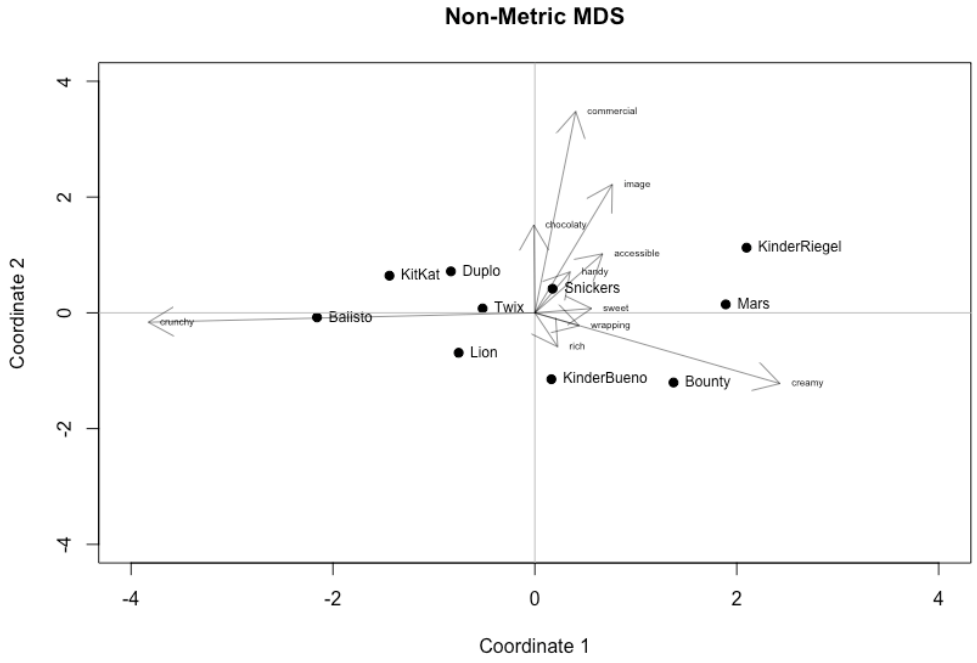
# Special Work Performance 2

Group 15 (Giorgi Modebadze (602191) – Radoslav Evtimov (570341) – Ecenaz Bal (429775))

The purpose of this report is to analyze survey data of 50 persons about 10 different chocolate brands in the German market. Each brand is assessed with 13 different attributes evaluated on a scale 1 to 5(1-lowest,5-highest) We used Kruskal's Non-metric Multidimensional Scaling and Principal Component Analysis to make the data more interpretable. For the purposes of data analysis, the missing values were replaced with the mean value for the given product.

SWP2a:

For Multidimensional Scaling we chose non-metric scaling method, which attempts to represent, as closely as possible, the pairwise dissimilarity between objects in a low-dimensional space. The reason for using non-metric scaling is that the survey data is non-metric, they do not possess a meter with which distance between scale values can be measured, meaning for that data only ranking matters and not actual differences between points. On the way to finding NMDS first the Euclidean distance is calculated using *daisy* formula from cluster library. Based on the result of Euclidean distance a two-dimensional perceptual map is created using *metaMDS* from vegan library. The formula is more robust than *isoMDS*, it also performs Nonmetric Multidimensional Scaling (NMDS), but unlike isoMDS it provides infrastructure to do the random starts and comparison of configurations for convergence. The comparison between these two methods on two-dimensional space gave drastic difference in terms of stress values. For *isoMDS* the stress value is 3.99, but as the rule of thumb it is never a good idea to use solutions with stress value more than 0.2 and a stress value approaching 0.3 indicates that the ordination is arbitrary. With every additional dimension, isoMDS also decreases stress value, but addition dimensionality makes interpretations more challenging. On the other hand, *metaMDS* using centering scale and PC rotation, after twenty random start found two convergent solutions with stress value of 0.04.



