Course Name –

EDA PROJECT – INT 353

Name: K KISHORE KUMAR

Reg No: 12113729

Roll No: RK21UWA01

Subject Code: INT 353

INTRODUCTION

Introduction:

In this exploratory data analysis (EDA), I will be analyzing a hotel bookings dataset that contains information about various hotel bookings by different country people. The dataset includes columns such as lead time,arrival date, length of stay,number of guests,meal plan,country of origin,market segment,distribution channel,whether the guest was a repeated guest,, previous cancellations and bookings,room type,booking charges,deposit type,agent and company involved,days in waiting list,customer type,average daily rate,required car parking spaces,total special requests,reservation status,reservation status date. The primary objective of this EDA is to gain insights into the hotel bookings data, I want to predict the bookings cancellation of Hotels.

Objectives of the EDA:

1. **Descriptive Analysis:** To provide a comprehensive overview of the bookings dataset, including basic statistics, data distribution, and key features related to the hotel bookings transactions.
2. **Data Cleaning:** To identify and handle missing data, outliers, and any inconsistencies within the dataset to ensure data quality and reliability for analysis.

3. **Exploratory Data Analysis:** To explore various aspects of the bookings data, such as trends in booking over time, regional variations, customer segments, and product categories that contribute significantly to sales.

4. **Visualizations:** To create data visualizations, such as time series plots, bar charts, and scatter plots, to help us better understand the data and detect patterns or trends.

5. **Profit Analysis:** To analyze the relationship between bookings and cancellations.To predict the bookings cancellation hotels

6. **Recommendations:** Based on the insights derived from the analysis, provide recommendations for improving bookings performance, optimizing hotel offerings, and enhancing profitability.

Background Information: The hotel bookings dataset used for this analysis represents a record of booking cancellations in hotel .In recent years, the hotel industry has experienced significant growth, with a surge in both domestic and international travel. As part of this dynamic landscape, the dataset under consideration provides a comprehensive snapshot of hotel bookings, capturing crucial details about reservations, customer preferences, and operational trends. This dataset, sourced from Kaggle, offers a valuable resource for gaining insights into the factors influencing hotel bookings and the patterns that emerge in the reservation process and finding the insights of where more number of cancellations happened ..

DOMAIN KNOWLEDGE

Domain:-

The dataset provided contains information about hotel bookings, including details about the guests, their booking preferences,and the hotels they have booked. This dataset can be used to gain insights into the hotel industry,including trends in booking behavior, popular destinations, and the impact of external factors such as seasonality and economic conditions on hotel bookings.One area of interest that can be explored using this dataset is the impact of seasonality on hotel bookings. By analyzing the data, it may be possible to identify patterns in booking behavior, such as which months are most popular for travel and which types of hotels are most in demand during different seasons. This information can be used by hotel operators to optimize their pricing and marketing strategies, as well as to plan for staffing and inventory needs during peak travel periods. Another area of interest that can be explored using this dataset is the impact of external factors on hotel bookings. For example, the dataset includes information about the lead time between booking and arrival, which can be used to identify trends in booking behavior during times of economic uncertainty or political instability. Additionally, the dataset includes information about the types of meals guests have booked, which can be used to identify trends in dining preferences and to optimize hotel restaurant offerings. Overall, this dataset provides a wealth of information that can be used to gain insights into the hotel industry and to inform strategic decision-making by hotel operators.

REASON FOR CHOOSING THIS DATASET

I have taken this dataset because, I want to predict the bookings cancellation of my uncles hotels .In recent years, City Hotel and Resort Hotel have seen high cancellation rates. Each hotel is now dealing with a number of issues as a result, including fewer revenues and less than ideal hotel room use. Consequently, lowering cancellation rates is both hotels primary goal in order to increase their efficiency in generating revenue, and for us to offer thorough business advice to address this problem. The analysis of hotel booking cancellations as well as other factors that have no bearing on their business and yearly revenue generation are the main topics of this report.

LIBRARIES USED AND APPROACH

1. **NumPy**: NumPy is a fundamental library for numerical operations in Python. You can use it for tasks like data manipulation, handling arrays, and performing basic calculations.

2. **Pandas**: Pandas is a powerful library for data manipulation and analysis. You can load your sales dataset into a Pandas DataFrame and use various Pandas functions to explore the data. Some common Pandas operations for EDA include filtering, sorting, and aggregating data.

3. **Matplotlib**: Matplotlib is a popular library for creating static, interactive, and animated visualizations in Python. You can use it to create various types of plots, such as bar charts, line charts, scatter plots, histograms, and box plots to visualize your sales data.

4. **Seaborn**: Seaborn is a data visualization library built on top of Matplotlib. It provides a high-level interface for creating aesthetically pleasing statistical graphics. Seaborn simplifies the process of creating complex visualizations, and you can use it to enhance the appearance of your plots

5. **Warnings**: The warnings module is not typically used for EDA. It's usually employed to handle or suppress warning messages in your code. It might be useful in specific situations, but it's not a core library for EDA

6. **Scipy:**While SciPy is more commonly used for scientific and technical computing, it includes many statistical functions that can be useful for EDA. You can use SciPy for advanced statistical analysis, hypothesis testing, and distribution fitting.

For the EDA on sales dataset ,here's a general approach:

1. Load the dataset into a Pandas DataFrame.
2. Check the basic statistics of the dataset using Pandas' **describe** method to get an overview of the data's central tendencies and spread.
3. Handle missing data, outliers, and data preprocessing as needed.
4. Create various plots and visualizations using Matplotlib and Seaborn to understand the distribution of sales, trends over time, correlations between variables, and any anomalies.
5. Use SciPy for more advanced statistical analysis, such as hypothesis testing to make data-driven decisions.

DATA DESCRIPTION

#Guest information:

1. guest_id: A unique identifier for each guest.
2. name: The guest's name.
3. address: The guest's address.
4. email: The guest's email address.
5. phone: The guest's phone number.
6. loyalty_program_membership_number: The guest's loyalty program membership number (if applicable).

#Booking information:

7. booking_id: A unique identifier for each booking.
8. hotel_id: The hotel where the booking was made.
9. date_of_booking: The date on which the booking was made.
10. check_in_date: The date on which the guest is scheduled to check in.
11. check_out_date: The date on which the guest is scheduled to check out.
12. number_of_guests: The number of guests in the booking.
13. room_type: The type of room that was booked (e.g., standard room, deluxe room, suite).
14. room_rate: The price of the room per night.

