Data Analysis on the Questionnaire about the Chocolate Bars

Xin Tan, 583833

Introduction

This report is structured by firstly demonstrate the summery of the survey results. Then followed by the Exploratory Factor Analysis (EFA) to the attributes rating. the EFA helps reducing the complexity by reducing the number of dimensions in the data, which enable us to discover the potential factors to explain e.g. what might influence people's judgment about the chocolate bars. The perceptual Mapping generated based on the similarity matrix will visualize the relationship between the relationship of the attributes and the brands. 这个需要重新调整：Respondents will be clustered into two groups according to main attributes rating afterwards. At last, the interesting findings about the consumers and products would be discussed.

About the Questionnaire and the Dataset

50 respondents from German were asked to complete the questionnaire about the satisfaction and the consumption behavior of chocolate bars. The questionnaire covered 10 chocolate bar brands and 13 attributes for each brand. The respondents were required to answered 35 questions, which touch upon three main components:

- **Satisfaction of the chocolate brands and attributes**. 50 respondents need to evaluate both brands preference and attribute preference by assigning the rating scores: In the brands rating, respondents are required to give a rating score between 1-7 for each of the ten brands; Whereas in the attribute rating, they need to evaluate the ten brands together with the 13 attributes by rating from 1 - 5 (13*10 = 130 rating scores are supposed to be given).
- **Consumption behavior** include e.g. when, why and under which circumstance to consumer a chocolate bar.
- **Demographic Questions** consists of several private information about the respondents.

Design of the Questionnaire (35 Questions)

Satisfaction of the chocolate brands and attributes	**Brands**	*How do you think **[Brand]** on the basis of the following **[Attribute]**?* 1 - strongly disagree; 5 - strongly agree	*brand1*	Snickers	*brand6*	Bounty
			brand2	Kinder Bueno	*brand7*	Kinder Riegel
			brand3	Twix	*brand8*	Balisto
			brand4	Mars	*brand9*	Lion
			brand5	KitKat	*brand10*	Duplo
	Attributes	*Preferences for **[Brands]*** 1 - not preferred at all; 7 - greatly preferred	*attr. 1*	Crunchy	*attr. 8*	Addiction
			attr. 2	Creamy	*attr. 9*	Accessible
			attr. 3	Sweet	*attr. 10*	Handy
			attr. 4	Chocolaty	*attr. 11*	Wrapping
			attr. 5	Healthful	*attr. 12*	Image
			attr. 6	Calorie	*attr. 13*	Commercial
			attr. 7	Rich		
Consumption behavior	**Frequency**	*How often do you consume Chocolate Bars?*				
	Place	*Where can you find yourself buying Chocolate Bars?*				
	Situation	*Under which circumstances do you consume Chocolate Bars?*				
	Consumed	*Which of the following Chocolate Bars have you ever consumed? (10 brands)*				
Demographic Questions (About the Respondents)	**Gender**	*What is your gender?*				
	Age	*How old are you?*				
	Occupation	*What is your main occupation?*				
	marital status	*What is your marital status?*				
	Children	*Do you have children? If yes, how many?*				
	City	*Where do you live?*				
	State	*In which state of Germany do you live?*				
	Sport	*How often do you practice any kind of sport?*				

Table 1. A Brief View of the Questionnaire

The main problem of this data set is small size with uneven distributed demographic variables, which limits the inference of the population. This data set would not suitable for studying e.g. whether having children is a significant reason to increase the probability of purchasing chocolate bars? Since there are only three respondents having children. However, with this data set, we can analyze some of the potential factors of consumer preferences by scoring each chocolate brand and its attributes by 50 respondents. And then cluster the respondents into different segments to see the characteristics of each cluster. This report focuses only on the application of the data analytical methods, e.g. factor analysis and cluster. Neither the data collection as well as the design of the questionnaire, nor the data quality will not be discussed.

Data Preparation

Missing values are mainly due to the no knowledge or experience about the corresponding brands attributes.

Table 2. Where are the Missing Values

	Crunchy	Creamy	Sweet	Choco.	Health.	Calorie	Rich	Addic.	Access.	Handy	Wrapp.	Image	Comm.	**Sum**
Balisto	7	7	6	6	7	7	12	12	11	7	9	7	11	**109**
Bounty	3	3	2	2	3	2	7	9	6	3	5	3	8	**56**
Duplo	0	0	0	0	1	0	7	5	3	1	3	0	1	**21**
KinderB	2	1	1	1	5	1	8	5	2	1	2	0	2	**31**
KinderR	0	0	0	0	2	0	5	5	3	1	3	0	3	**22**
KitKat	0	0	0	0	4	0	5	5	2	1	3	0	2	**22**
Lion	7	7	6	7	6	6	11	10	10	8	10	9	11	**108**
Mars	0	0	0	0	4	0	5	4	2	1	3	1	4	**24**
Snickers	2	2	1	2	5	2	5	6	4	2	3	1	4	**39**
Twix	0	0	0	0	3	0	5	4	3	1	3	0	3	**22**
Sum	**21**	**20**	**16**	**18**	**40**	**18**	**70**	**65**	**46**	**26**	**44**	**21**	**49**	**454**

In total there are 6.4% missing inputs in the ***Attribute Rating***. *Table 2* displays the exact 454 missing values with respect to products and attributes.

- None of the products received a complete feedback when the evaluation of the products together with the attributes. Balisto (109) and Lion (108) include the most missing values, which are significantly more than the other products. Duoplo (21) has the least missing value.
- Six attributes with the most missing values: rich (14% or 70 missing values), addiction (13% or 65), commercial (9.8% or 49), accessible (9.2% or 46), wrapping (8.8% or 44) and healthful (8% or 40). In these six attributes, half of the missing values are from the products Balisto, Lion and Bounty.

Imputation of Missing Data

The dataset is cleaned by imputing the missing values with corresponding attributes mean in terms of the brands. By checking the distribution of each attribute rating scores with respect to the brands, the mean and median are very close. The rating scaler is limited within 1 to 5, outliers would have very small impact. Imputation the missing value with mean or median deliver almost the same distribution for of the attributes (only a slightly impact on the calorie, rich and addiction).

Respondents

There are 29 women and 21 men have filled out a questionnaire. The average age are 26 years old (mean = median = mode). The youngest respondent is 18 years old, the oldest is 31 years old. Students and working people are almost half and half (25:23). Only two respondents have got married. For the non-married respondents, 36% are in a relationship and 56% are still single. 47 respondents have reported their family status, 44 do not have children. 50 respondents are from nine states of Germany. The majority of the respondents are from Berlin (38%), then followed by Sachsen-Anhalt (24%), Hessen (8%), Nordrhein-Westfalen (10%), Bayern (6%), Niedersachsen (6%), Sachsen (4%), Hamburg (2%) and Baden Württemberg(2%). 46 respondents live in the city and 4 live in a village. Above 80% respondents do exercise at least once a week, most respondents (58%) do sports 1-3 times per week. For these 50 respondents, students go to sport more often than working people (26:24), female go to excise more often than male (29:21).

