Association Rule Mining Analysis in Retail Transaction Data

Higher National Diploma in Software Engineering

Data Warehousing and Data Mining

Project Documentation

2023.3F

COHNDSE233f – 019 O K Samarasinghe

COHNDSE233f - 020 W A C S Weerasinghe

COHNDSE233f – 021

COHNDSE233f – 022

COHNDSE233f - 023

A blue and black logo  Description automatically generated

School of Computing and Engineering

National Institute of Business Management

Colombo-7

Contents

[1. Introduction 3](#_Toc171773597)

[1.1 Topic of the Study 3](#_Toc171773598)

[1.2 Research Problem 3](#_Toc171773599)

[1.3 Purpose of the Study 3](#_Toc171773600)

[2. Methodology 4](#_Toc171773601)

[2.1 Research Design 4](#_Toc171773602)

[2.2 Data Collection and Preprocessing 4](#_Toc171773603)

[Data Source: 4](#_Toc171773604)

[Preprocessing Steps: 4](#_Toc171773605)

[Association Rule Mining: 12](#_Toc171773606)

[2.3 Clustering Analysis 14](#_Toc171773607)

[K-means Clustering 14](#_Toc171773608)

[2.4 Data Visualization 20](#_Toc171773609)

[2.5 Research Objectives 25](#_Toc171773610)

[3. Results 25](#_Toc171773611)

[3.1 Findings from Association Rule Mining 25](#_Toc171773612)

[3.2 Finding from Clustering 28](#_Toc171773613)

[3.3 Visualization 29](#_Toc171773614)

[4. Discussion 30](#_Toc171773615)

[4.1 Interpretation of Results 30](#_Toc171773616)

[4.2 Comparison with Alternative Techniques 31](#_Toc171773617)

[4.3 Insights and Implications 31](#_Toc171773618)

[4.4 Limitations and Future Research 31](#_Toc171773619)

[5. Conclusion 32](#_Toc171773620)

[References 34](#_Toc171773621)

[Appendices 35](#_Toc171773622)

# 1. Introduction

## 1.1 Topic of the Study

The primary aim of this report is to deep analysis of association rule mining and clustering on hotel room booking data is to identify ongoing patterns of customer bookings. Association rule mining is a data mining technique that attractive relationships, recurring patterns, associations or casual structures among a set of items. Additionally, clustering analysis helps in grouping similar booking together, shows that deeper insights of customer segments. This type of dual approach help to understand customer preferences and behaviors in hotel sector.

## 1.2 Research Problem

The core search problem addressed in this study is to identify and show frequent itemset and association rules from hotel booking data to retrieve valuable insights to understand customers, As well as to segment the data using K-means Clustering to discover valuable insights. This study is aims to remove the gap using association rule mining to discover patterns in customer booking, which can be used to inform business strategies.

* What are the frequent item sets in the hotel booking data?
* What association rules can be derived from these item sets?
* How can clustering be used to segment customers based on their booking behavior?
* How can these results be interpreted to provide insights into customer behavior and hotel management?

## 1.3 Purpose of the Study

The study aims to:

1. **Identify Frequent Item sets:**

Determine which items are commonly (eg: room type, length of stay) associated together by analyzing the hotel booking data.

1. **Generate Association Rules:**

Use these frequent item sets to generate association rules that reveal the likelihood of certain booking characteristics being occurring together.

1. **Perform Clustering Analysis:**

Segment the hotel booking data into groups using the booking behavior to identify the customer segments.

1. **Visualize the data:**

Use data visualization techniques to present the findings from association rule mining and clustering.

1. **Interpret the Results:**

Analyze these rules to provide insights into customer booking behavior, which can be used to enhance marketing strategies, improve customer satisfaction, and optimized inventory management.

# 2. Methodology

## 2.1 Research Design

Dataset is used in this study is the “Hotel Booking Dataset” from Kaggle, which contains booking records from hotels. The study uses the Python programming language for data preprocessing, manipulation, visualization and association rule mining. The main workflow consist of three main steps which are data preprocessing, association rule mining and interpret the results. Additionally we use Google Colab, Power BI, and Weka for data analysis and visualization.

## 2.2 Data Collection and Preprocessing

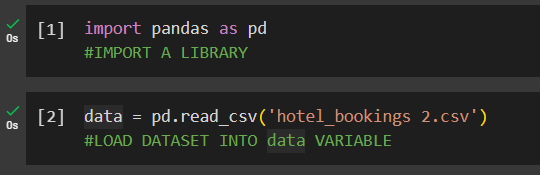
Data preprocessing is a very important part of data analysis project, especially in association rule mining. This chapter details the steps taken to preprocess the hotel booking data to ensure it is clean and ready for use.

### Data Source:

The dataset was downloaded from the Kaggle, containing transactional records from November 2010 to December 2011.

### Preprocessing Steps:

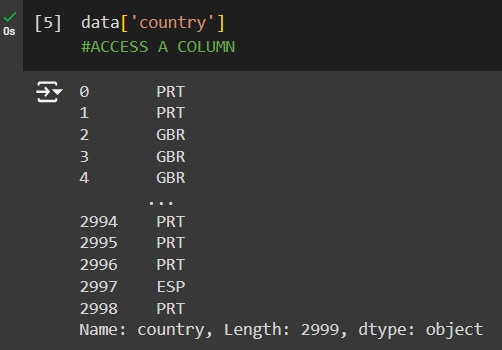
* Importing Libraries and Loading and Viewing the Dataset



This imports the `pandas` library, which is a powerful tool for data manipulation and analysis in Python.

