Cairo university
Faculty of Computers & Artificial Intelligence Operations Research & Decision Support Department

Hotel Data Insights

The Graduation Project Submitted to

The Faculty of Computers and Artificial

Intelligence, Cairo University

In Partial Fulfillment of the

Requirements for the

Bachelor Degree

In

Operations Research and Decision Support

Under Supervision of:

Prof: Ihab El Khodary

CAIRO

UNIVERSITY

February,2025

Cairo university
Faculty of Computers & Artificial Intelligence Operations Research & Decision Support Department

Hotel Data Insights

Abdelrahman Ahmed Barbary **20190282**

Mohamed Yousef Abdelfatah **20190791**

Reham Ashraf fathy **20201079**

Aya Aboelsood Hamdy **20200102**

Marwa Mahmoud Esmail **20201160**

Under Supervision of:

Prof: Ihab El Khodary

CAIRO

UNIVERSITY

February,2025

ABSTRACT

The hospitality industry has experienced significant transformations in recent years, driven by advancements in data analytics and changing consumer behaviors. This project presents a data-driven approach to uncover actionable insights from hotel data, aiming to optimize operational efficiency, enhance customer satisfaction, and maximize profitability. The analysis focuses on key factors such as booking patterns, customer preferences, room pricing strategies, and seasonal demand fluctuations. By utilizing machine learning techniques, statistical analysis, and advanced data visualization, the project identifies critical trends and provides predictive models to forecast demand and optimize pricing strategies.

The project leverages both structured and unstructured data, including historical booking data, customer reviews, and market trends, to provide a comprehensive understanding of the hotel's performance. A range of analytical techniques, including clustering, regression analysis, and sentiment analysis, are applied to extract valuable insights from the dataset. Predictive models, such as time series forecasting and machine learning algorithms, are used to predict occupancy rates, identify peak seasons, and suggest dynamic pricing models based on real-time market conditions.

Furthermore, the project emphasizes the importance of personalized customer experiences, utilizing data insights to propose strategies for improving customer retention and satisfaction. Case studies from real-world hotel datasets demonstrate the practical application of the analysis, showcasing the effectiveness of data-driven decision-making in optimizing hotel operations. This approach not only enhances profitability but also ensures that hotels remain competitive in an increasingly data-driven marketplace. By integrating these insights into strategic planning, hotel managers can make informed decisions, ultimately leading to more efficient operations, improved customer experiences, and increased financial performance.

DECLARATION

We hereby declare that our dissertation is entirely our work and genuine / original. We understand that in case of discovery of any PLAGIARISM at any stage, our group will be assigned an F (FAIL) grade and it may result in withdrawal of our Bachelor's degree.

Group members:

Name	Signature
Abdelrhman Ahmed Barbary	_____
Marwa Mahmoud Ismail	_____
Reham Ashraf Fathy	_____
Mohamed Youssef Abdelfatah	_____
Aya Aboelsood Hamdy	_____

TABLE OF CONTENTS

TABLE OF FIGURES

CHAPTER 1
Introduction

1.1 Introduction

In today's competitive hotel industry, it's crucial to understand customer behavior and optimize operations. This project uses a dataset of 140,000 hotel bookings to explore three key areas: customer segmentation, time series forecasting, and cancellation forecasting.

We'll use customer segmentation to personalize marketing and improve guest experiences by grouping customers based on their preferences. This allows hotels to tailor their services, boosting satisfaction and loyalty.

Time series forecasting will help predict future booking trends and occupancy rates, enabling better decisions about staffing, pricing, and resource allocation for maximum revenue.

Finally, cancellation forecasting will analyze past cancellation data to identify patterns and reasons for cancellations. This will help hotels proactively minimize financial losses.

Interactive dashboards will be created to visualize key metrics like Average Daily Rate (ADR) and cancellation insights, empowering managers to make data-driven decisions.

By combining these three strategies, this project offers a complete solution for improving operational efficiency and guest satisfaction in the hospitality industry.

1.2 – Project overview

The hotel industry is a crucial component of the global hospitality sector, encompassing a wide range of establishments from luxury resorts to budget accommodations. Hotels provide lodging, meals, and other guest services to travelers, contributing significantly to tourism and local economies. With the rise of global travel and tourism, the industry has seen substantial growth, but it also faces increasing competition and evolving customer expectations. As the hospitality industry becomes increasingly competitive, hotels must continuously seek ways to increase revenue and maintain a competitive edge. By leveraging advanced data analytics, hotels can:

- Understand Customer Preferences: Gain deeper insights into customer preferences and behavior patterns.
- Optimize Pricing Strategies: Implement dynamic pricing strategies that respond to market demand, maximizing revenue.
- Improve Marketing Efforts: Develop targeted marketing campaigns based on customer segmentation, leading to higher engagement and conversion rates.
- Enhance Operational Efficiency: Use accurate demand forecasts for better resource planning and management, reducing the risk of overbooking or underutilization

This project focuses on three key areas to help hotels enhance their operational efficiency and profitability:

1. Customer Segmentation and Personalization analyze the different customer segments within the hotel industry and gain a deeper understanding of the customer base and identify distinct groups based on various criteria such as demographics, behavior patterns, preferences, and past interactions with the hotel.
2. Forecasting Demand and Pricing build predictive models to forecast future hotel demand. The objective is analyzing historical data related to hotel bookings, market trends, seasonality, and other relevant factors that can predict future demand for hotel rooms.
3. Identifying Influential Factors on ADR investigate the factors that have the greatest impact on the average daily rate (ADR) in the hotel industry.

1.3–Objectives

This project aims to harness the power of data analytics to provide hotels with the tools they need to thrive in a competitive market. By focusing on customer segmentation, demand forecasting, and identifying influential factors on ADR, the project seeks to deliver actionable insights that enhance revenue management and customer satisfaction.

i. Customer Segmentation and Personalization Market segmentation is a crucial strategy for hotels to better understand their target audience and optimize their offerings. By categorizing guests based on various criteria, hotels can gain valuable insights that inform their marketing, pricing, and customer experience strategies.

The key aspects of this objective include:

- Customer Segmentation:
  - o Analyze guest data, such as demographics, booking channels, behavioral patterns, preferences, and past interactions.
  - o Identify distinct customer segments based on common characteristics and behaviors.
  - o Determine which guest segments are the most valuable in terms of revenue generation and loyalty

ii. Forecasting Demand and Pricing Accurate demand forecasting and dynamic pricing strategies are critical for hotels to maximize revenue and optimize their financial performance. By leveraging predictive analytics and integrating customer/operational data, hotels can make informed decisions to adjust rates, inventory, and marketing efforts in response to market conditions.

Key aspects of this objective include:

- Demand Forecasting:
  - o Analyze historical data on occupancy rates, booking patterns, seasonality, and market trends to identify demand patterns.
  - o Utilize advanced statistical models, machine learning algorithms, and data mining techniques to generate demand predictions.
  - o Continuously monitor and refine the forecasting models to improve their predictive capabilities.

- Pricing Optimization:
  - o Develop dynamic pricing strategies that adjust room rates in real-time based on demand forecasts and market conditions.
  - o Leverage customer segmentation and personalization insights to optimize pricing for different guest profiles.
  - o Continuously monitor the performance of pricing strategies and make adjustments to maximize revenue.

Identifying Influential Factors on ADR

Average Daily Rate (ADR) is a critical metric that measures the average revenue generated per occupied room per day. Accurately predicting and understanding the factors that influence ADR is essential for hotels to make informed decisions, optimize their pricing strategies, and improve financial performance.

Key aspects of this objective include:

iii. Defining ADR:

o ADR encompasses all room rates, including discounted rates, group rates, best available rates, and other price points.

o ADR provides a comprehensive view of the average revenue generated per occupied room, allowing hotels to evaluate their pricing strategies and financial performance.

- Predictive Analytics for ADR:

o Utilize machine learning and advanced statistical models to predict future ADR trends and patterns.

o Identify the key drivers and influential factors that impact ADR, such as:

- ♣ Seasonality and market trends
- ♣ Competitor pricing and market conditions
- ♣ Guest segmentation and booking behaviors
- ♣ Operational factors (occupancy, room availability, etc.)

iv. The primary objective of the cancellation forecasting component of this project is to develop a predictive model that accurately identifies factors influencing booking cancellations. By analyzing historical booking data and customer behavior, the goal is to:

Predict Cancellation Rates: Utilize statistical and machine learning techniques to forecast the likelihood of cancellations for upcoming bookings, allowing hotels to proactively manage resources.

1.4 – Problem statement

The hospitality industry is increasingly reliant on data-driven strategies to optimize operations and enhance customer experiences. However, many hotels face critical challenges in three key areas:

1. **Customer Segmentation**: Hotels often lack effective methods to segment their customers based on behaviors and preferences, leading to missed opportunities for targeted marketing and personalized services.
2. **Demand Forecasting**: Accurate prediction of booking demand through time series forecasting is essential for efficient resource allocation and operational planning. Many hotels struggle with fluctuating occupancy rates that hinder financial performance.
3. **Cancellation Management**: High rates of booking cancellations disrupt operational stability and result in significant revenue loss. Understanding the patterns and factors contributing to cancellations is crucial for developing effective mitigation strategies.

