Data-Driven Hotel Analytics and Recommendations

Fall'24

CSC4818 Data Sciences

A Project by

Ghazal E Ashar (2112143)

Shahzeb Ahmed Iqbal (2112166)

Mohammad Ali Hassan (2112148)

Muhammad Ammar Thahim (2112247)

Usama Habib (2112168)

Submitted to Dr Imran Amin & Sir Muhammad Ahsan Nisar

Abstract

The hospitality industry generates vast amounts of data across various dimensions, including hotel ratings, room pricing, amenities, and customer reviews. **Hotelytics** is a comprehensive data-driven solution designed to analyze and visualize hotel data to uncover actionable insights and provide AI-powered recommendations. This project begins with the **extraction** of hotel data from Expedia for **53 cities** across **36 countries**. The extracted data includes critical attributes such as hotel names, stars, ratings, room charges, services, and room details.

The data undergoes extensive **cleaning and transformation** using Power BI, including handling missing values, standardizing services, and building a **star schema** for optimal performance. Interactive Power BI visualizations, such as tree maps, bubble charts, and bar charts, are utilized to analyze trends, including room type distribution, pricing, and service popularity. A custom **AI-based recommendation model** using **TF-IDF Vectorization** and **Cosine Similarity** dynamically suggests the top hotels based on user-selected filters like city, price range, stars, and amenities.

The final output is an interactive Power BI dashboard that combines advanced analytics and machine learning to support decision-making in the hospitality sector. The project demonstrates the integration of **web scraping**, **data transformation**, **visualization**, and **AI modeling** to deliver a scalable and intelligent hotel analytics solution.

Table of Contents

Introduction

The hospitality industry thrives on customer satisfaction, pricing competitiveness, and personalized services. To stay ahead in an evolving market, data-driven insights are crucial for understanding trends, optimizing offerings, and recommending suitable hotels based on diverse criteria. **Hotelytics** is a hotel analytics and recommendation solution that leverages extracted hotel data, cleans it using advanced transformations, visualizes key metrics, and provides AI-powered recommendations.

This report explores the end-to-end process of data extraction, cleaning, analysis, visualization, and integration of a custom-built recommendation model within **Power BI**. The dataset, spanning **53 cities** across **36 countries**, forms the foundation for understanding room types, pricing categories, reviews, and service popularity.

Objectives

The primary objectives of **Hotelytics** include:

- Extracting hotel-related data from diverse sources.
- Cleaning and transforming data for analytical consistency.
- Visualizing trends such as pricing, services, and ratings.
- Developing an AI recommendation system based on user filters and preferences.
- Providing insights through an interactive Power BI dashboard for better decision-making.

Data Extraction

The data was extracted using **web scraping** techniques from **Expedia**, a hotel booking platform. Scraped data included:

- Hotel Names
- Country and City
- Star Ratings
- Review Keywords
- Guest Rating
- Services

- Property Offers

Individual Datasets

Initially, hotel data was scraped for multiple cities individually. Each dataset varied in size and format. For example:

- **London, UK Dataset:** Contained **250+ records**.
- **Istanbul, Turkey Dataset:** Included detailed room types and pricing.

Combined Dataset

The individual datasets were combined into a single dataset named **combined_hotel_data.csv**, resulting in **8,900+ rows**. During this process, duplicates were carefully removed while preserving accurate records.

Data Cleaning

The raw data underwent several transformations and standardization processes to ensure accuracy and consistency.

Standardization of Services

	Hotel ID	Services
1	1	['Breakfast available', 'Pool', 'Bar', 'Gym', 'Room service', 'Laundry']
2	2	['Pool', 'On private beach', 'Breakfast available', 'Bar', 'Spa', 'Air conditioning']
3	3	['On private beach', 'Pool', 'Breakfast available', 'Bar', 'Spa', 'Room service']
4	4	['Breakfast available', 'On private beach', 'Pool', 'Bar', 'Hot Tub', 'Spa']
5	5	['Pool', 'Spa', 'Gym', 'Laundry', 'Housekeeping', 'Air conditioning']
6	6	['On private beach', 'Pool', 'Bar', 'Spa', 'Room service', 'Housekeeping']
7	7	['Breakfast available', 'Bar', 'Pool', 'Parking available', 'Room service', 'Gym']
8	8	['Restaurant', 'Hot Tub', 'Pool', 'Bar', 'Spa', 'Room service']
9	9	['Hot Tub', 'Pool', 'Bar', 'All inclusive', 'Spa', 'Gym']
10	10	['Bar', 'Pool', 'Parking available', 'Spa', 'Gym', 'Room service']
11	11	['Breakfast available', 'Pool', 'Bar', 'Spa', 'Room service', 'Housekeeping']

- Services were normalized to a common format (e.g., “Parking Available” → “Parking”).
- A custom column for **Services** replaced redundant terms.

	Hotel ID	Services
1	1	Breakfast available, Pool, Bar, Gym, Room service, Laundry
2	2	Pool, On private beach, Breakfast available, Bar, Spa, Air conditioning
3	3	On private beach, Pool, Breakfast available, Bar, Spa, Room service
4	4	Breakfast available, On private beach, Pool, Bar, Hot Tub, Spa
5	5	Pool, Spa, Gym, Laundry, Housekeeping, Air conditioning
6	6	On private beach, Pool, Bar, Spa, Room service, Housekeeping
7	7	Breakfast available, Bar, Pool, Parking available, Room service, Gym
8	8	Restaurant, Hot Tub, Pool, Bar, Spa, Room service
9	9	Hot Tub, Pool, Bar, All inclusive, Spa, Gym
10	10	Bar, Pool, Parking available, Spa, Gym, Room service
11	11	Breakfast available, Pool, Bar, Spa, Room service, Housekeeping
12	12	Pool, Parking available, Spa, Breakfast available, Gym, Room service

PowerQuery:

= Table.AddColumn(#"Filtered Rows", "Service", each if Text.Contains(Text.Lower([Value]), "parking") then "Parking"

else if Text.Contains(Text.Lower([Value]), "breakfast") then "Breakfast"

else if Text.Contains(Text.Lower([Value]), "gym") then "Gym"

else if Text.Contains(Text.Lower([Value]), "pool") then "Pool"

else if Text.Contains(Text.Lower([Value]), "bar") then "Bar"

else if Text.Contains(Text.Lower([Value]), "room service") then "Room Service"

else if Text.Contains(Text.Lower([Value]), "laundry") then "Laundry"

else if Text.Contains(Text.Lower([Value]), "private beach") then "Private Beach"

else if Text.Contains(Text.Lower([Value]), "spa") then "Spa"

else if Text.Contains(Text.Lower([Value]), "air conditioning") then "Air Conditioning"

else if Text.Contains(Text.Lower([Value]), "hot tub") then "Hot Tub"

else if Text.Contains(Text.Lower([Value]), "housekeeping") then "Housekeeping"

else if Text.Contains(Text.Lower([Value]), "restaurant") then "Restaurant"

else if Text.Contains(Text.Lower([Value]), "all inclusive") then "All Inclusive"

else if Text.Contains(Text.Lower([Value]), "airport shuttle") then "Airport Shuttle"

else if Text.Contains(Text.Lower([Value]), "kitchen") then "Kitchen"

else if Text.Contains(Text.Lower([Value]), "front desk") then "Front Desk"

else if Text.Contains(Text.Lower([Value]), "pet friendly") then "Pet Friendly"

else if Text.Contains(Text.Lower([Value]), "wifi") then "WiFi"

else if Text.Contains(Text.Lower([Value]), "outdoor space") then "Outdoor Space"

else if Text.Contains(Text.Lower([Value]), "business services") then "Business Services"

else if Text.Contains(Text.Lower([Value]), "barbecue") then "Barbecue"

else if Text.Contains(Text.Lower([Value]), "washer") then "Washer"

else if Text.Contains(Text.Lower([Value]), "dryer") then "Dryer"

else if Text.Contains(Text.Lower([Value]), "city view") then "City View"

else if Text.Contains(Text.Lower([Value]), "fireplace") then "Fireplace"

else if Text.Contains(Text.Lower([Value]), "beach view") then "Beach View"

else if Text.Contains(Text.Lower([Value]), "lake view") then "Lake View"

else if Text.Contains(Text.Lower([Value]), "ocean view") then "Ocean View"

else if Text.Contains(Text.Lower([Value]), "all inclusive") then "All Inclusive"