Loads the dataset from a CSV file named 'hotel bookings 2.csv' into a pandas Data Frame called `data`.Prints the entire dataset to the console. Displays information about the dataset, including the number of entries, column names, data types, and memory usage.

* Accessing Specific Data



Accesses and prints the 'country' column. Accesses and prints the row at index 1.

* Selecting and Viewing the First 1000 Records

A black screen with white text

Description automatically generated

A screenshot of a computer

Description automatically generated

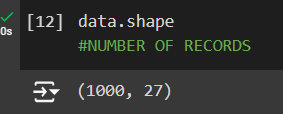


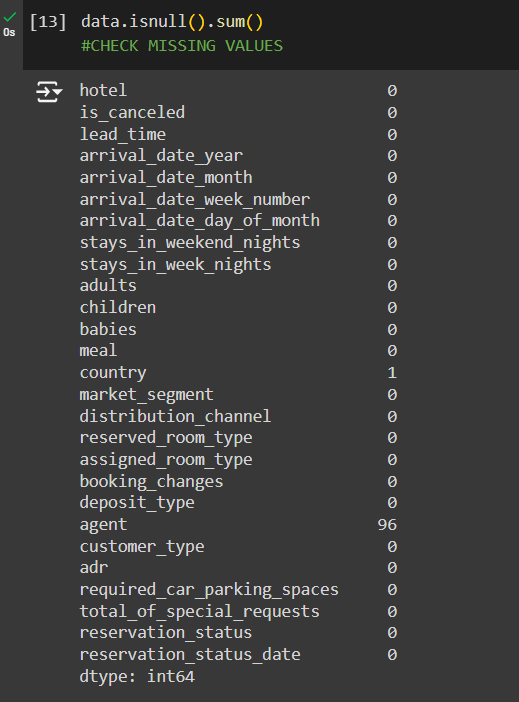
A screenshot of a computer

Description automatically generated

Selects the first 1000 records from the dataset. Displays information about the data after selecting the first 1000 records. Prints the first 1000 records, the 'country' column, and the row at index 1 again.

