UNSUPERVISED LEARNING ASSIGNMENT

Data Preparation, Exploration, and Modelling of Hotel Review Dataset using Association Rule Mining and Clustering Techniques

Name	Tia Isabel Solanki
Admin No.	220892L
Module	IT2384
Date of Submission	25/8/2024

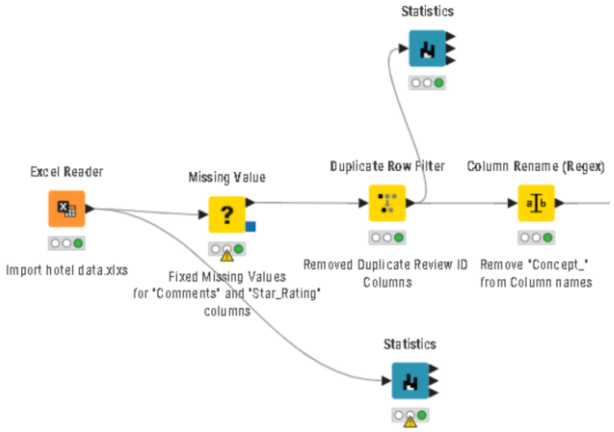
Table of Contents

TASK 1

DATA PREPARATION

Original Dataset: 52416 Rows, 63 Columns

Data Cleaning

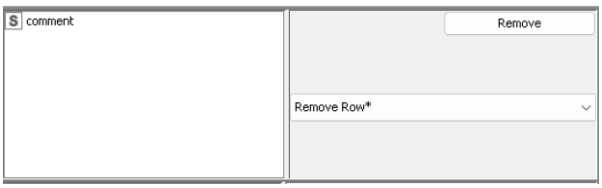


1. Checked and made sure all data types are correct

Name	Type
reviewID	Number (integ...
custID	Number (integ...
custName	String
cust_travel_ty...	String
travel_custer_...	Number (integ...
custCountry	String
countryRegion	String
datePosted	Local Date
cust_review_ti...	String
star_rating	Number (doub...
comment	String
hotelName	String

Name	Type
Concept_small	Number (integ...
Concept_place	Number (integ...
Concept_easy...	Number (integ...
Concept_expe...	Number (integ...
Concept_wifi	Number (integ...
Concept_like	Number (integ...
Concept_night	Number (integ...
Concept_helpf...	Number (integ...
Concept_pool	Number (integ...
Concept_probl...	Number (integ...
Concept_sing...	Number (integ...
Concept_com...	Number (integ...

2. Fixed Missing Values using the Missing Values Node
   a. <u>Removed Rows with no comments:</u> Since the analysis of the words in the comments are an essential part of this task, rows without comments were removed. No. of Rows reduced to 42795.



(Figure: Missing Value Node – Comments Column)

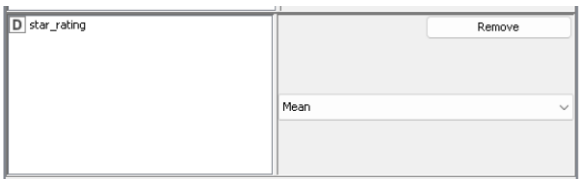
Name	Type	# Missing val...
comment	String	1476

Name	Type	# Missing val...
comment	String	0

(Figures: Statistics View – Comments Column Before and After)

   b. <u>Imputed star_rating column</u>

(Figure: Missing Value Node – star_rating_column)

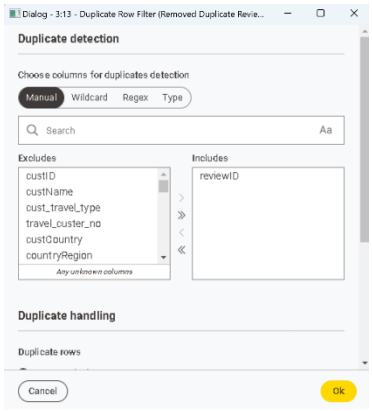


Name	Type	# Missing val...
star_rating	Number (doub...	9

Name	Type	# Missing val...
star_rating	Number (doub...	0

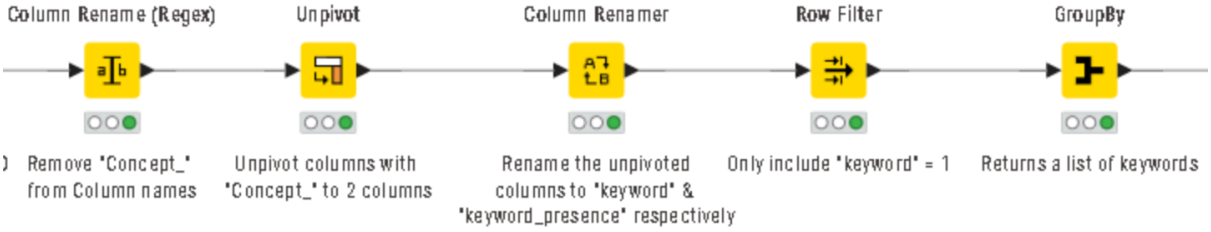
(Figure: Statistics View – star_rating column Before and After)

