2020

LMR Market Research

Techniques and Methodologies

DATA PROCESSING AND MULTIVARIATE MODELING

ADAM RUDMIK & TONY RUDMIK

Contents

Methodologies: Data processing & statistical modelling

Cross Tabulations & Banners:

The primary tool in any data processing arsenal is the ability to aggregate and display the distribution of data by tabulating results against other two-dimensional grids. A form of segmentation on its own, cross tabulations provide many of the essential statistics needed in order to discover the narrative hidden in the data. Our statistical packages and programming tools can handle up to 22 columns per banner per page of breaks. The aggregate data displayed can include;

Standard tabulations:

Frequency and vertical (or horizontal) percentages, averages, mean scores, standard deviation, Standard Error (for continuous/scale variables), Netting of responses into multiple layers of groups, ranking of values (when appropriate), Mode, Median.

Additional Analyses/Processes:

- **Significance testing** within/between column variables using efficient lettering system (e.g. Confidence level A=.95/a=.90)
- **Weighting** of data using simple cell by cell or, more complex, Iterative Proportional Fitting procedures
- **Volumetric Share Analysis** excluding outlier effects (e.g. Share of dollars spent)
- **Summary tables** with vertical/ordinal ranging (e.g. Mean scores, top box, top 2 box, etc.)
- **Indices** based on targeted cell/column cell (frequency or percentage indexing)
- **Optimization banner column selection** using CHAID/CART techniques (see below)
- Development of complex **data layered matrices** when category results are required
- Occasion/mention-based output for diary/chart data
- Integration of cross-tabs with multivariate output (e.g. segmentation, factors)
- Developing customer satisfaction/loyalty overall metrics
- **Fixed Sum allocation tables**
- **Tracking studies** across multiple datasets

The output of these tables can be delivered in either Excel or Word formatting depending on client preference.

Segmentation/Cluster Analysis:

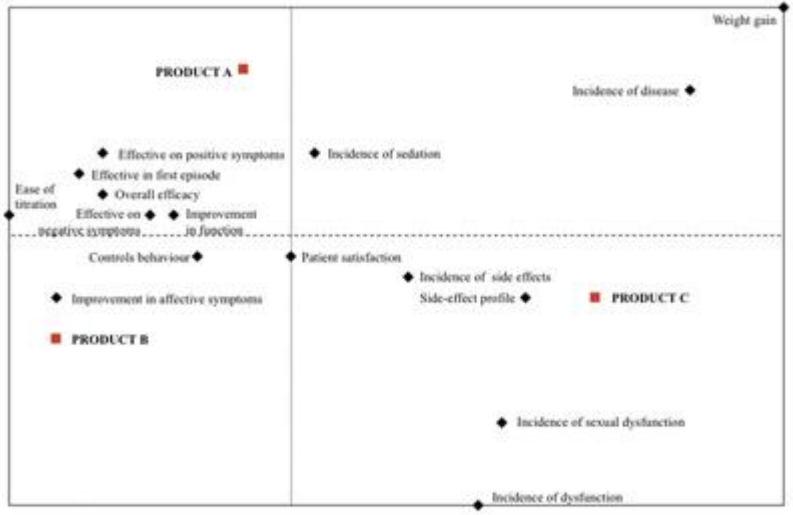
We use the most statistically accurate techniques that are designed to organize respondents into groups (profiles) with similar traits in order to produce preferences or anticipate behaviours. Segmentation allows for statistical confidence in the grouping of data with techniques that detect natural groupings.

1) **Hierarchical or linkage methods** – where cases will be joined together on the basis of highest similarity to any member of an existing group
2) **K-means** – a splitting method that clusters groups in order to produce the largest between group variance while minimizing within-cluster variances.
3) **Dominant Factor Segmentation** – Assigns each respondent to a segment by their "Dominant" factor score. This technique provides excellent results when the researcher is looking for the "big" picture trends.

Correspondence Analysis (Brand/Image Maps):

