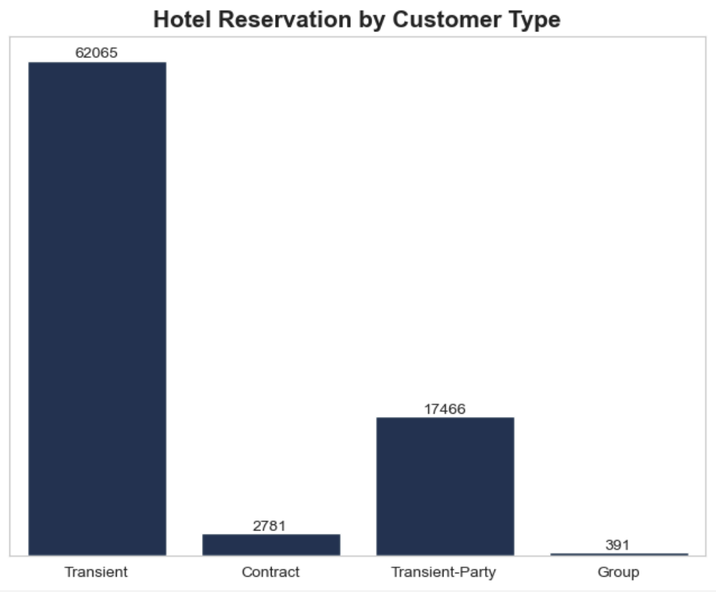
Question 2 :
What are the criteria for customers who frequently cancel hotel reservations?

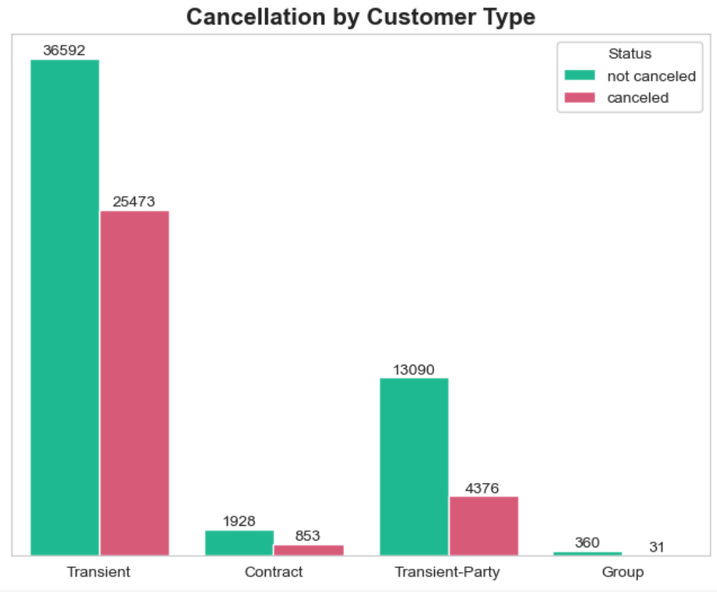
Hotel Reservation by Customer Category





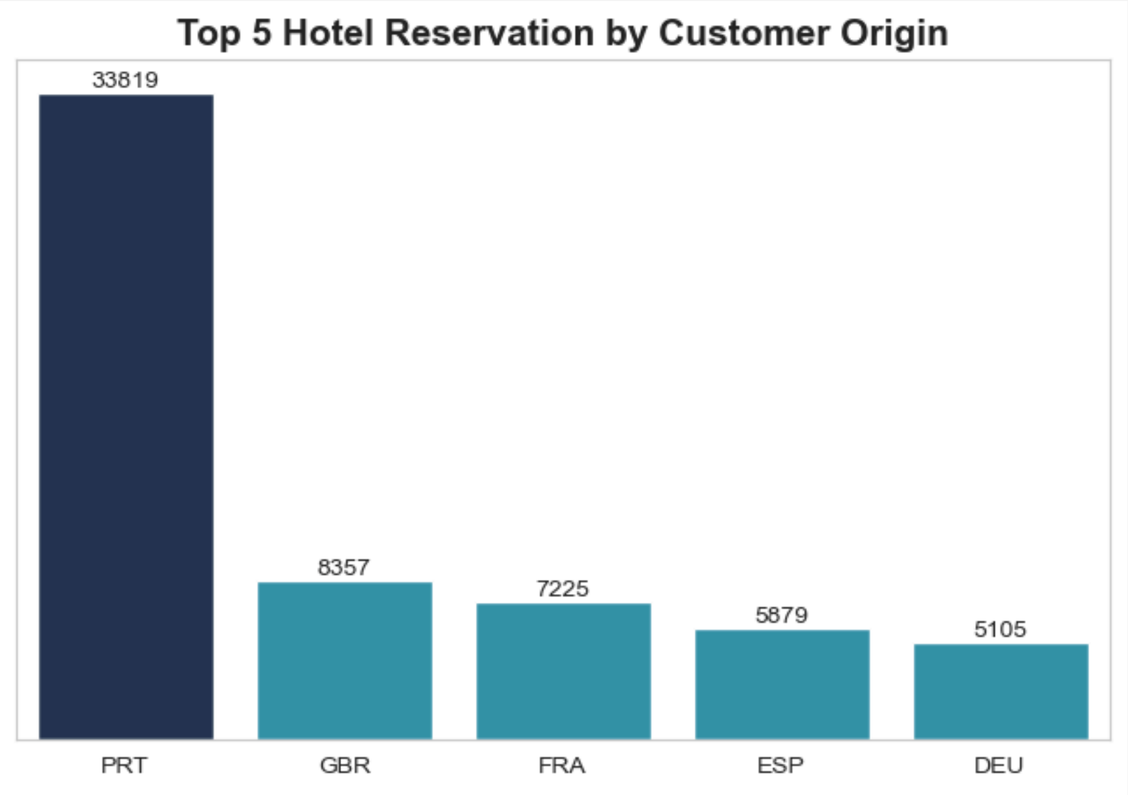
Based on the customer category, hotel reservations are mostly booked by couple customers and hotel reservation cancellations tend to be more than other categories, which is around 40% canceled.

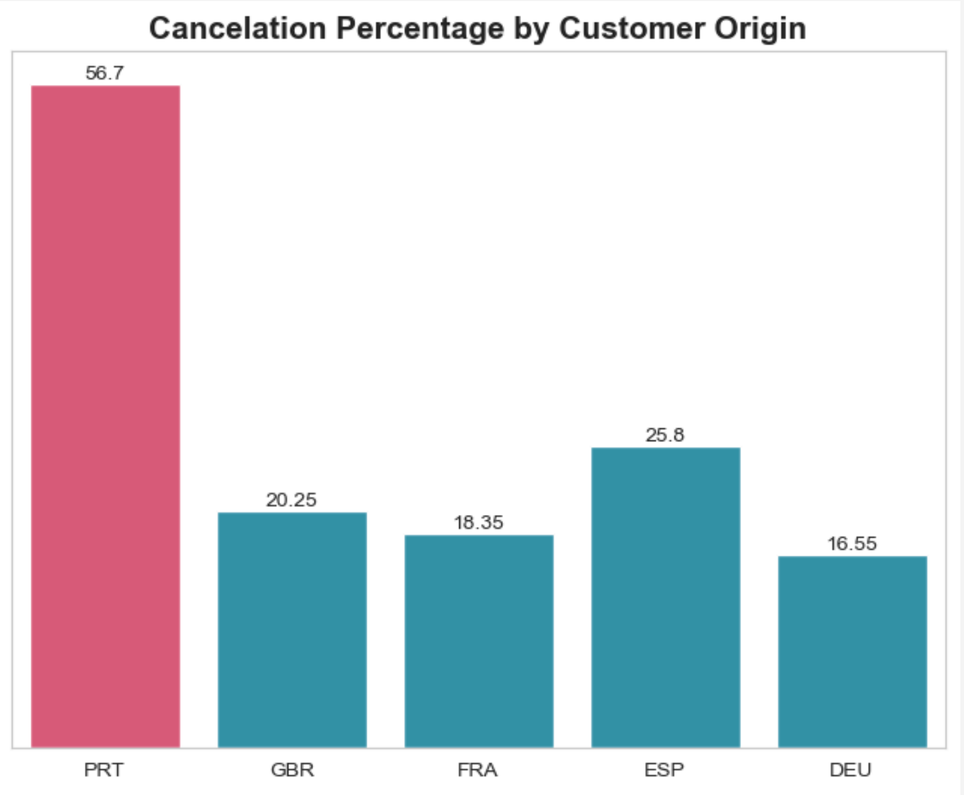




Based on customer type, hotel reservations are mostly booked by the Transient type and cancellation of reservations tends to be more than other types, which is 41%.

Hotel Reservation by Customer Origin





Most customers come from Portugal, it can be assumed that they are local residents and cancelation happens most in Portugal with a percentage of 56.7% compared to other countries.

Question 3 :
What hotel criteria and reservation types have the most cancellations?

