# Hotel Booking Cancellation Prediction

## Decision Systems

Date: 3/10/24

# Contents / Agenda

- Data Dictionary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Conclusions and Recommendations
- Appendix

# Data Dictionary

The data contains the different attributes of **customers' booking details**. The detailed data dictionary is given below:

- **Booking_ID:** the unique identifier of each booking
- **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 weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
- **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 INN Hotels Group

# Data Dictionary

- **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

# How to use this deck?

- This slide deck serves as a comprehensive template for your project submission
- Within this deck, you will come across various questions that are intended to test your ability to understand data visualizations, discover patterns / insights and postulate hypothesis. Think thoroughly and provide answers to these questions
- You are encouraged to modify this deck as required, by replacing the questions with suitable answers
- Please feel free to incorporate additional points if you deem necessary

**Note:** The data visualizations you see in this deck are obtained from RapidMiner

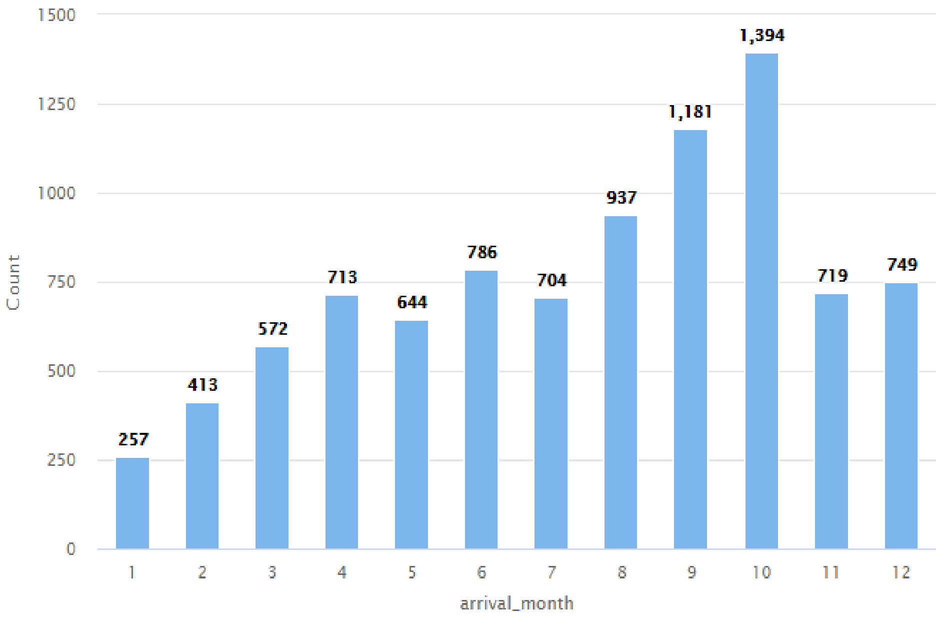
# Business Problem Overview and Solution Approach

- Please define the problem
  - Historically, hotel cancellations have been an area of concern for overall revenue because hotels are unable to determine with great precision, the likelihood that a reservation will be cancelled. If the reservation is cancelled, the hotel will lose money depending on how close it is to the check-in date. To better understand if a cancellation will occur, we will need to better understand the other variables and their correlation to a cancellation.

- Please mention the solution approach / methodology
  - We will determine the likelihood that a customer will cancel their reservation by using Decision Trees and Random Forests to understand the factors that lead to a cancellation.
  - Our Root Node is Lead time, but we will measure other factors such as Arrival Month, Special Requests, Meal Plan, Price, and Number of Adults
  - We will first conduct the Decision Tree process before using the pruned Decision Tree
  - Second, we will conduct the Random Forest process before using the pruned Random Forest

# EDA - Univariate Analysis

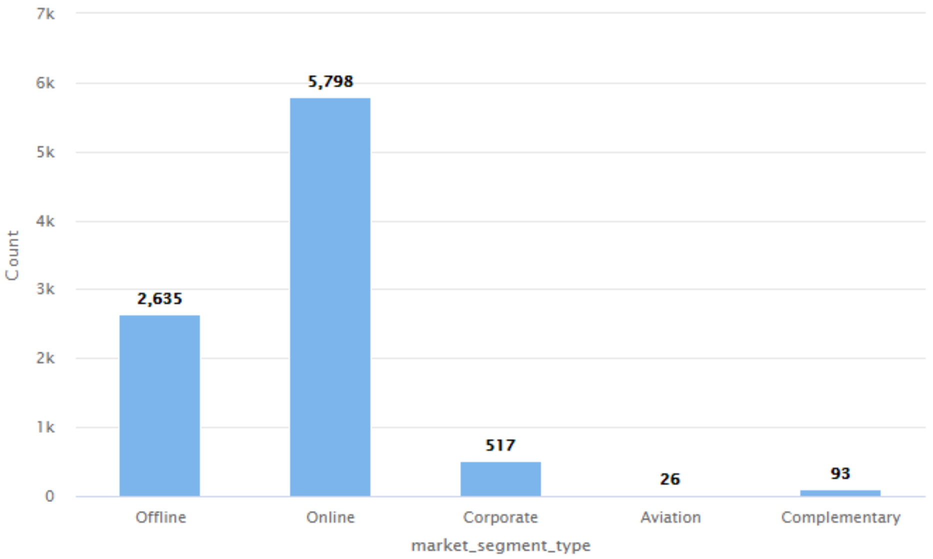
- Explore strategies to effectively manage and accommodate the increased demand
  - The number of bookings is lowest in the Winter months (November to February) and increases to its peak in October throughout the year
  - To manage this demand and maximize our profit, we can increase the lead time for bookings in the autumn since it is the best indicator of whether a Cancellation will occur. We can also provide discounts on other services when users are booking at this time
  - One area that we may want to investigate further is how we can improve the number of bookings in the Winter months.



