HOTEL BOOKING DEMAND DATASETS

Data Analysis Report and Predicting of Booking Status

May **2023** // Prepared by Muhannad Mansour Felemban

CONTEXT

Hotel booking cancellations can negatively impact revenue and profits by Additional costs of distribution channels or Loss of resources etc. By identifying the factors that lead to reservation cancellations, hotels can design targeted interventions to reduce cancellations and increase revenue.

In this project, I well analyze reservation data for Hotels, a chain of luxury hotels. I perform **Exploratory and Explanatory Data Analysis** to understand the properties of the data, identify trends and patterns of reservation cancellations. I then build **Machine Learning Algorithm (Classification models)** for Predicting Booking Status into cancellations vs non-cancellations. These models can be used to predict the likelihood of cancellation for new reservations and allow the hotels to take actions accordingly.

The data contains information on over 40,000 reservations made from 2017 to 2018 years. For each reservation, we have information such as arrival/departure dates, reservation status ,avg price per room , room type reserved, etc. We find that certain dates, hotels, and customer segments have higher tendencies to cancel reservations. These insights can help Hotels craft targeted policies to prevent cancellations and improve customer experiences.

Data Description:

no_of_adults: Number of adults

no_of_children: Number of Children

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

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

type_of_meal_plan: Type of meal plan booked by the customer:
0 – No meal plan
1 – Breakfast
2 – Half board (breakfast and one other meal)
3 – Full board (breakfast, lunch, and dinner)

required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)

room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by Star Hotels.

lead_time: Number of days between the date of booking and the arrival date

arrival_year: Year of arrival date

arrival_month: Month of arrival date

arrival_date: Date of the month

market_segment_type: Market segment designation.

repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)

no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking

no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking

avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)

no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)

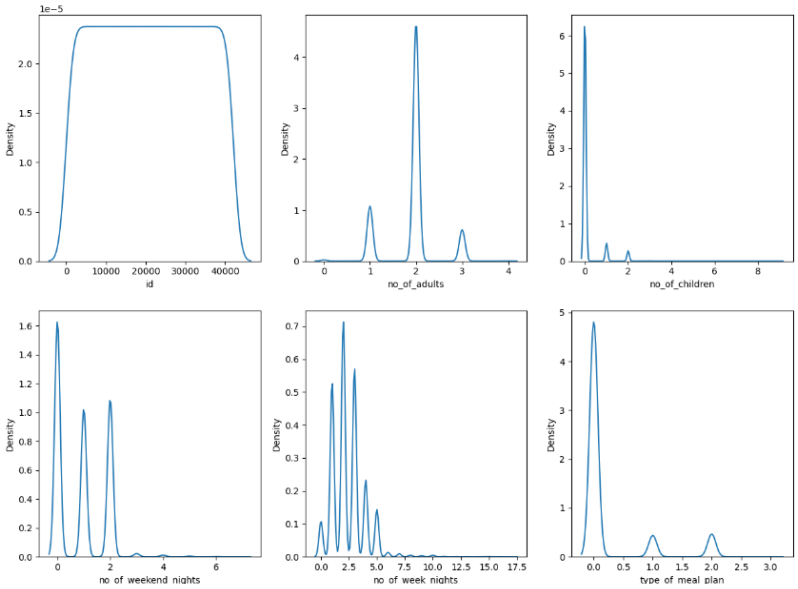
booking_status: Flag indicating if the booking was canceled or not (0 - No, 1 - Yes).

Exploratory Data Analysis :

Statistical description:

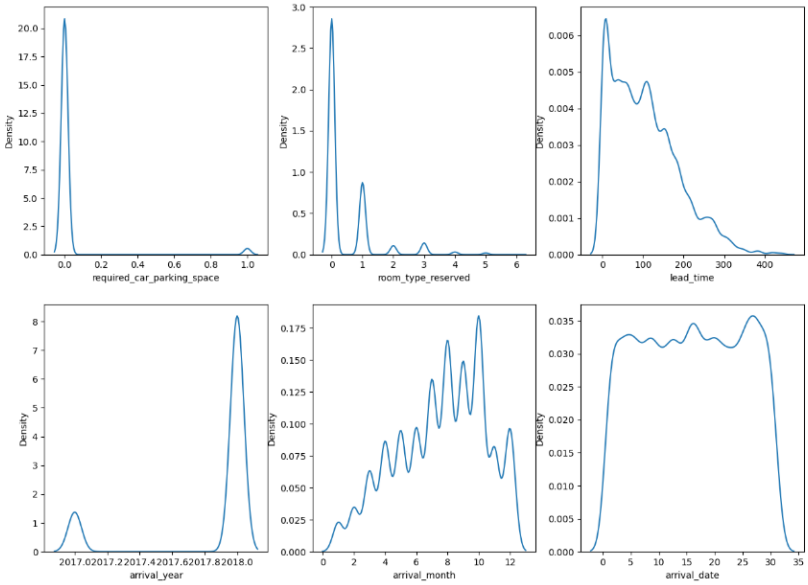
- "**no_of_adults**," the mean is 1.92 with a standard deviation of 0.52. The minimum value is 0, indicating that there may be bookings with no adults. The 25% is 2, the median is 2, and the 75% is 2, indicating that most bookings have 1 or 2 adults, with very few bookings having 3 or 4 adults. The distribution of this variable appears to be roughly symmetrical with no apparent outliers.
- "**no_of_children**," the mean is 0.14 with a standard deviation of 0.45. The minimum value is 0, indicating that many bookings do not include children. The 25% is also 0, the median is 0, and the 75% is 0, indicating that most bookings do not include children, with very few bookings having more than 1 child. The distribution of this variable appears to be heavily skewed to the right, with potential outliers, as indicated by the maximum value of 9.
- "**no_of_weekend_nights**," the mean is 0.88 with a standard deviation of 0.89. The minimum value is 0, indicating that some bookings did not include weekend nights. The 25% is 0, the median is 1, and the 75% is 2, suggesting that most bookings included either 0, 1, or 2 weekend nights. The distribution of this variable appears to be skewed to the right, with potential outliers reflected in the maximum value of 7.
- "**no_of_week_nights**," the mean is 2.40 with a standard deviation of 1.43. The minimum value is 0, indicating that some bookings did not include any weekday nights. The 25% is 1, the median is 2, and the 75t%is 3, suggesting that most bookings included either 1, 2, or 3 weekday nights. The distribution of this variable appears to be roughly symmetrical, with no apparent outliers.
- "**type_of_meal_plan**," the mean is 0.24 with a standard deviation of 0.59. The minimum value is 0, indicating that many bookings did not include any meal plan. The 25% is also 0, the median is 0, and the 75% is 0, indicating that most bookings did not include any meal plan. The maximum value is 3, which may indicate the presence of outliers. The distribution of this variable appears to be heavily skewed to the right.

