Beautiful Cosmetics

Mattia Brocco, Cecilia Giunta, Francesca Michielan, Giulio Piccolo

April 2020

INTRODUCTION

Beautiful is a company selling cosmetics, mainly in the UK. The company have been providing fragrances, skin care, and makeup since the nineties and has now decided to enter in the green marketing and widen their offer through a line of natural products. The main issues marketing managers want to understand, before proceeding with a particular strategy regard:

- The differences in attitudes and characteristics between consumers/non-consumers of natural products (today non-consumers can be tomorrow customers?)
- the factors impacting on the willingness to buy, and on the purchase habits
- the products to which customers are more interested
- the product characteristics on which customers focus their attention
- the way customers form their information about the products
- the existence of particular dimensions along which consumers of cosmetics perceive natural products
- the existence of particular profiles of customers, in terms of lifestyle, perceptions, sociodemographic characteristics

To understand the attitudes of prospect customers, they develop a questionnaire to investigate the most important factors influencing customers' choices and preferences, as well as a number of variables related to lifestyle, purchase and use habits, and a small number of sociodemographic variables. The questionnaire was administered through the CAWI (Computer Assisted Web Interviewing) technique, with recruitment through social networks (snow-ball sampling). The questionnaire was administrated to 209 respondents, obtaining 138 valid questionnaires. To facilitate respondents, the questionnaire was articulated in several sections:

1. Natural cosmetics The first section of the questionnaire was aimed at collecting general information on the perception of natural cosmetics and on the propensity to purchase the category by the respondents, in order to have a first classification of purchasers of natural cosmetics. In particular, the attitude of the latter has been examined with reference to the categories of products purchased, the frequency of shopping and the methods for finding out the characteristics of the products.
2. Facial care products The second section was aimed at detecting information regarding the purchase behaviour of facial care products, including a series of aspects such as, for example, distribution channels and relevant product attributes. This part was oriented to the evaluation of the variables concerning the entire process of purchase of skincare products by the interviewees, regardless of their propensity, in terms of purchase / non-purchase, towards natural cosmetics.
3. Face Care A section dedicated to the theme of "face care" was introduced, to understand what were the habits and styles of consumption of the respondents compared to the category investigated. For example, it was asked to indicate the type of product most purchased (hydrating, purifying, etc.), also assessing the level of interest and involvement of consumers for the beauty world.
4. Lifestyle Another section was dedicated to lifestyles, submitting to the assessment of the participants a series of statements about different topics such as nutrition, personal care, environmental sustainability and leisure time.

5. Personal Information Finally, the last part has been designated to collect the personal data of the respondents, such as sex, age, educational level and employment status.

ROADMAP:

1. Create an index summarizing the attitude towards natural cosmetics (V18-V26)
2. On the index, regress sociodemographic variables (V94-V97) to check if they have an impact
3. Add variables regarding the way they obtain information (V11-V17)
4. Factor analysis on general lifestyle questions (general: V70-V85, spending: V86-V93)
5. Cluster based on relevant factors found
    - – Add factors found to multiple regression model and check which ones have an impact
6. Profile them by looking at sociodemographic variables in each cluster
7. Further profiling using cross tabulation to look at attitude towards natural products, willingness to buy (V3, V4).
8. Logistic regression on willingness to buy to check if it depends on the distribution channels people choose (face care) and the way they retrieve information before they buy (face care)
    - – Interest in product categories (V5-V10)
9. Definition of natural products to understand how to market a new product, what to highlight

First of all, seeing as there were some missing values in the dataset, we remove the rows containing them and create a secondary dataset with them. This could potentially be used to look into people who did not respond to specific questions.

```
# Import data
df <- read.table("beautiful.txt", header = TRUE, sep = '\t', na.strings = c("NA", "NaN", "", " "))

# Drop the first column (respondent ID) as required
# note: all questions will be identified with (number on questionnaire-1) in the code
df <- df[,!(names(df) %in% colnames(df)[1])]

# Remove rows containing NaN values
beaut <- df[ complete.cases(df), ]

dim(beaut)
```

```
## [1] 122  96
```

16 out of 138 rows were removed, resulting in a dataframe with 122 rows and 96 columns.

We investigated the perception of natural products that people have with respect to traditional products, and we built an index measuring the attitude towards natural products, in such a way that a higher score implies more favourable attitude. In calculating this index we have summed the positive statements along with the inverse of the negative ones (pos. score = 10 - neg. score)

QUESTION 1: Do *sociodemographic* variables have an impact on the attitude people have towards natural cosmetics? In case they do not, does the attitude towards natural products depend on any other variables?

In order to create an "attitude towards natural products" index, we retrieve the information obtained from question 3 (V18-V26) which summarizes respondents' perceptions of natural products with respect to traditional products. In this index a higher score implies a more favourable attitude towards natural products. To calculate it, the positive statements are summed with the inverse of the negative ones such that positive score = 11 - negative score.

To do so, we need to investigate which statements are considered as negative by looking at their correlation with 'Just a marketing trick' (V26) which clearly indicates a negative attitude.

```
corr_exp <- cor(beaut[,17:25])
colnames(corr_exp) <- c(1:9); rownames(corr_exp) <- c(1:9)
round(corr_exp,3)
```

```
##        1      2      3      4      5      6      7      8      9
## 1  1.000  0.566 -0.132  0.537  0.370 -0.025  0.538  0.554 -0.358
## 2  0.566  1.000 -0.117  0.636  0.395  0.007  0.418  0.540 -0.356
## 3 -0.132 -0.117  1.000 -0.101  0.013  0.087 -0.165 -0.192  0.399
## 4  0.537  0.636 -0.101  1.000  0.450 -0.151  0.501  0.567 -0.397
## 5  0.370  0.395  0.013  0.450  1.000 -0.473  0.328  0.357 -0.336
## 6 -0.025  0.007  0.087 -0.151 -0.473  1.000 -0.066 -0.062  0.283
## 7  0.538  0.418 -0.165  0.501  0.328 -0.066  1.000  0.658 -0.263
## 8  0.554  0.540 -0.192  0.567  0.357 -0.062  0.658  1.000 -0.359
## 9 -0.358 -0.356  0.399 -0.397 -0.336  0.283 -0.263 -0.359  1.000
```

"Trendy" (row 3) and "Expensive" (row 6) seem to be positively correlated with "Just a marketing trick" (row 9) and mostly negatively correlated to other variables. "Trendy", "Expensive" and "Just a marketing trick" will be considered as negative statements in our index and therefore their score will be inverted.

```
# Investigate education.level
# it had 4 choices in the survey but no respondents selected 'Primary school' as their education level
unique(beaut[,94])
```

```
## [1] Master Degree    Bachelor Degree  Secondary school
## Levels: Bachelor Degree Master Degree Secondary school
```

Build the attitudes index

```
attitudes <- rowSums( cbind( beaut[17:18], (beaut[19]*-1)+11, beaut[20:21], (beaut[22]*-1)+11, beaut[23
# negative statements: take the inverse and sum 11 to obtain a low positive score

gender <- beaut[,93]
education.level <- beaut[,94]
occupation <- beaut[,95]
age <- beaut[,96]
```

MULTIPLE REGRESSION: regress *sociodemographic variables* (gender, education level, occupation, age) on the attitudes index to check if attitudes are related to sociodemographic characteristics of respondents.

```r
model <- lm(attitudes ~ gender + education.level + occupation + age)
summary(model)
```

```
##
## Call:
## lm(formula = attitudes ~ gender + education.level + occupation +
##     age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.460  -6.365   0.304   7.647  24.794
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     72.3998     4.5055  16.069   <2e-16 ***
## genderM                         -4.0420     6.7041  -0.603   0.5478
## education.levelMaster Degree    -3.5031     2.7996  -1.251   0.2134
## education.levelSecondary school  0.9867     2.5148   0.392   0.6955
## occupationstudent               -6.0328     2.8564  -2.112   0.0368 *
## occupationunemployed             0.6855     4.7228   0.145   0.8848
## age                             -0.2358     0.1044  -2.259   0.0258 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.08 on 115 degrees of freedom
## Multiple R-squared:  0.07323,    Adjusted R-squared:  0.02488
## F-statistic: 1.514 on 6 and 115 DF,  p-value: 0.1795
```

Note: genderF, education.levelBachelor Degree, occupationemployed are taken as a benchmark in the first three categorical variables (age is numerical).

Interpretation of the output:

- Gender and Education level do not have a significant impact on the attitudes index when considering the other variables. These variables should be removed from the model.
- Student is the only significant category in the occupation variable. It is significant at the 5% level: students have a less favourable (-6.03 points on the index) attitude towards natural products with relation to BA graduates.
- Age is significant the 5% level of significance: every additional year leads to a reduction of 0.23 points in the index.

Looking at the adjusted R^2 we can affirm that the model explains only 2.4% of the variability in the attitudes index and the p-value is not close to zero.

We can try to obtain a better result by removing the categories that were not significant from the model (Gender, Education, create dichotomous variable for occupation only indicating student/non-student):

```r
dich.occupation <- rep(0,96)
dich.occupation[which(beaut[,95]=="student")] <- "student"
dich.occupation[which(beaut[,95]!="student")] <- "non student"
contrasts(factor(dich.occupation))
```

```
##             student
## non student       0
## student           1
```

Estimate the model again:

```
model0 <- lm(attitudes ~ factor(dich.occupation) + age)
summary(model0)
```

```
##
## Call:
## lm(formula = attitudes ~ factor(dich.occupation) + age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.299  -6.766   0.101   7.575  23.047
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     71.2118     4.1717  17.070   <2e-16 ***
## factor(dich.occupation)student  -5.1257     2.5932  -1.977   0.0504 .
## age                             -0.2315     0.1039  -2.227   0.0278 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.06 on 119 degrees of freedom
## Multiple R-squared:  0.04554,    Adjusted R-squared:  0.0295
## F-statistic: 2.839 on 2 and 119 DF,  p-value: 0.06244
```

Now the variables are all significant: only being a student or not and the age of a respondent impact his/her attitude towards natural cosmetics.

- Occupation: students have a less positive attitude (-5.12 points in the index) towards natural products compared to non-students (p-value significant at the 5% level). This could potentially be explained by the fact that students have less money and natural products are generally more expensive.
- Age: with every additional year of age, the attitude towards natural products is 0.23 points lower. So, the older respondents are the less positive their attitude is towards natural product. This could possibly be because the "green market" seems to be a trend that is especially popular among young people and older people may be more skeptical.

Looking at the $adjustedR^2$ we can affirm that the model explains only 2.9% of the variability in the attitudes index and the p-value is still not close to zero.

This could be because the sociodemographic variables do not explain enough about the respondents' attitudes towards natural products. To improve this model, we can try adding additional variables.