**X-axis:** Arrival month
**Y-axis:** Number of bookings
**Description:** Number of bookings per month

# EDA - Univariate Analysis

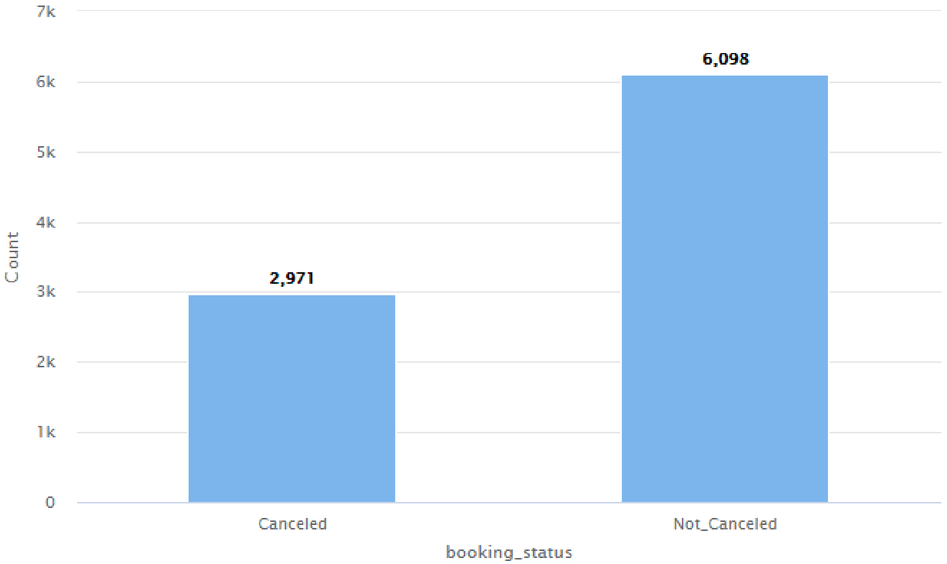
- Identify the percentage of bookings made online versus offline and highlight any trends or insights that can inform business strategies.
  - Based on the chart, there were 9,069 total bookings made across all Market Segments
  - Offline accounted for 29.1% of total bookings and Online accounts for 64% of total bookings
  - This would imply that Online is our key driver for hotel bookings and that is where we should focus our efforts to minimize the amount of Cancelaltions
  - We may want to look into Corporate, Aviation and Complementary segments further to see if we can grow those numbers as well



**X-axis:** Market Segment Type
**Y-axis:** Number of customers
**Description:** Number of Bookings received in each market segment

# EDA - Univariate Analysis

- Analyze the percentage of bookings that are cancelled. Are the cancellation trends a serious issue that needs to be looked at? If so, what potential solutions can be implemented to address this problem?
  - Approximately 33% of our total bookings end in a Cancellation
  - According to [WebRezPro](), the avg number of cancellations globally in 2021 was 25% and fell to 20% in 2023. This indicates we are well above the average
  - Some solutions that can be looked at involve the other variables (e.g. Special Requests, Meal Plan, and Number of Children) and building booking packages that include them



