**CAPSTONE PROJECT**

**DSE PROGRAM, GREAT LEARNING**



**PROJECT REPORT**

**ON**

**Hotel reservation cancellation prediction**

**Group – 3, DSE HYDERABAD, OCT – 2022 BATCH**

**TEAM MEMBERS**

**Khushbu Sharma**

**Ambati Venkata Bharath**

**Yashwanth**

**Geethu Krishna R**

**Veera Mahesh**

**Mohit Sharma**

**Manav Jalan**

**GROUP MENTOR**

**TRUPTI NAIK MATHAPATI**

**Guidelines for PGPDSE FT Capstone Project – Interim Report**

**Industry Review**

***Industry Review – Current practices, Background Research***

Matching the hotel's capacity with demand is one of the biggest challenges faced by hotel administrators. As stockpiling is not an option in the hospitality sector, hotels must manage demand and the restricted supply of rooms within a set time frame to avoid losing money on each vacant room. As a result, hotel owners try to boost revenue by maximising occupancy, which requires addressing future demand. Demand, however, is influenced by a number of outside factors, like the climate, political stability, intense competition, and others, making forecasting difficult. The task of organizing and arranging the hotel's capacity is made even more difficult by the possibility of cancelling already-made reservations.

The impact that cancellations have on hotel chains has resulted in the development of strategies designed specifically with the goal of reducing cancellations. One such strategy is overbooking, which consists in accepting more bookings than the hotel has the capacity for, relying on the fact that some cancellations will occur.

In order to prevent hotels from having idle capacity as much as possible, this strategy's primary goal is to match the number of customers who do not show up for the scheduled service (no-show), last-minute cancellations, and other cancellations that are informed in advance Hotels experience a loss of revenue if arrivals fall short of expectations due to unsold rooms.

In tourism and travel-related industries, most of the research on Revenue Management demand forecasting and prediction problems employ data from the aviation industry, in the format known as the Passenger Name Record (PNR). This is a format developed by the aviation industry.

Several managers found that applying machine learning in understanding customers’ behaviours and message customization leads to enhanced satisfaction and loyalty.

In this context, performing accurate forecasts is necessary for optimising operations, as well as supporting the decision-making process. Several managers found that applying machine learning in understanding customers’ behaviours and message customization leads to enhanced satisfaction and loyalty. Accurate forecasts help managers with medium- and long-term decisions not only for determining hotel polices, human resources required according to workloads or budget planning, but also for assisting in the development of short-term occupancy. Bearing in mind that forecasting models are based on historical data a number of investigations have encouraged management to consider the importance of having a reliable revenue management system through which past data can provide value to the organization. But cancellation models in the hospitality industry are a huge challenge because of the highly volatile and uncertain environment of this sector.

There has been a growing interest the development and application of machine learning models in this are to improve the customer experience and optimize hotel operations. current research in the hotel reservation area has focused on several key areas.

Online booking behaviour: Research has examined the behaviour of customers when booking hotel reservations online, including factors that influence their decision-making and the features they look for in a hotel.

* **Pricing strategies:** Research has explored various pricing strategies for hotel rooms, including dynamic pricing, which adjusts prices based on demand, and personalized pricing, which offers customized prices based on individual customer behaviour.
* **Recommender systems:** Research has focused on developing and improving recommender systems that suggest hotels to customers based on their preferences, search history, and previous bookings.
* **Customer feedback:** Research has examined customer feedback, including online reviews and social media comments, to identify areas for improvement in hotel services and operations.
* **Revenue management:** Research has explored various revenue management strategies, including forecasting demand, optimizing pricing, and managing inventory, to maximize hotel revenue.
* **Fraud detection:** Research has focused on developing fraud detection systems that use machine learning to identify and prevent fraudulent transactions in hotel reservations.

The 4- Star hotels in Portugal were able to save revenue of approximately € 39,000.00 using machine learning predicting techniques for cancellation.

In China, the machine learning-enabled, WeChat platform, cloud service platform, PMS, face/fingerprint device, networked door lock, room power controller, room guidance system, public area control system, room intelligent control system helped to save marketing cost of 115,000 Yuan

What are the machine learning techniques used in the hotel industry?

There are number of machine learning techniques and algorithms that have been used in the

hotel industry. Researchers have used XGBoost, Naive Bayes Classifier (NB), Generalized

Linear Model (GLM), Multinomial Naive Bayes (MNB), Extreme learning machine (ELM),

support vector regression (SVR), Boosted Regression Tree (BRT), Random Forest

Regression (RFR), Natural language processing (NLP), Convolutional Neural Network-based

Deep Learning (CNN-DL), and Nearest Neighbour.

**Literature Survey - Publications, Applications, past and undergoing research**

Bottom of FormThere have been several publications on hotel reservation machine learning models in recent years. Here are some examples:

A comparison of collaborative Filtering Algorithms for Hotel Recommendation “by Qiao et al. (2018): This paper compares different collaborative filtering algorithms for hotel recommendation and evaluates their performance using real – world hotel reservation data.

Deep Learning – Based Recommendation system using Implicit Feedback ‘by Lee et al. (2020): This paper proposes a hotel recommendation system based on deep learning models that can make recommendations using implicit feedback such as click – through rates and purchase history.

Predicting hotel Reservation cancellations using Machine Learning Techniques ‘by Zhang et al. (2019): This paper presents a machine learning approach to predicting hotel reservation cancellations using historical data and various features such as booking channel and room type.

Optimization of Hotel Room Prices Based on Machine Learning Methods ‘by xu et al. (2019): This paper proposes a machine learning approach to optimizing hotel room prices by analysing various factors such as seasonal demand and customer behaviour.

Machine Learning – Based Fraud Detection for online Hotel Reservations by wang et al. (2020): This paper presents a machine learning – based fraud detection system that can identify fraudulent hotel reservation transactions by analysing patterns in booking behaviour.

These publications demonstrate the wide range of applications for machine learning in hotel reservation systems, including recommender systems, price optimization, fraud detection, and more.

Hotel reservation machine learning models have been active area for several years now, and many on going research projects are currently exploring various applications of machine learning in this field. Here are some examples of both past and ongoing research in hotel reservation learning models:

**Recommender systems:** Machine learning algorithms are used to build recommender systems that suggest hotels to customers based on their preferences, search history, and previous bookings. ongoing research aims to improve the accuracy of these systems by incorporating new data sources and developing more advanced algorithms.

**Price optimization:** Machine learning be used to optimize hotel price strategies by analysing historical data and predicting demand patterns. Ongoing research in this area is exploring new approaches to pricing optimization, such as dynamic pricing and personalized pricing based on individual customer behaviour.

**Fraud detection:** Machine learning algorithms are used to detect and prevent fraud in hotel reservations by analysing patterns in booking behaviour and identifying anomalies. Ongoing research is focused on improving the accuracy of fraud detection systems and developing new techniques for identifying fraudulent transactions.

**Personalized customer service:** Machine learning models can be used to analyse customer data and provide personalized recommendations for dining, activities, and other services. Ongoing research in this area is focused on developing more advanced personalization algorithms that can better understand individual customer preferences and behaviour.

**Sentiment analysis:** Machine learning algorithms can be used customer reviews and feedback to understand customer sentiment and identify areas for improvement in hotel services. Ongoing research in this area is exploring new techniques for sentiment analysis and developing more accurate models for identifying customer sentiment.

**Forecasting demand:** Machine learning models can be used to forecast demand for hotel rooms based on historical data, seasonality, and other factors. Ongoing research is focused on developing more accurate forecasting models that can better predict demand patterns and improve revenue management for hotels.

**Image Recognition:** Machine learning models can be used to analyse hotel images and identify key features such as room type, amenities, and decor, making it easier for customers to find the right hotel. Ongoing research in this area is focused on developing more advanced image recognition algorithms that can accurately identify and classify hotel features.

Overall, ongoing research in hotel reservation machine learning models id focused on developing more advanced algorithms and techniques to improve customer experience, increase revenue, and optimize hotel operation.

