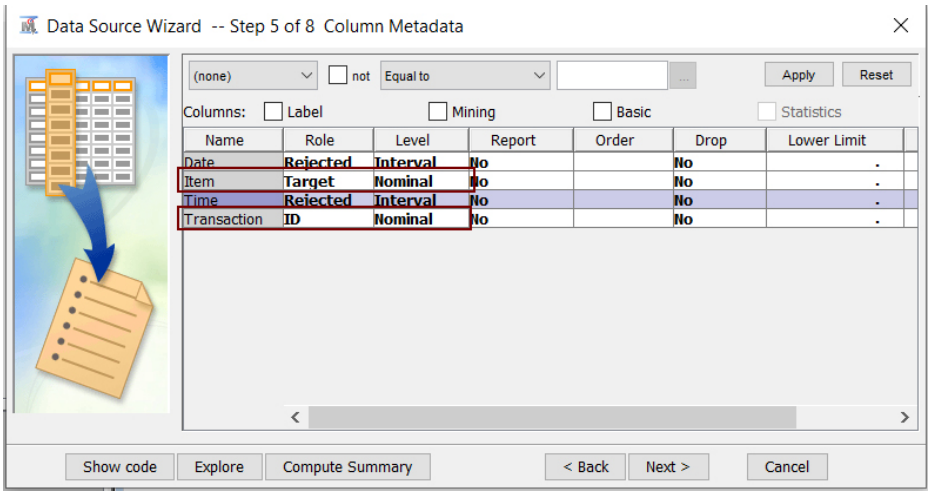
ASSOCIATIONS

1) Variables. For the association analysis (or market basket analysis), you need to have two kinds of variables: Target and ID. When creating a data source, set the role of the ITEM (the categories of goods people buy at the bakery) to TARGET and the role of the TRANSACTION to ID. We don't need the date and time, so just reject the variables.

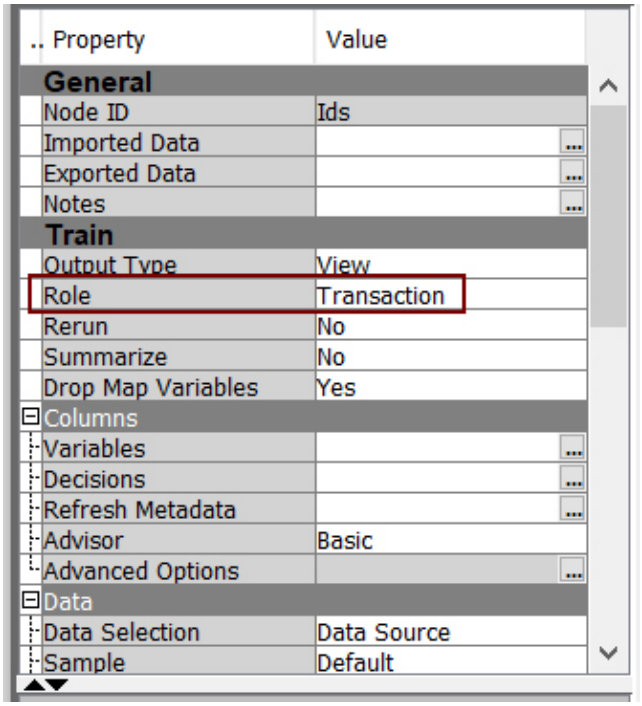


2) Create a new diagram. Drag and drop the data source and run the StatsExplore to understand your data better.

```
Variable Levels Summary
(maximum 500 observations printed)

                          Frequency
Variable       Role         Count

Item           TARGET          95
Transaction    ID            9531



Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                 Number
Data      Variable                 of                              Mode                 Mode2
Role        Name        Role      Levels    Missing      Mode      Percentage    Mode2    Percentage

TRAIN       Item        TARGET      95          0        Coffee      25.69       Bread      15.62
```

As you can see, the target variable – ITEM has 95 levels. Also, the ID variable has 9531 unique values out of 21,293 total observations.

3) Set the dataset role to TRANSACTION.



