Hotelytics

Ghazal E Ashar, Shahzeb Ahmed Iqbal, Mohammad Ali Hassan, Muhammad Ammar Thahim, Usama Habib
Department of Computer Science
Shaheed Zulfiqar Ali Bhutto Institue of Science and Technology
Karachi, Pakistan

Abstract— **The hospitality industry generates vast amounts of data, including hotel ratings, room pricing, amenities, and customer reviews. Hotelytics provides a comprehensive data-driven solution to analyze and visualize hotel data, offering actionable insights and AI-powered recommendations. The AI recommendation model, powered by TF-IDF Vectorization and Cosine Similarity, dynamically suggests hotels based on user-selected filters like price range, location, and amenities. This method was selected for its computational efficiency and interpretability compared to other approaches like collaborative filtering or neural models. The project integrates web scraping, data transformation, Power BI dashboards, and an AI recommendation engine to deliver an interactive and scalable solution for hotel analytics. Future enhancements include expanding datasets to improve accuracy and functionality.**

Keywords— AI Recommendations, Customer Insights, Data Analytics, Data Cleaning, Data Transformation, Data Visualization, Interactive Dashboards, Machine Learning, Power BI, Recommender Systems, Star Schema, TF-IDF, Web Scraping.

I. Introduction

The hospitality industry is rapidly evolving, driven by customer satisfaction, pricing competitiveness, and personalized services. To thrive in this environment, businesses must leverage data-driven insights to optimize their offerings and provide tailored recommendations. Hotelytics is a comprehensive hotel analytics and recommendation solution that integrates advanced data processing, visualization, and artificial intelligence.

Fig. 1. Logo for Hotelytics.

This project employs web scraping to collect hotel data from Expedia, encompassing over 53 cities across 36 countries. The data includes attributes such as hotel names, star ratings, reviews, pricing, and amenities. After extensive cleaning and transformation, the dataset is structured into a star schema for analytical efficiency. Using Power BI, interactive dashboards were developed to visualize trends, while an AI recommendation engine was integrated to suggest hotels dynamically based on user-selected preferences.

The solution demonstrates the potential of integrating data analytics and AI to deliver actionable insights and enhance decision-making in the hospitality sector. The following sections outline the methodologies used, including data extraction, cleaning, modeling, and visualization, along with key insights derived from the project.

II. Objectives

The primary goals of the Hotelytics project are outlined below. These objectives guided the methodology and design choices throughout the project:

A. *Primary Objectives*

- Data Extraction: Gather hotel-related data, including names, ratings, pricing, and services, through web scraping from Expedia.
- Data Cleaning and Transformation: Ensure data consistency by handling missing values, standardizing services, and creating a structured dataset.
- Trend Analysis: Visualize trends such as room type distribution, pricing patterns, and service popularity through interactive Power BI dashboards.
- AI-Powered Recommendations: Develop a recommendation model using TF-IDF Vectorization and Cosine Similarity to dynamically suggest hotels based on user-selected filters.
- Decision Support: Provide an interactive dashboard that enables users to make informed decisions using dynamic filters and drill-through functionalities.

B. *Secondary Goals*

- Design an intuitive user interface with Power BI to support exploration of detailed hotel metrics.
- Extract actionable insights into customer preferences and hotel performance.

III. Data Extraction

A. *Web Scraping Methodology*

The data for this project was collected through web scraping from the Expedia platform. Using Python's Selenium library, automated scripts were created to navigate hotel listing pages for cities like Abu Dhabi, extract relevant data, and store it in structured formats. The scraping process involved interacting with dynamic web elements and handling asynchronous content loading.

B. *Extracted Attributes*

Key attributes extracted during the process included:

- Hotel Names: Identifying unique hotels in the dataset.
- Star Ratings: Categorizing hotels based on quality standards.
- Room Pricing: Collecting nightly room rates in different currencies.
- Amenities and Services: Recording features such as WiFi, parking, and breakfast availability.

- Guest Ratings and Reviews: Gathering user feedback to assess customer satisfaction.

