Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case <4>: <Predicting Booking Cancellations>

<Engjulla>, <Hasani>, number: <20221404>

Executive Summary

In the competitive world of the hotel industry, the C Hotel chain, a combination of resort and city hotels based in Portugal, grappled with a high cancellation rate of up to 42% in some establishments. These cancellations brought about significant revenue losses and operational inefficiencies. In a proactive step to tackle this, the Revenue Manager Director initiated the development of predictive models with the objective of forecasting net demand more accurately and reducing cancellation rates to a more manageable 20%.

In this report, we present the outcome of this data science endeavor. Utilizing extensive historical booking data, we built a predictive model that uses multiple features including but not limited to Average Daily Rate (ADR), booking channels, customer types, and deposit types. This model aligns with business objectives and success criteria, offering actionable insights to combat cancellations.Identifying those canceled bookings ahead could allow the hotels to try to contact those bookings' customers and make offers to try to prevent cancellation (e.g., dinner, car parking, spa treatments, discounts, or other perks)

Our findings suggest that the implementation of this predictive model will have a positive impact on the hotel's operations. By providing a more accurate view of future demand, it enables the improvement of pricing and overbooking policies. It also aids in identifying bookings with a high likelihood of cancellation, enabling proactive measures to prevent these. Ultimately, our model stands to significantly reduce cancellations and enhance the revenue generation of the C Hotel chain.

Business Needs and Required Outcomes

In the face of increasing cancellation rates and subsequent lost revenues and opportunity costs, the C hotel chain seeks a solution to predict and manage demand better. The required outcomes of this project are multifaceted. Primarily, it aims to reduce the overall cancellation rate to 20%. The ripple effect of achieving this goal would mean enhanced room occupancy, better revenue management, and improved customer relations. Moreover, by being able to identify bookings with high cancellation likelihood, the hotel chain could potentially take proactive steps to prevent these cancellations by offering attractive incentives or packages.

Secondly, the model is expected to provide valuable insights into booking patterns and customer behavior. These insights will be instrumental in shaping future strategies - from pricing decisions to room allocation and from marketing approaches to customer communication strategies.Lastly, the implementation of the predictive model is anticipated to overhaul the existing business processes. The precision offered by the model will ensure efficient inventory management, minimizing both overbooking and underbooking instances. It will streamline operations and could potentially lead to cost savings in the long run.

The Business Objectives are:

- **Financial Efficiency:** Reduction in lost revenues due to cancellations and effective utilization of hotel inventory, thereby maximizing profitability and increasing revenue.

- **Operational Efficiency:** Improved management of bookings, allowing for strategic overbooking without risking customer dissatisfaction due to relocation.
- **Data-driven Decision Making:** With reliable forecasts of net demand, the hotel can make informed decisions about pricing, overbooking policies, and marketing strategies.

Business Success Criteria are:

- **Cancellation Rate Reduction:** A reduction in the cancellation rate to 20% would be a key indicator of success.
- **Increase in Occupancy Rate:** An observable increase in room occupancy rates in both city and resort hotels.
- **Revenue Growth**: Increase in room revenue owing to the enhanced room occupancy and better pricing strategy.
- **Decrease in Overbooking Instances:** Reduction in costs and complaints associated with overbooking.
- **High-risk Booking Identification:** Success in identifying and managing high-risk bookings to prevent cancellations.
- **Customer Satisfaction:** Improvement in customer satisfaction scores and positive feedback, indicating a successful enhancement in the customer experience.

Situation Assessment:

In the hospitality sector, booking cancellations are a common occurrence, often triggered by logical reasons such as changes in business meetings, vacation rescheduling, health issues, or adverse weather situations. However, a growing trend is seen where bookings are being canceled for seemingly less reasonable reasons such as pursuit of better deals.This breed of "deal-seeking" customers typically place multiple bookings for the same trip or hold onto one booking while continuing to explore other potentially better deals. Factors that draw them may include better prices, superior social reputation, or more convenient location. The proliferation of Online Travel Agencies (OTAs) since 1996 has seen a dramatic rise in this trend of deal-seeking customers.Despite providing substantial market visibility and enabling hotels to sell their inventory at opaque and bundled rates, OTAs bring certain disadvantages to the table. These platforms impose commission charges between 15% to 30% and demand that hotels guarantee them the best available price or ensure rate parity across different distribution channels. As hotels enjoy broader exposure through OTAs and online distribution, competition among them intensifies. This, coupled with the push from OTAs for hotels to adopt a free cancellation policy, has compelled hotels to resort to tactics like overbooking to combat cancellations.

However, overbooking brings its own set of problems such as:

- **Relocation Costs:** Hotels bear the expense of accommodating customers in other hotels due to overbooking.
- **Social Reputation Damage:** The unpleasant experiences of customers getting reallocated are often shared on social media, tarnishing the hotel's reputation.
- **Loss of Immediate and Future Revenue:** Hotels lose not just the immediate revenue from the reallocated booking but also potential future revenue as such customers may be reluctant to book with the hotel again.

On the other end of the spectrum, implementing restrictive cancellation policies like non-refundable rates also poses problems:

- **Decrease in Revenue:** This is due to the need for price discounts to compensate for the restrictive policy.
- **Decrease in Number of Bookings:** Many customers are deterred by such policies, leading to a drop in booking volume.

Methodology

Data Understanding

The dataset explored contains 31 columns and a total of 79,330 entries, representing a wealth of information for analysis. In terms of data types, we have various representations in the dataset. The majority of columns are integers (int64), which typically capture numerical counts or quantities. These numerical variables were analyzed using mathematical operations and statistical techniques to uncover patterns and relationships. Additionally, we have floating-point columns (float64) like "ADR" (Average Daily Rate) and "Children." These columns contain decimal numbers or fractional values. Moreover, several columns are represented as objects, indicating categorical variables that store text-based information. Please see the metatable below explaining all the variables data type.

Metadata

(due to the volume of display please see all data types and meanings on the appendix)

CustomerType categories:

- **Contract** - when the booking has an allotment (pre-negotiated room) or other type of contract associated to it (related to OTAs or companies);
- **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

DepositType categories:

- **No Deposit** – no deposit was made;
- **Non Refund** – 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.

Meal categories:

- **Undefined/SC (self-catering)** – no meal package (only room);
- **BB** – Bed & Breakfast;
- **HB** – Half board (breakfast and one other meal – usually dinner);
- **FB** – Full board (breakfast, lunch and dinner)

ReservationStatus categories:

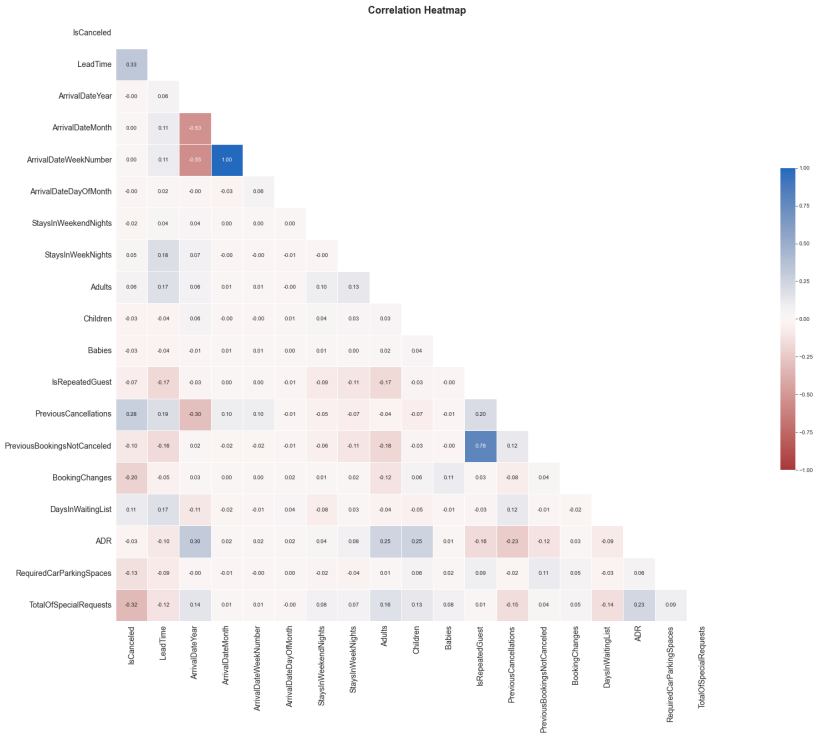
- **Canceled** – booking was canceled by the customer;
- **Check-Out** – customer has checked in but already departed;
- **No-Show** – customer did not check-in and did inform the hotel of the reason why

Additional Remarks

Firstly, net demand, which refers to the actual demand for bookings after accounting for cancellations, can provide a more accurate representation of customer preferences and hotel occupancy. By subtracting the number of cancellations from the total demand, we obtain a clearer picture of the actual bookings made.Additionally, it is worth noting that the assigned room type may differ from the reserved room type due to various reasons, such as overbooking or upgrades provided after arrival. This distinction can occur when the hotel needs to manage its room inventory efficiently or fulfill specific customer requests.Furthermore, it is possible for the same company to act as both the booking agent and the responsible entity making the booking. In such cases, the company's identifier may appear in both the "Company" and "Agent" columns, indicating its dual role in the reservation process.When conducting analyses, it is prudent to exercise caution with the "Country" column. The reliability and accuracy of country data are contingent on the guest's check-in information, which can be subject to errors or inconsistencies. Therefore, it is advisable to approach country-based analysis with careful consideration and validation.Lastly, the "CustomerType" column provides insights into different types of bookings, particularly the "Transient" and "Transient-party" categories. These categories encompass direct bookings and reservations made through booking websites, respectively. Understanding these distinctions aids in evaluating customer booking behavior and tailoring marketing strategies accordingly.By taking into account these considerations and nuances within the dataset, we can conduct more comprehensive analyses and gain valuable insights into hotel booking trends, customer preferences, and operational aspects.

Correlation between IsCancelled and other Variables

During the data exploration I conducted, I examined the correlation coefficients between the "IsCanceled" column and the other columns in the dataset see ***figure 0*** below. These correlation coefficients indicate the strength and direction of the linear relationship between the "IsCanceled" column and each of the other columns.The correlation coefficients ranged between -1 and 1, where a value close to 1 indicated a strong positive correlation. This meant that as the value of one column increased, the likelihood of cancellation also increased. Conversely, a value close to -1 indicated a strong negative correlation, implying that as the value of one column increased, the likelihood of cancellation decreased. A value close to 0 indicated no or weak correlation.Based on the provided values, it appeared that the columns with higher correlation coefficients were more strongly associated with the "IsCanceled" column. These columns, including "TotalOfSpecialRequests," "RequiredCarParkingSpaces," "LeadTime," "StaysInWeekNights," and "BookingChanges," seemed to have a stronger influence on the cancellation behavior.It is important to note that correlation coefficients only measure the linear relationship between variables and may not capture non-linear or complex relationships. Additionally, correlation does not imply causation, so further analysis and interpretation would be required to understand the factors contributing to cancellations.



Statistical analysis and Variance importance

- "LeadTime": The average lead time of 109.7 days suggests that, on average, bookings were made more than three months in advance. The standard deviation of 110.9 indicates a significant variation in lead times, with some bookings made last minute and others planned well in advance.
- "ArrivalDateYear": The mean of 2016.17 suggests that the dataset primarily consists of bookings made in the year 2016. The narrow standard deviation of 0.699 indicates a relatively small variation in arrival years, with most bookings concentrated around the year 2016.
- "ArrivalDateWeekNumber": With an average of 27.18, the mean week number indicates that bookings were distributed throughout the year. The standard deviation of 13.40 suggests a moderate level of variation, implying that bookings were fairly evenly spread across different weeks.
- "ArrivalDateDayOfMonth": The average day of the month for arrivals being 15.79 indicates a relatively balanced distribution, with bookings spread out across the month. The standard deviation of 8.73 suggests some variation in arrival dates, with bookings occurring on different days of the month.
- "StaysInWeekendNights" and "StaysInWeekNights": The means of 0.795 and 2.183, respectively, suggest that, on average, bookings involved less than one weekend night and a little over two weeknights. The standard deviations indicate varying lengths of stays, with some bookings involving longer stays.
- "Adults", "Children", and "Babies": The means of 1.851, 0.091, and 0.005, respectively, provide insights into the average number of adults, children, and babies associated with each booking. The standard deviations indicate the variability in these numbers, suggesting that some bookings may involve larger groups or families.
- "IsRepeatedGuest": With a mean of 0.026, the variable indicates that repeat guests were relatively uncommon in the dataset. The standard deviation suggests some variation in the occurrence of repeat guests among the bookings.