Consumption Behavior

The consumption behavior of the respondents, which includes consumption frequency, purchase location, consumption circumstances and consumed brands:

How often do you consume Chocolate Bars?

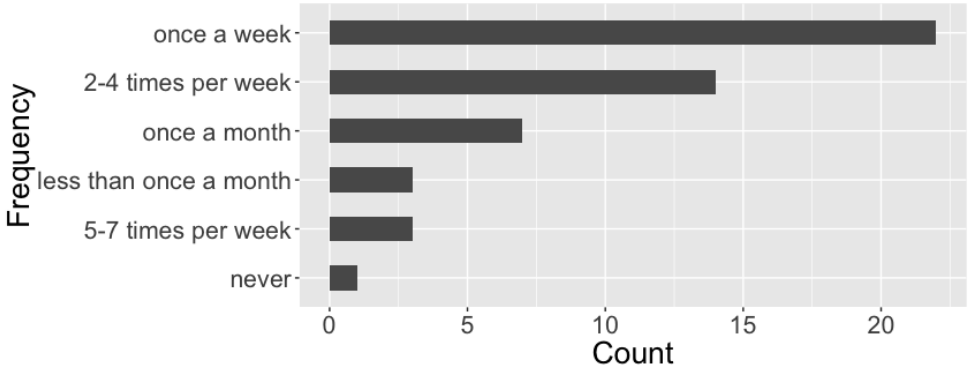


Figure 1. The Frequency of Chocolate Bar Consumption

78% respondents consume chocolate bar at least once a week. Only one respondent never consumes any chocolate bars.

Where can you find yourself buying Chocolate Bars?

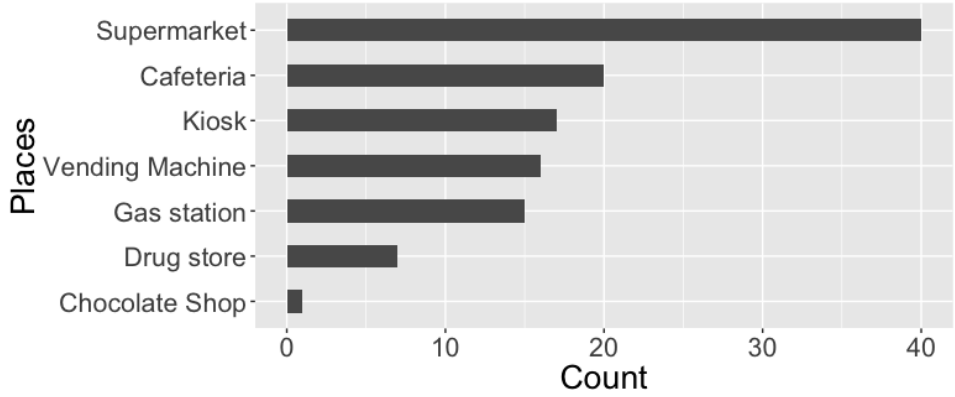


Figure 2. The Most Popular Places to Buy Chocolate Bars

Supermarket is the most popular place to buy chocolate bars. One quarter of the respondents only go to supermarket to buy chocolate bars. 66% (33/50) respondents purchase chocolate bar at more than one places. The second and the third most frequently places that to buy chocolate bars are cafeteria and kiosk. Quite few people buy chocolate bars at drug store or chocolate only shops like Ritter sport Shop.

Under which circumstances do you consume Chocolate Bars?

Same as the above question about where people buy chocolate bars, here the reason of consuming chocolate bars is also not unique for the most respondents. 74% respondents would like to eat chocolate bars under more than one circumstance. The most common reasons for consuming chocolate are:

- Being hungry (50% respondents)
- Under pressure (46%)
- Travelling or driving (44%)
- Watching TV (42%)
- As a treat (42%)
- As dessert (40%)

Which of the following Chocolate Bars have you ever consumed?

The questionnaire lists the 13 chocolate-bar-brands. In this part, consumers need to select out the chocolate brands they have ever consumed and then give their evaluation to each chocolate brands as well as the corresponding attributes in the next part. There are three levels of chocolate according to the popularity of the chocolate bars, which measured by the total number of the chocolate bars selected by the respondents.

Table 3 Feedback Amounts of with Respect to The Brands

Products	Duplo	Snickers	Twix	KitKat	KinderR	Mars	Bounty	Lion	Balisto	KinderB
Feedbacks	50	49	49	49	48	48	45	41	40	0

1. Most Popular. All of the respondents have ever consumed Duplo (50).
2. Very Popular: Snikers, Twix, KitKat, Kinder Riegel and Mars.
3. Well known: Bounty, Lion and Balisto.
4. Least Popular: Kinder Bueno.

Analysis based on the rating scores

1. Kinderriegel (5.9)
2. Snickers (5.5)
3. KinderBueno(5.4)
4. Twix (5.2)
5. KitKat (5.0)
6. Duplo (4.6)
7. Lion (4.5)
8. BalistoKornMix(4.3)
9. Bounty(4.0)
10. 10.Mars (3.9)

According to the rating on brands, most favored chocolate bars' brands is Kinder Riegel, whereas Mars is least preferred by the respondents.

If we look at the attributes-per-brands rating, we will have another result, since it tells another story: The **Brands with highest average mean scores** indicate that they have relative complete products lines which offer chocolate bars with different oriented attributes or combination of different attributes. Such as, Kinderriegel (3.9), Snickers (3.7), Twix (3.7) and Duplo (3.7).

The Attributes with highest average mean scores indicate that the people's perception of chocolate bars when people think about chocolate bars, these are the most common attributes one should have. Such as sweet (4.5), calorie (4.4) and accessible (4.1).

	crunchy	creamy	sweet	choco.	health.	calorie	rich	addic.	access.	handy	wrapp.	image	comm.	Brand.Avg
Balisto	4.5	1.9	4.0	3.9	2.3	4.1	3.4	2.9	3.5	4.0	3.4	3.2	3.2	3.4
Bounty	2.0	3.2	4.4	3.4	1.7	4.2	3.5	3.0	4.0	4.0	3.9	3.4	2.9	3.4
Duplo	4.0	2.5	4.5	4.3	1.6	4.3	3.1	3.3	4.0	4.1	3.6	4.1	4.4	**3.7**
KinderB.	3.8	4.4	4.6	3.8	1.5	4.3	3.5	3.6	3.8	3.8	3.7	3.4	3.0	3.6
KinderR.	1.7	4.1	4.7	4.5	1.7	4.3	3.4	4.0	4.5	4.4	4.0	4.5	4.4	**3.9**
KitKat	4.5	2.2	4.3	4.1	1.6	4.3	3.3	3.2	4.1	3.8	3.5	4.0	3.9	3.6
Lion	4.2	3.4	4.6	4.0	1.6	4.6	3.8	3.3	3.8	4.0	3.6	3.4	3.2	3.6
Mars	1.8	4.3	4.8	3.9	1.5	4.6	3.7	2.9	4.3	4.1	3.5	4.2	4.0	3.6
Snickers	3.6	3.4	4.4	4.0	1.4	4.6	3.6	2.9	4.3	4.3	3.3	4.4	4.3	**3.7**
Twix	4.1	3.4	4.5	3.8	1.5	4.5	3.6	3.3	4.3	4.0	3.6	4.0	4.1	**3.7**
Attri.Avg	3.4	3.3	**4.5**	4.0	1.6	**4.4**	3.5	3.2	**4.1**	4.0	3.6	3.8	3.7	