3. Removed Duplicate Rows with duplicate Row IDs: As there can only be 1 review per review ID, I check for and removed duplicate row IDs. The no. of rows reduced to 4755.
(Figure: Duplicate Row Filter Node)

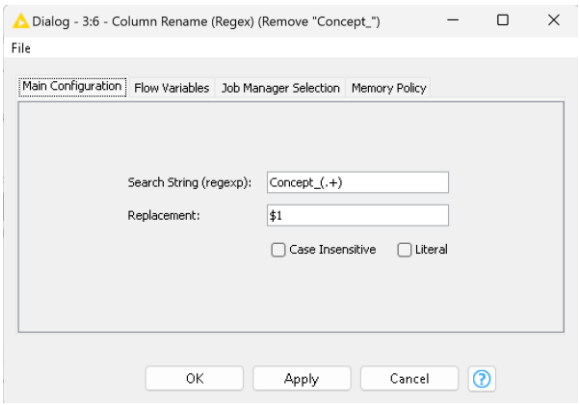


4. Removing Irrelevant Features: No need to remove irrelevant features, as the Association Rule learner will be applied later to produce the desired output.

Data Transformation



5. Removed the "Concept_" prefix from column names using Column Rename (Regex) node.
(Figure: Column Rename (Regex) Node)



Name
Concept_bad
Concept_small
Concept_place
Concept_easy...
Concept_expe...
Concept_wifi
Concept_like
Concept_night
Concept_helpf...
Concept_pool
Concept_probl...
Concept_sing...

Name
small
place
easy to access
expensive
wifi
like
night
helpful
pool
problem
singapore
comfortable

(Figures: Statisitcs View Before and After – Columns with "Concept_")

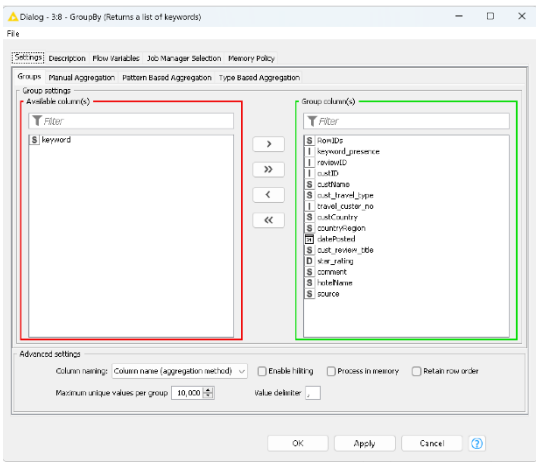
6. Unpivoted columns with the "Concept_" prefix and renamed the unpivoted columns to "keyword" and "keyword_presence" respectively.

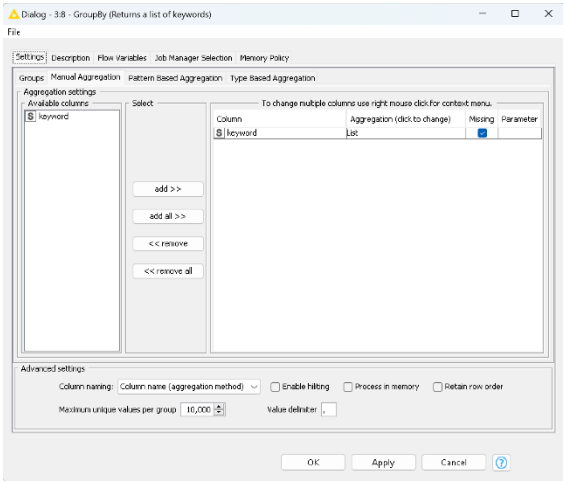
7. Applied a Row Filter to retain only rows where "keyword_presence" equals 1.