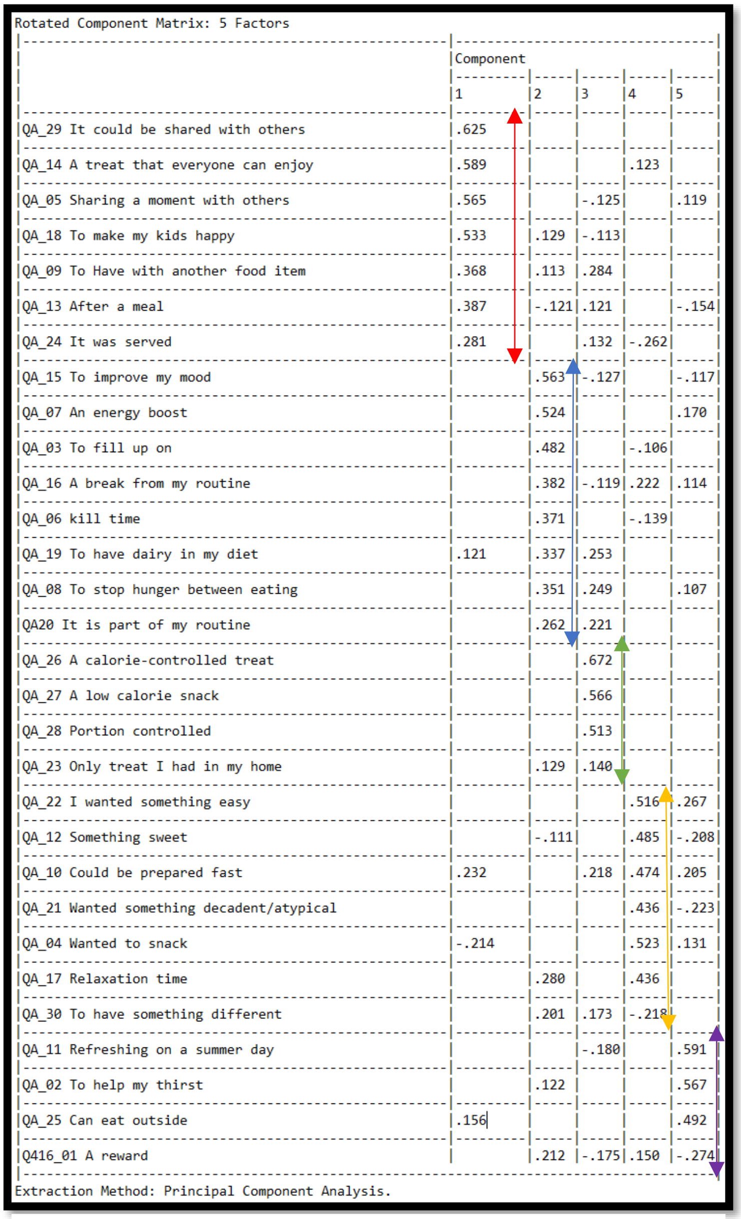
A visualization of a multivariate model that indicates a representation of a distance metric. Similar to Principal component analysis, Correspondence Analysis (CA) takes the group indicator scores and produces a two-way contingency table from data produced in the form of cross tabulations. Correspondence analysis produces graphs formed by the first principal axis taken two at a time. Each of the principal axes is associated with an eigen-structure which defines the projections on the axes, as well as the relative variance in the points explained by the axes (see example below).

Example: Illustrates the interaction between 3 Products and various attribute scores, associated with each product, whose visual distance represents a statistical metric of association.



Principal Component Analysis

Given an array of correlation coefficients, principle component analysis (PCA) enables us to see whether some underlying pattern of relationships exists such that data may be “rearranged” or “reduced” to a smaller set of factors or components that may be taken as source variables accounting for observed interrelations in the data. Not only does PCA reduce the number of variables into larger themes of correlated variables it also may reveal interesting relationships within a factor.



```
Rotated Component Matrix: 5 Factors
|-----------------------------------------------|-----------------------------|
|                                               |Component                    |
|                                               |-----|-----|-----|-----|-----|
|                                               |1    |2    |3    |4    |5    |
|-----------------------------------------------|-----|-----|-----|-----|-----|
|QA_29 It could be shared with others           |.625 |     |     |     |     |
|QA_14 A treat that everyone can enjoy          |.589 |     |     |.123 |     |
|QA_05 Sharing a moment with others             |.565 |     |-.125|     |.119 |
|QA_18 To make my kids happy                    |.533 |.129 |-.113|     |     |
|QA_09 To Have with another food item           |.368 |.113 |.284 |     |     |
|QA_13 After a meal                             |.387 |-.121|.121 |     |-.154|
|QA_24 It was served                            |.281 |     |.132 |-.262|     |
|QA_15 To improve my mood                       |     |.563 |-.127|     |-.117|
|QA_07 An energy boost                          |     |.524 |     |     |.170 |
|QA_03 To fill up on                            |     |.482 |     |-.106|     |
|QA_16 A break from my routine                  |     |.382 |-.119|.222 |.114 |
|QA_06 kill time                                |     |.371 |     |-.139|     |
|QA_19 To have dairy in my diet                 |.121 |.337 |.253 |     |     |
|QA_08 To stop hunger between eating            |     |.351 |.249 |     |.107 |
|QA20 It is part of my routine                  |     |.262 |.221 |     |     |
|QA_26 A calorie-controlled treat               |     |     |.672 |     |     |
|QA_27 A low calorie snack                      |     |     |.566 |     |     |
|QA_28 Portion controlled                       |     |     |.513 |     |     |
|QA_23 Only treat I had in my home              |     |.129 |.140 |     |     |
|QA_22 I wanted something easy                  |     |     |     |.516 |.267 |
|QA_12 Something sweet                          |     |-.111|     |.485 |-.208|
|QA_10 Could be prepared fast                   |.232 |     |.218 |.474 |.205 |
|QA_21 Wanted something decadent/atypical       |     |     |     |.436 |-.223|
|QA_04 Wanted to snack                          |-.214|     |     |.523 |.131 |
|QA_17 Relaxation time                          |     |.280 |     |.436 |     |
|QA_30 To have something different              |     |.201 |.173 |-.218|     |
|QA_11 Refreshing on a summer day               |     |     |-.180|     |.591 |
|QA_02 To help my thirst                        |     |.122 |     |     |.567 |
|QA_25 Can eat outside                          |.156 |     |     |     |.492 |
|Q416_01 A reward                               |     |.212 |-.175|.150 |-.274|
|-----------------------------------------------------------------------------|
Extraction Method: Principal Component Analysis.
```