C. *Challenges and Solutions*

- Dynamic Content Loading: Many elements on the Expedia website load dynamically, requiring the use of explicit waits to ensure all data is visible before extraction.
- HTML Structure Variations: Different city pages had slight variations in layout, necessitating adaptable scraping logic.
- Data Volume: Large datasets required more time.

D. *Outcome*

The extracted data was stored in CSV files for each city, ensuring that all attributes were consistently formatted. The resulting datasets provided a strong foundation for subsequent cleaning, analysis, and visualization stages.

IV. Data Sources and Preparation

A. *Data Cleaning and Transformation*

The raw dataset underwent extensive cleaning to ensure consistency and usability:

1) Handling Missing Values:

a) Missing star ratings were assigned a default value of 1.

b) Null reviews were filled with the placeholder "No Reviews."

c) Average ratings were used to replace missing guest scores.

2) Standardization of Services:

a) Services were normalized to reduce redundancy (e.g., "Parking Available" → "Parking").

b) Similar services were grouped under unified terms.

3) Splitting Property Offers:

a) The column containing combined room details and charges was split into distinct attributes like Room Name, Room Charges, and Amenities.

B. *Dataset Combination*

Individual city-level datasets were merged into a unified dataset named combined_hotel_data.csv, comprising over 8,900 records. Duplicates were carefully removed while ensuring the preservation of accurate entries.

C. *Feature Engineering*

Key features were engineered from the raw dataset to support advanced analysis:

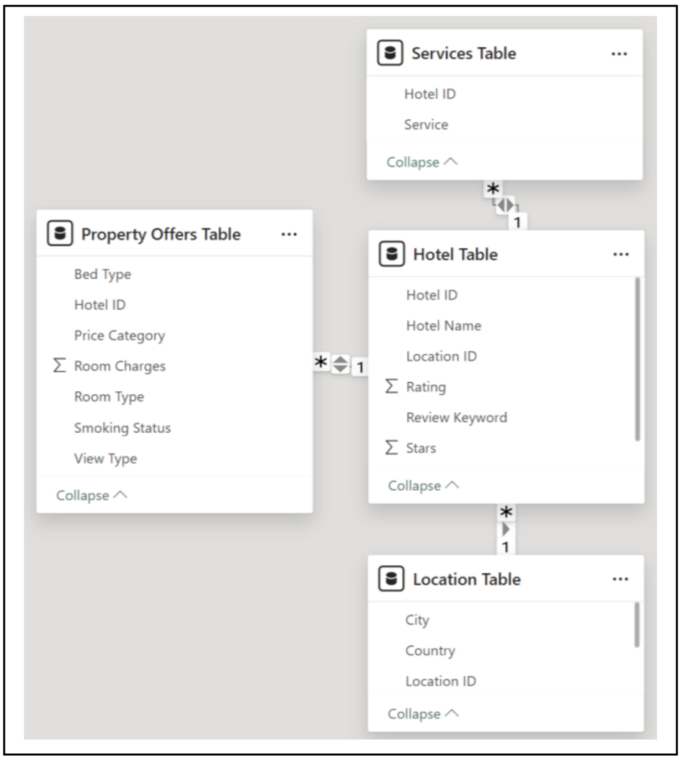
- Room Categories: Extracted based on keywords like "Suite," "Deluxe Room," and "Standard Room."
- Bed Types: Derived from room names using patterns (e.g., "King," "Queen," "Twin").
- View Preferences: Detected terms like "Sea View" and "City View" to categorize room views.

V. Data Analysis and Modeling

A. *Star Schema Design*

To optimize analytical performance and facilitate seamless querying in Power BI, a star schema was designed. This schema organizes the data into fact and dimension tables, ensuring efficient data retrieval.

Fig. 2. ERD for star schema.



