UNIVERSITY OF BIRMINGHAM | BIRMINGHAM BUSINESS SCHOOL

Assessment and Feedback: Student Template

Student ID Number(s): 2653531

Module: Marketing Analytics and Behavioral Sciences

Module Leader OR Dissertation/Extended Essay Supervisor: Dr Zizhou Peng

Assignment Title: Technical Report

Date and Time of Submission: 6 May 11:30 AM

Actual Word Count: 3040 approx (excluding ref, tables, citations)

Extension: N * **Extension Due Date:**

I do wish my *anonymised* assignment to be considered for including as an exemplar made available to UoB students. ***** *delete as appropriate*

Please ensure that you complete and attach this template to the front of all work that is submitted.

Declaration

By submitting your work, you are certifying that the submission is the result of your own work and does not contravene the University Code of Practice on Academic Integrity[1,2]. You must ensure that you have referred to valid sources of information to support your work, and that these are properly referenced in the required format (i.e. using Harvard referencing style).

If you have used a proofreader to review all or part of your work, you must declare this here:

☐ I have not used a proofreader

☒ I have used a proofreader. I confirm that the proofreader has not edited the text in an unacceptable manner as specified in Section A.1.6 of the Code of Practice on Academic Integrity[2] and School guidance.

If you have used Generative Artificial Intelligence (GenAI) to support the development of all or part of your work, you must declare this here:

☒ No content generated by GenAI tools has been used in the development of my final submission.

☐ I have used GenAI in the development of my final submission and confirm this has not been included as my own work. I have carefully checked and appropriately used the output according to the University's guidance on using Generative Artificial Intelligence tools ethically for study[3] and I take full responsibility of the entirety of the final submission. *If this option has been selected, please retain your outputs as these could be requested by the module leader grading your work.*

[1] https://intranet.birmingham.ac.uk/student/academic-support/academic-integrity-support-and-advice.aspx
[2] https://intranet.birmingham.ac.uk/as/registry/legislation/codesofpractice/index.aspx
[3] https://intranet.birmingham.ac.uk/as/libraryservices/asc/student-guidance-gai.aspx

CONTINUED BELOW

The purpose of this template is to ensure you make the most effective use of your feedback that will support your learning. It is a requirement to complete both sections, and to include this completed template as the first page of every assignment that is submitted for marking (your School will advise on exceptions).
Section One: Reflecting on the feedback that I have received on previous assessments, the following issues/topics have been identified as areas for improvement: (add 3 bullet points). *NB – for first year students/PGTs in the first term, this refers to assessments in your previous institution* • • •
Section Two: In this assignment, I have attempted to act on previous feedback in the following ways (3 bullet points) • • •

Table of Contents

Executive Summary

With Industry 4.0 technologies, the global tourism and hospitality industry is undergoing a major transformative journey as the existing business model have started to shift with the advent of home sharing models, along with rising costs. Yet, the industry continues to boom with expected revenue generation projected to be US$ 1.1 trillion by 2029. The disruptions alongside the shift towards novel technologies have put marketing analytics in the heart of this transformation.

This paper leverages data analytics tools from a marketing perspective to study hotel booking data collected from two Portuguese hotels, analyzing 36,000 observations. The core objectives of the study are identifying key factors influencing booking lead time and predicting booking cancellations.

Regression analysis reveals that business travelers book earlier compared to leisure travelers, with a weekday booking lead time of 8.98 days. However, leisure booking lead time increases by 15 days when the customers are offered a parking facility. Additionally, higher room prices correlate with last-minute bookings, suggesting luxury customers tend to book later.

Booking cancellation is detrimental to forecasting accuracy as it hinders the calculation of pricing strategy as well as resource wastage. As such, using a decision tree, a classification model has been run which shows a 72.4% sensitivity in predicting cancellations. Key predictors for cancellations are lead time, special request and market segments (booking channels). Online booking cancellation tends to be higher compared to other segments.

Based on the findings, hotels should consider promoting free parking facility for weekend travelers, and dynamic pricing for business travelers through implementing flexible deposit policies to reduce cancellations.

Introduction and Business Context

The Booming Hospitality And Tourism Industry

After a 75% drop between 2020 and 2022 due to the global pandemic, the global tourism and hospitality industry is set to reach a market revenue of US$ 955.94bn by the end of 2025, and with a CAGR of 3.91%, the market value will reach US$ 1.11 trillion by 2029 (Statista, 2024). Furthermore, the spending has seen a similar trend with travelers' outlays in 2024 reaching $8.6 trillion, 9% of the global GDP (Tufft *et al.*, 2024b). After the slowdown, the industry has seen a significant shift due to the introduction of Industry 4.0 technologies (Ben Youssef and Zeqiri, 2022), especially big data analytics, enabling the industry to look into customer behavior from a granular aspect (Jiwnani and Chemmanur, 2024).

Shifting Tide In The Business Model

The rising digital home-sharing business model has taken a sharp cut from the revenue of the existing hotels and resorts, as guests prefer an authentic experience and personalized offers with low costs(Um, Lee and Koo, 2025). This segment has grown from 10% to 14% between 2017 and 2023, with an estimated growth to US$ 232 billion by 2027 (Tufft *et al.*, 2024a). Besides, the big hotel brands prefer more franchise-based ownership rather than fully controlled and owned.

Role Of Marketing Analytics In Tourism And Hospitality

Big data has enabled hotels and businesses to harness the rich data and draw conclusions regarding customer behavior for enhanced services (Stylos et al., 2021, Melián-Alzola et al., 2020). On the other hand, marketing agility, the ability to adapt and quickly respond in marketing, has become mandatory in the tourism industry to adapt and recover from changing consumer behavior and environmental change (Melián-Alzola, Fernández-Monroy and Hidalgo-Peñate, 2020). Combining marketing agility and data analytics, marketing analytics derives marketing insights using descriptive, diagnostic, predictive, and prescriptive tools (Akter *et al.*, 2022) within the tourism industry, for drawing better insights into the transformative industry. Scholars have often described marketing analytics as the heart of this transformation (Chen and Wang, 2019).

Dataset Selection and Analysis

To gain closer insights into the hotel and tourism industry using marketing analytics tools, datasets were searched on multiple platforms, such as Kaggle and the UCI ML Repository. The secondary dataset was collected from Kaggle, which is an abridged version of two datasets collected by Antonio, de Almeida and Nunes (2019). The dataset contains 36,286 observations from two hotels in Portugal with important insights like lead time of booking, weekday, weekend stay, average price, meal plans, and finally whether the booking was accepted or cancelled. The data was collected between 2015 and 2017. Using the data, this paper utilizes data analytics approaches from a marketing perspective to draw insights into two important objectives:

1. To identify key factors influencing booking lead time.
2. To examine key predictors of booking cancellation.

Data Dictionary

The dataset consists of 16 factors and one unique customer identifier. It also consists of 36,286 observations with no missing or duplicate values. With a usability rating of 10/10 in Kaggle, it is believable and accurate, as this data has been used in multiple research studies for predicting booking cancellations using ML techniques and algorithms (Satu, Ahammed and Abedin, 2020; Castro-Martín, Rueda and Ferri-García, 2022; Chen *et al.*, 2023). The following table consists of a description of the dataset's variables.