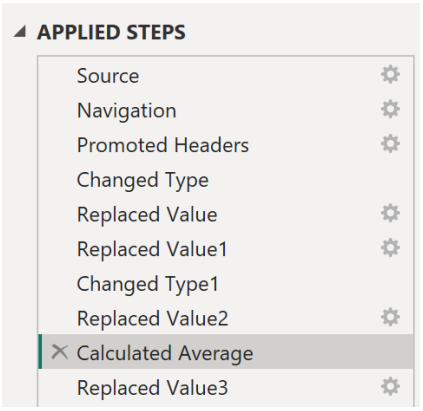
else "Other")

The result of the PowerQuery was as shown below:

	Hotel ID	Service
1	1	Breakfast
2	1	Pool
3	1	Bar
4	1	Gym
5	1	Room Service
6	1	Laundry
7	2	Pool
8	2	Private Beach
9	2	Breakfast
10	2	Bar
11	2	Spa
12	2	Air Conditioning
13	3	Private Beach
14	3	Pool
15	3	Breakfast
16	3	Bar
17	3	Spa
18	3	Room Service

Handling Missing Values

- **Stars**: Missing star ratings were assigned **1 Star**.
- **Ratings**: Null ratings were replaced with the **average rating** of the dataset.
- **Review Keywords**: Empty reviews were filled with **“No Reviews”**.



Cleaning Property Offers

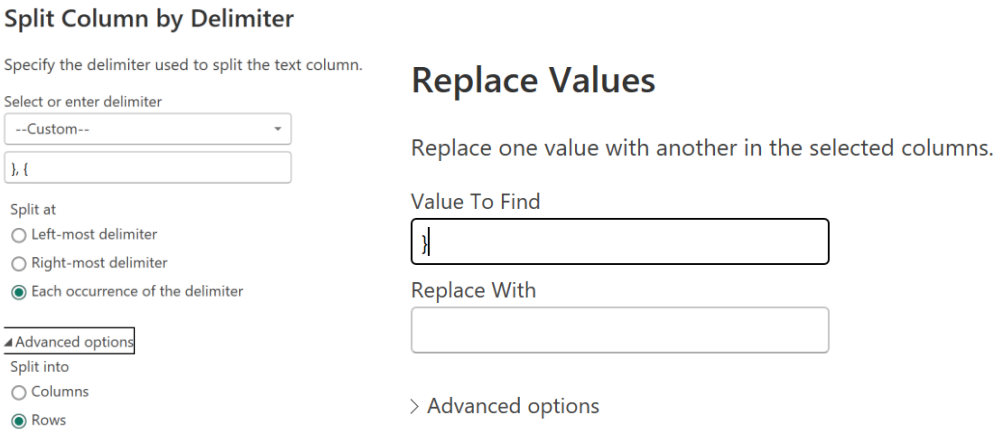
The **Property Offers** column contained combined details for **Room Name** and **Room Charges**.

	otel ID	Property Offers
1	1	{'Room Name': 'Deluxe Double Room, Non Smoking - 15% Discount on F&B', 'Room Charges': '$109'}, {'Room Name': 'Deluxe Double Room, 1 Ki...
2	2	{'Room Name': 'Room, 1 King Bed (Guest)', 'Room Charges': '$166'}, {'Room Name': 'Room, Kitchenette (Guest, Balcony)', 'Room Charges': '$17...
3	3	{'Room Name': 'Room, 1 King Bed, Balcony, Sea View', 'Room Charges': '$185'}, {'Room Name': 'Room, 2 Twin Beds, Balcony, Sea View', 'Room...
4	4	{'Room Name': 'Superior Room King - Free daily shuttle bus to Yas attractions, Yas Mall and Grand Mosque', 'Room Charges': '$190'}, {'Room Na...
5	5	{'Room Name': 'Deluxe Twin Room, Multiple Beds, Smoking', 'Room Charges': '$69'}, {'Room Name': 'Deluxe Suite, 1 King Bed, Smoking, 2 Bathr...
6	6	{'Room Name': 'Studio, Balcony, Sea View', 'Room Charges': '$190'}, {'Room Name': 'Apartment, 1 Bedroom, Sea View', 'Room Charges': '$323'},...
7	7	{'Room Name': 'Room, 2 Twin Beds, City View (Centro)', 'Room Charges': '$92'}, {'Room Name': 'Room, 2 Twin Beds (Centro, Stadium View)', 'Ro...
8	8	{'Room Name': 'Room, 1 King Bed', 'Room Charges': '$160'}, {'Room Name': 'Deluxe Room, 1 King Bed', 'Room Charges': '$173'}, {'Room Name':...
9	9	{'Room Name': 'Deluxe King City View', 'Room Charges': '$78'}, {'Room Name': 'DELUXE TWIN CITY VIEW', 'Room Charges': '$78'}, {'Room Name'...
10	10	{'Room Name': 'Standard Double Room', 'Room Charges': '$68'}, {'Room Name': 'Standard Twin Room', 'Room Charges': '$76'}, {'Room Name': '...
11	11	{'Room Name': 'Executive One Bedroom Apartment', 'Room Charges': '$156'}, {'Room Name': 'Junior Suite with Lounge Access', 'Room Charges'...
12	12	{'Room Name': 'Room, 1 King Bed (Guest)', 'Room Charges': '$107'}, {'Room Name': 'Room, 2 Twin Beds (Guest)', 'Room Charges': '$107'}, {'Roo...
13	13	{'Room Name': 'Superior Room, 2 Twin Beds', 'Room Charges': '$196'}, {'Room Name': 'Deluxe Room, 1 King Bed (Palace View)', 'Room Charges'...
14	14	{'Room Name': 'Room, 1 King Bed (Guest - Lounge Access)', 'Room Charges': '$175'}, {'Room Name': 'Suite, 1 King Bed (Lounge Access)', 'Room...
15	15	{'Room Name': 'Luxury Room, 1 King Bed, Sea View', 'Room Charges': '$215'}, {'Room Name': 'Luxury Club King Room with Club Access and Sea...
16	16	{'Room Name': 'Quartz Suite - 20% Off Spa', 'Room Charges': '$132'}, {'Room Name': 'Emerald Suite - 20% Off Spa', 'Room Charges': '$152'}, {'Ro...
17	17	{'Room Name': 'Double Room, Non Smoking (FREE Aprt Shuttle|Yas Island drop off)', 'Room Charges': '$119'}, {'Room Name': 'Family Room, No...
18	18	{'Room Name': 'Deluxe Room, 1 King Bed', 'Room Charges': '$425'}, {'Room Name': 'Deluxe Room, 1 King Bed (Waterfront)', 'Room Charges': '$4...
19	19	{'Room Name': 'Superior Room, 1 Queen Bed, Non Smoking', 'Room Charges': '$93'}, {'Room Name': 'Superior Room, 1 Queen Bed, Smoking', 'R...
20	20	{'Room Name': 'Studio (Twin)', 'Room Charges': '$88'}, {'Room Name': 'Studio (King)', 'Room Charges': '$92'}, {'Room Name': 'Deluxe Apartment,...
21	21	{'Room Name': 'Deluxe Room, 1 King Bed (Corniche view)', 'Room Charges': '$274'}, {'Room Name': 'Deluxe Room, 1 King Bed, Sea View', 'Room...
22		