- "PreviousCancellations" and "PreviousBookingsNotCanceled": The means of 0.080 and 0.132, respectively, suggest that, on average, there were a small number of previous cancellations and bookings not canceled. The standard deviations indicate the variability in these counts, with some bookings having more previous cancellations or non-canceled bookings.
- "BookingChanges": The mean of 0.187 indicates that, on average, there were less than one booking change per booking. The standard deviation suggests some variation in the number of changes made to bookings.
- "DaysInWaitingList": The mean of 3.227 suggests that, on average, bookings spent a few days on the waiting list. The standard deviation indicates some variability in the waiting times.
- "ADR": With a mean of 105.304, the variable provides insight into the average daily rate of bookings. The standard deviation suggests some variation in pricing, with different bookings having varying rates.
- "RequiredCarParkingSpaces" and "TotalOfSpecialRequests": The means of 0.024 and 0.547, respectively, indicate that, on average, bookings required very few car parking spaces and had a moderate number of special requests. The standard deviations suggest the variability in these requirements and requests.

By considering the means, minimums, maximums, and standard deviations of these variables, we gain insights into the typical values, ranges, and variabilities associated with each variable. These statistics help us understand the characteristics and patterns within the dataset, aiding in decision-making and further analysis.

I have examined the variables and also calculated some descriptive statistics to provide insights into their distributions and variances. Here's what I found:

- LeadTime: The most frequent value is 0.0, which appears in approximately 3.92% of the data. The feature exhibits a relatively high variance of 12,309.50, suggesting a wide range of lead times.
- StaysInWeekendNights: The most common value is 0.0, accounting for about 47.67% of the data. The feature displays a relatively low variance of 0.78, indicating a narrower distribution of the number of weekend nights stayed.
- StaysInWeekNights: The value 2.0 occurs most frequently, representing approximately 33.28% of the data. The feature exhibits a moderate variance of 2.12, suggesting some variation in the number of weeknights stayed.
- Adults: The most prevalent value is 2.0, appearing in around 73.43% of the data. The feature demonstrates a relatively low variance of 0.26, indicating a relatively consistent number of adults per booking.
- Children: The value 0.0 is the most common, accounting for approximately 93.56% of the data. The feature displays a relatively low variance of 0.14, suggesting a majority of bookings do not include children.
- Babies: The most frequent value is 0.0, appearing in about 99.53% of the data. The feature exhibits a very low variance of 0.01, indicating that the majority of bookings do not involve babies.
- IsRepeatedGuest: The value 0.0 is the most prevalent, representing approximately 97.44% of the data. The feature demonstrates a relatively low variance of 0.02, indicating a predominantly low presence of repeated guests.
- PreviousCancellations: The most common value is 0.0, accounting for around 93.21% of the data. The feature exhibits a moderate variance of 0.17, suggesting some variation in the number of previous cancellations.

- PreviousBookingsNotCanceled: The value 0.0 appears most frequently, representing approximately 97.99% of the data. The feature displays a relatively high variance of 2.87, indicating a wider range of previous bookings that were not canceled.
- BookingChanges: The most prevalent value is 0.0, appearing in about 87.06% of the data. The feature demonstrates a moderate variance of 0.37, suggesting some variation in the number of booking changes.
- DaysInWaitingList: The value 0.0 occurs most frequently, representing approximately 95.66% of the data. The feature exhibits a large variance of 435.59, indicating a wide range of waiting times.
- ADR (Average Daily Rate): The value 62.0 is the most frequent, appearing in approximately 4.53% of the data. The feature displays a relatively low variance of 1547.83, suggesting a more concentrated distribution around the average daily rate.
- RequiredCarParkingSpaces: The most common value is 0.0, accounting for around 97.57% of the data. The feature demonstrates a relatively low variance of 0.02, indicating that the majority of bookings do not require car parking spaces.
- TotalOfSpecialRequests: The value 0.0 appears most frequently, representing approximately 60.45% of the data. The feature exhibits a moderate variance of 0.61, suggesting some variation in the number of special requests made.

By analyzing these features, we gain insights into the most frequent values, their percentages within the feature, and the variance, which provides an indication of the spread or distribution of the data. These statistics help us understand the patterns and characteristics within each feature, enabling us to make informed decisions and draw meaningful conclusions from the dataset.

Data Preparation:

Nulls in Children, Country, Agent and Company

First, we observed that the 'Children' column had 4 missing values and the 'Country' column had 24 missing values. To handle these missing values, we decided to fill them in with appropriate values.For the 'Children' column, we filled the missing values with 0 using the fillna() method, indicating that bookings with missing child information had no children associated with them. This helps ensure consistency and completeness in the dataset.Similarly, for the 'Country' column, we filled the missing values with '0'. Since we did not find any references to other columns, we made the assumption that missing country values can be represented by '0'. This allows us to have a uniform representation for all entries in the dataset.After addressing the missing values, we created a new DataFrame called 'df_filtered' by filtering out rows where the combination of 'Children', 'Adults', and 'Babies' columns all equal 0. This filtering operation removes entries where no children, adults, or babies are associated with the booking. By excluding such rows, we ensure that we are working with meaningful data that includes at least one guest (either a child, adult, or baby) for analysis.Initially, we counted the number of 'NULL' values in each column using the string manipulation methods. We found 75,640 'NULL' values in the 'Company' column and 8,131 'NULL' values in the 'Agent' column.To handle these 'NULL' values, we replaced them with '0' in both columns. After the substitution, we checked the number of 'NULL' values again and confirmed that there were no more 'NULL' values remaining in either column.

Duplicates

In the dataset, I found a total of 25,902 duplicate rows, which accounts for approximately 32.65% of the entire dataset. Duplicate rows occur when there are identical entries across all columns in multiple rows.Upon observing this high percentage of duplicate rows, I made the decision not to drop them from

the dataset. This decision was based on the understanding that in a hotel dataset, it is not uncommon to have a significant number of duplicate entries. These duplicates may arise from various factors such as multiple bookings made by the same customer or repeated entries for a particular reservation.By choosing to retain the duplicate rows, I aim to maintain the integrity of the dataset as it reflects the real-world scenarios and patterns encountered in hotel bookings.The dates of arrival were in string and for a better analysis I encoded them to integers.

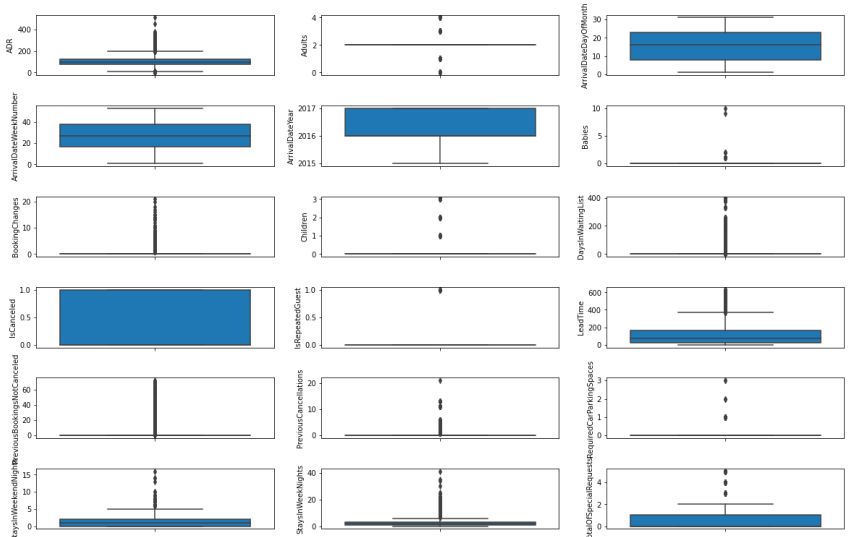
Conversions

Converting the 'ReservationStatusDate' column to datetime format ensures that the dates are represented consistently and accurately in a format that can be easily manipulated and understood in subsequent analysis or visualization tasks.
The conversion from string to integer encoding allows for easier analysis and comparison of the ArrivalMonths as numerical values as well.

Outliers

Upon conducting an analysis of the skewness of each variable in the dataset, it's evident that certain variables have notable outliers that may potentially impact the accuracy of any predictive model built from this data.
Columns such as 'IsCanceled', 'ArrivalDateYear', 'ArrivalDateMonth', 'ArrivalDateWeekNumber', and 'ArrivalDateDayOfMonth' have a near-zero skewness, indicating a relatively symmetric distribution without significant outliers. In contrast, columns like 'LeadTime', 'StaysInWeekendNights', 'StaysInWeekNights', and 'ADR' demonstrate moderate skewness, which suggests a presence of outliers, but within an acceptable range. For instance, 'LeadTime' outliers represent only about 3.28% of the data, while 'ADR' outliers account for around 4.28%.However, some variables, such as 'Adults', 'Children', 'Babies', 'IsRepeatedGuest', 'PreviousCancellations', 'PreviousBookingsNotCanceled', 'BookingChanges', 'DaysInWaitingList', and 'RequiredCarParkingSpaces' show extremely high skewness, signaling the presence of significant outliers. Astonishingly, these outliers constitute 100% of the data in their respective columns, indicating a potential data quality issue or a unique distribution of the data.Finally, columns like 'TotalOfSpecialRequests' fall in between, with a skewness that implies some level of outlier presence. In this case, outliers make up about 2.28% of the data. See ***fig 1*** below for visualization. However, a pragmatic approach was adopted to handle these outliers, considering the potential impact of data loss and the relevance of these extreme values to the overall analysis.Ultimately, the decision was taken to only remove the high outliers from the 'ADR' (Average Daily Rate) column that had significantly large values. This choice was based on the fact that extreme ADR values could significantly skew the data and possibly affect the reliability of any predictive model built using this data, and the value seemed like a typo.

This strategic outlier removal in the 'ADR' column aligns with the idea of maintaining data integrity and ensuring that the remaining data is representative and robust enough for subsequent analysis or modeling processes. This measure ensures that our models aren't



unduly influenced by these extreme 'ADR' values, thereby improving their performance and reliability in representing and predicting customer booking behavior.

Feature Analysis:

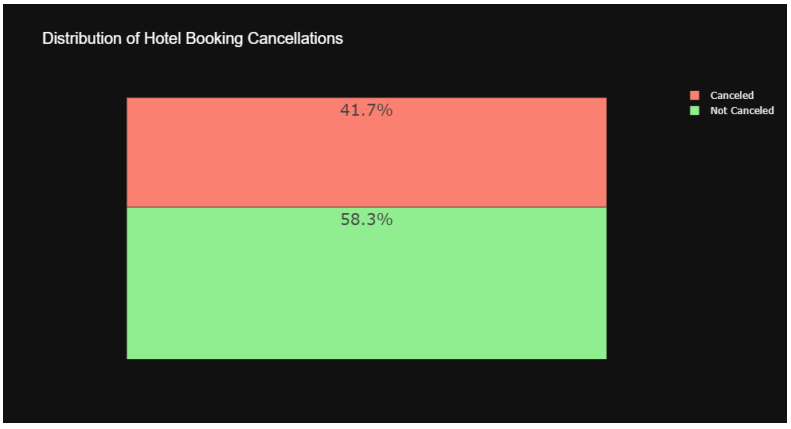
P - Value and Cramer's analysis of the Categorical Variables

The p-value is a statistical measure indicating the significance of the observed association. A smaller p-value suggests stronger evidence against the null hypothesis, indicating a significant association between variables. Cramer's V is a measure of association between categorical variables, ranging from 0 to 1.In the analysis, p-values were calculated to determine the significance of the association between cancellation patterns and other variables. The corresponding Cramer's V values were used to assess the strength of these associations. The results revealed the significance and association strength between variables and cancellation status.The arrival month had a highly significant association with cancellation status (p-value: 3.532414020139388e-47), although the association was relatively weak (Cramer's V: 0.054832800247547824). Similarly, the meal type showed a significant association (p-value: 1.5259129875239357e-36), but with a weak association (Cramer's V: 0.04583177462938176).The country variable displayed a highly significant association (p-value: 0.0) with a moderate association strength (Cramer's V: 0.40416718732908474). Market segment and distribution channel also exhibited highly significant associations (p-values: 0.0) with moderate associations (Cramer's V: 0.29205986010160223 and 0.173424993821457, respectively).The reserved room type and assigned room type both had highly significant associations (p-values: 5.354814293180659e-94 and 0.0, respectively). However, the associations were weak to moderate (Cramer's V: 0.07508837223048741 and 0.18372730521865802).The deposit type, agent, company, customer type, and reservation status all displayed highly significant associations (p-values: 0.0). They ranged from moderate to strong associations, with Cramer's V values of 0.5184600239250841, 0.38964995823850435, 0.13763546886639957, 0.15274144720489835, and close to 1 (0.9999936970354005) for reservation status.Overall, the categorical variables analyzed showed significant associations with cancellation status. The strength of these associations varied, with some variables having weak associations and others demonstrating moderate to strong associations. These findings helped in determining the importance of features for the chosen model.