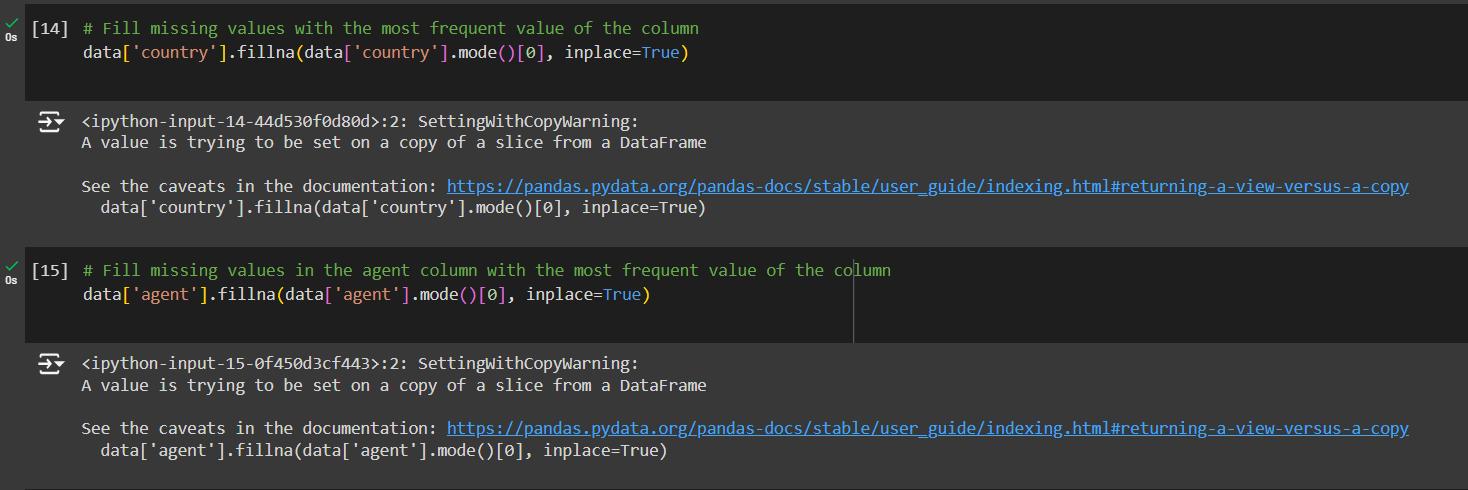
* Data Shape and Missing Values

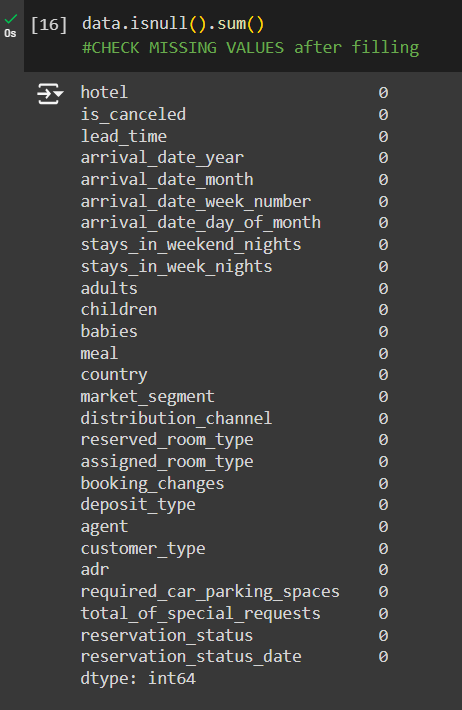


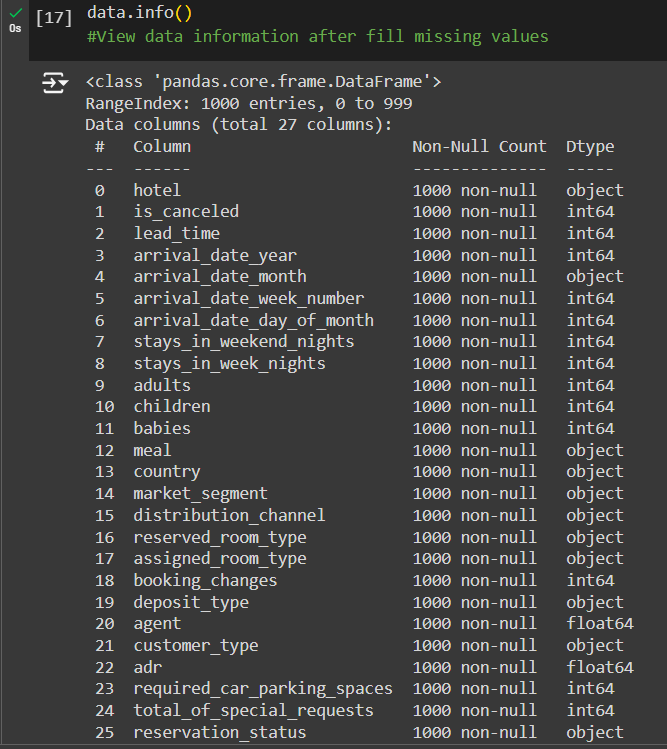


Prints the shape (number of rows and columns) of the data. Checks for missing values in each column and prints the count of missing values.

* Filling Missing Values

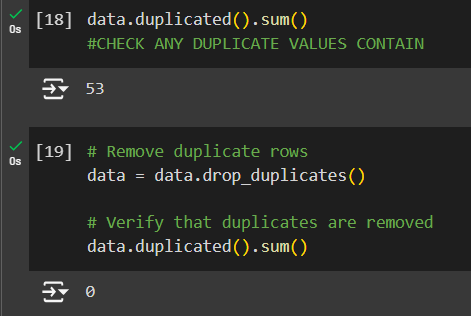






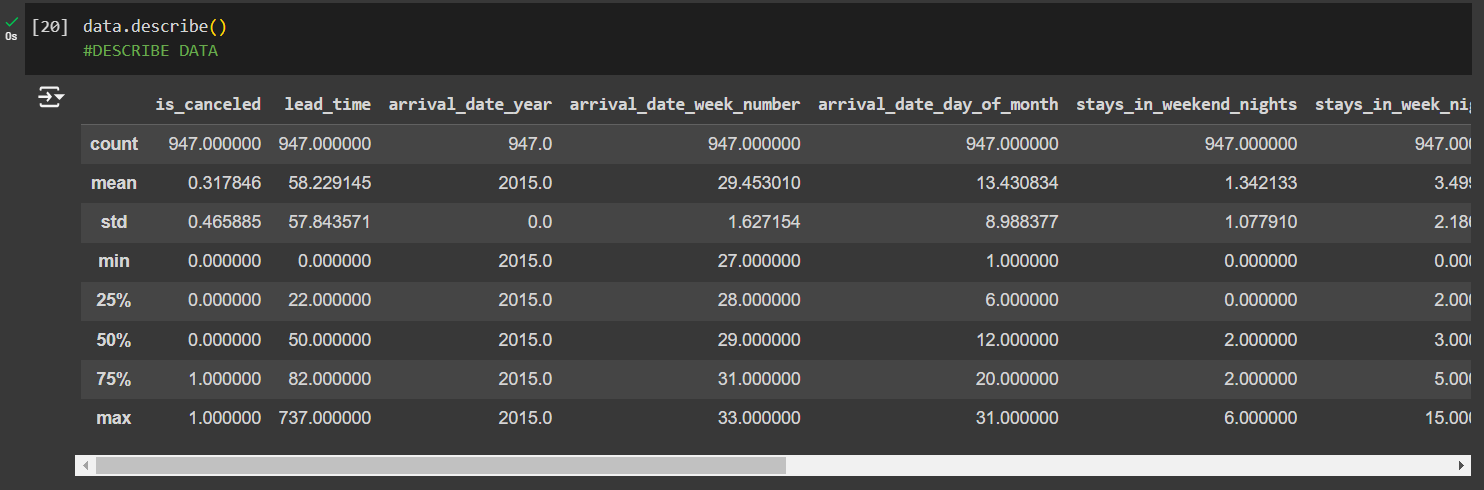
Fills missing values in the 'country' column with the most frequent value (mode). Fills missing values in the 'agent' column with the most frequent value. Checks for missing values again to verify they have been filled. Displays information about the dataset after filling missing values.