Table 4. Attributes Mean Scores

The heatmap based on the Attribute mean scores (Table. 4) re-arrange the order of the columns by putting the attributes with the most similar rating scores together, which simplify the inspection on:

• **Which attribute from which brand satisfies the respondents most/least?** (the darker of the color, the higher of the average rating score) e.g. Balisto is the only chocolate bar earns a relative higher score in healthful attribute which is also the least-sweet-taste chocolate bar (lowest sweet score).

• **Which attributes / brands are similar to each other in terms of rating scores?** If we look at the attributes, we can see respondents on average tend to give high rating scores on "sweet", "calorie", "accessible" and "handy", which indicate in general, chocolate bars are with these attributes. In general, people do not think chocolate bars are healthful. Brands, comparing with attributes, do not have very closed pairs as attributes (measure by the height of the dendrogram), the two brands with the most similar rating scores cross attributes are Duplo and KitKat. Balisto is the healthiest chocolate bar.

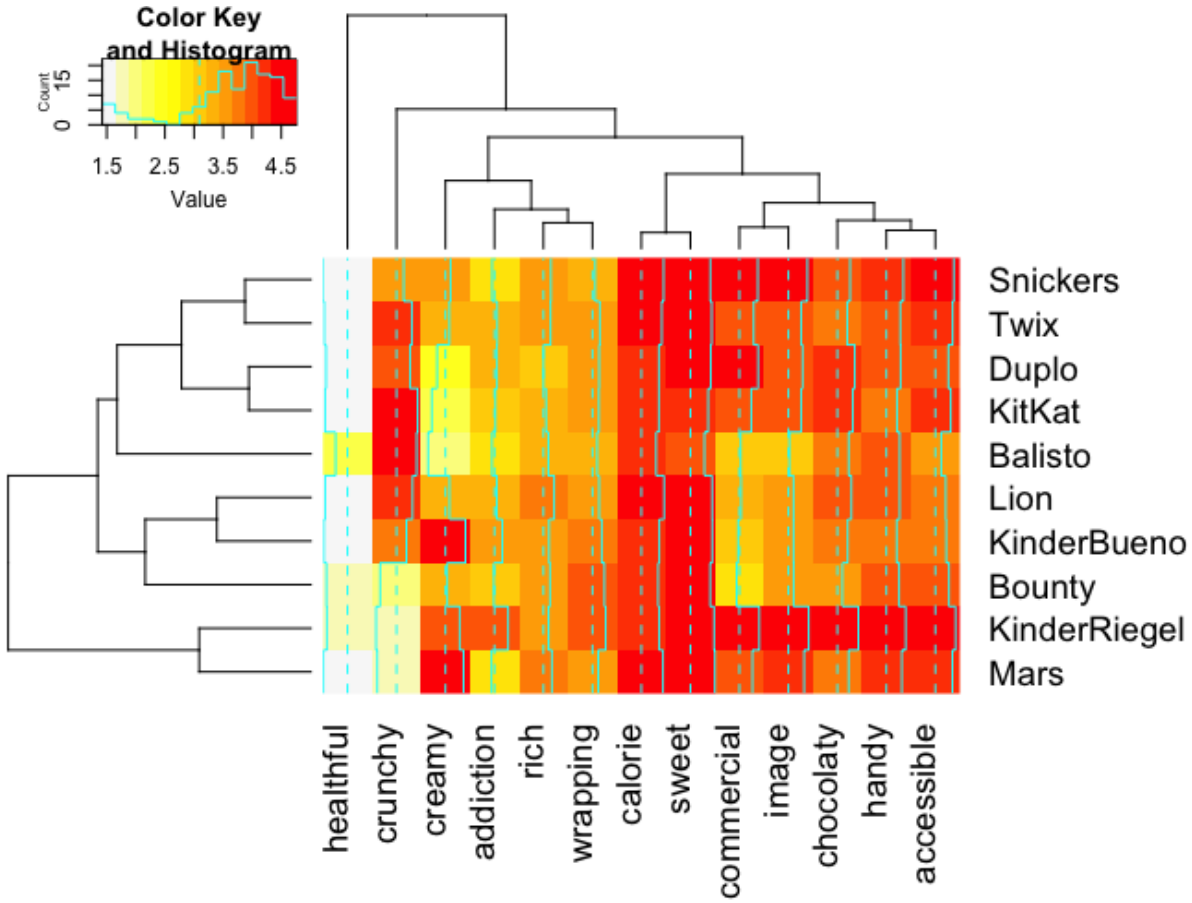


Figure x. The heatmap of the Attributes Mean Scores.

Rating score range from 1 to 5. 1 is white, 5 is red. The light blue line indicates the distribution of the rating score. The dotted line is the average score. The overall mean score is 3.1 and most of the attribute ratings are above average. In the big picture, if the solid line is on the left side of the dotted line, it means that the corresponding attribute is rated lower than the total average.

Correlation of the attributes allow us to check is there any relationship among attributes?

