Kombucha Survey Report

Federica Gazzelloni

2024-03-26

Overview

`Kombucha` is a fermented tea beverage that has gained popularity in recent years due to its potential health benefits and unique flavor profile. It is made by fermenting sweetened tea with a symbiotic culture of bacteria and yeast (SCOBY), which results in a tangy, slightly effervescent drink.

Customers who are more likely to try kombucha typically exhibit certain characteristics or preferences that align with the product's attributes and perceived benefits:

1. **Health-conscious Individuals**
2. **Adventurous Consumers**
3. **Environmentally Conscious Shoppers**
4. **Younger Demographic**
5. **Well-educated and Higher-income Individuals**

On the other hand, customers who are less likely to try kombucha may include:

1. **Traditional or Conservative Consumers**
2. **Skeptical or Risk-averse Individuals**
3. **Taste-sensitive Individuals**

Overall, understanding the characteristics and preferences of different customer segments can help businesses target their marketing efforts and product positioning to appeal to the right audience for kombucha. By identifying and appealing to the demographics and psychographic profiles of customers who are more likely to try kombucha, companies can maximize their market penetration and sales potential.

Introduction

As [brand] embarks on its next phase of growth, it is imperative to make informed decisions driven by data insights. To understand and target [brand]'s current and potential future audiences effectively, a comprehensive analysis of customer segments based on values and behaviors relevant to `kombucha consumption` is crucial.

Here is the result of the analysis of a set of survey data targeting a representative sample of 1,000 U.S. adults interested in health and trying new things.

Objective

The primary objective of this analysis is to identify segments of potential customers based on their mindsets and attitudes, as reflected in the responses to `51 psychographic survey` questions. Statistical clustering techniques are used to identify customer segments that exhibit similarities in their preferences, values, and behaviors related to kombucha consumption.

- group like-minded people together based on their mindsets and attitudes, not their demographic profile
- which type of customers are more or less likely to try `kombucha` and why

Data

Data is made of a set of responses related to `kombucha` propensity:

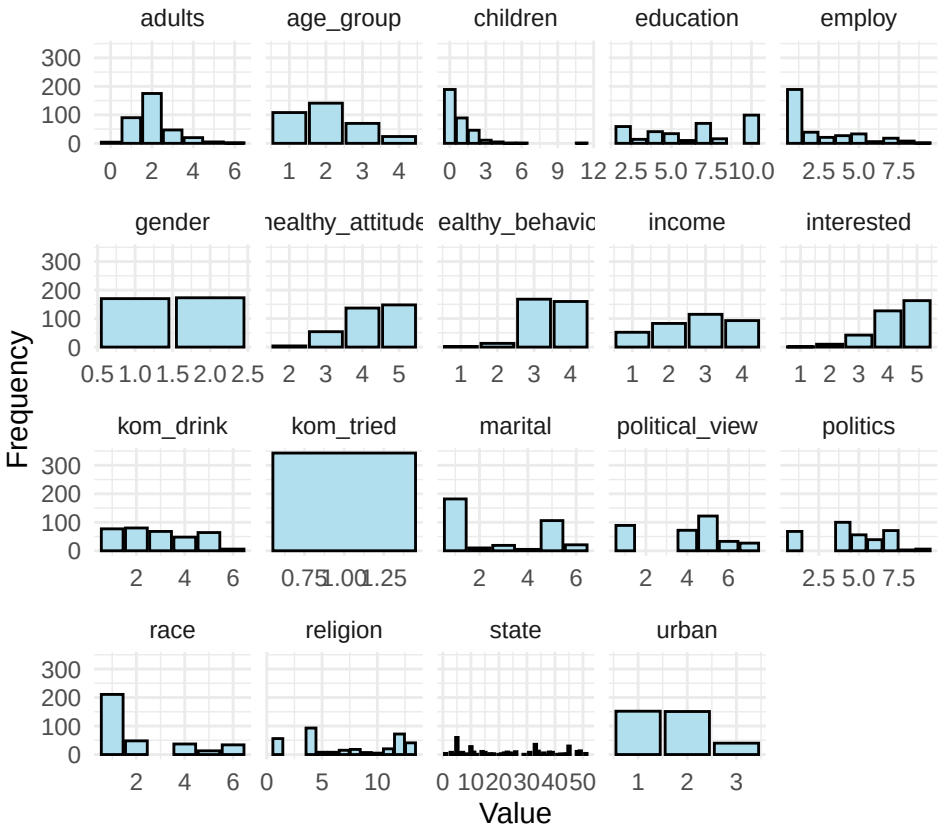
d_interested	d_kom_tried	d_kom_drink	d_kom_aware
Somewhat interested	Yes	Almost never	I don't know
Very interested	Yes	Often (a few times a month)	Kirkland
Very interested	Yes	Often (a few times a month)	Humm Kombucha
Very interested	Yes	Often (a few times a month)	Brew
Somewhat interested	Yes	Often (a few times a month)	kevita, gt's
Extremely interested	Yes	Often (a few times a month)	none