	count	mean	std	min	25%	50%	75%	max
no_of_adults	42100.0	1.920713	0.524950	0.0	2.0	2.0	2.0	4.0
no_of_children	42100.0	0.141093	0.450128	0.0	0.0	0.0	0.0	9.0
no_of_weekend_nights	42100.0	0.884632	0.885693	0.0	0.0	1.0	2.0	7.0
no_of_week_nights	42100.0	2.398005	1.427330	0.0	1.0	2.0	3.0	17.0
type_of_meal_plan	42100.0	0.239192	0.587674	0.0	0.0	0.0	0.0	3.0



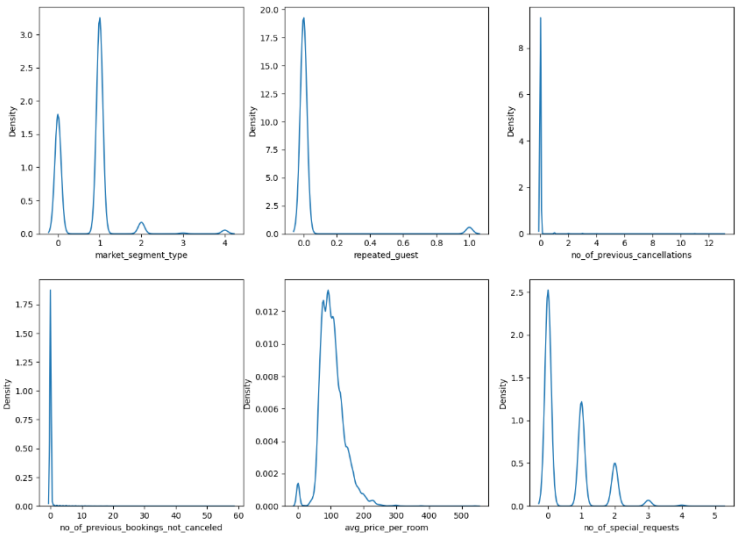
- "**required_car_parking_space**," the mean is 0.03 with a standard deviation of 0.16. The minimum value is 0, indicating that many bookings did not require any car parking space. The 25% is 0, the median is 0, and the 75% is 0, indicating that most bookings did not require any car parking space. The maximum value is 1, which suggests the presence of outliers. The distribution of this variable skewed to the right.
- "**room_type_reserved**," the mean is 0.43 with a standard deviation of 0.83. The minimum value is 0, indicating that many bookings did not specify a room type. The 25% is also 0, the median is 0, and the 75% is 1, indicating that most bookings specified either a standard ora superior room type. The maximum value is 6, which may indicate the presence of outliers. The distribution of this variable appears to be heavily skewed to the right.
- "**lead_time**," the mean is 103.89 with a standard deviation of 81.07. The minimum value is 0, indicating some bookings were made on the same day as arrival. The 25% is 37, the median is 93, and the 75% is 155, indicating that most bookings were made between 1 to 6 months prior to arrival. The maximum value is 443, which may indicate the presence of outliers.
- "**arrival_year**," the mean is 2017.86 with a standard deviation of 0.35. The minimum and maximum values are both 2017 and 2018, respectively, indicating that all bookings occurred within a two-year period. There is no apparent skewness or outliers in this variable.
- "**arrival_month**," the mean is 7.59 with a standard deviation of 2.83. The minimum value is 1, indicating that bookings occurred in all months of the year. The 25% is 6, the median is 8, and the 75% is 10, indicating that most bookings occurred during the summer or Fall months. There is no apparent skewness, but there may be some outliers in the higher values.
- "**arrival_date**," the mean is 15.90 with a standard deviation of 8.89. The minimum value is 1, indicating that there were bookings for every day of the month. The 25% is 8, the median is 16, and the 75%is 24, indicating that most bookings occurred between the 8th and 24th of the month. There is no apparent skewness or outliers in this variable.

	count	mean	std	min	25%	50%	75%	max
required_car_parking_space	42100.0	0.025249	0.156884	0.0	0.0	0.0	0.0	1.0
room_type_reserved	42100.0	0.428931	0.832605	0.0	0.0	0.0	1.0	6.0
lead_time	42100.0	103.888029	81.069343	0.0	37.0	93.0	155.0	443.0
arrival_year	42100.0	2017.856295	0.350795	2017.0	2018.0	2018.0	2018.0	2018.0
arrival_month	42100.0	7.593539	2.829395	1.0	6.0	8.0	10.0	12.0
arrival_date	42100.0	15.902945	8.888582	1.0	8.0	16.0	24.0	31.0