keyword String	keyword_... Number (inte...
bed	0
easy	0
bad	0
small	0
place	0
easy to access	0
expensive	0
wifi	0
like	0
night	0
helpful	0
pool	0
problem	0
singapore	0
comfortable	0
old	0
friendly	0
pleasant	0
would recom...	0
satisfied	0
hotel	0
quiet	0
service	0
location	0
excellent	0

keyword String	keyword_... Number (inte...
stay	1
easy to access	1
singapore	1
excellent	1
friendly	1
hotel	1
excellent	1
good	1
stay	1
hotel	1
excellent	1
pool	1
satisfied	1
good	1
staff	1
night	1
pool	1
comfortable	1
hotel	1
excellent	1
good	1
price	1
restaurants	1
stay	1
shopping	1

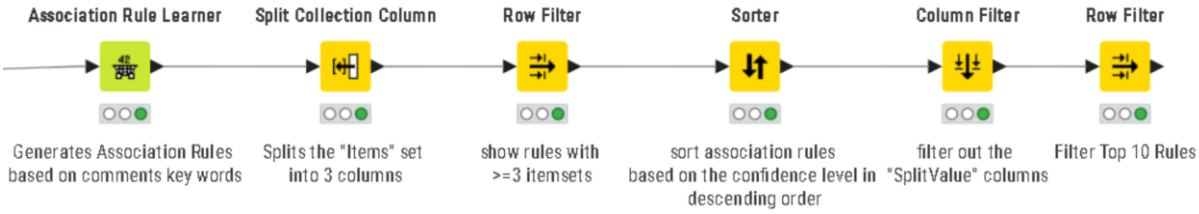
8. Used the GroupBy node to compile a list of keywords for each review_id. The keywords are aggregated into a list to put into the Association rule learner.





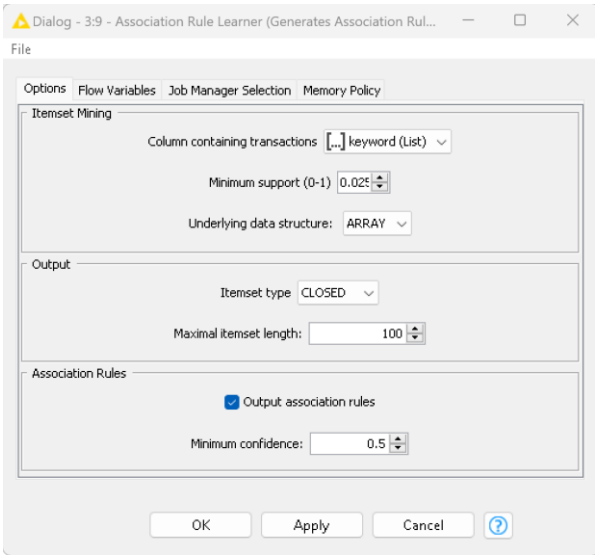
MODELLING PROCESS & TECHNIQUES

Association Rule Mining using the Association Rule Learner Node



Association Rule Mining

9. Used the Association Rule Learner Node to generate association rules from the dataset, aiming to discover significant relationships between variables. The parameters were set with a support threshold of 0.025 and a confidence threshold of 0.1 to balance the trade-off between rule specificity and generality. I set the maximal itemset length to 100 so that I can get every possible set.

