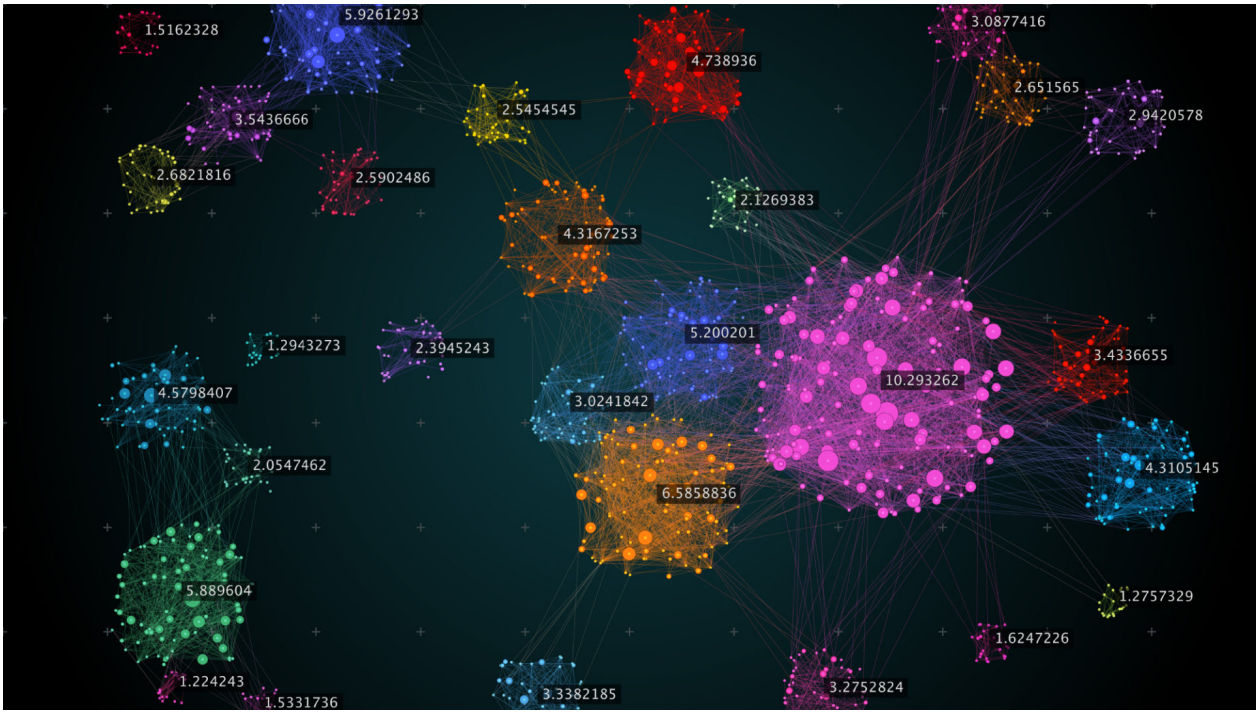
Project - Advanced Statistical Methods

Barcelona School of Management



Prepared by: Daniëlle Kotter
Prepared for: Michael Greenacre
Program: Msc in Management, Barcelona School of Management
Date: 2nd July 2021

Introduction

As a final project for the Master of Science in Management at Barcelona School of Management, I prepared a business plan. This entailed developing a business opportunity and performing consumer research to gain a better understanding of the market. The business concept is a sustainable specialty coffee company that has a subscription based model.

For the purpose of the business plan, I conducted a survey and collected a sample of 235 participants. The survey was designed to gain insights on coffee habits and purchasing decisions of coffee consumers. The goal of this particular research was to quantitatively confirm hypothesis and establish general patterns across several contexts. Ultimately, the objective of the primary research was to decrease the risk of the business opportunity, discover unexpected findings and provide opportunity to challenge the market through competitive advantages.

The following report includes a clustering approach to buyer personas, a principal component and a correspondence analysis. After initially evaluating univariate descriptions of the entire sample throughout the business plan, the next step was to perform cluster analysis for the purpose of identifying patterns in attitude and preferences to create segments of consumers for marketing purposes. These segments (or cluster groups) are then profiled to create the “buyer personas”. The most promising buyer persona will then be targeted in the positioning and marketing strategy.

The principal component analysis strives to find the attributes that differentiates the cluster groups. Lastly, the correspondence analysis evaluates the difference between the behaviors or habits within variation of explanatory variables.

Data

The survey covers several sections: basic demographics, socioeconomic, consumption and purchasing behavior and interest in social missions. The sample size was 235 and includes 0 missing values. Moreover, there are 25 mixed-scale variables included in the data set. These vary from continuous numerical, discrete numerical, and categorical variables. The next page displays an overview of all variable names, the question asked to consumers and the scale.

Data set

Field	Description	Scales
AmountWeek	How many cups of coffee do you typically consume weekly?	Ratio, Continuous
Am ountOutMonth	How frequently do you drink out-of-home per month on average?	Ratio, Continuous
MoneyCoffee	How much money on average do you estimate you spend on coffee per month?	Ratio, Continuous
Mo neyGroceries	How much on average do you spend on general groceries per month?	Ratio, Continuous
Machine	How do you brew your coffee at home?	Nominal
Brand change	How often do you switch between coffee brands?	Nominal
Purchase location	Where do you usually purchase your coffee?	Nominal
Supe rmarket_Posi tive_Reasons	When you purchase coffee from the supermarket what are your main reasons for doing so?	Nominal
Supe rmarket_Nega tive_Reasons	What would be reasons why you would not purchase coffee from the supermarket?	Nominal
Criteri a_Type_Cofee	What are your main criteria's or evaluation points for choosing the type of coffee?	Nominal
Kno wledgeCoffee	How would you describe your knowledge level regarding coffee in general?	Ordinal. 0-10, Discrete
Pu rchase_Price	I believe that the ____ is important to my decision on which coffee to purchase.	Ordinal, likert 0-5
Purchase_Su stainability	I believe that the ____ is important to my decision on which coffee to purchase.	Ordinal, likert 0-5
Purchase_Su stainability	I believe that the ____ is important to my decision on which coffee to purchase.	Ordinal, likert 0-5
Purcha se_Fairtrade	I believe that the ____ is important to my decision on which coffee to purchase.	Ordinal, likert 0-5
Purcha se_Packaging	I believe that the ____ is important to my decision on which coffee to purchase.	Ordinal, likert 0-5
Frequen cy_Specialty	How often do you drink specialty coffee?	Ordinal
Subscri ption_Likely	How likely are you to have an online subscription for (specialty) coffee?	Ordinal 0-10, Discrete
Subscriptio n_Not_Likely	What is the number one reasons why you would be hesitant?	Nominal
App_Likely	How likely are you to value and use an app for your online subscription?	Ordinal, 0-10, Discrete
Gender	What is your gender?	Nominal
AgeCategory	What is your age category?	Ordinal
Occupation	What is your occupational status?	Nominal
Education	What level of education have you completed?	Ordinal
Home	How would you describe the place you currently live in?	Nominal

Below the head of the data set is displayed to give an idea regarding choice options within variables.