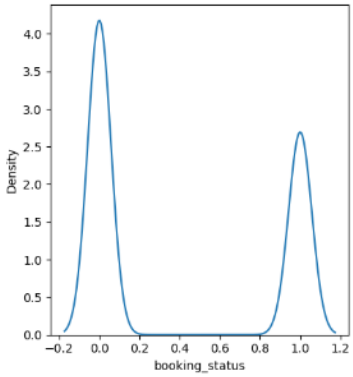
- "**market_segment_type**," the mean is 0.73 with a standard deviation of 0.63. The minimum value is 0, indicating that some bookings were not classified into any market segment. The 25% is 0, the median is 1, and the 75% is 1, indicating that most bookings were made through travel agents or tour operators. The maximum value is 4, which may indicate the presence of outliers with The distribution skewed to the right.
- "**repeated_guest**," the mean is 0.03 with a standard deviation of 0.17. The minimum value is 0, indicating that most bookings were made by first-time guests. The 25% is 0, the median is 0, and the 75% is 0, indicating that most bookings were made by first-time guests. The maximum value is 1, indicating that some bookings were madeby repeat guests.
- "**no_of_previous_cancellations**," the mean is 0.02 with a standard deviation of 0.33. The minimum value is 0, indicating that most bookings have not been cancelled before. The 25% is 0, the median is 0, and the 75% is 0, indicating that most bookings have not been cancelled before. The maximum value is 13, which may indicate the presence of outliers.
- "**no_of_previous_bookings_not_canceled**," the mean is 0.18 with a standard deviation of 1.73. The minimum value is 0, indicating that some bookings were made by first-time guests. The 25% is 0, the median is 0, and the 75% is 0, indicating that most bookings were made by first-time guests. The maximum value is 58, which may indicate the presence of outliers.
- "**avg_price_per_room**". the mean value of 104.57 and a standard deviation of 37.14. The minimum value is 0, indicating that in some cases, no price was charged for the room. The 25% is 80, the median value is 99.45, and the 75% is 123.3, indicating that most room prices fall within this range. The maximum value is 540, indicating that there are some rooms with a much higher price than the average. Overall, the distribution of this variable appears to be positively skewed, as the mean value is higher than the median value, and there are some high-priced rooms that skew the distribution to the right.
- "**no_of_special_requests**," the mean is 0.57 with a standard deviation of 0.78. The minimum value is 0, indicating that some bookings did not have any special requests. The 25% is 0, the median is 0, and the 75% is 1, indicating that most bookings had either no special requestsor only one special request. The maximum value is 5, which may indicate the presence of outliers with skewed to the right.

	count	mean	std	min	25%	50%	75%	max
market_segment_type	42100.0	0.728504	0.633529	0.0	0.0	1.00	1.0	4.0
repeated_guest	42100.0	0.029192	0.168347	0.0	0.0	0.00	0.0	1.0
no_of_previous_cancellations	42100.0	0.019715	0.325837	0.0	0.0	0.00	0.0	13.0
no_of_previous_bookings_not_canceled	42100.0	0.175772	1.732121	0.0	0.0	0.00	0.0	58.0
avg_price_per_room	42100.0	104.566377	37.139165	0.0	80.0	99.45	123.3	540.0
no_of_special_requests	42100.0	0.571734	0.775041	0.0	0.0	0.00	1.0	5.0



- "**booking_status**,"(Target variable) the mean is 0.39 with a standard deviation of 0.49. The minimum value is 0, indicating that some bookings were cancelled. The 25% is 0, the median is 0, and the 75% is 1, indicating that most bookings were not cancelled and were confirmed. There is no apparent skewness or outliers in this variable.

	count	mean	std	min	25%	50%	75%	max
booking_status	42100.0	0.392019	0.488207	0.0	0.0	0.0	1.0	1.0



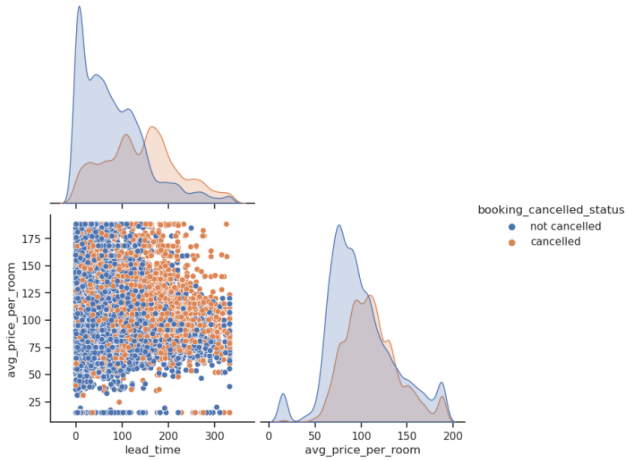
Correlation:

Here I've added four new features to Dataset
total nights = No. of weekend nights + No. of week nights
total cost = total nights * avg price per room
total bookings = No. of previous cancellations + No. of previous bookings not canceled
total families = No. of adults + No. of children

To improve the accuracy of these associations, outliers were removed from the data set. Outliers are data points that differ significantly from the majority of the data, which can skew the correlation. Here I used the IQR (Interquartile Range) method to identify and remove outliers from your data.

booking_status	1.000000
lead_time	0.377281
avg_price_per_room	0.163874
market_segment_type	0.148776
total_cost	0.141306
total_nights	0.067583
total_families	0.064336
no_of_week_nights	0.060340
no_of_weekend_nights	0.044368
arrival_month	0.007639
id	0.007264
arrival_date	0.003124
room_type_reserved	-0.015674
total_bookings	-0.081142
no_of_special_requests	-0.218567
no_of_adults	NaN
no_of_children	NaN
type_of_meal_plan	NaN
required_car_parking_space	NaN
arrival_year	NaN
repeated_guest	NaN
no_of_previous_cancellations	NaN
no_of_previous_bookings_not_canceled	NaN

- lead_time has a moderate positive correlation (0.377281) with booking_status, meaning that as lead time (time between booking and arrival) increases, there's a higher likelihood of the booking being canceled.
- avg_price_per_room has a weak positive correlation (0.163874) with booking_status, indicating that higher-priced rooms might be slightly more likely to be canceled.
- total_cost, market_segment_type, total_nights, no_of_week_nights, and no_of_weekend_nights also have weak positive correlations with booking_status, suggesting that these factors may contribute to cancellation likelihood to some extent.
- The NaN values in the data indicate that there's no correlation between those variables and booking_status, or the data is insufficient to calculate the correlation.