Among the variables we have, we attempt at using the information regarding *how respondents find out about the characteristics of the natural products they are interested in* (question 2.3, V11-V17) since our explorative research suggested that the way people gather information seems to impact their perceptions of natural products.

```r
ch_web <- factor(beaut[,10])  # dummy because it's a yes/no question
ch_wom <- factor(beaut[,11])
ch_social <- factor(beaut[,12])
ch_adv <- factor(beaut[,13])
ch_salesppl <- factor(beaut[,14])
ch_self <- factor(beaut[,15])
ch_pharma <- factor(beaut[,16])
```

```r
class(ch_web) # 1 is yes, 0 is no
```

```
## [1] "factor"
```

```r
model1 <- lm(attitudes ~ dich.occupation + age + ch_web + ch_wom + ch_social + ch_adv + ch_salesppl + ch
summary(model1)
```

```
##
## Call:
## lm(formula = attitudes ~ dich.occupation + age + ch_web + ch_wom +
##     ch_social + ch_adv + ch_salesppl + ch_self + ch_pharma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.3345  -7.2348  -0.1679   7.0878  24.1136
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             62.71201    5.29752  11.838  < 2e-16 ***
## dich.occupationstudent -7.09826    2.48092  -2.861  0.00504 **
## age                    -0.24486    0.09828  -2.491  0.01419 *
## ch_web1                 0.64457    2.00730   0.321  0.74872
## ch_wom1                -2.87422    2.08469  -1.379  0.17073
## ch_social1              5.57419    2.22971   2.500  0.01387 *
## ch_adv1                -6.74353    2.27535  -2.964  0.00371 **
## ch_salesppl1            2.84049    2.06513   1.375  0.17174
## ch_self1                7.37627    3.41967   2.157  0.03314 *
## ch_pharma1              2.12845    2.26136   0.941  0.34862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.27 on 112 degrees of freedom
## Multiple R-squared:  0.2249, Adjusted R-squared:  0.1626
## F-statistic:  3.61 on 9 and 112 DF,  p-value: 0.0005555
```

Interpretation of the output:

- Both student and age variables have improved significance (lower p-value), meaning they have gained explanatory power with relation to the attitudes towards natural products. Their coefficient slightly changed but not dramatically and still bring the same information.
- There are three relevant variables among the channels people use to fetch information about natural products they are interested in. The first one is 'Blogger/Forums/Social Networks', which is significant at the 5% level. People who retrieve information using social media platforms have a more positive attitude (5.57 points) towards natural products with relation to people who don't.

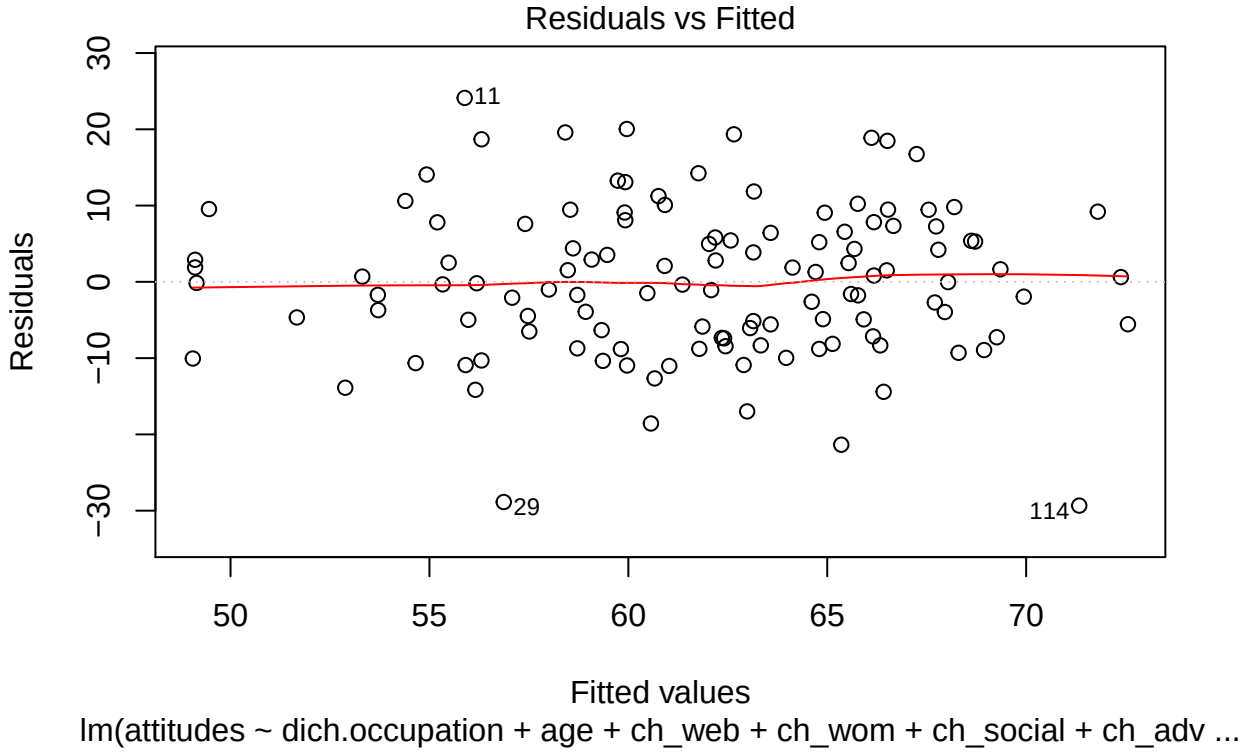
- The second one is "Advertising", with a significance at the 1% level and a negative influence on the index. People who retrieve information from advertising of natural products have a more negative attitude towards natural products (-6.74 points in the index) as compared to people who don't.
- The third and last one is 'Self-provided information', which has a significance at the 5% level. People who retrieve information by themselves have a more positive attitude towards natural products (7.38 points) with relation to people who don't.

Looking at the $adjusted R^2$ we can affirm that the model explains 16% of the variability in the attitudes index, which is a big improvement, and this is significantly different from zero since we obtain a small p-value (1% level).

DIAGNOSTIC CHECKING: Check if the hypothesis we imposed on the model are sensible for this dataset to understand if we can trust the results.

1. *Plot the residuals* over fitted values:

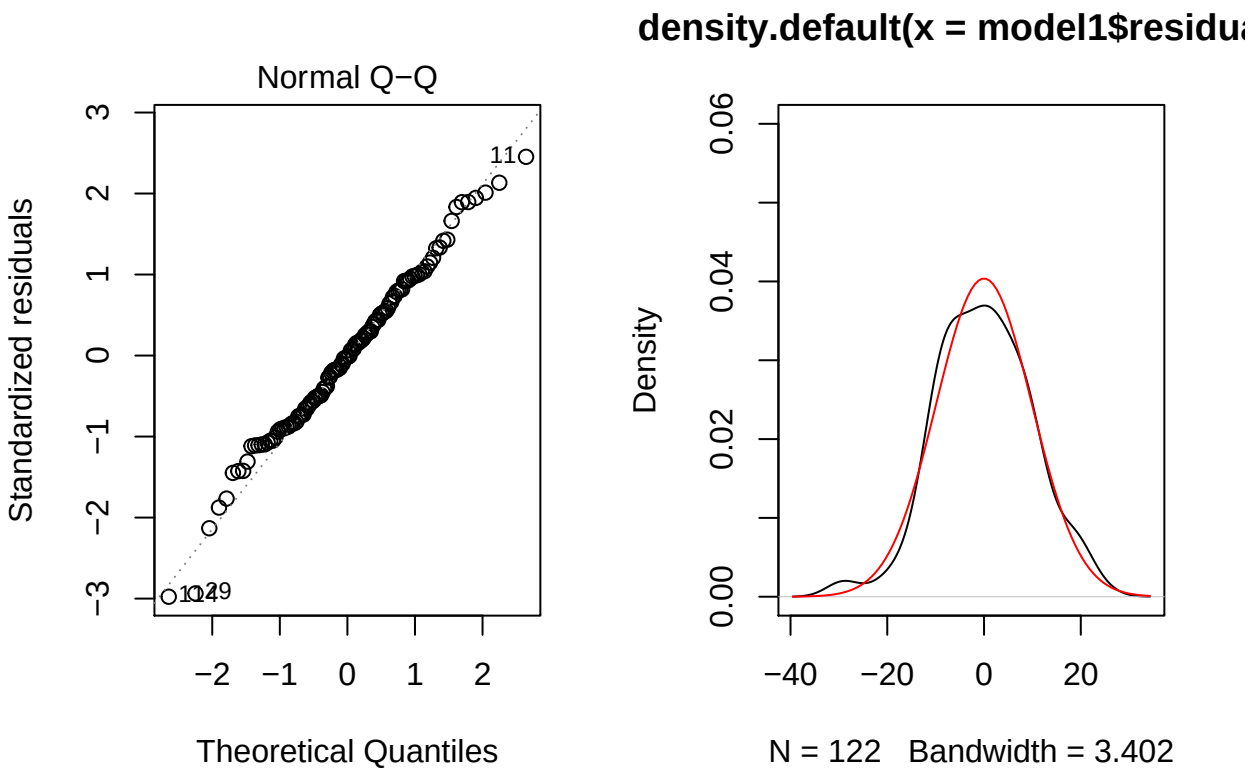
```
plot(model1, 1)
```



The line has no specific shape so there does not seem to be a systematic increase or decrease in the variance, indicating that residuals are homoscedastic. Outliers with large residuals with relation to the average residual are marked with the number of the observation. They could be influencing the estimates.

2. Control for *normal distribution of residuals*:

```r
par(mfrow = c(1, 2))

plot(model1, 2)

plot(density(model1$residuals), ylim = c(0, 0.06))
curve(dnorm(x, mean = mean(model1$residuals), sd = sd(model1$residuals)), add = T, col = 2)
```



Difference in normality only in the tails of the distribution but only slightly.

3. *Kolmogorov-Smirnov Test*:

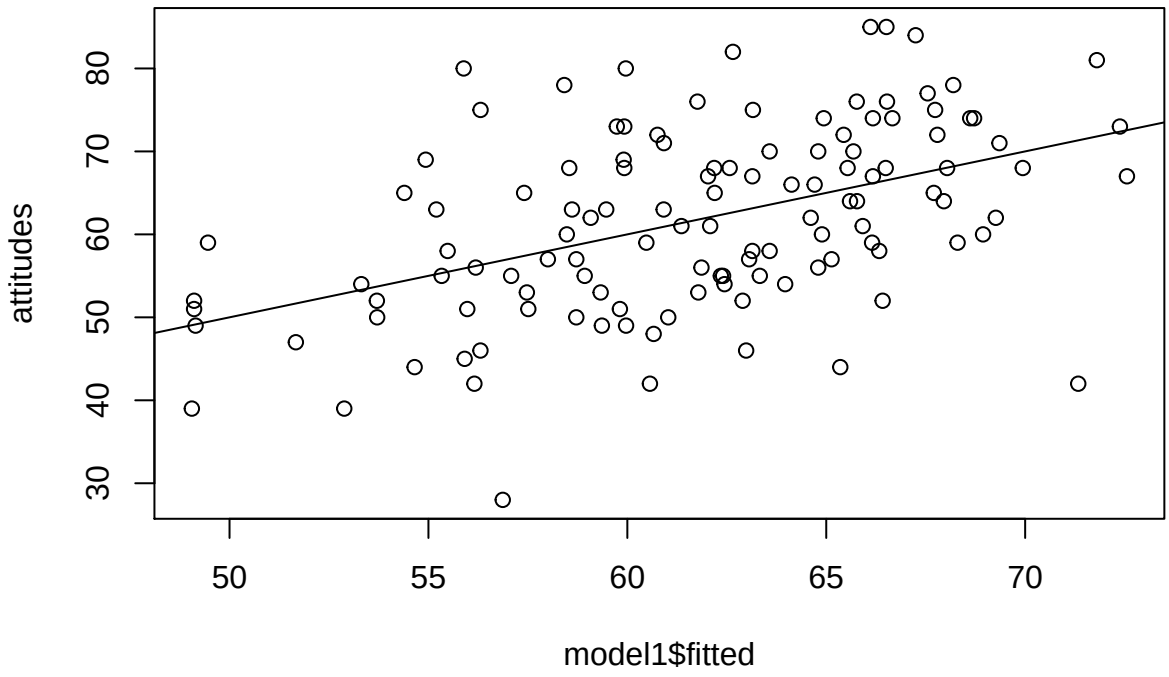
```r
ks.test(model1$residuals, "pnorm")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  model1$residuals
## D = 0.39842, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

The p-value close to zero. Reject the hypothesis that the empirical distribution we are observing comes from the normal distribution (KS test null hypothesis).

4. *Goodness of fit* between fitted and observed value:

```r
par(mfrow=c(1,1)); plot(model1$fitted, attitudes); abline(0,1)
```



Dots do not lie on the bisector so the model does not give a good fit, as anticipated by the low $adjustedR^2$.

To sum up: the analysis of residuals confirmed that they were homoscedastic, however, the hypothesis of normality is rejected by the KS test and the model does not give a good fit as we can see from the graph.

We can try removing the outliers and re-run the regression and test of normality to check that they were not impacting the estimates excessively.

REMOVE OUTLIERS:

```r
# Remove outliers from index and variables of interest
attitudes_nout <- attitudes[-c(11,29,114)]
dich.occupation_nout <- dich.occupation[-c(11,29,114)]
age_nout <- age[-c(11,29,114)]
ch_social_nout <- ch_social[-c(11,29,114)]
ch_adv_nout <- ch_adv[-c(11,29,114)]
ch_self_nout <- ch_self[-c(11,29,114)]
```

```r
model1_nout <- lm(attitudes_nout ~ dich.occupation_nout + age_nout + ch_social_nout + ch_adv_nout + ch_
summary(model1_nout)
```

```
##
## Call:
## lm(formula = attitudes_nout ~ dich.occupation_nout + age_nout +
```

```
##     ch_social_nout + ch_adv_nout + ch_self_nout)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.2960  -7.1162  -0.2734   6.5198  22.5494
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   64.33062    5.01565  12.826   <2e-16 ***
## dich.occupation_noutstudent -5.75581    2.44281  -2.356   0.0202 *
## age_nout                      -0.23034    0.09872  -2.333   0.0214 *
## ch_social_nout1                4.51671    2.15179   2.099   0.0380 *
## ch_adv_nout1                  -7.03866    2.14849  -3.276   0.0014 **
## ch_self_nout1                  6.47968    3.33412   1.943   0.0544 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.26 on 113 degrees of freedom
## Multiple R-squared:  0.1704, Adjusted R-squared:  0.1337
## F-statistic: 4.641 on 5 and 113 DF,  p-value: 0.0006861
```

Adjusted $R^2 = 13\%$ is lower than the previous one (16%), so the outliers were not interfering with the goodness of fit of the model. The low p-value indicates that the statistic is significantly different from zero.

Summary: This model cannot be used for prediction because the R^2 is low and the diagnostic checking was not successful, in particular the KS test has p-value close to zero and the last graph portrays a poor goodness of fit with the model. We can however still use it to investigate what influences attitudes towards natural products as we have investigated up until now.

QUESTION 2: Is people's lifestyle a unidimensional trait or does it have dimensions to it?

We want to perform factor analysis on the numerous questions investigating the respondents' lifestyle and habits (V70-V85) to reduce the number of variables used to indicate lifestyle, so that they can then be more easily used for other purposes such as clustering and adding them to the multiple regression model.

FACTOR ANALYSIS 1. Evaluate if there is any correlation worth exploring among the variables.

```r
# Correlation matrix
corr <- cor(beaut[,69:84])
colnames(corr) <- c(1:16); rownames(corr) <- c(1:16)
round(corr,3)
```

```
##         1     2     3     4     5     6     7     8     9    10    11
## 1   1.000 0.654 0.420 0.314 0.668 0.202 0.224 0.272 0.331 0.248 0.226
## 2   0.654 1.000 0.377 0.221 0.622 0.073 0.190 0.290 0.424 0.245 0.234
## 3   0.420 0.377 1.000 0.291 0.323 0.456 0.463 0.490 0.445 0.346 0.400
## 4   0.314 0.221 0.291 1.000 0.402 0.289 0.231 0.292 0.219 0.072 0.084
## 5   0.668 0.622 0.323 0.402 1.000 0.152 0.130 0.315 0.249 0.209 0.187
## 6   0.202 0.073 0.456 0.289 0.152 1.000 0.413 0.236 0.450 0.379 0.346
## 7   0.224 0.190 0.463 0.231 0.130 0.413 1.000 0.454 0.459 0.364 0.365
## 8   0.272 0.290 0.490 0.292 0.315 0.236 0.454 1.000 0.335 0.405 0.408
## 9   0.331 0.424 0.445 0.219 0.249 0.450 0.459 0.335 1.000 0.476 0.344
## 10  0.248 0.245 0.346 0.072 0.209 0.379 0.364 0.405 0.476 1.000 0.558
## 11  0.226 0.234 0.400 0.084 0.187 0.346 0.365 0.408 0.344 0.558 1.000
```

```
## 12 -0.092 -0.150 -0.244 0.046 -0.003  0.022 -0.207 -0.114 -0.148 -0.069  0.108
## 13 -0.004  0.039  0.014 0.063  0.056 -0.035  0.061  0.031  0.013 -0.169 -0.068
## 14  0.123  0.072  0.020 0.095  0.207 -0.001  0.070  0.139 -0.058  0.034 -0.007
## 15  0.331  0.272  0.039 0.189  0.367 -0.024  0.050  0.155  0.148  0.140  0.062
## 16  0.419  0.339  0.197 0.236  0.414  0.101  0.141  0.267  0.125  0.233  0.085
##        12     13     14     15     16
## 1  -0.092 -0.004  0.123  0.331  0.419
## 2  -0.150  0.039  0.072  0.272  0.339
## 3  -0.244  0.014  0.020  0.039  0.197
## 4   0.046  0.063  0.095  0.189  0.236
## 5  -0.003  0.056  0.207  0.367  0.414
## 6   0.022 -0.035 -0.001 -0.024  0.101
## 7  -0.207  0.061  0.070  0.050  0.141
## 8  -0.114  0.031  0.139  0.155  0.267
## 9  -0.148  0.013 -0.058  0.148  0.125
## 10 -0.069 -0.169  0.034  0.140  0.233
## 11  0.108 -0.068 -0.007  0.062  0.085
## 12  1.000  0.153 -0.052  0.031 -0.001
## 13  0.153  1.000  0.521  0.209  0.132
## 14 -0.052  0.521  1.000  0.543  0.403
## 15  0.031  0.209  0.543  1.000  0.546
## 16 -0.001  0.132  0.403  0.546  1.000
```

There seem to be some correlations worth exploring, although not too strong.

2. Evaluate if the correlation reported in the matrix is enough to proceed with factor analysis with KMO index and Bartlett test for independence among variables.

```
# KMO index
KMO(beaut[, 69:84])
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = beaut[, 69:84])
## Overall MSA =  0.77
## MSA for each item =
##      takecareimage        look.fundam     carenatmethods        regExercise
##               0.86               0.80               0.85               0.82
##         ImpGoodapp              Envir        organicfood           NatSuppl
##               0.83               0.74               0.86               0.86
##       BetterPerson           ReadLab         Difference         RefinedInd
##               0.77               0.79               0.77               0.35
## Highpricebrandqual      DesignClothes       FollowTrends        personality
##               0.45               0.56               0.72               0.84
```

```
# Bartlett test of sphericity
cortest.bartlett(beaut[, 69:84])
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 706.9655
##
```

```
## $p.value
## [1] 5.958138e-84
##
## $df
## [1] 120
```

It seems worthwhile to proceed with factor analysis because the variables are correlated as indicated by:

- Overall MSA = 0.77, bigger than 0.50 threshold. They share common variability.
- Bartlett test rejects the hypothesis of independence (equal to null matrix) with a p-value close to zero.

We start with the *hypothesis of 5 underlying factors* to summarize our 16 variables (number smaller than the number of variables).

```
factan5 <- factanal(beaut[,69:84], 5)
factan5
```

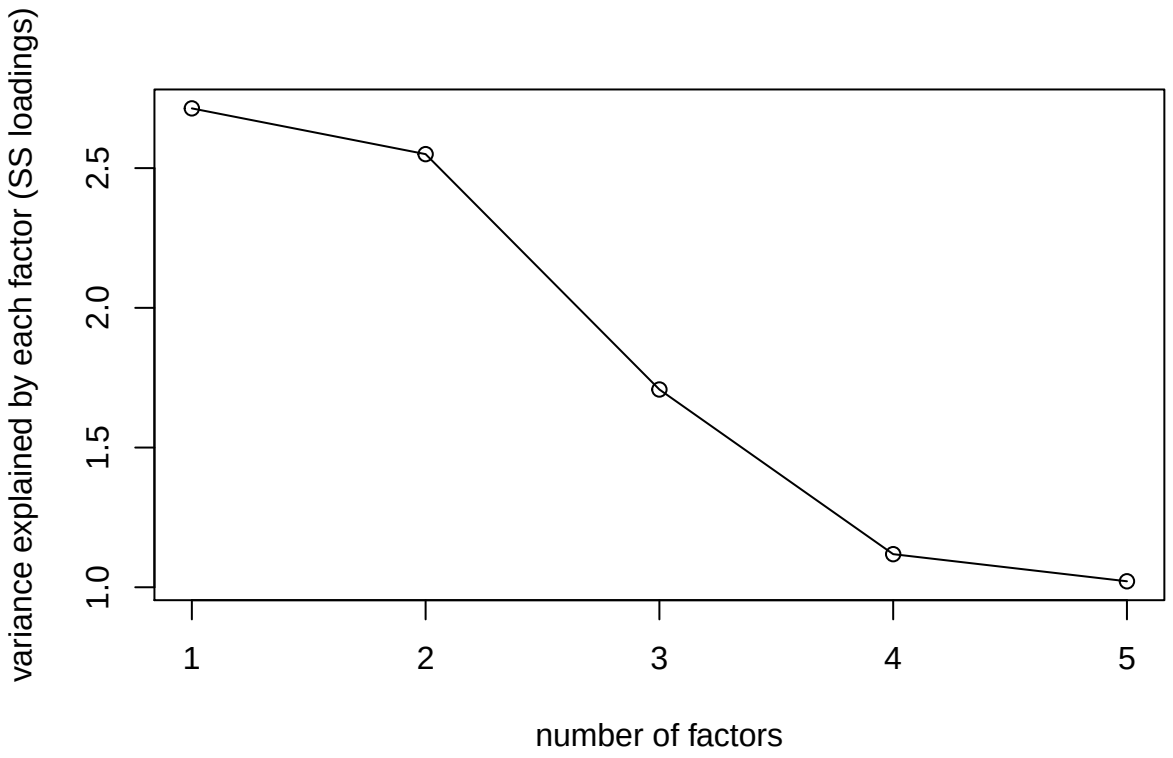
```
##
## Call:
## factanal(x = beaut[, 69:84], factors = 5)
##
## Uniquenesses:
##      takecareimage        look.fundam     carenatmethods        regExercise
##              0.314              0.390              0.404              0.745
##         ImpGoodapp              Envir        organicfood           NatSuppl
##              0.333              0.587              0.513              0.615
##       BetterPerson            ReadLab         Difference         RefinedInd
##              0.550              0.066              0.554              0.005
## Highpricebrandqual      DesignClothes       FollowTrends        personality
##              0.615              0.053              0.538              0.606
##
## Loadings:
##                   Factor1 Factor2 Factor3 Factor4 Factor5
## takecareimage       0.238   0.790
## look.fundam         0.206   0.744          -0.101
## carenatmethods      0.695   0.295          -0.159
## regExercise         0.319   0.342                  -0.142
## ImpGoodapp          0.158   0.787   0.138
## Envir               0.628
## organicfood         0.678                  -0.138
## NatSuppl            0.543   0.229   0.113          0.144
## BetterPerson        0.573   0.261                  0.197
## ReadLab             0.475   0.125                  0.832
## Difference          0.527   0.112           0.168  0.357
## RefinedInd         -0.106                  0.990
## Highpricebrandqual                 0.565   0.133 -0.212
## DesignClothes                      0.962
## FollowTrends                0.408  0.518          0.157
## personality                 0.469  0.371          0.174
##
##                Factor1 Factor2 Factor3 Factor4 Factor5
## SS loadings      2.714   2.550   1.708   1.118   1.021
```