{'Room Name': 'Deluxe Double Room, Non Smoking - 15% Discount on F&B', 'Room Charges': '$109'}, {'Room Name': 'Deluxe Double Room, 1 King Bed - 15% Discount on F&B', 'Room Charges': '$109'}, {'Room Name': 'Deluxe Double Room, Smoking - 15% Discount on F&B', 'Room Charges': '$109'}, {'Room Name': 'Deluxe Room, Accessible, Smoking - 15% Discount on F&B', 'Room Charges': '$116'}, {'Room Name': 'Deluxe Room, Accessible, Non Smoking - 15% Discount on F&B', 'Room Charges': '$116'}, {'Room Name': 'Executive Room, 1 King Bed, Smoking - 15% Discount on F&B', 'Room Charges': '$123'}, {'Room Name': 'Executive Double Room, 1 King Bed, Non Smoking - 15% Discount on F&B', 'Room Charges': '$123'}, {'Room Name': 'Comfort Suite, 1 Bedroom, Non

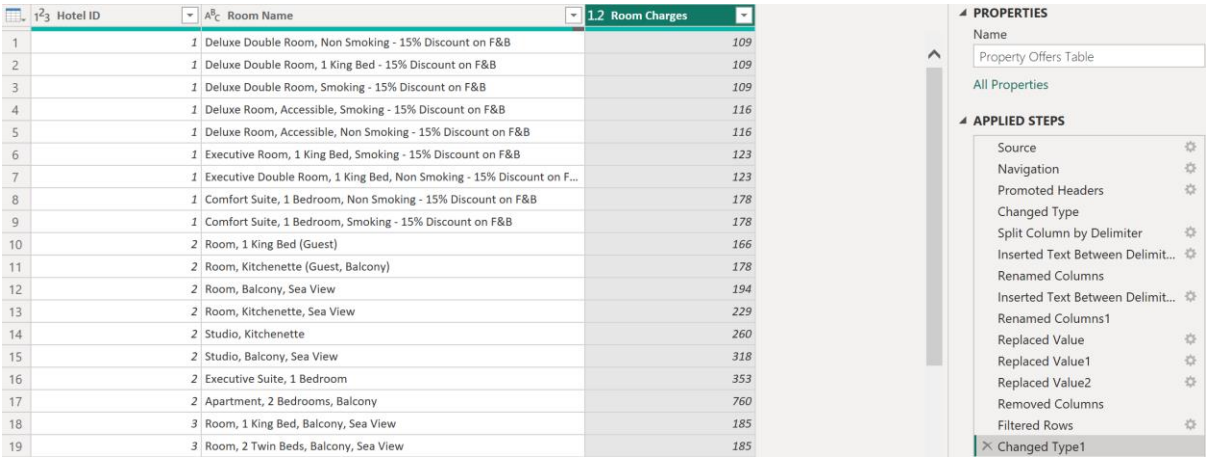
To clean and extract this data:

- **Splitting**: The column was split by delimiters such as }, { into individual room records.



- **Extracting Room Name and Charges**:
  - Using Power Query, **Room Name** and **Room Charges** were extracted from each split record using **Text Between Delimiters**.

- Charges were cleaned by removing symbols like $ and commas, then converted into numerical values.



	Hotel ID	Room Name	Room Charges
1	1	Deluxe Double Room, Non Smoking - 15% Discount on F&B	109
2	1	Deluxe Double Room, 1 King Bed - 15% Discount on F&B	109
3	1	Deluxe Double Room, Smoking - 15% Discount on F&B	109
4	1	Deluxe Room, Accessible, Smoking - 15% Discount on F&B	116
5	1	Deluxe Room, Accessible, Non Smoking - 15% Discount on F&B	116
6	1	Executive Room, 1 King Bed, Smoking - 15% Discount on F&B	123
7	1	Executive Double Room, 1 King Bed, Non Smoking - 15% Discount on F...	123
8	1	Comfort Suite, 1 Bedroom, Non Smoking - 15% Discount on F&B	178
9	1	Comfort Suite, 1 Bedroom, Smoking - 15% Discount on F&B	178
10	2	Room, 1 King Bed (Guest)	166
11	2	Room, Kitchenette (Guest, Balcony)	178
12	2	Room, Balcony, Sea View	194
13	2	Room, Kitchenette, Sea View	229
14	2	Studio, Kitchenette	260
15	2	Studio, Balcony, Sea View	318
16	2	Executive Suite, 1 Bedroom	353
17	2	Apartment, 2 Bedrooms, Balcony	760
18	3	Room, 1 King Bed, Balcony, Sea View	185
19	3	Room, 2 Twin Beds, Balcony, Sea View	185

Extracting Bed Type, Room Type, View Type, and Smoking Status

From the cleaned **Room Name** column:

- **Room Type**: Extracted and categorized as Deluxe Room, Suite, Apartment, Standard Room, etc., using **Text.Contains** conditions.

  if Text.Contains([Room Name], "Suite", Comparer.OrdinalIgnoreCase) then

  if Text.Contains([Room Name], "Junior", Comparer.OrdinalIgnoreCase) then "Junior Suite"

  else if Text.Contains([Room Name], "Presidential", Comparer.OrdinalIgnoreCase) then "Presidential Suite"

  else "Suite"

  else if Text.Contains([Room Name], "Apartment", Comparer.OrdinalIgnoreCase) then "Apartment"

  else if Text.Contains([Room Name], "Deluxe", Comparer.OrdinalIgnoreCase) then "Deluxe Room"

  else if Text.Contains([Room Name], "Standard", Comparer.OrdinalIgnoreCase) then "Standard Room"

  else if Text.Contains([Room Name], "Studio", Comparer.OrdinalIgnoreCase) then "Studio"

  else if Text.Contains([Room Name], "Family", Comparer.OrdinalIgnoreCase) or

  Text.Contains([Room Name], "Multiple Beds", Comparer.OrdinalIgnoreCase) then "Family Room"

else "Premium Room"

- **Bed Type**: Extracted based on keywords like King, Queen, Twin, Single, and Multiple Beds.

  if Text.Contains([Room Name], "King", Comparer.OrdinalIgnoreCase) then "King"

  else if Text.Contains([Room Name], "Queen", Comparer.OrdinalIgnoreCase) then "Queen"

  else if Text.Contains([Room Name], "Twin", Comparer.OrdinalIgnoreCase) then "Twin"

  else if Text.Contains([Room Name], "Single", Comparer.OrdinalIgnoreCase) then "Single"

  else if Text.Contains([Room Name], "Double", Comparer.OrdinalIgnoreCase) then "Double"

  else if Text.Contains([Room Name], "Multiple Beds", Comparer.OrdinalIgnoreCase) then "Multiple Beds"

  else if Text.Contains([Room Name], "Suite", Comparer.OrdinalIgnoreCase) then "King"

  else if Text.Contains([Room Name], "Apartment", Comparer.OrdinalIgnoreCase) then "Queen"

  else "Double"

- **View Type**: Derived by detecting words like Balcony, Sea View, Garden View, and City View.