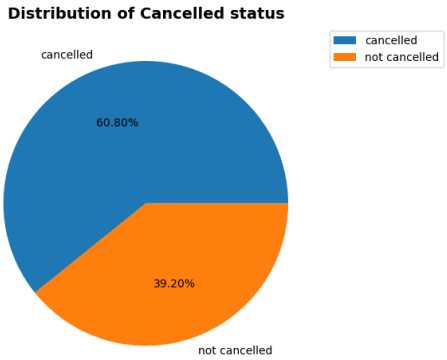


"**Note**" It is important to note that "**correlation does not imply causation"** and other factors may impact the relationship between variables. Nonetheless, these correlations provide valuable insights for understanding the patterns in the data and identifying potential areas forfurther investigation.

Explanatory Data Analysis :

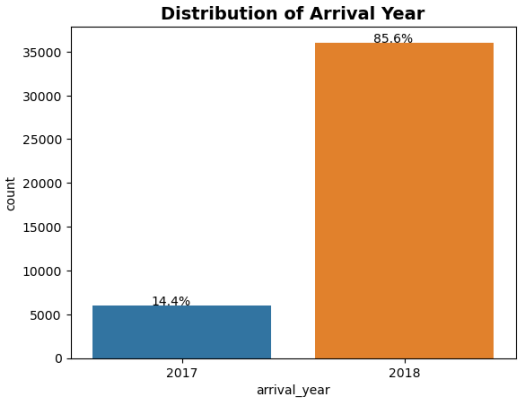
Cancelled status :

The percentage of bookings that were not cancelled was relatively high, at 60%, or about 25,596 bookings, while the cancellation rate was 40%, or about 16,504 of the total number of bookings. This could be an opportunity analyze the reasons behind the cancellations and work on addressing any issues that may be leading to high cancellation rates. It may also be useful to focus on customer retention strategies for the non-cancelled bookings, such as loyalty programs or customized promotions, to encourage more repeat bookings.



Arrival year:

- More bookings were made for the arrival year 2018 by about 36,050, or 86%, and in 2017, bookings were made much less, by about 6,050, or 14.4%. It is an excellent growth rate for bookings compared to the previous year

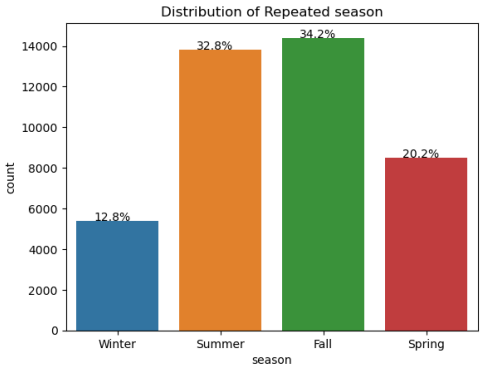


Arrival month:

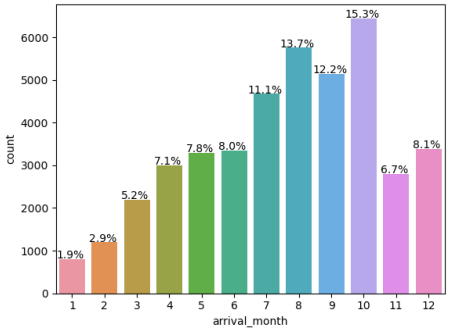
- More bookings are made for the arrival month during the months (August to October) from 5148 to 6453 (14% to 15.3%). Since the dataset contains entries from July 2017 to August 2018; July and August are counted for all three years while the remaining months are counted for two years only. This may also lead to an increase in the number of these months compared to the other months.

Season:

- It was noted that the largest growth rate of bookings in the seasons of the year was in the Summer by about 13800 and the Fallby 14404, i.e. an increase rate between these two seasons of 1.4%, as shown in the chart.

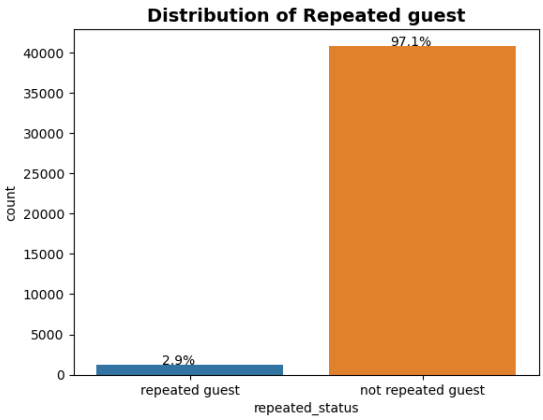


"**Note**" that these months are just a general guide and the actual weather and climate in hotel locations can vary from year to year. Additionally, certain months may have different weather patterns or events that make them more or less desirable for booking, so it's always a good idea to research the specific details of the season



Repeat guests:

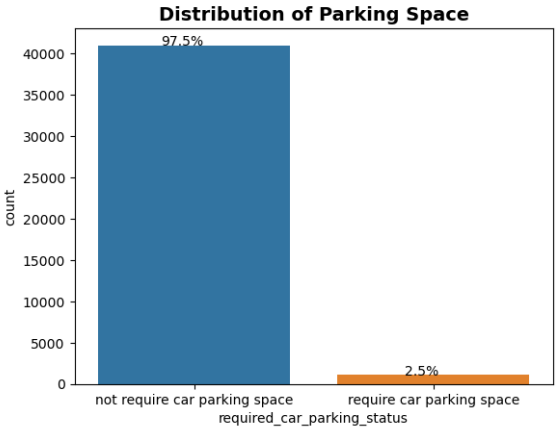
- The percentage of repeat guests was very small, 3%, or about 1,229 visitors, with the rate of not recurring guests being 97%, or about 40,871 of the total number of booking. This could be an opportunity for the hotel to focus on customer retention strategies, such as loyalty programs or customized promotions, to encourage more repeat visits or may be useful to investigate the characteristics of the guests who fall under the "repeated guest" category in order to identify promotional efforts targeted towards repeat or It may also be useful booking patterns and experiences repeated guests tend to book for longer durations room service or spa.



Parking Space:

- 97% of the guests who booked the hotel do not require parking car spaces. This may be an indication that the hotel is located in an area with good public transportation or walkability, or that guests prefer using ride-sharing services or taxis rather than renting a car.

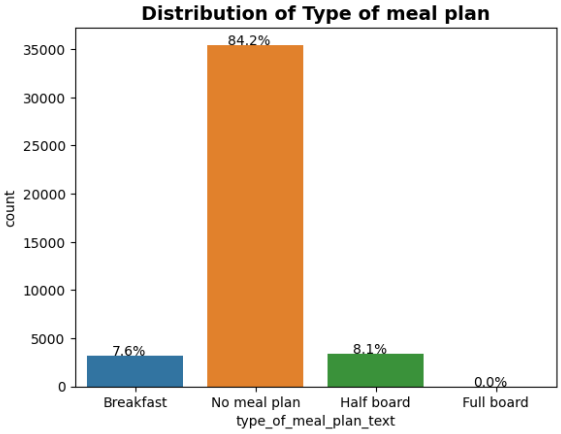
- A small percentage 2% of guests who booked the hotel required car parking space. This could be an opportunity for the hotel to focus on providing convenient and accessible parking options for these guests, such as valet parking or discounted rates at nearby



- it may be useful to track the "required_car_parking_status" variable over time to see if there are any changes in the proportion of guests who require car parking space. If the proportion of guests who require car parking space increases over time, it may suggest that the hotel needs to expand its parking options or adjust its pricing strategy. On the other hand, if the proportion of guests who require car parking space remains low, it may indicate that the hotel's location and accessibility are attractive to guests who prefer not to rent a car.

Type of meal plan:

- 84% of guests who booked the hotel did not select a meal plan. This could be an indication that the guests prefer to explore local restaurants or cafes, or that they have dietary restrictions that make it difficult to select a pre-set meal plan.



- It may be useful to investigate the characteristics of the guests who selected the different meal plan options in order to identify any patterns or trends. For example, do guests who select the "Half board" option tend to be families or couples on romantic getaways? Do guests who select the "Breakfast" option tend to be business travelers or solo travelers? This information could be used to inform targeted marketing.

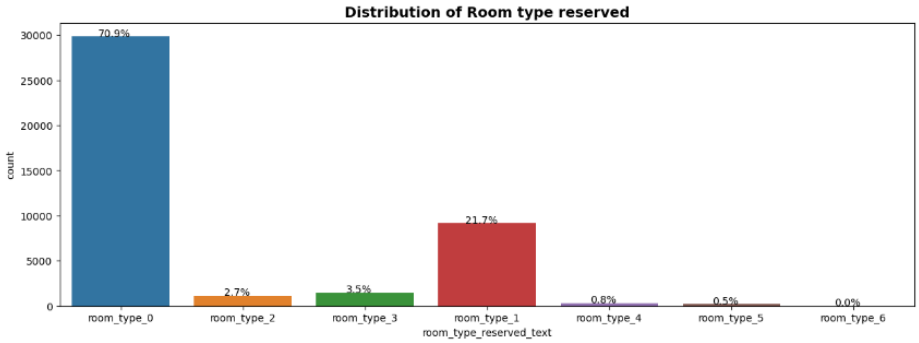
- The presence of the "Full board" option, even with a small count of 6 observations, may suggest that the hotel caters to guests who prefer an all-inclusive experience. It may be useful to investigate the reasons why only a small number of guests selected this option, such as whether the price is too high

Room type reserved:

- The high number of "room_type_0" bookings of 71% would indicate that this is the hotel's standard or default room type. It may be useful to investigate the characteristics and preferences of guests who have chosen this type of room to understand the factors driving their decisions.

- The "room_type_reserved" variable may be used to personalize the guest experience based on theirselected room type. For example, the hotel could offer customized recommendations for local attractions or services that are relevant to the guest's room type, such as nearby spas or outdoor activities for guests who reserved suites or rooms with balconies.

- it may be useful to compare the booking patterns and experiences of guests who selected different room types in order to identify any patterns or trends. For example, do guests who select larger or more luxurious room types tend to spend more on additional services such as room service

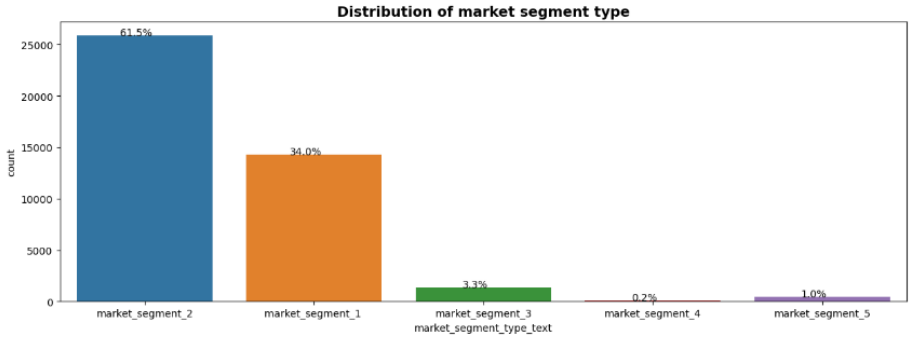


market segment type:

- The large percentage of "market_segment_2" 61% bookings would indicate that this is the most popular customer segment for a hotel. It can be helpful to investigate the characteristics and preferences of this customer group to understand the factors driving their decisions.