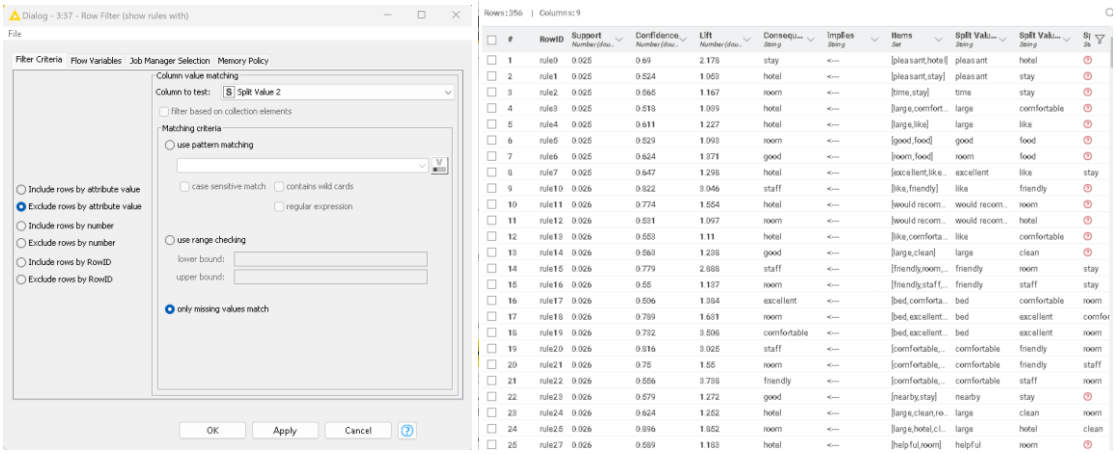


Filtering rows to achieve minimum of 3 items sets

10. Split the column to create three separate columns, each representing an item in the list.

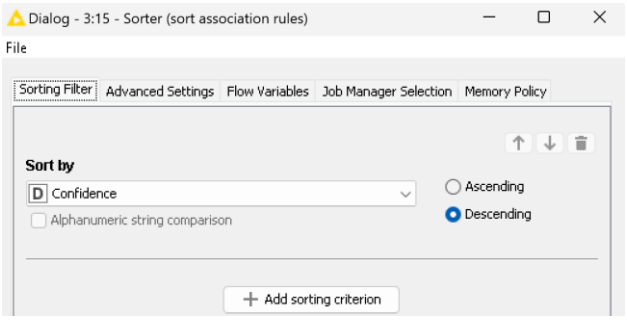
Consequ... String	Implies String	Items Set	Split Valu... String	Split Valu... String	Split Valu String
stay	<---	[pleasant,hotel]	pleasant	hotel	
hotel	<---	[pleasant,stay]	pleasant	stay	
room	<---	[time,stay]	time	stay	
hotel	<---	[large,comfort...	large	comfortable	
hotel	<---	[large,like]	large	like	
room	<---	[good,food]	good	food	
good	<---	[room,food]	room	food	
hotel	<---	[excellent,like...	excellent	like	stay
room	<---	[easy to acce...	easy to access		
free	<---	[wifi]	wifi		
staff	<---	[like,friendly]	like	friendly	
hotel	<---	[would recom...	would recom...	room	

11. Filter rows where SplitValue2 has missing values to ensure only itemsets with at least three items are included.



To ensure a minimum of three items per itemset, we exclude any rows where SplitValue2 or SplitValue3 is empty, as these indicate the itemset contains fewer than three items.

12. Applied the Sorter Node to rank the generated association rules based on their confidence levels in descending order. This step prioritized rules with the highest likelihood of being accurate, facilitating the identification of the most reliable patterns.



13. Used the **Row Filter Node** to extract and retain the top 10 association rules with the highest confidence scores with minimum 3 item sets.

Rows: 10 | Columns: 6

#	RowID	Support *Number (double)*	Confidence *Number (double)*	Lift *Number (double)*	Consequent *String*	implies *String*	Items *Set*
1	rule39	0.027	0.94	1.942	room	<---	[small,clean]
2	rule85	0.028	0.929	1.92	room	<---	[roomy,hotel]
3	rule2...	0.041	0.911	1.882	room	<---	[large,clean]
4	rule1...	0.031	0.9	3.335	staff	<---	[friendly,helpful]
5	rule25	0.026	0.896	1.852	room	<---	[large,hotel,clean]
6	rule2...	0.041	0.882	1.823	room	<---	[small,good]
7	rule78	0.028	0.878	1.815	room	<---	[clean,good,stay]
8	rule80	0.028	0.878	1.815	room	<---	[comfortable,clean,g...
9	rule2...	0.043	0.877	1.813	room	<---	[large,comfortable]
10	rule1...	0.032	0.867	1.792	room	<---	[large,excellent,hotel]

MODEL EVALUATION

The results from the Association Rule Learner Node show the top 10 rules with the highest confidence rates with a minimum of 3 item sets that provide valuable insights into customer preferences in hotel reviews

Support & Confidence

The model identifies the top 10 association rules with confidence levels ranging from 0.867 to 0.94. This indicates that a high percentage of the records containing the antecedents also match the consequent, showing a strong and reliable relationship between the identified keywords and the consequent.

The support values are moderate ranging from 0.026 to 0.043, which means that these rules can be found in 2.7% to 4.3% of the entire dataset. While not extremely common, these patterns are still significant enough to be actionable given a large dataset.

Lift

The lift values are all greater than 1 and range from 1.792 to 3.335. The suggests that the model is at least 1.792 times and at most 3.335 times better as compared to a random-choice model, which confirms the strength of the associations.