```
## Proportion Var   0.170   0.159   0.107   0.070   0.064
## Cumulative Var   0.170   0.329   0.436   0.506   0.569
##
## Test of the hypothesis that 5 factors are sufficient.
## The chi square statistic is 66.45 on 50 degrees of freedom.
## The p-value is 0.0596
```

Choice of number of factors: SS Loadings seem to suggest that a 5-component solution is sufficient since all 5 factors explain more than a single variable (loadings larger than 1) and the first 5 factors together explain 57% of the total variability (close to the 60% threshold). Moreover, the p-value accepts the hypothesis that 5 factors are sufficient.

Scree plot:

```
plot(seq(1:5), colSums(factan5$loadings^2), xlab="number of factors", ylab="variance explained by each f
lines(seq(1:5), colSums(factan5$loadings^2))
```



The scree plot suggests that maybe 4 factors are a better choice. However, when attempting a 4-component solution a p-value close to zero rejected that 4 factors were sufficient.

Therefore, *five factors were obtained and rotated*:

```
factan55 <- factanal(beaut[,69:84], 5, rotation = "varimax", scores = "regression")
factan55
```

```
##
```

```
## Call:
## factanal(x = beaut[, 69:84], factors = 5, scores = "regression",     rotation = "varimax")
##
## Uniquenesses:
##      takecareimage        look.fundam     carenatmethods        regExercise
##              0.314              0.390              0.404              0.745
##         ImpGoodapp              Envir        organicfood           NatSuppl
##              0.333              0.587              0.513              0.615
##       BetterPerson            ReadLab         Difference         RefinedInd
##              0.550              0.066              0.554              0.005
## Highpricebrandqual      DesignClothes       FollowTrends        personality
##              0.615              0.053              0.538              0.606
##
## Loadings:
##                    Factor1 Factor2 Factor3 Factor4 Factor5
## takecareimage       0.238   0.790
## look.fundam         0.206   0.744          -0.101
## carenatmethods      0.695   0.295          -0.159
## regExercise         0.319   0.342                  -0.142
## ImpGoodapp          0.158   0.787   0.138
## Envir               0.628
## organicfood         0.678                  -0.138
## NatSuppl            0.543   0.229   0.113           0.144
## BetterPerson        0.573   0.261                   0.197
## ReadLab             0.475   0.125                   0.832
## Difference          0.527   0.112           0.168   0.357
## RefinedInd         -0.106                   0.990
## Highpricebrandqual                 0.565   0.133  -0.212
## DesignClothes                      0.962
## FollowTrends                0.408   0.518           0.157
## personality                 0.469   0.371           0.174
##
##                Factor1 Factor2 Factor3 Factor4 Factor5
## SS loadings      2.714   2.550   1.708   1.118   1.021
## Proportion Var   0.170   0.159   0.107   0.070   0.064
## Cumulative Var   0.170   0.329   0.436   0.506   0.569
##
## Test of the hypothesis that 5 factors are sufficient.
## The chi square statistic is 66.45 on 50 degrees of freedom.
## The p-value is 0.0596
```

Here, the p-value is larger than 5% so we can accept the hypothesis that 5 factors are sufficient.

In order to try and achieve a cumulative variance higher that 60%, we attempt at taking out the variables that have the highest uniquenesses (we cannot reduce the number of factors). High uniquenesses: regExercise, NatSupply.

```
# Variables with high uniquenesses removed
factan551 <- factanal(cbind(beaut[,69:71],beaut[,73:75], beaut[,77:84]), 5, rotation="varimax", scores=
factan551
```

```
##
## Call:
## factanal(x = cbind(beaut[, 69:71], beaut[, 73:75], beaut[, 77:84]),     factors = 5, scores = "regres
```

```
##
## Uniquenesses:
##      takecareimage        look.fundam     carenatmethods         ImpGoodapp
##              0.299              0.374              0.428              0.355
##              Envir        organicfood       BetterPerson           ReadLab
##              0.552              0.542              0.528              0.005
##         Difference         RefinedInd Highpricebrandqual      DesignClothes
##              0.568              0.005              0.625              0.018
##       FollowTrends        personality
##              0.550              0.614
##
## Loadings:
##                     Factor1 Factor2 Factor3 Factor4 Factor5
## takecareimage        0.798   0.246
## look.fundam          0.755   0.212
## carenatmethods       0.300   0.675          -0.161
## ImpGoodapp           0.778   0.136   0.132
## Envir                        0.656
## organicfood                  0.651          -0.143
## BetterPerson         0.268   0.593                  0.186
## ReadLab              0.146   0.452                  0.876
## Difference           0.129   0.512           0.165   0.355
## RefinedInd                  -0.103           0.990
## Highpricebrandqual                   0.564   0.131  -0.193
## DesignClothes        0.107           0.977  -0.105
## FollowTrends         0.418           0.499           0.147
## personality          0.478           0.352           0.176
##
##                Factor1 Factor2 Factor3 Factor4 Factor5
## SS loadings      2.435   2.274   1.680   1.107   1.042
## Proportion Var   0.174   0.162   0.120   0.079   0.074
## Cumulative Var   0.174   0.336   0.456   0.535   0.610
##
## Test of the hypothesis that 5 factors are sufficient.
## The chi square statistic is 42.74 on 31 degrees of freedom.
## The p-value is 0.0781
```

We obtain a positive result since the cumulative variance for 5 factors is now up to 61%.

We proceed in *naming the factors* by looking at the correlation between the factors and the original variables.

- FACTOR 1: More correlated with the variables takecareimage ("I take care a lot of my image"), look.fundam ("Taking care of one's look is fundamental for his/her wellbeing"), and ImpGoodapp ("It's important to always have a good appearance"). We can call it an **Appearance Factor**.
- FACTOR 2: More correlated with the variables carenatmethods ("I take care of myself through natural methods"), Envir "Reducing impact on environment, with an environmentally friendly lifestyle"), organicfood ("Make use of organic food products"), BetterPerson ("Using natural products makes me feel like a better person"), Difference ("Know difference between natural/organic"). We can call it a **Sustainability Factor**.
- FACTOR 3: More correlated with the variables DesignClothes ("Prefer designer clothes and well known brand products"), Highpricebrandqual ("High price and well-known brand are synonyms of quality"), and FollowTrends ("Follow trends as seen in social media"). We can call it a **Trend Factor**.

- FACTOR 4: Most correlated to the variable RefinedInd ("Consume refined/industrial food products"). We can call it a **Traditional Factor**.
- FACTOR 5: Most correlated to the variable ReadLab ("I always read the labels of what I buy"). We can call it an **Attention Factor**.

To resume the findings, we can state that the respondents' lifestyles are characterized by five underlying dimensions: an appearance dimension, a sustainability dimension, a trend dimension, a traditional dimension and an attention one.

Question 2.1: Does the lifestyle of respondents lead influence their attitudes towards natural products? What factors drive attitude towards natural products? Our explorative research links the lifestyles led by people to their perceptions of natural products. We want to confirm or deny this by adding the factors found to the multiple regression model we have previously used to investigate what impacts the attitudes index.

```
# Extract factor scores for use
fact.scores <- factan551$scores
appearance <- fact.scores[,1]
sustainability <- fact.scores[,2]
trend <- fact.scores[,3]
traditional <- fact.scores[,4]
attention <- fact.scores[,5]
```

Adding lifestyle factors to the multiple regression on attitudes index

```
model1_factors <- lm(attitudes ~ dich.occupation + age + ch_social + ch_adv + ch_self + appearance + su
summary(model1_factors)
```