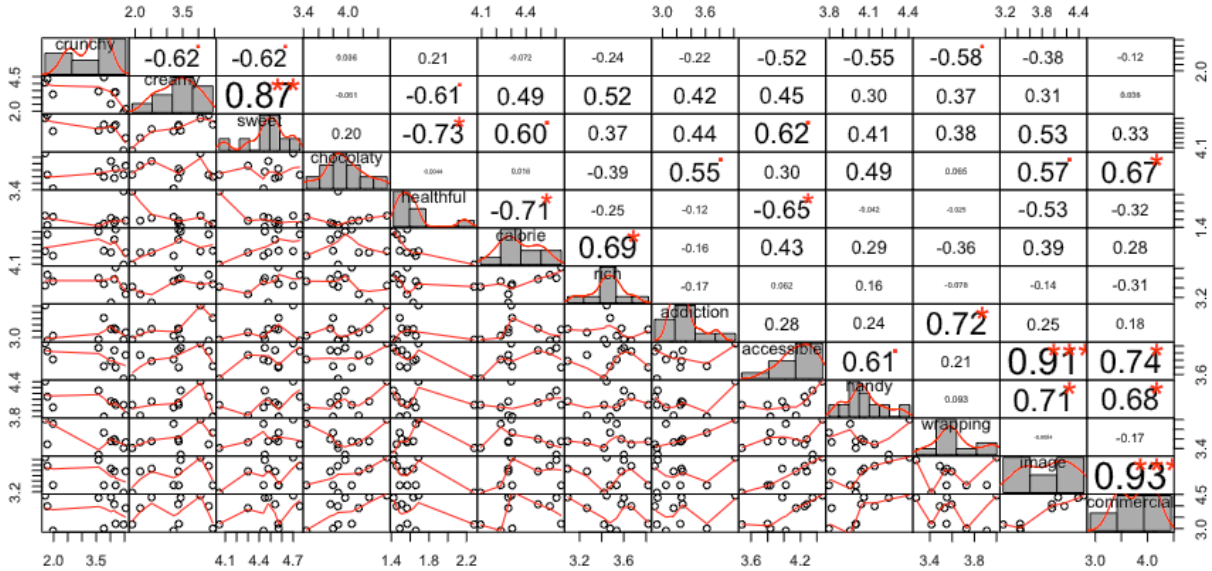


Figure x. Correlation Matrix of the attributes rating.

*The distribution of each variable is shown on the diagonal. On the bottom of the diagonal : the bivariate scatter plots with a fitted line are displayed On the top of the diagonal : the value of the correlation plus the significance level as stars Each significance level is associated to a symbol : p-values(0, 0.001, 0.01, 0.05, 0.1, 1) correspond symbols("***", "**", "*", ".", " ")*

• Both commercial and accessible are significantly positive correlated with image

• Creamy is positive correlated with sweet, but negative correlated with crunchy.

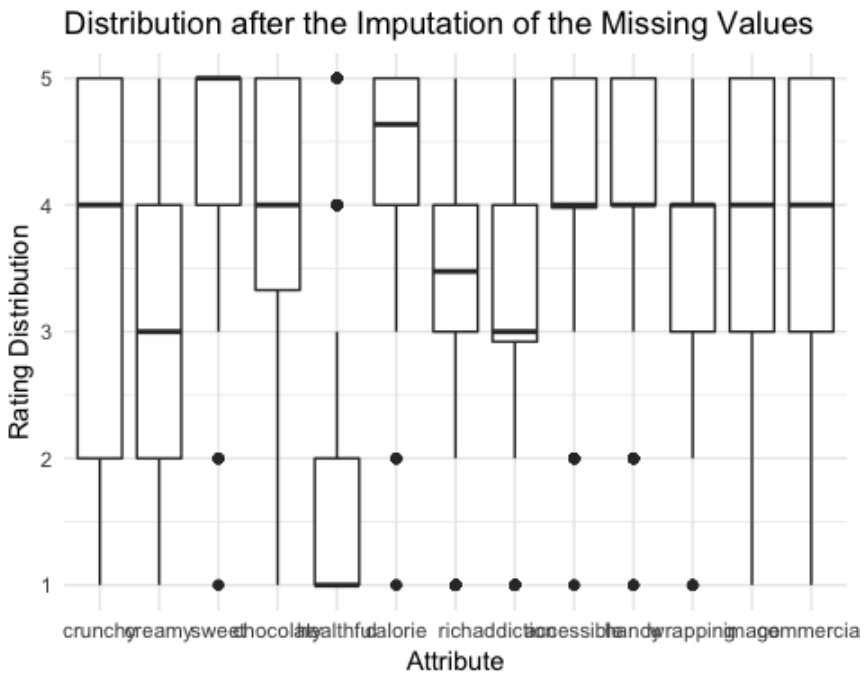
• Healthy is negative correlated with sweet and calorie.

• Healthful and accessible are negative correlated. Why?

+ Wrapping-Addiction: Packaging design has positive impact on the making chocolate bars more attractive and stimulate appetite.

关于设计，例如 Image, commercial, handy, accessible 会对分数有同向影响。

不同品牌巧克力在不同 attributes 上的投影看程度。Mds



Next, we want to know what the underlying factors are implied by consumers‘rating. The EFA method will be employed.

Mean Score for male and female

计算出率后（占比多少的人喜欢什么）还要计算 sd 和 ci sd = sqrt(10%(1-10%))/总调查人数，t_0.05 = 2.02 95%CI = 10% ± 2.02*sd. Interpretation: 如果用样本的喜欢率 10%来估计总体时，那么有 95%的可能在 10% ± 2.02sd 之间，ci 越接近 10%越可靠。

Part 2. Exploratory Factor Analysis

we use the brand-attribute rating data to ask the following questions: How many latent factors are there? How do the survey items map to the factors? How are the brands positioned on the factors? What are the respondents' factor scores?

EFA based on the correlation metrics is good at uncovering latent structure and attempt to find a factor structure. It produces results that are very interpretable in terms of the original variables. Furthermore, EFA offers the possibility to check whether the attributes in fact go together in a way that can be interpreted as a single factor, or whether they instead reflect multiple dimensions that we might not have considered

Dimension reduction through the factor analysis is achieved by extracting and synthesizing the overlapping parts among variables in the original dataset into several factors. This requires the correlation among variables not being zero.

Suitability of factor analysis

Factor analysis is based on a covariance matrix between variables and assumes that some factors linearly influence the observed model. In other words, the candidate variables (attributes) must have a certain correlation. If there is no correlation between the variables, or the correlation is small, the factor analysis will not be a suitable analysis method. The Kaiser-Meyer-Olkin measure of sampling adequacy (MSA) shows, this dataset is suitable for the degree of factor analysis, since KMO statistic is larger than 0.5. The MSA for individual variables are printed as the diagonal elements of the Anti-image Correlation matrix. The Bartlett's test for Sphericity compares the correlation matrix to the identity matrix. It checks if there exists relationship between variables that can be summarized with some factors. In this case, it rejects the null hypothesis that the correlation matrix is an identity matrix at 5% level of significance, which indicates the 13 attributes variables are related and therefore suitable for structure detection.