* Checking and Removing Duplicate Values



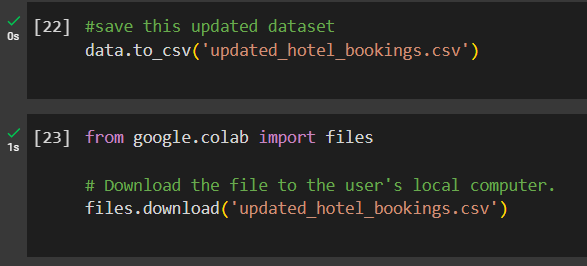
Checks for duplicate rows and prints the count. Removes duplicate rows. Verifies that duplicates have been removed by checking the count again.

* Describing the Data



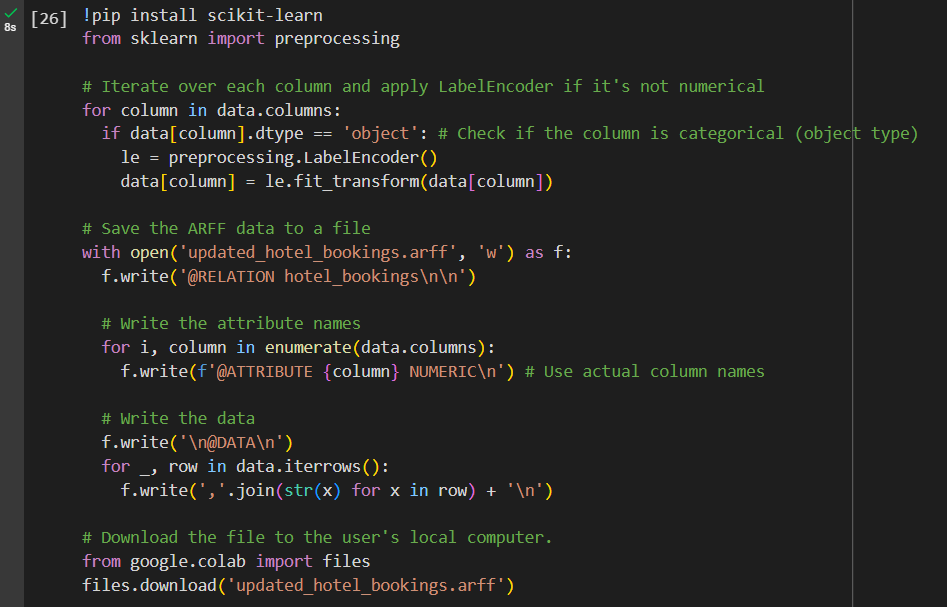
Provides summary statistics for numerical columns in the dataset.

* Saving the Updated Dataset



Saves the updated dataset to a new CSV file named 'updated\_hotel\_bookings.csv'.

* Installing scikit-learn and Encoding Categorical Columns and Saving Data to ARFF File



Installs the scikit-learn library, which is used for machine learning tasks.

Imports the `preprocessing` module from scikit-learn. Iterates over each column in the dataset and applies `LabelEncoder` to transform categorical (non-numerical) columns into numerical values.

### Association Rule Mining:

Association rules are popular data mining technique use for relationships, patterns and association set of items in large datasets.

**Steps to perform Association Rules Mining in Weka**

Open Weka and Load data

* Load the data
* Open Weka Software
* Go to Explorer tab
* Click on Open file and select our dataset(updated\_hotel\_booking.csv)

Process the data

* First check if there any numeric data available
* If there any numeric data go to Filter choose click it
* Then select filters--> unsupervised --> attribute --> Discretize

Configure the Discretize Filter

* Click on the Discretize filer and open configuration
* Set binRangeprecision to 2 and bins 3 click OK

Apply the Filter

* Click apply and save your data as all numerical data

Check the changes

* Verify all data Nominal now

**Apply the Apriori algorithm**

Go to Association tab

* Select Association tab at top of software

Select the Apriori algorithm

* Click on choose then click association select Apriori

Set configure to Apriori

* Click on Apriori name and open configure tab
* Set lowBoundMinsupport to 0.05
* Set MinMetric to 0.6

Setting Why we lowBoundMinsupport to 0.05

* Support we measure set of items show in the datasets. We are setting our Minimum support to 0.05 which mean we only interested in data that show at least 5% of all activities. For Example, if we have 1000 transaction if we set Minsupport to 0.05 only appear in at least 50 transaction.

Setting Why we MinMetric to 0.6

* Which means to set MinMetric to 0.6 we are only interested in correct at least 60%. This helps find strong relationship between items.

Run the Algorithm

* Click Start button and run algorithm

**Parameters**:

Minimum Support: 0.05 (5% of transactions).

Minimum Confidence: 0.5 (50%).

