



Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

Hotel Booking Intelligence Pipeline

The domain of the Project: Hotel Revenue and Customer Behavior Analytics

COURSE NAME: Data Analysis and Data Science

Team Mentors (and their designation):

Purnangshu Roy
AI Consultant at CSR BOX

Team Members:

Ms. Ajanta Ghosh

Period of the project

November 2025 to December 2025



Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

Declaration

The project titled "**Hotel Booking Intelligence Pipeline**" has been mentored by **Purnangshu Roy**, organised by **SURE Trust**, from **June 2025 to December 2025**, for the benefit of the educated unemployed rural youth for gaining hands-on experience in working on industry relevant projects that would take them closer to the prospective employer. I declare that to the best of my knowledge the members of the team mentioned below, have worked on it successfully and enhanced their practical knowledge in the domain.

Team Members:

Ms. Ajanta Ghosh

Mentor's Name
AI Consultant—CSR BOX

Prof. Radhakumari
Executive Director & Founder
SURE Trust



Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

Table of contents

1. Executive summary
2. Introduction
3. Project Objectives
4. Methodology & Results
5. Social / Industry relevance of the project
6. Learning & Reflection
7. Future Scope & Conclusion



Executive Summary

The **Hotel Booking Intelligence Pipeline** is an end-to-end data analytics project designed to convert raw hotel reservation data into meaningful insights and predictive intelligence. The project integrates multiple technologies—**Excel, SQL, Python, and Power BI**—to analyze booking behavior, cancellation patterns, and revenue performance.

Initially, raw booking data was cleaned and transformed using Excel, where duplicates, missing values, and inconsistencies were handled, and new derived features were created. The structured data was then stored and queried using SQL to enable efficient analysis of booking trends and cancellation drivers. Python was used to build machine learning models for predicting booking status, estimating cancellation probability, and forecasting revenue. Finally, Power BI dashboards were developed to present key performance indicators, trends, and insights in an interactive and business-friendly format.

The outcomes of this project help hotel management identify high-risk bookings, understand factors influencing cancellations, and evaluate revenue drivers. Overall, the system demonstrates how data-driven intelligence can support strategic decision-making, improve operational efficiency, and enhance revenue optimization in the hospitality domain.



Introduction

Background and Context

The hotel industry generates vast amounts of data from bookings, customer preferences, and cancellations. Efficient analysis of this data can help hotels optimize revenue, improve occupancy, and enhance customer satisfaction. With the growth of online reservations, leveraging data-driven insights has become crucial for strategic decision-making.

Problem Statement

Hotels often face challenges such as last-minute cancellations, uneven occupancy, and inefficient pricing strategies. Manual analysis of booking data is time-consuming and prone to errors. There is a need for an automated system that can process booking data, predict cancellations, identify high-value customers, and provide actionable insights to optimize operations.

Scope and Limitations

Scope: The project analyzes hotel booking data to identify trends, predict cancellations, and highlight high-value customers. It provides interactive dashboards and insights to help hotels optimize revenue, improve occupancy, and enhance customer satisfaction.

Limitations: The analysis is limited by the quality and completeness of the dataset. External factors like seasonal events or sudden market changes may affect predictions. Real-time data integration and external market variables are not included.

Innovation Component

The project introduces an automated data pipeline that cleans, analyzes, and visualizes hotel booking data. It leverages predictive analytics to forecast cancellations and identify high-value bookings. Additionally, an interactive dashboard provides actionable insights for revenue optimization, improved occupancy management, and better customer experience.



Project Objectives

The objectives of the project are as follows:

- To clean and preprocess raw hotel booking data for analysis
- To perform exploratory data analysis to identify key trends and patterns
- To analyze factors contributing to booking cancellations
- To build a predictive model to classify bookings as canceled or not canceled
- To create interactive dashboards for effective data visualization
- To demonstrate an end-to-end data science workflow using industry tools



Project Objectives

Data Cleaning & Preparation: Ensure booking data is accurate, complete, and ready for analysis.