**Dataset and Domain**

**Data Dictionary:**

### 1.Hotel Type Subgroups: Resort Hotel, City Hotel (Categorical)

### 2.is\_canceled (Target Variable) The value indicating if the booking was cancelled Subgroups: (1) is cancelled, (0) is not cancelled (Categorical)

### 3.lead\_time

Number of days that elapsed between the entering date of the booking into the PMS and the arrival date (Integer)

### 4.arrival\_date\_year

Year of arrival date - (Integer)

### 5.arrival\_date\_month

Month of arrival date with 12 categories  
Subgroups: “January” to “December” (Categorical)

### 6.arrival\_date\_week\_number

Week number of the arrival date - (Integer)

### 7.arrival\_date\_day\_of\_month

Day of the month of the arrival date - (Integer)

### 8.stays\_in\_weekend\_nights

Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel - (Integer)

### 9.stays\_in\_week\_nights

Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel - (Integer)

### 10.adults

Number of adults - (Integer)

### 11.children

Number of children - (Integer)

### 12.babies

Number of babies - (Integer)

### 13.meal

Type of meal booked. Categories are presented in standard hospitality meal packages  
Subgroups: Undefined/SC – no meal package, BB – Bed & Breakfast, HB – Half board (breakfast and one other meal – usually dinner), FB – Full board (breakfast, lunch and dinner) (Categorical)

### 14.country

Country of origin. Categories are represented in the ISO 3155–3:2013 format - (Categorical)

### 15.market\_segment

Market segment designation Subgroups: “TA” = “Travel Agents” “TO” = “Tour Operators” (Categorical)

Travel agents help individuals and groups plan and book their travel arrangements, while tour operators design and operate group tours and vacation packages.

### 16.distribution\_channel

Booking distribution channel Subgroups: “TA” = “Travel Agents” “TO” means “Tour Operators” (Categorical)

### 17.is\_repeated\_guest

Value indicating if the booking name was from a repeated guest Subgroups: (1) is repeated, (0) is not repeated (Categorical)

### 18.previous\_cancellations

Number of previous bookings that were cancelled by the customer prior to the current booking - (Integer)

### 19.previous\_bookings\_not\_canceled

Number of previous bookings not cancelled by the customer prior to the current booking - (Integer)

### 20.reserved\_room\_type

Code of room type reserved Code is presented instead of designation for anonymity reasons (Categorical)

### 21.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 to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons  
(Categorical)

### 22.booking\_changes

Number of changes/amendments made to the booking from the moment the booking was entered on the PMS (PMS stands for Property Management System. In the hotel industry, a PMS is a software application that enables hotels to manage their day-to-day operations, including managing reservations, check-ins and check-outs, guest information, room allocations, billing, housekeeping, and more. The PMS serves as the central hub for managing all the hotel's operational tasks, and it integrates with other systems such as online booking engines, point of sale systems, and revenue management systems.) until the moment of check-in or cancellation - (Integer)

### 23.deposit\_type

Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories’ Subgroups: No Deposit – no deposit was made, NonRefund – a deposit was made in the value of the total stay cost, Refundable – a deposit was made with a value under the total cost of stay (Categorical)

### 24.agent

ID of the travel agency that made the booking - (Categorical)

### 25.company

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

#### Notes: ID is presented instead of designation for anonymity reasons

(Categorical)

### 26.days\_in\_waiting\_list

Number of days the booking was in the waiting list before it was confirmed to the customer - (Integer)

### 27.customer\_type

Type of booking, assuming one of four categories’ Subgroups: Contract - when the booking has an allotment or other type of contract associated to it, Group – when the booking is associated to a group, Transient – when the booking is not part of a group or contract, and is not associated to other transient booking, Transient-party – when the booking is transient, but is associated to at least other transient booking (Categorical)

### 28.adr

Average Daily Rate Average Daily Rate (ADR) is a key performance metric used in the hotel industry to measure the average price that a hotel room is sold for per day. ADR is calculated by dividing the total room revenue generated by the number of rooms sold during a specific period of time, typically a day or a month. (Numeric)

### 29.required\_car\_parking\_spaces

Number of car parking spaces required by the customer - (Integer)

### 30.total\_of\_special\_requests

Number of special requests made by the customer (e.g. twin bed or high floor) - (Integer)

### 31.reservation\_status

Reservation last status, assuming one of three categories’ Subgroups: Cancelled – booking was cancelled by the customer, Check-Out – customer has checked in but already departed, No-Show – customer did not check-in and did inform the hotel of the reason why (Categorical)

### 32.reservation\_status\_date

Date at which the last status was set  
This variable can be used in conjunction with the Reservation Status to understand when was the booking cancelled or when did the customer checked-out of the hotel (Date)

***Variable categorization (count of numeric and categorical):***

* Count of Numerical Variables:
  1. Integer: 16
  2. Float: 4
* Count of Categorical Variables:
  1. Object: 12

**Preprocessing Data Analysis (count of missing/ null values, redundant columns, etc.):**

In a dataset with 1,19,390 records and 32 variables there were a total of 4 null values in children, 488 in country, 16340 in agent and 1,12,593 in company variable.

We are dropping, arrival\_date\_week\_number, arrival\_date\_day\_of\_month from the dataset because these columns are insignificant in the dataset according to the statistical tests performed with the target variable.

**Alternate sources of data that can supplement the core dataset (at least 2-3 columns):**

<https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset>

Online hotel reservation channels have dramatically changed booking possibilities and customers’ behaviour. A significant number of hotel reservations are called-off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Can you predict if the customer is going to honour the reservation or cancel it?

**Project Justification - Project Statement, Complexity involved, Project Outcome –Commercial, Academic or Social value:**

**Project Statement:**

This data collection comprises reservation details for a city hotel and a resort hotel, as well as details like the date the reservation was made, the duration of the stay, the number of adults, kids, and/or babies, and the number of parking spaces that are available. Due to changes in plans or no-shows, many hotel reservations are canceled. Changes in plans, scheduling issues, and other common causes of cancellations are listed below. The task is to predict whether the customer is going to honor the reservation or not.

**Complexity Involved:**

Hotel reservation cancellation prediction using machine learning involves several complexities. Here are some of the key factors to consider:

Data Collection: One of the biggest challenges in building a cancellation prediction model is collecting and preparing the data. Relevant data may include reservation details (e.g. check-in/out dates, room type, number of guests), customer demographics, booking history, and cancellation patterns.

Feature Selection: Once the data has been collected, it's important to select the most relevant features for the model. This may involve using domain knowledge or machine learning algorithms to identify the most important predictors of cancellation.

Model Selection: There are various machine learning algorithms that can be used to build a cancellation prediction model, such as decision trees, logistic regression, and neural networks. Choosing the right model depends on factors such as the size of the dataset, the complexity of the problem, and the desired level of accuracy.

Interpretation: After the model has been trained, it's important to interpret the results to gain insights into the factors that are driving cancellations. This may involve analyzing feature importance scores, generating visualizations, and conducting further analysis to identify trends and patterns.

**Commercial value:**

The 4- Star hotels in Portugal were able to save revenue of approximately € 39,000.00 using machine learning predicting techniques for cancellation (Antonio, et al., 2019).

In China, the machine learning-enabled WeChat platform, cloud service platform, PMS, face/fingerprint device, networked door lock, room power controller, room guidance system, public area control system, room intelligent control system, and service robot helped to save marketing cost of 115,000 Yuan.

**Data Exploration (EDA)**

**Relationship between variables:**

Check for

* Univariate Analysis
* Bi Variate Analysis
* presence of outliers and its treatment
* statistical significance of variables