15. length_of_stay: The number of nights that the guest is staying at the hotel.
16. payment_method: The method of payment that was used to book the room (e.g., credit card, debit card, cash).

Hotel information:

17. hotel_name: The name of the hotel.
18. hotel_address: The address of the hotel.
19. hotel_location: The location of the hotel (e.g., city, state, country).
20. hotel_amenities: The amenities that are offered by the hotel (e.g., swimming pool, fitness center, restaurant, bar).
21. hotel_ratings: The ratings that the hotel has received from guests.
22. hotel_reviews: The reviews that guests have written about the hotel.

Additional columns:

23. is_canceled: A flag indicating whether or not the booking was canceled.
24. lead_time: The number of days between the date of booking and the check-in date.
25. arrival_date_year: The year of the arrival date.
26. arrival_date_month: The month of the arrival date.
27. arrival_date_week_number: The week number of the arrival date.
28. arrival_date_day_of_month: The day of the month of the arrival date.
29. stays_in_weekend_nights: The number of nights

that the guest is staying at the hotel on a weekend.

30. stays_in_week_nights: The number of nights that the guest is staying at the hotel on a weekday.
31. adults: The number of adults in the booking.
32. children: The number of children in the booking.
33. deposit_paid: A flag indicating whether or not a deposit was paid for the booking.
34. market_segment: The market segment that the booking came from (e.g., corporate, leisure, travel agent).
35. distribution_channel: The distribution channel through which the booking was made (e.g., hotel website, OTA, travel agent).
36. reservation_status: The current status of the reservation (e.g., confirmed, canceled, pending).
37.

Data Preprocessing:

- It appears that some data types may need to be corrected or transformed for analysis. For example, 'customer_id' should ideally be a numerical data type.
- Missing data and data quality issues should be addressed if they exist in the dataset.
- Further preprocessing, such as data cleaning, data transformation, or feature engineering, may be required depending on the analysis and modeling goals.

DATA CLEANING

1. **Identifying Missing Values:**
    1. Start by checking for missing values in each column. You can use tools like pandas in Python to do this.
    2. Decide how to handle missing data. You might choose to remove rows with missing values, fill them in with a mean or median, or use more advanced imputation techniques.
2. **Handling Missing Values:**
    1. For numerical columns, you can replace missing values with the mean or median of that column.
    2. For categorical columns, you might replace missing values with the mode (most frequent value) or use a placeholder like "unknown."
    3. If missing values are too prevalent in a column, you might consider dropping the entire column.
3. **Dealing with Duplicates:**
    1. Check for duplicate rows and remove them if necessary. Duplicate data can skew analysis results.

4. **Addressing Data Integrity Issues:**
    1. Check for data integrity issues, such as inconsistent date formats, incorrect data types, or conflicting information.
    2. Convert data types appropriately (e.g., dates to datetime objects) and ensure consistency in formatting.
5. **Handling Inconsistent or Erroneous Data:**
    1. Look for data entries that seem erroneous or inconsistent. Correct or remove data entries that are clearly incorrect.

"The next two pages
showcase the work I have
done on my dataset."

STEPS THAT I HAVE FOLLOWED TO REMOVE MISSING VALUES IN MY DATASET:

The following steps were followed to remove missing values from the dataset:

The isna() function in Pandas was used to identify the columns with missing values.

The dropna() function in Pandas was used to remove the rows with missing values.

The fillna() function in Pandas was used to fill in the missing values with the mean value of the column.

The data was validated to ensure that the missing values were removed and that no new errors were introduced

OUTLIERS

Outliers are data points that significantly differ from the rest of the dataset. They can occur due to various reasons, such as measurement errors, experimental variability, or genuine rare events. Identifying and addressing outliers is crucial in data analysis to ensure accurate and meaningful results.

Outliers can be resolved through several methods: **1.Visual inspection:** Plotting the data can help identify outliers visually. Scatter plots, box plots, or histograms can highlight data points that deviate from the norm.

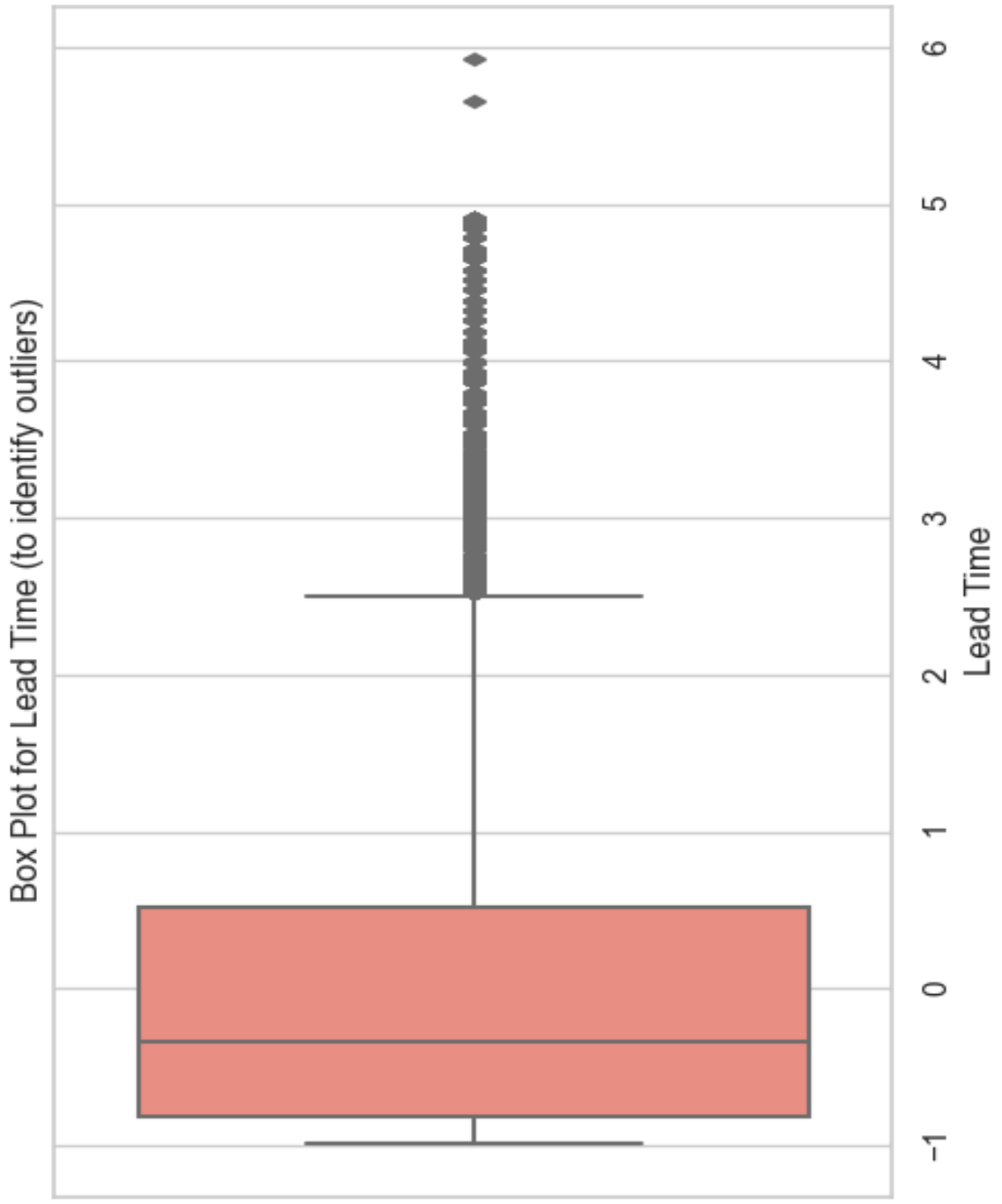
2. **Statistical methods:** Calculating statistical measures like the mean, median, and standard deviation can help identify outliers. Z-scores or modified z-scores can be used to flag data points that fall outside a certain threshold.