- The "market_segment_type" variable could be used to personalize the guest experience based on their selected market segment. For example, the hotel could offer customized recommendations for local attractions or services that are relevant to the guest's market segment, such as family-friendly activities for guests in the "market_segment_5" category or business services for guests in the "market_segment_1" category.

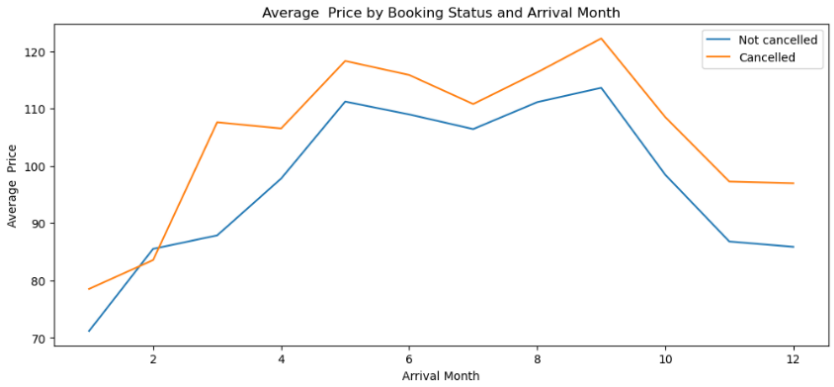
- Finally, it may be useful to compare the booking patterns and experiences of guests in different market segments in order to identify any patterns or trends. For example, do guests in the "market_segment_2" category tend to book for longer durations or spend more on additional services such as room service or spa treatments? This information could help the hotel identify opportunities to improve its offerings and attract more high-spending guests.



Average price by booking status and arrival month:

- For bookings that were not cancelled, the average price per room is lowest in the earlier months of the year (January to March), with a gradual increase from March to August, and a slight decrease in the later months of the year (November and December). The highest average price per room is 111.150985 in August.

- For bookings that were cancelled, the average price per room is also lowest in the earlier months of the year (January to February), with a significant increase in March and a peak in September. The highest average price per room is 122.277993 in September.



- These results suggest that there may be seasonal demand for hotel rooms for both not cancelled and cancelled bookings, with higher demand during the summer months for not cancelled bookings and higher demand during the spring and summer months for cancelled bookings. Hotels could use these insights to plan their pricing strategies and promotions to optimize their revenue and reduce the number of cancelled bookings.

Number of weekend nights Vs booking status :

The number 0 in the "no_of_weekend_nights" indicates bookings where no weekend nights were booked.

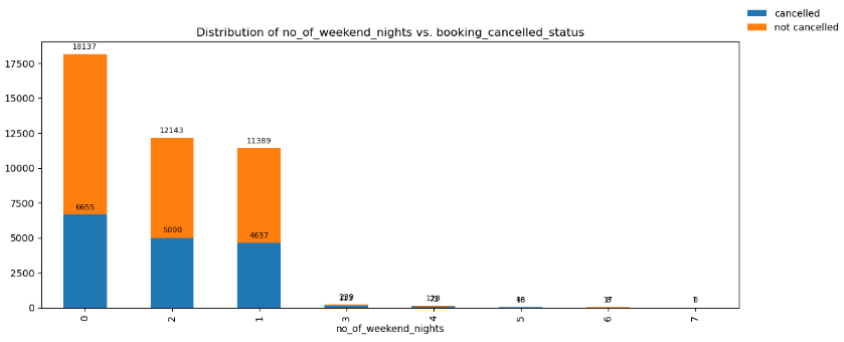
- The majority of bookings were for 0, 1, or 2 weekend nights, with a smaller number of bookings for 3 or more weekend nights.

- There were more bookings that were not cancelled than cancelled for all number of weekend nights categories, and The cancellation rate appears to be higher for bookings with more weekend nights.

- For example, for bookings with 3 weekend nights, there were only 126 not cancelled bookings compared to 113 cancelled bookings.This information could be useful for hotels to understand the cancellation patterns of their guests.

- hotels could consider offering promotions or incentives for guests who book for multiple weekend nights to encourage them to keep their bookings.

booking_cancelled_status	cancelled	not cancelled
no_of_weekend_nights		
0	6655	11482
2	5000	7143
1	4637	6752
3	113	126
4	73	55
5	18	28
6	8	9
7	0	1



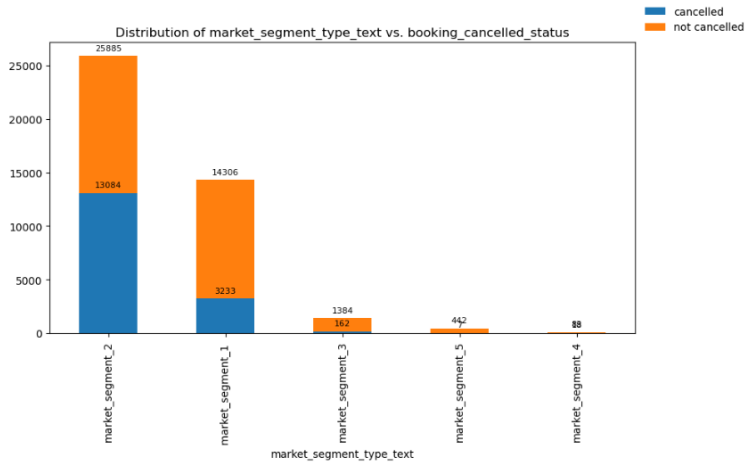
Market segment type Vs booking status:

- Market segment type 2 had the highest number of cancelled bookings, with 13,084 cancelled bookings. However, market segment type 1 had a higher cancellation rate, with 22.6% of bookings cancelled compared to 50.5% for market segment type 2.