AmountWeek	AmountOutMonth	MoneyCoffee	MoneyGroceries	Machine	BrandChange
3	5	15	320	Filter machine	Sometimes
21	4	10	125	Espresso machine	Sometimes
22	8	30	350	CupMachine	Sometimes
15	3	50	200	Espresso machine	Sometimes
6	2	9	350	Moka pot	Sometimes
24	0	10	300	Espresso machine	Sometimes

PurchaseLocation	Supermarket_Positive_ Reasons	Supermarket_Negative_ Reasons	Criteria_Type_Coffee
The supermarket	Time-saving	Not wanting to support big cooperations	Price
Specialty stores or cafés	I do not purchase coffee from the supermarket	It is not fresh, Better quality elsewhere	Origin, Flavour profile
The supermarket	Price, Time-saving	No reason	Roast level, Flavour profile
E-commerce	Convenience, Time-saving	No reason, Better quality elsewhere	Arabica or Robusta, Flavour profile
E-commerce	I do not purchase coffee from the supermarket	No reason	Flavour profile
The supermarket	Convenience, Time-saving	Better quality elsewhere	Origin, Roast level

KnowledgeCoffee	Purchase_Price	Purchase_Sustainability	Purchase_Certificate	Purchase_Fairtrade
4	2	5	1	5
7	2	4	1	5
5	3	3	3	3
6	1	1	3	1
8	5	5	5	5
6	1	1	1	1

Purchase_Packaging	Frequency_Specialty	Subscription_Likely	Subscription_Not_Likely
3	Only in cafes	3	The price
3	Always	10	No reason
3	Never	3	I am happy with my coffee now, I do not like being stuck with subscriptions
1	I do (did) not know what this is	1	I am happy with my coffee now
5	Sometimes	1	I do not like being stuck with subscriptions, I am happy with my coffee now
1	I do (did) not know what this is	7	The price

App_Likely	Gender	AgeCategory	Occupation	Education	Home
1	Male	18-25	Student	Bachelor’s degree	Urban (City)
9	Female	18-25	Student	Bachelor’s degree	Urban (City)
2	Male	18-25	Student	Bachelor’s degree	Suburbs
1	Female	> 60	Retired	Master	Urban (City)
1	Male	45-60	Unemployed	Master	Urban (City)
9	Female	25-45	Employed (Full time)	Master	Urban (City)

Methodology

Sample & data collection

A random probability-based sample technique was adopted due to time consideration and availability. The platform used for data collection is google sheets. The sample has been reached through multiple methods, including numerous Facebook groups and personal contacts near the region. The Facebook groups used for the research project were: Utrecht!, with almost 24 thousand members and expats in Utrecht, with over 31 thousand members. This is a form of convenience sampling where any member of the population is invited to participate without a dependent of the presence of the sampling frames.

Approach to data analysis

After the univariate analysis, the technique of k-means clustering is applied. Hereby patterns in the data are identified to find groups of respondents that are similar to one another and yet different from the others. These groups / clusters are used for profiling and thus determining the buyer personas. Subsequently, the buyer personas will be based on computational usage and theory ultimately for segmentation.

Moreover, I want to learn whether there is a variance in attitudes within demographic groups and what factors mostly drive the variance between consumers. Therefore, Correspondence analysis & Principal Component Analysis is additionally included in this analysis. The report follows several steps:

1. Preparing the data. Re-scaling & selecting the variables for analysis
2. Clustering analysis, excluding demographics
3. Individual cluster analysis - profiling
4. Principal Component Analysis
5. Correspondence analysis, bringing in demographics

Naturally this additionally entails visualization and hypothesis testing.