Table 1 Data Dictionary of the dataset

SL	Feature	Type	Level	Description
1	No of adults	Integer	Quantitative	Total number of adults in each visit
2	No of children	Integer	Quantitative	Total number of children in each visit
3	No of weekend nights	Integer	Quantitative	Weekend nights (Saturday/Sunday), the guests stayed or booked to stay
4	No of week nights	Integer	Quantitative	Weekend nights (Monday to Friday), the guests stayed or booked to stay
5	Type of meal	Categorical	Qualitative	Meal Plan 1 = BB (Bed and Breakfast) Meal Plan 2 = HB Half board (Breakfast and one other meal) Mal Plan 3 = FB Full Board (3 meals)
6	Car parking space	Categorical	Qualitative	Value 1 if the customer requires a parking space, otherwise 0
7	Room type	Categorical	Qualitative	7 different room types offered by the hotels
8	Lead time	Integer	Quantitative	Number of days between the booking date and arrival

SL	Feature	Type	Level	Description
9	Market segment type	Categorical	Qualitative	Market segment designation (channel of booking/ arrival) Types: Aviation, Complementary, Corporate, Offline, Online
10	Repeated	Categorical	Qualitative	Value 1 if the guest came before, otherwise 0
11	P-C	Integer	Quantitative	Number of booking cancellations made by the guest before the current booking
12	P-not-C	Integer	Quantitative	Number of bookings not cancelled by the guest before the current booking
13	Average price	Integer	Quantitative	Average daily price/night
14	Special request	Categorical	Qualitative	Value 1 if the guests made any special request, otherwise 0
15	Date of reservation	Date	Quantitative	Date of reservation
16	Booking status	Categorical	Qualitative	If the customer Cancelled/ Not cancelled

The dataset is a cross-sectional, structured dataset collected in a single instance. For analysis, categorical variables such as type of meal, room type, market segment type, and booking status had been changed to factors using R.

Descriptive Statistics

Guest Statistics

The total adult guests in the two-year period were 66,940 in total visits of 36,285. On average, 1.84 guests visited the hotels with median and mode of 2, meaning double by the adults is the most common. Furthermore, most of the guests visit without any children, with children average only 0.11 with median and mode 0.

Table 2 Descriptive Statistics of Guests

Description	Adults	Children
Mean	1.84	0.11
Standard Error	0.003	0.002
Median	2.000	0.000
Mode	2.000	0.000
Standard Deviation	0.519	0.403
Sample Variance	0.269	0.162
Kurtosis	0.813	36.947
Skewness	-0.333	4.708
Range	4.000	10.000
Minimum	0.000	0.000
Maximum	4.000	10.000
Sum	66940	3823
Count	36285	36285

The histograms further explain the occupancy behavior of the guests, which confirms dual occupancy as the highest and right skewness of the histogram of children, interpreting 0 to 2 children as maximum in each visit. The maximum number of adult guests is four, with a range of four. This echoes the minor standard deviation of the adult guests.

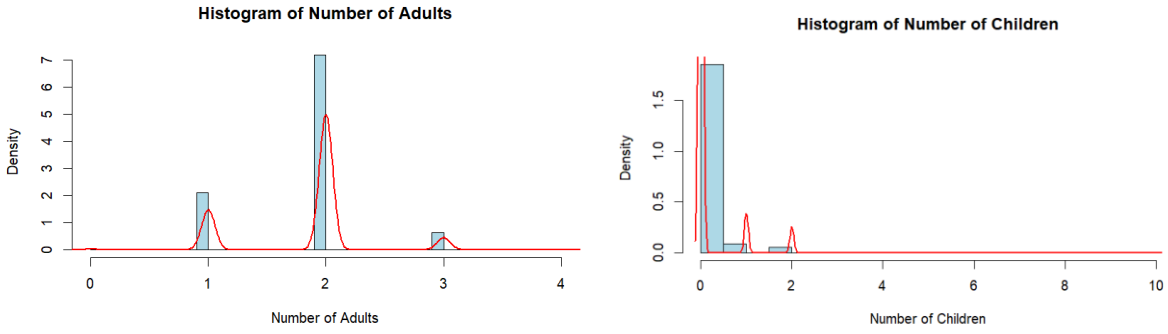

Figure 1 Histogram of Adult and Children guests

Occupancy Statistics

The hotel has two types of occupancy, weekends and weekdays. On average, guests stay 0.81 and 2.20 on weekdays, with median values of 1 and 2, respectively. This suggests that the hotels receive guests more on weekdays compared to weekends, indicating more business travelers compared to leisure travelers.

Besides, the standard deviation for weekday nights is 1.41 compared to weekends (0.87), illustrates more variability on weekdays. Although weekdays and weekends both have right-skewed distributions, weekdays have higher skewness (1.59) and kurtosis (7.794) compared to weekends with skewness (0.738) and kurtosis (0.299), suggesting the presence of longer tails and more outliers, illustrating guests staying longer on weekdays (max 17). On the contrary, weekends have more normally distributed graph with lower skewness and almost normal kurtosis (0.29)

Table 3 Descriptive Statistics of Occupancy (Weekends vs Weekdays)

Description	Weekend	Weekdays
Mean	0.81	2.20
Standard Error	0.005	0.007
Median	1.000	2.000
Mode	0.000	2.000
Standard Deviation	0.871	1.411
Sample Variance	0.758	1.991
Kurtosis	0.299	7.794
Skewness	0.738	1.599
Range	7.000	17.000
Minimum	0.000	0.000
Maximum	7.000	17.000
Sum	29416	79994
Count	36285	36285

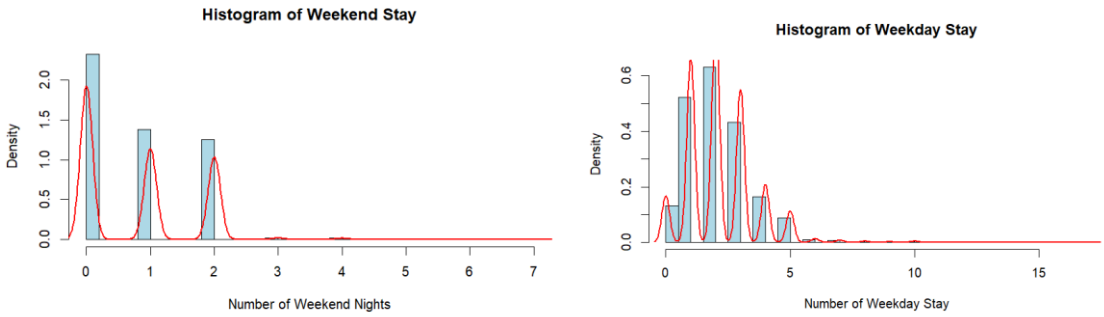

Figure 2 Histogram of Weekend and Weekday stay

The boxplots further reiterate the presence of more outliers in the weekday stays compared to the weekend stays, suggesting customers often stay more than average mostly due to business travel on the hotels.

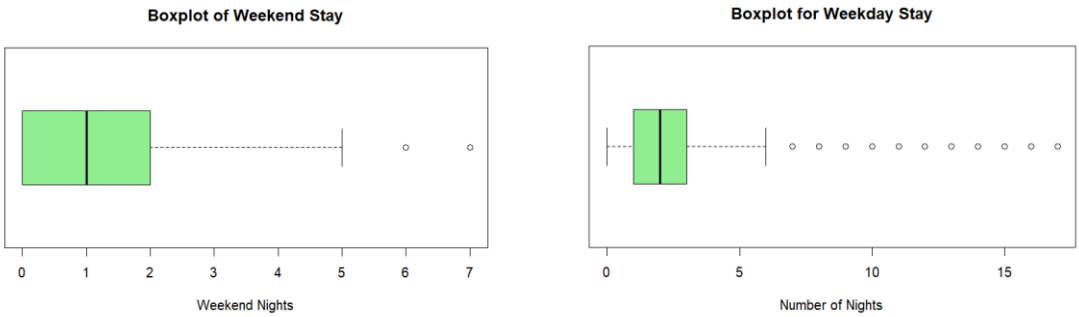

Figure 3 Boxplots of Weekend vs Weekdays

Spending by customers

The mean average spending by customers is £103.42, slightly more than the median £99.45, suggesting guests spend more than the other. This is highlighted with slightly right skewness (0.667). The price ranges are £540 and standard deviation of 35 indicates a spread of spending by the customers.