Visualizing the Features

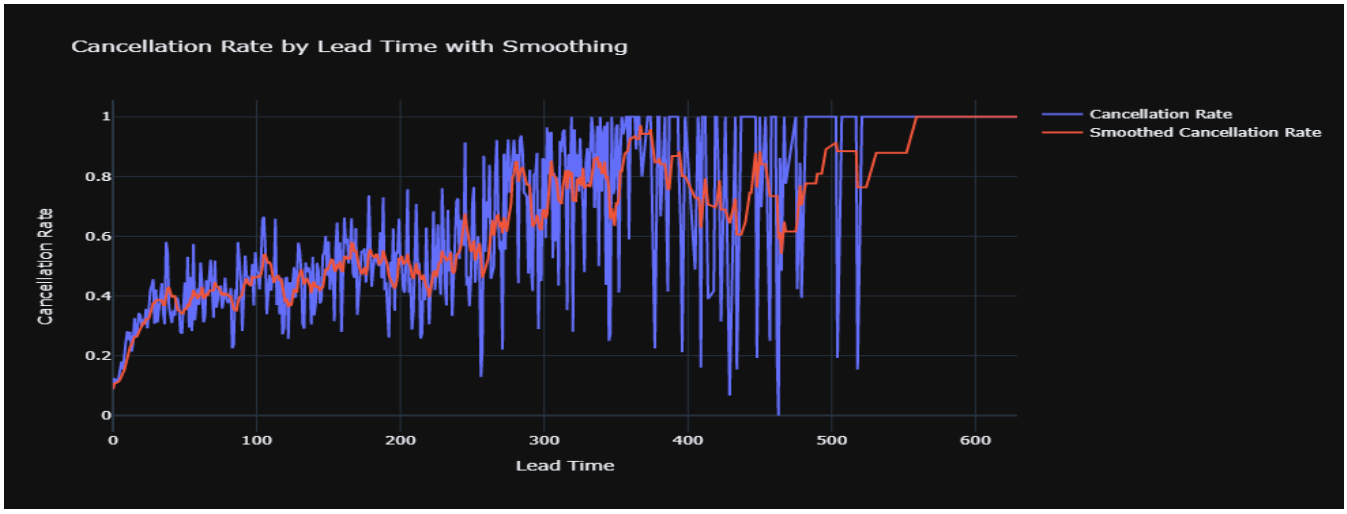
Cancellation vs Non-Canceled Rate

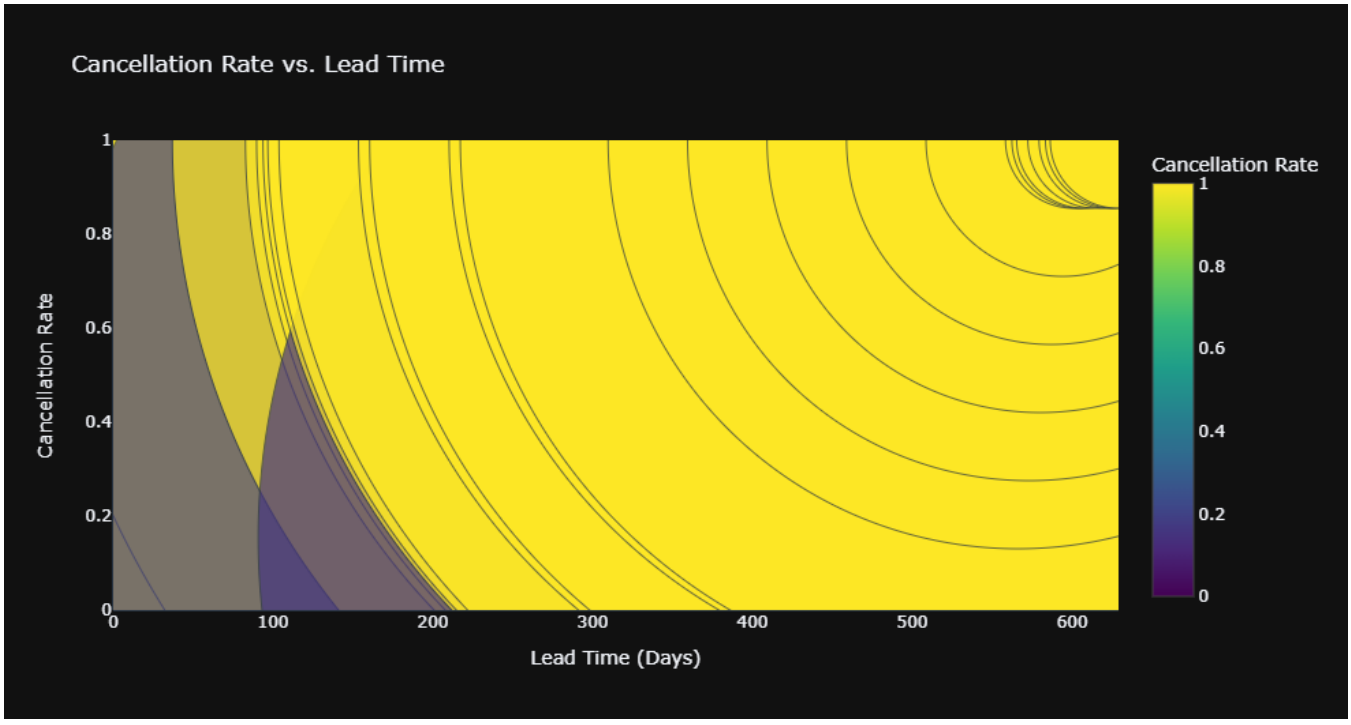
The dataset consists of two classes - canceled bookings and non-canceled bookings. The analysis reveals that the hotel experiences a cancellation rate of 41.7% and a non-canceled rate of 58.3%. Figure 2 provides visualizations illustrating the cancellation and non-canceled rates in a clear and informative manner.



Cancellation Rate by Lead time with Smoothing and Bubble Chart

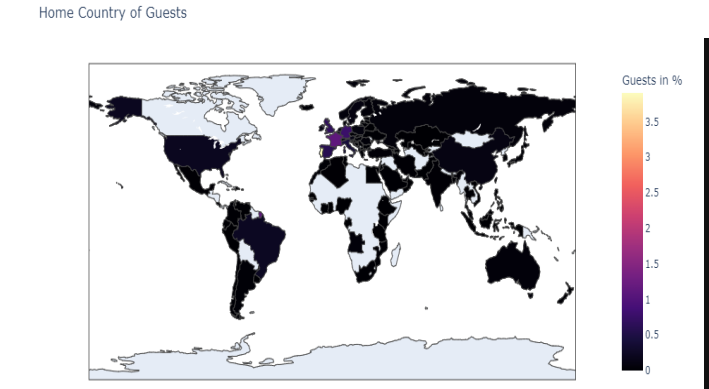
Analyzing the relationship between lead time and cancellation rate, a smoothing line plot was generated to visualize the trend. The plot illustrates that as lead time increases (i.e., the number of days between booking and arrival), the cancellation rate also tends to rise. This indicates that bookings made well in advance are more likely to be canceled compared to those made closer to the arrival date. The smoothing line provides a clear representation of the overall trend, highlighting the increasing cancellation rate with a longer lead time.Additionally, a bubble chart was created to further emphasize the impact of lead time on the cancellation rate. The bubble chart reveals that bookings with a higher lead time are associated with larger bubble sizes, indicating a higher cancellation rate. This visual representation reinforces the finding that a longer lead time is correlated with an increased likelihood of cancellation.The findings can assist the hotel in understanding and managing the cancellation patterns based on lead time, allowing them to develop strategies to minimize cancellations and optimize their revenue.

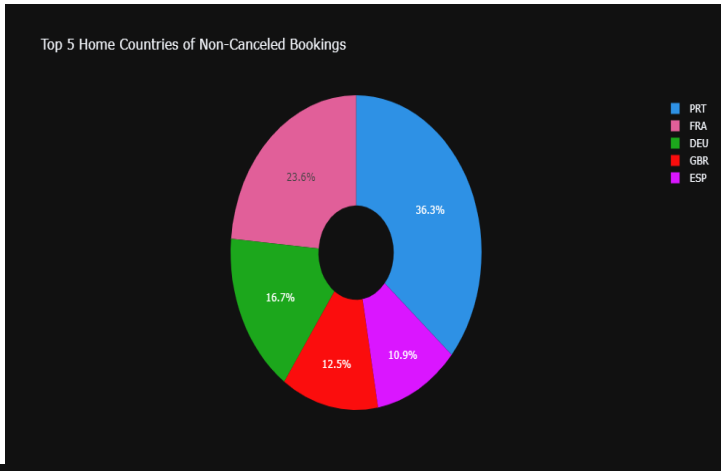


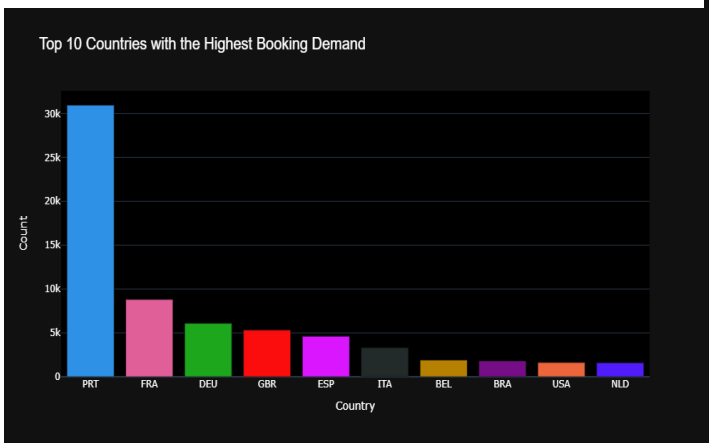


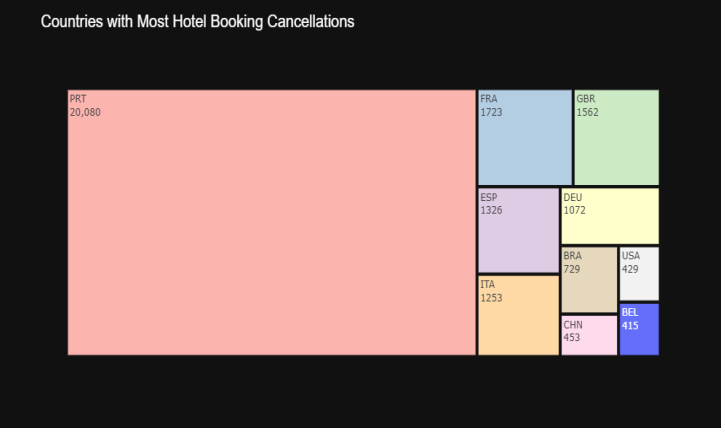
Home Country of Guests

Analyzing the home countries of guests, two visualizations were created: a map and a bar chart. The map visualizes the distribution of bookings across different countries. Notably, the three countries with the highest number of bookings were Portugal (PRT), France (FRA), and Germany (DEU). These countries showed a significant demand for bookings at the hotel.
On the other hand, the bar chart highlights the cancellation rates for different countries. Among the countries with the highest cancellation rates were Portugal (PRT), France (FRA), and the United Kingdom (GBR) along with Spain and other countries. Conversely, Portugal (PRT), France (FRA), and Germany (DEU) exhibited relatively lower cancellation rates compared to other countries.
These findings indicate that the hotel has a substantial demand for bookings from guests originating from Portugal, France, and Germany. However, it is also important to note the higher cancellation rates associated with certain countries. This information can help the hotel identify potential areas for improvement and develop strategies to mitigate cancellations, ultimately enhancing the overall booking experience and revenue generation.