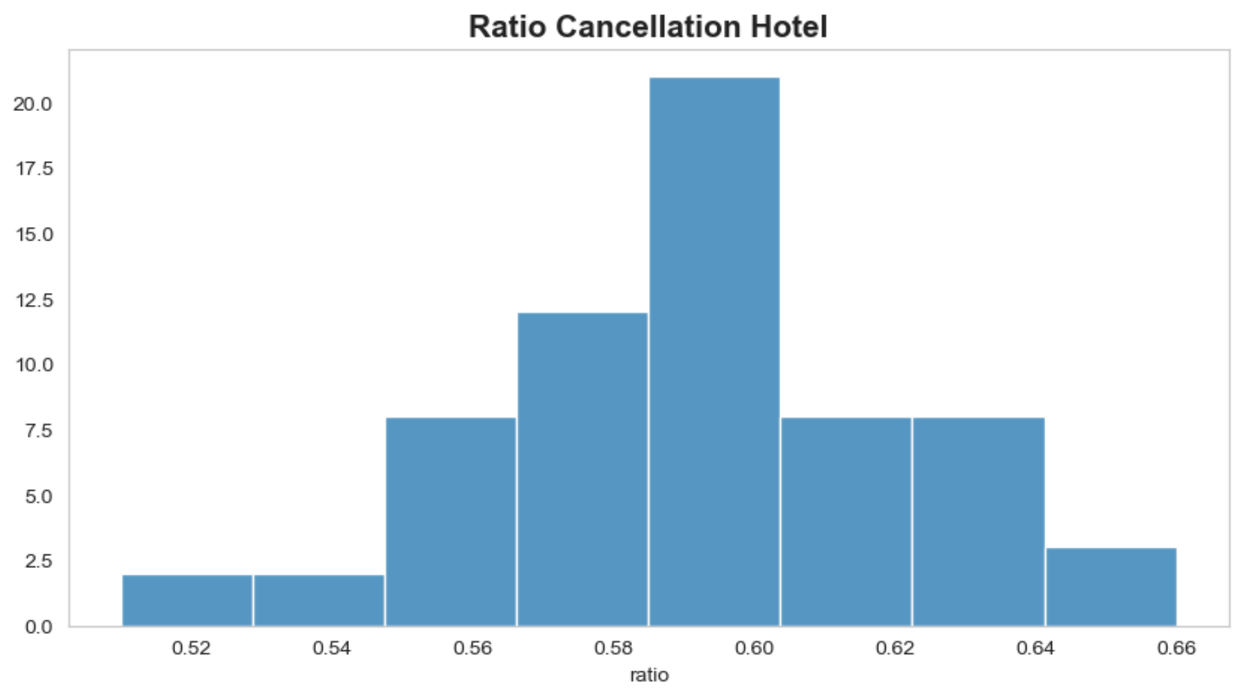
Most Reserved Hotel

Renaissance New York Times Square Hotel 1367 Orders

Renaissance New York Times Square Hotel is a frequently reserved hotel but this does not show that the hotel's performance is good, to see the hotel's performance is good or not, it must be compared between cancel and non cancel.

Therefore, the cancellation ratio of each hotel will be made and the smallest cancel ratio will be seen.

Ratio Cancel and non Cancellation

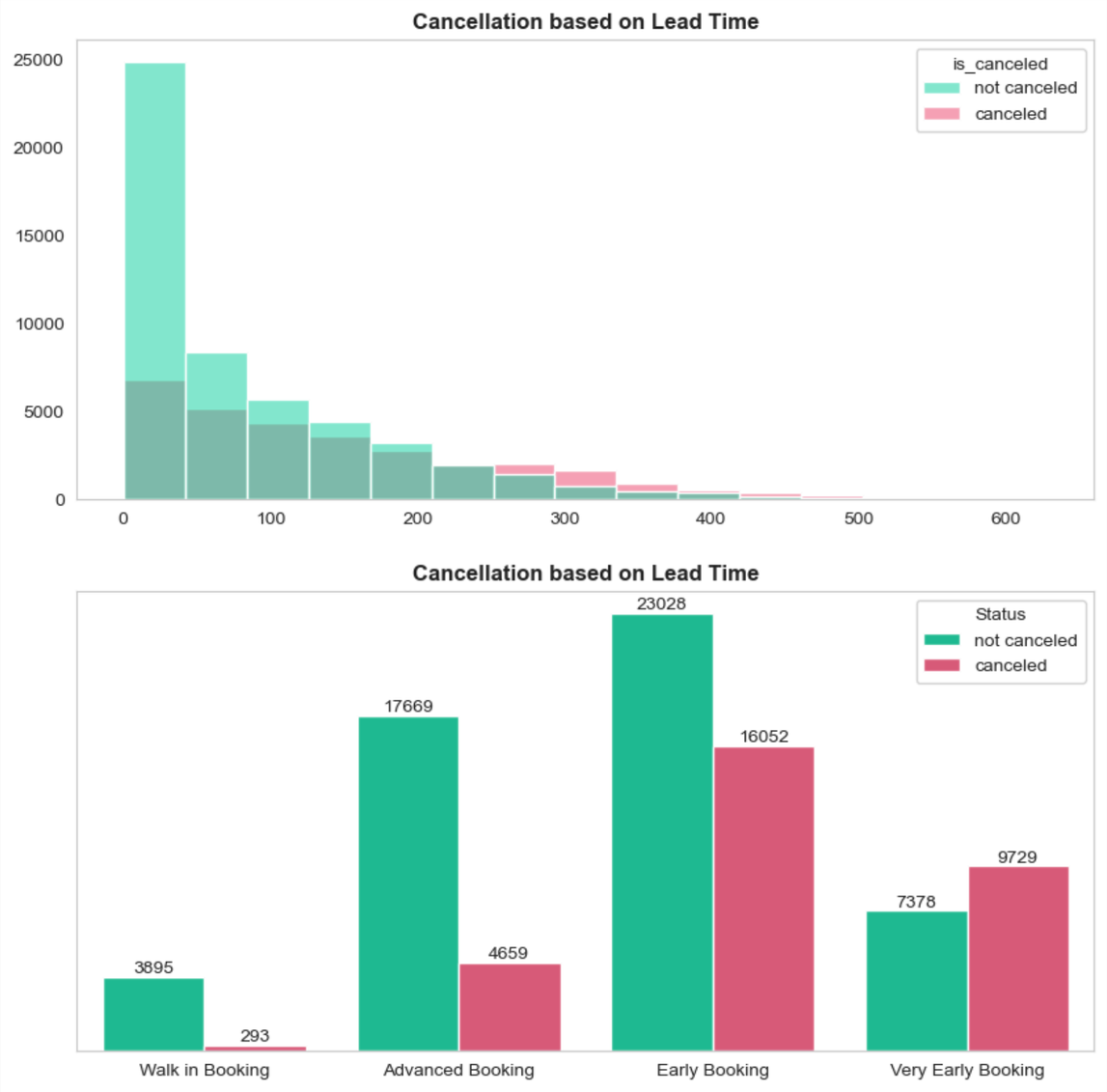


hotel	cancel	non cancel	ratio
Indianapolis Airport Courtyard Indianapolis, IN	434	847	0.51
Spokane Downtown at the Convention Center Cour...	434	841	0.52
Protea Hotel Fire & Ice! by Marriott Johannesb...	439	817	0.54
Greensboro Courtyard Greensboro, NC	439	812	0.54
The Ritz-Carlton, Tokyo Tokyo, Japan	452	820	0.55

It can be seen that the hotel cancellation ratio is between 0.52 - 0.66, which means that if there are 100 reservations, the possibility of canceling is 52 - 66 reservations.
based on cancellation ratio the hotels that have the lowest cancellation ratio are

- Indianapolis Airport Courtyard
- Downtown Spokane at the Convention Center
- Protea Hotel Fire & Ice! by Marriott
- Greensboro Courtyard