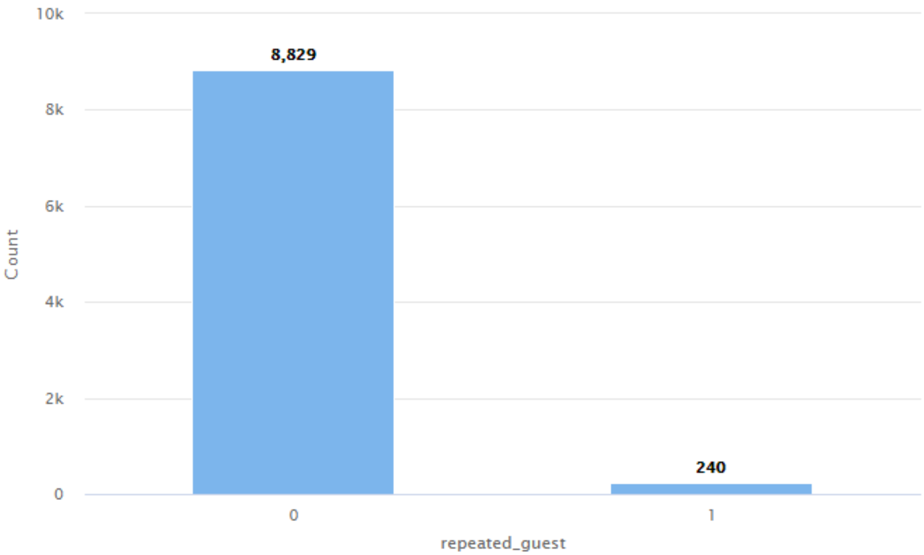
**X-axis:** Booking status

**Y-axis:** Number of bookings

**Description:** Count of bookings that are Cancelled v/s Not Cancelled

# EDA - Univariate Analysis

- To what extent are customers returning to the hotel? Is this indication considered positive or negative, and what factors may be influencing this behavior? Furthermore, what strategies can be implemented to address and potentially improve this situation?
    - Here we can see less than 3% (2.6%) of customers are repeat Customers
    - This is alarming and means that we have very poor retention of customers, and need to undestand why
    - We can address this situation by doing a better job to capture custmer feedback and implement solutions from it



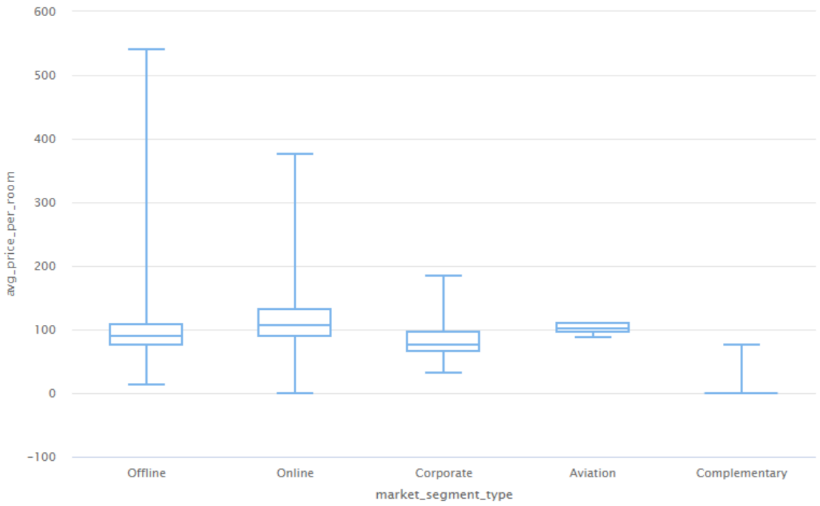
**X-axis:** Repeated guests

**Y-axis:** Number of Guests

**Description:** Customers who have visited the hotel more than once vs customers who have visited the hotel only once

# EDA - Univariate Analysis

- Explore how hotels can optimize their revenue management strategies considering the significant price variations observed across different market segments
  - The chart shows us that we have a wide range of prices for certain segments (Offline and Online) but that it is much smalled for Corporate, Aviation, and Complementary
  - However, regardless of how our prices are marketed, we are still averaging around $100 per night and that range does not vary
  - This means that our ‘surge’ pricing strategies may not be beneficial and are wasting resources (e.g. our marketing budget)



**X-axis:** Market segment type

**Y-axis:** Average price per room

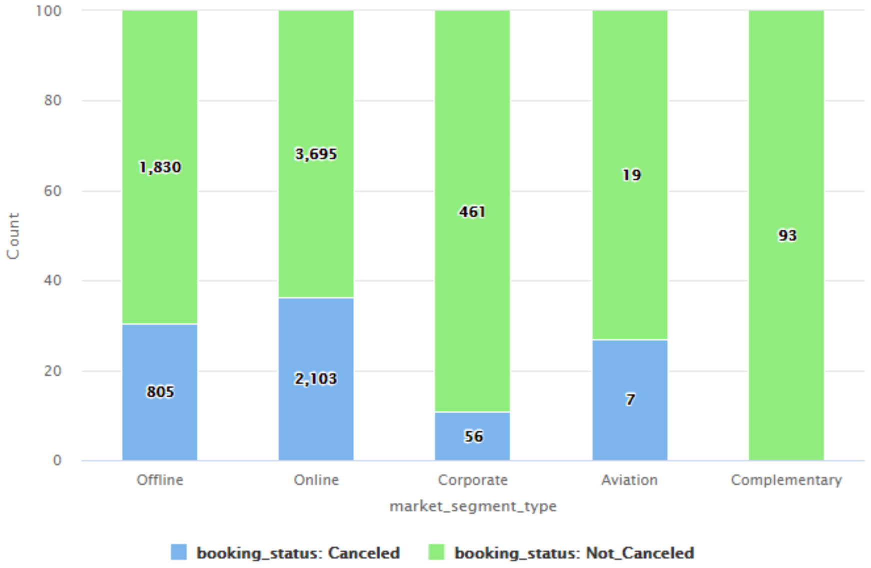
**Description:** Price per room across various  market segments