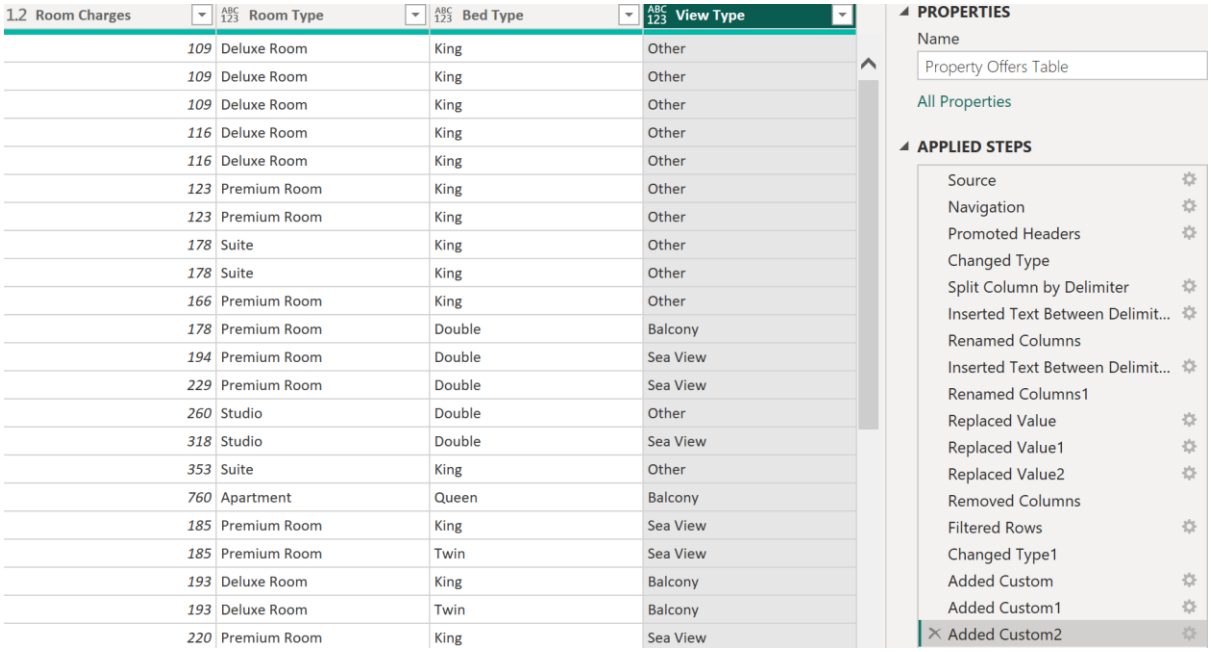
  if Text.Contains([Room Name], "Sea View", Comparer.OrdinalIgnoreCase) then "Sea View"

  else if Text.Contains([Room Name], "City View", Comparer.OrdinalIgnoreCase) then "City View"

  else if Text.Contains([Room Name], "Garden View", Comparer.OrdinalIgnoreCase) then "Garden View"

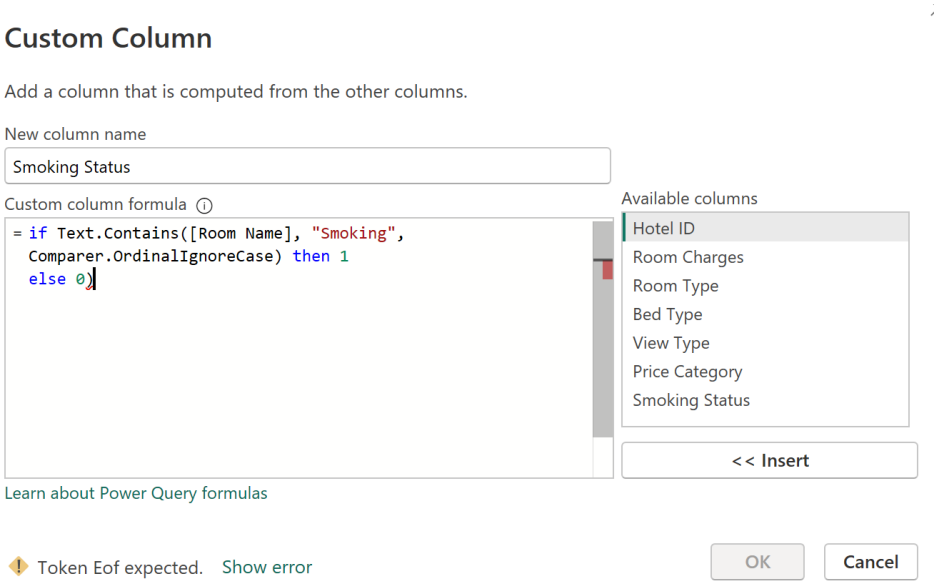
  else if Text.Contains([Room Name], "Balcony", Comparer.OrdinalIgnoreCase) then "Balcony"

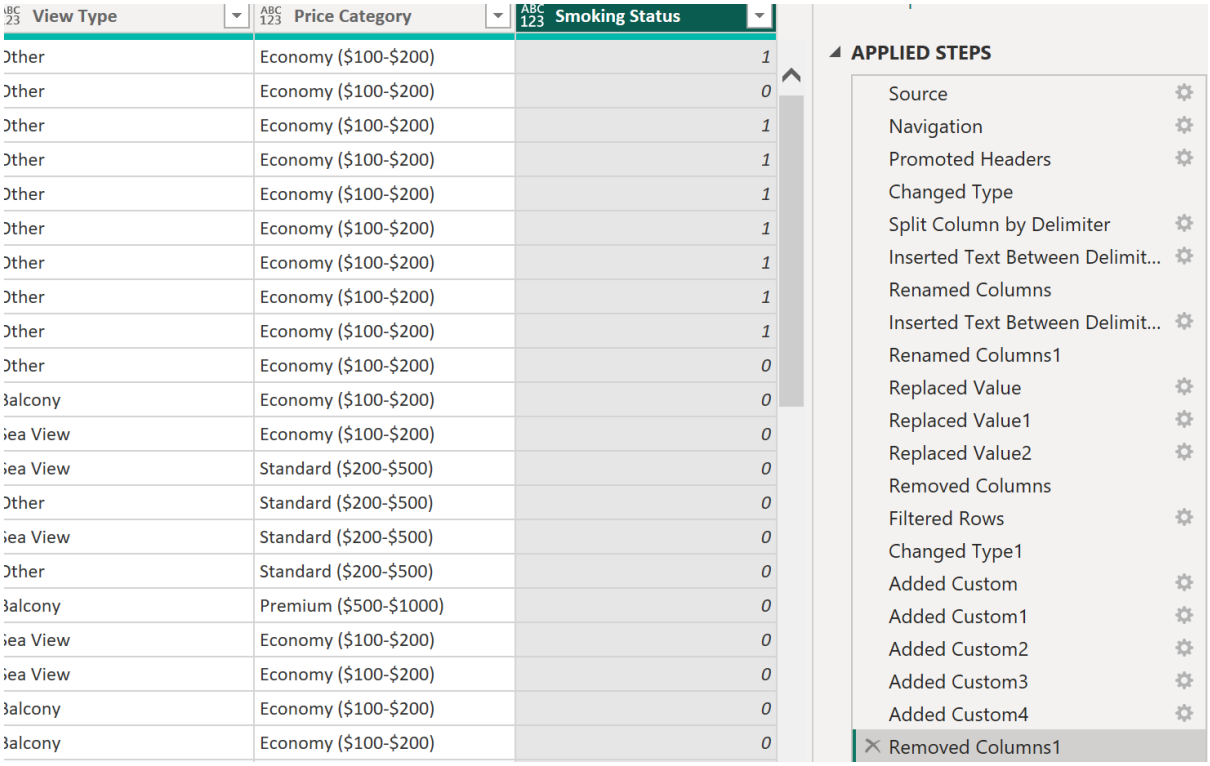
  else "Other"



