Alexandria University
Faculty of Computer and Data Science
Department : Data Science
Course Title: Data Science 2023-2024

Hotel Bookings Analysis

In

Introduction to Data Science

Course Code: 02-24-00104

Members Names and Role

Name	ID	Role
1. محمد محمود محمد عبد المجيد	22010230	Report & Documentation
2. مصطفى محمود محمد	22010260	Machine Learning
3. فؤاد رمزي فؤاد حسن	22010180	Visualization
4. اياد عنان عابد	22010061	Cleaning & EDA
5. محمد هشام إبراهيم احمد	22011634	UI

1. Introduction:

The project focuses on a comprehensive analysis of hotel booking data to derive insights into booking patterns, customer behavior, and factors influencing cancellations. The primary objective is to uncover trends, understand customer preferences, and identify key features affecting booking cancellations in the hospitality industry.

Objective:

The primary goal is to analyze a dataset containing hotel booking information, shedding light on various aspects such as booking frequency, lead time, customer demographics, and booking changes. The aim is to unveil patterns, correlations, and influential factors that contribute to booking cancellations. Insights derived from this analysis can aid in optimizing hotel operations, marketing strategies, and customer satisfaction initiatives.

Inputs and Outputs:

The dataset includes diverse parameters like arrival dates, lead times, customer demographics (such as adults, children, and babies), previous cancellations, booking changes, and more. The analysis output consists of visualizations, statistical summaries, and predictive models illustrating trends in booking behaviors, the relationship between different variables, and the impact of various factors on booking cancellations.

Dataset Description:

The dataset contains extensive records of hotel bookings, encompassing multiple hotels, different room types, and various booking-related attributes. It includes information about customer details and stay duration. Parameters such as lead time (the duration between booking and check-in), booking changes, previous cancellations, and demographic information are vital features within the dataset.

Parameters:

Key parameters in the dataset include arrival dates, lead time, stay duration, customer demographics (adults, children, babies), booking changes, and previous cancellations. These parameters serve as crucial elements for understanding customer behavior, preferences, and the factors that contribute to booking cancellations.

2. Methodologies used:

Descriptive Analysis: The initial step involved a comprehensive exploration of the dataset to understand its structure, data types, missing values, and basic statistics. This method helped in gaining insights into the dataset's characteristics, identifying key variables, and preparing the data for further analysis.

Data Cleaning: To ensure data quality, various data cleaning techniques like handling missing values, removing duplicates, and correcting inconsistencies were applied.

Feature Engineering: Derived features like total number of spent nights, conversion of textual months to numeric values, and transforming categorical variables were performed to enhance the dataset and extract more meaningful information for analysis.

Cluster Analysis (K-Means Clustering): Utilized K-Means clustering to segment customers into distinct groups based on similar characteristics. This technique allowed the identification of customer segments exhibiting similar booking behaviors, aiding in targeted marketing strategies and customer-centric services.

Visualization Techniques: Extensive use of data visualization techniques such as bar plots, histograms, box plots, and scatter plots were employed to visualize patterns, trends, and relationships between various features. Visualizations helped in understanding distributions, correlations, and the impact of different factors on booking cancellations.

Statistical Analysis: Statistical methods including mean, median, and frequency calculations were applied to summarize and analyze the dataset. This helped in understanding central tendencies and distributions within each variable, especially in the context of different clusters.

Predictive Modeling: For further analysis, predictive models such as logistic regression or decision trees might be employed to predict booking cancellations based on various features. These models can uncover the most influential factors contributing to cancellations and assist in devising strategies to reduce them.

Rationale for Methodologies:

Each methodology was chosen based on its suitability for extracting meaningful insights from the dataset:

- **Descriptive Analysis:** To gain an initial understanding of the dataset's structure and contents.
- **Data Cleaning:** To ensure the dataset's reliability and quality for accurate analysis.
- **Feature Engineering:** To derive new features that could reveal more about customer behavior.
- **Cluster Analysis:** For customer segmentation and personalized marketing strategies.
- **Visualization Techniques:** To present data patterns and relationships in an easily interpretable manner.
- **Statistical Analysis:** To understand the central tendencies and distributions of variables.
- **Predictive Modeling:** To forecast future cancellations and identify influential factors.

3. Challenges in the dataset:

Missing Data: Dealing with missing values in crucial columns like company, agent, or country posed a challenge. Strategies such as imputation or dropping rows with missing values had to be carefully considered to avoid bias or loss of important information.

