Data Analysis Based on the Questionnaire of the Chocolate Consumption

*Xin Tan, 583833*

# Part 1 Introduction

This report is based on the questionnaire of the chocolate bar consumption. Firstly, the questionnaire and the summery of the survey results will be demonstrated. Then the EFA (Exploratory Factor Analysis) will be introduced to help reducing the number of dimensions of the data, which enable us to discover the potential factors to explain e.g. what are the potential factors that could influence people’s judgment about the brands-attributes rating of the chocolate bars. The perceptual Mapping generated based on the distance matrix will visualize the relationship between the relationship of the attributes and the brands. At last, Respondents will be clustered into three groups according to their attribute-brand-rating scores and some interesting findings about the consumers and products would be discussed.

# Part 2. About the Questionnaire and the Dataset[[1]](#footnote-1)

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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Table 1. A Brief View of the Questionnaire*** | | | | | | |
|  | **Brands** | *How do you think* ***[Brand]*** *on the basis of the following* ***[Attribute]****?*  1 - strongly disagree;  5 - strongly agree |  |  |  |  |
| **Satisfaction of the chocolate brands and attributes** | *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?* | | | | |

**The main problem of this data set** is small sample 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 the survey sample include 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:

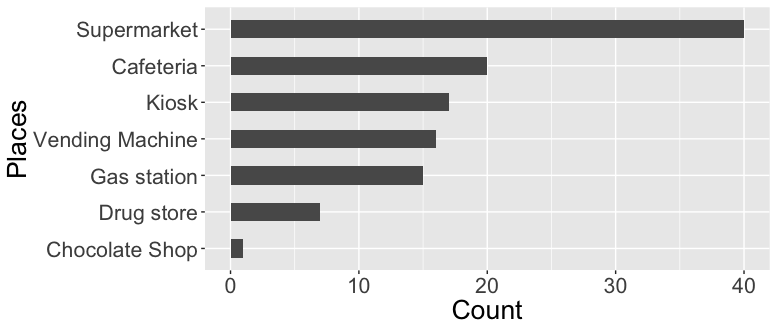
### How often do you consume Chocolate Bars?

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

|  |  |  |  |
| --- | --- | --- | --- |
| ***Table 3. The Frequency of Chocolate Bar Consumption*** | | | |
| **Low** | **never** | 1 | 4 (8%) |
| **less than once a month** | 3 |
| **Medium** | **once a month** | 7 | **29 (58%)** |
| **once a week** | 22 |
| **High** | **2-4 times per week** | 14 | 17 (34%) |
| **5-7 times per week** | 3 |

### Where can you find yourself buying Chocolate Bars?

***Figure 1. 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 4 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.

**The Direct Brand Rating Scores of the Chocolate Bars**

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.

**The Brand-Attribute Rating Scores of the Chocolate bars**

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).

***Table 5. Attributes Mean Scores***

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **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 |  |

**The heatmap** based on the Attribute mean scores (Table. 5) 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.

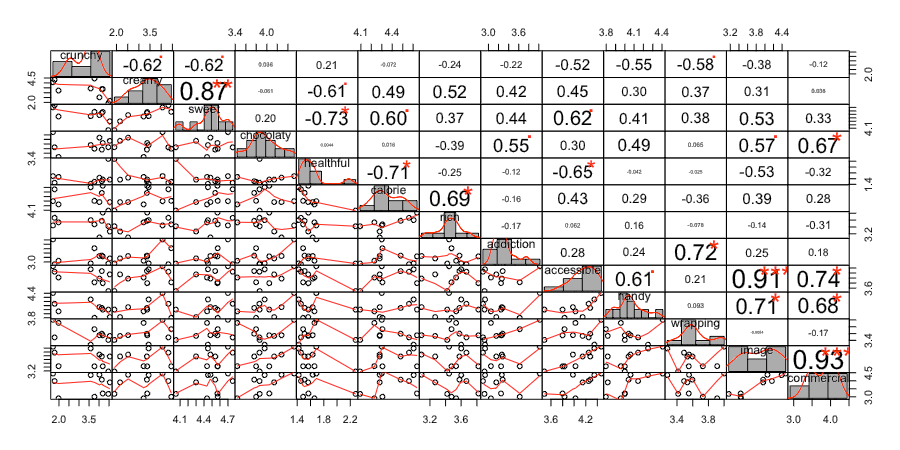
图片包含 文字

描述已自动生成

***Figure 2. 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 3. 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.

• Healthful is negative correlated with sweet and calorie.

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

In the next part, the EFA method will be employed to discover the underlying factors based on the respondents’ brands-attributes rating scores.