Rule Interpretations

Many of the rules appear to describe the overall room experience with keywords like “clean”, “large”, and “comfortable”. This shows that customers consider these as significant factors for their hotel experience and that these factors are a priority for them

We can see the keyword “staff” appears in rules with antecedents like “friendly” and “helpful” in the 4th rule, highlighting that customer service is an important factor in customer satisfaction.

Relevance & Robustness

The generate association rules appear highly relevant to the hotel industry and focuses on aspects that directly affect customer experience.

Rows: 10 | Columns: 6

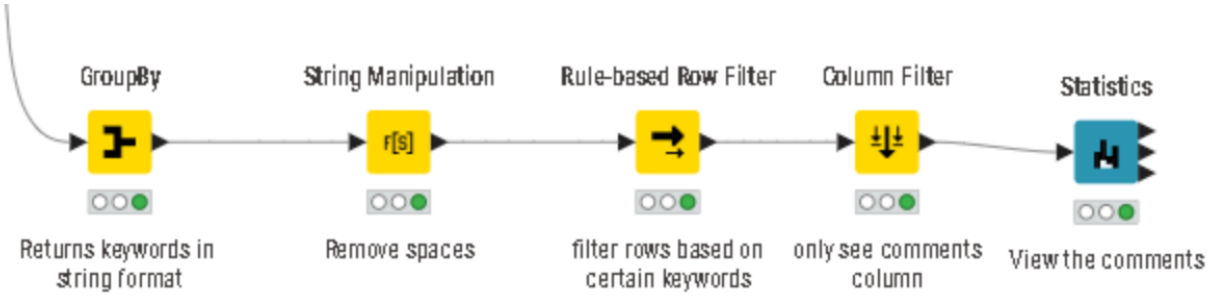
#	RowID	Support *Number (double)*	Confidence *Number (double)*	Lift *Number (double)*	Consequent *String*	Implies *String*	Items *Set*
1	rule39	0.027	0.94	1.942	room	<---	[small,clean]
2	rule85	0.028	0.929	1.92	room	<---	[roomy,hotel]
3	rule2...	0.041	0.911	1.882	room	<---	[large,clean]
4	rule1...	0.031	0.9	3.335	staff	<---	[friendly,helpful]
5	rule25	0.026	0.896	1.852	room	<---	[large,hotel,clean]
6	rule2...	0.041	0.882	1.823	room	<---	[small,good]
7	rule78	0.028	0.878	1.815	room	<---	[clean,good,stay]
8	rule80	0.028	0.878	1.815	room	<---	[comfortable,clean,g...
9	rule2...	0.043	0.877	1.813	room	<---	[large,comfortable]
10	rule1...	0.032	0.867	1.792	room	<---	[large,excellent,hotel]

(Figure: Top 10 Association Rules with minimum of 3 item sets)

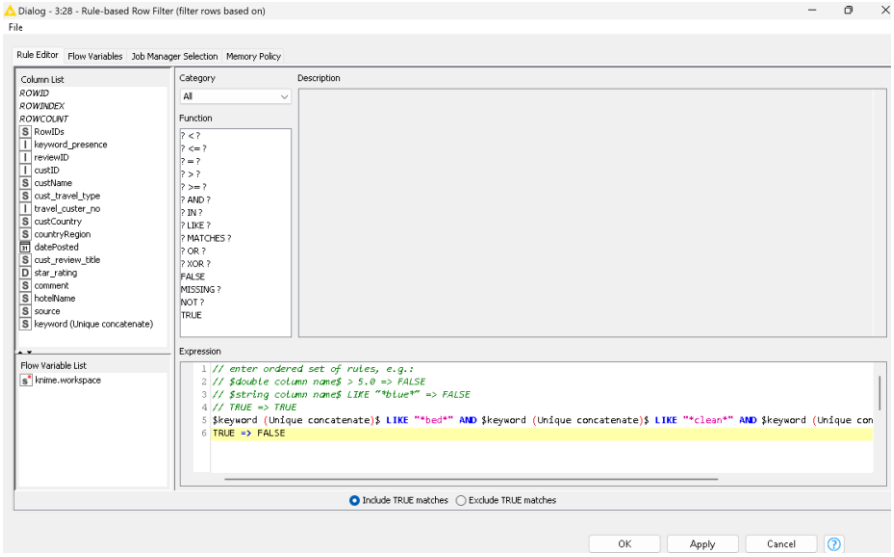
RECOMMENDATIONS

The Process

To formulate my recommendations, I grouped the keywords using the GroupBy node and standardized them by removing spaces through string manipulation. I then applied a rule-based filter to extract rows containing specific keywords, focusing on both antecedents and consequents in the association rules. By filtering the comments to include only those with relevant keywords, I conducted a detailed analysis using the Statistics node to gain deeper insights into guest feedback This approach, together with the top 10 rules from association rule mining, provided a solid foundation for my recommendations.