To address these challenges, this project will leverage a comprehensive dataset containing 140 variables to implement advanced analytics tools, including customer segmentation models, time series forecasting techniques, and cancellation prediction

CHAPTER 2

Data Exploration

2.1 Data Description

The success of the customer segmentation, demand forecasting, and ADR analysis in this project heavily relies on the quality and richness of the hotel data available. The size and complexity of the data factors play a crucial role in determining the insights that can be extracted and the accuracy of the models developed. this project, the hotel data is derived from two hotels in Portugal, spanning three years: 2018, 2019, and 2020. The dataset includes:

- 2018: 21,997 records
- 2019: 79,265 records
- 2020: 40,688 records

The dataset encompasses a wide range of features capturing different aspects of hotel operations, customer behavior, and market dynamics. These data factors can be categorized into several broad categories such as:

1. Booking details
   - hotel: The type of hotel, either "City Hotel" or "Resort Hotel."
   - booking_changes: Number of changes made to the booking before arrival.
   - market_segment: Market segment designation.
   - distribution_channel: Booking distribution channel.
   - days_in_waiting_list: Number of days the booking was on a waiting list before it
     - was confirmed or canceled.
     - agent: ID of the travel agency.
     - company: ID of the company.
2. Guest information and history
   - customer_type: Type of booking (transient, contract, group, etc.).
   - adults, children, babies: Number of guests categorized by age groups
     - o adults = Number of adults
     - o children = Number of children
     - o babies = Number infants
   - country: Customer's conuntry of origin.
   - is_repeated_guest: Binary value indicating whether the guest is a repeated guest or not.
   - previous_cancellations: Number of previous booking cancellations by a guest.
   - previous_bookings_not_canceled: Number of previous bookings not cancelled by a guest.
3. Reservation timing
   - arrival_date_year: The year of the arrival date
   - arrival_date_month: The month of the arrival date
   - arrival_date_week_number: Week number of arrival date.

- arrival_date_day_of_month: The day of the month of the arrival date
- lead_time: Number of days between booking date and arrival date.
- stays_in_weekend_nights: Number of weekend nights stayed or booked to stay at the hotel
- stays_in_week_nights: Number of week nights stayed or booked to stay at the hotel

4. Room and service information
   - meal: Type(s) food option(s) included in booking package.
   - reserved_room_type : Type of room was originally reserved
   - for each booking.
   - assigned_room_type : type of room was finally assigned for each
   - booking.
   - required_car_parking_spaces: Number of car parking spaces required.
   - total_of_special_requests: Number of special requests made.

5. Financial and Pricing
   - deposit_type: Type of deposit made.
   - adr: Average daily rate.

6. Reservation Status
   - is_canceled: Binary value indicating whether the booking was cancelled or not.
   - reservation_status: Reservation last status.
   - reservation_status_date: Date of the last status.

2.2 Data Cleaning

After collecting the hotel data from the years 2018, 2019, and 2020, the next crucial step is to clean the data. This ensures the dataset is ready for analysis by handling any missing or inconsistent values. In this project, four key factors contain null values, and the following steps were taken to address them:

Handling Missing Values

- Children:
  - o Most customers did not bring any children, making zero the most common (mode)
  - o Action: Replace all null values in the "children" column with zero (0).
- Country
  - o The majority of guests are from Portugal, with "PRT" being the most frequent value
  - o Action: Replace all null values in the "country" column with "PRT".
- Agent:
  - o Some bookings do not have an agent ID, which should be represented by zero to indicate the absence
  - o Action: Replace all null values in the "agent" column with zero (0).
- Company:
  - O Similar to the agent ID, bookings without a company ID should be indicated by zero.
  - Action: Replace all null values in the "company" column with zero (0).

Ensuring that missing values are appropriately handled allows for more accurate insights and predictions.

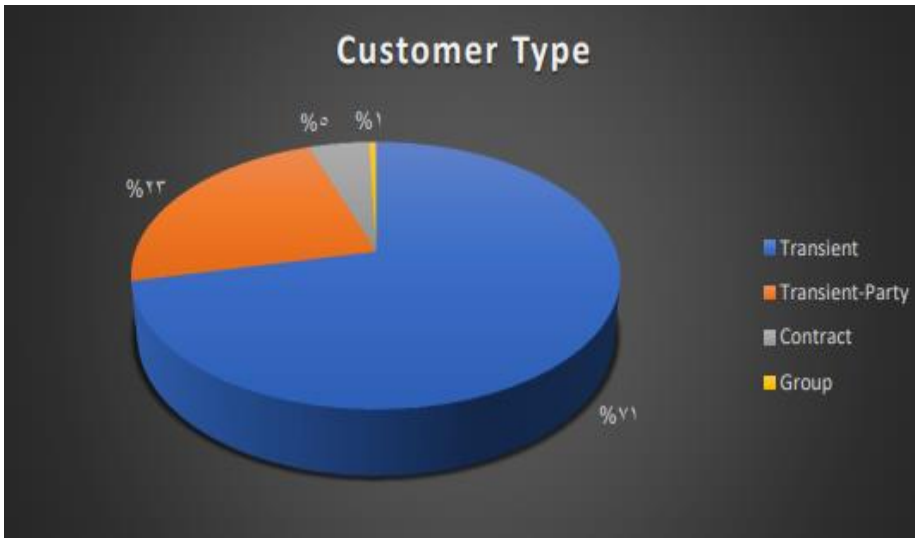
iii. Data analysis and Visualization

1. Customer Segmentation and Personalization

2.3-Exploratory data analysis

Customer Segmentation and Personalization

Customer Type

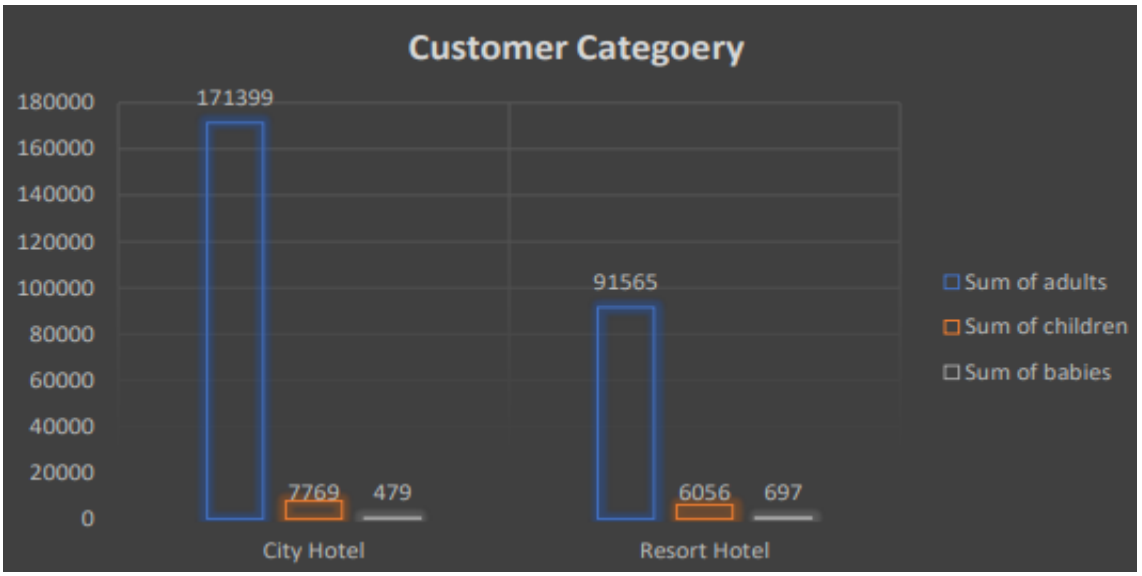


distribution of different customer types (e.g., transient, contract, group).

Most customers are transient who are usually walk-in guests, last minute or bookers or simply people that require a very short term stay in the hotel and that represent about 71 % of customer and And 21% of customer is of transient-party type.

By Understanding customer types helps in tailoring marketing strategies and service offerings to meet the needs of different customer segments.

Customer Categories



distribution of adults, children, and babies among guests at each hotel. o it is seen that most number of customer are containing 2 adults (probably newlyweds, or retirees), or 1 adult (probably single or a traveler).

Most customer did not bring any child or baby for the hotel booking.

City Hotel Primarily adult guests, with fewer children and babies However , Resort Hotel More family-oriented, with a higher proportion of children and babies.

This segmentation helps in customizing services and amenities to suit the demographic profile of the guests at each hotel.

Distribution Channels

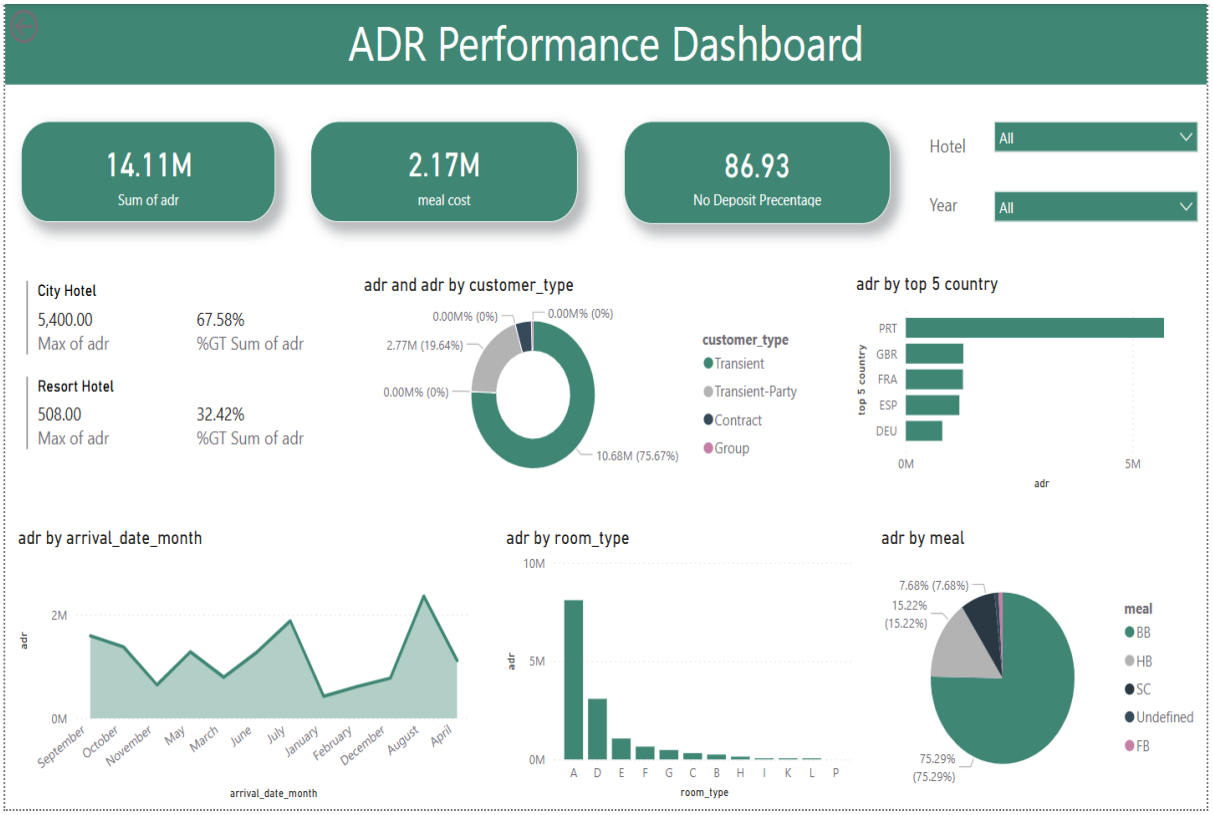
Row Labels	Count of distribution_channel
TA/TO	116042
Direct	17534
Corporate	8167
GDS	194
Undefined	10
Grand Total	**141947**