**Feature Engineering**

* Whether any transformations required
* Scaling the data

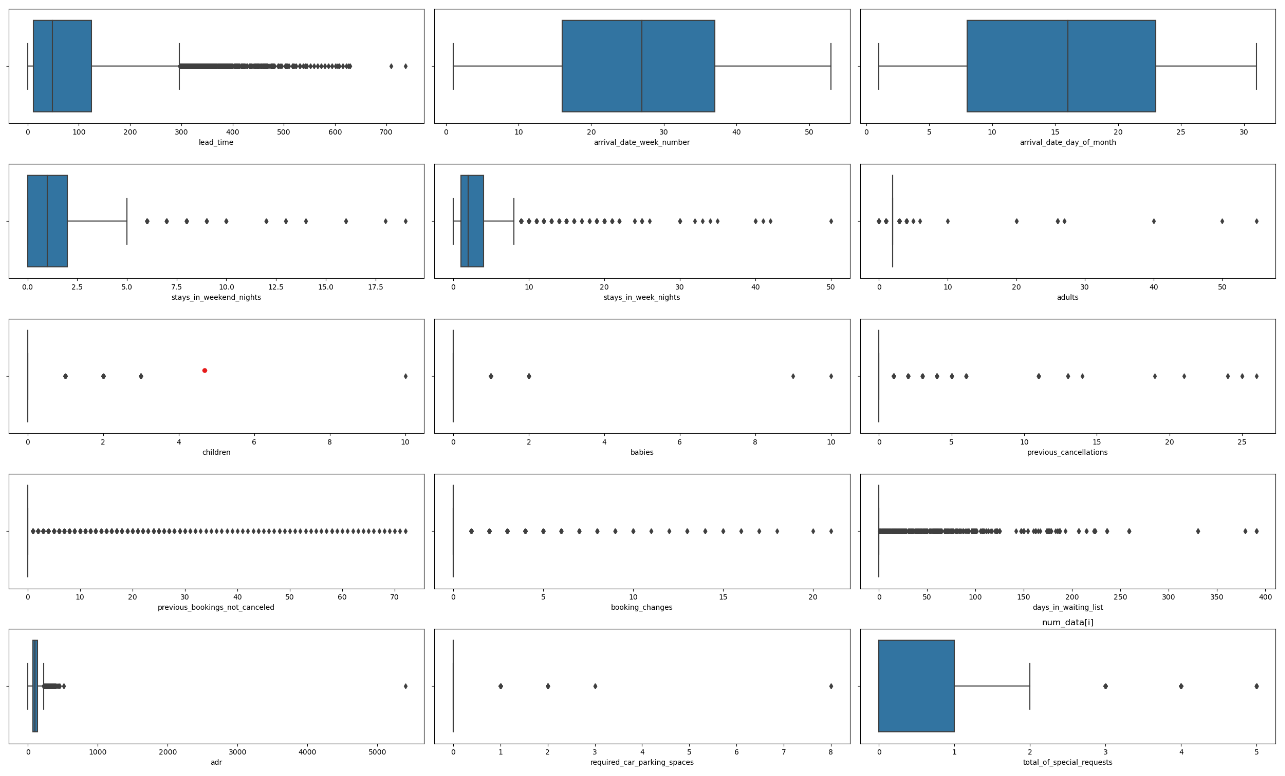
**Base Model**

Create a base model and evaluate the results

Explain in detail about further work.

**----------------------------- Interim Presentation Checkpoint----------------------------------------------------------**

**Univariate Analysis**

****

From the above box plots (apart from finding out the outliers) we find that:

1. **Median of lead time column is somewhere close to 40 days**

* The lead time column likely contains a set of numerical values that represent the time it takes for a product or service to be delivered or completed.
* The median is a statistical measure of central tendency that represents the middle value of a dataset. In other words, it's the value that's exactly halfway between the lowest and highest values in the dataset.
* Based on the statement, it suggests that the median value of the lead time column is around 40 days, indicating that half of the lead times are shorter than 40 days, and the other half are longer than 40 days.

1. **A lot of bookings are made from week numbers 17 to 38 which is understandable as these week numbers correspond to the start of Spring and end of Summer, respectively**

* Week numbers 17 to 38 are typically between late April to late September in the northern hemisphere, which is the spring and summer season. During this period, the weather is generally warmer, and people tend to take vacations, travel, and engage in outdoor activities.
* It’s common for businesses in the tourism and hospitality industry to experience a higher volume of bookings during this period due to the increased demand for travel and leisure activities.
* Therefore, it’s understandable that a lot of bookings are made during week numbers 17 to 38 as people tend to plan and book their spring and summer trips during this time, which can lead to increased revenue for businesses in the tourism and hospitality industry.

1. **Most bookings are neither made at the beginning nor at the end of the month**

* The statement suggests that there may be a pattern in the booking behaviour of customers where they are less likely to make bookings at the beginning or end of the month.
* This pattern could be due to a variety of factors, such as pay cycles, bill payments, or general monthly budgeting. For example, people may be more likely to make bookings around the middle of the month when they have a better understanding of their monthly finances.
* From a business perspective, it’s importing to understand these patterns in

Customer behaviour to optimize marketing and sales strategies, such as adjusting pricing or offering promotions at times when customers are more likely to book.

#### **There have been instances where a booking has comprised of more than 50 adults (this is probably a corporate event)**

* Large Group Bookings: There have been instances where a booking has consisted of more than 50 adults, indicating that the event is likely a corporate gathering or conference.
* Corporate Events: Such large group bookings are often made for corporate events where businesses bring together employees or clients for meetings, training sessions, or team building activities.
* Organized Planning: Organizing such events requires careful planning, communication, and coordination with the venue or hotel to ensure that all the requirements of the group are met. This includes arranging for meeting rooms, accommodation, meals, transportation, and other logistical needs.

1. **Most tourists do not have children accompanying them but there have been instances where a booking comprised of 10 children**

* The statement suggests that tourists typically do not have children accompanying them when they travel. This may be due to a variety of factors, such as personal preferences or the cost of traveling with children.
* However, there have been instances where a booking comprised of 10 children, indicating that some families do travel with larger groups of children. This could be due to various reasons, such as family reunions or vacations.
* From a business perspective, it's important to understand the demographics and preferences of customers to optimize marketing and sales strategies. For example, if a business notices a trend of larger groups of families traveling with children, they may want to consider offering group packages or family-friendly activities to cater to this market segment.

#### **Most tourists do not have babies accompanying them but there have been instances where a booking comprised of 10 babies**

* The statement suggests that tourists typically do not have babies accompanying them when they travel. This may be due to a variety of factors, such as the cost and logistical challenges of traveling with infants.
* However, there have been instances where a booking comprised of 10 babies, indicating that some families do travel with larger groups of infants. This could be due to various reasons, such as family reunions or events where multiple families with young children travel together.
* From a business perspective, it’s important to understand the demographics and preferences of customers to optimize marketing and sales strategies. For example, if a business notices a trend of larger groups of families traveling with infants, they may want to consider offering baby-friendly facilities or services to cater to this market segment. It's also important for businesses to have policies and procedures in place to ensure the safety and comfort of both the infants and other guests.

#### **7. There have been instances when a customer has cancelled more than 25 bookings, in the past**

* Frequent Cancellations: There have been instances where a customer has cancelled more than 25 bookings in the past.
* Reasons for Cancellations: Cancellations can occur for various reasons, such as changes in travel plans, unforeseen circumstances, or simply a change of mind.
* Impact on Establishment: Frequent cancellations can have an impact on the establishment, including loss of revenue, inconvenience to other guests, and difficulty in managing inventory and staffing. It is important for hotels and other establishments to have clear cancellation policies in place and to communicate them effectively to customers to minimize the impact of cancellations.

#### **8. There are customers which have made more than 20 changes to the booking**

* Frequent Changes: Some customers may make multiple changes to their bookings, with some making more than 20 changes.
* Flexibility: These changes could be due to a variety of reasons, such as changes in travel plans, preferences, or unforeseen circumstances.
* Customer Service: It is important for hotels and other establishments to be flexible and accommodating to these changes to ensure customer satisfaction.
* Communication: Proper communication and clear policies can help manage these changes, including any associated fees or restrictions, and provide customers with a smooth and hassle-free experience.

#### **9. Though most of the bookings in the waiting list have been cleared in a day, there have been instances where a booking was in the waiting list for close to 400 days!**

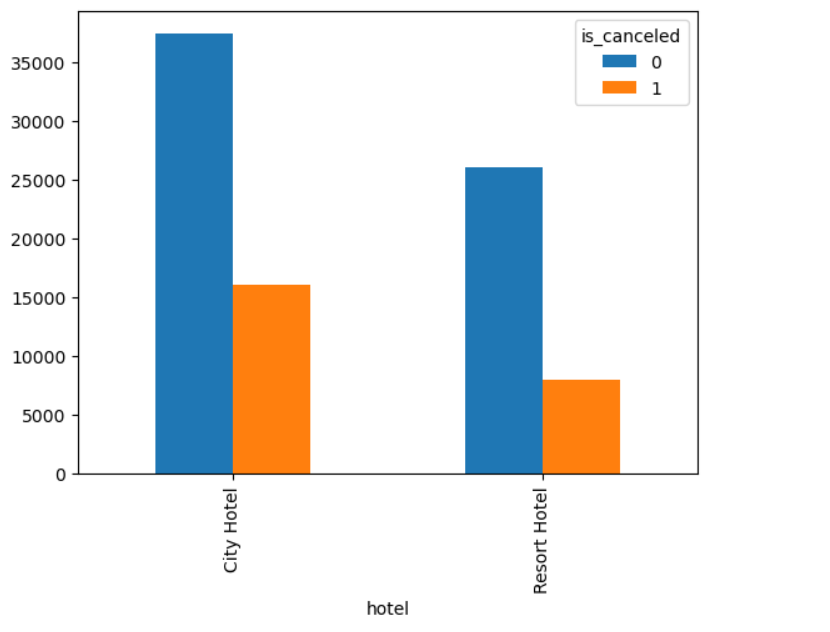
* Clearing Waiting List Bookings: Most hotels or establishments aim to clear their waiting list bookings within a day or two.
* Special Circumstances: However, there may be instances where a booking is on the waiting list for an extended period of time. This can occur due to special circumstances such as a high demand for the establishment, a limited number of rooms, or a particular event happening in the area.
* Long Waiting Periods: In rare cases, a booking may be in the waiting list for several months, or even close to a year, before being cleared.
* Customer Service: In such situations, it is important for hotels to provide excellent customer service and maintain clear communication with the customer about their booking status.
* Managing Expectations: Hotels may also need to manage the expectations of customers who are on the waiting list, informing them of the likelihood of their booking being cleared and providing alternatives if necessary.