- Market segment types 3, 4, and 5 had relatively low cancellation rates, with less than 15% of bookings cancelled.

- In summary, the results show the number of bookings that were cancelled and not cancelled by the market segment type. The highlights that market segment type 2 had the highest number of bookings, but also the highest number of cancellations. Market segment type 1 had a higher cancellation rate, while market segment types 3, 4, and 5 had relatively low cancellation rates. These insights could be useful for hotels to tailor their marketing and pricing strategies to attract and retain customers from different market segments, and to reduce the risk of cancellations.

booking_cancelled_status	cancelled	not cancelled
market_segment_type_text		
market_segment_2	13084	12801
market_segment_1	3233	11073
market_segment_3	162	1222
market_segment_5	7	435
market_segment_4	18	65



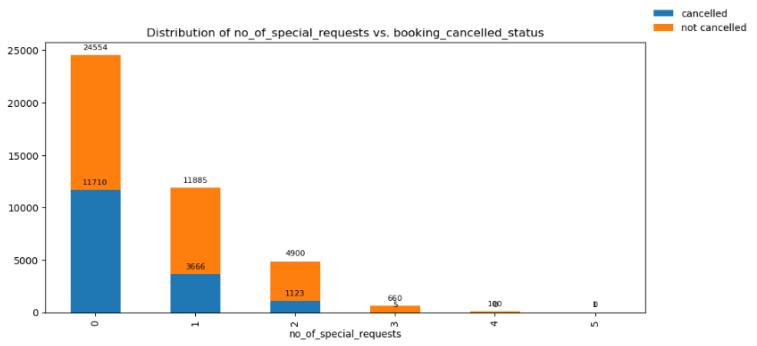
Number of special requests Vs booking status:

- The largest number of canceled reservations in relation to the remaining special requests (11,710) was for reservations without special requests (0). However, the cancellation rate was higher for bookings with three special requests, with about 48% of those bookings canceled and about 52% not cancelled.

- The cancellation rate for bookings with one or two special requests was relatively low, at around 15% and 5% respectively and not cancellation around 33% and 15% respectively .

- In summary, the results show the number of bookings that were cancelled and not cancelled by the number of special requests. The highlights that the majority of bookings did not have any special requests, and that the cancellation rate was highest for bookings with three special requests. These insights could be useful for hotels to tailor their services and amenities to meet the specific needs and preferences of their customers, and to reduce the risk of cancellations.

booking_cancelled_status	cancelled	not cancelled
no_of_special_requests		
0	11710	12844
1	3666	8219
2	1123	3777
3	5	655
4	0	100
5	0	1



Machine Learning models:

For the binary classification of the booking status, I built and evaluated three different machine learning models:

- **Decision Tree:** Decision trees are a simple yet powerful machine learning model. They partition the input data space into rectangular regions and make a prediction for each region. I trained a decision tree model on the training data set.

- **XGBoost:** XGBoost is an optimized distributed gradient boosting library. It builds an ensemble of weak prediction models, typically decision trees. By combining multiple weak learners, XGBoost creates a strong learner. I trained an XGBoost model on the training data set.

- **LGBM:** LightGBM is a fast, distributed, high-performance gradient boosting framework. It is another implementation of gradient boosting decision trees. I trained a LGBM model on the training data set.

To determine the best model, I evaluated all three models on the test data set. Based on the evaluation metrics such as accuracy, F1 score, etc., the LGBM model performed the best. Therefore, I selected the LGBM model for making final predictions on the booking status.

- In summary, I built three machine learning models based on decision trees and ensemble methods, evaluated them on the test data, and chose the LGBM model as the best model for the binary classification task based on its superior performance.

Model Performance Metrics :

To evaluate the performance of each model, we have used the following metrics: precision, recall, f1-score, and accuracy. Here, we present the results for each model:

LGBM Classifier:

LightGBM model Classifier model achieved an accuracy of 82% in predicting the booking status. The model had a precision of 0.83 and recall of 0.88 in predicting not cancelled (status 0) indicating it was good at detecting not cancelled. However, the precision and recall of predicting of cancelled (status 1) were 0.79 and 0.73 respectively showing some difficulty in identifying booking status.

	precision	recall	f1-score	support
0.0	0.83	0.88	0.85	7671
1.0	0.79	0.73	0.76	4959
accuracy			0.82	12630
macro avg	0.81	0.80	0.81	12630
weighted avg	0.82	0.82	0.82	12630

Models Evaluation :

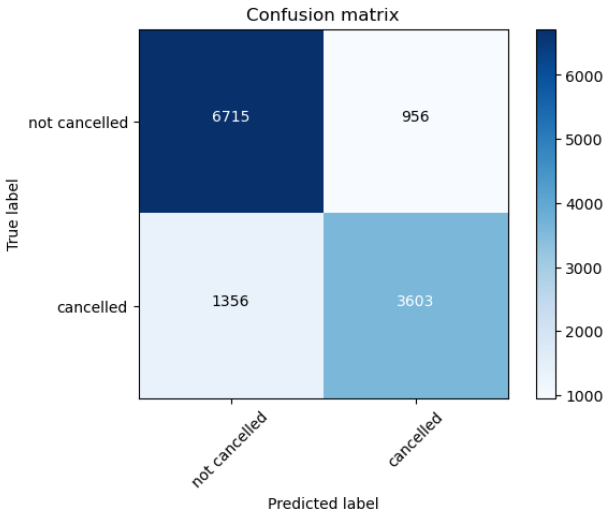
Confusion Matrix Explanation

A confusion matrix is a table that is often used to describe the performance of a classification model. In our case, it shows the true and predicted booking statuses in hotels. The values in the confusion matrix represent the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.