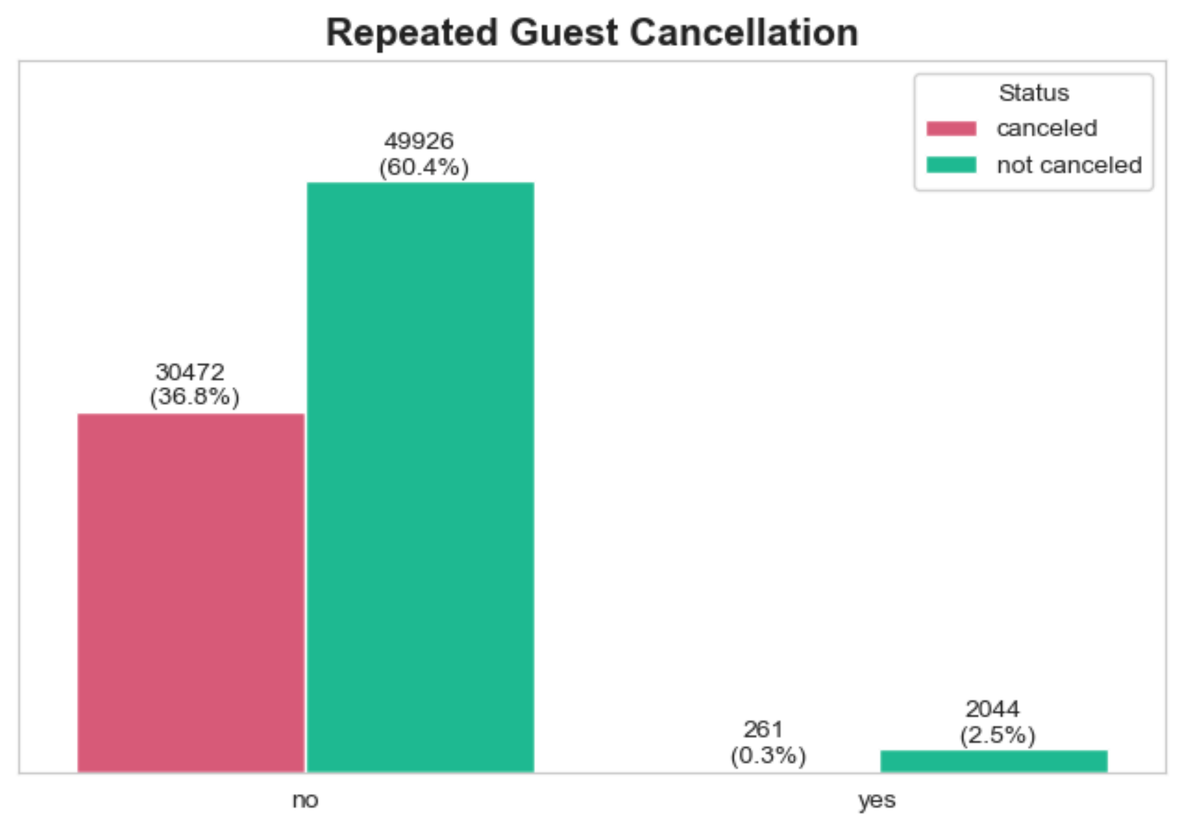
Cancellation by Lead Time



- If we look at it, the longer the waiting time, the higher the number of canceled reservations compared to non-canceled ones. Offering special pricing or discounts for customers who book advanced can be an effective way to incentivize them to book early and potentially reduce cancellations. Implementing a tiered pricing system that offers higher prices for bookings closer to the lead time can also help ensure that customers who book closer to their travel date are aware of the additional costs.

- Sending reminders to customers about their reservations can also be an effective way to reduce cancelations. Reminders can be sent via email, SMS, or other communication channels, depending on customer preferences and contact information.

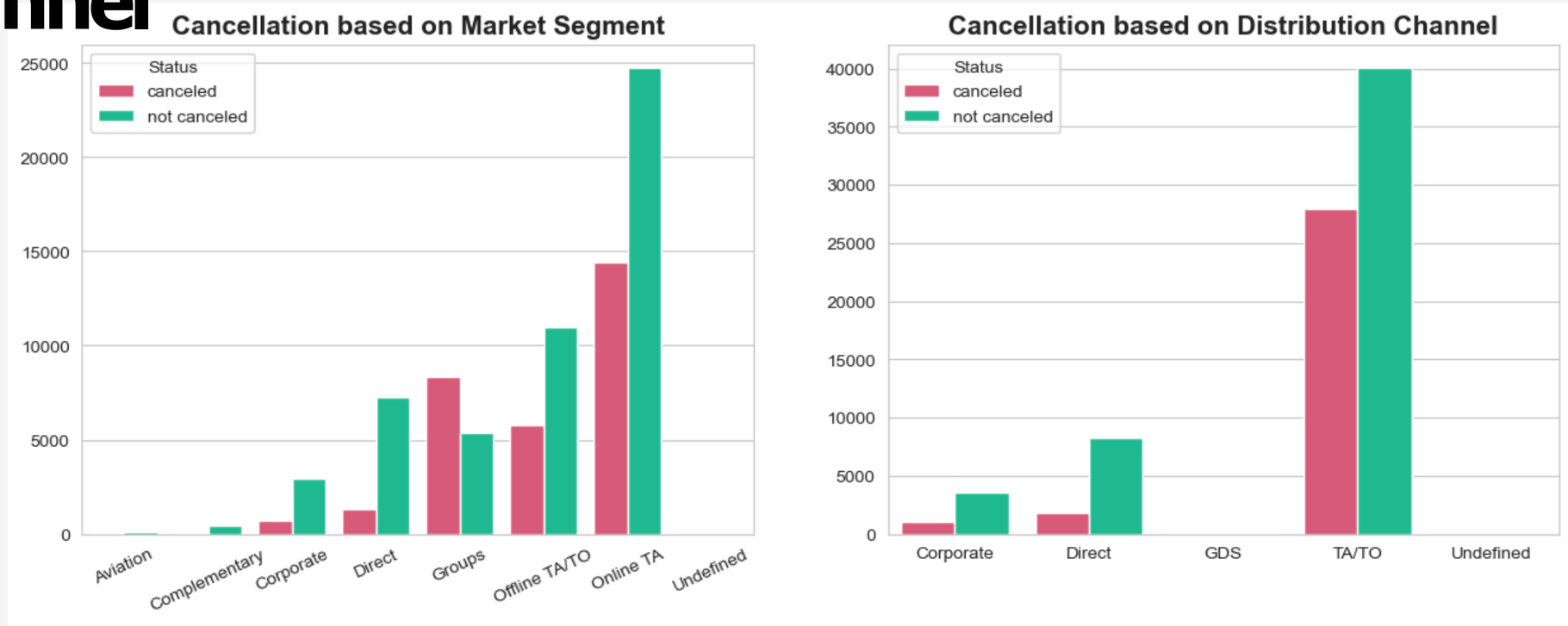
Cancellation by Repeated Guest



- Previous customers tend to be more loyal customers, with a very low reservation cancellation rate of 0.3%. This may indicate that the hotel service is good, so customers who have already stayed tend not to cancel their reservation.

- New customers tend to cancel their reservations more easily, with a reservation cancellation rate of 36.8%, possibly because they have not found a hotel that matches their wishes, the company can improve services and provide promos so that there are fewer cancellations.