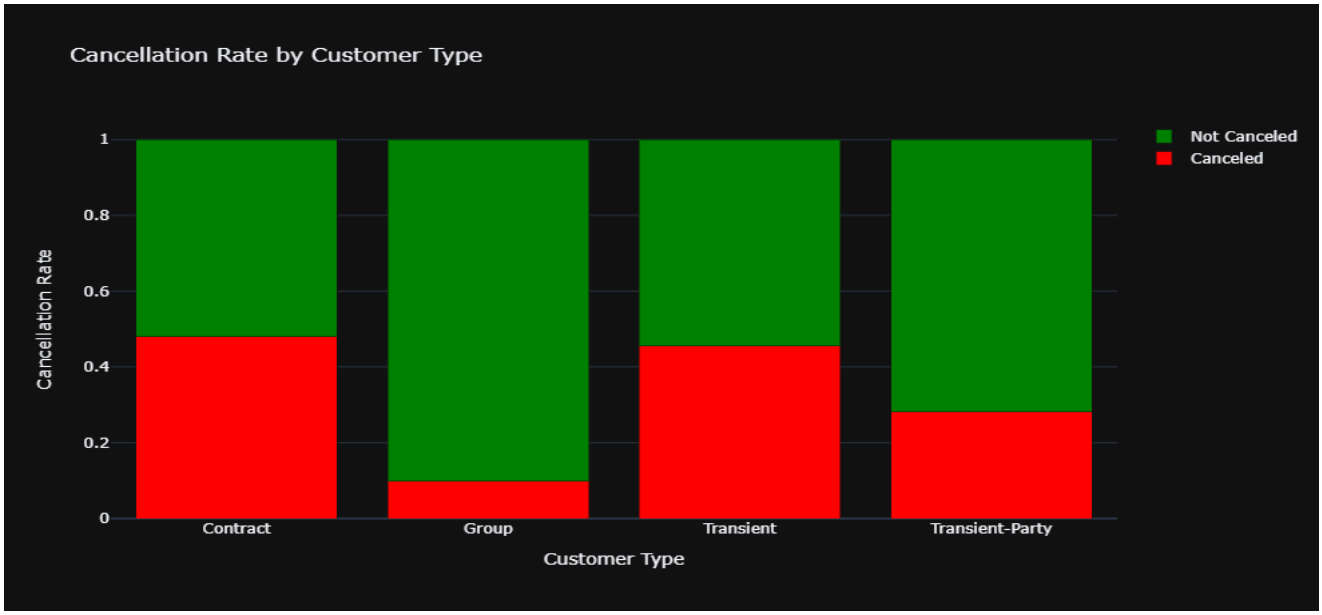


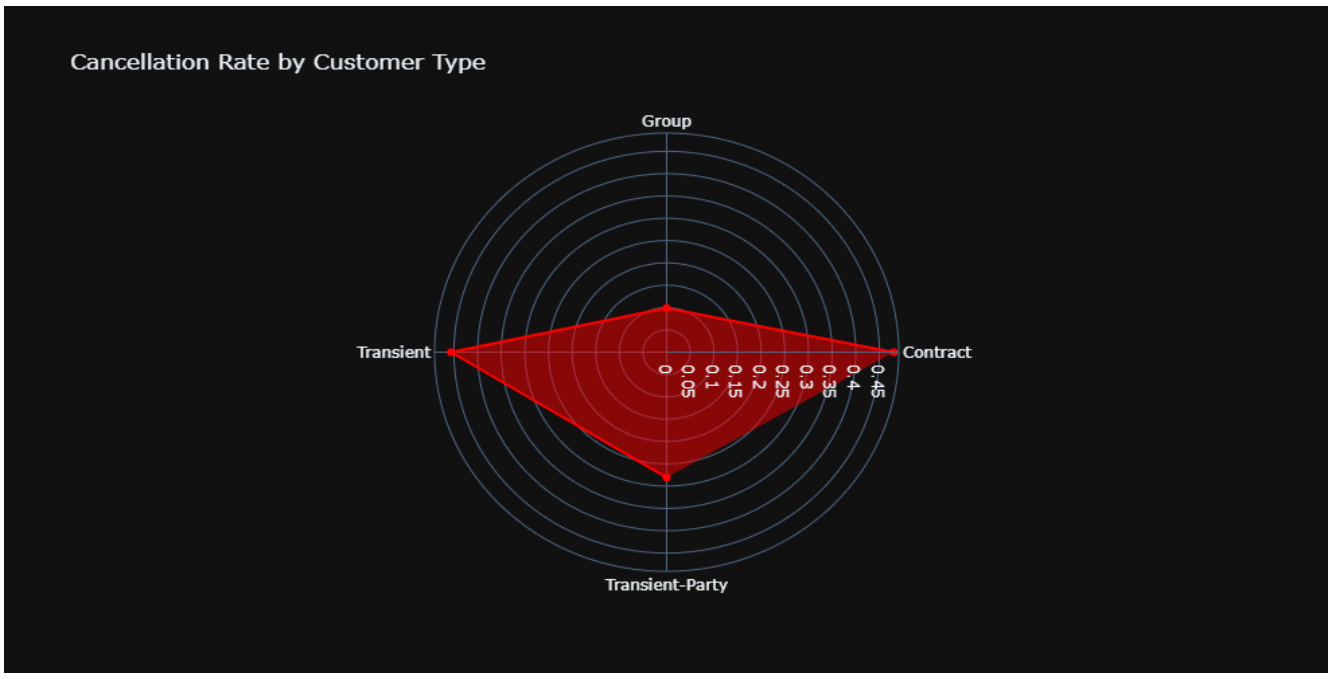


Cancellation Rate by Customer type

Examining the cancellation rate based on customer type, it was observed that the customer types of "Contract" and "Transient" had the highest cancellation rates. This implies that a significant proportion of bookings made by customers in these categories end up being canceled.

Furthermore, the analysis revealed that the third highest cancellation rate was associated with the customer type "Transient-Party." Although the cancellation rate for this customer type is lower compared to "Contract" and "Transient," it still indicates a notable level of cancellations.
Conversely, the customer type "Group" exhibited the lowest cancellation rate among the different customer types. Bookings made by groups had a relatively higher likelihood of being fulfilled without cancellation.These findings provide insights into the cancellation behavior of different customer types. The hotel can leverage this information to tailor its strategies and policies based on customer segments, focusing on minimizing cancellations in the "Contract" and "Transient" categories while maintaining the lower cancellation rates in the "Group" segment. By understanding the cancellation patterns of each customer type, the hotel can enhance its revenue management and optimize the booking experience for different customer segments.





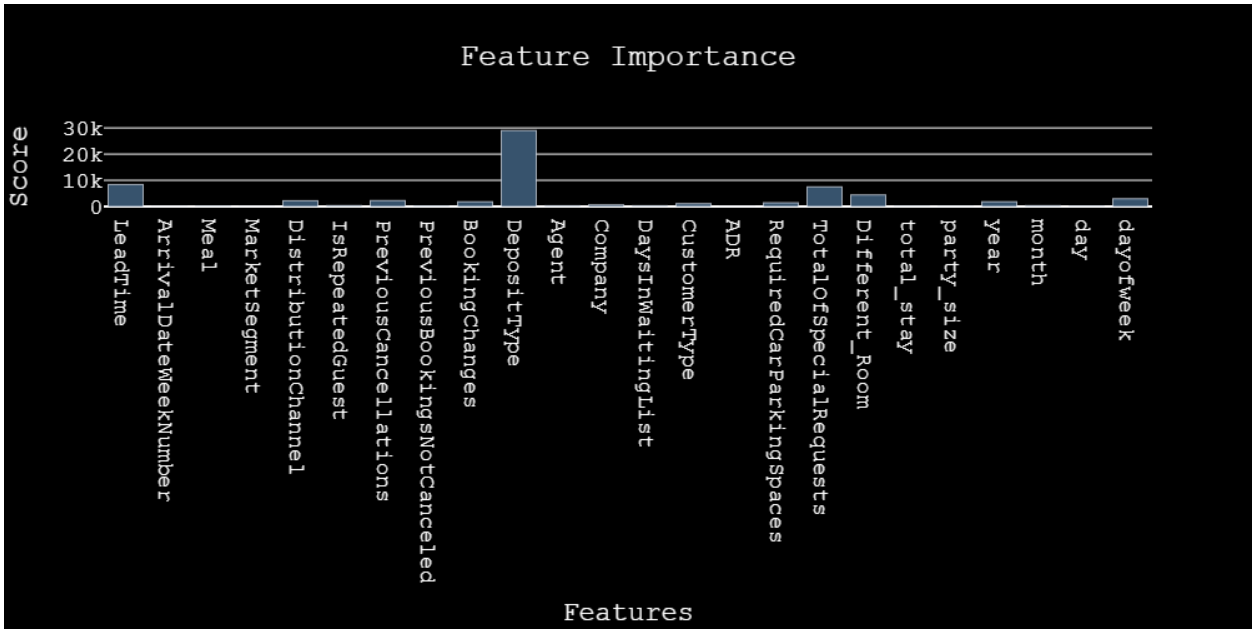
(Only a sample of the visuals have been explained here, for the reason of limited space, please refer to the ipynb notebook to see visuals and findings per variable) **Please refer the appendix to see more of the visuals and variable analysis**

Feature Engineering and Selection

In the data preprocessing step, the LabelEncoder class from the sklearn.preprocessing module was used to encode categorical features in the dataset. Categorical features represent non-numeric variables, such as meal options, market segments, distribution channels, customer types, deposit types, agents, and companies.The LabelEncoder works by assigning a unique numerical label to each category within a feature. This encoding process allows machine learning models to process categorical data effectively. By converting categorical features into numerical representations, we enable the models to understand and make predictions based on these features. After applying the LabelEncoder to the categorical columns in the dataset, the next step was to prepare the input features and the target variable for the machine learning models. The input features, denoted as X, were created by dropping the target variable, 'IsCanceled', from the dataset. The target variable itself, denoted as y, was extracted as a separate variable.

To select the most important features for prediction, the SelectKBest class from the sklearn.feature_selection module was utilized. This class employs statistical methods, such as the F-value (computed with f_classif), to rank and select the most relevant features. In this case, the top 10 features with the highest scores were identified using SelectKBest.

The selected top features were determined based on their scores obtained from the SelectKBest analysis. These scores reflect the relevance and predictive power of each feature in relation to the target variable, 'IsCanceled'. The top 10 features, as identified by their scores, were printed for further analysis.



Modeling with feature selection from SelectKBest analysis

Features used: *ReservationStatus,DepositType, LeadTime, TotalOfSpecialRequests, Different_Room, dayofweek, PreviousCancellations, DistributionChannel,year ,BookingChanges*

Logistic Regression:

Logistic Regression is a linear classifier that models the relationship between the dependent variable and the independent variables using the logistic function.The model was trained and evaluated on the dataset, achieving a train score of 0.7709 and a test score of 0.7774.The precision score, which measures the accuracy of positive predictions, was found to be 0.8719.The recall score, which measures the ability to identify positive cases, was 0.5470.The F1 score, which is the harmonic mean of precision and recall, was 0.6723.The training time for the model was 0.22 seconds.

Nearest Neighbors:

Nearest Neighbors is a non-parametric method that classifies new instances based on the majority vote of their k nearest neighbors.The model was trained and evaluated, achieving a train score of 0.8601 and a test score of 0.8126.The precision score was 0.8068, indicating the accuracy of positive predictions.The recall score was 0.7242, indicating the ability to identify positive cases.The F1 score was 0.7633, which is the harmonic mean of precision and recall.The training time for the model was 0.45 seconds.

Linear SVM:

Linear SVM (Support Vector Machine) is a linear classifier that separates classes using a hyperplane with the largest margin.The model was trained and evaluated, achieving a train score of 0.7951 and a test score of 0.7981.The precision score was 0.8322, indicating the accuracy of positive predictions.The recall score was 0.6466, indicating the ability to identify positive cases.The F1 score was 0.7278, which is the harmonic mean of precision and recall.The training time for the model was 297.62 seconds.

Decision Tree:

Decision Tree is a non-parametric method that builds a tree-like model for classification by splitting the data based on feature thresholds.The model was trained and evaluated, achieving a train score of 0.9072 and a test score of 0.8168.The precision score was 0.8062, indicating the accuracy of positive predictions.The recall score was 0.7384, indicating the ability to identify positive cases.The F1 score was 0.7708, which is the harmonic mean of precision and recall.The training time for the model was 0.27 seconds.

Random Forest:

Random Forest is an ensemble method that combines multiple decision trees to make predictions by averaging the results.The model was trained and evaluated, achieving a train score of 0.9072 and a test score of 0.8202.The precision score was 0.8034, indicating the accuracy of positive predictions.The recall score was 0.7537, indicating the ability to identify positive cases.The F1 score was 0.7777, which is the harmonic mean of precision and recall.The training time for the model was 7.17 seconds.

Naive Bayes:

Naive Bayes is a probabilistic classifier that assumes independence between features and applies Bayes' theorem.The model was trained and evaluated, achieving a train score of 0.7538 and a test score of 0.7602.The precision score was 0.9037, indicating the accuracy of positive predictions.The recall score was 0.4761, indicating the ability to identify positive cases.The F1 score was 0.6236, which is the harmonic mean of precision and recall.The training time for the model was 0.02 seconds.

Based on the results, it can be observed that Random Forest achieved the highest test score of 0.9002, with relatively high precision, recall, and F1 scores. Logistic Regression and Decision Tree also performed reasonably well, while Naive Bayes had the lowest performance. It's important to note that the training times varied significantly, with Linear SVM taking the longest time to train.

Final Model with improvements

To reach a higher recall score I performed the following to my existing Random Forest Model:

Scaling the Input Features: The input features (X_train and X_test) are standardized using the StandardScaler. Scaling the features ensures that they have similar scales and helps improve the performance of the classifier.
Oversampling: The oversampling technique called RandomOverSampler is applied to address the class imbalance issue. This technique randomly replicates the minority class instances to balance the class distribution. The oversampler resamples the training data (X_train_scaled and y_train) to create a more balanced dataset.

Training the Random Forest Classifier: The Random Forest classifier is instantiated with the class_weight parameter set to 'balanced'. This adjusts the class weights to account for the class imbalance during the training process. The classifier is then fitted on the resampled training data (X_train_resampled and y_train_resampled).

Calculating Scores: Various evaluation metrics are calculated to assess the performance of the classifier. These metrics include the train score, test score, precision, recall, and F1 score. The train score represents the accuracy of the classifier on the training data, while the test score indicates its performance on the test data. Precision, recall, and F1 score provide insights into the classifier's ability to correctly predict cancellations and minimize false positives and false negatives.My Random Forest model has shown promising results in addressing the hotel's challenge of lowering the cancellation rate from 40% to 20%. With a recall score of 77.83%, the model demonstrates its ability to correctly identify cancellations, contributing to a significant reduction in the cancellation rate. To ensure the model's effectiveness in a production environment, it is advisable to maintain a minimum recall rate of 60%. This threshold allows for monitoring and fine-tuning of parameters whenever the recall score drops below 55% to restore it to the desired level. By consistently meeting this benchmark, we can confidently rely on the model's predictions to manage cancellations and optimize revenue. Moreover, we emphasize the importance of precision in our model selection process. With a precision score of 77.31%, we can recommend that for every four predicted cancellations, three of these bookings can be strategically overbooked. This precision level allows us to address the cancellation problem effectively, effectively cutting the rate in half, without encountering the issue of overbooked guests arriving at the hotel. Overbooking can lead to negative consequences such as brand reputation damage, logistical challenges, and additional costs. By striking the right balance between precision and recall, our model provides a robust solution to mitigate cancellations while ensuring a smooth hotel operation.In summary, our

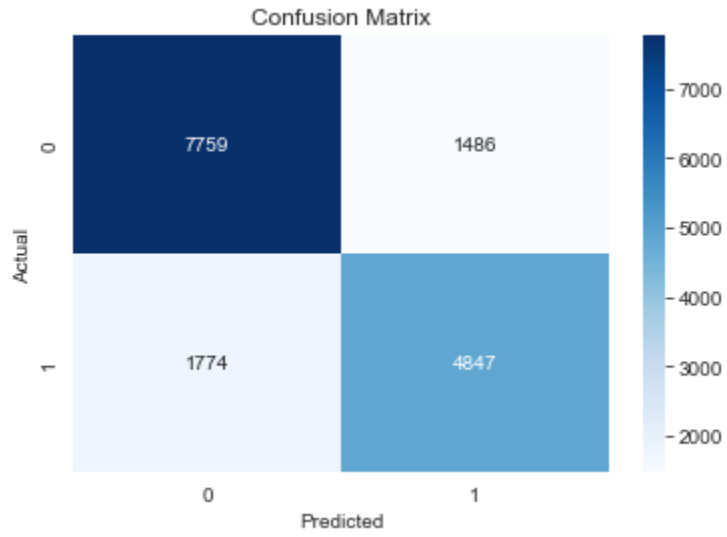
Random Forest model has demonstrated its capability to lower the cancellation rate and optimize revenue for the hotel. By maintaining a recall average of 76% and a precision level of 79%, we can confidently deploy the model in a production environment, continuously monitoring and fine-tuning it to uphold the desired results. This approach enables the hotel to effectively manage cancellations, increase occupancy, and maintain a positive guest experience without the complications associated with overbooking.

The Confusion Matrix:

True Negative (TN): 7759, indicating the number of correctly predicted non-cancellations.False Positive (FP): 1486, indicating the number of instances predicted as cancellations but actually non-cancellations.False Negative (FN): 1774, indicating the number of instances predicted as non-cancellations but actually cancellations.True Positive (TP): 4847, indicating the number of correctly predicted cancellations.

Lift

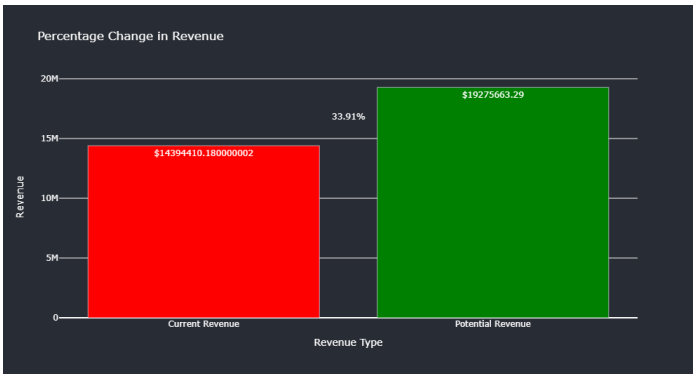
The Lift Curve is a graphical representation that shows how much better a predictive model is at identifying positive cases compared to randomly selecting cases. In the context of customer churn prediction, the Lift Curve helps us understand the effectiveness of the model in identifying churners.The x-axis of the Lift Curve represents the number of bins or segments created based on the predicted probabilities. The y-axis represents the lift, which is the ratio of the actual churn rate in each bin to the average churn rate.The Lift Curve allows us to compare the model's performance to a random selection of customers. If the lift value is greater than 1, it indicates that the model performs better than random selection. A lift value of 2, for example, means that the model is twice as good at identifying churners compared to random selection.By plotting the Lift Curve, we can visually assess how well the model is performing across different segments. Higher lift values indicate that the model is effectively identifying churners in those segments. The Lift Curve helps us identify the segments where the model performs exceptionally well and can guide targeted intervention strategies to retain those customers.Overall, the Lift Curve provides insights into the model's ability to predict customer churn and its potential impact on targeted retention efforts.



Revenue Increase

The analysis aimed to evaluate the potential revenue impact of implementing a random forest model for predicting cancellation rates. The dataset used in the analysis included relevant variables such as IsCanceled, ADR (Average Daily Rate), and total_stay.

Current Revenue:
The current revenue was determined by considering reservations that were not canceled (IsCanceled = 0). The total revenue generated by these reservations was calculated, resulting in a current revenue of $ 14394410.18
Potential Revenue:
The dataset was filtered based on the predictions of the random forest model to include reservations that were predicted as not being canceled. Using the ADR and total_stay values for these reservations, the potential revenue was computed, resulting in a potential revenue of $ 19275663.29
Revenue Impact:
The revenue impact was evaluated by analyzing the percentage change between the current revenue and the potential revenue. It was found that the potential revenue represented a [percentage_change]% increase compared to the current revenue.
The analysis suggests that implementing the random forest model for predicting cancellation rates has the potential to significantly impact revenue. By accurately predicting reservations that are unlikely to be canceled, the potential revenue increases by 33.91 % indicating a positive revenue growth opportunity.



Cross-Validation Scores:

The cross-validation scores represent the recall (also known as sensitivity or true positive rate) for each fold during cross-validation. The average recall score across the five folds is 0.7703. This suggests that, on average, the model correctly identifies approximately 77.03% of the actual cancellations.Overall, based on the confusion matrix and cross-validation scores, the cancellation prediction model shows a reasonable ability to identify cancellations.

Conclusion, Deployment of Model and Suggestions

Deployment (Technical suggestions for the hotel)

- Integration with the Booking System: Integrate the prediction model with the hotel's booking system. This can involve creating an API endpoint or a separate microservice that accepts booking information as input and returns the cancellation prediction as output. The booking system can then utilize this prediction to make informed decisions and manage resources effectively.
- Data Preprocessing and Feature Engineering: Implement the necessary data preprocessing steps and feature engineering techniques used during model training. This ensures that the input

data provided to the model during prediction is transformed and encoded appropriately to generate accurate predictions.

- Monitoring and Maintenance: Establish a monitoring system to track the model's performance and detect any potential issues or changes in data patterns. Regularly review the model's performance metrics and retrain the model periodically using new data to maintain its accuracy and relevance over time.
- Model Versioning and Rollbacks: Implement a version control system to manage different iterations of the model. This enables easy rollbacks to previous versions in case of any unforeseen issues or performance degradation with newer versions.
- Documentation and Communication: Prepare comprehensive documentation outlining the model's architecture, data preprocessing steps, and usage instructions. Share this documentation with relevant stakeholders to ensure transparency and effective communication regarding the model's capabilities and limitations.
- Security and Privacy: Implement appropriate security measures to protect sensitive data used by the model. Ensure compliance with privacy regulations and best practices to safeguard customer information.
- Performance Optimization: Continuously monitor and optimize the model's performance to enhance its speed and efficiency. Explore techniques like model compression, parallel processing, or hardware acceleration to improve the prediction speed, especially if large-scale predictions are required.
- User Training and Support: Provide training and support to the users who will interact with the model. This can include educating them on interpreting the model's predictions, understanding the factors that contribute to cancellations, and how to utilize the predictions effectively in their decision-making processes.

Business Suggestions based on the Random Forest Cancellation Prediction Model:

- Early Cancellation Identification: Utilize the predictions from the random forest model to identify bookings that are more likely to be canceled. By identifying these bookings early on, the hotel can take proactive measures such as offering incentives, personalized communication, or flexible cancellation policies to encourage guests to retain their bookings.
- Resource Allocation: The model can assist in optimizing resource allocation by considering the cancellation probabilities. For example, if a booking is predicted to have a high cancellation probability, the hotel can prioritize room assignments or allocate fewer staff resources to that particular booking until the cancellation risk decreases.
- Revenue Management: Incorporate the cancellation predictions into revenue management strategies. Adjust pricing or room availability based on the cancellation probabilities to maximize revenue and occupancy rates. For bookings with a high cancellation probability, the hotel can implement dynamic pricing strategies or offer promotions to attract other potential guests.
- Customer Retention: Leverage the cancellation predictions to focus on customer retention efforts. Identify guests with a high cancellation probability and personalize their experience by offering incentives, upgrades, or additional services to enhance their satisfaction and loyalty.
- Marketing Campaigns: Tailor marketing campaigns to target potential guests who are less likely to cancel their bookings. Utilize the model's insights to identify key demographics or customer segments that exhibit lower cancellation rates and develop targeted marketing strategies to attract more guests from those segments.

- Forecasting and Planning: Use the cancellation predictions to improve forecasting accuracy and planning. Incorporate the cancellation probabilities into demand forecasting models to better estimate future room availability, occupancy rates, and revenue projections.
- Operational Efficiency: The cancellation predictions can help optimize operational efficiency. For example, if a booking is highly likely to be canceled, the hotel can streamline check-in processes or allocate fewer resources for room preparation, resulting in cost savings and improved operational efficiency.
- Customer Communication: Improve customer communication and engagement by leveraging the cancellation predictions. Communicate with guests who have a high cancellation probability to address any concerns, provide relevant information, or offer personalized recommendations that increase the likelihood of them retaining their bookings.
- Data-Driven Insights: Continuously analyze the cancellation patterns and insights generated by the random forest model. Identify trends, factors, or common attributes associated with cancellations to gain a deeper understanding of guest behavior and preferences. Incorporate these insights into marketing strategies, service enhancements, or operational improvements.
- Continuous Model Improvement: Regularly update and retrain the random forest model with new data to improve its accuracy and adaptability. Monitor its performance and gather feedback from hotel staff to identify areas of improvement and potential enhancements to the model.
- In a business case where the cancellation prediction model is deployed, the understanding of True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP) has important implications. Here's how they can be interpreted in a business context:

True Negative (TN):
TN represents the instances that are correctly identified as non-cancellations by the model. In a business scenario, TN indicates the number of bookings that are predicted as non-cancellations and indeed turn out to be non-cancellations. These are the successful predictions where the model avoids unnecessary interventions or precautions for bookings that are likely to be honored.

False Positive (FP):
FP refers to the instances that are incorrectly predicted as cancellations by the model, but in reality, they are non-cancellations. In a business context, FP represents bookings that are flagged as potential cancellations but actually do not get canceled. This may result in unnecessary actions such as reserving additional resources or taking preventive measures, which can lead to additional costs or inconvenience for the business.

False Negative (FN):
FN represents the instances that are incorrectly predicted as non-cancellations by the model, but they actually turn out to be cancellations. In a business scenario, FN indicates bookings that are not flagged as potential cancellations but end up being canceled. This can lead to missed opportunities for proactive actions such as offering incentives or alternative arrangements to retain the bookings and minimize the impact of cancellations.

True Positive (TP):
TP represents the instances that are correctly identified as cancellations by the model. In a business context, TP indicates bookings that are predicted as cancellations and indeed get canceled. These are the successful predictions where the model accurately identifies potential cancellations, allowing the business to take appropriate actions such as reallocating resources, adjusting inventory, or contacting customers to manage the impact of cancellations.

Marketing Suggestions

- Targeted Marketing Campaigns: Create targeted marketing campaigns tailored to different market segments and customer types. Develop personalized messages and offers that resonate with each segment's preferences and needs. For example, for the transient segment, focus on offering exclusive deals and promotions to attract spontaneous bookings.
- Loyalty Programs: Implement loyalty programs to incentivize repeat bookings and customer loyalty. Offer rewards, discounts, or exclusive benefits to frequent customers or members of specific market segments. This can help increase customer retention and encourage repeat business.
- Customized Packages: Develop customized packages or experiences for specific market segments or customer types. For instance, create family-friendly packages for groups or special packages for corporate customers. Tailor the offerings to cater to the unique preferences and requirements of each segment.
- Personalized Communication: Leverage customer data and segmentation insights to deliver personalized communication. Use customer preferences, booking history, and demographic information to send targeted emails, newsletters, or promotional materials. Personalized communication enhances engagement and improves the likelihood of bookings.
- Partnerships and Collaborations: Form partnerships with relevant businesses or organizations to enhance the value proposition for specific market segments. For example, collaborate with local attractions or tour operators to offer bundled packages or exclusive discounts. This can attract customers from different segments and create cross-promotion opportunities.
- Online Presence and Reviews: Maintain a strong online presence and actively manage customer reviews and feedback. Encourage satisfied customers to leave positive reviews and respond promptly to any negative feedback. Positive online reviews can influence potential customers' decisions and help build trust in the brand.
- Social Media Engagement: Engage with customers through social media platforms to build brand awareness and foster a sense of community. Share engaging content, respond to customer inquiries, and run social media contests or giveaways. Social media can be a powerful tool to connect with customers, especially in the transient and social segment.
- Customer Service Excellence: Focus on providing exceptional customer service throughout the booking process and during the stay. Train staff to be attentive, friendly, and responsive to customer needs. Positive customer experiences can lead to repeat bookings and positive word-of-mouth recommendations.

By implementing these marketing suggestions, hotels can better target their audience, improve customer satisfaction, and drive bookings from various market segments and customer types. It is essential to continuously analyze market trends, monitor customer feedback, and adapt marketing strategies to stay competitive and meet evolving customer demands.

Avoiding High Cancellation Rate:

- Early Communication: Reach out to customers who have a higher likelihood of canceling as early as possible. Send personalized emails or messages to understand their concerns or reasons for

potential cancellation. Offer assistance, alternative options, or incentives to encourage them to keep their booking.

- Flexible Booking Policies: Implement flexible booking policies that allow customers to modify or cancel their reservations with ease. This can help reduce the likelihood of cancellations as customers feel more confident and comfortable knowing they have the flexibility to adjust their plans if needed.
- Special Offers and Incentives: Provide exclusive offers or incentives to customers who are at risk of canceling. Offer discounts, complimentary upgrades, or additional services to entice them to maintain their booking. These incentives can make them reconsider their decision and increase the chances of them keeping their reservation.
- Personalized Retention Strategies: Develop personalized retention strategies for customers who have a history of cancellations. Analyze their booking patterns, preferences, and feedback to understand their specific needs and concerns. Tailor special offers, benefits, or communication to address their individual concerns and provide a more personalized experience.
- Remarketing and Follow-ups: Implement remarketing campaigns to re-engage customers who have previously canceled bookings. Utilize targeted ads, email campaigns, or personalized offers to remind them of the value and benefits of staying at your hotel. Regular follow-ups and gentle reminders can help regain their interest and secure their booking.
- Proactive Customer Service: Provide exceptional customer service throughout the customer journey. Be proactive in addressing any concerns, answering inquiries, and resolving issues promptly. By demonstrating a commitment to customer satisfaction, you can build trust and loyalty, reducing the likelihood of cancellations.
- Analyze Booking Patterns: Continuously analyze booking patterns, cancellation trends, and customer feedback to identify patterns or common reasons for cancellations. Use this data to refine marketing strategies, improve service offerings, and address any pain points that may contribute to cancellations.
- Targeted Upselling and Add-ons: Offer targeted upselling opportunities and additional services to customers who are at risk of canceling. Present them with enticing options or upgrades that enhance their experience and make them reconsider their decision to cancel. This can increase the value of their reservation and reduce the likelihood of cancellation.

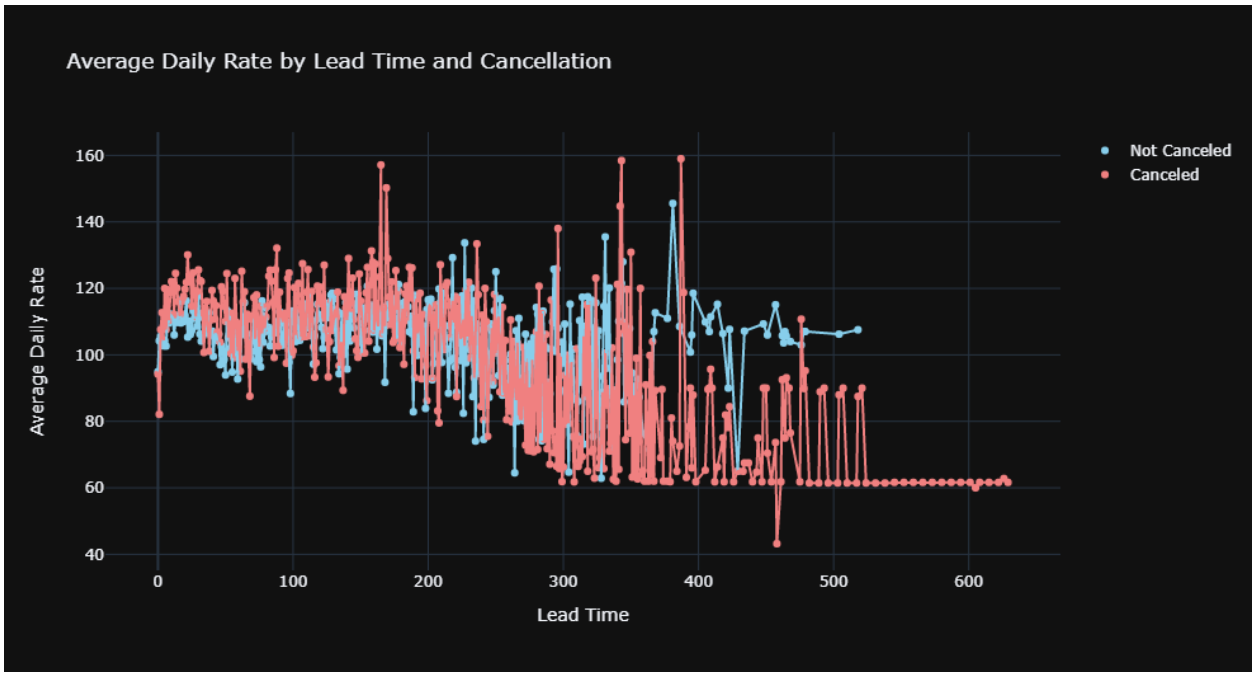
Appendix

MetaData

Variable Name	Meaning	Data Type
ADR	Average Daily Rate	**float64**
Adults	Number of adults	**int64**
Agent	ID of the travel agency	**object**
ArrivalDateDayOfMonth	Day of the month of the arrival date	**int64**
ArrivalDateMonth	Month of arrival date	**object**
ArrivalDateWeekNumber	Week number of the arrival date	**int64**
ArrivalDateYear	Year of the arrival date	**int64**
AssignedRoomType	Type of assigned room	**object**
Babies	Number of babies	**int64**
BookingChanges	Number of changes/amendments to the booking	**int64**
Children	Number of children	**int64**
Company	ID of the company/entity	**object**
Country	Country of origin	**object**
CustomerType	Type of booking	**object**
DaysInWaitingList	Number of days in the waiting list	**int64**

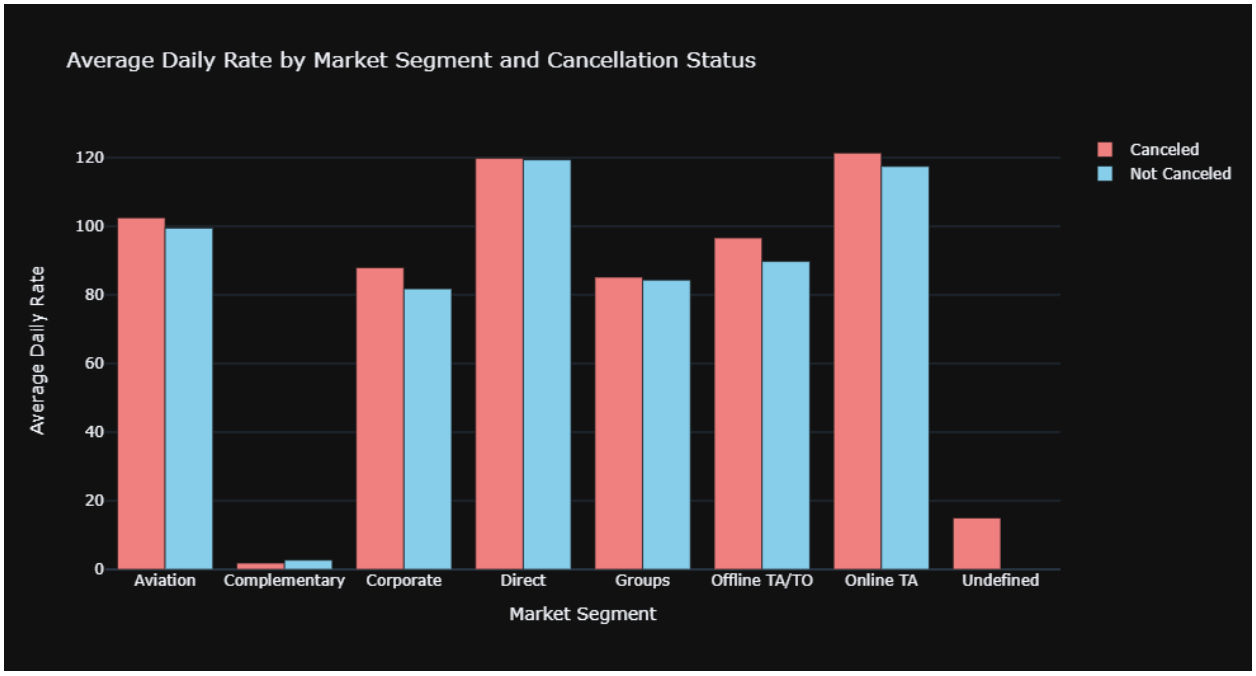
DepositType	Indication if a deposit was made	**object**
DistributionChannel	Booking distribution channel	**object**
IsCanceled	Value indicating if the booking was canceled	**int64**
IsRepeatedGuest	Value indicating if the booking is a repeat	**int64**
LeadTime	Number of days between booking and arrival	**int64**
MarketSegment	Market segment designation	**object**
Meal	Type of meal booked	**object**
PreviousBookingsNotCanceled	Number of previous bookings not canceled	**int64**
PreviousCancellations	Number of previous bookings canceled	**int64**
RequiredCarParkingSpaces	Number of car parking spaces required	**int64**
ReservationStatus	Reservation last status	**object**
ReservationStatusDate	Date of the last reservation status	**object**
ReservedRoomType	Type of reserved room	**object**
StaysInWeekendNights	Number of weekend nights stayed	**int64**
StaysInWeekNights	Number of weeknights stayed	**int64**
TotalOfSpecialRequests	Number of special requests made	**int64**

Average daily rate by lead time and cancellation rate



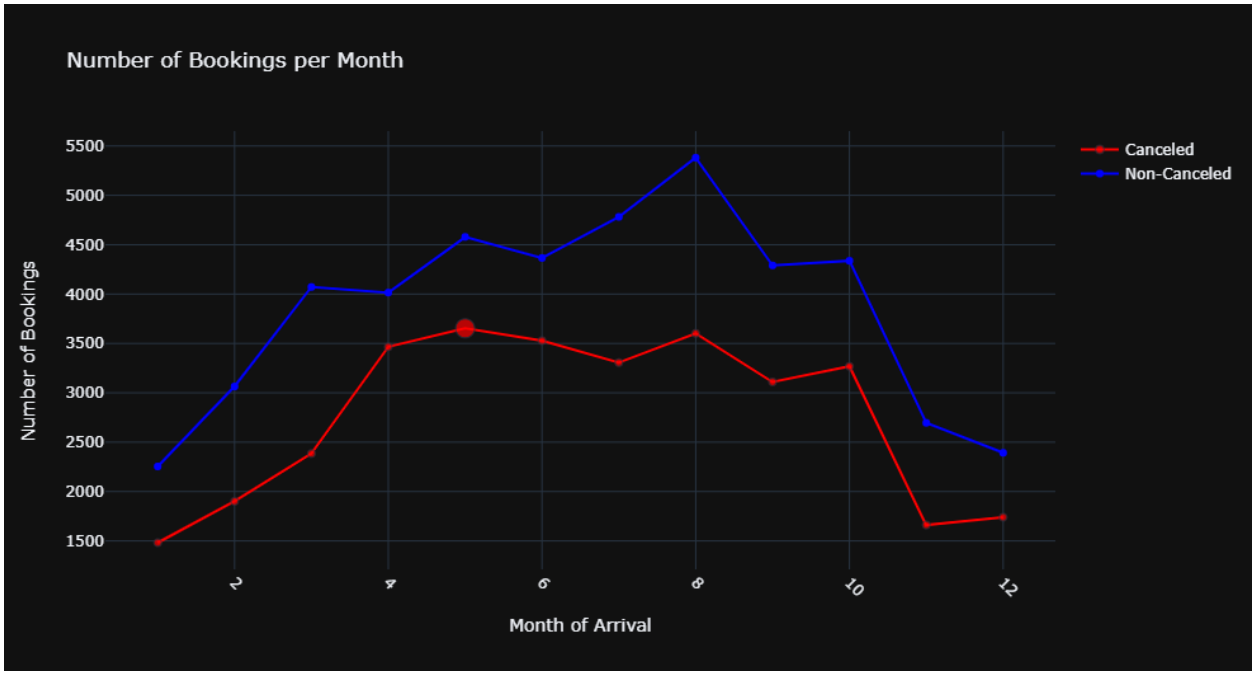
The visualization of the average daily rate by lead time and cancellation rate provides valuable insights into the relationship between these variables. By plotting the average daily rate against the lead time, we can observe how the pricing fluctuates as the lead time increases or decreases. This visualization helps us understand the pricing dynamics based on the time of booking and how it impacts the average daily rate.Additionally, incorporating the cancellation rate into the visualization allows us to identify any patterns or trends. We can observe if there is a correlation between the lead time, average daily rate, and the likelihood of cancellations. It helps us understand if longer lead times are associated with higher cancellation rates and how this affects the pricing strategy.By visualizing the average daily rate by lead time and cancellation rate, we can make data-driven decisions regarding pricing strategies, such as offering incentives or adjusting rates based on lead time, to optimize revenue and minimize cancellations. It provides a comprehensive view of the relationship between these variables and guides us in developing effective pricing and booking policies to maximize customer satisfaction and profitability.

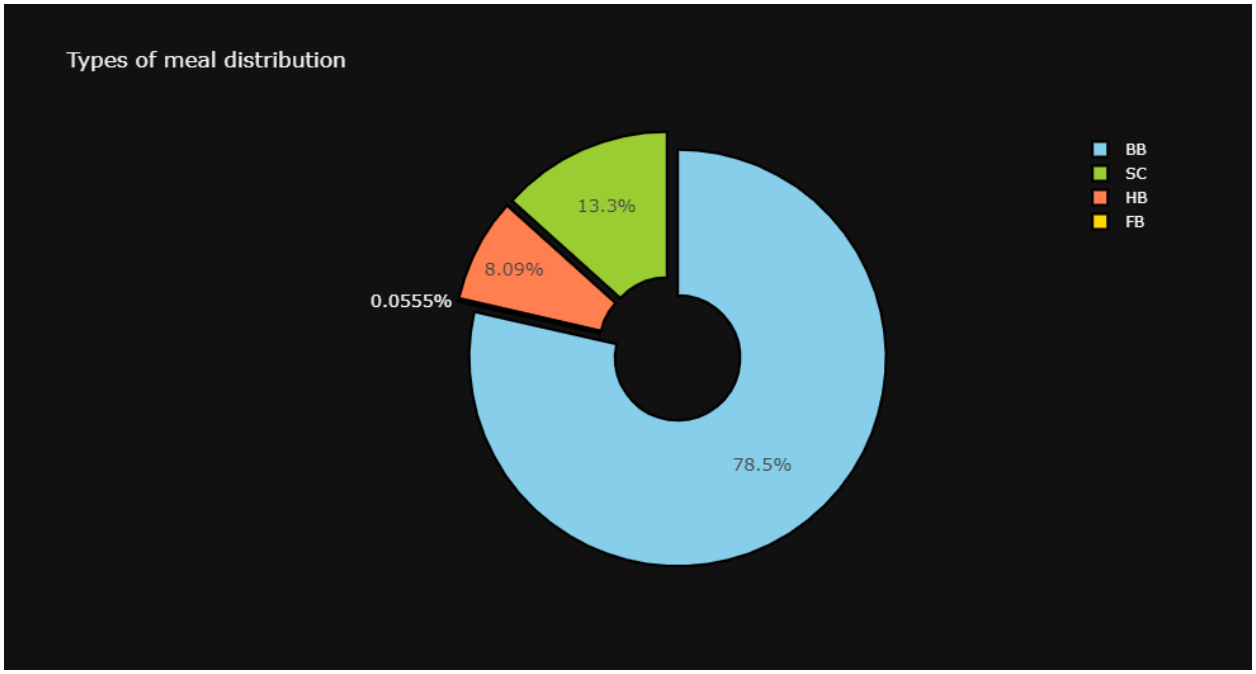
The average daily rate by market segment and cancellation rate

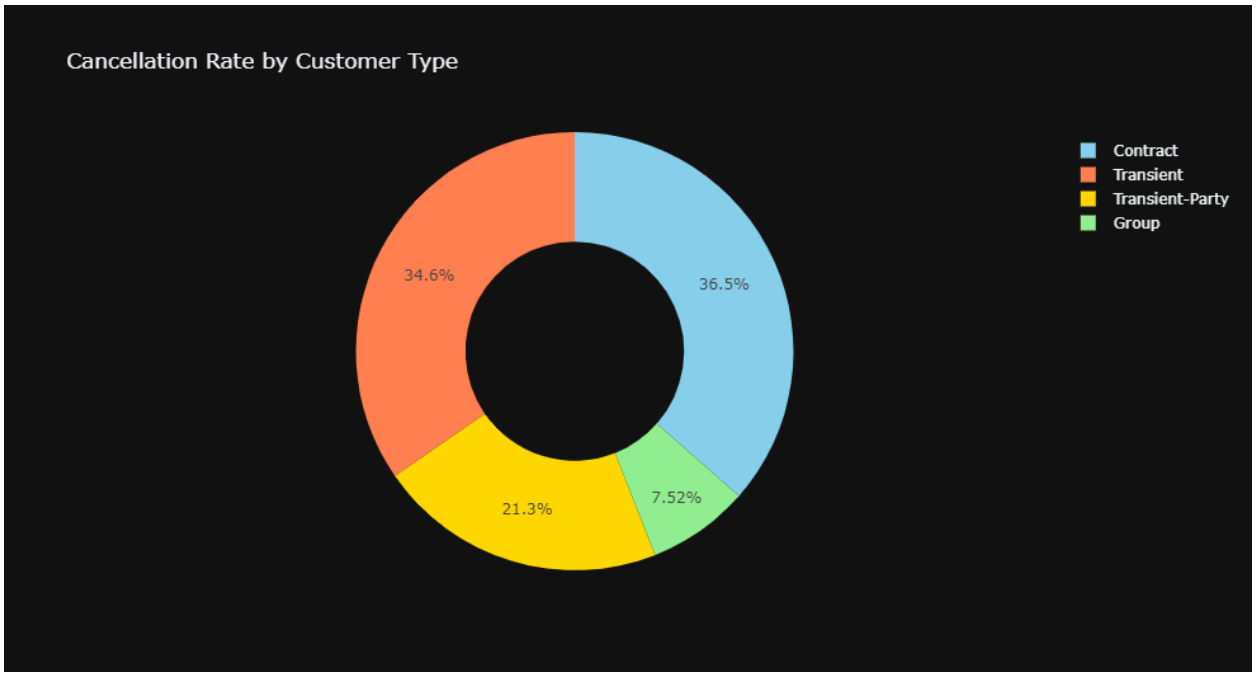


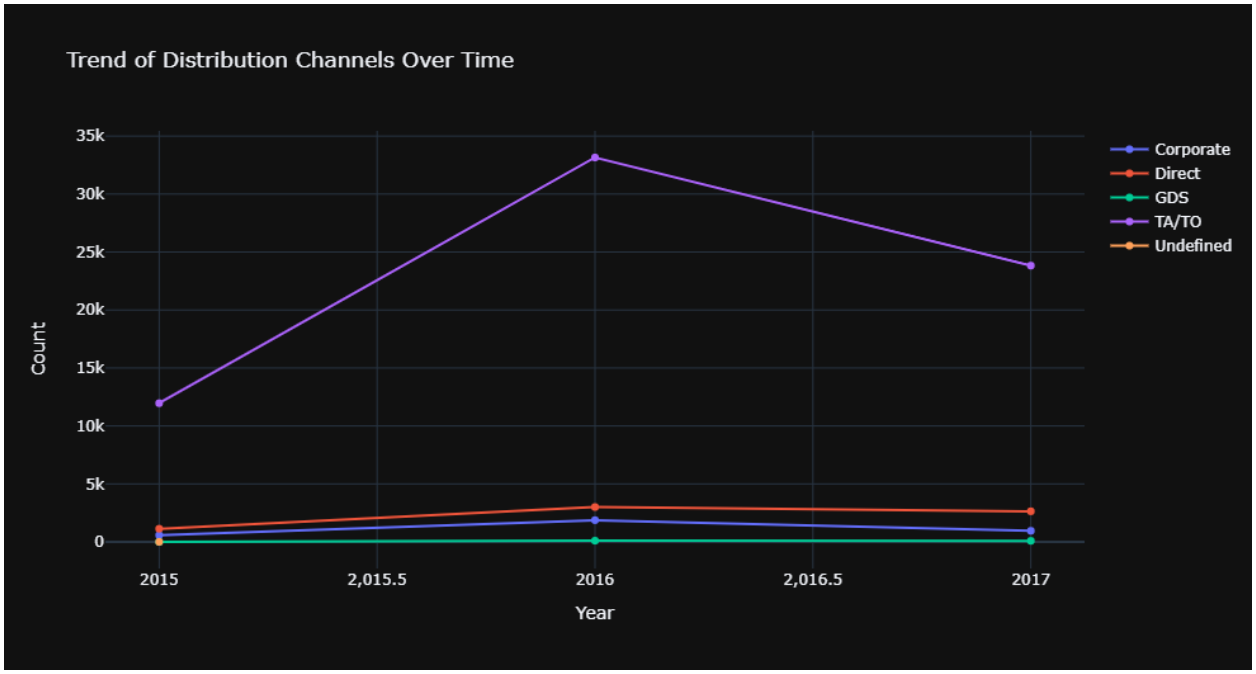
The average daily rate by market segment and cancellation rate visualization provides valuable insights into the relationship between these variables. By analyzing the average daily rate and cancellation behavior across different market segments, we can make informed decisions and tailor our strategies accordingly.

