Submission template – association rules

1. Introduction, understanding of the problem

This work's goal is to identify customer groups that are large enough and at the same time at least 90% of them choose United Kingdom as a destination, as well as identifying customer groups that are large enough and who choose United Kingdom 30% more often than is the average, with at least 20% of the group choosing this destination.

2. Understanding the data

Before starting the analysis, I excluded some data from the dataset that wasn't of interest such as ID, PassengerClass, FellowAdult, Newspaper, OnBoardEntertainmentSystem, Audio, Food, FellowChildren, Drink, Sport, Baggage_Count, Bagggage_CarryOn and Ticket.

3. Data preprocessing

Describe data preprocessing for each attribute used in the analysis using the following table.

attribute	preprocessing	Category with lowest support
Airport-countries	*Airports BRS, LCY and LHR merged into "UK"* *Airports BTS and KSC merged into "Slovakia"* *Airports CGN and FRA merged into "Germany"* *Airports HEL and VAA merged into "Finland"* *Airport ARN – Sweden* *Airport BFG – USA* *Airport VIE - Austria*	*Category "USA" with support 69.*
Age category	*Equidistant Enumeration*	*Category [40.8;52.2) with support 168*
Profession?	*Categories "Businessman" and "Employee" merged into "Employed"* *Categories "Student", "Jobless" and "Pensioner" merged into "Unemployed"* *Category "Other"*	*Category "Other" with support 137*

4. Modelling

a. Analytical question 1

Describe the final setting (minimum confidence and support thresholds).

interest measure	**minimum value**	**justification**
Confidence	*0.9*	*Higher value resulted in no rules found*
Support	*0.05*	*This support threshold corresponds to a minimum of 50 instances.*

Describe the list of attributes for the antecedent and consequent.

attribute	**position**	**justification**
Airport-countries	*Consequent*	*Target attribute, limited to „UK“*
ResidenceCountry	*Antecedent*	*Attribute of interest*
Age category	*Antecedent*	Equidistant categories, attribute of interest
FlightFrequence	*Antecedent*	*Attribute of interest*
Periodicity	*Antecedent*	*Attribute of interest*
PlannedTourLength	*Antecedent*	*Attribute of interest*
Profession?	*Antecedent*	*Attribute of interest*
Reason	*Antecedent*	*Attribute of interest*

a. Analytical question 2

Do the same for analytical question 2, listing also the value of lift.

5. **Model evaluation**

result type	*result*	*comment*
Task 1		
number of rules	*1*	
highest confidence	*0,9*	
highest support	*0,054*	
selected rule 1:	*ResidenceCountry(uk) → Airport-countries(UK)*	
rule 1 confidence	*0,9*	
rule 1 support	*0,054*	
interpret rule 1 confidence	*90% of UK citizens flying somewhere are flying to the UK*	
interpret rule 1 support	*54 UK citizens in this dataset are flying to the UK*	
Task 2		
number of rules	3	

highest confidence	0,9
highest support	0,054
selected rule 2:	*ResidenceCountry(poland) → Airport-countries(UK)(decided to choose this one to not list the same rule twice)*
rule 2 confidence	*0,544*
rule 2 support	*0,043*
interpret rule 2 lift	*2,991*

Lift in rule 2 means that the probability of the country of destination being the UK if the person is from Poland is 2,991 times bigger than the probability of any person in the dataset flying to the UK