Cancellation by Market Segment & Distribution Channel

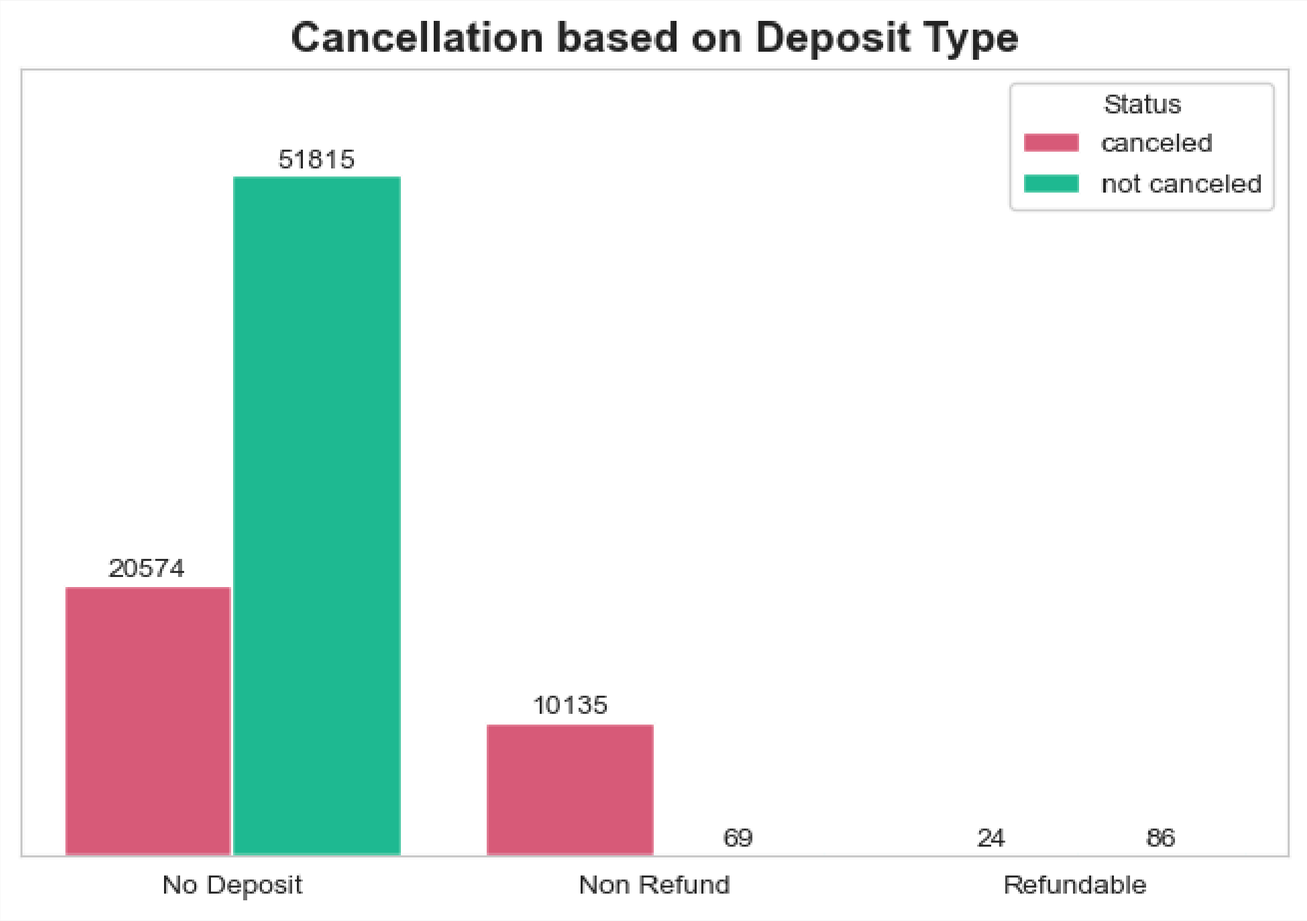


Cancellations happen mostly in the Online and Groups market segments while complementary none, there are several possibilities:

- There is no fee to be paid by the customer, so they may feel more committed to coming to the hotel and not want to cancel their reservation.
- Complimentary stays are usually given to customers who have loyalty to the hotel or are frequent guests, so they tend to plan their trip more carefully and are less likely to cancel their reservation.

Based on the distribution channel, the most cancelation occurs in travel agents and tour operators, this is in line with orders that are mostly made through travel agents.

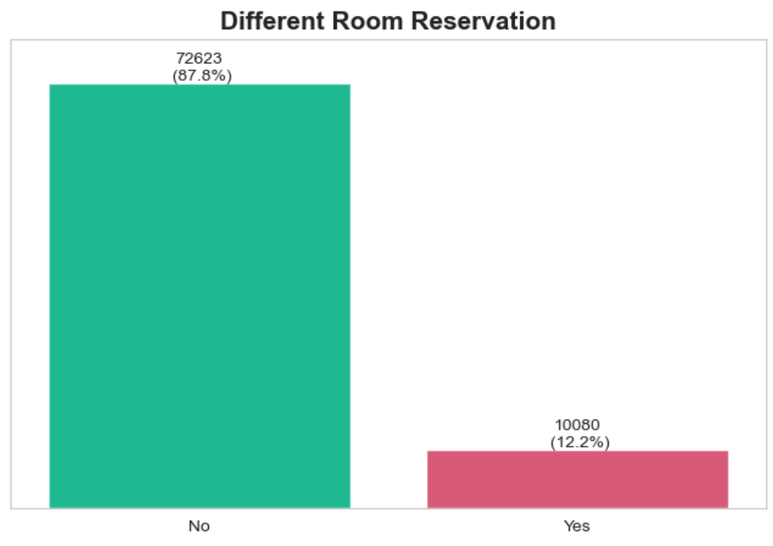
Cancellation by Deposit Type

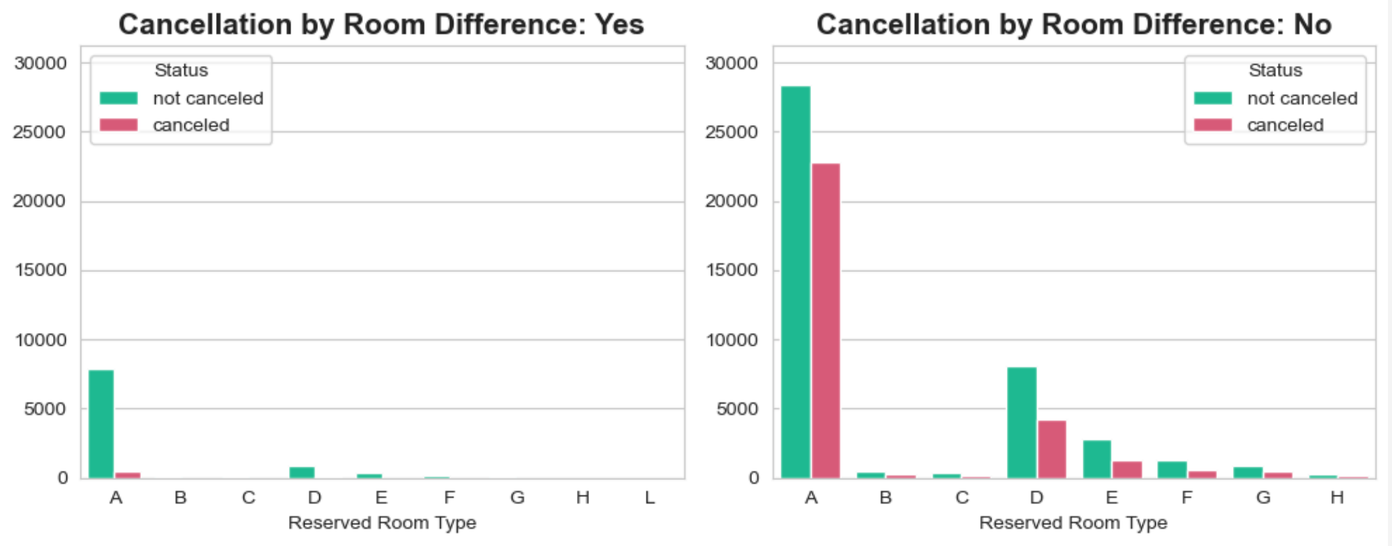


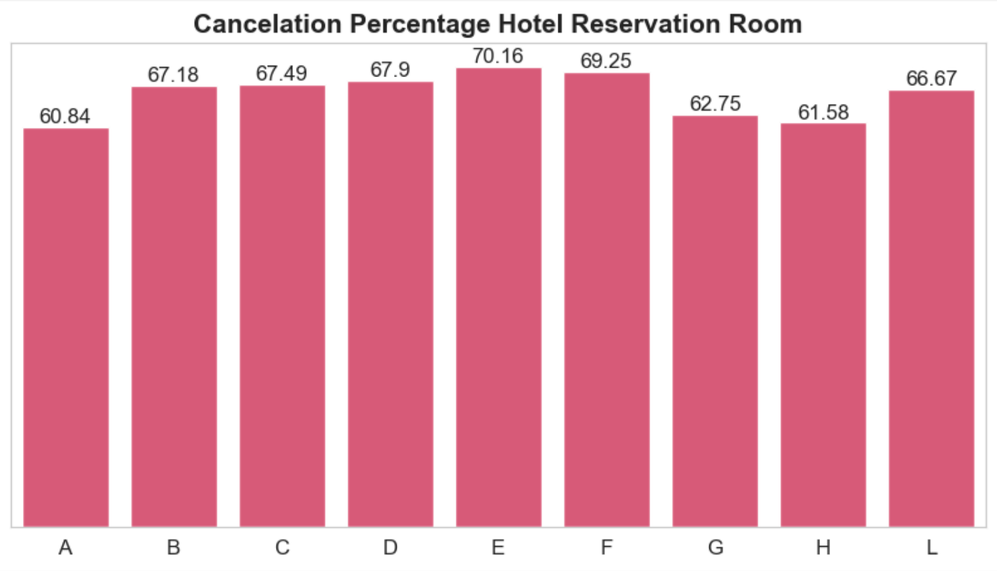
- Based on the deposit type, the number of reservations and cancelations is highest in the No Deposit type but compared to the Non Refund type, the highest percentage of cancelations is actually found in the Non Refund type where almost 100% of reservations are canceled, this is normal because customers do not want to lose their money due to order cancellation and definitely choose other options.

- For cancellation on No Deposit can be a concern by the hotel to notify those who want to book a hotel to make a down payment (DP) in advance so that it can reduce losses and the customers will not feel disappointed if the reservation is canceled one sidedly.