# Results EDA - Univariate Analysis

- Provide a concise overview of the discoveries derived from the Univariate Analysis.
  - Our bookings follow a cyclical cycle with Winter months having the fewest number of bookings, and it increases throughout the year until its peak in October
  - Online bookings make up nearly 2/3 of our bookings (64%) and Offline bookings accounting for 29% of our total bookings
  - 33% of our bookings end in Cancellation which is much greater than the industry average of 20%
  - More than 97% of our bookings are first time customers indicating that we have poor retention
  - Our price per night has a significant range for our Offline and Online bookings, but the prices still end up around the same average of $100 per night

# EDA - Bivariate Analysis

- Explore strategies to mitigate cancellations and consider how these insights can be leveraged to reduce cancellations across different booking channels
    - The greatest percentage of Cancellations was in the Online segment (36%) with Offline second (31%)
    - This tells us that we should focus on addressing the number of Online cancellations because it has the highest number of Cancellations and has the largest volume
    - We should also focus on increasing our volume in the Corporate and Complementary segments since they have the lowest Cancellations



**X-axis:** Market Segment type

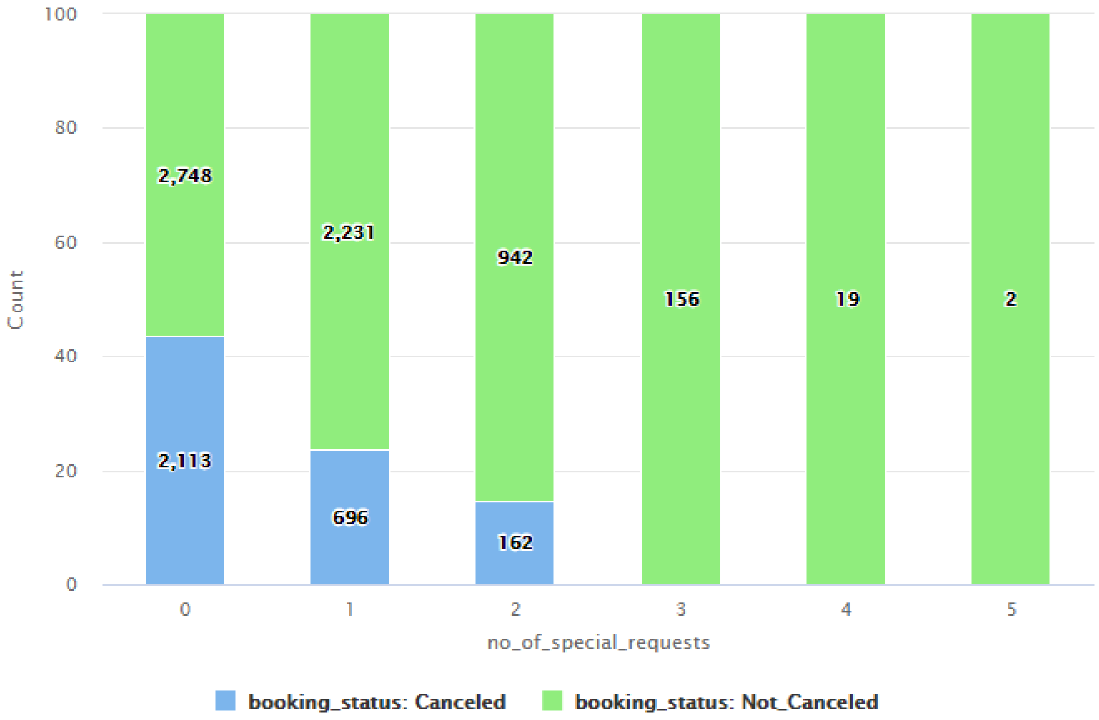
**Y-axis:** Number of customers

**Description:** Number of booking that are cancelled vs not cancelled across various market segments

# EDA - Bivariate Analysis

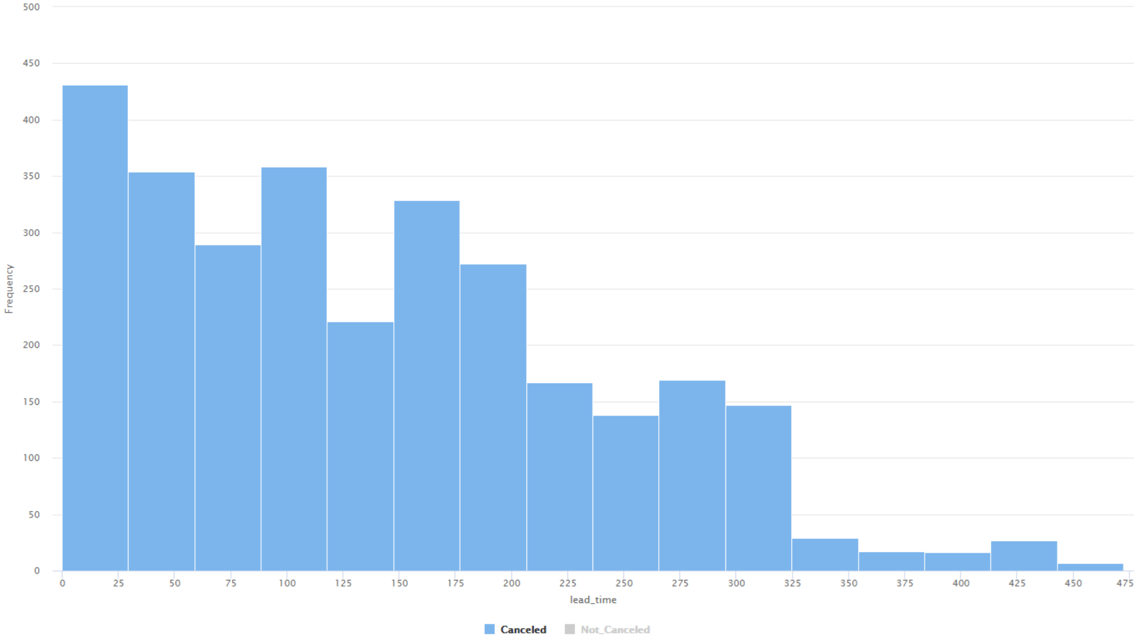