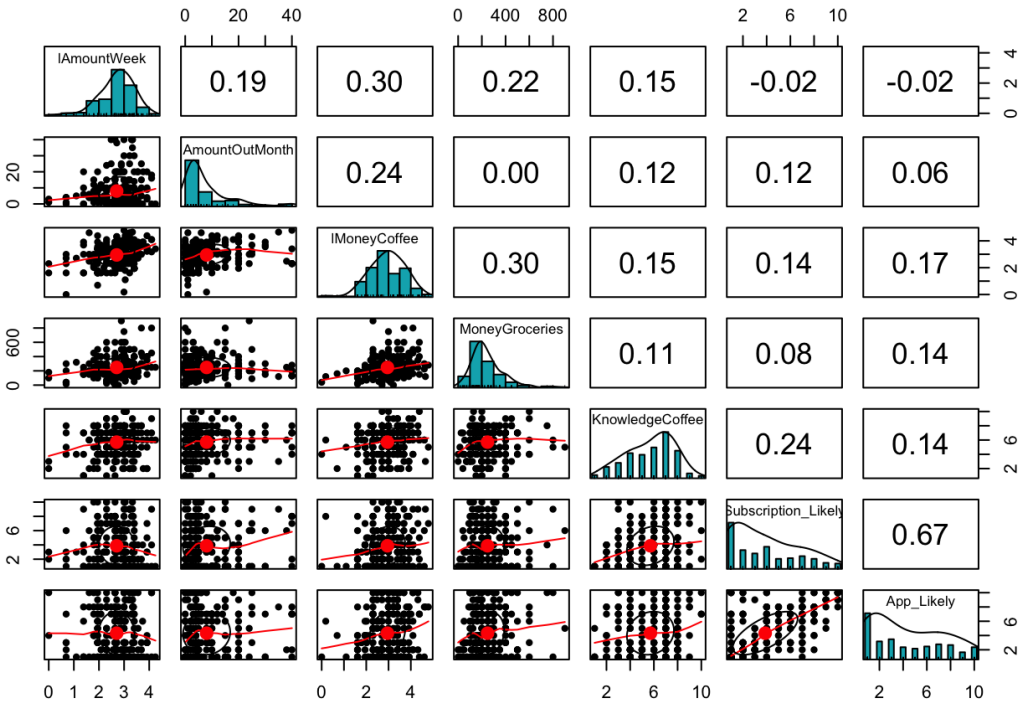
Results

Clustering Analysis

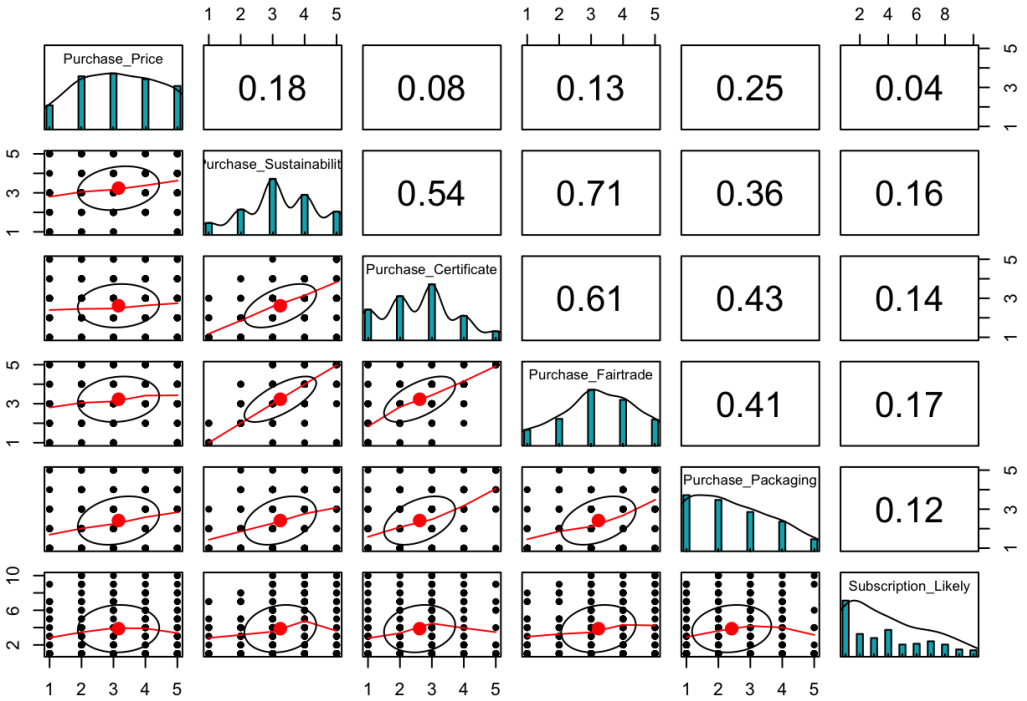
Selecting the variables

The first step to the clustering analysis was selecting the variables that could be valuable to distinguish groups of customers. Initially the pairs panel was used to analyze the numerical variables. The first observation is that the variables subscription likelihood and app likelihood have a strong positive relationship with a correlation coefficient of 0.67. Expectantly when consumers are more likely to set up a subscription they are also more likely to setup an app. This is further confirmed through a chi-square test where the variables are associated with a 1% significance level.

Moreover, the next highest correlation with subscription likelihood is knowledge of coffee. To further analyze this relationship, a chi-square test was performed which showed that there is a statistically significant relationship at a 5% significance level. Suggesting that having higher knowledge of coffee leads to a positive effect on the likelihood of setting up an subscription.



The highest linear correlation between the likeliness to setup an subscription is between the importance of coffee being fair trade and sustainable for purchasing decisions. Nevertheless, there is a weak positive correlation. What is evident is that the sustainability, fair trade and certificate variables are all moderately-highly positively correlated. This suggests that a higher score giving for one of these questions, also gives a higher score to the alternatives.



The demographic variables naturally were excluded from the cluster analysis considering the goal of the buyer personas was to find insights on consumer preferences and not demographic differences.

To ultimately decide which variables to keep for the analysis, all variables primarily were included and then tested on whether there was significant difference between the centroids of the cluster groups. The non-parametric Kruskal-Wallis chi-squared test was hereby applied, ranking variables on whether they are alike. The variables that would show the highest significance difference between cluster groups would be selected for the analysis. Ultimately this leads to the variables selected for clustering to be:

1. Purchasing Location

2. Frequency Specialty coffee consumption

3. Amount brand change

4. Amount consumed per week

5. Money spend on coffee

6. Likeliness to set up and app

7. Likeliness to set up and subscription

8. Purchasing importance - Fair trade

9. Purchasing importance - Certificate

10. Purchasing importance - Sustainability

11. Purchasing importance - Price

12. Purchasing importance - Packaging

Preparing the data

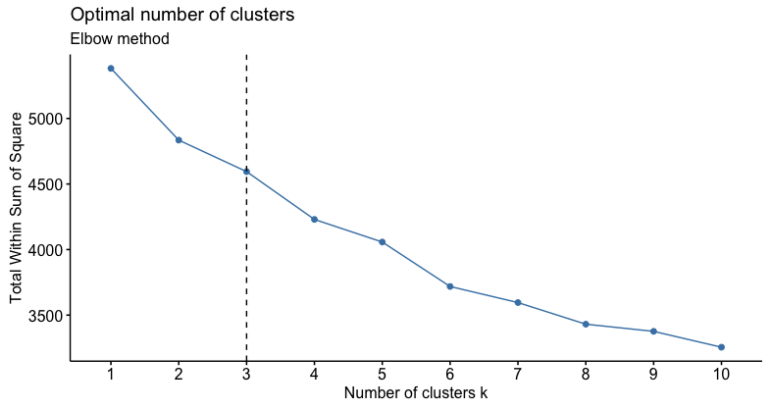
As can be observed, there are many different types of variables included in the data set. Therefore, this requires some scaling and transformation of variables to prepare a range-standardized data set. The categorical variables such as amount brand change are primarily transformed through one-hot encoding. Hereby if an option is selected, the value is taken as 1 = positive and if not selected: 0 = negative. Afterwards, all variables including other numerical variables are standardized between -1 and 1.

Finding optimal number clusters

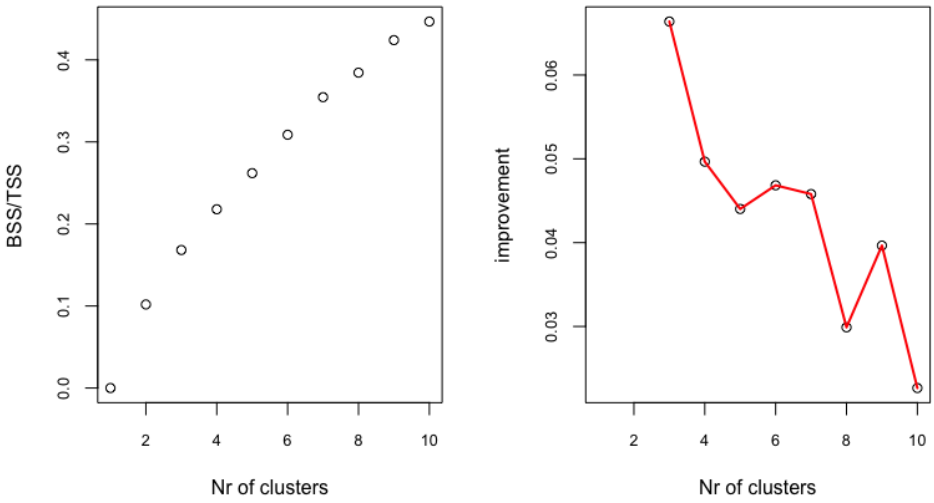
To find the optimal number of clusters for the k-means clustering approach, several aspects are evaluated.