Cancellation by Room Type

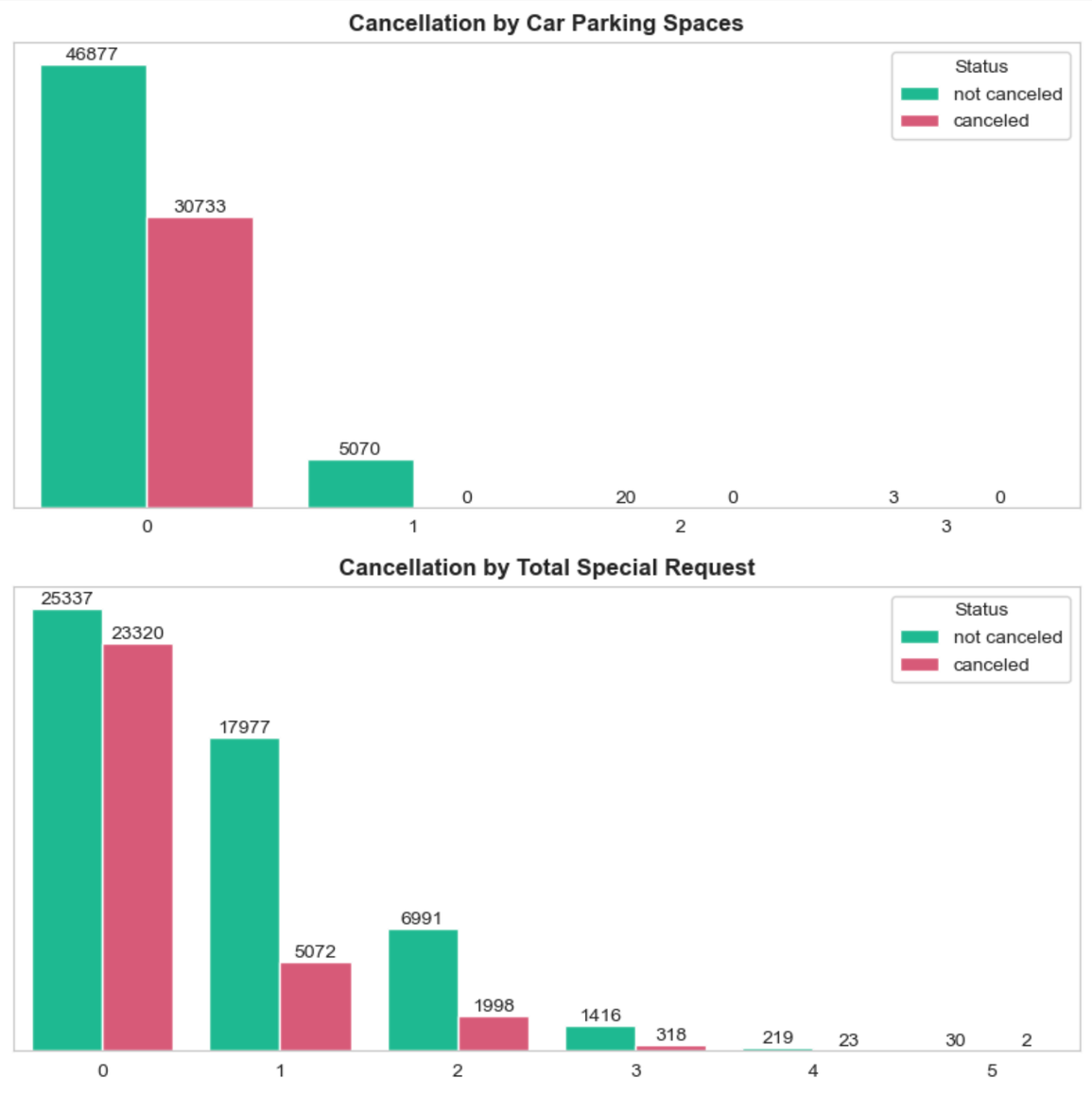






- About 12.2% of the difference between the rooms reserved and assigned,
  so it needs to be reviewed whether this causes cancellation or not.
- Cancellations happen in rooms that change but the comparison is very small so it can be said that cancellations are not influenced by changes in room type because rooms that do not change actually have more cancellations.
- It can be seen that room type A is the most reserved and the cancel rate is
  high, although the number of cancellations on room A is very high but based on the percentage of cancellations, the highest is room type E at 70.18% and the lowest is room A at 60.84%.

Cancellation by Request

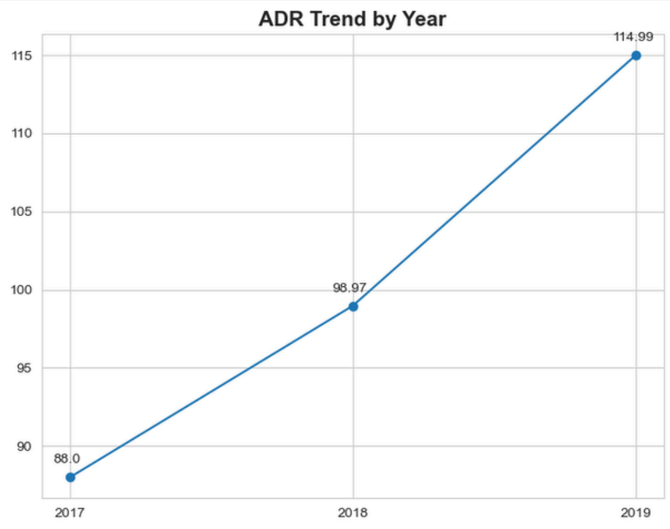


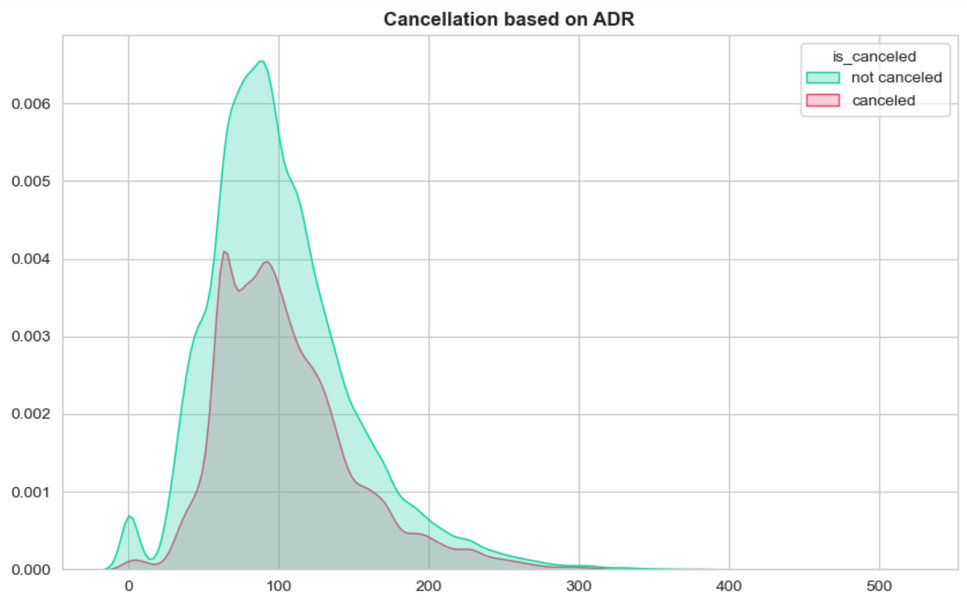
- That hotel reservations are more canceled for customers who do not bring a car.

- Most hotel order cancellations happen when customers do not ask for special requests. It is likely that customers who have reserved a hotel and asked for special requests have paid more for their hotel reservations so that the possibility of cancellation is very small.

Question 4 :
How does the booking status relate to hotel performance, specifically with ADR (Average Daily Rate)?

Cancellation by ADR





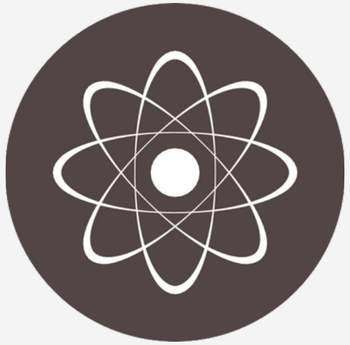
- Every year the price of hotel ADR increases
- Based on the monthly mean adr, it can be seen that the highest adr happens every August. It is recommended that the hotel can increase the average room rental price during peak hotel bookings and implement a cancellation policy during quiet bookings with only partial refunds to minimize losses when customers cancel their reservations. However, during peak bookings, reservation cancellations are free of charge due to high demand, and if there are any cancellations, they can be quickly replaced by other customers.
- Customers who do not spend money and spend more money on room costs tend to cancel fewer reservations, and the peak of canceled reservations is for customers who spend between 70-100 euros on room costs.