Table 4 Descriptive Statistics of Customer Spending/Night (Avg price)

Description	Customer Spending
Mean	103.42
Standard Error	0.184
Median	99.450
Mode	65.000
Standard Deviation	35.086
Sample Variance	1231.060
Kurtosis	3.155
Skewness	0.667
Range	540.000
Minimum	0.000
Maximum	540.000
Sum	3752654
Count	36285

The boxplot illustrates the presence of outliers on the right side, indicating guests staying more or spending more on the luxury suites of the hotels.

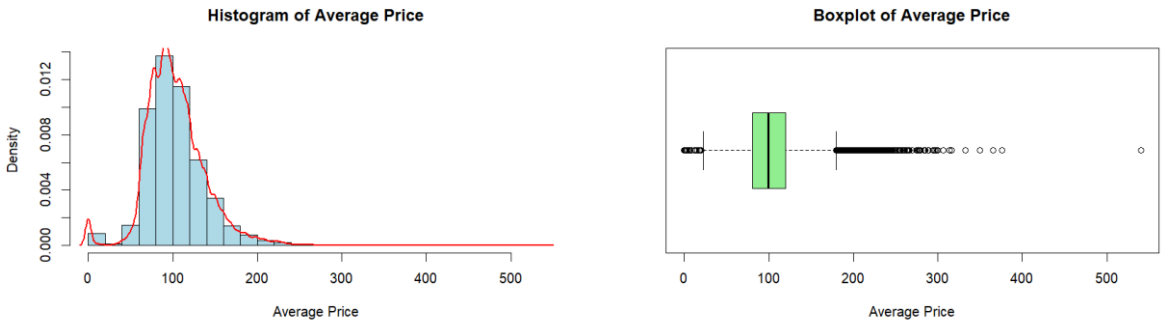

Figure 4 Histogram and Boxplot of Customer Spending (Average Price)

Booking Statistics:

Although the average lead time is 85 days, the median lead time is 57 days indicating many customers book quite earlier than they travel, illustrated by right skewness (1.292). However, mode 0 indicates many customers arrive without any booking. The standard deviation of 85 days and range of 443 days indicate the high variability of the booking along with lots of outliers.

Table 5 Descriptive Statistics of Lead Time

Description	Lead Time
Mean	85.24
Standard Error	0.451
Median	57.000
Mode	0.000
Standard Deviation	85.939
Sample Variance	7385.477
Kurtosis	1.179
Skewness	1.292
Range	443.000
Minimum	0.000
Maximum	443.000
Sum	3092928
Count	36285

This can be interpreted that many customers book their holidays prior, to avoid last minute miss as well as price relief.

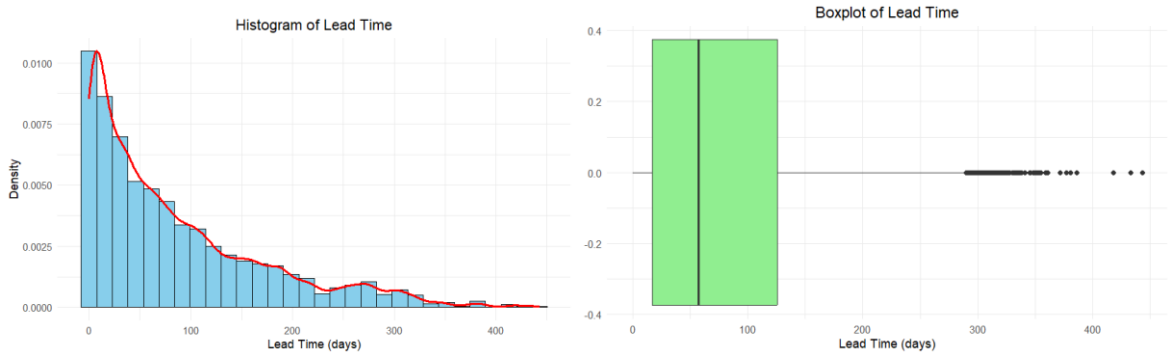

Figure 5 Histogram and Box Plot of Lead Time

Most of the customers booked their trips online, while a second majority booked offline or appeared without any booking.

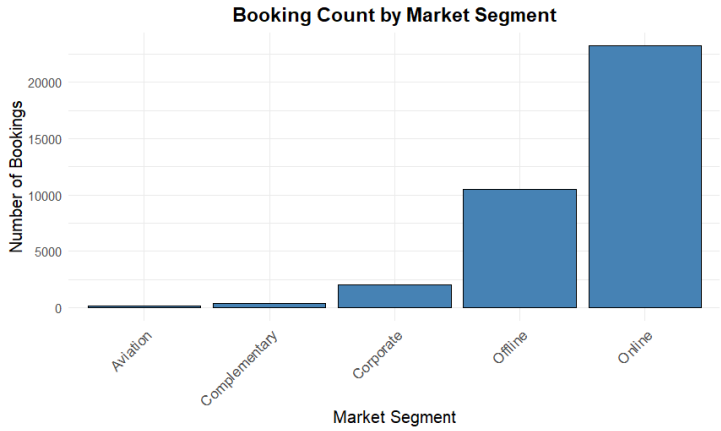

Figure 6 Booking Count by Market Segment

Here, only 3% of the customers asked for car parking space in the hotels, demonstrating most customers don't bring cars with them. Furthermore, a majority of the customers made special requests like twin bed or high floor.

Figure 7 Car Parking Space and Special Requests by Customers

Finally, around 33% of the customers cancelled their bookings which represents almost one-third of total bookings made. However, there are 986 repeat customers and most of them didn't cancel previous bookings indicating their loyalty to the hotels.

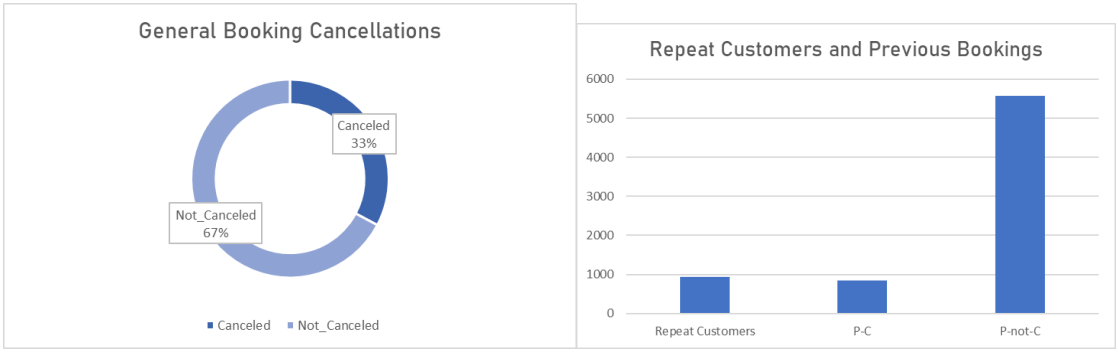

Figure 8 General Booking Cancellation and Cancellation by Repeat Customers

Hypothesis Formulation

Booking Lead Time in the Hotel and Tourism Industry:

Revenue management (RM) is a necessary instrument utilized using the information systems to match the supply and demand of the customers through allocating the right resources to the right customers to maximize the profitability of the organization (Ivanov and Zhechev, 2011). Within the hotel and tourism industry, RM is considered an important KPI for a firm's profitability(Guillet, 2020). Booking lead time is an important metric for calculating the revenue, as it enables forecasting pricing, inventory, and incoming demands(Webb *et al.*, 2021).

In context of marketing, booking lead time acts as an important instrument as it enables the existing hotel industries and Online Travel Agencies (OTA) to set price and promotions at different booking time (Guizzardi *et al.*, 2021). As a result, variables leading to influence lead time are worthy of analysis using the analytical tools for informed decision making.