1.2 Room Charges	Room Type	Bed Type	View Type
109	Deluxe Room	King	Other
109	Deluxe Room	King	Other
109	Deluxe Room	King	Other
116	Deluxe Room	King	Other
116	Deluxe Room	King	Other
123	Premium Room	King	Other
123	Premium Room	King	Other
178	Suite	King	Other
178	Suite	King	Other
166	Premium Room	King	Other
178	Premium Room	Double	Balcony
194	Premium Room	Double	Sea View
229	Premium Room	Double	Sea View
260	Studio	Double	Other
318	Studio	Double	Sea View
353	Suite	King	Other
760	Apartment	Queen	Balcony
185	Premium Room	King	Sea View
185	Premium Room	Twin	Sea View
193	Deluxe Room	King	Balcony
193	Deluxe Room	Twin	Balcony
220	Premium Room	King	Sea View

- **Smoking Status**: Classified into **Smoking** or **Non-Smoking** using keyword matches.

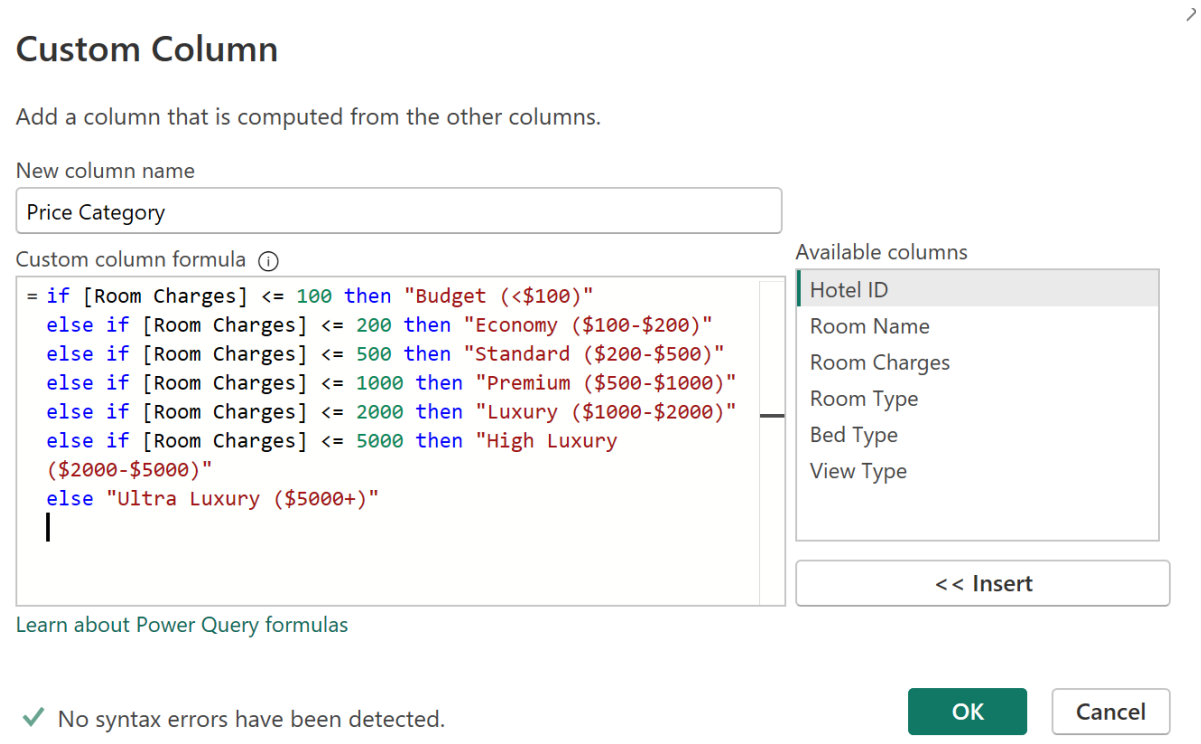




View Type	Price Category	Smoking Status
Other	Economy ($100-$200)	1
Other	Economy ($100-$200)	0
Other	Economy ($100-$200)	1
Other	Economy ($100-$200)	1
Other	Economy ($100-$200)	1
Other	Economy ($100-$200)	1
Other	Economy ($100-$200)	1
Other	Economy ($100-$200)	1
Other	Economy ($100-$200)	1
Other	Economy ($100-$200)	0
Balcony	Economy ($100-$200)	0
Sea View	Economy ($100-$200)	0
Sea View	Standard ($200-$500)	0
Other	Standard ($200-$500)	0
Sea View	Standard ($200-$500)	0
Other	Standard ($200-$500)	0
Balcony	Premium ($500-$1000)	0
Sea View	Economy ($100-$200)	0
Sea View	Economy ($100-$200)	0
Balcony	Economy ($100-$200)	0
Balcony	Economy ($100-$200)	0