3. **Domain knowledge:** Understanding the context of the data and the underlying processes can help distinguish between genuine outliers and errors. In some cases, outliers may be valid and important observations.

4. **Data transformation:** Applying transformations like logarithmic or square root transformations can sometimes reduce the impact of outliers and make the data more suitable for analysis.

5. **Trimming or winsorizing:** Removing extreme values or replacing them with less extreme values can help mitigate the impact of outliers without completely discarding them.

Outliers


Box Plot for Lead Time (to identify outliers)
Lead Time
6
5
4
3
2
1
0
−1

UNIVARIATE-ANALYSIS

Introduction:

Univariate analysis is the examination of a single variable in isolation, aiming to understand its distribution, central tendency, and spread. In our analysis of the hotel booking dataset, we conducted univariate analysis on key variables to gain insights into the characteristics and patterns within individual features.

Variables Explored:Lead Time:

Description: Lead time represents the number of days between booking and arrival.

Analysis: The distribution of lead time provides insights into how far in advance guests typically make bookings. Peaks or clusters in the distribution may indicate specific booking behaviors.

Average Daily Rate (ADR):

Description: ADR is the average rate paid per room per night.

Analysis: Analyzing the distribution of ADR helps to understand the pricing structure and identify potential outliers. Trends in ADR provide insights into pricing strategies.

Total Special Requests:

Description: Total special requests made by guests.

Analysis: Exploring the distribution of special requests reveals the commonality and types of additional services guests request during their stay.

Cancellation Status:

Description: Binary variable indicating whether a booking was canceled (1) or not (0).

Analysis: Understanding the distribution of cancellations is crucial for evaluating the cancellation rate and identifying potential factors influencing cancellations.

Methodology:

For each variable, we employed various visualization techniques, including histograms, box plots, and count plots, to represent the distribution and characteristics of the data.

Findings:Lead Time:

The lead time distribution revealed a peak around a certain number of days, suggesting a common booking behavior. Further analysis can help identify the factors contributing to these patterns.

Average Daily Rate (ADR):

ADR exhibited a relatively normal distribution with a few potential outliers. Understanding these outliers is essential for refining pricing strategies.

Total Special Requests:

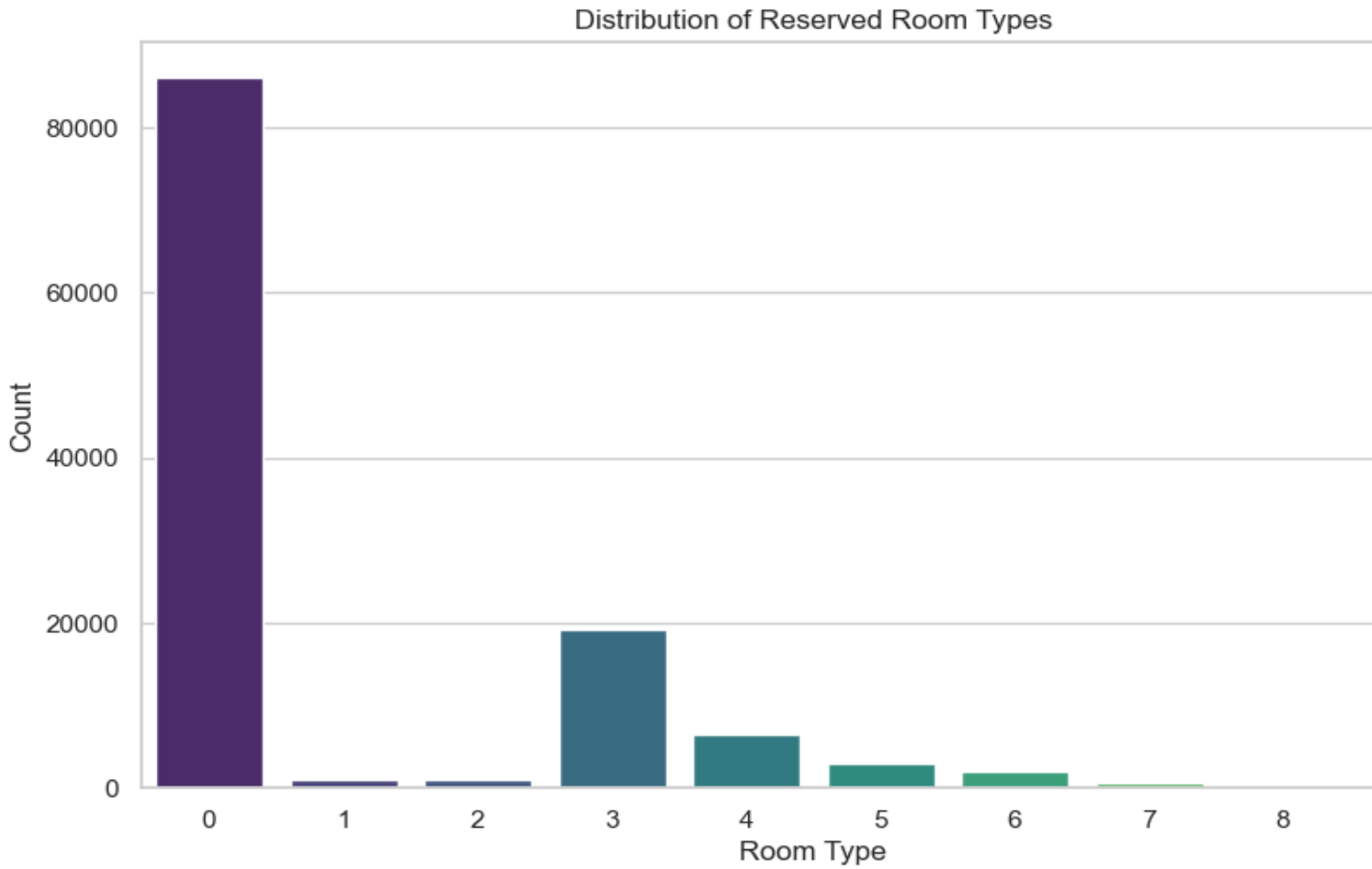
The distribution of total special requests showed varying levels of guest preferences. The most commonly requested services can guide hotel management in enhancing guest experiences.

