Data Analysis of the Chocolate Comsumption in Germany

Xin Tan, 583833

2019/3/11

# Introduction

This report is structred by firstly demostrate 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 jugement 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 fingdings about the consumers and products would be discussed.

# About the Questionnaire and the Dataset

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

插入表 Table 1. A Brief View of the Questionnaire.

* **Satisfaction of the chocolate brands and attributes**.50 respondents need to evaluate both brands preference and attrinute 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 how often they consumer a chocolate bars.
* **Demographic Questions** consists of several private information about the respondents.

这个数据集的问题：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. 我要干的事情：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 reasonability of the data collection as well as the design of the questionnaire , nor the data quality will not be discussed.

## Data Prepararition

The corresponding data store in the five table sheets:

1. **AttributeRatingsStacked** is the same table as the **Attribute Rating**, which has an inversed layout for research convenience. The datasets reflect consumer ratings of the ten chocolate bars with regard to the given thirteen attributes as expressed on survey items with the following form: • How do you think [Product] on the basis of the following [Attributes]? • Rating score is from 1 to 5. Where 1 is not preferred at all, 5 is greatly preferred.
2. In the **Direct Preference**, respondents ranked their preference only on the ten chocolate bars on survey items with the following form: On a scale from 1 to 7—where 1 is least preferred and 7 is most preferred.
3. **General Consumption** reports the consumption behavior of the respondents, which includes consumption frequency, purchase location, consumption circumstances and consumed brands.
4. **Social Demographic Questions** contains the basic personal information of the respondents.

Consumption behavior investigation (Gerneral Consumption); And the basic information of the respondents (Social Demographic Questions)

Except the **AttributeRatingsStacked** / **Attribute Rating** including missing value, the other three datasets appear to be clean and formatted appropriately.

## Data clean

### Before the imputation of missing data

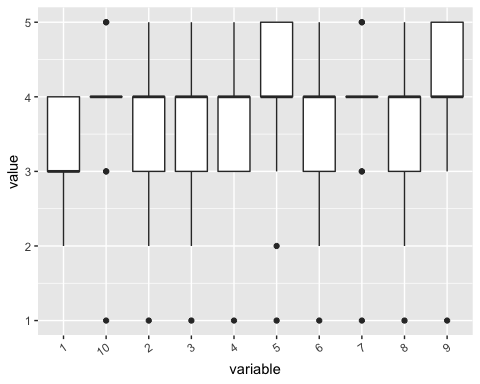
# average rating by person  
AvgRating\_person<- aggregate(products.rating[,3:15],by=list(products.rating$Person),FUN=mean, na.rm=TRUE) # used for cluster  
  
# average rating by products  
AvgRating\_products <- aggregate(products.rating[,3:15],by=list(products.rating$Product),FUN=mean, na.rm=TRUE) # mean  
AvgRating\_products <- aggregate(products.rating[,3:15],by=list(products.rating$Product),FUN=median, na.rm=TRUE) # median  
  
  
  
  
# boxplot by colnums (attribute means)  
library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────── tidyverse 1.2.1 ──

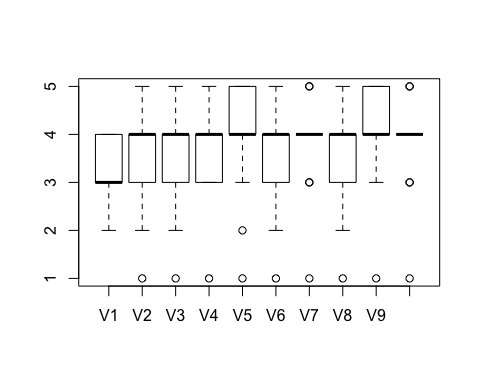
## ✔ ggplot2 3.1.0 ✔ purrr 0.2.5  
## ✔ tibble 2.0.0 ✔ dplyr 0.7.8  
## ✔ tidyr 0.8.2 ✔ stringr 1.3.1  
## ✔ readr 1.3.1 ✔ forcats 0.3.0

## ── Conflicts ───────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

AvgRating\_products%>%  
 #create new column and save the row.names in it  
 mutate(variable = row.names(.)) %>%   
 #convert your data from wide to long   
 tidyr::gather("var", "value", 2:14) %>%  
 #plot it using ggplot2  
 ggplot(., aes(x = variable, y = value)) +  
 geom\_boxplot()+  
 theme(axis.text.x = element\_text(angle=35,hjust=1))



# boxplot by row (products mean)  
tmp <- as.data.frame(t(AvgRating\_products[,2:14]))  
boxplot(tmp)



# imputate missing value by attribute of each brand (130 variables = 10 \* 13)  
  
library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

before.NA<- describe(attribute.rating)

表Table 2. The Distribution of Missing Values Across Products and Attributes

In total there are 6.4% missing inputs in the **AttributeRatingsStacked** / **Attribute Rating**. Table 1 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.

The existing data shows the most popular attribute is sweet with the average rating score 4.5. almost all of the chocolate products got their highest rating on “sweet”. In contrast, the overall chocolate bars got their lowest score at “healthful”, on average is 1.7. The most popular chocolate bar is KinderRiegel (3.9), which even leads in six attributes (wrapping, image, commercial, addiction, handy and accessible). Then followed by Lion and Twix. The least popular chocolate bars are Balisto (3.5) and Bounty (3.5) and KinderBueno (3.6).

### Imputation of Missing Data

The dataset has been cleaned by imputing the missing values with corresponding attribute mean of each brand. Since after 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 slighlt impact on the calorie, rich and addiction).