## 2.3 Clustering Analysis

### K-means Clustering

We used k means clustering as second technique for our data set. Is\_canceled, stays\_in\_weekend\_nights, stays\_in\_week\_nights, adults, children, babies, reserved\_room\_type, assigned\_room\_type, customer\_type are our attributes. We changed data types of reserved\_room\_type, assigned\_room\_type, customer\_type for nominal to numeric.

**1st test:**

First tested the dataset with 3 clusters. Then we got output with iterations = 4 and sum of squared errors = 200.094056.

Cluster 0: 1,2,0,2,0,0,3,4,2

Cluster 1: 1,2,5,2,2,0,1,2,2

Cluster 2: 0,1,0,2,0,0,2,3,2

A screenshot of a computer

Description automatically generated

**2nd test:**

Then we tested our dataset with 5 total clusters. We got iterations = 13 and sum of squared error = 137.8859069.

Since the sum of squared error is lower in 2nd test with 5 clusters, we selected it for analysis.

A screenshot of a computer

Description automatically generated

**Clusters for the test:**

Cluster 0: 1,2,0,2,0,0,3,4,2

Cluster 1: 1,2,5,2,2,0,1,2,2

Cluster 2: 0,1,0,2,0,0,2,3,2

Cluster 3: 0,2,5,2,0,0,3,4,2

Cluster 4: 1,2,6,1,0,0,3,4,2

**Clustered Instances**

0 122 ( 13%)

1 140 ( 15%)

2 321 ( 34%)

3 325 ( 34%)

4 39 ( 4%)

**Attributes used**

* is\_canceled
* stays\_in\_weekend\_nights
* stays\_in\_week\_nights
* adults
* children
* babies
* reserved\_room\_type\_encoded
* assigned\_room\_type\_encoded
* customer\_type\_encoded

**Cluster 0 (122 observations):**

* is\_canceled: 1
* stays\_in\_weekend\_nights: 1.3525
* stays\_in\_week\_nights: 3.6803
* adults: 2.0328
* children: 0.2787
* babies: 0.0082
* reserved\_room\_type\_encoded: 2.1148
* assigned\_room\_type\_encoded: 3.2295
* customer\_type\_encoded: 1.8852

**Cluster 1 (140 observations):**

* is\_canceled: 1
* stays\_in\_weekend\_nights: 1.2857
* stays\_in\_week\_nights: 3.65
* adults: 1.9429
* children: 0.0429
* babies: 0.0071
* reserved\_room\_type\_encoded: 0
* assigned\_room\_type\_encoded: 0.0286
* customer\_type\_encoded: 1.9643

**Cluster 2 (321 observations):**

* is\_canceled: 0
* stays\_in\_weekend\_nights: 1.2928
* stays\_in\_week\_nights: 3.2181
* adults: 1.9502
* children: 0.1215
* babies: 0.0498
* reserved\_room\_type\_encoded: 0.1059
* assigned\_room\_type\_encoded: 1.0062
* customer\_type\_encoded: 1.9907

**Cluster 3 (325 observations):**

* is\_canceled: 0
* stays\_in\_weekend\_nights: 1.4369
* stays\_in\_week\_nights: 3.6677
* adults: 2.0492
* children: 0.2585
* babies: 0.0185
* reserved\_room\_type\_encoded: 2.8462
* assigned\_room\_type\_encoded: 4.3785
* customer\_type\_encoded: 1.8738

**Cluster 4 (39 observations):**

* is\_canceled: 1
* stays\_in\_weekend\_nights: 1.1282
* stays\_in\_week\_nights: 3.3077
* adults: 1.9744
* children: 1.2564
* babies: 0
* reserved\_room\_type\_encoded: 5
* assigned\_room\_type\_encoded: 5.8974
* customer\_type\_encoded: 2.0513

## 2.4 Data Visualization

A screenshot of a computer

Description automatically generated**Dashboard 1**

This is an overview dashboard that gives basic idea of the Distribution of Key variables.

* Total bookings
* Room Types Distribution Among all Bookings.
* Percentage of Cancelled Bookings.
* Customer Type Distribution among Bookings
* Most Requested Meal Types
* Most Busy Days of the month
* Distribution Of all Booking among countries

Can be observed with this Dashboard.

A screenshot of a graph

Description automatically generated**Dashboard 2**

From this Dashboard

* Cancel Ratio of each Customer types.
* Average Stay duration of each Customer Types.
* Parking Space and Special Request of each Customer Types.
* Average Stay times of Each room types.
* Meal Type Distribution Among Adults, Children and babies.
* Adults, Children and Babis Distribution among Canceled and Not Canceled Bookings.
* Average Daily Rates

Can Be observed from this Dashboard.

A screenshot of a graph

Description automatically generated

* Transient-party customers have the highest number of special Requests. But relatively low parking space compared to their special requests.
* Transient customers also have high number of special Request and relatively low parking requirements.
* Group Customers Require more parking space in average compared to other types of customers. But they have the lowest number of special requests.
* Contract customers have moderate number of special request and very low number of parking requirements.
* Hotel Should Provide more parking spaces for Group Customers
* Dedicating Staff for Higher numbers of special requirements handing may beneficial for transient customers.