Survey Demographics

General demographics survey, including health attidudes:

```
 [1] "urban"            "gender"           "age_group"        "race"
 [5] "education"        "income"           "adults"           "children"
 [9] "marital"          "politics"         "political_view"   "employ"
[13] "religion"         "state"            "healthy_attitude" "healthy_behavior"
[17] "interested"
```

Here is a visualization of the distribution of the demographics variables, which help identify the structure of the interviewed sample population.

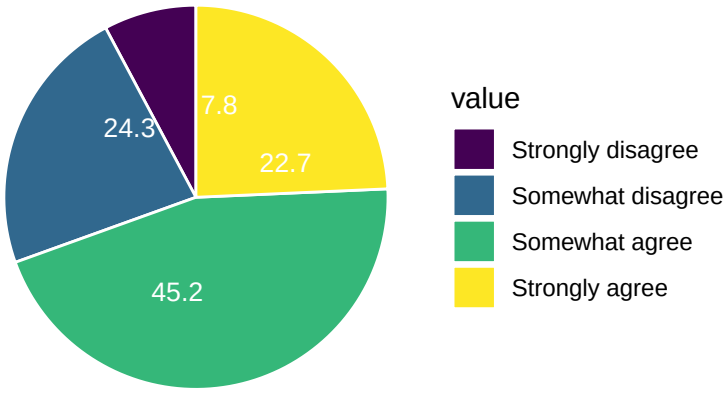


Survey Psychographic

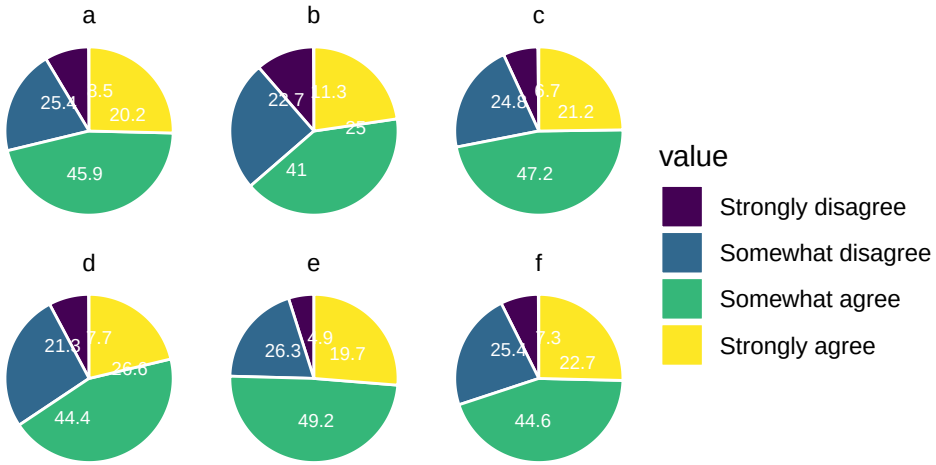
The psychographic segments are composed of a variety of different questions with responses varying from strongly disagree to strongly agree.

The extract psychographic variables for cluster frequencies and the overall agreement results show that 45.2 of interviewed people have answered `somewhat agree`, while `somewhat disagree` and `strongly agree` are 23% and 24% respectively.

value	frequency	percent
Strongly disagree	3994	7.8
Somewhat disagree	11572	22.7
Somewhat agree	23045	45.2
Strongly agree	12369	24.3



Agreement results based on type of question group (a, b, c, d, e, f)



Likert Plots

In the following graphical representations, used to visualize responses collected with Likert scale questions to measure attitudes, opinions, or perceptions of respondents towards Kombucha. The first plot is representing the overall tendency in agreement/disagreement to all questions. Overall results are positive.