the distribution of bookings across different channels (e.g., direct, OTA, travel agents).

About 82% from the hotel reservation come from Tour Agents, and Tour Operators, followed by Direct bookings by 12% and Corporate booking by 6%.

By analyzing distribution channels, hotels can optimize their sales strategies and allocate resources to the most effective channels.

2.4-Average Daily Rate Dashboard



1. Total ADR Performance: $14.11M

- The total ADR revenue over three years is **$14.11M**, which serves as the key performance metric for hotel revenue. This figure represents the total income generated per available room.

Key Insights:

- The high ADR value indicates that the hotels have maintained competitive pricing and strong demand.
- However, to sustain revenue growth, strategies should focus on **increasing occupancy rates** and **maximizing high-value bookings**

Recommendations:

- Implement **personalized marketing** to attract high-value guests and repeat customers.
- Use **dynamic pricing** during high-demand seasons to maximize ADR.
- Offer **long-stay discounts** to boost occupancy and maintain revenue consistency.

2. Meal Cost: $2.17M (Impact on Profitability)

Key Insights:

- The meal cost should be evaluated against the revenue generated by meal plans (BB, HB, FB, SC).
- If meal costs are too high relative to revenue from food services, profit margins may suffer.
- The most preferred meal plan is **BB (Bed & Breakfast), covering 75.29%** of the market.

Recommendations:

- **Optimize meal pricing** by aligning it with customer preferences.
- Introduce **more profitable meal plans**, such as premium dining experiences, to increase revenue.
- Reduce **food wastage and optimize supply chain costs** to maintain profitability.

3. No-Deposit Percentage: 86.93% with a 26% Cancellation Rate

- The **86.93% no-deposit rate**, combined with a **26% cancellation rate**, suggests a high risk of lost revenue. Customers booking without deposits may cancel more freely, directly affecting ADR and occupancy rates.

Key Insights:

- **26% of bookings get canceled**, meaning a quarter of potential revenue is lost.
- The **high no-deposit percentage encourages flexible bookings**, but this comes at the cost of **last-minute cancellations**, disrupting revenue forecasts.

Impact on Revenue:

- Assuming canceled bookings were part of the potential ADR, this could mean an opportunity loss of up to **$3.67M** (26% of $14.11M).
- The actual ADR could have been significantly higher if cancellation risks wxcere mitigated

Recommendations:

- **Reduce no-deposit bookings:** Introduce **partial or full deposit requirements** for high-demand dates.
- Offer **flexible but incentivized prepayment options** (e.g., "Pay now, save 10%")
- **Implement stricter cancellation policies:** For example, **tiered refunds** (100% refund 30 days out, 50% refund 14 days out, no refund within 7 days).

4. ADR Trend by Month

Key insight :

- **ADR peaks in December (M12: $7.76M)**, significantly higher than any other month.
- **July (M7: $1.88M) and August (M8: $2.36M) show high ADR**, indicating strong summer demand.
- **Lowest ADR months are January (M1: $421K) and November (M11: $643K)**, reflecting potential low-season demand.

Recommendations:

- **Peak Season Strategy (July, August, December):**
  - Implement **premium pricing** due to high demand.
  - **Offer premium services and upselling opportunities** to maximize revenue.
- **Low Season Strategy (January, November):**
  - Use **discounted packages and promotions** to attract bookings.
  - Focus on **corporate clients and long-stay incentives** to stabilize revenue.
  - Introduce **seasonal events or partnerships with travel agencies** to boost demand.

5. Customer Type Impact on ADR

Customer Type	ADR Contribution ($)	% of Total ADR
Transient	10.68M	75.67%
Transient-Party	2.77M	19.64%
Contract	598K	4.23%
Group	63K	0.45%

Key insights :

- **Transient guests dominate (75.67%)**, meaning most revenue comes from short-term, individual bookings.
- **Transient-party (19.64%) is also significant**, likely consisting of family/group travelers.

- **Contract and group bookings contribute minimally** (below 5%).

Recommendations:

- **Increase loyalty programs and direct booking discounts** for transient travelers.
- **Target group and corporate bookings** to diversify revenue.
- Implement **customized marketing for transient-party guests** (e.g., family packages).

6. Top 5 Countries Driving ADR

Country	ADR Contribution ($)
PRT (Portugal)	**569M**
GBR (UK)	1.27M
FRA (France)	1.26M
ESP (Spain)	1.18M
DEU (Germany)	804K

Key insights:

- **Portugal (PRT) dominates ADR ($569M)**, making it the most valuable market.
- **The UK (GBR), France (FRA), and Spain (ESP) have similar ADR contributions (~$1.2M each).**
- **Germany (DEU) still contributes significantly ($804K), but below the top four.**

Recommendations

Strengthen partnerships and targeted marketing campaigns in Portugal, as it leads in revenue.

Personalized promotions and loyalty programs tailored to Portuguese travelers.

Offer region-specific deals for UK, France, and Spain to maintain their strong contributions.

Leverage travel trends—analyze why Portugal is leading and replicate success in other markets.

7. Room Type Impact on ADR

Room Type	ADR Contribution ($)
Room A	9.27M
Room D	2.54M
Room E	927K
Room F	539K
Room G	409K
Others	Below 200K

Key insights:

- **Room A dominates ADR ($9.27M, ~66% of total ADR)**.
- **Room D is the second-highest performer ($2.54M, ~18%)**.
- **Rooms L and P have almost no contribution**.

Recommendations:

- **Optimize pricing for Room A and D** (high-value demand).
- **Evaluate and repurpose low-performing rooms (L & P)**—consider conversion or special promotions.
- **Upsell premium rooms (E, F, G) with value-added services**.

Chapter 3
Customer Segmentation

3.1-Methodology

3.1.1-Preprocessing

Sample selection: In this step, I took a reduced sample of the whole dataset for computation feasibility purposes. There are ~80,000 observations, making it very time-consuming to run different tests with different clustering algorithms. Therefore, our approach will select 20% of the rows in the dataset, stratifying them by the variable (Arrival_Date_Month) to get the most representative sample possible. Similarly, there are 37 variables in the original dataset, some of which are not relevant for the customer segmentation analysis. Those irrelevant variables are related to hotel processes rather than customer choices.

3.1.2Categorical to cardinal transformation

In present section, we will encode all categories with numeric variables to prepare the dataset for the scaling step. We will assign a new variable for each category but one to avoid collinearity. We can see that we end up with 91 variables in a very sparse dataset.

3.1.3-Feature scaling

Finally, I unified and recentered the distribution of the variables by scaling the whole dataset to a mean of 0 and unit variance

3.2-Modeling and Results

3.2.1-Dimensionality Reduction

This step entailed reducing features to allow faster computation without losing much information from those features eliminated. I used the Principal Component Analysis algorithm (PCA) to find synthetic attributes representing the most extensive variability in the data and contain the most valuable information

3.2.2-Modeling with K-Means Clustering

K-Means Clustering was the first algorithm used to classify our observations. Initially, I defined the number of desired clusters I was later going to use as input to the model. I carried out three different methods to figure out a priori this number:

- The Elbow Method
- The Silhouette Coefficient Method
- Clustering observation with t-SNE