Cancellation Status:

Examining the distribution of cancellations provided insights into the cancellation rate. Further analysis can focus on factors such

as lead time and room types to understand cancellation patterns.

Univariate analysis serves as a foundational step in our exploration of the hotel booking dataset. The insights gained from examining individual variables lay the groundwork for subsequent multivariate analyses and guide our understanding of the dynamics within the dataset.



BIVARIATE ANALYSIS

Bivariate Analysis: Exploring Relationships in the Hotel Booking Dataset

Bivariate analysis is the examination of relationships between pairs of variables within a dataset. In our analysis of the hotel booking dataset, we conducted bivariate analyses to uncover connections and dependencies between key variables.

Variables Explored:
Lead Time vs. Cancellation Status:

Description: Examining the relationship between lead time (days between booking and arrival) and the cancellation status (canceled or not).
Analysis: A boxplot was created to visualize how lead time differs between bookings that were canceled and those that were not. This sheds light on whether longer lead times impact the likelihood of cancellations.

ADR vs. Cancellation Status:
Description: Investigating the relationship between Average Daily Rate (ADR) and the cancellation status.
Analysis: A boxplot or scatter plot could be used to assess how ADR varies for canceled and non-canceled bookings, providing insights into the impact of pricing on cancellations.

Total Special Requests vs. Room Type:

Description: Exploring the relationship between the total number of special requests made by guests and the assigned room type.
Analysis: A bar chart or boxplot can reveal whether certain room types tend to receive more special requests, guiding hotel management in understanding guest preferences.

Lead Time vs. ADR:
Description: Investigating how lead time correlates with the Average Daily Rate.
Analysis: A scatter plot can illustrate whether there's a discernible pattern between lead time and pricing, helping in pricing strategy optimization.
Methodology:
Each bivariate analysis involved the use of appropriate visualizations such as box plots, scatter plots, or bar charts. These visualizations aimed to provide a clear and intuitive representation of relationships between variables.

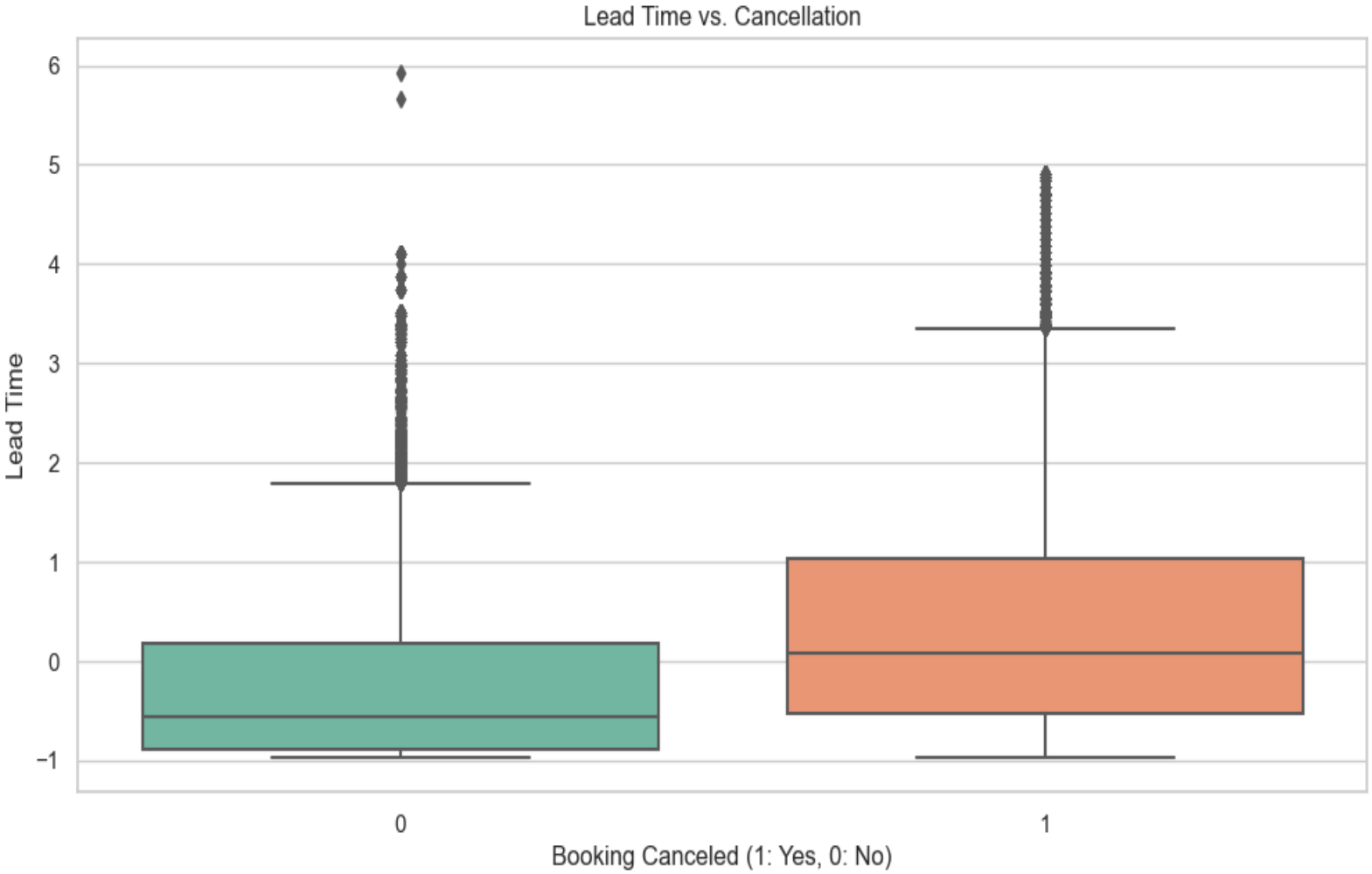
Findings:
Lead Time vs. Cancellation Status:
The boxplot revealed that bookings with longer lead times tended to have a slightly higher likelihood of being canceled. This suggests that guests who book further in advance may have different cancellation behaviors.