.. Property	Value
General	
Node ID	Ids
Imported Data	
Exported Data	
Notes	
Train	
Output Type	View
Role	Transaction
Rerun	No
Summarize	No
Drop Map Variables	Yes
Columns	
Variables	
Decisions	
Refresh Metadata	
Advisor	Basic
Advanced Options	
Data	
Data Selection	Data Source
Sample	Default

4) Run the association analysis:

EXPLORE → ASSOCIATION (attach to the dataset) –> RUN

NOTE: In an ideal situation, you are looking for cases with high CONFIDENCE and high SUPPORT, which doesn't occur very often.

When run with the support level set to 5%, we don't get as many rules as we could have. As we can see, we only have a little bit more than 10 rules. Sorting them by confidence, we may see that those who bought Medialuna also bought coffee.

Table: Statistics Plot

Relations	Expected Confidence(%)	Confidence(%) ▼	Support(%)	Lift	Transaction Count	Rule
2	47.51	56.92	3.49	1.20	333.00	Medialuna ==> Coffee
2	47.51	55.21	4.72	1.16	450.00	Pastry ==> Coffee
2	47.51	53.25	4.21	1.12	401.00	NONE ==> Coffee
2	47.51	53.24	3.80	1.12	362.00	Sandwich ==> Coffee
2	47.51	52.70	5.43	1.11	518.00	Cake ==> Coffee
2	47.51	51.84	2.80	1.09	267.00	Cookies ==> Coffee
2	47.51	50.72	2.94	1.07	280.00	Hot chocolate ==> Coffee
2	47.51	34.96	4.95	0.74	472.00	Tea ==> Coffee
2	32.49	33.87	2.90	1.04	276.00	Pastry ==> Bread
2	47.51	27.51	8.94	0.58	852.00	Bread ==> Coffee
2	32.49	19.70	2.79	0.61	266.00	Tea ==> Bread
2	32.49	18.82	8.94	0.58	852.00	Coffee ==> Bread
2	10.31	11.44	5.43	1.11	518.00	Coffee ==> Cake
2	14.16	10.42	4.95	0.74	472.00	Coffee ==> Tea

This table was generated by another node available in the EXPLORE tab: Market Basket, which gives you a primitive analysis of the associations. However, below, it will be used to explain how SAS calculates various values. Keep in mind the total number of unique customers in the dataset is **9531.**

Item Statistics

Item Name	Transaction Count	Support(%)
Bread	3097	32.4940
Scandinavian	275	2.8853
Hot chocolate	552	5.7916
Cookies	515	5.4034
Muffin	364	3.8191
Coffee	4528	47.5081
Pastry	815	8.5510
Medialuna	585	6.1379
Tea	1350	14.1643
NONE	753	7.9005
Farm House	371	3.8926
Juice	365	3.8296
Soup	326	3.4204
Cake	983	10.3137
Sandwich	680	7.1346
Alfajores	344	3.6093
Brownie	379	3.9765
Truffles	192	2.0145
Toast	318	3.3365
Scone	327	3.4309

For the association analysis **in SAS**, you need to know about the following:

Expected Confidence: corresponds to the **right-hand side** of the rule; the proportion of the customers who used the service from the whole dataset (measured in %). The proportion of customers who bought A out of all customers. For example, **COFFEE was bought 4528 times** out of 9531 unique customers. Thus, the expected confidence = **47.51%.**

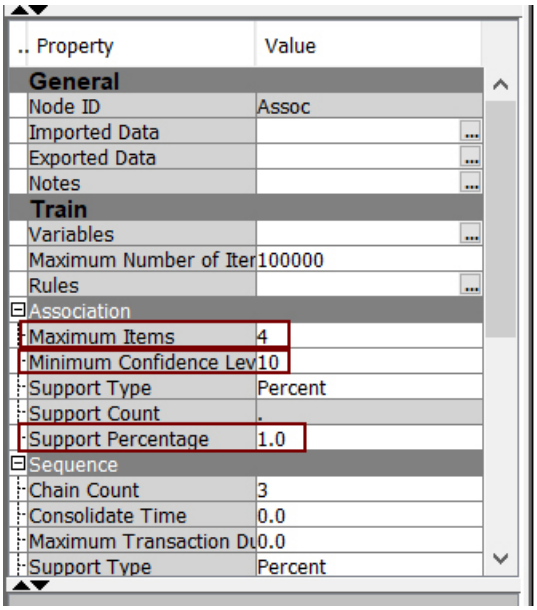
Support: measures the proportion of a service or a combination of services on the left hand side of the rule (the proportion of the antedencent, measured in %). The number of customers who bought both A and B out of the whole number of customers. The total count of those who bought Medialuna and Coffee together accounted for 333 customers out of 9531 (333/9531 = **3.49%**).