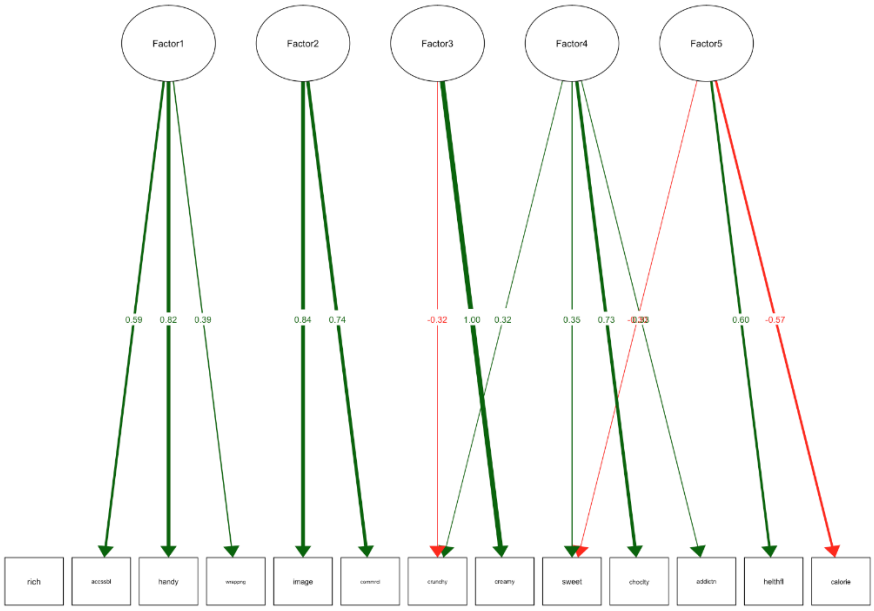
After creating two-dimensional perceptual map from the points *metaMDS* provided, the relative positions of each brand to each other can be observed. But it only provides us with information about similarity and dissimilarity between different brands and does not show which attributes contributed to this similarity. To get this information we used property fitting, which is a method of testing hypothesis about attributes that influence people`s judgement of the similarities among a set of chocolate bars. It tries to represent each attribute in the perceptual map as a vector using regression or correlation analysis. For our survey we used regression analysis. We are observing the problem as a classical regression model, where each attribute is the dependent variable and the two axis of the perceptual map are the independent variables.

First, we removed three attributes (healthy, addictive, calories) to make the plot more comprehensible. As they were similar for all of the brands and very close to the origin point, they did not play a significant role in the diferentiation process. Two main attributes, respondents differentiate chocolate bars with, are crunchiness and creaminess. This can be easily observed from the vectors which are describing them – the map shows that the brands are narrow on the y-axis and wide-spread on the x-axis. Crunchiest brands are Balisto and KitKat, while creamiest ones are Bounty and Mars. Kinder Riegel is also creamy, but creaminess is not the only quality it is perceived by people. But it is also more commercial than other brands in this category. Snickers and Twix are placed in the middle of the plot, meaning people don't differentiate them very much by any category. So this can be considered as a good sign for the brands, which are seeking stable position on the market. To evaluate the performance of the property fitting we can take a look at R-squared values of each attribute and as they are very close to 0 we can conclude that our model is not performing very well.

SWP2b:

Yet another way to decrease dimensionality and make results more readable is Factor Analysis or Principal Component Analysis. For our survey results PCA was chosen over FA, due to several reasons. One of the main reason behind is that the survey was not constructed to test any theoretical model of latent variables. Despite, FA still was run over the data, to prove our point empirically. With Factor Analysis we transformed our current variables into an equal number of variables, which combines current ones. Then we used eigenvalues and eigenvectos to make this transformation. In this step factors are sorted by decreasing order of the variances they explain. So the first factor explains most variance and second less then first and so on. As the analysis of eigenvalues of the data showed 5 factors are used to replicate the original data. This is due to the fact that, these 5 factors have values above 1 which means each of them can explain more variance than an original variable. Although using so many factors, it only explains 46% of the cumulative variance which is not acceptable. For this analysis we also used rotations other than orthogonal and varimax, but the performance of model was not significantly improved.

To exclude the possibility of latent variable and make it visually understandable we used function *semPaths*, which plots the path graph.




As this graph shows, there cannot be a significant logical link between the variables inside each of the factors. This proves there are no latent variables that should be taken into consideration.