General positive and negative responses to all questions.

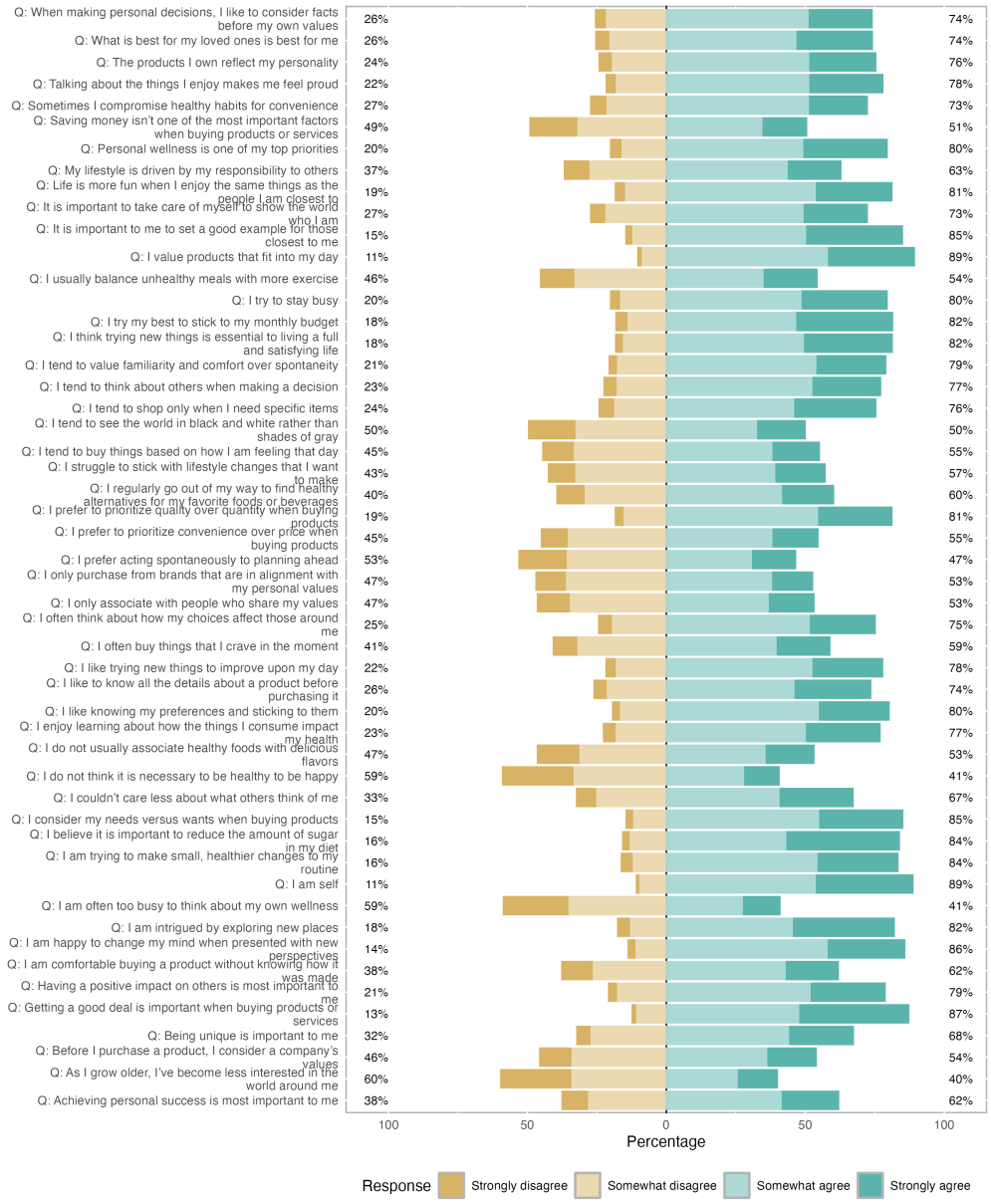


Figure 1: General positive and negative responses

Focusing attention on **kombucha tried and drink** responses based on a set of tailored questions relative to the attitudes of the respondents when buying new things, it can be clearly seen the positive agreements showing the sample population is keen to test and buy new things.