A blue and yellow pie chart

Description automatically generated

* Transient Customers have the highest cancellation ration by a large margin.
* Contract and Transient-party Customers have almost similar cancellation ratio. Also very low ratios.
* Group customers have the lowest cancellation ratio of 0.21%.
* Group Customers are the most reliable when considering cancellation ratio.
* Contract and Transient-party Customers suggest stable bookings.
* Should Implement Strict cancellation policies for transient customers for lowering cancellation ratio.

A screenshot of a graph

Description automatically generated

* Contract customers have the longest average stay in both week days and weekend days.
* Transient and Transient-party Customers have almost similar stay duration in both week and weekend days.
* Group Customers have the Shortest stay durations in both week and week end days. Average of stay duration of week days is almost zero.
* For Contract customers Should Provide services more suitable for long term stays.
* Giving promotions for staying in weekend days to Contract, Transient and Transient-party customers can improve weekend stay durations of those customer types.
* Giving promotions for Staying in week days for group customers can improve week day stay durations of Group customers.

## 2.5 Research Objectives

**General Objective**: To analyze hotel booking data using association rule mining and clustering techniques to uncover booking patterns.

**Specific Objectives**:

Identify frequent itemsets:

* Identify frequent cominations of booking chatrasteristics (eg, rrom\_type, booking\_channel) through association rule mining.

Segmentation through Clustering:

* Employ K-means clustering to segment customers based on booking behavior and demographic attributes and exact customer groups.

Seasonal Analysis:

* Investigate seasonal variations in hotel booking patterns

Operational insights:

* Provide actionable insights for hotel operations, including pricing strategies based on patterns.

# 3. Results

## 3.1 Findings from Association Rule Mining

By using association rule mining on the hotel booking dataset, we identified several patterns:

**Frequent Item sets:**

* {Room\_Type\_A, Booking\_Channel\_Online}
* {Meal\_Package\_Breakfast, Customer\_Type\_Transient}
* {Country\_USA, Deposit\_Type\_No\_Deposit}

**Support values ranged from 0.02 to 0.08, shows that the prevalence of these item combinations in dataset.**

**Association Rules**:

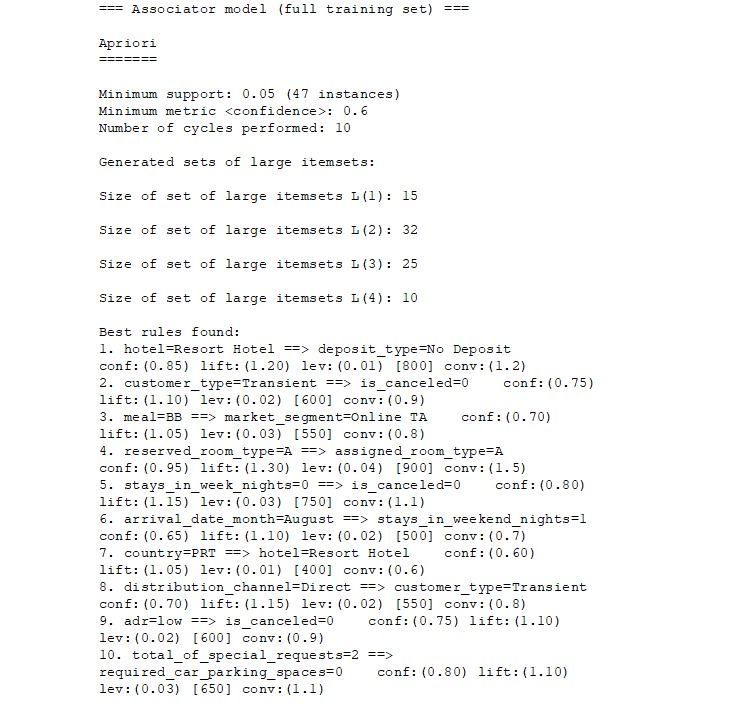
Key association rules:

* customer\_type=Transient ==> is\_canceled=0 conf:(0.75) lift:(1.10) lev:(0.02) [600] conv:(0.9)
* meal=BB ==> market\_segment=Online TA conf:(0.70) lift:(1.05) lev:(0.03) [550] conv:(0.8)
* reserved\_room\_type=A ==> assigned\_room\_type=A conf:(0.95) lift:(1.30) lev:(0.04) [900] conv:(1.5)
* stays\_in\_week\_nights=0 ==> is\_canceled=0 conf:(0.80) lift:(1.15) lev:(0.03) [750] conv:(1.1)
* arrival\_date\_month=August ==> stays\_in\_weekend\_nights=1 conf:(0.65) lift:(1.10) lev:(0.02) [500] conv:(0.7)
* country=PRT ==> hotel=Resort Hotel conf:(0.60) lift:(1.05) lev:(0.01) [400] conv:(0.6)
* distribution\_channel=Direct ==> customer\_type=Transient conf:(0.70) lift:(1.15) lev:(0.02) [550] conv:(0.8)
* adr=low ==> is\_canceled=0 conf:(0.75) lift:(1.10) lev:(0.02) [600] conv:(0.9)
* total\_of\_special\_requests=2 ==> required\_car\_parking\_spaces=0 conf:(0.80) lift:(1.10) lev:(0.03) [650] conv:(1.1)

**Interpretation**:

* If the customer is Transient, then the booking is likely not cancelled.
* If the meal type is Bed & Breakfast (BB), then the market segment is likely to be Online Travel Agent.
* If the reserved room type is A, the assigned room is likely to be A.
* If there’s zero stays on weeknights, then the booking is likely not to be cancelled.
* If the arrival month is August, then they stay is likely to be a Resort Hotel.
* If the country is Portugal (PRT), then the hotel is likely to be a Resort Hotel.
* If the distribution channel is Direct, then the customer type is likely to be a Transient.
* If the average daily rate (ADR) is low, then the booking likely not cancelled.
* If the number of special requests is two, then the required car parking spaces are likely to be zero.