```
##
## Call:
## lm(formula = attitudes ~ dich.occupation + age + ch_social +
##     ch_adv + ch_self + appearance + sustainability + trend +
##     traditional + attention)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.931  -5.771   0.932   6.230  20.752
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              64.5339     4.6531  13.869  < 2e-16 ***
## dich.occupationstudent  -5.1130     2.2496  -2.273 0.024956 *
## age                      -0.2031     0.0925  -2.195 0.030235 *
## ch_social1                5.2736     2.0152   2.617 0.010108 *
## ch_adv1                  -5.1207     2.1004  -2.438 0.016358 *
## ch_self1                  3.8063     3.1344   1.214 0.227191
## appearance                1.5075     0.9665   1.560 0.121659
## sustainability            4.1858     1.0706   3.910 0.000159 ***
## trend                     0.4808     0.8872   0.542 0.588983
## traditional              -0.9320     0.8819  -1.057 0.292905
## attention                 1.4915     0.9338   1.597 0.113062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 9.51 on 111 degrees of freedom
## Multiple R-squared:  0.3413, Adjusted R-squared:  0.282
## F-statistic: 5.752 on 10 and 111 DF,  p-value: 6.367e-07
```

Adding them has kept the significance of previous variables roughly the same. Only one of the factors is significant (at the 1% level): it is *Sustainability*. This means that attitudes towards natural products are positively related (positive coefficient) only to leading a particularly "ecological" lifestyle.

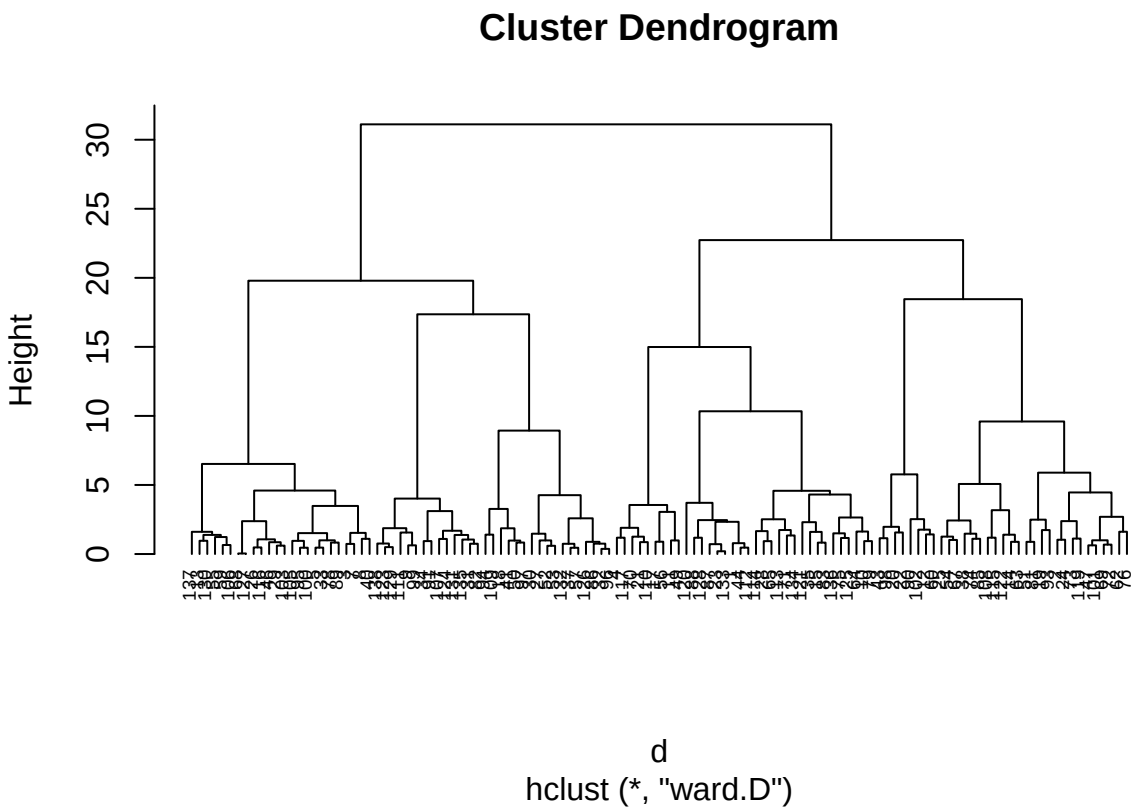
Looking at the adjusted R^2 we can affirm that the model explains 28% of the variability in the attitudes index, which is a considerable improvement, and this is significantly different from zero since we obtain a p-value that is very close to zero.

QUESTION 3: Can we identify different groups of respondents according to the characteristics of their lifestyle (resumed by the five factors)?

We are aiming to perform a market segmentation by (1) obtaining the clusters by applying a clustering algorithm, (2) naming the clusters by looking at the average scores for each (centroids), (3) profiling them according to external variables useful to understand consumer behavior and characteristics.

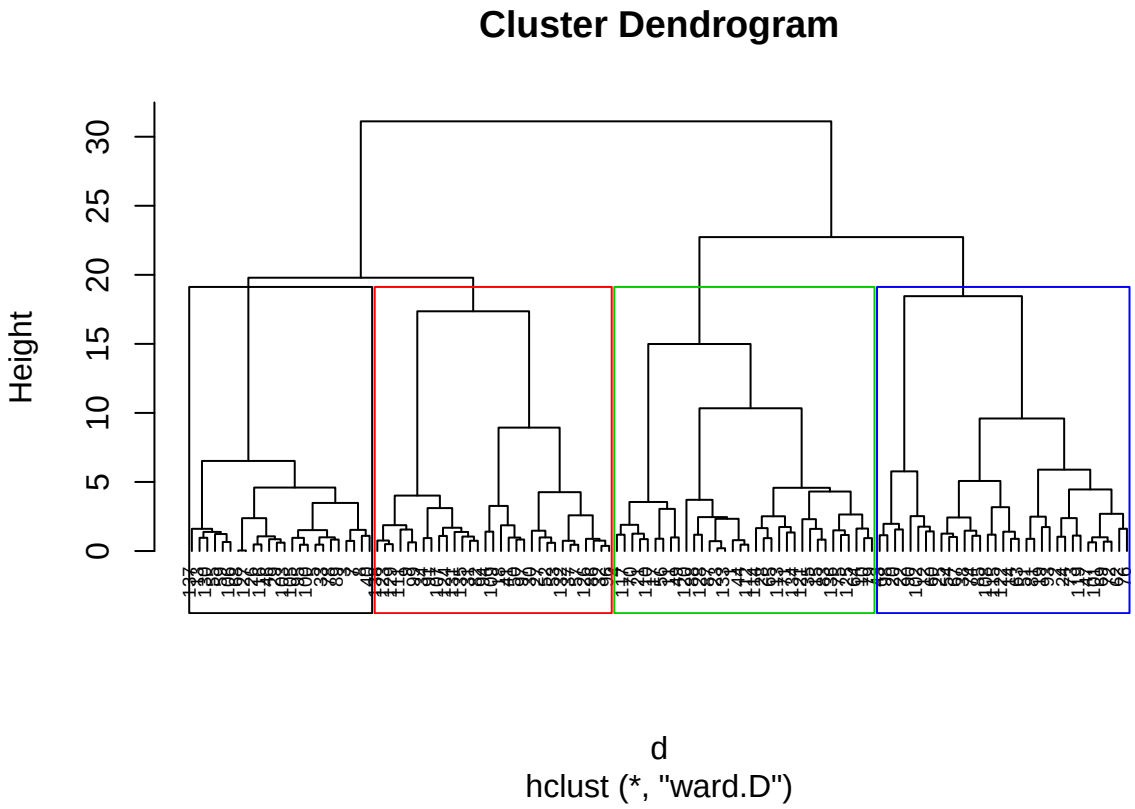
1) **Obtaining the clusters**

```r
# Calculate the distance matrix as the starting point for clustering
d <- dist(fact.scores, method = "euclidean")

clu <- hclust(d, method = "ward.D") # Ward method was the best performing

# Plot dendogram
plot(clu, hang = -1, cex = 0.6)
```



We have selected four as the number of clusters as a result of the difficult trade-off between cutting where we have the longest distance between subsequent agglomeration sets and searching for homogeneous and well-separated groups.

```
# Select four clusters and use rect.hclust to retrieve the group membership
plot(clu, hang = -1, cex = 0.6)
rect.hclust(clu, k = 4, which = c(1,2,3,4), border = 1:4)
```




```
memb <- cutree(clu, k = 4)
memb
```

```
##   1   2   3   4   5   7   8   9  10  11  12  13  14  15  17  18  19  20  21  22
##   1   2   3   1   3   3   3   4   1   1   2   3   2   1   1   4   2   2   1   2
##  23  24  25  26  27  28  29  30  31  32  33  34  35  37  38  40  41  43  44  46
##   1   2   2   3   4   1   3   4   4   4   3   4   1   4   2   3   4   2   1   3
##  47  49  50  51  52  53  54  55  56  57  59  60  62  63  64  65  67  68  69  70
##   2   1   4   2   4   4   2   3   1   4   3   2   2   2   1   1   2   3   2   1
##  71  72  73  74  75  76  78  79  81  82  83  85  86  87  88  89  90  91  93  94
##   1   2   2   2   1   2   3   1   2   4   1   2   4   1   1   3   2   4   2   4
##  95  96  97  98  99 100 101 102 103 105 106 107 108 109 110 111 112 113 114 115
##   3   4   4   2   4   3   2   2   3   3   3   4   2   4   1   4   1   1   1   2
## 116 117 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136
##   3   1   2   1   1   2   4   4   1   4   3   1   4   3   4   4   1   1   4   1
## 137 138
##   3   4
```

2) **Naming the clusters** by analyzing centroids

```
# Initialize a matrix with 4 rows (number of clusters) and 5 columns (number of factors used to cluster
clu.scores<-matrix(0,4,5)
rownames(clu.scores)<-c("cluster1","cluster2","cluster3","cluster4")
for (i in 1:4){
  for (j in 1:5){
```

```
    clu.scores[i,j]<- round(mean(fact.scores[memb==i,j]),3)

    }
  }

colnames(clu.scores)<-c("Appearance","Sustainability","Trendy","Traditional","Attention")
clu.scores
```

```
##          Appearance Sustainability Trendy Traditional Attention
## cluster1      0.263          0.023  1.173       0.155     0.065
## cluster2     -0.427         -0.747 -0.201      -0.215    -0.741
## cluster3      0.297          0.736 -0.612      -0.858     0.216
## cluster4     -0.064          0.200 -0.599       0.723     0.550
```

Note that factor scores are centered on zero. Therefore, the higher the score on the positive axis, the larger the level of agreement with that specific factor from respondents in that cluster; the opposite is true for negative values. Values close to zero will be considered as evaluated with neutrality by respondents.

- CLUSTER 1: **Fashionistas**. It is composed of consumers who have a very high score in the trendy factor, as well as significant score in the appearance factor. These people care about their image, follow the latest trends and prefer famous brands as an indicator of quality. They somewhat align with the traditional factor, meaning they buy industrial refined products. They are neutral with regards to sustainability and attention.
- CLUSTER 2: **Skeptically Indifferent**. It is composed of respondents who do not value appearance, sustainability and attention. They are somewhat on the negative side of the trendy and traditional factors too. These people do not lead an environmentally lifestyle and are not inclined to read labels on what they buy, but at the same time they do not necessarily buy refined products. They do not pay particular attention to their appearance and they also do not care too much about trends and brands. We imagine them as customers who will buy what is more convenient and/or effective rather than choosing based on what is natural or refined.
- CLUSTER 3: **Environmentalists**. It is composed of respondents who lead environmentally friendly lifestyles. In fact, we can observe low scores on the trendy factor and on the traditional one. They somewhat value appearance and pay attention to a certain extent to what they buy. These people are involved in actively reducing their impact on the environment. In fact, they do not follow trends and probably stay away of big famous brands since they also avoid buying industrial and refined products.
- CLUSTER 4: **Traditionals**. It is composed of respondents who do not behave according to the trendy factor, but they do follow the traditional factor. They value the attention factor and somewhat value the sustainability one. They are indifferent to appearance. These people are "traditional buyers" since they are probably conscious of what an eco-friendly lifestyle is, but they do not necessarily incorporate it in their behaviors. They do not follow trends or value famous brands. They buy industrial and refined products but tend to be careful of what they buy by reading the labels.

3) **Profiling the clusters**: characterize the clusters according to external variables other than the ones used to build them.

- Categorical variables: compare the percentage of each category with respect to the whole sample.
- Numerical variables: compare the average (median) of the variable with respect to the whole sample. In each case, understand which clusters differentiate themselves from the others with regard to any particular category and using the whole sample as a benchmark.

sociodemographic variables

Gender:

```
clusters <- c("Fashionistas", "Skeptically Indifferent", "Environmentalist", "Traditionals", "sample")

res_gender<-rbind(round(prop.table(table(memb, gender),1)*100,1),
                  round(prop.table(table(gender))*100,1))
rownames(res_gender)<-clusters
res_gender
```

```
##                             F   M
## Fashionistas            100.0 0.0
## Skeptically Indifferent  97.0 3.0
## Environmentalist         95.8 4.2
## Traditionals             96.8 3.2
## sample                   97.5 2.5
```

We can trust these results up to a certain extent since, as we can see from the sample values, the respondents were mainly females. Therefore, we may not have enough information about males to draw confident conclusions about them. It is also true, however, that, as verified in our exploratory research, usually the main purchasers of natural cosmetics are women. Anyway, from the table we can see that:

- Fashionistas are females in a higher percentage with respect to the whole sample. Actually, the cluster is composed solely of female respondents.
- Skeptically Indifferents are in line with the whole sample.
- Environmentalists are also roughly in line with the whole sample, with slightly less females and slightly more males.
- The same holds for Traditionals but to a lesser extent.

Education:

```
res_edu<-rbind(round(prop.table(table(memb, education.level),1)*100,1),
                  round(prop.table(table(education.level))*100,1))
rownames(res_edu)<-clusters
res_edu
```

```
##                         Bachelor Degree Master Degree Secondary school
## Fashionistas                       38.2          29.4             32.4
## Skeptically Indifferent            45.5          30.3             24.2
## Environmentalist                   41.7          29.2             29.2
## Traditionals                       25.8          25.8             48.4
## sample                             37.7          28.7             33.6
```

From the table we can see that:

- Fashionistas do not show any considerable difference from the sample.
- Skeptically Indifferents have a higher percentage of BA graduates as well as Master graduates, while a lower percentage of people with only a secondary school diploma.
- Environmentalists also have a higher percentage of BA graduates with relation to the whole sample. They also have slightly higher one for Master graduates and a lower one from secondary school.
- Traditionals have a lower percentage of BA and Master graduates with relation to the whole sample and a higher percent of secondary school graduates.

Occupation:

```
res_occ<-rbind(round(prop.table(table(memb, occupation),1)*100,1),
                  round(prop.table(table(occupation))*100,1))
rownames(res_occ)<-clusters
res_occ
```

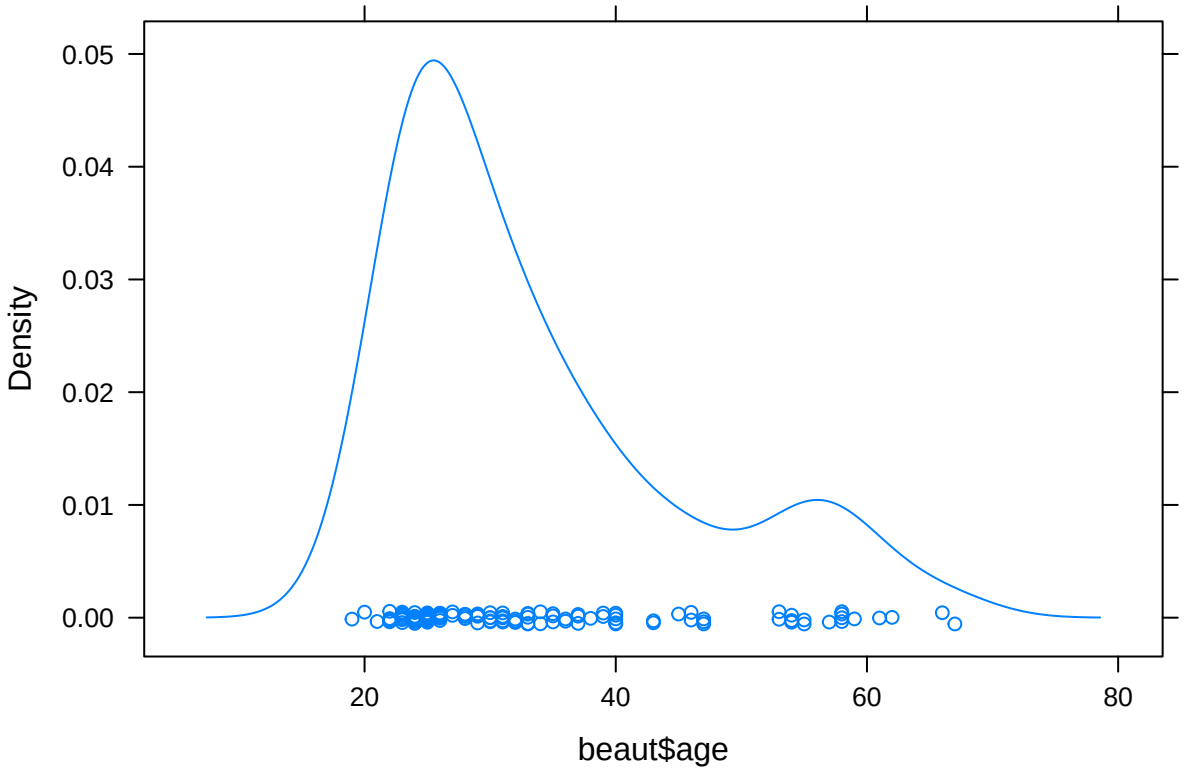
```
##                          Employed student unemployed
## Fashionistas                 67.6    32.4        0.0
## Skeptically Indifferent      57.6    39.4        3.0
## Environmentalist             54.2    33.3       12.5
## Traditionals                 64.5    29.0        6.5
## sample                       61.5    33.6        4.9
```

From the table we can see that:

- Fashionistas have a higher percentage of employed with relation to the whole sample. They are in line with the student percentage and they have no unemployed respondents.
- Skeptically Indifferents have a lower percentage of employed with relation to the whole sample. They also have a higher percentage of students and a slightly lower one of unemployed.
- Environmentalists have a lower percentage (the lowest) of employed with relation to the whole sample. Their percentage of students is in line with the sample, while they have a higher (highest) percentage of unemployed.
- Traditionals have a higher percentage of employed, a lower percentage of students and a higher percentage of unemployed with relation to the whole sample.

Age:

```
# Density plot to check where to place intervals for age
densityplot(beaut$age)
```




```
# Made age a categorical variable to portray generations (more informative than median)
short_age <- cut(beaut$age, breaks = c(18, 25, 35, 45, 55, Inf), labels = c('19-25', '26-35', '36-45',
```

```
res_age<-rbind(round(prop.table(table(memb, short_age),1)*100,1),
                  round(prop.table(table(short_age))*100,1))
rownames(res_age)<-clusters
res_age
```

```
##                          19-25 26-35 36-45 46-55 55-67
## Fashionistas              35.3  41.2   8.8   5.9   8.8
## Skeptically Indifferent  42.4  36.4   3.0   6.1  12.1
## Environmentalist         37.5  20.8  25.0  12.5   4.2
## Traditionals             25.8  32.3  19.4  16.1   6.5
## sample                   35.2  33.6  13.1   9.8   8.2
```

From the table we can see that:

- Fashionistas have a percentage of people between the ages of 19-25 and 55-67 that is in line with the whole sample. They show a higher percentage of people between the ages of 26-35 and a lower one between 36-45 and 46-55 with relation to the whole sample.
- Skeptically Indifferents have a higher percentage of people between 19-25, 26-35 and 55-67 with relation to the whole sample. On the other hand, they show a lower percentage between 36-45 and 46-55.

- Environmentalists have a slightly higher percentage of people aged 19-25 and 46-55 compared to the whole sample, and a higher one for people between 36-45. They also show a considerably lower percentage of people aged 26-35 and 55-67.
- Traditionals show a lower percentage of people aged 19-25 compared to the whole sample. They also have a percentage of 26-35 and 55-67 that is lower with respect to the sample. They have a higher percentage of people between the ages of 36-45 and 46-55 with relation to the whole sample.

Verify significance of relationships The cluster's membership is a categorical variable, so we can test if the described relationships are statistically significant (chi squared test of independence between two variables: categorical variable, membership to the cluster).

```
# build a list with tables for all our categorical sociodemographic variables all of which tabulated ac
res1 <-list(table(memb,gender), table(memb,education.level),
            table(memb, occupation), table(memb,short_age))
names(res1) <- c("gender","education","occupation", "age")
lapply(res1,summary) # apply summary to all of the elements of the list in a compact way
```

```
## $gender
## Number of cases in table: 122
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1.2698, df = 3, p-value = 0.7363
##  Chi-squared approximation may be incorrect
##
## $education
## Number of cases in table: 122
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 4.953, df = 6, p-value = 0.5499
##
## $occupation
## Number of cases in table: 122
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 5.951, df = 6, p-value = 0.4287
##  Chi-squared approximation may be incorrect
##
## $age
## Number of cases in table: 122
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 13.346, df = 12, p-value = 0.3444
##  Chi-squared approximation may be incorrect
```

With high p-values we cannot reject the hypothesis of independence, so we cannot state that gender, education, occupation and age have an impact on the membership to clusters. This means that membership to clusters does not depend on the sociodemographic variables of a respondent. Anyway, keeping this in mind, our clusters were still characterized by some striking differences from the whole sample in the sociodemographic variables mentioned that we coherent with other findings. We will mention this in the profiling purely as a measure of what we have found in our particular study about the distribution of the sociodemographic variables among the clusters, although the same may not necessarily hold for the whole population.

Attitude towards natural products

```
#attitude
m_att <-matrix(0,1,5)
colnames(m_att)<-clusters
rownames(m_att)<-c("attitude mean")

for (i in 1:4){m_att[1,i]<-mean(attitudes[memb==i])}
#mean for the whole sammple
m_att[5]<-mean(attitudes)
m_att
```

```
##               Fashionistas Skeptically Indifferent Environmentalist
## attitude mean     61.58824               56.12121         65.83333
##               Traditionals   sample
## attitude mean     64.67742 61.72951
```

From the table we can see that:

- Fashionistas have an attitude towards natural products that is line with the whole sample.
- Skeptically Indifferents have a lower attitude compared to the whole sample.
- Environmentalists have a higher attitude with relation to the whole sample.
- The same holds for Traditionals but to a lesser extent.

Intensity of purchase

```
# Creating binary variable for frequency of purchase
# 0 moderate buyer
# 1 intensive buyer
wtb <- rep(0,122)
wtb[which(beaut[, 3] == 3)] <- 1
```

```
res_wtb<-rbind(round(prop.table(table(memb, wtb ),1)*100,1),
                  round(prop.table(table(wtb))*100,1))

rownames(res_wtb)<-clusters
colnames(res_wtb)<-c("moderate","frequent")
res_wtb
```

```
##                          moderate frequent
## Fashionistas                 76.5     23.5
## Skeptically Indifferent      75.8     24.2
## Environmentalist             54.2     45.8
## Traditionals                 48.4     51.6
## sample                       64.8     35.2
```

From the table we can see that:

- Fashionistas have a larger percentage of moderate personal care products buyers with respect to the whole sample and a smaller one of frequent buyers. This was an unexpected result.

- The same holds for skeptically indifferents, but this was expected.
- Environmentalists present a lower percentage of moderate buyers with respect to the whole sample and a higher one for frequent buyers, as expected.
- Traditionals show a lower percent of moderate buyers compared to the whole sample, and a lower one for frequent buyers.

Verify significance of relationships

```
# Categorical variables: intensity of purchase
res2 <-list(table(memb,wtb))
names(res2) <- c("Frequency of Purchase")
lapply(res2,summary)
```

```
## $`Frequency of Purchase`
## Number of cases in table: 122
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 8.613, df = 3, p-value = 0.03491
```

The p-value is close enough to zero to affirm that a respondent's membership to one or the other cluster is dependent on his/hers frequency of purchase.