In our context:

- TP: Correctly predicted as cancelled
- TN: Correctly predicted as not cancelled
- FP: Incorrectly predicted as cancelled
- FN: Incorrectly predicted as not cancelled

```
Confusion matrix, without normalization
[[6715  956]
 [1356 3603]]
```



Here are the confusion matrices for LightGBM model:

- The confusion matrix for our booking status model shows that it correctly predicted 6715 bookings as not cancelled and 3603 bookings as cancelled. However, it incorrectly predicted 956 bookings as cancelled when they were actually not cancelled, and 1356 bookings as not cancelled when they were actually cancelled. Despite these errors, the model generally performed well and was able to accurately predict a large number of bookings.

- Our model correctly predicted the booking status (cancelled or not cancelled) for 69.1% of the bookings in the dataset. This means that our model is reasonably effective in predicting whether a booking will be cancelled or not." (Jaccard = 0.691)

- The F1-Score of our booking status model was 0.815, indicating that it has a good balance of precision and recall in predicting both cancelled and not cancelled bookings. This means that the model is able to accurately identify both types of bookings and is not biased towards one class over the other." (F1-Score = 0.815)

- Our model's predicted probabilities for booking status were quite accurate, with a log loss of 0.411. This means that the model's predicted probabilities were generally close to the actual probabilities of the bookings being cancelled or not cancelled. (Log loss = 0.411)

	Algorithm	Eval_Metric	Value
0	LGBMClassifier	Jaccard	0.691
1	XGBClassifier	Jaccard	0.686
2	DecisionTreeClassifier	Jaccard	0.681
3	LGBMClassifier	F1-Score	0.815
4	XGBClassifier	F1-Score	0.812
5	DecisionTreeClassifier	F1-Score	0.808
6	LGBMClassifier	Accuracy	0.817
7	XGBClassifier	Accuracy	0.813
8	DecisionTreeClassifier	Accuracy	0.809
9	LGBMClassifier	LogLoss	0.411
10	XGBClassifier	LogLoss	0.415
11	DecisionTreeClassifier	LogLoss	0.480

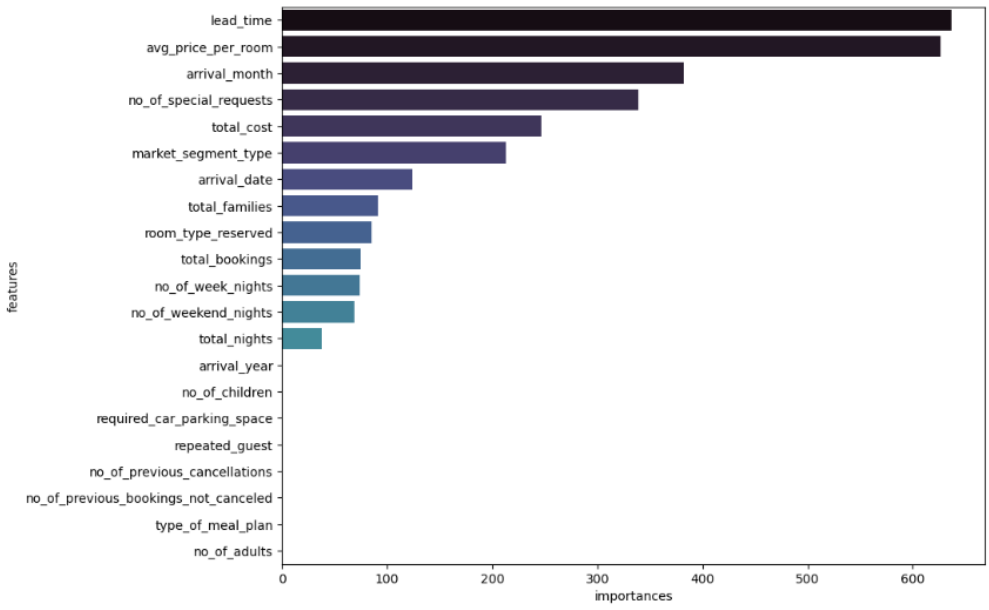
Feature Importance :

- The lead time variable, referring to the number of days between booking date and arrival date, was found to be the most significant predictor of booking status. A longer lead time increases the likelihood of cancellations. To mitigate this, I would recommend implementing policies to restrict bookings beyond a certain threshold prior to the arrival date, such as limiting bookings more than 60-90 days in advance. This can help reduce uncertainty from bookings made too far ahead in time and prevent overbooking scenarios due to cancellations.

- The average price paid per room was also an important feature impacting booking status. The hotel should adopt dynamic pricing strategies that optimize room rates based on several factors, including seasonality, events, demand, and other metrics. Higher rates for peak seasons and major events can capture additional revenue, while lower promotional rates for off-seasons can spur more bookings and foot traffic.

- The arrival month is a meaningful feature for predicting booking status, likely due to seasonal fluctuations in demand. Proactively planning seasonal marketing campaigns around months with higher and lower volumes, as well as laying out different promotional offers for various months may help tap into these patterns and bring more consistency to year-round bookings.

- The number of special requests and total costs perpersonalized experiences for larger bookings can be an effective strategy.



Thank You

Prepared by **Muhannad Mansour Felemban**

code : https://github.com/Muhannad0101/EDA-and-Predicting-of-Booking-Status/blob/main/eda-and-classification-models-for-cancelled-status.ipynb