HOTEL BOOKINGS

TABLE OF CONTENTS

BACKGROUND

Understanding business is crucial for anyone seeking to navigate the complex world of commerce. It involves comprehending the fundamental principles and practices that drive economic activity. By gaining a solid grasp of business concepts, individuals can make informed decisions, identify opportunities, and effectively contribute to the growth and success of organizations.

Business understanding encompasses several aspects, including the study of market dynamics, financial management, and strategic planning. Market dynamics include analyzing consumer behavior, market trends, and competition to identify potential target markets and develop effective marketing strategies. Financial management focuses on understanding financial statements, budgeting, and investment decisions. Strategic planning involves setting long-term goals.

INTRODUCTION & METHODOLOGY

The primary objective of this study project is to conduct a thorough examination and understanding of the intricate dynamics included within hotel reservation data spanning from 2015 to 2017. This study aims to discover the intricate factors that influence hotel bookings and their subsequent impacts on important aspects of the hospitality industry via a comprehensive investigation and application of several statistical models. The objective of this research is to discern fundamental patterns, trends, and interconnections in the data by scrutinizing the extensive dataset, which encompasses a diverse range of attributes such as visitor preferences and booking criteria.

PROBLEM STATEMENT

In this project, we delve into a comprehensive analysis of hotel bookings data spanning from July 2015 to August 2017. The primary objectives of our analysis revolve around employing statistical and machine learning models to uncover insights into the factors influencing the Average Daily Rate (ADR) and booking dynamics. This project aims to provide actionable insights for stakeholders in the hospitality industry, assisting them in making informed decisions based on a comprehensive understanding of the factors that influence hotel bookings and Average Daily Rate.

RESEARCH QUESTIONS & HYPOTHESIS

1)How does the arrival date of the month influence the average daily rate (ADR) for both resort and city hotels?

Null Hypothesis (H0): There is no significant difference in the average daily rate (ADR) among different arrival dates of the month for both resort and city hotels.

Alternative Hypothesis (H1): There is a significant difference in the average daily rate (ADR) among different arrival dates of the month for both resort and city hotels.

2)What is the combined impact of the country of origin and hotel type on the average daily rate (ADR)?

Null Hypothesis (H0): There is no interaction effect between the country of origin and hotel type on the average daily rate (ADR).

Alternative Hypothesis (H1): There is a significant interaction effect between the country of origin and hotel type on the average daily rate (ADR).

3)How does lead time affect the average daily rate (ADR)?

Null Hypothesis (H0): There is no linear relationship between lead time and the average daily rate (ADR).

Alternative Hypothesis (H1): There is a significant linear relationship between lead time and the average daily rate (ADR).

4)Do reserved room type and assigned room type impact the probability of a booking being cancelled?

Null Hypothesis (H0): Reserved room type and assigned room type do not have a significant impact on the probability of a booking being cancelled.

Alternative Hypothesis (H1): There is a significant impact of reserved room type and assigned room type on the probability of a booking being cancelled.

5)How does the average daily rate (ADR) change for resort and city hotels over the years?

Null Hypothesis (H0): There is no linear relationship between the arrival year and the average daily rate (ADR) for both resort and city hotels.

Alternative Hypothesis (H1): There is a significant linear relationship between the arrival year and the average daily rate (ADR) for both resort and city hotels.

DATASET SUMMARY

The dataset has 32 variables, as outlined in the data dictionary document. These variables include various categories, binary, and numerical data. The dataset contains a total of 119,391

observations, which is more than sufficient to conduct a rigorous and dependable statistical analysis.

S.no	Name	Description
1	Hotel Type	Type of hotel, whether it is "City" or "Hotel".
2	Is Canceled	Represents whether the hotel booking is canceled or not, with values "0" for not canceled and "1" for canceled.
3	Lead Time	Time period between the number of days and booking.
4	Arrival Date Year	Year of arrival date.
5	Arrival Date Month	Month of arrival date.
6	Arrival Date Week Number	Week of arrival date.
7	Year of Arrival	Denotes the year in which the visitor arrives.
8	Day of Arrival	Specifies the month in which the visitor will arrive.
9	Weekend Nights	Denotes how many nights the visitor intends to stay over the weekend (Saturday or Sunday).
10	Weeknights	Indicates the number of nights the guest intends to stay during the week (Monday through Friday).
11	Adults	Denotes the number of adult visitors. Children and Babies represent the number of children and babies included in the booking.
12	Meal Sort	Denotes the sort of meal booked, for example, Bed & Breakfast (BB).
13	Nation of Origin	Displays the nation from which the visitor is traveling.
14	Market Segment	Indicates the market segment to which the booking belongs, such as Online Travel Agent.
15	Distribution Channel	Denotes the method by which the booking was made, such as direct or via a travel agent/tour operator.
16	Repeated Guest	Indicates whether the guest has previously visited (1) or not (0).
17	Prior Cancellations	Displays the number of prior reservations that the visitor canceled.
18	Previous Bookings Not Canceled	Shows how many past reservations the guest did not cancel.
19	Reserved Room Type	Shows the type of room that was originally booked.
20	Assigned Room Type	Specifies what kind of room the person was given.
21	Booking Changes	Indicates how many changes have been made to the reservation.
22	Deposit Type	Shows what kind of deposit was made.
23	Agent	Shows the ID of the tour company that made the reservation.
24	Company	The ID of the business or organization that made the reservation.