Modelling

Workflow Modelling Process

Feature Engineering

Train Test Split



Model Training

Evaluation

Hyperparameter Tuning

Evaluation Best Model

Feature Engineering

Drop Data “Undefined

Drop Data “Undefined” in Distribution Channel and Market Segment Undefined

Add new Feature

- Feature “kids” : children + babies
- Feature “room_changes” : reserved_type_room == assigned_type_room
  If the reserved room is the same as the assigned room, fill in 0 otherwise fill in 1.
  (It is to reduce the complexity of the model in training and the features can be combined so as to reduce the number of features to be trained)
- Feature “total_stays” : stays_in_week_nights + stays_in_weekend_nights
- Feature “total_guest” : adults + kids

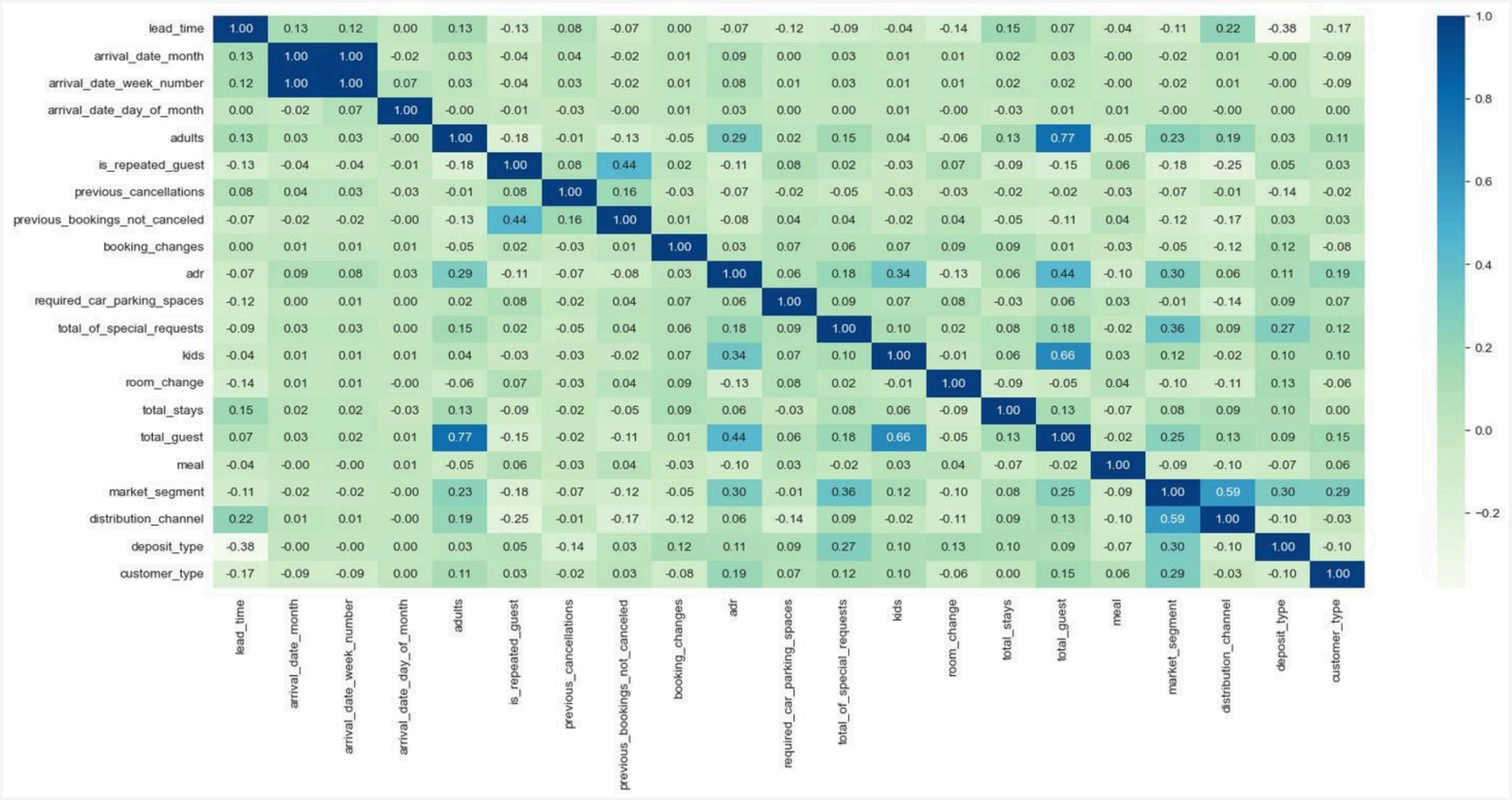
Encoding Categorical Features

- Feature “arrival_date_month” : ordinal encoding
- Feature “meal, market_segment, distribution_channel, deposit_type, customer_type” : Frequency encoding to avoid multicollinearity when using One Hot Encoding and avoid the curse of dimensionality because the features encoded by one hot encoding will create new features according to the number of categories.

Drop Unused Features

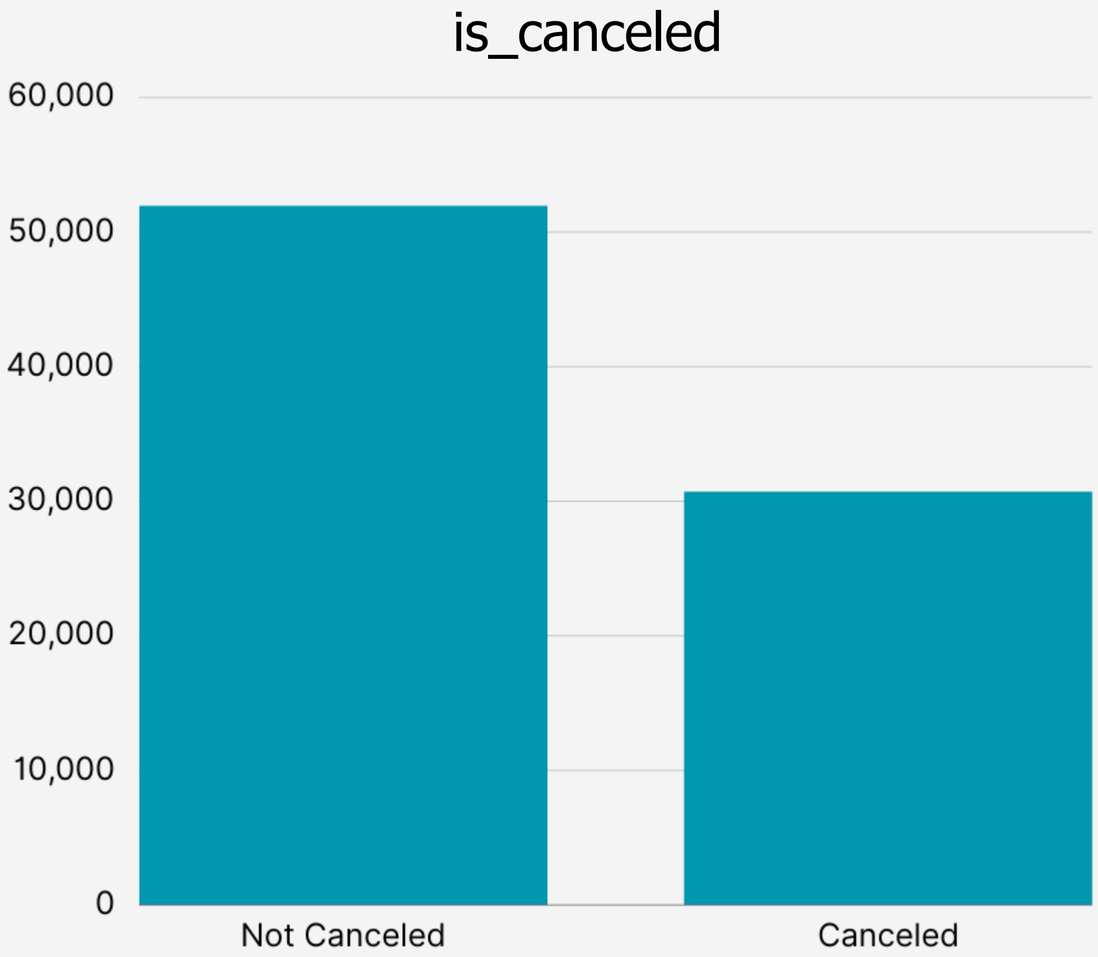
Drop Feature :
‘bookingID', 'reservation_status', 'reservation_status_date', arrival_date_year', 'assigned_room_type', 'reserved_room_type', 'babies' ,'children', 'hotel', 'country', 'agent', 'days_in_waiting_list', 'stays_in_weekend_nights', 'stays_in_week_nights'

Multicolienarity Checking



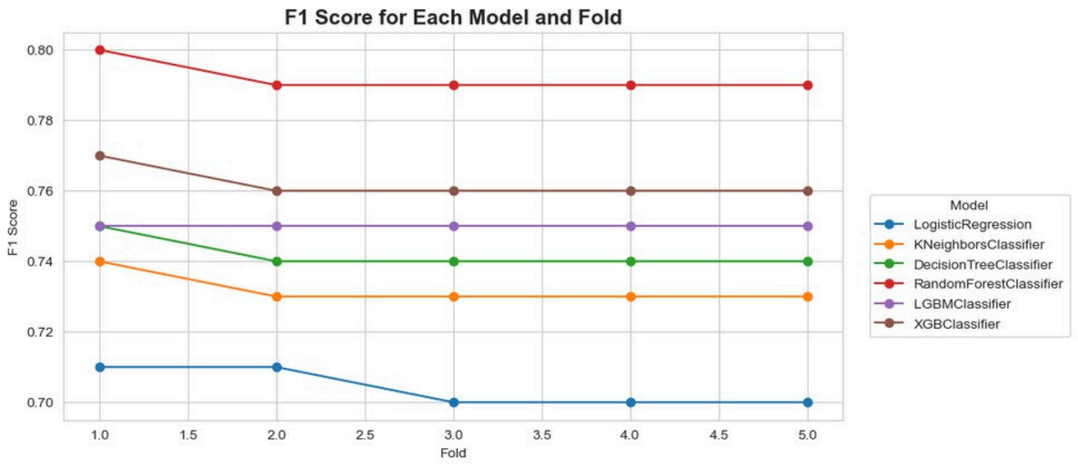
After multicolinearity checking there is a strong correlation between `arrival_date_month` and `arrival_date_week_number` of 1 meaning that the information possessed by both features is the same or redundant and after checking the correlation with the target feature `arrival_date_month` has a stronger correlation than `arrival_date_week_number` therefore the
`arrival_date_week_number` feature will be dropped.