Data Quality Issues: Inconsistent entries, duplicates, or outliers in variables like lead time, ADR, or booking changes required extensive data cleaning and validation processes.

Categorical Variables: Handling categorical variables like meal types or hotel types involved encoding or transforming these variables into a suitable format for analysis.

Interpreting Clusters: Post K-Means clustering, interpreting and understanding the distinct customer segments (clusters) required an in-depth analysis. It was essential to interpret each cluster's characteristics accurately to derive actionable insights.

Data Volume: Working with a sizable dataset comprising thousands of rows demanded efficient computational resources and programming techniques to perform analyses effectively within reasonable timeframes.

Visualization Complexity: Visualizing multi-dimensional data and communicating complex relationships between variables in a clear and concise manner posed a challenge.

4. Interpretations of the results

1. Clustering Analysis:

Utilized K-Means clustering to segment customers into distinct clusters based on booking behaviors.
Identified five clusters exhibiting varying booking patterns, such as “Early Bookers”, “With Babies”, “With Children”.

2. Cancellation Trends:

Early Bookers had the highest cancellation rate at 65%, keep in mind that this group averages a lead time of around 230 days, meaning that their status is volatile, a lot of things can change in 230 days which might be a possible reason for the high cancellation rate, however further analysis is required.

3. Booking Changes Impact:

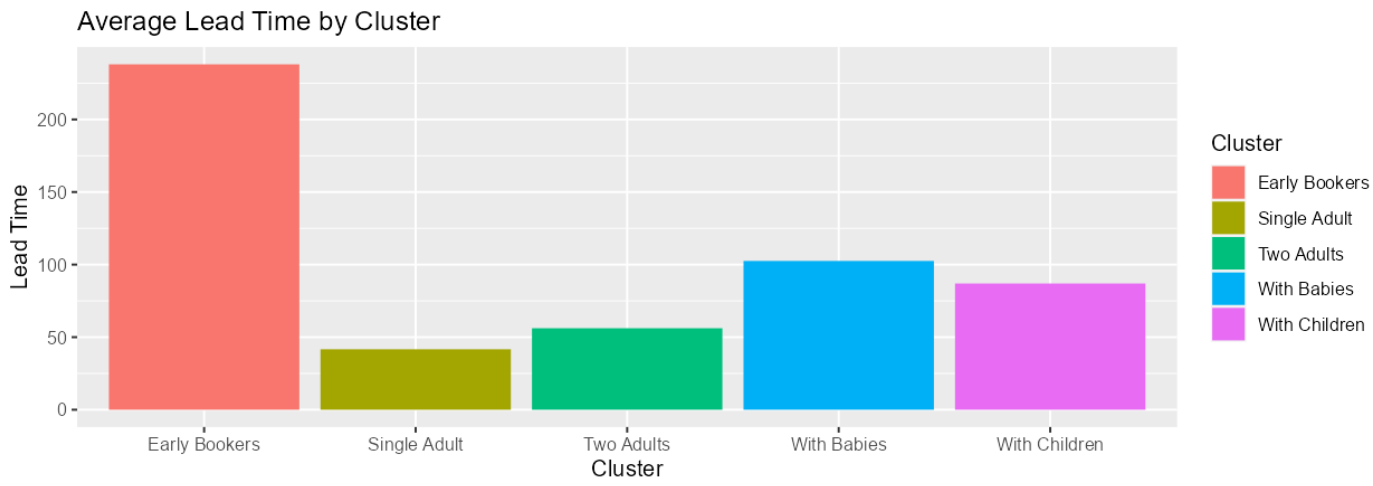
Booking changes on average had no significant correlation to cancellation rate.
Interestingly, “With Babies” cluster had the most booking changes on average, with most of the data falling in “2” booking changes, signifying that guests with babies are likely to not arrive at the initial booking date.