Part 3. Exploratory Factor Analysis

Recall, the brands-attributes rating includes 500 rows and 13 columns. Each respondent has evaluated 13 attributes for each ten brands (1 is disagree and 5 is strongly agree): How do you think **[Brand]** on the basis of the following **[Attribute]**?

*According to the brands-attributes ratings the following questions would be asked: How many latent factors are there? How do the survey items map to the factors? How are the brands positioned on the factors?*

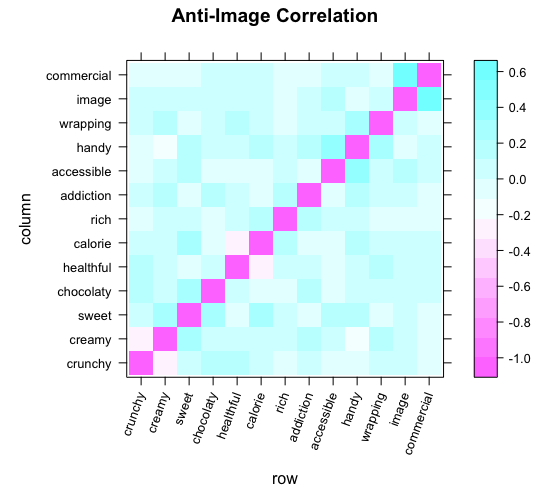
EFA based on the correlation metrics aims at uncovering latent structure and attempt to find a factor structure. It produces results that are very interpretable in terms of the original variables. Furthermore, EFA offers the possibility to check whether the attributes in fact go together in a way that 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.4 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.

图片包含 地图

描述已自动生成

**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 oblique 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).

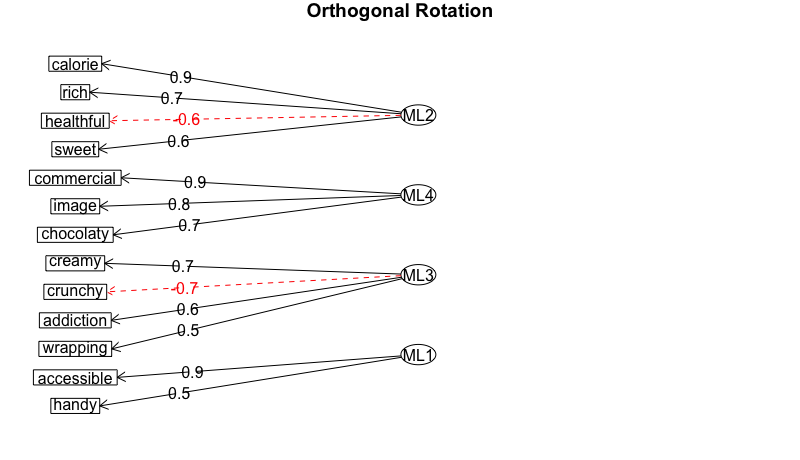
Commercial, image and chocolate load on the second factor (ML4). In most cases, the variability is captured by the four common factors achieve higher than 0.5 (measured by h2).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |  |  |  |

*Table 6. Loading of the factors*

**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. 6 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 energy 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**

图片包含 屏幕截图

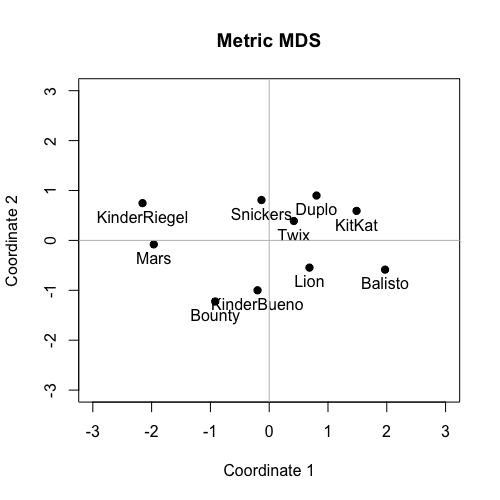
描述已自动生成

*Figure. 7*

If we plot the first two factors by their loading scores with respect to the brands, we can get the above figure. Since these first two factors captured the most variation of the dataset. Rating based on the attributes across brands are mainly influenced by the nature characteristics of the chocolate products, like sweet and high calories, at the same time the marketing strategy to promote the products could also be influential. 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. If we check the heatmap (figure 2) we can find, in general the respondents like creamy and sweet chocolate bars more than crunch and low calorie one.

