**IT2384 UNSupervised**

**Learning**

**ASSIGNMENT (40%)**

|  |  |
| --- | --- |
| **Submitted By:** | Jay Krish Vijendra (222197Y) |
| **Lecturer:** | Mr Kheng Kok MAR |
| **Deadline:** | 27/August/2023 |

# **Data Cleaning and Data Transformation**

The subsequent sections delineate the meticulous process of data cleansing and transformation applied to the "hotel data.xlsx" dataset in Python, in preparation for its utilization in the subsequent modelling using RapidMiner and SAS Viya.

## Import Libraries and Data

The initial phase involved importing the requisite libraries essential for data manipulation and analysis. Subsequently, the dataset was loaded into a data frame. The first five rows of the dataset were then retrieved to gain an initial insight into its structure and content.

A screenshot of a computer

Description automatically generated

The next phase involved going further into extensive dataset examination. Key information, including the number of columns and entries, was gathered by doing a thorough analysis of the data frame. The dataset had a large corpus of 52416 rows and 63 unique columns through ***df.info.***

## Data Understanding

Null value assessment was performed, revealing no instances of null values within the dataset.A screenshot of a computer program

Description automatically generated

## Data Cleaning

A computer screen shot of a black screen

Description automatically generated

Feature Removal: Certain columns were excluded from the analysis due to their limited relevance. Columns ['custID', 'custName','datePosted','hotelName','source','cust\_review\_title'] were deemed unnecessary for modelling purposes.

A screenshot of a computer program

Description automatically generated

The code removes duplicate rows from DataFrame (df) and displays the resulting clean DataFrame's contents. After dropping the multiple duplicate rows and dropping the above mentioned columns the dataset had 5825 rows and 57 columns.

Since there were Null Values from these columns :

A screenshot of a computer program

Description automatically generated['travel\_custer\_no','cust\_travel\_type', 'custCountry', 'star\_rating','comment']. Especially for comments as Null comments, especially for task 1, lack relevance for analysis since they translate to zeros across all concepts.

This is done to ensure that only complete and meaningful records are retained. This is vital for analyses and modelling, as missing data can lead to incorrect insights or biased results. Removing such rows with missing values enhances the reliability of subsequent analyses and ensures that the data used is consistent and reliable.

A screenshot of a computer screen

Description automatically generatedH The analysis focuses on extracting distinct values from the 'custCountry' and 'countryRegion' columns, pertaining to customer origins by regions and countries. An anomaly was detected during the review, where 'Meldives' was observed instead of the correct 'Maldives'.The discrepancy was to a single record, which was promptly amended to reflect the accurate spelling.

Furthermore, a consistency audit of the regional nomenclature highlighted that 'SEA' (South East Asia) was represented as an abbreviation different from the rest which had the spelling in full. In the interest of maintaining coherence, the abbreviated term was replaced with the full expression, 'South East Asia'. This guarantees precision and uniformity for our analytical insights.

A computer screen with white text

Description automatically generated

This code groups data by regions, capturing unique countries within each region. By doing this, we gain insights into which countries fall under specific regions, aiding in categorizing and understanding customer origins. The loop displays these insights for clear comprehension. There were many mistakes such as many Asian countries under South East Asia and countries wrongly grouped in their region.

A screenshot of a computer screen

Description automatically generatedA screen shot of a computer

Description automatically generated

A black background with colorful text

Description automatically generated

The left Image shows the countries which mapping had change and the right on which country kept constant. The dictionary 'country\_to\_region' is created, linking each country to its corresponding region. This mapping is used to replace 'countryRegion' values in the DataFrame with their appropriate regions. This helps organize and analyze data in a more meaningful way, providing clearer insights into customer origins and their regions. The DataFrame is then updated with this modified information, facilitating more accurate analysis and interpretation of results.

A computer screen with white text

Description automatically generated

The groupings are refined, including new regions like South Asia and East Asia to better accommodate corresponding countries.

## Exporting Data

A screenshot of a computer

Description automatically generated

After refining and enhancing the dataset through data cleaning and transformation, the next step is to save the cleaned data for modelling. This is accomplished by exporting the cleaned Data Frame into a new CSV file named "Cleaned\_Hotel\_Data.csv". After the Data Cleaning the dataframe has 4751 Rows and 57 Columns.

# **Task 1 Association Rule Mining (20%)**

## Modelling

A screenshot of a computer

Description automatically generated

The preceding section provides an overview of the RapidMiner design sequence, encompassing actions such as reading CSV data, implementing FP-Growth, and generating association rules. For the Read CSV Node, I initially selected columns beginning with 'concept' and then converted the chosen inputs into a binomial format. For FP-Growth it was set with min items per itemset being 2 and min support starting from 0.02. For Create Association Rules it has a min confidence starting from 0.3 with a gain theta at 2.0 and Laplace k value of 1.0.