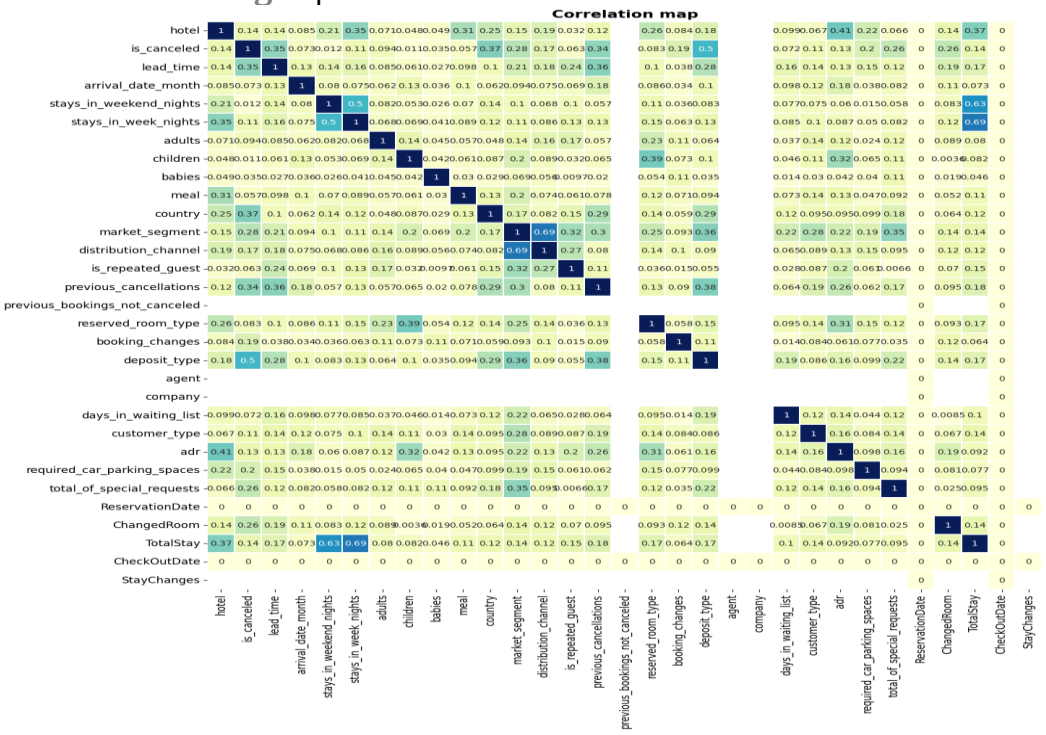
3.2.3-Modeling with alternative algorithms

In this step, I used other alternative algorithms to classify our samples:

- Agglomerative Hierarchical Clustering
- DBSCAN Clustering
- Mean-Shift Clustering
- Expectation-Maximization Clustering

3.3-Visualize Results

I plotted the dataset in 2D using PCA and labeled each sample with the results from the best performing model: K-Means clustering.



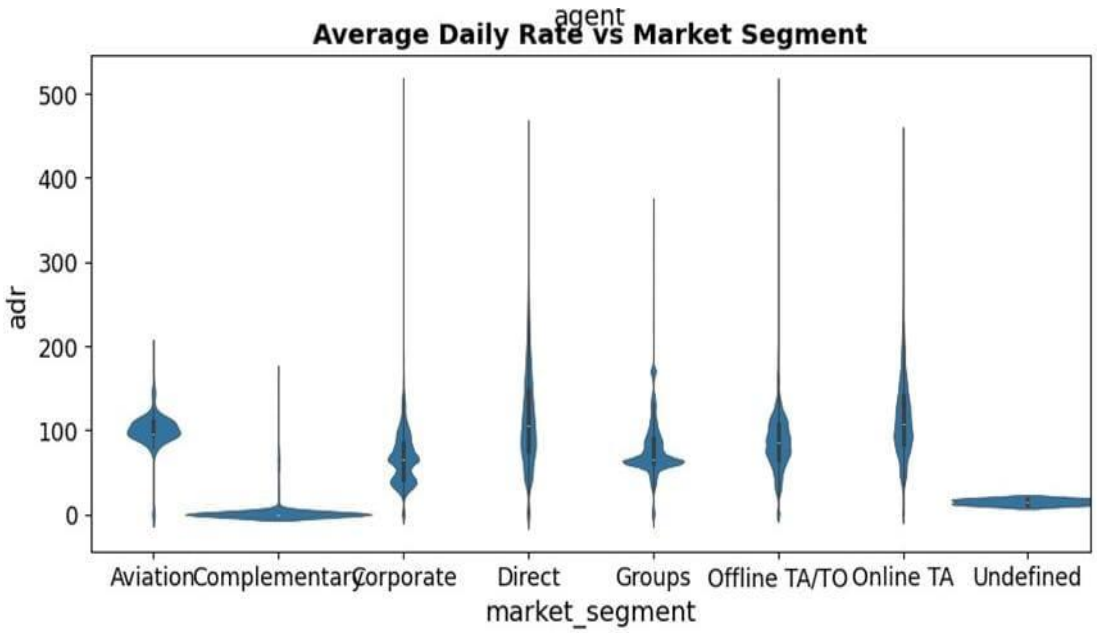
In graph 1, we can observe that in general, there is no association among the feature in the dataset, except for some exceptions:

Is_Repeated_Guests and Previous_Bookings_Not_Cancelled appears correlated with a 0.8 Cramer's V ratio

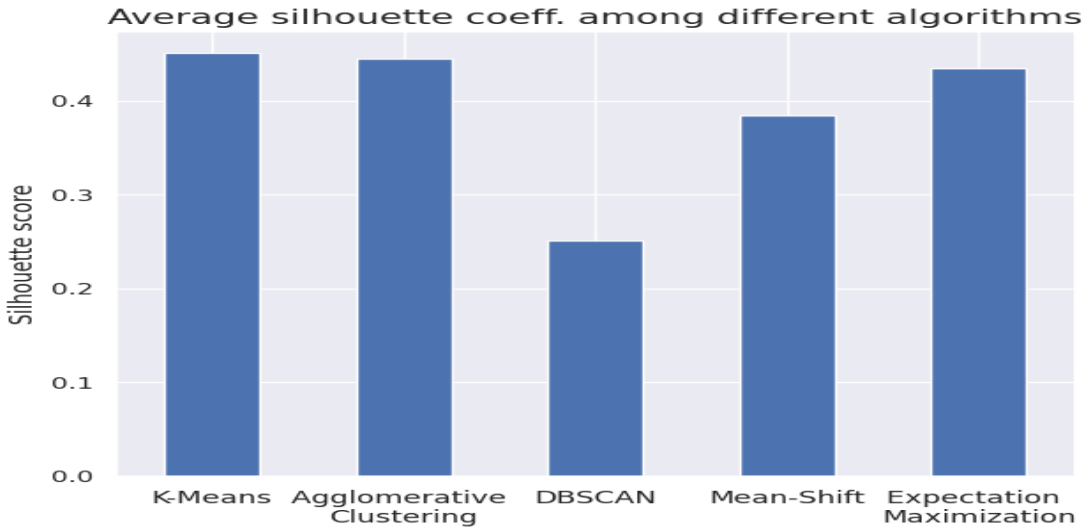
Distribution_Channel and Market_Segment are also associated with a 0.75 ratio, making sense since both are related to sales channels. Also, these features are moderately correlated with

Reservation_Status and Is_Cancelled are completely correlated, and the information they contain is wholly duplicated.

• Finally, another interesting association is Country and Is_Cancelled, meaning that guests from certain countries are more likely to cancel their reservation.



Regarding Market Segment, Corporate and Groups enjoyed the lower rates with roughly EUR 70 on average, contrasting with Online and Direct reservations which had EUR 100 rates on average



- we need at least four for the segmentation to be useful and accurate. K-Means is the best algorithm with a 0.48 average silhouette score. It is worth noting that DBSCAN performed poorly with this dataset, which is expected since our dataset is very sparse, and this algorithm works better for complex spatial structures but tighter datasets. Table 1 below from an article of the University of Berkeley3 shows that all algorithms except DBSCAN found some group structure in the data, showing that there are some customer segments with similar behaviors and preferences. However, the structure is not extraordinarily strong, and it is necessary to visualize it further to understand

The present study has shown an underlying segment structure in the data, and it is possible to establish four well defined customer groups with similar characteristics. However, not all these groups have the same consistency and cohesion. It can be seen in the scatter plot

 Segments 2 and 3 are closer to each other, meaning their attributes are more homogeneous. These segments form well-defined clusters with minimal overlap.

 Segment 1 is a dense and compact cluster, suggesting strong cohesion within the group.

 Segment 0 is more dispersed, indicating that the data points within this group are more spread out and less structured, which may suggest a weaker cluster formation.

The proximity of segments 2 and 3 suggests a possible similarity in characteristics, whereas the dispersion in segment 0 might indicate that it does not form a strong, cohesive group. This information is crucial in customer segmentation

3.4Customer Segments Profiling

- **Cluster 0:**
  - Profile: Long lead time bookings for City Hotel , mostly domestic customers booking through groups, preferring BB meal plan, and often canceling their reservations.
- **Cluster 1:**
  - Profile: Short lead time bookings for City Hotel, mostly European customers booking through online travel agents, preferring BB meal plan, staying over the weekend, and completing their stay.
- **Cluster 2:**
  - Profile: Short lead time bookings for City Hotel, mostly domestic customers booking through offline travel agents, preferring BB meal plan, and completing their stay.
- **Cluster 3:**
  - Profile: Short lead time bookings for Resort Hotel, mostly domestic customers booking directly, preferring BB meal plan, and completing their stay.

Chapter 4
Time Series Forecasting

4.1 Introduction to Time Series Forecasting

Time series forecasting is a statistical technique used to analyze time-ordered data points to identify patterns and predict future value0.0.

This method is highly valuable across various fields, including finance, economics, and notably, the hospitality industry. For city and resort hotels, accurate forecasts of occupancy rates, booking patterns, and seasonal demand fluctuations are crucial for strategic planning and revenue management.

4.2 Importance In hospitality

In the hospitality sector, effective time series forecasting enables hotel managers and revenue professionals to:

Optimize Pricing: Understanding seasonal trends and fluctuations in demand allows hotels to adjust pricing strategies dynamically. This maximizes revenue during peak times while maintaining occupancy in off-peak periods.

Enhance Customer Experience: Forecasting occupancy levels aids in staffing and service preparedness, ensuring that guest experiences are exceptional during busy seasons.

Strategic Planning: Hotels can allocate resources effectively, plan marketing campaigns, and make informed decisions about new investments or renovations based on expected demand.

4.3 Key Concepts

Several foundational concepts are essential for grasping the intricacies of time series forecasting:

Trends: These are long-term movement directions in data. For instance, an overall upward trend in hotel bookings could be indicative of increasing popularity in a destination.