Length of Stay in Midweeks and Weekends:

Customers staying in the hotel in midweek are mostly business travelers, whereas weekend customers stay mostly for leisure. Although these two major types of customers draw revenue for the hotels, it is often assumed that weekdays are busier than weekends (Lee *et al.*, 2011). As a result, these two distinct customer segments may have differentiating customer dynamics (Figure 2 & Figure 2Figure 3), and the relation with timeframe of booking should be analyzed.

Due to internal factors like short-term inventory, high fixed costs, and constrained capacity, hotels tend to sell out rooms by a target day, making length of stay and lead time important determinants for pricing(Guizzardi, Pons and Ranieri, 2017). Thus, it necessitates to understand the relation between length of stays in weekdays or weekend nights separately on the booking lead time. Based on the assumptions the following hypotheses have been formulated.

Hypothesis 1:

Null Hypothesis (H_0): There is no statistically significant relationship between the number of weekday nights and the lead time of a booking.

Alternative Hypothesis (H_1): There is a statistically significant relationship between the number of weekday nights and the lead time of a booking.

Hypothesis 2:

Null Hypothesis (H_0): There is no statistically significant relationship between the number of weekend nights and the lead time of a booking.

Alternative Hypothesis (H_2): There is a statistically significant relationship between the number of weekend nights and the lead time of a booking.

Average Price Per Night:

As the arrival date draws near, the price quoted for booking reduces for non-busy days; however, it increases exponentially during the final days due to the chances of selling out (Cho *et al.*, 2018). Price per night is attributed as the number one reason of booking a hotel(Jang, Chen and Miao, 2019). While people book earlier due to their desired attributes of the hotel, there has been a shift in the current

behavior. Due to the arrival of the OTA and relevant technologies, the last-minute bookings have soared in the past few years, further facilitated by the last-minute deals of the OTAs (Chen and Schwartz, 2013). As such, it is important to examine the relation between average price per night and lead time, leading to the third hypothesis.

Hypothesis 3:

Null Hypothesis (H_0): There is no statistically significant relationship between average price and the lead time of a booking.

Alternative Hypothesis (H_3): There is a statistically significant relationship between average price and the lead time of a booking.

Multivariate Regression Model

Prior to running the hypotheses for formulating a multivariate regression model, the existing relationship between the variables are required to be tested. The selected independent variables from the hypotheses are: number of weekdays stay, number of weekends stay and average price and the dependent variable is the booking lead time.

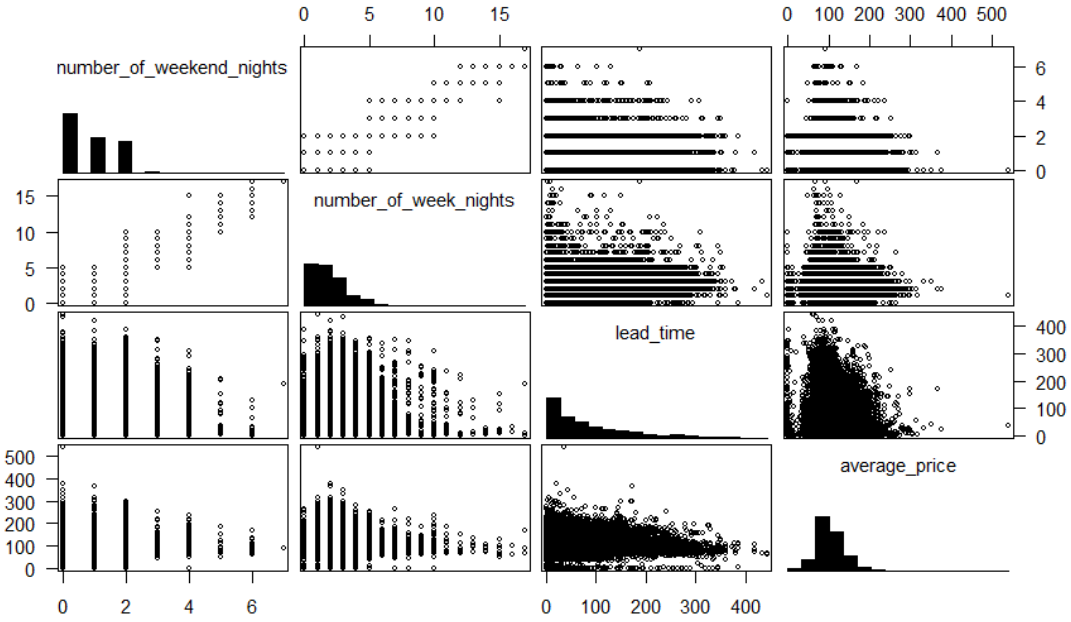

Figure 9 Pair Plots of variables

Although the pair plot doesn't reveal any specific linear relationship among the variables, there are chances of possible nuance relationship between the variables. For instance, we can see pattern between average price and lead time; number of week nights and average price. To verify that there is no multicollinearity among the independent variable, correlation is to be conducted among the independent variables.

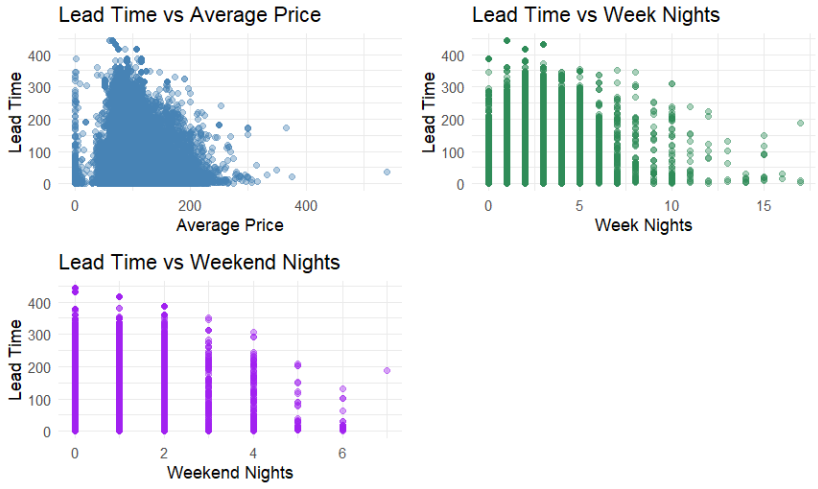

Figure 10 Pair plot of Lead Time with other Independent Variables

There is a very little to no correlation among the independent variables as highlighted from the correlation heatmap. However, the independent variables are slightly correlated with the dependent variable, i.e. lead time.