.



(*Figure: Rule-based Row filter filtering out rows with keywords in Association Rules)*

Suggested Recommendations

Enhance Room Size and Comfort

- Evidence: Reviews with the rule {large, comfortable} -> room shows that guests highly value spacious and comfortable rooms. The rule {large, clean} -> room also suggests that large, clean rooms significantly contribute to guest satisfaction.
- Action: The hotel management should ensure that rooms are designed to be spacious and comfortable and keep that in mind when designing rooms /room layout. They can also promote room size, cleanliness, and comfort in marketing to attract guests.

Maintain High Standards of Cleanliness

- Evidence: The rule {small, clean} -> room and {comfortable, clean, good} -> room show that cleanliness is a key factor in guest satisfaction, even in smaller or more budget-friendly rooms.
- Action: The hotel management should maintain high cleanliness standards across all rooms and common areas. Regularly training housekeeping staff and considering implementing a visible cleanliness certification (Eg. Cleanliness Grade/Rating) they can display to reassure assure the guests.

Cater to Large Groups and Families

- Evidence: Guests who mention {roomy, hotel} -> room and {large, hotel, clean} -> room feel that spacious, clean rooms are beneficial, particularly for families or larger groups.
- Action: The hotel management should offer options with larger rooms or suites for families or group, or perhaps giving free upgrades now and again. They should ensure these options are well-advertised and highlight their spaciousness and cleanliness in promotions.

Focus on Friendly and Helpful Staff

- Evidence: The rule {friendly, helpful} -> staff shows that guests appreciate friendly and accommodating staff, which enhances their overall experience.
- Action: During the staff's training, friendliness and helpfulness should be emphasised. They can also recognise and reward staff members who receive positive feedback from guests. The hotel management should inculcate the friendly and helpful culture in their staff making the guests experience a more positive one.

Address Concerns About Small Rooms

- Evidence: The rule {small, good} -> room suggests that guests may still be satisfied with smaller rooms if the overall experience is positive.
- Action: For hotels with smaller rooms, focus on maximizing the use of space and providing access to the amenities too. Highlight the quality and comfort of these rooms in advertising ensures potential guests see their value.

Promote High-Quality Bedding

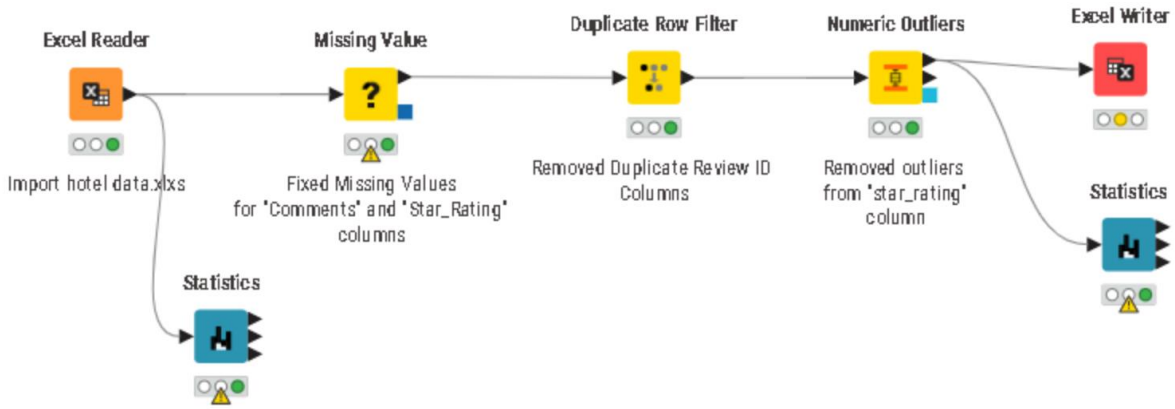
- Evidence: The rules {large, comfortable} -> room and {comfortable, clean, good} -> room emphasise the importance of comfortable and clean bedding to guest satisfaction.
- Action: The hotel management should invest in high-quality bedding and ensure that beds are large and comfortable. They can also offer different pillow and mattress options to cater to various guest preferences.