Seasonality: This refers to predictable fluctuations that occur at regular intervals. In hospitality, certain seasons, such as summer vacations or holiday periods, show increased demand, impacting booking behavior.

Cyclic Patterns: Unlike seasonality, these patterns occur at irregular intervals influenced by economic factors, such as a downturn or boom in tourism.

This document aims to provide a comprehensive analysis of these concepts, coupled with detailed historical booking patterns and methodologies that can inform decision-making in revenue management and operational strategies for hotels.

4.4 Forecasting Techniques

4.4.1 ARIMA

ARIMA (AutoRegressive Integrated Moving Average)

ARIMA stands out as one of the most widely used approaches in time series analysis due to its flexibility and effectiveness in modeling time-ordered data.

Components:

- **AR (AutoRegressive)**: Relies on the relationship between an observation and a specified number of lagged observations. This is crucial in capturing trends in hotel booking data.
- **I (Integrated)**: Represents the differencing of raw observations to allow for trend stabilization.
- **MA (Moving Average)**: Involves the relationship between an observation and a residual error from a moving average model.

Applicability:

- ARIMA is particularly effective for predicting future booking patterns based on historicaldata, allowing hotel managers to adapt strategies based on identified trends. It can handle

  seasonality when extended to Seasonal ARIMA models, making it suitable for cyclical booking periods, such as holidays and events.

4.4.2 Exponential Smoothing

Exponential smoothing is another popular technique, celebrated for its simplicity and effectiveness in forecasting time series data.

Components:

- Uses weighted averages of past observations where the weights decrease exponentially as observations get older.
- Various models exist, from simple exponential smoothing (suitable for data without trend or seasonality) to more advanced forms, such as Holt's Linear and Holt-Winters Seasonal, which account for trends and seasonality respectively.

Applicability:

- This method is particularly useful for hotels looking to forecast short-term booking demands, as it quickly adjusts to changes in patterns. For example, during unexpected spikes in demand due to local events, exponential smoothing can provide timely updates to forecasts.

4.4.2-Prophet

As the demand for accurate time series forecasting grows, Prophet has emerged as a powerful tool designed specifically for this purpose. Developed by Facebook, Prophet is optimized for forecasting time series data that exhibit seasonal patterns and trends, making it particularly useful in business contexts.

Components:

- **Additive and Multiplicative Models**: Prophet allows users to choose between additive and multiplicative seasonality, accommodating various types of seasonal effects in the data. This flexibility is crucial for accurately modeling trends that can change over time.
- **Automatic Holiday Effects**: The tool can incorporate effects of holidays and special events, which can significantly impact sales and bookings. Users can easily add custom holidays to enhance forecast accuracy.

4.5-Model Selection and Evaluation

Selecting the appropriate forecasting model is a critical step in deriving accurate predictions that inform revenue management strategies in the hospitality sector. The model selection process often involves evaluating several factors including model suitability to data characteristics, interpretability, and performance metrics.

Performance Metrics for Evaluation

Once potential models are identified, evaluating their performance is crucial. Common metrics include:

Mean Absolute Error (MAE): This metric calculates the average absolute errors between predicted and actual values. It provides a clear indication of forecast accuracy without penalizing large discrepancies excessively.

[MAE = \frac{1}{n} \sum_{i=1}^{n} |y_t - \hat{y}_t|]

where (y_t) represents the actual value and (\hat{y}_t) is the forecasted value.

Root Mean Squared Error (RMSE): Unlike MAE, RMSE squares the errors before averaging, which means it penalizes larger errors more heavily. This can be useful in recognizing models that make significant forecast missteps.

[RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_t - \hat{y}_t)^2}]

In the context of hotel bookings, consider a scenario where a model predicts occupancy rates for 29|page

an upcoming holiday season. If historical data indicates a consistent spike in bookings during this period, a well-performing model should yield a lower MAE and RMSE, based on past successful predictions.

Example Application

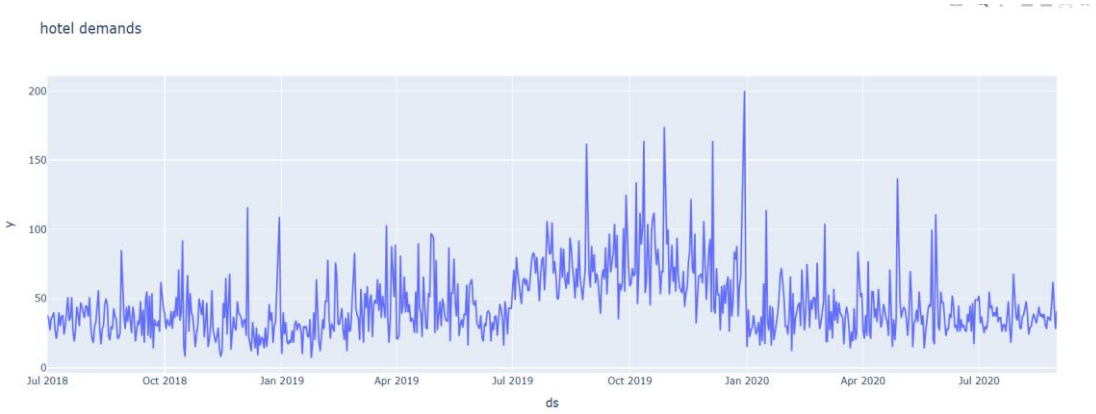
For instance, if a hotel utilizes ARIMA for forecasting quarterly bookings, the MAE and RMSE can be calculated using historical data from the previous years. If the MAE is consistently low, this indicates the model is capable of accurately forecasting expected occupancy, allowing management to adjust pricing strategies accordingly. Conversely, if the RMSE shows significant deviation during peak times, it may be necessary to switch to a more adaptive model like exponential smoothing or even a machine learning approach to capture unexpected fluctuations.

Utilizing these performance metrics helps hotel managers refine their forecasting methods, ensuring they not only derive accurate predictions but also enhance their overall operational strategies.

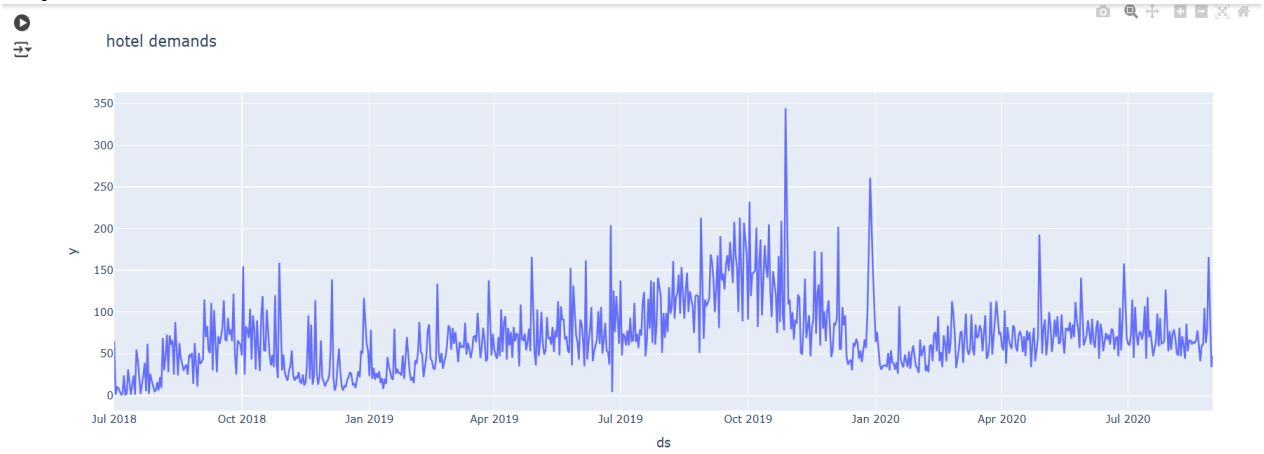
4.6-Application of Models

To streamline the analysis process, we specifically focus on the booking demand for the resort and city hotel and create a time series dataset where booking numbers are grouped according to the arrival date

Resort Hotel:



City hotel:



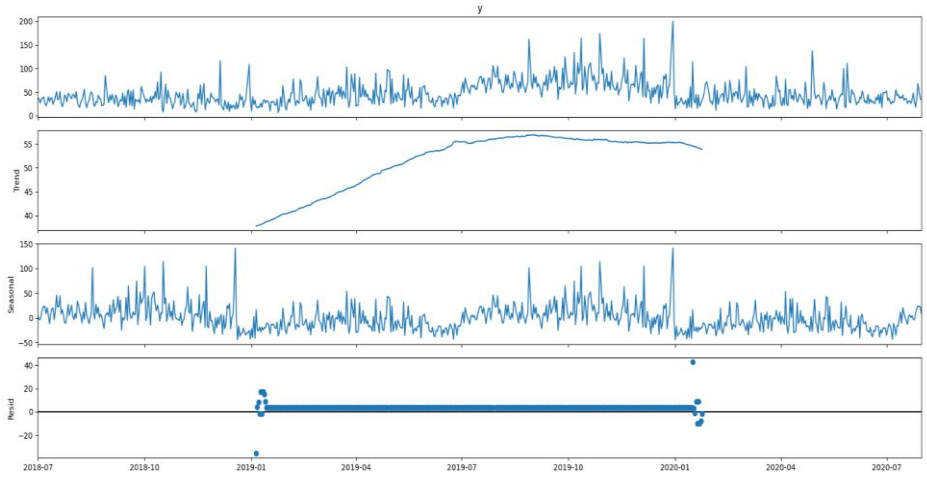
To assess the effectiveness of each forecasting method, we allocate a training period of 25 months, followed by a test period of 1 month specifically in August 2020.