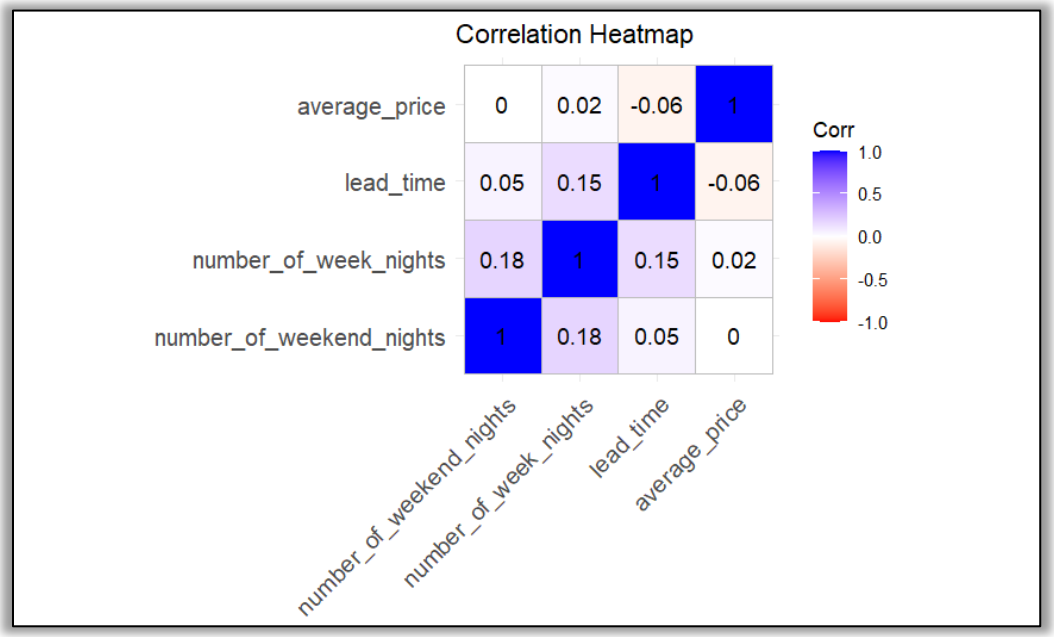

Figure 11 Correlation Heatmap among Variables

Regression Analysis:

With no multicollinearity, hypothesis 1, 2 and 3 from the previous section were tested to see any statistical significance between number of week nights, number of weekend nights, and average price with lead time separately.

Regression Summary:

Hypothesis 1: (Week Nights & Lead Time)

The p-value is smaller than 0.05 indicating a statistically significant relationship exists. So we reject the null value. Holding other variables constant, one day increase in weekday night stay is associated with approx. 9 days increase in lead time. This suggests that customers book earlier when planning for long stays in weekdays, mostly due to work related time constraints.

Hypothesis 2: (Weekend Nights & Lead Time)

Here, p-value<0.01, indicating the hypothesis is statistically significant. As such we reject the null value. The coefficient 1.94 indicates that with all other values constant, one extra weekend night stay increases the lead time by 1.94 days.

```
Call:
lm(formula = lead_time ~ number_of_week_nights + number_of_weekend_nights +
    average_price, data = booking)

Residuals:
    Min      1Q  Median      3Q     Max
-224.99  -63.14  -27.44   38.38  363.97

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              80.52761    1.56695  51.391  < 2e-16 ***
number_of_week_nights     8.98756    0.32071  28.024  < 2e-16 ***
number_of_weekend_nights  1.94622    0.51964   3.745  0.00018 ***
average_price            -0.16128    0.01269 -12.712  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84.77 on 36281 degrees of freedom
Multiple R-squared:  0.02711,   Adjusted R-squared:  0.02703
F-statistic:   337 on 3 and 36281 DF,  p-value: < 2.2e-16
```

Figure 12 Regression Summary

Hypothesis 3: (Average Price & Lead Time)

The coefficient for average price is – 0.161 with a p-value < 0.01 illustrates a statistically negative relationship between average price per night and lead time of booking. So we reject the null hypothesis. For each unit increase in the average price, the lead time decreases by 0.16, indicating that high-spending customers prefer delayed booking.

Without any of the variables, the booking lead time will be the same as the intercept, i.e., 80.52

Model Fit:

Although the regression is statistically significant (p-value<0.05), the multiple R^2 value is 0.0271 and the adjusted R^2 value is 0.027, indicating only 2.7% variability is explained by the three variables combined. This underscores the lower predictive power of the model and requires more variables. A step-wise regression (forward selection or backward elimination) can be utilized.

Multivariate Regression Equation:

Booking Lead Time = βo+ β1*number of weekday night stay+ β2*number of weekend night stay+ β3*average price + ε

Booking Lead Time = 80.52+ 8.98*number of weekday night stay+ 1.94*number of weekend night stay-0.16 average price + ε

Introducing Interaction

Facility for car parking plays a pivotal role in hotel management, and as a result, customers who are required to park their own vehicles can have distinctive booking behavior. To see the statistical significance of car parking on the overall multivariate regression model, the following hypothesis has been formulated.

Null Hypothesis (H_0): There is no statistically significant relationship between requesting a car parking space and the lead time of a booking.

Alternative Hypothesis (H_4): There is a statistically significant relationship between requesting a car parking space and the lead time of a booking.

Taking the car parking space as an independent variable, it was tested with the previously formulated regression model to evaluate its association with booking lead time.

```
Call:
lm(formula = lead_time ~ number_of_week_nights + number_of_weekend_nights +
    average_price + has_car_parking, data = booking)

Residuals:
    Min      1Q  Median      3Q     Max
-222.62  -62.91  -27.50   38.05  363.16

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               80.94525    1.56505  51.720  < 2e-16 ***
number_of_week_nights      8.83294    0.32055  27.556  < 2e-16 ***
number_of_weekend_nights   1.82401    0.51897   3.515 0.000441 ***
average_price             -0.15290    0.01269 -12.046  < 2e-16 ***
has_car_parkingTRUE      -27.26270    2.57335 -10.594  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84.64 on 36280 degrees of freedom
Multiple R-squared:  0.03011,   Adjusted R-squared:  0.03001
F-statistic: 281.6 on 4 and 36280 DF,  p-value: < 2.2e-16
```

Figure 13 Regression Summary with Car Parking Space as an Independent Variable

With p-value < 0.05, the model is statistically significant, with R^2 and adjusted R^2 increased from 2% to 3%, explaining 3% variability of booking lead time. Furthermore, car parking is negatively significant, with all variables constant, if a customer requires car parking, it reduces lead time by 27 days, meaning customers book last if they are travelling with their own car. As such the new equation is:

Booking Lead Time = 80.94+ 8.83*number of weekday night stay+ 1.82*number of weekend night stay-0.16 average price – 27.26 * has car parking + ε

To analyze the effect of car parking on the independent variables, it was tested using regression with each of the other independent variables to find an association with lead time.

Interaction	Coefficient	p-value	R-squared value
average price* car parking	0.38	<0.05	0.031
week nights * car parking	3.57	0.079	0.032
weekend nights * car parking	15.4	<0.05	0.03

Figure 14 Interaction with Independent Variables

The p-value of week nights and car parking is more than 0.05 which means it's statistically not significant. On the other hand car parking show statistically significant interaction with both average price and weekend nights. However, for weekend nights it increases the coefficient to 15.4, meaning customer requiring car parking space are likely to book 15 days prior for one extra day stay during the weekends.

```
Call:
lm(formula = lead_time ~ number_of_week_nights + number_of_weekend_nights +
    average_price + has_car_parking + (number_of_weekend_nights *
    has_car_parking), data = booking)

Residuals:
    Min      1Q  Median      3Q     Max
-220.98  -62.99  -27.30   38.39  362.81

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               81.34743    1.56684  51.918  < 2e-16 ***
number_of_week_nights                      8.86163    0.32051  27.649  < 2e-16 ***
number_of_weekend_nights                   1.41784    0.52571   2.697    0.007 **
average_price                             -0.15420    0.01269 -12.150  < 2e-16 ***
has_car_parkingTRUE                      -37.46684    3.34092 -11.215  < 2e-16 ***
number_of_weekend_nights:has_car_parkingTRUE  15.42394    3.22196   4.787  1.7e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84.61 on 36279 degrees of freedom
Multiple R-squared:  0.03072,   Adjusted R-squared:  0.03059
F-statistic:   230 on 5 and 36279 DF,  p-value: < 2.2e-16
```

Figure 15 Summary of Regression with car parking as an Interaction

Thus, the new multivariate regression equation, along with interaction, is

Booking Lead Time = 81.34+ 8.86*number of weekday night stay+ 1.41*number of weekend night stay-0.15* average price – 37.26 * has car parking +15.42(weekend night stay * has car parking) +ε

Marketing Implications from Regression Models:

The objective of the analysis is to analyze the ways to increase the booking lead time for optimized pricing and resource management. From the initial regression, we can assume that the business travelers book their journey quite early than leisure travelers, indicating their constrained and limited schedule. On the contrary, the weekend travelers are mostly last-minute booking guests. Special promotions, discounts, and marketing campaigns should be run for weekend travelers so that they are persuaded to book early. From the later regression, we have seen that the overall lead time decreases with the introduction of the parking facility. This highlights more local residents or distant travelers who may stay for the night while traveling. As a result, reserving some parking spaces can increase guest acquisition significantly. Finally, the interaction reveals that weekend travelers book

early when they have access to parking. Thus, guests who visit the hotels for leisure can be offered free parking if prior booking is made.