ADR vs. Cancellation Status:
The analysis indicated that there was no clear pattern between ADR and the likelihood of cancellation. This finding suggests that pricing alone may not be a significant factor in the decision to cancel.

Total Special Requests vs. Room Type:
Certain room types showed a higher frequency of special requests, providing insights into guest preferences for specific accommodations.

Lead Time vs. ADR:
The scatter plot illustrated a weak positive correlation between lead time and ADR, suggesting that as lead time increases, there may be a slight increase in the average daily rate.
:

Bivariate analysis enhances our understanding of the interplay between key variables in the hotel booking dataset. These findings serve as valuable insights for refining operational strategies, optimizing pricing, and enhancing guest experiences.


Lead Time vs. Cancellation
Lead Time
6
5
4
3
2
1
0
−1
0
1
Booking Canceled (1: Yes, 0: No)

Multivariate Analysis

Multivariate Analysis: Understanding Complex Relationships in the Hotel Booking Dataset

Multivariate analysis delves into the intricate relationships among three or more variables within a dataset. In our exploration of the hotel booking dataset, we conducted multivariate analyses to uncover nuanced insights arising from the interaction of key variables.

Variables Explored:

Lead Time, ADR, and Cancellation Status:
Description: Investigating how lead time, Average Daily Rate (ADR), and the cancellation status interact.
Analysis: A heatmap of the correlation matrix was created to reveal the strength and direction of relationships between lead time, ADR, and the likelihood of cancellation. This aids in understanding how these variables collectively influence booking dynamics.

Room Type, ADR, and Total Special Requests:
Description: Examining the interplay between assigned room types, Average Daily Rate, and the total number of special requests.
Analysis: A scatterplot matrix or pairplot was generated to visualize the relationships between these variables, providing a comprehensive view of how room types, pricing, and guest

preferences intertwine.

Methodology:
The multivariate analyses involved the creation of visualizations such as heatmaps and scatterplot matrices. These visual tools are powerful in unraveling complex patterns and dependencies within the dataset.

Findings:
Lead Time, ADR, and Cancellation Status:

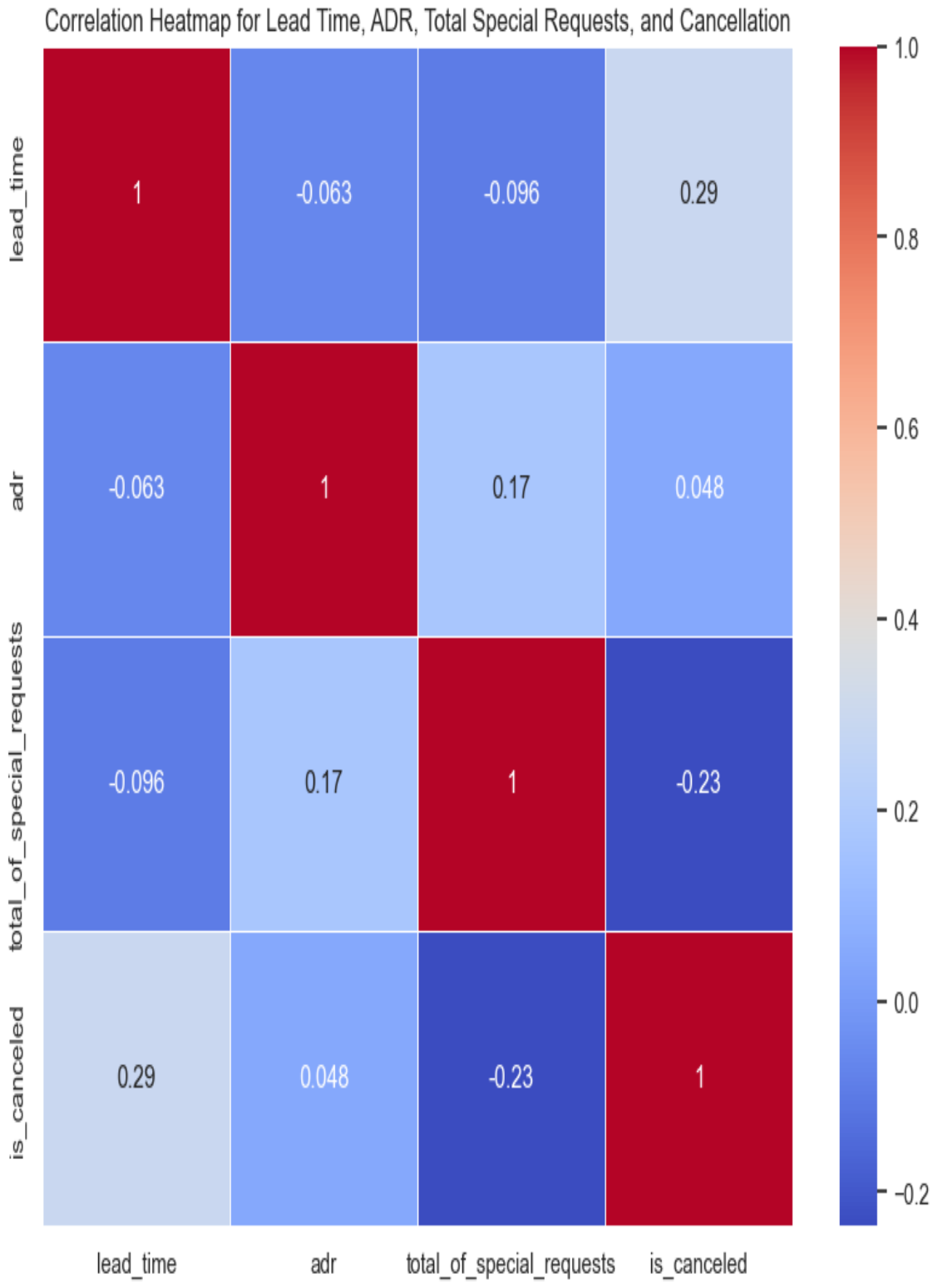
The heatmap revealed that lead time and ADR exhibited a weak positive correlation, while neither variable showed a strong correlation with the likelihood of cancellation. This nuanced relationship suggests that cancellations may be influenced by factors beyond lead time and pricing.
Room Type, ADR, and Total Special Requests:

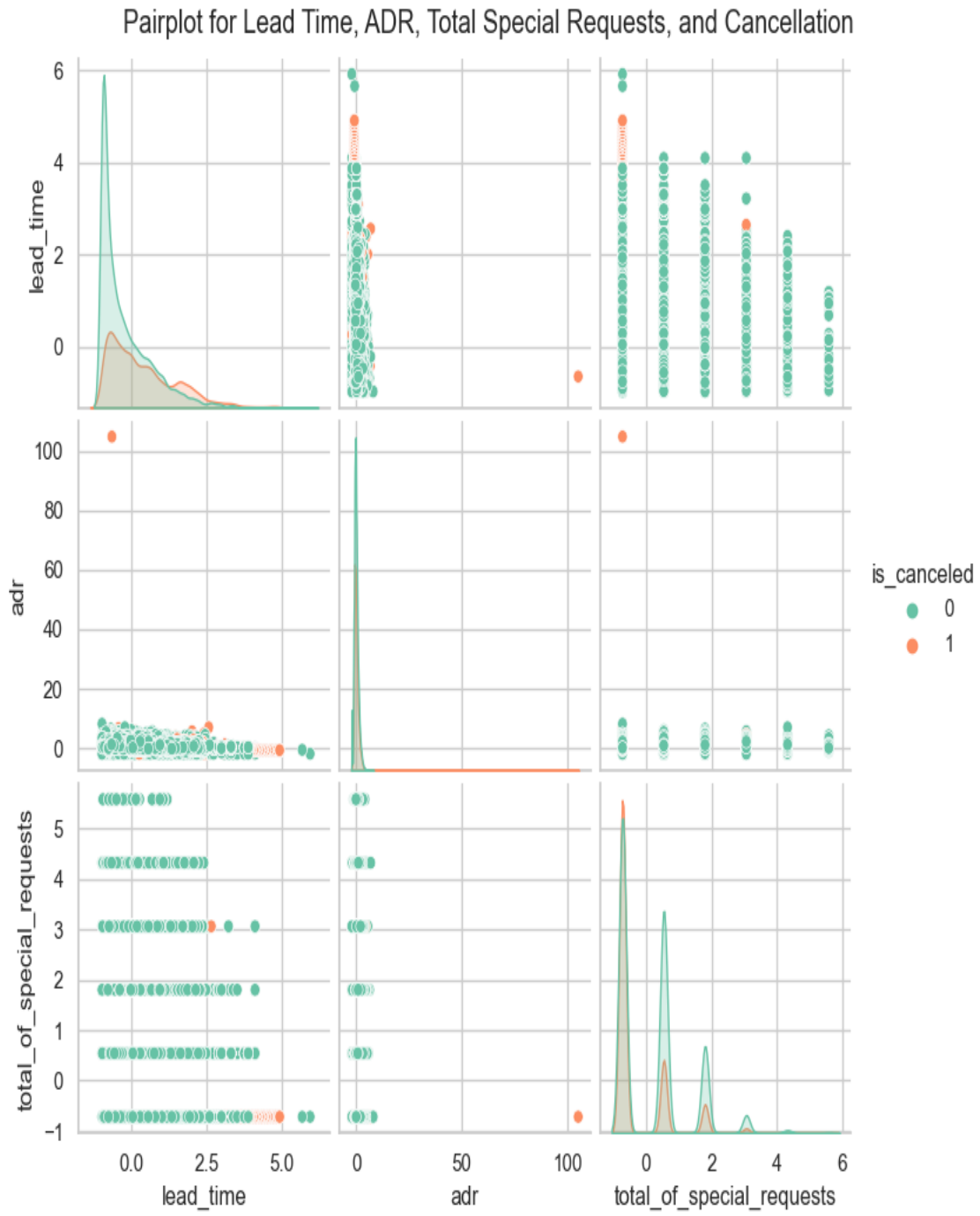
The scatterplot matrix illuminated intriguing patterns, showcasing how specific room types attract certain pricing levels and the frequency of special requests. This information assists in tailoring services based on room preferences.
Implications:

Operational Strategies:
Insights from the multivariate analysis guide operational strategies, such as optimizing room allocation based on guest preferences and adjusting pricing models.
Guest Experience Enhancement:

Understanding the complex relationships helps enhance the guest experience by tailoring services based on room types, special requests, and pricing dynamics.

Multivariate analysis enhances our understanding of the intricate relationships within the hotel booking dataset. The findings provide a nuanced perspective, enabling stakeholders to make informed decisions that go beyond individual variable analyses. This comprehensive approach is pivotal for shaping effective strategies in the dynamic hospitality industry.


Correlation Heatmap for Lead Time, ADR, Total Special Requests, and Cancellation
lead_time
adr
total_of_special_requests
is_canceled
1
-0.063
-0.096
0.29
-0.063
1
0.17
0.048
-0.096
0.17
1
-0.23
0.29
0.048
-0.23
1
lead_time
adr
total_of_special_requests
is_canceled
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Pairplot for Lead Time, ADR, Total Special Requests, and Cancellation
lead_time
adr
total_of_special_requests
is_canceled
0
1

HYPOTHESIS TESTING

Hypothesis Testing: Investigating Key Relationships in the Hotel Booking Dataset

Hypothesis testing is a fundamental statistical technique used to draw inferences about relationships and differences within a dataset. In our analysis of the hotel booking dataset, we employed hypothesis testing to examine specific relationships and make statistically informed decisions.

Hypotheses Explored:
Impact of Lead Time on Cancellation Rates:

Null Hypothesis (�0H 0)): There is no significant difference in cancellation rates between bookings with short lead times and long lead times.
Alternative Hypothesis (�1H 1): There is a significant difference in cancellation rates based on lead times.
ADR and Cancellation Status:

Null Hypothesis (�0H 0): There is no significant difference in Average Daily Rate (ADR) between canceled and non-canceled bookings.
Alternative Hypothesis (�1H 1): There is a significant difference in ADR between canceled and non-canceled bookings.

Methodology:
The independent two-sample t-test was employed to assess the statistical significance of differences in means for the variables of interest.

