Data Analysis on the Questionnaire about the Chocolate Bars

Xin Tan, 583833

# Introduction

这里要重新写

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)*** | | | | | | |
|  | **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?* | | | | |

*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

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

*Table 2. The Number of Missing Values Across Products and Attributes*

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).

# 

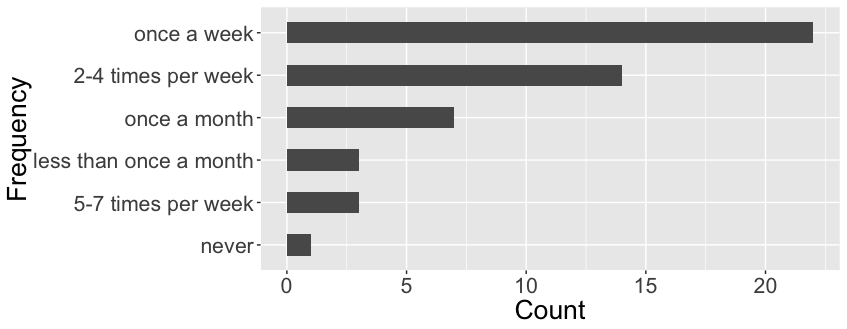
# Respondents

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:

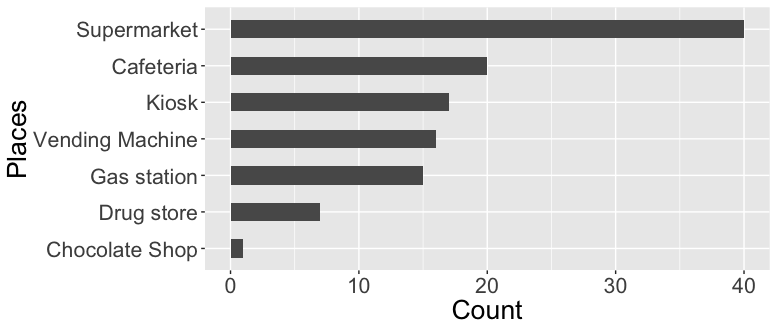
### How often do you consume Chocolate Bars?

**

*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.

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

*Table 3 Feedback Amounts of with Respect to The Brands*

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 results, 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.

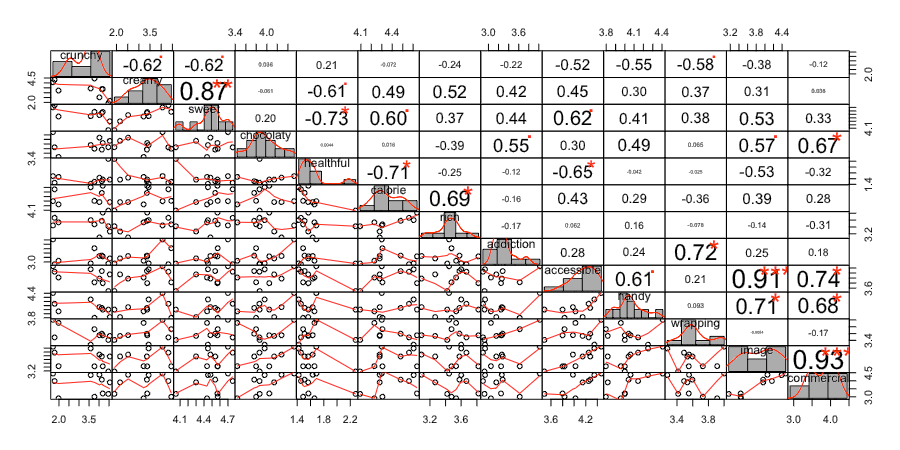
图片包含 文字

描述已自动生成

*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 attribute 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.

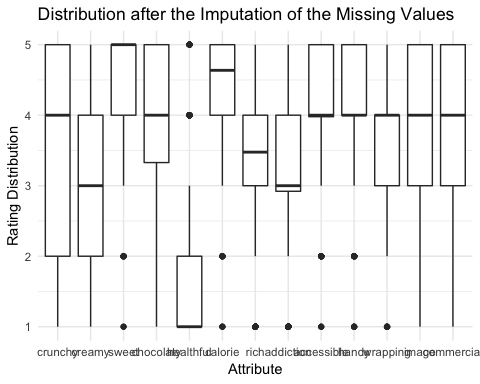
• Healthy is negative correlated with sweet and calorie.

• 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



# 透过打分看消费者哪些Attribute会影响评价。既大家看中的点什么。找出潜在因素。-> 因子分析

# 利用几个重要的影响因素对消费者进行归类，主要为了看怎么样的人，或是什么情况下，消费者会消费什么产品。

* 性别对偏好是否有影响？
* 年龄对偏好
* 学生和非学生会不会存在学生更喜欢吃（看书，休息，补充能量）
* 消费频率对口味和牌子？
* Sport与巧克力的消费

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。

*交叉分析* 是年龄、性别对“xx活动的喜爱”但个变量之间的关系。如果不分类统计，那么喜欢率是10%。 交叉分析后可能会发现不同年龄段，不同性别的喜好是不一样的。