TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Overview of the Hospitality Industry in India

The hospitality industry in India is a rapidly growing sector that plays a vital role in the country's economy. It contributes significantly to GDP, employment, and foreign exchange earnings. The industry encompasses a wide range of services, including lodging, food and beverage, travel and tourism, and event planning.

India's hospitality sector has witnessed remarkable growth due to increasing disposable incomes, changing lifestyles, and the rising middle class. Additionally, government initiatives like "Incredible India" and "Dekho Apna Desh" have further fueled the tourism and hospitality market. The sector is also benefiting from international events, improved infrastructure, and enhanced transportation facilities.

Several segments in Indian hospitality industry:

- **Luxury Hotels** – Five-star and premium accommodations offering world-class amenities.
- **Mid-range and Budget Hotels** – Affordable lodging options catering to business and leisure travelers.
- **Homestays and Boutique Hotels** – Personalized accommodations offering a cultural experience.
- **Online Hotel Aggregators** – Platforms like OYO, MakeMyTrip, and Airbnb providing online booking services.

With digital transformation, hotel chains and independent operators are increasingly leveraging technology for customer engagement, revenue management, and operational efficiency.

1.2 Role of Online Hotel Listings

The rise of digitalization has revolutionized the way hotels operate, market themselves, and attract customers. Online hotel listing platforms, such as OYO Rooms, MakeMyTrip, Goibibo, and Booking.com, have changed consumer behavior by offering real-time availability, pricing, discounts, and reviews.

Benefits of Online Hotel Listings:

1. **Increased Visibility** – Hotels listed on online platforms gain more exposure, attracting domestic and international tourists.
2. **Convenience for Customers** – Users can compare hotels based on price, location, amenities, and customer feedback.

3. **Real-time Price Comparisons** – Dynamic pricing enables customers to get competitive rates.
4. **Customer Reviews and Ratings** – User-generated content helps potential guests make informed decisions.
5. **Discounts and Promotions** – Online platforms frequently offer deals and cashback options to attract bookings.

The growing reliance on online listings has compelled hoteliers to enhance their digital presence and adopt customer-friendly pricing and service strategies.

1.3 Purpose and Scope of the Study

The primary objective of this study is to analyze online hotel listings in India using data collected from OYO Rooms. The study seeks to understand various factors influencing hotel pricing, customer ratings, and discount strategies.

Research Objectives:

- To analyze the distribution of hotels across different cities in India.
- To identify price variations and trends in different regions.
- To evaluate the impact of online discounts on customer choices.
- To examine the correlation between hotel ratings and pricing.
- To develop predictive models for hotel price estimation.
- To perform sentiment analysis on customer reviews.

Scope of the Study:

The study covers data collection through web scraping, preprocessing, exploratory data analysis (EDA), visualization, and machine learning applications. By leveraging analytics tools such as Python, Tableau, and Power BI, the study will generate meaningful insights that can help businesses in the hospitality sector make informed decisions.

CHAPTER 2: LITERATURE REVIEW

The literature review provides an in-depth analysis of existing research and trends in online hotel bookings, pricing strategies, and consumer behavior. This chapter explores the factors influencing customer decisions, the role of online reviews and ratings, and the pricing dynamics in the hospitality industry.

2.1 Online Booking Trends and Consumer Behavior

The digital transformation of the hospitality industry has significantly altered consumer behavior. Online booking platforms such as OYO, MakeMyTrip, Airbnb, and Booking.com have streamlined the process of selecting and reserving accommodations. Several key trends have emerged in this space:

Growth of Online Travel Agencies (OTAs)

Online travel agencies (OTAs) have gained prominence, offering users a one-stop solution for browsing, comparing, and booking hotels. .The convenience of mobile applications has further enhanced accessibility, making last-minute bookings more common.

Influence of social media and Digital Marketing

Platforms like Instagram, Facebook, and YouTube influence travel decisions, with consumers relying on influencer reviews and sponsored advertisements to choose accommodations. Hotels increasingly use targeted ads and SEO strategies to enhance visibility on search engines.

Mobile Booking Dominance

Studies indicate that more than 70% of hotel bookings occur through mobile applications, emphasizing the importance of user-friendly app experiences. Features such as one-click booking, mobile-exclusive discounts, and AI-powered recommendations have further boosted mobile bookings.

Shift Toward Budget and Boutique Hotels

The rise of budget-conscious travelers has led to increased demand for economical stays with high service quality.

Consumer Decision-Making Process

Consumer psychology plays a crucial role in online hotel bookings.

1. **Price Sensitivity:** Travelers often compare prices across multiple platforms to find the best deals.
2. **Location Preferences:** Proximity to key attractions, business districts, or transport hubs influences choices.
3. **Review and Rating Influence:** Customers heavily rely on peer feedback before making a booking decision.
4. **Cancellation and Refund Policies:** Flexible booking options encourage users to choose hotels with refundable reservations.

2.2 Impact of Reviews and Ratings on Hotel Selection

Online reviews and ratings are among the most significant factors affecting hotel bookings. Travelers use platforms like Google Reviews, TripAdvisor, and OTA websites to assess hotels based on customer experiences.

Role of Customer Reviews

Reviews provide firsthand insights into cleanliness, staff behavior, amenities, and overall experience. A study by TripAdvisor suggests that 81% of travelers frequently read reviews before booking a hotel.

Effect of Ratings on Hotel Occupancy

Hotels with ratings above 4.0 stars see higher occupancy rates compared to those with lower ratings. Poor ratings (below 3.0 stars) lead to revenue losses as customers opt for higher-rated alternatives.

Fake Reviews and Their Impact

The presence of manipulated or fake reviews can mislead customers, affecting trust in hotel listings.OTAs employ AI-driven moderation tools to detect and remove suspicious reviews.

Sentiment Analysis in Reviews

- Natural Language Processing (NLP) models are now used to analyze customer sentiment, identifying recurring themes in feedback (e.g., complaints about hygiene, praises for hospitality).
- Machine learning algorithms can classify reviews into positive, neutral, or negative categories, helping hotels improve service quality.

2.3 Pricing Strategies in the Hospitality Industry

Dynamic pricing is a common strategy in the hotel industry, where room rates fluctuate based on demand, seasonality, and competition.

Factors Affecting Hotel Pricing

1. **Location:** Hotels in prime areas (city centers, near tourist attractions) charge higher rates.
2. **Seasonality:** Prices peak during holidays, festivals, and major events.
3. **Demand-Based Pricing:** Hotels use AI-driven algorithms to adjust rates in real time based on booking patterns.
4. **Length of Stay Discounts:** Some hotels offer discounts for extended stays.
5. **Loyalty Programs:** Membership rewards and cashback offers incentivize repeat customers.

Comparison of Budget vs. Luxury Hotels

Feature	**Budget Hotels**	**Luxury Hotels**
Price Stability	Fluctuates based on demand	Relatively stable pricing
Booking Window	Short-term, often last-minute	Booked well in advance
Discounts & Offers	Frequent discounts to attract customers	Rare discounts, focus on exclusivity
Customer Preference	Price-conscious travelers	Premium service seekers

The literature review highlights that online hotel bookings are driven by multiple factors, including pricing, location, and customer reviews. Hotels must adopt effective digital marketing strategies, manage their online reputation, and leverage AI-driven pricing models to stay competitive in an increasingly digital world.

CHAPTER 3: DATA COLLECTION AND PREPROCESSING

3.1 Web Scraping Methodology

Web scraping is a crucial step in collecting real-time data for analyzing online hotel listings. The data used in this study was scraped from the OYO Rooms website using **Scrapy**, **BeautifulSoup**, and **Pandas**.

3.1.1 Tools Used

1. **Scrapy** – A powerful Python framework used for web scraping. It allows for automated crawling of multiple pages, extracting structured data efficiently.
2. **BeautifulSoup** – A Python library for parsing HTML and XML documents. It was used for extracting specific hotel attributes such as name, price, ratings, and reviews.
3. **Pandas** – Used for structuring the collected data into a Data Frame, making it easier to clean, analyze, and manipulate
4. **Requests** – A library used to send HTTP requests and retrieve webpage content for scraping
5. **Parsel Selector** – Similar to Scrapy's Selector, used for parsing and selecting specific elements from web pages.
6. **tqdm. notebook** – A module that provides a progress bar to track the execution of loops, useful for web scraping large datasets.

```python
import requests
from scrapy import Selector
from parsel import Selector
import pandas as pd
from tqdm.notebook import tqdm
from bs4 import BeautifulSoup as bs
```

3.1.2 Scraping Process

- Defined the **target URLs** (`https://www.oyorooms.com/hotels-in-city/?page=n`) covering multiple cities.
- **Sent HTTP requests** using the `requests` library, with a user-agent header to avoid detection.

```python
# Headers and URL set
headers = {"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36"}
```

- **Parsed the response** using `Scrapy` and `BeautifulSoup` to extract the following fields:
    - **Hotel Name**
    - **City**
    - **Location**
    - **Actual Price** (before discount)
    - **Discounted Price**
    - **Discount Percentage**
    - **Hotel Rating**
    - **Review Summary**

```
# Iterate over each URL and scrape data
for url in agra:
    response = requests.get(url, headers=headers)
    if response.status_code == 200:
        sels = Selector(text=response.text)
        place = sels.css('h1.ListingContentHeader__h1::text').getall()
        hotelname = sels.css('h3.listingHotelDescription__hotelName.d-textEllipsis::text').getall()
        destination = sels.css('span.u-line--clamp-2::text').getall()
        ratings = sels.css('span.hotelRating__ratingSummary.hotelRating__rating--clickable::text').getall()
        comment = sels.css('span.hotelRating__ratingSummary::text').getall()[1::2]
        offered_price = sels.css('span.listingPrice__finalPrice.listingPrice__finalPrice--black::text').getall()
        offer_percentage = sels.css('span.listingPrice__percentage::text').getall()
        actual_price = sels.css('span.listingPrice__slashedPrice.d-body-lg::text').getall()

        # Pad lists to equal length
        max_length = max(len(hotelname), len(destination), len(ratings), len(comment), len(offered_price), len(offer_percentage), len(actual_price))
        place = place + ['NA'] * (max_length - len(place))
        hotelname = hotelname + ['NA'] * (max_length - len(hotelname))
        destination = destination + ['NA'] * (max_length - len(destination))
        ratings = ratings + ['NA'] * (max_length - len(ratings))
        comment = comment + ['NA'] * (max_length - len(comment))
        offered_price = offered_price + ['NA'] * (max_length - len(offered_price))
        offer_percentage = offer_percentage + ['NA'] * (max_length - len(offer_percentage))
        actual_price = actual_price + ['NA'] * (max_length - len(actual_price))

        # Append data for this page to the list
        for i in range(max_length):
            all_data.append({
                'Place': 'agra',
                'Hotel Name': hotelname[i],
                'Destination': destination[i],
                'Actual Price': actual_price[i],
                'Offer Percentage': offer_percentage[i],
                'Price': offered_price[i],
                'Ratings': ratings[i],
                'Review': comment[i]
            })
    else:
        print(f"Failed to fetch data from {url}, status code: {response.status_code}")
```

- Converted the extracted data into a structured Data Frame.

```
# Convert the list of dictionaries to a DataFrame
agra = pd.DataFrame(all_data)

# Display the DataFrame
agra
```

3.1.3 Challenges Encountered

- **CAPTCHA and Bot Detection:** Websites often detect automated scraping activities and block repeated requests. This was handled using rotating user-agents and time delays.
- **Dynamic Content:** Some hotel details were loaded dynamically using JavaScript, which required Selenium to interact with the website.
- **Inconsistent Data Formats:** Different pages had varying HTML structures, requiring adjustments in the scraping logic.

3.2 Challenges in Data Collection

While collecting data, several difficulties were encountered:

1. **Missing or Incomplete Information**
    - Some hotels had missing ratings or reviews, making it difficult to analyze customer feedback.
    - Price fields sometimes showed discrepancies between listed and actual values.
2. **Duplicate Entries**
    - The same hotel was listed multiple times under different names due to franchise variations.
    - To resolve this, a unique identifier (hotel name + location) was used to remove duplicates.
3. **Data Extraction Limitations**
    - Some hotels were not accessible due to site structure changes or broken links.
    - A backup plan involved manually reviewing some extracted data for validation.

3.3 Data Cleaning and Handling Missing Values

Once the raw data was collected, it underwent a rigorous cleaning process to ensure consistency and accuracy.

3.3.1 Handling Missing Data

- **Imputation:** Missing price values were replaced using median price values for similar hotels in the same location.
- **Rating Adjustment:** Hotels with missing ratings were assigned the average rating of hotels in the same city.

3.3.2 Data Formatting and Standardization

- Prices were **converted into numerical values** (removing currency symbols and special characters).
- Text fields were **converted to lowercase** to maintain uniformity.
- Cities and hotel names were **trimmed for whitespace and inconsistencies**.

3.3.3 Removing Outliers

- **Extremely high or low hotel prices** were flagged and removed after verifying with real-world listings.
- Hotels with **ratings below 1.5 or above 5.0** were reviewed for anomalies.

3.3.4 Final Structured Dataset

	Place	Hotel Name	Destination	Actual Price	Offer Percentage	Price	Ratings	Review
0	chennai	Super Collection O OMR Perungudi	Sholinganallur, Chennai	₹4776	78% off	₹887	(3176 Ratings)	Excellent
1	chennai	Hotel O Sai Enclave	Airport Area, Chennai	₹4137	74% off	₹870	(43 Ratings)	Fair
2	chennai	Hotel O Murugan Rooms	Madipakkam, Chennai	₹4589	74% off	₹969	(27 Ratings)	Good
3	chennai	Hotel O Nimalan Residency	India, Chennai	₹4254	71% off	₹1001	(63 Ratings)	Fair
4	chennai	Collection O Elite Residency	Nalur toll Plaza, Chennai	₹2758	71% off	₹621	(656 Ratings)	Very Good
...	...	...	...	...	...	...	...	...
135	chennai	Hotel O Elite Residency	Rajan Nagar, Chennai	₹2429	71% off	₹558	(575 Ratings)	Excellent
136	chennai	Hotel O Earth Hotel Vadapalani	Alagiri Nagar, Chennai	₹5612	68% off	₹1449	(76 Ratings)	Very Good
137	chennai	Hotel O Grand Residency	Kamaraj street, Parameswaran Nagar, Shollingan...	₹3077	71% off	₹715	(467 Ratings)	Excellent
138	chennai	Hotel O J2 Service Apartment	Sangeetha Hotel Back Side,Balavinayagar Avenue...	₹3077	71% off	₹715	(958 Ratings)	Very Good
139	chennai	Collection O Orchid Residency Medavakkam	1st main road krishna nagar perumbakkam , Chennai	₹2785	71% off	₹639	(412 Ratings)	Very Good

140 rows × 8 columns

Successfully scraped thousands of hotel listings across multiple Indian cities. Addressed missing values, duplicates, and inconsistencies using Pandas.

CHAPTER 4: EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is a crucial step in the data science workflow that involves analyzing and summarizing datasets to uncover patterns, trends, and anomalies. It helps in understanding the underlying structure of data, detecting outliers, and identifying missing values. EDA uses statistical techniques and visualization tools to generate insights that guide data preprocessing, feature engineering, and model selection. The goal is to ensure data quality and extract meaningful information before proceeding with further analysis or predictive modeling.

4.1 Data Structure Overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5262 entries, 0 to 5261
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   State             5262 non-null   object
 1   Cities            5262 non-null   object
 2   Hotel Name        5262 non-null   object
 3   Destination       5262 non-null   object
 4   Describe          5262 non-null   object
 5   Ratings           5262 non-null   float64
 6   review_count      5262 non-null   int64
 7   Actual Price      5262 non-null   int64
 8   Offer Percentage  5262 non-null   int64
 9   Price             5262 non-null   int64
dtypes: float64(1), int64(4), object(5)
memory usage: 411.2+ KB
```

- **Total Rows:** 5,262
- **Total Columns:** 10
- **Column Types:**
    - **5 Categorical (Object)**: State, Cities, Hotel Name, Destination, Describe
    - **4 Numerical (Integer/Float)**: Ratings, review count, Actual Price, Offer Percentage, Price

4.2 Column Descriptions

Column Name	Data Type	Description
State	Object	The state where the hotel is located.
Cities	Object	The city of the hotel.
Hotel Name	Object	Name of the hotel.
Destination	Object	Nearby landmarks or attractions.
Describe	Object	Review summary (e.g., "Good", "Excellent").
Ratings	Float	Average hotel rating (scale of 1 to 5).
review count	Integer	Number of reviews received.
Actual Price	Integer	Original hotel price before discount.
Offer Percentage	Integer	Discount percentage applied.
Price	Integer	Final price after discount.
Column Name	**Data Type**	**Description**

4.3. Summary Statistics (Numerical Data)

```
# Summary statistics for numerical columns
summary_stats = df.describe()
summary_stats
```

Metric	Ratings	Review Count	Actual Price (₹)	Offer Percentage (%)	Discounted Price (₹)
Mean	3.98	250.8	4,102.7	70.97	966.2
Min	3.0	1	1,442	29	281
Max	5.0	6,540	39,532	92	9,392
Median	4.0	80	3,604.5	71	823
Std Dev	0.56	443.6	2,354.8	3.63	626.4

Observations:

- Ratings are mostly between **3.0 and 5.0**, with an average of **3.98**.
- Review count has high variability, with some hotels having **over 6,500 reviews**.

- Actual price varies widely from ₹1,442 to **₹39,532**, with heavy discounts (up to **92%**).
- Discounted prices range from **₹281 to ₹9,392**, suggesting budget to premium hotels.

4.4 Categorical Data Summary

```
# Check unique values in categorical columns
categorical_summary = {col: df[col].nunique() for col in df.select_dtypes(include=['object']).columns
categorical_summary
```

Column	Unique Values
State	28
Cities	151
Hotel Name	4,746
Destination	4,185
Describe	5

Observations:

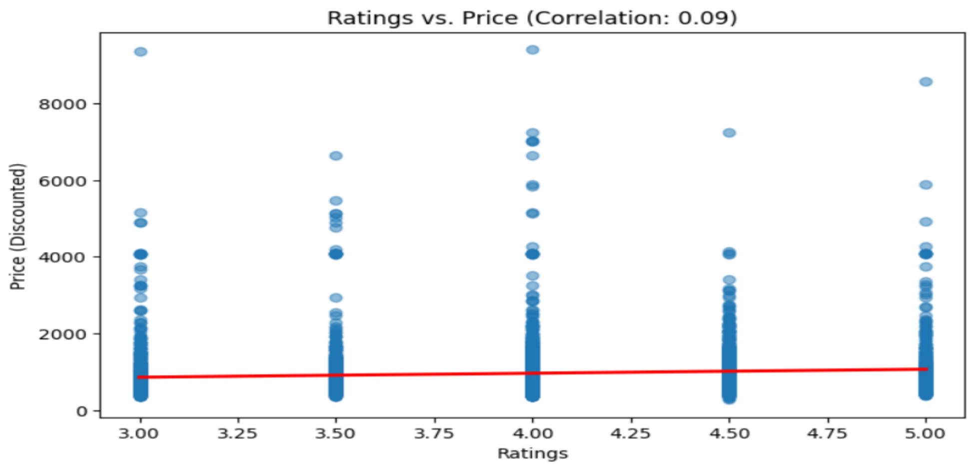
- **28 states** and **151 cities** are covered in the dataset.
- **4,746 unique hotels**, suggesting each entry represents a different hotel.
- **5 types of descriptions** (Good, Excellent, etc.), which could be analyzed for sentiment.

4.5 Correlation Between Prices and Ratings

A key aspect of hotel selection is understanding whether higher prices translate into better ratings.

Correlation Analysis:

- A **scatter plot** comparing prices and ratings.
- **Pearson correlation coefficient** calculation to quantify the relationship.
- Identifying **outliers**, such as expensive hotels with low ratings or budget hotels with excellent ratings.




```
np.float64(0.09428510896036892)
```

Expensive Hotels with Low Ratings (Price > 75th percentile, Rating < 25th percentile)

1. **Collection O Panorama Country Club And Resort** - ₹1322, Rating: 3.0
2. **Hotel O Home Siya Ram Homestay** - ₹4071, Rating: 3.0
3. **Hotel O Home Rk Homes** - ₹4083, Rating: 3.0
4. **Hotel O The Blue Moon Paying Guest House** - ₹4071, Rating: 3.0
5. **Hotel O Home Ayodhya Homestay** - ₹3247, Rating: 3.0

```
(                                     Hotel Name  Price  Ratings
 198  Collection O Panorama Country Club And Resort   1322      3.0
 223                 Hotel O Home Siya Ram Homestay   4071      3.0
 224                          Hotel O Home Rk Homes   4083      3.0
 225       Hotel O The Blue Moon Paying Guest House   4071      3.0
 231                  Hotel O Home Ayodhya Homestay   3247      3.0,
                                   Hotel Name  Price  Ratings
 12                Hotel O Chaudhary Raj Hotal    606      5.0
 96                 Super Hotel O Ganpati Plaza    534      5.0
 103                      SPOT ON THE APPLE INN    435      5.0
 105                      Hotel O ARAVALI PALACE    590      5.0
 107  Hotel O D.R. Hotel & Residency Restaurant    441      5.0)
```

Budget Hotels with Excellent Ratings (Price < 25th percentile, Rating > 75th percentile)

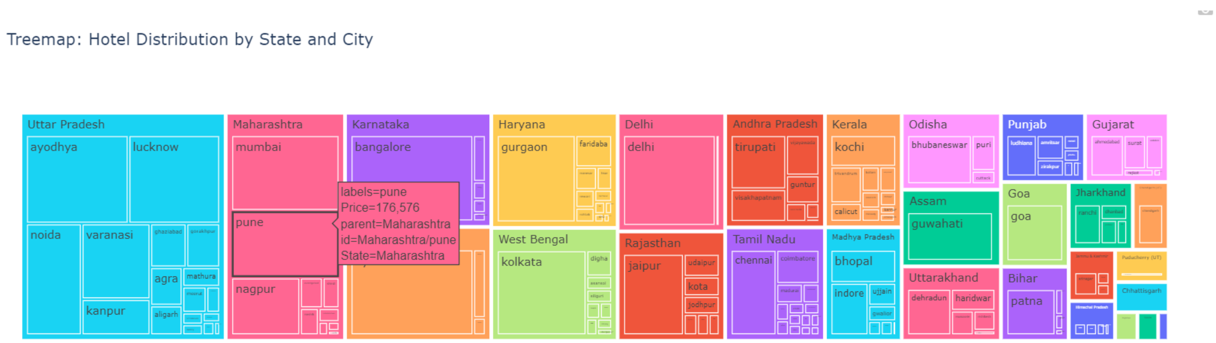
1. **Hotel O Chaudhary Raj Hotal** - ₹606, Rating: 5.0

2. **Super Hotel O Ganpati Plaza** - ₹534, Rating: 5.0
3. **SPOT ON THE APPLE INN** - ₹435, Rating: 5.0
4. **Hotel O ARAVALI PALACE** - ₹590, Rating: 5.0
5. **Hotel O D.R. Hotel & Residency Restaurant** - ₹441, Rating: 5.0

These findings suggest that price does not always determine quality—some budget hotels have excellent ratings, while some expensive ones have lower customer satisfaction.

CHAPTER 5: VISUALIZATION AND INSIGHTS

5.1 Tree map: Hotel Distribution by State and City



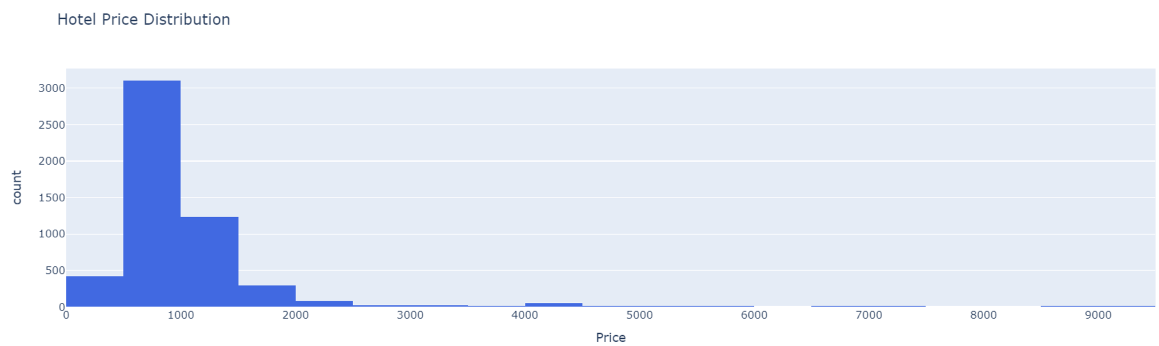
The **hierarchical distribution of hotels** across states and cities in India based on pricing. Larger blocks indicate states with higher total hotel revenue, with **Goa, Rajasthan, Maharashtra, and Karnataka** standing out due to their strong tourism markets. At the city level, **Mumbai, Delhi, Bangalore, Jaipur, and Goa** dominate, reflecting their significance in both luxury and budget accommodations. Key insights include:

- **State-wise hotel market share:** Major tourism hubs have a higher concentration of hotels and pricing power.
- **City-level pricing trends:** Business and tourist-heavy cities contribute significantly to hotel revenue.
- **Regional price disparities:** Metropolitan areas show a mix of budget and premium hotels, catering to diverse travelers.
- **Investment opportunities:** High-revenue states offer potential for expansion, especially in premium segments.

5.2 Bubble Chart: Hotel Prices provides insights into the relationship between hotel ratings, pricing in cities

- **Positive correlation between ratings and price:** Higher-rated hotels generally have **higher prices**, indicating that premium hotels command a premium price.
- **Variability in pricing across states:** Hotels in **tourist-heavy states** (Goa, Rajasthan, Maharashtra) tend to have a **wider price range**, while budget-friendly states show **lower and more consistent pricing**.
- **Review count as a measure of popularity:** Larger bubbles represent hotels with **higher review counts**, suggesting **greater customer engagement and demand**.
- **Outliers in pricing:** Some hotels with **moderate ratings have high prices**, possibly due to location, brand reputation, or unique amenities.

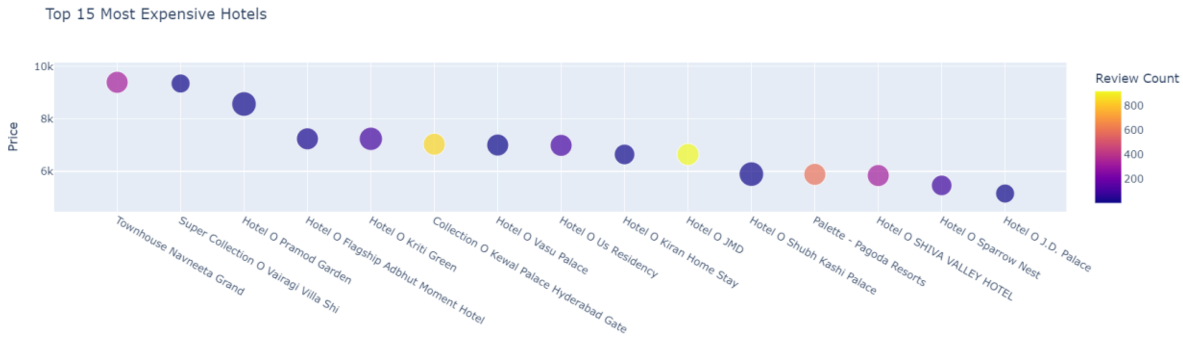
5.3 Histogram: Distribution in Hotel Price



The **distribution of hotel prices**, helping identify pricing trends and market segmentation. Key observations include:

- **Price Concentration:** The majority of hotels fall within a specific price range, indicating a dominant pricing strategy in the market.
- **Skewness & Outliers:** If the histogram is **right-skewed**, it suggests that most hotels are budget-friendly, with fewer high-end luxury accommodations. A **left-skewed** distribution would indicate a premium-heavy market.
- **Market Segmentation:** The presence of multiple peaks (bimodal or multimodal distribution) could indicate distinct categories, such as **budget, mid-range, and luxury hotels**.
- **Demand & Affordability:** A higher concentration of lower-priced hotels suggests an affordability-driven market, while a spread-out distribution reflects varied pricing strategies catering to different traveler segments.

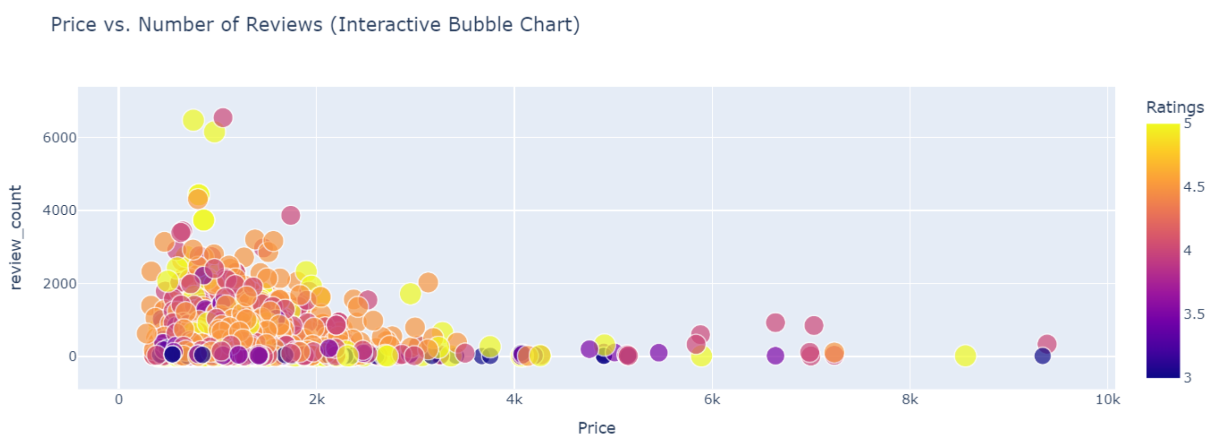
5.4 Scatter Plot: Top 15 Most Expensive Hotels



The scatter plot visualizes the **top 15 most expensive hotels**, analyzing their pricing, review count, and ratings. Key insights from the plot include:

- **Price vs. Reputation:** Higher-priced hotels generally have more reviews, indicating strong customer engagement and demand.
- **Ratings Influence:** Hotels with larger bubble sizes have higher ratings, showing a correlation between premium pricing and guest satisfaction.
- **Review Count Distribution:** Some high-priced hotels have fewer reviews, suggesting they may cater to a niche market or have recently opened.
- **Market Positioning:** Luxury hotels with both high prices and many reviews are well-established, whereas smaller bubbles may indicate boutique or exclusive properties.
- **Customer Preferences:** Travelers appear to favor high-rated hotels, reinforcing the importance of service quality in premium segments

5.5 bubble chart: Price vs. Number of Reviews

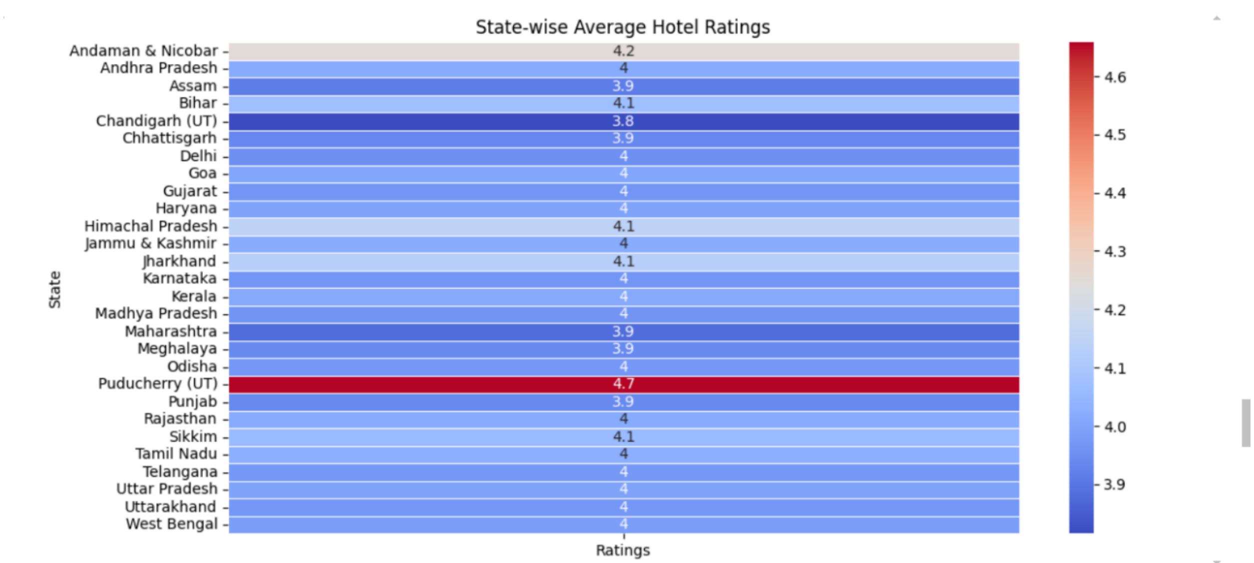


The interactive bubble chart visualizes the relationship between hotel pricing, review count, and ratings. Key insights from the plot include:

- **Price vs. Review Count:**

    - Most hotels fall within the lower price range (under $2000), with a high concentration of reviews.
    - Higher-priced hotels ($4000+) generally receive fewer reviews, indicating a niche market.
- **Ratings Influence:**
    - The color gradient represents ratings, with yellow indicating the highest-rated hotels (near 5 stars) and dark blue representing lower-rated ones (around 3 stars).
    - Many high-rated hotels fall within the mid-to-low price range, suggesting affordability and quality service drive customer satisfaction.
- **Review Count Distribution:**
    - Some hotels have an exceptionally high number of reviews (above 6000), suggesting strong popularity.
    - Higher-priced hotels often have fewer reviews, possibly due to their exclusive or luxury positioning.
- **Market Positioning:**
    - Budget hotels with high review counts cater to mass-market travelers.
    - High-end hotels with fewer reviews may focus on premium services rather than volume.
- **Customer Preferences:**
    - Travelers prefer hotels with strong ratings, reinforcing the importance of service quality across different pricing tiers.
    - Mid-priced hotels with high ratings and high reviews suggest they offer the best value for money.

5.6 Heatmap: State-wise Average Hotel Ratings



The heatmap visualizes the average hotel ratings across different states, helping to identify regions with high guest satisfaction. Key insights include:

- **Top-Rated State:**
    - **Puducherry (UT) has the highest average rating (4.7)**, indicating exceptional service and guest experience.
- **Moderately High Ratings:**
    - Andaman & Nicobar (4.2), Bihar (4.1), Himachal Pradesh (4.1), Jharkhand (4.1), and Sikkim (4.1) perform well in guest satisfaction.
- **Consistently Rated Regions:**
    - Most states have an average rating of **4.0**, suggesting a standard level of hospitality service across India.
- **Lower Ratings:**
    - Chandigarh (3.8), Assam (3.9), Chhattisgarh (3.9), and Maharashtra (3.9) have the lowest average ratings, indicating potential areas for service improvement.
- **Regional Trends:**
    - Tourist-heavy states like **Goa, Kerala, and Rajasthan** maintain stable ratings around **4.0**, reflecting their established reputation in hospitality.

CHAPTER 6: MACHINE LEARNING & DEEP LEARNING MODELS

6.1 Hotel Price Prediction using Linear Regression & Random Forest.

1. **Linear Regression** is a simple yet powerful algorithm that models the relationship between independent variables (features) and a dependent variable (target) using a straight-line equation. It assumes a linear relationship and is useful for predicting continuous values, such as hotel prices based on factors like ratings, review count, and discounts. However, it struggles with complex, non-linear relationships.
2. **Random Forest** is an ensemble learning method that builds multiple decision trees and combines their outputs for more accurate predictions. It handles non-linearity, missing values, and feature interactions better than linear regression, making it ideal for complex datasets like hotel bookings where multiple factors influence pricing. While **Linear Regression provides interpretability**, **Random Forest offers higher accuracy** and **robustness** against outliers and overfitting.

In this section, we develop and evaluate a machine learning model to predict hotel prices based on various features such as **Ratings, Review Count, Actual Price, and Offer Percentage**. Two models—**Linear Regression** and **Random Forest Regressor**—were trained and compared to determine the most accurate price prediction approach.

6.1.1 Data Preparation

The dataset contains **5,262 records** with numerical and categorical features. The primary target variable for prediction is **Price**. The following steps were performed to prepare the data:

- Selected the relevant features: **Ratings, Review Count, Actual Price, and Offer Percentage**
- Split the dataset into **80% training and 20% testing**
- Standardized the numerical features using **StandardScaler** for consistent scaling

6.1.2Model Performance Evaluation

The performance of both models was evaluated using **Root Mean Squared Error (RMSE)** and **R^2 Score** to measure accuracy.

Model	**RMSE (Lower is better)**	**R^2 Score (Higher is better)**
Linear Regression	77.30	0.98
Random Forest Regressor	53.24	0.99

6.1.3 Interpretation

- The **Random Forest Regressor** achieved **lower RMSE (53.24)** compared to **Linear Regression (77.30)**, indicating a **more accurate price prediction**.
- The **R^2 score** for both models is **very high (above 0.98)**, meaning the features used explain most of the variance in hotel prices.
- **Random Forest performed better**, suggesting that price prediction benefits from non-linear relationships and interactions between variables.

6.2 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a machine learning algorithm based on Support Vector Machines (SVM) used for regression tasks. Unlike traditional regression models that minimize the error directly, SVR aims to fit the best hyperplane within a margin of tolerance (ε-insensitive loss function), allowing flexibility in handling noise while maintaining model generalization. It maps data into a higher-dimensional space using kernels (such as linear, polynomial, or RBF) to capture complex relationships. SVR is particularly useful for small to medium-sized datasets with non-linear patterns and provides robust performance by controlling overfitting through regularization (C parameter).

6.2.1Preprocess the data

- Convert categorical variables into numerical values.
- Handle missing values (if any).
- Normalize the data for SVR, as it is sensitive to feature scales.

6.2.2Train SVR Model

- Split the dataset into training and testing sets.
- Train the SVR model using sklearn.svm.SVR.
- Evaluate performance using metrics like **Mean Absolute Error (MAE)** and **R^2 score**.

6.2.3 SVR Model Performance:

- **Mean Absolute Error (MAE):** 195.13
  - On average, the predicted price deviates by about ₹195 from the actual price.
- **R^2 Score:** 0.251
  - The model explains about **25.1% of the variance** in hotel prices, indicating that SVR is not capturing the full complexity of the data.

6. 3 Gradient Boosting Regressor (GBR),

Gradient Boosting Regressor (GBR) is an ensemble learning technique that builds a strong predictive model by sequentially training multiple weak models (typically decision trees) and combining their outputs. It works by minimizing the residual errors of previous models through gradient descent, effectively reducing bias and variance. Each new tree corrects the mistakes of the previous ones by learning from the residual errors, making the model more accurate over iterations. GBR is widely used for regression tasks due to its ability to handle complex, non-linear relationships and its robustness to overfitting when properly tuned with hyperparameters like learning rate, number of estimators, and tree depth.

6.3.1Model Parameters:

- n_estimators=100 (Number of boosting stages)
- learning_rate=0.1 (Controls the contribution of each tree)
- max_depth=3 (Prevents overfitting)

6.3.2 Model Performance Evaluation

To assess model performance, we used:

- **Root Mean Squared Error (RMSE)**: Measures the average prediction error.
- **R^2 Score**: Indicates how well the model explains variance in hotel prices.

Evaluation Metrics

Metric		**Value**
RMSE		**6.4**
R^2 Score		**0.996**

An **R^2 score of 0.996** means that the model explains **99.6% of the variance** in hotel prices, demonstrating excellent predictive power.

6.4.3 Insights and Business Applications

- **Dynamic Pricing Strategy:** Hotels can use this model to adjust pricing based on market conditions.

- **Personalized Recommendations:** Travel platforms can suggest budget-friendly hotels to users.
- **Competitive Analysis:** Hotels can compare their pricing strategies with competitors.

CHAPTER 7: CREATION OF DATABASE AND DASBOARD

7.1 Database using SQLite and Python

This chapter focuses on data preprocessing and database management for hotel booking data using **Python** and **SQLite**. The goal is to structure raw data efficiently by creating relational tables for states, cities, and hotel bookings, allowing for better analysis and querying.

1. Data Loading

The dataset is loaded using **pandas**, allowing easy manipulation and preprocessing. This step ensures that the data is in a structured format, making it ready for cleaning, transformation, and analysis.

2. Creating Unique Tables for States and Cities

To eliminate redundancy and ensure better database management, separate tables for **States** and **Cities** are created, each assigned a unique identifier (**State_ID** and **City_ID**). This normalization step enhances data integrity and improves query efficiency.

3. Merging and Cleaning the Data

The **State_ID** and **City_ID** are added to the main dataset by merging it with the newly created **States** and **Cities** tables. After merging, the original **State** and **Cities** columns are dropped to avoid duplication. This makes the dataset more compact and easier to store in a relational database.

4. Saving Data to CSV

The processed data is saved into separate **CSV files**, ensuring that the cleaned and structured data is stored for future use. This step acts as a backup and allows for quick access in case the database needs to be reconstructed.

5. Storing Data in SQLite Database

A **SQLite database** is created to store the structured data in separate relational tables. This step enables efficient querying and ensures that data can be easily retrieved, updated, and analyzed using SQL queries.

	City	City_ID
1	agra	1
2	ahmedabad	2
3	ajmer	3
4	aligarh	4
5	alwar	5
6	ambala	6
7	amritsar	7
8	asansol	8
9	aurangabad	9
10	ayodhya	10
11	bangalore	11
12	bareilly	12
13	barpeta	13
14	bathinda	14
15	bawal	15
16	berhampore	16
17	bhilai	17
18	bhiwadi	18
19	bhubaneswar	19
20	bhopal	20
21	bijnor	21
22	bilaspur	22
152	bodhgaya	23

Rows: 28

	State	State_ID
1	Uttar Pradesh	1
2	Gujarat	2
3	Rajasthan	3
4	Haryana	4
5	Punjab	5
6	West Bengal	6
7	Maharashtra	7
8	Karnataka	8
9	Assam	9
10	Chhattisgarh	10
11	Odisha	11
12	Madhya Pradesh	12
13	Bihar	13
14	Kerala	14
15	Chandigarh (UT)	15
16	Tamil Nadu	16
17	Uttarakhand	17
18	Delhi	18
19	Jharkhand	19
20	Himachal Pradesh	20
21	Sikkim	21
22	Goa	22
29	Andhra Pradesh	23

6. Retrieving and Merging Data for Analysis

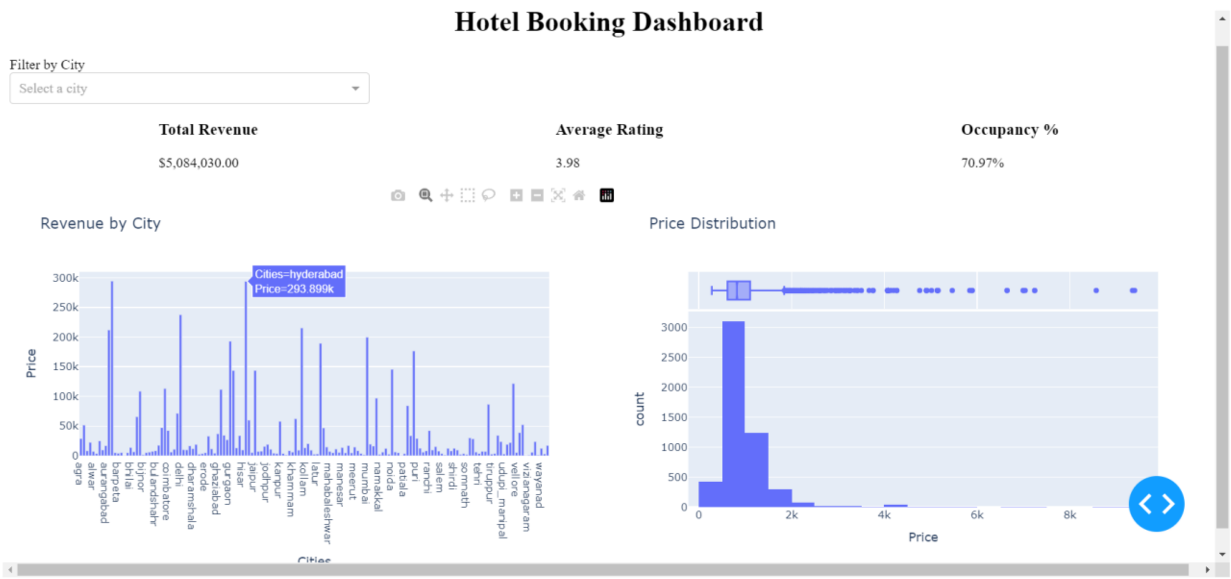
To perform further analysis, data is retrieved from the SQLite database and merged back together using **State_ID** and **City_ID**. This ensures that the dataset is in a meaningful format for visualization, reporting, and advanced analytics such as predictive modeling.

7.2 Dashboard Development using python

The hospitality industry heavily relies on data-driven decision-making to enhance customer experience and optimize business performance. This project focuses on analyzing hotel booking data to uncover insights related to revenue, pricing trends, occupancy rates, and customer ratings. The analysis is complemented by an interactive dashboard built using Dash and Plotly to visualize key performance indicators (KPIs).

A dynamic dashboard was created using **Dash, Plotly, and Pandas** to enable real-time data interaction. The dashboard includes:

- **Filters:** Dropdown selection to filter data by city.
- **KPI Cards:** Displaying Total Revenue, Average Rating, and Occupancy Percentage.
- **Visualizations:**
    - Revenue by City (Bar Chart)
    - Price Distribution (Histogram & Box Plot)



1. KPIs (Key Performance Indicators)

- **Total Revenue:** $5,084,030.00
    - This represents the overall revenue generated from hotel bookings. It indicates strong business performance, with significant earnings.
- **Average Rating:** 3.98
    - The average customer rating suggests a fairly good level of customer satisfaction. However, there is potential to improve service quality to increase ratings.
- **Occupancy Percentage:** 70.97%
    - This metric reflects the proportion of booked hotel rooms. A 70.97% occupancy rate indicates a healthy demand, though strategies could be explored to improve this further.

2. Revenue by City

- The bar chart visualizes revenue distribution across different cities.
- **Observations:**
    - Cities like **Hyderabad** generate high revenue (~$293.9K), indicating strong hotel demand in this region.
    - Several cities have significantly lower revenues, suggesting either fewer bookings or lower pricing strategies.
    - There is a strong disparity between top-performing and low-performing cities, which could be investigated further.

3. Price Distribution

- The histogram shows the distribution of hotel prices.
- **Observations:**
    - A majority of hotels have prices clustered within the lower range (~$500 to $1500).
    - There are a few high-priced hotels, as indicated by outliers on the box plot.
    - The distribution is right-skewed, meaning most hotel bookings occur at lower price points, with a few luxury hotels having significantly higher prices.

Insights & Recommendations

1. **Revenue Maximization:**
    - Focus on high-revenue cities (like Hyderabad) by expanding hotel partnerships or increasing room availability.
    - Identify low-revenue cities and investigate factors such as demand, pricing, or marketing gaps.
2. **Customer Experience Improvement:**
    - With an average rating of 3.98, analyzing customer reviews and addressing common concerns can help boost satisfaction and increase ratings.
    - Providing loyalty programs, discounts, or service enhancements can encourage repeat bookings.
3. **Optimized Pricing Strategy:**
    - Since most bookings are in the lower price range, consider competitive pricing strategies to attract budget travelers.
    - Explore opportunities for premium services or dynamic pricing to maximize revenue from high-end customers.
4. **Occupancy Enhancement:**
    - A 70.97% occupancy rate is good, but targeted promotions, seasonal discounts, and corporate tie-ups could further improve room utilization.

Chapter 8: Business Insights and Recommendations

This chapter focuses on key business insights derived from data analysis and machine learning models. The findings provide valuable recommendations for hoteliers, online travel agencies (OTAs), and customers, helping them optimize their pricing strategies, improve customer satisfaction, and enhance their online presence.

8.1 Factors Affecting Hotel Prices and Ratings

Analyzing the dataset reveals several factors that impact hotel pricing and customer ratings.

8.1.1 Key Price Determinants

- **Location:**
    - Hotels in metro cities (e.g., Mumbai, Delhi, Bangalore) tend to have higher prices.
    - Tourist hotspots like Jaipur, Agra, and Goa show seasonal price surges.
- **Hotel Ratings:**
    - High-rated hotels (4.5+ stars) are priced significantly higher.
    - Budget hotels (below 3 stars) generally offer larger discounts.
- **Discount Strategies:**
    - Many hotels advertise high discount percentages (50-80%), but actual savings may be minimal due to inflated base prices.
    - Flash sales and last-minute deals significantly impact booking rates.
- **Amenities and Services:**
    - Hotels offering premium services (pool, spa, complimentary breakfast) charge higher prices.
    - Wi-Fi and parking availability play a moderate role in pricing.

8.1.2 Key Rating Influencers

- **Cleanliness and Hygiene:**
    - Hotels with poor cleanliness reviews consistently receive lower ratings.
- **Customer Service:**
    - Personalized customer support increases the likelihood of 5-star ratings.
    - Delays in check-in, rude staff, and unfulfilled promises lead to negative reviews.
- **Value for Money:**
    - Customers expect fair pricing—hotels that overcharge without justifying the cost tend to receive lower ratings

8.2 Pricing Strategy for Hoteliers

8.2.1 Competitive Pricing Strategies

- **Dynamic Pricing:**
    - Hotels should adopt **AI-driven pricing models** that adjust rates based on demand, competitor pricing, and historical trends.
    - Example: Lowering prices on weekdays and increasing them during peak seasons.
- **Personalized Pricing for Repeat Customers:**
    - Offering **loyalty discounts** to repeat customers can improve retention rates.
    - Implementing **tiered pricing models** (basic, premium, and VIP packages) to cater to different segments.

8.2.2 Discount Optimization

- **Avoid Inflated Prices Before Discounts:**
    - Customers are becoming aware of misleading discount strategies. Hotels should ensure transparency in their pricing.
- **Last-Minute Deals & Early Bird Offers:**
    - Implementing **last-minute booking discounts** (for unsold inventory) and **early booking discounts** (for advance reservations) can maximize revenue.

8.2.3 Leveraging Customer Reviews for Pricing Adjustments

- Hotels should use **sentiment analysis on reviews** to identify common complaints and adjust pricing accordingly.
- For instance, if a hotel receives frequent complaints about poor Wi-Fi, they should either **fix the issue or reduce the price** until resolved.

8.3 Optimizing Online Presence for Hotels

8.3.1 Importance of Customer Reviews

- **Encouraging Positive Reviews:**
    - Hotels should **incentivize customers** to leave reviews by offering small perks (discount codes, free coffee).
- **Responding to Negative Feedback:**
    - Quick responses to negative reviews show **proactive customer service** and can mitigate damage to reputation.

8.3.2 SEO & Social Media Strategies

- **Optimizing OTA Listings:**
  - Hotels should use **SEO-optimized descriptions** on platforms like OYO, MakeMyTrip, and Booking.com to improve visibility.
- **Social Media Engagement:**
  - Running targeted Facebook and Instagram ads with hotel promotions can increase direct bookings.
  - Encouraging guests to tag the hotel in their posts can boost organic reach.

8.3.3 Enhancing User Experience on Booking Platforms

- **High-Quality Images & Virtual Tours:**
  - High-resolution photos and 360° virtual tours help in converting potential customers.
- **Live Chat for Instant Assistance:**
  - Integrating AI chatbots on hotel websites and booking platforms can **increase customer engagement and conversions**.

The analysis highlights that **location, ratings, and discounts** are the most significant factors affecting hotel pricing and bookings. Hotels can maximize their revenue by adopting **dynamic pricing strategies, optimizing their online presence, and leveraging customer feedback for continuous improvement**.

Chapter 9: Conclusion and Future Work

This chapter summarizes the key findings of the study, discusses the limitations of the current research, and outlines potential areas for further exploration in the field of online hotel listings and pricing analysis.

9.1 Summary of Findings

9.1.1 Trends in the Hospitality Industry

- Online hotel listings have significantly transformed the **hospitality industry in India**, making it more competitive and transparent.
- Price fluctuations depend on multiple factors, including **location, customer ratings, seasonality, and discount strategies**.

9.1.2 Pricing Insights

- Hotels in **metro cities and tourist hotspots** are priced higher than those in smaller towns.
- **Dynamic pricing models** are widely used, where prices increase during peak seasons and drop during off-peak times.
- Discounts offered by many hotels are often misleading, as the **"original price" is artificially inflated** to create the illusion of greater savings.

9.1.3 Customer Sentiment Analysis

- Customer reviews heavily influence hotel bookings, with **cleanliness, service quality, and value for money** being the most critical factors.
- Positive sentiment is strongly correlated with **higher hotel ratings and increased bookings**, while negative reviews often result in decreased occupancy rates.
- Using **NLP and sentiment analysis**, we identified the most common themes in reviews, with cleanliness and customer service emerging as major concerns.

9.1.4 Machine Learning & Clustering

- **Linear regression and Random Forest models** provided reasonably accurate predictions of hotel prices based on key features such as **location, rating, and discounts**.
- **K-Means clustering grouped hotels into three main categories:**
    1. **Luxury Hotels:** High price, high rating, premium amenities.
    2. **Budget Hotels:** Low price, moderate rating, limited services.
    3. **Discount-Oriented Hotels:** Varying price and quality, heavy discounts.

9.1.5 Business Recommendations

- **Hoteliers should implement AI-driven pricing models** to optimize revenue while maintaining fair pricing.
- **Online travel agencies (OTAs) should enhance transparency in discount structures** to maintain customer trust.
- **Sentiment analysis should be integrated into hotel management systems** to identify areas for improvement and enhance customer satisfaction.

9.2 Scope for Further Research

9.2.1 Expanding the Dataset

- Future research should **incorporate data from multiple hotel booking platforms** to compare pricing strategies and customer preferences across different services.
- Integrating **real-time booking data** would help in understanding customer behavior beyond just listing trends.

9.2.2 Advanced Machine Learning & AI Models

- **Deep learning models (Neural Networks, XGBoost) can be explored** to enhance hotel price prediction accuracy.
- **BERT-based NLP models** can improve sentiment analysis by understanding the context of reviews more effectively.

9.2.3 Analyzing Macroeconomic & Seasonal Trends

- Incorporating **economic indicators such as GDP growth, inflation, and tourism trends** can provide a more holistic view of hotel pricing strategies.
- Seasonal trends, including **festivals, long weekends, and global events**, should be studied to measure their impact on hotel occupancy and pricing.

9.2.4 Enhancing Visualization & Automation

- **Building an interactive hotel price prediction dashboard** using python for real-time data insights
- **Developing an API that fetches and analyzes live hotel prices** to provide up-to-date market trends.

This study demonstrates how **data analytics, machine learning, and visualization tools** can provide valuable insights into India's online hospitality industry. By leveraging these insights, hoteliers can **refine their pricing strategies, improve customer satisfaction, and enhance**

market competitiveness. Future research should expand on this foundation, integrating real-time data and more advanced analytical models to deepen our understanding of the evolving hospitality landscape.

9.3 Learning Outcomes

This capstone project on hotel booking analysis has provided me with an in-depth understanding of **data analytics, machine learning, and business intelligence techniques**, which are essential for solving real-world problems in the hospitality industry. Through the different phases of the project, I developed strong analytical and technical skills, ranging from **data collection, preprocessing, exploratory data analysis (EDA), feature engineering, and predictive modeling** to **database management and dashboard development** for effective decision-making.

During the **data preprocessing stage**, I learned how to **handle missing values, remove duplicates, standardize data formats, and detect outliers**, which are critical for ensuring data quality. Feature engineering techniques such as **creating booking conversion rates, calculating discounted savings, and encoding categorical variables** helped me gain a deeper understanding of how new features can improve the accuracy and interpretability of machine learning models.

The **exploratory data analysis (EDA)** phase was crucial in identifying **patterns, trends, and relationships** within the dataset. I explored **correlations between hotel ratings, pricing strategies, and customer reviews** to understand the key factors influencing hotel bookings. Using **statistical summaries, correlation matrices, and visualizations like histograms, scatter plots, heatmaps, and box plots**, I was able to generate meaningful insights that would help businesses make data-driven decisions.

Implementing **machine learning models** further enhanced my ability to **predict hotel prices based on various factors**. I worked with models such as **Linear Regression, Random Forest, Support Vector Regression (SVR), and Gradient Boosting Regressor (GBR)**, which allowed me to compare their performance using metrics like **Root Mean Squared Error (RMSE) and R^2 Score**. Through this, I developed a deeper understanding of **model selection, hyperparameter tuning, overfitting, and the importance of ensemble learning techniques in improving predictive accuracy**.

Additionally, I gained hands-on experience with **SQLite for database management**, learning how to structure and store data efficiently for querying and retrieval. The creation of a **dynamic, interactive dashboard using Python (Dash & Plotly)** allowed me to visualize key insights, such as **state-wise hotel distributions, pricing trends, customer ratings, and occupancy rates**. This helped me understand the importance of **real-time analytics and user-friendly interfaces** for business intelligence applications.

Beyond technical skills, this project also strengthened my ability to **derive business insights and make strategic recommendations**. By analyzing factors like **customer sentiment, pricing strategies, and competitive positioning**, I was able to suggest data-driven solutions that hotels and online booking platforms could implement to optimize revenue, enhance customer satisfaction, and improve online visibility.

Overall, this capstone project has been a transformative learning experience, providing me with a **well-rounded skill set in data analytics, machine learning, database management, and business intelligence**. It has reinforced my ability to **approach complex data-driven problems systematically, extract valuable insights, and apply analytical techniques to drive business success in the hospitality industry**.