Trend Analysis: Identify patterns in bookings, cancellations, and customer behavior.

Predictive Modeling: Forecast cancellations and detect high-value bookings using machine learning.

Revenue Optimization: Provide insights to improve pricing strategies, occupancy, and overall revenue.

Interactive Visualization: Create dashboards for easy exploration and decision-making by hotel management.

Decision Support: Enable data-driven strategies to enhance customer satisfaction and operational efficiency.



Methodology and Results

Methods / Technology Used:

Data Cleaning & Transformation: Handle missing values, duplicates, and outliers.

Data Analysis: Explore trends in bookings, cancellations, and customer behavior.

Predictive Modeling: Use machine learning models (e.g., Random Forest, Linear Regression) to forecast cancellations and identify high-value bookings.

Visualization & Dashboarding: Represent insights using interactive charts and slicers.

Tools / Software Used:

Excel / Power Query: Data cleaning, transformation, and initial analysis.

Python (Pandas, Scikit-learn, Matplotlib, Seaborn): Predictive modeling and visualizations.

MySQL / SQL: Data storage, querying, and integration.

Power BI : Dashboard creation for interactive insights.

Data Collection Approach:

Historical hotel booking dataset collected from online hotel reservation records.

Includes attributes such as booking dates, lead time, revenue, customer rating, number of guests, and cancellations.

Data was pre-processed to remove duplicates, handle missing values, and calculate derived metrics like net revenue and stay duration.

Project Architecture:

Data Input: Raw booking dataset from Excel/CSV.

Data Preprocessing: Cleaning, missing value handling, and feature engineering.

Analysis & Modeling: Trend analysis, predictive modeling for cancellations and high-value bookings.

Visualization & Dashboard: Interactive charts, slicers, and reports for decision support.



Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

Insights & Recommendations: Actionable outputs for revenue and occupancy optimization.

Results:

Predictive model accurately forecasts booking cancellations and highlights high-value customers.

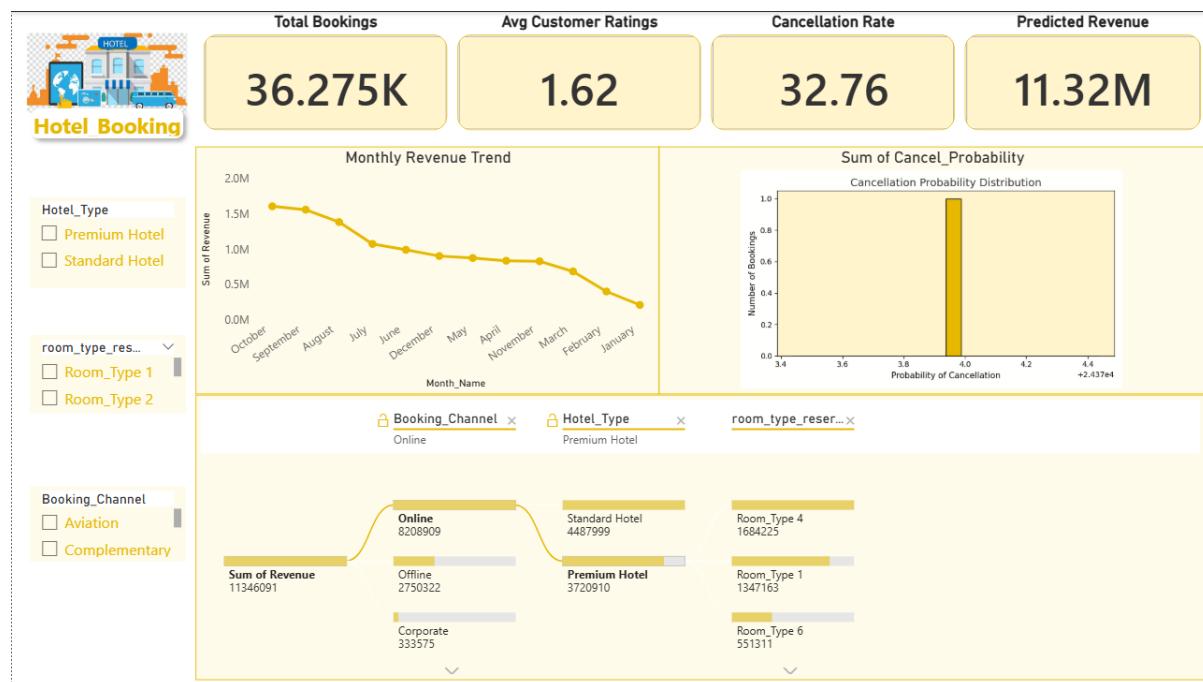
Interactive dashboard allows filtering by room type, country, and booking status to monitor performance.