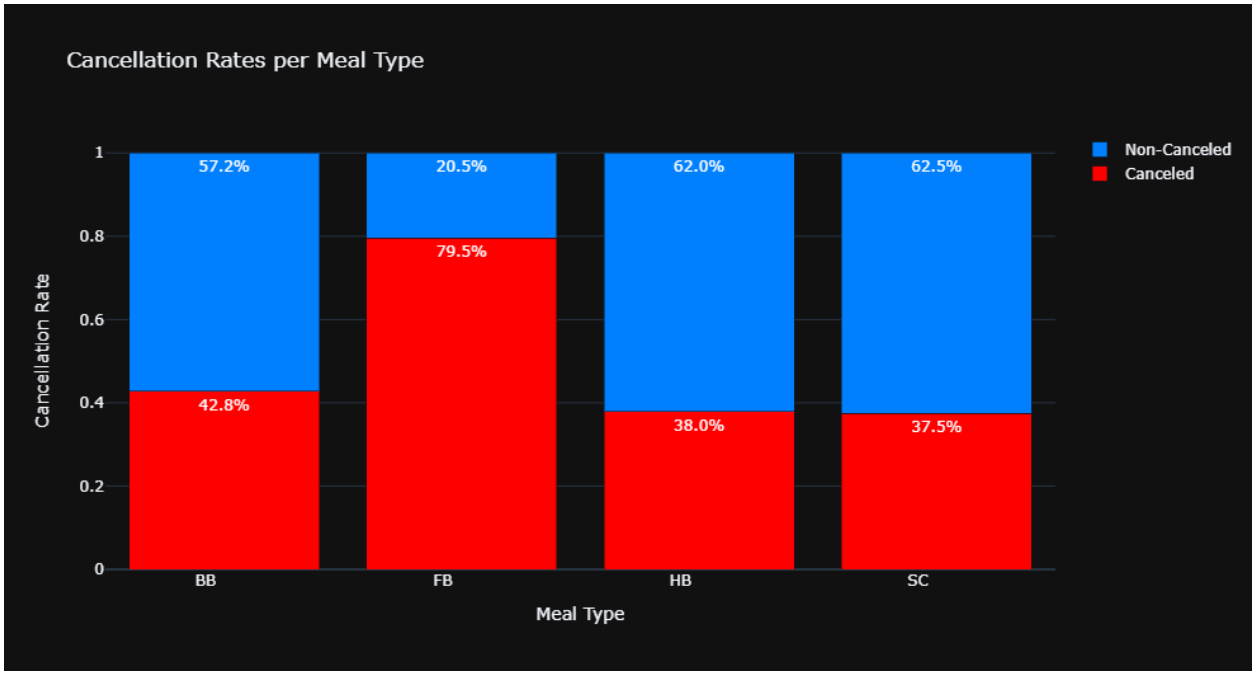
The visualization reveals that the direct market segment and the online travel agency (OTA) segment have distinct patterns in terms of cancellation rate and average daily rate. The direct segment shows a higher average daily rate, indicating that customers booking directly with the hotel are willing to pay more for their reservations. This higher price point may contribute to a lower cancellation rate in this segment.On the other hand, the OTA segment displays a lower average daily rate but a higher cancellation rate. This suggests that customers booking through online travel agencies may be more price-sensitive and flexible with their travel plans, leading to a higher likelihood of cancellations. Understanding this behavior allows us to devise strategies specific to the OTA segment, such as implementing effective communication strategies or offering incentives to minimize cancellations.Analyzing the average daily rate by market segment and cancellation rate empowers us to make data-driven decisions in marketing and revenue management. By targeting our efforts towards specific market segments and understanding their preferences and behaviors, we can optimize revenue generation and reduce the impact of cancellations on the hotel

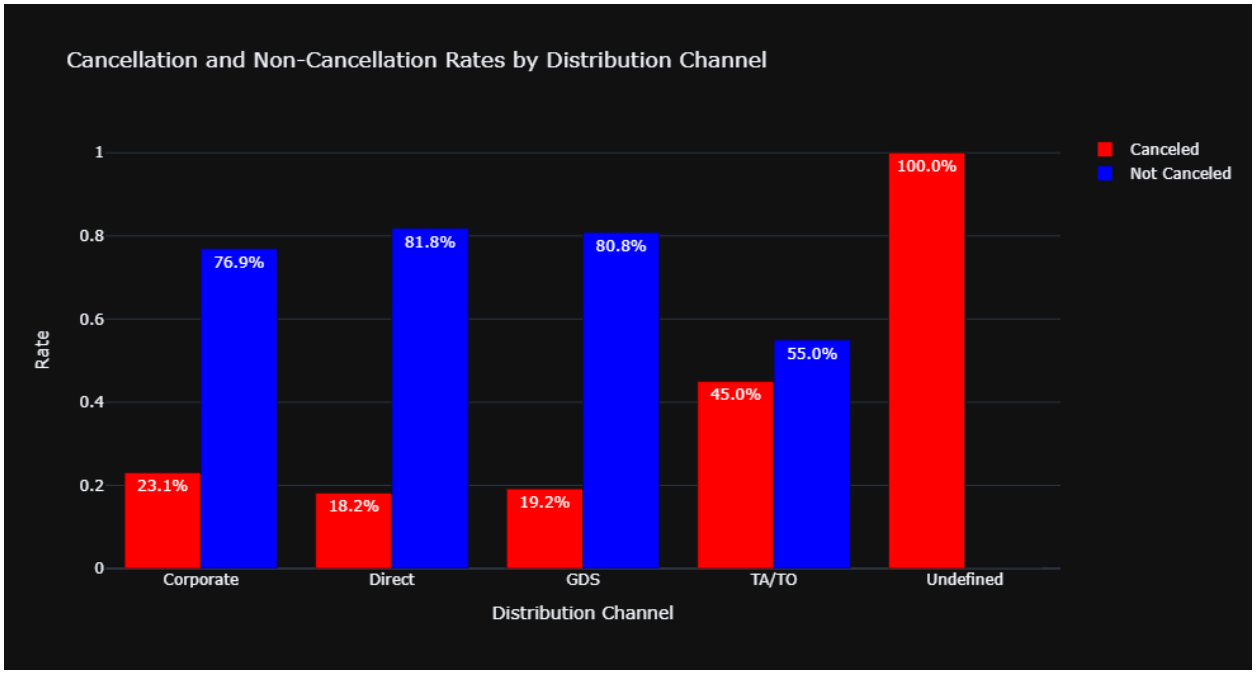


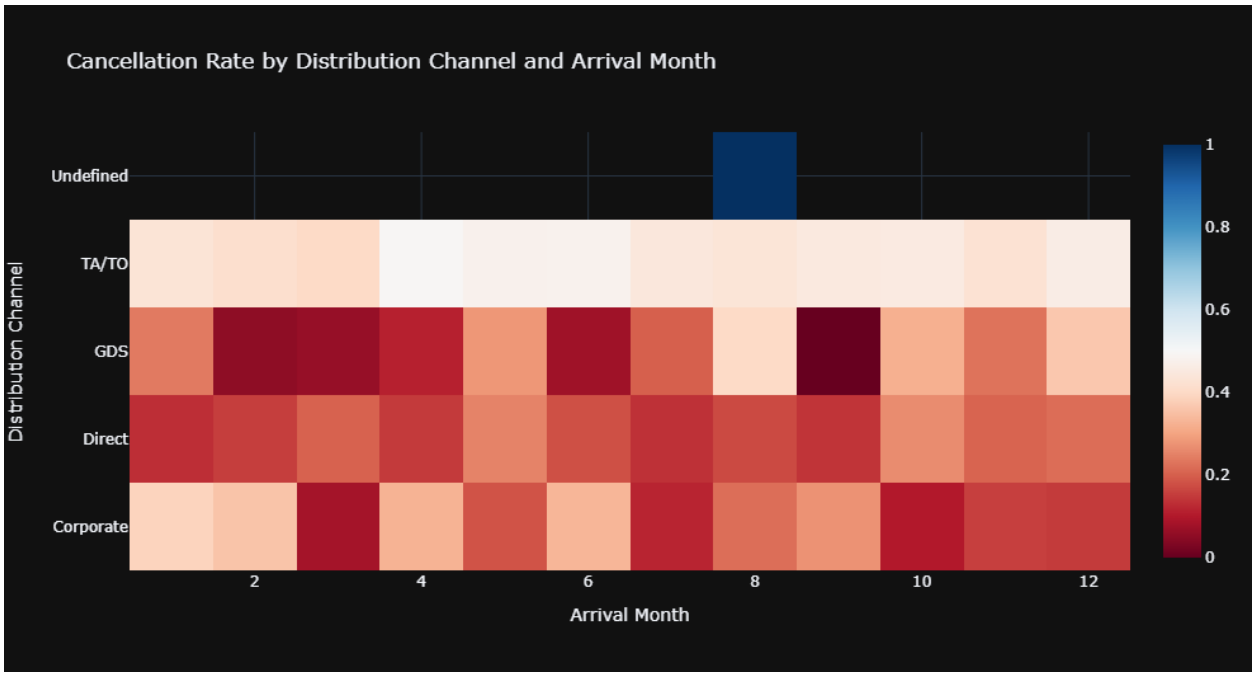

Types of meal distribution
BB
SC
HB
FB
13.3%
8.09%
0.0555%
78.5%


Cancellation Rate by Customer Type
Contract
Transient
Transient-Party
Group
34.6%
36.5%
21.3%
7.52%


Trend of Distribution Channels Over Time
35k
30k
25k
20k
15k
10k
5k
0
Count
2015
2,015.5
2016
2,016.5
2017
Year
Corporate
Direct
GDS
TA/TO
Undefined


Cancellation Rates per Meal Type
1
0.8
0.6
0.4
0.2
0
Cancellation Rate
57.2%
42.8%
20.5%
79.5%
62.0%
38.0%
62.5%
37.5%
BB
FB
HB
SC
Meal Type
Non-Canceled
Canceled


Cancellation and Non-Cancellation Rates by Distribution Channel
Canceled
Not Canceled
Rate
1
0.8
0.6
0.4
0.2
0
23.1%
76.9%
18.2%
81.8%
19.2%
80.8%
45.0%
55.0%
100.0%
Corporate
Direct
GDS
TA/TO
Undefined
Distribution Channel


Cancellation Rate by Distribution Channel and Arrival Month
Distribution Channel
Undefined
TA/TO
GDS
Direct
Corporate
2
4
6
8
10
12
Arrival Month
1
0.8
0.6
0.4
0.2
0