Elbow Method

The elbow method compares the total within sum of square for each k up to 10. Besides the requirement for statistical difference between clusters, there is also a managerial aspect needing to be taken into account. Hereby a small amount of clusters may not provide enough distinquishment for marketing purposes. However, too many clusters implies having to cater to too many different groups. Therefore, based on the following graph and taken the above into account, cutting at cluster 3 seems the most applicable. The jump from 2 to 3 clusters doesn't provide a significant drop in the within sum of squares.

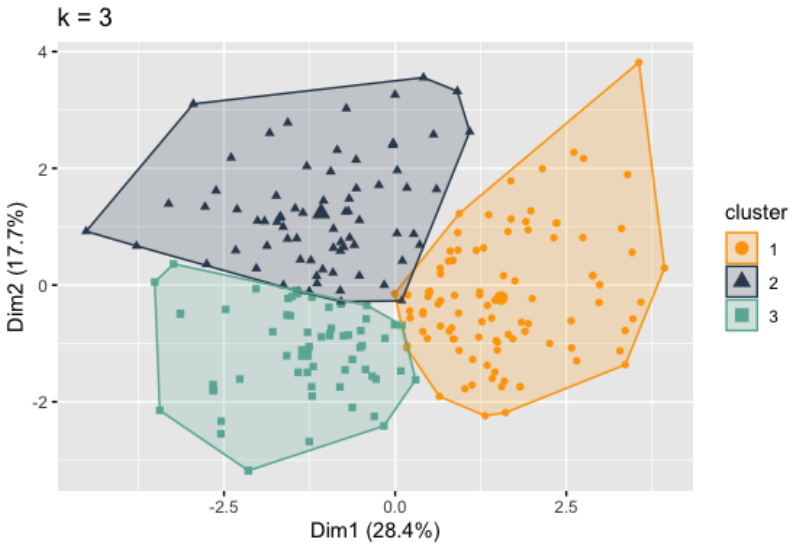


The following graph shows an overview of the between cluster sum of square vs the total sum of square and the improvement to be made. It can be observed that going from 3 to 4 clusters has a steep drop where the improvement goes from 0.07 to 0.05. Considering all factors, it has been decided to use k=3 for the clustering analysis.



Clusters

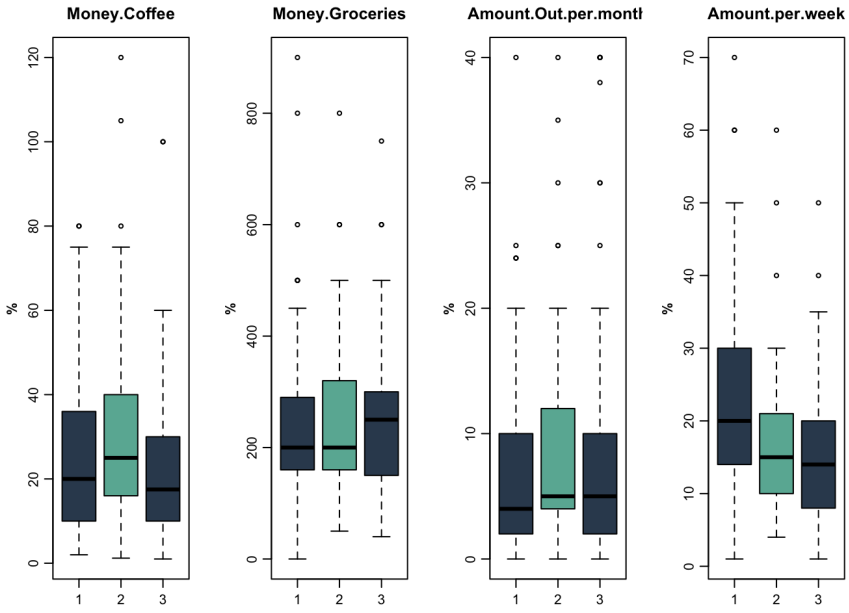
As previously explained, k-means clustering has been applied which ultimately resulted in 3 cluster groups. Below these groups are visualized in a two-dimensional map using principal component analysis to visualize the distance between data points. The optimal cluster amount has been determined through various approaches such as minimizing the total distance sum of square and the improvement on between-cluster variance. Moreover, the variables chosen for the clustering analysis had to show statistically significant variance to distinguish different perceptions between the several buyer personas. It can be observed that the 2-dimensional plot 46,2% of the variance. Nevertheless, little overlap between cluster groups is apparent.



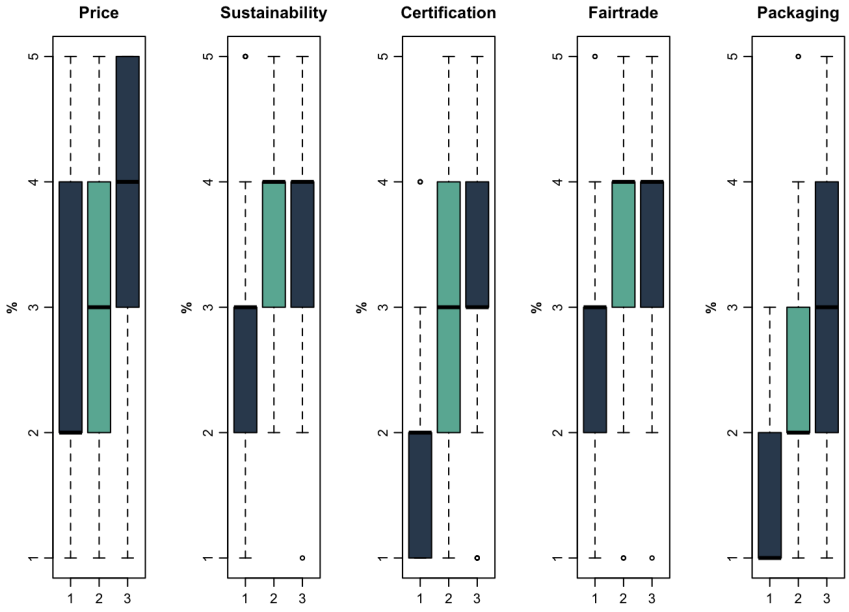
Profiling

Each cluster represents a buyer persona that is built on the insights and identifiers of this group. Therefore, the cluster centroids are analyzed to profile consumers who are alike.