S.no	Name	Description
25	Days in Waiting List	The number of days the reservation was on the waiting list before it was approved.
26	Customer Type	Describes what kind of booking it is, like "Transient." The usual price per room per day is shown by the usual Daily Rate (ADR).
27	Required Car Parking Spaces	Shows how many parking spots the guest has asked for.
28	Total Special Requests	The number of special requests the guest made.
29	Ticket State	Tells what the state of the ticket is, like "Checked Out" or "Canceled."
30	Reservation Status Date	Denotes when the reservation status was last changed.

EXPLORATORY DATA ANALYSIS

In this step, we perform the data preprocessing, that the data is transformed into a structured format from the raw unstructured data and removing null values, outliers and they are replaced by using central dispersions with mean, median and mode.

EDA RESULT

```
# Count the null values for each column in the dataset
print(df.isnull().sum())
# Replace the null values by filling with mean or median for numeric columns
dc = df.select_dtypes(include=['number']).columns
df[dc] = df[dc].fillna(df[dc].mean())
print(df.isnull().sum())
```

S.no	Variable Name	Null values found
1	Children	488
2	Agent	16340
3	Company	112593

IMPLEMENTATION

ONE-WAY ANOVA MODEL

ANOVA is a statistical method used to compare means between two or more groups to determine

if there are any statistically significant differences among them. Specifically, it conducts an ANOVA analysis to examine whether statistically significant differences exist in the average daily rate ('adr') based on various levels of the categorical variable 'arrival_date_day_of_month'. These values help assess the statistical significance of the impact of 'arrival_date_day_of_month' on the 'adr' variable.

TWO-WAY ANOVA MODEL

Analysis of Variance (ANOVA) for Assessing the Impact of 'Country' and 'Hotel' on Average Daily Rate (ADR) using Ordinary Least Squares (OLS) Method**.** In this analysis, we employ the Analysis of Variance (ANOVA) technique to evaluate the influence of both the 'country' and 'hotel' variables on the Average Daily Rate (ADR), denoted by 'adr'. The model encompasses a two-way interaction between 'country' and 'hotel', aiming to explore the joint effects of these factors.

LINEAR REGRESSION MODEL-1

This linear regression model analyzes the impact of lead time on the Average Daily Rate (ADR) in a dataset. The code essentially builds a linear regression model, trains it on a subset of the data, predicts ADR values on another subset, and evaluates the model's performance using coefficients, intercept, Mean Squared Error, and R-squared.

LOGISTIC REGRESSION

This logistic regression model assesses the impact of 'reserved_room_type' and 'assigned_room_type' on the probability of a booking being canceled. The results include the confusion matrix, providing insights into true positives, true negatives, false positives, and false negatives. Additionally, the classification report presents precision, recall, F1-score, and support for each class. The overall accuracy of the model is also displayed. Essentially, the aim is to model and evaluate the relationship between room types and the likelihood of booking cancellations using

logistic regression.