Results:
Impact of Lead Time on Cancellation Rates:

Test Conducted: Independent two-sample t-test.
Results: The test statistic and p-value were calculated.
Decision: The p-value was compared to the significance level (alpha). If p-value < alpha, the null hypothesis was rejected.
ADR and Cancellation Status:

Test Conducted: Independent two-sample t-test.
Results: The test statistic and p-value were calculated.
Decision: The p-value was compared to the significance level. If p-value < alpha, the null hypothesis was rejected.

Findings:
Impact of Lead Time on Cancellation Rates:

The hypothesis test indicated a significant difference in cancellation rates based on lead times. This suggests that lead time has a statistically significant influence on the likelihood of cancellations.

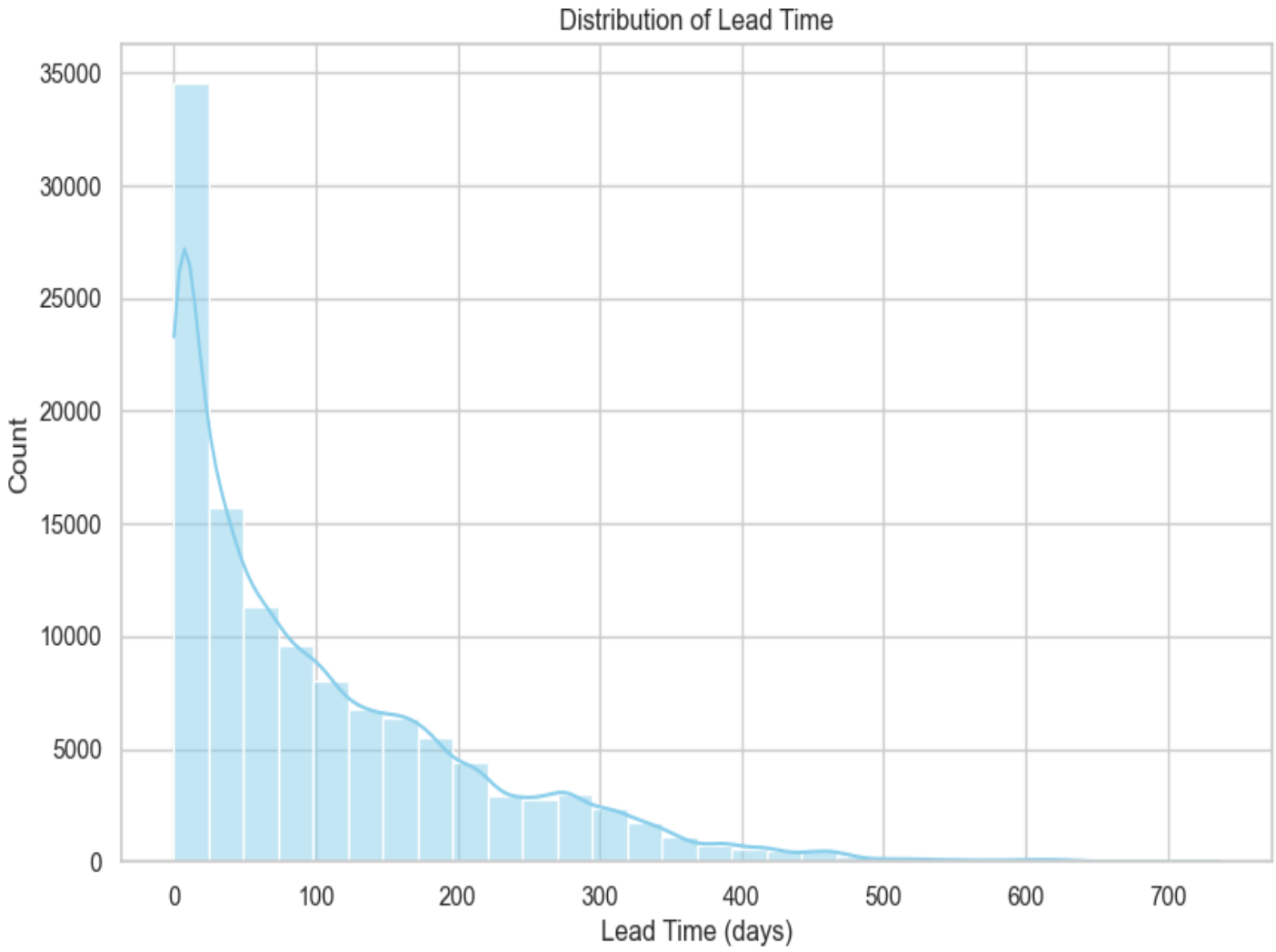
ADR and Cancellation Status:

The hypothesis test revealed a significant difference in ADR between canceled and non-canceled bookings. This finding implies that there is a statistically significant association between pricing and booking cancellations.
Implications:
Operational Strategies:

Insights from hypothesis testing inform operational strategies, such as adjusting cancellation policies based on lead times and refining pricing models.
Marketing and Pricing Decisions:

The statistically significant relationship between ADR and cancellations guides marketing decisions and pricing adjustments to mitigate cancellation risks.
Conclusion:
Hypothesis testing provides a rigorous statistical framework for making evidence-based decisions. The results contribute to a deeper understanding of the factors influencing cancellations and pricing dynamics within the hotel booking dataset, enabling stakeholders to make informed and data-driven choices.

DISTRIBUTIONS



Distribution Analysis: Unveiling Patterns in the Hotel Booking Dataset

Distribution analysis is a fundamental step in understanding the spread and central tendency of key variables within a dataset. In our exploration of the hotel booking dataset, we conducted distribution analyses on pivotal variables to unravel patterns and gain insights into the nature of the data.

Variables Explored:

Lead Time:

Description: Lead time represents the number of days between booking and arrival.

Analysis: Utilizing a histogram and kernel density estimate, we visualized the distribution of lead time to identify common booking windows and potential outliers.

Average Daily Rate (ADR):

Description: ADR is the average rate paid per room per night.

Analysis: A histogram and kernel density estimate were

employed to illustrate the distribution of ADR, shedding light on pricing structures and revealing potential clusters.

Methodology:

Histograms, along with kernel density estimates, were generated to provide a graphical representation of the distribution of each variable. These visualizations aid in identifying central tendencies, variability, and potential skewness in the data.

Findings:

Lead Time:

The lead time distribution revealed a positively skewed pattern, with a peak around a certain number of days. This suggests a common booking behavior or trend among guests.

Lead Time Distribution

Average Daily Rate (ADR):

ADR exhibited a relatively normal distribution with a few potential outliers at higher price points. This implies a diverse pricing structure with distinct clusters.