Principal Component Analysis

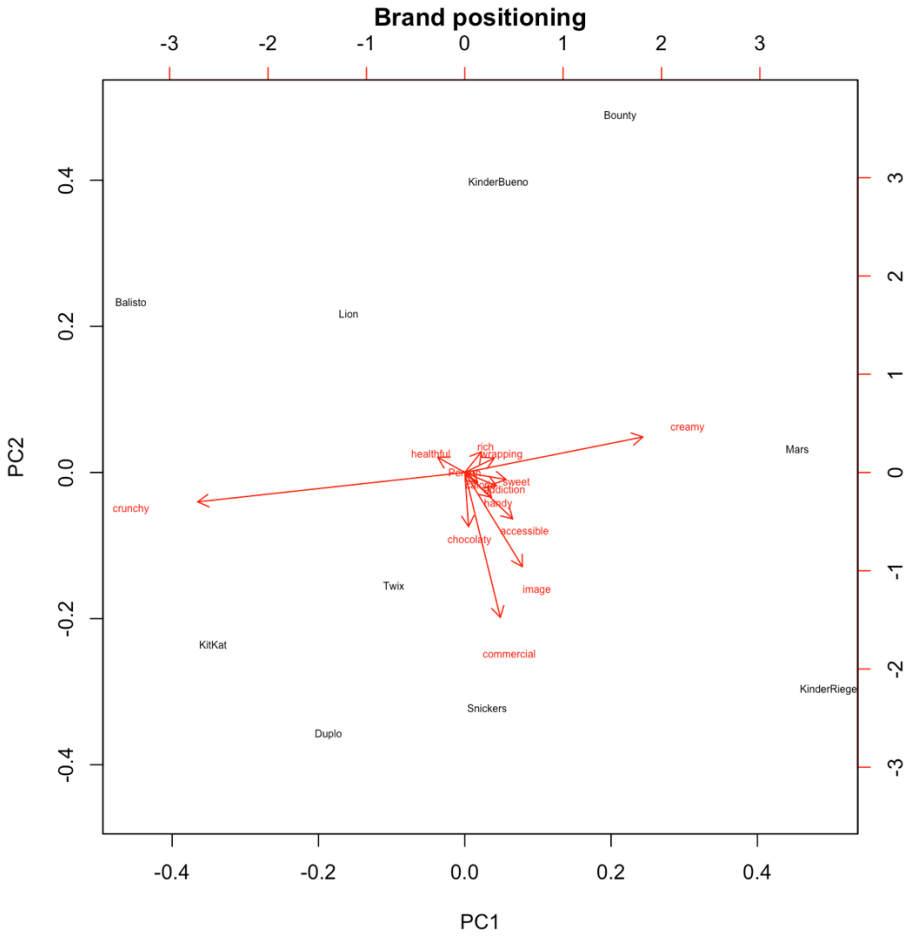
An alternative to Factor Analysis is Principal Component Analysis. PCA unlike Factor Analysis does not assume the possibility of latent variables. It is used to reduce a large set of variables as a dimension-reduction tool. This mathematical procedure transforms a number of variables into *principal components.* The first principal component accounts for as much of the variability in the data as possible as in the case of Factor Analysis.

PCA uses a method called singular value decomposition to extract linear combinations of different variables and explain as much variance as possible with fewer dimensions.

To determine number of possible principal components, plot of the eigenvalues is used. The elbow method showed that two or three components shall be selected replicating the data.

With two components, cumulative variance of 36 percent was explained, out of which first component explained 62 percent. When we moved to three components result improved only by 10 percent, making it to 46% in total, but it is also harder to interpret the data in the three dimensions.

Based on results of two-dimensional PCA, an MDS-like map can be built which also tries to explain the data visually.




The biplot delivers similar results about the perception of the participants for the different brands. Only using two dimensions it gives a basis to interpret the data easily. It is proven once more that the brands the respondents were asked about, differ mostly on the “crunchy-creamy” axis. Those are two opposite qualities and they are responsible for most of the dissimilarities between the chocolate bars.

SWP2c

PCA and FA are methods for data reduction while MDS is a class of analysis. PCA with a certain number of principal components that could be plotted is a particular case of MDS.

We generally recommend PCA as a more informative procedure than MDS for typical metric or near-metric (e.g., survey Likert scale) data. However, PCA will not work with non-metric data. In those cases, MDS is a valuable alternative.

MDS may be of particular interest when handling text data such as consumers’ feed- back, comments, and online product reviews, where text frequencies can be con- verted to distance scores. For example, if you are interested in similarities between brands in online reviews, you could count how many times various pairs of brands occur together in consumers’ postings. The co-occurrence matrix of counts—brand A mentioned with brand B, with brand C, and so forth—could be used as a measure of similarity between the two brands and serve as the distance metric in MDS.

For our particular case both non-metric MDS and Principal component analysis emphasized kind of similar tendencies, that respondents mainly differentiate chocolates on the scale of crunchy-creamy. But there was significant different between representing other attributes like commercial and image. When NMDS proposed Snickers as the overall medium brand, PCA showed that users perceive it as very commercial one.

Using main metrics to evaluate models, both NMDS and PCA showed below average performance, that can be explained by poor quality of the underlaying data and the fact that the difference between the question makes it hard to rescale data on lower dimension

Appendix:

Similarity Matrix:

| * | Balisto | Bounty | Duplo | KinderBueno | KinderRiegel | KitKat | Lion | Mars | Snickers |
|---|---|---|---|---|---|---|---|---|---|
| Bounty | 3.06 | | | | | | | | |
| Duplo | 2.02 | 2.87 | | | | | | | |
| KinderBueno | 2.96 | 2.31 | 2.59 | | | | | | |
| KinderRiegel | 4.42 | 2.62 | 3.05 | 3.03 | | | | | |
| KitKat | 1.57 | 3.11 | 0.91 | 2.74 | 3.75 | | | | |
| Lion | 1.95 | 2.42 | 1.80 | 1.37 | 3.35 | 1.74 | | | |
| Mars | 4.09 | 1.92 | 3.06 | 2.55 | 1.63 | 3.61 | 2.88 | | |
| Snickers | 2.83 | 2.58 | 1.42 | 2.25 | 2.50 | 1.89 | 1.79 | 2.10 | |
| Twix | 2.33 | 2.59 | 1.22 | 1.78 | 2.90 | 1.43 | 1.19 | 2.61 | 0.96 |

Factor analysis Eigen Values Test: