Hotel Reservation Analysis Project: Examining Cancellations and Key Influences

Prisicilla Udomprasert
Department of Computer Science
Auburn University at Montgomery
Montgomery, United States
pudompr1@aum.edu

Jun Thit
Department of Computer Science
Auburn University at Montgomery
Montgomery, United States
jthit@aum.edu

***Abstract*—The analysis of booking behavior and cancellations plays a crucial role in the hospitality industry, helping businesses optimize their operations and improve customer retention. This project uses the Hotel Reservations Dataset to explore factors such as room type, special requests, lead time, and market segments that influence booking cancellations. By utilizing MySQL Workbench, a database was created to manage and query the dataset effectively. Various SQL queries were developed to identify key trends and relationships between booking status and factors like special requests, market segments, and seasonal trends. The findings provide insights into how lead time, room type, and market segment affect the likelihood of cancellations. This project highlights the significance of data-driven decision-making in the hospitality industry and demonstrates how computer science skills can be applied to solve real-world problems. The process also involved refining database design, query optimization, and data analysis techniques to ensure accurate results and efficient execution.**

Keywords—MySQL, hotel reservations, cancellations, data analysis, SQL, market segment

I. INTRODUCTION

Booking a hotel often involves navigating a range of factors such as lead time, room type, special requests, and customer demographics. Using SQL queries, this project analyzes hotel reservation data to better understand booking patterns, cancellation trends, and customer behavior. The goals of this project are to: (1) explore factors that influence reservation statuses such as room type and special requests, (2) examine how lead time and market segment type affect cancellations, and (3) identify key trends in arrival dates and booking characteristics that hotels can use to improve retention. This project demonstrates how structured data analysis can help both hotels and travelers make more informed decisions about reservations.

II. LITERATURE REVIEW

The hotel industry has been extensively studied to understand factors influencing booking trends, cancellation rates, and customer behavior. This review explores how various factors, such as lead time, room type, and customer demographics, affect reservation outcomes. Before implementing the SQL-based approach for this project, a review of existing literature and data sources was conducted to identify key trends and opportunities for further exploration.

A. Causes of Hotel Booking Cancellations

Booking cancellations present a significant challenge for the hotel industry, often resulting in lost revenue and room availability inefficiencies. Research from DoWhy explores the underlying causes of cancellations, highlighting the impact of lead time on booking behavior. Longer lead times are often correlated with higher cancellation rates, as customers may change their plans or find alternative offers closer to their travel dates [1]. Additionally, factors such as the flexibility of cancellation policies and the booking platforms used play critical roles in determining whether a reservation will be canceled. This analysis aims to investigate how these factors, along with other customer attributes, contribute to cancellation rates.

B. Hotel Market Segmentation

Market segmentation is essential for targeting different customer groups effectively. Oaky discusses how understanding various segments such as business travelers, leisure tourists, and families can lead to better marketing strategies [2]. Different groups have distinct booking behaviors; for example, business travelers often book last-minute and are less likely to cancel, while leisure travelers tend to plan their trips further in advance and have higher cancellation rates. Analyzing customer types allows hotels to adjust their offerings and improve retention, especially by understanding how each segment interacts with reservation systems and booking policies.

C. Contribution of This Project

This project uses the “Hotel Reservations Dataset “from Kaggle website to explore how lead time, room type, special requests, and market segments influence cancellation rates and overall booking trends [3]. By analyzing these factors, the goal is to provide actionable insights for hotel management teams to reduce cancellations and enhance booking retention strategies. The structured, data-driven approach offers a more precise method for addressing the challenges of booking and

cancellation management, helping hotels make informed decisions about operational improvements.

III. METHODOLOGY

For this project, we used the "Hotel Reservations Dataset" from Kaggle website to analyze factors influencing hotel booking behaviors, cancellations, and retention [3]. The dataset was imported into a MySQL database system where it was structured and organized into tables, including Guests, Reservations, and ArrivalInfo. MySQL Workbench was used as the primary tool for querying, managing, and analyzing the database.