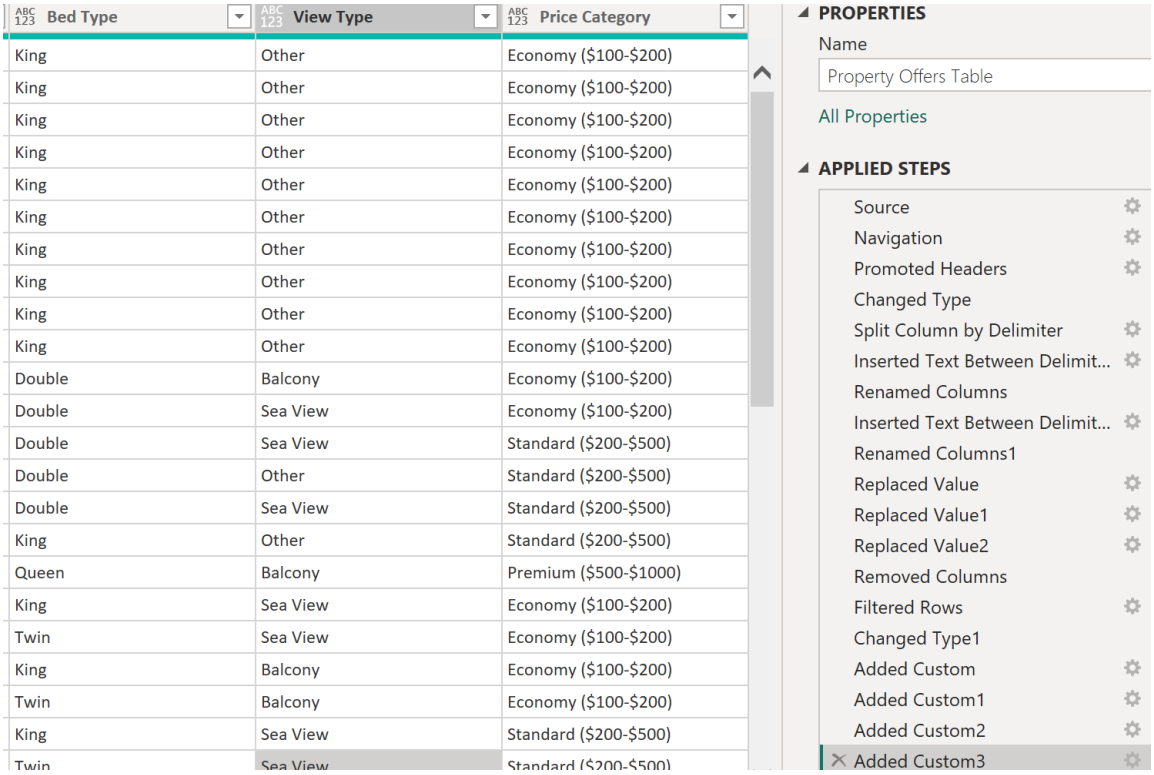
Price Cleaning and Categorization

Room charges were standardized by removing special characters like "$" and commas.
A new column **Price Category** was introduced to classify prices into ranges:



- Budget (<$100)
- Economy ($100–$200)

- Standard ($200–$500)
- Luxury ($500–$1000)



Bed Type	View Type	Price Category
King	Other	Economy ($100-$200)
King	Other	Economy ($100-$200)
King	Other	Economy ($100-$200)
King	Other	Economy ($100-$200)
King	Other	Economy ($100-$200)
King	Other	Economy ($100-$200)
King	Other	Economy ($100-$200)
King	Other	Economy ($100-$200)
King	Other	Economy ($100-$200)
King	Other	Economy ($100-$200)
Double	Balcony	Economy ($100-$200)
Double	Sea View	Economy ($100-$200)
Double	Sea View	Standard ($200-$500)
Double	Other	Standard ($200-$500)
Double	Sea View	Standard ($200-$500)
King	Other	Standard ($200-$500)
Queen	Balcony	Premium ($500-$1000)
King	Sea View	Economy ($100-$200)
Twin	Sea View	Economy ($100-$200)
King	Balcony	Economy ($100-$200)
Twin	Balcony	Economy ($100-$200)
King	Sea View	Standard ($200-$500)
Twin	Sea View	Standard ($200-$500)

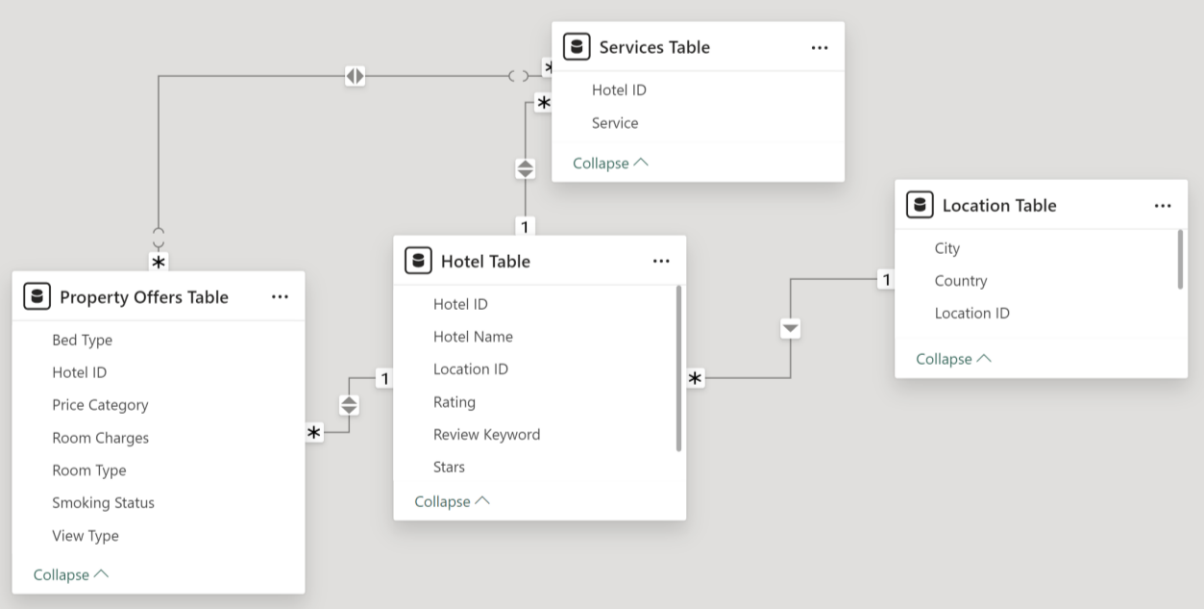
Data Analysis and AI Model

This section combines data analysis through Power BI charts and the implementation of an AI recommendation model.

Star Schema Design

To optimize analytical performance, a **star schema** was created. The tables included:

1. **Location Table**: Contains City, Country, and Location ID.
2. **Hotel Table**: Includes Hotel Name, Stars, Ratings, Review Keywords, and Location ID.
3. **Services Table**: Maps Hotel ID to available services.
4. **Property Offers Table**: Stores detailed room names and corresponding charges.



A **relationship diagram** linking these tables was established to enable smooth data flow between different components in Power BI.

Visualizations in Power BI

The Power BI dashboard provides insights into:

1. **Room Type Distribution** (Tree Map): Visualizing the count of each room type.
2. **Average Room Price vs. Ratings** (Scatter Plot): Relationship between pricing and guest ratings.
3. **Count of Hotels by Stars** (Funnel Chart): Distribution of hotels by star ratings.
4. **Room Charges by View Type and Price Category** (Stacked Bar): Analyzing pricing trends based on view type.
5. **Services Offered by Hotels** (Bar Chart): Highlighting the most common services such as breakfast, Wi-Fi, and parking.
6. **Key Influencers**: Identifying factors influencing excellent reviews (e.g., lower prices and service availability).

AI Recommendation Model

A **TF-IDF Vectorization** and **Cosine Similarity** model was implemented to recommend hotels dynamically based on filters such as:

- City, Stars, Price Range, Amenities, Room Type, and Smoking Status.

How It Works:

- **TF-IDF (Term Frequency-Inverse Document Frequency)**: Converts combined hotel attributes into numerical vectors.

- **Cosine Similarity**: Measures similarity between user inputs (filters) and hotel attributes.

Output: The model recommends the top **10 hotels** based on user-selected filters or overall dataset analysis if no filters are applied. The recommendations are displayed in a clean table format.