Bartlett's Test

chisq	1235	**p.value**	1.4e - 207	**df**	78

Kaiser-Meyer-Olkin factor adequacy

Overall MSA = 0.7
MSA for each item:

crunchy	creamy	sweet	chocolaty	healthful	calorie	rich
0.45	0.65	0.75	0.71	0.55	0.65	0.74
addiction	accessible	handy	wrapping	image	commercial	
0.76	0.79	0.74	0.77	0.69	0.65	

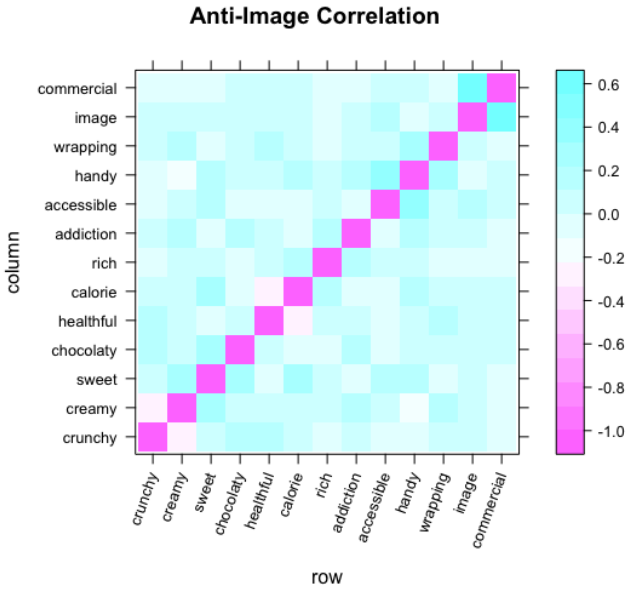


Figure.x Factorability check

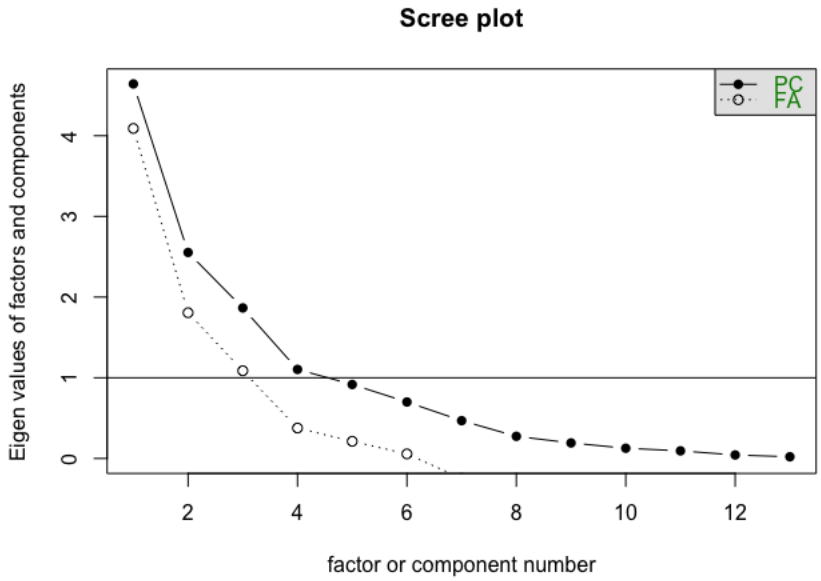
Sample size

The sample size should be large enough to yield reliable estimates of correlations among the variables. EFA can be reasonably with N/k (Cases / Items) > 5/1. The attribute-brand rating dataset include 13 variables and 500 rows (500/13 = 38).

The Number of Factors

We hope that the number of factors should be much smaller than the number of distant variables, while at the same time requiring the retained factors to keep as much information as possible of the original variables. Here we use the eigenvalue method to determine the number of retention factors.

The eigenvalues display the variation that can be explained by the corresponding factors. Factors with an eigenvalue greater than one would be retained. The scree plot visualizes the relationship between the number of factors and the corresponding eigenvalues. 4 components would be selected to do the following factor analysis. On the one hand, the increase in the number of the factors would not bring an increase in the marginal proportion of variance, on the other hand, although 3 factors are suggested, the interpretation power is less than 4 factors.



The Rotation of Factors

The rating variables themselves are ranged from 1 to 5, to normalize the data here is not necessary. Therefore, raw data will be used directly for factor analysis.

In order to interpret the results more easily, rotating is needed. After comparing the oblimin rotation, which allows the dependence among factors, with the orthogonal rotation, which artificially forced the factors to be uncorrelated, I rotate the four-factor solution using orthogonal rotation, since the results is more interpretable. The maximum likelihood approach is employed to extract common factors. The first four factors account for 69 percent of the variance in 13 attributes.

ML2 captures the highest proportion variance (21%) and then followed by ML4 (18%), ML1(15%) and ML3 (15%). Healthful and crunchy are negatively correlated with all the four factors. Loading represents the strength of relationship between a factor and a variable. The first two factors contain the most loadings. Image and commercial are very close to each other in all four factors. If we look at the factor loadings. We can see that calorie, rich and sweet load on the first factor (ML2).
Commercial, image and chocolate load on the second factor (ML4). In most cases, the variability is captured by the four common factors achieve higher than 0.5 (measured by h2).

Factor Analysis using method = Maximum Liklihood
rotate = "varimax", scores = "Anderson", max.iter = 1000, fm = "ml"
Standardized loadings (pattern matrix) based upon correlation matrix

	ML2	ML4	ML3	ML1	h2	u2	com
crunchy	-0.19	-0.01	-0.27	0.55	0.55	0.446	1.5
creamy	0.68	0.02	0.73	1.00	1.00	0.005	2.0
sweet	0.60	0.04	0.41	0.66	0.66	0.343	2.5
chocolaty	-0.09	0.71	0.02	0.52	0.52	0.484	1.1

healthful	-0.65	-0.18	-0.13	0.56	0.56	0.436	1.7
calorie	0.92	0.11	-0.27	1.00	1.00	0.005	1.4
rich	0.67	-0.26	0.07	0.56	0.56	0.441	1.5
addiction	-0.07	0.17	0.6	0.4	0.40	0.603	1.2
accessible	0.29	0-36	0.24	1.00	1.00	0.005	1.8
handy	0.15	0.3	0.08	0.33	0.33	0.667	2.0
wrapping	-0.16	-0.2	0.48	0.52	0.52	0.478	2.6
image	0.16	0.79	0.21	0.89	0.89	0.108	1.8
commercial	0.03	0.92	0.00	1.00	1.00	0.005	1.3

	ML2	**ML4**	**ML3**	**ML1**
SS loadings	2.74	2.39	1.95	1.89
Proportion Var	0.21	0.18	0.15	0.15
Cumulative Var	0.21	0.39	0.54	**0.69**
Proportion Explained	0.31	0.27	0.22	0.21
Cumulative Proportion	0.31	0.57	0.79	1.00

loadings of the factors on the variables means the relationship of the matrix of factors to the original variables. EFA attempts to find solutions that are maximally interpretable in terms of the manifest variables. In general, it attempts to find solutions in which a small number of loadings for each factor are very high, while other loadings for that factor are low. Different from the Principle Component Analysis (PCA), EFA produces results that are interpretable in terms of the original variables