Classification analysis

Predicting booking cancellations

Booking cancellations can cause a serious impact on the revenue management of the hotel and tourism industry, as it disrupts the forecasting, leading to over-inventory, overstaffing, and price mismanagement(Antonio, Almeida and Nunes, 2017). As a result, predicting cancellations can be the optimal solution for avoiding last-minute resource wastage.

Here, a decision tree model has been run to predict whether a customer will cancel the booking or not. All variables except "Booking ID" and "Date of Reservation" were removed before the modelling. In the data, 11889 (33%) of the bookings were cancelled, the remaining 24396(66%) were not cancelled. Initially, the C5.0 algorithm was run with 15 variables. It created a decision tree with a tree size, i.e., number of leaves, 350.

```
Call:
C5.0.formula(formula = booking_status ~ ., data = booking.df)

Classification Tree
Number of samples: 36285
Number of predictors: 15

Tree size: 350
```

Figure 16 Initial Decision Tree

With an error rate of 10.9%, the decision tree produced a complex decision structure with 339 nodes.

```
Evaluation on training data (36285 cases):

          Decision Tree
        ----------------
        Size      Errors

         339 3944(10.9%)   <<


          (a)   (b)    <-classified as
         ----  ----
         9471  2418    (a): class Canceled
         1526 22870    (b): class Not_Canceled


       Attribute usage:

       100.00% lead_time
        95.20% special_requests
        86.17% market_segment_type
        82.28% average_price
        80.29% repeated
        77.77% car_parking_space
        57.13% total_nights
        27.82% room_type
        19.15% type_of_meal
        18.86% number_of_weekend_nights
        17.88% number_of_adults
        15.29% number_of_week_nights
         1.34% number_of_children
```

Figure 17 Initial Decision Tree Result Summary

As a result, further pruning was done setting mincases to 2000, i.e. at least 2000 cases are required before splitting into a new node.

```
Evaluation on training data (36285 cases):

           Decision Tree
         ----------------
         Size      Errors

            5 7272(20.0%)   <<


          (a)   (b)    <-classified as
         ----  ----
         8606  3283    (a): class Canceled
         3989 20407    (b): class Not_Canceled


        Attribute usage:

        100.00% lead_time
         80.29% special_requests
         41.91% market_segment_type
```

Figure 18 Summary Result of Controlled Decision Tree

Although the error increased to 20%, the model is simpler with only five nodes. The three most important factors to decide whether a booking will be cancelled or not are lead time, special requests and market segment type.

Table 6 Confusion Matrix

Actual/ Predicted	**Cancelled**	**Not Cancelled**
Cancelled	8606 (True Positive)	3283 (False Negative)
Not Cancelled	3989 (False Positive)	20407 (Ture Negative)

Sensitivity: TP/(TP+FN)

The sensitivity of the model is 72.4%, indicating it can identify 72.4% of actual cancellation

Specificity: TN/(TN+FP)

The specificity is 83.7%, which identifies 83.7% of actually not cancelled bookings.

Thus, the model identifies noncancellation better than cancellation.

Marketing Implications:

Customers with lead times of more than 151 days are prone to cancel their bookings. In this case, extra discounts can be provided to those who book earlier than 151 days with an increased booking fee. If the special requests of customers who book less than 151 days in advance are not fulfilled, they are prone to cancel the order. In this scenario, the hotels can offer customizable options with pay-per-use or usage-based pricing models, allowing customers to choose what they need.

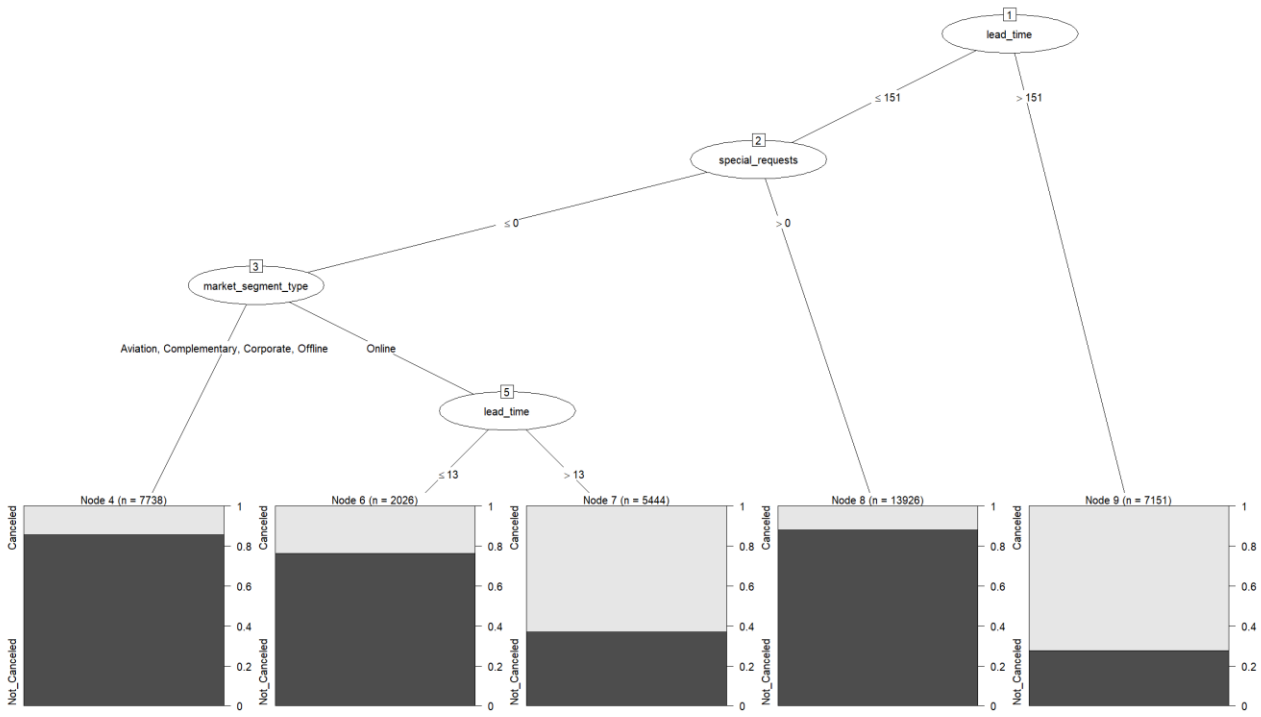

Figure 19 Decision Tree for Booking Cancellation Prediction

Next, reliable market segments such as aviation, complementary, corporate, and offline should also be prioritized. Currently, the number of customers in these segments is quite low compared to online (Figure 6). Special partnerships, promotions, and MOUs can be signed to increase the customers from those segments.

Conclusion:

From the analysis, we find that for explaining the variability of lead time, more variables should be considered which can be added through stepwise regression. The discussed independent variables, though statistically significant, perform poorly to describe the variability. On the other hand, lead time, special request and market segments turns out to be the most important for predicting booking cancellation.

References

Akter, S. *et al.* (2022) 'The future of marketing analytics in the sharing economy', *Industrial Marketing Management*, 104, pp. 85–100. Available at: https://doi.org/10.1016/j.indmarman.2022.04.008.