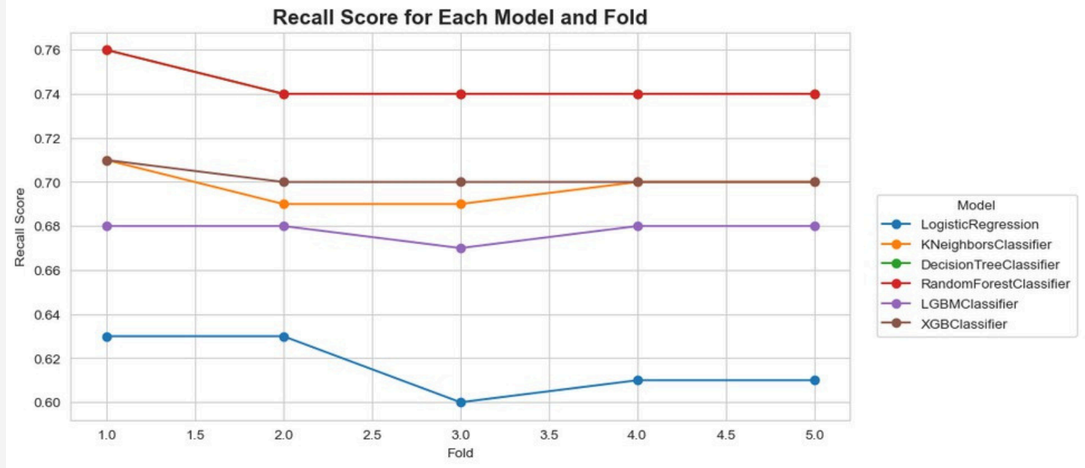
Modeling Train



- For Scaling Feature using **Robust Scaler** because it is seen that the data mostly has outliers so, this scaler is used in the hope of getting good performance results.
- The dataset is splitted into train data (80%) and test data (20%)
- There will be 2 models, the first will use the base model and the second will use the undersampling method for handling imbalance data and then will be compared

Base Model Training Evaluation



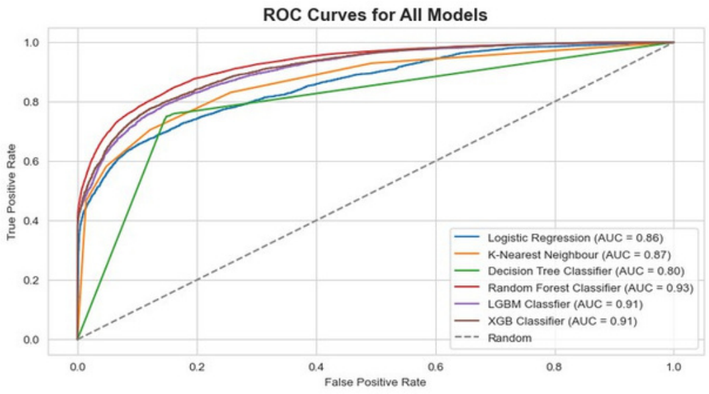


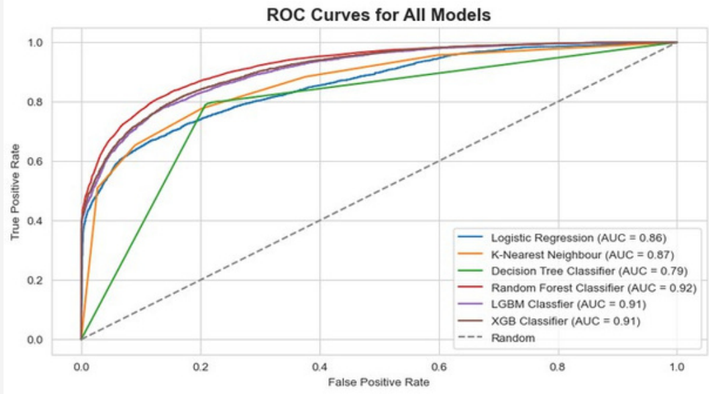
model	f1_score_mean	f1_score_all	recall_mean	recall_all
LogisticRegression(random_state=42)	0.705699	[0.71, 0.71, 0.7, 0.7, 0.7]	0.617646	[0.63, 0.63, 0.6, 0.61, 0.61]
KNeighborsClassifier()	0.734054	[0.74, 0.73, 0.73, 0.73, 0.73]	0.699573	[0.71, 0.69, 0.69, 0.7, 0.7]
DecisionTreeClassifier(random_state=42)	0.739613	[0.75, 0.74, 0.74, 0.74, 0.74]	0.744867	[0.76, 0.74, 0.74, 0.74, 0.74]
RandomForestClassifier(random_state=42)	0.792145	[0.8, 0.79, 0.79, 0.79, 0.79]	0.743362	[0.76, 0.74, 0.74, 0.74, 0.74]
LGBMClassifier(random_state=42)	0.753505	[0.75, 0.75, 0.75, 0.75, 0.75]	0.677617	[0.68, 0.68, 0.67, 0.68, 0.68]
XGBClassifier(base_score=None, booster=None, c...	0.763227	[0.77, 0.76, 0.76, 0.76, 0.76]	0.700468	[0.71, 0.7, 0.7, 0.7, 0.7]

Based on the cross validation results, the models that have good performance are Random Forest, LGBM and XGB Classifier. Furthermore, it will be compared to the test results

Base Model Evaluation Comparison

model	precision_base	precision_under	recall_base	recall_under	f1_score_base	f1_score_under	AUC_score_base	AUC_score_under
LogisticRegression	0.829002	0.711427	0.618745	0.718500	0.708606	0.714946	0.858	0.857
KNeighborsClassifier	0.773646	0.693673	0.705297	0.777343	0.737892	0.733128	0.869	0.866
DecisionTreeClassifier	0.747526	0.688047	0.750937	0.793806	0.749228	0.737153	0.803	0.792
RandomForestClassifier	0.850407	0.774665	0.749633	0.820375	0.796847	0.796865	0.926	0.923
LGBMClassifier	0.848307	0.766427	0.681826	0.781418	0.756009	0.773850	0.907	0.906
XGBClassifier	0.841700	0.756377	0.713284	0.797555	0.772190	0.776420	0.912	0.910





Base Model **Base Model + Undersampling Method**

It can be seen that the Random Forest, LGBM and XGB models have good performance when evaluated using test data. Then there is an increase in performance when using an undersampled model (handling imbalance) seen in the F1 score and recall score values that increase. Then the 3 models will be hyperparameters tuning to get the best model.

Hyperparameter Tuning Parameter

Base Model

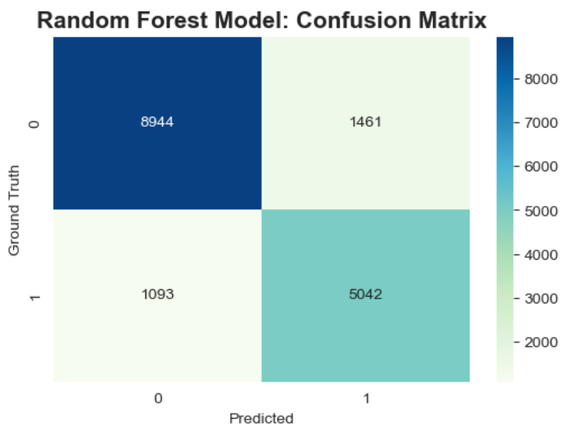
Grid Search CV = 5		
Model	**Best Scores**	**Best Parameters**
Random Forest	0.793360	‘max_depth': 50, 'min_samples_split': 2, 'n_estimators': 500
LGBM Classifier	0.779192	'learning_rate': 0.5, 'max_depth': 9, 'min_child_weight': 0.001, 'n_estimators': 400, 'objective': 'binary'
XGB Classifier	0.787115	'learning_rate': 0.1, 'max_depth': 9, 'min_child_weight': 0.001, 'n_estimators': 400, 'objective': 'binary:logistic'