Naming the Factors

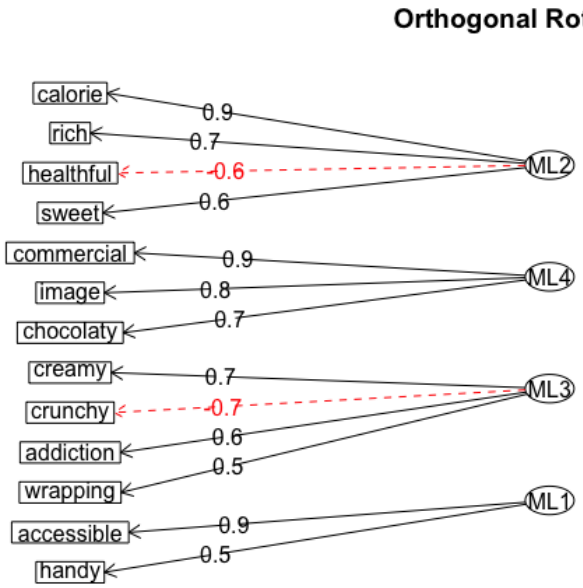
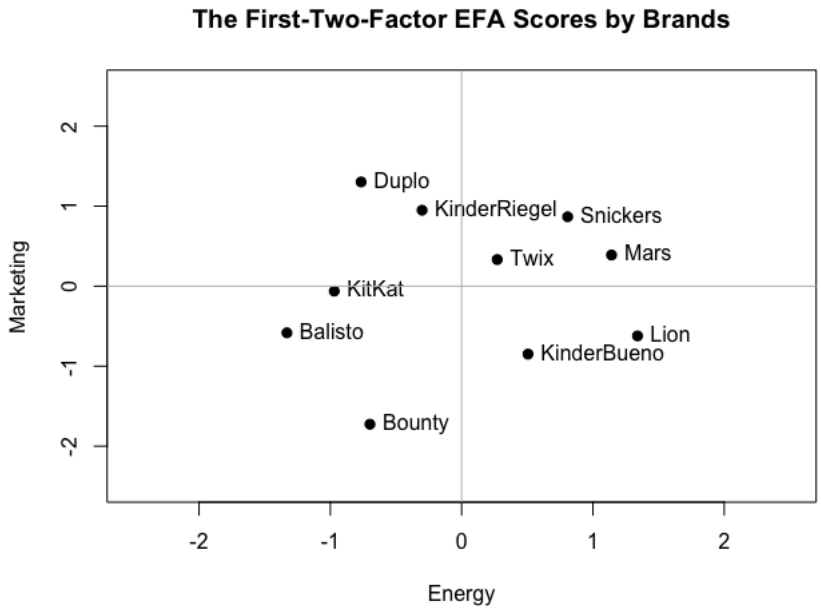


Figure.x loadings (|L| > 0.30)

In order to interpret the factors, let us focus on those attributes with loading > 0.3 by each factor.

• The loadings in ML2 are quite high, which seem very good. Calorie (0.92), rich (0.67) and sweet (0.603) have the highest loading in ML2 (0.92). Healthful (-0.65) locates its highest negative absolute loading also in ML2. We can conclude that ML2 is **the calories factor.**
• In ML4, except accessible, all the other attributes (commercial, image and chocolaty) have their largest loading there. Commercial and image are highly correlated with the advertisement and promotion in products, e.g. where to put the products in the supermarkets, how often and how is advertisement on TV? We can say that ML4 would be **the marketing factor**.
• In ML3, there are a pair opposite tasty attribute: crunchy (-0.67) and creamy (0.73). Wrapping also have a similar score in ML4, which does not contribute too much on explanation power to this factor. Since crunchy or creamy usually describe the filling taste within a chocolate bar, we consider ML3 be **the taste factor.**
ML1 can be seen as **the packaging factor**.

EFA Scores



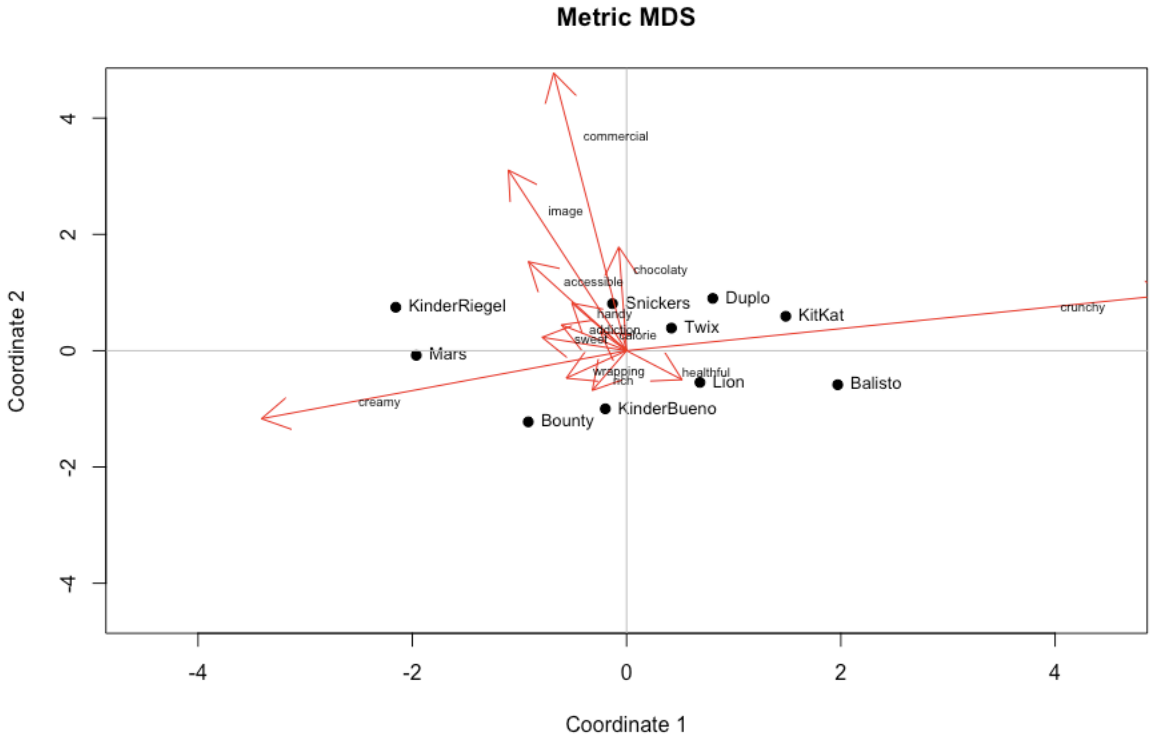
If we plot the first two factors by their scores, we can get the above figure, which is based on the Euclidean distance matrix (see Appendix), measures the similarity among brands. A lower Euclidean distance means the two products are similar to each other, taking all attributes into consideration. KitKat - Duplo (0.91), Snickers -Twix (0.96) are quite similar. KinderRiegel - Balisto (4.42) with the largest Euclidean distance means they are very dissimilar chocolate bars. The average Euclidean distance are 2.40 ± 0.81standard deviation.