TASK 2

Objective: To identify the profile of customers that gives the highest average ratings for the hotel reviews

Features Used for clustering analysis: Cust_travel_type, Star_rating, Travel_custer_no, countryRegion

DATA PREPARATION



Data Cleaning

1. Fixed all missing values
2. Removed Duplicates
3. Replaced Numeric Outliers in the star_rating column with the closest permitted value to avoid a skewed representation of the cluster

MODELLING PROCESS & TECHNIQUES

1. Imported the dataset into SAS Vivya and used the Clustering object
2. Set stars_rating to average aggregation
3. Set travel_cust_no to categorical input

Parameter Tuning

Original Parameters

No. of Clusters = 5, Seed = 1234, Initial Assignment = Forgy, Standardization = Standard Deviation, Parallel Coordinates Plot (No. of bins) = 10

No. of Clusters

Clusters	Cluster ID	Group	Avg. Distance	Centroid Distance	Min Obs-Centroid Dist.	Max Obs-Centroid Dist.
k = 5	1	Couples	0.6470	1.2806	0.0137	1.9537
k = 5	2	Couples	0.7960	1.5575	0.0071	2.0835
k = 5	3	Couples	0.9832	1.9707	0.0036	1.7453
k = 5	4	Business Travellers	0.6279	1.2014	0.0198	1.9480
k = 5	5	Solo Travellers	0.7313	1.2014	0.0050	1.5366
k = 4	1	Couples	0.8126	1.7023	0.0131	1.9466
k = 4	2	Couples	0.8106	1.7934	0.0006	2.0860
k = 4	3	Couples	0.9313	1.9950	0.0037	1.7453
k = 4	4	Business Travellers	0.7219	1.7023	0.0021	1.5974

After analysing both cluster configurations (k =4, k =5), I found that k =5 gives me a better distinction between the different travel groups, while k =4 only results in 2 main travel groups. K = 5 configuration might be better for achieving greater inter-class separation, as it creates more distinct clusters, allowing for better differentiation between groups.

Whereas for k =4, it offers a more balance solution with lower intra-class variance and acceptable inter-class separation. To maximise interclass separation and minimise intra-class separation, we will choose k=5 for further parameter tuning despite slight risk of redundancy

Initialization Method

I will be using the Forgy initialization method because it typically offers better initial centroid placement, leading to more distinct clusters from the outset. This method tends to produce more compact clusters, and it generally requires fewer iterations to converge, making the clustering process more efficient. Additionally, Forgy provides more consistent and reliable final cluster quality compared to the Random method, which can be more variable in its results.

Standardization Method

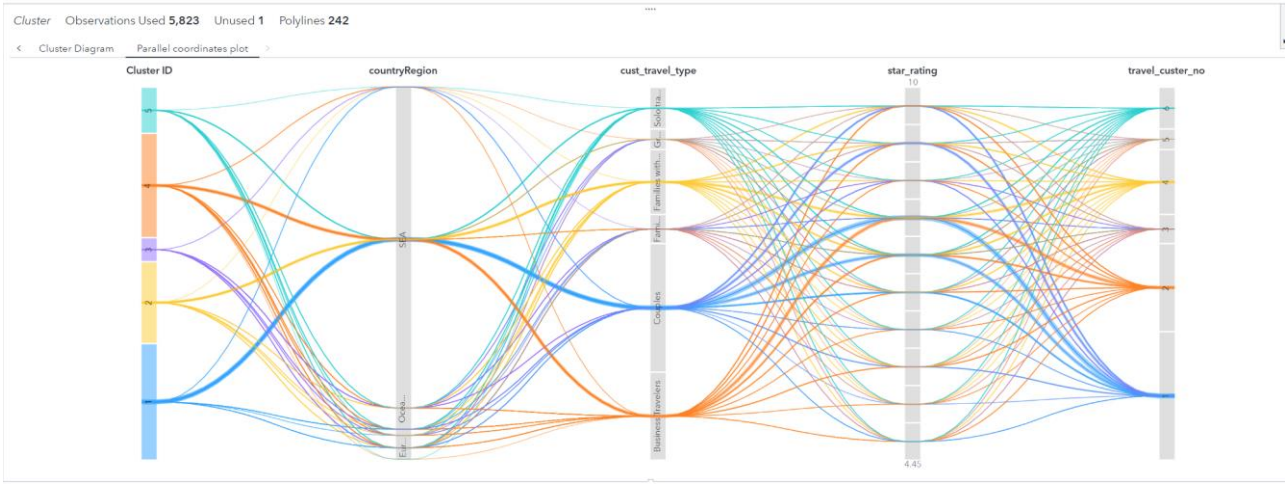
Range normalization was chosen for this clustering analysis to ensure all features are scaled to the same range of [0,1], which is crucial for the k-means algorithm. This method prevents features with larger scales from disproportionately influencing the clustering results, leading to a more balanced and accurate representation of customer profiles. By standardizing features to a common range, range normalization allows for equal contribution from all features in the distance calculations, facilitating a more meaningful and fair clustering outcome.