#### **10 Most customers at the most make only 1 special request**

* Single Special Request: The majority of customers only make one special request during their stay or visit.
* Specific Requests: Special requests may include anything from requesting a specific room location to dietary restrictions or special amenities.
* Importance of Meeting Requests: As a result, it is important for hotels and other establishments to meet these requests to ensure customer satisfaction and build loyalty. Proper communication and attention to detail can go a long way in making a guest's stay a memorable one.

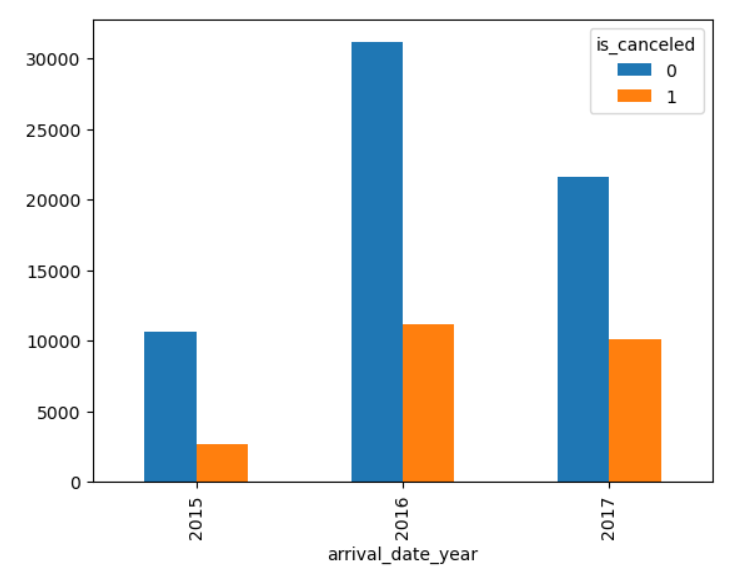
**Bivariant Analysis**

1. **City hotel getting more cancellations in number of customers when compared to resort hotel cancellations**



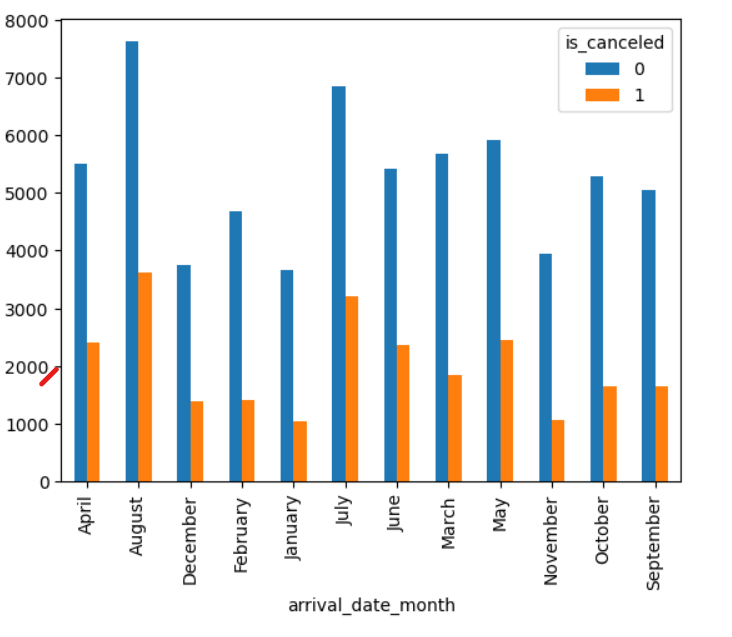
* City hotels typically cater to business travellers who often have last-minute changes in their plans, leading to higher cancellation rates.
* Resort hotels, on the other hand, primarily cater to leisure travellers who often book well in advance and have more predictable travel plans, resulting in lower cancellation rates.
* City hotels may have more strict cancellation policies due to higher demand and limited availability, which could dissuade some customers from making reservations in the first place.
* Resort hotels may offer more flexible cancellations policies, such as allowing cancellations up to a few days before arrival, which could encourage customers to book and then follow through with their plans.
* Overall, the difference in cancellation rates between city hotels and resort hotels is likely due to the different types of customers they attract and the different policies they implement

#### **2. We can see in arrival\_date\_year there are 2015,2016,2017. 2016 and 2017 had more number of bookings compared to 2015 and thus the number of cancellations were also high.**



* The data includes arrival dates for the years 2015,2016 and 2017.
* The number of bookings in 2016 and 2017 was higher than in 2015, indicating increased demand for hotel stays during those years.
* The higher number of bookings in 2016 and 2017 also led to a high number of cancellations, as some customers may have changed their plans or been unable to travel.
* Among the months included the data, August, July, and May had a higher number of cancellations compared to other months.
* The reasons for their higher number of cancellations during these months could be due to factors such as summer vacations, peak travel periods, or weather-related events.

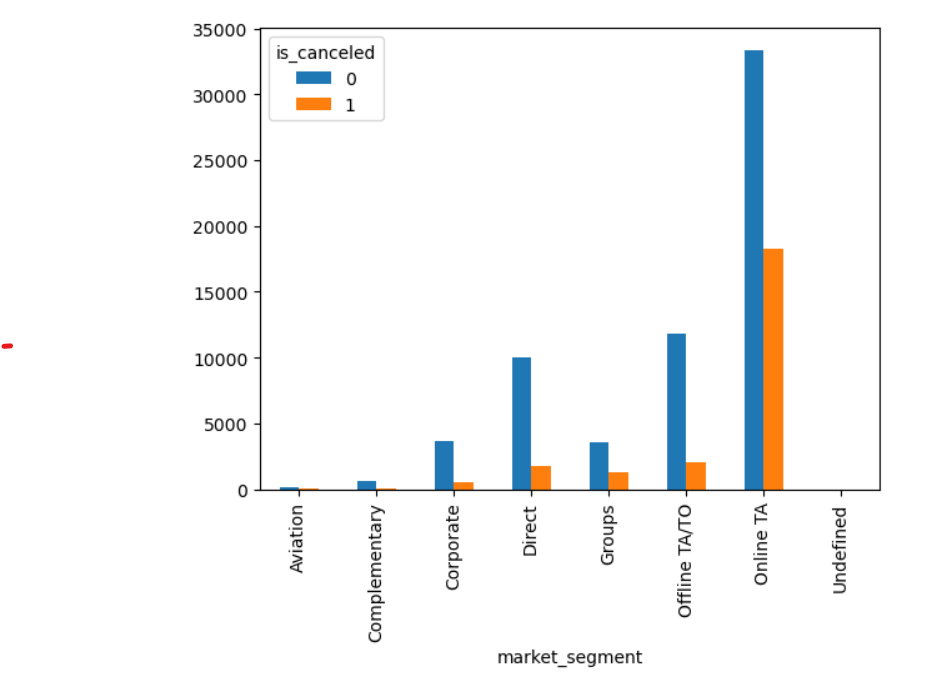
#### **3.August, July, may have high cancellations**



* The data analysis shows that the months of august, July, and may have higher cancellations
* Rates compared to other months.
* This could be due to various factors such as summer vacations, peak travel, periods or weather – related events.
* The higher cancellations rates during these months could result in a loss of revenue for hotels, as they may have to refund deposits or forgo potential bookings.