Antonio, N., Almeida, A. de and Nunes, L. (2017) 'Predicting hotel booking cancellations to decrease uncertainty and increase revenue', *Tourism & Management Studies*, 13(2), pp. 25–39. Available at: https://doi.org/10.18089/tms.2017.13203.

Antonio, N., de Almeida, A. and Nunes, L. (2019) 'Hotel booking demand datasets', *Data in Brief*, 22, pp. 41–49. Available at: https://doi.org/10.1016/j.dib.2018.11.126.

Ben Youssef, A. and Zeqiri, A. (2022) 'Hospitality Industry 4.0 and Climate Change', *Circular Economy and Sustainability*, 2(3), pp. 1043–1063. Available at: https://doi.org/10.1007/s43615-021-00141-x.

Chen, C. and Schwartz, Z. (2013) 'On revenue management and last minute booking dynamics', *International Journal of Contemporary Hospitality Management*, 25(1), pp. 7–22. Available at: https://doi.org/10.1108/09596111311290192.

Chen, Y. and Wang, L. (2019) 'Commentary: Marketing and the Sharing Economy: Digital Economy and Emerging Market Challenges', *Journal of Marketing*, 83(5), pp. 28–31. Available at: https://doi.org/10.1177/0022242919868470.

Cho, S. *et al.* (2018) 'Optimal dynamic hotel pricing', in *2018 Meeting Papers*.

Guillet, B.D. (2020) 'An evolutionary analysis of revenue management research in hospitality and tourism: Is there a paradigm shift?', *International Journal of Contemporary Hospitality Management*, 32(2), pp. 560–587. Available at: https://doi.org/10.1108/IJCHM-06-2019-0515.

Guizzardi, A. *et al.* (2021) 'Big data from dynamic pricing: A smart approach to tourism demand forecasting', *International Journal of Forecasting*, 37(3), pp. 1049–1060. Available at: https://doi.org/10.1016/j.ijforecast.2020.11.006.

Guizzardi, A., Pons, F.M.E. and Ranieri, E. (2017) 'Advance booking and hotel price variability online: Any opportunity for business customers?', *International Journal of Hospitality Management*, 64, pp. 85–93. Available at: https://doi.org/10.1016/j.ijhm.2017.05.002.

Ivanov, S.H. and Zhechev, V.S. (2011) 'Hotel Revenue Management – A Critical Literature Review'. Rochester, NY: Social Science Research Network. Available at: https://doi.org/10.2139/ssrn.1977467.

Jang, Y., Chen, C.-C. and Miao, L. (2019) 'Last-minute hotel-booking behavior: The impact of time on decision-making', *Journal of Hospitality and Tourism Management*, 38, pp. 49–57. Available at: https://doi.org/10.1016/j.jhtm.2018.11.006.

Jiwnani, L. and Chemmanur, A. (2024) *The hospitality industry guest of the future | Deloitte UK*. Available at: https://www.deloitte.com/uk/en/Industries/consumer/blogs/the-hospitality-guest-of-the-future.html (Accessed: 5 May 2025).

Lee, S. *et al.* (2011) 'Do you really know who your customers are?: A study of US retail hotel demand', *Journal of Revenue and Pricing Management*, 10(1), pp. 73–86. Available at: https://doi.org/10.1057/rpm.2009.8.

Melián-Alzola, L., Fernández-Monroy, M. and Hidalgo-Peñate, M. (2020) 'Information technology capability and organisational agility: A study in the Canary Islands hotel industry', *Tourism Management Perspectives*, 33, p. 100606. Available at: https://doi.org/10.1016/j.tmp.2019.100606.

Statista (2024) *Travel & Tourism - Worldwide | Statista Market Forecast*, *Statista*. Available at: http://frontend.xmo.prod.aws.statista.com/outlook/mmo/travel-tourism/worldwide (Accessed: 5 May 2025).

Tufft, C. *et al.* (2024a) *Six trends shaping new business models in tourism and hospitality*. Available at: https://www.mckinsey.com/industries/travel/our-insights/six-trends-shaping-new-business-models-in-tourism-and-hospitality (Accessed: 5 May 2025).

Tufft, C. *et al.* (2024b) *The trends shaping tourism in 2024 | McKinsey*. Available at: https://www.mckinsey.com/industries/travel/our-insights/now-boarding-faces-places-and-trends-shaping-tourism-in-2024 (Accessed: 5 May 2025).

Um, T., Lee, Y. and Koo, J. (2025) 'Economic impacts of digital home-sharing platform: Creative destruction in the hospitality industry', *Tourism Economics*, 31(2), pp. 201–220. Available at: https://doi.org/10.1177/13548166241253888.

Webb, T. *et al.* (2021) 'Hotel revenue management forecasting accuracy: the hidden impact of booking windows', *Journal of Hospitality and Tourism Insights*, 5(5), pp. 950–965. Available at: https://doi.org/10.1108/JHTI-05-2021-0124.

Appendix:

Appendix

Descriptive Statistics

```
booking.df <-read.csv('E:\\UoB\\Semester 2\\MABS\\Assessments\\Assignment 2\\booking_regression_q2.csv')


summary(booking.df)


# Factorizing the categorical variables
booking.df$type_of_meal <- factor(booking.df$type_of_meal)
booking.df$room_type <- factor(booking.df$room_type)
booking.df$market_segment_type <- factor(booking.df$market_segment_type)
booking.df$booking_status <- factor(booking.df$booking_status)


summary(booking.df)


# Histogram + Density for number_of_adults
hist(booking.df$number_of_adults,
    breaks = 30,
    prob = TRUE,
    main = "Histogram of Number of Adults",
    xlab = "Number of Adults",
    col = "lightblue")
lines(density(booking.df$number_of_adults, na.rm = TRUE), col = "red", lwd = 2)


# Boxplot for number_of_adults
boxplot(booking.df$number_of_adults,
      main = "Boxplot for Number of Adults",
      xlab = "Number of Adults",
      col = "lightgreen",
```

```
      horizontal = TRUE)

# Histogram + Density for number_of_children
hist(booking.df$number_of_children,
    breaks = 30,
    prob = TRUE,
    main = "Histogram of Number of Children",
    xlab = "Number of Children",
    col = "lightblue")
lines(density(booking.df$number_of_children, na.rm = TRUE), col = "red", lwd = 2)

# Boxplot for number_of_children
boxplot(booking.df$number_of_children,
      main = "Boxplot for Number of Children",
      xlab = "Number of Children",
      col = "lightgreen",
      horizontal = TRUE)

# Histogram for no_of_week_nights
hist(booking.df$number_of_week_nights,
    breaks = 30,
    prob = TRUE,
    main = "Histogram of Weekday Stay",
    xlab = "Number of Weekday Stay",
    col = "lightblue")
lines(density(booking.df$number_of_week_nights, na.rm = TRUE), col = "red", lwd = 2)

# Boxplot
boxplot(booking.df$number_of_week_nights,
```

```
      main = "Boxplot for Weekday Stay",
      xlab = "Number of Nights",
      col = "lightgreen",
      horizontal = TRUE)
# Histogram for weekend nights
hist(booking.df$number_of_weekend_nights,
    breaks = 30,
    prob = TRUE,
    main = "Histogram of Weekend Stay",
    xlab = "Number of Weekend Nights",
    col = "lightblue")
lines(density(booking.df$number_of_weekend_nights, na.rm = TRUE), col = "red", lwd = 2)

# Boxplot
boxplot(booking.df$number_of_weekend_nights,
      main = "Boxplot of Weekend Stay",
      xlab = "Weekend Nights",
      col = "lightgreen",
      horizontal = TRUE)
# Histogram of average price
hist(booking.df$average_price,
    breaks = 30,
    prob = TRUE,
    main = "Histogram of Average Price",
    xlab = "Average Price",
    col = "lightblue")
lines(density(booking.df$average_price, na.rm = TRUE), col = "red", lwd = 2)

# Boxplot
```

```
boxplot(booking.df$average_price,
        main = "Boxplot of Average Price",
        xlab = "Average Price",
        col = "lightgreen",
        horizontal = TRUE)

library(ggplot2)

ggplot(booking.df, aes(x = market_segment_type)) +
  geom_bar(fill = "steelblue", color = "black") +  # Black border for bars
  labs(title = "Booking Count by Market Segment",
       x = "Market Segment",
       y = "Number of Bookings") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),  # Centered, bold title
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),      # Slanted readable x labels
    axis.title = element_text(size = 12)
  )

#Histogram of lead time

library(ggplot2)



ggplot(booking.df, aes(x = lead_time)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "skyblue", color = "black") +
  geom_density(color = "red", size = 1) +
  labs(title = "Histogram of Lead Time", x = "Lead Time (days)", y = "Density") +
```

```
 theme_minimal() +
 theme(plot.title = element_text(hjust = 0.5))

#Boxplot

ggplot(booking.df, aes(x = lead_time)) +
 geom_boxplot(fill = "lightgreen") +
 labs(title = "Boxplot of Lead Time", x = "Lead Time (days)") +
 theme_minimal() +
 theme(plot.title = element_text(hjust = 0.5))
```