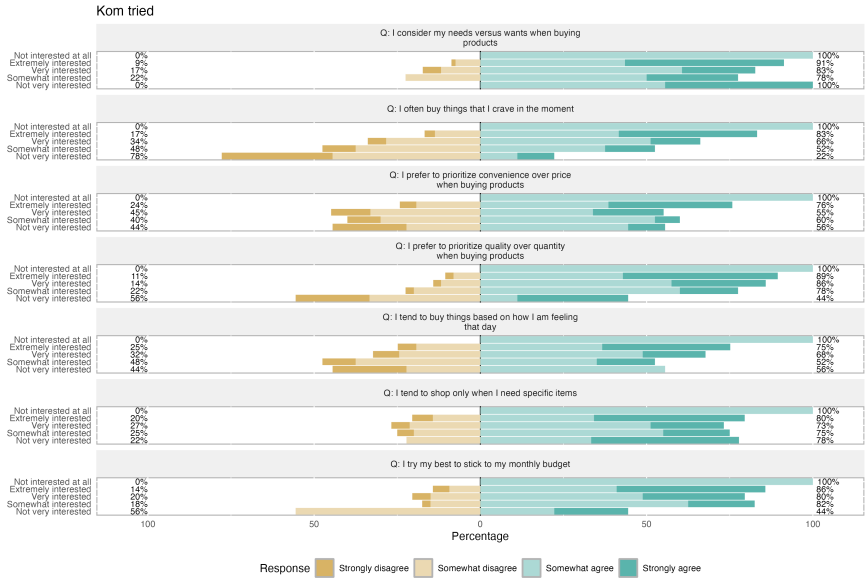


Figure 2: Kombucha tried

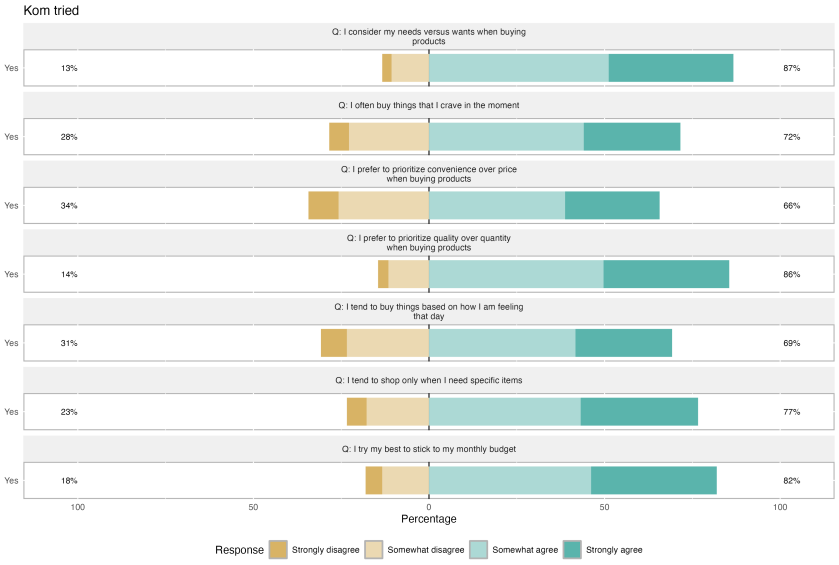


Figure 3: Kombucha tried

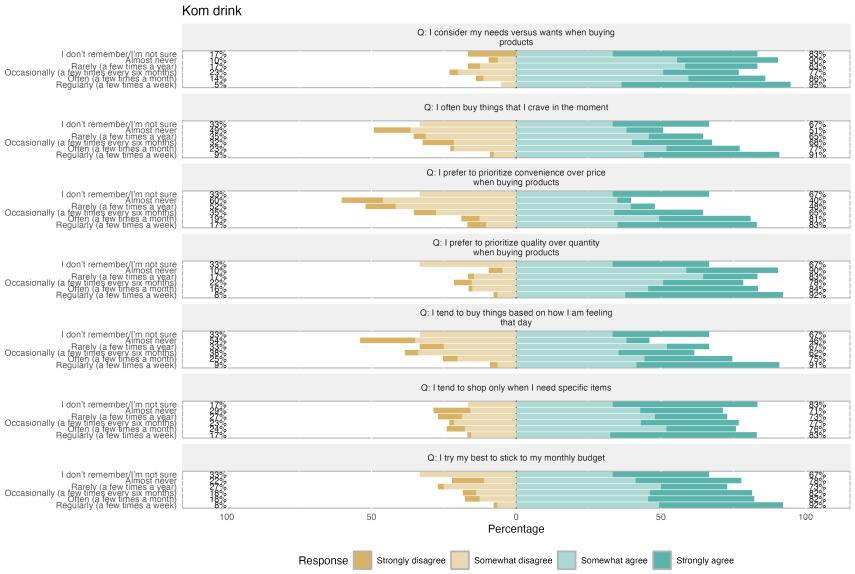


Figure 4: Kombucha drink

Customer general tendency in buying new things

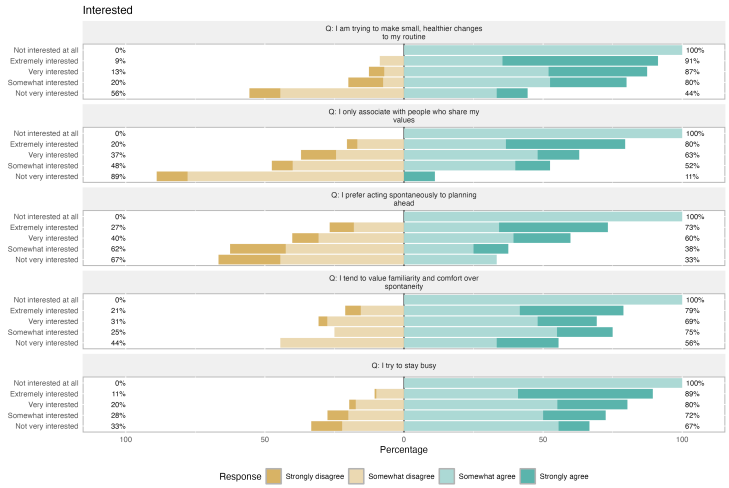


Figure 5: Interested

How strongly correlated are the responses

In this case the Cronbach's $\alpha = \frac{N\bar{c}}{\bar{v}+(N-1)\bar{c}}$ metrics are used to measure the internal consistency or reliability of the responses, used to show how strongly correlated they are. (If the values are low (below .7) it indicates the questions are not internally consistent). Other metrics that can be used are Guttman's (G6) λ, and Omega ω.

The data shows a very good $\alpha = 0.94$ internal consistency.

```
Reliability analysis
 raw_alpha std.alpha G6(smc) average_r S/N    ase mean  sd median_r
      0.93      0.94    0.95      0.22  15 0.0029  2.9 0.4     0.23

                raw_alpha std.alpha   G6(smc) average_r     S/N   alpha se
seg_battery_b_6 0.9349852 0.9363033 0.9496865 0.2271953 14.6994 0.00293265
                     var.r    med.r
seg_battery_b_6 0.01002374 0.230051
```

Scale structure

Information about this scale

Dataframe:	psychographic_data[qs[1:51]]
Items:	all
Observations:	980
Positive correlations:	1254
Number of correlations:	1275
Percentage positive correlations:	98

Estimates assuming interval level

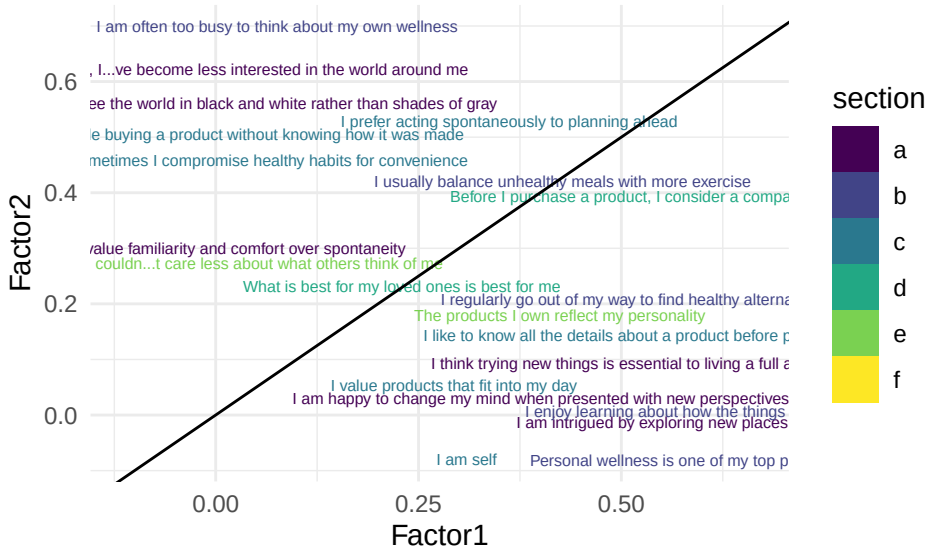
Omega (total):	0.95
Omega (hierarchical):	0.70
Revelle's Omega (total):	0.95
Greatest Lower Bound (GLB):	NA
Coefficient H:	0.94
Coefficient Alpha:	0.93

(Estimates assuming ordinal level not computed, as the polychoric correlation matrix has missing values.)

Note: the normal point estimate and confidence interval for omega are based on the procedure suggested by Dunn, Baguley & Brunsden (2013) using the MBESS function ci.reliability, whereas the psych package point estimate was suggested in Revelle & Zinbarg (2008). See the help ('?ufs::scaleStructure') for more information.

Factor analysis

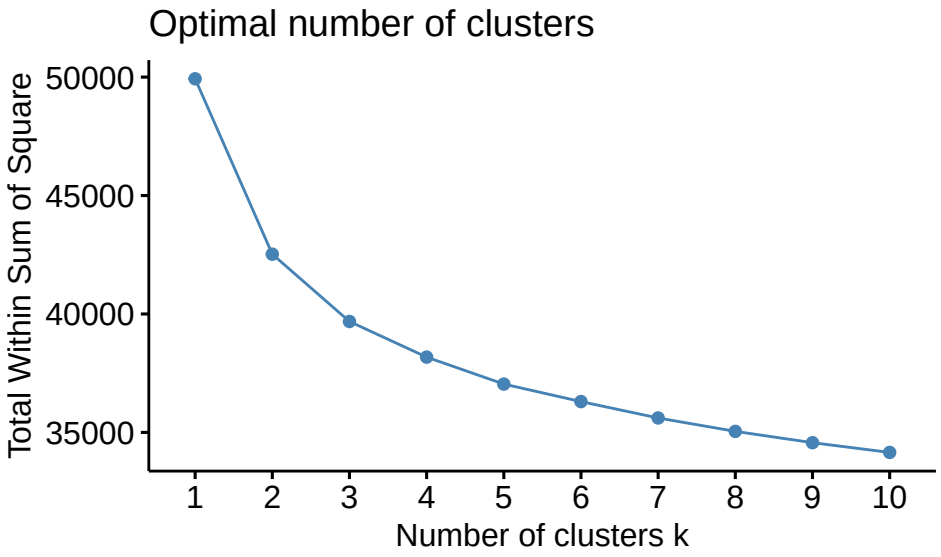
Other interesting thing to see is the application of factor analysis to the segments questions. Results show two main groups of questions.



Segmentation Analysis

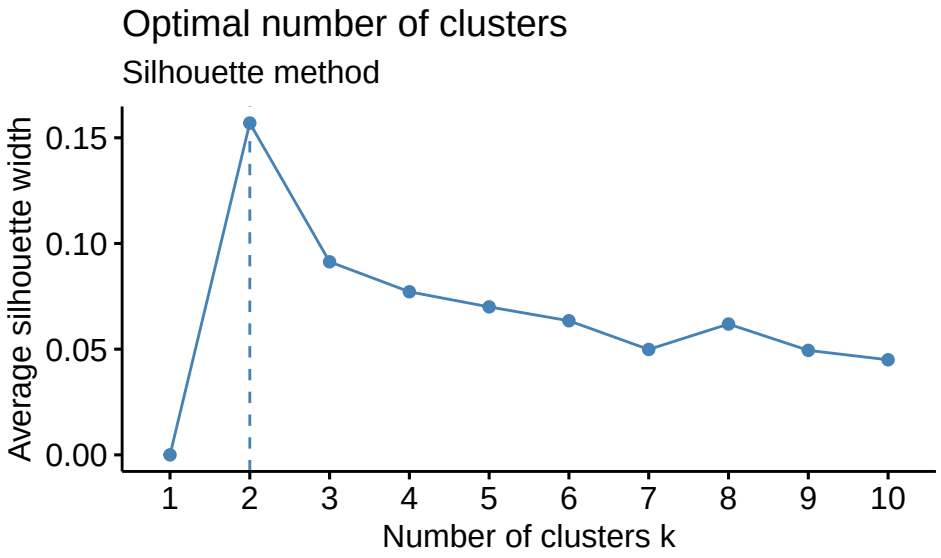
Finally, utilizing **k-means clustering** for partitioning data into distinct clusters, we can identified homogeneous groups of respondents based on similarities in their psychographic responses. The optimal number of clusters was determined through empirical methods such as the elbow method or silhouette method.

Determine the optimal number of clusters with the **Elbow method:**



There is no sharp elbow. For this reason, it draws attention as a result open to interpretation. An interpretation based on this elbow may therefore lead to incorrect results. It can be inferred that the optimal number of clusters is two.

Determine the optimal number of clusters with the **Silhouette method**:

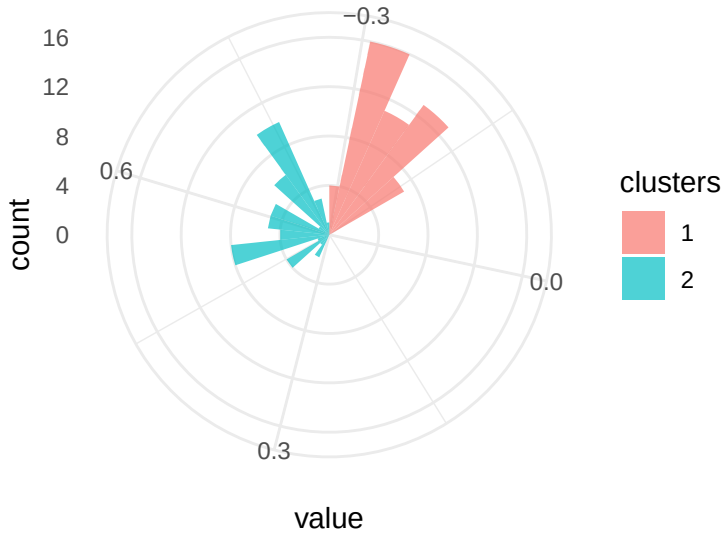


Another option as a further check of the most suitable number of clusters is the **Gap Statistics**.

K-means calculation

Start considering K=2, two clusters and extract the cluster assignments.

Clusters Visualization



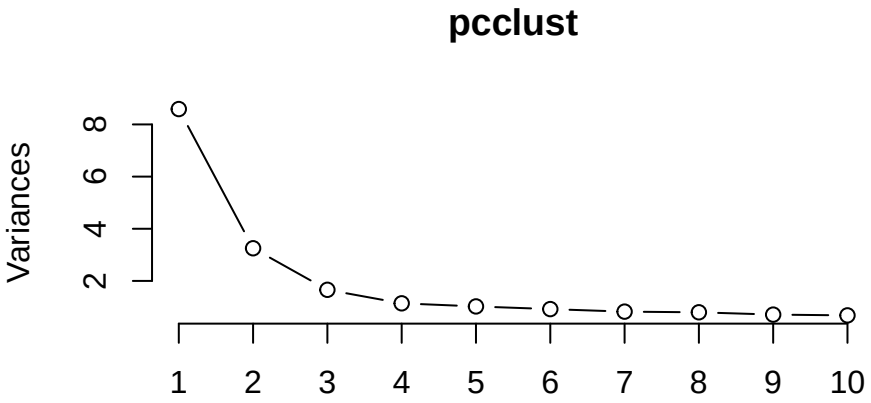
Demographic Analysis

We analyzed demographic differences between the identified clusters to understand how customer segments differ in terms of demographic characteristics such as age, gender, income, and geographic location.

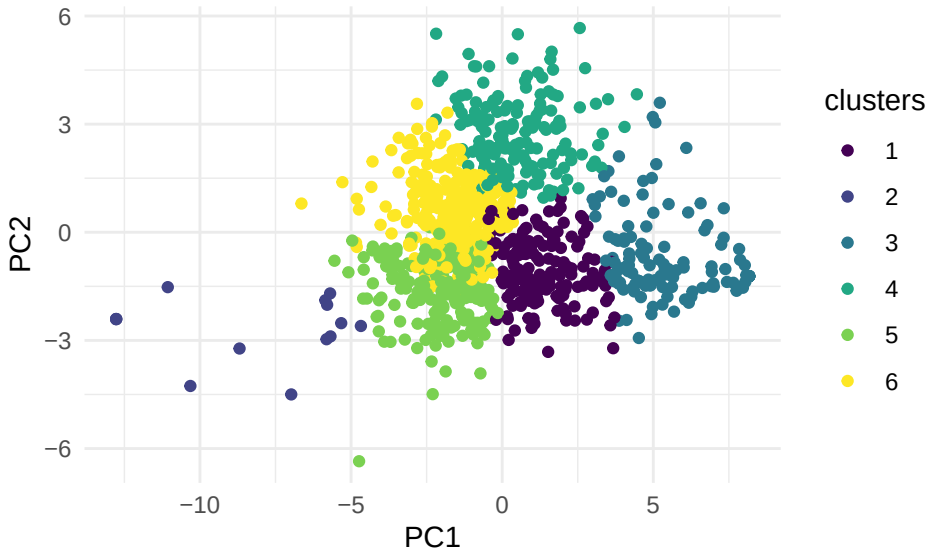
Principal Component Analysis (PCA)

Let's now increase the number of k-means clusters to 6, and calculate the PCA components to see in which group the highest variability is located. In case of dimentionality reduction need a subgroup can be selected to easily identify the behavioral that lead to the highest variability.

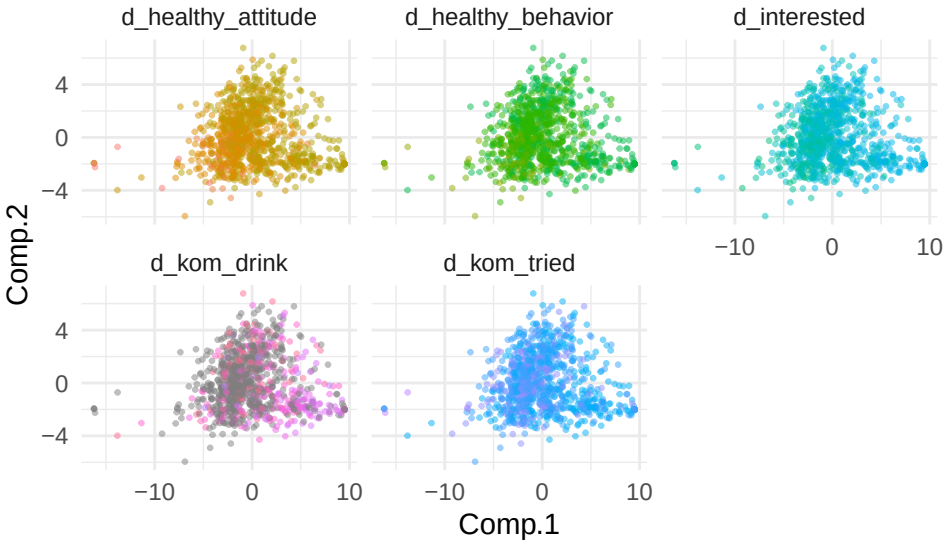
The max number of components that summarize the highest level of variability in the data is 2.



Finally, here is a visualization of the two components across the 6 clusters, we can see that the highest level of variability corresponds to clusters 3 and 4. Things changes if we focus only on two clusters.



A quick look at the Kombucha Demographics



Further analysis

We visualized the clustering results using a scatter plot, allowing for a clear understanding of how respondents are grouped into distinct segments based on their psychographic profiles.