Parallel Coordinates Plot – No. of bins

For the parallel coordinates plot, the number of bins was initially set to 10. After experimenting with different bin counts, it was observed that changes primarily affected the star_rating feature. Since star_rating is a numerical input, it can be effectively divided into different ranges based on the number of bins. Other variables, such as custTravelType and travel_custer_no, did not show significant changes in response to bin adjustments. Given that the adjustments did not impact the overall data interpretation significantly and to maintain consistency and clarity in the plot, the number of bins was retained at 10. This choice provides a balanced representation of the data while keeping the plot readable and informative.

MODEL EVALUATION

(Figure: Cluster Diagram)



(Figure: Parallel Coordinates Plot)

Centroids | Cluster Summary | Model Information | Within Cluster Statistics | Iteration History | Frequency | Standa

Cluster ID	countryRegion	cust_travel_type	star_rating ▾	travel_custer_no
3	Oceania	Couples	8.3510638298	1
2	SEA	Families with Young Children	8.1282401803	4
1	SEA	Couples	7.9043966085	1
5	SEA	Solo travelers	7.7799686103	6
4	SEA	Business Travelers	7.5255032914	2

(Figure: Centroids Table)

Cluster Description

Cluster ID 3: Couples (Cluster with the Highest Ratng)

Cluster ID 3 is dominated by couples, primarily from the Oceania region, with some travelers from Europe and Sub-Saharan Africa. This cluster features the highest overall average star rating, ranging from about 6.5 to 9, with an average centroid star rating of approximately 8.3.

Cluster ID 1: Couples
Cluster ID 1 primarily consists of couples, predominantly from the Southeast Asia (SEA) region, with some representation from North America and Sub-Saharan Africa. The majority of star ratings for this group fall in the moderate to high range of 7.5 to 8.5, indicating generally positive experiences. The average centroid star rating is approximately 7.7.

Cluster ID 2: Families with Young Children
Cluster ID 2 includes families with young children, mainly from the SEA region, with additional representation from Europe, the Middle East, North America, Oceania, and South America. The star ratings for this group are similar to those of Cluster ID 1, falling in the range of 7.5 to 8.5, with an average centroid star rating of approximately 8.1

Cluster ID 4: Business Travellers
Cluster ID 4 consists mainly of business travelers from the SEA region, with some representation from other regions. This cluster has the lowest average star rating, with ratings primarily between 7.5 and 8.5, and an average centroid star rating of approximately 7.5.

Cluster ID 5: Solo Travellers
Cluster ID 5 includes solo travelers from the SEA region, as well as from the Caribbean, Europe, the Middle East, North America, Oceania, and Sub-Saharan Africa. The average centroid star rating for this cluster is approximately 7.7, with ratings typically ranging from 7.5 to 8.5.

RECCOMENDATIONS

Couples
Since this cluster is mainly couples with a moderate to high star rating (7.5 to 8.5), enhancing experiences catered to couples, such as couples' retreats, special dinners, and exclusive couple-friendly activities, could attract more from this cluster. This cluster has people that come from different countries as compared to Cluster ID 1. Perhaps these experiences for this cluster can be tailored to their country perhaps having food and activities they are familiar with from their country.

Families with Young Children
This cluster has a slightly higher average star rating (around 8.1), indicating satisfaction among families. Providing additional family-oriented services, such as kids' clubs, family rooms, and child-friendly dining options, could further improve their experience

Couples (Cluster with the highest rating)
This cluster has the highest average star rating (around 8.3). This cluster come from OCEANIA which is different from the other clusters. Perhaps to improve the experience they could serve the food delicacies from OCEANIA giving the customers a sense of familiarity on their dates.

Business travellers
Business travellers have the lowest average star rating (around 7.5), which indicates some dissatisfaction. Enhancing business-oriented amenities such as high-speed Wi-Fi, comfortable workspaces, and efficient check-in/check-out processes could improve their experience.

Solo travellers
Solo travellers, with an average star rating of around 7.7, may benefit from enhanced safety measures, better connectivity (e.g., free Wi-Fi), and social spaces where they can interact with other travellers. Offering guided tours or social events specifically for solo travellers could also appeal to this segment.