**Metric Multidimensional scaling (MDS)**

Instead of extracting underlying components or latent factors, the MDS works instead with distances (here the Euclidean distance[[2]](#footnote-2) will be used). MDS attempts to find a lower-dimensional map that best preserves all the observed similarities between brands. A lower Euclidean distance means the two products are similar to each other, taking all attributes into consideration. In this case, the relative position of the brands has changed. E.g. Mars is used to very closed to the Snickers in the EFA marketing – energy coordination, but if we take all raw attributes distance matrix as the benchmark, Mars could be far from Snickers.



|  |  |  |
| --- | --- | --- |
|  | Dim.1 | Dim 2 |
| Balisto | 1.97093 | -0.58578 |
| Bounty | -0.91776 | -1.22449 |
| Duplo | 0.80465 | 0.89914 |
| KinderBueno | -0.19838 | -0.99981 |
| KinderRiegel | -2.15483 | 0.74704 |
| KitKat | 1.48594 | 0.59167 |
| Lion | 0.68481 | -0.54548 |
| Mars | -1.96296 | -0.07976 |
| Snickers | -0.13151 | 0.80897 |
| Twix | 0.41912 | 0.38850 |

Part 4. Cluster of Customers

Now we will go back to the respondents to see if there exist any consumption patterns for those behave similar in the brands-attributes evaluation. e.g. the consumption frequency, sports frequency, their favorite brands etc.

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.

By checking the *NbClust[[3]](#footnote-3)* for determining the best number of clusters, according to the majority rule (proposed by 9 indices), the best number of clusters is 2.

The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters. It ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. 2 cluster solution is 0.61 (medium structure) and 3 cluster solution is 0.47 (weak structure). However, if we look at the data together with the respondent’s information, we will find that 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 the 3 clusters solution, the cluster with 9 respondents is bad structured, in another word, the consistency within this cluster is bad. This is because the respondents, who consume few chocolate bars and have many missing values in some attributes or whole brands are clustered into this group.

图片包含 文字, 屏幕截图

描述已自动生成

*Figure 9. Silhouette Comparison for two and three cluster solution*

The fit statistics for the 3 clusters solution by kmeans is 72.5 %, which is calculated by within-cluster sum of squares by cluster (between\_SS / total\_SS)

图片包含 文字, 地图

描述已自动生成

*Figure. 10 Kmeans 3 Cluster*

In general, chocolate attracts women more than men. Female have consumed more chocolate bars in different brands than male. The cluster results mainly checked by the average factor means. The four factors are Energy (Calorie, rich, healthful, sweet), the marketing factor (Commercial, image, chocolaty), the taste factor (Creamy, crunchy, addiction, wrapping) and the packaging factor (Accessible, handy).

The majority of the respondents are clustered in the first cluster. This group of people in general are highly influenced by all the four factors. Their chocolate bars consumption is also the most frequent, nearly 90% of the respondents in this group consume every week at least once chocolate bars. The majority of them are do also quite frequent sports. The most favorite chocolate bars brands are Kinderriegel, Snickers and Twix.

The second cluster towards to give relative high scores in the attributes, where the other two clusters tend to give low scores, e.g. healthful and crunchy. Both chocolate consumption and sport frequency are average level by comparing with the other two clusters. The top three most favorite chocolate bars are Kinderriegel, KinderBueno, Duplo.

The third cluster is not sensitive to the commercial promotion, but sensitive to the calories, since they tend to give a high rating scores for overall chocolate bars with respect to the attributes belonging to energy factor. This group of people in general consume not quite often chocolate bars, but most of them are very engaging in sports.

**Part 5 Conclusion**

Half of the respondents consume a chocolate bars when they feel hungry, this is an important reason why give a higher rating scores for the brands that producing chocolate bars with high energy factor. Interesting is, in Germany(at least from this data sample), people like to associate cream with high calories, which enable them to replenish energy quickly, no matter from which group, people have ranked Kinderriegel at a quite top place, however, people do not like a chocolate bar with too sweet taste, e.g. Mars and Lion are considered have both high rating scores in cream and sweet, but respondents have assigned low brands scores to these two. From the factor analysis we know, that the marketing promotion increase influence of the products or brands image. The 2-dimensional metric MDS allow us to compare brands from all attributes, from it we can see which brands are similar to each other. If we increase to a higher dimensional MDS, e.g. 3D or 4D, we might observe different relative relationship among the brands.

By marketing data analysis, interpretation power plays an important role. We should not stick too much to the fitness or significant of statistical results but ignore the actual meaning in the real life.

**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   
##

## $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

1. *The questionnaire and the collected data are from Institute for Marketing of Humboldt University.* [↑](#footnote-ref-1)
2. Euclidean distance see Appendix [↑](#footnote-ref-2)
3. See Appendix [↑](#footnote-ref-3)