LINEAR REGRESSION MODEL- 2

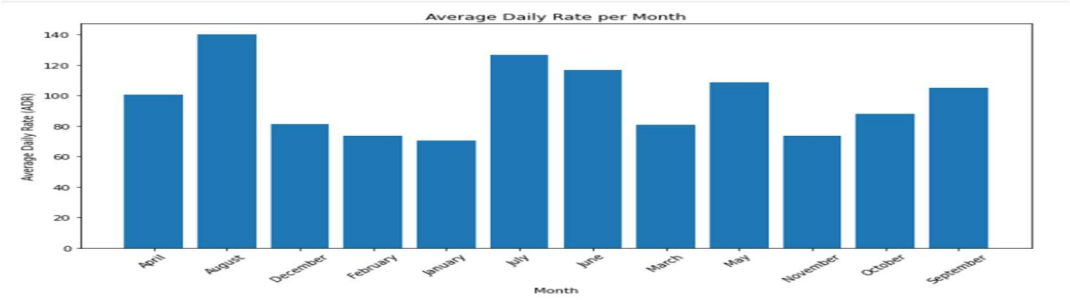
This linear regression model explores how the Average Daily Rate (ADR) changes over the years for resort and city hotels. It also provides a visual representation of how the ADR changes over the years for resort and city hotels and prints the regression coefficients and intercepts for each hotel type. This enables an understanding of the trend and relationship between 'arrival_date_year' and 'adr' for the specified hotel types.

EVALUATION AND DISCUSSION

ONE-WAY ANOVA MODEL

	sum_sq	df	F	PR(>F)
C(arrival_date_month)	5.843379e+07	11.0	2572.954399	0.0
Residual	2.464697e+08	119378.0	NaN	NaN

The F-statistic is a measure of the ratio of the variance between groups to the variance within groups. In this case, a high F-statistic (2572.95) indicates that there are significant differences in ADR across different arrival date months. The p-value (0.0) is less than the typical significance level of 0.05, suggesting strong evidence against the null hypothesis. Therefore, you can reject the null hypothesis that the means of ADR are equal across different arrival date months. In conclusion, there is evidence to suggest that the arrival date month has a significant impact on the Average Daily Rate.



The resulting bar plot visually represents the average daily rate (ADR) for each month, providing insights into how the ADR varies across different months based on the 'arrival_date_day_of_month'. The x-axis displays the months, the y-axis depicts the average daily rate, and each bar represents the mean ADR for a specific month.

TWO-WAY ANOVA MODEL

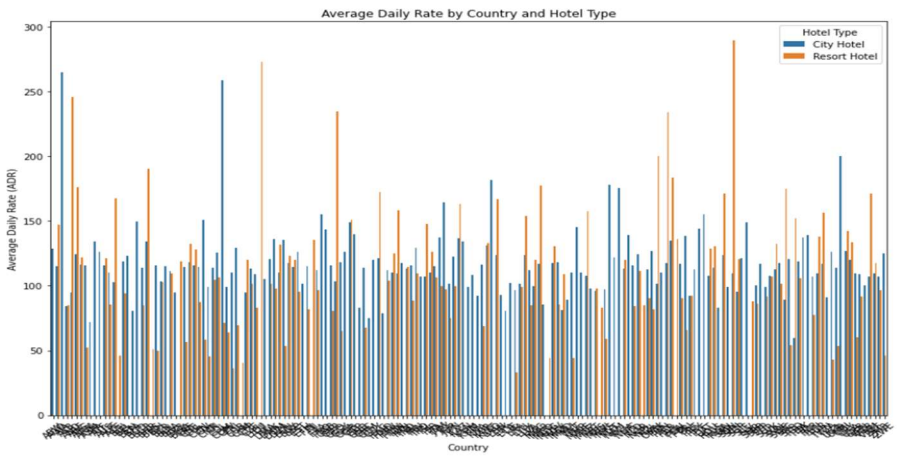
	sum_sq	df	F	PR(>F)
C(country)	1.115581e+07	176.0	26.683601	0.0
C(hotel)	-2.706684e+02	1.0	-0.113945	1.0
C(country):C(hotel)	1.081625e+07	176.0	25.871393	0.0
Residual	2.817534e+08	118611.0	NaN	NaN

C(country): The 'country' variable shows a statistically significant impact on Average Daily Rate (ADR) with a sum of squares of 1.115581e+07, 176 degrees of freedom, an F-statistic of 26.6836, and a p-value less than 0.05 ($p < 0.05$).

C(hotel): The 'hotel' variable does not show a statistically significant impact on ADR, as the sum of squares is negative (-2.706684e+02), the F-statistic is close to zero (-0.1139), and the p-value is greater than 0.05 ($p > 0.05$).

C(country):C(hotel): The interaction between 'country' and 'hotel' is statistically significant, with a sum of squares of 1.081625e+07, 176 degrees of freedom, an F-statistic of 25.8714, and a p-value less than 0.05 ($p < 0.05$).

Residual: The residual sum of squares is 2.817534e+08 with 118611 degrees of freedom. Nan values for F and p-value indicate that these are not applicable for the residual.