# 1. NA imputation ----  
dat <- read\_excel("Data\_Chocolate\_allinterviews.xlsx", sheet = "AttributeRatingsStacked")  
dat<-as.data.frame(dat)  
str(dat)

## 'data.frame': 500 obs. of 15 variables:  
## $ Person : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Product : chr "Snickers" "Snickers" "Snickers" "Snickers" ...  
## $ crunchy : num 3 3 NA 4 3 2 4 5 5 3 ...  
## $ creamy : num 3 4 NA 4 3 3 4 5 1 4 ...  
## $ sweet : num 4 5 NA 5 4 4 4 5 4 5 ...  
## $ chocolaty : num 3 3 NA 5 3 3 4 5 5 5 ...  
## $ healthful : num 1 1 NA 1 1 1 2 1 1 1 ...  
## $ calorie : num 4 5 NA 4 4 5 4 5 5 5 ...  
## $ rich : num 4 5 NA 4 3 4 4 2 NA 5 ...  
## $ addiction : num 2 1 NA 3 2 3 3 1 NA 4 ...  
## $ accessible: num 4 5 NA 4 4 4 4 5 5 4 ...  
## $ handy : num 4 5 NA 4 4 5 NA 3 5 3 ...  
## $ wrapping : num 4 4 NA 4 2 3 3 5 5 3 ...  
## $ image : num 4 5 NA 4 4 5 4 5 5 5 ...  
## $ commercial: num 5 5 NA 5 4 5 4 NA NA 5 ...

bars<-dat  
summary(dat)

## Person Product crunchy creamy   
## Min. : 1.0 Length:500 Min. :1.000 Min. :1.000   
## 1st Qu.:13.0 Class :character 1st Qu.:2.000 1st Qu.:2.000   
## Median :25.5 Mode :character Median :4.000 Median :3.000   
## Mean :25.5 Mean :3.397 Mean :3.296   
## 3rd Qu.:38.0 3rd Qu.:5.000 3rd Qu.:4.000   
## Max. :50.0 Max. :5.000 Max. :5.000   
## NA's :21 NA's :20   
## sweet chocolaty healthful calorie   
## Min. :1.00 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:4.00 1st Qu.:3.000 1st Qu.:1.000 1st Qu.:4.000   
## Median :5.00 Median :4.000 Median :1.000 Median :5.000   
## Mean :4.49 Mean :3.969 Mean :1.624 Mean :4.378   
## 3rd Qu.:5.00 3rd Qu.:5.000 3rd Qu.:2.000 3rd Qu.:5.000   
## Max. :5.00 Max. :5.000 Max. :5.000 Max. :5.000   
## NA's :16 NA's :18 NA's :40 NA's :18   
## rich addiction accessible handy   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:2.000 1st Qu.:4.000 1st Qu.:4.000   
## Median :4.000 Median :3.000 Median :4.000 Median :4.000   
## Mean :3.488 Mean :3.246 Mean :4.075 Mean :4.049   
## 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:5.000 3rd Qu.:5.000   
## Max. :5.000 Max. :5.000 Max. :5.000 Max. :5.000   
## NA's :70 NA's :65 NA's :46 NA's :26   
## wrapping image commercial   
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000   
## Median :4.000 Median :4.000 Median :4.000   
## Mean :3.607 Mean :3.868 Mean :3.767   
## 3rd Qu.:4.000 3rd Qu.:5.000 3rd Qu.:5.000   
## Max. :5.000 Max. :5.000 Max. :5.000   
## NA's :44 NA's :21 NA's :49

aggregate(dat[,-c(1,2)], by=list(bars$Product),mean, na.rm=TRUE)

## Group.1 crunchy creamy sweet chocolaty healthful calorie  
## 1 Balisto 4.465116 1.906977 4.045455 3.863636 2.279070 4.093023  
## 2 Bounty 1.978723 3.212766 4.437500 3.437500 1.659574 4.229167  
## 3 Duplo 4.020000 2.520000 4.480000 4.280000 1.632653 4.280000  
## 4 KinderBueno 3.770833 4.448980 4.571429 3.795918 1.488889 4.306122  
## 5 KinderRiegel 1.700000 4.060000 4.700000 4.520000 1.687500 4.300000  
## 6 KitKat 4.540000 2.160000 4.300000 4.140000 1.565217 4.280000  
## 7 Lion 4.186047 3.372093 4.568182 4.046512 1.568182 4.636364  
## 8 Mars 1.780000 4.320000 4.760000 3.860000 1.456522 4.560000  
## 9 Snickers 3.583333 3.437500 4.448980 3.958333 1.422222 4.604167  
## 10 Twix 4.140000 3.360000 4.540000 3.760000 1.510638 4.480000  
## rich addiction accessible handy wrapping image commercial  
## 1 3.421053 2.921053 3.487179 4.000000 3.414634 3.162791 3.179487  
## 2 3.534884 3.048780 4.022727 3.957447 3.911111 3.446809 2.904762  
## 3 3.093023 3.266667 3.978723 4.061224 3.553191 4.060000 4.367347  
## 4 3.476190 3.622222 3.833333 3.795918 3.729167 3.440000 3.000000  
## 5 3.444444 4.022222 4.468085 4.408163 4.000000 4.480000 4.446809  
## 6 3.266667 3.155556 4.145833 3.816327 3.531915 3.960000 3.895833  
## 7 3.820513 3.250000 3.800000 4.000000 3.550000 3.439024 3.205128  
## 8 3.666667 2.891304 4.250000 4.142857 3.489362 4.163265 3.978261  
## 9 3.600000 2.931818 4.347826 4.250000 3.276596 4.387755 4.347826  
## 10 3.577778 3.282609 4.276596 4.040816 3.595745 3.960000 4.063830

library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

# use data.table  
bars.dt = as.data.table(bars)  
# melt data.table to "long" format  
bars.dt.long = melt(bars.dt, id.vars = c("Person", "Product"),   
 variable.name = "Attribute", value.name = "Value")   
  
str(bars.dt.long)