- Examine the influence of special requests on booking cancellations and devise effective strategies to mitigate cancellation rates when customers make such requests.

- We can see clearly that the number of Special Requests greatly decreases the likelihood of a Cancellation
- As a result, we should find ways to encourage the number of Special Requests that a customer makes in the booking



# EDA - Bivariate Analysis

- What insights can be gained from the lead time of canceled bookings, and how can we leverage this information to improve our booking system and enhance customer satisfaction?



This chart is telling us that we should increase the Lead Time, a that >325 days is desirable. Once a booking is made that far in advance, we can say with great certainty that the reservation will not be Cancelled.

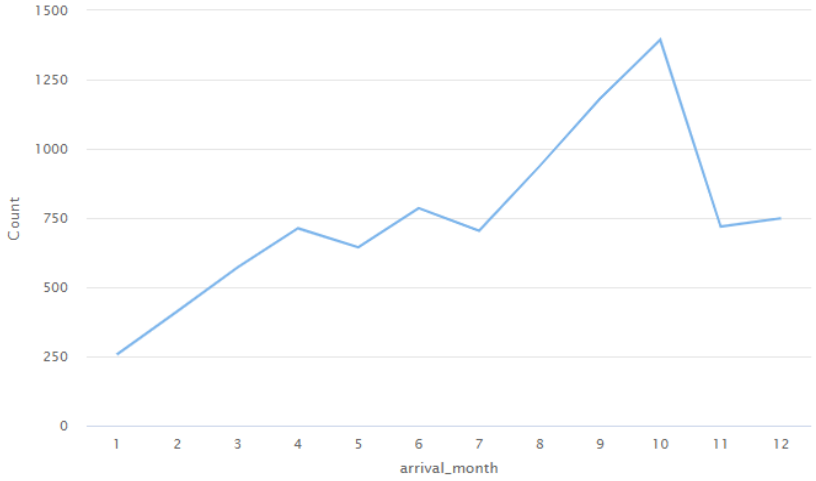
**X-axis:** Lead time

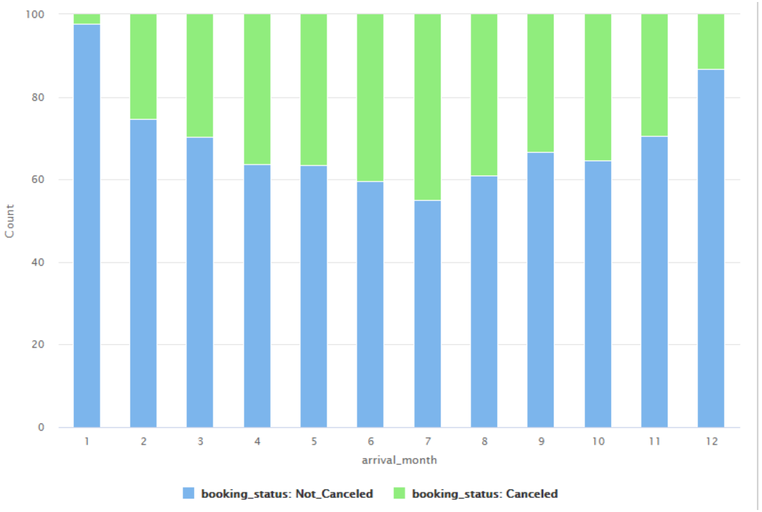
**Y-axis:** Number of bookings

**Description:** Lead time of the bookings that were cancelled

# EDA - Bivariate Analysis

- Analyze the cancellation rate for each month and compare it to other months to determine any patterns or reasons for variation.
    - These charts show us that while our number of Arrivals is greatest in October, that our greatest number of Cancellations is in July
    - The number of Cancellations during the Summer could be the reason why our Summer traffic is light compared to October. This could be because of a change in Summer vacation plans, which is why our Cancellations decrease around the Holidays