#### **4.Most of them have chosen online TA to be their market segment thus maximum**

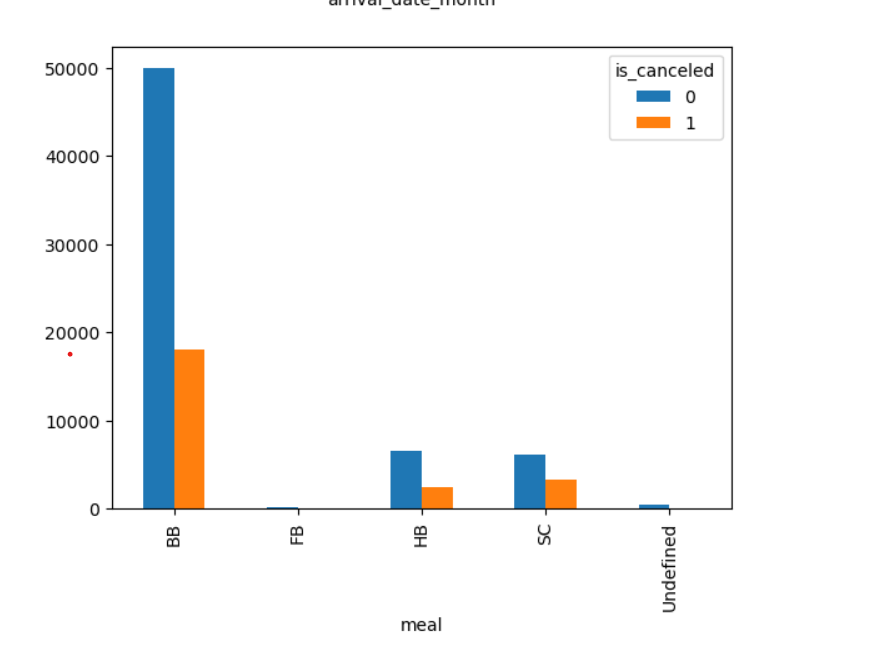
#### **cancellations are coming from there**



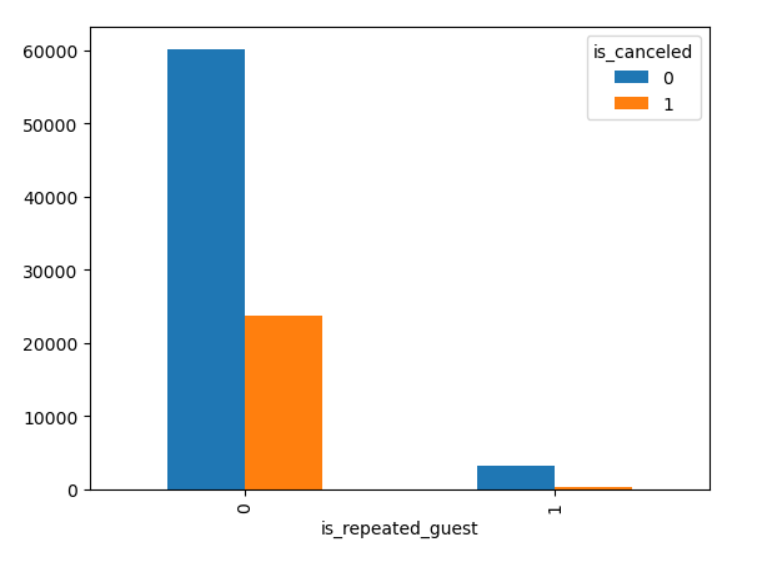
* Majority of the guests are not paying and deposit amount to the hotel. This is another reason for the cancellation to be high
* The data analysis shows that majority of guests are not paying any deposit amount to the hotel at the time of booking.
* This means that guests are not financially committed to their reservations and may be more likely to cancel.
* Without a deposit, there may be less incentive for guests to follow through with their plans, particularly if their travel plans change or they find a better deal elsewhere.
* Hotels may consider implementing a deposit policy to reduce the number of cancellations and ensure that guests are committed to their reservations.
* However, requiring a deposit may also deter some potential customers from making a reservation in the first place, so it is important to find a balance between minimizing cancellations and maintaining a competitive pricing strategy.
* Overall, the lack of a deposit requirement could be contributing to the higher cancellation rates observed in the data, and hotels may need to consider strategies to address this issue.

#### **5.BB – Bed & Breakfast has the high demand, HB – Half board has the 2nd high demand**

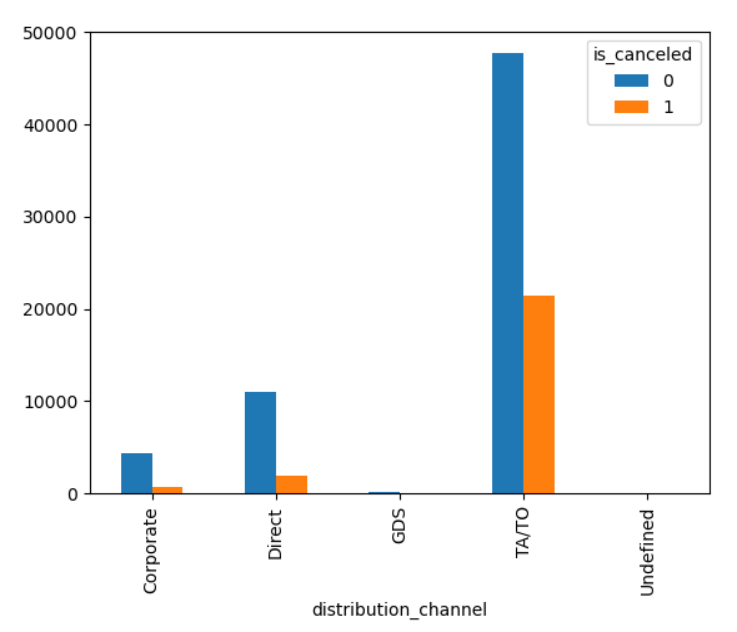
#### **FB – Full board has very less demand, Undefined/SC – no meal package**



* The data analysis indicates that Bed & Breakfast (BB) is the most popular meal package among hotel guests.
* Half board (HB), which includes breakfast and either lunch or dinner, is the second most popular meal package.
* Full board (FB), which includes all three meals, has very low demand among guests.
* Some guests may prefer the flexibility of not having a meal package, and therefore choose an undefined/SC (room only) option.
* The popularity of different meal packages may depend on factors such as the type of hotel, location, and target market

**6. Here we can compare is\_repeated\_guest and is canceled columns** The "is\_repeated\_guest" column has two possible values: 0 and 1. The value 0 appears much more frequently in the data, with 115,096 occurrences, while the value 1 appears only 3,806 times.

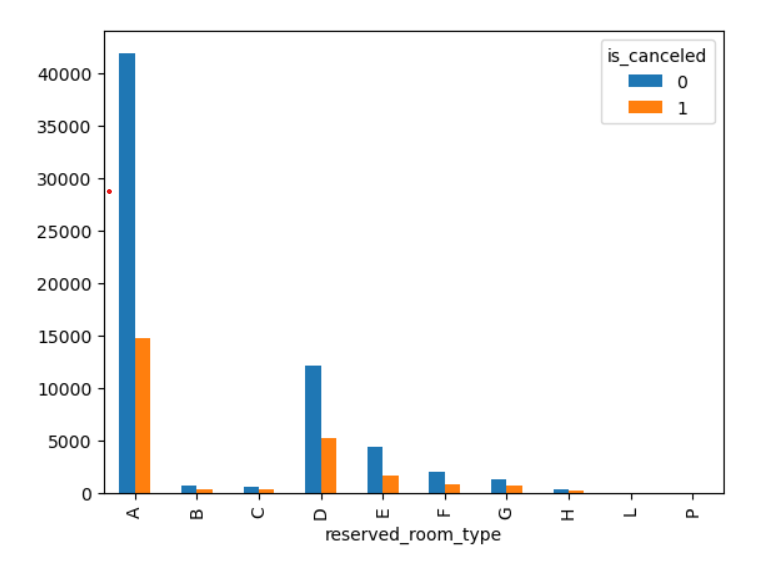
The "is\_canceled" column also has two possible values: 0 and 1. By comparing the two columns, we can see that guests who are not repeated guests (is\_repeated\_guest=0) are more likely to have their reservations canceled (is\_canceled =1) than guests who have stayed at the property before (is\_repeated\_guest=1).

**7.Most of them have chosen online TA to be their market segment thus maximum cancellations are coming from there**.

* The data analysis suggests that many customers have booked their hotel stays through online travel agencies (OTA).
* OTA customers may be more likely to cancel their reservations due to the convenience of booking online, which can result in overbooking or double-booking of rooms.
* Hotels may need to monitor the cancellation rates for OTA bookings closely and implement strategies such as dynamic pricing, overbooking management, or flexible cancellation policies to manage this segment effectively.