## Classes 'data.table' and 'data.frame': 6500 obs. of 4 variables:  
## $ Person : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Product : chr "Snickers" "Snickers" "Snickers" "Snickers" ...  
## $ Attribute: Factor w/ 13 levels "crunchy","creamy",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Value : num 3 3 NA 4 3 2 4 5 5 3 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

head(bars.dt.long)

## Person Product Attribute Value  
## 1: 1 Snickers crunchy 3  
## 2: 2 Snickers crunchy 3  
## 3: 3 Snickers crunchy NA  
## 4: 4 Snickers crunchy 4  
## 5: 5 Snickers crunchy 3  
## 6: 6 Snickers crunchy 2

# mean values for each attribute (by each product and attribute)----  
# 加了一列：每个product对应attribute的平均分.为之后一步做准备  
bars.dt.long[, Mean.by.prod.att := mean(Value, na.rm = TRUE), by = .(Product, Attribute)]  
head(bars.dt.long)

## Person Product Attribute Value Mean.by.prod.att  
## 1: 1 Snickers crunchy 3 3.583333  
## 2: 2 Snickers crunchy 3 3.583333  
## 3: 3 Snickers crunchy NA 3.583333  
## 4: 4 Snickers crunchy 4 3.583333  
## 5: 5 Snickers crunchy 3 3.583333  
## 6: 6 Snickers crunchy 2 3.583333

# test  
# 每个product对应attribute的平均分[130\*13]  
unique(bars.dt.long[, .(Product, Attribute, Mean.by.prod.att)])

## Product Attribute Mean.by.prod.att  
## 1: Snickers crunchy 3.583333  
## 2: KinderBueno crunchy 3.770833  
## 3: Twix crunchy 4.140000  
## 4: Mars crunchy 1.780000  
## 5: KitKat crunchy 4.540000  
## ---   
## 126: Bounty commercial 2.904762  
## 127: KinderRiegel commercial 4.446809  
## 128: Balisto commercial 3.179487  
## 129: Lion commercial 3.205128  
## 130: Duplo commercial 4.367347

# impute NA----  
bars.dt.long[is.na(Value), Value := Mean.by.prod.att]  
bars.dt.2 = dcast(bars.dt.long, Person + Product ~ Attribute, value.var = "Value")  
  
head(bars.dt.2)

## Person Product crunchy creamy sweet chocolaty healthful calorie  
## 1: 1 Balisto 5 2 3 3 3 4  
## 2: 1 Bounty 2 4 5 3 2 4  
## 3: 1 Duplo 4 3 4 4 2 4  
## 4: 1 KinderBueno 4 4 5 4 1 5  
## 5: 1 KinderRiegel 3 3 4 5 3 4  
## 6: 1 KitKat 4 2 3 4 2 4  
## rich addiction accessible handy wrapping image commercial  
## 1: 4 3 3 3 2 3 3  
## 2: 3 3 3 4 4 3 3  
## 3: 4 4 4 4 4 3 3  
## 4: 4 3 3 2 3 4 1  
## 5: 4 5 4 4 4 4 3  
## 6: 3 3 4 4 4 4 4

summary(bars.dt.2)

## Person Product crunchy creamy   
## Min. : 1.0 Length:500 Min. :1.000 Min. :1.00   
## 1st Qu.:13.0 Class :character 1st Qu.:2.000 1st Qu.:2.00   
## Median :25.5 Mode :character Median :4.000 Median :3.00   
## Mean :25.5 Mean :3.416 Mean :3.28   
## 3rd Qu.:38.0 3rd Qu.:5.000 3rd Qu.:4.00   
## Max. :50.0 Max. :5.000 Max. :5.00   
## sweet chocolaty healthful calorie   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:4.000 1st Qu.:3.328 1st Qu.:1.000 1st Qu.:4.000   
## Median :5.000 Median :4.000 Median :1.000 Median :4.636   
## Mean :4.485 Mean :3.966 Mean :1.627 Mean :4.377   
## 3rd Qu.:5.000 3rd Qu.:5.000 3rd Qu.:2.000 3rd Qu.:5.000   
## Max. :5.000 Max. :5.000 Max. :5.000 Max. :5.000   
## rich addiction accessible handy   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:2.921 1st Qu.:3.979 1st Qu.:4.000   
## Median :3.476 Median :3.000 Median :4.000 Median :4.000   
## Mean :3.490 Mean :3.239 Mean :4.061 Mean :4.047   
## 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:5.000 3rd Qu.:5.000   
## Max. :5.000 Max. :5.000 Max. :5.000 Max. :5.000   
## wrapping image commercial   
## Min. :1.000 Min. :1.00 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:3.00 1st Qu.:3.000   
## Median :4.000 Median :4.00 Median :4.000   
## Mean :3.605 Mean :3.85 Mean :3.739   
## 3rd Qu.:4.000 3rd Qu.:5.00 3rd Qu.:5.000   
## Max. :5.000 Max. :5.00 Max. :5.000

attribute <- bars.dt.2  
## attribute is the dataset to be used to do 'Reducing Data Complexity'. Cleaned!  
  
#### End of the NA imputation

# Respondents

## Basic Information

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 get 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 majorityn 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 exciese more often than male (29:21).