# Results EDA - Bivariate Analysis

- Provide a concise overview of the discoveries derived from the Bivariate Analysis.
- We should focus on minimizing the number of Online cancellations while increasing the volume of Corporate and Complementary bookings
- We should find ways to encourage customers to add Special Requests (>2) to accommodate their stay, which can help us significantly decrease our number of Cancellations
- We should increase the our Reservation lead time – the greater the lead time, the greater the likelihood of an Arrival
- We should find ways to encourage summer customers to keep their reservation since that is when we have an average Arrival volume, but experience the greatest number of Cancellations

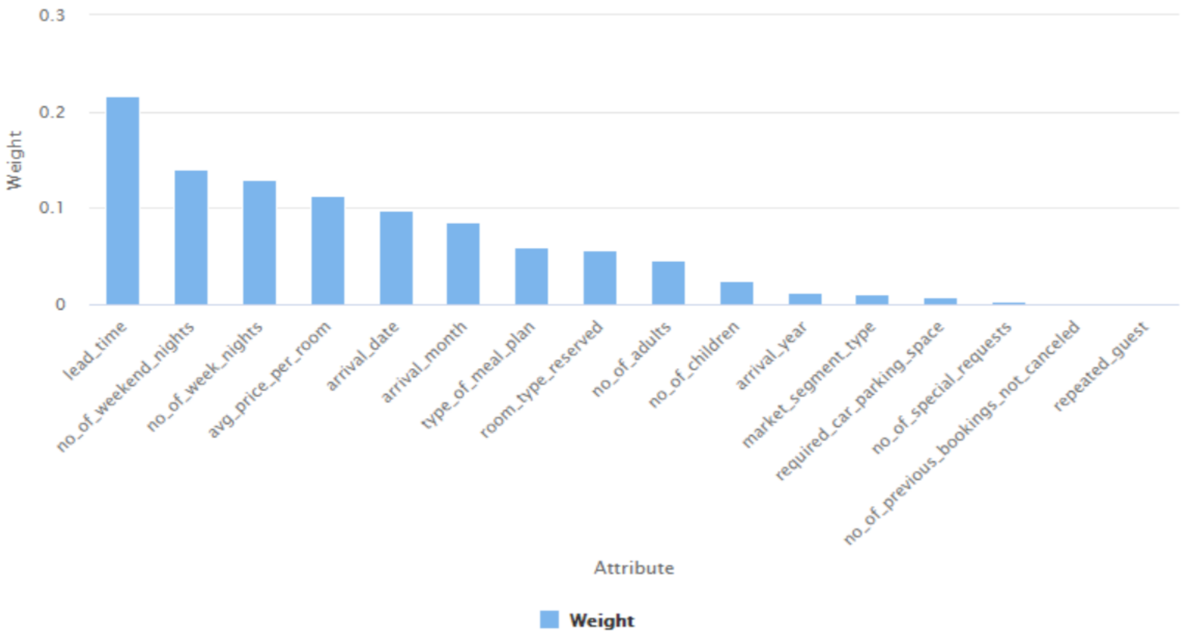
# Model Performance Evaluation Decision Tree

- Please comment on the model performance of the Decision Tree Model
  - The Decision Tree Model did extremely well with the Training data consisting of very high, positive scores (99%+)
  - However, all three categories fell significantly (Accuracy, Precision, and Recall) to below 85%
  - This indicates that the model was heavily overfitted and the model will not do well with Production data
  - As a result, we need to fine tune our model to improve it's performance with the Testing dataset

| Model | Train Accuracy | Test Accuracy | Train Recall | Test Recall | Train Precision | Test Precision |
|---|---|---|---|---|---|---|
| Decision Tree | 99.65 | 83.20 | 99.47 | 76.77 | 99.47 | 73.23 |

# Model Performance Evaluation Decision Tree

- To what extent and in what ways have the attribute weights influenced the overall performance of the model?