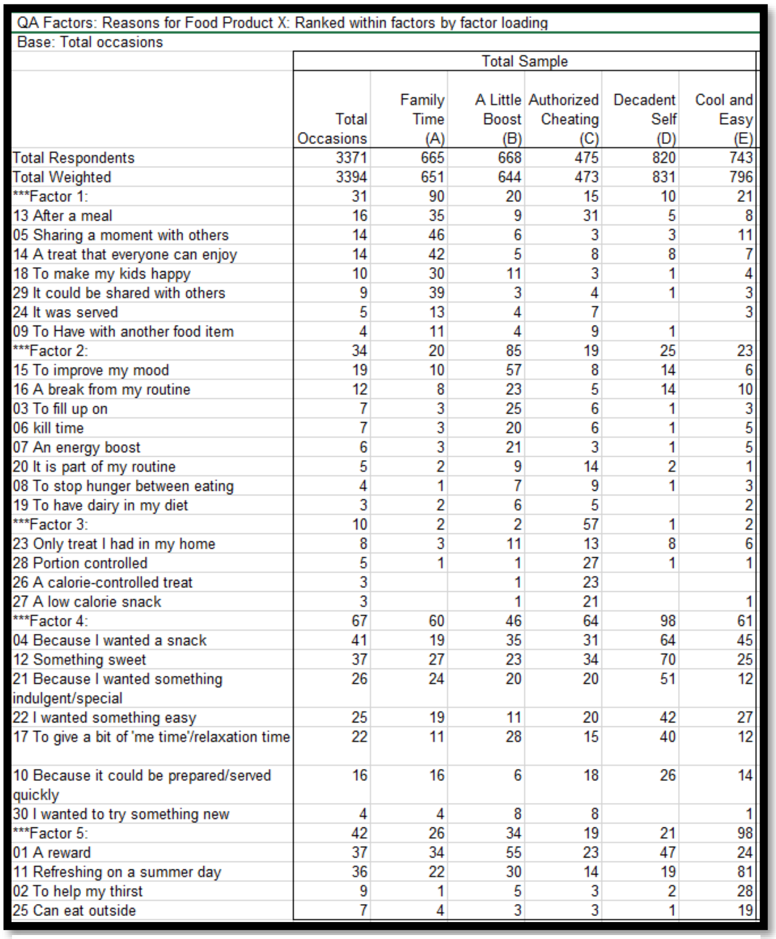
Figure 1

Figure 1 - Displays the final groupings that result from an iterative process that uses statistical output (i.e. Rotated component matrices, Cumulative % Variances), to determine the number of factors that represents a best fit of the data. In this example, 30 attributes associated with a food product were found to reduce to 5 significant groupings. Once the groups are decided upon, appropriate names can be created that describe the common thematic elements of each respective grouping.

Ex. Identifying groups

Group 1 – Family Time
Group 2 – A Little Boost
Group 3 – Authorized Cheating
Group 4 – Decadent Self
Group 5 – Cool and Easy

The model may be used in further confirmatory analysis with special cross tabulation tables that illustrate the degree to which the respondents appropriately fit within their respective segments. The results are also used as banner points through which to analyze all survey questions.