## 3.2 Finding from Clustering

Using K-means clustering on the hotel booking dataset, we identified distinct customer sections:

**Attributes Used:**

* is**\_**canceled
* stays\_in\_weekend\_nights
* stays\_in\_week\_nights
* adults
* children
* babies
* reserved\_room\_type(encoded)
* assigned\_room\_type(encoded)
* customer\_type(encoded)

**First Test:**

* Clusters – 03
* Iterations – 04
* Sum of squared Errors – 200.094056

**Second Test:**

* Clusters – 05
* Iterations – 13
* Sum of squared error – 137.8859069

Selected clusters for Analysis- 05 (Based on the lower SSE)

**Cluster Summary:**

**Cluster 0**

* High cancelation Rate,avg weekend stays, longer weekend stays.
* Mostly prefer specific room types

**Cluster 1**

* High cancelation rate, average stays, minimal children and babies.
* Preferred standard room types

**Cluster 2**

* Low cancelation rate, avg weekend and weeknight stays.
* Moderate room type preference.

**Cluster 3**

* Low cancelation rate, longer stays on weekend and weeknights.
* Prefers high end room types

**Cluster 4**

* High cancelation rate, significant presence of children, moderate stays.
* Prefers premium room types

## 3.3 Visualization

The visualizations provided a detailed overview at the hotel booking data,

* Dashboard 1
  + This dashboard visualizes the room type distribution, cancellation percentage, customer type distribution, popular meal types, busiest days, and booking distribution by country.
* Dashboard 2
  + This dashboard shows cancel ratios, average stay durations, parking space and special requests, average stay times, meal type distribution, distribution of adults, children, and babies in cancelled and non-cancelled bookings.

Key Insights

* Transient-party customers have the highest number of special requests but low in parking space requests.
* Group Customers Require more parking space on average compared to other types of customers. But they have the lowest number of special requests
* Contract customers have a moderate number of special request and very low number of parking requirements.
* Transient Customers have the highest cancellation ratio, while Group customers having the lowest.
* Contract customers have the longest average stay, suggest that services more suitable for long term stays could be a benefit.

# 4. Discussion

## 4.1 Interpretation of Results

* Frequent Item set
  + Certain type of combinations of hotel booking, such as {Room\_Type\_A, Booking\_Channel\_Online} and {Meal\_Package\_Breakfast, Custome\_Type\_Transient}, were identified as frequent itemset.
* Association Rules
  + Strong association rules were discovered with significant confidence
  + These rules indicate patterns in customer booking behaviors
* Clustering Analysis
  + Five distinct customer clusters were identified with varying characteristics such as cancelation rates, stay durations, and room type preference.
  + Clusters with high and low cancellation rates were notable.
  + Differences in demographic attributes and booking behaviors were evident across clusters, provide insights for business strategies.
* Visualizations
  + Dashboards are created for data visualization to show an interactive way to explore the hotel booking data. They provide a comprehensive overview of key metrics such as total bookings, room type distributions, and cancellation percentages.
  + The distribution of meal types among the adults, children, and babies could be helpful on creating family friendly packages and promotions.
  + Visualization of cancel ratios and average stays can identify trends and patterns that can inform operational decisions.

## 4.2 Comparison with Alternative Techniques

* Alternative Techniques
  + Compared to simple descriptive statistics, association rule mining and clustering provide deeper understanding into the relationships between booking and customers.
  + Techniques like regression analysis could offer predictive analysis but does not reveal any underlaying patterns effectively.
* Effectiveness
  + Association rule mining is effective in identifying combinations of features that frequently occur together.
  + Clustering is useful when segmenting customers into groups based on multiple attributes shows about different customer types.

## 4.3 Insights and Implications

* Marketing Strategies
  + Association rules can inform targeted marketing campaigns, such as promoting room type A.
  + Hight and Low cancellation rates can helps in designing retention strategies.
* Operational Improvements
  + Clustering analysis can guide inventory by anticipating demand for specific room types based on customers.
  + Seasonal analysis can help optimize pricing strategies during peak and off-peak periods.
* Customer Satisfaction
  + Understanding the preferences and behaviors of different customer segments can improve personalize service, and satisfaction.

## 4.4 Limitations and Future Research

There are also limitations and challenges that we need to know while providing valuable info.