Selected rule from task 1:
Contingency table:

	consequent	¬ consequent
Antecedent	*54*	*6*
¬ antecedent	*128*	*812*

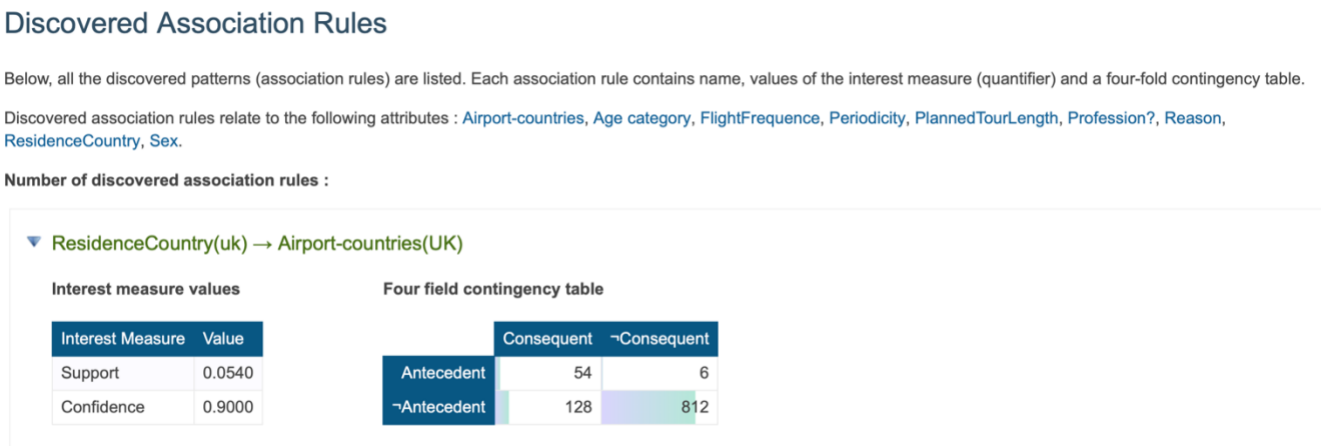
Selected rule from task 2:

	consequent	¬ consequent
Antecedent	*43*	*36*
¬ antecedent	*139*	*782*

6. Possibilities for the use of the model and conclusion.

By utilizing these models, companies can get an insight into the customer groups that are high enough and choose the UK as their destination more often than others and meet the specified criteria. This information helps companies better understand their clientele and tailor their services just for them, for example by using targeted marketing strategies and personalized offers. As a result customers satisfaction increases, as well as the business's revenue

Appendix



Discovered Association Rules

Below, all the discovered patterns (association rules) are listed. Each association rule contains name, values of the interest measure (quantifier) and a four-fold contingency table.

Discovered association rules relate to the following attributes : Airport-countries, Age category, FlightFrequence, Periodicity, PlannedTourLength, Profession?, Reason, ResidenceCountry, Sex.

Number of discovered association rules :

▼ ResidenceCountry(uk) → Airport-countries(UK)

Interest measure values

Interest Measure	Value
Support	0.0540
Confidence	0.9000

Four field contingency table

	Consequent	¬Consequent
Antecedent	54	6
¬Antecedent	128	812

Discovered Association Rules

Content

1. Datase
2. Create
3. Data M
4. Discov

Below, all the discovered patterns (association rules) are listed. Each association rule contains name, values of the interest measure (quantifier) and a four-fold conting table.

Discovered association rules relate to the following attributes : Airport-countries, Age category, FlightFrequence, Periodicity, PlannedTourLength, Profession?, Reason, ResidenceCountry, Sex.

Number of discovered association rules :

ResidenceCountry(uk) → Airport-countries(UK)

Interest measure values

Interest Measure	Value
Support	0.0540
Confidence	0.9000

Four field contingency table

	Consequent	¬Consequent
Antecedent	54	6
¬Antecedent	128	812

ResidenceCountry(poland) → Airport-countries(UK)

Interest measure values

Interest Measure	Value
Support	0.0430
Confidence	0.5443

Four field contingency table

	Consequent	¬Consequent
Antecedent	43	36
¬Antecedent	139	782

ResidenceCountry(other) → Airport-countries(UK)

Interest measure values

Interest Measure	Value
Support	0.0420
Confidence	0.5250

Four field contingency table

	Consequent	¬Consequent
Antecedent	42	38
¬Antecedent	140	780