```
# Numerical variables: attitudes
t.test(attitudes[memb==1], attitudes[memb==2], alternative = "greater") #ok p-value = 0.01 (5%)
```

```
##
##  Welch Two Sample t-test
##
## data:  attitudes[memb == 1] and attitudes[memb == 2]
## t = 2.3583, df = 59.735, p-value = 0.01082
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.593844      Inf
## sample estimates:
## mean of x mean of y
##  61.58824  56.12121
```

```
t.test(attitudes[memb==1], attitudes[memb==3], alternative = "less") #limit p-value = 0.09 (10%)
```

```
##
##  Welch Two Sample t-test
##
## data:  attitudes[memb == 1] and attitudes[memb == 3]
## t = -1.3142, df = 44.408, p-value = 0.09777
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 1.181329
## sample estimates:
## mean of x mean of y
##  61.58824  65.83333
```

```r
t.test(attitudes[memb==1], attitudes[memb==4], alternative = "less") #no p-value = 0.13
```

```
##
##  Welch Two Sample t-test
##
## data:  attitudes[memb == 1] and attitudes[memb == 4]
## t = -1.1191, df = 62.073, p-value = 0.1337
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 1.519926
## sample estimates:
## mean of x mean of y
##  61.58824  64.67742
```

```r
t.test(attitudes[memb==2], attitudes[memb==3], alternative = "less") #ok p-value = 0.001 (1%)
```

```
##
##  Welch Two Sample t-test
##
## data:  attitudes[memb == 2] and attitudes[memb == 3]
## t = -3.2821, df = 35.144, p-value = 0.001167
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -4.713073
## sample estimates:
## mean of x mean of y
##  56.12121  65.83333
```

```r
t.test(attitudes[memb==2], attitudes[memb==4], alternative = "less") #ok p-value = 0.0004 (1%)
```

```
##
##  Welch Two Sample t-test
##
## data:  attitudes[memb == 2] and attitudes[memb == 4]
## t = -3.5103, df = 53.063, p-value = 0.000461
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -4.475661
## sample estimates:
## mean of x mean of y
##  56.12121  64.67742
```

```r
t.test(attitudes[memb==3], attitudes[memb==4], alternative = "greater") #no p-value = 0.36
```

```
##
##  Welch Two Sample t-test
##
## data:  attitudes[memb == 3] and attitudes[memb == 4]
## t = 0.3485, df = 45.96, p-value = 0.3645
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
```

```
##  -4.412056       Inf
## sample estimates:
## mean of x mean of y
##  65.83333  64.67742
```

The test on the difference of the index values among different clusters was successful only for 3 pairs out of 6 (while one is at the 10% limit). Therefore, we can say that membership to a certain cluster does not statistically depend on respondents' attitudes towards natural products. However, since we obtained some positive results, the attitudes may still have a certain influence on a respondent's membership to one or the other cluster.

THE PROFILES

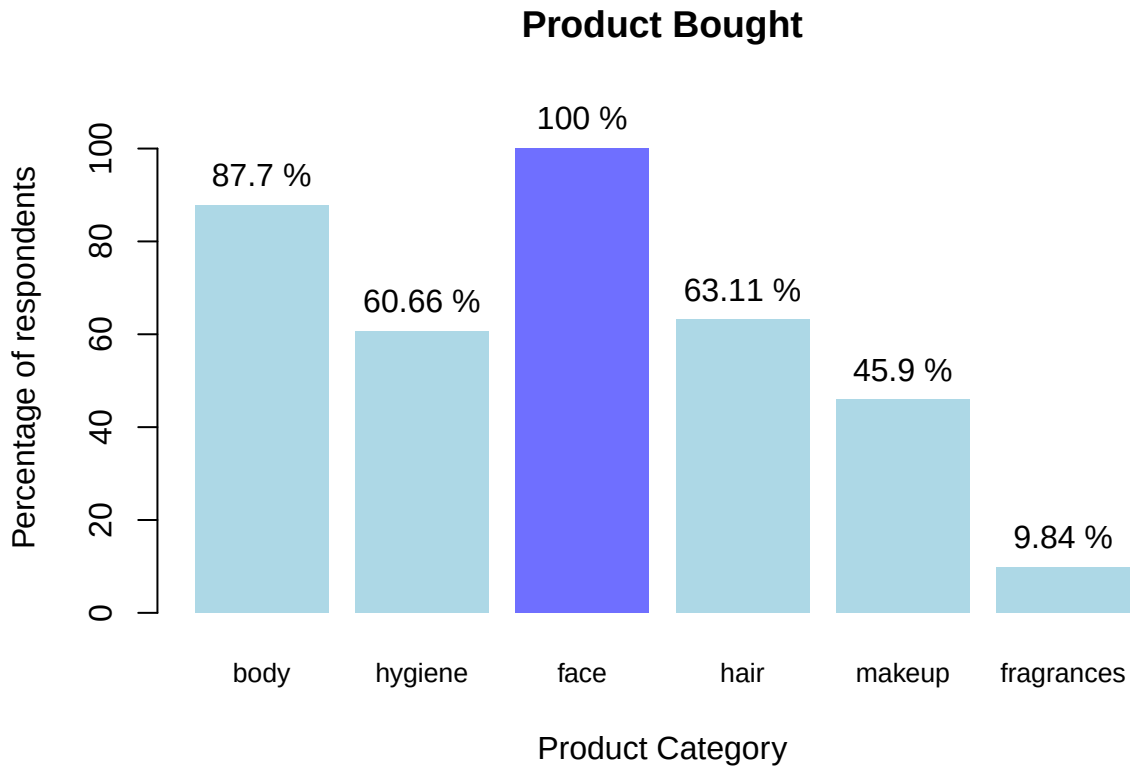
- *FASHIONISTAS*: more females, more people within the ages of 26-35 (millennials), most employed and very few unemployed, more moderate buyers. The rest of the variables are in line with the whole sample. Fashionistas are principally females who are interested in beauty and trends. Moreover, they are mostly employed but moderate buyers. They could be a segment on which to focus since, given their interests, they could potentially become frequent buyers and be interested in effective natural products.
- *SKEPTICALLY INDIFFERENT*: more educated, more students, their age group focuses on the extremes (Gen Z and Boomers), their attitude towards natural products is lower, more moderate buyers. This cluster is not promising in terms of targeting for natural products, especially the older part of it. It may not be worth it to try and change their minds.
- *ENVIRONMENTALISTS*: more unemployed, more people aged 36-45, better attitude towards natural products, more frequent buyers. This is certainly a segment to concentrate on since their habits make them possible consumers of natural products. However, they may need a more specific approach on pricing and communication channels, considering there are many unemployed and their age.
- *TRADITIONAL*: less educated, less students, more people aged 46-55, slightly higher attitude index, more frequent buyers. At the moment, this cluster may not be systematically buying natural products, but they may be interested in doing so for personal care since they are still frequent buyers in that market. Their attitude may be improved with campaigns designed to build awareness and/or underlining quality and effectiveness, so that they are encouraged to be more conscious about their choices not only in terms of reading labels but in terms of repercussions on the environment.

QUESTION 4: Can we predict frequency of purchase (proxy for willingness to buy)? What has an impact on the frequency of purchase (distribution channels, retrieving information)?

We would like to use logistic regression to test if the distribution channels customers choose and the way they gather information before buying a face care product have an impact on the frequency of purchase. However, this information is available only for the face care segment since out survey was centered mainly on face care products. From out explorative research we noted that face care is a particularly popular category of cosmetics in general and of natural cosmetics. To understand if this is true and if we can use the information regarding face care products as an indication of customer's behavior towards cosmetics in general, we investigate what are the product categories that people have bought in the last three months. This also underlines what the most popular segment would be among the cosmetics ones.

```
column_names = list()
values = list()
colors = vector()
for(i in 4:9){
  yes = sum(beaut[,i]==1)
```

```
  no = sum(beaut[,i]==0)
  bought = round( (yes/(yes+no))*100, 2)
  column_names = c(column_names,colnames(beaut)[i])
  values = c(values, bought)
  if(bought == 100){colors = c(colors, "#0000FF90")}else{colors = c(colors, "lightblue")}
}

bp <- barplot(unlist(values), ylim=c(0,110), names=column_names, cex.names=0.85, col=colors,
              xlab = "Product Category", ylab = "Percentage of respondents", border=NA, main="Product Bo
text(bp,values,labels=paste(values, "%"), pos=3)
```



As we can see from the graph, all respondents bought at least one face care product in the last three months. This suggests that it may be safe to use the information regarding face care products since it is the most popular category, and everyone has bought from it at least once.

LOGISTIC REGRESSION

As a proxy for willingness to buy we will use question 2.1 (V4) indicating the frequency of purchase in the last three months since all respondents stated they had bought at least one product for personal care in the last three months. In the dataset, the frequency of purchase variable has three categories and we need to reduce them to two.

Dependent variable: Y = *willingness to buy (frequency of purchase)*

```
# 0 moderate buyer
# 1 intensive buyer
```

```r
wtb <- rep(0,122)
wtb[which(beaut[, 3] == 3)] <- 1
```

1) MODEL 1 *Regressors*: x = *distribution channels* of choice for face care products (V29-V38).

```r
# Building the training set
set.seed(1234) # we set the seed for random sampling from the whole sample
index.tr<-sample(c(1:122), 90) # the training size of 90

train.index <- sort(index.tr)
test.index <- seq(1:122)[-train.index]
```

```r
log.reg.shop <- glm(wtb[index.tr] ~ beaut[index.tr,28] + beaut[index.tr,29] + beaut[index.tr,30] + beau

summary(log.reg.shop)
```

```
##
## Call:
## glm(formula = wtb[index.tr] ~ beaut[index.tr, 28] + beaut[index.tr,
##     29] + beaut[index.tr, 30] + beaut[index.tr, 31] + beaut[index.tr,
##     32] + beaut[index.tr, 33] + beaut[index.tr, 34] + beaut[index.tr,
##     35] + beaut[index.tr, 36] + beaut[index.tr, 37], family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8868  -0.6930  -0.2235   0.6943   2.3613
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.466114   1.370799  -1.799  0.07201 .
## beaut[index.tr, 28]  0.037708   0.128594   0.293  0.76934
## beaut[index.tr, 29] -0.144124   0.141171  -1.021  0.30729
## beaut[index.tr, 30]  0.223041   0.208929   1.068  0.28572
## beaut[index.tr, 31] -0.291966   0.137745  -2.120  0.03404 *
## beaut[index.tr, 32] -0.006579   0.105579  -0.062  0.95031
## beaut[index.tr, 33]  0.017973   0.098496   0.182  0.85521
## beaut[index.tr, 34] -0.126159   0.122673  -1.028  0.30376
## beaut[index.tr, 35]  0.302420   0.104858   2.884  0.00393 **
## beaut[index.tr, 36]  0.008507   0.109180   0.078  0.93789
## beaut[index.tr, 37]  0.222610   0.118935   1.872  0.06125 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 118.288  on 89  degrees of freedom
## Residual deviance:  76.115  on 79  degrees of freedom
## AIC: 98.115
##
## Number of Fisher Scoring iterations: 5
```

```r
round(exp(log.reg.shop$coefficients), 3)
```

```
##         (Intercept) beaut[index.tr, 28] beaut[index.tr, 29] beaut[index.tr, 30]
##               0.085               1.038               0.866               1.250
## beaut[index.tr, 31] beaut[index.tr, 32] beaut[index.tr, 33] beaut[index.tr, 34]
##               0.747               0.993               1.018               0.881
## beaut[index.tr, 35] beaut[index.tr, 36] beaut[index.tr, 37]
##               1.353               1.009               1.249
```

- *V[31] = Pharmacy*: With a higher intensity of shopping for face care products in a pharmacy, the odds of being an intensive buyer are reduced by 24.8%.
- *V[35] = Online shops*: With a higher intensity of shopping for face care products in online shops, the odds of being an intensive buyer are increased by 32.5%.
- *V[37] = Organic product shops*: With a higher intensity of shopping for face care products in organic product shops, the odds of being an intensive buyer are increased by 26%.

Goodness of fit: understand the extent to which we can trust this model. *Tests*