Confidence: measures the probability that a certain service is used given that another service or a combination of services is used (measured in %). The confidence percentage for A => B is the percentage of all customers who purchased both A and B, divided by the number of customers who purchased A.[1] The higher the confidence, the higher the association between the services. The rule Medialuna ===> Coffee should be read as "What is a probability of a customer buying a coffee given Medialuna?". The number of customers who bought Medialuna & Coffee = **333** (from the association analysis node) and the number of customers who bough only Medialuna = **585** (from the Market Basket node). P(Coffee|Medialuna) = P(Coffee&Medialuna)/P(Medialuna) = 333/585 = 0.**5692**, which will equal the CONFIDENCE of **56.92%.**

Lift: used to rank the rules. It is calculated as the Confidence over the Expected Confidence. Expected confidence = **47.51%** and Confidence = **56.92%**, then the lift will be (56.92%)/ (47.51%)=1.2.

[1] http://support.sas.com/documentation/cdl/en/emcs/66392/HTML/default/viewer.htm#n0omtukdp38gw4n14kudfgql5zao.htm

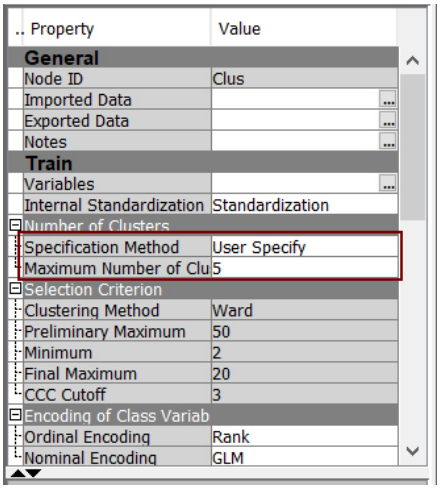
In general, we would be interested in finding rules that would have a high support and confidence levels as well as a high lift (higher than 1).
Note: If you have too few rules generated, you can also change the support or confidence level or the complexity of the rules as indicated below.



CLUSTERING/ SEGMENTING

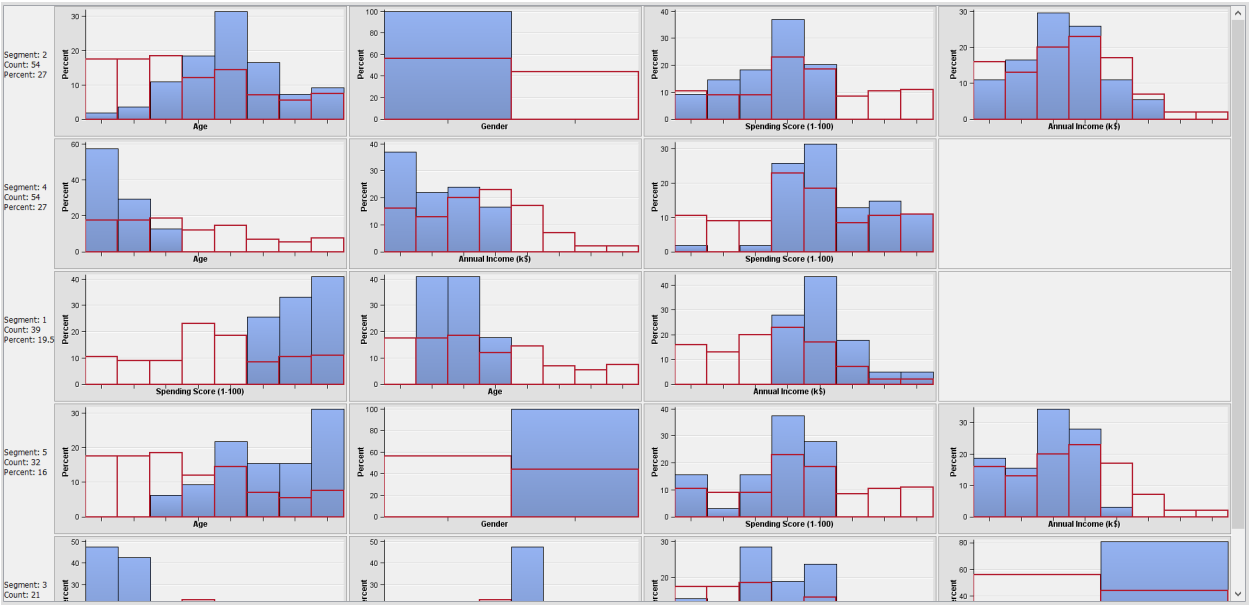
You need an ID variable whose role should be set to ID and its level should be set to nominal. The other variables should be interval (they shouldn't be skewed).