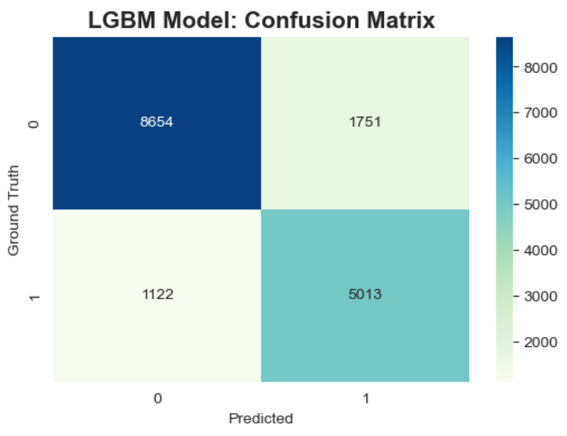
Using Undersampling Method

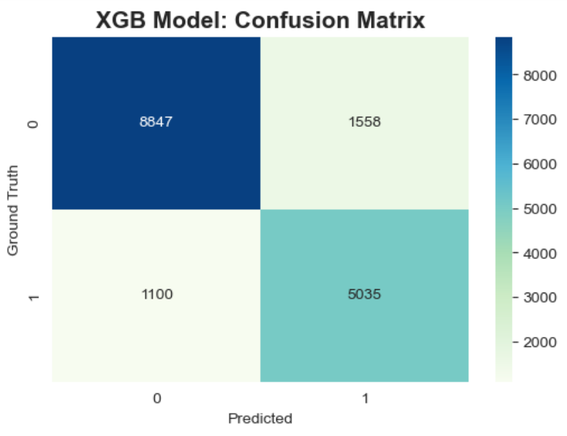
Grid Search CV = 5		
Model	**Best Scores**	**Best Parameters**
Random Forest	0.836103	'max_depth': 50, 'min_samples_split': 2, 'n_estimators': 300
LGBM Classifier	0.824227	'learning_rate': 0.5, 'max_depth': 8, 'min_child_weight': 1, 'n_estimators': 300, 'objective': 'binary'
XGB Classifier	0.831198	'learning_rate': 0.1, 'max_depth': 9, 'min_child_weight': 0.01, 'n_estimators': 400, 'objective': 'binary:logistic'

Hyperparameter Tuning Evaluation Comparison

Model	Precision Score_base	Precision Score_under	Recall Score_base	Recall Score_under	F1 Score_base	F1 Score_under	AUC Score_base	AUC Score_under
Random Forest Classifier	0.849585	0.775334	0.751263	0.821842	0.797405	0.797911	0.927440	0.924428
LGBM Classifier	0.816460	0.741130	0.751915	0.817115	0.782860	0.777270	0.914261	0.909910
XGB Classifier	0.840294	0.763689	0.744417	0.820701	0.789455	0.791169	0.922708	0.919413



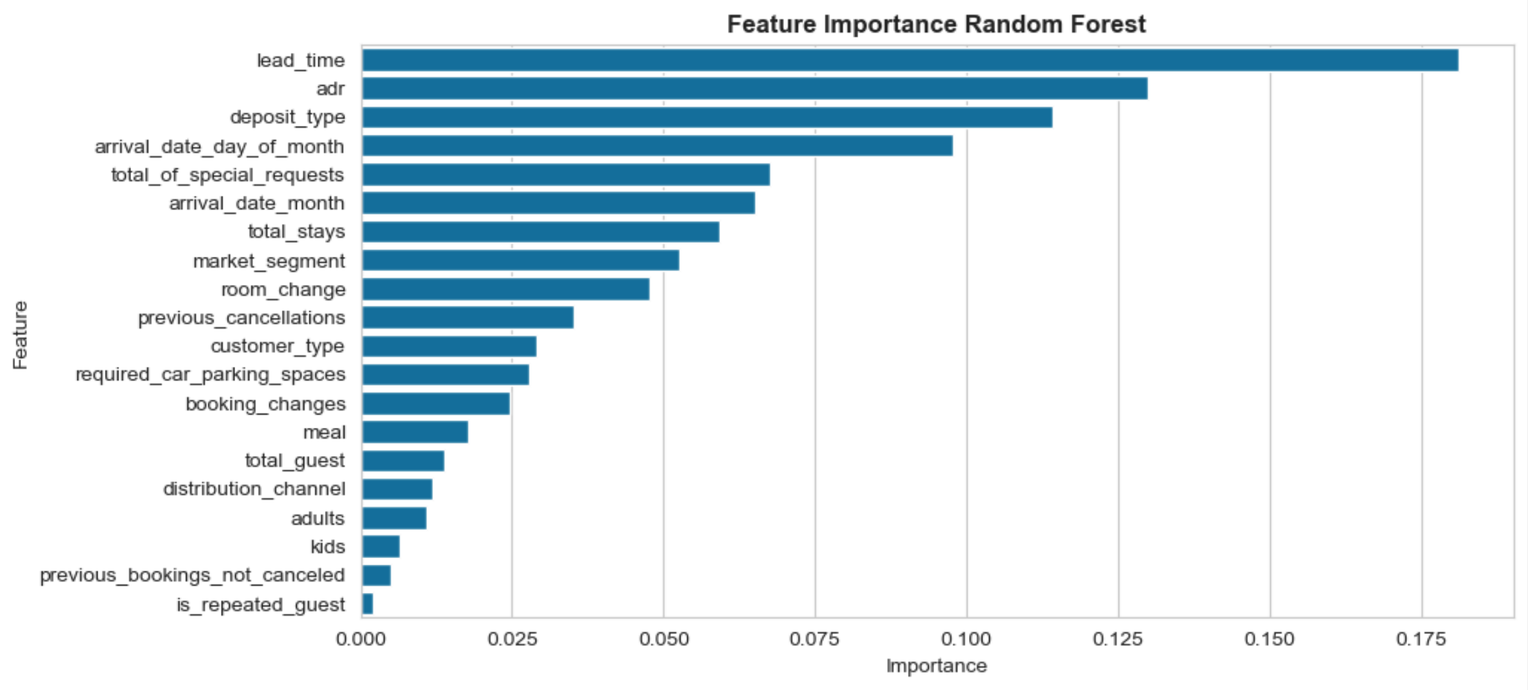




False Negative Rate (FNR): 0.1781 **False Negative Rate (FNR): 0.1828** **False Negative Rate (FNR): 0.1792**

After hyperparameter tuning, each of models has improved performance, it can be seen that recall and f1_score have increased and the best model is the **Random Forest model with undersampling**, where the **F1 Score is 0.7979** and has a smaller **False Negative rate** than other models, which is **0.1781**, meaning that the probability of a **prediction error is 17.81%**. With these results, it can be interpreted that if there are 100 who are predicted to cancel, 78 will really cancel **(Precision)**. 100 people who cancel, the model can predict 82 people who really cancel **(Recall)**.

Feature Importance



There are 3 importance features of the Random Forest model: lead_time, adr, deposit_type

- Through **lead_time**, it can be seen that the longer the guest booked from the check-in date, the higher the cancellation. Instead, guests who book close to the check-in date are less likely to cancel. This happens because guests who book close to the check-in date must be more confident to stay.
- Through **adr**, it can be seen that the higher the adr, the more guests cancel. However, at some point adr, the cancellation rate becomes fixed.
- Through **deposit_type**, it can be seen that the type of deposit type affects the cancel rate, especially for non-refund deposit types, there
  are fewer automatic orders because the booker avoids canceling the ticket and the money is not returned, while for non-deposit. it is hoped that the hotel will implement a DP system for bookers so that it can ensure that bookers are not just booking but really will stay.

Model Implementation

Scenario

- Assuming there are 5000 Hotel Reservations where 1000 have been cancelled
- If average cost per room per night is 102 euro (based on analysis)
  F1 Score our model is 0.80, with our model we can predict 800 hotel booking being canceled
- 102 x 800 = 81.600

We can minimize loss income by 81.600 euros

loss income from

102.000

loss income to

20.400

Business Recommendation

Business Recommendation

- Focus on managing reservations during peak season periods, by focusing on prospective customers who are likely to cancel
- Implement a deposit system (upfront payment) for customers who make early bookings so as to minimize the cancelation.
- Implement a stricter cancellation policy, where the cancellation policy determined can be differentiated between regular season and high season, by applying Down Payment for reservations and giving partial refunds when canceling in regular season to minimize losses. However, during the high season, cancellation of bookings is free of charge due to high demand, and if a cancellation occurs, it can be quickly replaced by another customer.
- Most customers reserve lower room categories. Therefore, hotels can reduce the price gap between the lower room category and the next room category. With a small price difference, customers can be compelled to pay a little more to book in a higher category.
- Hotels create a membership system by collecting points that can be redeemed for certain benefits, thus increasing customers to repeat bookings.
- Provide remainder to the customer if the time of arrival or stay is close so that the customer is aware of his order
- The hotel provides surveys to customers who stay overnight so that with this feedback the hotel can improve its services so that it can increase the number of customers making repeat bookings.


Thank You