1-ARIMA Model

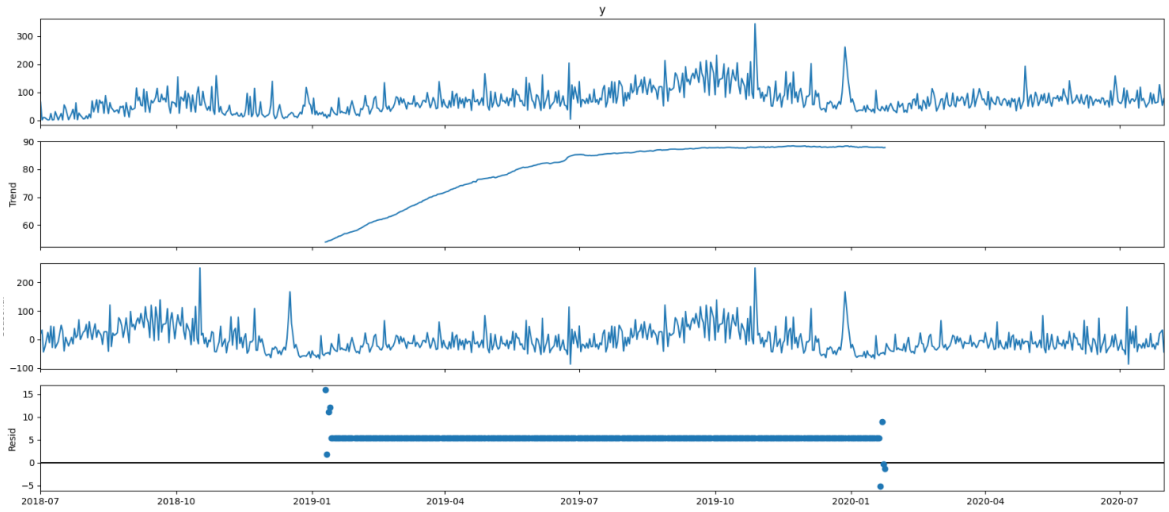
Initially, we employ a technique known as time-series decomposition to break down our time series into three distinct elements: trend, seasonality, and noise. However,

upon analysis, we do not observe a noticeable trend but we see weak seasonality within the data.

Resort Hotel:



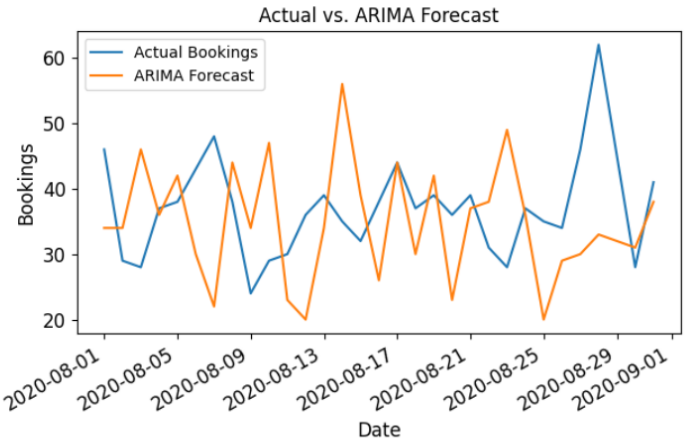
City Hotel:



Forecasted VS Actual

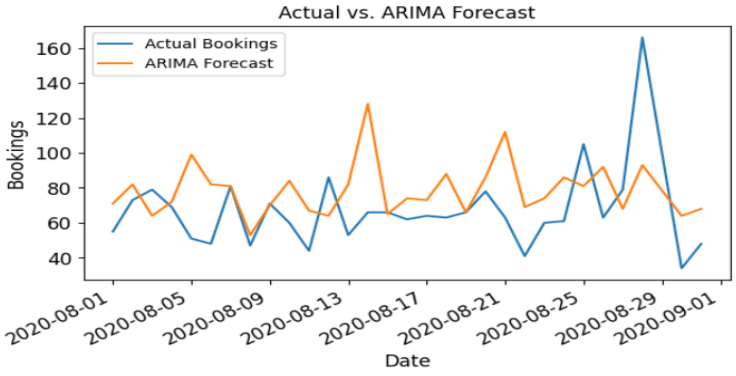
Resort Hotel:

```
Head of Merged Data (test_output):
           ds   y  ARIMA_forecast
0 2020-08-01  46              34
1 2020-08-02  29              34
2 2020-08-03  28              46
3 2020-08-04  37              36
4 2020-08-05  38              42
```



City Hotel:

```
Head of Merged Data (test_output):
           ds   y  ARIMA_forecast
0 2020-08-01  55              71
1 2020-08-02  73              82
2 2020-08-03  79              64
3 2020-08-04  69              72
4 2020-08-05  51              99
```



Resort Hotel:

ARIMA MAE: 10

ARIMA RMSE: 12

City Hotel:

ARIMA MAE: 21

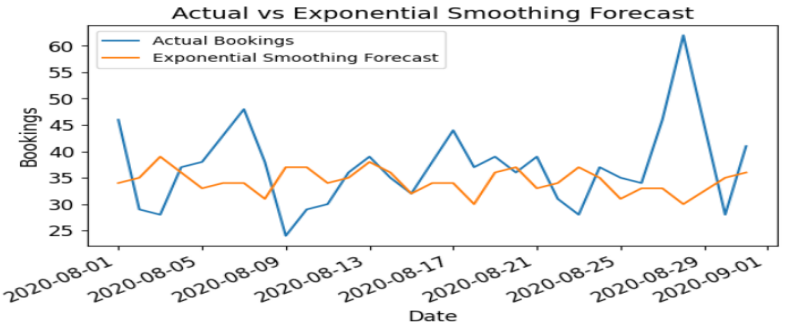
ARIMA RMSE: 27

- **Exponential Smoothing Model**

  Forecasted VS Actual

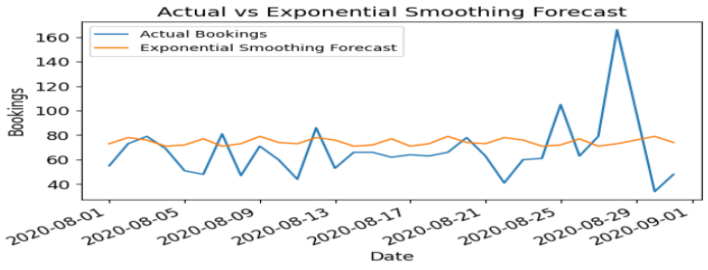
  Resort Hotel:

```
2020-08-01  46                          34
2020-08-02  29                          35
2020-08-03  28                          39
2020-08-04  37                          36
2020-08-05  38                          33
```



City Hotel:

```
Exponential Smoothing Forecasts:
                y  Exponential_Smoothing_Forecast
ds
2020-08-01  55                          73
2020-08-02  73                          78
2020-08-03  79                          76
2020-08-04  69                          71
2020-08-05  51                          72
```



Resort Hotel:

Exponential Smoothing MAE: 6

Exponential Smoothing RMSE: 8

City Hotel:

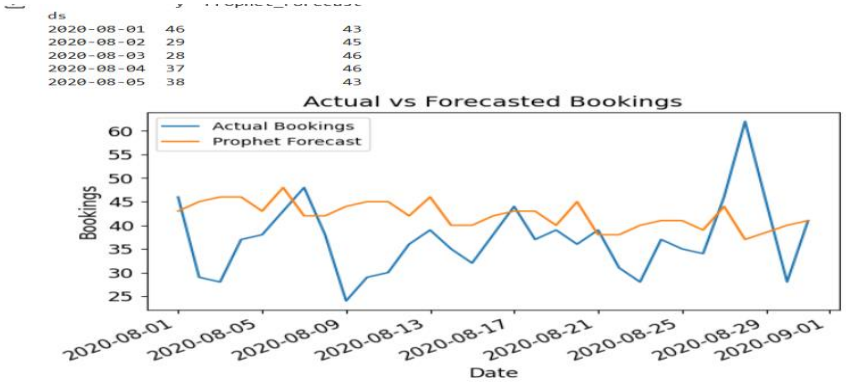
Exponential Smoothing MAE: 18

Exponential Smoothing RMSE: 25

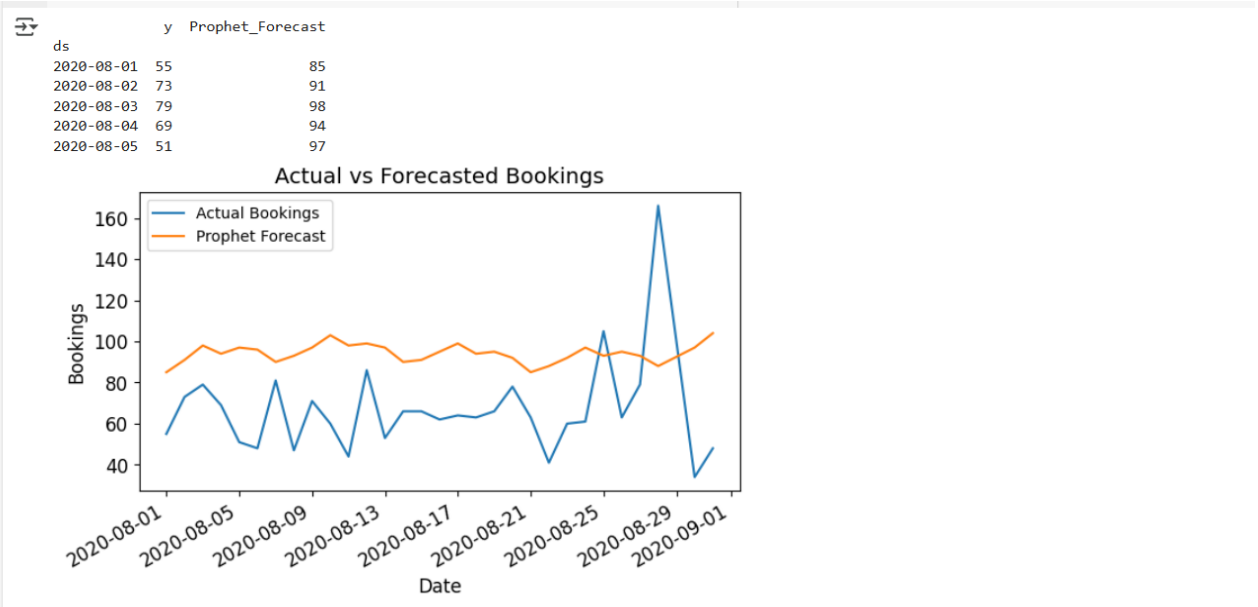
3- **Prophet Model**

Forecasted VS Actual

Resort Hotel:



City Hotel:



35|page

Resort Hotel:

Prophet MAE: 7

Prophet RMSE: 10

4.7 Performance analysis

Resort Hotel

Model	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
ARIMA	10	12
Exponential Smoothing	6	8
Prophet	7	10

City Hotel

Model	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
ARIMA	21	27
Exponential Smoothing	18	25
Prophet	33	37

In analyzing the forecasting performance between the two hotels, it is evident that the city hotel exhibits higher Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) when compared to the resort hotel. This discrepancy in accuracy can be attributed primarily to the

smaller size of the dataset associated with the resort hotel.

Chapter 5
Cancelation Forecasting

5.1 Introduction to Cancelation Forecasting

Booking cancellations represent a significant aspect of hotel management, particularly in the competitive landscape of resort and city hotels. When a guest cancels a reservation, it not only disrupts the hotel's operational planning but can also lead to substantial financial losses. Understanding cancellation patterns is crucial, as they can shed light on the underlying reasons behind guest decisions and help hotel management develop strategies to mitigate these occurrences.

5.2 Importance of Cancelation Patterns

Cancellations can have a profound impact on hotel operations, leading to underutilized resources and increased operational costs. For instance, when a booking is canceled, hotels may face challenges such as:

- Loss of Revenue: Unfilled rooms equate to lost income, which can dramatically affect profitability, especially during peak seasons.
- Operational Disruptions: Sudden cancellations can disrupt staffing levels and resource allocation, complicating operations and negatively influencing service quality.
- Increased Marketing Costs: To fill vacant rooms, hotels may incur additional marketing expenses, affecting overall financial performance.

Incorporating predictive analytics into cancellation management provides hotels with a framework to navigate these challenges, allowing for data- driven strategies that enhance both financial and operational outcomes. By harnessing advanced machine learning techniques, hoteliers can gain insights into customer behavior, ultimately leading to better decision-making and improved resilience against market fluctuations.

5.3 Data Collection and Preprocessing

In order to analyze hotel cancellations effectively, a comprehensive approach to data collection is essential. The datasets used for this analysis comprise several key components:

- **Booking Patterns**: This includes information regarding reservation dates and length of stay. It facilitates an understanding of trends and behaviors specific to customer bookings.
- **Customer Demographics**: Data on customer characteristics such as market segment, nationality, customer type and booking channels (online vs. offline) aid in segmenting the customer base and identifying differences in cancellation behavior.
- **Historical Cancellation Data**: Past records of cancellations, provide valuable insights into the circumstances that lead to cancellations, allowing for predictive model training.

PREPROCESSING STEPS

The preprocessing phase is critical in preparing the data for analysis, ensuring that the information is clean, reliable, and ready for modeling. Key steps include:

Handling Missing Values:

- Removal of Duplicates: Any duplicate records in the dataset were removed to maintain the integrity of the analysis and avoid skewed results.

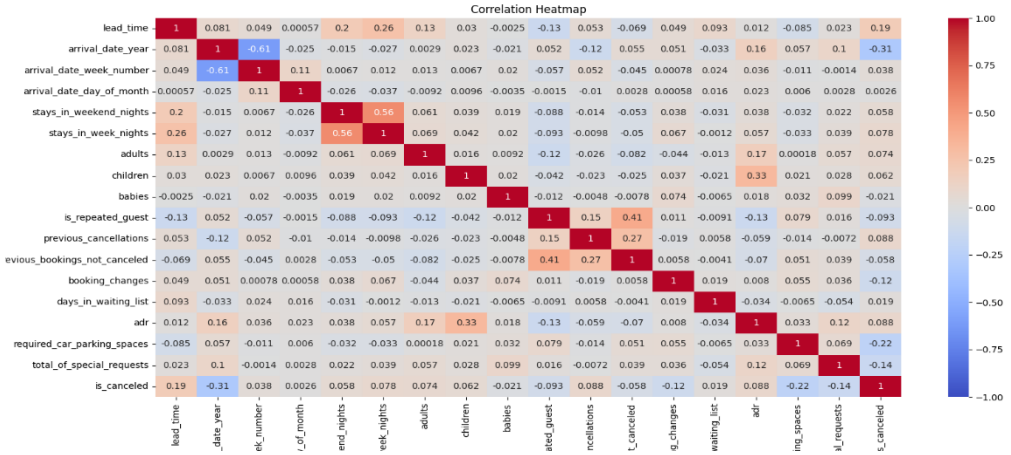
- Missing or null values can hinder the accuracy of predictive models.

Removal: Discarding records where crucial data points are missing, depending on the extent of missingness.

Feature Selection:

- Selecting the most relevant features is essential to improve model performance and reduce overfitting. Techniques include:
  - **Correlation Analysis**: we created a correlation matrix to evaluate the relationships between various features in our dataset and the target variable, is canceled, which indicates whether a booking was canceled. The correlation matrix is a powerful tool for understanding how different factors are interconnected and their potential influence on cancellation rates.

- Heatmap Visualization: We visualized the correlation matrix using a heatmap, which provides a clear representation of the strength and direction of the correlations.



Key Correlation Insights

From the correlation matrix, we observed the following key relationships with the target variable is_canceled:

1. **Lead Time (0.191)**: There is a moderate positive correlation, suggesting that longer lead times are associated with higher cancellation rates. This could imply that guests booking well in advance may have more uncertainties affecting their plans.

2. **Average Daily Rate (ADR) (0.088)**: A weak positive correlation indicates that higher average rates may slightly increase cancellation likelihood, possibly due to guest price sensitivity.

3. **Previous Cancellations (0.088)**: This feature also shows a weak positive correlation with cancellations, suggesting that guests with prior cancellations may be more likely to cancel again.

4. **Booking Changes (-0.117)**: This negative correlation suggests that a higher number of booking changes may be associated with a lower likelihood of cancellation, possibly indicating that more engaged guests are less likely to cancel.

5. **Total of Special Requests (-0.144)**: This indicates that guests making more special requests are less likely to cancel, likely because they have invested more in their booking.

6. **Required Car Parking Spaces (-0.218)**: A notable negative correlation suggests that the need for parking spaces might deter cancellations, as guests requiring this amenity may have more commitment to their stay.

7. **Arrival Date Year (-0.306)**: This strong negative correlation suggests that cancellations are more frequent in certain years, which could reflect broader trends or economic conditions affecting travel.

- **Recursive Feature Elimination:** Iteratively removing the least important features based on model performance metrics.

2. **Normalization:**

- Normalization is a crucial step in preparing data for predictive modeling. It ensures that all features contribute equally to the model's performance, particularly when different features have varying ranges and units. This step enhances model convergence and stability.

- In this project, we applied normalization through two primary techniques:

- **Label Encoding for Categorical Variables**:
  - We used LabelEncoder from the sklearn.preprocessing module to convert categorical columns into numerical format. This transformation enables the algorithm to interpret categorical data effectively.
  - The categorical columns included in this process are:
    - hotel
    - arrival_date_month
    - meal
    - country
    - market_segment
    - distribution_channel
    - reserved_room_type
    - deposit_type
    - customer_type
- **Standard Scaling for Numerical Variables**:
  - We applied StandardScaler to scale numerical features to a standard normal distribution with a mean of 0 and a standard deviation of 1. This ensures that all numerical features are on the same scale, which is essential for algorithms sensitive to feature scaling.
  - The numerical columns that underwent standardization include:
    - lead_time
    - arrival_date_year
    - arrival_date_day_of_month
    - stays_in_weekend_nights
    - stays_in_week_nights
    - adults
    - children
    - is_repeated_guest
    - previous_cancellations
    - required_car_parking_spaces

3. **Addressing Class Imbalance with SMOTEN**

In predictive modeling, particularly in classification tasks, class imbalance can significantly impact the model's performance. In our dataset, the target variable is_canceled exhibits a clear imbalance:

- Not Canceled (0): 73,961 instances
- Canceled (1): 26,775 instances

This disparity can lead to a model that is biased towards the majority class (not canceled), resulting in poor predictive performance for the minority class (canceled).

To address this issue, we employed **SMOTEN** (Synthetic Minority Over-sampling Technique for Nominal data), which is an effective technique for balancing class distributions by generating synthetic samples.

These preprocessing methods are vital in ensuring that the subsequent machine learning analyses yield accurate and reliable cancellation predictions, ultimately contributing to better decision-making in hotel management.

5.4 ML Models

In the context of predicting hotel cancellations, several machine learning models have gained prominence due to their effectiveness and versatility. This section provides an overview of four key models: Support Vector Machine (SVM), Logistic Regression, XGBoost, and Random Forest, along with their principles, strengths, and weaknesses.