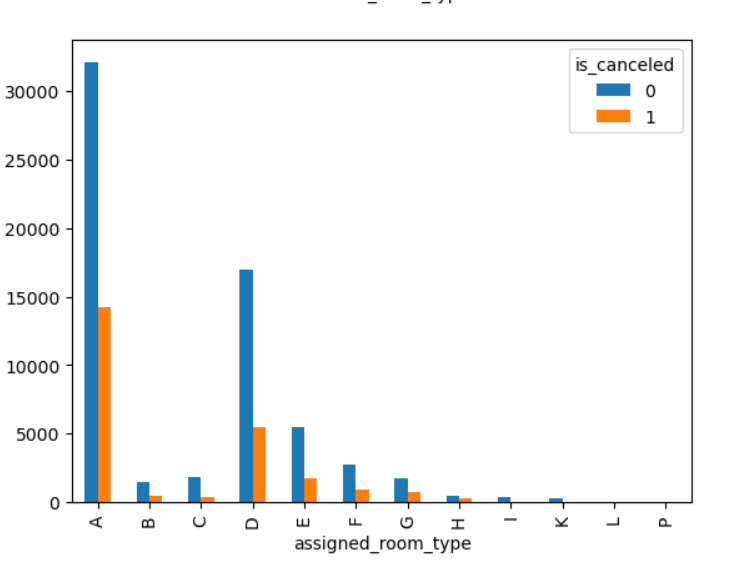
**8.Here we can compare reserved\_room\_type and is canceled columns the top 4 is**

**A, D, E, F**



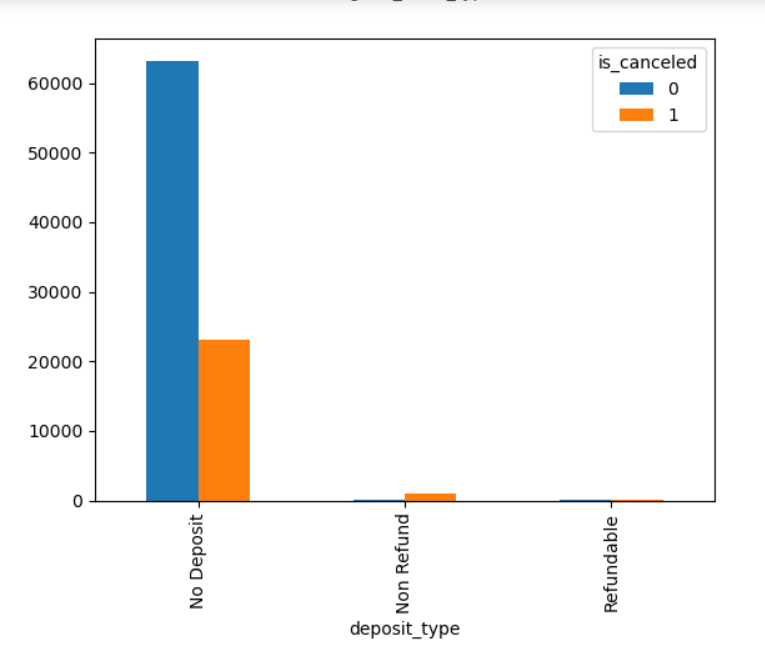
* The "reserved\_room\_type" column contains several different room types, with the most frequently reserved type being type A, which appears 85,601 times in the data.
* The "is\_canceled" column has two possible values: 0 and 1. By comparing the two columns, we can see that reservations for room types A, D, E, and F have all been cancelled at least once.
* Of the top four most frequently reserved room types (A, D, E, and F), type A has the highest number of cancellations, with 12,791 of its 85,601 reservations being cancelled.
* By contrast, type D has a lower cancellation rate, with only 2,406 of its 19,173 reservations being cancelled.
* This suggests that the cancellation rate may vary depending on the type of room reserved, and further analysis could be done to determine whether there are any factors that make certain room types more likely to be cancelled than others.

**9.Here we can compare reserved\_room\_type and is canceled columns the top 4 is**

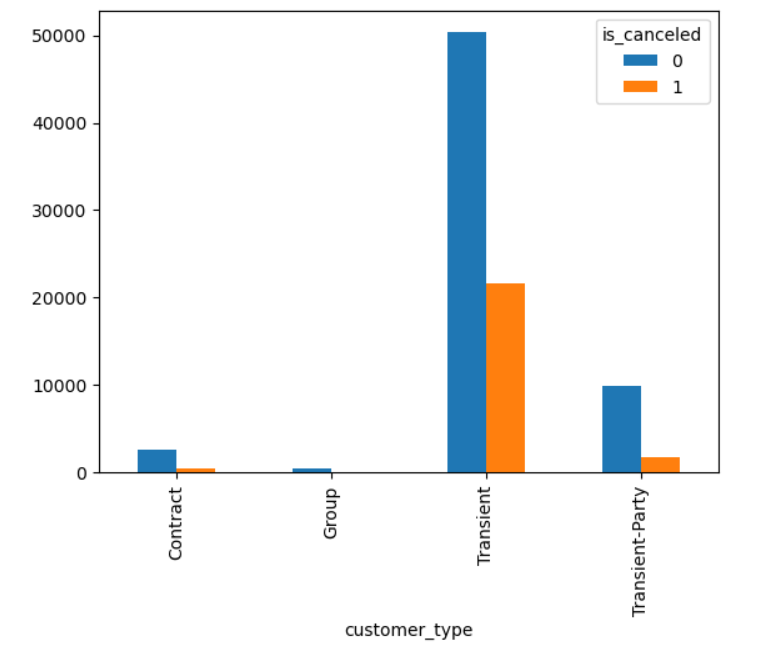
**A,D,E, F**

* The "assigned\_room\_type" column contains several different room types, with the most frequently assigned type being type A, which appears 73,863 times in the data.
* The "is\_canceled" column has two possible values: 0 and 1. By comparing the two columns, we can see that reservations for room types A, D, E, and F have all been canceled at least once.
* Of the top four most frequently assigned room types (A, D, E, and F), type A has the highest number of cancellations, with 10,347 of its 73,863 reservations being canceled.
* By contrast, type D has a lower cancellation rate, with only 2,720 of its 25,166 reservations being canceled.
* This suggests that the cancellation rate may vary depending on the type of room assigned, and further analysis could be done to determine whether there are any factors that make certain room types more likely to be canceled than others.
* It's worth noting that the number of reservations for each room type is not exactly the same in the "reserved\_room\_type" and "assigned\_room\_type" columns, which could affect the analysis of cancellation rates.

**10.Here we can compare reserved\_room\_type and is canceled columns the No Deposit is getting more cancellations than NonRefund, Refundable, No Deposit, NonRefund, Refundable**



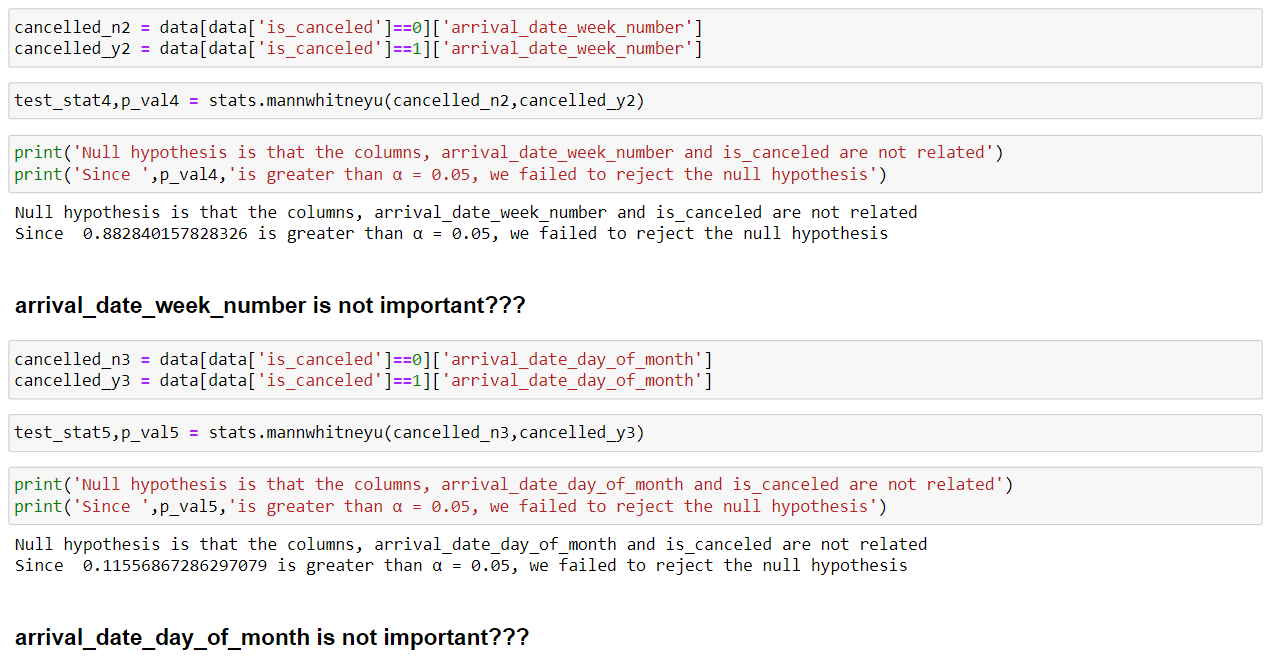
* The "deposit type" column contains three different types of deposits: No Deposit, Non-Refund, and Refundable.
* The "is\_canceled" column has two possible values: 0 and 1. By comparing the two columns, we can see that reservations made with No Deposit have the highest number of cancellations, with 44,809 of the 104,167 No Deposit reservations being cancelled.
* By contrast, reservations made with non-Refund deposits have a lower cancellation rate, with only 14,100 of the 14,573 Non refund reservations being cancelled. Similarly, reservations made with Refundable deposits have an even lower cancellation rate, with only 71 of the 162 Refundable reservations being cancelled.
* This suggests that the type of deposit made at the time of reservation may be an important factor in predicting whether a reservation will be cancelled, with No Deposit reservations being more likely to be cancelled than those made with Non refund or Refundable deposits. Further analysis could be done to determine the reasons behind this pattern.

**11.Most of the people who are visiting the hotel are transient type of customers so the cancellations happening are also more in number there**. 

* The data analysis indicates that most of the customers visiting the hotel are of the transient type.
* Transient customers typically stay for shorter periods of time and may be more likely to cancel their reservations.
* As a result, the number of cancellations is also higher among transient customers compared to other types of customers such as group or contract customers.
* Hotels may need to implement strategies to manage the higher cancellation rates among transient customers, such as offering flexible cancellation policies or overbooking rooms to account for potential cancellations.
* Understanding the booking and cancellation patterns of different types of customers can help hotels make informed decisions about pricing, inventory management, and revenue optimization.
* Overall, the high number of cancellations among transient customers is a common challenge for hotels and must be managed effectively to minimize the impact on

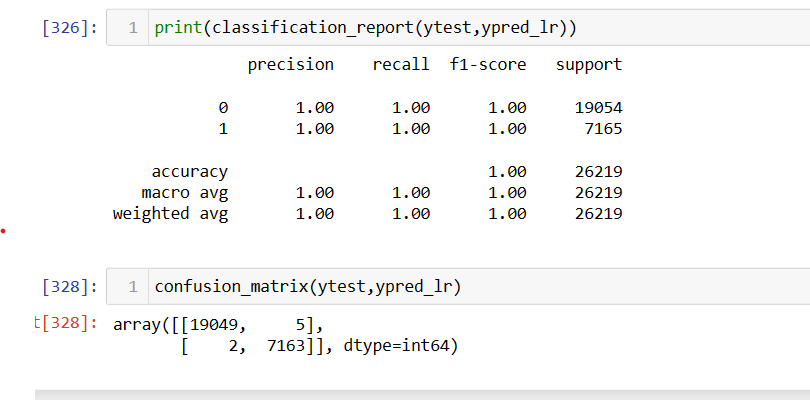
**Statistical Tests**

* The target variable being a categorical variable, ManWhitneyU for Numerical vs Categorical and Chi2\_square for Categorical vs Categorical were done in order to analyse the significance of the variable.
* From the above we could draw conclusions that the variables arrival\_date\_week\_number and arrival\_date\_day\_of\_month are least significant with respect to our target variable both being numerical columns.



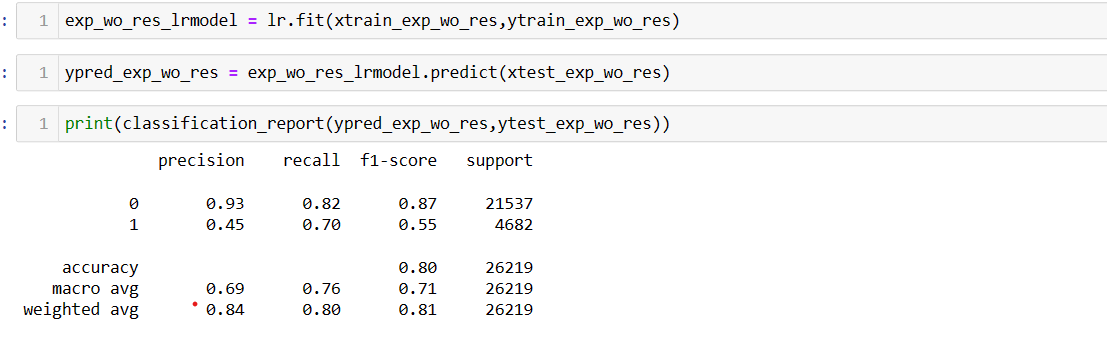
Here we can see every column is significant to is\_cancelled which is our target variable.

**Base model**



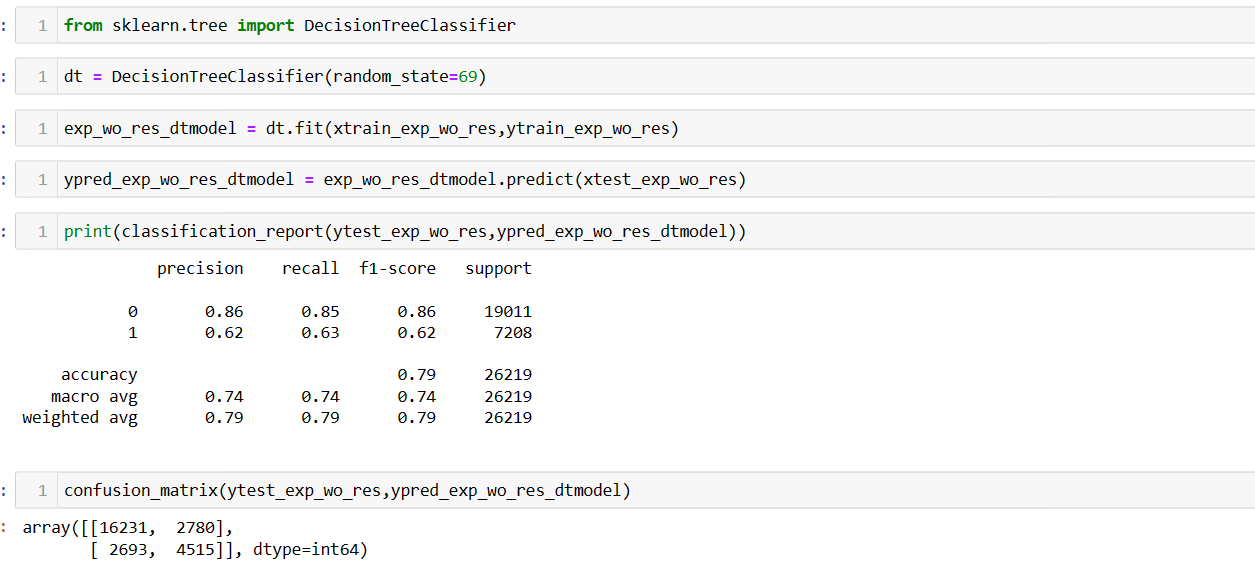
Here The Logistic Regression (base model) appears to have performed extremely well, achieving a precision, recall, and F1-score of 1.00 for both the positive (1) and negative (0) classes. This indicates that the model was able to accurately classify both types of instances with no false positives or false negatives. The overall accuracy of the model is also 1.00, meaning that it correctly classified all 26,219 instances in the test set. The macro average and weighted average metrics are also 1.00, indicating that the model's performance is consistent across both classes and weighted according to the number of instances in each class. Overall, these results suggest that the Logistic Regression model is a highly accurate classifier for this particular dataset.