4. Seasonal Booking Trends:

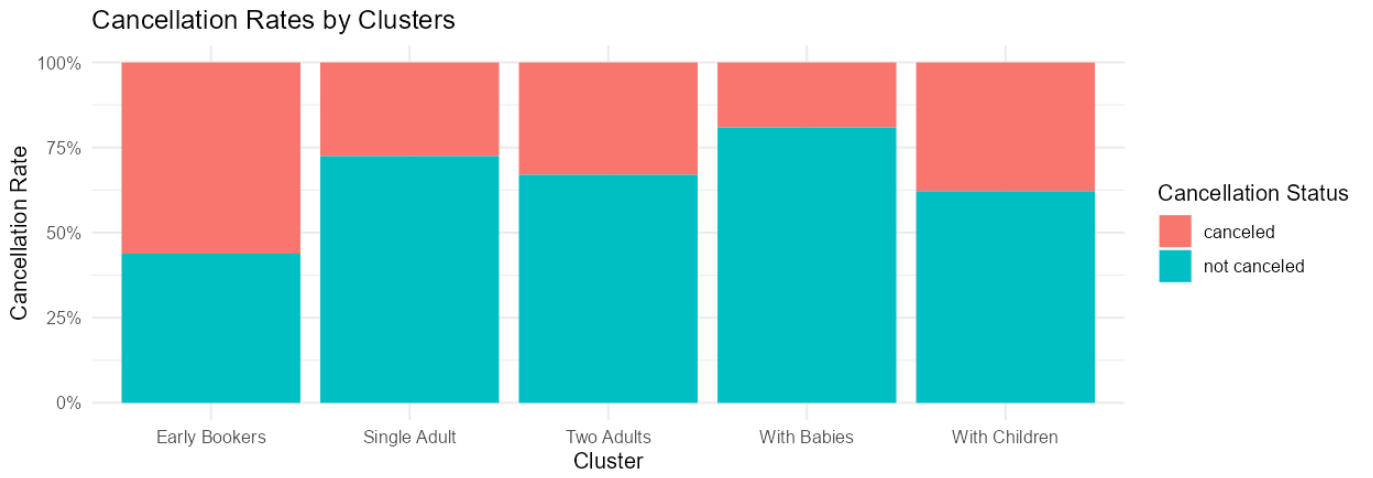
Discovered higher booking frequencies in summer months (June to August) across all clusters.
Clusters “With Children” and “With Babies” had more pronounced seasonal variations, suggesting different vacation preferences among these segments.
On the contrary, “Single Adult” cluster shyed away from these months.

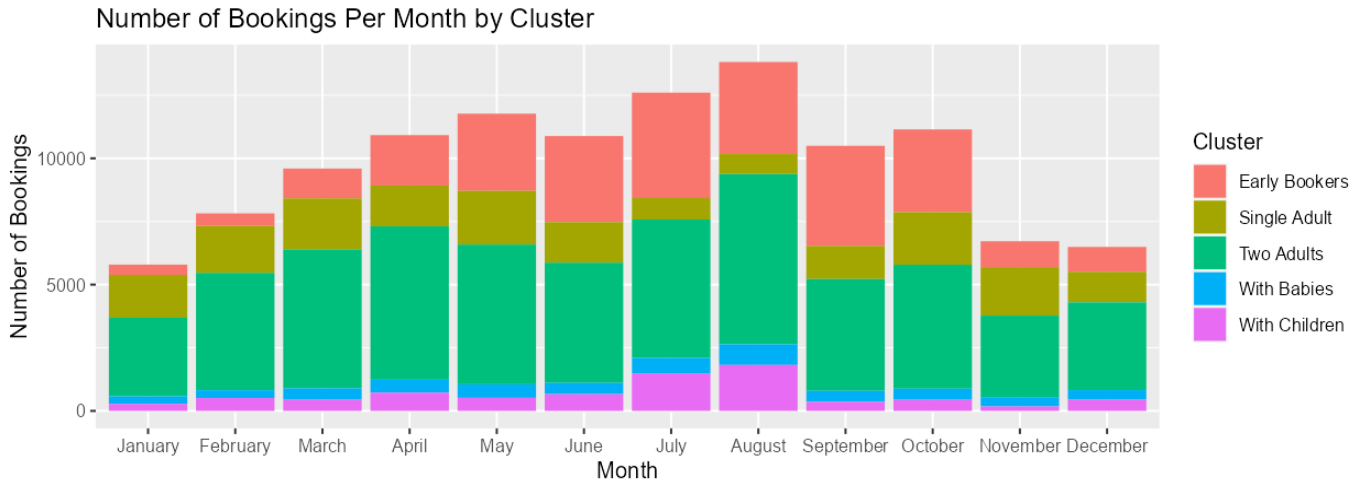
5. Lead Time Analysis:

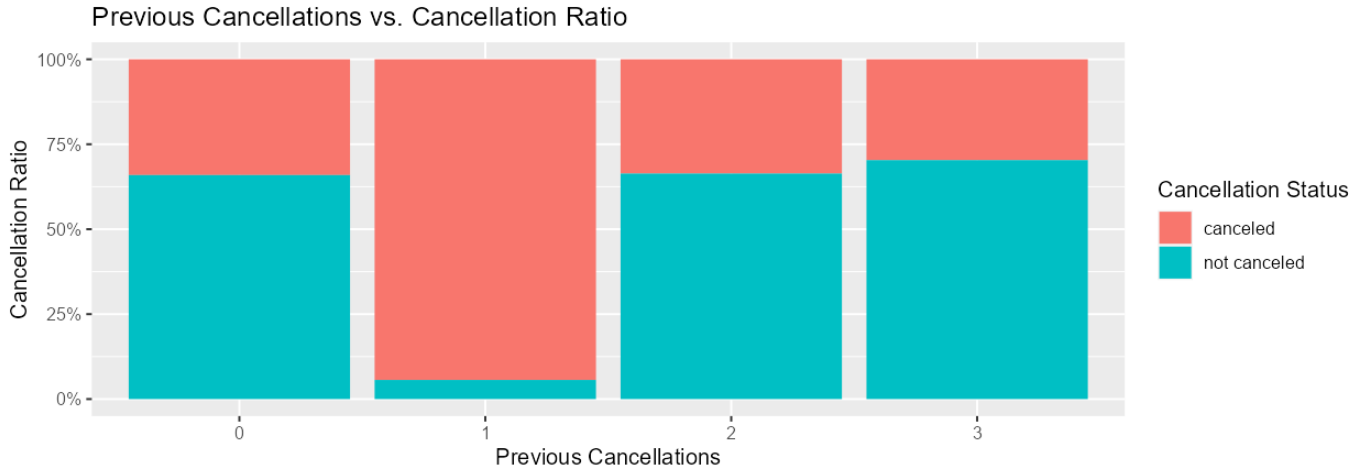
“Early Bookers” cluster obviously had the largest lead time at over 230 days, with “Single Adult” cluster having the lowest at 40 days, perhaps hinting at the more spontaneous nature of these bookings.


Average Lead Time by Cluster
Lead Time
200
150
100
50
0
Early Bookers
Single Adult
Two Adults
With Babies
With Children
Cluster
Cluster
Early Bookers
Single Adult
Two Adults
With Babies
With Children


Cancellation Analysis by Booking Changes
Early Bookers
Single Adult
Two Adults
With Babies
With Children
Count
Booking Changes
Cancellation Status
canceled
not canceled


Cancellation Rates by Clusters
Cancellation Rate
100%
75%
50%
25%
0%
Early Bookers
Single Adult
Two Adults
With Babies
With Children
Cluster
Cancellation Status
canceled
not canceled


Number of Bookings Per Month by Cluster
Number of Bookings
10000
5000
0
January
February
March
April
May
June
July
August
September
October
November
December
Month
Cluster
Early Bookers
Single Adult
Two Adults
With Babies
With Children


Previous Cancellations vs. Cancellation Ratio
Cancellation Ratio
100%
75%
50%
25%
0%
0
1
2
3
Previous Cancellations
Cancellation Status
canceled
not canceled

5. Conclusion:

Key Findings:

- **Cluster-Based Behavior:** Clustering analysis uncovered distinct customer segments exhibiting diverse booking habits. These segments varied in lead times, booking changes, and cancellation propensities.

- **Cancellation Dynamics:** Certain clusters demonstrated significantly higher cancellation rates, indicating potential areas for targeted interventions or service improvements.

- **Booking Changes Impact:** Booking Changes alone had no significant impact on cancellation rates.

- **Seasonal Trends:** Seasonal variations in booking frequencies highlighted different travel preferences among various customer segments.

Implications:

Understanding these diverse behaviors provides invaluable insights for the hotel industry to tailor strategies for different customer segments. Strategies aimed at reducing cancellation rates, optimizing booking experiences, and seasonal promotions can be devised based on these findings.

Future Considerations:

Future studies could delve deeper into the reasons behind booking changes and cancellations within each cluster, potentially uncovering underlying factors influencing these behaviors.