SUPPORT VECTOR MACHINE (SVM)

- **General Principles**: SVM is a supervised learning model that seeks to find the optimal hyperplane that separates data points of different classes in a high-dimensional feature space.
- **Strengths**:
  - Effective in high-dimensional spaces, making it suitable for datasets with numerous features.
  - Robust to overfitting, especially in cases where the number of dimensions exceeds the number of samples.
- **Weaknesses**:
  - Computationally intensive and can be less efficient with larger datasets.
  - Selection of the appropriate kernel function can significantly impact performance.

LOGISTIC REGRESSION

General Principles: Logistic Regression is a statistical method used for binary classification, predicting the probability that a given input belongs

to a particular category.

- **Strengths**:
  - Simple to implement and interpret, which is beneficial for stakeholders who are less technical.
  - Provides probabilities that can inform risk assessment and decision-making in hotel management.
- **Weaknesses:**
  - Assumes a linear relationship between features and the log-odds of the outcome, which may not capture complex patterns.
  - Prone to underperforming when dealing with non-linear data distributions.

XGBOOST

General Principles: Extreme Gradient Boosting (XGBoost) is an ensemble learning technique that builds models sequentially, correcting errors made by previous models through boosting.

- **Strengths**:
  - High performance and speed due to parallel processing and cache optimization.
  - Incorporates regularization techniques, reducing overfitting and enhancing generalization on unseen data.
- **Weaknesses**:
  - Complexity in tuning hyperparameters can lead to overfitting if not managed correctly.
  - Requires more computational resources compared to simpler models.

RANDOM FOREST

General Principles: Random Forest is another ensemble learning method that employs multiple decision trees to improve prediction accuracy by averaging their outputs.

- **Strengths**:

- Handles both numerical and categorical variables well, proving versatile across various datasets.
- Reduces overfitting by averaging multiple trees, thereby improving prediction stability.

• **Weaknesses**:

- The model can become less interpretable due to its complexity and the ensemble approach.
- Larger model size makes it slower for predictions compared to simpler models.

By understanding these models and their implications, hotels can make informed decisions regarding implementation strategies to better predict and reduce cancellation rates amongst their clientele.

5.5 Model Training and evaluation

The training and evaluation of models play a pivotal role in ensuring the effectiveness of cancellation predictions in hotel management. This section elaborates on the methodologies involved in training the selected models: Support Vector Machine (SVM), Logistic Regression, XGBoost, and Random Forest. Key processes include dataset splitting, cross-validation, hyperparameter tuning, and the assessment of model performance using specific evaluation metrics.

TRAINING DATASET SPLIT

Before training any machine learning model, it is essential to split the dataset into distinct subsets to prevent overfitting and ensure valid assessments. The typical approach involves:

Training Set: 80% of the data, it is employed to train the model.

Testing Set: The remaining 20% of the data, used exclusively for evaluating model performance post-training.

This division allows the model to learn patterns within the training set while being evaluated on unseen data from the testing set.

EVALUATION METRICS

The effectiveness of each model is assessed using specific evaluation metrics, focusing primarily on **accuracy** for this analysis. Accuracy represents the ratio of correctly predicted instances to the total instances:

- **Support Vector Machine (SVM)**: Achieved an accuracy of 0.779, demonstrating its capability in high-dimensional spaces.
- **Logistic Regression:** Reported an accuracy of 0.739, showcasing simplicity but limitations in complexity capture.
- **XGBoost:** Excelled with an accuracy of 0.818, benefiting from its ensemble learning approach and boosting technique.
- Random Forest: Matched XGBoost's performance with an accuracy of 0.818, illustrating the effectiveness of averaging predictions from multiple trees.

Model	Accuracy
Support Vector Machine	0.779
Logistic Regression	0.739
XGBoost	0.818
Random Forest	0.818

Through rigorous training, cross-validation and careful evaluation, these models demonstrate their predictive capabilities in understanding hotel cancellations. Each model's unique strengths cater to different aspects of the prediction problem, affirming the importance of choosing the right approach for a given dataset.

5.6 Results and Comparison

In analyzing the predictive power of various machine learning models for hotel cancellations, we evaluated their performance based on accuracy,

precision, recall, and the F1 score. The following model results highlight their effectiveness in providing robust predictions:

MODEL PERFORMANCE SUMMARY

Model	Accuracy	Precision(0)	Precision(1)	Recall(0)	Recall(1)	F1 Score(0)	F1 Score(1)
Support Vector Machine (SVM)	0.779	0.82	0.73	0.77	0.78	0.79	0.75
Logistic Regression	0.739	0.75	0.72	0.80	0.67	0.77	0.69
XGBoost	0.818	0.85	0.78	0.82	0.81	0.83	0.79
Random Forest	0.818	0.84	0.76	0.81	0.80	0.82	0.78

ANALYSIS OF MODEL PERFORMANCE

XGBoost and Random Forest

Both XGBoost and Random Forest exhibited superior performance, achieving an accuracy of **0.818.** This can be attributed to their ensemble methodologies:

- **XGBoost** improves upon traditional boosting techniques by incorporating regularization, which helps mitigate overfitting. It uses a sequential approach to correct the errors of prior models, leading to enhanced predictive capabilities.
- **Random Forest** utilizes bagging (bootstrap aggregating), which builds multiple decision trees and averages their outputs, thus reducing variance. This robust model effectively captures complex interactions within the data while maintaining a level of interpretability.

Why They Outperform Others:
The combination of these methods often results in higher precision and recall rates as seen in the metrics provided. The enhanced handling of different feature types and interactions allows these models to generalize better on unseen data.

Support Vector Machine and Logistic Regression

While Support Vector Machine (SVM) performed respectably with an accuracy of 0.779, it fell short compared to XGBoost and Random Forest. This is partly due to SVM's reliance on the choice of kernel function, which can significantly influence performance. Moreover, it may struggle with larger datasets, which are prevalent in hotel cancellations.

Logistic Regression, achieving an accuracy of 0.739, remains a fundamental baseline model. Although straightforward and interpretable, its assumption of linearity limits its ability to capture non-linear relationships in more complex datasets.

CONCLUSION OF MODEL COMPARISONS

In summary, the comparative analysis of these models illustrates that ensemble techniques like XGBoost and Random Forest are more effective in handling complex cancellation prediction tasks. Their ability to adaptively

learn from data variances enhances their overall robustness, making them suitable candidates for hotel cancellation management strategies.

Chapter 6
Conclusion and Future work

6.1 Summary Findings

Customer Segmentation:

Effective segmentation of hotel guests enhances personalized marketing strategies, improving customer satisfaction and loyalty. Four distinct customer segments were identified, each with unique behaviors and preferences.

Demand Forecasting:

Time series forecasting techniques, particularly ARIMA, Exponential Smoothing, and Prophet, were employed to predict booking trends and occupancy rates. These models provided valuable insights for optimizing pricing and resource allocation, contributing to better financial performance.

Cancellation Forecasting:

The project developed predictive models using machine learning techniques (e.g., XGBoost, Random Forest) to analyze cancellation patterns. Key factors influencing cancellations were identified, including lead time and previous cancellations, facilitating proactive resource management.

Operational Efficiency:

The insights gained from customer behavior and market dynamics enable hotels to enhance operational efficiency, reduce costs, and increase profitability. Data-driven decision-making is essential for staying competitive in the evolving hospitality landscape.

Personalization Strategies:

Emphasizing personalized customer experiences through data insights can improve customer retention and satisfaction, ultimately leading to a more loyal customer base.

6.2 Implications for Hotel management

Enhanced Customer Segmentation:

Tailoring marketing strategies and services to specific guest profiles improves satisfaction and loyalty.

Optimized Pricing Strategies:

Utilizing advanced forecasting models allows for dynamic pricing adjustments, maximizing revenue based on demand fluctuations.

Improved Resource Allocation:

Accurate demand forecasting aids in efficient staffing and inventory management, ensuring that hotels meet guest needs while minimizing costs.

6.3 Limitations of the study

Unclear Causality for High Cancellation Rates:

The study does not provide specific reasons for high cancellation rates, making it difficult to identify actionable strategies for mitigation. The lack of qualitative data on guest motivations contributes to this ambiguity.

Insufficient Financial Data on ADR:

The analysis lacks detailed financial information regarding pricing strategies and external market factors that influence the Average Daily Rate (ADR). Without this data, it is challenging to assess the true impact of pricing on hotel performance.

Limited Insight into Customer Behavior:

The absence of comprehensive data on customer preferences and booking habits may hinder a deeper understanding of the factors contributing to cancellations and pricing decisions.

6.4 Future Direction

Focus on Workforce Development:

Examine the role of employee training and development in enhancing service quality and operational effectiveness, recognizing the importance of human capital in the hospitality industry.

Enhanced Customer Insights:

Utilize advanced analytics to gain deeper insights into customer behavior and preferences, enabling more personalized marketing and service offerings.

Policy Impact Analysis:

Assess the impact of regulatory changes and policies on hotel operations and customer behavior, providing data-driven recommendations for compliance and adaptation.

7. References

1. **Kaggle**. (n.d.). *Hotel Booking Demand Dataset.* Retrieved from Kaggle https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

2. **Ridgeway, G.** (2016). *XGBoost: A Scalable Tree Boosting System*. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

3. **Breiman, L.** (2001). *Random Forests. Machine Learning*, 45(1), 5-32. doi: 10.1023/A:1010933404324

4**. Cortes, C., & Vapnik, V.** (1995). *Support-Vector Networks*. Machine Learning, 20(3), 273-297. doi:10.1007/BF00994018

5.**Katz, M.** (2017). *Logistic Regression for Machine Learning: Principles and Practical Applications*. Retrieved from https://machinelearningmastery.com/logistic-regression-for-machine-learning/

6. **Chen, T., & Guestrin, C.** (2016). *XGBoost: A Scalable Tree Boosting System.* arXiv Preprint arXiv:16