The resulting grouped bar chart provides a visual representation of how the Average Daily Rate (ADR) varies across different countries and hotel types. Each bar represents the mean ADR for a specific country, further divided into different hotel types, allowing for a comprehensive analysis of the impact of both country and hotel type on ADR.

LINEAR REGRESSION MODEL-1

```
Coefficients: -0.03
Intercept: 105.01
Mean Squared Error: 2277.39
R-squared: 0.004194573726446049
```

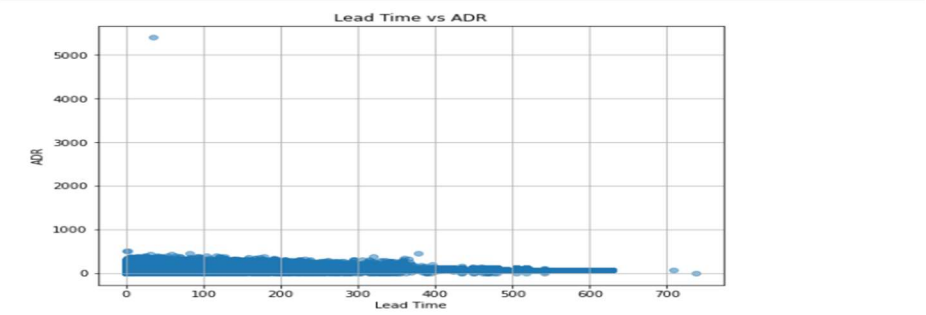
Coefficients: The coefficient of -0.03 suggests that, on average, for each unit increase in the lead time, the Average Daily Rate (ADR) decreases by 0.03 units.

Intercept: The intercept of 105.01 represents the estimated ADR when the lead time is zero.

Mean Squared Error: The Mean Squared Error of 2277.39 provides a measure of the average squared difference between predicted ADR values and actual ADR values. Lower values indicate better model performance.

R-squared: The R-squared value of 0.0042 indicates that the linear regression model explains

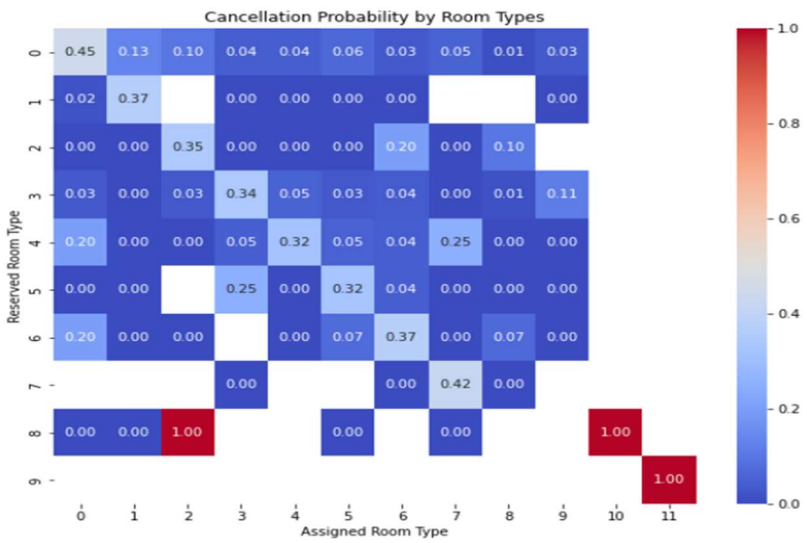
only a small proportion of the variance in the ADR. It suggests that the model may not be a strong predictor, as it explains a very limited portion of the variability in the target variable.



The resulting scatter plot visualizes the relationship between lead time and the Average Daily Rate (ADR). Each point on the plot represents a data point from the Data Frame, with lead time on the x-axis and ADR on the y-axis. The scatter plot allows for an exploration of any potential patterns or trends between lead time and ADR, helping to understand how these two variables may be correlated. The transparency of the points (alpha=0.5) can be particularly useful when dealing with many data points, as it helps to identify areas of higher point density.