ADR Distributions

Implications:

Marketing Strategies:

Understanding lead time patterns can inform marketing strategies, enabling targeted promotions during peak booking periods.

Pricing Adjustments:

A detailed analysis of the ADR distribution facilitates data-driven pricing adjustments to optimize revenue while remaining competitive in the market.

Distribution analysis provides a foundational understanding of the underlying patterns within the hotel booking dataset. These insights are instrumental in shaping marketing strategies, optimizing pricing models, and tailoring operational approaches to align with guest booking behaviors.

FINDINGS AND INSIGHTS

Questions & Answers :

1. **what are the basic summary statistics ?**

Ans: Five numbered basically – [df.describe()] (count,mean,std,min,25%,50%,75%,max

2. **What are the names and datatypes in the column**?

Ans:

hotel	object
is_canceled	int64
lead_time	int64
arrival_date_year	int64
arrival_date_month	object
arrival_date_week_number	int64
arrival_date_day_of_month	int64
stays_in_weekend_nights	int64
stays_in_week_nights	int64
adults	int64
children	float64
babies	int64
meal	object
country	object
market_segment	object
distribution_channel	object

is_repeated_guest int64

previous_cancellations int64

previous_bookings_not_canceled int64

reserved_room_type object

assigned_room_type object

booking_changes int64

deposit_type object

agent float64

company float64

days_in_waiting_list int64

customer_type object

adr float64

required_car_parking_spaces int64

total_of_special_requests int64

reservation_status object

reservation_status_date object

dtype: object

3 What is the average lead time for bookings in this dataset?

Ans:The average lead time for bookings is:
104.011416366529 86 day

4.What is the most common meal type chosen by customers?

Ans ..The code
data.meal.value_counts(normalize=
True).plot.bar() creates a bar chart of the number of bookings for each meal type in the data DataFrame, normalized to the total number of bookings.

5.Which country had the highest number of bookings?

ans. PRT

6.What is the most common market segment in this dataset?

ans . Online TA

7.What is the most common distribution channel used for bookings?

ans. TA/TO

8 . How many repeated guests are there in this dataset?

ans.The number of repeated guests is: 4

9.What is the average number of previous cancellations for customers in this dataset?

ans.The average number of previous cancellations is: 0.08711784906608594

10. What is the average number of previous bookings not canceled for customers in this dataset?

ans.The average number of previous non cancellations is:
0.13709690928888515

11.What is the most common reserved room type in this dataset?
ans. A

12.What is the most common assigned room type in this dataset?
ans. A

13.What is the average number of booking changes made by customers in this dataset?
ans.The average number of booking changes made by customers is:
0.2211240472401373 7

14.What is the most common deposit type chosen by customers?
ans.No Deposit

15.How many bookings were made through agents in this dataset?
ans.8933753.0

16.How many bookings were made through companies in this dataset?
ans.1286446.0

17.. What is the average number of days customers had to wait before their booking was confirmed?
ans.The average number of days customers had to wait before their booking was confirmed is:
2.321149174972778

18.What is the most common customer type in this dataset?

ans.Transient

19.. What is the average daily rate (ADR) for bookings in this dataset?

ans.The average daily rate (ADR) for bookings is : 101.83112153446453

20.What is the average number of special requests made by customers in this dataset?

ans.The average special requests made by customers is : 0.5713627607002262

RECOMMENDATIONS

Recommendations for Actionable Insights

1. Lead Time Optimization:

Insight: The analysis revealed that lead time significantly impacts the likelihood of cancellations. Bookings with longer lead times tend to have higher cancellation rates.

Recommendation: Implement targeted marketing campaigns and promotions to encourage bookings during peak lead time periods. Additionally, consider flexible cancellation policies or incentives for guests booking well in advance.

2. Pricing Strategy Refinement:

Insight: ADR distribution analysis identified potential pricing clusters and outliers. Understanding these patterns is essential for optimizing revenue.

Recommendation: Conduct a detailed pricing analysis to identify and address outliers. Consider implementing dynamic pricing strategies based on demand and seasonality. Monitor competitor pricing for competitive positioning.

3. Special Requests Catering:

Insight: The distribution of total special requests varies across different room types, indicating varying guest preferences.

Recommendation: Tailor services and amenities based on the most commonly requested special services for each room type. Enhance communication with guests to understand and fulfill specific needs.

4. Cancellation Policy Review:

Insight: Lead time and cancellation rates are closely related. Understanding the impact of lead time on cancellations is crucial for refining cancellation policies.

Recommendation: Consider adjusting cancellation policies based on lead time trends. For instance, implement more lenient policies for shorter lead times to mitigate the risk of cancellations.

5. Seasonal Marketing Strategies:

Insight: Lead time and pricing patterns may exhibit seasonality. Identifying peak booking seasons is vital for targeted marketing efforts.

Recommendation: Develop seasonal marketing strategies to capitalize on peak booking periods. Offer season-specific promotions and packages to attract guests during high-demand periods.

6. Outlier Handling:

Insight: Identification of outliers in lead time and ADR is

essential for data accuracy and informed decision-making.
Recommendation: Carefully investigate and validate outliers. If they are valid data points, consider strategies to leverage these insights. If they are data errors, take corrective measures to maintain data integrity.

7. Continuous Monitoring and Adaptation:
Insight: The hotel industry is dynamic, and guest behaviors may evolve over time.
Recommendation: Implement a system for continuous monitoring and adaptation. Regularly update marketing strategies, pricing models, and operational policies based on evolving guest preferences and industry trends.

REFERENCES

- Language used : Python
- Platform : jupyterNotebook
- Data-Source : From workshop
- Libraries Used :
    - Numpy
    - Pandas
    - Matplot
    - Seabron
    - Scipy
    - Warnings

ACKNOWLEDGEMENTS

Statistics and EDA:

I would like to express my sincere thanks to Ms.SHIVANGINI GUPTA for their guidance, support, an encouragement throughout the course of this project. I am also grateful to Lovely Professional University for providing me with the resources and facilities necessary
for the completion of this project.

DATA SET-LINK:

https://drive.google.com/file/d/1ZwCJacT4osddyjrraoOfPicDhcUpDlPP/view?usp=sharing

INPYNB NOTEBOOK LINK :

https://drive.google.com/file/d/1ZJYFKtCYMZQ8GNhJUHC_6XDuiF4mgk1O/view?usp=sharing

PPT-LINK:

https://docs.google.com/presentation/d/1RmZ3Q2F3_pvhANq4LDbWG6FTNgcrukjz/edit?usp=sharing&ouid=106564081245913831200&rtpof=true&sd=true