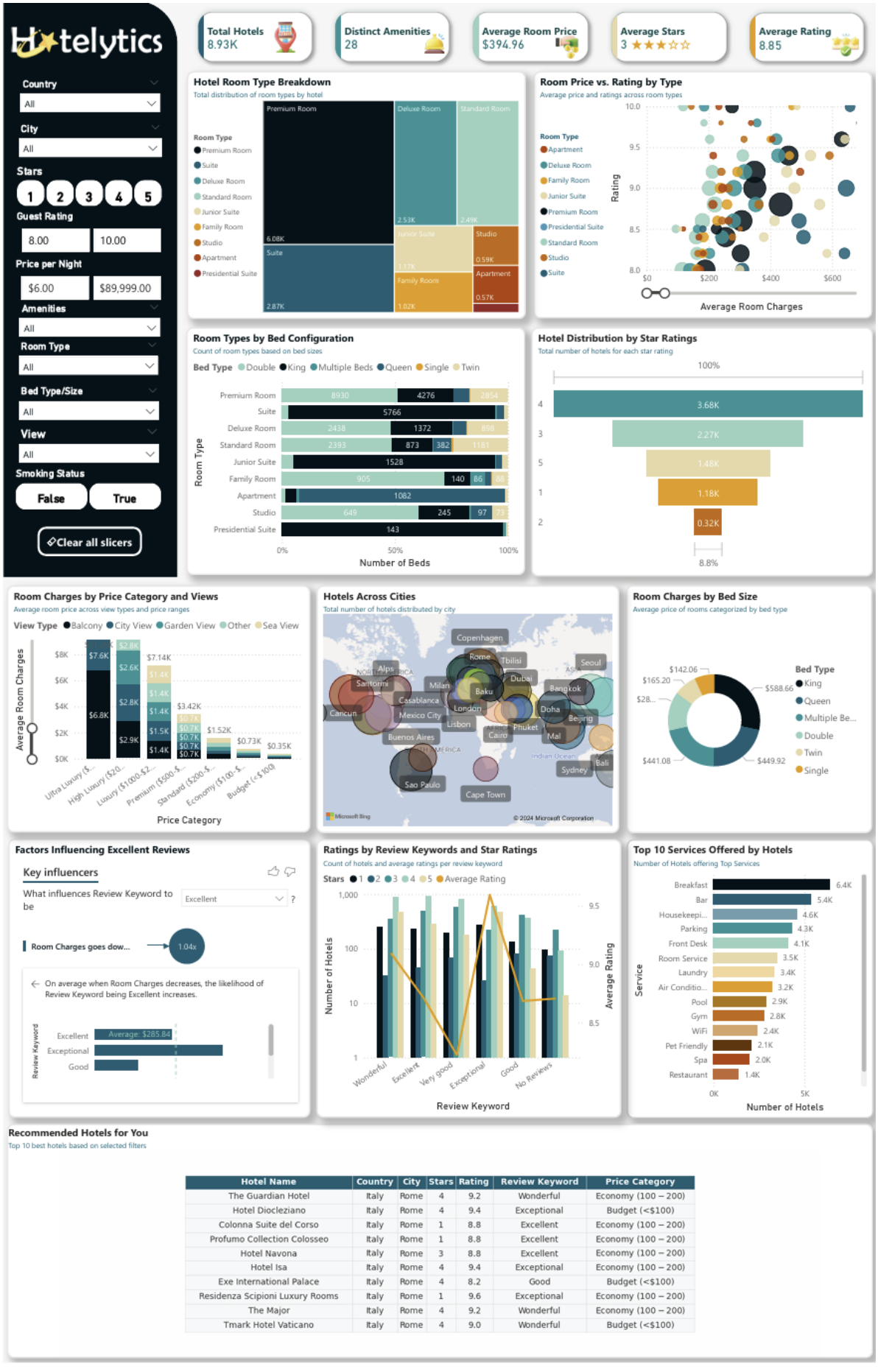
The table below compares TF-IDF with alternative models:

Model	Advantages	Disadvantages	Relevance to Hotelytics
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.

Justification: 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.

Power BI Dashboard

Below is the final dashboard created:



Recommended Hotels for You

Top 10 best hotels based on selected filters

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)

Dynamic Filters

The dashboard includes slicers for all critical dimensions, allowing users to filter data interactively.

- Filters include:
    - **Country** and **City**: Location-based insights.
    - **Star Rating**: Filter data by 1–5 stars.
    - **Guest Ratings**: Choose specific guest review scores (8.0–10.0).
    - **Price Range**: Adjust sliders to include hotel prices within desired budgets.
    - **Amenities**: Filter hotels based on services such as WiFi, Gym, Breakfast, etc.
    - **Room Type, Bed Type, and View**: Narrow down options based on room preferences.
    - **Smoking Status**: Identify smoking vs. non-smoking rooms.

Drill-Through Functionality

Users can click on visuals (e.g., Room Types, Cities) to "drill through" to detailed pages, such as:

- Hotel-level analysis with additional KPIs.
- Detailed room breakdown with price and services.

Insight: Enhances exploration by linking summary visuals to deeper analyses.

Interactive AI Recommendation Table

Powered by the AI model, this table dynamically suggests the **top 10 hotels** based on user selections.

- Highlights include:
    - **Hotel Name, Location, Star Rating, Average Rating, Price Category, and Review Keyword**.
    - Automatically recalculates recommendations when filters are applied.

KPI Cards for High-Level Metrics

Key performance indicators (KPIs) provide a quick snapshot of essential metrics:

- **Total Hotels**: Number of hotels in the dataset.

- **Count of Services**: Total number of services offered across all hotels.
- **Average Room Charges**: Average price per night for rooms.
- **Average Stars**: Mean stas rating across all hotels
- **Average Rating**: Mean guest rating across all hotels.

Visual Consistency and Design

- The dashboard follows a consistent theme, with a professional color palette of **dark blues** and **teal** for headers and text.
- **Hover Tooltips**: Display precise values and additional insights on mouse-over.
- **Clear Titles and Subtitles**: Each chart and visual is accompanied by a short descriptive title and subtitle.

Insights Derived from Visuals

The Power BI dashboard uncovers the following insights:

1. **Room Type Trends**: Premium Rooms and Suites dominate, indicating higher demand for luxury accommodations.
2. **Price vs. Rating**: Higher-rated rooms often fall into the **Economy** and **Luxury** price brackets, suggesting value-for-money trends.
3. **Geographic Patterns**: Cities like **Rome, Bali, and London** have the highest hotel density.
4. **Services Influence Reviews**: Essential services like **Breakfast, WiFi, and Parking** significantly affect customer satisfaction.
5. **Review Sentiment**: Positive reviews are associated with lower room prices and additional amenities.

User Experience and Interactivity

- The dashboard allows users to explore data at various levels, from high-level summaries to detailed views.
- Drill-throughs, filters, and AI recommendations provide a seamless and interactive experience, ensuring actionable insights for users.

Challenges and Solutions

Data Extraction

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

Data Preparation

- **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.
- **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.
- **Outliers**: Room prices included extreme outliers that impacted visualizations.
  Solution: Outliers were flagged and either removed or analyzed separately for specific insights.

Data Analysis and Modeling

- **Complex Relationships**: Establishing meaningful relationships between tables in the star schema required careful design.
  Solution: Dimension tables were meticulously created to ensure smooth integration and avoid redundancy.
- **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.

- **Scalability**: The large dataset posed performance challenges for some analyses.
  Solution: The star schema optimized query performance in Power BI and streamlined model computations.

Power BI Dashboard

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

Results and Insights

- **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.
- **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.

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.