CREATE A DIAGRAM → CREATE A DATA SOURCE → SET THE CUSTOMER_ID TO ID → ALL OTHER VARIABLES HAVE TO BE INTERVAL → DRAG AND DROP THE DATASET → EXPLORE → CLUSTER → SET THE NUMBER OF CLUSTERS TO 5 → ASSESS → SEGMENT PROFILE → RUN → ANALYZE THE RESULTS FOR BOTH NODES (CLUSTER AND SEGMENT PROFILE)



In the variable Worth plot you can study the importance of each variable in each segment (in the case below, you can see the variable worth for the second segment):

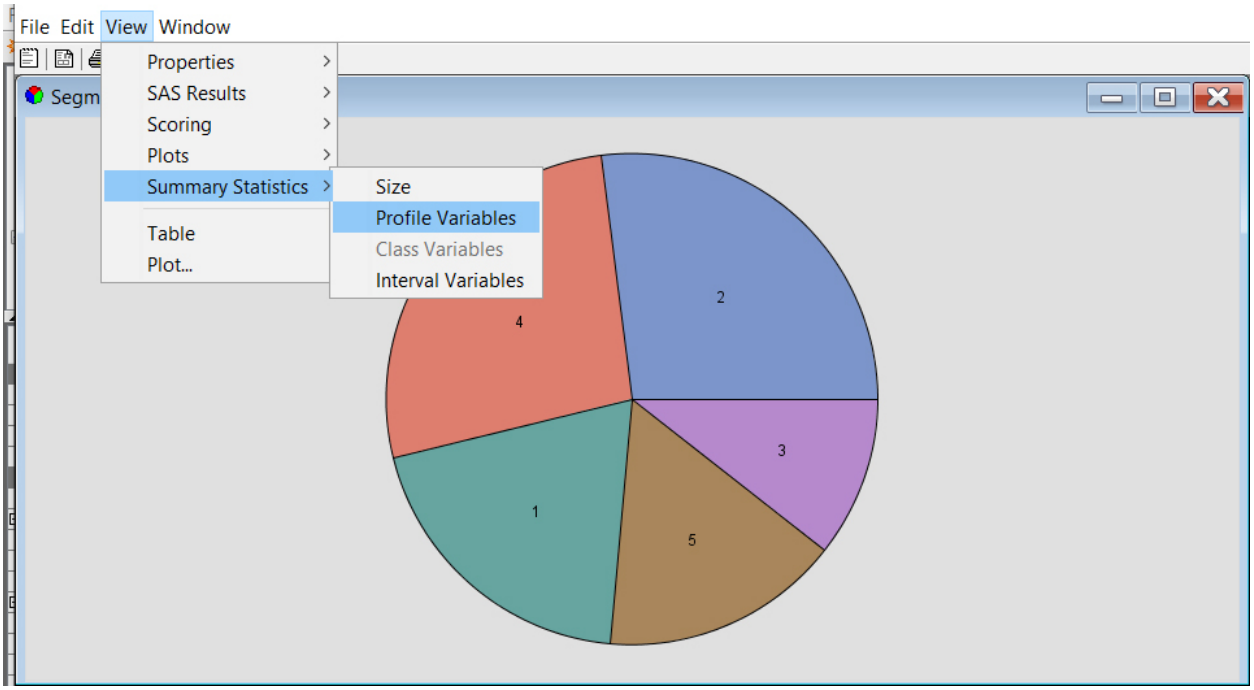
You can also study the distribution of variables in each segment by studying the plots. The red lines show the population distribution.



In the output you can estimate the size of the segment (proportion) and the importance of each variable in a specific segment.

```
25
26   Frequencies: _SEGMENT_
27
28                                          Percent of
29    Segment      Segment     Frequency       Total
30   Variable      Value       Count         Frequency
31
32   _SEGMENT_       2           54            27.0
33   _SEGMENT_       4           54            27.0
34   _SEGMENT_       1           39            19.5
35   _SEGMENT_       5           32            16.0
36   _SEGMENT_       3           21            10.5
37
38
39
40   Variable: _SEGMENT_ Segment: 2 Count: 54
41   Decision Tree Importance Profiles
42
43   Variable                    Worth      Rank
44
45   Age                       0.11796       1
46   Gender                    0.11456       2
47   Spending_Score__1_100_    0.09073       3
48   Annual_Income__k__        0.03157       4
49
```

You can estimate the cluster plots, cluster sizes, and also you can have a look at the table with values given for each cluster (Results → View → Summary statistics → Profile Variables). You can study the Interval variables included into your profiles in by studying the interval variables: Results → View →Summary statistics → Interval Variables.



Profile Variables

Type	Segment Variable	Segment Value	Variable	Rank	Worth	Label	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
	OVERA...	EMWS6....	Age	.	.	Age	0	18	70	38.85	13.96901	0.485569	-0.67157
	OVERA...	EMWS6....	Annual I...	.	.	Annual In...	0	15	137	60.56	26.26472	0.321843	-0.09849
	OVERA...	EMWS6....	Gender	.	.	Gender	0	0	1	0.44	0.497633	0.243578	-1.96038
	OVERA...	EMWS6....	Spending...	.	.	Spending...	0	1	99	50.2	25.82352	-0.04722	-0.82663
	SEGME...	2	Age	1	0.117964	Age	0	23	68	47.62963	10.53708	-0.06151	-0.08462
	SEGME...	2	Gender	2	0.114557	Gender	0	0	0	0	0	.	.
	SEGME...	2	Spending...	3	0.090732	Spending...	0	5	59	36.87037	15.88526	-0.51806	-0.80697
	SEGME...	2	Annual I...	4	0.031575	Annual In...	0	18	101	57.27778	19.91152	0.128441	-0.37303
	SEGME...	4	Age	1	0.226863	Age	0	18	35	24.90741	5.349197	0.497235	-1.00422
	SEGME...	4	Annual I...	2	0.128338	Annual In...	0	15	67	39.66667	17.03825	0.06388	-1.40188
	SEGME...	4	Spending...	3	0.067321	Spending...	0	6	99	61.2037	18.42003	0.003878	0.37774
	SEGME...	1	Spending...	1	0.195924	Spending...	0	63	97	82.12821	9.364489	-0.11264	-1.20215
	SEGME...	1	Age	2	0.122759	Age	0	27	40	32.69231	3.72865	0.42529	-0.88762
	SEGME...	1	Annual I...	3	0.11895	Annual In...	0	69	137	86.53846	16.31248	1.423666	1.736849
	SEGME...	5	Age	1	0.113639	Age	0	35	70	55.34375	10.52067	-0.26897	-1.02994
	SEGME...	5	Gender	2	0.065164	Gender	0	1	1	1	0	.	.
	SEGME...	5	Spending...	3	0.047657	Spending...	0	3	60	39.3125	16.96379	-1.06943	0.04227
	SEGME...	5	Annual I...	4	0.04418	Annual In...	0	19	77	48.9375	16.12839	-0.49388	-0.59299
	SEGME...	3	Spending...	1	0.08395	Spending...	0	1	28	13.47619	7.379831	0.127829	-0.31405
	SEGME...	3	Annual I...	2	0.048823	Annual In...	0	71	137	92.19048	18.56777	1.077539	0.352882
	SEGME...	3	Age	3	0.019926	Age	0	19	59	38.42857	11.45239	-0.14986	-0.38121
	SEGME...	3	Gender	4	0.012219	Gender	0	0	1	0.809524	0.402374	-1.70043	0.975232