A. *Database Setup*

The project began with the creation of a database, HotelReservationDB, in MySQL. Below are the key steps taken for setting up the database:

1) *Database Creation*

The database was created and named as HotelReservationDB using the following SQL command:

```
DROP DATABASE IF EXISTS HotelReservationDB;
CREATE DATABASE HotelReservationDB;
USE HotelReservationDB;
```

Fig. 1. HotelReservationDB Database Creation and Initialization Script

2) *Table Creation*

Three main tables were created: Guests, Reservations, and ArrivalInfo. These tables were designed to store the relevant data points for analysis.

a) Guests Table: Contains customer-related data, such as booking ID, number of adults, children, special requests, and cancellation history.

```
CREATE TABLE Guests (
    Booking_ID VARCHAR(20) PRIMARY KEY,
    No_of_Adults INT,
    No_of_Children INT,
    Required_Car_Parking_Space BOOLEAN,
    Repeated_Guest BOOLEAN,
    No_of_Previous_Cancellations INT,
    No_of_Previous_Bookings_Not_Canceled INT,
    No_of_Special_Requests INT
);
```

Fig. 2. Guests Table

b) Reservations Tabel: Contains booking-specific details, including lead time, room type reserved, and booking status.

```
CREATE TABLE Reservations (
    Booking_ID VARCHAR(20),
    Lead_Time INT,
    Type_of_Meal_Plan VARCHAR(50),
    Room_Type_Reserved VARCHAR(50),
    No_of_Weekend_Nights INT,
    No_of_Week_Nights INT,
    Avg_Price_Per_Room DECIMAL(10,2),
    Booking_Status VARCHAR(20),
    FOREIGN KEY (Booking_ID) REFERENCES Guests(Booking_ID)
);
```

Fig. 3. Reservations Tabel

c) ArrivalInfo Table: Contains details about the arrival date, year, month, and market segment type for each booking.

```
CREATE TABLE ArrivalInfo (
    Booking_ID VARCHAR(20),
    Arrival_Year INT,
    Arrival_Month INT,
    Arrival_Date INT,
    Market_Segment_Type VARCHAR(50),
    FOREIGN KEY (Booking_ID) REFERENCES Guests(Booking_ID)
);
```

Fig. 3. Reservations Tabel

3) *Data Insertion*

Data was manually inserted into the database, derived from the "Hotel Reservations" CSV file by Brian Risk from Kaggle [4]. This included guest details, booking information, and arrival-related data.

B. *Queries and Analysis*

Once the database was set up, we focused on developing SQL queries to extract meaningful insights related to booking behavior, cancellations, and customer retention. The queries used in this analysis were as follows:

1) *Cancellation Rate by Room Type and Special Requests:* The first query investigated the influence of Room Type Reserved and Number of Special Requests on booking cancellations. It joined the Reservations and Guests tables and grouped the data by room type and special requests.

```sql
USE hotelreservationdb;

/* Query 1: Cancellation rate by Room Type and Special Requests
(For Goal 1: Factors influencing reservation status) */
SELECT r.Room_Type_Reserved, g.No_of_Special_Requests,
r.Booking_Status, COUNT(*) AS Total_Bookings
FROM Reservations r JOIN Guests g
ON r.Booking_ID = g.Booking_ID
GROUP BY r.Room_Type_Reserved, g.No_of_Special_Requests, r.Booking_Status
ORDER BY r.Room_Type_Reserved, g.No_of_Special_Requests;
```

Fig. 4. SQL Query for Analyzing Cancellation Rates by Room Type and Special Requests

2) *Average Lead Time for Canceled vs. Not Canceled Bookings:* The second query explored the Average Lead Time for canceled and non-canceled bookings, grouped by the Booking Status.

```sql
/* Query 2: Average Lead Time for Canceled vs. Not Canceled Bookings
(For Goal 2: Lead time effect on cancellations) */
SELECT Booking_Status, AVG(Lead_Time) AS Avg_Lead_Time
FROM Reservations
GROUP BY Booking_Status;
```

Fig. 5. SQL query for calculating average lead time by booking status

3) *Cancellation Rate by Market Segment Type:* This query analyzed the Market Segment Type and its impact on booking cancellations. It joined the ArrivalInfo and Reservations tables and grouped by Market Segment and Booking Status.

```sql
/* Query 3: Cancellation rate by Market Segment Type
(For Goal 2: Market segment effect) */
SELECT a.Market_Segment_Type, r.Booking_Status, COUNT(*) AS Total_Bookings
FROM ArrivalInfo a JOIN Reservations r
ON a.Booking_ID = r.Booking_ID
GROUP BY a.Market_Segment_Type, r.Booking_Status
ORDER BY a.Market_Segment_Type;
```

Fig. 6. SQL Query for Analyzing Cancellation Rates by Market Segment Type

4) *Average Lead Time and Cancellation Rate by Market Segment Type (Combined Query):* This query combined the analysis of Average Lead Time and Cancellation Rate by Market Segment Type.

```sql
/* Combined Query (2 and 3): Average Lead Time and Cancellation Rate by Market Segment Type*/
SELECT a.Market_Segment_Type, r.Booking_Status,
AVG(r.Lead_Time) AS Avg_Lead_Time, COUNT(*) AS Total_Bookings
FROM ArrivalInfo a JOIN Reservations r
ON a.Booking_ID = r.Booking_ID
GROUP BY a.Market_Segment_Type, r.Booking_Status
ORDER BY a.Market_Segment_Type, r.Booking_Status;
```