QA Factors: Reasons for Food Product X: Ranked within factors by factor loading

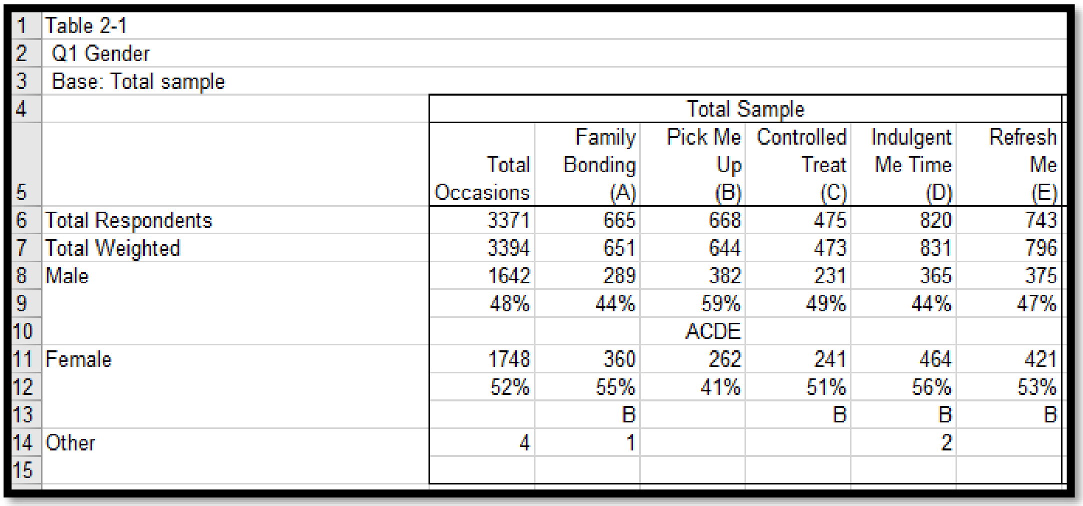
Base: Total occasions

	Total Sample					
	Total Occasions	Family Time (A)	A Little Boost (B)	Authorized Cheating (C)	Decadent Self (D)	Cool and Easy (E)
Total Respondents	3371	665	668	475	820	743
Total Weighted	3394	651	644	473	831	796
***Factor 1:	31	90	20	15	10	21
13 After a meal	16	35	9	31	5	8
05 Sharing a moment with others	14	46	6	3	3	11
14 A treat that everyone can enjoy	14	42	5	8	8	7
18 To make my kids happy	10	30	11	3	1	4
29 It could be shared with others	9	39	3	4	1	3
24 It was served	5	13	4	7		3
09 To Have with another food item	4	11	4	9	1	
***Factor 2:	34	20	85	19	25	23
15 To improve my mood	19	10	57	8	14	6
16 A break from my routine	12	8	23	5	14	10
03 To fill up on	7	3	25	6	1	3
06 kill time	7	3	20	6	1	5
07 An energy boost	6	3	21	3	1	5
20 It is part of my routine	5	2	9	14	2	1
08 To stop hunger between eating	4	1	7	9	1	3
19 To have dairy in my diet	3	2	6	5		2
***Factor 3:	10	2	2	57	1	2
23 Only treat I had in my home	8	3	11	13	8	6
28 Portion controlled	5	1	1	27	1	1
26 A calorie-controlled treat	3		1	23		
27 A low calorie snack	3		1	21		1
***Factor 4:	67	60	46	64	98	61
04 Because I wanted a snack	41	19	35	31	64	45
12 Something sweet	37	27	23	34	70	25
21 Because I wanted something indulgent/special	26	24	20	20	51	12
22 I wanted something easy	25	19	11	20	42	27
17 To give a bit of 'me time'/relaxation time	22	11	28	15	40	12
10 Because it could be prepared/served quickly	16	16	6	18	26	14
30 I wanted to try something new	4	4	8	8		1
***Factor 5:	42	26	34	19	21	98
01 A reward	37	34	55	23	47	24
11 Refreshing on a summer day	36	22	30	14	19	81
02 To help my thirst	9	1	5	3	2	28
25 Can eat outside	7	4	3	3	1	19

Figure 2

Figure 2 - Indicates the percentage, by factor, of respondents who selected each of the following as occasions for purchase. Netting by Factors allows for quicker reference but also shows directionality in terms of group fitting.