The components of the schema include:

1) Fact Table:

a) *Hotel Table: Stores Hotel Name, Star Ratings, and Review Keywords.*

2) Dimension Tables:

a) *Location Table: Contains attributes like City, Country, and Location IDs.*

b) *Services Table: Maps services (e.g., Parking, WiFi) to respective Hotel IDs.*

c) *Property Offers Table: Captures detailed room attributes, including Room Name and Charges.*

B. *Visualizations in Power BI*

The Power BI dashboard visualized key trends and metrics, providing actionable insights through some of the following:

- Tree Map (Room Type Distribution): Showcases the count of each room type (e.g., Suite, Standard Room) across the dataset.
- Scatter Plot (Average Price vs. Guest Ratings): Highlights the correlation between room pricing and guest satisfaction scores.
- Funnel Chart (Hotel Stars Distribution): Displays the count of hotels categorized by star ratings (1–5 stars).
- Stacked Bar Chart (Services by Popularity): Compares the availability of amenities like WiFi, Parking, and Breakfast across hotels.

- Dynamic filters allowed users to explore the dataset interactively, refining insights by criteria such as location, price range, and star ratings.

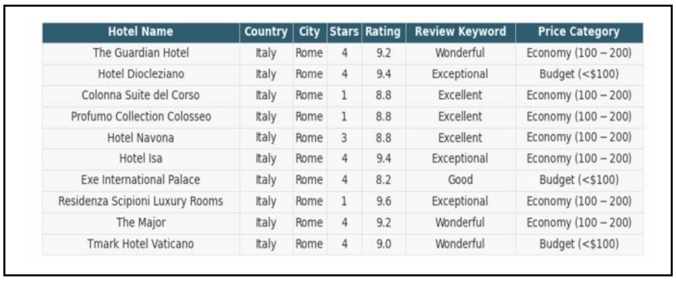
C. *AI Recommendation Model*

The AI recommendation system employed TF-IDF Vectorization and Cosine Similarity to suggest hotels based on user preferences. The process involves:

1) Data Preparation: Hotel attributes, including amenities, star ratings, and price categories, were converted into numerical vectors using TF-IDF.

2) Similarity Calculation: Cosine similarity was used to compare user-selected filters with hotel attributes, ranking hotels by relevance.

3) Recommendations: Top 10 hotels were dynamically suggested based on user input or overall dataset trends.



Hotel Name	Country	City	Stars	Rating	Review Keyword	Price Category
The Guardian Hotel	Italy	Rome	4	9.2	Wonderful	Economy (100 – 200)
Hotel Diocleziano	Italy	Rome	4	9.4	Exceptional	Budget (<$100)
Colonna Suite del Corso	Italy	Rome	1	8.8	Excellent	Economy (100 – 200)
Profumo Collection Colosseo	Italy	Rome	1	8.8	Excellent	Economy (100 – 200)
Hotel Navona	Italy	Rome	3	8.8	Excellent	Economy (100 – 200)
Hotel Isa	Italy	Rome	4	9.4	Exceptional	Economy (100 – 200)
Exe International Palace	Italy	Rome	4	8.2	Good	Budget (<$100)
Residenza Scipioni Luxury Rooms	Italy	Rome	1	9.6	Exceptional	Economy (100 – 200)
The Major	Italy	Rome	4	9.2	Wonderful	Economy (100 – 200)
Tmark Hotel Vaticano	Italy	Rome	4	9.0	Wonderful	Budget (<$100)

Fig. 3. Sample Recommendations.

D. *Model Comparison and Justification*

The table below compares TF-IDF with alternative models:

TABLE I. RECOMMENDATION MODEL COMPARISION

Model	***Advantages***	***Disadvantages***	***Relevance***
TF-IDF + Cosine Similarity	Simple, efficient, and interpretable.	Limited contextual understanding.	Highly suitable due to computational efficiency and the nature of structured data.
Collaborative Filtering	Learns from user behavior; effective for large datasets.	Requires historical user interaction data; suffers from cold-start issues.	Not applicable as user interaction data is unavailable.
Neural Networks	Captures complex relationships and patterns.	Computationally expensive; requires labeled training data.	Overly complex for the scope of this project and lacks interpretability for business users.
Hybrid Systems	Combines collaborative and content-based filtering.	High computational cost; complex implementation.	Unnecessary for the initial phase; TF-IDF adequately meets project requirements.

TF-IDF + Cosine Similarity was selected for its balance of simplicity, efficiency, and scalability. Unlike collaborative filtering, it does not rely on user behavior data, making it ideal for the project's requirements.

VI. Power BI Dashboard and User Interface

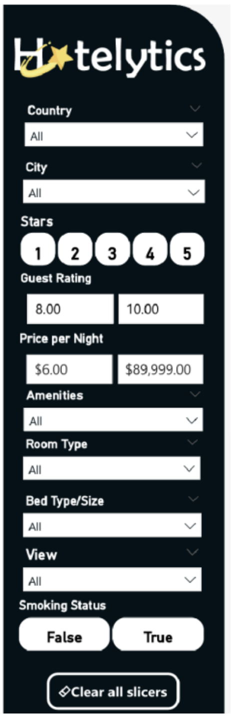
A. *Dashboard Overview*

The Power BI dashboard serves as the central interface for exploring hotel data and deriving actionable insights. Designed with a user-centric approach, the dashboard combines visual appeal with functionality to provide an intuitive and interactive experience.

B. *Key Features*

1) Dynamic Filters

a) Slicers for attributes like city, star ratings, price range, and amenities enable users to refine their view of the data interactively.



Country: All
City: All
Stars: 1 2 3 4 5
Guest Rating: 8.00 10.00
Price per Night: $6.00 $89,999.00
Amenities: All
Room Type: All
Bed Type/Size: All
View: All
Smoking Status: False True

Fig. 4. Filter Selection in the dashboard clearly highlighted.

b) Multi-select functionality allows users to apply multiple filters simultaneously for granular exploration.

2) Visual Consistency

a) A professional color palette of dark blue and teal ensures visual consistency.

b) Charts, graphs, and KPIs (Key Performance Indicators) maintain a clean and structured layout, making insights easy to interpret.

3) Interactive Visualizations

a) Tree Map: Displays the distribution of room types, offering a quick overview of accommodation trends.

b) Scatter Plot: Highlights the relationship between pricing and guest ratings, with tooltips providing additional data points.

c) Bar Charts: Show the frequency of amenities like WiFi, Parking, and Breakfast, offering insights into service availability.

d) Funnel Chart: Visualizes the distribution of hotels across different star ratings.

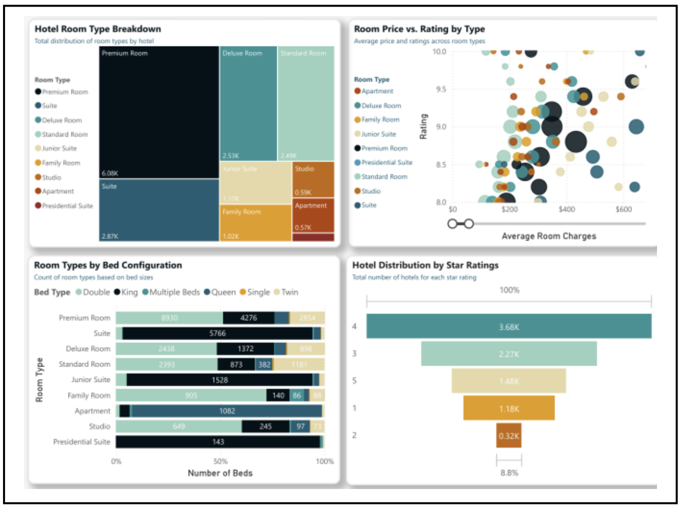


Fig. 5. Visual Charts and Themes.

4) Drill-Through Functionality

a) Users can click on specific data points (e.g., room types or cities) to access detailed views, such as hotel-level data or pricing trends.

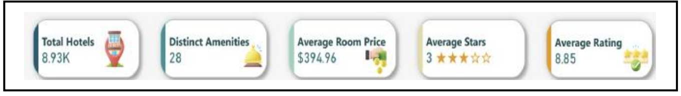
5) Enhancing User Interaction

a) Hover Tooltips provide detailed information about specific data points without overwhelming the interface.

b) For example, hovering over a bar in the service chart reveals the exact number of hotels offering the selected amenity.

6) KPI Cards

a) High-level metrics, such as the total number of hotels, average room charges, and mean guest ratings, are displayed prominently on the dashboard.



Total Hotels
8.93K

Distinct Amenities
28

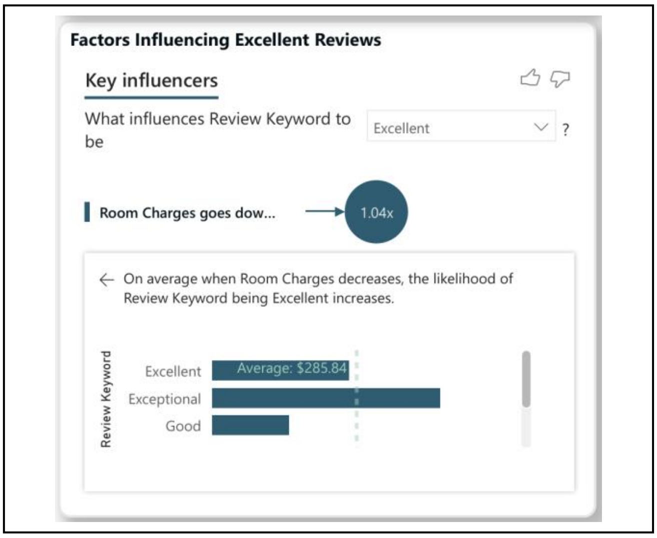
Average Room Price
$394.96

Average Stars
3 ★★★☆☆

Average Rating
8.85

Fig. 6. KPIs with relevant icons.

b) A KPI measure to understand what factors influence the reviews and rating of a hotel.



Factors Influencing Excellent Reviews

Key influencers

What influences Review Keyword to be: Excellent ?

Room Charges goes dow... 1.04x

On average when Room Charges decreases, the likelihood of Review Keyword being Excellent increases.

Fig. 7. Key influencers.

7) Navigation

a) The layout is designed for seamless navigation, with clear labels and logical groupings of visuals to minimize cognitive load.

b) Users can easily switch between summary-level insights and detailed views using slicers and drill-through links.

VII. Challenges and Solutions

A. Data Extraction

1) Dynamic Content Loading: Websites like Expedia utilize dynamic content loading, which delayed the scraping process.
Solution: Selenium's explicit wait feature was employed to ensure all elements were fully rendered before extraction.

2) HTML Structure Variations: Variations in the HTML structure across different city pages complicated data extraction.
Solution: XPath locators were dynamically adjusted to accommodate these differences.

3) Rate Limiting: Frequent requests risked blocking by website.
Solution: Delays (time.sleep) were introduced, and scraping was paced to prevent triggering rate limits.

B. Data Preparation

1) Inconsistent Formats: Attributes like services and pricing were inconsistently formatted in the raw data. Solution: Data was standardized using automated cleaning scripts and Power BI's transformation tools.

2) Handling Missing Values: Missing or incomplete records, such as star ratings and reviews, could skew analysis. Solution: Missing values were replaced with calculated averages or placeholders to maintain dataset integrity.

C. Data Analysis and Modeling

1) Complex Relationships: Establishing meaningful relationships between tables in the star schema required caution.
Solution: Dimension tables were meticulously created to ensure smooth integration and avoid redundancy.

2) Recommendation Model Interpretability: More complex models like neural networks were challenging to interpret and justify for business use. Solution: The TF-IDF + Cosine Similarity approach was chosen for its simplicity and transparency.

D. Power BI Dashboard

1) Visual Clutter: Balancing the inclusion of detailed insights without overcrowding the dashboard was challenging.
Solution: A clean layout and professional color palette ensured readability and logical grouping of visuals.

2) Interactivity: Providing users with intuitive and dynamic interactions across multiple filters and drill-through options.
Solution: Slicers and hover tooltips were implemented to enhance interactivity without adding complexity.

3) Data Loading Time: The volume of records slowed down data rendering in Power BI.

Solution: Query optimization and pre-processing ensured faster loading times for the dashboard.

E. Results and Insights

1) Generalization of Findings: Insights like price trends and room type distributions varied significantly across locations, making generalization difficult.
Solution: Visuals were categorized by location, allowing users to analyze data for specific cities or regions.

2) Limited Scope for Real-Time Updates: Static data limited the applicability of insights for rapidly changing markets.
Solution: The system was designed to accommodate real-time updates in future iterations by integrating APIs or live data feeds.

VIII. Learning Outcomes

The development of the Hotelytics project provided significant technical, analytical, and organizational learnings. These insights were gained across various phases, from data collection and preparation to advanced modeling and visualization. This section consolidates the key takeaways and their implications for future projects.

A. Technical Skills

1) Web Scraping:

a) Mastery of Python's Selenium library for automating web interactions and extracting structured data from dynamic websites like Expedia.

b) Understanding how to handle challenges such as dynamic content loading, rate limiting, and variations in HTML structure across pages.

c) Effective use of tools like pandas to process and store scraped data in manageable formats for analysis.

2) Data Cleaning and Transformation:

a) Hands-on experience in handling large datasets with missing, inconsistent, or redundant data points.

b) Implementation of standardization techniques, such as unifying service names and categorizing room types, to improve data quality.

c) Familiarity with tools like Power Query in Power BI for creating clean, analysis-ready datasets.

3) Star Schema Design:

a) Learning the principles of relational database design to structure data for analytical purposes.

b) Designing and implementing a star schema with fact and dimension tables to optimize query performance and maintain data integrity.

c) Leveraging schema relationships to simplify complex queries and improve dashboard responsiveness.

4) *AI Recommendation Systems:*

a) Developing a custom recommendation engine using TF-IDF Vectorization and Cosine Similarity, focusing on computational efficiency and interpretability.

b) Exploring alternative techniques such as collaborative filtering, neural networks, and hybrid systems to compare effectiveness and suitability.

c) Gaining insights into balancing simplicity and performance in machine learning models for real-world applications.

B. Analytical Skills

1) Data Visualization:

a) Building intuitive and visually appealing dashboards in Power BI, integrating slicers, drill-through functionality, and dynamic visuals.

b) Crafting meaningful visualizations like scatter plots, tree maps, and bar charts to uncover trends in room pricing, guest reviews, and service popularity.

c) Ensuring visual consistency and logical organization to enhance user understanding and engagement.

2) Trend Analysis:

a) Deriving actionable insights into room type distributions, pricing patterns, and geographic service preferences.

b) Identifying critical factors influencing guest ratings, such as pricing value, star ratings, and availability of amenities.

c) Recognizing patterns that inform business strategies, such as the correlation between luxury services and customer satisfaction.

3) Key Performance Indicators (KPIs):

a) Designing and implementing KPI cards to provide high-level summaries of essential metrics, including total hotels, average room charges, and mean guest ratings.

b) Learning to balance summary-level insights with detailed drill-through analyses for comprehensive decision-making.

C. Project Management and Team Building

1) Workflow Planning:

a) Coordinating a multi-phase workflow involving data collection, transformation, analysis, modeling, and visualization.

b) Allocating time effectively to each phase to meet deliverables while ensuring quality outputs.

2) Collaboration and Communication*:*

a) Working as part of a team to integrate contributions from multiple members, ensuring cohesion in the final output.

b) Regularly communicating progress and challenges with supervisors to align expectations and incorporate feedback.

3) Documentation and Presentation:

a) Preparing detailed documentation to support the technical aspects of the project, including scripts, visualizations, and model designs.

b) Presenting findings effectively through structured reports and visually rich dashboards, making insights accessible to both technical and non-technical audiences.

D. Broader Insights

1) Real-World Data Challenges:

a) Recognizing that real-world data is often messy, incomplete, and inconsistent, requiring significant preprocessing before analysis.

b) Developing a systematic approach to address data issues and ensure reliability for downstream tasks.

2) Scalability and Efficiency:

a) Understanding the importance of scalable solutions for handling growing datasets, especially in domains like hospitality with expanding offerings and user preferences.

b) Learning to optimize processes, such as query execution in Power BI and computational efficiency in recommendation models, to enhance performance.

3) User-Centric Design:

a) Focusing on the end-user experience in dashboard and model design, ensuring usability, interpretability, and engagement.

b) Prioritizing dynamic filters, intuitive navigation, and clear labeling to cater to diverse user needs.

IX. Future Work

The Hotelytics project successfully established a foundation for data-driven analytics and AI-powered recommendations in the hospitality sector. However, several enhancements can further improve its functionality, scalability, and impact.

A. Enhanced Recommendation System

1) Hybrid Models: Combine collaborative and content-based filtering to improve recommendation accuracy by leveraging both user behavior and hotel attributes.

2) Deep Learning Integration: Experiment with neural networks, such as embeddings or recurrent neural networks (RNNs), to capture complex patterns in preferences.

3) User Feedback Loop: Incorporate user feedback to refine recommendations based on satisfaction and interactions.

B. Advanced Data Analysis

1) Sentiment Analysis: Use NLP techniques to analyze guest reviews and derive customer satisfaction scores.

2) Predictive Analytics: Predict trends like price fluctuations and occupancy rates using machine learning.

3) Geographic Patterns: Include location-based insights, such as proximity to attractions or hubs.

C. Expanded Dataset and Scope

1) Broader Data Coverage: Extend to more cities, countries, and hotel chains for diverse market applicability.

2) Seasonality Analysis: Incorporate temporal dimensions to analyze seasonal pricing and demand trends.

D. Scalability and Performance

1) Cloud Deployment: Host on platforms like AWS or Azure to handle larger datasets efficiently.

2) Query Optimization: Refine Power BI queries and database structures for reduced loading times.

E. Integration with Emerging Technologies

1) IoT Data: Include real-time metrics like room occupancy using IoT sensors.

2) Blockchain for Authenticity: Validate reviews and ratings using blockchain technology.

3) Voice and Chat Interfaces: Develop conversational interfaces for insights.

These enhancements will elevate the Hotelytics system's capabilities, ensuring its relevance and scalability in an evolving industry.

X. Conclusion

The Hotelytics project demonstrated the power of integrating data analytics, visualization, and AI to address challenges in the hospitality sector. By combining web scraping, data cleaning, a star schema design, and an AI recommendation engine, the system provides valuable insights and personalized hotel suggestions.

The Power BI dashboard enhances decision-making through interactive visualizations and dynamic filters, while the recommendation model balances efficiency and interpretability. Though the project achieved its objectives, future improvements, such as real-time data integration, advanced analytics, and scalability enhancements, can further refine the system.

Hotelytics serves as a robust foundation for innovation in hotel analytics, with the potential to adapt to evolving needs and set new benchmarks in the industry.

Acknowledgments

The authors would like to thank Dr. Imran Amin and Sir Muhammad Ahsan Nisar for their guidance throughout the Hotelytics project. The team also acknowledges Expedia for providing access to publicly available data, which formed the basis for this analysis. Their contributions and resources were invaluable to the success of this work..

References

[1] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1986. Available online

[2] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*. Cambridge: Cambridge University Press, 2011, pp. 11–30. Available online

[3] F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*, 2nd ed. Free sample online

[4] T. K. Dasu and T. Johnson, *Exploratory Data Mining and Data Cleaning*. Hoboken, NJ: Wiley-Interscience, 2003. Available online

[5] P. Gupta and S. Gupta, “A robust TF-IDF based approach for similarity computation in large datasets,” *International Journal of Data Science and Analytics*, vol. 7, no. 2, pp. 121–134, Apr. 2018. Available online