Fig. 7. SQL Query For Combining Average Lead Time and Cancellation Rates by Market Segment Type

5) *Booking and Cancellation Trends by Arrival Month:* The final query examined Booking Trends and Cancellation Rates by month of arrival. It grouped the data by Arrival Month and Booking Status.

```sql
/* Query 4: Booking and Cancellation Trends by Arrival Month
(For Goal 3: Identify key trends to improve booking retention) */
SELECT
    CASE a.Arrival_Month
        WHEN 1 THEN 'Jan'
        WHEN 2 THEN 'Feb'
        WHEN 3 THEN 'Mar'
        WHEN 4 THEN 'Apr'
        WHEN 5 THEN 'May'
        WHEN 6 THEN 'Jun'
        WHEN 7 THEN 'Jul'
        WHEN 8 THEN 'Aug'
        WHEN 9 THEN 'Sep'
        WHEN 10 THEN 'Oct'
        WHEN 11 THEN 'Nov'
        WHEN 12 THEN 'Dec'
    END AS Arrival_Month_Name, r.Booking_Status, COUNT(*) AS Total_Bookings
FROM ArrivalInfo a JOIN Reservations r
ON a.Booking_ID = r.Booking_ID
GROUP BY Arrival_Month_Name, r.Booking_Status
ORDER BY MIN(a.Arrival_Month), r.Booking_Status;
```

Fig. 8. SQL Query for Identifying Booking and Cancellation Trends by Arrival Month

C. Software and Tools

1) *MySQL Workbench:* This was the the primary software tool used to create the database, run SQL quries, and perform data analysis. It provided an integrated environment for developing, testing, and visualizing query results.

2) *Kaggle Dataset:* The "Hotel Reservations Dataset" from Kaggle, sourced from the "Hotel Reservations" CSV file by Brian Risk, was used for data input. This dataset contains a variety of features related to hotel bookings, demographics, and booking details

3) *GitHub Repository: Github* was used to store and manage all project files, including the SQL code, database creation scripts, and documentation. This made it easy to track any changes, work together, and access project files [5].

\

IV. Conclusion

This project successfully analyzed the factors influencing booking cancellations and market segment effects using the "Hotel Reservations Dataset". By creating and querying a MySQL database, insights were derived from examining various aspects such as room type, special requests, lead time, and market segments. The queries helped uncover the relationship between booking status and factors like special requests, lead time, and arrival month, offering actionable insights into improving booking retention and minimizing cancellations. The analysis demonstrated how SQL queries can be effectively used to identify key patterns in large datasets, a

critical skill for any computer science student involved in data management and analysis.

As computer science students, this project provided a hands-on learning experience in database design, query development, and data analysis. It strengthened our understanding of relational databases, SQL, and the practical application of these technologies to real-world problems. Throughout the project, we developed our problem-solving skills by navigating through queries and ensuring accurate data handling and reporting. Additionally, this project enhanced our ability to draw meaningful conclusions from raw data, a key competence in data science and analytics. Ultimately, this experience solidified our foundation in both database management and data analysis, preparing us for future challenges in the field of computer science.

REFERENCES

[1] DoWhy, "Exploring Causes of Hotel Booking Cancellations." [Online]. Available: https://www.pywhy.org/dowhy/v0.9.1/example_notebooks/DoWhy-The%20Causal%20Story%20Behind%20Hotel%20Booking%20Cancellations.html. [Accessed: April 27, 2025].

[2] Oaky, "Hotel Market Segmentation: How to Appeal to the Right Guests," October 22, 2024. [Online]. Available: https://oaky.com/en/blog/hotel-segmentation. [Accessed: April 27, 2025].

[3] Ahsan Raza, "Hotel Reservations Dataset," January 4, 2023. [Online]. Available: https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset/code. [Accessed: April 27, 2025].

[4] Brian Risk, "Hotel Reservations Classification Dataset," March 26, 2025. [Online]. Available: https://www.kaggle.com/code/devraai/hotel-reservations-classification-analysis/input. [Accessed: April 27, 2025].

[5] Priscilla Udomprasert, "hotel_db_project." [Online]. Available: https://github.com/PriscillaUdomprasert/hotel_db_project. [Accessed: April 27, 2025].