Insights help in revenue optimization by identifying high-value bookings and understanding booking trends.

Provides actionable recommendations for hotel management to improve occupancy and customer satisfaction.

Project GitHub Link:

<https://github.com/sure-trust/AJANTA-GHOSH-g8-ds/tree/main/Final%20capstone%20project>



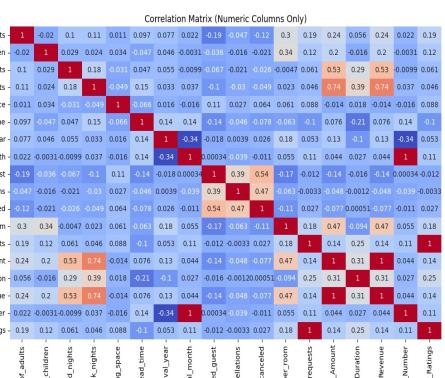
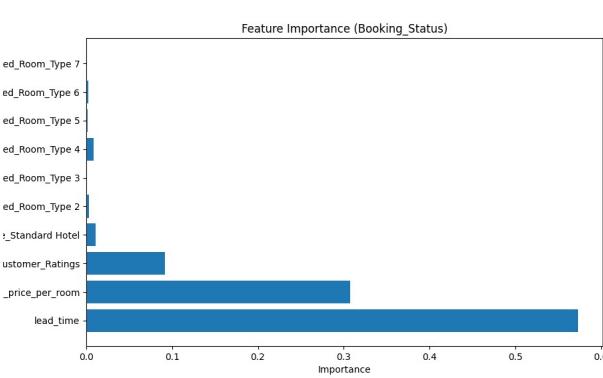


Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

	A	P	R	U	V	W	AB	AC	AD	AE	AF	AG
	Booking_ID	avg_price_per_room	booking_status	Revenue	Hotel_Type	Booking_Channel	Check_In_Date	Check_Out_Date			Column1	Column2
2	INN00001	65	Not_Cancelled	195	Standard Hotel	Offline	10/2/2017	10/2/2017			Total_Check_Ins_Today	
3	INN00002	107	Not_Cancelled	535	Standard Hotel	Online	11/6/2018	11/6/2018			Total_Revenue_Today	
4	INN00003	60	Cancelled	180	Standard Hotel	Online						
5	INN00004	100	Cancelled	200	Standard Hotel	Online						
6	INN00005	94	Cancelled	188	Standard Hotel	Online						
7	INN00006	115	Cancelled	230	Standard Hotel	Online						
8	INN00007	108	Not_Cancelled	432	Standard Hotel	Online	10/15/2017	10/15/2017				
9	INN00008	106	Not_Cancelled	424	Standard Hotel	Online	12/26/2018	12/26/2018				
10	INN00009	97	Not_Cancelled	388	Standard Hotel	Offline	7/6/2018	7/6/2018				
11	INN00010	133	Not_Cancelled	665	Premium Hotel	Online	10/18/2018	10/18/2018				
12	INN00011	85	Not_Cancelled	85	Standard Hotel	Online	9/11/2018	9/11/2018				
13	INN00012	140	Not_Cancelled	420	Premium Hotel	Online	4/30/2018	4/30/2018				
14	INN00013	88	Cancelled	264	Standard Hotel	Online						
15	INN00014	90	Cancelled	180	Standard Hotel	Online						
16	INN00015	94	Not_Cancelled	188	Standard Hotel	Online	10/20/2017	10/20/2017				
17	INN00016	115	Cancelled	230	Standard Hotel	Online						
18	INN00017	96	Not_Cancelled	96	Standard Hotel	Offline	10/5/2017	10/5/2017				
19	INN00018	96	Not_Cancelled	384	Standard Hotel	Online	8/10/2017	8/10/2017				
20	INN00019	65	Cancelled	260	Standard Hotel	Online						