The first two factors which highly influence the respondent's rating are the energy level (e.g. sweet and creamy chocolate bars are more attractive) and the marketing. Consistent with the earlier study, Balisto with the least energy level and has less marketing level, which might lead to respondents evaluate it with

a relative low mean rate score. Snickers, Mars and Twix have very high score on both factors. They are similar in such a way: high costumer recognition with rich, sweet and high calorie features, which replenish energy quickly. Most important information has been retained by EFA methods. e.g. Duoplo has the highest commercial EFA score, which is also consistent with the original mean score (4.4/5). Since EFA scores are driven by the correlation matrix, we can also interpret that consumers who like Mars would also like Snicker.

Metric-MDS Perceptual Map with Products and Attributes

Adding the regression attributes coefficients into the 2D metric-MDS perceptual map helps to know: How the products compare to each other? Which products are associated with which attributes? And which attributes are more closely related?



There are three obvious attributes point at three different directions: commercial to the north, creamy to the southwest, crunchy is opposite to the creamy. We can conclude that the coordinate 1 is the crunchy and creamy scale, whereas coordinate 2 indicates commercial level. Chocolate bars located on the left side of the coordinate 1 has creamy attribute, e.g. Mars, KinderRiegel. In contrast, chocolate bars like Kitkat, Twix, Lion and Balisto located on the right side of the coordinate 1 are crunchy. Similarly, chocolate bars on the upper side, e.g. KinderRiegel, Snickers and Duoplo are more commercial than the chocolate bars on the lower side of the map. Attributes with similar ratings are grouped together. Image and commercial

are closely related attributes, both of which are in the same direction and have contributed a lot to the explanation (long arrow). In contrast, healthy locates oppositely against commercial with very short arrow, whose direction contains few chocolate bars, e.g. Balisto. Healthy is an attribute with very low commercial level leads to customer's low rating score on it. The possible explanations are either the majority of the respondents don't think chocolate is healthy, so they will not give a high rating for this attribute, or healthy chocolate is a small niche that few producers are commercializing on. The products placement relative to the attributes helps to identify consumer impressions to certain products. KinderRiegel, Snickers and Mars are multi-attribute chocolate bars, they are in the direction with most attributes arrows. In contrast, Kitkat has quite simple attributes (crunchy). Chocolate bars with more attributes can satisfy a wider range of consumer preferences. Chocolate bars with a single attribute usually has very distinct characteristics. This can also be seen from the original attributes rating table. Usually the former will receive a more complete feedback. The latter will have a higher score in one specific attribute.

3. Cluster of Customers

In this section, respondents are going to be clustered into different groups based on their rating scores on the brands attributes. Different from the previous section, where missing values have been replaced by the brands attribute means, in this part, all the missing values would be replaced with zero, since NA means the respondents have no experience about the rating objective, e.g. they have never consumed Balisto, thus cannot give evaluations to any of the corresponding attributes. By doing so on the one hand, the important information of non-experience on the products has been kept, on the other hand, the distribution of attributes variables with non-zero-ratings are not be influenced.

The cluster results mainly checked **factor representative variables.** From the previous part we got the four factors (Calories, Marketing, Taste and Packaging), from which each factor, one represented variable would be picked out as the benchmark factors to compare among clusters. The variables are selected by going back to the original data set to check the attributes with the highest variance within each group. The factor representative variables tell us from which aspect the respondents' satisfaction are influenced? E.g. by the taste of the products or the marketing promotion of the companies?

Calorie,rich,healthful,sweet -> rich (1.563614)

Commercial, image, chocolaty -> commercial (1.569881)

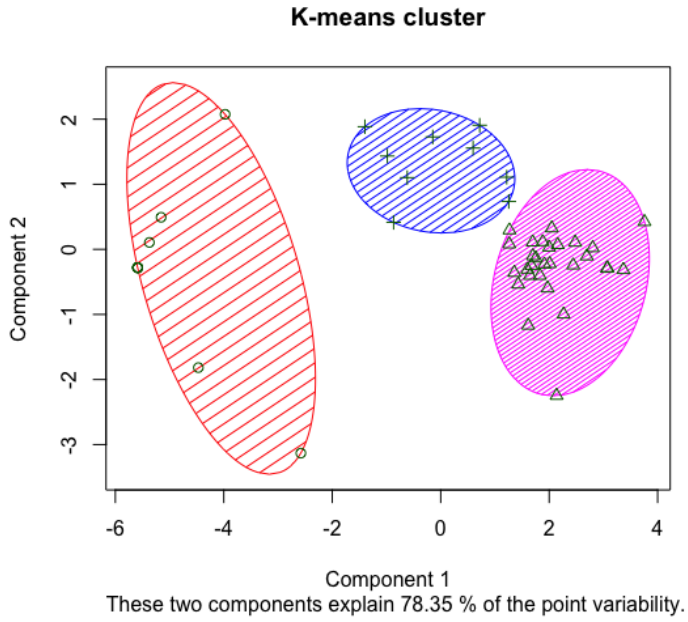
Creamy, crunchy, addiction, wrapping -> addiction (1.524589)

Accessible, handy -> accessible (1.427612)

Afterwards we will go back to the respondents to see the common characteristics of each cluster, e.g. the consumption frequency, age, gender, their favorite brands etc.

Cluster by Kmeans

By checking the *NbClust*[1] for determining the best number of clusters, according to the majority rule (proposed by 9 indices), the best number of clusters is 2. By silhouette structure check, 2 cluster solution is 0.61 (medium structure) and 3 cluster solution is 0.49 (weak structure). However, the 2 clusters solution just clustered the respondents into non-experienced and experienced groups, whereas from the interpretation point of view, the 3 clusters offered a better solution. Since we want to know more about e.g. what are the characteristics and consumption behavior within the experienced group? Thus 3 clusters would be chosen, which is the proposed by the second most indices (7 indices).



In general, chocolate attracts women more than men. Female have consumed more chocolate bars in different brands than male.