### **The Decision Tree model is overfitting (pretty obvious as it is not pruned)**

****

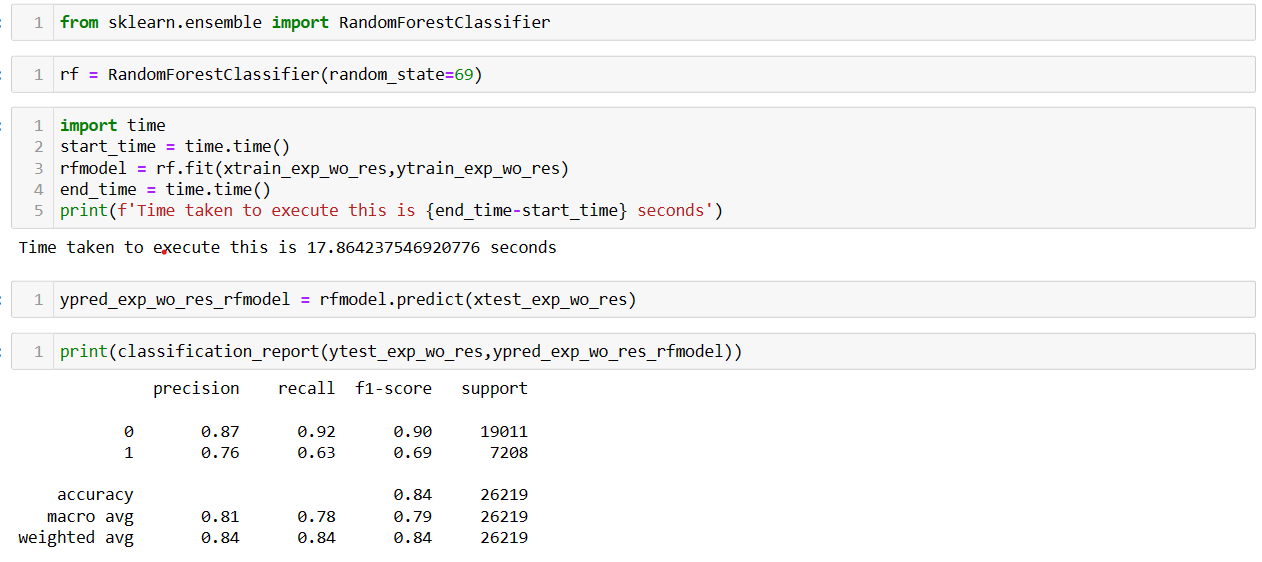
The model had an overall accuracy of 80%, meaning it correctly classified 80% of the samples. It performed well for class 0, but had more difficulty with class 1. The F1-score, which combines precision and recall, was higher for class 0 than for class 1. The macro average of precision, recall, and F1-score gives an average for both classes, while the weighted average takes into account the imbalance in the dataset. In this case, class 0 had a higher weight in the weighted average. The weighted average precision and recall were both higher than the macro average, indicating that the model performed better overall when taking the class imbalance into account.

### **Decision Tree Classifier without pruning on the data when reservation\_status has been dropped**



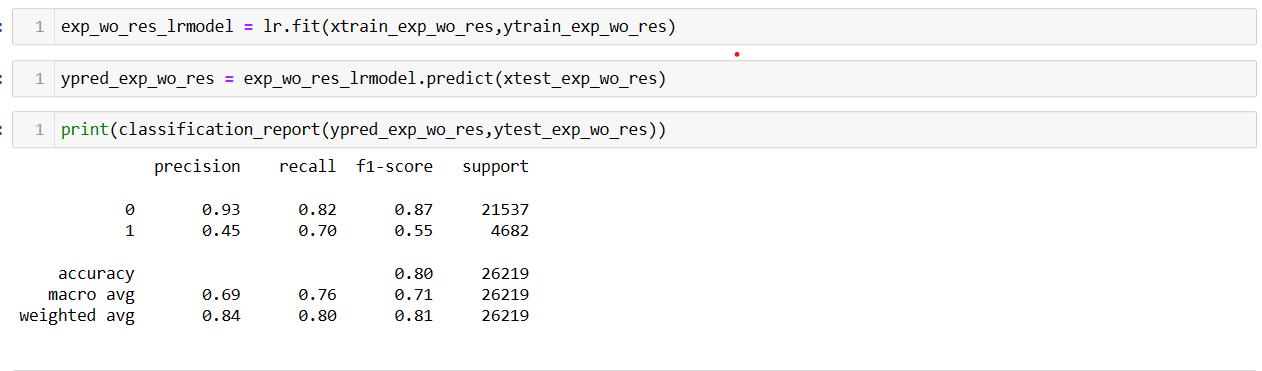
The model had an accuracy of 79%, performing slightly better for class 0. The F1-score was higher for class 0 than for class 1. The macro average precision, recall, and F1-score were all the same at 0.74. The weighted average precision, recall, and F1-score were all 0.79, taking into account the class imbalance. Overall, the model performed reasonably well but could potentially be improved with further analysis.

## **Random Forest Classifier without pruning**



The random forest model achieved an accuracy of 84%, performing slightly better for class 0. The F1-score was higher for class 0 than for class 1. The macro average precision, recall, and F1-score were 0.81, 0.78, and 0.79 respectively. The weighted average precision, recall, and F1-score were all 0.84, taking into account the class imbalance. Overall, the model performed well for both classes.

**Logistic Regression best fit model**

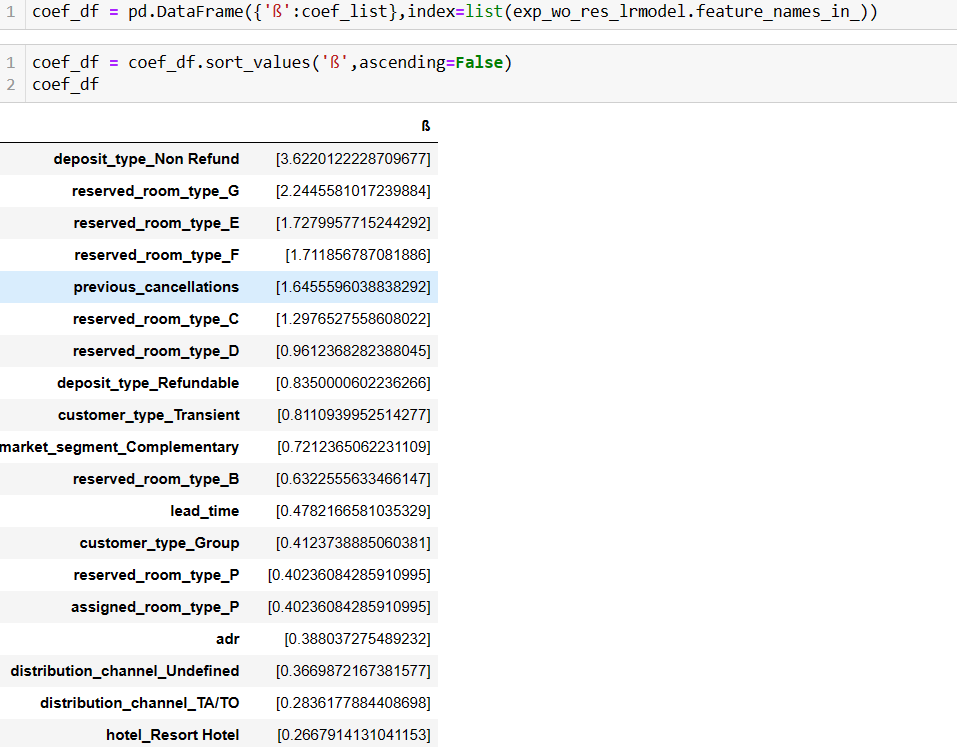
****

The report is about a model that predicts the outcome of two categories: 0 and 1. The model was tested on a dataset of 26,219 instances and achieved an overall accuracy of 80%, which means that it correctly predicted the outcome of 80% of the instances. The precision for class 0 was very high at 93%, which means that when the model predicted an instance to be in class 0, it was correct 93% of the time. However, the precision for class 1 was only 45%, indicating that the model made a lot of mistakes when predicting instances in this class.

The recall for class 0 was also high at 82%, indicating that the model was able to correctly identify 82% of the instances in class 0. The recall for class 1 was 70%, which means that the model correctly identified 70% of the instances in class 1.The F1 score, which is a measure of the model's overall performance, was higher for class 0 than for class 1. The weighted average F1 score was 81%, which indicates that the model performed reasonably well overall.

In summary, the model did a good job predicting instances in class 0 but struggled more with instances in class 1. Further investigation is needed to improve the model's performance on class 1.

These are the best beta coefficient values from Logistic Regression.



­

**Conclusion**

In our problem statement, incorrectly predicting that a customer will cancel their reservation (a false positive) could damage the hotel's reputation if the room is given to another customer, resulting in a booking conflict. Conversely, incorrectly predicting that a customer will not cancel (a false negative) can lead to revenue loss for the hotel. Therefore, both recall and precision metrics are crucial.

As our classes are balanced based on the domain, we consider accuracy as our primary metric, followed by f1\_score. After evaluating various models, we have found that Logistic Regression provides the highest accuracy of 80%, with proper explanations of the variables that affect the target variable. Therefore, it is our preferred model for this problem.