Multivariate Regression Model

```
install.packages("tidyverse")
install.packages("gridExtra")
install.packages("emmeans")
install.packages("reshape2")
install.packages("ggcorrplot")

library(tidyverse)
library(gridExtra)
library(emmeans)

#Loading dataset
booking <- as.data.frame(read_csv("E:/UoB/Semester 2/MABS/Assessments/Assignment 2/booking_regression_q2.csv"))

#summarizing dataset
head(booking)
summary(booking)
```

```
str(booking)

#calculating correlation
round(cor(booking[, c("number_of_weekend_nights", "number_of_week_nights", "lead_time", "average_price")]), digits = 2)

library(gridExtra)

p1 <- ggplot(booking, aes(x = average_price, y = lead_time)) +
 geom_point(alpha = 0.4, color = "steelblue") +
 geom_smooth(method = "loess", se = TRUE, color = "darkred") +
 labs(title = "Lead Time vs Average Price", x = "Average Price", y = "Lead Time") +
 theme_minimal()

p2 <- ggplot(booking, aes(x = number_of_week_nights, y = lead_time)) +
 geom_point(alpha = 0.4, color = "seagreen") +
 geom_smooth(method = "loess", se = TRUE, color = "darkred") +
 labs(title = "Lead Time vs Week Nights", x = "Week Nights", y = "Lead Time") +
 theme_minimal()

p3 <- ggplot(booking, aes(x = number_of_weekend_nights, y = lead_time)) +
 geom_point(alpha = 0.4, color = "purple") +
 geom_smooth(method = "loess", se = TRUE, color = "darkred") +
 labs(title = "Lead Time vs Weekend Nights", x = "Weekend Nights", y = "Lead Time") +
 theme_minimal()

grid.arrange(p1, p2, p3, ncol = 2)

#visualizing correlation with correlation matrix
library(ggplot2)
```

```
library(reshape2)
library(ggcorrplot)

# Select relevant numeric columns
selected_data <- booking[, c("number_of_weekend_nights", "number_of_week_nights", "lead_time", "average_price")]

# Compute correlation matrix
cor_matrix <- round(cor(selected_data, use = "complete.obs"), 2)

# Plot using ggcorrplot
ggcorrplot(cor_matrix,
        method = "square",
        lab = TRUE,
        lab_size = 4,
        colors = c("red", "white", "blue"),
        title = "Correlation Heatmap",
        ggtheme = theme_minimal())
#visualizing pairplot

install.packages("gpairs")
library(gpairs)

gpairs(selected_data)



#calculating regression (lead time by week.weekend.avg_price)

m1 <- lm(lead_time ~ number_of_week_nights + number_of_weekend_nights + average_price, data= booking)
```

```
summary(m1)
```

Regression with Interaction

```
#calculating correlation with car parking space

round(cor(booking[, c("number_of_weekend_nights", "number_of_week_nights", "lead_time", "average_price", "car_parking_space")]), digits = 2)


#visualizing correlation with correlation matrix

library(ggplot2)

library(reshape2)

library(ggcorrplot)


# Select relevant numeric columns

selected_data_2 <- booking[, c("number_of_weekend_nights", "number_of_week_nights", "lead_time", "average_price","car_parking_space")]


# Compute correlation matrix

cor_matrix_2 <- round(cor(selected_data_2, use = "complete.obs"), 2)


# Plot using ggcorrplot

ggcorrplot(cor_matrix_2,
        method = "square",
        lab = TRUE,
        lab_size = 4,
        colors = c("red", "white", "blue"),
        title = "Correlation Heatmap",
        ggtheme = theme_minimal())
#calculating regression (lead time by week.weekend.avg_price.parking)
```

```
m2 <- lm(lead_time ~ number_of_week_nights + number_of_weekend_nights + average_price+ car_parking_space, data= booking)

summary(m2)


#Introducing new variable : hotels with car parking


booking$has_car_parking <- factor(booking$car_parking_space > 0)


#calculating regression (lead time by week.weekend.avg_price.has_car_parking)


m3 <- lm(lead_time ~ number_of_week_nights + number_of_weekend_nights + average_price+ has_car_parking, data= booking)

summary(m3)


#adding interaction (car parking and average price)

m4 <- lm(lead_time ~ number_of_week_nights + number_of_weekend_nights + average_price+ has_car_parking+ average_price*has_car_parking , data= booking)

summary(m4)


#adding interaction (car parking and week nights)

m5 <- lm(lead_time ~ number_of_week_nights + number_of_weekend_nights + average_price+ has_car_parking+ (number_of_week_nights *has_car_parking) , data= booking)

summary(m5)


#adding interaction (car parking and weekend nights)

m6 <- lm(lead_time ~ number_of_week_nights + number_of_weekend_nights + average_price+ has_car_parking+ (number_of_weekend_nights *has_car_parking) , data= booking)

summary(m6)



summary(m1)$r.squared
```

```
summary(m2)$r.squared

summary(m3)$r.squared

summary(m4)$r.squared

summary(m5)$r.squared

summary(m6)$r.squared
```

Decision Tree

```
booking.df <-read.csv('E:\\UoB\\Semester 2\\MABS\\Assessments\\Assignment 2\\booking_regression_q2.csv')

summary(booking.df)

# Factorizing the categorical variables

booking.df$type_of_meal <- factor(booking.df$type_of_meal)

booking.df$room_type <- factor(booking.df$room_type)

booking.df$market_segment_type <- factor(booking.df$market_segment_type)

booking.df$booking_status <- factor(booking.df$booking_status)

summary(booking.df)

#removing irrelevant columns from decision tree

booking.df <- booking.df[, !(names(booking.df) %in% c("Booking_ID", "date_of_reservation"))]

str(booking.df)

table(booking.df$booking_status)

#running the C5.0 model

library(C50)
```

```
# Build the C5.0 decision tree model
booking_model <- C5.0(booking_status ~ ., data = booking.df)

booking_model

# View a summary of the model
summary(booking_model)

#plot the model
plot(booking_model)

# limiting the number of cases per leaf node
booking_model2 <- C5.0(
  booking.df[, -which(names(booking.df) == "booking_status")],
  booking.df$booking_status,
  control = C5.0Control(minCases = 2000)
)

# Display basic info about the tree
booking_model2

# Display detailed summary of the tree
summary(booking_model2)

# Plot the simplified tree
plot(booking_model2)
```