- The attributed weights helped us identify what nodes to use as Root Node (e.g. Lead Time) and what our Decision Nodes should be (e.g. Number of Weekend Nights, Number of Weeknights, Price Per Room)
- Assigning different weights to our Attributes may help improve the model

# Model Performance Evaluation Pruned Decision Tree

- Based on the evaluation metrics obtained from a pruned decision tree, do you perceive any notable enhancements in performance? What factors do you believe contributed to this improved performance compared to the previous version?
    - The model worked much better after Pruning, because even though its processing time was greater, the delta between the Accuracy, Recall, and Precision was much better
    - This means that we reduced the amount of Overfitting, even if our Training, Recall, and  Precision values decreased in the Training set.

| Model | Train Accuracy | Test Accuracy | Train Recall | Test Recall | Train Precision | Test Precision |
|---|---|---|---|---|---|---|
| Decision Tree | 99.65 | 83.20 | 99.47 | 76.77 | 99.47 | 73.23 |
| Decision Tree - Pruned | 95.64 | 84.34 | 90.77 | 76.09 | 95.69 | 76.09 |

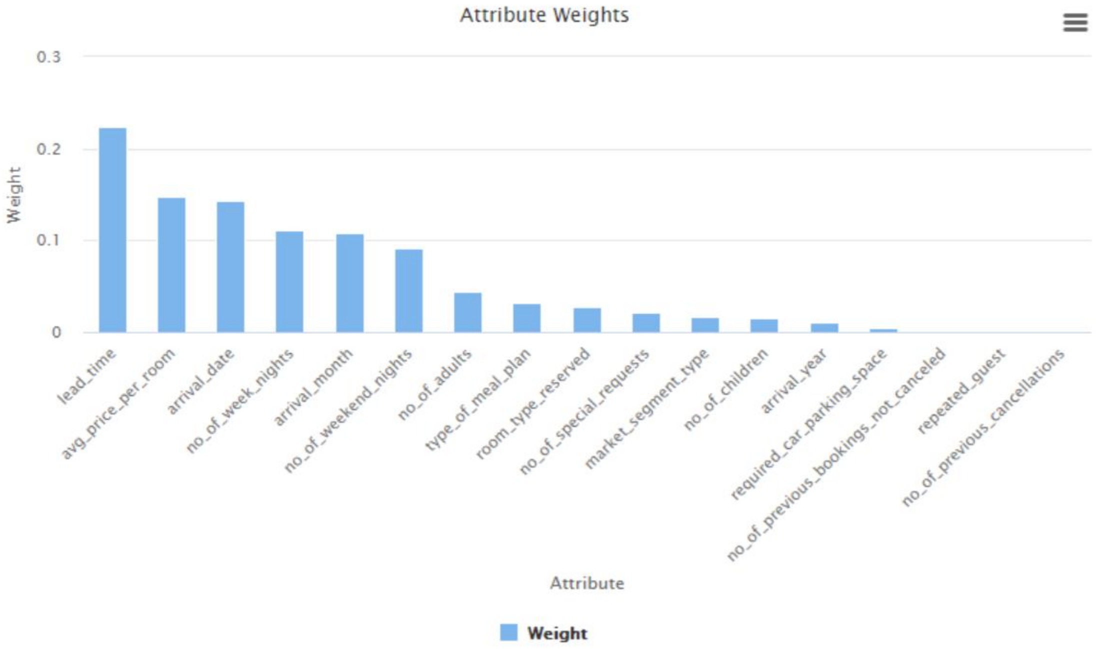
# Model Performance Evaluation Random Forest

- While the Random Forest model demonstrates high accuracy and precision on both the training and test sets, it appears to have a noticeable drop in recall on the test set compared to the training set. What potential factors could contribute to this difference in recall, and how might it impact the model's overall performance and practical applications?
  - The low Recall in our Testing set indicates that it is missing out on identifying the number of Positive cases – in this instance, the number of Cancellations
  - We can fix this issue by providing different weights to the Attributes, including changing our Root Node from Lead Time to other Attributes such as Price Per Room and Arrival Date
  - By changing our measured variables, we can improve the Test Recall

| Model | Train Accuracy | Test Accuracy | Train Recall | Test Recall | Train Precision | Test Precision |
|---|---|---|---|---|---|---|
| Random Forest | 98.47 | 87.94 | 96.44 | 77.10 | 98.87 | 84.71 |

# Model Performance Evaluation Random Forest

- To what extent and in what ways have the attribute weights influenced the overall performance of the model?



- The Weighted Attributes have impacted the Random Forest because we can build Decision Trees with the Root Node containing a different Weighted Attribute
- We can aggregate all that data to see if our model improves in all categories