PIVOT TABLE FOR SUM OF REVENUE BY HOTEL TYPE AND BOOKING CHANNEL

	A	B	C	D	E	F	G	H	I	J	K	L	M	N		
PIVOT TABLE FOR SUM OF REVENUE BY HOTEL TYPE AND BOOKING CHANNEL																
	room_type...	booking....	Room_Type 1	Room_Type 2	Cancelled	Not_Cancelled										
3	Sum of Revenue	Column Labels														
4			Premium Hotel		Premium Hotel Total	Standard Hotel										
5	Row Labels		Corporate	Offline	Online		Aviation		Complementary	Corporate	Offline	Online				
6	January		488	1600	2088				6	10978	60455	96695		168134		
7	February			3493	3493	352	24	17720	76697	106867		201660		205153		
8	March	139	1404	32313	33856	355	730	18425	87674	168604		275788		309644		
9	April		518	4013	34142	38673	6428	0	5234	78437	196884		286983		325656	
10	May		12075	15661	70175	97911	2185		0	22015	89534	120339		234073		331984
11	June		13581	79632	93213	1235	0	8971	177398	127314		314918		408131		
12	July		552	3194	85229	88975	158	111	10612	86430	159966		257277		346252	
13	August		318	7232	140225	147775		12	14292	131161	233331		378796		526571	
14	September	340	6693	50302	197899	255234	1150	548	29583	208275	168973		408529		663763	
15	October		6335	7782	124507	138624	1045	30	26171	287640	249683		564569		703193	
16	November		1647	3502	78174	83323	237	263	23125	86520	236634		346779		430102	
17	December			18668	34760	53428	100	26724	143451	289436		459711		513139		
18	Grand Total		479	28138	125827	882149	1036593	13145	1824	213850	1513672	2154726		3897217		4933810
19																
20																





Learning and Reflection

Technical Learnings:

Gained hands-on experience in data cleaning, preprocessing, and feature engineering.

Learned to implement predictive models using Python and evaluate their performance.

Developed skills in data visualization and creating interactive dashboards.

Understood integration of multiple tools (Excel, SQL, Python) for end-to-end data analysis.

Management & Professional Learnings :

Improved project planning and time management to meet milestones.

Learned to document processes and present insights in a clear, professional manner.

Gained experience in working with large datasets and translating technical results into business insights.

Developed problem-solving and decision-making skills in a real-world scenario.

Overall Experience :

The project provided a comprehensive understanding of hotel data analytics from raw data to actionable insights.

Strengthened both technical and professional competencies.

Enhanced confidence in using data-driven approaches for operational and strategic decisions.

Overall, it was a valuable learning experience bridging theory and practical application.



Conclusion and Future Scope

Conclusion

The “Hotel Booking Intelligence Pipeline” successfully transformed raw booking data into actionable insights. By analyzing trends, predicting cancellations, and identifying high-value customers, the project demonstrated how data-driven decision-making can optimize revenue, improve occupancy, and enhance customer satisfaction. The interactive dashboard enables hotel management to quickly interpret data and implement strategies effectively.

Future Scope

Integrate **real-time data** for dynamic prediction and monitoring.

Enhance predictive models with **advanced machine learning techniques** (e.g., XGBoost, Neural Networks).

Include **external factors** like seasonal events, promotions, or competitor pricing for more accurate forecasting.

Expand the dashboard to provide **automated alerts and recommendations** for management decisions.

Explore **customer segmentation and personalized marketing strategies** to further increase revenue.



Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)