```r
#pR2(log.reg.shop) # Mcfadden corresponds to pseudo R^2

nullmodel<-glm(wtb ~ 1, family = binomial(link="logit"))

# chisquared statistic applied to the comparison between null deviance and deviance of residuals
lt1 <- round(with(log.reg.shop, pchisq(null.deviance-deviance, df.null - df.residual, lower.tail = FALSE
print(paste("Likelihood test p-value:", lt1))
```

```
## [1] "Likelihood test p-value: 1e-05"
```

```r
# pseudo R sqaured
print(paste('pseudo R-squared:', round(1-logLik(log.reg.shop)/logLik(nullmodel),3)))
```

```
## [1] "pseudo R-squared: 0.519"
```

The p-value is low, so the statistic is actually significantly different from zero, meaning that the choice of distribution channel captures the frequency of purchase of personal care products. However, the R^2 is not close to 1.

Confusion Matrix

```r
# Retrieving test set
test.shop <- data.frame(beaut[test.index,28:37])
dim(test.shop); head(test.shop)
```

```
## [1] 32 10
```

```
##    supermkt profumery beauty.shop pharma spice.shop flagship.store
## 12        3         7          10      7          7              5
## 14        1         5           1      6          6              5
## 17        1         6           1      7          8              6
```

```
## 20        1        10           6     10         10              1
## 26        1         6           1      8          8              1
## 31        7         4           1      8          9              1
##    multibrand.store Online doortodoor org.shops
## 12                2      2          2         5
## 14                1      1          1         6
## 17                1      1          1         7
## 20                5      1          1         4
## 26                1      1          1        10
## 31                4      1          1         9
```

```
# Applying estimated coefficients to the whole test set
predicted1.shop <- exp(log.reg.shop$coefficients[1] + log.reg.shop$coefficients[2:11]%*%t(test.shop))/(

# Assign 1(intensive)/0(moderate) values to the predicted probabilities
predicted.shop <- ifelse(predicted1.shop > 0.5, 1, 0)

# Confusion matrix (absolute)
table(wtb[test.index], predicted.shop)
```

```
##    predicted.shop
##      0  1
##   0 19  3
##   1  5  5
```

```
#Confusion matrix percentages of correct and bad classification
round(prop.table(table(wtb[test.index],predicted.shop),1),2)
```

```
##    predicted.shop
##        0    1
##   0 0.86 0.14
##   1 0.50 0.50
```

The confusion matrix supports the validity of the model for prediction purposes and therefore the dependency of frequency of purchase on choice of distribution channel (although R^2 was low).

Therefore, a respondent's frequency of purchase depends on the distribution channels he/she choose.

Accuracy

```
miscl.shop <- mean(predicted.shop != wtb[test.index])
print(paste('Misclassification error:',miscl.shop))
```

```
## [1] "Misclassification error: 0.25"
```

```
print(paste('Accuracy:',round(1-miscl.shop, 2)))
```

```
## [1] "Accuracy: 0.75"
```

2) MODEL 2 *Regressors*: x = *channels trusts for information* before buying face care products (V39-V46).

```
log.reg.info <- glm(wtb[index.tr] ~ beaut[index.tr,38] + beaut[index.tr,39] + beaut[index.tr,40] + beau
summary(log.reg.info)
```

```
##
## Call:
## glm(formula = wtb[index.tr] ~ beaut[index.tr, 38] + beaut[index.tr,
##     39] + beaut[index.tr, 40] + beaut[index.tr, 41] + beaut[index.tr,
##     42] + beaut[index.tr, 43] + beaut[index.tr, 44] + beaut[index.tr,
##     45], family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9018  -0.7385  -0.4993   0.8892   2.4460
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.80732    1.33165  -0.606  0.54435
## beaut[index.tr, 38] -0.03645    0.09424  -0.387  0.69893
## beaut[index.tr, 39] -0.08986    0.10116  -0.888  0.37437
## beaut[index.tr, 40] -0.31139    0.10680  -2.916  0.00355 **
## beaut[index.tr, 41]  0.09788    0.09760   1.003  0.31589
## beaut[index.tr, 42] -0.09843    0.10493  -0.938  0.34821
## beaut[index.tr, 43]  0.11820    0.10737   1.101  0.27095
## beaut[index.tr, 44]  0.23435    0.15333   1.528  0.12643
## beaut[index.tr, 45] -0.09810    0.18369  -0.534  0.59330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 118.288  on 89  degrees of freedom
## Residual deviance:  93.743  on 81  degrees of freedom
## AIC: 111.74
##
## Number of Fisher Scoring iterations: 4
```

```
round(exp(log.reg.info$coefficients), 3)
```

```
##         (Intercept) beaut[index.tr, 38] beaut[index.tr, 39] beaut[index.tr, 40]
##               0.446               0.964               0.914               0.732
## beaut[index.tr, 41] beaut[index.tr, 42] beaut[index.tr, 43] beaut[index.tr, 44]
##               1.103               0.906               1.125               1.264
## beaut[index.tr, 45]
##               0.907
```

- *[V39] Pharmacist/Dermatologist*: With a higher intensity of getting advice from a pharmacist/dermatologist, the odds of being an intensive buyer are reduced by 18.9%.
- *[V40] Friends*: With a higher intensity of getting advice from a friend, the odds of being an intensive buyer are reduced by 27.2%.
- *[V41] Blogs/Social Networks*: With a higher intensity of getting advice from blogs/social networks, the odds of being an intensive buyer are increased by 19.4%.

- *[V44] List of ingredients*: With a higher intensity of reading the list of ingredients, the odds of being an intensive buyer are increased by 30.9%.

Goodness of fit

Tests

```
#pR2(log.reg.info)

lt2 <- round(with(log.reg.info, pchisq(null.deviance-deviance, df.null - df.residual, lower.tail = FALSE
print(paste("Likelihood test p-value:", lt2))
```

```
## [1] "Likelihood test p-value: 0.00186"
```

```
# pseudo R squared
print(paste('pseudo R-squared:', round(1-logLik(log.reg.info)/logLik(nullmodel),3)))
```

```
## [1] "pseudo R-squared: 0.408"
```

The p-value is close to zero but the pseudo R^2 is a bit low.

Confusion Matrix

```
test.info <- data.frame(beaut[test.index,38:45])
dim(test.info); head(test.info)
```

```
## [1] 32  8
```

```
##    advice.salespeople advice.pharma askfriends socialntw producer.website
## 12                  8             1          1         1                1
## 14                  4             4          1         7               10
## 17                  6             1          1         8                8
## 20                 10             1          7         1                1
## 26                  1             8          7        10                1
## 31                  9             5          6         1                1
##    comparison.brands ingredients label
## 12                 1           1     1
## 14                 7          10     7
## 17                 1           1     7
## 20                 1           1     1
## 26                10          10     7
## 31                 7           8     9
```

```
predicted1.info <- exp(log.reg.info$coefficients[1] + log.reg.info$coefficients[2:9]%*%t(test.info))/ (

predicted.info <- ifelse(predicted1.info > 0.5, 1, 0)

table(wtb[test.index], predicted.info)
```

```
##    predicted.info
##      0  1
##   0 16  6
##   1  3  7
```

```
round(prop.table(table(wtb[test.index], predicted.info), 1), 2)
```

```
##    predicted.info
##        0    1
##   0 0.73 0.27
##   1 0.30 0.70
```

The model is acceptable, only one out of the two categories is labeled at random. Even though it is not too bad, this model would probably be discarded for prediction. However, we can still affirm that the way people gather information before buying face care products influences the frequency of purchase of personal care products.

Accuracy

```
miscl.info <- mean(predicted.info!= wtb[test.index])
print(paste('Miscalssification error:', miscl.info))
```

```
## [1] "Miscalssification error: 0.28125"
```

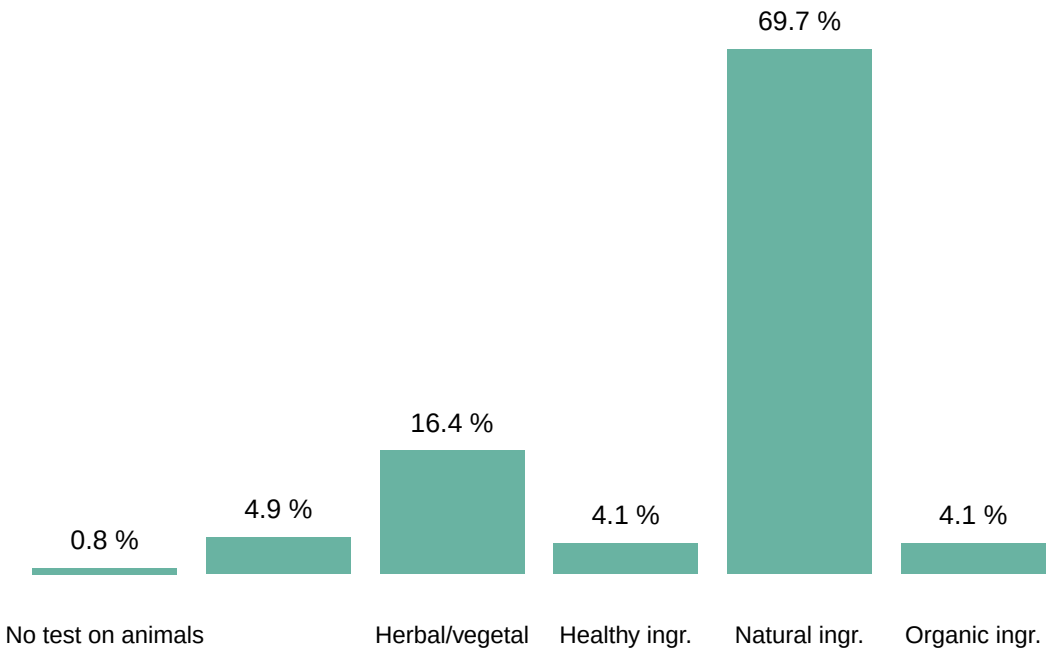
```
print(paste('Accuracy:', round(1-miscl.info, 2)))
```

```
## [1] "Accuracy: 0.72"
```

QUESTION 5: What are the most popular definitions of natural products in the eyes of respondents?

We want to understand how customers define natural products since our explorative research underlined the fact that definitions vary a lot. Once we know how people generally define natural products, this can serve as an indication about how to market natural products in order to highlight specific product characteristics.

```
fre <- table(beaut[,1])
fre.p <- round(fre/dim(beaut)[1]*100,1)

par(mar = c(2, 2, 0, 2))
bp <- barplot(fre.p, ylim=c(0,max(fre.p)*(1+0.5)), las=1, cex.names=0.71, names.arg=c("No test on anima
text(bp, fre.p, labels=paste(fre.p, "%"), cex = 0.8, pos = 3)
```



The most common definition of natural products in related to the fact that they contain natural ingredients, so this aspect should be stressed in a potential new product but in the labels and in its marketing. Other potentially beneficial characteristics to underline are the fact that preparations are herbal or vegetal based and that the products are non-polluting.