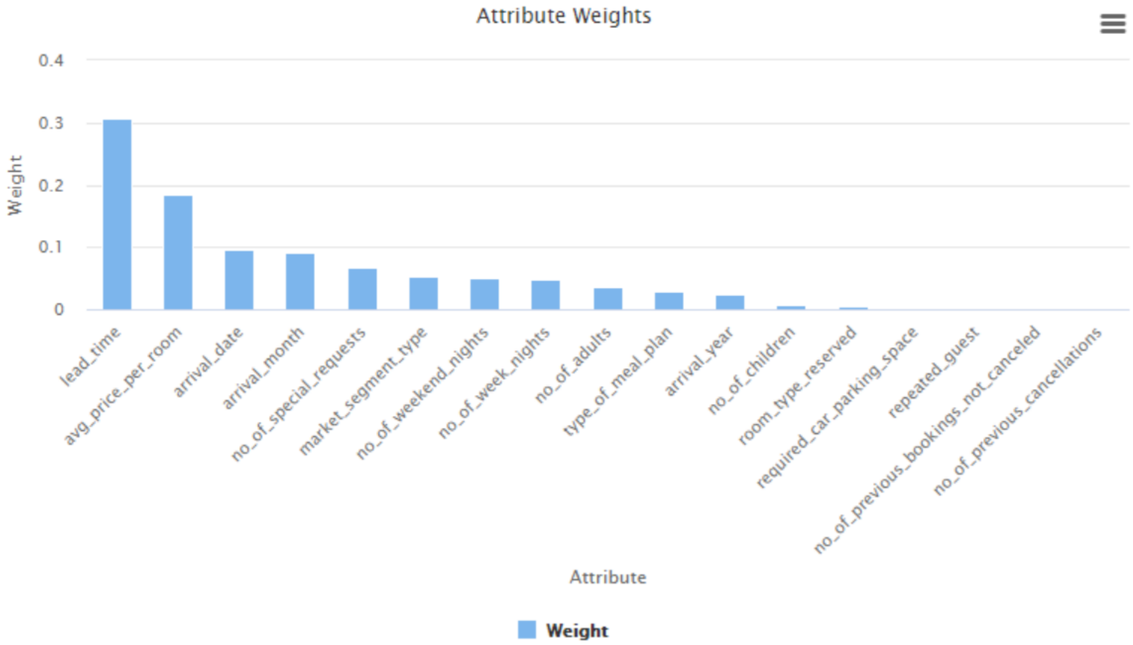
# Model Performance Evaluation Pruned Random Forest

- Based on the evaluation metrics obtained from a pruned Random Forest, do you perceive any notable enhancements in performance? What factors do you believe contributed to this improved performance compared to the previous version?
  - Here we can see that our Δ between Training and Testing data in the Pruned Random Forest is much less compared to the Random Forest model
  - This means that our model is now consistent across both data sets and will be more effective
  - This could be because we changed the weights of each of our Attributes

| Model | Train Accuracy | Test Accuracy | Train Recall | Test Recall | Train Precision | Test Precision |
|---|---|---|---|---|---|---|
| Random Forest | 98.47 | 87.94 | 96.44 | 77.10 | 98.87 | 84.71 |
| Random Forest - Pruned | 87.32 | 85.04 | 69.86 | 66.22 | 89.09 | 84.77 |

# Model Performance Evaluation Pruned Random Forest

- To what extent and in what ways have the attribute weights influenced the overall performance of the model?



- The change in how we weighted our Attributes may have improved the overall model's performance because we can now generate Decision Trees based on other Attributes rather than just Lead Time, and aggregate all the data together

# Model Performance Summary

- The best model for our objective is the Pruned Random Forest because it has the lowest variability in Accuracy, Recall, and Precision amogst the 4 models. Specifically, we want to limit the variance in Recall because we want to correctly yield Cancellation data.
- We were able to come to this conclusion because we can measure and change the weights of our Attributes in the model, and in the case of the Pruned Random Forest, aggregate the data

| Model | Train Accuracy | Test Accuracy | Train Recall | Test Recall | Train Precision | Test Precision |
|---|---|---|---|---|---|---|
| Decision Tree | 99.65 | 83.20 | 99.47 | 76.77 | 99.47 | 73.23 |
| Decision Tree - Pruned | 95.64 | 84.34 | 90.77 | 76.09 | 95.69 | 76.09 |
| Random Forest | 98.47 | 87.94 | 96.44 | 77.10 | 98.87 | 84.71 |
| Random Forest - Pruned | 87.32 | 85.04 | 69.86 | 66.22 | 89.09 | 84.77 |

# Conclusions and Recommendations

- Please mention actionable insights & recommendations
  - From the data, the Client should utilize the Pruned Random Forest to aggregate its data into understanding the factors that lead to a Cancellation
  - The client should understand how other variables such as Special Requests and Meal Plans can improve the number of Arrivals – this would especially be helpful in the Summer when the Cancellations are highest, but the number of Arrivals is around the average

- Below are some useful pointers that can guide you :

  ○ Identify the correlation between the Number of Adults and Avg. Price per Night to see if Revenue could rise in that segment

  ○ Focus on improving the Arrivals from Online bookings

  ○ Customer Retention is critical and a glaring flaw since 98% of our Arrivals are new

  ○ Maximize Lead Time and Special Requests – the longer the Lead Time and more Special Requests, the less likely a Cancellation

# APPENDIX

# Slide Header

- Please add any other pointers (if needed)

# Happy Learning !