Below the primary continuous variables are visualized through boxplots, separating the cluster groups. The range for any group is relatively large and there are many outliers identified. On average cluster group 2 spends the most amount on money on coffee, groceries and drinks the most coffee out-of-home. The out-of-home consumption is quite similar between groups. However, when the survey was conducted, restaurants and bars were closed. Therefore, it is possible this prediction is smaller than it would otherwise be. Cluster 1 drinks the most amount of coffee per week at home.

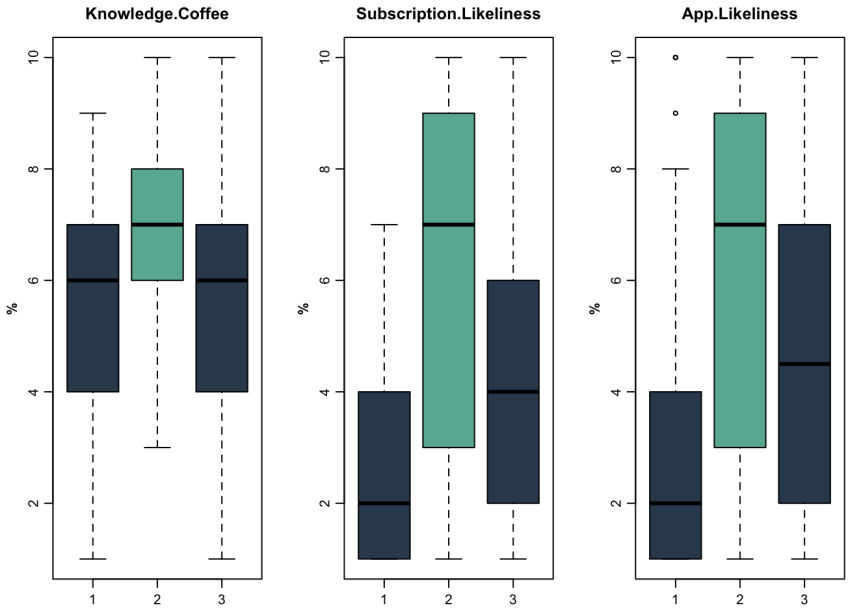


The following boxplot displays the criteria that consumers score importance for purchasing decisions. Cluster 1 on average gave lower scores to all criteria. Specially striking are the certification and packaging. However, cluster groups 2 & 3 gave relatively high scores to the factors sustainability, certification and fairtrade. The only factor that differentiates them between these three criteria, is that the range for certification is much higher. Therefore, cluster 2 has differing opinions regarding the need for certification of coffee. Cluster group 3 additionally gives higher weight to price and packaging.



Based on the buyer personas, a niche audience is selected. These represent those who would most benefit from the business model and will be segmented throughout marketing efforts and strategic decisions. From the three personas found, the coffee lover would be the most applicable to target for multiple reasons, one being that they have the highest likelihood of setting up a subscription. Above the boxplots are displayed for several attributes for each of the three cluster groups. As can be seen, the second cluster is much more likely to set up a subscription and app than the other two clusters.

Moreover, the highest median knowledge regarding coffee is 7, additionally by group 2. This cluster is the smallest cluster group with 54/235 (23%) of the respondents. However, we can observe that they are more likely to set-up an subscription, spend the most on coffee and their knowledge level is high. Subsequently, the niche audience to initially target is the coffee lover. Nevertheless, the limitations of the market research have to be accounted for and a larger scale research project with a true random sampling method could provide deviating results.



Profiling the clusters based on distinctive features, the cluster groups are classified as:

1. The casual drinker. Cluster size: 91.
2. The coffee lover. Cluster size: 54
3. The social buyer. Cluster size: 90.

Below the overview of the buyer persona, “the coffee lover” can be found. To provide a clarifying example, the largest proportion of cluster 2 is between 25-45 years old. Therefore, the buyer persona 2 (the coffee lover) can be identified as a 25–45-year-old. The numerical values represent the median response of each cluster.

The coffee lover

BASIC DEMOGRAPHICS	
AGE	25-45
EDUCATION	Bachelor's degree
OCCUPATION	Full-time employed
HOME DESCRIPTION	City

INSIGHTS

Criteria's for choosing type of coffee:
1. Flavour profile
2. Origin
3. Roast level

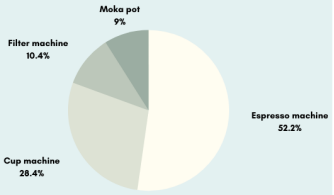
Where do you usually purchase your coffee?
1. Specialty stores or cafés
2. The supermarket
3. E-commerce

Reasons for not buying from the supermarket?
1. Better quality elsewhere
2. It is not fresh
3. Not enough variety

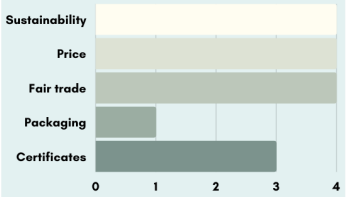
Why would you not set up a subscription?
1. Does not have a good reason
2. I already have a subscription
3. Happy with the coffee now

AMOUNT OF COFFEE CONSUMED PER WEEK	18
AMOUNT OUT-OF-HOUSE PER MONTH	10
AMOUNT SPEND ON COFFEE PER MONTH	31
AMOUNT SPEND ON GROCERIES PER MONTH	263