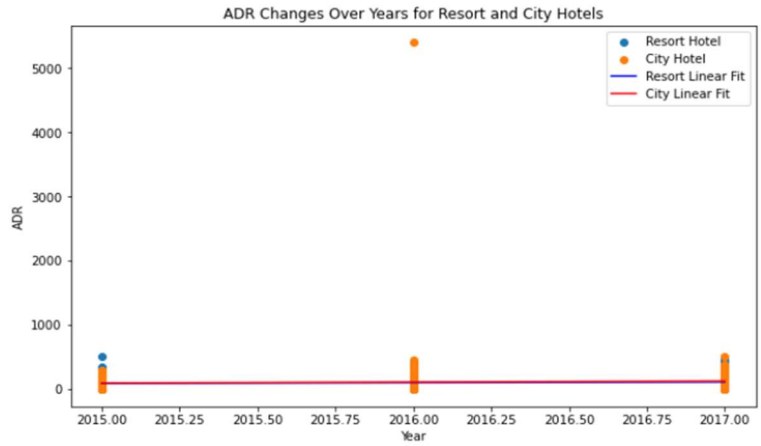
LOGISTIC REGRESSION MODEL

```
Confusion Matrix:
[[14810    97]
 [ 8969     2]]

Classification Report:
              precision    recall  f1-score   support

           0       0.62      0.99      0.77     14907
           1       0.02      0.00      0.00      8971

    accuracy                           0.62     23878
   macro avg       0.32      0.50      0.38     23878
weighted avg       0.40      0.62      0.48     23878

Accuracy: 0.6203199597956278
```



The resulting heatmap visually represents the cancellation probabilities for different combinations of reserved and assigned room types. Each cell in the heatmap corresponds to the cancellation probability for a specific pair of room types, providing insights into how these factors influence the likelihood of a booking being canceled. The color intensity indicates the cancellation probability, and annotations display the exact probability values for each cell.

LINEAR REGRESSION MODEL-2




```
Resort Hotel Coefficients: 10.869807881398792
Resort Hotel Intercept: -21819.89981958751
City Hotel Coefficients: 15.513206209169454
City Hotel Intercept: -31172.02296575127
```

The resulting plot and printed coefficients provide a visual and numerical representation of how

the Average Daily Rate (ADR) changes over the years for both resort and city hotels. The scatter plot with linear regression lines helps to identify trends and patterns in ADR variation over time for each hotel type.

Resort Hotel:

Coefficient: 10.87

Intercept: -21819.90

City Hotel:

Coefficient:15.51

Intercept: -31172.02

These coefficients indicate the change in ADR concerning each year for each hotel type.

For instance, for the city hotel, the ADR tends to increase by approximately 15.51 units per year, and the initial ADR at year 0(intercept) is around -31172.02. Similarly, for the resort hotel, the ADR tends to increase by approximately 10.87 units per year, with an initial ADR at year 0 of around -21819.90.

LIMITATIONS

The logistic model achieved high accuracy in predicting non-cancelled bookings (class 0) but performed poorly in predicting cancelled bookings (class 1), which is evident from the low precision, recall, and F1-score for class 1. This discrepancy suggests that the model might not effectively capture the patterns associated with cancelled bookings. Further feature engineering or model refinement might be necessary to improve predictions for cancellations. it seems that both country and the interaction between country and hotel type have a significant impact on the average daily rate, while the hotel type alone does not show a significant impact. Linear regression Model - 1 has a very low R-squared value (0.0042), indicating that the model explains only a small portion

of the variance in the data. Additionally, the coefficients being close to zero indicate a weak linear relationship between the predictor and the target variable. This suggests that the linear regression model might not be the best choice for capturing the relationship between lead time and ADR.

RECOMMENDATIONS AND SUGGESTIONS

The arrival date month has a statistically significant impact on the Average Daily Rate for hotels. In other words, there are significant differences in ADR among different months. Further post-hoc tests or detailed analysis could be conducted to identify specific months with significantly different ADRs. Overall, it seems that both country and the interaction between country and hotel type have a significant impact on the average daily rate, while the hotel type alone does not show a significant impact. There is a significant linear relationship between the arrival year and the average daily rate (ADR) for both resort and city hotels.

CONCLUSION

In conclusion, our project employed diverse statistical models to scrutinize key facets of the dataset. One-way ANOVA underscored the substantial impact of arrival date month on **Average Daily Rate (ADR).** The Two-way ANOVA model shed light on the influential role of 'country,' whereas 'hotel' did not significantly alter ADR. The interaction between 'country' and 'hotel' was also deemed significant. Linear Regression Model-1 depicted a modest negative association between lead time and ADR, albeit with limited explanatory power. Employing logistic regression, evaluated the influence of room types on booking cancellations, offering valuable insights via a confusion matrix and classification report. Lastly, Linear Regression Model-2 delved into ADR trends over the years for resort and city hotels, uncovering ascending patterns for both. These findings furnish valuable perspectives for decision-making within the hospitality industry.