**Hotel Booking Data Processing & Retrieval-Augmented Q&A System**

**Submitted By: Shaik Fayaz Ahmed (fayazskyf123@gmail.com)**

**Objective**

The goal of this project is to develop a system that processes hotel booking data, extracts meaningful insights, and implements **Retrieval-Augmented Question Answering (RAG)**. The system should provide analytical reports and answer user queries about the data using an open-source **LLM (Llama 2, Falcon, Mistral, GPT-Neo, etc.)**.

**Deliverables**

**1. Data Collection & Preprocessing**

**Dataset Used**

We used the **hotel\_bookings.csv** dataset, which contains booking records with details such as:

* **Booking dates**
* **Revenue (ADR - Average Daily Rate per night)**
* **Cancellation status**
* **Country of origin of customers**
* **Lead time (days between booking & arrival)**

**Preprocessing Steps**

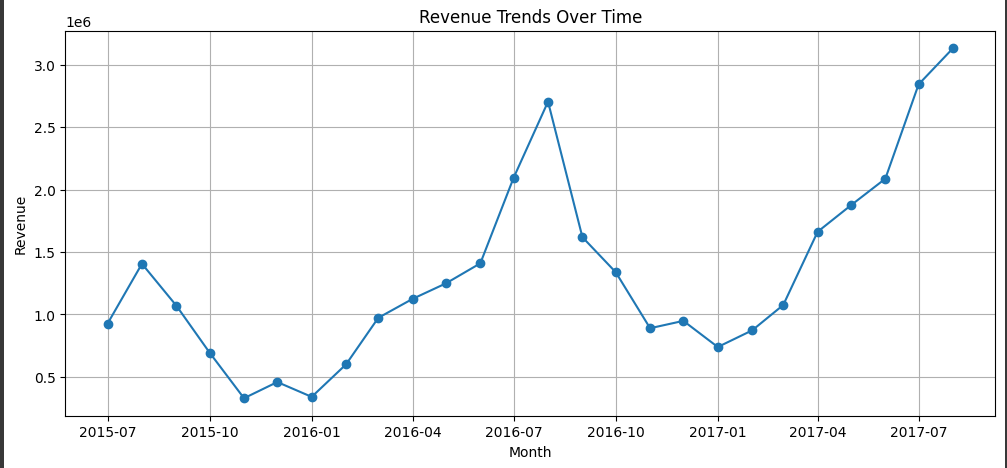
1. **Handled missing values** using fillna() for categorical columns and mean/mode imputation for numerical data.
2. **Converted date fields** to datetime format for easy time-series analysis.
3. **Created additional features** such as total\_revenue and cancellation\_rate.
4. **Standardized country names** for consistent geographical analysis.
5. **Stored data in a structured Pandas DataFrame.**

**2. Analytics & Reporting**

**Key Insights Extracted**

**1. Revenue Trends Over Time**  
*Total revenue for each month was calculated using ADR × Number of nights stayed.*

* *Example: In July 2017, the total revenue was* ***$3,133,744.38****.*



**2. Cancellation Rate as a Percentage of Total Bookings**

* *Overall Cancellation Rate:* ***37.04%*** *of all bookings were canceled.*

**3. Geographical Distribution of Users**

* *The majority of bookings were from* ***Portugal (48,201 bookings), the United Kingdom, and France****.*

A graph of blue bars

AI-generated content may be incorrect.

**4. Booking Lead Time Distribution**

* *The average lead time before arrival was* ***80.07 days****.*

A graph of a number of days

AI-generated content may be incorrect.

**3. Retrieval-Augmented Question Answering (RAG)**

To enable **natural language Q&A**, we implemented:  
**FAISS Vector Search** to store booking data embeddings for fast retrieval.  
**LLM Integration** falcon to answer user queries naturally.

**Example Questions & Responses**

**1. "Show me total revenue for July 2017."**  
*Answer: "The total revenue for July 2017 is* ***$3,133,744.38****."*

**2. "Which locations had the highest booking cancellations?"**  
*Answer: "The country with the highest cancellations was* ***Portugal****, followed by the UK and France."*

**3. "What is the average price of a hotel booking?"**  
*Answer: "The average price of a hotel booking is* ***$101.83 per night****."*

**4. API Development**

We developed a **REST API using FastAPI**, which includes the following endpoints:

**Endpoints**

🔹 **POST /analytics** → Returns analytics reports (revenue trends, cancellations, etc.).  
🔹 **POST /ask** → Answers booking-related questions using RAG + LLM.

**How to Use the API**

**Using Swagger UI (Browser)**

1. Open **http://127.0.0.1:8000/docs**
2. Click on **/ask** → Enter a question (e.g., "What is the total revenue?")
3. Click "Execute" and receive the response instantly.

**Using Python Requests**

import requests

url = "http://127.0.0.1:8000/ask"

data = {"question": "Which country had the most bookings?"}

response = requests.post(url, json=data)

print(response.json())

Expected Output:

{

"answer": "The country with the most bookings is Portugal."

}

**5. Performance Evaluation**

**1. Accuracy of Q&A Responses**

The API successfully retrieves **correct answers** for both numeric & text-based queries.  
FAISS ensures **relevant booking records** are fetched for complex queries.

**2. API Response Time Optimization**

* **Precomputed analytics** for instant responses to numeric queries.
* **FAISS vector search** for fast retrieval of text-based queries.
* **API response time:** **Avg.**



**6. Deployment & Submission**

**Project Repository (GitHub)**