[1] See Appendix

Appendix

NbClust Package for determining the best number of clusters

```
NbClust(attribute.rating[-1], method="kmeans")
## *******************************************************************
## * Among all indices:
## * 9 proposed 2 as the best number of clusters
## * 7 proposed 3 as the best number of clusters
## * 1 proposed 10 as the best number of clusters
## * 1 proposed 11 as the best number of clusters
## * 1 proposed 12 as the best number of clusters
## * 3 proposed 14 as the best number of clusters
## * 2 proposed 15 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************

## $All.index
##        KL      CH Hartigan    CCC     Scott      Marriot    TrCovW
## 2  7.5067 83.8665  15.5234 8.0232  267.6280 9.582151e+18 4619.9354
## 3  2.6561 61.9385   7.0823 5.0338  357.0975 3.601831e+18 2565.7535
## 4  2.8523 48.8142  -0.0195 4.5585  432.7966 1.408921e+18 1811.4313
## 5  0.3905 35.7948   5.7164 3.0340  453.0248 1.468948e+18 2342.6837
## 6  1.1582 32.6742   5.1216 3.0155  559.5705 2.511452e+17 1829.3550
## 7  0.5444 30.5412   9.6125 3.1472  641.1974 6.680613e+16 1324.4585
## 8  1.6617 32.6266   6.5756 4.6697  749.0790 1.008678e+16  923.0276
## 9  1.8176 33.0341   4.0346 5.6091  847.1322 1.796299e+15  649.2693
## 10 0.5182 31.9039   8.0826 5.8416  881.2995 1.119746e+15  512.6137
## 11 1.4125 34.4407   6.4664 7.3592  996.1750 1.361789e+14  395.4782
## 12 9.4244 36.1379   1.0943 8.4488 1080.4324 3.004941e+13  289.5879
## 13 0.1095 33.2723   6.9226 7.7966 1117.0910 1.694136e+13  276.4250
## 14 4.3940 35.9920   1.8811 9.3818 1210.4183 3.038692e+12  203.5956
## 15 0.1272 34.3212  -8.2907 9.0424 1275.8048 9.433506e+11  197.3841
##      TraceW  Friedman   Rubin Cindex     DB Silhouette    Duda Pseudot2
## 2  701.2186  585.0722  9.5317 0.4110 0.6297     0.6148  0.6921  16.0131
## 3  529.8598  642.9725 12.6143 0.3316 1.2477     0.4554  0.8335   2.7960
## 4  460.4722  716.2712 14.5151 0.3909 1.2747     0.4185  4.7663 -21.3352
## 5  460.6677  708.6422 14.5089 0.3325 1.5443     0.2397  0.8437   2.4081
## 6  408.7441  908.3555 16.3520 0.3883 1.5427     0.2420  2.4467  -5.9128
## 7  366.1267  942.2373 18.2554 0.3069 1.4097     0.2647  0.7392   5.2917
## 8  299.2340 1123.5740 22.3363 0.3217 1.2151     0.2921  0.8535   1.8877
## 9  258.7270 1377.5219 25.8334 0.3054 1.1564     0.3115  0.5733  14.8836
## 10 235.5478 1364.7490 28.3755 0.2710 1.1107     0.3224  1.0233  -0.0228
## 11 195.9528 1553.7318 34.1092 0.3078 0.9853     0.3520  2.0045  -7.0158
```

```
## 12 168.0836 1759.3607 39.7647 0.3025 0.9311     0.3648 12.2626   0.0000
## 13 163.3786 1984.2909 40.9098 0.3018 0.8620     0.3884  1.9100   0.0000
## 14 137.6286 2178.4997 48.5640 0.3826 0.7684     0.4250  0.9703   0.3668
## 15 130.7942 2473.4893 51.1016 0.3528 0.8893     0.4043  1.3645  -1.8699
##      Beale Ratkowsky     Ball Ptbiserial    Frey McClain   Dunn Hubert
## 2   3.8459    0.5514 350.6093     0.8651  1.7789  0.2305 0.6671 0.0010
## 3   1.5529    0.4813 176.6199     0.7812  0.5648  0.5025 0.3033 0.0010
## 4  -5.2665    0.4269 115.1181     0.7844 -7.1921  0.5265 0.3781 0.0010
## 5   1.3169    0.3830  92.1335     0.5015  0.1110  1.4420 0.1144 0.0010
## 6  -4.7289    0.3553  68.1240     0.5105  0.0842  1.4694 0.1426 0.0010
## 7   2.9390    0.3330  52.3038     0.5276  0.1680  1.4749 0.1520 0.0010
## 8   1.1437    0.3184  37.4042     0.5327  0.0626  1.4969 0.1944 0.0011
## 9   6.1407    0.3036  28.7474     0.5377  0.6127  1.4733 0.1944 0.0011
## 10 -0.1011    0.2905  23.5548     0.5026  0.0111  1.7024 0.1859 0.0011
## 11 -4.0821    0.2806  17.8139     0.5057 -0.9545  1.6719 0.2181 0.0011
## 12  0.0000    0.2708  14.0070     0.5181  0.9818  1.5801 0.2181 0.0011
## 13  0.0000    0.2608  12.5676     0.5168  0.0458  1.5881 0.2181 0.0011
## 14  0.2507    0.2529   9.8306     0.5178  0.7351  1.5771 0.2804 0.0011
## 15 -2.1580    0.2450   8.7196     0.4876  4.6122  1.7635 0.2804 0.0011
##    SDindex Dindex   SDbw
## 2   0.6262 3.2080 0.4255
## 3   0.8924 2.8361 0.4331
## 4   0.9210 2.6789 0.4305
## 5   1.3230 2.5975 0.3060
## 6   1.2827 2.4827 0.3431
## 7   1.2790 2.3101 0.3748
## 8   1.2576 2.0765 0.2882
## 9   1.2108 1.9591 0.2708
## 10  1.2383 1.8455 0.2464
## 11  1.1215 1.7025 0.1993
## 12  1.0813 1.6193 0.2022
## 13  1.0525 1.5579 0.1800
## 14  0.9402 1.4144 0.1176
## 15  1.1919 1.3781 0.1179
##
## $All.CriticalValues
##    CritValue_Duda CritValue_PseudoT2 Fvalue_Beale
## 2          0.7539            11.7537       0.0000
## 3          0.5459            11.6457       0.1140
## 4          0.4075            39.2588       1.0000
## 5          0.4549            15.5808       0.2337
## 6          0.5846             7.1064       1.0000
## 7          0.6563             7.8570       0.0006
## 8          0.4075            15.9943       0.3548
## 9          0.6372            11.3864       0.0000
## 10         0.2493             3.0108       1.0000
## 11         0.6139             8.8034       1.0000
## 12         0.0916             0.0000          NaN
## 13         0.0916             0.0000          NaN
## 14         0.6262             7.1631       0.9964
```

```
## 15         0.6002             4.6632       1.0000
##
## $Best.nc
##                       KL      CH Hartigan     CCC    Scott      Marriot
## Number_clusters 12.0000  2.0000  15.0000 14.0000  11.0000 3.000000e+00
## Value_Index      9.4244 83.8665  10.1718  9.3818 114.8754 3.787409e+18
##                     TrCovW   TraceW Friedman   Rubin Cindex     DB
## Number_clusters    3.000   3.0000  15.0000 14.0000 10.000 2.0000
## Value_Index     2054.182 101.9712 294.9896 -5.1165  0.271 0.6297
##                  Silhouette   Duda PseudoT2  Beale Ratkowsky     Ball
## Number_clusters     2.0000 3.0000    3.000 3.0000    2.0000   3.0000
## Value_Index         0.6148 0.8335    2.796 1.5529    0.5514 173.9893
##                  PtBiserial   Frey McClain   Dunn Hubert SDindex Dindex
## Number_clusters     2.0000 2.0000  2.0000 2.0000      0  2.0000      0
## Value_Index         0.8651 1.7789  0.2305 0.6671      0  0.6262      0
##                     SDbw
## Number_clusters 14.0000
## Value_Index      0.1176
```