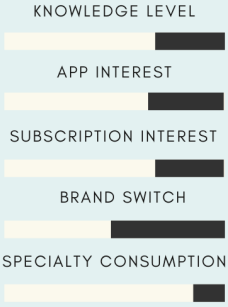
MACHINE



PURCHASING DECISION



INSIGHTS



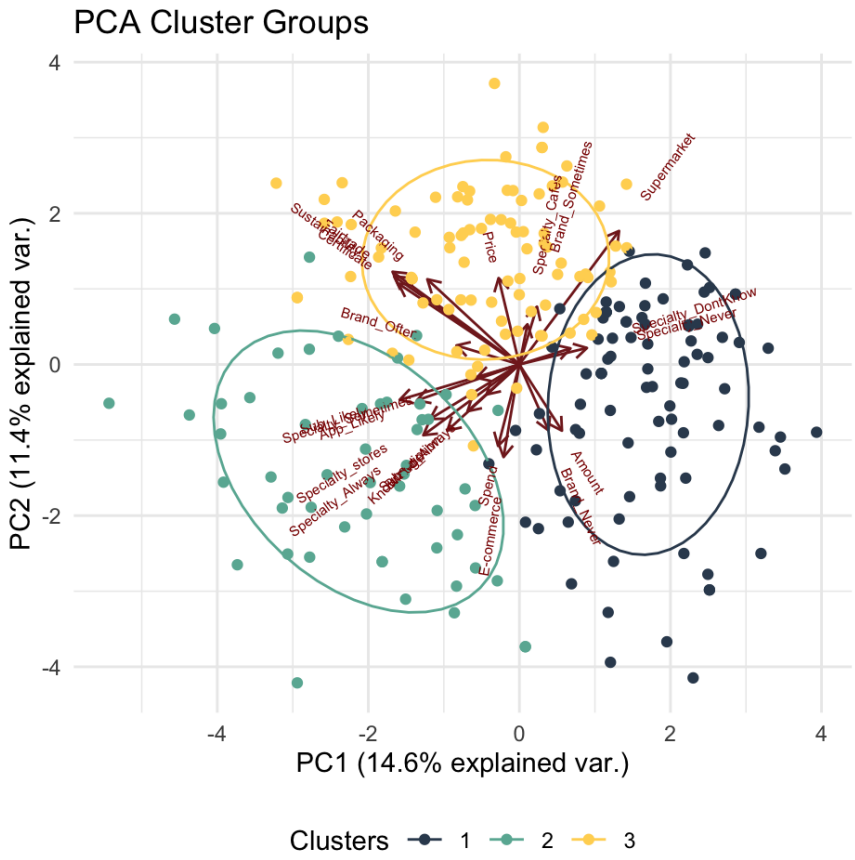
Principal Component Analysis

The principal component analysis displays what attributes moves the participants into different directions. The length of the arrow suggests how far away from the mean the opposing observations are. The two-dimensional plot captures 26% of the variance. The strongest variables that pulls the groups away from one another seem to be buying from the supermarket and the sustainability, fair trade, packaging and certificate criterion. In the next section I summarize the principal components between the clusters based on their average preferences/behaviors from the biplot.

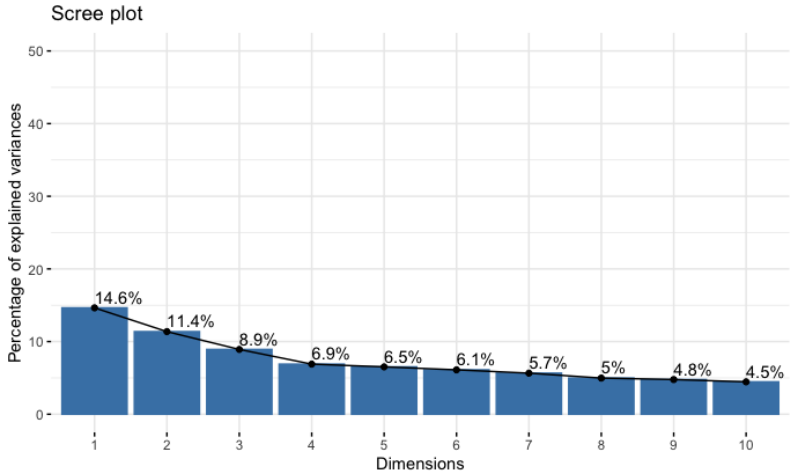
The casual drinker (cluster 1): Doesn't know specialty coffee or never drinks it. They mostly never change they brand of coffee and consume a large amount at home. Also looking at the polarizing directions, they do not value sustainability, fair trade factors etc. when making their purchasing decisions.

The social buyer (cluster 3): only drinks specialty coffee in cafes, changes brands sometimes/often and scores high on most criteria for coffee purchasing. They seem to drink the least amount of coffee and are closer to setting up an subscription than the casual drinker is.

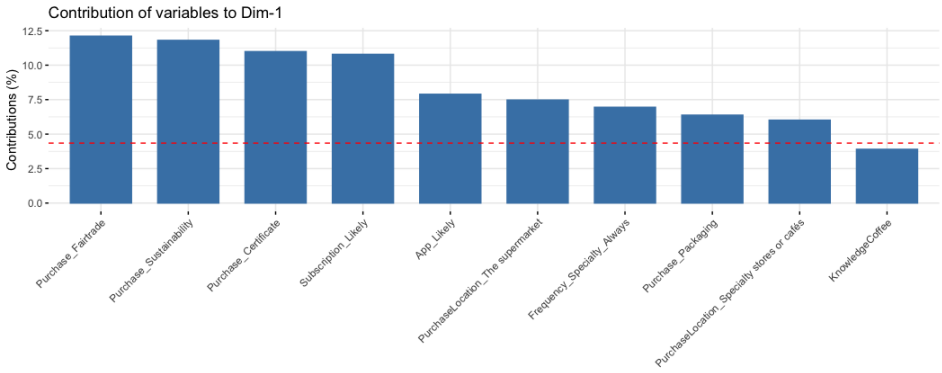
The coffee lover (cluster 2): This group clearly drinks the most amount of specialty coffee and is indeed the most likely to setup and subscription and an app. They also rarely buy from the supermarket and have the highest knowledge of coffee.



The scree plot below shows the percentage of variance explained by the amount of dimensions. The above only explains the 14,6% and the 11,4%. Nevertheless, we were able to capture valuable results.



The amount of variables in the biplot make it difficult to evaluate the length of the arrows. Therefore, below the % of contributions to explaining the variance, in dimension 1 is displayed. As can be seen the most variance between participants is captured by the purchasing criteria previously discussed: fair trade, sustainability and certificates. Moreover, subscription and app likeliness and purchasing from the supermarket coffee. For marketing purposes these are valuable results. Depending on who to target, it is clearly defined what distinguishes groups, what they value and how they behave between groups.



Correspondence Analysis

The next step is to evaluate the within group differences based on demographic variables. Hereby the response categories are discriminated by explanatory variables and the relationship analyzed. The response variables that I would like to evaluate most based on the explanatory variables are subscription likeliness and knowledge coffee. The objective is to discover what explanatory variables explain the variance in how likely people are to setup and subscription and their knowledge level.

In order to limit the data points included in the CA plot, several discrete variables are grouped together to make new points. For example, knowledge level has been decreased to the options: 1-2, 3-4, 5-6, 7-8 and 9-10. Then cross-tabulations of response variables and explanatory variables are made to ultimately analyze the distance between. Each plot has been re-scaled to improve the possibility of observing distinctions between groups. However, admittedly labels yet overlap, at times making it difficult to analyze results.