## Consumption Behaviour

**General Consumption** reports the consumption behavior of the respondents, which includes consumption frequency, purchase location, consumption circumstances and consumed brands:

#How often do you consume Chocolate Bars?  
table(consumption$`2. How often do you consume Chocolate Bars?`)/50

##   
## 2-4 times per week 5-7 times per week less than once a month   
## 0.28 0.06 0.06   
## never once a month once a week   
## 0.02 0.14 0.44

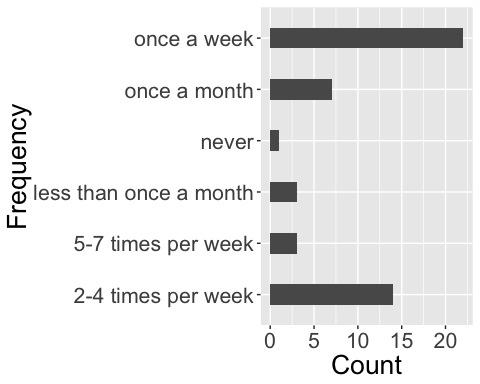
# Plot  
apply(consumption, 2, FUN = table)

## $Person  
##   
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   
## 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   
##   
## $`2. How often do you consume Chocolate Bars?`  
##   
## 2-4 times per week 5-7 times per week less than once a month   
## 14 3 3   
## never once a month once a week   
## 1 7 22   
##   
## $`4. Where can you find yourself buying Chocolate Bars?`  
##   
## cafeteria   
## 3   
## cafeteria, kiosk   
## 1   
## drug store, kiosk   
## 1   
## gas station   
## 1   
## supermarket   
## 12   
## supermarket, cafeteria   
## 3   
## supermarket, cafeteria, gas station   
## 2   
## supermarket, cafeteria, gas station, kiosk   
## 1   
## supermarket, cafeteria, kiosk   
## 2   
## supermarket, drug store   
## 1   
## supermarket, gas station   
## 2   
## supermarket, gas station, kiosk   
## 1   
## supermarket, kiosk   
## 3   
## supermarket, Rittersport Shop   
## 1   
## vending machine   
## 1   
## vending machine, cafeteria, gas station   
## 1   
## vending machine, cafeteria, kiosk   
## 1   
## vending machine, drug store   
## 1   
## vending machine, supermarket   
## 2   
## vending machine, supermarket, cafeteria   
## 1   
## vending machine, supermarket, cafeteria, gas station, kiosk   
## 1   
## vending machine, supermarket, cafeteria, kiosk   
## 2   
## vending machine, supermarket, drug store, cafeteria, gas station, kiosk   
## 2   
## vending machine, supermarket, drug store, gas station, kiosk   
## 2   
## vending machine, supermarket, gas station   
## 2   
##   
## $`5. Under which circumstances do you consume Chocolate Bars?`  
##   
## as a dessert   
## 1   
## as a dessert, as a treat, when I am watching TV   
## 1   
## as a dessert, as a treat, while travelling   
## 1   
## as a dessert, when I am in a bad mood, when I am watching TV, while travelling   
## 1   
## as a dessert, when I am under pressure (e.g. while preparing for an exam)   
## 1   
## as a dessert, when I am under pressure (e.g. while preparing for an exam), when I am bored, when I am in a bad mood, while travelling   
## 1   
## as a dessert, when I am under pressure (e.g. while preparing for an exam), when I am in a bad mood, when I am watching TV, while travelling   
## 1   
## as a dessert, when I am under pressure (e.g. while preparing for an exam), when I am watching TV, while travelling   
## 1   
## as a treat   
## 3   
## as a treat, when I am bored, and when i feel like it, no particular circumstance given   
## 1   
## when I am hungry   
## 1   
## when I am hungry, as a dessert, as a treat, when I am bored, when I am in a bad mood   
## 1   
## when I am hungry, as a dessert, as a treat, when I am in a bad mood, while travelling   
## 1   
## when I am hungry, as a dessert, when I am under pressure (e.g. while preparing for an exam)   
## 1   
## when I am hungry, as a dessert, when I am under pressure (e.g. while preparing for an exam), as a treat   
## 1   
## when I am hungry, as a dessert, when I am under pressure (e.g. while preparing for an exam), as a treat, when I am bored, when I am watching TV   
## 1   
## when I am hungry, as a dessert, when I am under pressure (e.g. while preparing for an exam), as a treat, when I am in a bad mood, when I am watching TV, while travelling   
## 1   
## when I am hungry, as a dessert, when I am under pressure (e.g. while preparing for an exam), when I am bored, when I am in a bad mood, when I am watching TV   
## 1   
## when I am hungry, as a dessert, when I am under pressure (e.g. while preparing for an exam), when I am bored, when I am in a bad mood, when I am watching TV, while travelling   
## 1   
## when I am hungry, as a dessert, when I am under pressure (e.g. while preparing for an exam), when I am bored, when I am watching TV   
## 1   
## when I am hungry, as a dessert, when I am under pressure (e.g. while preparing for an exam), when I am watching TV   
## 1   
## when I am hungry, as a treat   
## 1   
## when I am hungry, as a treat, while travelling   
## 2   
## when I am hungry, when I am bored, when I am watching TV   
## 2   
## when I am hungry, when I am bored, when I am watching TV, while travelling   
## 1   
## when I am hungry, when I am in a bad mood, while travelling   
## 1   
## when I am hungry, when I am under pressure (e.g. while preparing for an exam), when I am bored, when I am in a bad mood, when I am watching TV   
## 1   
## when I am hungry, when I am under pressure (e.g. while preparing for an exam), when I am in a bad mood   
## 2   
## when I am hungry, while travelling   
## 1   
## when I am under pressure (e.g. while preparing for an exam), as a treat   
## 1   
## when I am under pressure (e.g. while preparing for an exam), as a treat, when I am bored, when I am in a bad mood, when I am watching TV   
## 1   
## when I am under pressure (e.g. while preparing for an exam), as a treat, when I am bored, when I am watching TV   
## 1   
## when I am under pressure (e.g. while preparing for an exam), as a treat, when I am watching TV   
## 1   
## when I am under pressure (e.g. while preparing for an exam), when I am watching TV, while travelling   
## 2   
## when I feel like it   
## 1   
## while driving my car, as a dessert, while travelling   
## 1   
## while driving my car, as a treat, while travelling   
## 2   
## while driving my car, when I am hungry, as a dessert, when I am under pressure (e.g. while preparing for an exam), as a treat, when I am bored, when I am in a bad mood, when I am watching TV, while travelling   
## 1   
## while driving my car, when I am hungry, when I am bored, when I am watching TV   
## 1   
## while driving my car, when I am hungry, when I am in a bad mood   
## 1   
## while driving my car, when I am under pressure (e.g. while preparing for an exam), while travelling   
## 1   
## while travelling   
## 2   
##   
## $`7. Which of the following Chocolate Bars have you ever consumed?`  
##   
## Kinder Bueno, Twix, Mars, KitKat, Bounty, Kinderriegel, Lion, Duplo   
## 1   
## Snickers, Kinder Bueno, Twix, KitKat, Duplo   
## 1   
## Snickers, Kinder Bueno, Twix, Mars, KitKat, Bounty, Kinderriegel, Balisto Korn-Mix, Duplo   
## 4   
## Snickers, Kinder Bueno, Twix, Mars, KitKat, Bounty, Kinderriegel, Balisto Korn-Mix, Lion, Duplo   
## 34   
## Snickers, Kinder Bueno, Twix, Mars, KitKat, Bounty, Kinderriegel, Duplo   
## 2   
## Snickers, Kinder Bueno, Twix, Mars, KitKat, Bounty, Kinderriegel, Lion, Duplo   
## 3   
## Snickers, Kinder Bueno, Twix, Mars, KitKat, Kinderriegel, Balisto Korn-Mix, Duplo   
## 1   
## Snickers, Kinder Bueno, Twix, Mars, KitKat, Kinderriegel, Balisto Korn-Mix, Lion, Duplo   
## 1   
## Snickers, Kinder Bueno, Twix, Mars, KitKat, Kinderriegel, Duplo   
## 1   
## Snickers, Lion, Duplo   
## 1   
## Snickers, Twix, Mars, KitKat, Bounty, Kinderriegel, Lion, Duplo   
## 1

tmp<-as.data.frame(table(consumption$`2. How often do you consume Chocolate Bars?`))  
library(ggplot2)  
library(viridis)

## Loading required package: viridisLite

ggplot(tmp, aes(x=Var1,y=Freq)) + geom\_bar(stat="identity", width=.4) +   
 coord\_flip() + labs(y="Count", x="Frequency")+  
 theme(text = element\_text(size=20))



# Under which circumstances do you consume Chocolate Bars?  
  
library(stringr)  
str\_count(consumption, "driving")

## Warning in stri\_count\_regex(string, pattern, opts\_regex = opts(pattern)):  
## argument is not an atomic vector; coercing

## [1] 0 0 0 7 0

### 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.

### Where can you find yourself buying 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 do 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? (Brandcount)

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.

1. Most Popular. Most completed evaluation from respondents are is Duplo(50).
2. Very Popular: Snikers, Twix, KitKat, KinderRiegel and Mars.
3. Well known: Bounty, Lion and Balisto.
4. Least Popular: KinderBueno.

# Satisfaction of the chocolate brands and attributes

### 比较受欢迎的牌子、口味。每个牌子比较擅长或者受认可的attribute.(the products’ mean)

* **The most favored chocolate bars’ brands**

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.Mars (3.9)

# brands mean  
  
apply(preference.rating[,2:11], 2, mean)

## Snickers (b1) KinderBueno (b2) Twix (b3)   
## 5.54 5.40 5.18   
## Mars (b4) KitKat (b5) Bounty (b6)   
## 3.90 5.04 3.96   
## Kinderriegel (b7) BalistoKornMix(b8) Lion (b9)   
## 5.86 4.32 4.50   
## Duplo (bb)   
## 4.62

# calculate the products' mean----  
  
## What is the average (mean) position of the chocolat bars on each attribute.sc?  
products.mean <- aggregate(.~Product, data = attribute, mean)  
  
rownames(products.mean ) <- products.mean [, 1] # use brand for the row names  
products.mean <- products.mean [, -c(1,2)] # remove brand name column  
products.mean

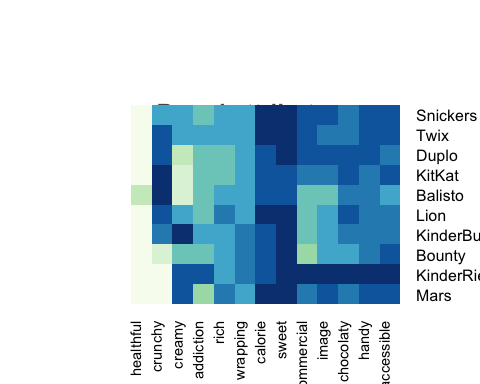
## crunchy creamy sweet chocolaty healthful calorie  
## Balisto 4.465116 1.906977 4.045455 3.863636 2.279070 4.093023  
## Bounty 1.978723 3.212766 4.437500 3.437500 1.659574 4.229167  
## Duplo 4.020000 2.520000 4.480000 4.280000 1.632653 4.280000  
## KinderBueno 3.770833 4.448980 4.571429 3.795918 1.488889 4.306122  
## KinderRiegel 1.700000 4.060000 4.700000 4.520000 1.687500 4.300000  
## KitKat 4.540000 2.160000 4.300000 4.140000 1.565217 4.280000  
## Lion 4.186047 3.372093 4.568182 4.046512 1.568182 4.636364  
## Mars 1.780000 4.320000 4.760000 3.860000 1.456522 4.560000  
## Snickers 3.583333 3.437500 4.448980 3.958333 1.422222 4.604167  
## Twix 4.140000 3.360000 4.540000 3.760000 1.510638 4.480000  
## rich addiction accessible handy wrapping image  
## Balisto 3.421053 2.921053 3.487179 4.000000 3.414634 3.162791  
## Bounty 3.534884 3.048780 4.022727 3.957447 3.911111 3.446809  
## Duplo 3.093023 3.266667 3.978723 4.061224 3.553191 4.060000  
## KinderBueno 3.476190 3.622222 3.833333 3.795918 3.729167 3.440000  
## KinderRiegel 3.444444 4.022222 4.468085 4.408163 4.000000 4.480000  
## KitKat 3.266667 3.155556 4.145833 3.816327 3.531915 3.960000  
## Lion 3.820513 3.250000 3.800000 4.000000 3.550000 3.439024  
## Mars 3.666667 2.891304 4.250000 4.142857 3.489362 4.163265  
## Snickers 3.600000 2.931818 4.347826 4.250000 3.276596 4.387755  
## Twix 3.577778 3.282609 4.276596 4.040816 3.595745 3.960000  
## commercial  
## Balisto 3.179487  
## Bounty 2.904762  
## Duplo 4.367347  
## KinderBueno 3.000000  
## KinderRiegel 4.446809  
## KitKat 3.895833  
## Lion 3.205128  
## Mars 3.978261  
## Snickers 4.347826  
## Twix 4.063830

# TEST/ 图 Cluster/ Heatmap of the mean score of each attribute.sc by product----   
library(gplots)

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(RColorBrewer)  
heatmap.2(as.matrix(products.mean),  
 col=brewer.pal(9, "GnBu"), trace="none", key=FALSE, dend="none",  
 main="\n\n\n\n\nBrand attribute.scs")



#write.csv(products.mean, file = "Products mean.csv")  
  
## visulized table shows which chocolate bar is specialied at which attribute  
  
########End of Cluster/ Heatmap

# boxplot by colnums (attribute)  
library(reshape2)

##   
## Attaching package: 'reshape2'

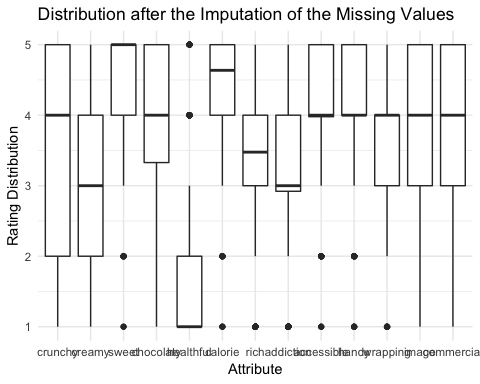
## The following objects are masked from 'package:data.table':  
##   
## dcast, melt

## The following object is masked from 'package:tidyr':  
##   
## smiths

library(ggplot2)  
x <- attribute[,c(2:15)]  
x <- as.data.frame(x)  
x <- melt(x)

## Using Product as id variables

plt <- ggplot(data =x , aes(x = variable, y = value))  
plt + geom\_boxplot() +theme\_minimal() + labs(x = "Attribute", y = "Rating Distribution") + ggtitle("Distribution after the Imputation of the Missing Values")



If we looking 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 healthy.

Table. Attribute Mean Scores

* **Brands with the most complete products lines** Here, the Brands with highest average mean scores are : KinderRiegel(3.9),Snickers(3.7), Twix(3.7) and Duplo(3.7). These brands have relative complete products line which offer chocolate bars with different oriented attributes or combination of different attributes.
* **People’s perception of chocolate bars** The Attributes with highest average mean scores are: sweet(4.5), calorie(4.4) and accessible(4.1). Considering the brands of various chocolate bars, a high score indicates when people think about chocolate bars, these are the most common attributes one should have.

The heatmap visulize which brands is good at making chocolate bars with what kind of attributes? (the darker of the color, the higher of the average rating score) e.g. Balisto is the only chaoclate bar earns a relative higher score in healthful attribute.

### Correlation of the attribute

Is there any relationship among attributes?

library("PerformanceAnalytics")

## Loading required package: xts

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Attaching package: 'xts'

## The following objects are masked from 'package:data.table':  
##   
## first, last

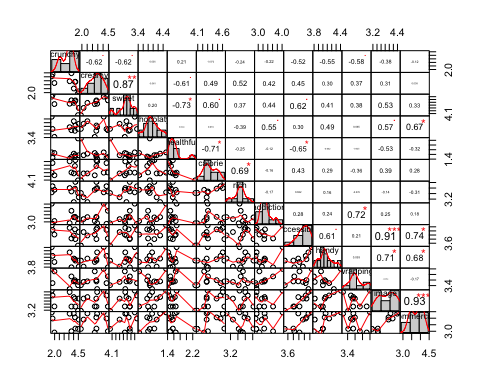
## The following objects are masked from 'package:dplyr':  
##   
## first, last

##   
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:gplots':  
##   
## textplot

## The following object is masked from 'package:graphics':  
##   
## legend

my\_data <- products.mean[, 1:13]  
chart.Correlation(my\_data, histogram=TRUE, pch=19)



In the above plot:

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) <=> symbols(“***”, “****”, “*”, “.”, " “)

Some findings in the correlation matrix: \* Both commercial and accessible are significantly positive corralted 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? 打分倾向可以放这里说？

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

# Mean Score for male and female  
# 3.0 Data Preparation----  
library(readxl)  
library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

setwd("~/Desktop/CACI/Special Work Performances/SWP 1+2/SWP2")  
  
dat <- read\_excel("Data\_Chocolate\_allinterviews.xlsx", sheet = "AttributeRatingsStacked")  
dat<-as.data.frame(dat)  
bars<-dat  
str(dat)

## 'data.frame': 500 obs. of 15 variables:  
## $ Person : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Product : chr "Snickers" "Snickers" "Snickers" "Snickers" ...  
## $ crunchy : num 3 3 NA 4 3 2 4 5 5 3 ...  
## $ creamy : num 3 4 NA 4 3 3 4 5 1 4 ...  
## $ sweet : num 4 5 NA 5 4 4 4 5 4 5 ...  
## $ chocolaty : num 3 3 NA 5 3 3 4 5 5 5 ...  
## $ healthful : num 1 1 NA 1 1 1 2 1 1 1 ...  
## $ calorie : num 4 5 NA 4 4 5 4 5 5 5 ...  
## $ rich : num 4 5 NA 4 3 4 4 2 NA 5 ...  
## $ addiction : num 2 1 NA 3 2 3 3 1 NA 4 ...  
## $ accessible: num 4 5 NA 4 4 4 4 5 5 4 ...  
## $ handy : num 4 5 NA 4 4 5 NA 3 5 3 ...  
## $ wrapping : num 4 4 NA 4 2 3 3 5 5 3 ...  
## $ image : num 4 5 NA 4 4 5 4 5 5 5 ...  
## $ commercial: num 5 5 NA 5 4 5 4 NA NA 5 ...

library(data.table)  
# use data.table  
bars.dt = as.data.table(bars)  
# melt data.table to "long" format  
bars.dt.long = melt(bars.dt, id.vars = c("Person", "Product"),   
 variable.name = "Attribute", value.name = "Value")  
str(bars.dt.long)

## Classes 'data.table' and 'data.frame': 6500 obs. of 4 variables:  
## $ Person : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Product : chr "Snickers" "Snickers" "Snickers" "Snickers" ...  
## $ Attribute: Factor w/ 13 levels "crunchy","creamy",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Value : num 3 3 NA 4 3 2 4 5 5 3 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

# mean values for each attribute (by each product and attribute)  
bars.dt.long[, Mean.by.prod.att := mean(Value, na.rm = TRUE), by = .(Product, Attribute)]  
head(bars.dt.long)

## Person Product Attribute Value Mean.by.prod.att  
## 1: 1 Snickers crunchy 3 3.583333  
## 2: 2 Snickers crunchy 3 3.583333  
## 3: 3 Snickers crunchy NA 3.583333  
## 4: 4 Snickers crunchy 4 3.583333  
## 5: 5 Snickers crunchy 3 3.583333  
## 6: 6 Snickers crunchy 2 3.583333

# test  
unique(bars.dt.long[, .(Product, Attribute, Mean.by.prod.att)])

## Product Attribute Mean.by.prod.att  
## 1: Snickers crunchy 3.583333  
## 2: KinderBueno crunchy 3.770833  
## 3: Twix crunchy 4.140000  
## 4: Mars crunchy 1.780000  
## 5: KitKat crunchy 4.540000  
## ---   
## 126: Bounty commercial 2.904762  
## 127: KinderRiegel commercial 4.446809  
## 128: Balisto commercial 3.179487  
## 129: Lion commercial 3.205128  
## 130: Duplo commercial 4.367347

# impute  
bars.dt.long[is.na(Value), Value := Mean.by.prod.att]  
  
  
  
  
bars.dt.2 = dcast(bars.dt.long, Person + Product ~ Attribute, value.var = "Value")  
  
str(bars.dt.2)

## 'data.frame': 500 obs. of 15 variables:  
## $ Person : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Product : chr "Balisto" "Bounty" "Duplo" "KinderBueno" ...  
## $ crunchy : num 5 2 4 4 3 4 3 2 3 4 ...  
## $ creamy : num 2 4 3 4 3 2 4 4 3 4 ...  
## $ sweet : num 3 5 4 5 4 3 4 4 4 4 ...  
## $ chocolaty : num 3 3 4 4 5 4 4 4 3 4 ...  
## $ healthful : num 3 2 2 1 3 2 2 2 1 1 ...  
## $ calorie : num 4 4 4 5 4 4 4 4 4 4 ...  
## $ rich : num 4 3 4 4 4 3 4 4 4 4 ...  
## $ addiction : num 3 3 4 3 5 3 4 4 2 4 ...  
## $ accessible: num 3 3 4 3 4 4 4 4 4 4 ...  
## $ handy : num 3 4 4 2 4 4 4 4 4 4 ...  
## $ wrapping : num 2 4 4 3 4 4 4 4 4 4 ...  
## $ image : num 3 3 3 4 4 4 4 3 4 4 ...  
## $ commercial: num 3 3 3 1 3 4 3 3 5 4 ...

bars<-bars.dt.2  
  
str(bars)

## 'data.frame': 500 obs. of 15 variables:  
## $ Person : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Product : chr "Balisto" "Bounty" "Duplo" "KinderBueno" ...  
## $ crunchy : num 5 2 4 4 3 4 3 2 3 4 ...  
## $ creamy : num 2 4 3 4 3 2 4 4 3 4 ...  
## $ sweet : num 3 5 4 5 4 3 4 4 4 4 ...  
## $ chocolaty : num 3 3 4 4 5 4 4 4 3 4 ...  
## $ healthful : num 3 2 2 1 3 2 2 2 1 1 ...  
## $ calorie : num 4 4 4 5 4 4 4 4 4 4 ...  
## $ rich : num 4 3 4 4 4 3 4 4 4 4 ...  
## $ addiction : num 3 3 4 3 5 3 4 4 2 4 ...  
## $ accessible: num 3 3 4 3 4 4 4 4 4 4 ...  
## $ handy : num 3 4 4 2 4 4 4 4 4 4 ...  
## $ wrapping : num 2 4 4 3 4 4 4 4 4 4 ...  
## $ image : num 3 3 3 4 4 4 4 3 4 4 ...  
## $ commercial: num 3 3 3 1 3 4 3 3 5 4 ...

socio<-read\_excel("Data\_Chocolate\_allinterviews.xlsx", sheet = "Social Demographic Questions")  
socio<-as.data.frame(socio)  
  
str(socio)

## 'data.frame': 50 obs. of 9 variables:  
## $ Person : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ 28. What is your gender? : chr "male" "female" "female" "female" ...  
## $ 29. How old are you? : num 27 25 25 24 21 18 29 20 29 27 ...  
## $ 30. What is your main occupation? : chr "I am an employee." "I am an employee." "I am a student." "I am a student." ...  
## $ 31. What is your marital status? : chr "single" "single" "es ist kompliziert" "single" ...  
## $ 32. Do you have children? If yes, how many? : num NA 0 NA NA 0 NA 0 NA 0 0 ...  
## $ 33. Where do you live? : chr "City" "City" "City" "Village" ...  
## $ 34. In which state of Germany do you live? : chr "Sachsen" "Berlin" "Sachsen-Anhalt" "Sachsen-Anhalt" ...  
## $ 35. How often do you practice any kind of sport?: chr "1-3 times a week" "1-3 times a week" "4-7 times a week" "4-7 times a week" ...

library(plyr)

## -------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## -------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following object is masked from 'package:purrr':  
##   
## compact

socio<-rename(socio,c("28. What is your gender?"="Gender"))  
  
  
bars<-merge(bars,socio,by="Person")  
str(bars)

## 'data.frame': 500 obs. of 23 variables:  
## $ Person : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Product : chr "Balisto" "Bounty" "Duplo" "KinderBueno" ...  
## $ crunchy : num 5 2 4 4 3 4 3 2 3 4 ...  
## $ creamy : num 2 4 3 4 3 2 4 4 3 4 ...  
## $ sweet : num 3 5 4 5 4 3 4 4 4 4 ...  
## $ chocolaty : num 3 3 4 4 5 4 4 4 3 4 ...  
## $ healthful : num 3 2 2 1 3 2 2 2 1 1 ...  
## $ calorie : num 4 4 4 5 4 4 4 4 4 4 ...  
## $ rich : num 4 3 4 4 4 3 4 4 4 4 ...  
## $ addiction : num 3 3 4 3 5 3 4 4 2 4 ...  
## $ accessible : num 3 3 4 3 4 4 4 4 4 4 ...  
## $ handy : num 3 4 4 2 4 4 4 4 4 4 ...  
## $ wrapping : num 2 4 4 3 4 4 4 4 4 4 ...  
## $ image : num 3 3 3 4 4 4 4 3 4 4 ...  
## $ commercial : num 3 3 3 1 3 4 3 3 5 4 ...  
## $ Gender : chr "male" "male" "male" "male" ...  
## $ 29. How old are you? : num 27 27 27 27 27 27 27 27 27 27 ...  
## $ 30. What is your main occupation? : chr "I am an employee." "I am an employee." "I am an employee." "I am an employee." ...  
## $ 31. What is your marital status? : chr "single" "single" "single" "single" ...  
## $ 32. Do you have children? If yes, how many? : num NA NA NA NA NA NA NA NA NA NA ...  
## $ 33. Where do you live? : chr "City" "City" "City" "City" ...  
## $ 34. In which state of Germany do you live? : chr "Sachsen" "Sachsen" "Sachsen" "Sachsen" ...  
## $ 35. How often do you practice any kind of sport?: chr "1-3 times a week" "1-3 times a week" "1-3 times a week" "1-3 times a week" ...

bars.gender<-aggregate(bars[,-c(1,2,16:23)], by=list(bars$Product,bars$Gender),mean, na.rm=TRUE)  
str(bars.gender)

## 'data.frame': 20 obs. of 15 variables:  
## $ Group.1 : chr "Balisto" "Bounty" "Duplo" "KinderBueno" ...  
## $ Group.2 : chr "female" "female" "female" "female" ...  
## $ crunchy : num 4.43 1.79 4.14 3.93 1.69 ...  
## $ creamy : num 1.86 3.39 2.31 4.45 3.97 ...  
## $ sweet : num 4.11 4.59 4.55 4.66 4.66 ...  
## $ chocolaty : num 3.86 3.31 4.31 3.62 4.52 ...  
## $ healthful : num 2.36 1.64 1.64 1.58 1.53 ...  
## $ calorie : num 4 4.34 4.38 4.28 4.34 ...  
## $ rich : num 3.37 3.47 2.94 3.36 3.43 ...  
## $ addiction : num 2.81 2.97 3.2 3.77 4.07 ...  
## $ accessible: num 3.58 4.21 4.14 3.99 4.6 ...  
## $ handy : num 4.07 3.97 4.17 4.03 4.55 ...  
## $ wrapping : num 3.69 4.09 3.64 3.91 4.24 ...  
## $ image : num 3.18 3.43 4.07 3.28 4.59 ...  
## $ commercial: num 3.34 2.85 4.48 2.83 4.6 ...

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。