Technical Skills

1. **Web Scraping Expertise**:

- Mastery of Python's Selenium library for automating web interactions and extracting structured data from dynamic websites like Expedia.
- Understanding how to handle challenges such as dynamic content loading, rate limiting, and variations in HTML structure across pages.
- Effective use of tools like pandas to process and store scraped data in manageable formats for analysis.

2. **Data Cleaning and Transformation**:
    - Hands-on experience in handling large datasets with missing, inconsistent, or redundant data points.
    - Implementation of standardization techniques, such as unifying service names and categorizing room types, to improve data quality.
    - Familiarity with tools like Power Query in Power BI for creating clean, analysis-ready datasets.
3. **Star Schema Design**:
    - Learning the principles of relational database design to structure data for analytical purposes.
    - Designing and implementing a star schema with fact and dimension tables to optimize query performance and maintain data integrity.
    - Leveraging schema relationships to simplify complex queries and improve dashboard responsiveness.
4. **AI Recommendation Systems**:
    - Developing a custom recommendation engine using TF-IDF Vectorization and Cosine Similarity, focusing on computational efficiency and interpretability.
    - Exploring alternative techniques such as collaborative filtering, neural networks, and hybrid systems to compare effectiveness and suitability.
    - Gaining insights into balancing simplicity and performance in machine learning models for real-world applications.

Analytical Proficiency

1. **Data Visualization**:
    - Building intuitive and visually appealing dashboards in Power BI, integrating slicers, drill-through functionality, and dynamic visuals.

- Crafting meaningful visualizations like scatter plots, tree maps, and bar charts to uncover trends in room pricing, guest reviews, and service popularity.
- Ensuring visual consistency and logical organization to enhance user understanding and engagement.

2. **Trend Analysis**:
    - Deriving actionable insights into room type distributions, pricing patterns, and geographic service preferences.
    - Identifying critical factors influencing guest ratings, such as pricing value, star ratings, and availability of amenities.
    - Recognizing patterns that inform business strategies, such as the correlation between luxury services and customer satisfaction.
3. **Key Performance Indicators (KPIs)**:
    - Designing and implementing KPI cards to provide high-level summaries of essential metrics, including total hotels, average room charges, and mean guest ratings.
    - Learning to balance summary-level insights with detailed drill-through analyses for comprehensive decision-making.

Project Management

1. **Workflow Planning**:
    - Coordinating a multi-phase workflow involving data collection, transformation, analysis, modeling, and visualization.
    - Allocating time effectively to each phase to meet deliverables while ensuring quality outputs.
2. **Collaboration and Communication**:
    - Working as part of a team to integrate contributions from multiple members, ensuring cohesion in the final output.
    - Regularly communicating progress and challenges with supervisors to align expectations and incorporate feedback.
3. **Documentation and Presentation**:
    - Preparing detailed documentation to support the technical aspects of the project, including scripts, visualizations, and model designs.

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

Broader Insights

1. **Real-World Data Challenges**:
    - Recognizing that real-world data is often messy, incomplete, and inconsistent, requiring significant preprocessing before analysis.
    - Developing a systematic approach to address data issues and ensure reliability for downstream tasks.
2. **Scalability and Efficiency**:
    - Understanding the importance of scalable solutions for handling growing datasets, especially in domains like hospitality with expanding offerings and user preferences.
    - Learning to optimize processes, such as query execution in Power BI and computational efficiency in recommendation models, to enhance performance.
3. **User-Centric Design**:
    - Focusing on the end-user experience in dashboard and model design, ensuring usability, interpretability, and engagement.
    - Prioritizing dynamic filters, intuitive navigation, and clear labeling to cater to diverse user needs.

Future Work

The Hotelytics project successfully laid the foundation for data-driven analytics and AI-powered recommendations in the hospitality sector. However, several enhancements and new directions can be pursued to further improve the system's functionality, scalability, and impact.

Enhanced Recommendation System

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

- **Deep Learning Integration**: Experiment with neural networks, such as embeddings or recurrent neural networks (RNNs), to capture complex patterns in user preferences and hotel features.
- **User Feedback Loop**: Incorporate feedback mechanisms to refine recommendations over time based on user satisfaction and interactions.

Advanced Data Analysis

- **Sentiment Analysis**: Analyze guest reviews using natural language processing (NLP) techniques to derive sentiment scores and insights into customer satisfaction.
- **Predictive Analytics**: Use machine learning models to predict future trends, such as price fluctuations, occupancy rates, or service demands, based on historical data.
- **Geographic Patterns**: Expand analysis to include location-based insights, such as proximity to tourist attractions or transportation hubs.

Expansion of Dataset and Scope

- **Broader Data Coverage**: Extend data collection to more cities, countries, and hotel chains, increasing the system's utility across diverse geographic and demographic markets.
- **Integration with Travel Platforms**: Partner with travel aggregators to include flight and activity data, providing a holistic view of travel planning.
- **Seasonality Analysis**: Add temporal dimensions to the analysis to uncover seasonal trends in pricing, demand, and customer preferences.

Scalability and Performance

- **Cloud Deployment**: Host the system on cloud platforms like AWS or Azure to ensure scalability and handle larger datasets efficiently.
- **Optimization of Queries**: Refine Power BI queries and database relationships to reduce loading times for large-scale visualizations.

Integration with Other Technologies

- **IoT Data**: Incorporate Internet of Things (IoT) data, such as smart hotel sensors, to enrich the dataset with real-time metrics like room occupancy or energy usage.
- **Blockchain for Authenticity**: Use blockchain technology to validate reviews and ratings, ensuring data authenticity and trustworthiness.

- **Voice and Chat Interfaces**: Develop voice or chatbot interfaces to allow users to query the system and receive insights in a conversational manner.

Conclusion

The Hotelytics project has demonstrated the transformative potential of data analytics and AI in the hospitality sector. By integrating data extraction, cleaning, analysis, and visualization with a custom recommendation engine, the system successfully provides actionable insights and personalized recommendations to users. This comprehensive solution underscores the value of combining advanced technologies with user-centric design to address industry challenges.

The project began with web scraping techniques that collected extensive data from the Expedia platform, encompassing diverse attributes such as hotel names, ratings, prices, and amenities. The extracted data was then systematically cleaned and transformed to address inconsistencies, standardize formats, and handle missing values. Through the implementation of a star schema, the data was optimized for analysis, enabling seamless querying and efficient visualization.

Power BI played a pivotal role in delivering a user-friendly interface, offering dynamic filters, intuitive visualizations, and interactive drill-through capabilities. Users can explore trends in room type distributions, pricing patterns, and service availability while gaining insights into customer satisfaction metrics like guest ratings. The dashboard not only provides high-level overviews but also enables detailed, drill-down analyses to support decision-making at every level.

The AI-powered recommendation model further enhanced the system by dynamically suggesting hotels tailored to user preferences. Using TF-IDF Vectorization and Cosine Similarity, the model offered a balance of computational efficiency and interpretability, ensuring scalability without compromising on accuracy. The recommendations align closely with observed data trends, validating the model's effectiveness and usability.

The learning outcomes from this project have been invaluable, encompassing technical skills like web scraping, data modeling, and machine learning, as well as analytical proficiencies in visualization and trend analysis. The team also gained experience in project management, user-centric design, and ethical considerations.

In conclusion, Hotelytics successfully bridges the gap between raw data and actionable insights, empowering users with an interactive, scalable, and intelligent platform for hotel analytics. Its modular architecture and flexibility make it adaptable to future enhancements, positioning it as a valuable tool for both businesses and travelers.