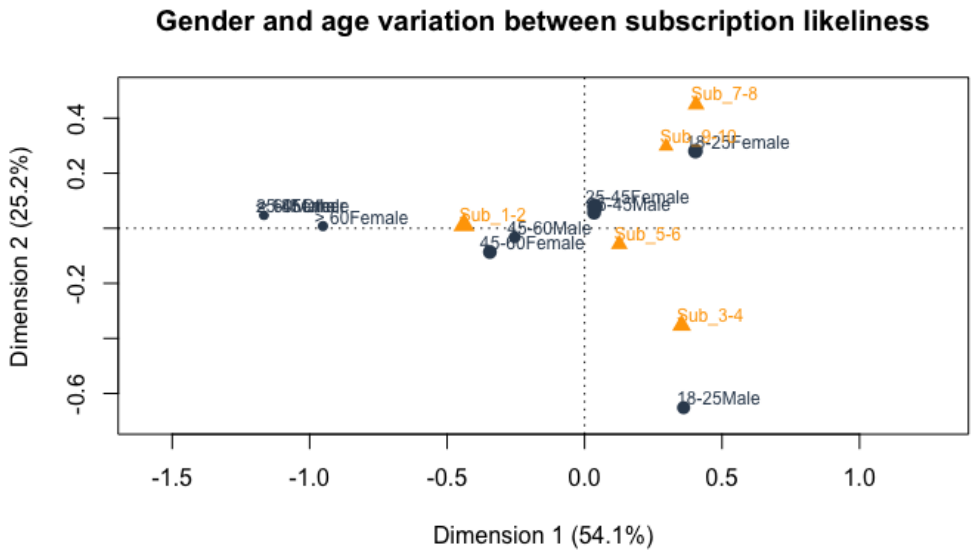
Gender and age variation between subscriptions likeliness

First I provide an example of one of the concatenated tables for this CA plot. Here we can already observe that there was no male below 18 that was a participant of the study. As previously mentioned, before using in production, the research should be conducted on a larger scale additionally to ensure to have a sample proportional to the population.

	Sub_1-2	Sub_3-4	Sub_5-6	Sub_7-8	Sub_9-10
< 18Female	2	0	0	0	0
> 60Female	6	0	1	0	0
> 60Male	4	0	0	0	0
18-25Female	11	12	4	13	6
18-25Male	6	14	4	1	1
25-45Female	24	12	13	10	5

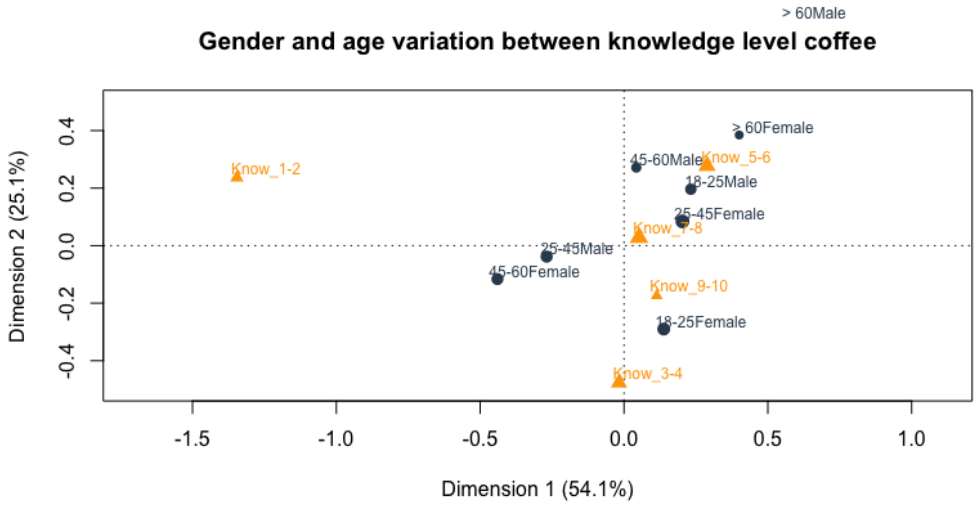
The first plot includes the variables gender and age based on subscription likeliness. The scale of the plot had to be adjusted in order to view each data point. The average appears to be subscription likeliness score of 5-6. The 25-45 year old male and females have expressed a similar view on this topic. Who stand out by providing the highest scores, are females between 18-25. However, males between this age have one of the lowest scores. These are quite interesting results where males and females from the same age group are on polar opposites sites. The least likely to setup an subscription are those over 45 years old for both males and females.

The inertia for this plot is 0.27. This measures the association between columns and rows. A high correlation is achieved by opposing those with strong opinions against those with moderate ones. In this case, this is a weak association suggesting that there might be other explanatory variables that could better explain the variance.



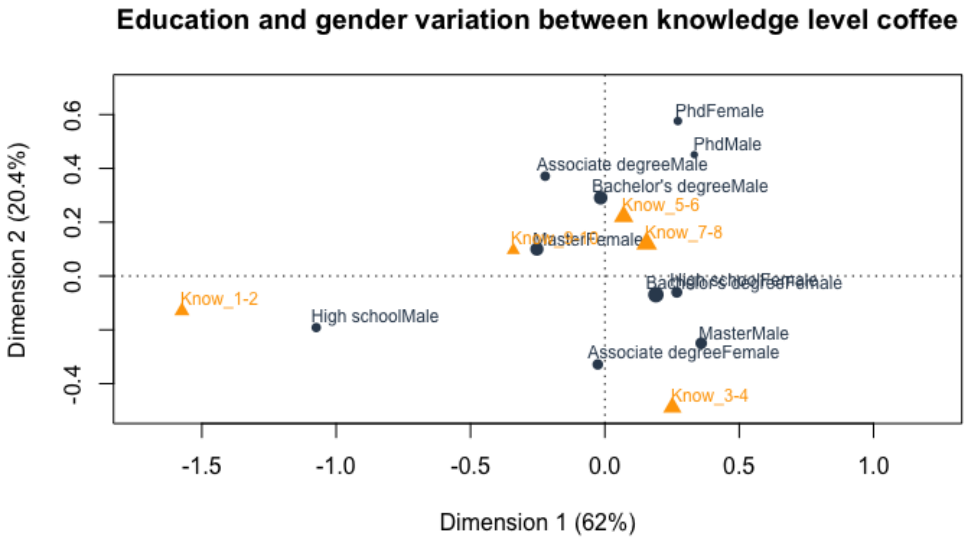
Gender and age variation between knowledge coffee

The same demographic features are selected to evaluate knowledge level of coffee. However, to analyze these relationships the scale excludes males above 60 as the distance is too far from the mean. This plot struggles to provide one clear pattern as increasing age group doesn't necessarily increase knowledge level. The lowest knowledge level is by 25-60 year old females and 25-45 males. The inertia of this plot is 0.30, slightly higher than the previous. Therefore, there is a moderate association and these explanatory variables slightly better explain the response variable knowledge.



Education and gender variation between knowledge coffee

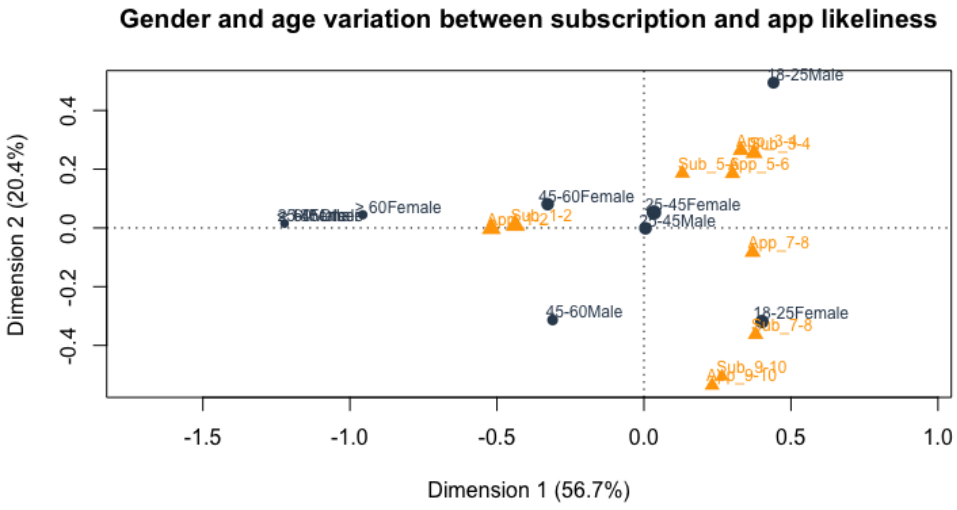
The inertia for this plot is one of the highest found with a value of 0.35 suggesting education better explains the variance on knowledge level. It is clear that high school males have the lowest knowledge level. However, other groups are closer together. The female obtaining a masters degree has the highest self-proclaimed knowledge level. However, it doesn't seem there is a linear relationship between the level of education and knowledge level. Otherwise, those with a Phd would have the highest level of knowledge regarding coffee.



Multiple Correspondence Analysis

Gender and age variation between subscription and app likeliness

When including both the variables subscription and app likelihood it can be observed that the answer to the first question is aligned with the answer of the second. This implies that if a participant is likely to setup an subscription they are also likely to setup and app no matter the age and gender group. Therefore, this plot yet shows similar results as plot 1 although flipped up side down. The 18-25 year old female is the most likely to set-up and app and subscription while those > 60 are the least likely. The association between these variables is weak with a inertia of 0.26.

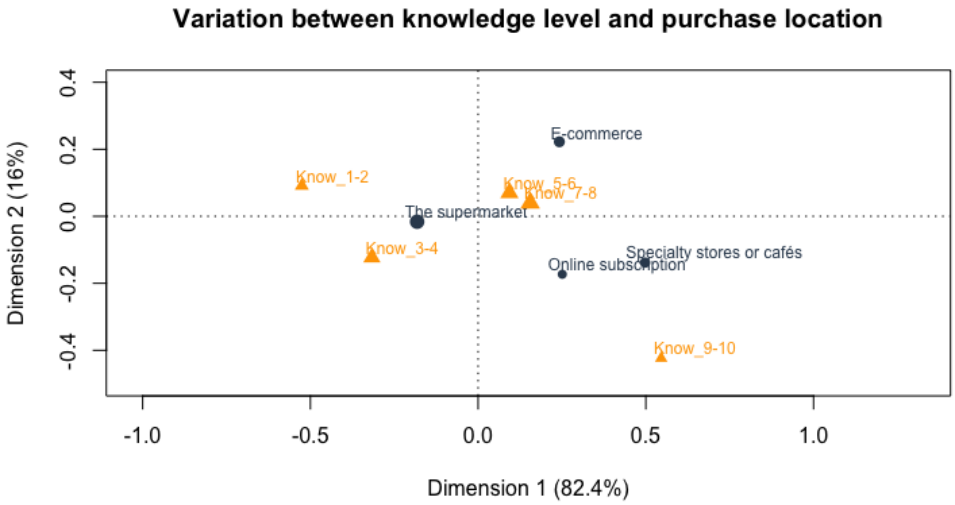


Variation between knowledge level coffee and purchase location

As expected, those with a lower knowledge level of coffee mostly are buying their coffee from the supermarket. Moreover, participants with a high knowledge level are buying from specialty coffee stores or cafes or have an online subscription. This suggests that when targeting consumers for a specialty coffee subscription, it might be valuable to look for those who have a substantial amount of knowledge regarding coffee. Nevertheless, a different approach

could be to educate those consumers that are currently buying from the supermarket to lead them to purchasing alternatives.

Important to note is that the inertia for this plot is only 0.08. This suggest that even though the direction of knowledge vs purchase location is as expected, the association between the two is very weak.



Discussion and conclusions

The results found throughout this research project are extremely valuable for marketing objectives. Not only does it provide an opportunity to target consumers based on specific characteristics, it additionally allows to understand the behavior better. The k-means clustering provides objectively distinguished groups whereas in previous marketing projects I tried to apply segmentation based on personal views. The principal component analysis visualized the distance between cluster based on the variables included in the clustering analysis. Therefore, profiling the buyer personas became much easier.

As stated at several moments, I would do this research on a large scale, perhaps introducing different variables that could explain the behaviors of the consumer. Moreover, due to time considerations I wasn't able to go more in-to-depth regarding plot analysis such as the correspondence analysis.

In further research I would try to incorporate digital data such as page viewings to apply clustering approaches. Naturally, conducting a survey requires a lot of resources and being able to automatically update the buyer personas, would improve the ability to use them in practice.

References

References DataCamp. 2021. PCA-Analysis-R. [online] Available at: https://www.datacamp.com/community/tutorials/pca-analysis-r [Accessed 1 July 2021].

Datanovia. 2021. Cluster Validation Statistics: Must Know Methods - Datanovia. [online] Available at: https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/ [Accessed 1 July 2021].

Datanovia. 2021. Determining The Optimal Number Of Clusters: 3 Must Know Methods - Datanovia. [online] Available at: http://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/#at_pco=wnm-1.0&at_si=609664423560aa01&at_ab=per-2&at_pos=0&at_tot=1 [Accessed 1 July 2021].

Essentials, P., 2021. PCA - Principal Component Analysis Essentials - Articles - STHDA. [online] STHDA. Available at: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/ [Accessed 1 July 2021].

Medium. 2021. Clustering Analysis in R using K-means. [online] Available at: https://towardsdatascience.com/clustering-analysis-in-r-using-k-means-73eca4fb7967 [Accessed 1 July 2021].

Medium. 2021. K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. [online] Available at: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a [Accessed 1 July 2021].

Vertica. 2021. Finding the “K” in K-means Clustering With a UDx | Vertica. [online] Available at: https://www.vertica.com/blog/finding-the-k-in-k-means-clustering-with-a-udf/ [Accessed 1 July 2021].