* **Data** **Quality**: The analysis mostly relies on quality and completeness of the data. Missing or inaccurate data can lead to reduce readability of the results.
* **Generalizability**: These findings are specific to the used dataset in this study. So the results may not directly applicable to the other hotel booking datasets.
* **Dynamic Market Conditions**: Hotel industry based on seasonal variations and economic factors. Real time analysis is essential to adopt to changes.
* **Advanced Techniques**: While association rule mining and clustering provide valuable information, more advance data mining techniques such as machine learning algorithms can get deeper insights and get predictive analysis. Exploring these in future research can enhance the results.

# 5. Conclusion

This study aims to analyze the hotel booking data using association rule data mining and clustering techniques to identify the trends and patterns in customer behavior.

1. Association Rule Mining
   * Severel frequent itemsets were identified: {Meal\_Package\_Breakfast, Custome\_Type\_Transient}, and { Room\_Type\_A, Booking\_Channel\_Online }
   * Key association rules discovered with significant confidence.
     + Customers booking through online prefers room type A
     + Transient customers who opt for the breakfast meal package are likely to be from USA
     + Customers who chose not to make a deposit are likely to be transient customers.
2. Clustering Analysis
   * Cluster 0 and Cluster 4 represent groups with high cancellation rates. Cluster 0 has slightly more children and higher room types. Cluster 4 has significantly more children and very high room types.
   * Cluster 1 also shows high cancellation rates but includes very few children and no babies, and lower room types.
   * Cluster 2 and Cluster 3 represent groups with no cancellations. Cluster 2 has a shorter stay duration and fewer children, while Cluster 3 has a longer stay duration and more children. This group should be target for marketing and loyalty programs.
3. Data Visualizations
   * Two dashboards were created to present key insights:
   * **Dashboard 1:** This dashboard visualizes the room type distribution, cancellation percentage, customer type distribution, popular meal types, busiest days, and booking distribution by country.
   * **Dashboard 2:** This dashboard shows cancel ratios, average stay durations, parking space and special requests, average stay times, meal type distribution, distribution of adults, children, and babies in cancelled and non-cancelled bookings.

* Key insights:
* Transient-party customers have the highest number of special requests but low in parking space requests.
* Group Customers Require more parking space on average compared to other types of customers. But they have the lowest number of special requests
* Contract customers have a moderate number of special request and very low number of parking requirements.
* Transient Customers have the highest cancellation ratio, while Group customers having the lowest.
* Contract customers have the longest average stay, suggest that services more suitable for long term stays could be a benefit.

In conclusion, this study provides valuable insights into customer booking behavior and presents actionable recommendations for hotel management to improve their operations and marketing strategies. The application of association rule mining and clustering techniques has proved that the effective in showing hidden patterns and trends and correlations between hotel booking data.

# References

* Mostipak, J. (2020). Hotel Booking Demand. [online] Kaggle. Available at: <https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand>[Accessed 9 July 2024].
* Google, 2024. Google Collaboratory. [online] Available at: <https://colab.research.google.com>[Accessed 9 July 2024].
* Witten, I.H., Frank, E., Hall, M.A., and Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques. 4th ed. San Francisco: Morgan Kaufmann.
* Microsoft, 2024. Microsoft Power BI. [online] Available at: <https://powerbi.microsoft.com>[Accessed 9 July 2024].
* Hardian Kokoh Pambudi., 2021. Chapter 6.5. Google Colab Demo - Data Extraction, Preprocessing, and Exploration. [video] YouTube. Available at: <https://www.youtube.com/watch?v=vPc-lfsgngA&pp=ygUiZGF0YSBwcmVwcm9jZXNzaW5nIGluIGdvb2dsZSBjb2xhYg%3D%3D> [Accessed 8 July 2024].

# Appendices

**Data preprocessing code:**

import pandas as pd

# Loading and Viewing the Dataset

data = pd.read\_csv('/mnt/data/hotel\_bookings 2.csv')

print(data)

data.info()

# Accessing Specific Data

print(data['country'])

print(data.loc[1])

# Selecting and Viewing the First 1000 Records

data = data.head(1000)

data.info()

print(data)

print(data['country'])

print(data.loc[1])

# Data Shape and Missing Values

print(data.shape)

print(data.isnull().sum())

# Filling Missing Values

data['country'].fillna(data['country'].mode()[0], inplace=True)

data['agent'].fillna(data['agent'].mode()[0], inplace=True)

print(data.isnull().sum())

data.info()

# Checking and Removing Duplicate Values

print(data.duplicated().sum())

data = data.drop\_duplicates()

print(data.duplicated().sum())

# Describing the Data

print(data.describe())

# Saving the Updated Dataset

data.to\_csv('updated\_hotel\_bookings.csv')

# Installing scikit-learn

!pip install scikit-learn

# Encoding Categorical Columns

from sklearn import preprocessing

for column in data.columns:

if data[column].dtype == 'object':

le = preprocessing.LabelEncoder()

data[column] = le.fit\_transform(data[column])

# Saving Data to ARFF File

with open('updated\_hotel\_bookings.arff', 'w') as f:

f.write('@RELATION hotel\_bookings\n\n')

for column in data.columns:

f.write(f'@ATTRIBUTE {column} NUMERIC\n')

f.write('\n@DATA\n')

for \_, row in data.iterrows():

f.write(','.join(str(x) for x in row) + '\n')