When crafting association rules, it's important to aim for high support and confidence values, indicating strong relationships. However, finding the right balance often involves trial and error. Setting these thresholds excessively high might yield no rules, while setting them too low can lead to prolonged computation times. Experimenting with different confidence values is recommended to strike the right balance and attain meaningful insights.

|  |  |  |
| --- | --- | --- |
| Min Support | Min Confidence | Top 10 association rules |
| 0.02 | 0.3 |  |
| 0.2 | 0.5 |  |
| 0.3 | 0.6 |  |
| 0.3 | 0.7 |  |

## Evaluation

The selection is a minimum support value of 0.3 and a minimum confidence value of 0.7 as parameters for generating association rules from the dataset. These choices were made to ensure a meaningful balance between the frequency of occurrence of items and the strength of the inferred relationships between them. A minimum support of 0.3 implies that only associations that appear in at least 30% of the transactions are considered, thereby filtering out relatively fewer common associations. Similarly, a minimum confidence of 0.7 indicates that the rules should have a confidence level of at least 70%, ensuring that the conclusions drawn from the relationships are reasonably accurate.

The presented association rules depict the top 10 relationships that meet these criteria. For example, the rule [Concept\_hotel, Concept\_staff] --> [Concept\_nearby] with a confidence of 0.898 signifies that when both "Concept\_hotel" and "Concept\_staff" are mentioned, there's a high confidence of 89.8% that "Concept\_nearby" will also be mentioned. This suggests a strong connection between positive feedback about the hotel and its staff with the proximity to nearby attractions or amenities.

In essence, the chosen thresholds and the showcased association rules aim to reveal noteworthy insights while upholding reliability. Adjusting these thresholds allows for a tailored exploration of associations, striking the right balance between comprehensiveness and accuracy in the identified relationships.

Based on this evaluation, the following recommendations are put forth:

1. Refinement of Thresholds: Continue exploring the impact of different threshold values for minimum support and minimum confidence. Fine-tuning these parameters could potentially reveal additional hidden patterns that might have been missed with the current settings.
2. Segmented Analysis: Consider segmenting the dataset based on specific attributes like customer demographics, travel purpose, or star ratings. This approach could unveil nuanced associations that are unique to different customer segments.

# **Task 2 Clustering (20%)**

## Modelling

A close-up of a graph

Description automatically generated

This model employs the K-prototypes clustering algorithm with a gamma value of 0.5 to handle mixed data types. It aims to uncover underlying patterns in the dataset by partitioning it into 5 clusters, using Euclidean distance and the "Cluster Change" stop criterion. The initialization is based on the "Forgy" method, with standardization, interval imputation, and nominal imputation not applied.

Iteration History:

With each iteration, the sum of distances between data points and their assigned centroids decreases. This reduction indicates the model's refinement in capturing data patterns. The "Distance Change" column represents the improvement in distances between iterations, while the "Stop Criterion" column indicates when the algorithm converges, as further iterations show negligible change. This process signifies the model's gradual convergence towards a solution that effectively groups data points, optimizing the clustering outcome.

A diagram of a graph

Description automatically generated with medium confidence

The chosen parameters for the parallel coordinates plot are designed to enhance the visualization of multivariate data. With 10 bins, the numerical values of each variable are discretized, offering insights into their distribution. By limiting the maximum number of polylines to 3,000, clutter is mitigated in cases of larger datasets. Five visible roles ensure that key attributes are represented along the vertical axis, optimizing the plot for comprehensively capturing relationships and patterns across multiple variables.

## Evaluation

The provided cluster data reveals distinct customer segments within the dataset based on their attributes. Cluster 2 primarily comprises couples who have given high star ratings, mainly originating from South Asia. Cluster 3 represents a diverse mix of customer types, including business travellers and couples, with varying star ratings from different regions. Cluster 4 is characterized by families with young children, displaying moderate star ratings and associations with East Asia and Oceania. Cluster 5 encompasses business travelers who provide high star ratings, primarily hailing from South East Asia. Lastly, Cluster 1 accommodates various customer groups with high star ratings, predominantly from North America and South East Asia. This clustering analysis enables the identification of customer preferences, geographical regions of origin, and potential target groups for personalized marketing and services. A comprehensive evaluation of these clusters can facilitate strategic decision-making, aiding in tailoring services and experiences to the specific needs of each segment for enhanced customer satisfaction and loyalty.

Notably, "Couples" stand out as the predominant customer type, making up a significant portion of the dataset. Their preferences are evident across regions like "South East Asia" and "Europe." In contrast, "Solo travelers" tend to hail from "South Asia" and "Europe." Geographically, the dataset is prominently populated by customers from "South East Asia," followed by "Europe" and "North America." These insights offer actionable information for tailored marketing strategies and service enhancements. By focusing on the prevalent customer segments and target regions, businesses can effectively enhance customer engagement and satisfaction, resulting in improved overall performance.

## **Conclusion**

In conclusion, the data cleaning and transformation process undertaken in this study has laid meaningful analysis and insights. The initial steps of importing libraries, examining the dataset's structure, and removing irrelevant features set the stage for a focused investigation.

Following association rule mining and clustering analysis, crucial information on client preferences and segmentation was made available. Targeted marketing strategies were made possible by the extraction of relevant and reliable connections and by defining acceptable thresholds for support and confidence in association rule mining.

However, there are a number of ways to improve the analysis that can be investigated. The insights will gain depth through comparative comparison with industry benchmarks, methods of validation to support credibility, and a deeper investigation of the practical effects of recommendations. Continuous progress will be facilitated by addressing drawbacks and suggesting new approaches, such as more sophisticated clustering algorithms or customer surveys.

Overall, this report provides actionable advice for strategic decision-making in addition to revealing interesting patterns and linkages within the data. The hotel can improve customer experiences, improve its products, and position itself for long-term success in the competitive hospitality industry by putting these findings into practice.