Figure 3 – The table below illustrates how the Factors apply to Gender groupings. Significance testing is applied to show which variables reach the threshold of significance when comparing differences between groups. The Capitol Letters Represent a 95% threshold of significance and the lower-case letters represent a 90% significance threshold



1	Table 2-1						
2	Q1 Gender						
3	Base: Total sample						
4		Total Sample					
5		Total Occasions	Family Bonding (A)	Pick Me Up (B)	Controlled Treat (C)	Indulgent Me Time (D)	Refresh Me (E)
6	Total Respondents	3371	665	668	475	820	743
7	Total Weighted	3394	651	644	473	831	796
8	Male	1642	289	382	231	365	375
9		48%	44%	59%	49%	44%	47%
10				ACDE			
11	Female	1748	360	262	241	464	421
12		52%	55%	41%	51%	56%	53%
13			B		B	B	B
14	Other	4	1			2	
15							

Figure 3

Key Drivers Analysis

After using various clustering techniques including those described previously, Key Drivers Analysis can be completed by using a regression analysis on a dependent variable (e.g. Brand appeal) and associated independent variables. The aim of this analysis is to determine what attributes are drivers/more important in describing Brand appeal. The closer the relationship the more significantly the independent variable is a driver for the dependent variable. (i.e. higher regression coefficient implies greater importance).

Data Layering/Stacking

Data layering, in survey research, is often used in situations where respondents are asked a repetitive battery of questions related to concepts such as; brand performance, usage occasions, patient profiles, daily dairies, etc. In these situations, layering the data gives a market perspective of **Total** brands, occasions, or patients as opposed to merely the individual brand perspective. It allows for a different dimension by which to examine the data that is based on total responses rather than individual respondents. It is an essential technique in order to transform data to be used in understanding how the market performs as a whole thereby creating a new matrix of variables that can be used on their own or within larger models. For example, layered data is often used in Key Drivers Analysis to answer questions that deal with what is driving the overall market versus what is driving individual brands.

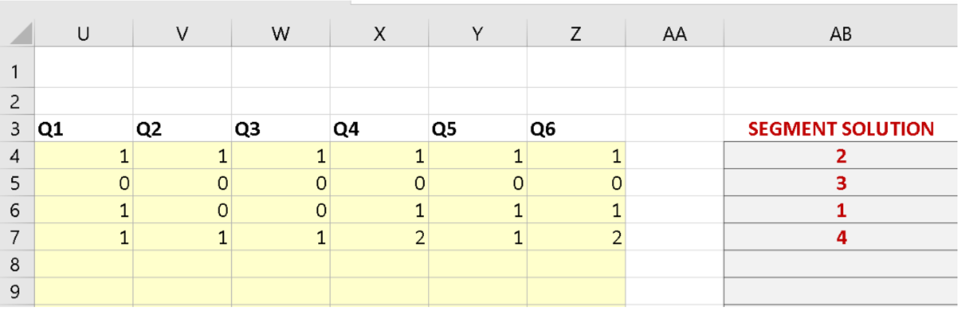
Overt (Stated)/Covert (Derived) Analysis

Researchers have argued over the relative merits of Stated importance (overt) versus Derived (Covert) importance for some time. Stated importance is simply asking respondents how important they rate a list of attributes using an importance scale (usually 1-10). By ranking the mean scores produced in a cross tabulation, one can order the importance of various attributes as it applies to the market. Stated importance often reflects the “ideal” situation and is not necessarily tied to real products/brands in the market. Derived importance is more complicated in that it seeks to understand what is important in the “market” using actual market brands/products to drive importance. In this case, importance is derived from **layering** that brand data together and correlating/regressing brand ratings with a given measure of overall ranking (e.g. Performance/Satisfaction). The correlation coefficients (which are derived irrespective of brands) indicate the relative importance of attributes within the market as defined by the brands. It is often quite useful to study the interaction between the 2 measures (stated vs. derived importance) by graphing and/or mapping the resulting quadrant data together. One can often observe opportunities that are not met by existing brands as well as address their relative strengths and/or weaknesses.

Segmentation Typing Tool

Once a segmentation algorithm has been successfully established, it is often useful to produce a user-friendly point and click solution for employees that allows for this segmentation model to be applied to new respondents. A segmentation typing tool allows anyone with limited knowledge in excel to answer survey questions, or apply an entire new set of respondent data, with simple copy and paste measures into a specified field to instantly produce corresponding segment results as the output.

Example: When data is input into the corresponding question fields on the left, an automatic segment solution is calculated based off a discriminant function created from the fisher’s coefficients of a previously tested model. This allows anyone to accurately segment respondents without the need for in depth statistical knowhow.



	U	V	W	X	Y	Z	AA	AB
1								
2								
3	Q1	Q2	Q3	Q4	Q5	Q6		SEGMENT SOLUTION
4	1	1	1	1	1	1		2
5	0	0	0	0	0	0		3
6	1	0	0	1	1	1		1
7	1	1	1	2	1	2		4
8								
9								